Data Mining with Incomplete Data

Hai Wang
Saint Mary's University, Canada

Shouhong Wang
University of Massachusetts Dartmouth, USA

INTRODUCTION

Survey is one of the common data acquisition methods for data mining (Brin, Rastogi & Shim, 2003). In data mining one can rarely find a survey data set that contains complete entries of each observation for all of the variables. Commonly, surveys and questionnaires are often only partially completed by respondents. The possible reasons for incomplete data could be numerous, including negligence, deliberate avoidance for privacy, ambiguity of the survey question, and aversion. The extent of damage of missing data is unknown when it is virtually impossible to return the survey or questionnaires to the data source for completion, but is one of the most important parts of knowledge for data mining to discover. In fact, missing data is an important debatable issue in the knowledge engineering field (Tseng, Wang, & Lee, 2003).

In mining a survey database with incomplete data, patterns of the missing data as well as the potential impacts of these missing data on the mining results constitute valuable knowledge. For instance, a data miner often wishes to know how reliable a data mining result is, if only the complete data entries are used; when and why certain types of values are often missing; what variables are correlated in terms of having missing values at the same time; what reason for incomplete data is likely, etc. These valuable pieces of knowledge can be discovered only after the missing part of the data set is fully explored.

BACKGROUND

There have been three traditional approaches to handling missing data in statistical analysis and data mining. One of the convenient solutions to incomplete data is to eliminate from the data set those records that have missing values (Little & Rubin, 2002). This, however, ignores potentially useful information in those records. In cases where the proportion of missing data is large, the data mining conclusions drawn from the screened data set are more likely misleading.

Another simple approach of dealing with missing data is to use generic "unknown" for all missing data items. However, this approach does not provide much information that might be useful for interpretation of missing data.

The third solution to dealing with missing data is to estimate the missing value in the data item. In the case of time series data, interpolation based on two adjacent data points that are observed is possible. In general cases, one may use some expected value in the data item based on statistical measures (Dempster, Laird, & Rubin, 1997). However, data in data mining are commonly of the types of ranking, category, multiple choices, and binary. Interpolation and use of an expected value for a particular missing data variable in these cases are generally inadequate. More importantly, a meaningful treatment of missing data shall always be independent of the problem being investigated (Batista & Monard, 2003).

More recently, there have been mathematical methods for finding the salient correlation structure, or aggregate conceptual directions, of a data set with missing data (Aggarwal & Parthasarathy, 2001; Parthasarathy & Aggarwal, 2003). These methods make themselves distinct from the traditional approaches of treating missing data by focusing on the collective effects of the missing data instead of individual missing values. However, these statistical models are data-driven, instead of problem-domain-driven. In fact, a particular data mining task is often related to its specific problem domain, and a single generic conceptual construction algorithm is insufficient to handle a variety of data mining tasks.

MAIN THRUST

There have been two primary approaches of data mining with incomplete data: conceptual construction and enhanced data mining.

Conceptual Construction with Incomplete Data

Conceptual construction with incomplete data reveals the patterns of the missing data as well as the potential

impacts of these missing data on the mining results based only on the complete data. Conceptual construction on incomplete data is a knowledge development process. To construct new concepts on incomplete data, the data miner needs to identify a particular problem as a base for the construction. According to (Wang, S. & Wang, H., 2004), conceptual construction is carried out through two phases. First, data mining techniques (e.g., cluster analysis) are applied to the data set with complete data to reveal the unsuspected patterns of the data, and the problem is then articulated by the data miner. Second, the incomplete data with missing values related to the problem are used to construct new concepts. In this phase, the data miner evaluates the impacts of missing data on the identification of the problem and develops knowledge related to the problem. For example, suppose a data miner is investigating the profile of the consumers who are interested in a particular product. Using the complete data, the data miner has found that variable *i* (e.g., income) is an important factor of the consumers' purchasing behavior. To further verify and improve the data mining result, the data miner must develop new knowledge through mining the incomplete data. Four typical concepts as results of knowledge discovery in data mining with incomplete data are described as follows:

(1) **Reliability:** The reliability concept reveals the scope of the missing data in terms of the problem identified based only on complete data. For instance, in the above example, to develop the reliability concept, the data miner can define index $V_M(i)/V_C(i)$ where $V_M(i)$ is the number of missing values in variable *i*, and V_C(i) is the number of samples used for the problem identification in variable *i*. Accordingly, the higher $V_M(i)/V_C(i)$ is, the lower the reliability of the factor would be.

(2) **Hiding:** The concept of hiding reveals how likely an observation with a certain range of values in one variable is to have a missing value in another variable. For instance, in the above example, the data miner can define index $V_M(i)|x(j) \in$ (a,b) where $V_M(i)$ is the number of missing values in variable *i*, x(*j*) is the occurrence of variable *j* (e.g., education years), and (a,b) is the range of x(*j*); and use this index to disclose the hiding relationships between variables *i* and *j*, say, more than two thousand records have missing values in variable income given the value of education years ranging from 13 to 19.

(3) **Complementing:** The concept of complementing reveals what variables are more likely to have missing values at the same time; that is, the correlation of missing values related to the problem being investigated. For instance, in the above example, the data miner can define index $V_M(i,j)/V_M(i)$ where $V_M(i,j)$ is the number of missing values in both variables *i* and *j*, and $V_M(i)$ is the number of missing values in variable *i*. This concept discloses the correlation of two variables in terms of missing values. The higher the value $V_M(i,j)/V_M(i)$ is, the stronger the correlation of missing values would be.

(4) **Conditional Effects:** The concept of conditional effects reveals the potential changes to the understanding of the problem caused by the missing values. To develop the concept of conditional effects, the data miner assumes different possible values for the missing values, and then observe the possible changes of the nature of the problem. For instance, in the above example, the data miner can define index $\Delta P|\forall z(i)=k$ where ΔP is the change of the size of the target consumer group perceived by the data miner, $\forall z(i)$ represents all missing values of variable *i*, and k is the possible value variable *i* might have for the survey. Typically, k={*max*, *min*, *p*} where *max* is the maximal value of the scale, *min* is the minimal value of the scale, and *p* is the random variable with the same distribution function of the values in the complete data. By setting different possible values of k for the missing values, the data miner is able to observe the change of the size of the consumer group and redefine the problem.

Enhanced Data Mining with Incomplete Data

The second primary approach to data mining with incomplete data is enhanced data mining, in which incomplete data are fully utilized. Enhanced data mining is carried out through two phases. In the first phase, observations with missing data are transformed into fuzzy observations. Since missing values make the observation fuzzy, according to fuzzy set theory (Zadeh, 1978), an observation with missing values can be transformed into fuzzy patterns that are equivalent to the observation. For instance, suppose there is an observation $\boldsymbol{A}=\mathbf{X}(x_1, x_2, \ldots x_c \ldots x_m)$ where x_c is the variable with missing value, and $x_c \in \{ r_1, r_2 \ldots r_p \}$ where r_j (j=1, 2, ... p) is the possible occurrence of x_c. Let $\mu_j = P_j(x_c = r_j)$, the fuzzy membership (or possibility) that x_c belongs to r_j (j=1, 2, ...p), and $\sum_j \mu_j = 1$. Then, $\mu_j [\mathbf{X} |(x_c = r_j)]$ (j=1, 2, ... p) are fuzzy patterns that are the equivalence to the observation $\boldsymbol{A}$.

In the second phase of enhanced data mining, all fuzzy patterns, along with the complete data, are used for data mining using tools such as self-organizing maps (SOM) (Deboeck & Kohonen, 1998; Kohonen, 1989; Vesanto & Alhoniemi, 2000) and other types of neural