Elevating the Accuracy of Missing Data Imputation Using Bolzano Classifier

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

[#] Research Scholar, Research Department of Computer Science,
NGM College, Pollachi 642001, Bharathiyar University, Coimbatore, India.
[1] kskanch@gmail.com

[*] Professor and Head, Research Department of Computer Science,
NGM college, Pollachi 642001, Bharathiyar University, Coimbatore, India.
[2] selvdoss@gmail.com

***Abstract*—Missing data occur in almost all serious statistical analyses. In statistics, imputation is the process of replacing missing data with substituted values. Simple imputation is attractive often used to impute missing data whereas multiple imputation generates right value to replace. This paper evaluates multiple imputation of missing data in large datasets and the presentation of MI focuses on several unsupervised ML algorithms like mean, median, standard deviation and Supervised ML techniques for probabilistic algorithm like NBI classifier. 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, then standard complete data procedures are register to each finalized dataset, and eventually the multiple sets of results are merge to produce a single inference. Main goal is to provide general guidelines on selection of suitable data imputation algorithms and also implementing Bolzano theorem in machine learning techniques to evaluate the performance of every sequence of rational and irrational number has a monotonic subsequence. To evaluate imputation of missing data, the standard machine learning repository dataset has been used. Experimental results shows the proposed approach have good accuracy and the accuracy measured in terms of percentage.**

Keyword - Bolzano Classifier, Imputation Algorithm, NBI Classifier, Supervised ML, Unsupervised ML.

I. INTRODUCTION

Missing data imputation is an actual and challenging issue confronted by machine learning and data mining [1]. 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 algorithm. Experimental results are separately imputed in each real datasets and checked for accuracy [2].

A simple techniques for handling with lost value is to bring forward all the values for any pattern removed one or more info items. The major issues among here content may be decreased. Especially this is applicable although the decreased pattern content be smaller to attain momentous outcome in the study. In parallel casing further sampling item sets can be collected. The mentioned issue hold an enormous data sets that might be noticeable. As an illustration assuming that an application along 5 query is about lost 10% of the item sets, later on moderate almost 60% of the sampling may obtain at minimum one query might be missing. These characteristics might be quite relevant to the analysis.

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 [3]. Missing Completely At Random (MCAR) point into several distinct data sets being removed are separate both of noticeable scalar and of unnoticeable argument. 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 the quantities or characters or symbols that is removed as a precise reasoning [4] [5].

A. Machine Learning Approach

In the data mining context, machine learning technique is generally classified as supervised and unsupervised learning technique both belong to machine learning technique [6]. 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 [7].

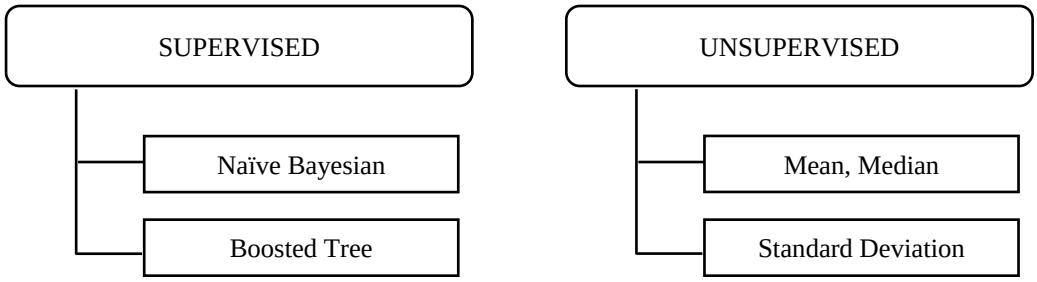


Fig.1. Structure of Machine Learning Approach

1) Supervised Machine Learning Approach: 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 calculate the scatter data concerning the mean value. It can be convenient in estimating the set of fact which can possess the identical aim but a different domain [8]. Estimate standard deviation based on sample and entire population data.

2) *Unsupervised Machine Learning Approach:* Another way of learning technique is classified as supervised learning that focus on the prediction based on known properties. Naïve Bayes technique [9] [10] is one of the most useful machine learning techniques based on computing probabilities. It analyses 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 [11] [12].

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.

II. Related Work

Little and Rubin [13] summarize the mechanism of imputation method. Also 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 in [14]. Classification of multiple imputation and experimental analysis are described in [14]. Min Pan et al. [15] summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Comparisons of different unsupervised machine learning technique are referred from survey paper [16]. To overcome the unsupervised problem Peng Liu, Lei Lei et al. applied the supervised machine learning techniques called Naïve Bayesian Classifier.

III. Analysis Of Multiple Imputation Method

The Multiple imputations 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 [17] 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.

A. Naïve Bayesian Classifier(NBC)

In Naïve Bayesian Classifier is one of the most useful machine learning techniques based on computing probabilities [19]. This classifier frequently executes especially strong and widely used because it continually execute further advanced classifying methods. 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. Figure 2 shows the structure of Naïve Bayesian Classifier approach.

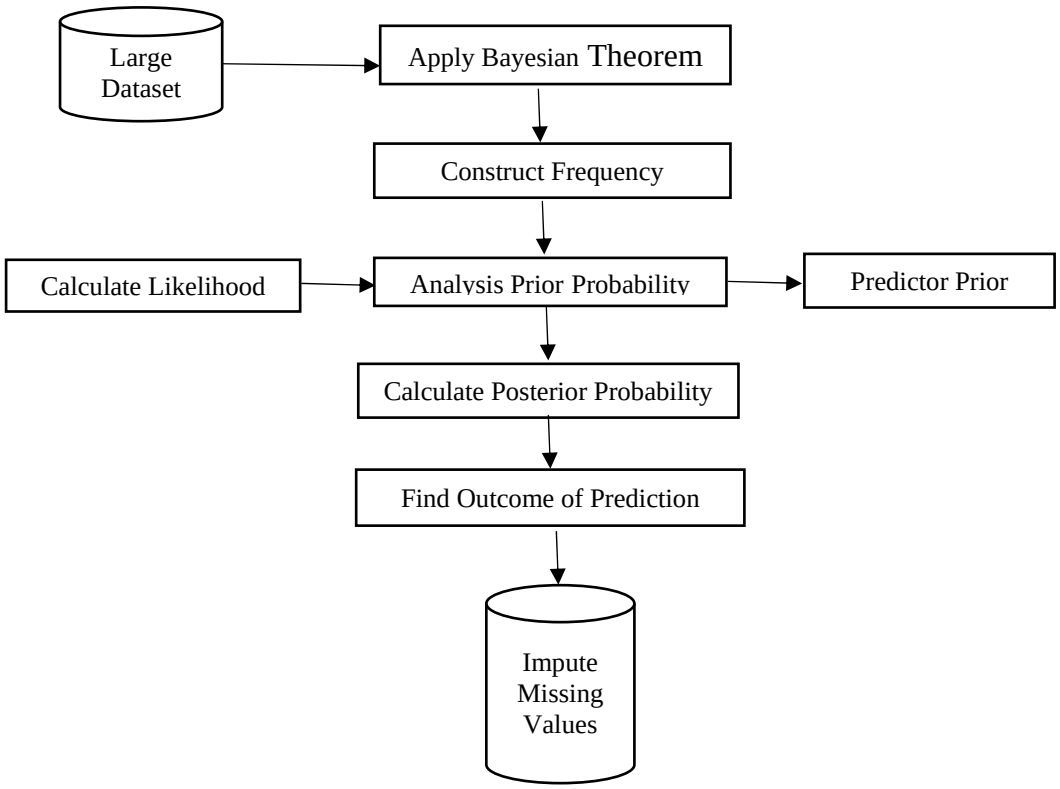


Fig.2. Process of Naïve Bayesian Classifier

B. Bolzano Weierstrass Theorem

The Bolzano Weierstrass theorem 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 $\mathbb{R}$ is compact if and only if it is closed and bounded. The set $S := Q \cap [0,1]$, since rational are countable, and treat $\mathbb{S}$ 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 $\mathbb{S}$. For example. $S_n := \frac{1}{n},\ n \in N$. $\mathbb{N}$ 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 $\mathbb{P}$, which is between the points $\mathbb{M}$ and $\mathbb{N}$. 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 $\mathbb{P}$ to be $\mathbb{N}$, 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 supremum. 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 supremum 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 everyε, 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.

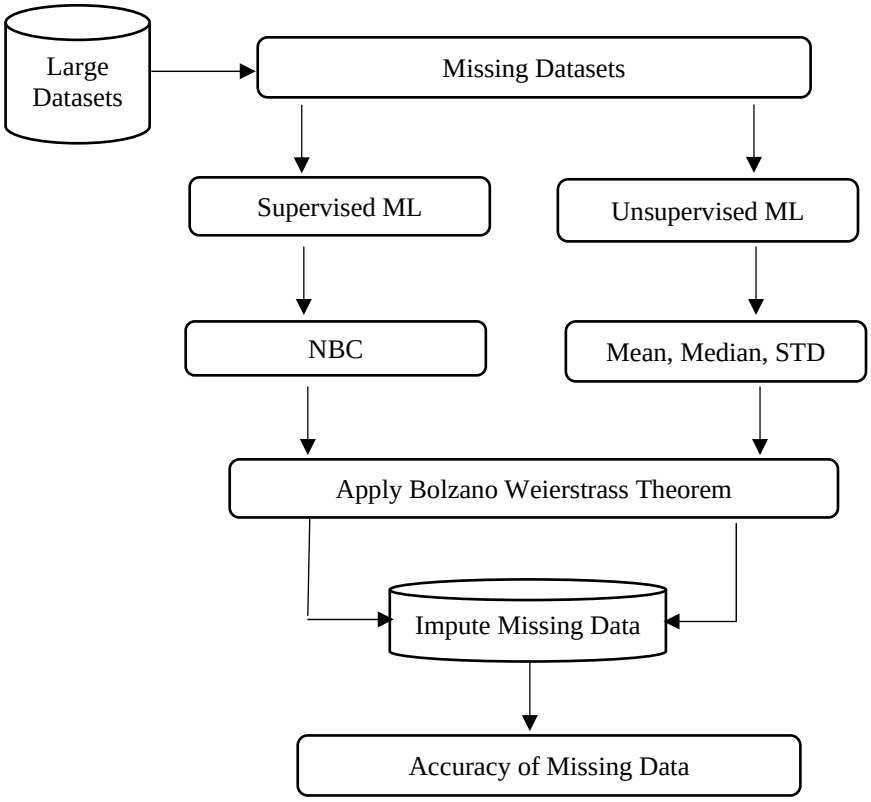


Fig.3. Bolzano Weierstrass Evaluation Process

C. Basic Idea

NBC technique is one of the widely used missing data treatment methods. The basic idea of NBC is first to define the attribute to be imputed, called imputation attribute and then, to construct NBC using imputation attribute as the class attribute. Other attribute in the dataset are used as the training subset. Hence the imputation problem is becoming a classification problem. Finally, the NBC is used to estimate and replace the missing data in imputation attribute. So this paper proposes a new method based on Naïve Bayesian Classifier to handle missing data.

Bayes theorem [20] afford a method of manipulating the rear probability $P(C / X)$ of category from $P(C)$ is the algorithmic probability of category, $P(X)$ is the algorithmic probability of rear and $P(X / C)$ is the likelihood of predictor for given category. Naïve Bayes classifier estimate that the outcome of the rate of a predictor (X) on a given category (C) is free from outside control of the point of other predictors called conditionally independent.

1) Algorithm for Posterior probability: Construct a frequency distributions for each credit across the destination. Transform frequency distribution to likelihood distribution. Certainly adopt the help of Naïve Bayesian equation to determine the posterior likelihood for every category.

2) Zero Frequency Problem: When a credit value doesn't exist with every category value increment 1 to the count for every aspect value category sequence.

3) Numerical Predictors: Arithmetic values need to be convert into their absolute analogue values since creating their frequency distribution. The classification with the greatest posterior likelihood is the result of the prediction [21].

IV. EXPERIMENTAL RESULTS

Experimental datasets were carried out from the Machine Learning Database UCI Repository. Table1. 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. The main objective of the experiments conducted in this work is to analyze the classification of machine learning algorithm. Datasets without missing values are taken and few values are removed from it randomly. The rates of the missing values removed are from 5% to 25%. In these experiments, missing values are artificially imputed in different rates in different attributes.

TABLE I. Dataset Used for Analysis

Dataset	Breast Tissue
Instances	106
Attributes	10(9 Features + 1 Class)
Missing Rates	5% to 25%
Unsupervised	Mean, Median, STD
Supervised	Naïve Bayesian

The following Figure 4 represents the classification of all attribute of original dataset using supervised machine learning techniques like NBI and unsupervised machine learning techniques like Mean, Median and STD without missing values.

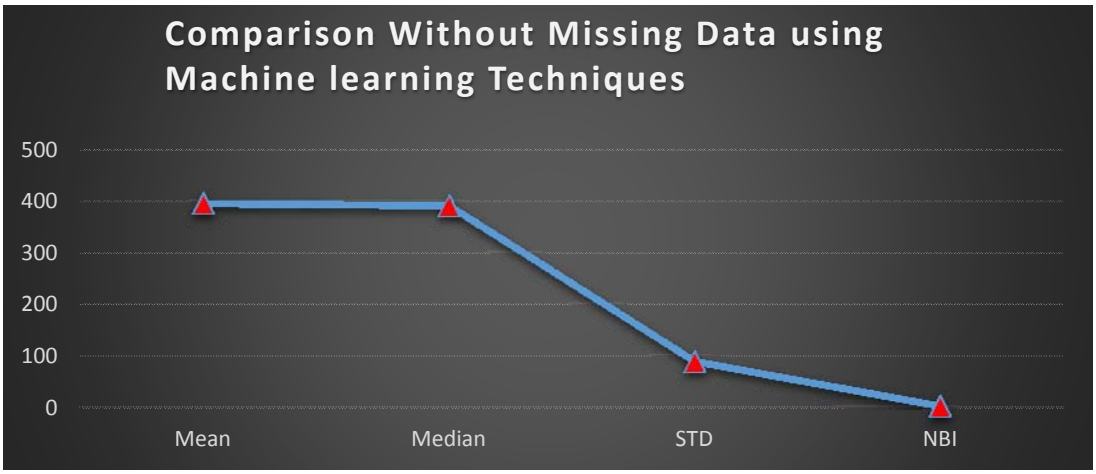


Fig.4. Original Dataset without Missing Values

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.

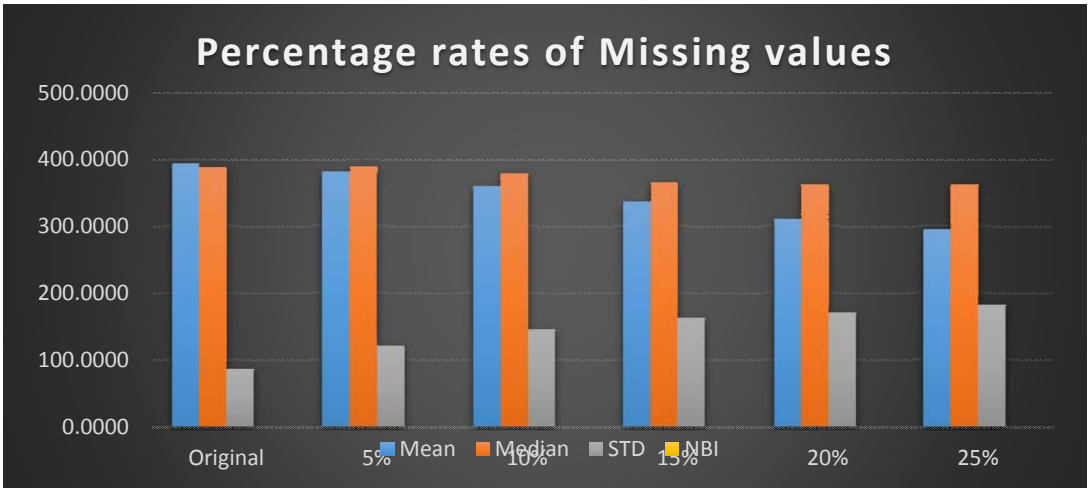


Fig 5. Percentage Rates of Missing Values

The following Figure 6 describes the different percentage rates of missing values for experimental analysis of unsupervised techniques like Mean, Median & STD with the missing rate of percentage.

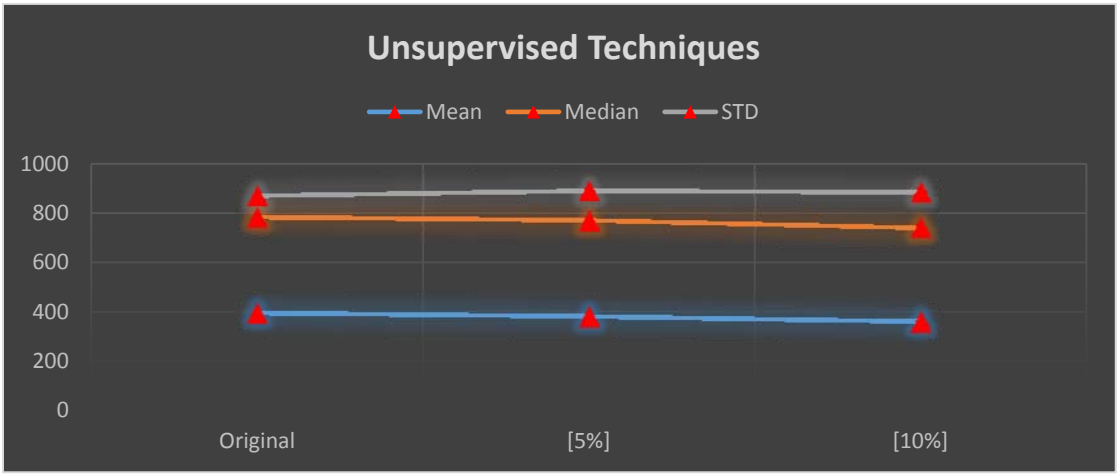


Fig 6. Percentage Rates of Unsupervised Method

Fig. 7 represent the experimental results of both supervised machine learning techniques like Naïve Bayesian Imputation using missing value with the rate of 5%, 10%, 15%, 20% & 25% respectively. The following figure 8 represents the comparison of both supervised NBI and unsupervised techniques Mean, Median and Standard Deviation using missing values for all the attributes contains different rate of percentage.

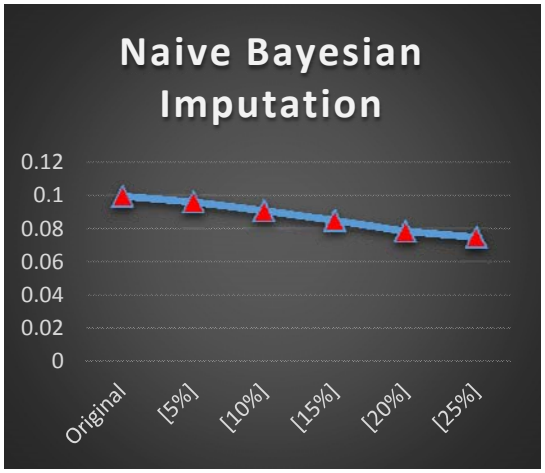


Fig 7. Experimental Results for Supervised Method

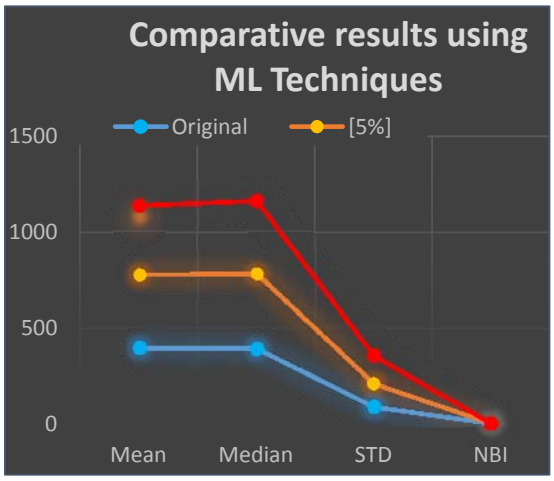
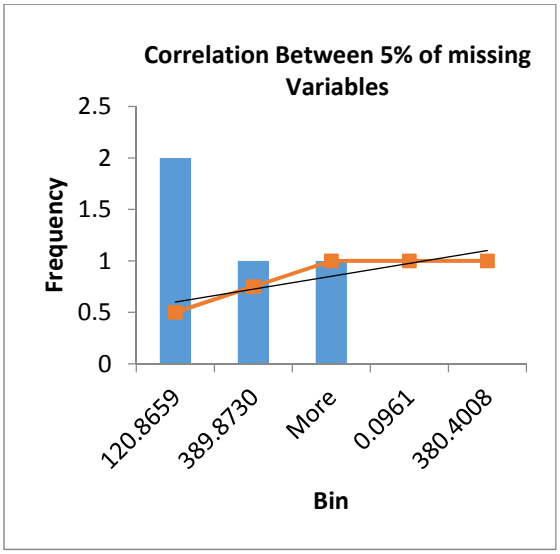


Fig 8. Comparative Results for both ML Methods

V. DISCUSSION

According to the previous discussion, Bayes imputation classifier consists of 2 process. Process 1. State the imputation of element and the imputation sequence. Process 2. Apply Bayes imputation classifier to assign missing values. For the sequence suitable approach, as stated above the imputation of element and the imputation sequence, Naïve Bayesian classifier assign the missing value in the first imputation element of the sequence and then assign the later on the altered new database. Bayes classifier construct classification model, however it can't be improved systemically also it can't automatically select suitable features like boosted tree as the performance of Bayes classifier lies on the rightness of the element selection in database. Since every imputation element, main facts of its function elements are determined. The most important drawbacks of Bayes classifier is that it has strong feature independence assumptions. Another one is if has no occurrences of a class label and a certain element value together then the frequency based probability estimate will be zero. According to conditional independence assumption, when all the probabilities are multiplied will get zero and this will affect the posterior probability estimate.



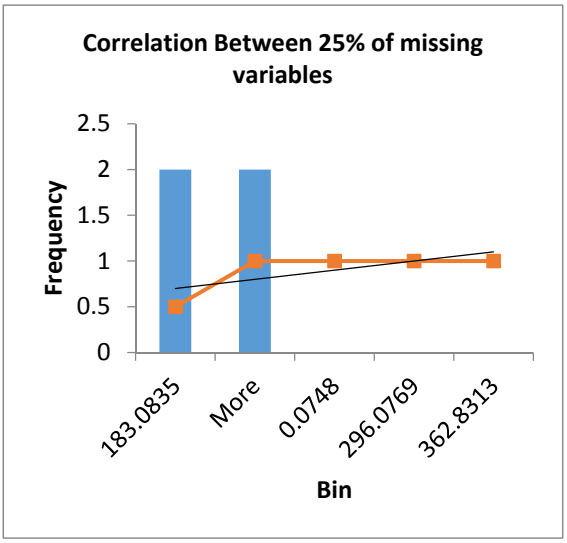


Fig 9 & Fig 10. Correlation between 5% and 25% of Missing Variables.

VI. CONCLUSION

In this paper, the proposed independence classifier has been implemented and evaluated. It gives the complete view about the multiple imputation of missing values in large dataset. Single imputation technique generates bias result and affects the quality of the performance. This paper focused multiple imputation using machine learning techniques of both supervised and unsupervised algorithms. The comparative study of mean, median, standard deviation in which standard deviation generates stable result in unsupervised algorithm. Also this paper shows the experimental result of standard deviation and Naïve Bayesian using limited parameter for their analysis and the performance evaluation stated, among the other missing value imputation techniques, the proposed method produce accurate result. In future it can be extended to handle categorical attributes and it can be replaced by other supervised machine learning techniques.

This paper presents an efficient and effective missing data handling method, Naïve Bayesian Classifier model. 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. According to the common imputation techniques, Bayes classifier is an effective missing data treatment model.

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 Science, 2910,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., Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.
[5] Ingunn Myrtveit, Erik Stensrud, “IEEE Transactions on Software Engineering”, Vol. 27, No 11, November 2001.
[6] 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.
[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, “Imputation of Missing Data in Industrial Databases”, Applied Intelligence, vol 11, pp., 259-275, 1999.
[9] K.Raja, G.Tholkappia Arasu, Chitra S.Nair, “Imputation Framework for missing value” International Journal of Computer Trends and Technology-Volume3 Issue2-2012.
[10] Lim Eng Aik and Zarita Zainuddin, “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., A Quantitative Study of the Effect of Missing Data in Classifiers, proceedings of CIT2005 by IEEE Computer Society Press, September 21-23,2005.
[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. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.
[14] R. Kavitha Kumar and Dr. R. M. Chandrasekar, “Missing Data Imputation in Cardiac data set”.
[15] 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.
[16] 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.

[17] Shichao Zhang, Xindong Wu, Manlong Zhu, "Efficient Missing Data Imputation for Supervised Learning" Proc, 9th IEEE conference on Cognitive informatics, 2010 IEEE.
[18] S.Hichao Zhang, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin, Chengqi Zhang, "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg,2008.
[19] 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.
[20] 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.
[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 pages 128-132.

AUTHOR PROFILE

Smt. S. Kanchana is a research scholar of the Department of Computer Science, NGM College under Bharathiyar University, Coimbatore. She had seven years of experience in Teaching and one year of experience in software Industry. She has published and presented many papers in International/National Journals. Her areas of interest include Data Mining, Cloud Computing and Artificial Intelligent. She is a life member of Indian Science Congress Association, India.

Dr Antony Selvadoss Thanamani is currently working as Professor and Head, Department of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore) and the Principal Investigator of UGC – MAJOR Research Project in Computer Science. He has published many papers in international/national journals and written many books. His areas of interest include E-Learning, Software Engineering, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 26 years of teaching and research experience. His current research interests include Grid Computing, Cloud Computing, Semantic Web. He is a life member of Computer Society of India, Life member of Indian Society for Technical Education, Life member of Indian Science Congress, Life member of Computer Science, Teachers Association, New York and Member of Computer Science, Teachers Association, India.