Data Mining

Sherry Y. Chen
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

There is an explosion in the amount of data that organizations generate, collect, and store. Organizations are gradually relying more on new technologies to access, analyze, summarize, and interpret information intelligently. Data mining, therefore, has become a research area with increased importance (Amaratunga & Cabrera, 2004). Data mining is the search for valuable information in large volumes of data (Hand, Mannila, & Smyth, 2001). It can discover hidden relationships, patterns, and interdependencies and generate rules to predict the correlations, which can help the organizations make critical decisions faster or with a greater degree of confidence (Gargano & Ragged, 1999).

There is a wide range of data mining techniques, which has been successfully used in many applications. This article is an attempt to provide an overview of existing data mining applications. The article begins by explaining the key tasks that data mining can achieve. It then moves to discuss applications domains that data mining can support. The article identifies three common application domains, including bioinformatics, electronic commerce, and search engines. For each domain, how data mining can enhance the functions will be described. Subsequently, the limitations of current research will be addressed, followed by a discussion of directions for future research.

BACKGROUND

Data mining can be used to achieve many types of tasks. Based on the kinds of knowledge to be discovered, it can be broadly divided into supervised learning and unsupervised learning. The former requires the data to be pre-classified. Each item is associated with a unique label, signifying the class in which the item belongs. In contrast, the latter does not require pre-classification of the data and can form groups that share common characteristics (Nolan, 2002). To achieve these two main tasks, four data mining approaches are commonly used: classification, clustering, association rules, and visualization.

Classification

Classification, which is a process of supervised learning, is an important issue in data mining. It refers to discovering predictive patterns where a predicted attribute is nominal or categorical. The predicted attribute is called the class. Subsequently, a data item is assigned to one of the predefined sets of classes by examining its attributes (Changchien & Lu, 2001). One example of classification applications is to analyze the functions of genes on the basis of predefined classes that biologists set (see the section on "Classifying Gene Functions").

Clustering

Clustering is also known as *exploratory data analysis* (EDA) (Tukey, 1977). This approach is used in those situations where a training set of pre-classified records is unavailable. Objects are divided into groups based on their similarity. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups (Roussinov & Zhao, 2003). From a data mining perspective, clustering is an approach for unsupervised learning. One of the major applications of clustering is the management of customers' relationships, which is described in the section "Customer Management."

Association Rules

Association rules that were first proposed by Agrawal and Srikant (1994) are mainly used to find out the meaningful relationships between items or features that occur synchronously in databases (Wang, Chuang, Hsu, & Keh, 2004). This approach is useful when one has an idea of different associations that are being sought out. This is because one can find all kinds of correlations in a large data set. It has been widely applied to extract knowledge from Web log data (Lee, Kim, Chung, & Kwon, 2002). In particular, it is very popular among marketing managers and retailers in electronic commerce who want to find associative patterns among products (see the section on "Market Basket Analysis").

Visualization

The visualization approach to data mining is based on an assumption that human beings are very good at perceiving structure in visual forms. The basic idea is to present the data in some visual form, allowing the human to gain insight from the data, draw conclusions, and directly interact with the data (Ankerst, 2001). Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary (Keim, 2002). This approach is especially useful when little is known about the data and the exploration goals are vague. One example of using visualization is author co-citation analysis (see the section on "Author Co-citation Analyses").

DATA MINING APPLICATIONS

As previously discussed, data mining can be used to achieve various types of tasks, such as classification, clustering, association rules, and visualization. These tasks have been implemented in many application domains. The main application domains that data mining can support include bioinformatics, electronic commerce, and search engines.

Bioinformatics

In the past decade, we have been overwhelmed with increasing floods of data gathered by the Human and other Genome Projects. Consequently, a major challenge in bioinformatics is extracting useful information from these data. To face this challenge, it is necessary to develop an advanced computational approach for data analysis. Data mining provides such potentials. Three application areas, which are commonly presented in the literature, are described next.

Clustering Microarray Data

Unsupervised learning produces clustering algorithms, which are being applied to DNA microarray data sets. Many algorithms are available for clustering, such as k-means and hierarchical clustering. These algorithms have different strengths and weaknesses, so they may cause the lack of inter-method consistency in assigning related gene-expression profiles to clusters. To overcome this problem, Swift et al. (2004) proposed a consensus strategy to produce both robust and consensus clustering of gene-expression data and assign statistical significance to these clusters from known gene functions. Robust clustering is compiling the results of different clustering methods reporting only the co-clustered instances grouped together by different algorithms—that is, with maximum agreement across clustering methods. On the other hand, consensus clustering relaxes the full agreement requirement by taking a parameter, "minimum agreement," which allows different agreement thresholds to be explored. It is reported that this approach can improve confidence in gene-expression analysis.

Classifying Gene Functions

Biologists often know a subset of genes involved in a biological pathway of interest and wish to discover other genes that can be assigned to the same pathway (Ng & Tan, 2003). Unlike clustering, which processes genes based on their similarity, classification can learn to classify new genes based on predefined classes, taking advantage of the domain knowledge already possessed by the biologists. Therefore, the classification approach seems more suitable than clustering for the classification of gene functions.

Earlier studies focus on classification of gene functions based on a single source of data, for example, Kuramochi and Karypis (2001). Recently, heterogeneous sources of data have been adopted. For example, Deng, Geng, and Ali (2005) presented a hybrid weighted naive Bayesian network model for the prediction of functional classes of novel genes based on multiple sources of data, such as DNA sequences, expressions, gene structures, database annotations, and homologies. This model can also be used to analyze the contribution of each source of data toward the gene function prediction performance.

Identifying Phenotype Data

In the aforementioned two approaches, the genes are treated as objects, while the samples are the attributes. Conversely, the samples can be considered as the objects and the genes as the attributes. In this approach, the samples can be partitioned into homogeneous groups. Each group may correspond to some particular phenotypes (Golub et al., 1999). Phenotype is observable and physical characteristics of an organism. Over the past decade, growing interest has surfaced in recognizing relationships between the genotypes and phenotypes. Tracing a phenotype over time may provide a longitudinal record for the evolution of a disease and the response to a therapeutic intervention. This approach is analogous to removing components from a machine and then attempting to operate the machine under different conditions to diagnose the role of the missing component. Function of genes can be determined by removing the gene and observing the resulting effect on the organism's phenotype.

Electronic Commerce

The widespread use of the Web has tremendous impact on the way organizations interact with their partners and customers. Many organizations consider analyzing customers'