Data Mining Project 2019

Clustering Market Baskets and Recipes data

Nicolo' Merzi
205150
Data Science 2nd year
"Do not know yet"
nicolo.merzi@studenti.unitn.it

Francesco Battista
214224
Data Science 2nd year
"Do not know yet"
francesco.battista@studenti.unitn.it

1. Introduction

This project's aim was to use data mining techniques to identify groups of customers based on the recipes they could prepare given what they bought at a supermarket. This is a challenging problem for many reasons. For starter, the same ingredients can be used for different recipes or for the same recipe different ingredients may be used. Also, a customer may only buy some of the needed ingredients because they have the rest at home. This makes it unclear exactly what recipe a person could make given that they bought a specific ingredient.

The input was a collection of supermarket transactions and a collection of different recipes. In order to be able to group customers that eat similarly based on what they buy, the unsupervised learning technique of K-means was used. This technique, using sparse ingredient feature vectors, was applied on the Recipes dataset, and then used to predict each transaction of the supermarket collection. The results showed that K-means was able to isolate specific types of macro-recipes, which can then be used to group customers together based on what they bought at the supermarket.

2. Related Work

There have been lots of separate studies on Market Basket data and on Recipes data. Many supermarket companies use data mining techniques to find structures, patterns and to discover important information [1]. Extensive research has also been done on recipes data to better understand the nature and composition of recipes. One such work is Learning to Cook: An Exploration of Recipes Data [2] in which they figure out different "types" of recipes based purely on what ingredients were included.

In this work we aim to combine the two different types of information: The Market Basket data and the Recipes data to group customers based on what they could eat given what they bought.

3. Problem Statement

As we said before, we aim to identify groups of customers that eat similarly based on the items they buy at the supermarket. We used two different datasets: A Market Basket dataset [3] containing 9835 transactions with 163 unique items, and a Recipes dataset [4] containing over 40.000 recipes with more than 6.000 unique ingredients. Extensive data processing and cleaning was done to make both datasets somewhat comparable. The end goal is to find several groups or clusters that represents a certain "type" of customer.

Most clustering algorithms work with some sort of distance or similarity measure between different observation to group them together. This can be problematic since we are dealing with lists of ingredients and it is hard to define a mathematical measure of distance between this type of data.

Another typical problem in clustering and generally in unsupervised settings is to assign/discover the meaning of the clusters found. Once the clustering is done, we are only left with an ID that identifies the different clusters and we need to interpret those clusters and give them some sort of meaning. What we want is a human level understanding of each cluster.

4. Solution

The datasets used in this paper have similar structure. Both the Market Basket dataset and the Recipes dataset are made of lists of ingredients, in the first case we have the items bought at the supermarket while in the second dataset we have a list of ingredients for each recipe.

Further processing was done on both datasets. Starting with the Recipes dataset, a dictionary of the 648 most common ingredients that occur in a least 100 recipes was used to clean the dataset. Each recipe was filtered using those word so that ingredients such as "chicken wings" and "condensed milk" were reduced to a simpler form such as "chicken" and "milk". Moreover, ingredients

that belonged in similar food groups were grouped together to match the same "language" used in the Market Basket dataset. This means that instead of having the specific ingredient used for that recipes, we substituted it with a more general description, for example instead of having "carrot" and "potato" we now have "vegetables". This was necessary so that the features of both datasets would match since the Market Basket dataset has much fewer ingredients.

In the end, each recipe was represented in a $\mathbb{R}^{34}$ binary vector, where the element in index *i* is 1 if ingredient *i* is present in the recipe, and 0 otherwise. This matrix represents the features that we used to train our model. The final dimension of this feature matrix was 39.599 x 34, where 39.599 is the total number of recipes after the cleaning procedure and 34 are the macro-ingredients used to describe each recipe.

K-means clustering was performed on these binary vectors representing each recipe in the Recipes dataset. In order to find the optimal cluster size, K-means was run with value of *K* between 2 and 20. The results in *Figure 1* shows no distinctive elbow meaning that there may not exist an optimal value to cluster recipes.

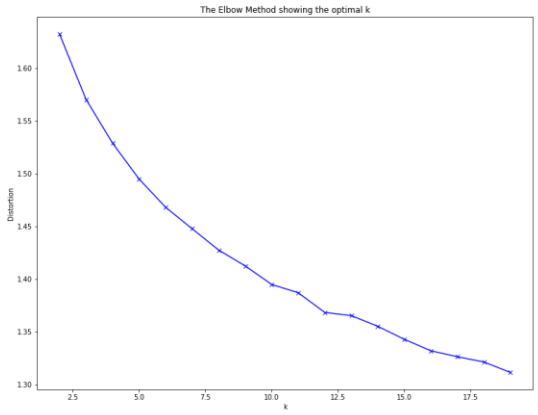

Figure 1: SSD of K-Means algorithm at different K.

By looking at the top ingredients for each cluster we can identify some macro-areas of cuisine types. For example, we may assume that the top ingredients of the following clusters represent respectively the "Dessert/Baked food cluster" and the "Fried dish cluster".

- Dessert/Baked food Cluster: flour, salt, butter, sugar, eggs, milk, water, fruits, oil, chocolate
- Fried dish Cluster: flour, oil, salt, herbs, butter, vegetables, eggs, chicken, bread, beer

Even though no single ideal clustering of recipes was found, analyzing the obtained clusters at different values of *K* showed interesting results: By increasing the number of clusters, the algorithms generalize less and less and "discovers" more specific types of cuisines.

A similar procedure was also done on the Market Basket dataset. Each ingredient was filtered in the previous step so that features would match between datasets. So, for example "long grain rice" was transformed in "rice", while all non-edible ingredients were removed from the dataset. Each recipe was then converted in a binary vector with the same features as the matrix used to train the model. The final dimension of the feature matrix of the Market Basket collection was 8.973 x 34, where 8.973 is the total number of transactions after the cleaning procedure and 34 are the same macro-ingredients used before.

Finally, the K-means model was used to predict in which cluster each transaction in the Market Basket dataset would belong with the effect of grouping customers based on similar macro culinary areas.

4.1 Algorithm (after cleaning procedure)

1. Convert each recipe in the Recipes dataset in binary vectors using as features the common ingredients between the Market Basket dataset and the Recipe dataset.
2. Apply K-means on the resulting matrix of binary vectors of the Recipe dataset.
3. Convert each transaction in the Market Basket dataset in binary vectors with the same features as before.
4. Using the previously trained model, predict cluster membership for each transaction in the Market Basket dataset.

5. Experimental Evaluation

Since there is no single optimal number of clusters to group recipes and by analyzing the top ingredients for each cluster for each k, we arbitrary decided to use the k-means model considering k=7. This clustering seemed to offer a good middle ground between group separation and ease of interpretation. Below are reported the top 10 ingredients for each of the seven clusters:

- Cluster 0: water, vegetables, herbs, salt, oil, chicken, sugar, wine, pasta, sauce
- Cluster 1: chicken, pork, salt, ham, beer, rum, beef, brandy, herbs, bread
- Cluster 2: flour, oil, salt, herbs, butter, vegetables, eggs, chicken, bread, beer
- Cluster 3: cheese, vegetables, herbs, salt, butter, sauce, cream, chicken, eggs, bread
- Cluster 4: flour, salt, butter, sugar, eggs, milk, water, fruits, oil, chocolate

- Cluster 5: vegetables, fruits, oil, sauce, butter, salt, water, milk, flour, pasta, bread
- Cluster 6: pasta, water, salt, rice, vegetables, herbs, cheese, chicken, pork, cream

By looking at the top ingredients for each cluster, we can identify some general trends in the composition of the different clusters. For example, Cluster 0 may group together all types of broths, so both vegetable and stock. Cluster 1 with almost all ingredients being meats or some type of alcoholic beverage may group together barbeque recipes. Cluster 2 may refer to some type of fried dish due to the presence of all ingredients used when frying (flour, eggs, salt, bread, oil, butter). Cluster 3 may group together dishes that have cheese as their main ingredient. Cluster 4 may refer to some type of dessert or baked food. Cluster 5 may group all vegetarian recipes since there is not a single meat in the top terms of this cluster and the first two ingredients are vegetables and fruits. And lastly, Cluster 6 may refer to a first course, such as a plate of pasta or rice with some vegetables and/or meats.

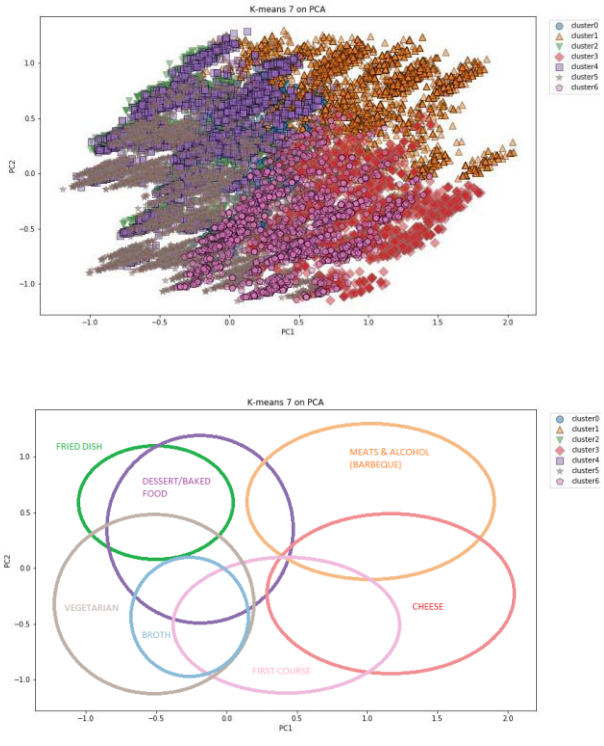

Figure 2-3: K-Means (7) results on PCA and highlighted clusters

Figure 2-3 shows the results of k-means projected on the 1st and 2nd principal component. As we can see from the image, there is a good separation between clusters. The distance between each cluster with all other clusters seems to validate our interpretation of them: If we consider Cluster 2 and 4, respectively the cluster of fried dished and the cluster of desserts/baked foods, we can see that they are almost overlapping, this makes sense since ingredients such as flour, salt, eggs, butter and oil are used extensively in both areas. In contrast, if we consider cluster 1 and 5, respectively the cluster of meats & alcohol and the vegetarian cluster are very far apart.

This model was then used to predict cluster membership for each transaction in the Market Basket dataset. Reported below are some example of the predicted transactions:

Items bought	**Items filtered**	**Predicted Cluster**
[whole milk]	[milk]	4 – Dessert/Baked food
[sausage, rolls/buns, soda, canned bear, specialty bar, shopping bags]	[pork, bread, soda, beer]	1 – Meats & Alcohol (Barbeque)
[flour, salt, bottled water, other vegetables, bottled beer]	[flour, salt, water, vegetables, beer]	2 – Fried dish
[root vegetables, butter, curd, whipped/sour cream, UHT-milk, hard cheese, rolls/buns, bottled water, long life bakery product]	[vegetables, butter, cream, milk, cheese, bread, water]	3 – Cheese
[chicken, hamburger meat, citrus fruits, root vegetables, butter, pasta, oil, detergent]	[chicken, beef, fruits, vegetables, butter, pasta, oil]	6 – First course
[rolls/buns]	[bread]	5 – Vegetarian

Most predictions seem to be accurate with respect to our interpretation of the clusters. Transaction with few ingredients are the “hardest” to predict due to some ingredients begin shared by multiple clusters. Take the last transaction in the table above as an example: if a customer buys rolls/buns (bread), the algorithm puts

this person in the Vegetarian cluster, but it could belong to basically any other cluster with as much sense.

K-means was trained on the entire Recipe dataset which consists of over 40.000 recipes (approximately 11 MB), in just under 3 seconds. To test the scalability of the algorithm, successive runs were performed with 30%, 50%, 200% and 400% the total number of recipes of the original dataset (respectively 3.3, 5.5, 22.2 and 44.4 MB). *Figure 4* below shows the running time of the algorithm considering the different sizes. As we can see from the graph, the algorithm behaves roughly linearly with the increase in size of the dataset.

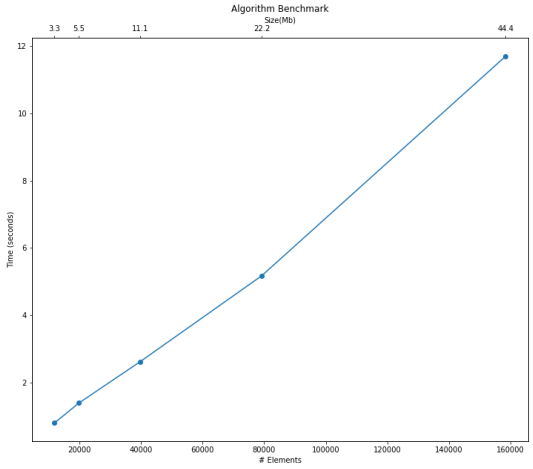

Figure 4: Algorithm running time at different dataset size.

6. Conclusions

The K-Means clustering technique proved to be effective in discovering the underlying structure of recipes based on ingredients and in predicting cluster membership given items bought at the supermarket.

After an extensive cleaning process, both datasets were converted in binary matrices using as features the common ingredients between them. K-Means was then applied on the resulting matrix of the Recipes dataset where we identified some general trends in the composition of recipes. The model was then used to predict each transaction of the Market Basket dataset with the effect of grouping customers based on recipes but given the ingredients bought at the supermarket.

The results are dependent on the number of clusters: we can have a lower number of cluster and have more general "type" of recipes or we can have a greater number of clusters and have more specific "type" of recipes. Testing K-Means with different *K* values seems to indicate that it does not exist an optimal number to cluster recipes. We choose *K*=7 because it seemed to produce consistent results, but further research could be done trying different number of clusters.

Another important aspect that has a significant impact on the algorithm is the data cleaning procedure since we have done extensive data manipulation to have matching feature between datasets. Improvements could be done on this part, either by using different datasets or by constructing a better data cleaning process to lose as little information as possible.

REFERENCES

[1] Data Mining in Supermarket: A Survey. (2017). D. Kaur, J. Kaur. *International Journal of Computational Intelligence Research*.
[2] Learning to Cook: An Exploration of Recipes Data. (2016). T. Arffa, R. Lim, J. Rachleff.
[3] Groceries Market Basket Dataset. https://www.kaggle.com/irfanasrullah/groceries
[4] Recipes Ingredients Dataset. https://www.kaggle.com/kaggle/recipe-ingredients-dataset
[5] NLTK package. WordNet Lemmatizer.

[6] List of culinary fruits: https://en.wikipedia.org/wiki/List_of_culinary_fruits
[7] List of vegetables: https://en.wikipedia.org/wiki/List_of_vegetables
[6] List of culinary herbs and spices: https://en.wikipedia.org/wiki/List_of_culinary_herbs_and_spices