Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

KNOWLEDGE DISCOVERY FROM MINING ASSOCIATION RULES FOR HEART DISEASE PREDICTION

[1]M.A.JABBAR, [2] DR PRITI CHANDRA ,[3] DR.B.L DEEKSHATULU

[1] Research Scholar ,JNTU Hyderabad

[2] Senior Scientist, Advanced Systems Laboratory,Hyderabad

[3] Distinguished Fellow, IDRBT (RBI Govt. of India)

E-mail: [1]jabbar.meerja@gmail.com, [2]priti_murali@yahoo.com, [3] deekshatulu@hotmail.com

ABSTRACT

Heart disease is the single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. Mortality data from the registrar general of India shows that coronary heart disease (CHD) are a major cause of death in india.studies to determine the precise cause of death in rural areas of Andhra Pradesh have revealed that CHD cause about 30% death are in rural areas .Although significant progress has been made in the diagnosis and treatment of heart disease further investigation is still needed. Data mining techniques have been used in medical diagnosis for many years and have been known to be effective. In this paper we proposed a new method to discover association rules in medical data to predict heart disease for Andhra Pradesh. This approach is expected to help physicians to make accurate decisions

Keywords: *Andhra Pradesh, Association Rule Mining, Boolean Matrix, Cardiovascular Disease*

1. INTRODUCTION

Hospitals and clinics accumulate a huge amount of patient data over the years. These data provide a basis for analysis of risk factors for many diseases.we can predict the level of heart attack to find patterns associated with heart disease.one of the major topics in data mining research is the discovery of interesting patterns in data[1].

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. Data mining also known as knowledge discovery in data bases (KDD) is the process of automatically discovering useful information in large data repositories [2].

Association rule mining, one of the most and well researched techniques of data mining was first introduced by agrawal etc all[3].it aims to extract interesting correlations, frequent patterns, associations among sets of items in transactional data bases or other data repositories.

Health care awareness and technology development have led to huge number of hospitals and health care centres.but still quality of health care services at affordable cost is still a challenging issue in developing countries.

World health organization in the year 2003 reported that 29.2%of total global deaths are due to CVD.by the end of this year,CVD is expected to be a leading cause of death 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.

In this paper we applied association rule mining on medical data to extract patterns for heart attack prediction.

The remaining sections of the paper are organized as follows. Section 2 describes association rule mining. Section 3 defines Boolean matrix of the transactional data base. Introduction about heart disease and its effects are given in section 4.data sets are explained in section 5.proposed method is presented in section 6.Results are in section 7.Evaluation in section 8, conclusion and future work are described in section 9.

2. ASSOCIATION RULE MINING

Association rule mining were primarily proposed for market based analysis to understand consumer purchasing patterns in retailing industry[3].since association rules are easy to understand and in

Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

widespread use ,they have been successful in business applications ,including finance, telecommunications and health care industry[4].Association rule is an unsupervised learning approach to discover the relationship or correlation among the items. An Association Rule is an implication expression of the form X=>Y, Where X and Y are disjoint sets.i.e X∩Y=Ø.The strength of an association rule can be measured in

terms of support and confidence. Both of them are in the form of percentage. Support determines how often a rule is applicable to a given data set ,while confidence determines how frequently items in Y appears in transactions that contain Y. SUPPORT(X=>Y)=P(X∩Y) COFIDENCE(X=>Y)=P(X∩Y)/P(X)

Minimum support and minimum confidence are the two thresholds in association rule mining .these thresholds can be set by researchers or domain expert. The function of threshold in mining tool is to help us to reduce minor rules efficiency. Association rule mining is decomposed into 2 major subtasks.

1) **Frequent item set generation**: to find all item sets that satisfy minimum support threshold. These item sets are called frequent item sets

2) **Rule generation**: This subtask is to extract high confidence rules from the frequent item sets. Rules which satisfy minimum support and minimum confidence are said to be strong rules [5].

Quantitative Association Rules:

Association rules that contain continuous attributes are commonly known as Quantitative Association Rules. Mining the continuous attributes may reveal useful insights about data

3. BOOLEAN MATRIX AND TRANSACTIONAL DATA BASE

3.1 TRANSACTIONAL DATA BASE

Let I={i1,i2,-----I n}be a set of items. Assume D is a set of transactions over I, called transaction data base, in which a transaction is a set of items, that is subset of I.A transaction has an associated unique identifier called transaction identifier (TID)

3.2 Boolean Matrix

The Boolean Matrix is a matrix with element values of "1" or "0".

1	0	0	1
0	0	0	1
1	0	0	1
1	1	1	0

3.3 Transforming Transaction Data Base into the Boolean matrix

For any transaction data base D,suppose

F:D=>R

R=F(D)=(r ij)$_{mXn}$`

And rij=1 Ii ∈ Tj

0 Ii∉ Tj

Where J=1,2,--n

Data base D can be mapped on to the Boolean matrix R with the effect of F by scanning the data base only once.

Transactional data **Boolean matrix**

TID	ITEMS
1	1,2,3
2	4,5
3	3,4
4	1,4
5	1,3,5

=>

1	1	1	0	0
0	0	0	1	1
0	0	1	1	0
1	0	0	1	0
1	0	1	0	1

3.4 Pruning the Boolean Matrix

Pruning the Boolean matrix means deleting rows and columns from it.

4. CARDIO VASCULAR DISEASES

Cardio Vascular Diseases are the number one cause of death globally. More people die annually from CVD than from any other cause.

An estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths of these deaths, an estimated 7.3 million were due to coronary heart disease. By 2030 almost 23.6 million people will die from CVD's mainly from heart disease and stroke [7].

Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up fatty deposit on the inner walls of the blood vessels that supply the heart or brain. Strokes can also be caused by bleeding from a blood vessel in the brain or from blood clots.

4.1 Risk factors for Coronary Heart Disease

Over 300 risk factors have been associated with coronary heart disease and stroke. The major risk factors are classified into four categories.

Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

A) Major modifiable risk factors:

1) High blood pressure 2) Tobacco use 3) physical inactivity 4) obesity 5) Unhealthy diets 6) Diabetes mellitus

B) Other modifiable risk factors:

1) Low socio economic status 2) mental ill health

3) Psychosocial stress 4) Alcohol use 5) use of certain medication

C) Non modifiable risk factors:

1) Advancing Age 2) Gender 3) Heredity 4) Family history5) Ethnicity

D) **Novel risk factors**

1) Inflammation 2) Excess homo cysteine blood

3) Abnormal blood coagulation [8].

Comprehensive and integrated action is the means to prevent and control Cardio Vascular Diseases.

5. DATA SETS

The features are collected for heart disease prediction in Andhra Pradesh based on the data collected from various corporate hospitals and opinion from expert doctors(table 1).In medical data each row represents patient id and column present heart attack attribute. Medical data is discretized into following attributes shown in table2

6. PROPOSED METHOD

Algorithm Description

Algorithm: HAPBM (Heart Attack Prediction using Matrix)

INPUT:

Transactional Data base D

Minimum Support S

OUTPUT:

Frequent patterns to predict heart attack

METHOD:

1) Discretize the medical data
2) Transform the discretized medical data in Boolean matrix.
3) Calculate all one item set support Ci and all one transaction item support Ri.the one item support ci can be gotten by counting the column with value ‘1’ and one transaction item sets Ri can be gotten by counting the rows with ‘1’
4) Generate frequent 1 item sets .If one item support Ci is less than minimum support threshold S prune corresponding items.
5) Recompute Ri to generate frequent 2 item sets .if transaction item support Ri is less than 2 prune corresponding Transaction.
6) Generate k frequent item sets .Repeat the steps 4 and 5 till no frequent item sets are generated.
7) Generate the association rules from the frequent item sets.

Sample C Functions for Proposed Method.

1. **Checking the columns which do not satisfy the minimum support**

```
for(i=1;i<n;i++)//col
{
        if(a[n][i]<=k)
        {
                continue;
        }
        col1++;
        count=count+1;
        //printf("%d\t%d\t",count,a[n][i]);
        for(j=0;j<n;j++)//row
        {
            a1[j][col1]=a[j][i];
        }
}
```

2. Finding the new row sum

```
for(i=1;i<n;i++)
{
        sum=0;
        for(j=1;j<col1;j++)
        {
                sum=sum+a1[i][j];
```

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

```
        }
a1[i][j]=sum;
        //printf("s%d,i %d,j %d",a1[i][j],i,j);
    }
```

3. **Checking the rows which do not satisfy the minimum support**

```
    for(i=1;i<n;i++)//row
    {
        if(a1[i][col1]<rtre)
        {
            continue;
        }
        row1++;
        for(j=0;j<col1+1;j++)
        {
            a2[row1][j]=a1[i][j];
        }
    }
```

6.1 Explanation of Proposed Method

This section describes execution of our proposed algorithm.discretized data is given in table 4..let minimum support is 6.

1) Transform the medical data into Boolean matrix.

2) Compute the Ci values of each column in the transactional data base. Prune the columns whose Ci<S.

3) After pruning columns the medical data will be shown in table

1 frequent item sets are {2,4,7,8,11,12,13,15,16,17,18,20}

<u>2 frequent item set generation</u>

check Ri for all the rows ,if ri<2,prune corresponding rows. Make the combinations of 1 itemsets,to get 2 itemsets and check support of 2 itemsets.if support of 2 item set<S prune the combination.

Frequent 2 item sets are

(2,8),(2,13),(2,16),(2,18),(4,8),(4,12),(4,13),(4,16),(4,17),(4,18)(7,8),(7,11),(7,12),(7,16),(7,17),(7,18) (8, 11), (8, 12), (8,15),(8,16),(8.17),(8,18),(8,20),(11,13),(11,16),(11,18),(11,20),(12,13),(12,16),(12,17),(12,18),(13,15),(13,16), (13,17),(13,18),(13,20),(15,16),(15,18),(16,17),(16,18), (17,18),(18,20)

<u>3- frequent item set generation</u>

Check Ri for all the rows, if ri<3,prune corresponding row. Make the combinations of 2 itemsets,to get 3 item sets and check support of 3 item sets. if support of 3 item set<S prune the combination. The transactional data base will be shown in table 5

<u>Frequent 3 item sets are</u>

(2,8,13),(2,8,16),(2,13,16),(2,13,18),(2,16,18),(2,13,15),(4,8,11),(4,8,16),(4,8,18),(4,13,16),(4,13,17),(4,13,18),(4,16,18),(4,17,18),(7,8,16),(7,8,17),(7,8,18),(7,11,18),(7,11,13),(7,11,20),(7,16,17),(7,16,18),

(7,17,18),(7,18,20),(8,11,12),(8,11,15),(8,11,18),(8,11,20), (8,11,13),(8,12,13),(8,12,16),(8,12,18),(8,15,16),(8,15,18),(8,16,17),(8,16,18),(8,17,18),(8,18,20),

(11,13,16),(11,13,18),(11,16,18),(11,18,20),(12,13,16),(12,13,17),(12,13,18),(12,13,20),(12,16,17),(12,16,18),(12,17,18),(13,15,16),(13,15,18),(13,16,17), (13,16,18),(13,17,18),(13,18,20),(15,16,17), (16,17,18),(16,18,20)(17,18,20**)**

<u>4- frequent item set generation</u>

Check Ri for all the rows, if ri<4,prune corresponding row. Make the combinations of 3 itemsets,to get 4 item sets and check support of 4 item sets. If support of 4 item set<S prune the combination. The transactional data base will be shown in table 5

(2,8,13,16),(2,8,13,18),(4,13,16,18),(4,13,17,18),(7,8,16,17)(7,8,16,18),(7,16,17,18),(8,11,12,18), (8,12,13,16),(8,12,13,18),(8,12,16,18),(8,15,16,18), (8,15,17,18),(11,13,16,18),(11,16,18,20),(12,13,16,17), (12,16,17,18),(13,15,16,17),(13,15,16,18),(13,15,16,20),(13,16,17,18),(13,17,18,20)

5 Frequent item set generation

Check Ri for all the rows, if ri<5,prune corresponding row. Make the combinations of 4 itemsets,to get 5 item sets and check support of 5 item sets. if support of 5 item set<S prune the combination.

5 frequent item sets are

1) (2, 8, 13, 16, 18),

2) (4, 13, 16, 17, 18)

3) (7, 8, 16, 17, 18)

4) (8, 12, 13, 16, 18)

Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

6- **frequent item set generation**

Check Ri for all the rows, if ri<6, prune corresponding row. Make the combinations of 5 itemsets, to get 6- item sets and check support of 6 item sets. If support of 6 item set<S prune the combination. Here no combination will have support>minimum Support threshold. Hence algorithm will be terminated.

Above 5 frequent item sets implies rules like

1) **Age between 41-65 and Hyper Cholestremia=yes and smoking=yes and family history=No and alcohol=yes =>Heart disease**
2) **Person=male and smoking=yes and family history=No and Rural=yes alcohol=yes =>Heart disease**
3) **Hypertension=No and Hyper Cholestremia=yes and family history=No and Rural=yes alcohol=yes =>Heart disease**
4) **Hyper Cholestremia=yes and resting ECG>2 and smoking=yes and family history=No and alcohol=yes =>Heart disease**.

7. RESULTS AND DISCUSSION

The following section describes different categories of association analysis for heart disease prediction. A total of 70 patients records are collected from the cardio thoracic departments of various corporate hospitals in Andhra Pradesh. Based on the data collected, we analyzed medical data using our proposed approach.

Table 6 lists the top 10 association rules having highest support values

The patterns of CHD in Andhra Pradesh has been reported as follows.

- Out of 70 samples 52 rural population have associated with heart disease.
- Males are affected more than females
- 75% of population who had Hypertension are associate with heart disease.
- Population with age group >45 are correlated with heart disease.
- Smoking has been identified as a major risk factor.
- More than 50%of males who lives in rural are correlated with heart disease.
- Males with age group>45 and who have smoking habit are correlated with CHD.
- 20%of Diabetic patients are associated with heart disease.

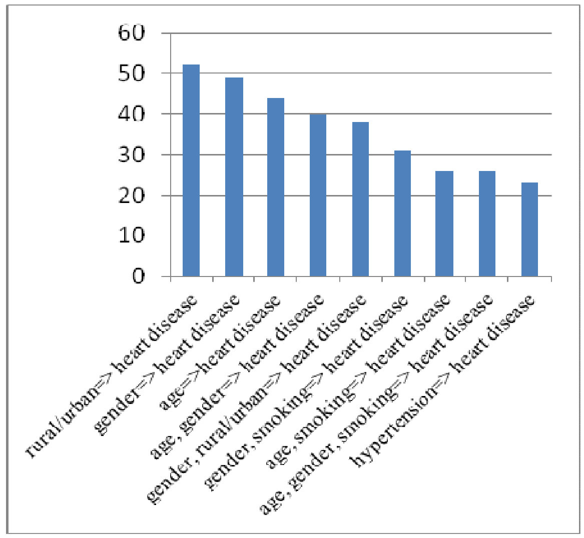

Figure 1 Association rule representing heart disease

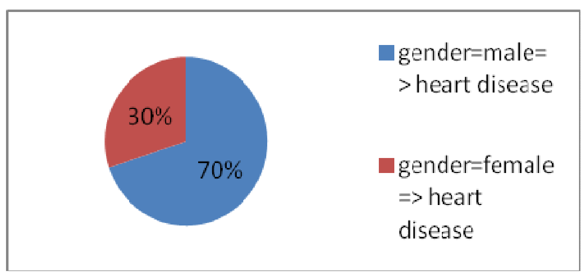

Figure 2 Representing correlation between gender and heart disease

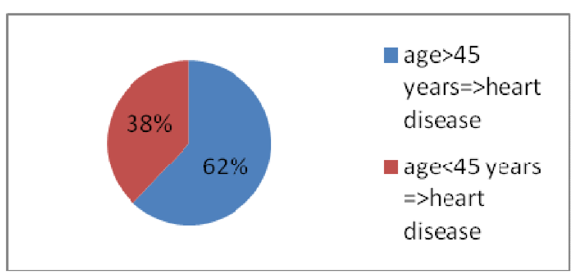

Figure 3 Representing relation between age >45 and heart disease

Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

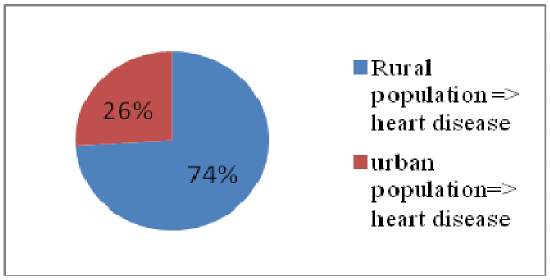

Figure 4 Correlation between rural population and heart disease

8. PERFORMANCE EVALUATION

To assess the performance of our proposed algorithm for discovering frequent item sets we have taken some comparisons tests between our algorithm and Apriori.Experimental environment: CPU, Pentium-IV Dual core, RAM 2GB, Operating system: windows XP,Programing language C.To compare with apriori we have taken the Data sets from FIMI(Frequent item set mining implementation)[9].

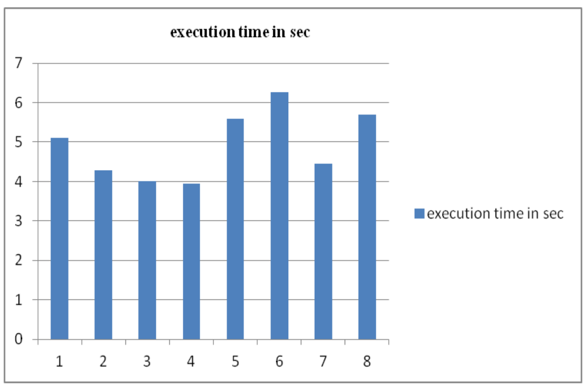

Figure 5 Support vs execution time for our proposed algorithm

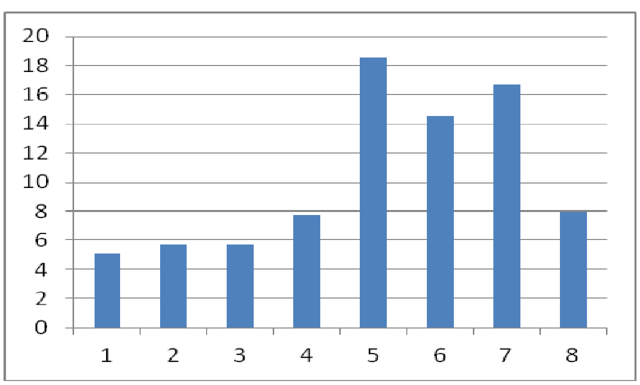

Fig 6 Support vs. execution time for Apriori algorithm

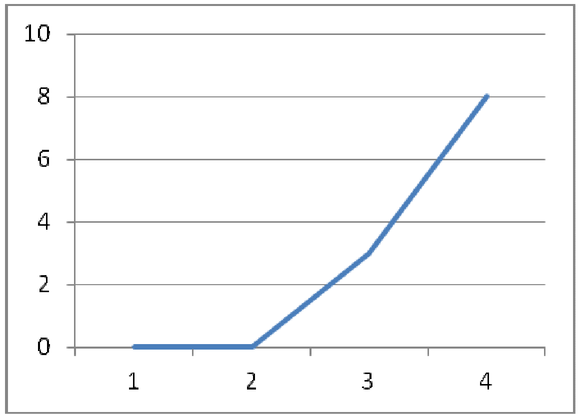

Fig 7 item sets vs transactions reduced

X-Axis Represents Item set no

Y-Axis Represents no of Transactions Reduced

Experimental results shows that as we goes on generating item sets, transaction reduction also increases thus reduces search space and saves time to generate item sets.

9. CONCLUSION

The objective of our work is to predict more accurately the presence of heart disease for Andhra Pradesh population. In our work we used matrix based approach to reduce no. of scans of data base. Our approach is simple and efficient for extracting significant patterns from the heart disease data for the efficient prediction of heart Disease in Andhra Pradesh. Our algorithm takes less time to generate patterns and reduces transactions at each stage thus reduces search space. The need to contain the epidemic of CHD as well as its combat its impact and minimize its toll on population of Andhra Pradesh is obvious and urgent. Strategies to meet the objective of prevention and control of CHD must be developed and efficiently implemented. In our future work we will try to incorporate to generate patterns using Advanced data mining techniques for heart disease prediction.

REFERENCES

[1]Hint wint Khaing,"Data Mining based Fragmentation and Prediction of Medical Data",IEEE 2011.

[2] Pang ning Tan,Steinbach,Vipin Kumar,"Introduction to data mining" Pearson education 2006

[3] R.Agrawal, T.Imielinski, A.Swami1. "Mining association rules between Sets of items in large databases". ACM SIGMOD Int'l Conf.

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

on Management of Data, Washington, D. C., 1993

[4] Bose, I.Mahapatra R.K "Business Data Mining a Machine Learning Perspective information and management 2001,pp 211-225

[5] Han J, Kamber M.Data Mining concepts and techniques, Morgan and Kaufmann 2000

[6] Zhang Zhanglin,Liu Jun etc all"A Fasr algorithm for mining association rules based on Boolean matrix",IEEE 2008

[7] WHO Report on Non Communicable Diseases, September 2011

[8] WHO Report on Cardio Vascular Disease and Risk Factors Chapter 3 Page 24-25,2011

[9] Data Sets from FIMI Repository http://fimi.ua.ac.be/data/

[10] Arun Pujari, "data mining techniques" by University press 2008

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

Table 1 Attributes of Heart Disease Data Sets

No	Attribute Name	Description
1	Age	Age in years
2	Sex	Male=1,Female=0
3	hypertension	is a condition in which the blood pressure in the arteries is chronically elevated
4	Blood pressure	Resting Blood pressure upon hospital admission
5	Hyper Cholestremia	High blood cholestrol
6	Diabetes	Diabetes is a lifelong (chronic) disease in which there are high levels of sugar in the blood.
7	Resting ECG	Resting Electrographic Results
8	Smoking	**CAD** is associated with smoking
9	Alcohol consumption	**CAD** is associated with alcohol
10	Family history of CAD	A **family history** of early **CAD** is one of the predictors of **CAD**
11	Rural/Urban	Lives in urban /Rural
12	Concept class	Concept class yes or no

Table 3 Medical data is discretized into following attributes

Attribute No	Name
1)	Age<40
2)	Age 41-65
3)	Age>65
4)	Male
5)	Female
6)	Hypertension=yes
7)	Hypertension=no
8)	Hyper Cholestomia=yes
9)	Hyper Cholestromia=no
10)	Diabetes=yes
11)	Diabetes =no
12)	ECG =yes
13)	Smoking=YES
14)	Smoking=NO
15)	Family history of CAD=Yes
16)	Family history of CAD=NO
17)	Rural =yes
18)	Alcohol=YES
19)	Alcohol=No
20)	Rural =no
21)	Concept class=yes
22)	Concept class=no

Table 2 Medical Data Transformed to Binary Data

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

Journal of Theoretical and Applied Information Technology
31st July 2012. Vol. 41 No.2

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

Table 4 Discretization of Medical Data

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	**Ri**
T1	0	1	0	1	0	1	0	1	0	1	0	0	1	0	1	1	1	1	0	0	1	1	**13**
2	0	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	0	0	**13**
3	0	0	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	**12**
4	1	0	0	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0	1	**13**
5	0	1	0	0	1	0	0	1	0	0	1	1	1	0	1	1	0	1	0	1	0	1	**12**
6	0	1	0	1	0	1	0	1	0	0	1	1	1	0	1	1	0	1	0	1	0	1	**11**
7	0	1	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0	0	**13**
8	0	1	0	0	1	0	1	0	1	0	1	1	1	1	0	0	0	1	0	1	0	1	**10**
9	0	1	0	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	0	**13**
10	0	1	0	1	0	0	1	1	0	1	0	0	1	1	0	1	1	1	0	0	1	0	**12**
Ci	**1**	**7**	**2**	**7**	**3**	**3**	**7**	**9**	**5**	**2**	**8**	**7**	**10**	**4**	**6**	**9**	**7**	**10**	**0**	**6**	**4**	**5**	

Table 5 Transactional data base after pruning rows and columns.

T	2	4	7	8	9	11	12	13	15	16	17	18	20	Ri
1	1	1	0	1	0	0	0	1	1	1	1	1	0	**8**
2	0	1	1	1	1	1	1	1	0	1	1	1	1	**11**
3	0	1	1	1	0	1	1	1	0	1	1	1	0	**9**
4	0	1	1	1	1	1	1	1	1	1	1	1	1	**12**
5	1	0	0	1	0	1	1	1	1	1	0	1	1	**9**
6	1	1	0	1	0	1	1	1	1	1	0	1	1	**10**
7	1	0	1	1	1	1	1	1	1	1	1	1	1	**12**
8	1	0	1	0	1	1	1	1	0	0	0	1	1	**8**
9	1	1	1	1	1	1	1	1	1	1	1	1	0	**12**
10	1	1	1	1	0	0	0	1	0	1	1	1	0	**8**
Ci	**7**	**7**	**7**	**9**	**5**	**8**	**7**	**10**	**6**	7	**9**	**10**	**6**	

Table 6 Top 10 association rules

No	**Association rule**	**support**	**confidence**
1	rural/urban=> treatment	52	100
2	gender=> treatment	49	100
3	age=> treatment	44	100
4	age, gender=> treatment	40	100
5	gender, rural/urban=> treatment	38	100
6	gender, smoking=> treatment	31	100
7	BP=> treatment	30	100
8	age, smoking=> treatment	26	100
9	age, gender, smoking=> treatment	26	100
10	hypertension=> treatment	23	100