Data Mining in Healthcare using Hybrid Approach

Monica Sharma
PG Student
Department Of Computer Science and Engineering
Lovely Professional University, Punjab

Rajdeep Kaur
Assistant Professor
Department Of Computer Science and Engineering
Lovely Professional University, Punjab

ABSTRACT
Modern medicine generates a great deal of information stored in the medical database. Extracting useful data and making scientific decision for diagnosis and treatment of disease from the database increasingly becomes necessary. We propose a Heart diseases Prediction System for the society to prevent the cause of the death. So we are analyzing heart disease patient to identify which treatment is most effective one and provide better result.

Keywords
Cardiac autonomic neuropathy, Cardiovascular disease, Naïve bayes, Decision table.

1. INTRODUCTION
Nowadays healthcare are enough capable to generate and collect large amount of data. Due to the generation of huge amount of data it may require most frequent way to extract this healthcare type of data when needed. Using data mining approaches, it is possible and useful to extract interesting and meaningful information and their regularities. To acquire this type of knowledge using mining it can be used in various areas so as to improve the work efficiency or performance and extend the quality of decision making process. As there is much need for the coming generation related to computer theories and tools to help people in extracting the useful information from the continuously growing of large data.

Day by day information technologies are increasing to implement various problems in healthcare organization in order to take action as needed by the doctors in taking decision making actions usually. Various technologies have developed number of tools of data mining which are useful to manage the limitation of the people as occurred errors due to fatigue and offer them the indication for decision making process. The main goal of data mining technique is to identify relationship, different patterns and model which provide support in medical health. These can be referred as predictive model and they have been included into information system of various hospitals as a representation or prototype as to make decision, reduce content and time for decision making. Using information in healthcare enables the management of medical data and its safe switching between the users and the providers of various healthcare services[10].

1.1 Heart Diseases Patient Chosen From Healthcare For Prediction
Heart Diseases remain the biggest cause of deaths for the last two epochs. Recently computer technology develops software to assistance doctors in making decision of heart disease in the early stage. Diagnosing the heart disease mainly depends on clinical and obsessive data [4]. Prediction system of Heart disease can assist medical experts for predicting heart disease current status based on the clinical data of various patients [2].In biomedical field data mining plays an essential role for prediction of diseases. For diagnosing, the information which has been provided by the patients may include similar data and interrelated symptoms and shows especially when the patients suffering from more than one type of diseases of the similar category. The physicians are not capable enough to diagnose it correctly.

1.2 Data mining
Data Mining is referred as to extract the valuable or useful data from the huge amount of dataset. It is said that it is defined as to mine the knowledge from complex data and mined information can be used for various applications.

1.3 Need of data mining
Data is most important assets of the organization but finding out useful information from the data is a complex task for finding useful information from given data set we have to apply data mining techniques. Data mining is the field where we study and research techniques for the mining data more effectively and efficiently give more realistic information which provides help in decision planning for example a clothing company preformed mining on his yearly sale and find out in which month he have to provide offers to the consumer which increase overall profit. Today we have lot of unstructured data we need such efficient techniques which help providing efficient analysis of this data [3]. Data mining provides is method which helps us mapping unstructured data to structured data with the help several techniques Data Cleaning, Data Integration, Data Transformation. These Techniques provide us information which helps in various fields of our daily life for decision making for in all the business and educational field it reduces the un-useful information and provide us authentic information which is necessary for the decision making because storing the data is costly task.

1.4 How data mining works
By observing and analyzing the data of patient, various data miners experts uncover the hidden patterns or facts .For example in Washington, D.C hospital wants to identify why their patient got soon sick after discharge, after many researches data mining technique reveals the fact that the patient who are staying in the same hospital room afterwards developing the same infection. It was an example to show how various data mining technique helps us to analyze and solve the problems in healthcare. Commonly, data miner experts use the method called cross industry standard process for data mining i.e. CRISP-DM [5].

It involves six steps are as follows:

- **Understanding the business-**Identifying the objective and the requirements from the perspective of business and defining the problem of data mining.
- **Understanding the collected data-**Collecting the raw data, study the data and looking into it for any problem related to data quality.
- **Data preparation-** Build the evaluating dataset from the collected raw data.
- **Modeling-**Different Data mining software were used for the purpose of analyzing.
- **Evaluation-**Evaluate the accomplishment of the objective by comparing data mining model and their result.
- **Deployment-** Implementing the result.

1.5 Benefits of healthcare in Data mining

In healthcare, Data mining has become more and more popular as it offers benefits to patients, health, organization, care provider, researchers, insurers.

- **Patient-**They receive better and more affordable healthcare. Healthcare manager use data mining technique for identifying and tracking the chronicle disease and high risk patient, reduces the number of hospital admission and claiming.
- **Healthcare Organization-**In this mining influence cost revenue and operating efficiency. It provides information by guiding the patient interaction, by giving patient preferences, usage pattern, current and future requirement all these helps in improving patient satisfaction.
- **Care provider-**They have used various data analyzing technique to identify the useful treatment and best practices.
- **Insurers-**They can detect insurance fraud and abuse with the help of using the mining by creating norms and then identify unusual claims pattern

2. PROBLEM

In today's world people want to live very luxurious life so they work hard in order to earn lot of money and live comfortable. Due to this, people forget to care of themselves which result in the change in their lifestyle and food habits which led's to high blood pressure, sugar problem at very young age. They don't even worry if they are sick neither go for their own meditation. As a result of these, it led's to major problem called heart diseases. As in human body heart is most essential it may spoil the human health system. Therefore it is very important to diagnose the heart diseases. Due to availability of huge amount of data, the information can't be retrieved easily, so data mining approaches are implemented in order to extract knowledgeable information for the survival of patient or to analyse major cause of disease.So, we have proposed system in which hybrid method is used which involves combination of naïve bayes and decision table so as to improve the performance i.e. accuracy.

3. METHODOLOGY

In this proposed work, a heart disease prediction system has to be developed so that it can used for extracting useful information on the basis of symptoms. Due to availability of huge amount of unstructured data on different type of diseases the information can't retrieved easily and also data mining approaches can't apply on all amount of database. The hidden relationship on different diseases and their causes are not easy to extract from unstructured information.

3.1 Data source

Dataset were mainly collected from UCI repository and from various hospitals of heart diseases. Around 1080 patients data were collected which contains 50 attributes but we have selected only 14 of them in order to obtain the accurate results.

3.2 Preprocessing

In preprocessing step, here numeric to nominal filtration is applied in order to sort values i.e. it just take all numeric value and add them to the list of nominal values of that attribute.

3.3 Classification

Classification is a technique for machine learning by which it is used to predict the grouping membership of different data instances. It will perform the task by which it will generalize the well-known structure so as to apply it on new data. Here naïve bayes classifier has been used quality measurement of dataset will be consider on the basis of percentage of correctly classified instances. For validation phase we use 10 fold cross validation method. Naive bayes classifier helps in identifying the characteristics of patient with heart diseases. It gives the probability of each selected attribute for the predictable state.

3.4 Hybrid Approach

In this hybrid approach Decision table algorithm stores the data based on the preferred set of attributes and using that model as a lookup table while making the predictions for the data. Every entry in the table has class probability related with it. The main challenge of Decision table is to select a part of set that has highly discriminative attributes. Naïve Bayes is depend on Bayes' theorem with independence assumptions between predictors.

1. In proposed approach, Decision table represent the conditional probability table for naive bayes.
2. Each point in the search, the algorithm estimates the values by isolating the qualities of attributes into two different parts: one for each of naïve Bayes and decision table.
3. Initially every attributes are modeled by the decision table. At every step, forward selection search is used, and the attributes selected from this are build by naive Bayes and the rest by the decision table.

At every point in this exploration it estimates the value which is associated with the attributes splitted into two sets. The class probability estimates must be combined to produce overall class probability estimates.

Split the attributes into two groups based on the searching at each point each selected attribute is modeled using NB and the remaining is modeled by DT. The

results of each model are evaluated and probability is calculated**.**

3.5 Parameter evaluation Evaluation

Calculate the accuracy, precision, recall , f- measure, ROC. Compare the result for naïve bayes and hybrid approach.

1. Accuracy-Accuracy is termed as correctly classified instances in percentage.

   Accuracy= True positive +true negative/ (true pos +false pos +true neg +false neg).

2. Precision -Precision is the fraction of retrieved instances that is measure of exactness[20].

   Precision=TP/(TP+FP).

3. Recall- Recall is the fraction of relevant instances that is retrieved that is measure of completeness i.e. true positive rate of class[20].

   Recall=TP/(TP+FN).

4. F-measure-F-measure is the harmonic mean of precision and recall[20].

   F=2*precision*recall/(precision+recall).

5. ROC Area- Receiver operating characteristics curve shows the relationship between false positives and true positives[20].

Confusion matrix- Confusion matrix is a matrix include information about actual and predicted classifications[20].

Parameters of confusion matrix

1. TP - It indicates the number of records classified as true though they were actually true. Such as patients having heart disease correctly identified as heart disease.
2. FP - It denotes the number of records classified as true while they were actually false. Such as people who are healthy are incorrectly identified as heart disease.
3. TN- It denotes the number of records classified as false while they were actually false. Such as people who are healthy are correctly identified as healthy.
4. FN- It denotes the number of records classified as false while they were actually true. Patient having heart disease are incorrectly identified as healthy.

3.6 Attribute description

Table 1. Description

Name	Type	Description
Age	continuous	Age of patient
Sex	discrete	1=male 0=female
Cp	discrete	Chest pain • typical angina= 1 • atypical angina = 2 • non-angina = 3 • asymptomatic= 4
Trestbps	continuous	Resting blood pressure(in mm Hg)
Chol	continuous	Serum cholesterol in mg/dl
Fbs	discrete	Fasting blood sugar>120mg/dl • True = 1 • False = 0
Restecg	discrete	Resting electrocardiographic results • Normal indicate= 0 • Having ST-T wave abnormality indicate =1 • Showing probable or define left ventricular hypertrophy by estes criteria =2
Thalach	Continuous	Maximum heart rate achieved
Exang	discrete	Exercise induced angina • Yes indicates value =1 • No indicates value=0
Slope	discrete	The slope of the peak exercise segment • up sloping value=1 • flat value=2 • down sloping value=3
oldpeak	Continuous	ST depression induced by exercise
Thal	discrete	• Normal represent value=3 • fixed defect represent value=6 • reversible defect represent value=7
CA	discrete	Number of major vessels colored by floursopy(0-3)
Class attr	discrete	Diagnosis classes • Present= having heart disease • Absent=not having heart disease

4. EXPERIMENTAL RESULTS

Here, analysis on heart data set is done in WEKA tool based on each attributes and also distributing the values shown as follow.

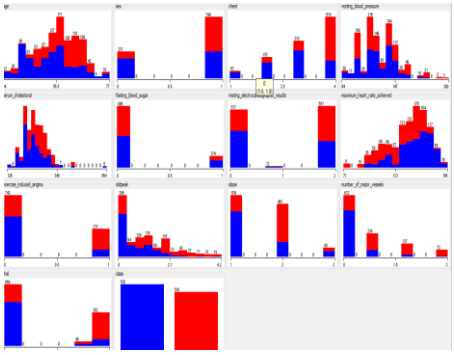

Figure 1 Visualization of the heart patients

Here we have three different algorithms to compare the results that are Naïve bayes, Decision table, hybrid i.e. combination of both naïve bayes and decision table. This algorithm is apply on heart dataset in WEKA tool. Here, we can see the performance of the algorithm i.e. how much accurate result is provided by these algorithms.

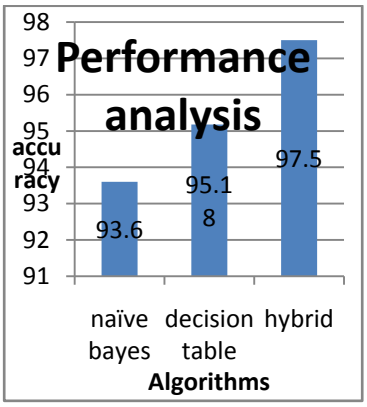

Figure 2 Performance analysis based on accuracy

The dataset contains 1080 instances and on which classification algorithm is applied to measure the performance of algorithm that from these many instances how many are correctly classified instances with the help of WEKA tool that can be seen from performance analysis.

5. CONCLUSION AND FUTURE SCOPE

Day by day healthcare data is increasing and having this huge amount of data that is being difficult to manage so mining techniques apply on it. We proposed a Heart disease prediction system that provides the important tool for physicians to take decisions from this huge and mined data for analysis based on previous data. The research undertake an experiment on application of various data mining algorithms to predict the heart attack and to compare the best method of prediction. Different classification algorithms is used to analyze on heart disease patient data, it will check all the symptoms to predict the presence of heart disease and also measure the accurate result based on the performance of the algorithm. The predictive accuracy determined by Naïve Bayes, Decision Table, Hybrid algorithms are measured and finds that hybrid provides best result while comparing with others. For the future study, analysis of heart disease patient based on the treatment and medicine provided by the doctors to find the best and effective treatment for the risky patient.

6. ACKNOWLEDGEMENT

I owe a great many thanks to a great many people who helped and supported me in writing this research paper.

7. REFERENCES

[1] Abhishek Taneja 2013 ”Heart Disease Prediction System Using Mining Techniques”, Oriental Journal of Computer Science and Technology.

[2] Ahmed T. sadiq alobaidi, Noor thamer mahmood 2013.”Modified full Bayesian network classifiers for medical diagnosis*”* IEEE.

[3] David Cornforth, Mika Tarvainen, Herbert F.Jelinek. 2013 ”Computational intelligence methods for the identification of early cardiac autonomic neuropathy”, IEEE.

[4] G.Karthiga, C.Preethi, R.Delshi Howsalya Devi 2014 ”Heart Disease Analysis Sytem Using Data Mining Techniques”, International Conference on Innovations in Engineering and Technology IEEE

[5] Hanaa Elshazly, Ahmed taher azar, Abeer el-korany and aboul ella hassanien. 2013 .”Hybrid System for lymphatic diseases diagnosis”, IEEE.

[6] Hezlin Aryani Abd Rahman, Yap Bee Wah. 2012”Comparison of predictive models to predict survival of cardiac surgery patients”

[7] Hian Chye Koh and Gerald Tan 2012 ”Data Mining Applications in Health care”, Journal of healthcare information management vol.19,No.2

[8] Hlaudi Daniel Masthe, Mosima Anna Masethe. 2014 “Prediction of heart disease using classification algorithms

[9] Lin Li, Saeed Bagheri ,Helena Goote 2013 “Risk adjustment of patient expenditures ,Philips Research North America Briarcliff Manor, US IEEE

[10] Mai Shouman, Tim Turner ,Rob Stocker 2012 ”Using Data Mining Techniques In Heart Disease Diagnosis And Treatment*”*, School of Engineering and Information Technology University of New South Wales At the Australian Defence Force Academy Northcott Drive,ACT 2600 IEEE.

[11] M.Akhil jabbar,Priti Chandra,B.L Deekshatulu 2012 ”Prediction of Risk Score for Heart Disease Using Associative Classification and Hybrid Feature Subset Selection*”*, Aurora’s Engineering College Bhongir A.P,India,Advanced System Laboratory Hyderabad,IDRBT,RBI(Govt of INDIA) Hyderabad, IEEE.

[12] Mohammad Taha Khan,Dr. Shamimul Qamar and Laurent F.Massin. 2012 “A prototype of cancer/heart disease prediction model using data mining “

[13] Mythili, Dev Mukherji, Nikita Padalia ,and Abhiram Naidu. 2013 “Heart disease prediction model using SVM Decision tree logistic regression”.

[14] Ranganatha S., Pooja Raj H.R,Anusha C, Vinay S.K 2013 ”Medical data mining and analysis for heart diseases dataset using classification techniques”, Govt .Engineering College, Hassan INDIA,PES Institute of Technology, Banglore, INDIA

[15] R.Chitra and V. Seenivasagam 2013 ”Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques*”,* Department of Computer Science and Engineering, Noorul Islam Center for Higher Education ,India.ICTACT Journal on Computing.

[16] Saba Bashir, Usman Qamar, M.Younus Javed. 2014 “Ensemble based decision support framework for intelligent heart disease diagnosis”, IEEE.

[17] Tina R. Patil, Mrs.S.S.Sherekar 2013 ”Performance analysis of Naïve bayes and J48 classification algorithm for data classification”.

[18] Vikas Chaurasia 2013 ”Early Prediction of Heart Diseases Using Data Mining Techniques”,Carib.j. Science technology,vol.1.

[19] V.Manikantan & S.Lanthan 2013 ”Predicting The Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods”, Department of Computer Science and Engineering, Mahendra Institute of Technology Tiruchengode, Namakkal, India, ISSN ,Vol -2.

[20] Xiao Fu,Yinzi Guiqiu, Qing Pan. 2011 ”A Computational modelfor heart failure stratification”.

[21] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algodecisiontree.htm#DmC0N019

[22] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#110057465 available at: Data Mining Concepts

[23] http://stats.stackexchange.com/questions/23490/why-do-naivebayesian classifiers available at: cross validated.

[24] http://www.techopedia.com/defination/18829/decisiontabledetab

[25] http://www.cdc.gov/dhdsp/action_plan/pdf/action_planfull.pdf

[26] http://www.nimh.nih.gov/health/publications/depressionandheart-disease/depressionandheartdisease_142318.pdf.

[27] http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/217118/93872900853-cvd-outcomes-web1.pdf.