Objectif

Dans de très nombreuses analyses, il est souvent nécessaire de caractériser un groupe d'individus. Le composant GROUP CHARACTERIZATION joue ce rôle en calculant des statistiques descriptives comparatives entre le groupe qui nous intéresse et les autres groupes ou la totalité de l'échantillon. Bien que très instructif, cet outil est limité par le fait qu'il effectue essentiellement des comparaisons univariées, variable par variable, il ne permet pas de tenir compte des interactions entre les variables.

Dans ce didacticiel, nous montrons l'utilisation d'un autre composant qui permet de réaliser des caractérisations multivariées, mettant en œuvre la conjonction de plusieurs variables. La méthode, très simple, s'appuie essentiellement sur le mécanisme des règles d'association dans lequel nous fixons la modalité que nous voulons voir dans le conséquent des règles.

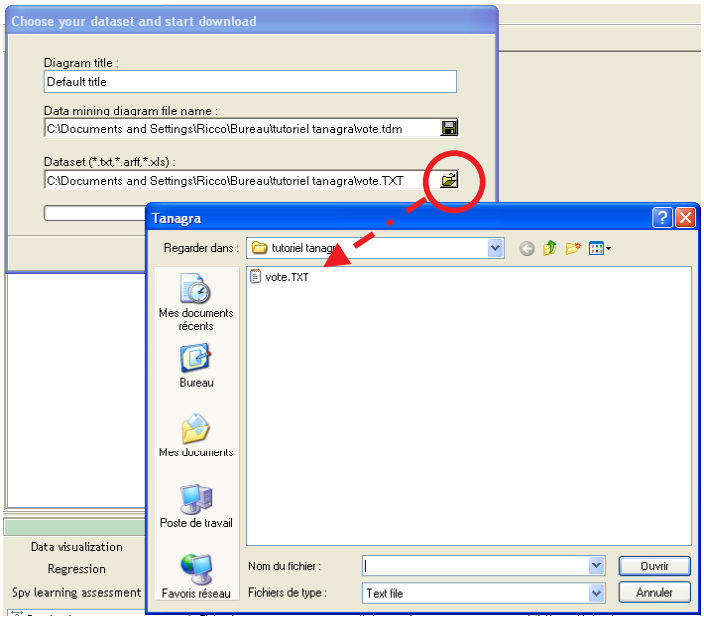
Fichier

Il s'agit du fichier des « Votes au Congrès » (VOTE.TXT). Nous voulons caractériser le comportement de vote des « républicains » sur différents thèmes qui leur ont été soumis.

Caractérisation d'un groupe d'individus

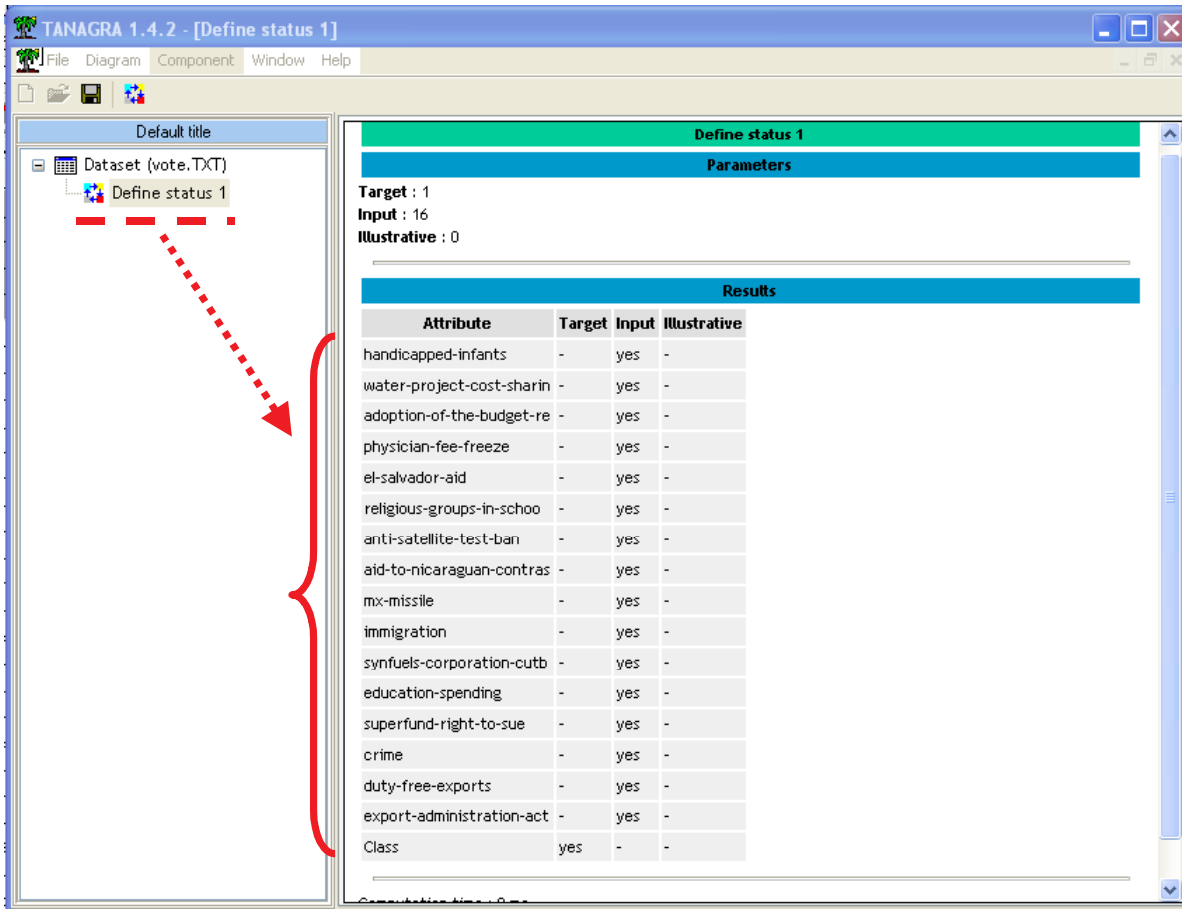
Importer les données

Première étape toujours, créez un nouveau diagramme et importez les données à l'aide du menu FILE / NEW.



Définir le rôle des variables

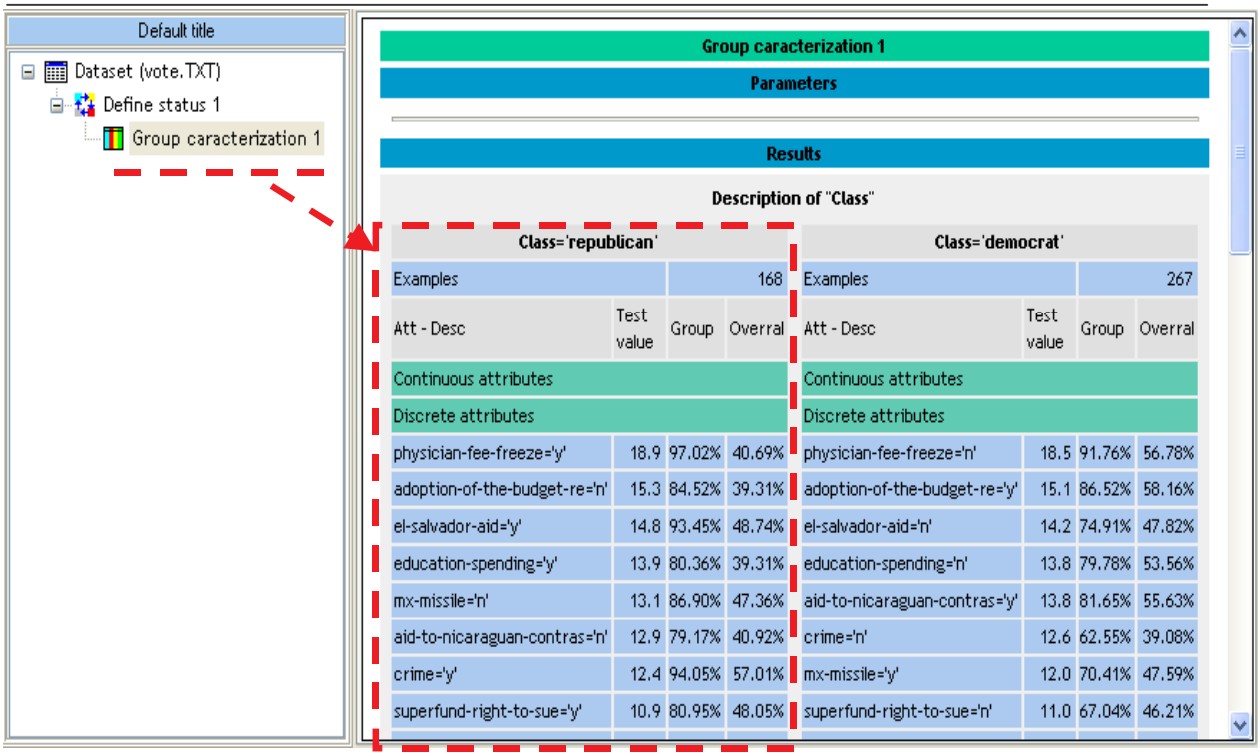
Placer par la suite le composant DEFINE STATUS, vous mettrez en TARGET la variable CLASS qui décrit l'appartenance politique, et toutes les autres variables, le comportement lors de chaque vote, en INPUT.



Caractérisation univariée

Dans un premier temps, nous allons réutiliser le composant GROUP CHARARCTERISATION. Il réalise un test de comparaison de proportions sur chaque variable puis les classe selon l'importance de la différence matérialisée par la valeur test, qui évolue de la même manière que la p-value du test de comparaison.

Bien que nous voulions étudier plus particulièrement le groupe des « républicains », le composant a été défini de manière à ce qu'il réalise la caractérisation sur chaque modalité de la variable définie comme TARGET. C'est pour cette raison que nous trouvons deux colonnes de résultats, une pour chaque modalité de la variable CLASS. Seule la lecture de la première colonne nous intéresse réellement dans ce didacticiel.



Class='republican'

Examples			168
Att - Desc	Test value	Group	Overral
Continuous attributes			
Discrete attributes			
physician-fee-freeze='y'	18.9	97.02%	40.69%
adoption-of-the-budget-re='n'	15.3	84.52%	39.31%
el-salvador-aid='y'	14.8	93.45%	48.74%
education-spending='y'	13.9	80.36%	39.31%
mx-missile='n'	13.1	86.90%	47.36%
aid-to-nicaraguan-contras='n'	12.9	79.17%	40.92%
crime='y'	12.4	94.05%	57.01%
superfund-right-to-sue='y'	10.9	80.95%	48.05%

Class='democrat'

Examples			267
Att - Desc	Test value	Group	Overral
Continuous attributes			
Discrete attributes			
physician-fee-freeze='n'	18.5	91.76%	56.78%
adoption-of-the-budget-re='y'	15.1	86.52%	58.16%
el-salvador-aid='n'	14.2	74.91%	47.82%
education-spending='n'	13.8	79.78%	53.56%
aid-to-nicaraguan-contras='y'	13.8	81.65%	55.63%
crime='n'	12.6	62.55%	39.08%
mx-missile='y'	12.0	70.41%	47.59%
superfund-right-to-sue='n'	11.0	67.04%	46.21%

Le fichier contient 435 parlementaires, 168 d'entre eux (# 38%) sont estampillés « républicains ».

Prenons la première variable qui semble caractériser le mieux ce groupe : 40.69% des députés (177 députés) ont répondu OUI (« Y ») à la question « PHYSICIAN-FEE-FREEZE » ; parmi les républicains, ce pourcentage monte à 97.02%, c'est-à-dire, 97.02% x 168 = 163 députés. Le fait d'être républicain a déterminé le comportement de vote sur cette question, ce comportement caractérise donc l'appartenance politique.

Il est possible de déduire deux probabilités conditionnelles à partir de ces résultats, la première calculée directement par TANAGRA

$$P(physician = y / republican) = 97.02\%$$

la seconde que l'on peut facilement déduire

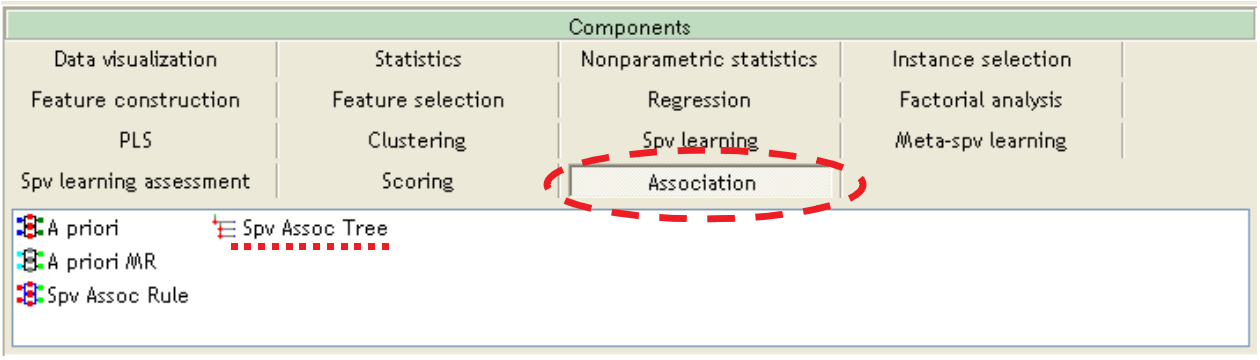
$$P(republican / physician = y) = \frac{P(republican \cap physician = y)}{P(physician = y)} = \frac{163}{177} = 92.1\%$$

On peut lire les résultats de la même manière pour chaque variable INPUT.

Caractérisation multivariée

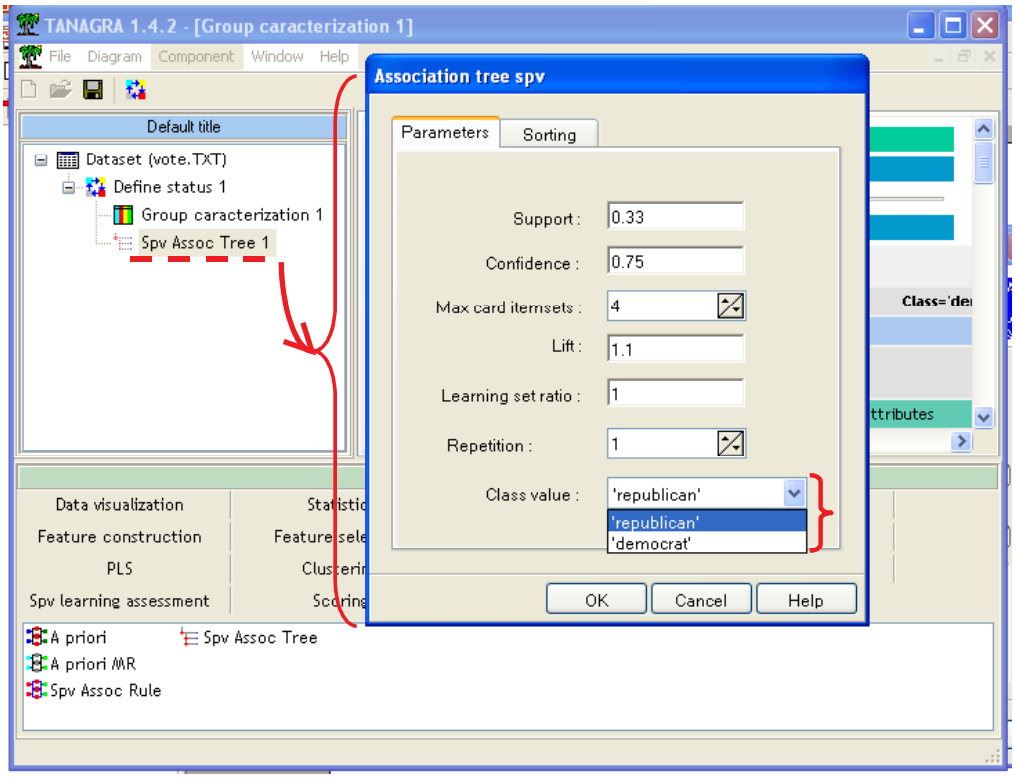
Si cette approche permet déjà de bien comprendre les caractéristiques d'un groupe d'individu, elle reste peu satisfaisante dans la mesure où elle évalue le rôle de chaque variable individuellement, sans tenir compte des éventuelles interactions entre deux ou

plusieurs variables qui permettrait de circonscrire avec plus de précision le groupe que l'on veut étudier.



Le composant SPV ASSOC TREE de la palette ASSOCIATION RULE permet de réaliser cette opération. En réalité il s'agit d'un cas particulier des règles d'association sauf que l'on définit à l'avance l'item que l'on veut voir apparaître dans le conséquent de la règle. Le composant SPV ASSOC RULE effectue exactement les mêmes calculs, il s'appuie sur une autre implémentation et produit un grand nombre d'indicateurs, peu en rapport avec notre propos.

Plaçons ce composant dans notre diagramme et voyons quels en sont les principaux paramètres.



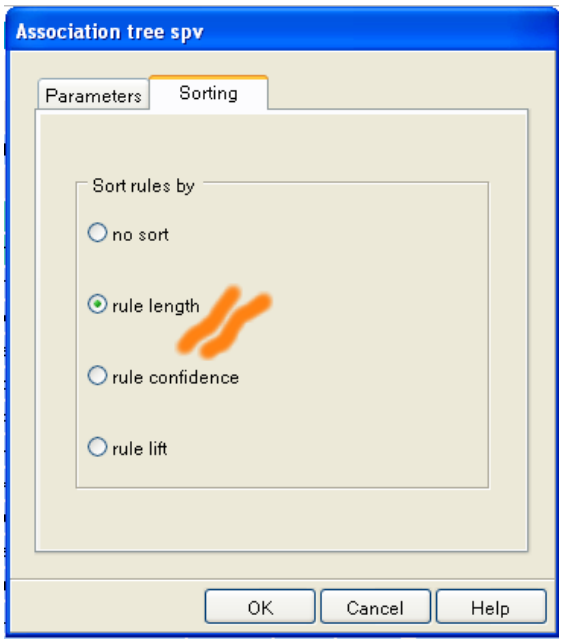
Les quatre premiers paramètres (SUPPORT, CONFIDENCE, MAX ITEMSET, LIFT) sont standards dans la construction des règles d'association. Ils permettent de contrôler le nombre de règles extraites.

LEARNING SET RATIO permet de subdiviser les données en « échantillon apprentissage » et « échantillon test ». Son rôle n'est pas déterminant ici.

REPETITION est un paramètre expérimental lié à nos travaux de recherche. Le mieux est de l'ignorer.

Enfin, dernier paramètre, très important dans le contexte de notre étude, CLASS VALUE définit la modalité de la variable TARGET que nous voulons étudier, ce qui délimite le groupe d'individu que nous voulons caractériser. Dans notre cas, il s'agit bien du groupe des républicains.

Dans le second onglet, il est possible de définir le tri des règles, cela peut s'avérer primordial si le nombre de règles est très important. Prenons le plus simple, trier les règles selon leur longueur pour obtenir des règles de complexité croissante.



Les résultats sont alors affichés dans la fenêtre adéquate après exécution du composant.

Results

Rules

"Class" is "'republican'" -- IF ...

N°	Antecedent	Length	Support	Confidence	Lift
1	physician-fee-freeze='y'	1	0.375 (0.00)	0.921 (0.00)	2.384 (0.00)
2	el-salvador-aid='y' - mx-missile='n'	2	0.333 (0.00)	0.788 (0.00)	2.040 (0.00)
3	el-salvador-aid='y' - physician-fee-freeze='y'	2	0.359 (0.00)	0.929 (0.00)	2.404 (0.00)
4	mx-missile='n' - physician-fee-freeze='y'	2	0.333 (0.00)	0.942 (0.00)	2.438 (0.00)
5	crime='y' - physician-fee-freeze='y'	2	0.356 (0.00)	0.923 (0.00)	2.389 (0.00)
6	crime='y' - el-salvador-aid='y'	2	0.343 (0.00)	0.768 (0.00)	1.989 (0.00)
7	religious-groups-in-schoo='y' - physician-fee-freeze='y'	2	0.338 (0.00)	0.919 (0.00)	2.379 (0.00)
8	religious-groups-in-schoo='y' - el-salvador-aid='y' - physician-fee-freeze='y'	3	0.331 (0.00)	0.923 (0.00)	2.390 (0.00)
9	crime='y' - el-salvador-aid='y' - physician-fee-freeze='y'	3	0.340 (0.00)	0.925 (0.00)	2.395 (0.00)
10	el-salvador-aid='y' - mx-missile='n' - physician-fee-freeze='y'	3	0.331 (0.00)	0.941 (0.00)	2.437 (0.00)

Nous retrouvons en partie les résultats du composant précédent, mais nous disposons également d'informations complémentaires.

La première règle est la même, avec un point de vue un peut différent :

- le SUPPORT de la règle nous indique $P(physician = y \cap republican) = \frac{163}{435} = 37.5\%$
- la CONFIDENCE est la probabilité conditionnelle $P(republican / physician = y) = 92.1\%$
- le LIFT indique qu'on a 2.384 de chances d'être un républicain lorsqu'on a voté « Y » à la question « PHYSICIAN-FEE-FREEZE », il s'agit du rapport de probabilité $\frac{P(republican / physician = y)}{P(republican)} = \frac{92.1\%}{38.6\%} = 2.384$

Nous aurions pu calculer ces différents ratios à partir du composant GROUP CHARACTERIZATION, le principal intérêt de ce nouveau composant est qu'il possible de prendre en compte le rôle conjoint de plusieurs variables.

Prenons l'exemple de la règle n°4. La variable MX-MISSILE a été rajoutée dans la règle. Elle nous permet d'obtenir une règle plus précise, matérialisée par un accroissement de la confiance : $P(republican / physician = y \cap mx - missile = n) = 94.2\%$

La règle est plus précise, bien que sur cet exemple, le gain reste quand même assez minime.

Modifier les paramètres de l'analyse

Nous comprenons aisément que ce type d'analyse, multivarié, peut se révéler extrêmement puissant. En revanche, son utilisation requiert le réglage délicat des paramètres usuels utilisés dans les règles d'association. Si nous sommes trop restrictifs, nous risquons de passer à côté de règles « intéressantes », en revanche, si nous sommes trop permissifs, nous serons vite noyés sous un nombre incalculable de règles.

Essayons par exemple de diminuer notre SUPPORT MIN à 10%, la CONFIANCE MIN à 90%, et trions les cette fois-ci selon le LIFT. Nous obtenons les résultats suivants.

Spv Assoc Tree 1

Parameters

A-Priori parameters	
Support min	0.10
Confidence min	0.90
Max rule length	4
Lift filtering	1.10
Learning set ratio	1.00
Value to predict	'republican'
Sort criteria	rule lift

Results

Rules

"Class" is '"republican"' -- IF ...

N°	Antecedent	Length	Support	Confidence	Lift
1	immigration='y' - physician-fee-freeze='y' - adoption-of-the-budget-re='n'	3	0.168 (0.00)	1.000 (0.00)	2.589 (0.00)
2	synfuels-corporation-cutb='n' - immigration='y' - physician-fee-freeze='y'	3	0.175 (0.00)	1.000 (0.00)	2.589 (0.00)
3	export-administration-act='y' - water-project-cost-sharin='n' - physician-fee-freeze='y'	3	0.101 (0.00)	1.000 (0.00)	2.589 (0.00)
4	immigration='y' - water-project-cost-sharin='n' - physician-fee-freeze='y'	3	0.103 (0.00)	1.000 (0.00)	2.589 (0.00)
5	synfuels-corporation-cutb='n' - duty-free-exports='n' - physician-fee-freeze='y'	3	0.267 (0.00)	0.991 (0.00)	2.567 (0.00)
6	synfuels-corporation-cutb='n' - physician-fee-freeze='y' - adoption-of-the-budget-re='n'	3	0.267 (0.00)	0.991 (0.00)	2.567 (0.00)
7	synfuels-corporation-cutb='n' - physician-fee-freeze='y' - education-spending='y'	3	0.262 (0.00)	0.991 (0.00)	2.567 (0.00)

270 règles ont été produites ! Les 4 premières s'avèrent exactes, sans contre-exemples -- avec une confiance de 100% -- elles permettent de déterminer avec certitude l'appartenance au groupe des républicains. Notons néanmoins que ces règles couvrent peu d'observations.

Cette méthode est donc particulièrement performante dès lors que l'on veut caractériser une sous-population. En revanche, toute médaille a son revers, il faut définir judicieusement ses paramètres pour obtenir des résultats convaincants.