NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 2, ISSUE 12, DEC.-2015

COMPARISION OF PERCENTAGE ERROR BY USING IMPUTATION METHOD ON MID TERM EXAMINATION DATA

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

The issue of incomplete data exists across the entire field of data mining. In this paper, Mean Imputation, Median Imputation and Standard Deviation Imputation are used to deal with challenges of incomplete data on classification problems. By using different imputation methods converts incomplete dataset in to the complete dataset. On complete dataset by applying the suitable Imputation Method and comparing the percentage error of Imputation Method and comparing the result

INTRODUCTION

Missing data are the absence of data items; they hide some information that may be important. The presence of missing data is a general and challenging problem in the data analysis field. Fortunately, missing data imputation techniques can be used to improve the data quality. Missing data imputation techniques refer to any strategy that fills in missing values of a dataset so that standard data analysis methods can be applied to analyzed completed dataset [1].
Information quality is important to organization. People use information attribute as a tool for accessing information quality. Information quality is measured based on users as well as experts opinion on the information attributes. The commonly known information attributes for information quality including accuracy, objectivity, reputation, access, security, relevancy, value added, timeliness, completeness, amount of data, and ease of understanding and consistent representation. These attribute can also be applicable to data quality. Commonly, one can rarely find a data set that contains complete entries [2].

1.1 Related Work- Methods for dealing with missing values can be classified into three categories 1) Case Deletion, 2) Learning Without Handling of Missing Values, and 3) Missing Value Imputation. The case deletion is to simply omit those cases with missing values and only to use the remaining instances to finish the learning assignments. The second approach is to learn without handling of missing data.

Missing data imputation methods advocates filling in missing values before a learning application. Missing data imputation is a procedure that replaces the missing values with some possible values [3].

1.2 Missing Data- Different methods have been applied in data mining to handle missing values in database. Data with missing values could be ignored, or a global constant could be used to fill the missing values, such as attribute mean, attribute mean of the same class, or an algorithm could be applied to find the missing values. Missing data imputation techniques means a strategy to fill the missing values of a dataset in order to apply the standard methods which require completed data set for analysis. These techniques retain data in incomplete cases, as well as impute values of correlated variables.

Missing data imputations techniques are classified as ignorable missing data imputations methods, which include single imputation methods and multiple imputation methods, and non-ignorable missing data imputations methods which include likelihood based methods and the non-likelihood based methods. Single imputation methods could fill one value for each missing values and it is more commonly used at present than multiple imputations which replace each missing value with several possible values and better reflects sampling variability about actual value [4].

1.3 Patterns of Missing Values- There are a number of ways to know how missing data arises. Little and Rubin introduced specific missing data terminology as a standard framework to deal with missing data mechanisms and their effect on data analysis.

(A) Missing completely at Random (MCAR)- If the probability that a response is missing is independent of both the observed data for that case and the unobserved responses are simple a random sample from the observed data. An example of MCAR missing data arises when investigators randomly assign research participants to complete two-thirds of a survey instrument

(B) Missing at Random (MAR)- If the probability that a response is missing depends on the observed data, but not on the unobserved data. This assumes the parameters of the model for the data are distinct from the parameters of the missingness mechanism. The missingness mechanism is ignorable. For example, in a reading comprehension test at the beginning of a survey administration session, research participants with lower reading comprehension scores may be less likely to complete the entire survey. The missing data are due to some other external influence.

(C) Not-Missing at Random. (NMAR)- When respondents and non-respondents, with the same values of some variables observed for both, differ systematically with respect to values of the variable missing for the non-respondent. In other words, the pattern of data missingness is non-random and it is not predictable from other variables in the database. For example, a participant in a weight-loss study does not attend a weight-in due to concerns about his/her weight loss; his/her data are missing due to non-ignorable factors.

In practice it is usually difficult to meet the MCAR assumption. It is mentioned that when making sampling distribution inferences about the parameter of the data, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the observed data are observed at random, but the inferences are generally conditional on the observed pattern of missing data [5].

FOUR DIFFERENT METHODS TO DEAL WITH MISSING VALUES

A. Case Deletion (CD) - It is Also known as complete case analysis. It is available in all statistical packages and is the default method in many programs. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and deletes the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with a high degree of missing values for other situations where the sample size is insufficient or some structure exists in the missing data, CD has been shown to produce more biased estimates than alternative methods. CD should be applied only in cases in which data are missing completely at random.

B. Mean Imputation (MI). The method replaces the missing data for a given feature (attribute) by mean of all the known values for that particular attribute.
Let us consider that the value of x_{ij} of the k_{th} class, C_k, is missing then x_{ij} is calculated as

$$x^{ij} = \underset{i:x^{ij}\text{€ck}}{\sum \frac{x^{\;ij}}{n^{k}}} \qquad (1)$$

Where n_k represents the number of non-missing values in the j_{th}feature of the k_{th} class.
However, According to Little and Rubin, the drawback of MI are

a) Sample size is overestimated
b) Variance is underestimated
c) Correlation is negatively biased
d) The distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean. Replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical test based on it.

C. Median Imputation (MDI). This method uses median of all known values of the feature or attribute in the class where the missing instance with missing value belongs.
Consider the value x_{ij} of the k_{th} class, C_k, is missing. It will be calculated as

$$x_{ij=}median\{i: x_{ij}\text{€}c_k\}\{x_{ij\}}\} \qquad (\ 2)$$

Instead of mean and median, mode also can be in imputation. Imputation method is applied separately for many attribute. However, imputation does not consider co-relation structure of the data [6].

D. Standard Deviation

The standard deviation spread 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

$$S_{N}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{i}-\hat{x})2} \qquad (3)$$

Where{x_1,x_2,…,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 ANALYSIS

We use three techniques for data retrieval

1) Imputing the missing values using attribute mean value.
2) Imputing the missing values using Median Imputation
3) Imputing the missing values using Standard Deviation

3.1 Imputing Missing Values Using Attribute Mean Value

This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instances with missing attribute belongs. In this method by replacing the missing values by attribute mean and find out percentage accuracy with original values by using the Mean Imputation Method.

3.2 Imputing Missing Values Using Median Imputation

In this method the missing values of instances are imputed. This method uses median of all known values of the feature or attribute in the class where the missing instance with missing value belongs.

Instead of mean and median, mode also can be in imputation. Imputation method is applied separately for many attribute. However, imputation does not consider co-relation structure of the data.

3.3 Imputing Missing Values Using Standard Deviation Imputation

The standard deviation spread data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. The missing values can be replaced by Standards Deviation value at respective attribute and then find out the percentage accuracy with original value.

DATASET USED

In this paper dataset consists of records of class test marks of engineering college. In dataset contains four attribute Roll No., Imputed value by Mean Imputation, Imputed value by Median Imputation and Imputed value by Standard Deviation Imputation. In dataset Roll No. are imaginary and generated for data analysis. Dataset contains some missing values randomly distributed at particular record number by using imputation techniques fill up the missing value so that dataset becomes complete dataset so that it is easy to analyze the dataset. Dataset is shown in following table.

Table 1. Dataset with Mean imputation, Median Imputation, Standard Deviation Imputation

Roll. No.	**Imputed Value by MI**	**Imputed value by MDI**	**Imputed value by SDI**
121	8.45	8	4.14
140	11.16	12	4.53
165	8.45	8	4.14
200	8.45	8	4.14
230	10.56	11	4.34
300	11.16	12	4.53
534	11.16	12	4.53
1000	11.16	12	4.53
1429	8.45	8	4.14
1800	11.16	12	4.53
2187	9.75	10	4.34
2220	11.16	12	4.53
2247	10.56	11	4.34
2350	11.16	12	4.53
2557	10.56	11	4.14
3000	11.16	12	4.53
3160	10.56	11	4.34
3225	8.45	8	4.14
3295	10.56	11	4.34
3350	11.16	12	4.53
3467	10.56	11	4.34
3550	11.16	12	4.53
3679	11.16	12	4.53
3700	10.56	11	4.34
3773	11.16	12	4.53
3900	9.75	10	3.9
4113	11.16	12	4.53
4125	11.16	12	4.53
4132	10.56	11	3.9
4150	8.45	8	4.14
4175	10.56	11	4.34
4190	11.16	12	4.53
4202	8.45	8	4.14
4250	11.16	12	4.53

4352	11.16	12	4.53
4500	8.45	8	4.14
4600	8.45	8	4.14
4762	9.75	10	3.9
4850	10.56	11	4.34
4920	11.16	12	4.53
4984	8.45	8	4.14
4998	11.16	12	4.53
Percentage Error	2.1	0.04	0.05

4.1 Comparison of Percentage Errors

Percentage Error calculates by observing the Experimental value and the Actual Value. By taking the difference between Experimental Value and Actual Value Divided by Actual Value and Multiplied by hundred so that we calculate the Percentage Error by using the following formula.

$$\text{Percentage Error} = \frac{\text{Experimental Value} - \text{Actual Value}}{\text{Actual Value}} * 100$$

Table 2. Comparison of Percentage Error

Imputation Technique	Percentage Error
Mean imputation	2.1
Median Imputation	0.04
Standard Deviation Imputation	0.05

By comparing the percentage error of imputation method. The percentage error of Mean Imputation is 2.1 %.The percentage error of Median Imputation is 0.04 and percentage error of Standard Deviation Imputation is 0.05 so the Median Imputation method is having lowest percentage of error as compare to Mean Imputation method and Standard Deviation Imputation method. So the Median Imputation Method is more suitable as compare to other method.

CONCLUSION AND FUTURE WORK

Missing values are regarded as serious problems in most of the information systems due to unavailability of data and must be impute before the dataset is used. To handle these missing values three techniques are used named as Mean imputation, Median Imputation and Standard Deviation Imputation. Median Imputation method is having lowest percentage of error as compare to Mean Imputation method and Standard Deviation Imputation method. So the Median Imputation Method is more suitable as compare to other method.

The proposed work handles missing values only for the numerical attributes. Further it can be extended to handle a categorical attribute. Different classification algorithm can be used for comparative analysis of missing data techniques. Missing data techniques can also be implemented in mat lab.

REFERENCES

[1] Dinesh J. Prajapati, Jagruti H. Prajapati, and Handling Missing Values: Application to University Data set. Issue 1, Vol. 1(August-2011), ISSN 2249-6149

[2] Shamsher Singh, Prof. Jagdish Prasad, Estimation of Missing Values in the Data Mining and comparison of Imputation Methods. Mathematical Journal of Interdisciplinary Sciences Vol. 1, Issue 1, March 2013, pp. 75–90

[3] Xiao Feng Zhu, Shichao Zhang, Senior Member, IEEE, Zhi Jin, Zili Zhang, and Zhuoming Xu, Missing Value Estimation for Mixed-Attribute Data Sets. IEEE Transactions on Knowledge And Data Engineering, Vol. 23, No. 1, January 2011.

[4] T.R.Sivapriya, V. Thavavel, A.R.Nadira Banu Kamal, Imputation and classification of Missing Data Using Least Square Support Vector Machines- A New Approach in Dementia Diagnosis. International Journal of Advanced Research in Artificial Intelligence, Vol.1, No.4, 2012

[5] Yann-Yann Shieh, Imputation Methods on General Linear Mixed Models of Longitudinal Studies, American Institutes for Research

[6]] Edgar Acu˜Na1 And Caroline Rodriguez, The Treatment Of Missing Values And Its Effect In The Classifier Accuracy Studies In Classification, Data Analysis, And Knowledge Organization, 2004, Springer.Com

[7] MS. R. Malarvizhi, Dr. Antony Thanamani, Comparision of Imputation Techniques after Classifying the Dataset Using Knn Classifier for the Imputation of Missing Data, International Journal of Computational Engineering Research (IJCER online.com) ISSN 2250-3005, Janaury-2013

[8] Anjana Sharma, Naina Mehta, Iti Sharma, Reasoning With Missing Values in Multi Attribute Datasets. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue5, May 2013 ISSN: 2277 128X