university of groningen / faculty of science and engineering

Information System
Lab 5
Associative Rule Mining

T. Back s3218147
R. Bwana s3232352

18th January, 2018

1 Introduction

In this weeks' lab, a look is taken on associative rule mining. On the basis on the provided list of groceries transactions, we should identify which products are often bought together.

In the preprocessing, we converted the file groceries.txt into an actual groceries.csv file to ease handling it. This was done by opening the file in Microsoft Excel and save it again as a csv file. This new file is included in our submission folder.

2 Histogram

Within the grocery transaction file, we found 179 unique items. The most common item is whole milk with 2513 occurrences followed by other vegetables (1903), rolls (1809) and soda (1715). The histogram also reveals around 20-30 items are also frequently sold and other items are almost never sold. The complete histogram is shown in Figure 1. Due to the large amount of items, the labels of the individual item are difficult to read.

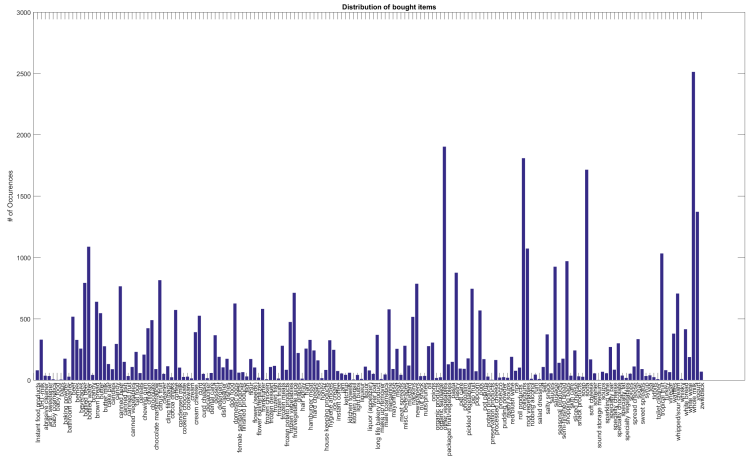

Figure 1: Histogram of all items

3 Associative Rule Mining

The apriori algorithm uses, as in its name, the apriori principle. This proposes that if an itemset, rule, or combination of rules are not common, as measured by support and confidence, than any subsequent set based on these need not be considered. This allows for a much faster calculation of rules than were all possibilities to be considered.

3.1 Implementation

We implemented the Apriori algorithm in Python, the code of which can be found in our submission folder.

3.2 How many rules?

Firstly, the itemsets.

With a support value of 0.001, the algorithm produced 157 1-item itemsets, 2981 2-item itemsets, 6831 3-item itemsets, 3137 4-item itemsets, 376 5-item itemsets, and 10 6-item itemsets for a total of 13,492 itemsets.

As for the rules, taking these 13,492 itemsets and applying a minimum confidence of 0.8, we obtained 409 rules.

3.3 30 most confident rules

Sorting the list based on the returned confidence we hereby list the 30 rules with the highest confidence:

- 'sugar', 'rice' ⇒ 'whole milk'
- 'canned fish', 'hygiene articles' ⇒ 'whole milk'
- 'domestic eggs', 'soft cheese', 'butter' ⇒ 'whole milk'
- 'pip fruit', 'hygiene articles', 'root vegetables' ⇒ 'whole milk'
- 'domestic eggs', 'sugar', 'curd' ⇒ 'whole milk'
- 'pip fruit', 'whipped/sour cream', 'brown bread' ⇒ 'other vegetables'
- 'whipped/sour cream', 'hygiene articles', 'root vegetables' ⇒ 'whole milk'
- 'pip fruit', 'butter', 'hygiene articles' ⇒ 'whole milk'
- 'rice', 'butter', 'root vegetables' ⇒ 'whole milk'
- 'domestic eggs', 'cream cheese ', 'napkins' ⇒ 'whole milk'
- 'soft cheese', 'citrus fruit', 'root vegetables' ⇒ 'other vegetables'
- 'domestic eggs', 'sugar', 'cream cheese' ⇒ 'whole milk'
- 'flour', 'whipped/sour cream', 'root vegetables' ⇒ 'whole milk'
- 'tropical fruit', 'whole milk', 'yogurt', 'grapes' ⇒ 'other vegetables'
- 'tropical fruit', 'whipped/sour cream', 'butter', 'fruit/vegetable juice' ⇒ 'other vegetables'
- 'tropical fruit', 'sausage', 'rolls/buns', 'root vegetables' ⇒ 'whole milk'
- 'white bread', 'butter', 'other vegetables', 'root vegetables' ⇒ 'whole milk'
- 'newspapers', 'whole milk', 'rolls/buns', 'soda' ⇒ 'other vegetables'
- 'bottled water', 'pip fruit', 'other vegetables', 'root vegetables' ⇒ 'whole milk'
- 'yogurt', 'other vegetables', 'oil', 'root vegetables' ⇒ 'whole milk'
- 'tropical fruit', 'pip fruit', 'ham', 'whole milk' ⇒ 'other vegetables'
- 'tropical fruit', 'pip fruit', 'yogurt', 'ham' ⇒ 'other vegetables'
- 'whipped/sour cream', 'citrus fruit', 'pastry', 'rolls/buns' ⇒ 'whole milk'
- 'whipped/sour cream', 'butter', 'other vegetables', 'pork' ⇒ 'whole milk'
- 'domestic eggs', 'whipped/sour cream', 'butter', 'other vegetables' ⇒ 'whole milk'

- 'tropical fruit', 'oil', 'yogurt', 'root vegetables' ⇒ 'whole milk'
- 'tropical fruit', 'whipped/sour cream', 'citrus fruit', 'root vegetables' ⇒ 'other vegetables'
- 'root vegetables', 'tropical fruit', 'oil', 'yogurt', 'other vegetables' ⇒ 'whole milk'
- 'sugar', 'other vegetables', 'cream cheese ' ⇒ 'whole milk'
- 'tropical fruit', 'yogurt', 'sausage', 'root vegetables' ⇒ 'whole milk'

From a quick look at this list we can see that despite several combination of items, 'whole milk' can be considered a near definite buy alongside them. In fact, in the top 30 rules, whenever 'whole milk' is not bought, it is almost always 'other vegetables' which are bought instead. Overall, the items mentioned in the top 30 rules represent commonly bought grocery items, therefore they may not be standout rules which a business would not know of already.

4 Rule Measures

The result of the associative rule mining are rules. But their importance matters and so also different measures exist to provide a ranking of the different rules. Additionally, the measures oftentimes also include information about whether the found rule is well supported by the data or rather a bit better guess than just random guessing.

4.1 Support

The support measures how often a specific item X appears in all transactions. Therefore, items that are bought a lot have a high support. The formula is:

$$supp(X) = \frac{|t \in T; X \subseteq t|}{|T|}$$

4.2 Confidence

The confidence defines how often the found rule is supported by the data. A confidence of 1 implies that the rule is true for all transaction, therefore a high confidence is desirable. It is defined by:

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

4.3 Lift

The lift evaluates the independence of rules. If two items are independent of each other, then they are bought together randomly and no rule can be derived. The higher the lift, the more important the rule. Lift is composed using the support in the following way:

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}$$

4.4 Conviction

Conviction is similar to Lift. It evaluates how much more often a rule is better than pure random chance. It is defined by:

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

A value of 1 indicated pure randomness; 1.5 that the rule is 50% more often found to be correct.

4.5 Leverage

Leverage is also similar to the Lift. Instead of a division, a subtraction is used:

$$leverage(X \Rightarrow Y) = supp(X \cup Y) - supp(X) * supp(Y)$$

4.6 Comparison

All of these measures have their advantages and disadvantages and all are based on the measure of support. The support is a very simple measure. It supports common items in transactions, but hides important rules of niche items.

The confidence solves part of the mentioned issue by checking how often a rule is supported by the data. This often leads to niche items recommending common items.

Finally, the lift also takes into account a recommendation from a niche to a niche product, which are however often bought together.

In practise, often support and confidence are sufficient. Sometimes also the lift is used. Conviction and Leverage are ratio measures which allow ranking, but on an open scale (no fixed range). Therefore, it is harder to interpret individual values of these measures. Also, these measures take longer to compute.