Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019)
IEEE Xplore Part Number: CFP19P17-ART; ISBN:978-1-7281-2119-2

A Novel Algorithm for Missing Data Imputation on Machine Learning

G. Madhu, B.Lalith Bharadwaj, G.Nagachandrika, K.Sai Vardhan
Department of Information Technology,
VNR Vignana Jyothi Institute of Engineering and Technology, Telangana, Hyderabad-90, INDIA.
[1] madhu_g@vnrvjiet.in, [2] lalithbharadwaj313@gmail.com, [3] nchandrika_g@vnrvjiet.in, [4] kanumolusaivardhan@gmail.com

Abstract— Missing data value plays a significant role in medical research and its presence causes an adverse effect on machine learning and AI models which leads to the wrong insights for decision making. Past few decades, researchers have developed and applied various imputation approaches to real-world applications. In addition, imputation methods help us to build effective models to discover hidden patterns in medical applications that can provide insightful outcomes for better decision-making. In this paper, a new approach is proposed to impute the missing data value using XGBoost (eXtreme Gradient Boosting) of ensemble learning method for continuous attributes in medical datasets. The proposed methods are continuous type attribute imputations for continuous and discrete data attributes. In this approach, we impute each missing data attribute value by predicting its data value from non-missing data attributes. The experiments are conducted on benchmark medical datasets missing values ranging from 1.98% to 50.65% and compared with iterative imputation, KNN imputation, and missForest imputation. In our study, we observe that missXGBoost can successfully handle missing data attributes of continuous types of attributes and it outperforms other imputation methods.

Keywords – Dataset; Imputatio;, Normalized Mean; Missing Data Value;

I. INTRODUCTION

The concept of missing value problem plays a significant role in data analysis and predictions in medical applications, wireless sensor networks, and electrical power networks [1]. The presence of the missing data value has a major issue for many machine learning algorithms, which is not capable to handle missing data values [2]. In general, medical datasets are collected from various legacy systems which are incomplete form due to the data entry procedures, inappropriate measurements, and equipment malfunctions, etc., [3]. In this view, data imputation methods play a significant role in handling incomplete dataset. In literature, we have two types of methods to deal with missing data value problem i.e. delete case or imputed with the plausible values [4]. The first one is the best ways of concerns with the missing data value which can delete the incomplete data record in the dataset [2][5]. This is applicable only a limited number of data records, and which leads to bias on classification problems [2][4]. Another method is imputation that means the process of substituting the missing data attribute of plausible data attributes [6]. The traditional imputation approaches are employed well-known statistical measures such as hot and cold deck imputations, listwise and

pairwise deletion, mean imputations and other imputation approaches [7][4]. However, these procedures are single imputation or multiple imputation methods, in a single imputation method, a missing attribute is substituted by one probable data value. In multiple imputations one or more attribute values are imputed, then the multiple imputations are performed better results in terms of modeling the uncertainty for data analysis [8]. In addition, some of the imputation approaches are fixed type attributes, random values, the nearest neighbor attribute values and mean attribute values [9]. Using statistical methods for data imputation may be reduced the bias of model with higher precision of the estimated data value which impacts on the classification model [10]. The important issue in the data analysis of missing value problem which is to determine the most plausible value [9]. However, the missing data value can be creating a serious impact on data analysis under the decision-making process.

In general, imputation techniques are completely accountable for uncertainty while forecasting the missing data value by imputing plausible variability into multiple imputed values [11]. In view of the above issues and challenges, we introduce a new imputation method that can be handle continuous type of attributes as input data values. Tianqi Chen [12] developed an eXtreme Gradient Boosting also known as XGBoost and it is a popular classification method that is implemented over the gradient boosting with greedy passion method, also deal with continuous type attributes. We present the missing data value problem using an imputation strategy by training an XGBoost algorithm [13] based on the complete dataset and predicting the missing data values then imputation of a plausible value then it proceeds an iterative manner.

During this process, first, split the dataset into a complete dataset and missing dataset and train XGBoost algorithm on this complete dataset and to predict the missing data values iteratively then it replaces the plausible values in given dataset. For predicting the missing value, we pass non-missing instances in the complete dataset and pass this into the XGBoost algorithm, so it predicts a plausible value. This attribute value is imputed in a suitable place in dataset i.e. attribute value is imputed without creating a new dataset after

imputation. This process is recast until all the missing data values are imputed in the given dataset respectively.

II. Related Works

In the literature, many authors have been studied and applied numerous imputation methods for concerns with the missing data values. Rubin (1978) [14] presented a multiple imputation model for missing data value problem. Little and Schluchter (1985) [15] presented a maximum likelihood-based approximation by merging multivariate models for categorical data imputation. Ragunathan et al., (2001) [16] discussed a variable-wise conditional distribution for missing value problemusing sequential regression. Sterne J A C et al., (2009) [17] presented a review on multiple imputations for missing value problem and its limitations. Jerez et al., (2010) [18] presented statistical-based imputation methods such as mean, hot-deck and multiple imputations on the breast cancer problem. Ross Larsen (2011) [19] presented a comparative study on multiple imputations and full information maximum likelihood of inferior SAS simulation. Daniel J. Stekhoven and Peter Buhlmann (2012) [20] presented a new algorithm called missForest imputation which can handle mixed types of attributes and of missing data value, Karpievitch et al., (2012) [21] presented several imputation methods for missing value problemon the microarray data and mass spectrometry-based data. de França et al., (2013) [22] discussed a Bi-clustering based data imputation technique utilizing the mean squared residual metric that estimates the degree of coherence between each recorded cell of the dataset. Govardhan et al., (2014) [23] proposed a new imputation algorithm for continuous attributes using non-parametric based discretization with the z-score statistical measure. Xiaobo Yan et al., (2015) [24] presented the imputation method for missing data value in IoT data applying context-based linear mean, binary search method, and Gaussian mixture model. Yea J, Chu et al., (2016) [25] developed patent works for imputing missing value problem of one or more predictor variables. Imputation process is built on the information of a target variable using statistics of the predicted variable and the target variable. Myneni Madhu Bala et al., (2017) [26] discussed imputation framework using correlated based clustering technique and computed the correlation between each record in dataset w.r.t missing attributes and to impute the missing data value based on cluster mean value. UshaRani Yelipe et al.,(2018) [27] presented an efficient imputation approach for missing data value problem using IM-CBC method. Wei-Chao Lin et al., (2019) [28] presented a comparative study of the imputation algorithm for missing value problems and various issues were addressed for the missing value problem. Minseok Lee et al., (2019) [29] proposed a deep learning-based imputation algorithm for continuous attributes in IoT data smart space. The deep learning-based imputation network uses the multiple long short-term memory is design to the correlation information of each IoT data.

In the view of aforementioned imputation methods are either to impute or filled with plausible values based on various distance, statistical techniques or clustering techniques. But these approaches do not yield an effective in the classification model.

III. Proposed Methods

Motivated by the aforementioned missForest [20], we proposed a new imputation method called 'missXGBoost' that can be handle continuous type of attributes as input data values. We present the missing value problem using an imputation strategy by training XGBoost algorithm on the complete dataset and predicting the missing data values then fill the plausible value then it proceeds an iterative manner.

Let D=(D1, D2,….Dm) be an n×mdataset, we propose a new methodology which input data values are missing attributes. We use XGBoost for imputation of the missing value. In general, XGBoost algorithm has an in-built strategy to compute missing data values using default strategy distance metric and KNN imputation procedure. However, this approach leads to bias when the presence of a missing value in the test set which not the same for the train set. Instead, directly predict the missing data value by using XGBoost algorithm on the complete data values of the dataset. The comprehensive representation of missXGBoost method presented in Algorithm-1, 2 & 3.

Algorithm-1: Remove Missing Records

Algorithm for $S^{CC}: Input \rightarrow D$
Define $S^{CC}(D)$:
  If $D_j^i = Missing\ data$ then *do*
    delete D^i;
  return (D) // Complete dataset

The algorithm-1 identify the missing values and remove the complete column data records in the dataset.

Algorithm-2: Training the given Dataset

Algorithm for $trainX : Input \rightarrow D$, $Instance\ i$
  Define $TrainX\,(D, I)$:
    array $\leftarrow [\,]$ // An empty array
      for j in 0,1,2,..m :
        if $D_j^i \neq Missing\ data\ value$ then *do*
        array.add (D_j);
return transpose (array)

The algorithm-2 to be considered dataset as the input dataset. In that attributes are not missing values and generate a transpose matrix of the given dataset.

Algorithm-3: New Imputation Algorithm

$D \leftarrow Dataset\ with\ size\ mxn$
  Define Impute_XGB(D):
   for i in 0,1,2,…n
     for j in 0,1,2,…m
   If $D_j^i = Missing\ data$ then *do*
   $y_{train} \leftarrow S^{CC}$;
   $x_{train} \leftarrow TrainX\,(D_j^i)$;
   fit XGB algorithm $y_{train} \sim x_{train}$;
   $P \leftarrow concatenate\ (D^i_{0\rightarrow j}, D^i_{j+1\rightarrow q})$;
   $x_{mis} \leftarrow remove\ missing\ records\ from\ P$;

$y_{mis} \leftarrow D_j^i$;
$D_j^i \leftarrow$ *to predict* y_{mis} *using* x_{mis} ;
endif;
end for;
end for;
return D // Imputed dataset

IV. EXPERIMENTS AND RESULTS

This study, we demonstrated an overall analysis and the impact of the proposed method with other imputation methods such as iterative and KNN imputation methods on benchmark medical datasets. We have taken a group of six datasets are downloaded from the KEEL repository [30].

TABLE-I: MEDICAL DATASETS USED FOR EXPERIMENTATION.

Datasets	# Instances	# Features	#Classes	% MV's
Cleveland	13	303	5	1.98
Diabetes Dataset	8	768	2	50.65
Dermatology	34	366	6	2.19
Hepatitis	19	155	2	48.39
Mammographic	5	961	2	13.63
Wisconsin	9	699	2	2.29

In Table-I, summarize the properties and statistics of their datasets, the column categorized as "%MVs" indicates the percentages of missing values in the dataset. The initial observation is that the KEEL datasets such as Cleveland dataset, Diabetes dataset (Pima), Dermatology dataset, Hepatitis dataset, Mammographic dataset, and Wisconsin dataset. The percentage of MVs varied from 50.65% to 1.98% and we have used multi-class datasets with prompted MVs shown in table-I.

We have employed various imputation methods with these datasets. We used iterative imputation, KNN imputation, and missForest imputation and XGboost classifiers for classification simulations. The classification accuracies are performed using 10-fold cross-validation test, the RMSE value obtained for various test Splits (0.1, 0.15, 0.2, 0.25, 0.3, 0.35) using XGB Algorithm and variance scores values a are presented in Table-II to Table-IV.

TABLE-II: TEST CLASSIFIER ACCURACY WITH XGBOOST CLASSIFIER.

Datasets	Proposed Method + XGBoost		
	RMSE (%)	Accuracy (%)	Variance (%)
Cleveland	1.01	53.73	8.4
Diabetes	0.19	78.22	2.89
Dermatology	0.18	96.45	2.72
Hepatitis	0.11	82.45	8.6
Mammographic	0.13	82.5	3.7
Wisconsin	0.16	95.71	4.2

TABLE-III: TEST CLASSIFIER WITH ITERATIVE IMPUTATION AND XGBOOST.

Datasets	Iterative Method + XGBoost		
	RMSE (%)	Accuracy (%)	Variance (%)
Cleveland	0.99	54.7	7.94
Diabetes	0.19	75.7	5.59
Dermatology	0.19	96.45	2.73
Hepatitis	0.15	81.66	11.24
Mammographic	0.13	82.7	4.2
Wisconsin	0.16	95.8	3.9

TABLE-IV: TEST CLASSIFIER WITH MISSFOREST IMPUTATION AND XGBOOST

Datasets	MissForest Method + XGBoost		
	RMSE (%)	Accuracy (%)	Variance (%)
Cleveland	1.01	54.71	7.94
Diabetes	0.19	77.45	6.04
Dermatology	0.19	96.45	2.72
Hepatitis	0.11	85.00	8.6
Mammographic	0.13	82.30	3.77
Wisconsin	0.16	95.70	3.78

From the summary of Table II to IV, we can say that the proposed algorithms (Algorithms) are superior to other imputation methods like iterative and missForset imputation with XGBoost classifier.

However, to compute the missing data value, and the implementation of various imputation approaches have been failed to deal with more than 50% percent of missing data values. Thus, we tried to successfully impute the missing data value by using the proposed algorithm-1,2 &3. The imputation methods and XGboost algorithm of test split accuracies are shown in figure-1 on Pima dataset.

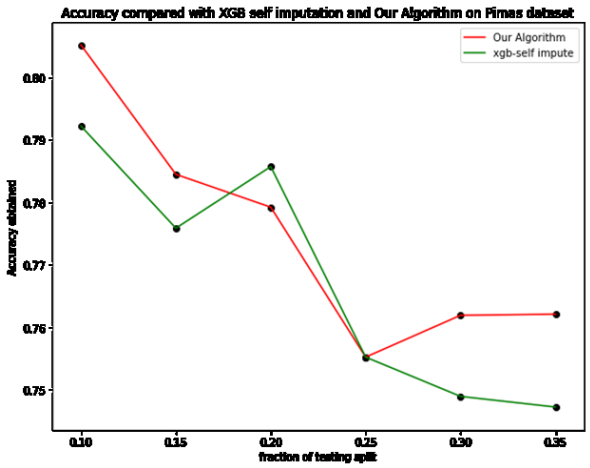

Fig.1: Classifier accuracy comparison with XGBoost self-imputation the proposed method on Pima dataset.

The root mean square error rate (RMSE) results are given in Figure-2, from this results on Pima dataset then we can say that miss XGBoost performs well.

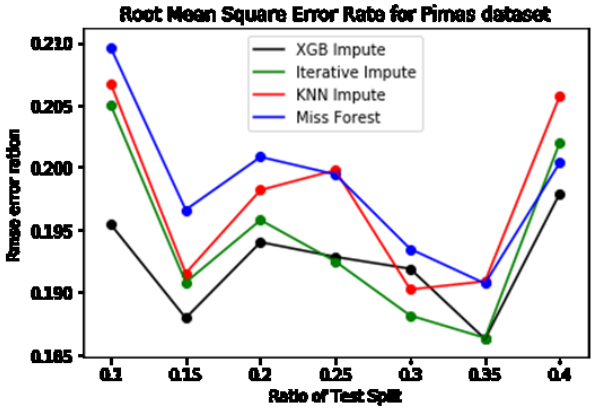

Fig.2: Root means square error rate compared with the proposed method vs other methods on Pima dataset.

In addition, variance scores are computed on Pima dataset with the proposed method and other imputation methods that are shown in figure-3.

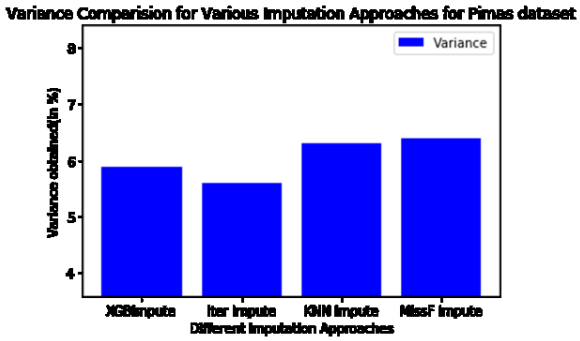

Fig.3: Variance scores with proposed method vs other methods on Pima dataset.

V. CONCLUSIONS

In this paper, developed a novel imputation technique called missXGBoost imputation technique. This imputation is very much suitable for continuous attributes of medical applications. After imputation strategy, this methodology has been tested on benchmark medical datasets with the percentage of MVs varied from 1.98% to 50.65%. The proposed missXGBoost method is imputed plausible data value in the original dataset and evaluated the classifier accuracy with XGBoost, variance scores, and RMSE rates are computed. In addition, experiment results are shown that proposed imputation method accuracy better than the other traditional imputation methods. Furthermore, missXGBoost can be applied on mixed-type attributes and to high dimensional datasets.

REFERENCES

[1] Aizaz Chaudhry et al., (2019). "A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness," Wireless Communications and Mobile Computing, vol. 2019, Article ID 4039758, 13 pages, 2019.

[2] G. Madhu and Dr.T.V. Rajinikanth (2012) "A novel index measure imputation algorithm for missing data values: A machine learning approach", IEEE International Conference on Computational Intelligence & Computing Research, pp.1-7,2012.

[3] Huang, M. W., Lin, W. C., & Tsai, C. F. (2018). Outlier Removal in Model-Based Missing Value Imputation for Medical Datasets. Journal of healthcare engineering, 2018, 1817479. doi:10.1155/2018/1817479

[4] Farhangfar Alireza et al., (2007). "A novel framework for imputation of missing values in databases." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 37, no. 5 (2007): 692-709.

[5] Chiewchanwattana et al., (2007). "Imputing incomplete time-series data based on varied-window similarity measure of data sequences." Pattern recognition letters 28, no. 9 (2007): 1091-1103.

[6] Rubin, Donald B. "Inference and missing data." Biometrika 63, no. 3 (1976): 581-592.

[7] https://en.wikipedia.org/wiki/Imputation_(statistics).

[8] Rezvan, P. H., Lee, K. J., & Simpson, J. A. "The rise of multiple imputation: a review of the reporting and implementation of the method in medical research", BMC medical research methodology, 2015, 15(1), 30.

[9] Hayati Rezvan, Panteha et al. "The rise of multiple imputation: a review of the reporting and implementation of the method in medical research." BMC medical research methodology vol. 15 30. 7 Apr. 2015, doi:10.1186/s12874-015-0022-1

[10] Wood, Angela M., Ian R. White, and Simon G. Thompson. "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." Clinical trials 1, no. 4 (2004): 368-376.

[11] Sterne, Jonathan AC, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." BMJ 338 (2009): b2393.

[12] Chen, Tianqi, and Tong He. "Higgs boson discovery with boosted trees." In NIPS 2014 workshop on high-energy physics and machine learning, pp. 69-80. 2015.

[13] Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM. pp. 785–794.

[14] Rubin D. (1978) Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse, Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, American Statistical Association, pp. 20-34.

[15] Little R. (1985) Schluchter M. Maximum likelihood estimation for mixed continuous and categorical data with missing values, Biometrika , 1985, vol. 72, pp. 497-512.

[16] Raghunathan T., et al.(2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, Surv. Methodol, 2001, vol. 27, pp. 85-96.

[17] Sterne Jonathan A C et al., (2009) "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." Bmj 338 (2009): b2393.

[18] Jerez et al., (2010) "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine 50, no. 2 (2010): pp.105-115.

[19] Larsen, Ross (2011). "Missing data imputation versus full information maximum likelihood with second-level dependencies." Structural Equation Modeling: A Multidisciplinary Journal 18, no. 4 (2011): 649-662.

[20] Stekhoven Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." Bioinformatics 28, no. 1 (2011): 112-118.

[21] Karpievitch, Yuliya V., Alan R. Dabney, and Richard D. Smith. "Normalization and missing value imputation for label-free LC-MS analysis." BMC Bioinformatics 13, no. 16 (2012): S5.

[22] de França, (2013). "Predicting missing values with bi-clustering: A coherence-based approach." Pattern Recognition 46, no. 5 (2013): 1255-1266.

[23] Govardhan, A., G. Madhu, and T. V. Rajinikanth (2014). "A non-parametric discretization-based imputation algorithm for continuous attributes with missing data values", International Journal of Information Processing, 8(1), pp.64-72, 2014.

[24] Yan Xiaobo, et al., (2015). "Missing value imputation based on Gaussian mixture model for the internet of things." Mathematical Problems in Engineering 2015 (2015).

[25] Chu, Yea J., Sier Han, Jing-Yun Shyr, and Jing Xu. (2016) "Missing value imputation for predictive models." U.S. Patent 9,443,194, issued September 13, 2016.

[26] Myneni Madhu Bala, et al., (2017). "Correlated Cluster-Based Imputation for Treatment of Missing Values." In Proceedings of the First International Conference on Computational Intelligence and Informatics, pp. 171-178. Springer, Singapore, 2017.

[27] Yelipe UshaRani et al., (2018). "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records." Computers & Electrical Engineering 66 (2018): 487-504.

[28] Wei-Chao, and Chih-Fong Tsai. (2019) "Missing value imputation: a review and analysis of the literature (2006–2017)." Artificial Intelligence Review (2019): 1-23.

[29] Lee, Minseok, et al., (2019). "Missing-Value Imputation of Continuous Missing Based on Deep Imputation Network Using Correlations among Multiple IoT Data Streams in a Smart Space." IEICE TRANSACTIONS on Information and Systems 102, no. 2 (2019): 289-298.

[30] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.