Data Preparation for Data Mining

Magdi Kamel
Naval Postgraduate School, USA

INTRODUCTION

Practical experience of data mining has revealed that preparing data is the most time-consuming phase of any data mining project. Estimates of the amount of time and resources spent on data preparation vary from at least 60% to upward of 80% (SPSS, 2002a). In spite of this fact, not enough attention is given to this important task, thus perpetuating the idea that the core of the data mining effort is the modeling process rather than all phases of the data mining life cycle. This article presents an overview of the most important issues and considerations for preparing data for data mining.

BACKGROUND

The explosive growth of government, business, and scientific databases has overwhelmed the traditional, manual approaches to data analysis and created a need for a new generation of techniques and tools for intelligent and automated knowledge discovery in data. The field of knowledge discovery, better known as data mining, is an emerging and rapidly evolving field that draws from other established disciplines such as databases, applied statistics, visualization, artificial intelligence and pattern recognition that specifically focus on fulfilling this need. The goal of data mining is to develop techniques for identifying novel and potentially useful patterns in large data sets (Tan, Steinbach, & Kumar, 2006).

Rather than being a single activity, data mining is as an interactive and iterative process that consists of a number of activities for discovering useful knowledge (Han & Kamber, 2006). These include data selection, data reorganization, data exploration, data cleaning, and data transformation. Additional activities are required to ensure that useful knowledge is derived from the data after data-mining algorithms are applied such as the proper interpretation of the results of data mining.

MAIN FOCUS

In this article, we address the issue of data preparation—how to make the data more suitable for data mining. Data preparation is a broad area and consists of a number of different approaches and techniques that are interrelated in complex ways. For the purpose of this article we consider data preparation to include the tasks of data selection, data reorganization, data exploration, data cleaning, and data transformation. These tasks are discussed in detail in subsequent sections.

It is important to note that the existence of a well designed and constructed data warehouse, a special database that contains data from multiple sources that are cleaned, merged, and reorganized for reporting and data analysis, may make the step of data preparation faster and less problematic. However, the existence of a data warehouse is not necessary for successful data mining. If the data required for data mining already exist, or can be easily created, then the existence of a data warehouse is immaterial.

Data Selection

Data selection refers to the task of extracting smaller data sets from larger databases or data files through a process known as sampling. Sampling is mainly used to reduce overall file sizes for training and validating data mining models (Gaohua & Liu, 2000). Sampling should be done so that the resulting smaller dataset is representative of the original, large file. An exception of this rule is when modeling infrequent events such as fraud or rare medical conditions. In this case oversampling is used to boost the cases from the rare categories for the training set (Weiss, 2004). However, the validation set must approximate the population distribution of the target variable.

There are many methods for sampling data. Regardless to the type of sampling it is important to construct a *probability sample*, one in which each record of the data has a known probability of being included in the sample. The importance of using probability samples is that it allows us to calculate a level of error on the statistics calculated from the sample data when using statistical inferential techniques.

The two main types of sampling include simple random sampling and stratified random sampling. In *simple random sampling*, each record in the original data file has an equal probability of being selected. In *stratified random sampling*, records are selected such that they are proportional to the segments of population they represent. A similar type of sampling is sampling over time. In this type, samples are selected so as to adequately represent all time periods of interest.

Another aspect of data selection is the selection of a subset of variables, an approach knows as dimensionality reduction (Liu & Motoda, 1998). With a smaller number of variables many data mining algorithms work better, and the resulting models are more understandable.

As data is being prepared, it is usually difficult to determine which variables are likely to be important for data mining. A good understanding of the business and the goals of data mining should provide at this stage a general idea on what variables might be important and therefore should be included in the analysis.

Data Reorganization

Data reorganization is used to change the case basis of a data set. Data can be reorganized through summarization or other complex file manipulation. Summarization replaces the current dataset with one containing summary information based on one or more subgroups. For example, a bank transactions dataset can be summarized by customer, account type, or branch.

Complex file manipulation tasks can include one or many types of operations, such as matching, appending, aggregating, and distributing data from one variable into several other new variables (Bock & Diday, 2000).

It is important to note that data needs to be explored, cleaned and checked before it is summarized. The reason for doing so is that failure to eliminate errors at a lower level of summarization will make it impossible to find at a higher level of summarization.

Data Exploration

Data exploration is the preliminary investigation of the data in order to better understand its characteristics. There is nothing particularly distinctive about data exploration for data mining than that used for Exploratory Data Analysis (EDA), which was created by John Tukey in the 1970's (1977). It usually involves the following tasks:

1. Examining the distribution of each variable, summary statistics, and the amount and type of missing data. This can be accomplished by creating frequencies table and/or appropriate graphs such as bar charts and histograms.
2. Studying the relationship between two variables of interest using techniques such as crosstabulations, correlations, and On-Line Analytical Processing (OLAP), as well as graphs such as clustered bar charts, scatterplots, and web graphs.
3. Determining the extent and patterns of missing data.

We briefly examine three major areas of data exploration: summary statistics, data visualization, and OLAP.

Summary Statistics

Summary statistics capture many characteristics of large set of values with a single number or a small set of numbers (Devore, 2003). For categorical data, summary statistics include the *frequency* which is the number of times each value occurs in a particular set of data, and the *mode* which is the value that has the highest frequency. For continuous data, the most widely used summary statistics are the mean and median, which are measures of central tendency. The *mean* is the average value of a set of values, while the *median* is the middle value if there are an odd number of values and the average of the two middle values if the number of values is even.

Another set of commonly used summary statistics for continuous data measures the variability or spread of a set of values. These measures include the range, the standard deviation, and the variance. The *range* is the difference between the highest and lowest values in a