International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 13, Issue 01) and (Publishing Month: May 2014)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Enhanced Heart Disease Prediction Using Decision Tree

Madhavi Kumari[1] and Amit Verma[2]

[1]M.Tech. Student and [2]Assistant Professor
[1,2]Computer Science and Engineering Department
[1,2]Samalkha Group of Institutions, Hathwala, Panipat (Haryana)
[1]madbld16@gmail.com

Abstract

Healthcare information systems containing huge number of medical records are ideal targets for data mining. Many works have applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. Data mining is to extract hidden rules and relationships between diseases from a real world Healthcare Information System. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. This paper proposes an algorithm that uses error back propagation along with cart. The proposed algorithm increases the classification accuracy by 7%.

Keywords: *Data Mining, Heart Disease, Classification technique, CART.*

I. Introduction

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining". Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining. The following definition is given [1]:

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. In [2], the following definition is given: Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

Data mining is an interdisciplinary subfield of computer science which involves computational process of large data sets' patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence.

Data Mining is about solving problems by analyzing data already present in databases. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns. Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in two categories-descriptive and predictive.

Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions. The purpose of a data mining effort is normally either to create a descriptive model or a

predictive model. A descriptive model presents, in concise form, the main characteristics of the data set. It is essentially a summary of the data points, making it possible to study important aspects of the data set. Typically, a descriptive model is found through undirected data mining; i.e. a bottom-up approach where the data "speaks for itself". Undirected data mining finds patterns in the data set but leaves the interpretation of the patterns to the data miner [1].

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and interpretation of data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data" [3].

II. Heart Disease and Data Mining

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data.

The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases bur are hidden among large amounts of data [4].

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?" , "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However they cannot answer complex queries like "Given patient records, predict the probability of patients getting a heart disease." Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database.

This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The system that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions [4].

Coronary heart disease is a narrowing of the small blood vessels that supply blood and oxygen to the heart. This is also called as coronary artery disease. Coronary heart disease is usually caused by a condition called atherosclerosis, which occurs when fatty material and a substance called plaque builds up on the walls of arteries. This causes them to get narrow. As the coronary arteries narrow, blood flow to the heart can slow down or stop, causing chest pain, shorteness of breath, heart attack, and other symptoms. Men in their 40's have higher risk of Coronary heart disease than women, but as women gets older, their risk increases so that it is almost equal to a man's risk. Major risk factors for Coronary heart disease are 1) Diabetes 2) High blood pressure 3) High LDL (bad) cholesterol 4) low LDL (good) cholesterol 5) Not getting enough physical activity6) Obesity7) Smoking [5].

Coronary artery disease cause severe disability and more death than any other disease including cancer. Coronary artery disease is due to athermanous narrowing and subsequent occlusion of the coronary vessel. It manifests as angina, silent ischemia, unstable angina, myocardial infraction, arrhythmias, heart failure and sudden death [6].

The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The

term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death [7]. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease (CHD).

Myocardial infarctions, generally known as a heart attacks, and angina pectoris, or chest pain are encompassed in the CHD. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pain arise when the blood received by the heart muscles is inadequate [7].High blood pressure, coronary artery disease, valvular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease.[6]

III. Classification Technique

CART is one of the more popular methods of constructing the decision tree. It builds a binary decision tree by splitting the records at each node according to a function of a single attribute. Classification and Regression Trees (CART) is a flexible method to describe how the variable Y distributes after assigning the forecast vector X. This model uses the binary tree to divide the forecast space into certain subsets on which Y distribution is continuously even. Tree's leaf nodes correspond to different division areas which are determined by Splitting Rules relating to each internal node. By moving from the tree root to the leaf node, a forecast sample will be given an only leaf node, and Y distribution on this node also be determined [8].

- **Splitting criteria:** CART uses GINI Index to determine in which attribute the branch should be generated. The strategy is to choose the attribute whose GINI Index is minimum after splitting.

- **GINI index:** Assuming training set T includes n samples, the target property has m values, among them, the ith value show in T with a probability Pi, so T’s GINI Index can be described as below:

$$GINI(T) = 1 - \sum_{i=1}^{m} P_i^2$$

Assuming the A be divided to q subsets, {T1, T2, · · · Tq}, among them, Ti’s sample number is ni, so the GINI Index divided according to property A can be described below:

$$GINI(T) = 1 - \sum_{i=1}^{q} GINI(T_i)$$

CART divides the property which leads a minimum value after the division.

IV. Proposed Technique

The proposed algorithm is divided in two phases, one is training phase and the other is the testing phase. In the training phase the error back propagation is used to train the network.

Training Phase

Initialize the weight vector of length N, learning rate say R, and the epochs counter say ep with random values.

Assume Wep is weight vector at the start of any ‘ep’ epoch.

Save the current weight value

Wold=Wkep.

For n=1,2,...N

Apply the error back propagation.

Update the weights

Wi(ep+1)=Wi(ep)−R*d(Wi)*d(E)

Here d(E) is the partial derivative of E, and E is the error.

ep=ep+1

if ep< mac epochs

then go to step 5.

Else end

The final weight vector is taken as the GINI INDEX say G and The E is maximum error possible.

Testing Phase

Calculate the GINI index of root node Gr. (GINI Index calculation already described).

If Gr>G

then split the node in two parts. One if Gr>=G and other is Gr<G

else label the node as LEAF node.

End if

Take new node as root node and go to step 12.

Repeat the process until all node processed.

The above algorithm uses the error back propagation and the binary splitting of CART. As the error back propagation defines the GINI INDEX and the error precisely so the accuracy of algorithm must be increased as compared to existing algorithm CART.

V. Results

The dataset used to analyze the proposed algorithm over WEKA is the Heart Disease Databases. This dataset is downloaded from the UCI repository [9]. This directory contains 4 databases concerning heart disease diagnosis. This data set contains 76 raw attributes instances and only 14 attributes of them are used. The attributes are as: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4 [10,11,12,13].

a. Performance Evaluation Metrics

This work adopted Precision and Recall, ROC and Lift as the performance metrics for estimating the accuracy of a given classification model [14] [15]. Each of these was used where appropriate in the analysis of the performances. Apart from the major performance criteria mentioned, the work will also measure the speed and the robustness of the classifiers.

(i) Accuracy

- **Precision and Recall**

The general percentage accuracy as a performance measure has been proven to be misleading [16] [14]. For example, a classifier that labels all regions as the majority class will achieve an accuracy of 94%, because 96% of the majority may belong to that region. Meanwhile, the classifier may have incorrectly classified some of the minority class instance as the majority because of the bias nature of the dataset but may appear to be accurate. It is therefore imperative to compare the accuracy using an alternative method - Precision and Recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times \quad 100\% \qquad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times \quad 100\% \, (2)$$

Where, TP, TN, FP, and FN are as represented in the confusion matrix. Precision in this context refers to the actual percentage of responses to mails that were predicted by the classification model, which translates into the returns on cost of mailing. The Recall, on the other hand, measures the percentage of customers that were identified and needed to be targeted.

The proposed algorithm is compared with the simple CART and the J48 algorithm over the described dataset.

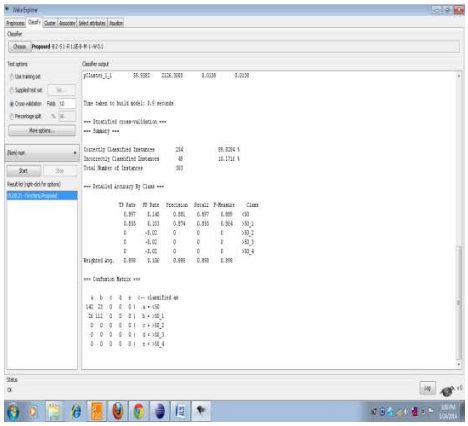

Figure 1 Screen Shot of Proposed algorithm Results

The table 1 shows the parameters comparison of the J48, simple CART and the proposed algorithm. The parameters are TP i.e. true positive rate and FP i.e. false positive rate, classification accuracy, precision,

recall and the F-measure. These parameters are already defined.

Table 5.1 Parameter Analysis of Various Algorithms

Algorithm Name	Classification accuracy	Tp rate	Fp rate	Precision	Recall	F-measure
J48	77.7578	0.776	0.235	0.776	0.776	0.774
Simple Cart	80.8581	0.809	0.202	0.809	0.809	0.808
Proposed	89.8284	0.898	0.106	0.898	0.898	0.898

The comparison can also be done graphically as shown in the following figures.

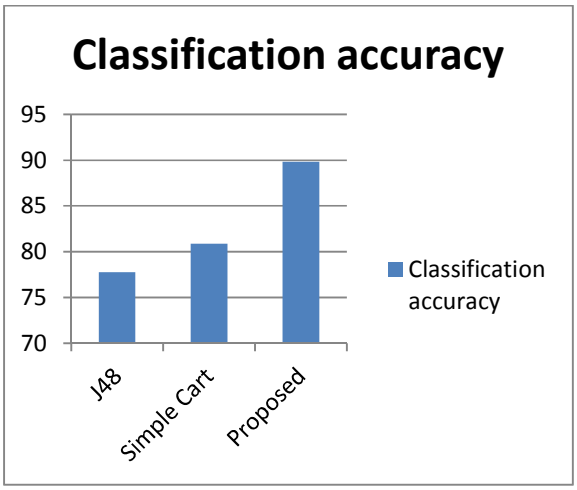

Figure 2: Classification Accuracy comparison between J48, simple CART and proposed algorithm

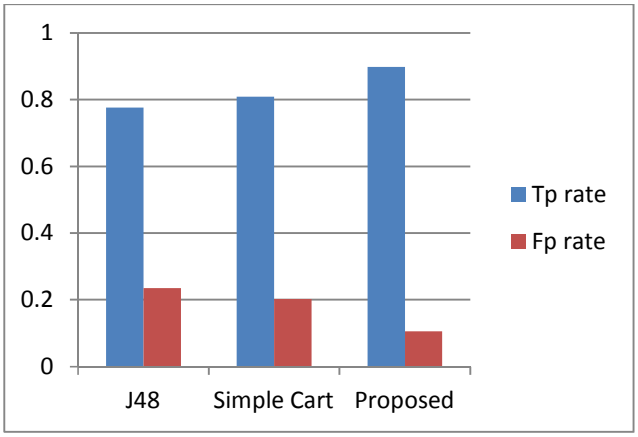

Figure 3: Tp rate and Fp rate comparison between J48, simple CART and proposed algorithm

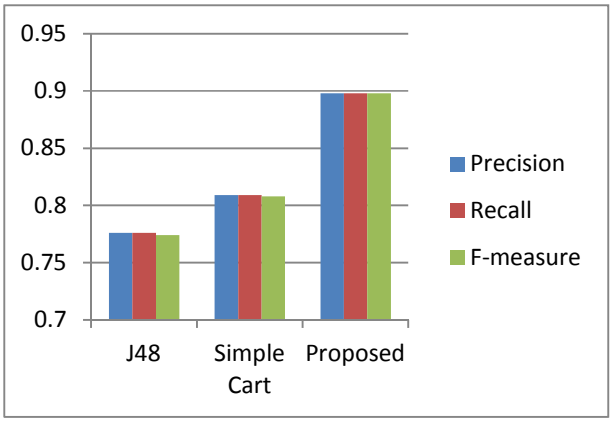

Figure 4: Precision, Recall and F-measure comparison between J48, simple CART and proposed algorithm

The figure 2, 3 and 4 shows the comparison of the various parameters between the J48 , simple cart and the proposed algorithm. A significant increase of almost 7% in the classification accuracy can be analyzed using the above the figures. The Fp rate get decreased and the true positive rate get increased. The Precision, Recall as well as the F-measure of the proposed algorithm are better than the existing algorithms. It means the performance of the proposed algorithm is better than the existing algorithms.

VI. Conclusion

The paper uses the error back propagation and the binary splitting of CART. The error back propagation defines the GINI INDEX precisely so the accuracy of algorithm gets increased as compared to existing algorithm CART. The simple CART algorithm classifies the tree on the basis of the GINI index. The simple CART is basically a binary splitting method that split a node on the basis of the information gain i.e. GINI index. The split node is further split and process continues. If the node has the information gain i.e. GINI index lower than the training only then the node can't be split. The nodes which cannot be split are named leaf nodes. The proposed technique classifies the tree binary but the training for each node is done by the error back propagation. The results shows the comparison of the various parameters between the J48 , simple cart and the proposed algorithm. A significant increase of almost 7% in the classification accuracy of proposed algorithm is analyzed as compared to existing algorithms. The Fp rate get decreased and the true positive rate get increased. The Precision, Recall as well as the F-measure of the proposed algorithm are better than the existing algorithms. It means the performance of the proposed algorithm is better than the existing algorithms. In future following work can be done: The proposed algorithm can be analyzed on various other datasets. The algorithm can use fuzzy or the neuro-fuzzy to increase the classification accuracy. The algorithm can be extended for regression process.

References

[1] Jain, N., & Srivastava, V. (2013),Data Mining Techniques: A Survey Paper. International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11.

[2] Xingquan Zhu, Ian Davidson, (2007) Knowledge Discovery and Data Mining: Challenges and Realities, ISBN 978- 1-59904-252, Hershey, New York.

[3] Singh, A., & Das, K. K. (2007). Application Of Data Mining Techniques In Bioinformatics (Doctoral dissertation).

[4] Subbalakshmi, G., Ramesh, K., & Chinna Rao, M. (2011). Decision Support In Heart Disease Prediction System Using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), 170-176.

[5] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart Disease Prediction System using Associative Classification and Genetic Algorithm.arXiv preprint arXiv:1303.5919.

[6] Kavitha, K. S., Ramakrishnan, K. V., & Singh, M. K. (2010). Modeling And Design Of Evolutionary Neural Network For Heart Disease Detection. International Journal of Computer Science Issues (IJCSI), 7(5).

[7] Cheng, J., & Greiner, R. (1999, July). Comparing Bayesian Network Classifiers. In Proceedings Of The Fifteenth Conference On Uncertainty In Artificial Intelligence (pp. 101-108). Morgan Kaufmann Publishers Inc

[8] Prof. Nilima Patil (July-August 2012)Customer Card Classification Based on C5.0 & CART Algorithms, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4,

[9] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

[11] University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

[12] University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

[13] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

[14] Guo, H. and Viktor, H. L., (2004) Learning From Imbalanced Data Sets With Boosting And Data Generation: The Databoost-IM Approach., SIGKDD Explorations;6(1): 30-39.

[15] Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools And Techniques. Morgan Kaufmann.

[16] Provost, F., Fawcett, T., and Kohavi, R.,(1998) The Case Against Accuracy Estimation for Comparing Classifiers., In Proceedings of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann.

[17] Provost, F. and Fawcett, T., (2001) Robust Classification for Imprecise Environments, Machine Learning; 42(3): 203-231.

[18] Flach, P. and Gamberger, D., (2001) Subgroup Evaluation And Decision Support For A Direct

Mailing Marketing Problem, Aspects of Data Mining, Decision Support and Meta-Learning,[Online]available from: http://www.informatik.unifreiburg.de/~ml/ecmlpkdd/WSProceedings/w04/paper8.pdf., 2001.