# Classification of Efficient Imputation Method for Analyzing Missing Values

S.Kanchana[#1], Dr. Antony Selvadoss Thanamani[#2],

[#1]*Research Scholar,*[#2]*Professor and Head of the Department, Research Department of Computer Science, NGM College, 90 Palghat Road, Pollachi, Bharathiyar University, Coimbatore, India.*

***Abstract*— In Statistical analysis, missing data is a common problem for data quality. Many real datasets have missing data. Imputation preserves all cases by replacing missing data with a probable value based on other available information. Once all missing values have been imputed, the data set can be analyzed using standard techniques for complete data. This paper aim is to describe the efficient imputation method like Mean, Median, Refined Mean, Standard Deviation, Linear Regression, Discretization based method and some of clustering techniques like K-Mean and KNN methods which are used for imputing missing values in the dataset. The datasets are taken from the UCI ML repository. The results are compared in terms of accuracy.**

***Keywords*— Clustering Techniques, Discretization, K-Mean, KNN, Mean, Median, Refined Mean, Standard Deviation.**

## I. INTRODUCTION

Most of the real world datasets are characterized by an unavoidable problem of incompleteness, in terms of missing values. Missing data are simply observations that we intended to be made. For example, an individual may only respond to certain questions in a survey, or may not respond at all to a particular wave of a longitudinal survey. A variety of different reasons result in introduction of incompleteness in the data. Examples include manual data entry procedures, incorrect measurements, equipment errors, and many others. Existence of errors, and in particular missing values, makes it often difficult to generate useful knowledge from data, since many of data analysis algorithms can work only with complete data. However, data mining algorithms always handle missing data in very simple way. The most traditional missing value imputation techniques are deleting case, mean value imputation, maximum likelihood and other statistical methods. Nowadays research has explored the use of machine learning techniques as method for missing value imputation.

The scope of the paper section wise i.e. Section I: describe the introduction of imputation techniques. Section II: produce the survey experience about imputation. Section III: analysis of missing data mechanism. Section IV: classification of imputation technique available. Section V: Comparison of efficient imputation techniques for analyzing the missing data. Section VI: Experiment and Results. Section VII: Conclusion.

## II. BACKGROUND WORK

This section shares the survey experience about imputation mechanisms as per Little and Rubin [1] [2].Classification of imputation methods and accuracy experimental analysis [3].Types of imputation & missing data Analysis describes [4]. Multiple imputations approach for missing data and the drawback [5].Introduction of missing data and its problem [6]. Comparison of K-Means and KNN clustering algorithm from [7][8][9]. Discretization based method implementation [10] [11].comparison of efficient imputation methods with experimental analysis [12].

## III. MISSING DATA MECHANISMS

### A. *Types of Mechanisms*

Missing data can be divided into Missing Completely at Random (MCAR), missing at Random (MAR), Not Missing at Random (NMAR).

1) *Missing completely at Random (MCAR):* The level of randomness is high in MCAR. There is no any reason based on what the data are missing. If any missing variable M that is not depended on any other variable N. It cannot predict the missing variable M from any other variable in dataset. So the probability of the missing variable is same for all the missing variables.

2) *Missing at Random (MAR):* This is completely different than the MCAR. To predict the value on missing variable M is depend on the other variable N in given dataset. But not the value of missing data itself. Missing values are depending on the value of observed information or values in the dataset.

3) *Not Missing at Random (NMAR):* The missing variables are not random and also cannot predict from other variables in the dataset.

### B. *Missing Data Problem*

Missing data occurs when the person missed to answer for some question in survey or missing may occur in the period of data entry or in medical field some patients missed their regular check up. These problems have to be faced by imputing values by using efficient imputation technique.

### C. *Treatment of Missing Data*

The treatment of missing data proposed by Rubin [1]

*1) Ignoring and discarding data*: These methods determining the missing data on each instance and delete those instances. Another thing is determining each attribute /instances and to remove the whole attribute/instances which having high level of missing data. This method is applicable only when the dataset is MCAR.

*2) Parameter estimation:* This method is used to find the parameters for the complete data. This method is use the Expectation maximization algorithm for handling the parameter estimation of the missing data.

*3) Imputation technique imputation:* This is one kind of procedure in which replaces the missing values based on estimated values.

## IV. CLASSIFICATION OF IMPUTATION METHODS

This section share the survey experience about imputation techniques like Mean, Refined Mean, Median, Standard Deviation, Discretization and the clustering Techniques.

### *A. Mean Value Substitution Method*

. Mean imputation method is one of the most frequently used methods [1]. It consists of replacing the missing data for a given feature or attribute by the mean of all known values of that attribute in the class where the instance with missing attribute belongs.

### *B. Refined Mean Value Substitution Method*

This method starts with mean value substitution [3]. But, by assuming that the initially imputed values are not accurate, this method again re-estimates the new values based on the Euclidean distance of the missing value records and the remaining records. For mean value calculations, the records with minimum Euclidean distance with the missing value record were not taken in to account.

### *C. Median Value Substitution Method*

The median is virtually as same as the mean [2]. Median substitution is calculated by grouping up of data and finding average for the data. It requires the lower class boundary of median class, the size of median class and the frequency of median class.

### *D. Standard Deviation*

The standard deviation measures the spread of the data about the mean value [4]. It is useful in comparing sets of data which may have the same mean but a different range. It is the square root of the variance which makes it easier to interpret. It is the most frequently used measure of dispersion.

### *E. Discretization Method*

A typical Discretization process broadly consists of 4 steps [10]: (1) sorting the continuous values of the feature to be discredited, (2) evaluating a cut-point for splitting or adjacent intervals for merging, (3) according to some criterion, splitting or merging intervals of continuous value, and (4) finally stopping at some point. It reduce the learning complexity and help to understand the dependencies between the attributes and the target class.

### *F. Clustering Techniques*

*1) Imputation with K-Means clustering method*: K-Means is to classify or to group the objects based on attributes/features into k number of group [6] [7]. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It provides fast and accurate way of estimating missing values.

*2) Imputation with K-NN clustering method*: K-Nearest Neighbour is a method for classifying cases based on their similarity to other cases [8]. Similar cases are near each other and dissimilar cases are distant from each other. The distance between two cases is a measure of their dissimilarity. Cases that are near each other are said to be neighbours.

*3) Imputation with K-Medoid clustering method*: It is similar to K- Means [9]. In both the method, the dataset gets partitioned into several datasets by minimizing the distance between points. In k-Medoid, when the cluster number is high, it gives poor result.

## V. COMPARISON OF IMPUTATION TECHNIQUES

TABLE I

| Sr . No | Pros & Cons: Imputation methods | | |
|---|---|---|---|
| | **Method** | **Pros** | **Cons** |
| 1 | Mean Value Substitution | Most frequently used method. Easy to impute missing data. | Imputed values are not accurate because it substitutes missing data with artificially created average data points. |
| 2 | Refined Mean Value Substitution | Start with mean value substitution and again re-estimate the new values based on the Euclidean distance of the missing value records and the remaining records. | It reduces the variability of the characterization of the imputed dataset. It is very time consuming process because the re-estimation of all instance in the dataset. |
| 3 | Median Value Substitution | Easy to use. Give better estimates when compare to mean substitution because it calculate the data by grouping up and finding the average for the data. | Applicable only for small size of data. Takes a long time to calculate for a very large set of data. |
| 4 | Standard Deviation | It gives better result of data than the mean. Easy to compare the data which have the same mean but a different range. | It doesn't give the full range of the data. It can be effect by outliers to give a skewed result. |
| 5 | Discretization Method | It reduces the learning complexity. Help to understand the dependencies between the attributes and | The estimated values are same for all missing values. Not efficient for all types of dataset. |

| | | | |
|---|---|---|---|
| | | the target class. | |
| 6 | K-Means | Easily classify or group the data. Fast and provide accurate estimating missing values. | Difficult to predict K value. It didn't work well with the cluster of global data, different size and density. |
| 7 | KNN | It can predict both quantitative and qualitative data. Easily handle multiple missing values. | It estimates the most similar values. Time consuming process because it searches all instances of similar dataset. |
| 8 | K-Medoid | Similar to K-Means. Easily partitioned into several datasets by minimizing the distance between points. | When the cluster number is high, it gives poor result. It does not provide accurate result when comparing to other cluster algorithms. |

## VI. Experimental Results

Our experiments were carried out Primary Tumor datasets taken from the Machine Learning Database UCI Repository. Dataset contains 339 numbers of instances and 18 numbers of attributes including the class attributes.

The main objective of the experiments conducted in this work is to analyze the classification of efficient imputation methods. Missing values are artificially imputed in different rates in different attributes. Fig.1 shows the experimental evaluations of imputation method.

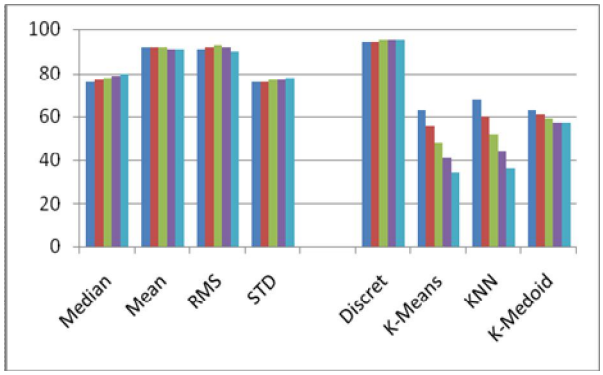


Fig. 1 Comparison result of different imputation method.

The following Table II shows the different imputation method performance in terms of accuracy and the following chart Fig. 2 shows the average performance in terms of accuracy.

TABLE III
Performance in terms of accuracy

| Imputation Methods | Accuracy (%) |
|---|---|
| Mean Substitution | 91.38 |
| Median Substitution | 78 |
| Refined Mean Substitution | 91.73 |
| Standard Deviation | 76.8 |
| Discretization | 94.6 |
| K-Means | 48.4 |
| KNN | 52 |
| K-Medoid | 59.4 |

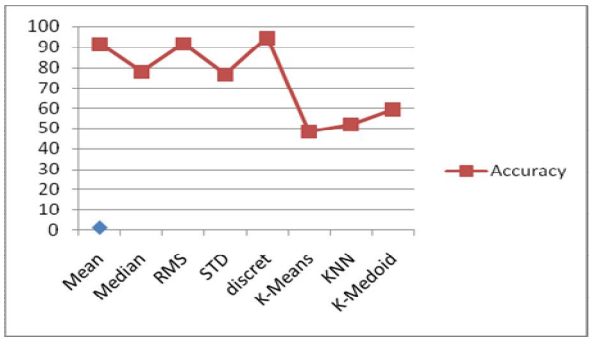


Fig. 2 Percentage of Missing Values vs. Accuracy.

## VII. Conclusions

We trust that this paper gives the complete view about the efficient imputation methods for finding the missing values from the dataset. Also present the advantage and disadvantage of the different imputation methods for analysing the missing value from the dataset in the field of data mining.

## References

[1] R. J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.

[2] R. Kavitha Kumar and Dr. R. M. Chandrasekar. Missing data imputation in cardiac data set (Survival progonosis).

[3] R.S. Somasundaram, R. Nedunchezhian, "Evaluation on Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications, Vol21-No. 10, May 2011, pp14-19.

[4] Graham, J.W,"Missing Data Analysis: Making it work in the real world. Annual Review of Psychology", 60, 546-576, 2009.

[5] Jeffrey C.Wayman, "Multiple Imputation for Missing Data: What Is It And How Can I Use It?", Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL, pp. 2-16,2003.

[6] A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo St ijnen, Kernel G.M Moons, "Review: A gentle introduction to imputation of missing values", Journal of Clinical Epidemiology 59, pp.1087-1091,2006.

[7] Kin Wagstaff, "Clustering with Missing Values: No Imputation Required"-NSF grant IIS-0325329, pp.1-10.

[8] S.Hichao Zhang, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin, Chengqi Zhang, "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg, 2008.

[9] Shalini S. Singh, N C Chauhan – "K-means v/s K-medoids: A comparative Study". National Conference on Recent Trends in Engineering & Technology,(13-14 May 2011).

[10] Blessie, C.E.; Karthikeyan,E.: Selvaraj,B. (2010): NAD – A Discretization approach for improving interdependency, Journal of Advanced Research in Computer Science, 2910, pp. 9-17.

[11] Liu, H.; Setiono,R. (1997): Feature selection via discretization, IEEE Transaction on Knowledge and Data Engineering 9(4), pp. 642-645.

[12] Ms.R.Malarvizhi, Dr. Antony Selvadoss Thanamani- "K-Nearest Neighbor in Missing Data Imputation", International Journal of Engineering Research and Development, Volume 5 Issue 1- November-2012.