Archive of Biomedical Science and Engineering

Maryam Negahbani[1], Sanaz Joulazadeh[2], Hamid Reza Marateb[3]* and Marjan Mansourian[4]

[1]Department of Mathematical Sciences, University of Isfahan, Isfahan, Iran
[2]Department of Electrical Engineering, University of Isfahan, Isfahan, Iran
[3]Department of Biomedical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran
[4]Department of Biostatistics and Epidemiology, School of Public Health, Isfahan University of Medical Sciences, Isfahan, Iran

Dates: **Received:** 04 April, 2015; **Accepted:** 04 June, 2015; **Published:** 08 June, 2015

***Corresponding author:** Hamid Reza Marateb, Biomedical Engineering Department, Engineering Faculty, the University of Isfahan, Hezar Jerib st, 81746-73441, Isfahan, Iran; Tel: +98-31-37935616; Fax: +98-31-37932771; E-mail: h.marateb@eng.ui.ac.ir

www.peertechz.com

Keywords: Coronary artery disease; Differential search algorithm; Generalized minkowski metrics; Multiple logistic regression; Supervised fuzzy C-means

ISSN: 2641-3027

Research Article

Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics

Abstract

Introduction: Coronary Artery Disease (CAD), one of the leading causes of death, is narrowing the walls of the coronary arteries. Angiography is the most accurate but invasive and costly CAD diagnosis method associated with mortality. The aim of this study was to design a computer-based non-invasive CAD diagnosis system.

Methods: In this work, a dataset from Cleveland clinic foundation, containing 303 patients and 20 features, was used. Supervised Fuzzy C-means (SFCM) classification was used to design a classifier for CAD diagnosis. The Generalized Minkowski Metrics (GMM) was used to handle objects containing different measurement scale features. The performance of the SFCM was assessed with/without Statistical Feature Selection (SFS). The weights of the GMM, i.e. the significance of different features, beside other classifier parameters were tuned using Differential Search Algorithm (DSA), and the validity of the proposed classifier was further investigated. The hold-out and 10-fold cross validation were used for the performance assessment.

Result: The average accuracy of the base classifier (SFCM + GMM) was 79% (hold-out validation). It increased to 82% when using SFS. The average accuracy, sensitivity and specificity of the DSA-based classifier were 88%, 86% and 88%, respectively (cross-validation).

Conclusion: The most important features were the number of major vessels colored by fluoroscopy, the family history of CAD, peak exercise systolic blood pressure, maximum exercise heart rate achieved, chest pain type, resting heart rate, Fasting Blood Sugar and gender. This classifier showed substantial agreement with the angiographic results. The hybrid diagnosis system is thus promising. However, it is necessary to improve its reliability.

Abbreviations

CAD: Coronary Artery Disease; DSA: Differential Search Algorithm; GMM: Generalized Minkowski Metrics; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; MLR: Multiple Logistic Regression; PSO: Particle Swarm Optimization; SFCM: Supervised Fuzzy C-Means

Introduction

Coronary Artery disease (CAD), the most common type of heart disease, is one of the leading causes of death in industrialized countries and is rapidly achieving the same dubious distinction in developing nations as well [1].

CAD is the result of the accumulation of plaques within the walls of the coronary arteries supplying blood to the myocardium [2]. Blockage of one or more coronary arteries interrupts the flow of the blood to the heart, which causes heart attack [3]. The CAD is considered when narrowing of at least one of the coronary arteries is more than 50% [4].

The CAD risk factors have been identified over the past several decades include abnormal levels of circulating cholesterol, hypertension, cigarette smoking, diabetes, male gender, postmenopausal state, advancing age, sedentary lifestyle, obesity, and a positive family history of premature vascular disease. Moreover, new risk factors have been emerged as elevated blood levels of homocysteine, fibrinogen, inflammation and infection, atherogenic lipoprotein phenotype, elevated levels of lipoprotein, insulin resistance syndrome, psychosocial factors and a number of genetic polymorphisms [5].

There are several diagnostic tools for CAD [6-8]. Some of the general diagnostic tests include physical examination, lab tests, Electrocardiogram (ECG), echocardiogram, stress test, electron beam computed tomography, coronary angiography and cardiac catheterization. One of the major limitations of ECG is the undiagnosed symptoms of CAD. On the other hand, another alternative invasive methodology, angiogram, is painful and discomfort to the patients. Furthermore, the above mentioned procedures take a lot of cost, time and effort [9].

Computer aided diagnostic methods which extract relevant features and use them in classifiers for automated detection of diseases, can overcome such difficulties. Such techniques are noninvasive and provide reproducible and objective diagnosis, and hence, can prove to be valuable adjunct tools in clinical practice [9].

Yan et al. used an improved back propagation algorithm to train the CAD medical diagnosis system [10]. A novel inference engine named fuzzy-evidential hybrid inference engine proposed by Khatibi et al., used Demister–Shafer theory of evidence and fuzzy sets theory to diagnose CAD. This hybrid engine precisely modeled the information's vagueness and decision making's uncertainty and through information fusion, provided accurate results [11]. A fuzzy expert system based on particle swarm optimization (PSO) was developed by Muthukaruppan et al., in order to classify heart disease and healthy condition. In the proposed method, the significant attributes and fuzzy rules were extracted using the decision tree algorithm [12]. Giri et al., proposed a methodology for the automatic detection of normal and CAD using heart rate signals. It was shown that Gaussian Mixture Model classifier had the best results among the three other classifiers Support Vector Machine, Probabilistic Neural Network and K-Nearest Neighbor [9]. Using feature selection and extraction algorithm, Alizadehsani et al., enriched the dataset. Then, Information Gain and confidence were used to determine the effectiveness of features on CAD [13].

In this study, we proposed an automated medical diagnosis system based on the statistical feature selection, supervised fuzzy c-means (SFCM) and Generalized Minkowski Metrics (GMM). Since the features used for CA diagnosis have different measurement scales (nominal, ordinal or interval), a mixed-type data distance metric was used. A statistical feature selection method was used to reduce the feature space. Alternatively, the weights of the input features were tuned on the GMM using a novel stochastic optimization method called Differential Search Algorithm (DSA) and the important features were selected. The data-set, methodologies and the validation procedure will be studied at the following sections.

Materials and Methods

Experimental methods

In this work, the CAD dataset from the University of California (UCI, Irvine), available online, taken from the Cleveland Clinic Foundation datasets [14-17], was used. This database consisted of 303 records with 76 attributes (features), among which 13 to 20 features have been widely used in the literature [12]. The experimental protocol of recording the dataset was mentioned elsewhere in details [14,18]. A number of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984, without the history of prior myocardial infarction or known volvuli or cardiomyopathy diseases, participated in the experiment. Different demographic and clinical attributes (some of which were listed in (Table 1)) were recorded from the CAD (case) and healthy (control) subjects [19]. When at least one of the coronary arteries narrowed more than 50%, shown by angiography, the CAD was considered in the subjects [4]. The aim of the study was to design a computer-based CAD diagnosis system using 30 recorded features whose outcome had acceptable agreement with that of coronary angiography. In the next section, the Fuzzy C-means (FCM) data mining techniques are introduced.

Fuzzy C-means (FCM) Algorithm in clinical applications

Risk factors are the smallest units indicating the existence of a disease. A syndrome, on the other hand, is a collection, a set, or a cluster of concurrent risk factors, which together indicate the presence and the nature of the disease. Here the main question is that what the relation between these risk factors and a specific syndrome is. Classification and clustering are therefore basic concerns in medicine. Classification depends on the definition of the classes and on the required degree of affiliation of their elements [20].

Clustering algorithms are generally divided into two groups. First, hard partitioning algorithms which are based on classical set theory; they require that an object either does or does not belong to a cluster. Soft clustering methods however, allow the objects to belong to several clusters simultaneously with different degrees of membership [21]. Fuzzy clustering methods are one of the well-known methods of soft clustering which are vastly used in solving medical diagnosis problems. In medicine, there are usually imprecise conditions and highly overlapping classes and therefore fuzzy methods seem to be more suitable than crisp ones [20].

FCM algorithm was first introduced by Bezdek as the enhancement to the classical K-means clustering [22]. This algorithm estimates the membership function of the object k to the clusters i ($u_{ik} \geq 0$) to minimize the following cost function [23]:

$$J_r = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}^{\,2} \quad (1)$$

Where d_{ik} is the distance measure of k^{th} data point from i^{th} cluster center and parameters c, n and $m \geq 1$ are the number of clusters and objects in the dataset and the fuzzy coefficient, respectively. In the probabilistic FCM, the following constrain must be met when optimizing the above cost function:

$$\sum_{i=1}^{c} u_{ik} = 1;\ k = 1,\ldots,n \quad (2)$$

The FCM algorithm iteratively estimates the cluster centers and the membership functions to minimize the cost function (J_r) which could be found elsewhere in details [23].

Despite all the benefits of using FCM as the clustering core algorithm, it is still a blind method and may misclassify the input data. Thus it is necessary to train the algorithm in a way that induces a meaningful convergence. As a result, the classification will be even more accurate.

In the original FCM the data distance to the cluster centers are normally calculated using standard Euclidean distance. In our case, each object is a vector of multiple risk factors with various measurement scales types in different ranges, hence using the classic Euclidean distance is not appropriate [24]. Basically, there are three major data types in clinical data sets: nominal, discrete ordinal, and Interval. Nominal scales are only used for non-ranked qualitative

Citation: *Negahbani M, Joulazadeh S, Marateb HR, Mansourian M (2015) Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics. Biomed Sci Eng 1(1): 006-0014.*

Table 1: The attributes of the raw Cleveland dataset for normal and Coronary Artery Disease (CAD) groups, along with their categories (percentage) for qualitative variables and (min-max) mean±SD for quantitative variables.

Attribute	Measurement Scale	Definition	Categories*	Demographic attributes	
				Normal	CAD
Age	Interval	Age in years	-	(29-76) 53±9	(35-77) 56±8
Gender	Nominal	Sex	Male/Female	Male (54.8%), female (45.2%)	Male (84.0%), female (16%)
Trestbps	Interval	Resting blood pressure (mmHg)	-	(94-180) 129±17	(100-200) 134±19
CHOL	Interval	Serum cholesterol (mg/dl)	-	(126-564) 244±53	(149-409) 256±48
FBS	Nominal	Fasting Blood Sugar > 120 (mg/dl)	True/False	False (85.4%), True (14.6%)	False (84.8%), True (15.2%)
Restecg	Nominal	resting electrocardiographic results	(1) Normal; (2) Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); (3)Showing probable or definite left ventricular hypertrophy by Estes' criteria	1 (56.7%), 2 (0.6%), 3 (42.7%)	1 (39.2%), 2 (0.8%), 3 (60.0%)
Thalrest	Interval	resting heart rate (bpm)	-	(49-119) 77±14	(40-109) 73±13
Cigs	Interval	number of cigarettes per day	-	0-99 (15±19)	0-80 (17±20)
Years	Interval	Number of years as a smoker	-	(0-50) 14±14	(0-54) 17±16
Famhist	Nominal	family history of CAD	Yes/No	No (42.0%), Yes (58.0%)	No (32.8%), Yes (67.2%)
Cp**	Nominal	chest pain type	(1) Typical angina pectoris; (2) Atypical angina; (3) Nonanginal pain; (4) No pain	1 (10.2%), 2 (22.3%), 3 (42.7%), 4 (24.7%)	1 (4.8%), 2 (6.4%), 3 (13.6%), 4 (75.2%)
Tpeakbps	Interval	peak exercise systolic blood pressure (mmHg)	-	(84-232) 170±23	(90-230) 165±25
Tpeakbpd	Interval	peak exercise diastolic blood pressure (mmHg)	-	(26-120) 78±14	(50-120) 79±12
Thalach	Intreval	maximum exercise heart rate achieved (bpm)	-	(96-202) 158±19	(71-195) 139±23
Exang	Nominal	exercise induced angina	Yes/No	No (85.4%), Yes (14.6%)	No (44.8%), yes (55.2%)
Oldpeak	Interval	ST depression induced by exercise relative to rest	-	(0-4) 0.6±0.8	(0-6) 1.6±1.3
Slope	Ordinal	the slope of the peak exercise ST segment	(1) Upsloping; (2) Flat; (3) Downsloping	1 (64.3%), 2 (30.6%), 3 (5.1%)	1 (27.2%), 2 (64.8%), 3 (8.0%)
Ca	Interval	number of major vessels (0-3) colored by fluoroscopy	-	(0-3) 1±1	(0-3) 1±1
Thal***	Nominal	thallium-201 stress scintigraphy	(1) Normal; (2) Fixed defect; (3) Reversible defect	1 (79.5%), 2 (3.8%), 3 (16.7%)	1 (28.2%), 2 (6.5%), 3 (65.3%)
Num	Nominal	diagnosis of heart disease (angiographic disease status)	(1) Normal : < 50% diameter narrowing; (2) CAD: > 50% diameter narrowing		

*: The categories were shown for nominal or ordinal features; **: (1) Typical angina pectoris: Pain that occurs in the anterior thorax, neck, shoulders, jaw, or arms is precipitated by exertion and relieved within 20 min by rest. (2) Atypical angina. Pain in one of the above locations and either not precipitated by exertion or not relieved by rest within 20 min. (3) Nonanginal pain. Pain not located in any of the above locations, or if so located not related to exertion, and lasting less than 10 sec or longer than 30 min. (4) No pain; ***: (1) Normal, (2) Fixed abnormality (defects observed during exercise that persisted at redistribution), and (3) Reversible abnormality (defects present during exercise and significantly corrected during redistribution).

classification e.g. gender, blood type, and health condition. A discrete–ordinal scale is a nominal variable, but the different states are ordered in a meaningful sequence e.g. the slope of the peak exercise ST segment. Interval scales are measured on a linear scale e.g. BMI (Body Mass Index) and age. It is important to define a distance measure to balance all these differences in a way that no feature lessens the other features' effect or vice versa.

SFCM algorithm

The class labels provide a useful guidance during training procedure. Hence, it is necessary to use the labeled samples in training phase and unlabeled samples in testing phase to improve the performance of FCM. This idea led to the development of a new algorithm called Supervised *Fuzzy C-Means* (*SFCM*) algorithm, a slight modification of FCM [25].

The main goal of SFCM is to use the labeled data samples to guide the iterative optimization procedure. In this method, a known fixed set of categories and category-labeled training data are used to induce a classification function. The determination of fuzzy partition matrix U (dividing *N* data sets into *C* classes) using Supervised Fuzzy C-Means clustering is an iterative optimization procedure. The objective function of SFCM classification is defined as:

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\, d_{ik}{}^2+a\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}-f_{ik})^m\, d_{ik}{}^2 \qquad (3)$$

Citation: *Negahbani M, Joulazadeh S, Marateb HR, Mansourian M (2015) Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics. Biomed Sci Eng 1(1): 006-0014.*

Where U is the fuzzy partition matrix, V is the cluster center, f_{ik} is the membership degree of k^{th} labeled sample belonging to the i^{th} cluster (value is either 0 or 1).

The coefficient *'a'* denotes the scaling factor. The role of *'a'* is to maintain a balance between supervised and unsupervised component within the optimization procedure and parameter *'m'* controls the amount of fuzziness in the classification. The typical value of m is 2 and $a=L/n$, L denotes the size of labeled samples [25]. However, it is better to tune these two parameters based on the properties of the dataset. Function J_m can take a large number of values, the smallest one being associated with the best clustering.

An effective algorithm for supervised fuzzy classification is discussed herein. The steps of algorithm are as below [20,25]:

1) Initiate fuzzy partition matrix, $U^{(0)}$, with random values between 0 and 1 and fix the number of cluster centers as the number of outcome classes.

2) Start the iterative procedure and set the iteration counter to one.

3) Calculate the cluster centers using the following equation:

$$v_{ij}^{(t)} = \frac{\sum_{k=1}^{n}\left(U_{ik}^{(t-1)}\right)^{m} Z_{kj}^{(train)}}{\sum_{k=1}^{n}\left(U_{ik}^{(t-1)}\right)^{m}} \quad (4)$$

Where, $V_{ij}^{(t)}$ represents the i^{th} cluster center of j^{th} feature which j changes from 1 to m (number of features), and $Z_{kj}^{(train)}$ is the k^{th} data instance corresponding to the m^{th} selected feature variable.

4) Calculate the distance between i^{th} cluster center and k^{th} dataset, distance measured with Euclidean Distance as follows:

$$d_{ik}^{(t)} = \left\| Z_k - v_i^{(t)} \right\| = \sqrt{\sum_{j=1}^{m}\left(Z_{kj}^{(train)} - v_{ij}^{(t)}\right)^2} \quad (5)$$

5) Update the fuzzy partition matrix for the next iteration given by the following equation:

$$u_{ik}^{(t+1)} = (1-a)\left[\sum_{j=1}^{c}\left(\frac{d_{ik}^{(t)}}{d_{jk}^{(t)}}\right)^{\frac{2}{m-1}}\right]^{-1} + af_{ik} \quad (6)$$

For test set samples, whose class labels are unknown, the fuzzy partition matrix is calculated as follows:

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{c}\left(\frac{d_{ik}^{(t)}}{d_{jk}^{(t)}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (7)$$

6) When $\|U^{(t+1)} - U^{(t)}\| \le \varepsilon$ (ε is the iterative accuracy) has achieved, Stop the iteration; In this case outputs will be v (cluster center) and U (fuzzy matrix), otherwise return to step 3.

Statistical Feature Selection (SFS)

In many classification problems, especially in the biomedical domain, high dimensional data with few observations are used [26]. This can lead to lower classification accuracy and clusters of poor quality. High dimensional data is also a serious problem for many classification algorithms due to its high computational cost, memory usage andthe curse of dimensionality [27].

Since most of the features are redundant or irrelevant, feature selection method (FS) is used to pick a subset of features that are relevant to the target concept [28]. In this work, a statistical FS method entitled as Multiple Logistic Regression (MLR) was used, which is widely used to identify relevant risk factors in epidemiological studies. MLR, known as feature vector machine in machine learning, can be used to select statistically significant features. It not only considers significant features that provide acceptable discrimination between two classes, but it also takes into account the correlation between features. After running MLR on the input features (excluding the intercept point in the analysis), the selected features were used in the tested classifier [2,29,30].

Measurement scale

It is impossible to perform any arithmetic operation on nominal data because it has no order. The only Operation defined here is the equality. The distance of two nominal instances A and B is 1, if A equals B, and 0 otherwise.

For interval scales, it is possible to calculate the distance with standard norm definitions. The distance between two data samples A and B from a given interval I, is defined as $|A-B|$. As the interval size could be different between multiple features, it is important to normalize the distances relative to the interval size given as $|A-B| / |I|$.

Discrete–ordinal scale is a nominal variable, but the different states are ordered in a meaningful sequence. Ordinal data has order, but the intervals between scale points may be uneven. But still, the distance of two samples lying in the same interval is computed similar to that of interval distance i.e. $|A-B| / |I|$. Now, each feature is normalized to a value between 0 and 1.

Note that $|I|$ is calculated by subtracting the maximum and minimum values herein. The l1-norm was then used to combine the distance between different transformed features, simply known as GMM in the literature [31,32]. Thus, the GMM distance definition between two feature vectors A and B, could be given as below:

$$d_{GMM}(A,B) = \sum_{k=1}^{d} c_k \psi(A_k, B_k) \quad (8)$$

Where d is the number of dimensions, Ψ is the distance function for each feature which varies according to its measurement scale, and C_k are weights. The weights could be either set to the value of unity or tuned using an optimization algorithm. Accordingly, the features' weights were set to unity when using GMM+SFCM with/without MLR. Alternatively, instead of using SFS, all of the features were used and their weights were calculated and the features with small value of weights were neglected. For the later approach, it is necessary to use an efficient optimization algorithm, discussed at the next section.

Differential Search Algorithm (DSA)

DSA is an optimization algorithm developed by P. Civicioglu simulating the Brownian-like random-walk movement used by an organism to migrate [33]. The motivation of DSA, like many

population-based stochastic optimization algorithms, was taken from the nature. Many living organisms show annual migration. In this migration, super organism is constituted containing large number of individuals. The movement of a super organism could be illustrated by a Brownian-like random-walk model [33,34]. In DSA, the population contains random solutions. The migration is performed to the global optimum of the cost function. At each iteration, some of the populations are selected. They move based on a Brownian-like random walk model [34]. DSA is simple to implement and was shown to have acceptable performance in variety of the optimization problems in comparison with that of other traditional optimization algorithms while it is not too sensitive to the initialization of its parameters [33].

We have used DSA to optimize the parameters of GMM and SFCM. The cost function was the absolute error rate of the classifier on the training set. The initial setting of the DSA used in our study was similar to that of P. Civicioglu [33]. Briefly, the size of the population was set to 30, and the maximum number of function evaluation value (i.e. the number of times that the cost function was called in the program) as the only stopping criterion was 2,000,000.

Performance measures for classification

The performance of a classifier could be evaluated by computing the number of correctly recognized CAD subjects (TP: True Positives), the number of correctly recognized healthy subjects (TN: True Negatives), and examples that either were incorrectly assigned to the CAD class (FP: False Positives) or that were missed as class examples (FN: False Negatives). These four counts constitute the information-theory formulas to accurately measure the performance of the classification [35,36].

Results

The demographic information of the Cleveland CAD data was shown in (Table 1) for the case (CAD) and control (healthy) groups. To assess the performance of the base classifier, the main dataset was randomly divided into two roughly equal size datasets, namely dataset 1 and dataset 2 (hold-out validation method [23]). The best Accuracy (*Acc*) achieved when tuning on the dataset 1 and testing on the dataset 2 and vice versa in ten runs of the SFCM algorithm with GMM distance measure (unity weights) was shown in (Table 2). The maximum number of iterations was set to 100 in all the classifiers. The scaling factor (a) was tuned in the training set using exhaustive grid search (a=0.8) while the value of the fuzziness parameter (m) was set to 2 in the base classifier. The overall percentage accuracy (the average *Acc* of the classifier on dataset 2 when it was tuned on dataset 1 and vice versa) in the base classifier was 79%. The average Sensitivity and Specificity of the base classifier were 71% and 84%, respectively.

Table 3 shows the results of running the algorithm with SFS. MLR revealed that following significant features: gender, *cp* (chest pain type), *trestbps* (resting systolic blood pressure), *thalach* (maximum exercise heart rate achieved), *slope* (the slope of the peak exercise ST segment), *ca* (number of major vessels (0-3) colored by fluoroscopy), and *thal* (thallium-201 stress scintigraphy). The overall accuracy for SFCM-SFS was 82%. The average Sensitivity and Specificity of this classifier were 85% and 82%, respectively.

Finally, DSA optimization method was used to tune the features' weights, the fuzziness parameter and the scaling factor. The cost function was set as the absolute error rate of the classifier (SFCM+GMM) (i.e. 1-*Acc*) on the training set. Guarding against Type III error [37], 10-fold cross-validation [23] was used to assess the performance of the proposed hybrid classifier (Table 4). The average Accuracy, Sensitivity and Specificity of this classifier were 88%, 86%, and 88% respectively. The McNemar's test [23] revealed that the performance of the hybrid classifier was significantly better than the base classifier (p_value<0.05) but comparable with that of SFS+SFCM. Meanwhile, Multi-fold cross validation was used instead of leave-one-out, since it is proven to have better performance in terms of accuracy and efficiency [23].

The values of the parameters of the hybrid classifier tuned using DSA were shown in Table 5. The most important features (feature weight w>0.5) were listed in the descending order: the number of major vessels colored by fluoroscopy (w=1), the family history of CAD (w=1), peak exercise systolic blood pressure (w=0.89),

Table 2: The performance of the base Supervised Fuzzy c-means (SFCM) classifier using Generalized Minkowski Metrics (GMM).

Sets	Training				Test			
Indices	Se	Sp	Pr	Acc	Se	Sp	Pr	Acc
Scenario #1	100	100	100	100	67	76	64	73
Scenario #2	79	93	91	87	81	92	94	86

Scenario #1: the classifier was trained on the dataset n.1 and tested on the dataset n.2; Scenario #2: the classifier was trained on the dataset n.2 and tested on the dataset n.1; Se: Sensitivity (%), Sp: Specificity (%), Pr: Precision (%), Acc: Accuracy (%)

Table 3: The performance of the Supervised Fuzzy c-means (SFCM) classifier using Generalized Minkowski Metrics (GMM) with Statistical feature Selection (SFS).

Sets	Training				Test			
Indices	Se	Sp	Pr	Acc	Se	Sp	Pr	Acc
Scenario #1	100	100	100	100	81	80	77	80
Scenario #2	92	90	88	91	88	83	76	85

Scenario #1: the classifier was trained on the dataset n.1 and tested on the dataset n.2; Scenario #2: the classifier was trained on the dataset n.2 and tested on the dataset n.1; Se: Sensitivity (%), Sp: Specificity (%), Pr: Precision (%), Acc: Accuracy (%)

Table 4: The performance of the hybrid Supervised Fuzzy c-means (SFCM) classifier with Differential Sequential Algorithm-based Generalized Minkowski Metrics (GMM).

Sets	Training			
Indices	Se	Sp	Pr	Acc
Average	86±10	88±8	85±10	88±8

Se: Sensitivity (%), Sp: Specificity (%), Pr: Precision (%), Acc: Accuracy (%); 10-fold cross validation was performed and the average values of the indices were shown in Mean±SD

Table 5: The value of the system parameters obtained based on tuning on the training data set by the Differential Sequential Algorithm (DSA) optimization in the hybrid classifier.

Parameter	Value	Parameter	Value
Age	0.085382	thalrest	0.674363
Gender	0.539178	tpeakbps	0.891240
cp	0.817916	tpeakbpd	0.032989
trestbps	0.000000	trestbpd	0.417742
chol	0.289871	exang	0.311712
cigs	0.068179	oldpeak	0.406722
years	0.376728	slope	0.000000
fbs	0.640574	ca	1.000000
famhist	1.000000	thal	0.275304
restecg	0.074037	m	2.638523
thalach	0.865246	a	0.532491

Parameters are Generalized Minkowski Metrics (GMM), weights except a: scaling factor; m: fuzziness coefficient

maximum exercise heart rate achieved (w=0.87), chest pain type (w=0.82), resting heart rate (w=0.67), Fasting Blood Sugar (w=0.64) and gender (w=0.54). However, the list significant attributes (w<0.1) were age, resting blood pressure, number of cigarettes per day, resting electrocardiographic results, peak exercise diastolic blood pressure, and the slope of the peak exercise ST segment. For the definition, and number of categories of the above attributes the reader is referred to (Table 1).

The overall performance of the hybrid classifier was shown in Table 6. It includes the contingency table (confusion matrix) on the total of 303 subjects. The agreement rate between the results of this classifier and those of the gold standard (i.e. CAD diagnosis using angiography) was assessed based on the Cohen's *kappa* coefficient [38]. Substantial agreement was shown between the outcomes of the proposed hybrid classifier and angiography (*kappa*=0.73) [39].

Discussion

In this paper, three classification systems were designed for non-invasive CAD diagnosis; among which the hybrid classifier showed better performance (Tables 2-4). This diagnosis system was based on the SFCM classifier in which the distance between objects were calculated using GMM and the parameters of the system (SFCM parameters and GMM weights) were estimated using DSA optimization. Other approaches such as PSO [40] were used for optimization, but DSA showed more accurate results. The Type I error and the power of the hybrid classifier were 0.1 and 86%, respectively. Since the data-set was not totally balanced (i.e. the number of cases and controls were not identical), F1-score measure might be more accurate than the accuracy. The average F1-score during 10-fold cross-validation was 85±10 (%), indicating that the proposed system is accurate. The comparison between the performance of the postposed system and some of the other systems designed on the CAD dataset was shown (Table 7). Some methods had higher accuracy that n that of the proposed system. We compared the result of the method proposed by Muthukaruppan et al. [12]. Although its accuracy was 93%, McNamara's test showed that it was not significantly higher than our hybrid system (p_value>0.05). Another issue is that among the methods listed in (Table 7), those in which Fuzzy classification was used, showed higher accuracies. Since most or all classificatory concepts in medicine are fuzzy, fuzzy taxonomy was used in our study. Meanwhile, it is very difficult to define sharp borders between various symptoms in the set of all symptoms and between various diseases in the set of diseases [41]. Thus, the framework of fuzzy systems is very useful to deal with the absence of sharp boundaries of the sets of symptoms, diagnoses, and phenomena of diseases [19,42].

The significant features selected by the DSA, were known to be directly involved in CAD. Fluoroscopy is one of the most popular non-invasive CAD diagnosis methods whose accuracy ranges between 35% and 75% in comparison with that of the gold standard (i.e. angiography) in the literature [43,44]. We permed a univariate (i.e. the number of major vessels (0-3) colored by fluoroscopy) classification based on the Receiver Operating Characteristic (ROC) plot. Its accuracy was 75% (Area under Curve: AUC=0.75; cut-off=0.5). The average number of vessels colures were statistically different in the CAD and normal group (independent-samples t-test; p_value 0.05). High value of GMM weights are in agreement with the statistical test. Thus, it was a suitable feature but not enough for accurate classification. The other traditional non-invasive CAD diagnosis method is thallium-201 stress scintigraphy. The prevalence of CAD in three groups of scintigraphy was statistically different (Chi-square test; p_value<0.05). Having designed a decision-tree classifier with scintigraphy feature, the accuracy was 76%. However, due to

Table 6: The overall performance of the hybrid classifier including the contingency table (confusion matrix) of the Coronary Artery Disease (CAD) diagnosis.

		Angiographic result	
		CAD	Healthy
The hybrid classifier t	CAD	114 (TP)	20 (FP)
	healthy	19 (FN)	150 (TN)

TP: True Positives; TN: True Negatives; FN: False Negatives; FP: False Positives

Table 7: Comparison of the proposed system with other similar CAD diagnosis systems.

Author	Method	Accuracy
(Joulazadeh et al., 2015)	Our Proposed FS-SFCM	87
(Detrano et al., 1989)	Probability theory (logistic regression)	77
(Gennari et al, 1989)	Clustering (CLASSIT conceptual system)	79
(Kukar et al., 1999)	Bayesian classification and neural network	80
(Haddad et al., 1999)	Neural network	48
(Cheung et al., 2001)	BNNF	81
(Cheung et al., 2001)	BNND	81
(Cheung et al., 2001)	Naïve Bayes	81
(Khatibi et al., 2010)	Fuzzy sets and evidence theories	91
(Senthil Kumar et al., 2011)	ANFIS	91
(Senthil Kumar et al., 2012)	Fuzzy resolution mechanism	92
(Muthukaruppan et al., 2012)	PSO based fuzzy expert system	93
(Kahramanli & Allahverdi, 2008)	Hybrid neural network system	87
(Das et al., 2009)	Neural network ensembles	89
(Polat et al., 2007)	Fuzzy-AIRS-Knn based system	87

CAD: Coronary Artery Disease

the directional correlation between fluoroscopy and scintigraphy (Eta=0.3), DSA estimated the fluoroscopy and the scintigraphy wegihts as 1.00 and 0.28. The other clinical variable is ST segment depression used in cardiography. Its eight was zero, indicating that no further information could be extracted by adding this variable. In the lietrature, the ranked order of CAD predictive were cardiac fluoroscopy score, thallium score and extent of Electrocardiography (e.g. ST segment depression) [45,46] which is in agreement with our findings.

CAD is associated with higher morbidity and mortality in women than in men [47]. It was also shown that the incidence of CAD in women aged less than 70 years is lower than their male counterparts [48]. In our study, the percentage of men and women having CAD were 84% and 16%, respectively. Considering that women in the CAD group had the age of 66 years old or lower, this is in agreement with our study. However, women usually have CAD 7 to 10 years later than men [49]. In our data-set, the average age of women and men who had CAD was 60±5 and 55±8 years, respectively. Moreover, gender was a significant feature (w=0.539). Meanwhile, the age by itself was not a significant feature in our study (Table 5; w=0.085). This is in agreement with the fact that the average age of people in the CAD and normal groups was 56±8 and 53±9 years, respectively (Table 1). This is related to the stratified age sampling used in our study.

Although, it is proved that high blood pressure increases the risk of CAD [50], it was not significant in our study. Meanwhile elevated resting heart rate is known as a CAD risk factor, which is in agreement with our findings (w=0.674) [51]. In the literature, the family history of CAD is a major CAD risk factor in adults [52]. This is in agreement with our findings where it had the highest GMM weight (w=1; Table 5). Meanwhile, fasting blood sugar was known as an important determinant of CAD [53], in agreement with our findings where its GMM weight was estimated as 0.641. A high total cholesterol level can increase your risk of cardiovascular disease. However, decisions about when to treat high cholesterol are usually based upon the level of LDL or HDL cholesterol, rather than the level of total cholesterol. This might explain the fact that the weight of the cholesterol was 0.289 in our study. Moreover, total cholesterol/ high-density lipoprotein cholesterol ratio were shown to be associated to CAD rather than cholesterol, by itself [54,55].

Chest pain type (GMM weight of 0.818) was divided into the following categories: Typical angina pectoris, atypical angina, non-anginal pain, and no pain. Typical angina (pain that occurs in the anterior thorax, neck, shoulders, jaw, or arms is precipitated by exertion and relieved within 20 min by rest) was the most common symptom of CAD [18]. It occurs when blood flow to an area of heart is decreased, impairing the delivery of oxygen and vital nutrients to the heart muscle cells. The byproduct of using this inefficient fuel is producing lactic acid that builds up in the muscle and causes pain [56].

The key to a good classification is a dataset containing all the

possibly relevant features (i.e. risk factors mentioned in the literature) with enough cases (i.e. suitable sample size). Although the sample size of the Cleveland dataset is rather high, some important features such as BMI, LDL and HDL are missing. Meanwhile, we are going to design an automated CAD risk assessment program, based on the findings of this study, in collaboration with Isfahan Healthy Heart Program [57]. Such a large database, could allow us to investigate the accuracy of the proposed diagnosis system in a broader sense. Another issue is that the performance of the base classifier with/without FS (Tables 2,3) was so different in the first and second scenarios. Having calculated the cluster representatives for the healthy and CAD groups in the first and second datasets, the dataset 1 showed better discrimination in comparison with dataset2 on the whole 20 features and also those selected by the MLR. This is why that performance of the base classifier with/without FS was higher on the data set 1 in the entire training and test procedure. Also, the discrimination with/without FS was not that different. This, in fact, shows that the FS could have selected features with significant discrimination power.

Another step would be developing a web-based online system with which patients/ medical doctors could assess their risk of having CAD at home. These Web-based diagnostic decision support systems have been recently focused in Medicine and are proven to be valuable in identifying the correct diagnosis in complicated cases [58]. There might be two possible approaches to improve the performance of the proposed diagnosis system. First, further features could be defined by considering the interactions between input risk factors/predictors [59] e.g. simply multiplication of the predictors. Second, multiple clusters could be formed for each healthy and CAD class by using mixed-type data clustering methods [24]. Then, supervised FCM could be used with multiple clusters corresponding with two healthy and CAD classes. Extracting supervised classification rules on groups of similar objects could potentially reduce the misclassification rate especially close to the class borderlines. These two approaches will be the focus of our future work.

Conclusion

The hybrid classifier showed the average accuracy of 87%. The power of the designed diagnosis system was 86%. Type I error (α) was 0.1 and the F-score was 85%. Although the power of the method is acceptable, type I error must be reduced down to 0.05, to introduce a reliable and accurate clinical test which is the focus of the future work. One possible strategy to improve the accuracy of the proposed diagnosis system is using classifier fusion. Combining different reliable classifiers, might improve the accuracy though the fusion procedure. In conclusion, we designed an automated non-invasive CAD diagnosis system based on the Fuzzy theory. The results showed that the proposed system is promising. However, further improvements are needed to be able to use it in clinical laboratories.

Acknowledgments

This work was supported by the University of Isfahan (MN, SJ, HM) and Isfahan University of medical Sciences (MM).

References

1. Squeri (2012) Coronary Artary Disease – New Insights and Novel Approaches: In Tech.
2. Zhao L, Chen Y, Schaffner DW (2001) Comparison of logistic regression and linear regression in modeling percentage data, Appl. Environ. Microbiol 67: 2129–2135.
3. Romaine DS, Randall OS (1956) The Encyclopedia of the Heart and Heart Disease
4. Setiawan NA, Venkatachalam PA, Hani AFM (2009) Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System, Proceedings of the International Conference on Man-Machine Systems (ICoMMS).
5. Shah PK (2006) Risk Factors in Coronary Artery Disease Fundamental and Clinical Cardiology: CRC Press
6. Kahramanli H, Allahverdi N (2008) Design of a hybrid system for the diabetes and heart disease.Expert Systems with Applications 35: 82-89.
7. Polat K, Şahan S, Güneş S (2007) Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing, Expert Systems with Applications 32: 625-631.
8. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications 36: 7675-7680.
9. Giri D, Acharya UR, Martis RJ, Sree SV, Lim TC (2013) Automated diagnosis of Coronary Artery Disease affected patients using LDA, PCA, ICA and Discrete Wavelet Transform, Knowledge-Based Systems 37: 274–282.
10. Yan H, Jiang Y, Zheng J, Peng C, Li Q (2006) A multilayer perceptron-based medical decision support system for heart disease diagnosis, Expert Systems with Applications 30: 272–281.
11. Khatibi V, Montazer GA (2010) A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment, Expert Systems with Applications 37: 8536–8542.
12. Muthukaruppan S, Er MJ A (2012) hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease, Expert Systems with Applications 39: 11657–11665.
13. Alizadehsania R, Habibia J, Hosseini MJ, Mashayekhi H, Boghrati R, et al. (2013) A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed 3: 52–61.
14. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid J, et al. (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease, American Journal of Cardiology 64: 304--310.
15. Aha DW, Kibler D (1988) Instance-based prediction of heart-disease presence with the Cleveland database. University of California, CA1988.
16. Gennari JH, Langley P, Fisher D (1989) Models of incremental concept formation, Artificial Intelligence 40: 11--61.
17. Janosi A University of California, Irvine; Machine Learning Repository; Heart Disease Data Set.
18. Detrano R, Yiannikas J, Salcedo E, Rincon G, Go RT, et al. (1984) Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease, Circulation 69: 541-547.
19. Marateb HR, Goudarzi S (2015) A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system, Journal of Research in Medical Sciences 20.
20. Berks G, v. Keyserlingk DG, Jantzen J, Dotoli M, Axer H (2000) Fuzzy Clustering - A Versatile Mean to Explore Medical Databases
21. Babuska R (2001) Fuzzy and Neural Control-DISC Course Lecture Notes.
22. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms: Kluwer Academic Publishers
23. Webb AR, Copsey KD (2011) Statistical pattern recognition, 3rd ed. Hoboken: Wiley.

24. Marateb HR, Mansourian M, Adibi P, Farina D (2014) Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies, J Res Med Sci 19: 47-56.

25. Kalyani S, Swarup KS (2010) Supervised fuzzy C-means clustering technique for security assessment and classification in power systems, International Journal of Engineering, Science and Technology 2: 175-185.

26. Zakharov R, Dupont P (2011) Ensemble logistic regression for feature selection. Pattern Recognition in Bioinformatics, Springer 7036: 133-144.

27. Janecek AGK, Gansterer WN, Demel MA, Ecker GF (2008) On the Relationship Between Feature Selection and Classification Accuracy, MLR: Workshop and Conference Proceedings. New challenges for feature selection 4: 90-105.

28. Dash M, Liu H (1997) Feature Selection for Classification, Intelligent Data Analysis1: 131–156.

29. Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res1: 211–244.

30. Marateb HR, Mansourian M, Faghihimani E, Amini M, Farina D (2014) A hybrid intelligent system for diagnosing microalbuminuria in type 2 diabetes patients without having to measure urinary albumin, Computers in Biology and Medicine 45: 34–42.

31. Ichino M, Yaguchi H (1994) Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis, ieee transactions on systems, man, and cybernetics 24: 698-708.

32. Ichino M (1988) General metrics for mixed features-the Cartesian space theory for pattern recognition. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics 1: 494-497.

33. Civicioglu P (20120 Transforming geocentric Cartesian coordinates to geodetic coordinates by using differential search algorithm. Computers & Geosciences 46: 229-247.

34. Trianni V, Tuci E, Passino KM, Marshall JA (2011) Swarm cognition: an interdisciplinary approach to the study of self-organizing biological collectives. Swarm Intelligence 5: 3-18.

35. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks, Information Processing and Management 45: 427–437.

36. Marateb HR, Mansourian M, Adibi P, Farina D (2014) Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. J Res Med Sci 19: 47-56.

37. Mosteller F (2006) A k-sample slippage test for an extreme population. In Selected Papers of Frederick Mosteller, ed: Springer 101-109.

38. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Computational linguistics 22: 249-254.

39. Gwet KL (2014) Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters: Advanced Analytics, LLC

40. Marateb HR, McGill KC (2009) Resolving superimposed MUAPs using particle swarm optimization. IEEE transactions on bio-medical engineering 56: 916-919.

41. Sadegh-Zadeh K (2011) Handbook of analytic philosophy of medicine. Springer Science & Business Media 113.

42. Seising R (2006) From vagueness in medical thought to the foundations of fuzzy reasoning in medical diagnosis. Artificial Intelligence in Medicine 38: 237-256.

43. Detrano R, Simpfendorfer C, Day K, Salcedo EE, Rincon G, et al. (1985) Comparison of stress digital ventriculography, stress thallium scintigraphy, and digital fluoroscopy in the diagnosis of coronary artery disease in subjects without prior myocardial infarction. Am J Cardiol 56: 434-440.

44. Vliegenthart R (2004) Detection and Quantification of Coronary Calcification, in Coronary Radiology, M. Oudkerk, Ed., ed: Springer Berlin Heidelberg 175-184.

45. Hung J, Chaitman BR, Lam J, Lesperance J, Dupras G, et al. (1984) Noninvasive diagnostic test choices for the evaluation of coronary artery disease in women: a multivariate comparison of cardiac fluoroscopy, exercise electrocardiography and exercise thallium myocardial perfusion scintigraphy. J Am Coll Cardiol 4: 8-16.

46. Nallamothu N, Ghods M, Heo J, Iskandrian AS (1995) Comparison of thallium-201 single-photon emission computed tomography and electrocardiographic response during exercise in patients with normal rest electrocardiographic results. J Am Coll Cardiol 25: 830-836.

47. Eastwood JA, Doering L (2005) Gender Differences in Coronary Artery Disease. J Cardiovasc Nurs 20: 340-351.

48. Wake R, Yoshiyama M (2012) Gender Differences in Coronary Artery Disease, in Coronary Artery Disease - Current Concepts in Epidemiology, Pathophysiology, Diagnostics and Treatment, D. Gaze, Ed., ed: In Tech

49. M Maas AHE, Appelman YEA (2010) Gender differences in coronary heart disease. Netherland Heart Journal 18: 598–603.

50. Lieb W, Jansen H, Loley C, Pencina MJ, Nelson CP, et al. (2013) Genetic predisposition to higher blood pressure increases coronary artery disease risk. Hypertension 61: 995-1001.

51. Fox K, Ferrari R (2011) Heart rate: a forgotten link in coronary artery disease? Nat Rev Cardiol 8: 369-379.

52. Hoseini K, Sadeghian S, Mahmoudian M, Hamidian R, Abbasi A (2008) Family history of cardiovascular disease as a risk factor for coronary artery disease in adult offspring. Monaldi Arch Chest Dis 70: 84-87.

53. Lawes CM, Parag V, Bennett DA, Suh I, Lam TH, et al. (2004) Blood glucose and risk of cardiovascular disease in the Asia Pacific region. Diabetes Care 27: 2836-2842.

54. Nair D, Carrigan TP, Curtin RJ, Popovic ZB, Kuzmiak S, et al. (2009) Association of total cholesterol/ high-density lipoprotein cholesterol ratio with proximal coronary atherosclerosis detected by multislice computed tomography. Prev Cardiol 12: 19-26.

55. Castelli WP (1988) Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study. Can J Cardiol 4: 5A-10A.

56. Immke D, McCleskey E ASIC3: a lactic acid sensor for cardiac Pain. Scientific World Journal 1: 510-512.

57. Sarraf-Zadegan N, Sadri G, Afzali HM, Baghaei M, Fard NM, et al. (2003) Isfahan Healthy Heart Program: A comprehensive integrated community-based program for cardiovascular disease prevention and control. Acta cardiologica 58: 309-320.

58. Graber ML, Mathew A (2008) Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 23: 37-40.

59. Jacobsen SJ, Freedman DS, Hoffmann RG, Gruchow HW, Anderson AJ, et al. (1992) Cholesterol and coronary artery disease: age as an effect modifier. J Clin Epidemiol 45: 1053-1059.

Citation: *Negahbani M, Joulazadeh S, Marateb HR, Mansourian M (2015) Coronary Artery Disease Diagnosis Using Supervised Fuzzy C-Means with Differential Search Algorithm-based Generalized Minkowski Metrics. Biomed Sci Eng 1(1): 006-0014.*