Experimental Analysis of Imputation of Missing Data Using Machine Learning Techniques

S. Kanchana
Research Scholar, Research Department of
Computer Science, NGM College, Pollachi 642001
Bharathiyar University, Coimbatore, India

Dr. Antony Selvadoss Thanamani
Professor and Head, Research Department of
Computer Science, NGM College, Pollachi 642001
Bharathiyar University, Coimbatore, India

Abstract—**Missing data arise a common problem for data quality. Most important issue faced by researchers and practitioners who use industrial and research databases is incompleteness of data, usually in terms of missing or erroneous values. Some of the data analysis algorithms can work with incomplete data, a large portion of work require complete data. Therefore, variety of machine learning (ML) techniques are developed to reprocess the incomplete data. This paper concentrates on different imputation techniques and also proposes supervised and unsupervised machine learning techniques Naïve Bayesian imputation method in MI model. The analysis is carried out using a comprehensive range of databases, for which missing values were introduced randomly. The goal of this paper is to provide general guidelines on selection of suitable data imputation algorithms based on characteristics of the data. Experimental analysis on real datasets are taken from the Machine Learning repository and the results are compared in terms of accuracy.**

Index terms -Bayesian classifier, MI model, ML techniques, Supervised ML, Unsupervised ML.

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 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.

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) 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 literature review 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 [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. Comparisons of different unsupervised machine learning technique 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.

III. 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 [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.

A. Unsupervised Machine Learning Techniques

Mean imputationis the process of replacing the missing data from the available data where the instance with missing

attribute belongs.

Median Imputationis 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 Deviationmeasures 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.

B. 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 techniques 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.

IV. EVALUATION OF MULTIPLE IMPUTATION METHOD

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 [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.

A. Naïve Bayesian Classifier (NBC)

Naïve Bayesian Classifier is one of the most useful machine learning techniques 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. NBC is a popular classifier, notonly for its good performance, simple form and high calculation speed, but also for its insensitivity to missing data. It can build models on dataset with any amount of missing data. 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).

B. Naïve Bayesian Imputation (NBI)

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. Naïve Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticate classification methods.

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 assumes 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. Figure 1 shows the pictorial representation of proposed system.

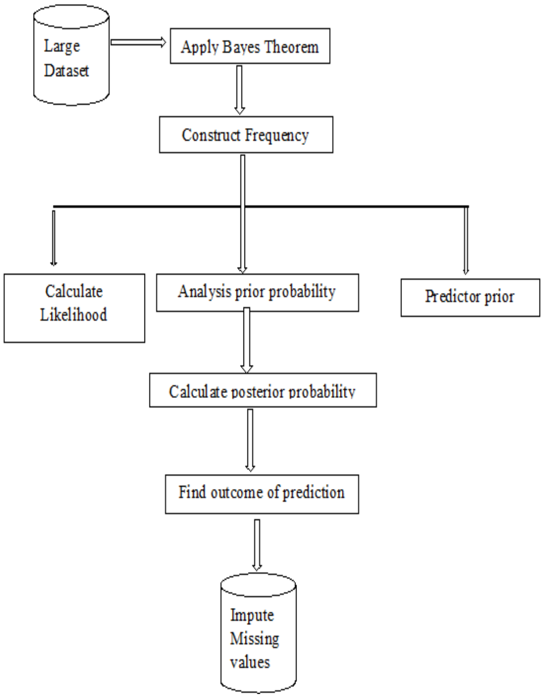


Figure 1. Flowchart of the Proposed System

1) **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.

2) **Zero-Frequency Problem***:* When an attribute value doesn't occur with every class value adds 1 to the count for every attribute value class combination.

3) **Numerical Predictors:** Numerical variables need to be transformed to their categorical counterparts before constructing their frequency tables.

V. EXPERIMENTALRESULTS

A. Design

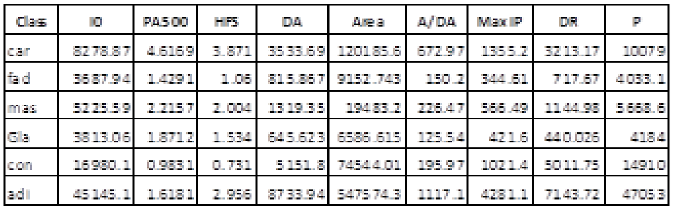
Experimental datasets were carried out from the Machine Learning Database UCI Repository. Table1. describes the dataset with electrical impedance measurements in samples of freshly excised tissue dataset contains number of instances and number of attributes about the datasets used in this paper. The 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.

Datasets	Breast Tissue
Instances	106
Attributes	10 (9features + 1 classes)
Missing rates	5% to 25%
Unsupervised	Mean, Median, Standard Deviation
Supervised	Naïve Bayesian

Table1. Datasets used for Analysis

B. Experimental Evaluation

Table2 describe the complete structure of all the attributes.



Clas	I0	PA500	HFS	DA	Area	A/DA	Max IP	DR	P
car	8278.87	4.6169	3.871	3533.69	120185.6	672.97	1355.2	3213.17	10079
fad	3687.94	1.4291	1.06	815.867	9152.743	150.2	344.61	717.67	4033.1
mas	5225.59	2.2157	2.004	1319.35	19483.2	226.47	566.49	1144.98	5668.6
Gla	3813.06	1.8712	1.534	645.623	6586.615	125.54	421.6	440.026	4184
con	16980.1	0.9831	0.731	5151.8	74544.01	195.97	1021.4	5011.75	14910
adi	45145.1	1.6181	2.956	8733.94	547574.3	1117.1	4281.1	7143.72	47053

Table2. Original datasets without missing values

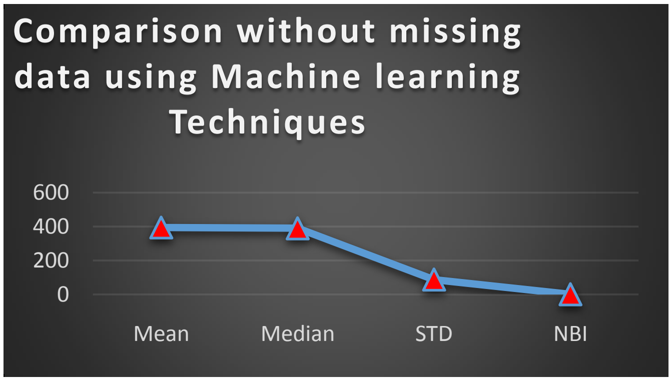


Figure2. Original Datasets without missing values

The above Figure 2 represents the classification of all attribute of original dataset using both the machine learning techniques without missing values.

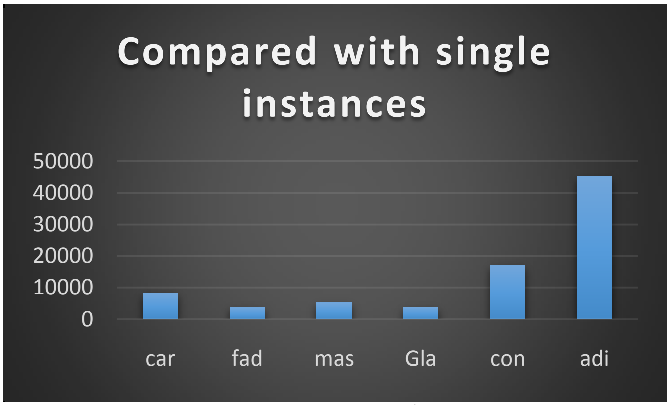


Figure3. Single instance of original datasets

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

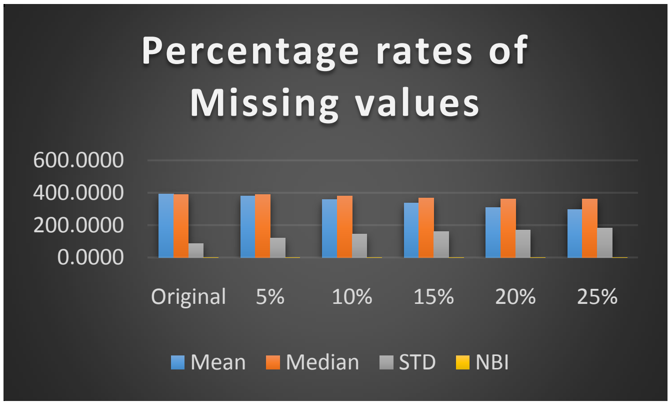


Figure4. Missing value rates for experimental analysis

Figure 4 specifies the different percentage rates of missing values for experimental analysis.

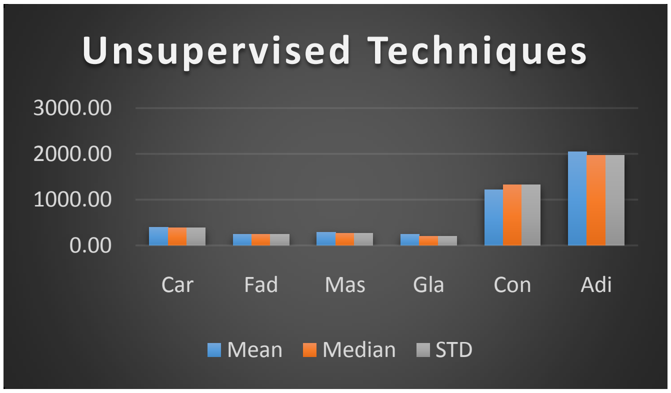


Figure5. Experimental results for Mean, Median and STD

Figure 5 & 6 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.

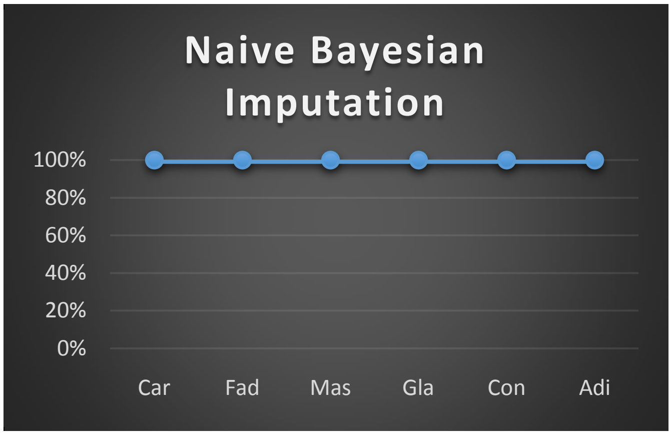


Figure6. Experimental results for Supervised Techniques

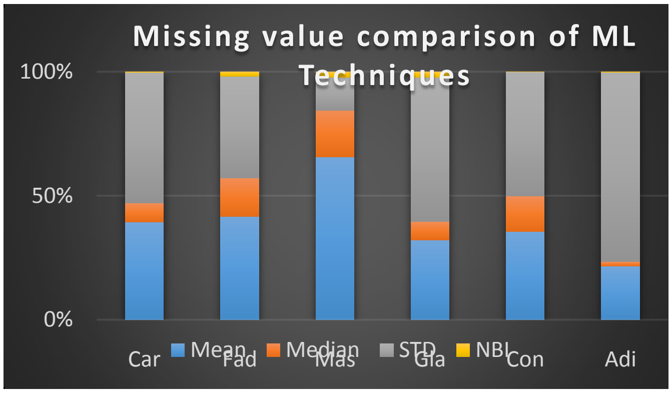


Figure7. Comparative results using missing values for both ML Techniques

Figure 7 specifies the comparison of both ML techniques using missing value and Table 3 describes the percentage of missing value occur in the original dataset.

Class	Original	5%	10%	15%	20%	25%
Car	394.232	380.4008	359.5065	336.3794	310.0995	296.0769
	389.873	389.873	380	366.9424	362.8313	362.8313
	87.04574	120.8659	145.6537	162.2337	170.0887	183.0835
	0.099589	0.096095	0.090817	0.084974	0.078336	0.074793
Fad	245.8626	229.5293	212.2036	212.2036	188.5369	168.45
	245	243.294	211	211	200	196.8567
	69.76127	94.33182	110.7889	110.7889	115.9052	120.9645
	0.044363	0.041416	0.03829	0.03829	0.034019	0.030395
Mas	290.3108	266.1389	253.0278	246.3056	216.047	200.4179
	267.6355	256.1388	256.1388	256.1388	251	223.1825
	111.9575	125.0618	139.8981	149.2123	140.0982	147.8644
	0.06286	0.057626	0.054787	0.053332	0.04678	0.043396
Gla	238.3162	228.8162	216.5037	216.5037	204.9412	195.8787
	197	197	191	191	187.5	187.5
	119.1858	131.9038	143.7354	143.7354	153.5449	161.3969
	0.045868	0.04404	0.04167	0.04167	0.039445	0.0377
Con	1212.864	1135.418	1135.418	1088.99	1009.574	891.8643
	1328.166	1328.166	1328.166	1328.166	1328.166	1020.334
	386.4724	504.7634	504.7634	577.5002	646.4499	670.8443
	0.204258	0.191215	0.191215	0.183396	0.170022	0.150198
Adi	2052.05	1956.596	1849.778	1774.778	1683.869	1524.778
	1974.559	1924.559	1875	1875	1850	1825
	342.4865	555.123	686.394	791.3788	874.7565	1003.077
	0.543062	0.517801	0.489532	0.469684	0.445625	0.403523

Table3. Percentage of missing values occur in original datasets

VI. Conclusion

This paper 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 less parameter for their analysis and the performance evaluation express among the other missing value imputation techniques the proposed method performs best. In future it can be extended to handle categorical attributes and it can be replaced by other supervised machine learning techniques.

References

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

[2]. 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.

[3]. 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.

[4]. S. Kanchana, Dr. Antony Selvadoss Thanamani, "Multiple Imputation Of Missing Data Using Efficient Machine Learning Approach", Internation Journal of Applied Engineering Research, ISSN 0973-4562 Volume 10, Number 1 (2015) pp.1473-1482.

[5]. 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.

[6]. 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,

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

[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]. 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.

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

[11]. 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.

[12]. R. Kavitha Kumar and Dr. R. M. Chandrasekar, “Missing Data Imputation in Cardiac data set”.

[13]. Ingunn Myrtveit, Erik Stensrud, “IEEE Transactions on Software Engineering”, Vol. 27, No 11, November 2001.

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

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

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

[17]. 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.

[18]. Ingunn Myrtveit, Erik Stensrud, “IEEE Transactions on Software Engineering”, Vol. 27, No 11, November 2001

[19]. K.Raja, G.Tholkappia Arasu, Chitra S.Nair, “Imputation Framework for missing value” International Journal of Computer Trends and Technology-Volume3 Issue2-2012.

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

Authors Profile

Smt. S.Kanchanais a research scholar of theDepartment 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, Dept 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.