

Step1: Load the Supermarket Dataset

Load the Supermarket dataset (*data/supermarket.arff*). This is a dataset of point of sale information. The data is nominal and each instance represents a customer transaction at a supermarket, the products purchased and the departments involved.

The data contains 4,627 instances and 217 attributes. The data is denormalized. Each attribute is binary and either has a value (“t” for true) or no value (“?” for missing). There is a nominal class attribute called “total” that indicates whether the transaction was less than \$100 (low) or greater than \$100 (high).

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | Forecast

Open fil... | Open U... | Open D... | Generat... | Undo | Edit... | Save...

Filter: Choose None [Apply]

Current relation
 Relation: supermarket
 Instances: 4627
 Attributes: 217
 Sum of weights: 4627

Attributes
 All | None | Invert | Pattern

No.	Name
1	department1
2	department2
3	department3
4	department4
5	department5
6	department6
7	department7
8	department8
9	department9
10	department10
11	department11
12	grocery misc
13	baby needs
14	bread and cake

[Remove]

Selected attribute
 Name: department1
 Missing: 3580 (77%)
 Distinct: 1
 Type: Nominal
 Unique: 0 (0%)

No.	Label	Count	Weight
1	t	1047	1047.0

Class: total (Nom) [Visualize All]

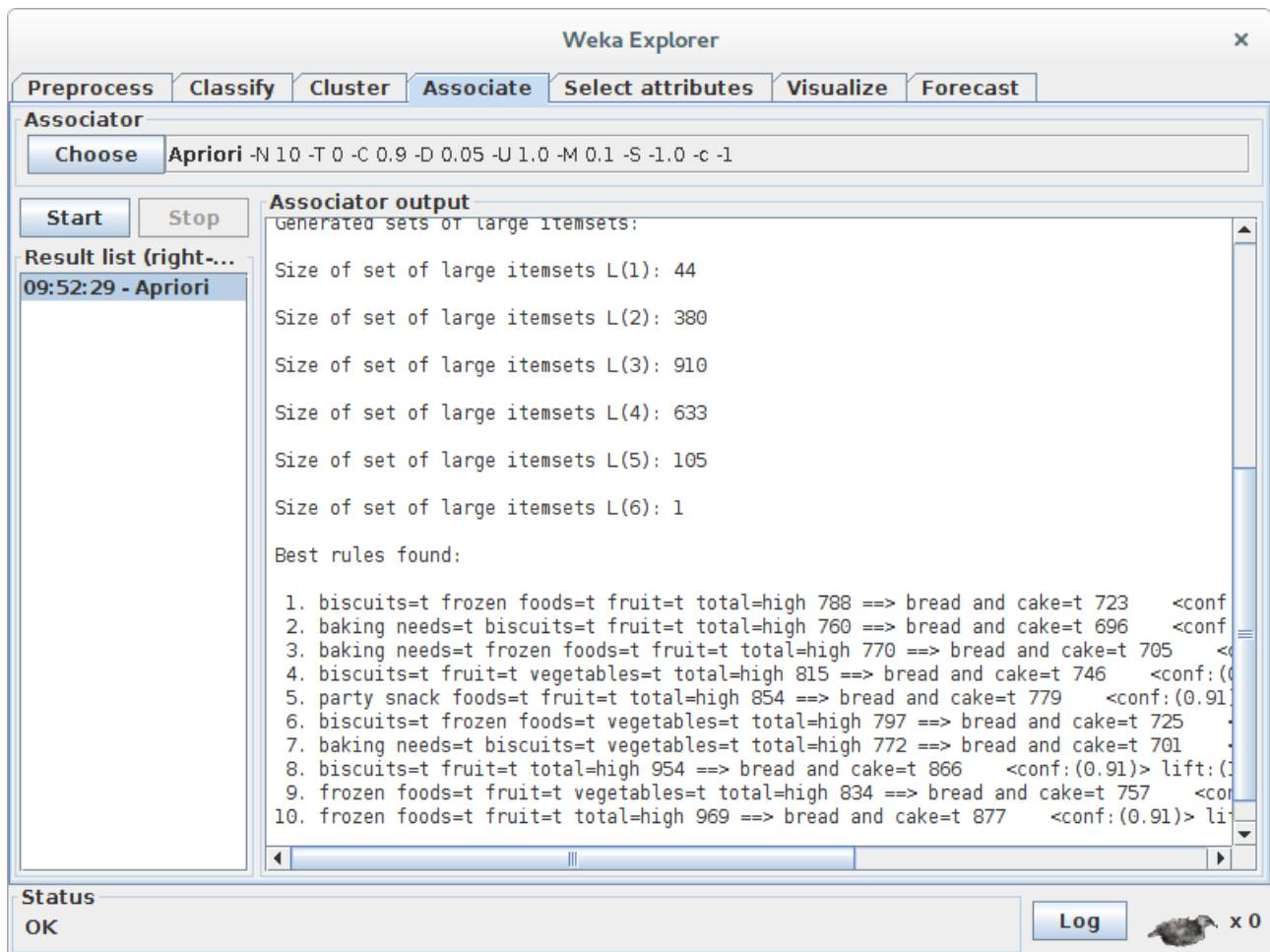
1047

Status: OK [Log] x 0

Step2: Discover Association Rules

Click the “Associate” tab in the Weka Explorer. The “Apriori” algorithm will already be selected. This is the most well known association rule learning method because it may have been the first (Agrawal and Srikant in 1994) and it is very efficient.

Click the “Start” button to run Apriori on the dataset (you can set the parameters if needed).



Step4: Observations

You can see rules are presented in antecedent => consequent format. The number associated with the antecedent is the absolute coverage in the dataset (in this case a number out of a possible total of 4,627). The number next to the consequent is the absolute number of instances that match the antecedent and the consequent. The number in brackets on the end is the support for the rule (number of antecedent divided by the number of matching consequents). You can see that a cutoff of 91% was used in selecting rules, mentioned in the “Associator output” window and indicated in that no rule has a coverage less than 0.91.

We can see that all presented rules have a consequent of “bread and cake”.

- All presented rules indicate a high total transaction amount.
- “biscuits” an “frozen foods” appear in many of the presented rules.