

# SURVEY OF TOOLS FOR DATA MINING

Natasa Miskov

## CATEGORIES

In evaluating data mining tools you must look at a whole constellation of features, described below. You cannot put data mining tools into simple categories such as “high-end” versus “low-end” because the products are too rich in functionality to divide along just one dimension.

There are three main types of data mining products.

- First are tools that are analysis aids for OLAP. They help OLAP users identify the most important dimensions and segments on which they should focus attention. Leading tools in this category include Business Objects, Business Miner and Cognos Scenario.
- The next category includes the “pure” data mining products. These are horizontal tools aimed at data mining analysts concerned with solving a broad range of problems. Leading tools in this category include (in alphabetical order) IBM Intelligent Miner, Oracle Darwin, SAS Enterprise Miner, SGI MineSet, and SPSS Clementine.
- The last category is analytic applications which implement specific business processes for which data mining is an integral part.

For example, while you can use a horizontal data mining tool as part of the solution of many customer relationship management problems, you can also buy customized packages with the data mining imbedded. However, even packaged solutions require you to build and tune models that match your data. In some cases, the package requires a complete model development phase that can take months.

No matter how comprehensive the list of capabilities and features you develop for

describing a data mining product, nothing substitutes for actual hands-on experience. While feature checklists are an essential part of the purchase decision, they can only rule out products that fall short of your requirements. Actually using a product in a pilot project is necessary to determine if it is the best match for your problem and your organization.

Depending on your particular circumstances — system architecture, staff resources, database size, problem complexity — some data mining products will be better suited than others to meet your needs. Evaluating a data mining product involves learning about its capabilities in a number of key areas that may not be addressed in standard marketing materials.

## A COMPERISON OF LEADING TOOLS

Categories used for comparisons are: platforms supported, algorithms included (Decision Trees, Neural Networks, etc.), data input and model output options, usability ratings, visualization capabilities and modeling automation methods.

After comparing, differences between tools which are determined are presented through their strenghts and weaknesses:

- *Clementine from SPSS*, leading visual rapid modeling environment for data mining. Strenghts: visual interface; algorithm breadth. Weaknesses: scalability
- *IBM Intelligent Miner for Data*. embodies latest data mining technology with capabilities to support the full range of mining processes, from data analysis and preparation tasks through mining and assimilation of results. Strenghts: algorithm breadth; graphical tree/cluster output. Weaknesses: few algorithm options; no automation.

- *Oracle Darwin* is powerful, easy-to-use enterprise data mining software that finds meaningful patterns hidden within corporate data. Strengths: efficient client-server intuitive interface options. Weaknesses: no unsupervised; limited visualization.
- *Data Cruncher from Data Mind*. Strengths: ease of use. Weaknesses: single algorithm.
- *Enterprise Miner from SAS Institute*, an integrated suite which provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process. Strengths: depth of algorithms; visual interface. Weaknesses: harder to use; new product issues.
- *Gain Smarts from Urban Science*, uses predictive modeling technology that can analyze past purchase, demographic, and lifestyle data, to predict the likelihood of response and develop an understanding of consumer characteristics. Strengths: data transformations, built on SAS; algorithm option depth. Weaknesses: no unsupervised; limited visualization.
- *MineSet from Silicon Graphics*. Strengths: data visualization. Weaknesses: few algorithms; no model export.
- *Model 1 from Group 1/Unica Technologies*, Response Modeler, Segmenter and Profiler, Customer Valuator, and Cross-Seller modules with a wizard GUI. Strengths: ease of use; automated model discovery. Weaknesses: really a vertical tool.
- *Model Quest from AbTech Corp*. Strengths: breadth of algorithms. Weaknesses: some non-intuitive interface options.
- *PRW from Unica Technologies*. Strengths: extensive algorithms; automated model selection. Weaknesses: limited visualization.
- *CART from Salford Systems*. Strengths: depth of tree options. Weaknesses: difficult file I/O; limited visualization.
- *Neuro Shell from Ward Systems Group*. Strengths: multiple neural network architectures. Weaknesses: unorthodox interface; only neural networks.
- *OLPARS from PAR Government Systems*. Strengths: multiple statistical algorithms; class-based visualization. Weaknesses: dated interface; difficult file I/O.
- *Cognos Scenario*, innovative data mining tool that enables managers and knowledge workers to discover hidden trends and patterns, and unearth previously unsuspected correlations, without requiring them to have expert knowledge of statistical techniques. Strengths: ease of use. Weaknesses: narrow analysis path.
- *See 5 from RuleQuest Research*. Strengths: depth of tree options. Weaknesses: limited visualization; few data options.
- *S-Plus from MathSoft*. Strengths: depth of algorithms; visualization; programable/extendable. Weaknesses: limited inductive methods; steep learning curve.
- *WizWhy from WizSoft* is a data mining tool for revealing trends, rules and interesting phenomena in the data and issuing predictions of new cases. Strengths: ease of use; ease of model understanding. Weaknesses: limited visualization.

## REFERENCES

1. KDNuggets, Software Suites for Data Mining and Knowledge Discovery, 2002 (<http://www.kdnuggets.com>)
2. Knowledge Storm, Data Mining, 2002, (<http://www.knowledgestorm.com>)
3. Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, 1999 (<http://www.twocrows.com>)
4. John F. Elder IV & Dean W. Abbott, A Comparison of Leading Data Mining Tools, 1998 (<http://www.datamininglab.com>)
5. <http://www.thearling.com>
6. Veljko Milutinovic, Frederic Patricelli, E-Business and E- Challenges, 2002