

Faculté des Nouvelles Technologies de l'Information et de la Communication

Département : Informatique Fondamentale et ses Applications

Année Universitaire : 2023/2024

Module : DAMI (Master 1 SDIA)

TD N° 4.2 : Clustering – Classification Ascendante Hiérarchique

❖ **Exercice 1 :**

Considérons un ensemble de données représentant les moyennes globales de différents étudiants : E1(16,1), E2(13,3), E3(9,7), E4(11,5), E5(14,8), E6(12,5).

1. Effectuer une Classification Hiérarchique Ascendante sur l'ensemble des étudiants en utilisant la distance de Manhattan entre les instances et la distance minimale (single link method) entre clusters.

Il faut spécifier :

- Le tableau des distances à chaque regroupement d'instances ou de clusters (chaque itération).
 - Les données regroupées à chaque étape.
 - Le dendrogramme obtenu.
2. Effectuer une autre Classification Hiérarchique Ascendante sur l'ensemble des étudiants en utilisant la distance de Manhattan entre les instances, mais cette fois ci en utilisant la distance maximale (complet link method) entre clusters.

❖ **Exercice 2 :**

Supposons que nous disposions des données suivantes représentant la taille (en centimètres) et le poids (en kilogrammes) de huit objets :

	A	B	C	D	E	F	G	H
Taille (Cm)	160	165	152	170	155	180	168	162
Poids (Kg)	50	58	43	68	46	80	70	55

Utilisez la méthode de classification hiérarchique agglomérative avec la distance euclidienne pour regrouper ces objets en trois clusters. Calculez les distances entre les clusters à chaque étape et dessinez le dendrogramme correspondant.

Solution Exercise 1

CAH (Single link) :

Tableau des distances de la première itération :

	E1	E2	E3	E4	E5	E6
E1	0					
E2	2,8	0				
E3	6,4	3,6	0			
E4	4,6	1,8	1,8	0		
E5	1,3	1,5	5,1	3,3	0	
E6	3,6	0,8	2,8	1	2,3	0

Fusionnons les deux clusters les plus proches $G1 = \{E2 \text{ et } E6\}$

	E1	G1	E3	E4	E5
E1	0				
G1	2,8	0			
E3	6,4	2,8	0		
E4	4,6	1	1,8	0	
E5	1,3	1,5	5,1	3,3	0

Fusionnons les deux clusters les plus proches $G2 = \{G1, E4\}$

	E1	G2	E3	E5
E1	0			
G2	2,8	0		
E3	6,4	1,8	0	
E5	1,3	1,5	5,1	0

Fusionnons les deux clusters les plus proches $G3 = \{E1 \text{ et } E5\}$

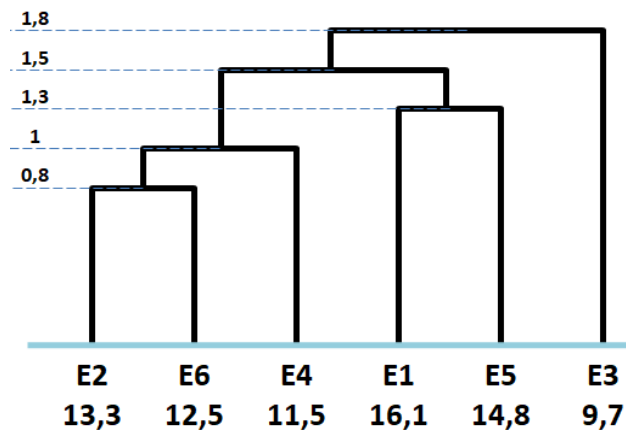
	G3	G2	E3
G3	0		
G2	1,5	0	
E3	5,1	1,8	0

Fusionnons les deux clusters les plus proches $G4 = \{G3, G2\}$

	G4	E3
G4	0	
E3	1,8	0

Fusionnons les deux clusters les plus proches $G5 = \{G3, G4\}$

Le dendrogramme (arbre hiérarchique) correspondant serait un diagramme qui montre les étapes de fusion des clusters et les distances associées.



CAH (Compleat link) :

Tableau des distances de la première itération :

	E1	E2	E3	E4	E5	E6
E1	0					
E2	2,8	0				
E3	6,4	3,6	0			
E4	4,6	1,8	1,8	0		
E5	1,3	1,5	5,1	3,3	0	
E6	3,6	0,8	2,8	1	2,3	0

Fusionnons les deux clusters les plus proches $G1=\{E2 \text{ et } E6\}$

	E1	G1	E3	E4	E5
E1	0				
G1	3,6	0			
E3	6,4	3,6	0		
E4	4,6	1,8	1,8	0	
E5	1,3	2,3	5,1	3,3	0

Fusionnons les deux clusters les plus proches $G2=\{E1 \text{ et } E5\}$

	G2	G1	E3	E4
G2	0			
G1	3,6	0		
E3	6,4	3,6	0	
E4	4,6	1,8	1,8	0

Fusionnons les deux clusters les plus proches (E3 et E4) ou ((E2, E6) et E4)

	E1, E5	E2, E6	E3, E4
E1, E5	0		
E2, E6	3,6	0	
E3, E4	6,4	3,6	0

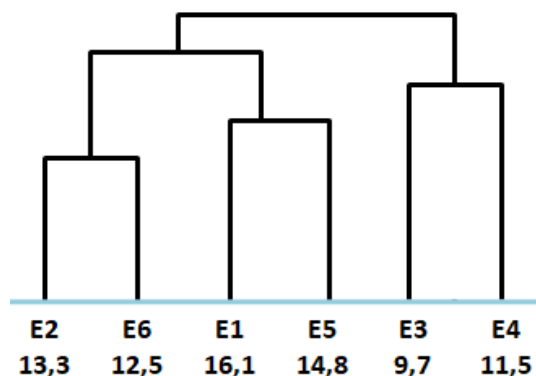
Fusionnons les deux clusters les plus proches ((E1, E5) et (E2, E6) ou (E2, E6) et (E3, E4))

	E1, E5, E2, E6	E3, E4
E1, E5, E2, E6	0	
E3, E4	6,4	0

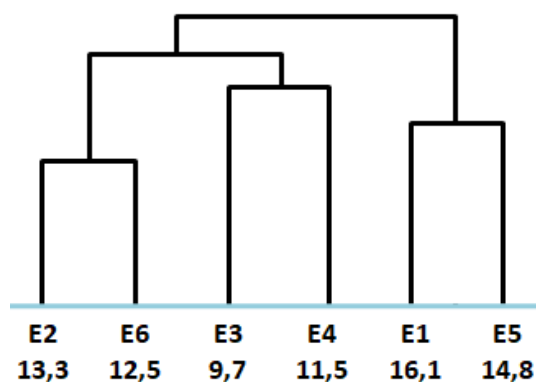
Ou

	E1, E5	E2, E6, E3, E4
E1, E5	0	
E2, E6, E3, E4	6,4	0

Le dendrogramme (arbre hiérarchique) correspondant serait un diagramme qui montre les étapes de fusion des clusters et les distances associées.



Ou



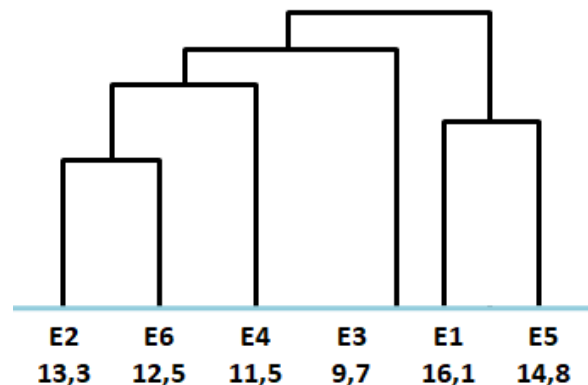
	E1, E5	E2, E6, E4	E3
E1, E5	0		
E2, E6, E4	4,6	0	
E3	6,4	3,6	0

Fusionnons les deux clusters les plus proches ((E2, E6, E4) et E3)

	E1, E5	E2, E6, E4, E3
E1, E5	0	
E2, E6, E4, E3	6,4	0

Fusionnons les deux clusters les plus proches ((E1, E5) et (E2, E6, E4, E3))

Le dendrogramme (arbre hiérarchique) correspondant serait un diagramme qui montre les étapes de fusion des clusters et les distances associées.



Solution Exercise 2

Nous calculons les distances euclidiennes entre tous les objets et commençons avec chaque objet comme un cluster individuel.

	A	B	C	D	E	F	G	H
A	0							
B	9,43	0						
C	10,63	19,85	0					
D	20,59	11,18	30,81	0				
E	6,40	15,62	4,24	26,63	0			
F	36,06	26,63	46,40	15,62	42,20	0		
G	21,54	12,37	31,39	2,83	27,30	15,62	0	
H	5,39	4,24	15,62	15,26	11,40	30,81	16,15	0

Fusionnons les deux clusters les plus proches (D et G)

Répétons les calculs des distances à chaque fois entre le nouveau cluster fusionné et les autres clusters jusqu'à ce qu'il ne reste qu'un seul cluster contenant tous les points.

NB : La définition de "proche" dépend de la mesure de distance utilisée (ex. : distance minimale (single link), maximale (complete link), moyenne (average link), etc.).

NB : Prenons l'exemple de la méthode Single Link où la distance entre deux clusters est déterminée par la distance la plus courte entre les points des deux clusters.

Supposons que nous avons fusionné les clusters (D) et (G) dans un nouveau cluster (DG). Maintenant, pour calculer la distance entre ce nouveau cluster (DG) et un autre point ou cluster, nous calculons la distance la plus courte entre n'importe quel point de (DG) et n'importe quel point du point ou du cluster en question.

Par exemple, si nous calculons la distance entre le nouveau cluster (DG) et le point (A), nous devons calculer la distance la plus courte entre tous les points de (DG) et le point (A). Ensuite, nous pouvons utiliser cette distance comme résultat de la distance entre (DG) et (A).

	A	B	C	D , G	E	F	H
A	0						
B	9,43	0					
C	10,63	19,85	0				
D, G	20,59	11,18	30,81	0			
E	6,40	15,62	4,24	26,63	0		
F	36,06	26,63	46,40	15,62	42,20	0	
H	5,39	4,24	15,62	15,26	11,40	30,81	0

Fusionnons les deux clusters les plus proches, soit (B et H) ou (C et E)

	A	B, H	C	D, G	E	F
A	0					
B, H	5,39	0				
C	10,63	15,62	0			
D, G	20,59	11,18	30,81	0		
E	6,40	11,40	4,24	26,63	0	
F	36,06	26,63	46,40	15,62	42,20	0

Fusionnons les deux clusters les plus proches (C et E)

	A	B, H	C, E	D, G	F
A	0				
B, H	5,39	0			
C, E	6,40	15,62	0		
D, G	20,59	11,18	26,63	0	
F	36,06	26,63	42,20	15,62	0

Fusionnons les deux clusters les plus proches (A et (BH))

	A, B, H	C, E	D, G	F
A, B, H	0			
C, E	6,40	0		
D, G	11,18	26,63	0	
F	26,63	42,20	15,62	0

Fusionnons les deux clusters les plus proches ((A, B, H) et (C, E))

	A, B, H, C, E	D, G	F
A, B, H, C, E	0		
D, G	11,18	0	
F	26,63	15,62	0

Fusionnons les deux clusters les plus proches ((A, B, H, C, E) et (D, G))

	A, B, H, C, E, D, G	F
A, B, H, C, E, D, G	0	
F	15,62	0

Fusionnons les deux clusters les plus proches ((A, B, H, C, E, D, G) et F)

Le dendrogramme (arbre hiérarchique) correspondant serait un diagramme qui montre les étapes de fusion des clusters et les distances associées.

Cluster 1 : {C, E, B, H, A}.

Cluster 2 : {D, G}.

Cluster 3 : {F}.

