Handling Missing Values in Data Mining

Submitted By:
Bhavik Doshi
Department of Computer Science
Rochester Institute of Technology
Rochester, New York 14623-5603, USA
Email: bkd4833@rit.edu
Abstract

Missing Values and its problems are very common in the data cleaning process. Several methods have been proposed so as to process missing data in datasets and avoid problems caused by it. This paper discusses various problems caused by missing values and different ways in which one can deal with them. Missing data is a familiar and unavoidable problem in large datasets and is widely discussed in the field of data mining and statistics. Sometimes program environments may provide code for missing data but they lack standardization and are rarely used. Thus analyzing the impact of problems caused by missing values and finding solutions to tackle with them is an important issue in the field of Data Cleaning and Preparation. Many solutions have been presented regarding this issue and handling missing values is still a topic which is being worked upon. In this paper we discuss various hitches we face when it comes to missing data and see how they can be resolved.

1. Introduction

Anyone who does statistical data analysis or data cleaning of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes. For example in surveys people generally tend to leave the field of income blank or sometimes people have no information available and cannot answer the question. Also in the process of collecting data from multiple sources some data may be inadvertently lost. For all these and many other reasons, missing data is a universal problem in both social and health sciences. This is because every standard statistical method works on the fact that every problem has information on all the variables an it needs to be analyzed. The most common and simple solution to this
problem is if any case has missing data for any of the attribute to be analyzed we can simply ignore it. This will give us a dataset which will not contain any missing value and we can then use any standard methods to process it further. But this method has a major drawback which is deleting missing values sometimes might lead to ignoring a large section of the original sample. This paper first illustrates different types of missing values and analyzes their consequences on datasets. After that we study two approaches taken by researchers to identify missing data from datasets in different scenarios.

This paper reviews some of the problems caused by missing values and how we can tackle them. Section 2 describes different types of missing data while section 3 describes the consequences of missing values in monotonous data sets. In section 4 we discuss the impact of disguised missing values and discuss a heuristic approach to identify and eliminate them. Section 5 consists of the future work going on in handling missing values followed by conclusion in section 6.

2. Different Types Missing Data

The problem of missing data resides in almost all the surveys and designed experiments. As stated before one of the common method is to ignore cases of missing values. Ignoring cases of missing values may sometimes lead to elimination of a major portion of the dataset thus leading into inappropriate results. Also the use of default values, results into disguised missing values which will be discussed further in section 4. The different types of missing mechanisms are stated as below:
MCAR
The term “Missing Completely at Random” refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset [1]. Here the data are collected and observed arbitrarily and the collected data does not depend on any other variable of the dataset. Such type of missing data is very rarely found and the best method is to ignore such cases.

MAR
Sometimes data might not be missing at random but may be termed as “Missing at Random”. We can consider an entry $X_i$ as missing at random if the data meets the requirement that missingness should not depend on the value of $X_i$ after controlling for another variable. As an example, depressed people tend to have less income and thus the reported income now depends on the variable depression. As depressed people have lower income the percentage of missing data among depressed individuals will be high.

NAMR
If the data is not missing at random or informatively missing then it is termed as “Not missing at Random”. Such a situation occurs when the missingness mechanism depends on the actual value of missing data [1]. Modeling such a condition is a very difficult task to achieve. When we have a data with NMAR problem the only way to attain an estimate of parameters is to model the missingness. This means we need to write a model for missing data and then integrate it into a more complex model for estimating missing values. As mentioned earlier this is easier said than done.
3. Dealing with Missing Values in Monotone Datasets

3.1 Monotone Datasets
Missing values come in the process of knowledge discovery not only by human mistakes and omissions of data but also when data for certain variables is hard, costly or even impractical to obtain. A dataset with the set of attributes $A$, and a labeling $\lambda$ is called monotone if the value of each attribute are ordered and for each two data points $x, y$ such that $x \leq y$ (y dominates x on all attributes in $A$) it is true that $\lambda(x) \leq \lambda(y)$ [2]. Monotone troubles appear in various domain areas like credit rating, bankruptcy prediction, bond rating etc. It can thus be stated that using a monotone classifier not only maximizes returns but also helps in motivating the decision in front of internal or external parties. Algorithms designed to study monotonous datasets cannot handle non-monotonous data. Sometimes additional expense is required for handling non-monotonous data. Thus in order to achieve simplicity and maximum benefit it is advisable to work on fully monotone datasets. This brings into picture the importance of filling missing values so that we can achieve fully monotone datasets and thus eliminate the burden of handling non-monotone data. In [2] the authors of the paper propose a simple and easy preprocessing method which can further be used supplement to several other approaches for filling in missing values so that the monotonicity property of the resulting data is maintained.

3.2 Missing values in Monotone Datasets
The monotonicity property of classification has many features like it provides information about attributes arriving from ordered domains. Also it states that a monotone function of independent attributes symbolizes a target variable. We discussed the importance of handling missing values in the previous section. In [2] the authors of the paper assume that the missing values are present
only in conditional attributes. If the value of a decision attribute is missing that we cannot get enough information from the object whereas if the value of a condition attribute is missing we still can retrieve enough information from the remaining attributes along with the decision attribute. Thus ignoring objects with missing values is not a suitable approach as it might lead to wrong results. The authors in [2] propose an extension of preprocessing methods which makes sure that the final dataset is monotonic. The algorithm computes the possible values of the interval taking into consideration only fully defined objects using the formulas stated in [1]. If the calculated interval contains only one value then we assign the object with the missing value. Otherwise, we either ignore the value or fill in the value depending on the conditions. The author states many other approaches to fill in missing values in the paper. The above algorithm fills the missing values and gives the output as a monotone dataset.

In the case of noisy data with some monotone inconsistencies, the above algorithm can still be applied but it might not necessary result into a monotone dataset. But we might decrease the monotone inconsistencies by discarding objects where empty values are calculated. This will improve the degree of monotonicity of the dataset. In [2] the authors conduct two experiments to validate and prove their methods but were successful to some extent only. They suggest more extensive experiments so as to predict the accuracy of the monotone classifier. Thus filling missing values in the context of monotone datasets can be done using the suggested algorithm complimented by some preprocessing methods.
4. Dealing with Disguised Missing Data

The problem of disguised missing data arises when missing data values are not explicitly represented as such, but are coded with values that can be misinterpreted as valid data [3]. The paper explains how missing, incomplete and incorrect metadata can lead to disguised missing data. Consequences of disguised missing data can be severe and hence detecting and recognizing missing data is a very important part in the data cleaning process. As defined disguised missing data are a set of disguised missing entries which can be considered as valid data by mistake. In the data collection process many times values provided by users lead to disguised missing data. Fake values are recorded in the table if a user knowingly or unknowingly provides an incorrect value or sometimes just does not provide any value.

4.1 Sources of Disguised Missing Data

There are many possible ways which can lead to fake or disguised values being recorded in the dataset. The most obvious but uncommon possibility is someone deliberately providing or entering false values in the dataset. Alternatively default values can become a source of disguised missing data. As an example, consider an online form having the default sex as male and the default country as United States of America. A customer filling the form may not want to disclose his/her personal information and hence it might lead to missing values disguising themselves as default values. Such data entry errors accompanied by rigid edit checks form the sources of forged data. The lack of standard code to enter data into tables opens the door for factually incorrect data into the dataset. The ultimate source of most disguised missing data is probably the lack of a standard missing data representation [3]. Sometimes even within a single
data file there might be multiple codes representing the same missing data. Each individual or organization has their own way of representing data and this facilitates the rise of disguised missing data. Developing a standard way to represent and handle missing values will only lead to reduction fake or false values entering into the dataset.

4.2 Discovering the Disguise
Assuming the presence of disguised missing data in the datasets the most important question comes in one’s mind is how to detect them. If the data is adequately disguised in the dataset then sometimes even domain knowledge or best of the methods known cannot detect them. But the approach is to identify abnormal values or patterns in the datasets with the help of domain knowledge or other methods and try to distinguish real from disguised data. The basic step is to identify suspicious values in the datasets which may look factual but are actually fake or false data. With the background knowledge a preliminary analysis of data can be done thus coming up with the range of values for each attribute. Domain knowledge might also prove useful in the above process. Once we have the range of attributes we can examine the data to find suspicious values and thus detect disguised values. Alternatively partial domain knowledge can also prove useful in exposing disguised missing data. For example, even if we do not have any knowledge of lower or upper bounds of data we can still come to a conclusion for variables like age that they can never be negative. Detecting outliers can sometimes help in uncovering disguised missing data but not always. If the values selected to encode missing data are sufficiently far outside the range of the nominal data to appear as outliners, we can apply standard techniques to look for disguised missing data [3].
4.3 Approaches to clean disguised missing data.
One approach to clean such data is to have some domain knowledge. A domain expert can screen entries with suspicious values, such as blood pressure of a patient being 0 [4]. Considering outliers as disguised missing data is another way but might not be feasible if the volume of disguised missing data is large. Again having some domain knowledge may help in identifying disguised missing data. For example, if the number of males exceeds the number of females in the dataset and we know that the population of males and females is nearly equal in the dataset, then we can come to the conclusion that some male values in the dataset may be disguised missing [4]. The above described methods heavily depend on domain knowledge which is not always available. Or sometimes if missing values disguise themselves as inliers then domain knowledge may also not be useful in detecting them [3]. The authors in [4] propose a framework for identifying suspicious frequently used disguised values. Additionally the paper defines an Embedded Unbiased Sample Heuristic approach to discover missing values. The framework is divided into two phases, namely the mining phase and the post processing phase. In the mining phase each attribute is analyzed and checked based on the heuristic approach. The first phase outputs some probable disguised missing values which can be confirmed in the post processing phase with the help of domain knowledge or other data cleaning methods. Thus identifying disguised missing data and then eliminating for the dataset constitute a very important step in the process of data cleaning and preparation.
5. Future Scope

Presence of missing values in the dataset can sometime be really annoying. It becomes important to identify them and eliminate their presence so as to get the right results. The authors in [2] propose a simple preprocessing technique to eliminate missing values from monotonous datasets and also conducted few experiments on the same. Extensive experiments are needed so as to check the correctness of the algorithm by taking a wide variety of datasets and then analyzing the output. This will help in identifying any faults in the existing algorithm and can also help in predicting the accuracy of the classifier. In the paper [4] the authors propose a heuristic approach to identify missing values and then develop an approach to eliminate them. These experiments were performed on both real datasets and synthetic dataset and the algorithm developed was successful in cleaning the datasets with missing values. Also the authors intend to run the algorithm for large amount of datasets so as to test it in real world applications and ensure its correctness. Wide amount of research is going on in handling missing values in large datasets. In [5] the authors of the paper propose new measures for item sets and association rules to be used in incomplete datasets and then propose a novel algorithm.

6. Conclusion

Data cleaning and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss two approaches to deal with missing data in different scenarios. This paper addresses the issues of handling missing values in datasets and methods in which missing values can be tackled. We first discuss the different types of missing
data and analyze their impact on the dataset. We now look into the problem of missing values in monotonous datasets. We suggest a simple preprocessing method which when used with other techniques help in eliminating missing values and help in maintaining the dataset monotonous. The authors in paper [2] conduct simple experiments to test the algorithm and find that taking the most frequent value and replacing it in place of missing values give better results. Missing data sometimes also disguise themselves as valid data and are difficult to identify. We therefore propose a heuristic approach to tackle a practical and challenging issue of cleaning disguised missing data. With the help of this approach we identify suspicious sample of data and then develop an unbiased sample heuristic approach to discover missing values.

References:


