KENNESAW STATE UNIVERSITY
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Market Basket Analysis with Apriori

An unsupervised learning technique to recognize purchasing patterns of consumers

Musfiqus Salehine, MSAS – Faculy Supervisor: Lili Zhang

INTRODUCTION

In today's data-driven world, almost every company has huge database of purchase transaction. Therefore, a big question arises that what products are more likely to be purchased together. To answer this, Market Basket Analysis with Apriori can be used.

It is an unsupervised learning technique that can be used to analyze the **purchasing patterns of consumers**. It can be used as a recommendation mechanism, for example, product recommendation, music recommendation, and others promotional strategies can be developed.

METHODS

- Apriori is a pattern mining algorithm.
  - All subsets of a frequent itemset must be frequent
  - For any infrequent itemset, all its supersets must be infrequent.
- To apply the algorithm, at first, I transformed the dataset into transactional data. Apriori uses transactional data to design association models. In transactional data, there is a one-to-many relationship between the case identifier and the values for each case.
- The Apriori algorithm has been applied with support = 0.001, confidence = 0.8, and minlen=2.
- To check association for an specific item, the apriori has been applied with supp=0.001,conf = 0.05, minlen=2, and appearance = list(default="rhs",lhs=("rice")). The inspection of this rules only displays the association between 'Rice' and other items.

RESULTS

- The dataset has 9835 rows (number of transactions) and 169 columns (unique products)
- After implementing apriori on the dataset, 410 rules are generated. A length of 5 items has the most rules. The summary of quality is measured by ranges of support, confidence, and lift.
- The **Figure 05** displays the Top 5 rules based on confidence. From the figure
  - 100% consumers who bought rice and sugar bought whole milk.
  - 100% consumers who bought butter domestic eggs, and soft cheese also bought whole milk.
- The **Figure 06** displays the Top 5 rules for a specific item 'Rice' based on confidence. From the figure,
  - 61.3% consumers who bought rice bought milk.
  - 30.6% consumers who bought rice bought yogurt.
- The **Figure 07** displays the Top 5 rules for a specific item 'Rice' based on lift. From the figure,
  - Consumers who bought rice are nearly 5.5 times more likely to buy hard cheese.
  - Consumers who bought rice are nearly 4 times more likely to buy chickens.

```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
whole milk   other vegetables   rolls/buns   soda   yogurt  (Other)
  2513            1903             1809      1715    1372    34055
```

Figure 01: Summary of RawTransactional Data

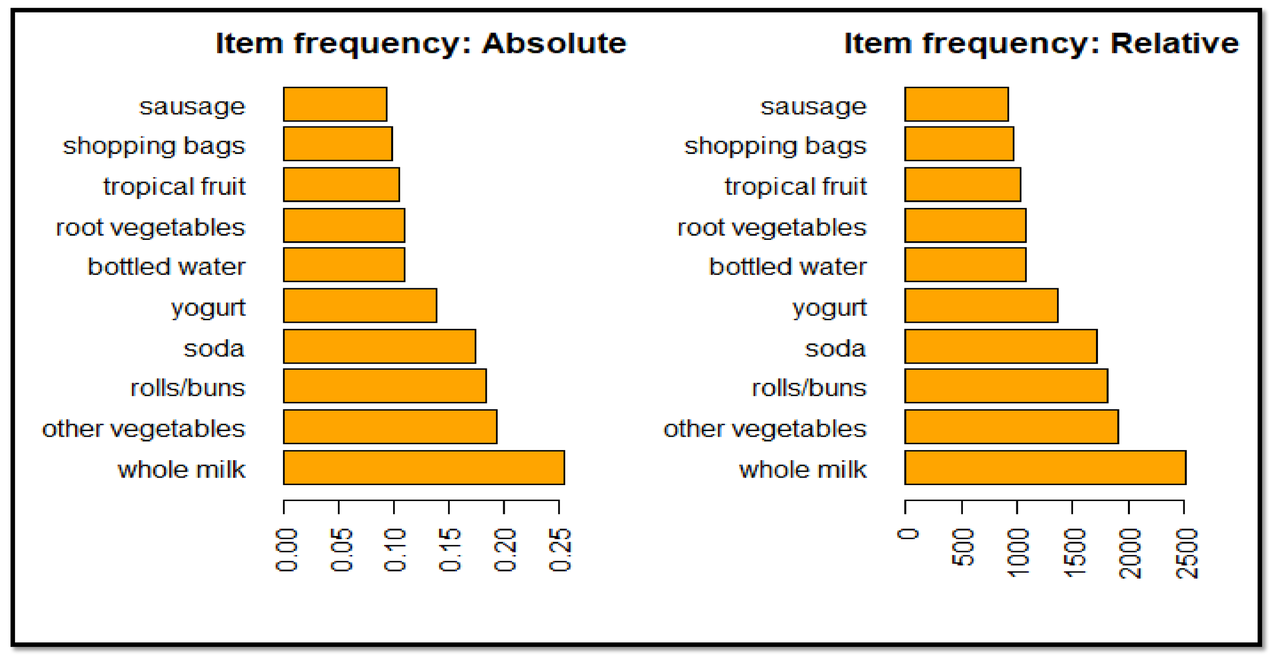

Figure 02: Absolute and Relative Distribution of Top 10 Selling Items

```
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.329   5.000   6.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001017   Min.   :0.8000   Min.   : 3.131   Min.   :10.00
 1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.: 3.312   1st Qu.:10.00
 Median :0.001220   Median :0.8462   Median : 3.588   Median :12.00
 Mean   :0.001247   Mean   :0.8663   Mean   : 3.951   Mean   :12.27
 3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.: 4.341   3rd Qu.:13.00
 Max.   :0.003152   Max.   :1.0000   Max.   :11.235   Max.   :31.00

mining info:
    data ntransactions support confidence
 master2          9835   0.001        0.8
```

Figure 03: Summary of Rules after Applying Apriori

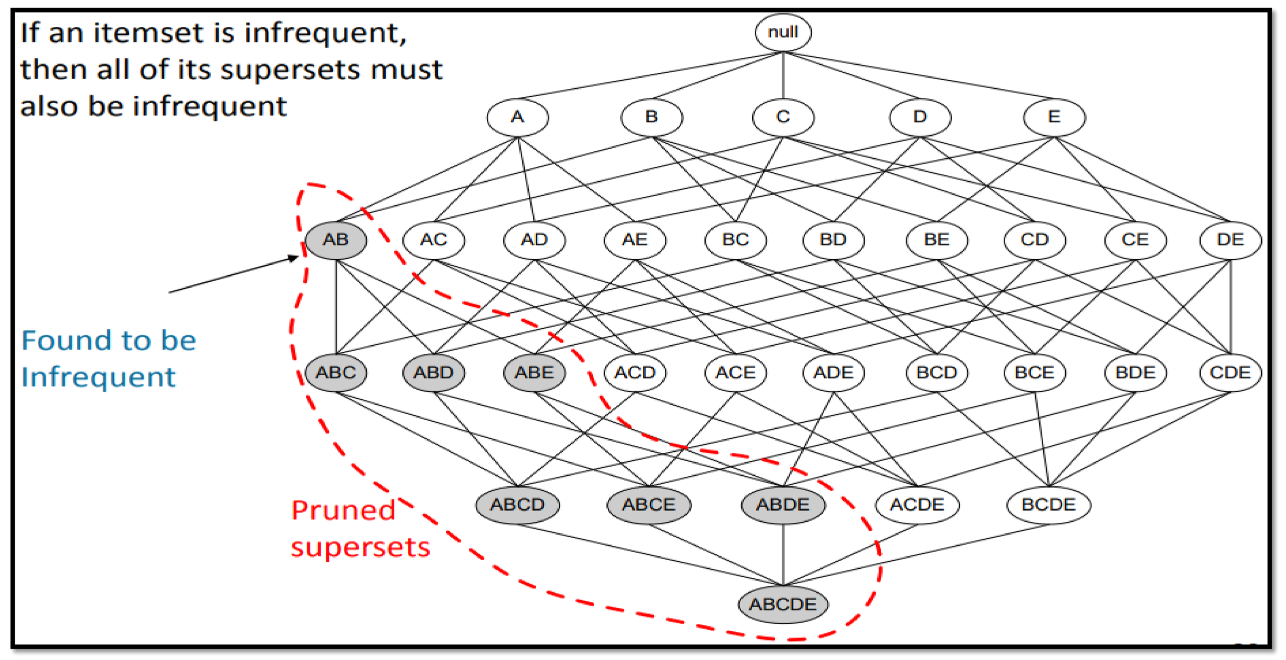

Figure 04: Apriori Principles (Source: Frequent Itemset Generation Using Apriori Algorithm by Chih-Ling Hsu)

lhs		rhs	support	confidence	lift
{rice,sugar}	=>	{whole milk}	0.001220132	1	3.913649
{canned fish,hygiene articles}	=>	{whole milk}	0.001118454	1	3.913649
{butter,rice,root vegetables}	=>	{whole milk}	0.001016777	1	3.913649
{flour,root vegetables,whipped/sour cream}	=>	{whole milk}	0.001728521	1	3.913649
{butter,domestic eggs,soft cheese}	=>	{whole milk}	0.001016777	1	3.913649

Figure 05: Association Check for Top 5 Rules by Confidence

lhs		rhs	support	confidence	lift
{rice}	=>	{whole milk}	0.004677173	0.6133333	2.400371
{rice}	=>	{other vegetables}	0.003965430	0.5200000	2.687441
{rice}	=>	{root vegetables}	0.003152008	0.4133333	3.792102
{rice}	=>	{yogurt}	0.002338587	0.3066667	2.198299
{rice}	=>	{fruit/vegetable juice}	0.001931876	0.2533333	3.504266

Figure 06: Association Check for Top 5 Rules for Rice by Confidence

lhs		rhs	support	confidence	lift
{rice}	=>	{hard cheese}	0.001016777	0.1333333	5.441217
{rice}	=>	{sugar}	0.001220132	0.1600000	4.725526
{rice}	=>	{butter}	0.001830198	0.2400000	4.331009
{rice}	=>	{chicken}	0.001321810	0.1733333	4.039652
{rice}	=>	{hamburger meat}	0.001016777	0.1333333	4.010194

Figure 07: Association Check for Top 5 Rules for Rice by Lift

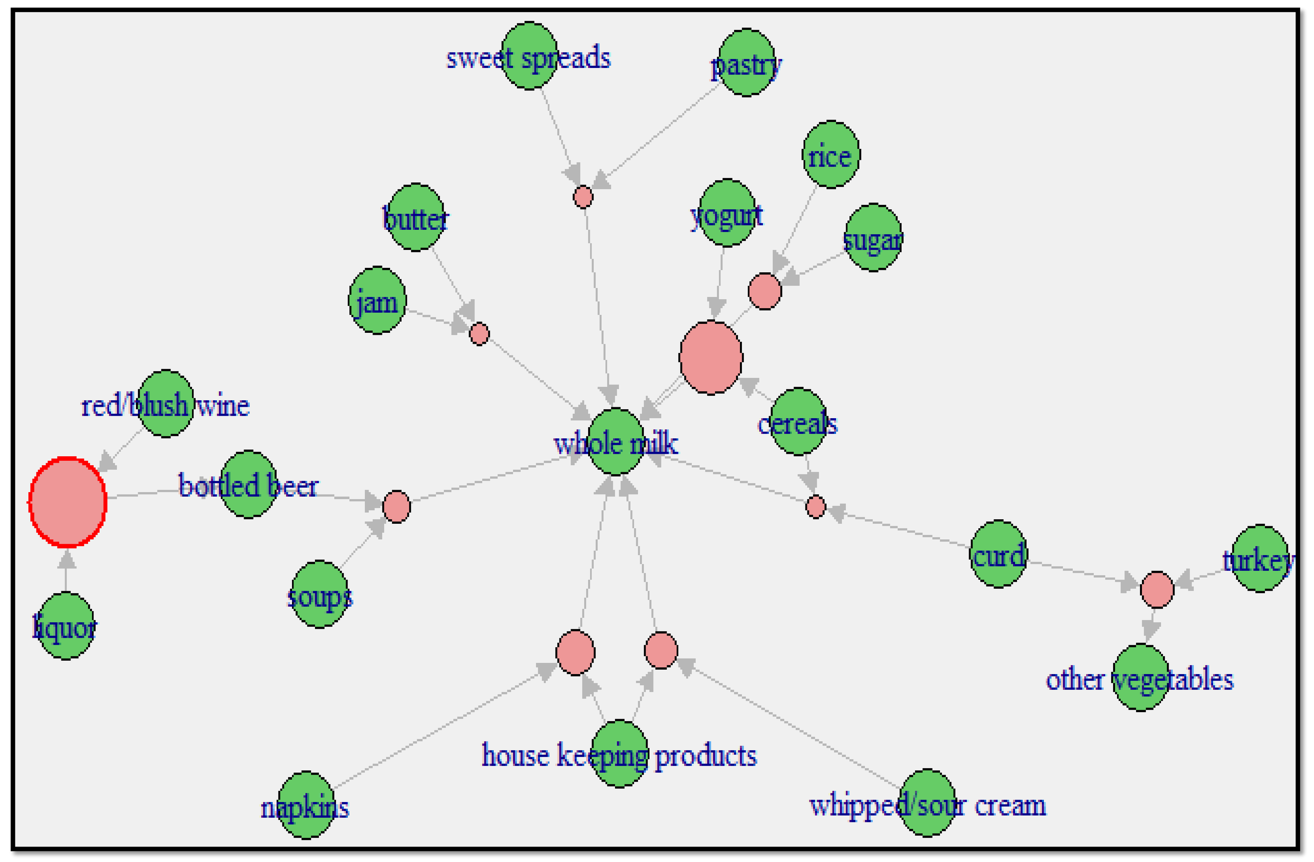

Figure 08: Overall Top 10 Association Rules

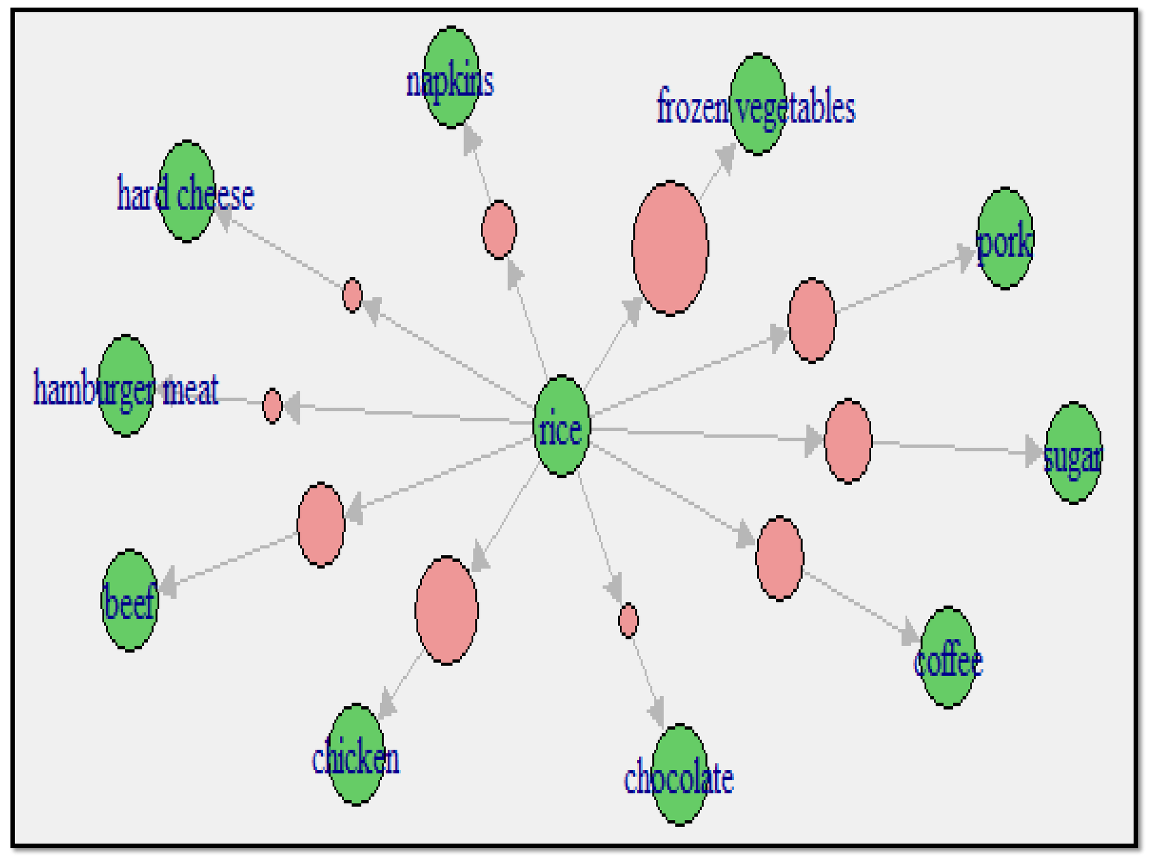

Figure 09: Overall Top 10 Association Rules for Rice

DISCUSSION

- **Support:** The fraction of transactions in data set that contain that product or set of products.
- **Confidence:** It is conditional probability implies that a customer buy product A will also buy product B.
- **Lift:** If someone buys Product A, what % of chance of buying product B would increase.
- **Density:** It refers to the proportion of non-zero cells in the matrix..

From this Market Basket Analysis, a lot of different strategic questions can be answered. For example,

- What are the most purchased items? **Figure 02** displays the top 10 selling items
- What are the least selling items?
- What items should be kept together to boost sales?
- What items should not be kept together?
- Which products boost the sales of other products? Etc.

This model can also be applied in lot of other sectors. For example,

- Unusual credit card purchases to detect fraud.
- Finding potential clients in insurance industries.
- Finding the pattern of different diseases based on human behavior and provide health recommendation.

A major limitation of Association Technique with Apriori is that it is slow when there are a large number of transactions. Therefore, specifying the number of rules to be inspected is often necessary.

R CODE

```
# Install package "arules"
install.packages("arules")
library(arules)

# create path
path <- "C:\\Users\\msalehin\\Desktop\\FALL 2017\\Programming in R\\R Project\\R Project"
setwd(path)

# import file
master1 <-  read.csv("groceries.csv", header = FALSE)
summary(master1)
master2 <- read.transactions("groceries.csv", sep = ",")
summary(master2)

# Diplay Top 10 in Relative and Absolute Frequency in Chart
par(mfrow=c(1,2))
itemFrequencyPlot(master2,type="relative",topN=10,horiz=TRUE,col='orange', xlab='',
                  main='Item frequency: Absolute')
itemFrequencyPlot(master2,type="absolute",topN=10,horiz=TRUE,col='orange',xlab='',
                  main='Item frequency: Relative')
# Apply Apriory
Apply_Apriori1 <- apriori(master2, parameter = list(supp = 0.001, conf = 0.8, minlen=2))

# Display overall summary
summary(Apply_Apriori1)

# inspect top 5 rules by confidence and by lift
inspect(sort(Apply_Apriori1, by = 'confidence')[1:5])
inspect(sort(Apply_Apriori1, by = 'lift')[1:5])

# Apply Apriory for Specific item
Apply_Apriori2 <- apriori (master2, parameter=list (supp=0.001,conf = 0.05, minlen=2),
          appearance = list(default="rhs",lhs=("rice")), control = list (verbose=FALSE))

# inspect top 5 rules for specific tem by confidence and by lift
inspect(sort (Apply_Apriori2 , by="confidence")[1:5])
inspect(sort (Apply_Apriori2 , by="lift")[1:5])

# Install package "arulesViz"
install.packages("arulesViz")

# plot Top 10 Rules for Specific item
library(arulesViz)
plot(Apply_Apriori1[1:10],interactive=TRUE,method="graph",shading=NA)

# plot Top 10 Rules for Specific item
library(arulesViz)
plot(Apply_Apriori2[1:10],interactive=TRUE,method="graph",shading=NA)
```