Apriori Algorithm

Contents

- **Introduction**
- **What is the Apriori Algorithm?**
- **Apriori algorithm – The Theory**
- **Support**
- **Confidence**
- **Lift**
- **How does the Apriori Algorithm in Data Mining work?**
- **Pros, Cons and Limitations**
- **How to Improve the Efficiency of the Apriori Algorithm?**
- **References**

Introduction to APRIORI

- Apriori is the most famous frequent pattern mining method. It scans dataset repeatedly and generate item sets by bottom-top approach.
- Apriori algorithm is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties

Contd…

- This small story will help you understand the concept better. You must have noticed that the local vegetable seller always bundles onions and potatoes together. He even offers a discount to people who buy these bundles.
- Why does he do so? He realizes that people who buy potatoes also buy onions. Therefore, by bunching them together, he makes it easy for the customers. At the same time, he also increases his sales performance. It also allows him to offer discounts.
- Similarly, you go to a supermarket, and you will find bread, butter, and jam bundled together. It is evident that the idea is to make it comfortable for the customer to buy these three food items in the same place.
- The Walmart beer diaper parable is another example of this phenomenon. People who buy diapers tend to buy beer as well. The logic is that raising kids is a stressful job. People take beer to relieve stress. Walmart saw a spurt in the sale of both diapers and beer.
- These three examples listed above are perfect examples of Association Rules in Data Mining. It helps us understand the concept of apriori algorithms.

What is the Apriori Algorithm?

- Apriori algorithm, a classic algorithm, is useful in mining frequent itemsets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket.
- It helps the customers buy their items with ease, and enhances the sales performance of the departmental store.
- This algorithm has utility in the field of healthcare as it can help in detecting adverse drug reactions (ADR) by producing association rules to indicate the combination of medications and patient characteristics that could lead to ADRs.

Apriori algorithm – The Theory

- Three significant components comprise the apriori algorithm. They are as follows.
- Support
- Confidence
- Lift
- This example will make things easy to understand.
- As mentioned earlier, you need a big database. Let us suppose you have 2000 customer transactions in a supermarket. You have to find the Support, Confidence, and Lift for two items, say bread and jam. It is because people frequently bundle these two items together.
- Out of the 2000 transactions, 200 contain jam whereas 300 contain bread. These 300 transactions include a 100 that includes bread as well as jam. Using this data, we shall find out the support, confidence, and lift.

Support

- Support is the default popularity of any item. You calculate the Support as a quotient of the division of the number of transactions containing that item by the total number of transactions. Hence, in our example,
- Support (Jam) = (Transactions involving jam) / (Total Transactions)

= 200/2000 = 10%

Confidence

- Confidence is the likelihood that customers bought both bread and jam. Dividing the number of transactions that include both bread and jam by the total number of transactions will give the Confidence figure.
- Confidence = (Transactions involving both bread and jam) / (Total Transactions involving jam)

  = 100/200 = 50%
- It implies that 50% of customers who bought jam bought bread as well.

Lift

- Lift is the increase in the ratio of the sale of bread when you sell jam. The mathematical formula of Lift is as follows.
- Lift = (Confidence (Ja$\underset{\longrightarrow}{\text{m –}}$ Bread)) / (Support (Jam))

  = 50 / 10 = 5
- It says that the likelihood of a customer buying both jam and bread together is 5 times more than the chance of purchasing jam alone. If the Lift value is less than 1, it entails that the customers are unlikely to buy both the items together. *Greater the value, the better is the combination.*

How does the Apriori Algorithm in Data Mining work?

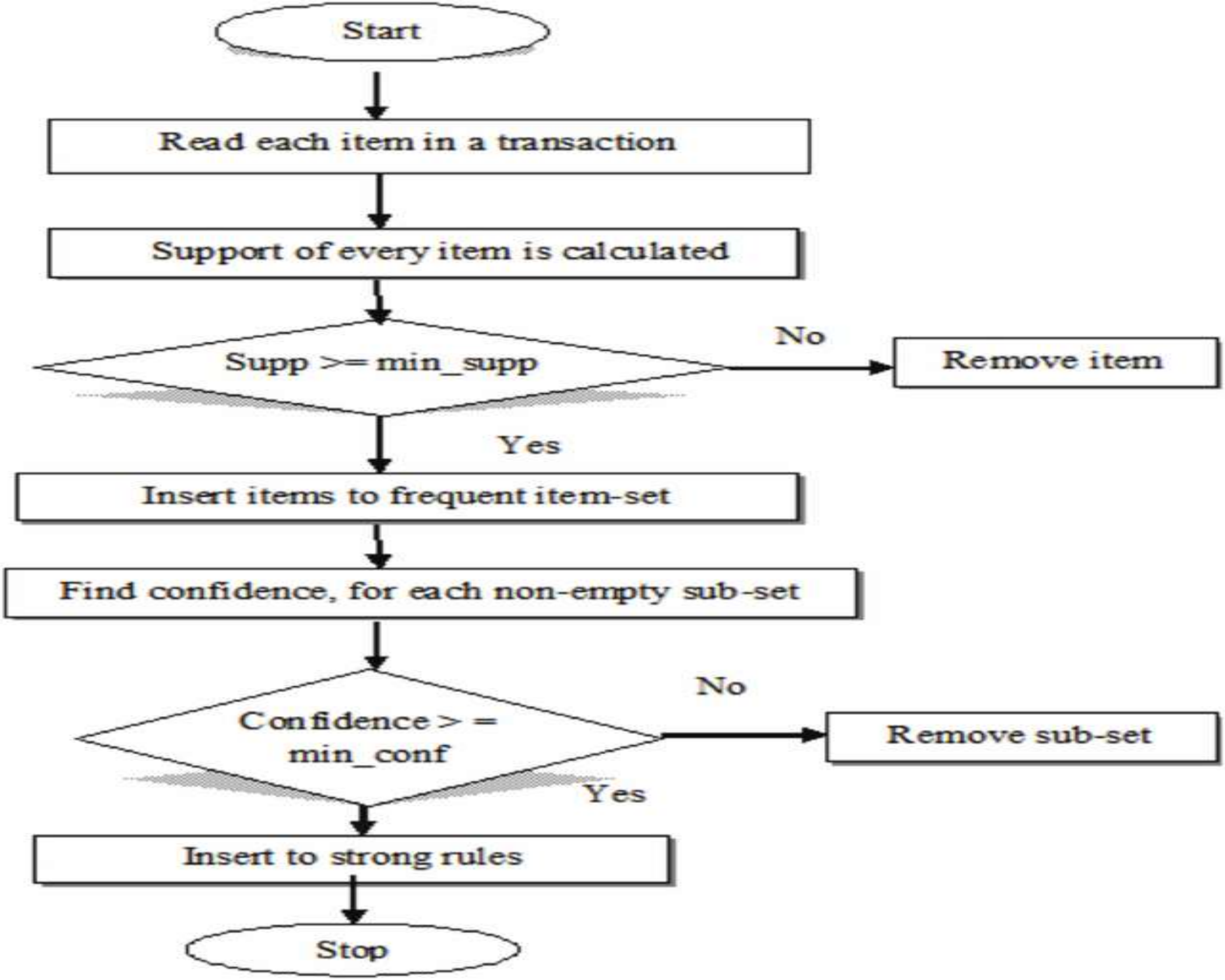
- We shall explain this algorithm with a simple example.
- Consider a supermarket scenario where the itemset is I = {Onion, Burger, Potato, Milk, Juice}. The database consists of six transactions where 1 represents the presence of the item and 0 the absence.

Contd...

Transaction ID	Onion	Potato	Burger	Milk	Juice
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	0

The Apriori Algorithm makes the following assumptions

- All subsets of a frequent itemset should be frequent.
- In the same way, the subsets of an infrequent itemset should be infrequent.
- Set a threshold support level. In our case, we shall fix it at 50%


Start
Read each item in a transaction
Support of every item is calculated
Supp >= min_supp
No
Remove item
Yes
Insert items to frequent item-set
Find confidence, for each non-empty sub-set
Confidence > = min_conf
No
Remove sub-set
Yes
Insert to strong rules
Stop

Steps

- Create a frequency table of all the items that occur in all the transactions. Now, prune the frequency table to include only those items having a threshold support level over 50%.
- We arrive at this frequency table.

Threshold Support >= 50%

Step 1 : One Itemset

C1

Itemset	Frequency
Onion	4
Potato	5
Burger	4
Milk	4
Juice	2

Pruning

Step 2

F1

Itemset	Frequency
Onion	4
Potato	5
Burger	4
Milk	4

Make pairs of items such as OP, OB, OM, PB, PM, BM. This frequency table is what you arrive at.

Step 3 : Two Itemset C2

Itemset	Frequency
Onion Potato	4
Onion Burger	3
Onion Milk	2
Potato Burger	4
Potato Milk	3
Burger Milk	2

Pruning

Apply the same threshold support of 50% and consider the items that exceed 50% (in this case 3 and above).
Thus, you are left with OP, OB, PB, and PM

Step 4 F2

Itemset	Frequency
Onion Potato	4
Onion Burger	3
Potato Burger	4
Potato Milk	3

Note: Consider items from F1 only for making C2

Look for a set of three items that the customers buy together. Thus we get this combination.

OP and OB gives OPB

PB and PM gives PBM

Step 6 : Three Itemset C3

Itemset	Frequency
Onion Potato Burger	3
Potato Burger Milk	2

Pruning →

Determine the frequency of these two itemsets. You get this frequency table.

If you apply the threshold assumption, you can deduce that the set of three items frequently purchased by the customers is OPB.

We have taken a simple example to explain the apriori algorithm in data mining. In reality, you have hundreds and thousands of such combinations.

Step 7 F3

Itemset	Frequency
Onion Potato Burger	3

Note: Consider items from F2 only for making C3

Subset Creation

I = (O P B)

S = (OP) (OB) (PB) (O) (P) (B)

For every subset S of I, Output the rule:

S ⟶ (I – S) (S recommends I-S)

If Support(I)/ Support(S) >= min_conf value

Apriori Algorithm – Pros

- Easy to understand and implement
- Can use on large itemsets

Apriori Algorithm – Cons

- At times, you need a large number of candidate rules. It can become computationally expensive.
- It is also an expensive method to calculate support because the calculation has to go through the entire database.

Apriori Algorithm – Limitations

- The process can sometimes be very tedious.

How to Improve the Efficiency of the Apriori Algorithm?

Use the following methods to improve the efficiency of the apriori algorithm.

- **Transaction Reduction –** A transaction not containing any frequent k-itemset becomes useless in subsequent scans.

- **Hash-based Itemset Counting –** Exclude the k-itemset whose corresponding hashing bucket count is less than the threshold is an infrequent itemset.
- There are other methods as well such as partitioning, sampling, and dynamic itemset counting.

References:

- Apriori Algorithms and Their Importance in Data Mining (digitalvidya.com)
- Flow chart of Apriori-algorithm | Download Scientific Diagram (researchgate.net)
- What is the Apriori algorithm? (educative.io)
- Apriori Algorithm – GeeksforGeeks
- Niu, K., Jiao, H., Gao, Z., Chen, C., & Zhang, H. (2017). *A developed apriori algorithm based on frequent matrix. Proceedings of the 5th International Conference on Bioinformatics and Computational Biology - ICBCB '17.* doi:10.1145/3035012.3035019
- Fast Algorithms for Mining Association Rules (vldb.org)
- Apriori Association Rules | Grocery Store | Kaggle
- Implementing Apriori algorithm in Python - GeeksforGeeks