Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Study of Missing Value Imputation Methods – A Comparative Approach

Naresh Rameshrao Pimplikar*, Asheesh Kumar, Apurva Mohan Gupta

SCSE, VIT University

India

Abstract*— Data mining is the field of studying experimental data sets for the discovery of interesting and potentially useful relationships. The complete set of data is necessary without any missing values (MVs) in it. Practically, it's not easy to get complete data. Many researches are going today in this area where data is missing. Several methods are proposed for imputation of missing values using available values in the data set. In this study, different methods are overviewed and compared to solve the MV problem, along with their advantages and disadvantages. By observing the performance of these methods, the efficient method is then selected to fill the MVs of different data sets.

***Keywords*— Data mining, Missing values, Imputation methods, Comparison**

I. INTRODUCTION

The aim of data mining is extracting the knowledge out of huge set of data. The knowledge that is mined should be useful and advantageous. This method involves many areas such as medical diagnosis, databases, learning machine and statistical analysis. So, the complete dataset is crucial which does not contain any MVs. For instance, data collection is carried out normally by survey in statistics, but it is not necessary that people will answer for all the questions. They preferably answer to only those questions which don't need any personal or privacy related information, like which is favourite mobile brand or favourite tourist place; But people normally never answer to those questions regarding their income, email, or contact number. Thus, the data collected contains some missing values for particular feature(s). In addition to this, because of error in measurement tool MVs can come. If feature value that is needed to measure lies beyond the limitations of measurement tool then that feature value remains null i.e. missing in the collected dataset. Missing values in dataset produce incomplete dataset; and the dataset is called as complete dataset if it contains no MVs. Further, it is challenging to collect feature values for clinical data because of noise, large dimensions and different errors [3]. Moreover, in data mining, it can be said that clinical data probably is uncertain and needs efficient mining methods [6]. So, for handling MVs in dataset different imputation methods are proposed with their advantages and disadvantages. In this survey, such imputation methods are compared and the method with efficient performance is then can be selected to fill the missing values for particular dataset.

II. RELATED WORKS

The main purpose of database is data integrity. Missing data tends to reduce of integrity. Thus to prevent the data integrity from such degradation and to get better result from it, the MVs in data have to impute in proper manner. For imputing the MVs number of imputation methods is studied with the detail survey about them describing their types, pros and cons. Some important imputation methods that are discussed are given below.

A. Deletion Methods

Deletion method is categorized into two types, first is listwise deletion and second is pairwise deletion. In listwise deletion the MV records are deleted and further evaluation is carried out. Thus, this method is simple to use and understand. But it involves many disadvantages. The available data is discarded by deletion of records with MVs. Thus, there can be much reduction in dataset size. Also, this will result in precision loss and the induce bias. The mean value of original dataset and its mean after deletion vary greatly. Listwise deletion also has high effect on variability.

Pairwise deletion, whereas, will not delete the whole record. The available values are kept as it is so that each feature can be examined in independent way. Only MVs are deleted. Therefore, this method has less effect on mean and variability as compared to listwise deletion [1].

B. Mean Imputation Method OR Most Common Imputation (MCI)

The mean imputation method or MCI is also simple to use. The MV is filled using the mostly common value of attribute in case of nominal attributes. If the values in dataset are numerical then MV is filled with the mean of attributes [2, 5, 6], [H. Park, G. Golub, H. Kim, 2005][P. Allison, 2001].

C. Concept Most Common Imputation (CMCI)

CMCI is like the MCI. It also fills the MV by most common repeated attribute in case of nominal attributes and fills the MV with mean of attributes in case of numerical attributes. The only difference is CMCI takes into consideration the records of same class as reference records or cases [2, 5], [H. Park, G. Golub H. Kim, 2005][P. Allison, 2001].

D. Regression Imputation

Using regression method for imputation, the values from the features are observed and then predicted values are used for filling MVs. The output (response) variable 'Y' is determined based on the input (explanatory) variable 'X'. Regression determines the relationship between these two variables. Thus regression equation can be written as

$Y = \alpha_0 + \alpha_1 X + e$. Where, '$\alpha_0$' and '$\alpha_1$' are called as coefficients of regression and 'e' denotes the error that gives the data variation for the line, above and below. The aim is to minimize this variation and to determine the straight line which fits best for the data [1].

E. Expectation Maximization (EM)

Expectation maximization [1, 2, 6, 9] is a method for estimating maximum likelihood in various kinds of problems such as MV problem. It is based on the iteration. In each iteration, there are two steps i.e. Expectation (E step) and second is Maximization (M step). "M step" evaluates max likelihood estimation (MLE) for determining the best factor whereas "E step" determines the conditional expectation with the help of the factor and data available. The EM iterative algorithm is as below:

1) Initialization of the MV
2) Use MLE for determining the current best factor
3) Again impute the MV by new factor
4) Go to 2) till the result gets converge.

F. KNN Hot Deck Methods

This method use alike records to fill the MVs. The hock deck method with k nearest neighbour is described here. In this method, mainly the metric is to be defined for calculating the distance in between data records. Three different metrics are used in the study, Euclidean distance, Mahalanobis distance and Grey distance.

Consider a record Y to impute. Entire distances in between the Y and other record with no MV are calculated. After that the topmost k nearest data records for Y are chosen. Then average of entire values of such k data records is found out which becomes the value of Y to be imputed [1, 2, 6].

G. K-Means Imputation (KMI)

K-means algorithm is used for classification of the data. It classifies objects into K groups on the basis of attributes. To make a group, the sum of the squares of the distances in between data and centroid of cluster is minimized. The centroid denotes the mean of objects present inside the cluster. After clusters are converged, data objects which are part of the alike cluster would be considered as the nearest neighbour to one another. Then KMI uses algorithm called nearest neighbour to impute the MVs in the same way as KNNI [2, 4, 7, 8], [M. Cantor, O. Troyanskaya, 2001] [M. Monard, G. Batista, 2003].

H. Fuzzy K-Means clustering Imputation (FKMI)

In FKMI, membership function plays an important role. Membership function is assigned with every data object that depicts in what degree the data object is belonging to the particular cluster. Data objects would not get allotted to concrete cluster which is denoted by centroid of cluster (as in the case of K means), this is due to the various membership degrees of every data with entire K clusters. Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values [2, 10]

I. Support Vector Machine Imputation (SVMI)

The SVMI is regression based method to impute the MVs. It takes condition attributes (here, decision attribute i.e. output) and decision attributes (here, conditional attributes). SVMI then would be applied for prediction of values of missed condition attribute [2, 3].

J. Artificial Neural Network with Rough Set Theory (ANNRST)

The ANNRST imputation method is divided into: Reducing attributes of RST and ANN construction for missing attribute value prediction. Reduction combines attributes which distinguish between the objects and attributes of decision system S = (U, A,d); here 'U' denotes object set, 'A' is conditional attribute set and 'd' denotes decision attributes. Then computing minimum reduct, construction of ANN is done where inputs used are reduct attributes (conditional) and decision attributes. 'n' number of ANN topologies constructed for imputing 'n' attributes missing value. In case of k number of conditional attributes, l decision attributes, ANN would have input attributes k-1+l containing one conditional attribute for decision attribute which has MV as the output [11].

III. IMPUTATION METHODS SUMMARY

Different imputation methods used in this study can be summarized as shown in Table I as follows:

TABLE II
SUMMERIZATION OF IMPUTATION METHODS

Imputation Methods	**Research Papers**	**Description**	
Deletion Methods	[1]	***Listwise Deletion***	***Pairwise Deletion***
		Deletion of cases containing missing values (entire row is deleted)	Deletion of records only from column containing missing values
		High loss of information due to deletion of entire row	Less loss of information by keeping all available values
		High effect on variability	Less effect on variability
		Loss of precision and induce bias	Less Loss of precision and induce bias
Mean Imputation (Most	[2, 5, 6]	Replace MVs with the arithmetic mean of data	
		Resultant Mean and SD after imputation may be much higher than that of	

Common Imputation) (MCI)		original
		Not a good substitution method
Concept Most Common Imputation (CMCI)	[2, 5]	Same as Mean Imputation, but replaces MV by mode in case of nominal or by the mean value in case of numerical
		Takes into account the records of same class as reference records
Regression Imputation	[1]	Replace MVs with the values predicted from observed values
		Regression Equation: $Y = \alpha_0 + \alpha_1 X$
		To avoid lack of variability: $Y = \alpha_0 + \alpha_1 X + e$
Expectation-Maximization (EM)	[1,2,6,9]	Iterative method, finds maximum likelihood
		Two steps: Expectation (E step), Maximization (M step)
		Iteration goes on until algorithm converges
KNN Hot Deck Methods	[1, 2, 6]	Use alike records to fill the MVs
		Euclidean distance (KNN-ED) between records $d(X_i, X_j) = \sqrt{(x_{i1} - x_{j1})^2 + ... + (x_{ik} - x_{jk})^2}$
		Missing value term in X_i or X_j is not counted in the equation
		Mahalanobis distance (KNN-MA) $d(X_i, X_j) = (X_i - X_j)^T C_X^{-1} (X_i - X_j)$
		Grey distance (KNN-Grey) $d(X_i, X_j) = 1 - GRG(X_i, X_j).$
		The topmost k nearest data records for Y are chosen
		Average of entire values of such k data records becomes the impute value of Y
K-Means Imputation (KMI)	[2,4,7,8]	The centroid denotes the mean of objects present inside the cluster.
		After clusters are converged, data objects which are part of the alike cluster would be considered as the nearest neighbour to one another
		Data object within same cluster are considered as nearest neighbour to each other
		Then KMI uses algorithm called nearest neighbour to impute the MVs in the same way as KNNI
Fuzzy K-Means clustering Imputation (FKMI)	[2, 10]	Membership function is assigned with every data object that depicts in what degree the data object is belonging to the particular cluster
		Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values
Support Vector Machine Imputation (SVMI)	[2, 3]	Regression based method to impute the MVs
		Takes condition attributes (here, decision attribute i.e. output) and decision attributes (here, conditional attributes)
		SVMI then would be applied for prediction of values of missed condition attribute
Artificial Neural Network with Rough Set Theory (ANNRST)	[11]	Divided into : Reducing RST attribute and ANN construction for missing attribute value prediction
		Reduction combines attributes which distinguish between the objects and attributes of decision system S = (U, A,d)
		'n' number of ANN topologies constructed for imputing 'n' attributes missing value

IV. COMPARISONS

The following Table IIIIV give the comparison of the imputation methods depicted above in terms of advantages, disadvantages and their studies:

TABLE V VI
COMPARISON OF IMPUTATION METHODS

Imputation Methods	Advantages	Disadvantages	Studies by
Listwise Deletion	Simple to use	Loss of huge data, loss of precision, high effect on variability, induce bias	Staelin and Gleason (1975), Curry and Kim (1977), Malhotra (1987), Roth (1994) and [1]
Pairwise Deletion	Simple, keeping all available values i.e. only missing values	Loss of data, not a better solution as compared to other	

	are deleted	methods	
Mean Imputation (Most Common Imputation) (MCI) & Concept Most Common Imputation (CMCI)	Simple to use, it is built in most of the statistical packages	Resultant Mean and SD after imputation may be much higher than that of original	P. Allison (2001), H. Park, G. Golub and H. Kim (2005) and [1,2,6]
Regression Imputation	Calculated data saves deviations from mean and distribution shape	Degree of freedom gets distort and may raises relationships	Cohen (1983), Rubin and Little (1987), Roberts and Raymond (1987) and [1]
Expectation Maximization (EM)	Accuracy increases if the model is right	Takes time for converging, very complex	Rubin and Little (1987), Malhotra (1987), Laird (1988), Azen (1989), Ruud (1991), Donaldson and Graham (1993) and [1,2,9]
KNN Hot deck Imputation	MVs are imputed by realistically obtained values which avoids distortion in distribution	Bit empirical work for accuracy estimation, creates problem if any other sample has no close relation in entire manner of the dataset	Roth (1999), Ford (1983) and [1,2,6]
K-Means Imputation (KMI)	Fast and hence good for running big datasets. Reduces intra cluster variance to minimum.	Does not assure the global min. variance. Difficult to predict 'K' value.	M. Cantor, O. Troyanskaya (2001), M. Monard, G. Batista (2003) and [2,7,8]
Fuzzy K-means clustering Imputation (FKMI)	Best outcome for overlapping data, better than k means imputation. Data objects may be part of more than one cluster center.	High computation time. Noise sensitive i.e. low or no membership degree for noisy objects.	Rodriguez C and Acuna E (2004), Shuart B, Spaulding W and Li D (2004) and [2,10]
Support Vector Machine Imputation (SVMI)	Efficient in large dimensional spaces. Efficient memory consumption	Poor performance if number of samples are much lesser than number of features	Yang B, Guoshun C, Chen Y, Feng H, Cheng Y (2005) [2, 3]
Artificial Neural Network with Rough Set Theory (ANNRST)	Generally ANNRST dataset yields better accuracy than CMCI and KNN classifiers.	Complex computations and time consuming	Ohrn (1999), Hu M, Grzymala-Busse J (2001) [11]

V. OBSERVATIONS

Based on study performed using imputation methods, it is clear that different imputation methods gives different performance for different types of datasets. So, the observation include following results:

1) Deletion method for handling MVs is good option if very less attribute values are missing in case of huge dataset. However, pairwise deletion is better than listwise deletion. [Blood pressure dataset, 1].
2) Regression imputation results in much approximate MVs filling than MCI and CMCI. [Blood pressure dataset, 1].
3) Imputation using EM performs better than Regression imputation and MCI or CMCI because of its high convergence. [Blood pressure dataset,1], [Clinical Heart Failure Data,2].
4) KNN Hot deck imputation method is better than EM and Regression for mpg but worse for body fat [1].
5) KNN method has better accuracy than MCI and K-means for filling MVs [4].
6) SVMI and EM imputation is recommended for clinical heart failure dataset because of its consistent precision and recall [2, 6].
7) ANNRST shows best imputation as compared to KNN and CMCI for coronary heart disease dataset [1, 8, 11].

VI. CONCLUSION AND FUTURE WORK

The need of extracting useful knowledge from the dataset leads to have a complete dataset before mining. This dataset must not contain any missing value. Thus, imputation methods are widely used to fill the missing values of different kinds of datasets. In this survey, the overall views on the imputation methods and their categories are discussed. Thus it can be clearly seen that many methods are proposed for handling missing values present in the dataset. The brief description of each method is given. Further, these imputation methods are compared along with their advantages and

disadvantages. There are also some other methods that are not included in this study. The comparative study of them is the future work here.

ACKNOWLEDGMENT

We would like to thank to the reviewers of our paper. We are thankful to the seminar course of School of Computing Science and Engineering (SCSE), VIT University, Vellore, Tamil Nadu, India.

REFERENCES

[1] Chih-feng Liu, Thao-Tsen Chen, Shie-Jue Lee, "A Comparion of approaches for dealing with missing values", in ICMLC, July 2012, pp. 1576-1582.

[2] Y. Zhang, C. Kambhampati, D. N. Davis, K. Goode, J. G. F. Cleland, "A Comparative Study of Missing Value Imputation with Multiclass Classification for Clinical Heart Failure Data", in IEEE 9th FSKD, 2012, pp. 2840-2844.

[3] A. K. Tanwani, M. J. Afridi, M. Z. Shafiq, M. Farooq "Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets," EvoBIO 2009, pp. 128-139.

[4] Ms.R.Malarvizhi, Dr.Antony Selvadoss Thanamani, "K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation", in IOSRJCE, Vol. 6, 2012, pp. 12-15.

[5] J. Grzymala-Busse, L. Goodwin, W. Grzymala-Busse, and X. Zheng, "Handling missing attribute values in preterm birth data sets," in 10th RSFDGrC, 2005, pp. 342-351.

[6] N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland, "Handling Missing Values in Data Mining - A Case Study of Heart Failure Dataset", in IEEE 9th FSKD, 2012, pp 2934-2938.

[7] Kehar Singh, Dimple Malik and Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", in IJCEM, April 2011, Vol. 12, pp 105-109.

[8] B. Mehala, K. Vivekanandan, P. Thangaiah, "An Analysis on K-means Algorithm as an Imputation Method to Deal with Missing Values", in AJIT 2008, Vol. 7, pp. 434-441.

[9] Shichao Zhang, Xindong Wu, Manlong Zhu, "Efficient Missing Data Imputation for Supervised Learning", in IEEE 9th ICCI, 2010, pp. 672-679.

[10] Jing Tian, Bing Yu, Dan Yu, Shilong Ma "A Fuzzy Clustering Approach for Missing Value Imputation with Non-Parameter Outlier Test", in BUAA, 2012, Vol. 4, pp. 33-42.

[11] N.A. Setiawan, P.A. Venkatachalam and A.F.M. Hani, "A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set", in Springer IFMBE Proceedings 2008, Vol. 21, pp. 266-269.

