Imprecise Data and the Data Mining Process

Marvin L. Brown
Grambling State University, USA

John F. Kros
East Carolina University, USA

INTRODUCTION

Missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection. The management of missing data in organizations has recently been addressed as more firms implement large-scale enterprise resource planning systems (see Vosburg & Kumar, 2001; Xu et al., 2002). The issue of missing data becomes an even more pervasive dilemma in the knowledge discovery process, in that as more data is collected, the higher the likelihood of missing data becomes.

The objective of this research is to discuss imprecise data and the data mining process. The article begins with a background analysis, including a brief review of both seminal and current literature. The main thrust of the chapter focuses on reasons for data inconsistency along with definitions of various types of missing data. Future trends followed by concluding remarks complete the chapter.

BACKGROUND

The analysis of missing data is a comparatively recent discipline. However, the literature holds a number of works that provide perspective on missing data and data mining. Afifi and Elashoff (1966) provide an early seminal paper reviewing the missing data and data mining literature. Little and Rubin's (1987) milestone work 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. Lee and Siau (2001) present an excellent review of data mining techniques within the knowledge discovery process. The references in this section are given as suggested reading for any analyst beginning their research in the area of data mining and missing data.

MAIN THRUST

The article focuses on the reasons for data inconsistency and the types of missing data. In addition, trends regarding missing data and data mining are discussed along with future research opportunities and concluding remarks.

REASONS FOR DATA INCONSISTENCY

Data inconsistency may arise for a number of reasons, including:

- Procedural Factors
- Refusal of Response
- Inapplicable Responses

These three reasons tend to cover the largest areas of missing data in the data mining process.

Procedural Factors

Data entry errors are common and their impact on the knowledge discovery process and data mining can generate serious problems. Inaccurate classifications, erroneous estimates, predictions, and invalid pattern recognition may also take place. In situations where databases are being refreshed with new data, blank responses from questionnaires further complicate the data mining process. If a large number of similar respondents fail to complete similar questions, the deletion or misclassification of these observations can take the researcher down the wrong path of investigation or lead to inaccurate decision-making by end users.

Refusal of Response

Some respondents may find certain survey questions offensive or they may be personally sensitive to certain

questions. For example, some respondents may have no opinion regarding certain questions such as political or religious affiliation. In addition, questions that refer to one's education level, income, age or weight may be deemed too private for some respondents to answer.

Furthermore, respondents may simply have insufficient knowledge to accurately answer particular questions. Students or inexperienced individuals may have insufficient knowledge to answer certain questions (such as salaries in various regions of the country, retirement options, insurance choices, etc).

Inapplicable Responses

Sometimes questions are left blank simply because the questions apply to a more general population rather than to an individual respondent. If a subset of questions on a questionnaire does not apply to the individual respondent, data may be missing for a particular expected group within a data set. For example, adults who have never been married or who are widowed or divorced are likely to not answer a question regarding years of marriage.

TYPES OF MISSING DATA

The following is a list of the standard types of missing data:

- Data Missing at Random
- Data Missing Completely at Random
- Non-Ignorable Missing Data
- Outliers Treated as Missing Data

It is important for an analyst to understand the different types of missing data before they can address the issue. Each type of missing data is defined next.

[Data] Missing At Random (MAR)

Rubin (1978), in a seminal missing data research paper, defined missing data as MAR "when given the variables X and Y, the probability of response depends on X but not on Y." Cases containing incomplete data must be treated differently than cases with complete data. For example, if the likelihood that a respondent will provide his or her weight depends on the probability that the respondent will not provide his or her age, then the missing data is considered to be Missing At Random (MAR) (Kim, 2001).

[Data] Missing Completely At Random (MCAR)

Kim (2001), based on an earlier work, classified data as MCAR when "the probability of response [shows that] independence exists between X and Y." MCAR data exhibits a higher level of randomness than does MAR. In other words, the observed values of Y are truly a random sample for all values of Y, and no other factors included in the study may bias the observed values of Y.

Consider the case of a laboratory providing the results of a chemical compound decomposition test in which a significant level of iron is being sought. If certain levels of iron are met or missing entirely and no other elements in the compound are identified to correlate then it can be determined that the identified or missing data for iron is MCAR.

Non-Ignorable Missing Data

In contrast to the MAR situation where data missingness is explained by other measured variables in a study; non-ignorable missing data arise due to the data missingness pattern being explainable — and only explainable — by the very variable(s) on which the data are missing.

For example, given two variables, X and Y, data is deemed Non-Ignorable when the probability of response depends on variable X and possibly on variable Y. For example, if the likelihood of an individual providing his or her weight varied within various age categories, the missing data is non-ignorable (Kim, 2001). Thus, the pattern of missing data is non-random and possibly predictable from other variables in the database.

In practice, the MCAR assumption is seldom met. Most missing data methods are applied upon the assumption of MAR. And in correspondence to Kim (2001), "Non-Ignorable missing data is the hardest condition to deal with, but unfortunately, the most likely to occur as well."

Outliers Treated As Missing Data

Many times it is necessary to classify these outliers as missing data. Pre-testing and calculating threshold boundaries are necessary in the pre-processing of data in order to identify those values which are to be classified as missing. Data whose values fall outside of predefined ranges may skew test results. Consider the case of a laboratory providing the results of a chemical compound decomposition test. If it has been predetermined that the maximum amount of iron that can be