Data Mining Transactions At a Bakery

Brandon Taylor and **Renato Stoco**
477: Data Mining
Western Washington University

Abstract

This paper reports on our research about bakery shop items. The bakery has 40 baked goods, 10 beverages, and over 70,000 receipts available to use. We will be using our data mining knowledge to try to infer interesting rules or classifications that would lead us to suggesting clever ways that the bakery could improve its marketing strategy to better sell their products or to have a better understanding of their customers and transaction patterns.

Introduction

The extended bakery shop dataset was chosen for our research because it contained sales receipts of three different sizes: 5000, 20000 and most importantly, 75000. In addition, the bakery has 40 baked goods and 10 beverages, which seemed like a good variety of products without being so large as to cause the data to be too sparse. Our goal has been to implement and use the Apriori algorithm to mine association rules about what goods and beverages are most often purchased in the same transaction. We first experimented with the raw data and then applied different binning techniques to improve our results. We then went on to explore Weka, a data mining tool, which allowed us to confirm our results conducted with our self implemented code as well as experiment with the data in ways we wouldn't have had time to do otherwise. This further research included exploring several clustering techniques where we attempted to form meaningful groups of customers based on the type of products they purchased.

Background

Apriori

We implemented our own version of apriori in python after reading the chapter and following the instructions in the book, Machine Learning in Action by Perter Harrington [2].

Weka 3

A collection of machine learning algorithms for various types of data mining. Complete with a GUI interface and visualization, this tool helped us interpret and confirm our results. The most difficult part about dealing with Weka was just that it had a very specific data format (.arff) that we had to import the data in which lead to some difficulties.

Bakery Data Analysis

In order to get good results from the apriori algorithm, after several trials, we decided to set the Support and Confidence levels to 5% and 60% respectively.

Support: 0.05 Confidence: 0.6

For the first round of association rule exploration, we applied apriori to the dataset without any binning. Unfortunately, with so many products in the dataset, the transactions were too sparse and didn't result in any rules generated.

No binning: no rules generated

Then we realized that it would be better to group up all the subcategories of baked goods and beverages into their “type” category. This translated our data from having 50 products to just 18.

Bin by general category:
5000 receipts
['frap'] --> ['tart'] conf: 0.7220
['bottled water'] --> ['tart'] conf: 0.7227
['croissant'] --> ['tart'] conf: 0.6059
['danish'] --> ['tart'] conf: 0.7151

20000 receipts
['frap'] --> ['tart'] conf: 0.7348
['bottled water'] --> ['tart'] conf: 0.7297
['danish'] --> ['tart'] conf: 0.7041
['espresso'] --> ['tart'] conf: 0.710

75000 receipts
['frap'] --> ['tart'] conf: 0.7247
['bottled water'] --> ['tart'] conf: 0.7347
['danish'] --> ['tart'] conf: 0.7048

After binning the data, we were able to generate some meaningful rules. By taking a look at the results, we can infer that the customers are much more likely to buy tarts if they buy a danish, bottled water, or frappuccino. After these results, we considered how the rules might change if we altered some of our binning techniques. So, in addition to the general binning we had already implemented, we categorized all ten of the drinks into two categories, hot and cold.

Bin by beverages (hot/cold) + general binning:
5000 receipts
['danish'] --> ['tart'] conf: 0.7151
['croissant'] --> ['tart'] conf: 0.6059

20000 receipts
['danish'] --> ['tart'] conf: 0.7041

75000 receipts
['danish'] --> ['tart'] conf: 0.7048

Unfortunately, and contrary to our intuition, this resulted in fewer rules than were generated without drink binning. Next, by taking a close look at the results generated by both binning techniques, tarts really dominate the results. So, we thought that changing all the tarts in the main category “Tart” back to their sub tart categories might lead to seeing something different.. Specifically, it could be the case that one type of tart is dominating the transaction sales but with binning applied, is resulting in all tarts appearing “popular.

Bin by beverages (hot/cold) + general binning - tart binning:
5000 receipts
['Cherry tart'] --> ['cake'] conf: 0.6760
['Cherry tart'] --> ['danish'] conf: 0.6152
['Blueberry tart'] --> ['croissant'] conf: 0.6619

20000 receipts
['Lemon tart'] --> ['cake'] conf: 0.6813
['Cherry tart'] --> ['cake'] conf: 0.6526
['Berry tart'] --> ['cold beverage'] conf: 0.6032
['Cherry tart'] --> ['danish'] conf: 0.6257
['Blueberry tart'] --> ['croissant'] conf: 0.6084

75000 receipts
['Lemon tart'] --> ['cake'] conf: 0.6824
['Cherry tart'] --> ['cake'] conf: 0.6539
['Berry tart'] --> ['cold beverage'] conf: 0.6020
['Cherry tart'] --> ['danish'] conf: 0.6201
['Blueberry tart'] --> ['croissant'] conf: 0.6333

Now, we are back to getting interesting results, and oddly enough, tarts are no longer the consequent, but instead the initial product. In addition, if we compare, 20000 receipt data and the 75000 receipt data, they match up. Furthermore, with the rules no longer being skewed by the tart category we can do some predictions of what the customers are going to buy together. This could lead to useful business

analytics that could hint at what sort of products to advertise together or put on sale. For example, Knowing that people who buy Cherry tarts are often going to buy a cold beverages as well, the bakery can increase the price of the cold beverage and decrease at the same time the price of the berry tart, and vice versa.

Clustering

Clustering is an unsupervised learning technique that allow us to perform a classification task even though we do not have a labeled data. This research will use Weka clustering algorithms with the focus on two different clustering algorithms: CLOPE and Hierarchical. They were chosen because not only they work well on transactional data(As some articles describe) but after some empirical test they proved to be the ones to pursue further.

As described in CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data [1] this research will be using repulsion of approximately 1.2. After running CLOPE on 75,000 transactions we got 6 different clusters with the first four being more dense than the other two.

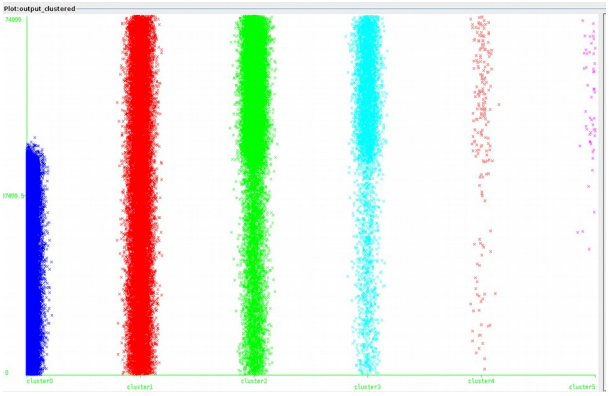

Illustration 1: Illustration 1: CLOPE Clusters with 75,000 transactions

```
Time taken to build model (full training data) : 0.58 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      25541 ( 34%)
1      24230 ( 32%)
2      19780 ( 26%)
3       5205 (  7%)
4        193 (  0%)
5         51 (  0%)
```

Weka allows us to see which transaction in the plot above contains the desired product. And, by using this tool we manually extracted and annotated the information from the plot figures, by so generating a table containing the potential customers based on the clusters and their products in common.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cake	Cake	Cake	Cake	Tart	Tart
Tart	Tart	Tart	Eclair	Twist	Lemonade
Cookie	Cookie	Cookie	Tart	Water	Water
Danish	Juice	Meringue	Pie	Espresso	Espresso
Lemonade	End	Danish	Twist	End	End
Coffee		Bearclaw	Juice		
End		⋮	⋮		

We can observe that clusters number 2 and 3 have so many items contained that its probably not very useful for good predictions.

For comparison, we observed results from 20,000 transactions with the Hierarchical Clustering Algorithm. 75,000 transactions were not used in this research because it costs too much memory due to its high time complexity, $O(n^3)$. Nevertheless, the results from this algorithm were more evenly distributed than the results of the CLOPE algorithm.

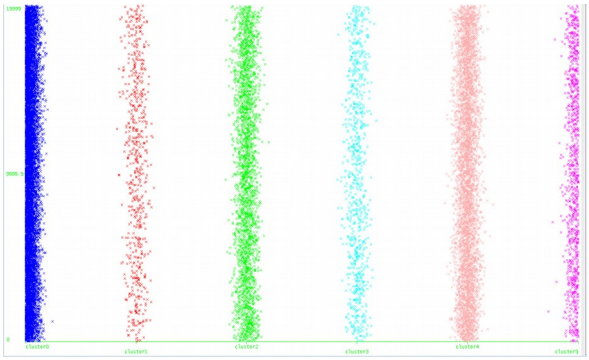

Illustration 2: Hierarchical Cluster with 20,000 transactions

```
Time taken to build model (full training data) : 933.86 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9763 ( 49%)
1        628 (  3%)
2       2657 ( 13%)
3        867 (  4%)
4       5110 ( 26%)
5        975 (  5%)
```

The above plot shows a better composition of clusters and that they are more balanced in terms of density. With all these results, we can then compare which one is better at predicting potential costumer types based on the clusters.

Potential Customer 1	Potential Customer 4	Potential Customer 5
Cake	Tart	Tart
Tart	Twist	Lemonade
Cookie	Water	Water
Juice	Espresso	Espresso

Illustration 3: Potential Customer Groups based on CLOPE

Future Work

In the future, we would like to find a way to produce results for the set of 75,000 transactions using the hierarchical clustering algorithm. We believe this would cluster the data better than CLOPE and would produce more meaningful results than the 20,000 transaction dataset that we are able to currently perform hierarchical clustering on.

Conclusion

In this research we used some of the well known data mining concepts and algorithms to better understand and predict information regarding this bakery dataset. After some preprocessing and binning to better classify the types of items contained in the dataset we found rules suggesting that the customers are much more likely to buy tarts. However, after altering our binning techniques a bit we were able to collect more meaningful rules, concluding that cold beverages will most likely be bought together with berry tarts, and that lemon and cherry tarts will be bought together with cakes.

Furthermore, after clustering the data using two different techniques we found that customers can be grouped into several meaningful customer types. Although, some clusters were quite “noisy” and didn't create a very unique customer profile, at least 3 or 4 meaningful profiles were developed from each study.

All of this data can be turned into useful marketing strategies that could help the bakery shop see more profit and target specific customer profiles.

References

[1] Yiling Yang, Xudong Guan, and Jinyuan You. *CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data*. Shanghai, China 2002. Web.

[2] Harrington, Peter. *Machine Learning in Action*. Shelter Island, NY: Manning Publications, 2012. Print.