Internet Data Mining Using Statistical Techniques

Kuldeep Kumar
Bond University, Australia

INTRODUCTION

Data mining has emerged as one of the hottest topics in recent years. It is an extraordinarily broad area and is growing in several directions. With the advancement of the Internet and cheap availability of powerful computers, data is flooding the market at a tremendous pace. However, the technology for navigating, exploring, visualising, and summarising large databases is still in its infancy. Internet data mining is the process of collecting, analysing, and decision making while the data is being collected on the Internet. In most of the financial applications, data is updated every second or minute or hour. By using Internet data-mining tools, a decision can be made as soon as data are updated.

The quantity and diversity of data available to make decisions have increased dramatically during the past decade. Large databases are being built to hold and deliver these data. Data mining is defined as the process of seeking interesting or valuable information within large data sets. Some examples of data mining applications in the area of management science are analysis of direct mailing strategies, sales data analysis for customer segmentation, credit card fraud detection, mass customization, and so forth. With the advancement of the Internet and World Wide Web, both management scientists and interested end users can get large data sets for their research from this source. The Web not only contains a vast amount of useful information, but also provides a powerful infrastructure for communication and information sharing. For example, Ma, Liu, and Wong (2000) have developed a system called DS-Web that uses the Web to help data mining. A recent survey on Web-mining research can be seen in the paper by Kosala and Blockeel (2000).

Both statistics and data mining are concerned with drawing inferences from data. The aim of inference may be to understand the patterns of correlation and causal links among the data values (explanation) or making predications for the future data values (generalization). At present, data-mining practitioners and statisticians seem to have different approaches to solving problems of a similar nature. It appears that statisticians and data miners can profit by studying each other's methods and using a judiciously chosen combination of them.

Data-mining techniques can be broadly classified into four areas.

1. **Exploratory Data Analysis (EDA) and Inferential Techniques:** As opposed to traditional hypothesis testing designed to verify an a priori hypothesis about relations between variables, EDA is used to identify systemic relationships between variables when there are no a priori expectations as to the nature of those relations. Computational EDA includes both simple and basic statistics and more advanced, multivariate exploratory techniques designed to identify patterns in multivariate data sets.
2. **Sampling Techniques:** Where an incomplete data set is available, sampling techniques are used to make generalizations about the data. Various considerations need to be accounted for when drawing a sample, not the least of which is any a priori knowledge about the nature of the population.
3. **Neural Networks:** Neural networks are analytical techniques modeled after the process of learning in the cognitive system and the neurological functions of the brain. These techniques are capable of predicting new observations from other observations after executing a process of so-called learning from data. One of the major advantages of neural networks is that they are capable of approximating any continuous function and the researcher does not need to have any hypothesis about the underlying model or even, to some extent, which variables matter.
4. **Decision-Tree Techniques:** These techniques successively split the data into subgroups in order to improve the prediction or classification of the dependent variable. These techniques can handle a large number of independent variables and, being nonparametric in nature, can capture the relationship where traditional statistical techniques fail.

Table 1. Summary of exploratory data analysis techniques

Graphical Representation of Data
- Histogram
- Stem and Leaf Plot
- Box-Cox Plot

Descriptive Statistics
- Measures of Central Tendency (Mean, Median, Mode)
- Measures of Dispersion (Variance, Standard Deviation, Coefficient of Variation)
- Skewness
- Kurtosis

Data-Driven Modeling
- Correlation
- Multiple Regression

Data-Reduction Techniques
- Principal Component Analysis
- Factor Analysis

Classification Techniques
- Discriminant Analysis
- Cluster Analysis

Influential Observations

Forecasting Techniques
- Box- Jenkins Analysis
- Exponential Smoothing Techniques
- State Space Modeling

EXPLORATORY DATA ANALYSIS AND DATA-MINING TECHNIQUES

Keogh and Kasetty (2003) have published a survey on time series data mining. References of these techniques can be seen in Flury (1997), Kumar and Baker (2001), McLachlan (1992), Mendenhall and Sincich (2003), and so forth.

SAMPLING TECHNIQUES FOR DATA MINING

Sampling techniques are essential tools for business data mining. Although there are several sampling techniques, simple random sampling is commonly used in practice. There are quite a few drawbacks of using simple random sampling, and it may not be an appropriate technique in many situations. In this section we have reviewed various sampling techniques with their relative merits and demerits. We have also proposed a new sampling technique.

Sampling techniques play a key role in data mining. These techniques are also widely used in market research, economics, finance, and system analysis and design. Sampling is the process of selecting representative elements of a population from a huge database. When these selected elements are examined closely, it is assumed that the selected sample will reveal useful information about the whole database.

For example, given a huge database of customers, it may be difficult to interview each and every one because of the enormous cost and it will also take lots of time. Obviously, the best way is to draw a small representative sample and interview these customers to assess the preferential pattern of consumers for different types of products, the potential demand for a new product, scope