IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8926

Chapter VII

The Impact of Missing Data on Data Mining

Marvin L. Brown
Hawaii Pacific University, USA

John F. Kros
East Carolina University, USA

ABSTRACT

Data mining is based upon searching the concatenation of multiple databases that usually contain some amount of missing data along with a variable percentage of inaccurate data, pollution, outliers, and noise. The actual data-mining process deals significantly with prediction, estimation, classification, pattern recognition, and the development of association rules. Therefore, the significance of the analysis depends heavily on the accuracy of the database and on the chosen sample data to be used for model training and testing. The issue of missing data must be addressed since ignoring this problem can introduce bias into the models being evaluated and lead to inaccurate data mining conclusions.

THE IMPACT OF MISSING DATA

Missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection. More historical data is being collected today due to the proliferation of computer software and the high capacity of storage media. In turn, the issue of missing data becomes an even more pervasive dilemma. An added complication is that the more data that is collected, the higher the likelihood of missing data. This will require one to address the problem of missing data in order to be effective.

This chapter appears in the book, *Data Mining: Opportunities and Challenges,* edited by John Wang.

During the last four decades, statisticians have attempted to address the impact of missing data on information technology.

This chapter's objectives are to address the impact of missing data and its impact on data mining. The chapter commences with a background analysis, including a review of both seminal and current literature. Reasons for data inconsistency along with definitions of various types of missing data are addressed. The main thrust of the chapter focuses on methods of addressing missing data and the impact that missing data has on the knowledge discovery process. Finally, trends regarding missing data and data mining are discussed in addition to future research opportunities and concluding remarks.

Background

The analysis of missing data is a comparatively recent discipline. With the advent of the mainframe computer in the 1960s, businesses were capable of collecting large amounts of data on their customer databases. As large amounts of data were collected, the issue of missing data began to appear. A number of works provide perspective on missing data and data mining.

Afifi and Elashoff (1966) provide a review of the literature regarding missing data and data mining. Their paper contains many seminal concepts, however, the work may be dated for today's use. Hartley and Hocking (1971), in their paper entitled "The Analysis of Incomplete Data," presented one of the first discussions on dealing with skewed and categorical data, especially maximum likelihood (ML) algorithms such as those used in Amos. Orchard and Woodbury (1972) provide early reasoning for approaching missing data in data mining by using what is commonly referred to as an expectation maximization (EM) algorithm to produce unbiased estimates when the data are missing at random (MAR). Dempster, Laird, and Rubin's (1977) paper provided another method for obtaining ML estimates and using EM algorithms. The main difference between Dempster, Laird, and Rubin's (1977) EM approach and that of Hartley and Hocking is the Full Information Maximum Likelihood (FIML) algorithm used by Amos. In general, the FIML algorithm employs both first- and second-order derivatives whereas the EM algorithm uses only first-order derivatives.

Little (1982) discussed models for nonresponse, while Little and Rubin (1987) considered statistical analysis with missing data. Specifically, Little and Rubin (1987) defined three unique types of missing data mechanisms and provided parametric methods for handling these types of missing data. These papers sparked numerous works in the area of missing data. Diggle and Kenward (1994) addressed issues regarding data missing completely at random, data missing at random, and likelihood-based inference. Graham, Hofer, Donaldson, MacKinnon, and Schafer (1997) discussed using the EM algorithm to estimate means and covariance matrices from incomplete data. Papers from Little (1995) and Little and Rubin (1989) extended the concept of ML estimation in data mining, but they also tended to concentrate on data that have a few distinct patterns of missing data. Howell (1998) provided a good overview and examples of basic statistical calculations to handle missing data.

The problem of missing data is a complex one. Little and Rubin (1987) and Schafer (1997) provided conventional statistical methods for analyzing missing data and discussed the negative implications of naïve imputation methods. However, the statistical