*Available online at www.ijiere.com*

**International Journal of Innovative and Emerging Research in Engineering**

e-ISSN: 2394 - 3343 e-ISSN: 2394 - 5494

# Analysis of Various Techniques to Handling Missing Value in Dataset

Rajnik L. Vaishnav [a], Dr. K. M. Patel [b]

[a] School of Engineering, RK University, Rajkot, India, rajnik.vaishnav55@gmail.com +91-9033319249.

[b]Associate Professor, School of Engineering, RK University, Rajkot, India, kamlesh.patel@rku.ac.in, +91-9099240604

**ABSTRACT:**

**Data mining has made a great progress in recent year but the problem of missing data or value has remained great challenge for data mining. data mining is the field of studying experimental data sets for the discovery of interesting and potentially useful relationships. Missing data or value in a datasets can affect the performance of classifier which leads to difficulty of extracting useful information from datasets. There are a number of alternative ways of dealing with missing data. Several methods like Listwise Deletion, Pairwise Deletion, Mean Imputation, Regression Imputation, K-Means Imputation(KMI), Fuzzy K-means clustering Imputation (FKMI), Support Vector Machine Imputation (SVMI) for imputation of missing values using available values in the data set. In this study, different methods are reviewed and compared with their advantages and disadvantages**

**Keywords: missing value, data mining, KMI, FKMI, SVMI**

## 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. Data Integrity is the foremost aim of database missing data helps to degrade the integrity and to avoid the degradation or to improve or to maintain the data integrity missing data to be imputed properly. Handling of imputation causes the three major issues 1. Loss of information, as a consequence, a loss of efficiency. 2. Data handling is an issue, computation and analysis due to irregularities in the data structures. 3. Systematic difference among the data.[3]

Missing data are the absence of data items for a subject; they hide some information that may be important. In practice, missing data have been one major factor affecting data quality. 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 data quality. Missing data imputation techniques refer to any strategy that fills in missing values of a data set so that standard data analysis methods can be applied to analyze the completed data set.

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. 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 .[6]

Missing data randomness is classified in three classes.

**Missing completely at random (MCAR):** Missing values are scattered randomly across all instances. In this type of randomness, any missing data handling method can be applied without risk of introducing bias on the data. This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can be applied without risk of introducing bias on the data.[2]

**Missing at random (MAR):** Missing at random (MAR) is a condition, which occurs when missing values are not randomly distributed across all observations but are randomly distributed within one or more . When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself;[8]

**Not missing at random (NMAR):** When the probability of an instance having a missing value for an attribute could depend on the value of that attribute.. It is also called as non-ignorable missingness. The probability of an instance with a missing value for an attribute might depend on the value of that attribute.[9]

## II. PREVIOUS WORK

In this section we would like to share the survey experience about imputation Techniques like Mean- Mode imputation, Hotdeck imputation, K-nearest neighbour's imputation, multiple imputation, Multivariate imputation by chained equations (MICE).
**Mean-Mode imputation (MMimpute):** MMimpute filling the missing data by the mean or mode(qualitative) from all Know data set.

**Hotdeck (HD) imputation:** Given an incomplete pattern, HD replaces the missing data with values form input data vector that is closest in terms of the attributes that are Know in both patterns.. HD attempts to preserve the distribution by substituting different observed values for each missing .The similar method of HD is Cold deck imputation method which takes other data source than current dataset[3]

**K-nearest neighbor`s imputation (KNNimpute):** Training dataset incomplete pattern where missing data K will be selected with the help of known data such that they minimize some distance measure. Once the K nearest neighbours have been found, are placement value to substitute for the missing attribute value must be estimated. How the replacement value is calculated depends on the type of data.

**Regression Imputation:** Using regression method for imputation, the values from the features are observed and then predicted values are used for filling MVs[11].

**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.[9]

**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.

**Classification algorithm:** Classification is a supervised learning method. It means that learning of classifier is supervised in that it is told to which class each training tuples belongs. Data classification is a two step process. First step, a classifier is build describing a predetermined set of data classes or concepts. The data classification process has two phases, these are:- [7]

**Learning-** Classification algorithm analyzed the training data. Classifier is represented in the form of classification rules. This phase is also viewed as learning of a mapping.

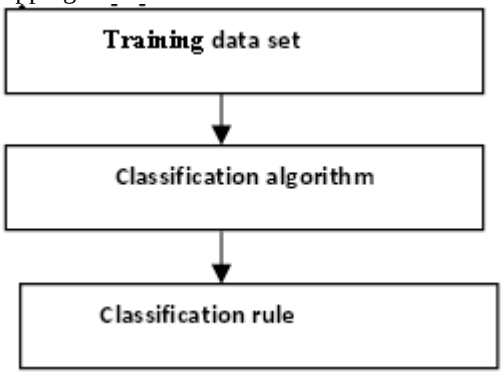


**Figure 1. Learning State**

**Classification-** To estimates the accuracy of classification algorithm test data is used. The rules can be applied to classification of new data tuples. Accuracy of a classifier on a given test set is percentage of test set that are correctly classified by classifier. The associated class labels of each test tuples is compared with learning classifier class prediction for that tuple.[14]

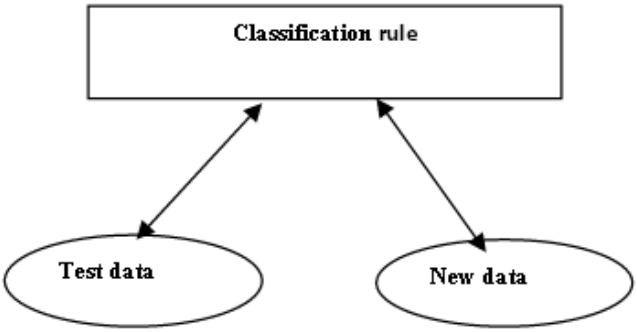


**Figure 2. Classification**

## III. THE TREATMENT OF MISSING VALUES

There are several methods for treating missing data, some methods are described below. Missing data treatment methods can be divided into three categories, as proposed in[6]

### A. Ignoring and discarding data

There are two main ways to discard data with missing values. The first method is known complete case analysis it is available in all statistical programs and is the default method in many programs. This method consists of discarding all instances with missing data. The second method is known as discarding instances and/or attributes. This method consists of determining the extent of missing data on each instance and attribute, and deleting the instances and/or attributes with high levels of missing data. Unfortunately, relevant attributes should be kept even with high degree of missing values

### B. Parameter estimation

Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data

### C. Imputation

Imputation method is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set assist in estimating the missing values. This work focuses on imputation of missing data.

## IV. CLUSTERING AND MISSING VALUE IMPUTATION

In this phase, clusters will be generated by passing every record in every block as the seed-point-record. The records in the training set will be formulated into number of clusters in which the missing-attribute records are the seed points. In most of the clustering models, the concept of similarity is based on distance. Here we explore a new approach for measuring the similarity, wherein a record is said to be similar to the seed-point- record only if it has the maximum weight. The training dataset is scanned for measuring the similarity. The standard deviation of the attributes is taken into account for measuring the similarity. The weights are assigned to the records based on the number of similar attributes. The cluster is then formed by grouping the records which have higher degree of weights. Finally, the missing attribute- value(s) in each cluster resulted by computing the mean value of the respective attribute(s) in the cluster and complete dataset is generated without any missing any data.[15]

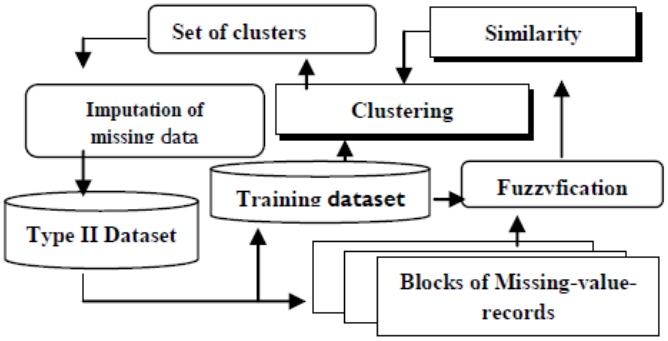


***Fig. 3**: Flow of clustering algorithm to impute the value.*

## V. PRONE AND CONS OF DIFFERENT TECHNIQUES TO HANDLE MISSING VALUE

| Techniques of Missing Value | Short Review | Prone | Cones |
|---|---|---|---|
| **Listwise Deletion** | - Deletion of cases containing missing values (entire row is deleted) High loss of information due to deletion of entire row High effect on variability Loss of precision and induce bias. | - Simple to use . | -Loss of huge data, loss of precision, high effect on variability, induce bias |
| **Pairwise Deletion** | - Deletion of records only from column containing missing values Less loss of information by keeping all available values Less effect on variability Less Loss of precision and induce bias[2] | -Simple, keeping all available values i.e. only missing values are deleted | -Loss of data, not a better solution as compared to other methods |
| **Mean Imputation (Most Common Imputation) (MCI)** | - Replace MVs with the arithmetic mean of data Resultant Mean and SD after imputation may be much higher than that of original Not a good substitution method.[2] | - 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. |
| **Regression Imputation** | - Replace MVs with the values predicted from observed values Regression Equation: $Y = \alpha 0 + \alpha 1 X$ | -Calculated data saves deviations from mean and distribution shape | - Degree of freedom gets distort and may raises relationship |
| **Expectation Maximization (EM)** | -Iterative method, finds maximum likelihood Two steps: Expectation (E step), Maximization (M step) Iteration goes on until algorithm converges | -Accuracy increases if the model is right | -Takes time for converging, very Complex |
| **KNN Hot deck Imputation** | -Use alike records to fill the MVs 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[6] | -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 |
| **K-Means Imputation (KMI)** | -Then KMI uses algorithm called nearest neighbour to impute the MVs in the same way as KNNI | -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. |
| **Fuzzy K-means clustering Imputation (FKMI)** | -Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values[8] | -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. |
| **Support Vector Machine Imputation (SVMI)** | -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 | -Efficient in large dimensional spaces. Efficient memory consumption | -Poor performance if number of samples are much lesser than number of features |

| | | | |
|---|---|---|---|
| | Attribute[10] | | |
| **Artificial Neural Network with Rough Set Theory(ANNRST)** | -Divided into : Reducing RST attribute and ANN construction for missing attribute value prediction[2] | -Generally ANNRST dataset yields better accuracy than CMCI and KNN classifiers. | -Complex computations and time Consuming. |

## VI. CONCLUSIONS

The need of extracting useful knowledge from the dataset leads to have a complete dataset before mining. This dataset contain less number of 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. Further, these imputation methods are compared along with their advantages and disadvantages.

## REFERENCES

[1] B. Mehala, P. Ranjit Jeba Thangaiah , and K. Vivekanandan "Selecting Scalable Algorithms to Deal With Missing Values ". International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009
[2] Dipak V. Patil., "Multiple Imputation of Missing Data with Genetic Algorithm based Techniques." IJCA Special Issue on "Evolutionary Computation for Optimization Techniques" ECOT, 2010
[3] 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 2012
[4] Naresh Rameshrao Pimplikar*, Asheesh Kumar, Apurva Mohan Gupta., " Study of Missing Value Imputation Methods – A International Journal of Advanced Research in Computer Science and Software Engineering 4(3), March - 2014, pp. 1487-1491.
[5] YamanishiLuis E. Zairate, Bruno M. Nogueira, Tadeu R. A. Santos and Mark A. J. Songe, "Techniques for Missing Value Recovering in Imbalanced Databases: Application in a Marketing Database with Massive Missing Data". 2006 IEEE International Conference on Systems, Man, and Cybernetics
[6] N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland "Handling Missing Values in Data Mining - A Case Study of Heart Failure Datasetl"2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).
[7] Preeti Patidar,Durga Toshniwal,"Handling Missing Value In Decision Tree Algorithm" ACEEE.
[8] Kaiser,"Algorithm For Missing Values Imputation In Categorical Data With Use Of Association Rules" The Ninth International Conference on Web-Age Information Management 2012 IEEE(2008).
[9] T.V Rajnikanth "An Enhanced Approach On Handling Missing Values Using Bagging K-Nn Imputation" 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 06,
[10] Sandeep Kumar Singh ,Archana Purwar "Empirical Evaluation of Algorithms to impute Missing Values for FinancialDataset" 2014 IEEE International Conference
[11] Y. Kou, C.-T. Lu, and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining,pp. 614–618, Bethesda, Maryland, USA, 2006.
[12] Jerzy W. Grzymala-Busse and Ming Hu "A Comparison of Several Approaches to Missing Attribute Values in Data Mining". W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp. 378−385, 2001. □ Springer-Verlag Berlin Heidelberg 2001
[13] Dan Li, Jitender Deogun, William Spaulding, and Bill Shuart "Towards Missing Data Imputation:A Study of Fuzzy K-means Clustering Method". Springer-Verlag Berlin Heidelberg 2004
[14] Feng Honghai, Chen Guoshun, Yin Cheng,Yang Bingru, and Chen Yumei "A SVM Regression Based Approach to Filling in Missing Values". Springer-Verlag Berlin Heidelberg 2005
[15] Shamsher Singh, Prof. Jagdish Prasad "Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods". W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp. 378−385, 2001. □ Springer-Verlag Berlin Heidelberg 2001