Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Estimation of Missing Attribute Value in Time Series Database in Data Mining

By Swati Jain & Mrs. Kalpana Jain

College of technology and Engineering,India

Abstract- Missing data is a widely recognized problem affecting large database in data mining. The substitution of mean values for missing data is commonly suggested and used in many statistical software packages, however, mean substitution lead to large errors in correlation matrix and therefore degrading the performance of statistical modeling. The problems arises are biasness of result data base, inefficient data in missing data when anomalous data is also present. In proposed work there is proper handling of missing data values and their analysis with removal of the anomalous data.This method provides more accurate and efficient result and reduces biasness of result for filling in missing data. Theoretical analysis and experimental results shows that proposed methodology is better.

Keywords : *missing data method, imputation, outlier, inference, anova test.*

GJCST-C Classification : *H.2.8,J.4*

Strictly as per the compliance and regulations of:

Estimation of Missing Attribute Value in Time Series Database in Data Mining

Swati Jain [α] & Mrs. Kalpana Jain [σ]

Abstract- Missing data is a widely recognized problem affecting large database in data mining. The substitution of mean values for missing data is commonly suggested and used in many statistical software packages, however, mean substitution lead to large errors in correlation matrix and therefore degrading the performance of statistical modeling. The problems arises are biasness of result data base, inefficient data in missing data when anomalous data is also present. In proposed work there is proper handling of missing data values and their analysis with removal of the anomalous data.This method provides more accurate and efficient result and reduces biasness of result for filling in missing data. Theoretical analysis and experimental results shows that proposed methodology is better.

Keywords: missing data method, imputation, outlier, inference, anova test.

I. Introduction

Data cleaning is a step for discovery of database. Data cleaning, it is also known as data cleansing, it is a phase in which noisy data, anomalous data and irrelevant data are removed from the collection of various data. Missing data are defined as some of the values in the data set which are either lost or not observed or not available due to natural or non natural reasons. Data with missing values confuses both the data analysis and the submission of a solution to fresh data. Thus, three main problems arise when dealing with incomplete data. First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to the irregulaties in data structure and the impossibility of using standard software. Third, and most important, there may be bias due to systematic differences between observed and unobserved data. Deal with missing data is major task for cleaning data. Noor et all [1] In this paper, three types of mean imputation techniques introduced on missing data. Rubin [7] explored about inference and missing data and multiple imputations for non-response in the survey. Allison [8] investigated estimates of linear models with incomplete data and on missing data. Smyth [9] and Zhang [10] have considered that data preparation is a fundamental stage of data analysis. Therefore, this research focuses on anomalous and missing data values. In our research we create a novel method to replace the missing values.

Author α σ: College of technology and Engineering,India.
e-mail: swati.subhi.9@gmail.com

II. Missing Data Methods

There are several methods for treating missing data. Missing data treatment methods can be divided into three categories, as proposed in [7].

a) Ignoring and discarding data

In this method the two main ways to discard data with missing values. The first method is known as list wise deletion. It consists of discarding all instances with missing data. The second method is known as pair wise deletion method. It consists of discarding instances or attributes before deleting any attribute, it is necessary to evaluate its relevance to the analysis.

b) Parameter estimation

In this missing data treatment method, Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data.

c) Imputation

Imputation method is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set assist in estimating the missing values [9].

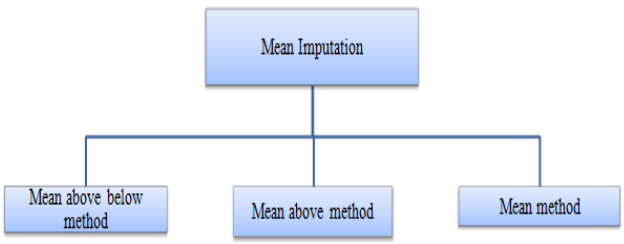

Figure 1: Types of mean imputation method

Mean Imputation Method: In this technique, It consists of replacing the missing data for a given feature by the mean of all known values of that attribute in the class where the instance with missing attribute belongs mean of each attribute that contains missing values is calculated and is replaced in the place of missing values. Each missing value is substituted with calculated mean value which is same for all.

Mean Above Below Method: [1] this method replaces all missing values with the mean of the data above the missing value and one data below the missing value.

Mean Above Method [1]: This method replaces all missing values with the mean of all available data above the missing values.

Mean Method[1]: This method replaces all missing values with the mean of all available data.

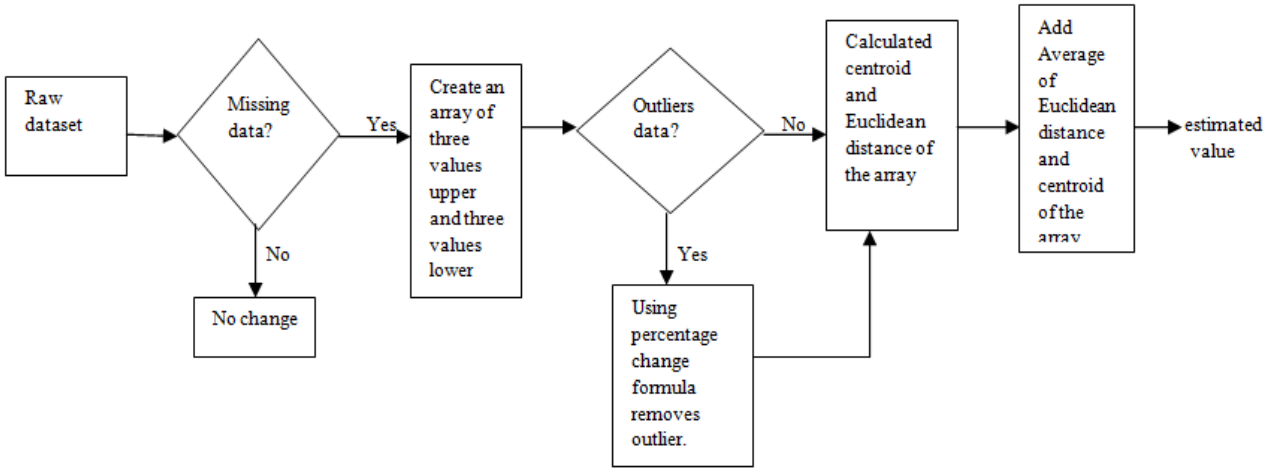

Figure 2: Design flow of proposed methodology

III. Proposed Method for Inference of Missing Attributes Value in Data Mining

The proposed Methodology works in two stages. The first stage is localizing missing data and remove anomalous data, in next stage we substitute the estimated value in the place of missing values by using proposed method. This calculation gives the effective result and decreases the biasness of result.

The working stream of proposed work is shown in figure 2, if there are missing values in the raw data set, then the small subset/array is created from the input data sheet in which missing data value is existing, along with this we work out for anomalous value, according if anomalous value is presented, replace anomalous value with the new calculated value, last step of the work is estimation of missing values using Euclidean distance.

a) Missing value check & outlier detection

This step is preprocessed step of missing data in the input data sheet. The missing values locations are checked in entire data set.

As per the figure 3, the missing value case is pointed by the subscript of the attribute and denoted by the variable x_i after pointing missing value case, we have to record the three upper value(x_{i-1},x_{i-2},x_{i-3}) and three lower value(x_{i+1},x_{i+2},x_{i+3}) from the missing value subscripts. Now the anomalous value in this subset is detected by the percentage change formula. After computing the percentage change of the subset. Now, we find the outlier range, value of outlier range define as per the suitable of the array value. If the anomalous value is detected in the data set, remove that value from the array.

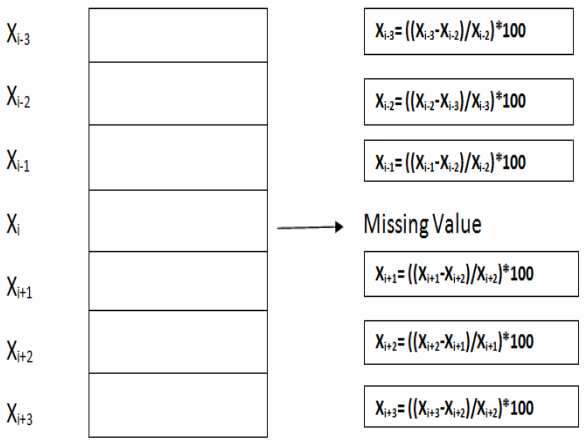

Figure 3: Calculation of percentage change for outlier detection

b) Calculation for missing values

Estimation of missing values in the last phase, when we have the outlier-free data. we process to fill missing values in this array, firstly calculate centroid of the subset ,centroid is generated by the mean of subset. At the further stage Euclidean distance is calculated between centroid of the data and the each value of the data. Euclidean distance is the square root of the sum of squared differences between corresponding elements

of the two vectors. The distance between vectors X and Y is defined as follows:

$$\text{Euclidean distance (d)} = \sqrt{\sum_{i=1}^{n}(\text{Xa} - Ya)\,2}$$

Here,
X_a is centroid of the array
Y_a is particular value of the array

at the last we compute the average of the Euclidean distance and add centroid with average value of distance this is the estimated value of the missing value. The value of X_{est}(estimated value) is separately computed for every missing value in the complete datasheet.

IV. Results and Discussion

Our experiments were carried out for time series datasets taken from Earthpolicy.org site. In proposed work we used different Datasheet like Hydroelectric Generation in India 1965-2013, Average Global Temperature 1880-2014, U.S. Motor Gasoline Consumption 1950-2014, World Wood Production 1961-2011 and few more. Here, we evaluate the U.S. Motor Gasoline Consumption 1950-2014 contains 50 number of instances and two attributes.

Below graph figure 4 shows comparison with respect to mean of all method. The U.S. Motor Gasoline Consumption respectively for the years 1950-2014 for million barrels attribute. The mean consumption of u.s. motor gasoline of million barrels are 2714.the variables are observed and missing values it may be noted that in the planned way 20% values are missing in the random manner for all the variables and in this dataset value of outliers is greater than 5.The mean calculated from incomplete data sets are 2379 this value is slightly lower than the mean values. The proposed methodology applied on the data sets to fill up missing values and the value is 2714.It is observed that the mean values are obtained after replacing missing values by proposed work are close to the actual mean. The results from the proposed method are compares with the techniques like mean above below(MAB), mean above(MA),mean imputation(MI), mean comparison method proposed by Noor et all [1] and analyze shows that proposed method value substitute missing values are more close to the original method with respect to the other method.

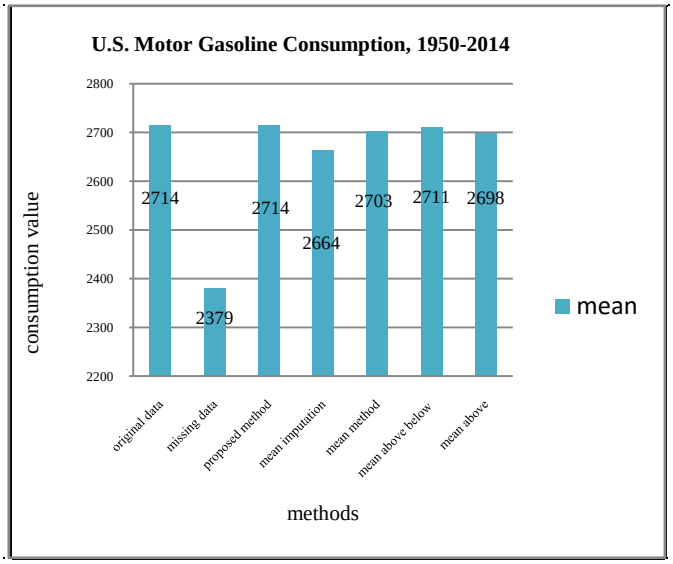

Figure 4: mean value

In figure 5 & figure 6 shows the comparison with respect to standard deviation value and coefficient of variance value of all methods. The proposed method performed significantly better than all other methods.

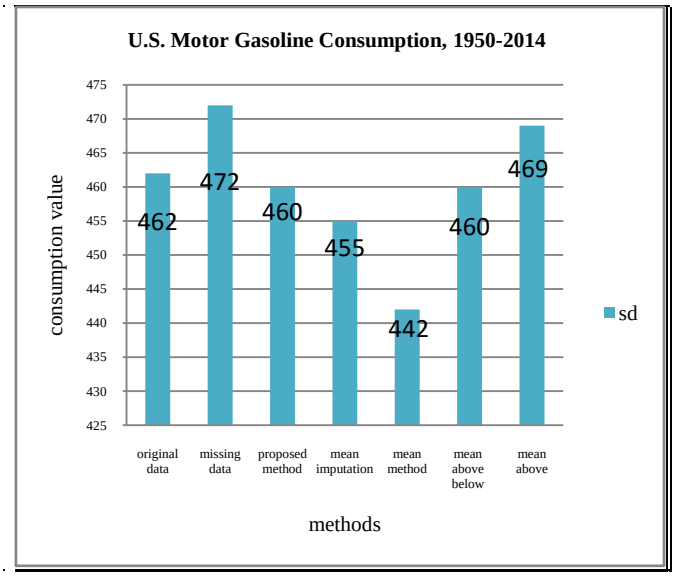

Figure 5: standard deviation value

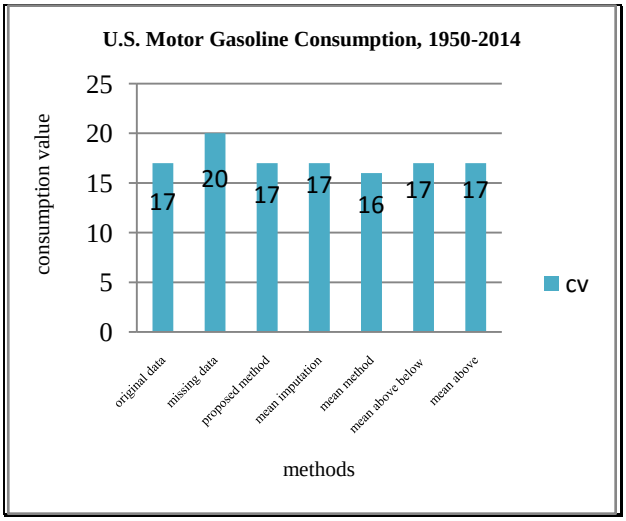

Figure 6: coefficient of variance value

Figure 4: Performance comparison of proposed methodology with MI (mean imputation), mean method, mean above below(MAB) method, mean above(MA) method. (a) Mean value (b) standard deviation value(c) coefficient of variance value

The data sheets are imported in the SPSS and necessary tests for the data validation and significance were applied. On the SPSS software the results are checked by using the ANOVA test for the data sheet and significance value is 99.1% that shows the result is efficient and more compatible with original data.

V. Conclusion

The work focuses on imputing missing values using proposed methodology for numerical attribute in time series data sheet. This method is suitable to handling missing data alone in presence of anomalous data. In this work, performance of proposed method is more reliable as comparing to other mean imputation technique for data analysis in the data mining field.

References Références Referencias

1. Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M. 2014. Mean imputation techniques for filling the missing observations in air pollution dataset. *Key Engineering Materials* **594-599**: 902-908 Trans Tech Publications.
2. Cho, H. Y., Oh, J. H., Kim, K. O., & Shim, J. S. 2013. Outlier detection and missing data filling methods for coastal water temperature data. *Journal of Coastal Research*, **65**:1898-1903.
3. Porter, J. R., Cossman, R. E., & James, W. L. 2009. imputing large group averages for missing data,
4. using rural-urban continuum codes for density driven industry sectors. *Journal of Population Research*, ***26***(3): 273-278.
5. Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology*, **60**: 549-576.
6. Barnett, V., & Lewis, T. 1994. Outliers in Statistical Data ,John Wiley and Sons. *New York*.
7. Genolini, C., & Jacqmin-Gadda, H. 2013. Copy mean: a new method to impute intermittent missing values in longitudinal studies. *Open Journal of Statistics*, **3**: 26-40.
8. Rubin, D. B. 1976. Inference and missing data. *Biometrika*, ***63***(3): 581-592.
9. Allison, P. D. 1987. Estimation of linear models with incomplete data. *Sociological methodology*, 71-103.
10. Smyth, P. 2001. Data mining at the interface of computer science and statistics. In *Data mining for scientific and engineering applications* 35-61. Springer US.
11. Zhang, S., Zhang, C., & Yang, Q. 2003. Data prepara tion for data mining. *Applied Artificial Intelligence*, **17(5-6):** 375-381.
12. Chen, L., Toma-Drane, M., Valois, R. F., & Drane, J. W. 2005. Multiple imputation for missing ordinal data. *Journal of Modern Applied Statistical Methods*, **4(1):** 26.
13. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. 2011. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, ***23***(1): 110-121.
14. Song, Q., & Shepperd, M. 2007. A new imputation method for small software project data sets. *Journal of Systems and Software*, **80(1)**: 51-62.
15. Grzymala-Busse, J. W. 2004. Data with missing attribute values: Generalization of indiscernibility relation and rule induction. *Transactions on Rough Sets* **1***:* 78- 95. Springer Berlin Heidelberg.
16. Han, J., Pei, J., & Kamber, M. 2011. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 225 Wyman Street,Waltham, USA pp. 83-91.