DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES

M.A.Nishara Banu[1], B Gomathy[2]
[1]PG Scholar, [2]Assistant Professor (Sr. G)
Department of Computer Science and Engineering
Bannari Amman Institute of Technology
Sathyamangalam, India

Abstract— **The successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. The medical environment is still information rich but knowledge weak. There is a wealth of data possible within the medical systems. However, there is a lack of powerful analysis tools to identify hidden relationships and trends in data. Heart disease is a term that assigns to a large number of heath care conditions related to heart. These medical conditions describe the unexpected health conditions that directly control the heart and all its parts. Medical data mining techniques like association rule mining, classification, clustering is implemented to analyze the different kinds of heart based problems. Classification is an important problem in data mining. Given a database contain collection of records, each with a single class label, a classifier performs a brief and clear definition for each class that can be used to classify successive records . A number of popular classifiers construct decision trees to generate class models. The data classification is based on MAFIA algorithms which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree. The heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database.**

Keywords—Data mining; MAFIA (Maximal Frequent Itemset Algorithm); C4.5 Algorithm; K-means clustering

I. INTRODUCTION

Data mining is process of extracting hidden knowledge from large volumes of raw data.Datamining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans

Disease Prediction plays an important role in data mining. Data Mining is used intensively in the field of medicine to predict diseases such as heart disease, lung cancer, breast cancer etc.

This paper analyzes the heart disease predictions using different classification algorithms . Medicinal data mining has high potential for exploring the unknown patterns in the data sets of medical domain .These patterns can be used for medical analysis in raw medical data.Heart disease was the major cause of casualties in the world. Half of the deaths occur in the countries like India, United States are due to cardiovascular diseases. Medical data mining techniques like Association Rule Mining, Clustering, Classification Algorithms such as Decision tree,C4.5 Algorithm are implemented to analyze the different kinds of heart based problems. C4.5 Algorithm and Clustering Algorithm like K-Means are the data mining techniques used in medical field [1]. With the help of this technique, the accuracy of disease can be validated

II. RELATED WORKS

The difficult of recognizing constrained association rules for heart illness prediction was studied by Carlos Ordonez. The data mining techniques have been engaged by various works to analyze various diseases, for instance: Hepatitis, Cancer, Diabetes, Heart diseases. According to WHO (World Health Organization), heart disease is the main cause of death in the UK, USA, Canada, England [2]. Heart disease kills one in every 32 seconds in USA .25.4% of all deaths in the USA today are caused by heart disease. Jyothi Soni et.al [3] proposed for predicting the heart disease using association rule mining technique, they have generated a large number of rules when association rules are applied to dataset .Frequent Itemset Mining is used to find all frequent itemsets. Association rule mining methods like Apriority and FPgrowth are used frequently[4].Genetic algorithm have been used in [6], to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is one of the supervised learning methods to extract models describing important classes of data. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the Presence of heart disease in patients.

III. MAFIA

The association rule mining is a very important problem in the data-mining field with numerous practical applications, including consumer medical data analysis and network intrusion detection .

Maximal Frequent Itemset Algorithm (MAFIA) is an algorithm used for mining maximal frequent item sets from a transactional database [7].It integrates a depth-first traversal of the itemset lattice with effective pruning mechanisms. MAFIA efficiently stores the transactional database as a series of vertical bitmaps. . If support(X)=minSup, we say that X is a frequent itemset, and we denote the set of all frequent itemsets by FI .The process for finding association rules has two separate phases. In the first step, we find the set of frequent itemsets (FI) in the database T. In the second step, we use the set FI to generate “interesting” patterns, and various forms of interestingness have been proposed. In practice, the first step is the most time-consuming. Smaller alternatives to FI that still contain enough information for the second phase have been proposed including the set of frequent closed itemsets FCI.

```
Pseudocode:Simple(Current nodeC, MFI){
        For each itemiinC.tail {
newNode=C U i
if newNode is frequent
        Simple(newNode,MFI)}
if (Cis a leaf and C.head  is not in MFI)
        AddC.head to MFI
}
```

IV. C4.5 ALGORITHM

Decision trees are powerful and popular tools for classification and prediction [5]. Decision trees produce rules, which can be inferred by humans and used in knowledge system such as database. C4.5 is an algorithm for building decision trees .It is an extension of ID3 algorithm and it was designed by Quinlan .It converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. It handles discrete and continuous attributes. C4.5 is one of widely-used learning algorithms.

C4.5 algorithm builds decision trees from a set of training data using the concept of information entropy. C4.5 is also known as a statistical classifier.

- Check for base cases.
- For each element x, discover the normalized information gain from dividing on x.
    - Let x_best be the element with the highest normalized information gain.
- Create a decision node that breaks on a best.
- Repeats on the sublists obtained by dividing on x_best, and add those nodes as children of node.

V. K-MEAN CLUSTERING

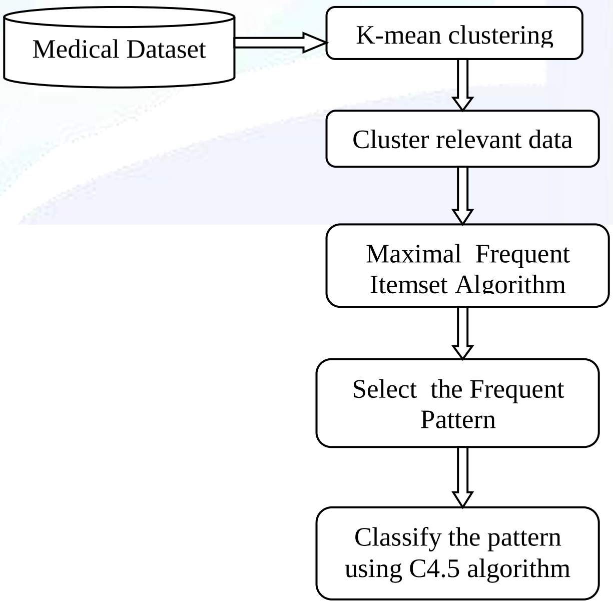
Clustering is a technique in data mining to find interesting patterns in a given dataset .The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters informations into k groups, where k is considered as an input parameter. It then assigns each information’s to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then more computed and the process begins again. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging and related fields. .K-Means algorithm is a divisive, unordered method of defining clusters. The phases convoluted in a k-means algorithm are given consequently:

Prophecy of heart disease using K – Means clustering techniques

- The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
- Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
- Each cluster center is recomputed as the average of the points in that cluster.
- Steps 2 and 3 repeat until the clusters converge. Convergence may be explained differently depending upon the performance, but it regularly explains that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

The clustering is performed on preprocessed data set using the K-means algorithm with the K values so as to extract relevant data to heart attack. K-Means clustering produces a definite number of separate, non-hierarchical clusters. K-Means algorithm is a disruptive, non-hierarchical method of defining clusters.

VI. SYSTEM ARCHITECTURE


Medical Dataset
K-mean clustering
Cluster relevant data
Maximal Frequent Itemset Algorithm
Select the Frequent Pattern
Classify the pattern using C4.5 algorithm

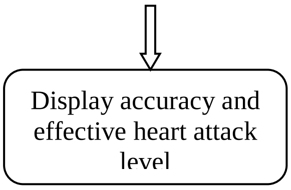


TABLE I. HEART DISEASE DATASET

KEY ID	KEY ATTRIBUTE
1	PatientId – Patient's identification number
2	Age in Year
3	Sex (value 1: Male; value 0: Female)
4	Chest Pain Type (value 1: typical type 1 angina,
5	typical type angina, value
6	non-angina pain, value 4: Asymptomatic)
7	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:
8	Serum Cholesterol (mg/dl)
9	Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave Abnormality: value 2: showing probable
10	Maximum Heart Rate Archieved; value (0.0) :> 0.0 and <=80, value (1.0) : >81 and <119,
11	Fasting Blood Sugar; 120
12	Exang - exercise induced angina (value 1: yes;
13	Old peak – ST depression induced by exercise
14	Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
15	CA – number of major vessels colored by floursopy (value 0-3)
16	Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)

VII. EXPERIMENTAL RESULTS

The results of our experimental results in identifying important patterns for predicting hear diseases are presented n this section. The heart disease database is preprocessed successfully by deleting corresponding records and providing missing values as shown in table I. The well mannered heart disease data set, resulting from preprocessing, is then collected by K-means algorithm with the K value of 2.One collection contains of the data related to the heart disease as shown in table II and the further contains the left over data. Then the regular forms are mined efficiently from the collection suitable to heart disease, using the MAFIA algorithm. The model consortiums of heart attack parameters for general and risk level along with their values and levels are listed below In that, ID lesser than of (#1) of weight contains the normal level of prediction and higher ID other than #1 comprise the higher risk levels and mention the prescription IDs. Table III display the parameters of heart attack prediction with equivalent prescription ID and their levels. Table IV show the example of training data to foresee the heart attack level and then figure 1 shows the efficient heart attack level with tree using the C4.5 by information gain

ID	REFERENCE ID	ATTRIBUTE
1	#1	Age
2	#2	Sex
3	#9	painloc: chest pain location
4	#16	Relrest
5	#18	cp: chest pain type
6	#21	trestbps: resting blood pressure
7	#24	chol: serum cholesterol in mg/dl
8	#27	Smoke
9	#28	cigs (cigarettes per day)
10	#31	years (number of years as a smoker)
11	#33	fbs: (fasting blood sugar > 120 mg/dl)
12	#36	dm (1 = history of diabetes; 0 = no such history)
13	#38	famhist: family history of coronary artery disease
14	#42	thalach: maximum heart rate achieved
15	#44	exang: exercise induced angina
16	#45	Sedentary Lifestyle/inactivity
17	#47	ca: number of major vessels (0-3) colored by fluoroscopy
18	#49	Hereditary
19	#51	num: diagnosis of heart disease

TABLE II.CLUSTER RELEVANT DATA BASED ON HEART DISEASE DATASET

C4.5 DECISION TREE STRUCTURE

If Age=<30 and Overweight=no and Alcohol Intake=never
Then
Heart attack level is Low
(Or)
If Age=>70 and Blood pressure=High and Smoking=current
Then
Heart attack level is high

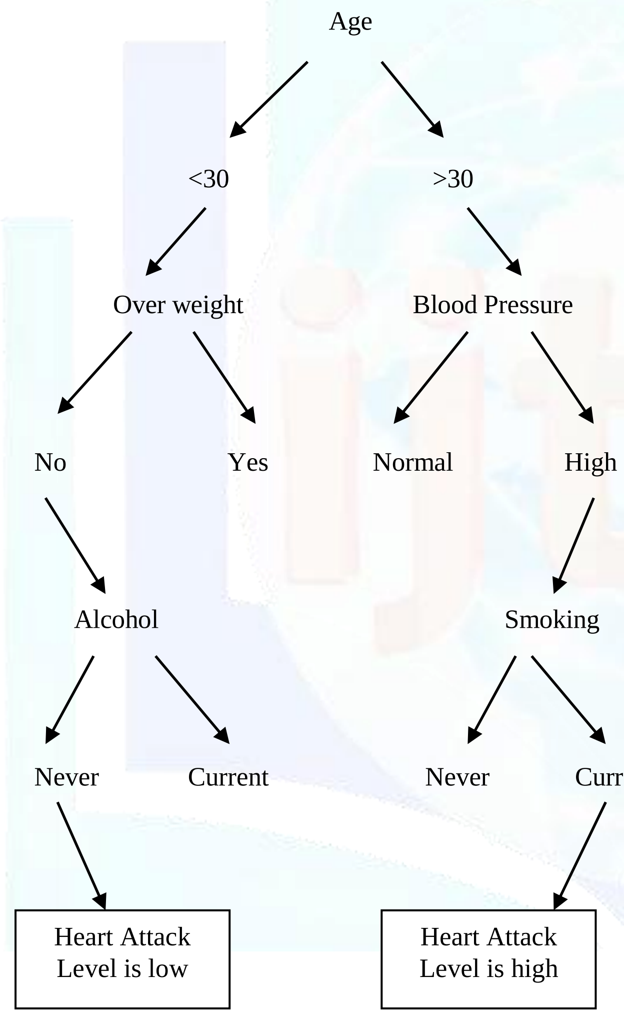


Figure 1: A decision tree for the concept heart attack level by information gain (C4.5)

TABLE III. HEART ATTACK PARAMETERS WITH CORRESPONDING PRESCRIPION IDS AND CONDITIONS

Parameter	Weight	Risk level
Male and Female	Age<30 Age> 30	#1 #8
Smoking	Never Past Current	#1 #3 #6
Overweight	Yes No	#8 #1
Alcohol Intake	Never Past Current	#1 #3 #6
High Salt Diet	Yes No	#9 #1
High saturated diet	Yes No	#9 #1
Exercise	Regular Never	#1 #6
Sedentary Lifestyle/inactivity	Yes No	#7 #1
Hereditary	Yes No	#7 #1
Bad cholesterol	High Normal	#8 #1
Blood Pressure	Normal (130/89) Low (< 119/79) High (>200/160)	#1 #8 #9
Blood sugar	High (>120&<400) Normal (>90&<120) Low (<90)	#5 #1 #4
Heart Rate	Low (< 60bpm) Normal (60 to 100) High (>100bpm)	#9 #1 #9

The experimental results of our approach as presented in Table IV. The goal is to have high accuracy, as well as high precision and recall metrics. These can be easily converted to true-positive (TP) and false-positive (FP) metrics.

Precision = TP/(TP+FP)
Recall = TP/(TP+FN)

- True Positive (TP): Total percentage of members classified as Class A belongs to Class A
- False Positive (FP): Total percentage of members of Class A but does not belong to Class A.
- False Negative (FN): Total percentage of members of Class A incorrectly classified as not belonging to Class A
- True Negative (TN): Total percentage of members which do not belong to Class A are classified not a part of Class A .It can also be given as(100%-FP)

TABLE IV COMPARISION BETWEEN SIMPLE MAFIA AND K-MEAN BASED MAFIA

Technique	Precision	Recall	Accuracy(%)
K-mean based MAFIA	0.78	0.67	74%
K-mean based MAFIA with ID3	0.80	0.85	85%
K-mean based MAFIA with ID3 and C4.5	0.82	0.94	94%

VI .CONCLUSION AND FUTURE WORK

Health care relevant data are enormous in nature and they arrive from various birthplaces all of them not wholly relevant in structure or quality. These days, the performance of knowledge, observation of various specialists and medicinal screening data of patients grouped in a database during the analysis process, has been widely accepted. In this paper we have presented an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses for the efficient prediction of heart attack. In our future work, we have planned to conduct experiments on large real time health datasets to predict the diseases like heart attack and compare the performance of our algorithm with other related data mining algorithms.

VII REFERENCES

[1] V. Manikantan and S. Latha ,“Predicting the analysis of heart disease symptoms using medicinal data mining methods”,International Journal of Advanced Computer Theory and Engineering, vol. 2 ,pp.46-51,2013.

[2] Shadab Adam Pattekari and Alma Parveen,“ Prediction system for heart disease using naïve bayes”, International Journal of Advanced Computer and Mathematical Sciences, vol.3,pp 290-294,2012.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni “Predictive data mining for medical diagnosis: an overview of heart disease prediction” International Journal of Computer Science and Engineering, vol. 3 ,2011

[4] R. Agrawal,T Imielinski ,and A. Swami ,‘Mining association rules between sets of items in large databases’

[5] Hnin Wint Khaing, “Data Mining based Fragmentation and Prediction of Medical Data”, International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2,2011

[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, pp.5370- 5376,2010.

[7] Douglas Burdick, Manuel Calimlim, Johanne Gehrke,“MAFIA: A Maximal Frequent Item set Algorithm For Transactional Databases”, Proceedings of the 17th International Conference on Data Engineering.

[8] K.Srinivas, Dr. G.Raghavendra Rao and Dr. A. Govardhan, “ A Survey On Prediction Of Heart Morbidity Using Data Mining Techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, No.3, May 2011

[9] S.Vijayarani, M. Divya, “ An Efficient Algorithm for Generating Classification Rules”, IJCST ,vol. 2, Issue 4, 2011