Volume 7, No. 5, September-October 2016
tional Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Classification models on cardiovascular disease detection using Neural Networks, Naïve Bayes and J48 Data Mining Techniques

Mudasir M Kirmani
SKUAST-K,
J&K, India

Syed Immamul Ansarullah
MANUU,
Hyderabad, India

Abstract: The huge amounts of data generated by healthcare transactions are complex and voluminous which needs to be processed and analyzed by different traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. In today's modern world cardiovascular disease is the most lethal one. Diagnosis of heart disease is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. This research paper investigates three different classification models of Data Mining Techniques for detection of cardiovascular disease to facilitate experts in the healthcare domain. This research paper highlights the performance of all the three classifications models on cardiovascular disease detection and the same has been justified with the results of different experiments conducted using WEKA machine learning software.

Key words: Data mining, Decision Tree, Multilayer Perception, Naive Bayes, cardiovascular disease, Coronary heart disease.

Introduction

The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the United States, England, Canada and Wales as in 2007 [1] it was reported that heart disease kills one person every 34 seconds in the United States [1]. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases [1]. Cardiovascular disease (CVD) results in severe illness, disability and in extreme cases may result in death of a patient. Narrowing of the coronary arteries leads to the Coronary heart disease (CHD). Myocardial infarctions are generally known as heart attacks and angina pectoris or chest pain are encompassed in the Coronary heart disease [1]. Making a diagnosis of heart disease includes taking a complete medical evaluation, history, physical examination and early diagnosis of heart

disease can help reduce the rate of mortality (Thaksin University, 2006) [17]. One of the best ways to diagnose a heart disease is by using echocardiography. Echocardiography or echo is a painless test that uses sound waves to create pictures of the heart. The test gives information about the size & shape of the heart and how well the heart chambers & valves are working. The test also can identify areas of heart muscles that are not contracting normally due to poor blood flow or injury from a previous heart attack. In addition, a type of echo test called Doppler ultrasound shows how well blood flows through the chambers and valves of the heart (Joel and Robert, 1976) [18].

As described by National Heart Lung and Blood Institute (2008) echo can detect possible blood clots inside the heart; fluid buildup in the pericardium; and problems with the aorta. However, the interpretation of echo recordings remains a challenge as no precise rules that are deduced from databases are present. The analysis of Echo data by experts is time consuming and this is in connection with the shortage of experts possessing knowledge on the analysis of Echo data. Therefore, methods to automate the interpretation of Echo recordings by minimizing human efforts are important for diagnosis of heart disease in patients [2]. In order to solve this problem in the healthcare sector different researchers have developed an assistant tool to help cardiologists in diagnosing heart diseases based on echo readings.

Related Work

Different studies have been carried out by researchers which focus on diagnosis of heart disease using the data available from different Echo test results. They have applied different data mining techniques for diagnosis of cardioVascular disease and the results have shown different probabilities for different methods. The different methods developed and used by researchers have been explained below:

The researchers have used pattern recognition and data mining methods as predicting models in the domain of cardiovascular diagnoses. The experiments were carried out using classification algorithms Naïve Bayes, Decision Tree, KNN and Neural Network. The results reported have shown Naive Bayes technique has performed better than the other techniques [3].

The researchers have used K-means clustering algorithm on a heart disease warehouse to extract data relevant to heart disease and applied MAFIA (Maximal Frequent Item set Algorithm) algorithm to calculate weightage of the frequent patterns significant to heart attack predictions [4].

The researchers had proposed a layered neuro-fuzzy approach to predict occurrences of coronary heart disease simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced error rate which was very low and high work

efficiency in performing analysis for coronary heart disease occurrences [5].

The researchers had proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions. The implementation of the proposed approach was implemented in C programming language and reduced main memory requirement by considering a small cluster at a time in order to be considered scalable and efficient [6].

The researchers used the data mining algorithms Decision Trees, Naive Bayes, Neural Networks, Association Classification and Genetic Algorithm for predicting and analyzing heart disease from the given dataset [7].

An experiment was performed by the researchers on a dataset using Neural Networks and Hybrid Intelligent Algorithm, and the results shows that the Hybrid Intelligent Technique did improve accuracy of the prediction on different datasets [8].

The researcher used Association Rules representing a technique in Data Mining to improve disease prediction with higher levels of efficiency and effectiveness. An algorithm with search constraints was also introduced to reduce the number of Association Rules and validated using "train and test" approach [9].

The researchers used classifiers such as Decision Trees, Naive Bayes, and Neural Network to predict heart disease with 15 popular attributes as risk factors that are included in the medical literature [10].

The researchers [11] implemented a Hybrid System that uses global optimization benefit of Genetic Algorithm for initialization of Neural Network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high- cholesterol, smoking, alcohol-intake and obesity [10].

A model Intelligent Heart Disease Prediction System (IHDPS) was developed with the aid of Data Mining Techniques like Decision Trees, Naive Bayes and Neural Network was proposed by Palaniappan and Awang, (2008) [12], they used a CRISP-DM methodology to develop the mining models on a dataset obtained from the Cleveland Heart Disease database. IHDPS was capable of answering queries that the conventional Decision Support Systems were not able to cater. IHDPS facilitated the establishment of vital knowledge patterns and relationships amid medical factors connected with heart disease [12].

Guru et al. [13] experimented on a sample database of patient records. The Neural Network technique was tested and trained with 13 input variables. The supervised network had been recommended for diagnosis of heart diseases. Training was carried out with the aid of Back Propagation algorithm. The system developed by the researcher did generate list of probable diseases that a patient may be vulnerable to by

comparing the unknown datasets with the trained datasets. The results of the predictions for unknown datasets were promising and the success rate was more than 99.0%.

To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by Heon Gyu Lee et al. [14]. To achieve this, they have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).

Kiyong Noh et al. [15] used a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern. The dataset consisted of dataset corresponding of 670 persons, distributed into two groups namely "normal-people" and "patients with heart disease" were employed to carry out the experiment for the associative classifier.

Based on the literature reviewed the research work is an effort to work on the classification model for prediction of heart disease based on patterns generated from International Cardiovascular Hospital database.

Classification Techniques Used For Heart Disease Predictions

The main objective of this research is to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop this system, 15 input parameters of ECG attributes were used. The three different Data Mining Classification Techniques Neural Networks, Decision Trees and Naive Bayes were used to analyze the dataset.

Neural Networks

An Artificial Neural Network (ANN) also known as Neural Network (NN) is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [16].

Decision Trees

The Decision Tree [16] approach is one of the best approaches for classification problems. There are two steps in this technique which include building a tree & applying the tree to the dataset. There are many popular Decision Tree Algorithms like CART, ID3, C4.5, CHAID and J48. J48 algorithm has been used in this system for better performance.

Naive Bayes

Naive Bayes classifier [16] is based on Bayes theorem. This classifier algorithm uses conditional independence by assuming that an attribute value on a given class is independent of the values of other attributes.

Performance Measures

The overall accuracy of classifiers is estimated by 10-Fold cross validation and confusion matrix. However, performance measures like RECALL (SENSITIVITY), SPECIFICITY and F-measure are also used for calculating other aggregated performance measures (e.g., area under the ROC curves).

10-Fold Cross Validation

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold.

Confusion Matrix

In classification problems, the primary source of performance measurements is a confusion matrix Given m classes, a confusion matrix is a table of at least size 'm' by 'm' (Olson and Delen, 2008) [19].

If the instance is **POSITIVE** and it is classified as **POSITIVE**, it is counted as a **TRUE POSITIVE (TP);**

If the instance is **POSITIVE** and If it is classified as **NEGATIVE**, it is counted as a **FALSE NEGATIVE (FN).**

If the instance is **NEGATIVE** and if it is classified as **NEGATIVE**, it is counted as a **TRUE NEGATIVE (TN);**

If the instance is **NEGATIVE** and if it is classified as **POSITIVE,** it is counted as a **FALSE POSITIVE (FP).**

Table 1: Confusion Matrix

		PREDICTED CLASS	
ACTUAL CLASS		***CLASS 1***	***CLASS 2***
	CLASS 1	TRUE POSITIVES	FALSE NEGATIVE
	CLASS 2	FALSE POSITIVES	TRUE NEGATIVES

Table 1 shows a confusion matrix for a two-class classification problem. The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. The equations of most commonly used metrics that can be calculated from the coincidence matrix are discussed below.

Accuracy

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$\text{ACCURACY} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP – True Positive, TN – True Negative, FP – False Positive and FN – False Negative.

Han and Kamber (2006) [16] gave an example to emphasize on using other measures as an alternative to the accuracy measure. Assume that a trained classifier for classifying medical data tuples as either "Heart Disease" or "Not Heart Disease" an accuracy rate of 90% will make the classifier seem quite accurate. However, in case of 3–4% of the training tuples valued as "Heart Disease", an accuracy rate of 90% may not be acceptable—the classifier could be correctly labeling only the "Not Heart Disease" tuples. Instead it would be preferred to access how well the classifier can recognize "Heart Disease" tuples (the positive tuples) and how well it can recognize "Not Heart Disease" tuples (the negative tuples). The SENSITIVITY and SPECIFICITY measures can be used for this purpose.

Sensitivity

SENSITIVITY (also referred to as the TRUE POSITIVE or RECOGNITION or RECALL rate) is the proportion of positive tuples that are correctly identified.

$$\text{TRUE POSITIVE RATE} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Where TP – True Positive and FN – False Negative.

Specificity

SPECIFICITY (TRUE NEGATIVE) rate is the proportion of negative tuples that are correctly identified.

$$\text{TRUE NEGATIVE RATE} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Where TN – True Negative and FP – False Positive.

Precision

PRECISION is another performance measuring unit where PRECISION for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$\text{PRECISION} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{F}\text{P}}$$

where TP – True Positive and FP – False Positive.

F-measure

F-measure is a measure of test's accuracy. It considers both the PRECISION and the RECALL of the test to compute the score. It can be interpreted as a weighted average of the PRECISION and the RECALL, where 1 is the best value and 0 is the worst. The F-Measure only produces a high result when PRECISION and RECALL are both balanced which makes it very significant.

$$\mathrm{F-measure} = \frac{2}{\frac{1}{\mathrm{PRECISION}} + \frac{1}{\mathrm{RECALL}}}$$

Area under the ROC Curve

A Receiver Operating Characteristics (ROC) curve is a technique for visualizing, organizing and selecting classifiers based on their performance. This technique is used as a performance evaluation technique for classification models and is very useful tool for comparing two or more classification models. As Han and Kamber (2006) [16] described ROC curve showing the trade-off between the true positive rate and the false positive rate for a given model. The area under the ROC curve is a measure of the accuracy of the model.

Selecting the Target Dataset

The transthoracic echocardiography report of 7,708 patients with a size of 300 MBs was selected as a target dataset including 15 echocardiography attributes and the list is given in table 2.

Table 2: Echocardiography attributes and their description

S.No	Attributes	Description	Type
1	Age	Age of the patient in years	Numeric
2	Sex	Sex of the patient (Male / Female)	Nominal
3	Aortic root – diameter	Size of Aortic root – diameter in mm	Numeric
4	Left atrium: (systole) diameter	Size of Left atrium: (sys) diameter in mm	Numeric
5	Left ventricle in: diastole	Left ventricle in: diastole in mm	Numeric
6	Left ventricle in systole	Size of Left ventricle in systole in mm	Numeric
7	Posterior wall of LV	Size of Posterior wall of LV in mm	Numeric
8	Interventricular septum	Size of in Interventricular septum in mm	Numeric
9	LV- ejection fraction	Fraction of blood pumped out of ventricles with each heart beat in percentage	Numeric

10	Main Pulmonary Artery diameter	Size of Main Pulmonary Artery diameter in cm	Numeric
11	Pericardial effusion	Presence of an abnormal amount and/or character of fluid in the pericardial space	Ordinal
12	TR Velocity	Tricuspid Regurgitation Velocity in cm/sec	Numeric
13	Em/Am velocity ratio	The ratio between myocardial early and atrial peak velocities	Numeric
14	Rhythm	Type of the heart rhythm observed	Nominal
15	Diagnosis	Does the patient has a heart disease (Yes or No)	Nominal

RESULTS

As the objective of this research is to detect heart disease using Data Mining technique therefore, a Classification Data Mining Technique was adopted to develop a predictive model. The model was developed with three different supervised machine learning algorithms Decision Tree, Naive Bayes and Neural Network using WEKA 3.6.4 machine learning software. Three different experiments were conducted on the dataset using two scenarios one will 15 attributes and second with 08 selected attributes. The performance of the experiments conducted is listed in table 3.

Table 3: Performance of supervised machine learning algorithms ;Decision Tree, Naïve Bayes and Neural Network

Classifier	Instances	Attribute		Time to Build Model	Accuracy	True positive Rate	True Negative Rate	Precision	F-measure	ROC Area
		15	8							
J48	7339	Yes	×	0.89 sec	94.29 %	0.932	0.95	0.943	0.943	0.942
	7339	×	Yes	0.36 sec	95.52 %	0.944	0.963	0.955	0.955	0.965
es	7339	Yes	×	0.11 sec	91.96 %	0.865	0.959	0.92	0.919	0.97

	7339	×	Yes	0.05 sec	92.42 %	0.871	0.962	0.925	0.924	0.972
Neural Network	7339	Yes	×	158.94 sec	93.83 %	0.915	0.955	0.938	0.938	0.969
	7339	×	Yes	34.14 sec	94.85 %	0.932	0.96	0.948	0.948	0.974

Based on the results given in table 3, performance of three different classifiers was analyzed and it was observed that all the three algorithms J48, Naïve Bayes and Neural Network performed best in True Negative Rate. Therefore, the models are best in identifying Negative cases. The results also showed that the Naïve Bayes and Neural Network Classifiers perform better on 08 (eight) selected attributes when compared with 15 attributes.

1. Specific Rule Extraction

The model developed with J48 classifier was the best model for this research. Researchers extracted four rules and showed their influence and importance on prediction of heart disease.

Rule 1: IF Left Atrium Systole Diameter > 40 millimeter **AND** LV Ejection Fraction<=51% **THEN** Diagnosis = YES.

This rule is the strong rule for predicting patients with heart disease. Success fraction of this rule is 99.79%.

Rule 2: IF Left Atrium Systole Diameter > 40 millimeter **THEN** Diagnosis = YES.

Based on this rule patient with Left Atrium Systole Diameter > 40 millimeter are in high risk of having Heart Disease. Success fraction of this rule is 97%. From this rule it can be concluded that the attribute Left Atrium Systole Diameter is a key attribute in determining patients with heart disease.

Rule 3: IF Left Atrium Systole Diameter <= 40 millimeter **AND** LV Ejection Fraction>51% **AND** posterior wall of LV <=11 **AND** Main pulmonary Artery Diameter<=2 **AND** Pericardial Effusion=”Normal” **AND** TR Velocity<2.6 **AND** EM_AM Velocity ratio>0.6 **THEN** Diagnosis=”NO”.

This rule is strong in identifying normal patients i.e. patients who are free from heart disease. The success fraction =97.85%.

Rule 4: IF Main Pulmonary Artery diameter > 1.8 millimeter **AND** TR Velocity > 2.3 **THEN** Diagnosis = "Yes". The success fraction of this rule 4 is 53.62%.

CONCLUSION

The main aim of this research work was to design a diagnostic model for the prediction of heart disease using three different Data Mining supervised machine learning algorithms Decision Tree, Naive Bayes and Neural Network tested using WEKA machine learning software. The performance of the models was evaluated using the standard metrics of ACCURACY, PRECISION, RECALL and F-measure. J48 classifier performed better in predicting the heart disease with 95.52% of Accuracy. Significant rules were extracted from dataset that will be useful in predicting heart disease. This model can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease effectively. Furthermore, the resulting model has a high specificity rate which makes it a handy tool for junior cardiologists to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis. The research work carried out will work as a platform for future research in an effort to make heart disease detection using Data Mining Techniques effective and efficient.

REFERENCES

1. World Health Organization. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. 2011
2. Felner M., J. and Schlant C., R. (1976). Echocardiography: A Teaching Atlas. Grune & Stratton Publishers, New York.
3. T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012.
4. S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.
5. A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, vol. 2, no. 9, pp. 1663–1671, 2013.
6. M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.
7. K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

8. R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605–609, 2013.

9. C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction.," IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, vol. 10, no. 2, pp. 334–43, Apr. 2006.

10. K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in The 5th International Conference on Computer Science & Education, 2010, pp. 1344–1349.

11. S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. Ict, pp. 1227–1231.

12. Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

13. Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

14. Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

15. Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721-727, 2006.

16. Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco

17. Thaksin University, Faculty of Science (2006), Available at http://www.sci.tsu.ac.th/sciptl/radio/tape043/colesterol.doc (Accessed15 December 2010)

18. Felner M., J. and Schlant C., R. (1976). Echocardiography: A Teaching Atlas. Grune & Stratton Publishers, New York

19. David L. Olson and Dursun Delen, "Advanced Data Mining Techniques" springer.com 2008.