



CADERNO DE EXERCÍCIOS – PARTE 02

4 CAPÍTULO 04: AGRUPAMENTO DE DADOS

4.1 EXERCÍCIOS CONCEITUAIS

4.1.1 Qual é a definição e o objetivo da tarefa de agrupamento de dados?

Resposta: A tarefa de agrupamento de dados tem por objetivo descobrir grupos homogêneos de objetos utilizando métodos numéricos de análise de dados multivariados.

Ela pode ser definida como a organização de um conjunto de objetos, normalmente representados por pontos em um espaço multidimensional em grupos baseada na similaridade entre eles.

4.1.2 A avaliação da saída de um algoritmo de agrupamento de dados pode determinar a qualidade do agrupamento resultante. Para isso utilizamos medidas de avaliação de desempenho que são responsáveis por aferir quantitativamente o agrupamento resultante. As medidas podem ser categorizadas em dois tipos. Quais são os tipos de medidas de avaliação e como eles funcionam?

Resposta: Os dois tipos de medidas são: internas e externas.

- **Internas:** são medidas que utilizam apenas informações intrínsecas aos objetos do agrupamento baseando-se em medidas de similaridade e avaliando as distâncias intragrupos e/ou intergrupos.
- **Externas:** são medidas que avaliam quão correto está um agrupamento dado um agrupamento ideal que se deseja alcançar. O cálculo dessas medidas requer o conhecimento prévio do grupo ao qual cada objeto pertence.

4.1.3 Existem diversos algoritmos de agrupamento disponíveis, mas de forma abrangente estes algoritmos podem ser divididos em três categorias. Quais são estas categorias e suas características?

Resposta: As três categorias são: hierárquicos, particionais e não exclusivos. Abaixo estão as características de cada um.

- **Hierárquicos:** os métodos hierárquicos criam uma decomposição hierárquica dos dados. Esses métodos podem ser aglomerativos ou divisivos, baseados em como o processo de decomposição é efetuado.
- **Particionais:** a partir de um número n de partições, o método constrói k partições dos dados, sendo cada partição representa um cluster onde $k \leq n$.
- **Não Exclusivos:** esse método permite que um objeto pertença completamente ou parcialmente a mais de um grupo ao mesmo tempo.

4.2 EXERCÍCIOS NUMÉRICOS

4.2.1 Para a base de dados apresentada na tabela abaixo execute dois passos do algoritmo k -Médias, com $k = 2$ e distância Euclidiana. Considere os objetos 1 e 4 como centroides iniciais.

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
1	5,4	3,9	1,7	0,4
2	5,0	3,4	1,5	0,2
3	6,6	2,9	4,6	1,3
4	5,2	2,7	3,9	1,4

Resposta:

Centroides

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C1	1	5.4	3.9	1.7	0.4	(5.4, 3.9, 1.7, 0.4)
C2	4	5.2	2.7	3.9	1.4	(5.2, 2.7, 3.9, 1.4)

Cluster	Centroides
C1	(5.4, 3.9, 1.7, 0.4)
C2	(5.2, 2.7, 3.9, 1.4)

Iteração 1

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
2	5.0	3.4	1.5	0.2

$$\begin{aligned}
 d(x_2, c_1) &= \sqrt{(5.0 - 5.4)^2 + (3.4 - 3.9)^2 + (1.5 - 1.7)^2 + (0.2 - 0.4)^2} \\
 d(x_2, c_1) &= \sqrt{0.16 + 0.25 + 0.04 + 0.16} \\
 d(x_2, c_1) &= \sqrt{0.61} \\
 d(x_2, c_1) &= \mathbf{0.78}
 \end{aligned}$$

$$\begin{aligned}
 d(x_2, c_2) &= \sqrt{(5.0 - 5.2)^2 + (3.4 - 2.7)^2 + (1.5 - 3.9)^2 + (0.2 - 1.4)^2} \\
 d(x_2, c_2) &= \sqrt{0.4 + 0.49 + 5.76 + 1.44} \quad d(x_2, c_1) = \sqrt{0.61} \\
 d(x_2, c_1) &= 2.84
 \end{aligned}$$

Objeto	Cluster
1	C1
2	C1
3	
4	C2

Recalculando Centroide C1

Atual

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C1	1	5.4	3.9	1.7	0.4	(5.4, 3.9, 1.7, 0.4)

Novo

$$\begin{aligned}
 C_1 &= \left(\frac{5.4 + 5.0}{2}, \frac{3.9 + 3.4}{2}, \frac{1.7 + 1.5}{2}, \frac{0.4 + 0.2}{2} \right) \\
 C_1 &= (5.2, 3.65, 1.6, 0.3)
 \end{aligned}$$

Cluster	Centroides
C1	(5.2, 3.65, 1.6, 0.3)
C2	(5.2, 2.7, 3.9, 1.4)

Iteração 2

Objeto	Atributo A	Atributo B	Atributo C	Atributo D
3	6.6	2.9	4.6	1.3

$$d(x_3, c_1) = \sqrt{(6.6 - 5.2)^2 + (2.9 - 3.65)^2 + (4.6 - 1.6)^2 + (1.3 - 0.3)^2}$$

$$d(x_3, c_1) = \sqrt{1.96 + 0.56 + 9 + 1}$$

$$d(x_3, c_1) = \sqrt{12.52}$$

$$d(x_3, c_1) = 3.54$$

$$d(x_3, c_2) = \sqrt{(6.6 - 5.2)^2 + (2.9 - 2.7)^2 + (4.6 - 3.9)^2 + (1.3 - 1.4)^2}$$

$$d(x_3, c_2) = \sqrt{1.96 + 0.04 + 0.49 + 0.01}$$

$$d(x_3, c_2) = \sqrt{2.5}$$

$$d(x_3, c_2) = \mathbf{1.58}$$

Objeto	Cluster
1	C1
2	C1
3	C2
4	C2

Recalculando Centroide C2 Atual

Cluster	Objeto	Atributo A	Atributo B	Atributo C	Atributo D	Centroides
C2	4	5.2	2.7	3.9	1.4	(5.2, 2.7, 3.9, 1.4)

Novo

$$C_2 = \left(\frac{5.2 + 6.6}{2}, \frac{2.7 + 2.9}{2}, \frac{3.9 + 4.6}{2}, \frac{1.4 + 1.3}{2} \right)$$

$$C_2 = (5.9, 2.8, 4.25, 1.35)$$

Cluster	Centroides
C1	(5.2, 3.65, 1.6, 0.3)
C2	(5.9, 2.8, 4.25, 1.35)

Conclusão

Cluster com $k = 2$

Objeto	Cluster
1	C1
2	C1
3	C2
4	C2

Centroides finais

Cluster	Centroides
C1	(5.2, 3.65, 1.6, 0.3)
C2	(5.9, 2.8, 4.25, 1.35)

4.2.2 Para o agrupamento resultante do exercício anterior, determine o valor do índice de Dunn. Utilize a distância Euclidiana para identificar os objetos mais similares. Mostre passo-a-passo a realização do cálculo do índice.

Resposta:

Clusters

Cluster	Centroides
C1	(5.2, 3.65, 1.6, 0.3)
C2	(5.9, 2.8, 4.25, 1.35)

Cálculo da medida intragrupo: $Intra(g_i) = \max_{x,y \in g_i} \{d(x,y)\}$

$$\begin{aligned}
 x &= \max(C1) = 5.2 \\
 y &= \max(C2) = 5.9 \\
 d(x,y) &= \sqrt{(5.2 - 5.9)^2} \\
 d(x,y) &= 0.7 \\
 Intra(g_i) &= \mathbf{0.7}
 \end{aligned}$$

Cálculo da medida intergrupo: $Inter(g_i, g_j) = \frac{1}{|g_i| \times |g_j|} \sum d(x,y) \mid x \in g_i, y \in g_j$

$$\begin{aligned}
 d(x,y) &= \sqrt{(5.2 - 5.9)^2 + (3.65 - 2.8)^2 + (1.6 - 4.25)^2 + (0.3 - 1.5)^2} \\
 d(x,y) &= \sqrt{0.49 + 0.72 + 7.02 + 1.44} \quad d(x,y) = 3.11 \quad Inter(g_i, g_j) = \frac{1}{4 \times 4} \times 3.11 \\
 Inter(g_i, g_j) &= \mathbf{0.38}
 \end{aligned}$$

Cálculo do Índice de Dunn: $DU(g) = \min_{i=1,\dots,k} \left\{ \min_{j=1,\dots,k; i \neq j} \left\{ \frac{Inter(g_i, g_j)}{\max_{l=1,\dots,k} \{Intra(g_l)\}} \right\} \right\}$

$$\begin{aligned}
 DU(g) &= \frac{0.38}{0.7} \\
 DU(g) &= \mathbf{0.54}
 \end{aligned}$$

4.2.3 Aplique o método de agrupamento single-link e desenhe o dendrograma da matriz de distâncias abaixo.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1										
P2	14									
P3	10	16								
P4	25	27	15							
P5	28	26	18	7						
P6	24	22	18	17	18					
P7	11	15	7	16	21	15				
P8	24	20	14	21	22	10	19			
P9	20	30	14	15	14	20	15	20		
P10	22	30	16	23	26	20	21	12	14	

Resposta: **Cluster Inicial:** $[P4, P5] = 7$

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1										
P2	14									
P3	10	16								
P4	25	27	15							
P5	28	26	18	7						
P6	24	22	18	17	18					
P7	11	15	7	16	21	15				
P8	24	20	14	21	22	10	19			
P9	20	30	14	15	14	20	15	20		
P10	22	30	16	23	26	20	21	12	14	

$$d(P1, [P4P5])$$

$$\min(d(P1, P4), d(P1, P5))$$

$$\min(d(25, 28)) \Rightarrow 25$$

$$d(P2, [P4P5])$$

$$\min(d(P2, P4), d(P2, P5))$$

$$\min(d(27, 26)) \Rightarrow 26$$

$$d(P3, [P4P5])$$

$$\min(d(P3, P4), d(P3, P5))$$

$$\min(d(15, 18)) \Rightarrow 15$$

$$d([P4P5], P6)$$

$$\min(d(P4, P6), d(P5, P6))$$

$$\min(d(17, 18)) \Rightarrow 17$$

$$d([P4P5], P7)$$

$$\min(d(P4, P7), d(P5, P7))$$

$$\min(d(16, 21)) \Rightarrow 16$$

$$d([P4P5], P8)$$

$$\min(d(P4, P8), d(P5, P8))$$

$$\min(d(21, 22)) \Rightarrow 21$$

$$d([P4P5], P9)$$

$$\min(d(P4, P9), d(P5, P9))$$

$$\min(d(15, 14)) \Rightarrow 14$$

$$d([P4P5], P10)$$

$$\min(d(P4, P10), d(P5, P10))$$

$$\min(d(23, 26)) \Rightarrow 23$$

	P1	P2	P3	[P4 P5]	P6	P7	P8	P9	P10
P1									
P2	14								
P3	10	16							
[P4 P5]	25	26	15						
P6	24	22	18	17					
P7	11	15	7	16	15				
P8	24	20	14	21	10	19			
P9	20	30	14	14	20	15	20		
P10	22	30	16	23	20	21	12	14	

Próximo cluster: $[P3P7] = 7$

	P1	P2	P3	[P4 P5]	P6	P7	P8	P9	P10
P1									
P2	14								
P3	10	16							
[P4 P5]	25	26	15						
P6	24	22	18	17					
P7	11	15	7	16	15				
P8	24	20	14	21	10	19			
P9	20	30	14	14	20	15	20		
P10	22	30	16	23	20	21	12	14	

$$d(P1, [P3P7])$$

$$\min(d(P1, P3), d(P1, P7))$$

$$\min(d(10, 11)) \Rightarrow 10$$

$$d(P2, [P3P7])$$

$$\min(d(P2, P3), d(P2, P7))$$

$$\min(d(16, 15)) \Rightarrow 15$$

$$d([P3P7], [P4P5])$$

$$\min(d(P3, [P4P5]), d(P7, [P4P5]))$$

$$\min(d(15, 16)) \Rightarrow 15$$

$$d([P3P7], P6)$$

$$\min(d(P3, P6), d(P7, P6))$$

$$\min(d(18, 15)) \Rightarrow 15$$

$$d([P3P7], P8)$$

$$\min(d(P3, P8), d(P7, P8))$$

$$\min(d(14, 19)) \Rightarrow 14$$

$$d([P3P7], P9)$$

$$\min(d(P3, P9), d(P7, P9))$$

$$\min(d(14, 15)) \Rightarrow 14$$

$$d([P3P7], P10)$$

$$\min(d(P3, P10), d(P7, P10))$$

$$\min(d(16, 21)) \Rightarrow 16$$

	P1	P2	[P3 P7]	[P4 P5]	P6	P8	P9	P10
P1								
P2	14							
[P3 P7]	10	15						
[P4 P5]	25	26	15					
P6	24	22	15	17				
P8	24	20	14	21	10			
P9	20	30	14	14	20	20		
P10	22	30	16	23	20	12	14	

Próximo cluster: $[P1[P3P7]] = 10$

	P1	P2	[P3 P7]	[P4 P5]	P6	P8	P9	P10
P1								
P2	14							
[P3 P7]	10	15						
[P4 P5]	25	26	15					
P6	24	22	15	17				
P8	24	20	14	21	10			
P9	20	30	14	14	20	20		
P10	22	30	16	23	20	12	14	

$$d([P1[P3P7]], P2)$$

$$\min(d(P1, P2), d(P2, [P3P7]))$$

$$\min(d(14, 15)) \Rightarrow 14$$

$$d([P1[P3P7]], [P4P5])$$

$$\min(d(P1, [P4, P5]), d([P3P7], [P4P5]))$$

$$\min(d(25, 15)) \Rightarrow 15$$

$$d([P1[P3P7]], P6)$$

$$\min(d(P1, P6), d([P3P7], P6))$$

$$\min(d(24, 15)) \Rightarrow 15$$

$$d([P1[P3P7]], P8)$$

$$\min(d(P1, P8), d([P3P7], P8))$$

$$\min(d(24, 14)) \Rightarrow 14$$

$$\begin{aligned}
& d([P1[P3P7]], P9) \\
& \min(d(P1, P9), d([P3P7], P9)) \\
& \min(d(20, 14)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d([P1[P3P7]], P10) \\
& \min(d(P1, P10), d([P3P7], P10)) \\
& \min(d(22, 16)) \Rightarrow 16
\end{aligned}$$

	[P1 P3 P7]	P2	[P4 P5]	P6	P8	P9	P10
[P1 P3 P7]							
P2	14						
[P4 P5]	15	26					
P6	15	22	17				
P8	14	20	21	10			
P9	14	30	14	20	20		
P10	16	30	23	20	12	14	

Próximo cluster: $[P6P8] = 10$

	[P1 P3 P7]	P2	[P4 P5]	P6	P8	P9	P10
[P1 P3 P7]							
P2	14						
[P4 P5]	15	26					
P6	15	22	17				
P8	14	20	21	10			
P9	14	30	14	20	20		
P10	16	30	23	20	12	14	

$$\begin{aligned}
& d([P1P3P7], [P6P8]) \\
& \min(d([P1P3P7], P6), d([P1P3P7], P8)) \\
& \min(d(15, 14)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d(P2, [P6P8]) \\
& \min(d(P2, P6), d(P2, P8)) \\
& \min(d(22, 20)) \Rightarrow 20
\end{aligned}$$

$$\begin{aligned}
& d([P4P5], [P6P8]) \\
& \min(d([P4P5], P6), d([P4P5], P8)) \\
& \min(d(17, 21)) \Rightarrow 17
\end{aligned}$$

$$\begin{aligned}
& d([P6P8], P9) \\
& \min(d(P6, P9), d(P8, P9)) \\
& \min(d(20, 20)) \Rightarrow 20
\end{aligned}$$

$$\begin{aligned}
& d([P6P8], P10) \\
& \min(d(P6, P10), d(P8, P10)) \\
& \min(d(20, 12)) \Rightarrow 12
\end{aligned}$$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	P9	P10
[P1 P3 P7]						
P2	14					
[P4 P5]	15	26				
[P6 P8]	14	20	17			
P9	14	30	14	20		
P10	16	30	23	20	12	

Próximo cluster: $[P9P10] = 12$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	P9	P10
[P1 P3 P7]						
P2	14					
[P4 P5]	15	26				
[P6 P8]	14	20	17			
P9	14	30	14	20		
P10	16	30	23	20	12	

$$\begin{aligned}
& d([P1P3P7], [P9P10]) \\
& \min(d([P1P3P7], P9), d([P1P3P7], P10)) \\
& \min(d(14, 16)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d(P2, [P9P10]) \\
& \min(d(P2, P9), d(P2, P10)) \\
& \min(d(30, 30)) \Rightarrow 30
\end{aligned}$$

$$\begin{aligned}
& d([P4P5], [P9P10]) \\
& \min(d([P4P5], P9), d([P4P5], P10)) \\
& \min(d(14, 23)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d([P6P8], [P9P10]) \\
& \min(d([P6P8], P9), d([P6P8], P10)) \\
& \min(d(20, 20)) \Rightarrow 20
\end{aligned}$$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P3 P7]					
P2	14				
[P4 P5]	15	26			
[P6 P8]	14	20	17		
[P9 P10]	14	30	14	20	

Próximo cluster: $[[P1P3P7]P2] = 14$

	[P1 P3 P7]	P2	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P3 P7]					
P2	14				
[P4 P5]	15	26			
[P6 P8]	14	20	17		
[P9 P10]	14	30	14	20	

$$\begin{aligned}
& d([P1P3P7]P2, [P4P5]) \\
& \min(d([P1P3P7], [P4P5]), d(P2, [P4P5])) \\
& \min(d(15, 26)) \Rightarrow 15
\end{aligned}$$

$$\begin{aligned}
& d([P1P3P7]P2, [P6P8]) \\
& \min(d([P1P3P7], [P6P8]), d(P2, [P6P8])) \\
& \min(d(14, 20)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d([P1P3P7]P2, [P9P10]) \\
& \min(d([P1P3P7], [P9P10]), d(P2, [P9P10])) \\
& \min(d(14, 30)) \Rightarrow 14
\end{aligned}$$

	[P1 P2 P3 P7]	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P2 P3 P7]				
[P4 P5]	15			
[P6 P8]	14	14		
[P9 P10]	14	14	20	

Próximo cluster: $[P1P2P3P7][P6P8] = 14$

	[P1 P2 P3 P7]	[P4 P5]	[P6 P8]	[P9 P10]
[P1 P2 P3 P7]				
[P4 P5]	15			
[P6 P8]	14	14		
[P9 P10]	14	14	20	

$$\begin{aligned}
& d([P1P2P3P7][P6P8], [P4P5]) \\
& \min(d([P1P2P3P7], [P4P5]), d([P4P5], [P6P8])) \\
& \min(d(15, 14)) \Rightarrow 14
\end{aligned}$$

$$\begin{aligned}
& d([P1P2P3P7][P6P8], [P9P10]) \\
& \min(d([P1P2P3P7], [P9P10]), d([P6P8], [P9P10])) \\
& \min(d(14, 20)) \Rightarrow 14
\end{aligned}$$

	[P1 P2 P3 P6 P7 P8]	[P4 P5]	[P9 P10]
[P1 P2 P3 P6 P7 P8]			
[P4 P5]	14		
[P9 P10]	14	14	

Próximo cluster: $[P1P2P3P4P6P7P8][P4P5] = 14$

$$\begin{aligned}
& d([P1P2P3P6P7P8][P4P5], [P9P10]) \\
& \min(d([P1P2P3P6P7P8], [P9P10]), d([P4P5], [P9P10])) \\
& \min(d(14, 14)) \Rightarrow 14
\end{aligned}$$

	[P1 P2 P3 P6 P7 P8]	[P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8]		
[P9 P10]	14	

Próximo cluster: $[P1P2P3P4P5P6P7P8][P9P10] = 14$

	[P1 P2 P3 P4 P5 P6 P7 P8]	[P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8]		
[P9 P10]	14	

	[P1 P2 P3 P4 P5 P6 P7 P8 P9 P10]
[P1 P2 P3 P4 P5 P6 P7 P8 P9 P10]	

Dendrograma

