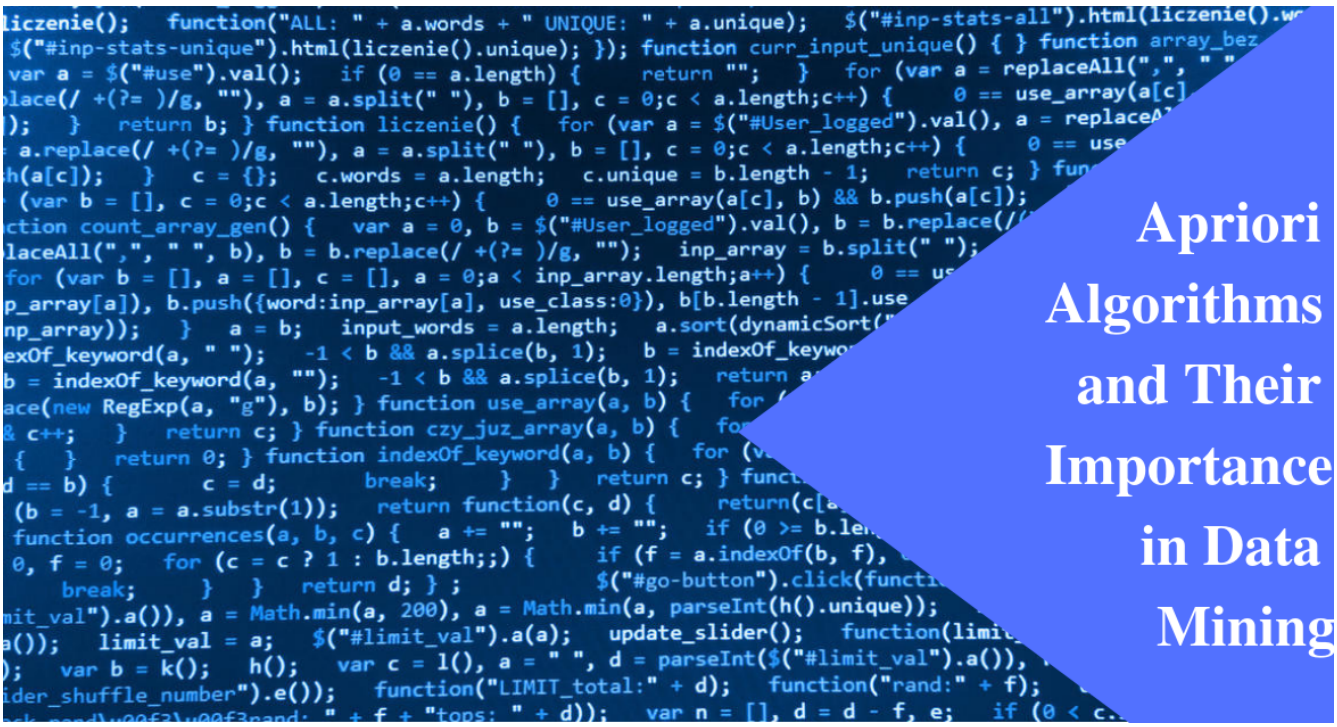

INSAID



Apriori Algorithms And Their Importance In Data Mining

When you talk of data mining, the discussion would not be complete without the mentioning of the term, 'Apriori Algorithm.'

This algorithm, introduced by R Agrawal and R Srikant in 1994 has great significance in data mining. We shall see the importance of the apriori algorithm in data mining in this article.

An Introduction to Apriori Algorithms

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 realises 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 – An Odd Name

It has got this odd name because it uses 'prior' knowledge of frequent itemset properties. The credit for introducing this algorithm goes to Rakesh Agrawal and Ramakrishnan Srikantin 1994. We shall now explore the apriori algorithm implementation in detail.

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

In our example, Confidence is the likelihood that customer 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.

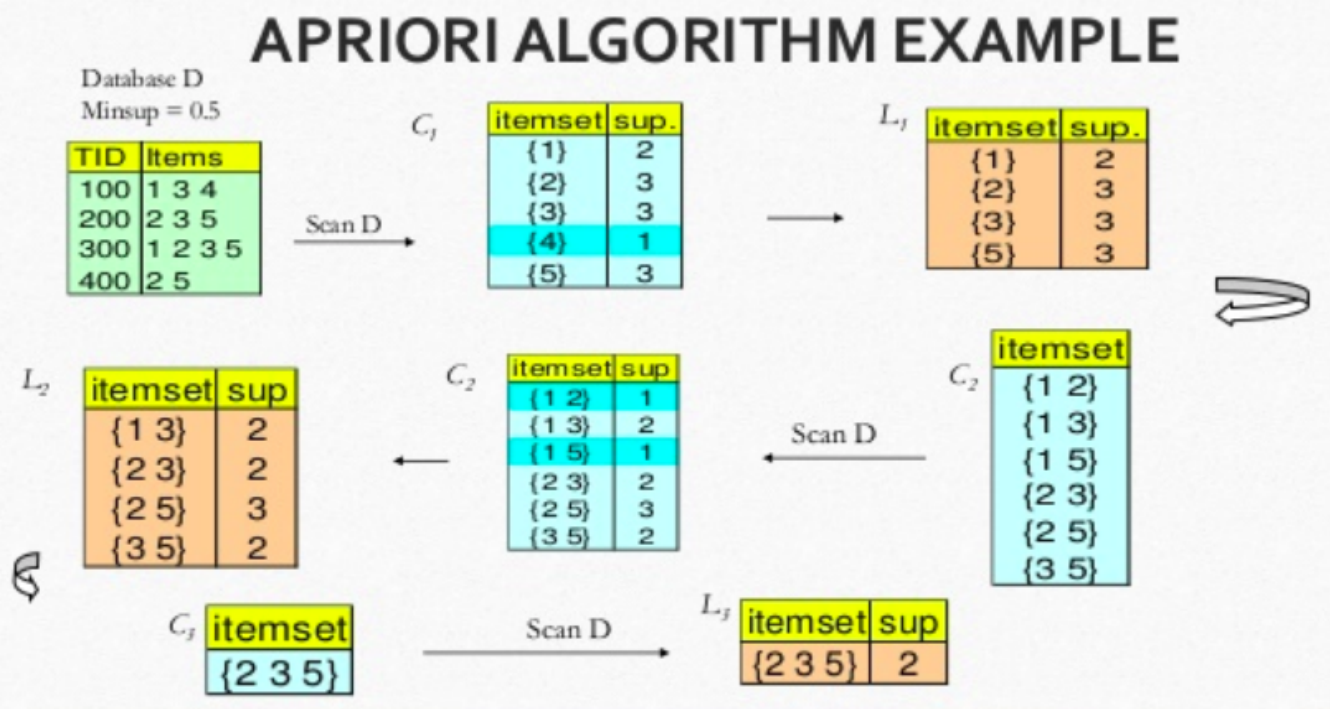
Lift

According to our example, 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 (Jam_– 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.



Apriori Algorithm Example

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, Beer}. The database consists of six transactions where 1 represents the presence of the item and 0 the absence.

Simple Apriori Algorithm Example

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%

Step 1

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.

Simple Apriori Algorithm Example

This table signifies the items frequently bought by the customers.

Step 2

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

Simple Apriori Algorithm Example

Step 3

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

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 5

Simple Apriori Algorithm Example

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.

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.

Apriori Algorithm in Python

This video will help you understand the importance of the apriori algorithm in python.

https://youtu.be/-mg2rMqmyNA

Apriori Algorithm in Data Mining

We have seen an example of the apriori algorithm concerning frequent itemset generation. There are many uses of apriori algorithm in data mining. One such use is finding association rules efficiently.

The primary requirements for finding association rules are,

- Find all rules having the Support value more than the threshold Support
- And Confidence values greater than the threshold confidence

There are two ways of finding these rules.

- **Use Brute Force –** List out all the rules and determine the support and confidence levels for each rule. The next step is to eliminate the values below the threshold support and confidence levels. It is a tedious exercise.
- **The Two-Step Approach –** This method is a better one as compared to the Brute Force method.

Step 1

We have seen earlier in this article how to prepare the frequency table and find itemsets having support greater than the threshold support.

Step 2

Use binary partition of the frequent itemsets to create rules. You have to look for the ones having the highest confidence levels. You also refer to it as the candidate rules.

In our example, we found out that the OPB combination was the frequent itemset. Apply Step 2 and find out all the rules using OPB.

OP—B, OB—P, PB—O, B—OP, P—OB, O—PB

You find that there are six combinations. Therefore, if you have k elements, there will be 2k – 2 candidate association rules.

Final Words

In this age of eCommerce retail shops and globalisation, it becomes imperative for businesses to use machine learning and artificial intelligence to stay ahead of the competition. Hence, data analysis has a great scope in today's environment. The use of apriori algorithms has great value in data analysis.

We have seen how you can deduce various kinds of data and enhance the sales performance of the supermarket. That was one example of the utility of the apriori algorithm.

This concept has been used in other critical industries like the healthcare industries, and so on. It enables the industry to bundle drugs that cause the least ADRs depending on the patient's characteristics. Read Digital Vidya Blogs to know more about machine learning aspects.