Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 12, December-2015

Single Imputation Method to Handle Missing Values in Large Dataset

Shradha Prajapati[1], Shruti Patel[2], Heemani Chaudhari[3]

[1] *M.Tech, IT Department., Ganpat University, Gujarat (INDIA).*
[2] *M.Tech, IT Department., Ganpat University, Gujarat (INDIA).*
[3] *M.Tech, IT Department., Ganpat University, Gujarat (INDIA).*

Abstract: - *In real world, data may be incomplete, inconsistent or noisy. Missing values may occur due to several reasons. Data preprocessing 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. Out of different ways of method we have proposed new efficient single imputation method Advance Mean Imputation (AMI). We evaluate imputed data with Naive Bayes classification algorithm. We compare proposed method AMI with Mean based Single Imputation (MI) and Standard Deviation Imputation (SDI) for effectiveness and improvement.*

Keywords**:** *Data mining, Preprocessing, Imputation, Mean Imputation.*

I. INTRODUCTION

Data mining refers to extracting or mining" knowledge from large amounts of data"[1]. Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. Missing data, or missing values, occur when no data value is stored for the variable in an observation[1]. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Missing Data is a widespread problem that can affect the ability to use data to construct effective predictions systems Handling of imputation causes the three major issues: Loss of information as a consequence a loss of efficiency, Data handling is an issue, computation and analysis due to irregularities in the data structures, Systematic difference among the data[4].
Missing data is absence of data items that hide some information that may be important. Missing data mechanism can be divided into 3 categories [5][7]:

1. "Missing at Random" (MAR),
2. "Missing completely at Random", (MCAR)
3. "Missing Not at Random" (MNAR)

II. THE NORMAL AND THE PROPOSED IMPUTATION ALGORITHM

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.

2.1 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.

Procedure: MI

Input: Original Dataset Ð

Output: Modified Dataset Ð"

Do

Ð' ← Generate missing valued dataset from dataset *Ð*

ℜ ← 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"*

2.2 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''

Procedure: SDI

Input: Original Dataset Ð

Output: Modified Dataset Ð"

Do

Ð' ← Generate missing valued dataset from dataset *Ð*

ℜ ← 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

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

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
ß = (a[j] - Ê) ^2
Next J;
End
SDI= sqrt (ß / (Ñ -1));
Fill up dataset *Ð'* using SDI
Generate Dataset *Ð"*.

2.3 Advance Mean Imputation (AMI)

Let *Ð* have a Original dataset with random missing values. Now we have apply min-max normalization to dataset Ð and modified to Ð'. Apply the MI method and modified the dataset with Ð". Find the distance Ü between current row and other rows in dataset Ð" with help of distance metric function and store into the one total distance variable then find the mean of that variable with one less than the total number Ñ. For mean value calculations, the records with minimum distance with the missing value were not taken into the dataset. Find the mean distance higher than missing value is taken and store into the instance of the element. Now again the mean imputation is applied with the Dataset Ð' and again the mean is calculated with instance I and the one less than the total number Ñ and the value is stored in the new dataset Ð"'.

Procedure: AMI

Input: Original Dataset Ð

Output: Modified Dataset Ð"'
Do
Ð' ← Generate missing valued dataset from dataset *Ð*
ℜ ← 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' = { R1, R2, R3… Rm }
For each attribute R in *Ð'*
Find the missing value Mi in *Ð'*
$r_i = R_i \cap mi$
$r_i = \mu\ r_i / Ñ$
Fill up dataset *Ð'* using r_i
End For
End For
New Dataset Ð"
Let
Ð'' = {R1, R2, R3… Rm}
For each attribute R in Ð'' ÜÝÍ
Find the Distance Metric in Ð'' with each row
Û = Ü (Ð'', R_j)
Í = Ð' > µ (Û)
End For
For

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 which are Indian liver patient dataset, Liver disorder, Page-Block 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	STDI	AMI
Indian liver patient dataset	63.1218 %	68.6106 %	68.9537 %	69.2537 %
Liver disorder	58.5507 %	60.00%	60.00%	61.8696 %
Glass Identification	58.6854 %	63.3803%	63.3803 %	64.9533 %

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

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 and new generated dataset after applying proposed methods *MI, STDI and AMI* shown in table 1.

Evaluation Parameter	MI	STDI	AMI
Root mean squared error	0.5094	0.5094	0.5069
Mean absolute error	0. 4333	0. 4333	0.4301
Precision	0. 599	0. 599	0. 608
Recall	0.600	0.600	0. 609

Table 2: Comparison of parameters with Liver Disorder Dataset V/S imputed dataset

Evaluation Parameter	MI	STDI	AMI
Root mean squared error	0. 4717	0. 4699	0. 4692

Mean absolute error	0. 3486	0. 3469	0. 3459
Precision	0. 657	0. 660	0. 659
Recall	0. 686	0. 690	0. 690

Table 3: Comparison of parameters with Indian Liver Patient Dataset V/S imputed dataset

Evaluation Parameter	MI	STDI	AMI
Root mean squared error	0.5338	0.5338	0.493
Mean absolute error	0.3888	0.3888	0.3891
Precision	0.639	0.639	0.654
Recall	0.634	0.634	0.650

Table 4: Comparison of parameters with Glass Identification V/S imputed dataset

From following Parameter it has proven that proposed algorithm works better than MI algorithm and it has observed on three different datasets from the Table 2, Table 3, and Table 4.

IV. CONCLUSION

Missing data imputation is a procedure that replaces 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 Proposed method works better than other imputation methods. In future we will implement these proposed 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. 4 No.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 Journal of Engineering Research and Development, Volume 4, Issue 7, November 2012.
[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.
[11] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/.

