Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms

Ajit Solanki, Mehul P. Barot

Abstract - In the survey paper, different techniques of mining for forecasting of heart risk are discussed. Heart disease cause millions of death every year, It's rapidly increasing. Mining methods are too much helpful detect and diagnose heart risk. Data mining in medical domain has great potential to uncover the pattern which are hidden in the medical dataset [2]. For this reason, different mining methods can be used to abstract knowledge for forecasting heart disease [4]. In this paper, survey is carried on various single data mining techniques and hybrid mining techniques to identify the best suited technique to achieve high accuracy in prediction of heart disease [5]. Here, Potential of many classification techniques was evaluated, namely Naïve Bayes, SVM, Decision tree, K-nearest neighbour, and even hybrid approach of classifiers. Analysis on various methods proved that techniques based on classification obtain high accuracy compared to previous methods [14].

Keywords- Data mining, Classification, Disease Diagnosis, prediction, Accuracy

I. INTRODUCTION

According to World Health Organization, Hypertensive heart attack is one of the prime causes of death. [1] Health Care industry has huge amount of data, which is not mined. It has hidden patterns which is necessary for prediction of heart disease risk. We know, heart is very important part of human body. If, the organs of the body that is brain and heart there is an insufficiency in blood circulation, heart stop working immediately and it causes death. The risk parameters associated with heart risk are age, family history, hypertension, high cholesterol, diabetes, smoking, tobacco, alcohol consumption, obesity, poor diet and chest pain [15].

Even about sixty percent of total population are suffering from the heart disease, so detection of the heart disease earlier can prevent the heart disease. In many cases heart disease almost noticed at the final stage or after death. It's difficult to cure heart disease at the final stages so people are very reluctant to treat at the early stages of heart disease [2].

Health care industries have very huge amount of data and which is not mined to discover hidden pattern. Solution to this problem is data mining techniques. It is the process of analysis of large dataset and extraction useful information from it. Mining techniques are:

Manuscript published on 30 January 2019.
* Correspondence Author (s)
Ajit Solanki, Research Scholar, Department of Computer Engineering, LDRP Institute of Technology and Research, Gandhinagar, Gujarath, India. (e-mail: ajit.solanki02@gmail.com)
Mehul P. Barot, Assistant Professor, Department of Computer Engineering, LDRP Institute of Technology and Research, Gandhinagar, Gujarath, India. (e-mail: mehulce@gmail.com)

1. Association Mining
2. Classification Techniques
3. Clustering Techniques

Association rule mining is a method to discover interesting relationship between variables in large set of data.

Classification is used to extract a model describing useful classes. There are many classification techniques which are used in diagnosis of heart disease, like Decision tree, Naïve Bayes, SVM, K- Nearest neighbour and many more.

Clustering is procedure of grouping similar features datasets into cluster. K-mean algorithm is used for clustering. It is faster than other [2].

In this paper different kind of classification methods which are applied in the forecasting of heart disease has been talked and also comparison made on the classifiers to justify which algorithm achieve the high accuracy.

II. BACKGROUND THEORY AND RELATED WORK

In this section, we will discuss about the various classification techniques used to diagnose the heart disease.

A. Support Vector Machine (SVM)

Classification technique, which uses hyper planes which achieves largest distance of two classes. It overcomes the high dimensionality problems.[5]

In 2010, Youn - jung son and Hong-Gee Kim used SVM for heart failure diagnosis. It was performed on data with 11 variables and author achieved 77.63% high accuracy using this model.

B. Decision Tree

Decision tree has tree like framework. It divides dataset to small sets. Leaf node represents the decision. while top most node is the root node. [5]

In 2013, Vikas Chaurasia and Saurabh Pal gave many mining methods used for heart disease prognosis. They used various classification techniques like Naïve Bayes, Decision Tree and Bagging Algorithm. DT is very easy to understand but is too much sensitive towards noise.

C. K- Mean

K-means is a technique of iterations. It divides the n data objects into the K- clusters. Here, K is any predefined value to cluster the data into k clusters. K means places the object closest to cluster center as per Function of Euclidean Distance. This algorithms computation is fast but it is difficult to find value of K. [5]

D. Naïve Bayes

This algorithm has roots in Bayes Theorem with few Assumptions. It is based on the conditional probability. Main advantage of this algorithm is that it requires small dataset. [5]

In 2012 Shadab Adam Pattekari and Asma parveen both offered the Naïve Bayes to diagnose heart risk. This system provides effective result for prediction of heart disease.

In 2017, Sushmita Manikandan used Gaussian Naïve Bayes for the classification. It gives an accuracy of 81.25% [1].

E. Neural Network

NN is an Iterative method which works on only non-linear dataset. In neural network random weight is assigned to each of the input then calculation and comparison is made of corresponding output. Artificial Neural Network gives the best accuracy with non-linear data but has limitations for linear datasets. [5]

F. MLP (Multi-Layer Perceptron)

MLP uses backpropagation for classification. MLP Architecture is of three layers input, Hidden and Output layer. This network has neurons as node. Here input dataset is given to input layer further process is carried out by hidden layer and output layer generates the decision. [5]

In previous efforts, many classification models are compared, Naïve Bayes showed highest accuracy of 81.25% for 14 attribute dataset of heart disease [1] and SVM showed the accuracy of 83% when applied on dataset of National Health and Nutrition Exam Survey [1]. There are many mathematical models were created and analyzed on earlier work.

Also in 2017, Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M has proposed the algorithm Decision Tree by successively applying of Genetic algorithm on dataset to lower the real size of data to get favorable set of attributes which are accepted for prediction of disease [2].

Ankita Dewan and Meghna Sharma in 2017 used Back Propagation but it is only used for Non-linear datasets. They have not achieved maximum profit using this technique because of its drawback to get stuck in local minima. To solve this problem, they used genetic algorithm's mutation and cross over phenomenon. The weights which are used for Back propagation need to be optimized before given as input to network [3].

Gnaneswar B. and Ebenezar Jebarani M.R. conducted survey on various mining techniques for diagnosis of the heart disease. They found some of the limitations of algorithms like Neural network only works well with linear datasets. Also, Decision Tree performs poor with large datasets [5].

B. Jin*, Senior Member, IEEE , C. Che*, Z. Liu, Shulong Zhang, Xiaomeng Yin and X.P. Wei in in 2017 proposed a predictive model using Lon Short-term Memory Network (LSTM) methods. It has proven to give superior performance compared to Linear Regression, Random Forest and AdaBoost. In data analysis they preferred One-hot encoding and word embedding vectors to show person diagnosis events [7]. In 2017, Kanika Pahwa and Ravinder Kumar proposed hybrid method of selecting features to optimize classification issue. This approach reduces the size of data along with enhancement in the performance of hybrid classifiers SVM-RFE and Naïve Bayes- Random Forest [8].

Rashmi G. Saboji and Prem Kumar Ramesh, in 2017, implemented algorithm of Random Forest for heart disease prediction, they able to achieve 98% accuracy, but with very small dataset of 600 instances [10].

Meenal Saini, Niyati Baliyan, and Vinita Bassi proposed Hybrid models using Bagging and Boosting and proved that Hybrid model achieves higher accuracy than normal classifiers like Linear model and Support Vector Machine [11].

In 2016, Theresa Princy R. and J. Thomas used KNN and ID3 algorithm to detect risk of heart disease by using various number of attributes, it achieved efficiency of model by adding important attributes to datasets [13].

III. FRAMEWORK OF EXISTING SYSTEM

There are many mining algorithms used for the prediction of heart disease in past. In existing model, researchers used dataset as input which may or may not be appropriate format. To make compatible they applied pre-processing techniques like data cleaning, data transformation, data reduction and many more on dataset. Afterwards they applied various data mining approaches to achieve high efficiency of model like classification, clustering and association rule mining.

Use of various simulation tools is carried out on pre-processed dataset using different algorithms. Researchers used single classifiers as well as hybrid classifiers by combining two or more than two classifiers to achieve efficacious accuracy for diagnosis of the heart disease.

Framework of the existing methods for prediction of Heart Risk is as given below:

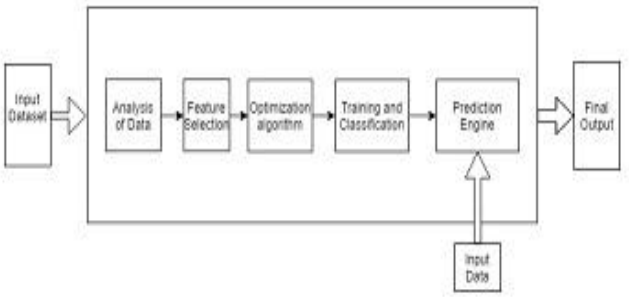

Figure 1: General Architecture of Existing Algorithms [16]

IV. PERFORMANCE EVALUATION AND RESULTS

Performance of the algorithm can be evaluated on the basis of various parameters like accuracy, Confusion matrix, precision and many more.

A. Accuracy measure

Accuracy measure mentioned to the how many percent of right prediction achieved using proposed method by comparing it with actual real classification made to the test dataset. It's a calculation of model's efficiency to perfectly label the unknown data.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Say, if data is categorical accuracy can be measured as the rate at which data will be labelled with true category of data. While, if data is continuous, accuracy can be measured by the distance between predicted value and the correct value.

B. Confusion Matrix

It is measured by the count of right and wrong prediction made by model in comparison with real classifiers in case of test dataset. This matrix would of n*n, n is count of classes.

V. COMPARISON OF EXISTING METHODS & RESULTS

This study has made the comparison between various methods used for the prediction of heart disease. Table given below shows the comparison of various methods used for the diagnosis of heart disease:

	Author	Algorithm	Accuracy	Observation
1.	Sushmita Manikandan	Naïve Bayes	81.25%	Gaussian Naïve Bayes algorithm was used for the classification on 14 attributes dataset.
2.	Youn-jung Son , Hong-Gee Kim	Support Vector Machine	77.63%	SVM obtained the accuracy of 77.63% on dataset of 11 variables but speed of SVM is very slow.
3.	Kemal Polat, Seral S ahan	K- nearest Neighbour	87%	Before use of main algorithm they used K- nearest neighbour model in pre-processing step.
4.	Kanika Pahwa , Ravinder Kumar	Naïve Bayes Random Forest	84.1584% 84.1604%	They proposed the approach of feature selection which does not only reduce the size but also increase the efficiency. For Naïve Bayes they have selected 10 attributes and 12 for Random forest
5.	K. Srinivas et al	SVM, DT, MLP	82.5% 82.5% 89.7%	Scholars Applied various classifiers on dataset, MLP achieved the highest accuracy of 89.7% but it's slow in performance. And it can only be applied for linear data sets.

Figure 2: Comparison of Existing Algorithms

VI. FUTURE WORK

From the survey we made on various research papers on diagnosis of heart disease we understand that more research work can include use of hybrid approach on heart disease dataset, also to use an efficient optimization technique is to decrease the real data size for the optimal set of attributes which are allowable for heart disease forecasting. To develop an algorithm to achieve high efficiency using both kind of dataset (linear and non-linear) with low computation cost.

VII. CONCLUSION

The main aim of this paper is to provide an insight of heart disease risk diagnosis using classification techniques. From the analysis, many authors used various techniques of classification using different number of attributes for study. It has been proven that for heart disease diagnosis classification algorithms achieve high efficiency compared to other. From study it has been concluded that Multilayer Perceptron achieved the highest accuracy, but drawback of MLP is that it is very slow in performance and it can only be applied for linear data set. In future, risk of heart failure can be diagnosed using less number of attributes and also accuracy can be enhanced using some other algorithms.

REFERENCES

1. Sushmita Manikandan, "Heart Attack Prediction System" International conferenceon Energy, communication, Data Analytics and Soft Computing. (ICECDS-2017)
2. Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017
3. Ankita Dewan, Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 978-9-3805-441 6-8/15/$31.00 c 2015 IEEE
4. Monika Gandhi, Dr Shailendra Narayan sinh, "Predictions in Heart Disease Using Techniques of Data Mining", 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015)
5. Gnaneswar B., Ebenezar Zebarani M.R., "A review on prediction and diagnosis of heart failure", 2017 (ICIIECS)
6. M.A.Jabbar, Shirina samreen, "Heart Disease prediction System based on Hidden Naïve Bayes Classifier"
7. B. Jin*, Senior Member, IEEE , C. Che*, Z. Liu, Shulong Zhang, Xiaomeng Yin and X.P. Wei, "Predicting the risk of Heart Failure with HER sequential data modelling", DOI 10.1109/ACCESS.2017.2789324, IEEE Access
8. Kanika Pahwa, Ravinder Kumar. "Prediction of Heart Disease using Hybrid Technique for selecting Features" 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, Oct 26-28, 2017
9. Purushottam, Prof. (Dr.) Kanak Saxena,Richa Sharma, "Efficient Heart Disease Prediction system using Decision Tree" International Conference on Computing, Communication and Automation (ICCCA2015)
10. Rashmi G Saboji, Prem Kumar Ramesh, " A Scalable Solution for Heart Disease Prediction using Classification Mining Techniques" International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)
11. Meenal Saini, Niyati Baliyan, Vineeta Bassi,,"Prediction of Heart Disease Severity with Hybrid Data Mining." 2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)
12. Jayshril S. Sonawane, D.R. Patil, "Prediction of Heart Disease using Multilayer Layer Perceptron Neural Network", ICICES2014 - S.A.Engineering College, Chennai, Tamil Nadu, India
13. Theresa Princy R., J. Thomas, "Human Heart Disease Prediction System Using Data Mining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]
14. C. Sowmiya, Dr. P.Sumitra, "Analytical Study of Heart Disease Diagnosis using Classification Techniques", 2017 IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL,OPTIMIZATION AND SIGNAL PROCESSING

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)