CS584 A Study of Instacart Market Basket Analysis through Association Rules Mining

Binhui Xu (G01172196)

December 7, 2020

Abstract

Basket Market analysis is a technology that helps retailers understand the behavior of customers when shopping in their stores. In this project I am going to make use of Instacart's customer transaction data and focus on descriptive analysis on the customer purchase patterns, items which are bought together and to identify customer groups and subgroups with similar buying behavior, and visualize the data to provide productive recommendations that focus on improving the user experience by suggesting the next likely product to purchase to the customer during the order process.

1 Introduction

With the rise of the Internet of Things in modern society, there is increasingly high demand on the innovation. Based on the statistic portal, the number of people purchase goods online is continuous to increase. More research works and innovation are underway to meet people's needs and satisfaction. Understanding customer purchasing behavior through analyzing their transaction details turn to be one of widely used techniques in retail business[3].

Knowing which products are sold together is very useful for any business. The most obvious effect is facilitating impulse buying by reorganizing their products, so that things sold together can be found together. In addition, this has the side effect of improving customer satisfaction-once they find what they want, they don't need to research that particular product, which saves more time of customers and make the shopping process more effective[6]. In this project, I will analyze the Instacart transactional data in order to have better understand on market basket ananlysis. Instacart is an e-commerce website where users can buy groceries online from a local grocery store, and then let Instacart's individual shoppers come to pick up and deliver the goods on the same day[1].

With the provided dataset, I could come up with the best possible models for the below mentioned business objectives: To predict the next likely product, the customer would purchase during the ordering process.

2 Problem Statement

One of the major problems is how to choose the threshold values? The association rule analysis needs several thresholds to be specified in front. If the value is too low, we would get too much meaningless rules. In contrast, if the value is too high, it might turn to lose many possible recommendations that could be provided to customers.

In addition, the two major drawbacks of Apriori algorithm, that is, the size of itemset from candidate generation could be extremely large, and there needs a large amount of time on computing the support by scan the entire itemsets over and over again, which cause it is hard to find all rules accurately in this program.

3 Literature Review

Raorane A.A [3] analyze a large amount of data to discover consumer purchasing habits, so as to make correct business decisions to gain an advantage in the competition. The experimental analysis was carried out employing association rules by Market Basket Analysis, which proved the value of association rules compared with traditional methods.

Rakesh Agrawal et al [2] describe an effective algorithm that generates all essential association rules between items in the datasets. The algorithm incorporates buffer management and novel estimation and pruning techniques. Also, they present the experimental result of applying this algorithm to transactional data gain from a large retailer, which proven the effectiveness of this algorithm.

Yildiz, Baris et al [3] compare Matrix Apriori and FP-Growth algorithms. Using the generated two synthetic data sets, the different phases of the algorithm operation process were analyzed with examples,then to see their performance on datasets with distinct characteristcs, and understand the reason why different phase has different performance. The three conclusions they obtain: first,there is a close relationship between performance of algorithm and the characteristic of the given dataset and threshold value. Second, when the threshold value below 10%, Matrix Apriori has better perferance than FP-Growth in entire process. Third, although building matrix data structure more expensive, finding itersets is faster.

4 Methods and Techniques

4.1 Association Rules

Association rule can be thought of as an if-then relationship, just to elaborate on that we have come up with a rule, suppose if an item A is being bought by the customer, then the chances of item B being picked by the customer to under same transaction ID is found out.

Association rule mining is all about building the rules and we have just seen one rule that if you buy A then there's a slight possibility or there's a chance that you might buy B also[3].

Since this dataset contains a huge number of products and orders, so here I will implement the Apriori algorithm manually. Based on the orders and products data, it could generate the association rules from the relational information and transactions.

The goal is to predict the next product that customer would add to their cart during ordering processing based on their prior orders and purchasing pattern. Therefore, the product was bought before would be included in the antecedent, and the predict product would be included in the consequence. To finish the goal, there are three major step to go through:

- Step1: create frequent itemset within minimum support value.
- Step2: create association rule from frequent itemset with minimum lift.
- Step3: filter the rules which match the goal's description above.

I will go further for every step in result section.

4.2 Aprior Algorithm

For the Apriori algorithm, we use the degree of support as the standard to measure frequent itemset. The goal of Apriori is to find the largest frequent set of k items. This has two meanings. One is that we need to find a frequent set that meets the supported criteria. But there may be many such sets. The second level means that we need to find the largest number of frequent sets. For example, if we find that both the frequent set AB and ABE meet the support requirements, we will abandon AB and only keep ABE. Therefore, how does Apriori find the frequent set of k items?

The Apriori algorithm applies an iterative method. First, it finds a candidate set containing only one item and the corresponding support, and removes those items that do not meet the support requirements to obtain frequent one itemet. Then combine these frequent one itemets to get the candidate set of frequent two itemset. Similarly, filter out those frequent two itemets that do not meet the support requirements, and get the true frequent binomial itemsets, and so on. Iterate this operation until frequent k+1 itemset are found, and its corresponding frequent k itemset is the output result of the algorithm.

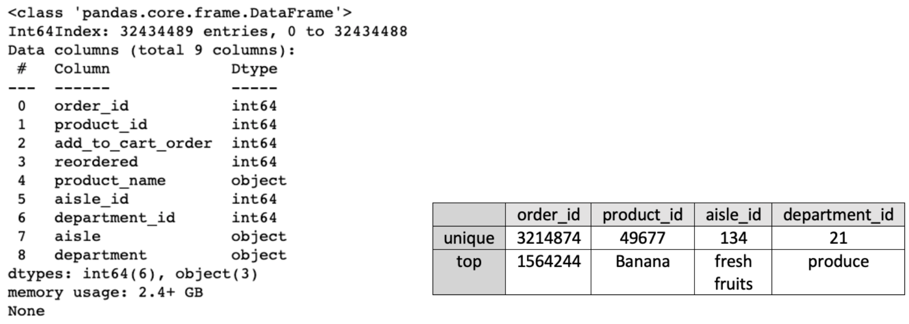
5 Discussion and Results

5.1 Datasets

The datasets were provided by Instacart Technology Company and was taken from Kaggle to perform the analysis. The datasets provided by Instacart had complete information of over 3 million grocery orders from more than 200,000

Instacart users. Both product data and customer data from Instacart includes 50,000 unique products, week and the time of purchase, different product aisle and departments[1].

Before starting on association analysis, I would do a data exploration first to understand the data types and the dimension of the original dataset. There were no null or empty values for the variables like aisle, departments, Order_product_prior, order_product_train and products datasets. The transaction details are contained in the six datasets, and 'orders_product_prior' and 'prodcuts' datasets formed the base of the complete transactions. I merged those datasets to be a single datasets group by the common product_id and order_id variables accordingly. Also, merged the aisle and departments information into the order and product master dataset for further commence analysis. Figure 1 shows the details about the master dataset.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 32434489 entries, 0 to 32434488
Data columns (total 9 columns):
 #   Column             Dtype
---  ------             -----
 0   order_id           int64
 1   product_id         int64
 2   add_to_cart_order  int64
 3   reordered          int64
 4   product_name       object
 5   aisle_id           int64
 6   department_id      int64
 7   aisle              object
 8   department         object
dtypes: int64(6), object(3)
memory usage: 2.4+ GB
None
```

	order_id	product_id	aisle_id	department_id
unique	3214874	49677	134	21
top	1564244	Banana	fresh fruits	produce

Figure 1: Information of the order and product combined dataset

5.2 Evaluation Metrics

It can be start to mining association rules after generating itemsets by Apriori. The rules generated will be look like the form $\{A\} \Rightarrow \{B\}$. Then we could use those association rules to make some recommendations to the retailer, where customers who brought item A are recommended item B. Here are three key metrics to consider when evaluating association rules.

- Measure 1: Support. It's refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item divided by total number of transactions. Suppose we want to find support for item B[7]. This can be calculated as:

$$Support = \frac{freq(A, B)}{N}$$

- Measure 2: Confidence refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought[7]. Mathematically, it can be represented as:

$$confidence = \frac{freq(A, B)}{freq(A)}$$

- Measure 3: Lift. This says how likely item B is purchased when item A is purchased, while controlling for how popular item B is. $LiftA(\Rightarrow B)$ refers to the increase in the ratio of sale of B when A is sold. $Lift(A \Rightarrow B)$ can be calculated by dividing $Confidence(A \Rightarrow B)$ divided by Support(B)[7]. Mathematically it can be represented as:

$$lift = \frac{support}{(supp(A)) \times (supp(B))}$$

For this large size of dataset, when the Apriori algorithm tied to filter rules from all possible combination of items, it would spend extremely long time to deal with the number of such combinations. Therefore, to speed up the process, the following steps need to be performed[7]:

- Set a threshold value for support and confidence. This means that we are only interested in finding rules for the items that have certain default existence (e.g. support) and have a minimum value for co-occurrence with other items (e.g. confidence).
- Extract all the subsets having higher value of support than minimum threshold.
- Select all the rules from the subsets with confidence value higher than minimum threshold.
- Order the rules by descending order of Lift.

5.3 Experimental Results

Figure 2 shows the histogram that shows the frequency of products in transactions. The graph shows us that more than 42599 products appear less than 0.001 times. Because of computer memory and running time, I will remove those products that are not appear frequently,and only search for those product occurred frequently.Therefore, I choose to generate the rules by setting support threshold to be 0.01, 0.002, 0.001 separately.

1) Rules by set support threshold 0.01, max-length of itemset is 5. It generated 11 rules totally.The Figure 3 shows the association between some products. Based on the result, we can recommend the product B to any customers who have already bought the product in A accordingly. In the association rules we found, the highest lift can be near to 3.00.

   The above is the visualization of the association rules between products. Yellow nodes represent the index of rule while green nodes represent the

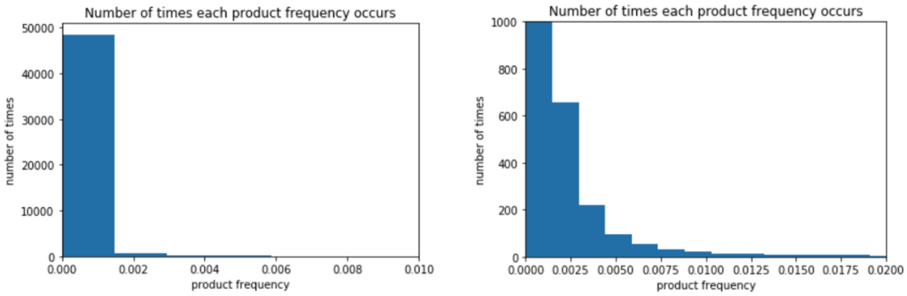
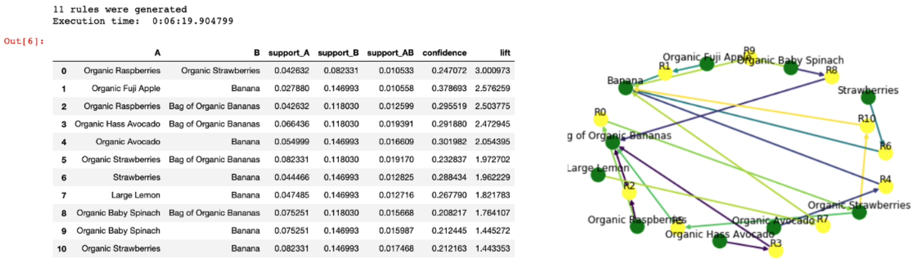


Figure 2: the frequency of products in transactions



```
11 rules were generated
Execution time:  0:06:19.904799
```

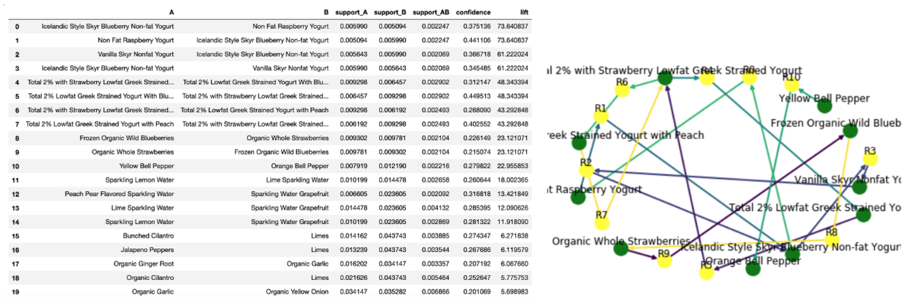
Out[6]:

	A	B	support_A	support_B	support_AB	confidence	lift
0	Organic Raspberries	Organic Strawberries	0.042632	0.082331	0.010533	0.247072	3.000973
1	Organic Fuji Apple	Banana	0.027880	0.146993	0.010558	0.378693	2.576259
2	Organic Raspberries	Bag of Organic Bananas	0.042632	0.118030	0.012599	0.295519	2.503775
3	Organic Hass Avocado	Bag of Organic Bananas	0.066436	0.118030	0.019391	0.291880	2.472945
4	Organic Avocado	Banana	0.054999	0.146993	0.016609	0.301982	2.054395
5	Organic Strawberries	Bag of Organic Bananas	0.082331	0.118030	0.019170	0.232837	1.972702
6	Strawberries	Banana	0.044466	0.146993	0.012825	0.288434	1.962229
7	Large Lemon	Banana	0.047485	0.146993	0.012716	0.267790	1.821783
8	Organic Baby Spinach	Bag of Organic Bananas	0.075251	0.118030	0.015668	0.208217	1.764107
9	Organic Baby Spinach	Banana	0.075251	0.146993	0.015987	0.212445	1.445272
10	Organic Strawberries	Banana	0.082331	0.146993	0.017468	0.212163	1.443353

Figure 3: Rules by set support threshold 0.01

itemset. The edge out from nodes are antecedents and get into nodes are consequences[8]. I used the same way of visualization in the following parts. And we can obviously see the presence of Banana and organic foods. Then we might know that customers tend to take organic food together with Banana in their purchase.

2) Rules by set support threshold 0.002, max-length of itemset is 3.



	A	B	support_A	support_B	support_AB	confidence	lift
0	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt	0.005990	0.005094	0.002247	0.375136	73.640837
1	Non Fat Raspberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0.005094	0.005990	0.002247	0.441106	73.640837
2	Vanilla Skyr Nonfat Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0.005643	0.005990	0.002069	0.366718	61.222024
3	Icelandic Style Skyr Blueberry Non-fat Yogurt	Vanilla Skyr Nonfat Yogurt	0.005990	0.005643	0.002069	0.345485	61.222024
4	Total 2% with Strawberry Lowfat Greek Strained...	Total 2% Lowfat Greek Strained Yogurt With Blu...	0.009298	0.006457	0.002902	0.312147	48.343394
5	Total 2% Lowfat Greek Strained Yogurt With Blu...	Total 2% with Strawberry Lowfat Greek Strained...	0.006457	0.009298	0.002902	0.449513	48.343394
6	Total 2% with Strawberry Lowfat Greek Strained...	Total 2% Lowfat Greek Strained Yogurt with Peach	0.009298	0.006192	0.002493	0.268090	43.292848
7	Total 2% Lowfat Greek Strained Yogurt with Peach	Total 2% with Strawberry Lowfat Greek Strained...	0.006192	0.009298	0.002493	0.402552	43.292848
8	Frozen Organic Wild Blueberries	Organic Whole Strawberries	0.009302	0.009781	0.002104	0.226149	23.121071
9	Organic Whole Strawberries	Frozen Organic Wild Blueberries	0.009781	0.009302	0.002104	0.215074	23.121071
10	Yellow Bell Pepper	Orange Bell Pepper	0.007919	0.012190	0.002216	0.279822	22.955853
11	Sparkling Lemon Water	Lime Sparkling Water	0.010199	0.014478	0.002658	0.260644	18.002365
12	Peach Pear Flavored Sparkling Water	Sparkling Water Grapefruit	0.006605	0.023605	0.002092	0.316818	13.421849
13	Lime Sparkling Water	Sparkling Water Grapefruit	0.014478	0.023605	0.004132	0.285395	12.090626
14	Sparkling Lemon Water	Sparkling Water Grapefruit	0.010199	0.023605	0.002869	0.281322	11.918090
15	Bunched Cilantro	Limes	0.014162	0.043743	0.003885	0.274347	6.271838
16	Jalapeno Peppers	Limes	0.013239	0.043743	0.003544	0.267686	6.119579
17	Organic Ginger Root	Organic Garlic	0.016202	0.034147	0.003357	0.207192	6.067660
18	Organic Cilantro	Limes	0.021626	0.043743	0.005464	0.252647	5.775753
19	Organic Garlic	Organic Yellow Onion	0.034147	0.035282	0.006866	0.201069	5.698983

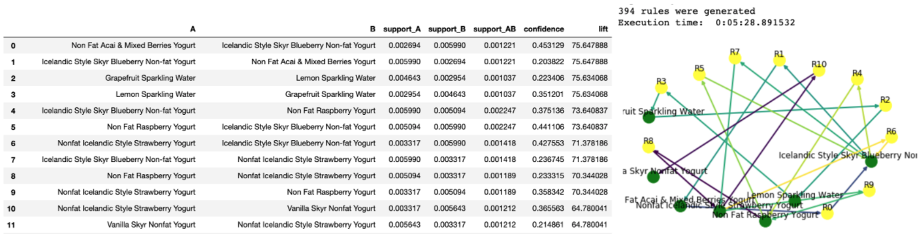
al 2% with Strawberry Lowfat Greek Strained Yogurt
R4
R0
R10
R6
Yellow Bell Pepper
R1
Frozen Organic Wild Blueb
reek Strained Yogurt with Peach
R3
R2
Vanilla Skyr Nonfat Yo
t Raspberry Yogurt
Total 2% Lowfat Greek Strained Yo
R7
R8
Organic Whole Strawberries
Icelandic Style Skyr Blueberry Non-fat Yogurt
R9
Orange Bell Pepper
R5

Figure 4: Rules by set support threshold 0.002

This time it generated 295 rules. Here the lift and the confidence show relatively higher values. If we analyze the first 20 rules we can notice that the proportion of one product occurring, and the other product also occurring is pretty high, which showing a close relationship between each other.

Rules 11 to 14 show the customer's preference for water, that is, show the customer's buying habits. When one product is purchased, the probability of another being purchased is also high. This shows that consumers may not only like to drink the same kind of water.

3) rules by set support threshold 0.001, max-length of itemset is 2



	A	B	support_A	support_B	support_AB	confidence	lift
0	Non Fat Acai & Mixed Berries Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0.002694	0.005990	0.001221	0.453129	75.647888
1	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Acai & Mixed Berries Yogurt	0.005990	0.002694	0.001221	0.203822	75.647888
2	Grapefruit Sparkling Water	Lemon Sparkling Water	0.004643	0.002954	0.001037	0.223406	75.634068
3	Lemon Sparkling Water	Grapefruit Sparkling Water	0.002954	0.004643	0.001037	0.351201	75.634068
4	Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt	0.005990	0.005094	0.002247	0.375136	73.640837
5	Non Fat Raspberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0.005094	0.005990	0.002247	0.441106	73.640837
6	Nonfat Icelandic Style Strawberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0.003317	0.005990	0.001418	0.427553	71.378186
7	Icelandic Style Skyr Blueberry Non-fat Yogurt	Nonfat Icelandic Style Strawberry Yogurt	0.005990	0.003317	0.001418	0.236745	71.378186
8	Non Fat Raspberry Yogurt	Nonfat Icelandic Style Strawberry Yogurt	0.005094	0.003317	0.001189	0.233315	70.344028
9	Nonfat Icelandic Style Strawberry Yogurt	Non Fat Raspberry Yogurt	0.003317	0.005094	0.001189	0.358342	70.344028
10	Nonfat Icelandic Style Strawberry Yogurt	Vanilla Skyr Nonfat Yogurt	0.003317	0.005643	0.001212	0.365563	64.780041
11	Vanilla Skyr Nonfat Yogurt	Nonfat Icelandic Style Strawberry Yogurt	0.005643	0.003317	0.001212	0.214861	64.780041

```
394 rules were generated
Execution time:  0:05:28.891532
```

Figure 5: Rules by set support threshold 0.001

In this final example, it generate 394 rules, and there is a high frequency of yogurt. Rules 2 and 3 bring us back to the waters but the yorgutes show a similar relationship to the previous case, consumers tend to always catch a yorgute if they have already caught another one. Here the lift is even greater than in the example shown above, the chances of these products being taken together are even higher.

6 Conclusion

Because of limitations of the computer memory, it is hard to run more examples. However, it is likely to notice some certain patterns of purchases such as organic products accompanied by bananas. Besides, cold product like yorgut which are often bought together. The product from same department also have a pattern such as organic food products always bought together with other food products. For instance, flavored waters always bought together. According to the last rules, banana seems appear frequently and tend to be the first product to be added to the cart. It could be predict that the market has a great demand for organic and healthy food products. Like it shown in the rule tables, those orders and major part in the market basket.

6.1 Directions for Future Work

The Apriori algorithm of association rule mining is the one that boosted data mining research, however, it has a fatal bottleneck that is the process of generate candidate requires multiple many times passes over the source data. The further learning direction is to explore more algorithm that can overcome this shortfall.

References

[1] Datasets: https://www.kaggle.com/c/instacart-market-basket-analysis/notebooks

[2] Agrawal, Rakesh Imielinski, Tomasz Swami, Arun. (1993). Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference. 10.1145/170036.170072.

[3] A.A. Raorane, R.V. Kulkarni, B.D. Jitkar, Association Rule – Extracting Knowledge Using Market Basket Analysis, Research Journal of Recent Sciences, 1 (2) (2012), pp. 19-27

[4] Yildiz, Baris Ergenç, Belgin. (2010). Comparison of two association rule mining algorithms without candidate generation.

[5] Kotsiantis, Sotiris Kanellopoulos, D.. (2005). Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering. 32. 71-82.

[6] Website - Margaret Rouse, Basic understanding of Market basket analysis. Retrieved from https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis

[7] Website - Association Rule Mining via Apriori Algorithm https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

[8] Website - How to Create Data Visualization for Association Rules in Data Mining. Retrieved from https://intelligentonlinetools.com/blog/2018/02/10/how-to-create-data-visualization-for-association-rules-in-data-mining/

[9] Website - Apriori: Association Rule Mining In-depth Explanation https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6