: How does data mining provide knowledge?

by Daniel J. Power

Editor, DSSResources.com

In the 1970s, companies employed business analysts who used statistical packages like SAS and SPSS to perform trend analyses and cluster analyses on data to find patterns and draw conclusions. As it became possible and affordable to store large amounts of data, managers wanted to access and analyze transaction data like that generated at a retail store cash register. Bar coding and the World Wide Web have also made it possible for companies to collect large amounts of new data. In 2021, business and data analytics professionals are more common. The Internet of Things (IoT) generates "big data", a large volume of varied, and often volatile and high velocity (streaming) data. If data is accurate, then data mining can provide knowledge and create value.

Much of this discussion is from Power (2002) with a few updates. Data mining is a process that uses a broad range of analytical techniques to find anomalies, patterns and correlations in data sets to predict outcomes and discover knowledge. A related term Machine Learning (ML) is either a supervised or an unsupervised computer-based process that develops an algorithm, i.e., a set of rules, from data sets. Both data mining and machine learning are tools used in business and data analytics projects.

When a statistician looks at data, he or she makes a hypothesis about a relationship, then performs a query on a database and uses statistical techniques to prove or disprove the hypothesis using that data. This has been called the “verification mode” (IBM, 1998). Data mining software is used in a “discovery mode” and an analyst tries to find meaningful patterns. In data mining, a hypothesis is not stated before the data is analyzed.

In general, there are two main kinds of models in data mining: predictive and descriptive. Predictive models can be used to forecast explicit values, based on patterns determined from known results. For example, from a database of customers who have already responded to a particular offer, a model can be built that predicts which prospects are likeliest to respond to the same offer. The predictive model is then used in a DSS. Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters. Once a descriptive model is identified it may be used for target marketing or other decision support tasks.

There are a wide variety of tools for data mining. The decision about which technique to use depends on the type of data and the type of questions that

managers want answered by that data. Let's examine some major tools:

Case-Based Reasoning

Case-based tools find records in a database that are similar to specified records. A user specifies how strong a relationship should be before a new case is brought to her attention. This category of tools is also 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. This technique has been successful in analyzing relationships in free-form text. The Web site www.ai-cbr.org is a resource for the artificial intelligence and case-based reasoning technology fields.

A five-step problem-solving process is used with case-based tools: 1) Presentation: a description of the current problem is input to the system; 2) Retrieval: the system retrieves the closest-matching cases stored in a database of
cases; 3) Adaptation: the system uses the current problem and closest-matching cases to generate a solution to the current problem; 4) Validation: the solution is validated through feedback from the user of the environment; 5) Update: if appropriate, the validated solution is added to the case base for use in
future problem solving (cf., Allen, 1994).

Data Visualization

Visualization tools graphically display complex relationships in multidimensional 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. Decision support 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 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.

Genetic Algorithms

Genetic algorithms are optimization programs similar to the linear programming models. Genetic algorithm software conducts random experiments with new solutions while keeping the “good” interim results. A sample problem is to find the best subset of 20 variables to predict stock market behavior. To create a genetic model, the 20 variables would be identified as “genes” that have at least two possible values. The software would then select genes and their values randomly in an attempt to maximize or 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 genes. This quantitative model finds a specific type of pattern, like other data mining techniques.

Neural Networks

Neural network tools are used to predict future information by learning patterns from past data. 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. For example, with appropriate input data a neural network could be trained to predict the price or net asset value of a mutual fund in the next quarter. Neural networks attempt to learn patterns from data directly by repeatedly examining the data to identify relationships and build a model. Neural networks build 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 learning process involves three iterative steps: predict, compare, and adjust. Neural networks are commonly used in a knowledge-driven DSS to classify data and to make predictions. Neural Networks are a class of models within the general machine learning literature.

Data Mining Process

Data mining and knowledge discovery attempt to identify predictive relationships and provide managers with descriptive information about the subject of a database. There are a number of prescribed data mining processes. To make the best use of data mining, one must first make a clear statement of objectives. Researchers at IBM have described data mining as a three-phase process of data preparation, mining operations, and presentation. Analysts at the Gartner Group describes it similarly as a five-stage process:

1. Select and prepare the data to be mined.

2. Qualify the data via cluster and feature analysis.

3. Select one or more data mining tools.

4. Apply the data mining tool.

5. Apply the knowledge discovered to the company's specific line of business to achieve a business goal (Gerber, 1996).

These alternative processes can guide a special decision support study that uses data mining. In general, the first step is to select and prepare the data to be mined. Some data mining software packages include data preparation tools that can handle at least some of the preparation that needs to be done to the data. The second step is qualifying or testing the data using cluster and feature analysis software. This step takes some business knowledge about the question that one is trying to answer. This is the step where bias in the data should be detected and removed (IBM, 1998). In the third step an appropriate data mining tool is selected and used. Finally, the results are presented to decision makers, and if the results seem useful, a decision is influenced and one hopes business goals are achieved.

References

Power, D. J., Decision Support Systems: Concepts and Resources for Managers, Westport, CT: Greenwood/Quorum Books, 2002, ISBN: 156720497X. See https://scholarworks.uni.edu/facbook/67

: How does data mining provide knowledge?

Author: Daniel Power
Last update: 2021-01-17 02:33