International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Finding Pattern using Apriori Algorithm through WEKA Tool

Chaman Verma

Research Scholar, Faculty of Informatics, EOTVOS Lorand University, Budapest (Hungary)

ABSTRACT

In many field including agriculture, health, transportation and education, there is huge need of promotion. Today's, in promotion of business, there is need to multidisciplinary science to perform large scale analysis includes product recommendation, market analysis and demand forecasting by applying scientific efforts. These efforts are basically skill of extracting of knowledge from large or diverse data set. Data Science is multidisciplinary science of deriving significant insights from data and to produce data products. Data Science is collective field that joints skills from various domains such as Software engineering, statistics and data mining. It focuses more on data rather than programming code. It also emphasizes on storing, transforming, cleaning and processing of unstructured data. Data science also concentrates on reuse the function of existing data. Further, it plays important role in predictive analysis of any business firm. Association rules play a vital role in every research domain. In order to predict frequent used pattern many algorithm are used by various researchers. This paper describes execution of popular data mining algorithm named Apriori using WEKA 3.8 tool.

Keywords: - Association Rule, Confidence, Data Set, Support

I. INTRODUCTION

Data mining is the extraction of implicit, previously unknown and potentially useful information from huge data. Recently, data mining studies have been carried out in many engineering disciplines. Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Many of tool and language can be used to to creation association rules such as Python, R Language and Weka. There are many algorithms have been developed to extract the association rules from the large databases such FP Growth, CBPNARM and Context Based association rule. Apriori algorithm is the most popular algorithm to extract the association rules from the databases [1].

The University of Waikato in New Zealand developed WEKA tool in JAVA language that implements data mining algorithms. WEKA is abbreviated for Waikato Environment For Knowledge Analysis. WEKA is open source software which is freely available on internet. It has the huge storage of machine learning algorithms for

data mining problems. We can easily apply these algorithms on desired dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules. In WEKA, Explorer is used for exploring and extracting the dataset on which the operations has to be performed. Experimenter is used to perform experiments or statistical tests on the dataset. Knowledge Flow provides same functionalities as provided by Explorer but with a drag-and-drop interface. It helps in incremental learning. Simple CLI provides simple Command Line Interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface [2].

II. THEORETICAL OF APRIORI ALGORITHM

As we know that Apriori algorithm is used to find frequent item sets from transaction database. It requires the two major measurement support and confidence. The support also has a desirable property that can be exploited for the efficient discovery of association rules. Confidence, on the other hand, measures the reliability of the inference made by a rule. For a given rule X→Y the higher the confidence, the more likely it is for Y to be present in transactions that contain X. Confidence also provides an estimate of the conditional probability of Y given X. This study finds the frequently characters used in the given four transactions as depicted in Table 1.

III. CANDIDATES' GENERATION PROCESS

In Apriori algorithm, we are used to perform the candidate generation. For this, we require two measures support count and confidence level. We have assumed both the Minimum support and minimum confidence is 50%. We shall find out the frequent item set in given transaction data set. The following are the steps involve into candidate generation in Apriori algorithm.

1. Select the transaction data set and assume both support and confidence level.
2. Calculate the minimum support count for comparison.
3. Generate candidate C1 for single item set with total occurrence.
4. Retain item set with higher support count as compare to minimum support count.
5. Generate candidate C2 for double item set with total occurrence.
6. Retain item set with higher support count as compare to minimum support count.

Table 1. Transaction Data Set

T	Item Sets
T1	A,B,C
T2	A,C
T3	A,D
T4	B,E,F

(Source: Author)

Minimum support Count = (Min support percentage/100)* no. of transactions

(50/100)*4=2

Table 2. Candidate Generate C1

Item	Support count
{A}	3
{B}	2
{C}	2
{D}	1
{E}	1
{F}	1

Table 3. Frequent Item Set L1

Item	Support count
{A}	3
{B}	2
{C}	2

(Source: Author)

(Source: Author)

We have calculated minimum support count which is 2. Table 2 shows the each character with its frequently occurrence in transaction data set. We have removed the characters whose support count is less than the minimum support count. Table 3 shows the frequent item set with single set of frequent characters.

Table 4. Candidate Generate C2

Item	Support count
{A,B}	1
{B,C}	1
{A,C}	2

(Source: Author)

Table 5.Frequent Item Set L2

Item	Support count
{A,C}	2

(Source: Author)

Table 4 depicts the frequency of double character set in transaction data set. Again we have removed the characters set whose support count is less than the minimum support count. Finally, Table 5 shows the frequent item set with double characters occurrence which is 2. Later on, we have built the association rule table which contains our final association rules between frequent characters used in transaction data set. Following is the formula to calculate the confidence for association rule.

Confidence (**A→C**)=Support/occurrence of A in data set

2/3=6.6

Confidence (**C→A**)=Support/occurrence of C in data set

2/2=1

Table 6. Association Rule

Association Rule	Support	Confidence	Confidence in %age
A→C	2	0.66	66%
C→A	2	1	100%

(Source: Author)

Data from Table 6 shows final association rules **A→C and C→A** which has the confidence 0.66 and 1 respectively. These are the final rules which infer that A stands for A and A stands for C. It means occurrence of A with C and occurrence of C with A in transaction data set. These two characters are frequently used most in transaction database. Table 1 shows also occurrence of A is thrice and occurrence of C is twice. Table 2 also infers that in two transactions T1 and T2 both are used together.

IV. EXECUTION OF APRIORI ALGORITHM USING WEKA

The Apriori algorithm was first proposed by Agrawal and Srikant in 1994. The Association rules are created by analyzing data for frequent patterns and using the criteria support and confidence to identify the most important

relationships. They are divided into separate categories in the data mining and used in the Weka to perform the operations [3]. The following are the major steps to create association rules Using WEKA tool.

1. Built CSV (comma delimited file) contains transaction data set.
2. Click on Explorer Tab in WEKA 3.8 to preprocess the data set.
3. Load ARFF file by open file tab in explorer.
4. Select Associate Tab and apply Apriori using start button.

	A	B	C	D
1	T1	T2	T3	T4
2	A	A	A	B
3	B	C	D	E
4	C			F

Fig.1 Data Set in CSV File (Source: Author)

Fig.1 shows four attribute T1,T2,T3 and T4 and and maximum three instances. A row of data is called an instance, as in an instance or observation from the problem domain and a column of data is called a feature or attribute, as in feature of the observation. All transaction data set are stored into comma delimited file in MS-Excel. Another file format ARFF (attribute relation file format) can be also used as well.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (1 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 13

Size of set of large itemsets L(3): 8

Size of set of large itemsets L(4): 2

Best rules found:

1. T2=A 1 ==> T1=A 1    <conf:(1)> lift:(3) lev:(0.22) [0] conv:(0.67)
```

Fig.2 Association Rule (Source: Author)

Data from Fig.2 reveals the outcomes of Apriori algorithm on our transaction data set. The Apriori algorithm brings out one best association rule with 50% minimum support and confidence. It shows the association

between two transactions such as T1 and T2. It contains frequent characters A and C in both transaction T1 and T2 as given in Table 6.

V. CONCLUSION

This paper focuses on finding association rules between data set transactions with the help of support and confidence measurements in WEKA tool. The pattern finding algorithm Apriori discovers only one association rule. It reveals that presence of A with C and presence of C with A in transaction data set. It is also found that Transaction T1 and T2 are associated with each other well due to occurrence of both frequent characters A and C.

VI. LIMITATIONS

1. This study is delimited to Apriori algorithm only.
2. It is confined to 50% support and confidence measure.
3. This study is also limited to virtual transaction data set.

VII. RECOMMENDATIONS

This paper is written for apparent the concept of the execution of Apriori Algorithm in WEKA tool for researchers. The future recommendation for researcher is to apply FP growth algorithm for frequent pattern finding as it is fast as compared to Apriori algorithm.

REFERENCE

[1] Rakesh Agrawal and Ramakrishnan Srikant , " Fast algorithms for mining association rules in large databases", Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[2]http://eisc.univalle.edu.co/cursos/web/material/750061M/1/WekaManual.pdf

[3] Ajay Kumar and R.N Panda (2014),Implementation of Apriori Algorithm using WEKA, International Journal of Intelligent Computing and Informatics, Vol. 1, Issue 1,Page 12-15.