ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Resolution of Heart Disease Prevision System

E. Nalini, T.Elavarasi and C. Ezhil Kumaran
MS. Software Engineering (Project Students)
VIT University,
Vellore, India.
c.ezhil@yahoo.co.in

Mr. D .Rajesh*
Assistant Professor
SITE, VIT University,
Vellore, India.
drajesh@vit.ac.in

Abstract: In our day to day life, we can see that number of heart disease patients keep on increasing but heart disease prediction methods are limited. Resolution of Heart Disease Prevision System (RHDPS) is used to discover hidden patterns of information for the doctors in making their decisions accurately. Our system initiates by gathering the patients details like age, sex, ECG results, cholesterol level, BP, etc. and updates it regularly. To analyze the various hidden patterns we apply the data mining algorithms like regression analysis, clustering methods and association in an comparative nature. Our system finally predicts the possibility of heart disease very quickly and easily. The system is implemented and tested on java platform.

Keywords: Data mining, Clustering, Association Rules, Prediction, AprioriHybrid.

1. INTRODUCTION

Our system acts as a prototype to predict the heart attack disease in humans and also encompasses the diverse disease that affects the heart. Diagnosis of heart disease is significant and tedious task in medicine. Thus our system compares other factor and symptoms to provide effective care for patients. Many information systems are not utilized properly by doctors since it contains huge amount of information. There are many tools to predict heart disease but they are not efficient. But our system RHDPS is used efficiently with data mining technology.

Our system focuses on how the system answers complex queries. Rather than doctors own Knowledge and experience, the system used the Data mining technique like clustering, Association rule and prediction which provides good clinical decision support. Our system extracts hidden information from databases which provides valuable knowledge. Appropriate computer-based information and/or decision support system can aid in achieving clinical tests at a reduced cost.

The existing heart disease patient database contains the patients details with attributes related to heart disease. The user input consists of same attributes. The input is compared with database in order to predict the heart disease. If it is found to contain similar data, the data mining techniques are applied. If it is found to be dissimilar, it is compared with another database which contains generic symptoms of heart disease patients. The basic attributes include

A. Input Attributes

a. Sex (value 1: Male; value 0 : Female)
b. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
c. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
d. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
e. Exang – exercise induced angina (value 1: yes; value 0: no)
f. Slope – the slope of the peak exercise ST segment (value 1: up sloping; value 2: flat; value 3: down sloping)
g. CA – number of major vessels colored by floursopy (value 0 – 3)
h. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
i. Trest Blood Pressure (mm Hg on admission to the hospital)
j. Serum Cholesterol (mg/dl)
k. Thalach – maximum heart rate achieved
l. Oldpeak – ST depression induced by exercise relative to rest
m. Age in Year

II. CLUSTER ANALYSIS

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

A. K-MEANS Algorithm

The k-means algorithm follows a simple and easy way to classify a given dataset through certain number of

clusters. The main idea is to define k centroids, one for each cluster. The better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given dataset and associate it to the nearest centroid. We need to recalculate k new centroids for the clusters resulting from the previous step. After we have these k new centroids, a new bindings has to be done between the same data set points and the nearest new centroid.

Finally, this algorithm aims at minimizing an objective function, a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j.

The algorithm is composed of following steps:

a. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids.

b. Assign each object to the group that has the closest centroid.

c. When all objects have been assigned, recalculate the positions of the k centroid.

d. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of objects into groups.

III. ASSOCIATION RULE MINING

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules are employed today in many application areas including detection and bioinformatics.

Association rules are required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. To achieve this, association rule generation is a two-step process. First, minimum support is applied to find all frequent item sets in a database. In second step, these frequent item sets and the minimum confidence constraints are used to form rules. While second step is straight forward, the first step needs more attention.

A. Apriori Hybrid Algorithm

The AprioriHybrid algorithm is combination of Apriori and AprioriTid. The Apriori algorithm makes multiple passes over the database. In the first pass individual item support is counted and the item that have support greater than or equal to minimum support are considered as large items. For each pass K after that, the large item sets in the previous pass is grouped in sets of k items and these from candidate item sets. The support for the different candidate item sets is counted and if the support then that item set is considered to be large. This process continues until the large item set in a particular pass comes out to be an empty set.

Similar to Apriori algorithm, AprioriTid algorithm also uses the Apriori-gen function to determine the candidate item sets but the difference is that the database is not used for counting support after the first pass. Instead set of candidate item sets is used for this purpose for k>1. In case a transaction does not have any candidate k-item set then the set of candidate item sets would not have any entry for that transaction which will eventually decrease the number of transaction in the set containing the candidate item sets as compared to the database. As value of k increases each entry will be smaller than the corresponding transaction will decrease. In Apriori Hybrid algorithm, Apriori algorithm is used for initial phases and AprioriTid used in later phases for better performance.

B. Apriori Algorithm

The AprioriHybrid algorithm is combination of Apriori and AprioriTid. In Apriori algorithm, the first pass consists of counting the occurrences of itemsets. A pass say k consists of two phases. First, the large itemsets L_{k-1} found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the Apriori gen function. Next the database is scanned and the support of candidate inC_k. For fast counting, we need to efficiently determine the Candidates in C_k that are contained in a given transaction t.

K-item set ***An item set having k items.***

Lk	Set of large k-itemsets (those with minimum support). Each member of this set has two fields: i) Item set and ii) support count.
Ck	Set of candidate k-itemsets (potentially large itemsets). Each member of this set has two fields: i) item set and ii) support count.
Ck	Set of candidate k-itemsets when the TIDs of the generating transactions are kept associated with the candidates.

C. Pseudo Code

a) L1 = large 1-itemsets;
b) for (k = 2; $L_{k-1} \neq 0$; k++) do begin
c) C_k=Apriori-gen (L_{K-1}); // New candidates
d) For all transactions t € D do begin
e) C_t= subset (C_k, t); // Candidates contained in t
f) For all candidates c € C_t do
g) c.count++;
h) End
i) L_k= c € C_k | c.count ≥minsup
j) End
k) Answer = U_K L_k;

D. Apriori Candidate Generation

Insert into C_k
Select p.$item_1$, p.$item_2$, ..., p.$item_{k-1}$, q.$item_{k-1}$
From L_{K-1} p, L_{K-1} q
Where p.$item_1$ = q.$item_1$, . . ., p.$item_{k-2}$ = q.$item_{k-2}$, p.$item_{k-1}$ < q.$item_{k-1}$;

E. Prune Step

In prune step, we delete all itemsets c € C_k such that some (k-1) subset of c is not in L_{K-1}.

C_k= X U X | X, X € L_{K-1},| X ∩ X | = k-2

C_k= { X € $\overline{C_k}$ | X contains k members of L_{k-1} }

F. AprioriTid Algorithm

Aprioritid also uses gen function to determine the candidate itemsets. The feature of this algorithm is that the database {D} is not used for counting support after the first pass. Instead, the set C_k is used for this purpose. Each set in C_k is of the form < TID, X_k >, where X_k is a large k-itemset present in the transaction with identifier TID. For k=1, C_1 corresponds to the database D, each item I is replaced by the itemset {i}. For k>1, C_k is generated by the algorithm. If a transaction does not contain any candidate k-itemset, then C_k will not have an entry for this transaction. Thus the no of entries in C_k may be smaller than the no of transactions in the database.

a) L_1 = { large 1- itemsets };
b) C_1 = database D;
c) For (k=2;L_{K-1}≠0;k++) do begin
d) C_k= Apriori-gen(L_{K-1}); // New candidates
e) C_k=0;
f) For all entries t € C_{k-1} do begin
g) // determine candidate itemsets in C_k contained
// in the transaction with identifier t.TID
C_t={ c € C_k| (c –c[k]) € t.set of itemsets □ (c – c[k-1]) € t.set of itemsets };
h) for all candidates c € C_t do
i) c.count ++;
j) if (C_t≠0) then C_k += <t.TID, C_t>;
k) End
l) L_k= { c € C_k| c. count ≥ minsup}
m) End
n) Answer= $U_K L_k$;

Example

Database

TID	Items
TID1	1 3 4
TID2	2 3 5
TID3	1 2 3 5
TID4	2 5

C_1

TID	Set of itemsets
TID1	{ {1},{3},{5} }
TID2	{ {2},{3},{5} }
TID3	{ {1},{2},{3},{5} }
TID4	{ {2},{5} }

L_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	support
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

$\overline{C_2}$

T I TID	Set of itemsets
TID1	{ {1 3} }
TID2	{ {2 3},{2 5},{3 5} }
TID3	{ {1 2},{1 3},{1 5},{2 3},{2 5},{3 5} }
TID4	{ {2 5} }

L_2

Itemset	support
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

Itemset	Support
{2 3 5}	2

$\overline{C_3}$

TID	Set of itemsets
TID2	{ {2 3 5} }
TID3	{ {2 3 5} }

L_3

Itemset	support
{2 3 5}	2

The patterns significant to heart disease is extracted from datasets. It is extracted based on maximum support count. The result is predicted based on the significant patterns.

IV. CONCLUSION

The Resolution of Heart Disease prediction system is developed using data mining techniques like Clustering,Association rule and prediction.Our system extracts hidden knowledge from Heart Disease database.Structured Query Language is used to create database. Oracle (or) Mysql server tool is used to execute queries. The symptoms with highest support count are considered as the relevant patterns related to heart disease.If the input founds to contain those extracted patterns, then the heart disease of patient is identified. The result is updated to patient database so that other user inputs can also be compared.

V. FUTURE WORK

In our paper we used clustering and assosiation algorithm.In future, Text mining can also be used and other datamining techniques like Time Series and regression can be used.

VI. REFERENCES

[1] Application of data mining techniques in healthcare and prediction of heart attacks. K.Srinivas, B.Kavihta Rani, Dr. A.Govrdhan 2010.
[2] Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network-Shantakumar B.PatilY.S.Kumaraswamy 2009.
[3] Extraction of significant patterns from heart disease warehouse for heart attack prediction Shantakumar, B.Patil Dr.Y.S.Kumaraswamy 2009
[4] Empirical study on applications of data mining techniques in healthcare Harleen Kaur and Siri Krishan Wasan 2006
[5] From Data Mining to Knowledge Discovery in Databases Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth
[6] Data Mining and data warehousing Han J.kamber
[7] CLICK: Clustering Categorical Data using K-partite Maximal Cliques Markus Peters and Mohammed J. Zaki Computer Science Department Rensselaer Polytechnic Institute Troy NY 12180
[8] Fast Algorithms for mining Association rules Rakesh Agrawal, Ramakrisnan Srikant IBM Almaden Research Center