ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Effective Heart Disease Prediction using Frequent Feature Selection Method

S.Saravanakumar[1], S.Rinesh[2]

M.E-Third year, Department of Computer Science, Karpagam University, Coimbatore[1]

Asst. Professor (Sr), Department of Computer Science, Karpagam University, Coimbatore[2]

ABSTRACT: The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. The diagnosis of heart disease is a significant and tedious task in medicine. This research paper proposed a frequent feature selection method for Heart Disease Prediction. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral. The none additively of the fuzzy 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.

KEYWORDS: Knowledge discovery, fuzzy measure, non linear integral

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. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

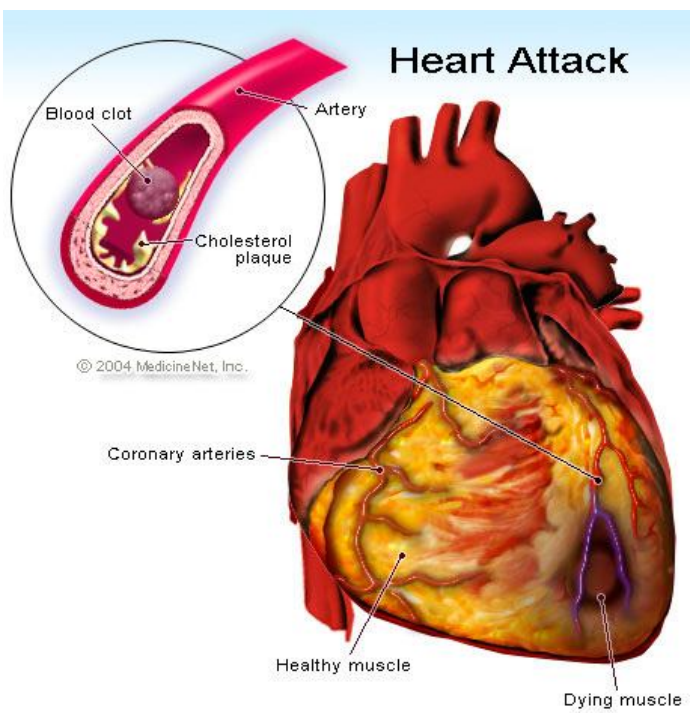


Fig .1 Heart

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[10] 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. 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 [8] 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.

However, we aim at global performance of the whole framework, so we adopt a simpler algorithm based on frequent method to select initial features. Frequent item sets capture all the dominant relationships between items in a dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Comparing to earlier techniques this paper has the advantage of using Clustering technique with the feature selection process. This is to group the attributes which are similar to each other..

3.1 Weighted Support

Weighted support as in [7]. WSP of X→classlabel, Where X is set of non empty subsets of attribute value set, is fraction of weight of the record that contains above attribute value set relative to the weight of all transactions.

S.No.	Attribute	Value
1	Age	Young,Middle, Old
2	Sex	Male,Female
3	Smoking	High,Medium, Low
4	Overweight	High,Medium, Low
5	Alcohol intake	High,Medium, Low
6	Hih Salt Diet	Yes,NO
7	High Saturated Fat diet	Yes,No
8	Exercise	High,Medium, Low
9	Sedentary Life style	Yes,No
10	Hereditary	Yes,No
11	Bad Cholestrol	High,Medium, Low
12	Blood Pressure	High,Medium, Low
13	Blood Sugar	High,Medium, Low
14	Heart Rate	High,Medium, Low

Table 1.Heart Disease Dataset

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

$$WSP(X \rightarrow classlable) = \frac{\sum_{i=1}^{|x|} weight(r_i)}{\sum_{k=1}^{|n|} weight(r_k)}$$

Example: Consider a rule R (Smoking="yes") →heart attack ="yes" then weighted support of R is calculated as:

$$WSP(R) = \frac{(Sum\ of\ record\ weight\ having\ the\ condition\ smoking = "yes"\ true\ and\ also\ given\ class\ label\ heart\ disease)}{Sum\ of\ weight\ of\ all\ transactions}$$

Table 2.Selected Attribute

S. No.	Attribute
1	Age
2	Sex
3	Smoking
4	Overweight
5	Alcohol Intake
6	Bad Cholestrol
7	Blood Pressure
8	Heart Rate

The Table 1 represents the relevant heart disease attributes with their values. The Table 2 represents the attributes which are selected by using the frequent feature selection algorithm.

3.2Frequent Pattern Mining Using MAFIA

Frequent Item set Mining (FIM) [8] 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Pseudo code for MAFIA [8]:

```
MAFIA(C, MFI, Boolean IsHUT) {
name HUT = C.head C.tail;
if HUT is in MFI
stop generation of children and return
Count all children, use PEP to trim the tail, and recorder by increasing support,
For each item i in C, trimmed_tail {
IsHUT = whether i is the first item in the tail
newNode = C I
MAFIA (newNode, MFI, IsHUT)}
if (IsHUT and all extensions are frequent)
Stop search and go back up subtree
If (C is a leaf and C.head is not in MFI)
Add C.head to MFI
}
```

The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm[8] 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.

3.3Significance Weightage Calculation

After mining the frequent patterns using MAFIA algorithm, the significance weightageof 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].
The formula used to determine the significant weightage (SW) is as follows:

$$Swi = \sum_{i=1}^{n} Wi\,Fi$$

Subsequently the patterns having significant weightage greater than a predefined threshold are chosen to aid the prediction of heart attack

$$SFP = \{x : Sw(x) \geq \Phi\}$$

Where SFP represents significant frequent patterns and Φ represents the significant weightage. This SFP can be used in the design of heart attack prediction system. Where SFP represents significant frequent patterns and Φ represents the significant weightage. This SFP can be used in the design of heart attack prediction system.

IV. CLASSIFICATION BASED ON NON LINEAR INTEGRALS

In classification, we are given a data set consisting of N example records, called the training set, where each record contains the value of a classifying attribute Y and the value of feature attributes $x_1, x_2, \ldots ; x_m$. Positive integer N is the data size.[1] The classifying attribute indicates the class to which each example belongs, and it is a categorical attribute with values coming from an unordered finite domain. The set of all possible values of the decisive attribute is denoted by $C = c_1, c_2, \ldots c_k$, where each $c_k, k = 1,2, \ldots, K$, refers to a specified class. The feature attributes are numerical, and their values are described by an m-dimensional vector,$(f(x_1),f(x_2),\ldots,f(x_m))$. The range of the vector, a subset of n-dimensional euclidean space, is called the feature space. [1]Thus, the j example record consists of the jth observation for all feature attributes and the classifying attribute, and is denoted by $(f_j(x_1),f_j(x_2),\ldots,f_j(x_m),Y_j),j=1,2,..N$ In this

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

section, a method of classification based on nonlinear integrals will be presented. It can be viewed as an idea of projecting the points in the feature space onto a real axis through a nonlinear integral, and then using a one dimensional classifier to classify these points according to a certain criterion optimally. Our classifying attributes holding the value of low, medium, high is memorialized to be a virtual variable. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral, since the nonadditivity of the fuzzy measure reflects the importance of the feature attributes, as well as their inherent interactions, toward the discrimination of the points. In fact, each feature attribute has its, respective, important index reflecting its amount of contribution toward the decision. A combination of the feature attributes may have a mutually restraining or a complementary synergy effect on their contributions towards the classification decision[9]. A relevant non linear integral is a good fusion tool to aggregate the information coming from the individual and the combinations of the feature attributes for the classification.

V. EXPERIMENTAL RESULT

Experiments were conducted with Weka 3.6.0 tool. Data set of 1000 records with 8 attributes is used. 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. With the help of the dataset, the patterns significant to the heart attack prediction are extracted using the approach discussed. The sample combinations of heart attack parameters for normal and risk level along with their values and weight ages are mentioned. In that, lesser value (0.1) of weightage comprises the normal level of prediction and higher values other than 0.1 comprise the higher risk levels.[8]

If
Male And age < 30 And Smoking = Never And Overweight = No And Alcohol = Never And
Stress = No And High saturated fat diet (hsfd) = No And High salt diet (hsd) = No And
Exercise = Normal And Sedentary Lifestyle (Inactivity) = No And Hereditary = No And Bad Cholesterol = Low And
Blood Sugar = Normal And Blood Pressure = Normal And Heart Rate
= Normal
Or
Male And age > 50 and age < 70 And Smoking = Current And Overweight = No And Alcohol = Past And Stress = No
And High saturated fat diet (hsfd) = No And High salt diet (hsd) =
Yes And Exercise = High And Sedentary Lifestyle (Inactivity) = No And Hereditary = No
And Bad Cholesterol = Low And Blood Sugar = Normal And Blood Pressure = Normal And
Heart Rate = Normal
Then
Risk Level = Normal
Otherwise If
Male And Age > 30 and age < 50 And Smoking = Current And Overweight = Yes And Alcohol=Current
And Stress = Yes And High saturated fat diet (hsfd) = No And High salt diet (hsd) =
Yes And Exercise = High And Sedentary Lifestyle (Inactivity) = Yes And Hereditary = Yes And Bad Cholesterol =
High And Blood Sugar = High And Blood Pressure = Low And Heart
Rate = Low Or High
Or
Female And Age >70 And Smoking = Never And Overweight = Yes And Alcohol = Past And
Stress = No And High saturated fat diet (hsfd) = Yes And High salt diet (hsd) = Yes And
Exercise = Never And Inactivity = No And Hereditary = Yes And Bad Cholesterol = High And
Blood Sugar = High And Blood Pressure = High And Heart Rate = High
Then
Risk Level

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

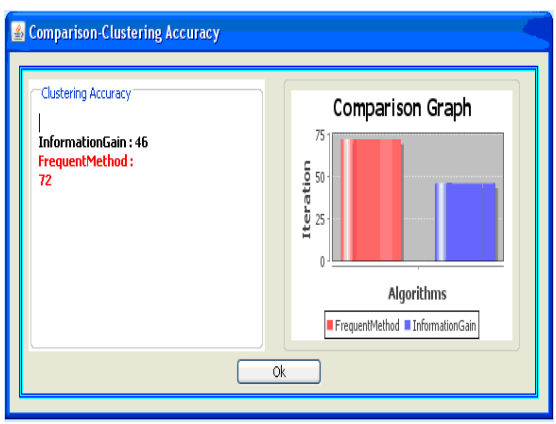
(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Fig 2 represents the accuracy between the information gain and the frequent feature selection method



VI. 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.

VII. CONCLUSION

This paper proposed a frequent feature selection method for Heart Disease Prediction. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral. The nonadditivity of the fuzzy 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. 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. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[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] Sunita Soni , Jyothi Pillai, O.P.Vyas, An Associative Classifier Using Weighted Association Rule , IEEE proceedings of the World Congress on Nature and Biologically Inspired Computing (NaBIC'09), December 09- 11, 2009, 1492-1496.

[8] 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

[9] Asha Rajkumar, G.Sophia Reena, Diagnosis Of HeartDisease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver.1.0 September 2010.

[10] 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.