What is Association Rule Mining and What are its Applications

Edris Safari

Bellevue University, Nebraska U.S.A.

Association rule in machine learning is an if-then statement between the features in a dataset. These if-then statements or rules, give the most accurate indication that the variables in the if-then statement (i.e. if Milk then Bread) are related. This is evident in the online shopping experience when shoppers are shown suggested items at the check out or even during search of a product.

Market Basket Analysis is a popular description and indeed an application of Association rule mining. It is often called association rule mining, itemset mining, or frequency itemset mining. Item sets are set of items in say a transaction such as “if milk then bread” or a set of symptoms such as “if high blood pressure and headache and cholesterol and smoker then heart attack”. In general, association rules show probability of relationship between the items in a set of rules. In finding those rules, we must search and make sense of many combinations of the items in the dataset.

It immediately becomes evident the number of combinations of items in the dataset that the algorithms must make to evaluate all combinations and produce meaningful rules. The larger the number of items in the data set, the more the number of association rule. For this reason, any algorithm that processes these datasets must be able to filter the data so that valid or pertinent rules are identified. For example, the algorithm should throw out milk and a five gallon can of white paint and keep milk and cookies. How would an algorithm do this? Well it painstakingly goes through all the transactions and then by association, it finds that milk and cookies have met the proper criteria to be included and milk and paint did not because they simply did not appear together or very seldom.

The criteria that the rules must meet to be included is a fuzzy area and is use case specific. The algorithms that are used in association rule mining produce 3 measures between the items in every rule. The analysts then decide on the threshold value for these measures. These threshold s are completely use-case specific. The measures are Support, Confidence and lift as described below:

- Support shows the probability of A and B occurring together among the entire dataset or total number of transactions. For example, in a dataset with 50 transactions, if Milk and Bread have been found 6 times, then support for the rule if milk then bread is 6/50=12%.

- Confidence is probability of A and B occurring together among those transactions where A occurred. For example if Milk and Bread appeared together 6 times but milk also appeared with two other items (paint, and butter) then confidence is 6/(6+2)=75%.

- Lift is the ratio of confidence to expected confidence (Ref 3). Expected confidence is the confidence divided by frequency of B which is the number of times B has appeared without A divided by total number of records or observations. The lift tells us the how much better a rule is at predicting the result than just assuming the result in the first place. The greater the value of lift, the stronger the association.

The association rules will each have these values, and it is up to the analysts to decide which ones to keep or consider. Just going through the dataset and indiscriminately creating rules and calculating their support, confidence and lift value could be time consuming and resource intensive. Algorithms must accommodate for that. For example, the Apriori algorithm performs the following steps to speed up the process:

1. Set a minimum value for support and confidence. This means that we are only interested in finding rules for the items that have certain default existence (e.g. support) and have a minimum value for co-occurrence with other items (e.g. confidence).
2. Extract all the subsets having higher value of support than minimum threshold.
3. Select all the rules from the subsets with confidence value higher than minimum threshold.
4. Order the rules by descending order of Lift.

The python 'apyori' package provides the function of creating rules based on the steps above. In the code snippet below, the min values are the thresholds to maintain above those values. The min and max length parameters specify the minimum and maximum number of items in the rule.

```
from apyori import apriori
rules = apriori(transactions = transactions, min_support = 0.003, min_confidence =
0.2, min_lift = 3, min_length = 2, max_length = 3)
```

The number of rules generated are all governed by the min /max length. In the example above, min length of 2 and max length of 2 generates 9 rules and changing the max to 4 generates 132 rules and max of 6 generates 154 rules. The analysts must consider the use case to fine-tune these parameters to get the best rules identified.

Association rule mining is prevalent in areas where the variables are categorical, and the goal is to find patterns that have high confidence of repeating. The application of it in the medical profession in diagnosing and even predicting diseases is noteworthy. We see it being applied in our everyday shopping experiences be it online or at the grocery store. Now I know why(and how) I get those coupons at the check out and why the items in the coupons are in my shopping basket or have been sometime in the past.

References

1. Abbott, Dean. **Applied Predictive Analytics**. Wiley. Kindle Edition.

2. Association Rule Mining - https://towardsdatascience.com/association-rule-mining-be4122fc1793

3. Building a market basket model - https://infocenter.informationbuilders.com/wf80/index.jsp?topic=%2Fpubdocs%2FRStat16%2Fsource%2Ftopic49.htm

4. Association rules – https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining