Cornell University
Cornell Statistical Consulting Unit

StatNews #55

What is Data Mining
December 2002

Data mining has been increasingly popular in a wide number of fields due to the large volume of data collected and a dramatic need for analysis of these data. For example, in the pharmaceutical industry, data mining is used to analyze massive amounts of data for selecting chemical compounds for producing new drugs. In bioinformatics, data mining techniques are applied to cluster genes by similarities in gene expression. This newsletter discusses some of the most commonly used techniques in data mining.

Data mining is the process of data exploration to extract consistent patterns and relations among variables that can be used to make valid predictions. Data mining is a modern data analysis approach that does not replace traditional statistical techniques; rather it combines statistical methods with increasing computing power to process huge volumes of available data. While parts of data mining can be automated, they form only a small part of the process. Understanding the problem, selecting the relevant data, transforming data to bring the information content to the surface, and interpreting the results are activities that have not been automated and are not likely to be any time soon. Different types of data call for different techniques. For example, if your goal is to predict a value for a continuous response variable, then regression is appropriate. If the desired outcome is assigning a case to a predefined class, then classification is appropriate. Several data mining techniques are discussed next. Neural networks are non-linear predictive models that learn through training and resemble biological neural networks in structure. The advantage of this technique is that it can handle complex problems, with a large number of predictors that have many interactions. The results are not easy to interpret. This method requires that all variables be numeric. The neural network method is a good choice when the miner is more interested in the results of the model than in understanding how the model works.

Decision trees are tree-shaped structures that represent sets of decisions, leading to a class or value. Specific decision tree methods include Classification and Regression Trees (CART) (see StatNews #19) and Chi Square Automatic Interaction Detection (CHAID). The trees produced by these different algorithms differ in the number of splits allowed at each level of the tree, how the splits are chosen, and how tree growing is limited to prevent over-fitting. Decision trees are a good choice when there are a number of input variables and the goal is to classify the data (classification trees) or to make prediction of outcomes (regression trees). This method handles non-numeric data, works well with a large number of variables, and is not sensitive to outliers and skewed distributions. Nearest neighbor method classifies each record in a data set based on a combination of the classes of the records most similar to it. It is relatively easy to use for numeric data, but categorical data require special handling.

Genetic algorithms use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. One generation of models competes in passing on characteristics to the next generation of models, until the best model is found. Genetic algorithms are useful in guiding data mining algorithms such as neural networks and decision trees.

Successful data mining involves a series of decisions that need to be made before the data mining process starts. It involves setting up a clearly defined goal, selecting the type of prediction, e.g. regression or classification, appropriate for the data, and the type of model to be used. For example, one may choose a decision tree model for a classification problem using the CART algorithm. An upcoming newsletter will demonstrate the implementation of a data-mining example using SAS Enterprise Miner (SAS EM).

References:

- Edelstein, Herbert A. (1999). Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation.
- Berry, Michael J.A., Linoff Gordon (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management. Wiley, NY.

Author: Simona Despa

Back to StatNews Table of Contents

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).