Hw2 IEMS 308

SHENGJIE LIU

**Association Rule**

**Executive Summary**

Association rule is a rule-based machine learning method for discovering in interesting relations between variables in large database. Nowadays, this method is largely applied in many business regions. For example, in Amazon, when you decide to buy something to check out, there is a reminder near the button ‘check out’ that XX% bought same thing also buy XX. What behinds this function is the association rule.

This data mining technique can also help retail stores to boost their sales volume. By performing association rule, retailer stores can know the relationship among products. Therefore, they can make the promotion on one of the products or rearrange the floors of the stores to improve their sales volume.

Our client is a retail chain that has over 400 stores across the country. Besides, it has over 100000 Skus in their stores. In this case, our client wants to rearrange the floors of the stores to improve their sales volume. However, it is difficult for our client to rearrange all floors due to the constraints of the cost, number of staff. Therefore, it is useful to use association rule to find strong relationship among products.

In this case, due to large data set, we select 200,000 data (random sampling) to do association rules analysis. Based on the request from client, we need to get the top 20 associations from 100 SKU candidates.

By performing association rules, we get 52 rules which are 104 products).Then, by fixing the supp and conf, we can get our top 20 assoication rules. For these 20 rules, they have high confidence and lift.

**Problem Statement**

Our client wants to rearrange the floors of the stores to improve their sales volume. However, it is difficult for our client to rearrange all floors due to the constraints of the cost, number of staff. Therefore, our client ask us to find the 100 SKUs that are best candidates to modify the planograms and the best 20 moves.

**Assumption**

1. This data set is general. Based on our association rules, we assume that all customer behaviors and all retailer stores are relatively consistent.
2. In this case, our goal is to boost its sales volume, not the profit.
3. In this case, we do not care about the return product. We take these products into account when we analyze the dataset.

**Methodology**

In this case, basically we can divide our methodology into four steps.

The first step is that we use library (DBI, RPostgresql) in R to connect our database to get the entire data we want.

The second step is that we explore the features of the data to make sure which columns we need to analyze and rename the column name, since table pos.trnsact column’s name is not defined well for our analysis. Then we transfer the type of the column, so we can easily combine the columns (‘Store’,’Register’,’Trannum’,’Saledate’) into our column ‘basket’. After this, we can write the modified data into csv.file, which is helpful for us to use read.transaction function.

The third step is to use the association rule library (apriori) t to get all the association rules. Then we prune the redundant rules to get the final association rules. After this, we fix the supp and conf to get our 20 best association rules.

The fourth step is the data visualization for the association rules. In this step, we use library (arulesViz).

**Analysis:**

**The top 20 rules (move):**

**lhs rhs support confidence lift count**

**[1] {,"9837444} => {,"9847561} 5.938659e-05 0.05583756 43.274579 11**

**[2] {,"9849902} => {,"9839902} 4.858903e-05 0.05232558 38.613986 9**

**[3] {,"9838844} => {,"9848844} 2.159512e-05 0.02702703 37.640114 4**

**[4] {,"9838227} => {,"9848227} 2.699390e-05 0.03184713 35.112792 5**

**[5] {,"9848059} => {,"9838059} 1.149940e-03 0.20821114 35.092198 213**

**[6] {,"9837606} => {,"9847606} 6.478537e-05 0.07792208 32.952677 12**

**[7] {,"9849530} => {,"9839530} 4.319025e-05 0.04232804 32.397919 8**

**[8] {,"9850179} => {,"9840179} 3.779147e-05 0.03255814 19.903124 7**

**[9] {,"9848246} => {,"9838246} 2.159512e-05 0.02877698 19.173645 4**

**[10] {,"9836307} => {,"9846307} 3.779147e-05 0.03017241 16.485975 7**

**[11] {,"9848074} => {,"9838074} 2.699390e-05 0.02164502 16.364255 5**

**[12] {,"9838419} => {,"9848419} 4.319025e-05 0.02259887 11.925701 8**

**[13] {,"9849975} => {,"9839975} 2.159512e-05 0.01777778 11.677037 4**

**[14] {,"9837337} => {,"9847337} 6.478537e-05 0.02580645 9.755207 12**

**[15] {,"9839890} => {,"9849890} 6.478537e-05 0.02362205 8.240002 12**

**[16] {,"9838665} => {,"9848665} 3.779147e-05 0.01794872 6.455509 7**

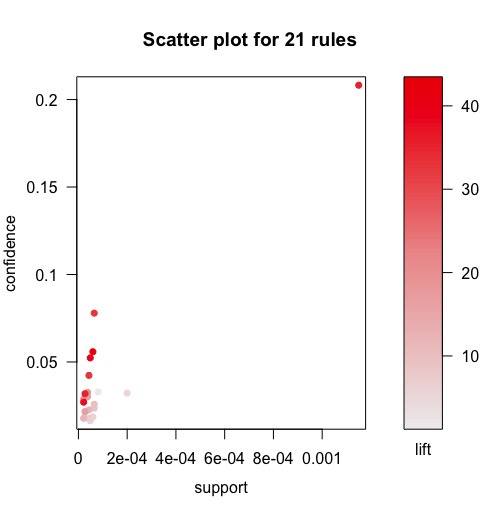
**[17] {,"9847494} => {,"9837494} 5.938659e-05 0.01864407 5.698655 11**

**[18] {,"9837292} => {,"9847292} 4.858903e-05 0.01636364 5.490919 9**

**[19] {,"9847534} => {,"9837534} 1.997549e-04 0.03220191 5.141952 37**

**[20] {,"9836812} => {,"9846812} 2.159512e-05 0.01818182 2.707205 4**

**[21] {,"9844793} => {,"9844108} 8.098171e-05 0.03289474 1.816093 15**

In this case, these 20 rules can be observed from our analysis. We can notice that the support maybe relatively low. This is because we got these data from a very large dataset. 

Based on this plot, we can know that there are some rules that have relatively large fit and confidence, which means that they have a strong relationship.

**Conclusion:**

We get about 21 rules in our case. The retailer can choose top 20 as their move plan. We notice that in these 21 rules, the supp is relatively low compared with lift and conf. It is not necessary for us to analyze the products with high supp. Maybe they are just common products such as sauce, butter and so on. For those products, there is no need for us to analyze, because the demand is steady.

**Next step:**

Every results of data analysis need validation plans for both retailers and us. The next step for the retails will be execute these rules to rearrange the floors of the store. We should keep communication with the staff in retail store to validate if our recommendation rules are effective to boost the sales volume of products.