

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.

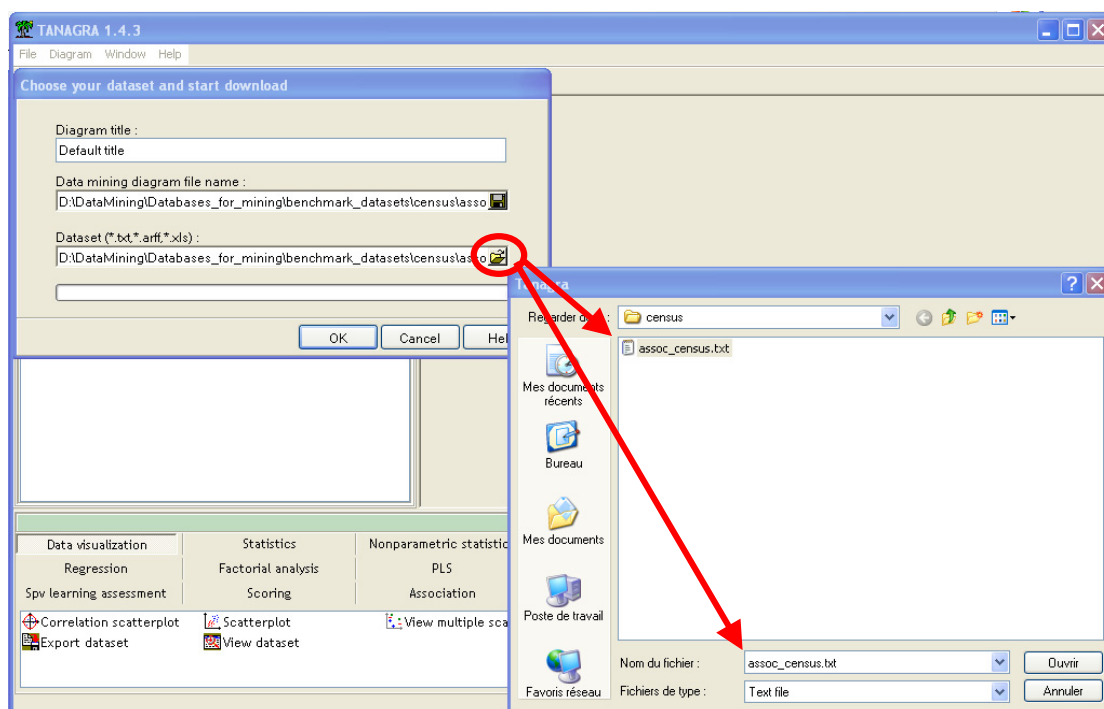
Dataset

We use the CENSUS¹ 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 seconds²), we see the characteristics of the dataset.

Dataset (assoc_census.txt)

Parameters

Database :
D:\DataMining\Databases_for_mining\benchmark_datasets\census\assoc_census.txt

Results

Download information

Datasource processing

Computation time 8125 ms } !
Allocated memory 5716 KB } !

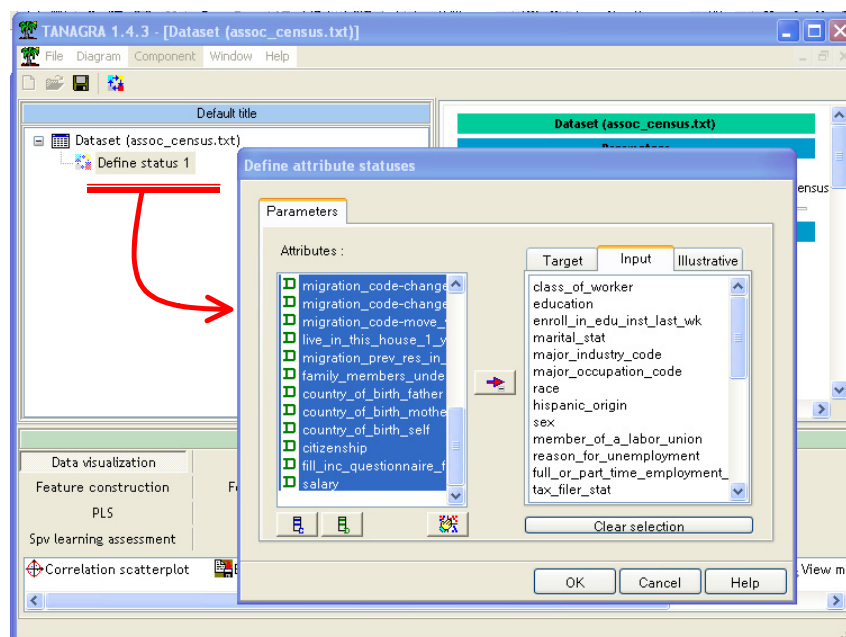
Dataset description

29 attribute(s) } !
200000 example(s) } !

Attribute	Category	Informations
class_of_worker	Discrete	9 values
education	Discrete	17 values
enroll_in_edu_inst_last_wk	Discrete	3 values
marital_stat	Discrete	7 values
major_industry_code	Discrete	24 values

Define INPUT attributes

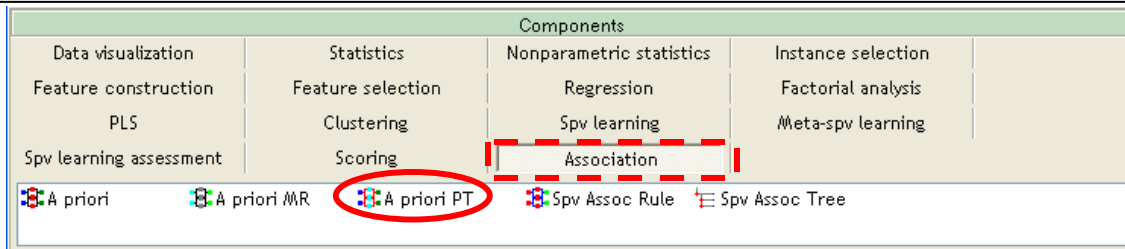
We set all attributes as INPUT.



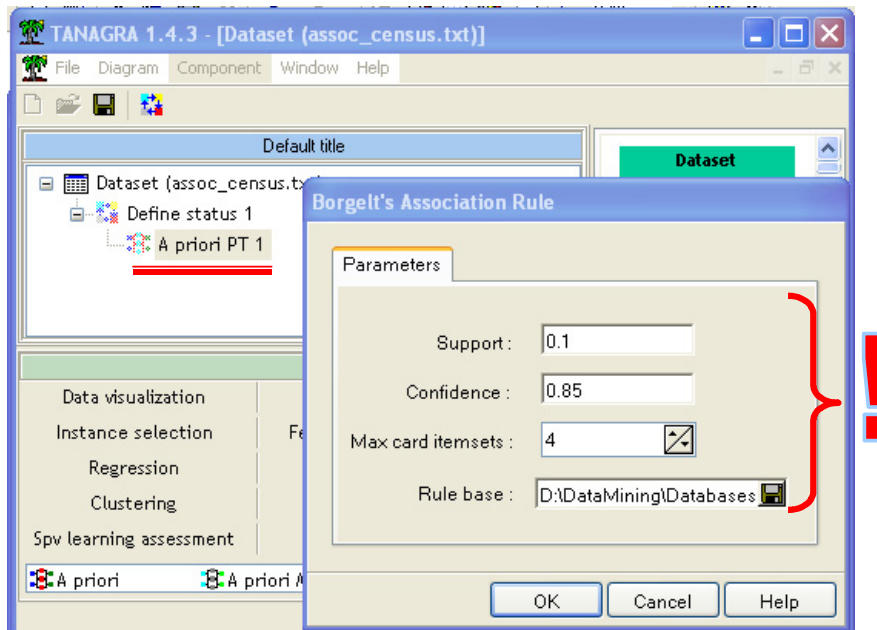
Add the A PRIORI PT component

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 APRIORILEXE 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!

Execution log...

```
D:\Temp\Exe\apriori.exe - find association rules with the apriori algorithm
version 4.27 (2005.06.20) (c) 1996-2005 Christian Borgelt
reading C:\DOCUMENTS\11Home\LOCALS\11Temp\dat16.tmp ... [398 item(s), 200000 transaction(s)] done [11.50s].
sorting and recoding items ... [52 item(s)] done [0.59s].
creating transaction tree ... done [2.58s].
checking subsets of size 1 2 3 4 done [9.61s].
writing D:\DataMining\Databases_for_mining\benchmark_datasets\census\census.rul ... [137607 rule(s)] done [2.97s].
```

Rules [#137607 association rules loaded]					
N°	Antecedent	Consequent	Support	Confidence	Lift
1	race=Black	country_of_birth_father=United-States	10.2	90.6	113.5
2	race=Black	country_of_birth_mother=United-States	10.2	90.6	112.7
3	race=Black	hispanic_origin=All_other	10.2	97.0	112.5
4	race=Black	country_of_birth_self=United-States	10.2	93.4	105.3
5	race=Black	citizenship=Native-Born_in_the_United_States	10.2	93.4	105.3
6	race=Black	member_of_a_labor_union=Not_in_universe	10.2	91.2	100.8
7	race=Black	region_of_previous_residence=Not_in_universe	10.2	91.0	98.6
8	race=Black	state_of_previous_residence=Not_in_universe	10.2	91.0	98.6
9	race=Black	enroll_in_edu_inst_last_wk=Not_in_universe	10.2	93.1	99.3

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

Rules [#137607 association rules loaded]					
N°	Antecedent	Consequent	Support	Confidence	Lift
613	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	member_of_a_labor_union=Not_in_universe	99.0	90.4	100.0
643	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	reason_for_unemployment=Not_in_universe	99.0	96.9	100.0
587	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	country_of_birth_self=United-States	99.0	88.6	99.9
601	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	citizenship=Native-Born_in_the_United_States	99.0	88.6	99.9
571	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	hispanic_origin=All_other	99.0	86.1	99.9
631	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	state_of_previous_residence=Not_in_universe	99.0	92.2	100.0
641	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	salary<less_50000	99.0	93.9	100.1
637	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	enroll_in_edu_inst_last_wk=Not_in_universe	99.0	93.7	99.9
623	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	region_of_previous_residence=Not_in_universe	99.0	92.2	100.0
585	reason_for_unemployment=Not_in_universe	country_of_birth_self=United-States	97.0	88.9	100.2
629	reason_for_unemployment=Not_in_universe	state_of_previous_residence=Not_in_universe	97.0	92.4	100.2
599	reason_for_unemployment=Not_in_universe	citizenship=Native-Born_in_the_United_States	97.0	88.9	100.2
569	reason_for_unemployment=Not_in_universe	hispanic_origin=All_other	97.0	86.4	100.2
611	reason_for_unemployment=Not_in_universe	member_of_a_labor_union=Not_in_universe	97.0	90.1	99.7
639	reason_for_unemployment=Not_in_universe	salary<less_50000	97.0	93.7	99.9
635	reason_for_unemployment=Not_in_universe	enroll_in_edu_inst_last_wk=Not_in_universe	97.0	93.9	100.2
642	reason_for_unemployment=Not_in_universe	fill_inc_questionnaire_for_veteran_s_admin=Not_in_universe	97.0	99.0	100.0
552	reason_for_unemployment=Not_in_universe	race=White	97.0	84.1	100.2