Objectif

Clustering avec la CAH mixte (Classification ascendante hiérarchique mixte).

La CAH est une méthode de classification qui permet de mettre en évidence un regroupement « naturel » d'un ensemble d'individus décrits par des caractéristiques (les variables). Elle propose une série de partitions emboîtées représentées sous forme d'arbres appelés dendogrammes. L'algorithme procède par agrégations successives, partant de la partition la plus fragmentaire, un individu est égal à une classe, jusqu'à la partition triviale, le regroupement de tous les individus dans une et une seule classe.

Le principal avantage de la CAH par rapport aux autres méthodes de classification réside dans cette représentation sous forme d'arbre qui met en évidence une information supplémentaire : l'augmentation de la dispersion dans un groupe produit par une agrégation. L'utilisateur peut dès lors avoir une idée du nombre adéquat de classes en choisissant la partition correspondant au saut le plus élevé dans l'augmentation de la dispersion au sein des classes.

Le principal inconvénient de la CAH est qu'elle nécessite le calcul des distances entre individus pris deux à deux. Ce qui est très rapidement prohibitif dès que la taille du fichier excède le millier d'individus.

La CAH mixte permet de lever cette limitation tout en bénéficiant des avantages de la CAH classique. Partant du constat que l'on veut très souvent produire un nombre limité de classes, on délègue la création de la partie basse du dendogramme à des méthodes de ré-allocations nettement moins gourmands en ressources machines (K-Means, SOM, …).

L'algorithme comporte donc deux étapes :

- La création d'une partition grossière avec une méthode de ré-allocation, avec un nombre assez élevé de classes, l'expérience montrer que des valeurs autour de 20 sont amplement suffisantes ;
- CAH à partir des noyaux constitués par les groupes mis en avant par le premier algorithme.

Il est à noter que dans TANAGRA, le premier niveau de partitionnement peut être produit par n'importe quel algorithme, il peut même être introduit par l'utilisateur via une variable discrète.

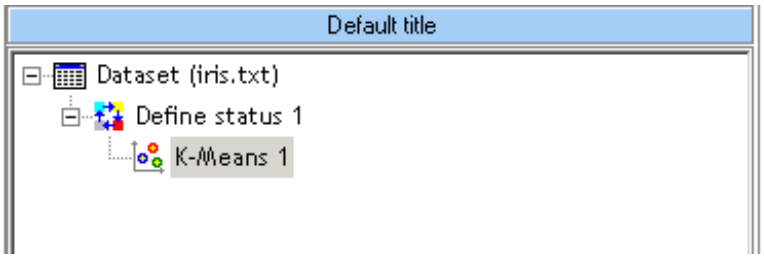
Enfin, plus que l'arbre lui-même, c'est bien le saut entre chaque nœud représentant un regroupement qui intéresse l'utilisateur. Afin de simplifier les affichages, TANAGRA a choisi de n'afficher que la hauteur des sauts entre les nœuds en les listant dans un tableau, associés au nombre de classes produits si l'on décidait de couper l'arbre à ce stade.

Fichier

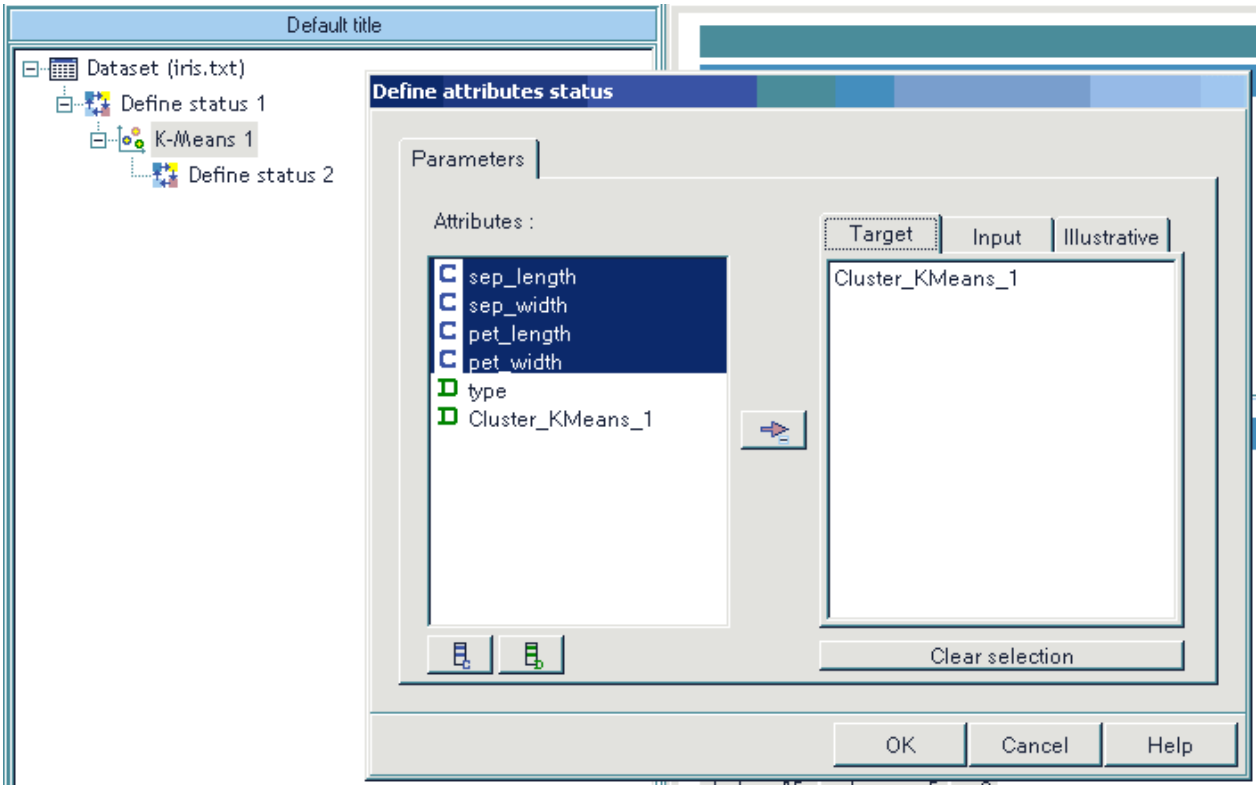
Le fameux fichier des IRIS recensant des fleurs (FISHER 1936), le principal intérêt est que l'on connaît à l'avance les résultats à obtenir.

Construire une CAH

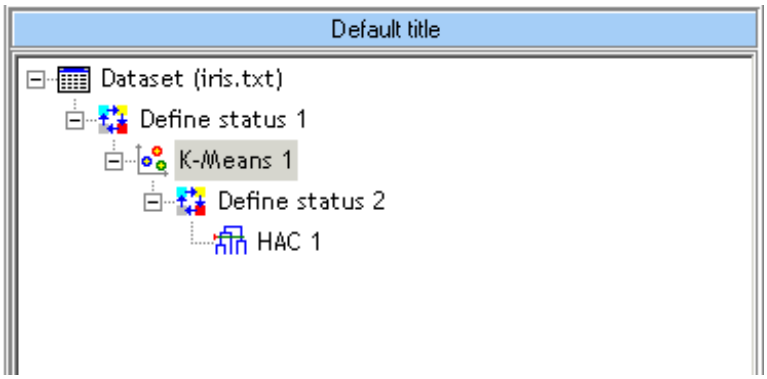
1. Charger le fichier IRIS_HAC.BDM
2. Insérer le composant « Define Status », attribuer aux variables continues le statut INPUT.
3. Insérer dans le diagramme un composant K-MEANS et fixer un nombre de classes à produire égal à 20, laisser les autres paramètres à leur valeur par défaut. ***Vous devez exécuter la filière à ce stade.*** Votre diagramme doit avoir l'aspect suivant :



4. Introduire à la suite du diagramme un composant « Define Status », mettez les attributs continus en INPUT et introduisez en TARGET l'attribut qui a été produit par le K-MEANS



5. Insérer alors le composant HAC dans la filière puis lancer les calculs.



6. Les résultats apparaissent dans la fenêtre de droite, cette méthode se démarque des autres algorithmes de classification en proposant un tableau recensant la hauteur du saut entre deux regroupements. Il surligne automatiquement le saut le plus élevé.

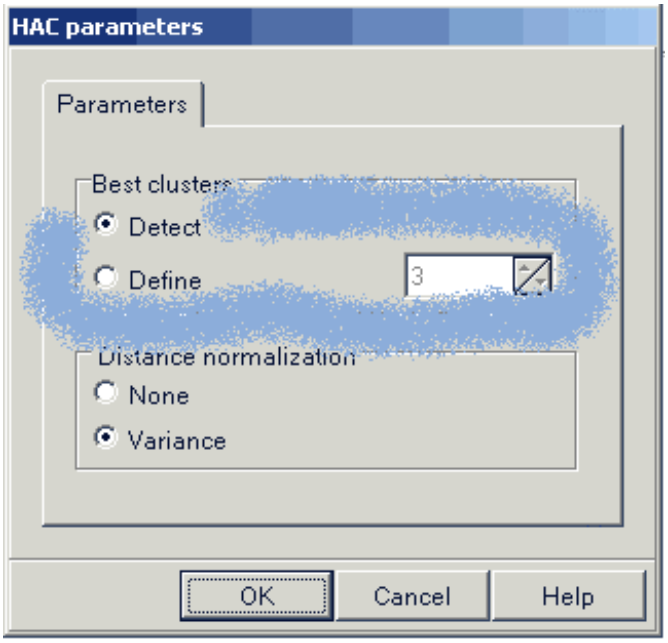
Best cluster selection

Clusters	BSS ratio	Gap
2	0.6271	2.0102
3	0.7517	0.3247
4	0.7951	0.0026
5	0.8378	0.0897
6	0.8581	0.0146
7	0.8748	0.0011
8	0.8911	0.0060
9	0.9060	0.0045
10	0.9198	0.0272
11	0.9268	0.0067
12	0.9321	0.0023
13	0.9368	0.0019
14	0.9411	0.0045
15	0.9442	0.0013
16	0.9470	0.0001
17	0.9498	0.0006
18	0.9524	0.0040
19	0.9541	0.0022

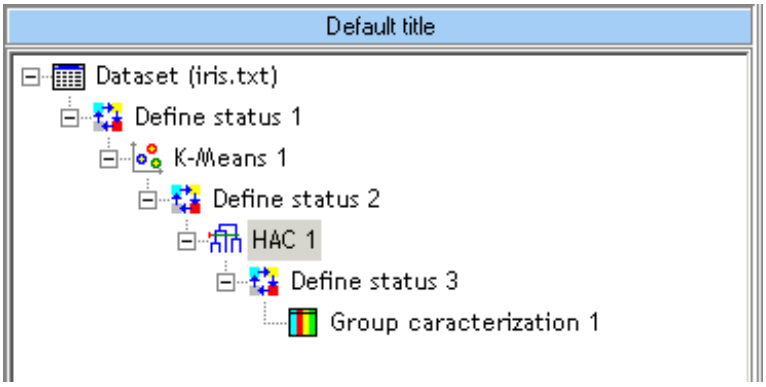
7. Le saut le plus intéressant (GAP) qui a été mis en évidence correspond à un partitionnement en 3 classes (CLUSTERS = 3). Deux remarques :
   - La classification en deux classes a été ignorée alors qu'elle correspond au saut le plus élevé, cela est naturel car à l'usage on constate que cette subdivision en deux classes proposera toujours le saut le plus élevé, ce qui est normal dans le sens où il s'agit là de la première subdivision possible de l'ensemble de données, la dispersion dans les deux groupes produits chute mécaniquement sans que cela corresponde forcément, dans la plupart des cas, à un partitionnement intéressant ;

- Dans le cas présent, le partitionnement en trois classes se démarque fortement des autres. Il se peut que dans certains cas, plusieurs « hauteur de saut » soient proches, mettant en concurrence plusieurs solutions possibles.

8. L'utilisateur a la possibilité de fixer lui-même le nombre adéquat de partitions en paramétrant l'algorithme.



9. Enfin, il faut caractériser la classification produite. Sur ce fichier, *qui est un cas particulier*, on dispose des classes « réelles » d'appartenance des individus, il est possible de les utiliser pour décrire les classes produites par l'algorithme. Pour ce faire, il faut insérer un composant « Define status » dans la filière, mettre en TARGET « Cluster_HAC_1 » produit par la CAH, et en INPUT l'attribut « Type » fournit avec le fichier. Brancher alors à la suite le composant « Group Characterization » (Descriptives stats).



10. Les résultats montrent bien que les classes produites correspondent à des types particuliers d'IRIS (*voir page suivante…*).

Group caracterization 1

Parameters

Results

Description of "Cluster_HAC_1"

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples			50	Examples			37	Examples			63
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes				Continuous attributes				Continuous attributes			
Discrete attributes				Discrete attributes				Discrete attributes			
type=Iris-setosa	12.2	100.00%	33.33%	type=Iris-virginica	7.9	86.49%	33.33%	type=Iris-versicolor	8.4	71.43%	33.33%
[illegible]				[illegible]				[illegible]			