ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

An Approach for Early Diagnosis of Cardio-Vascular Disease Using Modified Decision Tree

Zarna Parekh[1], **Avaniba Parmar**[2]
Hasmukh Goswami College Of Engineering, Ahmedabad, India

Abstarct: The heart disease is a major cause of mortality and death-rate in modern society. Medical assessment is very important but the labyrinthine task should be performed efficiently and accurately. So, in this era of computing and intelligence it is an easy yet complicated task to estimate the probability of disease on the basis of data and fact provided to the system. With the growing research on heart disease predicting system, it has become important to categorize the research outcomes and provide readers with an overview of the existing heart disease prediction techniques in each category. The data mining algorithms are typically used to identify the disease that occurs in original from the database

So in this research work, we have used Decision Tree for the detection of Cardiovascular Disease.

Keywords
Data Mining, Decision Tree, C5.0,Heart Disease Diagnosis

I. INTRODUCTION

Cardiovascular disease which is also called as heart disease comprises a class of disease which consists of heart the blood vessels or both. Heart diseases are the major cause of death globally: more people die annually from heart diseases than from any other disease [1]. An estimate is done that 17.3 million people died because of heart disease in 2012, which comprise of 30% of all global deaths. From these death an estimate was done that resulted with 7.3 million deaths were due to coronary heart disease and 6.2 million were due to stroke[1]. So to make a study and research on prevention of heart disease risk has become a important task nowadays for the researchers. Health care industry generates a large amount of complex data comprising of patients, electronic patient records, disease diagnosis etc. These large amount of data collected is a key source which is to be processed and analyzed for knowledge extraction which reduces the cost and helps for decision making. Extracting necessary information and providing scientific decision-making for the diagnosis and treatment of disease from the dataset becomes important. Data mining in medical field brings a set of tools and techniques which can be used to process the data to discover unseen patterns which provides healthcare professionals an additional useful source of information for making decisions. The database produced by the biological researchers can be massive [2]. However, using data mining techniques to develop a suitable treatment for prediction of heart disease has received a less attention so researchers have been identifying the effect of hybridizing more than one technique giving enhanced results in diagnosis of heart disease. Artificial neural network has emerged as an efficient tool giving accurate and efficient results when hybrided with data mining [3].

II. DATA MINING

Data Mining and Knowledge discovery is the process of getting important and valuable information that has been unknown in the raw data previously. It is a process of capturing knowledge with the help of computer. Data Mining is an exploration of large dataset to extract hidden and previously unknown patterns, relationships and knowledge which are not easy to detect with traditional statistics [6]. Researchers are trying to obtain satisfactory results in a reasonable time with help of searching techniques because many problems are difficult to be solved in a feasible time by analytically. Data Mining is divided into two tasks such as Predictive Tasks and Descriptive Tasks. The value of a specific attribute based on other attribute is been predicted by Predictive task. Classification, Regression and Deviation Deduction come under Predictive Tasks. Descriptive Tasks derive pattern that summarize the relationship between data and consists of Clustering, Discovery. Data Mining Association Rule Mining and Sequential Pattern involves few steps from raw data collection to some form of new knowledge. The iterative cycle consists of steps like Data cleaning, Data Integration, Data Selection, Data transformation, Data Mining, Pattern Evaluation, and Knowledge Representation.

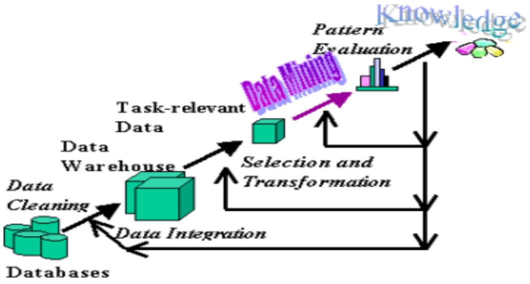


Figure 1: Data Mining is the core of Knowledge Discovery Process[4]

The Figure 1 shows that the Data Mining is the core for Knowledge Discovery process. Data Mining techniques like Naïve Bayes, Decision Tree, SVM(Support Vector Machine), K- NN, CMAR (Classification Based on Multiple Association Rules) etc are used for prediction of heart disease.[4]

III. LITERATURE SURVEY

A. *Feature Selection using Artificial Bee Colony for Cardiovascular disease Classification [2]*

In this paper B. Subanya has proposed feature selection approach in which ABC (Artificial Bee Colony) algorithm is used to optimize the process of feature selection and gives best optimal feature subset that is used to increase the predictive accuracy of the classifier. The use of algorithm is done as a feature selector to generate the feature subset and a classifier is used to evaluate every feature subset that is made by the onlookers. Here the use of SVM is done with two key concepts to solve the problem: large margin separation and kernel function. The idea of large margin separation can be motivated by classification of points in 2 dimensions by drawing a straight line and call points on one side as positive and other as negative. Kernels are used for non-linear classification. Accuracy obtained is 86.76% with only 7 attributes.

B. *Genetic Neural Network based Data Mining in Prediction of Heart disease using Risk Factors[12]*

In this paper Syed Umar Amin initially analyzed the dataset from various sources [15]-[16] and then composes a dataset of 12 important risk factors. Then by using the genetic algorithm which is specialized global searching algorithm [17] initial weights of neural network were optimized and use of back propagation algorithm was done to train the network using weight optimized by GA. Lastly a multilayered Feed Forward network is used. Here the 1st step is to initialize the weights of network using 'configure' function of MATLAB and then this weights are passed through GA to optimize according to fitness function and after weights are optimized BP algorithm is used train and learn them. The predicted output will tell us the presence or absence of heart disease. 89% accuracy is obtained.

C. *An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Technique [14]*

In this paper t John Peter has made use of pattern reorganization and data mining technique into risk prediction models in clinical domain of cardiovascular medicine and has proposed the system in which two methods are proposed: (1)ARFF creation and (2)attribute selection and classification.(1) In Attribute –Relation File Format file is an ASCII text file that describes a list of instances sharing a set of attributes. (2) A heart disease dataset consists large quantity of data and applying classifier to this data is time consuming and gives less accuracy so in this paper dimensionality of data is reduced using various attribute selection methods and then the data is

D. *Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining [11]*

In this paper Nabeel Al-Milli has proposed a heart disease prediction system using BP algorithm which uses 13 attributes to predict heart disease. Algorithm uses two passes to pass through different layers of network: forward pass and backward pass. An activity pattern is given to input nodes of network in forward pass and its effect is propagated through the network layer by layer which produces a set of output which is actual response of the network. The actual response obtained is subtracted from a desired response to produce error signal which is propagated backward through the network. The synaptic weights are adjusted to make actual response of network come closer to desired response in true sense.

E. *Feature Selection in Ischemic Heart Disease Identification Using Feed Forward Neural Network [3]*

In this paper K Rajeswari et al proposed Feed Forward Neural Network trained with BP algorithm to identify a person affected by IHD or not and further the classification if IHD is done as 'high' , 'medium' and 'low' risk levels. Two phases are included: 1)learning phase where network learns by modification of weights and 2) testing phase where unknown input is tested for proper learning of neural network. Here initially the data was collected and analyzed for heart risk score prediction based on extensive study and expert opinion. Based on data collected classification of immediate risk analysis was done as 'no risk', 'low risk', 'medium risk' and 'high risk'. Next step was feature selection in which features were selected from given dataset. Use of multilayer perception network was done. Everytime combinations of some features were taken and accuracy was checked. In the end 12 features were selected with training accuracy 89.4% and testing 82.2%.

F. Neural Network Approach in Diagnosis of Patient: A Case Study [8]

In this paper Farhad Soliemanian Gharehchopogh et al has made use of a decision support model. Here the author has proposed Neural Network for patients information and has made use of four attributes. Data is collected of 40 persons from health center in Tabriz region. Here multilayer perceptron neural network architecture with 6 input nodes, 4 hidden nodes and 2 output nodes is used. Back propagation algorithm is used to find beneficial and effective information from data. 85% accuracy was obtained.

G. Early diagnosis of Heart Disease using Classification and Regression Trees [10]

In this paper the author has proposed a method to automatically classify PCG (Phonocardiograms) and used classification and Regression tress(CART) to identify pathological murmur.3 methods are used: 1) Pre-Processing which consists of filtering of heart sounds is performed with goal of removing unwanted noise and segmentation – identifies the heart sound components and timing interval between them. 2) Feature Extraction – this phase is focused on extracting signal features that better highlight the properties of the PCG signal, with the goal of identifying those that are more suitable for classification purpose. 3) Classification and Regression Trees – a step by step process in which decision tree is constructed by either splitting each node on tree in two daughter nodes. Objective of portioning is to find partitions of data such that terminal nodes are such homogeneous as possible.

H. Context Aware Cardiac Monitoring for Early Detection of Heart Diseases [13]

In this proposed method initially the use of ECG waveforms was done to determine many cardiac arrhythmias [19]. When the values of the interval do not fall within expected range then detection of different cardiac abnormalities was done. Then by using wearable sensors some vital signs were measured continuously which can say the kind of illness which may have direct impact on some heart disease. When abnormal ECG is detected and and classified the CMS picks the rules associated with arrhythmia, which were defined by experts and stored in service providers cloud repository. If context matched with rule then necessary action is taken to help user to overcome the problem.

I. HDPS : Heart Disease Prediction System [4]

In this paper the author has proposed a tool which is built in C language to implement heart disease classification and prediction via ANN. Learning Vector Quantization (LVQ), a prototype based supervised learning algorithm is used. The clinical data obtained is separated into two equal parts randomly. One is used for training and the other is used for testing. An initial weight is assigned randomly to each feature. The weights of all features are adjusted using calculated errors. The final weight of every feature is determined when the errors meet with the termination conditions. The testing data are then used to calculate the performance of this model. The process is repeated for 100 times. The output results include the average value of accuracy, specificity, and sensitivity. Lastly, the ROC curve is calculated in order to check the decency of the model.

IV. PROPOSED METHODOLOGY

In this section we are going to present a new method C5 with S-T Entropy calculation for diagnosis of cardiovascular disease.

C5.0 is expansion of C4.5 algorithm. C5.0 works by splitting the sample based on attribute that provides maximum gain ratio. The attribute which is obtained from former split is splitted afterwards. The process prolongs until the sample subset cannot be split and leaf node reaches class label. Finally, examine the lowest level split, those attribute that don't have remarkable contribution to the algorithm will be rejected.

Algorithm

Step 1: Start

Step 2: select training data set from the database

Step 3: Calculate information gain using Sharma-Taneja entropy calculation method

Step 4: [Calculate entropy of each attribute using Sharma-Taneja entropy calculation method]:

$$H_{\alpha,\beta}(\mathrm{P}) = \frac{1}{2^{1-\beta}-2^{1-\alpha}} \sum_{i=1}^{n}(P_i^{\beta} - P_i^{\alpha})$$

Where $H_{\alpha,\beta}(\mathrm{P})$ is calculated entropy and is α-entropy, β-entropy is its inherent parameter..

Step 5: [Calculate the information Gains of all attributes]

Gain (A) = Info (D) – $Info_A(D)$

//where Gain (A) tells us how much would be gained by branching on A attribute

Step 6: [Find SplitInfo of attribute]

$$SplitInfo_{(A)}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2\left(\frac{|D_j|}{|D|}\right)$$

Step 7: [Find Gain Ratio based on Gain(A) and SplitInfo]

$$Gain\ Ratio\ (A) = \frac{Gain(A)}{SplitInfo(A)}$$

Step 8: [Discard the Attributes whose Gain Ratio is below threshold]

Step 9: [split tree which has maximum value of gain ratio]

// set child node according the maximum gain ratio.

Step 10: For each child of the root Node, apply algorithm recursively until reach node that has entropy zero or reach leaf node.

Step 11: Display generated final optimal Decision tree.

Step 12: Stop

4.1 Flowchart

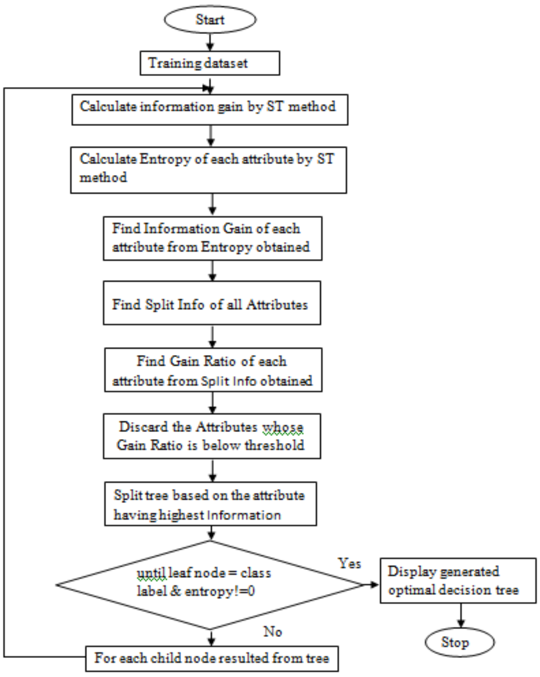
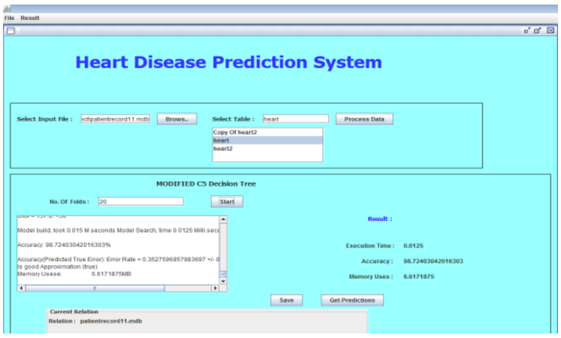

Start
Training dataset
Calculate information gain by ST method
Calculate Entropy of each attribute by ST method
Find Information Gain of each attribute from Entropy obtained
Find Split Info of all Attributes
Find Gain Ratio of each attribute from Split Info obtained
Discard the Attributes whose Gain Ratio is below threshold
Split tree based on the attribute having highest Information
until leaf node = class label & entropy!=0
Yes
Display generated optimal decision tree
Stop
No
For each child node resulted from tree

Figure 2 : Flowchart for overall process

V. PERFORMANCE EVALUTION

We have predicted heart disease by passing heart disease dataset, taken from UCI repository[8], to Modified C5 algorithm using Sharma-Taneja Entropy calculation method. We have got accuracy near about 99.2% .



Heart Disease Prediction System

Select Input File : :cf\patientrecord11.mdb

Select Table : heart

Copy Of heart2
heart
heart2

MODIFIED C5 Decision Tree

No. Of Folds : 20

```
Model build, took 0.015 M seconds Model Search, time 0.0125 Milli secc
Accuracy: 98.72403042016303%
Accuracy(Predicted True Error): Error Rate = 0.3527596957983697 +/- 0
is good Approximation (true)
Memory Usese:              6.6171875MB
```

Result :

Execution Time : 0.0125

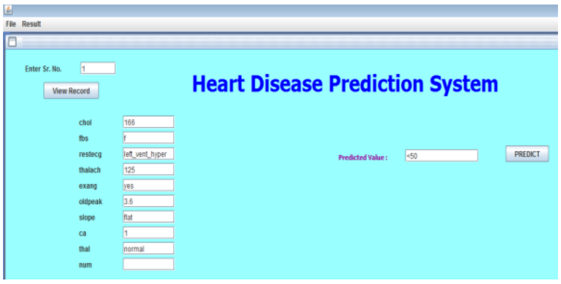
Accuracy : 98.72403042016303

Memory Uses : 6.6171875

Current Relation

Relation : patientrecord11.mdb

Figure 3 : Modified C5 for 20 number of folds



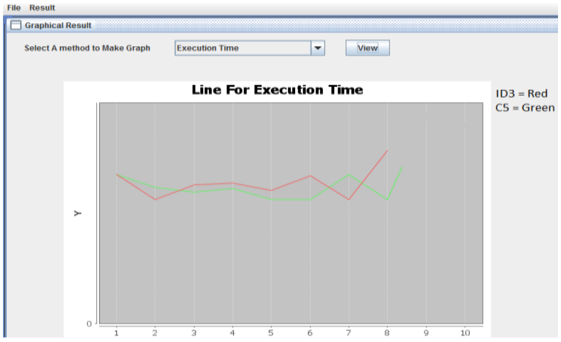
Heart Disease Prediction System

Enter Sr. No. 1

chol 166
fbs f
restecg left_vent_hyper
thalach 125
exang yes
oldpeak 3.6
slope flat
ca 1
thal normal
num

Predicted Value : <50

Figure 4 : Prediction of Heart disease after applying threshold value



Select A method to Make Graph: Execution Time

Figure 5 : Line for Execution Time of ID3 and modified C5

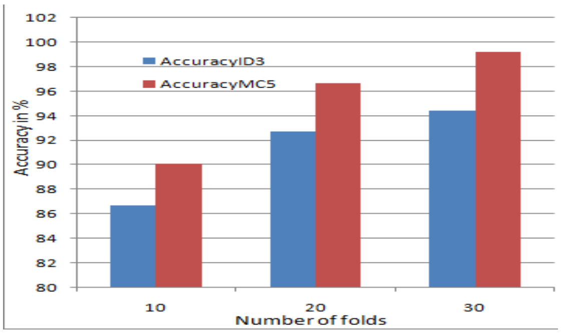


Figure 6: Comparative accuracy results of ID3 and Modified C5

VI. CONCLUSION

Heart diseases are the major cause of death globally: more people die annually from heart diseases than from any other disease. So to make a study and research on prevention of heart disease risk has become a important task nowadays for the researchers. So in this work we have implemented ID3 algorithm which will generate an accurate classified decision tree. The main work of ID3 algorithm is to calculate entropy and information gain. Based upon these values it generate decision tree. Basic ID3 algorithm selects attributes in terms of information entropy, which is less as compared to entropy obtained from Sharma-Taneja method, resulting in more time and less accuracy.

VII. FUTURE WORK

In place of decision tree other techniques of data mining can be used.

VIII. ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my guide, Asst Prof.Avaniba Parmar for her valuable guidance and useful suggestions.

IX. REFERENCES

[1] “Global atlas on cardiovascular disease prevention and control”,WHO,2011

[2] B.Subanya, Dr.R.R.Rajalaxmi ”Feature Selection using Artificial Bee Colony for Cardiovascular Disease Classification” IEEE (ICECS-2014)

[3] K.Rajeswari , Dr.V.Vaithiyanathan , Dr. T.R. Neelakantan “Feature Selection in Ischemic Heart Disease Identification Using Feed Forward Neural Network “ Procedia Engineering 41 (2012) Science Direct

[4] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin “HDPS Heart Disease Prediction System” IEEE Computing in Cardiology 2011

[5] George Cybenk,, (1996)“Neural Networks in Computational Science and Engineering”, IEEE Computational Science and Engineering, pp.36-42

[6] Liangxiao. J, Harry.Z, Zhihua.C and Jiang.S “One Dependency Augmented Naïve Bayes”, ADMA, pp 186-194, 2005.

[7] “Centre for Disease Control and Prevention, http://www.cdc.gov/heartdisease/risk_factors.htm.”

[8]http://repository.seasr.org/Datasets/UCI/arff/hear-c.arff. Heart disease dataset UCI Repository

[9] Amir Mohammad Amiri and Giuliano Armano “Early diagnosis of Heart Disease using Classification and Regression Trees “ IEEE

[10] Amir Mohammad Amiri and Giuliano Armano “Early diagnosis of Heart Disease using Classification and Regression Trees “ IEEE

[11] Brijain R Patel , Kaushik K Rana “Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining” IJARCSMS 2014

[12] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg “Genetic Neural Network based Data Mining in Prediction of Heart disease using Risk Factors” 978-1-4673-5758-6/13 2013 IEEE

[13] Abdul Forkan, Ibrahim Khalil, Zahir Tari “Context Aware Cardiac Monitoring for Early Detection of Heart Diseases” IEEE Computing in Cardiology 2013

[14] T.John Peter , K. Somasundaram “An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Technique “IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012)

[15] “Centre for Disease Control and Prevention, http://www.cdc.gov/heartdisease/risk_factors.htm.”

[16] “American Heart Association [Online], http://www.heart.org/HEARTORG/Condition”

[17] Y. J. Lei, and X. W. Zhang, Genetic Algorithm Toolbox of MatLab and its Application, Xian University of Electronic Science and Technology Press, 2005.

[18] Farhad Soleimanian Gharehchopoghi, Zeynab Abbasi Khalifelu ”Neural Network Approach in Diagnosis of Patient: A Case Study “978-1-61284-941-6/11 2011 IEEE

[19] Owis MI, Abou-Zied AH, Youssef A, Kadah YM. “Study of features based on nonlinear dynamical modeling in ecg arrhythmia detection and classification”. Biomedical engineering IEEE Transcations on 2002;49(7):733-736