IMPROVEMENT OF CLASSIFIER ACCURACY ON CLINICAL DATA SETS BY OPTIMAL SELECTION OF IMPUTATION METHODS

[1]C.UshaNandhini, [2]Dr. P.R.Tamilselvi,

[1]Research Scholar in Periyar University, Assistant Professor in Computer Applications,
Vellalar College for Women(Autonomous),Erode - 638012, Tamil Nadu, India.
[2]Research Supervisor, Assistant Professor in Computer Science,
Govt. Arts and Science College, Komarapalayam, Namakkal District,
Tamil Nadu, India.
[1]cushanandhini1970@gmail.com , [2]selvipr2003@gmail.com

Abstract: Missing value imputation is one of the biggest tasks of data pre-processing whenperforming data mining. Most clinical datasets are usually incomplete. Simplyremoving the incomplete cases from the original datasets can bring more problemsthan solutions. A suitable method for missing value imputation can help to producegood quality datasets for betteranalyzing clinical trials. In this paper we explore theuse of a machine learning technique as a missing value imputation method forincomplete cardiovascular data. Mean imputation, Group mean imputation, kNN imputation and Multi-Linear Regression Imputation are used as missing value imputation and the imputed datasets are subject to classification and prediction using C5.0 and Random Forest classifier. The experiment shows that final classifier performance is improvedwhen Multi-Linear Regression Imputation is used to predict missingattribute values for Random Forest and in most cases, the machine learningtechniques were found to perform better than the standard mean imputationtechnique.

Keywords: Missing value imputation - cardiovascular data - Mean imputation -Group mean imputation -kNN imputation -Multi-Linear Regression Imputation- C5.0 – Random Forest – Performance Measures.

I. INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses or other information repositories. The data stored in a database may reflect noise, exceptional cases, or incomplete data objects. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases. There are number of data preprocessing techniques. Data cleaning can be applied to correct inconsistencies in the data. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers in the data. Imputation 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 to assist in estimating the missing values. This paper focuses on imputation of missing data.

II. RELATED WORK

Rahman M. M. and Davis D. N. (2012) have investigated a Fuzzy Unordered Rule Induction Algorithm to predict the missing value and compared with imputation using machinelearning algorithms such as Decision Tree, SVM, KNN. The imputed datasets areclassified using decision tree, fuzzy unordered rule induction, KNN and K-Meanclustering. The experiment showed that final classifier accuracy is improvedwhen the fuzzy unordered rule induction algorithm is used to predict missingattribute values for K-Mean clustering and in most cases, the machine learningtechniques were found to perform better than the standard mean imputationtechnique [1].

Mohammad Al Khaldy, Chandrasekhar Kambhampati(2016) have presented six scalable imputation methods such as KNN, Expectation Maximization imputation (EM), K-mean imputation, Most Common Imputation (MCI), Concept Most Common Imputation (CMCI), Support Victor Machine (SVM) and are implemented on a Heart Failure dataset. The comparison is done by the performance metrics of three different classifiers namely J48, REPTree, and Random Forest. The results showed thatthe Random Forest classification achieves the best results in comparison to the decision tree J48 and REP Tree [2].

M.N.M. Salleh and N.A. Samat(2017) have proposed an imputation approachbased on the incorporation of FCM and PSO are used to findthe optimum value for finding the best value to replace the missing value in the dataset. In this paper, they have experimented no imputation, mean imputation, KNN imputation, and Fuzzy C-Means imputation along with proposed approach. The performance of the imputation methods were analyzed using Decision Tree classifier. The accuracy of Decision Tree resultsclearly showed that the imputed dataset using the proposed approach has improved compared to no imputation, Mean imputation, k-NNimputation and FCM method [3].

Dr. M. Sujatha, SallaAnusha&GundaBhavani(2018) have analyzed the missing values imputation in medical dataset by the proposed method IMVC. The experiment isconducted on Cleveland heart disease datasetusing IMV classifier. IMV classifier is used to imputemissing values in medical datasets through Naive Bayesclassifier, Neural Network classifier, C4.5 classifier. Results showed that the accuracy of neural network and C4.5 classifier is 87.3% [4].

S.Anitha&M.Vanitha(2019) have presented a comparison of four different types of imputation methods such as Mean, Singular Value Decomposition (SVD), K-Nearest Neighbors (KNN), Bayesian Principal Component Analysis (BPCA). Comparison was performed in the real VASA dataset and also evaluated the performance usingMean Square Error (MSE) and Root Mean Square Error (RMSE). While comparing the algorithms using the evaluation methods based on the real dataset, BPCA produced lower error rate than other methods [5].

Taeyoung Kim, WoongKo and Jinho Kim (2019) have applied four different missing value imputation for PV forecasting applications. The imputation methods experimented is Linear Interpolation (LI), Mode Imputation (MI),K-Nearest Neighbors (KNN) and Multivariate Imputation of Chained Equations (MICE). The results concluded that the most appropriate missing data imputation for application to PV forecasting is the KNN method [6].

Anil Jadhav, DhanyaPramod, and Krishnan Ramanathan(2019) evaluates performance of fourapproaches, for estimating missing values in numeric data sets namely meanimputation, median imputation, kNN imputation, predictive meanmatching, Bayesian Linear Regression (norm), Linear Regression,non-Bayesian (norm.nob), and random sample. They have used fivedifferent numeric datasets obtained from UCI machine learningrepository for analyzing and comparing performance of the dataimputation methods. Performance of the imputationmethodis evaluated using Root Mean Square Error (RMSE)method. The results of analysis showed that kNN imputation methodoutperforms the other methods. It has also been found that performanceof the imputation method isindependent of dataset and percentage ofmissing values [7].

AdityaSundararajan and Arif I. Sarwat(2019) have conducted statistical analyses to understand missing value imputation mechanism in data of a real grid-tied photovoltaic (PV) systemsat Miami. Theyhave compared the imputation performance of different methods: random imputation, multiple imputation using expectation maximization, *k*NN, and random forests and evaluated using error metrics.Imputed values are used in a multilayer perceptron topredict and compare PV generation with observed values. Results showed that among the six methods, *k*NN and random forests performed the best, followedclosely by multiple imputations using expectation maximization [8].

III. CLINICAL DATA SETS

In this paper, the following clinical benchmark data sets were collected from UCI repository and are subjected to imputation. All attributes are numeric valued.

1. Hungarian Institute of Cardiology, Budapest (hungarian.data)
2. University Hospital, Zurich, Switzerland (switzerland.data)

Table 1: Attribute Information

S.No	Attribute	Description
1	age	age in years
2	sex	sex (1 = male; 0 = female)
3	cp	chest pain type:- 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
4	trestbps	resting blood pressure
5	chol	serum cholestoral in mg/dl
6	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	restecg	resting electrocardiographic results- 0: normal,1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	maximum heart rate achieved

Received: 02 March 2020 | Revised: 21 April 2020 | Accepted:05 May 2020

9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment- 1: upsloping, 2: flat, 3: downsloping
12	ca	number of major vessels (0-3) colored by flourosopy
13	thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14	num	Label

The chol, fbs, ca predictors of Switzerland data set have more than 60% of missing value. Therefore they are subjected to remove for analysis. The predictors such asTrestbps,Restecg, Thalach, Exang, Oldpeak, Slope and Thal in Switzerland heart disease dataset have missing value in some of the objects. Similarly, In Hungarian heart disease dataset Trestbps, Chol, Fbs, Restecg, Thalach, Exang attributes have missing value in some of the objects. These attributes are subjected to imputation before applying classification and prediction.

IV. IMPUTATION METHODS

Imputation methods involve replacing missing values with estimated ones based on some information available in the data set. There are a variety of methods to substitute the missing value by imputation varying from naïve methods like mean imputation to some more robust methods based on relationships among attributes. This section surveys some widely used imputation methods, although other forms of imputation are available. In this paper, the author concentrated on fourestimation methods that are experimented.

a) Mean Imputation:

Mean imputation is a method in which the missing value on a certain variable is replaced by the mean of the available cases. This method maintains the sample size and is easy to use, but the variability in the data is reduced, so the standard deviations and the variance estimates tend to be under estimated. The magnitude of the covariances and correlation also decreases by restricting the variability and this method often causes biased estimates, irrespective of the underlying missing data mechanism.Let X^j_i be the j^{th} missing attribute of thei^{th} instance, which is imputed by

$$X^j_i = \sum_{k\in I}(X_k^j)/n$$

Where $\sum_{k \in I}(X_k^j)$the sum of values of j^{th}attribute having value other than missing attribute and n is the total number of instances of j^{th} attribute has values[9].

b) Group mean Imputation method:

The process for this method is the same as that for mean imputation. However, the missing values are replaced with the group (or class) mean of all known values of that attribute. Each group represents a target class from among the instances (recorded) that have missing values. Let $X^j_{n,i}$ be the j^{th} missing attribute of thei^{th} instance of the m^{th} class, which is imputed by

$$X^j_{m,i} = \sum_{k \in I}(X^j_{m,k})/n_m$$

Where $\sum X^j_{m,k}$ is the sum of values of set of m^{th} class of instances that has values in the j^{th} attribute and n_m is the total number of instances where the j^{th} attribute of the m^{th} class is not missing[9].

c) KNN Imputation Method

KNN is one of the simplest machine learning algorithm. Each missing values are imputed using the mean value from k nearest neighbors found in the training set. By default, a Euclidean Distance metric is applied to find the nearest neighbors.The missing value instance is approximated by selecting the most similar instances. K-NN is a lazy model, and its drawback is that this algorithm searches through all the dataset looking for the most similar instances, which is critical in the analysis of large datasets [10].

d) Multiple Linear Regression Method

Multiple Linear Regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.In this method, the functional relationship between multiple input variables and single or multiple target variables of the given data is represented in the form of a linear equation. This method sets attributes that have missing values as dependent variables and other attributes as independent variables in order to allow prediction of missing values by creating a regression model using those variables. For target variable Y_i, the multiple linear regressions with n predictor and m training instances can be represented as

$$Y_i = C + M_1X_{i1} + M_2X_{i2} + M_3X_{i3} + \ldots + M_nX_{in} \quad \{\text{for } I = 1\ldots m\}$$

V. SYSTEM ARCHITECTURE

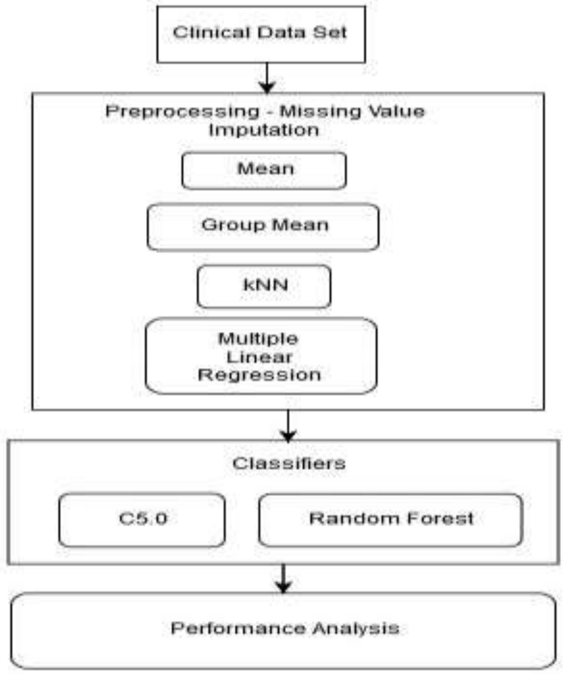


Figure 1. System Architecture of Proposed Methodology

VI. CLASSIFICATION ALGORITHMS

In this paper, two popular tree based classifiers namely C5.0 and Random Forest are applied to evaluate the performance of missing value imputation. The decision tree is one of the main methods of learning a classification applied across a wide range of problems. We chose the decision tree algorithms because they are the most commonly used techniques. The two decision trees selected here have different features. C5.0 is one of the most effective classification methods and Random Forest though giving high accurate results, has a tendency to be very slow.

The C5.0 algorithm has become the industry standard for producing decision trees and compared to more advanced and sophisticated machine learning models such as Neural Networks and Support Vector Machines, the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy.It uses the concept of entropyfor measuring purity. The entropy of a sample of data indicates how mixed the class values are; the minimum value of 0 indicates that the sample is completely homogenous, while 1 indicates the maximum amount of disorder. The entropy can be specified as

$$\text{Entropy (S)} = \sum_{i=1}^{c} -P^{i} log_2(P^{i})$$

In equation 1, for a given segment of data (S), the term c refers to the number of different class levels, and p_i refers to the proportion of values falling into the class level i. One of the benefits of this algorithm is that it is opinionated about pruning; it takes care of many of the decisions automatically using fairly reasonable defaults. Its overall strategy is to post prunethe tree. It does this by first growing a large tree that overfits the training data. Afterwards, nodes and branches that have little effect on the classification errors are removed.Random Forest consists of a large number of individual decision trees that operate as an ensemble. It creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. The procedures followed in this algorithm are:

1. Randomly select "n" features from total "k" features from the given data set where n< **k**
2. Construct a decision tree using best split point.
3. Repeat steps (1) and (2) until 'm' number of trees has been reached.
4. Obtain the prediction result from every decision tree.
5. Voting will be performed for every predicted result.
6. Select the most voted prediction result as the final prediction result.

VII. EXPERIMENT AND RESULTS

In this experiment, the medical data related to heart disease is considered because the heart disease is one of the leading causes of death in human. The performance evaluation of the imputation methods and classification algorithms described in the previous section are conducted using actual datasets taken from the UCI machine learning repository which is publicly available. The approaches are experimented using R tool. In this study, C5.0 and Random Forest are chosen to analyze the heart disease datasets and Random forest provides better accuracy for medical data sets than C5.0. With an intension to find out whether the same imputation method may lead to best accuracy for various data sets of same domain, various experiments are conducted on two different heart disease datasets. The results are compared and analyzed. The performances of the classifiers are analyzed in terms of accuracy, precision, recall and f-measure. A confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes.

Accuracy is the percentage of test tuples that are correctly classified by the classifier[11].

Accuracy = (TP+TN)/(TP+TN+FP+FN).

Precision is a metric that quantifies the number of correct positive predictions made. That is the proportion of positive identifications was actually correct

Precision = (TP)/(TP+FP)

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. That is the proportion of actual positives was identified correctly.

Recall = (TP/(TP+FN)

F-Measureis a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

The highlighted best classifier method corresponding to a particular imputation method for the two data sets is posted in the following tables 2 and 3. By comparing all the imputation methods with two classifiers, the Random Forest classifier on multiple linear regression imputation method with the accuracy of 84.93% is best of Hungarian heart patient dataset and 100% for Switzerland heart patient dataset.

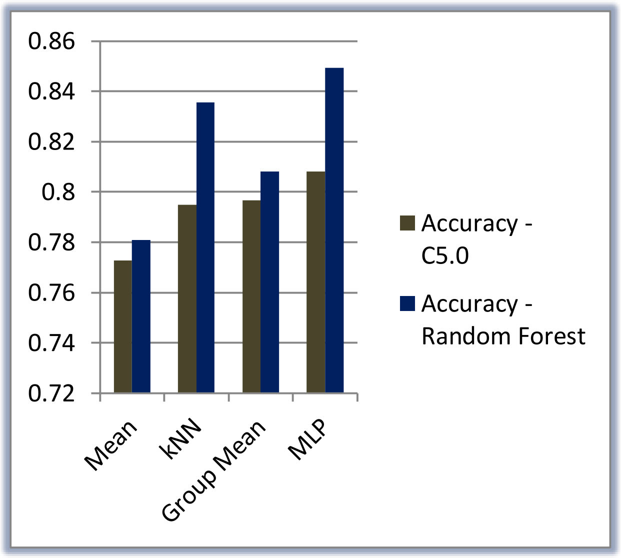
Table 2: Performance of Hungarian Heart Patient Dataset

Classifier	Imputation Methods	Accuracy	Precision	Recall	F-Measure
C5.0	**Mean**	0.7727	0.8333	0.8036	0.8181
	kNN	0.7948	0.8030	0.9463	0.8688
	Group Mean	0.7966	0.7708	0.9737	0.8605
	Multiple Linear Regression	**0.8082**	**0.9024**	**0.7872**	**0.8409**
Random Forest	**Mean**	0.7808	0.8444	0.8085	0.8261
	kNN	0.8356	0.8888	0.8511	0.8696
	Group Mean	0.8082	0.8367	0.8723	0.8542
	Multiple Linear Regression	**0.8493**	**0.9091**	**0.8511**	**0.8791**

Table 3: Performance of Switzerland Heart Patient Dataset

Classifier	Imputation Methods	Accuracy	Precision	Recall	F-Measure
C5.0	**Mean**	0.9594	0.75	0.6	0.6667
	kNN	0.9459	0.6667	0.4	0.5

	Group Mean	0.9594	0.6667	0.8	0.7273
	Multiple Linear Regression	**0.9729**	**1**	**0.6**	**0.75**
Random Forest	**Mean**	0.9459	0.6	0.6	0.6
	kNN	0.9729	1	0.6	0.75
	Group Mean	0.9729	1	0.6	0.75
	Multiple Linear Regression	**1**	**1**	**1**	**1**



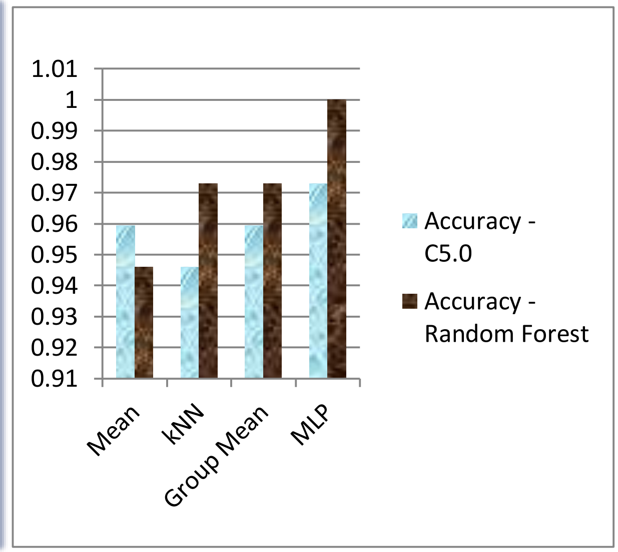


Figure 1.Accuracy of Hungarian Dataset Figure 2. Accuracy of Switzerland Dataset

The figure 1 shows the accuracy of Hungarian Dataset and figure 2 shows the accuracy of Switzerland Dataset.

VIII. CONCLUSION

Accuracy is most important in the field of medical diagnosis to diagnose the patient's disease. Experimental results show that an association between the performance of classification algorithms and the characteristics of missing data. We conclude that the classifier accuracy has been greatly enhances the accuracy of classificationby the use of any of imputation methods.The factors affecting the performance of classification algorithms were identified as follows: characteristics of missing values, dataset features, and imputation methods. Moreover, we assume that the chosen imputation method regulates the interconnection between these factors. Using benchmark data we found that several factors were significantly associated with the performance of classification algorithms. First, the results show that the missing data ratio ispositively associated with the performance of the classification algorithms. Second, we observed that the number of missing values in each record was more sensitive in affecting the classification performance than the number of missing cells in each feature. The results of this study suggest that multiple linear regression approach is the optimal selectionof the imputation method according to the characteristicsof the dataset improves the accuracy ofcomputing applications.

REFERENCES

1. Rahman, M. M. and Davis, D. N. (2013) “Machine Learning-Based Missing Value Imputation Method for Clinical Datasets”, IAENG Transactions on Engineering Technologies, Springer Netherlands, 245-257.
2. Mohammad Al Khaldy (2016) “Performance Analysis of Various Missing Value Imputation Methods on Heart Failure Dataset”, SAI Intelligent Systems Conference 2016, IEEE, September 20-22, 2016 | London, UK
3. M.N.M. Salleh and N.A. Samat(2017),” An Imputation for Missing Data Features Based on Fuzzy Swarm Approachin Heart Disease Classification”, © Springer International Publishing AG 2017, Y. Tan et al. (Eds.): ICSI 2017, Part II, LNCS 10386, pp. 285–292, 2017.
4. Dr. M. Sujatha , SallaAnusha and GundaBhavani(2018), “A STUDY ON PERFORMANCE OFCLEVELAND HEART DISEASE DATASET FOR IMPUTING MISSING VALUES”, International Journal of Pure and Applied Mathematics, Volume 120 No. 6 2018, 7271-7280, ISSN: 1314-3395 (on-line version)
5. S.Anitha, M.Vanitha (2019), “Imputation Methods for Missing Data for a Proposed VASA Dataset”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019, Blue Eyes Intelligence Engineering & Sciences Publication
6. Taeyoung Kim, WoongKo and Jinho Kim (2019), “Analysis and Impact Evaluation of Missing DataImputation in Day-ahead PV Generation Forecasting”, Appl. Sci. 2019, 9, 204; doi:10.3390/app9010204, Published: 8 January 2019
7. Anil Jadhav, DhanyaPramod& Krishnan Ramanathan(2019), “Comparison of Performance of Data Imputation Methods for Numeric Dataset”, Applied Artificial Intelligence 33:10, 913-

933, DOI: 10.1080/08839514.2019.1637138, An International Journal ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20, Published online: 04 Jul 2019

8. AdityaSundararajan and Arif I. Sarwat(2019), "Evaluation of Missing Data ImputationMethods for an Enhanced Distributed PV Generation Prediction", Springer Nature Switzerland AG 2020,K. Arai et al. (Eds.): FTC 2019, AISC 1069, pp. 590–609, 2020. https://doi.org/10.1007/978-3-030-32520-6_43
9. C. UshaNandhini, Dr.P.R.Tamilselvi.,"An Ensemble Approach for Performance Analysis of Preprocessing Techniques on Classification for Heart Disease Datasets", by IMRF, International Research Journals (UGC approved), 2018.
10. Meenakshi, Dr..RajanVohra, Gimpsy(2014), "Missing value Imputation in Multi Attribute Date Set", International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol. 5(4), 2014, 5315-5321.
11. Jiawei Han and MichelineKamber, Data Mining Concepts and Techniques,2nd Edition, An imprint of Elsevier
12. Margaret H.Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2014
13. C. UshaNandhini, Dr.P.R.Tamilselvi, "A Review on Feature Selection Approaches for Heart Disease Classification", International Journal of Theoretical & Applied Sciences, Special Issue 10(1a): 63-67(2018).
14. Jared P. Lander, R for Everyone-Advanced Analytics and Graphics, 2nd Edition, Pearson India Education Services Pvt., Ltd.,
15. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R",Springer Texts in Statistics, 1st Edition, 2017.