ISSN: 2393-8528

IJICSE
Open Access

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 4 Issue 3; May-June-2017; Page No. 110-116

A Hybrid model for Missing Data Imputation

Swarnendu Kundu, Bidisha Pyne, Ilango P

SCOPE, VIT UNIVERSITY, Vellore

Email:jphswarnendu@gmail.com, pyne_bidisha@yahoo.in

Received 04 May 2017; Accepted 01 June 2017

ABSTRACT

Data mining is also known as the procedure of mining useful knowledge from large amount of data. This procedure found its application in various field like in making Business strategy, Market analysis, advancing medical treatments etc. But in order to do this deal this analysis, data scientist have to deal with real world dataset which consist of noisy, inconsistent as well as missing data. Thus the presence of such missing data can give rise to invalid and inaccurate decisions in knowledge extraction. The aim of this paper is to propose a new methodology in dealing with missing values. The methodology is named it as Exponential Clustering technique as this methodology is hybridization of clustering and exponential prediction of data and applied on Pima Indians Type II Diabetes dataset to analyze the performance with the existing techniques. The performance measured in this methodology is analyzed better than that of the existing technique of clustering.

Keywords: Pima dataset, missing data, imputation, exponential, clustering

INTRODUCTION

Within this decade the vast use of technology leads to abundant of data. This data can be used very efficiently to predict the future economic or health condition of the world. In order to do this prediction, various tools have been introduced to handle this extensive data. The classical knowledge extraction from this data consists of seven steps: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern evaluation and Knowledge presentation.

Data Cleaning is carried in order to remove the noisy and inconsistent data from the bulk of datasets. Followed with Data Integration, this step considered as the process of combining multiple inter related to dataset to get a more meaningful dataset. Next step follows Data Selection; it is done to retrieve the relevant data from the set of integrated dataset to do the required analysis. Then this selected data is transformed into an appropriate form for performing various mining operation, which is known as Data Transformation. All this four steps is together known as the process of Data Pre-processing. As these steps ensures proper implementation of various data mining algorithm.

Data mining algorithm is then implemented in this dataset for extraction of data patterns. Next step is required to identify the patterns represented from the extracted data, also known Pattern Evaluation. This mined knowledge is represented using various visualization techniques to give more helpful information to the user.

Hence, in order to extract the proper patterns hidden in the datasets, the process of Data Pre-processing is very important as the real time dataset suffer from various quality problem like incompleteness, redundancy as well as inaccurate data. According the researches don before it was found that 60 percent of the effort is spent in data pre processing as only an efficient process can improve the quality of data. Hence, with improve data quality assures better quality of knowledgeable patterns extraction of the data.

The most important aspect of the data pre processing step is the Data cleaning. This step makes the dataset free from noisiness and irrelevant data making the dataset more reliable for data mining algorithms. It also identifies the inconsistent data as well as missing data and replaces them with most appropriate imputed value from remaining dataset. Hence, handling of this missing data is one of the significant tasks of Data cleaning process.

The incompleteness of data leads to the issue of missing value problem. The reason of missingness can

Corresponding author: Swarnendu Kundu |

be many as for example human error, malfunction of system, deletion of data, and inconsistency of data with other records or refusal to respond by the user. The major causes are illustrated with help of examples:

Procedural factors: Data entry can have many faults because this work is majorly done by humans. Inaccuracies in classification of new data can results to error identification of data wherever illegal codes are passed into the database. The error in the data estimates the prediction of invalid patterns and deductions have been made. If the relation factors or relations are skewed, it gives rise to error in association rules. Sometimes database are restored with new set of data, and there is no response from questionaries' make the process further more complicated. When a greater part of appellant unable to answer any particular question. If respondent fails to complete this kind of data then the investigator goes to wrong or invalid path of investigation which may lead to deletion or misclassification to flawed data. [1]

Refusal of response: Some question can be such that it becomes personal to the respondents and they become very sensitive in answering those questions. As for example an individual may have no views about political affairs or may not have any religion faith, hence the answers of such question have no data to be stored. Further some other question of education level, income or age factor can some of the private question which may some respondent not willing to answer. Student or some inexperienced person may have insufficient knowledge to answer certain questions. They are often giving polls for their future hopes and aims for their career objectives. [1]

Inapplicable responses: Some answer of the question remained blanks such question are general public question rather than any individual question. For example, any post graduate or scholar student may leave the social activities to be blank as they don't have much time for this. Like, the adult who are unmarried or widow or divorced may not likely to answer questions regarding their marital life. [1]
Handling of missing data can be categorized into three different classes:

Missing completely at random (MCAR):

Here, the missing values are in complete level of haphazardness .i.e. the possibility of finding pattern is quite low. So, imputation in this kind of dataset is very risky as we cannot find the appropriate pattern for imputation. This kind of missingness is found in real time dataset[2]

2. Missing at random (MAR).

When the possibility of having a missing value of an attribute may or may not depend on the familiar variables, but not on the value of the missing data itself. [2]

3. Not missing at random (NMAR).

When the possibility of having a missing value of an attribute could directly or indirectly depends on the value of some other variable in the data set. [3]

Literature Survey

The data mining have various methods for prediction; a researcher (DursunDelen in 2004) came forward with comparative study between three methods of data mining using breast cancer dataset. The popular methods are logistic regression, decision tree and artificial neural network. Logistic regression is known as the extension of linear regression. It is mainly used in predicting the dependent attribute in a multi-variable dataset. This model is built such that in problem class of two if number of odds is greater than 50%, it would assign 1 otherwise it builds the model to0 would be assigned. [4]

Another algorithm widely used for the imputation of missing values is random forest (in 2009 Adam Pantanowitz). He used HIV seroprevalence data for comparison among five methods. They are random forest, auto associative neuro-fuzzy, hybridisation of neural network and random forest and also auto associative neuro-fuzzy with genetic algorithm.
Random Forest algorithm has better performance with time perspective with high accuracy over single classification and regression trees and gives an efficient estimation for missing value and outlier location. [5]

The weight based clustering is proposed (Ilango Paramasivam 2014) in order to impute missing value for Type II diabetes data and then its performance is measure by Average Imputation Error. The clustering can be defined as the grouping if data based some similar properties. Thus this algorithm can be again classified into wo types. One is similarity based clustering and other one is model based clustering. The similarity based model shows how much similar the objects of dataset are whereas the model based methods se probabilistic approaches to evaluate the model. [6]

Another algorithm known as Miss Forest based on non-missing variables in the dataset in order to impute missing data. The Miss Forest is an open source package in R tool. This algorithm mostly outperforms multiple imputation method like K-

nearest neighbors for continuous data as well as categorical data.[7]

The important technique used is clustering (i.e. Grouping), this cluster analysis is the collection of different groups based some observed patterns. These patterns are is valid if the value of same cluster shows more similar than the values of different cluster. The pattern representation depends on number of stages like feature selection, feature extraction and the inter pattern similarity etc. [8]

Another alternative better clustering algorithm proposed by Greg Hamerly and Charles Elkan[9] ,in their paper demonstrates the test after effecting the predominance of the k-harmonic means algorithm (KHM) for discovering cluster of high caliber in low dimension.

Yasunari Yokota[10] proposed a new entropy estimator which minimizes mean square error between its estimate and true entropy .It is uses when an information source outputs two kinds of source symbols independently.

Jiye Li and Nick[11] proposed rough sets fit for prediction of missing data. By matching quality esteem sets among a similar center of the first data set, the appointed esteem saves the attributes of the first information set.

Samuel Kaski[12] introduce a clustering algorithm to explore the dependency among attribute. Tests information set are clustered with the end goal that the conditions between gatherings of various sets catch however much of pairwise conditions between the specimens as could reasonably be expected. The author additionally formalized the issue with a novel way, presenting advancement for Bayes factor.

Logistic Regression:

It is one of the popular regression models in order to find the dependent attribute, also for categorical data. This method is derived from that of standard logistic function. This function is helpful as it takes any value as an input i.e. from negative to positive esteems, but this yield values in-between zero and one and henceforth is interpreted as a probability.
The logistic function $\rho(t)$ can be defined with the following equation.

$$\rho(t) = \frac{1}{1 + e^{-t}}$$

The value of t is expressed as value can be expressed as follows where x is the given variable and α_1 and α_2 as the constants:

$$t = \alpha 1 + \alpha 2.x$$

Now, this function can be explained as:

$$\phi(x) = \frac{1}{1 + e^{-(\alpha 1 + \alpha 2.x)}}$$

Here, $\phi(x)$ represents the probability of an attribute being dependent, i.e. if the value is near to 1, it indicates a positive condition whereas the value near to 0 represents non dependency of variable.

K-means Clustering Algorithm

In the proposed technique, K-means clustering is used in order to impute missing value in the Pima dataset. Generally, clustering is done on unsupervised training data to discover patterns among the data. In this process, data of each cluster is highly similar with each other but data points of different groups are highly distinguishable. K-means algorithm is most popular technique of clustering. It works by calculating the centroid i.e. the mean for each of the k clusters. Initially, k numbers of cluster centers are randomly selected within the available data points. The remaining data points assigned to the cluster which is closest it. Iteratively, this algorithm improves the variation of clusters till the shifting of data points is observed as mean of the cluster is computed each time. (Han and Kamber 2006) In this paper, the value of k is assumed to be 2 and the clustering algorithm is applied on each of the dependent attribute which results in two clusters in each case.

Algorithm

1. The dependent attribute of dataset is found using logistic regression
2. The dependent attribute is clustered based on K-means algorithm
3. The group is identified having corresponding missing value and then the mean is calculated which is represented by $\mathcal{T}$.
4. Also the mean of the previous element and next element of missing value is calculated and it can be represented by ţ.

$$t = \frac{\left(\varphi_{prev} + \varphi_{next}\right)}{2}$$

where φ_{prev} represents the previous element of missing value
and φ_{next} represents the next element of missing value if any of them are null then it is considered zero.

5. Then, the value is predicted using this below formula:

$$\mathcal{T}_{imp} = t + (1 - \lambda)\mathcal{T}$$

Where λ is the smoothing factor and it is treated as a constant (lies between 0 and 1)

6. The steps 3 to 6 are repeated, until all the missing value of the attribute is imputed.

Dataset used:

Pima Indians Diabetes Dataset collected from UCI Repository with 768 records and 8 attributes with missing values. In this experiment, all the incomplete records are neglected and only 336 complete cases with 8 attributes is considered. At one instance of time, only one attribute is considered as missing with 5%, 10%, 15%, 20%, 25%, 30% and 35% of missing values iteratively for the entire dataset by missing the data completely at random. And then the performance of the proposed methodology is measured by computing the mean imputation error. Finally, we have compared this result with existing semi-supervised clustering technique.

Experiment and Result:

The techniques like Miss Forest and Exponential Clustering are performed with Pima Indians Diabetes Dataset collected from UCI Repository having 336 complete cases with 6 attributes which is used as training dataset for the above mentioned techniques. So, in order to analyse the performance of proposed methodology specific amount of data is missed completely at random and the average imputation error is computed. Table I shows the calculated average imputation error for the six attributes of Diabetes Type II dataset with different percentage of missing values with two clusters i.e. k=2. The average value of imputation error varies from 0.0170077 to 0.0787685 as the amount of missing value varies from 5% to 35%.

Table 1: Average imputation error for Pima Type II dataset

	No. preg	Plama glucose	Diastolic BP	Tricep.thickness	serum insulin	Age	Average
5%	0.0208306	0.0209474	0.0171043	0.0164151	0.013808	0.0129405	0.0170077
10%	0.0156716	0.0257182	0.0240198	0.0239174	0.0213822	0.0200343	0.0217906
15%	0.0233538	0.0402205	0.0386645	0.0393831	0.034802	0.0322938	0.0347838
20%	0.0279851	0.0494745	0.0527931	0.0528389	0.0459728	0.0417565	0.0451368
25%	0.0363806	0.06295	0.0734451	0.0675089	0.0580514	0.0517681	0.0583507
30%	0.0419776	0069984	0.0781368	0.0709431	0.0608388	0.0540905	0.0626618
35%	0.0531716	0.0989219	0.0974383	0.0857266	0.0730707	0.0642817	0.0787685

Figure 1: Visualization of Performance of the proposed methodology

Figure 1 shows the effect on performance of proposed methodology is inversely proportional to the amount of missing data. Each attributes results in little difference in performance which depends on the

size of missing data. As the size of missing data increases, the number of training records decreases which results in poor performance imputation. With less percentage of missing values, the performance of this methodology is quite high. This is due to the unavailability of similar records to make an efficient cluster for imputation. Hence, from this experiment we observe that as the percentage of missing value decreases, the performance of this methodology becomes quite higher than the existing techniques which is discussed in the next section.

Comparative Analysis:

A comparative analysis is made on the performance of proposed technique with that of the existing ones like Mean Imputation, EM Imputation, NORM, 10-NN, LS Impute Rows and Semi supervised Clustering technique [8] In order to test the performance of the algorithm, up to 35% of the data is missed completely at random and compared with those of existing techniques.

The below Table 2 shows the Average Mean Imputation ± Standard Deviation of the imputed dataset approximately up to first decimal place.

Table 2: Comparative analysis of average imputation error with standard deviation [8]

Method	Size of Missing Data						
	5%	10%	15%	20%	25%	30%	35%
10-NN	10.2±10.0	10.4±10.2	10.4±10.0	10.7±10.1	10.9±10.2	11.3±10.3	11.7±10.5
NORM	16.0±13.4	15.9±13.6	15.7±13.5	16.1±13.6	16.2±13.6	16.3±13.7	16.4±13.9
EMImpute_Columns	8.5±23.8	8.5±23.6	8.6±23.4	8.5±23.3	8.5±23.1	8.7±22.9	8.1±22.3
LSImpute_Rows	10.9±23.9	11.6±23.9	11.6±24.0	11.8±23.9	11.6±23.9	11.8±23.9	12.2±23.3
Mean Imputation	12.5±10.5	12.3±10.5	12.4±10.4	12.4±10.5	12.4±10.5	12.4±10.4	12.3±10.3
SESU_ CLUST	11.0±7.2	10.9±7.4	11.0±7.3	11.1±7.5	11.1±7.4	11.1±7.5	11.0±7.4
EXPO_CLUST	1.7±12.4	2.1±12.5	3.4±12.6	4.5±12.7	5.8±12.9	6.2±12.9	7.8±13.4

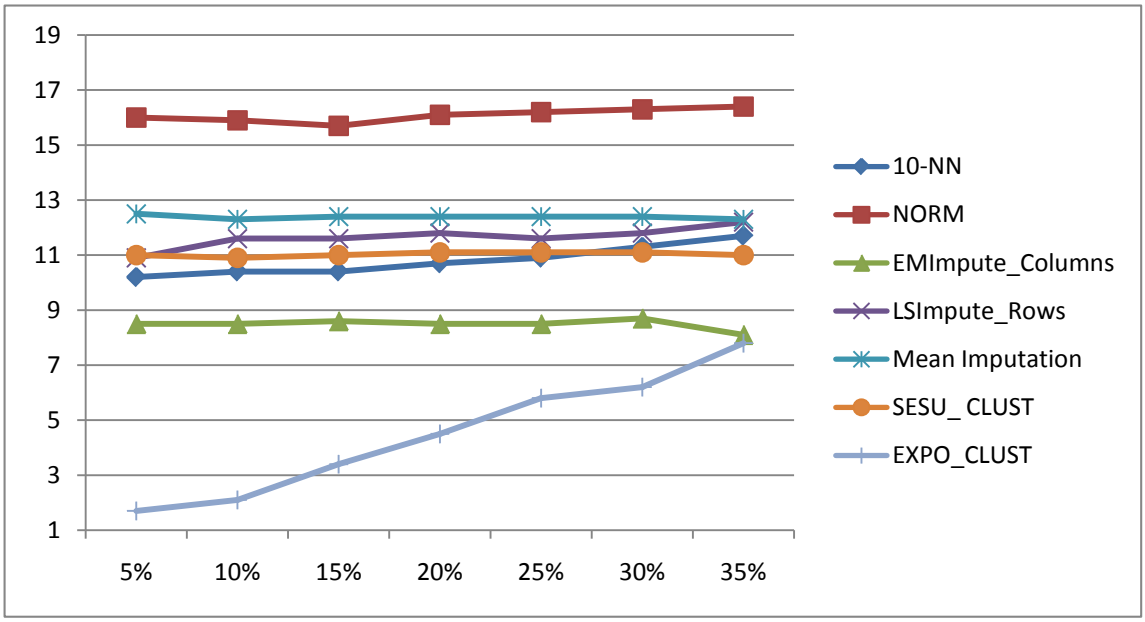

Figure 2: Visualization of performance different techniques vs exponential clustering

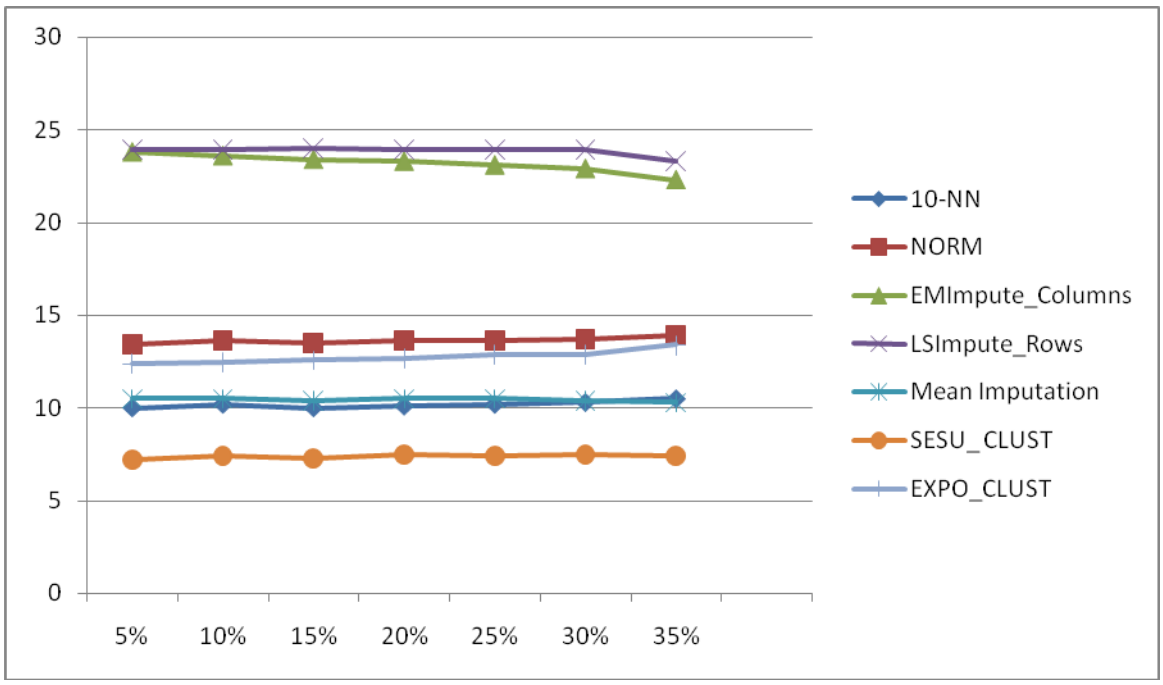

Figure 3: Comparative Analysis of different imputation methods w.r.t Standard Deviation

Figure 2 and 3 represents the nature of imputation method with respect to that of performance (mean imputation error) and standard deviation respectively. The mean imputation error is computed to be highest in NORM methodology with the range of 16.0 to 16.4. This performance followed with Mean Imputation, LSImpute Rows, 10-NN, SESU_CLUST with the range 12.3 to 12.5, 10.9 to 12.2, 10.2 to 11.7, 7.2 to 7.4 respectively. Whereas, the proposed method EXPO_CLUST have much lower mean imputation error having range 1.7 to 7.8 i.e. error increases with increase amount of missing values.

LSImpute Rows and EM columns are noticed to have the highest standard deviation whereas the SESU_CLUST is observed to have the lowest standard deviation. In the NORM technique for ascription the standard deviation shifts from 13.4 to 13.9 with the mean error rate differing from 15.7 to 16.4. The Mean Imputation performs with the mean error rate varies from 12.3 to 12.5 and the standard deviation changing from 10.3 to 10.5 which is higher than SESU_CLUST. But, the proposed methodology is observed to have the standard deviation 12.4 to 13.4 which quite near to the standard deviation of the original dataset.

Conclusion:

The efficient utilization of this missing data imputation is critical for associations to remain focused in today's complex, advancing environment. The associations confront a great deal of difficulties when attempting to manage extensive, assorted, and frequently complex databases. They embrace a few methodologies to enhance the nature of information in the database. The missing value, one of the inescapable issues in information examination, is taken care of utilizing different methodologies in view of the issue setting. In this paper, we have analysed different methodology and compared their performance with the help of mean imputation error. The result shows that this error is much lower than that of the existing technique in the selected dataset. The performance of this proposed methodology is highly varies with dependent variable and the number of clusters. The variation of performance of this methodology with respect to different number of clusters can be considered as future work of this paper.

Reference:

1. Marvin L. Brown John F. Kros, (2003),"Data mining and the impact of missing data", Industrial Management & Data Systems, Vol. 103 Iss 8 pp. 611 – 621
2. Sentas, P., & Angelis, L. (2006). Categorical missing data imputation for software cost estimation by multinomial logistic regression. Journal of Systems and Software, 79(3), 404-4
3. Batista, G. E., &Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533
4. Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34.2

(2005): 113-127.Evaluating the Impact of Missing Data Imputation through the use of the Random Forest Algorithm

5. Pantanowitz, Adam, and Tshilidzi Marwala. "Missing data imputation through the use of the Random Forest Algorithm." Advances in Computational Intelligence. Springer Berlin Heidelberg, 2009. 53-62.
6. Ilango, Paramasivam, Hemalatha Thiagarajan, and Nickolas Savarimuthu. "A semi-supervised clustering by λ_Cut for imputation of missing data in Type II diabetes databases." (2009).
7. Osadnik, Tadeusz, et al. "Comparison of modification of diet in renal disease and chronic kidney disease epidemiology collaboration formulas in predicting long-term outcomes in patients undergoing stent implantation due to stable coronary artery disease." Clinical Research in Cardiology 103.7 (2014): 569-576.
8. Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.
9. Hamerly, Greg, and Charles Elkan. "Alternatives to the k-means algorithm that find better clusterings." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.
10. Yokota, Yasunari. "An entropy estimator with least square error [biological signals]. " Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/ BMES Conference, 2002. Proceedings of the Second Joint. Vol. 1. IEEE, 2002.
11. Li, Jiye, and Nick Cercone. "Assigning missing attribute values based on rough sets theory." 2006 IEEE International Conference on Granular Computing. IEEE, 2006.
12. Kaski, Samuel, et al. "Associative clustering for exploring dependencies between functional genomics data sets." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2.3 (2005): 203-216.