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
Related Content

Data Warehousing for Association Mining
www.igi-global.com/chapter/data-warehousing-association-mining/10881?camid=4v1a

Data Mining and the Text Categorization Framework
www.igi-global.com/chapter/data-mining-text-categorization-framework/10850?camid=4v1a

Active Learning with Multiple Views
www.igi-global.com/chapter/active-learning-multiple-views/10790?camid=4v1a

Data Warehouse Back-End Tools
www.igi-global.com/chapter/data-warehouse-back-end-tools/10878?camid=4v1a