PREDICTION OF MISSING VALUE USING PROPOSED METHOD WITH IMPUTATION METHOD

V.B. Kamble¹ and S.N. Deshmukh²

¹P.E.S. College of Engineering, Aurangabad. (M.S.) India.
²Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. (M.S.) India.

Abstract—Missing values creates different type of problems in the dataset leads to difficult for data analysis. In this research work, student database is taken contains marks of four different subjects in engineering college. Mean, Mode and Median Imputation were used to deal with challenges of incomplete data. Using imputation methods like Mean, Mode and Median Imputation on the dataset convert incomplete data in to complete data and found out to be predicted value by using Proposed Method with Mean Imputation Method in place of missing value. Similarly found out to be predicted value by using Proposed Method with Mode Imputation Method in place of missing value. Also found out to be predicted value by using Proposed Method with Median Imputation Method in place of missing value. Compares Predicted value and original value from the missing dataset at respective feature and found out to be a percentage Accuracy. By comparing the result Proposed Method with Mean Imputation Method shown the best result as compare as to other Imputation Method.

Keywords—Incomplete data, Missing Values, Imputation, Mean Imputation, Median Imputation and Mode Imputation, Percentage Accuracy, Predicted Value etc.

I. INTRODUCTION

Missing data imputation methods can be utilized for to make dataset complete. Missing data imputation techniques replacing missing values of a dataset at respective feature so that data analysis methods can be applied to retrieve the necessary knowledge from the dataset [1].

In this research work, student database is taken contains marks of four different subjects in engineering college. Mean, Mode and Median Imputation were used to deal with challenges of incomplete data. Using imputation methods like Mean, Mode and Median Imputation on the dataset convert incomplete data in to complete data and found out to be predicted value by using Proposed Method with Mean Imputation Method in place of missing value. Similarly found out to be predicted value by using Proposed Method with Mode Imputation Method in place of missing value. Also found out to be predicted value by using Proposed Method with Median Imputation Method in place of missing value.

Compares Predicted value and original value from the missing dataset at respective feature and found out to be a percentage Accuracy. By comparing the result Proposed Method with Mean Imputation Method shown the best result as compare as to other Imputation Method.

The organization of the paper is Section 1: Introduction, Section 2: Literature Reviews, Section 3: Dataset Used, Section 4: Methodology, Section 5: Experimental Result and Analysis, Section 6: Conclusions.

DOI:10.21884/IJMTER.2017.4360.AZS50
II. LITERATURE REVIEWS

2.1 Four Different Methods to Deal with Missing Values

A. Case Deletion (CD) - It is also known as complete case analysis. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and deletes the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with a high degree of missing values for other situations where the sample size is insufficient or some structure exists in the missing data, CD has been shown to produce more biased estimates than alternative methods. CD should be applied only in cases in which data are missing completely at random.

B. Mean Imputation (MI). The method replaces the missing values for a given feature (attribute) by mean of all the known values for that particular attribute by using the following equation. Let us consider that the value of $x_{ij}$ of the $k_{th}$ class, $C_k$, is missing then $x_{ij}$ is calculated as

$$
\hat{x}_{ij} = \frac{\sum_{i} x_{ij}}{n_k}
$$

Where $n_k$ represents the number of non missing values in the $i_{th}$ feature of the $k_{th}$ class.

C. Median Imputation (MDI). This method uses median of all known values of the feature or attribute in the class where the missing value present in the feature by using the following equation. Consider the value $x_{ij}$ of the $k_{th}$ class, $C_k$, is missing. It will be calculated as

$$
x_{ij} = \text{median}\{i:x_{ij}\in C_k\}\{x_{ij}\}
$$

Instead of mean and median, mode also can be in imputation. Imputation method is applied separately for many attribute. However, imputation does not consider co-relation structure of the data [6].

D. Mode Imputation Method (MI)

Mode Imputation Method replacing missing values with the mode of all attributes which having missing value. To reduce the influence of exceptional data, median can also be used. This is one of the most commonly used methods [6].

III. DATASET USED

In this work dataset having characteristics is given below.
Number of Instances: 5000, 10,000, 15,000, 20,000
Number of Attributes: 05
(Record No., M1, ECE, EM, EE)

Dataset contains marks of four different subjects of engineering college. In dataset randomly distributed the missing values in each attribute to become the incomplete dataset. Record. No. in the Dataset is used are imaginary and generated for the data analysis purpose in data mining process. In dataset M1, ECE, EM, EE are the subject in engineering college and class test marks for each subject is out of twenty marks for each subject respectively. The structure of Dataset as shown in the table 1.
### Table No. 1. Dataset

| Record. No. | Subject1 | Subject2 | Subject3 | Subject4 |
|-------------|----------|----------|----------|----------|
| 01          | X1       | X2       | X3       | X4       |
| ..          | ..       | ..       | ..       | ..       |
| N           | XN       | XN       | XN       | XN       |

**IV. METHODOLOGY**

**4.1 Proposed Method with Mean Imputation**

A. This method finds out mean of all corresponding feature or column

B. Multiply mean value of corresponding feature where missing value appears with values of the next feature and divided by mean of the second feature by following equation

$$A = \frac{\text{MEAN}_x}{\text{MEAN}_2}$$  \hspace{1cm} (3)

C. Multiply mean value of corresponding feature where missing value appears with values of the next feature and divided by mean of the third feature by following equation

$$B = \frac{\text{MEAN}_x}{\text{MEAN}_3}$$  \hspace{1cm} (4)

D. Multiply mean value of corresponding feature where missing value appears with values of the next feature and divided by mean of the fourth feature by following equation

$$C = \frac{\text{MEAN}_x}{\text{MEAN}_4}$$  \hspace{1cm} (5)

E. By taking average value of A, B and C from equation no. 3, 4, 5 respectively by using following equation

$$\text{Average} = \frac{A + B + C}{3}$$  \hspace{1cm} (6)

F. After finding the average value place that value at respective missing values in the data set so as to become complete data set for analysis purpose

**4.2 Proposed Method with Mode Imputation**

A. This method finds out mode of all corresponding feature or column

B. Multiply mode value of corresponding feature where missing value appears with values of the next feature and divided by mode of the second feature by following equation

$$A = \frac{\text{MODE}_x}{\text{MODE}_2}$$  \hspace{1cm} (7)

C. Multiply mode value of corresponding feature where missing value appears with values of the next feature and divided by mode of the third feature by following equation

$$B = \frac{\text{MODE}_x}{\text{MODE}_3}$$  \hspace{1cm} (8)

D. Multiply mode value of corresponding feature where missing value appears with values of the next feature and divided by mode of the fourth feature by following equation

$$C = \frac{\text{MODE}_x}{\text{MODE}_4}$$  \hspace{1cm} (9)

E. By taking average value of A, B and C from equation no. 7, 8, 9 respectively by using following equation
F. After finding the average value place that value at respective missing values in the data set so as to become complete data set for analysis purpose

### 4.3 Proposed Method with Median Imputation

A. This method finds out median of all corresponding feature or column

B. Multiply median value of corresponding feature where missing value appears with values of the next feature and divided by median of the second feature by following equation

\[
A = \frac{\text{MEDIAN}_X}{\text{MEDIAN}_2}
\]  
(11) C.

Multiply median value of corresponding feature where missing value appears with values of the next feature and divided by median of the third feature by following equation

\[
B = \frac{\text{MEDIAN}_X}{\text{MEDIAN}_3}
\]  
(12)

D. Multiply median value of corresponding feature where missing value appears with values of the next feature and divided by median of the fourth feature by following equation

\[
C = \frac{\text{MEDIAN}_X}{\text{MEDIAN}_4}
\]  
(13)

E. By taking average value of A, B and C from equation no. 11, 12, 13 respectively by using following equation

\[
\text{Average} = \frac{A + B + C}{3}
\]  
(14) F.

After finding the average value place that value at respective missing values in the data set so as to become complete data set for analysis purpose

### V. EXPERIMENTAL RESULT AND ANALYSIS

5.1 For Experimental Result Student dataset is taken which contains marks of four different subjects of engineering college. Mean, Mode and Median Imputation were used to deal with challenges of incomplete data.

5.2 Using imputation methods like Mean, Mode and Median Imputation on the student dataset convert incomplete databases in to the complete database and found out to be predicted value by using Proposed Method with Mean Imputation Method in place of missing value.

5.3 Similarly found out to be predicted value by using Proposed Method with Mode Imputation Method in place of missing value.

5.4 Also found out to be predicted value by using Proposed Method with Median Imputation Method in place of missing value.

5.5 Predicted value compares with original value from the missing dataset at respective feature, Percentage Accuracy also found out. By comparing the result Proposed Method with Mean Imputation
Method shown the best result as compare to other Imputation Method. The result is shown in the Table No.2

VI. COMPARISON OF PERCENTAGE ACCURACY BY USING PROPOSED METHOD WITH IMPUTATION TECHNIQUE

| Proposed Method with Imputation Technique | Percentage Accuracy |
|------------------------------------------|----------------------|
| Predicted value by Proposed Method with Mean Imputation | Five Thousand | Ten Thousand | Fifteen Thousand | Twenty Thousand |
|                                          | 87.72                | 88.37        | 89.33           | 91.25          |
| Predicted value by Proposed Method with Mode Imputation | 84.00                | 86.33        | 89.34           | 90.80          |
| Predicted value by Proposed Method with Median Imputation | 86.99                | 87.15        | 90.45           | 91.34          |

Figure 1. Graphical Representation of Percentage Accuracy

VII. CONCLUSIONS

In this research paper using proposed method with imputation technique like Mean, Mode and Median Imputation on the student dataset and found out to be predicted value in place of missing value from the dataset. When comparing the result between Predicted value and original value also found out Percentage Accuracy. Experimental result observed that whenever size of the database grows then Percentage Accuracy also increases and proposed Method with Mean Imputation Technique is found to more suitable as compared to the other proposed method with imputation techniques to handle the missing values from the dataset.
REFERENCES

[1] Dinesh J. Prajapati, Jagruti H. Prajapati, “Handling Missing Values: Application to University Data set”. Issue 1, Vol. 1(August-2011), ISSN 2249-6149

[2] Shamsher Singh, Prof. Jagdish Prasad, “Estimation of Missing Values in the Data Mining and comparison of Imputation Methods”. Mathematical Journal of Interdisciplinary Sciences Vol. 1, Issue 1, March 2013, pp. 75–90

[3] Xiao Feng Zhu, Shichao Zhang, Senior Member, IIEEE, Zhi Jin, Senior Member, IIEEE, Zili Zhang, and Zhuming Xu, “Missing Value Estimation for Mixed-Attribute Data Sets”. IEEE Transactions on Knowledge And Data Engineering, Vol. 23, No. 1, January 2011.

[4] T.R.Sivapriya, V. Thavel, A.R.Nadira Banu Kamal, “Imputation and classification of Missing Data Using Least Square Support Vector Machines, A New Approach in Dementia Diagnosis”. International Journal of Advanced Research in Artificial Intelligence, Vol.1, No.4, 2012

[5] Yann-Yann Shieh, “Imputation Methods on General Linear Mixed Models of Longitudinal Studies”, American Institutes for Research

[6] Edgar Acuña, And Caroline Rodriguez, “The Treatment Of Missing Values And Its Effect In The Classifier Accuracy Studies In Classification”, Data Analysis, And Knowledge Organization, 2004, Springer.Com

[7] MS. R. Malarvizhi, Dr. Antony Thanamani, “Comparison of Imputation Techniques after Classifying the Dataset Using Knn Classifier for the Imputation of Missing Data”, International Journal of Computational Engineering Research (IJCER online.com) ISSN 2250-3005, January-2013

[8] Anjana Sharma, Naina Mehta, Iti Sharma, “Reasoning With Missing Values in Multi Attribute Datasets”. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue5, May 2013 ISSN: 2277 128X

[9] Luai Al Shalabi, “A comparative study of techniques to deal with missing data in data set”, In Proceedings of the 4th International Multiconference on Computer Science and Information Technology /CSIT 2006.

[10] A. Pujari, “Data Mining Techniques”, Universities Press, India, 2001.

[11] Ragel, A. and Cremilleux, B., “MVC A preprocessing method to deal with missing values”, In Proceedings of Knowledge Based Systems 1999, 285-291.

[12] Chih-Hung Wu, Chian-Huei Wun, Hung-Ju Chou, “Using Association Rules for Completing Missing Data”, Fourth International Conference on Hybrid Intelligent Systems (HIS’04), 2004 pp.236-241.

[13] Lakshminarayan, K., Harp, S., Goldman, R., and Samad, “Imputation of missing data using machine learning techniques, In Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining. T. 1996

[14] Zhang, S.C., et al., “Information Enhancement for Data Mining”, IEEE Intelligent Systems, 2004, Vol. 19(2): 12-13, (2004).

[15] Qin, Y.S., “Semi-parametric Optimization for Missing Data Imputation”, Applied Intelligence, 2007, 27(1): 79-88.

[16] Zhang, C.Q., ” An Imputation Method for Missing Values”, PAKDD, LNAI, 4426, 2007: 1080-1087.

[17] Han, J., and Kamber, M., “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2002, 2nd edition.

[18] A. Dempster, N.M. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, J. Royal Statistical Soc., vol. 39, pp. 1-38, 1977.

[19] R. Little and D. Rubin, “Statistical Analysis with Missing Data”, second ed. John Wiley and Sons, 2002.

[20] D. Rubin, “Multiple Imputations for Nonresponsive in Surveys”, Wiley, 1987.

[21] J.R. Quinlan, “C4.5: Programs for Machine Learning”. Morgan Kaufmann, 1993.

[22] Q.H. Wang and R. Rao, “Empirical Likelihood-Based Inference under Imputation for Missing Response Data”, Annals of Statistics, vol. 30, pp. 896-924, 2002.

[23] S.C. Zhang, “Par imputation: From Imputation and Null-Imputation to Partially Imputation”, IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.

[24] V.B.Kamble, S.N.Deshmukh, “Comparison of Percentage Error by using Imputation Method On Mid Term Examination Data”, International Journal of Innovations in Engineering Research and Technology (IIJERT),Impact Factor 2.77, Volume 2, Issue 12,2015

[25] V.B.Kamble, S.N.Deshmukh,”Comparative Analysis Of Standard Error Using Imputation Method”, ICITDCEME’15 Conference Proceedings on International Conference on Innovations and Technological Developments in Computer,Electronics and Mechanical Engineering, 28-29, December 2015, VACOE Ahmednagar.ISNN NO.2394-3696

[26] V.B.Kamble, S.N.Deshmukh,” A Novel Hybrid Approach For Prediction Of Missing Values In Numeric Dataset” Global Journal of Engineering Science and Research Management, Impact Factor: 2.265, ISSN 2349-4506