Chapter 1
A Framework to Detect Disguised Missing Data

Rahime Belen
Informatics Institute, METU, Turkey

Tuğba Taşkaya Temizel
Informatics Institute, METU, Turkey

ABSTRACT

Many manually populated very large databases suffer from data quality problems such as missing, inaccurate data and duplicate entries. A recently recognized data quality problem is that of disguised missing data which arises when an explicit code for missing data such as NA (Not Available) is not provided and a legitimate data value is used instead. Presence of these values may affect the outcome of data mining tasks severely such that association mining algorithms or clustering techniques may result in biased inaccurate association rules and invalid clusters respectively. Detection and elimination of these values are necessary but burdensome to be carried out manually. In this chapter, the methods to detect disguised missing values by visual inspection are explained first. Then, the authors describe the methods used to detect these values automatically. Finally, the framework to detect disguised missing data is proposed and a demonstration of the framework on spatial and categorical data sets is provided.

INTRODUCTION

Information management has become challenging with the ever-increasing data volumes. This data deluge has made the data miners and decision makers more enthusiastic than ever about discovering hidden and precious information by applying sophisticated data mining algorithms. However, once they realize that the data quality is poor, these databases often turn out to be data tombs that are rarely or no longer used.

Data quality ensures the completeness, timeliness, accuracy, validity and consistency of data. The systems having high-quality data are usually systems that implement and follow a data quality management plan in a timely fashion. Data quality problems arise when some systems lack of a plan or when for some, a plan is carried out during the design and implementation phases but neglected afterwards. Data quality also suffers in

DOI: 10.4018/978-1-60960-067-9.ch001

systems that change or evolve in time with a data quality management plan that does not take into consideration the new constraints (Hipp, Guntzer, & Grimmer,2001). As Geiger (2004) states, "The viability of the business decisions is contingent on good data and good data is contingent on effective approach to data quality management". Data quality is a multidimensional, complex and morphing concept (Dasu, 2003). In the last decade, it has become a popular issue in the areas of database statistics, workflow management, and knowledge engineering.

Poor data quality is pervasive. It makes it difficult to understand the data in relation to the nature of the phenomena in databases and make appropriate decisions concerning the customers. As a result, the customer satisfaction may be affected. Implementing data warehouses with poor data quality levels is, at best, very risky. Despite of all these risks, a proper data quality management plan can be a unique source of competitive advantage (Redman, 1997).

A well-known data quality problem is that of explicitly missing data that is indicated by using special codes such as "NaN" or "0" which arises when data is not provided or unknown. There are many algorithms to deal with this problem in the literature. On the other hand, missing values can appear as valid values that disguise themselves within the true values. Since they are not explicitly represented, disguise values have less chances of becoming detectable and may easily become a part of an analysis which may lead to biased and inaccurate results. Therefore, disguised missing data impair the data quality surreptitiously. For an example, consider a case where users are asked to select their "gender" in a form where the default selected value in the select box is "female". If the users do not want to reveal their gender information, they may skip the question. Consequently, the default value is recorded incorrectly for male users who have skipped the question. Another example is a website requiring registration in which users tend to leave the default values as they are or select the first entries in the select box lists. Fields like date of birth or place of birth can be given as examples that are frequently left out and cause disguised missing values to emerge. In such datasets, many people are recorded as if they were born in 'Alabama' (first state in the list of U.S.) or on January 1 (the first value in the pop-up lists of month and day, respectively), which is formally valid but factually incorrect.

A well-known example of disguised missing data is that of Pima Indian diabetes dataset from UCI Machine Learning Example (Pima Indians Diabetes Data Set, 2009). Its metadata file indicates that there are no missing data values. However, Berault (2001) points out that five of seven attributes exhibit biological implausible *zero* values, suggesting that this metadata is incorrect and many analyses were conducted without taking into consideration these values in which some constitute 48% of the data set. When they inspected the data set, they realized that these values disrupted the mean and standard deviation of the distribution of variables, in some cases severely. For example while the mean of serum insulin concentration values was 79.8 mu U/ml on raw data, the mean increased to 155.55 after the removal of disguise values.

The disguise values bring about serious drawbacks in data analysis and mining tasks. For example a hypothesis test was conducted to measure whether there was any difference on diastolic blood pressures between diabetic and nondiabetic patients (Pearson, 2006). The tests showed significantly different results when performed on raw and cleansed data sets even though the discrepancy between the two sets only constituted 5% of them. So hypothesis results can be affected severely even with the existence of small amount of disguise values. Data mining results can also be affected negatively. For example, in k-means clustering algorithm, the disguise values can produce centroids that are shifted towards the disguise values thus form inaccurate clusters. In the presence of disguised missing data, some simple