Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques et Informatique

Module: Data Mining **Année: 2022-2023**
Nature de document: TD2-Solution
Niveau: L3-STID **M^me^ N.BERMAD**

Exercice1:

1. L'algorithme des centres mobiles

	X_1	X_2
W_1	-2	2
W_2	-2	-1
W_3	0	-1
W_4	2	2
W_5	-2	3
W_6	3	0

Etape0 :

La matrice des distances entre les individus et les centres C_1^0de coordonnées (1, 1) et C_2^0de coordonnées (2,3) est :

	W1	W2	W3	W4	W5	W6
c_1^0	3.16	3.16	2.24	1.41	3.60	2.24
C_2^0	4.12	5.66	4.47	1	4	3.16

En utilisant la distance euclidienne :

$$d(w1, c_1^0) = \sqrt{(-2-1)^2 + (2-1)^2} = 3.16$$

D'où les deux groupes :

A= {W1, W2, W3, W6, W5}, B= {W4}

Etape01:

On considère deux nouveaux centresC_1^1, C_2^1 lesquels sont les centres de gravité des deux groupes A et B. Donc :

$$C_1^1 = \left(\frac{-2+(-2)+0+(-2)+3}{5}, \frac{2+(-1)+(-1)+0}{5}\right) = (-0.6, 0.6)$$

$$C_2^1 = (2,2)$$

Le tableau des distances entre les individus et ces centres est :

	W1	**W2**	**W3**	**W4**	**W5**	**W6**
C_1^1	1.98	2.13	1.71	2.95	2.78	3.65
C_2^1	4	5	3.60	0	4.12	2.24

D'où les deux groupes :

A= {W1, W2, W3, W5}, B= {W4, W6}

Etape02 :

On considère deux nouveaux centresC_1^2, C_2^2 lesquels sont les centres de gravité des deux groupes A et B. Donc

$$C_1^2 = \left(\frac{-2+(-2)+0+(-2)}{4}, \frac{2+(-1)+(-1)+3}{4}\right) = (-1.5, 0.75)$$

$$C_2^2 = \left(\frac{2+3}{2}, \frac{2+0}{2}\right) = (2.5, 1)$$

Le tableau des distances entre les individus et ces centres est

	W1	**W2**	**W3**	**W4**	**W5**	**W6**
C_1^2	1.34	1.82	2.30	3.72	2.30	4.56
C_2^2	4.61	4.92	3.20	1.12	4.92	1.12

D'où les deux groupes :

A= {W1, W2, W3, W5}, B= {W4, W6}

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.

2. a. L'algorithme CAH

Etape0:

On va regrouper les individus avec l'algorithme CAH et la méthode du voisin le plus proche munie de la distance euclidienne. Au départ chaque individu est un cluster, donc :

$$P_0 = \{\{w1\}, \{w2\}, \{w3\}, \{w4\}, \{w5\}\}$$

- La matrice de distance associée à $P_0 = \{\{w1\},\{w2\},\{w3\},\{w4\},\{w5\}\}$ est :

	w1	**w2**	**w3**	**w4**	**w5**
w1	0				
w2	5.85	0			
w3	**1.41**	4.60	0		
w4	3.35	7.07	3.20	0	
w5	4.47	1.50	3.16	5.59	0

Les éléments (individus) w1 et w3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les rassemble pour former le groupe : A= {w1, w3}. On a une nouvelle partition de P :

$$P_1 = \{\{w2\},\{w4\},\{w5\},A\}$$

Etape1:

- La matrice de distance associée à P_1 est :

	w2	**w4**	**w5**	**A**
w2	0			
w4	7.07	0		
w5	**1.50**	5.59	0	
A	4.60	3.20	3.16	0

On a:

e(w2,A)=Min(e(w2,w1),e(w2,w3))=Min(5.85,4.60)=4.60

e(w4,A)=Min(e(w4,w1),e(w4,w3))=Min(3.35,3.20)=3.20

e(w5,A)=Min(e(w5,w1),e(w5,w3))=Min(4.47,3.16)=3.16

Les éléments (individus) w2 etw5 sont les plus proches. On les rassemble pour former le groupe : B = {w2, w5}. On a une nouvelle partition de P :

$$P_2 = \{\{w4\}, A, B\}$$

Etape2:

- La matrice de distance associée à P_2 est :

	w4	**A**	**B**
w4	0		
A	3.2 0	0	
B	5.59	**3.16**	0

On a:

e(B,w4)=Min(e(w2,w4),e(w5,w4))=Min(7.07,5.59)=5.59

Et

e(B,A)=Min(e(w2,A),e(w5,A))=Min(4.60,3.16)=5.27

Les éléments (individus) A et A sont les plus proches. On les rassemble pour former le groupe : C = {A, B}= {w1, w3, w2, w5}. On a une nouvelle partition de P :

$$P_3=\{w4, C\}$$

Etape3 :

- La matrice de distance associée à P_3 est :

	W4	C
W4	0	
C	**3.20**	0

On a :

e(C, w4)=Min(e (A,w4), e (B,w4))=Min (3.20, 5.59)=3.20

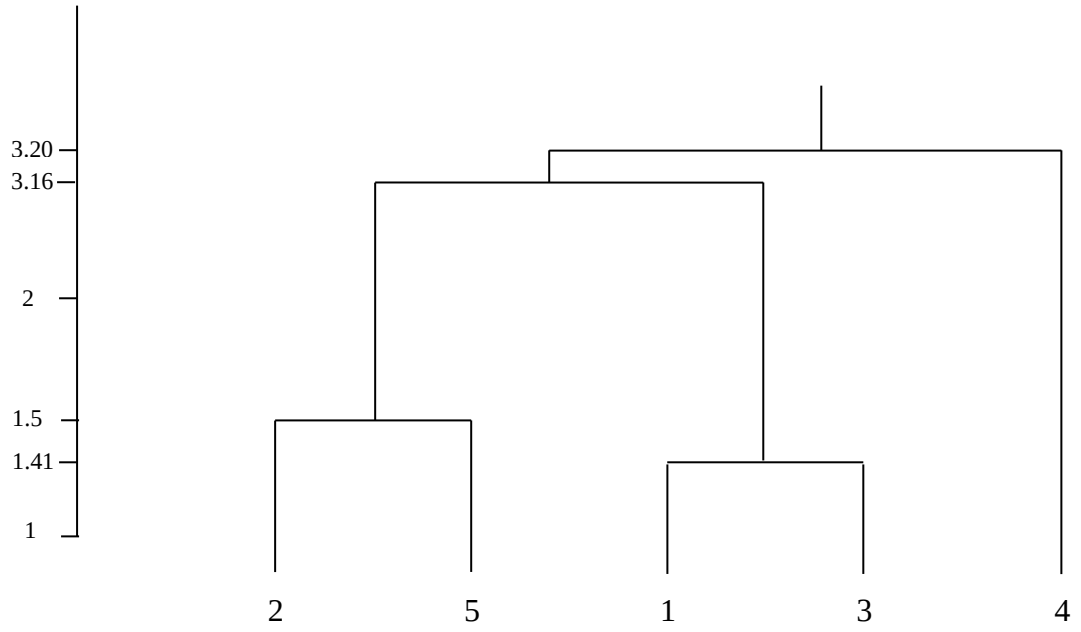
Il ne reste plus que 2 éléments, w4 et C ; on les regroupe. On obtient la partition $P_4=\{w1, w2, w3, w4, w5\} = P$. Cela termine l'algorithme de CAH.

2. b.Construction du dendrogramme

- Les individus {w1}, {w3} ont été regroupés avec un écart de **1.41**

- Les individus {w2}, {w5} ont été regroupés avec un écart de **1.50**
- Les individus A= {w1, w3} et B={w2,w5} ont été regroupés avec un écart de **3.16**
- Les individus C= {A, B} et w4 ont été regroupés avec un écart de **3.20**

On peut donc construire le dendrogramme associé:



Exercice 2 :

1. La table de contingence (TDC) associée :

	F	H	R	U	C	M	A
W_1	0	1	0	1	1	0	0
W_2	1	0	0	1	1	0	0
W_3	1	0	1	0	0	1	0
W_4	1	0	0	1	0	0	1
W_5	0	1	1	0	0	1	0
W_6	0	1	1	0	0	0	1

d (w_1, w_2):

		W_2		
		1	0	sum
W_1	1	$a_{1,2}$=2	$b_{1,2}$=1	a+b=3
	0	$c_{1,2}$=1	$d_{1,2}$=3	c+d=4
	sum	a+c=3	b+d=4	

On a:

$$a_{1,2} = 2, b_{1,2} = 1, c_{1,2} = 1, \text{et} d_{1,2} = 3.$$

Donc

d (w1, w2) = $\frac{b_{1,2}+c_{1,2}}{a_{1,2}+b_{1,2}+c_{1,2}} = \frac{2}{4} = 0.5$// La distance de Jaccard mésure la dissimilarité entre les individus(sim(w1,w2)=1-d(w1,w2)=1-0.5=0.5

d (w_3, w_6) :

On a:

$$a_{3,6} = 1, b_{3,6} = 2, c_{3,6} = 2, \text{et} d_{3,6} = 2.$$

Donc

d (w_3, w_6) = $\frac{b_{3,6}+c_{3,6}}{a_{3,6}+b_{3,6}+c_{3,6}} = \frac{4}{5} = 0.8$

2. Comme

d (w1, w2) <d (w_3, w_6)

w_1 est plus proche de w_2 que w_3 de w_6.