|  |
| --- |
|  |
| TDS3301 Data Mining |
| Report |
|  |
|  |
| **13-Jan-17** |

**Group Member :**

Tan Chong Raen 1121115725

Mah Siew Chin 1132702455

Oh Soon Kit 1132702874

Benjamin Au Wen Wei 1141327105

Table of Contents

[Objective 1](#_Toc472116012)

[What is the domain and what are the potential benefits to be derived from association rule mining. This is high level - not find patterns, but what would improve because of the use of the patterns. 1](#_Toc472116013)

[Data Description Rule 2](#_Toc472116014)

[What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices. 2](#_Toc472116015)

[Rule Mining Process 2](#_Toc472116016)

[Parameter settings, choice of algorithm, and the time required. 2](#_Toc472116017)

[Resulting Rules 3](#_Toc472116018)

[Summary (number of rules, general description), and a selection of those you would show to a client. 3](#_Toc472116019)

[Recommendation 7](#_Toc472116020)

[What should the client do because of the rules discovered? 7](#_Toc472116021)

# Objective

## What is the domain and what are the potential benefits to be derived from association rule mining. This is high level - not find patterns, but what would improve because of the use of the patterns.

Association rule mining algorithms focus on the discovery of valid rules by testing all

items or elements in the domain, instead doing testing only on some known elements, which making it inefficient because it creates a lot of candidates. The association rules mainly discovers significant and valid correlations among items belonging to a specific domain. Some researched works in the past showed that association rule mining was used for forecasting in classification problems. In our opinion, by adapting association rule in data mining field, it helps us in understanding how strong the relationship of one domain to another relates to real life.

# Data Description Rule

## What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices.

The Extended Bakery dataset about one year worth of sales information for a couple of small bakery shops. There are varieties in sold products such as chocolate cake, blueberry tart, cheese croissant, and etc. The training dataset used contains 1000 unprocessed data. For the pre-processing steps in this particular dataset, the first step done was to check duplication of data. Followed by checking missing values, replacing numbers to names of the product. The dataset is already in basket format with first column as an ordered unique identifier for each transaction, along with second column is the set of items bought in that transaction. Therefore all the items bought at the same time in one row which is needed in finding association rules, the dataset is read as transaction. The main reason in ignoring the columns or commas(,) is because due to the irrelevance while doing association ruling, whereby it is absolutely have the outcome of results not affected.

# Rule Mining Process

## Parameter settings, choice of algorithm, and the time required.

In the rule mining process, we set our minimum parameter settings as 0.01 for support, 0.5 for confidence, 2 for minimum length of rules and selected Aprior as our choice of algorithm on the transaction for the dataset “bakery”. For the time required, we continue in trying different parameter until we found interesting rule then only proceed to pruning process. The detail of the process steps are as stated below:

i) First, The number of rules has to go through interesting rule process.

i) Second, Viewing the total number or summary of rules output from modified “tran1rules” and

followed by printing the association rules involved.

ii) Third, View the sorted high chances rules item purchased by “lift” in order to find the high support and high confidence rule. In order to visualize the rules clearly, different categories graphs are plotted such as default, grouped, grouped, and paracoord.

iii) Fourth, A pruning redundant rule is implemented to eliminate those redundant rules. The pruning redundant rule begins with “which(redundant)” that is used to display redundant rules and “rules.pruned<- tran1rules[!redundant]” that is used to remove the redundant rules.

iv) Next, The rule is sorted according to confidence, support, and lift and followed by plotting graph for the removed redundant rules.

# Resulting Rules

## Summary (number of rules, general description), and a selection of those you would show to a client.

In general, there are remaining 28 rules left to be used after the elimination of redundancy rule from the overall of 124 rules and these are number of rules that will present to the client. Before starting the pruning process, number of rules have to go through interesting rule. The interesting rule is judged from insight why we decide this support and confidence value as an interesting rule. Figure below shows the display for interesting rule.

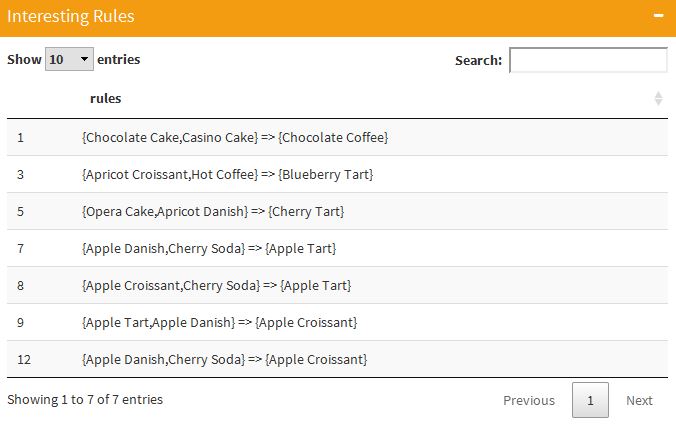


Figure 1: Interesting rule.

Figure below shows the number of rules of bakery dataset that undergo process of pruning.

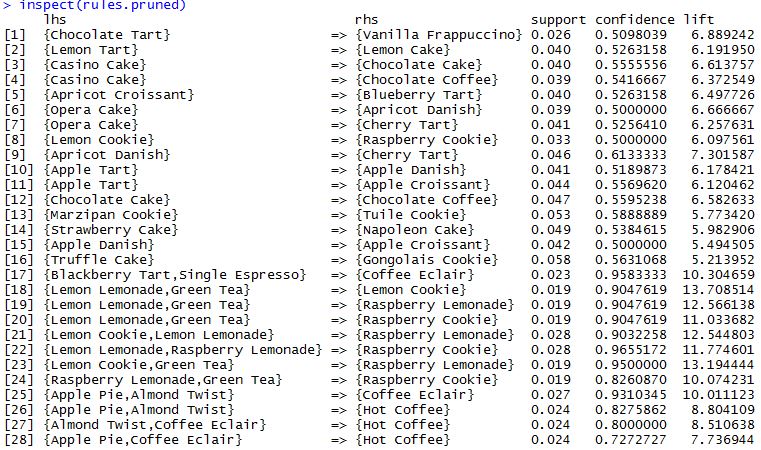


Figure 2: Number of rules of bakery dataset that is pruned

A description on the chart reading for instance rule number 1 shows LHS {Chocolate Tart} => {Vanilla Frappuccino} and RHS with a support of 0.026, a confidence level of 0.5098039 and a lift of 6.889242. In general, This implies that for customers who purchase chocolate tart might also buy vanilla frappucino.

Support indicates how popular an item set is as measured by the proportion of transactions in which an item set appears. Rule number 1 has a support of 0.026.

Confidence measure indicates how high the possibility an item B will be purchased after an item A is purchased. For instance the rule number 1 has a confidence level of 0.9655172 which it means that it is normal.

Lift indicates how high the possibility an item B will be purchased after item A is purchased with a supervision on the popularity of item B. Rule number 1 has a lift of 6.889242 which it means that it is normal.

All the graph’s plot type for the number of rules that is pruned as stated below:

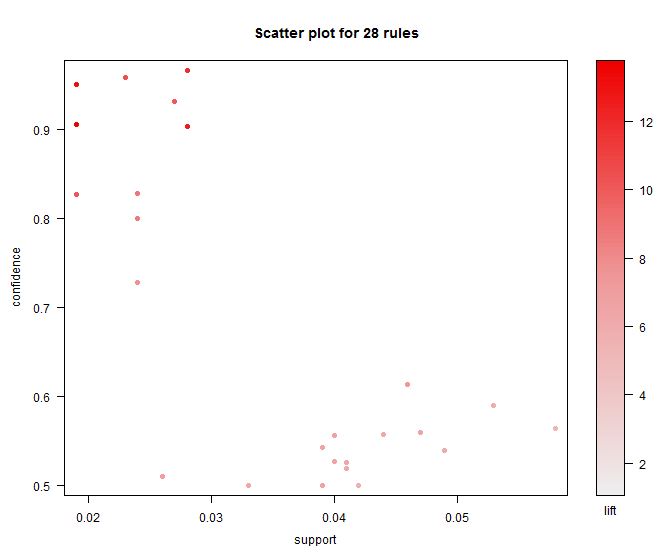
i) The first one is the default plot type which is known as scatter plot 

Figure 3: scatter plot for 28 rules

ii) The second one is the graph plot type which is known as network graph in which larger circles indicates higher support and red circles indicates higher lift.

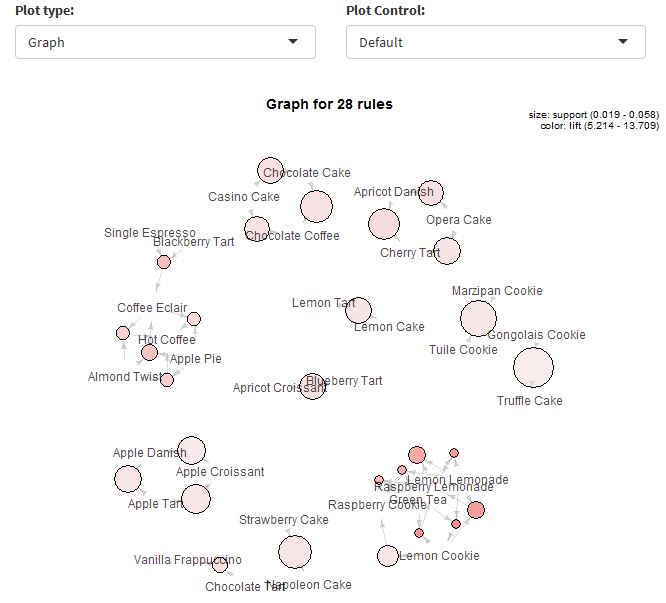


Figure 4: Graph for 28 rules

iii) The third one is the grouped plot type which is known as grouped matrix.

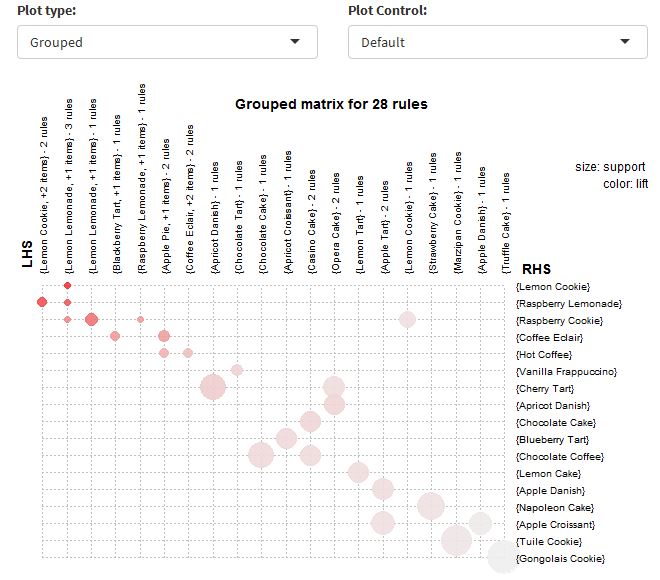


Figure 5: Graph matrix for 28 rules

iv) The fourth one is the paracoord plot type which is known as parallel coordinates plot.

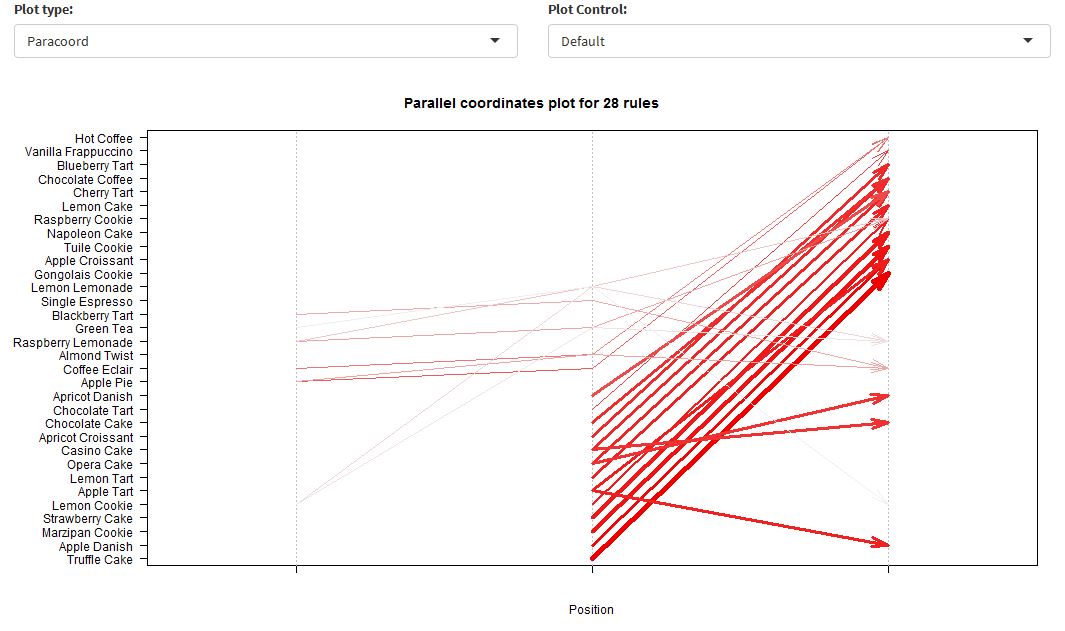


Figure 6: Parallel coordinates for 28 rules

# Recommendation

## What should the client do because of the rules discovered?

Based on the tabulated results, in our opinion the client should increase the items of blackberry tarts, single espresso, lemon cookie, raspberry lemonade, raspberry cookie, lemon lemonade, green tea, apple pie, lemon twist, coffee éclair as the prediction of buying one of them will increase the chance in buying other. Thus it would be a good move for them to focus in bringing up the sales in these specific products.