Chapter XII
A Comparison and Scenario Analysis of Leading Data Mining Software

John Wang
Montclair State University, USA

Xiaohua Hu
Drexel University, USA

Kimberly Hollister
Montclair State University, USA

Dan Zhu
Iowa State University, USA

ABSTRACT

Finding the right software is often hindered by different criteria as well as by technology changes. We performed an analytic hierarchy process (AHP) analysis using Expert Choice to determine which data mining package was best suitable for us. Deliberating a dozen alternatives and objectives led us to a series of pair-wise comparisons. When further synthesizing the results, Expert Choice helped us provide a clear rationale for the decision. The issue is that data mining technology is changing very rapidly. Our article focused only on the major suppliers typically available in the market place. The method and the process that we have used can be easily applied to analyze and compare other data mining software or knowledge management initiatives.

INTRODUCTION

Based on the *knowledge life cycle* model, four stages of knowledge creation, knowledge storage/retrieval, knowledge transfer, and knowledge application have been proposed by Alavi and Leidner

(2001) and confirmed by Jennex (2006). "To be effective knowledge management systems, KMS, must support the various knowledge management functions of knowledge capture, storage, search, retrieval, and use" (Jennex, 2006, p.3). Knowledge discovery is generally one of the important stages or phases of KM. And while this incorporates identifying critical knowledge (this may also be what this stage is called), using data mining to aid in knowledge discovery is appropriate as being a useful KM tool.

Data mining is a promising tool that assists companies to uncover patterns hidden in their data. These patterns may be further used to forecast customer behavior, products and processes. It is important that managers who understand the business, the data, and the general nature of the analytical methods are involved. Realistic expectation can yield rewarding results across a wide range of applications, from improving revenues to reducing costs (Davenport & Harris, 2007; Porter & Miller, 2001). It is crucial to properly collect and prepare the data, and to check the models against the real figures. The best model is often found after managers build models of several different types or by trying different technologies or algorithms. This alone demonstrates the active role managers play in the data mining or other knowledge management processes.

Selecting software is a practical yet very important problem for a company (James, Hakim, Chandras, King, & Variar, 2004). However, not enough attention is given to this critical task. Current literature is quite limited because selecting software is such a complex problem, due to many criteria and frequent technology changes (Elder IV & Abbott, 1998; Giraud-Carrier & Povel, 2003). Haughton, Deichmann, Eshghi, Sayek, Teebagy, and Topi (2003) generally reviewed several computer *software packages* for *data mining, including* SPSS Clementine, XLMiner, Quadstone, GhostMiner, and SAS Enterprise Miner. Corral, Griffin, and Jennex (2005) examined the potential of knowledge management in data warehousing from an expert's perspective. Jennex (2006) introduced technologies in support of knowledge management systems.

Firstly, this article will take a brief look at data mining today, through describing some of the opportunities, applications and available technologies. We will then discuss and analyze several of the most powerful data mining software tools available on the market today. Ultimately, we will also attempt to provide an analytical analysis and comparison among the brands we have selected. Our selection is based, in part, on our own experience using data mining software as well as writing data mining code, SQL code and our work as relational database administrators. For our analytical comparison we will be using *Expert Choice* (Version 11) advanced decision support software.

DATA MINING SOFTWARE

Data mining software analyzes- based on open-ended user queries- relationships and patterns that are stored in transaction data. Available are several types of analytical software: statistical, machine learning and neural networks, decision trees, Naive-Bayes, K-Nearest Neighbor, rule induction, clustering, rules based, linear and logistical regression time sequence, and so forth. Along the lines of Mena (1998) and Martin (2005), the basic steps of data mining for knowledge discoveries are:

1. Define business problem
2. Build data mining data base
3. Explore data
4. Prepare data for modeling
5. Build model
6. Evaluate model
7. Deploy model
8. Results

Note: Each of these steps contains managerial issues which must be addressed.

The key to knowledge discovery is a true understanding of your data and your business. Without this understanding, no algorithm is going to provide you with a result in which you should confide. Moreover, without this background you will not be able to identify the problems you are trying to solve, prepare the data for mining, or correctly interpret the results. There are many tasks involved in the construction of a database: data collection, data description, selection, data quality assessment and data cleansing, consolidation and integration, metadata construction, and maintaining the database. In exploring the data, the manager must choose the appropriate hardware to accomplish this feat. The goal is to identify the most important fields in predicting an outcome, and determine which derived values may be useful. According to O Chan (2005), a good interface and fast computer response are very important in this phase because the very nature of your exploration is changed when you often have to wait up to 30 minutes for some graphs to be created.

Preparing data for modeling consists of four main parts: selecting variables, selecting rows, constructing new variables and transforming variables. The managerial decision in this case focuses on identifying key variables to examine, nonfully functional variables inclusive. The time it takes to build a model increases with the number of variables while blindly including extraneous columns can lead to incorrect models. The most important thing to remember about data model building is that it is an interactive process. Many alternative models may have to be examined to find one that is most appropriate in solving your business problem. A manager searching for a good model may go back and amend the data he or she is using or even modify his or her problem statement. In the evaluation and interpretation process, the accuracy rate found during testing applies only to the data on which the model was built. The accuracy may vary if the data to which the model is applied differ in important and unpredictable ways from the original data set.

Once a data mining model is built and validated, it can be used in two main ways. First, the manager may recommend actions based on simply viewing the model and its results. For example, the manager may look at the clusters the model has identified, the rules that define the model or the lift and ROI charts that depict the effect of the model. The second process involves applying the model to different datasets. The manager may use the model to flag records based on their classification or assign a score such as the probability of an action.

Data Mining Software Alternatives

As stated earlier in our introduction, there are numerous data mining software alternatives that vary in the number of modeling and visualization nodes as well as in price. We have elected the following eight software vendors for comparison due to a limitation of trial version of *Expert Choice* (Version 11):

- Clementine from SPSS
- DB2 Intelligent Miner from IBM
- Enterprise Miner from SAS
- GhostMiner by Fujitsu
- Insightful Miner V5.2 for Insightful
- Megaputer PolyAnalyst
- Microsoft SQL Server 2005 Enterprise Edition
- Oracle Data Miner

Although there are various other comparable programs available, we were limited in our selection. One of the limiting factors was the inadequacy of alternatives in our decision tools' evaluation copy.

Decision Tool

To aid in comparing our software choices, we used an evaluation copy of *Expert Choice* version 11, a leading software solution construed to analyze, categorize, prioritize, select, allocate and choose a selection based on relevant criteria (Expert Choice Inc, 2007). *Expert Choice* incorporates a process known as *Analytical Hierarchical Process* (AHP) into its software (Saaty & Vargas, 2006; Saaty, 1980, 1996, 2001, 2005). Research has demonstrated that AHP is a powerful decision-making tool that can help organizations avoid making costly mistakes caused by bad decisions (Hemaida & Schmits, 2006). AHP was developed by Saaty and Kearns and consists of four stages (Roper-Lowe & Sharp, 1990). The first stage is to construct a hierarchy where the primary objective, or goal, is at the highest level. Criteria, which can also be subdivided, follow in decreasing order. At the bottom of the hierarchy are the alternatives to be evaluated. The second stage calculates weights for the criteria using pair-wise comparisons. In the next stage, the alternatives are also compared to each other in respect to each criterion. Finally, all weighted scores are tallied to yield a final score. The alternative with the highest score is considered the best alternative.

SAMPLE PRODUCTS

In this section, we will analyze the various features and benefits offered by each of our software alternatives that are to be considered, as well as researching product reviews and professional opinions. The information gathered from this analysis will serve as the basis for the pair-wise comparisons of the software alternatives with respect to each of our criteria In other words, we will investigate how important one choice is over the other alternative given a specific criterion, when comparing any two software alternatives,

CART by Salford Systems

CART is an easy to use decision tree tool that uses the CART algorithm and boosting. Its main objective is to rifle through databases, identifying significant patterns and relationships, which are then used to generate predictive models. CART uses an exhaustive, recursive partitioning routine to generate binary splits by posing a series of yes-no questions. It searches for questions that split nodes into relatively homogenous child nodes. As the tree evolves, the nodes become more homogenous, identifying segments. CART supports more than 80 file formats, including SAP, SPSS databases such as Oracle and Informix, Excel spreadsheets, and Lotus 1-2-3 spreadsheets.

CART was formulated from the original CART code developed by Stanford University and University of California at Berkeley statisticians. The frequent addition of new features and capabilities continually enhances the procedure, strengthening the accuracy and reliability of the results. CART has a no-stopping rule, which makes it unique. This means that more data are read and compared, and it ensures that important data are not overlooked by stopping too soon. It produces an over-grown tree, and immediately prunes it back for the most optimal results. CART also uses a powerful binary split search approach. This means the trees are more sparing with data and detect more structure before too little data are left for learning. Next, CART uses automatic self-validation procedures, which are essential in avoiding the trap of finding patterns that apply only to the training data.

CART was designed for both technical and nontechnical users. It can quickly identify important data relationships. It offers users some flexibility, with the choice of how to split criteria. It also offers different models for scalability. The results are easy to read and understand, with decision tree diagrams drawn out. Salford Systems understands that expert and timely technical support is a critical part of the business, which is why they offer many means of customer training and support. The company offers both public and private on-site instruction, user seminars, hand-on training courses, consultation services, and e-mail, NetMeeting, and phone support from offices worldwide.

SPSS—Clementine

Clementine data mining software by SPSS is useful for organizations using SPSS infrastructure as well as those with mixed platforms. This program supports both client and server platforms, including the Windows family of products and Sun Solaris, HP-UX11i, IBM AIX, and OS/400 server platforms.

With regard to reliability, Clementine by SPSS supports decision trees, neural networks, regression, self-organizing maps, clustering, and association rules (Lampe & Garcia, 2004). However, the author states that Clementine implements "a broad set of statistical algorithms, but fewer than in the SAS and IBM packages" (Lampe & Garcia, 2004, p. 18).

In the area of efficiency, Clementine works with SPSS, SAS and SQL and can export to C code and Predictive Model Markup Language (PPML) (Angus, 2006). It can also handle critical data preparation, rapid modeling, and model scoring tasks. These tasks are all performed using GUI graphical layouts, workflow diagrams, scatterplots, distribution, histogram, multiplot and Web charts (which are unique to Clementine (Haughton et al., 2003)).

Training and support for Clementine are available from SPSS in the form of online tutorials, downloadable overview and demos of the program, along with online technical support and excellent help screens. The price of Clementine starts at $75,000 (Angus, 2006).

Enterprise Miner by SAS

Enterprise Miner was developed by SAS Corporation, which was originally called Statistical Analysis System. Enterprise Miner is an integrated software product that provides widespread business solutions for data mining based on SEMMA (Sample, Explore, Modify, Model, Assess) methodology. It has many different statistical tools including decision trees, clustering, linear and logistic regression and neural networks. Data preparation tools include outlier detection, variable transformations, random sampling, and the partitioning of data sets into training, test, and validation data sets. Its advanced GUI allows you to review large amounts of data in multidimensional histograms with ease, as well as compare modeling results graphically.

Enterprise Miner includes several procedures that automate traditional analysis tasks, such as choosing the variables to include in a model and applying the appropriate transformations. The system also provides extensive visualization to help users explore the data to decide on additional manipulations that the system itself does not recommend. Enterprise Miners' graphical user interface and automated framework indicate that the user does not have to know how the tools work to use them. Release 8.2 provides cross-platform national language support that is especially important to international customers.

GhostMiner by Fujitsu

GhostMiner is a data mining software product from Fujitsu that not only supports common databases (or spreadsheets) and mature machine learning algorithms, but also assists with data preparation and selection, model validation, multimodels (like committees or k-classifiers), and visualization. GhostMiner provides a large range of **data preparation** techniques and a broad scope of **selection of featured** methods. Choice of data mining algorithms and v**isualization techniques** are integrated.

GhostMiner offers several project features unique to their platform, which enables users to create simple interfaces for their specific needs. GhostMiner has a human machine interface (HMI) that is fairly user friendly and easy to start up right out of the box. The system is so easy to use that it actually has a feel of being too user friendly and may be missing some of the power of the larger server based data mining software tool such as Darwin IBM and Oracle. GhostMiner can be loaded directly onto a Windows based PC and is equally adept at data mining a system database as it is a series of spreadsheets, text or ASCII files.

GhostMiner contains both data preprocessing capabilities as well as data visualization capabilities. Data preprocessing includes data normalization, standardization and many preliminary statistical analysis functions, such as variance, standard deviation, mean and median across the entire database. GhostMiner does not have the inherent flexibility of some of the larger, more robust products, and also does not offer the same levels of support as the products offered by Oracle and IBM. GhostMiner is a product marketed to small to mid size users who are looking for a simple to use product at a lower price than some of its larger, well known counterparts.

Insightful-Insightful Miner

Insightful Miner is a cost effective data mining software program. The software has numerous model types, algorithms and visualizers, including decision trees, *Block Model Averaging*, linear and logistic regression, neural networks, Naïve Bayes, Cox proportional hazard models, K-means clustering and others. Insightful Miner offers highly scalable algorithms, which train models on very large data sets without the need for sampling or aggregation. Insightful also offers data preprocessing and data cleansing as well as exploratory data analysis and visualization. Insightful Miner's cost is typically $12,000.

According to a product review in *DM Review* (Lurie, 2004) the main strength of Insightful Miner is its ability to scale large data sets in an accessible manner. It provides the analytic tools required to transform fragmented raw data into "actionable knowledge" (Lurie, 2004, p. 88). Insightful Miner provides cutting edge analytics and reporting tools to identify patterns, trends and relationships in data. Insightfuls' simplicity allows users to quickly aggregate, clean and analyze data. Its powerful reporting and modeling capabilities allow users to deliver clear, usable analytics to designers and producers. Simple visual work maps make it easy for users to become productive relatively quickly. Insightful Miner provides excellent product support and its documentation is complete and easy to understand.

In another product review of Insightful Miner (Deal, 2004), the software was found to be a comprehensive data mining application that includes extensive data input, data manipulation, and analysis capabilities. Insightful Miner can efficiently process large amounts of data by using a chunking and processing algorithm that is intended to be scalable to the mass of data used for each analysis. Insightfuls' ability to integrate S-Plus strengthens and extends its functionality. Deal (2004) stated that Insightful Miner "is a very simple and intuitive process" (p. 46).

IBM-Intelligent Miner

IBM Intelligent Miner V7R1 is very user friendly software. It is an essential e-commerce tool, as it can aid in handling transactions as they come in. It has business intelligence applications, which allow it to make decisions that would be good for any business, large or small. The intelligence part of the software could cut costs and increase profits. The data screens help with decision making and improvement on processes that are out of date. It also maximizes the business to customer relationship, because of the personalization the software can provide for each client. This software package is also compatible with Windows, AIX, Solaris and Linux servers.

IBM Intelligent Miner V7R1 has IM Scoring, in which the user has an advantage, because scores and ranks are done in real time. This means that as a new transaction takes place, it would then reorganize the scores/ranks of the customers' information. For example, when a customer buys an item off the Internet, the software would update for the payment due and when it should be posted. The same would apply for a dentist visit: after 6 months it would indicate that it is time for another checkup for the particular patient. As the appointment would approach, the higher the person would be on the list, that is, moving up the ranks. Another advantage with the IM Scoring is the high performance and scalability of mining functions, thus making sampling obsolete.

The best aspect of this product is its user friendliness. The whole staff would be content with it. It can also be updated easily, without any disruption to the business. IBM is currently promoting the DB2 Query Management Facility version 8.1, because in March 2006, IBM withdrew Intelligent Miner from all marketing and ended all its support for Intelligent Miner tools.

Megaputer-PolyAnalyst

Another data mining software package is PolyAnalyst (the newest version is 4.6) made by the Megaputer Company. This company is quite small, especially compared to some of the other companies we have profiled. Megaputer Intelligence Inc. is a leading developer and distributor of advanced software tools for data mining, text mining, and intelligent e-commerce personalization. The tools help reveal knowledge hidden in data. They add intelligence and insight to every step of the business decision-making process.

Because the Megaputer Company focuses primarily on data mining programs, they can offer a more comprehensive program than other companies who simply have a data mining component to existing products. They offer a vast array of algorithms from which a consumer can choose the ones they need specifically, making the product ready to be customized. The price for an older version of PolyAnalyst (the most recent pricing data found) is an affordable $2,300 for the base version and can go up to $14,900 with all the algorithms. Also, the developer kit for PolyAnalyst is available for $16,000.

PolyAnalyst can be run either on a stand-alone system or in a client/server configuration, where the server would handle the data processing. It only works with the Microsoft Windows O/S, which shows that it is not as portable as some other products analyzed. Also, Megaputer offers possible users a free evaluation version to decide if this is the software right for them. The program offers a rich set of features. PolyAnalyst by Megaputer seems to be a feature rich data mining software package. The price and ala carte feature set seem more suited for a smaller company that cannot afford to use a more expensive data mining solution that would require the use of highly trained employees.

Oracle–Oracle Data Mining

Oracle Data Miner is the graphical user interface for Oracle Data Mining (Release 10.1 and above) that helps data analysts mine their Oracle data to find valuable hidden information, patterns, and new insights. Oracle Data Mining is a powerful data mining software embedded in the Oracle Database that enables you to discover new information hidden in your data and helps businesses target their best customers and find and prevent fraud.

Oracle provides unique portability across all major platforms including Windows, Solaris, HP-UX, IBM AIX, Compaq Tru64, and Linux and ensures that applications run without modification after changing platforms. There are two common ways to architect a database: client/server or multitier. Two basic memory structures are associated with Oracle software: the system global area and the program global area.

Oracle Data Miner facilitates interactive data preparation, data mining model creation, evaluation, refinement and model scoring. Oracle Data Mining provides the following supervised data mining algorithms: Naïve Bayes, Adaptive Bayes Network, decision trees, Support Vector Machines, and attribute importance. Unsupervised algorithms are: clustering, association rules, feature selection, anomaly detection, text mining and unstructured data, and life sciences algorithm. Mining Activity Guides provide structured templates for all users to explore and mine their data.

Oracle Data Mining (ODM) enables companies to extract information efficiently from the very largest databases, and build integrated business intelligence applications and support data mining problems such as: classification, prediction, regression, clustering, associations, attribute importance, feature extraction and sequence similarity searches and analysis. When the capabilities of Oracle Data Mining are combined with the ability of the RDBMS to access, preprocess, retrieve and analyze data, they create a very powerful platform for data analysis.

Oracle Data Mining can generate valuable new insights and reports that can help proactively manage your business, according to the Oracle Discoverer report. Oracle Data Miner models can be visualized graphically and can be display in tables, histograms, line graphs and pie graphs. Data may be in either Excel or the Database. Significant productivity enhancements are achieved by eliminating the extraction of data from the database to special-purpose data mining tools (Berger & Haberstroh, 2005).

Data size is unlimited. The expert analyst can adjust some or all of the parameters manually, but the option to allow the algorithms to optimize the parameters intelligently, with no intervention, is available. There are free demos available: Oracle Data Mining, Integration with Oracle BI EE, Spreadsheet Add-in for Predictive Analytics, and Text Mining. The tutorial Oracle by Example series and online training provides valuable hands-on experience, step-by-step instructions on how to implement various technology solutions to business problems. Oracle Data Mining significantly reduces the cost of data mining. Savings are realized in the avoidance of additional hardware purchases for computing and storage environments, redundant copies and multiple versions of the data and duplication of personnel who perform similar functions. Database analytics includes: engine, basic statistics (free), data mining, and text mining.

SQL Server 2005

SQL server 2005 is Microsoft's solution to database management and data mining. SQL Server database platform provides enterprise-class data management with integrated business intelligence (BI) tools. SQL Server 2005 combines analysis, reporting, integration, and notification. SQL server is closely integrated with Microsoft Visual Studio, the Microsoft Office System, and a suite of new development tools, including the Business Intelligence Development Studio (Bednarz, 2005).

Microsoft SQL Server series utilizes the Windows operating system and features four discrete algorithms. HMI features include a Windows' interface, as well as complete integration with the Microsoft Office suite. Reports that are served by the Report Server in Reporting Services can run in the context of Microsoft SharePoint Portal Server and Microsoft Office System applications such as Microsoft Word and Microsoft Excel (Fontana, 2005). SharePoint can be used to subscribe to reports, create new versions of reports, and distribute reports. SQL Server 2005 also supports rich, full-text search applications. Query performance and scalability have been improved dramatically, and new management tools will provide greater insight into the full-text implementation.

SQL Server also features an online restore function, database encryption and a fast recovery option. It also has a system with built-in scalability features such as parallel partition processing, creation of *remote relational online analytical processing* (ROLAP) or *hybrid online analytical processing* (HOLAP) partitions, distributed partitioned cubes, persisted calculations, and proactive caching.

COMPARISON

We use *Expert Choice* in the evaluation process and will attempt to analytically quantify the aspects of data mining software that best define overall product quality. Before we describe the decision making process, we would like to present several assumptions on which our decision will be based:

1. In addition to our experience, we will rely on manufacture specifications, descriptions and described attributes, along with third party reviews where available.
2. We will base our needs on fundamental business goals such as business-related decision making and business-driven information analysis. Although this definition may seem overly broad, we will attempt to further limit our scope by eliminating research and development, educational and political as well as most human resource applications.
3. Because we are using a trial version of *Expert Choice* advanced decision making software, there will be limits with respect to importing and exporting data as well as with printing and possibly some advanced analytical tools. Therefore, we will utilize screen captures embedded into this document, and will manually write any necessary data as opposed to systematic imports or exports.

Criteria Revisited

Our selection process will be centered on the below mentioned software quality criteria. We will attempt to compare all of our selections based on the specified criteria. Using *Expert Choice*, we will make objective ratings of each product, comparing in a pair-wise manner, attributes that define each element.

- **Portability:** the amount of platform independence; the number of support platforms and supported software architectures as well as any software requirements needed to run the software.
- **Reliability:** the degree of completeness, accuracy and consistency, any stated warranty and support provided by the vendor. The number of data models and algorithms available with the software as well as any templates or custom models available for creation of projects.
- **Efficiency:** the degree of efficiency and accessibility; the degree in which the product supports the general business goal assumptions and the number of tools available for data preprocessing.
- **Human engineering:** how well the software interfaces and communicates with the outside world, plus the quality of the human machine interface (HMI). Testability – how well the software is structured; how results are displayed and how results are reintroduced into the process if applicable.
- **Understanding:** degree of self-descriptiveness; the degree of simplicity of the machine interface, the use of graphical user interfaces, visual programming ability, summary reports, and data model visualization.
- **Modifiability:** the degree of augmentation ability and the ability to change over time

and expand; the use of batch processing and any expert options as well as data size limitations.

- **Price, training and support:** price of product, availability of evaluation or demo versions, and the amount of post purchase support included in the package.

Evaluation Model

Our evaluation criteria, as entered in the Expert Choice, are as follows:

- Portability: evaluated in terms of:
  - Hardware platform (PC, Unix/Solaris workstation, etc.).
  - Software Architecture (standalone, client/server, thin client).
  - Software requirements (DB2, SAS, Base, Java/JRE, Oracle, and so forth.
- Reliability - evaluated in terms of:
  - What model classes does the tool support?
  - How many algorithms does the tool use?
  - Does the tool allow custom model creation or simply uses templates?
  - What is the reputation of the vendor supplying the tool?
- Efficiency evaluated in terms of:
  - How well does the product support our general business goal assumption?
  - Ability to perform data preprocessing.
- Human Engineering evaluated in terms of:
  - Simplicity of HMI (human machine interface)
  - Graphical layout
  - Visual programming ability
- Testability evaluated in terms of dissemination and deployment:
  - How well the results are reintroduced into the process "closing the loop"
  - How results are displayed
- Understanding in terms of evaluation and interpretation of data:

Table 1. Weights assigned to criteria

Category	Priorities
Human Engineering	0.22
Training and Support	0.193
Understandability	0.19
Reliability	0.142
Portability	0.128
Modifiability	0.051
Efficiency	0.039
Testability	0.022
Price	0.016

Table 2. Pair-wise comparison grid WRT hardware platform

	CART by Salford Systems	SAS Enterprise Miner	Oracle 8i	GhostMiner	SQL Server 2005
CART by Salford Systems		3.0	3.0	3.0	7.0
SAS Enterprise Miner			2.0	3.0	4.0
Oracle 8i				3.0	3.0
GhostMiner					4.0
SQL Server 2005					
Inconsistency: 0.73					

Table 3. Class weightings for overall hardware platform independence

Vendor	Class Weighting
CART by Salford Systems	0.268
SAS Enterprise Miner	0.223
Oracle 8i	0.215
GhostMiner	0.141
SQP Server 2005	0.152
Overall inconsistency: 0.28	

Table 4. Results from "Choosing a data mining software vendor"

Vendor	Overall Weight
CART by Salford Systems	0.191
SAS Enterprise Miner	0.215
Oracle 8i	0.222
GhostMiner	0.109
SQP Server 2005	0.263
Inconsistency: 0.73	

  - Are summary reports available?
  - Can the model be visualized graphically?
- Modifiability in terms of scalability and upgrades:
  - What is the data set size limit?
  - Are there expert options or batch processing?
- Training and support evaluated in terms of:
  - Is a free demo available?
  - Is any free training or support available with the purchase?
- Price (where available) – if pricing is not available we will note our evaluation as price neutral.

PROCEDURE OF EXPERT CHOICE

Let's use five products to demonstrate the whole process of *Expert Choice* on a small scale. We commence with pair-wise comparisons for each of our criteria. Figure 1 is a screen capture of

Table 5. Dynamic sensitivity analysis

Category	Category Weight	Vendor	Vendor Preference Weight
Portability	12.8%	CART by Salford Systems	26.8%
Reliability	14.2%	SAS Enterprise Miner	22.3%
Efficiency	3.9%	Oracle 8i	21.5%
Human Engineering	22.0%	GhostMiner	14.1%
Testability	2.2%	SQL Server 2005	15.2%
Understandability	19.0%		
Modifiability	5.1%		
Training and Support	19.3%		
Price	1.6%		

Table 6. Dynamic sensitivity analysis with different constraints

Category	Category Weight	Vendor	Vendor Preference Weight
Portability	0.1%	CART by Salford Systems	24.0%
Reliability	2.6%	SAS Enterprise Miner	18.5%
Efficiency	3.2%	Oracle 8i	17.1%
Human Engineering	5.8%	GhostMiner	20.8%
Testability	3.3%	SQL Server 2005	19.6%
Understandability	21.5%		
Modifiability	4.1%		
Training and Support	15.7%		
Price	33.5%		

our initial results of priorities. As can be seen in Table 1, we placed a great deal of importance on Human Engineering (weight of .220), slightly less on Training and Support (w=. 193) and then on Understanding (w=. 190). Our main driver was that for the software to be successful, people had to know and understand it. Our next highest priority was Reliability, with a relative weight of .142, followed by Portability, which is platform and hardware independence, with a relative weight of .128.

Next, we perform a pair-wise comparison of each software tool for each criterion; that is, we compare the components of each criterion on a case-by-case basis, assigning relative strengths and weaknesses to each product. Although this process is quite tedious, it provides an accurate measurement for each product. Table 2 shows an example of a pair-wise comparison for the contribution of hardware platform independence to overall platform independence, which is a contributor to overall portability within our quality structure.

Table 3 shows a graphical representation from the pair-wise comparison between all products for the hardware contribution, to overall portability. The screen capture shows a weight for each product with SQL Server as the best in class (with a weight of .269) and GhostMiner as last in class (with a weight of .109). These criteria are also weighted individually so as to roll up into the overall contribution toward portability.

Overall Results

Table 4 shows the overall results from *Expert Choice* advanced decision support software. From the first iteration of our selection process, the best solution for our chosen attributes and assigned priorities is the CART product, with an overall weight of .268, followed by SAS Enterprise with an overall weight of .223. We also performed several iterations, changing the weights of our criteria.

Table 5 shows the assigned weights of each category along with the overall score for all of the objects. This tool allows dynamic sensitivity analysis with respect to changing priorities. We used this tool to look at how much a change in one weight changes the overall goal. Using this tool is similar to the sensitivity analysis performed in Excel Solver; however, instead of listed ranges the Expert Choice tool allows for dynamic manipulation. From the chart, we can see our weighted emphasis on Human Engineering (22%), Training and Support (19.3%) and Understandability. The window on the right shows which system best fits our stated criteria.

In Table 6 we change our requirements in order to verify the strength of our decision. We increase the importance of Price from 1.6% all the way up to 33.5%. We also change our Human Engineering requirement from 22% down to 5.8%, and also reduce Training and Technical Support, Portability, Reliability, Efficiency and Modifiability (flexibility) substantially and still came up with CART systems as our best overall choice (24% weight).

SCENARIO ANALYSIS

We now start to compare eight leading data mining packages based on seven criteria. Determining the best software is a multiple objective decision-making process because different companies may have completely different needs. An array of software may each be the best choice because their design and performance are defined within a certain type of institution. Usually, one data mining software cannot be the best for every scenario. This is because specific software cannot meet the expectations of every type of institution; therefore, the creation of scenarios is a very important tool in term of decision-making process.

Table 7. Weights assigned to each alternative for both a small and large-sized company

Synthesized Weights - with respect to criteria	Efficiency		Human Engineering\ Understandability		Modifiability		Portability	
	Large	**Small**	**Large**	**Small**	**Large**	**Small**	**Large**	**Small**
Clementine	0.213	0.153	0.154	0.154	0.159	0.159	0.200	0.059
Enterprise Miner	0.176	0.174	0.155	0.155	0.151	0.151	0.221	0.059
GhostMiner	0.110	0.102	0.124	0.124	0.069	0.069	0.069	0.235
Insightful Miner	0.142	0.168	0.135	0.135	0.108	0.108	0.097	0.235
Intelligent Miner	0.106	0.130	0.112	0.112	0.155	0.155	0.097	0.059
Megaputer	0.086	0.091	0.090	0.090	0.060	0.060	0.067	0.235
Oracle Data Miner	0.102	0.103	0.104	0.104	0.151	0.151	0.148	0.059
SQL Server 2005	0.066	0.078	0.126	0.126	0.147	0.147	0.079	0.059

Synthesized Weights – continued	Reliability		Training and Support\Price		Testability	
	Large	**Small**	**Large**	**Small**	**Large**	**Small**
Clementine	0.194	0.194	0.040	0.036	0.156	0.156
Enterprise Miner	0.206	0.206	0.027	0.025	0.149	0.149
GhostMiner	0.069	0.069	0.170	0.194	0.116	0.116
Insightful Miner	0.107	0.107	0.261	0.266	0.147	0.147
Intelligent Miner	0.110	0.110	0.029	0.024	0.099	0.099
Megaputer	0.147	0.147	0.261	0.349	0.151	0.151
Oracle Data Miner	0.101	0.101	0.108	0.051	0.096	0.096
SQL Server 2005	0.066	0.066	0.104	0.054	0.086	0.086

In order to make this project more accurate and realistic, we have combined different scenarios. Factors, such as size, budget, type of business, and the type of data we have to manipulate, can affect the software we attempt to choose and the reasons why we choose it. Having simulated many scenarios, we found that size is a decisive factor. For instance, if we choose two companies with different sizes and follow a traditional reasoning, the tentative result may contradict with our intuition.

After researching all the alternatives, we used our decision tool, *Expert Choice*, to make pair-wise comparisons for each of them. A total of 392 pair-wise comparisons were required to compare each of the alternatives with respect to each of the criteria (196 comparisons for each scenario). This was in addition to the 42 pair-wise comparisons required to assign weights to each criterion with respect to the goal (21 comparisons for each scenario). After completing all of the pair-wise comparisons the software synthesized all of the weights of the alternatives with the weights of the criteria and selected the best alternative of each of the two scenarios. Table 7 below summarizes the weights of the pair-wise comparisons for each of the alternatives with respect to each criterion. As with the weights of the criteria, there is also a direct relationship between the calculated weights in each column and the respective criterion. In other words, the higher a number is in a given column the more important that sized company views that particular software for that specific criterion.

Based on our research of all the alternatives and the weighted criteria calculated by our decision tool, the software has determined that for a small-sized company the top three alternatives are Insightful Miner, Megaputer PolyAnalyst, and SAS Enterprise Miner. These results were somewhat unexpected. We had anticipated that Insightful, Megaputer, and GhostMiner would be among the top three or four alternatives for a small-sized company, primarily because each of these vendors offer stand-alone versions of their software and are also the least expensive among all the alternatives. However, GhostMiner ranked lower than expected while SAS Enterprise Miner ranked higher. A closer analysis of the pair-wise comparisons shows that SAS was more efficient and had better human engineering than GhostMiner. Because of the weights given to these two criteria, SAS Enterprise Miner was able to beat all of the other alternatives despite its cost.

According to our decision tool, the top three alternatives for a large company are SPSS Clementine, followed by SAS Enterprise Miner and Insightful Miner. There was not one overwhelming choice for a large company. The differences in weight among the top three alternatives were relatively small (16.7%, 15.9%, and 13.5%, respectively). These results were also a little unanticipated. All of the software reviews that we have read rated SPSS and SAS among the top leading data mining software that are commercially available. We had not expected Insightful Miner to rank among the top three alternatives for a large company. We anticipated that Microsoft, IBM, or Oracle would round out the top three alternatives because these vendors offer enterprise class DBMS. Upon closer examination of the software analysis, Insightful Miner's visualization and modeling nodes were comparable to those of an enterprise class data mining software program such as Enterprise Miner and Clementine. In addition, a closer review of the pair-wise comparisons showed that Insightful Miner ranked higher than the other alternatives (with the exception of SAS and SPSS) in human engineering and efficiency. It also tied for top weight for Training and Support/Price. Table 8 below summarizes the results of the Expert Choice software.

Certainly, different companies may have different priorities, preferences, and prerequisites. We have explored a few individual scenarios.

Special Case 1: This is a large international company, with thousands of employees, doing business between U.S., Mexico and other countries in Latin America. It is an import and export company. The main goal of using data mining software is to determine the best distribution methods in order to maximize profits. Keeping costs down so it can compete with other companies is always a concern. The three main criteria this company is looking for in a software package are: Training and Support with multilingual support because of the language difference between the countries, Human Engineering to assure that employees in different countries with possibly different computer skill levels, will be able to adapt and use the software, and Portability because it is an established company with an IT department and different platforms that include Microsoft, IBM and Sun Microsystems, as well as a range of desktop operating systems that includes Windows 2000 and XP, Linux, and old legacy equipment. They must be certain that the software is compatible with all existing platforms. Because large amounts of data are processed, software must be robust and reliable as well.

- **Our goal is:** To find the best data mining software.
- **Our criteria are:** Listed seven factors. The three most important criteria are Training and Support, Human engineering, and Portability.
- **Our alternatives are:** Clementine, Enterprise Miner, Oracle, Microsoft SQL, IBM DB2, Salford CART, Megaputer, and Insightful Miner.

Criteria	Weights
Portability	.215
Modifiability	.181
Training and Support	.147
Human Engineering/Testability	.130
Understandability	.119
Reliability	.109
Efficiency	.100

Software	Ranking
Clementine	.136
Enterprise Miner	.126
Oracle	.162
Microsoft SQL	.109
IBM DB2	.117
Salford CART	.123
Megaputer	.111
Insightful Miner	.117

Special Case 2: This is a large national corporation, between 500 and 1,000 employees, in the retail industry with many branches throughout the country. They already have an IT department and different types of platforms, including Unix Servers and Microsoft 2000 and 2003, as well as XP and 2000 Workstations. Portability is very important to make sure that the software is able to run with the platforms already in place. This company already has a well-established customer base, so the goal of choosing data mining software is to find the best way to maximize customer retention, while lowering costs. The three most important criteria that this company is looking for in a software package are: Portability, Efficiency to assure it supports the general business goal assumption, and Modifiability because it is a growing business, and they want to be sure that they can go back and customize the software if necessary.

- **Our goal is:** To find the best data mining software.
- **Our criteria are:** Listed seven factors. The three most important criteria are portability, efficiency, and modifiability.
- **Our alternatives are:** Clementine, Enterprise Miner, Oracle, Microsoft SQL, IBM DB2, Salford CART, Megaputer, and Insightful Miner.

Criteria	Weights
Modifiability	.217
Portability	.171
Efficiency	.153
Reliability	.132
Training and Support	.126
Human Engineering/Testability	.123
Understandability	.078

Software	Ranking
Clementine	.134
Enterprise Miner	.127
Oracle	.163
Microsoft SQL	.110
IBM DB2	.115
Salford CART	.125
Megaputer	.108
Insightful Miner	.117

Special Case 3: This is a small start-up landscaping and construction company with less than 50 employees. The employees have limited knowledge of computers and software. The goal of using data mining software is to find the best way to attract new customers.

- **Our goal is:** To find the best data mining software.
- **Our: criteria are:** Listed seven factors. The three most important criteria we are looking for are Human engineering, Training and Support, and Understandability.
- **Our alternatives are:** Clementine, Enterprise Miner, Oracle, Microsoft SQL, IBM DB2, Salford CART, Megaputer, and Insightful Miner.

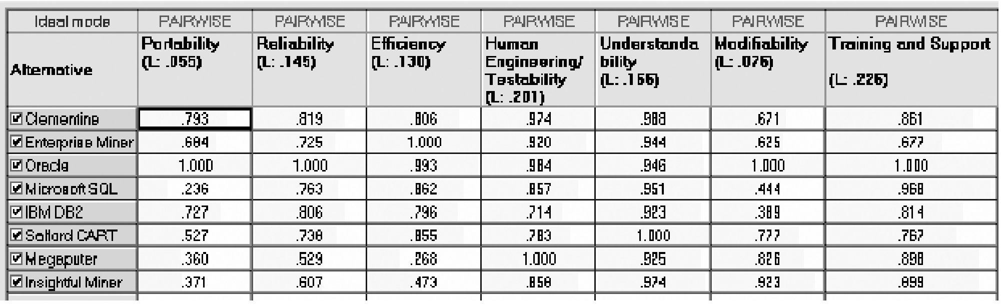
Criteria	Weights
Training and Support	.226
Human Engineering/Testability	.201
Understandability	.166
Reliability	.145
Efficiency	.130
Modifiability	.076
Portability	.055

Software	Ranking
Clementine	.132
Enterprise Miner	.124
Oracle	.150
Microsoft SQL	.124
IBM DB2	.117
Salford CART	.122
Megaputer	.114
Insightful Miner	.118

Other Cases

Online Company/E-commerce

An online/e-commerce company in the recently growing industry: First of all, because an Internet company has both actual and potential customers, it needs a tool that can hold and analyze



Ideal mode	PAIRWISE	PAIRWISE	PAIRWISE	PAIRWISE	PAIRWISE	PAIRWISE	PAIRWISE
Alternative	Portability (L: .055)	Reliability (L: .145)	Efficiency (L: .130)	Human Engineering/ Testability (L: .201)	Understanda bility (L: .166)	Modifiability (L: .076)	Training and Support (L: .226)
☑ Clementine	.793	.819	.806	.974	.988	.671	.861
☑ Enterprise Miner	.684	.725	1.000	.920	.944	.625	.677
☑ Oracle	1.000	1.000	.993	.984	.946	1.000	1.000
☑ Microsoft SQL	.236	.763	.862	.857	.951	.444	.968
☑ IBM DB2	.727	.806	.796	.714	.923	.389	.814
☑ Salford CART	.527	.738	.855	.783	1.000	.777	.767
☑ Megaputer	.360	.529	.268	1.000	.925	.826	.898
☑ Insightful Miner	.371	.607	.473	.858	.974	.923	.899

large amounts of data. Secondly, it might have engineers or a technical department, so it may not put weight on human engineering and training and support. Consequently, we put more weight on modifiability and less weight on engineering and support. As a result, Oracle would be the best tool for an online/e-commerce company because it scores the highest among eight tools. If Oracle is not available, IBM would be the second choice.

Educational Institutions

Data mining software is used worldwide in the educational industry. One of Megaputer's data mining software called PolyAnalyst gets a majority of its business from educational industry. Microsoft SQL came in first place with GhostMiner as the runner up.

Even though these scenarios can be used as references, they did not apply to every type of institution. Thus, it will be interesting to see what other choices are available in term of the best data mining software. What would be the best data mining software for a medical institution?

CONCLUSION

With the use of *Expert Choice* we were able to analytically evaluate eight products within a complex yet controlled environment. The detailed analysis included prioritizing our constraints, evaluating the contributing criteria, entering comparative data and performing relevant sensitivity analysis. The software, *Expert Choice,* performed the analysis, based on our definition, priorities and data.

Data mining technology is changing very rapidly. Our article focused only on the major suppliers typically available in the market place. There is no definite and explicit answer as to which tool is better suited to potential clients, mainly due to their unique priorities. As there are so many variables to quantify, the problem needs to be defined. Based on what approach the problem requires, then and only then can tools start being quantified. Certainly, the method and the process that we have used can be easily applied to analyze and compare other data mining software for each potential user. Although there is no pattern for pairing the correct software with the proper institution, with the use of this process, every institution should be able to determine which data mining software is the best for their operations.

ACKNOWLEDGMENT

We would like to thank Dr. Jennex, the *Edit-in-Chief of IJKM*, for his tremendous help and guidance during the period of revising our manuscript for more than one year. Dr. Jennex has suggested *Scenario Analysis*, a practical and wonderful idea, and offered many other specific suggestions.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Knowledge management systems: Emerging views and practices from the field. In *Proceedings of the 32nd Hawaii International Conference on Systems Sciences. IEEE Computer Society.*

Angus, J. (2006). Clementine 8.1 melds BA with BI. *InfoWorld*, *26*(19), 28-29.

Bednarz, A. (2005). Microsoft beefs up SQL Server database. *Network World, 22*(13), 12.

Berger, C., & Haberstroh, B. (2005). *Oracle data mining 10g release 2: Know more, do more, spend less*. Oracle White Papers. Retrieved November 8, 2007 from http://www.oracle.com/technology/products/bi/odm/pdf/bwp_db_odm_10gr2_0905.pdf

Corral, K., Griffin, J., & Jennex, M.E. (2005). Expert's perspective: The potential of knowledge management in Data Warehousing. *Business Intelligence Journal*, *10*(1), 36-40.

Davenport, T., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business School Press.

Deal, K. (2004). The quest for prediction. *Marketing Research*, *16*(4), 45-47.

Elder IV, J.F., & Abbott, D.W. (1998, August 28). A comparison of leading data mining tools. In *Proceedings of the Fourth International Conference on Knowledge Discovery & Data* 5*Mining*, New York.

Expert Choice Inc. (2007). *Expert Choice 11*. Retrieved November 8, 2007, from http://www.expertchoice.com/software/

Fontana, J. (2005). Microsoft's future in BI market unclear. *Network World*, *22*(43), 9-14.

Giraud-Carrier, C., & Povel, O. (2003). Characterizing data mining software. *Intelligent Data Analysis*, *7*(3), 181-192.

Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, A. (2003). A review of software packages for data mining. *The American Statistician*, *57*(4), 290-309.

Hemaida, R., & Schmits, J. (2006). An analytical approach to vendor selection. *Industrial Management*, *48*(3), 18-24.

James, G., Hakim, J., Chandras, R., King, N., & Variar, G. (2004). Reviewers' choice: Only the best survive. *Intelligent Enterprise*, *7*(1), 34-38.

Jennex, M.E. (2006, April). Technologies in support of knowledge management systems. In *Proceedings of the 6th International Forum on Knowledge Management*, Tunis.

Lampe, J. C., & Garcia, A. (2004). Data mining: An in-depth look. *Internal Auditing, 19*(2), 4-20.

Lurie, I. (2004). Product Review: Insightful Miner. *DM Review*, *14*(6), 88.

Martin, W. E. (2005). *Managing information technology* (5th ed.). Saddle River, NJ: Prentice Hall.

Mena, J. (1998). Data mining FAQ's. *DM Review.*

O Chan, J. (2000). Enterprise information system strategy and planning. *Journal of American Business, Cambridge*, *6*(2), 148-154.

Porter, M. E., & Miller, V. (2001). Strategy and the Internet. *Harvard Business Review*, *72*(3), 62-68.

Roper-Lowe, G. C., & Sharp, J. A. (1990). The analytic hierarchy process and its application to an information technology decision. *The Journal of the Operational Research Society*, *41*(1), 49-59.

Saaty, T.L. (1980). *Multicriteria decision making: The analytic hierarchy process*. RWS Publications.

Saaty, T.L. (1996). *Decision making with dependence and feedback: The analytic network process.* Pittsburgh, PA: RWS Publications.

Saaty, T.L. (2001). *The analytic network process* (2nd version). Pittsburgh, PA: RWS Publications.

Saaty, T.L. (2005). *Theory and applications of the analytic network process*. Pittsburgh. PA: RWS Publications.

Saaty, T.L., & Vargas, L.G. (2006). *Decision making with the analytic network process: Economic, political, social and technological applications with benefits, opportunities, costs and risks*. New York: Springer-Verlag.

This work was previously published in International Journal of Knowledge Management, Vol. 4, Issue 2, edited by M. E. Jennex, pp. 17-34, copyright 2008 by IGI Publishing, formerly known as Idea Group Publishing (an imprint of IGI Global).