: What are common data mining tools?

by Dan Power

Editor, DSSResources.com

As it became possible and affordable to store large amounts of data, managers wanted to access and analyze the proprietary data gathered by their company to identify patterns that would be useful in making decisions. Sometimes data is analyzed as part of a special study; in other cases the analysis is conducted periodically or routinely. Patterns that are identified may be incorporated into a DSS or may inform a specific decision. The data mining process involves identifying an appropriate data set to examine or sift through to discover data content relationships. Data mining tools include case-based reasoning, data visualization, fuzzy query and analysis, genetic algorithms, and neural networks (cf., Greenfield, 1999). Let's briefly review each of the common data mining tools.

Case-Based Reasoning

According to Wikipedia, "Case-based reasoning (CBR), broadly construed, is the process of solving new problems based on the solutions of similar past problems." Case-based tools find cases/records in a database that are similar to a specified pattern. A user specifies how strong a relationship should be to identify a case match. This approach has also been called memory-based reasoning. Software tries to measure the "distance" based on a measure of one record to other records and cluster records by similarity. In general, CBR is an empirical classification method.

Data Visualization

These tools graphically display complex relationships in multi-dimensional data from different perspectives. Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. Data visualization tools are data mining tools that translate complex formulas, mathematical relationships or data warehouse information into graphs or other easily understood models. Statistical tools like cluster analysis or classification and regression trees (CART) are often part of data visualization tools. Analysts can visualize the clusters or examine a binary tree created by classifying records. In marketing, an analyst may create "co-occurrence" tables or charts of products that are purchased together. A good visualization is easy to understand and interpret and it is a reasonably accurate representation of the relationships in underlying data.

Fuzzy Query and Analysis

Fuzzy data mining tools allow users to look at results that are "close" to specified criteria. The user can vary what the definition of "close" is to help determine the significance and number of results that will be returned. This category of data mining tools is based on a branch of mathematics called fuzzy logic. The logic of uncertainty and "fuzziness" provides a framework for finding, scoring, and ranking the results of queries. Examples of fuzzy criteria include AGE = "VERY YOUNG" or SALARY = "MORE OR LESS HIGH".

Genetic Algorithms

Genetic algorithms are optimization programs similar to linear programming models. In general a genetic algorithm (GA) is a search technique to find a solution. Genetic algorithm software automatically conducts random experiments with new solutions while keeping the "good" interim results. A example problem would be to find the best subset of 20 variables to predict the stock market. To create a genetic model, the 20 variables would be identified as "genes" that have at least 2 possible values. The software would then select genes and their values randomly in an attempt to maximize of minimize a performance or fitness function. The performance function would provide a value for the fitness of the specific genetic model. Genetic optimization software also includes operators to combine and mutate variables.

Neural Networks

Neural network tools are used to predict future information by learning patterns and then applying them to predict future relationships. According to Berry and Linoff (1997), neural networks are the most common type of data mining technique. Some people even think that using a neural network is the only type of data mining. Vendors make many claims for neural networks. One claim that is especially questionable is that neural networks can compensate for a lower quality of data. Neural networks attempt to learn patterns from data directly by repeatedly examining the data to identify relationships and build a model. The algorithm builds models by trial and error. The network guesses a value that it compares to the actual number. If the guess is wrong, the model is adjusted. This process involves three iterative steps: 1) predict, 2) compare, and 3) adjust. An artificial neural network involves a network of simple processing elements that can exhibit complex behavior, determined by the connections between the processing elements and element parameters. Neural networks are commonly used in a DSS to classify data and, as noted, to make predictions. The various inputs are transformed by a network of simple processors. The processors combine and weight the inputs and produce an output value.

: What are common data mining tools?

Data mining techniques and tools are **NOT** fundamentally different from the older quantitative model-building techniques. The methods used in data mining are extensions and generalizations of analytical methods known for decades. Neural networks are a special case of what is called projection pursuit regression, a method developed in the 1940s. For example, classification and regression tree (CART) methods were used by social scientists in the 1960s. The computing technology used to implement these underlying methods has however greatly improved.

References

Berry, Michael J. A. and Gordon Linoff. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: Wiley Computer Publishing, 1997.

Dhar, V. and R. Stein, Intelligent Decision Support Methods: The Science of Knowledge, Upper Saddle River, NJ: Prentice-Hall, 1997.

Greenfield, Larry. Data Mining. LGI Systems, Inc. January 12, 2000.

Power, D., What is data mining and how is it related to DSS? DSS News, Vol. 2, No. 25, December 2, 2001.

Power, D. J. Decision Support Systems Hyperbook. Cedar Falls, IA: DSSResources.COM, HTML version, Fall 2000, accessed on 12/06/2009 at URL http://dssresources.com/dssbook/.

Thearling, Kurt. Data Mining, Decision Support and Database Marketing. (URL http://www3.shore.net/~kht/index.htm)

Thearling, Kurt. Data Mining and Advanced DSS Technology, an on-line Data Mining Tutorial. (URL http://www3.shore.net/~kht/dmintro/dmintro.htm).

Wikipedia. "Case-based reasoning". URL http://en.wikipedia.org/wiki/Case-based_reasoning.

Author: Daniel Power
Last update: 2009-12-06 05:08