Asian Journal of Information Technology 15 (14): 2509-2515, 2016
ISSN: 1682-3915

Identification of Factors Affecting Ischemic Heart Disease Using Data Mining Algorithms

[1]Fatemeh Rangraz Jeddi, [2]Majid Nikougoftar Nateghm,
[3]Gholamabbass Mosavi and [1]Zahra Shahabinia
[1]Health Information Management Research Center,
Kashan University of Medical Sciences, Kashan, Iran
[2]Department of Information Technology Engineering,
University of Qom, Qom, Iran
[3]Department of Biostatistics and Public Health, Faculty of Health,
Kashan University of Medical Sciences, Kashan, I.R. Iran

Abstract: Cardiovascular disease is now a major cause of mortality throughout the world. Ischemic heart disease with an increasing rate and increased mortality is considered as one of the most expensive and controversial topics in the field of healthcare in the country. Several factors are considered as risk factors. This study aimed to assess factors associated with ischemic heart disease using data mining in 2016. This prospective study of the diagnostic value was carried out in Kashan's Shahid Beheshti hospital in 2015-16. The 345 cardiac patients admitted to Kashan's Shahid Beheshti hospital were selected based on purposive sampling. The data collection tool was a valid and reliable researcher-made questionnaire and checklist. The questionnaire included questions related to lifestyle and eating habits and checklist included questions regarding to the physiological and laboratory parameters. Patient information was collected through interviews. The data was analyzed then using rapidminer data mining Software, Version 5. The results showed that the accuracy of the data mining algorithm in both decision trees and support vector machine were 87.16 and 97.25%, respectively. All algorithms were able to predict factors in this disease with various degrees of accuracy. To examine factors associated with ischemic heart disease, SVM classification model had the least amount of errors and highest accuracy compared to other models.

Key words: Data mining, risk factors for cardiovascular disease, decision tree, support, vector

INTRODUCTION

The heart is one of the vital organs of the body that its decreased oxygen even for short periods of ischemia causes tissue damage and dysfunction and damage to humans (Alizadehsani *et al*., 2013. Now cardiovascular disease and especially coronary heart disease or in medical term ischemic heart disease account for 30% of fatal diseases around the world. In developing countries, due to lifestyle, work culture and eating habits, the disease has been reported as the first and most important cause of death (Han *et al*., 2006; Heidari *et al*., 2012; Higgins, 1988). Unfortunately, 25% of patients with coronary artery disease die suddenly without any previous symptoms (Higgins, 1988). Therefore, early diagnosis, especially in the early stages of life leads to early treatment and reduces the severity of symptoms (Kajabadi *et al*., 2009; Kudyba and Gregorio, 2010). Cardiac diagnosis is determined based on clinical and pathological data set (Kajabadi *et al*., 2009). In most cases, there is a direct relationship between coronary artery disease and the number and severity of risk factors of atherosclerosis as independent risk and changeable factors such as hypertension, diabetes, smoking cigarettes, unchangeable factors like age and sex (Lakshmi and Kumar, 2013). In general, it can be said that coronary artery disease is the result of convergence of a number of risk factors (Little, 1998). Thus, considering the prevalence of ischemic heart disease, identification of risk factors will increase the efficiency of the process of diagnosis and treatment (Kudyba and Gregorio, 2010). Extraction of a set of risk factors that can predict more precise and accurate results and indicating the relationship between factors has increased the need to use new techniques because in spite of abundant data, the knowledge gained from this data with new technique

Corresponding Author: Zahra Shahabinia, Health Information Management Research Center, Kashan University of Medical Sceiences, Kashan, Iran

was poor. However; the volume of medical data is increasing day by day and physicians usually gain precious data on diseases and their relationship with each other and causes of diseases (Longo *et al*., 2012). The raw data is not valuable by itself and must be analyzed and turned into useful knowledge (Mahmoudi *et al*., 2013).

Given the prevalence of cardiovascular disease worldwide and the need to use new methods in medical research and besides, since data mining technology provides a new approach to determine medical and hidden new patterns from the distributed and heterogeneous raw data, its advanced techniques may be used for discovering useful knowledge from databases as well for medical research and specifically heart disease (Mahmoudi *et al*., 2013). So it can be used to uncover risk factors for heart disease, particularly that the diagnosis and treatment of heart disease require angiography for diagnosis of coronary artery atherosclerosis and taking the necessary measures to treatment (Mann *et al*., 2014) which in addition to imposing high costs have several side effects. Therefore, using data mining techniques for the detection of coronary disease are taken into consideration (Kudyba and Gregorio, 2010; Mann *et al*., 2014; Mehdi *et al*., 2011; Mirershadi *et al*., 2010). The aim of this study was to investigate the factors associated with the use of data mining method in ischemic heart diseasethrough classification and estimation techniques including data mining algorithm of C4.5 decision tree and support vector machine. We hope that the result of the study helps anticipation and early detection of ischemic heart disease and increased survival.

MATERIALS AND METHODS

Data source: This study applied a descriptive method and was conducted through a diagnosis (diagnostic) and prospective procedure in Kashan's Shahid Beheshti hospital. The study population consisted of cardiac patients admitted to Kashan's Shahid Beheshti. To collect data, single purposive random sampling was employed. To collect data, the researcher accompanied the patient from the beginning of hospitalization and after the arrival of patient in the intensive care unit; the behavioral factors questionnaire was completed by interview. Physiological and other factors related data were extracted from medical records. Failure to complete the file, the information was collected from doctors, nurses and lab. The instrument for collecting data and physiological parameters was non-modifiable checklist that was developed based on the determined features. Behavioral data was collected by researcher made questionnaire and was developed by refereeing to the questionnaires in (reference) and other relevant studies. Face and content validity of the questionnaire was obtained by professional experts, including specialists in cardiology, nurses and training, dietitians, clinical psychologists, respectively. A sample of 345 patients referred for coronary angiography was considered.

Pre-processing data

Primary data collection: Based on non-systematic review of references, the required features, clinical guidelines for heart disease and valid papers were developed and confirmed by comments of two cardiologists. Physiological parameters and non-modifiable factors with data extracted from medical records and behavioral factors and also data obtained through interviews of researcher with patients were collected.

Data preparation: Preparation is one of the most important and time consuming stages in data mining. It includes all activities that require the final database to build the raw data.

Data selection: Based on primary data collected in the previous section, we selected the data set being suitable for data mining purposes.

Data cleaning: Involves closer look at the data used for analysis. At this stage, data was cleaned, for example for missed values, removed rows or special features or replacements are carried out with an estimated value.

Regarding data errors, logic was used for detecting and removing features. Programming items were carried out by a programming style and then the values were converted or replaced. Regarding to the missing and bad data, fields were manually checked and replaced them with correct values, if possible (Jeddi *et al*., 2013).

Data mining algorithms: Finding undetected information and useful patterns in a database is called data mining. Data mining is widely used in the health sector and medical applications such as patient's prognosis and management. Some of the most common prediction methods in data mining are defined as follows.

Decision tree: Decision tree classifies samples in such a way that they grow from the roots to downwards and finally reaches the leaf nodes. Each interior node or leaf is determined with a characteristic. This question is posed in relation to such input. In each internal node, there are branches based on the number of possible answers to this question and each are identified with its answer. The leaves of this tree are determined with a class or a bunch

of answers (Jeddi and Rezaiimofrad, 2013). Because nodes and branches are organized hierarchically, they are easy to understand and interpret. They are reliable and have better accuracy in clinical decision making (Jeddi *et al*., 2016).

C4.5 algorithm: The algorithm is a generalization of ID3 using Gain Ratio criteria for selection of especial traits. The algorithm is stopped as the number of samples is less than the specified amount. It utilizes post-pruning technique and also accepts the numerical data (Salarifar and Kazemeini, 2008).

Support vector machine: SVMs are generally used for issues in which there are two categories, yet different methods were proposed for multi-class classification algorithm. In this two-plain classification algorithm, two classes of data are located on the border and the problem is finding the maximum between these two plains and finally between the two sets of data. This means that two data are far from each other clashing data. As shown in the figure, the goal is to find two plains with maximum distance and thus the plain will be split. SVM generalization is of high accuracy (Sitar-Taut and Sitar-Taut, 2010).

Artificial network: One of the simplest and yet the most efficient layout proposals for use in modeling real nerve is MLP (Multi-layer perceptron) consisted of an input layer, one or more hidden layers and an output layer (Sonawane *et al*., 2013). In this structure, all neurons in a layer are connected to all next neurons. This arrangement forms a network with full connectivity. The following figure shows the scheme of a 3-layer perceptron network. Simply it can be concluded that the number of neurons in each layer is independent of the number of neurons in other layers. It is worth noting that in the figure, each aggregated circle is the product of summation operation and thresholding (crossing the sigmoid nonlinear function). In fact, any solid circle in Figure is a model of collector and thresh holding block shown in the figure for facilitation of optimal viewing. According to the figure, the output neuron i M (the last layer) can be shown as follows (Soni *et al*., 2011).

Implementation and evaluation criteria: The most important criteria used in the medical field include sensitivity, accuracy and features. In this thesis, the accuracy, sensitivity and specificity criteria were used for algorithm evaluation.

For pre-processing database, a particular algorithm was used to select the variables and then applying this algorithm, the data of patients who lacked sufficient

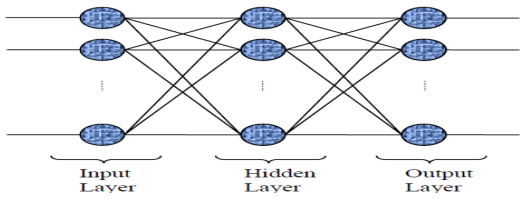

Fig. 1: Output reuron layers

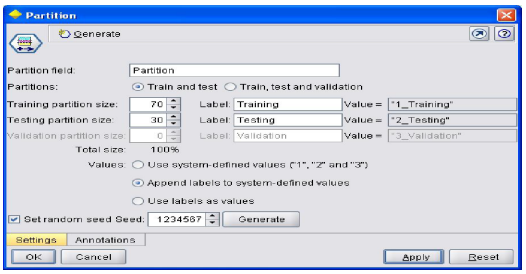

Fig. 2: Percentage of training and testing data

variables were identified and deleted from the data base. Concerning to the eliminating variables, variables that had overlapping research results were eliminated or in existing records, patient information was very limited. After removing the variables listed in accordance with two modes (overlapping with other variables or lack of maximum data records), patient's records were analyzed. Patient's records with limited variables were deleted. The remaining records with less missed variables were replaced through the expected value maximization.

Table 1 show forecast index for disease modeling. To test the validity of this study, the data was divided into two sets of test data and training data which its final outputs were assessed and validity was confirmed. In general, theoretical basis of this study included similar experiences also supports the validity of the data collection tool. Consequently, there is no significant difference between final output among test data and its training data. Data was partitioned into 70-30% for training and test data respectively. It means that 70% of the total number of sets belonged to the training data and the remaining 30% were considered as test data. The figure below shows both training and testing data. Also, the percentage of training and test data was shown in Figure. Thus training data was divided into two parts: the first is built on the basis of the model and the second models are used to evaluate the model. The second group of data by category is determined by the data model and is compared with the categories specified by the model. Then the accuracy of the model is extracted. In fact, a well-known label of test sample is compared with the results of classification (Fig. 1 and 2).

Table 1: Confusion matrixs

	Real class	
Predicted class	C1	C2
C1	True positive	False positive
C2	False negative	True negative

The accuracy of the model refers to the percentage of times that the test samples are packed with success. If the model was acceptable for data classification, it can be used for packaging of unspecified category. Training and testing data is divided into 10 parts and is divided in 10 different experiments. This data is then divided in two classes, disease and lack of disease. This data is then divided in two classes, disease and lack of disease. The relationship between real and predicted classes is called confusion matrix. Disorder matrix or complexity matrix show the efficiency of algorithms. For a two-class, problem with the class matrix of two columns and rows shows the number of False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN) (Table1). Real positive is the number of the samples properly diagnosed. False negative is the number of C2 incorrectly classified and false positive is the number of C2 incorrectly classified:

- Accuracy: the number of samples that are correctly diagnosed, relative to the total sample
- Sensitivity: the probability of correct disease prediction by algorithms (true positives divided by the true positive+false negative)
- Feature: the probability of correct disease prediction by algorithms (false positive + true negative divided by the actual negative):

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

Where:

TP = The number of positive samples correctly diagnosed

TN = The number of negative samples correctly diagnosed

FP = The number of positive samples incorrectly diagnosed

FN = The number of negative samples incorrectly diagnosed

RESULTS AND DISCUSSION

In this study, three methods of data mining models based on accuracy were compared and the ultimate goal was achieving a model with the highest accuracy. After removing records containing missing values, 345 records were remained. To find the best forecasting performance, data mining classification techniques including decision trees, neural networks and support vector machine were used and several parameters were selected randomly. Results of accuracy of models show the numerical values obtained from calculations by rapidminer Software and its final analysis. In similar studies, the accuracy of SVM model was higher than the other models. This means that this model is more accurate than the other models as well as in determining the variables affecting disease, certain variables were consistent with real-world variables. As our research is reflected in the table _ SVM algorithm has the highest accuracy rate (97.25%) among other classification algorithms in this study. Table shows the degree of accuracy, sensitivity and specificity for classifying three different ways. In connection with support vector machine algorithm, it should be mentioned that a classification algorithm has been of interest to many researchers. This algorithm has high efficiency in two class's data classification and can correctly classify records with acceptable accuracy. SVM algorithm is very similar to neural networks, yet its potential is much higher than in these networks. For example, SVM with sigmoid kernel function like MLP neural network consists of two hidden layers. This algorithm will find a general solution for every problem while neural networks for local answer (Tohidi *et al*., 2012). In this study, the C5 algorithm was used to construct a decision tree and rule set. Decision tree model appears as a series of if-then rules using C5 algorithm which shows information in a form acceptable to details and completeness. The speed of construction of this model is also high due to the use of parallel processing. Parameters for these models were set 10-fold cross-validation and 20 trials use boosting. Boosting is a method to improve prediction accuracy rate. This method operates by building multiple models in succession. The first model should be made in the usual way, then the second model focuses on the records that have to be made by the first unspecified model. The third model is the second model to be built with a focus on errors and so on. Finally the records are classified based on models and are combined using weighted vote process for single model predictions. Boosting can significantly improve accuracy in C5 model but it requires longer training. Using the cross-validation option, algorithm employs part of the training data to estimate model accurately. In this study, depression variable has been chosen as the starting point

Table 2: Comparison of the results of three data mining models

Variables	Accuracy	Sensitivity	Feature
The decision tree (C5)	87.16	0.10	0.90
Artificial neural network	95.42	0.94	1.00
Support vector machine	97.25	0.96	1.00

Table3: Confusion matrix for different algorithms by selecting and creating features

Algorithm	Variables	Really normal	Sick really
C5	Patient prediction	233	14
	Normal	3	95
	Patient prediction	236	3
SVM	Normal prediction	0	106
	Patient prediction	234	5
Neural network	Normal prediction	2	104

for failure or tree split which somehow represents one of the most important factors in the diagnosis of coronary artery disease. Features like a previous history of heart disease, respiratory problems and gastrointestinal disease, other heart disease like volvuli and diabetes are considered risk factors of coronary heart disease. These extracted rules show that simultaneously cardiovascular risk factors will develop coronary heart disease in individuals.

The performance of three classification methods was studied in terms of accuracy, sensitivity (the proportion of people who have the disease and properly detected), features (the proportion of those who do not have the disease properly diagnosed). As seen in the table, the comparison of performance evaluation of categorization model shows that almost all the algorithms produce high specificity (>90%), respectively. In addition, the highest sensitivity in our tests was obtained for SVM with the 96%. Confusion matrix for classification algorithm of features has been shown in the table. The table shows that the algorithm SMO detected 236 samples correctly diagnosed as CAD and 106 samples as correctly normal, 3 samples were diagnosed normal and samples were detected false CAD, so they show that the greatest number of CAD detect support vector machine algorithm with correctly RBF kernel detection and it correctly detects the same algorithm normal (Table 2 and 3) .

Common factors affecting patients in all algorithms in this study included: chest pain, level of education in people, feelings of depression, age, family history of cardiovascular disease, body mass index, comorbidities, the use of fiber in daily dishes, methods of cooking and smoking.

The most important factors to determine the physiological, behavioral and non-modifiable associated with ischemic heart disease based on the C5 algorithm of decision tree models in the table including feeling depressed, level of education, fruit and vegetable intake and family history of heart disease are considered the most important risk factors for coronary disease (Table 4).

Table 4: The most important physiological and non-modifiable behavior factors associated with ischemic heart disease based on C4.5 algorithm from decision tree mode

Factor type	Importance weight	Feature
Behavior and lifestyle habits	0.23	Feeling depressed
Behavior and lifestyle habits	0.20	Level of Education
Behavior and lifestyle habits	0.19	fruits and vegetables
Demographic features	0.12	Family history of heart disease
Behavior and lifestyle habits	0.11	Salt consumption
Details examinations andlaboratory factors	0.07	Hypertension
Details examinations andlaboratory factors	0.06	Fasting blood glucose
Demographic features	0.02	Smoking
Behavior and lifestyle habits	0.02	Exposure to cigarette smoking
Demographic features	0.01	BMI

Table 5: The most important physiological and non-modifiable behavior factors associated with ischemic heart disease based on support vector machine

Factor type	Importance weight	Feature
Details examinations and laboratory factors	0.04	Local pain chest
Demographic features	0.04	Family history of heart disease
Behavior and lifestyle habits	0.04	Cooking method
Behavior and lifestyle habits	0.04	Feeling depressed
Details examinations and laboratory factors	0.03	Diabetes
Demographic features	0.03	BMI
Details examinations and laboratory factors	0.03	Hypertension
Behavior and lifestyle habits	0.03	Salt consumption
Behavior and lifestyle habits	0.03	Cereals consumption
Behavior and lifestyle habits	0.02	fruits and vegetables

Table 6: Extraction of the most important features influencing coronary artery disease variable using KNN

Factor type	Importance weight	Feature
Behavior and lifestyle habits	0.14	Level of education
Details examinations and laboratory factors	0.08	The blood platelets
Demographic features	0.05	Age
Details examinations and laboratory factors	0.05	The blood neutrophil
Behavior and lifestyle habits	0.05	sugary drinks
Behavior and lifestyle habits	0.04	Salt consumption
Behavior and lifestyle habits	0.04	fruits and vegetables
Demographic features	0.04	BMI
Details examinations and laboratory factors	0.03	Hypertension
Behavior and lifestyle habits	0.03	Cereals consumption

Based on support vector machine and according to table, the family history of heart disease, localized pain in the chest, feeling depressed and cooked food are considered the most important risk factors for coronary disease by the algorithm (Table 5).

Based on KNN algorithm, level of education, age and blood platelets and neutrophils in the blood of non-modifiable and modifiable factors combinations are considered the most important risk factors for coronary disease by the algorithm (Table 6).

CONCLUSION

Prediction of the future behavior of patient in a given disease can be one of the most important applications of data mining techniques (Witten *et al*., 2005). Hence, use of data mining in cardiovascular data analysis has provided a good opportunity to examine the relationships between variables related to cardiovascular disease, prediction of cardiovascular disease and identification of factors and reduced risk of heart disease. In this research, modeling and prediction of coronary artery disease were examined using data mining techniques on 345 heart patients in Shahid Beheshti Medical Center. Data mining is able to discover and exploit new knowledge from retrospective data, that is necessary for a good treatment (Wu *et al*., 2008; Yoon *et al*., 2011). The preprocessing of data and variables has significant impact on the discovery of knowledge.There are various techniques of data that are used to predict disease risk factors. In this article, 54 variables were used in data mining model including artificial neural networks, C5.0 decision trees and support vector machine. Experimental results show the efficiency and effectiveness of all three methods were compared based on the sensitivity, specificity and accuracy. To find the optimal structure and increased the accuracy of the results, pruning and strengthening methods were used. Examined existing database and the data were divided into two parts, training and testing results of 70-30%, respectively. The accuracy, sensitivity and transparency of each of the algorithms used in the prediction of coronary artery disease in patients of Shahid Beheshti hospital showed that all algorithms had the ability to predict coronary artery disease with varying degrees of accuracy. Accurate support vector machine 97.25% was the best predictive accuracy of the test data to assist in the classification and artificial neural networks and decision trees methods, respectively and is included 95.42, 87.16%.

In order to identify the relative importance of predictors, artificial neural network model were analyzed using sensitivity analysis. In the analysis, some records were included missing variables and attributes.

The results indicate that three classifications SVM, ANN and decision tree had the minimum error and the highest accuracy was related to SVM. In a decision tree model, accurate prediction showed the lowest among the three prediction models. It is in accordance with other study.

The results showed that the most appropriate model among the three models, SVM was the predictor of effective factor diseases. In addition, variables influencing the disease by the software were recognized by cardiologists as the predictor. Data mining can guide clinicians in predicting ischemic heart disease and factors affecting the accuracy of the results of the models is also quite close to reality.

In future research, merge of multiple databases increased the number of variables.

Data mining and knowledge discovery rules in this research project can be found at health centers to be used for prevention and prediction of coronary disease in other words, knowledge discovery from data can help doctors predict future behavior by patient's record. Also the development of the system and gathering detailed information from patients helped to improve public health.

ACKNOWLEDGEMENTS

The study is the result of a master thesis in health Information Technology. Vice Chancellor for Research of Kashan University of Medical Sciences is highly appreciated for financial support for carrying out the study (Project No 93216) and also, we would like to thank the participants in the collaborative project.

REFERENCES

Alizadehsani, R., J. Habibi, M.J. Hosseini, H. Mashayekhi and R. Boghrati et al., 2013. A data mining approach for diagnosis of coronary artery disease. Comput. Methods Prog. Biomed., 111: 52-61.

Han, J., J. Pei and M. Kamber, 2006. Data Mining, Southeast Asia Edition. 2nd Edn., Morgan Kaufmann, Burlington, ISBN: 9780080475585, Pages: 800.

Heidari, R., M. Sadeghi, H. Sanei, K. Rabiei and M. Shiri, 2012. The effects of trinitroglycerin injection on early complications of angiography. ARYA Atherosclerosis, 8: 50-53.

Higgins, C.B., 1988. Coronary angiography: A decade of advances. Am. J. Cardiol., 62: K7-K10.

Jeddi, F.R. and M.R. Rezaiimofrad, 2013. Development of common data elements to provide tele self-care management. Acta Inf. Med., 21: 241-245.

Jeddi, F.R., F. Abazari, A. Moravveji and M. Nadjafi, 2013. Evaluating the ability of hospital information systems to establish evidence-based medicine in Iran. J. Med. Syst., 37: 1-7.

Jeddi, F.R., M. Arabfard, Z. Arabkermany and H. Gilasi, 2016. The diagnostic value of skin disease diagnosis expert system. Acta Inf. Med., 24: 30-33.

Kajabadi, A., M.H. Saraee and S. Asgari, 2009. Data mining cardiovascular risk factors. Proceedings of the International Conference on Application of Information and Communication Technologies, October 14-16, 2009, Baku, Azerbaijan, pp: 1-5.

Kudyba, S. and T. Gregorio, 2010. Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. Health Inf. J., 16: 235-245.

Lakshmi, K.R. and S.P. Kumar, 2013. Utilization of data mining techniques for prediction and diagnosis of major life threatening diseases survivability-review. Int. J. Sci. Eng. Res., 4: 923-932.

Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc., 83: 1198-1202.

Longo, D.L., A.S. Fauci, D.L. Kasper, S.L. Hauser, J.L. Jameson and J. Loscalzo, 2012. Harrison's Principles of Internal Medicine. 18th Edn., McGraw-Hill, New York, USA.

Mahmoudi, I., M.H. Moazzam and S. Sadeghian, 2013. Prediction model for coronary artery disease using neural networks and feature selection based on classification and regression tree. J. Shahrekord Univ. Med. Sci., 15: 47-56.

Mann, D.L., D.P. Zipes, P. Libby and R.O. Bonow, 2014. Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. Elsevier, London, UK.

Mehdi, K., F. Mehri, S.R. Panoe, P. Zeinab and A. Safollah *et al*., 2011. Evidence-based information resources management skill among Iranian residents, internship and nursing students in urgent care. Sci. Res. Essays, 6: 4708-4713.

Mirershadi, F., M. Faghihi and A.R. Dehpour, 2010. Effect of endogenous nitric oxide on cardiac ischemic preconditioning in rat. Bimonthly J. Hormozgan Univ. Med. Sci., 14: 13-21.

Salarifar, M. and S.M. Kazemeini, 2008. Prevalence of coronary artery disease and related risk factors in first degree relatives of patients with premature CAD Tehran heart center. Tehran Univ. Med. J. TUMS Publications, 65: 49-54.

Sitar-Taut, D.A. and A.V. Sitar-Taut, 2010. Overview on how data mining tools may support cardiovascular disease prediction. J. Applied Comput. Sci., 4: 57-62.

Sonawane, J.S., D.R. Patil and V.S. Thakare, 2013. Survey on decision support system for heart disease. Int. J. Adv. Technol., 4: 89-96.

Soni, J., U. Ansari, D. Sharma and S. Soni, 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int. J. Comput. Appl., 17: 43-48.

Tohidi, M., M. Assadi, Z. Dehghani, K. Vahdat, S.R. Emami and I. Nabipour, 2012. High sensitive C-reactive protein and ischemic heart disease, a population-based study. Iran South Med. J., 15: 253-262.

Witten, H.I., E. Frank and M.A. Hall, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann Publ., Massachusetts.

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh and Q. Yang *et al*., 2008. Top 10 algorithms in data mining. Knowledge Inform. Syst., 14: 1-37.

Yoon, H., S.C. Jun, Y. Hyun, G.O. Bae and K.K. Lee, 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol., 396: 128-138.