International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-2, Issue-1,January 2016
ISSN: 2395-3470
www.ijseas.com

A PROFICIENT HEART DISEASE PREDICTION METHOD USING FUZZY-CART ALGORITHM

S.Suganya[1], P.Tamije Selvy[2]
[1] PG Scholar ,Department of CSE, Sri Krishna College Of Technology, Tamil Nadu, India.
[2] Professor, Department of CSE, Sri Krishna College Of Technology, Tamil Nadu, India

Abstract
Data mining technique in the history of medical data found with enormous investigations resulted that the prediction of heart disease is very important in medical science. The data from medical history has been found as heterogeneous data and it seems that the various forms of data should be interpreted to predict the heart disease of a patient. Various techniques in Data Mining have been applied to predict the patients of heart disease. But, the uncertainty in data was not removed with the techniques available in data mining. To remove uncertainty, it has been made an attempt by introducing gfuzziness in the measured data. A membership function was designed and incorporated with the measured value to remove uncertainty. Further, an attempt was made to classify the patients based on the attributes collected from medical field. Minimum distance CART classifier was incorporated to classify the data among various groups. It was found that CART classifier suits well as compared with other classifiers of parametric techniques.

Keywords: *CART Classifier, Membership function, Cleveland HeartDisease Data Base, Statlog Heart Disease Database,DataMining,Heart disease.*

1. Introduction

HEART attack diseases remains the main cause of death worldwide, including South Africa and possible detection at an earlier stage will prevent the attacks. Medical practitioners generate data with a wealth of hidden information present, and it's not properly being used effectively for predictions. For this purpose, the research converts the unused data into a dataset for modeling using different data mining techniques. People die having experienced symptoms that were not taken into considerations. There is a need for medical practitioners to predict heart disease before they occur in their patients. The features that increase the possibility of heart attacks are smoking, lack of physical exercises, high blood pressure, high cholesterol, unhealthy diet, harmful use of alcohol, and high sugar levels. Cardio Vascular Disease (CVD) incorporates coronary heart, cerebrovascular (Stroke), hypertensive heart, congenital heart, peripheral artery, rheumatic heart, inflammatory heart disease.

Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information. The current research intends to predict the probability of getting heart disease given patient data set. Predictions and descriptions are principal goals of data mining, in practice. Prediction in data mining involves attributes or variables in the data set to find an unknown or future state values of other attributes. Description emphasize on discovering patterns that explains the data to be interpreted by humans. The purpose of predictions in data mining is to help discover trends in patient data in order to improve their health. Due to change in life styles in developing countries, like South Africa, Cardio Vascular Disease (CVD) has become a leading cause of deaths. CVD is projected to be a single largest killer worldwide accounting for all deaths. An endeavor to exploit knowledge, experience and clinical screening of patients to diagnose or recognize heart attacks is regarded as a treasured opportunity. In the health sectors data mining play an important role to predict diseases. The predictive end of the research is a data mining model.

The diagnosis of diseases is a vital and intricate job in medicine. The recognition of heart disease from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. The health care industry collects huge amount of health care data which unfortunately are "not mined" to discover hidden info for effective decision making.

In India, about 25% of deaths at the age group of 25 - 69 years occur because of heart disease. In urban areas, 32.8% deaths occur because of heart ailments, while this percentage in rural areas is about 22.9%. The proposed method predicts the heart disease by using Fuzzy CART algorithm.

Most of the authors in data mining classifications techniques proposed for the prediction of heart disease, but the prediction system did not considered the uncertainty in the data measure. So, to remove the ambiguity and uncertainty, we made an experiment with fuzzy approach by introducing a membership function to the classifier. The fuzzy *K-NN classifier* results show promising in nature for removing the redundancy of data and to improve the accuracy of classifier as compared with other classifiers of supervised and un-supervised learning methods in data mining.

The term, Heart disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the United States, England, Canada and Wales as in 2007. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease; Cardiomyopathy and Cardiovascular diseases 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. Cardio Vascular Disease (CVD) results in severe illness, disability and death. 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 attack and angina pectoris or chest pain are encompassed in the CHD. A sudden blockage of a coronary artery, generally, due to a blood clot that results a heart attack. The chest pains arise when the receiving of the blood by the heart muscles is inadequate.

High blood pressure, coronary artery disease, valvular heart disease, stroke or rheumatic fever/rheumatic, heart diseases are the various forms of cardiovascular disease. The World Health Organization (WHO) has estimated that 12 million of deaths occurred worldwide every year due to the cardiovascular diseases. Half of the deaths in the United States and other developed countries occur due to cardio vascular diseases. It is also the chief reason of the deaths in numerous developing countries. On the whole, it is regarded the primary reason behind the deaths in adults.

2. RELATED WORK

There are different measures of centrality used in heart analysis:

- Dataset collection and Preprocessing.
- Attribute selection using ID3 algorithm.
- Classification using CART algorithm.
- Optimization using Particle Swarm Optimization algorithm

Three constraints were introduced to reduce the number of patterns, they are as follows:

1. The attributes have to appear on one side of the rule only.
2. It separates the attributes into uninteresting groups.
3. The ultimate constraint restricts the number of attributes in a rule.

The inhibited problem to identify and predict the association rules for the heart disease according to Carlos Ordonez presented in "Improving Heart Disease Prediction Using Constrained Association Rules,". The assessed set of data encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing.Besides decreasing the running time as per the experiments illustrated the constraints of exposed rules have been extremely decreased the number. The presence or absence of heart disease in four specific heart arteries have been anticipated two groups of rules. Data mining methods may help the cardiologists in the predication of the survival of patients and the practices adapted consequently. The work of Franck Le Duff et. al. might be executed for each medical procedure or medical problem and it would be feasible to make a wise decision tree fast with the data of a service or a physician. Comparison of traditional analysis and data mining analysis have been illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the present study. A chief drawback of the process was knowledge acquisition and the need to collect adequate data to create an appropriate model.

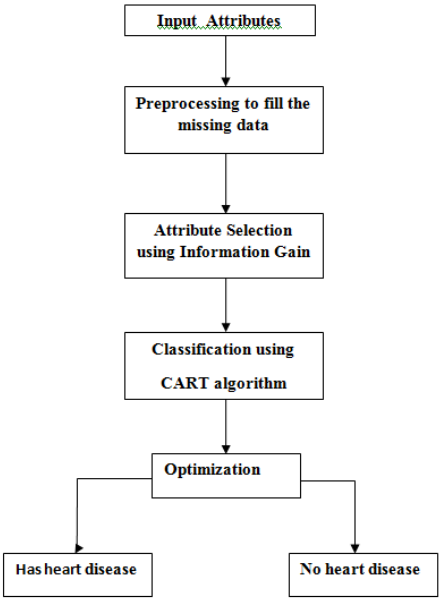
2.1 Preprocessing

Preprocessing involves adding the missing values and applying fuzzy membership function to remove data uncertainty.

Advantage: Improves data accuracy and removes uncertainty in data

Steps:

1. Collect the heart disease data set from Statlog heart disease database
2. Divide the dataset based on the attributes that increase the risk of heart disease.
3. Data should be preprocessed in order to remove any missing data.
4. Fuzzy Membership function is applied to remove the uncertainity in the data.



2.2 Attribute Selection using Information Gain

Steps:

1. To determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.
2. Calculate the entropy of the target.
3. Information Gain = entropy(parent) – [average entropy(children)]
4. Choose attribute with the largest information gain as the decision node.
5. The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

2.3 Classification by CART Algorithm

To classify the patient with heart disease.

Steps:

1. Tree building by using the attributes that increase the risk of heart diseases.
2. Prune and choose the final tree using cross validation.
3. Optimal Tree Selection.

3. RESEARCH METHODOLOGY

The goal of the prediction methodology is to design a model that can infer characteristic of predicted class from combination of other data [20]. The task of data mining in this research is to build models for prediction of the class based on selected attributes. The research applies the following algorithms: J48, Bayes Net, and Naive Bayes, Simple Cart, and REPTREE algorithm to classify and develop a model to diagnose heart attacks in the patient data set from medical practitioners.The objective of the research is to predict possible heart attacks from the patient dataset using data mining techniques and determines which model gives the highest percentage of correct predictions for the diagnoses.

Prediction of heart disease with high accuracy

The important objectives are:

1. High accuracy of prediction.
2. To help the health care sector.
3. To reduce the death rate of people without any awareness.

3.1 Advantages

1. Cost Effective.
2. Heart disease can be predicted earlier from the given attributes.
3. High accuracy in prediction.
4. Helps health sector in predicting the disease earlier.
5. Death rate of our country can be reduced.

3.2 Attribute Information:

1. Age
2. Sex
3. Chest Pain Type (4 Values)
4. Resting Blood Pressure
5. Serum Cholesterol In Mg/Dl
6. Fasting Blood Sugar > 120 Mg/Dl
7. Resting Electrocardiographic Results

8. Maximum Heart Rate Achieved
9. Exercise Induced Angina
10. Old peak = St Depression Induced By Exercise Relative To Rest
11. The Slope Of The Peak Exercise St Segment
12. Number Of Major Vessels (0-3) Colored By Fluoroscopy
13. Thal: 3 = Normal; 6 = Fixed Defect; 7 = Reversable Defect

3.3 Attributes types

Real	1,4,5,8,10,12
Ordered	11
Nominal	7,3,13
Binary	2,6,9

3.4 Experimental Results

The results of our experimental analysis in finding significant patterns for heart attack prediction are presented in this section. We have implemented our proposed approach in Java. The heart disease dataset that we have used for our experiments was obtained from [45].The detailed description of the dataset and the extracted significant patterns are given in the following subsections.

3.5 Heart Disease Dataset Description

The dataset used in our approach contains the parameters like blood pressure, cholesterol, chest pain, maximum heart rate and more. The detailed description of the parameters and their ranges are given as follows:

Value 1: upsloping
Value 2: flat
Value 3: downsloping

12 ca: number of major vessels (0-3) colored by flourosopy

13 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

14 num: diagnosis of heart disease (angiographic disease status)

Value 0: < 50% diameter narrowing
Value 1: > 50% diameter narrowing

4. Results

With the help of the dataset, the patterns significant to theheart attack prediction are extracted. The heart disease dataset is preprocessed successfully by removing duplicaterecords and supplying missing values. The refined heartdisease data set, resultant from preprocessing, is thenclustered using CART algorithm.

5. Conclusions and Future Work

Data mining in health care management is unlike the otherfields owing to the fact that the data present areheterogeneous and that certain ethical, legal, and socialconstraints apply to private medical information. Healthcare related data are voluminous in nature and they arrivefrom diverse sources all of them not entirely appropriate instructure or quality. These days, the exploitation ofknowledge and experience of numerous specialists andclinical screening data of patients gathered in a databaseduring the diagnosis procedure, has been widelyrecognized. In this paper we have presented an efficientapproach for extracting significant patterns from the heartdisease data warehouses for the efficient prediction ofheart attack.

The preprocessed heart disease data warehouse was clustered to extract data most relevant to heart attack using CART-means clustering algorithm. The frequent items have been mined effectively using CART algorithm. Based on the calculated significant weight age, the frequent patterns having value greater than predefined threshold were chosen for the valuable prediction of heart attack. In our future work, we have planned to design and develop an efficient heart attack prediction system with the aid

of these selected significant patterns using artificial intelligence techniques.

6. Reference

[1] Lee, H.G., Noh, K.Y., Ryu, K.H.: Mining biosignal data: Coronary artery disease diagnosis using linear and nonlinear features of HRV. In: Washio, T., Zhou, Z.-H., Huang, J.Z., Hu, X., Li, J., Xie, C., He, J., Zou, D., Li, K.-C., Freire, M.M. (eds.).

[2] Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network.

[3] Guru, N., Dahiya, A., Rajpal, N.: Decision Support System for Heart Disease Diagnosis Using Neural Network. Delhi Business Review 8(1) (January - June 2007).

[4] Szymanski, B., Han, L., Embrechts, M., Ross, A., Sternickel, K., Zhu, L.: Using Efficient Supanova Kernel For Heart Disease Diagnosis. In: Proc. ANNIE 2006, Intelligent Engineering Systems Through Artificial Neural Networks, vol. 16, pp. 305–310 (2006)

[5] Parthibanand, L., Subramanian, R.: Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm. International Journal of Biological, Biomedical and Medical Sciences 3(3) (2008).

[6] Frawley and Piatetsky-Shapiro, Knowledge Discovery inDatabases: An Overview. The AAAI/MIT Press, MenloPark, C.A, 1996.

[7] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, andWilliam Hersh, "Knowledge Management, Data Mining,and Text Mining In Medical Informatics", Chapter 1, pgs 3-34

[8] S Stilou, P D Bamidis, N Maglaveras, C Pappas , Miningassociation rules from clinical databases: an intelligentdiagnostic process in healthcare. Stud Health TechnolInform 84: Pt 2. 1399-1403, 2001.

[9] T Syeda-Mahmood, F Wang, D Beymer, A Amir, M
Richmond, SN Hashmi, "AALIM: Multimodal Mining forCardiac Decision Support", Computers in Cardiology,pages:209-212, Sept. 30 2007-Oct. 3 2007

[10] Anamika Gupta, Naveen Kumar, and VasudhaBhatnagar,"Analysis of Medical Data using Data Mining and FormalConcept Analysis", Proceedings Of World Academy OfScience, Engineering And Technology ,Volume 6, June2005,.

[11] SellappanPalaniappan, RafiahAwang, "Intelligent HeartDisease Prediction System Using Data Mining Techniques",IJCSNS International Journal of Computer Science andNetwork Security, Vol.8 No.8, August 2008

[12] Andreeva P., M. Dimitrova and A. Gegov, InformationRepresentation in Cardiological Knowledge Based System,SAER'06, pp: 23-25 Sept, 2006.

[13] LathaParthiban and R.Subramanian, "Intelligent Heart
Disease Prediction System using CANFIS and GeneticAlgorithm", International Journal of Biological, Biomedicaland Medical Sciences 3; 3, 2008

[14] HianChyeKoh and Gerald Tan ,"Data Mining Applicationsin Healthcare", Journal of healthcare informationmanagement, Vol. 19, Issue 2, Pages 64-72, 2005.

[15] L. Goodwin, M. VanDyne, S. Lin, S. Talbert ,"Data miningissues and opportunities for building nursing knowledge"Journal of Biomedical Informatics, vol:36, pp: 379-388,2003.

[16] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "MiningBiosignal Data: Coronary Artery Disease Diagnosis

usingLinear and Nonlinear Features of HRV," LNAI 4819:Emerging Technologies in Knowledge Discovery and DataMining, pp. 56-66, May 2007.

[17] Cristianini, N., Shawe-Taylor, J.: An introduction to SupportVector Machines. Cambridge University Press, Cambridge ,2000.

[18] Li, W., Han, J., Pei, J.: CMAR: Accurate and EfficientClassification Based on Multiple Association Rules. In:Proc. of 2001 Interna'l Conference on Data Mining .2001.

[19] Chen, J., Greiner, R.: Comparing Bayesian NetworkClassifiers. In Proc. of UAI-99, pp.101–108 ,1999.

[20] Quinlan, J.: C4.5: Programs for Machine Learning. MorganKaufmann, San Mateo 1993.

[21] "Hospitalization for Heart Attack, Stroke, or CongestiveHeart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.