Journal of
Computer Science & Systems Biology

Waqas et al., J Comput Sci Syst Biol 2016, 9:2
DOI: 10.4172/jcsb.1000221

Review Article Open Access

Treatment of Missing Values in Data Mining

Waqas I*, Syed Saeed-Ur-Rahman S, Imran MJ and Rehan A

Superior University, Lahore, Pakistan

Abstract

Data mining has pushed the realm of information technology beyond predictable limits. Data mining has left its permanent marks on decision making in just in few years of its inception. Missing value is one of the major factor, which can render the obtain result beyond use attained from specific data set by applying data mining technique. There could be numerous reasons for missing values in a data set such as human error, hardware malfunction etc. It is imperative to tackle the labyrinth of missing values before applying any technique of data mining; otherwise, the information extracted from data set containing missing values will lead to the path of wrong decision making. There are several techniques available to control the issue of missing values such as replacing the missing value with: (a) closest value, (b) mean value and (c) median value etc. Some algorithms are also used to deal with the problem of missing values such as *k*-nearest neighbour. Paper reviewed certain techniques and algorithms to deal with the puzzle of missing values whereby achieving pure data set (i.e., data set without missing value) which in-turn will lead to path of correct and accurate decision making.

Keywords: Data mining; Missing value; Treatment of missing value

Introduction

Missing data is one of the issues which are to be fathomed for real-time application. Improper imputation produces predisposition result. For example, manual information entrances system, inaccurate estimations, gear blunders, and numerous others. Hence legitimate consideration is expected to credit the missing values. Missing Data will not make any impact on the result if its percentage is less 1%, if missing data's range within the range of 1-5% then it is somehow manageable; however in case of 5-15% complex techniques are used for handling the problems of missing data but if it exceeds from 15% then it will surely hinder the result achieved after applying data mining techniques [1]. Often, missing values appears as "NULL" in databases or it can be represented as empty cells in spreadsheet. Whereas, some flat-files use others symbols as well to indicate the missing values like "?" etc. If missing values are represented from any above depicted symbol then it is somewhat easier to identity and elucidates them but missing data can also be appeared as outliers or wrong data. These wrong data must be removed before performing analysis to get the result according to expectation [2].

Ordinary and modern strategies are there for handling this issue. The variables may be: (a) missing completely at random, (b) missing at random and (c) missing not at random. Each variable should be managed freely. Imputation techniques help to credit the missing values. Pre-processing must be carried out before imputing the values by imputation techniques. *K*-NN algorithm is utilized to gathering the data set into distinctive gatherings. After grouping of data, missing information in every gathering is imputed by mean, median and standard deviation. The results are compared in distinctive rate of precision [3].

Categories of Missing Data

Statistician categorized missing data into three categories as: (a) Missing at Random (MAR) (b) Missing Completely at Random (MCAR) and (c) Missing Not at Random (MNAR) [4].

According to Rubin, MAR is to be a condition in which the probability that data are missing depends only on the observed data but on the missing data, after controlling for observed data [5,6]. Missing completely at random (MCAR) is the prospect of a record possessing a missing value of the attribute but it does not depends on the missing data or the observed data [6]. Not missing at Random (NMAR) is the probability of a record containing missing value of field that depends on the value of attire [2].

Missing Values Problem

According to Luengo J, three issues are interlinked in the domain of missing values: (a) decrease in efficiency, (b) hurdles in analyzing and managing the data (c) differences between missing and complete data that impact on the result [2]. Missing Values issue must be tackled by attributing values by utilizing effective imputation method [7]. Loss of effectiveness is brought about by tedious procedure of managing missing values. It is somehow coupled with the second issue difficulties in managing and exploring information from data. Difficulties in tackling and data dissecting lie truth told that many techniques or calculations are normally not able for tackling missing values and issue of missing values should be comprehended preceding investigations amid data readiness stage. The other issue inclination coming about because of contrasts in between of missing values and non-missing values data lies indeed that attributed values are not the similar as estimated finished dataset [2].

Methods of Imputation

The process of replacing attributed values from the available data is known as Imputation. There are some imputation techniques which lie in the regime of: (a) regulated and (b) unsupervised. However, List-wise and Pairwise deletion erases entire line [3].

***Corresponding author:** Waqas I, Superior University, Lahore, Pakistan, E-mail: waqasilyas12@gmail.com

Received February 06, 2016; **Accepted** February 29, 2016; **Published** March 03, 2016

Citation: Waqas I, Syed Saeed-Ur-Rahman S, Imran MJ, Rehan A (2016) Treatment of Missing Values in Data Mining. J Comput Sci Syst Biol 9: 051-053. doi:10.4172/jcsb.1000221

Standard deviation

It calculates the feast of information to the mean worth. Standard deviation is helpful while looking at dataset having the similar mean yet an alternate extent. It is the square base of the change which makes it easier to interpret. It is the most often utilized measure of scattering [8].

Mean substitution

Mean imputation technique is a standout amongst the most commonly used strategies. Mean substitution replaces missing values on a variable with the mean estimation of the observed values. The imputed missing values are dependent upon one variable– the between subjects mean for that variable based on accessible data. Mean substitution jelly the mean of a variables distribution; nonetheless, mean substitution regularly twists different attributes of a variables circulation. Utilizing constants to swap missing data will change the feature for the original dataset; overlooking the relationship among properties will predisposition the accompanying data mining algorithm. A variation of this technique is to swap the missing value of a characteristic with mean of acknowledged estimations of those values where chance of missing data has a place [9].

$$\frac{\sum X}{N}$$

Median substitution

Mean or median substitution of covariates and result variables is still regularly utilized. The median is essentially as same as the mean. It is determined through gathering of data and calculating average. Median imputation brings about the average of the whole data set being the same as it would be with case deletion, yet the variability between individual's reactions is diminished, biasing differences and covariance's toward zero. Since the mean is influenced by the outliers appears common to utilize the median rather to just guarantee power. In this scenario, the absent information of a characteristic is replaced through the median of acknowledged values [2].

Median can be calculated by the given formula.

$$\textbf{Mean} = \frac{25+28+45+29+49+33+11+20}{12}$$

Where L is the length and n is the total number of items.

Regression imputation

This is one of the expansive strategies for imputation missing values. There are different regression techniques.

The linear regression: The linear regression which utilized for numeric variables and logistic regression is utilized for categorical data. The linear regression is chip away at linear capacity in view of likelihood and the logistic regression is deal with logistic function in view of likelihood yet it has just two possible outcomes for likelihood [10].

The linear regression can be calculated by the formula as follow:

$$Z=c+dY$$

Where Y is the explanatory variable and Z is dependent variable, d represents the slope of line and c represent the intercept.

The random regression: The random regression imputation which discover missing values for any variable in view of conditional distribution. It imputes the value in view of contingent dispersion of Y given X. It is more successful for numeric data [10].

Closest fit

The closet fit algorithm depends upon exchanging absent values with present value of the similar attribute of other likewise cases. Main notion is to find out from dataset likewise scenarios and select the likewise case to the case in discussion with missing attribute values [2].

K-nearest neighbour algorithm

It is a strategy for gathering cases in view of likeness from different cases. In the prospect of machine learning, it had created as an approach to perceive samples without obliging a precise match with any stockpile designs, or scenarios. Comparable scenarios are close to one another and unique cases are inaccessible from one another. Consequently, the difference of two scenarios is a degree of their uniqueness. Likewise scenarios are said to be "neighbours". When another scenario is brought into consideration, its variation from each of the scenario in the model is processed. Orders of the comparable cases–closest neighbours– are matched and the new case is set that possess the best amount of closest neighbours [1].

In this system the principle component is Distance measurements. In 1NN imputation system we can supplant the missing values with the closest neighbour. At the same time if the estimation of K is more than one then supplant the missing value with the mean of K-closest neighbours. Here we talk about the one attribution technique which is said to be regression imputation which we as of now talked about above. The imputation techniques for missing value incorporate parametric regression imputation techniques and nonparametric regression imputation strategies. Non-parametric imputation models are K-closest neighbour or kernel regression.

In which we have no relationship in the middle of depended and autonomous variables. Parametric imputation model is Expectation Maximization (EM). In which we know the piece of relationship in the middle of depended and autonomous variables. Preference of Non-parametric imputation contrast with EM-parametric imputation. More effective for moderate size datasets. In parametric models the data are not fitted because it is less defenceless to slips. At the point when models of the information are not Apriori then this system is more exact [10].

The K- nearest neighbour calculation does not make express models. There are a few progressions for using acknowledged values of K-closest neighbour. One plausibility for numeric attributes is to attribute a mean of the closest neighbours' characteristic values. Weights are contrariwise proportional to the distance from the neighbouring sections [2].

Implementation

At Section Methods of Imputation, we elaborate different techniques to tackle and handle the missing values before applying the Data Mining tricks (at Pre Data Mining Stage). In this section, implementation of two techniques amongst the above mentioned methods is done as under.

Mean substitution

Small Data Set as a sample is selected from huge Database for the treatment of Missing Values in Table 1; that can be used for the censuses of the population.

The missing value of ***Age*** field is represented with red background. In the above small Dataset of 11 Rows, Missing value is presented in 4 rows; which is 36% of the whole dataset. As the ratio of missing value is

too high, so we have to handle the missing value presented in the said data set at pre-data mining phase for the correct pattern generation.

Now, we are going to implement the Mean Substitution technique to handle the Missing Value of Age Field.

$$\text{Mean} = \frac{25+28+45+29+49+33+11+20}{12}$$

$$\text{Mean} = \frac{240}{12}$$

$$\text{Mean} = 20$$

So, we can replace the missing value of Age fields with 20. Table 2 shows the Dataset after treatment of Missing Values.

Closest fit

Now, we are going to elaborate the closest fit technique for the treatment of missing values. We will implement the closest fit technique in data set mentioned in Table 3, which is relating to do the analysis on data regarding property.

In the above Dataset, the tuple having missing value of Rent Field is represented with red background. In closest fit approach, we will replace the missing value with 12000 (from the case with Area of 850). So, the Dataset after applying the closest fit approach is showing in Table 4.

ID	Registration No	Name	Gender	Age
14213	M-382619	Aleeze	Female	25
17629	M-382638	Alan	Female	28
18261	R-2947261	Eleena	Female	45
10372	I-3736262	Dhawan	Male	29
18351	E-3836273	Shakira	Female	
19362	M-286323	Medona	Female	
18362	E-3763637	David	Male	
19463	N-353826	D'couza	Male	49
18362	E-3828273	Diana	Female	33
29362	E-3726182	Joseph	Male	11
19372	Z-387329	Jannifer	Female	
19271	Z-253651	Edge	Male	20

Table 1: Dataset for Mean Substitution with Missing Values.

ID	Registration No	Name	Gender	Age
14213	M-382619	Aleeze	Female	25
17629	M-382638	Alan	Female	28
18261	R-2947261	Eleena	Female	45
10372	I-3736262	Dhawan	Male	29
18351	E-3836273	Shakira	Female	20
19362	M-286323	Medona	Female	20
18362	E-3763637	David	Male	20
19463	N-353826	D'couza	Male	49
18362	E-3828273	Diana	Female	33
29362	E-3726182	Joseph	Male	11
19372	Z-387329	Jannifer	Female	20
19271	Z-253651	Edge	Male	20

Table 2: Dataset for Mean Substitution after treatment of Missing Values.

Area Sq. ft	Rent
275	8000
500	10000
850	12000
900	
1000	17000
1225	19000
1500	20000

Table 3: Dataset for Closest Fit with Missing Values.

Area Sq. ft	Rent
275	8000
500	10000
850	12000
900	12000
1000	17000
1225	19000
1500	20000

The treated missing value (12000) is represented with green background.

Table 4: Dataset for Closest Fit after treatment of Missing Values.

Conclusions and Future Enhancement

This paper gives the complete view about the effective imputation techniques for discovering the missing values from the dataset. *K*-NN algorithm is one of the well-known classifier for gathering up of data. It is likewise examined that when the absent value percentage is high then the strategy is the exactness diminishes. Mean Substitution and Closest Fit techniques to handle Missing Values on small dataset works effectively and efficiently. Whereas, *K*-NN algorithm is better to handle missing value on the large dataset.

It can be further improved via contrasting and some other procedures i.e., MLP and SOM. Mean Substitution can be interchanged with mode, average, standard deviation or by applying Expectation value– Expansion, regression based strategies.

References

1. Acuna E, Rodriguez C (2004) The treatment of missing values and its effect on classifier accuracy. In Classification, Clustering, and Data Mining Applications. David B, Leanna H, Frederick R, McPhipps A, Wolfgang G (eds). Springer Berlin Heidelberg, USA.
2. Kaiser J (2014) Dealing with Missing Values in Data. Journal of Systems Integration 5: 42-51.
3. Malarvizhi MR, Thanamani AS (2013) Comparison of Imputation Techniques after Classifying the Dataset Using K-NN Classifier for the Imputation of Missing Data. International Journal of Computational Engineering Research 3: 101-104.
4. Malarvizhi M, Thanamani A (2012) K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation. IOSR Journal of Computer Engineering 6: 12-15.
5. Rubin DB (1976) Inference and Missing Data. Biometrika 63: 581-592.
6. Dong Y, Peng CJ (2013) Principled missing data methods for researchers. Springer Plus 2: 222.
7. Kanchana S, Antony ST (2014) Classification of Efficient Imputation Method for Analyzing Missing Values. International Journal of Computer Trends and Technology 12: 193-195.
8. Malarvizhi MR, Thanamani DAS (2012) K-nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development 5: 1-5.
9. Liu PLL (2005) A review of missing data treatment methods. Int Journal of Intel Inf Manag Syst Tech 1: 3-10.
10. Suthar BH, Hemant P, Goswami A (2012) A Survey: Classification of Imputation Methods in Data Mining. International Journal of Emerging Technology and Advanced Engineering 2: 309-312.