Journal of
Health & Medical Informatics

Moghaddasi et al, J Health Med Informat 2017, 8:2
DOI: 10.4172/2157-7420.1000250

Research Article Open Access

Comparing the Efficiency of Artificial Neural Network and Gene Expression Programming in Predicting Coronary Artery Disease

Moghaddasi H[1]*, Mahmoudi I[2] and Sajadi S[3]

[1]Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Health Information Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Department of English Language, School of Allied Medical Sciences Shahid Beheshti University of Medical Sciences, Tehran, Iran

Abstract

Background: Angiography, as the gold standard for the diagnosis of coronary artery disease, has made an attempt to predict coronary artery disease by comparing the efficiency of gene expression programming, as a new data mining technique, and artificial neural network, as a conventional technique. Besides, the study went further to present the results of feature selection based on stepwise backward elimination, classification and regression tree.

Methods: The subjects were assessed for nine coronary artery disease risk factors to develop a prediction model for the disease. They included 13,288 patients who were chosen to undergo angiography for the diagnosis of coronary artery disease; from this sample, 4059 subjects were free from the disease while 9169 were suffering from it. Modeling was carried out based on gene expression programming and artificial neural network techniques. The Delong's test was then used to choose the final model based on the area under the Receiver Operating Characteristic (ROC) curve.

Results: The model, developed based on artificial neural network, had AUC of 0.719, accuracy of 73.39%, sensitivity of 93.44% and specificity of 28.34%. On the other hand, the model, formulated based on gene expression programming, had AUC of 0.720, accuracy of 73.94%, sensitivity of 93.29% and specificity of 31.43%. Delong's test showed no significant difference between the two models (p value=0/789). Then, feature selection method was used to choose a model with four risk factors and an accuracy rate of 73.26%.

Conclusion: Comparison of the results showed no significant difference between the two modeling techniques. The gene expression programming model was very easy to present and interpret; it could also be easily converted to other programming languages; so, with these features in mind, the researchers preferred to choose this technique.

Keywords: Coronary artery disease; Gene expression programming; Artificial neural network; Classification and Regression Tree (CART)

Introduction

Coronary artery disease is the most common cardiovascular disease [1] and the most frequent cause of death in the world [2]. In Iran it is known as the first leading cause of death [3]. The disease results from the convergence of a number of contributing risk factors [4]. Studies on different medical resources show that the risk factors for this disease mainly include smoking, hypertension, lipid disorders (high cholesterol, high triglycerides, high LDL, low HDL, diabetes, physical inactivity, obesity, abdominal obesity, age, sex, family history, alcohol consumption, psychological factors, menopause, high fasting blood glucose, fibrinogen, lipoprotein a, C Reactive Proteins (CRP) and homocysteine [4-11]. Coronary angiography is considered as a gold standard for diagnosis of Coronary Artery Disease (CAD) [11]. Angiography, however, is an expensive and invasive procedure, which is associated with some risks [6]. On the other hand, non-invasive tests might yield false negative or false positive results that could be dangerous for the patient. Hence adoption of decision support systems, along with other procedures which are done before angiography, is essential to reduce the false results [12]. Decision support systems, that can help to solve complex problems effectively and to make proper decisions [13], have been recommended by many researchers for disease detection. These systems detect patterns in medical data, improve the decision making process and, as a result, affect costs [14] while enhancing the quality of health care [15]. Decision support systems are created by a variety of data mining techniques of which Artificial Neural Network (ANN), which is inspired by biological neural networks, serves as a mathematical model in human diagnostic systems that are widely used in various fields especially medicine [16]. Among different data mining techniques, GEP is a genotype/phenotype genetic algorithm (linear and ramified) that is presented as a new technique for the creation of computer programs. Gene expression programming uses character linear chromosomes that are composed of genes structurally organized in heads and tails. The chromosomes encode expression trees which are the object of selection. The creation of these separate entities (genome and expression tree, with distinct functions) allows the algorithm to perform with such high efficiency that can greatly surpass the existing adaptive techniques [17].

Numerous studies were done to predict CAD based on data mining techniques. One study, for example, compared performances of three techniques, known as logistic regression, decision tree and neural network, to predict CAD. In this study, the multilayer perceptron neural network model, with an accuracy rate of 78.7%, was shown to be the best model [18]. In two other studies, Mobley and his colleagues created two models for CAD by using neural networks. They worked on a set of data, different in size and risk factors, to develop CAD models; they developed their own models with accuracy rates of 89% and 72%

***Corresponding author:** Moghaddasi H, Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran, Tel: 0098 2122747373; E-mail: moghaddasi@sbmu.ac.ir

Received February 15, 2017; **Accepted** March 26, 2017; **Published** April 1 , 2017

Citation: Moghaddasi H, Mahmoudi I, Sajadi S (2017) Comparing the Efficiency of Artificial Neural Network and Gene Expression Programming in Predicting Coronary Artery Disease. J Health Med Informat 8: 250. doi: 10.4172/2157-7420.1000250

[12,19]. Some other studies used the data already stored in repository machines at the University of California, Irvine [20]. The results varied depending on the type of data mining techniques used in these studies [21-23]. According to what was mentioned before, the risks involved in invasive diagnostic procedures, like angiography, have to be dealt with properly. One way of overcoming such risks could be data mining techniques with their promising outcomes. In this study, the results obtained from the comparison of GEP will be presented as a new data mining technique; the ANN will then be introduced as a

Conventional technique; in the end, a diagnostic model to predict CAD will be followed.

Methods

To obtain a prediction model for CAD, the angiography database of Tehran Heart Center, with 13,228 records, was used. The database included nine risk factors known as age, sex, obesity, abdominal obesity, family history, smoking, high cholesterol, diabetes and hypertension. Descriptive statistics for this database appear in Table 1. To avoid over-fitting and to evaluate generalizability power of the model, the data set was classified into two subsets of training (70%) and validation (30%) [24]. Then, modeling was done by using GENEXPRO and MATLAB applications, based on GEP and ANN. The steps involved in GEP were as follows: First, an initial population of chromosomes (solutions) was randomly generated. Then each chromosome was expressed and its fitness value was calculated. It is worth mentioning that one of popular Fitness functions is "Hits with penalty" that acts based on the number of samples that is to be properly classified and penalties considered for models that have True Positive (TP) or True Negative (TN) with values equal to zero, but their total number of success is high. When in a generation a model is obtained that has higher accuracy than the models produced in previous generations, that model would survive. If the termination condition of the algorithm (e.g. achieving the greatest fitness) was fulfilled, the best solution, among the existing options, would be selected and the algorithm would then be terminated; otherwise, the procedure would continue by producing another generation of solutions [25]. For a GEP-based modeling, some setting initials are necessary, as can be seen in Table 2. Modeling based on ANN was done using a Multilayer Perceptron (MLP) neural network. Also, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, developed based on a quasi-Newton algorithm, was used for learning the network. This learning algorithm has a faster convergence rate than the gradient descend and the conjugate gradient algorithms and is one of the appropriate learning algorithms [26]. Since there is no equation for estimating parameters such as the number of neurons in the hidden layer, the layer activation function and error function of a neural network model could be adopted. So, with this point in mind, we created 100 neural network models by randomly selecting the parameter value, as can be seen in Table 3. The area under the ROC curve, in the current work, was used to compare the efficiency of models. This method has been widely used in recent years to evaluate machine learning algorithms [27]; this method has also been used in the field of medicine, as an effective method, to evaluate the performance of diagnostic tests against the gold standard [28]. Following the modeling procedure based on ANN and GEP, the model obtained from each technique was compared against the AUC value in an attempt to select the best models and techniques. Based on the feature selection technique, the intended variables were obtained by removing extraneous and irrelevant variables [29,30]. In line with this procedure, the stepwise backward elimination method was adopted to compare the results of ANN and GEP and to select the best possible model and technique. As such, the least important risk factors were also removed and the modeling process was carried out with the remaining risk factors. This process continued until there was no significant change in the accuracy of the model in the following steps. The Classification and Regression Tree (CART) were adopted to determine the importance of the variable. Upon the completion of the modeling procedure, the best models from different modeling stages were compared and the final model was selected.

Results and Discussion

In the first stage of modeling, which considered all the relevant risk factors, modeling was done based on GEP and ANN. In GEP a total of

Dependent Variable	**Domain**	**Operational definition**		
Coronary artery disease	0.1	non CAD (0), CAD (1)		
Independent Variable	**Domain**	**Operational definition**	**CAD (69%)**	**Non CAD (31%) P value**
Age	18-100	Age	61.32±10.52	56.28±11.26 <0.001
Sex	0.1	M (0) F (1)	M (69%)	M (45%) <0.001
Family History	0.1	No (0), Yes (1)	No (82%)	No (85%) <0.002
Cigarette Smoking	0.1	No (0), Yes (1), Withdrawal (2)	No (60%), Yes (24%)	No (76%), Yes (14%) <0.001
Hyperlipidemia	0.1	No (0), Yes (1)	No (33%)	No (41%) <0.001
Hypertension	0.1	No (0), Yes (1)	No (41%)	No (49%) <0.001
Diabetes Mellitus	0.1	No (0), Yes (1)	No (64%)	No (79%) <0.001
Abdominal Obesity	0.1	No (0), Yes (1)	No (45%)	No (31%) <0.001
BMI	0.1	No (0), Yes (1)	No (26%)	No (22%) <0.001

Table 1: Descriptive statistics of the data set under study.

52 models were produced; these models were then evaluated to select the best model with AUC of 0.72. Also, in ANN a total of 100 models were produced using different parameter values; these models were then evaluated to select the best model with AUC of 0.719. The results of modeling based on ANN and GEP techniques appear in Table 4. To compare the models based on GEP and ANN, Delong's test was used. The test is to know whether or not there is any significant difference between various levels of AUC [31]. In this study, comparison of the AUC levels of GEP and ANN models, using Delong's test, shows no significant difference (p=0.789). However, the ANN model cannot be presented and interpreted in great detail as it is composed of a black box. Nonetheless, because of their unique nature (i.e., expression trees), the GEP-based models can easily be presented, interpreted or converted to other programming languages; so with these features in mind, the current study preferred to choose the GEP technique. As mentioned in the method section, the current study adopted a feature selection procedure to achieve a simple model. In line with this procedure, with the help of CART technique, the risk factors were sequenced in order of importance as follows: age (100%), diabetes (86%), hypertension (52%), sex (49%), high cholesterol (37%), consumer smoking (36%), obesity (17%), and family history (13%). In the second stage of modeling, family history as the least important risk factor was removed; the modeling was then repeated with the remaining risk factors. Twenty-Four models were then generated; after evaluation of these models, the best one for AUC, with an area under the curve of 0.700, was selected. There was a little difference between the best model in the first and second stage of modeling. So the modeling process continued until the researchers were left with just few models in the third time of modeling. After evaluation of these models, the best one with AUC of 0.677 was selected. This value is slightly different from that of the previous stages of modeling. As in previous stages, the least important risk factors were further removed and the modeling process continued with the remaining risk factors. At the seventh stage of modeling, the area under the curve obtained for the best model of AUC was significantly different from that obtained for the AUC at the sixth stage of modeling. This means that the risk factors available at stage seven were no longer sufficient for further modeling; so owing to insufficient risk factors the modeling process was abandoned at stage seven. After seven stages of modeling, the models obtained at the first and sixth stages were shown to be the best models with some salient features, as can be seen in Table 5. As shown in Table 5, the model produced at the first stage was the best model in terms of accuracy and area under ROC curve; this model was therefore considered as the selected model. However, the model created at the sixth stage of modeling has few negligible differences with the selected model as it is composed of only four risk factors, making it simpler than the selected one. In the following steps, in order to obtain a simpler model with greater accuracy, the lastly selected model was considered as an input for gene expression programming algorithm, resulting in the shortening of model sizes from 33 to 25, while there was no change in accuracy. The final model, shown as a tree diagram in Table 1, can be easily converted into any programming language. A noticeable point, following its modification, is that, in addition to getting shorter in size, the model does not include hypertension as a risk factor. As such, the final model is composed of eight risk factors known as age, sex, obesity, abdominal obesity, family history, smoking, hyperlipidemia, and diabetes. The ROC curve of this model is comparable to Figures 1 and 2.

Setting	Function Set
Chromosomes:30	-
Genes:3	+ , - , * , /
Linking Function: Addition	Sqrt, Ln, Exp
Mutation Rate:0.044	Sin, Cos, Tan
Inversion Rate:0.1	Asin, Acos, Atan
IS Transposition Rate:0.1	Not, OR, AND
RIS Transposition Rate:0.1	-
Gene Transposition Rate:0.1	X2: (x^2)
One-Point Recombination Rate:0.3	GOE2C:
Two-Point Recombination Rate:0.3	If x ≥ y, Then (x+y), Else (x-y)
Gene Recombination Rate:0.3	-
Fitness Function: Hits with Penalty	L2TB
0/1 Rounding Threshold:0.5	If x<y, Then 1, Else 0
Generation number:30000	-

Table 2: Initial setting for modeling base on GEP.

Hidden layer activation function	Identity, Logistic, Tanh, Exponential
Output layer activation function	Identity, Logistic, Tanh, Exponential Softmax
Error function	Sum of squares, Cross entropy
Number of neurons in the hidden layer	Min: 3; Max: 13

Table 3: Parameter value for creating a model based on ANN.

Techniques	Data set	Accuracy	AUC	Specificity	Sensitivity
ANN	Learn	75.58%	0.717	33.65%	93.83%
	Test	73.39%	0.719	28.34%	93.44%
GEP	Learn	75.03%	0.728	32.28%	93.74%
	Test	73.94%	0.720	31.43%	93.29%

Table 4: Results of modeling based on ANN and GEP.

A number of studies have produced certain models with high accuracy rates, using data sets available in Machine Learning Repository of University of California. The reason for the production of these models, in addition to CAD-related risk factors, is the performance of physical examinations, and electrocardiography and stress imaging tests. This shows that diagnostic tests carried out before angiography could be very effective in developing models with high rates of accuracy. The prediction models developed in the current research and other similar studies, based on the CAD risk factors, tend to vary in accuracy rates [12,18,32]. Limited access to sufficient risk factors in some of these studies, including the current work, has resulted in a low accuracy rate. Another important issue, in the current research and other similar studies, is that the specificity of the models is less than their sensitivity; that is, the models are more successful in diagnosing patients than healthy individuals. The reason could be that

Stage of modeling	Risk factors	Data set	Accuracy	AUC	Specificity	Sensitivity
First	Age, Sex, BMI, Abdominal obesity, Family history, Smoking, Hyperlipidemia, Diabetes, Hypertension	Learn	75.03%	0.728	32.28%	93.74%
		Test	73.94%	0.720	31.43%	93.29%
Sixth	Age, Sex, Diabetes, Hypertension	Learn	74.06%	0.704	29.56%	93.53%
		Test	73.26%	0.700	30.22%	92.85%

Table 5: The results of prediction model for coronary artery disease based on gene expression programming.

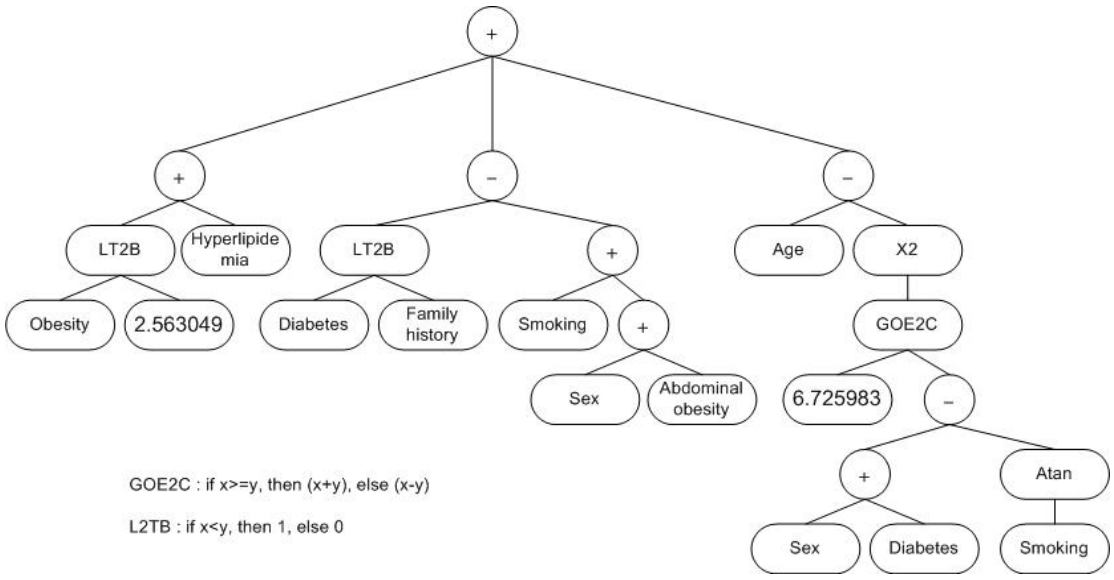

Figure 1: Prediction model for coronary artery disease based on gene expression programming.

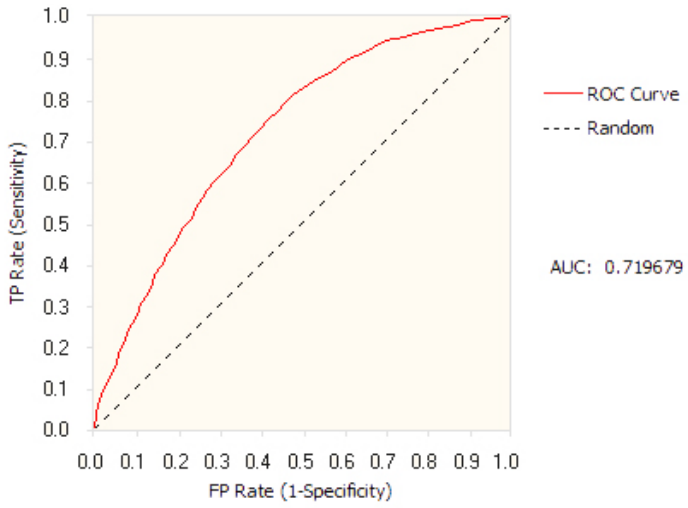

Figure 2: ROC curves of the selected model for predicting coronary artery disease.

these studies have not used suitable risk factors in sufficient numbers. Also, in the present research, feature selection led to the production of a model with four risk factors of age, sex, diabetes and high blood pressure at an accuracy rate of 73.26%, which slightly differs from the final model with an accuracy of 73.94%. According to the researchers' reviews, in some studies, the ROC curve analysis was the main measure to evaluate the proposed models while in others the model's accuracy in relation to the sum of the data was the prime evaluation criterion. Bearing in mind that a model's accuracy alone is not a suitable criterion for its evaluation, the current study used the ROC curve analysis, as the best criteria for evaluating and generating the intended models.

Conclusion

Comparison of the results of ANN and GEP showed no significant difference between the two models although the latter (i.e., GEP), was easier to present or interpret and more convenient to be converted into a programming language. So the model obtained for coronary artery disease in this study was create during gene expression programming technique; the model includes different risk factors such as age, sex, obesity, abdominal obesity, family history, smoking, high blood fats and diabetes. The model enjoys an accuracy of 73.94%, specificity of 31.43% and sensitivity of 93.29%. The study's limitation in getting access to suitable risk factors in sufficient numbers has possibly affected the model's accuracy. Some research studies have managed to produce certain models with high accuracy rates by investigating a number of factors such as physical examination, electrocardiography, imaging and stress tests together with risk factors from coronary artery disease. This shows that diagnostic tests before angiography could be very effective in obtaining more accurate models. The current research used the classification and regression tree technique, and the stepwise backward elimination method for feature selection, resulting in the production of a model with four risk factors of age, sex, diabetes and hypertension with 73.26% accuracy rate, which was slightly different from the final model. The model presented in this study selected 390

subjects, out of a total of 1242, who were free of CAD; the model, however, failed to diagnose 183 patients, out of a total of 2727, suffering from coronary artery disease. This indicates that such models are on their way to develop and improve further; then they will be able to make a better distinction between patients and non-patients. Given the importance of parametric methods, like Logistic Regression Analysis, and development of ensemble methods, the authors recommend new comparative studies in line with the objectives of the current work.

Authorship

Issa Mahmoud collected the data, carried out statistical analyses and interpreted the data. Hamid Moghaddasi proposed the topic, designed the study, and formulated the research problem. Samad Sajjadi revised the article critically and provided numerous insightful comments.

Funding

The researcher on this manuscript received no particular funds from any particular organization or research body.

Conflict of Interest

There are no conflicts of interest.

Ethical Approval

This article did not need any ethical approval as it did not deal with human participants or animals.

References

1. Mendis S, Puska P, Norrving B (2011) Global Atlas on Cardiovascular Disease Prevention and Control, World Health Organization, Geneva.
2. World Health Organization (WHO) (2011) Fact sheet No. 310: The top ten causes of death.
3. Amani F, Kazemnejad A, Habibi R, Hajizadeh E (2011) Pattern of mortality trend in Iran during 1970-2009. J Gorgan Uni Med Sci 12: 85-90.
4. Bonow RO, Mann DL, Zipes DP, Libby P (2012) Braunwald's Heart Disease A Textbook of cardiovascular Medicine. Volume 1, Saunders, Philadelphia.
5. Cecil RL, Goldman ML, Schafer AI (2012) Goldman's Cecil medicine, Elsevier/ Saunders, Philadelphia.
6. Braunwald E, Kasper D, Hauser S, Longo D, Jameson J, et al. (2008) Harrison's Principles of Internal Medicine (17th edn.). McGraw-Hill Professional.
7. American Heart Association (AHA) (2010) Heart disease & stroke statistics: 2010 update.
8. Factors CER, Erqou S, Kaptoge S, Perry PL, Angelantonio ED, et al. (2009) Lipoprotein (a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. J Am Med Assoc 302: 412-23.
9. Kaptoge S, Angelantonio ED, Lowe G, Pepys MB, Thompson SG, et al. (2010) C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet 375: 132-140.
10. Humphrey LL, Fu R, Rogers K, Freeman M, Helfand M (2008) Homocysteine level and coronary heart disease incidence: a systematic review and meta-analysis. Mayo Clinic 83: 1203-1212.
11. Crawford MH (2009) Current Diagnosis & Treatment in Cardiology, Third Edition, McGraw-Hill Medical, NY, USA.
12. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE (2000) Predictions of coronary artery stenosis by artificial neural network. Artif Intell Eng 18: 187-203.
13. Shim JP, Warkentin M, Courtney JF, Power DJ, Sharda R, et al. (2002) Past, present, and future of decision support technology. Decis Support Syst 33: 111-126.
14. Koh HC, Tan G (2005) Data mining applications in healthcare. J Health Inf Manag 19: 65.
15. Chae YM, Kim HS, Tark KC, Park HJ, Ho SH (2003) Analysis of healthcare quality indicator using data mining and decision support system. Expert Syst Appl 24: 167-172.
16. Fausett LV (1994) Fundamentals of Neural Networks: Architectures, Algorithms, and Applications, Prentice-Hall.
17. Ferreira C (2001) Gene expression programming: A new adaptive algorithm for solving problems. Compl Syst 13: 87-129.
18. Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 34: 366-374.
19. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE, et al. (2005) Neural network predictions of significant coronary artery stenosis in men. Artif Intell Eng 34: 151-61.
20. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
21. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 36: 7675-7680.
22. Khatibi V, Montazer GA (2010) A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Syst Appl 37: 8536-8542.
23. Muthukaruppan S, Er MJ (2012) A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Syst Appl 39: 11657-11665.
24. Larose DT (2006) Data mining methods and models. Wiley Online Library.
25. Ferreira C (2006) Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence (Studies in Computational Intelligence), Springer-Verlag Inc, NY, USA.
26. Bishop CM, Hinton G (1998) Neural Networks for Pattern Recognition. Clarendon Press.
27. Fawcett T (2006) An introduction to ROC analysis, Pattern. Recogn Lett 27: 861-874.
28. Kumar R, Indrayan A (2011) Receiver operating characteristic (ROC) curve for medical researchers. Ind Pediatr 48: 277-287.
29. Han J, Kamber M, Pei J (2006) Data Mining Concepts and Techniques (2nd edn.), Morgan Kaufman.
30. Tan PN, Steinbach M, Kumar V (2006) Introduction to Data Mining, Pearson Addison Wesley.
31. Delong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837-845.
32. Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV (2008) Automated diagnosis of coronary artery disease based on data mining and fuzzy modelling. Trans Inf Technol Biomed 12: 447-458.