International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.162.01

POA Apriori Algorithm and its Applications in Datamining

T. Ashalatha [1], N. Nagalakshmi [2], Dr. A. Prashanth Rao [3]
[1] Assistant Professor, [2] Assistant Professor, [3]Professor,
Anurag Group of Institutions, Venkatpur, Ghatkesar, Hyderabad

Abstract - Data is the vital to understand customers and market. Sustaining of an organization without proper data is very difficult. Data is an assortment of realities like numbers, words, estimations, perceptions, and so on that has been converted into a structure that computer can process. As a result of computerization of activities in an organization the storing and processing of data has been increased. As a result data should be maintained accurately. In this situation data mining is playing an important role. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information using intelligent methods from a data set and transform the information into a comprehensible structure for further use [1]. One of the data mining technique association rules play a crucial role in data mining; Using association rule we will find the relation between the large data sets. Association rules are used to find the frequent datasets, Apriori algorithm is one the best algorithm in association rules for finding frequent data sets. There are many algorithms associated with Apriori algorithm. In this paper, we are introducing a new filed and work on it. Using Python programming language using Orange Associate, we will implement the new process.

Keywords: Apriori, association rules, python, advanced Apriori

I. INTRODUCTION

Data mining is the process of mining the present raw data into required information. Mining is done to extract the useful information from the present useless data. Extracting the knowledge is the basic motto behind mining the data. Knowledge discovery process is the process used for finding the data from the huge databases. Knowledge Discovery in databases (KDD) process is necessary for huge datasets and for big organizations. Association rules are applied on the data for finding the frequent data sets. Apriori algorithm is one the best association rule that is applied for finding frequent data sets and to find the required item sets. In this paper, the present Apriori algorithm is implemented in a new manner for better results. Apart from the new method, this paper shows how to apply it by using python programming language. With this the drawbacks of the Apriori algorithm will be blocked. In the nearest future we can develop the algorithm by including much more new advancement in python for decreasing the time complexity to the least value.

II. LITERATURE COLLECTION

2.1 APRIORI ALGORITHM

In Data Mining, Apriori is a classical algorithm. Apriori algorithm is used to find frequent item sets basing on association rules. Apriori algorithm is applied on large databases to find the frequent data sets. For example, there is a supermarket, we will use algorithm to find the list of the customers and the list of the items that most of the customers brought from the store.

Apriori is very effective in case of huge data sets, in organizations, in super markets, in universities and in big educational organizations. In medical field also we can use this for finding the type of medicines customers are buying most frequently; we can find the percentage of the medicines which are being used by the customers.

Apriori algorithm is all full of association rules, without association rules Aprior is nothing. So, in the following section association rules are discussed.

2.2 ASSOCIATION RULES

Association rules are used to find the relations, we can use them according to your requirement. Based on the concept of strong rules, Rakesh Agrawal, Tomasz Imieliński and Arun Swami introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarket [2].

Let $I = \{i_1, i_2, i_3, ..., i_n\}$ be a set of n attributes called items and $D = \{t_1, t_2, ..., t_n\}$ be the set of transactions [3]. It is called database. Every transaction, t_i in D has a unique transaction ID, and it consists of a subset of item sets in I.

Consider an example from more market. Here, this is a small example where the data set consists of five items and the no of transactions or users are six.

Suppose that the Item sets may be I= {coco powder, baking soda, condensed milk, icing sugar, vanilla essence}. As mentioned earlier the number of transactions or the number of users is six. If the user buys the product then the value of the files is 1 if not 0.

In this example, as per association rule, the case may be {Coco Powder, Icing Sugar} => {Vanilla Essence}, which means that if coco powder and icing sugar are bought, customers also buy a vanilla essence.

User No	Coco powder	Baking Soda	Condensed Milk	Icing Sugar	Vanilla Essence
U1	1	1	1	1	1
U2	0	1	1	0	0
U3	1	1	0	1	1
U4	1	1	1	0	0
U5	0	0	1	1	1
U6	1	0	1	0	1

To obtain the accurate item set we will calculate support, confidence, lift and conviction.

2.3 SUPPORT

The support of an item set X, $supp(X)$ is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an item set.

$$supp(X) = \frac{\text{Number of transaction in which } X \text{ appears}}{\text{Total number of transactions}}$$

2.4 CONFIDENCE

Confidence of a rule is defined as follows:

$$conf(X \longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

2.5 LIFT

The lift of a rule is defined as:

$$lift(X \longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

This signifies the likelihood of the item set Y being purchased when item X is purchased while taking into account the popularity of Y.

2.6 CONVICTION

The conviction of a rule can be defined as:

$$conv(X \longrightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \longrightarrow Y)}$$

By calculating all the values we will find the appropriate frequent data sets, which will help us to understand the large databases and the relation between the attributes in the databases.

III. IMPLEMENTATION

Apriori algorithm is used for finding item sets; Main concept of the algorithm is anti –monotonicity of the support measure.

Assumptions in the algorithm are as follows:

1. All subsets of a frequent item set must be frequent
2. Similarly, for any infrequent item set, all its supersets must be infrequent too

Example:

Consider the above example

User No	**Coco powder**	**Baking Soda**	**Condensed Milk**	**Icing Sugar**	**Vanilla Essence**
U1	1	1	1	1	1
U2	0	1	1	0	0
U3	1	1	0	1	1
U4	1	1	1	0	0
U5	0	0	1	1	1
U6	1	0	1	0	1

User ID	Items
U1	CP,BS,CM,IS,VE
U2	BS,CM
U3	CP,BS,IS,VE
U4	CP,BS,CM
U5	CM,IS,VE
U6	CP,CM,VE

The following table shows the item set and support.

Item Set	Support
CP	4
BS	4
CM	5
IS	3
VE	4

CP,BS	3
CP,CM	3
CP,IS	2
CP,VE	3

CP,BS,CM	2
CP,CM,IS	1
CP,IS,VE	2

3.1 PROPOSED ALGORITHM

In the original algorithm, we will have data sets and the support value. But here we will add another filed naming group id, by that we can name the group for each user and check whether the item is present in the group or not. This reduces number of computer scans. This is implemented by some researchers but we will implement it in python especially orange associate.

User ID	Items	Group Id
U1	CP,BS,CM,IS,VE	G1
U2	BS,CM	G2
U3	CP,BS,IS,VE	G3
U4	CP,BS,CM	G4
U5	CM,IS,VE	G5
U6	CP,CM,VE	G6

From the above table we will find the group of the item and check whether the item is present in the group or not and find the remaining support sets.

Item Set	Support	Group Id
CP	4	G1,G3,G4,G6
BS	4	G1,G2,G3,G4
CM	5	G1,G2,G4,G5,G6
IS	3	G1,G3,G5
VE	4	G1,G3,G5,G6

3.2 PYTHON IMPLEMENTATION

We will implement python on the above data a new proposed algorithm minimizes the transactions and memory scans.

To run the program with dataset provided and default values for *minSupport* = 0.15 and *minConfidence* = 0.6

python apriori.py -f

INTEGRATED-DATASET.csv

To run program with dataset

python apriori.py -f INTEGRATED-DATASET.csv -s 0.17 -c 0.68
Best results are obtained for the following values of support and confidence:
Support: Between 0.1 and 0.2
Confidence: Between 0.5 and 0.7

3.3 ORANGE ASSOCIATE

This module will implement the proposed algorithm.

Orange associate includes two algorithms, one for association rules and the other for standard Apriori algorithms proposed by Agrawal Sriakanth. Orange supports both the algorithms for finding the frequent data sets.

For example, consider a simple market basket data:

Bread, Milk
Bread, Diapers, Beer, Eggs
Milk, Diapers, Beer, Cola
Bread, Milk, Diapers, Beer
Bread, Milk, Diapers, Cola

The following script induces association rules with items that appear in at least 30% of data instances (transactions):

```
import Orange
data = Orange.data.Table("market-basket.basket")
rules = Orange.associate.AssociationRulesSparseInducer(data, support=0.3)
print "%4s %4s  %s" % ("Supp", "Conf", "Rule")
for r in rules[:5]:
print "%4.1f %4.1f  %s" % (r.support, r.confidence, r)
```

The code reports on support and confidence first five rules found:

```
Supp Conf  Rule
 0.4  1.0  Cola -> Diapers
 0.4  0.5  Diapers -> Cola
 0.4  1.0  Cola -> Diapers Milk
 0.4  1.0  Cola Diapers -> Milk
 0.4  1.0  Cola Milk -> Diapers
```

In Apriori, association rule induction is two-stage algorithm first finds itemsets that frequently appear in the data and have sufficient support, and then splits them to rules of sufficient confidence. Function *get_itemsets* reports on itemsets alone and skips rule induction:

```
import Orange
data = Orange.data.Table("market-basket.basket")
ind = Orange.associate.AssociationRulesSparseInducer(support=0.4, storeExamples = True)
itemsets = ind.get_itemsets(data)
for itemset, tids in itemsets[:5]:
    print "(%4.2f) %s" % (len(tids)/float(len(data)),
                          " ".join(data.domain[item].name for item in itemset))
```

The above script lists frequent itemsets and their support:

```
(0.40) Cola
(0.40) Cola Diapers
(0.40) Cola Diapers Milk
(0.40) Cola Milk
```

(0.60) Beer

3.4 ASSOCIATION RULES INDUCTION ALGORITHMS

AssociationRulesSparseInducer induces frequent itemsets and association rules from sparse data sets. These can be either provided in the basket format or in an attribute-value format where any entry in the data table is considered as presence of a feature in the transaction (an item), and any unknown (empty) entry signifies its absence. AssociationRulesInducer works feature-value data, where am item is a combination of feature and its value (e.g., *astigmatic=yes*).

Sparse (basket) data sets class Orange.associate.AssociationRulesSparseInducer support

Minimal support for the rule. Depending on the data set it should be set to sufficiently high value to avoid running out of working memory (default: 0.3).

Confidence Minimal confidence for the rule.

store_examples Store the examples covered by each rule and those confirming it.

max_item_sets The maximal number of itemsets induced. Orange will stop with inference of frequent itemsets once this number of itemsets is reached.

__call__(*data*, *weight_id*) Induce rules from the provided data set.

get_itemsets(*data*) For a given data set, return a list of frequent itemsets. List elements are pairs, where the first element includes indices of features in the item set (negative for sparse data) and the second element a list of indices supporting the itemset. If **store examples** is False, the second element is None.

To test this rule inducer, we will first create a sparse data sets consisting of list of words in sentences from a brief description of Spanish Inquisition, given by Palin et al.:

Nobody expects the Spanish Inquisition! Our chief weapon is surprise...surprise and fear...fear and surprise.... Our two weapons are fear and surprise...and ruthless efficiency.... Our three weapons are fear, surprise, and ruthless efficiency...and an almost fanatical devotion to the Pope

IV. CONCLUSION

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. The Apriori algorithm learns association rules and is applied to a database containing a large number of transactions. The features of the python programming language and orange associate identified the required data sets with less complexity. The further development of the algorithm is possible by adding more fields and implement them using python programming language. The Apriori algorithm can be implemented in many ways and using different techniques. The latest module allows us to reduce the memory scans and also saves time.

REFERENCES

[1] Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.
[2] https://en.wikipedia.org/wiki/Apriori_algorithtm
[3] https://en.wikipedia.org/wiki/Association_rule_learning
[4] http://blog.hackerearth.com/beginners-tutorial-apriori-algorithm-data-mining-r-implementation
[5] https://docs.orange.biolab.si/2/reference/rst/Orange.associate.html
[6] Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case, Springer, 2010.

[7] S,Sphulari, P.U.Bhulchadra, Dr.S.D.Khamitkar &S.N.Lokhande ,Understanding rule behavior through apriori algorithm over social network data , Globle journal of computer science and technology software &data engineering,2012.
[8] Xiao Cui, Hao Shi, Xun Yi, Application of association rule mining theory in SinaWeibo, Journal of computer and communication,2014.
[9] Paula R.C.Silva, Wladmir C,Brandao, Mining Professional profile from LinkedIn Using Association rules, The 7th international conference on Information, Process, and knowledge management,2015.
[10] P. Nancy, R. Geetha Ramani &Shomona Gracia Jacob, Mining of Association Patterns in Social Network Data (Face Book 100 Universities) through Data Mining Techniques and Methods,Springer,2013.
[11] Gayana Femanda & Md Gapar MdJohar, Framework for social network data mining, International Journal of computer application, April,2015.
[12] R.Sathya,A.Aruna devi, S.Divya, Data mining and analysis of online social network, International Journal of Data mining Techniques and applications, June,2015.
[13] Muhammad Mahbubur Rahman, .Mining social data to extract Intellectual Knowledge, I.J, Intelligent System and application, 2012.
[14] Paresh Tanna, Dr.Yogesh Ghodasara, Using Apriori with WEKA for frequent pattern mining, International Journal of Engineering Trends Technology (IjETT), June,2014.
[15] D.Magdalene Delighta Angeline, I.Samuel Peter James, Efficient aprior mend algorithm for pattern extraction process, International Journal of Computer Science and InformationTechnologies,2011.
[16] Nergis Y1, limaz ans Gulfem1 and Klar Alptekin, The Effect of Clustering in the Apriori Data Mining algorithm: A case study, Proceedings of the World Congress on Engineering , July,2013.
[17] Nancy.P,Dr.R. Geetha Ramani, Discovery of classifiaction and rules in prediction of application usage in social network data (Facebook application data) using application algorithms, International journal in human machine Interaction (IJHMI),June,2014.
[18] Vipulm Mangla, Chandni Sarda,Sarthal Madra, VIT University, Vellore(842014), Improving the efficiency of Apriori Algorithm, International Journal of Engineering and Innovative Technology(IJEIT), September,2013.