RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

IMPORTANCE OF PREPROCESSING IN PREDICTION

Shobha K[*1] & Nickolas S[2]

[*1&2]Department of Computer Applications, National Institute of Technology, Tiruchirapalli Tamilnadu, India -620015

ABSTRACT

Missing values in data collected over real world is a common phenomenon. These missing values need to be handled properly to achieve good analytical results. Missing values can be handled by deleting or by imputation methods. An imputation approach based on ensemble technique on consensus clusters is proposed in this paper. These clusters are outcome of unsupervised competitive neural network Adaptive Resonance Theory 2(ART2). Publicly available ”Titanic” dataset from Kaggle repository is used to compare the proposed methodology with existing ones. Experimental results show that proposed method outperformed existing methods by achieving lower error rates and higher prediction accuracy.

Keywords: *Missing values, Ensemble technique, competitive neural network, Imputation.*

I. INTRODUCTION

Data collection, the first phase of data mining or knowledge discovery, is being performed from single or multiple sources. In real world scenario, the datasets collected are often incomplete due to various reasons like human errors, incorrect measurements, equipment errors, etc. Applying data mining techniques to incomplete data and getting accurate result is very difficult [1]. Hence, missing value needs to be treated with some plausible values. Several research works on handling missing data exist, which either deletes missing values or imputes them.

Imputation algorithms are usually application specific and got its diverse application in different fields like, imputation in traffic data [2], meteorological data, medical data, genome data etc.,[3] [4]. Two most important factors that need to be considered in imputation are the missing type and the type of data distribution. Existing algorithms like mean, k-nearest neighnor, k-means clustering , self organizing map(SOM), fuzzy based imputation, genetic algorithm [2][5] [6] consider only the type of missing values, but not the type of distribution of missing data. Considering combined effect of missing type and data distribution in imputation is not being carried out in the literature. In this paper, we propose an algorithm which has a combined effect of missing type and data distribution for imputation of classification dataset. The proposed algorithm uses ensemble approach on unsupervised cluster and it also employs one of the best classification algorithms for imputed dataset.

II. LITERATURE SURVEY

Existing imputation algorithms are based on dependent and independent attributes. These algorithms treat missing attribute as dependent attribute and imputes them using non missing attributes (or independent attributes). Missing data are classified into three different types: (i) Missing Completely at Random (MCAR) (ii) Missing at Random (MAR) and (iii) Not Missing at Random (NMAR)[7]. Data deletion and data imputation are the two widely used treatments for handling these missing types. In data deletion approach, records with missing values, or attributes with missing values are removed. Removal of attributes is done only when particular attribute is not needed for analysis. Loss of information that is used for analysis is the major drawback of this approach. Inspite of this drawback, data deletion proves to be helpful in the case of MCAR.

Researchers have overcome this drawback by various imputation techniques. Imputation refers to the process of replacing missing values with possible values. This can be done with different algorithms which are further divided into two categories: single imputation and multiple imputation [8]. In single imputation method, as the name suggests, missing value is treated by a single value [9]. Whereas, in multiple imputation, several different values are

imputed based on method's likelihood. Multiple imputation process is divided into three phases namely (i) imputation (ii) analysis and (iii) pooling. This makes multiple imputation methods complex and computationally more expensive than single imputation. But this technique accommodates uncertainty and sample variability associated with imputation.

Imputation can be carried out in three different manners, viz., data driven, model based and machine learning based. Data driven imputation uses complete data to compute missing values. It includes methods like mean, hot deck and cold deck imputation. Model based imputation methods use some data models to fill missing values. They work on the assumption that the data are generated by a model and are governed by unknown parameters. Machine learning based models use various machine learning algorithms on available complete data to impute missing values.

Need for improved accuracy and better scalability are the major reasons that demands revision in existing imputation algorithms. Existing multiple imputation and machine learn-ing based methods provides high accuracy only for smaller datasets. Hence, we need an algorithm to work in scalable en-vironment. Our proposed model work in scalable environment by clustering the data and by applying imputation algorithm on these clusters. The proposed work considers imputation of different type of attributes and it also chooses suitable classification algorithm for dataset.

III. PROPOSED METHOD

In this paper, we propose an ensemble imputation algorithm on different clusters of data set, these clusters are the outcome of competitive neural network. Different steps involved in proposed method is as shown in Fig 1.

Proposed method has two phases: the model building phase and the evaluation phase.The model building phase starts by choosing dataset of interest, for analytics, from the repository. Data chosen are clustered using a competitive self-organizing neural network. The competitive neural network used in this work is the second generation network of Adaptive Resonance Theory (ART) and is called as Adaptive Resonance Theory 2(ART2)[10][11]. ART2 is capable of handling continuous variables. Architecture of ART2 consists of comparison field C_1, recognition field C_2, composed of neurons and with two deciding elements (a) threshold recognition field and, (b) reset module. C_1 and C_2 are connected with bottom up and top down weights respectively. Input data undergoes feature enhancement in comparison field and then passed to its best match in subsequent layer(C_2) with bottom up weights. Best match is the neuron whose set of weights closely matches the feature enhanced input data. Data reaching recognition field C_2 turn on competition among all neurons of C_2 field. Each C_2 field neuron outputs a negative signal to other neurons and thus inhibits their output. In this way, each neuron in recognition field represent a category to which input data are clustered.

The reset module handles the process of matching output signal from the recognition layer to the vigilance parameter(a threshold value). If the threshold is overcome, the training commences. Otherwise, the recognition neuron is inhibited and adjustment in weights are carried out towards matching the input vector.

Input vectors that fall under same class are clustered together.In this work, we have used a combination of similar clusters from different runs of ART2 on same dataset, this method of clustering is called as consensus clustering. Each run of algorithm is set with different vigilance parameter set by the user. Best clusters are obtained when expectation patterns are between 0.7-1, with an increment of 0.05 in each run. Final clusters are formed based on the similarity match of clusters.

Table I mean squared error(MSE) of proposed and existing methods

Method	MSE P	MSE M	MSE KNN
5%	0.100059	0.554713	0.161741
10%	0.236971	1.039114	0.384876
15%	0.314402	1.56221	0.50747
20%	0.405279	2.198652	0.679831
25%	0.509243	2.697023	0.776607
30%	0.714282	3.27892	1.420718

Table II mean absolute error(MAE) of proposed and existing methods

Method	MAE P	MAE M	MAE KNN
5%	0.0602	0.1543	0.0733
10%	0.1325	0.2899	0.1486
15%	0.2041	0.4381	0.2513
20%	0.3076	0.5982	0.3416
25%	0.3395	0.7409	0.4343
30%	0.4887	0.8988	0.5474

Ensemble algorithm is used on each cluster, obtained as a result of the application of ART2 neural network, to impute the missing values. Ensemble algorithm is a heuristic approach in which we pick different combinations of algorithm to impute. The combination which gives lowest error rate is taken as final imputation algorithm. These algorithms are evaluated using statistical approaches like Mean Squared Error (MSE) and Mean Absolute Error(MAE). Lower the error rates higher the accuracy of imputation. If error rates are higher, then ensemble algorithm need to be redesigned. This process continues until there is no much variation in error rates. For the dataset considered in this paper, the ensemble algorithm for imputation is formed by considering dependencies of each attribute among clustered data, mean of non-missing values in clusters and k-nearest neighbor(k-NN) in each cluster. Second phase of the proposed work is to find the best classification algorithm for imputed dataset.

IV. RESULTS AND DISCUSSION

Efficiency of the proposed systems are evaluated on a 3.50 GHz Intel Xeon processor E3-1270 machine, with python 2.7 as a programming environment.

Experiments were conducted using ”Titanic” dataset from Kaggle repository. Aim of the data set is to predict what sorts of people were likely to survive titanic disaster. This data set is a multivariate attribute with 1310 records and 14 attributes having missing values.To check effectiveness of proposed imputation method, missing values and non-missing values were initially separated. Later, with non-missing data the missing data were generated with Missing Completely at Random (MCAR) pattern by randomly deleting values from each attribute in data set ranging from 5% to 30%.

Dataset with variable missing percentages is imputed with proposed ensemble method and comparisons are made with two other existing techniques, ”mean method” and ”k-nearest

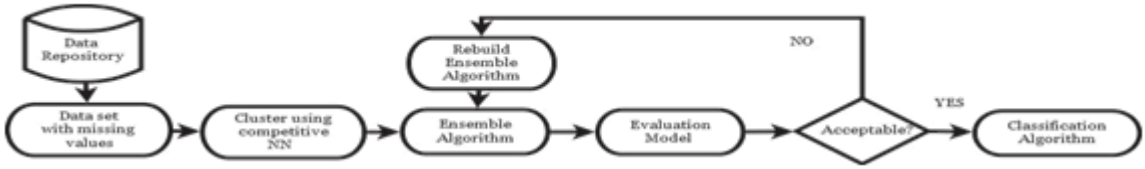

Fig. 1. Data flow of imputation process

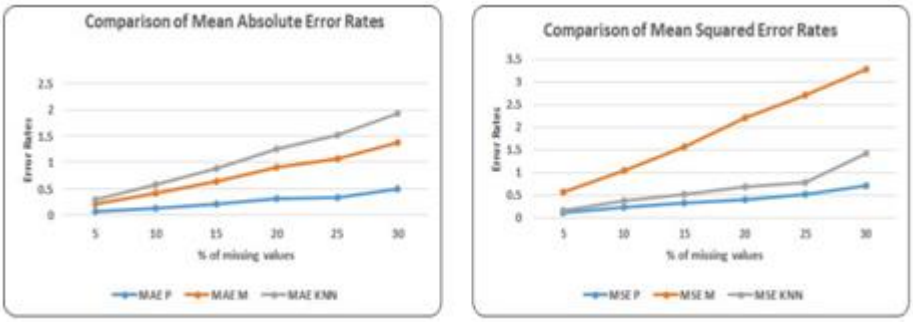

Fig. 2. Mean Squared Error & Mean Absolute Error of proposed and existing methods

Table III comparison of classification results

Method	Accuracy %			Precision %			Recall %			F-Measure %			AUC %		
	RF	RT	RTWS	RF	RT	RTWS	RF	RT	RTWS	RF	RT	RTWS	RF	RT	RTWS
Proposed Method	84	79	93	86	76	94	69	65	87	86	79	90	88	77	97
KNN	82	72	88	82	74	86	92	84	87	86	79	87	87	74	95
Mean	81	75	90	80	67	92	68	70	81	73	69	86	85	78	95

- RF- Random Forest
- RT- Random Tree
- RTWS- Random Tree with resampling

neighbor method (k-NN)". This comparison reveals that how ensemble based imputation outperforms other two methods for "Titanic" dataset. Mean Absolute Error (MAE) and Mean Squared Error (MSE) of data missing rates are shown in Table I and Table II and corresponding graphs in Fig2.

From these results it's observed that ensemble approach on clusters formed through consensus approach achieves smaller error rate than the single imputation approach. Along with statistical approaches, we have evaluated dataset with dif-ferent classification algorithms in "WEKA" tool to find the best classification algorithmic match for "Titanic" dataset. In this experiment, "Random Tree with sampling" classifier outperformed in majority of the cases, among three imputation techniques. Evaluation metrics of classification are shown in Table III. In this classification evaluation, imputed dataset with proposed method outperforms classification results of other methods of imputation.

V. CONCLUSION AND FUTURE WORK

In this paper, an algorithm for missing value imputation is proposed based on ensemble approach on consensus cluster. Consensus clusters are the outcome of unsupervised com-petitive neural network. Ensemble approaches are heuristic combination of different algorithms which gives low error rates and high accuracy. For Titanic data set,

ensemble of mean and k-NN imputation method results in lower error rates of MAE and MSE for different missing percentages. Integration of this ensemble method with ”Random Tree with sampling” classifier results in higher accuracy classification results than the non-ensemble models for imputation in ”Titanic” dataset. As a future work, the proposed method can be upgraded to impute different missing pattern in big data environment.

REFERENCES

1. *Alessandro Colantonio, Roberto Di Pietro, Alberto Ocello, and Nino Vincenzo Verde. Abba: Adaptive bicluster-based approach to impute missing values in binary matrices. In Proceedings of the 2010 ACM Symposium on Applied Computing, pages 1026–1033. ACM, 2010.*
2. *Wei Chiet Ku, George R Jagadeesh, Alok Prakash, and Thambipillai Srikanthan. A clustering-based approach for data-driven imputation of missing traffic data. In Integrated and Sustainable Transportation Systems (FISTS), 2016 IEEE Forum on, pages 1–6. IEEE, 2016.*
3. *George P Cressman. An operational objective analysis system. Mon. Wea. Rev, 87(10):367–374, 1959.*
4. *Ceylan Yozgatligil, Sipan Aslan, Cem Iyigun, and Inci Batmaz. Com-parison of missing value imputation methods in time series: the case of turkish meteorological data. Theoretical and applied climatology, 112(1-2):143–167, 2013.*
5. *Chandan Gautam and Vadlamani Ravi. Data imputation via evolution-ary computation, clustering and a neural network. Neurocomputing, 156:134–142, 2015.*
6. *Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using a fuzzy clustering-based em approach. Knowledge and Information Systems, 46(2):389–422, 2016.*
7. *Bing Zhu, Changzheng He, and Panos Liatsis. A robust missing value imputation method for noisy data. Applied Intelligence, 36(1):61–74, 2012.*
8. *Iffat A Gheyas and Leslie S Smith. A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing, 73(16-18):3039–3065, 2010.*
9. *Alireza Farhangfar, Lukasz A Kurgan, and Witold Pedrycz. A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(5):692–709, 2007.*
10. *Jianhong Luo and Dezhao Chen. An enhanced art2 neural network for clustering analysis. In Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on, pages 81–85. IEEE, 2008.*
11. *Gail A Carpenter and Stephen Grossberg. Adaptive resonance theory. Springer, 2017.*