FCMPSO: An Imputation for Missing Data Features in Heart Disease Classification

Mohd Najib Mohd Salleh, Nurul Ashikin Samat

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), 86400, Parit Raja, Batu Pahat, Johor, Malaysia.

Corresponding author: najib@uthm.edu.my, gi140022@siswa.uthm.edu.my

Abstract. The application of data mining and machine learning in directing clinical research into possible hidden knowledge is becoming greatly influential in medical areas. Heart Disease is a killer disease around the world, and early prevention through efficient methods can help to reduce the mortality number. Medical data may contain many uncertainties, as they are fuzzy and vague in nature. Nonetheless, imprecise features data such as no values and missing values can affect quality of classification results. Nevertheless, the other complete features are still capable to give information in certain features. Therefore, an imputation approach based on Fuzzy C-Means and Particle Swarm Optimization (FCMPSO) is developed in preprocessing stage to help fill in the missing values. Then, the complete dataset is trained in classification algorithm, Decision Tree. The experiment is trained with Heart Disease dataset and the performance is analysed using accuracy, precision, and ROC values. Results show that the performance of Decision Tree is increased after the application of FCMPSO for imputation.

1. Introduction

Heart is an important human organ which works as a pump to circulate the blood throughout the whole body that transporting oxygen, nutrients, and expel waste such as carbon dioxides. Disorder of heart to function well leads to heart disease. Heart Disease can also be known as cardiovascular disease is a group of disease connected to heart and blood vessel defect. According to World Health Organization (WHO), Heart Disease is the leading cause of death in the world over the past 10 years and the mortality from heart disease in Malaysia is 296 per 100 000 population in a year 2012. Thus, the expected value of deaths globally on year 2030 due to heart disease is 23.3 million. Early prevention and precaution can help to reduce the mortality of heart disease. Thus, motivated by the increasing number of heart disease patients’ data around the world, especially, researchers in computer science area have done various studies to help the experts to diagnose early detection of heart disease with the help of data mining tools [1].

Using data mining tools such as classification and clustering can work as the uncovering the knowledge of pattern and hidden information from analyze patients’ symptom, health condition data, and medical history [2, 3]. Knowledge and rules will be extracted out from the huge amount of data. In general, Decision Tree classifier grows to become popular and competent classification techniques among the researchers [4, 5]. The clear visualization of tree gives advantages to the user to know the most important class influence the result of the position of the attribute in the tree from top to the bottom. Yet, before any data mining models can run, it needs dataset to analyze. Collecting dataset is...
the first step in using any data mining tools. Problem arise when the data collected is not complete thus there will have problems arising in decision making. Real data especially medical data is hard to collect the accurate, precise, and complete medical data. Apart from that, incomplete dataset also will affect the accuracy of data mining models, might create bias result, and loss efficiency of computational process due to the holes in the dataset.

Thus, preprocessing method will play an important role in data mining task. Preprocessing is essential step for filtering and cleaning the dataset before it can be train in the data mining stage [6]. Cleaning the dataset from missing data by using imputation method is one of the reliable ways to solve the problem. So, this paper emphasis that imputation method in preprocessing stage will improve the ability and accuracy the data mining machine learning, Decision Tree. To achieve this goal, imputation method based on Fuzzy C-Means and Particle Swarm Optimization, FCMPSO has been applied towards missing dataset to get the complete dataset. Then, complete dataset will be trained using Decision Tree to examine the results.

This paper is organized as follows. A brief overview of missing data and imputation method will be given in Section 2. In Section 3, brief explanation of FCM and PSO. The methodology of proposed method, FCMPSO imputation and Decision Tree will be described in Section 4. The experiments, results, and evaluation are presented in Section 5. Section 6 contains the conclusion of this study.

2. Preliminaries Work

In this section, the literature review regarding missing data and imputation treatments towards missing data in preprocessing stage has been discussed. The type of missing data and the type of imputation methods that used to substitute the missing data also has been discussed.

2.1. Missing Data

In the real world, missing data are unfavorable by the researcher and expert because it may lead to errors and confusion in interpreting the data. But, to just ignore the missing data also been disadvantages as it may contain other important information. Thus, it has been shows significant research interest in recent years. Preprocessing stage is a way to clean and filter the target data because the data collection not always complete and perfect.

According to Rubin [7] missing data can be categorized into three which is

1. Missing Completely at Random (MCAR). The missing value has no relationship or dependency towards other data set or variable.
2. Missing at Random (MAR). The missing value that depended on other variables but the missing value can be obtained by estimated other complete variables.
3. Not Missing at Random (NMAR). The missing value that depended on other missing value, therefore the missing value cannot be estimated from existing data.

There are several ways to treat the incomplete dataset such as (1) Delete and ignore the missing data, (2) Imputation [7]. In first treatment, it means that, no imputation values will be substitute to the missing data. It also called as filter-based treatment. Although it is simplest way by deleting the records that contain missing values. Subsequently, the number of records will be less and the information cannot be fully utilized. In this paper, no imputation-classification result was compared with proposed method. Thus, in this paper preprocessing process focused on handling the missing value in the dataset by using imputation method. Imputation method is where the missing values being imputed or substituted with estimated value.

2.2. Imputation Method

Imputation has capability to manipulate and maximize the available information to fill up the missing value with the most plausible value. Mean imputation is the most basic technique used by past researchers [8]. Author replaced the missing values with the mean value for the attribute. Meanwhile, $k$-NN imputation method is a common method to impute the missing data with an actual range of
datasets [9]. However, to get a better range of dataset values to substitute, grouping or clustering the data with same similarity features or data will increase the accuracy of imputation values. Selecting the right imputation is significant as different methods will give different performance.

Thus, in this study, implementation of imputation based on feature cluster by Fuzzy C-Means (FCM) clustering was preferred. Through this method, useful and complete feature can be grouping into same subsets based on their similarities. Thus, the imputation values will be more accurate and relevant compared to just fill in randomly or using mean and mode values. The study by [10] examined that by using FCM, imputation values are more accurate. Meanwhile, [11] applied FCM in psychological scenario because author stated that, exploring and discover relation between the data is significant as they can cover valuable information. Apart from that, [12] used FCM with Nearest Neighbor (NN) intervals as the representation of missing data apart from numerical representation. They find the range of NN instead of using basic NN. It shows that minimum the range can increase the rate of accurate of imputation.

The main challenge is to find the best imputation values, which can help Decision Tree classify the Heart Disease better. Although the clustering can find the subsets of same features, there is more than one candidate to fill in to the missing spots. Thus, optimization is needed to find the best plausible data to replace the missing data.

3. Fuzzy C-Means and Particle Swarm Optimization

3.1. Fuzzy C-Means (FCM)

FCM is a clustering algorithm which group the dataset based on the similarity measures [13]. It is repetitive steps until the condition is fulfilled. The purpose to used FCM instead of other clustering algorithm is because of the ability to partition or grouping the ambiguity data by having the membership function values. FCM partitions set of n dataset \( x = \{x_1, x_2, \ldots, x_n\} \) in \( R^d \) dimensional space into fuzzy cluster \( c, 1 < c < n \) with \( r = \{n_1, n_2, \ldots, n_c\} \) cluster centers or centroids. The fuzzy clustering dataset is described in by fuzzy matrix \( \mu \) with n rows and c columns which n is number of dataset and c is the number of clusters. While, \( \mu_{ij} \) is the element in \( i^{th} \) row and \( j^{th} \) column in \( \mu \), shows the membership function of the \( i^{th} \) dataset with the \( j^{th} \) cluster. Therefore, authors exploit cluster analysis to complete a database with an estimate of the missing values [11, 12, 14].

3.2. Particle Swarm Optimization (PSO)

PSO is swarm metatheuristic algorithm that can optimize the solutions of problems. Particle Swarm Optimization developed and introduced by Kennedy and Eberhart in 1995 based on the natural behavior of bird flocking or fish schooling to find the food [15]. 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. Apart from that, due to the simplicity framework of PSO, the algorithms can find the optimization solution directly within acceptable computation time [16]. The D-dimensional position for the particle \( i \) at iteration \( t \) can be represented as follows \( x_i^t = \{x_{i1}, x_{i2}, \ldots, x_{iD}\} \). 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 \( p_i^t = \{p_{i1}^t, p_{i2}^t, \ldots, p_{iD}^t\} \) represent the best solution that particle \( i \) has obtained until iteration \( t \), and \( p_g^t = \{p_{g1}^t, p_{g2}^t, \ldots, p_{gD}^t\} \) 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 (1) and (2), respectively:
\[ v_{id}^t = v_{id}^{t-1} + c_1 r_1 (P_{id}^t - x_{id}^t) + c_2 r_2 (P_{g_{id}}^t - x_{id}^t), d = 1,2,\ldots, D \] (1)

\[ X_{id}^{t+1} = X_{id}^t + v_{id}^t, d = 1,2,\ldots, D \] (2)

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.

Thus, an alternative imputation method named FCMPSO which combined the benefit of Fuzzy C-Means algorithm, known to have good capabilities for their ability to cluster the data [17] and Particle Swarm Optimization algorithm, renowned to have a worthy ability and simple framework in optimization [18] is implemented for training the imputation method in preprocessing stage to overcome the missing data handicaps.

4. Proposed Method, FCMPSO

Figure 1 shows the flowchart of proposed method to impute the missing data and subsequently classified the data using Decision Tree. It consists of three main functional modules: 1) imputation in preprocessing stage; 2) application of Decision Tree; and 3) evaluations.

![Figure 1. Structure of Proposed Method.](image)

The main process is to group the data in the similar features with FCM to get the centroid values of each feature and to calculate the distance between each missing data with each cluster. After that, the impute value will be optimized with PSO according to the information from missing dataset. The method of the proposed method is as follows:

Step 1: Normalize the dataset using min-max normalization as Equation 3 where \( x_{\text{min}} \) is the minimum values and \( x_{\text{max}} \) is the maximum values for the features respectively. Separate the Complete and Incomplete data.

\[ x_{\text{normalize}} = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \] (3)

Step 2: For all Complete dataset,

(2.1) Calculates the cluster center using Equation 4

\[ r_j = \frac{\sum_{i=1}^{n} \mu_{ij}^m x_i}{\sum_{i=1}^{n} \mu_{ij}^m} \] (4)

(2.2) Computes the Euclidean distance, \( \| x_i - r_j \|_2^2 \)
(2.3) Updates the membership function using Equation 5 and 6
\[
\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{jk}} \right)^{m-1}}
\]
\[
J_m = \sum_{j=1}^{n} \sum_{i=1}^{n} \left( \mu_{ij} \right)^{2} \| x_i - r_j \|^{2}
\]

(2.4) If condition is not met, repeat Step 2

Step 3: For all Incomplete dataset
Calculate the imputation value according Equation 7 where
\[
x_{i\text{Miss}} = \sum_{k=1}^{k} U(x_{i\text{Com},C_k}) \times (C_k)
\]
End for;

Step 4: For each Imputed dataset,
(4.1) Initialize the population size, N = 50; Maximum Iterations, Iters = 100; C1 and C2 = 2; and create a swarm with P particles
(4.2) Calculates the fitness each particle using Equation (6);
(4.3) Calculates the pbest for the particles and gbest for the swarm;
(4.4) Update velocity for using Equation (1) and update the position of matrix using Equation (2);
(4.5) Terminating if condition met otherwise repeat step 4.

Step 5: The optimum value for Imputed dataset is obtained.

4.1. Decision Tree
After the incomplete dataset, has been imputed using proposed technique, FCMPSO, the complete impute dataset will be trained in Decision Tree algorithm in Waikato Environment for Knowledge Analysis (WEKA) Version 3.6.11. This study used C4.5 algorithm by Quinlan and WEKA J48 was a suitable algorithm as it generates a pruned or unpruned C4.5 Decision Tree [19]. The data will be partitioned into 60% for training and 40% for testing phase. Training set will build the classification algorithm model and the model will be tested with remaining percentage for testing set.

5. Experiments, Results, and Evaluations
In this section, some experiments were conducted to compare the performance of Decision Tree on different missing values imputation methods. Thus, classification techniques, Decision Tree accuracy, precision, and ROC values were analysed on proposed imputation, FCMPSO against No Imputation, Mean Imputation, k-NN imputation and FCM Imputation. The other imputation methods were implemented in preprocessing stage.

5.1. Heart Disease Dataset
The Framingham Heart Study is supported by the National Heart Lung and Blood Institute (NHLBI), a part of the National Institutes of Health (NIH). NIH has been committed to classify the collective causes that contribute to Heart Disease. Framingham Heart dataset was got from 5209 men and women between the ages of 32 and 70 from the town of Framingham, Massachusetts. In these experiments, there are no artificial missing ratios were inserted into the following dataset. Table 1 shows the summary of Heart Disease dataset.
Table 1. Summary of Heart Disease dataset.

| Dataset Name       | Framingham Heart Disease Dataset |
|--------------------|----------------------------------|
| Class              | Two (1-Absent; 2-Presence)       |
| Instances          | 4012 (Absent-2945; Presence-1067) |
| Attribute          | 10                               |
| Age                | 32-70                            |
| Gender (0-Male; 1-Female) | 0-1                       |
| Blood Pressure (Systolic) | 83.5-295         |
| Blood Pressure (Diastolic) | 48-142.5               |
| Heart Rate         | 44-143                         |
| Blood Sugar        | 40-394                         |
| BMI                | 15.54-56.8                     |
| Cholesterol        | 107-696                        |
| Smoking (0-No; 1-Yes) | 0-1                     |
| Chest Pain (0-No; 1-Yes) | 0-1                   |

5.2. Results
The result of proposed method is compared with No Imputation, Mean Imputation, $k$-NN Imputation, and FCM Imputation. The performances of the five methods are compared with respect to their corresponding Accuracy, Precision, and ROC values.

![Figure 2. Accuracy of Decision Tree.](image)

The results from Figure 2 shows the accuracy value of proposed method is higher with 86.3% compared to other imputations. Thus, proves that FCMPSO helps the Decision Tree to classify well with better imputation values for missing data. It also shows that missing data can lead to error and confusion in interpreting the data by ignores it. It is disadvantages as the data may contain other important information in the dataset. Apart from that, although Mean and $k$-NN Imputation easy to use, the accuracy is lower due to higher number of missing values in the dataset. Hence, it will affect the correlation between the features during classification process. Although, FCM shows high accuracy, but with the help of optimization in FCMPSO, it gives more accurate results. FCMPSO
help to improve the imputation values and help Decision Tree get better and accurate classification. It shows that with the imputation approach to group the features and optimize the features values can help Heart Disease classification with missing data problems.

![Figure 3. Precision of Decision Tree.](image)

The results from Figure 3 show the precision comparison between No Imputation, Mean Imputation, k-NN Imputation, FCM, and FCMPSO respectively. From the figure clearly shows that FCMPSO gives highest precision value compared to other imputation which means that, it offers better accuracy and more precise reading towards Heart Disease problems.

ROC is a measure tools to evaluate the efficiency of classifiers through the determination of the rates of true positives (elements correctly classified as positive class) and false positives (elements incorrectly classified as positive class). From Table 2 it shows that, ROC curve are more to value 1. It shows that, the ability of the ability of the test to correctly classify those with and without the Heart Disease.

|       | No Imputation | Mean | k-NN | FCM | FCMPSO |
|-------|---------------|------|------|-----|--------|
| ROC Values | 0.671         | 0.690| 0.710| 0.750| 0.830   |

6. Conclusions
The presence of missing values in database is a common fact and can generate difficulties on the knowledge extractions. Thus, imputation in preprocessing stage will benefit the data mining algorithm works more accurate. By using the proposed imputation method, FCMPSO, it can help to get better imputation by substitute the most possible values to missing value. The usage of PSO which is an optimization algorithm allows FCM to develop the ability to find the best imputation values where the results have led to a significant improving in the classification. The accuracy of Decision Tree results clearly showed that the imputed dataset using FCMPSO has improved the classification algorithm analysis compared to no imputation dataset, Mean imputation dataset, k-NN imputation dataset and FCM dataset. To conclude, FCMPSO can be considered as a promising imputation method in preprocessing stage for future research.

Acknowledgement
This paper was funded by Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), UTHM.
7. References

[1] Shouman M, Turner T and Stocker R 2012 Using data mining techniques in heart disease diagnosis and treatment. In: Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on, pp 173-7

[2] Deb R and Liew A W-C 2016 Missing value imputation for the analysis of incomplete traffic accident data Information Sciences 339 274-89

[3] Ali A and Senan N 2016 A Review on Violence Video Classification Using Convolutional Neural Networks. In: International Conference on Soft Computing and Data Mining: Springer) pp 130-40

[4] Mohamed W N H W, Salleh M N M and Omar A H 2012 A comparative study of Reduced Error Pruning method in decision tree algorithms. In: Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on, pp 392-7

[5] Tomar D and Agarwal S 2013 A survey on Data Mining approaches for Healthcare International Journal of Bio-Science and Bio-Technology 5 241-66

[6] Dhevi A T S 2014 Imputing missing values using Inverse Distance Weighted Interpolation for time series data. In: 2014 Sixth International Conference on Advanced Computing (ICoAC), pp 255-9

[7] Rubin D B 1976 Inference and missing data Biometrika 63 581-92

[8] Ravi V and Krishna M 2014 A new online data imputation method based on general regression auto associative neural network Neurocomputing 138 106-13

[9] Rahman S A, Huang Y, Claassen J, Heintzman N and Kleinberg S 2015 Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data Journal of biomedical informatics 58 198-207

[10] Aydilek I B and Arslan A 2013 A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm Information Sciences 233 25-35

[11] Di Nuovo A G 2011 Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario Expert Systems with Applications 38 6793-7

[12] Li D, Gu H and Zhang L 2010 A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data Expert Systems with Applications 37 6942-7

[13] Bezdek J C, Ehrlich R and Full W 1984 FCM: The fuzzy c-means clustering algorithm Computers & Geosciences 10 191-203

[14] Hathaway R J and Bezdek J C 2001 Fuzzy c-means clustering of incomplete data Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 31 735-44

[15] Eberhart R C and Yuhui S 2001 Particle swarm optimization: developments, applications and resources. In: Evolutionary Computation, 2001. Proceedings of the 2001 Congress on, pp 81-6 vol. 1

[16] Rana S, Jasola S and Kumar R 2011 A review on particle swarm optimization algorithms and their applications to data clustering Artificial Intelligence Review 35 211-22

[17] Niu Q and Huang X 2011 An improved fuzzy C-means clustering algorithm based on PSO Journal of Software 6 873-9

[18] Rasip N M, Basari A, Ibrahim N K and Hussin B 2015 Advanced Computer and Communication Engineering Technology: Springer) pp 459-69

[19] Quinlan J R 1986 Induction of decision trees Machine learning 1 81-106