Objectif

Montrer la démarche à suivre pour un K-Means sur variables qualitatives. Les classes obtenues peuvent être « validées » en les comparant avec un regroupement a priori « naturel ».

Fichier

Le fichier des votes au congrès américain (UCI). Des parlementaires votent sur différents sujets, chaque vote est recensé (« oui », « non » ou « on ne sait pas ce qu'il a voté»). La variable naturelle de regroupement est l'appartenance politique du député (« démocrate » ou « républicain »). L'objectif dans ce fichier est de montrer (1) qu'il est possible de regrouper les députés selon leur comportement lors des votes, (2) que ce regroupement peut être rapproché avec leur appartenance politique.

Etapes de l'expérimentation

1. Charger les données, il comporte 435 individus et 17 variables, dont la variable « class » qui représente l'appartenance politique.
2. Comme il n'existe pas de méthodes dans TANAGRA pour effectuer une classification sur variables discrètes, nous allons dans un premier temps construire les axes factoriels avec une analyse des correspondances multiples. Avec un « Define Status », définir le statut de toutes les variables, sauf « class », en input, puis brancher à la suite une ACM. Laisser les paramètres par défaut.
3. On se rend compte que les cinq premiers axes factoriels résument 50% de l'information disponible. Nous allons les utiliser pour lancer un K-Means sur ces axes.
4. Placer un « Define Status » dans la chaîne, puis sélectionnez les 5 axes en « Input ».
5. Placer un composant « K-Means », avec les paramètres suivants : Nombre de clusters = 2 ; Nombre maximum d'itérations = 10 ; Nombre d'essais = 5 ; Normalisation des données pour le calcul des distances = none (ne pas pondérer car la variance d'un axe représente également l'inertie qu'il représente !!!) ; Calcul des centres lors des itérations = McQueen (à chaque passage).
6. L'opération propose deux clusters, l'une de #240 individus, l'autre de #135 (le résultat dépend des initialisations qui sont aléatoires), 40% de l'inertie est expliquée.
7. Reste maintenant à caractériser ces groupes. Nous plaçons à la suite du K-Means, un nouveau « Define Status », avec comme « Target », la classe produite « Cluster_Kmeans_1 », et comme « Input », toutes les variables initiales de la base, y compris l'appartenance politique qui joue le rôle de variable illustrative dans ce cas. Puis placez le composant « Group Characterization », il permet de comparer les statistiques descriptives calculées sur tout l'échantillon et dans chaque groupe défini par « Target ». Les résultats sont édifiants, pour caractériser le premier groupe, la modalité « democrat » de « class » apparaît en très bonne place, avec une valeur test très élevée (cf. Lebart et al. pour une discussion sur les valeurs tests, pp. 181-184), on constate que dans tout l'échantillon, il y a 61% de démocrates, dans ce groupe ils

sont 95% ; il en est de même pour le second groupe, il y a 39% de républicains en tout, ils sont 79% dans ce groupe.

Description of "Cluster_KMeans_1"							
Cluster_KMeans_1=c_kmeans_1				**Cluster_KMeans_1=c_kmeans_2**			
Examples			239	Examples			196
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes				Continuous attributes			
Discrete attributes				Discrete attributes			
el-salvador-aid='n'	17.7	86.19%	47.82%	el-salvador-aid='y'	18	96.43%	48.74%
aid-to-nicaraguan-contras='y'	17.6	93.72%	55.63%	aid-to-nicaraguan-contras='n'	17.2	85.71%	40.92%
physician-fee-freeze='n'	16.4	92.05%	56.78%	physician-fee-freeze='y'	16.9	84.69%	40.69%
Class='democrat'	**15.7**	**94.56%**	**61.38%**	mx-missile='n'	16	89.80%	47.36%
adoption-of-the-budget-re='y'	15.4	91.21%	58.16%	adoption-of-the-budget-re='n'	15.8	80.10%	39.31%
mx-missile='y'	14.7	79.50%	47.59%	**Class='republican'**	**15.7**	**79.08%**	**38.62%**
crime='n'	14.3	69.46%	39.08%	education-spending='y'	14.6	77.04%	39.31%

8. Une autre manière de visualiser ces résultats est de construire un tableau croisé entre le regroupement produit par le clustering et la supposée classe naturelle. Pour ce faire, placez un n-ième « Define Status », dans laquelle vous définissez « class » en « target » et « Cluster_Kmeans_1 » en « input ». Puis, insérer à la suite un composant « Cross-Tabulation » de la palette « Descriptive Stats ». Les résultats sont cohérents avec ce que l'on a vu plus haut.

Row (Y)	Column (X)	Statistical indicator		Cross-tab			
		Stat	**Value**		**c_kmeans_1**	**c_kmeans_2**	**Sum**
		Tschuprow's t	0.752565	**'republican'**	13	155	168
		Cramer's v	0.752565	**'democrat'**	226	41	267
		Phi²	0.566354	**Sum**	239	196	435
		Chi²	246.364086				
Class	Cluster_KMeans_1	Pr(Chi²)	0				

9. Voici le diagramme de traitements associé.