



## Missing Data Imputation for Medical Database: Review

Shalu Bhati, Manoj Kumar Gupta

Department of Computer Science and Engineering, Sharda University, Greater Noida,  
Uttar Pradesh, India

*Abstract: Data mining has made a great progress in almost every domain in recent years. But the problem of missing data and the values remain untreated has been a great challenge because of this it is the area for research for a quite long time. The missing values in a database can affect the accuracy and performance of the classifier which results in difficulty to extract the meaningful information, loss of efficiency. It may be very difficult to obtain the data quality mining results from the incomplete datasets, hence this missing gaps need to be treated. As the amount of data is increasing in the medical database also so there is a need for effective technique to extract the information. In this paper we explore the different measures and techniques involved in handling the missing values in datasets. And a review on the different techniques for missing value imputation for the medical database.*

*Key words- Data Mining, Missing data, K-means clustering, Regression Imputation, Multilayer perceptron.*

### I. INTRODUCTION

Missing data creates various problems in analyzing and processing data in the database. Missing value is a very common problem in the data cleaning process and we need to have a strategy for handling them. Missing data is an issue associated with data mining research. Missing data occurs due to no attribute associated for any instance, or the values are not relevant or the values are not collected properly at the time of data has been subjected. The area of missing value imputation has been an area of research for over decades in decision making having more emphasis on statistical method. Missing data imputation is a problem of dealing with incomplete data or missing values in a specific data set or filling in gaps in a data set with the help of different techniques [2]. The easiest technique to remove the instances containing missing values in their attributes. Other way is by estimating the parameters such as variance and covariance depend on complete data using maximum likelihood methods. Or the missing values are calculated by analyzing the relationship within the attributes [9]. The data set analyzing has become more competitive as the amount of data has been increasing. The big data aimed at improving the public healthcare system by clinical decision support system, Public sector, Retail industries, business modeling and marketing etc [4]. Now, the problem here comes to handling the issues in processing the datasets i.e. data cleaning which is a step of KDD knowledge discovery from data by performing some operations to extract the meaningful information. Out of these issues there is an important issue handling the missing values in a large database i.e. missing data imputation [4]. "Missing values are very common occurrence in a real world database, and statistical methods have been developed to deal with this problem referred to as missing data imputation"[10]. Some machine learning methods are also developed for handling this problem. But the amount of data is increasing in the medical database so there is a need for effective technique to extract the information. This missing value occurs in the database due to various reasons such as manual entry mistakes, equipment errors and measurements errors this leads to decrease the performance of classification and address some complexities in the database which results in difficulty for the outcome of useful information. All of these data help in analyzing complex clinical decision support system, personalized medicine, analyze disease patterns etc the paper presents an overview of missing data and the strategies involved in dealing with the missing values. On the other part, a review on the missing value imputation for the medical database has been done. The problem of missing data is very common occurring problem in various types of databases. The handling and managing of missing data mechanism is classified in three classes:-

- **Missing completely at random:** In this type of randomness the missing data occurs completely At the random, it occurs when the probability of a case having missing value for an attribute does not depend on the known values or the missing data. In this, any of the data treatment method can be applied without the risk of bias result.

- **Missing at Random:** In this type of randomness the missing value is not randomly distributed among all the observations but are randomly distributed within one or more. The probability of a record containing missing value for an attribute depends on the known value but not on the values of missing data itself [11].

- **Not missing at random:** This type of mechanism is known as not missing at random. It is also considered as Non-ignorable. It occurs when the probability of an instance containing the missing value doesn't depend on the value of the attribute. We can solve this by going back to the source data and then analyze more information about the mechanism [13]. In order to analyze the best imputation techniques from various experiments are done to choose the best one to handle missing value in a data set. Here are some approaches that have been empirically assessed to find the missing values [10].

- **Case deletion:** - In this technique, the entire row or column is deleted having missing value attributes.

- **Most common method:** - In this the missing value is replaced by the mean of the known attributes in the data set. It is only applicable for the numeric attributes.
- **Concept most common:** - This technique is quiet similar to previous method, but it considers only the same class in which MV is missing.
- **Treating missing attribute as special value:** - This is totally different approach. Instead of finding some new value, the missing value is considered itself a special value for the instance that containing missing value.
- **Closest fit:-** This method based on replacing a missing attribute value with an existing attribute from the other case of the same attribute that resembles as much as possible the case with missing attribute values.
- **K-nearest neighbor:** - This method finds the K-nearest neighbors, and among all neighbors the most common value is considered for nominal attributes.
- **Weighted imputation with K-nearest neighbor:** - In this the distance of each missing value instances from its neighbors is calculated. Here, the distance used for calculating weight. Missing values are calculated by weighted mean for numerical attributes.
- **K-means clustering imputation:** - This method is similar to "K-Nearest Neighbor" the instances are clustered by using K-means clustering. And the instances in each cluster are considered nearest neighbors of each other.
- **Fuzzy k-means clustering imputation:** - This method is better than K means imputation. Here the data object can be a part of more than one cluster centre. It is best method for overlapping data. In this unreformed attributes for every - uncompleted data are substituted. Missing values are calculated by weighted sum of all entroids on the basis of membership degrees.
- **Support vector machines imputation:** - This method is efficient in memory consumption and large dimensional spaces. This model is used to predict missing attributes from the complete instances which do not have missing values. Here the condition attributes and decision attributes are considered.
- **Local least squares imputation:** - This fits a least squares model between the instances and known part of the record with missing values.
- **Regression Imputation:** - In this method the missing values are first observed and the predicted values are used for handling the missing values
- **Hot-Deck Imputation:** - This is a traditional method failed to give a complete simulation associated with the missing data. In this type of method each missing value is replaced with an observed response from a similar unit.
- **Missing Data Using Neural Networks-** A neural network which is an information processing paradigms that is inspired by the biological neural nervous system it consist of four parts- processing units having a certain activation level at any point in time. The activation of one unit leads to the input for another which is resolved by weighted interconnections between different processing units. It works on an activation rule and a learning rule which specified how the weights are adjusted for a given input/output pair. They have the ability to analyze meaning from complicated data and also used to extract patterns. They model the various non-linear applications because as their ability to adjust to a non linear data networks [7].

## II. RELATED WORK

The missing data estimation can be done using theory of dynamic programming with neural network and genetic algorithm. The concept of optimality and bellman's equation for missing data estimation has been used. The genetic algorithm used to solve optimization problems [17]. The problem is handled step by step, acquiring the optimal solution for each step, keeping tracks of recurrence and overlaps. This is advantageous as it improves its performance. Here the single classifier is used in the base model. The field of missing data researches also took of machine learning learned by a re-distributing control problem using dynamic programming. The estimation of missing data with the help of some tools in machine learning those investigate the problem by using the decision trees. They evaluate this method by using dynamic programming which includes two types of approaches [2]:-

Top down approach- The dynamic programming approach breaks the problem in sub-problem and then solution is acquired for missing values and stored them for future use this approach is called as Top-down approach.

Bottom up approach- In this type of approach, the sub-problems are anticipated and solved beforehand these approaches are usually used to solve the bigger problem called as Bottom-up approach. But it is also not possible to solve all the problems beforehand.

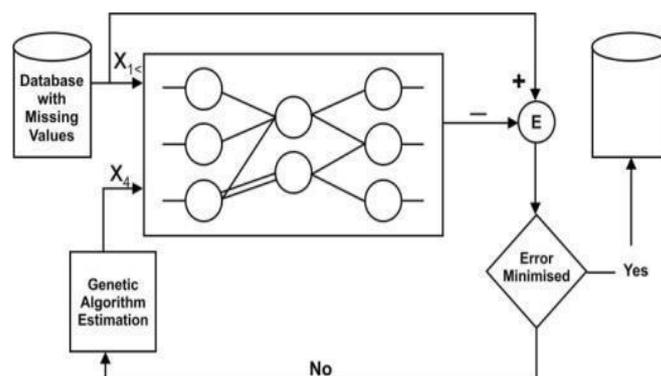


Fig.1 Auto-associative neural network and genetic algorithm model for missing data imputation.

Another model for missing value imputation introduced using statically methods such as hot-deck, regression techniques and machine learning methods such as artificial neural networks, self-organized map and k-nearest neighbor. This method based on machine learning is best suited for missing value imputation rather than statistical method. Statistical method need enhancement for gaining maximum accuracy [15]. The patterns are classified by using multi-task learning perceptron. It is also a method of dealing with missing values in a dataset which are used in real-world classification problem.

There is a need of additional research to develop robust clinical decisions problems support system. So the three techniques used for filling the missing value is by filling the missing value with zero, or by replacing the missing value with mean or using multiple imputation methods. The analyzing of differences between single imputation and multiple imputations methods to copy the different levels of missing data operated by an artificial neural network trained on incomplete dataset [12].

The technique for data preparation using missing data analysis in neural modeling for complex data analysis has been introduced. Although, the neural network analysis is an important task for data preparation. As some of the neural network existing before the data preparation but novel integrated data preparation approach with the help of neural network analysis have been introduced. These also provide some intelligent solution technique for data preparation and finding the solution for the issues involved in the integrated scheme. Here cost benefit analyzing technique for the complex data analysis. The enhancement for neural network models is made the complexities for the neural networks have been reduced which helps in complex data analysis. This is beneficial for neural network analysis for preparation of data. The integrated data preparation approach is made as promising results for improving the performance of neural data analysis [6].

Another technique for the missing data imputation present at completely at random using multi layer perceptions which provide a methodological framework for the development of an automated data [3]. The three traditional imputation procedures are mean/ mode, regression and hot-deck. The MLF requires less skills and efforts than a self – organized map because it performs better than regression and nearest neighbor. The component analysis and genetic algorithms have two architectures to handle the missing values based on the auto associative neural networks and principal component analysis.

A new approach introduce for managing medical database to improve modeling performance. This also demonstrated for improving modeling performance in a simulated test bed. As the two popular methods for analyzing missing data are to impute or delete but they cause bias results. Here, the researchers used a fuzzy classifier followed by fuzzy modeling to predict which missing data have to impute and which doesn't. Here, the statistical classifier is used. When the data in a medical database is stored the time is recorded in a database sometimes the time series is misaligned when the samples are not registered with same sampling time. These are in the form of evenly and unevenly misalignment. Therefore for handling the misalignments the two techniques are used for templating and gridding. This technique is useful for managing the missing data is an essential and useful step for the enhancement of predictive risk modeling [12].

A methodological research have focused on two type of missing data analysis i.e. multiple imputation and maximum like hood these techniques have advantage as they overcome the pitfalls of traditional techniques. The main focus is on maximum like hood estimation and present two analysis examples from the longitudinal study of America youth data. Here, the auxiliary variables are included to made description for one of these examples. And the underpinnings of missing data analysis have been explained [16].

The clinical data often contains missing value as a common problem and imputation is the excellent method to deal with the disadvantages occurred due to the missing values in a data mining task. Here the comparison of various imputation techniques by analyzing their performance when applied to different types of classification algorithms. The result concludes that there is not a single imputation method which can perform excellent for all classifier. Here the aim is to investigate the behavior of different missing value techniques when they applied for the preprocessing for multiclass classification. And to find suitable imputation method for development of predictive model involves in managing the dataset of heart failure patients. Furthermore result may be enhanced by adjusting the parameters utilized in the classification procedure [5].

The presence of missing gaps in the database affects the interpretation of a classifier developed by using the dataset as a training sample. Different approaches have been introduced for the handling of absent data and the one which is more recently deleted from the instances that containing at least one missing value of a feature. Here the experiments are carried out with twelve datasets to evaluate in the consequences of the misclassification error rate of four procedures to deal with missing values that is case deletion techniques, in which the entire instance is deleted containing the missing attribute, mean imputation, where the gap is filled by mean of the presented value median imputation and KNN imputation method. The classifier used here is the linear discriminated analysis (LDA) and KNN classifier. The two types of classifiers parametric classifier and nonparametric classifier have been used. The main aim is to evaluate the difference between both the classifiers parametric and non parametric .To evaluate the affects of missing data imputation on the accuracy of the classifier the work have been done only on the relevant classifier. This also enhances the speed of the imputation procedure. The most relevant features are selected using the RELIEF this is a process used in feature selection supervised classification. This method is mostly applicable on the large database rather than on small database [1].

A novel hybrid prediction model analyzing the various imputation techniques using simple K-means clustering and apply best one to the data set. This is the first time the simple k-means is used with the combination of multilayer

perceptron. This is very useful in prediction when the number of missing value is large in an instance. It supports medical decisions with pattern extraction and multilayer perceptron.

Here the evaluation has been using three medical dataset. Further the focus is on the improving and testing the multiclass imbalanced classification problems. Multilayer perceptron is a feed forward artificial neural network. The prediction has been done in three stages selecting the best imputation technique by analyzing 11 imputation techniques using simple k-means clustering on the database. Then the patterns extracted and apply multilayer perceptron using backpropagation as training algorithm and the evaluating the performance measures of classifier that is accuracy, sensitivity, specificity [10]

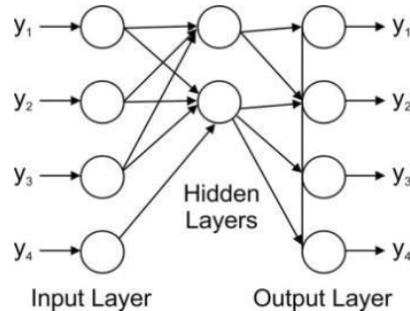


Fig. 2 Multilayer Perceptron

A concept that is mainly applied to optimal control problems, because the control problems decisions are mostly taken with partial information about a particular condition in the system [8]. The experimental setup is carried out by the multi-layer perceptron consists of multiple layers of nodes in a directed graph and each layer is connected to the next layer. It is made up of an input layer, hidden layer, output layer. Each node is made up of a neuron with a non-linear activation function. Here the supervised learning is used called as back propagation. In multilayer perceptron the output of hidden layer is considered as input to output layer the output of the input layer is calculated by using the sigmoidal function. And then the output layer does the prediction by using the instances of data set as "yes" and "no" for binary classification problem [10].

### III. CONCLUSION

For the extraction of useful information from the dataset we need to have complete datasets first that is the dataset containing the missing values then the different techniques are analyzed to handle missing values.. This paper introduces the overview of the missing value imputation techniques and the strategies involved in handling them. Apart from been eliminating the rows or columns from the datasets containing the missing value attributes here the solution for them is described by analyzing different imputation techniques to enhance the data quality, accuracy and efficiency. The selection of missing value imputation technique depends on the type of data set and the structure of the attributes. Most of the missing data imputation techniques are mainly applicable to only numeric type of data, but the real world ata also contains the mixed, alphabetical, categorical data where some of the techniques failed.

### REFERENCES

- [1] Acuna, Edgar, and Caroline Rodriguez. "The treatment of missing values and its effect on classifier accuracy." *Classification, clustering, and data mining applications*. Springer Berlin Heidelberg, 2004. 639-647.
- [2] Nelwamondo, Fulufhelo V., Dan Golding, and Tshilidzi Marwala. "A dynamic programming approach to missing data estimation using neural networks." *Information sciences* 237 (2013): 49-58.
- [3] Silva-Ramírez, Esther-Lydia, et al. "Missing value imputation on missing completely at random data using multilayer perceptrons." *Neural Networks* 24.1 (2011): 121-129.
- [4] Bakshi, Kapil. "Considerations for big data: Architecture and approach." *Aerospace Conference, 2012 IEEE*. IEEE, 2012.
- [5] Zhang, Y., et al. "A comparative study of missing value imputation with multiclass classification for clinical heart failure data." *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*. IEEE, 2012.
- [6] Yu, Lean, Shouyang Wang, and Kin Keung Lai. "An integrated data preparation scheme for neural network data analysis." *Knowledge and Data Engineering, IEEE Transactions on* 18.2 (2006): 217-230.
- [7] Nelwamondo, Fulufhelo V., Shakir Mohamed, and Tshilidzi Marwala. "Missing data: A comparison of neural network and expectation maximisation techniques." *arXiv preprint arXiv:0704.3474* (2007).
- [8] Cogill, Randy, et al. "An approximate dynamic programming approach to decentralized control of stochastic systems." *Control of uncertain systems: Modelling, approximation, and design*. Springer Berlin Heidelberg, 2006. 243-256.
- [9] Farhangfar, Alireza, Lukasz A. Kurgan, and Witold Pedrycz. "A novel framework for imputation of missing values in databases." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 37.5 (2007): 692-709.
- [10] Purwar, Archana, and Sandeep Kumar Singh. "Hybrid prediction model with missing value imputation for medical data." *Expert Systems with Applications* 42.13 (2015): 5621-5631.

- [11] Horton, Nicholas J., and Ken P. Kleinman. "Statistical Computing and Graphics-Statistical Computing Software Reviews-Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *American Statistician* 61.1 (2007): 79.
- [12] Markey, Mia K., et al. "Impact of missing data in evaluating artificial neural networks trained on complete data." *Computers in Biology and Medicine* 36.5 (2006): 516-525.
- [13] Kumutha, V., and S. Palaniammal. "An enhanced approach on handling missing values using bagging k-NN imputation." *Computer Communication and Informatics (ICCCI), 2013 International Conference on.* IEEE, 2013.
- [14] Cismondi, Federico, et al. "Missing data in medical databases: Impute, delete or classify?." *Artificial intelligence in medicine* 58.1 (2013): 63-72.
- [15] Jeraz, J.m., Molina I., Garcia, P.j albila, Ee, Ribelles, n., martin.m & Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem" *Artificial intelligence in medicine*, 50, 105-115,2010.
- [16] Araldi, a.n. & Enders, c.k. " An introduction to modern missing data analyses" *Journal of School Psychology*, 48, 5-37,2.010.
- [17] Patil, Dipak V., and R. S. Bichkar. "Multiple imputation of missing data with genetic algorithm based techniques." *IJCA Special Issue on" Evolutionary Computation for Optimization Techniques* (2010): 74-78.