Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.421 – 426

RESEARCH ARTICLE

Design & Analysis of Computational Features Prediction Model for Heart Disease Diagnosis

Atul Kumar Pandey*
atul.pandey.it2009@gmail.com *
Research Scholar
Department of Physics
Govt. PG Science College
Rewa (M.P.)-India*

Prabhat Pandey**
prabhatpandey51@gmail.com **
OSD
Additional Directorate
Higher Education, Division
Rewa (M.P.)-India**

K.L. Jaiswal***
drkanhaiyalaljaiswal@gmail.com ***
Assistant Professor
Department of Physics
Govt. PG Science College
Rewa (M.P.)-India***

Abstract— Heart disease prediction is designed to support clinicians in their diagnosis. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease prediction. Since the data is huge attribute selection method used for reducing the dataset. Then the reduced data is given to the classification. We also propose a new feature selection method algorithm which is the hybrid method combining CFS and RandomTree followed by part rule. The proposed algorithm provides better accuracy compared to the traditional algorithm and the hybrid Algorithm CFS. This research paper proposed a frequent feature selection method for Heart Disease Prediction. Good performance of this method comes from the use of the RandomTree and the PART rule. The nonadditivity of the RandomTree against different target attributes measure reflects the importance of the feature attributes as well as their interactions. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. Clustering the objects which have similar meaning, the proposed approach improves the accuracy and reduces the computational time.

Key Terms: - Data mining; Feature selection; classification

I. INTRODUCTION

The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition commonly called "Heart Attack" and the factors, which lead to such condition. Cardiomyopathy and Cardiovascular disease [2] 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) [3] results in severe illness, disability, and death. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack [3]. Chest pains arise when the blood received by the heart muscles is inadequate. High blood pressure, coronary artery disease, valvular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease. Life itself is completely dependent on the efficient operation of the heart. Cardiovascular disease is not contagious; you can't catch it like you can the flu or a cold. Instead, there are certain things that increase a person's chances of getting cardiovascular disease.

Knowledge discovery in databases is well-defined process consisting of several distinct steps [5]. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the nontrivial extraction of implicit previously unknown and potentially useful information about data". Data mining technology [5] provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service.

Cardiovascular disease (CVD) refers to any condition that affects the heart. Many CVD patients have symptoms such as chest pain (angina) and fatigue, which occur when the heart isn't receiving adequate oxygen. As per a survey nearly 50 percent of patients, however, have no symptoms until a heart attack occurs. A number of factors have been shown to increase the risk of developing CVD[4] Some of these are :

- Family history of cardiovascular disease
- High levels of LDL (bad) cholesterol
- Low level of HDL (good) cholesterol
- Hypertension
- High fat diet
- Lack of regular exercise
- Obesity

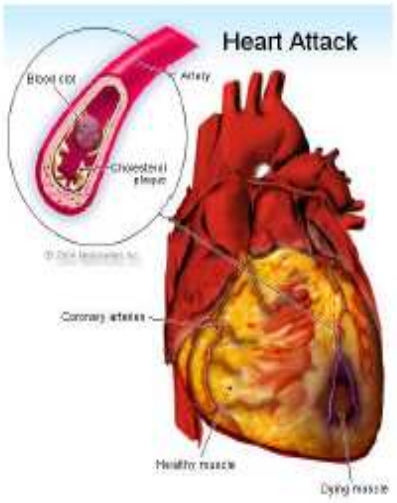

Fig .1 Heart

II. CLUSTERING

Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are "similar" and are "dissimilar" to the objects belonging to other clusters. This technique may be used as a preprocessing step [8] before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes.

III. CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

A learning classifier is able to learn based on a sample. The dataset used for training consists of information x and y for each data-point, where x denotes what is generally a vector of observed characteristics for the data-item and y denotes a group-label. The label y can take only a finite number of values.

IV. FEATURE SELECTION

The main purpose of feature selection [6] is to reduce the number of features used in classification while maintaining acceptable classification accuracy. For example, the Sequential Forward Floating Selection (SFFS)

algorithm [7] proposed by Pudil et al. was one of the commonly used algorithms. The main advantage of this method is that it produces a hierarchy of feature subsets with the best selection for each dimension. In our previous work, information gain is used to find the relevant features. Information gain [1] is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier.

Frequent Item set Mining (FIM) [7] is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-occur frequently in a dataset .It is of vital significance in a variety of Data Mining tasks that aim to mine interesting patterns from databases, like association rules, correlations, sequences, episodes, classifiers, clusters and the like. Numerous algorithms like the Apriori and FP-Tree have been proposed to support the discovery of interesting patterns. The proposed approach utilizes an efficient algorithm called MAFIA[8](Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm. The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of very long itemsets specifically. The depth-first traversal of the itemset lattice and effective pruning mechanisms are incorporated in the search strategy of the proposed algorithm.

The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm [7] to mine the frequent patterns present in it. Then the significance weightage of each pattern is calculated using the approach described in the following subsection. After mining the frequent patterns using MAFIA algorithm, the significance weightage of each pattern is calculated. It is calculated based on the weightage of each attribute present in the pattern and the frequency of each pattern [8].

V. PROPOSED METHOD

A. CFS and Decision Tree

We proposed a new hybrid feature selection method by combining CFS and Decision tree. The CFS algorithm reduces the number of attributes based on the SU measure, In CFS each attributes are compared pair wise to find the Similarity and the Attributes are compared to class attribute to find the amount of contribution it provides to the class value , based on these the attributes are removed. The selected attributes from the CFS algorithm is fed into Decision tree for further reduction. Decision tree calculates the conditional probability for each attribute and the attribute which has highest conditional probability is selected.

B. Dataset used in the Experiment

The following is the sample of the Heart Disease Data.arff @relation heart-statlog
@attribute age real
@attribute sex real
@attribute chest real
@attribute resting_blood_pressure real
@attribute serum_cholestoral real
@attribute fasting_blood_sugar real
@attribute resting_electrocardiographic_results real
@attribute maximum_heart_rate_achieved real
@attribute exercise_induced_angina real
@attribute oldpeak real
@attribute slope real
@attribute number_of_major_vessels real
@attribute thal real
@attribute class {absent, present}
@data
70,1,4,130,322,0,2,109,0,2.4,2,3,3,present
67,0,3,115,564,0,2,160,0,1.6,2,0,7,absent
57,1,2,124,261,0,0,141,0,0.3,1,0,7,present

The Heart Disease data after applying traditional method in Weka, The number high number of attributes reduced is 6 and then these attributes can be fed to various classifiers. The CFS+ Decision tree algorithm is coded, where the attribute afte**r CFS** is 6 and the selected attributes after RandomTree is only 4. CFS Feature selection method which selects the attributes based on the symmetrical uncertainty reduces the number of attributes from 13 to 4.The reduced attributes is fed to Decision tree followed by PART rule for further reduction.

The heart ARFF will contain large quantity of data and applying classification algorithms to this dataset is time consuming and also gives result with less accuracy. Hence we have to reduce the data set by using attribute selection method. Then this reduced dataset is fed into the four classification algorithm and which algorithm is

best fit for this prediction is investigated. Likewise, all other attribute selection and classification algorithms are applied for heart disease dataset. From that we identified that RandomTree classification algorithm gives better accuracy after applying the CFS attribute selection method.

C. Feature Selector

The best Feature Selection methods CFS are applied in sequence with different target attribute. (i.e), in this method the reduced number of attributes using frequent pattern mining method is 6. After applying the hybrid feature selector, the data is applied to the classification algorithm in which Random tree gives higher Accuracy comparing to the other classifiers.

After applying into Decision tree, the incorrectly classified instances are separated. The correctly classified samples are kept as training set and the incorrectly classified samples as test set are fed into various other classifiers, where the RandomTree gives greater accuracy.

D. CFS and Decision Tree

We proposed a new hybrid algorithm that is CFS+ Decision tree. When applying this feature selection algorithm, the attributes are reduced as 4. Then reduced dataset is given to classifiers. Here Random tree gives the greater accuracy compared to other classifiers.

VI. RESULT AND DISCUSSION

1. Attribute Selection

The feature selector method is automated, where the number of reduced attributes by CFS with different target attributes. Theorem is shown in Table 6.1.

Table 6.1: Reduced attributes by CFS with different Target Attributes

S.No.	Class Attribute against 14 total Attribute	Reduced Attributes
1	age	2,4,5,6,7,8,10,12,14 : 9
2	sex	5,13,14 : 3
3	Cp	6,8,9,10,14 : 5
4	trestbps	1,2,6,7,9,10,13,14 : 8
5	chol	1,2,3,7,9,12,14 : 7
6	fbs	3,7,11,13 : 4
7	restecg	3,6,11,14 : 4
8	thalach	1,9,11,14 : 4
9	exang	3,8,14 : 3
10	oldpeak	4,11,14 : 3
11	slope	6,8,10,14 : 4
12	ca	1,6,14 : 3
13	tha	2,6,8,14 : 4
14	num	3,7,8,9,10,12,13 : 7
Reduced Attribute with Minimum Support Value=5		3,6,8,9,10,14

Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. If the label is categorical (classification), accuracy is commonly reported as the rate which a case will be labeled with the right category. If the label is continuous, accuracy is commonly reported as the average distance between the predicted label and the correct value.

A confusion matrix displays the number of correct and incorrect Predictions made by the model compared with the actual classifications in the test data. The matrix is *n*-by-*n*, where *n* is the number of classes. From that we calculated the accuracy of each classification algorithms.

Table 6.2: Classifiers Accuracy with full dataset

S.No	Classifiers	Correctly Classified Samples	Incorrectly Classified Samples	Time(Seconds)	Accuracy (%)
1.	RandomForest	302	1	0.05	99.67
2.	RandomTree	303	0	0.00	100
3.	REFTree	255	48	0.02	84.1584
4.	J48	279	24	0.03	92.0792

5.	Simple k-means	175-0 cluster	128-1 cluster	0.03	58
6.	DensityBasedCluster	172-0 cluster	131-1 cluster	0.03	57
7.	Part	286	17	0.05	94.3894
8.	Decision table	251	52	0.09	82.8383

2. CFS and RandomTree

We proposed a new hybrid Feature selector combining CFS and RandomTree.

Table 6.3: Hybrid feature selection

Attribute Selection methods	Selected attributes
CFS+ RandomTree	4(3, 6, 9, 14)

Then this reduced data is given to the classification algorithms and calculate the accuracy for identifying the best algorithm.

Table 6.4: Reduced attributes by RandomTree classifier with different Target Attributes

S.No. of Attribute	Target Attribute	Correctly Classified Instances	Accuracy against different Target Attribute (%)
2	sex	303	100
3	Cp	303	100
6	fbs	301	99.3399
7	restecg	303	100
9	exang	303	100
11	slope	303	100
13	tha	301	100
14	num	303	100

From the above investigation, we have to conclude that CFS + Random Tree gives the better accuracy compared to the other algorithms.

VII. ISSUES AND CHALLENGES

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's 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. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

VIII. CONCLUSION

This paper proposed an efficient frequent feature selection method for Heart Disease Prediction. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. Real data from Health care organizations and agencies needs to be collected and all the available techniques will be compared for the optimum accuracy. The experimentation is conducted on dataset of health care domain. The new hybrid feature selection namely CFS and DT followed by Part rule was proposed. The proposed algorithm gives better accuracy for Random Tree and RandomForest classifier. We conclude that CFS, Decision Tree and part rule based feature selector is best suitable for heart disease data prediction.

We intend to extend our work applying various classification methods to predict the heart disease more efficiently.

REFERENCES

[1] Kwong-Sak Leung,kin hong Lee,Jin-Feng Wang,Eddie Y.T.Ng,Henry L.Y.Chan,Stephen K.W.Tsui,Tony S.K.Mok,Pete Chi-Hang Tse,Joseph Jao-yui Sung Data Mining on DNA Sequences of Hepatitis B virus IEEE/ACM Transactions on Computational Biology and Bioinformatics,Vol 8,No 2,March/April 2011

[2] Sunita Soni, Jyoti Soni,Ujma Ansari,Dipesh Sharma, Predictive Data Mining for Medical Diagnosis:An Overview of Heart Disease Prediction, International Journal of Computer Application (IJCA, 0975 – 8887) Volume 17– No.8, March 2011.

[3] Minas A. Karaolis, Member, IEEE, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis, Senior Member, IEEE, Assessment of the Risk Factors of Coronary Heart Events Based on

Data Mining With Decision Trees, IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 3, May 2010.

[4] Milan Kumari and Sunila Godara, Comparative study of Data Mining Classification Methods in Cardiovascular Disese Prediction ,IJCST Vol 2, Issue 2, June 2011.

[5] K.Srinivas, B.Kavihta Rani , A.Govrdhan , Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

[6] M.Anbarasi,E.Anupriya,N.Ch.S.N.Iyengar,Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[7] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

[8] Sellappan Palaniappan Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

