1 Topic

Mining Association Rules from Transactions File.

Association rule learning is a popular method for discovering interesting relations between variables in large databases. It was often used in market basket analysis domain. But in fact, it can be implemented in various areas where we want to discover the associations between variables. The association is described by a "IF THEN" rule. The IF part is called "antecedent" of the rule; the THEN part correspond to the "consequent" e.g. IF onions AND potatoes THAN burger (http://en.wikipedia.org/wiki/Association_rule_learning) i.e. if a customer buys onions and potatoes then he buys also burger.

It is possible to find co-occurrences in the standard attribute - value tables that are handled with the most of the data mining tools. In this context, the rows correspond to the baskets (transactions); the columns correspond to the list of all possible products (items); at the intersection of the row and the column, we have an indicator (true/false or 1/0) which indicates if the item belongs to the transaction. But this kind of representation is too naive. A few products are incorporated in each basket. Each row of the table contains a few 1 and many 0. The size of the data file is unnecessarily excessive. Therefore, another data representation, says "transactions file", is often used to minimize the data file size. In this tutorial, we treat a special case of the transactions file. The principle is based on the enumeration of the items included in each transaction. But in our case, we have only two values for each row of the data file: the transaction identifier, and the item identifier. Thus, each transaction can be listed on several rows of the data file.

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

For our example, the items B, E and H are incorporated into the transaction 1; the items A, B, E and F into the transaction 2; etc. This data representation is more space-saving. Only the items included in each transaction are listed.

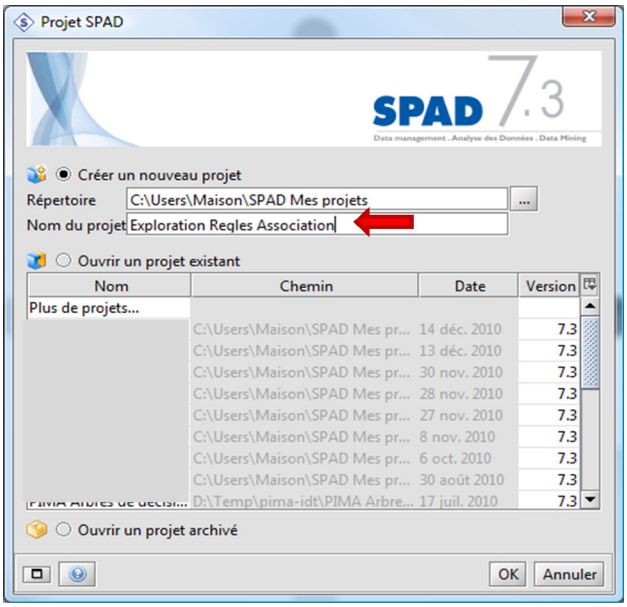
This data representation is quite natural considering the problem we want to treat. It also has the advantage of being more compact since only the items really present in each transaction are enumerated. However, it appears that many tools do not know manage directly this kind of data representation. We observe curiously a distinction between professional tools and the academic ones. The first ones can handle directly this kind of data file without special data preparation. This is the case of **SPAD 7.3** and **SAS Enterprise Miner 4.3** that we study in this tutorial. On the other hand, the academic tools need a data transformation, prior the importation of the dataset. We use a small program written in VBA (Visual Basic for Applications) under Excel to prepare the dataset. Thereafter, we perform the analysis with **Tanagra 1.4.37** and **Knime 2.2.2** (***Note**: a reader told me that we can transform the dataset with Knime without the utilization of external program. This is true. I will describe this approach in a separate section at the end of this tutorial*).

Attention, we must respect the original specifications i.e. focus only on rules indicating the simultaneous presence of items in transactions. We must not, consecutively to a bad "presence - absence" coding scheme, to generate rules outlining the simultaneous absence of some items. This may be interesting in some cases may be, but this is not the purpose of our analysis.

2 Dataset

The « transactions.txt » data file describes 10.000 baskets, 8 goods are referenced. Of course, it is an artificial dataset intended for the evaluation the association rule learning tools. We have already used this dataset (http://data-mining-tutorials.blogspot.com/2008/11/association-rule-learning-from.html). The originality here, at least with SPAD and SAS, is that we treat directly the data organized in the transactions file format.

3 Association Rule Mining with SPAD 7.3



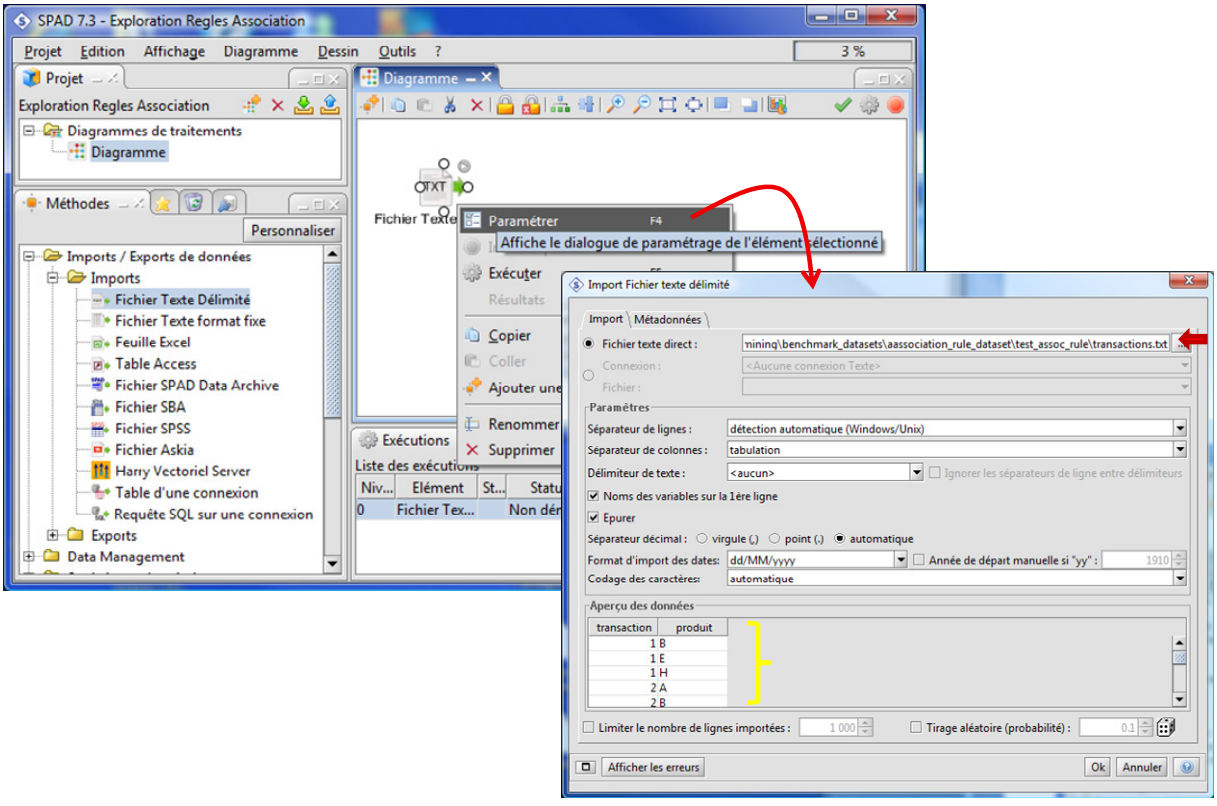
SPAD is a popular tool for the French data miners. Its modus operandi is consistent with the standards of the domain. A data mining process is represented as a stream. The nodes represent the operations performed on the data.

3.1 Creating a new diagram

When we launch SPAD, a dialog setting allows us to create a new diagram.

3.2 Importing the dataset

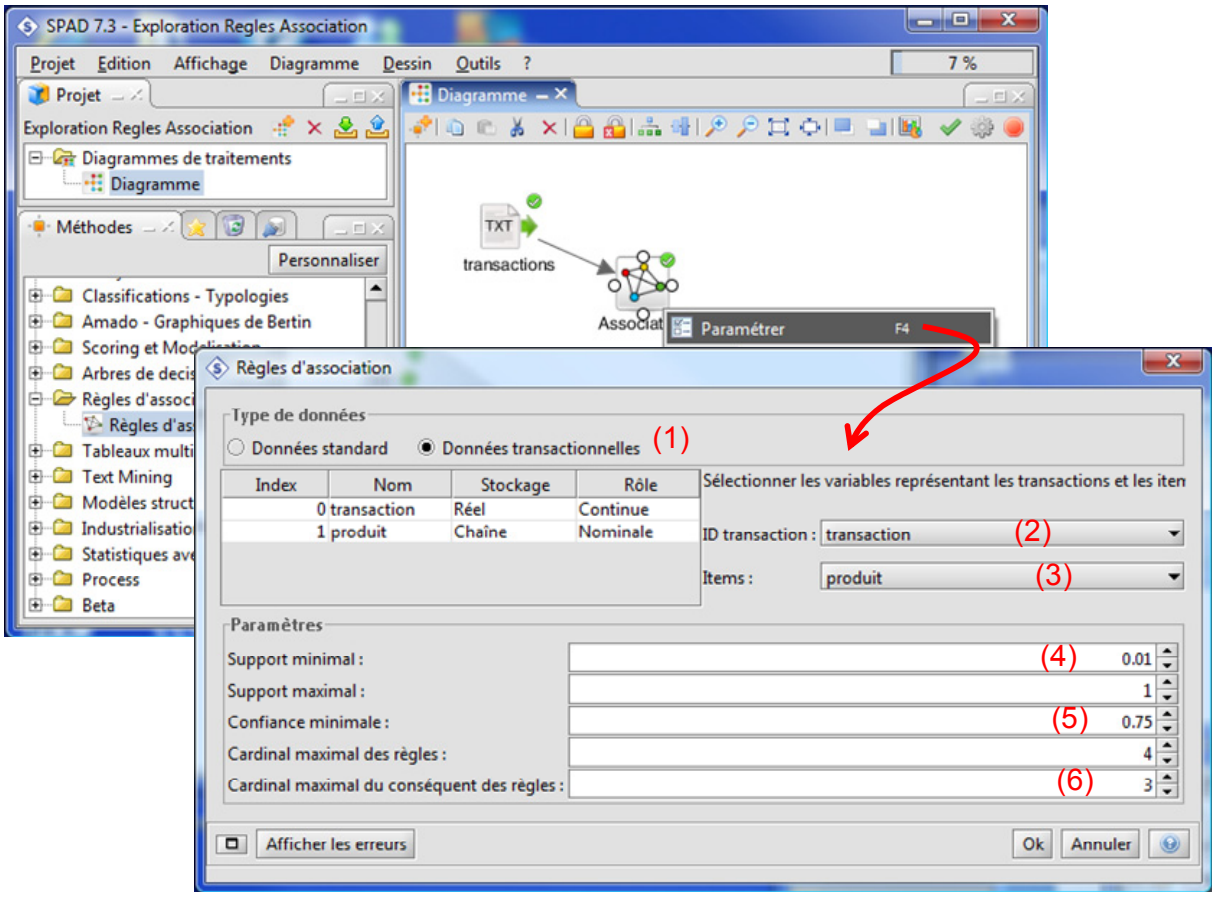
We add the "Fichier Texte Délimité" tool into the stream canvas. We set the appropriate settings (PARAMETERS contextual menu) to import the "transactions.txt" data file.



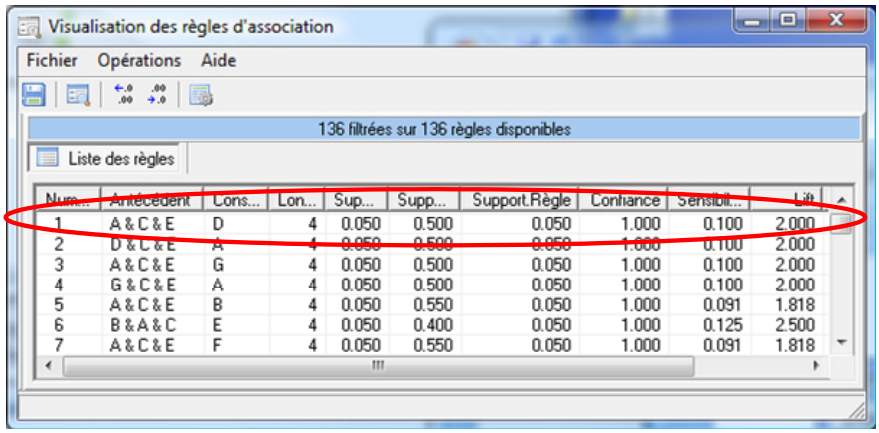
There are 39,000 rows and 2 columns as specified by SPAD.

3.3 Extracting the rules

The "Règles d'Association" tool enables to extract the rules. We link the data access component to it. We set the following parameters (PARAMETERS contextual menu):



We use the transactions file format (1). The identifier of the transaction is the first column "transaction" (2). The items are described in the second column "produit" (3). We set the minimum support to 0.01 i.e. a rule must cover at least (0.01 x 10000) 100 transactions (4). The minimum confidence is 0.75 (5). Last, the max cardinal of a rule (i.e. the max number of items authorized in a rule) is set to 4; we want rules up to 3 items in the consequent (6). The calculations are started as soon as we validate the parameters: 136 rules are extracted. We visualize them by clicking on the RESULTATS / VISUALISATION DES REGLES D'ASSOCIATION contextual menu.



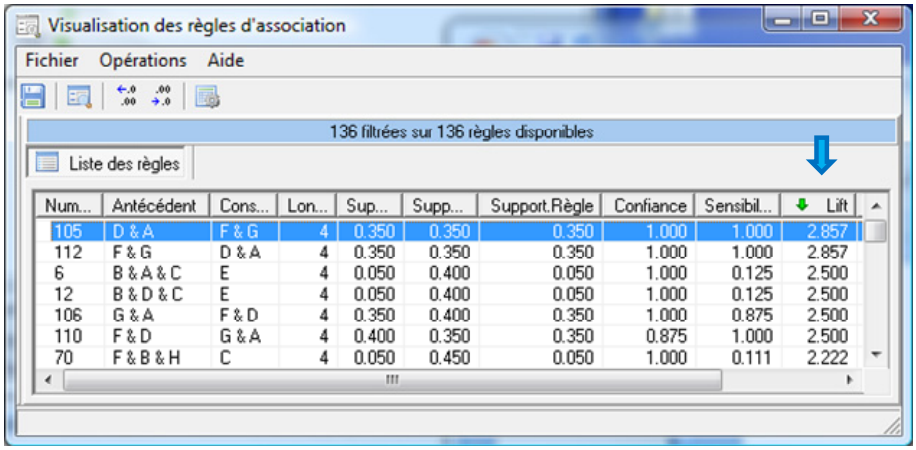
Num...	Antécédent	Cons...	Lon...	Sup...	Supp...	Support.Règle	Confiance	Sensibil...	Lift
1	A&C&E	D	4	0.050	0.500	0.050	1.000	0.100	2.000
2	D&C&E	A	4	0.050	0.500	0.050	1.000	0.100	2.000
3	A&C&E	G	4	0.050	0.500	0.050	1.000	0.100	2.000
4	G&C&E	A	4	0.050	0.500	0.050	1.000	0.100	2.000
5	A&C&E	B	4	0.050	0.550	0.050	1.000	0.091	1.818
6	B&A&C	E	4	0.050	0.400	0.050	1.000	0.125	2.500
7	A&C&E	F	4	0.050	0.550	0.050	1.000	0.091	1.818

For instance, the rule n°1 means

If the consumer buys (A, C and E) then it buys also (D)

The description of the rule is completed with some interestingness measures which outline the relevance of the rule. We have described some of them in a previous tutorial (http://data-mining-tutorials.blogspot.com/2009/02/interestingness-measures-for.html).

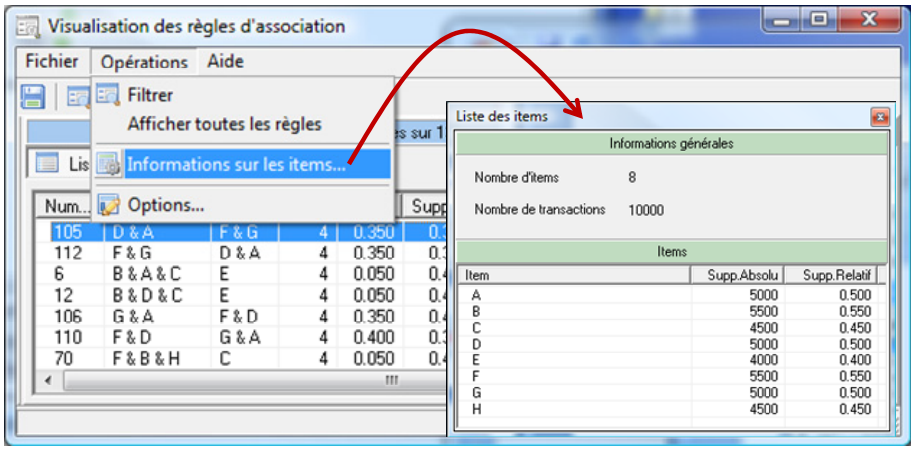
Each measure has its specificity. If we want to sort the rules according to the LIFT indicator (in a decreasing order, we click on the header of the corresponding column (if we click again, the rules are ordered in an increasing order).



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 Interactive exploration of the rules

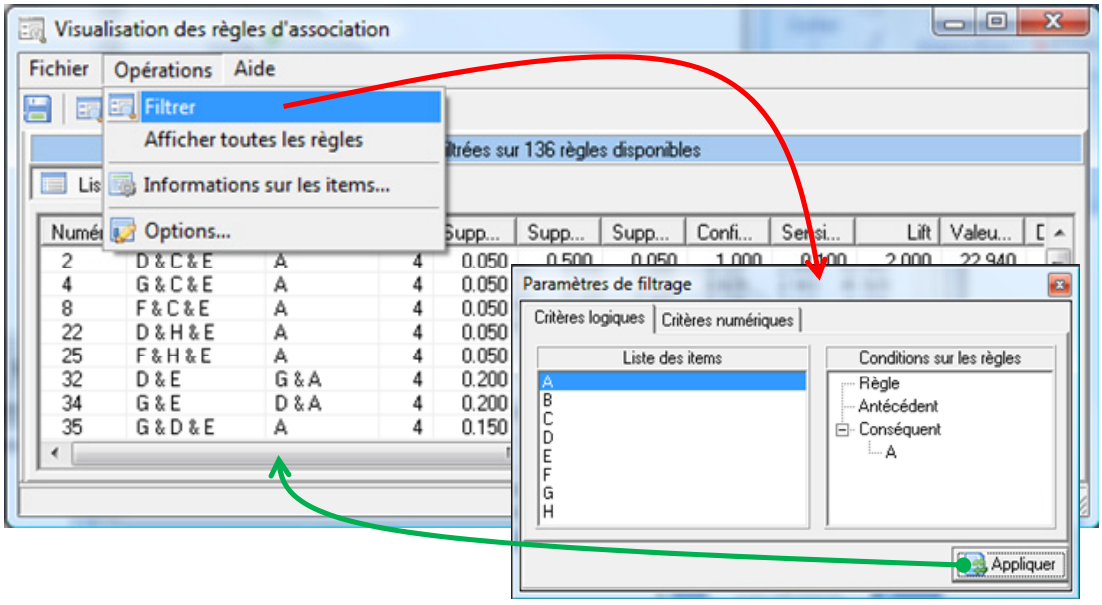
Additional tools enable to better identify the content of the results. We click on the menu OPERATIONS / INFORMATIONS SUR LES ITEMS. We obtain the list of items.



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

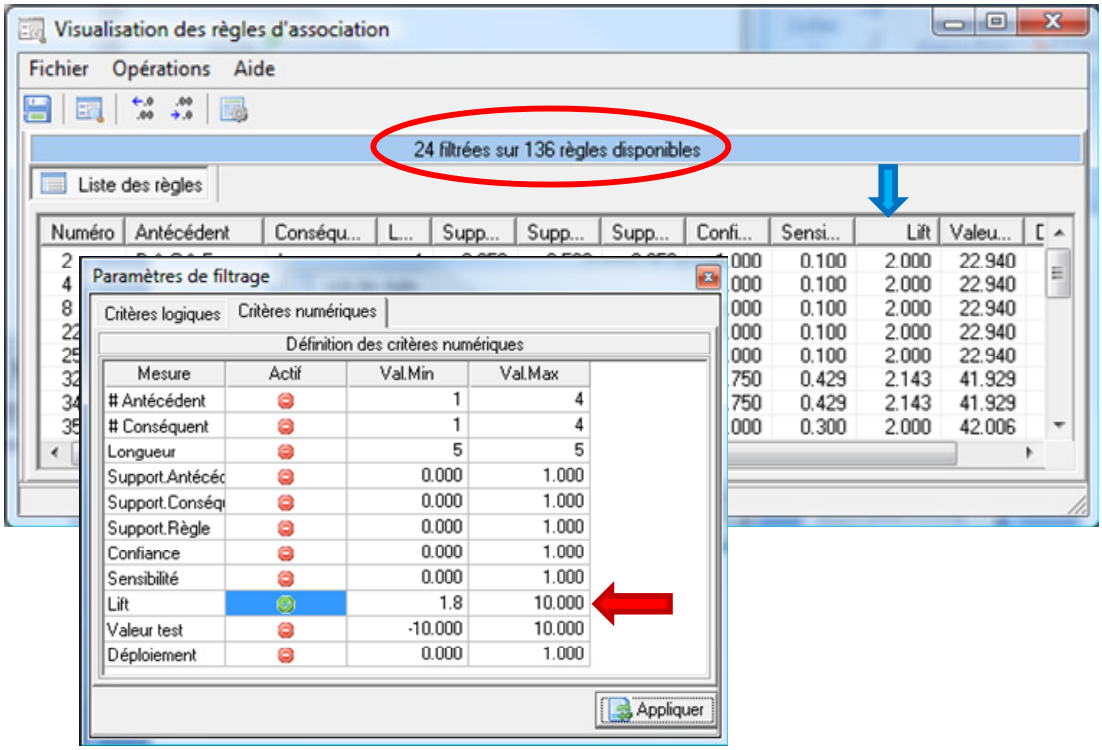
SPAD provides tools for filtering the rules. By clicking on the OPERATIONS / FILTRER menu, a dialog box appears. We can filter the rules according the presence of some items into the rules, or according numerical criteria.

Let us visualize the rules containing the item A into the consequent part. We drag the item A from the list of the items and drop it into the "Consequent" branch. We click on the "Appliquer" button: 30 rules are highlighted.



Now, we want to filter this subset of rules by retaining only those rules with a LIFT upper than 1.8 (and less than 10, but all the rules anyway have a LIFT less than 10).

In the "Critère numérique" tab, we set the appropriate bounds. Then, we activate these new filtering parameters.



We click on the "Appliquer" button. We obtain 24 rules.

Like this, we can highlight the subset of rules which is the most relevant according to the goals and the constraints of our analysis.

4 Association Rule Mining with SAS EM 4.3

I am not an expert in SAS, even less with regard "Enterprise Miner". The description proposed here can appear simplistic sometimes. The most important thing is to obtain the desired result.

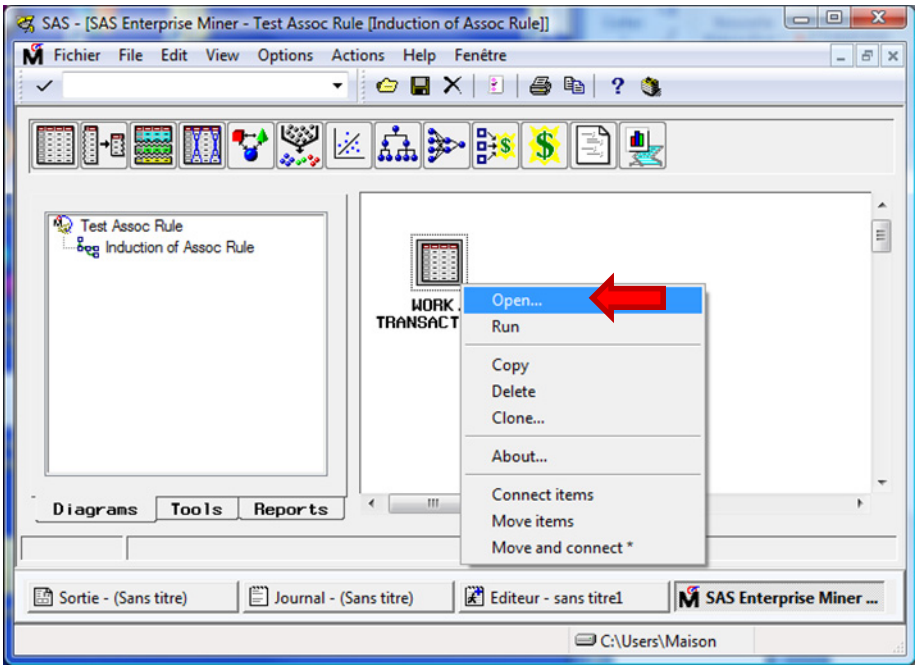
4.1 Importing the data file

First step, we want to import the "transactions.txt" data file into the WORK data library (***Note**: it is more appropriate to create a specific data library, but we use the most basic way in this tutorial*). To do this, after we launch SAS, we click on the File / Import Data menu. We select the "Tab Delimited File (*. txt)" format. We set the TRANSACTIONS as the name of the dataset into the WORK library.

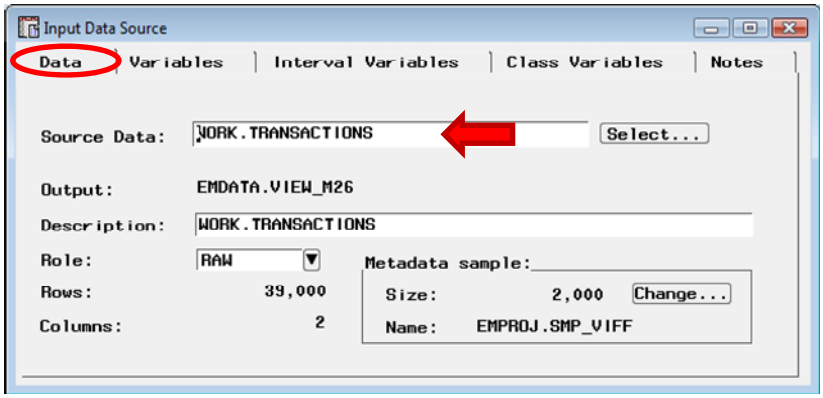
4.2 SAS Enterprise Miner

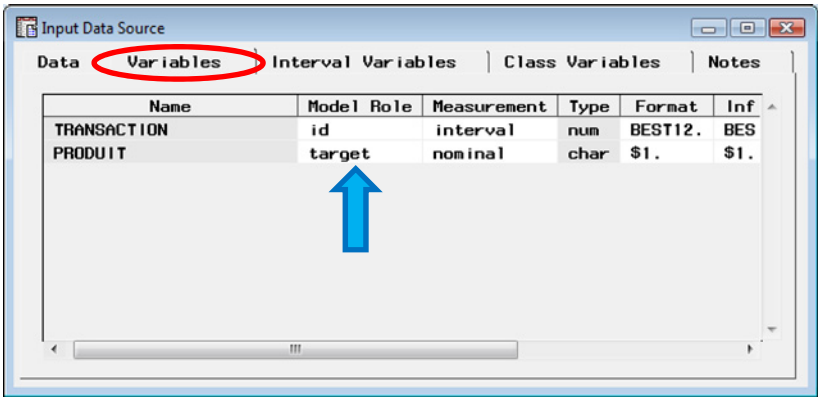
We click on the SOLUTIONS / ANALYSE / ENTERPRISE MINER menu to launch the Data Mining module of SAS. Usually a project has already been created with a default diagram. We set "Induction of Assoc Rule" as the name of the diagram.

The "Input Data Source" component enables to import the dataset. We add it into the workspace. We open the dialog settings by clicking on the OPEN menu.

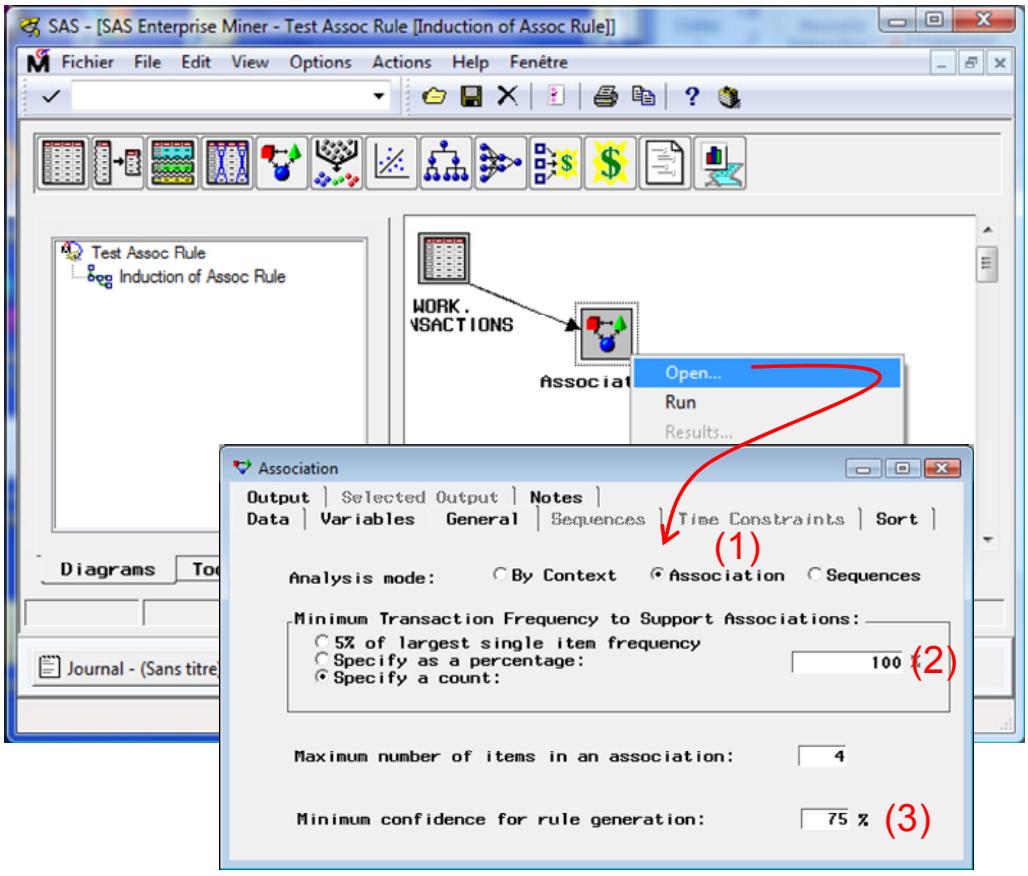


We select TRANSACTIONS from the WORK library as "Source Data". Into the VARIABLES tab, we specify the role of the columns using the SET MODEL ROLE contextual menu: TRANSACTION is the ID; PRODUIT is the TARGET.



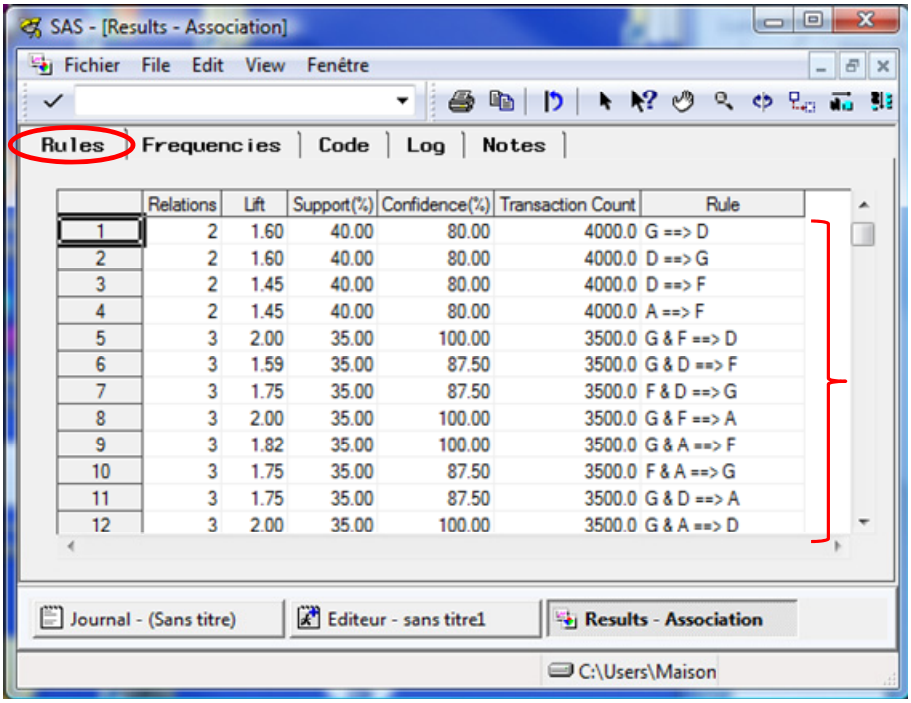


Then we add the ASSOCIATION component into the stream canvas. We set the following settings.



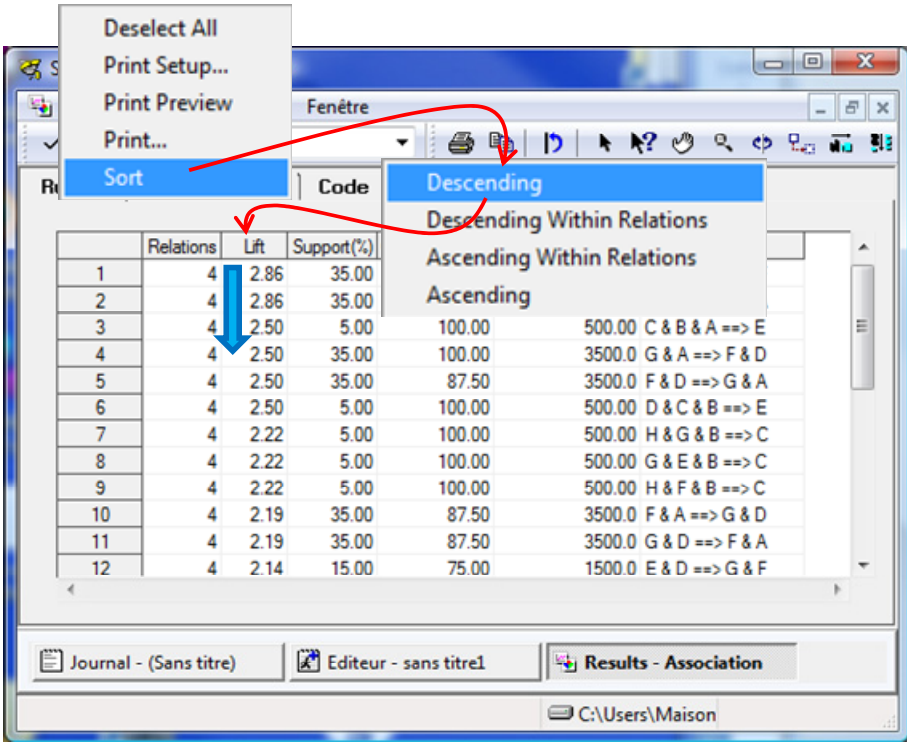
We want to obtain association rules (1). The minimum support of the rule is 100 (2). The maximum cardinal of the rules is 4. And the confidence minimum is 75% (3).

We confirm these choices. Then we click on the RUN menu. The rules are listed in a new visualization window.



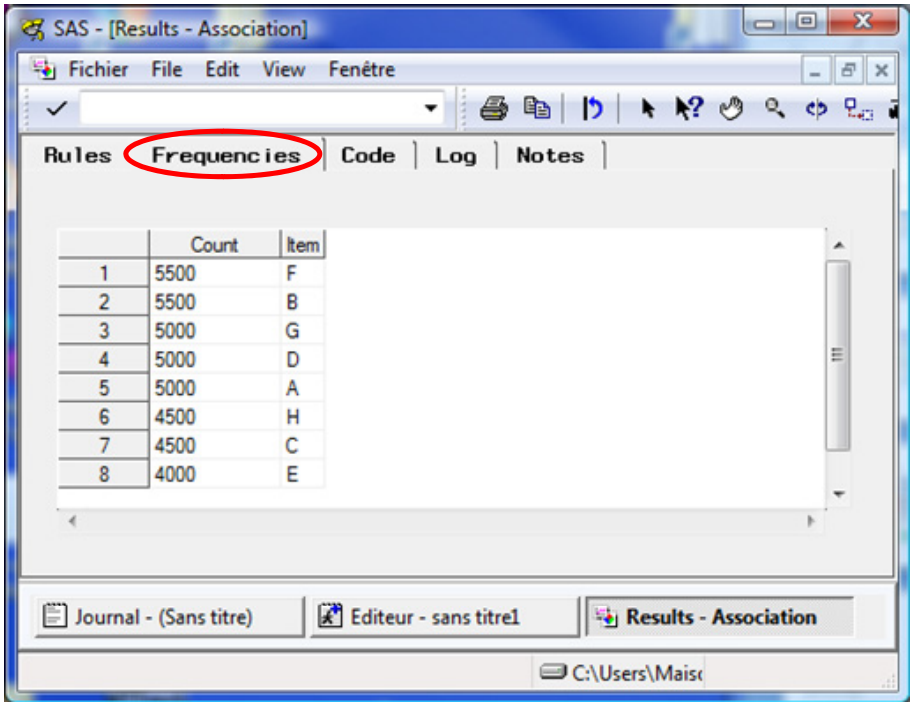
	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.60	40.00	80.00	4000.0	G ==> D
2	2	1.60	40.00	80.00	4000.0	D ==> G
3	2	1.45	40.00	80.00	4000.0	D ==> F
4	2	1.45	40.00	80.00	4000.0	A ==> F
5	3	2.00	35.00	100.00	3500.0	G & F ==> D
6	3	1.59	35.00	87.50	3500.0	G & D ==> F
7	3	1.75	35.00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

Like SPAD, 136 rules are extracted. We can sort them also according to the LIFT criterion.

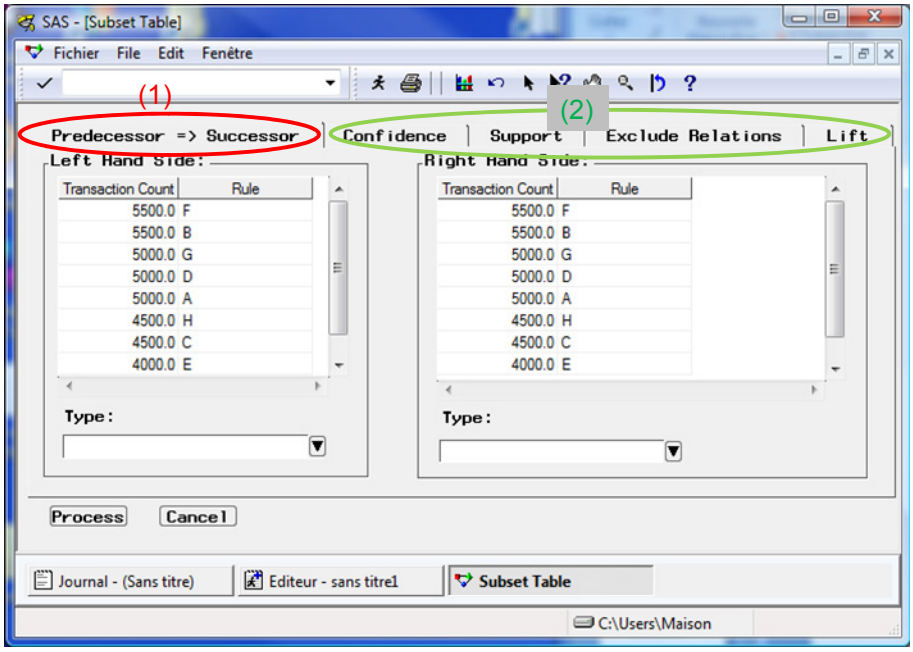


	Relations	Lift	Support(%)			
1	4	2.86	35.00			
2	4	2.86	35.00			
3	4	2.50	5.00	100.00	500.00	C & B & A ==> E
4	4	2.50	35.00	100.00	3500.0	G & A ==> F & D
5	4	2.50	35.00	87.50	3500.0	F & D ==> G & A
6	4	2.50	5.00	100.00	500.00	D & C & B ==> E
7	4	2.22	5.00	100.00	500.00	H & G & B ==> C
8	4	2.22	5.00	100.00	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

Into the FREQUENCIES tab, we have the list of the items and their count.



Like SPAD again, we can display only a part of the rule set using logical or numerical criteria. We click on the VIEW / SUBSET TABLE for that.



5 Association Rule Mining with Tanagra and Knime

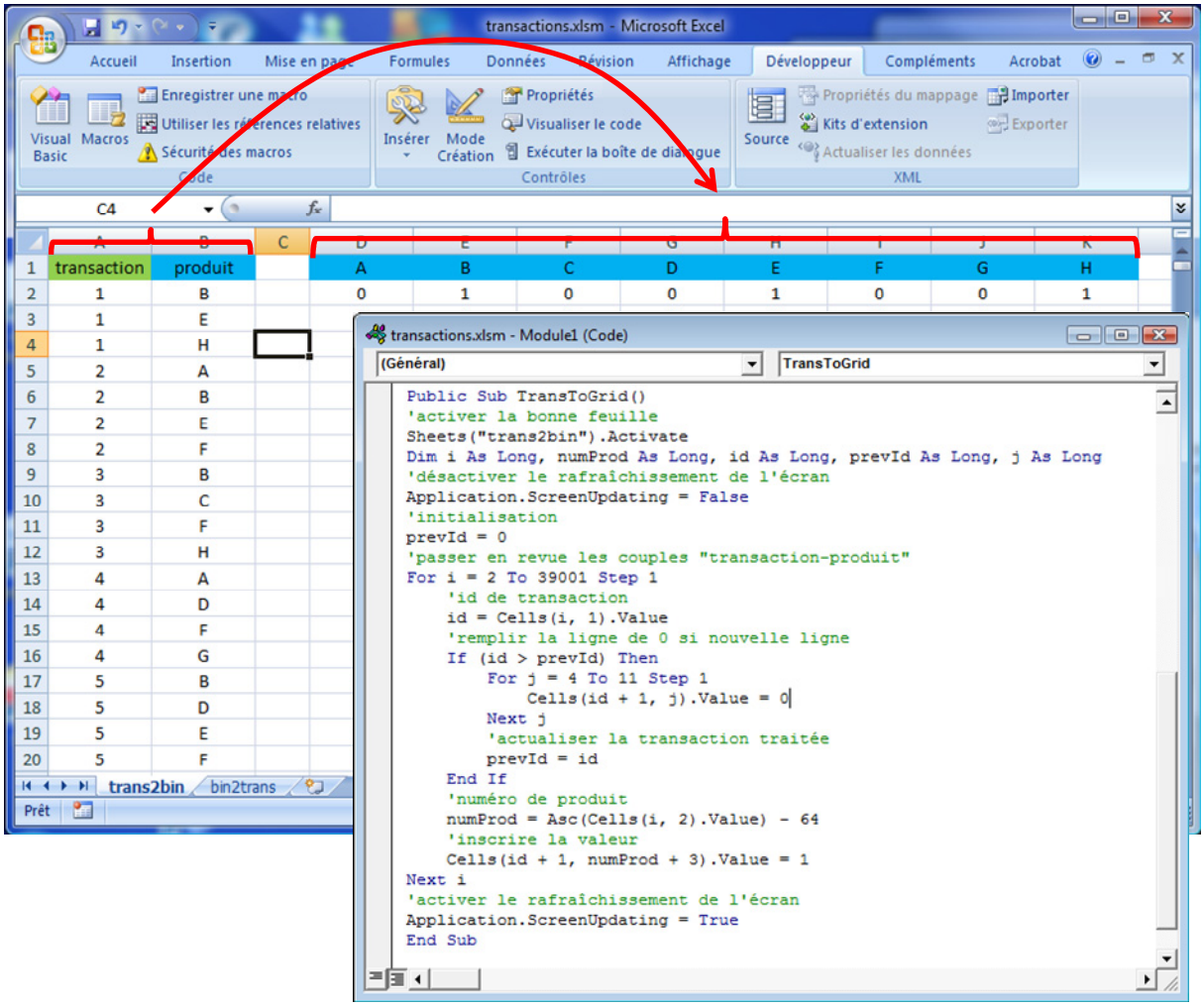
In this section we show how to transform the dataset prior to the importation into Tanagra and Knime. Again, the objective of the analysis is to extract only the positive rules (the items co-occurrence in the transactions). We create an intermediate data file where we use a 0/1 coding scheme to specify the presence of the items into the transactions.

We have already shown how Tanagra handles this kind of format for the extraction of association rule (http://data-mining-tutorials.blogspot.com/2008/11/association-rule-learning-from.html). In

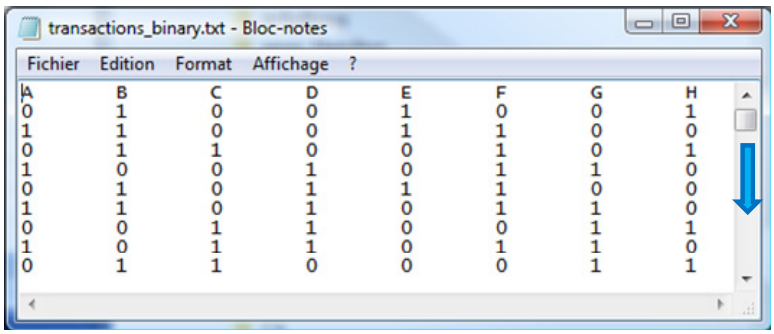
this tutorial, we propose a very simplistic program (in VBA) to transform the transactions file into this format. Thereafter, we show how to treat the transformed data file with Tanagra and Knime. *A priori*, the obtained rules must be the same as the ones extracted with SPAD and SAS.

5.1 Transforming the data file

We use the following program to transform the dataset under Excel.

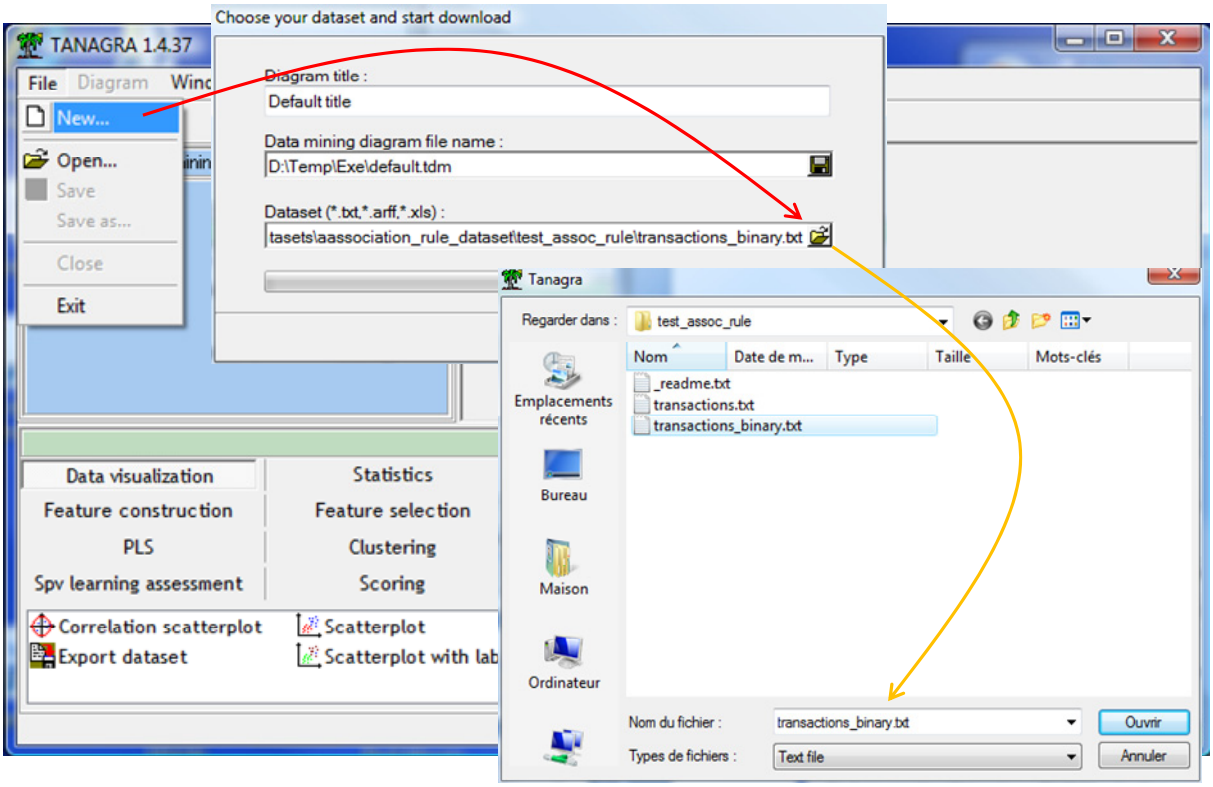


The program is rather simple. It relies on two elements to generate the binary table: the transaction ID indicates the row that must be completed; the ASCII code of the product name (this solution is highly specific to our data representation) to detect the column to fill. All the blanks cells are filled with zeros. We export the new data table into the "transactions_binary.txt" text file. Here are the first rows of the data file.

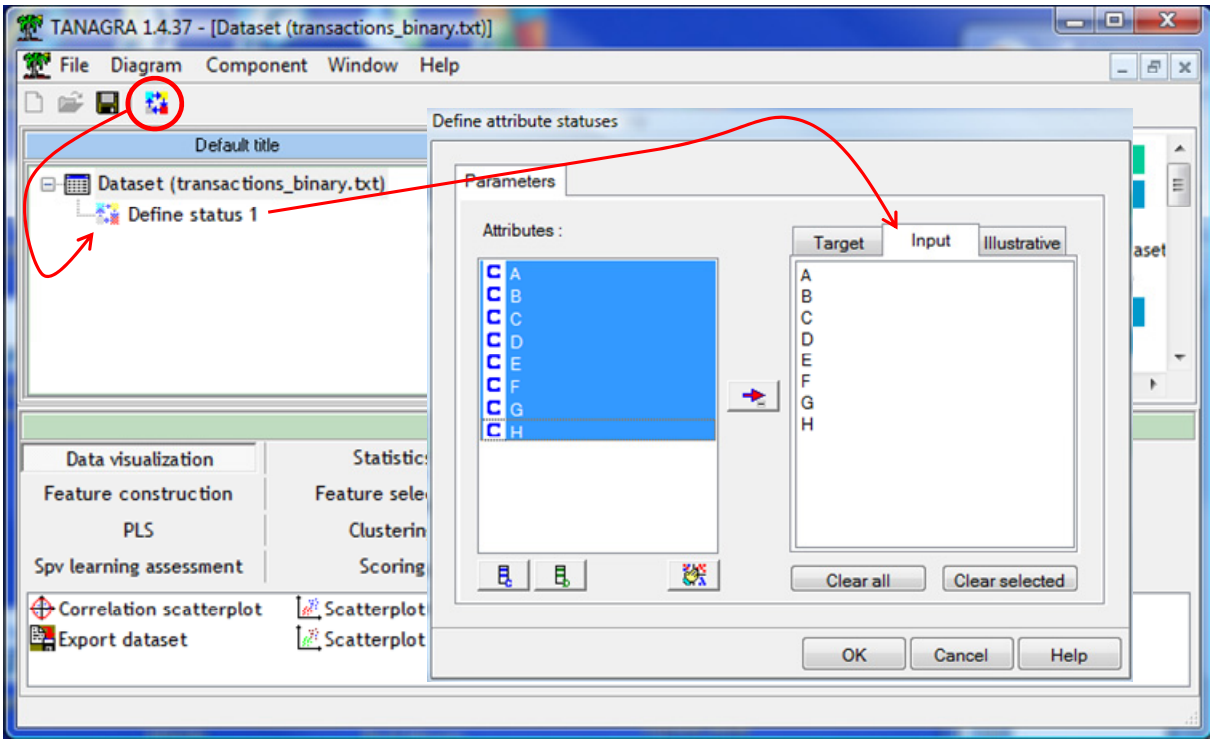


5.2 Performing the analysis with Tanagra

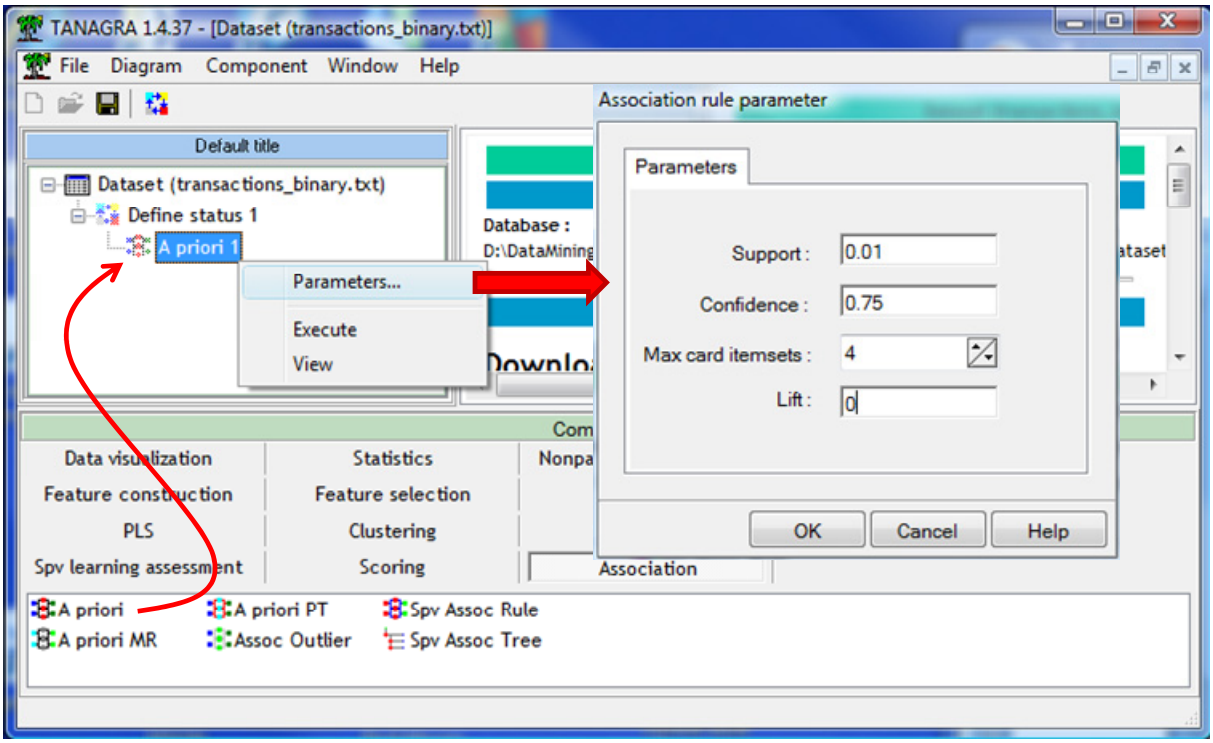
We launch Tanagra. Then, we create a new diagram by clicking on the FILE / NEW menu. We select the "transactions_binary.txt" data file.



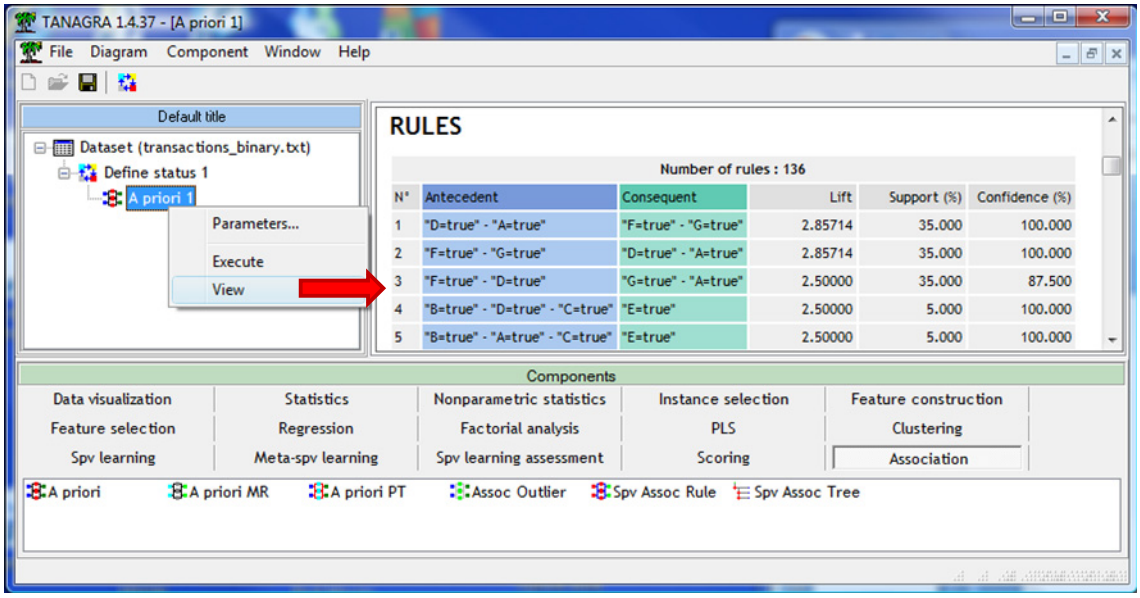
We must specify the role of the columns. We use the DEFINE STATUS component. We set the 8 variables as INPUT.



We insert the A PRIORI component (ASSOCIATION tab) into the diagram. We set the following settings (PARAMETERS contextual menu): (Support min = 0.01; Confidence min = 0.75; Max Cardinal = 4; Lift min = 0.0).



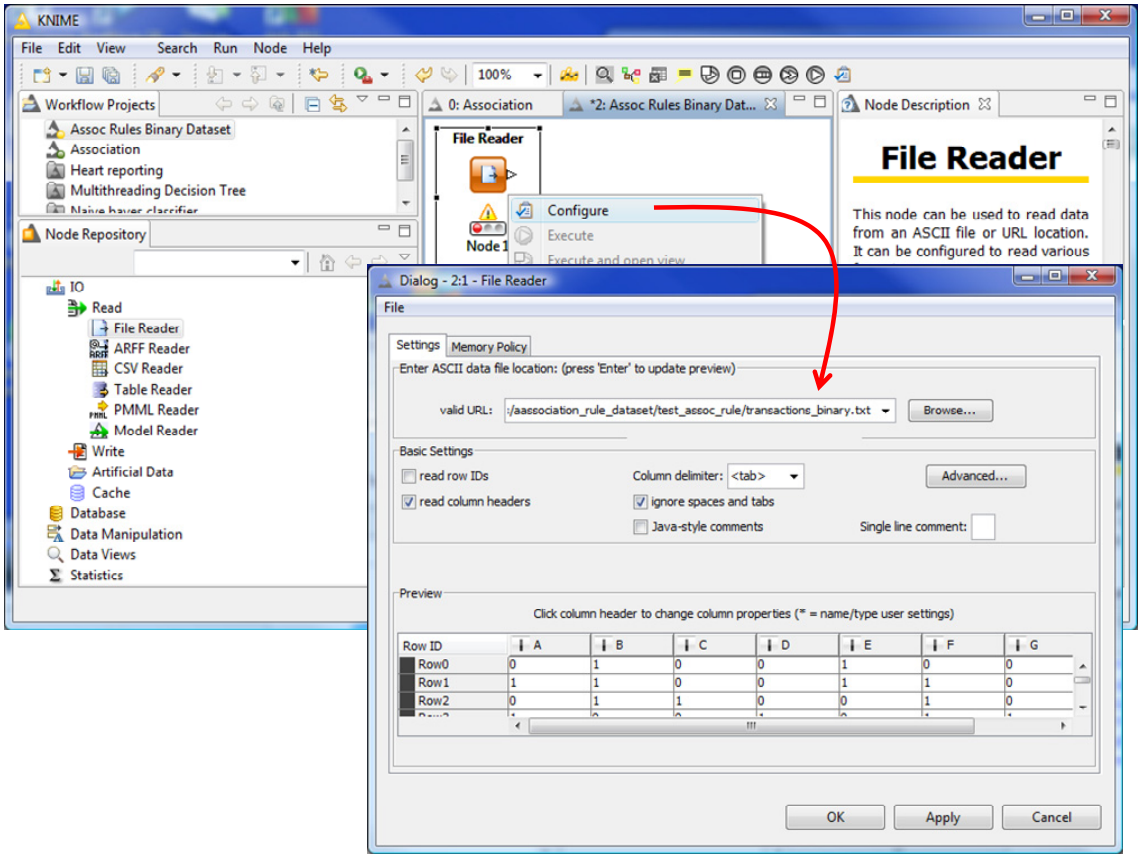
We confirm and we click on the VIEW menu to launch the calculations.



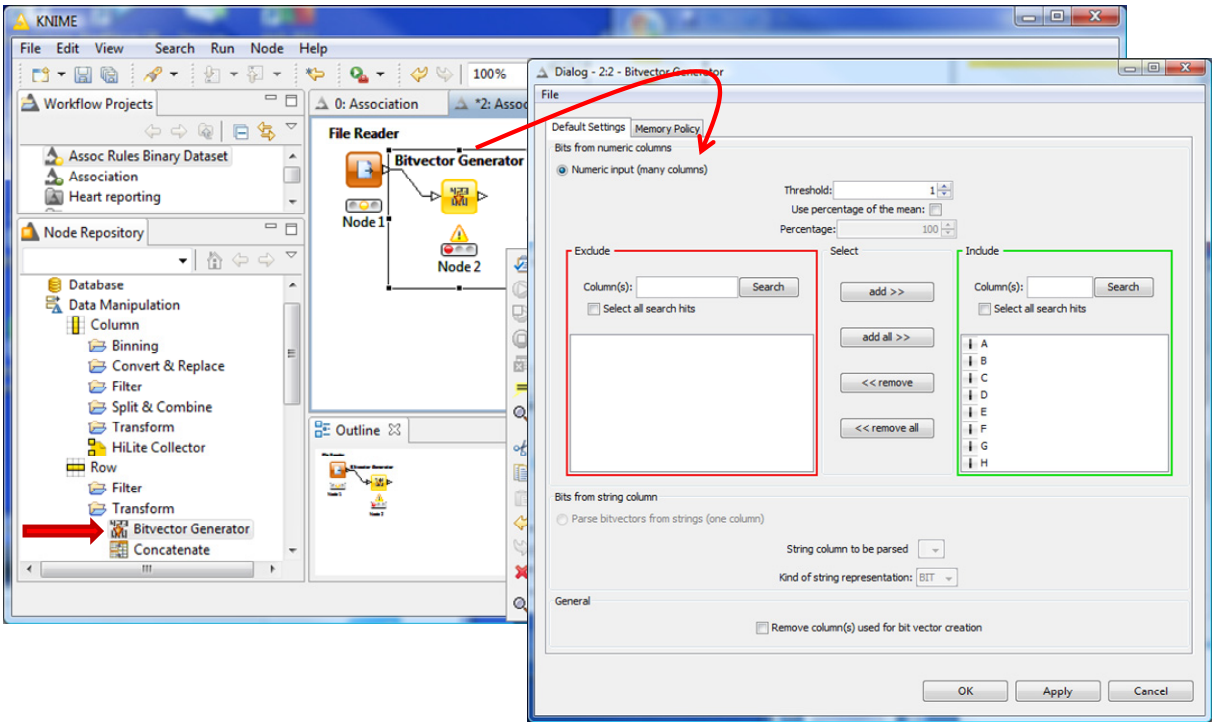
We obtain 136 rules, like SPAD and SAS. We note that Tanagra generates only the positive rules which highlight the simultaneous presence of the items in transactions. The rules are sorted according the LIFT criterion (decreasing order). Unlike SPAD or SAS, we cannot handle interactively the rules.

5.3 Performing the analysis with Knime

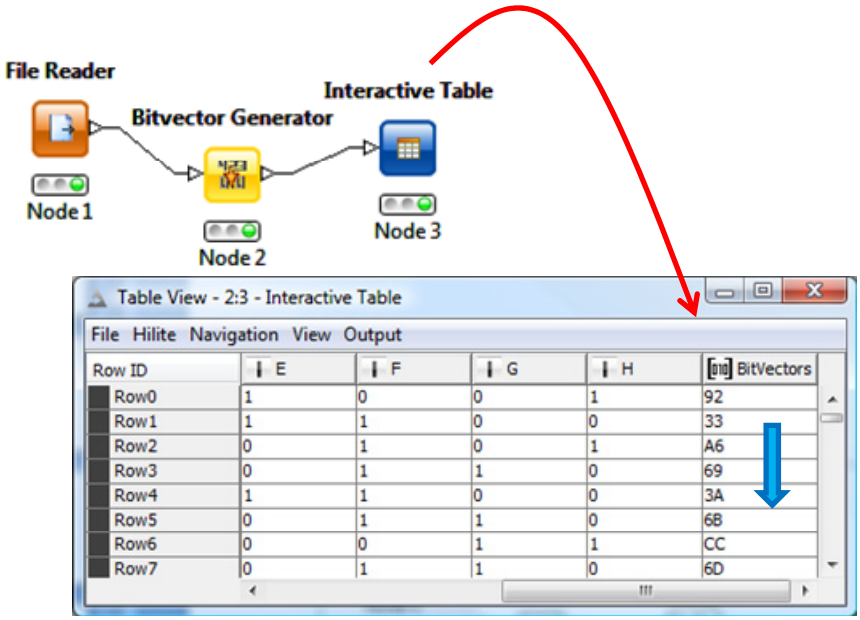
We create a diagram into Knime. We import the data file using the FILE READER component.



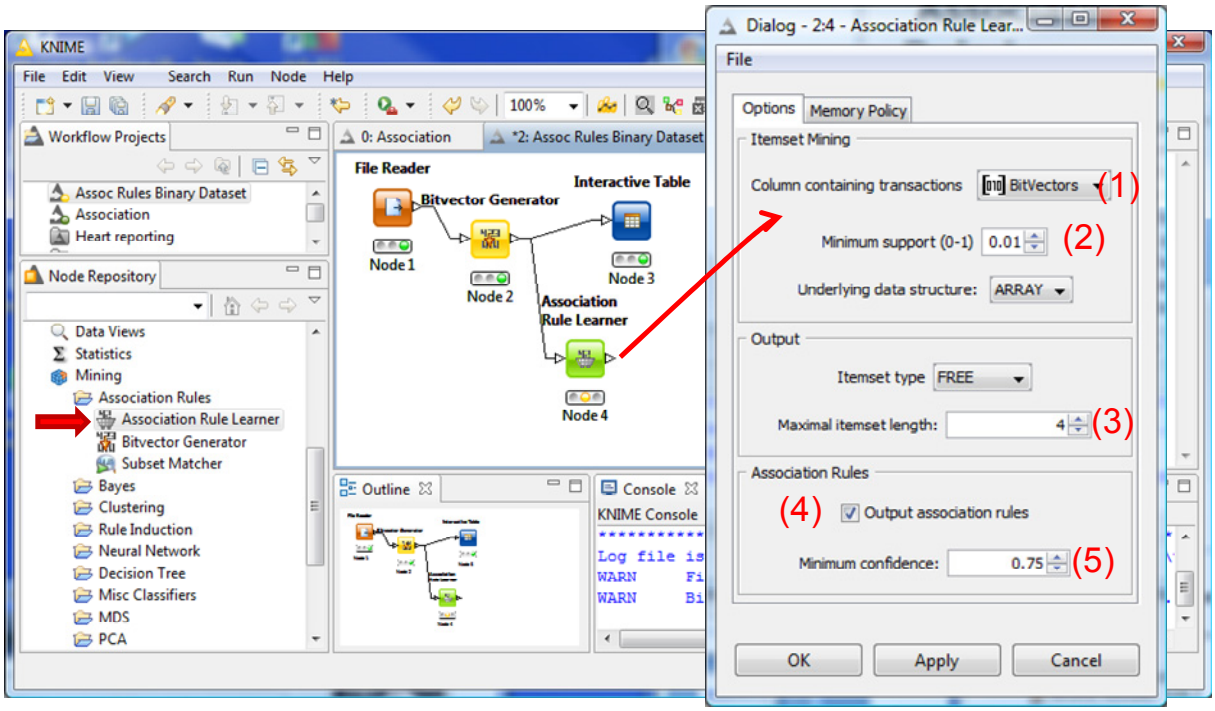
Another transformation is needed to generate association rules. We use the BITVECTOR GENERATOR component for this. It transforms each transaction into a vector of bits. We set the following parameters (CONFIGURE menu)



We use the INTERACTIVE TABLE component to visualize the new column. BITVECTORS corresponds to a vector of bits in a hexadecimal encoding scheme.

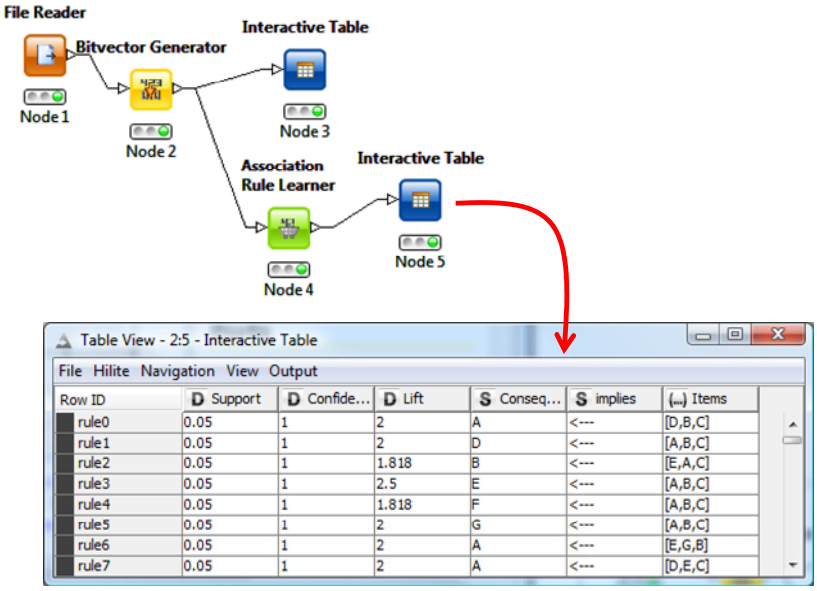


We can launch the extraction of the rules using the ASSOCIATION RULE LEARNER component. We set the following settings (CONFIGURE menu).



The transactions are described by the BITVECTORS column (1). The minimum support of the generated rules is 0.01 (2). The maximal itemset length is 4 (3). Last, we want to generate rules (4) with a confidence upper than 0.75 (5).

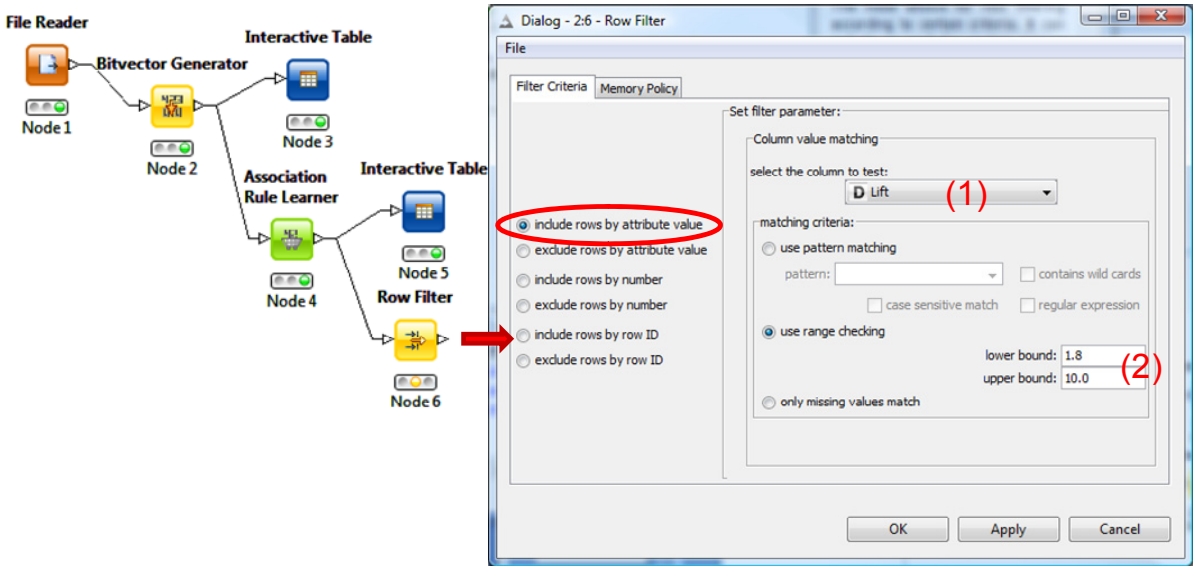
We click on the EXECUTE menu to launch the calculations. We can visualize the rules with the INTERACTIVE TABLE component.



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]

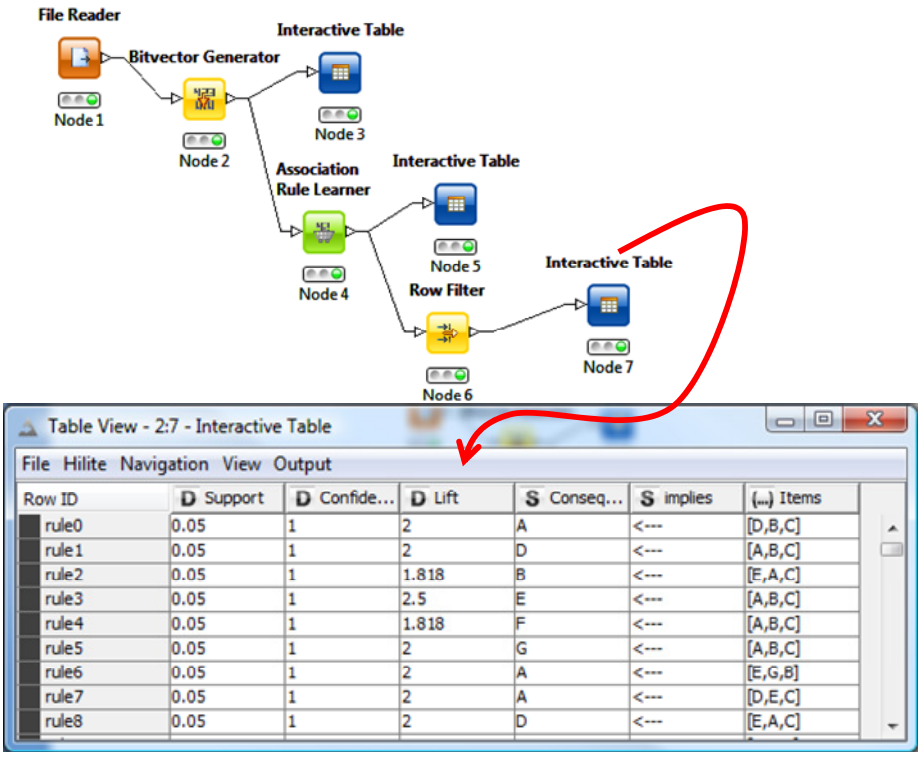
Unlike the other tools, Knime generates only the rules with one item into the consequent. We obtain 94 rules.

We can filter the rules with the ROW FILTER component. We want to visualize the rules with a LIFT upper than 1.8 (and lower than 10.0). We parameterize the component as follows.



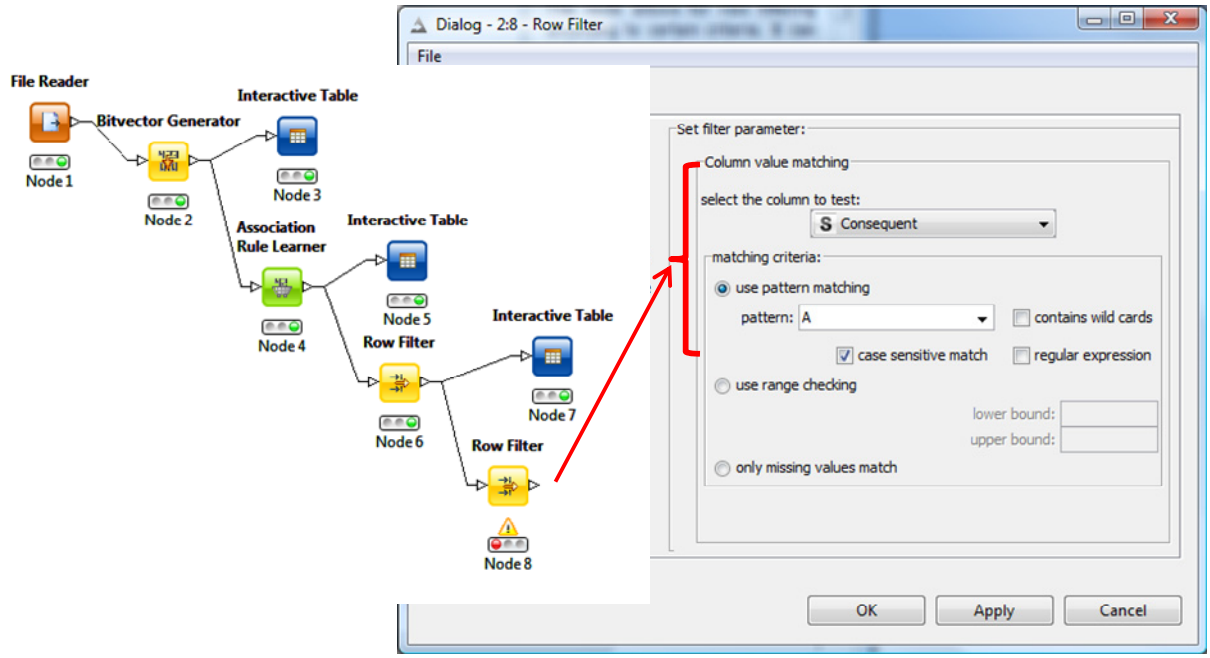
The filtering is based on the LIFT criterion (1). We set the bounds of the values (2).

We use another INTERACTIVE TABLE component in order to visualize the filtered rule set.

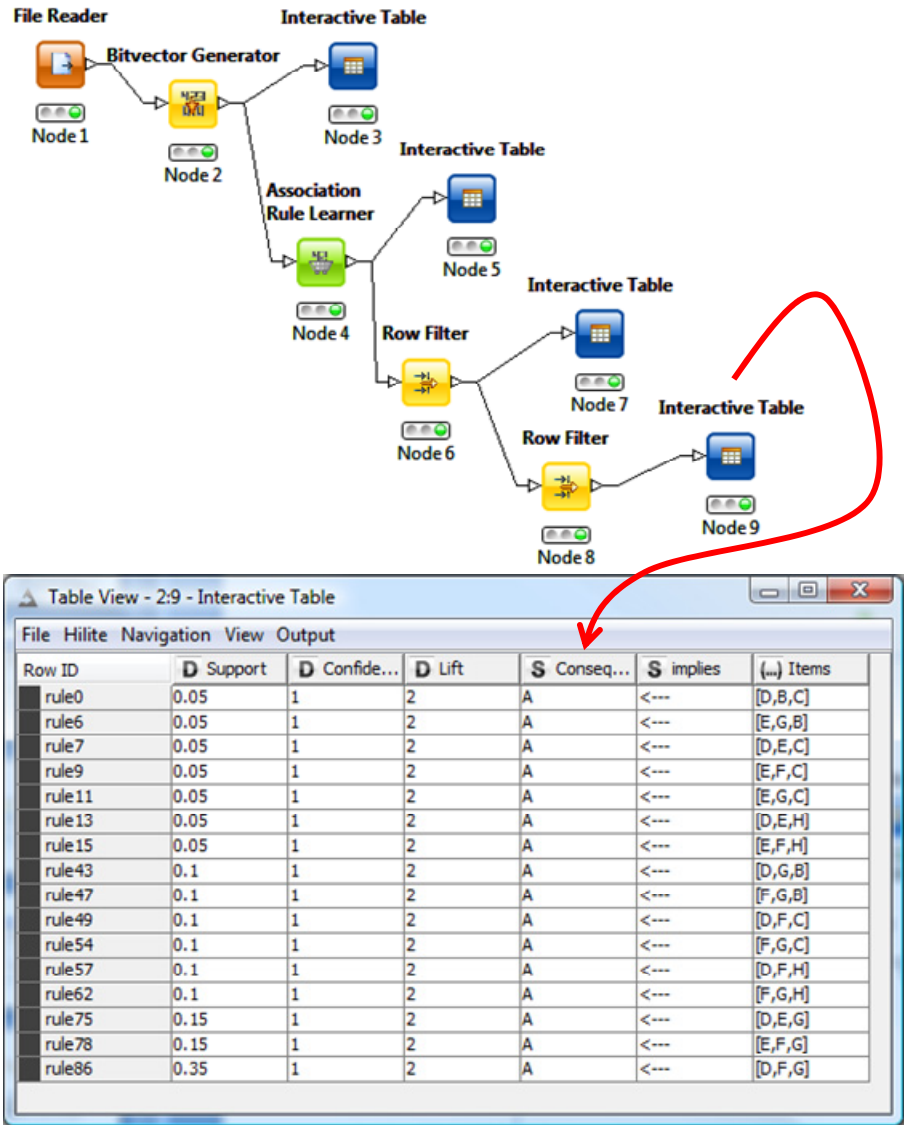


Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.05	1	2	A	<---	[D,B,C]
rule1	0.05	1	2	D	<---	[A,B,C]
rule2	0.05	1	1.818	B	<---	[E,A,C]
rule3	0.05	1	2.5	E	<---	[A,B,C]
rule4	0.05	1	1.818	F	<---	[A,B,C]
rule5	0.05	1	2	G	<---	[A,B,C]
rule6	0.05	1	2	A	<---	[E,G,B]
rule7	0.05	1	2	A	<---	[D,E,C]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

We can refine the filtering. For instance, based on this new set of rules, we can highlight the ones with the item A into the consequent. We add the ROW FILTER component. The settings are specified about the consequent of the rule here.



Now, we have only 16 rules.

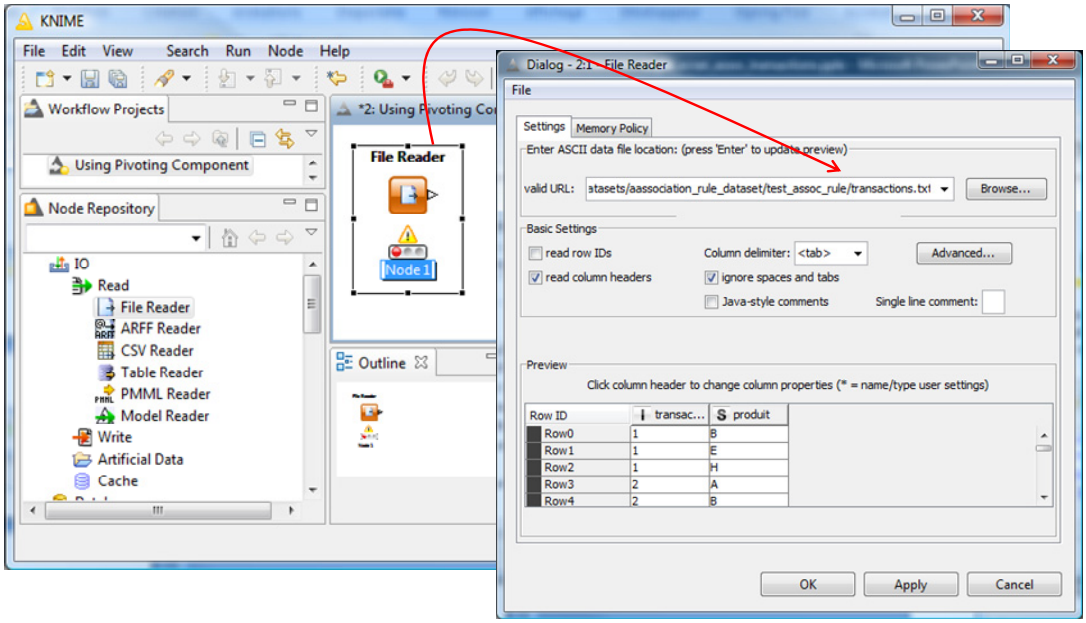


Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.05	1	2	A	<---	[D,B,C]
rule6	0.05	1	2	A	<---	[E,G,B]
rule7	0.05	1	2	A	<---	[D,E,C]
rule9	0.05	1	2	A	<---	[E,F,C]
rule11	0.05	1	2	A	<---	[E,G,C]
rule13	0.05	1	2	A	<---	[D,E,H]
rule15	0.05	1	2	A	<---	[E,F,H]
rule43	0.1	1	2	A	<---	[D,G,B]
rule47	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]

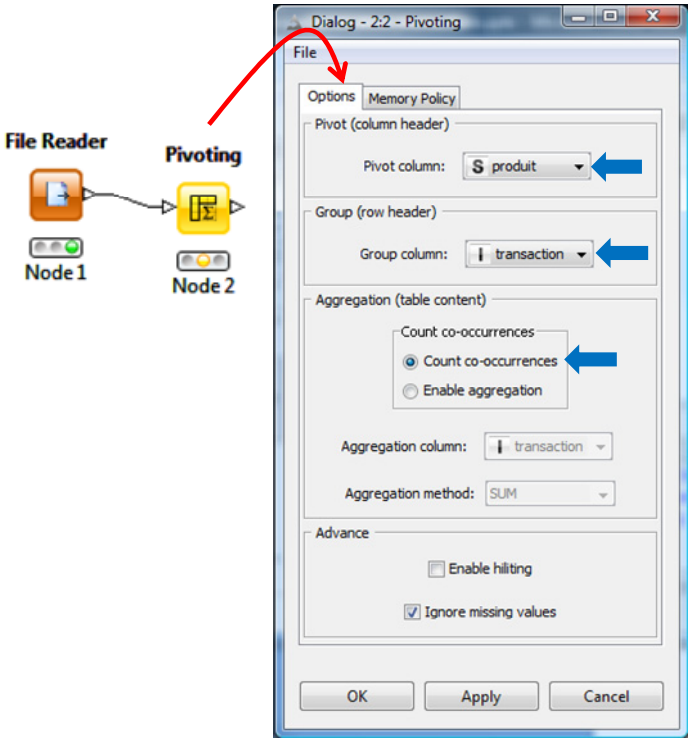
5.4 Performing the analysis with Knime (again)

After the publication of the French version of this tutorial, Loïc LUCEL told me that it was possible to generate within Knime the 0/1 data table. We can therefore avoid the priori recoding using an external program (in VBA here). The operation can be included directly into the stream diagram. Thank you very much for these indications Loïc.

Let's see this new approach. We create a new Workflow Project. Using the FILE READER component, we import the transactions file "transactions.txt". We note that Knime recognizes Knime automatically the transaction ID as integers [Integer], the products column as strings [String].



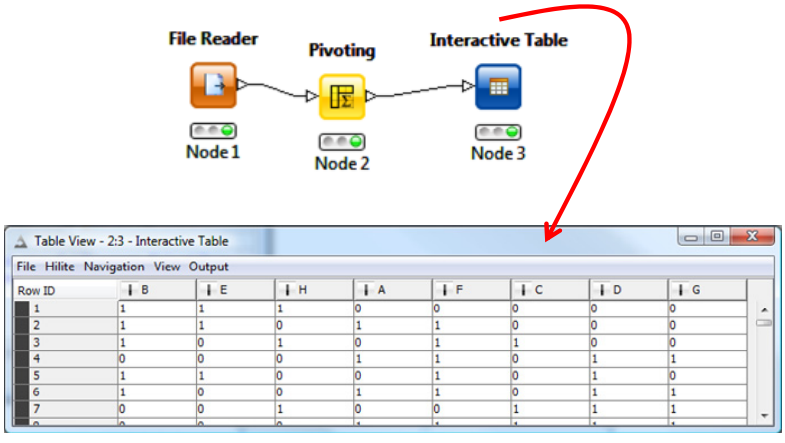
Then, we add the PIVOTING component (DATA MANIPULATION / ROW / TRANSFORM branch).



About the settings, we set PRODUIT as PIVOT, TRANSACTION as GROUP. Thus, we count the presence of each item for each transaction. Since the transactions file counts only the presence (or absence) of the item, the possible values are 0 or 1.

We check this with the INTERACTIVE TABLE. We have indeed the appropriate table with the 0/1 coding scheme.

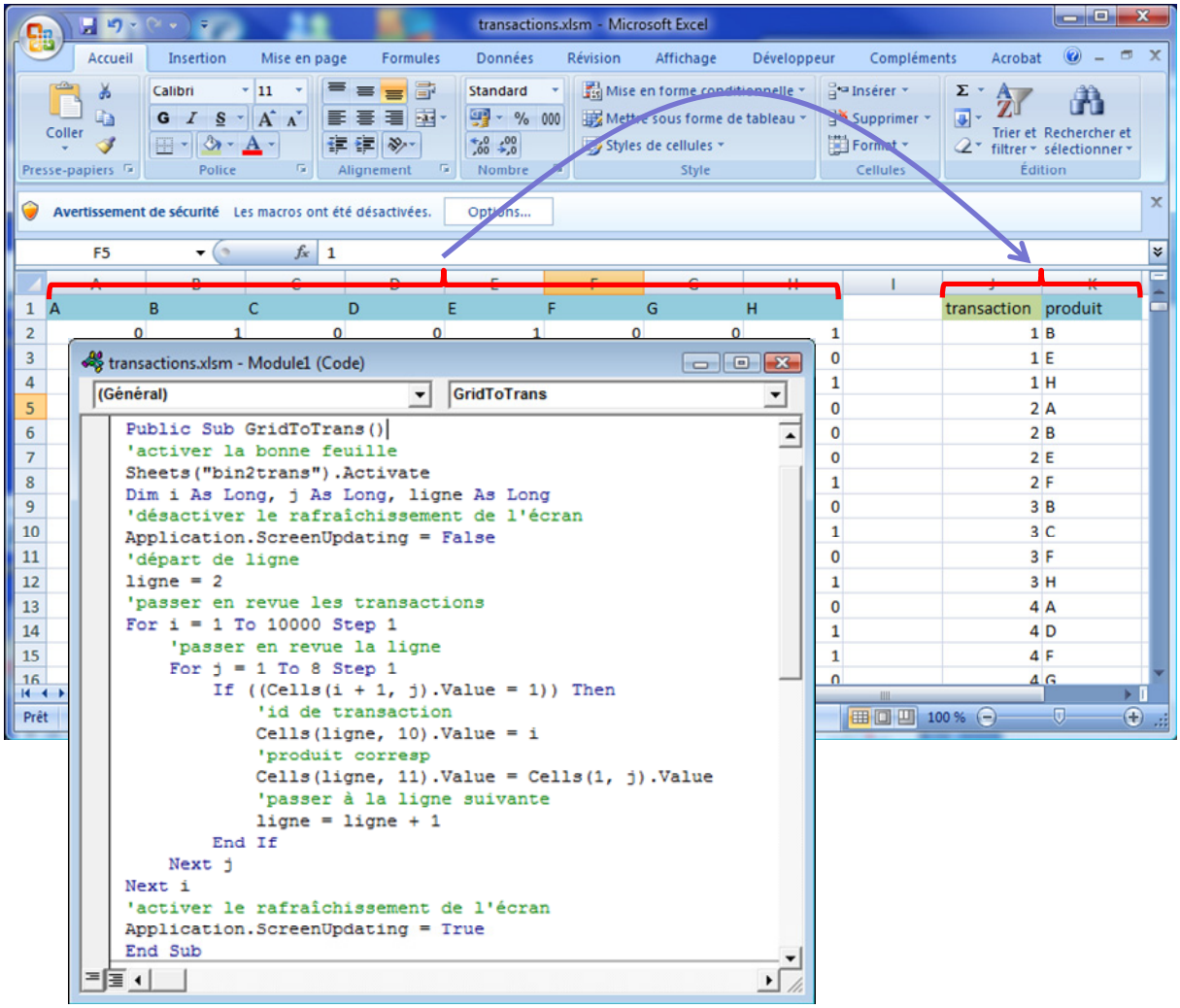
Therefore, as in the previous section, we have to add a BIT VECTOR GENERATOR after the PIVOTING component and extract the association rules by following the same approach.



The main advantage of the procedure described in this section is that the data transformation is included into the automated process. Thus, it remains applicable if the source file is modified i.e. if we want to refresh the transactions file and perform a new extraction of the rules.

5.5 From the binary table to the transactions table

Out of curiosity, we show in this subsection the reverse transformation, from the binary table to the transactions table. As we can see, the VBA code is very simple also. The generalization to other databases, with more items and transactions is easy.



6 Conclusion

The ability to handle transactions file is an asset for the extraction of association rules. The ease of operations with SPAD and SAS shows this. Curiously, this feature is lacking in some generalists (academic) tools.

In this tutorial, we show that the data transformation in order that free tools such as Tanagra can handle this kind of dataset is rather simple. We can use a little program written in VBA (under Excel). The limitation is not too restrictive. We remind that Excel (version 2007 and later) can treat 1,048,575 "transaction id - item id" pairs (we do not count the first row) and 16,381 items (if the start the binary table form the D column into the worksheet). We can do many things with that.