Two Single Imputation Method a Comparison for Handle Missing Values in Large Dataset

A.Finny Belwin[1] Dr.G.P.Rameshkumar[2]
[1]Ph.D Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]S.N.R Sons College, Coimbatore-641006

Abstract— *In real world, data may be incomplete, inconsistent or noisy. Missing values may occur due to several reasons. Data pre-processing is required in order to improve the efficiency of an algorithm. One of the challenging issues in data pre-processing is to handle the missing values in machine learning and data mining. There is a need for quality of data, thus it is ultimately important. To recover the solution of missing values the imputation techniques such as single, multiple and iterative imputations are there. The performance of the proposed algorithm has been compared with the other simple and efficient imputation methods. We compare Mean based Single Imputation (MI) and Standard Deviation Imputation (SDI) for effectiveness and improvement.*

Keywords— Data mining, Pre-processing, Imputation, Mean Imputation

I. INTRODUCTION

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans [1]. Data Mining is the notion of all methods and techniques, which allow analysing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies [3].

A. Missing data:

Missing data or missing values occur when no data value is stored for an instance in the current record. Missing data might occur because value is not relevant to a particular case, could not be recorded when data was collected or ignored by users because of privacy concerns [14]. Most information system usually has some missing values due to unavailability of data. Sometimes data is not presented or get corrupted due to inconsistency of data files. Missing data is a common problem that has a significant effect on the conclusion that can be drawn from the data. Missing data is absence of data items that hide some information that may be important [1]. 1) Types of missing data: There are basically three types of missing data, these are: MCAR- It is probability of missing data on any attribute does not depend on any value of attribute [7]. The term "Missing Completely at Random" refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset [2] MAR- The probability of missing data on any attributes does not depends on its own value but value of other attribute [7]. Sometimes data might not be missing at random but may be termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable [2].

B. MNAR-

Missing data depends on the values that are missing [7]. Sometimes data might not be missing at random but may be termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable [2], To recover the solution of missing values the imputation techniques such as single, multiple and iterative imputations are there. The performance of the proposed algorithm has been compared with the other simple and efficient imputation methods. Out of different ways of method we have proposed new efficient single imputation method Advance Mean Imputation (AMI) [11].

II. SINGLE IMPUTATION ALGORITHMS

There are various methods to impute the missing values introduced by the researchers such as case-deletion, Mean Substitution, Single Imputation Methods, Multiple Imputation methods, Iterative Imputation Methods and so on[10]. In case deletion method the values which are missing will ignore or delete the instances or attribute. In Single imputation method the values are impute by particular value. In Multiple imputations procedure it replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute [7].

In this section, a standard mean based imputation technique [8] as well as out proposed imputation techniques are addressed.

A. Mean Imputation

The procedure of the MI algorithm is as follows. Let D have an Original Numerical dataset with random missing values.

Now we have apply min-max normalization to dataset Ð and modified to dataset Ð'. In order to handle the missing values

- M in dataset Ð' the attribute containing
- Missing value m_i take the mean μ of that
- Attribute a_i and fill the missing value m_i
- with correspond attributes Values a_i [9].
- Here in this algorithm we
- Have taken attribute as A. we have find the
- Mean μ of attributes A and stored in value
- a_i, Using a_i fill up the dataset Ð' and
- Generate new modified dataset.

1) Procedure: MI

2) Input: Original Dataset Ð
3) Output: Modified Dataset Ð"
Do
Ð' ←Generate missing valued dataset from dataset *Ð*
Normalize each attribute value e_i using
Min-max normalization in dataset *Ð'*

$$Normalized\ (e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

Let
Ð' = { A1, A2, A3, ….. An }
For each attribute Ai in *Ð'*
Find the missing value M in *Ð'*
a_i = Ai ∩ mi
a_i = µ (a_i / N)
Fill up dataset *Ð'* using a_i
End For
End For
Generate Dataset *Ð"*

B. Standard Deviation Imputation

Let Ð have a Original dataset with random missing values.

Now we have apply min-max normalization to dataset Ð and modified to Ð'. In order to handle the missing values M in dataset Ð' the first loop adds each element É or number in the data array a [i] together. It is then divided by the total number Ñ of elements to create the mean Ê µ. In second loop the mean Ê has been subtracted and the result has been squared ß together. Finally, this number ß is divided by one less than the total number Ñ of data entries before being square-rooted. We have find the mean SDI of attributes A and stored in value SDI. Using SDI fill up the dataset Ð' and generate new modified dataset Ð''

1) Procedure: SDI
2) Input: Original Dataset Ð
3) Output: Modified Dataset Ð"
Do
Ð' ← Generate missing valued dataset from dataset *Ð*
Normalize each attribute value e_i using min-max normalization in dataset *Ð'*

$$Normalized\ (e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

Compute standard deviation

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \qquad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Let *Ð'* = { A1, A2, A3, ….. An }
For i = 0 to Ñ
É = É + a [i] ;
To count the value of (x-x');
Next I;
End
Ê = µ É + Ñ;
For j=0 to Ñ
End
SDI= sqrt (ß / (Ñ -1));
Fill up dataset *Ð'* using SDI
Generate Dataset *Ð".*
Ð' = { R1, R2, R3… Rm }
For k=1 to Ñ
Impute Mean With Í in Ð'
Let μ_j be the mean of elements Ð'(Í, Ñ)
R_j (k) = µ R_j (k) / Ñ
Fill up dataset *Ð"* using R_j (k)
End For
End For Generate Dataset Ð"'

III. Framework and Experimental Analysis

The experiments will conducted on UCI data sets at different missing ratios. Imputation is the process of finding a feasible or plausible value for a missing value. After imputing all the missing values, the numerical dataset [10] can be analyzed using standard techniques for complete data. For comparing datasets we will use the classification technique known as Naive-Bayes Classification Algorithm to measure the accuracy of all datasets and check the performance of the proposed algorithm and other algorithm. Datasets are taken from Standard UCI Machine learning data repository [11] and we have conducted three datasets Classification. These all datasets are of numeric attributes. Original datasets does not have missing values we have randomly moves the data from the dataset with 5% missing in the dataset.

DATASET	ORIGINAL	MI	SDI
wine	63.1218 %	68.616 %	68.9537 %
Iris	58.5507 %	60.00%	60.300%
E.Coli	58.6854 %	63.3803%	63.4803 %

Table 1: Comparison of Accuracy with Original dataset V/S Imputed dataset

Evaluation parameter	MI	SDI
RMSE	0.7586	0.7678
MAE	0.4777	0.4978
Precision	0.5723	0.5997
Recall	0.6870	0.7512

Table 2: Evaluation parameter accuracy

Now we are comparing the Original dataset and imputed datasets. We have taken the readings from standard tool known as Weka 3.7 Version which stands for (Waikato Environment). We have applied the Naive Bayes Classification algorithm from the Weka Software. The Naive Bayes classification algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The values are missing in class attributes and conditional attributes [5][6]. The percentage of the dataset are correctly classified in original dataset shown in table 1.

From following Parameter it has proven that SDI algorithm works better than MI algorithm and it has observed on three different datasets from the Table 2.

IV. Conclusion

Missing data imputation is a procedure that eplaces the missing values with some possible values. Missing values are regarded as serious problem in most of the information system due to unavailability of data and must be impute before the dataset is used. The SDI method works better than other imputation methods. In future we will implement these methods with categorical attributes to get the better accuracy, confidence intervals and to fill the missing value by comparing original dataset.

REFERENCES

[1] Han and Kamber, "Data Mining Concepts and Techniques", 2nd edition, 2006.
[2] Ludmila Himmelspach and Stefan Conrad , "Clustering Approaches for Data with Missing Values: Comparison and Evaluation" 2010
[3] Nambiraj Suguna and Keppana Gowder Thanushkodi, "Predicting Missing Attribute Values Using k-Means Clustering", Journal of Computer Science 7 (2): 216-224, 2011.
[4] Bhavisha Suthar, Hemant Patel and Ankur Goswami, "A Survey: Classification of Imputation Methods in Data Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 1, January 2012.
[5] E. Chandra Blessie, Dr. E. Karthikeyan and Dr. V.Thavavel "Improving Classifier Performance by Imputing Missing Values using Discretization Method" , International Journal of Engineering Science and Technology (IJEST), Vol. 4No.03 March 2012.
[6] Kavitha.P and T.Senthil Prakash, " Missing Value Estimation For Mixed Attribute Data Sets Using Higher Order Kernels", International Journal of Communications and Engineering Volume 05– No.5, Issue: 01 March2012.
[7] K. Raja, G. Tholkappia Arasu, Chitra. S. Nair, "Imputation Framework for Missing Values", International Journal of Computer Trends and Technology- volume3 Issue2- 2012 [VOL
[8] S.Thirukumaran and Dr. A.Sumathi, "Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm To Improve The Efficiency Of Imputation", IEEE- Fourth International Conference on Advanced Computing, ICoAC December 2012.
[9] Ms.R.Malarvizhi and Dr.Antony Selvadoss Thanamani, "Framework for Missing Value Imputation", International
[10] Anjana Sharma, Naina Mehta and Iti Sharma," Reasoning with Missing Values in Multi Attribute Datasets"International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.
[1] Shradha Prajapati1, Shruti Patel2, Heemani Chaudhari3 "Single Imputation Method to Handle Missing Values in Large Dataset" Volume 2, Issue 12, December-2015