ISSN 2319 – 1953
International Journal of Scientific Research in Computer Science Applications and Management Studies

Missing Data Imputation Using Bayesian Classifier

A.Linda Sherin[1] Dr.Niraimathi
[1]*Research Scholar, NGM College, Pollachi, Tamilnadu*
[2]*Assistant Professor of Computer Applications, NGM College, Pollachi, Tamilnadu*
[1]a.selvadoss@gmail.com

Abstract - Multiple imputations provide a useful strategy for dealing with data sets with missing values. Multiple imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in valid statistical inferences that properly reflect the uncertainty due to missing values. This paper reviews methods for analyzing missing data, including basic concepts and applications of multiple imputation techniques. The paper also presents new SAS R procedures for creating multiple imputations for incomplete multivariate data and for analyzing results from multiply imputed data sets. We have approached the data completion problem using two well-known supervised machine learning techniques – Booster Algorithm. The primary objective is to highlight the features on selection of suitable data imputation algorithms and also implementing Darboux's theorem in machine learning techniques to evaluate the performance of every sequence of rational and irrational number has a monotonic subsequence. To generate the inference levels of imputation of missing data, the standard UCI repository dataset is deployed. Experimental results reveal a significant improvement of accuracy in the proposed approach.

Keyword - Imputation Algorithm, Naive Bayesian Classifier, Supervised Machine Learning, Unsupervised Machine Learning.

I INTRODUCTION

Missing data imputation 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 data are simply observations that we intended to be made. 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. A simple technique for handling with missing 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 is smaller to attain momentous outcome in the study. 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. 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.

Machine Learning Approach

In machine learning, such solutions are called target or output and situations are called input or *unleveled data*. Situation and solution in combination it is called *leveled data*.

Supervised: So, if you are training your machine learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training. Your learning algorithm seeks a function from inputs to the respective targets. If the targets are expressed in some classes, it is called classification problem. Alternatively, if the target space is continuous, it is called regression problem.

Unsupervised: Contrary, if you are training your machine learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs.

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

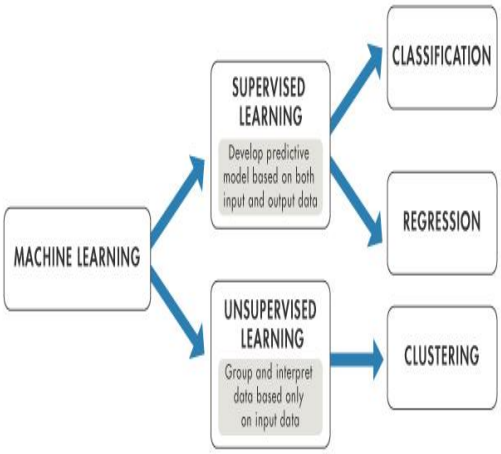

Figure 1 Structure of Machine Learning

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 calculates 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. 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 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.

II Literature Review

The missing-data mechanism has three classifications (Rubin, 1976): missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR). Data are said to be missing at random (MAR) if other variables in the dataset can be used to predict missingness on a given variable. For example, in surveys, men may be more likely than women to refuse to answer some questions. Here, data will be missing completely at random (MCAR) because the process that causes missingness does not depend on the values of variables in the dataset subject to analysis (Little, 1988; Rubin, 1976; Zhang, 2003). MCAR is a fairly strong assumption, and tends to be relatively rare. For instance, in the context of survey data, MCAR data might occur when a respondent simply skips an item or a question, perhaps because of neglecting to turn the page of a questionnaire booklet. MAR is a less restrictive assumption than MCAR. Finally, data are said to be missing not at random (i.e., MNAR, also called nonignorable missing data) if the value of the unobserved variable itself predicts missingness. A classic example of this is income. Individuals with very high incomes generally refuse to answer questions about their earnings. This is not the case for individuals with more modest incomes. Miroslaw Pawlak computed a number of nonparametric kernel classification rules, consistency and speed of convergence of kernel classification rules established from missing data. Prediction density approach, the deletion techniques, and stochastic mechanism of mechanism of generation of missing values can be introduced by imposing probability distribution on the variables. ShenQiping et al. described problem solving processes in value management (VM) workshop in the construction industry are experience-based, and the quality workshops depends on the experience of the team members. Experimental results show that DM techniques can help team members in VM workshops to understand the problems more clearly and to generate more information for recent issues in feasibility studies, risk analyses, resource allocation, site layout, and time/cost predictions. Jort Florent Gemmeke et al. focused effective way to increase the noise robustness of automatic speech recognition to label noisy speech features either reliable or unreliable and to replace the missing data by clean speech estimates. Little and Rubin 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. Educational researchers have become increasingly aware of the problems and biases which caused by missing data. Multiple imputations is not implemented by many researchers who could benefit from it, very possibly because of lack of familiarity with the technique. Therefore, the main objective of this author, to help familiarize researchers with the basic process of multiple imputations. Classification of multiple imputation and experimental analysis are described by Min Pan et al., summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. To overcome the unsupervised problem Peng Liu, Liu Lei et al. applied the supervised machine learning techniques called Naïve Bayesian Classifier. Yuri Pirola et al, introduced minimum-recombinant haplotype configuration problem (MRHC) provide highly successful in sound

combinatorial formulation for genotype phasing on pedigrees. An experimental analysis demonstrated the biological soundness of the phasing model and the effectiveness of the algorithm under several contexts.

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 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. 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.

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 ϵ (M, N).Now we take an arbitrarily small ε. Observe the segment [M, R+ε]. R+ε cannot belong to S since it is higher than the supremum. Hence [M, R+ε] contains an infinite number of (R_n) members. Now the segment [M, R-ε]. R-ε must belong to S, since it is smaller than the supremum of the segment S. Thus [M, R-ε] contains a finite number of members from (R_n). But [M, R-ε] is a subset of [M, R+ε]. 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-ε, R+ε) 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 in R^n has a convergent subsequence

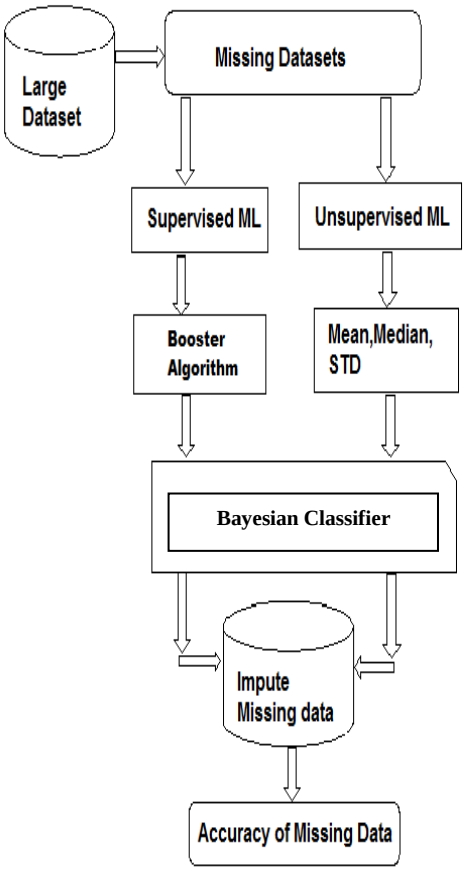

Fig.2. Enhanced accuracy of Missing Data using Bayesian Classifier.

Concept used in this research

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. In addition to NBC is used to fix the infimum and supremum in the data sequence. Hence the imputation problem is becoming a problem of classified data sequence. Finally, the NBC along with the 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 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 estimates 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 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 converting into their absolute analogue values since creating their frequency distribution. The classification with the greatest posterior likelihood is the result of the prediction.

IV Experimental Results

Experimental datasets were carried out from the University Data set of the UCI Repository. Table1 describes the dataset with Multivariate Data Characteristics with Categorical Integer Attributes which contains 285 number of instances and 17 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.

The following Figure 4 represents the classification of missing value Imputation of original dataset using supervised machine learning techniques like Naïve Bayesan, Booster Algorithm, NBC and unsupervised machine learning techniques like Mean, Median and STD.

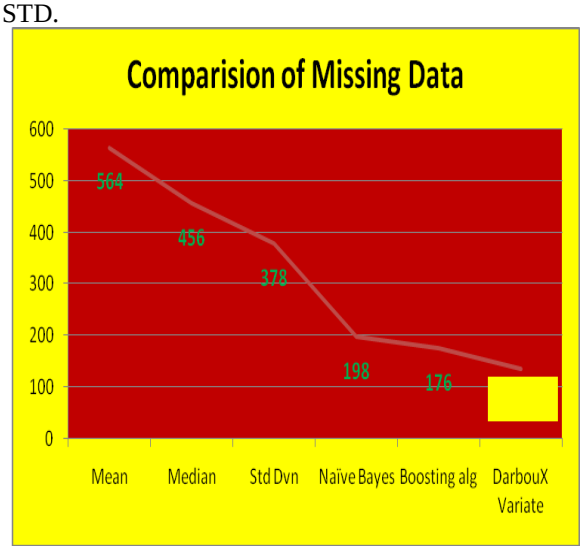

Fig.3. Missing Values imputation in Original Dataset

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. It also represents the comparison of both supervised techniques - NBC, Boosting Algorithm, NBC- and unsupervised techniques - Mean, Median and Standard Deviation using missing values for all the attributes contains different rate of percentage.

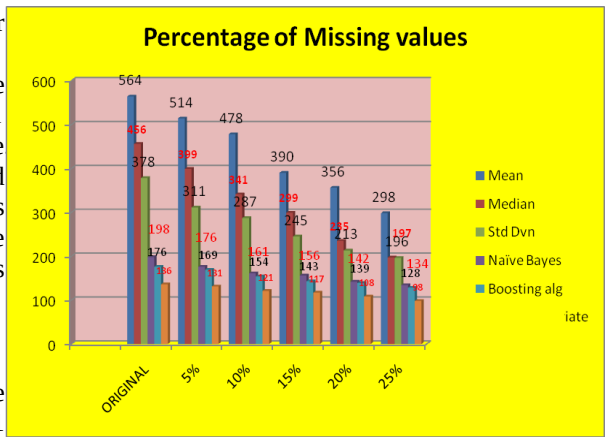

Fig 4. Percentage Rates of Missing Values

Table 1: Dataset Used for Analysis

Data Set Characteristics:	Multivariate	**Number of Instances:**	285
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	17
Associated Tasks:	Classification	**Missing Values?**	Yes

Discussion

According to the previous discussion, Naives Bayesan imputation classifier consists of 2 process. Process 1. State the imputation of element and the imputation sequence. Process 2. Apply NBC to assign missing values. 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. It helps in construction of the classification model with infimum and supremum bounds, however it can't be improved systemically also it can't automatically select suitable features like boosted tree as the performance 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 assumption. Another one is it 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. Thus this drawback is overcome with applying in NBC to fix the infimum and supremum of the data sequence.

V 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, NBC, Booster Algorithm and - NBC in which standard deviation generates stable result in unsupervised algorithm. Also this paper shows the experimental result of standard deviation and - NBC 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.

REFERENCES

[1]. Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, "Experimental Analysis of Methods for Imputation of Missing Values in Databases.
[2]. Apostol, Tom M.: Mathematical Analysis: A Modern Approach to Advanced Calculus, 2nd edition, Addison-Wesley Longman, Inc. (1974), page 112.
[3]. 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.
[4]. Bruckner, Andrew M: Differentiation of real functions, 2 ed, page 6, American Mathematical Society, 1994
[5]. Ciesielski, Krzysztof (1997). Set theory for the working mathematician. London Mathematical Society Student Texts. **39**. Cambridge: Cambridge University Press. pp. 106–111. ISBN 0-521-59441-3. Zbl 0938.03067.
[6]. Damian Dechev, Pierre Laborde, and Steven D. Feldman, "LC-DC: Lockless Containers and Data Concurrency A Novel Nonblocking Container Library for Multicore Applications" IEEE Access Practical Innovations: Open Solutions Vol. 1, 2013
[7]. 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.
[8]. Enders CK. *Applied Missing Data Analysis (Methodology in the Social Sciences)* ISBN-13: 978-1606236390. Available from: *https://www.amazon.com/Applied-Missing-Analysis-Methodology*
[9]. Han J. and Kamber M., Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.
[10]. HTTP Link: *https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Darboux.html*
[11]. Ingunn Myrtveit, Erik Stensrud, "IEEE Transactions on Software Engineering", Vol. 27, No 11, November 2001.
[12]. 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.
[13]. Jort Florent Gemmeke, Hugo Van Hamme, Bert Cranen, and Lou Boves "Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition" IEEE Journal of Selected Topics in Signal Processing, Vol. 4, No. 2, April 2010
[14]. K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases", Applied Intelligence, vol 11, pp., 259-275, 1999.
[15]. K.Raja, G.Tholkappia Arasu, Chitra S.Nair, "Imputation Framework for missing value" International Journal of Computer Trends and Technology-Volume3 Issue2-2012.
[16]. Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad, "Imputation of Missing Data Using Machine Learning Techinques", from KDD-96 Proceedings.
[17]. Lim Eng Aik and Zarita Zainuddin, "A Comparative Study of Missing Value Estimation Methods: Which Method Performs Better?" 2008 International Conference on Electronic Design.
[18]. Link(most recent): /ipns/QmdJiuMWp2FxyaerfLrtdLF6Nr1EWpL7dPAxA9oKSPYYgV/wiki/Darboux's_theorem_(analysis).html
[19]. 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 21-23,2015.
[20]. Min Pan "Based on Kernel Function and Non-Parametric Multiple Imputation Algorithm to Solve the Problem of Missing Data" IEEE MSIE 2014
[21]. MiroslawPawlak "Kernel Classification Rules from Missing Data" IEEE Transactions on Information Theory, Vol. 39, No.3, May 2003
[22]. Olsen, Lars: A New Proof of Darboux's Theorem, Vol. 111, No. 8 (Oct., 2014) (pp. 713–715), The American Mathematical Monthly