1 Introduction

Mise en œuvre et paramétrage du composant A PRIORI MR.

L'extraction des règles d'association est une approche très populaire pour dégager les interdépendances entre les caractéristiques des individus. Elle a beaucoup été utilisée pour étudier les achats concomitants chez les consommateurs. Le résultat se présente sous la forme d'une règle logique du type « SI un individu a acheté tel ou tel produit ALORS il achètera également tel et tel produit ». Bien entendu, il est possible d'étendre le champ d'application de la méthode à d'autres domaines.

Nous avons présenté les règles d'association à plusieurs reprises dans nos didacticiels[1]. La méthode A PRIORI est certainement la plus connue[2]. Malgré ses qualités, l'approche présente un écueil fort : le nombre de règles produites peut être très élevé. La capacité à mettre en avant les « meilleures » règles, celles qui sont porteuses d'informations « intéressantes », devient ainsi un enjeu fort.

Ces dernières années, on a vu fleurir un nombre impressionnant de publications cherchant à proposer des mesures d'intérêt des règles. Leur mise en œuvre est simple : on assigne un score (mesure d'intérêt) à chaque règle, on trie alors la base de règles de manière à ce que celles qui sont les plus informatives apparaissent en premier.

Le composant A PRIORI MR (onglet ASSOCIATION) est un outil expérimental qui propose plusieurs mesures d'évaluation des règles. Il met en avant, entres autres, le concept de « valeur-test ». C'est une mesure statistique développée par A. Morineau (1984), décrite dans un ouvrage (Lebart, Morineau et Piron, 2000), et largement utilisée dans le logiciel commercial SPAD (http://www.spad.eu/).

A PRIORI MR propose plusieurs variantes de la valeur test issues de travaux récents (Morineau et Rakotomalala, 2006). Elles diffèrent par l'écriture de l'hypothèse nulle et le schéma d'échantillonnage utilisé. La description technique des mesures est disponible dans un document spécifique en ligne (http://tutoriels-data-mining.blogspot.com/2009/02/mesures-dinteret-des-regles-dans-priori.html). Dans ce tutoriel, nous nous attarderons avant tout sur la mise en œuvre, le paramétrage et la lecture des résultats fournis par A PRIORI MR.

2 Données

Nous utilisons une version modifiée du fichier GERMAN CREDIT[3]. Il décrit les caractéristiques de demandeurs de crédit. Nous avons discrétisé les variables quantitatives.

Nous attribuerons un statut identique à toutes les variables dans ce didacticiel. Nous cherchons les interdépendances qui peuvent exister entre les caractéristiques des individus. Le fichier est accessible en ligne (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_assoc.xls).

---

[1] http://tutoriels-data-mining.blogspot.com/search/label/R%C3%A8gles%20d%27association

[2] http://tutoriels-data-mining.blogspot.com/2008/04/rgles-dassociation-algorithme-priori.html

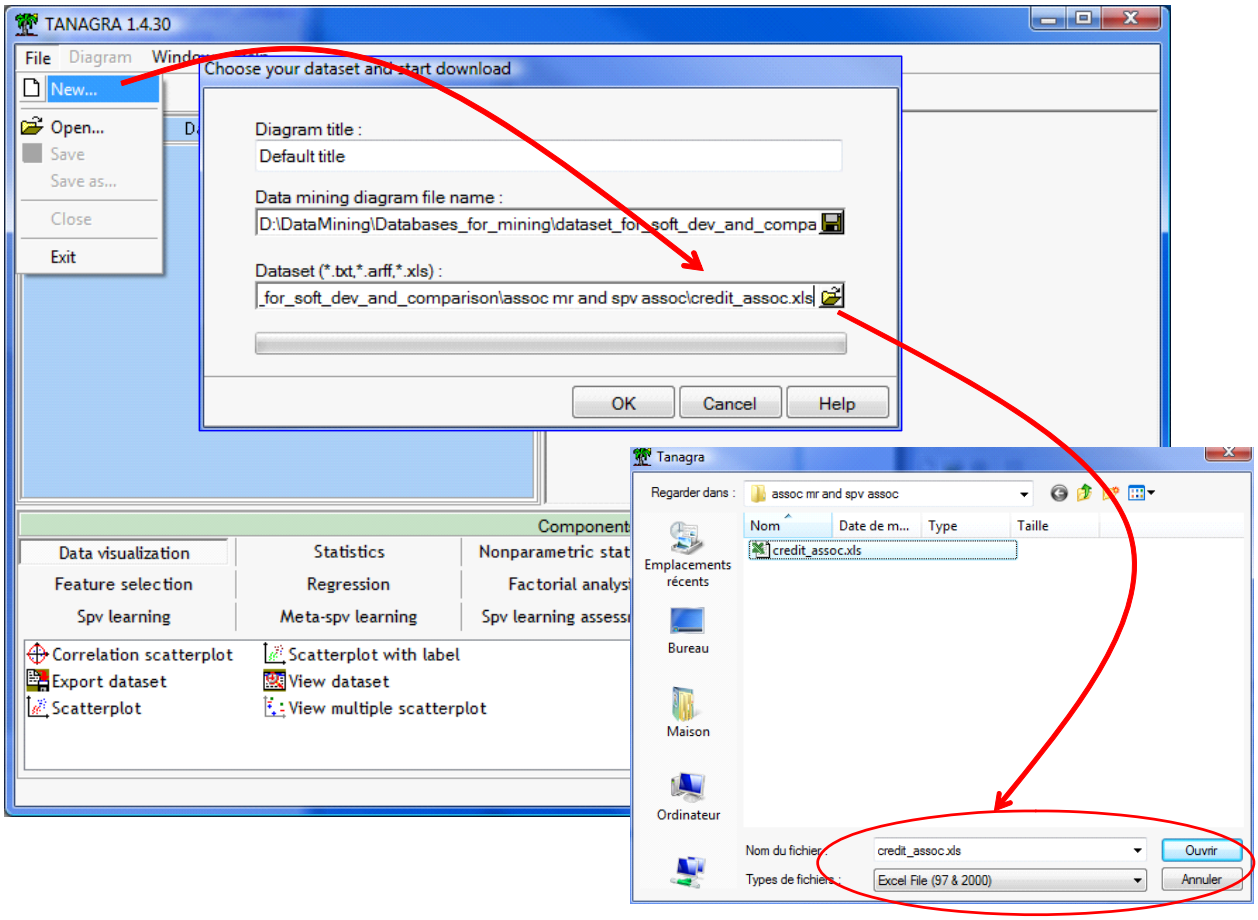
[3] http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

3 Le composant A PRIORI MR

3.1 Créer un diagramme et importer les données

Il y a plusieurs manières d'importer un fichier Excel (XLS) dans Tanagra. Nous pouvons notamment le lire directement pourvu que (1) le fichier ne soit pas en cours d'édition, (2) que les données soient situées dans la première feuille du classeur.

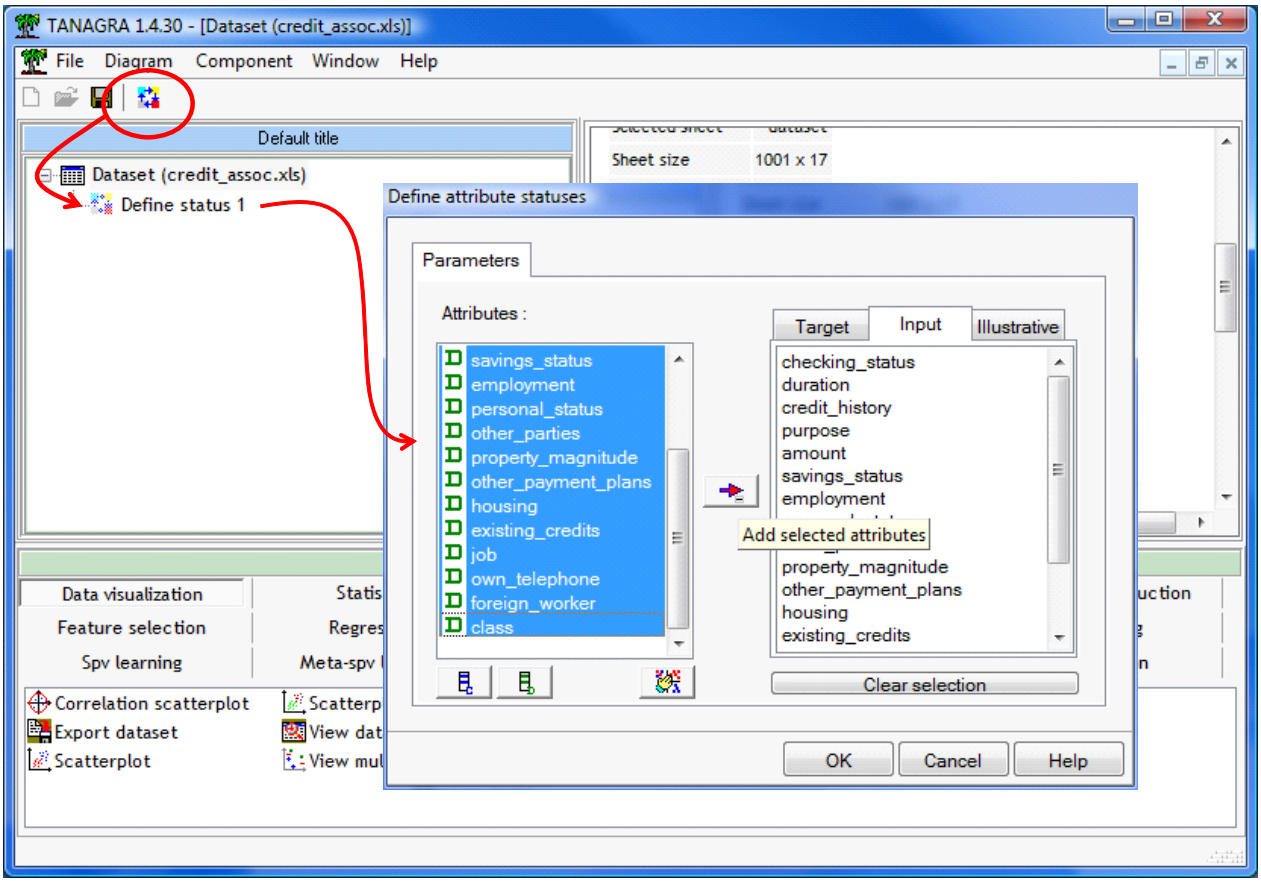
Nous actionnons le menu FILE / NEW pour créer un nouveau diagramme. Nous sélectionnons le fichier CREDIT_ASSOC.XLS.



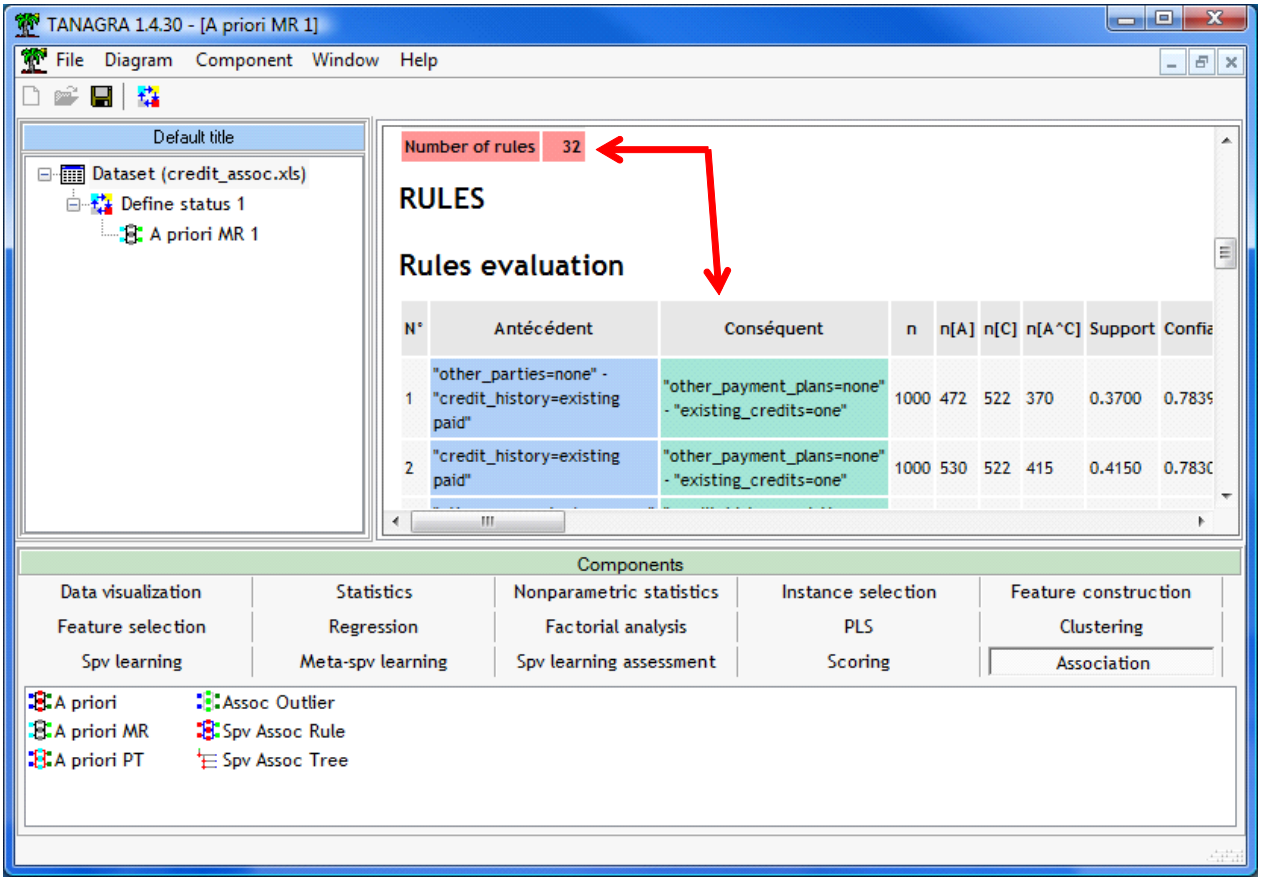
17 attributs et 1000 observations en provenance de la feuille DATASET sont disponibles pour les traitements.

3.2 A PRIORI MR

Pour désigner les variables à analyser, nous introduisons le composant DEFINE STATUS via le raccourci dans la barre d'outils. Nous les plaçons toutes en INPUT.



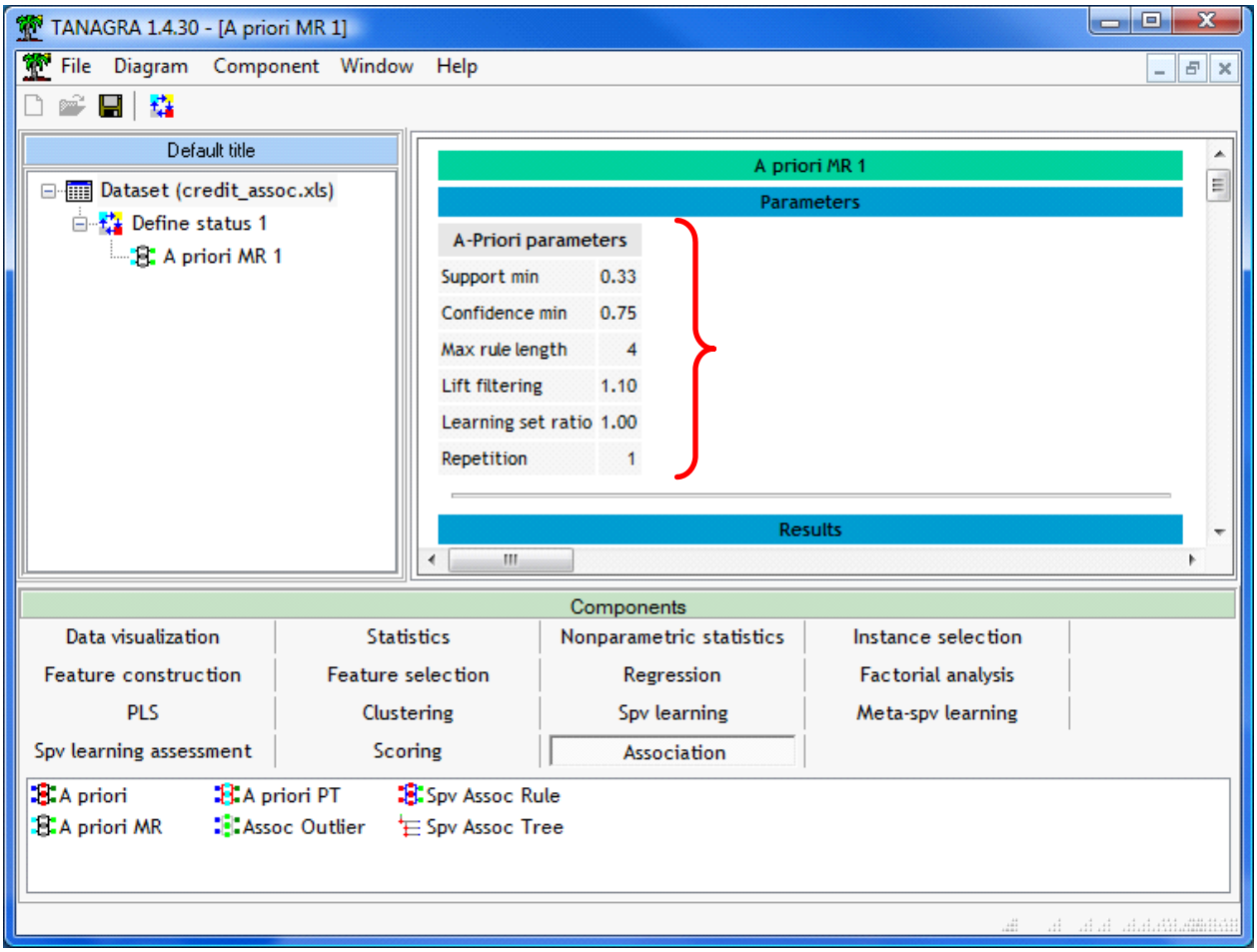
Nous insérons le composant A PRIORI MR (onglet ASSOCIATION) dans le diagramme. Pour obtenir un premier résultat avec les paramètres par défaut, nous actionnons le menu contextuel VIEW.



Tanagra produit 32 règles énumérées dans la partie basse de la fenêtre de visualisation. Elles sont triées selon le LIFT décroissant.

3.3 Les paramètres de A PRIORI MR

Plusieurs informations importantes sont listées dans la partie PARAMETERS de la fenêtre de visualisation :



- SUPPORT MIN indique le support minimum des règles produites, en deçà de ce seuil la règle n'est pas acceptée ;
- CONFIDENCE MIN est la confiance minimum ;
- MAX RULE LENGTH est le nombre maximum d'items (de couples attribut = valeur) dans les règles ;
- LIFT FILTERING est le LIFT minimum.

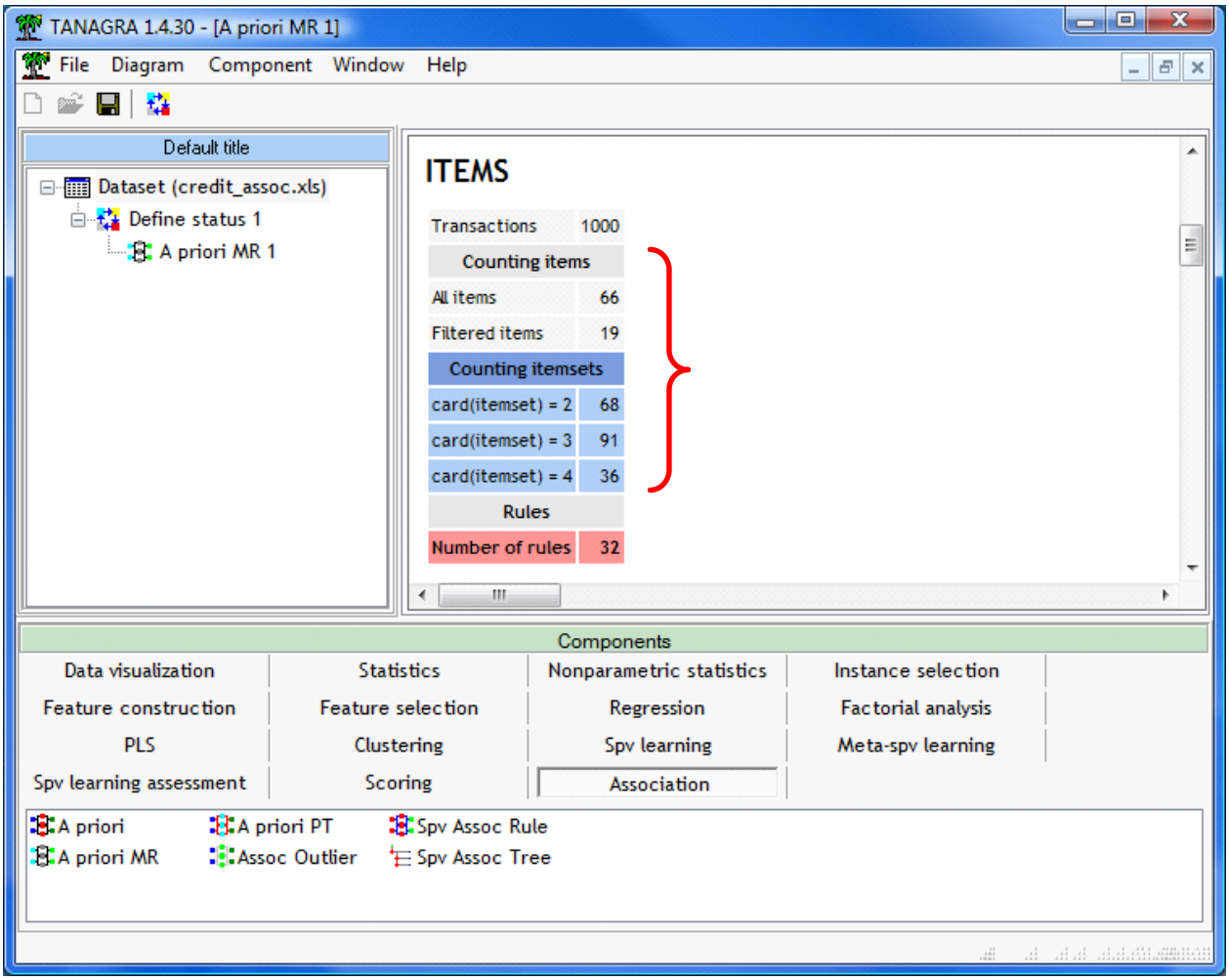
Ces paramètres permettent de limiter le nombre de règles. Les 3 premières permettent même de contrôler la durée des calculs et l'occupation mémoire. Le critère LIFT en revanche n'agit qu'a posteriori, pour filtrer les règles déjà produites.

- LEARNING SET RATIO indique la proportion des données utilisées pour construire les règles. En effet, Tanagra sait scinder les données en 2 parties : la première, l'échantillon d'apprentissage, sert à élaborer les règles ; la seconde, l'échantillon test, sert à les évaluer. Cette idée, très pratiquée en apprentissage supervisé, est moins usuelle dans l'extraction des règles d'association. Elle permet d'évaluer la stabilité des solutions proposées par l'algorithme. La valeur par défaut du paramètre est 1 c.-à-d. la totalité des observations sont dévolues à l'apprentissage.

- REPETITION est le nombre de réplications utilisées lors du calcul de la VT-100 à l'aide de la procédure de Monte-Carlo.

3.4 Les résultats fournis par A PRIORI MR

La partie ITEMS des résultats indique le nombre d'itemsets fréquents (dont le support est supérieur au SUPPORT MIN) regroupés par cardinalité. Le nombre total d'items (ATTRIBUT = VALEUR) est 66 dans ce fichier, 19 sont fréquents. Pour les couples d'items, nous en avons 68, pour les triplets, 91, etc.

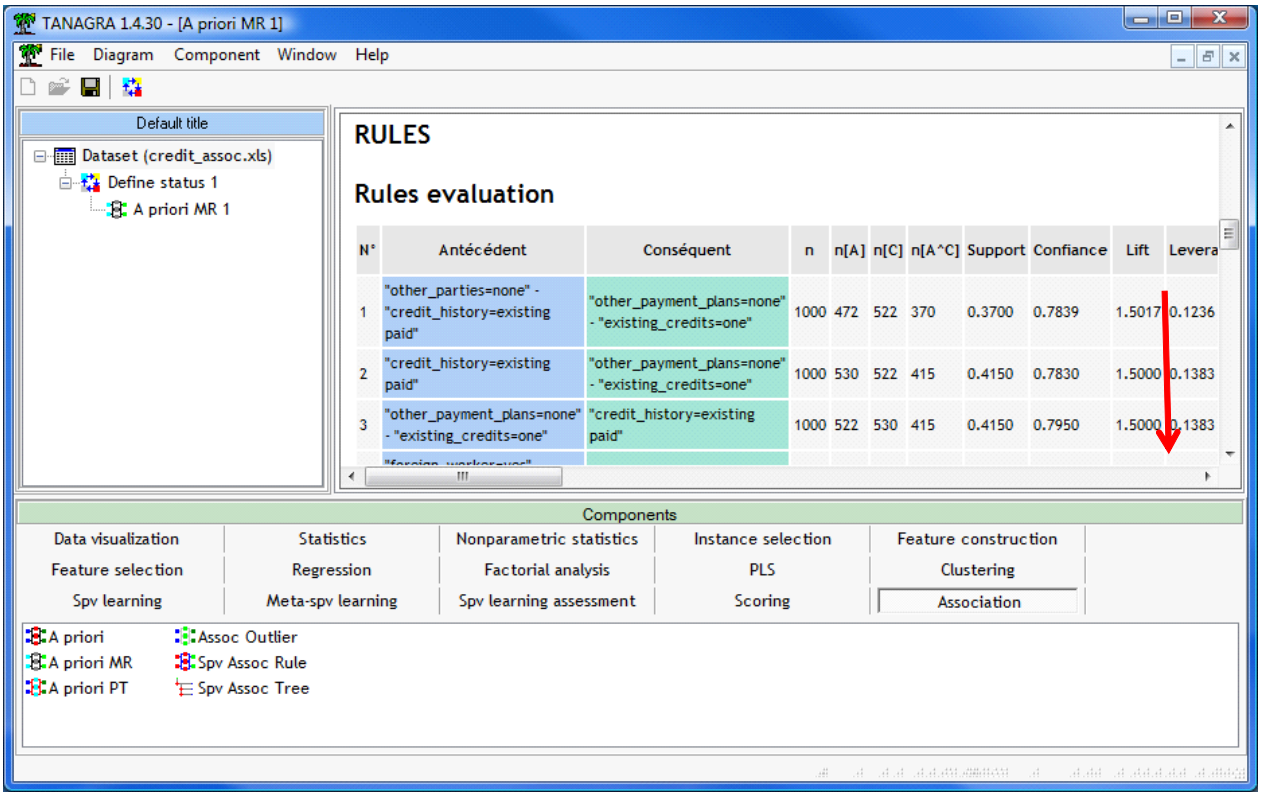


Nous pouvons diminuer le nombre d'itemsets de 2 manières : en augmentant le SUPPORT MIN ; en diminuant le MAX RULE LENGTH, qui correspond ici au cardinal maximum des itemsets que nous explorons.

Dans un deuxième temps, Tanagra va tenter d'extraire des règles à partir des itemsets de cardinal supérieur à 2. Le critère CONFIDENCE MIN joue à ce stade. Si nous l'augmentons, nous serons plus exigeant sur les règles à produire. Nous en obtiendrons moins.

La section RULES décrit les règles extraites par l'algorithme. Nous avons l'antécédent de la règle (la partie SI…) et le conséquent (la partie ALORS…).

Suivent une série d'indicateurs qualifiant la crédibilité de la règle. Nous les décrivons de manière détaillée dans un document accessible en ligne (…). Les indicateurs usuels que l'on retrouve dans la majorité des logiciels sont le SUPPORT, la CONFIANCE et le LIFT. La particularité de Tanagra est de proposer plusieurs variantes de la mesure VALEUR TEST.

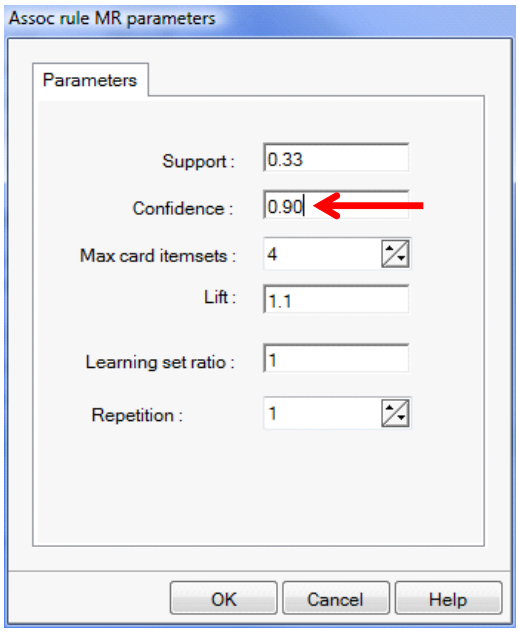


4 Manipuler les paramètres de A PRIORI MR

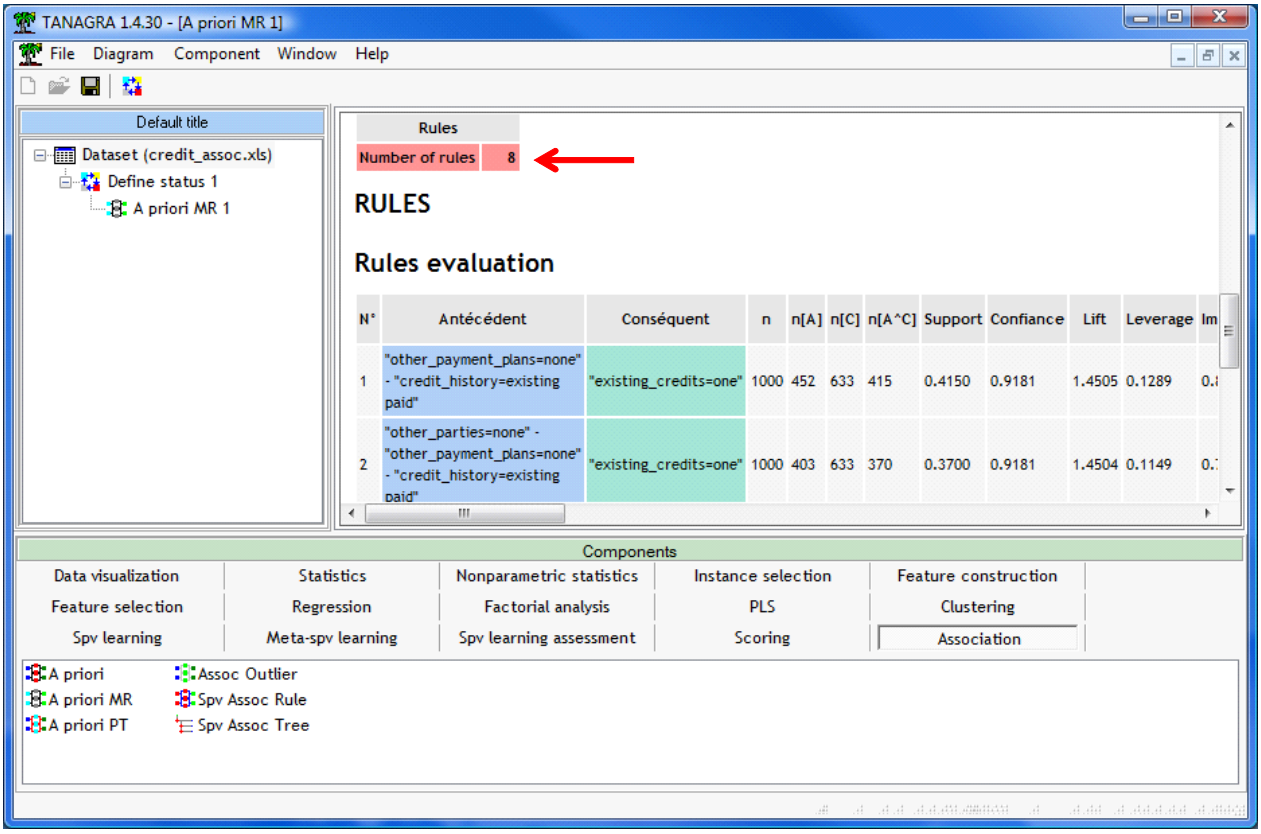
4.1 Obtenir moins de règles

La profusion des règles est le principal défaut des algorithmes d'extraction des règles d'association. Pour en limiter le nombre, nous pouvons agir sur plusieurs paramètres.

Mettons que nous souhaitons obtenir des règles plus précises, avec une confiance supérieure à 90%. Nous actionnons le menu contextuel PARAMETERS, nous augmentons la valeur de CONFIDENCE MIN à 0.90 (attention au point décimal, il dépend de la configuration de votre système).



Nous validons ce choix. Nous cliquons de nouveau sur VIEW.



Nous obtenons maintenant 8 règles.

4.2 Explorer une base de règles

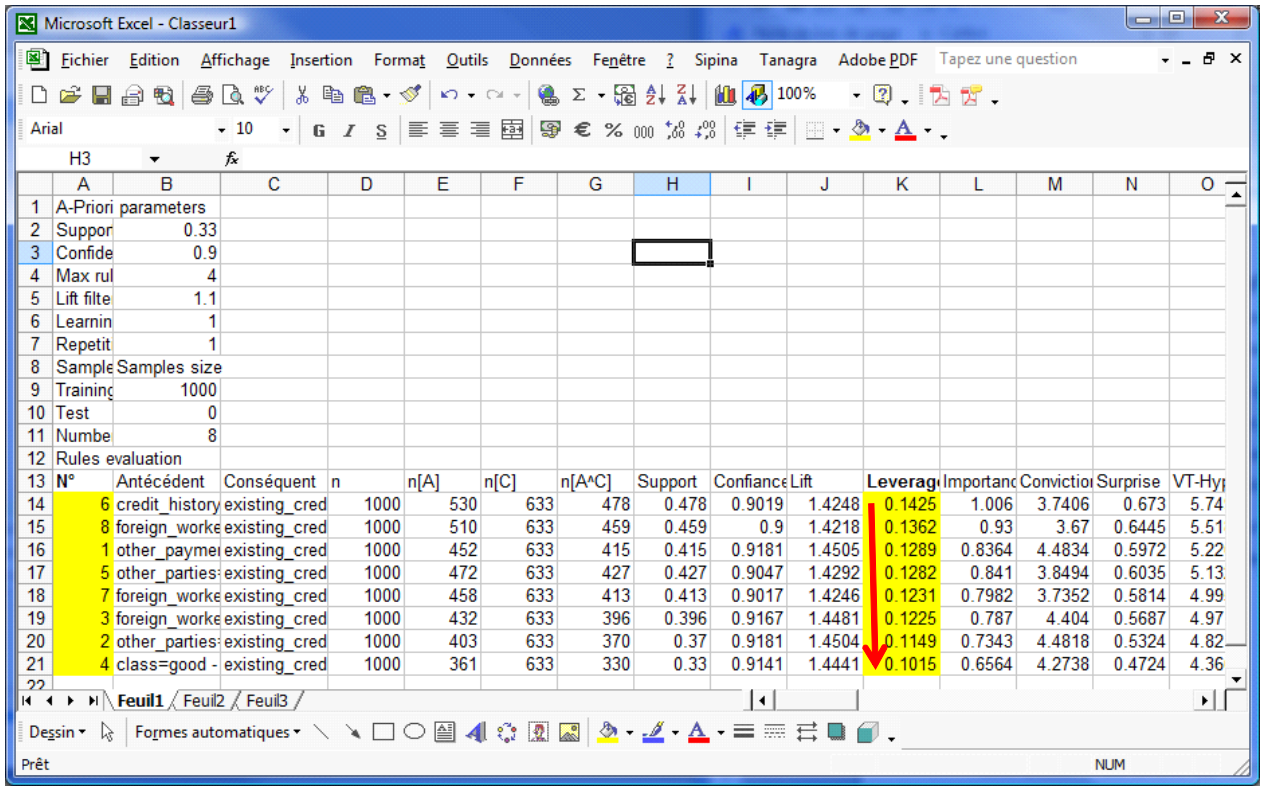
Filtrer les règles est une option intéressante. Encore faut-il savoir comment fixer les valeurs adéquates des paramètres. Procéder par tâtonnement peut se révéler très gourmand en temps de calcul.

Tanagra propose une fonctionnalité qui permet d'aller plus loin dans l'exploration des règles. Nous pouvons copier les résultats dans un tableur, et ainsi bénéficier des fonctionnalités de ce dernier en matière de filtrage et de tri des listes. Nous actionnons le menu COMPONENT / UNFORMATTED COPY. Puis nous lançons le tableur EXCEL, nous y copions alors le tableau de résultats[4].

A l'aide des outils du tableur, nous pouvons filtrer les règles en nous basant sur des combinaisons de critères. Nous pouvons également les trier. Dans la copie d'écran ci-dessous, les règles sont triées selon le critère LEVERAGE décroissant.

[4] COMPONENT / UNFORMATTED COPY réalise une copie brute, sans formatage. Les valeurs sont simplement positionnées dans les différentes cellules du tableur. C'est l'outil à privilégier lorsque le nombre de règles est élevé.

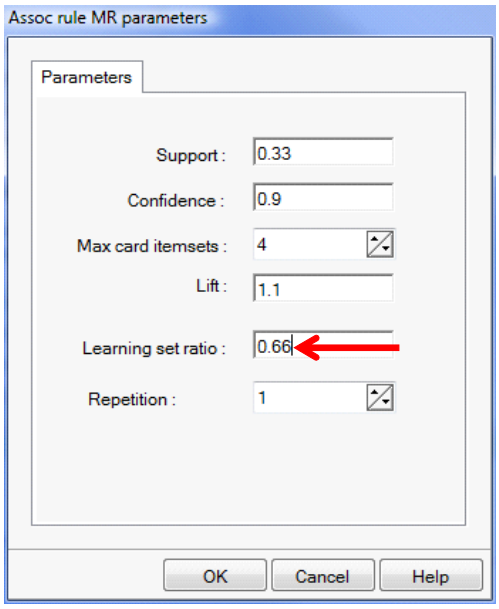
COMPONENT / COPY RESULTS réalise également une copie, mais au format HTML. Nous pouvons coller les résultats dans le tableur EXCEL, le formatage initial sera respecté (couleurs, mise en forme, etc.).



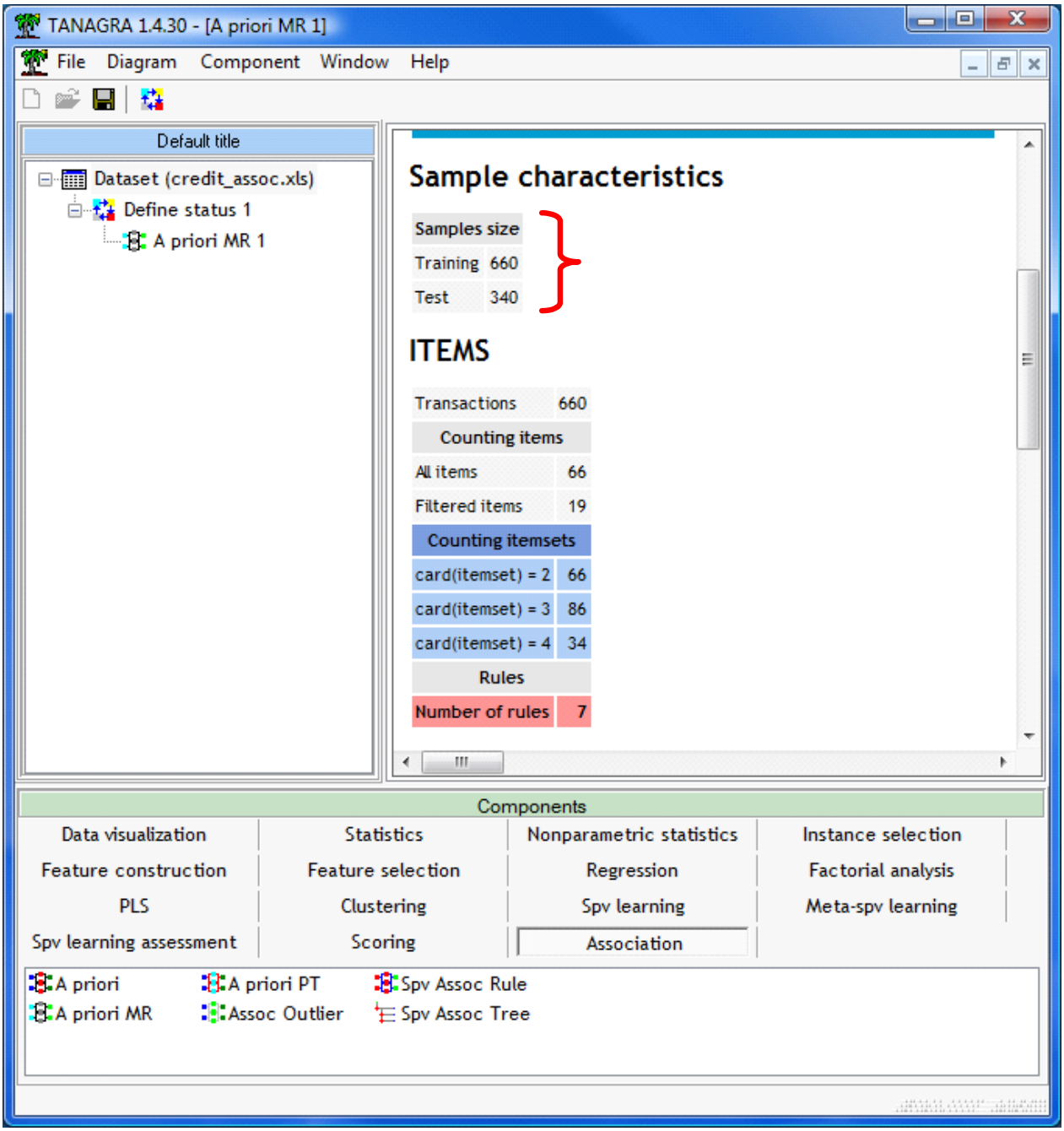
4.3 Partitionner les données en échantillons d'apprentissage et de test

Tanagra propose une option originale s'agissant de l'exploration des règles d'association. Nous pouvons subdiviser la base de données en 2 parties pour la construction et l'évaluation des règles. L'idée est de pouvoir éprouver la crédibilité de la règle en la confrontant à un échantillon n'ayant pas servi à son élaboration.

Revenons sur le menu contextuel PARAMETERS, nous passons le LEARNING SET RATIO à 0.66 (attention toujours au point décimal, il dépend de votre système) : 660 observations seront dédiées à la construction des règles, les 340 autres seront uniquement utilisés pour calculer les indicateurs de qualité.



Nous validons et nous cliquons sur VIEW.



7 règles sont produites. Dans le tableau énumérant les règles, la colonne TEST (en rouge) sépare les valeurs obtenues en apprentissage et en test. Si l'on souhaite étudier finement la stabilité d'une règle, on pourra comparer les valeurs obtenues en apprentissage et en test.

5 Conclusion

La composant A PRIORI MR de Tanagra extrait des règles d'association en se basant sur l'algorithme A PRIORI. Il se démarque des autres logiciels libres en proposant des outils supplémentaires (plusieurs mesures d'évaluation des règles entre autres) qui nous offrent la possibilité d'étudier finement les résultats.