Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms

Conference Paper · December 2012

DOI: 10.1109/ICDMW.2012.29

Roohallah Alizadehsani
Sharif University of Technology
19 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Asma Ghandeharioun
Massachusetts Institute of Technology
15 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Reihane Boghrati
University of Southern California
16 PUBLICATIONS **95** CITATIONS
SEE PROFILE

2012 IEEE 12th International Conference on Data Mining Workshops

Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms

Roohallah Alizadehsani
Department of Computer Engineering, Sharif University of Technology
e-mail:
Alizadeh_roohallah@yahoo.com

Mohammad Javad Hosseini
Department of Computer Engineering, Sharif University of Technology
e-mail:
mjhosseini@ce.sharif.edu

Corresponding author:
Zahra Alizadeh Sani
Tehran University of Medical Science
e-mail:
d_zahra_alizadeh@yahoo.com

Asma Ghandeharioun
Department of Computer Engineering, Sharif University of Technology
e-mail:
asma.ghandeharioun@gmail.com

Reihane Boghrati
Department of Computer Engineering, Sharif University of Technology
e-mail:
r.boghrati@gmail.com

***Abstract*— One of the main causes of death the world over are cardiovascular diseases, of which coronary artery disease (CAD) is a major type. This disease occurs when the diameter narrowing of one of the left anterior descending, left circumflex, or right coronary arteries is equal to or greater than 50 percent. Angiography is the principal diagnostic modality for the stenosis of heart vessels; however, because of its complications and costs, researchers are looking for alternative methods such as data mining. This study conducts data mining algorithms on the Z-Alizadeh Sani dataset which has been collected from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center. In this paper, the reason of effectiveness of a preprocessing algorithm on the dataset is investigated. This algorithm which has been merely introduced in our previous works, extracts three new features from the dataset. These features are then used to enrich the primary dataset in order to achieve more accurate results. Moreover, despite the fact that misclassification of diseased patients has more side effects than that of healthy ones, to the best of our knowledge cost-sensitive algorithms have yet to be used in this field. Therefore, in this paper 10-fold cross validation on cost-sensitive algorithms along with base classifiers of Naïve Bayes, Sequential Minimal Optimization (SMO), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and C4.5 were employed. As a result, the SMO algorithm has yield to very high sensitivity (97.22%) and accuracy (92.09%) rates, the likes of which have not been reported simultaneously in the existing literature.**

Keywords-component; Naïve Bayes algorithm; Data Mining; Coronary Artery Disease; C4.5 algorithm; Cost Sensitive Algorithms; Feature Extraction

I. INTRODUCTION

The morality rates from diseases are much greater than those of accidents and natural disasters. The World Health Organization estimates that 17 million deaths worldwide each year occur due to cardiovascular diseases [1]. A major type of such diseases is coronary artery disease (CAD), which is reported to account for 7 million deaths over the world per annum [1].

Mining is the extraction of knowledge from a set of data. In other words, data mining is a process that uses intelligent techniques whereby knowledge of a set of data can be extracted [2].

Angiography is the modality of choice for the diagnosis of CAD. Angiography determines the location and extent of the stenotic arteries; nevertheless, its high costs and risks for the patient have prompted researchers to seek less expensive and more effective methods with the aid of data mining. Moreover, Cost-sensitive algorithms can be of huge value in this field as misclassification of diseased or healthy patients has different costs. Pedreira *et al.* [3] using a neural network on UCI [4] datasets, attained an accuracy rate of 80% for CAD diagnosis. Das *et al* .[5] applied Neural Network on the datasets of Cleveland and reported an accuracy rate of 89.01%. Babaoglu *et al.* [6] used the Support Vector Machine (SVM) algorithm on an exercise test data and achieved an accuracy rate of 79.17%. Tsipouras *et al.* [7] used the Fuzzy Model to detect CAD. Itchhaporia *et al.* [8] drew upon the Neural Network to analyze an exercise test data for the diagnosis of CAD.

The purpose of the present study is to use MetaCost which is a cost-sensitive [9] algorithm, so as to distinguish CAD patients from healthy individuals. The Sequential Minimal Optimization (SMO) [10], Naïve Bayes [11], C4.5

978-0-7695-4925-5/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDMW.2012.29

[12], and K-nearest Neighbors (KNN) [13] algorithms are employed to analyze the Z-Alizadeh Sani dataset with no feature normalization. The performance of all mentioned algorithms was calculated using 10-fold cross validation. This dataset contains information on 303 random visitors to Shaheed Rajaei hospital in Tehran, Iran. The dataset is enriched with three created features which are extracted from the other features prior to the applying the cost-sensitive algorithms on the datasets. The effect of the created features is investigated both theoretically and practically. First, an assumption is made about the created features. Then a lemma is stated which provides a subset of sample which satisfy the assumption. Afterwards, another lemma is presented which using the assumption 1, discusses the effectiveness of the created features. In the experiments, the correctness of assumption 1 and the effectiveness of the created features are studied. As a result, high rates of both accuracy and sensitivity are obtained which, to best of our knowledge, are superior to the existing studies in this area.

The rest of this paper is organized as follows: Section 2 describes the medical dataset. The used data mining methods are presented in section 3 and section 4 discusses the reason of effectiveness of the proposed method. The methods are evaluated in section 5 and finally, section 6 concludes the paper and discusses some future research directions.

II. Used Medical Dataset

The Z-Alizadeh Sani dataset is collected from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center and contains 54 features [14]. The features along with their valid ranges are given in Tables I, II, III, and IV.

TABLE I. Demographical Features

Demographic features	**Range**
Age	30-86
Weight	48-120
Sex	Male, Female
BMI (Body Mass Index Kg/m^2)	18-41
DM (Diabetes Mellitus)	Yes, No
HTN (Hypertension)	Yes, No
Current Smoker	Yes, No
Ex Smoker	Yes, No
FH (Family History)	Yes, No
Obesity	Yes if MBI>25, No otherwise
CRF (Chronic Renal Failure)	Yes, No
CVA (*Cerebrovascular Accident)*	Yes, No
Airway Disease	Yes, No
Thyroid Disease	Yes, No
CHF (Congestive Heart Failure)	Yes, No
DLP (*Dyslipidemia*)	Yes, No

TABLE II. Symptoms and Examination Features

Symptoms and Examination features	**Range**
BP (Blood Pressure)	90-190
PR (Pulse Rate)	50-110
Edema	Yes, No
Weak peripheral pulses	Yes, No
Lung Rales	Yes, No
Systolic Murmur	Yes, No
Diastolic Murmur	Yes, No
Typical CP (Typical Chest Pain)	Yes, No
Dyspnea	Yes, No
Function Class	1, 2, 3, 4
Atypical CP	Yes, No
Non-anginal CP	Yes, No
Exertional CP (Exertional Chest Pain)	Yes, No
LowTh Ang (low Threshold angina)	Yes, No

TABLE III. ECG Features

ECG Features	**Range**
Rhythm	Sin, AF
Q Wave	Yes, No
ST Elevation	Yes, No
ST Depression	Yes, No
T inversion	Yes, No
LVH (Left Ventricular Hypertrophy)	Yes, No
Poor R Progression (Poor R Wave Progression)	Yes, No

TABLE IV. Laboratory and Echocardiography Features

Laboratory Features	***Range***
FBS (Fasting Blood Sugar)	62- 400
Cr (Creatine)	0.5- 2.2
TG (Triglyceride):	37- 1050
LDL (Low-Density Lipoprotein)	18- 232
HDL (High-Density Lipoprotein)	15- 111
BUN (Blood Urea Nitrogen)	6- 52
ESR (Erythrocyte Sedimentation Rate)	1- 90
Hb (Hemoglobin)	8.9- 17.6
K (Potassium)	3.0- 6.6
Na (Sodium)	128- 156
WBC (White Blood Cell)	3700- 18000
Lymph (Lymphocyte)	7- 60
Neut (Neutrophil)	32- 89
PLT (Platelet)	25- 742
EF (Ejection Fraction)	15- 60
Region with RWMA (Regional Wall Motion Abnormality)	0,1,2,3,4
VHD (Valvular Heart Disease)	Normal, Mild, Moderate, Severe

The details of the features of the tables and how much they influence on CAD can be found in [14]. The discretization ranges provided in Braunwald's Heart Book [1] is used, and some additional features are added to dataset and they are introduced in Index 2. Some of these categories are given in Table V.

III. Methods

In this section, the data mining algorithms used to analyze the dataset and also the definition of information gain which is needed in the next sections are described.

A. *MetaCost*

MetaCost builds a classification model using cost values from a given matrix. This operator uses a given cost matrix to compute label predictions according to classification costs. It is a wrapper method for making classifiers cost-sensitive [9]. In the experiments, RapidMiner [15] is used to apply this method and C4.5, Naïve Bayes, KNN and SMO classifiers are used as the base classifier of this method.

TABLE V. ADDITIONAL ADDED FEATURES USING CATEGORIZATION

Feature	Low	Normal	High
Cr2	Cr<0.7	0.7≤Cr≤1.5	Cr>1.5
FBS2	FBS<70	70≤FBS≤105	FBS>105
LDL2		LDL≤130	LDL>130
HDL2	HDL<35	HDL>=35	-
BUN2	BUN<7	7≤BUN≤20	BUN>20
ESR2		if male & ESR≤age/2 or if female & ESR≤age/2+5	if male & ESR>age/2 or if female & ESR>age/2+5
Hb2	if male & Hb<14 Or If female & Hb<12.5	if male & 14≤Hb≤17 or if female & 12.5≤Hb<=15	if male & Hb>17 or if female & Hb>15
K2	K<3.8	3.8≤K≤5.6	K>5.6
Na2	Na<136	136≤Na≤146	Na>146
WBC2	WBC<4000	4000≤WBC≤11000	WBC>11000
PLT2	PLT<150	150≤PLT≤450	PLT>450
EF2	EF≤50	EF>50	
Regional wall motion abnormality (RWMA)[2]	-	Region with RWMA=0	RWMA≠0
Age2[a]		if male & age≤45 or if female & age≤55	if male & age>45 or if female & age>55
BP2	BP<90	90≤BP≤140	BP>140
PR2	PR<60	60≤PR≤100	PR>100
Neut 2		Neut≤65	Neut>65
TG2		TG≤200	TG>200
Function Class2		1	2, 3, 4

[a] *Given that women under 55 years* and *men under 45 years are less affected by CAD, the range of age is partitioned at these values.*

B. *Feature Selection*

The "weights by SVM" [16] on all samples is used to select important features. First, this method is used to assign a weight to the feature and then the features are selected for the task of classification according to their weights.

The "weights by SVM" method uses the coefficients of the normal vector of a linear SVM as feature weights. In contrast to most of the SVM based operators available in RapidMiner, this one works for multiple classes, too. The attribute values, however, still have to be numerical [15].

Among many features, 34 of them with weights higher than 0.6, which means they have more effect on separation of CAD and normal patients, were selected and the algorithms were applied on them.

C. *Feature Creation*

In this part, an algorithm is presented which was introduced in our previous works [14] and creates three new features besides the existing ones: the left anterior descending artery (LAD) recognizer, left circumflex coronary artery (LCX) recognizer, and right coronary artery (RCA) recognizer. The LAD recognizer is created as follows: using train data, any feature is converted to binomial by discretizing it if numerical and merging groups of values if polynomial. The features are converted to binomial in such a way that the patients with value 1 of each feature tend to have CAD more than the patients with value 0 of the feature. Then a weight is assigned to each feature *f*. It is the fraction of number of LAD stenotic patients that their *f* value is 1, divided by all patients with *f=1* value. The *k* features with highest *w* value are selected. Then LAD recognizer is calculated as follows:

$$\text{LAD recognizer } = \sum_{i=1}^{K} w(i) f(i), \qquad (1)$$

where, *w(i)* is the weight of feature *i* and *f(i)* is the value of feature *i* for the sample:

Similarly, the LCX recognizer and RCA recognizer are calculated for each sample. CAD happens when at least one of these arteries, i.e. the LAD, LCX, or RCA, is blocked. Therefore, these three features will be expected to have a great importance in CAD diagnosis.

The effectiveness of the three new features is discussed in sections 4 and 5. However, feature creation method can only be conducted on datasets containing information about stenosis of LAD, LCX, and RCA vessels. This is a property of the introduced data set that makes the use of this method applicable.

D. *Information Gain*

Information gain of a feature can be defined as the difference of the entropy of the dataset after it is split over a value of the feature. The entropy of a dataset with *n* classes is:

$$entropy = -\sum_{c=1}^{n} p_c \log(p_c), \qquad (2)$$

where p_c is the prior probability of the class *c* in the dataset. Then, the information gain of a feature *t* is defined as:

$$IG_t = entropy(t) - \sum_{i=1}^{K} \frac{N_i}{N} entropy(i), \qquad (3)$$

where *K* is the number of sets of data produced after the dataset is split, *N* is the number of samples in the dataset, N_i is the number of samples of each of the subsets and entropy(*i*) is the entropy of the i^{th} set [17].

IV. EFFECTIVENESS OF THE FEATURE CREATION METHOD

Before applying the classification algorithms on the dataset, the dataset is enriched with the three created features, namely LAD, LCX and RCA recognizers. The new dataset may lead to higher performance than the primary dataset. It is shown in the experiments that although just a little increase in performance can be seen for some of the classifiers, the performance is increased much for some other classifiers. In this section, we seek for the reason of the effectiveness of the new three features for the classification task.

According to the definition of the created features, it is expected that the patients whose LAD, LCX or RCA vessels are stenotic have higher values of LAD, LCX or RCA recognizers, respectively. Intuitively it is because the recognizers are the weighted summations of the binomial features with two properties: a) the value 1 for each feature and vessel shows that the probability of the stenosis of the vessel may be higher than that of the value 0, and b) the features which have more effect on the stenosis of the vessels are weighted more in the summation. Therefore, the higher values of the created features show that the corresponding vessels may be stenotic with more probability in general. Therefore, they may lead the classifiers to higher accuracies in CAD diagnosis, since the stenosis of each vessel means that the patient has CAD.

The above discussion can be more clarified according to *Assumption 1*, *lemma 1* and *lemma 2* which are presented below.

Assumption 1. If the LAD, LCX or RCA vessel of a random patient is stenotic and the corresponding recognizer of the vessel has value x, then the same vessel of a patient with value $y>x$ of recognizer is stenotic with high probability. Similar arguments hold for vessels which are normal: If the LAD, LCX or RCA vessel of a random patient is normal and the corresponding recognizer of the vessel has value x, then the same vessel of a patient with value $y<x$ of recognizer is normal with high probability.

This assumption is another statement of the positive correlation between the probability of the stenosis of the vessels and their recognizers which was stated before. *Lemma 1* introduces a set of instances which satisfy this assumption according to two simple rules. Afterwards, *lemma 2* uses the assumption and deduces that the information gains of the created features are high, so they can increase the performance of the classifiers.

Before stating *lemma 1*, two simple rules are presented which can be considered valid in most circumstances. These rules are about the probability of the stenosis of the LAD, LCX or RCA vessels of two patients *p1* and *p2* with the selected features $f_{1,1}$, …, $f_{1,k}$ and $f_{2,1}$, …, $f_{2,k}$ for the corresponding vessel.

rule 1. If $f_{1,i}=f_{2,i}$ for all $1\leq i\leq k$, $i\neq j$; $f_{1,j}=1$ *and* $f_{2,j}=0$; then the probability of the stenosis of the corresponding vessel of *p1* is higher than that of *p2*.

rule 2. If $f_{1,i}=f_{2,i}$ for all $1\leq i\leq k$, $i\neq j$, $i\neq l$; $f_{1,j}=1$ *and* $f_{2,j}=0$; $f_{1,l}=0$ *and* $f_{2,l}=1$; $w_j>w_l$; then the probability of the stenosis of the corresponding vessel of *p1* is higher than that of *p2*.

These two rules are needed for validity of *lemma 1*.

lemma 1. Consider two patients *p1* and *p2* with the selected features $f_{1,1}$, …, $f_{1,m}$, $f_{1,m+1}$, …, $f_{1,n}$, $f_{1,n+1}$, …, $f_{1,p}$ and $f_{2,1}$, …, $f_{2,m}$, $f_{2,m+1}$, …, $f_{2,n}$, $f_{2,n+1}$, …, $f_{2,p}$ for the corresponding vessel, where we have:

1) $f_{1,i}=f_{2,i}$ for all $1\leq i\leq m$,
2) $f_{1,i}=0$ and $f_{2,i}=1$ for all $m+1\leq i\leq n$,
3) $f_{1,i}=1$ and $f_{2,i}=0$ for all $n+1\leq i\leq p$,
4) $(n-m)\leq(p-n)$,
5) $w_{m+i}\leq w_{n+i}$ for all $1\leq i\leq(n-m)$;

then the probability of the stenosis of the corresponding vessel of *p1* is higher than that of *p2* according to *rule 1* and *rule 2*.

proof. Applying *rule 1* on the i^{th} features where $2n-m+1\leq i\leq p$ and *rule 2* on the $(m+i)^{th}$ and $(n+i)^{th}$ features where $1\leq i\leq(n-m)$ assures that the conclusion of *lemma 1* is valid. ■

Considering that the features can be reordered before applying the rules of *lemma 1* and that the recognizer of the corresponding vessel for *p1* is higher than that of *p2* according to the conditions of the lemma, a subset of the samples which satisfy *assumption 1* can be obtained. In addition, it can be seen that the assumption is valid for a larger set of samples as it will be discussed in section 5. *lemma 2* is based on *assumption 1* and discusses why the created features can increase the performance of the classifiers.

lemma 2. If assumption 1 is considered valid, the information gain of the LAD, LCX and RCA recognizers are high in determining the stenosis of the three vessels, respectively.

proof. Consider the threshold on the recognizer which leads to the highest value of information gain for the vessel (any of the three vessels) is τ. Define the patient P_{high} as the patient with lowest recognizer value among the patients with higher recognizer values than τ. Similarly, define P_{low} as the patient with highest recognizer value among the patients with lower recognizer values than τ. According to the definition of information gain, the vessel of P_{high} must be stenotic and the vessel of P_{low} must be normal; otherwise the threshold τ can be substituted by the recognizer value of P_{high} or P_{low} to obtain a higher information gain which is a contradiction. Therefore the probability that a patient's vessel with higher recognizer value than P_{high} is stenotic will be high and also the probability that a patient's vessel with lower recognizer value than P_{low} is patent will be high, too. Therefore, the probability that a patient's vessel with higher recognizer value than τ is stenotic and the probability that a patient's vessel with lower recognizer value than τ is normal are high. Therefore, the information gain of the recognizers (for any of the vessels) is high according to the definition of information gain. ■

V. The Experimental Results

In this section, first the RapidMiner and performance measures are discussed in order to evaluate the algorithms described above and thereafter the actual results are presented. Finally the correctness of *assumption 1* which is the base of reasoning of effectiveness of the created features is investigated.

A. RapidMiner

RapidMiner [15] is a tool for experimenting with machine learning and data mining algorithms. An experiment is a set of operators that perform different tasks in the data. The experiments can be described visually as a process.

RapidMiner is an environment for machine learning and data mining process. It follows a modular operator concept which allows the design of complex nested operator chains for a huge number of learning problems. It also allows for the data handling being transparent to the operators. RapidMiner introduces new concepts of transparent data handling and process modeling, which eases process configuration for the end users. Additionally, clear interfaces and a sort of scripting language based on XML turns RapidMiner into an integrated developer environment for data mining and machine learning [15].

B. Performance measure

The accuracy, sensitivity, and specificity are of great significance in the medical field. Consequently, for measuring the performance of algorithms, accuracy, sensitivity, and specificity were used.

1) Confusion matrix

A confusion matrix contains information on actual and predicted classifications done by a classification system.

In this table:

- a_1 is the number of correct predictions for positive instances
- a_2 is the number of incorrect predictions for positive instances
- a_3 is the number of incorrect of predictions for negative instances
- a_4 is the number of correct predictions for negative instances

TABLE VI. CONFUSION MATRIX

	Actual Positive	**Actual Negative**
Predicted Positive	a_1	a_3
Predicted Negative	a_2	a_4

2) Accuracy

Accuracy is the proportion of the total number of predictions that are correct. It is determined using the following equation:

$$Accuracy = \frac{a_1 + a_4}{a_1 + a_2 + a_3 + a_4} \qquad (4)$$

3) Sensitivity and specificity

Sensitivity and specificity are the ratio of correctly diagnosed CAD and normal samples.

$$Sensitivity = \frac{a_1}{a_1 + a_2} \qquad (5)$$

$$Specificity = \frac{a_4}{a_4 + a_3} \qquad (6)$$

C. Results

In Tables VII-IX, cost matrix has been shown with different CAD and normal wrong diagnosis costs. In Table VII, wrong diagnosing cost of CAD and normal are the same. In Table VIII, wrong diagnosing cost of CAD is twice the normal case and in Table IX, it is triple.

TABLE VII. COST MATRIX WITH EQUAL COST FOR WRONG DIAGNOSING OF CAD AND NORMAL

	True CAD	**True Normal**
Predicted CAD	0	1
Predicted Normal	1	0

TABLE VIII. COST MATRIX WITH TWICE THE COST FOR WRONG DIAGNOSING OF CAD RELATIVE TO NORMAL

	True CAD	**True Normal**
Predicted CAD	0	1
Predicted Normal	2	0

TABLE IX. COST MATRIX WITH THREE TIMES THE COST FOR WRONG DIAGNOSING OF CAD RELATIVE TO NORMAL

	True CAD	**True Normal**
Predicted CAD	0	1
Predicted Normal	3	0

Among the different cost Matrices in Tables VII, VIII, and IX the highest rate of sensitivity was related to cost matrix in Table IX. Accordingly, for all the algorithms in Tables X, XI, and XII the results were calculated by using it.

Using feature selection, the following features were selected: The LAD recognizer; LCX recognizer; RCA recognizer; Typical CP; Regional with RWMA2; Age; EF2; HTN; DM; BP2; T inversion; ESR; Q wave; ST Elevation; BMI; CR2; Dyspnea; FH; Function Class; Hb; HDL; Hb2; LVH; Lymph; Na2; PR; Sex; TG2; VHD; WBC2; Airway Disease; FBS; PLT2 and LDL

For comparing SMO and SVM algorithms, SVM output is also shown in Tables X, XI, and XII. The accuracy rates of the algorithms after feature selection and without the created features are shown in Table X.

TABLE X. COMPARISON BETWEEN THE ACCURACY OF THE ALGORITHMS WITH THE SELECTED FEATURES AND WITHOUT THE CREATED FEATURES USING COST MATRIX OF TABLE IX

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**
SMO	91.43%	95.83%	80.46%
SVM	89.10%	98.15%	66.67%
C4.5	83.82%	93.98%	58.62%
Naïve Bayes	69.91%	63.89%	85.06%
KNN(k=1)	70.31%	87.50%	27.59%
KNN(K=2)	70.99%	93.06%	16.09%
KNN(K=10)	72.62%	100%	4.6%

In Table X, the best accuracy was related to the SMO algorithm (91.43%) and SVM and C4.5 algorithms were in the second and third place respectively (89.10% and 83.82%). The accuracy rates of the other algorithms were at least 10% lower than that of the C4.5 algorithm. Also, the

Naïve Bayes achieved nearly the same accuracy as the KNN algorithm.

To illustrate the impact of the new features, the accuracy rates of the algorithms after feature selection with the created features are depicted in Table XI.

TABLE XI. COMPARISON BETWEEN THE ACCURACY OF THE ALGORITHMS USING THE SELECTED AND CREATED FEATURES USING COST MATRIX OF TABLE IX

Algorithm	Accuracy	Sensitivity	Specificity
SMO	92.09%	97.22%	79.31%
SVM	89.11%	91.20%	83.91%
C4.5	83.85%	95.37%	55.17%
Naïve Bayes	80.15%	74.54%	94.25%
KNN(k=1)	74.61%	93.06%	28.74%
KNN(K=2)	74.94%	98.15%	17.24%
KNN(K=10)	72.62%	100%	4.6%

Tables X and XI show that the accuracy rates of all the algorithms increased with the new created features. Also, in these tables, the SMO algorithm had the best accuracy rate (92.09%). Since diagnosing CAD is very vital, any method which improves accuracy of algorithms even slightly is valuable. The new created features increased the accuracy of some of the classification methods and the accuracy of the others were almost the same. For example, accuracy of SMO was 0.66% increased and the accuracy of Naïve Bayes was increased about 10%. In Table XI, as opposed to Table X, the accuracies of KNN for K=1 and K=2 are higher than that of K=10 as these accuracies have improved almost 4% in Table XI.

To illustrate the impact of the new features, the accuracy rates of the algorithms only with the created features are depicted in Table XII.

TABLE XII. COMPARISON BETWEEN THE ACCURACY OF THE ALGORITHMS USING ONLY CREATED FEATURES AND COST MATRIX OF TABLE IX

Algorithm	Accuracy	Sensitivity	Specificity
SMO	87.15%	93.52%	71.26%
SVM	81.23%	98.15%	39.08%
C4.5	86.82%	96.30%	63.22%
Naïve Bayes	85.15%	87.96%	78.16%
KNN(k=1)	83.51%	92.13%	62.07%
KNN(K=2)	85.15%	96.30%	57.47%
KNN(K=10)	84.82%	96.76%	55.17%

Comparing Tables XII and X shows that accuracies of all algorithms except for SMO and SVM have increased using the three created features alone rather than all the selected important features.

In Tables XIII, XIV, XV, and XVI the results of the algorithms with the three different cost Matrices are shown. In these tables, 1-1 means Cost Matrix for equal cost for the wrong diagnosis of CAD and normal cases (Table VII); 2-1 means cost matrix for twice the cost for the wrong diagnosis of CAD (Table VIII); and 3-1 means cost matrix for three times the cost for the wrong diagnosis of CAD (Table IX). For example in Table XIII, SMO 1-1, SMO 2-1, and SMO 3-1 mean using the cost matrices from Tables VII, VIII, and IX, respectively.

TABLE XIII. COMPARING PERFORMANCE OF SMO IN DIFFERENT COST MATRICES

Algorithm	Accuracy	Sensitivity	Specificity
SMO 1-1	92.74%	96.30%	83.91%
SMO 2-1	92.42%	96.30%	82.76%
SMO 3-1	92.09%	97.22%	79.31%

As is shown in Table XIII, the SMO algorithm had the highest sensitivity (using Table IX), while the highest accuracy was related to using cost matrix in Table VII. Additionally, in all the cases, sensitivity was higher than specificity, which means that these algorithms tend to diagnose patients as class CAD. Moreover, by changing cost matrices, the SMO accuracy almost stays the same. This is due to the high value of sensitivity of this algorithm. Therefore, many false positive predictions are needed to further decrease its false negative that is avoided by the cost sensitive algorithm.

TABLE XIV. COMPARING PERFORMANCE OF NAÏVE BAYES FOR DIFFERENT COST MATRICES

Algorithm	Accuracy	Sensitivity	Specificity
Naïve Bayes 1-1	74.52%	65.28%	97.70%
Naïve Bayes 2-1	78.49%	71.30%	96.55%
Naïve Bayes 3-1	80.15%	74.54%	94.25%

As is demonstrated in Table XIV, the highest sensitivity and accuracy values of the Naïve Bayes algorithm were obtained using the cost matrix in Table IX. However, in this algorithm, unlike the others, specificity was higher than sensitivity, which means that this algorithm tends to diagnosing patients as healthy.

TABLE XV. COMPARING PERFORMANCE OF C4.5 FOR DIFFERENT COST MATRICES

Algorithm	Accuracy	Sensitivity	Specificity
C4.5 1-1	85.80%	91.20%	72.41%
C4.5 2-1	84.51%	93.52%	62.07%
C4.5 3-1	83.85%	95.37%	55.17%

Table XV illustrates that the highest sensitivity was related to the C4.5 algorithm with the cost matrix in Table IX, while the highest rate of accuracy was related to the cost matrix in Table VII. Also, in all the algorithms, sensitivity was higher than specificity, which means these algorithms tend to diagnose patients as class CAD.

TABLE XVI. COMPARING PERFORMANCE OF KNN FOR DIFFERENT COST MATRICES

Algorithm	Accuracy	Sensitivity	Specificity
KNN(K=1) 1-1	71.34%	84.26%	39.08%
KNN(K=1) 2-1	74.96%	90.74%	35.63%
KNN(K=1) 3-1	74.61%	93.06%	28.74%
KNN(K=2) 1-1	74.65%	90.74%	34.48%
KNN(K=2) 2-1	74.96%	95.37%	24.14%
KNN(K=2) 3-1	74.94%	98.15%	17.24%
KNN(K=10) 1-1	75.94%	94.91%	28.74%
KNN(K=10) 2-1	73.95%	99.54%	10.34%
KNN(K=10) 3-1	72.62%	100%	4.6%

In Table XVI, sensitivity is higher than specificity. Furthermore, when K increases, the algorithm tends to diagnose patients as CAD. For K=1 and K=2, accuracy with the cost matrix in Table VIII is higher than that in the others; whereas for K=3, accuracy with the cost matrix in Table VII is the highest.

As Tables XIII-XVI demonstrate, the sensitivity of almost all the algorithms with the cost matrix in Table IX is higher than that of the other Cost Matrices.

Thus, as is shown in Table XI, the highest performance was related to the SMO algorithm. Naïve Bayes and C4.5 nearly had the same performance.

Figure 1 shows sensitivity, specificity, and accuracy changes of SMO algorithm with respect to cost of wrong CAD diagnosis.

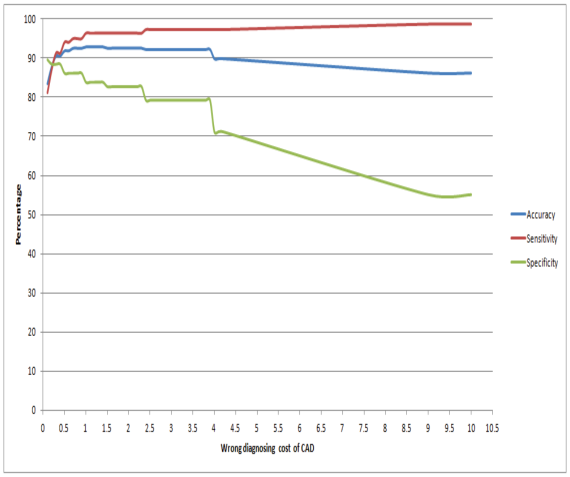


Figure 1. Comparing accuracy, sensitivity, and specificity of SMO algorithm with cost marix changes

For comparing measurement values, cost of wrong normal diagnosis is considered constant as 1 and cost of wrong CAD diagnosis has changed from 0.1 to 10. In this figure, x axis is the cost of wrong CAD diagnosis and y axis, shows sensitivity, specificity, and accuracy. As seen in the figure, the more the cost of CAD diagnosis is, the more the algorithm is probable in correct diagnosing of CAD. Therefore, sensitivity increases and specificity decreases. But accuracy is mildly constant. The highest accuracy is 92.74% which is achieved with x=1, meaning the same cost for wrong CAD and normal diagnosis. The highest specificity is 89.66% which is reached with x=0.1, and the highest sensitivity is 98.61% which is reached with x=9. In x=0.25 sensitivity, specificity, and accuracy have the same value, 88.51%.

D. *Investigation the validity of assumption 1*

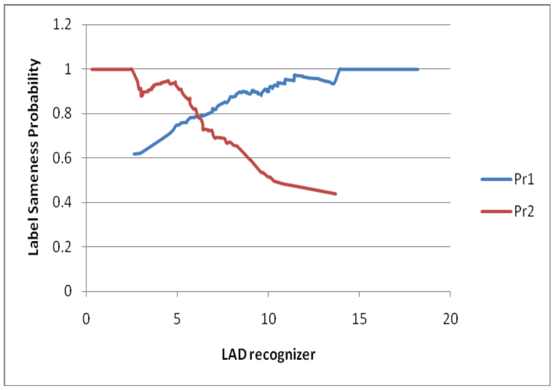
In order to check the correctness of *assumption 1*, for each vessel the probabilities discussed in this assumption were calculated, using the respective recognizer alone: 1) Pr1, i.e. the probability that a patient's vessel is stenotic given that its recognizer value is higher than that of a specific patient with stenotic vessel; 2) Pr2, i.e. the probability that a patient's vessel is normal given that its recognizer value is less than that of a specific patient with normal vessel. Parts (a) through (c) of Figure 2 show the probabilities for LAD, LCX and RCA vessels, respectively. If the vessel (LAD, LCX or RCA) of a patient is stenotic, Pr1 was calculated for him/her, otherwise Pr2 was calculated. One of the 303 records of the dataset seemed to be exception; therefore it was eliminated from the dataset prior to this experiment. These probabilities were then plotted against the value of the recognizer. As expected, probabilities Pr1 have positive correlation with the recognizer values and probabilities Pr2 have negative correlation. These probabilities are almost high, especially when the recognizer values are high or low for probabilities Pr1 and Pr2, respectively. Therefore, *assumption 1* is valid to some extent, depending on the recognizer values.

E. *Result and discussion*

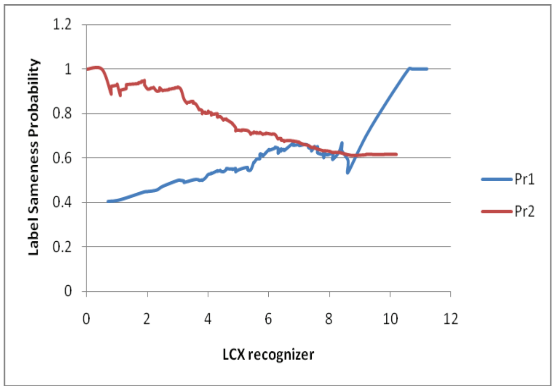
In this study, the MetaCost algorithm, which is a cost-sensitive algorithm, was used. First, from a total of 54 features, 34 were selected using feature selection algorithm. Then three created features were added to the dataset. The C4.5, KNN, Naïve Bayes, SVM and SMO algorithms were thereafter used in MetaCost. As Table XI shows, the accuracy of the SMO algorithm was better than that of the other Algorithms. The accuracy of the KNN is not as well as the other algorithms since the number of patients who have CAD is about 2.5 times more than the number of normal ones and also comparison of Euclidean distance between patients cannot accurately discriminate them so this algorithm is more likely to diagnose patients as CAD.

In order to study the cost sensitive algorithms, first the cost matrix was set with no difference between the two classes. Next, taking two times and third times the cost for the wrong CAD diagnosis, it was seen that third case leaded to the best sensitivity.

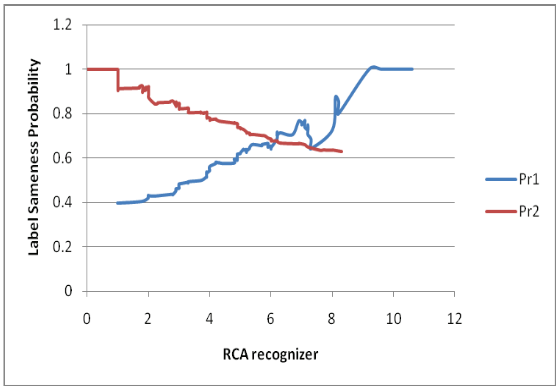
In addition, the feature creation method was investigated. This method increased the accuracy of some of the classification algorithms, substantially. The *assumption 1* was also shown to be almost valid.



(a) The probabilities stated in *assumption 1* for LAD



(b) The probabilities stated in *assumption 1* for LCX



(c) The probabilities stated in *assumption 1* for RCA

Figure 2. The probabilites discussed in *assumption 1* for (a) LAD, (b) LCX and (c) RCA vessels against the recognizer values.

VI. CONCLUSTION AND FUTURE WORK

In this paper, the MetaCost algorithm was run on the Z-Alizadeh Sani dataset to for CAD diagnosis and yielded the highest accuracy rate when employed alongside the SMO algorithm. The sensitivity was also high as the cost sensitive algorithms were applied. A feature creation method which can be used to add three new features regarding LAD, LCX and RCA vessels to the dataset was shown to be effective for the task of classification.

The future goal is to add stress heart Magnetic Resonance Imaging (MRI) and cardiac radionuclide imaging features to examine their effects on CAD. In addition, the validity of the assumption and the effectiveness of the features can be investigated furthered, both theoretically and practically. Also, the algorithms can be applied on more datasets to obtain more reliable and interesting results. More accurate comparisons could also be obtained by applying existing state of the art methods on the introduced data set. Finally, the proposed cost sensitive algorithm can be used on other diseases such as cancer.

REFERENCES

[1] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, “Braunwald’s Heart Disease: A Textbook of Cardiovascular Medicine”, 9th edition: New York, Saunders, 2012.

[2] S. Bickel and T. Scheffer, “Multi-view clustering”. In Proc. Of the IEEE Int’l Conf. on Data Mining, pp. 19–26, 2004.

[3] C. E. Pedreira, L. Macrini, and E. S. Costa, "Input and Data Selection Applied to Heart Disease Diagnosis", Proceedings of International Joint Conference on Neural Networks, IEEE, 2005.

[4] UCI KDD Archive, [online]. Available from <http://archive.ics.uci.edu/ml/> (last accessed: July 2, 2012).

[5] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, pp. 7675–7680, 2009.

[6] I. Babaoglu, O. Fındık, and M. Bayrak, "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine", Expert Systems with Applications, pp. 2182–2185, 2010.

[7] M. Tsipouras, T. Exarchos, D. Fotiadis, A. Kotsia, K.Vakalis, K. Naka, L. Michalis, ”Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling”, IEEE Transactions on information technology in biomedicine , Vol.12, NO.4, pp.447-458, 2008**.**

[8] D. Itchhaporia, R. Almassy, L. Kaufman, P. Snow, and W. Oetgen, "Artificial neural networks can predict significant coronary disease", J.Am. Coll. Cardiol., Vol.28, NO.2, pp.515-521, 1995.

[9] P. Domingos, “MetaCost: A general method for making classifiers costsensitive”, In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155-164, 1999.

[10] J. C. Platt,”sequential minimal optimization:A fast algorithm for training support vector machines”.Technical report MSR-TR-98-14, Microsoft Research, 1998.

[11] R. Caruana, and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms”, Proceedings of the 23rd international conference on Machine learning, pp. 161 – 168, 2006.

[12] J. R. Quinlan, ”Improved use of continuous attributes in c4.5”, Journal of Artificial Intelligence Research, vol.4, pp.77-90, 1996.

[13] D. T. Larose, "Discovering knowledge in data: an introduction to data mining"*,* John Wiley & Sons*,* 2005.

[14] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al. (Unpublished resutls). “A Data Mining Approach for Diagnosis of Coronary Artery Disease”, submitted to Artificial Intelligence in Medicine.

[15] http://sourceforge.net/projects/rapidminer/ (last accessed: July 10, 2012).

[16] A.Ben-Hur, and J. Weston, A User's Guide to Support Vector Machines, Methods in Molecular Biology, 2010, pp.223-239.

[17] P.N. Tan, M. Steinbach, V. Kumar, “Introduction to data mining”, Pearson Addison Wesley Boston, 2006.