International Journal For Technological Research In Engineering
Volume 1, Issue 10, June-2014 *ISSN (Online): 2347 - 4718*

PREDICTION THE DISEASE ANALYSIS USING DATA MINING

Manju Saini[1], Prof. Dharmender Kumar[2]
Department of Computer Science and Application, Mewar University, Chittogarh Rajasthan

Abstract: Medicinal data processing strategies square measure wont to analyze the medical knowledge data resources. Medical data processing content mining and structure strategies square measure wont to analyze the medical knowledge contents. the trouble to develop information and skill of frequent specialists and clinical choice knowledge of patients collected in databases to facilitate the designation method is taken into account a valuable choice. Designation of cardiopathy could be a vital and tedious task in drugs. The term Heart malady encompasses the varied diseases that have an effect on the heart. The exposure of cardiopathy from varied factors or symptom is a problem that isn't complimentary from false presumptions usually in the midst of unpredictable effects. Association rule mining procedures square measure wont to extract item set relations. Item set regularities square measure utilized in the rule mining method. The information classification is predicated on MAFIA algorithms that end in accuracy, the information is valuated victimization entropy primarily based cross validations and partition techniques and also the results square measure compared. Here victimization the C4.5 rule because the coaching rule to indicate rank of heart failure with the choice tree. Finally, the center malady information is clustered victimization the K-means bunch rule, which can take away the information applicable to heart failure from the information. The results showed that the healthful prescription and designed prediction system is capable of prophesying the center attack with success.

Keywords: MAFIA(Maximal Frequent Itemset Algorithm), C4.5, K-means.

I. INTRODUCTION

Data Mining is a dynamic area. One of the most popular approaches to do data mining is discovering Medical decision support systems are designed to support clinicians in their diagnosis. The prediction of heart disease pattern with classification algorithms is proposed here. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease classification. A dimensionality of the data is reduced by attribute selection methods. This cleaned data is classified by different classification algorithms such as MAFIA, K-Means, C4.5 Algorithm. This kind of classification be likely to optimize the use of data storage for numerous purposes - practical, directorial, legal, data can be classified according to any criteria, not only based on the relative position or regularity of use.

A. Causes of casualties

Heart disease was the major cause of casualties in the many countries including India. Heart diseases kill one person every 32 seconds in the United States. 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.

B. Medical Data mining techniques

Major medical data mining techniques are implemented to analyze the different kinds of heart based problems. The data about the combinations of heart disease should be used in finding the efficient data retrieval in data mining. C4.5 algorithm and K-Means Clustering algorithm are widely used in the basic medical data mining techniques. This technique widely used to validate the accuracy of medical data.

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 in the works to analyze various diseases, for instance: Hepatitis, Cancer, Diabetes, Heart diseases. Frequent Item set Mining (FIM) is measured to be one of the basic data mining difficulties that expects to discern collections of items or values or forms that co-occur regularly in a dataset. The term Heart illness covers the various diseases that affect the heart. Heart disease was then major source of fatalities in the United States of America, England, and Canada. Heart disease kills one in every 32 seconds in the United States of America. This technique is used while prescribing the patient and this system predicts which remedy in the form of medicines and medical test suits best.

III. MAXIMAL FREQUENT ITEM SET ALGORITHM

The association rule problem is a very important problem in the data-mining field with numerous practical applications, including consumer medical data analysis, inferring patterns from web page access logs, and network intrusion detection. The association rule model and the support-confidence framework were originally proposed by few authors. Let I be a set of items (we assume in the remainder of the paper without loss of generality I={1,…,N}).We call X ⊆ I an itemset, and we call X a k-itemset if the cardinality of itemset X is k. Let database T be a multiset of subsets of I, and let support (X) be the percentage of itemsets Y in T such that X ⊆ Y. Informally, the support of an itemset measures how often X occurs in the database. If support(X)=minSup, we say that X is a frequent itemset, and we denote the set of all frequent itemsets by FI. If X is recurrent and no superset of X is frequent, we say that X is a maximally frequent itemset, and we denote the set of all maximally frequent itemsets by MFI. The process for finding association rules

has two separate phases. In the first phase, 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 phase 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 itemsetsFCI. An itemset X is closed if there does not exist an itemset X' such that $X \subseteq X'$ and $t(X) = t(X')$,witht(Y) defined as the set of transactions that contain itemset Y. It is straightforward to see that the following relationship holds: MFI⊆FCI⊆FI.

A. Sample algorithms and pseudocodes

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 inMFI)
AddC.head to MFI
}

IV. C4.5 ALGORITHM

C4.5 is a program that contributions a set of labeled data and produces a decision tree as output. This follow-on decision tree is then verified against unseen labeled test data to enumerate its generalization. C4.5 is a program used for generating taxonomy rules using decision trees from a set of given data. C4.5 algorithm is an extension of the basic ID3 algorithm and it was designed by Quinlan. C4.5 is one of widely-used learning algorithms. C4.5 algorithm builds decision trees from a set of training data similar to the ID3algorithm, using the concept of information entropy. C4.5 is also known as a statistical classifier.

Sample algorithm

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

At each node of the tree, C4.5 chooses one attribute of the data that splits its set of samples into subsets enriched in one class or the other.

V. K-MEANS CLUSTERRING ALGORITHM

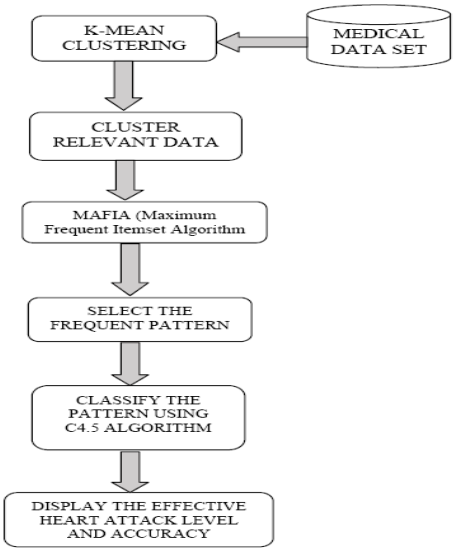
The classification of objects into various groups or the segregating of dataset into subcategories so that the data in each of the subcategory share a common article, often the proximity with respect to some definite space degree is known as Clustering. The clustering difficult has been well-known in numerous situations and addressed being proven beneficial in many medical data mining applications. Clustering the medical data into small with meaningful data can aid in the discovery of forms by supporting the abstraction of several suitable features from each of the collections thereby introducing party into the data and helping the application of orthodox data mining techniques. Copious procedures are presented in the works for clustering and engaged the renowned K-Means clustering algorithm in this approach. K-means groups the data in accord with their individual values into k distinct collections. Data categorized into the identical cluster have alike feature values. K, the positive number representing the number of collections, needs to be delivered in advance. The phases convoluted in a k-means algorithm are given consequently:

Prophecy of heart disease using K–Means clustering techniques

- K points denoting the data to be bunched are positioned into the space. These points signify the primary collection centroids.
- The data are consigned to the group that is nearby to the centroid.
- The points of all the K centroids are again calculated as swiftly as all the data are allotted.
- Steps 2 and 3 are repeated until the centroids stop affecting any further. This results in the isolation of data into groups from which the metric to be diminished can be reflected.

The preprocessed heart illness data is grouped using the K-means algorithm with the K values. Clustering is a type of multivariate statistical examination also known as cluster analysis, unsupervised classification analysis, or numerical taxonomy. K-Means clustering produces a definite number of separate, flat (non-hierarchical) clusters. It is sound suitable to producing orbicular constellations. The K-Means method is numerical, unsubstantiated, non-deterministic and iterative

VI. SYSTEM ARCHITECTURE



VII. EXPERIMENTAL RESULTS

The results of our experimental analysis in discovering substantial forms for heart attack prophecy are presented in this section. With the help of the database, the forms substantial to the heart attack prophecy are mined using the method deliberated. The heart disease database is

preprocessed successfully by eliminating identical records and providing missing values as shown in table I. The polished 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 recurrent forms are mined efficiently from the collection applicable to heart disease, using the MAFIA algorithm. The model consortiums of heart attack parameters for ordinary and risk level along with their values and levels are detailed 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.

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, value
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 or definite left ventricular hypertrophy)
10	Maximum Heart Rate Archieved; value (0.0) :> 0.0 and <=80, value (1.0) : >81 and <119, value (2.0):=120;
11	Fasting Blood Sugar; 120
12	Exang - exercise induced angina (value 1: yes; value 0: no)
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)

Table 1: Heart Attack Dataset

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 2. Grouping Relevant Data Based On Heart Attack 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

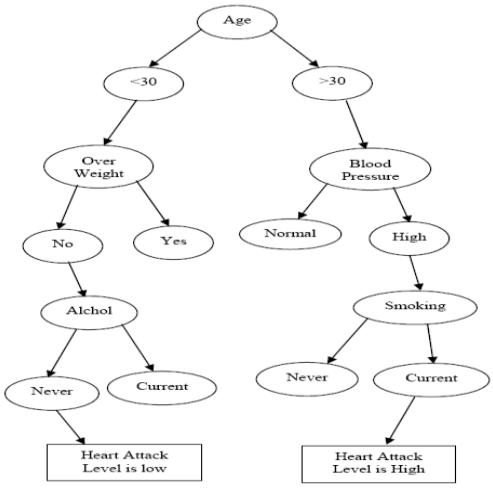


FIg.1: A decision tree for the concept heart attack level by information gain (c4.5)

PARAMETER	CONDITIONS	PRESCRIPTION ID
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

Table 3 : Heart attack parameters with corresponding prescripion ids and conditions

The experimental results of our approach as presented in Table IV. The goal is to have high accuracy, besides high precision and recall metrics. These metrics can be derived from the confusion matrix and can be easily converted totrue-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).

Technique	Precision	Recall	Accuracy(%)
K-Mean based MAFIA	0.78	0.67	74%
K-Mean base MAFIA with ID3	0.80	0.85	85%
K-Mean based MAFIA with ID3 and C4.5	0.82	0.92	92%

Table 4: Comparison Between Simple Mafia and Proposed K-Mean Based Mafia

VIII. CONCLUSION AND FUTURE WORK

Health care related data are huge in nature and they arrive from various birthplaces all of them not wholly suitable in structure or quality. These days, the utilization of knowledge and experience of copious specialists and medical screening data of patients collected in a database during the diagnosis 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 design and develop an efficient heart attack prediction system with Patient Prescription Support using the web mining and data warehouse techniques.

REFERENCES

[1] Mai Shouman, Tim Turner and Rob Stocker, Using Data Mining Techniques In Heart Disease Diagnosis And Treatment, Japan-Egypt Conference on Electronics, Communications and Computers, 2012

[2] T.John Peterand K. Somasundaram, An Empirical Study On Prediction Of Heart Disease Using Classification Data Mining Techniques, ISBN: 978-81-909042-2-3, 2012

[3] Hnin Wint Khaing, Data Mining based Fragmentation and Prediction of Medical Data, ISBN: 978-1-61284-840-2, 2011

[4] S Satapathy and S Chattopadhya, Mining Important Predictors Of Heart Attack, International Conference On Advances In Recent Technologies In Communication And Computing 2011

[5] S. el Rafaie, Abdel-Badeeh M. Salem and K. Revett, On the Use of SPECT Imaging Datasets for Automated Classification of Ventricular Heart Disease, The 8th International Conference on INFOrmatics and Systems (INFOS2012)

[6] S.Vijayarani, M.Divya, An Efficient Algorithm for Generating Classification Rules, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print), IJCST Vol. 2, Issue 4, 2011

[7] Geetika, A Survey of Classification Methods and its Applications, International Journal of Computer Applications (0975 – 8887) Volume 53– No.17, September 2012

[8] G.Subbalakshmi, K. Ramesh, M. Chinna Rao , Decision Support in Heart Disease Prediction System using Naive Bayes, ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[9] Bala Sundar V, T Devi, N Saravanan, Development of a Data Clustering Algorithm for Predicting Heart, International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012

[10] Shadab Adam Pattekari and Asma Parveen, Prediction System For Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[11] K.Srinivas, Dr. G.Raghavendra Rao and Dr. A.Govardhan, 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

