The Issue of Missing Values in Data Mining

Malcolm J. Beynon
Cardiff University, UK

INTRODUCTION

The essence of data mining is to investigate for pertinent information that may exist in data (often large data sets). The immeasurably large amount of data present in the world, due to the increasing capacity of storage media, manifests the issue of the presence of missing values (Olinsky *et al*., 2003; Brown and Kros, 2003). The presented encyclopaedia article considers the general issue of the presence of missing values when data mining, and demonstrates the effect of when managing their presence is or is not undertaken, through the utilisation of a data mining technique.

The issue of missing values was first exposited over forty years ago in Afifi and Elashoff (1966). Since then it is continually the focus of study and explanation (El-Masri and Fox-Wasylyshyn, 2005), covering issues such as the nature of their presence and management (Allison, 2000). With this in mind, the naïve consistent aspect of the missing value debate is the limited general strategies available for their management, the main two being either the simple deletion of cases with missing data or a form of imputation of the missing values in someway (see Elliott and Hawthorne, 2005). Examples of the specific investigation of missing data (and data quality), include in; data warehousing (Ma *et al*., 2000), and customer relationship management (Berry and Linoff, 2000).

An alternative strategy considered is the retention of the missing values, and their subsequent 'ignorance' contribution in any data mining undertaken on the associated original incomplete data set. A consequence of this retention is that full interpretability can be placed on the results found from the original incomplete data set. This strategy can be followed when using the nascent CaRBS technique for object classification (Beynon, 2005a, 2005b). CaRBS analyses are presented here to illustrate that data mining can manage the presence of missing values in a much more effective manner than the more inhibitory traditional strategies. An example data set is considered, with a noticeable level of missing values present in the original data set. A critical increase in the number of missing values present in the data set further illustrates the benefit from 'intelligent' data mining (in this case using CaRBS).

BACKGROUND

Underlying the necessity to concern oneself with the issue of missing values is the reality that most data analysis techniques were not designed for their presence (Schafer and Graham, 2002). It follows, an external level of management of the missing values is necessary. There is however underlying caution on the ad-hoc manner in which the management of missing values may be undertaken, this lack of thought is well expressed by Huang and Zhu (2002, p. 1613):

Inappropriate treatment of missing data may cause large errors or false results.

A recent article by Brown and Kros (2003) looked at the impact of missing values on data mining algorithms, including; *k*-nearest neighbour, decision trees, association rules and neural networks. For these considered techniques, the presence of missing values is considered to have an impact, with a level external management necessary to accommodate them. Indeed, perhaps the attitude is that it is the norm to have to manage the missing values, with little thought to the consequences of doing this. Conversely, there is also the possibilities that missing values differ in important ways from those that are present.

While common, the specific notion of the management of missing values is not so clear, since firstly it is often necessary to understand the reasons for their presence (De Leeuw, 2001), and subsequently how these reasons may dictate how they should be future described. For example, in the case of large survey data, whether the missing data is (*ibid*.); Missing by design, Inapplicable item, Cognitive task too difficult, Refuse to respond, Don't know and Inadequate score. Whether the data is survey based or from another source, a typi-

cal solution is to make simplifying assumptions about the mechanism that causes the missing data (Ramoni and Sebastiani, 2001). These mechanisms (causes) are consistently classified into three categories, based around the distributions of their presence, namely;

- *Missing Completely at Random (MCAR)*: The fact that an entry is missing is independent of both observed and unobserved values in the data set (e.g. equipment failure).
- *Missing at Random (MAR)*: The fact that an entry is missing is a function of the observed values in the data set (e.g. respondents are excused filling in part of a questionnaire).
- *Missing Not at Random (MNAR)*: An entry will be missing depends on both observed and unobserved values in the data set (e.g. personal demographics of a respondent contribute to the incompleteness of a questionnaire).

Confusion has surrounded the use of these terms (Regoeczi and Riedel, 2003). Further, the type of missing data influences the available methods to manage their presence (Elliott and Hawthorne, 2005). Two most popular approaches to their management are case deletion and imputation (next discussed), in the cases of the missing values being MCAR or MAR then imputation can produce a more complete data set that is not adversely biased (*ibid*.).

The case deletion approach infers cases in a data set are discarded if their required information is incomplete. This, by its nature incurs the loss of information from discarding partially informative case (Shen and Lai, 2001), De Leeuw (2001) describes the resultant loss of information, less efficient estimates and statistical tests. Serious biases may be introduced when missing values are not randomly distributed (Huang and Zhu, 2002). A further problem occurs if there is a small sample so deletion of too many cases may reduce the statistical significance of conclusions. Also associated with this approach is re-weighting, whereby the remaining complete cases can be weighted so that their distribution more closely resembles that of the full sample (Schafer and Graham, 2002; Huang and Zhu, 2002).

Imputation infers an incomplete data set becomes filled-in by the replacement of missing values with surrogates (Olinsky et al., 2003). It is potentially more efficient than case deletion, because no cases are sacrificed, retaining the full sample helps to prevent loss of power resulting from a diminished sample size (Schafer and Graham, 2002). De Leeuw (2001) identifies the availability of modern and user-friendly software encourages the use of imputation, with approaches that include; hot deck imputation, cold deck imputation, multiple imputation, regression and stochastic regression imputation.

Beyond these approaches the commonest is mean imputation, whereby the missing value for a given attribute in a case is filled in with the mean of all the reported values for that attribute (Elliott and Hawthorne, 2005). One factor often highlighted with the utilisation of mean imputation is that the distribution characteristics (including variance) of the completed data set may be underestimated (El-Masri and Fox-Wasylyshyn, 2005). The accompanying danger here is that this approach lulls the user into the plausible state of believing that the data are complete after all. The possible effectiveness of mean imputation, and other approaches, depends of course on the level of incompleteness inherent, with many studies testing results with differing percentages of missing data present (ibid.).

MAIN THRUST

To demonstrate the effect of the management of missing values in data mining, the nascent CaRBS technique is employed (for a detailed elucidation see Beynon, 2005a, 2005b). The aim of the CaRBS technique is to construct a body of evidence (BOE) for each characteristic value that purports levels of exact belief (mass values – $m_{j,i}(\cdot)$) towards the classification of an object to a given hypothesis ($m_{j,i}(\{x\})$), its complement ($m_{j,i}(\{\neg x\})$) and concomitant ignorance ($m_{j,i}(\{x, \neg x\})$). More formally, the mass values come from:

$$m_{j,i}(\{x\}) = \frac{B_i}{1 - A_i} cf_i(v) - \frac{A_i B_i}{1 - A_i},$$

$$m_{j,i}(\{\neg x\}) = \frac{-B_i}{1 - A_i} cf_i(v) + B_i,$$

and $m_{j,i}(\{x, \neg x\}) = 1 - m_{j,i}(\{x\}) - m_{j,i}(\{\neg x\})$, where $cf_i(v)$ is a confidence value (usually sigmoid function). A graphical representation of this process is given in Figure 1.