ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

A Case Study of Association Rule for the Confidence and Support Using Apriori Algorithm

[1] S.Uma Gowri, [2] Dr V. Lakshmi Praba
[1] M.Phil. Scholar, Department of Computer Science, Rani Anna Government College, Tirunelveli,India,
[2] Assistant Prof, Department. of Computer Science, Rani Anna Government College, Tirunelveli.

Abstract: **- The data mining approach, a relatively new technique, is deployed in large databases to find novel and useful patterns that might otherwise remain unknown. Association rule learning which follows the apriori algorithm is a popular and well research technique for discovering interesting relations between variables in large databases. Among various algorithms available for mining frequent item set we include apriori algorithm. Main objective of apriori algorithm is to find out hidden information which is the major goal data mining. In this study, we use United States Congressional Voting dataset and analyze the information based on predefined rules.**

Keywords: **Apriori, Association Rule, Support and Confidence.**

I. INTRODUCTION

The Apriori algorithm is based on the fact that if a subset 'S' appears 'k' times, any other subset 'S' that contains S will appear k times or less. So if S is doesn't pass the minimum Support threshold value, neither does S'. There is no need to calculate S', it is discarded apriori. The Apriori algorithm needs n+1 scans if a database is used, where 'n' is the length of the longest pattern. Item set is set of items, a group of element that represents together as a single entity [1]. A frequent item set is an item set that occurs frequently. In frequent pattern mining to check whether a item set occurs frequently or not, we have a parameter called Support of an item set.

Support(S) of an association rule is described as the percentage of record that hold union of X and Y to the total number of record in the database

Support It is the probability of item or item sets in the given transactional data base:

$$Support(X) = \frac{n(X)}{n} \qquad \text{..... (1)}$$

Where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set X. Therefore, Support (X-->Y) = Support(XUY).

Confidence

It is conditional probability, for an association rule X-->Y and defined as

.....(2)

Confidence(C) of an association rule is defined as the percentage of the number of transaction that contains union of X and Y to the total number of records that include X. confidence is a calculated of strength of the association rule. The problem of mining association rule is to generate all association rules that have Support and confidence greater than the user-specified minimum Support (called minsup) and minimum Confidence (called minconf) respectively. The problem of discovering all association rules can be decomposed into two sub problems:

(1) Finding all the frequent itemsets (whose Support is greater than minsup), also called large item sets.
(2) Generating the association rules derived from the frequent itemsets. If and X are frequent itemsets, the rule holds if the ratio of Support( ) to Support(X) is, at least, as large as minconf.

Lift

The definition of lift is compared to the probability of 'ante' and 'conseq' happening together separately to the observed frequency of such a combination, as a measure of interestingness. In lift value is 1, then the probabilities of 'ante' and 'conseq' its occurring together is independent and there is no special relationship.

Class Predictiveness and Predictability

The definition of the class predictiveness and predictability is between a class measure and comparison of the products. Predictability component is relative to a specific class and attributes.

II. LITERATURE REVIEW

K. Padmavathi, R. Aruna Kirithika[1] in this paper uses Fp-growth and Apriori algorithm. Both algorithms are

scanning the complete vote dataset.the Fp-growth algorithm is done with two times Scanning for vote dataset, but apriori algorithm Scanning is performed multiple times. After scanning process the fp-growth algorithm is efficient than apriori algorithms.

Minal G. Ingle and N. Y. Suryavanshi [2] in this paper used Apriori algorithm and Improved Apriori algorithm. This algorithmis evaluated on textile dataset. The Apriori applied dataset to find out frequent itemset decision tree can be used for the multileavel association rule mining Textile database contains the attributes like item and itemsets.

Mohammed Al-Maolegi and Bassam Arkok [3] in this paper analyze improved Apriori algorithm. The association rules is divided into two phases: detection the frequent itemsets and generation of association rules. In the first phase, every set of items is called itemset, if they occurred together greater than the minimum support threshold, this itemset is called frequent itemset. Finding frequent itemsets is easy but costly so this phase is more important than second phase. In the second phase, it can generate many rules from one itemset.

Pratibha Mandave, Megha Mane and Prof. Sharada Patil [4] In this paper used Apriori and improved Apriori algorithmthat proposed to compute the frequency of frequent k-item sets when k-item sets are generated from (k-1)-item sets. If k is greater than the size of transaction T, there is no need to scan Transaction T which is generated by (k-1)-item sets according to the nature of Apriori algorithm, and we can remove it.

Himani Bathla and Ms. Kavita Kathuria [5] in this paper to Proposed the comparative of two algorithms: Apriori algorithm and Filter Associator. They have analyzed the generation of frequent itemset using sales transaction dataset.

Jiao Yabing [6]in this paper focused on how to solve the efficient problems of Apriori algorithm and raise another association rules mining algorithm. In this Apriori algorithm Ck-1 is compare with support level once it was found. In this optimized algorithm is that before the candidate item sets Ck come out, further prune Lk-1, count the times of all items occur in Lk-1, delete item sets with this number less than k-1 in Lk-1. The number of connecting items sets will decrease, so that the number of candidate items will decline.

Jayshree Jha and Leena Ragha [7] in this paper analyze improved Apriori algorithm using educational dataset. Apriori algorithm with a main motivation of reducing time and number of scans required to identify the frequent itemset. They presented an Improved Apriori algorithm based on Bottom up approach and Support matrix to identify frequent item set. The proposed algorithm replaces arbitrary user defined minimum support with functional model based on Standard Deviation.

M.S. Mythili and A.R. Mohamed Shanavas [8] in thispaper used Aprioriand FP-Growth algorithm. In the aims of Association Mining to extract attention-grabbing correlations, frequent patterns, and association structures among set of things or objects in transaction data based relational databases or different data repositories. The distinction between apriori and FP growth algorithm generate candidate frequent itemset and also the FP growth algorithm avoids candidate generation.

Smitha.T and V.Sundaram [9] in this paper used different three strong association algorithms. Studied to show how association rules will be effective with the dense data and low support threshold. Association rules allows to identify the behavior pattern with respect to a particular event where as frequent items are used to find how a group are segmented for a specific set.

Sikha Baguiet, Dustin Mink, and Patrick Cash [10] in this paper used apriori algorithm that can be used to study or mine voting patterns in the US House of Representatives. We have shown the whole data mining processing – from processing input data to preprocessing to attribute relevance analysis to the use of advanced data mining techniques like association rule mining and decision tree generation and analysis to presenting information (in the form of rules) and conclusions.

III. DATA SET DESCRIPTION

- Title: 1984 United States Congressional Voting Records Database
- Source Information:

(a) Source: Congressional Quarterly Almanac, 98th Congress,
2nd session 1984, Volume XL: Congressional Quarterly Inc.
Washington, D.C., 1985.
(b) Doner: Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
(c) Date: 27 April 1987

- Past Usage

- Predicted attribute: party affiliation (2 classes)

- Relevant Information

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac (CQA). The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflictof interest, and did not vote

ISSN (Online) 2394-2320

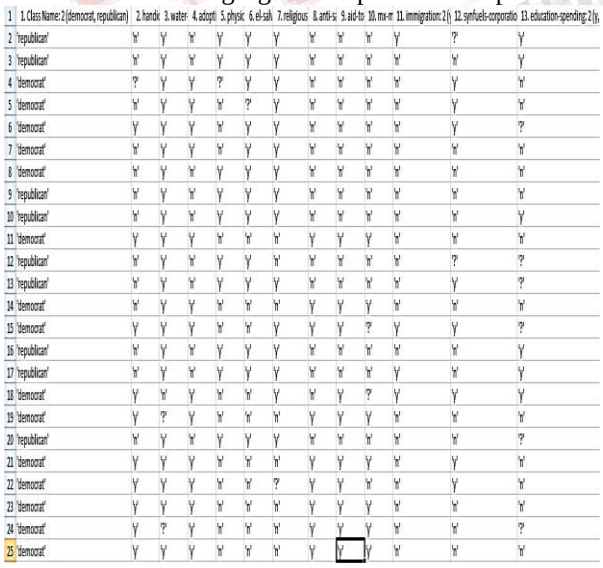
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

or otherwise make a position know (these three simplified to an unknown disposition).

- Number of Instances: 435 (267 democrats, 168 republicans)
- Number of Attributes: 16 + class name = 17 (all Boolean valued)
- Attribute Information: All the attribute information are listed below:

- ✓ Class Name: 2 (Democrat, Republican)
- ✓ handicapped-infants: 2 (y, n)
- ✓ water-project-cost-sharing: 2 (y, n)
- ✓ adoption-of-the-budget-resolution: 2 (y, n)
- ✓ physician-fee-freeze: 2 (y, n)
- ✓ el-Salvador-aid: 2 (y,n)
- ✓ religious-groups-in-schools:2 (y, n)
- ✓ anti-satellite-test-ban: 2 (y, n)
- ✓ aid-to-Nicaraguan-contras: 2 (y, n)
- ✓ mx-missile: 2 (y, n)
- ✓ immigration: 2 (y, n)
- ✓ synfuels-corporation-cutback:2 (y, n)
- ✓ education-spending: 2 (y, n)
- ✓ superfund-right-to-sue: 2 (y, n)
- ✓ duty-free-exports: 2 (y, n)
- ✓ export-administration-act-South-Africa: 2 (y,n)

- Missing Attribute Values: Denoted by "?" "?" in this database indicates that it's categorized under unknown dispost.
- Class Distribution: (2 classes)
  45.2 percent are Democrat
  54.8 percent are Republican
- The following figure 1 depicts the sample dataset



IV. METHODOLOGY

Data mining is the important method to increase efficiency in mining dataset. Association rule mining is used to get association rules that convince the predefined minimum support and confidence from a given datasets.

V. ASSOCIATION RULE-APRIORI ALGORITHM

The Apriori algorithm is a well-known association rule algorithm and is used in most commercial products. In this paper voting dataset of frequent itemset is used. That generates two different types of classes. Republican and Democratbased on minimum Support and Confidence that generates four rules. Each rule is different from one and another.
The Working functionality of the study as illustrated below:
1. Generate frequent item sets
2. Generate rules

Similarly set values as 0's/1's for all attributes form a matrix with 0's and 1's as follows:
Votes = cell(size(dem));
For i = 1:length(dem)
Votes{i} = find(arr(i,:));
End
First Step we changed the dataset values N and Y into 0's and 1's
arr = strcmp(C(:,2:end), 'y');
Second column value if y then 1 else 0
arr = [arr, strcmp(C(:,2:end), 'n')];
Second column value if n then 1 else 0
dem = strcmp(C(:,1),'democrat');
rep = strcmp(C(:,1),'republican');
arr = [dem,rep,arr];
Displaysnon zero elements for each row.
By using Apriori algorithm the following result are obtained based [Table I] on minimum Support and Confidence

Table I: Results Obtained

Resultant type	Value
Minimum Support	0.30
Frequent Item sets Found	1026
Number of Support Data	2530
Max Level Reached	7-itemsets
Minimum Confidence	0.90
Rules Found	2942

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

VI. SAMPLE SCREEN SHOT OF THE RESULT OBTAINED IN FIG 1

```
Minimum Support       : 0.30
Frequent Itemsets Found: 1026
Max Level Reached     : 7-itemsets
Number of Support Data : 2530
Minimum Confidence    : 0.90
Rules Found           : 2942

{El Salvador = Yes, Budget Resolution = No, Mx Missile = No} => {Republican}
  Conf: 0.91 Lift: 2.36 Sup: 0.30
  Correct!  Expected Conf 0.91

{Budget Resolution = Yes, Mx Missile = Yes, El Salvador = No} => {Democrat}
  Conf: 0.97 Lift: 1.59 Sup: 0.36
  Correct!  Expected Conf 0.97

{Physician Fee Freeze = Yes, Right To Sue = Yes, Crime = Yes} => {Republican}
  Conf: 0.94 Lift: 2.42 Sup: 0.30
  Correct!  Expected Conf 0.94

{Physician Fee Freeze = No, Right To Sue = No, Crime = No} => {Democrat}
  Conf: 1.00 Lift: 1.63 Sup: 0.31
  Correct!  Expected Conf 1.00
```

Fig 1 Rule and estimated execution time

From the figure it is observed that if a rule is applied then the following results are obtained [Table II].

Table II: Confidence, Lift and Support rules applied

RULES	CONFIDENCE	LIFT	SUPPORT
Rule 1	0.91	2.36	0.91
Rule 2	0.97	1.59	0.36
Rule 3	0.94	2.42	0.30
Rule 4	1.00	1.63	0.31

The elapsed time for executing the algorithm is 21.662582 milliseconds.

VII. CONCLUSION

Data mining deals with processing of large, complex and noisy data. Association rule discovery is one of the most popular and successful tool in data mining. The association rule technique is used in Apriori algorithm. In this paper, we have focused on analysis of association rule using Apriori algorithm. For the considered attributes the are confidence and support level are implemented in benchmark data of US Congressional Voting dataset. In future other association rule algorithm can also be implemented and compared.

REFERENCES

[1] K.Padmavathi,R.ArunaKirithika, ”Performance Based Study of Association Rule Algorithms On Voter DB”, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 4, June 2014.

[2] Minal G. Ingle and N. Y. Suryavanshi, ”Association Rule Mining using Improved Apriori Algorithm”, International Journal of Computer Applications (0975 – 8887) Volume 112 – No 4, February 2015.

[3] Mohammed Al-Maolegi and Bassam Arkok, ”An Improved Apriori Algorithm For Association Rules”, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, February 2014.

[4] Pratibha Mandave, Megha Mane and Prof. Sharada Patil,” Data mining using Association rule based on APRIORI algorithm and improved approach with illustration”, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 3 Issue2 November 2013.

[5] Himani Bathla, Ms. Kavita Kathuria,” Association Rule Mining: Algorithms Used”, International Journal of Computer Science and Mobile Computing”, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.6, June- 2015, pg. 271-277.

[6] Jiao Yabing,” Research of an Improved Apriori Algorithm in Data Mining Association Rules”, International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[7] Jayshree Jha and Leena Ragha, “International Journal of Information and Computation Technology”, ISSN 0974-2239 Volume 3, Number 5 (2013), pp. 411-418.

[8] M.S. Mythili and A.R. Mohamed Shanavas,” Performance Evaluation of Apriori and FP-Growth Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 79 – No10, October 2013.

[9] Smitha.T and V.Sundaram, ”Association Models For Prediction With Apriori Concept”, International Journal of Advances in Engineering & Technology, Nov. 2012.

[10] Sikha Bagui, Dustin Mink, and Patrick Cash,”Data Mining Techniques To Study Voting Patterns In The Us”, Data Science Journal, Volume 6, 20 April 2007.