International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.13513

Scopus

Original Research

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Imputation as a Technique for Enhancing the Quality of Medical Data

Vinutha MR[1], Chandrika J[2]

[1]Assistant Professor, Department of Information Science and Engineering, Malnad College of Engineering, Hassan, Karnataka, India; [2]Professor, Department of Computer Science and Engineering, Malnad College of Engineering, Hassan, Karnataka, India.

ABSTRACT

Introduction: In the field of Medical data analysis Data Mining plays an influential role. We should be capable of extracting fruit-bearing information from wealthy medical data. Extraction of effective information from wealthy medical data and making the valuable decision for predicting the diseases increasingly becomes necessary. Missing data or incomplete data pose a great problem in analysis. There is a good number of traditional methods available for taking care of data cleaning.

Objective: In this paper, we have attempted to throw light on various methods/tools existing for data cleaning. In our work we have imputed the missing values using different machine learning techniques and also have performed a comparative study of different machine learning techniques used.

Methods: A total of five hundred records of liver cirrhosis patients is collected. Two tasks have been carried out here, one is imputing missing values and the other is finding classification accuracy. The data set with no missing values for the predictor variables are used to generate the regression equation. In Random forest multiple decision trees were built and then these trees are merged to get the more accurate and stable prediction.

Results: We observed accuracy of class prediction before imputing the missing values and after imputing the missing value by using different algorithms.

Conclusion: It is noticeable that the accuracy of class prediction is high when missing values are handled properly. Also, the efficiency of class prediction is very high when the random forest is used both for imputing the missing values as well as for predicting the class.

Key Words: Decision Tree, K-NN Imputation, Linear Regression, Pre-Processing, Random Forest

INTRODUCTION

Medical diagnosis is extremely important which has to be carried out with a lot of care. Medical diagnosis should be performed conclusively. In the field of medicine, data mining is used to make prognosis, diagnosis and decision making. Predictions made should be Precise. Sometimes undesirable clinical decision risks the life of an individual. So there should be no compromise of the decision taken towards the health of the patient. For this one should have quality data. Any noisy, defective, inappropriate data may lead to defective results. Clinical data sets[1] have constituted a unique challenge, especially for data mining algorithms. Some Medical data set has large dimensionality, data may be noisy, there may be multiple classes, multiple values, and irrelevant values may be because of human errors. Data preprocessing has to be carried out with utmost care and also data preprocessing is an important step in KDD process. One simple way of taking care of missing value is imputation. The most common imputation method is the mean imputation. However, data mining techniques can also be used for taking care of missing values in an appropriate manner. Different data mining techniques like association rules, clustering methods can be used to predict the missing value. Some of the other imputation methods are ANN imputation, K-NN imputation, Regression imputation, Mean Imputation, Hot Deck Imputation.

Usha et al.[1] Focus on handling data quality problems using data mining techniques. The common sources of errors are lexical errors, syntactical errors, irregularities, duplicates and data entry problems. Researchers proposed a new algorithm

Corresponding Author:
Ms. Vinutha MR, Assistant Professor, Department of Information Science and Engineering, Malnad College of Engineering - 573201, Hassan, Karnataka, India; Email: mrv@mcehassan.ac.in

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 18.09.2020 **Revised:** 02.11.2020 **Accepted:** 05.12.2020 **Published:** 03.03.2021

called HECT - Hybrid Error Correction Techniques which combines the concept of context-independent and context-dependent for data standardization and correction. In HECT researchers used the parameters such as occurrence relation, minimum support threshold, distance threshold, Levenshtein distance and modified Levenshtein distance.

Swapna et al.[2] Proposed an algorithm which constructs a decision tree for every attribute and missing values are to be replaced with the leaf node values. Decision trees consist of a set of nodes. Each node is a test for an attribute and the output is given as two branches from each node except for the leaf nodes. One branch will be associated with yes and the other branch with no. Researchers endorsed that they can have more quality data on which mining techniques can be applied for a quality analysis

Multiple imputation approach was developed by Houari.[3] This approach is based on sampling techniques to take care of missing values problems. Multiple imputation approach has two objectives. The first objective is to guess missing values, which is performed by using the most effective method. The second objective is to improve the efficiency of the data mining process. The method proposed by the researcher's works in 2 steps. The raw data sets are compared by taking into account heterogeneous data in the first step. Estimation of missing data is done in the second step and the outcome of the second step is the estimated value and this value is used to fill in the missing data.[3]

Krishnamoorthy et al.[4] Contemplated a new technique called Effective Data Cleaning -EDC. EDC aims to identify irrelevant instance and relevant instances from a very large data set. This identification is carried out through the degree of missing value. Then reconstruction of missing value is performed through its closest instance especially within the instance set. EDC comprises 2 methods. The first method carries out the task of identifying relevant instances (IRV). Missing values are reconstructed in the second method (RMV). The reconstruction of the missing value is based on the distance metric.

Mohammed et al.[5] Developed a system whose framework is based on ETL process - Extract, Transform and Load Processes. The data cleaning process is carried out by splitting into three smaller sub-processes. This process helps to reduce the complexity of data cleaning. Researchers urged that system developed by them is capable of detecting errors, perform data deletion and programmatically create a knowledge base for valid values and table merging functions.

Kavitha Kumar et al.[6] Applied Data Mining Techniques for cleaning the data set. Techniques such as Association Rule Mining used as context-dependent attribute correction and Clustering techniques used for context-independent attribute correction. The algorithms were applied to cardiology dataset. To generate the rules entire data set was used but to carry out the task of corrections of attributes only random samples were used.

Vateekul et al.[7] proposed an ITree or Imputation tree. First, a missing pattern tree is constructed, which is also a binary classification tree. Study of predictability of missingness is made. This is carried out by making all the observations. If there are any missing values in the terminal node or cluster then such missing values are estimated by a regression tree. Researchers stated that it is advantageous compared to numerous other imputation methods. The reason is ITree states that the missingness of a variable is influenced by the existence of other variables in the data set. It also gives the knowledge of missingness type by analyzing the data, rather than assuming.

Kusumasariet et al.[8] used Open Refine Tool. Various techniques were outlined by the researchers for profiling data. Data quality was assured through data profiling. The type of analysis for data profiling is data profile classification. Researchers carried out both multi-column analysis and single-column analysis. Single column analysis is carried out for checking data completeness. Single column analysis consists of checking for duplication of text, text pattern and null value. The multi-column analysis is been done using association rules and identifying correlation. The researchers have written the Metadata rules on metadata. Data profiling is performed using built-in features of Open Refine or custom regular expression or combination of both.

Widera et al.[9] presented a classification method that differentiates the issues of cleaning into two classes of problems. One class of problem is found during the process of cleaning of single data sources and another class of problem is with several data sources. Both classes of problems are then divided into schema level problems and instance-level problems. The single source instance-level problems may be missing values, misspelling, cryptic values, missed values, violated attribute dependencies. Multisource schema level problems are naming and structural conflicts. The prototype application has been implemented using Python programming language.

Katherine et al.[10] Proposed a framework called BIOAJAX. BIOAJAX uses principles of data cleaning to improve data quality. BIOAJAX is a data cleaning tool Kit for the biological information system. BIOAJAX is designed to enhance the quality of data at the schema level as well as at the data level. It adopts and modifies the conceptual operations. BIOAJAX was originally developed in the declarative framework AJAX.

Wang et.al. [11] Used Gibbs sampling for refining missing values. It is useful in analysing stochastic systems. It uses multi-variant distribution for generating random numbers.

To handle the missing values, simulation of data is achieved which is based on existing population information. Researchers stated that Gibbs sampling has better performance when large samples were taken. Missing values processing in Gibbs is good even though all the relevant data of related attributes are missing. When the Gibbs sampling is used it is assumed that the simulation data still comply with existing data's statistical properties.

After carrying out the extensive literature survey we have planned to make use of different machine learning techniques for imputing the missing values and to do the comparative study of different algorithms used.

Imputation

Imputation is a process of filling the missing value with some valid related value and then carrying out the analysis as if there were no missing values.[12] Numerous imputation methods are available.

Mean Imputation

Mean Imputation calculates the mean of the observed value of a particular feature. This value is used to fill in all the missing value. It is one of the simplest methods of all the imputation methods

Cold deck Imputation

A value is chosen systematically from an individual who has similar values on other variables.[13] Random variation can be removed.

Hot deck Imputation

In this method, missing value is filled in with a value of an estimated distribution for the missing value from the current data. The implementation of the hot deck is a two-step process. Data is partitioned into different clusters in the first step. In the second step associate each record of missing data with one cluster. Then missing values are filled in with the complete cases in a cluster by calculating mode or mean of the attribute, especially within the cluster.

MATERIAL AND METHODS

Data Set: A total of five hundred records of liver cirrhosis patients is collected and that is taken for the study. Forty-one attributes are considered such as Duration of alcohol consumption, Quantity of alcohol consumption, Triglycerides(TG), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), MCV, MCH, Polymorphs, Lymphocytes, Albumin, Globulin etc. Records with no missing values for any features is taken as a training data set and records with missing values or records for which the values have to be imputed are taken as testing data set. Two tasks have been carried out here, one is imputing missing values and the other is finding classification accuracy. We have used three different techniques for imputing the missing values and to find the efficiency of classification, they are decision tree, random forest and linear regression.

Linear Regression Imputation: The data set with no missing values for the predictor variables are used to generate the regression equation. Then this equation is used for predicting the missing values. The relationship between the variables that have been used in the imputation model is conserved.

Decision Tree: Decision tree classification technique[14] is based on the concept of splitting basis. The decision tree is a flow chart like structure where the classification of records is carried out by sorting the instances based on the values of attributes. Each node represents' an attribute, all branches of an intermediate node represent an outcome of the test, each leaf node represents' the class label. Construction of Decision Tree is processed in a top-down approach and a greedy method. The process of constructing the tree starts with training set recursively finding a split feature by maximizing some local criterion. Methods like Gini Index, Information Gain Ratio etc. can be used for finding the feature which best splits the data. The data set is split into two subsets. Records having all the values are taken as one subset and the other subset has the records with the missing values. Then the decision tree is used for constructing the model. This model is used for imputing missing values.

Random Forest: It is a supervised learning algorithm. In Random forest multiple decision trees are built and then these trees are merged to get the more accurate and stable prediction. Random Forest method take care of missing values in two ways they are: i) If there are missing values then such data points are dropped ii) If there are missing values and they are of numeric value type then those values are filled with mean value iii) If there are missing values and they are of categorical type then those values are filled with the mode value.

RESULTS

The data pre-processing is the first step and it is one of the most important steps in KDD. The quality results are completely dependent on quality data. Many times, data that is collected is of not good quality. The data collected may be incomplete, inappropriate, noisy data and redundant. So the collected data has to be cleansed. In this work, we have implemented a decision tree algorithm, random forest algorithm and linear regression for imputing the missing values

in the collected medical data set. A total of five hundred records of liver cirrhosis patients were taken.

First case: We have used decision tree for imputing the missing values. After imputing the prediction of the missing value of classes are carried out by using three different algorithms and the corresponding accuracy of classification is expressed in the table1 and graph1.

Second case: Random forest is used for imputing the missing values and class prediction is carried out by using three different methods and the corresponding accuracy of classification is exhibited in table 2 and graph 2.

Third case: Linear Regression is used for imputing missing values and class prediction is accomplished by using three different methods and the corresponding accuracy of classification is plotted in table 3 and graph 3

Table 4 and graph 4 shows the accuracy of class prediction before imputing the missing values and after imputing the missing value by using different algorithms.

DISCUSSION

The outcome of the study is, from table 4 and graph 4 it is noticeable that the accuracy of class prediction is high when missing values are handled properly. Also, the efficiency of class prediction is very high when the random forest is used both for imputing the missing values as well as for predicting the class. It is evident from the above tables that when the data values are not imputed the accuracy is comparatively less.

In this paper, an attempt is made to review the valuable work carried out by various researchers in imputing the missing values. We have summarized various approaches and tools used for cleansing the data. In this work, we have applied the Decision tree, Random forest, and Linear Regression algorithms for imputing the missing values. From the results shown in sec V, it is clear that the accuracy of predicting the diseases is completely dependent on the quality of the data collected. This study unfolds the importance of having good data. Disease prediction should be highly accurate. The incorrect prediction will have an unfavourable effect on patients. To achieve good accuracy of prediction it is very much important to have quality data. In our future work, we have planned to use the ensemble methods for imputing missing values on the large number of liver cirrhosis patient data set.

CONCLUSION

This work was focused on applying data mining techniques for imputing the missing values. In our work, we have used three different data mining techniques that is the random forest, decision tree and linear regression. For the liver data set that we have collected random forest performs considerably better than the decision tree and linear regression. However, still there is a scope for applying various other data mining techniques and do the comparative study and arrive at the best method for imputing the missing values.

REFERENCES

1. Usha T. Data Cleaning of Medical Datasets using Data Mining Techniques. Int J Adv Res Comput Comm Engg 2018;7(6):283.
2. Swapna S, Niranjan P. Data Cleaning for data quality. IEEE International Conference on Engineering and Technology. 2016;3(1): 56.
3. Rima H, Ahcene B. Handling Missing Data Problem with Sampling Methods, IEEE International Conference on Engineering and Technology, 2014.
4. Krisgnamoorthy R, Sreedhar KR. A New Approach for Data Cleaning Process, IEEE International Conference on Engineering and Technology, 2014; 3(1): 123.
5. Hasimah HM, Tee LK, Chee C.A Data Cleaning Framework for Patient Data. International Conference on Informationand Computer Intelligence. 2011;12(2):189.
6. Kavitha Kumar R, Chandrashekaran RM. Attribute correction -Data Cleaning using Association Rule and Clustering Methods. Int J Data Mining Knowl Mgmt Process 2011;2(1):781.
7. Peerapon V, Tree-based Approach to Missing Data Imputation, IEEE International Conference on Engineering and Technology 2009;2(3):834.
8. Adam W, Michal W, Daniel F. Data Cleaning of medical data sets, J Med Informa Tech 2004;8(3):123.
9. Tien FK, Fitria D. Data profiling for data quality Improvement with Open Refine, Int. conf. Info-Tech Sys. Inno, 2016; 4(5): 871.
10. Herbert KG, Gehani NH, Piel WH, Wang JTL, Wu CH. BIO-AJAX: An Extensible Framework for Biological Data cleaning. ACM Sigmod 2004: 33(2): 51-57
11. Wang Yi, Zhou Li. Processing of missing values using Gibbs Sampling. Third International Conference Measuring Tech Mech Auto 2011; 23(4):46-49
12. Gustavo EP. Batista A, Maria CM. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Info Comm Embedd System 2007;4(2): 562.
13. Sivagowry .S, Dr Durairaj.M and Persia.An et al. An Empirical Study on Applying Data Mining Techniques for the Analysis and Prediction of Heart Disease. International Conference on Information Communication Embedd System 2013;10(4):82-86.
14. Han J, Kamber M, Pei J. Data Mining -concepts and Techniques, Third Edition. 2016;561.

Table 1: Efficiency after imputing the missing values using decision tree

Methods used for Class Prediction	Method used for imputing missing values	Efficiency after imputing the missing values
Decision tree		78%
Random forest	Decision Tree	90%
Linear Regression		88%

Table 2: Efficiency after imputing the missing values using Random Forest

Methods used for Class Prediction	Method used for imputing missing values	Efficiency after imputing the missing values
Decision tree		85%
Random forest	Random Forest	93%
Linear Regression		82%

Table 3: Efficiency after imputing the missing values using Linear Regression

Methods used for Class Prediction	Method used for imputing missing values	Efficiency after imputing the missing values
Decision tree		85%
Random forest	Linear Regression	92%
Linear Regression		87%

Table 4: Efficiency before imputing and after imputing Missing values

Methods used for Class Prediction	Efficiency before imputing the missing values	Efficiency after imputing the missing values
Decision tree	75%	85%
Linear Regression	80%	87%
Random forest	88%	93%

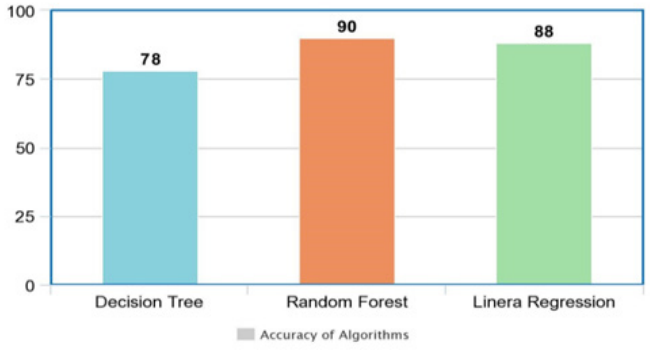

Figure 1: Efficiency after imputing missing values decision tree.

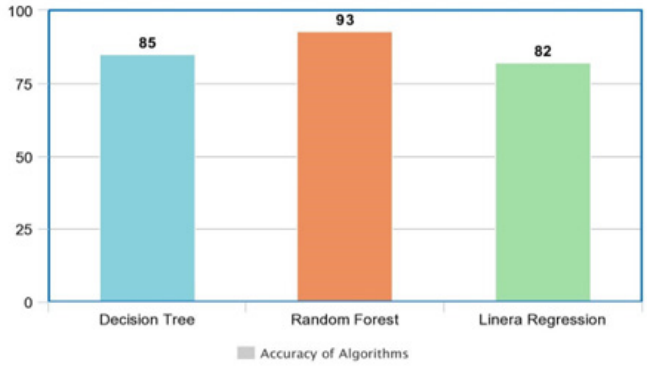

Figure 2: Efficiency after imputing missing values using random forest.

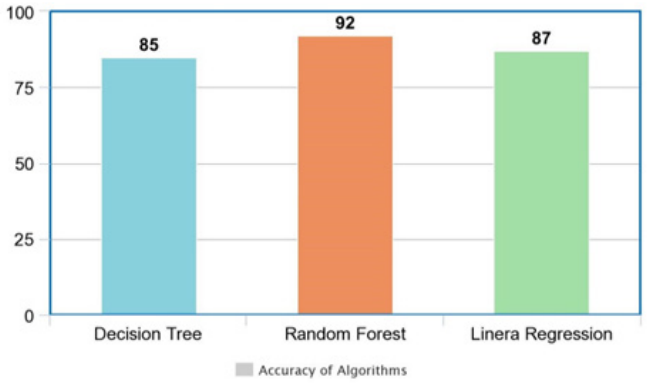

Figure 3: Efficiency after imputing missing values using linear regression.

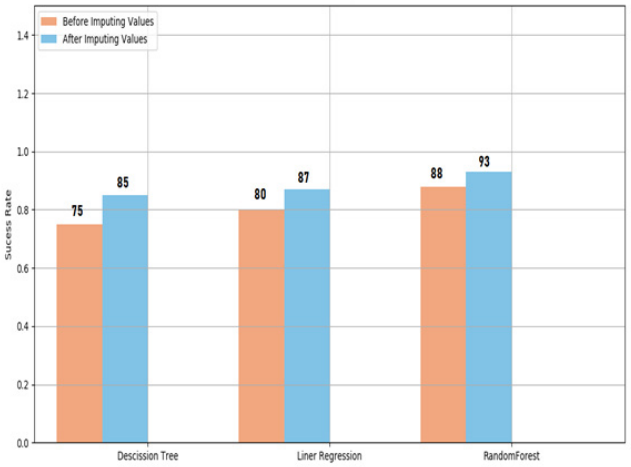

Figure 4: Success rate of algorithms.