Fouille de données - Clustering hiérarchique

G. Forestier

Fouille de données

C. Wemmert, S. Lebre

Classification hiérarchique ascendante (CHA)

- **Principe :** créer, à chaque étape, une partition obtenue en agrégeant deux à deux les éléments les plus proches.
- Eléments :
  - individus ou objets à classer
  - regroupements d'individus générés par l'algorithme.
- Chaque individu ou cluster est progressivement absorbé par le cluster le plus proche

Classification hiérarchique ascendante (CHA)

- **Résultat** : hiérarchie de partitions, se présentant sous la forme d'arbres contenant $n-1$ partitions.
- **Définition** : l'ensemble D des données, partitionné en K classes, est une hiérarchie H si :

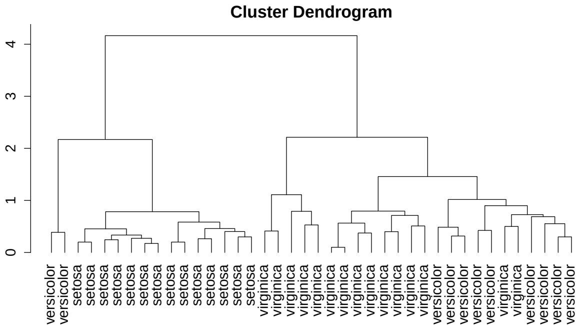
$$D \in H$$

$$\forall x \in D, \{x\} \in H$$

$$\forall h, h' \in H, h \cap h' = \varnothing$$

Classification hiérarchique ascendante (CHA)

Exemple sur Iris :

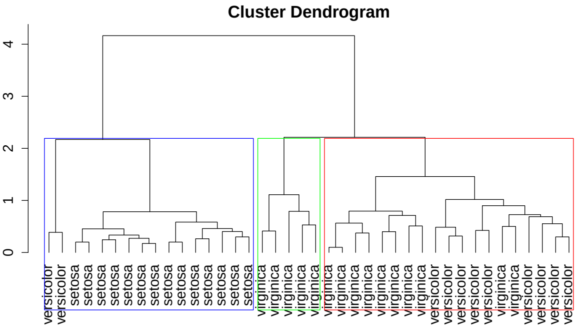


Classification hiérarchique ascendante (CHA)

- **Intérêt** : ces arbres donnent une idée du nombre de classes existant effectivement dans la population.
- En "coupant" l'arbre par une droite horizontale, on obtient une partition, d'autant plus fine que la section est proche des éléments terminaux.
- Une hiérarchie permet donc de fournir une chaîne de *n* partitions ayant de 1 à *n* classes.

Classification hiérarchique ascendante (CHA)

Exemple sur Iris :



Classification hiérarchique ascendante (CHA)

- Au départ : l'ensemble des individus est muni d'une distance.
- Sur quelle base calculer
  1. une distance entre un individu et un groupe ?
  2. une distance entre deux groupes ?
- Définir une stratégie de regroupements des éléments: calcul des distances entre groupements disjoints d'individus
- Critères d'agrégation

Classification hiérarchique ascendante (CHA)

Exemple :

- On peut définir la distance de H à y par la plus petite distance des éléments de H à y

$$d(H, y) = min\{d(x_i, y)\}x_i \in H$$

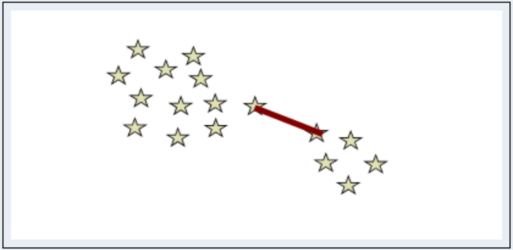
- Saut minimal (*single linkage*)
- On peut définir la distance entre deux regroupements $H1$ et $H2$ par

$$d(H1, H2) = min\{d(x_i, y_j)\} \quad \text{où} \quad x_i \in H1, y_j \in H2$$

Classification hiérarchique ascendante (CHA)

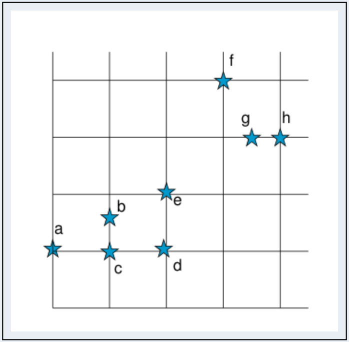
- Indice du lien minimum (plus proche voisin)

$$D(h_1, h_2) = \min_{x_i \in h_1, x_j \in h_2} d(x_i, x_j)$$



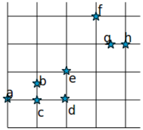
Classification hiérarchique ascendante (CHA)

- Exemple avec l'indice de lien minimal



Classification hiérarchique ascendante (CHA)

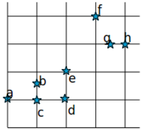
Matrice des distances



	a	b	c	d	e	f	g	h
a	0	1.118	1	2	2.23	4.24	4.03	4.47
b	1.118	0	0.707	1.118	1.118	3.201	2.91	3.35
c	1	0.707	0	1	1.414	3.605	3.201	3.605
d	2	1.118	1.118	0	1	3.162	2.121	2.828
e	2.23	1.118	1.414	1	0	2.236	1.802	2.236
f	4.24	3.20	3.605	3.162	2.236	0	1.118	1.414
g	4.03	2.91	3.201	2.121	1.802	1.118	0	0.707
h	4.47	3.35	3.605	2.828	2.236	1.414	0.707	0

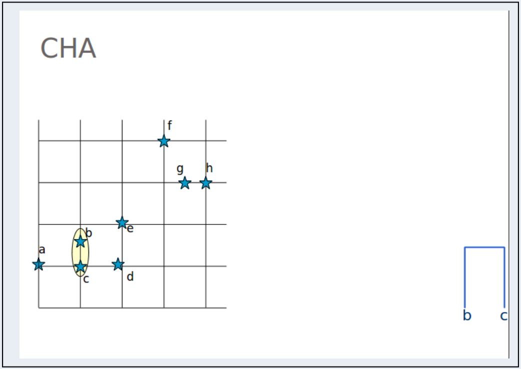
Classification hiérarchique ascendante (CHA)

Matrice des distances



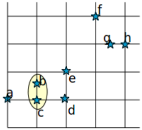
	a	b	c	d	e	f	g	h
a	0	1.118	1	2	2.23	4.24	4.03	4.47
b	1.118	0	$\boxed{0.707}$	1.118	1.118	3.201	2.91	3.35
c	1	0.707	0	1	1.414	3.605	3.201	3.605
d	2	1.118	1.118	0	1	3.162	2.121	2.828
e	2.23	1.118	1.414	1	0	2.236	1.802	2.236
f	4.24	3.20	3.605	3.162	2.236	0	1.118	1.414
g	4.03	2.91	3.201	2.121	1.802	1.118	0	0.707
h	4.47	3.35	3.605	2.828	2.236	1.414	0.707	0

Classification hiérarchique ascendante (CHA)



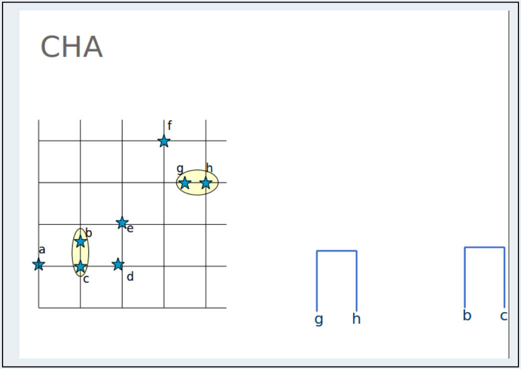
Classification hiérarchique ascendante (CHA)

Matrice des distances



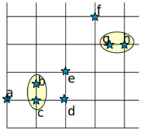
	a	{b,c}	d	e	f	g	h
a	0	1	2	2.23	4.24	4.03	4.47
{b,c}		0	1	1.118	3.201	2.91	3.35
d			0	1	3.162	2.121	2.828
e				0	2.236	1.802	2.236
f					0	1.118	1.414
g						0	$\boxed{0.707}$
h							0

Classification hiérarchique ascendante (CHA)



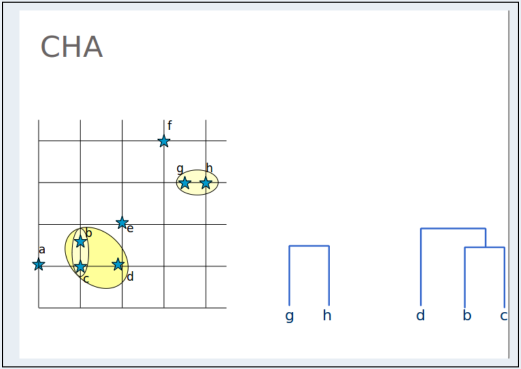
Classification hiérarchique ascendante (CHA)

Matrice des distances



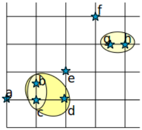
	a	{b,c}	d	e	f	{g,h}
a	0	1	2	2.23	4.24	4.03
{b,c}		0	$\boxed{1}$	1.118	3.201	2.91
d			0	1	3.162	2.121
e				0	2.236	1.802
f					0	1.118
{g,h}						0

Classification hiérarchique ascendante (CHA)



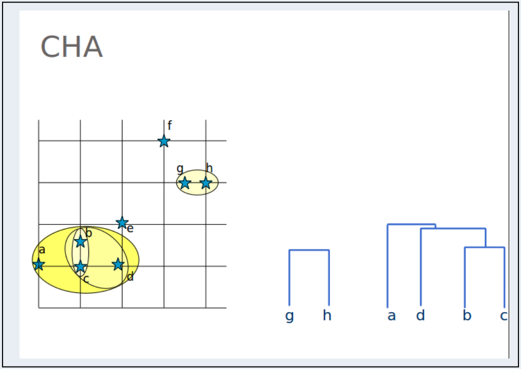
Classification hiérarchique ascendante (CHA)

Matrice des distances



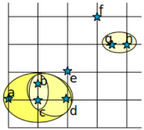
	a	{{b,c},d}	e	f	{g,h}
a	0	$\boxed{1}$	2.23	4.24	4.03
{{b,c},d}		0	1.118	3.201	2.121
e			0	2.236	1.802
f				0	1.118
{g,h}					0

Classification hiérarchique ascendante (CHA)



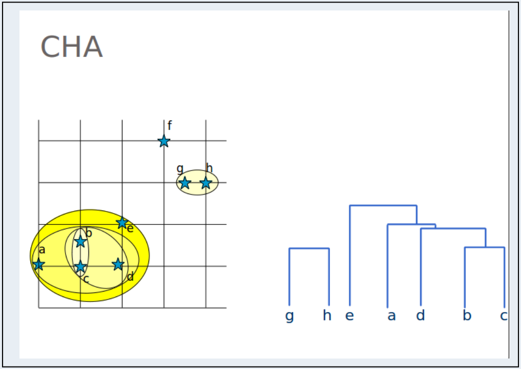
Classification hiérarchique ascendante (CHA)

Matrice des distances



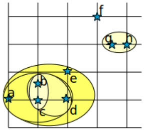
	{a,{{b,c},d}}	e	f	{g,h}
{a,{{b,c},d}}	0	1.118	3.201	2.121
e		0	2.236	1.802
f			0	1.118
{g,h}				0

Classification hiérarchique ascendante (CHA)



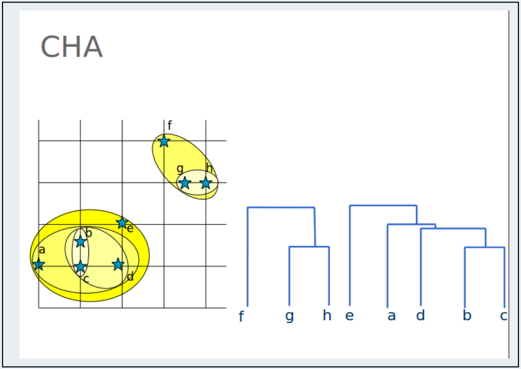
Classification hiérarchique ascendante (CHA)

Matrice des distances



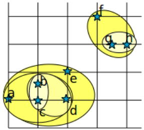
	{{a,{{b,c},d}},e}	f	{g,h}
{{a,{{b,c},d},e}	0	3.201	2.121
f		0	$\boxed{1.118}$
{g,h}			0

Classification hiérarchique ascendante (CHA)



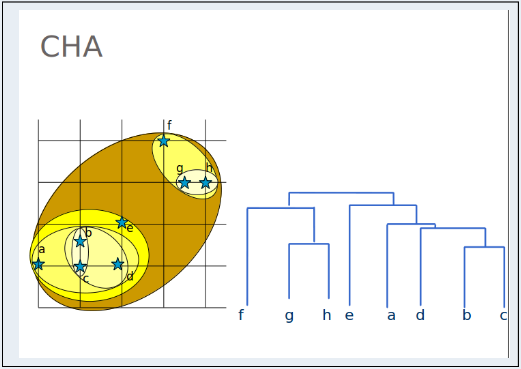
Classification hiérarchique ascendante (CHA)

Matrice des distances



	{{a,{{b,c},d}},e}	{f,{g,h}}
{{a,{{b,c},d},e}	0	$\boxed{1.802}$
{f,{g,h}}		0

Classification hiérarchique ascendante (CHA)



Classification hiérarchique ascendante (CHA)

- On peut définir la distance de $H1$ à $H2$ par la plus grande distance des éléments de $H1$ à $H2$

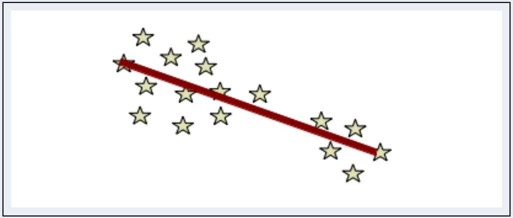
$$d(H1, H2) = max\{d(x_i, y_j)\} \quad \text{où} x_i \in H1, y_j \in H2$$

- Saut maximal (*complete linkage*)
- **Attention :** on fusionne quand même les classes les plus proches

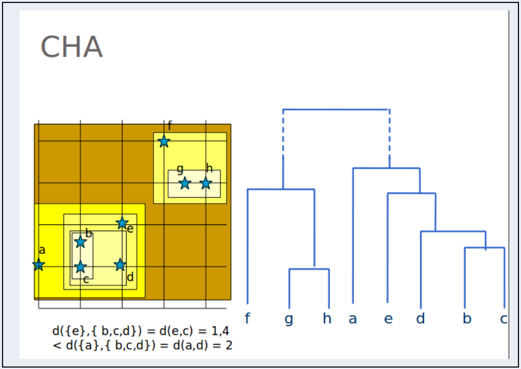
Classification hiérarchique ascendante (CHA)

- Indice du lien maximal (diamètre maximum)

$$d(H, y) = max\{d(x_i, y)\}x_i \in H$$

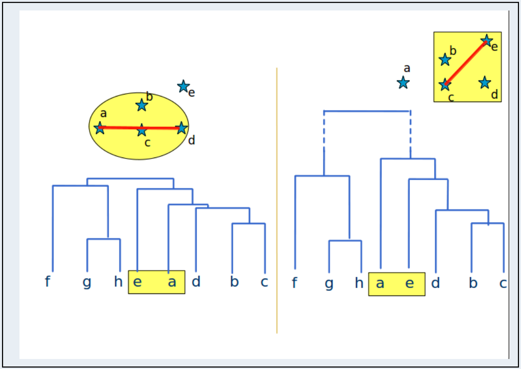


Classification hiérarchique ascendante (CHA)



Remarque : attention autre matrice de distances que pour single link

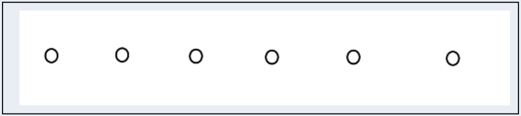
Classification hiérarchique ascendante (CHA)



Remarque : attention autre matrice de distances que pour single link

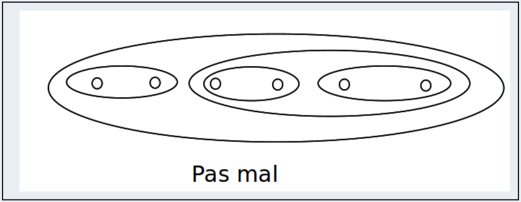
Classification hiérarchique ascendante (CHA)

- Problème des "chaînes" dans la classification



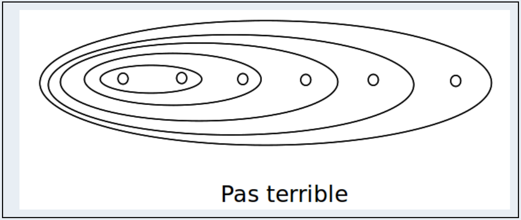
Classification hiérarchique ascendante (CHA)

- Problème des "chaînes" dans la classification



Classification hiérarchique ascendante (CHA)

- Problème des "chaînes" dans la classification



Problème des chaînes

- Tenir compte de la variance des classes
- Indice de Ward réactualisation :
  - Après avoir fusionné $h1$ et $h2$ on calcule la distance avec un autre cluster :

$$\delta_1(h_1 \cup h_2, h) = \frac{|h| + |h_1|}{|h| + |h_1| + |h_2|}\delta_1(h_1, h)$$

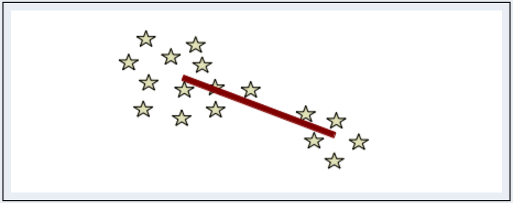
$$+\frac{|h| + |h_1|}{|h| + |h_1| + |h_2|}\delta_1(h_2, h) + \frac{|h| + |h_1|}{|h| + |h_1| + |h_2|}\delta_1(h_1, h_2)$$

  - On peut montrer qu'à chaque étape, la nouvelle partition est celle qui limite l'augmentation de l'inertie intra-classe

Classification hiérarchique ascendante (CHA)

- Indice des centres de gravité (distance moyenne)

$$D(h_1, h_2) = d(g_1, g_2)$$



Classification hiérarchique ascendante (CHA)

- Indice d'agrégation de la variation de l'inertie :

$$\delta_1(h_1, h_2) = \frac{p(h_1).p(h_2)}{p(h_1) + p(h_2)} d(g_1, g_2)$$

- Indice de la vraisemblance du lien :

$$\delta_1(h_1, h_2) = -log(-log([d(h_1, h_2)]^{(n_1, n_2,)^{\epsilon}}))$$

où $d(h_1, h_2)$ est l'indice du lien simple

- Facilite la fusion des classes à variance faible

Ultramétrique

- On appelle ultramétrique sur M, toute application définie par :

$$d : MxM \rightarrow R^+$$

- Ayant les propriétés suivantes :

$$\forall (x, y) \in M^2, d(x, y) = 0 \leftrightarrow x = y$$

$$\forall (x, y) \in M^2, d(y, x) = d(x, y)$$

$$\forall (x, y, z) \in M^3, d(x, z) \leq d(x, y) + d(y, z)$$

et

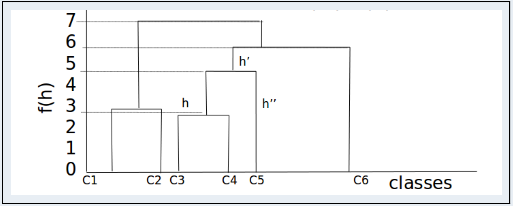
$$\forall (x, y, z) \in M^3, d(x, z) \leq max\{d(x, y), d(y, z)\}$$

Ultramétrique

- une hiérachie est indicée s'il existe f telle que

$$\forall x \in H, f(\{x\}) = 0$$

$$\forall h, h' \in H, h \neq h', h' \subset h \rightarrow f(h') < f(h)$$



Ultramétrique

On peut montrer que :

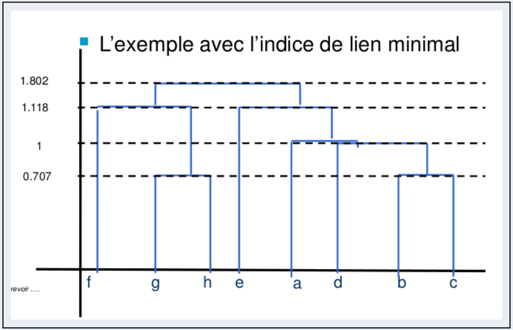
- Toute hiérarchie indicée permet de définir une ultramétrique
- Toute ultramétrique permet de définir un hiérarchie indicée

Ultramétrique

{b};{c} = 0.707 / {h},{g} = 0.707{b,c}.{d} = 1
{{b,c},{d}}.{a} = 1 {{a},{b,c},{d}}.{e} = 1.118
{g,h}.{f} = 1.118 {{a},{b,c},{d}.{e}}{{f},{g,h}} =
1.802

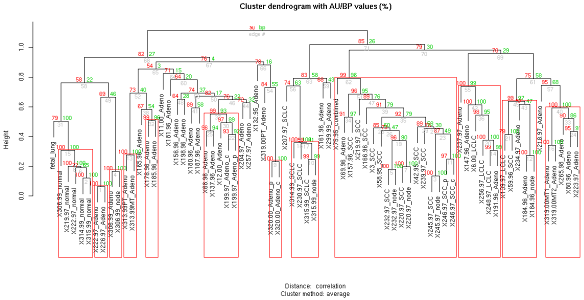
	a	b	c	d	e	f	g	h
a	0	1	1	1	1.118	1.802	1.802	1.802
b	1	0	0.707	1	1.118	1.802	1.802	1.802
c	1	0.707	0	1	1.118	1.802	1.802	1.802
d	1	1	1	0	1.118	1.802	1.802	1.802
e	1.118	1.118	1.118	1.118	0	1.802	1.802	1.802
f	1.802	1.802	1.802	1.802	1.802	0	1.118	1.118
G	1.802	1.802	1.802	1.802	1.802	1.118	0	0.707
H	1.802	1.802	1.802	1.802	1.802	1.118	0.707	0

Ultramétrique



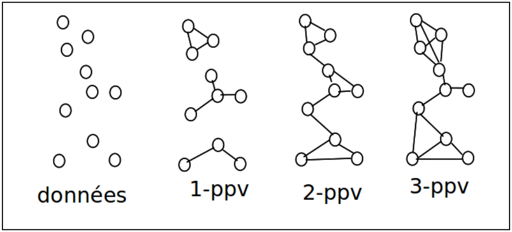
Choix de la coupe dans la hiérarchie

- Utiliser un critère statistique (par exemple la p-value)
- AU : Approximately Unbiased
- BP : Bootstrap Probability



CHA : Chameleon

- **Principe** : estimer la densité intra-cluster et extra-cluster à partir du graphe des k plus proches voisins



CHA : Chameleon

Trouver les clusters initiaux :

- En partitionnant le graphe k-ppv en m partitions "solides" (où la distance entre les points est minimisée)

Fusionner dynamiquement les sous-clusters :

- En fonction des deux critères :
  - RI(C,C') : inter connectivité relative
  - RC(C,C') : proximité relative

CHA : Chameleon

$$RC(C, C') = \frac{(|C| + |C'|)DC(C, C')}{|C|DC(C) + |C'|DC(C')}$$

$$RI(C, C') = \frac{2x|EC(C, C')|}{|EC(C)|xEC(C')|}$$

- $EC(C, C')$: ensemble des arêtes qui relient C et C' $\rightarrow$ inter-connectivité absolue entre 2 clusters
- $EC(C)$: plus petit ensemble d'arêtes qui partitionne C en 2 clusters de taille proche $\rightarrow$ inter-connectivité interne
- $DC(C, C')$: distance moyenne entre les points de C et C'
- $DC(C)$: distance moyenne entre les points de C

CHA : Chameleon

Problèmes :

- Coût en $O(n^3)$ en calcul de distances
- Arbre binaire : en effet agréger k classes d'un coup nécessite
  - de déterminer k
  - de vérifier que les indices entre chaque couple des k classes sont bien minimaux

Construction descendante

CHD : Classification hiérarchique descendante :

- **Principe :** par division de classes, on construit une suite de partitions emboîtées dont les classes forment la hiérarchie H recherchée

Problèmes :

- Comment sélectionner la classe à diviser ?
- Et en combien de sous-classes ?

Classification hiérarchique descendante (CHD)

Comment sélectionner la classe à diviser ?

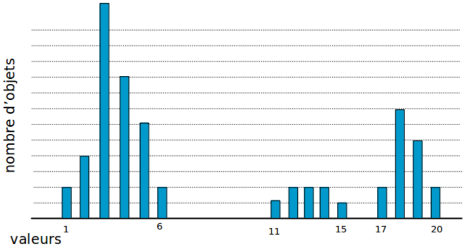
- Soit à chaque niveau, toutes les classes pouvant être développées suivant un certain critère, le sont
- Soit seule la classe de plus fort critère est développée

Critères possibles :

- Variance minimale à respecter
- Nombre d'objets
- Etude des histogrammes des valeurs prise par les données

Classification hiérarchique descendante (CHD)

- Utilisation des histogrammes des valeurs



CHD : étude d'histogramme

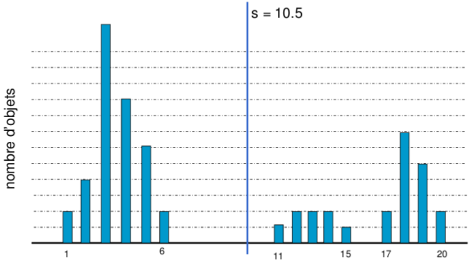
Principe : pour chaque attribut

1. Trouver un seuil séparant l'histogramme en deux "sous-classes"
2. Séparer ces deux sous-classes s'il y a lieu
3. Itérer

CHD : étude d'histogramme

Calcul d'un seuil "théorique"

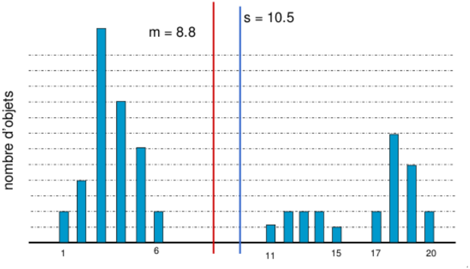
- $s = (max + min)/2 = 10,5$



CHD : étude d'histogramme

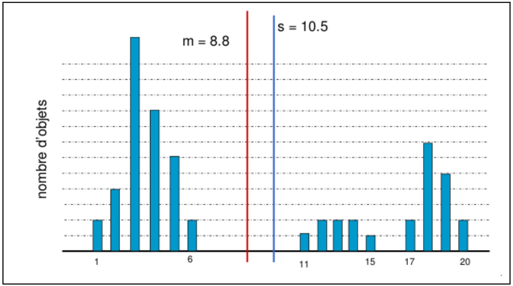
Calcul de la moyenne pondérée

- $m = 526/60 = 8.8$



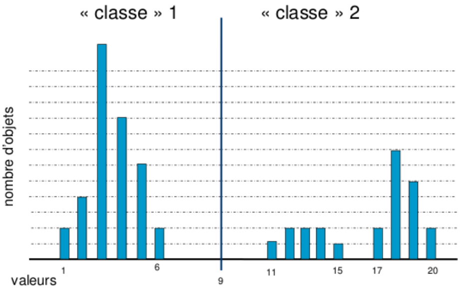
CHD : étude d'histogramme

- $10{,}5 \neq 8{,}8 \rightarrow$ on sépare à 9



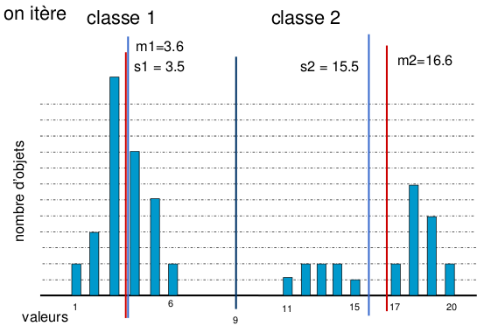
CHD : étude d'histogramme

Exemple :



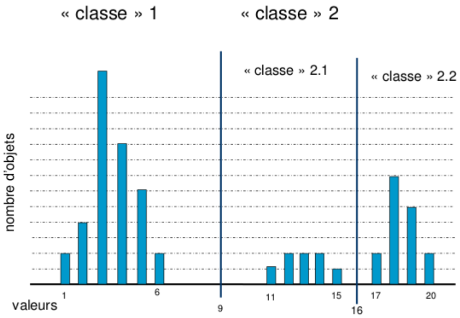
CHD : étude d'histogramme

Exemple :



CHD : étude d'histogramme

Exemple :

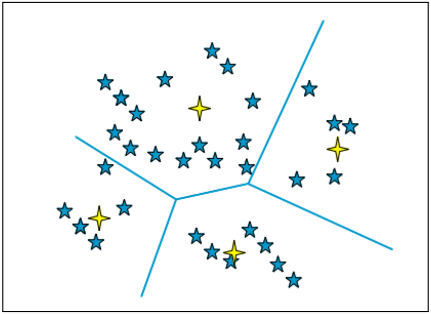


Classification hiérarchique descendante (CHD)

Exemple de Kmeans :

- Deux approches lorsqu'on développe une classe :

1. Soit seuls les objets de la classe sont reclassés par rapport aux nouveaux centres

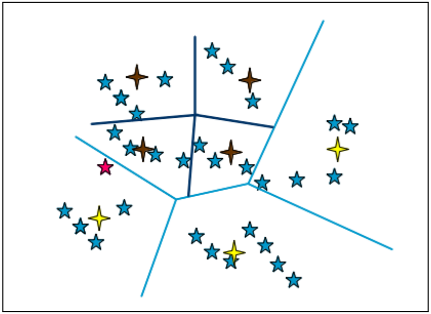


Classification hiérarchique descendante (CHD)

Exemple de Kmeans :

- Deux approches lorsqu'on développe une classe :

1. Soit seuls les objets de la classe sont reclassés par rapport aux nouveaux centres

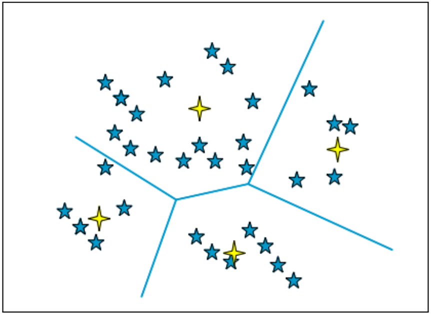


Classification hiérarchique descendante (CHD)

Exemple de Kmeans :

- Deux approches lorsqu'on développe une classe :

2. Soit tous les objets sont reclassés par rapport à tous les centres

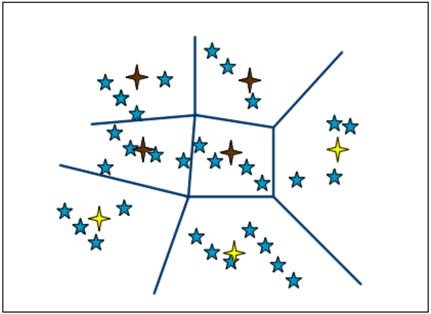


Classification hiérarchique descendante (CHD)

Exemple de Kmeans :

- Deux approches lorsqu'on développe une classe :

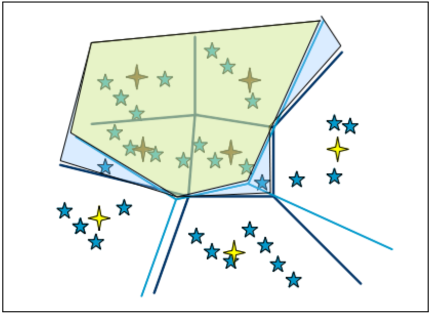
2. Soit tous les objets sont reclassés par rapport à tous les centres



Classification hiérarchique descendante (CHD)

Inconvénient :

- La classe initiale ne recouvre pas l'ensemble des objets affectés aux nouveaux noyaux → on perd le lien de hiérarchie



Classification hiérarchique descendante (CHD)

Comment déterminer K :

- Soit K est à l'aide de critères statistiques
- Soit K est calculé par étude des histogrammes
- Soit K est fixe : en général dans ce cas $K = 2$

Exemple pour Kmeans

- Les deux premiers cas correspondent à Isodata
- Pour le troisième cas, reste le problème de l'initialisation des nouveaux centres

Classification hiérarchique descendante (CHD)

Kmeans hiérarchique :

1. Générer dans la classe à découper en 2 points par une petite perturbation du centre de gravité
2. Appliquer un algorithme Kmeans aux objets de la classe à découper
3. Itérer