1 Classification

Définition 1 *La classification réside dans le fait de construire des classes qui représentent des individus statistique possédant des traits des caractères communs.Il existe deux types de classifications*

Hiérarchique C.A.H.

Méthode de partitionnement (k-means)

1.1 Principe de la C.A.H.

Ce type de classification se fait sur un tableau de type A.C.P. (n individus et p variables)

L'objectif de la classification A.M est :

la construction d'un arbre hérarchique (ascendent,déscendent) qui permet de mettre évidence les liens hérarchiques entre les individus ou groupes d'individus.

La détection d'un nombre de classe partitionnement comme pour l'A.C.P. ;A.F.C. et l'A.C.M. nous nous avons besoin de définir une distance qui permet de calculer la ressemblance entre individus.

Ici,la distance naturelle est la distance euclidienne du fait que les individus

admettent des coordonnées quantitatives.

Il existe aussi d'autre type de distance que nous allons voir dans la suite (indice de similarité)

*Dans le cas d'étude de la distance entre*2 *groupes d'individus on peut utilisé plusieurs indices de similarité comme la distance min d,euclidienne ,max d, ou bien la distance de ward....*

Maintenant rest à décider le nombre de classes qui me permet de coupé l'arbre hérarchique pour cela il faut définir le niveau ou seuil de coupure.

Exemple 1

	v_1	v_2	v_3
A	0	1	2
B	2	1	0
C	1	2	0
D	0	1	1

La premiere étape :

$d_{AB} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$

$d_{AB} = \sqrt{(-2)^2 + (0)^2 + (2)^2} = 2$

$d_{AC} = \sqrt{(-1)^2 + (-1)^2 + (2)^2} = \sqrt{6}$

$d_{BC} = \sqrt{(1)^2 + (-1)^2} = \sqrt{2}$

$d_{AD} = \sqrt{(0)^2 + (0)^2 + (1)^2} = 1$

$d_{BD} = \sqrt{(2)^2 + (0)^2 + (-1)^2} = \sqrt{5}$

$d_{CD} = \sqrt{(1)^2 + (1)^2 + (1)^2} = \sqrt{3}$

	A	B	C	D
A	0	2	$\sqrt{6}$	1
B	2	0	$\sqrt{2}$	$\sqrt{5}$
C	$\sqrt{6}$	$\sqrt{2}$	0	$\sqrt{3}$
D	1	$\sqrt{5}$	$\sqrt{3}$	0

	A	B	C	D
A	0	2	2.45	1
B	2	0	1.41	2.23
C	2.45	1.41	0	1.73
D	1	2.23	1.73	0

Deuxieme étape :

$d_{(AD,B)} = min(d_{(AB)}; d_{(DB)} = d_{(DB)}) = 2$

$d_{(AD,C)} = min(d_{(AD)}; d_{(DC)}) = 1.73$

	AD	B	C
AD	0		
B	2	0	
C	1.73	1.41	0

Troisieme étape :

$d_{(AD,BC)} = min(d_{(AD,B)}; d_{(AD,C)}) = 1.73$

	AD	BC
AD	0	1.73
BC	1.73	0

ici on voit bien que le niveau de coupure est 1.41 *et on obtient* 2 *classes.*

Remarque 1 *Il faut bien savoir détecté la ressemblance dans les classes.*

2 *individus proches*$\Longleftrightarrow$*meme classe* $\Longleftrightarrow$ *variabilité petite (Intra-classe)*

2 *individus éloignées* $\Longleftrightarrow$ *classe différente* $\Longleftrightarrow$ *variabilité grande (Interclasse)*

Donc on obtient 2 *critère de choix du nombres de classe.Ces deux critères jouent le meme role(comme pour* $I_G = I_\Delta + I_{\Delta^*} Hygens$)

1.2 Inertie

On a

$$I_G = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{p}(x_{ij} - \overline{x_{\bullet j}})^2$$
$$= I_{Intra} + I_{Inter}$$
$$I_{Intra} = I_G = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{g}(x_{ij} - \overline{x_{\bullet g}})^2$$

Donc il faut ou bien minI_{Intra} *ou max*I_{Inter} $\Longleftrightarrow 0 \leq \frac{I_{Inter}}{I_G} \leq 1$

plus $\frac{I_{Inter}}{I_G} \to 1$ *plus la partition est bien.*

$\frac{I_{Inter}}{I_G} \to 0 \Rightarrow$ *toutes les classes ont meme moyenne (pas de classification).*

$\frac{I_{Inter}}{I_G} = 1 \Rightarrow$ *les individus sont identiques donc les classes sont homogène c'est le cas idéal.*

Remarque 2 *Tous critère , il admet un défaut et qui est qu'il dépand du nombre d'individus dans la meme classe donc du nombre de classe.*

Pour parliée à ce probleme nous nous avons le critère de Ward.