Novateur Publication's
International Journal of Innovation in Engineering, Research and Technology [IJIERT]
ICITDCEME'15 Conference Proceedings
ISSN No - 2394-3696

COMPARATIVE ANALYSIS OF STANDARD ERROR USING IMPUTATION METHOD

V.B. Kamble
P.E.S. College of Engineering,Aurangabad. (M.S.), India

S.N. Deshmukh
Dr. Babasaheb Ambedkar Marathwada University,Aurangabad. (M.S.) India.

ABSTRACT

Presence of missing values in the dataset remains great challenge in the process of knowledge extracting. Also this leads to difficulty for performance analysis in data mining task. In this research work, student dataset is taken that contains marks of four different subjects of engineering college. Mean Imputation, Mode Imputation, Median Imputation and Standard Deviation Imputation were used to deal with challenges of incomplete data. By implementing imputation methods for example Mean Imputation, Mode Imputation, Median Imputation and Standard Deviation Imputation on the student dataset and find out standard errors for each imputation method then analyze obtained result. Mean Imputation with standard error is less as compare with other imputation method with standard error. Hence Mean Imputation Method with standard error is more suitable to handling the missing values in the dataset.

INTRODUCTION

Most of the real world datasets are incomplete, due to presence of missing values. Missing data is randomly distributed in the dataset. Incompleteness in the dataset occurs due to various reasons like manual data entry procedures, incorrect measurements, equipment errors etc. When data is incomplete, it is very difficult to extract useful knowledge from dataset as data analysis algorithms can work efficiently only with complete data. In this research work minimize errors by using imputation methods with standard error. Mean Imputation with standard error is less as compare with other imputation method with standard error. Hence Mean Imputation Method with standard error is more suitable to handling the missing values in the dataset.

The organization of the paper is Section1: Introduction of Imputation Techniques. Section 2: Related Work, Section 3: Missing Data Imputation Technique, Section 4: Experimental result and Analysis, Section 5: Conclusion and future work.

1.1 MISSING VALUES

Missing data occur due to nonresponsive field items when the data was collected and ignored by users. Missing values lead to the difficulty of extracting useful information from data set [2]. Missing data is the absence of data items that hide some information that may be important [1]. Most of the real world databases are characterized by an unavoidable problem of incompleteness, in terms of missing values. [3].

1.2 HANDLING MISSING VALUES

A) HANDLING MISSING VALUES BY IGNORING THE TUPLE

This method is not effective, unless the tuple contains several attributes with missing values. It is poor when the percentage of missing values as per attribute varies considerably.

B) HANDLING MISSING VALUES BY FILLING MISSING VALUE MANUALLY

This approach is time consuming and may not be feasible given a large data set with many missing values.

C) HANDLING MISSING VALUES BY USING GLOBAL CONSTANT

In this method missing values replace by some constant, for example Unknown. This method is simple, but it is not used frequently because it will bias the useful information from dataset. By using this technique cannot do the proper data analysis.

D) HANDLING MISSING VALUES BY USING ATTRIBUTE MEAN

This method uses mean to fill missing value. In this method missing values can be replaced by taking the respective attribute mean of incomplete dataset. This method is frequently used for data analysis but result may be biased [4].

1.3 SURVEY ON MISSING VALUE IMPUTATION METHODS

Missing value imputation is challenging issue in data mining. Missing value may generate bias result and affect the quality of the data mining task. But in practice a suitable method for dealing missing values is necessary. Missing values may appear either in conditional attributes and in class attribute. There are many approaches to deal with missing values described.

(a) Ignore tuple containing missing values;
(b) Fill missing value manually;
(c) Substitute the missing value by a global constant or the mean of the attribute;
(d) Most probable value to fill in the missing values.

First approach usually lost more useful information, whereas the second one is time- consuming and expensive, so it is not suitable in many applications. Third approach assumes that all missing values are with the same value, probably leading to considerable distortions in data distribution. The method of imputation is a popular strategy. In comparison to other methods, it uses as more information as possible from the observed data to predict missing values [14, 17]. Missing value imputation techniques can be classified into parametric imputation and non-parametric imputation. The parametric regression imputation is applicable if a dataset can be adequately modeled parametrically and users can correctly specify the parametric forms for the dataset [16, 17].

Non-parametric imputation can provide superior fit by capturing structure in the dataset, offers a nice alternative if users have no idea about the actual distribution of a dataset [7]. Statistical methods consist of linear regression, replacement under same standard deviation, and mean-mode method. But these methods are not completely satisfactory ways to handle missing value problems [5].

RELATED WORK

2.1 TREATMENT OF MISSING DATA

A) IGNORING TUPLE: This method determine the missing data on each instance and delete instances and remove the whole attribute which having high percentage of missing data in the dataset. This method is applicable only when the dataset is MCAR pattern.

B) PARAMETER ESTIMATION: This method is used to find out parameters for the complete data. This method uses the Expectation-maximization algorithm for handling the parameter estimation of the missing data.

C) IMPUTATION TECHNIQUE: This technique is very popular and frequently used in which replaces the missing values based on estimated values. Imputation techniques are classified into 1. Mean Imputation 2. Mode Imputation, 3.Median Imputation 4. Standard Deviation Imputation etc. [6].

2.2 RESEARCH INTO MISSING VALUE IMPUTATION

To impute missing value following techniques are used:

A) PARAMETRIC REGRESSION IMPUTATION METHODS: The parametric methods [18], [19], and [20], are superior while the data set are adequately modeled but in real applications it is impossible to know the distribution of the data set. Therefore, the parametric estimators can lead to highly bias, and the optimal control factor settings may be miscalculated [18, 19, and 20].

B) NON PARAMETRIC REGRESSION IMPUTATION METHODS: By using The Nonparametric imputation method [21], [22], [23] we know the structure of the data set. These imputation methods are designed for either continuous or discrete independent attributes. And these estimators cannot handle discrete attributes efficiently. Some methods, such as C4.5 algorithm [13], association rule based method [16], and rough set based method are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always makes discrete before imputing. This possibly leads to a loss of useful information of the continuous attributes [17].

MISSING DATA IMPUTATION TECHNIQUES

3.1 LIT WISE DELETION: This method deletes those instances with missing data and does analysis on the dataset. It is the most common method, it has two drawbacks: a) A substantially decreases the size of dataset available for the data analysis. b) Data are not always missing completely at random.

3.2 MEAN/MODE IMPUTATION (MMI)

By replacing a missing data with the mean or mode of all attribute which having missing value. To reduce the influence of exceptional data, median can also be used. This is one of the most commonly used methods.

3.3 K-NEAREST NEIGHBOR IMPUTATION (KNN)

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are a) it can estimate both qualitative attributes and quantitative attributes; b) It is not necessary to build a predictive model for each attribute with missing data [2].

3.4 MEDIAN SUBSTITUTION

Median Substitution is calculated by grouping up of data and finding average for the data. Median can be calculated by using the formula

$$\text{Median} = L + h/f\ (n/2 - c) \qquad (1)$$

where L is the lower class boundary of median class h is the size of median class i.e. difference between upper and lower class boundaries of median class f is the frequency of median class, c is previous cumulative frequency of the median class, n/2 is total no. of observations divided by 2

3.5 STANDARD DEVIATION: The 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. The Standard Deviation is given by the formula

$$S_{N=}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{x})2} \quad (2)$$

Where$\{x_1,x_2,\ldots,x_n\}$ are the observed values of the sample items and is the mean value of these observations, while the denominator N stands for the size of the sample [7].

EXPERIMENTAL RESULTS AND ANALYSIS

4.1 GENERATION OF MISSING DATA IN THE DATASET

In this work dataset having characteristics is given below.
Number of Instances: 5000
Number of Attributes: 08
(Year, Class, Semester, Roll No., M1, ECE, EM, EE)
Dataset contains marks of four different subjects of engineering college. In dataset randomly distributed the missing values in each attribute to become the incomplete dataset. Roll. No. in the Dataset is used are imaginary and generated for the data analysis purpose in data mining process.

Table1: Dataset

Year	Class	Semester	Roll. No.	Subject1	Subject2	Subject3	Subject4
Y1	C1	S1	01	X1	X2	X3	X4
..	..	..	..	..	..	..	..
YN	CN	SN	N	XN	XN	XN	XN

4.2 EVALUATION OF RESULT

Standard Error is calculated by find out the Standard Deviation of the respective feature/column in the dataset. After getting the standard Deviation (S.D.) it is divided by the square root of sample size (N) of the dataset, so in this way Standard Error is calculated by using the following formula.

$$\text{Standard Error (S.E.)} = \frac{\text{Standard Deviation (S.D.)}}{\text{Square Root of sample size (N)}} \quad (3)$$

As per experimental result following observations are found out

Table2: Out put

Sr. No.	MI Standard Error	Mode Standard Error	MDI Standard Error	SD Standard Error
01	0.2064	0.4317	0.2486	0.2714

As per result show in the above table Mean Imputation with Standard Error is 0.2064 is less as comparing to another Mode Imputation with Standard Error is 0.4317, Median Imputation with Standard Error is o.2486 and Standard Deviation with standard error 0.2714 so the performance of efficiency of Mean Imputation with Standard Error is more efficient as compare to other imputation technique, so as to handle the missing data which is randomly distributed in the dataset.

CONCLUSION AND FUTURE WORK

To handle the missing value in the dataset Mean Imputation with Standard Error is more efficient as comparing to other Imputation Technique for data analysis in the data mining field. In this work, performance of proposed method is most reliable as compare with the previous work.

In this work main focus is on how to handle missing data from the numerical dataset. In future work it is demanded to work on to handle the missing data from categorical dataset and also on image dataset for doing data analysis work to retrieve the useful and needed information from the missing dataset so as to take policy decision.

REFERENCES

[1]S.Kanchana, Dr. Antony Selvadoss Thanaman, Classification of Efficient Imputation Method for Analyzing Missing Values, International Journal of Computer Trends and Technology (IJCTT) – volume 12 number 4 – Jun 2014
[2]Minakshi , Dr. Rajan Vohra, Gimpy, Missing Value Imputation in Multi Attribute Data Set,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5315-5321,
[3]Alireza Farhangfara , Lukasz Kurganb , Witold Pedrycz, Experimental analysis of methods for imputation of missing values in databases
[4]Dinesh J. Prajapati, Jagruti H. Prajapati, Handling Missing Values: Application to University Data Set, International journal of emerging trends in Engineering and Development Issue1, Vol 1August2011
[5]R.S. Somasundaram, R. Nedunchezhian, Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values, International Journal of Computer Applications (0975 – 8887) Volume 21– No.10, May 2011
[6]Santosh Dane, Dr. R. C. Thool, Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013, ISSN: 2277 128X
[7]Ms.R.Malarvizhi, Dr. Antony Selvadoss Thanamani, Comparison of Imputation Techniques after Classifying the Dataset Using Knn Classifier for the Imputation of Missing Data, International Journal Of Computational Engineering Research (ijceronline.com) Vol. 3 Issue. 1, ISSN 2250-3005(online), January 2013
[8]Arnaud Ragel, Bruno Cremilleux & J. L. Bosson, an Interactive and Understandable Method to Treat Missing Values: Application to a Medical Data Set, In ACM Computer. Survey. 1985.
[9] Luai Al Shalabi, A comparative study of techniques to deal with missing data in data sets, In Proceedings of the 4th International Multiconference on Computer Science and Information Technology /CSIT 2006.
[10]A. Pujari, Data Mining Techniques, Universities Press, India, 2001.
[11]Ragel, A. and Cremilleux, B., MVC - a preprocessing method to deal with missing values, In Proceedings of Knowl.-Based Syst 1999, 285-291.
[12]Chih-Hung Wu, Chian-Huei Wun, Hung-Ju Chou, Using Association Rules for Completing Missing Data, Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 2004 pp.236-241.
[13]Lakshminarayan, K., Harp, S., Goldman, R., and Samad, T. 1996, Imputation of missing data using machine learning techniques, In Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining.
[14]Zhang, S.C., et al., Information Enhancement for Data Mining, IEEE Intelligent Systems, 2004, Vol. 19(2): 12-13, (2004).
[15]Qin, Y.S., Semi-parametric Optimization for Missing Data Imputation, Applied Intelligence, 2007, 27(1): 79-88.
[16]Zhang, C.Q., an Imputation Method for Missing Values, PAKDD, LNAI, 4426, 2007: 1080-1087.
[17]Han, J., and Kamber, M., "Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2006, 2nd edition.
[18]A. Dempster, N.M. Laird, and D. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Royal Statistical Soc., vol. 39, pp. 1-38, 1977.
[19]R. Little and D. Rubin, Statistical Analysis with Missing Data, second ed. John Wiley and Sons, 2002.

[20]D. Rubin, Multiple Imputations for Nonresponsive in Surveys, Wiley, 1987.
[21]J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[22]Q.H. Wang and R. Rao, Empirical Likelihood-Based Inference under Imputation for Missing Response Data, Annals of Statistics, vol. 30, pp. 896-924, 2002.
[23] S.C. Zhang, Par imputation: From Imputation and Null-Imputation to Partially Imputation, IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.
[24] V.B.Kamble, S.N.Deshmukh, Comparison of Percentage Error by using Imputation Method On Mid Term Examination Data, International Journal of Innovations in Engineering Research and Technology (IJIERT),Impact Factor 2.766, Volume 2, Issue 12,2015