Objectif

Construire des règles d'association sur de gros fichiers de données, utilisation d'un programme externe.

La construction de règles d'association pose de problèmes de performances, tant en occupation mémoire qu'en temps de traitement. L'implémentation actuelle dans TANAGRA est relativement rapide, en revanche, elle est très gourmande en mémoire, au point de la saturer très rapidement dès que l'on a à produire un grand nombre de règles. De plus, l'affichage des règles en HTML est tributaire du composant d'affichage, un peu limité, au point que le temps consacré à l'affichage est parfois aussi important que le temps consacré à l'élaboration des règles.

Il fallait donc se tourner vers un module très performant de construction des règles et proposer une nouvelle fenêtre d'affichage peu sensible au nombre de règles, fussent-elles de plusieurs centaines de milliers.

Sur la création des règles, j'ai découvert les travaux de Christian BORGELT (http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html). Il propose une implémentation réellement impressionnante. Traduire le code en DELPHI m'exposait aux risques de mauvaises interprétations de son travail, et donc d'introduction d'erreurs ; le passage par des DLL est également séduisant mais m'oblige à faire un travail de traduction des structures en C vers DELPHI, toujours hasardeux, pour la définition des unités d'import. J'ai donc décidé d'inaugurer une nouvelle approche avec cette nouvelle version, l'appel à un programme externe avec passage de fichiers temporaires. La rapidité de l'ensemble dépend en grande partie du temps consacré à l'écriture et à la lecture des fichiers temporaires. Force est de constater que le travail de BORGELT est réellement faramineux, son temps de « parsing » et de lecture est plus rapide que mon temps d'écriture du fichier. Au final, cette approche semble viable, du moins tant qu'il n'est pas nécessaire de produire des données qui seront par la suite utilisées dans le diagramme. Nous en montrons un exemple dans ce didacticiel.

L'autre point important était de créer une fenêtre de visualisation des règles qui ne s'effondre pas dès que leur nombre excède la centaine de milliers de règles, et qui par ailleurs, comporte des fonctionnalités de tri selon différents critères. Nous avons donc élaboré un outil simple qui permet de récupérer les sorties de BORGELT et d'afficher simplement les règles dans une fenêtre conviviale.

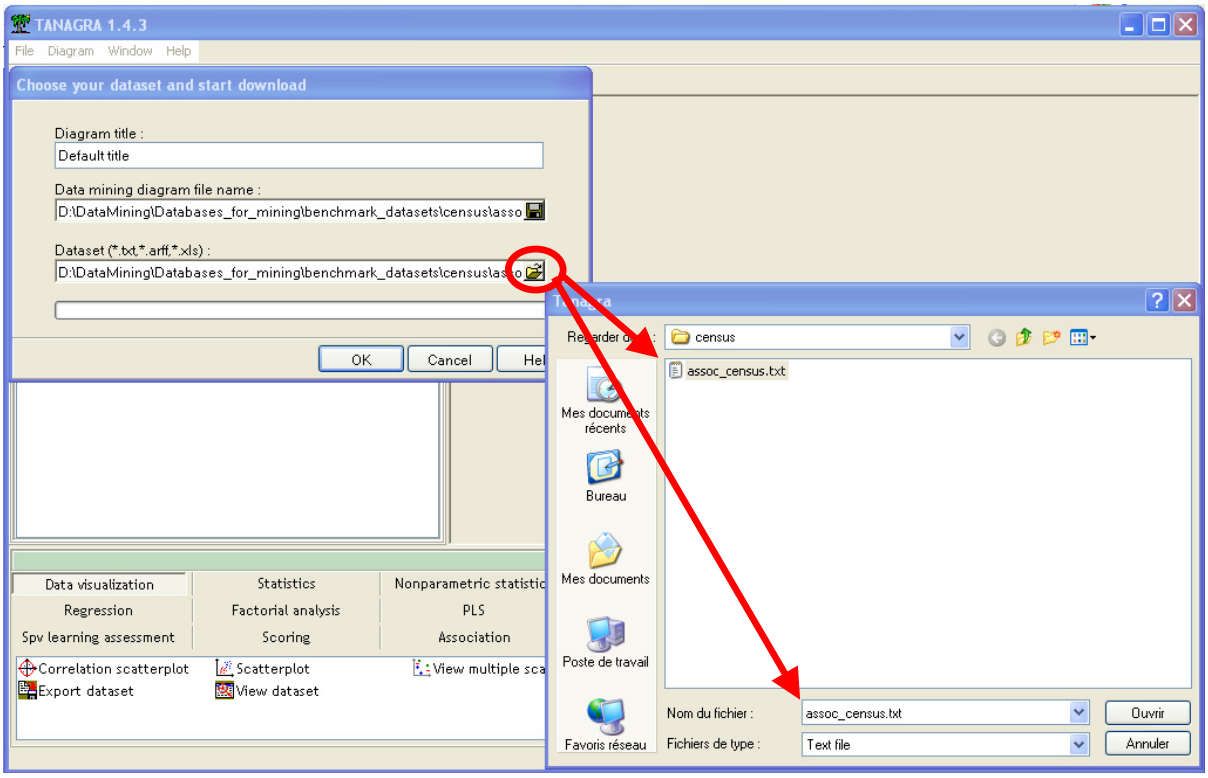
Fichier

Nous avons utilisé le fichier CENSUS[1] (ASSOC_CENSUS.TXT). Nous l'avons expurgé de toutes les variables continues et sélectionné au hasard un sous ensemble d'observations : il contient maintenant 29 variables discrètes et 200 000 observations. Le fichier texte fait approximativement 90 Mo.

Construire des règles d'association

Importer les données

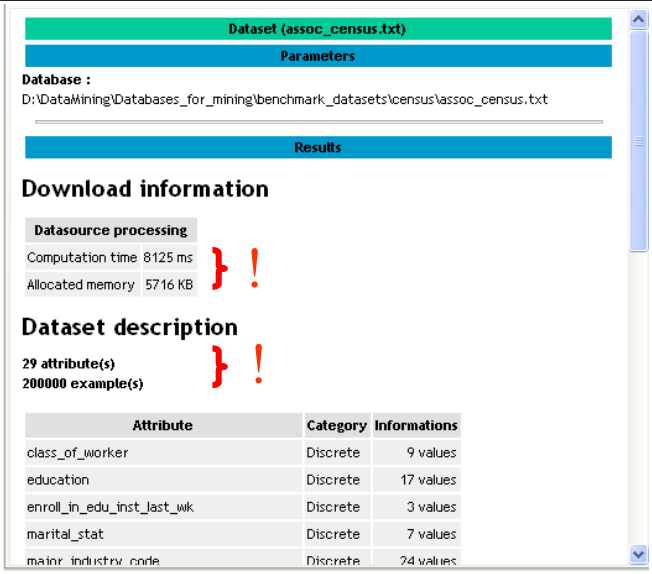
Première étape toujours, importer les données. Nous activons le menu FILE / NEW.



Nous constatons que l'importation est assez rapide (# 8 secondes[2]), les caractéristiques des données sont affichées dans la fenêtre de résultats.

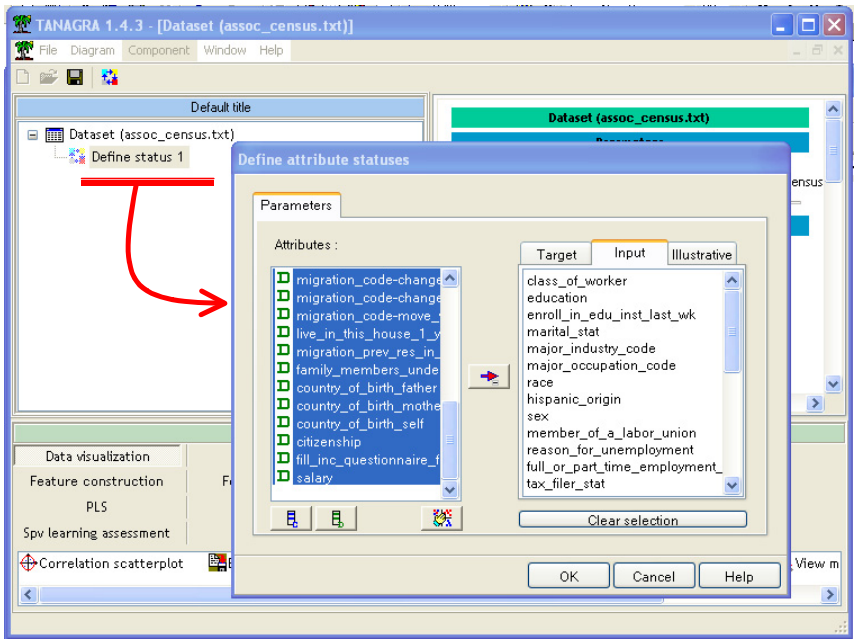
[1] This data was extracted from the census bureau database found at | http://www.census.gov/ftp/pub/DES/www/welcome.html. Donor: Terran Lane and Ronny Kohavi.

[2] CELERON 2,53 Ghz sous XP



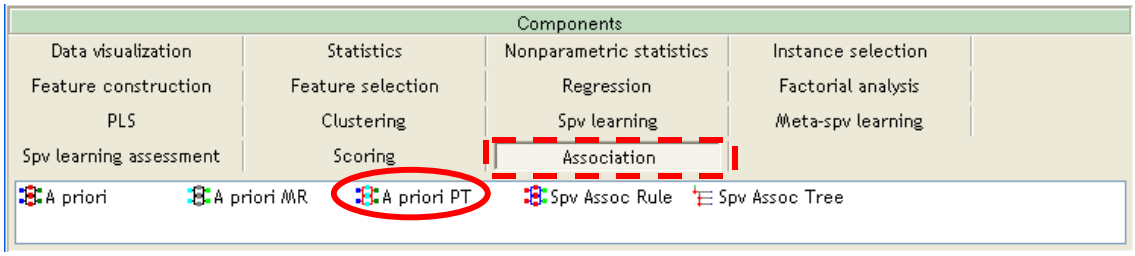
Sélectionner les attributs

Nous plaçons tous les attributs en INPUT.

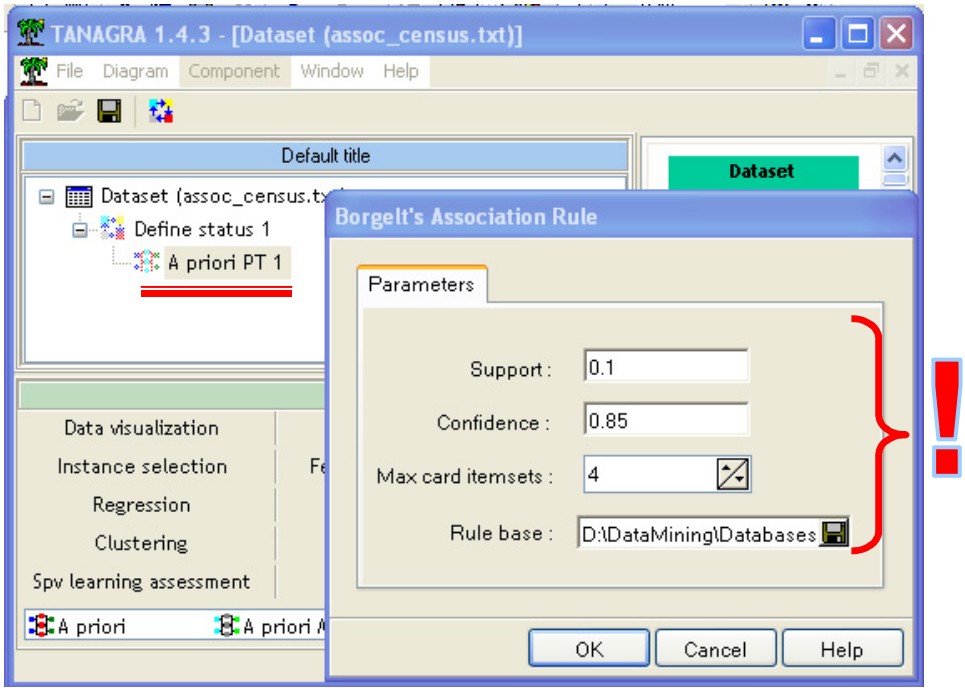


Placer et paramétrer le composant « Règle d'Association »

Nous devons maintenant placer le composant A PRIORI PT, nous le trouvons dans la palette ASSOCIATION.



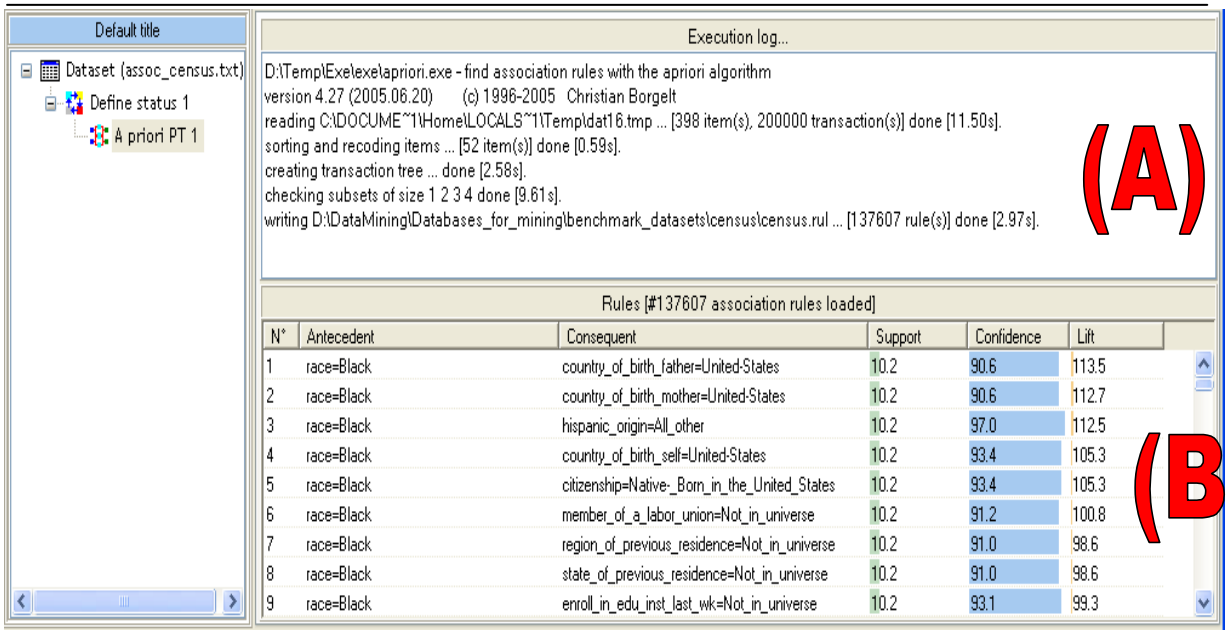
Nous allons configurer le composant en fixant les paramètres de l'analyse (MENU CONTEXTUEL + PARAMETERS).



Nous fixons le SUPPORT MIN à 0.1 (10%), la CONFIANCE MIN à 0.85 (85%), le nombre maximum d'itemsets dans les règles à 4 (MAX CARD ITEMSETS), tous ces paramètres servent à limiter le nombre de règles produites par la méthode ; enfin, nous précisons le nom de la base de règle qui est créée sur le disque dur (RULE BASE), le nombre de règles pouvant être considérable, il est nécessaire de prévoir un emplacement approprié (n'allez pas générer la base de règles sur une disquette ou même sur une clé USB).

Exécution et visualisation des règles

Il nous reste à lancer l'exécution. Après avoir généré le fichier de données temporaire, TANAGRA affiche la fenêtre de résultats, il nous permet de surveiller la progression du programme externe (A) ; lorsque le travail est terminé, les règles sont automatiquement dans la partie basse de la fenêtre (B). Notons au passage le temps de traitement, réellement impressionnant, compte tenu de la taille de la base et du nombre de règles (137 607) générées.



Nous retrouvons dans la fenêtre (B) les informations usuelles sur les règles : l'antécédent de la règle ; le conséquent de la règle ; le support, la confiance et le lift qui sont fournis automatiquement par le programme de Christian BORGELT.

L'intérêt de cette nouvelle fenêtre de visualisation, outre le fait qu'elle gère sans problème un très grand nombre de règles, est qu'elle permet de les trier assez simplement en cliquant dans l'en-tête des colonnes. Si nous voulons par exemple trier les règles selon le SUPPORT, il nous faut tout simplement cliquer dans l'en-tête correspondant. Voici le résultat associé.

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_u	97.0	99.0	100.0
552	reason_for_unemployment=Not_in_universe	race=White	97.0	84.1	100.2