Missing Data Imputation Method for Autism Prediction

Kamatchi Priya L, Baranidharan C

Abstract: ***Missing data imputation is essential task becauseremoving all records with missing values will discard useful information from other attributes. This paper estimates the performanceof prediction for autism dataset with imputed missing values. Statistical imputation methods like mean, imputation with zero or constant and machine learning imputation methods like K-nearest neighbour chained Equation methods were compared with the proposed deep learning imputation method. The predictions of patients with autistic spectrum disorder were measured using support vector machine for imputed dataset. Among the imputation methods, Deeplearningalgorithm outperformed statistical and machine learning imputation methods. The same is validated using significant difference in p values revealed using Friedman's test.***

Keywords: K-nearest neighbour, chained Equation, deep learning, support vector machine, Friedman's test, p values

I. INTRODUCTION

Autism is a neuro development disorder accompanying substantial health care costs. To reduce the same early diagnosis is required. Regrettably, waiting times for an autism analysis are extensive and processes are not cost effective. Thefinancial influence of autism and the progress in the number of autism cases discloses an imperative requirement for the development of easily realized and operationalselection methods. Therefore autism predictionhelps health professionals to notify the patients whether they should follow prescribedscientific diagnosis. Therapid increase in the number of autism cases throughout the world demands a perfect diagnosis tool and correct data. However, such datasets with missing value affects the performance of analyses in terms of specificity, sensitivity, efficiency and predictive accuracy of the autism identification process. Currently, very less datasets are available pertaining to autism identificationis accessible and most of them correspond to genetic data. Hence, Autism screening of adults obtained from UCI repository that contained 20 features to be utilized for more exploration particularly in evaluating dominant characters which influence autism and improving the performance of classification for identifying autism cases. This dataset consists of 10 social features and 10 specific characteristics for predicting the autism cases. Two of the features have many missing values which cannot be ignored, hence imputation of those values are essential to perform prediction.

Revised Manuscript Received on January 15, 2020

Kamatchi Priya L, Associate Professor, New Horizon College of Engineering, Bangalore, India.

Baranidharan C, AVEVA Solution.

The missing data bias the results of predictionhence it is important to either ignore the whole record or impute the missing values. When records with missing values are ignored there is a possibility of insufficient training data. In such cases imputation is preferred. Missing data are categorized as Associated Missing Value Imputation (AMVI) [1] and Independent Missing Value Imputation (IMVI) [2]. AMVI refers there is neither anassociation between a missing value and othermissing value in the dataset nor a missing value and an observed value in the dataset. IMVIhaseither an association between a missing value and other missing value in the dataset or a missing value and an observed value in the dataset. In this paper, AMVI is considered and there are several techniques to explore the same. Among the AMVI methods list wise deletion imputation method is adopted to analyze autistic spectrum disorder.

This paper discusses imputation of missing values to improve the prediction of autistic spectrum disorder. In Section II, Description of the dataset autistic spectrum disorder is detailed. In Section III, illustration of different missing value imputation methods using statistical methods and machine learning methods is presented along with various methods of estimation of missing value proportion in the dataset. Support Vector Machine (SVM) classification algorithm is used for predicting the patients with disorder is also elaborated. In Section IV, results pertaining to the execution of missing value imputation of categorical variables in the autistic spectrum disorder dataset using various statistical and machine learning methods are presented. Discussion on these methods and their performance is detailed. Finally, in conclusion, significant inferences stating the efficiency of machine learning imputation methods are discussed.

II. LITERATURE SURVEY

In this section, briefly describesthe approaches appliedto impute missing values in the dataset and also defineswhich prediction method is used after imputation of missing values. There aresubsections that illustratenumerous considerationsthat are made and also give detailsof how the imputation methods have beenimplemented. Various imputation techniques applied aredescribed which includes boththe statistical methods, covering Imputation using Mean/Median values, Imputation using most frequent or zero or constant values, Stochastic regression imputation, Extrapolation and Interpolation and Hot-Deck imputation,the machine learning based methods, coversImputation using k nearest neighbors (k-NN), Imputation using Multivariate

Imputation by Chained Equation (MICE), and Imputation using deep learning (DNN). Finally, the SVM model to predict survival probabilities isexplained with statistical methods used inmodel accuracy evaluation.

A. Missing Data Imputation for Incomplete Data Set

Before describing the methods used for missing data imputation,it is important to address the remarks concerning all the imputation methods.

Missing data pattern: Autism data set can be considered missing completely at random (AMVI, since there the probability that an observation is missing is neither dependent on the observed set nor the missing set.

Imputation on categorical variables using dummies: Anencodingscheme is adopted to replace categorical values with dummy variables. For instance, in Autism data set, the feature relation has five categories as self,relative,parent, heath care professional wereencoded as 01, 02,03, 04 and 05 using dictionary in python. Dummies have to be treated properly otherwise they will be considered as numeric values.

B. *Statistical Methods*

Imputation of missing value using statistical method includes, (1) Imputation using Mean/Median values, and (2) Imputation using most frequent or zero or constant values.

- Imputation using Mean/Median values:This is done by manipulating the mean or median of the non-missing values for that feature and then substituting the missing values within each feature or column unconnectedlyfrom the other features. It is efficient only forsmall numeric data. The disadvantage is it does not consider the correlation among the features and moreover cannot be used for categorical features [3].
- Imputation using most frequent or zero or constant values: This method works for features with categorical values, where the most frequent occurring label is assigned. The disadvantage of this method isthat it also does not consider the correlation among the features [4].
- Stochastic regression imputation:This method works like regression imputation method. Itattempts to predict the missingvalues by degenerating it from other interrelated variables in the dataset and also fromsome random residual value [5].
- Extrapolation and Interpolation (E&I):This method tries to assess the values from other records within the range of adiscrete set of known data points [6].
- Hot-Deck imputation:This method arbitrarilyselects the missing value from a set of associated andsimilar variables [7].

C. *Machine Learning Methods*

Imputation of missing value using Machine learning method includes, (1) Imputation using k nearest neighbors (k-NN), (2) Imputation using Multivariate Imputation by Chained Equation (MICE), and (3) Imputation using deep learning (DNN).

- Imputation using k nearest neighbors (k-NN): The k nearest neighbour is a classification algorithm that uses 'feature similarity' to predict the valuesof data points with unknown class. The k-NN algorithm looks for how close it is with the training data, this feature can be used for predicting missing values.By finding the k'sclosest neighbour to the data point with missing value and then imputingthe values based on the non-missing values in the neighborhood data points. This is implemented by creating a basic mean impute, then it uses the resulting complete list toconstruct a KDTree. Theresulting KDTree is used to compute nearestneighbour (NN). After it finds the k-NNs, it takes the weighted average ofthem to impute the same in the place of missing value. Though the algorithm is better than mean, median or most frequent imputation methods, it is computationally expensive in terms of memory. Unlike SVM, it is highly sensitive to outliers [8].

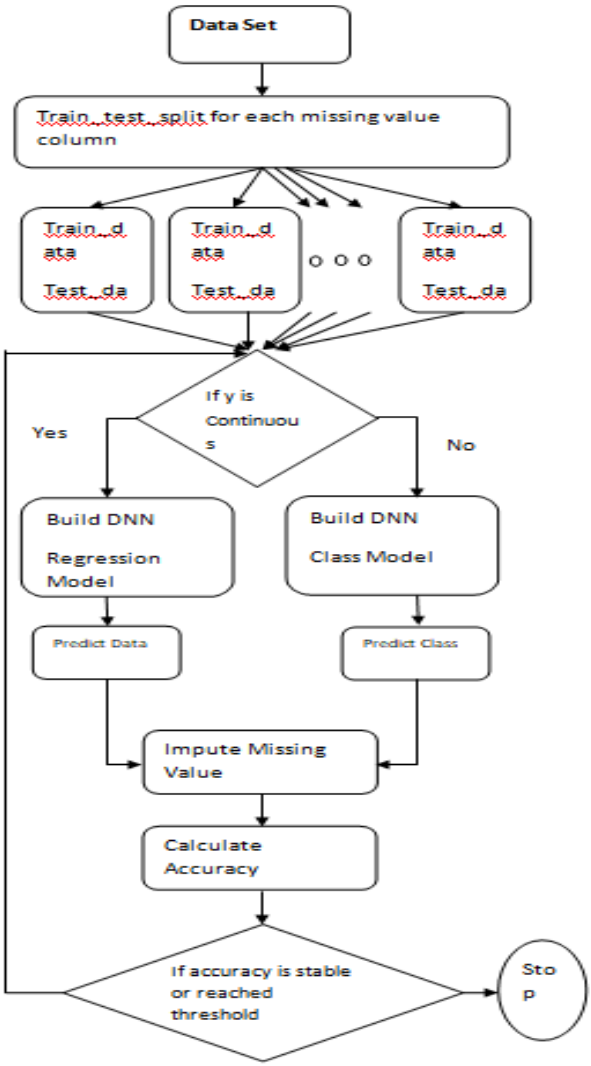

Fig 1. Flow Diagram of the DNN for Missing Value Imputation.

- Imputation using Multivariate Imputation by Chained Equation (MICE): This is done by filling the missing data multiple times.Multiple Imputations (MIs) are much better than a single imputation, as itmeasures the uncertainty of the missing values in a better way.

- The chainedequations approach is generic and can be adapted for different variablesof different data types (i.e., continuous or binary). It also works fine for complexitiessuch as bounds or survey skip patterns [9].

III. PROPOSED METHOD

A. Imputation using deep learning (DNN)

This method works for both categorical and non-numerical features. Deep Neural Networks [10] are used in learning and prediction; the same can be used for imputing missing values in the data frame.

The proposed method has the following activities: The dataset is segregated based on the data type, viz., numerical, categorical. Appropriate encoding is chosen based on the type of the variable/feature. For example if it is numerical variable, correlation of the variable with other variable are calculated. The regression model is built to find the missing value. If it is a categorical variable, the classification model is built to find the missing value. Input / Output split are made randomly and prediction is done based on the model created using Deep Neural Networks. The Fig 1 illustrates the process of missing value imputation using DNN in detail. But it consumes more time for imputation of missing values in a single column for large datasets. It is more accurate than other methods because it not only considers that column rather the whole dataset is being used for predicting that value.

B. The SVM Prediction Model

SVM is one of the prediction models which have high performance. SVM is widely used in application such as regression estimation, performance in pattern recognition, healthcare problems, financial forecasting, manufacturing yield prediction, text classification, facial detection using image processing, hand written digit recognition, etc., SVM generates separating hyper planes for classification, through mapping of features in high-dimensional data. SVM utilizes support vectors to construct a model which is linear in nature. It estimates the class boundaries using a non-linear decision function.

SVM invents an ideal hyper plane which splits the data of extreme distance between the hyperplane and the neighboring training points. The training points that are nearby to the best unraveling hyperplane are called support vectors. All other physical activity examples like unrelated records for defining the binary class boundaries. In general cases where the data is not linearly separated, SVM uses non-linear machines to find a hyperplane that minimize the number of errors for the training set [11]. Feature selection [12] is performed before classification for high dimension data to eliminate redundancy among feature and choose the most relevant feature to the target vector, thus improving the classification accuracy.

The learning process of SVM to build decision function is similar to that of Back Propagation Neural Network (BPN). SVM enjoys many benefits, one is it has two free parameters, that is kernel parameter and upper bound. The second benefit is that it assurances the presence of inimitable, best and universal solution which is measured alike to resolving a quadratic programming. Third, SVM is used to decrease the generalization error instead of decreasing the training error. Fourth, the hypothesis space is examined by data to improve the generalization performance. Fifth, SVM is comparatively steady compared to other classification techniques and the flexibility in the decision boundary is substantially small. Finally, SVM is built with the small training data set size [13], since it learns by taking geometric picture conforming to the kernel function. Moreover, no matter how large the training set size is, SVM is capable of extracting the optimal solution.

C. Estimation of Missing Value Proportions

The proportion p is estimated by five different techniques, they are (1) Complete Case Analysis (List wise Deletion) [14], (2) Linear Imputation without rounding [15], (3) Linear Imputation with Rounding [16], (4) Logistic Regression Imputation [17] and (5) Discriminant Function Imputation [18]. To estimate the proportion p, Complete Case Analysis also termed as List wise Deletion is applied to each sample.

D. Model Evaluation

The performance of the prediction models is estimated by two leading properties: (1) Classification Accuracy and (2) Statistical test to measure significance of the methodology. The amount of the nearness of the experimental value to the factual value is known as accuracy. Statistical test which does not assume normality in the distribution is known as non-parametric test [19]. To difference in the treatment across two tests is evaluated using Mann-Whitney test and Wilcoxon signed rank test. The difference in the treatment across multiple tests is evaluated using Friedman's test [20]. Multiple pair wise comparisons are made to difference in every pair of treatment considered. Here Friedman's test is used to show the differences.

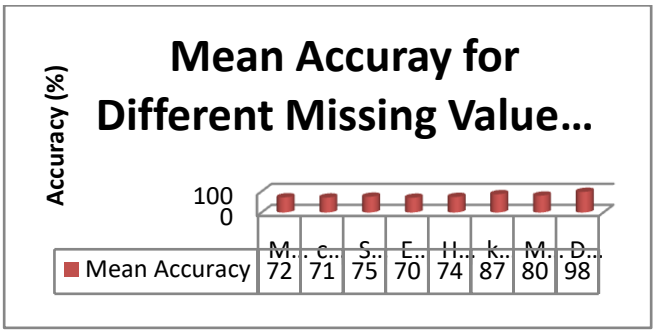

Fig 2. Mean accuracy of different missing value imputation techniques

TABLE I. DATASETDESCRIPTION

Dataset Name	Autism Identification for Adolescent	Autism Identification for Children
Type	Biological Data	Biological Data
No. of Observations	104	292
No. of features	21	21
Missing Value Characteristics	Categorical	Categorical

TABLE II. PAIRWISE COMPARISONS: P VALUES INDICATING SIGNIFICANT DIFFERENCE

Methods	Mean	Const	SR	E&I	HD	k-NN	MICE	DNN
Mean	1	1	0.997	1	0.999	0.7	0.79	0.02
Const	1	1	0.975	1	0.984	0.72	0.89	0.034
Stochastic	0.997	0.975	1	0.995	1	0.795	0.812	0.046
E&I	1	1	0.995	1	0.997	0.81	0.79	0.023
Hot-Deck	0.999	0.984	1	0.997	1	0.797	0.834	0.001
k-NN	0.7	0.72	0.795	0.81	0.797	1	0.112	0.049
MICE	0.79	0.89	0.812	0.79	0.834	0.112	1	0.041
DNN	0.02	0.034	0.046	0.023	0.001	0.049	0.041	1

IV. RESULTS AND DISCUSSION

The performance of the many imputation techniques is tested using Autistic Spectrum Disorder Prediction dataset from the UCI Repository (as shown in Table I). All the missing values are categorical. Various experiments were conducted by varying the training and testing set ratio as 50:50,55:45, 60:40, 65:35, 70:30, 75:25, 80:20, 85:15 and 90:10 for classification. SVM is used for classifying the data. The average of the accuracy for various training and testing ratio obtained is considered for evaluation of missing value imputation techniques. The same has been tabulated and visualized in Figure 1. To determine the statistical validation, Friedman's test for missing value imputation on classification accuracy test is conducted with the results obtained. The p values of Friedman's test are shown in Table II.From the table it is evident than there is a significant difference in p-values of Datawig method compared to others.

Null Hypothesis: The records/observationsare derived from the same population (H_0).

Alternate Hypothesis: The records/observationsare derived from the different population (H_a).

The Table II illustrates that, the p-value is lesser than the significance level **0.05**, Hence the null hypothesis $\boldsymbol{H_0}$ is rejected and the alternative hypothesis $\boldsymbol{H_a}$ is accepted.

V. CONCLUSIONS

As far as missing value imputation is concerned, there is no faultlesstechnique to impute missing valuesin a dataset. Each technique or strategy can perform better for certain criteria depending on the datasets and missing data types but may perform much worse on other scenarios.There aresome set rules to decide which strategy to use for particular typesof missing values, but beyond that, you should experiment and check whichmodel works best for your dataset. As far as the autistic spectrum disorder prediction dataset is concerned, DNN outperforms other statistical and machine learning state of art methods considered. The same has been statistically validated using significance value obtained to check the difference in the mean accuracy.Though the autism dataset has only categorical variables, the proposed DNN imputation is feasible for both categorical and numerical variables.

REFERENCES

1. Little, Roderick JA. "A test of missing completely at random for multivariate data with missing values." Journal of the American statistical Association 83, no. 404 (1988): 1198-1202.
2. Allison, Paul D. "Multiple imputation for missing data: A cautionary tale." Sociological methods & research 28, no. 3 (2000): 301-309.
3. Engels, Jean Mundahl, and Paula Diehr. "Imputation of missing longitudinal data: a comparison of methods." Journal of clinical epidemiology 56, no. 10 (2003): 968-976.
4. Follmann, Dean, Paul Elliott, I. L. Suh, and Jeffrey Cutler. "Variance imputation for overviews of clinical trials with continuous response." Journal of clinical epidemiology 45, no. 7 (1992): 769-773.
5. Gold, Michael Steven, and Peter M. Bentler. "Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization." Structural Equation Modeling 7, no. 3 (2000): 319-355.
6. Zhu, Xiaofeng, Shichao Zhang, Zhi Jin, Zili Zhang, and Zhuoming Xu. "Missing value estimation for mixed-attribute data sets." IEEE Transactions on Knowledge and Data Engineering23, no. 1 (2010): 110-121.
7. Myers, Teresa A. "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data." Communication Methods and Measures 5, no. 4 (2011): 297-310.

8. García-Laencina, Pedro J., José-Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal, and Michel Verleysen. "K nearest neighbours with mutual information for simultaneous classification and missing data imputation." Neurocomputing72, no. 7-9 (2009): 1483-1493.
9. Buuren, S. van, and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." Journal of statistical software (2010): 1-68.
10. Whitehead, T. M., B. W. J. Irwin, P. Hunt, M. D. Segall, and G. J. Conduit. "Imputation of Assay Bioactivity Data Using Deep Learning." Journal of chemical information and modeling 59, no. 3 (2019): 1197-1204.
11. Kamatchi Priya, L, Kavitha Devi, MK &Nagarajan, S, "Improvising classification performance for high dimensional and small sample data sets", Asian journal of Information and technology17, no. 4(2018), 261-270.
12. Kamatchi Priya, L, Kavitha Devi, MK &Ilakkiya, J, "Laplancianfisher clustering based feature subset selection for semi-supervised data", International Journal of Applied Engineering Research (IJAER)10, no. 75 (2015), pp. 263-269.
13. Kamatchi Priya, L,Kavitha Devi, MK &Nagarajan, S, "Data analytics: Feature extraction for application with small sample in classification algorithms", International journal of Business Information Systems 26, no. 3 (2018), pp. 378-401.
14. Little, Roderick JA, and Donald B. Rubin. Statistical analysis with missing data. Vol. 793. John Wiley & Sons, 2019.
15. Hasler, Caren, and Radu V. Craiu. "Nonparametric imputation method for nonresponse in surveys." Statistical Methods & Applications (2019): 1-24.
16. Allison, Paul D. "Imputation of categorical variables with PROC MI." SUGI 30 proceedings 113, no. 30 (2005): 1-14.
17. Harrell Jr, Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015.
18. Yuan, Yang. "Multiple imputation using SAS software." Journal of Statistical Software 45, no. 6 (2011): 1-25.
19. Kamatchi Priya, L, Kavitha Devi,MK &Nagarajan, S, "Spending on education determinant of economic growth using structural equation modeling", International Journal of Applied Engineering Research (IJAER), 10, no. 7 (2015), pp. 17991-18005.
20. McCrum-Gardner, Evie. "Which is the correct statistical test to use?" , British Journal of Oral and Maxillofacial Surgery, 46, no. 1 (2008): 38-41.

AUTHORS PROFILE

Dr. L. Kamatchi priyacompleted her Ph.D. - Decision Making Strategies in Anna University, Chennai, India. She is currently working as Associate Professor in New Horizon College of Engineering, Bangalore, India. She has published around 10 papers in different International journals and has organized CSIR Sponsored Faculty development Program and two IEEE sponsored Conferences. She holds CSI membership and Life membership in ISTE

Mr. C. Baranidharan is a Software Engineer with 14 years of experience and currently working in AVEVA Solution He completed his MCA in 2005. His area of expertise is Industrial Automation.