Data Mining Tools:
Association Rules

Sanjiv K. Bhatia
University of Missouri, USA

Jitender S. Deogun
University of Nebraska, USA

INTRODUCTION

The current times have witnessed an exponential increase in the popularity of the Internet as well as advanced data collection tools across a wide variety of application domains. This has led to an explosive growth of data accumulation from terabytes to petabytes at a dramatic pace, and is already trending towards exabytes. The data flows from various sources including business (Web, e-commerce, transactions, stocks, and business intelligence), science (remote sensing, bioinformatics, Large Hedron Collider, and scientific simulations), and society (news, blogs, forums, and social networks). The heterogeneous nature of such voluminous data dictates the requirement to extract and reveal the summary of knowledge that can be used in tasks such as decision making, event prediction, and pattern extraction. More recently, data mining techniques have also been applied to predict vulnerabilities of Web applications, involving scripts running on multiple sites, for computer security (Shar & Tan, 2012; Shar & Tan, 2012).

BACKGROUND

The current generation of hardware and database technologies provides efficient and inexpensive ways to store and access large amounts of data. This data in its raw form may not be directly useful. However, the data can be made useful by extracting knowledge contained therein. For example, large databases of sales in a consumer goods company can be processed to extract the correlation between increase in sales for an item in relation to national holidays or in response to a certain marketing campaign. Such extracted knowledge is definitely more useful for the business in contrast to the large amount of data in raw form. A corporation can implement new marketing strategies and campaigns to maximize the sales during a certain time. Such large databases may be dormant but potentially useful resources, and can yield substantial benefits with intelligent processing and analysis.

The processing and analysis of large amounts of raw data by manual methods is a slow and expensive process, in addition to being highly dependent on a domain. The high number of records and attributes makes this processing an almost impractical task in terms of its computational complexity. The complexity of the task leads to the search for new theories that can be applied along with the cutting edge tools to extract useful information and knowledge from a large pool of data. In the rest of the chapter, we'll describe the concepts and tools to facilitate such knowledge extraction.

MAIN FOCUS OF THE CHAPTER

This chapter will focus on different data mining tools, such as knowledge discovery in databases, association rules, and probability logic.

DOI: 10.4018/978-1-4666-5202-6.ch059

Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) refers to the process of searching for knowledge in large data sets. The notion of *knowledge* on a data set is defined through a specification of measures and thresholds over many observed patterns. Knowledge is extracted through data mining methods, and may be of interest to researchers in a variety of domains such as machine learning, pattern recognition, intelligent databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. KDD involves the evaluation and possible interpretation of patterns to decide on what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projection of the data prior to the data mining step. KDD has been aptly described as "*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The KDD process starts with learning about the relevant prior knowledge regarding an application. This involves understanding the domain of interest, the required prior knowledge for the application, and the expected goal of the model being developed.

The learning process is followed by preparation of the target data set. The target data set is characterized by a sample data set as well as a careful study of the relevant variables in the data set. It provides an overview of the variables used in the KDD process.

At this point, the data set may require some preprocessing to prepare it for further discovery. The preprocessing step involves removal of noise and outliers in data, and fixing any missing data. It includes generalizing the attribute values in specific ranges depending on domain and expected goals. For example, consider the data set to be the crime reports collected over years that contains absolute time value of the occurrence of crime. We can probably perform better correlation of data in these reports if the sampling of time is changed to time intervals such as Late Night [midnight-4am], Early Morning (4-8am) Morning (8am-noon), Afternoon (noon-4pm), Evening (4-8pm), and Night (8pm-midnight).

The preprocessed data is then reduced in dimensions and projected onto useful features. This step involves the representation of data in terms of features that are useful towards the final goal(s). The dimensionality reduction may also involve transformation techniques to minimize the number of variables under consideration. This can be illustrated by invariant representation of time line series data (Udechukwu, Barker, & Alhajj, 2004) where the authors provide domain-independent techniques to encode trends in time line series data using an angle. The angle is normalized in the range of [-90°, +90°] between the time line axis and the line representing a measure of relative change in two consecutive values in time-line. Other techniques for dimensionality reduction have involved the use of embedding data in the form of a graph (Yan, et al., 2007).

After reducing the dimensionality of data and enclosing the trends in time-line series data, we finalize the goal of the KDD process, such as regression or clustering. Then, we have to choose a suitable data mining algorithm. The last step involves the selection of an appropriate technique to find patterns in the data set. We need to decide upon the feasibility of parameters and models to be used. At the same time, all these tasks must be performed keeping in view the overall criterion of the KDD process. Some authors have also investigated the application of multiple iterations of *K*-means algorithm and random projection to cluster high-dimensional data (Cardoso & Wichert, 2012).

We apply data mining techniques to discover knowledge for tasks such as classification rules, association rules, dependency, and the degree of interest. We interpret the rules and analyze them for patterns while removing the redundant and irrelevant rules. We can then translate the extracted rules and patterns in an understandable manner. Finally, we apply the discovered knowledge into