Knowledge Mining and Big Data Part 4

Pedram Ghazi

Student Number: 267640

1) An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. 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.

2) Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

3) Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It moves forward by identifying each of frequent items in the database and extending them to larger and larger item sets till these item sets be present sufficiently often in the database.

4) Index 13 best rules found:

```
1. milk-cream=t fruit=t 2038 ==> bread and cake=t 1684    conf:(0.83)
2. milk-cream=t vegetables=t 2025 ==> bread and cake=t 1658    conf:(0.82)
3. fruit=t vegetables=t 2207 ==> bread and cake=t 1791    conf:(0.81)
4. margarine=t 2288 ==> bread and cake=t 1831    conf:(0.8)
5. biscuits=t 2605 ==> bread and cake=t 2083    conf:(0.8)
6. milk-cream=t 2939 ==> bread and cake=t 2337    conf:(0.8)
7. tissues-paper prd=t 2247 ==> bread and cake=t 1776    conf:(0.79)
8. fruit=t 2962 ==> bread and cake=t 2325    conf:(0.78)
9. baking needs=t 2795 ==> bread and cake=t 2191    conf:(0.78)
10. frozen foods=t 2717 ==> bread and cake=t 2129    conf:(0.78)
```

Index 61 best rules found:
1. bread and cake=t baking needs=t fruit=t 1564 ==> milk-cream=t 1161   conf:(0.74)
2. bread and cake=t baking needs=t vegetables=t 1586 ==> milk-cream=t 1169   conf:(0.74)
3. bread and cake=t fruit=t vegetables=t 1791 ==> milk-cream=t 1311   conf:(0.73)
4. total=high 1679 ==> milk-cream=t 1217   conf:(0.72)
5. bread and cake=t fruit=t 2325 ==> milk-cream=t 1684   conf:(0.72)
6. bread and cake=t vegetables=t 2298 ==> milk-cream=t 1658   conf:(0.72)
7. bread and cake=t baking needs=t 2191 ==> milk-cream=t 1580   conf:(0.72)
8. baking needs=t fruit=t 1900 ==> milk-cream=t 1365   conf:(0.72)
9. bread and cake=t tissues-paper prd=t 1776 ==> milk-cream=t 1275   conf:(0.72)
10. baking needs=t vegetables=t 1949 ==> milk-cream=t 1392   conf:(0.71)

Index 17 best rules found:
1. total=high 1679 ==> tea=t 475   conf:(0.28)
2. baking needs=t margarine=t 1645 ==> tea=t 465   conf:(0.28)
3. baking needs=t biscuits=t 1764 ==> tea=t 476   conf:(0.27)
4. bread and cake=t tissues-paper prd=t 1776 ==> tea=t 468   conf:(0.26)
5. bread and cake=t margarine=t 1831 ==> tea=t 479   conf:(0.26)
6. biscuits=t frozen foods=t 1810 ==> tea=t 472   conf:(0.26)
7. baking needs=t frozen foods=t 1835 ==> tea=t 473   conf:(0.26)
8. tissues-paper prd=t 2247 ==> tea=t 575   conf:(0.26)
9. margarine=t 2288 ==> tea=t 584   conf:(0.26)
10. biscuits=t fruit=t 1837 ==> tea=t 463   conf:(0.25)

Confidence value is in fact number of supporting true values of the one classIndex divided by total number of matches for a defined rule for that classIndex. For example, for the first rule of classIndex 61, confidence = 1161/1564. For classIndex 17, at first there were no rules found so I decreased the minMetric to 0.1 and after this change there were some rules found with low confidences.

5) From the findings that we had, for example in classIndex 13, we can say 83% of people who bought “milk-cream” and “fruit” also have purchase “bread and cake” or in classIndex 61, we can say 74% of people who bought “bread and cake”, “baking needs”, and “fruit” have also purchased “milk-cream”. We can discover more shopping habits by studying more rules. We as a shop owner, can make packages of products in one of these rules so we can sell these products at once.