

1 Theme

Updating of the APRIORI PT component, based on the Borgelt's apriori.exe 5.57 program.

A PRIORI PT is a tool dedicated for the extraction of association rules. This is one of the few components of Tanagra based on external library. We use the Borgelt's "apriori.exe" program. Until the version 1.4.40 of Tanagra, we used the 4.31 version of "apriori.exe". From the Tanagra 1.4.41, we introduce the latest update 5.57 (2011/09/02). Even if the settings of the tool are slightly modified, we observe that the extracted rules and the readings of the results are identical.

We take again a former tutorial to describe the behavior of this component ([Association Rule Learning using A PRIORI PT](#)). Thus, we do not detail the construction of the diagram here. We try above all to highlight the improvement of the library, especially about the computation time. We observe that this improvement is really impressive.

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

We import the [assoc_census.txt](#) data file. We define the diagram. We use the default settings for A PRIORI PT (min support = 0.33, min confidence = 0.75, max itemset length = 4). We launch the calculations, we obtain the following results.

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].
```

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

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

The 'Components' section at the bottom shows the 'Association' component selected, with 'A priori' and 'A priori PT' also visible in the component list.

30.495 rules are extracted.

In a hidden manner, Tanagra creates a temporary file (transaction file format) that is sent to the "apriori.exe" library. The calculation time for this preparation is about 4 seconds. Then, the external library takes over for the processing. We observe that the rule extraction duration is rather fast. The calculation time is mainly penalized by the data file importation (*reading*). We see below that the improvement concerns this step.

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

With the latest version of Tanagra (1.4.41), we obtain the following results.

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

```
D:\Temp\Exe\laxelapriori.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\assoc\output.rul ... [30495 rule(s)] done
```

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

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_65	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 analysis tools 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.

With the same settings, we obtain the same set of rules.

We observe that an additional step is detailed during the process (*sorting and reducing transactions*). We observe above all that the importation (*reading*) duration is drastically reduced. **On the same computer, the reading time fell from 22.93 seconds to 1.66 seconds!** We note that the temporary transaction file size is about 220 MB. It would appear difficult to make better.

The screenshot shows a Windows File Explorer window with the following table of files:

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
Tuxiao001.Mty	24/09/2011 11:27	Fichier MTY	1 Ko

4 Conclusion

Usually, I do not like to running after the versions of the libraries. Because, when we deal with a new version, we must check if the old treatments are still viable, if we obtain the same results than with the older versions. In the case of "apriori.exe", the checks that I made confirm the good behavior of the tool. And we have found that the calculation time is improved in spectacular proportions.