Improving the classification accuracy using hybrid techniques

Mamdouh Abdel Alim Saad Mowafy and
Walaa Mohamed Elaraby Mohamed Shallan
Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Ain Shams University, Cairo, Egypt

Received 21 October 2020
Revised 23 December 2020
Accepted 16 January 2021

Abstract

Purpose – Heart diseases have become one of the most causes of death among Egyptians. With 500 deaths per 100,000 occurring annually in Egypt, it has been noticed that medical data faces a high-dimensional problem that leads to a decrease in the classification accuracy of heart data. So the purpose of this study is to improve the classification accuracy of heart disease data for helping doctors efficiently diagnose heart disease by using a hybrid classification technique.

Design/methodology/approach – This paper used a new approach based on the integration between dimensionality reduction techniques as multiple correspondence analysis (MCA) and principal component analysis (PCA) with fuzzy c means (FCM) then with both of multilayer perceptron (MLP) and radial basis function networks (RBFN) which separate patients into different categories based on their diagnosis results in this paper, a comparative study of the performance performed including six structures such as MLP, RBFN, MLP via FCM–MCA, MLP via FCM–PCA, RBFN via FCM–MCA and RBFN via FCM–PCA to reach to the best classifier.

Findings – The results show that the MLP via FCM–MCA classifier structure has the highest ratio of classification accuracy and has the best performance superior to other methods; and that Smoking was the most factor causing heart disease.

Originality/value – This paper shows the importance of integrating statistical methods in increasing the classification accuracy of heart disease data.

Keywords Principal component analysis, Heart disease, Fuzzy c-means, Multilayer perceptron, Multiple correspondence analysis, Radial basis function networks

Paper type Research paper

1. Introduction

Early diagnosis of heart disease is a difficult issue from a medical point of view because there are many diseases that share symptoms of the disease. Also, the health-care industry

The authors gratefully thank the hospital for cardiac, chest and vascular diseases at Ain Shams University for facilitating access to the data used in this article.

Compliance with Ethical Standards: Conflict of interest on behalf of all authors, the corresponding author states that there is no conflict of interest.

The authors received no funding support for this research.

Review of Economics and Political Science
Vol. 6 No. 3, 2021
pp. 223-234
Emerald Publishing Limited
e-ISSN: 2631-3561
p-ISSN: 2356-9980
DOI 10.1108/REPS-10-2020-0161

has large amounts of complex data about patients, disease diagnosis, electronic patient records (Haq *et al.*, 2018), etc.

Factors affecting heart disease are characterized by their multiple levels known as multiplicity, so that it is not easy to discriminate between them, which leads to a decrease in the classification accuracy of heart disease data, so this paper evaluates integrating dimensionality reduction techniques, fuzzy c-means (FCM) and both multilayer perceptron (MLP) and radial basis function networks (RBFNs) for improving the classification accuracy of heart disease data. The proposed structure comprises three stages; at the first stage, the study applied the multiple correspondence analysis (MCA) and principal component analysis (PCA) (dimensionality reduction techniques) on the heart disease data set with the aim of arranging relationships between the variables and reducing them to a smaller number of dimensions that have the most variances. In the second stage, the study used the dimensions obtained from PCA and MCA as inputs for FCM to ease separation of clusters and increase the ability to classify observations more precisely and raised FCM classifier performance (Hwang *et al.*, 2010; Ziasabounchi and Askerzade, 2014), thus they interpret the relationships between the variables correctly. In the third stage, the dimensions obtained from the MCA and PCA (both separately) represent the input layer in MLP and RBFNs, whereas the clusters obtained from FCM via MCA and FCM via PCA classifier act as the output layer; this is because FCM is responsible for grouping data with different values of membership. Based on these membership values, the MLP backpropagation algorithm and RBFNs classify heart disease data into two groups (infected and uninfected), thereby reducing the training period of the neural networks and increasing the accuracy of classification.

Finally, all the methods used in this paper were compared after and before integrating, and it is found that FCM via MCA is the best primary classifier, and that MLP via FCM–MCA is the best final classifier in this study.

2. Literature review

Many studies focused on the diagnosis and classification of heart diseases data; these studies have applied different statistical methods to a specific problem and have achieved a high classification accuracy of 75% or higher, and some studies have gone into integrating between classification methods and cluster methods to improve classification accuracy. Some examples of such studies are as follows.

Bhatla and Jyoti (2012) explain the results of applying neural networks, decision trees, fuzzy logic and genetic algorithm that have been closely associated with heart disease diagnosis. In recent years, the result of classification accuracy shows that the neural network was the best.

Dalvi *et al.* (2016) applied an integrating among neural networks and dimensionality reduction technique on the electrocardiogram (ECG) database of the Massachusetts Institute of Technology arrhythmia, and the achieved classification accuracy was 96.97%. The results obtained confirmed that using PCA reduced classification complexity without major changes within the performance.

Deng (2020) proposed generating new insight into an improvement on general clustering algorithms through this inspection of one specific clustering algorithm (FCM) help. This paper clarifies that there are three common problems of clustering algorithms, one of them is the noise problem and then it explained that the solution to this problem lies in combining adversarial learning with the FCM algorithm.

The study by El-Bialy *et al.* (2015) applied integration of the outcomes of the machine learning analysis applied to coronary artery heart disease data sets to compare the classification accuracy. The results clarified that the accuracy of classification of the collected data set is 78.06% higher than that of all separate data sets.

Haq *et al.* (2018) this study used machine learning algorithms as a hybrid system, to diagnose heart diseases, such as logistic regression, k-nearest neighbors algorithm, artificial neural network (ANN), support vector machine (SVM), Naive Bayes Algorithm, decision tree algorithm, and the random forest used with three feature selection algorithms Relief, minimum redundancy maximum relevance, and least absolute shrinkage and selection operator to select the important features, the result gives that the logistic regression performance with Relief; is the best predictive system which gives 89% of the classification accuracy.

In this paper, Jabbar *et al.* (2013) had applied integrating associative classifications algorithms and genetic approach for heart disease prediction; this integration gives high classification accuracy and best heart disease prediction compared to integration of Naive Bayes and neural networks methods

Kumar *et al.* (2018) aimed to build a classification model for patients of heart diseases; they used four classification methods, such as Naive Bayes, MLP, Random Forest and Decision Table, to classify whether a patient is tested positive or negative for heart diseases. The results illustrated that the Naive Bayes has the highest percent of classification accuracy (87.20%) for diagnosing heart patients.

Kumari and Godara (2011) made a comparative study of four classification techniques; they are the Ripper Algorithm, Decision Tree, ANN and SVM in data mining to predict cardiovascular disease applied to coronary heart illness data. The results show that the SVM predicts that cardiovascular disease has the least error rate with classification accuracy equals to 84.12%.

Kurt *et al.* (2008) made a comparative study using the receiver operating characteristic curve, hierarchical cluster analysis and multidimensional scaling between MLP, logistic regression, a classification and regression tree, radial basis function (RBF) and self-organizing feature map of the performances of classification to predict coronary artery disease (CAD) presence. MLP gives classification accuracy of 75.3% and it was the best technique to predict the presence of CAD in this data set.

Le (2019) applied a fuzzy c-means clustering interval Type-2 cerebellar model articulation neural network (FCM-IT2CMANN) method to help physicians improve the accuracy of diagnostic for breast cancer and liver disease. The proposed method combines two classifiers, where the IT2CMANN is the primary classifier and the FCM algorithm is the preclassifier; the results illustrated that the proposed classifier better than other methods.

Patra and Pradhan (2008) aimed to integrate FCM, independent component analysis (ICA) and neural networks (NN) and then compared with the structure ICA–NN. The performance of proposed FCM–ICA–NN was faster and more accurate than that of ICA–NN. Patra *et al.* (2009) made a comparative study of the performance of four structures such as FCM–NN, PCA–NN, FCM–ICA–NN and FCM–PCA–NN to investigate the classification of ECG arrhythmias. They confirmed that the performance of FCM–PCA–NN structure was faster and better than other techniques.

Wiharto and Suryani (2020) aimed to make a comparison between FCM and clustering algorithms K-means for segmentation retinal blood vessels. The statistical test results of comparison between them based on area under the ROC curve values resulted in p-values <0.05 with a confidence level of 95%. They confirmed that retinal vascular segmentation with the FCM method is significantly better than k-means.

Ziasabounchi and Askerzade (2014) integrated fuzzy clustering and k-means with PCA to diagnose heart disease patients; they showed that classification based on k-means via PCA is the best with 87% of classification accuracy.

The above results show that the studies that used the integration between the classification techniques gave higher classification results for heart disease data or other data, so we will integrate the artificial neural networks, fuzzy cluster and dimensionality reduction techniques to help increase the classification accuracy of heart disease data of Egyptian patients. The study produced a comparison between the algorithms that integrated, to illustrate the Importance of integrating between earlier methods mentioned.

3. Proposed methodology

3.1 Multiple correspondence analysis

MCA is an analytical tool that shows how strong the relationships are between large groups of variables, and it is a method of dimensionality reduction techniques used to organize the multi-level data into reduced dimensions based on the percentages of explained variance that each variable interprets, that is, it increases the homogeneity between these variables (Hwang *et al.*, 2006). In addition, object scores used as a preliminary stage for other techniques (Hwang *et al.*, 2010). Initially, by applying MCA to the indicator matrix, we will obtain the scores for the rows and columns factors, and then these scales are to be measured again to get another table $J \times J$ called Burt Matrix ($B = X^T X$) to use for obtaining MCA. In addition, note that we choose the dimension that has an Eigenvalue greater than 1 and if it is less than 1, it should be rejected, and that in this study, six dimensions were obtained. The calculation is done using SPSS 16 and STATA 15 (Abdi and Valentin, 2007).

3.2 Principle components analysis

PCA considered one of the most popular methods of dimensionality reduction techniques, as it converts the original data into a new coordinate system (uncorrelated variables) or which called reduced dimensions space by combining the variables that related to each other in one factor for using it in another analysis, while the unnecessary variables that have no effect on the target variable are eliminated without losing much of the information (Ziasabounchi and Askerzade, 2014).

Steps of PCA are as follows:

- assuming a set of data is $X = \{x_1, x_2, \ldots, x_n\}$, X is converted to standard variables;
- calculate correlation matrix or the covariance matrix;
- calculate the eigenvectors and eigenvalues;
- computing the principal components then forming a feature vector; and
- reducing the dimensions of the data set (Bhateja *et al.*, 2018).

3.3 Fuzzy C-mean method

FCM method is one of the most widely used fuzzy clustering analysis techniques, especially in medical research; this technique is based on the idea of partial membership and fuzzy partitioning.

Assuming that $X = \{x_1, x_2, \ldots, x_n\}$ represents the set of data elements (inputs) and assuming that k is the number of clusters and is an integer such that $2 \leq k \leq n$, by a membership grade u_{ij}.

The membership grade quantifies the grade of membership of the element to the fuzzy set. The value 0 means that it is not a member of the fuzzy set; the value 1 means that it is fully a member of the fuzzy set. The values between 0 and 1 characterize fuzzy members, which belong to the fuzzy set only partially (Hunt, 2012).

FCM technique divides the data set X to c of the fuzzy clusters, the idea of FCM is to find the center of each cluster and reduce the objective function J_m, which takes the following form (Patra *et al.*, 2009):

$$J_m = \sum_{i=1}^{N}\sum_{j}^{C} U_{ij}{}^{m}\|_{xi} - c_i\|^2, 1 \le m < \infty \tag{1}$$

where d_{ij} is the Euclidean distance between the X_i data point and the center c_i, u_{ij} is the degree of membership of the data point X_i to the cluster center j which is in the range [0, 1], m is the weighting exponent or fuzziness exponent, x_i a set of data with a $d-$ dimensional $(n \times p)$, c_i cluster centers with $d-$ dimensional $(k \times p)$, $\|\ ^{*}\ \|$ indicates the measure of similarity between any measured data and the center. To reduce the objective function and reach the previous equation, with the update of membership u_{ij} and the cluster centers c_i, the following conditions must be met (Bezdek, 2013):

$$c_i = \frac{\sum_{k=1}^{n} u_{ij}{}^{m} X_i}{\sum_{k=1}^{n} u_{ij}{}^{m}} \text{ for all } i \tag{2}$$

$$u_{ij} = \frac{1}{\sum_{j=1}^{c}\left(\frac{|x_i - c_i|}{|x_i - c_j|}\right)^{2/(m-1)}} \text{ for all } k \tag{3}$$

The iteration will stop if $\| U(k+1) - U(k) \| < \varepsilon$

where ε is a termination tolerance between zero and one, and k is the iteration.

FCM algorithm consists of several stages as follows:

- initializing partition matrix U, which is expressed as $[u_{ik}]$ matrix U^0;
- at k-step: calculate the centers vectors $C(k) = [cj]$ *with* $U(k)$; and
- update $U^{(k)}$ *to* $U^{(k+1)}$.
- if $\| U(k+1) - U(k) \| < \varepsilon$. stop; otherwise return to Step 2 (Liu *et al.*, 2011).

3.4 Multilayer networks

MLP is one of the most important of the feed-forward neural network classification methods, which are trained by the backpropagation algorithm that relies on the training of a nonlinear and feed-forward neural network, and it is a generalization of the training method in the pattern of error reduction, as this algorithm aims to reduce the error value between the targeted outputs and the network output by adjusting weights, where the algorithm depends on the spread of errors from the back to the front to adjust the network weights and the implementation of the backpropagation on three stages.

Steps of the backpropagation algorithm (Chakraverty *et al.*, 2019) *are as follows:*

- small random values generated for the weights (*initial values*);
- display inputs and target outputs, and then prepare input vector values for (x(1), x (2),. . ..x(N)), which correspond to the target outputs (d(1), d(2),. . .., d(N)), where N is the number of training patterns;

- calculate the actual outputs using the activation function, also to calculate the output signals, which are $(y_1, y_2, \ldots, y_{N_M})$ by the following formula:

$$y_i = \varphi\left(\sum_{j=1}^{N_{M-1}} w_{ij}^{(M-1)} x_j^{(M-1)} + b_i^{(M-1)}\right), i = 1, \ldots, N_{M-1} \tag{4}$$

- modification of weights (w_{ij}) and biases (b_i), where the algorithm begins to work by modifying weights between the output layer and the hidden layers:

$$\Delta w_{ij}^{(l-1)}(n) = \mu x_j(n) \delta_i^{(l-1)}(n), \tag{5}$$

$$\Delta b_i^{(l-1)}(n) = \mu \delta_i^{(l-1)}(n), \tag{6}$$

Where:

$$\delta_i^{(l-1)}(n) = \begin{array}{c} \varphi'(NET^{(l-1)})\left[d_i - y_i(n)\right], 1 = M \\ \varphi'(NET^{(l-1)}) \sum_k w_{ki} \delta_k^{(l)}(n), 1 \le l \le M \end{array} \tag{7}$$

where $x_j(n)$ is the output of the node j in the cycle n and l is the layer and k is the number of outputs of the nodes in the neural network, and M represents the output layer and φ represents the activation function, and μ expresses the learning rate (*used to modify weights during the training process*) which increased the convergence faster but may also cause the network oscillation around the extreme values and may not get the desired benefit from the training, and to achieve a faster convergence with the minimum oscillation, the Momentum Term added to the basic formula to update weight, as it improves the efficiency and speed of the training process through continuous adjustment then the effect of the learning rate passed, and after completing the training process for the neural network, the weights of the multi-layer networks are frozen and ready for use during the testing phase (Patra *et al.*, 2009).

3.5 Radial basis functions network

RBFN networks represent an attractive alternative to other neural network models, while they are one of the usual function approximations, it has a successfully important role in medical diagnostics. Note that the idea of the RBFN networks derives from the theory of function approximation, where the Euclidean distance calculated from the point that evaluated to the center of each neuron, and RBFN applied to distance with the goal of calculating the weight (effect) of each neuron (Riahi-Madvar *et al.*, 2019).

Training steps in the radial base function are as follows:

Starting to generate random values for weights of the layer. In this step, each unit j of the hidden layer was calculated using the following equation:

$$y_i(x) = \sum_{j=1}^{k} w_{ij} \varphi_j\left(||x - c_j||\right), i = 1, \ldots m`j = 1, .k \tag{8}$$

where x is the number of input dimensions vector, $\varphi(.)$ is the base function (Gauss activation function) which is described by $x - c_j$, c_j is the central vector of hidden j neurons having the same number of dimensions with x, w_{ij} is the weight that connects the node jth of the hidden layer and the node ith of the output layer, m is the number of neurons nodes in the output layer and k is the number of nodes in the hidden layer.

Applying a sigmoid function to each node in each output layer using the following equation:

$$\Psi_k(x) = f\left[\sum_{j=1}^{L} W_{jk}\varphi_j(x)\right] \tag{9}$$

where L is the number of hidden nodes.

The weights that link the output and the hidden layers are updated using the following equations:

$$w_{jk}^{new} = w_{jk}^{old} + \Delta w_{jk} \tag{10}$$

$$\Delta w_{jk} = \eta(t_k(x) - \Psi_k(x))\varphi_j \tag{11}$$

where η is the learning rate which takes the value between (0.1), the actual output of the network expressed by $\Psi_k(x)$, while $t_k(x)$ expresses the required output of the target vector for each pair. The earlier steps repeated from Step 3 until a small and acceptable error rate reached in the event that the desired goal did not reach (Liu *et al.*, 2011).

3.6 Processing steps used in the study to reach the hybrid technique

Step 1: Get the input data.

Step 2: Calculate each of the PCA or MCA.

Step 3: Obtaining the reduced dimensions from the previous techniques and using them as inputs for the FCM analysis.

Step 4: Using the clusters obtained from the previous step as inputs to the analysis of both multilayer networks and RBFN.

Note that dimensions of PCA or MCA are the input layer at both multilayer networks and RBFN and clusters obtained from FCM analysis is the output layer.

4. Data source

The population of the study is the data obtained from the Hospital for cardiac, chest, and vascular diseases at Ain Shams University about heart disease patient's records from the year 2010 to 2020.

All the available observations in the population selected to consist of a sample size of 216 Observations to reduce sampling error as possible, with 17 attributes used as shown in Table 1.

5. Result and discussion

5.1 Multiple correspondence analysis and principal component analysis

Primarily, by applying MCA procedure, an explanation of 94.126% of the total variance (by SPSS 16) is obtained, and 81.58% (by STATA 15) based on six dimensions confirmed that the scree plot is as shown in Figure 1.

Age	20 to 35:1, >35 to 65:2, > 65:3
Gender	Male: 1, female: 0
Cholesterol	Yes: 1, No: 0
Hypertension	Yes: 1, No: 0
Family history	Yes: 1, no: 0
Obesity rate	>25 :1, 25:30: 2, >30: 3
Diabetes	Yes: 1, no: 0
Smoking	Yes: 1, no: 0
Place of residence	Village: 1, urban: 0
Marital status	Married:1, unmarried: 2, widower: 3, Divorcee: 4
Alcohol abuse	Yes: 1, no: 0
Sleep apnea	Yes: 1, no: 0
Number of family members	0 to2: 1, 3 to 5: 2, >5:3
Working hours	35 to 40: 1, >40 to 48: 2, >48: 3
Physical activity	Inactive: 1, sedentary: 2, moderately active: 3, very active: 4, extremely active: 5
Level of blood urea	Low blood urea: 0, high blood urea:1
Uric acid ratio	Normal: 0, high:1

Table 1.
Attributes description

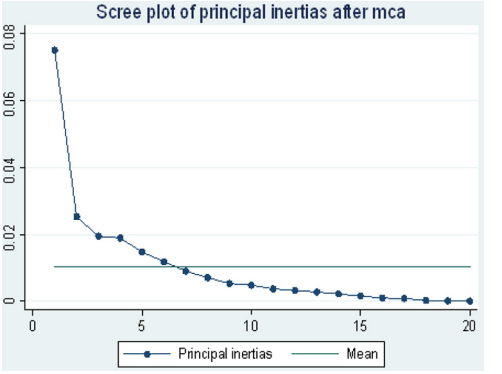

Figure 1.
Scree plot of eigenvalues obtained from MCA

By applying PCA procedure, it reduced the size of a heart disease patient data set into six principal components, and the eigenvalues of the six principals explained 80.99% and 78.27% of variance (NCSS11 and STATA), respectively, the eigenvalues of the all six dimensions are greater than 1, as shown in Figure 2.

5.2 Fuzzy c- means, fuzzy c-means via multiple correspondence analysis and fuzzy c-means via principal component analysis

There is more than one measure-to-evaluate goodness of fit of a fuzzy clustering solution. The first is the average silhouette per cluster, which has a range between $(-1, +1)$ and an average silhouette ≥ 0.71 meaning that a strong structure has been found. An average silhouette that has the range from 0.51 to 0.70 shows a reasonable structure; the value from 0.26 to 0.50 means that the structure is weak and try other methods on this database, and the value from 0.25 to -1 means no substantial structure. The second measure is the normalized Dunn partition coefficient $Fc(U)$, which is in the range from 0 (completely fuzzy) to 1 (hard clustering) and by the normalized Kaufman coefficient $Dc(U)$ that ranges from 0 (hard

clustering) to 1–(1/K) (completely fuzzy), and K indicates the number of clusters. The number of clusters should choose so that Fc (U) is large, and Dc (U) is small, because of that the results of the FCM procedure were unaccepted, and Table 2 clarifies that FCM is merged with both MCA and PCA and running the analysis based on their dimensions improves the performance of the FCM clustering method.

According to the results presented in Table 2, it is concluded that FCM via MCA provides better performance than FCM via PCA (*but they yield close results*) and thus FCM method, and it is obvious that FCM method performance has improved. SPSS 16 and NCSS 11 Statistical Software were used for calculation, the best results can be obtained if the analysis was carried out using two clusters.

The results were compared based on each value of the average silhouette that was 0.72, the normalized Dunn partition coefficient $Fc(U)$ that was 0.72, the normalized Kaufman coefficient Dc (U) and its value is 0.08, hence all results were the highest than each of FCM and FCM via PCA.

5.3 Classification by neural networks

Primarily, note that the input layer of MLP and RBFN is the number of dimensions extracted from MCA and PCA analyses; it was six dimensions, and then these six dimensions were used in FCM to get a new variable that separates the study observation into two clusters, where the new variable represents the output layer in MLP and RBFN.

Table 3 shows that classification accuracy increased in all cases of merging, whereas the results of MLP and RBF before merging have less performance which confirms the importance of integrating these methods, but it is noticed that the relative error percent in training and testing sample of MLP via FCM–MCA structure is least in comparison to MLP, RBF, MLP via FCM–PCA, RBF via FCM–MCA and RBF via FCM–PCA. Table 3 also shows

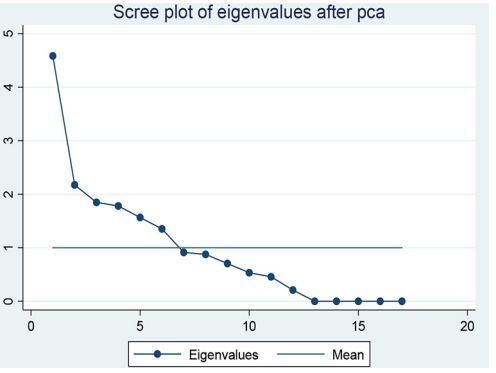

Figure 2.
Scree plot of eigenvalues obtained from PCA

Method	No. clusters	Average silhouette	The normalized Dunn partition coefficient Fc (U)	The partition coefficient Dc (U)
FCM	2	0.12	0.00	0.91
FCM via MCA	2	0.72	0.74	0.08
FCM via PCA	2	0.71	0.74	0.15

Table 2.
Performance results comparison

that the ratio of classification accuracy is higher in the case of MLP via FCM–MCA technique that is comparable with the techniques referred to earlier.

The training and test results also depicted for all structures in Table 2. It indicates that the training time for MLP via FCM–MCA is 00.016 s. On the other hand, MLP, RBF, MLP via FCM–PCA, RBF via FCM–MCA and RBF via FCM–PCA structures have the training time 00.039, 00.603, 00.065, 00.065, 00.155 and 00.170 s, respectively, which is much more than that of MLP via FCM–MCA structure. It is noticeable that MLP via FCM–MCA performance was superior to other methods.

It is noticeable that the performance of MLP via FCM–MCA was superior to other methods, so the normalized importance chart presented for it only.

The results were compared based on each value of percent of relative error in training with value 2.5%, percent of relative error in testing with value 5.1%, ratio of classification

Table 3.
Comparative of classification accuracy of heart disease data for different techniques

Methods	% of relative error in training	% of relative error in testing	Ratio of classification accuracy in sample training (%)	Ratio of classification accuracy in sample testing (%)	Training time in sec.
MLP	21.3	23.0	78.7	77.0	00:00:00.039
RBF	23.9	36.1	76.1	63.9	00:00:00.603
MLP VIA FCM AND MCA	1.9	3.4%	98.1	96.6	00:00:00.016
MLP VIA FCM AND PCA	2.5	5.1	97.5	94.9	00:00:00.065
RBF VIA FCM AND MCA	4.6	6.3	95.4	93.7	00:00:00.155
RBF VIA FCM AND PCA	5.9	7.9	94.1	92.1	00:00:00.170

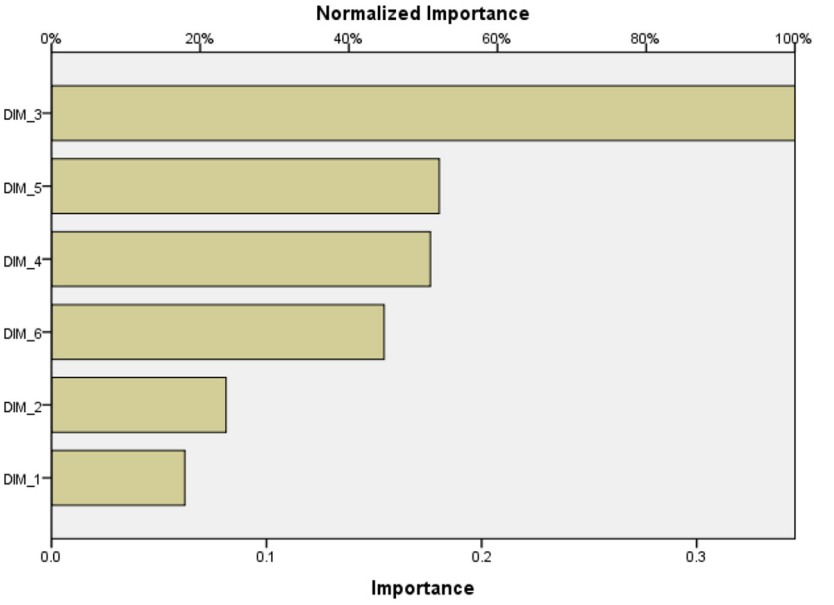

Figure 3.
Independent variable importance

accuracy in sample training with value 97.5%, ratio of classification accuracy in sample testing with value 93.7%, all results were superior to other methods.

5.4 Importance chart shows

Figure 3 emphasizes that the results dominated by the third dimension, which has the highest percent of normalized importance, included only smoking, followed by the fifth dimension, which included alcohol abuse, followed by the fourth dimension, which included both (age, marital status and sleep apnea), followed by the sixth dimension, which included both cholesterol and hypertension, followed by the second dimension, which included (gender, obesity rate, and physical activity) and finally, the first dimension included both (family history, diabetes, place of residence, number of family members, working hours, level of blood urea and uric acid ratio) and this dimension has the lowest percentage of normalized importance.

6. Conclusion

This proposed work is known as the hybrid technique, which uses both FCM–MCA and FCM–PCA as a preliminary stage of MLP and RBFN, the classifier FCM with PCA and MCA give better performance than FCM only but FCM–MCA gives higher performance, and MLP, RBF with FCM–MCA and FCM–PCA gives more accuracy than the classification techniques MLP and RBFN. As a way to validate the proposed system, it has been tested with a focus on those infected and uninfected with heart disease using six structures, which show that hybrid classifier structures improve the accuracy than traditional classifiers, and the comparison of performance between the six classifiers MLP, RBF, MLP via FCM–MCA, MLP via FCM–PCA, RBFN via FCM–MCA and RBFN via FCM–PCA shows that proposed hybrid classifier structure MLP via FCM–MCA performs faster with a high classification accuracy of heart disease data. Finally, the results showed that smoking is the most important variable that caused heart disease; hence, the obtained prediction model will help doctors to efficiently diagnose heart diseases.

In the future, results will be used to create a monitoring plan for heart patients because heart patients usually are not identified until a later stage of the disease or the event of complications and will be integrated between other methods to clarify the importance of integrating.

References

Abdi, H. and Valentin, D. (2007), "Multiple correspondence analysis", *Encyclopedia of Measurement and Statistics*, Vol. 2, pp. 651-666.

Bezdek, J.C. (2013), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer Science and Business Media.

Bhateja, V., Le Nguyen, B., Nguyen, N.G., Satapathy, S.C., and Le, D.-N. (2018), *Information Systems Design and Intelligent Applications: Proceedings of Fourth International Conference India 2017*, Springer.

Bhatla, N. and Jyoti, K. (2012), "An analysis of heart disease prediction using different data mining techniques", *International Journal of Engineering*, Vol. 1, pp. 1-4.

Chakraverty, S., Sahoo, D.M., and Mahato, N.R. (2019), *Concepts of Soft Computing: fuzzy and ANN with Programming*, Springer.

Dalvi, R.D.F., Zago, G.T. and Andreão, R.V. (2016), "Heartbeat classification system based on neural networks and dimensionality reduction", *Research on Biomedical Engineering*, Vol. 32 No. 4, pp. 318-326.

Deng, S. (2020), "Clustering with fuzzy C-means and common challenges", *Journal of Physics: Conference Series*, Vol. 1453, p. 012137

EL-Bialy, R., Salamay, M.A., Karam, O.H. and Khalifa, M.E. (2015), "Feature analysis of coronary artery heart disease data sets", *Procedia Computer Science*, Vol. 65, pp. 459-468.

Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R. (2018), "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms", *Mobile Information Systems*, Vol. 2018.

Hunt, V.D. (2012), *Artificial Intelligence and Expert Systems Sourcebook*, Springer Science and Business Media.

Hwang, H., Dillon, W.R. and Takane, Y. (2010), "Fuzzy cluster multiple correspondence analysis", *Behaviormetrika*, Vol. 37 No. 2, pp. 111-133.

Hwang, H., Montréal, H., Dillon, W.R. and Takane, Y. (2006), "An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents", *Psychometrika*, Vol. 71 No. 1, pp. 161-171.

Jabbar, M.A., Deekshatulu, B.L. and Chandra, P. (2013), "Heart disease prediction using lazy associative classification", *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, *IEEE*, pp. 40-46.

Kumari, M. and Godara, S. (2011), "Comparative study of data mining classification methods in cardiovascular disease prediction 1".

Kumar, M., Shambhu, S., and Sharma, A. (2018), *Classification of Heart Diseases Patients Using Data Mining Techniques*, IJRECE.

Kurt, I., Ture, M. and Kurum, A.T. (2008), "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease", *Expert Systems with Applications*, Vol. 34 No. 1, pp. 366-374.

Le, T.-L.J.I.A. (2019), "Fuzzy C-Means clustering interval type-2 cerebellar model articulation neural network for medical data classification", *IEEE Access*, Vol. 7, pp. 20967-20973.

Liu, D., Zhang, H., Polycarpou, M., Alippi, C. and He, H. (2011), *Advances in Neural Networks–ISNN 2011: 8th International Symposium on Neural Networks, ISNN 2011, Guilin, China, May 29–June 1, 2011, Prodceedings*, Springer Science and Business Media.

Patra, D. and Pradhan, S. (2008), "Integration of FCM, ICA and neural network for ECG signal classification", *Proc. of IEEE International Conference on Soft Computing (ICSC 2008), IET Alwar*, pp. 8-10.

Patra, D., DAS, M.K. and Pradhan, S. (2009), "Integration of FCM, PCA and neural networks for classification of ECG arrhythmias", *IAENG International Journal of Computer Science*, Vol. 36.

Riahi-Madvar, H., Dehghani, M., Seifi, A., Salwana, E., Shamshirband, S., Mosavi, A. and Chau, K.-W. (2019), "Comparative analysis of soft computing techniques RBF, MLP, and ANFIS with MLR and MNLR for predicting grade-control scour hole geometry", *Engineering Applications of Computational Fluid Mechanics*, Vol. 13 No. 1, pp. 529-550.

Wiharto, W. and Suryani, E. (2020), "The comparison of clustering algorithms k-means and fuzzy c-means for segmentation retinal blood vessels", *Acta Informatica Medica*, Vol. 28 No. 1, p. 42.

Ziasabounchi, N. and Askerzade, I.N. (2014), "A comparative study of heart disease prediction based on principal component analysis and clustering methods", *Turkish Journal of Mathematics and Computer Science (TJMCS)*, Vol. 16, p. 18.

Corresponding author
Walaa Mohamed Elaraby Mohamed Shallan can be contacted at: walaaelaraby@bus.asu.edu.eg