Chapter XV

Internet Data Mining

Kuldeep Kumar
Bond University, Australia

John Baker
Universal College of Learning, New Zealand

Gold mining is a process for sifting through lots of ore to find valuable nuggets. Data mining is a process for discovering patterns and trends in large data sets to find useful decision-making information.

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, visualizing and summarizing large databases are still in their infancy.

The quantity and diversity of data available to make decisions has 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, etc. 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 predictions 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 in three areas:

Exploratory Data Analysis (EDA): As opposed to traditional hypothesis testing designed to verify 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.

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.

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.

EXPLORATORY DATA ANALYSIS AND DATA MINING TECHNIQUES

Graphical Representation of the Data

When a large number of observations (discrete or continuous) is available, it becomes necessary to condense the data without losing any information. The condensed data can be put in a tabular form, known as frequency distribution, or in graphical form. There are quite a few techniques available to represent data in a graphical form such as bar diagram, histogram, frequency polygon, stem and leaf plot and box-cox plot. The graphical representation can give us a lot of information about data, for example, whether it is symmetric or skewed, and also if there is any outlier or influential observation present in the data. Some of these techniques are described briefly in the following paragraphs.
SERREA: A Semantic Management System for Retail Real Estate Agencies

[www.igi-global.com/article/serrea-semantic-management-system-retail/43562?camid=4v1a](www.igi-global.com/article/serrea-semantic-management-system-retail/43562?camid=4v1a)