International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6| ISSN : 2456-3307

An Ensemble Classifier for the Prediction of Heart Disease

Ria A Kurian
Department of Information and Technology, Rajagiri School of Engineering and Technology, Ernakulam, Kerala, India

ABSTRACT

Heart disease has become a silent killer among people of all ages. The major risk factors of heart disease include smoking, blood pressure, cholesterol, diabetes etc. Early diagnosis and treatment can reduce morbidity rate to an extent by identifying patients at higher risk of having a heart disease and providing them right care at right time. However provisioning of quality services at reasonable costs is a major concern of every healthcares. Poor clinical decisions can pose adverse effects on human health. This paper introduces a method based on data mining according to the information of patients' medical records to predict heart disease. An ensemble classifier approach is being used, that is the combination of three classifiers (KNN, Decision Tree, NaiveBayes) composing an ensemble, so that the overall model can be used to give predictions with greater accuracy than that of individual classifiers.

Keywords : Heart Disease, KNN, decision Tree, Naive Bayes, Classifier Ensemble, Majority Voting.

I. INTRODUCTION

Heart disease has become a leading cause of death over the years. An estimated 17 million people die of heart diseases (CVD) every year [1]. Heart disease is mainly caused by the blocked blood vessels to heart. Unhealthy diet, smoking and alcohol consumption leading to raised blood pressure, blood lipids, overweight and obesity forms the major risk factors. A congestive heart failure can even lead to severe other complications including heart attack and stroke.

Data mining is a powerful technique that helps to retrieve relevant information from huge set of data. It has been applied in the healthcare industries for the effective prediction of diseases. Data mining helps healthcare professionals to analyse huge set of patient records and find out the relationship between risk factors and disease providing a way of earlier detection of patients at higher risk. Research in the field of cardiovascular diseases using data mining has been an ongoing effort involving prediction, treatment, and risk score analysis with high levels of accuracy[2]. Various classification techniques such as Decision Tree, Artificial Neural Networks, Support Vector Machines, Naive Bayes, Logistic Regression, etc. have been used to build models in healthcare research[2] .

Classification algorithms are found out to be very useful in categorizing the data for making appropriate decisions. In this study, an ensemble classifier approach has been used for the prediction of heart disease which shows better predictive accuracy than that of individual classifiers. The main objective is to reduce the morbidity rate by providing early diagnosis of heart disease based on risk factors. It also reduces the accompanied cost by sidestepping the unnecessary medical tests to be performed.

II. LITERATURE SURVEY

Several classifiers such as SVM, Neural networks, Logistic Regression have been used for identifying

people at higher risk of having a heart disease. These classifiers were tested based on their accuracies.

Mythili T et al [2] showed a framework using combinations of support vector machines, logistic regression, and decision trees to arrive at an accurate prediction of heart disease. Most efficient model of the multiple rule based combinations by training and testing the system was being achieved. Comparison between performance measures which include sensitivity, specificity, and accuracy was also performed. SVM proved to be the best model with accuracy of 90.5%, followed by decision tree with 77.9%, and logistic regression 73.9%.

Rafiah et al [3] developed a heart disease prediction system using three data mining techniques such as Decision Trees, Naive Bayes, and Neural Network. The result obtained after prediction using the Cleveland Heart disease Database indicated that Naive Bayes performed well followed by Neural Network and Decision Trees. It was also observed that the relationship obtained between attributes using Neural Network is more difficult to understand than that of the other models used to predict heart disease.

Another approach by Nitin Kumari et al [4] compared three soft computing techniques ANNs, fuzzy logic and neuro-fuzzy integrated approach for the diagnosis of coronary heart disease. The evaluated result showed that neuro-fuzzy integrated approach which comprises the merits of both Artificial Neural Network and Fuzzy Logic is found out to be the best one among the three. Latha Parthiban et al.[5] implemented a similar heart disease prediction system based on coactive neuro-fuzzy inference system(CANFIS) integrated with genetic algorithm. The CANFIS model is developed to find a best solution for this problem by combining the advantages of neural network and the fuzzy logic approach. Genetic optimization helped in the selection of most relevant features from the training data by removing the redundant variables. Some of the works done are given below, in Table1

Authors	Classifiers	Accuracy
Das et al.[6]	ANN ensemble	89.01%
Nidhi Bhatla et al.[11]	Decision Tree Naive Bayes ANN	89% 86.53% 85.53%
T. John et al.[13]	Naive Bayes Multi Layered Feed Forward	85.18% 78.88%
A.Q. Ansari et al.[10]	Neuro-fuzzy intergrated system	Maximum accuracy obtained
Muhammed et al. [7]	CLIP3 CLIP4 CLIP4 ensemble	84.0% 86.1% 90.4%
Polat et al. [8]	AIS	84.5%
Tu et al. [12]	J4.8DecisionTree Bagging Algorithm	78.9% 81.41%
Detrano et al. [9]	Logistic regression	77%

TABLE I : Various Classifiers and their accuracies applied for the prediction of Heart Disease

All these works showed how different classification models are being used for the prediction of heart disease. However there is a need for more precise predictions and hence this paper presents a unique model, a combination of three classifiers namely KNN, Decision Tree and Naive Bayes forming an ensemble that can give more accurate predictions than that of individual classifiers.

III. METHODS AND MATERIAL

Data mining has been used in various fields over the years, but now only, it has become well recognized and more reliable. Data mining helps to extract relevant informations from a huge set of data. This paper proposes a new method based on data mining

that focuses on a better predictive performance than the existing models. A combination of three classifiers such as KNN, Naive Bayes and Decision Tree have been used so that the overall model forming an ensemble gives accurate predictions than that of individual classifiers.

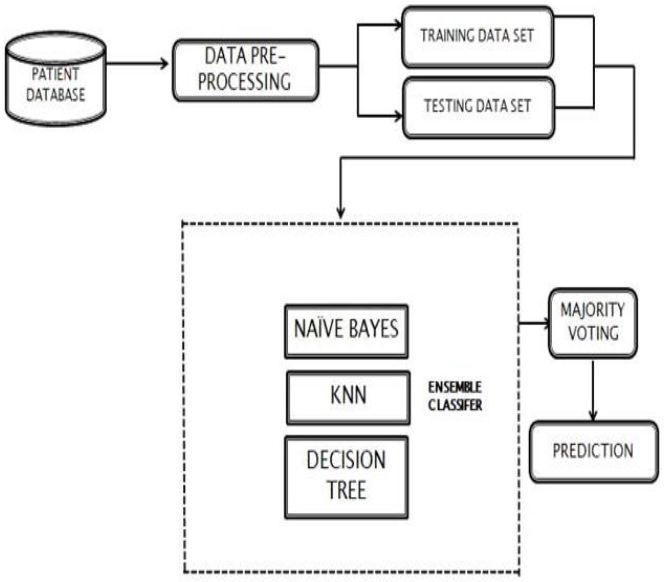

Figure 1 : System Overview

A. *Data Source*

Patient Database is collected from UCI Repository [7]. Dataset has 303 instances and 13 attributes which are important for heart disease prediction. The input data source is fed into Data pre-processing stage which involves removal of missing fields and outliers, normalization and transforming the data into the appropriate form.

B. Classifier Ensemble

An ensemble of classifiers is a set of classifiers by which decisions of each classifiers are combined together to classify a new set of data by performing a voting among the obtained predictions. In this study, mainly three classifiers are being used: KNN, Naive Bayes and Decision Tree. A majority voting approach is applied to this ensemble to get more accurate predictions of heart disease.

1. KNN Classifier

Neighbors Algorithm is considered as a lazy learning algorithm that classifies data set based on the similarity with neighbors. Working of KNN is simple. Minimum distance from test samples to training samples is calculated using distance metric such as Euclidean Distance to determine the K nearest neighbor. After gathering the nearest neighbors, majority of the K-nearest neighbors is taken to obtain the prediction of the test sample.

2. Naive Bayes Classifier

Naive Bayes performs probabilistic prediction with an assumption that there exist a strong independence among predictors.

Bayesian classification predicts the class of new set of data, following the Bayes theorem.

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)}$$

P(C|X) - Posterior probability of class C given a predictor X.

P(X|C) - Probability of predictor X given a class C, calculated from training data.

P(C) - Prior probability of class C.

P(X) - Evidence or the predictor prior probability X

Naive Bayes proved to be more accurate and faster to train. It find application in medical diagnosis and often outperforms even more complex classification methods.

3. Decision Tree Classifier

Decision Tree creates a training model to predict the class of new set of data based on the decision rules inferred from the training data. Given a data of attributes together with classes, a decision tree requires to calculate its best split node. There are many different splitting criterions that can be used, such as information Gain or using Gini Coefficient.

For larger datasets, most preferred is the Gini Coefficient. Gini index of each attribute is estimated and the attribute with largest reduction is taken as the root node. This procedure is repeated until the leaf node having the predicted class label is reached.

C. Majority Voting

The ensemble of classifiers is performed using majority voting rule which is suitable for datasets having two class labels(0 and 1)Majority voting also known as plurality voting uses the technique of highest number of votes for classifying a new set of data instances. In this step, all the three classifiers are combined together and a majority voting is applied to predict the class labels. The class label thus obtained will be the majority of class labels predicted by each individual classifiers. A Majority Classifier proved to be a well performing model, balancing out the individual weakness of classifiers thereby increasing the accuracy of overall model in heart disease prediction.

IV. EVALUATION

Efficiency of the overall model is measured using a tool known as confusion matrix that measures the performance of each classifier on a set of data for which the true values are known. It forms a matrix layout where each row of the matrix represents the variables in a actual class while each column represents the variables in an predicted class(or vice versa).Table II represents the result variables obtained from classification.

		Predicted	
		P	N
Actual	P	TP	FN
	N	FP	TN

TABLE II: Definition of confusion matrix

true positives (TP) : person is having the disease and correctly predicted the same

true negatives (TN) : person is not having the disease and correctly predicted the same

false positives (FP) : person is not having disease but incorrectly predicted with having disease

false negatives (FN) : person is having the disease but incorrectly predicted with no disease

TP, TN, FP, FN are the values that can be applied to evaluate the performance of classification against specificity, sensitivity and F1 score. Sensitivity identifies the people having heart disease whereas Specificity correctly identifies the people with no disease. F1score measures the performance of the test for positive class.

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

A. Classification Result

Following shows the accuracies obtained using each classifier. The result shows that ensemble classifier is having the higher rate of accuracy about 90.8% followed by Naive Bayes (86.4%), KNN(81.2%) and Decision Tree(79.06%). Hence, an ensemble classifier approach helps in the prediction of heart disease with greater accuracy than that of individual classifiers.

1. KNN Classifier

TP=106	FN=23
FP=24	TN=97

TABLE III : Evaluation performance of heart disease prediction using KNN

ACCURACY	SPECIFICITY	SENSITIVITY	F1 SCORE
81.2%	80.16%	82.17%	81.8%

TABLE IV : Evaluation performance of accuracy, specificity, sensitivity and F1score of KNN

2. Decision Tree Classifier

TP=98	FN=26
FP=25	TN=101

TABLE V : Evaluation performance of heart disease prediction using Decision Tree

ACCURACY	SPECIFICITY	SENSITIVITY	F1 SCORE
79.06%	80.1%	79.03%	79.35%

TABLE VI : Evaluation performance of accuracy, specificity, sensitivity and F1score of Decision Tree

3.Naive Bayes

TP=116	FN=20
FP=14	TN=100

TABLE VIII : Evaluation performance of heart disease prediction using Naive Bayes

ACCURACY	SPECIFICITY	SENSITIVITY	F1 SCORE
86.4%	87.71%	85.29%	87.21%

TABLE VIIII : Evaluation performance of accuracy, specificity, sensitivity and F1score of Naive Bayes

4.Ensemble Classifier

TP=117	FN=15
FP=8	TN=110

TABLE IX : Evaluation performance of heart disease prediction using Ensemble classifier

ACCURACY	SPECIFICITY	SENSITIVITY	F1 SCORE
90.8%	93.22%	88.63%	91.05%

TABLE X : Evaluation performance of accuracy, specificity, sensitivity and F1score of Ensemble Classifier

V. CONCLUSION

Early detection and treatment of heart disease can save human life up to a great extent. Existing methods were based on the experience and knowledge of healthcare professionals for which the accuracy is always limited. So ,this paper introduced a method based on data mining that helps the medical professionals to retrieve relevant information from patient database and thereby performing an early diagnosis of disease. An ensemble classifier approach , that is, the combination of three classifiers forming an ensemble is being implemented so that the overall model has a better predictive performance than that of individual classifiers. It can discover relationship between the various risk factors and heart disease by analyzing the patient records. It allows healthcares to identify people at higher risk of having a heart disease and provide them better care at affordable costs. Moreover, this paper performs a comparison between performance measures which include sensitivity, specificity, and accuracy. It shows that Ensemble classifier has outperformed with 90% accuracy in the prediction of heart disease as compared to that of individual classifiers. This model can also be extended to improve its performance in the medical field for diagnosing diseases.

VI. REFERENCES

1. Mackay,J., Mensah,G. 2004 "Atlas of Heart Disease and Stroke" Nonserial Publication, ISBN-13 9789241562768 ISBN-10 9241562765
2. Mythili T, Dev Mukherji, Nikita Padalia and Abhiram Naidu"A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)" International Journal of Computer Applications Volume 68-No.16, April2013
3. Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (IJCSNS), Vol.8 No.8, August 2008.
4. Nitin Kumari, Sunita, Smita "Comparison of ANNs, Fuzzy Logic and Neuro- Fuzzy Integrated Approach for Diagnosis of Coronary Heart Disease : A survey",International Journal of Computer Science and Mobile Computing Vol.2 Issue. 6, June- 2013, pg. 216-224
5. Latha Parthiban, R.Subramanian "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", World Academy of Science, Engineering and Technology International Journal of Medical and Health Sciences Vol:1, No:5, 2007 .
6. Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, pp 7675-7680.
7. Lamia Abed Noor Muhammed 2012 "Using Data Mining technique to diagnosis heart disease" Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on March 2012, pp 173-177.
8. Polat, K., Sahan, S., Kodaz, H., Gnes, S. 2005 "A new classification method to diagnose heart disease: Supervised artificial immune system". In Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN), pp 2186-2193..
9. Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15 J.
10. Ansari.A.Q. and Neeraj Kumar Gupta., Automated Diagnosis of coronary heart disease

using Neuro-Fuzzy integrated system. World Congress on Information and Communication Technology, pp 1383-1388, 2011.

11. Nidhi Bhatla and Kiran Jyoti., 2012. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering Research & Technology, Vol 1(8), pp 1-4, Oct 2012.
12. My Chau Tu, Dongil Shin, Dongkyoo Shin 2009 "Effective Diagnosis of Heart Disease through Bagging Approach" Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference, pp 1- 4D.
13. TJohn Peter and K.Somasundaram, Study and development of novel feature selection framework for heart disease prediction. International Journal of Scientific and Research Publications, Vol 2(10), 2012.
14. Shouman, M., Turner, T. and Stocker.R 2011 "Using Decision Tree for Diagnosing Heart Disease Patients" Australasian Data Mining Conference (AusDM 11) Ballarat 2011, pp 23-30 .
15. Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.