International Journal of Advanced Scientific Technologies in Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Multi kernel fuzzy clustering with Lion neural network for missing Data Imputation and classification

R RAJANI[1], PROF T.SUDHA[2]
1MCA Department, NARAYANA ENGINEERING COLLEGE, NELLORE, INDIA
2MCA Department, Sri Padmavathi Mahila University, TIRUPATI, INDIA
1rajaninec@gmail.com, 2 thatimakula_sudha@gmail.com

ABSTRACT: The primary intention of my research is to design and develop an approach for missing data imputation and classification using hybrid prediction model. The hybrid model combines the multi kernel fuzzy clustering with the proposed lion neural network. Multi kernel fuzzy clustering is a recent clustering mechanism for grouping the data records. Lion neural network will be newly developed by combining the neural network with lion optimization. So, a new training algorithm will be applied to feed forward neural network to predict the missing data. Overall, the input data having the missing data will be given as input for the proposed approach. Here, missing attributes are considered as class attribute. Initially, lion neural network will be trained by considering the missing attribute as class attribute. In the testing phase, the missing data will be predicted by the lion neural network. Also, the same missing attribute will be computed through multi fuzzy clustering using the averaging process of centroids. Then, these two imputed values will be effectively combined to find the final value of missing data. Once the missing attributes are identified, the classification will be performed using the proposed lion neural network.

Keywords: multi kernel fuzzy clustering, lion neural network, lion optimization

I. INTRODUCTION

Data mining algorithms are frequently used for knowledge discovery in databases since they are non-trivial processes of exploring the new facts and identifying helpful relationships or patterns in data [7]. Data incompleteness or missing data is a pervasive data quality problem in all kinds of databases in the data mining [1]. Missing data imputation aims at providing estimations for missing values by reasoning from observed data. Because missing values can result in bias that impacts on the quality of learned patterns and the performance of classifications, missing data imputation has been a key issue in learning from incomplete data [11]. There are three main types of missing data mechanisms; they are (1) Missing At Random (MAR): the probability of missing values depends on other observed responses (2) Missing Completely At Random (MCAR): the missing values do not have any relationship with other responses in the dataset. (3) Not Missing At Random (NMAR): sometimes, missingness occurs only on specific values of interest and hence the missingness depends on the missing values themselves [16]. The simplest way of dealing with missing values is practical only when i) the data contain a relatively small number of observations containing MVs, or when ii) the analysis of the complete examples will not lead to a serious bias during the inference [10].

The process of filling the missing data in the dataset is termed as data imputation. The goal of the imputation method is to reduce the bias of survey estimates. Imputation of missing data minimizes bias and allows for analysis using a dataset, so that standard analysis can then proceed [13]. There are many approaches to deal with the problem of missing values: (a) Ignore the objects containing missing values; (b) Fill the gaps manually; (c) Substitute the missing values by a constant; (d) Use the mean or the mode of the objects as a substitution value; and (e) Get the most probable value to fill the missing ones [12]. An important aspect in missing data imputation is the pattern of missing values because it determines the selection of an imputation procedure. Then, the data imputation is divided into two types, which are single imputation and multiple imputations [14]. Single imputation techniques indicate the substitution of a single value for each missing data. Such as mean imputation, regression and hot-deck imputation. Multiple imputation substitutes each missing value with a set of plausible values that are obtained from observed data, resulting in multiple completed data sets that allow imputation uncertainty to be incorporated into statistical inferences [15].

In missing data imputation, the learning task becomes classification if the missing attribute is nominal, whereas it becomes regression if the missing attribute is continuous. For each instance with a missing attribute, a machine learning algorithm is trained based on the instances without missing values and the non-missing values of the instance are used by the model to predict the target missing attribute value [19]. Recently, the machine learning imputation methods have been developed. They estimate missing values by constructing a predictive model to estimate the absent values from information in the dataset [4]. However, there are two other important restrictions: 1) imputation methods evaluation cannot be properly evaluated apart from the modelling task and 2) complete-case analysis, should be avoided, where information of instances or attributes with missing values are removed [22]. Thus, the machine learning based methods include self organizing feature map (SOM) [5],

K-nearest neighbour [2], multi-layer perception [17], fuzzy-neural network, auto- associative neural network imputation with genetic algorithms [18] etc. Machine learning-based classification is one of the main tasks in machine learning, data mining, and pattern recognition. Over the years, many solutions to classification task have been developed and several quite interesting classifiers are now available for various applications like, sensor networks, monitoring, traffic management, telecommunication, web log analysis, medical data and so on [5].

II. REVIEW AND COMPARISON

I studied Retrieving and inferring based methods by Zhixu Li, Lu Qin, Hong Cheng *et al.*[1]. The advantages of these methods include Inference rules and retrieving queries are used to other attribute dependencies. The drawbacks are of an un-inferable missing value cannot be inferred even if all the other missing values in the table are known. The K-Nearest Neighbour by Jose Luis Sancho-Gomez *et al.* Says that Efficiency is improved in both artificial and real classification datasets. But, whenever the KNN impute for the similar instances, it is difficult to search through all the dataset. To enhance the performance by using activation function suggested by Particle Swarm Optimization by Chandan Gautam and Vadlamani Ravi [3] includes the drawback that it does not preserve the covariance structure of the original data. If we prefer Fuzzy Clustering by Loris Nannia *et al.* [4], the performance of classification is significantly improved, whereas it is not suitable when the number of features and a sample is very high. The missing data of different variables is also imputed at the same time as per Kohenen Self-organizing maps Laura Folguera *et al.* [5], but, the large proportions of missing values might lead to erroneous results. When we try to go with classification by referring Adaptive matching of classifiers by Jaemun Sim *et al.* [6], it is superior in terms of scalability and accuracy. But, while increasing the number of matching, it requires more elapsed time. When I tried to study clustering based on Fuzzy clustering based EM approach by Md. Geaur Rahman and Md Zahidul Islam [8], it achieves higher quality of imputation, but, it is difficult to formulate when the missing values have a random nature. When tried to decrease imputation errors a paper on by Nearest neighbour method by Gerhard Tutz and Shahla Ramzan [7], The performance gets degraded when the correlation is low.

III. PROBLEM STATEMENT

The causes of missing values are the most diverse and related to the application domain, such as drawbacks in the data acquisition, measurement errors, sensors network problems, data migration failures and unwillingness to respond to survey questions [22].

The presence of missing data [15] is a general and challenging problem in the data analysis field. Thus, the data imputation techniques refer to any strategy that fills in missing values of a data set so that standard data analysis methods can be applied to analyze the completed data set.

The missing data imputation becomes the challenging task in data mining since the missing value leads to degrade the quality of the data, generate bias and mitigate the quality of the data [18, 21].

In data classification, the presence of missing values can lead to critical problems during the learning process, such as a loss of efficiency, biased data structure, analytical difficulties, and prediction performance degeneration [19].

Incomplete data in either the training set or test set or in both set leads to affect the prediction accuracy of learned classifiers. The seriousness of this problem depends in part on the proportion of missing data and high dimensionality of the problem [23, 24].

The challenge of the efficient classifiers is that the handling of datasets which are constantly changing their patterns of the missing data, data volume and data structure [6].

IV. PROPOSED METHODOLOGY

The primary intention of my research is to design and develop an approach for missing data imputation and classification using hybrid prediction model. The hybrid model combines the multi kernel fuzzy clustering with the proposed lion neural network. Multi kernel fuzzy clustering is a recent clustering mechanism for grouping the data records [26]. Lion neural network will be newly developed by combining the neural network with lion optimization [27]. Here, feed forward neural network will be taken and the training algorithm of levenberg marquardt algorithm will be modified with the lion optimization. So, a new training algorithm will be applied to feed forward neural network to predict the missing data. Overall, the input data having the missing data will be given as input for the proposed approach. Here, missing attributes are considered as class attribute. Initially, lion neural network will be trained by considering the missing attribute as class attribute. In the testing phase, the missing data will be predicted by the lion neural network. Also, the same missing attribute will be computed through multi fuzzy clustering using the averaging process of centroids. Then, these two imputed values will be effectively combined to find the final value of missing data. Once the missing attributes are identified, the classification will be performed using the proposed lion neural network. The proposed model will be implemented using MATLAB and the results will be compared with existing algorithms using MSE and classification

accuracy. The UCI datasets like, Iris, Wine, and heart disease datasets will be utilized for the experimentation.

REFERENCES

[1] Zhixu Li, Lu Qin, Hong Cheng, Xiangliang Zhang, and Xiaofang Zhou, "TRIP: An Interactive Retrieving-Inferring Data Imputation Approach", IEEE Transaction on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2550-2563, August 2015.

[2] Jose Luis, Sancho-Gomez, Anibal R.Figueiras Vidal and Michel Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation", Neurocomputing, vol. 72, no. 9, pp. 1483-1493, 2009.

[3] Chandan Gautam and Vadlamani Ravi, "Data imputation via evolutionary computation, clustering and a neural network", Neurocomputing, vol. 156, pp. 134-142, May 2015.

[4] Loris Nanni, Alessandra Lumini and Sheryl Brahnam, "A classifier ensemble approach for the missing feature problem", Artificial Intelligence in Medicine, vol. 55, no. 1, pp. 37-50, May 2012.

[5] Laura Folguera, Jure Zupan, Daniel Cicerone and Jorge F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices", Chemometrics and Intelligent Laboratory Systems, vol. 143, pp. 146-151, 2015.

[6] Jaemun Sima, Ohbyung Kwon and Kun Chang Lee, "Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in datasets" Expert Systems with Applications, vol. 46, pp. 485–493, 2016.

[7] Gerhard Tutz a,∗, Shahla Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods", Computational Statistics and Data Analysis, vol. 90, pp. 84-99, October 2015.

[8] Md. Geaur Rahman and Md Zahidul Islam, "Missing value imputation using a fuzzy clustering based EM approach", Knowledge and Information Systems, vol. 46, no. 2, pp. 389-422, February 2016.

[9] Jing Tian, Bing Yu, Dan Yu and Shilong Ma, "Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering", Applied Intelligence, vol. 40, no. 2, pp. 376-388, March 2014.

[10] Marcilio CP de Souto, Pablo A Jaskowiak and Ivan G Costa, "Impact of missing data imputation methods on gene expression clustering and classification", BMC Bioinformatics, vol. 16, February 2015.

[11] Xiaofeng Zhu, Shichao Zhang, Zili Zhang, and Zhuoming Xu, "Missing Value Estimation for Mixed-Attribute Data Sets", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 110-121, January 2011.

[12] Estevam R. Hruschka Jr, Eduardo R. Hruschka and Nelson F. F. Ebecken, "Bayesian networks for imputation in classification problems", Journal of Intelligent Information Systems, vol. 29, no. 3, pp. 231-252, December 2007.

[13] Ruey-Ling Yeh, Ching Liu, Ben-Chang Shia, Yu-Ting Cheng and Ya-Fang Huwang, "Imputing manufacturing material in data mining", Journal of Intelligent Manufacturing, vol. 19, no. 1, pp. 109-118, February 2008.

[14] Tae Yeon Kwon and Yousung Park, "A new multiple imputation method for bounded missing values" Statistics & Probability Letters, vol. 107, pp. 204-209, December 2015.

[15] Jianhua Wu, Qinbao Song and Junyi Shen, "An Novel Association Rule Mining Based Missing Nominal Data Imputation Method", In proceedings of IEEE International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel or Distributed Computing", pp. 244-249, August 2007.

[16] R. Devi Priya and R. Sivaraj, "Imputation of Discrete and Continuous Missing Values in Large Datasets Using Bayesian Based Ant Colony Optimization", Arabian Journal for Science and Engineering, pp. 1-13, May 2016.

[17] Jose M. Jerez, Ignacio Molina and Pedro J. Garcia-Laencina, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem", Artificial Intelligence in Medicine, pp. 105-115, 2010.

[18] Ibrahim Berkan Aydilek and Ahmet Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm", Information Sciences, vol. 233, pp. 25–35, 2013.

[19] Pilsung Kang, "Locally linear reconstruction based missing value imputation for supervised learning", Neurocomputing, vol. 116, pp. 65-78, October 2013.

[20] Ingunn Myrtveit, Erik Stensrud and Ulf H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods", IEEE Transactions on Software Engineering, vol. 27, no. 11, pp. 999-1013, August 2002.

[21] Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang and Chengqi Zhang, "Semi-parametric optimization for missing data imputation", Applied Intelligence, vol. 27, no. 1, pp. 79-88, August 2007.

[22] Fabio Lobato, Claudomiro Sales, Igor Araujo, Vincent Tadaiesky, Lilian Dias, Leonardo Ramos and Adamo Santana, "Multi-Objective Genetic Algorithm For Missing Data Imputation", Pattern recognition letters, vol. 68, pp. 126-131, December 2015.

[23] Dusan Sovilj, Emil Eirola, Yoan Miche and Kaj-Mikael Bjork, "Extreme learning machine for missing data using multiple imputations", Neurocomputing, vol. 174, pp. 220-231, January 2016.

[24] Julian Luengo, Jose A. Saez and Francisco Herrera, "Missing data imputation for fuzzy rule-based classification systems", Soft computing, vol.16, no. 5, pp. 863-881, 2012.

[25] C. J. Carmona, J. Luengo and P. Gonzalez and M. J. del Jesus, "A Preliminary Study on Missing Data Imputation in Evolutionary Fuzzy Systems of Subgroup Discovery", In proceedings of IEEE International Conference on Fuzzy Systems, pp. 1-7, June 2012.

[26] B R Rajakumar, "Lion Algorithm for Standard and Large Scale Bilinear SystemIdentification: A Global Optimization based on Lion's Social Behavior ", IEEE Congress on Evolutionary Computation (CEC), Beijing, China, 2014.

[27] Hsin-Chien Huang, Yung-Yu Chuang,Chu-Song Chen, "Multiple Kernel Fuzzy Clustering, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 20, NO. 1, pp. 120-134, FEBRUARY 2012

BIOGRAPHY

Mrs. R. RAJANI is an Associate Professor and heading the department of MCA, Narayana Engineering College, Nellore, AP, India. She guided many projects for B.Tech and PG students. Her research interests include Datamining, Query Optimization, Computer Networks and Software Engineering etc., She is pursuing her Ph.D from Sri Padmavathi Mahila University under the guidance of Prof.T.Sudha.