Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Review of Association Rule Mining Using Apriori Algorithm

**Shelly Ahuja, Gurpreet Kaur**
*Chandigarh University*
*India*

***Abstract— Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Association rule is one of the popular techniques used for mining data for pattern discovery is the KDD[16]. Rule mining is an important component of data mining. To find regularities/patterns in data, the most effective class is association rule mining[9]. Mining has been used in many application domains. In this work, an efficient mining based algorithm for rule generation is presented. By using Apriori algorithm we improve the precision and recall and F-measure value.***

***Keywords— Data mining, classification, association rule mining, apriori, Optimization,KDD.***

## I. INTRODUCTION

The demand of data mining is growing increasingly fast for extracting useful information from a data set. Therefore now days, the theory of data mining becomes more and more significant. Data mining is the analysis step of the "Knowledge Discovery in Databases" process (KDD) [9]. Data mining is the computational process of discovering patterns in large data sets which involves methods that utilized artificial intelligence, statistics, machine learning and database systems. Extraction of information from a data set and transformation of this information in to some understandable form for further use is the main goal of the data mining process. Aside from the raw analysis step, it makes utilization of database and its data management aspects. Mining of data is done through data pre-processing, various models and complexity considerations, inference considerations, post-processing of that per-processed data, visualization and finally updating.
Data mining finds its applications in various fields. These fields include financial data analysis, Telecommunication industries, Retail industries, health care and biomedical research and science and engineering. The tasks of data mining are very diverse and distinct because there are many pattern in large data set. Different types of methods and techniques are needed to find different kinds of pattern. Based on kinds of pattern, tasks in data mining can be classified in to summarization, regression, classification, clustering, association rule learning, and trend analysis. From the various fields of data mining, we chose classification field. Classification is the process of classifying the data set into different classes where data belongs to similar classes. [9] Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). Classification is a supervised learning method used for mining data. Classification is the task of generalizing known structure to apply to new unclassified data. For example, an image classifier differentiates or classifies various images on the bases of various features. Classification problems try to determine the characteristics which correctly identify the class to which each instance belongs to. Classification is similar to clustering, the main difference being that, in classification, the class to which each instance in the dataset belongs to is known as [2] apriori. Classification process is basically rule based and name based. Rule based classification is of our choice. as the name suggest rule based classification techniques are based on some specific rules. Rule based classification can be achieved using

- Decision tree
- Nearest neighbor
- Neural networks
- Apriori

## II. ASSOCIATION RULE MINING

A popular method for discovering interesting relations between variables and patterns in large databases is Association rule learning. Association rule mining is a widely-used technique for classification in data mining[11]. It is intended to identify strong rules discovered in databases using different measures of interest. Association rules [10] are usually required to satisfy a user specified minimum confidence and a user-specified minimum support simultaneously. Association rule generation is usually divided into two separate steps:

1. Minimum support is applied to find all frequent itemsets in a database.
2. Thus frequent itemsets and the minimum confidence constraint are used to form rules [9].

The first step needs more attention, while the second step is straight forward. Complexity is a major problem in association rule mining. Even for moderate-sized databases it is intractable to find all the relationships. This is why a

 

mining approach defines an interestingness measure to guide the search and prune the search space.[18] Therefore, the result of an association rule mining algorithm is the set of all interesting ones but not the set of all possible relationships [9]. However, the term interesting depends on the application. The large number of rules makes it difficult to compare the output of different association rule mining algorithms.

It is basically useful for discovering relationships among items from large databases. A basic association rule is in the form X→Y which says that if X is true for some instance in a database, than Y is also true for the same instance, with a particular level of significance as measured by two coefficients, support and confidence. Some support and coverage thresholds are given, all these rules whose support and confidence are respectively above this threshold are taken for output. The selected attributes of related association rule in the database has been encapsulated, for e:g bread→butter: 0.02 support; 0.70 coverage denotes that in the database, 70% of the people who buy bread also buy butter, and these buyers constitute 2% of the database. This rule signifies a positive (directional) relationship between bread and butter buyers . [8]

Apriori is a well known algorithm because of its effective results in knowledge discovery. Apriori is a classic algorithm for frequent item set mining and association rule learning over unstructured databases[1]. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database[3]. The frequent item sets resulted through Apriori can be used to determine association rules. Apriori algorithm is the best-known algorithm to mine association rules. In Apriori algorithm the breadth-first search strategy is used to count the support of itemsets [12].

## II. PROPOSED APPROACH USING ASSOCIATION RULE MINING

In this work, we present an efficient mining based on association rule generation. By using predefined efficient algorithm we find the positive and negative association rules. Results show the effectiveness of our approach. Mining of patients' data base on Weka simulator and checking the precision and recall with maximum number of attribute and forming clusters, analysis of data base with different parameters on the basis of classification using rule based classification. In rule based classification, we generates rule using filtered associator which works along with apriori algorithm. It generates candidate item sets of length $k$from item sets of length [9]. Then it prunes the candidates which have an infrequent sub pattern. Data base shows patients classification which defines person is sick or not. The purpose of work is to present an efficient classification using apriori algorithm for association rule generation. Thus, the objectives of our proposed method are as follow:

- Generating relevant data base.
- An efficient algorithm for rule based classification and simulation for data mining
- Analysis of mining on the basis of precision, recall and f measure.
- Weka simulator has been used.

**DATA SET TAKEN**

**TItle: Thyroid disease records**

```
@data
41,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,1.3,t,2.5,t,125,t,1.14,t,109,f
,?,SVHC,negative
23,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,4.1,t,2,t,102,f,?,f,?,f,?,othe
r,negative
46,M,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.98,f,?,t,109,t,0.91,t,120,f,
?,other,negative
70,F,t,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.16,t,1.9,t,175,f,?,f,?,f,?,o
ther,negative
70,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.72,t,1.2,t,61,t,0.87,t,70,f,
?,SVI,negative
18,F,t,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.03,f,?,t,183,t,1.3,t,141,f,?
,other,negative
59,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,?,f,?,t,72,t,0.92,t,78,f,?,oth
er,negative
80,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,2.2,t,0.6,t,80,t,0.7,t,115,f,?
,SVI,sick
66,F,f,f,f,f,f,f,f,f,f,f,f,f,t,f,f,t,0.6,t,2.2,t,123,t,0.93,t,132,f
,?,SVI,negative
68,M,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,2.4,t,1.6,t,83,t,0.89,t,93,f,?
,SVI,negative
84,F,f,f,f,f,f,f,f,f,f,f,f,f,t,f,f,t,1.1,t,2.2,t,115,t,0.95,t,121,f
,?,SVI,negative
67,F,t,f,f,f,f,f,f,f,f,f,f,f,f,f,t,0.03,f,?,t,152,t,0.99,t,153,f,
?,other,negative
71,F,f,f,f,t,f,f,f,f,f,t,f,f,f,f,f,t,0.03,t,3.8,t,171,t,1.13,t,151,
f,?,other,negative
59,F,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,2.8,t,1.7,t,97,t,0.91,t,107,f,
?,SVI,negative
28,M,f,f,f,f,f,f,f,f,f,f,f,f,f,f,t,3.3,t,1.8,t,109,t,0.91,t,119,f
```

Fig1:Database of Thyroid pateints

 

Relevant Information:

This data file contains details of patients.

Number of Instances: 3772

Number of attributes: 30 (overall)

Attribute Information- Some of the attributes are:

- 'age'- define age of the patient
- 'on thyroxine'- defines some disease factor and shows results only by true or false.
- 'sick'- define about the person is sick or not.

Information about the dataset

CLASSES: sick, negative.

- 'sick' – if person having some positive factors of infection.
- 'negative' – if person having no infection.

## III. BASIC FLOW DESIGN

In our work, the data set that we take is of Thyroid disease patient. We apply an association rule mining technique named Apriori for mining our data set. Firstly, candidate generation phase comes after that no. of occurrence of frequent itemsets is counted. If the support factor is greater than the decided threshold, positive and negative rules are generated for large itemset with support. By applying these rules classification of data set is done.

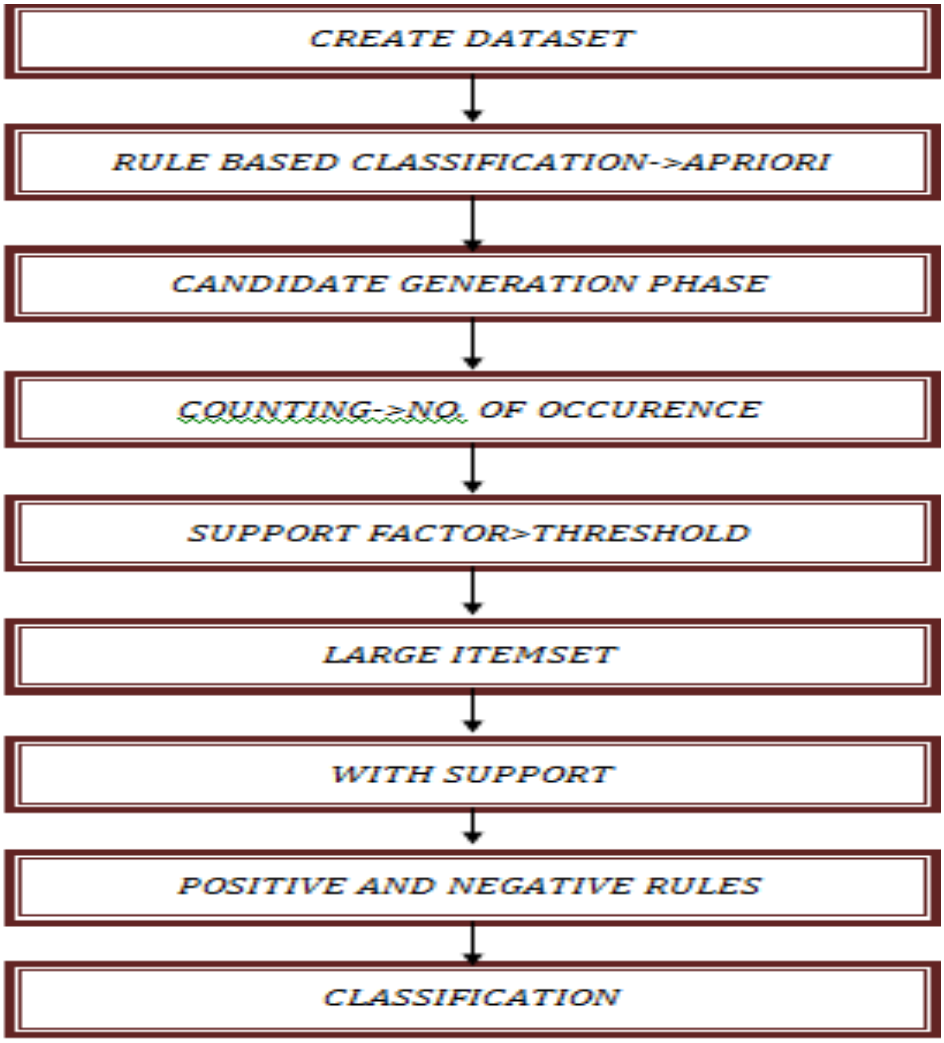


Fig 2: Basic flow design of proposed work

## IV. RESULTS

In this classification, we generates rule using filtered associator which works along with apriori algorithm As a result we get the better Precision ,Recall ,F-measure value. As a result dataset has been categorized in two classes sick and negative.

- 'negative'–If a person is having no infection. Given a class label "a".
- 'sick'– If a person is having some positive factors of infection. Given a class label "b".

 

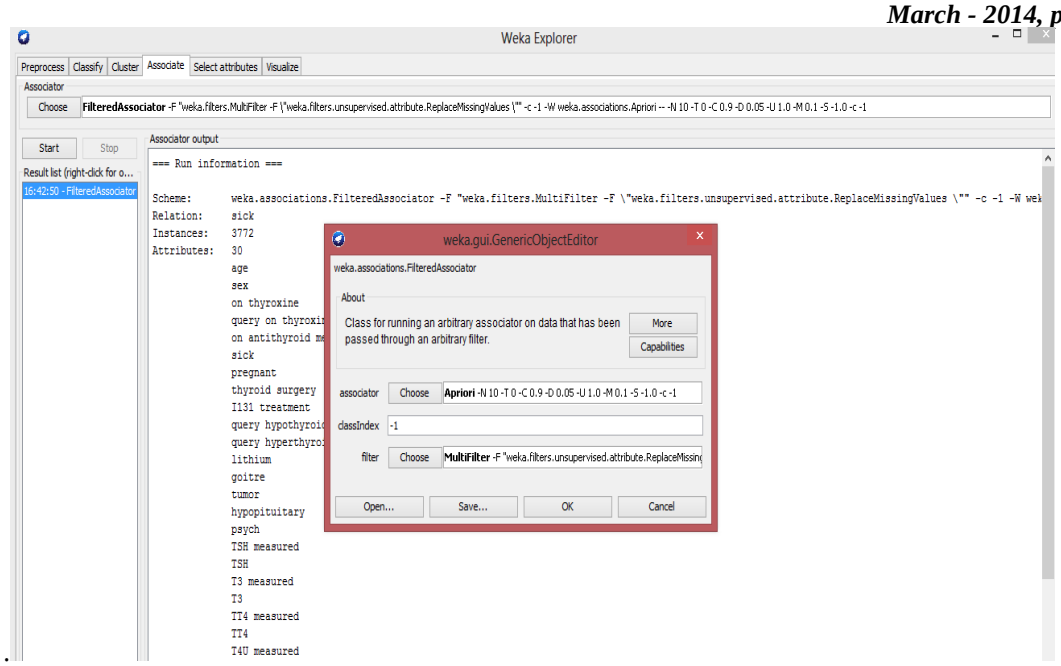


Fig.3: Showing used associator as Apriori and filter used as Mulfilter

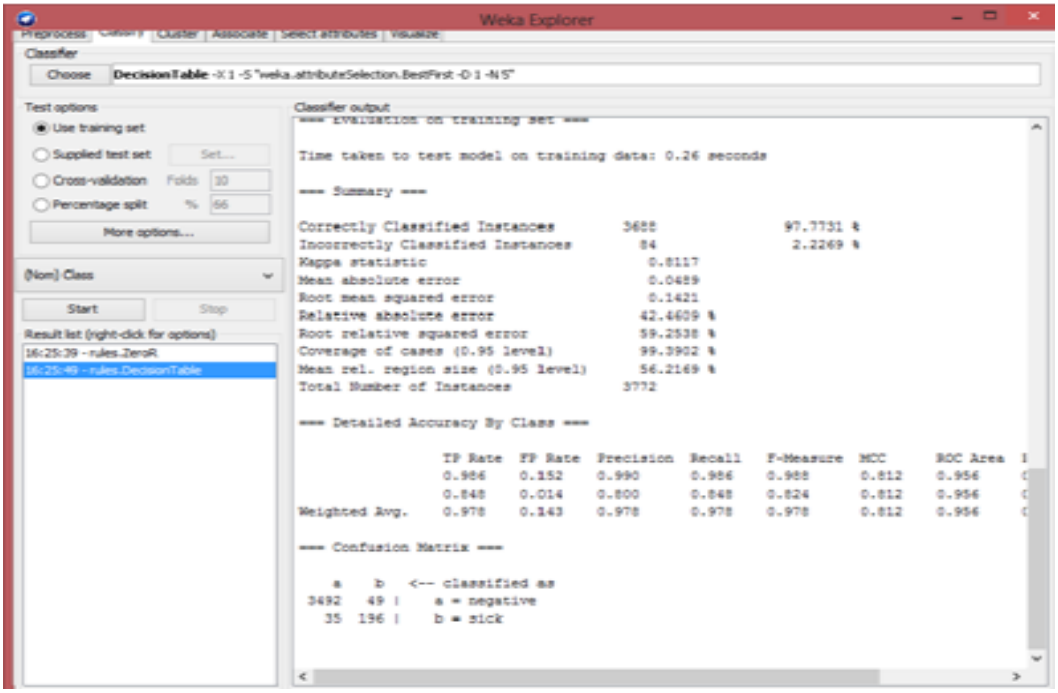


Fig.4 : Showing Precision ,Recall and F-measure and generated classes “a” and “b”.

## VI. CONCLUSION AND FUTURE WORK

Data-mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making [10]. Rule is an important component of data mining. Rules are an important class of methods of finding regularities/patterns in data. Association rule mining is perhaps the most important model invented and extensively studied by databases and data mining community. In this work, we present an efficient mining using apriori for rule generation. In this work we apply efficient algorithm to achieve better precision and recall values. Most of the Association rules are optimized using Ant colony Optimization and Biogeography based optimization till now. Pollination based optimization is still unexplored area for research. In future PBO will be used for optimization of association rules.

## REFERENCES

[1] Bansal Divya., Bhambhu Lekha., “Execution of Apriori Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

 

[2] Goswami D.N., Chaturvedi Anshu., Raghuvanshi C.S., "An Algorithm for Frequent Pattern Mining Based On Apriori" (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 04, 2010, 942-947.
[3] Hassan M., Mohammed Al-Maolegi., Bassam Arkok., "An Improved Apriori Algorithm for Association Rules" International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08.
[4] Wu Huan., Lu Zhigang., Pan Lin., Xu Rongsheng., "An Improved Apriori-based Algorithm for Association Rules Mining" Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.
[5] Aggarwal Shruti., Kaur Ranveer., "Comparative Study of Various Improved Versions of Apriori Algorithm" International Journal of Engineering Trends and Technology (IJETT), Volume4,Issue4,April 2013.
[6] Joshi Sunil., Dr. Jadon R. S., Dr. Jain R. C., "An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function" International Journal of Computer Applications, Volume 9– No.9, November 2010.
[7] Sheila A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation" International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.
[8] Smitha, T., Sundaram, V., "Association Models For Prediction With Apriori Concept" International Journal of Advances in Engineering & Technology, Nov. 2012.
[9] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499.
[10] Yang Q, Wu X (2006) "10 challenging problems in data mining research" International Journal of Information Technology & Decision Making ,Vol 5 No. 4,2006.
[11] Sheila A. Abaya, "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation" International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.
[12] Sanjeev Rao, Prianka Gupta, "Implimenting Improved Algorithm Over Apriori Data Mining Association Rule Algorithm", In: preceeding of IJCST,VOL.3, Issue 1, 2012.
[13] Han, J., Kamber, M. 2001. " Data Mining: Concepts and Techniques", Morgan Kaufman.
[14] I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools And Techniques* Morgan Kauffman, 2005.
[15] Angeline Delighta D.Magdalene and I.Samuel Peter James "Association Rule Generation Using Apriori Mend Algorithm for Student's Placement" IJES, March 2012.
[16] Hipp, J., Guntzer, U., Gholamreza, N. (2000). Algorithm for Association Rule Mining: A General Survey and Comparison, ACM SIGKDD, volume 2 (Issue 1), p. 58.
[17] XindongWu, Vipin Kumar et al., "Top 10 algorithms in data mining" Knowledge Information System, Springer (2007).
[18] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview" GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

 