ISSN 2278-3091
Volume 9, No.1.3, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2191.32020.pdf
https://doi.org/10.30534/ijatcse/2020/2191.32020

Fuzzy C Means Imputation of Missing Values with Ant Colony Optimization

Farahida Hanim Mausor, Jafreezal Jaafar, Shakirah MohdTaib
Universiti Teknologi PETRONAS, Malaysia, farahida_17010728@utp.edu.my

ABSTRACT

Missing value is an error that always happened, and it is unavoidable. This error should be handled correctly before data is processed into processing model. This paper proposes a improved method of imputation by employing a new version of Fuzzy c Means (FCM) which hybridized with Evolutionary Algorithm to handle missing values problem. Missing values can be treated by imputing the values. The advantage of FCM is it can provide a better separation of instances where it is not well separated. It is a well-known classification method that can provide highest accuracy. It can be benefit from Ant Colony Optimization that can help to select only highly related feature to be process as an estimation for a missing value. Here, a traditional FCM basic is test as a cluster technique for imputed data.

Key words: Ant Colony Optimization, Evolutionary Algorithm, Fuzzy c Means, Imputation, Missing Values.

1. INTRODUCTION

Data mining which is also known as a knowledge discovery is a process of discovering a useful information through pattern of data[1]. Data mining is currently getting more attention from modern organization for a decision making process[2]. This happen because through a data mining process, an organization can be benefit from the result of large dataset pattern. It can be done through technique of computational or statistical. However, before these technique is executed to the collected data it must be prepared through data pre-processsing. Data pre-processing is a process that can be use to prepare the data before being process in a data mining technique [3]. Because most of the data is likely to be imperfect and contain noise(meaningless data)[3] one of this imperfection can be in term of missing values.

Missing values or incomplete data is happend because of the faulty during data sampling process and also because of the noise.
From previous research missing values is treated by using statistical, machine learning and also deletion methods. These method including action of deletion, zero mean estimation and also imputation.

Imputation technique for a missing values is proved by [4] can give a significant impact and influence the quality of data mining result. Regression Imputation for example having an advantage in approximation of missing values the missing values[5] however having a difficulties in terms of variable relation. K-nearest neighbour (KNN) imputation technique on the other hand, does not have variable issue but only use complete case to impute missing values because of its limited performance[6]. Other imputation technique such as artificial neural network which passing instances through the network is computationally costly[7].

In recent years, many researcher trying to improve imputation techique by looking up for the accuracy and also brigning down the computational cost. Because of that, a hybrid techniques of imputation was proposed. Hybrid Regression with ANFIS (Adaptive Neuro Fuzzy Inference System) for example which proposed by [8] was successfully improving the accuracy compared to other methods but having a computational cost issue. Through the literature review that has been conducted it shows that not all types of existing method can handle to process data especially inconsistent and big data. Current imputation technique such as maximum likelihood[9], KNN [10], SVM(Support Vector Machine)[11] and Regression [12] is less efficient for some types of missing values.

Therefore, in this research we focusing on improving an imputation technique that using Fuzzy C Means(FCM) imputation that will hybridized with Ant Colony Optimization Techniques.

Fuzzy C means is a clustering technique that can be manipulate to be use in imputation technique and it was invented because of the drawback that occurs in the traditional clustering[13]. Fuzzy clustering is modelled by a set of fuzzy that assigning each data item to membership function from 0 to 1. Membership function will show how strong the item or instance belongs to each cluster. The idea behind the clustering is to predict the missing values within a record using the information of the cluster which missing values located in that cluster[7]. The selection of cluster can be later use for a selection of subset.

Table 1: Evolutionary in Imputation Technique

Evolutionary Algorithm	Remarks	References
Ant Colony Optimization (ACO)	Turn feature selection problem into path finding problem Suitable for big data Low execution time, suitable for MCAR, MAR, MNAR data	[14]
Genetic Algorithm	Can be used to optimize membership function in FCM Not stated clearly whether it is suitable for big data	[15]
Genetic Programming	Used to build a classifier for each pattern of missing values. Works well with multiple imputation for data classification	[16]
Particle Swarm Optimization	Use to minimize functional error Works well with covariance matrix	[17]

Here, Ant Colony Optimization (ACO) is proposed to improve the selection. ACO is one of the Evolutionary Algorithm(EA) [18] that has been previously used in many research field including in data imputation. EA provide a solution to undefined problem by maintaining a population of 'individuals' and each of the individual is a candidate solution[19]. Table 2 showing EA that previously has been deployed in data imputation research field. From the EA that was listed in Table 2, ACO having an advantage when it comes to solving a feature selection problem. Moreover, ACO has a potential of turning feature selection problem into path finding solution that are suitable for a big data and have low execution time. [14].

Therefore, the objective of this research is to formulate imputation technique base on fuzzy c means and ACO.

Second is to impute missing values by optimizing feature selection approach and membership function in FCM. It is found that the selection of method to solve missing values is important because imputed data can have a big impact to a data mining process. In the view of the fact that, imputation technique can increase the capabilities of data mining model and type of missing data and option for imputation method can directly influence the prediction method when it is applied.

2. METHODOLOGY

This section comprises research activities, research framework and proposed model. Figure 1 shows the research activities that was conducted and will be conducted for further study.

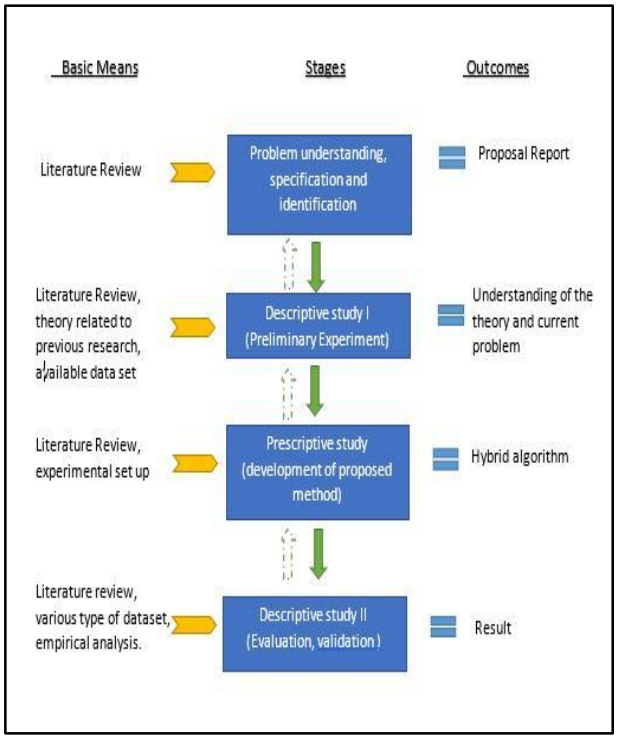

Figure 1: Research Activities

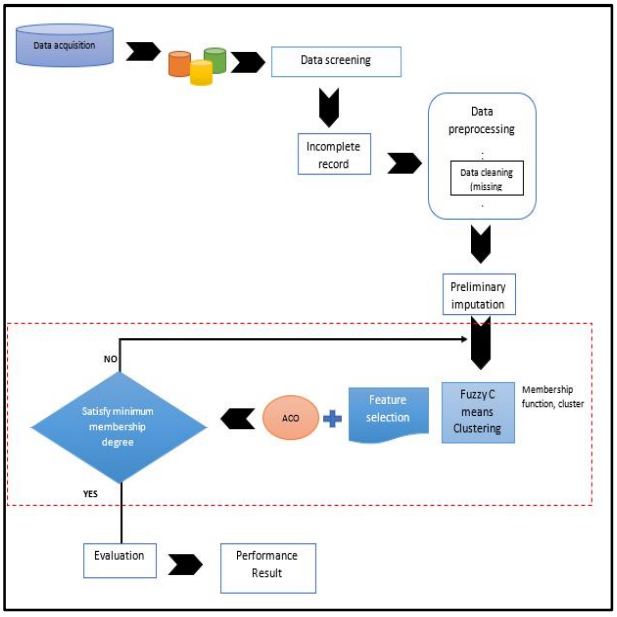

Figure 2: Proposed Framework

At this stage, problem identification, understanding and specification was identified through a meticulous literature review. Also, a preliminary experiment to proof the theory was conducted to shows that this study is feasible to be done. Before precede with the data processing, data acquisition is important to know the state of the data that need to be process.

	Cluster 1	Cluster 2	Cluster 3
1	0.0010720343	0.996623586	0.0023043797
2	0.0074979471	0.975852543	0.0166495094
3	0.0064145785	0.979825922	0.0137594999
4	0.0101075228	0.967427446	0.0224650314
5	0.0017679352	0.994470355	0.0037617094
6	0.0206196544	0.934574112	0.0448062334
7	0.0065045178	0.979491667	0.0140038150
8	0.0001412048	0.999547263	0.0003115325
9	0.0219024180	0.930379787	0.0477177955
10	0.0053415614	0.982722963	0.0119354753
11	0.0102009404	0.968042314	0.0217567454
	:		
	:		
	:		

Figure 3: Clustering Result

Figure 4: Cluster Plot of IRIS Data

Different types of data need a different approach of data cleaning. In data screening the type of data preparation including data cleaning, transformation, integration, normalization and missing data will be identified. For this research, as we focusing on missing data, so the process of treating missing data will be highlighted.

As mentioned earlier, imputation is one of the ways to treat missing data. Imputation using FCM technique having an advantage by providing a better separation of instances. Besides, providing a better feature selection of identification is foreknow to have a better imputation outcome.

As shows in Figure 2 preliminary imputation must be conducted before FCM techniques can take place. It is imperative because FCM algorithm cannot handle with missing values[7]. Beside, FCM also having a problem with the slower convergence speed[20] and sometimes it may lead to inaccurate data description because of the one point based membership[21]. Therefore, the propose method is to improve the objective function and distance function to increase the convergence speed. The advantage of ACO can help by choosing the best feature to help FCM functionality of convergence speed issue. For the preliminary experiment, K- Nearest Neighbor is used for basic imputation. Subsequently, the imputed data is divided into cluster using FCM algorithm as shown in Figure 3 and 4. Membership degree in Figure 3 is the main outcome for FCM. Whereas, for figure 4 cluster plot with *fviz cluster* in R software is used to show the dispersion of cluster. This research propose that a membership degree must satisfies a minimum condition. Then, feature selection will be determined and ACO will be injected to help to choose the best selection of feature for the imputation step to be complete. The next step is as shown in Figure 2.

3.RESULTS AND DISCUSSION

This section present experimental result on validating performance of the proposed method.

As to shows that FCM is feasible to be use in imputation we injected 10% of missing values in IRIS dataset. The performance was measure using RMSE (Root Mean Square Error).

RMSE is a well known performance indicator for quantitative measure. The measurement is based on the actual values and estimation of imputation model[2]. RMSE is measure as below

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_1 - y_2)^2} \qquad (1)$$

Where y_1 is the actual values and y_2 is the estimated values from the imputation model. Whereas *n* is the total number of prediction. Accoding to [7] the lower RMSE the better imputation result.

Table Below is the RMSE result from the conduxted experiment.

Table 2: RMSE Result of the proposed Method

MISSING RATE	MISSINGNESS MECHANISM	RMSE
10%	MAR(Missing At random)	0.26

From the Table 2, it is shows that RMSE for the proposed method is 0.26. This means that FCM without hybridized with other method having only 0.26 error when it is used for imputation with missingness mechanism of Missing at Random. However, having one dataset tested is not enough to quantify the result. In Reality, there are three types of missing mechanism first is Missing Completely at Random (MCAR), which happened when the probability or the observed missing values does not depend on any attribute. Second is Missing At Random (MAR) happen when the probability of the observed missing values depends on the other attribute but does not depends on the missing values. Third is Missing Not at Random (MNAR) when the probability of the observed missing values could depend on that attribute itself.

Therefore, our aim in future to formulate a method that capable to solve these three types of missingness.

4. CONCLUSION

Issues in data processing such as curse dimensionality, large dataset, variety of data types and missing values is always happening and continuous. The states of these type of data can contribute to ineffectiveness and slow down data processing model[22-23]. Some of the reason for this issue to arise is large dataset in mining algorithm. This issue needs to be catered and cannot simply ignored. This to ensure that the processing data will have an unbiased result. To overcome this issue many researchers come with the solution but it is still ongoing.

In this research, we proposed a solution for a missing value which is data imputation. Data imputation is an important technique that can be used to replace missing values in dataset. Fuzzy C means (FCM) imputation method is proposed to hybridized with Ant Colony Optimization (ACO) from Evolutionary Algorithm. The strength of FCM in terms of soft clustering making it suitable to be used with ACO that was proved to provide a better of selecting the feature in clustering of FCM. This combination expectantly to have a better imputation accuracy. There is a lot of improvement need to be done in a way to achieve the objective of this study. In this paper, the result that was showed previously is just a preliminary experiment.

This framework has not yet been tested as a whole technique that as shown in Figure 2. It is too early to prove that this experiment will outperform the other hybrid method. However, from the extensive literature review FCM and ACO has shown a significant advantage that can be used together to have a better result. Next, a larger dataset with the three types of missingness mechanism will be tested and the use of feature selection method will be applied for the further study.

REFERENCES

1. Available: https://www.techopedia.com/definition/1181/data-mining
2. M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," Knowledge and Information Systems, vol. 46, pp. 389-422, 2016. https://doi.org/10.1007/s10115-015-0822-y
3. S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," Big Data Analytics, vol. 1, p. 9, 2016.
4. U. Garciarena and R. Santana, "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers," Expert Systems with Applications, vol. 89, pp. 52-65, 2017. https://doi.org/10.1016/j.eswa.2017.07.026
5. S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing value imputation based on data clustering," in Transactions on computational science I, ed: Springer, 2008, pp. 128-138.
6. J. Van Hulse and T. M. Khoshgoftaar, "Incomplete-case nearest neighbor imputation in software measurement data," Information Sciences, vol. 259, pp. 596-610, 2014.
7. A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," Expert Systems with Applications, vol. 115, pp. 68-94, 2019.

8. H. Shahbazi, S. Karimi, V. Hosseini, D. Yazgi, and S. Torbatian, "A novel regression imputation framework for Tehran air pollution monitoring network using outputs from WRF and CAMx models," Atmospheric Environment, 2018. https://doi.org/10.1016/j.atmosenv.2018.05.055
9. P. T. von Hippel, "New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples," Structural Equation Modeling: A Multidisciplinary Journal, vol. 23, pp. 422-437, 2016.
10. H. M. de Silva and A. S. Perera, "Evolutionary k-nearest neighbor imputation algorithm for gene expression data," ICTer, vol. 10, 2017.
11. M. Zieba and J. Swiatek, "Ensemble SVM for imbalanced data and missing values in postoperative risk management," in e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on, 2013, pp. 95-99.
12. J. Chen, X. Zhang, K. Hron, M. Templ, and S. Li, "Regression imputation with Q-mode clustering for rounded zero replacement in high-dimensional compositional data," Journal of Applied Statistics, vol. 45, pp. 2067-2080, 2018/08/18 2018.
13. L. Himmelspach and S. Conrad, "Fuzzy clustering of incomplete data based on cluster dispersion," in International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2010, pp. 59-68. https://doi.org/10.1007/978-3-642-14049-5_7
14. R. D. Priya, R. Sivaraj, and N. S. Priyaa, "Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases," Knowledge-Based Systems, vol. 133, pp. 107-121, 2017.
15. J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation," Transportation Research Part C: Emerging Technologies, vol. 51, pp. 29-40, 2015.
16. C.T. Tran, M. Zhang, P. Andreae, and B. Xue, "Multiple imputation and genetic programming for classification with incomplete data," in Proceedings of the Genetic and Evolutionary Computation Conference, 2017, pp. 521-528. https://doi.org/10.1145/3071178.3071181
17. M. Krishna and V. Ravi, "Particle swarm optimization and covariance matrix based data imputation," in Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, 2013, pp. 1-6.
18. S. Dreyer, "Evolutionary feature selection," Institutt for datateknikkoginformasjonsvitenskap, 2013.
19. A.A. Freitas, "Basic Concept of Evolutionary Algorithm" in data Mining and Knowledge discovering with evolutionary algorithm ed new york Springer 2002, pp. 79-81.
20. J. Nayak, B. Naik, and H. Behera, "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014," in Computational intelligence in data mining-volume 2, ed: Springer, 2015, pp. 133-149. https://doi.org/10.1007/978-81-322-2208-8_14
21. S. Gan, S. Liang, K. Li, J. Deng, and T. Cheng, "Trajectory length prediction for intelligent traffic signaling: A data-driven approach," IEEE Transactions on Intelligent Transportation Systems, vol. 19, pp. 426-435, 2018. https://doi.org/10.1109/TITS.2017.2700209
22. S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, pp. 111-117, 2006.
23. Mohamad Haider Abu Yazid, Mohamad Shukor Talib, Muhammad Haikal Satria. Heart Disease Classification Framework Using Fuzzy and Flower Pollination Neural Network. International Journal of Advanced Trends in Computer Science and Engineering. Volume 8, No.1.6, 2019