BOOSTING THE ACCURACY OF WEAK LEARNER USING SEMI SUPERVISED CoGA TECHNIQUES

Kanchana S. and Antony Selvadoss Thanamani
Department of Computer Science, NGM College, Pollachi, Bharathiyar University, Coimbatore, India
E-Mail: kskanch@gmail.com

ABSTRACT

This article elucidate and appraise a technique for imputing missing values using right machine learning approach for predictive analytics solutions. Using supervised and unsupervised learning techniques make predictions based on historical dataset. This survey carried out using comprehensive range of databases, for which missing cases are first filled by several sets of reasonable values to create multiple finalized datasets, later standard data procedures are inserted to each destination dataset, parallel multiple sets of output are merge to produce a single inference. In statistics, the Naïve Bayesian approach provide supplemented information in the form of a prior probability distribution, prior information about the function to generate and estimates misplaced parameters. The main goal of this article provides suitable data imputation algorithms and also implementing Bolzano Weierstrass in machine learning techniques to evaluate the performance of every sequence of rational and irrational number has a monotonic subsequence. To reducing bias data, implementing Boosting algorithms to perform the process of turning the noisy classifier into final classifier then to correlate with true classification. This articles represent AdaBoost techniques to improve the performance of the final classifier. Experimental results shows the proposed approach have good accuracy and results of simulation studies are also presented.

Keywords: AdaBoost techniques, bolzano weierstrass, boosting algorithm, naïve bayesian, supervised, unsupervised learning.

INTRODUCTION

Numerous existing industrial and research datasets contain missing values. Specifying various reasons such as system errors, data entry process and incorrect analysis. Simplest part of dealing missing values is to discard the instances that contain the noisy datasets and its produce bias result. Missing value may process bias and affect the quality of the supervised learning process. Though, the absence of data may effect data analysis, simple way to deal with missing data are, supervised and unsupervised machine learning techniques. Multiple imputation of missing value is an effective way to find and analysis the missing values based on other information in the datasets [1]. Various imputation techniques are available in data mining such as machine learning, statistic and database system. Main goal of this mining process is to determine knowledge from large database and modify into user understandable format. This article concentrate on various algorithm includes missing data mechanisms, imputation techniques and machine learning techniques. Experimental analysis shows that the results are separately imputed in datasets and checked with the proposed techniques have good accuracy.

How to deal missing data

Almost in the real world datasets are classified by an unavoidable issue of incompleteness, in spite of missing data. A small method for handling missing data is to bring forward all the values for any pattern removed more than one items from the dataset [2] [3]. The major problem among here content may be decreased. Especially this is applicable though the decreased pattern content be smaller to attain momentous outcome in the study. In parallel case further sampling item sets can be collected.
The most traditional missing value techniques [4] are mean value imputation, standard deviation imputation, deleting case, maximum likelihood and other statistical approach. At present research has explored the use of machine learning techniques for imputation of missing values. The major issues hold huge data sets that might be noticeable [5]. As an instance assuming that an application along 5 query is about 10% of the item sets, later on moderate almost 60% of the sampling may obtain at minimum one query might be missing.

Missing data mechanism

The mechanism of missing data can influence the performance of both imputation and complete data methods. Three different ways to categorize missing data as defined in missing completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).These characteristics might be quite relevant to the analysis.

LITERATURE SURVEY

Classification of multiple imputation and experimental analysis are described. Min Pan et al. summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Little and Rubin summarize the mechanism of imputation method. Introduces mean imputation 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. Comparisons of different unsupervised machine learning technique are referred from survey paper. To overcome the unsupervised problem Peng Liu, Lei Lei *et al*. applied the supervised machine learning techniques of Naïve Bayesian Classifier.

MULTIPLE IMPUTATION TECHNIQUES

Multiple imputations of each missing values generated a set of possible values, each missing value used to fill the data set, resulting in a number of representative sets of complete data set for statistical methods and analysis [6]. The main application of multiple imputation process produces more intermediate interpolation values, that can use the variation between the values interpolated reflects the uncertainty has no answer, including the case of no answer to the reasons given sampling variability. This technique simulates the distribution that preserves the relationship between variables [7]. This process requires performing in three different steps called imputation, analysis and pooling. Rubin in the year 1987 has shown that to generate imputations is valid then the resulting inferences will be statistically valid.

MACHINE LEARNING TECHNIQUES

Machine learning technique is a model of analytical building model [8]. With the help of this technique, it allows computers to find missing data without being programmed where to view. This technique is most important because as models are exposed to new data, it can able to separately adapt. Nowadays more interest in machine learning due to the factors that provide data mining and Bayesian analysis more powerful than other. Because huge volumes and different varieties of data, computational processing is more powerful, cheaper and economical storage of data. The most broadly used machine learning about 70% of supervised learning and about 10 to 20% of unsupervised learning techniques [9]. Very rarely used techniques are semi-supervised and reinforcement learning. The below Figure-1 shows the structure of machine learning techniques.

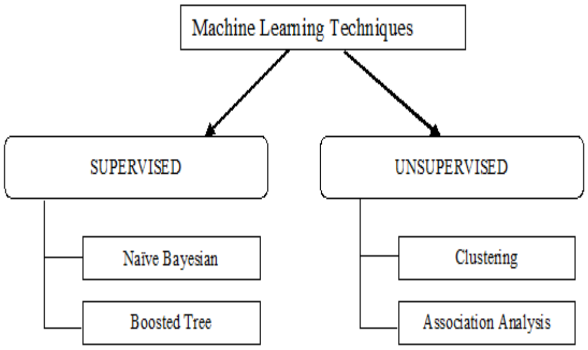

Figure-1. Structure of machine learning techniques.

Supervised machine learning

This learning technique uses different patterns to predict the values of the number on added unlabeled data. This algorithm is generally used to predict historical data for future purpose. This technique accepts a set of input data along with the relevant output to find noisy data in the data set. With the help of classification model, regression model, prediction model and gradient boosting it can modify the input data with corresponding proper output data. Mean, Median and Standard Deviation calculate the scatter data concerning the mean value [10]. It can be convenient in estimating the set of fact which can possess the identical aim but a different domain. Estimate standard deviation based on sample data.

Unsupervised machine learning

Unsupervised has no historical labels to predict the values of the unlabeled data. This algorithm goal [19] is to exhibit the data and find some structure within the dataset. Most popular techniques include Nearest Neighbor techniques, K-Means clustering algorithm and singular value decomposition. These techniques also used for segment text topics, recommend items and identify data outliers. Another way of learning technique is classified as supervised learning that focus on the prediction based on known properties. Naïve Bayesian technique [14] [15] is one of the most useful machine learning techniques based on computing probabilities. It analysis 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 [20].

ANALYSIS OF PROPOSAL TECHNIQUES

The multiple imputations for each missing values generated a set of possible values, each missing values is used to fill the data set, resulting in a number of representative sets of complete data set for statistical methods and statistical analysis. This technique simulate the distribution that preserve the relationship between variables [11]. It can give a lot of information for uncertainty of measuring results of a single interpolation is relatively simple. The main application of multiple imputation process produces more intermediate interpolation values. 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.

Naïve Bayesian classifier

In Naïve Bayesian Classifier is one of the most useful machine learning techniques based on computing probabilities. This classifier frequently executes especially strong and widely used because it continually execute further advanced classifying methods [12]. Naïve Bayesian Classifier uses probability to represent each class and tends to find the most possible class for each sample. It is a popular classifier, not only for its good performance, simple form and high calculation speed, but also for its insensitivity to missing data is called Naïve Bayesian Imputation classifier to handle missing data.

Bolzano Weierstrass Theorem

The Bolzano Weierstrass theorem [13] [18] states that every defined group in (R_n) consist of a concurrent subgroup. For instance [18], 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 $\mathrm{S} \coloneqq \mathrm{Q} \cap [0,1]$, 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. $\mathrm{S_n} \coloneqq \frac{1}{n}$, $\mathrm{n} \in \mathrm{N}$. 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 $[\mathrm{M,N}]$, then it has a subsequence which converges to a point in $[\mathrm{M,N}]$. Let's have an arbitrary point P, which is between the points M and N. Then observe the segment $[\mathrm{M,P}]$. 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 $[\mathrm{M,N}]$ would contain an infinite number of members from the sequence [21].

If take the point P to be M, the segment $[\mathrm{M,N}]$ would contain at most only one point from the sequence. Let's introducing the set $\mathrm{S} = \{\mathrm{P} \epsilon [\mathrm{M,N}][\mathrm{M,P}]\}$ contains a finite number of (R_n) members. M belongs to set S. If a point P belongs to S, it mean that $[\mathrm{M,N}]$ has a finite number of members from (R_n), and it will mean that any subset of $[\mathrm{M,P}]$ 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 $[\mathrm{M,N}]$. Now let's move to next step $\mathrm{R} = \mathrm{Sup}(\mathrm{S})$ it means R is an accumulation point of (R_n). According to the special case $\mathrm{R} = \mathrm{M}$, and assume that $\mathrm{R} \in (\mathrm{M,N})$. Now we take an arbitrarily small ε. Observe the segment $[\mathrm{M,R} + \varepsilon]$. $\mathrm{R} + \varepsilon$ Cannot belong to S since it is higher than the supremum. Hence $[\mathrm{M,R} + \varepsilon]$ contains an infinite number of (R_n)members. Now the segment $[\mathrm{M,R} - \varepsilon]$. $\mathrm{R} - \varepsilon$ Must belong to S, since it is smaller than the supremum of the segment S. Thus $[\mathrm{M,R} - \varepsilon]$ contains a finite number of members from(R_n). But $[\mathrm{M,R} - \varepsilon]$ 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 $(\mathrm{R} - \varepsilon, \mathrm{R} + \varepsilon)$ 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 $(\mathrm{R} - 1, \mathrm{R} + 1)$ to be the first member. This theorem proof that every bounded sequence of real numbers has a convergent subsequence, every bounded sequence in $\mathrm{R^n}$ has a convergent subsequence and every sequence in a closed and bounded set S as $\mathrm{R^n}$ has a convergent subsequence. The following Figure-2 specifies Bolazano weierstrass model.

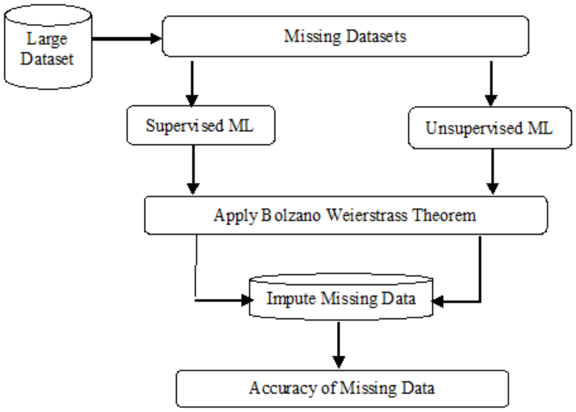

Figure-2. Bolzano Weierstrass techniques.

SEMI-SUPERVISED TRAINING ALGORITHM

Machine learning techniques provide ensemble meta-algorithm to reducing biased data for a particular input X and it is incorrect when predicting the correct output X [16] [17]. The classification of the data is slightly correlated with the predicted classification is said to be weak dataset and it is arbitrarily correlated with predicted classification is known as strong dataset. Boosting algorithms consist of iteratively weak classifiers and adding into a final strong classifier. When weak classifier is added into a strong classifier then it became weighted in some way that is normally related to the weak learner accuracy.

Boosting algorithm

To improve the performance of the weak learners, boosting algorithm combines weak learners into a single strong learner. With the help of gradient boosting technique, produces prediction model in the process of obtaining better predictive performance i.e. ensemble prediction model F that predicts values $\hat{g}$=F(x), x is perfect model minimizing the average of the squares of the deviations $(\hat{g}\text{-}g)^2$ to the correct values g. Every stage $1 \leq m \leq M$ of gradient boosting, it may assumed some imperfect model F_m, instead it improves a new model that adds an estimator ê to provide a strong model $F_{m+1}(x) = F_m(x) + e(x)$. e(x) is said to be weighted sum of functions for some weak learners. Every time choosing arbitrary loss function L is an optimization problem so it is much easier to resolve and it added more weightage to the true dataset. To increase the performance of the weighted true classification, focuses on Adaptive boosting techniques. It initialize the weights for each items in a dataset, fit a classification tree with the trained dataset, analysis the misclassification error, update the weights for each weak learners and finally generate the effective performance of strong learners. Figure-3 represent proposal model of CoGA boosting techniques.

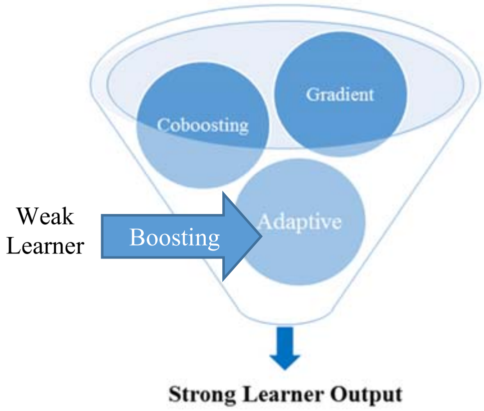

Figure-3. CoGA Boosting techniques.

```
Algorithm

Input:
$x_1 … x_n, w_1 …w_n$

Output:
$y_1 … y_n$, $y \epsilon \{-1, 1\}$

Initialize:
 i, j =0, g (j, $x_i$) = 0, w = 1 … 1/n

Loop: t= 1 … T, j= 1, 2

▪ Set Virtual Distribution $D_t^j$ (I)
▪ Find weak learner $h_t^j$
▪ Find the real valued y and find similarity $\hat{g}(x)$
▪ Set weak learner $h_m(x)$ where $x\epsilon \{-1, 1\}$
▪ Calculate multiplier $\lambda_m$ to minimize the loss function
▪ Update the model $G_m(x) =G_{m-1}(x) +\lambda_m h_m(x)$
▪ Find Error function ¢ ($\hat{g}(x)$, y, i) =$¢^{-y}{}_i\hat{g}$ ($x_i$)
▪ Update weight $w_{i,}$ t+1 for all i

Output: simplify new weights and produce strong learner.
```

Basic idea

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. 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.

Table-1. Datasets used for analysis.

S. No.	Parameter	Instances
1	Datasets	Breast Tissue
2	Instances	106
3	Attributes	10 (9features + 1 classes)
4	Missing rates	5% to 25%
5	Unsupervised	Mean, Median, Standard Deviation
6	Supervised	Naïve Bayesian

EXPERIMENTAL RESULTS

The below Figure-4 gives the coefficient of correlation value R^2=0.8511 for the original dataset which indicates high positive relation between variables.

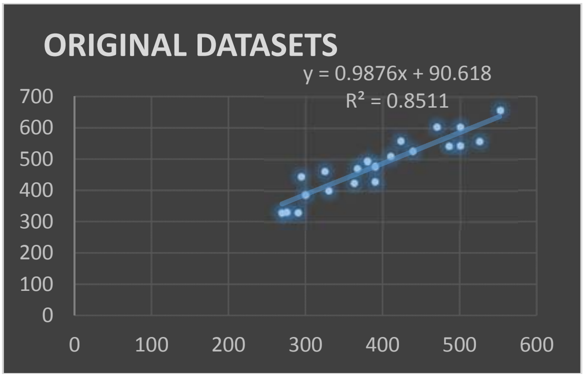

Figure-4. Correlation chart of original dataset.

The below Figure-5 represents the percentage rates of missing values using both the techniques like supervised and unsupervised using missing values with the rate of 5%, 10%, 15%, 20% and 25% respectively.

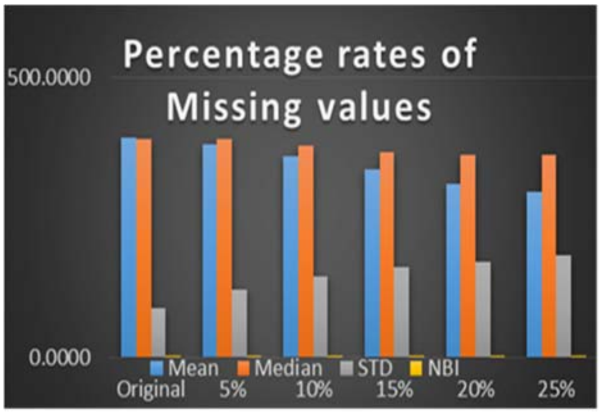

Figure-5. Percentage rates of missing values.

The following 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.

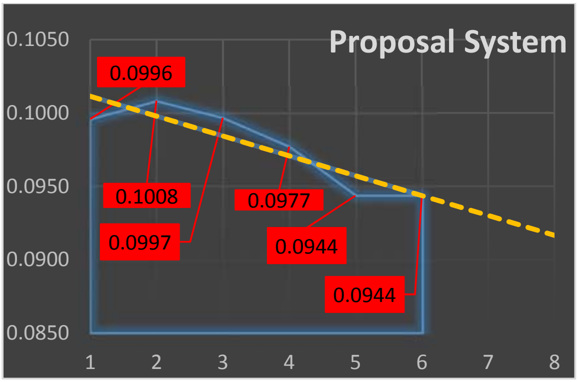

Figure-6. Experimental results for Supervised techniques.

Table-2 describes the percentage of missing value occur in the original dataset using machine learning techniques of supervised like Naïve Bayesian Classifier and unsupervised Mean, Median and Standard Deviation compared with all the other attributes.

Table-2. Percentage of missing values in dataset.

	Orig	5%	10%	15%	20 %	25%
Mean	394.2	380.4	360	336.4	310	296.1
Medi	389.9	389.9	380	366.9	363	362.8
STD	87.0	120.9	146	162.2	170	183.1
NBI	0.1	0.1	0.1	0.1	0.1	0.1

DISCUSSIONS

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. The evaluation results show that NBC is superior to multiple imputations. 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 most important drawbacks of Bayes classifier is that it has strong feature independence assumptions.

The performance of missing values can be measured by using central tool called Bolzano Weierstrass, which proved the several properties of continuous function. This problem can be solved by using Boosting algorithm which translates weak dataset to strong dataset. The following Figure-7 shows the outcome of weak learner.

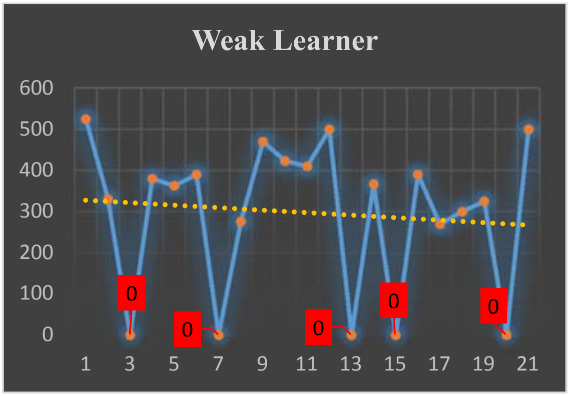

Figure-7. Percentage rates of missing values.

Figure-8 shows the performance of the weak learners, boosting algorithm combines weak learners into a single strong learner with the help of CoGA boosting technique.

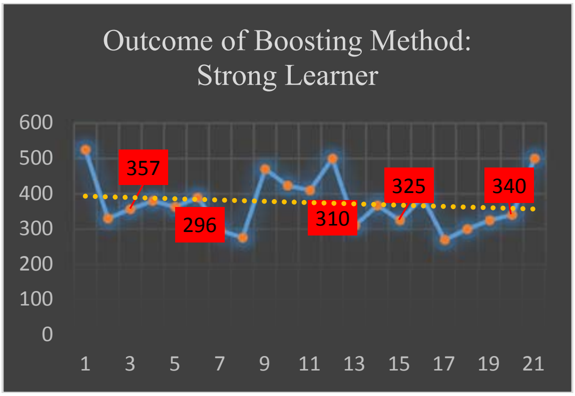

Figure-8. CoGA Boosting techniques.

CONCLUSIONS

This article focused the performance of Bolzano Weiestress theorem imputed missing values of every sequence of real numbers has a monotonic subsequence and a bounded monotonic sequence always has a finite limit. It also generate 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.

Bolzano theorem proved that every continuous function on a closed bounded interval is bounded and also analysis the distance between a closed bounded set and a closed set in R^n. Machine learning techniques provide ensemble meta-algorithm to reducing biased data. This problem can be solved by using Boosting algorithm which translates weak dataset to strong dataset. To improve the performance of the weak learners, boosting algorithm combines weak learners into a single strong learner with the help of CoGA boosting technique.

REFERENCES

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

[2] Blessie C.E., Karthikeyan E, Selvaraj. B. 2010. NAD - A Discretization approach for improving interdependency, Journal of Advanced Research in Computer Scienc. pp. 9-17.

[3] E. Chandra Blessie, DR.E. Karthikeyan and DR.V.Thavavel. Improving Classifier Performance by Imputing Missing Values using Discretization Method. International Journal of Engineering Science and Technology.

[4] Han J. and Kamber M. 2001. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

[5] Ingunn Myrtveit, Erik Stensrud. 2001. IEEE Transactions on Software Engineering. 27(11).

[6] Jeffrey C.Wayman. 2003. 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.

[7] Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad. Imputation of Missing Data Using Machine Learning Techinques. from KDD-96 Proceedings.

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

[9] K. Raja, G. Tholkappia Arasu, Chitra S. Nair. 2012. Imputation Framework for missing value. International Journal of Computer Trends and Technology. 3(2).

[10] Lim Eng Aik and Zarita Zainuddin. 2008. A Comparative Study of Missing Value Estimation Methods: Which Method Performs Better? 2008 International Conference on Electronic Design.

[11] Liu P., Lei L. and Wu N. 2005. A Quantitative Study of the Effect of Missing Data in Classifiers, proceedings of CIT2005 by IEEE Computer Society Press.

[12] 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.

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

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

[15] R. Malarvizhi, Dr. Antony Selvadoss Thanamani. 2012. K-Nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development. 5(1).

[16] R.S. Somasundaram, R. Nedunchezhian. 2011. 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.

[17] Shichao Zhang, Xindong Wu, Manlong Zhu. 2010. Efficient Missing Data Imputation for Supervised Learning. Proc, 9th IEEE conference on Cognitive informatics. IEEE.

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

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

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

[21] 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. pp. 128-132.