1 Objectif

Induction de règles d'association à partir de bases transactionnelles.

L'extraction des règles d'association est une des applications phares du data mining. L'idée est de mettre à jour des régularités, sous forme de cooccurrences, dans les bases de données. L'exemple emblématique est l'analyse des tickets de caisses des grandes surfaces : on veut découvrir des règles de comportement du type « si le client a acheté des couches et des lingettes, il va acheter du lait de croissance ». Auquel cas, il est peut être opportun de mettre les rayons adéquats dans la même zone du magasin (c'est le cas en ce qui concerne l'hypermarché que je fréquente habituellement). La partie « si » de la règle est appelée « antécédent », la partie « alors » est le « conséquent ».

transaction	produit
1	B
1	E
1	H
2	A
2	B
2	E
2	F
3	B
3	C
3	F
3	H

Il est possible de rechercher des cooccurrences dans les tableaux individus – variables que l'on manipule avec les logiciels de Data Mining usuels. Mais bien souvent, surtout dans le cadre de l'induction des règles d'association, les données peuvent se présenter sous la forme d'une base transactionnelle. Si l'on reprend l'exemple de l'analyse des tickets de caisse, nous disposons d'une liste de produits par caddie.

Dans notre tableau exemple ci-joint, nous lisons : dans le caddie n°1 sont présents les produits B, E et H ; dans le caddie n° 2, nous avons trouvé les produits A, B, E et F ; etc.

Cette représentation des données est assez naturelle eu égard au problème que l'on souhaite traiter. Elle présente aussi l'avantage d'être plus compacte puisque seuls sont effectivement listés les produits observés dans chaque caddie. Nous n'avons pas besoin de nous préoccuper des produits qui n'y sont pas, surtout qu'ils peuvent être très nombreux si l'on se réfère aux nombre d'articles que peut proposer une enseigne de grande distribution.

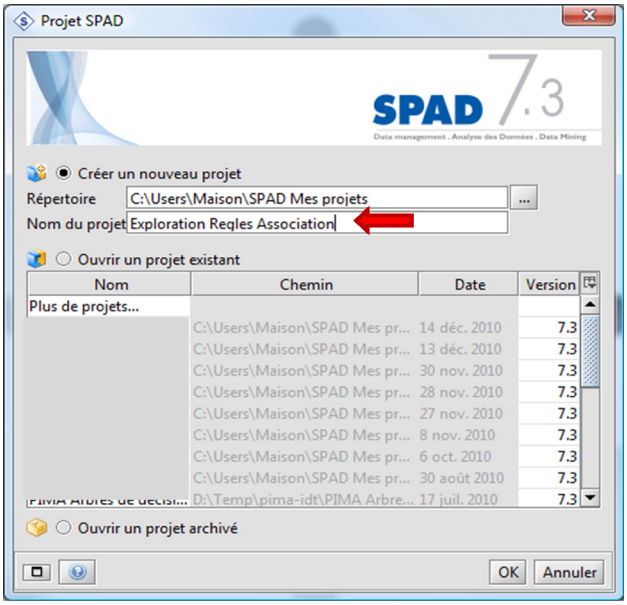
Pour autant que ce mode de description soit naturel, il s'avère que de nombreux logiciels ne savent pas l'appréhender directement. On observe curieusement un vrai clivage entre les outils à vocation professionnelle et ceux issus du monde universitaire. Les premiers savent pour la plupart manipuler ce type de fichier. C'est le cas des logiciels **SPAD 7.3** et **SAS Enterprise Miner 4.3** que nous étudions dans ce didacticiel. Les seconds en revanche demandent une transformation préalable des données pour pouvoir fonctionner. Nous utiliserons une macro VBA fonctionnant sous Excel pour transformer nos données en base « individus – variables » binaire propice au traitement sous **Tanagra 1.4.37** et **Knime 2.2.2**. Attention, nous devons respecter le cahier des charges initial, à savoir s'intéresser uniquement aux règles signalant la présence simultanée des produits dans les caddies. Il n'est pas question, consécutivement à un codage « présent – absent » mal maîtrisé, de produire des règles mettant en évidence l'absence simultanée de certains produits. Cela peut être intéressant dans certains cas peut être, mais ce n'est pas l'objectif de notre analyse.

2 Données

La base « transactions.txt » décrit la composition de 10.000 caddies. 8 produits sont référencés. Il s'agit bien entendu d'un jeu de données synthétique utilisé pour évaluer les algorithmes d'extractions de règles. Nous l'avons déjà exploité dans un précédent didacticiel (http://tutoriels-

data-mining.blogspot.com/2008/04/priori-sur-les-bases-transactionnelles.html). L'originalité ici, du moins dans un premier temps avec SPAD et SAS, est que nous traitons directement les données organisées sous la forme d'une base de transactions.

3 Analyse avec SPAD 7.3



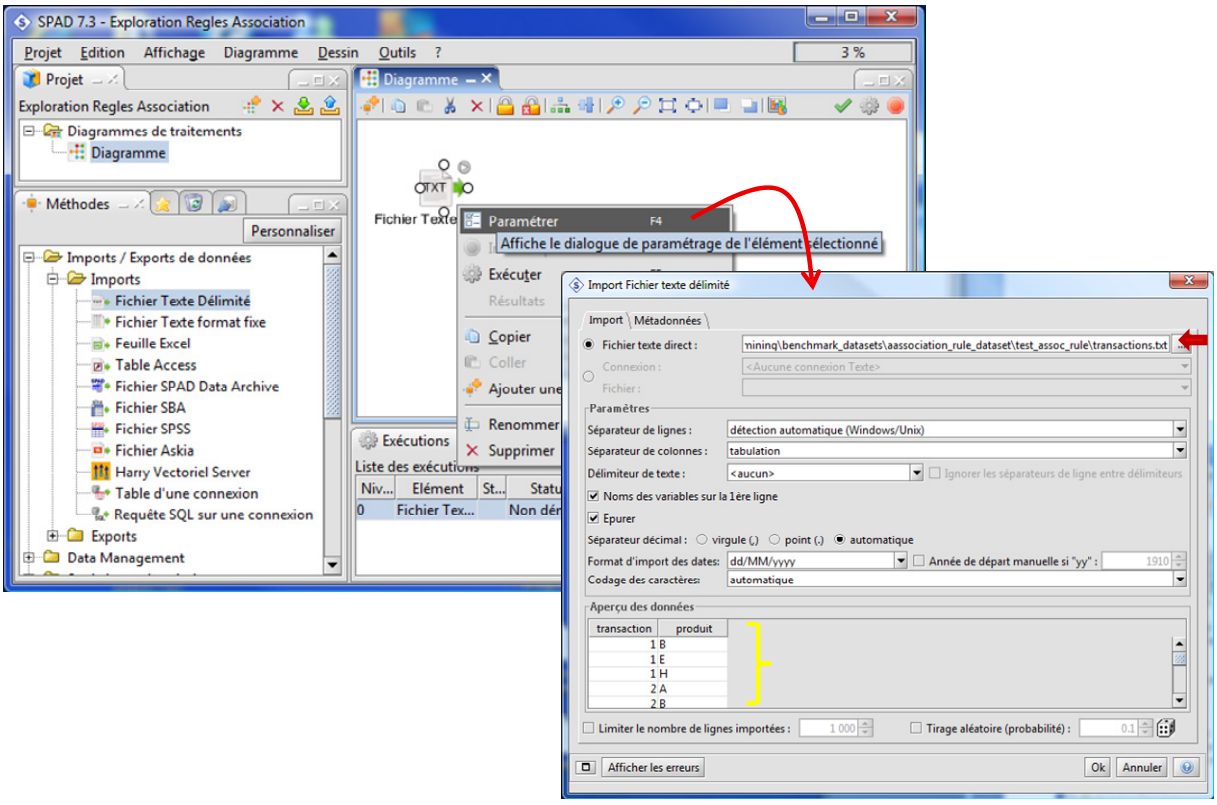
Le logiciel SPAD est un logiciel connu et reconnu dans le monde de l'Analyse de Données et du Data Mining. Nous utilisons la version 7.3 dans ce didacticiel.

3.1 Création d'un diagramme

Au démarrage du logiciel, une fenêtre d'accueil nous permet de charger un diagramme existant ou d'en créer un nouveau. Nous choisissons cette seconde option, nous lui attribuons un nom « Exploration Règles Association » par exemple.

3.2 Importation des données

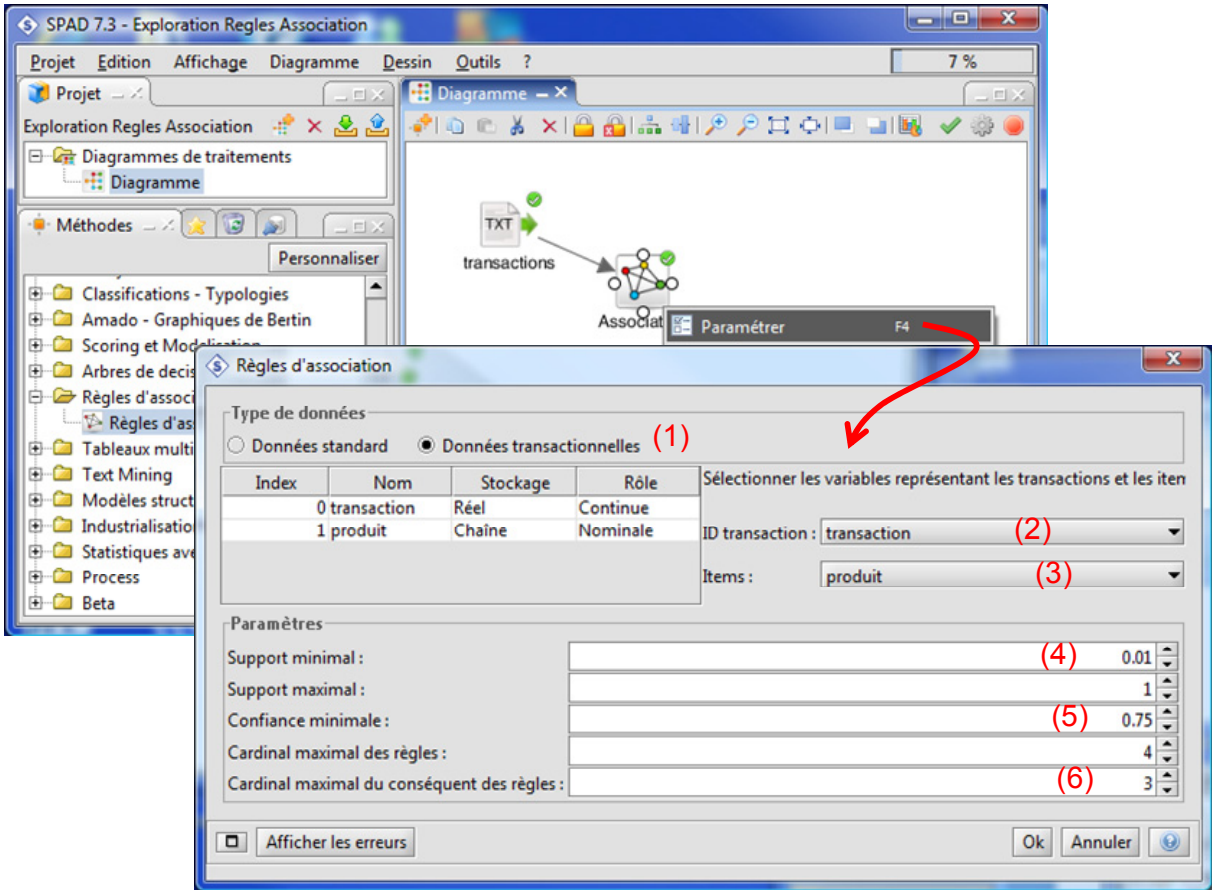
Nous utilisons l'outil « Fichier Texte Délimité » pour charger les données. Nous le paramétrons (menu contextuel PARAMETRES) de la manière suivante pour appréhender notre fichier « transactions.txt ».



Il comporte 39000 lignes (sur 2 colonnes), confirmé par l'indicateur fourni par SPAD.

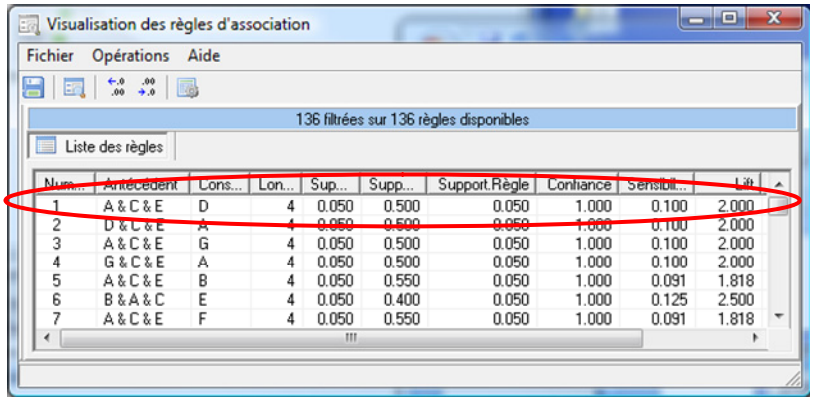
3.3 Extraction des règles

L'outil « Règles d'Association » nous servira à extraire les règles. Nous lui relions le composant d'accès aux données. Nous spécifions les paramètres suivants (menu PARAMETRES) :



Nous utilisons bien des « données transactionnelles » (1). L'identifiant de la transaction est la première colonne « transaction » (2). Les produits sont décrits sur la seconde colonne « produit » (3). Nous baissons le support minimal des règles à 0.01 c.-à-d. une règle doit couvrir au moins (0.01 x 10000) 100 individus (4). La confiance minimale des règles est fixée à 0.75 (5). Enfin, si le cardinal (le nombre d'items) maximum des règles est laissé à 4, nous souhaitons pouvoir obtenir des règles proposant au plus 3 items dans le conséquent (6).

Les calculs sont démarrés dès que nous validons les paramètres : 136 règles sont extraites. Nous les visualisons en cliquant l'icône dans la grille d'exécution (ou via le menu contextuel RESULTATS / VISUALISATION DES REGLES D'ASSOCIATION).

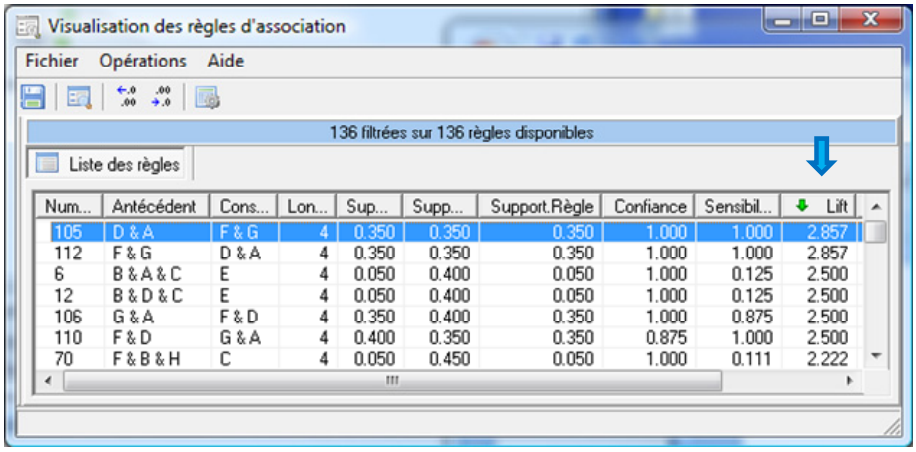


Nous observons la liste des règles. La n°1 par exemple revient à dire

Si le client a acheté les produits (A, C et E) alors il achètera le produit (D)

Chaque règle est accompagnée d'une série d'indicateurs censés mesurer leur pertinence. Nous en avons présenté quelques uns dans un de nos tutoriels décrivant la méthode A PRIORI MR implémentée dans Tanagra (http://tutoriels-data-mining.blogspot.com/2009/02/mesures-dinteret-des-regles-dans-priori.html).

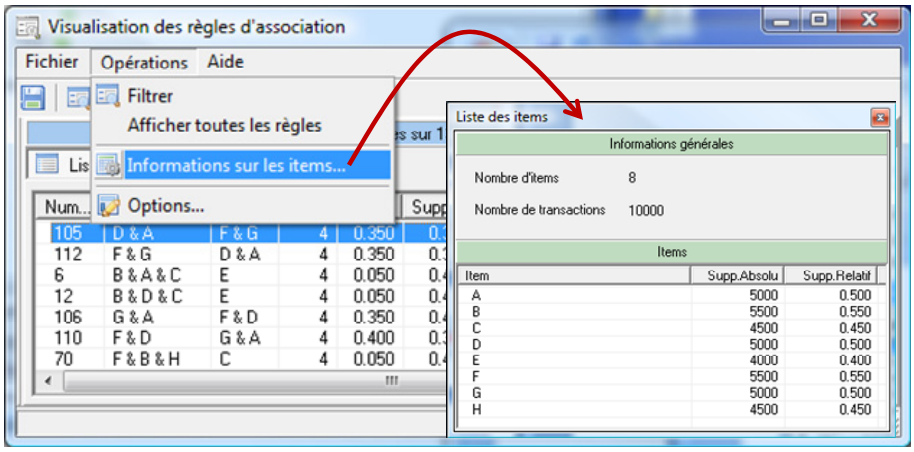
Chaque mesure a sa spécificité. Pour ma part, je pense que le simple LIFT nous permet déjà de très bien appréhender le surcroît d'information fourni par l'antécédent sur la connaissance du conséquent de la règle. En cliquant (plusieurs fois) sur l'en-tête de la colonne LIFT, nous constatons que les règles sont triées par ordre croissant ou décroissant. Nous pouvons ainsi mettre facilement en première ligne les règles (in)intéressantes au sens du LIFT.



Num...	Antécédent	Cons...	Lon...	Sup...	Supp...	Support.Règle	Confiance	Sensibil...	Lift
105	D & A	F & G	4	0.350	0.350	0.350	1.000	1.000	2.857
112	F & G	D & A	4	0.350	0.350	0.350	1.000	1.000	2.857
6	B & A & C	E	4	0.050	0.400	0.050	1.000	0.125	2.500
12	B & D & C	E	4	0.050	0.400	0.050	1.000	0.125	2.500
106	G & A	F & D	4	0.350	0.400	0.350	1.000	0.875	2.500
110	F & D	G & A	4	0.400	0.350	0.350	0.875	1.000	2.500
70	F & B & H	C	4	0.050	0.450	0.050	1.000	0.111	2.222

3.4 Exploration interactive des règles

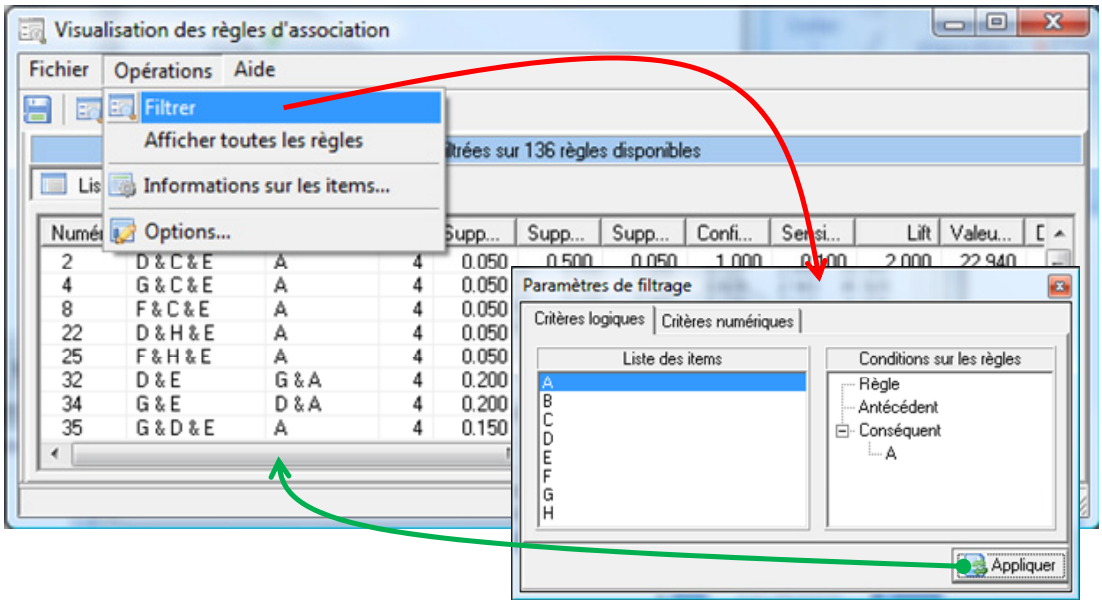
Des outils supplémentaires permettent de mieux cerner la teneur des résultats. En actionnant le menu OPERATIONS / INFORMATIONS SUR LES ITEMS, nous obtenons la liste des items (des produits) et leur nombre d'apparition dans l'ensemble des transactions (caddies).



Item	Supp.Absolu	Supp.Relatif
A	5000	0.500
B	5500	0.550
C	4500	0.450
D	5000	0.500
E	4000	0.400
F	5500	0.550
G	5000	0.500
H	4500	0.450

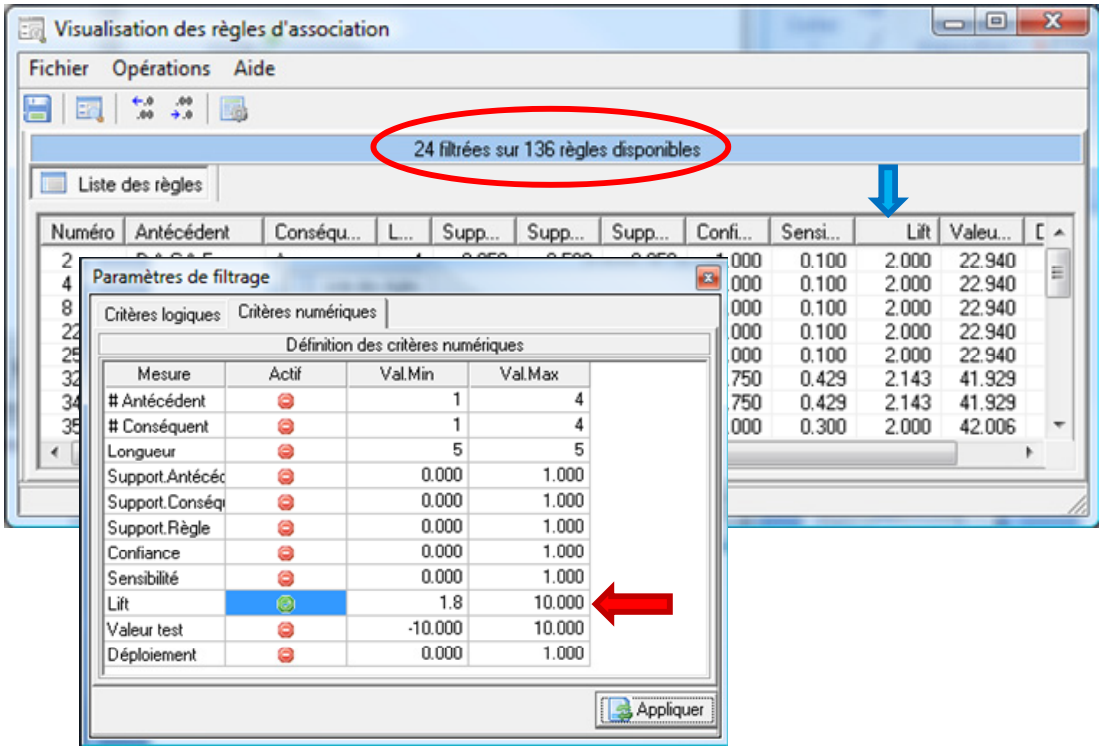
SPAD propose des outils supplémentaires pour filtrer a posteriori les règles. En actionnant le menu OPERATIONS / FILTRER, une boîte de pilotage apparaît. Nous pouvons filtrer à partir des items (en spécifiant les produits que l'on veut voir présents dans les règles) et/ou à partir de critères numériques.

Mettons que nous souhaitons visualiser les règles comportant le produit « A » dans leur conséquent. Nous glissons « A » de la liste des items vers le « Conséquent ». Nous cliquons alors sur le bouton « Appliquer ». 30 règles sont mises en avant.



Mettons que nous voulons filtrer ce sous-ensemble en retenant uniquement les règles ayant un LIFT supérieur à 1.8 (et inférieur à 10, mais toutes les règles de toute manière ont un LIFT inférieur à 10).

Dans l'onglet « Critères Numériques » de la boîte de paramétrage, nous fixons les bornes adéquates, et nous activons le filtrage sur le LIFT.



Nous cliquons sur le bouton « Appliquer ». Nous obtenons 24 règles.

En procédant ainsi, nous pouvons mettre en évidence le sous-ensemble de règles le plus approprié compte tenu du problème que l'on cherche à résoudre.

4 Analyse avec SAS EM 4.3

Je ne suis pas un spécialiste de SAS, encore moins de « Enterprise Miner » (le module spécialisé dans le Data Mining de SAS), loin de là. Le descriptif proposé ici est donc très simpliste, issu de tâtonnements divers, sans jamais avoir consulté la documentation. Les experts pousseront surement des cris d'orfraie. Qu'ils me pardonnent par avance. Pour ma part, je trouve que le résultat obtenu est plutôt pas mal. Mû par mes connaissances sur les règles d'association, et avec un peu de bon sens, il m'a finalement été facile de trouver les bonnes options dans le logiciel. Cela rejoint le message que je rabâche tous les jours à mes étudiants (les pauvres, je suis désolé, je me répète souvent) : c'est à l'utilisateur de spécifier au logiciel ce qu'il doit faire ; et non pas l'inverse, le logiciel ne doit jamais guider notre analyse en fonction de ce qu'il peut faire.

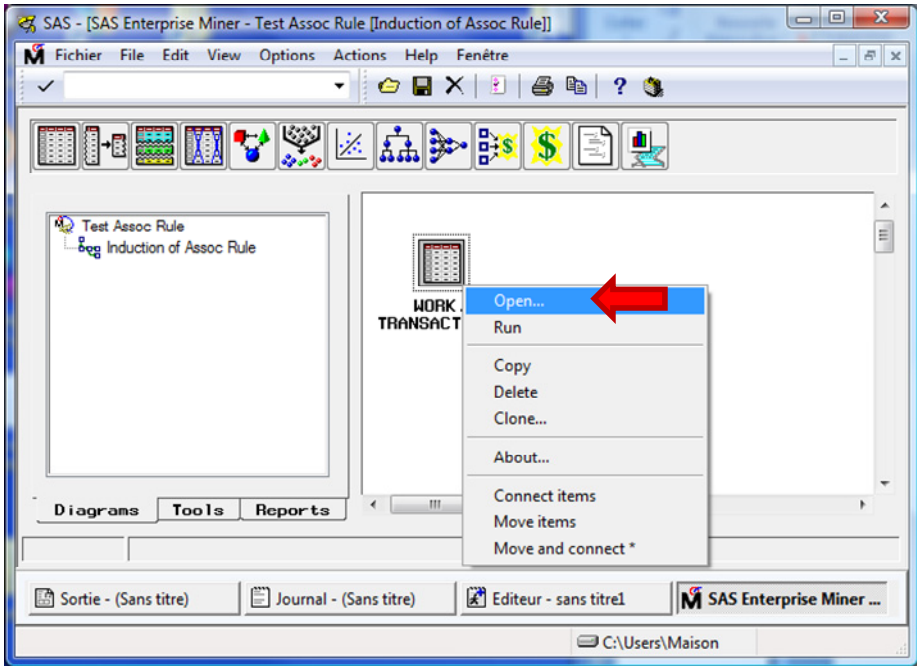
4.1 Importation des données

Première étape, nous importons le fichier « transactions.txt » directement dans la bibliothèque WORK (Note : *il est plutôt conseillé de créer sa propre bibliothèque de données si on veut pouvoir revenir sur son projet plus tard, nous allons au plus simple dans ce didacticiel*). Pour ce faire, après avoir démarré SAS, nous actionnons le menu FICHIER / IMPORTER LES DONNEES. Nous sélectionnons le format « Tab Delimited File (*.txt) ». Nous lui attribuons le nom TRANSACTIONS dans la bibliothèque WORK.

4.2 SAS Enterprise Miner

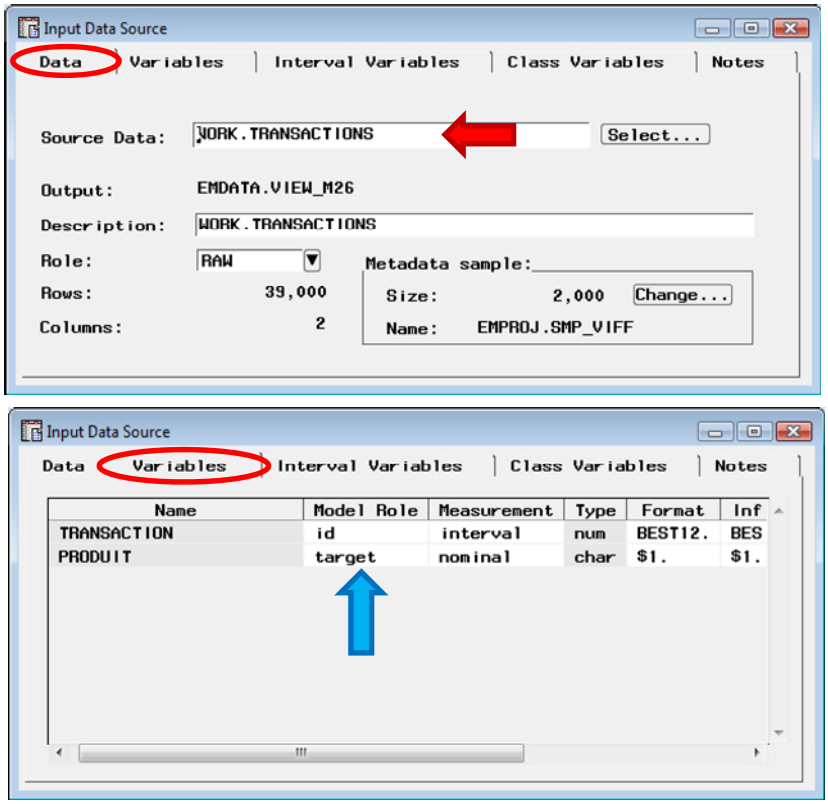
Nous cliquons sur le menu SOLUTIONS / ANALYSE / ENTERPRISE MINER pour lancer le module Data Mining de SAS. Habituellement un projet a déjà été créé, avec un diagramme par défaut. Nous les renommons « Induction of Assoc Rule ».

Le composant « Input Data Source » sert à se connecter aux données. Nous le plaçons dans l'espace de travail, puis nous le paramétrons en actionnant le menu contextuel OPEN.

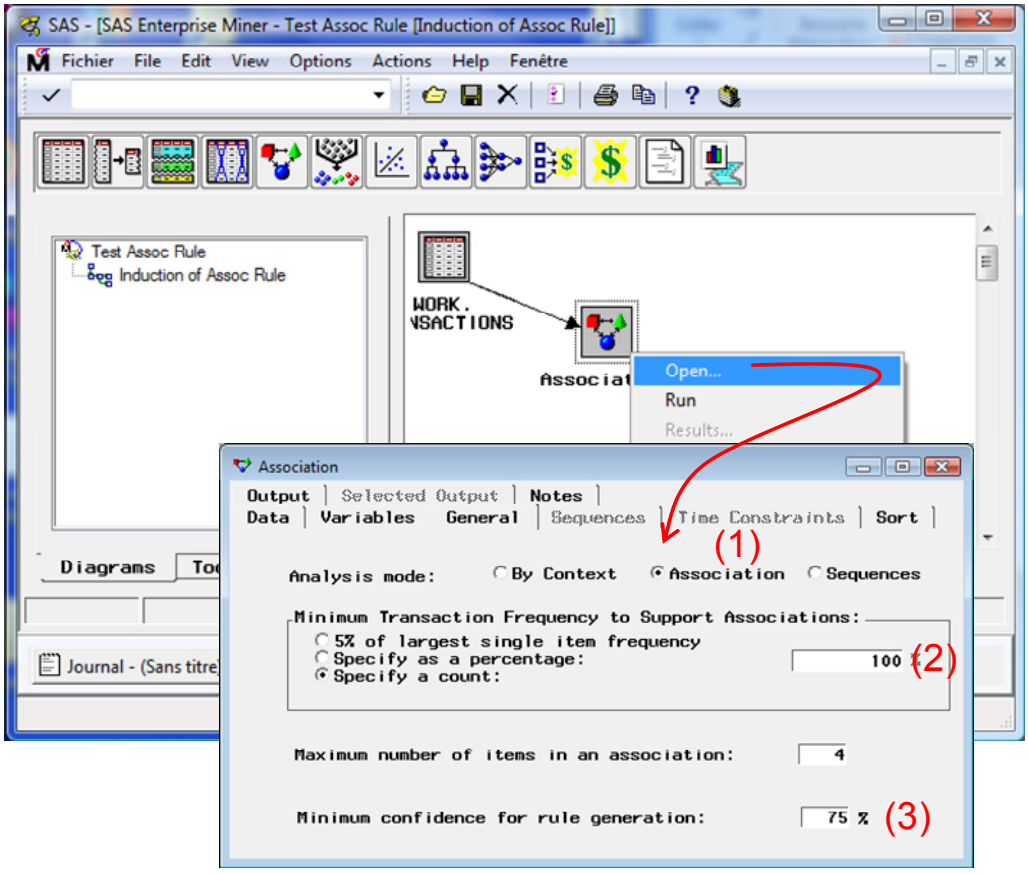


Dans « Source Data », nous sélectionnons la base TRANSACTIONS dans la librairie WORK.

Dans l'onglet VARIABLES, nous spécifions le rôle des colonnes via le menu contextuel SET MODEL ROLE : TRANSACTION est ID ; PRODUIT est TARGET.

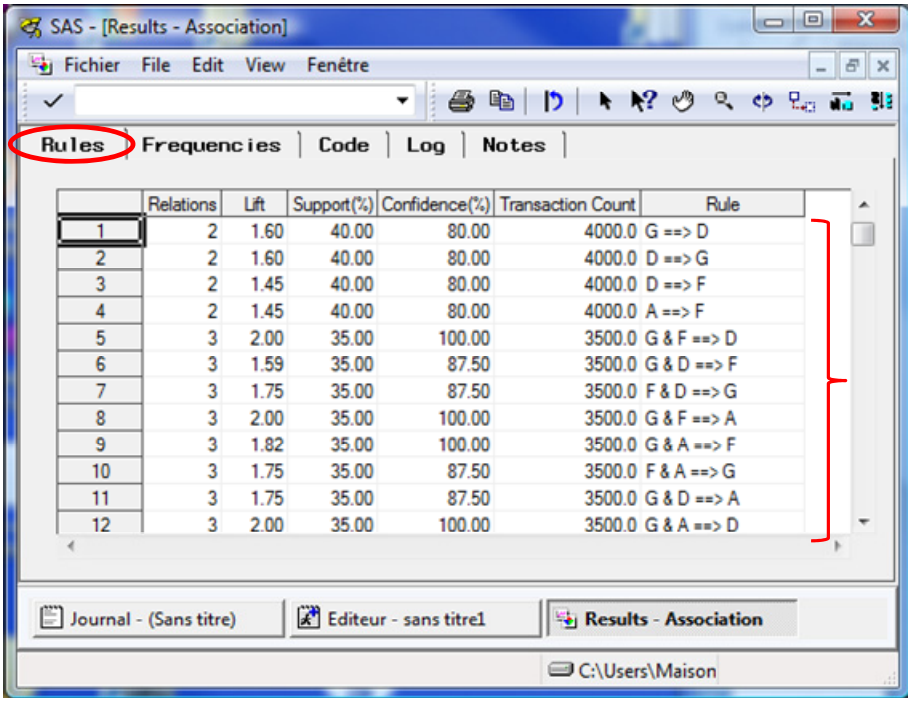


Reste à placer le composant ASSOCIATION. Nous le paramétrons également.



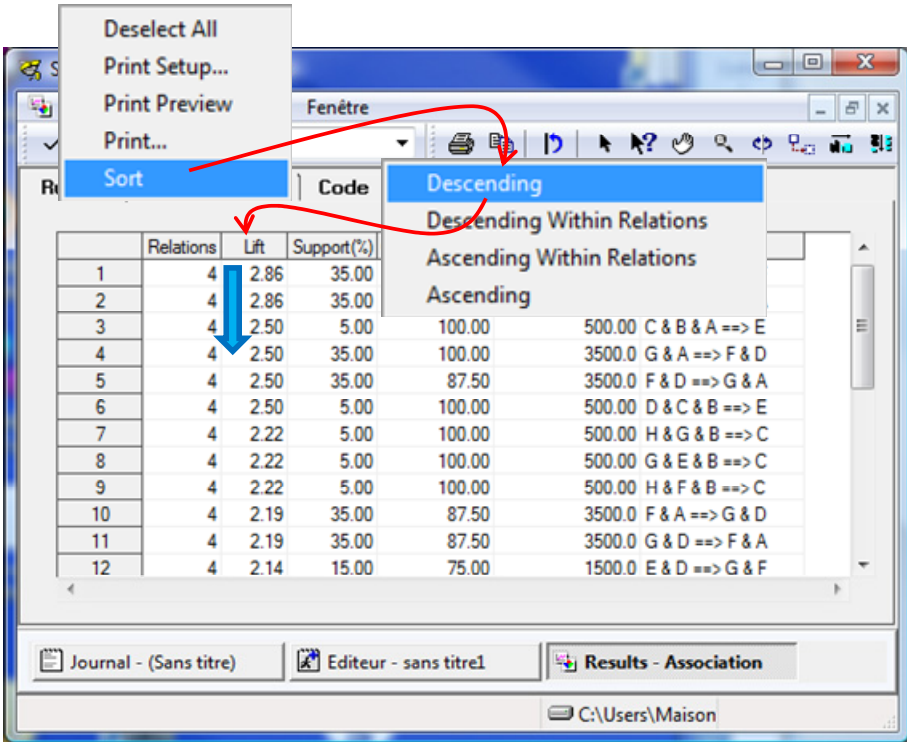
Nous souhaitons obtenir des règles d'association (1). Chaque item doit être présent dans au moins 100 transactions pour être introduit dans les règles (2). Le cardinal maximal des règles est 4. Et la confiance minimale est de 75% (3).

Nous validons et nous actionnons le menu contextuel RUN. Au bout d'un bref moment, SAS nous annonce que les calculs ont été menés à bien. Un affichage des règles est proposé.



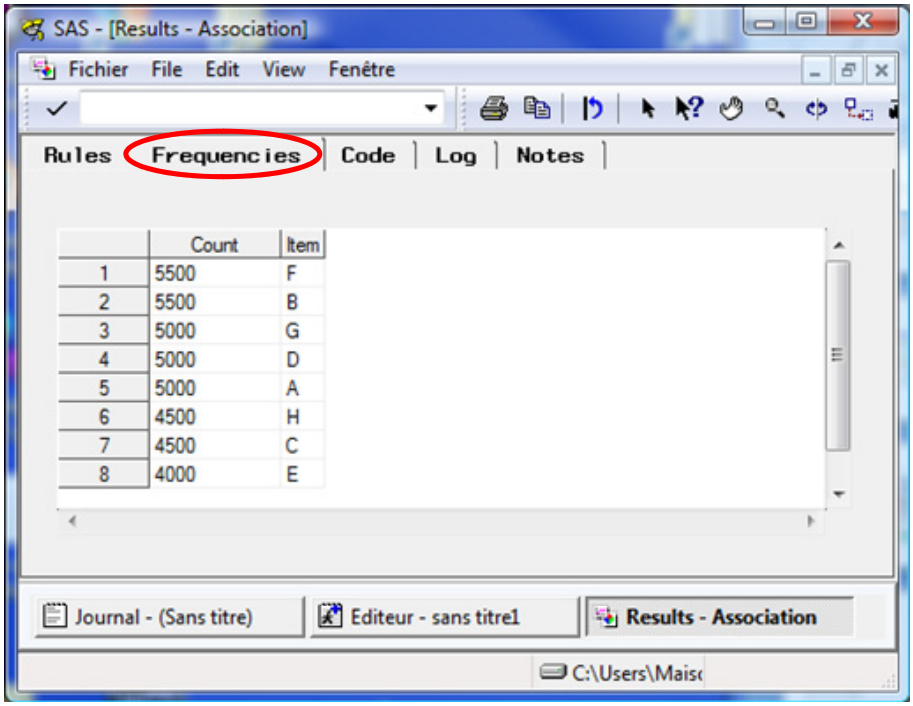
	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.60	40.00	80.00	4000.0	G ==> D
2	2	1.60	40.00	80.00	4000.0	D ==> G
3	2	1.45	40.00	80.00	4000.0	D ==> F
4	2	1.45	40.00	80.00	4000.0	A ==> F
5	3	2.00	35.00	100.00	3500.0	G & F ==> D
6	3	1.59	35.00	87.50	3500.0	G & D ==> F
7	3	1.75	35.00	87.50	3500.0	F & D ==> G
8	3	2.00	35.00	100.00	3500.0	G & F ==> A
9	3	1.82	35.00	100.00	3500.0	G & A ==> F
10	3	1.75	35.00	87.50	3500.0	F & A ==> G
11	3	1.75	35.00	87.50	3500.0	G & D ==> A
12	3	2.00	35.00	100.00	3500.0	G & A ==> D

136 règles ont été produites, à l'instar de SPAD. Il est possible de trier les règles de manière décroissante selon le LIFT avec le menu contextuel (SORT / DESCENDING).

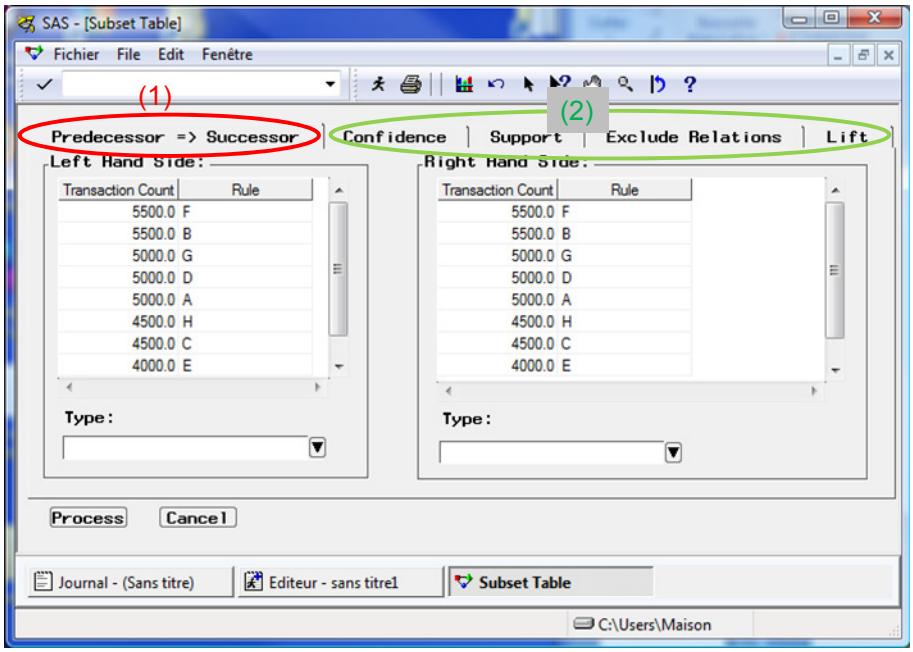


	Relations	Lift	Support(%)		
1	4	2.86	35.00		
2	4	2.86	35.00		
3	4	2.50	5.00	100.00	500.00 C & B & A ==> E
4	4	2.50	35.00	100.00	3500.0 G & A ==> F & D
5	4	2.50	35.00	87.50	3500.0 F & D ==> G & A
6	4	2.50	5.00	100.00	500.00 D & C & B ==> E
7	4	2.22	5.00	100.00	500.00 H & G & B ==> C
8	4	2.22	5.00	100.00	500.00 G & E & B ==> C
9	4	2.22	5.00	100.00	500.00 H & F & B ==> C
10	4	2.19	35.00	87.50	3500.0 F & A ==> G & D
11	4	2.19	35.00	87.50	3500.0 G & D ==> F & A
12	4	2.14	15.00	75.00	1500.0 E & D ==> G & F

Dans l'onglet FREQUENCIES, nous avons la fréquence de chaque item.



Tout comme dans SPAD, il est possible de restreindre le nombre de règles affichées via des critères logiques et/ou numériques. Il faut actionner le menu VIEW / SUBSET TABLE pour cela. On outil de pilotage permet de spécifier les conditions portant sur l'antécédent ou le conséquent des règles (1) ou sur les indicateurs numériques (2).



5 Analyse avec Tanagra et Knime

Très curieusement, les logiciels initialement développés dans les labos de recherche ne savent pas (ou ne veulent pas) manipuler directement les bases de transactions. Nous devons les transformer avant de pouvoir les importer et extraire les règles.

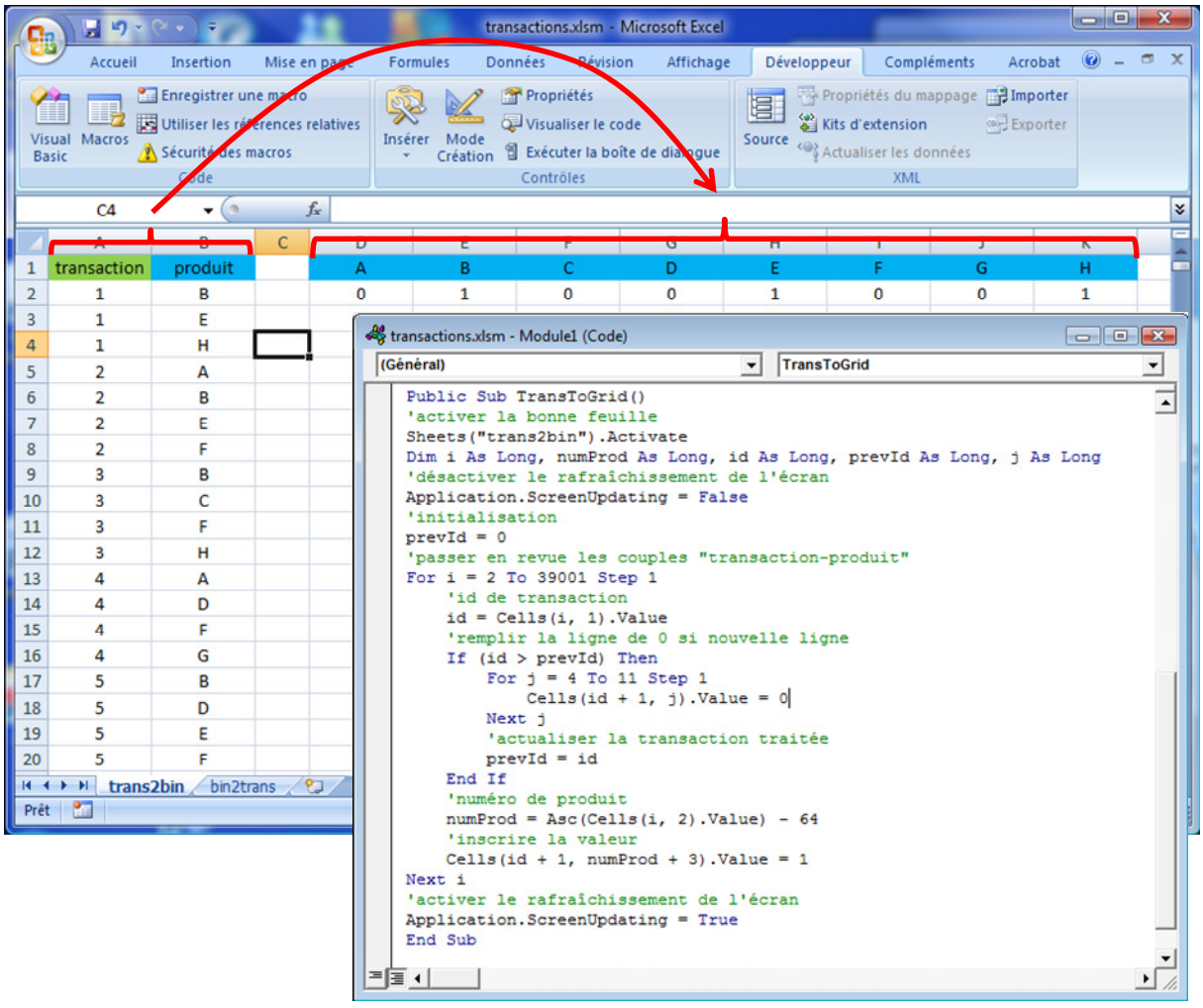
La transformation doit être en phase avec l'objectif de l'étude : nous souhaitons découvrir les cooccurrences des produits dans les caddies. Tanagra et Knime optent pour le même codage binaire 0/1 pour la manipulation de ce type de données. Nous avons autant de colonnes de produits dans la base, en ligne nous avons les transactions (les caddies). Nous en avions déjà parlé dans un précédent didacticiel, nous montrions comment Tanagra manipulait ce type de format (http://tutoriels-data-mining.blogspot.com/2008/04/priori-sur-les-bases-transactionnelles.html).

Dans ce qui suit, nous présentons tout d'abord une petite macro VBA (Visual Basic pour Applications) tournant sous Excel pour transformer les données. On se rend compte que le code source est très simple (il ne fonctionne que sur notre base, mais la transposition dans d'autres contextes ne devrait pas poser de problèmes). Par la suite, nous montrons le traitement de la base transformée (0/1) sous Tanagra et Knime.

J'avoue avoir un peu cherché quand même pour les autres logiciels que j'ai l'habitude de manipuler (RapidMiner, Weka, etc.) mais je n'ai pas trouvé (avis aux internautes curieux).

5.1 Transformation de la base avec une macro VBA

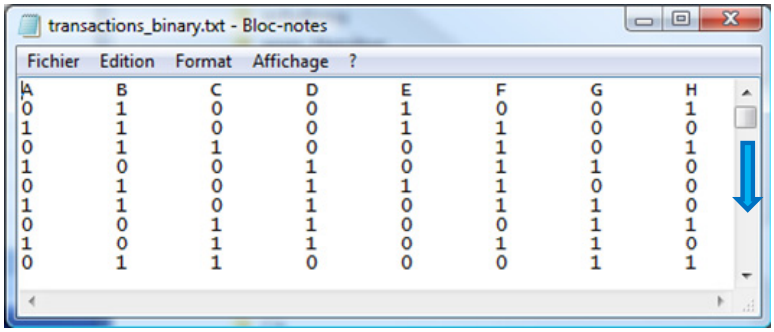
Nous utilisons la macro suivante pour transformer la base.



Elle est vraiment très simple. On s'appuie sur deux éléments pour produire le tableau binaire : le numéro de transaction indique la ligne que l'on doit remplir ; le code ASCII du nom du produit (qui

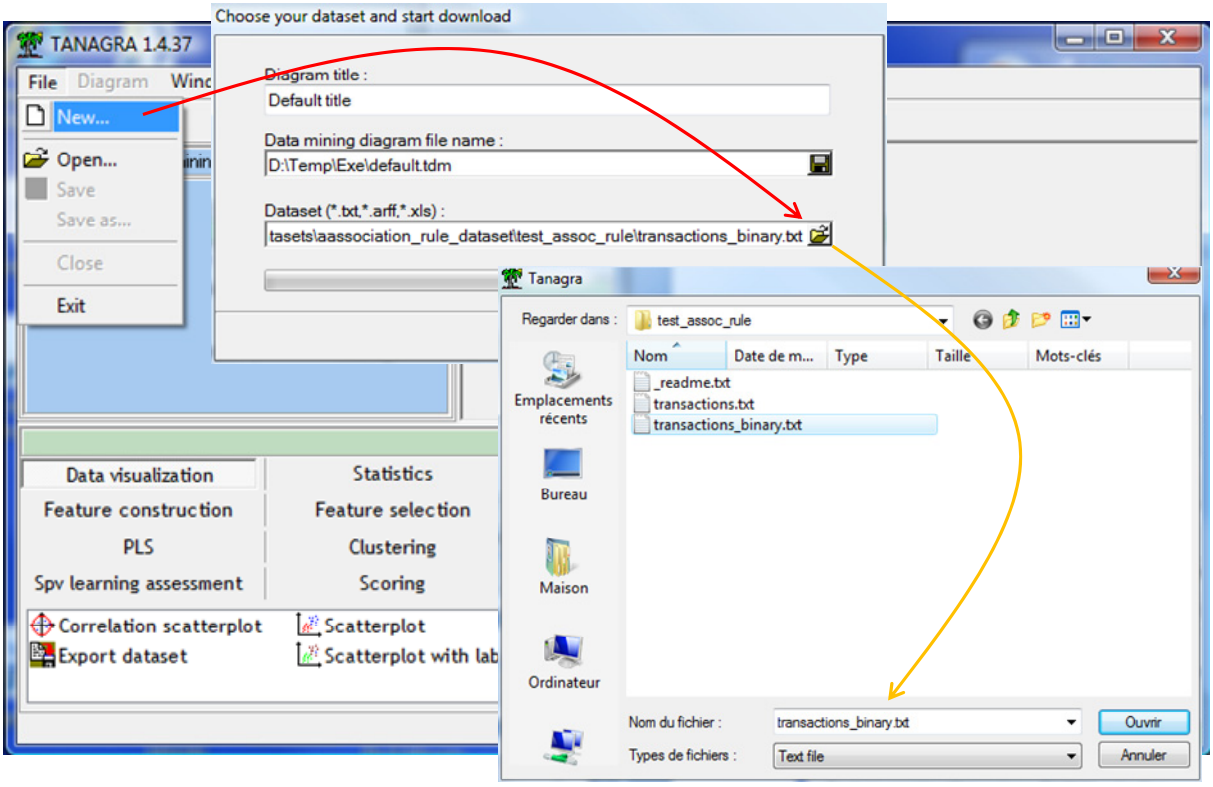
portent des noms qui nous arrangent bien pour le coup) pour détecter la colonne à remplir. Toutes les cases vides sont remplies par des zéros.

Nous exportons le tableau des données 0/1 dans le fichier texte « transactions_binary.txt ». Voici les premières lignes du fichier.



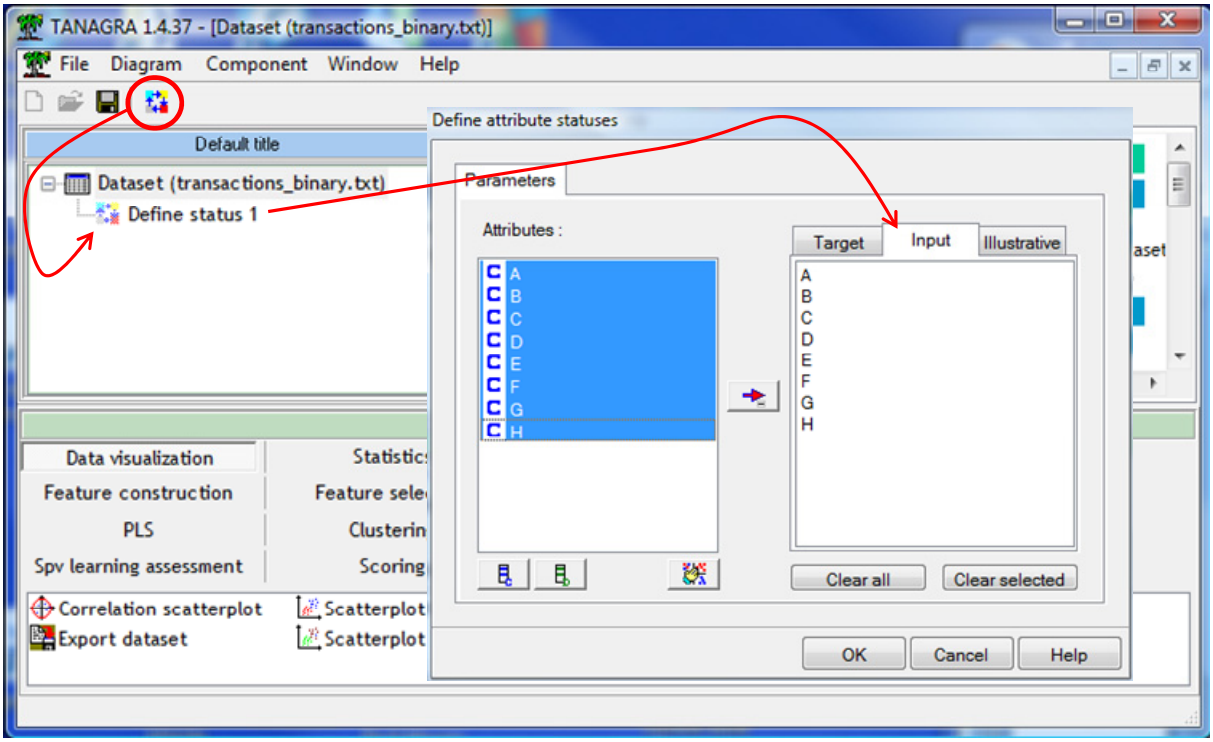
5.2 Analyse avec Tanagra

Nous démarrons Tanagra, et nous créons un nouveau diagramme en actionnant le menu FILE / NEW. Nous sélectionnons le fichier « transactions_binary.txt ».

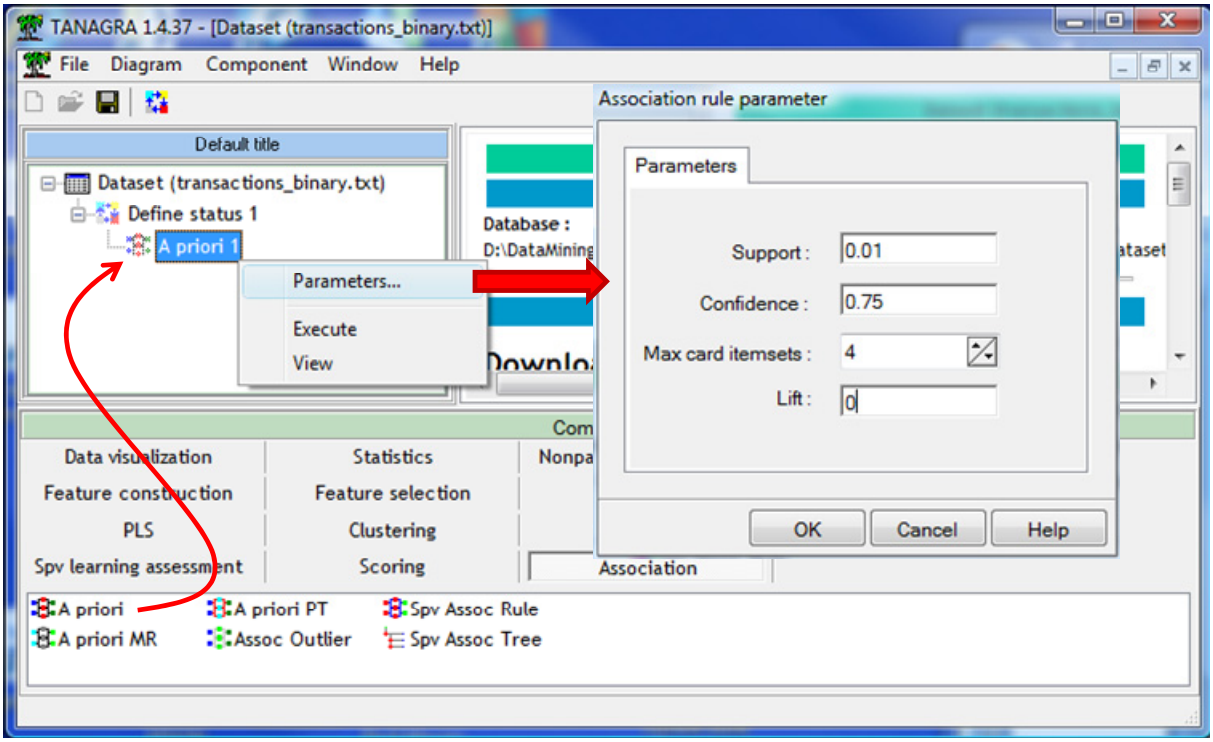


10000 observations (caddies) et 8 attributs (produits) ont été importés.

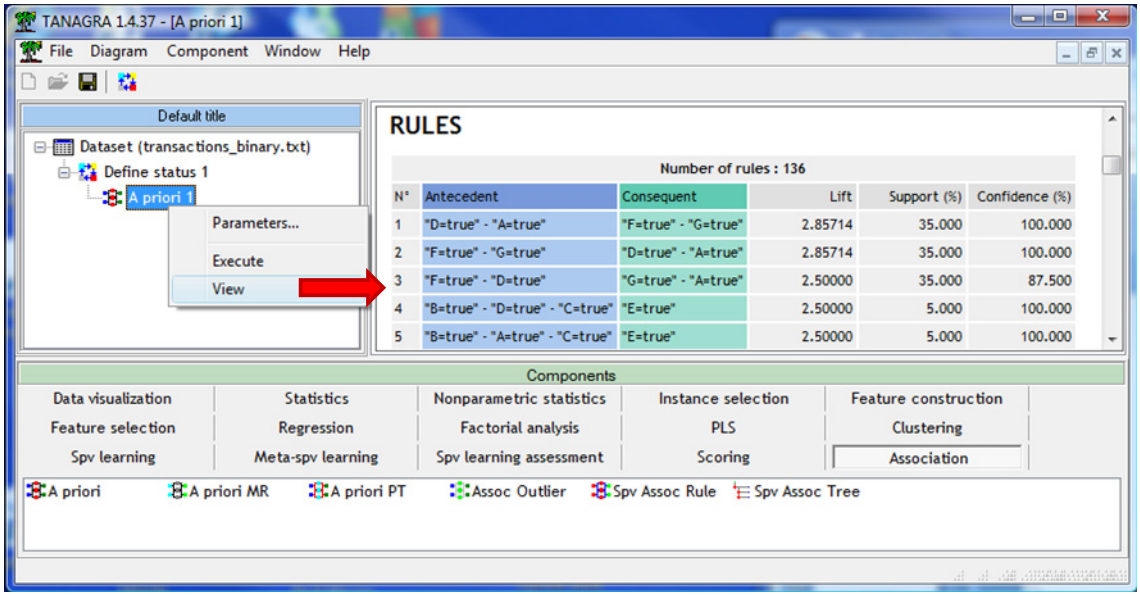
Nous devons spécifier le rôle des variables avant de lancer les traitements. Nous utilisons le composant DEFINE STATUS pour cela. Nous plaçons les 8 variables en INPUT.



Enfin, nous insérons le composant A PRIORI (onglet ASSOCIATION) dans le diagramme. Nous le paramétrons (menu contextuel PARAMETERS) de la manière suivante : (Support min = 0.01 ; Confiance min = 0.75 ; Cardinal max des règles = 4 ; Lift min = 0.0).



Il ne reste plus qu'à valider et à actionner le menu contextuel VIEW.

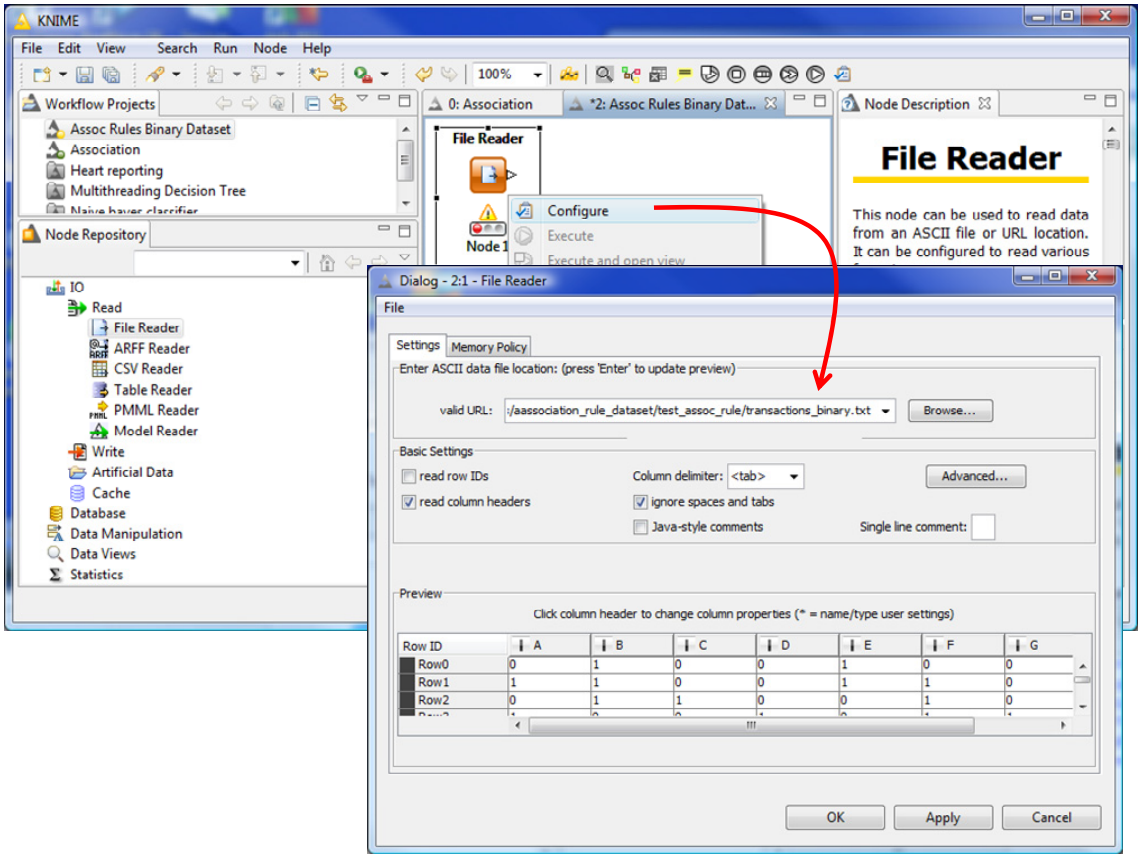


Nous obtenons 136 règles, comme SPAD et SAS. On notera que Tanagra a compris la nature du problème lorsqu'on lui présente des données binaires : il ne génère que les règles « positives » c.-à-d. il s'intéresse uniquement à la coprésence des produits dans les caddies.

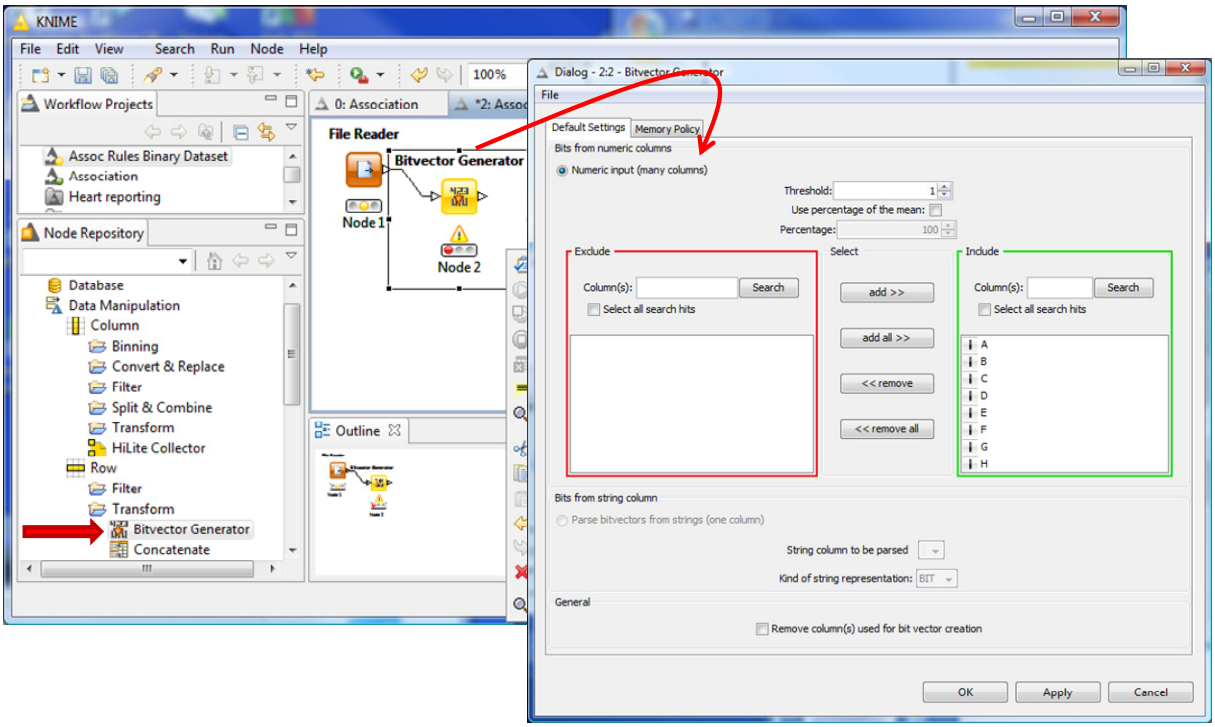
Enfin, Tanagra trie automatiquement les règles de manière décroissante sur le LIFT. Les manipulations interactives ne sont pas possibles en revanche.

5.3 Analyse avec Knime

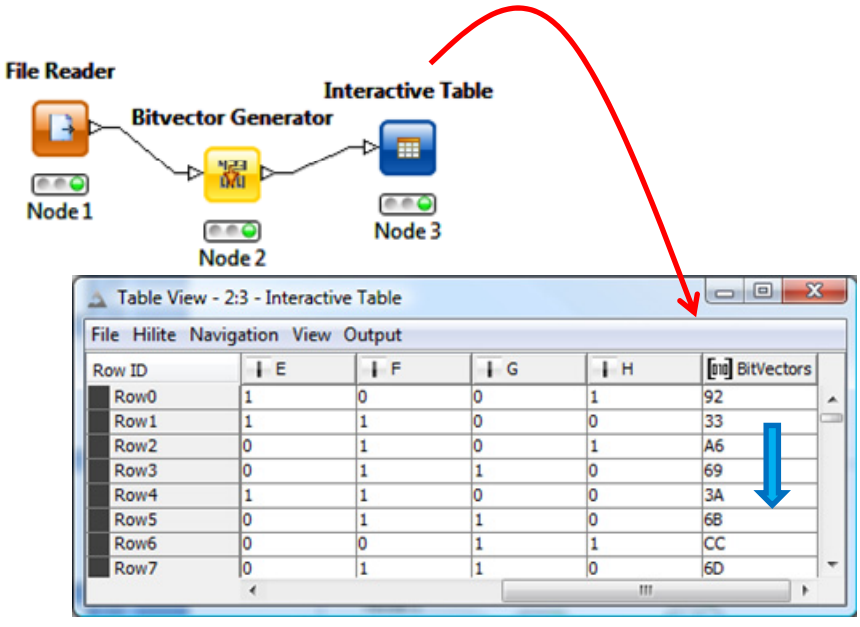
Après avoir démarré Knime, nous créons un nouveau diagramme. Nous importons les données à l'aide du composant FILE READER. Nous sélectionnons le fichier « transactions_binary.txt ».



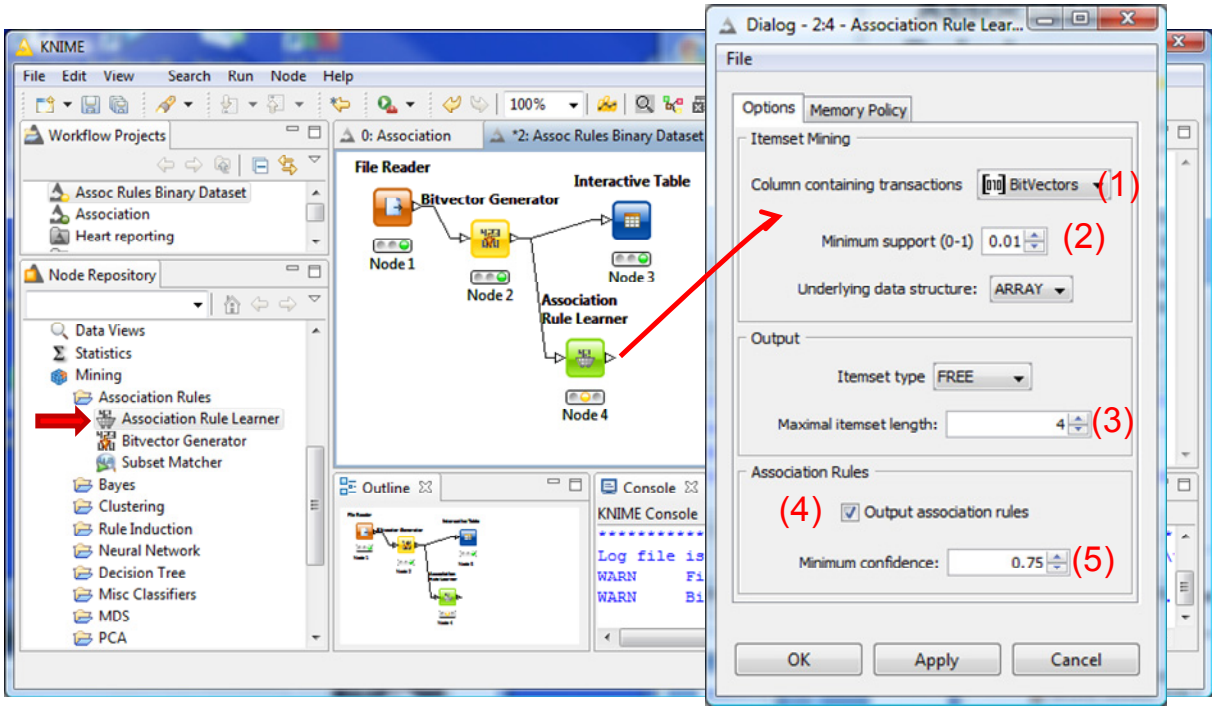
Une autre transformation est nécessaire pour que Knime puisse générer les règles d'association. Nous utilisons le composant BITVECTOR GENERATOR pour cela. Son rôle est de transformer chaque transaction en vecteur de bits propice à la recherche des itemsets (les sous-ensembles de produits). Nous le paramétrons comme suit (menu CONFIGURE)



Nous utilisons le composant INTERACTIVE TABLE pour visualiser la nouvelle variable générée. BITVECTORS représente un vecteur de bits codée en hexadécimal.

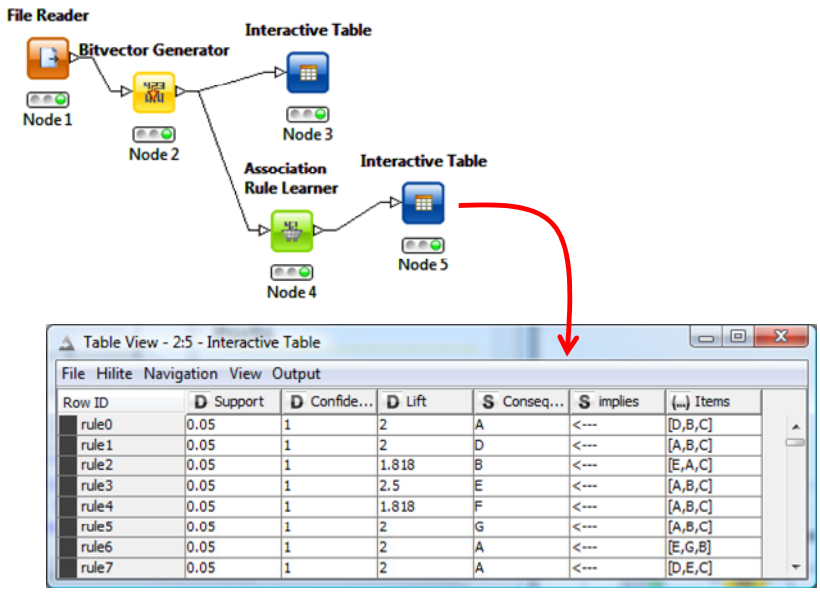


Nous pouvons initier l'induction des règles en introduisant le composant ASSOCIATION RULE LEARNER. Nous accédons à la boîte de paramétrage via le menu contextuel CONFIGURE.



Les transactions sont décrites à l'aide de la colonne BITVECTORS (1). Le support minimal des règles générées est 0.01 (2). Le nombre maximal d'items dans une règle est 4 (3). Enfin, nous souhaitons produire des règles (4) avec une confiance supérieure ou égale à 0.75 (5).

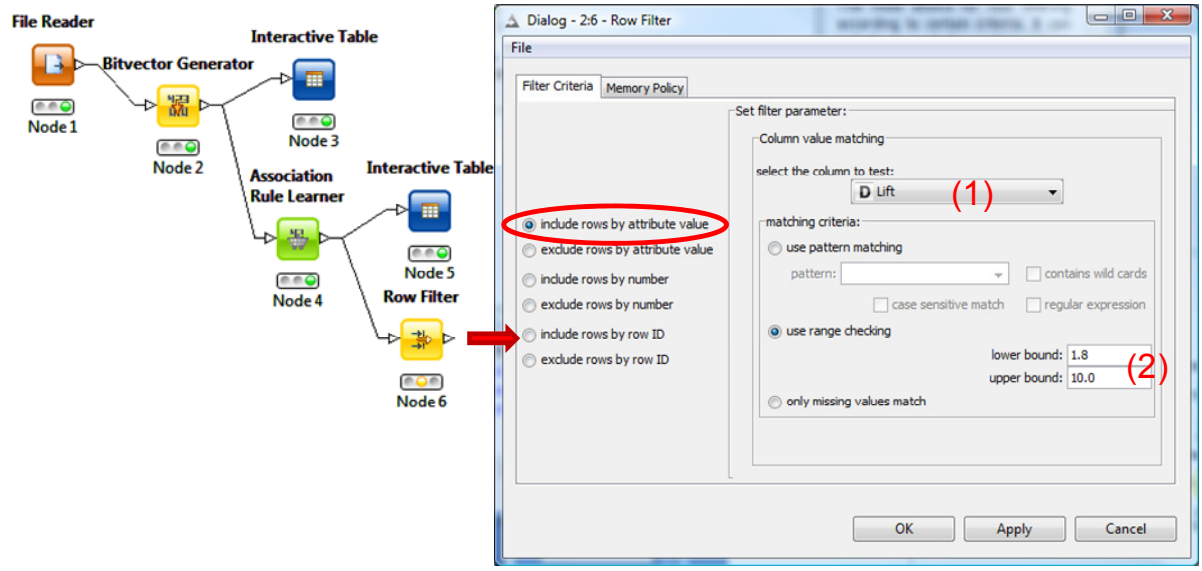
Nous cliquons sur EXECUTE pour lancer les calculs. Les règles peuvent être visualisées à l'aide du composant INTERACTIVE TABLE.



Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.05	1	2	A	<---	[D,B,C]
rule1	0.05	1	2	D	<---	[A,B,C]
rule2	0.05	1	1.818	B	<---	[E,A,C]
rule3	0.05	1	2.5	E	<---	[A,B,C]
rule4	0.05	1	1.818	F	<---	[A,B,C]
rule5	0.05	1	2	G	<---	[A,B,C]
rule6	0.05	1	2	A	<---	[E,G,B]
rule7	0.05	1	2	A	<---	[D,E,C]

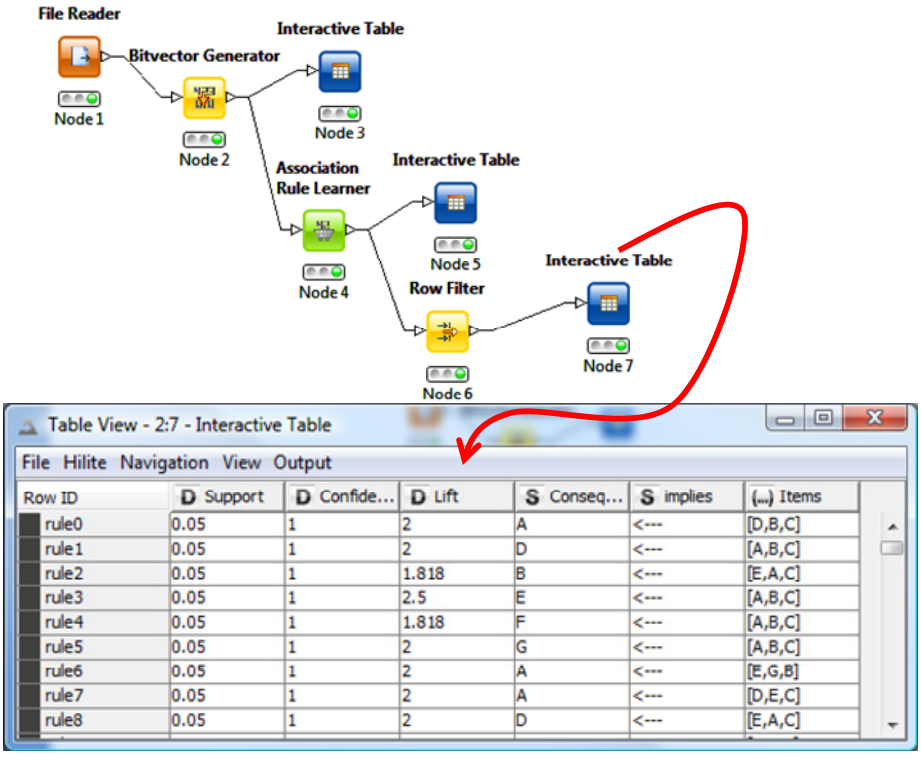
Knime ne produit que les règles à conséquents simples c.-à-d. composés d'un seul item. Il a produit 94 règles d'association.

Nous pouvons filtrer les règles avec le composant ROW FILTER. Nous souhaitons conserver celles qui présentent un LIFT supérieur ou égal à 1.85. Nous paramétrons le composant de la manière suivante.



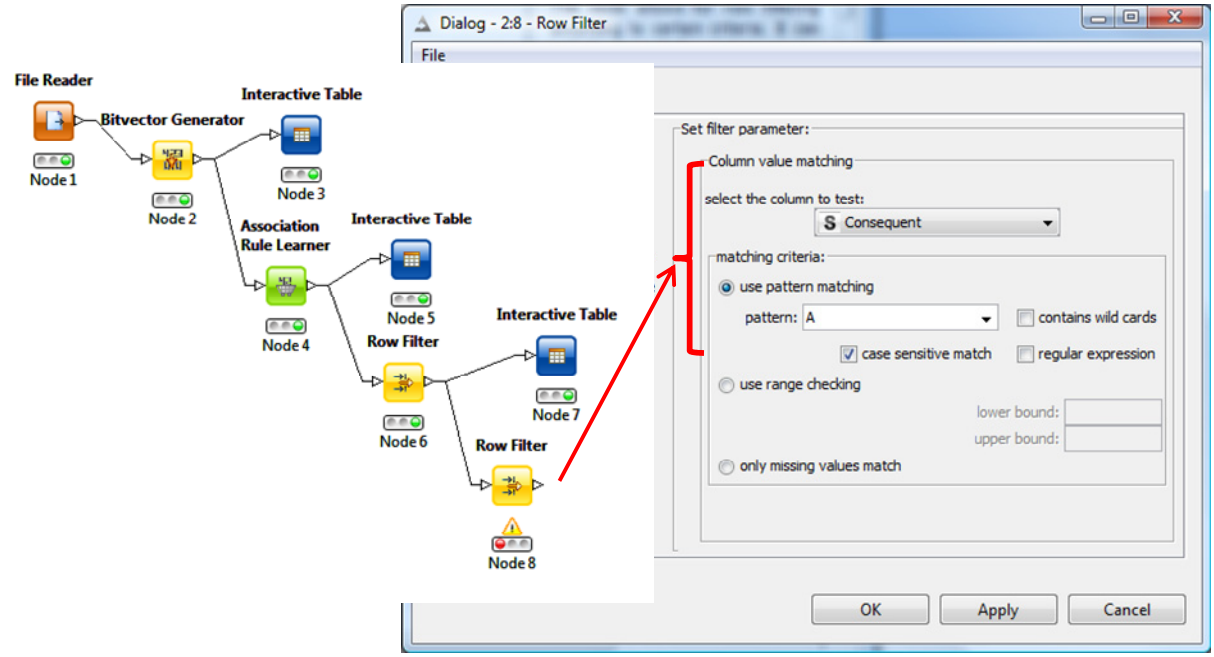
La sélection est basée sur le LIFT (1). La plage de valeurs acceptée est 1.8 à 10.0 (2).

Nous utilisons un autre INTERACTIVE TABLE pour visualiser les 90 règles correspondant à cette restriction.

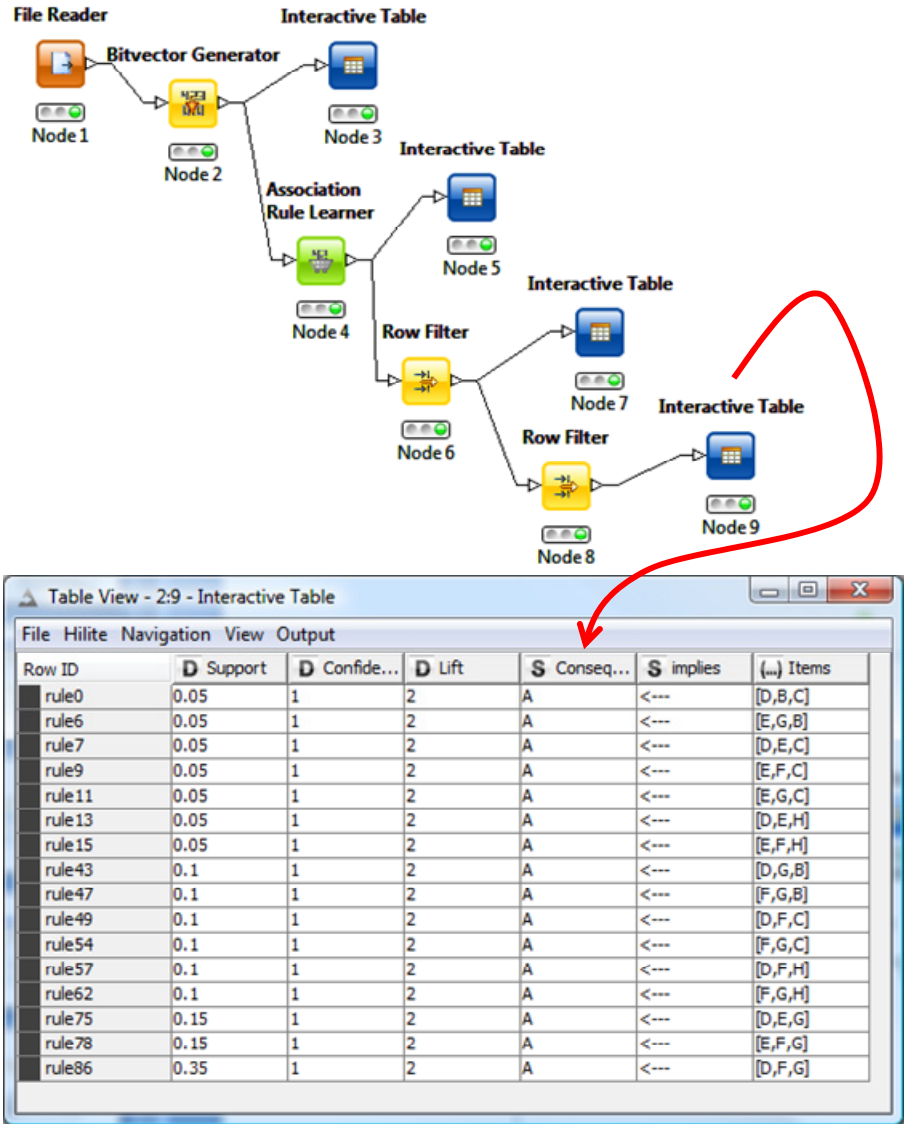


Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.05	1	2	A	<---	[D,B,C]
rule1	0.05	1	2	D	<---	[A,B,C]
rule2	0.05	1	1.818	B	<---	[E,A,C]
rule3	0.05	1	2.5	E	<---	[A,B,C]
rule4	0.05	1	1.818	F	<---	[A,B,C]
rule5	0.05	1	2	G	<---	[A,B,C]
rule6	0.05	1	2	A	<---	[E,G,B]
rule7	0.05	1	2	A	<---	[D,E,C]
rule8	0.05	1	2	D	<---	[E,A,C]

Nous pouvons affiner les restrictions sur les règles. Si nous souhaitons ne visualiser que les règles ayant un LIFT ≥ 1.8 et dont le conséquent correspond au produit A. Nous ajoutons un second ROW FILTER, le paramétrage est défini sur le conséquent cette fois-ci.



Via un autre INTERACTIVE TABLE, nous constatons qu'il ne reste plus que 16 règles.

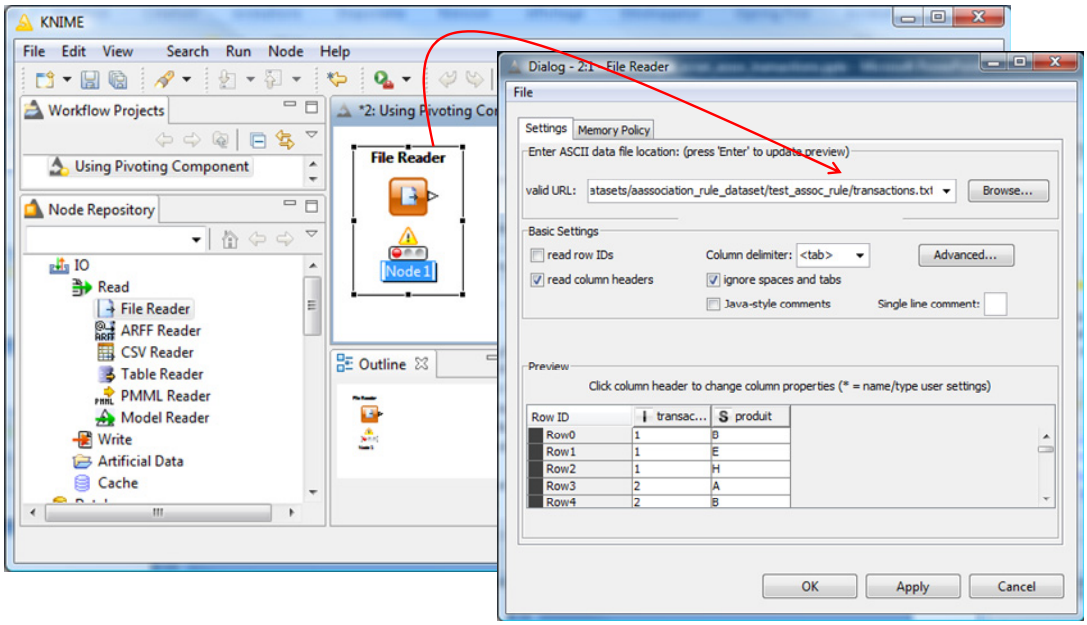


Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.05	1	2	A	<---	[D,B,C]
rule6	0.05	1	2	A	<---	[E,G,B]
rule7	0.05	1	2	A	<---	[D,E,C]
rule9	0.05	1	2	A	<---	[E,F,C]
rule11	0.05	1	2	A	<---	[E,G,C]
rule13	0.05	1	2	A	<---	[D,E,H]
rule15	0.05	1	2	A	<---	[E,F,H]
rule43	0.1	1	2	A	<---	[D,G,B]
rule47	0.1	1	2	A	<---	[F,G,B]
rule49	0.1	1	2	A	<---	[D,F,C]
rule54	0.1	1	2	A	<---	[F,G,C]
rule57	0.1	1	2	A	<---	[D,F,H]
rule62	0.1	1	2	A	<---	[F,G,H]
rule75	0.15	1	2	A	<---	[D,E,G]
rule78	0.15	1	2	A	<---	[E,F,G]
rule86	0.35	1	2	A	<---	[D,F,G]

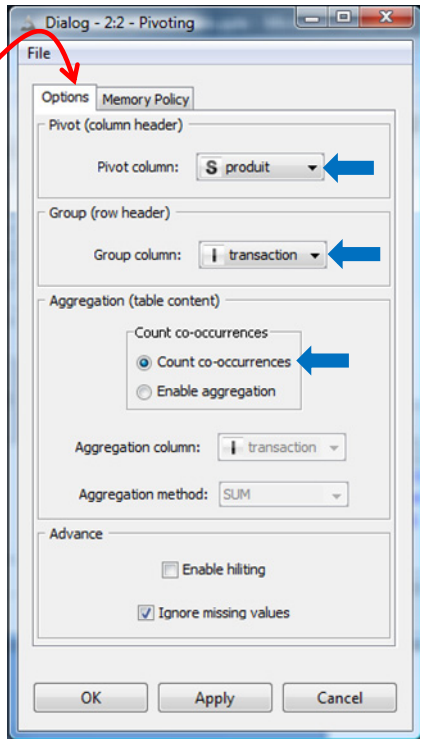
5.4 Analyse avec Knime (bis)

A la suite de la publication de ce didacticiel, Loïc LUCEL m'a signalé qu'il était possible de générer à l'intérieur même de Knime le tableau des données 0/1. On peut donc éviter le recodage préalable via un programme externe (en VBA dans mon exemple), l'opération peut être intégrée directement dans le diagramme de traitements. Merci beaucoup pour ces indications Loïc.

Voyons ce qu'il en est. Nous créons un nouveau « Workflow Project ». Via un FILE READER, nous accédons aux données au format transactionnel « transactions.txt ». Nous notons lors du typage automatique que Knime reconnaît les valeurs de TRANSACTIONS comme des entiers [**I**nteger], celles de PRODUIT comme des chaînes de caractères [**S**tring].



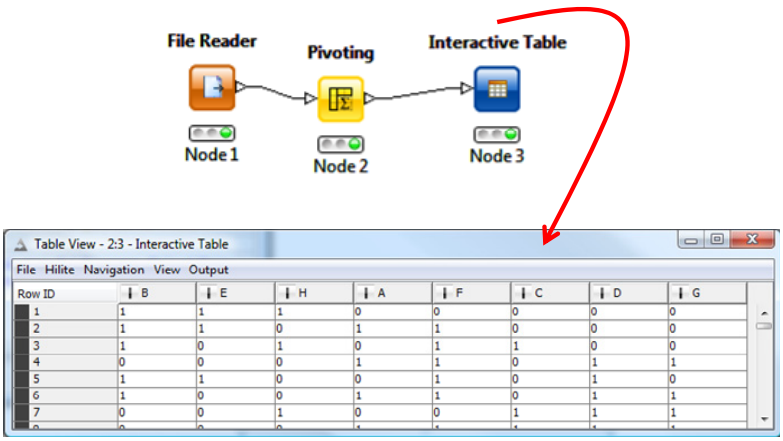
Ensuite, nous insérons le composant PIVOTING (de la branche DATA MANIPULATION / ROW / TRANSFORM) que nous paramétrons.



Très important, en PIVOT nous mettons la colonne PRODUIT, et en GROUP nous plaçons la colonne TRANSACTION. Ainsi, pour chaque transaction, nous comptons le nombre d'apparition de chaque produit. A priori, la valeur sera soit 0 (le produit est absent du caddie), soit 1 (il est présent).

Vérifions cela via un INTERACTIVE TABLE. En effet, nous retrouvons la matrice de données 0/1 avec en ligne le numéro de transaction, et en colonne les produits. Ils ne sont

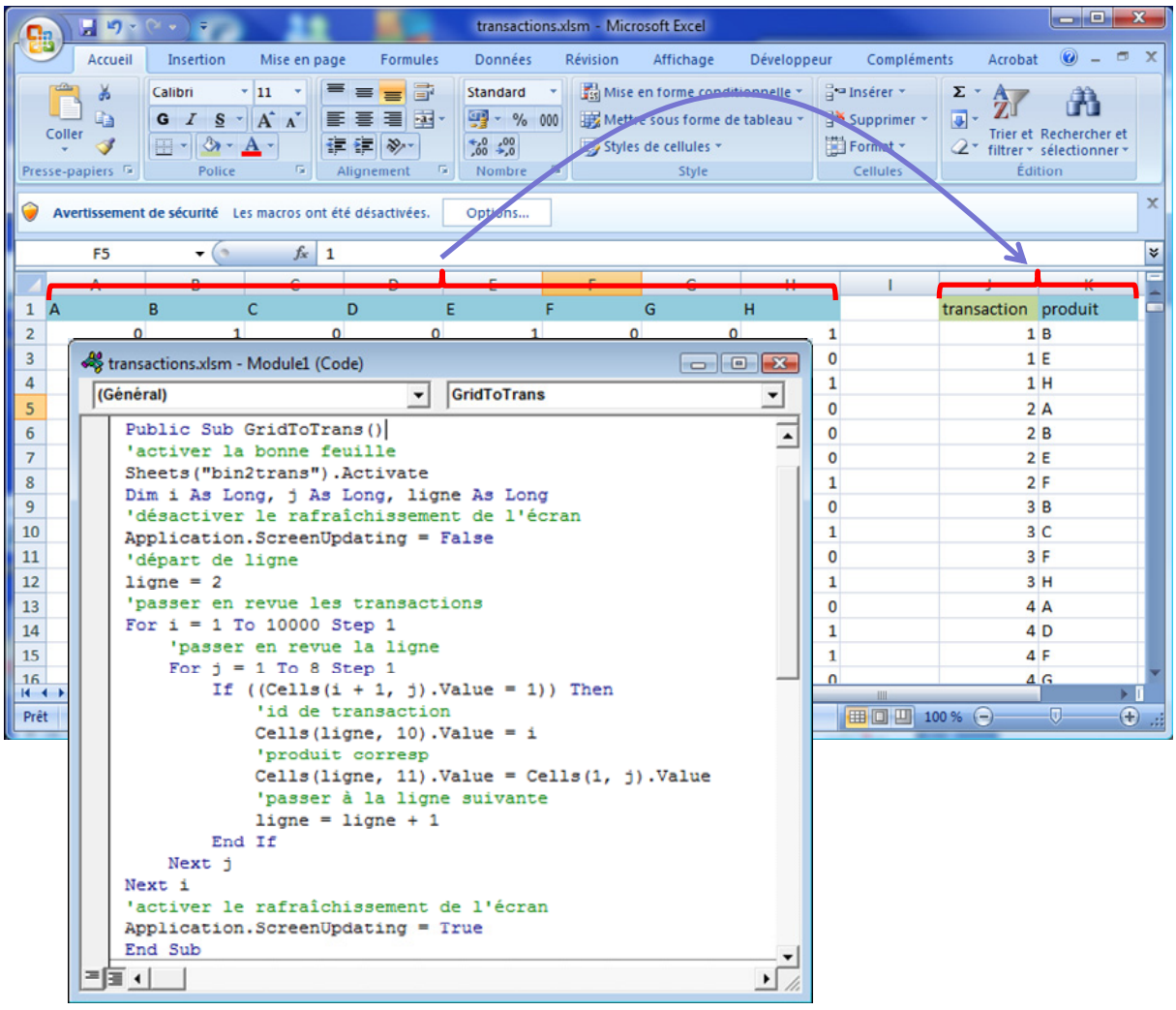
pas ordonnés mais ça n'a pas d'importance dans l'analyse que nous souhaitons mener.



Dès lors, comme dans la section précédente, il ne nous reste plus qu'à placer un BIT VECTOR GENERATOR à la suite du PIVOTING et procéder à la génération des règles d'association en suivant la même démarche. **L'énorme avantage de la procédure décrite dans cette section est qu'elle reste applicable si le fichier source est modifié c.-à-d. si l'on veut intégrer de nouvelles de transactions ou de nouveaux produits**.

5.5 Codage inverse : de la base binaire vers la base de transactions

A titre de curiosité, nous montrons dans cette sous-section la transformation inverse, partant de la base binaire vers la base transactionnelle. Comme nous pouvons le constater, le code VBA est très simple lui aussi. Et la généralisation à d'autres bases, avec plus de produits (d'items) et de transactions, ne pose aucun problème.



6 Conclusion

Pouvoir manipuler directement des bases transactionnelles est un atout pour l'extraction des règles d'association. La facilité des manipulations sous SPAD et SAS le montrent. Curieusement, cette fonctionnalité est absente de la majorité (ceux que j'étudie souvent en tous les cas) des logiciels universitaires « généralistes ». Parce que l'affaire est loin d'être évidente. Dans un fichier individus-variables, on ne sait pas trop comment appréhender la colonne des produits. Il ne s'agit pas d'une variable catégorielle normale, le nombre de modalités déjà peut être très élevé.

Pour contourner cet écueil, nous montrons les transformations adéquates pour que des logiciels tels que Tanagra et Knime puissent répondre au cahier des charges de l'analyse (***Mise à jour*** *: Knime peut s'occuper en interne du recodage des données transactionnelles en tableau 0/1 ; voir la section 5.4*). Nous constatons que le programme de transformation est très simple, on peut l'implémenter dans un tableur. Sous Excel 2007, la limitation serait de 1 048 575 couples « id caddie – produit » (en comptant la ligne d'en-tête) et 16381 produits (en ne commençant le tableau binaire qu'à la colonne D). On a une bonne marge quand même.