Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 1 (2016), pp. 1131-1140

http://www.ripublication.com

A Monotonic Sequence and Subsequence Approach in Missing Data Statistical Analysis

S. Kanchana
Research Scholar, Research Department of Computer Science,
NGM College, Pollachi-642001, Bharathiyar University, Coimbatore, India.

Dr. Antony Selvadoss Thanamani
Professor and Head, Research Department of Computer Science,
NGM College, Pollachi-642001, Bharathiyar University, Coimbatore, India.

Abstract

Missing values are ubiquitous issues in research design stage. This paper focus on the guidelines for researchers deal with analyzing partially perceive datasets describes the issues that need to be considered. In this chapter we discuss a variety of methods to handle missing data using machine learning techniques for the imputation of missing values in large datasets. Multiple imputation produce right value to replace whereas single value imputation produce biased results. This article deal with several algorithms in supervised and Unsupervised machine learning techniques like Mean, Median, Standard Deviation, Regression and Naïve Bayesian classifier. The performance of above method has been compared by using correlation statistics analysis gives the imputed values are positively related or negatively related or not related with each other. To evaluate the performance of missing values can be measured by using central tool called Bolzano Weierstrass, which is proving several properties of continuous function. Based upon the performance of Monotonic sequence and subsequence, can able to find the imputed missing values are increasing or decreasing or a bounded monotonic sequence of finite limit and also analyzing that every bounded sequence of missing values has a convergent subsequence. To evaluate the performance, the standard machine learning repository dataset has been used. This article focuses primarily on how to implement Bolzano Weiestress theorem to perform imputation of missing values.

Keywords: Bolzano Weiestress, Bounded Monotonic sequence, Machine Learning Techniques, Monotonic sequence, Naïve Bayesian, Supervised Machine Learning Techniques

Introduction

Missing data problem is a common aspect in many practical situations. The treatment of missing data in research is not a simple issue. A different methods have been developed to compensate for missing data. The imputation of missing data is an actual and challenging issue confronted by machine learning and data mining. Most of the real world datasets are characterized by an unavoidable problem of incompleteness, in terms of missing values. Missing value may generate bias and affect the quality of the supervised learning process. Missing value imputation is an efficient way to find or guess the missing values based on other information in the datasets. Data mining consists of the various technical approaches including machine learning, statistic and database system. The main goal of the data mining process is to discover knowledge from large database and transform into a human understandable format. This paper focuses on several algorithms such as missing data mechanisms, multiple imputation techniques and supervised machine learning method. Experimental results are separately imputed in each real datasets and checked for accuracy.

The mechanism causing the missing data can influence the performance of both imputation and complete data methods. There are three different ways to categorize missing data as defined in [1]. Missing Completely At Random (MCAR) if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters. Missing At Random (MAR) is the alternative, suggesting that what caused the data to be missing does not depend upon the missing data itself. Not Missing At Random (NMAR) is data that is missing for a specific reason.

In the rest of this paper gives the background work or the related work in section II, machine learning technique concepts in Section III, Section IV introduces new methods based on Naïve Bayesian Classifier to estimate and replace missing data. Experimental analyses of NBI model in Section V and the Conclusions are discussed in Section VI.

Literature Survey

Little and Rubin [1] summarize the mechanism of imputation method. Also introduces mean imputation [2] method to find out missing values. The drawbacks of mean imputation are sample size is overestimated, variance is underestimated, correlation is negatively biased. For median and standard deviation also replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Different types of machine learning techniques are supervised and unsupervised machine learning techniques summarized in [3]. Classification of multiple imputation and experimental analysis are described in [4]. Min Pan et al. [5] summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Comparison of different unsupervised machine learning techniques are referred from survey paper [6]. To overcome the unsupervised problem Peng Liu, Lei Lei et al. [7] applied the supervised machine learning techniques called Naïve Bayesian Classifier. Figure 4 states that NBC produce accurate results compare to the existing supervised method.

Machine Learning concepts

In the data mining context, machine learning technique is generally classified as supervised and unsupervised learning technique both belong to machine learning technique [8]. Supervised classification focus on the prediction based on known properties and the classification of unsupervised focus on commonly used classification algorithm known as Naïve Bayesian imputation techniques.

A. *Unsupervised Machine Learning Concepts*

Mean Imputation is the process of replacing the missing data from the available data where the instance with missing attribute belongs.

Median Imputation is calculated by grouping up of data and finding average for the data. Median can be calculated by finding difference between upper and lower class boundaries of median class.

Standard Deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. Estimate standard deviation based on sample and entire population data.

Regression is a statistical package for determining and evaluating the correlation among variables. It contain many techniques for modeling and analyzing several variables, focusing on the relationship between a dependent variable and one or more independent variables. It is used for prediction and forecasting.

Correlation is a powerful statistical technique which gives us if two variables are related. In this articles represent the coefficient of correlation which used to describe the relationship is positive or negative and also can understand about the strength of relationship. Experimental analysis compare 2 variables from large dataset in different ways. First compare the variables without missing values, with missing values and the imputation of missing values. Every task it analysis the correlation and gives us the result of these two variables. Find the mean of X, and the mean of Y, Subtract the mean of X from every X value and the same process for Y also. Calculate a X b, a^2 and b^2 for every value, sum up a X b, sum up a^2 and sum up b^2. Divide the sum of a X b by the square root of [(sum of a^2) X (sum of b^2)]

B. *Supervised Machine Learning Concepts*

Another way of learning technique is classified as supervised learning that focus on the prediction based on known properties. Naïve Bayes technique [9] is one of the most useful machine learning technique based on computing probabilities. It analyzes relationship between each independent variable and the dependent variable to derive a conditional probability for each relationship. A prediction is made by combining the effects of the independent variables on the dependent variable which is the outcome that is predicted. It requires only one pass through the training set to generate a classification model, which makes it very efficient. The Naïve Bayesian generates data model which consists of set of conditional probabilities, and works only with discrete data.

Multiple Imputation Techniques

Multiple imputation for each missing values generated a set of possible values, each missing value is used to fill the data set, resulting in a number of representative sets of

complete data set for statistical methods and statistical analysis. The main application of multiple imputation [10] process produces more intermediate interpolation values, can use the variation between the values interpolated reflects the uncertainty that no answer, including the case of no answer to the reasons given sampling variability and non- response of the reasons for the variability caused by uncertainty. Multiple imputation simulate the distribution that well preserve the relationship between variables. It can give a lot of information for uncertainty of measuring results of a single interpolation is relatively simple.

C. *Naïve Bayesian Imputation*

Naïve Bayesian Imputation is one of the most useful machine learning technique based on computing probabilities [11]. It uses probability to represent each class and tends to find the most possible class for each sample. It analyzes relationship between each independent variable and the dependent variable to derive a conditional probability for each relationship. A prediction is made by combining the effects of the independent variables on the dependent variable which is the outcome that is predicted Naïve Bayesian Classifier generates full use of all the data in the present dataset. This paper focus a new method based on Naïve Bayesian classifier to handle missing data called Naïve Bayesian Imputation (NBI).

Bayes theorem [12] provides a way of calculating the posterior probability P(C/X) of class from P(C) is the prior probability of class, P(X) is the prior probability of predictor and P(X/C) is the likelihood which is the probability of predictor given class. Naïve Bayes classifier assume that the effect of the value of a predictor (X) on a given class (C) is independent of the values of other predictors called conditional independence. Fig 1. Shows the pictorial representation of proposed system.

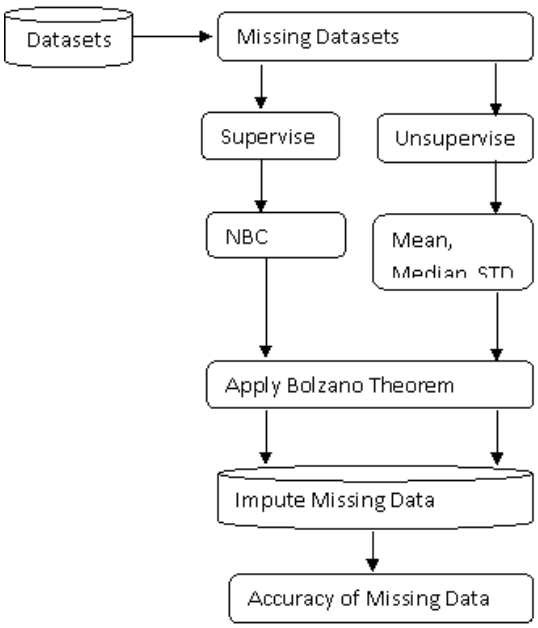

Figure 1: Flowchart of the Proposed System

D. *Bolzano Theorem*

The Bolzano Weierstrass theorem states that every defined group in (R_n) consist of a concurrent subgroup. For instance [13], a subgroup is a group that can be derived from another group by deleting any items without modifying the order of the resting items. Every bounded real sequence has a convergent subsequence. A subset of R is compact if and only if it is closed and bounded. The set $S \coloneqq Q \cap [0,1]$, since rational are countable, and treat S as a bounded sequence from 0 to 1. Then it gives the following results for each statement are listed 1. There is a convergent subsequence in S. For example. $S_n \coloneqq \frac{1}{n},\ n \in N$. N Is not compact since it is not closed. Bolzano Weierstrass require an infinite construction, and it has no exception. The infinite construction is easier than the constructions in other proof. If (R_n) is a sequence of numbers in the closed segment $[M,N]$, then it has a subsequence which converges to a point in $[M,N]$. Let's have an arbitrary point P, which is between the points M and N. Then observe the segment $[M,P]$. It may contain a finite number of members from the sequence (R_n) and it may contain an infinite number of them. If take the point P to be N, the segment $[M,N]$ would contain an infinite number of members from the sequence.

If take the point P to be M, the segment $[M,N]$ would contain at most only one point from the sequence. Let's introducing the set $S = \{P \epsilon [M,N][M,P]\}$ contains a finite number of (R_n) members. M belongs to set S. If a point P belongs to S, it mean that $[M,N]$ has a finite number of members from (R_n), and it will mean that any subset of $[M,P]$ would also have only a finite number of members from (R_n). Therefore for any P that belongs to S, all the point between that P and M would also belongs to S. The set S is actually a segment, starting at M and ending in some unknown location $[M,N]$. Now let's move to next step $R = Sup(S)$ it means R is an accumulation point of (R_n). According to the special case $R = M$, and assume that $R \in (M,N)$. Now we take an arbitrarily small ε . Observe the segment $[M,R+\varepsilon]$. $R+\varepsilon$ Cannot belong to S since it is higher than the supermom. Hence $[M,R+\varepsilon]$ contains an infinite number of (R_n) members. Now the segment $[M,R-\varepsilon]$. $R-\varepsilon$ Must belong to S, since it is smaller than the supermom of the segment S. Thus $[M,R-\varepsilon]$ contains a finite number of members from (R_n). But $[M,R-\varepsilon]$ is a subset of $[M,R+\varepsilon]$. If the bigger set contains an infinite number of (R_n) members and its subset contains only a finite amount, the complement of the subset must contain an infinite number of members from (R_n). Proved that for every ε, the segment $(R-\varepsilon, R+\varepsilon)$ contains an infinite number of members from the sequence. Construct a subsequence of (R_n) that converges to R. Take ε to be 1. Take any (R_n) member in $(R-1, R+1)$ to be the first member. This theorem proof that every bounded sequence of real numbers has a convergent subsequence, every bounded sequence in R^n has a convergent subsequence and every sequence in a closed and bounded set S as R^n has a convergent subsequence.

Results Analysis

E. *Experimental Design*

Experimental datasets were carried out from the Machine Learning Database UCI Repository. Table 1. describes the dataset with electrical impedance measurements in

samples of freshly excised tissue dataset contains number of instances and number of attributes about the datasets used in this paper. Datasets without missing values are taken and few values are removed from it randomly. The main objective of the experiments conducted in this work is to analyze the multiple imputation of machine learning algorithm. Then the performance of this method has been compared by using Correlation statistics analysis which produces the imputed values are positively related or negatively related or not related with each other. The rates of the missing values removed are from 5% to 25%.

Table 1: Datasets used for Analysis

Datasets	Breast Tissue
Instances	106
Attributes	10 (9features + 1 classes)
Missing rates	5% to 25%
Unsupervised	Mean, Median, Standard Deviation, Correlation, Regression
Supervised	Naïve Bayesian

F. *Experimental Analysis*

The below Figure 2. Represent the single instance of original datasets with 5% missing values M and it's to be implement using mixed learning techniques. The bounding line indicate a pivot point that aligns between interval periods.

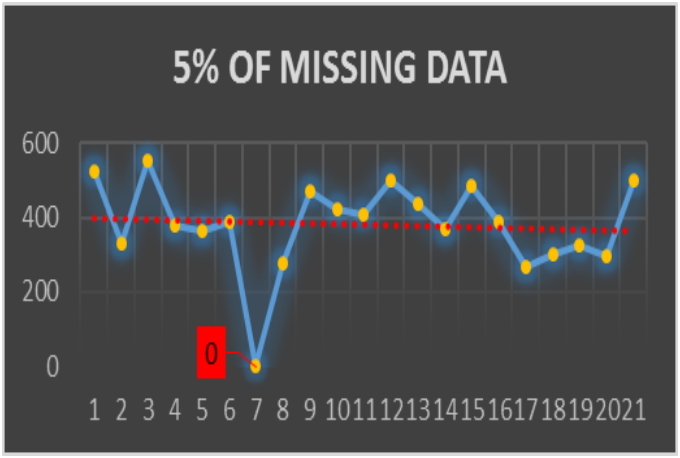

Figure 2: 5% of missing data

Figure 3. Represent the imputation of 5% missing data using supervised and unsupervised machine learning techniques. The sequence of missing data M is a bounded monotonic sequence and has a finite limit.

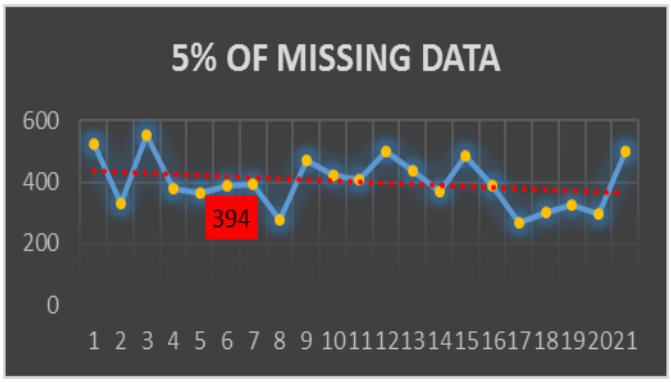

Figure 3: Imputation of missing data (5%)

The below Figure 4. Represent the single instance of original datasets with 25% of missing M and it's to be implement using mixed learning techniques.

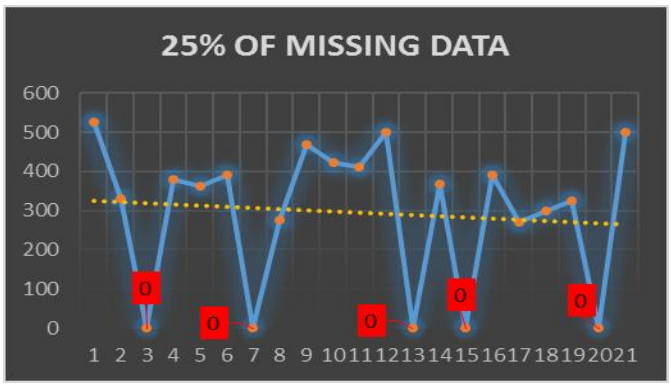

Figure 4: 25% of Missing data

Figure 5. Represent the imputation of 25% missing data using supervised and unsupervised machine learning techniques.

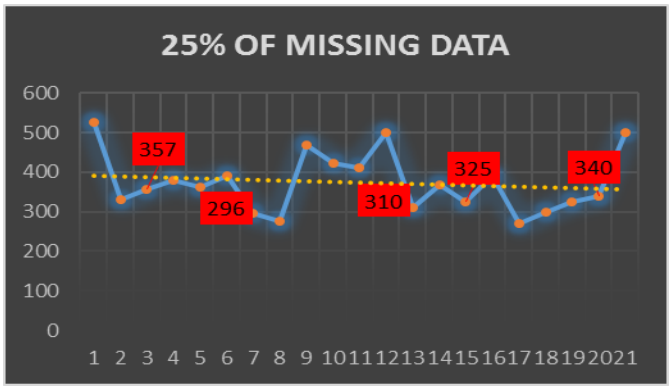

Figure 5: Imputation of 25% of Missing data

Figure 6 represent the experimental results of supervised machine proof that every sequence of real numbers is monotonic if it is either increasing or decreasing. A bounded monotonic sequence always has a finite limit.

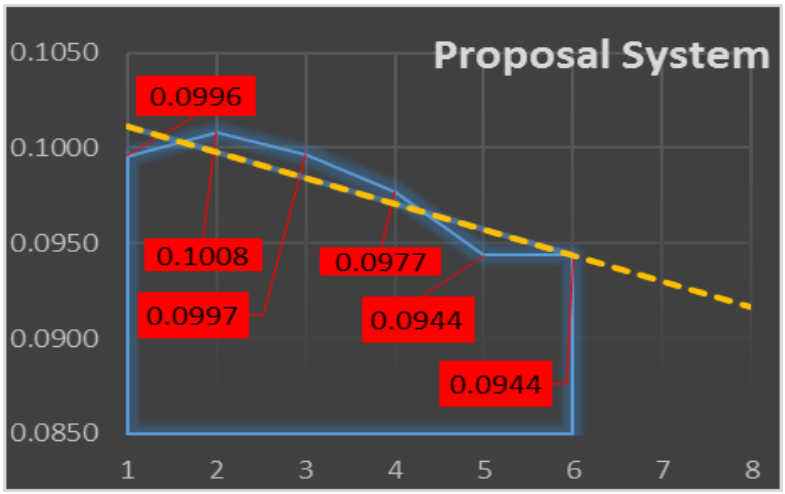

Figure 6: Experimental results for Supervised Techniques

The below Figure 7. Gives the coefficient of correlation value R^2=0.8511 for the original dataset which indicates high positive relation between variables.

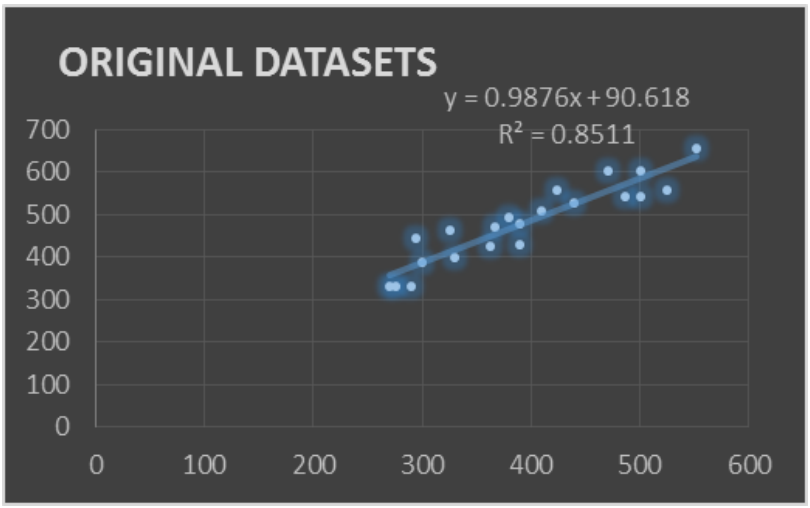

Figure 7: Correlation chart of original dataset

Figure 8 specify the imputation of missing data using NB Techniques and the correlation value R^2=0.027, no correlation between variables.

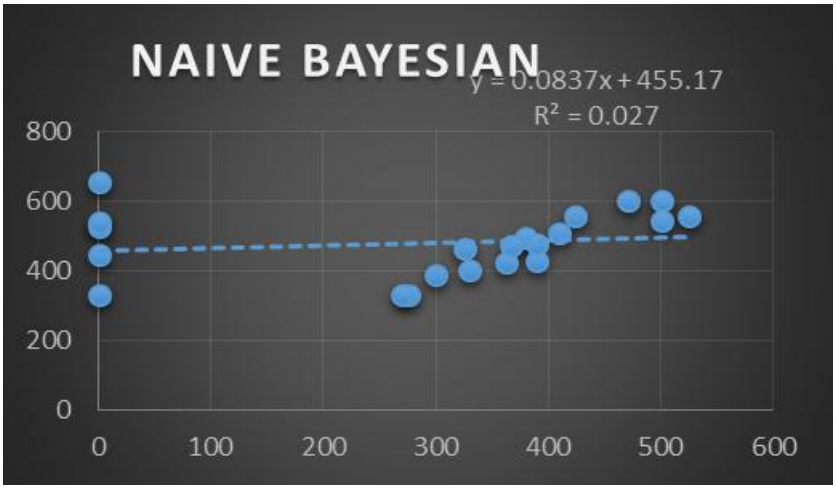

Figure 8: Correlation value of Naïve Bayesian Techniques.

Discussion

This paper gives the complete view about the multiple imputation of missing values in large dataset. This article proposed multiple imputation using machine learning techniques of both supervised and unsupervised algorithms and also shows the experimental results of correlation between variables. Several master plan of NBI are examined in the experiments. The evaluation results show that NBC is superior to multiple imputation. The performance of NBC is improved by the attribute selection. When the imputation attribute has been defined, the order of irrelevant master plan is recommended. The performance of missing values can be measured by using central tool called Bolzano Weierstrass, which proved the several properties of continuous function.

Conclusion

The performance of Monotonic sequence and subsequence approach of missing data, analysed that the imputed missing values of every sequence of real numbers has a monotonic subsequence and a bounded monotonic sequence always has a finite limit and also analysed that every bounded sequence of missing values has a convergent subsequence. This article focused primarily on how to implement Bolzano Weiestress theorem to perform imputation of missing values.

This paper presents an efficient and effective missing data handling method, Naïve Bayesian Classifier model. According to the common imputation techniques, Byes classifier is an effective missing data treatment model. In future it can be extended to handle categorical attributes and it can be replaced by other supervised machine learning techniques.

References

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

[2] 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.

[3] 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.

[4] Mrs.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,

[5] Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, "Experimental Analysis of Methods for Imputation of Missing Values in Databases.

[6] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases", Applied Intelligence, vol 11, pp., 259-275, 1999.

[7] Peng Liu, Lei Lei, "Missing Data Treatment Methods and NBI Model", Sixth International Conference on Intelligent Systems Design and Applications, 0-7695-2528-8/06.

[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] 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.

[10] R. Kavitha Kumar and Dr. R. M. Chandrasekar, "Missing Data Imputation in Cardiac data set".

[11] Ingunn Myrtveit, Erik Stensrud, "IEEE Transactions on Software Engineering", Vol. 27, No 11, November 2001.

[12] S. Kanchana, Dr. Antony Selvadoss Thanamani, "Classification of Efficient Imputation Method for Analyzing Missing values", International Journal of Computer Trends and Technology, Volume-12 Part-I, P-ISSN: 2349-0829.

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

[14] S. Kanchana, Dr. Antony Selvadoss Thanamani, "Multiple Imputation of Missing Data Using Efficient Machine Learning Approach", International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 10, Number 1 (2015) pp.1473-1482.

[15] S. Kanchana, Dr. Antony Selvadoss Thanamani, "Experimental Analysis of Imputation of Missing Data Using Machine Learning Techniques", International Journal of Advanced Information Science and Technology, ISSN 2319-2682 pages 128-132.