Application of Data Mining in Health Care

S. L. Nalawade[1], Dr. R.V. Kulkarni[2]

[1]Assistant Professor, Department of Computer Application, K.G.D.B.L.M., Kundal, Maharashtra, India

[2] Professor and Head of Department of Computer Studies, CSIBER Kolhapur, Maharashtra, India

Abstract: ***Data mining is the process of selecting, exploring and modeling a large database in order to discover model and pattern that are unknown. Data mining refers to extracting knowledge from large amount of data. Real life data mining approaches are interesting because they often present a different set of problems for data miners. The huge amounts of data generated by healthcare transactions are too complex and in large volume to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. The purpose of this study is to review the relevant data mining tool and its applications in healthcare units. This paper focuses on various models and techniques used in data mining for health care and its applications for better health policy-making and in decision making. It provides recommendation for future research in the application of data mining in health care units.***

Keywords: Data mining tools, Blood bank, Health Care Information, Weka, R tool.

1. Introduction

In healthcare data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. The huge amounts of data generated by healthcare transactions are too complex and in large volume to be processed and analyzed by traditional methods. The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. Massive healthcare data needs to be converted into information and knowledge, which can help control cost and maintain high quality of patient care. Without data mining it is difficult to realize the full potential of data collected within healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable.

In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. The outcomes of Data Mining technologies are to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides them effective treatments.

Data mining has been used by many organizations and in healthcare; data mining is becoming increasingly popular. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, evaluation of treatment effectiveness, management of healthcare, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services, for quality control and maintenance scheduling. It is useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management. Recent technologies are used in medical field to enhance the medical services in cost effective manner. In this paper the different relevant data mining tools used for Healthcare are reviewed and proposes a data model for monitoring individual's information for population based health care management.

2. Data Mining - Knowledge Discovery

The main difference of data mining from knowledge discovery is that the data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process −

- **Data cleaning** −In this step, the noise and inconsistent data is removed.
- **Data Integration** − In this step, multiple data sources are combined.
- **Data Selection** − In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** − In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** − In this step, data patterns are evaluated.
- **Knowledge Presentation** − In this step, knowledge is represented.

3. The Data Mining Models

Following figure shows the data mining models and tasks:

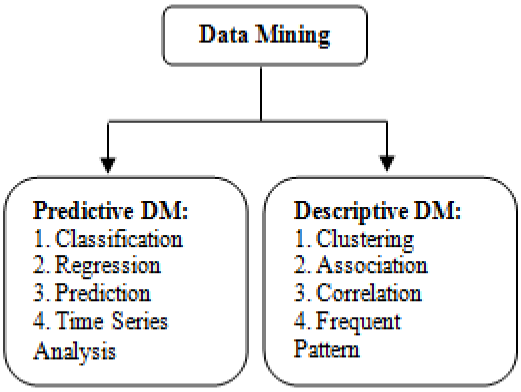


Figure 1: Data Mining Models and Tasks

The predictive model makes prediction about unknown data values by using the known values. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Next section discusses about the work done so far in this field.

4. Literature Survey

The work [1] focuses on data mining and trends associated with it. In this paper, main purpose of the system is to increase blood donor's rate as well as to attract more blood donors to donate blood. [2] The work has been made to classify and predict the number of blood donor's according to their age and blood group. In this work, the WEKA data mining tool and J48 algorithm is used to classify the data and evaluation of the data. In this work [3], the main aim of this paper is to provide a detailed introduction of weka clustering algorithms. It provides the past project data for analysis. With the help of figures researchers are showing the working of various algorithms used in weka.[4], It provides the work and the main purpose of the system is to guide diabetic patients during the disease. In this work, the WEKA data mining tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared.

In this research work [5], five medical databases are used and experimental results are computed using data mining software tool. In this research work [6]. The researcher presents the process of designing a model that can help in blood platelet transfusion database maintained in Maxcare Hospital, which has a great significance in the health care field. Research work [7], mainly focuses on an analysis, which was performed by a team of physicians and computer science researchers, using a commercially available on-line analytical processing (OLAP) tool in conjunction with proprietary data mining techniques. The initial objective of the analysis was to discover how to use data mining techniques to make business decisions that can influence cost, revenue, and operational efficiency while maintaining a high level of care. Another objective was to understand how to apply these techniques appropriately and to find a repeatable method for analyzing data and finding business insights. The process used to identify opportunities and effect changes is described. The work [8], focus on a brief introduction of data mining techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. The purpose of this paper is to provide an insight towards requirements of health domain and about suitable choice of available technique. The study mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management [9].

5. Proposed Methodology

The whole work is completed through the following steps:

- Collection of blood data from Hindratna Prakashbapu Patil Blood Bank in Sangli, Maharashtra, India.
- Applying Weka tool for checking association between the number of blood donors through their age and blood group, and blood group and their disease.
- Application of weka tool for checking significant relationship between patients and tests using pathological data in healthcare.

In this methodology, Researcher suppose that a blood bank management system has the following objectives:

- To study role of data mining in different sections of healthcare (Blood Bank, Pathology) to bring the relevant inferences.
- To study and analyze the risk factors of HIV infections among blood using Data Mining (Weka Tool).
- To study the contribution of pathological treatment in healthcare using Weka tool in data mining techniques.

The estimation and prediction may be viewed as types of association. The problem usually is to evaluate the work through the training data set and then verify the result by using a test set of data. The following table1 shows different association algorithms:

Table 1: Types of Data Mining Algorithms

Type	Algorithm
Association	Apriori
Decision Tree	ID3 J48 C4.5 CART REP Tree NB Tree
Neural Network	Propagation NN Supervised learning
Statistical	Regression Bayesian

In this work researcher used the Weka as a data mining tool. No user needs to install Weka in his workstation, but it do already enough to be installed on server machine. In this work researcher collected 150 dataset from a Blood Bank Centre. The dataset has the following 10 attributes:

Table 2: Attributes of Blood Donors Dataset

Name of Attribute	Description
Date of Bleeding	Date of bleeding
Expiry Date	Expiry date of blood
Age	Age of the donors in years
Sex	M = male, F = female
Height	Height of the donors in cm
Weight	Weight of the donors in kg
BP	Blood Pressure of the donors
Hb	Hemoglobin of the donors in gm
Pulse Rate	Pulse rate of the donors
Blood Group	Blood group of the donors in years

This data set has been implemented in WEKA which is a very popular data mining tool used to check association between variables in the dataset. Algorithm for attribute selection was applied on dataset for pre-processing. After pre-processing of dataset, association is performed. It is a technique which helps to check significant relationship between variables in the dataset.

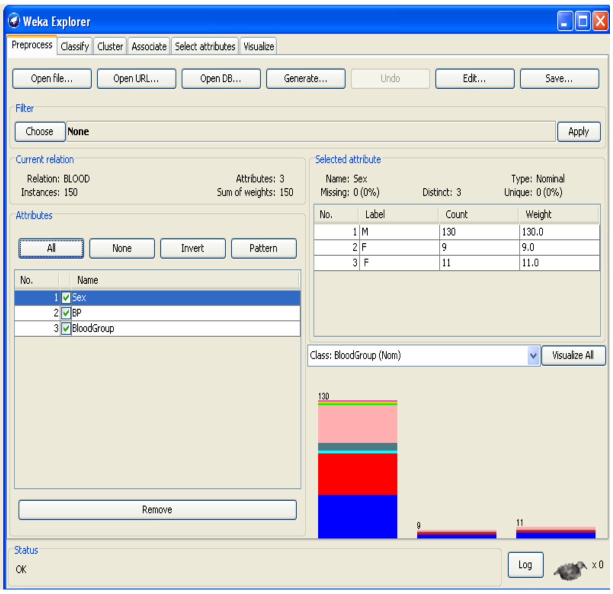
6. Experiments and Results

Researchers aim in this work is to check significant relationship between variables in the blood bank and pathology dataset using WEKA. Researcher has data file containing attribute values for 150 samples for blood bank data in .csv format. The data set has been collected from Blood Bank center. The data file contains all 10 attributes respectively. Researcher had implemented the Apriori algorithm for association relationship. The steps of Apriori algorithm in WEKA tool are as follows:

Step 1: Create a data file in the .csv format with name AllBlood.csv and convert it into .arff format with given name blood.arff and disease.arff.

Step 2: Open the Weka application.

Step 3: Loading data into WEKA and open the blood.arff, disease.arff and urine.arff file as well as viewing that data in same window as shown in following fig.



Filter: None

Relation: BLOOD
Instances: 150
Attributes: 3
Sum of weights: 150

Name: Sex
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

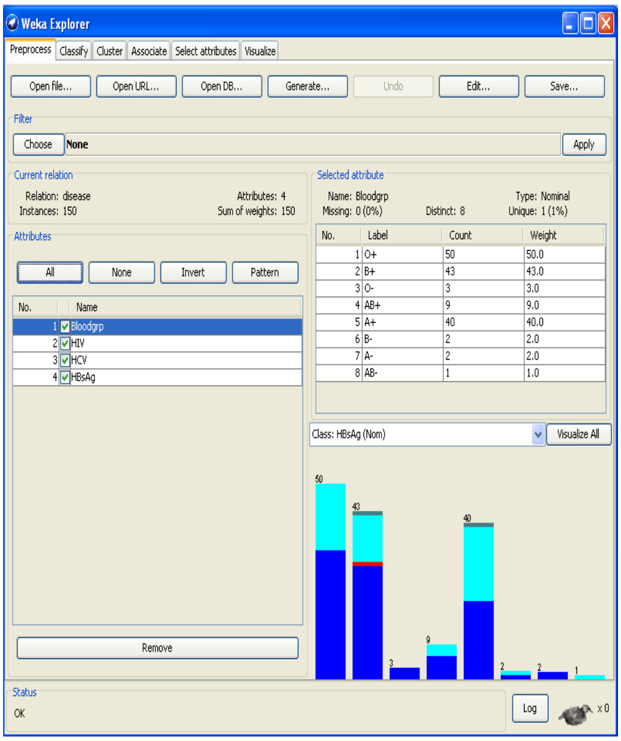
No.	Label	Count	Weight
1	M	130	130.0
2	F	9	9.0
3	F	11	11.0

No.		Name
1	☑	Sex
2	☑	BP
3	☑	BloodGroup

Class: BloodGroup (Nom)

Status: OK

Figure 2a: blood.arff files in Weka explorer



Filter: None

Relation: disease
Instances: 150
Attributes: 4
Sum of weights: 150

Name: Bloodgrp
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 1 (1%)

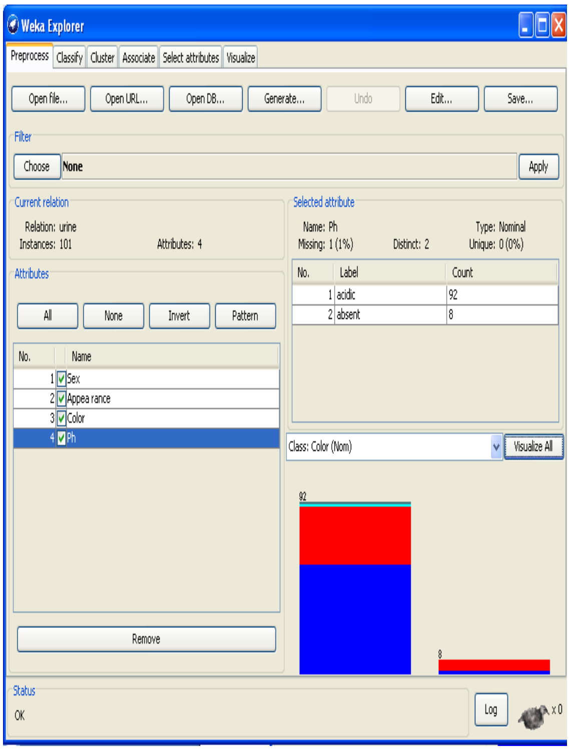
No.	Label	Count	Weight
1	O+	50	50.0
2	B+	43	43.0
3	O-	3	3.0
4	AB+	9	9.0
5	A+	40	40.0
6	B-	2	2.0
7	A-	2	2.0
8	AB-	1	1.0

No.		Name
1	☑	Bloodgrp
2	☑	HIV
3	☑	HCV
4	☑	HBsAg

Class: HBsAg (Nom)

Status: OK

Figure 2b: disease.arff files in Weka explorer



Filter: None

Relation: urine
Instances: 101
Attributes: 4

Name: Ph
Missing: 1 (1%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	acidic	92
2	absent	8

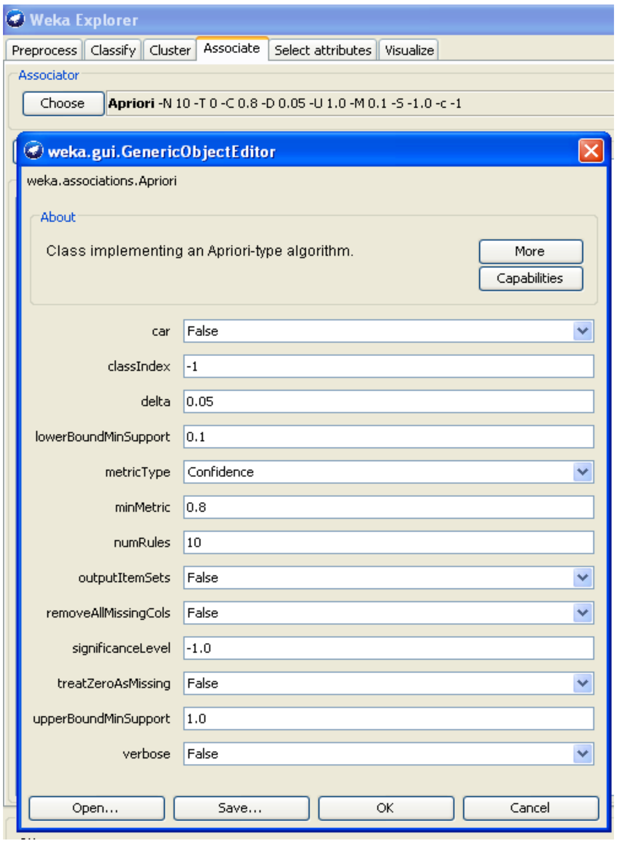
No.		Name
1	☑	Sex
2	☑	Appea rance
3	☑	Color
4	☑	Ph

Class: Color (Nom)

Status: OK

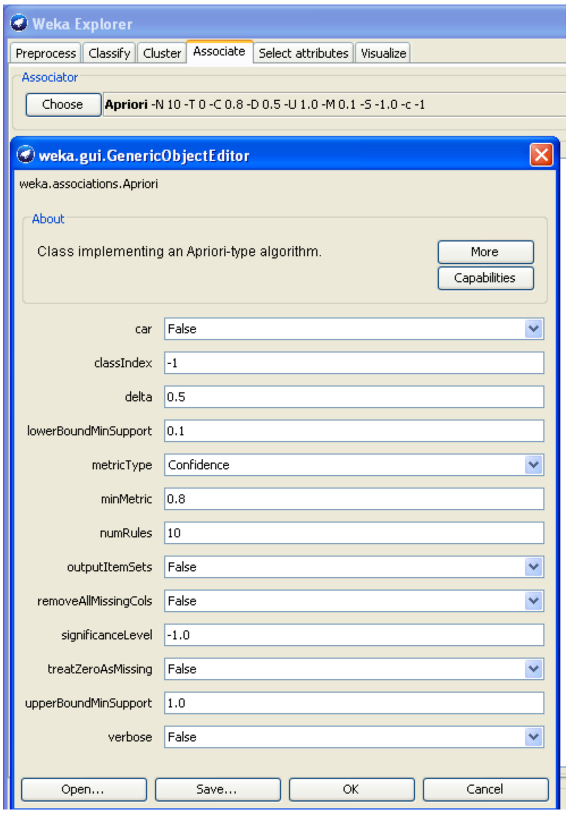
Figure 2c: urine.arff files in Weka explorer

Step 4: Click on Associate tab and we set up the configuration as shown in figure.



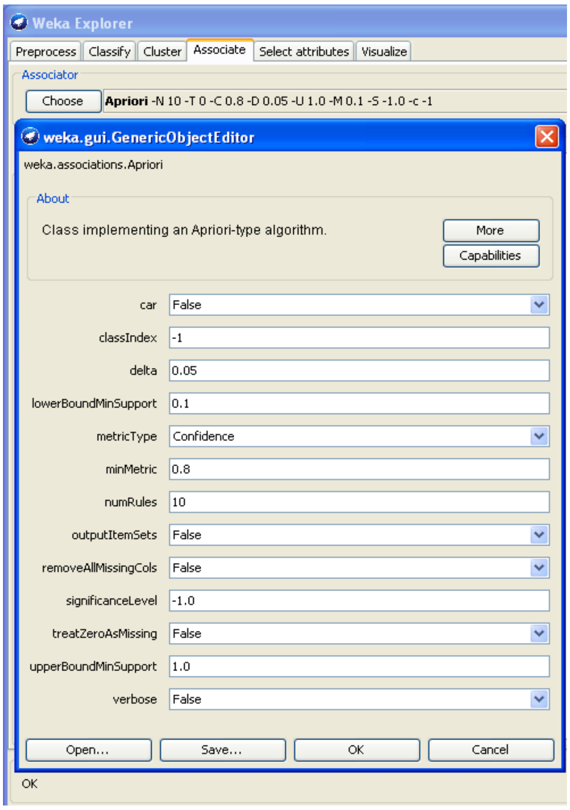
```
Associator
Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
                 car  False
          classIndex  -1
               delta  0.05
lowerBoundMinSupport  0.1
          metricType  Confidence
           minMetric  0.8
            numRules  10
      outputItemSets  False
removeAllMissingCols  False
   significanceLevel  -1.0
  treatZeroAsMissing  False
upperBoundMinSupport  1.0
             verbose  False
```

Figure 3a: Associate tab for configuration for blood.arff



```
Associator
Apriori -N 10 -T 0 -C 0.8 -D 0.5 -U 1.0 -M 0.1 -S -1.0 -c -1
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
                 car  False
          classIndex  -1
               delta  0.5
lowerBoundMinSupport  0.1
          metricType  Confidence
           minMetric  0.8
            numRules  10
      outputItemSets  False
removeAllMissingCols  False
   significanceLevel  -1.0
  treatZeroAsMissing  False
upperBoundMinSupport  1.0
             verbose  False
```

Figure 3b: Associate tab for configuration for disease.arff



```
Associator
Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
                 car  False
          classIndex  -1
               delta  0.05
lowerBoundMinSupport  0.1
          metricType  Confidence
           minMetric  0.8
            numRules  10
      outputItemSets  False
removeAllMissingCols  False
   significanceLevel  -1.0
  treatZeroAsMissing  False
upperBoundMinSupport  1.0
             verbose  False
OK
```

Figure 3c: Associate tab for configuration for urine.arff

Figure represents associate tab for confidence and support factors for blood data.

Step 5: Run the Apriori algorithm.

Left click on chosen Apriori algorithm to open Weka editor. Select Support=0.05 and confidence=0.8 values to get the result.
Click on "Start" to start association. After completing the algorithm, we get the following results in separate window:

=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.8 - D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: BLOOD
Instances: 150
Attributes: 3
 Sex
 BP
BloodGroup
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.2 (30 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 16

Generated sets of large itemsets:
Size of set of large itemsetsL(1): 5
Size of set of large itemsetsL(2): 7
Size of set of large itemsetsL(3): 2

Best rules found:
1. BP=120/80 BloodGroup=B+ 36 ==>Sex=M 34 <conf:(0.94)> lift:(1.09) lev:(0.02) [2] conv:(1.6)

```
2. BloodGroup=B+ 43 ==> Sex =M 39
<conf:(0.91)> lift:(1.05) lev:(0.01) [1] conv:(1.15)
 3. BP=120/80 121 ==> Sex=M 109
<conf:(0.90)> lift:(1.04) lev:(0.03) [4] conv:(1.24)
4. BloodGroup=A+ 40 ==> Sex =M 35
<conf:(0.88)> lift:(1.01)lev:(0) [0] conv:(0.89)
 5. BP=120/80 BloodGroup=O+ 40 ==>Sex=M 35
<conf:(0.88)> lift:(1.01)lev:(0) [0] conv:(0.89)
 6. Sex =M BloodGroup=B+ 39 ==> BP=120/80 34
<conf:(0.87)> lift:(1.08) lev:(0.02) [2] conv:(1.26)
 7. Sex=M BloodGroup=O+ 41 ==> BP=120/80 35
<conf:(0.85)> lift:(1.06) lev:(0.01) [1] conv:(1.13)
 8. Sex =M 130 ==> BP=120/80 109
<conf:(0.84)> lift:(1.04)lev:(0.03) [4] conv:(1.14)
 9. BloodGroup=B+ 43 ==> BP=120/80 36
<conf:(0.84)> lift:(1.04) lev:(0.01) [1] conv:(1.04)
10. BloodGroup=O+ 50 ==> Sex =M 41
<conf:(0.82)> lift:(0.95) lev:(-0.02) [-2] conv:(0.67)
```

Figure 4 a: Apriori output of Wekafor blood.arff

```
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.5
-U 1.0 -M 0.1 -S -1.0 -c -1
Relation: disease
Instances: 150
Attributes: 4
Bloodgrp
 HIV
 HCV
HBsAg
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.6 (90 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 1

Generated sets of large itemsets:
Size of set of large itemsetsL(1): 3
Size of set of large itemsetsL(2): 3
Size of set of large itemsetsL(3): 1

Best rules found:
1. HBsAg=NR 94 ==> HIV=NR 94
<conf:(1)> lift:(1.38) lev:(0.17) [25] conv:(25.69)
 2. HCV=NR HBsAg=NR 92 ==>HIV=NR 92 <conf:(1)
lift:(1.38) lev:(0.17) [25] conv:(25.15)
 3. HBsAg=NR 94 ==> HCV=NR 92
<conf:(0.98)>lift:(1.32)lev:(0.15)[22] conv:(8.15)
 4.HIV=NR HBsAg=NR 94 ==>HCV=NR 92
<conf:(0.98)>lift:(1.32)lev:(0.15)[22] conv:(8.15)
 5.HBsAg=NR 94 ==>HIV=NR HCV=NR 92
<conf:(0.98)>lift:(1.38)lev:(0.17)[25] conv:(9.19)
 6. HIV=NR 109 ==> HCV=NR 106
<conf:(0.97)>lift:(1.31)lev:(0.17)[25] conv:(7.09)
 7. HCV=NR 111 ==> HIV=NR 106
<conf:(0.95)>lift:(1.31)lev:(0.17)[25] conv:(5.06)
 8. HIV=NR HCV=NR 106 ==> HBsAg=NR 92
<conf:(0.87)>lift:(1.38)lev:(0.17)[25] conv:(2.64)
 9. HIV=NR 109 ==> HBsAg=NR 94
<conf:(0.86)>lift:(1.38)lev:(0.17)[25] conv:(2.54)
10. HIV=NR 109 ==> HCV=NR HBsAg=NR 92
<conf:(0.84)>lift:(1.38)lev:(0.17)[25] conv:(2.34)
```

Fig.4b: Apriori output of Wekafor disease.arff

```
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.8 -D
0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: urine
Instances: 101
Attributes: 4
 Sex
 Appearance
 Color
 Ph
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (25 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 15
Generated sets of large itemsets:
Size of set of large itemsetsL(1): 7
Size of set of large itemsetsL(2): 12
Size of set of large itemsetsL(3): 5

Best rules found:

1. Sex=M Color=pale yellow 32 ==>Ph=acidic 32
<conf:(1)> lift:(1.1) lev:(0.03) [2] conv:(2.85)
2. Appearance=Hazy 39 ==>Ph=acidic 38 <conf:(0.97)>
lift:(1.07) lev:(0.02) [2] conv:(1.74)
3. Appearance=clear Color=pale yellow 32 ==>Ph=acidic
31 <conf:(0.97)> lift:(1.06) lev:(0.02) [1] conv:(1.43)
4. Color=pale yellow 61 ==>Ph=acidic 59 <conf:(0.97)>
lift:(1.06) lev:(0.03) [3] conv:(1.81)
5. Appearance=Hazy Color=pale yellow 27 ==>Ph=acidic
26 <conf:(0.96)> lift:(1.06) lev:(0.01) [1] conv:(1.2)
6. Sex=F Color=pale yellow 29 ==>Ph=acidic 27
<conf:(0.93)> lift:(1.02) lev:(0.01) [0] conv:(0.86)
7. Sex=M 51 ==> Ph=acidic 47
<conf:(0.92)> lift:(1.01)lev:(0.01) [0] conv:(0.91)
8. Sex=F 49 ==> Ph=acidic 45 <conf:(0.92)>
lift:(1.01)lev:(0) [0] conv:(0.87)
9. Sex=M Appearance=clear 32 ==> Ph=acidic 29
<conf:(0.91)> lift:(0.99) lev:(0) [0] conv:(0.71)
10. Appearance=clear 52 ==> Ph=acidic 47 <conf:(0.9)>
lift:(0.99) lev:(0) [0] conv:(0.77)
```

Figure 4c: Apriori output of Weka for urine.arff

Step 6: Analysis of the Apriori algorithm.

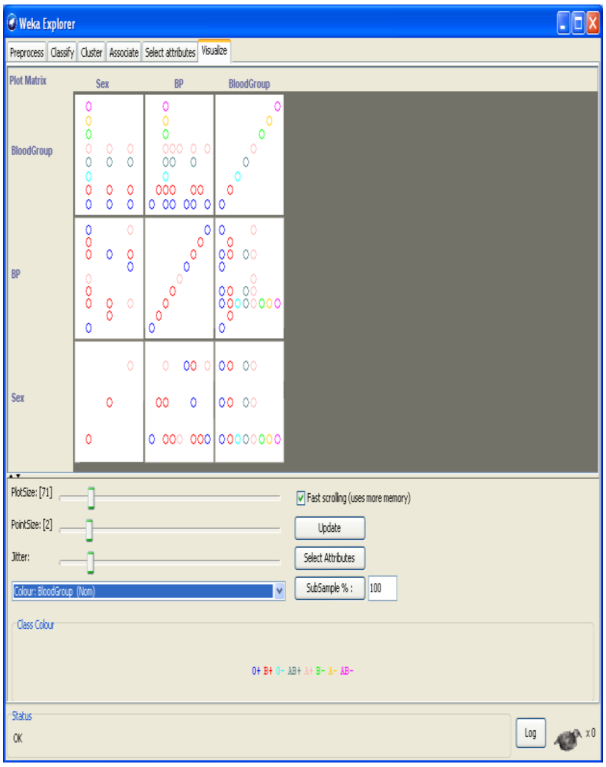
a) From the apriori output we can find that From above algorithm it is clear that there is association between donor having sex male with Blood group=B+. Similarly also there is association between donor having sex male with Blood group=A+ and Blood group=O+ having BP=120/80.
b) Confidence refers to the likelihood of the Sex of donor and Support refers to the % of involvement in the donation.

   From above algorithm it is clear that there is association between donors having disease HBsAg with disease

HCV. Similarly also there is association between donor having disease HBsAg with disease HIV.
Confidence refers to the likelihood of the blood group of donor and Support refers to the % of involvement in the disease.

c) From above algorithm it is clear that there is association between person having pH acidic with color pale yellow and appearance clear. Similarly also there is association between person having pH acidic with Sex male and female. Confidence refers to the likelihood of the patient and Support refers to the % of involvement in the test.

Step 7: Visualization of result of the Apriori algorithm.



PlotSize: [71]

PointSize: [2]

Jitter:

Colour: BloodGroup (Nom)

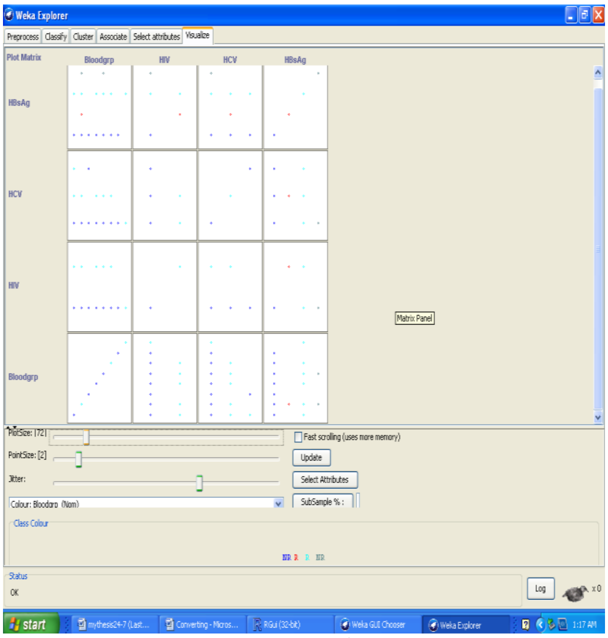
SubSample % : 100

O+ B+ O- AB+ A+ B- A- AB-

Status

OK

Figure 5a: visualization result for Apriori algorithm in Weka for blood.arff



PlotSize: [72]

PointSize: [2]

Jitter:

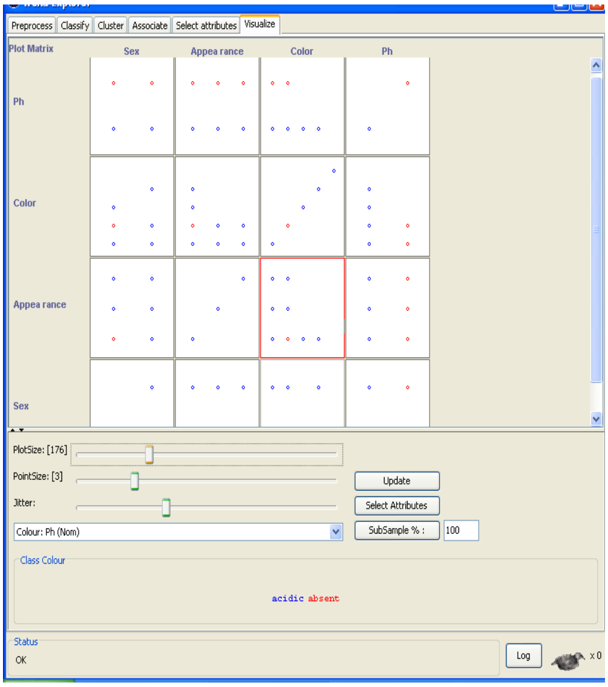
Colour: Bloodgrp (Nom)

SubSample % :

Status

OK

Figure 5b: visualization result for Apriori algorithm in Weka for disease.arff



PlotSize: [176]

PointSize: [3]

Jitter:

Colour: Ph (Nom)

SubSample % : 100

acidic absent

Status

OK

Figure 5c: visualization result for Apriori algorithm in Weka for urine.arff

7. Conclusion

This paper is used to check relationship between the number of blood donors of a particular age and blood group as well as blood group of donors and disease.The purpose of this work is to analyze a data to extract knowledge of blood donor's association to aid clinical decisions in blood bank center. This study utilized real world data collected from blood bank department of Hindratna Prakashbapu Patil Blood Bank in Sangli, Maharashtra and used Apriori algorithm for the association of donors, which can help the blood bank owner to make proper decisions faster and more accurately.

From blood bank department it is found that blood donors having blood group O+, A+ and B+ are more available so blood bank keep stock of the blood, while blood bank need to encourage other blood group donors of both male and female sex with 20-50 age group and as well as with standard Hb range donors(13.8-17.2 for male and 12.1-15.1 for female). With respect to disease, there is no positive relationship between blood group of donor and disease. In pathology department researcher found that maximum respondents for test are having age group below 40. The results had drawn by applying data mining algorithm like Apriori on healthcare data.

The future research can be made to analyze the impact of data mining on all departments of any hospital. The foremost intention for the research is to improve the performance of healthcare. The study suggested that to improve the performance, it should enhance the functionalities with the help of data mining. Data mining algorithms with WEKA programming helps in improving healthcare functionality.

References

[1] Ankit Bhardwaj, Arvind Sharma, V.K. Shrivastava, Data Mining Techniques and Their Implementation in Blood Bank Sector-A Review,International Journal of Engineering Research and Application(IJERA) ISSN:2248-9622, Volume 2, Issue 4, July-August 2012, pp.1303-1309.

[2] Arvind Sharma, P.C. Gupta, Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool, International Journal of Communication and Computer technologies, ISSN: 2278-9723, Volume 1, Issue 2, September 2012.

[3] SonamNarwal and Mr. KamaldeepMintwal, Comparison the Various Clustering and Classification Algorithms of WEKA Tools,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013 ISSN: 2277 128X .

[4] vana D. Radojevic et al., Total coliforms and data mining as a tool in water quality monitoring,African Journal of Microbiology Research, Vol. 6(10), 16 March 2012.

[5] Alaahamouda,Automated Red Blood Cells counting,International Journal of Computing Science, Vol.1, No.2, February 2012.

[6] DevchandChaudhari& Dr. Ravindra S. Hegadi, Data Mining in Blood Platelets Transfusion Using Classification Rule:",pp: 1-8.

[7] Michael Silver TaikiSakataHua.Steven B. Dolins, Michael J. O'Shea, Journal of Healthcare Information Management, 2001, vol. 15, no. 2.

[8] DivyaTomar and SonaliAgarwal, A survey on Data Mining approaches for Healthcare ,International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266

[9] HianChyeKoh and Gerald Tan, —Data Mining Applications in Healthcare, Journal of Healthcare Information Management – Vol 19, No 2.

Author Profile

Miss. Swati Laxman Nalawade has completed MCA, M.Phil in Computer Application from CSIBER Kolhapur. She is working as Assistant Professor in the Krantiagrani G.D. Bapu Lad Mahavidyalaya, Kundal, BCA Dept., Shivaji University, Kolhapur. Her fields of research interest are data mining, Weka and R tool. She has published papers in the international journals and presented research papers in international and national conferences.

Dr. R. V. Kulkarni has completed M. Sc. (Stats), Ph.D. He is currently working as head of department in CSIBER Kolhapur, Maharashtra, India. He is working as Ph.D. and M. Phil guide for number of students from CSIBER as well as Shivaji University Kolhapur. He has published papers in the international journals and presented research papers in international and national conferences.