Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLUSTER BASED ASSOCIATION RULE MINING FOR HEART ATTACK PREDICTION

[1]MA.JABBAR, [2] Dr.PRITI CHANDRA, [3] *B.L.DEEKSHATULU*
[1]Research scholar, JNTUH, Hyderabad, India
[2]Scientist, Advanced System Laboratory, Hyderabad, India
[3]Distinguished Fellow IDRBT, RBI (Govt.of INDIA), Hyderabad, India
E-mail: jabbar.meerja@gmail.com, priti_murali@ yahoo.com,deekshatulu@hotmail.com

ABSTRACT

This Paper Focuses on Analyzing Medical Data using Association Rule mining for Heart Attack Prediction. Association rule mining is one of the fundamental research topics in data mining and knowledge discovery that finds interesting association or correlation relation ship among a large set of data items and predicts the associative and correlative behaviors for new data. The Association rule mining algorithms must perform efficiently. In this paper we suggest a new approach for association rule mining based on sequence number and clustering the transactional data base for heart attack prediction. Experimental results shows effectiveness of our proposed algorithm

Keywords: *Association rule mining, Sequence number, Clustering, Heart attack*

1. INTRODUCTION

Data mining is one of the component of the broader process known as knowledge discovery in data bases(KDD).Berry and linoff [1] define data mining as the evaluation and analysis by automatic or semi automatic means of large quantities of data in order to discover meaningful patterns and rules[2].Knowledge discovery in data bases(KDD) was initially named by Gregory Piatetsky Shapiro in a work shop at 1989 international joint conference on AI.In 1996 Fayyad et all[2] defined KDD as nontrivial process of identifying valid,novel,potentially useful and ultimately understandable patterns in data. Association rule mining one of the most important and well researched techniques in data mining was introduced by R.Agrawal in 1993[3].

The problem of association rule mining can be decomposed into two sub problems.

1) Discovering frequent item sets

2) Generating rules from frequent item sets

Overall performance of mining association rules are determined by 1st step.

Association rules must satisfy two important basic measures namely Support and Confidence. Assume the minimum support and minimum confidence are ms and mc respectively which are given by experts and users. Then X=>Y is a valid association rule if support>=ms and confidence>=mc

Apriori algorithm is the most well known association rule mining algorithm and is used in most commercial products. It uses the following property, which we call the large item set property [4].

"Any subset of a large item set must be large"

The large item sets are also said to be downward closed because if an item set satisfies the minimum support requiremets,so do all of it's subsets. Apriori algorithm suffers from the following bottlenecks

1) Repeated number of data base scans

2) Huge candidate sets.

Heart attack is called as myocardial infarction in medical terminology. Heart is supplied by right and left coronary arteries. Whenever these arteries are blocked, blood supply to heart stops and wall of heart damages, resulting in heart attack.

Coronary heart disease is epidemic in India and one of the major causes of disease burden and deaths. Data from registrar general of India shows that heart attacks are major cause of death in India. Studies to determine the precise cause of death in urban Chennai and rural areas of A.P have revealed that cardio vascular disease cause about 40%of the deaths in urban and 30%in rural areas [5]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

World health organization in the year 2003 reported that 29.2%of total global deaths are due to Cardio Vascular Disease (CVD). Cardio vascular disease is expected to be leading cause of deaths in developing countries due to change in life style, work culture and food habits. Hence more careful and efficient methods of cardiac diseases and periodic examination are of high importance [6]

Many hospital systems are designed to support patient billing, inventory management and simple statistics. They can't answer complex queries like "given patient records, predict the probability of patient getting a heart disease". Data mining have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

The objective of this paper is to mine the patterns and relationships associated with heart disease from heart disease data base using association rule mining.

2. PRELIMANARIES

Definition 1(Association rule): Given a set of items $I=\{I_1, I_2, I_3,.....I_N\}$and data base of transactions D={T1,T2…Tn}where Ti={I1,I2,I3,I4…In}.An association rule is an implication of the form X=>Y where $X,Y \subseteq I$ are sets of items called item sets and $X \cap Y=\emptyset$

Definition 2(Support): the support for an association rule X=>Y is the percentage of transactions in the data base that contain XUY.

Support(X=>Y) =P (XUY)

Definition 3(Confidence): confidence or strength is the ratio of number of transactions that contain XUY to the number of transactions that contain X.

Confidence= Support (XUY)/support(X)

Definition 4(Frequent item set): item set whose occurrences is above support threshold

Definition 5(sequence number): it is group of numbers, here these numbers may be repeated, and each number is called a sub item of sequence number. [7]

Let SN= {23, 45, and 67,126,178} be a Sequence number 67 is called sub item of SN.

Definition 6(sequence number Degree): Sequence number degree (SND) is an integer, which is equal to sum of items SID contained by SN.

For example let SN=={23,45,67,126,78,65,111}

Then

SND=SID(23)+SID(45)+SID(67)+SID(126)+SID(78)+SID(65)+SID(111)=28

Definition 7(Sub item degree): it is an integer, which is equal to the number of" 1"contained by binary code of sub item. Let 45 be a sub item, and then SID (45) =SID (101101) =4

Definition 8(Skipping Fragments): if D is a database consists of 12 transactions, and if D is partitioned into 3 skipping fragments that have more dissimilar transactions

P1={T1,T4,T7,T10},P2={T2,T5,T8,T11},P2={ T3,T6,T9,T12}

3. RELATED WORK

The very first well known algorithm for frequent item set generation is Apriori algorithm. It works on the principle of Apriori property, which states that the subset of any frequent item set is also frequent. Apriori algorithm adopts layer by layer search iteration method to mine association rules which is frequent k-1 item set L K-1 is used to search k item set L K.The Apriori algorithm suffers from the following 2 problems

1) Candidate generation
2) Repeated no. of scans

Various techniques have been adopted to improve the efficiency of the of the apriori algorithm

They are 1) hash based technique 2) transaction reduction 3) partitioning4) sampling5) dynamic item set counting

Second well known algorithm is FP-Tree algorithm which is proposed by j.w Han et.all for frequent pattern mining, which compresses data items into FP-Tree [8]. It removes the candidate generation and test approach. It uses recursive, divide and conquer top down approach for frequent item set generation. It scans data base two times .so it's efficiency is 10 times than apriori algorithm. Although FP-Tree algorithm is more efficient than apriori algorithm it suffers from the following problems.

1) space requirement is more
2) More scalability is needed for large applications.

Gang Fang [7] etc all proposed an algorithm for improved association rule mining. They used the method of binary Boolean calculation to generate candidate frequent item sets by computing sequence number degree, which is

Gained through computing these sequence number of all these items obtained by candidate frequent item sets .

Wael a.Alzoubi etc all[9] proposed scalable and

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Efficient method for mining association rules based On Clustering.

Sellappan Palaniappn etc all[10] proposed intelligent Heart disease prediction system using data mining Techniques namely decision trees, naïve bayes and Neural networks .their results shows that each Technique has its unique strength in realizing the Objectiveness of the defined Mining goals.

Carlos Ordonez[11] implemented efficient searchfor Diagnosis of heart disease comparing association Rules with decision trees.

4. DATA SOURCE

The objective of this Paper is to discover association Rules to Predict Heart Disease from heart disease Data base. We have taken 14 attributes from medical Data [12].

Table 1

Attributes of Heart Disease Data Sets

no	Attribute Name	Description
1	Age	Age in years
2	Sex	Male=1,Female=0
3	cp	Chest pain type
4	Blood pressure	Resting Blood pressure upon hospital admission
5	Cholesterol	Serum Cholesterol in mg/dl
6	Fasting blood sugar	Fasting blood sugar>120 mg/dl true=1 and false=0
7	Resting ECG	Resting electrocardiographic results
8	Thalach	Maximum Heart Rate
9	Induced Angina	Does the patient experience angina as a result of exercise
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	CA	Number of major vessols colored by fluoroscopy
13	Thal	Normal ,fixed defect, reversible defect
14	Concept class	Angiographic disease status

Table 2

Cleveland Heart Disease Data Sets

T/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	63	1	1	145	233	1	2	150	0	2	3	0	6	0
2	67	1	4	160	286	0	2	108	1	1	2	3	3	2
3	67	1	4	120	229	0	2	129	1	2	2	2	7	1
4	37	1	3	130	250	0	0	187	0	3	3	0	3	0
5	41	0	2	130	204	0	2	172	0	1	1	0	3	0
6	56	1	2	120	236	0	0	178	0	0	1	0	3	0
7	62	0	4	140	268	0	2	160	0	3	3	2	3	3
8	57	0	4	120	354	0	0	163	1	0	1	0	3	0
9	63	1	4	130	254	0	2	147	0	1	2	1	7	2
10	53	1	4	140	203	1	2	155	1	3	3	0	7	1

We have taken 10 patients data where row indicate patient and column indicate responding attribute

The values corresponding to each attribute in table 2 are mapped into binary Transaction table based on the following criteria(possible conditions for heart attack)

Age>45, BP>120, Cholestral range>240, Thal Value>3, thalach value>100 Beats/Minute, Chest Pain type>=3, Resting Ecg>1, Induced Angina=1, Old peak>0, Slope>=2, CA>3 and Concept class>0

Table 3

Binary Transaction Table

T/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	0	1	0	0	1	1	0	1	1	0	1	0
2	1	1	1	1	1	0	1	1	1	1	1	0	1	1
3	1	1	1	1	0	0	1	1	1	1	1	0	1	1
4	0	1	1	1	1	0	0	1	0	1	1	0	1	0
5	0	0	0	1	0	0	1	1	0	1	0	0	1	0
6	1	1	0	1	0	0	0	1	0	1	0	0	1	0
7	1	0	1	1	1	0	1	1	0	1	1	0	1	1
8	1	0	1	0	1	0	0	1	1	0	0	0	1	0
9	1	1	1	1	1	0	1	1	0	1	1	0	1	1
10	1	1	1	1	0	0	1	1	1	1	1	0	1	1

5. METHODOLOGY

In this paper we proposed a new algorithm which Combines the concept of sequence numbers and Clustering. Our proposed method consists of two Sub steps

1) Transform the medical data into binary

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2) Apply proposed method on binary transactional data

Algorithm Description

Algorithm: CBARBSN Cluster Based Association Rule Mining Based on Sequence Number

Input:

- **D**,a transaction database
- **ms**, The Minimum-Support threshold

Output: Frequent item sets

Method:

1) Scan the database **D** and partition the transaction table into equal no. of clusters using skipping Fragments. Cluster size should be well balanced in order to have same distribution of transaction fragments. Apply the method from step 2 to 6 on each cluster**.**

2) Find SN and SID of each item

3) The set of frequent 1 item sets say L1, can then be determined. It consists of candidate 1 item Sets which satisfy minimum support **(SID≥ms)**

4) To discover the set of frequent 2- item sets say L2, Join **L1** ⋈ **L2** and perform logical AND .If Support of 2 -item sets is less than minimum Support **ms** prune 2 -item sets. Construct the Graph by Drawing Edge between each pair of 2 - Frequent item sets.

5) To determine frequent 3 item sets, traverse the graph as if there is a path among the nodes {i,j},{j,k}then the set {i,j,k}will be frequent 3 item sets.

6) The algorithm iterates to find upto n- frequent item sets

7) From each cluster find out the n-frequent item sets. These item sets are said to be local frequent Item sets

8) Intersect the set of frequent item sets from each cluster to get global frequent item sets (Item Sets which appear in both the Clusters are said to be global frequent item sets)[13].

Explanation of Algorithm

The Transactional Table is partitioned into two clusters based on skipping fragments cluster 1 and cluster 2.we applied our proposed algorithm on cluster 1 and cluster 2.**Applying on cluster 1**.

T/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	0	1	0	0	1	1	0	1	1	0	1	0
3	1	1	1	1	0	0	1	1	1	1	1	0	1	1
5	0	0	0	1	0	0	1	1	0	1	0	0	1	0
7	1	0	1	1	1	0	1	1	0	1	1	0	1	1
9	1	1	1	1	1	0	1	1	0	1	1	0	1	1
SN	27	25	11	31	3	0	31	31	8	31	27	0	31	11

Cluster 1 (Table 4)

Calculate Sequence Number of each Item.

SN(1)=27,SN(2)=25,SN(3)=11,SN(4)=31SN(5)=3SN(6)=0SN(7)=31SN(8)=31SN(9)=8SN(10)=21 SN(11)=27SN(12)=0SN(13)=31SN(14)=11

Find out SID of each item

Then SID(1)=4 SID(2)=3 SID(3)=3 SID(4)=5 SID(5)=2 SID(6)=0 SID(7)=5 SID(8)=5 SID(9)=1 SID(10)=5SID(11)=4SID(12)=0SID(13)=5 SID(14)=3

Let Minimum Support **ms** be 3

Discover Frequent 1- item sets

Item sets whose SID satisfy minimum support(**ms)**

Frequent 1 items are 1,2,3,4,7,8,10,11,13,14

Discover Frequent 2- item sets

To discover the set of frequent 2- item sets say L2, Join **L1** ⋈ **L2** and perform logical AND .If Support of 2 -item sets is less than minimum Support **ms** prune 2 -item sets. Construct the Graph by Drawing Edge between each pair of 2 - Frequent item sets

SID(1,2)=3SID(1,3)=2SID(1,4)=4SID(1,7)=4 SID(1,8)=4SID(1,10)=4SID(1,11)=4SID(1,13)=4 SID(1,14)=3.

Frequent 2- item sets are(1,2),(1,4) ,(1,7),(1,8),(1,10)(1,11),(1,13),(1,14)

SID(2,3)=2SID(2,4)=3SID(2,7)=3SID(2,8)=3,SID(2,10)=3SID(2,11)=3SID(2,13)=4SID(2,14)=2.

Frequent 2 item sets are (2,4),(2,7),(2,8),(2,10),(2,11),(2,13),(3,4)(3,7),(3,8),(3,10),(3,11),(3,13),(3,14)

Frequent 2 item sets are (4,7),(4,8),(4,10),(4,11)(4,13),(7,8),(7,10)(7,11) (7,13)

SID(8,10)=5 SID(8,11)=4 SID(8,13)=5 SID(10,11)=4, SID(10,13)=5,SID(11,13)=4

Frequent 2 item sets are (8,10),(8,11),(8,13),(10,11),SID(10,13),(11,13)

Discover Frequent 3- item sets

To compute 3 item sets follow step 5 of algorithm. Traverse the Graph if there is a path among the nodes {i,j},{j,k}then the set {i,j,k}will be frequent 3 item sets. for example (2,4),(2,3)and (4,7)are 2 frequent item sets

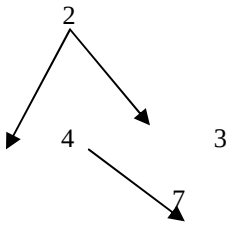

Figure 1

Traverse the graph of figure 1.There is a

Journal of Theoretical and Applied Information Technology
31st October 2011. Vol. 32 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

direct path among the three nodes 2-4-7.Then the set (2, 4, 7) is a 3 frequent item set.
As the data base contains no. of transactions and different attributes, constructing graph for all items here is not practical. Algorithm is iterated and we will get frequent item sets as
(1,4,7,8,10) (1,4,8,10,13),(1,4,7,8,11),
(1, 4, 8, 10, 11)

Applying the same procedure on cluster 2

Cluster 2 Table 5

T/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	1	1	1	1	1	0	1	1	1	1	1	0	1	1
4	0	1	1	1	1	0	0	1	0	1	1	0	1	0
6	1	1	0	1	0	0	0	1	0	1	0	0	1	0
8	1	0	1	0	1	0	0	1	1	0	0	0	1	0
10	1	1	1	1	0	0	1	1	1	1	1	0	1	1
SN	23	29	27	29	26	0	17	31	19	29	25	0	31	17

Calculate Sequence Number of each Item
SN(1)=23,SN(2)=29,SN(3)=27,SN(4)=29SN(5)=26
SN(6)=0SN(7)=17SN(8)=5SN(9)=3SN(10)=4
SN(11)=3SN(12)=0SN(13)=5SN(14)=2

Find out SID of each item
Then SID(1)=4 SID(2)=3 SID(3)=4 SID(4)=4
SID(5)=3 SID(6)=0 SID(7)=2 SID(8)=5 SID(9)=3
SID(10)=4SID(11)=3SID(12)=0SID(13)=5
SID(14)=2
Let Minimum support =3

Discover Frequent 1- item sets
Item sets whose SID satisfy minimum support(**ms)**
Frequent 1 item sets are 1,2,3,4,5,8,10,11,13

Discover Frequent 2- item sets
To discover the set of frequent 2- item sets say L2,
Join **L1** ⋈ **L2** and perform logical AND .If Support of 2 -item sets is less than minimum Support **ms** prune 2 -item sets. Construct the Graph by Drawing Edge between each pair of 2 -Frequent item sets
SID(1,2)=3SID(1,3)=3 SID(1,4)=3SID(1,5)=2
SID(1,8)=3SID(1,9)=3SID(1,10)=3,SID(1,11)=2
SID(1,13)=4,SID(2,3)=3,SID(2,4)=4,SID(2,5)=2
SID(2,8)=4,SID(2,9)=2,SID(2,10)=4,SID(2,11)=3
SID(2,13)=4 Similarly the algorithm will be processed.
frequent 2-item sets are (1,2), (1,3), (1,4), (1,8), (1,9), (1,10), (1,13), (2,2), (2,4), (2,8), (2,10), (2,11), (2,13), (3,4), (3,5), (3,8), (3,10), (3,11), (3,13), (4,8), (4,10), (4,11), (4,13), (5,8), (5,13), (8,9), (8,10), (8,11), (8,13), (9,13), (10,11), (10,13), (11,13)

Discover Frequent 3- item sets
To compute 3 item sets follow step 5 of algorithm. Traverse the Graph if there is a path among the nodes {i,j},{j,k}then the set {i,j,k}will be frequent 3 item sets. From the above list (1,3), (3,10) are 2 frequent item sets..

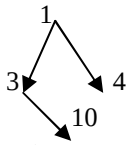

Figure 2

Traverse the graph of figure 2.There is a direct path Among the three nodes 1-3-10.Then the set (1, 3, 10) is a 3- frequent item set. Algorithm will be Iterated and finally we will get the following Frequent item sets are (1,2,4,8,10), (1,2,4,8,13) (1,2,4,10,13),(1,4,8,10,13) By observing each cluster we will find (1,4,8,10,13) is common frequent item set appear in both the clusters ,which we call global frequent item set. The Combination of different attributes found in frequent item sets are the attributes listed in table 1. The attributes which are frequent in medical data are Age, BP, Max Heart Rate, Old peak, Thal. Hence the rule relating to heart disease arteries is

Age>45and Blood pressure>120 and Max Heart rate>100 and old Peak>0 and Thal>3 =>Heart attack

6. EXPERIMENTAL RESULTS

To assess the efficiency of the proposed method, we implemented the algorithm in C language on Pentium 4 processor. The fig shows the support vs. execution time in ms.The experimental results show that our algorithm performs better than the other algorithms.

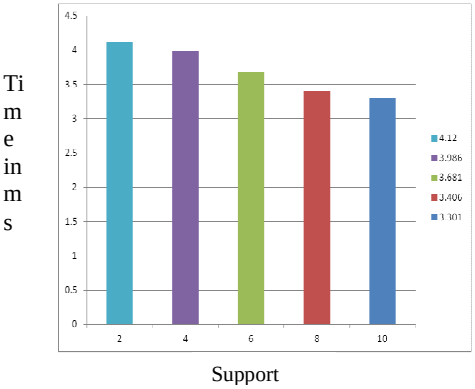

Figure 3 ARNBSN algorithm of reference 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

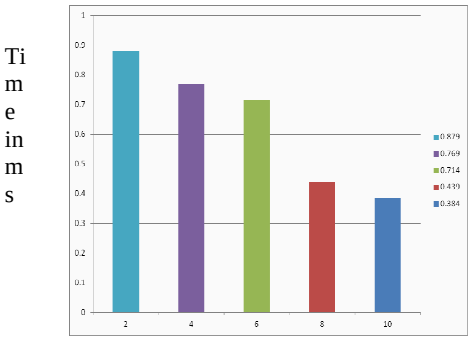

Figure 4 Our Proposed Algorithm CBARBSN

In figure 3 the execution time to mine association Rules is high and as the support increases execution Time slightly changes but in our proposed algorithm Execution time to mine association rules is less and as support increases execution time changes Drastically.

7. CONCLUSION

In this paper, we proposed new algorithm to mine association rules from medical data based on digit sequence and clustering .The entire data base is divided into partitions of equal size, each partition will be called cluster. Each cluster is considered one at a time by loading the first cluster into memory and calculating frequent item sets. Then the second cluster is considered similarly and calculating frequent item sets .This approach reduces main memory requirement since it considers only a small cluster at a time and it is scalable and efficient.

REFERENCES

[1]Berry, Michael J.A and Gordon linoff, Data Mining techniques: for marketing, sales, and customer support.wiley, 1997

[2]Fayyad, usamam, Gregory, piatetsky Shapiro etc all, advances in knowledge discovery and data mining, AAAI Press/MIT Press1996

[3] R.Agrawal, T.Imielinski, A.Swami1. "Mining association rules between sets of items in large databases". ACM SIGMOD Int'l Conf. Management of Data, Washington, D. C., 1993

[4] Agrawal, R Srikant. "Fast algorithms for mining association rules". In: Proc. Of the 20th Int'l Conf. Very Large Data Bases (VLDB'94).1994.487-499

[5] Rajeev Gupta, Recent trends in coronary heart disease epidemiology in India, Indian heart journal,2010

[6] M.Anbarasi, Enhanced prediction of heart disease with feature subset selection, IJEST 2010

[7] Gang Fang, Zu-Kuan Wei, Yu-Lu Liu," An algorithm of improved association rules mining", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

[8] Pan J, Pei J,Yin Y:Mining frequent patterns without candidate generation ACM-SIGMOD,2000

[9] Wael A. Al Zoubi, Azuraliza Abu Bakar, Khairuddin Omar," Scalable and Efficient Method for Mining Association Rules", 2009 International Conference on Electrical Engineering and Informatics

[10] Sellappan palaniappan etc all, Intelligent Heart Disease Prediction System Using Data Mining IJCSNS Vol 8,No 8,August 2008

[11] Carlos Ordonez, Comparing association rules and decision trees for Disease Prediction, ACM 2006

[12] UCI Machine learning Repository from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[13]SON N.Nguyen,Maria Orlowska,"A Further study in the data partitioning approach for frequent item sets mining",17th Australian Database Conference 2006.