Subject

In this tutorial, we show how to build association rule on a big dataset using an external program.

Our implementation of A PRIORI is fast but needs a lot of memory that limits its performances when we treat a big dataset or generate numerous rules. I have discovered the Christian BORGELT's work, he proposes a very powerful association rule generator, which can handle large dataset and is very fast (http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html).

To execute its implementation, we integrated a new approach in TANAGRA: the launching and the control of an external program. At the time of the execution, we create a temporary file, which we transmit to the APRIORI.EXE executable file. Then the rules are automatically downloaded and displayed.

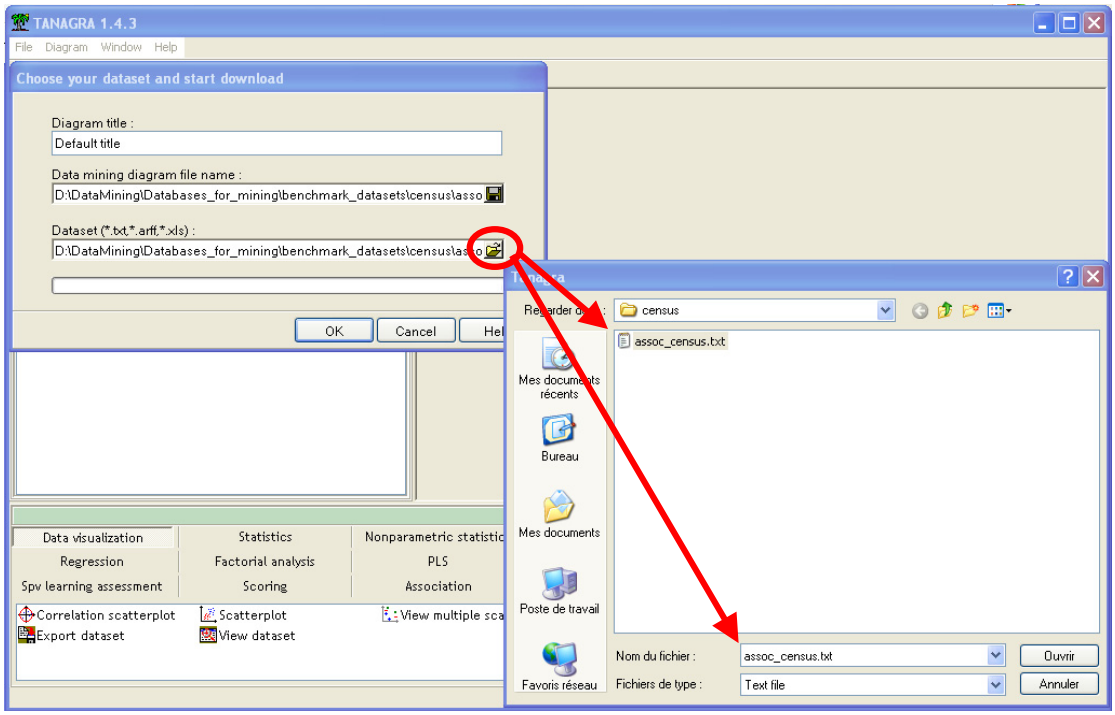
Dataset

We use the CENSUS[1] dataset (ASSOC_CENSUS.TXT). We removed the continuous attributes and select 200 000 examples. The size of the dataset is 90 MB.

Borgelt's APRIORI

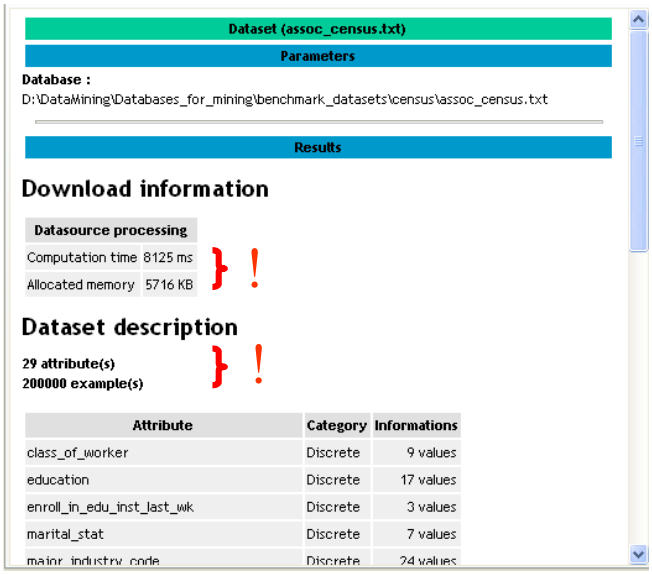
Import the dataset

First, we must import the dataset with the FILE / NEW menu.



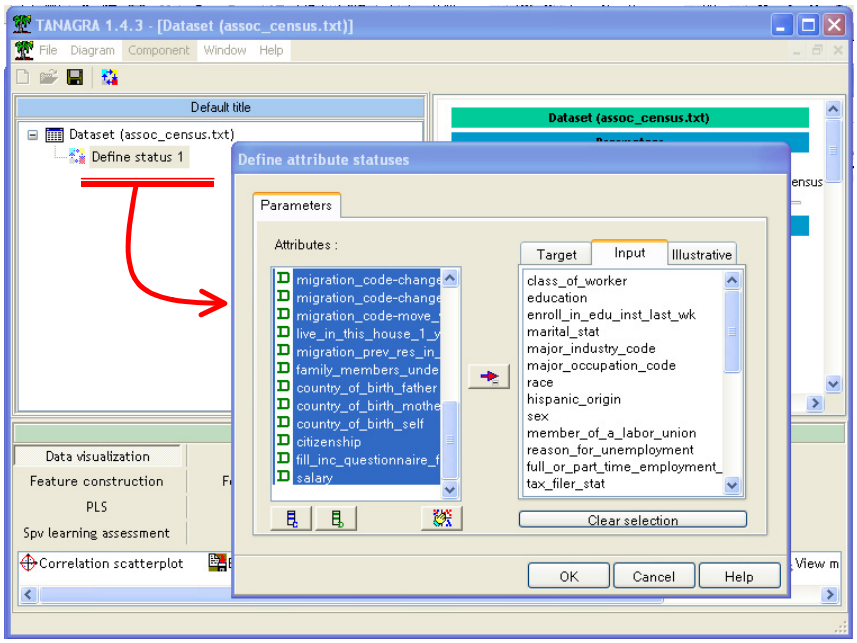
[1] This data was extracted from the census bureau database found at | http://www.census.gov/ftp/pub/DES/www/welcome.html. Donor: Terran Lane and Ronny Kohavi.

The data importation is moderately fast (# 8 seconds[2]), we see the characteristics of the dataset.



Define INPUT attributes

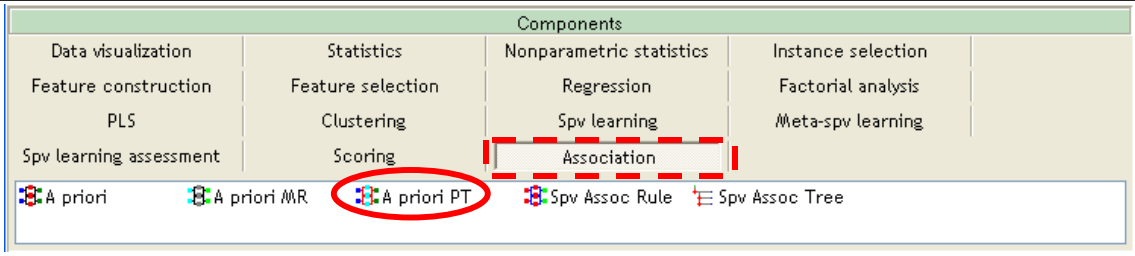
We set all attributes as INPUT.



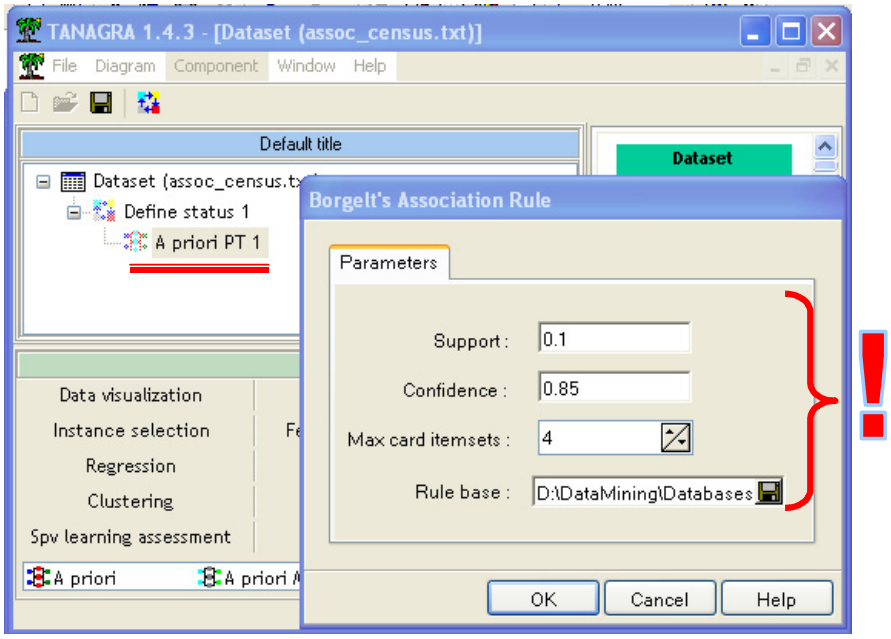
Add the A PRIORI PT component

We add the A PRIORI PT component. We can find the component into the ASSOCIATION tab.

[2] CELERON 2,53 Ghz Under XP



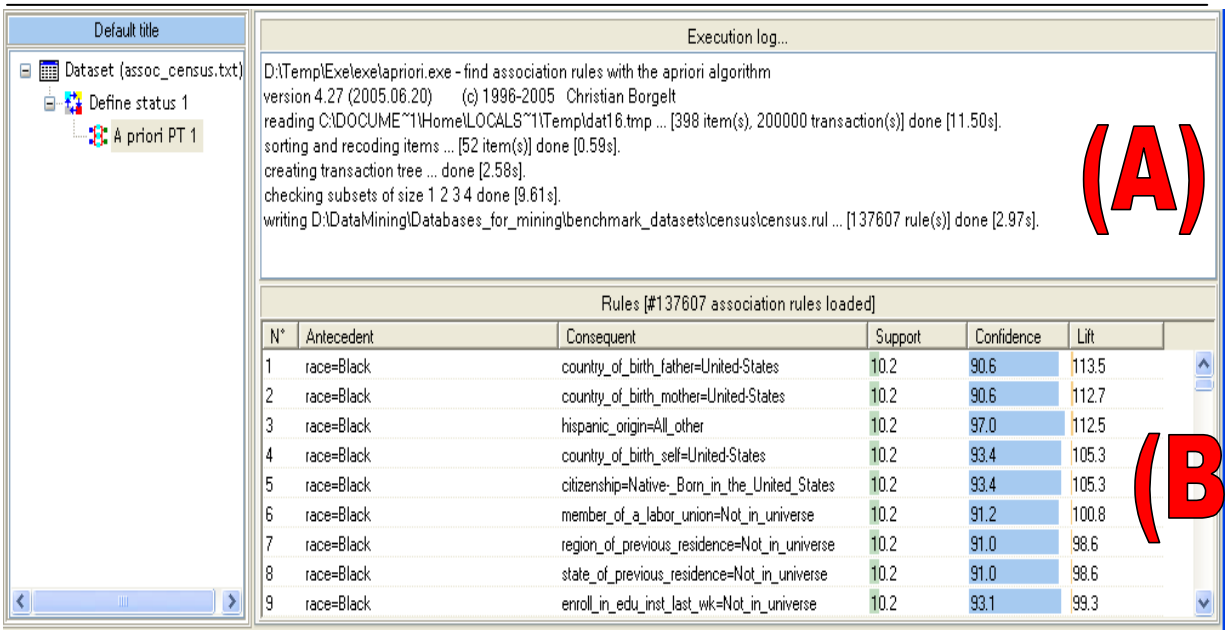
We set the settings as the following.



We set the MIN SUPPORT at 0.1 (10%), the MIN CONFIDENCE MIN at 0.85 (85%), the max cardinal of itemsets as 4 (MAX CARD ITEMSETS), all these parameters enables to restrict the number of the generated rules; at last, we set the rule base filename. Because, we can obtain a great number of rules, save always the rules on a hard disk.

Run the component and visualize the rules

When we run the component, TANAGRA creates a temporary file and run the Christian BORGELT's APRIORI.EXE program. A window enables us to see the execution progress (A), and see the results (B). Christian's program is extremely fast, it needs several seconds to generate 137 607 rules!



We can sort the rules according some measures. For instance, if you click on the header of SUPPORT, you obtain the following results.