IEMS 308 Data Science & Analysis: Homework 2

Yiheng Pan

**Association Rule**

**Executive Summary:**

How to organize the position of each item in store is one of the most concerned topics for managers. Since the correct order for items can contribute high profits. For example, customers often purchase shampoo and conditioner together, so put both items together can make customers more convenient, which is customer-friendly strategy.

Dillard’s is an American department store chain with 453 stores and 104876 SKUs. Their point-of-sales data over period of time is available in database. And there are 5 information tables about the transaction, store location, SKU and department.

The Dillard’s retailer is interested in rearranging the floors of the sores. For manpower limited, they can only make 20 moves across the entire chain. One move means moving one SKU to a different position. Since this is the first time for Dillard’s to use this strategy and we assume that most associated SKUs are not close to each other and need to be moved. We are going to find out the 100 SKUs that are best candidates.

Association rule is a kind of data analysis and data mining technique that discovers co-occurrence relationships among activities performed by specific individuals or groups. And Market basket analysis reveals affinities between individual products or product groupings.

**Problem Statement:**

The economic behavior is always affected by the psychology of customers. If the company wants to increase their profit not only by adding new product or putting promotion, it’s a great strategy to locate the items related together so that the customer might purchase another one which she or he at first did not realize it’s needed. Due to the limit of human resource, only 20 moves can be made, and we are going to analyzing the top 100 items can be relocated in the store.

**Assumption:**

1. The transaction data includes large number of information and covers all the stores. We assume that the customer behaviors from all the district in the United States are the same.

2. This is the first time for Dillard’s to use the relocating strategy. We assume that the original location for each product is random. And it’s highly unlikely that the best candidates are close to each other.

3. Even though some purchased products were returned, we assume that the returned product are intentionally bought by customers. The record of any purchased product is meaningful for our analysis.

4. Assume that we are going to have the largest purchased quantity instead of selling something with low cost and high price. The most concerned topic here is how to sell products as many as we can.

5. Since the data set is too large, we need to sample part of it to analyze and we suppose that the sample data can represent as well as the whole data set.

**Methodology:**

We store the data in POSTGRES as database and use the software R to analyze data.

First, we understand the data by looking through the database diagram. There are five information tables. The STRINFO is information about the store, including store number, city, state and zip code. The SKSTINFO is information about the price, since different stores might sell the same SKU in different price, we use SKU and store number to mark one product price. The SKUINFO table shows the detailed information about SKU, such as department, UPC, style, color, size and so on. The DEPTINFO table shows the description for a certain department. And the TRNSACT table records each transaction. We specify the SKU in certain transaction by store number, register number, transaction number, sale date and sequence.

Second, we explore the data in POSTGRES. There are five tables. There’re 453 records about the store. It offers 760212 SKUs. The TRNSACT table is the main table we are going to use, which is about 10Gb. So we need to select the useful data. Since we are going to find the 100 top purchased SKUs to remove the redundant rows. By using SQL, we count the number of sold quantity and chose the first 100 SKUs.

In R software, first we connect with the POSTGRES and read data from database. The information we interested most is products sold best. We only load the transaction information about the top 100 purchased SKUs, which can remove redundant information. Since the data set is too large, we still need to randomly sample data in limited number 100000.

In order to put the products purchased in the same transaction together, we create a new label as basket ID, which marks one transaction by matching transaction number, register number, sale date, store together.

Then we use association rules to find the affinity between each product during purchase transaction. We set the minimum of support as 0.00001, confidence as 0.0125, and minimum length as 2 since we need at least the relationship between 2 items.

Because there might be some subset rules of each other, which is redundant information. We find out all the reproduced information and remove them to refine data. Print all the association rules and plot them all to analyze.

Analysis:

lhs rhs support confidence lift count

[1] {,"5851141} => {,"5849753} 1.063581e-05 0.33333333 1119.309524 1

[2] {,"5858066} => {,"5848066} 2.127162e-05 0.06896552 360.237548 2

[3] {,"5849901} => {,"5849757} 1.063581e-05 0.06250000 189.560484 1

[4] {,"5848350} => {,"5848583} 1.063581e-05 0.05263158 154.641447 1

[5] {,"5857729} => {,"5858086} 1.063581e-05 0.06250000 133.553977 1

[6] {,"5848163} => {,"5858163} 2.127162e-05 0.02739726 33.453834 2

[7] {,"5848327} => {,"5858095} 1.063581e-05 0.01923077 30.135256 1

[8] {,"5848459} => {,"5857437} 1.063581e-05 0.14285714 24.465782 1

[9] {,"5858011} => {,"5856921} 1.063581e-05 0.01886792 23.342105 1

[10] {,"5848038} => {,"5858038} 3.190743e-05 0.02777778 23.112586 3

[11] {,"5849539} => {,"5848473} 1.063581e-05 0.07142857 21.252712 1

[12] {,"5858151} => {,"5848636} 1.063581e-05 0.02702703 19.547193 1

[13] {,"5857658} => {,"5850202} 1.063581e-05 0.02000000 19.385979 1

[14] {,"5858120} => {,"5848120} 1.063581e-05 0.01449275 19.192080 1

[15] {,"5858049} => {,"5848049} 2.127162e-05 0.02000000 14.245758 2

[16] {,"5848893} => {,"5848956} 1.063581e-05 0.01265823 13.524453 1

[17] {,"5852768} => {,"5852854} 1.063581e-05 0.01818182 11.396606 1

[18] {,"5848524} => {,"5848665} 1.063581e-05 0.01250000 7.631656 1

[19] {,"5848208} => {,"5858208} 1.063581e-05 0.01265823 5.749526 1

[20] {,"5848788} => {,"5851778} 2.127162e-05 0.02531646 0.401739 2

[21] {,"5858310} => {,"5851778} 1.063581e-05 0.01449275 0.229981 1

21 rules:

Min. 1st Qu. Median Mean 3rd Qu. Max.

2 2 2 2 2 2

summary of quality measures:

support confidence lift count

Min. :1.064e-05 Min. :0.01250 Min. : 0.23 Min. :1.000

1st Qu.:1.064e-05 1st Qu.:0.01818 1st Qu.: 13.52 1st Qu.:1.000

Median :1.064e-05 Median :0.02532 Median : 21.25 Median :1.000

Mean :1.367e-05 Mean :0.05071 Mean : 105.92 Mean :1.286

3rd Qu.:1.064e-05 3rd Qu.:0.06250 3rd Qu.: 33.45 3rd Qu.:1.000

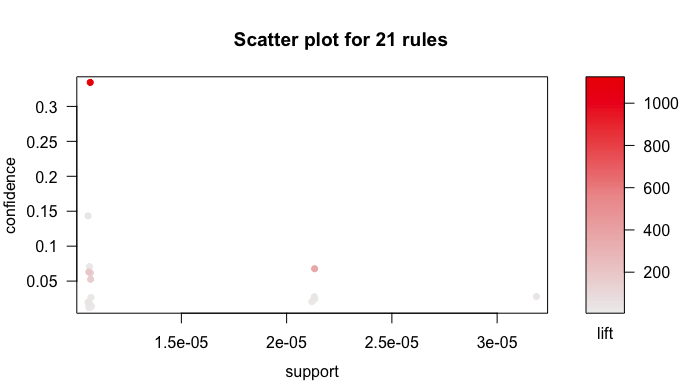
Max. :3.191e-05 Max. :0.33333 Max. :1119.31 Max. :3.000

mining info:

data transactions support confidence

transaction 94022 1e-05 0.0125

The information of rules above has been refined by removing all the redundant rules. All the lhs and rhs are one SKU. Since there so many data and transaction, the maximum of support is 3.191e-05 and the maximum of confidence is 0.33. But there are five rules’ lifts is above 100, which means high correlation between lhs and rhs.



Due to the large database and large amount of items, most rules are in low support as expected. As we discussed above, the products we selected are all best seller, which means they might be the most common items in reality, like tissue and salt. So people buy them just because they are always needed. And when we try to find out the best sellers’ confidence between each other, the result is low just because people buy them with a lot of other products. So that’s why in this project the confidence is also low. What we really care about is the lift, so we sort the rules with lift then try to find out the most coherent relationship between best sellers.

**Conclusion:**

There are 21 rules found. By following the coherence instruction, we should put each correlated products together. Look over the summary of the rules, we notice that there might be low confidence between best seller products just because they are too common and always be sold with other products in different combination. At first, we try to use the most popular products to remove too many data. But the best seller products might not be highly coherent with each other. Take tissue and salt for example, they are both the most common products in our life, but they are not related to each other like shampoo and conditioner.

**Next step:**

In this project, we focus on the top products whose sale numbers are large. But this doesn’t mean that the retailer can make most profits. Although some products are popular among customer, their net margin might be low. And some electronic products like mobile phone, which are cost low and sold at high price, should be focused on. In the next step, we are going to find out the rules between the high net margin products. These products have less quantity in stock, so it will be convenient to relocate. In order to make largest profits, the number of move will increase in spite of budget constraint.