Intelligent Data Analysis 7 (2003) 181–192
IOS Press

Characterising Data Mining software

C. Giraud-Carrier and O. Povel
ELCA Informatique SA, Av. de la Harpe, 22-24, CH-1000 Lausanne 13, Switzerland
Tel.: +41 21 613 21 11; Fax: +41 21 613 21 00; E-mail: cgc@elca.ch

Received 13 September 2002
Revised 19 October 2002
Accepted 29 October 2002

Abstract. The ever-increasing number of fielded Data Mining applications is evidence that the technology works and produces added value in a variety of business areas. Most of the research-lab generated algorithms have found their way under various guises in a number of commercial software packages. When considering the use of Data Mining, the average business user is now faced with a plethora of DM software to choose from. In order to be informed, such a choice requires a standard basis from which to compare and contrast alternatives along relevant, business-focused dimensions, as well as the location of candidate tools within the space outlined by these dimensions. This paper aims at meeting this business requirement. It presents a standard schema for the characterisation of Data Mining software tools and the results of a recent survey of 41 popular Data Mining tools described within this schema.

Keywords: Data mining, tool characterisation, tool comparison, business decision support

1. Introduction

It has been argued that all one needs to engage in Data Mining (DM) is data and a willingness to "give it a try". Although this view is attractive from the perspective of enthusiastic DM consultants who wish to expand the use of the technology, it can only serve the purposes of one-shot proofs of concept or preliminary studies. It is not representative of the reality of deploying DM within existing business processes. In such contexts, one needs two additional ingredients: a process model or methodology, and supporting tools.

Several Data Mining process models have recently been developed [2,4,10,12,18]. Although each sheds a slightly different light on the process, their basic tenets and overall structure are essentially the same (e.g., see [13]). A recent survey suggests that virtually all practitioners follow some kind of process model when applying DM and that the most widely used methodology is CRISP-DM [15].

In this paper, we therefore focus on the second ingredient, namely, supporting tools. The past few years have seen a proliferation of DM software packages. Whilst this makes DM technology more readily available to non-expert end-users, it also creates a critical decision point in the overall business decision-making process. When considering the application of Data Mining, business users now face the challenge of selecting a tool adequate to their needs and expectations.

As good decisions rely on relevant and accurate information, we describe a template for the characterisation of DM software along a number of complementary dimensions, together with a dynamic database of 41 of the most popular DM tools as of February 2003. Ours is clearly not the only such work, nor is the list of tools exhaustive. A section of this paper is dedicated to a survey of most other relevant related work.

However, our contribution distinguishes itself in two sigificant ways:

1088-467X/03/$8.00

1. *Business Orientation* The target audience of our work is business decision-makers. The proposed characterisation and accompanying database emphasise the complete Data Mining process and are intended to provide the basis for informed, business-driven tools comparison and selection. As such, we have attempted to motivate each of the characteristics with a reality-action scenario, where the reality states a fact about the application of DM and the action binds this reality to a precise activity in the decision-making process.
2. *Timeliness* The information in our dynamic database is up-to-date as of February 2003 and will be maintained through time. This is particularly relevant in the very dynamic DM software industry where acquisitions, mergers and product upgrades have been frequent. Most related studies seem to be at least two years out-of-date.

The paper is organised as follows. Section 2 briefly summarises relevant work. Section 3 presents an intuitive view of our proposed characterisation for DM tools. An appendix contains the more formal database schema used in our survey of existing tools. Section 4 lists the DM tools currently included in our dynamic database. Finally, Section 5 concludes the paper.

2. Previous studies

This section briefly outlines, in chronological order, some of the most relevant work on DM tool characterisation and evaluation.

Catenate Consulting has compiled a very thorough list of criteria to define data mining tools, organised, as ours is, along the main phases of the data mining process (i.e., data management/pre-processing, exploration/visualisation, data mining operations, interpretation and deployment) as well as business-related information [5]. However, they do not provide any specific tool evaluation.

Information Discovery, Inc. published in 1997 a taxonomy of data mining techniques with a short list of products for each category [21]. The focus was restricted to implemented DM algorithms.

Elder Research, Inc. produced in 1998 two lists of commercial desktop DM products (one containing 17 products and the other only 14), defined along a few, yet very detailed, dimensions [9,16].

Another 1998 study, described in [13], contains an overview of 16 products, evaluated against pre-processing, data mining and post-processing features, as well as additional features such as price, platform, release date, etc. The originality of this study is its very interesting application of multidimensional scaling and cluster analysis to position 12 of the 16 evaluated tools in a four-segment space.

The Laboratoire de Recherche en Informatique evaluated in 1999 a list of 42 products along a large number of dimensions [11]. Unfortunately, the resulting matrix is rather sparse, with most information missing for a number of tools.

In 1999, the Data & Analysis Center for Software (DACS) also released one of its state-of-the-art reports, consisting of a thorough survey of data mining techniques, with emphasis on applications to software engineering, which includes a list of 55 products with both summary information along a number of technical as well as process-dependent features and detailed descriptions of each product [19].

Another study, including a list of 21 products, defined mostly by the algorithms they implement and a few additional technical dimensions, was compiled by Exclusive Ore, Inc. in 2000 [8]. A small study offering a simple “OK-Good-Excellent” rating of 10 products for the techniques they implement, their visualisation capabilities and their perceived usability is also found in [1].

Finally, it is worth mentioning a number of lists of DM tools that, although not including any characterisation/evaluation, provide an useful starting point for tool evaluation and selection exercises by centralising (and generally maintaining in time) basic information for each tool and links to the vendor’s homepage for further details (e.g., see [6,7,17]).

3. A hierarchical view

This section describes, rather intuitively, our business-oriented proposal for the characterisation of DM tools.

The characteristics proposed for the description of DM tools are organised naturally in a hierarchical fashion. At the top level, we define the following general categories:

– *Business Goal*: what kinds of business problems can the selected tool solve?
– *Model Type*: what classes of DM models does the selected tool implement?
– *Process-dependent Features*: how well and with what methods does the selected tool support the overall DM process?
– *User Interface Features*: how easy is it to use and interact with the selected tool?
– *System Requirements*: what specific system requirements must be met for the selected tool to be deployed?
– *Vendor Information*: who is behind the selected tool? how much do they sell the tool for? what kind of support can they offer? etc.

Each of these categories is further motivated and detailed in the following subsections.

3.1. Business goal

We define Data Mining as a *business-driven process* aimed at the *discovery* and *consistent use* of *profitable* knowledge from corporate data.

– *Reality*: From a commercial standpoint, knowledge obtained through DM must add value to the business.
– *Action*: It is essential that users articulate a specific, preferably measurable objective.

Within our proposed schema, the following classes of business goals are defined:

– *Customer Acquisition*: targeting those prospects most likely to become customers.
– *Cross- and Up-selling*: targeting those customers most likely to buy more/respond positively to a specific offer.
– *Product Development*: segmenting and profiling customers and prospects to define new relevant offers of services/products.
– *Churn Prediction*: predicting how likely an existing customer is to take his/her business to a competitor.
– *Fraud Detection*: detecting fraudulent activities, such as credit card fraud, abusive insurance claims, etc.
– *Market-basket Analysis*: looking for associations among data items (e.g., products bought, activities performed).
– *Risk Assessment*: determining the risk level associated with a particular decision (e.g., granting a mortgage, underwriting some insurance cover).
– *Prediction/Forecasting*: predicting future trends/behaviours (e.g., toxicology, stock market, web site usage).
– *Outcomes Measurement*: examining clinical encounter information, insurance claims and billing data to measure the results of past treatments and processes, so as to capitalise on existing strengths and identify areas of improvement.

– *Condition Monitoring*: capturing the "normal" behaviour of plant equipment (e.g., rotating machinery such as gas turbines), raising alarms only when significant deviations occur and remedial action is required.
– *Public Health Monitoring*: detecting significant deviations from the norm (e.g., infectious outbreaks, bio-terrorist attacks)
– *Discovery*: seeking "new" insight by exploring data in a semi-directed way.

3.2. Model type

The generative aspect of Data Mining consists of the building of a model from data. There are many available (machine learning) algorithms, each inducing one of a variety of types of models.

– *Reality*: Different business objectives call for different modelling approaches.
– *Action*: Users must match their business objective to an adequate model type.

The following model types are standard in DM and several algorithms exist for each type:

– *Predictive Modelling*: mapping a set of "input" values (i.e., the independent variables) to an "output" value (i.e., the dependent variable). This takes two forms depending on the type of the output, as follows.

  ∗ *Classification:* learning a function that associates with each data object one of a finite number of pre-defined classes (e.g., rock-mine discrimination).
  ∗ *Regression:* learning a function that maps each data object to a real value (e.g., propensity to buy).

– *Descriptive Modelling* (*clustering/segmentation*): discovering groups or categories of data objects that share similarities and help in describing the data space (e.g., customer segments).
– *Dependency Modelling*: learning a model that describes significant associations or dependencies among features (e.g., contents of subscription orders, market baskets).
– *Anomaly/Change Detection*: detecting the most significant deviations from previous measurements/behaviour or norms (e.g., network intrusion, fraud).
– *Time-series Analysis*: detecting patterns or trends in time-dependent data (e.g., nosocomial infections).

The dependency between model types and business objectives is summarised in Table 1, where a 'Y' indicates that the corresponding model type is suitable for the corresponding business objective.

3.3. Process-dependent features

As stated above, Data Mining is a business *process*. As such, it is iterative by nature and involves a number of complementary activities.

– *Reality*: Business environments differ and DM tools differ in their functionality.
– *Action*: Users must ensure that the tool they choose supports their needs, standards and practices.

The following aspects cover the overall DM process:

– *Input Formats*: Users' data generally originates in a number of heterogeneous databases in different formats. To avoid unnecessary "interfacing" work, users should strive to select a tool able to read their various data formats. Most commonly supported formats/connectivity include:

Table 1
Business objectives vs model types

	Predictive modelling		Descriptive	Dependency	Anomaly/Change	Time-series
	Classification	Regression	modelling	modelling	detection	anlysis
Customer acquisition	*Y*	Y				
Cross- and up-selling	Y	Y	Y	Y		
Product development	Y	Y	Y			
Churn prediction	Y	Y	Y			
Fraud detection	Y	Y	Y		Y	Y
Market-basket analysis				Y		
Risk assessment	Y	Y				
Outcomes measurement	Y	Y		Y		
Condition monitoring					Y	Y
Public health monitoring					Y	Y
Discovery	Y	Y	Y	Y	Y	Y

  * Flat File (e.g., comma-separated)
  * ODBC/JDBC
  * SAS
  * XML

- *Pre-processing*: Raw data is seldom readily usable for DM. Instead, before it can be further analysed, it may be usefully enriched and transformed, e.g., to add new features or to remove noise. Users must not underestimate the need for pre-processing and data preparation. Experience suggests that such activities represent the bulk of the DM process, reaching as much as 80% of the overall process' time. It is thus critical to ensure that the tool selected is able to assist effectively, both from an usability point of view and from a functionality perspective. Pre-processing techniques can be broadly categorised as follows.

  * *Data Characterisation:* statistical and information theoretic measurements such as mean value, mutual entropy, kurtosis, etc.
  * *Data Visualisation:* graphical representations to "eye-ball" underlying distributions, data spread, etc.
  * *Data Cleaning:* automatic detection of undesirable conditions, such as outliers, single-valued attributes, identifier attributes (i.e., key attributes with as many values as there are records), records or attributes with too many missing values, etc.
  * *Record Selection:* mechanisms to select specific records (or rows) based on *ad-hoc* queries.
  * *Attribute Selection:* mechanisms to select specific attributes (or columns) based on *ad-hoc* queries.
  * *Data Transformation:* mechanisms to transform data such as discretisation, definition of new attributes, thresholding, etc.

- *Modelling Algorithms*: There are a number of algorithmic techniques available for each DM approach, with features that must be weighed against data characteristics and additional business requirements. Users must ensure that their chosen tool implements algorithms that meet their needs. The following are the standard classes of DM modelling algorithms.

  * Decision Trees
  * Rule Learning
  * Neural Networks

Table 2
Modelling algorithm classes vs model types

	Predictive modelling		Descriptive	Dependency	Anomaly/Change	Time-series
	Classification	Regression	modelling	modelling	detection	analysis
Decision trees	Y	(Y)				
Rule learning	Y					
Neural networks	Y	Y	(Y)		Y	Y
Linear/logistic regression		Y				
Association learning				Y		
Instance-based learning	Y		Y		Y	
Unsupervised learning			Y		Y	
Probabilistic learning	Y				Y	

∗ Linear/Logistic Regression
∗ Association Learning
∗ Instance-based/Nearest-neighbour Learning
∗ Unsupervised Learning
∗ Probabilistic Learning

Note that there is a natural dependency between modelling algorithm classes and model types, as each modelling algorithm class is targeted at a specific type of model. This dependency is summarised in Table 2. ('(Y)' indicates that some algorithms in the class support the corresponding model type).

– *Evaluation/Interpretation*: DM will produce results. These, however, are not always actionable or profitable. Ultimately, the "buck stops with the business users." Hence, users must be able to "see" enough of what they need in order to assess its potential value to the business. The following are the standard methods of DM results evaluation.

∗ Hold-out/Independent Test Set
∗ Cross-validation
∗ Lift/Gain Charts
∗ ROC Analysis
∗ Summary Reports
∗ Model Visualisation

– *Dissemination/Deployment*: In order to "close the loop," knowledge acquired through Data Mining must become part of business-as-usual. Hence, users must decide on what and how they wish to disseminate/deploy results, and how they integrate DM into their overall business strategy. The selected tool, or custom implementation, must support this view. Available methods are as follows.

∗ Save/Reload Models
∗ Produce Executable
∗ PMML/XML Export
∗ Comment Fields

– *Scalability*: It is not unusual to be dealing with huge amounts of data (e.g., bank transactions) and/or to require near real-time responses (e.g., equipment monitoring). In such cases, users must consider how well the selected tool scales with the size of their data sets and their time constraints. Three features capture this issue.

∗ Data Set Size Limit

  * Support for Parallelisation
  * Incrementality (of the modelling algorithms)

- *Miscellaneous*: In addition to the "core" elements described above, several tools offer a number of special features that facilitate or enhance the work of users. We consider the following in our characterisation.

  * Expert Options
  * Batch Processing

3.4. *User interface features*

Most of the algorithms (both for pre-processing and modelling) used in Data Mining originate from research labs. Only in the past few years have they been incorporated in commercial packages.

- *Reality*: Users differ in skills, and DM tools vary greatly in their type and style of user interaction.
- *Action*: Users must think of their DM target audience and be sure to select a tool that is adapted to the skill level of such users.

In this dimension, we focus on three characteristics:

- *Graphical Layout*: does the selected tool offer a modern GUI or is it command-line driven?
- *Drag&Drop/Visual Programming*: does the selected tool support simple visual programming based on selecting and sequencing icons on the screen?
- *On-line Help*: does the tool include any on-line help? how much of it and of what quality for the unexperienced user?

3.5. *System requirements*

Software tools execute in specific computer environments. It is therefore important for users to be aware of the specific requirements of their selected tool. The following characterises such requirements.

- *Hardware Platform*: PC, Unix/solaris workstation, etc.
- *Software Requirements*: DB2, SAS Base, Oracle, Java/JRE, etc.
- *Software Architecture*: standalone, client/server, thin client.

3.6. *Vendor information*

In addition to, or sometimes in spite of, the technical aspects described above, there are important commercial issues which must be addressed when selecting a software package.

The following characteristics attempt to summarise these issues.

- *Contact Details*: how may the tool vendor/distributor be contacted?

  * Headquarters' Phone Number
  * Company's URL
  * General Inquiry's Email Address

- *Level of Technical Support*: how well is the vendor/distributor able to carry out such activities as maintenance, upgrades, technical support, help desk, etc.?
- *Price*: how much does the tool cost (for 1 license on 1 machine)?

  * Freeware
  * Contact Vendor
  * Price Range

- *Free Demo/Evaluation Version*: is it possible to download (or otherwise obtain) a free evaluation version of the software?
- *Market Penetration*: how much of the DM tool market does the vendor hold (i.e., number of customers/sites)?
- *Longevity*: how long has the vendor been in business? how likely is it to continue?

4. Putting it all to work

Using the above schema, we have already characterised the following 41 popular DM tools. Note that when a vendor offers more than one tool, these have been lumped together to avoid duplication. Except for Weka and SwissAnalyst, which are licensed under the GNU GPL license and are thus freeware, all other tools are commercial.

- Affinum Model
- Alice / AC2
- AnswerTree
- BusinessMiner
- CART / MARS
- Clementine
- Critical Action Software
- DataDetective
- Data Mining Suite
- Datascope
- DD Series 3.X
- DecisionHouse
- Enterprise Miner
- FOCUS Product Family
- GainSmarts
- GhostMiner
- iModel
- Insightful Miner (formerly S-PLUS)
- Intelligent Miner
- Knowledge Miner
- KnowledgeSTUDIO
- K.Wiz / K.Wiz AE / thinkCRA
- KXEN Analytic Framework
- MATLAB Neural Network Toolbox
- Model 1
- NeuralWorks Predict/Professional II
- Neuroshell
- Oracle 9i Data Mining Suite (formerly Darwin)
- PolyAnalyst

– Re:cognition (formerly MinEit)
– Scenario / 4Thought
– See5 / Cubist / Magnus Opus
– SPAD
– SQL Server 2000
– STATISTICA Data Miner
– SuperQuery Discovery
– SwissAnalyst
– Teradata Warehouse Data Mining
– Weka
– WizRule / WizWhy
– XpertRule Miner

The corresponding database is available as a dynamic Excel document on request from datamining@elca.ch. The information is up to date as of February 2003. We plan to implement the dynamic nature of this document within a publicly-available Web-enabled Internet front-end.

The data gathered about the 41 tools studied can serve as input to such methodologies as [3] to assist users in evaluating and selecting an appropriate tool. To narrow down such advanced evaluations, simple queries can be issued to the database to find the subset of tools that match specific selection criteria (e.g., a given price range, a specific business goal, etc.).

The above list is by no means exhaustive. It is believed however to be representative of those tools that explicitly implement Data Mining algorithms and that have attracted the most popularity so far (see for example the June 2002 survey [14]). New tools may, of course, be added as information becomes available.

5. Conclusion

We have described a general schema for the characterisation of Data Mining tools. The dimensions proposed are meant to be accessible to business users and to assist them in the tool selection process.

The dimensions are intended for Data Mining *tools* and not *algorithms*. When dealing with algorithms, one should also consider issues such as computational complexity, noise handling, comprehensibility vs opacity, etc. Although these are often critical in applying Data Mining, they can not be generally dealt with at the tool level since any given tool may implement a number of different algorithms, each with a different behaviour with respect to these issues.

However, with a standard schema and corresponding database, users are able to select a DM software package with respect to its ability to meet high-level business objectives. Automatic advice strategies such as METAL's Data Mining Advisor [20] can then be used to assist users further in the selection of the most appropriate algorithms/models for their specific tasks.

Appendix: Formal DB schema

This appendix shows the formal database schema implementing our characterisation and used for the creation and population of our current dynamic database of commercial tools.

```
TABLE BUSINESS-GOALS
   ATTR Tool-ID                Key
   ATTR Cust-Acquisition       {Yes, No}
   ATTR CrossUp-selling        {Yes, No}
   ATTR Product-Dev            {Yes, No}
   ATTR Churn-Prediction       {Yes, No}
   ATTR Fraud-Detection        {Yes, No}
   ATTR Market-basket-Anal     {Yes, No}
   ATTR Risk-Assessment        {Yes, No}
   ATTR Predict/Forecast       {Yes, No}
   ATTR Outcomes-Measure       {Yes, No}
   ATTR Condition-Monitor      {Yes, No}
   ATTR Health-Monitor         {Yes, No}
   ATTR Discovery              {Yes, No}

TABLE MODEL-TYPES
   ATTR Tool-ID                Key
   ATTR Classification         {Yes, No}
   ATTR Regression             {Yes, No}
   ATTR Descriptive-Model      {Yes, No}
   ATTR Dependency-Model       {Yes, No}
   ATTR Anomaly-Change         {Yes, No}
   ATTR Time-Series            {Yes, No}

TABLE PROCESS-FEATURES
   ATTR Tool-ID                Key
   ATTR Flat File              {Yes, No}
   ATTR ODBC/JDBC              {Yes, No}
   ATTR SAS                    {Yes, No}
   ATTR XML                    {Yes, No}
   ATTR Data-Char              {Yes, No}
   ATTR Data-Visual            {Yes, No}
   ATTR Record-Sel             {Yes, No}
   ATTR Attribute-Sel          {Yes, No}
   ATTR Data-Transform         {Yes, No}
   ATTR Decision-Tree          {Yes, No}
   ATTR Rule-Learning          {Yes, No}
   ATTR Neural-Networks        {Yes, No}
   ATTR Lin/Log-Regression     {Yes, No}
   ATTR Association-Learn      {Yes, No}
   ATTR IB/NN-Learn            {Yes, No}
   ATTR Unsupervised-Learn     {Yes, No}
   ATTR Probabilistic-Learn    {Yes, No}
   ATTR HO/ITSet               {Yes, No}
   ATTR Cross-Validation       {Yes, No}
```

```
   ATTR Lift/Gain-Charts     {Yes, No}
   ATTR ROC-Analysis         {Yes, No}
   ATTR Summary-Reports      {Yes, No}
   ATTR Model-Visual         {Yes, No}
   ATTR Save/Reload-Models   {Yes, No}
   ATTR Produce-Executable   {Yes, No}
   ATTR PMML/XML-Export      {Yes, No}
   ATTR Comment-Fields       {Yes, No}
   ATTR DS-Size-Limit        Integer
   ATTR Parallelisation      {Yes, No}
   ATTR Incrementality       {Yes, No}
   ATTR Expert-Options       {Yes, No}
   ATTR Batch-Processing     {Yes, No}

TABLE UI-FEATURES
   ATTR Tool-ID              Key
   ATTR Layout               {GUI, Command Line}
   ATTR Visual-Prog          {Yes, No}
   ATTR OnLine-Help          {Yes, No}

TABLE SYS-REQ
   ATTR Tool-ID              Key
   ATTR PC                   {Yes, No}
   ATTR Unix-Solaris         {Yes, No}
   ATTR Soft-Req             {DB2, SAS Base, Oracle, Java/JRE}
   ATTR Standalone           {Yes, No}
   ATTR Client/Server        {Yes, No}
   ATTR ThinClient           {Yes, No}

TABLE VENDOR-INFORMATION
   ATTR Tool-ID              Key
   ATTR HQ-Phone#            Text
   ATTR URL                  Text
   ATTR Email                Text
   ATTR Level-TechSup        {Excellent, Good, Average, Poor, Unaccep
                              table}
   ATTR Price                {Freeware, Contact Vendor} OR Num(6,2)
   ATTR Free-Eval            {Yes, No}
   ATTR Market-Pen           {High, Average, Low}
   ATTR Long                 Num(2)
```

References

[1] F. Bihan, *Comparison of Leading Data Mining Tools*, Available at http://www.ksurf.net/cs761/.

[2] R.J. Brachman and T. Anand, The Process of Knowledge Discovery in Databases: A Human-centered Approach, in: *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, eds, AAAI/MIT Press, 1996, pp. 33–51.

[3] K. Collier, B. Carey, D. Sautter and C. Marjaniemi, *A Methodology for Evaluating and Selecting Data Mining Software*, in Proceedings of the Thirty Second Hawaii International Conference on System Sciences, Available at http://www.cse.nau.edu, 1999.

[4] *CRISP-DM 1.0: Step-by-step data mining guide*, Available as an SPSS White Paper at http://www.spss.com.

[5] *Criteria for Data Mining Tool Evaluation*, Available at http://www.catenate.net/a_basic_checklist_for_a_high_per.htm.

[6] *Data Mining*, Available at http://www.dwinfocenter.org/datamine.html.

[7] *Data Mining Catalogue*, Available at http://datamining.csiro.au/catalogue.html.

[8] *Data Mining Product Features*, Available at http://www.xore.com/prodtable.html.

[9] J.F. Elder and D.M. Abbott, *A Comparison of Leading Data Mining Tools*, in Proceedings of the Fourth Annual Conference on Knowledge Discovery and Data Mining (KDD'98), Available at http://www.datamininglab.com, 1998.

[10] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM* **39**(11) (1996), 27–34.

[11] L. Fontaine, *Industrial KDD Systems*, Available at http://www.lri.fr/blf/results, 1999.

[12] *From Data to Business Advantage: Data Mining, The SEMMA Methodology and SAS Software*, Available as a SAS Institute White Paper at http://www.sas.com.

[13] W. Gaul and F. Saeuberlich, Classification and Positioning of Data Mining Tools, in: *Classification in the Information Age*, W. Gaul and H. Locarek-Junge, eds, Springer, Berlin, Heidelberg, 1999, pp. 143–152.

[14] *KDnuggets Poll: Data mining tools you regularly use*, Available at http://www.kdnuggets.com/polls/data_mining_tools_2002_june2.htm.

[15] *KDnuggets Poll: What main methodology are you using for data mining?* Available at http://www.kdnuggets.com/polls/methodology.htm.

[16] M.A. King and J.F. Elder, *Evaluation of Fourteen Desktop Data Mining Tools*, in Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics, 1998, Available at http://www.datamininglab.com.

[17] *List of Software*, Available at http://www.mlnet.org.

[18] H. Mannila, *Methods and Problems in Data Mining*, in Proceedings of the 1997 International Conference on Database Theory, 1997.

[19] M. Mendonca and N.L. Sunderhaft, Mining Software Engineering Data: A Survey, *Data & Analysis Center for Software* (*DACS*) *State-of-the-Art Report*, No. DACS-SOAR-99-3, 1999, Available at http://www.dacs.dtic.mil/techs/datamining/index.html.

[20] *METAL: A Meta-learning Assistant for Providing User Support in Machine Learning and Data Mining*, European ESPRIT Project Nr 26.357, See http://www.metal-kdd.org.

[21] K. Parsaye, A Characterization of Data Mining Technologies and Processes, *Journal of Data Warehousing*, December 1997, Available at http://www.datamining.com/dm-tech.htm.