: What are common data mining mistakes?

by Dan Power

Editor, DSSResources.com

Data mining is a broad term for identifying patterns in data, analyzing data and summarizing it into useful information. These tasks are often challenging. Although technology has made it easier to access large data sets and use various tools to analyze the data set, making sense of the data is difficult. Now almost any manager or staffer can conduct analyses that were previously conducted only by statisticians, market researchers and other data analysis professionals. The lack of expertise of many new users increases the likelihood of identifying false relationships and over generalizing from the data. What are common mistakes?

1. **Asking the wrong question.** Data mining should involve testing hypotheses and trying to answer a research question. Defining hypotheses and questions is very important.

2. **Failing to test the reasonableness of the results.** Testing reasonableness means asking does the result make sense. It is not enough to only apply statistical tests of significance.

3. **Ignoring discrepancies in the data** and discarding data points. It is common to have preconceived notions about the patterns in data and ignore divergence in data values. One should be very careful about discarding data points and that should be documented.

4. **Ignoring simple explanations and building overly complex models.** Including more variables in a predictive model may seem better and more realistic, but the complexity of the model may actually limit its usefulness.

5. **Over generalizing from the results**, extrapolating into the future when circumstances have changed. One needs to avoid reaching too general a conclusion from a data set. For example, if the data is from the winter months the relationship may not hold for the summer months.

6. **Using insufficient or inadequate data.** Data analysis is only as good as the data that is analyzed. Emphasize data quality in data collection.

(c) 2022 Daniel J. Power, Power Enterprises <power@dssresources.com>

7. **Using a single data analysis tool.** There are many data mining tools and each tool serves a somewhat different but often complementary purpose.

In 1996 Bill Palace wrote "Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations." According to Wikipedia, "Data mining is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery." In 2010, the various data mining tools are broadly used and misused by many large organizations. Improved training is one solution to improving the quality of data analyses. Awareness of common mistakes also helps one avoid them.

Andrew Moore has a number of Statistical Data Mining Tutorials at www.cs.cmu.edu. The tutorials cover foundations of probability, the foundations of statistical data analysis, and most of the classic machine learning and data mining algorithms. Classification algorithms include decision trees, neural nets, Bayesian classifiers, Support Vector Machines and cased-based (aka non-parametric) learning. They include regression algorithms such as multivariate polynomial regression, MARS, Locally Weighted Regression, GMDH and neural nets. And they include other data mining operations such as clustering (mixture models, k-means and hierarchical), Bayesian networks and Reinforcement Learning.

Data mining techniques can help managers discover hidden relationships and patterns in data. Some analysts feel data mining can help a company gain a competitive advantage. Data mining tools can be used for both hypothesis testing and knowledge discovery. When vendors discuss data mining, they may be selling a set of end-user tools or a decision support capability or both.

References

Palace, B. "Data Mining," Technology Note prepared for Management 274A,
Anderson Graduate School of Management at UCLA, Spring 1996 at URL
http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm .

Power, D. J. Decision Support Systems Hyperbook. Cedar Falls, IA: DSSResources.COM, HTML version, 2000, accessed on (today's date) at URL
http://dssresources.com/subscriber/password/dssbookhypertext.

(c) 2022 Daniel J. Power, Power Enterprises <power@dssresources.com>

Power, D. J., Decision Support Systems: Concepts and Resources for Managers, Westport, CT: Greenwood/Quorum Books, 2002.

Power, D., What is the "true story" about data mining, beer and diapers? DSS News, Vol. 3, No. 23, November 10, 2002.

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

Data mining, from Wikipedia, the free encyclopedia, URL http://en.wikipedia.org/wiki/Data_mining .

Author: Daniel Power
Last update: 2010-12-19 08:02

(c) 2022 Daniel J. Power, Power Enterprises <power@dssresources.com>