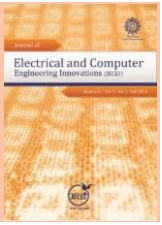
SRTTU

Journal of Electrical and Computer Engineering Innovations

JECEI, Vol. 4, No. 2, 2016

Regular Paper



Diagnosis of Heart Disease Based on Meta Heuristic Algorithms and Clustering Methods

Sadaf Roostaee[1,*] and Hamid Reza Ghaffary[1]

[1]Islamic Azad University, ferdows Branch, ferdows, Iran.

[*]Corresponding Author's Information: sadafroostaee@gmail.com

ARTICLE INFO

ARTICLE HISTORY:
Received 27 August 2016
Revised 11 October 2016
Accepted 30 October 2016

KEYWORDS:
Heart disease
Support vector machine
Binary cuckoo optimization
Algorithm
Features selection

ABSTRACT

Data analysis in cardiovascular diseases is difficult due to large massive of information. All of the features are not impressive in the final results. So, it is very important to identify more effective features. In this study, the method of feature selection with binary cuckoo optimization algorithm is implemented to reduce property. According to the results, the most appropriate classification for support vector machine is featured diagnoses heart disease. The main purpose of this article is feature reduction and providing a more precise diagnosis of the disease. The performance of the proposed method is evaluated using three measures: accuracy, sensitivity and specificity. For comparison, a data set of Machine Learning Repository database including information about 303 people with 14 features was used. In addition to the high accuracy of current methods, they are expensive and time-consuming. The results indicate that the proposed method is superior to the other algorithms in terms of performance, accuracy and run time.

1. INTRODUCTION

In most cases, the heart disease term is used instead of cardiovascular disease term. Cardiovascular diseases are the leading cause of death in most countries [1]. Based on the latest statistics by World Health Organization (WHO), over 37% of world deaths are caused by cardiovascular disease. At the beginning of the twentieth century, cardiovascular diseases cause more than 10% of all deaths throughout the world. The results of predictions show that by 2020, cardiovascular diseases cause more than 75% of deaths in the world [2].

In the recent years, using a combination of evolutionary algorithms and data mining techniques to improve diagnosis of the disease is very common. The methods can help doctors to use the large amounts of data much easier. Data mining techniques include collecting, processing and analyzing the data. Evolutionary algorithms are a class of optimization algorithms that are inspired by nature. In this paper, two binary cuckoo optimization algorithm and support vector machine classification are used for diagnosis of cardiovascular diseases.

By comparing the obtained results, it is indicated that binary cuckoo optimization algorithm has the highest accuracy among used algorithms including genetic algorithm, particle swarm optimization algorithm and cuckoo search algorithm. With developing models and using new parameters, we can reach more accurate and reliable results. By using the results of this model, it is possible to diagnosis the probability of heart disease with less error. The results could help medical specialists, hospitals and emergencies.

This paper is organized as follows. First an introduction about previous works in the field of heart disease is presented. Then, algorithms and heart disease databases used in this paper are described briefly. Afterward, the proposed method is introduced and finally the results of evaluation are presented and the conclusion is expressed.

2. Previous Works

In this section, an introduction about previous works in the field of heart disease is presented. In this regard, some studies on Heart Disease Dataset (HDDB) are evaluated.

In [3], a model for large data sets is proposed that can reduce features by using type2 fuzzy logic system (IT2FLS) and chaos algorithm. This model is a combination of fuzzy c-means clustering algorithm with the parameters tuned by genetic chaos with firefly algorithm and is compared with the other machine learning methods such as simple bayesian, support vector machine and neural network. The results indicate the superiority of this model over the other models.

In [4], a model based on decision tree is provided to reduce the size of data and produce valid rules. In this model, fuzzy expert system is used to classify data for heart disease. Adjusting the fuzzy membership functions is done by using Independent Component Analysis (ICA) and improved ICA.

In [5], genetic algorithm and neural network have been used for evaluating diagnosis of heart disease. The main purpose of this research was developing a prototype to identify and extract unknown knowledge related to heart disease database. The disadvantage of Back Propagation (BP) algorithm is getting stuck in local optimum. This problem is solved by optimization. The results indicate that neural network performs better in comparison with the other classification methods.

In [6], Genetic Algorithm-Multilayer Perceptron (GA-MLP) method is introduced for medical diagnosis. In this way, the improved genetic algorithm is used to determine the optimization parameters automatically and simultaneously through evolutionary process. The parameters include some hidden nodes, initial weights and feature subsets of multi-layer perception. Compared with previous works, the mentioned algorithm has the best accuracy among other algorithms for diabetes, heart disease and cancer.

3. The Proposed Method

3.1. The Use Database

In this study, University of California, Irvine (UCI) dataset is used to evaluate the proposed method [7]. This data includes test results of 303 people. Number of common features for the diagnosis of heart disease, according to World Health Organization standard, is 76. All of these features are not useful and the helpful features should be selected. This dataset contains two classes, one class for healthy people and the other class for people with heart disease. In table (1), the features of this data set are introduced.

TABLE 1
SPECIFICATIONS OF THE DATASET USED IN THIS PAPER

Feature	Description
Age	Age in years
Sex	1 = male 0 = female
Cp	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Resting blood pressure (in mm Hg)
Chol	Serum cholesterol in mg/dl
Fbs	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or defines left ventricular hypertrophy by Estes'criteria
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina: 1 = yes 0 = no
Oldpeak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Number of major vessels colored by fluoroscopy
Thal	Heart condition: 3: normal 6: fixed defect 7: reversible defect
Num	0: if less than 50% diameter narrowing in any major vessel (CAD, no) 1: if more than 50% (CAD, yes)

3.2. Support Vector Machine

Support vector machine (SVM) is one of the supervised learning methods that have been used for classification [8]. The main purpose of linear classification methods is separating data by building a hyper plan. SVM classification method, that is one of the linear classification methods, tries to find the best hyper plan with maximum margin for separating data to two classes. Training data are shown with pairs of

(x_i,y_i). X_i is the n-dimensional feature and y_i ϵ{-1,1}is it's label. The goal is to find a hyper plan that can separate two classes with labels of -1 and 1 with maximum margin. SVM is shown in Fig. 1.

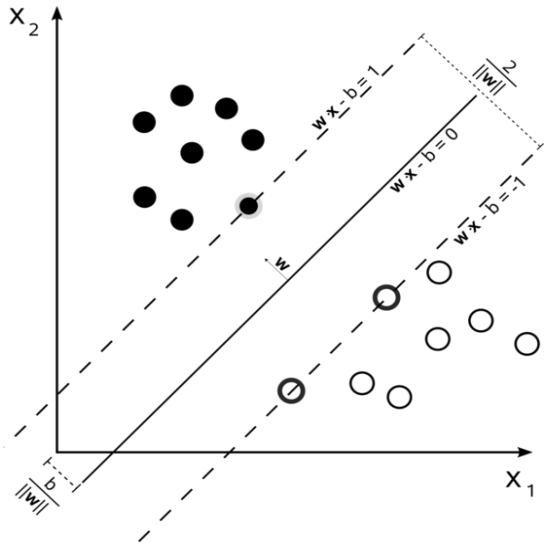


Figure 1: Linear optimal boundary when two classes are completely separated from each other.

b, is a y-intercept for a hyper plan with maximum separating boundary. Regardless of b, the hyper plans that have been passed from the source are the only answers. The hyper plan vertical distance to the source is obtained by dividing the absolute value of parameter b on the length of w. The main idea is selecting a suitable separator that has maximum distance from the neighbors of each class. This solution actually has the maximum distance from the points of two classes and can be bounded with two parallel hyper plans that pass from at least one class points. These vectors are called machine vectors. Mathematical formula for these two parallel hyper plans that have formed a separable boundary, are presented in equations (1) and (2):

$$w.x - b = 1 \quad (1)$$

$$w.x - b = -1 \quad (2)$$

It is noticeable that if training data are linearly separable, it is possible to choose two hyper plans such that no data are located between them. The distance between two parallel hyper plans must be at maximum value. By using geometrical theorems, this distance is equal to 2 /|w|, and thus |w| must be minimized. Also, it is necessary to avoid placing data points within boundary area. For this purpose, a mathematical constraint is added to the formal definition. For each i, applying the following constraints ensure that no point is on the boundary:

For data in the first class	$w.x - b > 1$
For data in the second class	$w.x - b \leq 1$

This constraint is shown as equation (3):

$$C_i(w.x - b) = 1 \qquad 1 \leq i \leq n \quad (3)$$

So optimization problem is described as minimization of *w*, with respect to the implicit constrain in equation (4):

$$C_i(w.x - b) \geq 1 \qquad 1 \leq i \leq n \quad (4)$$

3.3. Binary cuckoo optimization algorithm

In [9], binary cuckoo optimization algorithm called binary cuckoo optimization algorithm (BCOA) is proposed. For converting continuous Cuckoo optimization algorithm (COA) to binary version, COA migration operator is defined as follows. Let XGoal and XCurrenPostion be the current target point and the current position of a cuckoo in the population, respectively. The next position of cuckoo (XNextHabitat) is calculated according to equation (5):

$$X_{NextHabitat} = X_{CurrenPostion} + rand \times (X_{Goal} - X_{CurrenPostion}) \quad (5)$$

, If sigmoid function is used to mapping XNextHabitat into range of 0 and 1 as follow, new position is appropriate for binary space (6):

$$S = \frac{1}{1 + e - X_{NextHabitat}} \quad (6)$$

According to equation (7), the position values are maped to binary values 0 and 1 (rand in the following equation is a uniform random number).

$$If\ S < rand\ Then\ X_{NextHabitat} = 1$$
$$If\ S < rand\ Then\ X_{NextHabitat} = 0 \quad (7)$$

3.4. Feature Selection

The purpose of feature selection is selecting a subset of features that can improve diagnosis accuracy [10]. Feature selection is the process of selecting a subset of features from original set of features by which the quality of selected subset is equal or better than with compared to all features. Removing useless features, results in lower dimensions that increases computing speed and performance of diagnosis system. The advantages of feature selection include reducing cost, reducing complexity of problem and increasing accuracy of the model.

3.5. The Proposed Method

Modeling is performed by Matlab R2015a. First a feature subset is selected as solution by using one of evolutionary algorithms. In the end, final selected features are applied for classification. One evolutionary algorithm called BCOA is selected and SVM is used as the ultimate classifier.

Specifications of BCOA parameters are as follows: The number of initial cuckoos in problem space is 5. Training process is repeated 50 times. In creating diagnosis model, first the data set is partitioned into 5 parts for train and test, by using 5-fold cross validation. Fig. 2 describes the proposed model.

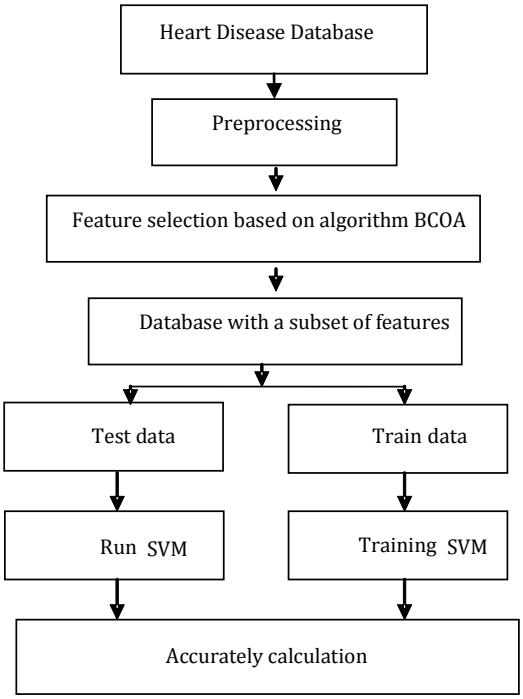


Figure 2: The proposed model struture.

4. Simulation Results

The heart disease dataset (HDDB) has been used for evaluation of algorithms. Models have been evaluated according to accuracy, sensitivity, and specificity measures. Each method has been evaluated by using 6-fold cross-validation algorithm. In each section, the results obtained by each method are presented.

4.1. Results Obtained from Neural Network, Evolutionary Algorithms and Fuzzy C-Means Clustering

Tables 2, 3 and 4, show the results of combining GA, Particle Swarm Optimization (PSO), and COA with neural network, fuzzy c-means clustering algorithm and Principal Component Analysis (PCA), respectively.

According to Table 2, accuracy of COA is higher than the other two methods. This prominence is also evident for the other two measures.

Table 2
The Results of the Combination of Artificial Neural Networks and Evolutionary Algorithms

Method	Accuracy	Sensitivity	Specificity
GA-ANN	81.11	77.74	80.66
PSO-ANN	77.41	70.27	81.64
COA-ANN	82.22	78.64	83.74

Table 3
The Results of the Combination of Artificial Neural Networks and Evolutionary Algorithms and FCM Algorithms

Method	Accuracy	Sensitivity	Specificity
GA-ANN+FCM	80.74	78.08	81.57
PSO-ANN+FCM	77.78	75.57	79.66
COA-ANN +FCM	79.26	80.42	78.00

Sample reduction has the most negative effect on the performance of COA which is equal to 2.96%. However, GA and PSO have been less affected. Again, the best accuracy is obtained for COA.

Table 4
The Results of the Combination of Artificial Neural Networks and Evolutionary Algorithms and Feature Selection

Method	Feature Number	Accuracy	Sensitivity	Specificity
GA-ANN+PCA	10	80.74	74.41	84.46
PSO-ANN+PCA	10	80.37	78.72	80.36
COA-ANN +PCA	10	82.96	80.37	83.59

According to Table 4, applying PCA improves the classification ability of evolutionary algorithms. The most positive effect has been occurred for PSO and then for COA. The average classification accuracy of PSO and COA have been improved by 2.96% and 0.74%, respectively. But, the average of classification accuracy has been decreased by 0.37% for GA.

4.2. The Results Obtained by Using SVM Along with Binary Cuckoo Optimization Algorithm

For improving the classification ability, BCOA is used for feature selection and SVM is used for constructing the model. The results obtained by the proposed method are presented in Table 5. In simulations, the number of cuckoos in the initial population has been set as 5.

TABLE 5
OBTAINED RESULTS OF THE PROPOSED MODEL BY USING BINARY CUCKOO ALGORITHM

Measure	Support Vector Machine
Accuracy	84.44
Sensitivity	86.49
Specificity	81.49

4.3. Evaluation of the results

Cuckoo optimization algorithm (COA), binary cuckoo optimization algorithm (BCOA), genetic algorithm (GA) and particle swarm optimization (PSO) algorithm have been used for adjusting the weights of neural network. Each algorithm has been used once with PCA and once with sample reduction methods. The proposed method has been constructed by SVM and BCOA as feature selection. The average accuracy of the proposed method and the other algorithms have been shown in Fig. 3.

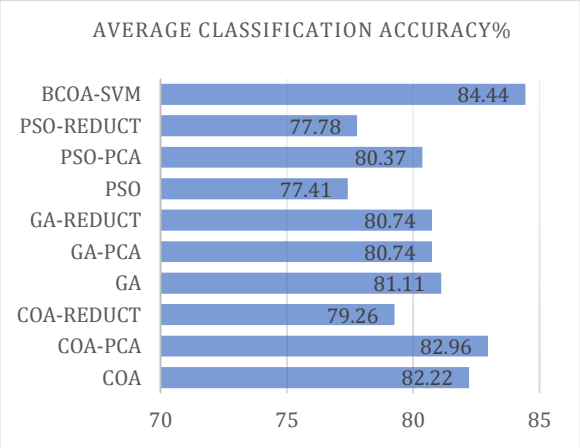


Figure 3: Average classification accuracy between the proposed method and trained neural network with evolutionary algorithms including: cuckoo, genetic, particle swarm optimization and combining them with PCA and sample reduction.

According to Fig. 3, the binary cuckoo optimization algorithm has been achieved to an average accuracy of 84.44%; while, the average accuracy of COA, GA, and PSO is 82.22%, 81.11%, and 77.41%, respectively. Therefore, the proposed model has the best accuracy comparing to the mentioned algorithms.

Besides the average of classification accuracy, the standard deviation of results is also very important. The less standard deviation means the method is more stable. In Fig. 4, the standard deviation of each method along with the average of accuracy is depicted.

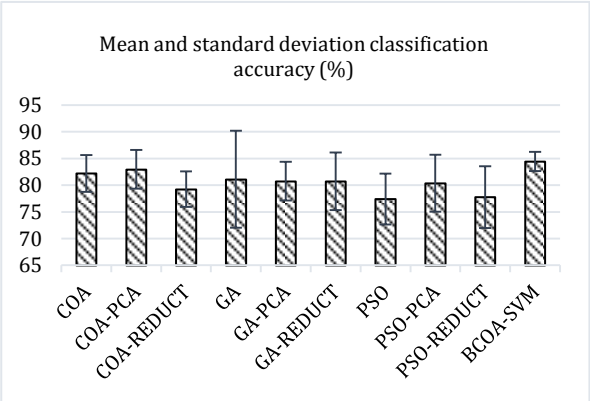


Figure 4: Mean and standard deviation of classification accuracy between the proposed method and the neural network trained with evolutionary algorithms including: cuckoo, genetic, particle swarm optimization and combining them with PCA and sample reduction.

TABLE 6
THE RESULTS OF COMPARING THE PROPOSED METHOD WITH PREVIOUS METHODS

Research Work	Method	Accuracy	Sensitivity	Specificity
[11]	GAANNFS	78.52	78.84	80.60
[4]	DTFISICA	78.89	78.42	78.47
Suggested Method	BCOA-SVM	84.44	86.49	81.49

According to Fig. 4, the least standard deviation occurs for different combinations of COA, while the most deviation is related to different combinations of GA. In other words, the COA is a more stable method comparing with GA and PSO. The proposed method has the least standard deviation equal to 1.

4.4. Comparing with previous research

In this section, the performance of the proposed method is compared with some reference methods on heart disease dataset and the results are presented in Table 6.

In Table 7, the results of sample reduction by fuzzy c-means algorithm are presented for each algorithm.

5. CONCLUSION

According to the results presented in this paper, BCOA-SVM has the best accuracy in diagnosing heart disease. In this paper, binary cuckoo optimization algorithm (BCOA) is used for feature selection and

SVM is used for constructing the model. The constructed model and the other hybrid algorithms have been applied on heart disease dataset with the same conditions. Accuracy, sensitivity, and specificity measures have been calculated for each algorithm.

TABLE 7
THE RESULTS OF COMPARING THE PROPOSED METHOD WITH PREVIOUS METHODS BY USING FEATURE SELECTION

Research Work	Method	Accuracy	Sensitivity	Specificity
[11]	GAANNFS-REDUCT	72.22	69.90	72.13
[4]	DTFISICA-REDUCT	77.78	76.07	78.47
Suggested Method	BCOA-SVM	84.44	86.49	81.49

In this research, the powerful classification method has been used for diagnosing heart disease. The final model has an accuracy equal to 84.44%, a sensitivity equal to 86.49%, and a specificity equal to 81.49%. A high percentage of patients could be diagnosed correctly with this model. The specificity measure is lower than sensitivity measure which is in agreement with the other similar research. This means that models are more powerful in diagnosing patients with compared to healthy people.

For feature research, we propose using the feature selection methods such as Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA). Also, new hybrid models such as firefly algorithm could be used for constructing the model. The proposed method can also be applied on the other medical datasets such as diabetes and different types of cancer.

REFERENCES

[1] MAYO CLINIC. 2014. *Diseases and Conditions Heart disease*. http://www.mayoclinic.org/diseases-conditions/heart disease/ basics/ definition/ con-20034056.

[2] World Health Organization. 2014. Reviewed June 2016. WHO cardiovascular disease. http://www.who.int/mediacentre/factsheets/fs317/en/.

[3] N. Cong Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Systems with Applications*, vol. 42, pp. 8221-8231, 2015.

[4] Z. Mahmoodabai and S. Shaerbaf Tabrizi, "A new ICA-based algorithm for diagnosis of coronary artery disease," *Intelligent Computing, Communication and Devices*, vol. 2, pp. 415-427, 2014.

[5] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," *Computing for Sustainable Global Development (INDIACom)*, pp. 704-706, 2015.

[6] F. Ahmad, N. Ashidi Mat Isa, Z. Hussain, and M. Khusairi Osman, "Intelligent medical disease diagnosis using improved hybrid genetic algorithm-multilayer perceptron network," *Journal of Medical Systems*, vol. 37, pp. 9934, 2013.

[7] https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart).

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*" vol. 20, pp. 273-297, 1995.

[9] S. Mahmoudi, R. Rajabioun, and S. Lotfi, "Binary cuckoo optimization algorithm," *National Conference on New Approaches in Computer Engineering and Information Retrieval*, Iran, Roudsar, 2013.

[10] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[11] A. Fadzil, M. Nor Ashidi, H. Zakaria, O. Muhammad Khusairi, and S. Siti Noraini, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," *Pattern Analysis and Applications*, vol. 14, pp. 861-870, 2014.

BIOGRAPHIES

Sadaf Roostaee was born in Mashhad, Iran, in 1990. She received the M.Sc. degree in computer engineering from Islamic Azad University, Ferdows, Iran, in 2016. Her research interests are revolutionary algorithms and data mining.

Hamid Reza Ghaffari was born in Ferdows, Iran, in 1974. He received the Ph.D. degree in computer engineering from Ferdowsi University, Mashhad, Iran, in 2014. His research interests are revolutionary algorithms and data mining.

How to cite this paper:

S. Roostaee and H. R. Ghaffary, "Diagnosis of heart disease based on meta heuristic algorithms and clustering methods," Journal of Electrical and Computer Engineering Innovations, vol. 4. no. 2, pp. 105-110, 2016.

DOI: 10.22061/jecei.2016.570

URL: http://jecei.srttu.edu/article_570.html