Subject

Computing association rule with TANAGRA, ORANGE and WEKA.

We must respect the following steps if we want to compute association rules from a dataset:

- Import the dataset;
- Select the descriptors;
- Set the parameters of the association rule algorithm i.e. the minimal support and the minimal confidence;
- Execute the algorithm and visualize the rules.

Our three packages use attribute-based dataset. Each attribute-value couple becomes an item which be used for generating rules.

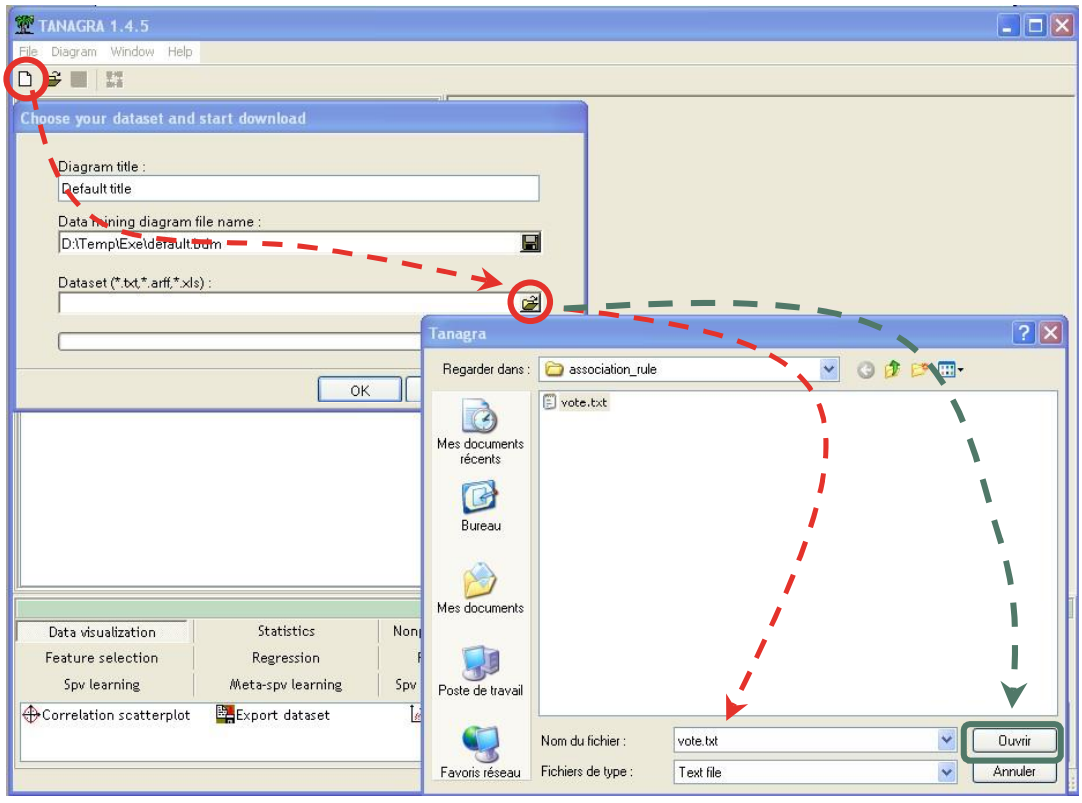
Dataset

We use the VOTE.TXT dataset from the UCI IRVINE repository.

Association rules with TANAGRA

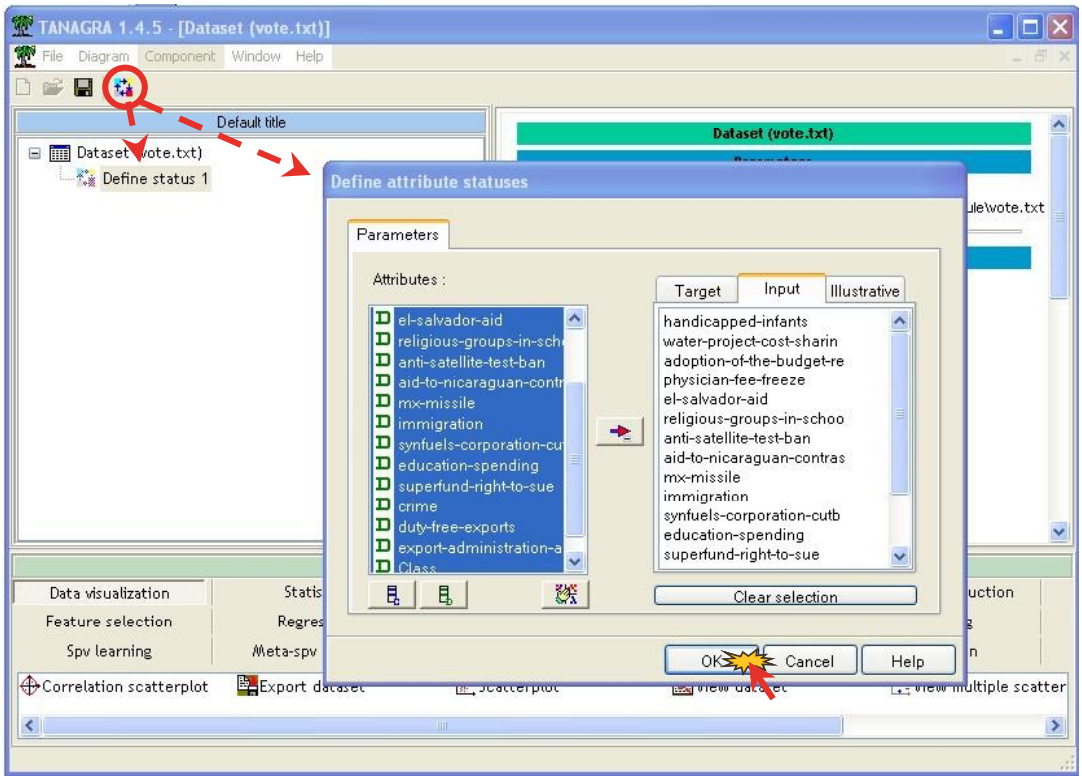
Import the dataset

First, we must create a new diagram and import the dataset with the FILE/NEW menu. We select the VOTE.TXT dataset.



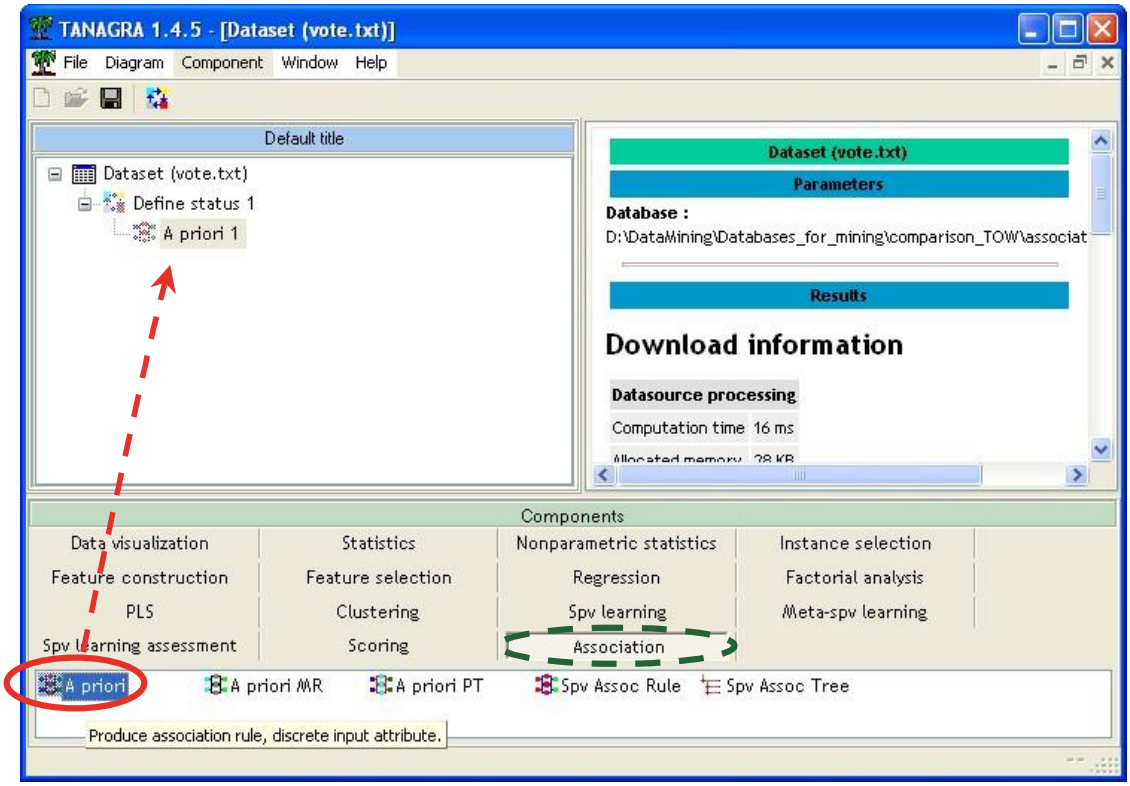
Defining the attributes for the analysis

We add a DEFINE STATUS component in the diagram; we set all attributes as INPUT.

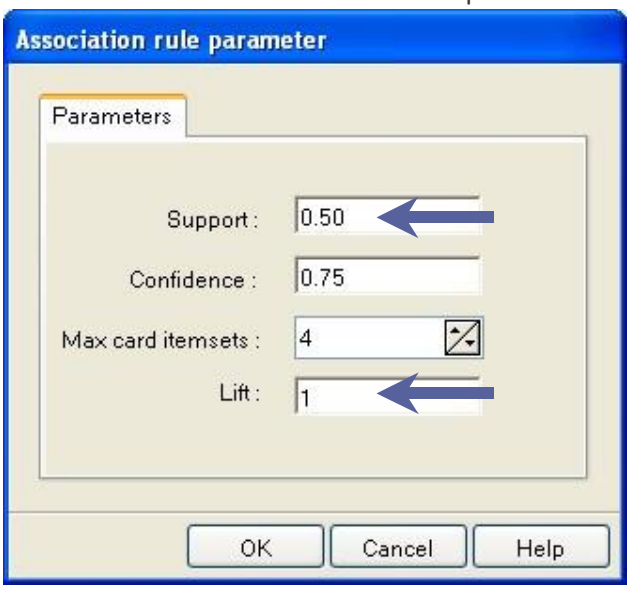


A PRIORI algorithm

There are various algorithms in TANAGRA; some of them come from external libraries. In this tutorial, we use the standard A PRIORI algorithm.



Then we click on the PARAMETERS contextual menu of the component.



The minimal support is set to 0.5; the minimal confidence to 0.75; we use only frequent itemsets of cardinal lower or equal to 4; the rules with a LIFT lower than 1 are removed.

Computing the rules

We select the contextual VIEW menu in order to see the rules. We obtain 14 rules.

Results

ITEMS

Transactions	435
Counting items	
All items	50
Filtered items	12
Counting itemsets	
card(itemset) = 2	4
card(itemset) = 3	1
Rules	
Number of rules	**14**

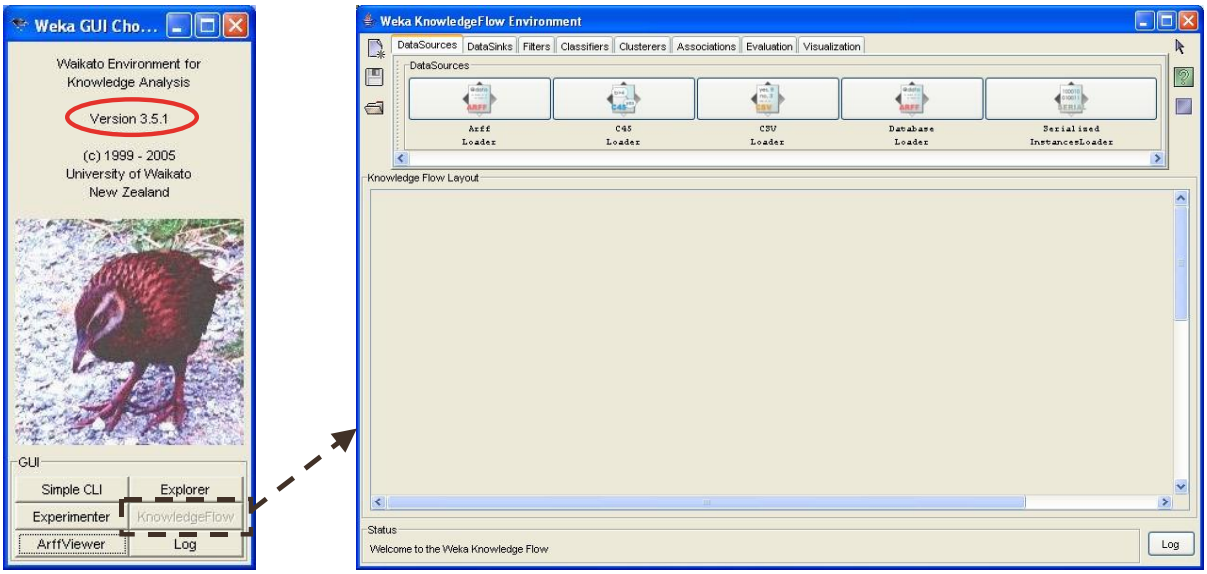
RULES

Number of rules : 14

N°	Antecedent	Consequent	Lift	Support	Confidence
1	"physician-fee-freeze=n"	"Class=democrat" - "adoption-of-the-budget-re=y"	1.670	0.503	0.887
2	"Class=democrat" - "adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.670	0.503	0.948
3	"Class=democrat"	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	1.629	0.503	0.820
4	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	"Class=democrat"	1.629	0.503	1.000
5	"Class=democrat"	"physician-fee-freeze=n"	1.616	0.563	0.918
6	"physician-fee-freeze=n"	"Class=democrat"	1.616	0.563	0.992
7	"adoption-of-the-budget-re=y"	"Class=democrat" - "physician-fee-freeze=n"	1.537	0.503	0.866
8	"Class=democrat" - "physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.537	0.503	0.894
9	"adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.524	0.503	0.866
10	"physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.524	0.503	0.887
11	"Class=democrat"	"adoption-of-the-budget-re=y"	1.488	0.531	0.865
12	"adoption-of-the-budget-re=y"	"Class=democrat"	1.488	0.531	0.913
13	"Class=democrat"	"aid-to-nicaraguan-contras=y"	1.468	0.501	0.816
14	"aid-to-nicaraguan-contras=y"	"Class=democrat"	1.468	0.501	0.901

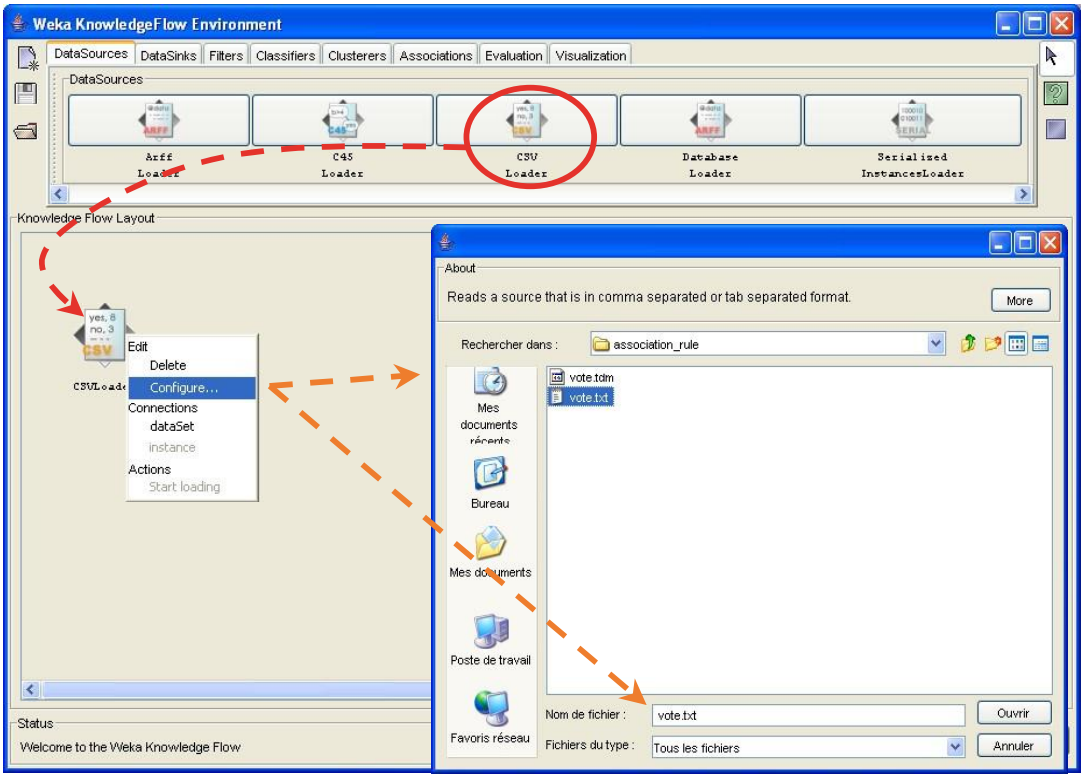
Association rules with WEKA

A dialog box appears when we execute WEKA; we choose the **KNOWLEDGE FLOW** paradigm. We have used the **3.5.1** version.



Import the dataset

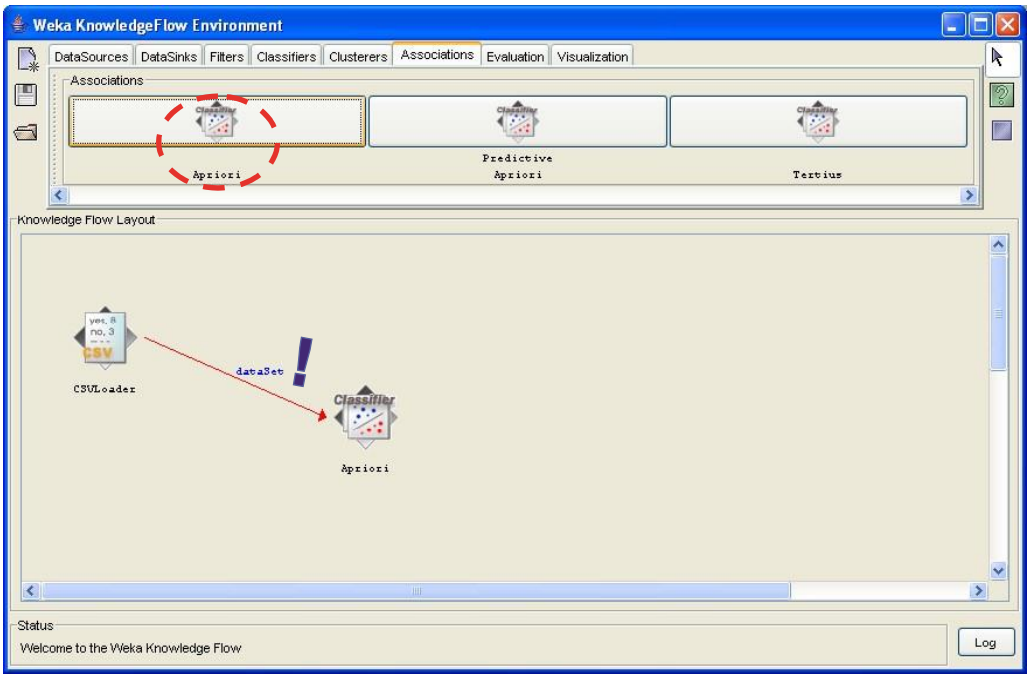
The CSV LOADER enables to handle text file format. We select the VOTE.TXT dataset with the CONFIGURE contextual menu.



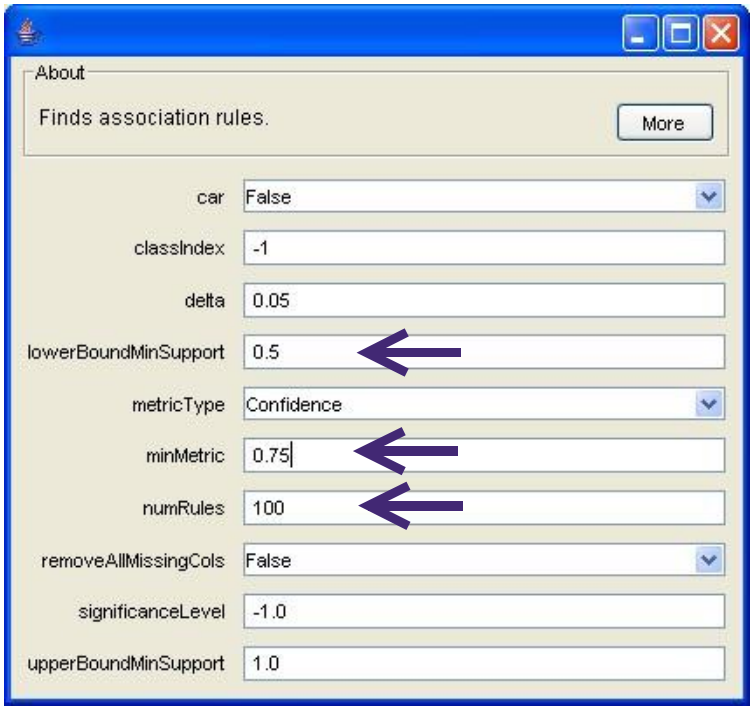
A PRIORI algorithm

The default selections are all instances and all attributes, so we must add only the A PRIORI component from the ASSOCIATION tab in the diagram.

We use the DATASET connection type.

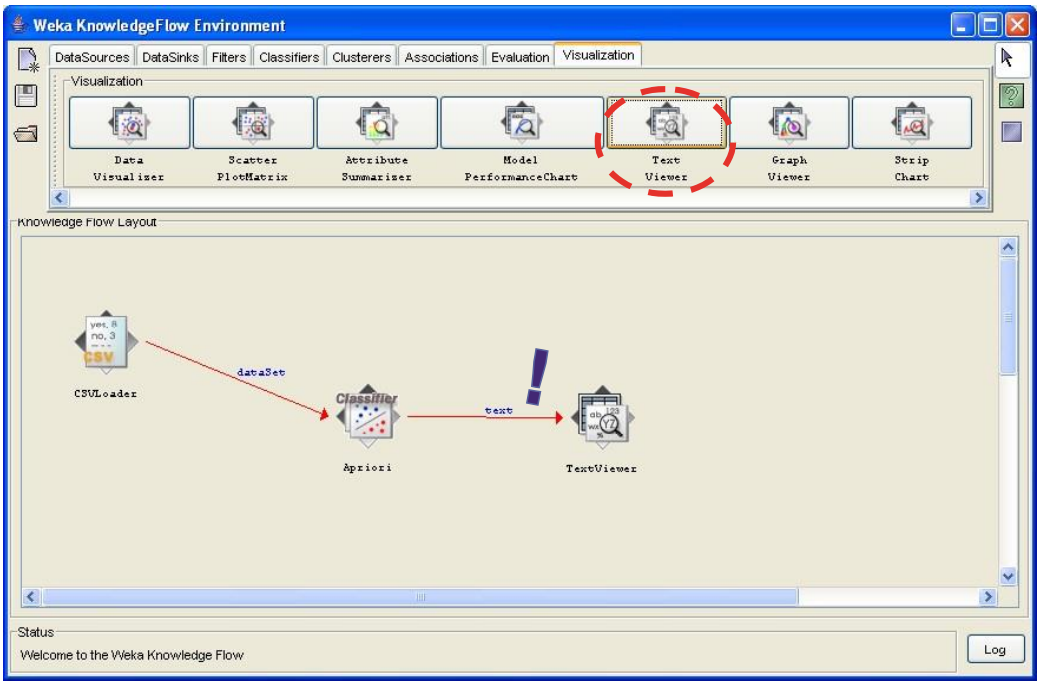


The CONFIGURE contextual menu allows to set the parameters values.

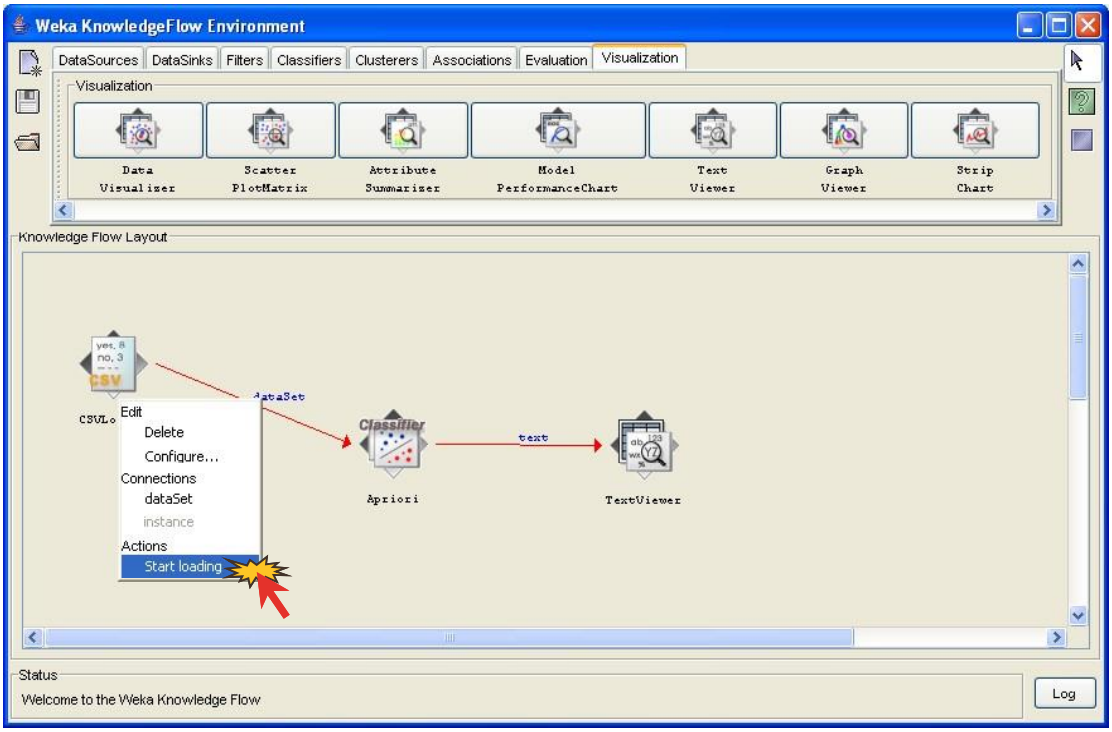


LOWERBOUNDMINSUPPORT set the minimal support of rules; MINMETRIC is the minimal confidence, if we set CONFIDENCE as METRIC TYPE; NUMRULES set the maximal number of rules that we can generate.

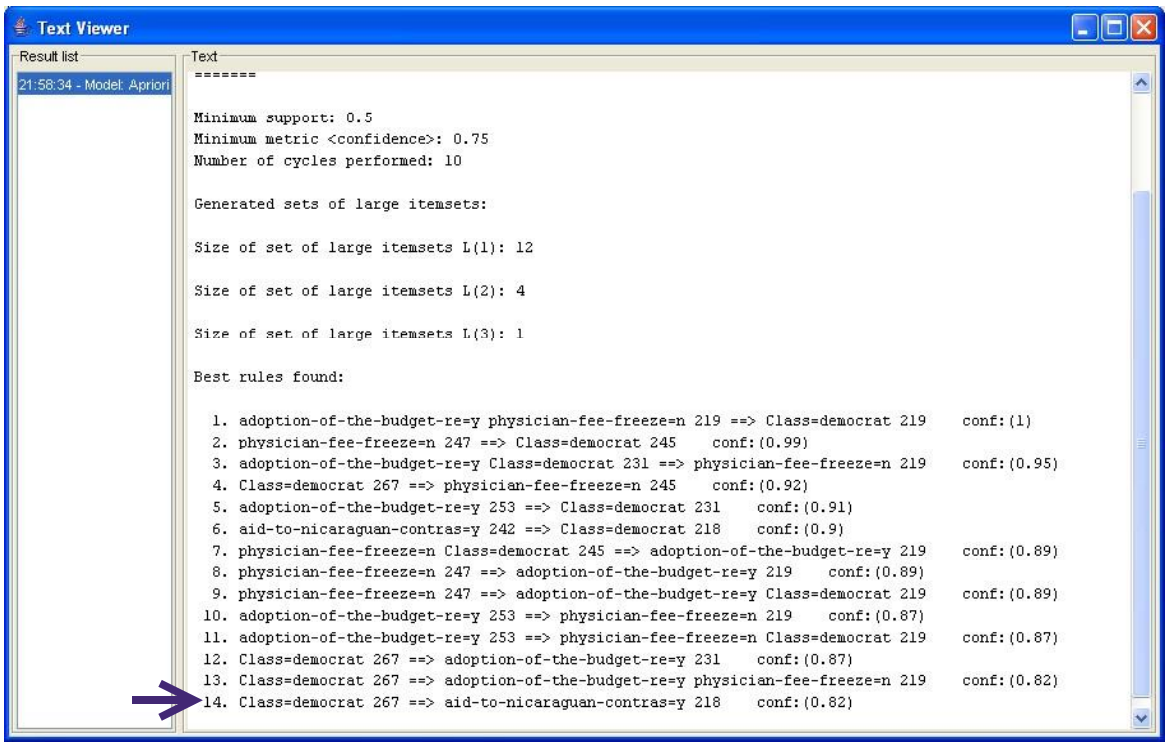
In order to visualize the rules, we add the TEXT VIEWER component in the diagram; we use the TEXT connection.



To execute the computation, we click on the START LOADING of the first component (CSV LOADER).



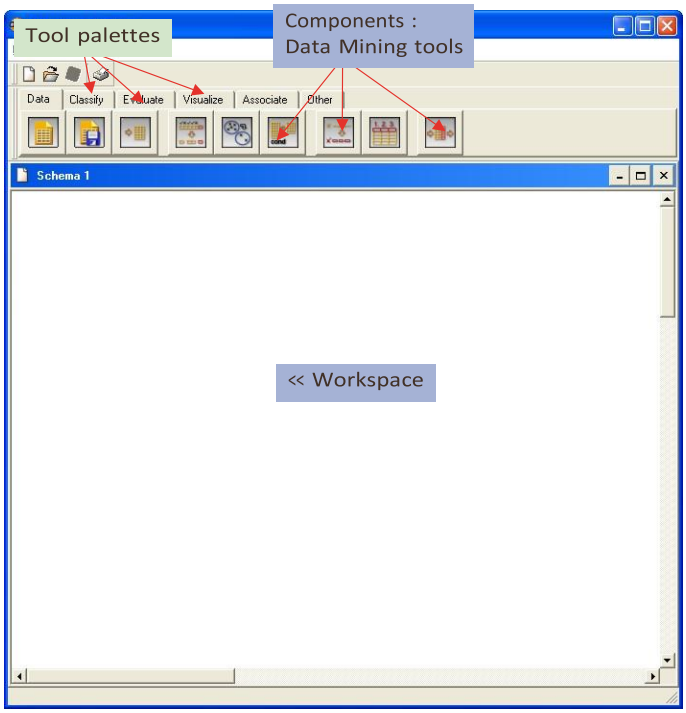
We can see the rule by clicking the SHOW RESULTS menu of the TEXT VIEWER component.



We obtain the same 14 rules as TANAGRA.

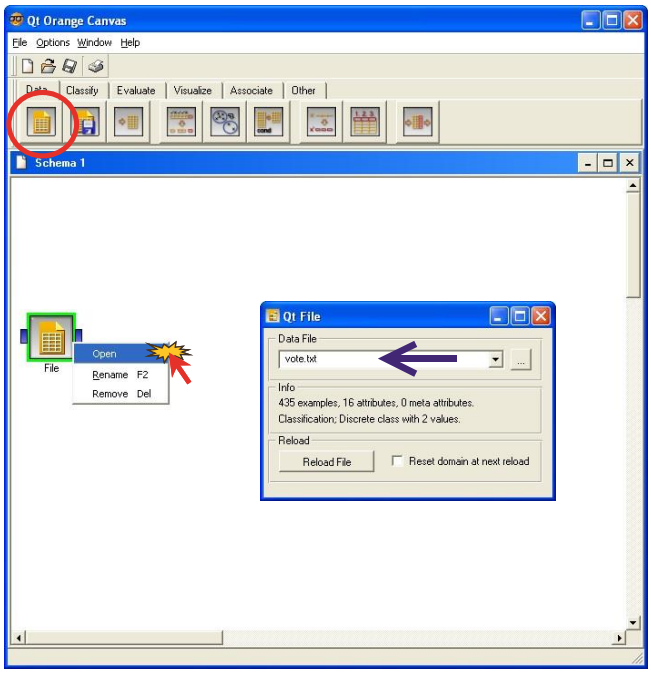
Association rules with ORANGE

When we execute ORANGE, we have the following interface.



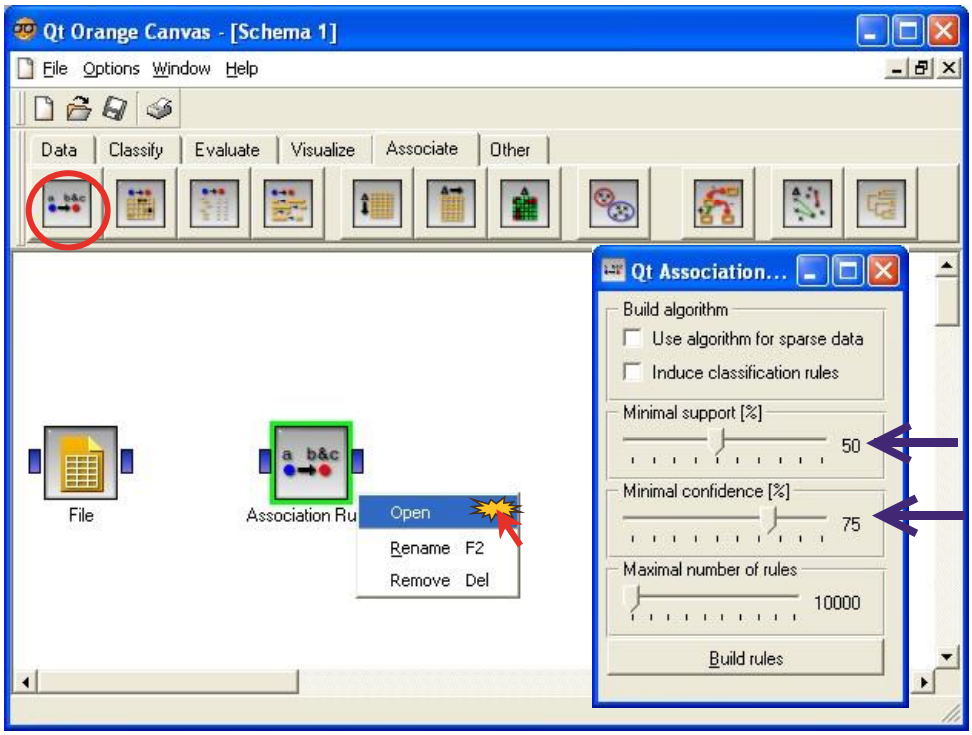
Importing the dataset

ORANGE can handle text file format (tabulation separator). When we select the tool, a new component is inserted in the diagram. We can select the file with the OPEN contextual menu.

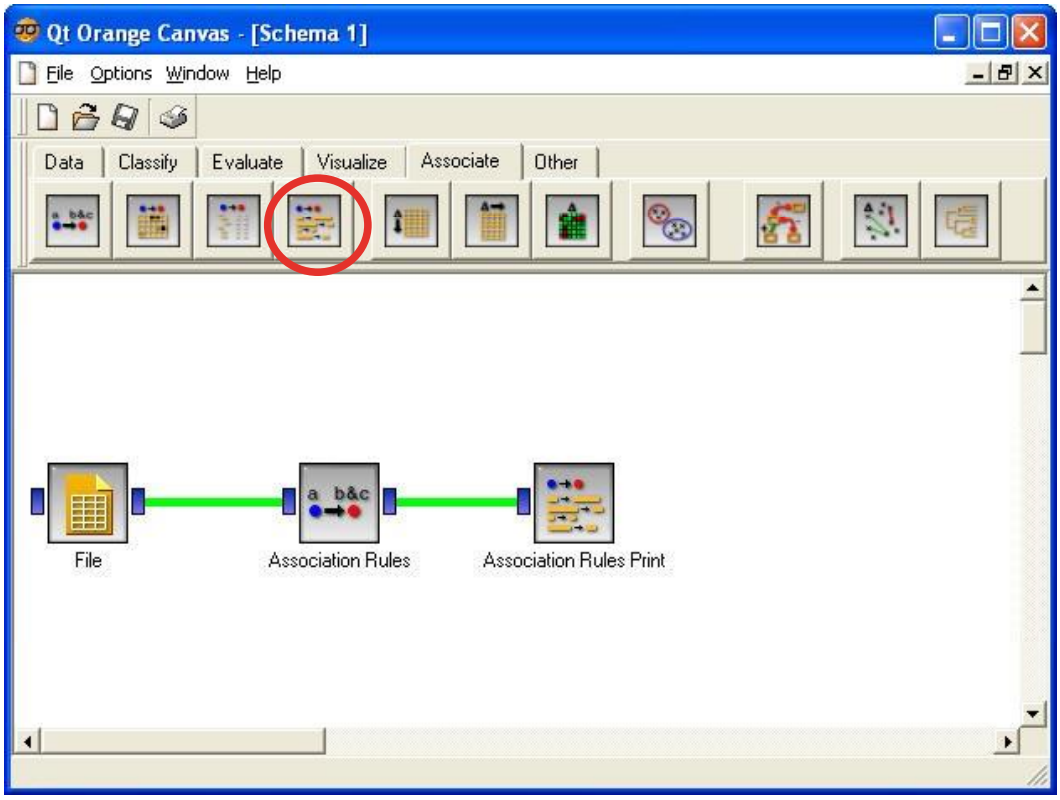


A PRIORI algorithm

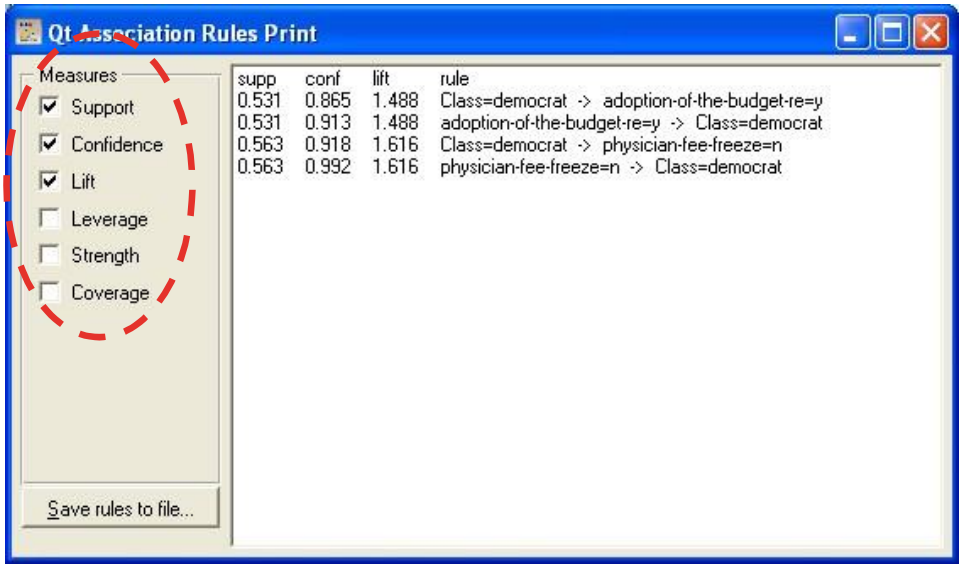
In order to compute the rules, we add ASSOCIATION RULES component. All examples and attributes are used. We click on the OPEN menu for parameters setting.



The rules are automatically computed when we connect the FILE component to ASSOCIATION RULE.
We add the ASSOCIATION RULE PRINT component for rules visualization.



We click on the OPEN menu in order to view the rules.



We obtain only 4 rules in ORANGE; we had found 14 with TANAGRA and WEKA.

ORANGE seems to have a preference for "shorter" rules. I did not find the reasons of this difference.

Conclusion

Our three packages are very simple to use for association rules induction.

These packages are largely sufficient for the majority of the analyses. The situation is a little more difficult if we wish to treat big databases with thousands of items. The number of generated rules can become very high and, the performances, the possibility of even carrying out calculations, very strongly depend on the RAM memory available of the machine used.