Objectif

Construire des règles d'association à partir d'un fichier binaire.

Fichier

Si les règles d'association peuvent être construites à partir d'un fichier « attribut-valeur » que l'on a recodé en interne, il arrive aussi que le fichier soit binaire par nature. Dans ce cas, le recodage n'est absolument pas nécessaire, il faut traiter le fichier tel quel en se focalisant sur la présence des items dans les transactions.

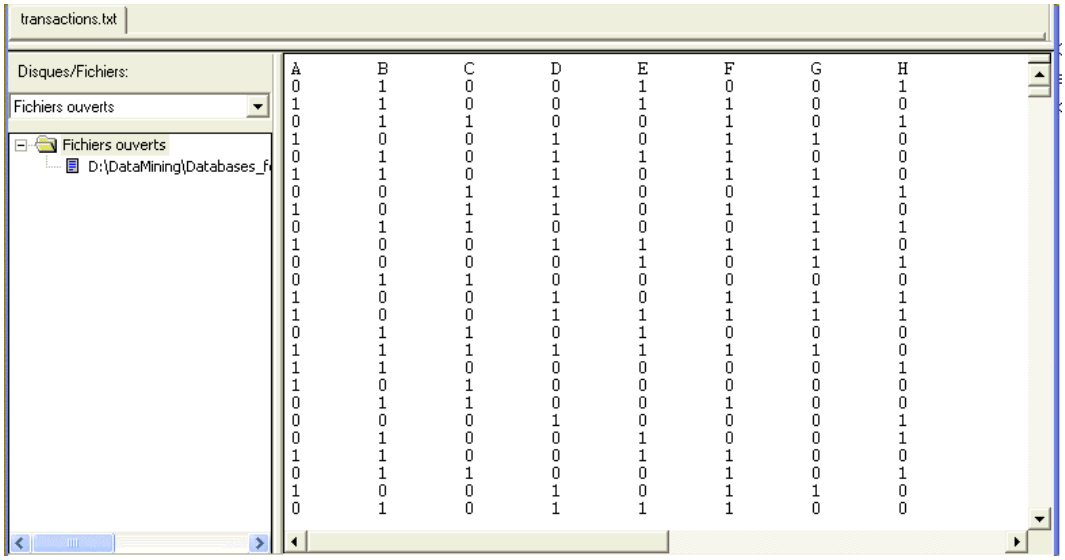
TANAGRA ne gère que les fichiers tabulaires, pour spécifier la présence ou absence d'un item par chaque transaction, on utilisera la valeur 0 (absence) et 1 (présence). Lors de l'importation, ces données étant numériques, il les importe comme une variable continue.

TANAGRA peut donc construire des règles d'association sur des variables continues, sachant qu'il interprète la valeur 0 comme une absence d'un item dans la transaction, et toute autre valeur (on s'attend à ce que ce soit 1 si le fichier est correctement construit) comme une présence.

Construire des règles d'association sur un fichier binaire

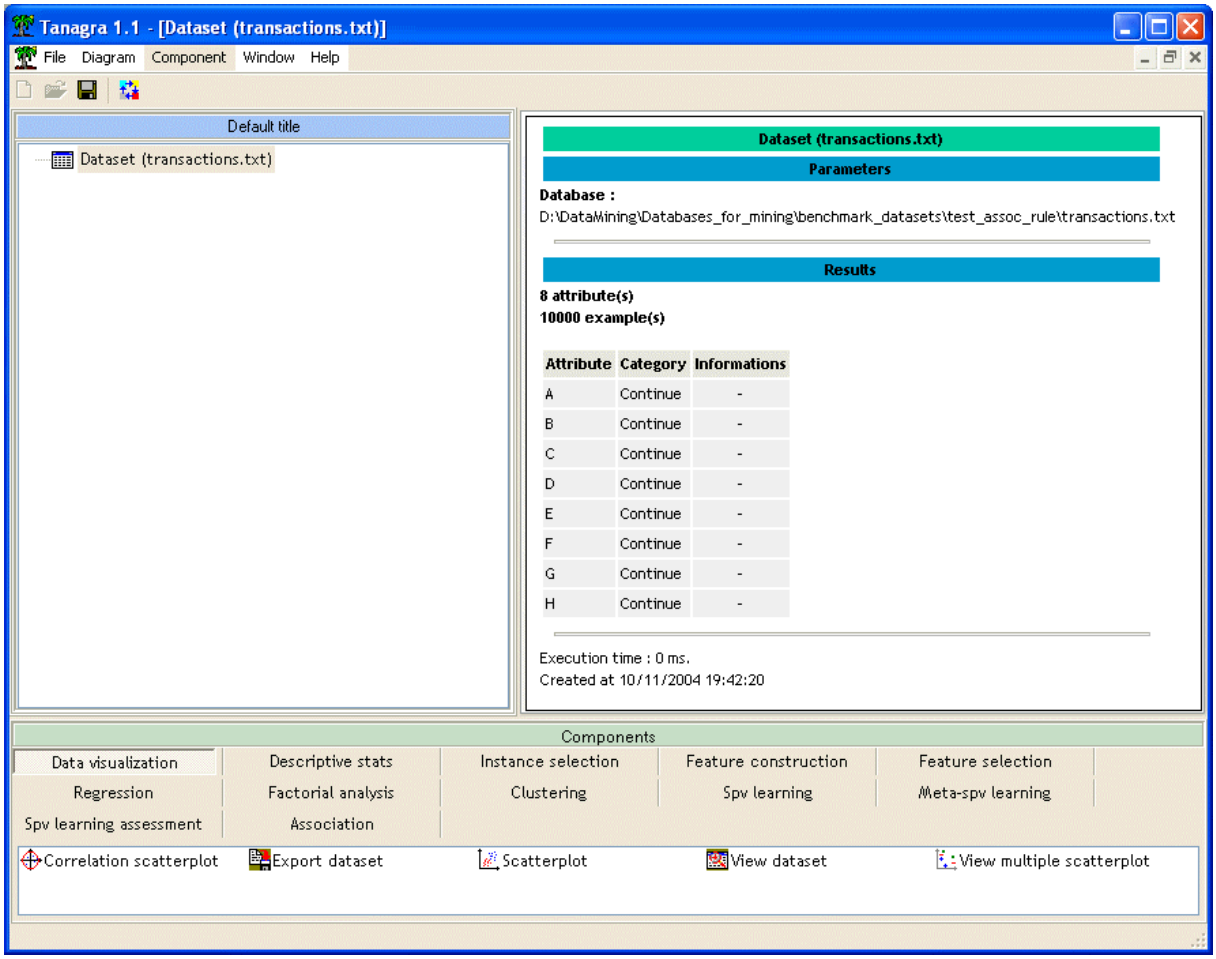
Données

Les données proviennent du web[1], elles représentent la présence de 8 items dans 10000 transactions. Pour que TANAGRA puisse charger le fichier, il faut le préparer sous une forme tabulaire de présence (1) / absence (0) d'un attribut pour chaque observation. Utilisez la tabulation comme séparateur de colonnes et enregistrez-le au format texte.



[1] http://www2.cs.uregina.ca/~dbd/cs831/notes/itemsets/itemset_prog1.html

Il est dès lors possible d'importer le fichier dans TANAGRA, il considère que les variables sont toutes continues.

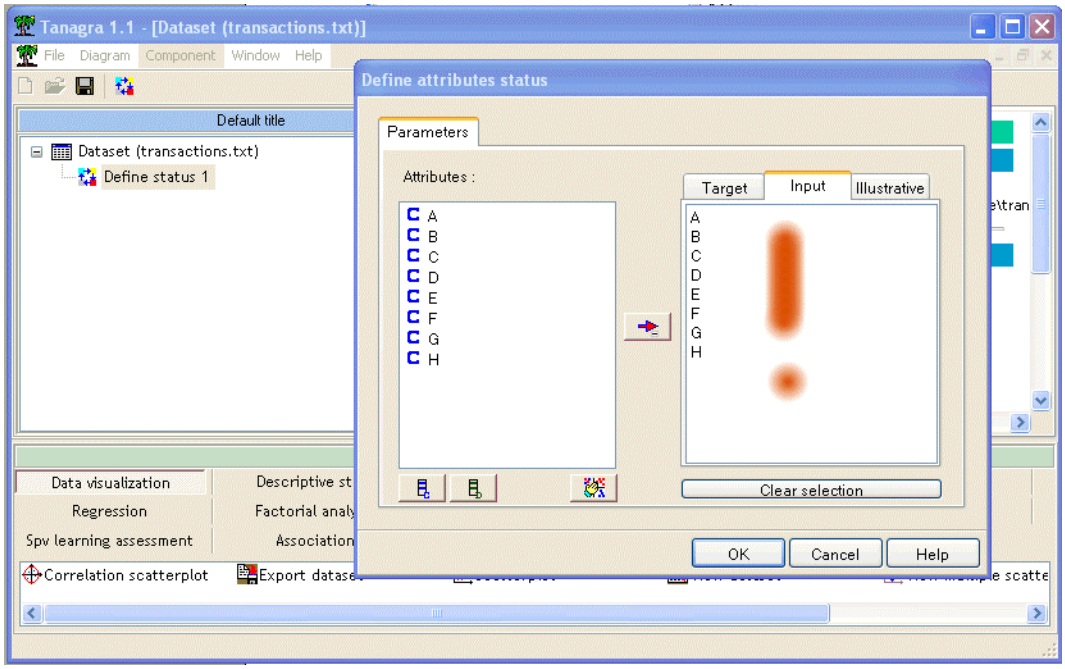


Sélection des attributs

L'étape suivante consiste à choisir les variables pour l'analyse, dans notre cas, il faut toutes les sélectionner.

TANAGRA accepte deux types de sélection pour la construction des règles d'association :

1. Toutes les variables sont discrètes. Dans ce cas il procède à un recodage pour obtenir un tableau interne de présence-absence. Chaque valeur d'une variable correspond à un item.
2. Toutes les variables sont continues. Dans ce cas, le recodage interne consiste à associer un item à chaque variable, et à coder absence la valeur 0, présence toute autre valeur (1 si le fichier est correctement construit).



Analyse

Il ne reste plus qu'à placer le composant A PRIORI qui construit les règles d'association. Seules les règles « positives », correspondant à la présence des items, sont extraites.