

Data Mining Overview

By: Dr. Michael Gilman, CEO, [Data Mining Technologies Inc.](#)

With the proliferation of data warehouses, data mining tools are flooding the market. Their objective is to discover hidden gold in your data. Many traditional report and query tools and statistical analysis systems use the term "data mining" in their product descriptions. Exotic Artificial Intelligence-based systems are also being touted as new data mining tools. What is a data mining tool and what isn't?

The ultimate objective of data mining is knowledge discovery. Data mining methodology extracts hidden predictive information from large databases. With such a broad definition, however, an online analytical processing (OLAP) product or a statistical package could qualify as a data mining tool.

Data mining methodology extracts hidden predictive information from large databases.

That's where technology comes in: for true knowledge discovery a data mining tool should unearth hidden information automatically. By this definition data mining is data-driven, not user-driven or verification-driven.

Verification-Driven Data Mining: An Example

Traditionally the goal of identifying and utilizing information hidden in data has proceeded via query generators and data interpretation systems. A user formats a theory about a possible relation in a database and converts this hypothesis into a query.

For example, a user might hypothesize about the relationship between industrial sales of color copiers and customers' specific industries. He or she would generate a query against a data warehouse and segment the results into a report. Typically, the generated information provides a good overview.

This verification type of data mining is limited in two ways, however. First, it's based on a hunch. In our example, the hunch is that the industry a company is in correlates with the number of copiers it buys or leases. Second, the quality of the extracted information depends on the user's interpretation of the results—and is thus subject to error.

Multifactor analyses of variance and multivariate analyses identify the relationships among factors that influence the outcome of copier sales. Pearson product-moment correlations measure the strength and direction of the relationship between each database field and the dependent variable.

One of the problems with this approach, aside from its resource intensity, is that the techniques tend to focus on tasks in which all the attributes have continuous or ordinal values. Many of the attributes are also parametric. A linear classifier, for instance, assumes that a relationship is expressible as a linear combination of the attribute values.

Statistical methodology assumes normally distributed data—an often tenuous assumption in the real world of corporate data warehouses.

Manual vs. Automatic

Manual data mining stems from the need to know facts, such as regional sales reports stratified by type of business—while automatic data mining comes from the need to discover the factors that influence these sales.

One way to identify a true data mining tool is by how it operates on the data: is it manual (top-down) or automatic (bottom-up)? In other words, who originates the query, the user or the software?

Even some sophisticated AI-based tools that use case-based reasoning, a nearest neighbor indexing system, fuzzy (continuous) logic, and genetic algorithms don't qualify as data mining tools since their queries also originate with the user. Certainly the way these tools optimize their search on a dataset is unique, but they do not perform autonomous data discovery. Neural networks, polynomial networks, and symbolic classifiers, on the other hand, do qualify as true automatic data mining tools because they autonomously interrogate the data for patterns.

Neural networks, however, often require extensive care and feeding—they can only work with preprocessed numeric, normalized, scaled data. They also need a fair amount of tuning such as the setting of a stopping criterion, learning rates, hidden nodes, momentum coefficients, and weights. And their results are not always comprehensible.

Another Paradigm

Symbolic classifiers that use machine learning technology hold great potential as data mining tools for corporate data warehouses. These tools do not require any manual intervention in order to perform their analysis. Their strength is their ability to automatically identify key relationships in a database -- to discover rather than confirm trends or patterns in data and to present solutions in usable business formats. They can also handle the type of real-world business data that statistical and neural systems have to "scrub" and scale.

Most of these symbolic classifiers are also known as rule-induction programs or decision-tree generators. They use statistical algorithms or machine-learning algorithms such as ID3, C4.5, AC2, CART, CHAID, CN2, or modifications of these algorithms. Symbolic classifiers split a database into classes that differ as much as possible in their relation to a selected output. That is, the tool partitions a database according to the results of statistical tests conducted on an output by the algorithm instead of by the user. Machine learning algorithms use the data -- not the user's hypothesis -- to automate the stratification process.

To start the process, this type of data mining tool requires a "dependent variable" or outcome, such as copier sales, which should be a field in the database. The rest is

automatic. The tool's algorithm tests a multitude of hypotheses in an effort to discover the factors or combination of factors, (e.g., business type, location, number of employees) that have the most influence on the outcome.

The algorithm engages in a kind of "20 Questions" game. Presented with a database of 5,000 buyers and 5,000 nonbuyers of copiers, the algorithm asks a series of questions about the values of each record. Its goal is to classify each sample into either a buyer or nonbuyer group.

The tool processes every field in every record in the database until it sufficiently splits the buyers from the nonbuyers and learns the main differences between them. Once the tool has learned the crucial attributes it can rank them in order of importance. A user can then exclude attributes that have little or no effect on targeting potential new customers.

Rule Generation

Most data mining tools generate their findings in the format of "if then" rules. Here's an example of a data mining process that discovers ranges for targeting potential product buyers.

```
CONDITIONA
IF CUSTOMER SINCE = 1978 through 1994
AND REVOLVING L/M/T = 5120 through 8900
AND CREDIT/DEBITRAT/O =67
THEN Potential Buyer = 89%
CONDITIONZ
IF CUSTOMER SINCE= 1994 through 1996
AND REVOLVING LIMIT = 1311 through 5120
AND CREDIT/DEB/TRAT/O =67
THEN Potential Buyer=49%
```

Advantages of Symbolic Classifiers

Symbolic classifiers do not require an intensive data preparation effort. This is a convenience to end-users who freely mix numeric, categorical, and date variables.

Another advantage of these tools is the breadth of the analyses they provide. Unlike traditional statistical methods of data analysis which require the user to stratify a database into smaller subgroups in order to maximize classification or prediction, data mining tools use all the data as the source of their analysis.

Still another advantage is that these tools formulate their solutions in English. They can extract "if-then" business rules directly from the data based on tests they conduct for statistical significance. They can optimize business conditions by providing answers to decision-makers on important questions.

Almost all of the current symbolic classifier-type data mining tools incorporate a methodology for explaining their findings. They also tabulate model error-rates for estimating the goodness of their predictions. In a business environment where small

changes in strategy translate to millions of dollars, this type of insight can quickly equate to profits. Some of these tools can also generate graphic decision trees which display a summary of significant patterns and relationships in the data.

The Bottom Line

Many of today's analytic tools have tremendous capabilities for performing sophisticated user-driven queries. They are, however, limited in their abilities to discover hidden trends and patterns in a database. Statistical tools can provide excellent features for describing and visualizing large chunks of data, as well as performing verification driven data analysis. Autonomous data mining tools, however, based on machine-learning algorithms, are the only tools designed to automate the process of knowledge discovery.

The Ten Steps of Data Mining

Here is a process for extracting hidden knowledge from your data warehouse, your customer information file, or any other company database.

1. Identify The Objective

Before you begin, be clear on what you hope to accomplish with your analysis. Know in advance the business goal of the data mining. Establish whether or not the goal is measurable. Some possible goals are to

- find sales relationships between specific products or services
- identify specific purchasing patterns over time
- identify potential types of customers
- find product sales trends

2. Select The Data

Once you have defined your goal, your next step is to select the data to meet this goal. This may be a subset of your data warehouse or a data mart that contains specific product information. It may be your customer information file. Segment as much as possible the scope of the data to be mined.

Here are some key issues:

- Are the data adequate to describe the phenomena the data mining analysis is attempting to model?
- Can you enhance internal customer records with external lifestyle and demographic data?
- Are the data stable—will the mined attributes be the same after the analysis?
- If you are merging databases can you find a common field for linking them?
- How current and relevant are the data to the business goal?

3. Prepare The Data

Once you've assembled the data, you must decide which attributes to convert into usable formats. Consider the input of domain experts—creators and users of the data.

- Establish strategies for handling missing data, extraneous noise, and outliers
- Identify redundant variables in the dataset and decide which fields to exclude
- Decide on a log or square transformation, if necessary
- Visually inspect the dataset to get a feel for the database
- Determine the distribution frequencies of the data

You can postpone some of these decisions until you select a data mining tool. For example, if you need a neural network or polynomial network you may have to transform some of your fields.

4. **Audit The Data**

Evaluate the structure of your data in order to determine the appropriate tools.

- What is the ratio of categorical/binary attributes in the database?
- What is the nature and structure of the database?
- What is the overall condition of the dataset?
- What is the distribution of the dataset?

Balance the objective assessment of the structure of your data against your users' need to understand the findings. Neural nets, for example, don't explain their results.

5. **Select The Tools**

Two concerns drive the selection of the appropriate data mining tool—your business objectives and your data structure. Both should guide you to the same tool.

Consider these questions when evaluating a set of potential tools.

- Is the data set heavily categorical?
- What platforms do your candidate tools support?
- Are the candidate tools ODBC-compliant?
- What data format can the tools import?

No single tool is likely to provide the answer to your data mining project. Some tools integrate several technologies into a suite of statistical analysis programs, a neural network, and a symbolic classifier.

6. **Format The Solution**

In conjunction with your data audit, your business objective and the selection of your tool determine the format of your solution.

The Key questions are

- What is the optimum format of the solution—decision tree, rules, C code, SQL syntax?
 - What are the available format options?
 - What is the goal of the solution?
 - What do the end-users need—graphs, reports, code?
7. **Construct The Model**

At this point that the data mining process begins. Usually the first step is to use a random number seed to split the data into a training set and a test set and construct and evaluate a model. The generation of classification rules, decision trees, clustering sub-groups, scores, code, weights and evaluation data/error rates takes place at this stage. Resolve these issues:

- Are error rates at acceptable levels? Can you improve them?
 - What extraneous attributes did you find? Can you purge them?
 - Is additional data or a different methodology necessary?
 - Will you have to train and test a new data set?
8. **Validate The Findings**

Share and discuss the results of the analysis with the business client or domain expert. Ensure that the findings are correct and appropriate to the business objectives.

- Do the findings make sense?
 - Do you have to return to any prior steps to improve results?
 - Can use other data mining tools to replicate the findings?
9. **Deliver The Findings**

Provide a final report to the business unit or client. The report should document the entire data mining process including data preparation, tools used, test results, source code, and rules. Some of the issues are:

- Will additional data improve the analysis?.
 - What strategic insight did you discover and how is it applicable?
 - What proposals can result from the data mining analysis?
 - Do the findings meet the business objective?
10. **Integrate The Solution**

Share the findings with all interested end-users in the appropriate business units. You might wind up incorporating the results of the analysis into the company's business procedures. Some of the data mining solutions may involve

- SQL syntax for distribution to end-usersC
- code incorporated into a production system
- Rules integrated into a decision support system.

Although data mining tools automate database analysis, they can lead to faulty findings and erroneous conclusions if you're not careful. Bear in mind that data mining is a business process with a specific goal—to extract a competitive insight from historical records in a database.

applications pertaining to Internet e-commerce and direct marketing, healthcare, stock prediction and financial services. Our core product is Nuggets®, a desktop data mining toolkit, using the most powerful rule induction engine on the market