

1 Objectif

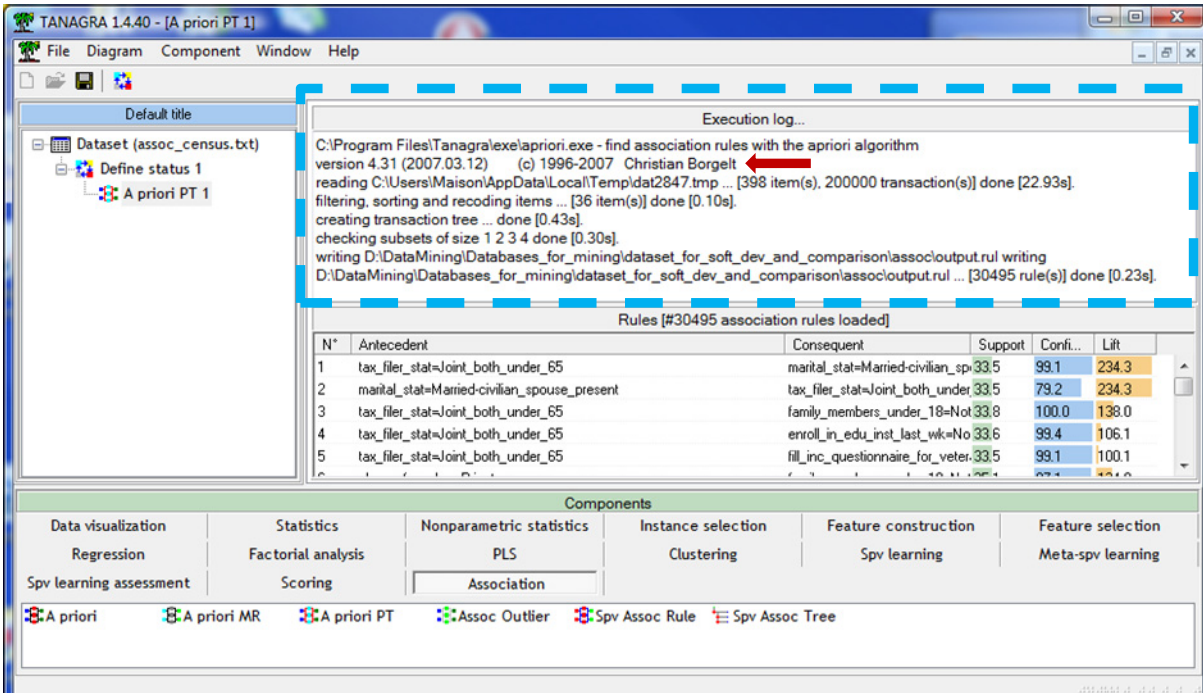
Mise à jour du composant APRIORI PT basé sur le programme apriori.exe 5.57 de Borgelt.

APRIORI PT est un des rares composants de Tanagra basé sur une bibliothèque externe, le programme « apriori.exe » de Borgelt en l'occurrence¹. Jusqu'à la version 1.4.40 de Tanagra, nous utilisons la version 4.31 de l'exécutable (du 12/03/2007). Nous introduisons une version autrement plus récente (5.57 du 02/09/2011) dans Tanagra 1.4.41. Les paramètres étant légèrement modifiés, il a fallu adapter le programme appelant. Néanmoins, le fonctionnement reste identique, il en est de même en ce qui concerne la lecture des résultats.

Nous reprenons un ancien tutoriel (<http://tutoriels-data-mining.blogspot.com/2008/04/rgles-dassociation-avec-les-prefix-tree.html>) pour décrire le comportement de cette nouvelle mouture. Nous ne revenons pas sur le détail (importation des données, choix des variables, paramétrage) de l'utilisation du composant APRIORI PT, puisque cela a déjà été fait. Nous essayons surtout de mettre en évidence les progrès du module en termes de temps de traitements. Force est de constater qu'ils sont particulièrement impressionnants.

2 APRIORI.EXE 4.31 (via la version 1.4.40 de Tanagra)

Nous importons la base [assoc_census.txt](#). Nous définissons les traitements, puis nous lançons l'extraction avec les paramètres par défaut (support min = 0.33, confiance min = 0.75, cardinal maximal des règles = 4). Nous obtenons les résultats suivants.



The screenshot shows the TANAGRA 1.4.40 interface. The 'Execution log...' window displays the following text:

```
C:\Program Files\tanagra\exe\apriori.exe - find association rules with the apriori algorithm
version 4.31 (2007.03.12) (c) 1996-2007 Christian Borgelt
reading C:\Users\Maison\AppData\Local\Temp\dat2847.tmp ... [398 item(s), 200000 transaction(s)] done [22.93s].
filtering, sorting and recoding items ... [36 item(s)] done [0.10s].
creating transaction tree ... done [0.43s].
checking subsets of size 1 2 3 4 done [0.30s].
writing D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\assocoutput.rul writing
D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\assocoutput.rul ... [30495 rule(s)] done [0.23s].
```

The 'Rules [#30495 association rules loaded]' table is as follows:

N°	Antecedent	Consequent	Support	Confu..	Lift
1	tax_filer_stat=Joint_both_under_65	marital_stat=Married-civilian_sp	33.5	99.1	234.3
2	marital_stat=Married-civilian_spouse_present	tax_filer_stat=Joint_both_under	33.5	79.2	234.3
3	tax_filer_stat=Joint_both_under_65	family_members_under_18=Not	33.8	100.0	138.0
4	tax_filer_stat=Joint_both_under_65	enroll_in_edu_inst_last_wk=No	33.6	99.4	106.1
5	tax_filer_stat=Joint_both_under_65	fill_inc_questionnaire_for_veter	33.5	99.1	100.1

The 'Components' section at the bottom shows the 'Association' component selected.

30.495 règles ont été générées.

En sous main, le temps de préparation du fichier temporaire envoyé à « apriori.exe » est de 4 secondes. Par la suite, nous notons que la recherche des règles proprement dite est très rapide. Le

¹ <http://www.borgelt.net/apriori.html>

temps de traitements est surtout grevé par le chargement du fichier temporaire des données de transactions en mémoire. Nous verrons que le module a surtout évolué sur ce point.

3 APRIORI.EXE 5.57 (via la version 1.4.41 de Tanagra)

Avec la nouvelle version 1.4.41 de Tanagra, nous obtenons les résultats suivants.

The screenshot shows the TANAGRA 1.4.41 interface. The 'Execution log...' window displays the following text:

```
D:\Temp\Exe\exe\apriori.exe - find frequent item sets with the apriori algorithm
version 5.57 (2011.09.02) (c) 1996-2011 Christian Borgelt
reading C:\Users\Maison\AppData\Local\Temp\dat3C83.tmp ... [398 item(s), 200000 transaction(s)] done [1.66s].
filtering, sorting and recoding items ... [36 item(s)] done [0.02s].
sorting and reducing transactions ... [9332/200000 transaction(s)] done [0.37s].
building transaction tree ... [25531 node(s)] done [0.01s].
checking subsets of size 1 2 3 4 done [0.24s].
writing D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\assocoutput.rul ... [30495 rule(s)] done
```

Below the log, a table titled 'Rules [#30495 association rules loaded]' is shown:

N°	Antecedent	Consequent	Support	Conf...	Lift
1	tax_filer_stat=Joint_both_under_65	marital_stat=Married-civilian_sp	33.5	99.1	234.3
2	marital_stat=Married-civilian_spouse_present	tax_filer_stat=Joint_both_under	33.5	79.2	234.3
3	tax_filer_stat=Joint_both_under_65	family_members_under_18=Not	33.8	100.0	138.0
4	tax_filer_stat=Joint_both_under_65	enroll_in_edu_inst_last_wk=No	33.6	99.4	106.1
5	tax_filer_stat=Joint_both_under_65	fill_inc_questionnaire_for_veter	33.5	99.1	100.1
6	class_of_worker=Private	family_members_under_18=Not	35.1	97.1	134.0
7	class_of_worker=Private	enroll_in_edu_inst_last_wk=No	33.4	92.3	98.5

The 'Components' section at the bottom lists various data mining tasks such as Data visualization, Statistics, Nonparametric statistics, Instance selection, Feature construction, Feature selection, Regression, Factorial analysis, PLS, Clustering, Spv learning, Meta-spv learning, Spv learning assessment, and Scoring.

Nous obtenons les mêmes 30.495 règles à paramétrage égal. C'est heureux.

Nous remarquons dans la fenêtre log que les calculs sont organisés différemment, avec une étape supplémentaire (« *sorting and reducing transactions* »). Nous notons surtout que le temps de chargement du fichier temporaire transmis à « apriori.exe » a été réduit dans des proportions absolument impressionnantes. A machine strictement égale, nous passons de 22.93 secondes à 1.66 secondes. Faire mieux me paraît difficile, le fichier temporaire pèse quand même ~220 Mo.

The screenshot shows a Windows Explorer window with the address bar set to 'C:\Users\Maison\AppData\Local\Temp'. The file list is as follows:

Nom	Date de modificati...	Type	Taille
pptf629.tmp	24/09/2011 11:47	Fichier TMP	233 Ko
dat3C83.tmp	24/09/2011 11:31	Fichier TMP	220 485 Ko
Twaip001.Mtx	24/09/2011 11:37	Fichier.MTX	1 Ko

4 Conclusion

Habituellement, je ne suis pas trop enclin à courir après les versions des logiciels. Je préfère une version ancienne que je connais et qui marche correctement plutôt qu'une version, certes récente, mais qu'il faut valider de nouveau. Dans le cas de « apriori.exe », les différents tests que j'ai menés

montrent que le dispositif est tout à fait satisfaisant. Et les avancées en termes de temps de calculs sont vraiment remarquables.