CHƯƠNG 3: PHÂN CỤM DỮ LIỆU

3.1. KHÁI NIỆM VỀ PHÂN CỤM DỮ LIỆU

3.2. ĐỘ ĐO SỬ DỤNG TRONG PHÂN CỤM

3.3. PHÂN CỤM DỮ LIỆU VỚI GIẢI THUẬT K-MEANS

(Phân cụm từ trên xuống)

3.4. PHÂN CỤM DỮ LIỆU VỚI GIẢI THUẬT HAC

(Phân cụm từ dưới lên)

3.5. SO SÁNH GIẢI THUẬT K-MEANS VÀ HAC

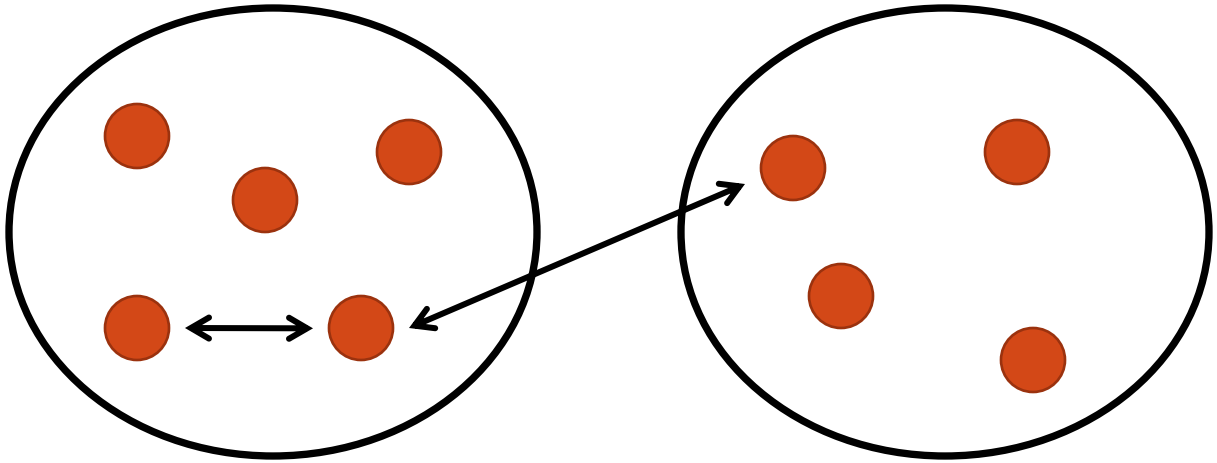
3.1. KHÁI NIỆM VỀ PHÂN CỤM DỮ LIỆU

3.1.1. Phân cụm dữ liệu (clustering) là gì?

- Phân cụm dữ liệu là quá trình phân chia các đối tượng dữ liệu (bản ghi) vào các nhóm (cụm) sao cho các đối tượng thuộc về cùng một cụm thì có các đặc điểm "tương tự" nhau ("gần" nhau) và các đối tượng thuộc về các cụm khác nhau thì có các đặc điểm "khác" nhau ("xa" nhau).

Đại lượng nào xác định sự "tương tự" và "khác" nhau giữa các đối tượng?

- Khác với phân lớp, phân cụm được xem quá trình học không có giám sát (unsupervised learning). Dữ liệu được phân vào các cụm mà không cần có tập mẫu học (training sample).



3.1.2. Ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu có thể ứng dụng trong nhiều lĩnh vực:

- **Nghiên cứu thị trường** (***Marketing***): Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng lớn, phân loại và dự đoán hành vi khách hàng,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn.
- **Sinh học** (***Biology***): Phân nhóm động vật và thực vật dựa vào các thuộc tính của chúng.
- **Quản lý thư viện** (***Libraries***): Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…
- **Tài chính, Bảo hiểm** (***Finance and Insurance***): Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds).
- **Khai phá web (*Web Mining*)**: Phân loại tài liệu (document classification), phân loại người dùng web (clustering weblog),…

3.2. ĐỘ ĐO SỬ DỤNG TRONG PHÂN CỤM

- Để xác định tính chất tương đồng giữa các đối tượng dữ liệu, người ta thường sử dụng khái niệm "khoảng cách" (distance).
- Hai đối tượng có "khoảng cách" càng nhỏ thì càng "tương tự" (giống) nhau và có "khoảng cách" càng lớn thì càng "khác" nhau.

Xét hai đối tượng dữ liệu (bản ghi) r_i và r_j , mỗi đối tượng có n thuộc tính:

$$r_i = (x_{i1}, x_{i2}, ..., x_{in})$$

$$r_j = (x_{j1}, x_{j2}, ..., x_{jn})$$

Khoảng cách Euclid (*Euclidean Distance*):

$$d(r_i, r_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2}$$

Khoảng cách Manhattan (*Manhattan Distance*):

$$d(r_i, r_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{in} - x_{jn}|$$

3.3. PHÂN CỤM VỚI GIẢI THUẬT K-MEANS

3.3.1. Khái niệm về trọng tâm cụm

Xét cụm dữ liệu C_j gồm m đối tượng thuộc cụm: $C_j = \{r_1, r_2, r_3, ..., r_m\}$

Mỗi đối tượng có n thuộc tính: $r_i = (x_{i1}, x_{i2}, x_{i3}, ..., x_{in})(1 \le i \le m)$

Trọng tâm cụm (mean/centroid) là đối tượng m_j được xác định:

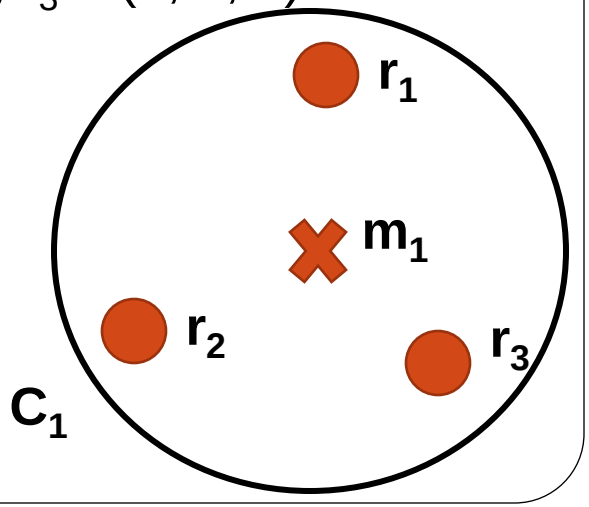
$$m_j = \left(\frac{1}{m}\sum_{i=1}^{m} x_{i1}, \frac{1}{m}\sum_{i=1}^{m} x_{i2}, ..., \frac{1}{m}\sum_{i=1}^{m} x_{in} \right)$$

Ví dụ:

Cho cụm $C_1 = \{r_1, r_2, r_3\}$ với $r_1 = (1, 2, 1)$, $r_2 = (1, 3, 2)$, $r_3 = (1, 1, 3)$.

Trọng tâm cụm là:

$$m_1 = \left(\frac{1+1+1}{3}, \frac{2+3+1}{3}, \frac{1+2+3}{3} \right) = (1, 2, 2)$$



3.3.2. Nội dung giải thuật K-means

Input: Tập dữ liệu D gồm m đối tượng dữ liệu (bản ghi): $r_1, r_2, \ldots, r_m$.

Số lượng cụm k.

Output: k cụm dữ liệu.

Begin

Chọn ngẫu nhiên k đối tượng làm trọng tâm cho k cụm;

Repeat

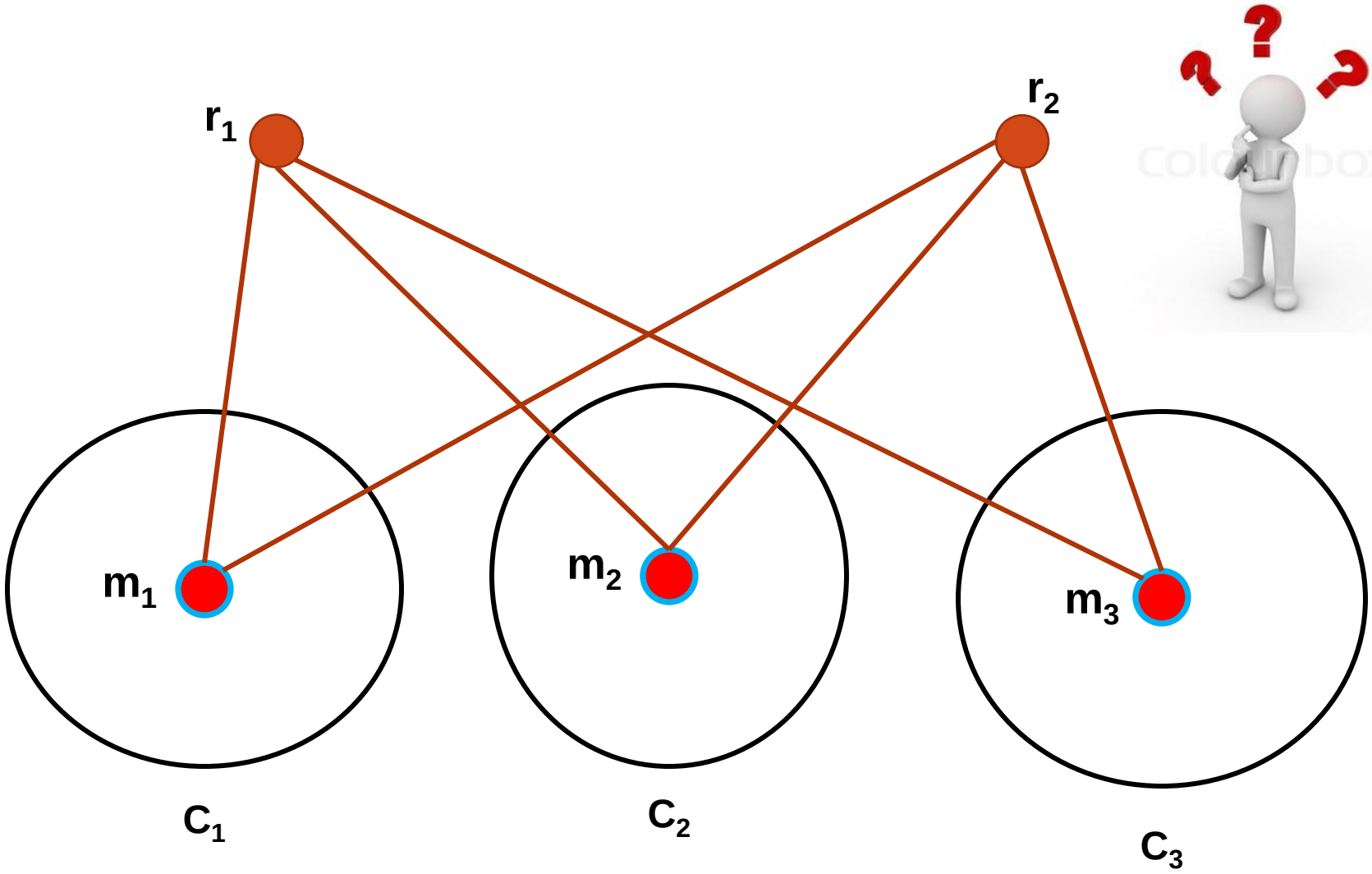
Gán mỗi đối tượng r_i cho cụm mà khoảng cách từ đối tượng đến trọng tâm cụm là nhỏ nhất trong số k cụm;

Xác định lại trọng tâm cho mỗi cụm dựa trên các đối tượng được gán cho cụm;

Until *(Không còn sự thay đổi);*

End;

(xem "Fundamentals of Database Systems – 4th Edition" trang 680)



$d(r_1,m_1) < d(r_1,m_2) < d(r_1,m_3) \rightarrow r_1$ thuộc C_1

$d(r_2,m_3) < d(r_2,m_2) < d(r_2,m_1) \rightarrow r_2$ thuộc C_3

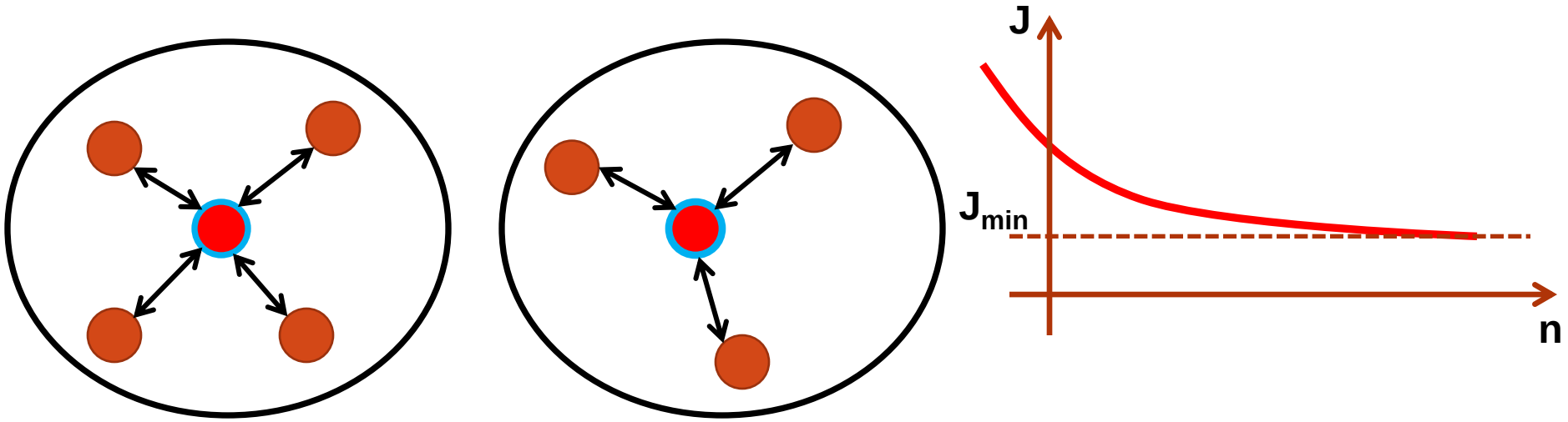
3.3.3. Điều kiện dừng của giải thuật K-means

Có hai kết cục có thể xảy ra đối với giải thuật K-means:

Giải thuật hội tụ: **không còn sự phân chia lại** các đối tượng giữa các cụm, hay **trọng tâm các cụm là không đổi**. Lúc đó tổng các tổng khoảng cách nội tại từ các đối tượng thuộc cụm đến trọng tâm cụm là cực tiểu:

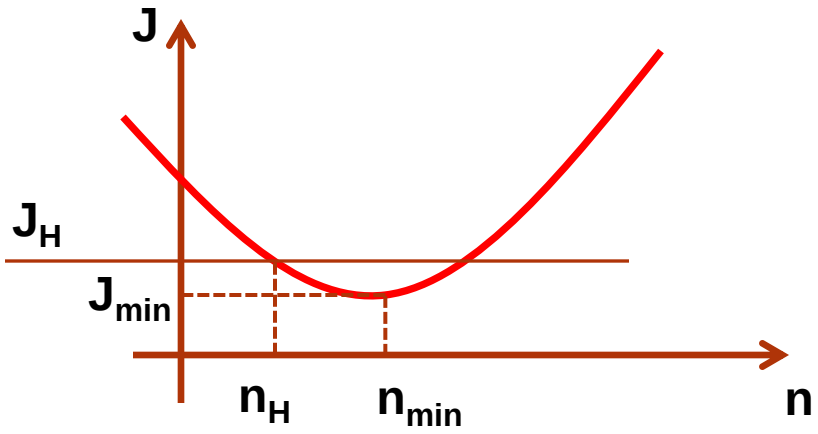
$$J = \sum_{j=1}^{k} \sum_{r_i \in C_j} d\left(r_i, m_j\right) \rightarrow \min$$

Đây là điều kiện dừng "lý tưởng".



Giải thuật không hội tụ: trọng tâm của các cụm cứ liên tục thay đổi. Lúc đó có 3 lựa chọn:

- Dừng giải thuật khi số lượng vòng lặp vượt quá một ngưỡng nào đó định trước.
- Dừng giải thuật khi giá trị J nhỏ hơn một ngưỡng nào đó định trước.



- Dừng giải thuật khi hiệu giá trị của J trong hai vòng lặp liên tiếp nhỏ hơn một ngưỡng nào đó định trước: $|J_{n+1} - J_n| < \varepsilon$

BÀI TẬP ÁP DỤNG

Bài tập số 1: Cho tập dữ liệu D như sau:

	X_1	X_2
r_1	1	2
r_2	2	2
r_3	2	3
r_4	3	3
r_5	3	4
r_6	2	4

Hãy phân cụm tập dữ liệu D với k = 2.

Chọn $m_1 = r_1 = (1, 2)$, $m_2 = r_6 = (2, 4)$.

Lần lặp 1:

$r_2 = (2, 2)$

$d(r_2, m_1) = |2 - 1| + |2 - 2| = 1$, $d(r_2, m_2) = |2 - 2| + |2 - 4| = 1 \Longrightarrow r_2 \in C_1$

$r_3 = (2, 3)$

$d(r_3, m_1) = |2 - 1| + |3 - 2| = 2$, $d(r_3, m_2) = |2 - 2| + |3 - 4| = 1 \Longrightarrow r_3 \in C_2$

$r_4 = (3, 3)$

$d(r_4, m_1) = |3 - 1| + |3 - 2| = 3$, $d(r_4, m_2) = |3 - 2| + |3 - 4| = 2 \Longrightarrow r_4 \in C_2$

$r_5 = (3, 4)$

$d(r_5, m_1) = |3 - 1| + |4 - 2| = 4$, $d(r_5, m_2) = |3 - 2| + |4 - 4| = 1 \Longrightarrow r_5 \in C_2$

Ta thu được 2 cụm: $C_1 = \{r_1, r_2\}$ và $C_2 = \{r_3, r_4 , r_5 , r_6\}$

Cập nhật trọng tâm cụm:

$$m_1 = \left(\frac{1+2}{2}, \frac{2+2}{2}\right) = (1.5, 1)$$

$$m_2 = \left(\frac{2+3+3+2}{4}, \frac{3+3+4+4}{4}\right) = (2.5, 3.5)$$

Với $m_1 = (1.5, 2)$, $m_2 = (2.5, 3.5)$

Lần lặp 2:

$r_1 = (1, 2)$

$d(r_1, m_1) = |1 - 1.5| + |2 - 2| = 0.5$, $d(r_1, m_2) = |1 - 2.5| + |2 - 3.5| = 3 \Longrightarrow r_1 \in C_1$

$r_2 = (2, 2)$

$d(r_2, m_1) = |2 - 1.5| + |2 - 2| = 0.5$, $d(r_2, m_2) = |2 - 2.5| + |2 - 3.5| = 2 \Longrightarrow r_2 \in C_1$

$r_3 = (2, 3)$

$d(r_3, m_1) = |2 - 1.5| + |3 - 2| = 1.5$, $d(r_3, m_2) = |2 - 2.5| + |3 - 3.5| = 1 \Longrightarrow r_3 \in C_2$

$r_4 = (3, 3)$

$d(r_4, m_1) = |3 - 1.5| + |3 - 2| = 2.5$, $d(r_4, m_2) = |3 - 2.5| + |3 - 3.5| = 1 \Longrightarrow r_4 \in C_2$

$r_5 = (3, 4)$

$d(r_5, m_1) = |3 - 1.5| + |4 - 2| = 3.5$, $d(r_5, m_2) = |3 - 2.5| + |4 - 3.5| = 1 \Longrightarrow r_5 \in C_2$

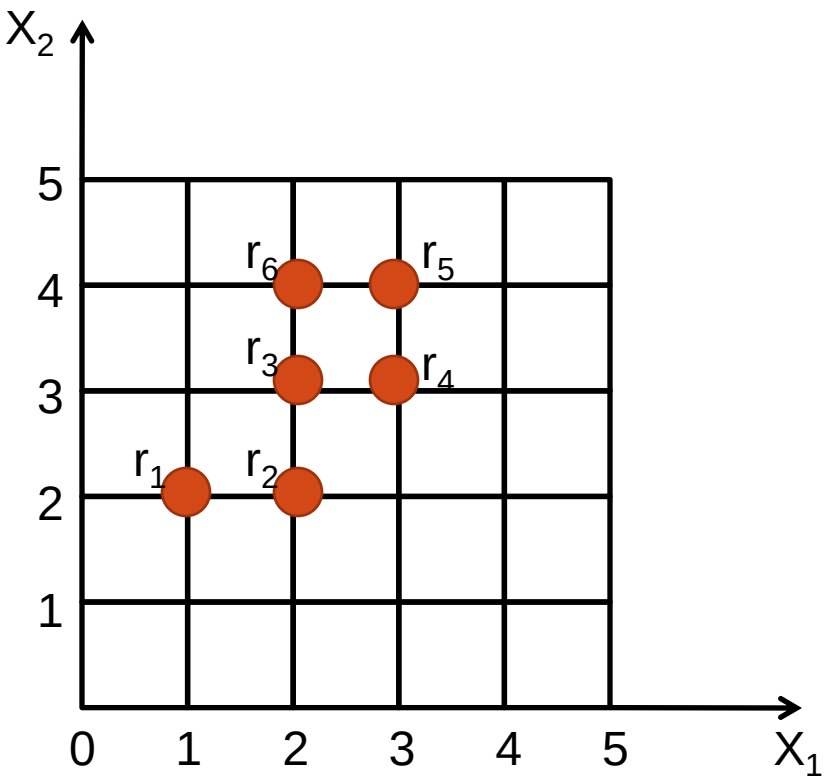
$r_6 = (2, 4)$

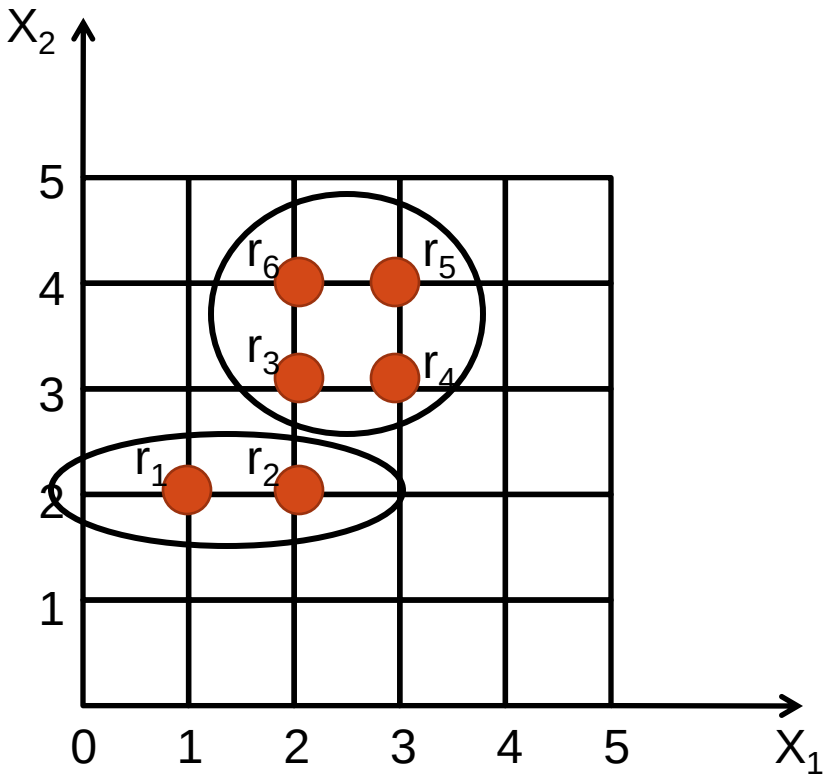
$d(r_6, m_1) = |2 - 1.5| + |4 - 2| = 2.5$, $d(r_6, m_2) = |2 - 2.5| + |4 - 3.5| = 1 \Longrightarrow r_6 \in C_2$

Ta thu được hai cụm $C_1 = \{r_1, r_2\}$ và $C_2 = \{r_3, r_4 , r_5 , r_6\}$.

Sau lần lặp 2 không có sự phân bố lại các đối tượng giữa các cụm (điều kiện dừng lý tưởng). Giải thuật kết thúc và kết quả của quá trình phân cụm là:

$$C_1 = \{r_1, r_2\} \text{ và } C_2 = \{r_3, r_4 , r_5 , r_6\}.$$


X2
5
4
3
2
1
r6
r5
r3
r4
r1
r2
0
1
2
3
4
5
X1


X2
5
4
3
2
1
r6
r5
r3
r4
r1
r2
0
1
2
3
4
5
X1

Bài tập số 2: Cho tập dữ liệu D như sau:

	X_1	X_2
A	1	1
B	2	1
C	4	3
D	5	4

Hãy phân cụm tập dữ liệu D với k = 2.

Chọn $m_1 = A = (1, 1)$, $m_2 = C = (4, 3)$.

Lần lặp 1:

$B = (2, 1)$

$d(B, m_1) = |2 - 1| + |1 - 1| = 1$, $d(B, m_2) = |2 - 4| + |1 - 3| = 4 \Longrightarrow B \in C_1$

$D = (5, 4)$

$d(D, m_1) = |5 - 1| + |4 - 1| = 7$, $d(D, m_2) = |5 - 4| + |4 - 3| = 2 \Longrightarrow D \in C_2$

Ta thu được 2 cụm: $C_1 = \{A, B\}$ và $C_2 = \{C, D\}$

Cập nhật trọng tâm cụm:

$$m_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5, 1) \qquad m_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5, 3.5)$$

Lần lặp 2: $m_1 = (1.5,1)$, $m_2 = (4,5, 3.5)$

$A = (1, 1)$

$d(A, m_1) = |1 - 1.5| + |1 - 1| = 0.5$, $d(A, m_2) = |1 - 4.5| + |1 - 3.5| = 6 \Longrightarrow A \in C_1$

$B = (2, 1)$

$d(B, m_1) = |2 - 1.5| + |1 - 1| = 0.5$, $d(B, m_2) = |2 - 4.5| + |1 - 3.5| = 5 \Longrightarrow B \in C_1$

C = (4, 3)

$d(C, m_1) = |4 - 1.5| + |3 - 1| = 4.5$, $d(C, m_2) = |4 - 4.5| + |3 - 3.5| = 1 \Longrightarrow C \in C_2$

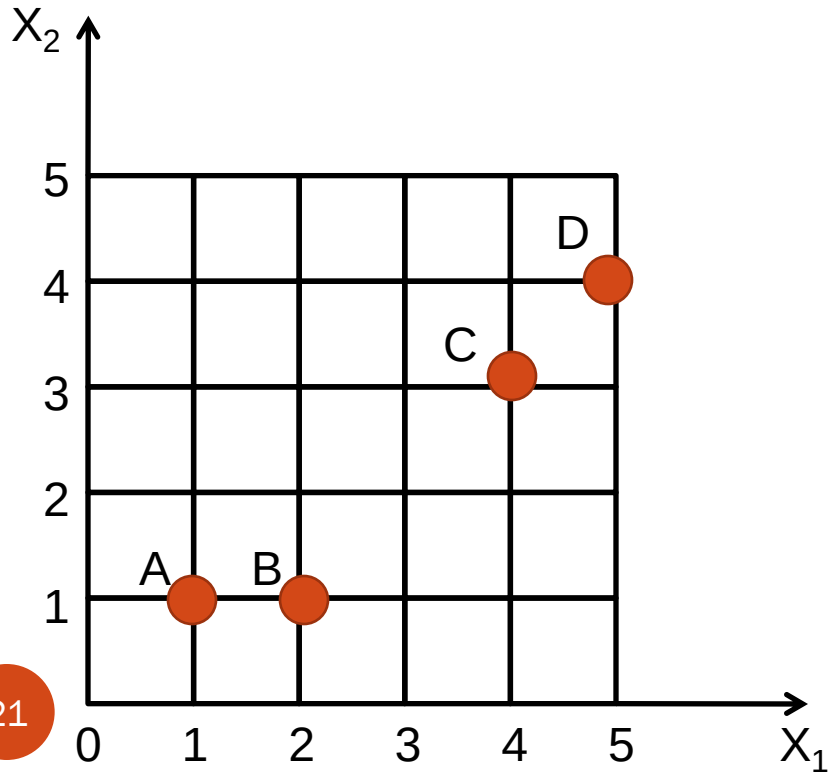
D = (5, 4)

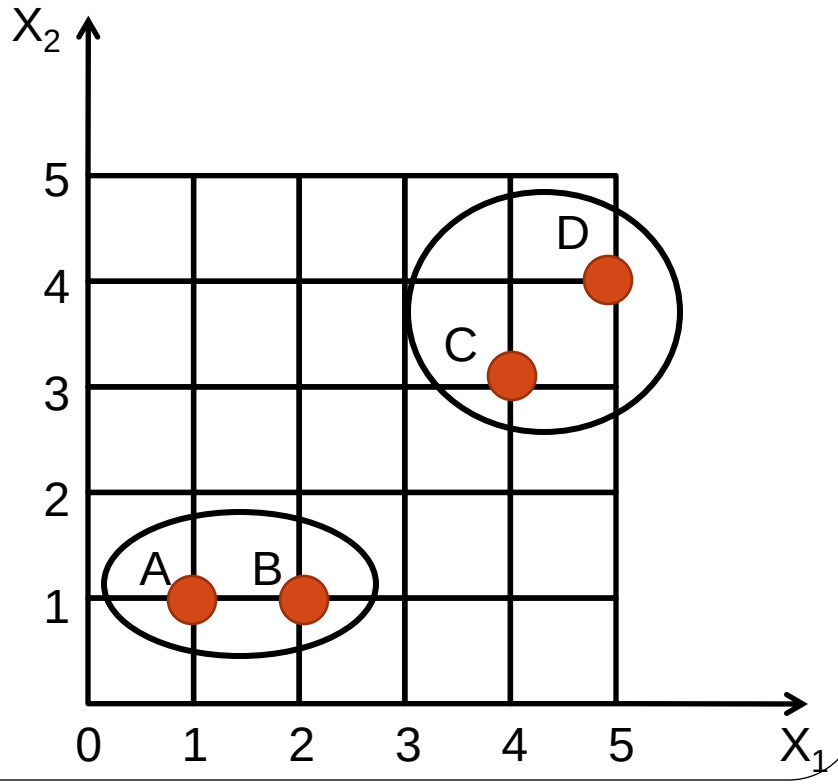
$d(D, m_1) = |5 – 1.5| + |4 - 1| = 6.5$, $d(D, m_2) = |5 - 4.5| + |4 - 3.5| = 1 \Longrightarrow D \in C_2$

Ta thu được 2 cụm: $C_1 = \{A, B\}$ và $C_2 = \{C, D\}$

Sau lần lặp 2 không có sự phân bố lại các đối tượng giữa các cụm (điều kiện dừng lý tưởng). Giải thuật kết thúc và kết quả của quá trình phân cụm là:

$C_1 = \{A, B\}$ và $C_2 = \{C, D\}$





3.4. PHÂN CỤM VỚI GIẢI THUẬT HAC
(HAC - Hierarchical Agglomerative Clustering)

3.4.1. Nội dung giải thuật HAC

Tích tụ dần "từ dưới lên" (Bottom-Up)

Tư tưởng giải thuật:

1. Ban đầu, mỗi đối tượng (bản ghi) dữ liệu được coi là một cụm.
2. Từng bước kết hợp các cụm đã có thành các cụm lớn hơn với yêu cầu là khoảng cách giữa các đối tượng trong nội bộ cụm là nhỏ.
3. Dừng thuật toán khi đã đạt số lượng cụm mong muốn, hoặc chỉ còn một cụm duy nhất chứa tất cả các đối tượng hoặc thỏa mãn điều kiện dừng nào đó.

G: tập các cụm.

D: tập các đối tượng (bản ghi) dữ liệu cần phân cụm.

k: số lượng cụm mong muốn.

d_o: ngưỡng khoảng cách giữa 2 cụm.

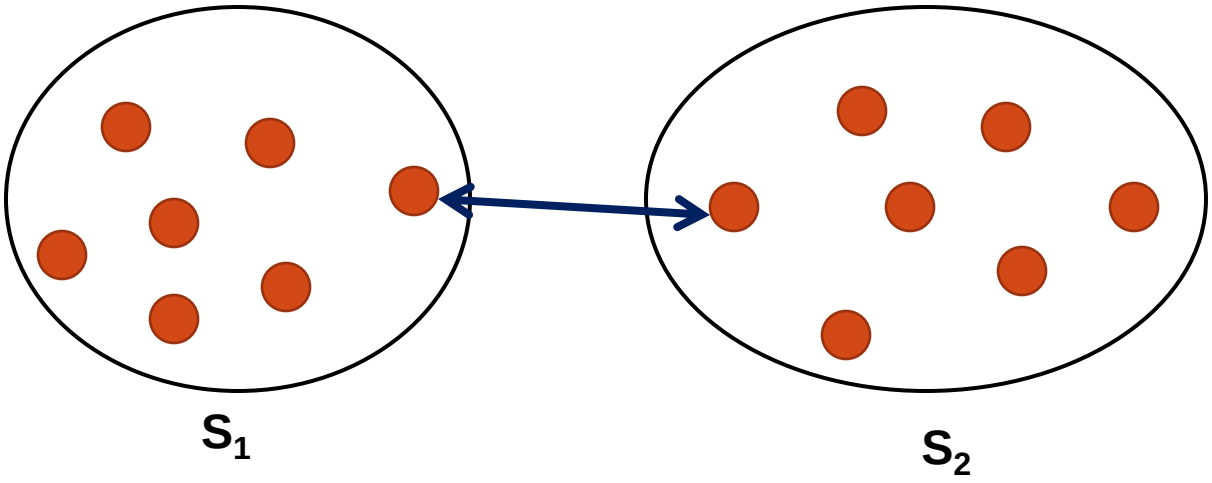
1. **G = {{r} | r ∈ D};** *//Khởi tạo G là tập các cụm chỉ gồm 1 đối tượng*
2. **Nếu |G| = k thì dừng thuật toán;** *//Đã đạt số lượng cụm mong muốn*
3. **Tìm hai cụm S_i , S_j ∈ G có khoảng cách $d(S_i, S_j)$ là nhỏ nhất;**
4. **Nếu $d(S_i, S_j) > d_o$ thì dừng thuật toán;** *//Khoảng cách 2 cụm gần nhất đã lớn hơn ngưỡng cho phép*
5. **G = G\{S_i, S_j};** *//Loại bỏ 2 cụm S_i ,S_j khỏi tập các cụm*
6. **S = S_i ∪ S_j;** *//Ghép S_i, S_j thành cụm mới S*
7. **G = G ∪ {S};** *//Kết nạp cụm mới vào G*
8. **Nhảy về bước 2.**

3.4.2. Độ đo "khoảng cách" giữa 02 cụm

A. Độ đo khoảng cách gần nhất (single-link)

Khoảng cách giữa 02 cụm được xác định là khoảng cách giữa 02 phần tử "gần" nhau nhất của 02 cụm đó:

$$d(S_1,S_2) = \min_{r_i \in S_1,\, r_j \in S_2} d(r_i,r_j)$$

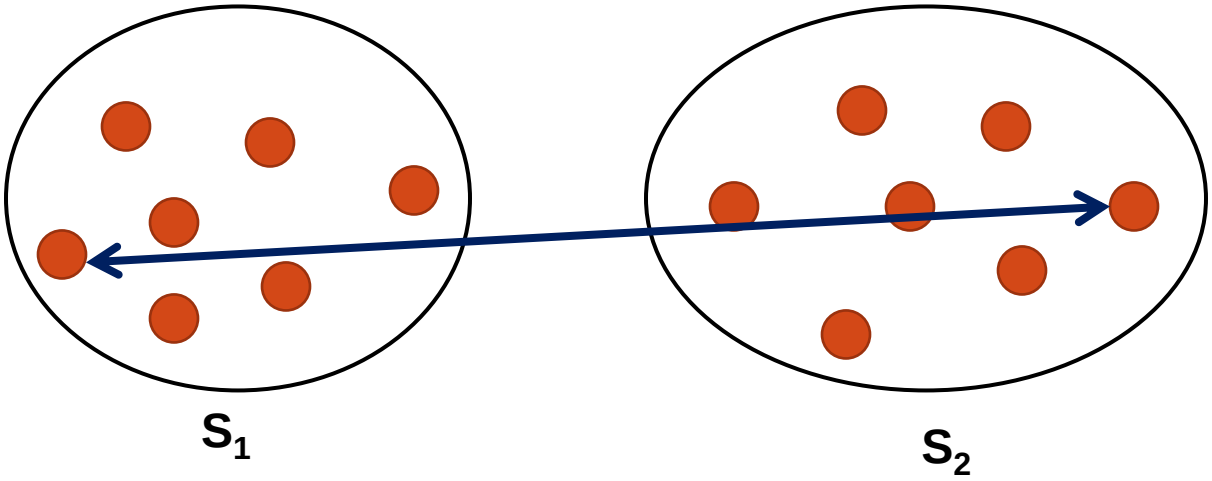


3.4.2. Độ đo "khoảng cách" giữa 02 cụm

B. Độ đo khoảng cách xa nhất (complete-link)

Khoảng cách giữa 02 cụm được xác định là khoảng cách giữa 02 phần tử "xa" nhau nhất của 02 cụm đó:

$$d(S_1,S_2) = \max_{r_i \in S_1,\, r_j \in S_2} d(r_i,r_j)$$

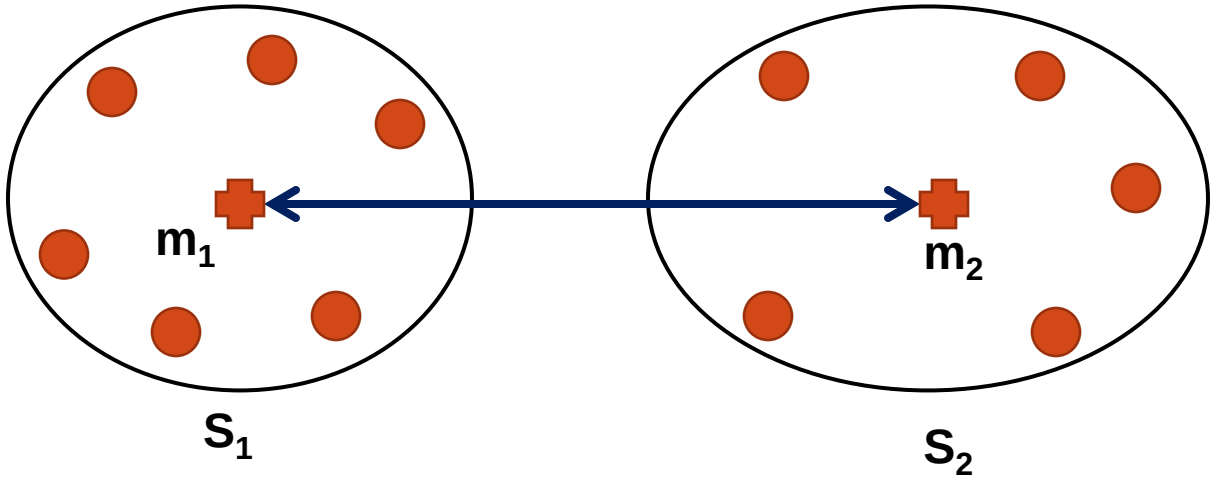


3.4.2. Độ đo "khoảng cách" giữa 02 cụm

C. Độ đo khoảng cách trọng tâm (centroid-link)

Khoảng cách giữa 02 cụm được xác định là khoảng cách giữa 02 trọng tâm của 02 cụm đó:

$$d(S_1, S_2) = d(m_1, m_2)$$

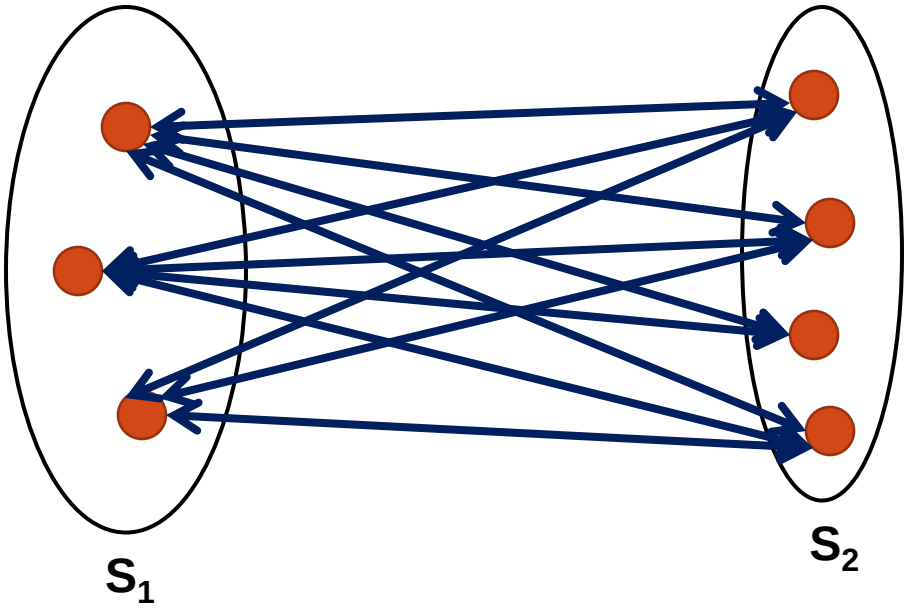


3.4.2. Độ đo "khoảng cách" giữa 02 cụm

D. Độ đo khoảng cách trung bình nhóm (group-average)

Khoảng cách giữa 02 cụm được xác định là khoảng cách trung bình giữa các phần tử thuộc về 02 cụm đó:

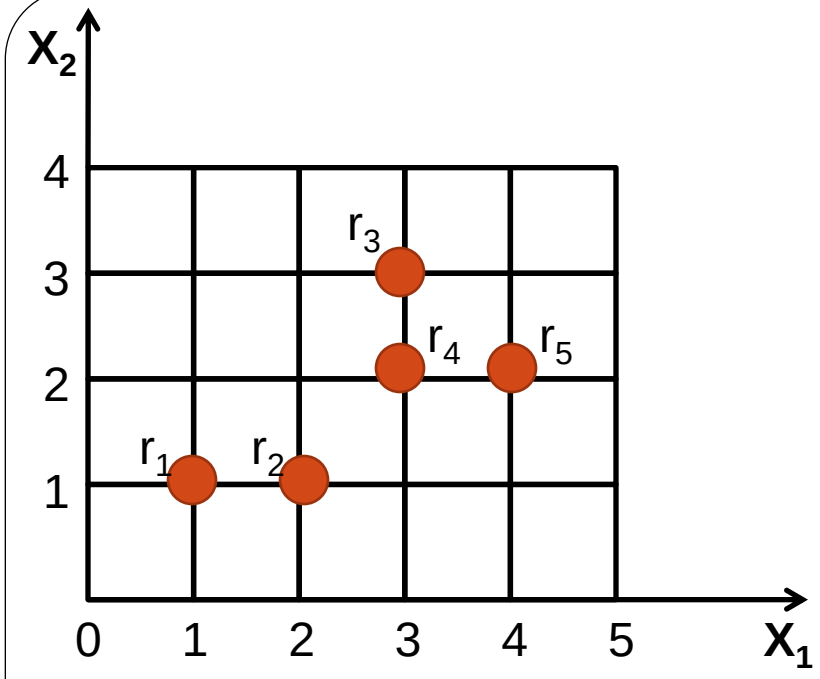
$$d(S_1, S_2) = \frac{1}{|S_1||S_2|} \sum_{r_i \in S_1, r_j \in S_2} d(r_i, r_j)$$



Ví dụ: Cho tập dữ liệu D gồm các bản ghi:

r	X_1	X_2
1	1	1
2	2	1
3	3	3
4	3	2
5	4	2

Xét 2 cụm dữ liệu $C_1 = \{r_1, r_2\}$, $C_2 = \{r_3, r_4, r_5\}$. Xác định khoảng cách $d(C_1,C_2)$ giữa 2 cụm dựa trên các độ đo khác nhau.



Ma trận khoảng cách:

	r_3	r_4	r_5
r_1	4	3	4
r_2	3	2	3

Nếu sử dụng single-link:

$d(C_1, C_2) = d(r_2, r_4) = 2$

Nếu sử dụng complete-link:

$d(C_1, C_2) = d(r_1, r_3) = d(r_1, r_5) = 4$

Nếu sử dụng group-average-link:

$d(C_1, C_2) = 19/6 = 3.17$

Nếu sử dụng centroid-link:

$$m_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(\frac{3}{2}, 1\right) \qquad m_2 = \left(\frac{3+3+4}{3}, \frac{3+2+2}{3}\right) = \left(\frac{10}{3}, \frac{7}{3}\right)$$

$$d(C_1, C_2) = d(m_1, m_2) = \left|\frac{3}{2} - \frac{10}{3}\right| + \left|1 - \frac{7}{3}\right| = \frac{19}{6}$$

3.4.2. Độ đo "khoảng cách" giữa 02 cụm

E. Nhận xét về các độ đo

Với độ đo single-link:

- ***Mang tính chất cục bộ: Chỉ quan tâm đến những vùng mà ở đó có phần tử của 2 cụm gần nhau nhất, không quan tâm đến các phần tử khác trong cụm cũng như cấu trúc tổng thể của các cụm.***
- ***Chất lượng phân cụm kém khi chỉ có 2 phân tử trong 2 cụm là rất gần nhau trong khi các phần tử khác ở phân tán rất xa nhau.***

Với độ đo complete-link:

- ***Khoảng cách 2 cụm dựa trên khoảng cách 2 phần tử xa nhau nhất $\Longrightarrow$ Việc ghép 2 cụm sẽ tạo ra cụm mới có đường kính nhỏ nhất.***
- ***Chất lượng phân cụm kém khi 2 phần tử trong 2 cụm ở rất xa nhau nhưng thực tế trọng tâm 2 cụm lại ở rất gần nhau.***

Với độ đo group-average:

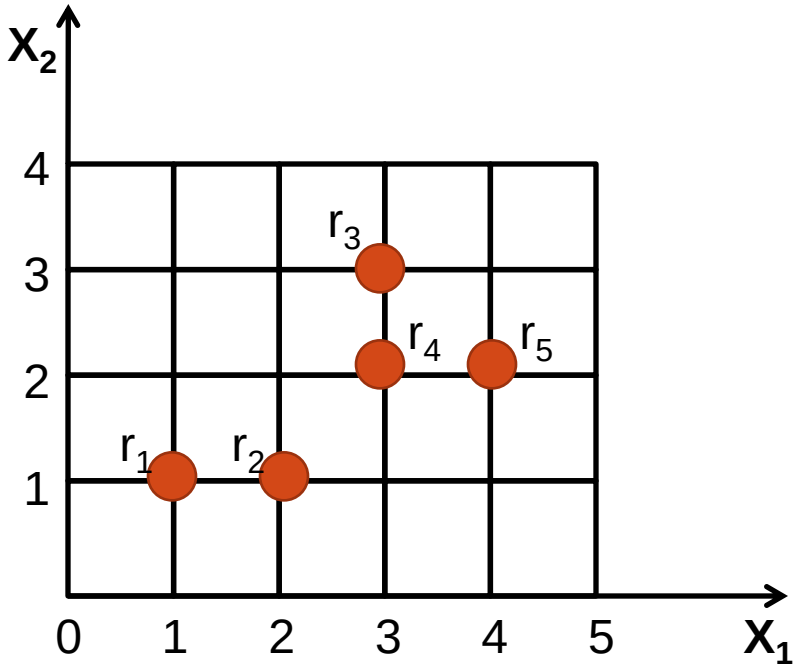
- *Tính toán khoảng cách của 2 cụm dựa trên khoảng cách của toàn bộ các cặp phần tử trong 2 cụm chứ không chỉ dựa trên một cặp phần tử duy nhất ⟹ tránh được nhược điểm của single-link và complete-link.*

Với độ đo centroid-link:

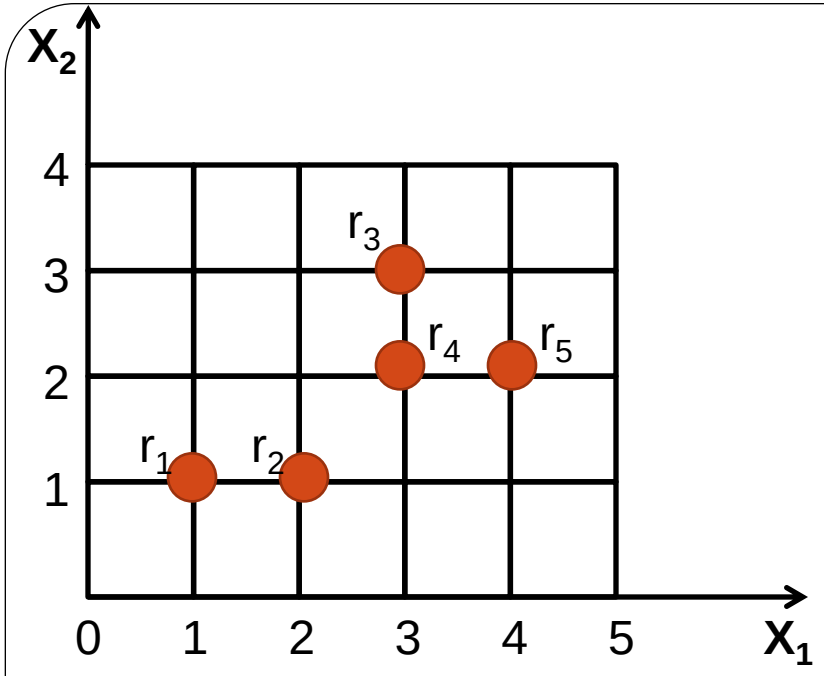
- *Khắc phục được nhược điểm của single/complete-link.*
- *Vẫn có nhược điểm là khoảng cách giữa các cụm khi từ đi mức dưới lên mức trên của cây phân cấp có thể là không tăng dần (do trong tâm các cụm ở mức cao nhiều khi gần nhau hơn các cụm ở mức dưới) ⟹ Trái với giả thiết về độ kết dính "Các cụm nhỏ thường có độ kết dính cao hơn các cụm có kích thước lớn hơn".*

Ví dụ: Cho tập dữ liệu gồm các đối tượng với 02 thuộc tính X_1, X_2 như sau:

r	X_1	X_2
1	1	1
2	2	1
3	3	3
4	3	2
5	4	2



Áp dụng giải thuật HAC hãy phân chia tập dữ liệu trên thành 02 cụm. Biết khoảng cách giữa 02 đối tượng được đo bằng độ đo Manhattan và khoảng cách giữa 02 cụm sử dụng độ đo single-link.



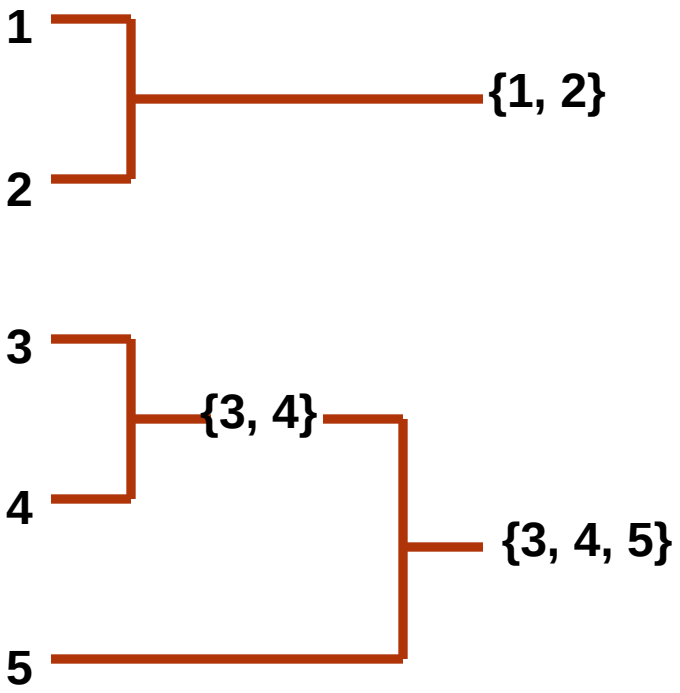
	1	2	3	4	5
1	0	1	4	3	4
2		0	3	2	3
3			0	1	2
4				0	1
5					0

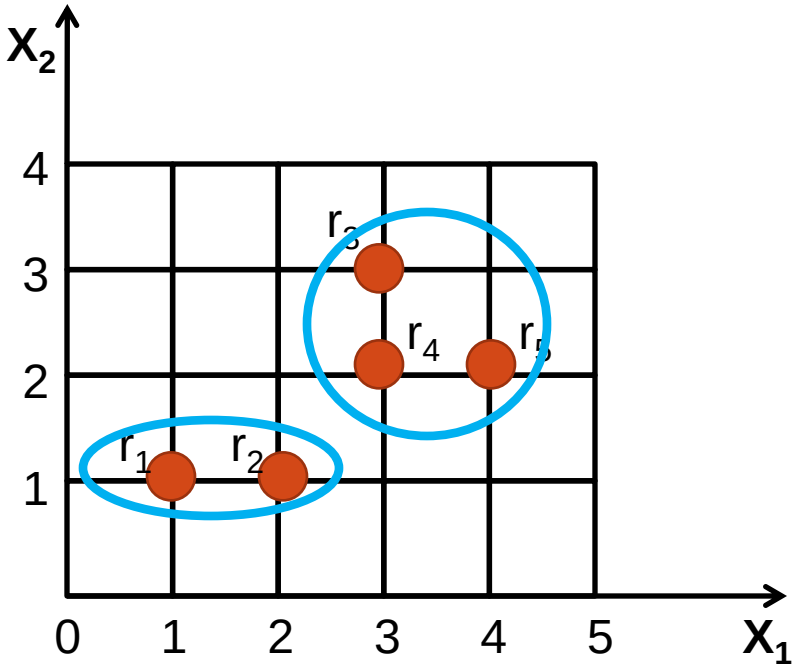
	{1,2}	3	4	5
{1,2}	0	3	2	3
3		0	1	2
4			0	1
5				0

	{1,2}	{3,4}	5
{1,2}	0	2	3
{3,4}		0	1
5			0

Ghép {3,4} với {5} thu được 02 cụm là {1,2} và {3,4,5}

Đã đạt số lượng cụm cần thiết. Kết thúc thuật toán


1
2
{1, 2}
3
4
{3, 4}
{3, 4, 5}
5


X2
4
3
2
1
r3
r4
r5
r1
r2
0
1
2
3
4
5
X1

3.5. SO SÁNH GIẢI THUẬT K-MEANS VÀ HAC

GIẢI THUẬT HAC	GiẢI THUẬT K-MEANS
Độ phức tạp thuật toán ❖ Độ phức tạp thuật toán là $O(N^2)$ trong đó N là số đối tượng được phân cụm. **Ưu, nhược điểm** ***Ưu điểm***: ❖ Khái niệm đơn giản. ❖ Lý thuyết tốt. ❖ Khi cụm được trộn hay tách thì quyết định là vĩnh cửu vì thế các phương pháp khác nhau cần được xem xét được rút giảm. ***Nhược điểm***: ❖ Quyết định trộn tách các cụm là vĩnh cửu nên thuật toán không có tính quay lui, nếu có quyết định sai thì không thể khắc phục lại. ❖ Độ phức tạp thuật toán cao, thời gian thực hiện phân cụm lâu. **Áp dụng tạo cây phân cấp** ❖ Thuật toán tạo ra cây phân cấp ngay trong quá trình phân cụm.	**Độ phức tạp thuật toán** ❖ Độ phức tạp thuật toán O(NkT) trong đó N là số đối tượng được phân cụm, k số cụm và T là số vòng lặp trong quá trình phân cụm. ❖ Thường T, k << N nên ta có thể coi độ phức tạp của thuật toán là O(N). **Ưu, nhược điểm** ***Ưu điểm***: ❖ Tính mở rộng cao, phù hợp với lượng dữ liệu lớn. ❖ Thời gian thực hiện thuật toán ít. ❖ Kết thúc ở điểm tối ưu cục bộ, có thể dùng thuật toán di truyền để tìm tối ưu toàn cục. ***Nhược điểm***: ❖ Cần chỉ định trước k cụm. ❖ Không thể xử lý dữ liệu chuỗi và ngoại lệ. ❖ Không phù hợp với miền dữ liệu không lồi hay cụm có kích thước khác nhau. ❖ Chỉ thực hiện tốt khi xác định được trị số trung bình của các đối tượng. **Áp dụng tạo cây phân cấp** ❖ Tạo ra cây phân cấp từng bước một. ❖ Tạo cây phân cấp ở mức một sau khi tiến hành phân cụm lần một bộ dữ liệu lớn. ❖ Tiếp tục tạo chủ đề mức hai và các mức sau sau khi tiếp tục tiến hành phân cụm cho bộ dữ liệu thuộc từng chủ đề con. ❖ Cây phân cấp được tạo ra bằng cách kết hợp các lần tiến hành phân cụm.

Q & A