www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -04 April, 2016 Page No. 16070-16072

Early Diagnosis Of Heart Disease Using Alternating Decision Tree And KNN Algorithm

Piyush Jangle[1], Sarang Narayankar[2], Kapil Suryawanshi[3], Saurabh Bhandare[4], Prof. Sanjay Ghodke[5]

[1]Dept. of Computer Engineering,
MIT-AOE, Alandi, Maharashtra, India
piyushjangle@gmail.com

[2]Dept. of Computer Engineering,
MIT-AOE, Alandi, Maharashtra, India
snarayankar200@gmail.com

[3]Dept. of Computer Engineering,
MIT-AOE, Alandi, Maharashtra, India
kapil.s0492@gmail.com

[4]Dept. of Computer Engineering,
MIT-AOE, Alandi, Maharashtra, India
saurabhbhandare41@gmail.com

[5]Dept. of Computer Engineering,
MIT-AOE, Alandi, Maharashtra, India

Abstract: ***The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data. Because of this complexity, there exists a significant amount of interest among clinical professionals and researchers regarding the efficient and accurate prediction of heart disease. In case of heart disease time is very crucial to get correct diagnosis in early stage. Patient having chest pain complaint may undergo unnecessary treatment or admitted in the hospital. In most of the developing countries specialists are not widely available for the diagnosis. Hence, automated system can help to medical community to assist doctor for the accurate diagnosis well in advance. So the decision support systems play an important role in the diagnosis of heart disease. However, accurate diagnosis at an early stage followed by proper subsequent treatment can result in significant life saving.***

Keywords: Heart disease, KNN, Alternating decision trees, Data mining, Classification

1. Introduction

The data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. The most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data.These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this research.Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths. The Economical and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, , most lives are lost to non communicable diseases such as cardiovascular diseases, cancers, diabetes and chronic respiratory diseases. According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. In 2008, five out of the top ten causes for mortality worldwide, other than injuries, were non-communicable diseases; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs). Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non-communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths . Cardiovascular disease includes coronary heart disease (CHD), cerebrovascular disease (stroke), Hypertensive heart disease, congenital heart disease, peripheral artery disease, rheumatic heart disease, inflammatory heart disease. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease . Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques

This technique is routinely use in large number of industries like engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining.

2. LITERATURE SURVEY

Data mining has been played an important role in the intelligent medical systems. The relationships of disorders and the real causes of the disorders and the effects of symptoms that are spontaneously seen in patients can be evaluated by the users via the constructed software easily. Large databases can be applied as the input data to the software by using the extendibility of the software. The effects of relationships that have not been evaluated adequately have been explored and the relationships of hidden knowledge laid among the large medical databases have been searched in this study by means of finding frequent items using candidate generation. The sets of sicknesses simultaneously seen in the medical databases can be reduced by using our non candidate approach. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. The three major causes of heart diseases are chest pain, stroke and heart attack . Data mining and KDD (Knowledge Discovery in Databases) are related terms and are used interchangeably. According to Fayyad et al., the knowledge discovery process are structured in various stages whereas the first stage is data selection where data is collected from various sources, the second stage is preprocessing of the selected data, the third stage is transformation of the data into appropriate format for further processing, the fourth stage is Data mining where suitable Data mining technique is applied on the data for extracting valuable information and evaluation is the last stage The data mining methods like artificial neural network technique is used in effective heart attack prediction system. First the dataset used for prediction of heart diseases was preprocessed and clustered by means of K-means clustering algorithm . Then neural network is trained with the selected significant patterns. Multi-layer Perceptron Neural Network with Back propagation is used for training. The results indicate that the algorithm used is capable of predicting the heart diseases more efficiently. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The outcome shows that Decision Tree performance is more and few times Bayesian classification is having similar accuracy as of decision tree but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering will not perform well. By using the Weighted Associative Classifier (WAC), a slight change has been made, instead of considering 5 class labels, only 2 class labels are used. One for "Heart Disease" and another one for "No Heart Disease". The maximum accuracy (81.51%) has been achieved. When genetic algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset of 909 patient records with heart diseases has been collected and 13 attributes has been used for consistency. The patient records have been spitted equally as 455 records for training dataset and 454 records for testing dataset. After applying genetic algorithm the attributes has been reduced to 6 and decision tree performs more efficiently with 99.2% accuracy when compared with other algorithms.

3. ALTERNATING DECISION TREE

Berry and Linoff defined decision tree as "a structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another." Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain. Decision trees are produced by algorithms that are used to identify various ways of splitting a data set into segments. These segments form an inverted decision tree. That decision tree originates with a root node at the top of the There are many types of Decision trees. The Difference between them is mathematical model that is used to select the splitting attribute in extracting the Decision tree rules. Three most commonly used research tests types: 1) Information Gain, 2) Gini index and 3) Gain ratio Decision Trees.

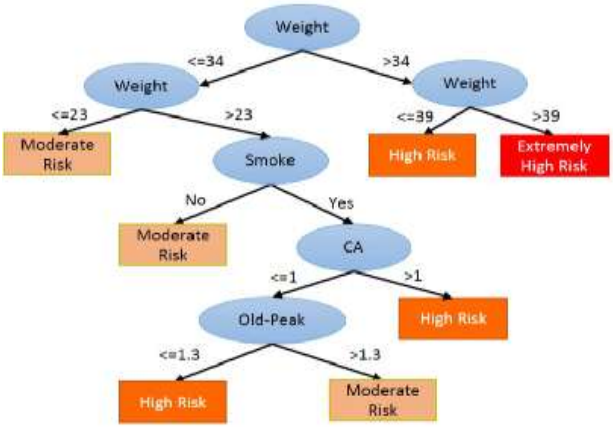


Figure 1: Decision tree classifying heart disease data.

Figure1 shows decision tree classifying weather data set. Among other classification algorithms decision tree algorithms are most commonly used because they are easy to understand and cheap to implement. Decision tree are machine learning methods, which combines decision trees and boosting to generate classification rules. The structure of an decision tree represents decision paths, when a path reaches decision node it continues with the child node which corresponds to the outcome of the decision associated with the node. When reaching a prediction node, the path continues with all of the children of the node.

4. HEART DISEASE

Heart disease refers to any condition in which the heart and blood vessels are injured and do not function properly, which results in fatal health problems. Different types of heart diseases among which are 1) congential 2) myocarditis 3) coronary 4) angina 5) rheumatic heart disease.

Risk factors of Coronary Heart Disease: Scientists are still unclear about specific cause of heart disease, but gender, age, ethnic back ground, family history, high BP, abnormal blood lipids, use of tobacco, obesity, hypertension, and diabetes mellitus are the major risk factors of heart disease.

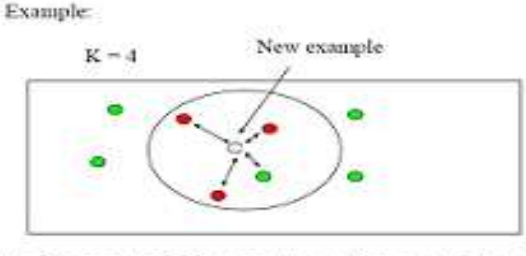
5. METHODOLOGY

This section describes our proposed methodology. We applied alternating decision trees with PCA as filter on heart disease data set collected from various corporate hospitals. Features are selected based on the opinion from expert doctors. A total of 96 patients records with 10 features are used in our experiments. Attributes used in our experiments are shown in Table

Sr. No.	Name Of Attributes
1	Age
2	Sex
3	Chest pain type
4	Resting blood pressure
5	Cholesterol
6	Fasting blood sugar
7	Resting electrocardiograph
8	Max. heart rate
9	Exercise
10	Slope of exercise peak
11	Diagnosis Classes

6. KNN ALGORITHM

In our research paper we use KNN algorithm after construction of decision tree. KNN algorithm is used for better result and prediction. It makes sense to weight the contribution of each example according to the distance to the new query example. Weight varies inversely with the distance, such that examples closer to the query points get higher weight Instead of only k examples, we could allow all training examples to contribute



Find the *k* nearest neighbors and have them vote. Has a smoothing effect. This is especially good when there is noise in the class labels.

7. CONCLUSION

In this paper, different classifiers are studied and the experiments are conducted to find the best classifier for predicting the patient of heart disease. we propose an approach to predict the heart diseases using data mining techniques. Three classifiers such as ID3, KNN and DT were used for diagnosis of patients with heart diseases. Observation shows that decision tree with KNN performance is having more accuracy, when compared with other classification methods. Choice tree is one of the vital information mining procedures that is generally utilized as a part of conclusion of Heart Disease. Exchanging choice trees are new representation for order standard which are anything but difficult to execute, early and interpretable. We also shows that the most important attributes for heart diseases are *cp (Chest pain), slope (The slope of the peak exercise segment), Exang (Exercise induced angina),* and *Restecg (Resting electrocardiographic)*. These attributes were found using three tests for the assessment of input variables: Chi-square test, Info Gain test and Gain Ratio test. utilizing ADTrees will assume an imperative part to help medicinal services experts.

References

[1]India today, Indias No 1 killer; Heart disease

[2]Sevith Rao, Cardiac disease among south Asians: A silent epidemic Indian heart association(last accessed 30/5/2014)

[3]Vikas chaurasia and saurab palData mining approach to detect heart diseaseIJACSIT,Vol no 2,no 4,pp56-66(2013)

[4]M.A .Jabbar, Deekshatulu, Priti Chandra , Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection,GJCST, Vol 13, issue 3, version1.0 pp15-25(2013)

[5]Swathi shilaskar et.al, Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease , IJCA, Vol 61,No 5,pp 1-8(2013)

[6]reventing Chronic Disease: A Vital Investment. World Health Organization Global Report, 2005.

[7]Global Burden of Disease. 2004 update (2008). World Health Organization.

[8]Srinivas, K., Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques, IEEE Transaction on Computer Science and Education (ICCSE),p(1344-1349),2010.

[9]S. K. Yadav and Pal S., Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.

[10]Anbarasi, M., E. Anupriya, et al. (2010). "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm."International Journal of Engineering Science and Technology Vol. 2(10).

[11] R. Agrawal, T.Imielinski, A.Swami, "Mining association rules between sets of items in large databases", In ACM SIGMOD Conference, pp. 207216, 1993.