A pragmatic approach of data imputation using fuzzy-based swarm for heart disease classification

Mohd Najib Mohd Salleh[1], Nurul Ashikin Samat[1], Kashif Hussain[1] and Abdul Mutalib Leman[2]

[1]*Faculty of Computer Science and Information Technology*
[2]*Faculty of Engineering Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia*
Email: najib@uthm.edu.my, gi140022@siswa.uhm.edu.my

In this study, we investigate the classification problem of heart disease with incomplete datasets. Our pragmatic approach is to exploit the potential of complete data for selecting relevant features in incomplete datasets. We define our approach by implementing fuzzy-based particle swarm optimization to impute missing values with tuning the exist structure with the data which leads to better solution. The FCM clustering is applied to identify the similar records in the complete dataset. We compared the Root Mean Square Error (RMSE) results of three different datasets with seven different ratios with range 1% to 20% of missing data. The experimental results provide the evidence that our approach performs better accuracy compared to other approach. The proposed method makes it possible to select relevant feature by offering good combination of its setting to classification of heart disease problem.

Keywords: FCM; PSO; Imputation; Preprocessing.

1. Introduction

Heart disease is the leading cause of death for human. Recently, hundreds of data mining works in various industry to extract the knowledge from a pile of data collection [1, 3, 4, 5]. Most data mining methods such as classification, clustering and regression have been implemented into clinical datasets for the accurate diagnosis of disease and proper treatment [2]. Thus, data quality is very important concern for the researcher towards knowledge extraction in healthcare industry. However, the real world data collection Stroke and other Cardiovascular Diseases is not always complete and accurate due to various uncertainty [6]. The collection process might be involved and tangled with uncertain situations and consequently incomplete dataset have been appeared. This reason might arise from different type of error, either human error, equipment error [7] and noise data, or irrelevant input data and missing value in highly dimensionality recorded in the dataset [8]. Therefore, preprocessing

process is considered important before training the data into the classifier models to maximize the accuracy and efficiency of machine learning techniques [9]. In this paper, incomplete dataset due to missing value recorded in the dataset is focused. If the missing dataset is not treated, apart from less accuracy problems, it might create bias result, and loss efficiency of computational process due to the holes in the dataset.

There are several ways to treat the incomplete dataset such as (1) Delete and ignore the missing data, (2) Parameter expectation, and (3) Imputation [10]. For delete and ignore method, it is not practical in several area of study as the data might contribute to the knowledge practical only when the data contain relatively small number of examples. While, for the second method are not efficient in computational time consuming and expensive. Imputation methods have various types, from mean imputation until study the relationship among attributes. Rubin [10] categorized missing data as: (1) Missing completely at random (MCAR), (2) Missing at random (MAR), and (3) Missing not at random (MNAR). Cohen and Cohen [13] have suggested developing missing data dummy codes for each variable with missing data, and using this missing variable code as a predictor in a regression model.

This paper is organized as follows: Section 2 will reviews the related works regarding FCM and PSO in imputation works, Section 3 will overviews the Fuzzy C-Means Clustering algorithm and Particle Swarm Optimization algorithm. It also describes about the proposed methodology, FCMPSO used in this paper. Section 4 will be explained about experiments and result analysis. Finally, Section 5 will provide the conclusion and recommendations.

2. Overview

Several clustering approaches have been proposed recently to handle the missing values directly. We presents a brief summary related to imputation based on clustering methods and focused on Fuzzy C-Means and Particle Swarm Optimization. From last decade, Hathaway and Bezdek [11] proposed that incomplete missing data can be obtained by calculate the distance between the missing and complete dataset. This estimation are then be proposed by other researcher in imputation method. Li and Gu [12] used FCM with Nearest Neighbor (NN) intervals as the representation of missing data apart from numerical representation. Nishanth et al [14] used K-Mean algorithm and multilayer perceptron (MLP) in solving missing data using imputation method. Authors claimed that the proposed method yield better results in accuracy of data mining algorithm due to the imputed dataset. In addition, Aydilek and Arslan [15] used FCM together with support vector regression (SVM) and

genetic algorithm (GA) to find the impute value. Author optimizes the FCM parameters with GA which is the cluster size and weighting factor.

Tab. 1 A Summary of Existing work

Authors	Imputation Methods	Findings
Li and Gu [12]	FCM & NN	Optimum the range for NN using FCM
Di Nuovo [13]	FCM	Applied in psychological scenario with FCM
Aydilek and Arslan [15]	FCM & SVM, GA	Optimizes FCM parameters with GA
Krishna *et al* [16]	PSO & Covariance Matrix	Covariance matrix for imputation and PSO will optimize the value with lowest RMSE
Gautam *et al* [17]	PSO & ECM, AAELM	PSO will be choosing optimum Dthr value forECM imputation.
Tang *et al* [18]	FCM & GA	GA has been used to optimize FCM parameters; membership function and centroids.
Rahman and Islam [19]	FCM & EM	Optimizes FCM with EM concept for better imputation
Proposed Approach	FCM & PSO	Optimizes FCM with PSO to find the optimum imputation value.

Krishna et al [16] use Particle swarm optimization with covariance matrix for imputation. Author want to preserve the covariance matrix of missing data and used PSO to minimize the mean squared error (MSE) and absolute difference. Furthermore, Tang et al [18] applied FCM with GA for missing traffic volume data. Author used GA to optimize the membership function and the centroids value in FCM model. With the help of GA, optimize FCM can give good imputation results for the problem. Rahman and Islam [19] proposed that missing value been imputed using fuzzy clustering based on Expectation Maximation (EM) approach. These approach been applied towards numerical and categorical missing values and they optimizes the FCM imputation based on the EM concept and yield a good results especially in time series tailoring problems.

Sherin [24] propose evolutionary-based Fuzzy Particle Swarm data Clustering (EFPSC) dynamically adapt to the changes and does not require a prior knowledge of the number of clusters in the datasets. The result showed that an important strength of the algorithm to automatically determine the number of cluster within the data. Lou. et al [25] introduced Selection in Incomplete Data (SID) based on uncertainty margin-based objective function to select relevant feature in its own relevant subspace. The author select weight for each feature was obtained ad using an EM algorithm to

perform feature selection directly from incomplete datasets. We summarize our work as in Table.1. It shows that FCM have an ability to find the imputation value with great results. But, FCM just impute the missing data once in one time.

3. The Proposed Method

Given an incomplete datasets, our approach replaces the missing value to optimum value using FCM and PSO and the best combination of its values to generate good clustering result in the solution space. The following section will be discussed about the concept of FCM and PSO approach before to the proposed method.

3.1. *Fuzzy c-means clustering (FCM)*

FCM algorithm was developed by Dunn in 1972 and improved by Bezdek in 1981. Traditional clustering algorithm are always partitioned the data into crisp division which membership function is either 0 or 1. Fuzzy clustering algorithm assign data object partially into multiple clusters. The degree of membership in the fuzzy clusters depends on the closeness of the instances to the cluster centre. However, FCM cannot handle missing value directly, thus, many researcher have done improvement towards this problem. With the ability of FCM to separate the data towards different group with certain value, thus the missing value can be obtain by calculate the distance from complete dataset and used it.

FCM partitions set of n dataset $x=\{x_1, x_2, \ldots\ldots, x_n\}$ in R^d dimensional space into fuzzy cluster c, $1<c<n$ with $r=\{r_1, r_2, \ldots\ldots, r_c\}$ cluster centers or centroids. Fuzzy clustering dataset is described in by fuzzy matrix μ with n rows and c columns which n is number of dataset and c is the number of clusters. While, μ_{ij} is the element in th i^{th} row and j^{th} column in μ, shows the membership function of the i^{th} dataset with the th j cluster. While μ was defines as follows,

$$\mu_{ij} \in [0,1] \ , \ \sum_{j=1}^{c} \mu_{ij} = 1, \qquad \forall i = 1,2,\ldots,n; \ \forall i = 1,2,\ldots,c \tag{1}$$

The objective function of FCM algorithm to minimize iteratively,

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} \left\| x_i - r_j \right\| \tag{2}$$

In which m(m> 1) is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and $\|x_i - r_j\|_2^2$ stands for the Euclidean distance from dataset, x_i to the cluster center, j r . The centroid j r of the th j cluster is obtained using,

$$r_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \quad (3)$$

3.2. *Particle swarm optimization (PSO)*

Particle swarm optimization is an extremely simple algorithm seems to be effective for optimize a wide range of functions; it has obviously ties with evolutionary computation. Particle Swarm Optimization introduced by Kennedy and Eberhart in 1995 based on the natural behavior of bird flocking or fish schooling to find the food. The flock of bird fly in a group follows the member that has closest distance to destination. In traditional PSO, population is called the swarm and the candidate of solutions in swarm is called particles while the food is called objective function.

The D-dimensional position for the particle i at iteration t can be represented as follows $x_1^t = \{x_{i1}^t, x_{i2}^t, \ldots, x_{id}^t\}$ Each elements of particle contains parameter; own position, own velocity, and own historical information. Each particle is given random position in search space and random velocity for the particles to fly within the search space.

Let $p_i^t = \{p_{x1}^t, p_{i2}^t, \ldots, x_{id}^t\}$ represent the best solution that particle i has obtained until iteration t, and $p_8^t = \{p_{8^1}^t, p_{8^2}^t, \ldots, p_{8^D}^t\}$ denote the best solution obtained from p_i^t in the population at iteration t. To search for the optimal solution, at each time step, each changes its velocity according to the pbest and gbest parts according to equation (4) and (5), respectively:

$$v_{id}^t = v_{id}^{t-1} + c_1 r_1 (P_{id}^t - x_{id}^t) + c_2 r_2 (P_{gd}^t - x_{id}^t), d = 1,2,\ldots,D \quad (4)$$

$$X_{id}^{t+1} = X_{id}^t + V_{id}^t, d = 1,2,\ldots,D \quad (5)$$

Where c_1 indicates the cognition learning rates for individual ability, c_2 indicates the social learning factor and r^1, r^2 are random numbers uniformly distributed in the interval 0 and 1.

3.3. *Our approach (FCMPSO)*

Fuzzy c-means clustering is an effective algorithm but the iterative process falling into local optimal. We proposed imputation approach based on FCM PSO to escape from the drawback and overcome the shortcoming of slow convergence speed. The main operation is to group the data in the similar

features with FCM, thus get the centroid values of each attribute and then to calculate the distance between each missing data with each clusters. After that, the impute value will be optimized with PSO according to the information from missing dataset.

4. Results and Discussions

4.1. *Datasets & Result*

To test the performance of imputation method, two datasets from UCI Machine Learning Repository Dataset [20] were used as bench mark datasets for the experimental purpose. In this paper, we used complete dataset only to have control over the missing data in the dataset. The artificial missing ratio range 1% to 25% was inserted into dataset to analyse how the imputation method worked. The attribute is normalized using min-max normalization and accelerate the training phase [15,18]. Cleveland Heart Disease dataset was based on heart disease diagnosis in Cleveland Clinic Foundation by Dr. Robert Detrano [21]. This dataset contains 297 records and 13 attributes, while the class contains two classes. On second data sets, Breast cancer dataset was created based on Wisconsin Breast Cancer problem by Dr. William H. Wolberg [23]. This dataset contain 683 records and 9 attributes. This dataset contains 2 classes.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_1 - y_2)^2} \tag{6}$$

Table 2 shows the RMSE value of proposed method with the other 2 imputation methods. It is clearly shows that FCMPSO outperforms compared to other two imputation methods when the missing ratios are increased from range 1% and 5%. The performance of proposed method will be evaluated with Root Mean Square Error (RMSE). In this paper, we compared the FCMPSO with 2 other ways of imputation which is Nearest-Neighbor Imputation and Fuzzy C-Means imputation.

Tab. 2 RMSE Results from Cleveland Heart Disease and Breast Cancer

Missing Percentages	Cleveland Heart Disease			Breast Cancer		
	NN	FCM	FCMPSO	NN	FCM	FCMPSO
1%	0.05432	0.05207	0.05231	0.44794	0.44795	0.44793
5%	0.0577	0.05500	**0.05360**	0.44790	0.4477	**0.44780**
10%	0.0631	0.06130	**0.06070**	0.44780	0.44730	0.44720

15%	0.0695	0.06620	**0.06540**	0.44790	0.44660	**0.44660**
20%	0.0765	0.07110	**0.06950**	0.44790	0.44660	**0.44650**
25%	0.0832	0.07510	**0.07350**	0.44790	0.44660	0.44640
30%	0.0873	0.07780	**0.07650**	0.44800	0.44620	**0.44580**

Consequently, from overall findings, it shows that FCMPSO gives a lower RMSE compared to other imputation methods over the increasing of missing ratios. It shows that, the missing dataset ratios is important and can give information for better imputations. Thus, further studies can be done towards the sensitivity of FCMPSO towards the dataset properties and condition.

5. Conclusion and Future Works

In this study, an imputation method using FCM and PSO are proposed which is aimed at optimizes estimating missing values in dataset using PSO and the effectiveness of proposed method is demonstrated on three different datasets which contains seven different missing ratios. We utilizes the PSO ability to find the best value for imputation from complete impute dataset by searching the best position of distance value and the result gives promising results as traditional clustering imputation. Due of that ability, further study to improve the clustering imputation with PSO in weight parameters of elements that belongs to each cluster for better imputation results and the sensitivity of dataset problems can be done.

Acknowledgments

This research was funded by Universiti Tun Hussein Onn Malaysia Graduate Incentive Scheme (VOT No U309) and Ministry of Higher Education Malaysia MyBrain15.

References

1. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery in databases. AI magazine, 1996. 17(3): p. 37.
2. Canlas Jr, R.D., Data Mining in Healthcare: Current Applications and Issues. [MS in Information Technology thesis], 2009.
3. Raju, P.S., D.V.R. Bai, and G.K. Chaitanya, Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International Journal of Innovative Research in Computer and Communication Engineering, 2014. 2(1).
4. Salleh, M.N.B.M., N.B. Mohd Nawi, and P. Boursier. Incomplete Information of Decision Support System for Planting Material Selection. in

INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on. 2009. 5. Romero, C. and S. Ventura, Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013. 3(1): p. 12-27.

5. Salleh, M.N.M. Implementing fuzzy modeling of decision support for crop planting management. in Fuzzy Theory and Its Applications (iFUZZY), 2013 International Conference on. 2013.
6. Zhang, S., W. Xindong, and Z. Manlong. Efficient missing data imputation for supervised learning. in Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on. 2010.
7. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, Supervised machine learning: A review of classification techniques. 2007.
8. Saleem, A., K.H. Asif, A. Ali, S.M. Awan, and M.A. Alghamdi. Pre-processing Methods of Data Mining. in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on. 2014.
9. Rubin, D.B., Inference and missing data. Biometrika, 1976. 63(3): p. 581-592.
10. Bezdek, J.C., R. Ehrlich, and W. Full, FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984. 10(2): p. 191-203.
11. Li, D., H. Gu, and L. Zhang, A fuzzy c-means clustering algorithm based on nearestneighbor intervals for incomplete data. Expert Systems with Applications, 2010. 37(10): p. 6942-6947.
12. Cohen, J., & Cohen, P. Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences. New York: John Wiley 1975.
13. Nishanth, K.J., V. Ravi, N. Ankaiah, and I. Bose, Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications, 2012. 39(12): p. 10583-10589.
14. Aydilek, I.B. and A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences, 2013. 233(0): p. 25-35.
15. Krishna, M. 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. IEEE.
16. Tang, J., 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, 2015. 51: p. 29-40.
17. Rahman, M.G. and M.Z. Islam, Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems, 2015:

p. 1-34.
18. Frank, A. and A. Asuncion, UCI machine learning repository. 2010.
19. Aha, D. and D. Kibler, Instance-based prediction of heart-disease presence with the Cleveland database. University of California, 1988.
20. Fisher, R.A., The use of multiple measurements in taxonomic problems. Annals of eugenics, 1936. 7(2): p. 179-188.
21. Street, W.N., W.H. Wolberg, and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. in IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. 1993. International Society for Optics and Photonics.
22. Sherin M.Y. A New hybrid Evolutionary-based Data Clustering Using Fuzzy Particle Swarm Optimization. in I2011 23rd International Conference on Tools with Artificial Intelligence ICTAI. 2011.
23. Luo Q, Obradovic, Z. Margin-based Feature Selection in Incomplete Data. In Proceedings of the Twenty-Sith AAAI Conference on Artificial Intelligence 2011.