

## Objectif

Comparer TANAGRA, ORANGE et WEKA lors de la recherche de règles d'association.

S'agissant de la construction de règles d'association, quel que soit le logiciel utilisé, nous devons impérativement passer par les étapes suivantes :

- Importer les données dans le logiciel ;
- Définir le problème à résoudre, c.-à-d. sélectionner les descripteurs ;
- Paramétrer la recherche des règles, il s'agit essentiellement de fixer leur support et leur confiance minimale ;
- Lancer l'apprentissage et visualiser les règles.

Nos trois logiciels s'accordent sur un point : ils prennent en entrée des données tabulaires (individus x variables) et procèdent à un codage disjonctif complet : chaque couple attribut-valeur devient alors un item candidat à la génération des règles.

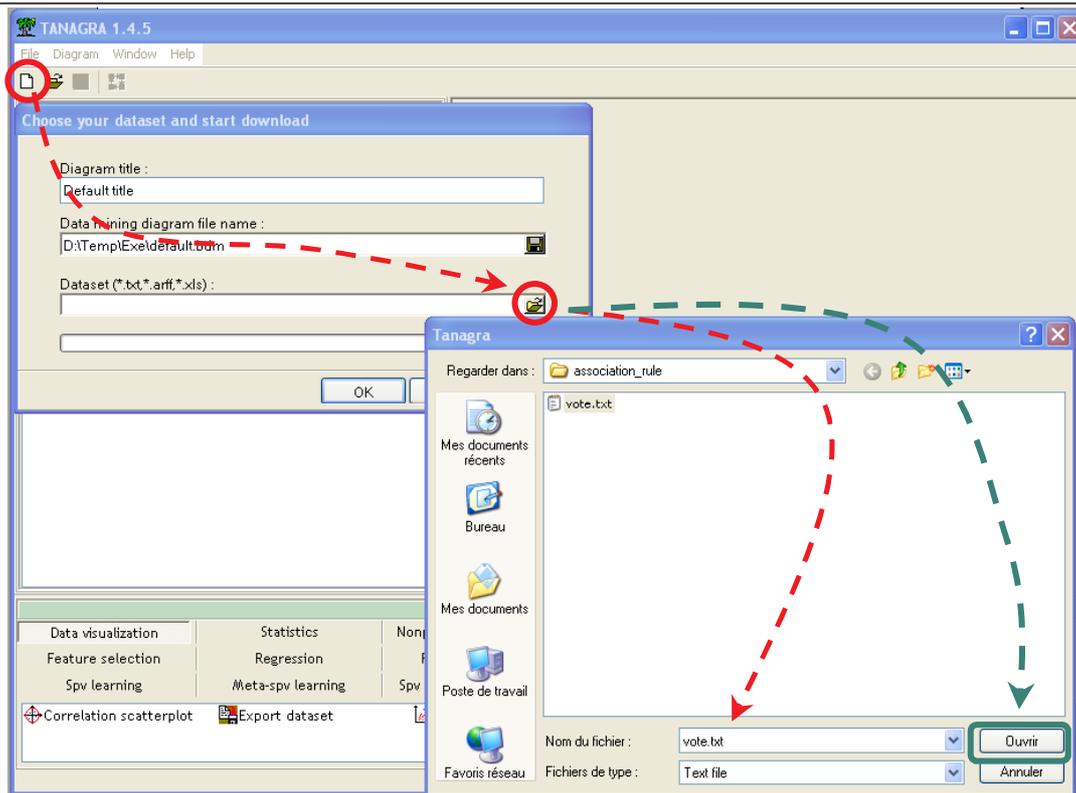
## Fichier

Nous utilisons le fichier VOTE.TXT (source UCI IRVINE), le fichier contient 435 observations.

## Construire des règles d'association avec TANAGRA

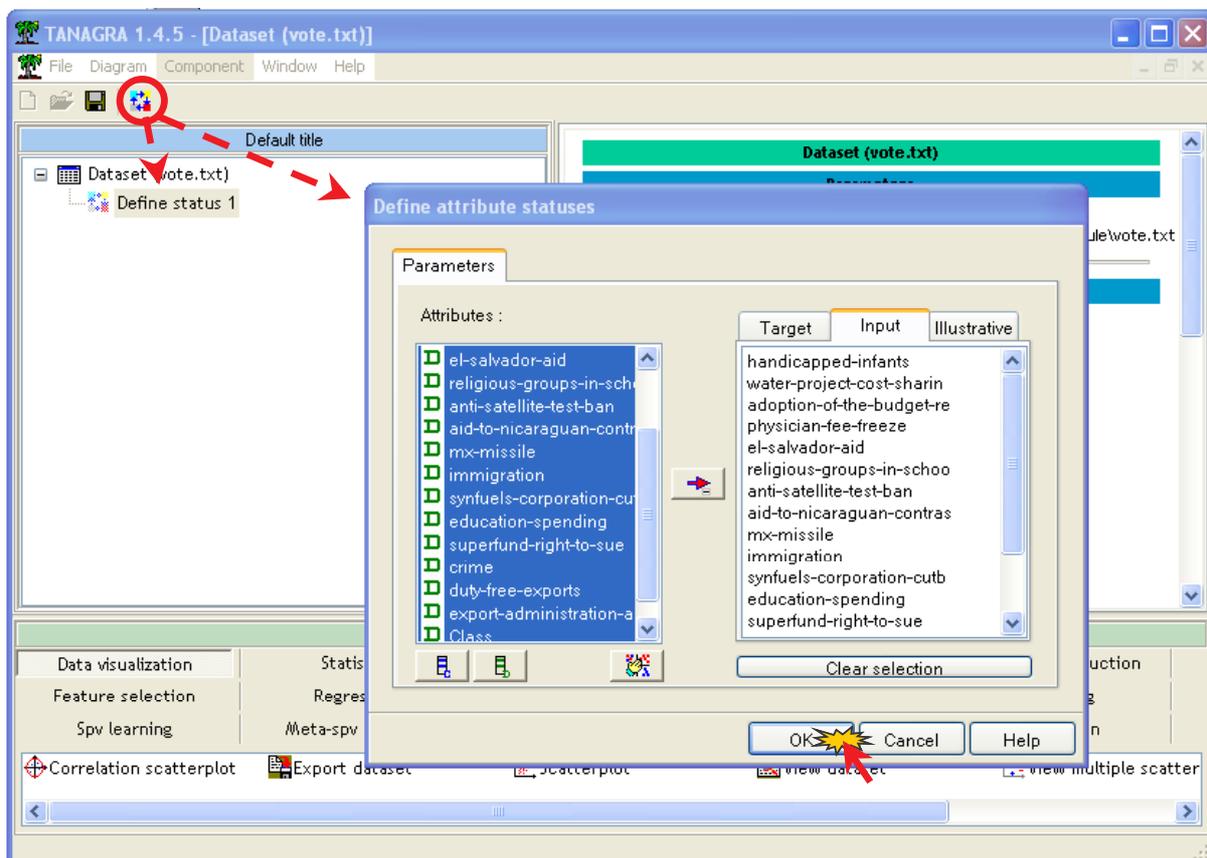
### Charger les données

Première étape toujours, après le lancement du logiciel, nous devons charger les données. Nous activons le menu FILE/NEW pour créer un nouveau diagramme et nous sélectionnons le fichier VOTE.TXT.



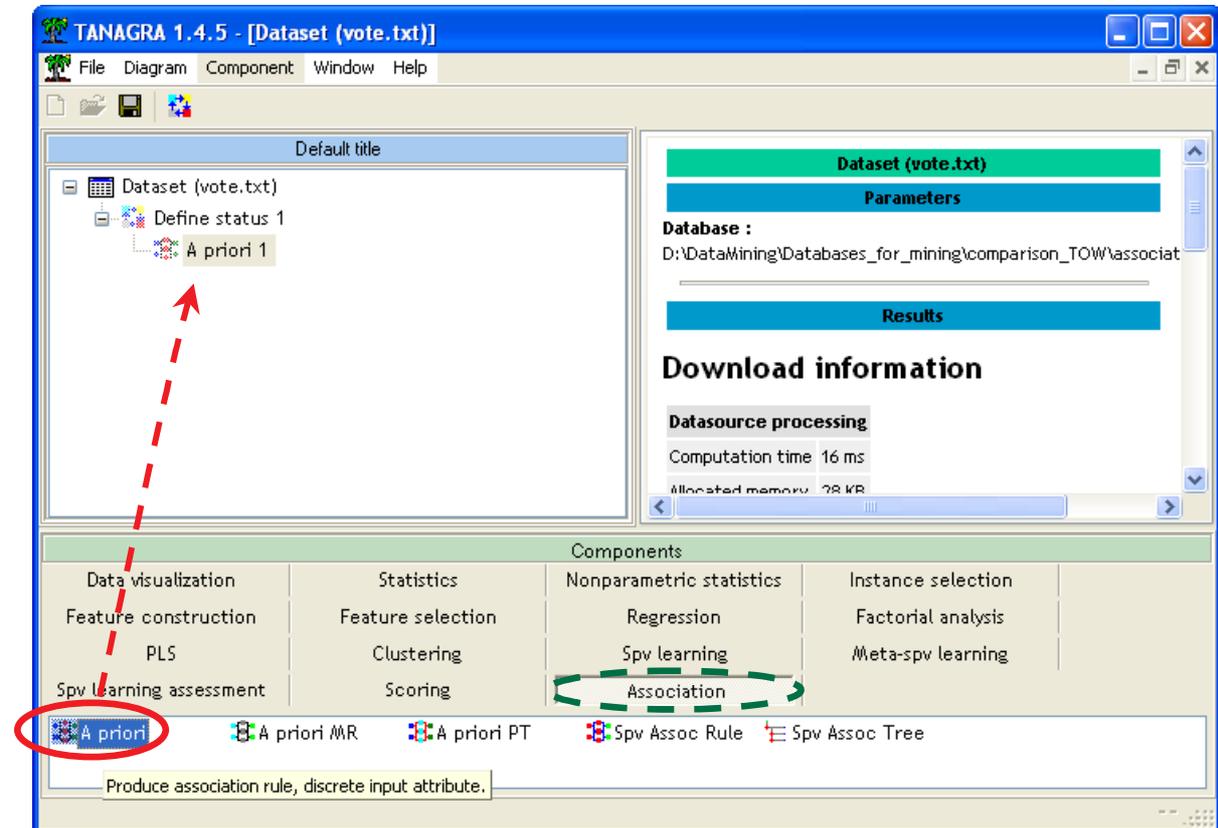
### Définir les variables de l'analyse

Nous ajoutons le composant DEFINE STATUS dans le diagramme en cliquant sur le raccourci dans la barre d'outils. Nous plaçons toutes les variables en INPUT.

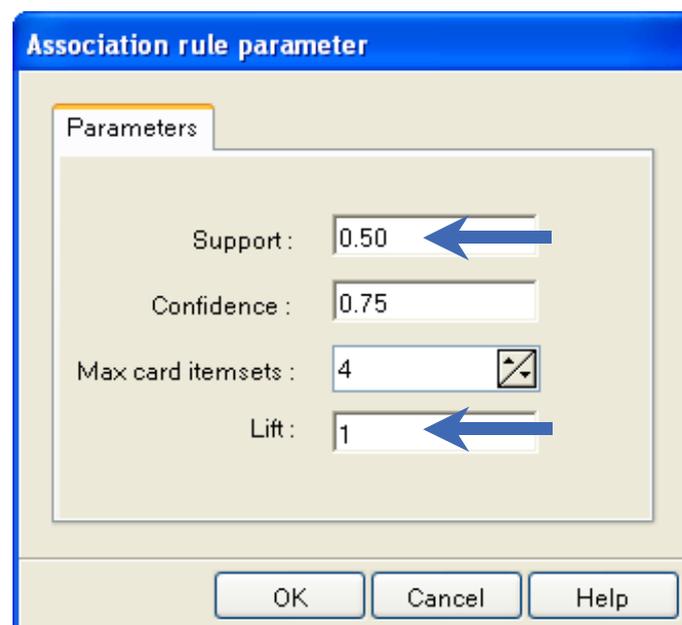


## Définir et paramétrer l'algorithme de génération des règles d'apprentissage

Il existe toute une série d'algorithmes de génération de règles d'association dans TANAGRA, certaines d'ailleurs proviennent de bibliothèques externes. Dans ce didacticiel, nous choisirons la méthode A PRIORI qui correspond aux caractéristiques définies dans les articles originels d'AGRAWAL.



Après l'avoir placé dans le diagramme, nous activons le menu contextuel PARAMETERS pour accéder aux paramètres, nous définissons les valeurs suivantes.



Le support minimal des règles est fixé à 0.5, leur confiance minimale à 0.75, nous générons les règles à partir des itemsets fréquents ayant un cardinal inférieur ou égal à 4, les règles présentant un LIFT inférieur à 1 sont éliminés.

## Exécution

Pour voir les résultats, nous devons activer le menu VIEW du composant A PRIORI, les règles apparaissent dans la fenêtre de droite, 14 règles ont été générées.

Results	
<b>ITEMS</b>	
Transactions	435
<b>Counting items</b>	
All items	50
Filtered items	12
<b>Counting itemsets</b>	
card(itemset) = 2	4
card(itemset) = 3	1
<b>Rules</b>	
<b>Number of rules</b>	<b>14</b>

## RULES

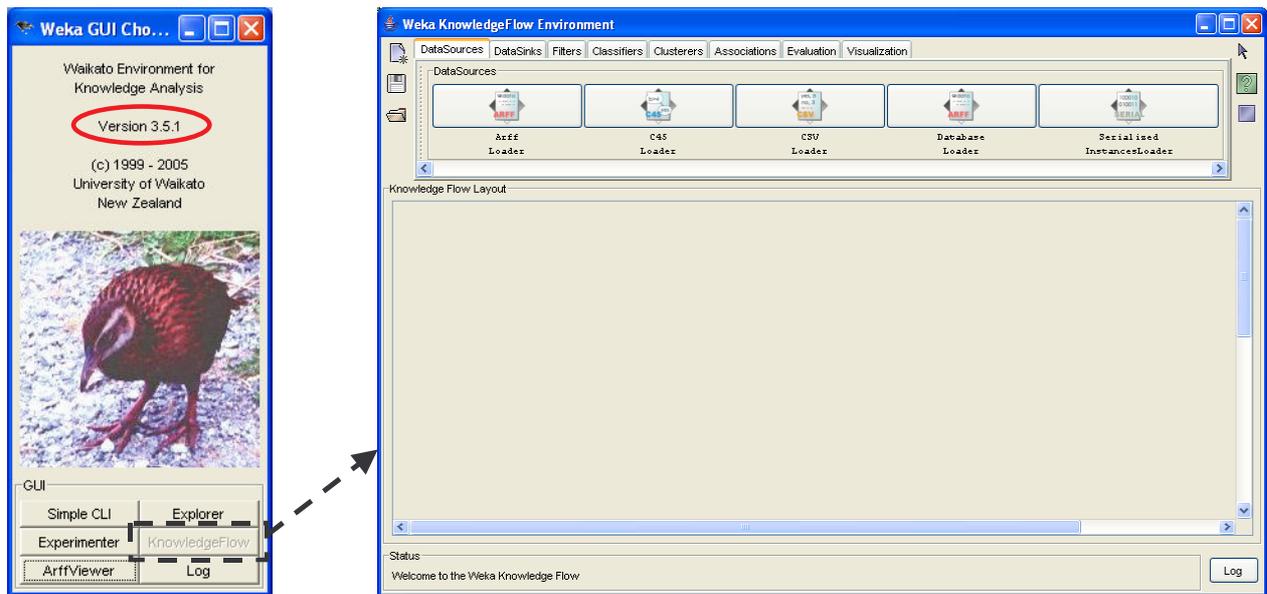
Number of rules : 14					
N°	Antecedent	Consequent	Lift	Support	Confidence
1	"physician-fee-freeze=n"	"Class=democrat" - "adoption-of-the-budget-re=y"	1.670	0.503	0.887
2	"Class=democrat" - "adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.670	0.503	0.948
3	"Class=democrat"	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	1.629	0.503	0.820
4	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	"Class=democrat"	1.629	0.503	1.000
5	"Class=democrat"	"physician-fee-freeze=n"	1.616	0.563	0.918
6	"physician-fee-freeze=n"	"Class=democrat"	1.616	0.563	0.992
7	"adoption-of-the-budget-re=y"	"Class=democrat" - "physician-fee-freeze=n"	1.537	0.503	0.866
8	"Class=democrat" - "physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.537	0.503	0.894
9	"adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.524	0.503	0.866
10	"physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.524	0.503	0.887
11	"Class=democrat"	"adoption-of-the-budget-re=y"	1.488	0.531	0.865
12	"adoption-of-the-budget-re=y"	"Class=democrat"	1.488	0.531	0.913
13	"Class=democrat"	"aid-to-nicaraguan-contras=y"	1.468	0.501	0.816
14	"aid-to-nicaraguan-contras=y"	"Class=democrat"	1.468	0.501	0.901

Remarque : Les règles sont triées selon un LIFT décroissant. Il est possible de les trier différemment en les copiant dans un tableur, bien entendu il faut pour cela qu'il y ait moins de 65536 règles générées.

## Construire les règles d'association avec WEKA

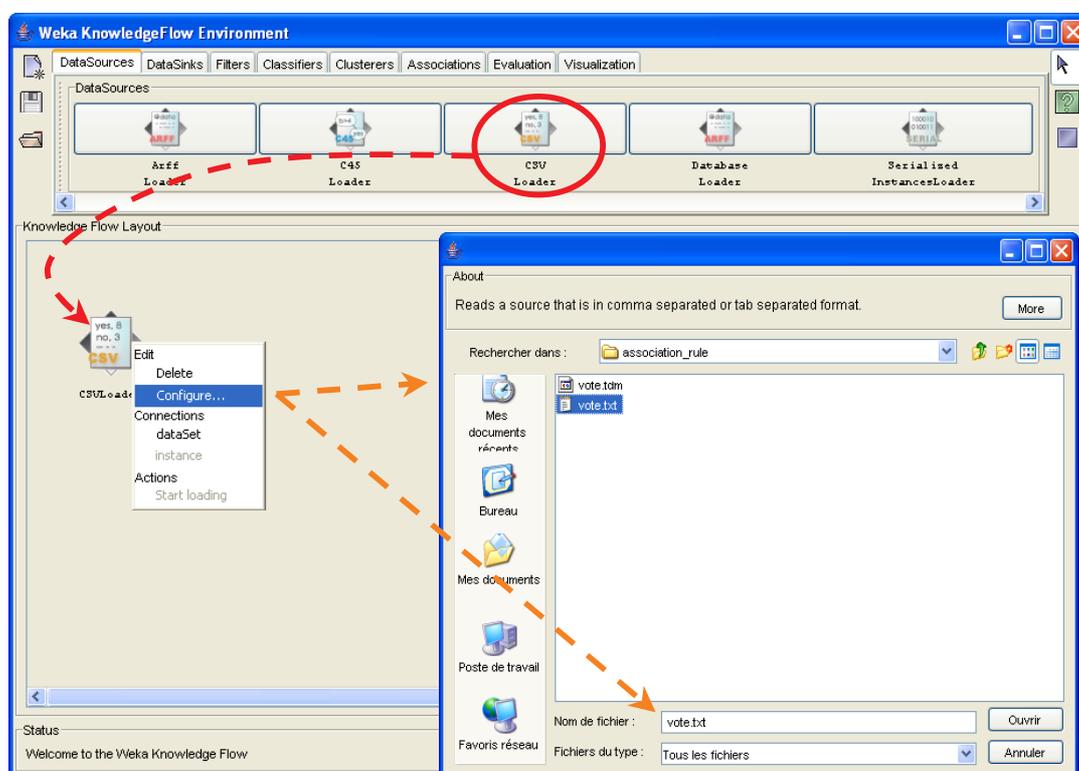
Au lancement de WEKA, un panneau permet de choisir le mode d'exécution du logiciel. Nous choisissons le mode **KNOWLEDGE FLOW**. Nous avons utilisé la version **3.5.1** dans ce didacticiel. Nous parvenons alors dans l'espace de travail dans lequel nous allons définir nos

traitements. Dans la partie haute, nous trouvons les icônes organisées dans des palettes, ils représentent les opérateurs de traitement.



## Charger les données

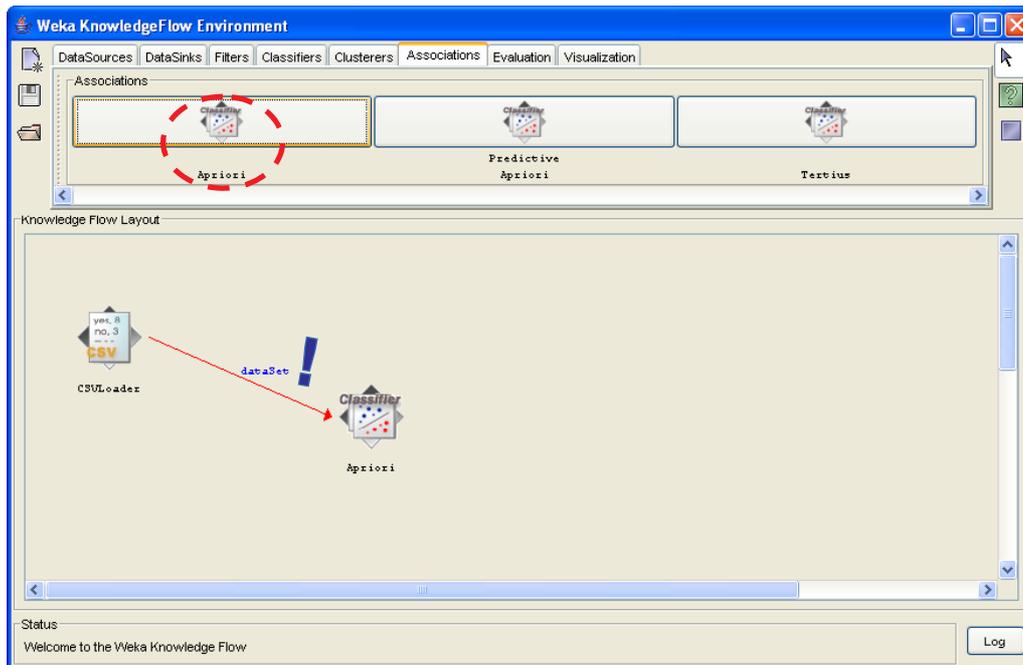
Le composant CSV LOADER (onglet DATASOURCES) permet de charger les données. Nous ajoutons donc ce composant dans notre espace de travail, nous pouvons le paramétrer à l'aide de l'option CONFIGURE de son menu contextuel. Nous chargeons le fichier VOTE.TXT.



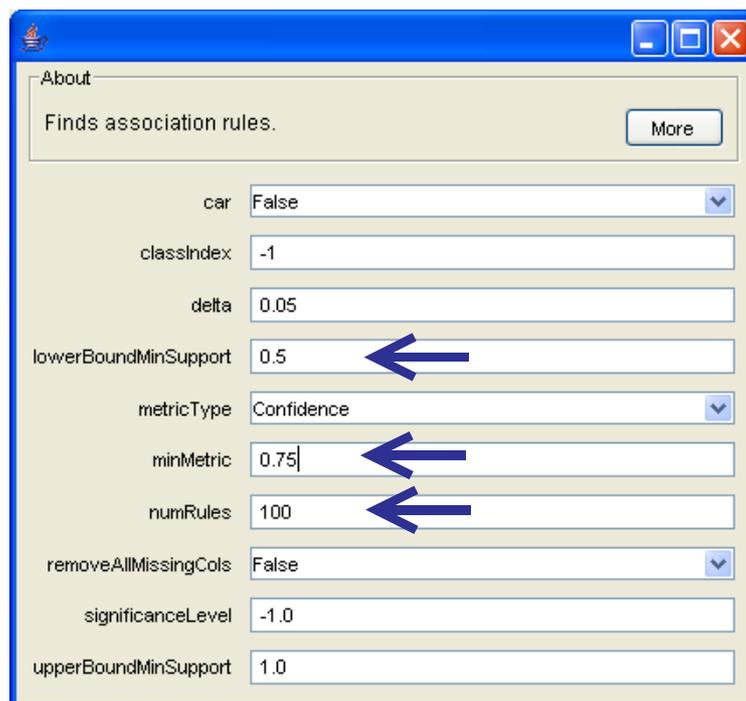
## Apprentissage des règles d'association

Par défaut, toutes les observations et toutes les variables sont sélectionnées dans WEKA. Il nous reste donc à placer le composant permettant de générer les règles. Nous trouvons le composant A PRIORI dans l'onglet ASSOCIATION.

DATASET est le type de connexion à utiliser lorsque nous relient les données à la méthode de génération de règles.

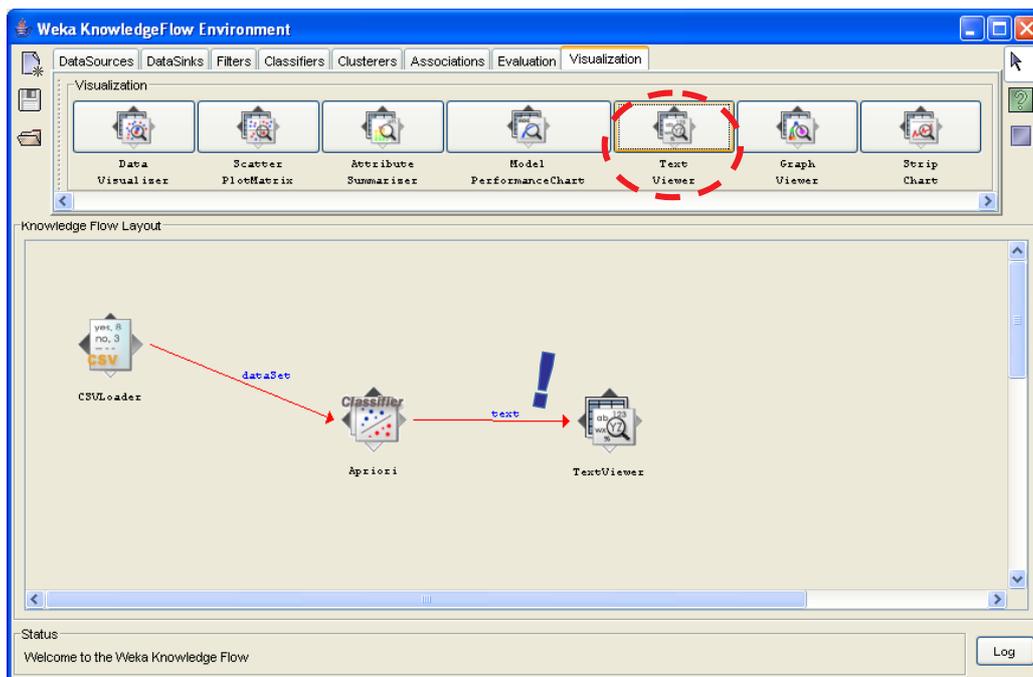


Le menu contextuel CONFIGURE permet de paramétrer la recherche des règles.

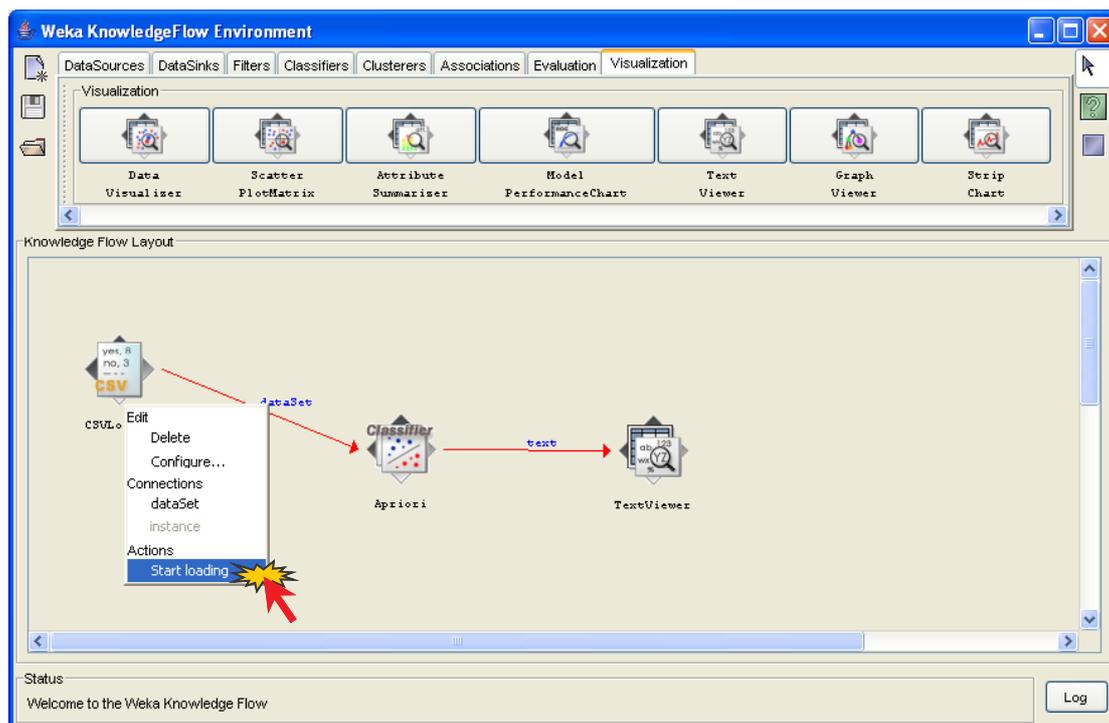


LOWERBOUNDMINSUPPORT permet de fixer le support minimal des règles ; MINMETRIC correspond à la confiance minimale si on a bien sélectionné CONFIDENCE dans la zone METRICTYPE ; enfin, NUMRULES permet de limiter le nombre maximal de règles que l'on peut générer.

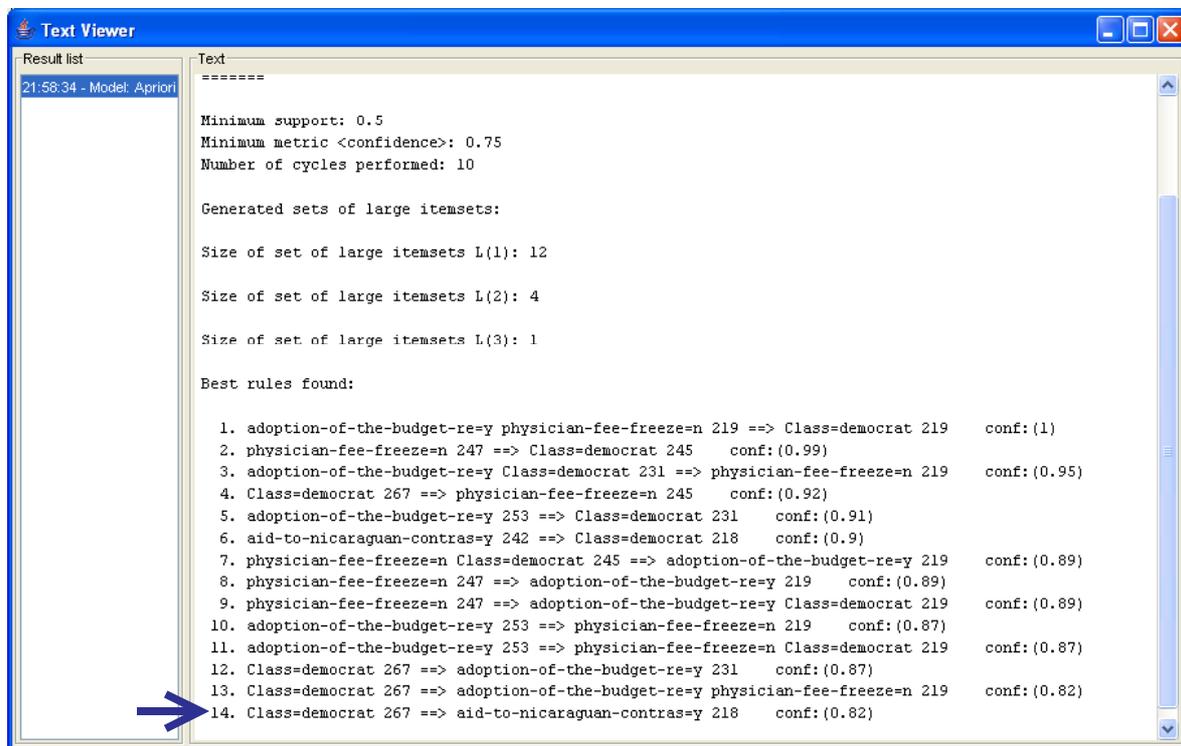
Il faut placer l'outil de visualisation des résultats pour avoir accès aux règles. Nous utilisons le composant TEXT VIEWER, nous la relierons à A PRIORI en utilisant la connexion TEXT.



Pour lancer les calculs, il nous faut revenir sur le composant de données et activer le menu START LOADING.



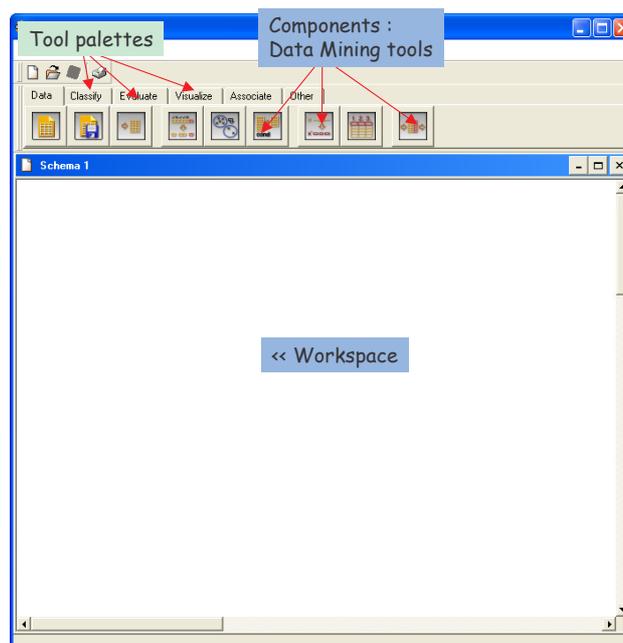
Les règles sont accessibles dans le composant TEXTVIEWER en cliquant sur le menu SHOW RESULTS.



Nous obtenons les mêmes 14 règles qu'avec TANAGRA.

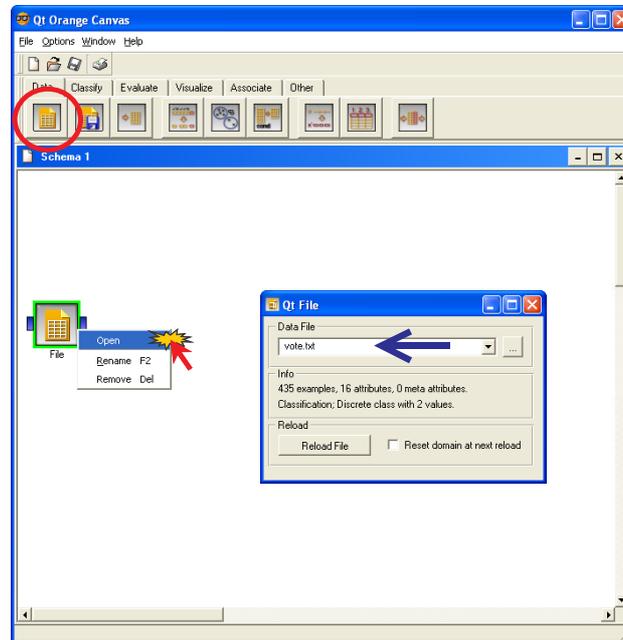
## Construire les règles d'association avec ORANGE

ORANGE propose une interface composée de deux parties distinctes : un espace pour définir les traitements ; une palette d'outils située dans la partie haute de la fenêtre principale.



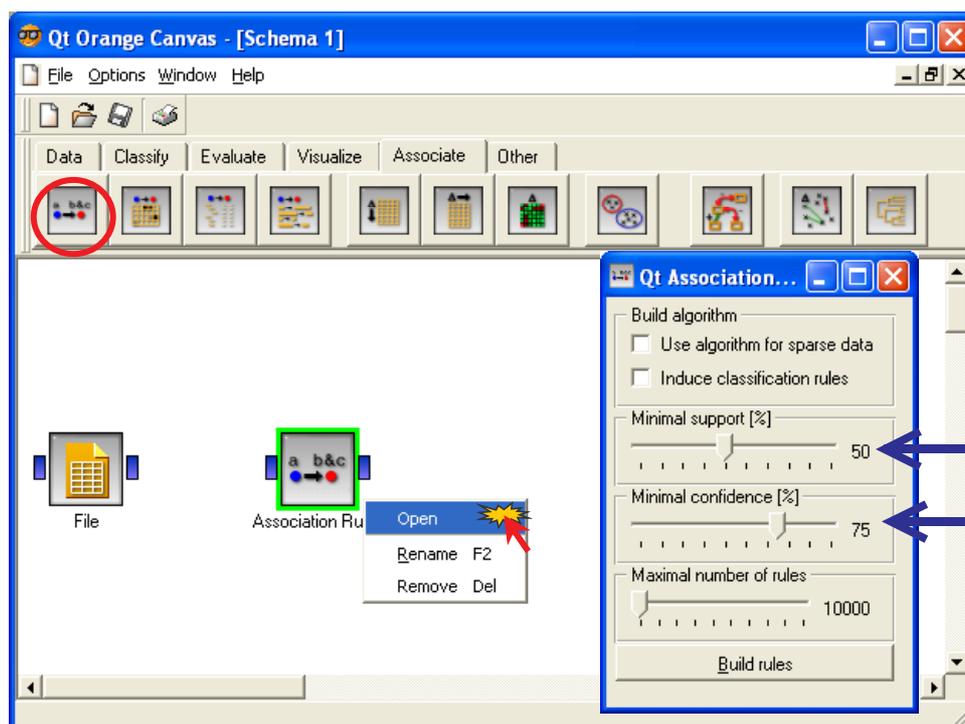
## Importer les données

ORANGE peut importer des données au format texte (séparateur tabulation). Il suffit de cliquer sur l'outil pour qu'il s'insère automatiquement dans l'espace de travail. Nous sélectionnons donc notre fichier VOTE.TXT en cliquant sur le menu contextuel OPEN.

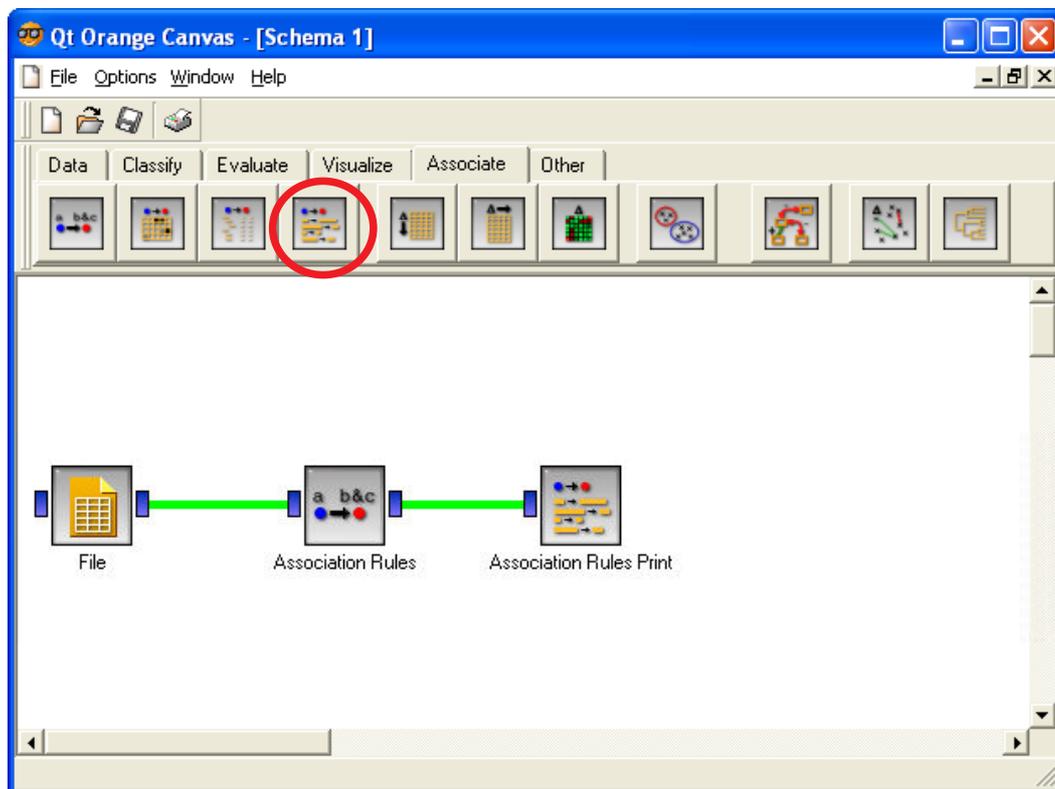


## Placer la méthode d'apprentissage

Par défaut, toutes les observations et toutes les variables sont sélectionnées dans ORANGE. Il nous faut donc placer le composant ASSOCIATION RULES de l'onglet ASSOCIATE. Nous le paramétrons en cliquant sur le menu OPEN, les paramètres sont très explicites dans ORANGE.



Les calculs sont automatiquement exécutés lorsque nous relierons le composant FILE à ASSOCIATION RULE. Pour voir les règles, nous ajoutons le composant ASSOCIATION RULE PRINT dans le diagramme.



Nous cliquons sur le menu OPEN pour voir les règles.

supp	conf	lift	rule
0.531	0.865	1.488	Class=democrat -> adoption-of-the-budget-re=y
0.531	0.913	1.488	adoption-of-the-budget-re=y -> Class=democrat
0.563	0.918	1.616	Class=democrat -> physician-fee-freeze=n
0.563	0.992	1.616	physician-fee-freeze=n -> Class=democrat

Assez curieusement, nous n'obtenons que 4 règles dans ORANGE, contre 14 avec nos deux précédents logiciels. ORANGE semble avoir une préférence pour les règles plus « simples ». Je n'ai pas trouvé les raisons de cette différence.

---

## **Conclusion**

Pour une fois, nos logiciels rivalisent de simplicité pour produire des résultats.

Ces logiciels sont largement suffisants pour la plupart des études. La situation est un peu plus difficile si nous souhaitons traiter de grandes bases de données avec des milliers d'items. Le nombre de règles générées peut devenir très important et, les performances, la possibilité de réaliser les calculs même, dépendent très fortement de la taille mémoire de la machine utilisée.