JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 8, No. 2, June 2014, pp. 87-93

CADICA: Diagnosis of Coronary Artery Disease Using the Imperialist Competitive Algorithm

Zahra Mahmoodabadi*
Computer Engineering Department, Imam Reza University, Mashad, Iran
za.mahmoodabadi@gmail.com

Mohammad Saniee Abadeh
Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
saniee@modares.ac.ir

Abstract
Coronary artery disease (CAD) is currently a prevalent disease from which many people suffer. Early detection and treatment could reduce the risk of heart attack. Currently, the golden standard for the diagnosis of CAD is angiography, which is an invasive procedure. In this article, we propose an algorithm that uses data mining techniques, a fuzzy expert system, and the imperialist competitive algorithm (ICA), to make CAD diagnosis by a non-invasive procedure. The ICA is used to adjust the fuzzy membership functions. The proposed method has been evaluated with the Cleveland and Hungarian datasets. The advantage of this method, compared with others, is the interpretability. The accuracy of the proposed method is 94.92% by 11 rules, and the average length of 4. To compare the colonial competitive algorithm with other metaheuristic algorithms, the proposed method has been implemented with the particle swarm optimization (PSO) algorithm. The results indicate that the colonial competition algorithm is more efficient than the PSO algorithm.

Category: Smart and intelligent computing

Keywords: CAD; Decision tree; Fuzzy; ICA; Membership functions; PSO

I. INTRODUCTION

In the cardiovascular disease group, coronary artery disease (CAD) is the most prevalent disease that leads to death. CAD causes roughly 1.2 million heart attacks each year, and more than 40% of those suffering from a heart attack will die. Even more worrisome, 335,000 people with heart attacks will die in an emergency department, or before ever reaching the hospital. According to the American Heart Association, over 7 million Americans have suffered a heart attack in their lifetime [1].

In the CAD, the coronary arteries become narrower, and the heart muscles are deprived of an adequate supply of blood and oxygen.

When the blood flow slows down, the heart doesn't receive enough oxygen and nutrients. This usually results in chest pain, called angina. When one or more of the coronary arteries are completely blocked, the result is a heart attack [2].

To diagnose the CAD, many factors have to be considered, which makes the detection difficult and time-consuming. Moreover, some of the detection tests, such as

Open Access http://dx.doi.org/10.5626/JCSE.2014.8.2.87 http://jcse.kiise.org

Received 21 January 2014; **Revised** 10 April 2014; **Accepted** 13 May 2014
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

electrocardiography (ECG), echocardiography, and the exercise test, are not very accurate. Currently, the gold standard for the diagnosis of CAD is angiography, which is an expensive and invasive procedure.

In recent decades, many experts have tried to make CAD diagnosis using computer-aided techniques, such as neural networks [3], the Bayesian model and decision tree [4], support vector machine [5], and the naive Bayes classifier [6]. The purpose of such researches is to help physicians to make CAD diagnosis by a non-invasive procedure.

This article proposes an optimization system using two evolutionary algorithms, the imperialist competitive algorithm (ICA) and particle swarm optimization (PSO), based on a fuzzy expert system that involves four stages. The first stage is to preprocess the data used for training. This primarily involves dealing with missing values, outliers, and feature selection. In the second stage, a decision tree is created, and a set of rules are extracted from it. In the third stage, the rules are transformed into fuzzy rules, using fuzzy membership functions. Finally, in the fourth step the fuzzy membership functions are separately tuned by ICA and PSO, and the results are then compared. The fuzzy model with the optimized parameters results in the final fuzzy expert system (FES). Since the generated FES is based on a set of rules, the decisions that are made provide interpretability.

Generally, the innovations of this paper are as follows. This papar proposes a hybrid method for rule learning, using membership functions based on the ICA and fuzzy set; proposes a method with acceptable performance and sensitivity, in comparison with other interpretable methods; and proves the superiority of ICA, in comparison with PSO, from the viewpoints of performance and convergence.

This paper is organized as follows. Section II briefly introduces the datasets employed, and the decision tree algorithm. In Section III, the fuzzy expert system and its parameters are presented. Simulation details and the results are reported in Section IV. The conclusion and future works are provided in Section V.

II. DATA SETS

To evaluate the proposed system, two datasets in heart disease have been used. Both datasets are provided by the University of California at Irvine. The datasets are the Hungarian Institute of Cardiology, Budapest, and the Cleveland Clinic datasets [7]. There are 597 records in total. There are 76 attributes in the datasets, of which all researches just use 14: 13 of them as inputs and 1 attribute as a result. The 13 input attributes include age, blood pressure, serum cholesterol, maximum heart rate, sex, chest pain type, fasting blood sugar, resting ECG, exercise-induced angina, oldpeak, slope, fluoroscopy, and thallium scan. The output variable is the angiography status.

A. Preprocessing the Data

The first step is to preprocess the data sets. This primarily includes the filling of missing values, removal of outliers, and feature selection. The different preprocessing steps that are used are: filling the missing values, dealing with outliers, and normalizing the data and attribute selection.

Real databases frequently include missing data for many reasons, such as the tests not having been performed entirely, or the data being unavailable. Handling them is a very significant step, because they could lower the accuracy of classification. There are different methods for this purpose. Two important methods are removing attributes, including missing data, and data imputation, which estimates and calculates the missing values, by methods like mean and mode values. The first one is used when there are a small number of missing values. In this system, the second method is employed, because of the relatively high number of missing data from the CAD dataset. The categorical values are replaced with the mode, and the numerical values with the mean [8].

Having accurate results, the data preprocessing steps were thoroughly performed: filling the missing data, removing outliers, and normalizing the data had all been done.

Distance-based outliers methods are used to detect the outliers, using the k-nearest neighbor and Euclidean distance. In the normalization step, the intervals of all data were changed to between [0,1]. Data were normalized with the Eq. (1).

$$Normalize\ (x) = \frac{x - X_{min}}{x_{min} - x_{max}} \tag{1}$$

B. Decision Tree Algorithm

A decision tree is a decision support tool that uses a tree-like model. It is a flowchart like tree structure, in which each node represents a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. A decision tree can easily be converted to classification rules [9]. Decision trees have several advantages. They are simple to understand, can model large and complex data, and can be combined with other methods. A path from root to leaf represents a classification rule. A decision tree is used for feature selection, and as a rule producer in the proposed system.

Sometimes, the training and learning of decision trees produces large trees. If the tree is too large, the class of a new instance is hard to determine. Pruning decision trees is an important step in the learning phase. The size of the decision tree can be reduced by removing unnecessary

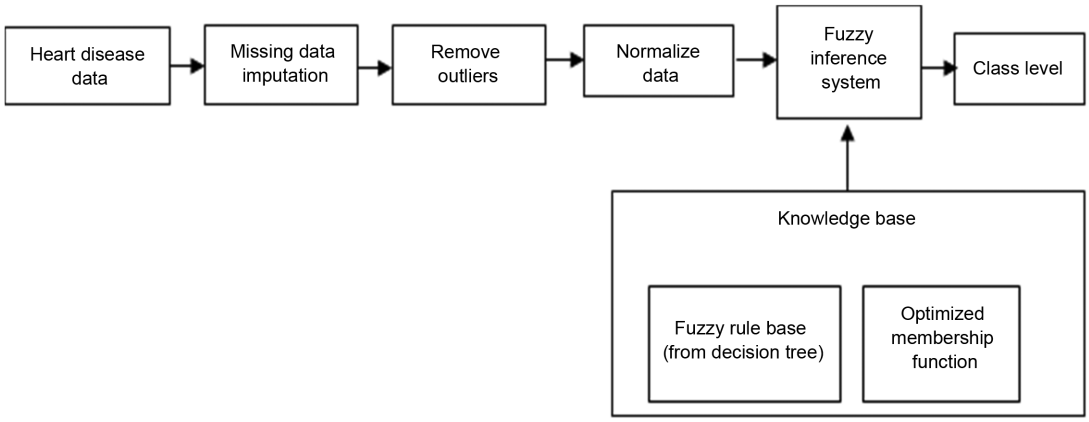

Fig. 1. Components of the proposed fuzzy expert system.

features that do not affect the decision tree classification accuracy. Tree pruning will increase the classification accuracy for future instances. To obtain high classification accuracy, pruning the decision tree was done in the proposed system.

III. FUZZY EXPERT SYSTEM

The extracted rules from the decision tree will change to fuzzy rules. This requires having a fuzzy model. To create a fuzzy model, there are three main steps to perform: fuzzification, fuzzy inference system, and deffuzification process. The fuzzy model design process includes the following definitions: 1) input and output variables, 2) fuzzy membership functions for each variable, and 3) fuzzy rules.

The input variables are designed based on features selected by a decision tree. The output variable is the class field in the dataset. Fuzzy membership functions are defined based on the boundary values of each branch of the decision tree. Fuzzy rules are designed based on decision tree rules, and membership functions. The proposed fuzzy expert system components are illustrated in Fig. 1. As shown in Fig. 1, after the preprocessing step, heart disease data are entered into the fuzzy inference system, and are classified. This classification and inference are based on knowledge base information.

The Sugeno fuzzy inference system was used for the proposed system. In the fuzzification process, the set of rules extracted from a decision tree is transformed into fuzzy rules. The main part of a fuzzy model is the membership function of each feature. The desired datasets have 13 features, among which just 7 features were employed that have more influence than others in decision-making. The feature selection task is performed by the decision tree. The features that are selected are: resting blood pressure (trestbps), chest pain type (cp), thal, number of major vessels colored by fluoroscopy (ca), serum cholesterol (chol), ST depression induced by exercise relative to rest (oldpeak), and maximum heart rate achieved (thalach).

With the normalization step, the intervals of all data were changed to between [0,1]. The output variable refers to the presence of heart disease in the patient. It has two values (0 and 1), which stand for health and heart disease status.

A. Membership Function Optimization

Designing fuzzy membership functions and fuzzy rules are very important phases in a fuzzy model, because they directly impact system performance. To optimize membership functions, an intelligent method is proposed that uses evolutionary algorithms, namely the imperialist competitive algorithm and particle swarm optimization.

In contrast with most optimization algorithms, such as PSO, genetic algorithms, and the ant colony algorithm, which are each inspired by natural processes, ICA is the first algorithm that is inspired by a social-political behavior, that of imperialism.

The ICA was proposed in 2007, and like other evolutionary algorithms, starts with an initial population. Population individuals, called countries, are of two types: colonies and imperialists, which together form empires [10]. The competition among these empires forms the basis of the ICA. During this competition, weak empires collapse, and powerful ones take possession of their colonies. The algorithm continues, until just one empire remains.

In addition to the ICA, PSO algorithm also refers to a relatively new family of evolutionary algorithms, and is inspired by the social behavior of certain animals, such as flocking birds, schools of fish, swarms of bees, and even human behavior.

There are three parameters for each triangular fuzzy membership function. Fig. 2 shows the membership parameters: C (center), L (left), and R (right) correspond to the original membership function, where C', L', and R'

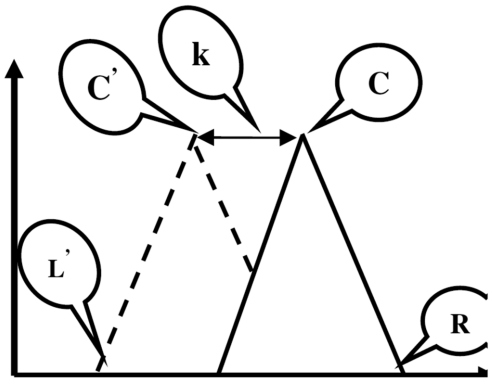

Fig. 2. Triangular fuzzy membership function parameters. Adapted from [11].

refer to the center, left, and right of the adjusted membership function, respectively.

The following equations are defined to adjust the membership functions:

$$C' = (C + k_i) \tag{2}$$

$$L' = (L + k_i) - w_i \tag{3}$$

$$R' = (R + k_i) - w_i \tag{4}$$

where, k_i and w_i are adjustment coefficients. k_i moves each membership function to the left or right. The membership function shrinks or expands through the parameter w_i. ICA and PSO were used to find appropriate values for the k_i and w_i parameters.

Fig. 3 illustrates a step-by-step flowchart of the proposed approach.

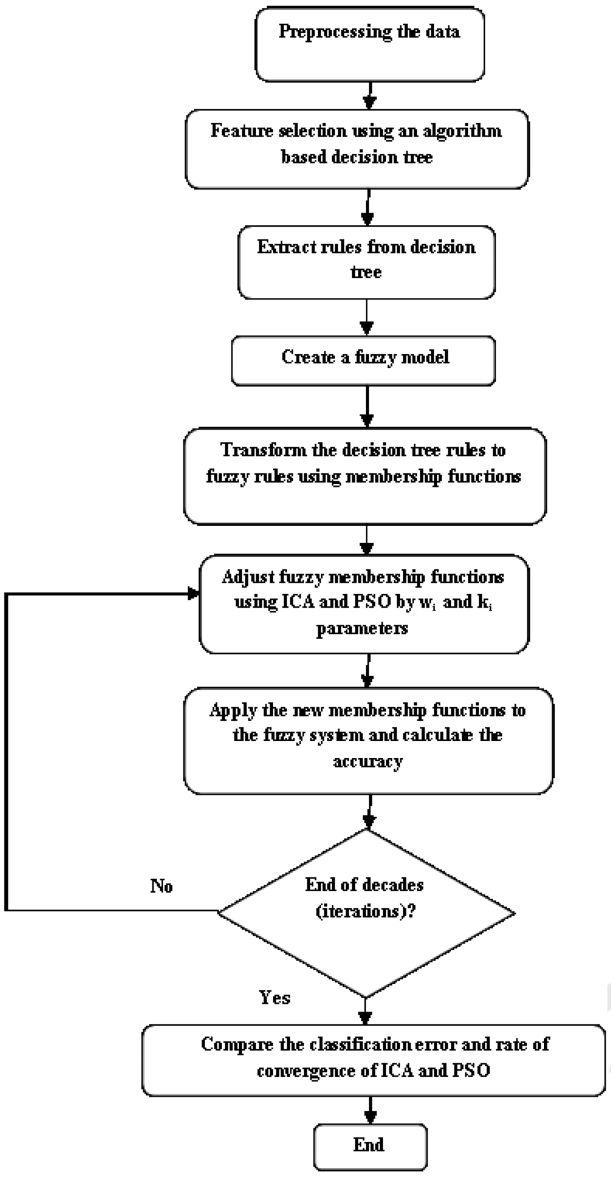

Fig. 3. Flowchart of the proposed system. ICA: imperialist competitive algorithm, PSO: particle swarm optimization.

IV. RESULTS AND DISCUSSION

The proposed fuzzy system is designed using MATLAB 7.12. The designed fuzzy system has 11 rules, and the average length of 4.

A 10-fold cross validation method was used to determine the training and test sets. Each time, 1/10 of the data was the test data, and the others were used for training. The test data had 59 instances out of 597 instances, i.e., about 1/10 of all instances. The proposed fuzzy system was evaluated using the test sets, and its performance was given as a confusion matrix (CM).

$$\boldsymbol{CM} = \begin{pmatrix} \boldsymbol{TP} & \boldsymbol{FP} \\ \boldsymbol{FN} & \boldsymbol{TN} \end{pmatrix} \tag{5}$$

True positive (TP): Number of instances in which the classifier detects them as positive, and the detection is true.

True negative (TN): Number of instances in which the classifier detects them as negative, and the instances don't belong to the positive class.

False negative (FN): Number of instances in which the classifier doesn't detect them, but they are positive.

False positive (FP): Number of instances in which the classifier detects them as positive, but they are not positive.

Using the confusion matrix, five evaluation measures, accuracy, sensitivity (recall), specificity, precision, and F-measure were evaluated, as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{6}$$

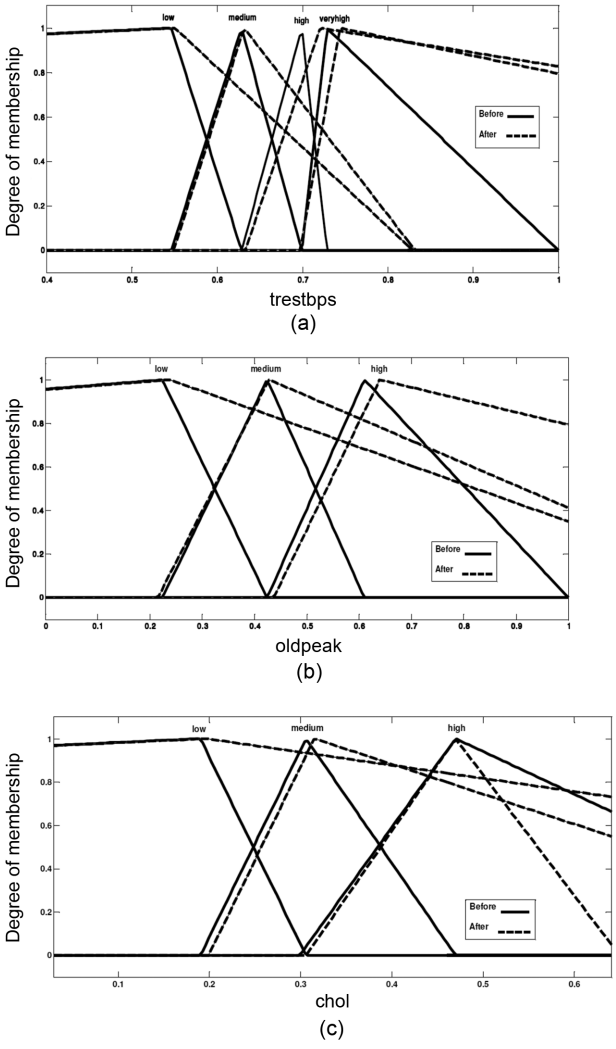

Fig. 4. The membership functions of (a) trestbps, (b) oldpeak, and (c) cholesterol attributes, before and after adjustment.

$$Sensitivity\ (recall) = \frac{TP}{TP+FN} \tag{7}$$

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

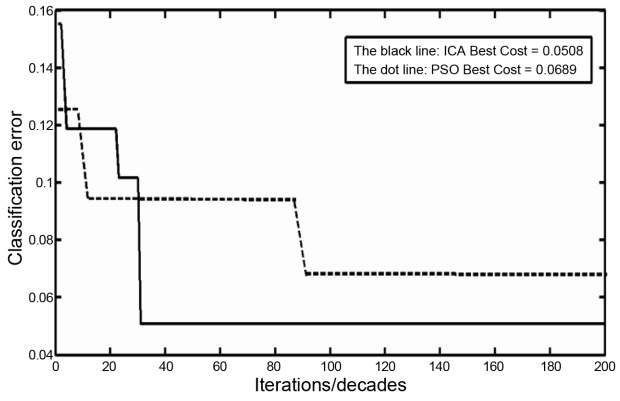

Fig. 5. Classification error using the imperialist competitive algorithm (ICA) and particle swarm optimization (PSO) algorithms.

$$F-measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

The accuracy determines the number of samples that were evaluated correctly. Sensitivity specifies the number of positive samples that were correctly classified. Specificity specifies the number of negative samples that were correctly classified.

The membership functions of attributes trestbps, oldpeak, and cholesterol are shown in (Fig. 4), before and after adjusting the membership functions.

The classification errors for ICA and PSO are illustrated in Fig. 5. The black line shows the classification error that is converted in about the 30th decade. The dotted line shows the classification error using the PSO algorithm. The convergence in PSO occurs later, than in ICA. As shown, the convergence is fixed in the 90th iteration. This proves that in adjusting the membership function problem, the speed of convergence of the ICA is better than that of the PSO algorithm.

From the viewpoint of performance, the ICA has 94.92% classification accuracy, and PSO has 93.11% classification accuracy. The comparison between ICA and PSO in this problem shows that the accuracy of classification using ICA is more than that of PSO; moreover, the convergence in ICA occurs earlier than in PSO.

Table 1 illustrates a comparison between the measured parameters for the CAD data set. Table 2 presents a comparison between the proposed system, and other systems that have worked on the CAD problem.

Table 1. Comparison parameters for the CAD dataset (unit: %)

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-measure
ICA	94.92	94.11	92.30	96.97	95.51
PSO	93.11	93.75	92.59	93.75	93.75

CAD: coronary artery disease, ICA: imperialist competitive algorithm, PSO: particle swarm optimization.

Table 2. Comparison of the proposed method with similar works

Method	Year	Accuracy (%)	Sensitivity (%)	Specificity (%)
Weighted fzzy rules [12]	2011	57.85	45.22	68.75
Decision tree [6]	2009	78.9	72.01	84.48
Support vector machine [13]	2010	79.17	-	-
Bagging [6]	2009	81.41	74.93	86.64
k-NN [14]	2009	81.50	-	-
Baysian model [4]	2010	82.00	87.00	-
Decision tee (C4.5) [4]	2010	82.50	87.17	-
Fuzzy [15]	2012	84.20	95.85	83.33
Ensemble neural network [16]	2008	89.01	80.95	95.91
PSO and fuzzy (proposed)	2013	93.11	93.75	92.59
Ensemble PSO and fuzzy [17]	2011	92.59	-	-
PSO and fuzzy [11]	2012	93.20	93.20	93.30
ICA and fuzzy (proposed)	2013	94.92	94.11	92.30

V. CONCLUSION AND FUTURE WORKS

Diagnosis of coronary artery disease has been performed by computer-aided techniques in recent decades. Although these techniques had an acceptable classification error, they do not have the ability for interpretation. This paper proposed a system with the ability for interpretation. In the proposed method, classification accuracy is 94.92%. The proposed system for the first time uses an ICA to diagnose CAD. A fuzzy expert system and decision tree are employed to obtain valid rules from data. The ICA is used to adjust fuzzy membership functions.

Generally, this paper presented the following tasks:

- *A hybrid method was proposed for rule learning, using the membership functions based on the ICA and fuzzy set*: the system adjusted the fuzzy membership functions using the ICA, and applied them to the fuzzy model. With such a new fuzzy model, we have a different classification error.
- *A method was proposed with an acceptable performance, in comparison with other interpretable methods*: interpretability is an important factor in medical systems. The proposed system presented this ability by 11 rules, with average length of 4. Results showed that the highest accuracy and sensitivity was obtained with such a number of rules and average length.
- *The superiority of ICA in comparison with PSO was proven, from the viewpoint of performance and convergence*: the diagram in Fig. 5 shows the superiority of the ICA, in comparison with PSO. ICA convergence occured in the 30th decade, while PSO achieved it in the 90th iteration. Moreover, the classification error by the ICA was far less, than by PSO.

The admissible results indicate that using an evolutionary algorithm and a fuzzy expert system would be efficient in problems of the diagnosis of disease.

Future works will focus on filling the missing values with intelligence methods. This leads to higher classification accuracy. Determining membership functions is a time-consuming process, and was performed manually in the proposed system. To reduce the errors that are caused by manual definition, we could define an intelligent procedure to automatically determine the membership functions. This could increase the classification accuracy.

REFERENCES

1. Webmd.com. "Risk factors for heart disease," http://www.webmd.com/heart-disease/risk-factors-heart-disease.
2. National Heart, Lung, and Blood Institute, "What is coronary heart disease?" http://www.nhlbi.nih.gov/health/health-topics/topics/cad/.
3. S. B. Patil and Y. S. Kumaraswamy, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of Scientific Research*, vol. 31, no. 4, pp. 642-656, 2009.
4. K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *Proceedings of the 5th International Conference on Computer Science and Education*, Hefei, China, 2010, pp. 1344-1349.
5. I. Babaoglu, O. K. Baykan, N. Aygul, K. Ozdemir, and M. Bayrak, "Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2562-2566, 2009.
6. M. C. Tu, D. Shin, and D. Shin, "Effective diagnosis of heart disease through bagging approach," in *Proceedings of*

2nd International Conference on Biomedical Engineering and Informatics, Tianjin, China, 2009, pp. 1-4.
7. C. Blake and C. J. Merz, "UCI repository of machine learning databases," Department of Information and Computer Science, University of California, Irvine, CA, 1998.
8. S. Sajja, "Data mining of medical datasets with missing attributes from different sources," Doctoral dissertation, Youngstown State University, Youngstown, OH, 2010.
9. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Amsterdam: Morgan Kaufmann, 2006.
10. E. Atashpaz-Gargari, and C. Lucas, "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition," in *Proceedings of IEEE Congress on Evolutionary Computation*, Singapore, 2007, pp. 4661-4667.
11. S. Muthukaruppan and M. J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11657-11665, 2012.
12. P. K. Anooj, "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules," *Journal of King Saud University-Computer and Information Sciences*, vol. 24, no. 1, pp. 27-40, 2012.
13. I. Babaoglu, O. Findik, and M. Bayrak, "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2182-2185, 2010.
14. N. A. Setiawan, P. A. Venkatachalam, and M. H. Ahmad Fadzil, "Rule selection for coronary artery disease diagnosis based on rough set," *International Journal of Recent Trends in Engineering*, vol. 2, no. 5, pp. 198-202, 2009.
15. D. Pal, K. M. Mandana, S. Pal, D. Sarkar, and C. Chakraborty, "Fuzzy expert system approach for coronary artery disease screening using clinical parameters," *Knowledge-Based Systems*, vol. 36, pp. 162-174, 2012.
16. R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7675-7680, 2009.
17. N. G. Hedeshi and M. S. Abadeh, "An expert system working upon an ensemble PSO-based approach for diagnosis of coronary artery disease," in *Proceedings of the 18th Iranian Conference of Biomedical Engineering*, Tehran, Iran, 2011, pp. 249-254.

Zahra Mahmoodabadi

Zahra Mahmoodabadi received her B.S. degree in computer engineering from the Azad University of Mashhad, Iran, in 2010, and an M.S. degree in computer software engineering from Imam Reza International University, Mashhad, Iran, in 2013. Her research interests include intelligent computing systems, evolutionary algorithms, data mining techniques and knowledge discovery.

Mohammad Saniee Abadeh

Mohammad Saniee Abadeh received his B.S. degree in Computer Engineering from Isfahan University of Technology, Isfahan, Iran, in 2001, the M.S. degree in Artificial Intelligence from Iran University of Science and Technology, Tehran, Iran, in 2003 and his Ph.D. degree in Artificial Intelligence at the Department of Computer Engineering in Sharif University of Technology, Tehran, Iran in February 2008. Dr. Saniee Abadeh is currently a faculty member at the Faculty of Electrical and Computer Engineering at Tarbiat Modares University. His research has focused on developing advanced meta-heuristic algorithms for data mining and knowledge discovery purposes. His interests include data mining, bio-inspired computing, computational intelligence, evolutionary algorithms, fuzzy genetic systems and memetic algorithms.