Data Mining

Mark Brown, SAS Institute Inc.
John Brocklebank, SAS Institute Inc.

This subject of this paper is ***Data Mining****, an area of ongoing development at SAS Institute. At the opening session of SUGI 22,* ***a new Data Mining Solution*** *from SAS Institute was announced and demonstrated. At the presentation time for this paper a* ***technical overview*** *of the new data mining solution will be presented, not the information below. The contents of this overview are available on the WWW at* ***http://www.sas.com/datamining/***

Due to development efforts beyond the deadline for these proceedings, a business overview of data mining is provided below. However, for the latest information of the ***features and benefits*** *of the new data mining solution, attend this paper presentation and visit the SAS Institute web site.*

Overview of Data Mining

To convert the value of the data warehouse or data mart into strategic business information, many companies are turning to data mining, an emerging technology based on a new generation of software. Data mining combines techniques including statistical analysis, visualization, induction, and neural networks to explore large amounts of data and discover relationships and patterns that shed light on business problems. In turn, companies can use these findings for more profitable, proactive decision making and competitive advantage.

Data mining was designed for exploiting massive amounts of data. This process can be more efficient if you first define what the business problem is, and then determine the amount of data you will need to solve the problem. By taking this "bottom up" approach to data mining and involving upper management in the understanding of business problems and the potential ROI, the process will be much more acceptable and the goals attainable.

SAS Institute defines data mining as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns for a business advantage. As a sophisticated decision support tool, data mining is a natural outgrowth of a business' investment in data warehousing. The data warehouse provides a stable, easily accessible repository of information to support dynamic business intelligence applications.

As the next step, organizations employ data mining to explore and model relationships in the large amounts of data in the data warehouse. Without the pool of validated and "scrubbed" data that a data warehouse provides, the data mining process requires considerable additional effort to pre-process data. Although the data warehouse is an ideal source of data for data mining activities, the Internet can also serve as a data source. Companies can take data from the Internet, mine the data, and distribute the findings and models throughout the company via an Intranet.

Although data mining tools have been around for many years, data mining became feasible in business only after new hardware and software technology advances became available.

Hardware advances--reduced storage costs and increased processor speed--paved the way for data mining's large-scale, intensive analyses. Inexpensive storage also encouraged businesses to collect data at a high level of detail, consolidated into records at the customer level.

Software advances continued data mining's evolution. With the advent of the data warehouse, companies could successfully analyze their massive databases as a coherent, standardized whole. To exploit these vast stores of data in the data warehouse, new exploratory and modeling tools--including data visualization, neural networks, and decision trees--were developed. Finally, data mining incorporated these tools into a systematic, iterative process.

SEMMA

Data mining is often seen as an unstructured collection of methods, or as one or two specific analytic tools, such as neural networks. However, data mining is not a single technique, but an iterative process in which many methods and techniques may be appropriate. And--like data warehousing--data mining requires a systematic approach.

Beginning with a statistically representative sample of the data, you can apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and affirm the model's accuracy.

To clarify the data mining process, SAS Institute has mapped out an overall plan for data mining. This step-by-step process is referred to by the acronym SEMMA: sample, explore, modify, model, and assess.

Step 1: Sample

Extract a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly.

For optimal cost and performance, SAS Institute advocates a sampling strategy, which applies a reliable, statistically representative sample of the full detail data. Mining a representative sample instead of the whole volume drastically reduces the processing time required to get crucial business information.

If general patterns appear in the data as a whole, these will be traceable in a representative sample. If a niche is so tiny that it's not represented in a sample and yet so important that it influences the big picture, it can be discovered using summary methods.

Step 2: Explore

Search speculatively for unanticipated trends and anomalies so as to gain understanding and ideas.

After sampling your data, the next step is to explore them visually or numerically for inherent trends or groupings. Exploration helps refine the discovery process.

If visual exploration doesn't reveal clear trends, you can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering. For example, in data mining for a direct mail campaign, clustering might reveal groups of customers with distinct ordering patterns. Knowing these patterns creates opportunities for personalized mailings or promotions.

Step 3: Modify

Create, select, and transform the variables to focus the model construction process.

Based on your discoveries in the exploration phase, you may need to manipulate your data to include information such as the grouping of customers and significant subgroups, or to introduce new variables. You may also need to look for outliers and reduce the number of variables, to narrow them down to the most significant ones.

You may also need to modify data when the "mined" data change. Because data mining is a dynamic, iterative process, you can update data mining methods or models when new information is available.

Step 4: Model

Search automatically for a variable combination that reliably predicts a desired outcome.

Once you prepare your data, you are ready to construct models that explain patterns in the data. Modeling techniques in data mining include neural networks, tree-based models, logistic models, and other statistical models--such as time series analysis and survival analysis.

Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data. For example, neural networks are good at combining information from predictors which support nonlinear associations with a target.

Step 5: Assess

Evaluate the usefulness and reliability of findings from the data mining process.

The final step in data mining is to assess the model to estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage sometimes known as validation data. For a model to be considered successful and useful, it should work for this validation sample as well as for the training data used to construct the model.

Similarly, you can test the model against known data. For example, if you know which customers in a file had high retention rates and your model predicts retention, you can check to see whether the model selects these customers accurately. In addition, practical applications of the model, such as partial mailings in a direct mail campaign, help prove its validity.

The Future

By all accounts, data mining is a technology that is quickly gaining momentum in the marketplace. The Gartner Group estimates that over the next 10 years the use of data mining in target marketing applications will increase from less than 5% to more than 80%. The META Group estimates that the data mining market will grow to $300 million by 1997 and to $800 million by the year 2000.

However, the real promise of data mining is that software products will increasingly be focused on business solutions. Data mining functionality will be packaged to integrate seamlessly with existing data warehouse and business intelligence software--with the accent on solving business problems rather than on the enabling technology. As a result, organizations using data mining techniques will be able to understand key business issues more thoroughly and to present the results of analysis meaningfully to specialist marketing analysts and general users alike.

In learning more about themselves and their customers, these organizations will see a shift towards true one-to-one relationships with the customers--ensuring complete customer relationship management. Accurate anticipation of the customers' actions can lead to increased effectiveness of marketing activities and decreased financial risks.

The core technology for data mining already exists--in fact, thousands of SAS Institute customers have been using it for years. What is new is the application of these techniques--in methodical ways--to the data warehouses that many organizations have built in recent years.

For more information and success stories of how different companies are using SAS for data mining including Reader’s Digest, Mellon Bank, and Meredith Corporation, check the Internet at: **http://www.sas.com/feature/4qdm/dmprofiles.html**