Association Rules

Greeshma Gopinathan (GG33287)

2024-08-18

Discovering hidden patterns in grocery purchases using association rule mining

We are going to use apriori algorithm to study customer's purchasing decisions and find some rules to apply to business decisions

```
library(tidyverse)
library(igraph)
library(arules)
library(arulesViz)
```

Loading the dataset as required by apriori the algorithm

```
groceries <- read.transactions('groceries.txt', format = 'basket', sep = ',')
```

Data exploration

```
summary(groceries)
inspect(groceries[1:5])  # View the first 5 transactions
```

Calculate item frequency

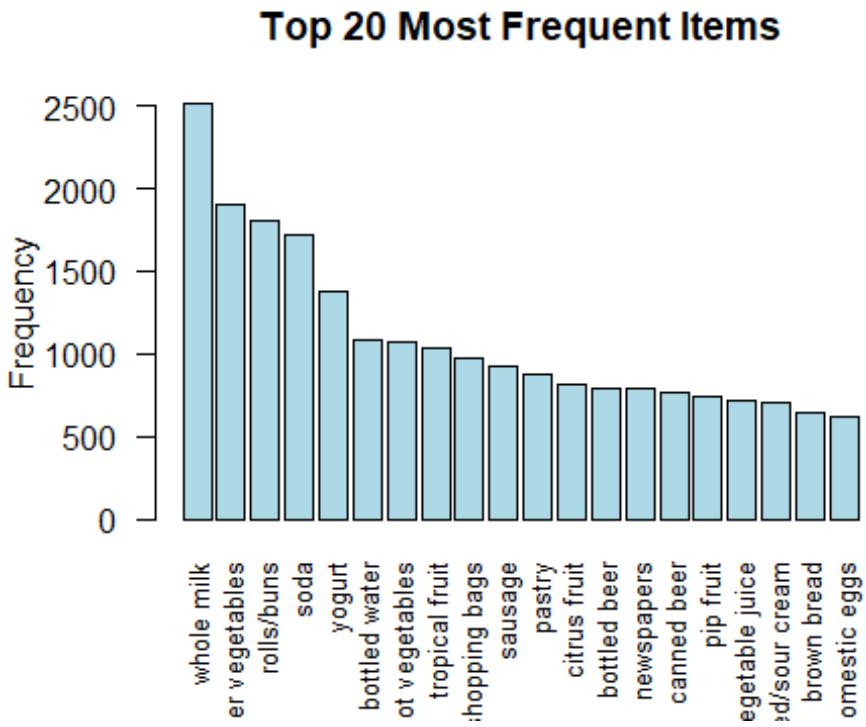
```
item_freq <- itemFrequency(groceries, type = "absolute")
```

Get the top 20 most frequent items

```
top_items <- sort(item_freq, decreasing = TRUE)[1:20]
```

Create a bar plot of the top 20 items

```
barplot(top_items, las = 2, cex.names = 0.8, col = "lightblue",
        main = "Top 20 Most Frequent Items", ylab = "Frequency")
```



The bar plot shows Whole milk and Other vegetables as the most frequent items in the dataset. Followed by rolls/buns, soda, and yogurt.

Apply the apriori algorithm

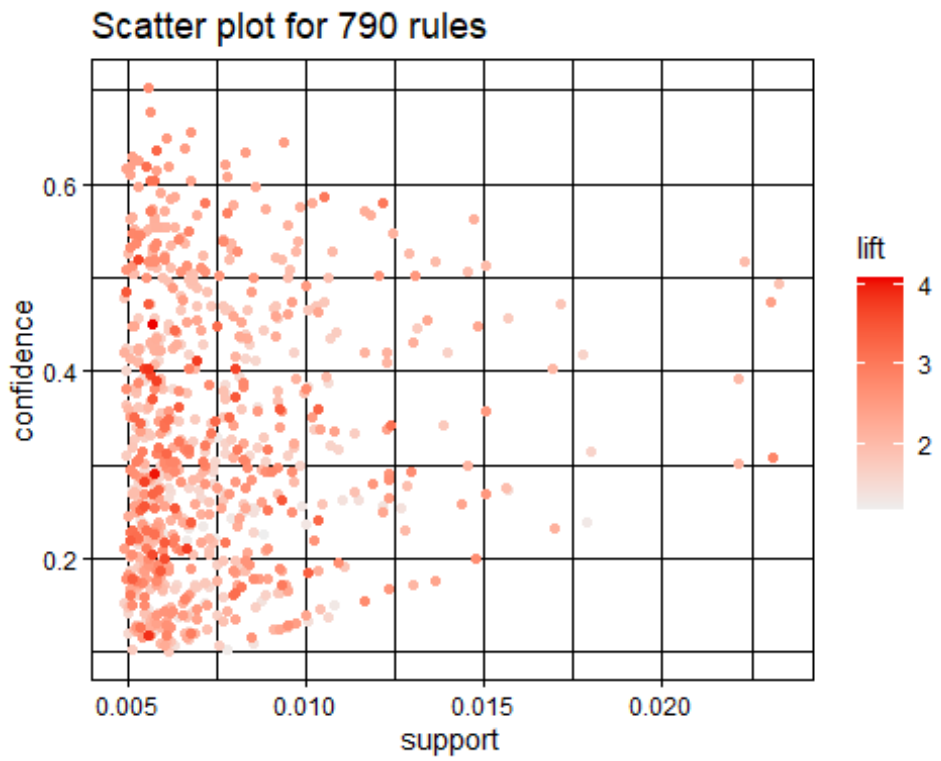
```
rules <- apriori(groceries,
                 parameter = list(supp = 0.005, conf = 0.1, minlen = 3))
```

Filter rules with a lift greater than 1.2

```
rules <- subset(rules, lift > 1.2)
inspect(rules)
```

plot all the rules in (support, confidence) space

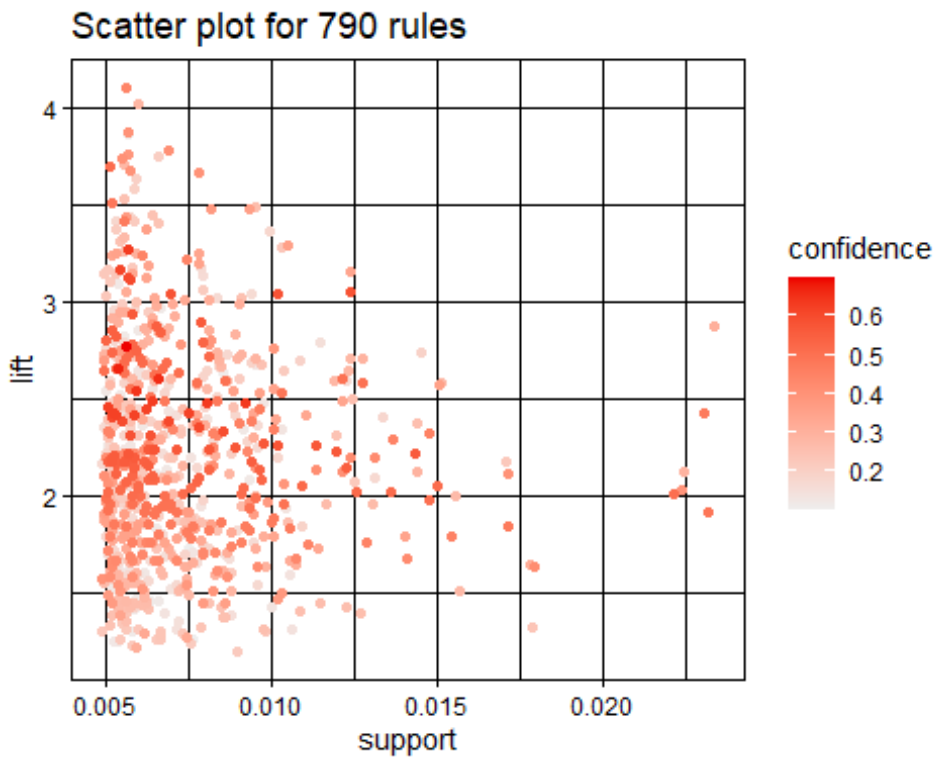
```
plot(rules)
```



- Most rules have low support (ranging from 0.005 to 0.015) but varying levels of confidence. This indicates that while these rules do not apply to a large portion of the dataset, they are still relatively reliable when they do occur
- The plot uses a color gradient to represent lift, with darker shades of red indicating higher lift. Rules with higher lift (darker red) are more significant as they suggest a stronger association between the items
- There are some rules with relatively high lift, making them potentially valuable for targeting or promotion, even if their support is low
- The rules with high lift (dark red) could reveal product combinations that are strongly associated and should be considered for bundling or cross-promotions

Can swap the axes and color scales

```
plot(rules, measure = c("support", "lift"), shading = "confidence")
```



- Focus on high-lift, high-confidence rules for bundling or special offers. For instance, if a rule has high confidence and lift, it suggests that customers who buy one item are likely to buy another, making it ideal for cross-promotion

"two key" plot: coloring is by size (order) of item set

```
plot(rules, method='two-key plot')
```

- The cyan points representing 4-item rules are generally clustered towards the lower end of the support axis. This makes sense because as more items are involved, the chance that all of them appear together in a transaction decreases, leading to lower support.
- Understanding that more complex rules (higher-order) are less frequent but sometimes highly confident can help prioritize where to apply these insights—perhaps in niche marketing strategies rather than widespread promotions.

Can now look at subsets driven by the plot

```
inspect(subset(rules, support > 0.005))
inspect(subset(rules, confidence > 0.5))
inspect(subset(rules, lift > 3))
```

Subset rules where support > 0.005, confidence > 0.5, and lift > 3

```
filtered_rules <- subset(rules, support > 0.005 & confidence > 0.5 & lift > 2)
```

Inspect the filtered rules

```
inspect(filtered_rules)
```

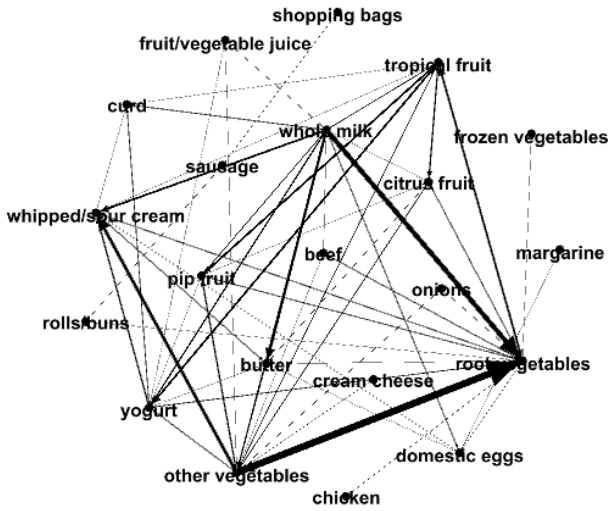
Extract the LHS (antecedent) and RHS (consequent) of each rule

```
lhs_items <- labels(lhs(filtered_rules))
rhs_items <- labels(rhs(filtered_rules))

# Combine LHS and RHS into a data frame for easy viewing
```

```
rules_lhs_rhs <- data.frame(LHS = lhs_items, RHS = rhs_items)

# View the first few rows
rules_lhs_rhs
```

Graph-based visualization

```
grocery_graph = associations2igraph(subset(rules, lift>3),
associationsAsNodes = FALSE)
igraph::write_graph(grocery_graph, file='grocery.graphml', format =
"graphml")
```



High-frequency items “whole milk” or “other vegetables” dominate the association rules, #we will address this issue by adjusting the parameters and techniques used in the Apriori algorithm.

Increase the Support Threshold for High-Frequency Items

Apply the Apriori algorithm with higher support for high-frequency items

```
rules1 <- apriori(groceries, parameter = list(supp = 0.01, conf = 0.2,
minlen=2,maxlen = 5),
                  appearance = list(none = c("whole milk", "other
vegetables")))
```

Filter rules by lift

```
filtered_rules <- subset(rules, lift > 3)
```

Inspect the top 10 rules

```
inspect(filtered_rules)
```

Identify redundant rules

```
redundant <- is.redundant(filtered_rules)
```

Remove redundant rules

```
Rrules <- filtered_rules[!redundant]
```

Sort the remaining rules by lift

```
Rrules <- sort(Rrules, by = "lift", decreasing = TRUE)
```

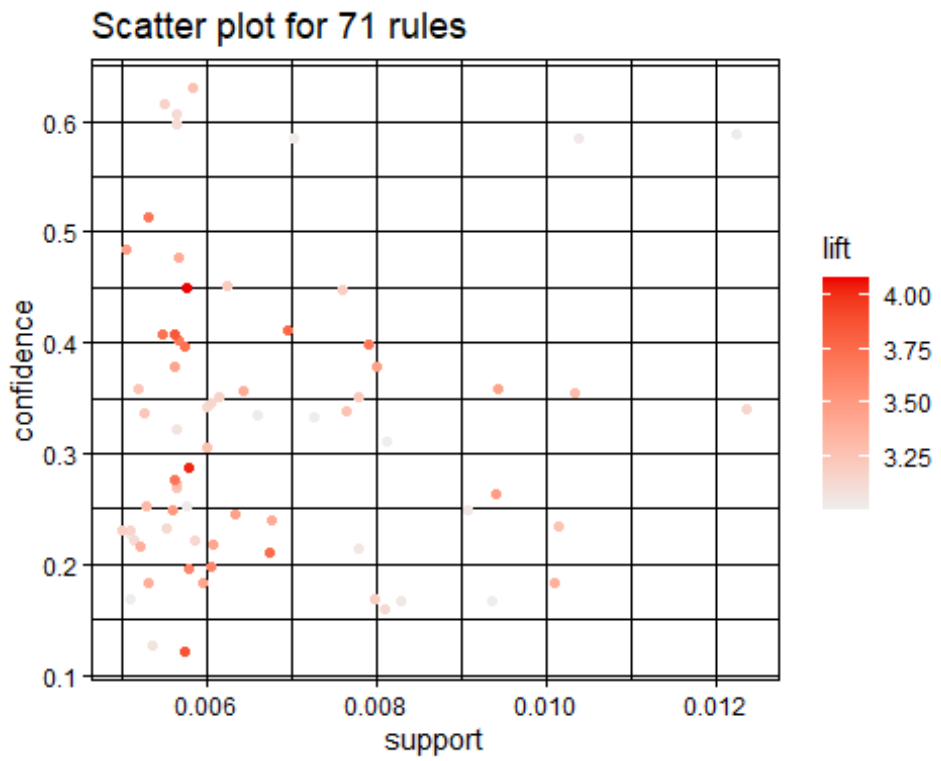
Inspect the top 10 rules

```
inspect(Rrules)
```

Visualization

Visualization 1:

```
plot(Rrules, method = "scatterplot", measure = c("support", "confidence"),
shading = "lift")
```



- The presence of rules with lift values between 3.25 and 4.0 suggests that the items in these rules are more strongly associated with each other than expected by chance. This makes these rules potentially useful for marketing strategies, as they indicate combinations of products that tend to be purchased together
- The plot shows a few rules with higher lift values (darker red points) around specific support and confidence levels. These high-lift rules are particularly interesting as they suggest very strong associations between the items involved, even if they occur in a small fraction of transactions

Graph-based visualization

```
grocery_graph = associations2igraph(subset(Rrules, lift>3),
associationsAsNodes = FALSE)
igraph::write_graph(grocery_graph, file='grocery_V1.graphml', format =
"graphml")
```

Insights

1.The Vegetable and Dairy Synergy:

Insight: Root vegetables are a cornerstone of our customers' grocery baskets, often purchased with other vegetables and dairy products. For example, when customers buy a combination of "citrus fruit," "other vegetables," and "whole milk," they are over four times more likely to also purchase "root vegetables" (lift = 4.09).

Actionable Strategy: "Veggie and Dairy Power Packs":

- Create a bundled promotion that combines root vegetables with dairy items like butter, yogurt, and sour cream.
- Highlight this bundle in-store with a "Farm Fresh Essentials" display, encouraging customers to buy these items together.

2.High-Impact Dairy Pairings:

Insight: Dairy products drive complementary purchases. For instance, when "whipped/sour cream" is bought with "whole milk," customers are over three times more

likely to add “butter” to their cart (lift = 3.76). Similarly, “butter” and “other vegetables” frequently lead to the purchase of “whipped/sour cream” (lift = 4.04).

Actionable Strategy: Dairy Pair Promotions:

- Position dairy products like butter, sour cream, and milk together in a “Perfect Pairings” section.
- Offer discounts on combined purchases, such as “Buy 2, Get 1 Free” on select dairy products.

3.Fruity Trio Trends:

Insight: Fruits exhibit strong co-purchase behavior, particularly tropical and citrus varieties. For example, customers buying “citrus fruit” and “tropical fruit” are nearly four times more likely to also buy “pip fruit” (such as apples and pears) (lift = 3.71).

Actionable Strategy: “Fruit Trio Specials”:

- Organize a seasonal fruit display featuring these three categories.
- Offer a promotional price when customers purchase all three types together, promoting recipes or snacks that combine these fruits.

4.Reinforcing the Core: Whole Milk as a Key Basket Driver:

Insight: Whole milk is central to many shopping baskets, frequently linking with both vegetables and fruits. A notable example is the rule {whole milk, yogurt} => curd with a lift of 3.37, indicating strong demand for dairy product clusters.

Actionable Strategy: “Milk and More” Campaign:

- Feature whole milk as the anchor product in various cross-category promotions.
- For instance, “Buy Whole Milk and Get 10% Off on Any Yogurt or Curd” could drive additional sales and increase basket size.

5.Breakfast Essentials: Leveraging Morning Staples:

Insight: Bread and eggs are often paired with protein products like sausage. The association {rolls/buns, shopping bags} => sausage with a lift of 3.27 suggests a strong breakfast trend among our shoppers.

Actionable Strategy: “Build Your Breakfast” Promotion:

- Cluster rolls, eggs, and breakfast meats together in a dedicated section.
- Offer combo deals that encourage customers to stock up on all their breakfast essentials in one go.

6.The Power of Cross-Promotions in High-Traffic Areas:

Insight: Certain vegetables and dairy combinations are highly predictive of other purchases. For instance, buying “other vegetables,” “root vegetables,” and “whole milk” significantly increases the likelihood of purchasing “citrus fruit” (lift = 3.02).

Actionable Strategy: Cross-Promotion Displays:

- Place related products near each other with clear signage promoting the cross-purchase.
- For example, a refrigerated display that combines root vegetables, dairy, and citrus fruits can drive multiple-item purchases