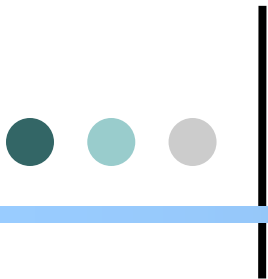


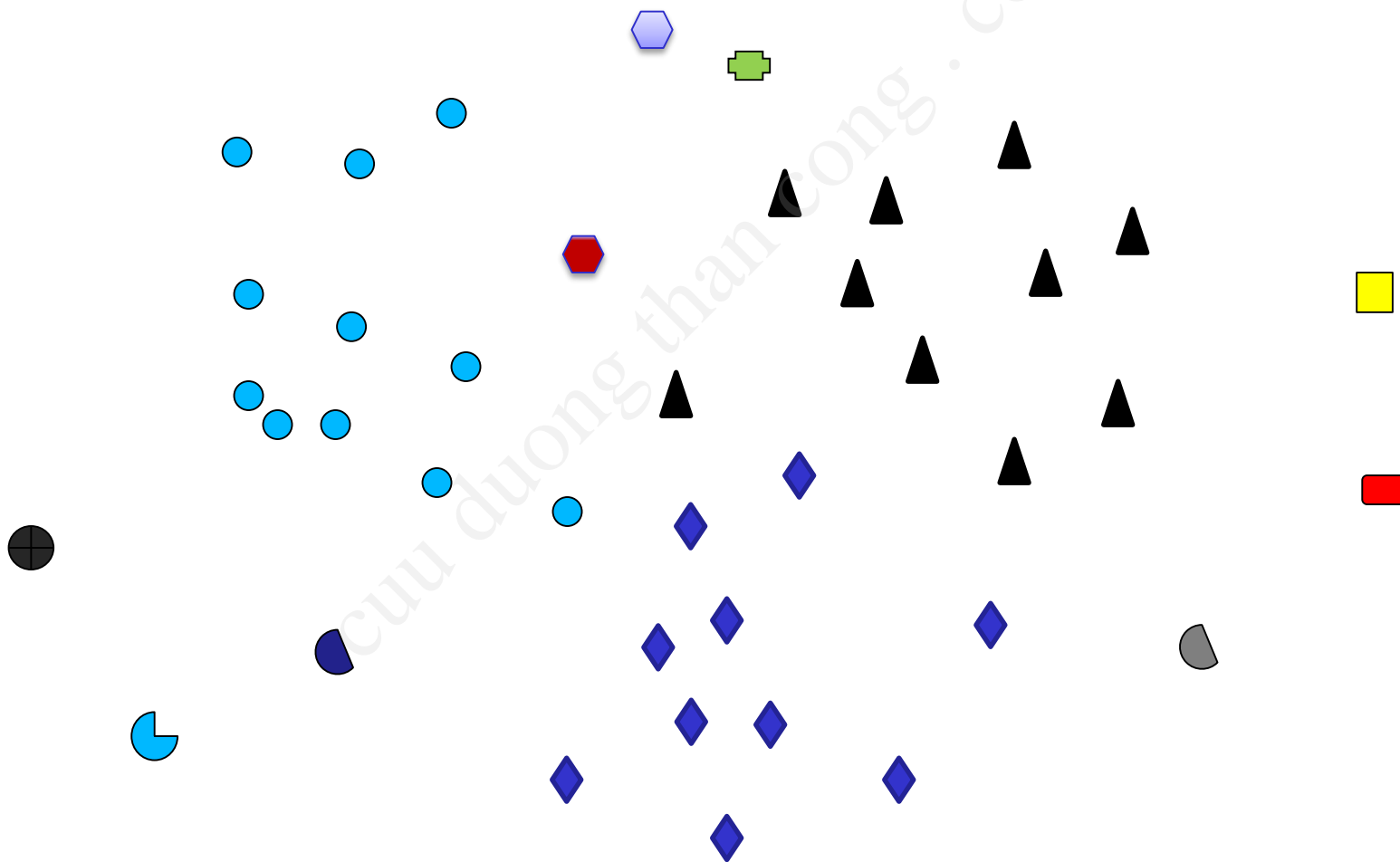
Chương 4

Khai phá dữ liệu

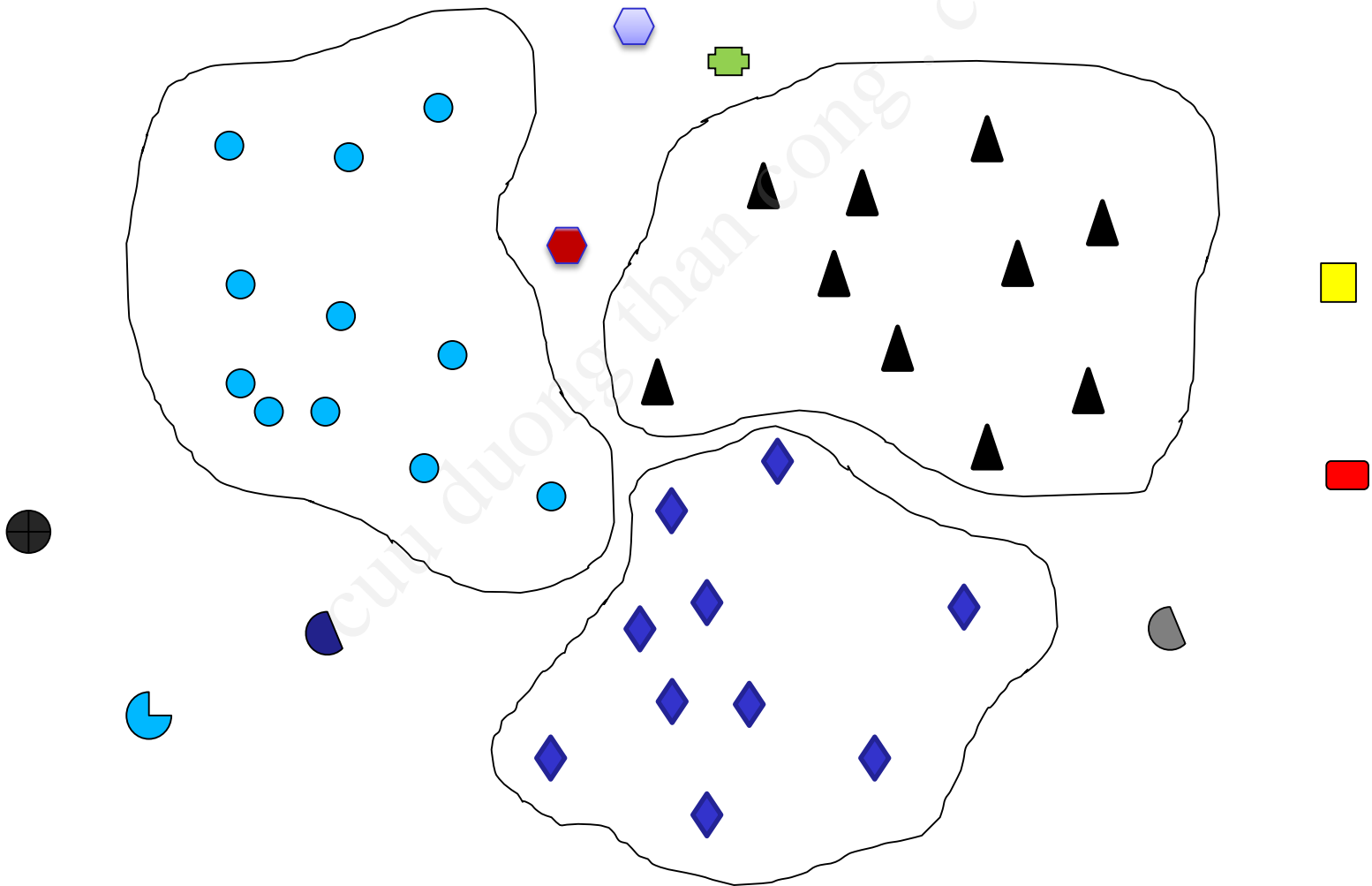
1. Tiền xử lý dữ liệu.
2. Phương pháp khai phá bằng luật kết hợp.
3. Phương pháp cây quyết định.
4. Các phương pháp phân cụm.
5. Các phương pháp khai phá dữ liệu phức tạp.



Gom cụm dữ liệu

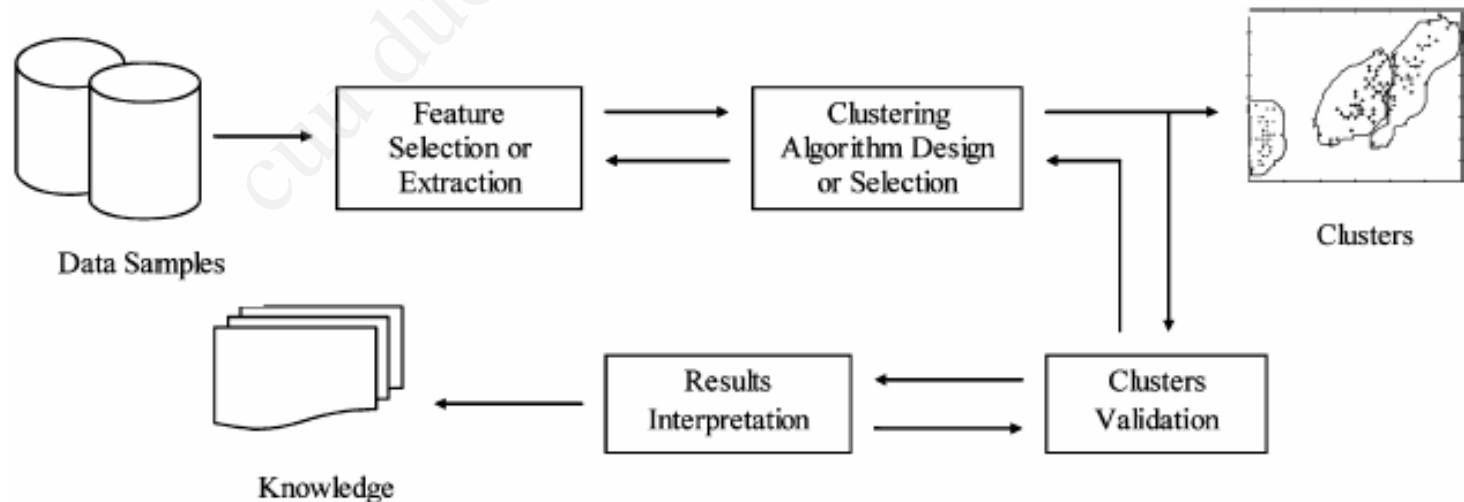


Gom cụm dữ liệu



Gom cụm dữ liệu

- ❖ Gom cụm: **Clustering**
- ❖ Dữ liệu phát sinh trong quá trình tác nghiệp gọi là dữ liệu thô,
- ❖ Để có thể khai phá các khía cạnh khác của dữ liệu chúng cần phải biến đổi về dạng thích hợp,



Độ đo trong gom cụm dữ liệu

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

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

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

➤ Khoảng cách Euclidean

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

➤ Khoảng cách Manhattan

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

Độ đo trong gom cụm dữ liệu

Trọng tâm cụm (mean/centroid):

Cụm C có m phần tử; mỗi phần tử có n thuộc tính:

$$C = \{r_1, r_2, \dots, r_m\},$$

$$R_i = (x_{i1}, x_{i2}, \dots, x_{in}).$$

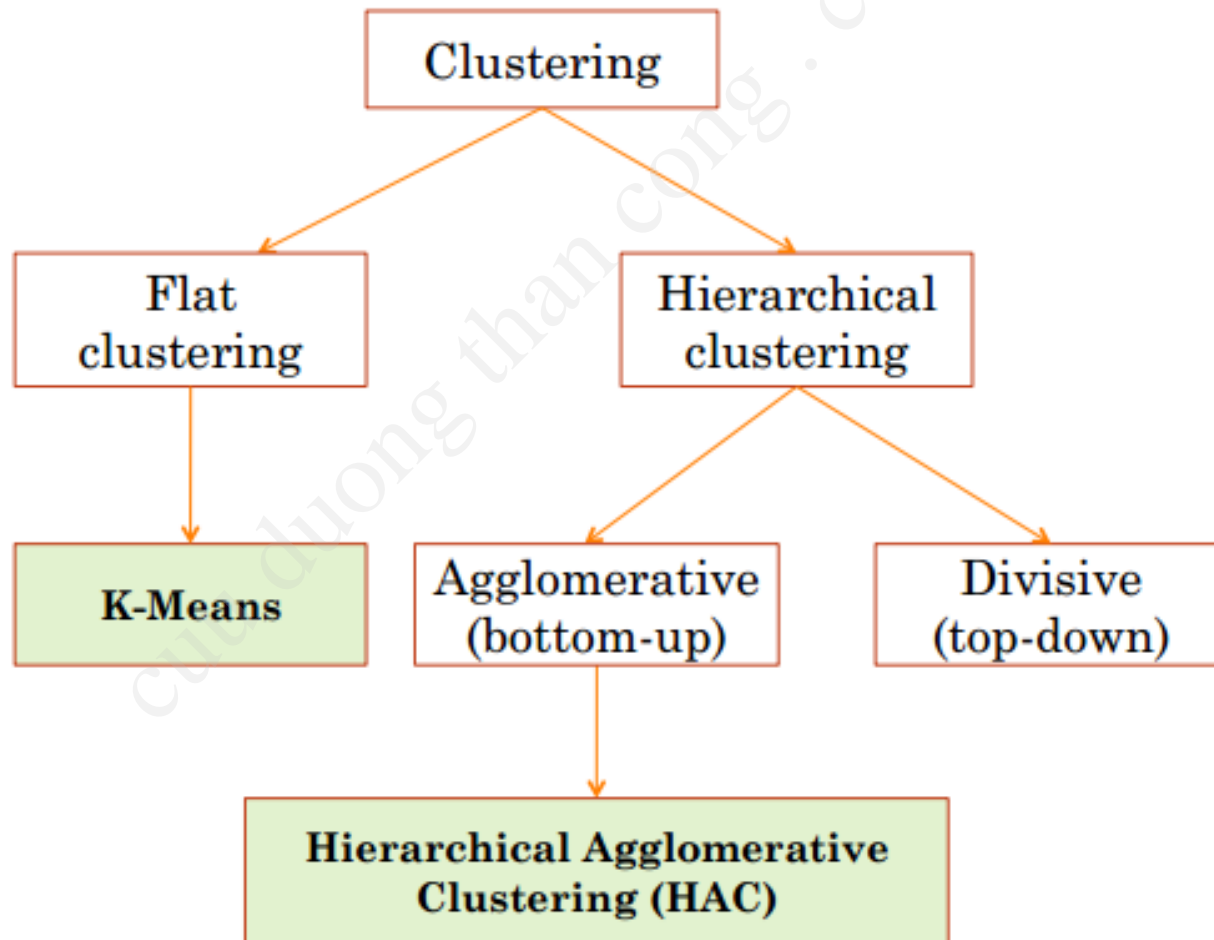
Trọng tâm m của cụm C xác định như sau:

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

Một số thuật giải gom cụm dữ liệu

- ❖ Hierarchical Agglomerative Clustering (HAC)
 - Single Link
 - Complete Link
 - Centroid
 - Group Average
- ❖ K-means

Một số thuật giải gom cụm dữ liệu



Một số thuật giải gom cụm dữ liệu

❖ Giải thuật K-means

Input: Tập dữ liệu D gồm m đối tượng dữ liệu (bản ghi): r_1, r_2, \dots, 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 Hội tụ (không còn sự thay đổi);

End;

Một số thuật giải gom cụm dữ liệu

- ❖ Giải thuật K-means – Điều kiện dừng:
 - 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 từ các đối tượng thuộc cụm đến trọng tâm cụm là cực tiểu:

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

Một số thuật giải gom cụm dữ liệu

- ❖ Giải thuật K-means – Điều kiện dừng:
 - Giải thuật không hội tụ: trọng tâm của các cụm liên tục thay đổi. Khi này có các 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:
$$|J_{n+1} - J_n| < \varepsilon$$

Thuật giải K-means

- Phân dữ liệu sau thành 2 cụm ($K=2$).

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4

Thuật giải K-means

Bước 1: Chọn tâm ban đầu $c_1 = A, c_2 = B$

- ✓ Dùng công thức tính khoảng cách (Euclidean) để lần lượt tính khoảng cách từ các tâm đến từng đối tượng.
- ✓ Gán đối tượng vào cụm mà khoảng cách từ đối tượng đến tâm là gần hơn

$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = \sqrt{18}$$

$$\rightarrow D \in \{B\}$$

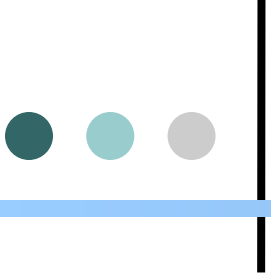
$$d(C, c_1) = \sqrt{(4-1)^2 + (3-1)^2} = \sqrt{13}$$

$$d(C, c_2) = \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{8}$$

$$\rightarrow C \in \{B\}$$

Bước 2: Tính lại tâm mới của cụm

Bước 3: Lặp lại các **Bước 1** và **Bước 2**



Thuật giải HAC (Hierarchical Agglomerative Clustering)

Ý tưởng: tích lũy từ dưới lên

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.

Thuật giải HAC (Hierarchical Agglomerative Clustering)

G: tập các cụm.

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

k: số lượng cụm mong muốn.

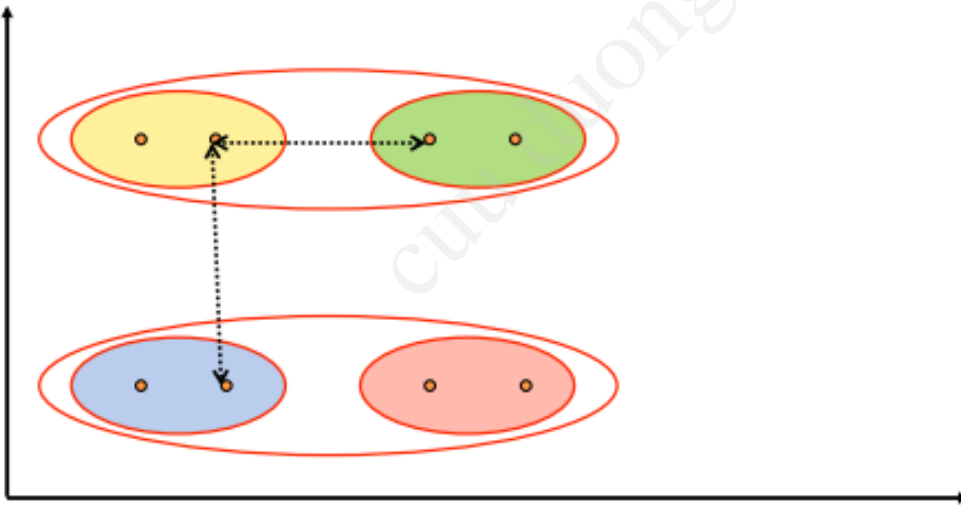
d₀: ngưỡng khoảng cách giữa 2 cụm.

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; // Đạt số lượng cụm mong muốn
3. Tìm hai cụm $S_i, S_j \in \mathbf{G}$ có khoảng cách $d(S_i, S_j)$ là nhỏ nhất;
4. Nếu $d(S_i, S_j) > d_0$ 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. $S = S_i \cup S_j$; // Ghép S_i, S_j thành cụm mới S
7. **G** = **G** ∪ { S }; // Kết nạp cụm mới vào **G**
8. Quay về bước 2.

Thuật giải HAC (Hierarchical Agglomerative Clustering)

❖ Single Link (đo khoảng cách gần nhất):

- Khoảng cách giữa hai cụm được xác định là khoảng cách giữa hai phần tử “gần” nhau nhất của hai cụm đó.

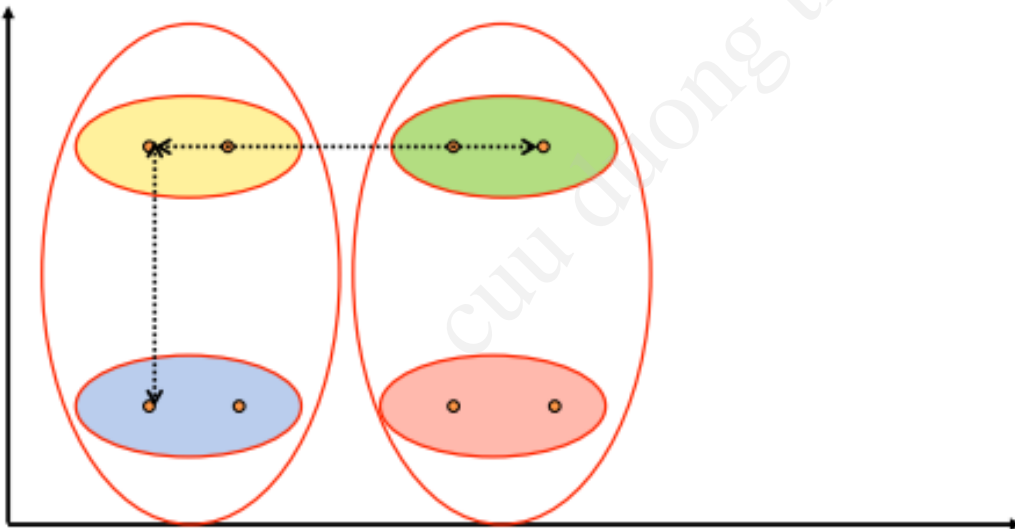


$$d(S_1, S_2) = \min_{r_i \in S_1, r_j \in S_2} d(r_i, r_j)$$

Thuật giải HAC (Hierarchical Agglomerative Clustering)

❖ Complete Link (đo khoảng cách xa nhất):

- Khoảng cách giữa hai cụm được xác định là khoảng cách giữa hai phần tử “xa” nhau nhất của hai cụm đó

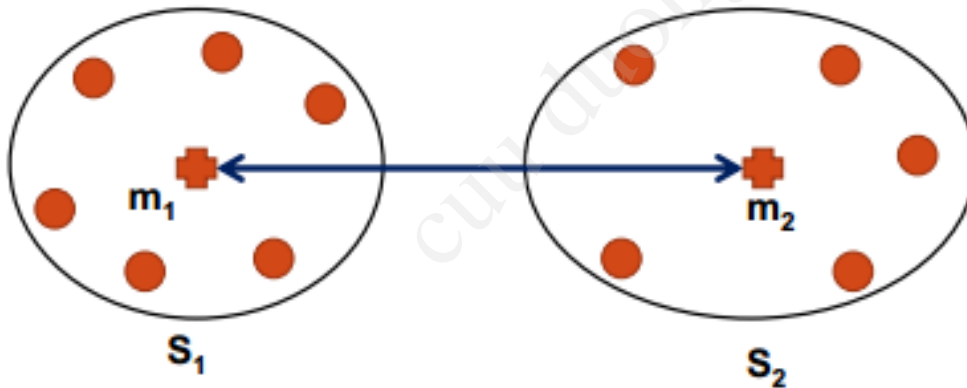


$$d(S_1, S_2) = \max_{r_i \in S_1, r_j \in S_2} d(r_i, r_j)$$

Thuật giải HAC (Hierarchical Agglomerative Clustering)

❖ Centroid Link (đo khoảng cách trọng tâm):

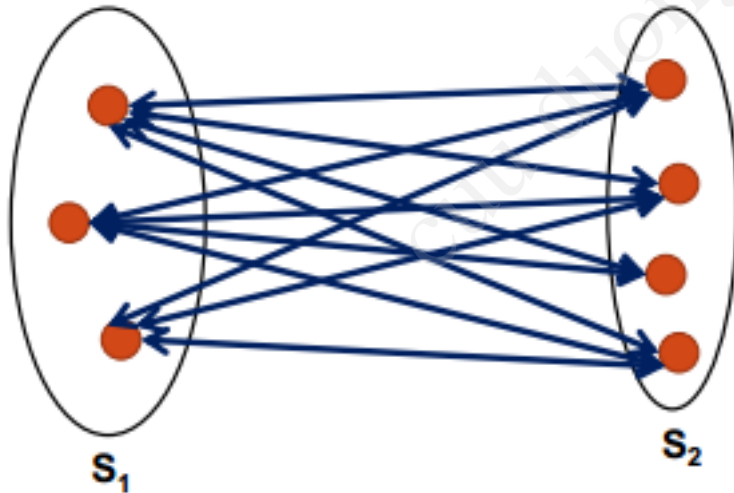
- Khoảng cách giữa hai cụm được xác định là khoảng cách giữa hai trọng tâm của hai cụm đó



$$d(S_1, S_2) = d(m_i, m_j)$$

Thuật giải HAC (Hierarchical Agglomerative Clustering)

- ❖ **Group Average Link** (đo khoảng cách trung bình nhóm):
 - Khoảng cách giữa hai cụm được xác định là khoảng cách trung bình giữa các phần tử thuộc về hai cụm đó



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

Một số thuật giải gom cụm dữ liệu

❖ Ứng dụng:

- Hierarchical Agglomerative Clustering (HAC)
 - ✓ Tạo ra cây phân cấp ngay trong quá trình phân cụm,
 - ✓ Độ phức tạp cao ($O(n^2)$).
- K-means
 - ✓ Tạo cây phân cấp từng bước một,
 - ✓ Độ phức tạp thấp hơn HAC ($O(nkt)$)