Journal of Cardiovascular Diseases & Diagnosis

Chaithra and Madhu, J Cardiovasc Dis Diagn 2018, 6:6
DOI: 10.4172/2329-9517.1000348

Research Article Open Access

Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques

Chaithra N[1*] **and Madhu B**[2]

[1]*Faculty of Life Sciences, Division of Medical Statistics, Jagadguru Sri Shivarathreeshwara University, Mysore, Karnataka, India*

[2]*Department of Community Medicine, Jagadguru Sri Shivarathreeshwara University, Mysore, Karnataka, India*

[*]**Corresponding author:** Chaithra N, Assistant Professor, Faculty of Life Sciences, Division of Medical Statistics, Jagadguru Sri Shivarathreeshwara University, Mysore, Karnataka, India, Tel: +91 9742119518, E-mail: chaithra.mstats@jssuni.edu.in

Received: November 13, 2018**; Accepted:** November 27, 2018**; Published:** November 30, 2018

Abstract

Background: The huge amounts of data generated by healthcare transactions are complex and voluminous. They need to be processed and analysed by different traditional methods. Data Mining provides the methodology and technology to transform these amounts of data into useful information for decision making. Cardiovascular diseases are one of the highest-flying diseases of the modern world. The treatment of the said disease is quite high and not affordable by most of the patients particularly in India. To solve this problem, Data Mining is the best available technique for classification and prediction.

Aim: Research work was aimed to analyse the various data mining techniques introduced in recent years to design a predictive model for cardiovascular diseases from the data obtained by transthoracic echocardiography.

Methods: A total of 336 records with 24 attributes were highly relevant in predicting heart disease from echocardiography dataset were analysed by applying techniques prospectively. This study investigates three different classification models: J48 Decision Tree, Naive Bayes and Neural Network on cardiovascular disease prediction and the same has been justified with the results of different experiments conducted and the performance of the models was evaluated using the standard metrics of Accuracy, Precision, Recall and F-measure.

Discussion and Conclusion: The results of all the three algorithms performed best in true negative rate which makes it a handy tool to train medical students and junior cardiologists to diagnose patients with heart disease.

Keywords: Cardiovascular disease; Echocardiography; Classification models

Introduction

In today's modern world cardiovascular disease is the most lethal one [1]. According to World Health Organization about more than 12 million deaths occurs worldwide, every year due to heart problems [2]. With the turn of the century, cardiovascular diseases (CVDs) have become the leading cause of mortality in India [3]. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels, and the way blood is pumped and circulated through the body, also are considered forms of heart disease [4]. This disease attacks a person so instantly that it hardly gets any time to get treated with. 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 [5].

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 [6,7]. So, diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. The Healthcare industry today generates huge amounts of complex data about patients, disease diagnosis, hospitals resources and medical devices, which is difficult to process by manual methods [8]. Data mining provides a set of tools and techniques to find patterns and extract knowledge to provide better patient care and it combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [9]. 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. Effective and efficient automated heart disease prediction can benefit healthcare sector and this automation will save not only cost but also time [10]. This research paper highlights the utility and application of three different classification models of data mining techniques for prediction of cardiovascular disease to facilitate experts in the healthcare domain [11-13].

Methods

A total of 336 records with 24 attributes were obtained from the Echocardiography database and list is given in the Table 1. The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. The present study conducted by using simple random sampling (SRS) method [14,15], with an SRS each patient has an equal chance of being chosen. Every patient who comes for the ECHO are included and paediatric patient are excluded for the study, patient's personal information is collected such as Age, Sex, Smoking, Alcohol intake, Diabetics, Hypertension, Family History and echocardiography measurements are recorded measured by ECHO technician.

S. No.	Attributes	Description	Type
1	Age	Age of the patient in years	Numeric
2	Sex	Gender of the patient (Male/Female)	Nominal
3	Smoking	Smoking habit of the patient (Never, Current, Past)	Nominal
4	Alcohol intake	Alcohol intake of the patient (Never, Current, Past)	Nominal
5	Diabetics	Diabetics of the patient (Yes, No)	Nominal
6	Hypertension	Hypertension of the patient (Yes, No)	Nominal
7	Family History	Family history of the patient (Yes, No)	Nominal
8	AO (mm)	Aorta root	Numeric
9	LA (mm)	Left atrium	Numeric
10	RV (mm)	Right ventricle	Numeric
11	LVID_d (mm)	Left ventricular internal diameter end diastole	Numeric
12	LVID_s (mm)	Left ventricular internal diameter end systole	Numeric
13	IVS_d (mm)	Interventricular septum end diastole	Numeric
14	IVS_s (mm)	Interventricular septum end systole	Numeric
15	LVPW_d (mm)	Left ventricular posterior wall end diastole	Numeric
16	LVPW_s (mm)	Left ventricular posterior wall end systole	Numeric
17	EDV (ml)	End diastolic volume	Numeric
18	ESV (ml)	End systolic volume	Numeric
19	SV (ml)	Stroke volume	Numeric
20	EF %	Ejection fraction	Numeric
21	FS %	Fractional shortening	Numeric
22	MPA (mm)	Main pulmonary artery	Numeric
23	Rhythm	Type of the heart rhythm observed	Nominal
24	Diagnosis	Those the patient has the heart disease (Yes/No)	Nominal

Table 1: Echocardiography attributes and their description.

To understand classifier's behaviour, we should calculate metric Confusion Matrix. It is a visualization tool which is commonly used to present the accuracy of the classifiers in classification [16].

Actual class	Predicted class		
		Positive	Negative
	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Table 2: Confusion matrix.

Table 2 shows a confusion matrix for a two-class classification problem. It is a contingency table that contains information about actual and predicted classifications done by a classification system. It is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The entries in the confusion metrics that can be calculated from the coincidence matrix, we use hypothesis below:

- True Negative (TN) is the number of correct predictions that an instance is negative.
- False Positive (FP) is the number of incorrect predictions that an instance is positive.
- False Negative (FN) is the number of incorrect of predictions that an instance negative.
- True Positive (TN) is the number of correct predictions that an instance is positive.

Results

The experimental results have shown that Neural Network outperformed J48 Decision tree and Naïve Bayes in the domain of predicting heart diseases cases. Three different experiments were conducted on the echocardiography report dataset, the experiment was designed to evaluate the performance of a J48 Decision tree, Neural Network and Naïve Bayes to investigate the effect of attribute selection on the model. Neural Network has proved its performance as a powerful classifier in term of accuracy (97.91%), Sensitivity (97.2%) and Specificity (98.4%), which makes it a good classifier to be used in the medical field for classification and prediction.

Discussion

In this research, the data mining classifiers J48 Decision tree, Naïve Bayes, and Neural Network are considered for the comparisons to classify and diagnose heart diseases for the patient data set from medical practitioners. For better understanding results of confusion matrix for all the three algorithms given in Table 3.

Classification Matrix displays the frequency of correct and incorrect predictions [17]. It compares the actual values in the test dataset with the predicted values in the trained model. Table 3 shows the results of the Classification Matrix for all the three algorithms, 88%, 97% and 50% patients are correctly diagnosed that they have disease and predicted as having the disease. 12%, 3% and 50% patients are wrongly diagnosed as they don't have but, they had disease, it is very dangerous saying the disease patient is free from the disease. 4%, 2% and 7% patients are diagnosed as they have the disease but, they were free from the disease. 96%, 98% and 93% patients are correctly diagnosed as they don't have disease and predicted as not having the disease for J48 Decision Tree, Neural Network and Naïve Bayes algorithms respectively.

Algorithm	Actual Class	Predicted Class	
		Yes	No
J48 Decision Tree	Yes	0.88	0.12
	No	0.04	0.96
Neural Network	Yes	0.97	0.03
	No	0.02	0.98
Naïve Bayes	Yes	0.5	0.5
	No	0.07	0.93

Table 3: Results of Confusion Matrix for all the three algorithms.

Classifier	Instances	Attributes	Time to build model (sec)	Accuracy (%)	True Positive Rate	True Negative Rate	Precision	F-measure	ROC Area
J48 Decision Tree	336	24	0.02	92.55	0.87	0.96	0.95	0.91	0.94
Neural Network	336	24	1.81	97.91	0.97	0.98	0.98	0.98	0.97
Naïve Bayes	336	24	0.02	74.40	0.5	0.93	0.84	0.62	0.79

Table 4: Performance measures of decision tree, neural network and naïve Bayes.

The performances of the models in this study were evaluated using the standard metrics of accuracy, precision, F-measure which were calculated using the predictive classification table, ROC area was also used to compare the performances of the classifiers [18-20]. Based on the results given in Table 4.

Three different experiments were conducted on the dataset of 336 instance 24 attributes using three algorithms: J48 Decision Tree, Naive Bayes and Neural Network, respectively it took 0.02, 1.81- and 0.02-seconds time to build the models. The True positive rate for J48 Decision Tree algorithm (0.87), Neural Network (0.97) and Naive Bayes (0.5). Whereas Neural Network performed best in True Positive Rate 0.97 and Naive Bayes performed lowest in True Positive Rate 0.5. The True Negative Rate for J48 Decision Tree algorithm (0.96), Neural Network (0.98) and Naive Bayes (0.93), it was observed that all the three algorithms J48 Decision Tree, Naïve Bayes and Neural Network performed best in True Negative Rate. Therefore, the models are best in identifying Negative cases. The comparative ROC curves based on risk of heart diseases. Neural Network has outperformed than J48 Decision Tree, Naïve Bayes with area under curve (AUC) 0.97, AUC for J48 Decision Tree was 0.94 and Naive Bayes 0.79. Overall, these results of area under curve reveals better performance of Neural Network.

Conclusion

The analysis shows that Neural Network performed better in predicting the heart disease with 97.91% of accuracy, this model will have high true negative rate which makes it a handy tool for junior cardiologists and echo technicians to screen out patients who have a high probability of having the disease and transfer those patients to senior cardiologists for further analysis.

Conflicts of Interest

There are no conflicts of interest for the present study.

References

1. Reddy RV, Raju KP, Kumar MJ, Sujatha CH, Prakash PR (2016) Prediction of heart disease using decision tree approach. Int J Adv Res Comput Sci Softw Eng 6: 530-532.
2. Soni J, Ansari U, Sharma D, Soni S (2011) Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int J Comput Appl 17: 43-48.
3. Prabhakaran D, Jeemon P, Roy A (2016) Cardiovascular diseases in India: Current epidemiology and future directions Circulation 133: 1605-1620.
4. Taneja A (2013) Heart disease prediction system using data mining techniques. Orient J Comp Sci & Technol 6: 457-466.
5. Dey M, Rautaray SS (2014) Study and analysis of data mining algorithms for healthcare decision support system. Int J Comput Sci Inf Technol 5: 470-477.
6. Kaddoura S (2009) Echo made easy. (2nd edn). Churchill Livingstone, Elsevier pp:1-219.
7. Kirmani MM, Ansarullah SI (2016) Classification models on cardiovascular disease detection using neural networks, naïve Bayes and J48 data mining techniques. Int J Adv Res Comput Sci 7: 52-61.
8. Rohilla J, Gulia P (2013) Analysis of data mining techniques for diagnosing heart disease. Int J Adv Res Comput Sci Softw Eng 3: 717-724.
9. Abdar M, Kalhori SR, Sutikno T, Subroto IM, Arji G (2015) Comparing performance of data mining algorithms in prediction heart diseases. International Journal of Electrical and Computer Engineering 5: 1569-1576.
10. Bhatla N, Jyoti K (2012) An analysis of heart disease prediction using different data mining techniques. Int J Adv Res Technol 1: 1-4.
11. Sudhakar K, Manimekalai DM (2014) Study of heart disease prediction using data mining. Int J Adv Res Comput Sci Softw Eng 4: 1157-1160.
12. Srinivas K, Rani BK, Govrdhan A (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering 2: 250-255.
13. Kajal ES, Nishika P (2016) Prediction of heart disease using data mining techniques. International Journal of Advance Research, Ideas and Innovations in Technology 2: 1-7.
14. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 39: 561-577.
15. Boukenze B, Mousannif H, Haqiq A (2016) Predictive analytics in healthcare system using data mining techniques. Comput Sci Inf Technol 1: 1-9.
16. Kim J, Lee J, Lee Y (2015) Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Health Informatics J 21: 167-174.
17. Sundar NA, Latha PP, Chandra MR (2012) Performance analysis of classification data mining techniques over heart disease database. Int J Eng Sci Adv Technol 2: 470–478.

18. Alzahani SM, Althopity A, Alghamdi A, Alshehri B, Aljuaid S (2014) An overview of data mining techniques applied for heart disease diagnosis and prediction. Eng Technol Publ 2: 310–315.

19. Masethe HD, Masethe MA (2014) Prediction of heart disease using classification algorithms. In:"Proceedings of the world Congress on Engineering and computer Science 2: 22-24.

20. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J (2015) Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Inform 216: 40-44.