TDS3301 DATA MINING ASSIGNMENT

Part 2 Association Rule Mining

A)

The aim of an association analysis is to find association rules. Association rules are similar to the classification rules, except that the prediction is not limited to the goal attribute. It also searches correlations between any attributes.

Association Rule Mining is part of the supervised learning. One of the most used cases is the shopping cart analysis. There you want to find frequent patterns.

IF a person buys bread THEN she will buy ….

Frequent pattern: a pattern (a set of items, sub sequences, substructures, etc.) that occurs frequently in a data set.

- Motivation: Finding inherent regularities in data
- What products were often purchased together?—Beer and diapers?!
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we automatically classify web documents?
- Applications
- Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

What improve because of the use of pattern?

With the Association Rule Mining, you can make correlations between different goods. You can recognise and analyse the customer behaviour. The association analysis is a predictive data mining process. It analyses the data to ensure regularities and to predict the behaviour of new records.

When you use the Basket data analysis, you can solve following questions:

- How do I arrange my goods optimally?
- In which categories can I classify my Customers?
- Which items should be deleted from the assortment?

The result of solving these questions is to maximize **the profit.**

B)

For this portion of the assignment, you will be using the Extended Bakery dataset, which describes transactions from a chain of bakery shops that sell a variety of drinks and baked goods.

This particular dataset is made up of smaller subsets that contain 1,000, 5,000, 20,000 and 75,000 transactions each. Furthermore, each of these subsets is given in two different formats.

Sparse Vector Format: XXX-out1.csv files. Each line of the file has the following format:

• First column: transaction ID
• Subsequent columns: list of purchased goods (represented by their ID code)

Example: 3,0,2,4,6 Transaction 3 contained items 0, 2, 4, 6

Full Binary Vector Format: XXX-out2.csv files. Each line has the following format:

• First column: transaction ID
• 50 subsequent columns: A sequence of binary values that indicate if item i (represented in column i) has been purchased in that transaction.

Example:

3,0,0,1,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Transaction 3 contained items 0, 2, 4, 6

Pre-processing Tasks:

- Deleting of wrong values who are out of the range (51,52,….)
- Deleting NA values
- Transform the Data set in a Transaction set
- In this case, we can ignore the first attribute that just contains the number of the transaction.

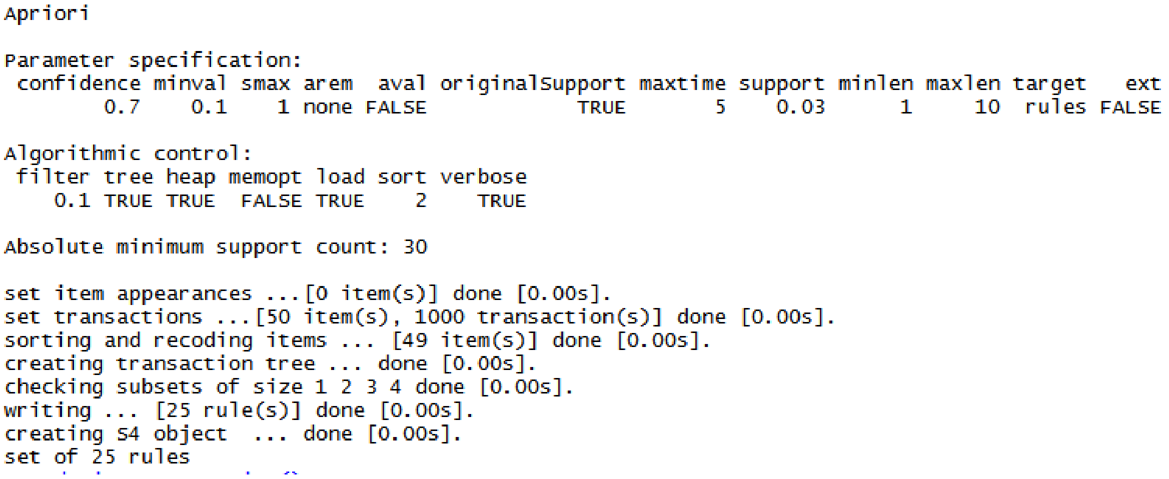
C)

Algorithm: Apriori

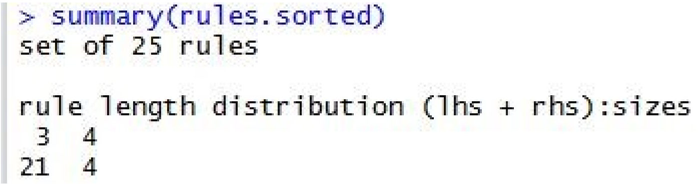
Parameters: support=0.03, confidence=0.7

Time Required: 0.01318192 secs

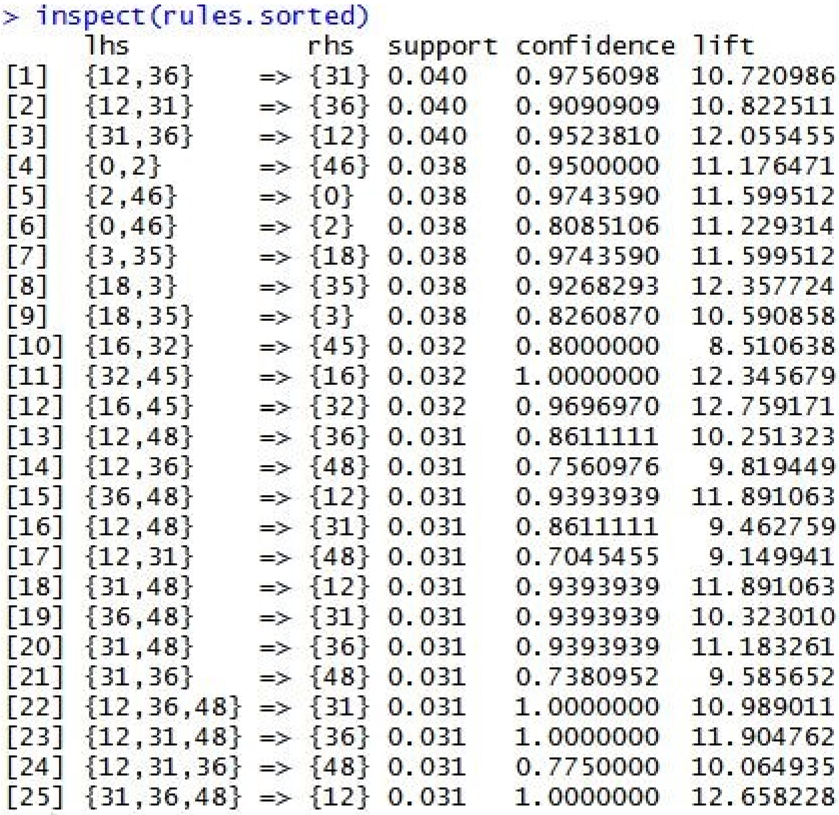
D)



```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.7    0.1    1 none FALSE            TRUE       5    0.03      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 30

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[50 item(s), 1000 transaction(s)] done [0.00s].
sorting and recoding items ... [49 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [25 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
set of 25 rules
```

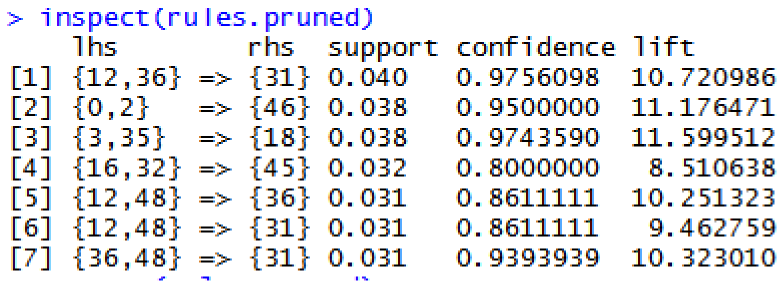


```
> summary(rules.sorted)
set of 25 rules

rule length distribution (lhs + rhs):sizes
 3  4 
21  4 
```



```
> inspect(rules.sorted)
     lhs             rhs  support confidence lift     
[1]  {12,36}      => {31} 0.040   0.9756098  10.720986
[2]  {12,31}      => {36} 0.040   0.9090909  10.822511
[3]  {31,36}      => {12} 0.040   0.9523810  12.055455
[4]  {0,2}        => {46} 0.038   0.9500000  11.176471
[5]  {2,46}       => {0}  0.038   0.9743590  11.599512
[6]  {0,46}       => {2}  0.038   0.8085106  11.229314
[7]  {3,35}       => {18} 0.038   0.9743590  11.599512
[8]  {18,3}       => {35} 0.038   0.9268293  12.357724
[9]  {18,35}      => {3}  0.038   0.8260870  10.590858
[10] {16,32}      => {45} 0.032   0.8000000   8.510638
[11] {32,45}      => {16} 0.032   1.0000000  12.345679
[12] {16,45}      => {32} 0.032   0.9696970  12.759171
[13] {12,48}      => {36} 0.031   0.8611111  10.251323
[14] {12,36}      => {48} 0.031   0.7560976   9.819449
[15] {36,48}      => {12} 0.031   0.9393939  11.891063
[16] {12,48}      => {31} 0.031   0.8611111   9.462759
[17] {12,31}      => {48} 0.031   0.7045455   9.149941
[18] {31,48}      => {12} 0.031   0.9393939  11.891063
[19] {36,48}      => {31} 0.031   0.9393939  10.323010
[20] {31,48}      => {36} 0.031   0.9393939  11.183261
[21] {31,36}      => {48} 0.031   0.7380952   9.585652
[22] {12,36,48} => {31} 0.031   1.0000000  10.989011
[23] {12,31,48} => {36} 0.031   1.0000000  11.904762
[24] {12,31,36} => {48} 0.031   0.7750000  10.064935
[25] {31,36,48} => {12} 0.031   1.0000000  12.658228
```

I would show these seven rules to the client. This is after removing of redundant rules. They have the highest support.



```
> inspect(rules.pruned)
    lhs        rhs  support confidence lift
[1] {12,36} => {31} 0.040   0.9756098  10.720986
[2] {0,2}   => {46} 0.038   0.9500000  11.176471
[3] {3,35}  => {18} 0.038   0.9743590  11.599512
[4] {16,32} => {45} 0.032   0.8000000   8.510638
[5] {12,48} => {36} 0.031   0.8611111  10.251323
[6] {12,48} => {31} 0.031   0.8611111   9.462759
[7] {36,48} => {31} 0.031   0.9393939  10.323010
```

E)

Appel Tart, Apple Danish → Apple Croissant when we see the other rules it seems like the people like Apple.

- They could make special offers for the Appel Croissant
- Put all the products with apple in a special shelf.

For the 2 Rule

Chocolate cake, Casino Cake → Chocolate Coffee

- Make a “Menu” deal

Blueberry Tart, Apricot Croissant →Hot Coffee. Seems like a Breakfast

- Make a breakfast Deal
- Fast Coffee To Go so the customer don’t need to wait → Better Customer behavior

Sources

http://www3.nd.edu/~rjohns15/cse40647.sp14/www/content/assignments/assignment2.pdf

https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery

http://www.nyu.edu/classes/jcf/g22.3033-002/slides/session6/MiningFrequentPatternsAssociationAndCorrelations.pdf