
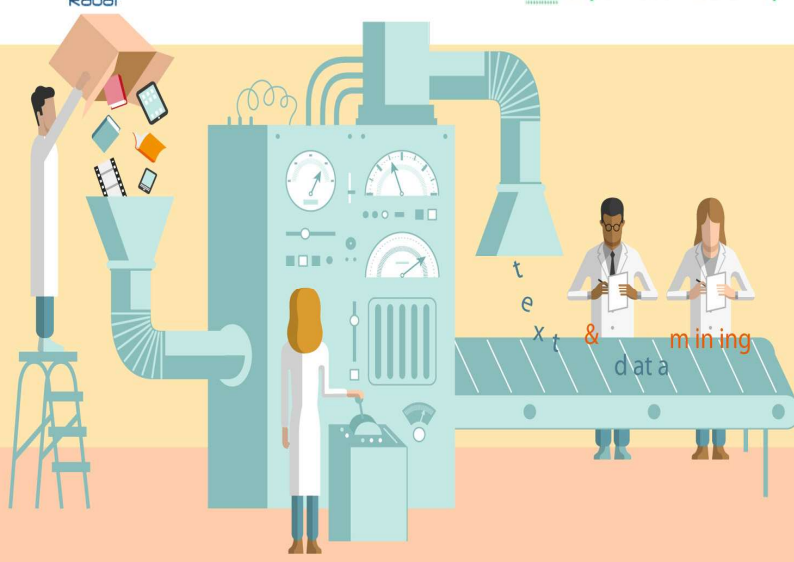


Royaume du Maroc
 Université Mohammed V de Rabat
 Faculté des Sciences
 Département d'Informatique



IPSS
 Intelligent Processing
 Systems & Security



Data Mining & Machine Learning

Master IPS
 Faculté des sciences – Rabat
 Université Mohamed V

127


Chapitre 2 :

Techniques descriptives

Classification/(Clustering)

- Qu'est-ce que le clustering?
- Catégories de clustering
- Méthodes hiérarchiques
- Méthodes de partitionnement
- Méthodes basées sur la densité
- Méthodes mixtes

(1ère partie)



ipes.boutyour@gmail.com

128

128

Techniques Descriptives (rappel)

- Trois grandes familles
 - Classification
 - Règles d'association
 - Analyse factorielle

Qu'est ce que le clustering?

- Objectif:
 - Rassembler les objets qui se ressemblent et/ou
 - Séparer les objets qui diffèrent
 - Regroupement des objets selon leurs ressemblances
- Finalité:
 - Mieux traiter et analyser un grand volume de données en le découpant en classes homogènes

Qu'est ce que le clustering?

Propriétés :

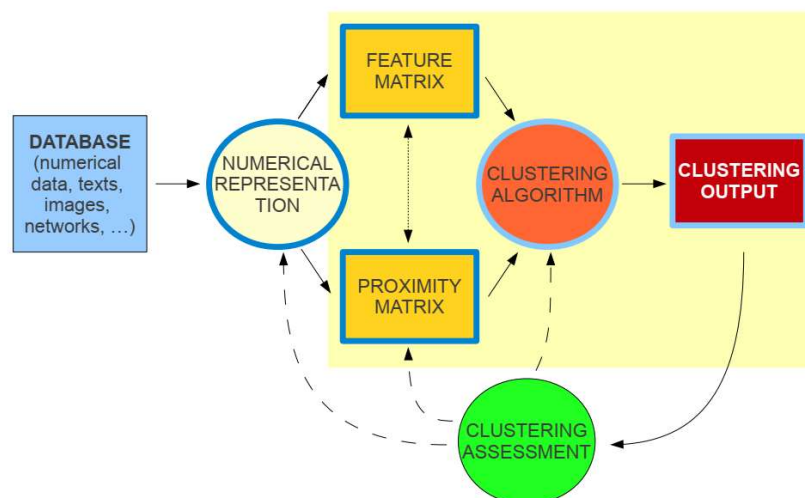
- Le nombre de classes n'est pas à fixer à l'avance
- Les classes ne sont pas connues à l'avance mais découvertes par le processus de classification
- Intérêt :
 - Utiliser les méthodes de classification pour trouver une partition répondant au problème sans avoir à parcourir toutes les partitions possibles.

ipes.boutyour@gmail.com 131

131

Qu'est ce que le clustering?

Le processus de clustering



ipes.boutyour@gmail.com 132

132

Qu'est ce que le clustering?

Représentation des objets à classer

- Matrice des similarités ou distances entre individus (resp. variables)
 - Deux individus se ressemblent ou sont proches s'ils ont des valeurs proches pour toutes les variables considérées.
 - Deux variables se ressemblent ou sont proches → sont corrélées

ipes.boutyour@gmail.com 133

133

Qu'est ce que le clustering?

Représentation des objets à classer

- Matrice des données

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Matrice des distances (similarité/dissimilarité)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

ipes.boutyour@gmail.com 134

134

Qu'est ce que le clustering?

Similarité et dissimilarité

- Métrique de similarité/dissimilarité exprimée en terme d'une fonction de distance, typiquement $d(i; j)$.
- Fonction de distance dépend du type des données: binaire, catégoriel, ordinal ou continu
- Pondération des dimensions selon l'application et la sémantique des données

ipes.boutyour@gmail.com 135

135

Qu'est ce qu'un bon clustering?

- Une bonne méthode va produire des clusters dont les éléments ont:
 - une forte similarité intra-classe
 - une faible similarité inter-classe
- Détecter les structures (patterns) présentes (mais cachées) dans les données
- Permettre de déterminer le nombre optimal de classes
- Fournir des classes :
 - Bien différenciées
 - Stables vis-à-vis de légères modifications des données
- Traiter efficacement un grand volume de données

ipes.boutyour@gmail.com 136

136

Qu'est ce que le clustering?

Représentation des objets à classer

- Matrice des similarités ou distances entre individus (resp. variables)
 - Deux individus se ressemblent ou sont proches s'ils ont des valeurs proches pour toutes les variables considérées.
 - Deux variables se ressemblent ou sont proches → sont corrélées

Quelle distance / mesure de similarité utiliser?

Matrice des distances

Individus $X=(X_1, \dots, X_n)$ et $Y=(Y_1, \dots, Y_n)$

- Variables quantitatives continues

- **Distance de Minkowski :**

$$d(X; Y) = \left(\sum_{i=1}^n |X_i - Y_i|^q \right)^{1/q}$$

- Si $q = 1$: **Distance de Manhattan :**

$$d(X; Y) = \sum_{i=1}^n |X_i - Y_i|$$

- Si $q = 2$: **Distance de Euclidienne :**

$$d(X; Y) = \left(\sum_{i=1}^n |X_i - Y_i|^2 \right)^{1/2}$$

Matrice des distances

Individus $X=(X_1, \dots, X_n)$ et $Y=(Y_1, \dots, Y_n)$

○ Variables quantitatives continues

▪ **Distance de Canberra :**

$$d(X; Y) = \sum_{i=1}^n \frac{|X_i - Y_i|}{|X_i + Y_i|}$$

▪ **Distance de Chebyshev :**

$$d(X; Y) = \max_{i \in [1, n]} |X_i - Y_i|$$

ipes.boutyour@gmail.com

139

139

Matrice des distances

Individus $X=(X_1, \dots, X_n)$ et $Y=(Y_1, \dots, Y_n)$

○ Variables binaires

• **Jaccard** : $S_{jaccard}(x, y) = \frac{a}{a+b+c} \in [0, 1]$

• **Dice** : $S_{dice}(x, y) = \frac{2a}{2a+b+c} \in [0, 1]$

• **Ochiai** : $S_{ochiai}(x, y) = \frac{a}{\sqrt{(a+b)(a+c)}} \in [0, 1]$

• **Kulczynski** : $S_{kulczynski}(x, y) = \frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right) \in [0, 1]$

• **Sokal-Michener** : $S_{soc-mich}(x, y) = \frac{a+d}{a+b+c+d} \in [0, 1]$

• **Rogers-Tanimoto** : $S_{rog-tan}(x, y) = \frac{a+d}{a+2(b+c)+d} \in [0, 1]$

• **Phi** : $S_{phi}(x, y) = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1, 1]$

		Y	
		1	0
X	1	a	b
	0	c	d

ipes.boutyour@gmail.com

140

140

Matrice des distances

Individus $X=(X_1, \dots, X_n)$ et $Y=(Y_1, \dots, Y_n)$

○ Variables mixtes

■ Indice de Gower :

$$\frac{1}{n} \sum_{i=1}^n s_i(X, Y)$$

■ Si i_e variable qualitative :

$s_i(X, Y) = 1$ si même valeur pour X et Y; 0 sinon

■ Si i_e variable quantitative :

$$\frac{1 - |X_i - Y_i|}{\text{Max}_{U,V}(|U_i - V_i|)}$$

Matrice des distances

Exemple de calcul de la matrice des distances

○ Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

Matrice des distances

Exemple de calcul de la matrice des distances

- Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

	Sara	Karim	Ali	Rania
Sara	0	-	-	-
Karim		0	-	-
Ali			0	
Rania				0

ipes.boutyour@gmail.com

143

143

Matrice des distances

Exemple de calcul de la matrice des distances

- Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

	Sara	Karim	Ali	Rania
Sara	0	-	-	-
Karim	?	0	-	-
Ali			0	
Rania				0

$$d(\text{Karim}, \text{Sara})^2 = (20-19)^2 + (15-17)^2 + (12-15)^2 + (10-12)^2 \\ = 1 + 4 + 9 + 4 = 18$$

$$d(\text{Karim}, \text{Sara}) = 4,24$$

ipes.boutyour@gmail.com

144

144

Matrice des distances

Exemple de calcul de la matrice des distances

- Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

	Sara	Karim	Ali	Rania
Sara	0	-	-	-
Karim	4,24	0	-	-
Ali			0	
Rania				0

ipes.boutyour@gmail.com

145

145

Matrice des distances

Exemple de calcul de la matrice des distances

- Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

	Sara	Karim	Ali	Rania
Sara	0	-	-	-
Karim	4,24	0	-	-
Ali			0	
Rania				0

$$d(\text{Ali}, \text{Sara})^2 = (-.-)^2 + (-.-)^2 + (-.-)^2 + (-.-)^2$$

$$= . + . + . + . = .$$

$$d(\text{Ali}, \text{Sara}) = .$$

ipes.boutyour@gmail.com

146

146

Matrice des distances

Exemple de calcul de la matrice des distances

- Calculer la distance euclidienne pour les observations suivantes :

	Mathématiques	Physique	Français	Histoire Géo
Sara	19	17	15	12
Karim	20	15	12	10
Ali	10	14	19	14
Rania	12	10	18	18

	Sara	Karim	Ali	Rania
Sara	0	-	-	-
Karim	4,24	0	-	-
Ali	10.49	12,88	0	-
Rania	11.96	13,75	6,08	0

ipes.boutyour@gmail.com

147

147

Matrice des distances

Exemple de calcul de l'indice de Jaccard:

- Soit X et Y avec les valeurs suivantes :

- $X = (1, 0, 1, 0, 0, 0, 0)$

- $Y = (1, 0, 0, 1, 0, 1, 1)$

Calculer Indice de Jaccard.

On calcule tout d'abord a, b, c et d.

$a =$; $b =$; $c =$; $d =$

$$J = \frac{a}{a+b+c} =$$

		Y	
		1	0
X	1	a	b
	0	c	d

ipes.boutyour@gmail.com

148

148

Matrice des distances

Exemple de calcul de l'indice de Jaccard:

- Soit X et Y avec les valeurs binaires suivantes :

■ $X = (1, 0, 1, 0, 0, 0, 0)$

■ $Y = (1, 0, 0, 1, 0, 1, 1)$

Calculer Indice de Jaccard.

On calcule tout d'abord a, b, c et d.

$a = 1$; $b = 1$; $c = 3$; $d = 2$

$$J = \frac{a}{a+b+c} = \frac{1}{1+1+3} = \frac{1}{5} = 0.2$$

		Y	
		1	0
X	1	a	b
	0	c	d

Matrice des distances

Exemple de calcul de l'indice de Jaccard:

- Soit X et Y avec les valeurs binaires suivantes :

■ $X = (1, 0, 1, 0, 0, 0, 0)$

■ $Y = (1, 0, 0, 1, 0, 1, 1)$

Calculer Indice de Jaccard.

On calcule tout d'abord a, b, c et d.

$a = 1$; $b = 1$; $c = 3$; $d = 2$

$$S_{jaccard} = \frac{a}{a+b+c} = \frac{1}{1+1+3} = \frac{1}{5} = 0.2$$

Calculer les autres indices : dice, ochiai, kulczynski, sok-mich, rog-tan, phi.

		Y	
		1	0
X	1	a	b
	0	c	d

Matrice des distances

Exemple de calcul de l'indice de Jaccard:

- Pour calculer les indices de similarité pour les données binaires avec R, on utilise:

```
> install.packages("proxy")
> library(proxy)
> x=c(1,1,1,1,0,1,0)
> y=c(1,0,1,1,1,0,0)
> X=data.frame(rbind(x,y))
> simil(X,method="Jaccard")
      x
y 0.5
> simil(X,method="Dice")
      x
y 0.6666667
> simil(X,method="Ochiai")
      x
y 0.6708204
> simil(X,method="Kulczynski2")
      x
y 0.675
> simil(X,method="Sokal/Michener")
      x
y 0.5714286
> simil(X,method="Rogers")
      x
y 0.4
> simil(X,method="Phi")
```

ipes.boutyour@gmail.com 151

151

Normalisation des données

Variables continues

- En général, les données brutes contiennent des caractéristiques qui sont des mesures différentes à des échelles différentes (par exemple cm, kg, euros, . . .). Dans ce cas, toute mesure de proximité peut être biaisée.
- Avant d'appliquer toute mesure de proximité ou tout algorithme de clustering, il faut normaliser les données. Notons \mathbf{X}^* l'ensemble des données brutes :

- Afin de normaliser les données brutes, nous pouvons soustraire une mesure de position et diviser une mesure de dispersion pour chaque caractéristique j :

$$x_{ij} = \frac{x_{ij}^* - L_j^*}{M_j^*}$$

$$\mathbf{X}^* = \begin{pmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1p}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^* & x_{n2}^* & \cdots & x_{np}^* \end{pmatrix}$$

ipes.boutyour@gmail.com 152

152

Normalisation des données

Variables continues

Notons pour chaque caractéristique j son :

- Moyenne : $\mu_j^* = \frac{1}{n} \sum_{i=1}^n x_{ij}^*$
- Écart-type : $\sigma_j^* = \left(\frac{1}{n-1} \sum_{i=1}^n (x_{ij}^* - \mu_j^*)^2 \right)^{1/2}$
- Range : $\max_i \{x_{ij}^*\} - \min_i \{x_{ij}^*\}$

Les deux normalisations de données les plus utilisées à l'aide de l'équations précédentes sont :

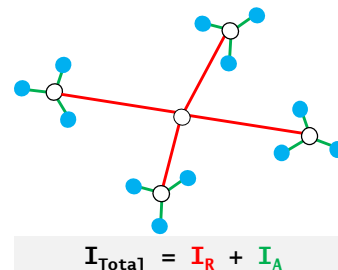
- **Z-score** : $x_{ij} = \frac{x_{ij}^* - \mu_j^*}{\sigma_j^*}$; $\forall j: \mu_j = 0$ et $\sigma_j = 1$
- **Range** : $x_{ij} = \frac{x_{ij}^* - \min_i \{x_{ij}^*\}}{r_j^*}$; $\forall j: \mu_j = \frac{\mu_j^* - \min_i \{x_{ij}^*\}}{r_j^*}$ et $\sigma_j = \frac{\sigma_j^*}{r_j^*}$
- La méthode range est sensible aux outliers.

ipes.boutyour@gmail.com 153

153

Critère d'homogénéité

- Soit $C = \{e_i \mid i = 1, \dots, n\}$ un groupe de n individus de barycentre g , partitionné en k classes C_i d'effectifs n_i et de barycentres $g_i (i = 1, \dots, k)$
- Inertie totale de C : $I_{Total} = \frac{1}{n} \sum_{i=1}^n d^2(e_i, g) = I_R + I_A$
- Inertie inter-classe : $I_R = \frac{1}{n} \sum_{i=1}^k n_i d^2(g_i, g)$
- Inertie intra-classe : $I_A = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} d^2(e_j, g_i)$



ipes.boutyour@gmail.com 154

154

Principales méthodes du clustering

- Méthodes hiérarchiques :
 - Ascendante
 - Descendante
- Méthodes de partitionnement :
 - Nuée dynamique
 - Centres mobiles
 - K-means
 - Réseaux de Kohonon
- Méthodes à estimation de densité :
 - DBSCAN
 - DENCLUE
- Méthodes mixtes :
 - Algorithme BIRCH
 - Méthode hybride de Wong

ipes.boutyour@gmail.com 155

155

Chapitre 2 : Techniques descriptives

Classification/(Clustering) Méthodes hiérarchiques

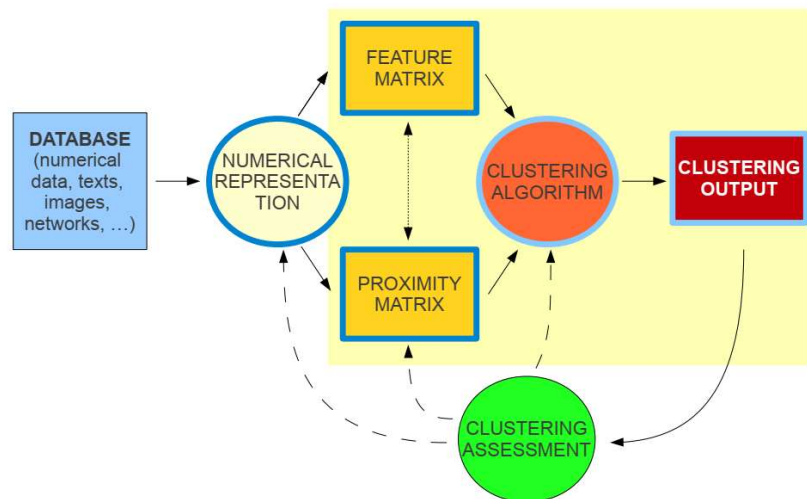
- Classification Ascendante Hiérarchique (CAH)
- Classification Descendante Hiérarchique (CDH)

(2^{ème} partie)

ipes.boutyour@gmail.com 156

156

Le processus de clustering



ipes.boutyour@gmail.com

157

157

Clustering : Méthodes hiérarchiques

○ Classification Ascendante Hiérarchique (CAH) :

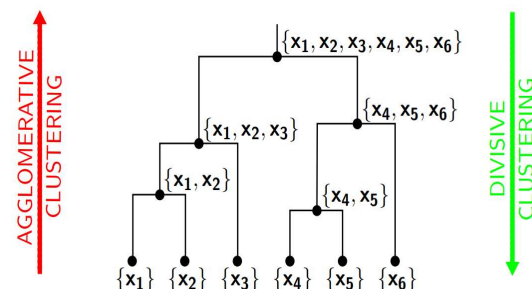
(Agglomerative hierarchical clustering)

- Chaque classe est progressivement absorbée par la classe la plus proche jusqu'à n'avoir qu'une seule classe qui regroupe tous les objets
- Souvent utilisé

○ Classification Descendante Hiérarchique (CDH):

(Divisive hierarchical clustering)

- Part d'une seule classe contenant toutes les observations qui est divisée progressivement jusqu'à obtenir une observation par classe
- Pratiquement jamais utilisé



ipes.boutyour@gmail.com

158

158

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Pseudo code :

- Initialiser la représentation de l'arbre avec n feuilles
- Tant que tous les points de données ne soient pas regroupés faire
 - Fusionner les deux classes les plus proches
 - Mettre à jour la matrice de distances selon le **critère d'agrégation** adopté

ipes.boutyour@gmail.com

159

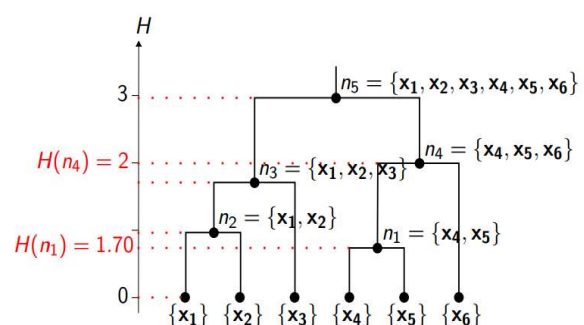
159

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Dendrogramme :

- Le résultat est représenté sous la forme d'un arbre de classification ou dendrogramme
- Le nombre de classes est déterminé par le niveau où on coupe l'arbre de classification



ipes.boutyour@gmail.com

160

160

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Pseudo code :

- Initialiser la représentation de l'arbre avec n feuilles
- Tant que tous les points de données ne soient pas regroupés faire
 - Fusionner les deux classes les plus proches
 - Mettre à jour la matrice de distances selon le **critère d'agrégation** adopté

Quel critère d'agrégation utiliser?

ipes.boutyour@gmail.com

161

161

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

- On peut distinguer les algorithmes de CAH en fonction du type de mesures de distance utilisées. Il existe deux approches :

Les méthodes graphiques :

- Méthode du lien unique(simple)
- Méthode du lien complet
- Méthode de la moyenne des groupes (UPGMA)
- Méthode de la moyenne pondérée des groupes (WPGMA)

Les méthodes géométriques :

- Méthode de Ward
- Méthode des centroïdes
- Méthode de la médiane

Pour plus d'info, chercher la formule de « Lance-Williams »

Dans les méthodes basées sur les graphes, les distances entre les clusters reposent sur les distances entre les points de données dans les clusters, tandis que dans les méthodes basées sur la géométrie, les clusters sont représentés par des centroïdes et la distance entre eux repose sur la distance entre les centroïdes.

ipes.boutyour@gmail.com

162

162

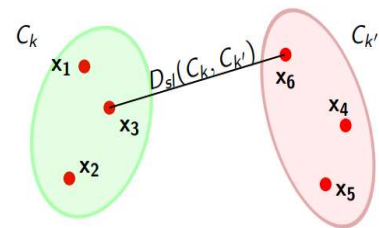
Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Lien unique ou simple :

- L'une des méthodes CAH les plus simples proposée par Sneath en 1957
- Saut minimum ou single linkage
- Également connue sous le nom de méthode du plus proche voisin, car elle utilise le plus proche voisin pour mesurer la dissimilarité entre deux clusters.
- Distance minimale entre deux classes $C_1=\{a, b\}$ et $C_2=\{e, f\}$:

$$d_{\min}(C_1, C_2) = \min(d(a, e), d(a, f), d(b, e), d(b, f))$$



ipes.boutyour@gmail.com

163

163

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Lien unique ou simple :

■ Avantages :

- Simple à calculer
- Ne nécessite pas de recalculer la matrice des distances
- Permet de détecter les classes allongées ou sinueuses

■ Inconvénients :

- A tendance à créer un petit nombre de clusters à grands effectifs
- Moins adapté pour détecter les classes sphériques
- Sensible à l'effet de chaîne

ipes.boutyour@gmail.com

164

164

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

Exemple du lien unique ou simple :

Initialiser la matrice des distances entre les observations

Objet	1	2	3	4	5	6
1	0	-	-	-	-	-
2	0.27	0	-	-	-	-
3	0.11	0.23	0	-	-	-
4	0.22	0.33	0.37	0	-	-
5	0.38	0.31	0.4	0.29	0	-
6	0.17	0.19	0.25	0.13	0.12	0

5 classes au départ : $\{1, 3\}$; $\{2\}$; $\{4\}$; $\{5\}$; $\{6\}$

ipes.boutyour@gmail.com

167

167

Clustering : Méthodes hiérarchiques

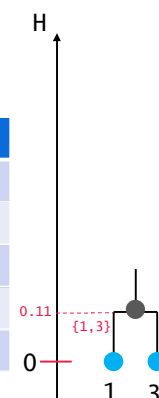
Classification Ascendante Hiérarchique (CAH)

Exemple du lien unique ou simple :

Fusionner la classe 1 et 3

Mettre à jour la matrice des distances entre les observations

Objet	$\{1, 3\}$	2	4	5	6
$\{1, 3\}$	0	-	-	-	-
2	0.23	0	-	-	-
4	0.22	0.33	0	-	-
5	0.38	0.31	0.29	0	-
6	0.17	0.19	0.13	0.12	0



5 classes au départ : $\{1, 3\}$; $\{2\}$; $\{4\}$; $\{5\}$; $\{6\}$

ipes.boutyour@gmail.com

168

168

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

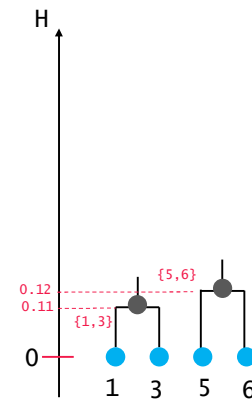
Exemple du lien unique ou simple :

Fusionner la classe 5 et 6

Mettre à jour la matrice des distances entre les observations

Objet	{1,3}	2	4	{5,6}
{1,3}	0	-	-	-
2	0.23	0	-	-
4	0.22	0.33	0	-
{5,6}	0.17	0.19	0.13	0

4 classes au départ : {1, 3} ; {2} ; {4} ; {5, 6}



ipes.boutyour@gmail.com

169

169

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

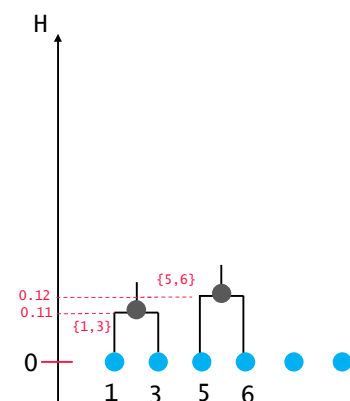
Exemple du lien unique ou simple :

Fusionner la classe {4} et {5,6}

Mettre à jour la matrice des distances entre les observations

Objet	{1,3}	2	4	{5,6}
{1,3}	0	-	-	-
2	0.23	0	-	-
4	0.22	0.33	0	-
{5,6}	0.17	0.19	0.13	0

3 classes au départ : {1, 3} ; {2} ; {4, 5, 6}



ipes.boutyour@gmail.com

170

170

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

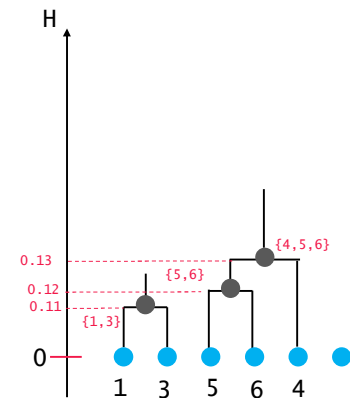
Exemple du lien unique ou simple :

Fusionner la classe $\{4\}$ et $\{5,6\}$

Mettre à jour la matrice des distances entre les observations

Objet	$\{1,3\}$	2	$\{4,5,6\}$
$\{1,3\}$	0	-	-
2	0.23	0	-
$\{4,5,6\}$	0.17	0.19	0

3 classes au départ : $\{1, 3\}$; $\{2\}$; $\{4, 5, 6\}$



ipes.boutyour@gmail.com

171

171

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

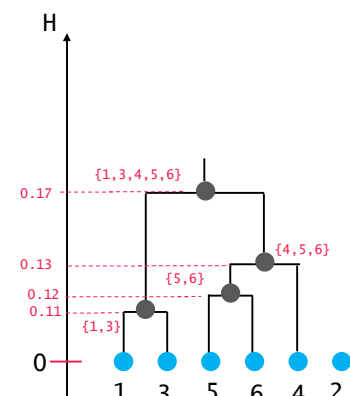
Exemple du lien unique ou simple :

Fusionner la classe $\{4\}$ et $\{5,6\}$

Mettre à jour la matrice des distances entre les observations

Objet	$\{1,3\}$	2	$\{4,5,6\}$
$\{1,3\}$	0	-	-
2	0.23	0	-
$\{4,5,6\}$	0.17	0.19	0

3 classes au départ : $\{1, 3\}$; $\{2\}$; $\{4, 5, 6\}$



ipes.boutyour@gmail.com

172

172

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

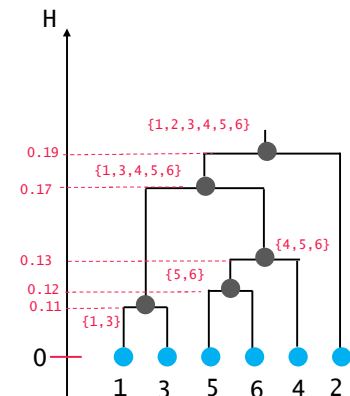
Exemple du lien unique ou simple :

Fusionner la classe $\{1,3\}$ et $\{4,5,6\}$

Mettre à jour la matrice des distances entre les observations

Objet	$\{1, 3, 4, 5, 6\}$	2
$\{1, 3, 4, 5, 6\}$	0	-
2	0.19	0

2 classes au départ : $\{1, 3, 4, 5, 6\}$; $\{2\}$



ipes.boutyour@gmail.com

173

173

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

Exemple du lien unique ou simple :

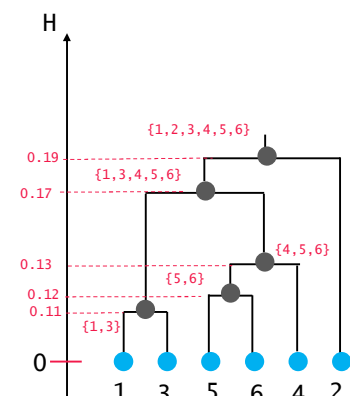
Fusionner la classe $\{2\}$ et $\{1,3,4,5,6\}$

Mettre à jour la matrice des distances entre les observations

Objet	$\{1, 3, 4, 5, 6\}$	2
$\{1, 3, 4, 5, 6\}$	0	-
2	0.19	0

1 classes au départ : $\{1, 3, 4, 5, 6, 2\}$

○ Arrêt de l'algorithme



ipes.boutyour@gmail.com

174

174

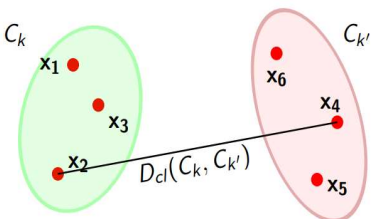
Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Lien complet :

- Introduite par McQuitty en 1960
- Saut maximum ou complete linkage
- Distance maximale entre deux classes $C_1=\{a, b\}$ et C_k
 $C_2=\{e, f\}$:

$$d_{min}(C_1, C_2) = \max(d(a, e), d(a, f), d(b, e), d(b, f))$$
- Contrairement aux méthodes à lien unique, elle utilise le voisin le plus éloigné pour mesurer la dissimilarité entre deux clusters, mais prend la distance minimale à partir de la matrice des distances.
- A tendance à produire des classes à diamètres égaux
- Peu utilisé



ipes.boutyour@gmail.com

175

175

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Lien complet :

- Avantages :
 - Simple à calculer
 - Ne nécessite pas de recalculer la matrice des distances
 - Moins sensible à l'effet de chaîne
- Inconvénients :
 - Très sensible aux valeurs extrêmes/aberrantes
 - Crée un grand nombre de clusters à petits effectifs

ipes.boutyour@gmail.com

176

176

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

Exemple du lien complet

Même exemple qu'auparavant

Objet	1	2	3	4	5	6
1	0	-	-	-	-	-
2	0.27	0	-	-	-	-
3	0.11	0.23	0	-	-	-
4	0.22	0.33	0.37	0	-	-
5	0.38	0.31	0.4	0.29	0	-
6	0.17	0.19	0.25	0.13	0.12	0

6 classes au départ : {1} ; {2} ; {3} ; {4} ; {5} ; {6}

- Créer l'arbre CAH avec le lien complet et dessiner-le.

ipes.boutyour@gmail.com 177

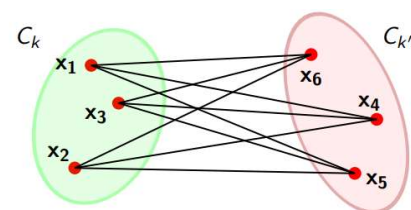
177

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Méthode de Ward

- Proposée par Ward en 1963
- Basé sur la perte d'information quantifiée en termes de critère de la somme des carrés des erreurs (ESS).
- Agréger les individus en minimisant l'inertie (variance) intra-classe I_A et maximisant l'inertie interclasse I_R
- Fusionner les classes C_1 et C_2 qui minimisent l'indice de dissimilarité



g_A = centre de gravité de la classe A (poids p_A)
 g_B = centre de gravité de la classe B (poids p_B)

$$\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

ipes.boutyour@gmail.com 178

178

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Méthode de Ward

■ Avantages :

- Le plus utilisé des critères d'agrégation
- Le plus performant des critères d'agrégation selon Griffiths and al.

■ Inconvénients :

- Peu adapté pour détecter des classes allongées
- Très sensibles aux valeurs extrêmes/aberrantes

ipes.boutyour@gmail.com

179

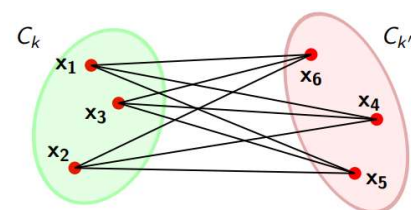
179

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Méthode de la moyenne des groupes (UPGMA)

- Proposée par McQuitty en 1967
- Saut moyen ou average linkage
- Également appelée UPGMA pour "Unweighted Pa Group Method using Arithmetic mean" (méthode de groupes de paires non pondérées utilisant la moyenne arithmétique).
- Distance moyenne entre les observations de deux classes C1 et C2



$$D_{upgma}(C_k, C_{k'}) = \frac{1}{|C_k||C_{k'}|} \sum_{x \in C_k, y \in C_{k'}} D(x, y)$$

ipes.boutyour@gmail.com

180

180

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Méthode de la moyenne des groupes (UPGMA)

■ Avantages :

- Bon compromis entre la distance minimale et la distance maximale
- A tendance à produire des classes de variance proche

■ Inconvénients :

- Moins simple à calculer
- Nécessite de recalculer les distances

ipes.boutyour@gmail.com

181

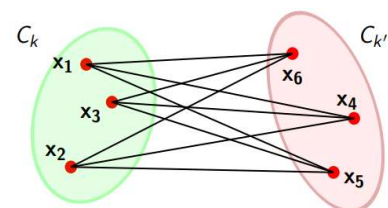
181

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Méthode de la moyenne pondérée du groupe (WPGMA)

- Proposée par McQuitty en 1966
- Également appelée WPGMA pour "Weighted Pair Group Method using Arithmetic mean" (méthode des groupes de paires pondérées utilisant la moyenne arithmétique).
- Distance moyenne entre les observations de deux classes C_1 et C_2



$$D_{wpgma}(C_k, C_l \cup C_{l'}) = \frac{1}{2}D_{wpgma}(C_k, C_l) + \frac{1}{2}D_{wpgma}(C_k, C_{l'})$$

ipes.boutyour@gmail.com

182

182

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Exemples de calcul : Warn, UPGMA et WPGMA

- Pour plus d'information et d'exemples , consulter le site :

- <http://www.slimsuite.unsw.edu.au/teaching/upgma/>
- https://www.mun.ca/biology/scarr/UPGMA_vs_WPGMA.html
- https://cs.slu.edu/~goldwasser/courses/slu/csci1020/2019_Spring/lectures/UPGMA/
- <http://www.wjheeringa.nl/thesis/thesis06.pdf>

ipes.boutyour@gmail.com

183

183

Clustering : Méthodes hiérarchiques

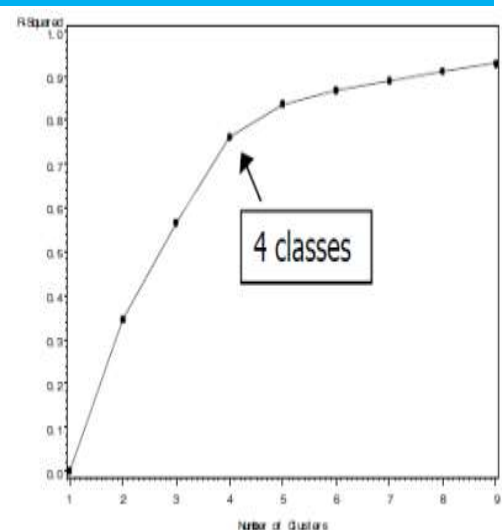
Classification Ascendante Hiérarchique (CAH)

○ Qualité d'une classification: R^2

- Proportion de la variance (inertie) expliquée par les classes.

$$R^2 = \frac{I_R}{I_{Total}}$$

- $0 \leq R^2 \leq 1$
- Plus c'est proche de 1, plus la classification est bonne
- Critère d'arrêt de fusion des classes: arrêter après le dernier changement de valeur important du R^2



ipes.boutyour@gmail.com

184

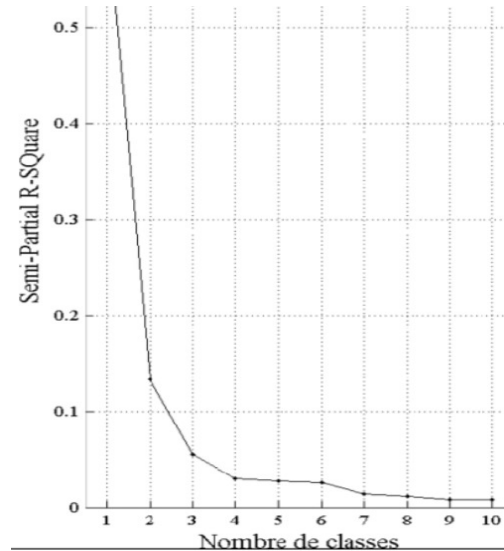
184

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Qualité d'une classification: SPRSQ

- Mesure de la perte d'inertie interclasse causée par la fusion de deux classes
- Proportionnelle à la hauteur d'une branche du dendrogramme
- Bonne classification → inertie interclasse maximum
- Chercher un faible SPRSQ à $k+1$ classes suivi d'un fort SPRSQ à k classes
→ bonne classification en $k+1$ classes



ipes.boutyour@gmail.com

185

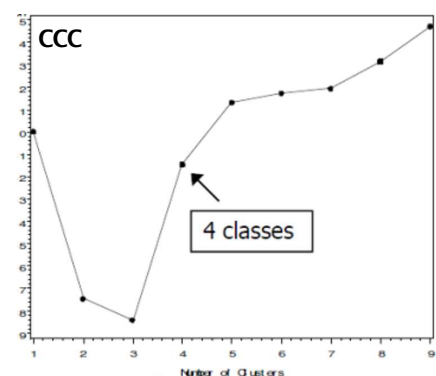
185

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Qualité d'une classification: CCC

- Cubic Clustering Criterion
- Indicateur de comparaison entre la classification obtenue et celle relative à un R^2 espéré (d'une distribution uniforme)
 - $CCC > 2 \rightarrow$ bonne classification
 - $0 < CCC < 2 \rightarrow$ classification acceptable, mais à vérifier
 - $CCC < 0 \rightarrow$ présence de valeurs extrêmes / aberrantes qui biaisent l'analyse (surtout si $CCC < -30$)
- Chercher une chute en k classes suivi d'un pic en $k+1$ classes → bonne classification en $k+1$ classes
- Ne pas utiliser si le critère d'agrégation est lien simple



ipes.boutyour@gmail.com

186

186

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ **Avantages :**

- La CAH est simple et polyvalente puisqu'elle peut être appliquée à tout type d'objet, à condition de disposer d'une matrice de distance.
- Il existe de nombreuses approches et la CAH peut gérer des clusters de forme non sphérique (en utilisant la méthode du lien unique, par exemple).
- Le schéma de classification est un dendrogramme, c'est-à-dire un ensemble de partitions imbriquées. qui pourrait être plus informatif qu'une partition plate. De plus, en fonction du nombre de clusters que nous souhaitons, nous pouvons couper le diagramme de l'arbre en conséquence et obtenir une partition plate.

ipes.boutyour@gmail.com

187

187

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ **Inconvénients :**

- Dans le CAH, une fois que deux objets ont été regroupés, on ne peut pas les dégroupier plus tard au cours de l'algorithme.
- Complexité temporelle : étant donné qu'à chaque itération, nous devons trouver la distance la plus faible entre $n(n-1)$ paires de points de données, et qu'il y a n itérations, la complexité temporelle est de $O(n^3)$.
- Complexité de stockage : comme nous devons stocker la matrice de distance, sa complexité est de $O(n^2)$.
- En raison de leur complexité temporelle et de stockage, la plupart des algorithmes de CAH ne peuvent être utilisés sur de grands ensembles de données.
- → Pour surmonter ces limites, certaines approches récentes de Classification Hiérarchique (CH) ont été proposées, voir par exemple [Han et al., 2006, Murtagh et Contreras, 2012 | | Xu et Wunsch, 2005].

ipes.boutyour@gmail.com

188

188

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

○ Conseils de mise en place

- Bien choisir la distance entre les individus
- Bien choisir le critère d'agrégation
- Centrer-réduire les variables de préférence
- Ecarter les valeurs extrêmes/aberrantes de préférence
- Calculer plusieurs indicateurs pour déterminer le nombre de classes à retenir

ipes.boutyour@gmail.com

189

189

Clustering : Méthodes hiérarchiques

Classification Ascendante Hiérarchique (CAH)

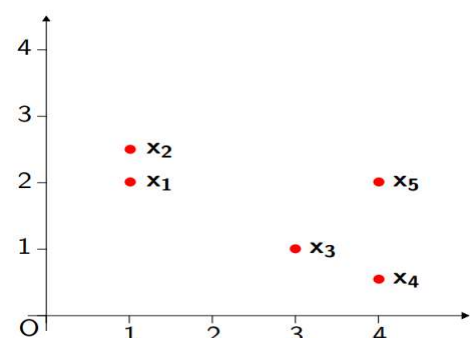
○ Exercice d'application

- On considère 5 points de données dans \mathbb{R}^2 :

- $X_1 = (1 ; 2)$
- $X_2 = (1 ; 2.5)$
- $X_3 = (3 ; 1)$
- $X_4 = (4 ; 0.5)$
- $X_5 = (4 ; 2)$

- On choisit le critère de distance euclidienne.

- Réaliser une CAH avec les différents critères d'agrégations et dessiner les dendrogrammes correspondants.




ipes.boutyour@gmail.com

190


190

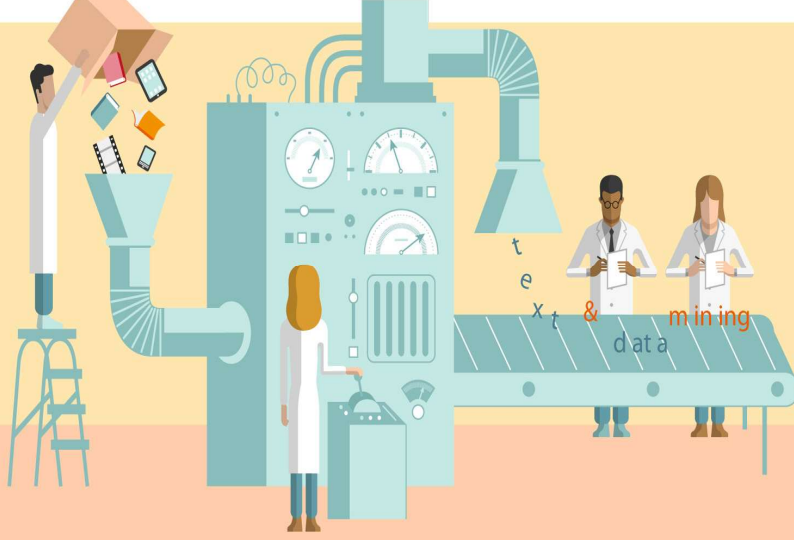


191



Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences
Département d'Informatique





Data Mining & Machine Learning

Master IPS
Faculté des sciences – Rabat
Université Mohamed V

192