International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 1 (2015) pp. 1473-1482
© Research India Publications
http://www.ripublication.com

Multiple Imputation Of Missing Data Using Efficient Machine Learning Approach

S. Kanchana

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

Dr. Antony Selvadoss Thanamani

Professor and Head, Research Department of Computer Science,
NGM College, Pollachi 642001
Bharathiyar University, Coimbatore, India
selvdoss@gmail.com

Abstract

Missing values are an unavoidable problem in dealing with most of the real world data sources, and the missing values in a dataset may affect the quality of the supervised learning process. This paper focuses machine learning techniques Naïve Bayesian Classifier method for the imputation of missing values in large datasets. Single value imputation produce biased results, whereas multiple imputation generates right value to replace. The study was conducted to implement several algorithms in unsupervised techniques like Mean, Median, Standard Deviation, and Regression and supervised machine learning techniques like Naïve Bayesian classifier provides different methods to calculate Class prior probability, Posterior probability, Predictor probability and Maximum Likelihood. Both methods are trying to find hidden structure in unlabeled data. 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. To evaluate the performance, the standard machine learning repository dataset has been used. We experimentally shows that our approach significantly outperforms some standard machine learning methods for handling missing values in classification tasks.

Keywords—Correlation, Supervised, Maximum Likelihood, Naïve Bayesian, Unsupervised, Regression.

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 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 Techniques

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.

Unsupervised Machine Learning Techniques

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)]

Supervised Machine Learning Techniques

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.

The Method of Multiple Imputation

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.

Naïve Bayesian Classifier(NBC)

Naïve Bayesian Classifier 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).

Naïve Bayeian Classifier Algorithm

The Naïve Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. This model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets..

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.

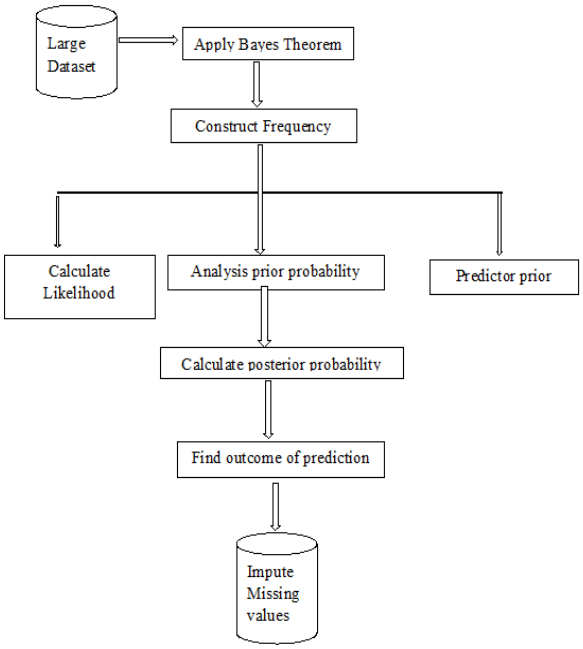

Figure 1. Flowchart of the Proposed System

Algorithm for posterior probability

- Construct a frequency table for each attribute against the target.
- Transform frequency table to likelihood tables
- Finally use the Naïve Bayesian equation to calculate the posterior probability for each class.
- The class with the highest posterior probability is the outcome of prediction.

Zero-Frequency Problem

When an attribute value doesn't occur with every class value add 1 to the count for every attribute value class combination.

Numerical Predictors

Numerical variables need to be transformed to their categorical counterparts before constructing their frequency tables.

Experimental Results

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%.

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

Table 1. Datasets used for Analysis

Experimental Evaluation

The below Figure 2. Represent the single instance of original datasets without any missing values and it is implemented using Mean, Median Standard Deviation and Naïve Bayesian.

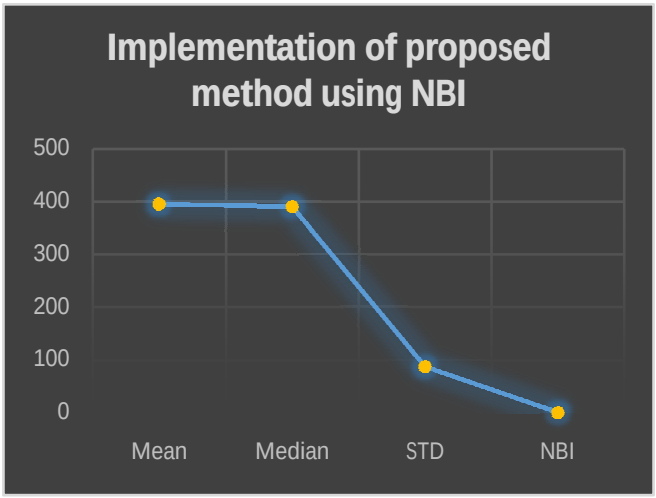

Figure 2. Single instance of original datasets

Figure 3 represent the experimental results of both supervised and unsupervised machine learning techniques using missing value with the rate of 5%, 10%, 15%, 20% & 25% respectively.

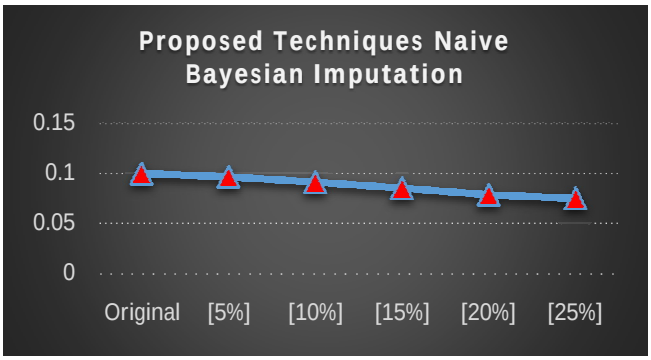

Figure 3. Experimental results for Supervised Techniques

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.

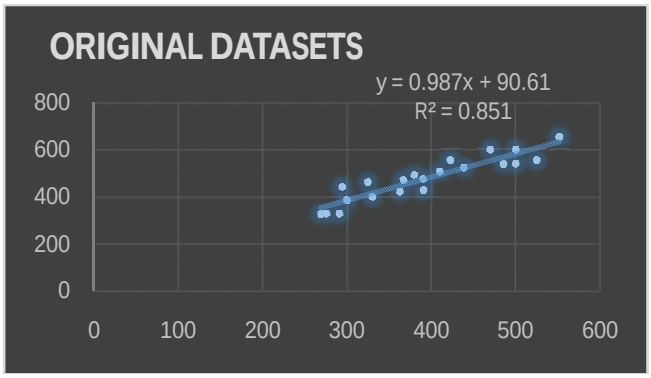

Figure 4. Correlation chart of original dataset

Figure 5. represent the imputation of missing values using Standard Deviation techniques and analysis the correlation value R^2= 0.0592 which indicates no correlation between variables.

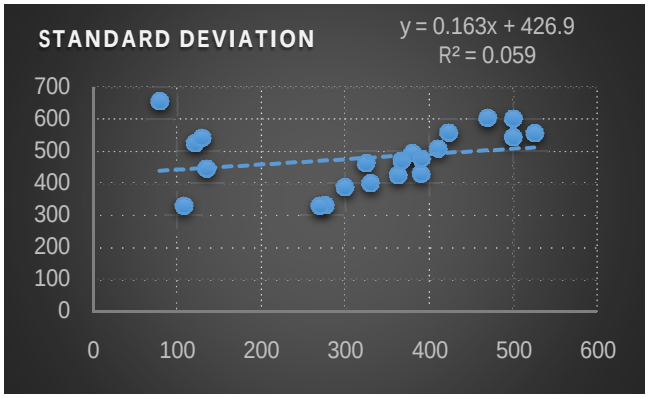

Figure 5. Correlation chart of Standard Deviation Method

Figure 6 represent the correlation value of Regression method R^2=0.8886 which means high positive correlation between variables.

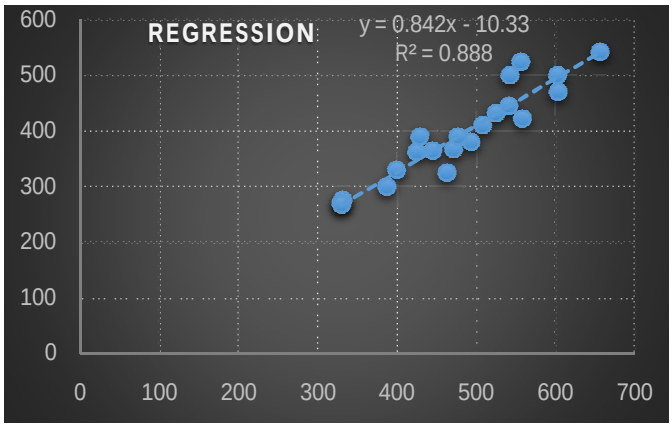

Figure 6. Correlation chart using Regression Method

Figure 7 specify the imputation of missing data using NB Techniques and the correlation value R^2=0.027 which means no correlation between variables.

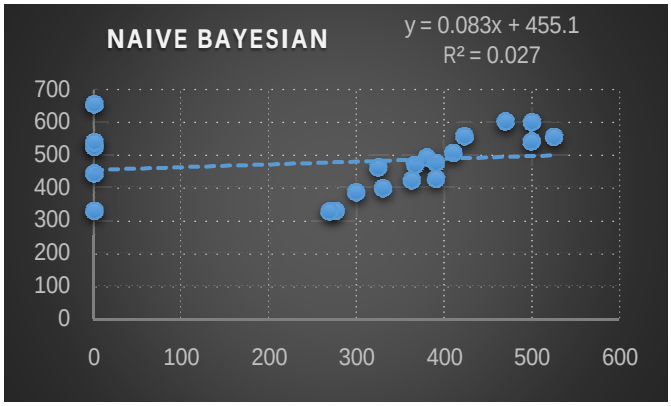

Figure 7. Correlation value of Naïve Bayesian Techniques.

Figure 8 indicates the correlation chart of different ML methods and specify Regression method gives high positive correlation value from other method.

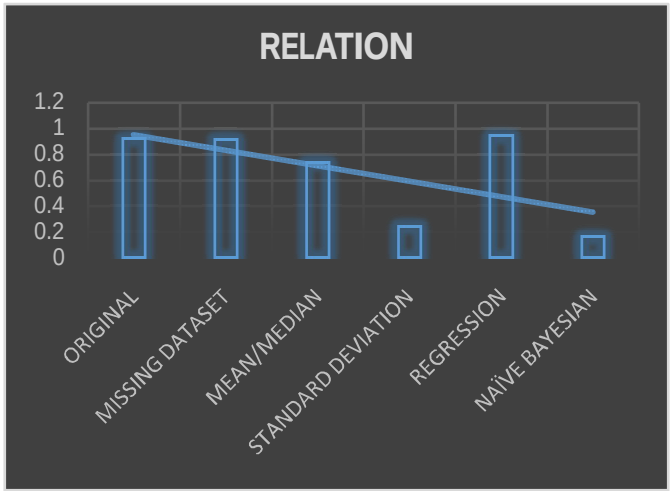

Figure 8. Correlation chart of Machine Learning Techniques.

Conclusions

This paper gives the complete view about the multiple imputation of missing values in large dataset. Single imputation technique generate bias result and affect the quality of the performance. This article proposed multiple imputation using machine learning techniques of both supervised and unsupervised algorithms and also analysis the experimental results of correlation between variables and different machine learning Techniques. The comparative study of unsupervised algorithms like mean, median, standard deviation etc., in which standard deviation generates stable results. The comparative result of correlation value shows that among the other techniques Regression method evaluate high positive correlation value. In future it can be extended to handle categorical attributes.

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.