SURVEY ON HEART DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUE

AMRUTA POWAR

Department of Computer Science and Engineering, DYPCET, Shivaji University,Kolhapur, Maharashtra, India. amrutapowar15@gmail.com

PROF. DR. VIJAY GHORPADE

Department of Computer Science and Engineering, DYPCET, Shivaji University, Kolhapur, Maharashtra, India. vijayghorpade@rediffmail.com

ABSTRACT

According to WHO maximum death in worldwide are happened due to heart disease. So, advanced data mining techniques can be used to take out hidden patterns from healthcare industry, so that it becomes easy to predict heart disease. In this paper we consider the concepts used for prediction of heart disease using data mining techniques. Human heart disease Prediction using various data mining techniques, their comparative result and their limitations are discussed here. This paper gives an idea about the key topics for heart disease prediction system which will help in predicting heart disease so that diagnosing it can take less medical tests and provide effective treatments.

KEYWORDS: Naive Bayes, Data Mining, Heart Disease, k means.

I. INTRODUCTION

Data mining deals with finding the relationships and global patterns from large databases which are unseen among large amounts of data. These days healthcare industry has large amounts of patient data but these data are not mined in order to give some hidden information, and thus to make effective decisions, advanced data mining techniques can be used. Using medical profiles such as age, gender, chest pain type, fasting blood sugar, resting blood pressure, resting electrographic results, cholesterol, maximum heart rate achieved, exercise induced angina, old peak, slope, number of vessels colored, defect type, obesity and smoking we can predict the probability of patients getting a heart disease.

Quality of service is the serious challenge that the healthcare industry has face to. Quality of service deals with diagnosing disease correctly & provides efficient treatments to patients. Heavy loss can happen because of poor diagnosis. Diagnosis is an important task that must be executed correctly and efficiently. The diagnosis is always based on doctor's experience and knowledge. This leads to unwanted results and excessive medical costs of treatment provided to patients.

Therefore an automatic medical diagnosis system needs to be designed that can take benefit of heart disease database which is publicly available.

Our paper is organized as follows: Firstly explanation of some data mining techniques used for heart disease prediction system then conclusion.

II. DIFFERENT DATA MINING TECHNIQUES TO PREDICT HEART DISEASE

1. Sellappan Palaniappan [2] has recommended intelligent heart disease prediction system with the use of data mining techniques.

This paper uses three data mining classification modeling techniques namely Naïve Bayes, Decision Trees and Neural Network and developed a prototype called heart disease prediction system. This system extracts hidden knowledge from heart disease database. They used DMX query language and functions which is used to build and access the models. The models are trained and verified against a test dataset. To evaluate the effectiveness of the models two methods namely classification matrix and lift chart are used. All three models are intelligent to extract patterns in response to the predictable state. Their result shows that Naïve Bayes model most effectively predict the patients with heart disease then it is followed by Neural Network and Decision Trees.

Authors defined five mining goals as follows:

Goal 1: is to predict those who are likely to be diagnosed with heart disease.

Goal 2: is to identify the interaction in the medical inputs related with the predictable state – heart disease.

Goal 3: is to identify the impact among the medical attributes related with the predictable state – heart disease.

Goal 4: is to identify characteristics of patients having heart disease.

Goal 5: is to decide the attribute values that distinguish between patients having heart disease and patients with no heart disease.

These 5 goals are evaluated against the trained models.

According to result Naïve Bayes possibly will answer four out of the five goals; Decision Trees, three; and Neural Network, two.

Some limitations of this HDPS system are as follows: First the list of used attributes needs to be extended to provide a more broad diagnosis system. Another limitation is that it only uses categorical data. For some diagnosis, continuous data can be used. The small size dataset is used. For better result large dataset can be used.

2. Aigerim Altayeva [3] has proposed medical decision making diagnosis system by integrating Naïve Bayes and k-means algorithms.

In this paper, with the use of Naïve bayes data mining technique we can find many patterns so that in future it can be used in various areas such as decision-making, we can also predict value for the prediction and calculation. Authors proposed a Decision Support in Heart Disease Prediction System (HDPS) using Naïve Bayes data mining technique and K-means clustering algorithms where the initial choice of the centroid strongly influences the final result. It also finds various methods of initial centroid selection of the K-means clustering such as inlier, range, outlier, random attribute method, and random row methods in the diagnosis of the patients.

This paper explores the combination of K-means clustering with naive Bayes in predicting heart disease patients. Finally the result shows that it improves accuracy in diagnosis of the patient. In this system they used human's knowledge and artificial agents.

In the experiment, they compared the accuracy of each method, for the number of clusters from 2 clusters till 5 clusters and showed that the number of clusters could increase their accuracy. And the results show, most of the methods return high accuracy for the clusters number 2 or 3.

3. Syed Umar Amin [4] has introduced genetic neural network based data mining in predicting heart disease by using risk factors [4]

This paper developed an intelligent data mining system based on genetic algorithm optimized neural networks for the prediction of heart disease on the basis of major risk factors. This technique contains two most useful data mining tools namely genetic algorithms and neural networks. This system is executed in Matlab and result shows that 89% accurately it can predict the risk of heart disease.

The dataset contains 12 important risk factors like sex, age, hereditary, blood pressure, blood cholesterol, Smoking, use of alcohol, physical inactivity, diabetes, poor diet, obesity and stress. The data for 50 people was collected from American Heart Association.

This system uses back propagation algorithm for learning and training the neural network, but there are two limitations with back propagation algorithm. First is that the initialization of the NN weights is a blind process hence there is a possibility that the network output would run towards local optima and because of this the overall tendency of the network to find out a global solution is affected. Second problem is that there is a possibility that network never converges because back propagation algorithm is little slow in convergence.

We can use this system in clinical decision support and as a risk indicator so that it helps people to reduce the risks of having any heart disease in future. Results show that genetic algorithm and neural network gives better prediction accuracy than the traditional ANN.

4. S. Radhimeenakshi [5] has used two data mining technique namely support vector machine and artificial neural network and proposed classification and prediction of heart disease risk.

The purpose of this paper is to break the use of AI devices for order and expectation of heart diseases. This work uses the classes of Heart Disease utilizing Support Vector Machine (SVM) and Artificial Neural Network (ANN).

Experiment is completed among two strategies based on accuracy and training time. The author introduces a medical choice backing framework for coronary illness characterization in a purpose, precise and fast manner. Here Cleveland Heart Database and Statlog Database are used which are taken from UCI Machine learning dataset. Here proposed system model arranges the data records into two classes in Support Vector Machine as well as Artificial Neural Network. And finally analyzes the performance of the both the datasets. This work includes pre-processing, preparing, training and testing with individual models, evaluation of results and the prediction of heart disease risk and is implemented in Matlab R2010.

The performance evaluation of support vector machine has shown great percentage in terms of accuracy and precision like 84.7% and 85.06% respectively. In this system the authors have utilized maximum records for training and testing. The testing procedure can prove the effect in the terms of specificity, precision, accuracy and sensitivity. A result shows that SVM is more viable than the ANN. But ANN performs better in terms of specificity and adaptability to the extensive training. The authors compared the output obtained from SVM classification with the output from ANN classification and found that ANN are extremely empowering.

5. C. Kalaiselvi [6] has recommended diagnosing of heart diseases by using average k-nearest neighbor algorithm of data mining.

This paper proposed a new approach called average k-nearest neighbor algorithm which is used to get improved efficiency and classification accuracy. The dataset used for prediction is obtained from UCI machine learning repositories. The main purpose of this proposed system is to detect heart disease with less number of attributes that are relevant to heart diseases.

The proposed system uses KNN algorithm which is used to categorize the data using training examples available in feature space. The k nearest training set data are found for every row of testing data and then classification is done by finding the majority vote by breaking the ties at random. To remove the drawback of voting like frequent classes dominates the prediction of new example a faster algorithm AKNN is proposed. In Average KNN, super sample is created for each class. When the test samples are given the AKNN searches the sample data and find the closest neighbor to the input. To identify closest neighbor, Euclidean distance for every training sample is computed. And finally classification of sample is done based on majority of class among the neighbors.

Result gives higher efficiency and reduces complexity based on attribute reduction. It reduces the cost for different medical tests and helps the patients to take precautions in advance.

III. CONCLUSION

The main motivation of this paper is to review various data mining techniques that can be used to predict heart disease. It is seen that many authors use various techniques for heart disease prediction with the use of different number of attributes. Some techniques are naïve bayes, artificial neural network, support vector machine, decision tree, and K-nearest neighbor algorithm and so on. Each system has its own advantages. Hence different methods give different precision depending on the attributes used.

REFERENCES

I. Blake, C. L., Mertz, C. J. :"*UCI Machine Learning Databases*" http:// mlearn.ics. uci.edu/ databases/ heartdisease

II. 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

III. Aigerim Altayeva; Suleimenov Zharas; Young Im Cho. "*Medical decision making diagnosis system integrating k-means and Naïve Bayes algorithms.*" 2016 16th International Conference on Control, Automation and Systems (ICCAS), oct 2016.

IV. Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg *"Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors." 2013 IEEE Conference on Information and Communication Technologies (ICT 2013)*

V. S. Radhimeenakshi *"Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network." 2016 International Conference on Computing for Sustainable Global Development (INDIACom)*

VI. C. Kalaiselvi **"***Diagnosing of heart diseases using average k-nearest neighbour algorithm of data mining***"** *2016 International Conference on Computing for Sustainable Global Development (INDIACom), October 2016*