ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

Multiple Imputation Of Missing Data Using Naïve Bayesian Classifier

Priya.S, Dr.Antony Selvadoss Thanamani

Research Scholar, Bharathiar University and Assistant Professor, Department of Computer Science,

Government First Grade, College, KGF, India

Professor and Head, Department of Computer Science, NGM College, Pollachi, Coimbatore, Tamilnadu, India

ABSTRACT: In a large database missing data and inconsistent data are pervasive and lasting problem .Missing data occur in almost all serious statistical analysis. In statistics, imputation is the process of replacing missing data with substituted values. Traditional and modern methods are there for solving this problem. Multiple imputation generates the right value to replace. Variety of machine learning techniques is developed to reprocess the incomplete information. This paper evaluates multiple imputation of missing data in large datasets by using a supervised machine learning technique for probabilistic algorithms like naïve Bayesian classifier, which is very sensitive and provides a good accuracy to handle missing data

KEYWORDS*:* Missing data, Accuracy, Imputation, naive Bayes classifier.

I. INTRODUCTION

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouse, or other information repositories [1]. The main goal of the data mining process is to discover knowledge from large database and transform into a human understandable format. There are a lot of functions of DM, such as description, association analysis, classification and prediction, clustering analysis etc. Among all, classification and prediction are widely used in many fields. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc. All these serious data quality problems affect the performance of DM algorithms [2].

Missing data imputation is a real and challenging issue confronted by machine learning and information mining. Most of the real world datasets are characterized by an unavoidable problem of incompleteness, in terms of dropping values. Missing values may generate bias and affect the caliber of the supervised learning procedure. Missing value imputation is an efficient means to detect or estimate the missing values based on other data in the data sets. Data mining consists of the various technical approaches including machine learning, statistic and database system.. 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 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.

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) 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.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

In the rest of this paper gives a brief explanation of classification and prediction in section II, Introduces new method based on Naïve Bayesian Classifier to estimate and replace missing data. Experimental analyses of NBI model in Section III and the Conclusions are discussed in Section IV.

II. CLASSIFICATION AND PREDICTION

A. *Classifier*

Classification means constructing a classifying function or model from the known data. Such function or model can also be called 'classifier', which can classify the records in database into given classes, thus can predict the unknown variables under some given conditions [3]. Classifiers differ greatly in prediction accuracy, training time and number of leaves (Decision Trees). There is not a classifier which performs best in all aspects [4]. Prediction accuracy of classifiers can be affected by the factors as follows [3].

Number of records in training subset. Classifier needs to learn from training set. Therefore, larger training set makes the classifier more reliable. But the training time is also become longer.

Data quality. Problems such as noise data, missing data, data inconsistency etc. bring a lot of wrong information which will lead to wrong classify. It is impossible to build a convictive classifier with incompleteness or wrong data.

Attribute quality. Attributes provide information for classifying. The prediction accuracy can be improved by including more attributes. However, more attributes means calculating more attribute combinations and more training time. It is essential to choose attributes which are valuable for classification.

Characteristics of the records to be predicted. If Characteristics of the records to be predicted are different from records in training set, it may lead to high incorrect rate.

B. *Missing Treatment Methods Using Classifier*

In the popular methods for missing data handling, using classifier to predict and fill in missing data is a set of fast developing methods.

Naïve Bayesian Classifier (NBC)

Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian Classifier to impute the missing data. It is consisted of two phases:

a) Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.;

b) Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data (usually 3~4 attributes) and the times for iterative can be reduced [7]

This method is effortless to construct and no complex iterative argument estimation, that forms the specific functional for extremely big datasets. This classifier frequently executes especially strong and widely used because it continually execute further advanced classifying methods[11]. Figure 1 shows the structure of Naïve Bayesian Classifier approach.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

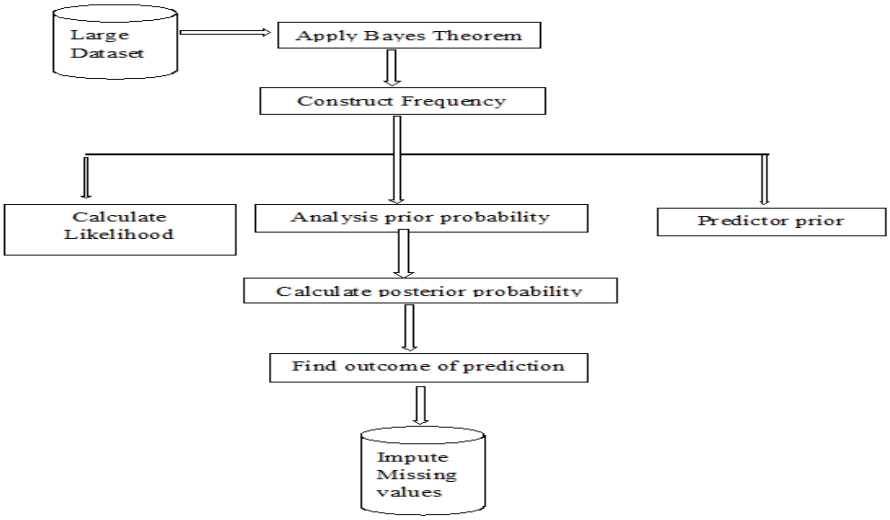

Fig 1. Framework of NB Classifier

III. EXPERIMENT AND ANALYSIS

A. *Sensitivity Analysis*

Sensitivity analysis (SA) is to study the impacts of one or more input variables on the outputs of a model, that is, the sensitivity of the model to one parameter or a combination of parameters [8]. If a tiny change of an input leads to great changes of the output, the model is highly sensitive to that input. SA can help to identify the decisive input parameter of the model [8]. In our experiments, the proportion of missing data in the datasets is the parameter which affects the results of the classification models. The effect of missing data on the prediction accuracy will be investigated through the tiny changing of the missing rate.

IV .EXPERIMENT RESULTS

Design

Experimental datasets were carried out from the Machine Learning Database UCI Repository. It 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.

Experimental evaluation

Experimental evaluation provides the complete structure of all the attributes and classes without any missing values. Below datasets describe the electrical measurements in samples of freshly excised tissue from the breast also includes 106 instances, 10 attributes of 9 features and 1 class attribute. Impedance measurements were made at the frequencies 15.625, 31.25, 62.5 etc. KHZ.

The following Figure 2 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. Figure 3 describes the single instance of Breast tissue dataset without missing values.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

The below figure 4 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 specifies the different percentage rates of missing values for experimental analysis of unsupervised techniques like Mean, Median and STD with the missing rate of different percentage.

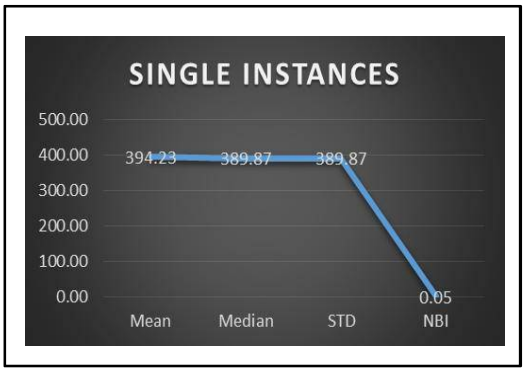

Fig 2. Original Datasets without Missing value values

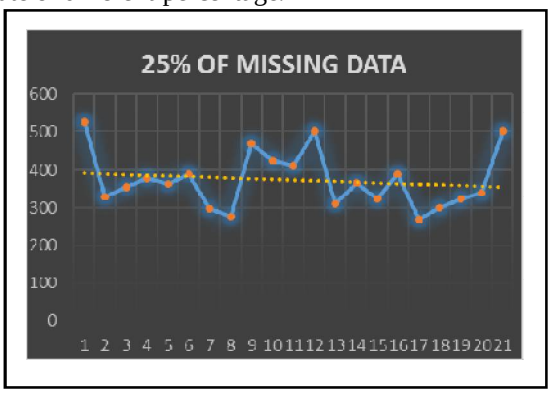

Fig 3. Original Datasets with Missing

Figure 6 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. Figure 7 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.

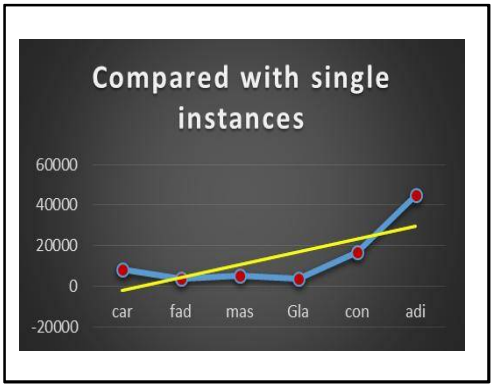

Fig 4 Missing value rates for Experimental Analysis

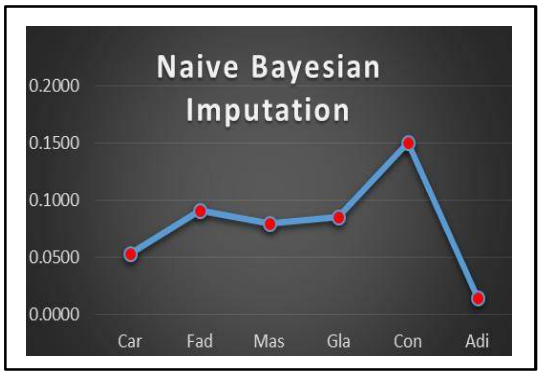

Fig 5. Experimental Results for Supervised Technique

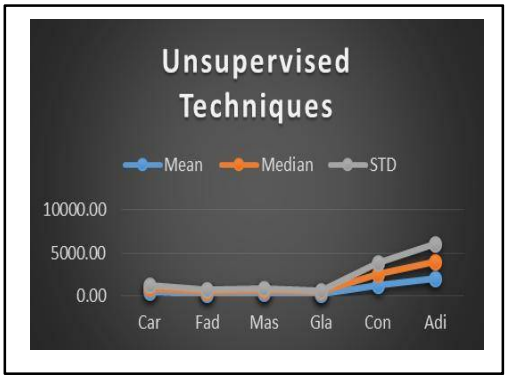

Fig 6. Experimental Results for Mean, Median and STD

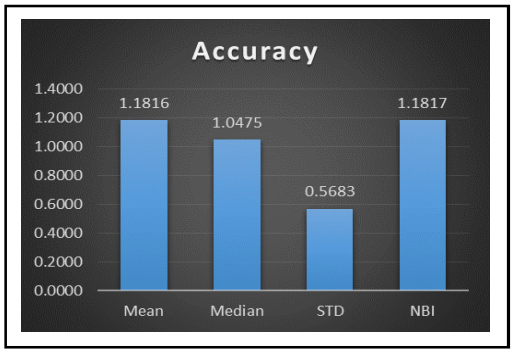

Fig 7. Comparative Results using Both ML Techniques

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

IV. CONCLUSION

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The sensitivity of classifiers to missing data is analyzed . The results showed that, with the increasing of the missing rate, the classification accuracies of all the classification algorithms have an obvious trend of decrease. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers. The evaluation results indicate that NBC is superior to multiple imputation. The performance of NBC is improved by the attribute selection. When the imputation attribute has been specified, the decree of the irrelevant master plan is commended. Granting to the common imputation techniques, Bayes classifier is an effective missing data treatment model.

REFERENCES

[1] Han J., Kamber M. Data Mining Concepts and Technique. Morgan Kaufmann Publishers, 2000
[2] Cios K.J., Kurgan L. Trends in Data Mining and Knowledge Discovery. In N.R. Pal, L.C. Jain
[3] Tian Jinlan, Li Ben. Tools for Data Mining: Classifiers. Department of Computer Science, Tsinghua University. Computer World, 1999, 20th Periodical
[4] Tjen Sienlim, Wei Yinloh, Yu ShanShih. A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and NewClassification Algorithms. Machine Learning, 40, 2000, 203~229, Kluwer Academic Publishers, Boston
[5] Marvin L., John F. Kros. Chapter VII-The Impact of Missing Data on Data Mining, Data Mining: Opportunities and Challenges, Idea Group Publishing, 2003
[6] Quinlan J. R., C4.5 Programs for Machine Learning. Morgan Kaufmann, CA, 1988.
[7] Liu P., Lei L. and Zhang X.F., A Comparison Study of Missing Value Processing Methods, Computer Science, 31(10):155-156 & 174, 2004.
[8] J.T. Yao. Sensitivity Analysis for Data Mining. Proceedings of The 22nd International Conference of NAFIPS, July 24-26, Chicago, U SA, 2003, 272~277
[9] Peng Liu, Elia El-Darzi et al. Comparative analysis of Data Mining Algorithms for Predicting Inpatient Length of Stay. Proceedings of the Eighth Pacific-Asia Conference on Information Systems July 2004
[10] Peng Liu, Lei Lei, Naijun Wu, A Quantitative Study of the Effect of Missing Data in Classifiers.
[11] S.kanchana, Dr. Antony Selvadoss Thanamani et al. A Magnified Application of Deficient Data Using Bolzano Classifier. Invention Journal of Research Technology in Engineering & Management 2455-3689 Volume 1 Issue 4 ǁ June. 2016 ǁ PP 32-37
[12] R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.
[13] R. Kavitha Kumar and Dr. R. M. Chandrasekar, "Missing Data Imputation in Cardiac data set".
[14] 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,
[15] 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.
[16] Shichao Zhang, Xindong Wu, Manlong Zhu, "Efficient Missing Data Imputation for Supervised Learning" Proc, 9 th IEEE conference on Cognitive informatics, 2010 IEEE.
[17] S.Hichao Zhang, Jilian Zhang, Xiaofeng Zhu, Yongsong Qin, Chengqi Zhang, "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg,2008.

BIOGRAPHY

Priya.S is a Research scholar of Bharathiar University, Coimbatore, and working as an Assistant Professor epartment of Computer Science, Government First grade College ,Kolar Gold Fields, Karnataka. She received Master of Computer Application (MCA) degree in 2001 from Bharathidasan university, Trichy ,TN, India. And M.Phil degree in 2006 from Periyar University ,Salem, TN, India. Her research interests are Data Mining and Cloud computing