International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Missing Values Imputation using Feed Forward Neural Network

Saravanan P, Justin Samuel S, Nirmalrani V, Mathivanan G

Abstract*: *Data cleaning intends to make the data as highly qualified one in terms of completeness and noisy free to makes the pattern outcome as high quality. Since most of the real world data possess these crucial issues, finding the most probable value for the holes introduced in data collection becomes a challenging task. This paper attempts to employ the feed forward neural network to make the collected dataset as complete which in turn the pattern outcome also complete. The collected dataset which possess the missing values is used to generate the identity matrix where the filled cells might get one and rest of the cells as zero. The given dataset gets normalized using minmax variety after replacing the missing cells as zero which will become a target matrix. By adjusting the weight values for the edges across the various edges the net value gets computed. The process gets repeated with a small increment done over the input to reach the target till the loss function yields the desirable value. The method is experimented with various UCI machinery dataset for different standard missing ratios. The proposed system performance is evaluated through RMSE parameter and the above method shows better accuracy with other popular methods.

Keywords* : *Data Cleaning, Missing Values, Imputation, RMSE, Pattern Discovery.

I. INTRODUCTION

Data Mining aims to unwrap the pattern from the collected data. The quality of pattern is influenced by the quality of the data completeness, correctness and consistency. Missing data is the common and major issue in various research and scientific domains like Data Analytics, Medicine, and IoT. Nowadays Data Collection became automated through various methods. Missing data arise because of various reasons and sources. The various sources for missing data include non-response, data transmission and misinterpretation. Rubin classified the missingness into three classes based on the nature of missingness: namely MCAR-Missing Completely At Random, MAR – Missing At Random and MNAR – Missing Not At Random. Dataset which has the data instances with missing cells which is completely independent of other complete instances and the instances which possess the degree of missingness are said to inhibit the properties of MCAR whereas if there exists certain rate of dependency between the record instance which possess the missingness and other complete or incomplete instances are said to have MAR attribute. The third variety is somehow interesting, where there exists a dependency among the instances which possess the missing values and no missing values. Missingness in the dataset makes the data analysis as ambiguous because of incompleteness which in turn the discovered pattern as incomplete. The issue of incompleteness leads to severe consequences in the quality of pattern discovered in terms of completeness after the mining process, which in turn the quality of decision gets affected. A variety of classical and enhanced techniques have been adopted and proposed which aims to fill up the missing cells with most probable values using the other instances which don't have the missing cells. The classical methods include substituting missing values with mean, median and other statistical estimated values. Statistical prediction can also be performed by designing a better model using the set of complete instances in the dataset. The entire process of making the dataset as complete is referred as missing data imputation. Many articles published in this field deal with the development of new effective methods for imputing the missing values; however only few studies were done over the high dimensional dataset to evaluate the effectives of the existing imputation methods which gives a set of guidelines to get a better methodological choice, which can be adopted in practice. Many methods were proposed conventionally to handle the missing values which includes list wise deletion also called as case exclusion where missing cases (tuples) gets excluded for analysis. Even though this method is simple and straight forward it results greater loss of data for analysis which makes the process to loss its generality. In order to avoid the case loses many imputation methods were proposed where the reasonable guess made for the missing cells and makes the data for analysis as complete. Most commonly used methods include marginal mean imputation, conditional mean imputation. Marginal mean imputation compute the mean for the attribute and substitute the value for missingness whereas in conditional mean imputation the model gets designed between the variable which possess the unfilled cells with other attributes which doesn't have that issue in the linear or nonlinear fashion.

Manuscript published on 30 September 2019
* Correspondence Author
Saravanan P*, Research Scholar, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India. Email: saravanan.it@sathyabama.ac.in
Justin Samuel,Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, India.
Nirmalrani V, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India. Email: nirmalv76@gmail.com
Mathivanana G, School of Computing, Sathyabama Institute of Science and Technology, Chennai, India.

The designed model gets used to predict the probable value to make the dataset as complete for analysis which in turn the analysis results become complete.

II. RELATED WORK

A new and effective missing value imputation based on clustering approach called CRI (Clustering-based Random Imputation) fill up the instances which possess the missing values with the plausible values that are generated from the data similar to this instance using a kernel-based random method. For the given dataset which possess the filled values are grouped using the clustering concept. The instances with missing values are assigned to a group which is most related to it. Finally, missing values of an instance gets calculated and made as member of that group [1]. Missing value problem in data mining gets considered and evaluates some of the methods generally used for missing value imputation. Three simple missing value imputation methods are implemented namely Constant substitution, Mean attribute value substitution and Random attribute value substitution [2]. First, grey relational analysis is employed to determine the nearest neighbors of an instance with missing attribute values. The known attribute values derived from these nearest neighbors are used to infer those missing values. Experimental results indicate that the accuracy of classification is maintained or even increased [3]. Multiple imputations present a useful approach for dealing with missing values in dataset. Instead of filling in a single value for each missing value, this procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute [4]. Deals the missing data imputation with heterogeneous attributes by contributing for both continuous and discrete data. This study proposes a higher order spherical kernel based iterative estimator to impute mixed-attribute data sets. Spherical kernel based estimator will give better results than other estimators. The Authors didn't consider the other correlated samples to the samples which possess the missing values [5]. The Fuzzy Possibilistic C Means Clustering Algorithm is optimized by Support Vector Regression and Genetic algorithm and used for establishing the clusters and based on the cluster member similarity the missing values gets imputed. The Authors assumed that the missing data are Random [6]. Complete datasets are clustered using KFCM (Kernel Fuzzy C Means) technique. Following which the incomplete cells are computed using GRA and entropy based multiple imputations [8].

III. PROPOSED METHOD

The algorithm takes the dataset which possess the missing values due to any of the sources and reasons. The identity matrix gets framed for the given dataset by assigning one for the missing cells and the zero for all the remaining filled entries. Following that in order to make the numerical computation the missing cells in the given dataset D gets filled with zero. The normalization process gets applied over the D to make the values in wide range as small range ie., between -1 to 1 which makes the computation as easier and normalize the residual error. The normalized dataset D called as X gets fixed as a target matrix to which the feed forward neural network need to be trained by adjusting the weight values across the various layers. The weight values between the input layer and hidden layer say U and between the hidden to output layer weight values say V are assigned with small random values. For the input dataset values and weight values the net value across the various layers gets computed by using net function. The calculated result ie., the result of net function produced by the output layer is multiplied with the identity matrix to compute the gradient which need to be introduced for the subsequent iteration. For the produced output the cost function MSE is computed: if the cost function results in tolerable value the computation process stopped, the algorithm returns Z as a result else with the small adjustment with the X values the process continues.

Input: *Dataset D which possess the missing values.*

Output: *Complete Dataset D.*

Step:1. *Form an identity matrix I for missing values*
$I_{ij} = 1$ *if* D_{ij} *is missing else 0*

Step: 2. *Replace all the missing values in D as zero.*

Step: 3. *Normalize D*
X= map minmax(D,-1,1)

Step: 4. *Set Target Matrix T=X.*

Step: 5. *Initialize weight matrices U & V with small random values.*

Step: 6. *Compute*
Net1= U.X , Net2 = V.X
Y=f(Net1), Z=f(Net2)

Step: 7. *Compute mv = Z * I and X = X+mv*

Step: 8. *Evaluate err = MSE (e) where e = T – Z*

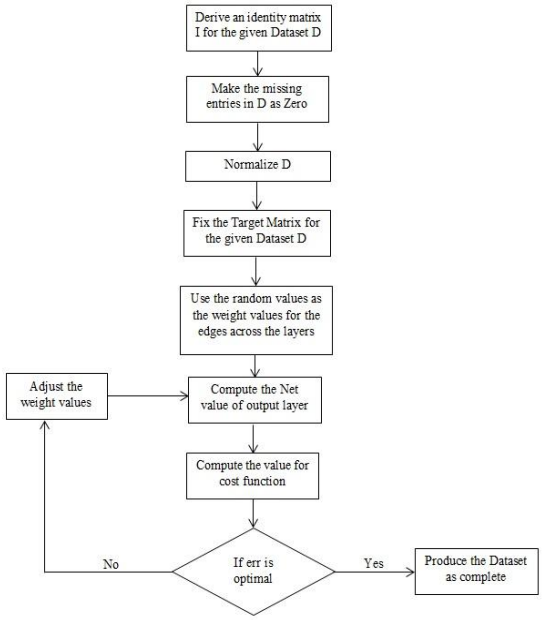

Fig. 1 Process Flow of the Proposed Model

Retrieval Number: B2503078219/19©BEIESP
DOI:10.35940/ijrte.B2503.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. EXPERIMENTAL SETUP AND PERFORMANCE ANALYSIS

RMSE is most widely and commonly used parameter to measure the value of cost function which is the difference between values predicted by a proposed method and the values actually observed. It also helps to evaluate the predictive capability of the imputation methods for the quantitative attributes.

$$RMSE = \sqrt{\frac{(X_i - Y_i)^2}{m}} \quad (1)$$

Where,

X_i is the estimated value

Y_i is the original value

m is the no. of estimated values

Wine dataset is selected to test the validity of the proposed algorithms. Wine database contains 178 instances and 13 attributes. The variable values are real. Upon the selected Dataset the proposed algorithm is used to estimate the values for various standard missing ratios and accuracy is summarized with the contrasting system.

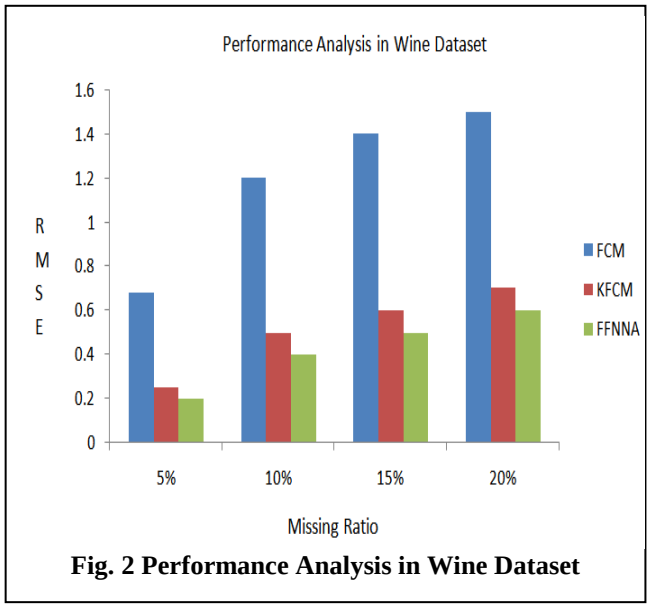

Fig. 2 Performance Analysis in Wine Dataset

V. CONCLUSION

Missingness in data collection is the major and common issue which distorts the quality of pattern mined. In order to resolve the incompleteness in pattern which is used for decision making the proposed method is intended to make the dataset as complete by estimating the probable values. All the complete instances are used to train the neural network and optimize the weight values for the edges across the various layers. The trained feed forward network is used for instances which possess the incomplete cells to make it as complete. The missingness gets induced purposely at random for various standard missing ratios and the algorithm gets tested with the contrasting one. The accuracy gets calculated and it shows that the proposed method outperforms.

REFERENCES

1. Amanda N. Baraldi, Craig K. Enders Arizona State University, "An introduction to modern missing data analyses United States 2009.
2. Eduardo R. Hruschka1, Estevam R. Hruschka Jr.2, and Nelson F. F. Ebecken "Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values", Advances in Artificial Intelligence, pp. 723 – 734, 2003.
3. R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," IEEE Transactions on Systems, Man, and Cybernetics B, vol. 31, no. 5, pp. 735–744, 2001.
4. Joseph L. Schafer and Maren K. Olsen The Pennsylvania State University "Multiple imputation for multivariate missing-data problems: a data analyst's perspective", 1998.
5. Ganga.A.R, B.Lakshmipathi, "Higher Order Kernel Function Algorithm for Imputing Missing Values", International Journal of Advanced Research in Computer Science, Vol. 3, No. 3, pp. 271-275, 2012.
6. P.Saravanan, P.Sailakshmi, "Missing Value Imputation Using Fuzzy Possibilistic C Means Optimized with Support Vector Regression and Genetic Algorithm", Journal of Theoretical and Applied Information Technology, Vol. 72, No. 1, pp. 34 – 39, 2015.
7. Nirmalrani V, Sakthivel P , "A Hybrid Access Control Model with Multilevel Authentication and Delegation to Protect the Distributed Resources", Journal of Pure and Applied Microbiology (JPAM), Vol. 9 (Spl. Edn. 2), pp. 595 – 609, 2015.
8. P.Saravanan, S. Justin Samuel, V. Nirmalrani, "Missing Data Imputation Using Kernel Fuzzy C Means Clustering Via Gray Relational Analysis", International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing" – EECMC-2018.

AUTHORS PROFILE

Saravanan P, Research Scholar of Sathyabama Institute of Science and Technology, Chennai, He Completed M.E.(C.S.E.) and his areas of interest includes Data Analytics, Data Mining, Algorithms and Optimization.

Justin Samuel S, Professor, Department of C.S.E., P.S.N. College of Engineering and Technology, Tirunelveli. He completed his Ph.D., degree in Web Services. His areas of interest include Web Services, Data Analytics, Big Data.

Nirmalrani V, Associate Professor, Department of I.T., Sathyabama Institute of Science and Technology, Chennai. She completed her Ph.D., degree in Network Security. Her areas of interest include Network Security & Role Delegation, Data Analytics, Big Data.

Mathivanan G, Associate Professor, Department of I.T., Sathyabama Institute of Science and Technology, Chennai. He completed his Ph.D., degree in Neuro Fuzzy Systems. His areas of interest include Neuro Fuzzy Systems, Data Analytics, Big Data.