1 Introduction

Implementing the APRIORI MR component.

Association rule learning is a popular method for discovering interesting relations between variables in large databases (http://en.wikipedia.org/wiki/Association_rule_learning). It was often used in market basket analysis domain e.g. *if a customer buys onions and potatoes then he buys also beef*. But, in fact, it can be implemented in various application areas where we want to discover the association between variables.

We were already described the association rule mining tools of Tanagra in several tutorials (http://data-mining-tutorials.blogspot.com/search/label/Association%20rules). The A PRIORI approach is certainly the most popular. But, despite its good properties, this method has a drawback: the number of obtained rules can be very high. The ability to underline the most interesting rules, those which are relevant, becomes a major challenge.

These last years, numerous interestingness measures for association rules are studied. The goal is to associate a numerical indicator to rules. The A PRIORI MR component (ASSOCIATION tab) is an experimental tool which supplies several interestingness measures for evaluating rules. There are widespread measures such as confidence, support, etc; there are also less known measures such as those based on the test value principle. The theoretical foundations of these last measures are described in various tutorials available online http://data-mining-tutorials.blogspot.com/2009/02/interestingness-measures-for.html and http://data-mining-tutorials.blogspot.com/2009/05/understanding-test-value-criterion.html.

In this tutorial, we show to implement the A PRIORI MR component, how to set the parameters in order to obtain more or less rules, and how to read the results.

2 Dataset

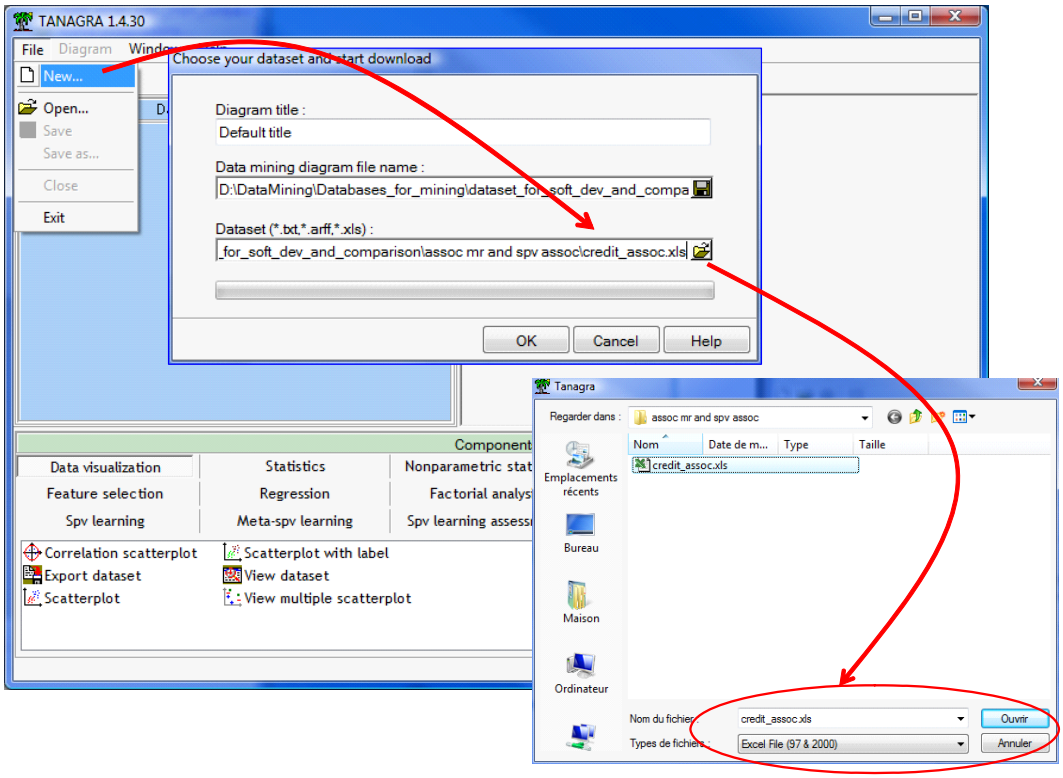
We use a modified version of the GERMAN CREDIT dataset in this tutorial. It depicts the characteristics of customers [http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)]. We do not try to explain in particular a variable in our framework. We try mainly to underline the interdependence between the variables (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_assoc.xls).

3 The A PRIORI MR Component

3.1 Creating a diagram and importing the dataset

Tanagra can load directly an Excel file format (XLS) even if the Excel software is not available on our computer. There are two restrictions for handling the data file: it must not be currently opened in other tool; the dataset must be in the first worksheet (see http://data-mining-tutorials.blogspot.com/search/label/Data%20file%20handling).

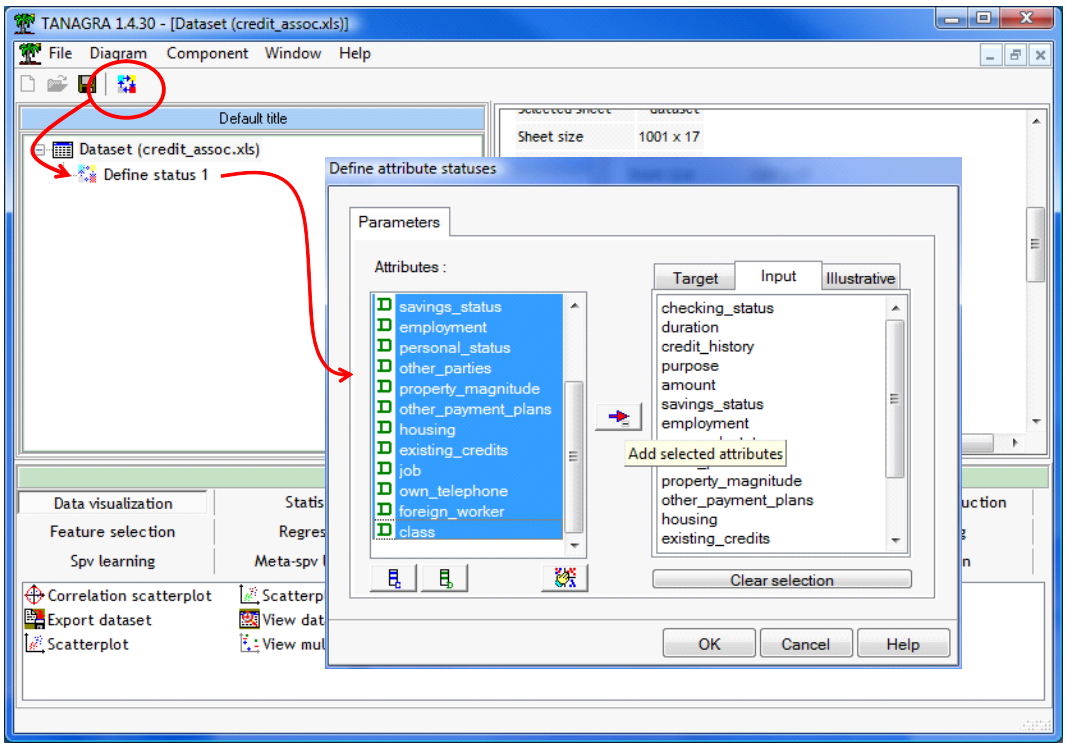
Into Tanagra, we click on the FILE / NEW menu; we select the CREDIT_ASSOC.XLS data file.



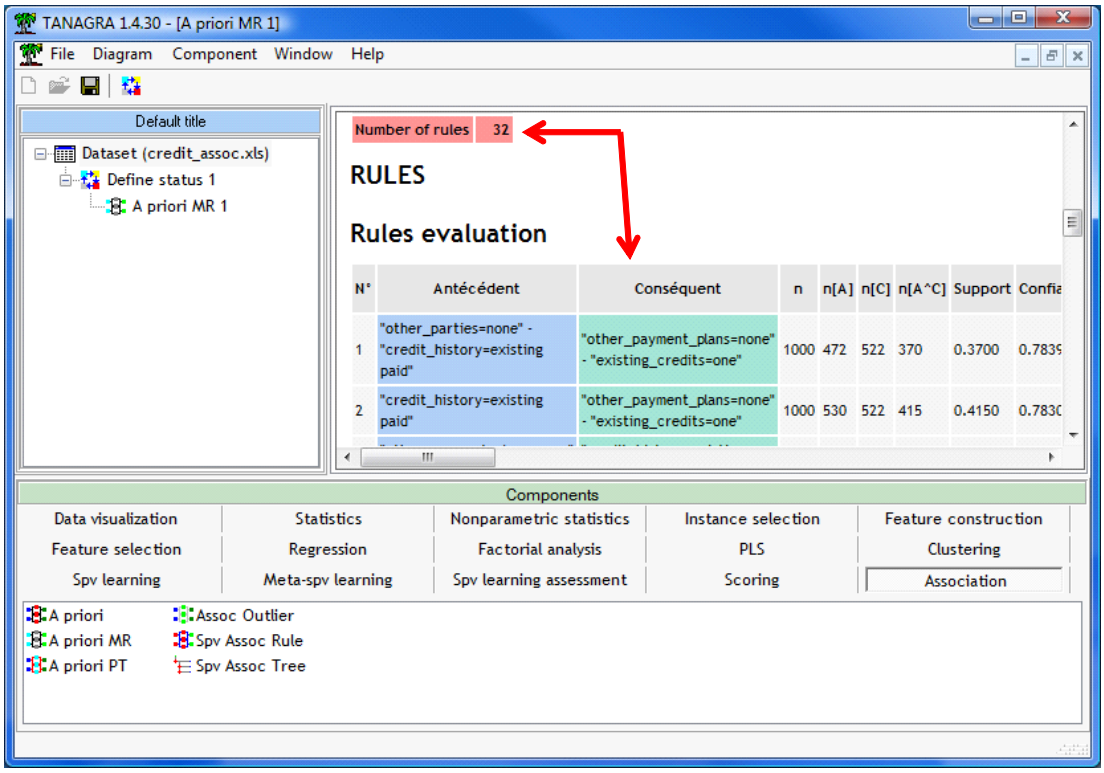
17 attributes and 1000 individuals from the DATASET sheet are now available for the analysis.

3.2 A PRIORI MR

We insert the DEFINE STATUS component into the diagram, using the shortcut into the tool bar. We set all the variables as INPUT.



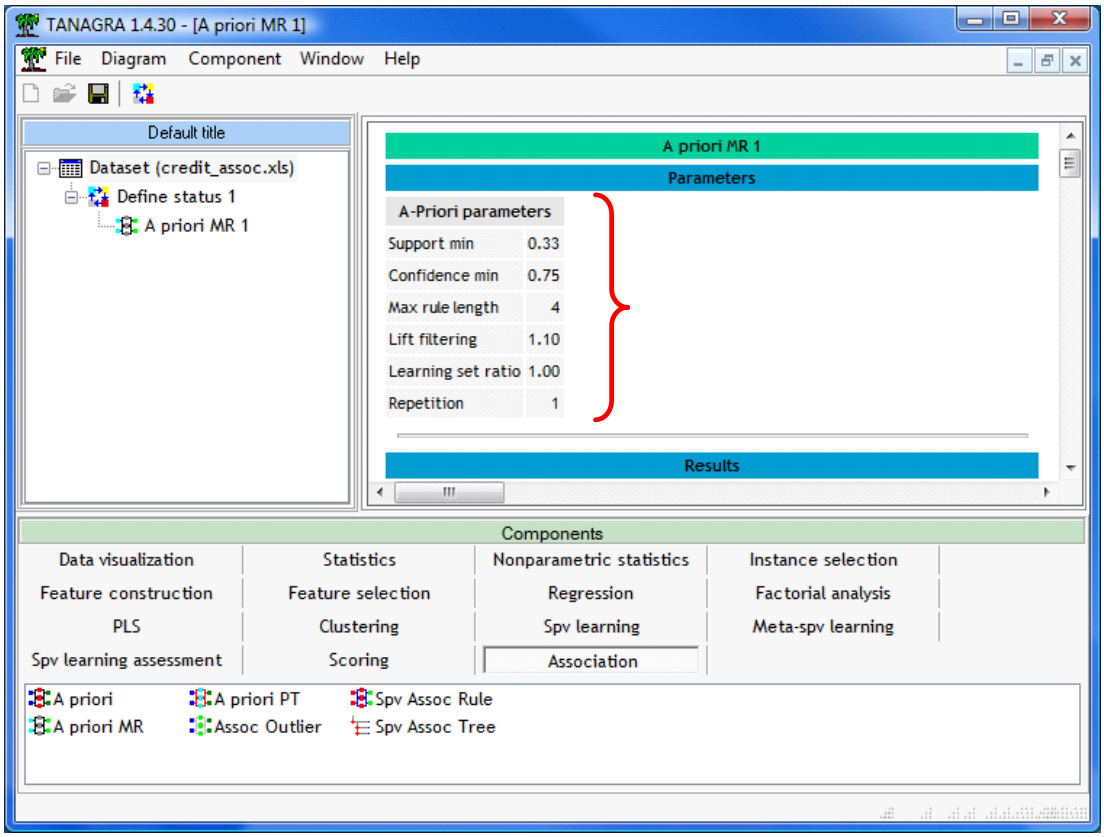
Then, we add the A PRIORI MR component (ASSOCIATION tab). We click on the VIEW menu. We obtain a first result with the default settings.



32 rules are generated. They are sorted in a decreasing order according the LIFT measure.

3.3 A PRIORI MR settings

Several essential settings are displayed in the PARAMETERS part of the report.



- SUPPORT MIN states the minimum support of the mined rules;
- CONFIDENCE MIN states the minimum confidence;

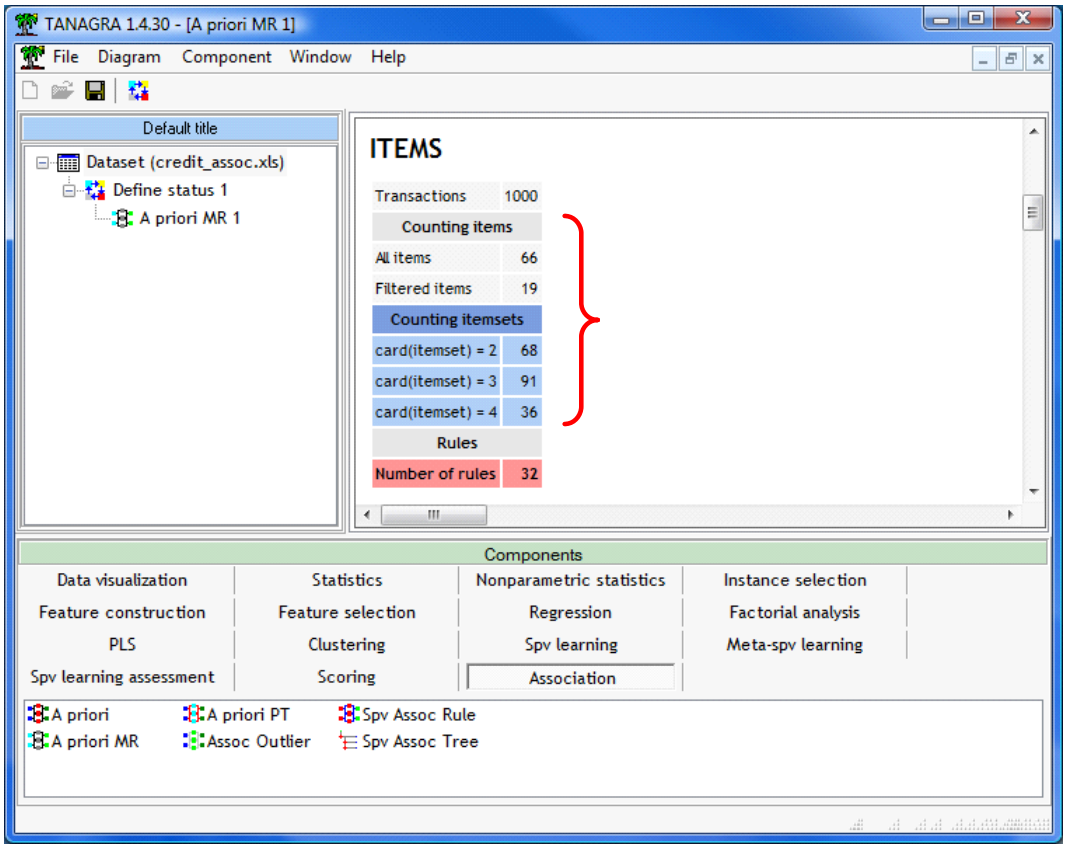
- MAX RULE LENGTH states the maximum number of items (attribute = value) of rules ;
- LIFT FILTERING states the minimum of LIFT.

These settings allow to restrict the number of mined rules. The first 3 parameters determines the computation time and the memory occupation; the last one (LIFT) filters only the displayed rules.

- LEARNING SET RATIO states the proportion of the dataset used for the learning phase. Indeed, Tanagra can subdivide the dataset into two parts: the first is used to extract the rules; the second is used to asses them. This framework is often used in the supervised learning task. Here, it allows to evaluate the reliability of the numerical indicator associated to the rules. The default value is 1 i.e. all the observations are used in the learning phase.
- REPETITION is the number of the replication of the Monte-Carlo procedure for the computation of the VT-100.

3.4 The results supplied by the A PRIORI MR Component

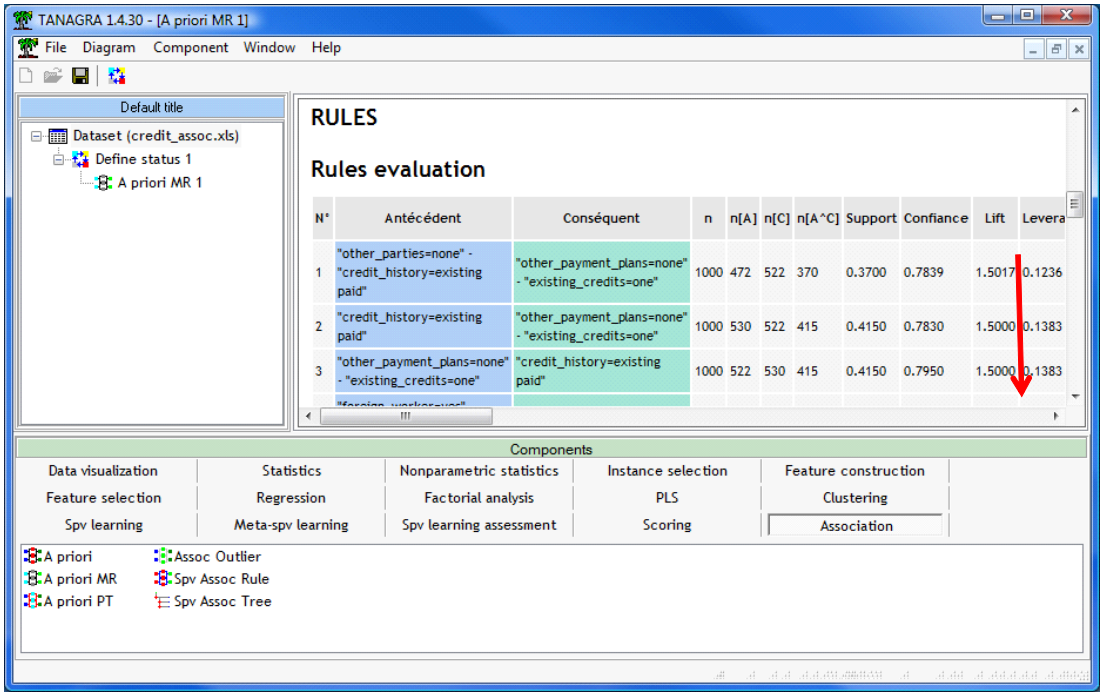
The ITEMS part describes the number of mined frequent itemsets (those of which the support is higher than SUPPORT MIN), gathered by cardinality. The total number of items is 66, 19 of them are frequent. For the itemsets of cardinality equal to two, we have 68; etc.



We can decrease the number of extracted itemsets: increasing the SUPPORT MIN parameter, it influences the number of itemsets extracted whatever the cardinality; decreasing the MAX RULE LENGTH, the itemsets with high cardinality are not computed.

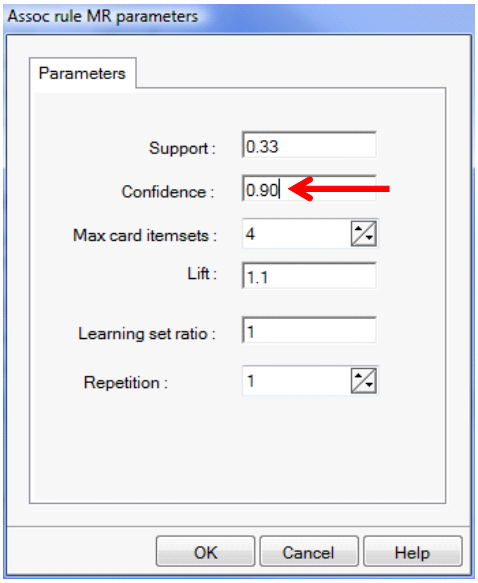
In a second phase, Tanagra computes the rules starting from the itemsets with cardinality higher to two. The CONFIDENCE MIN can take effect here. If we increase its value, we obtain fewer rules.

The **RULES section** describes the rules. The rule is subdivided in two parts: the antecedent part corresponds to the premise of the rule; the consequent part corresponds to the conclusion. Then, we have various numerical indicators which allow to assess the reliability of the rule. Some of them are well known (confidence, support, lift ...), the others are supplied only by Tanagra (test value).

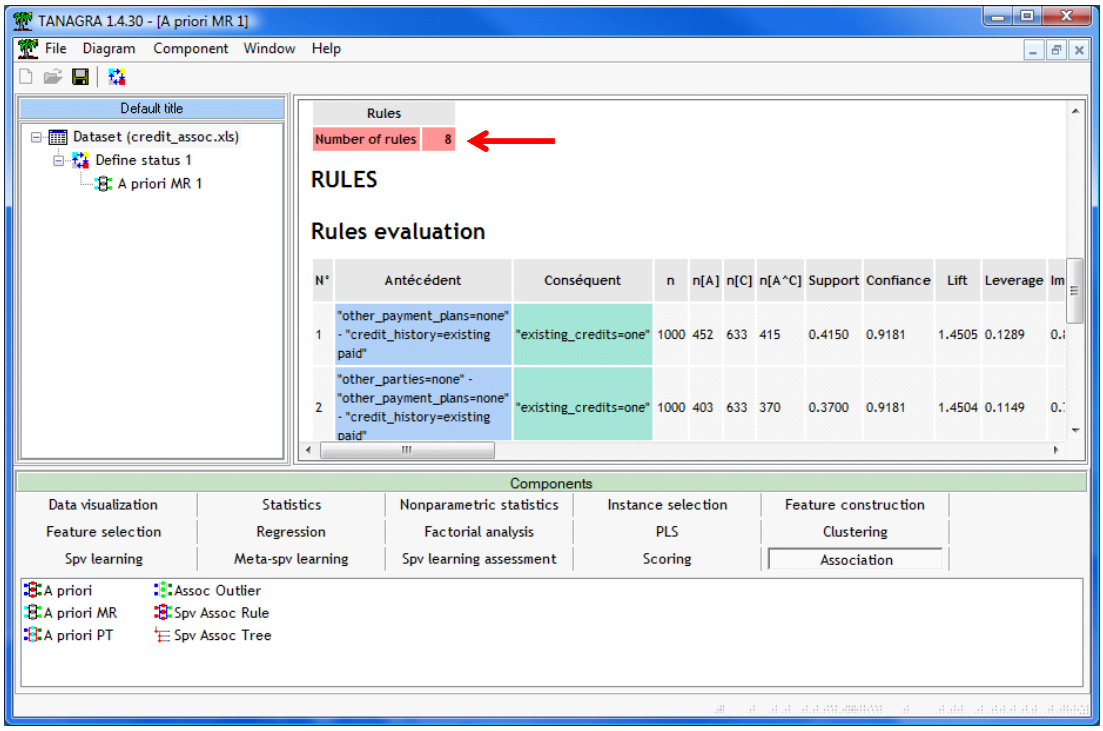


4 Fine tuning the parameters of A PRIORI MR

4.1 How to obtain fewer rules?



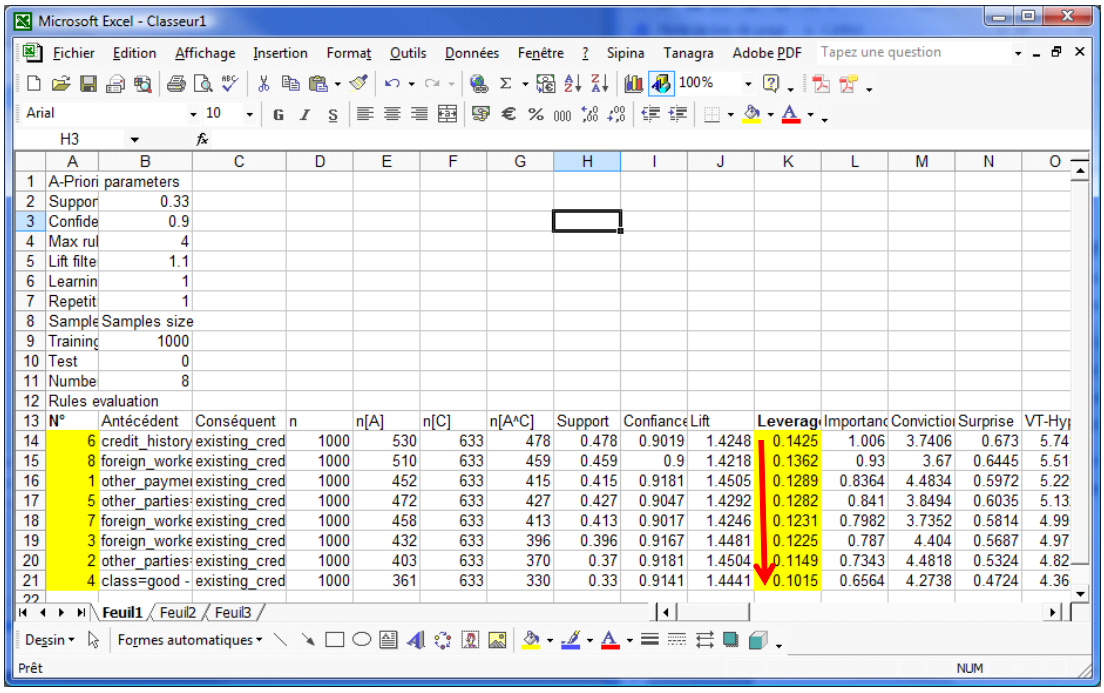
We want to obtain fewer rules using the MIN CONFIDENCE parameter. We click on the PARAMETERS contextual menu. We set 0.9 i.e. the itemsets with a support lower than 90% are not extracted. Then, we click on the VIEW menu. We obtain now 8 rules.



4.2 Exploring the rules

Filtering the rules is an interesting functionality. But, we must know how to set appropriately the parameters. It is not obvious. Tanagra supplies an option which allows to deeply explore the extracted rules. We can copy the results in a spreadsheet, then we can use the abilities of Excel to organize (sort) the rules in different ways, according various interestingness measures.

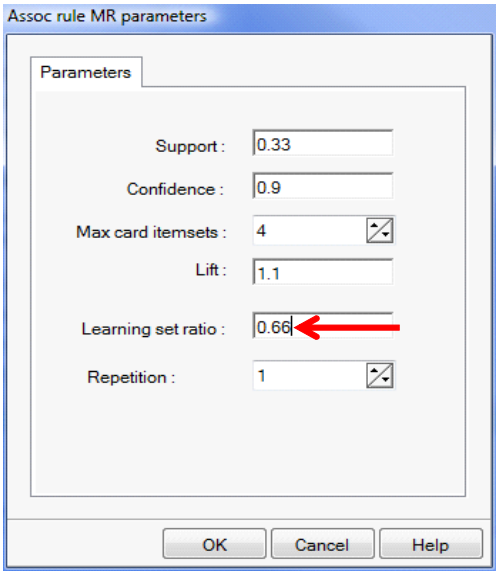
We click on the COMPONENT / UNFORMATTED COPY in the main menu. Then, we launch EXCEL. We paste the results. With the various tools supplied by Excel, we can explore deeply the rule base. In the screenshot below, we have sorted the rules according the LEVERAGE criterion.



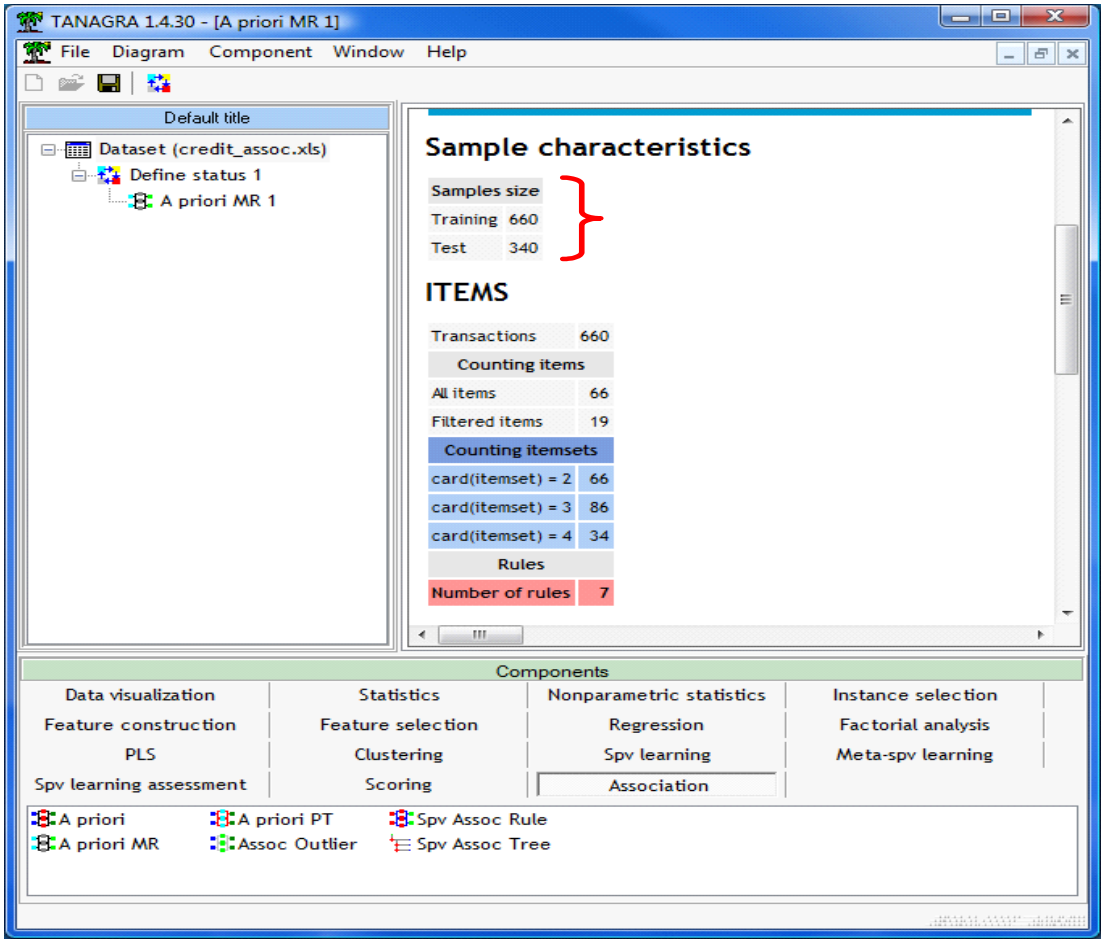
4.3 Subdividing the dataset into training and test sets

We can subdivide the dataset into two parts for computing the rules (train set) and assessing them (test set). We obtain thus an honest estimation of the interestingness measure associated to the rule.

We click on the PARAMETERS menu. We set the LEARNING RATIO to 0.66 i.e. 660 observations on 1000 are used in the training phase; 340 in the testing phase.



We validate and we click on the VIEW menu.



Tanagra extracts 7 rules. In the right part of the table enumerating the rules, after the TEST column (in red), we have the interestingness measures computed on the test set. The comparison of the values obtained on the train and the test set allows to assess the stability of the rules.

5 Conclusion

The A PRIORI MR component of Tanagra extracts rules from data using the A PRIORI algorithm. It differentiates oneself from other by offering additional tools for exploring and assessing the mined rules: original measures based on the “test value” principle allow to evaluate differently the rules; the ability to copy the results into a spreadsheet allows a more detailed exploration of the rule base; by subdividing the dataset into train and test sets, we obtain a more reliable values of the interestingness measures of rules.