Association Rules Mining using the Retail Market Basket Data Set

Apostol, Edward Nataniel C.
2010-27627 | CS 176 THX
edward.nataniel@gmail.com

1. INTRODUCTION

Association rule mining is useful for discovering relationships among data in a database. It has many applications in marketing and sales. One example is supermarket shelf management.

In this programming assignment, the Apriori and FP-growth algorithms will be used to mine association rules from the provided dataset. The rules generated will be interpreted.

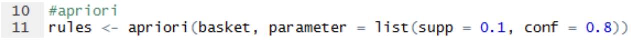
2. OBJECTIVES

The objective of this analysis is to be able to apply the apriori and fp-growth algorithms for mining association rules to a market basket data. The parameters and correlation measures will also be modified and analyzed to look for interesting results.

3. METHODOLOGY

3.1 Apriori

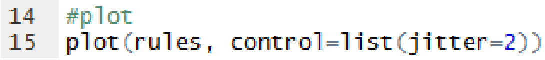
A package named *arules* and *aruleViz* was installed in R to do the mining and visualization of the rules.



```
10  #apriori
11  rules <- apriori(basket, parameter = list(supp = 0.1, conf = 0.8))
```

Figure 1. Line of code that mines the association rules

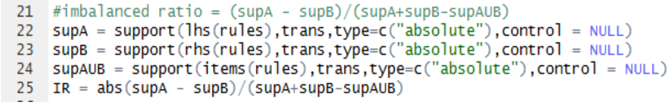
The line 11 of the file *basket-apiori-ApostolEdwardNataniel.R* is the part where the association rules where mined. The *apriori* function from the *arules* package was used to mine the association rules. The parameters *sup* and *conf* are the minimum support and confidence that we will allow.



```
#plot
plot(rules, control=list(jitter=2))
```

Figure 2. Plotting of the rules

We can also plot the rules by using the *plot* function in the *arulesViz* package. A scatter plot of confidence-support will be shown.



```
21  #imbalanced ratio = (supA - supB)/(supA+supB-supAUB)
22  supA = support(lhs(rules),trans,type=c("absolute"),control = NULL)
23  supB = support(rhs(rules),trans,type=c("absolute"),control = NULL)
24  supAUB = support(items(rules),trans,type=c("absolute"),control = NULL)
25  IR = abs(supA - supB)/(supA+supB-supAUB)
```

Figure 3. Computing for the imbalanced ratio (IR)

The rules are stored in the rules variable. The rules are of the form LHS => RHS. In order to compute the Imbalanced Ratio (IR), we need to calculate for the support of LHS (line 22), support of RHS (line 23) and the support of LHS U RHS (line 24).

$$IR = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

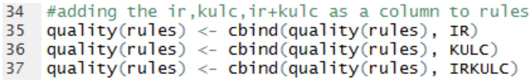
Figure 4. Formula for IR

$$Kulc(A, B) = \frac{1}{2}(P(A|B) + P(B|A))$$

Figure 5. Kunlczynski Measure

IR is then computed using the formula in Figure 4. A is LHS and B is the RHS. Kulczynski measure can also be easily computed using the formula in Figure 5. It is because P(A|B) = support_count(AUB) / support_count(A). So Kulc can be computed as:

Kulc (A,B) = ½ (supAUB/supA + supAUB/supB)



```
#adding the ir,kulc,ir+kulc as a column to rules
quality(rules) <- cbind(quality(rules), IR)
quality(rules) <- cbind(quality(rules), KULC)
quality(rules) <- cbind(quality(rules), IRKULC)
```

Figure 6. Adding a new column for the measures

We add this measures to *rules* as a new column. The parameters were varied and the observations were reported in section 4.

3.2 FP-Growth

In the FP-growth mining, the program can easily be run from the command line. The general syntax of the program invocation is

```
fpgrowth [options] infile [outfile]
```

The program was run with the ff options:

-m1: The minimum number of items per itemset is 1

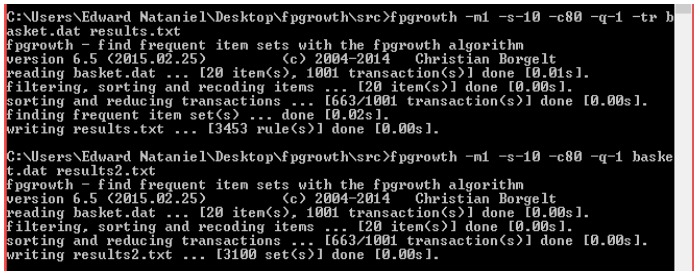
-q -1: Items will be sorted according to item frequency in descending order

-s10: Minimum support is 10%

-c80: Minimum confidence is 80%

-tr : The program will generate the association rules.

The program was run twice. The first run was to generate the frequent itemsets (without the –tr option). The second was to output the association rules.The observations were reported in the next section.



```
C:\Users\Edward Nataniel\Desktop\fpgrowth\src>fpgrowth -m1 -s-10 -c80 -q-1 -tr b
asket.dat results.txt
fpgrowth - find frequent item sets with the fpgrowth algorithm
version 6.5 (2015.02.25)         (c) 2004-2014   Christian Borgelt
reading basket.dat ... [20 item(s), 1001 transaction(s)] done [0.01s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [663/1001 transaction(s)] done [0.00s].
finding frequent item set(s) ... done [0.02s].
writing results.txt ... [3453 rule(s)] done [0.00s].

C:\Users\Edward Nataniel\Desktop\fpgrowth\src>fpgrowth -m1 -s-10 -c80 -q-1 baske
t.dat results2.txt
fpgrowth - find frequent item sets with the fpgrowth algorithm
version 6.5 (2015.02.25)         (c) 2004-2014   Christian Borgelt
reading basket.dat ... [20 item(s), 1001 transaction(s)] done [0.00s].
filtering, sorting and recoding items ... [20 item(s)] done [0.00s].
sorting and reducing transactions ... [663/1001 transaction(s)] done [0.00s].
writing results2.txt ... [3100 set(s)] done [0.00s].
```

Figure 7.Running the fpgrowth program

4. EXPERIMENTAL RESULTS, ANALYSIS OF RESULTS

4.1 Apriori

With min_support = 10% and min_conf = 80%, this plot was generated.

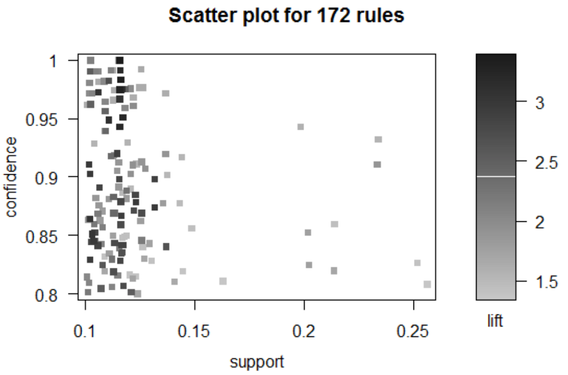
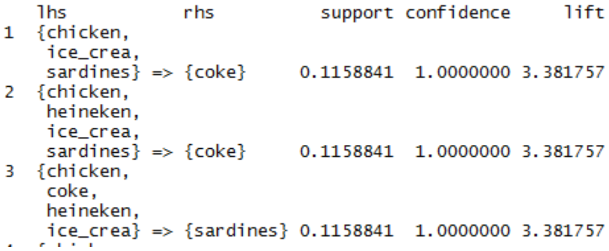


Figure 8. Scatter plot of rules (minsup=10%, minconf=80%)

Note that in the scatter plot, there are rules with a very high support but has positive lift. Positive lift means that the rules are positively correlated (presence of one implies occurrence of another). The rule on the rightmost side is the rule {soda}=>{Heineken}.

The rules with the highest lift are listed in Figure 8.



```
    lhs              rhs          support  confidence    lift
1   {chicken,
     ice_crea,
     sardines} => {coke}      0.1158841  1.0000000 3.381757
2   {chicken,
     heineken,
     ice_crea,
     sardines} => {coke}      0.1158841  1.0000000 3.381757
3   {chicken,
     coke,
     heineken,
     ice_crea} => {sardines} 0.1158841  1.0000000 3.381757
```

Figure9. Rules with the highest lift (min_sup=10%, minconf=80%)

These rules has a support of 11.59% and a confidence of 100%. It means that all of the 116 records that contains the items on the LHS, the item of the RHS appears together with them. And since the lift is also >1, we can say that there is a positive correlation between the items on the LHS and RHS.

By decreasing the min_sup to 5% and min_conf to 60%, more rules were generated. Here is the scatter plot of these rules:

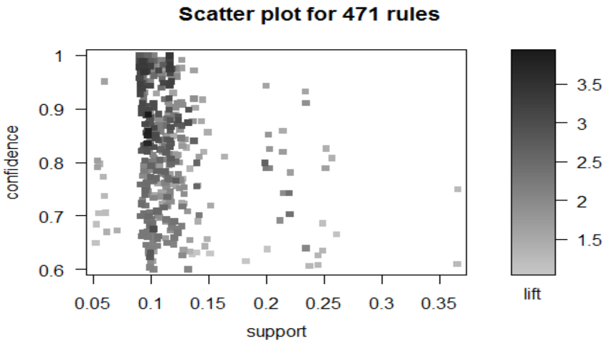
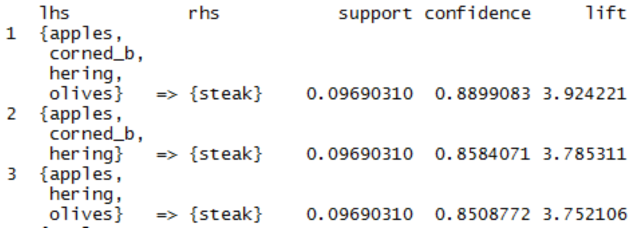


Figure 10.Figure 7. Scatter plot of rules (minsup=5%, minconf=60%)

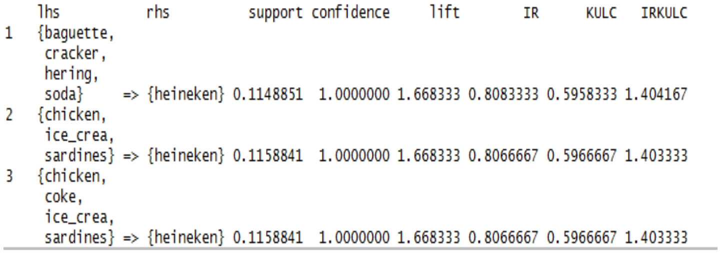
By decreasing the min_support and min_conf, we were able to find new rules with higher lift.



```
    lhs          rhs        support  confidence    lift
1   {apples,
     corned_b,
     hering,
     olives}   => {steak}  0.09690310  0.8899083 3.924221
2   {apples,
     corned_b,
     hering}   => {steak}  0.09690310  0.8584071 3.785311
3   {apples,
     hering,
     olives}   => {steak}  0.09690310  0.8508772 3.752106
```

Figure 11.Rules with the highest lift (min_sup=5%, minconf=60%)

The top 3 rules using the IR + Kulc measure:



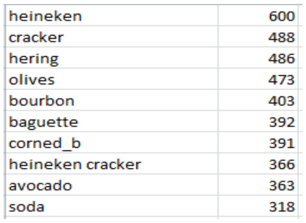
```
    lhs            rhs           support confidence    lift        IR      KULC   IRKULC
1   {baguette,
     cracker,
     hering,
     soda}      => {heineken} 0.1148851  1.0000000 1.668333 0.8083333 0.5958333 1.404167
2   {chicken,
     ice_crea,
     sardines} => {heineken} 0.1158841  1.0000000 1.668333 0.8066667 0.5966667 1.403333
3   {chicken,
     coke,
     ice_crea,
     sardines} => {heineken} 0.1158841  1.0000000 1.668333 0.8066667 0.5966667 1.403333
```

Figure 12. Rules with high Kulc + IR measure. (min_sup=10%, min_conf=80%)

Note that the support of the top 3 rules are a bit lower when we use Kulc + IR than lift. IR and Kulc are both correlation measures. Higher values means that there is closer relationship between the LHS and RHS of the rule. Kulc + IR is a recommended measure.

4.2 FP-growth

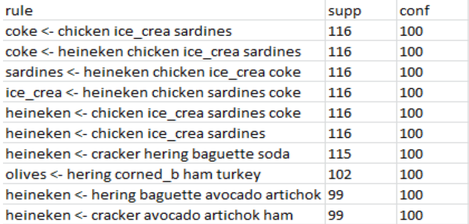
A total of 1000 frequent itemsets satisfy the minimum support of 10.



heineken	600
cracker	488
hering	486
olives	473
bourbon	403
baguette	392
corned_b	391
heineken cracker	366
avocado	363
soda	318

Figure 13. Top 10 itemsets with highest support

The ff. association rules was generated. The ff. was sorted accdg to highest confidence and support.



rule	supp	conf
coke <- chicken ice_crea sardines	116	100
coke <- heineken chicken ice_crea sardines	116	100
sardines <- heineken chicken ice_crea coke	116	100
ice_crea <- heineken chicken sardines coke	116	100
heineken <- chicken ice_crea sardines coke	116	100
heineken <- chicken ice_crea sardines	116	100
heineken <- cracker hering baguette soda	115	100
olives <- hering corned_b ham turkey	102	100
heineken <- hering baguette avocado artichok	99	100
heineken <- cracker avocado artichok ham	99	100

Figure 14. Generated rules of fp-growth (min_sup=10%, min_conf=80%)

5. CONCLUSION

Association Rule mining can be a big help in Market Analysis. Two techniques of mining Association Rules were used here. FP Growth is more efficient/faster than apriori. It can also be seen that the parameters like minimum support and minimum confidence can be modified to achieve the desired result. There were also correlation measures that determines whether the generated rule is interesting (meaning, the LHS and RHS of the rule are positively correlated). Lift is a good measure but it is unstable. The Kulczynski measure in junction with IR measure is a recommended measure.