Subject

How to perform a K-MEANS clustering on discrete attributes? Validate clusters with external criteria, i.e. to compare our clusters with preexistent classes.

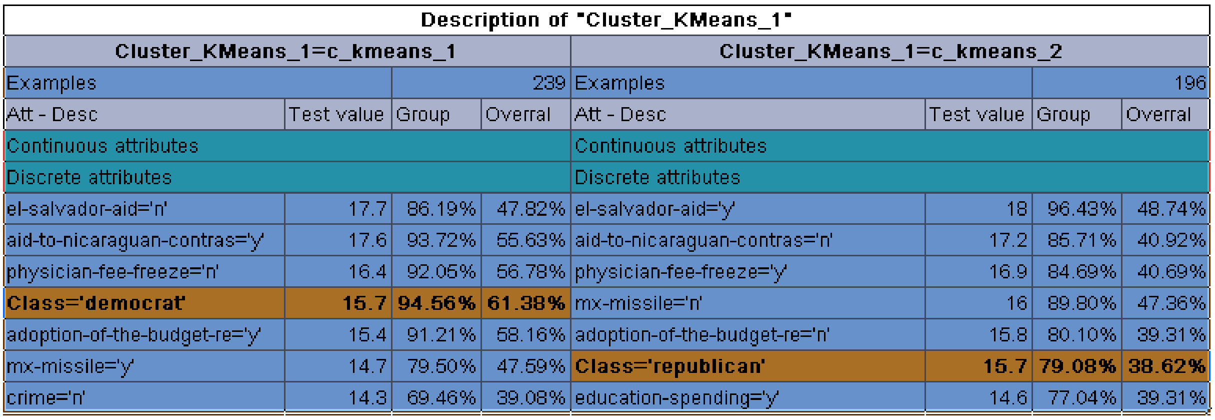
Dataset

The famous US CONGRESS VOTE (UCI): pre-existing class attribute is political affiliation of congress members; descriptors are their vote behavior on various subjects.

We want to build homogenous groups (clusters) of members from their behavior and compare these clusters with their political affiliation.

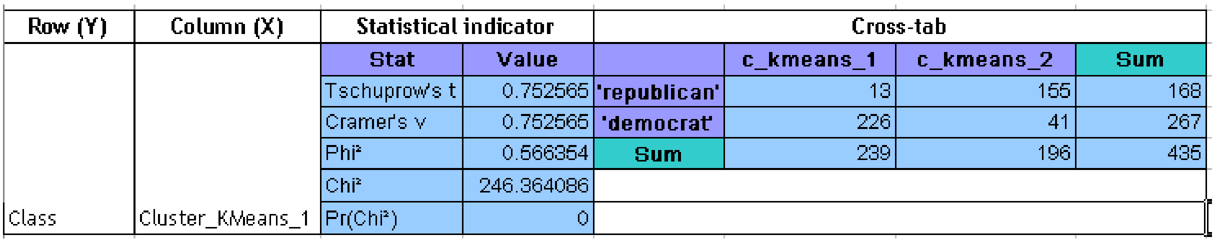
Experimentation steps

1. Load dataset, there is 435 examples and 17 attributes; "class" is the political affiliation.
2. There is not clustering method into TANAGRA that handles directly discrete attributes. We perform in the first time a feature construction using factorial analysis (Correspondence multiple analysis) and use them as new attributes for K-MEANS.
3. Add as "Define Status" component in the diagram and select all attributes except "class" as INPUT. Add an ACM component and use default parameters.
4. The 5 first factorial axis (dimensions) summarize 50% of available information. It indicates the quality of representation of points in theses 5 dimensions. We use these axis as descriptors for K-MEANS.
5. Add a "Define Status" component and set as INPUT the factorial axis.
6. Add a K-MEANS component end set the following parameters: Number of clusters = 2; Max number of iteration = 10; Trials = 5; *Distance Normalization = None (Variance of an axis is the "weight" of this axis, we do not standardize the data)*; Average computation = Mc Queen; Seed random number generation = Standard.
7. We have two clusters: #240 examples for the first, and #135 examples for the second (the exact clusters size relies on the random number generator used and your computer). Explained inertia ratio is 40%.
8. How to characterize these clusters? Add an another "Define Status" in the diagram and set as TARGET the cluster attribute "Cluster_Kmeans_1", set as INPUT all other native attributes including the political affiliation (Class). Don't select factorial axis.
9. Add a "Group characterization" component, this component performs comparative descriptive statistics between the whole dataset and examples in the clusters. A ratio called "Test value" shows the strength of the differences.
10. Above all, we note that clusters strongly correspond to the political affiliation: there is 61% of democrats in the congress, they are 95% in the first cluster; in the second cluster, there is a majority (79%) of republican.



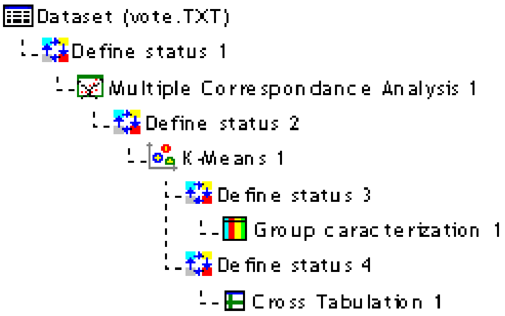
Description of "Cluster_KMeans_1"							
Cluster_KMeans_1=c_kmeans_1				Cluster_KMeans_1=c_kmeans_2			
Examples		239		Examples		196	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes				Continuous attributes			
Discrete attributes				Discrete attributes			
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%

11. There is another way to compare clusters and political affiliation. Add a "Define Status" and set as TARGET "Class", set as INPUT "Cluster_Kmeans_1". Add a "Cross-tabulation" component, we have a result that is coherent with the previous one.



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				

12. Here the data mining diagram.



```
Dataset (vote.TXT)
 └─Define status 1
    └─Multiple Correspondance Analysis 1
       └─Define status 2
          └─K-Means 1
             ├─Define status 3
             │  └─Group caracterization 1
             └─Define status 4
                └─Cross Tabulation 1
```