Heart Disease Prediction System using Data Mining Method

Keerthana T K [#1]

[#]PG student, Dept. of Computer Science, Jyothi Engineering College
Thrissur, Kerala, India

Abstract *— Heart disease is most common in present era. The treatment cost of heart disease is not affordable by most of the patients. So we can reduce this problem by a Heart Disease Prediction System (HDPS).It is helpful for earlier diagnosis of heart disease. Data mining techniques are used for the construction of HDPS. In health care field some systems use large healthcare data in varied forms such as images, texts, charts and numbers .But this datas are hardly visited and are not mined. This problem can be avoided by introducing HDPS. This system would enhance medical care and it can also reduce the costs. The system can handle complex queries for detection of heart disease and thus help to make intelligent medical decisions. This paper proposes a HDPS based on three different data mining techniques. The various data mining methods used are Naive Bayes, Decision tree (J48), Random Forest and WEKA API. The system can predict the likelihood of patients getting a heart disease by using medical profiles such as age, sex, blood pressure, cholesterol and blood sugar. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall. The overall system provides high performance and better accuracy.*

Keywords *— HDPS; WEKA; Random Forest; Naïve Bayes; J48*

I. INTRODUCTION

Heart is one of the important organs in blood circulatory system of all living organism. There are many elements which make problems to heart. They are smoking, poor eating methodology, high pulse, cholesterol and high blood pressure etc. The diagnosis of heart disease in earlier stage is a challenging problem for the medical industry. Data mining based heart disease prediction system can help in determining the heart disease during early stages. The prediction system helps to reduce the high risk of heart disease. Prediction is done based on the current data given to the system. For building Heart Disease Prediction System we use WEKA tool. It is open source data mining software in Java. Here system is being developed using three different data mining techniques; Nave Bayes, J48, Random forest with WEKA API. Here different classification algorithms analyses the input dataset and accuracy is compared for analysis. Cleveland heart disease dataset with 14 attributes and 303 instances is the training dataset used for analysis.

II. RELATED WORK

Diagnosis of heart disease is a complicated task in medical field. So it is needed to develop an efficient disease prediction system for the earlier detection of disease.

One of the earliest systems for heart disease detection was proposed by Meghna Sharma et al. and they propose a hybrid technique in data mining for heart disease prediction. Here a prototype which can extract unknown data related with heart disease from a past heart disease database record is developed. They put an idea of hybrid technique methodology which can be implemented in future to have accuracy of almost 99% or with least error.

In [2] authors analyses various papers on heart on using different data mining technologies. Also they make comparative study on the performance of three classifiers like Naïve Bayesian classifier, Decision trees and Probabilistic Neural Network (PNN). The analysis showed that artificial neural networks gave accuracy of 94.6 percentages in heart disease prediction.

In [3] author developed an intelligent Heart Disease Prediction System. This system is builds using Naive Bayes algorithm and it also uses a smoothing technique (Jelinek mercer smoothing) to improves the performance. Here the system is proposed using Cleveland heart disease database as the input dataset. Each attribute of the dataset were fed to the Naive Bayesian classifier and it produces the prediction results based on the classification process. We can conclude that efficiency can be improved with the use of smoothing technique. This model could answer complex queries which traditional decision support systems cannot.

In [5] Bhuvaneswari Amma proposes a system using genetic algorithm and neural network, which is helpful for cardiovascular disease prediction. Observed demerits of this system are: less accuracy and no extraction of hidden data.

III. PROPOSED MODEL

The proposed architecture of heart disease prediction system is given below.

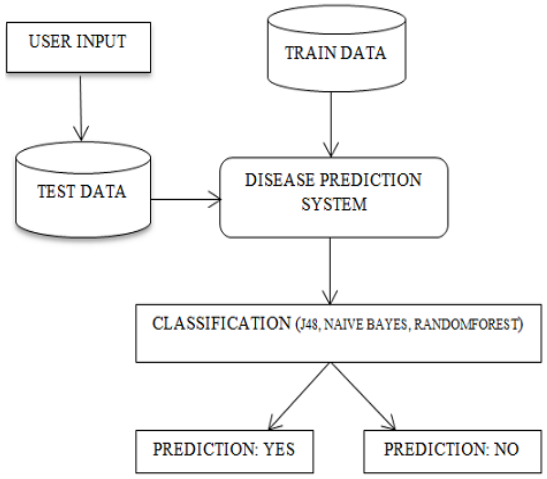

Fig 1: Proposed model

It consists of training dataset and user input as the test dataset. Weka data mining tool with api is used to implement the heart disease prediction system. The source code of Weka is in java. The system is designed with java swing and use Weka api to call the different methods of Weka. The components used are instances, different classifiers and methods for evaluation. Supervised learning method is used here. A supervised learning algorithm analyses the training data and deduces a function from the labelled training set. It can be used for mapping new examples. The training data obtained from cleveland heart disease database is the training example. This training data consist of the class label and its corresponding value. Naive Bayesian, J48 and Random Forest classifiers are supervised learning algorithms. They learn from the provided training examples. When a new instance with same attributes as in training data with different values other than those in the training example comes, these algorithms correctly classify the new instance based on the generalization created from the training set. Naive Bayesian, J48 and Random Forest classifiers are classify the new observation into two categories on the basis of training dataset. The training dataset is in the ARFF format. The training set consists of 14 attributes including the class attribute. Heart disease prediction system accepts input from the user through a graphical user interface. All the attributes needed for classification is received from a text field. The graphical user interface is built using swing. The next process is to transfer the user input obtained from graphical user interface into a file of CSV (Comma separated Value) extension. Then the CSV file is converted into ARFF file. Weka api provide native methods for converting from CSV to ARFF. The converted user input is treated as test data. The test data set will contain all the attributes of training dataset. If the user did not enter an attribute value a '?' will be assigned at the value of that corresponding attribute. Weka will handle this missing value. This test data is run on Naive Bayes algorithm, Random Forest algorithm, and J48 algorithm. These algorithms classify the instances received from the user and predict the chance to have heart disease. Netbeans IDE is used to code in Java.

A. Naïve Bayes algorithm

Naïve Bayes algorithm is a classification method based on Bayes theorem (1). It is a probabilistic classifier.

Bayes theorem: consider X and H

X: is an evident, X=x1, x2...,xn

H: is the hypothesis

Then

$$P(H/X) = (P(X/H)*P(H))/P(X) \quad (1)$$

B. J48

J48 algorithm is a decision tree that builds a tree by using pruning method. Pruning method is used for getting real trees by removing redundant data. The resulting tree is small in size. By J48 algorithm the large datasets can be divided in to smaller sets. So it reduces the complexity and improves the performance of classification.

At first it will set the root node with high information gain. Then remaining nodes are also taken based on the information gain. The tree is built until the clear classification is done. J48 method provides new data addition rather than other decision tree methods.

C. Random Forest

Grow a forest of many trees. Grow each tree on an independent sample from the training data.

At each node,

1. Select k variables at random out of all K possible variables.
2. From the selected m variables, find the best split
3. Grow the tree to maximum depth (classification).
4. Average the trees to get prediction for new data.

IV. EXPERIMENTAL RESULT

This section describes the performance proposed heart disease prediction system. Experiments show that the proposed method gives the accurate diagnosis of heart disease than the existing methods.

A. TABLE I

NAÏVE BAYES RESULTS

Parameter	Value
Correctly classified instances	253
Incorrectly classified instances	50
Mean absolute error	0.1846
Root mean squared error	0.3634
Kappa statistics	0.6661
Relative absolute error	0.372065
Total number of instances	303
Root relative squared error	0.729767
Precision	83.33%
Recall	0.7971
Error rate	0.1650

B. TABLE II

J48 RESULTS

Parameter	Value
Correctly classified instances	236
Incorrectly classified instances	67
Mean absolute error	0.2589
Root mean squared error	0.431
Kappa statistics	0.5508
Relative absolute error	0.5219
Total number of instances	303
Root relative squared error	0.8654
Precision	78.4%
Recall	0.7101
Error rate	0.2211

C. TABLE III

RANDOM FOREST RESULTS

Parameter	Value
Correctly classified instances	251
Incorrectly classified instances	52
Mean absolute error	0.2668
Root mean squared error	0.3549
Kappa statistics	0.6536
Relative absolute error	0.537772
Total number of instances	303
Root relative squared error	0.712652
Precision	82.8%
Recall	0.823
Error rate	0.221122

V. CONCLUSION

This paper introduces a heart disease prediction system for diagnosing heart disease in earlier stage. The system uses data mining techniques such as Naïve Bayes, J48, and Random Forest along with Weka api to call different methods of Weka. The classification process inside the system is performed with attributes like age, sex, heart beat rate, cholesterol level etc. The prediction is then made based on this classification results. Here the machine learning capability of the computer system can be extended into the medical field. The proposed system is best for reducing the error occurrence during the disease prediction. In this paper the precision and accuracy of three different classifiers are measured. The result shows Naive Bayesian classification possesses high precision and less error rate. Random Forest classification method produces better result than J48 classification.

ACKNOWLEDGMENTS

I take this opportunity to express my heartfelt gratitude to all respected personalities who had guided me and also The Lord Almighty for guiding me in this endeavor.

REFERENCES

[1] Ankita Dewan and Meghna Sharma "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classifıcation" IEEE2010

[2] Majali J, Niranjan R, Phatak V,Tadakhe O.Data mining techniques for diagnosis and prognosis of cancer. International Journal of Advanced Researchin Computer and Communication Engineering.2015;4(3):613–6.

[3] Ms.Rupali R.Patil "Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing" IJARCCE 2014

[4] Monika Gandhi and Dr. Shailendra Narayan Singh "Predictions in Heart Disease Using Techniques of Data Mining" International Conference Conference on Futuristic trend in ComputationalAnalysis and Knowledge Management (ABLAZE-2015)

[5] Bhuvaneswari Amma N.G.," Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network" 2014

[6] Chaitrali S. Dangare et. al.,"Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", (IJCA) (0975 – 8887), Vol. 47, No. 10, June 2012, page no. 44-48.

[7] S. Vijiyarani et. al., "An Efficient Classification Tree Technique for Heart Disease Prediction", (ICRTCT - 2013) Proceedings published in (IJCA) (0975 – 8887), 2013, page no. 6-9.

[8] Y.E.Shao, C.-D. Hou, and C.-C. Chiu,"Hybrid intelligent modelling schemes for heart disease classification, "Applied Soft Computing, vol. 14,pp. 47–52, 2014.

[9] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.

[10] Guru, N., Anil, D., Navin, R., Decision Support System FoR Heart Disease Diagnosis Using Neural Network.Delhi Business Review. **8**(1): (2007).

[11] Shouman M, Turner T, Stocker R. Using decision tree for diagnosing heart disease patients. Proceedings of the 9th Australasian Data Mining Conference(AusDM'11); Ballarat, Australia. 2011. p. 23–30.

[12] RaniKU.Analysis of heart diseases dataset using neural network approach.IJDKP. 2011; 1(5):1–8.