Khai phá dữ liệu
(Data mining)

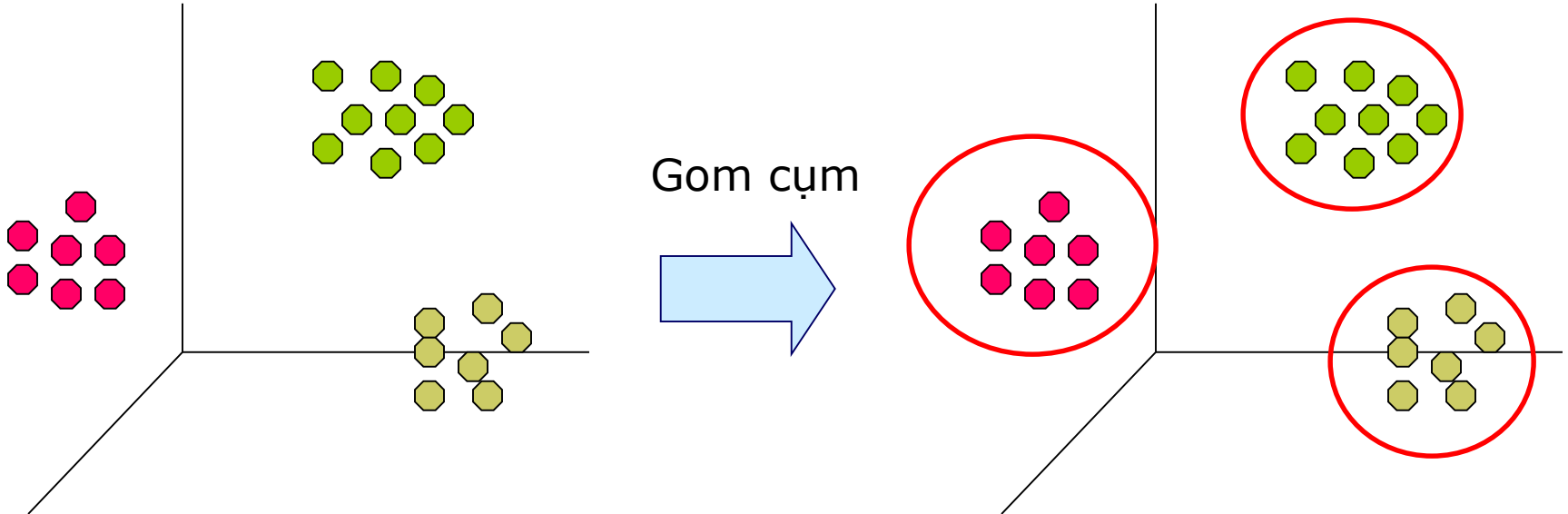
CHƯƠNG 4: PHÂN CỤM DỮ LIỆU

Nội dung

- Khái niệm về phân cụm dữ liệu
- Một số phương pháp phân cụm dữ liệu
  - Phân cụm phân hoạch (K-Means)
  - Phân cụm phân cấp
- Ứng dụng của phân cụm dữ liệu

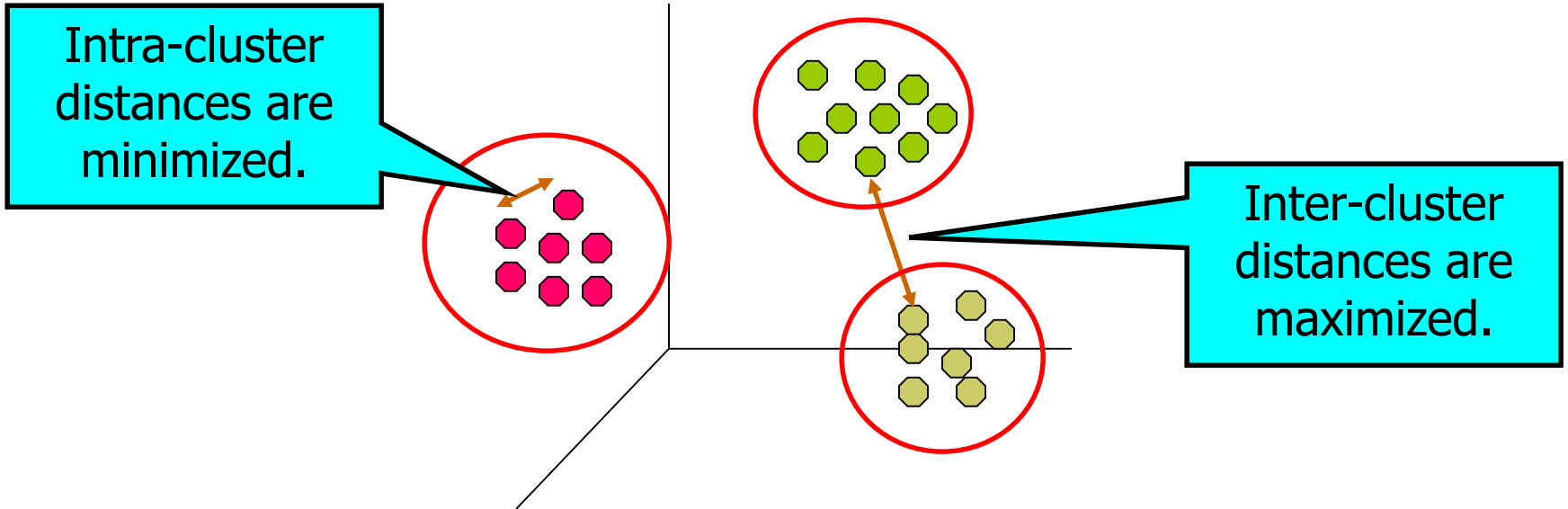
Khái niệm phân cụm

- Phân cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



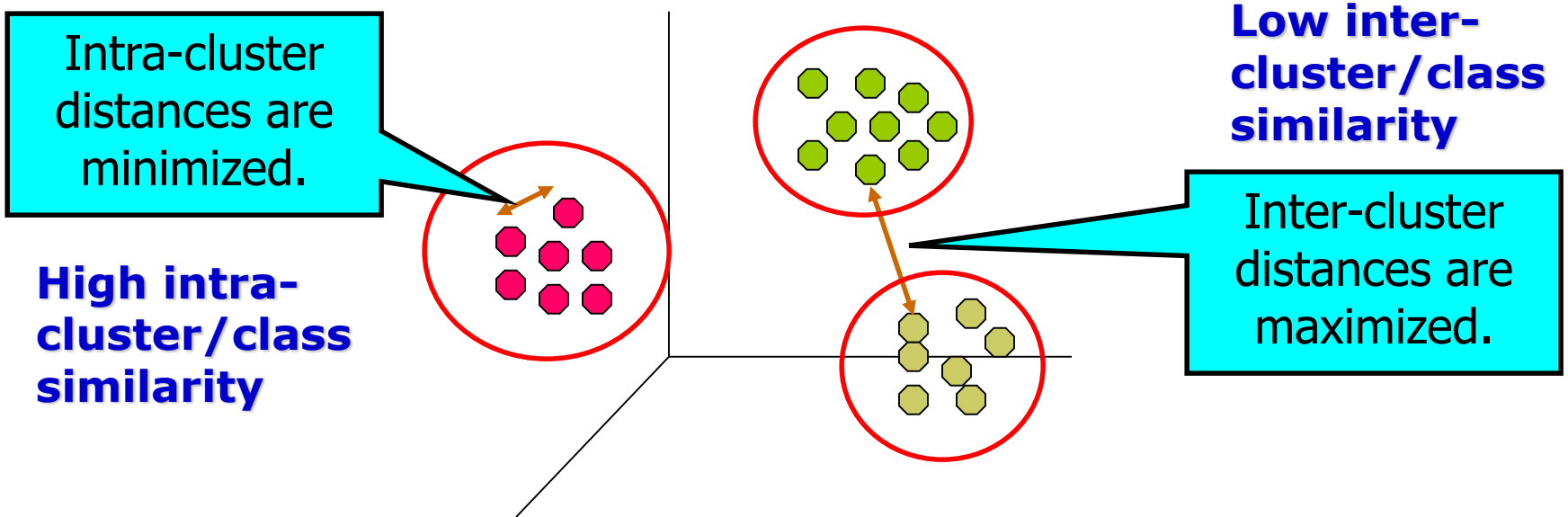
Khái niệm phân cụm

- Phân cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



Khái niệm phân cụm

- Phân cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



Độ đo trong phân cụm dữ liệu

- Minkowski

$$\sum_{i=1}^{n} (|| x_i - y_i ||^p)^{\frac{1}{p}}$$

- Euclidean – p = 2

- Độ đo tương tự (gần nhau): cosin hai vectơ

$$\cos\mu = \frac{v.w}{\| v \| . \| w \|}$$

Một số phương pháp phân cụm

- Phân cụm phân hoạch
- Phân cụm phân cấp
- Phân cụm dựa trên mật độ
- Phân cụm dựa trên lưới
- Phân cụm dựa trên mô hình
- Phân cụm có ràng buộc

Phân cụm phân hoạch

Phân cụm phân hoạch

- Phân 1 tập dữ liệu có n phần tử cho trước thành k tập con dữ liệu ($k \leq n$), mỗi tập con biểu diễn 1 cụm.
- Các cụm hình thành trên cơ sở làm tối ưu giá trị hàm đo độ tương tự sao cho:
  - Các đối tượng trong 1 cụm là tương tự.
  - Các đối tượng trong các cụm khác nhau là không tương tự nhau.
- Đặc điểm:
  - Mỗi đối tượng chỉ thuộc về 1 cụm.
  - Mỗi cụm có tối thiểu 1 đối tượng.
- Một số thuật toán điển hình : K-mean, PAM, CLARA,…

Thuật toán K-Means

Phát biểu bài toán:

- Input
  - Tập các đối tượng $X = \{x_i | i = 1, 2, \ldots, N\}$, $x_i \in R^d$
  - Số cụm: K
- Output
  - Các cụm C_i ($i = 1 \div K$) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Thuật toán K-Means (tt)

- Thuật toán hoạt động trên 1 tập vectơ d chiều, tập dữ liệu X gồm N phần tử:

$$X = \{x_i \mid i = 1, 2, \ldots, N\}$$

- K-Mean lặp lại nhiều lần quá trình:
  - Gán dữ liệu.
  - Cập nhật lại vị trí trọng tâm.

- Quá trình lặp dừng lại khi trọng tâm hội tụ và mỗi đối tượng là 1 bộ phận của 1 cụm.

Thuật toán K-Means (tt)

- Hàm đo độ tương tự sử dụng khoảng cách Euclidean

$$E = \sum_{i=1}^{N} \sum_{x_i \in C_j} (\| x_i - c_j \|^2)$$

trong đó c_j là trọng tâm của cụm C_j

- Hàm trên không âm, giảm khi có 1 sự thay đổi trong 1 trong 2 bước: gán dữ liệu và định lại vị trí tâm.

Thuật toán K-Means (tt)

- Bước 1 - Khởi tạo

  Chọn K trọng tâm $\{c_i\}$ (i = 1÷K).

- Bước 2 - Tính toán khoảng cách

$$S_i^{(t)} = \{ x_j : \| x_j - c_i^{(t)} \| \leq \| x_j - c_{i^*}^{(t)} \| \text{ for all } = 1, .i^*, k\}$$

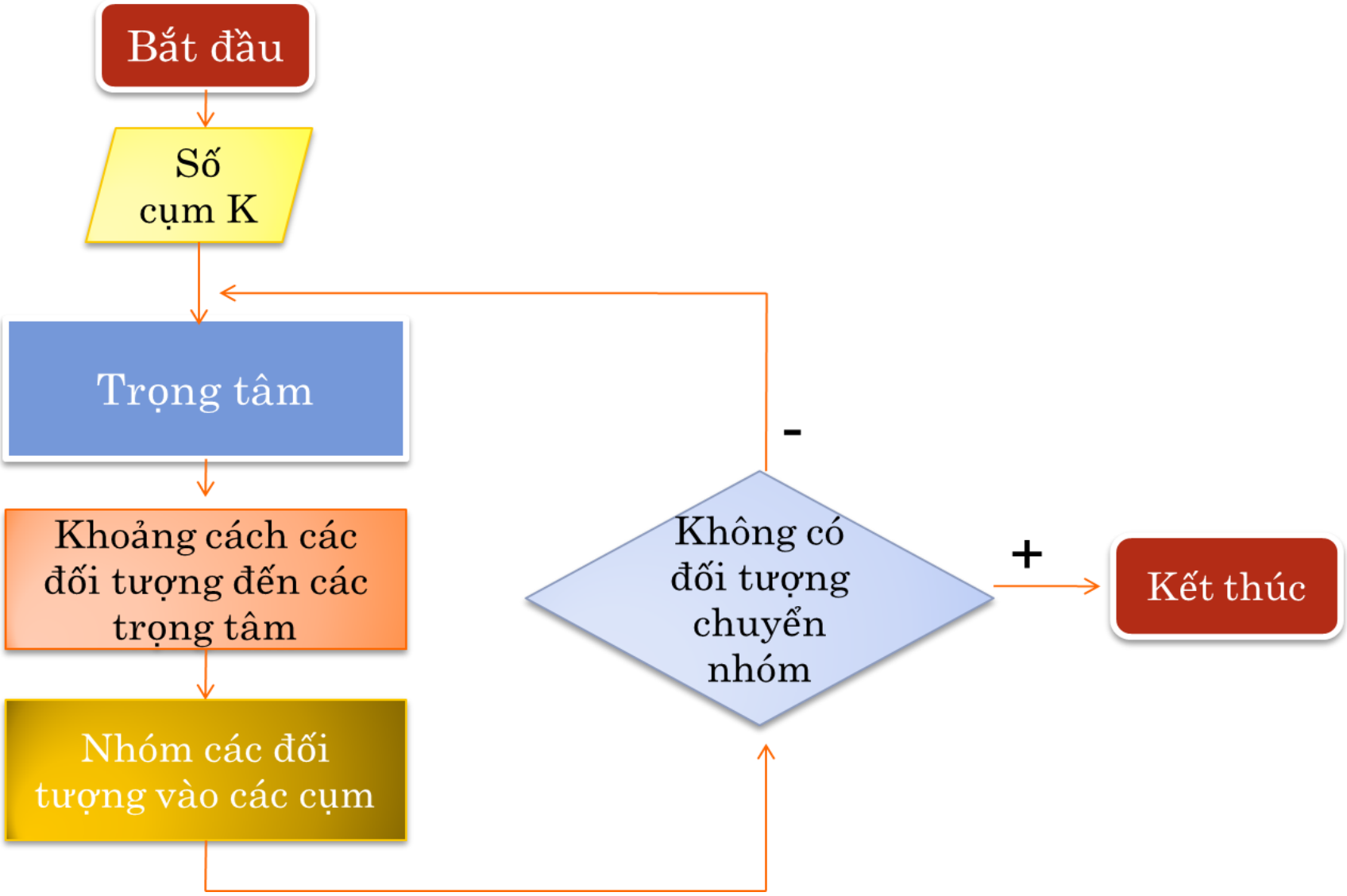
- Bước 3 - Cập nhật lại trọng tâm

$$c_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

- Bước 4 – Điều kiện dừng

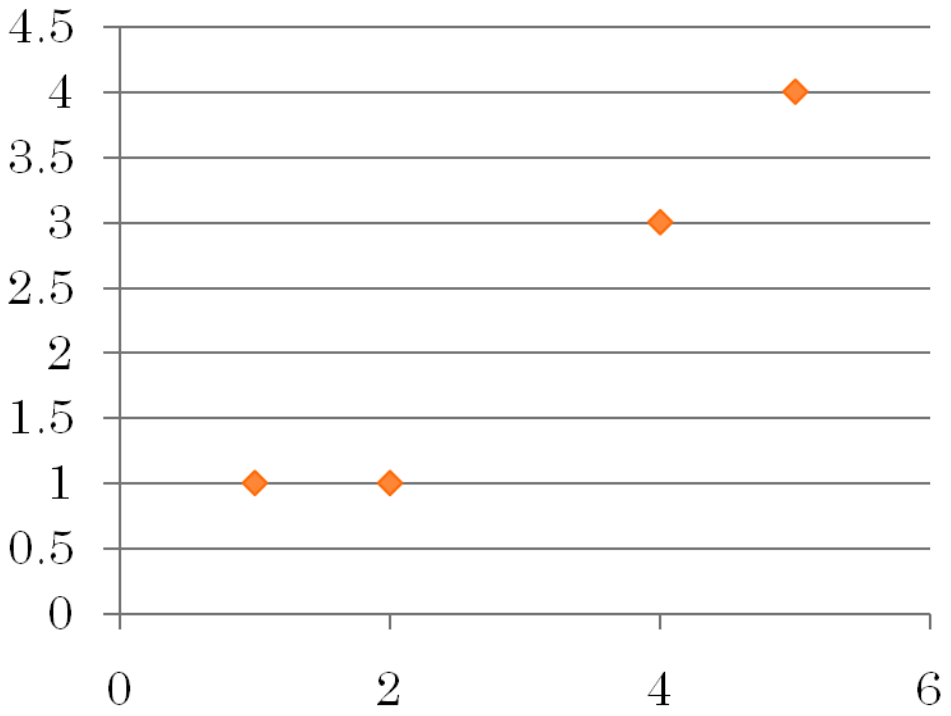
  Lặp lại các bước 2 và 3 cho tới khi không có sự thay đổi trọng tâm của cụm.

Thuật toán K-Means (tt)



Ví dụ

Đối tượng	Thuộc tính 1 (X)	Thuộc tính 2 (Y)
A	1	1
B	2	1
C	4	3
D	5	4



Ví dụ

- **Bước 1:** Khởi tạo

  Chọn 2 trọng tâm ban đầu:

  $c_1(1,1) \equiv A$ và $c_2(2,1) \equiv B$, thuộc 2 cụm 1 và 2

Ví dụ

- **Bước 2:** Tính toán khoảng cách

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

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

  $d(C, c_1) > d(C, c_2)$ ➡ C thuộc cụm 2

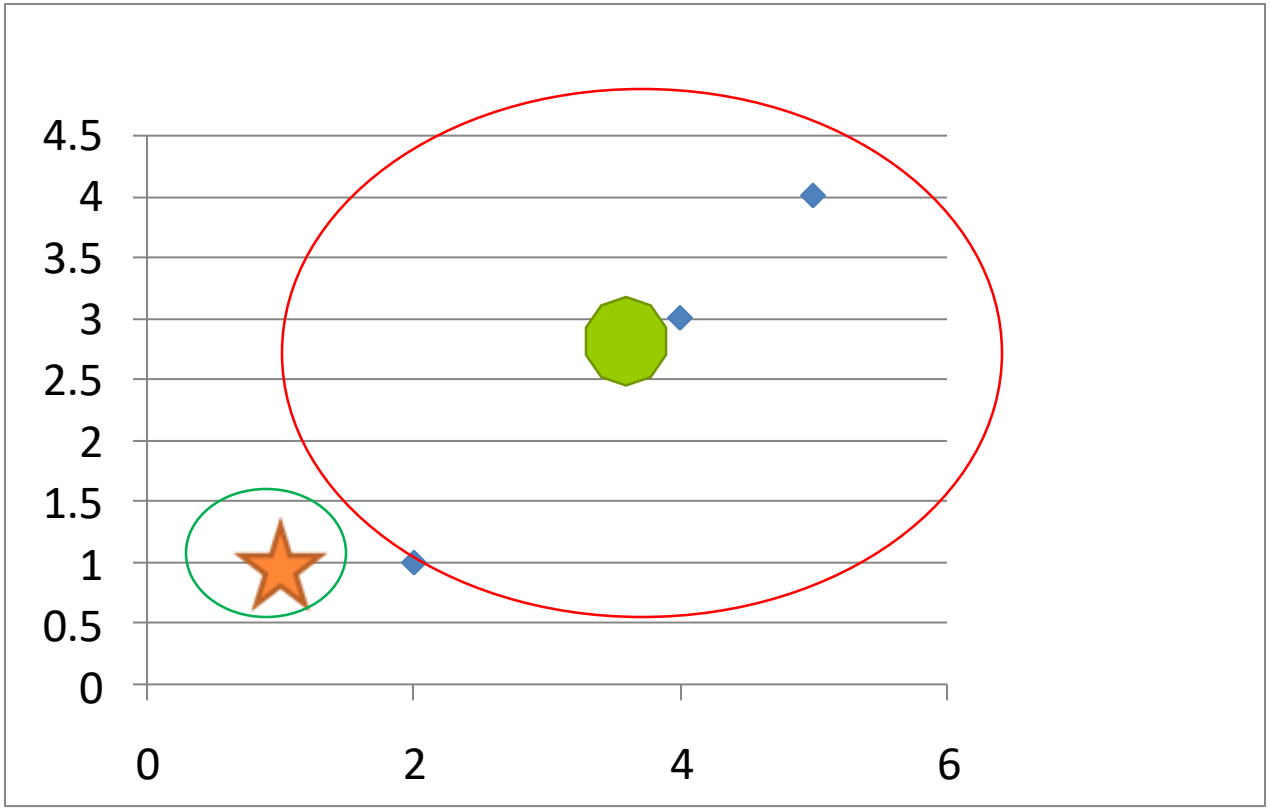
- $d(D, c_1) = \sqrt{(5-2)^2+(4-1)^2} = 25$

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

  $d(D,c_1) > d(D, c_2)$ ➡ D thuộc cụm 2

Ví dụ

- **Bước 3:** Cập nhật lại vị trí trọng tâm
  - Trọng tâm cụm 1 $c_1 \equiv A\ (1, 1)$
  - Trọng tâm cụm 2 c_2 $(x,y) = (\frac{2+4+5}{3}, \frac{1+3+4}{3})$



Ví dụ

- **Bước 4-1:** Lặp lại bước 2 – Tính toán khoảng cách
- $d(A, c_1) = 0 < d(A, c_2) = 9.89$

  A thuộc cụm 1
- $d(B, c_1) = 1 < d(B, c_2) = 5.56$

  B thuộc cụm 1
- $d(C, c_1) = 13 > d(C, c_2) = 0.22$

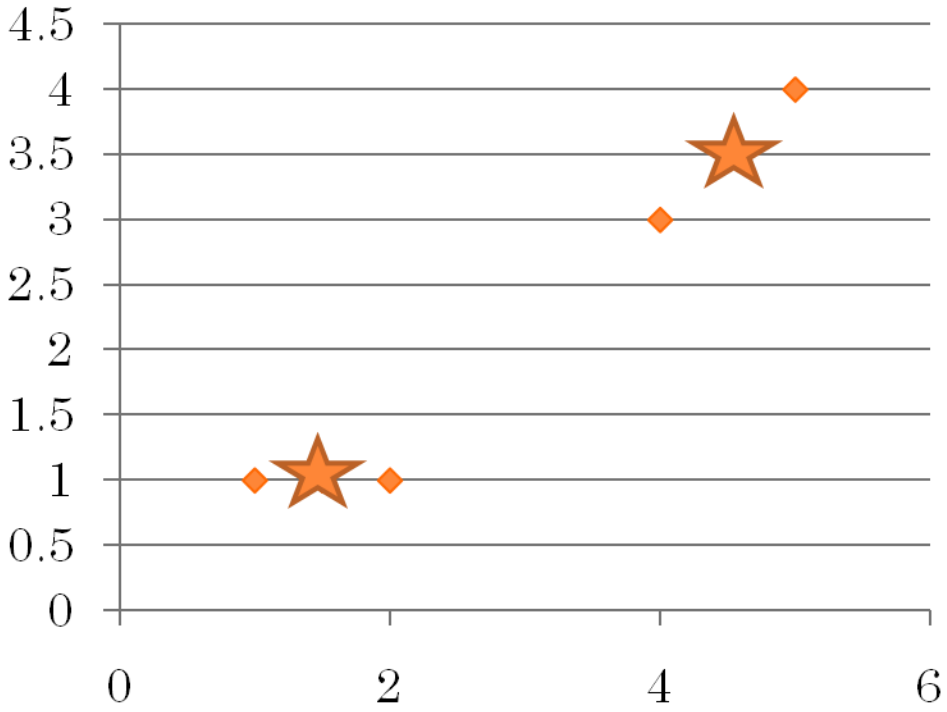
  C thuộc cụm 2
- $d(D, c_1) = 25 > d(D, c_2) = 3.56$

  D thuộc cụm 2

Ví dụ

- **Bước 4-2:** Lặp lại bước 3-Cập nhật trọng tâm

  $c_1 = (3/2, 1)$ và $c_2 = (9/2, 7/2)$



Ví dụ

- **Bước 4-3:** Lặp lại bước 2
  - $d(A, c_1) = 0.25 < d(A, c_2) = 18.5$

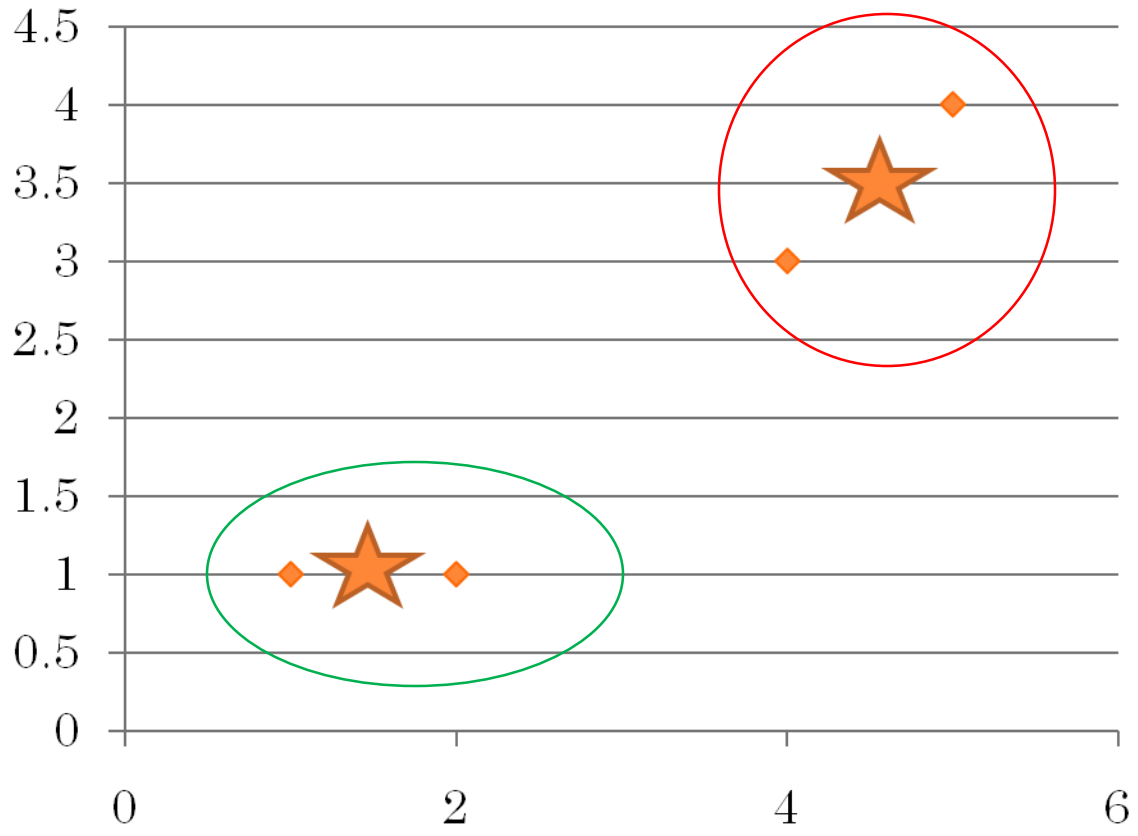
    A thuộc cụm 1
  - $d(B, c_1) = 0.25 < d(B, c_2) = 12.5$

    B thuộc cụm 1
  - $d(C, c_1) = 10.25 < d(C, c_2) = 0.5$

    C thuộc cụm 2
  - $d(D, c_1) = 21.25 > d(D, c_2) = 0.5$

    D thuộc cụm 2

Ví dụ



Đánh giá thuật toán K-means

1. Độ phức tạp: O($K.N.l$) với l: số lần lặp
2. **Ưu điểm**
   - Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.
   - Bảo đảm hội tụ sau 1 số bước lặp hữu hạn.
   - Luôn có K cụm dữ liệu
   - Luôn có ít nhất 1 điểm dữ liệu trong 1 cụm dữ liệu.
   - Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau.
   - Mọi thành viên của 1 cụm là gần với chính cụm đó hơn bất cứ 1 cụm nào khác.

Đánh giá thuật toán K-means

3. **Hạn chế:**

- Không có khả năng tìm ra các cụm không lồi hoặc các cụm có hình dạng phức tạp.
- Khó khăn trong việc xác định các trọng tâm cụm ban đầu

  - Chọn ngẫu nhiên các trung tâm cụm lúc khởi tạo

  - Độ hội tụ của thuật toán phụ thuộc vào việc khởi tạo các vector trung tâm cụm
- Khó để chọn ra được số lượng cụm tối ưu ngay từ đầu, mà phải qua nhiều lần thử để tìm ra được số lượng cụm tối ưu.
- Rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.
- Không phải lúc nào mỗi đối tượng cũng chỉ thuộc về 1 cụm, chỉ phù hợp với đường biên giữa các cụm rõ.

Biến thể của thuật toán K-Means

1. Thuật toán K-medoid:
   - Tương tự thuật toán K-mean
   - Mỗi cụm được đại diện bởi một trong các đối tượng của cụm.
   - Chọn đối tượng ở gần tâm cụm nhất làm đại diện cho cụm đó.
   - K-medoid khắc phục được nhiễu, nhưng độ phức tạp lớn hơn.

Biến thể của thuật toán K-Means

2. Thuật toán Fuzzy c-mean (FCM):
   - Chung chiến lược phân cụm với K-mean.
   - Nếu K-mean là phân cụm dữ liệu cứng (1 điểm dữ liệu chỉ thuộc về 1 cụm) thì FCM là phân cụm dữ liệu mờ (1 điểm dữ liệu có thể thuộc về nhiều hơn 1 cụm với 1 xác suất nhất định).
   - Thêm yếu tố quan hệ giữa các phần tử và các cụm dữ liệu thông qua các trọng số trong ma trận biểu biễn bậc của các thành viên với 1 cụm.
   - FCM khắc phục được các cụm dữ liệu chồng nhau trên các tập dữ liệu có kích thước lớn hơn, nhiều chiều và nhiều nhiễu, song vẫn nhạy cảm với nhiễu và các phần tử ngoại lai.

Phân cụm phân cấp

Phân cụm phân cấp

- Phân cụm dữ liệu bằng phân cấp (hierarchical clustering): nhóm các đối tượng vào cây phân cấp của các cụm
  - Agglomerative: bottom-up (trộn các cụm)
  - Divisive: top-down (phân tách các cụm)

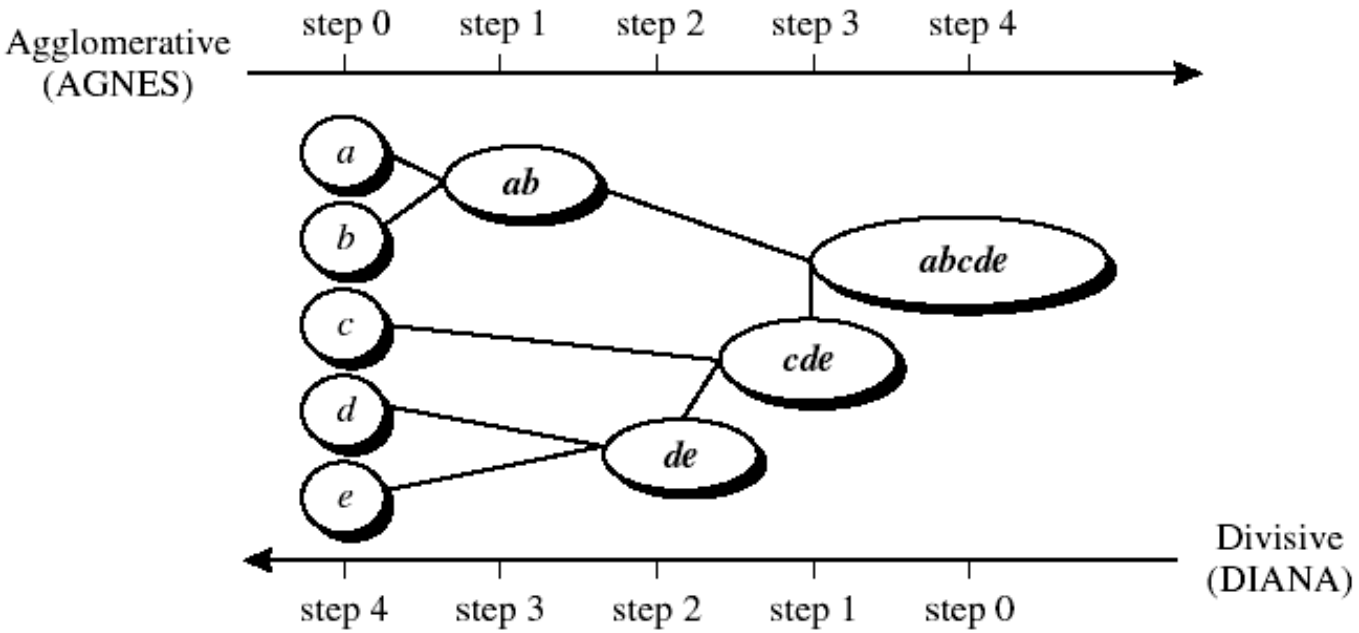
→ Không yêu cầu thông số nhập k (số cụm)

→ Yêu cầu điều kiện dừng

→ Không thể quay lui ở mỗi bước trộn/phân tách

Phân cụm phân cấp

- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting) → bottom-up
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis) → top-down

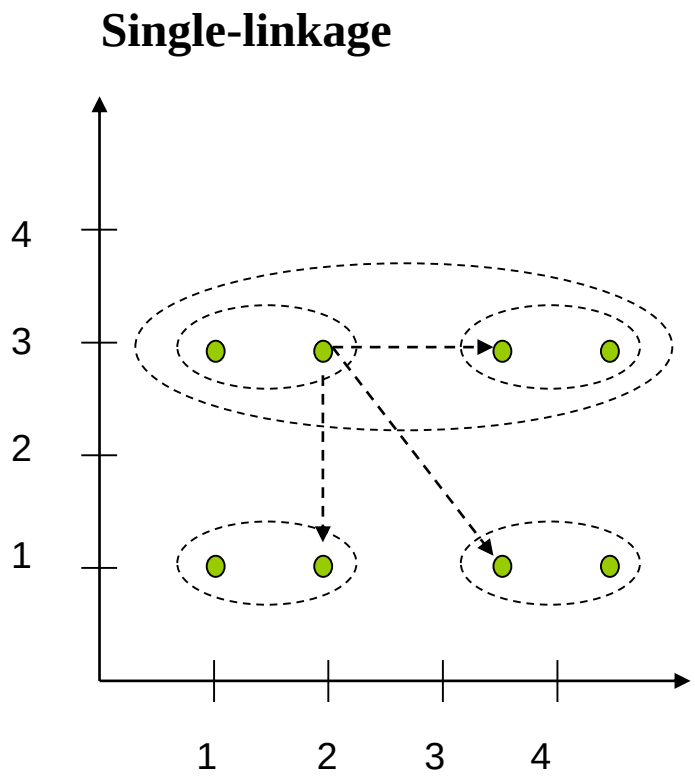


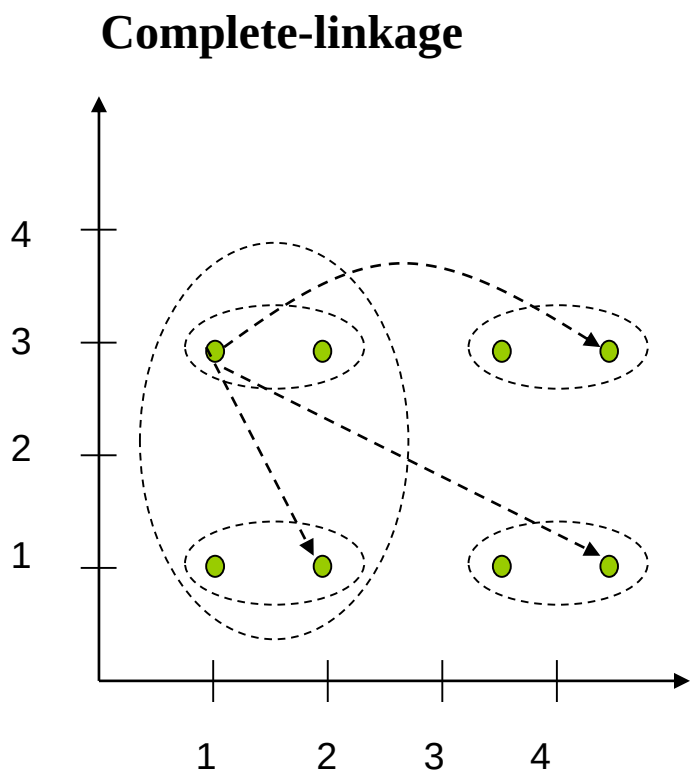
Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

Phân cụm phân cấp

- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting)
  - Khởi đầu, mỗi đối tượng tạo thành một cụm.
  - Các cụm sau đó được trộn lại theo một tiêu chí nào đó.
    - Cách tiếp cận single-linkage: cụm C1 và C2 được trộn lại nếu khoảng cách giữa 2 đối tượng từ C1 và C2 là ngắn nhất.
  - Quá trình trộn các cụm được lặp lại đến khi tất cả các đối tượng tạo thành một cụm duy nhất.
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis)
  - Khởi đầu, tất cả các đối tượng tạo thành một cụm duy nhất.
  - Một cụm được phân tách theo một tiêu chí nào đó đến khi mỗi cụm chỉ có một đối tượng.
    - Khoảng cách lớn nhất giữa các đối tượng cận nhau nhất.

Phân cụm phân cấp

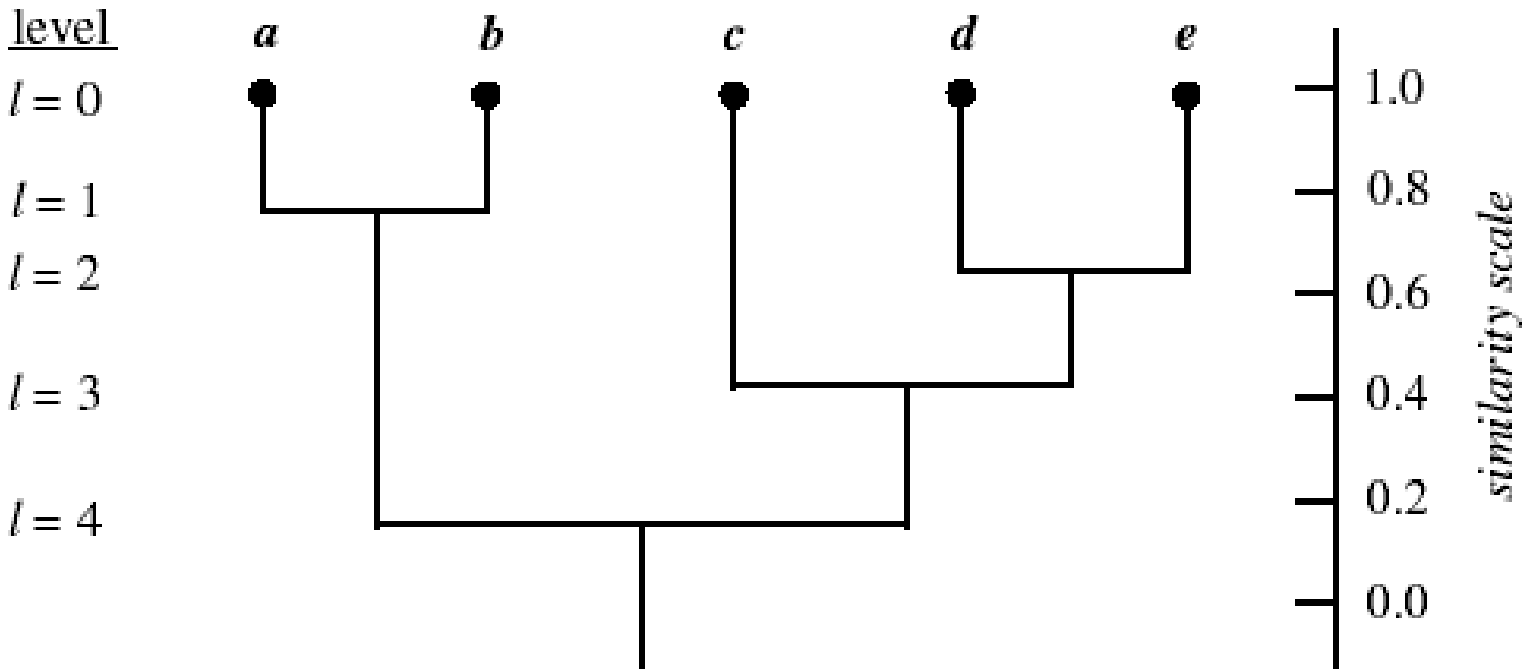




Tiêu chí trộn các cụm: single-linkage và complete-linkage

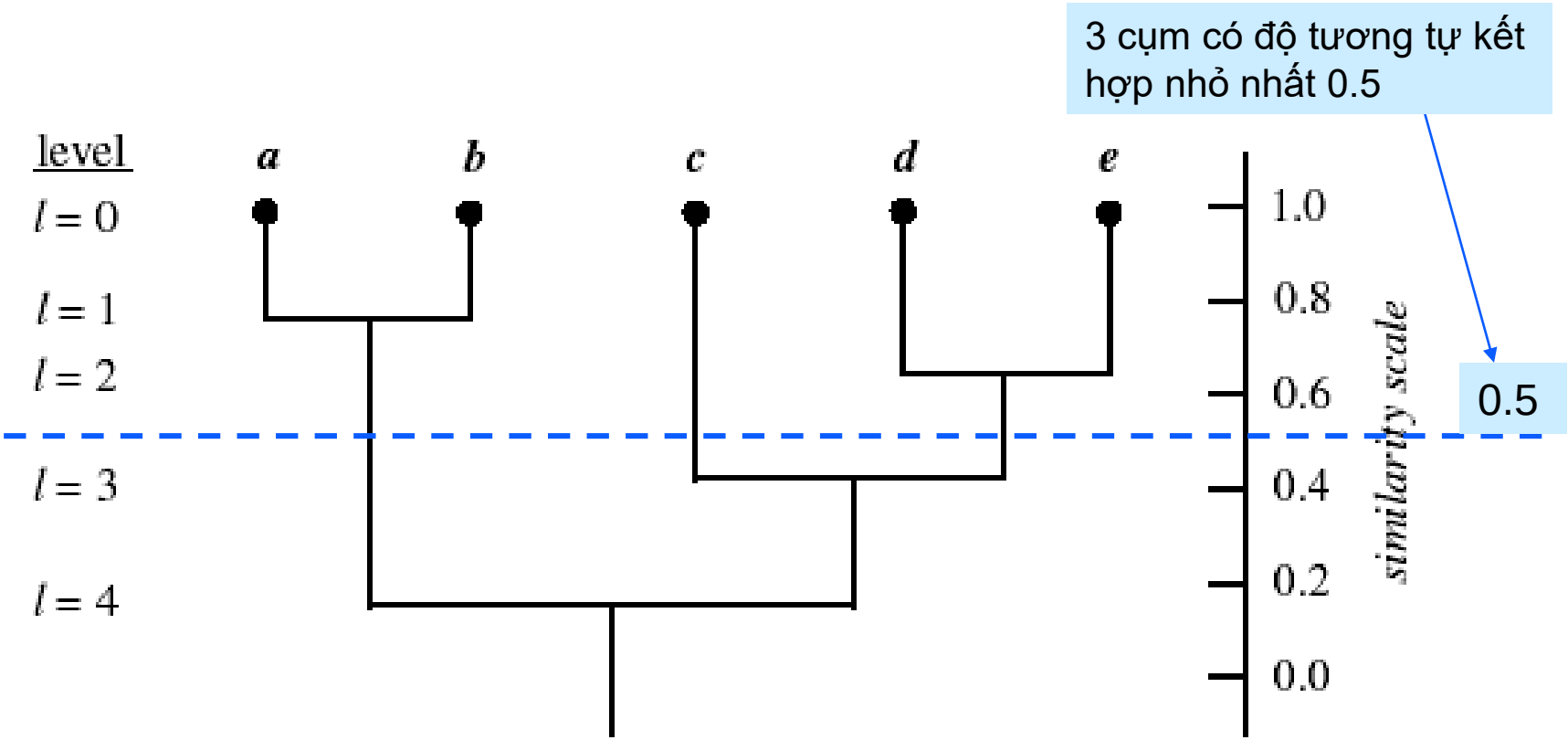
Phân cụm phân cấp

- Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).



Phân cụm phân cấp

- Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).



Phân cụm phân cấp

- Các độ đo khoảng cách giữa các cụm C_i và C_j

$$\text{Minimum distance}: \quad d_{min}(C_i, C_j) = min_{p \in C_i,\, p' \in C_j} |p - p'|$$

$$\text{Maximum distance}: \quad d_{max}(C_i, C_j) = max_{p \in C_i,\, p' \in C_j} |p - p'|$$

$$\text{Mean distance}: \quad d_{mean}(C_i, C_j) = |m_i - m_j|$$

$$\text{Average distance}: \quad d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

p, p': các đối tượng

|p-p'|: khoảng cách giữa p và p'

m_i, m_j: đối tượng trung bình của C_i, C_j, tương ứng

n_i, n_j: số lượng đối tượng của C_i, C_j, tương ứng

Phân cụm phân cấp

- Một số giải thuật gom cụm dữ liệu bằng phân cấp
  - BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): phân hoạch các đối tượng dùng cấu trúc cây theo độ co giãn của phân giải (scale of resolution)
  - ROCK (Robust Clustering using linKs): gom cụm dành cho các thuộc tính rời rạc (categorical/discrete attributes), trộn các cụm dựa vào sự kết nối lẫn nhau giữa các cụm
  - Chameleon: mô hình động để xác định sự tương tự giữa các cặp cụm

Phân cụm phân cấp

- Một số vấn đề với gom cụm dữ liệu bằng phân cấp
  - Chọn điểm trộn/phân tách phù hợp
  - Khả năng co giãn (scalability)
    - Mỗi quyết định trộn/phân tách yêu cầu kiểm tra/đánh giá nhiều đối tượng/cụm.

  → Tích hợp gom cụm dữ liệu bằng phân cấp với các kỹ thuật gom cụm khác

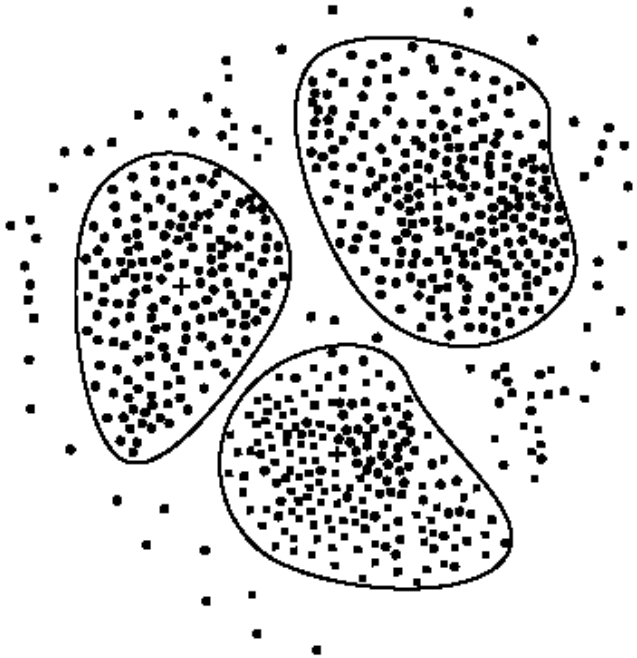
  → Gom cụm nhiều giai đoạn (multiple-phase clustering)

Ứng dụng của phân cụm dữ liệu

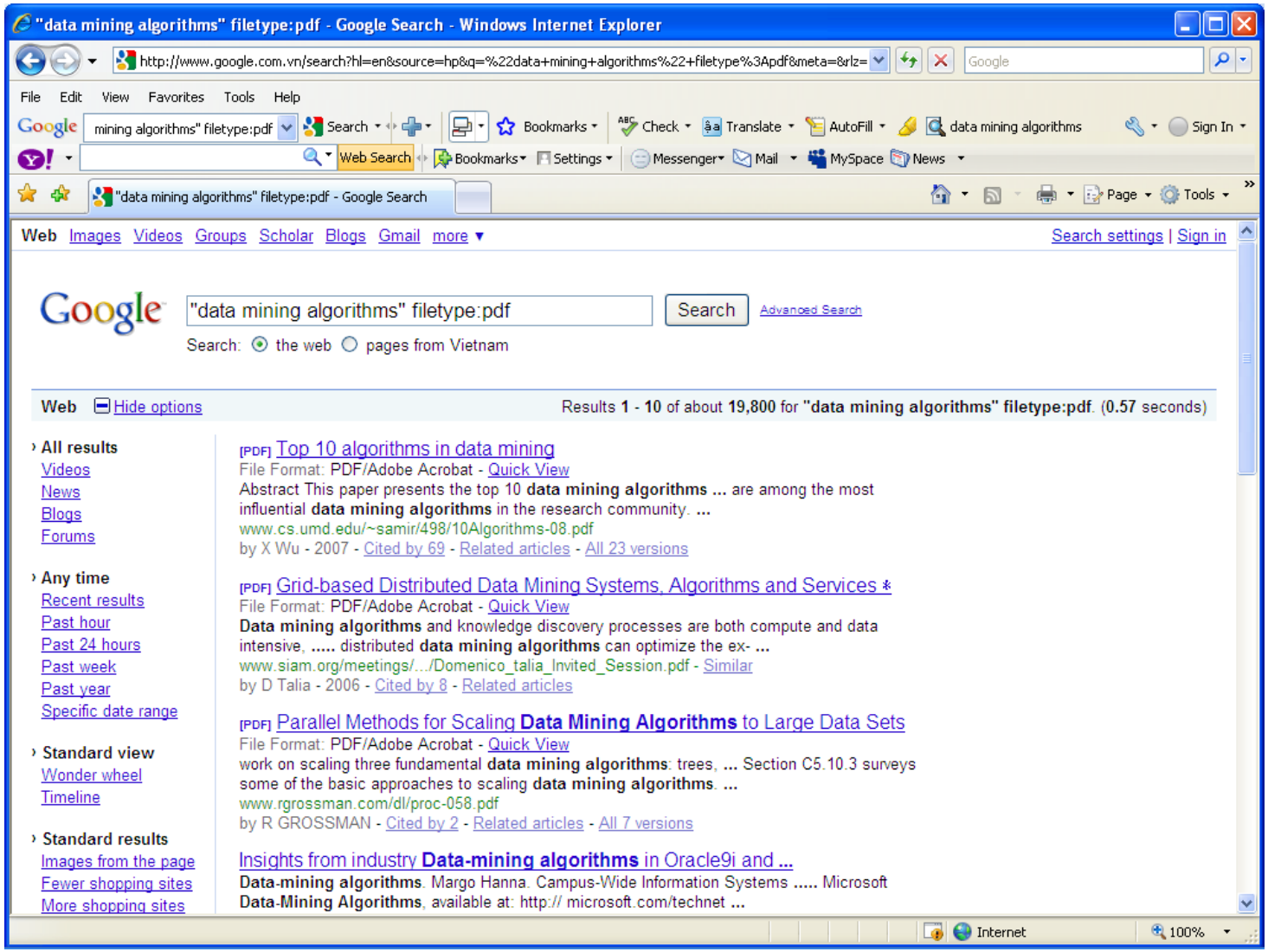
- Hỗ trợ tiền xử lý dữ liệu (làm sạch dữ liệu)
- Mô tả sự phân bố dữ liệu/đối tượng (data distribution)
- Nhận dạng mẫu (pattern recognition)
- Phân tích dữ liệu không gian (spatial data analysis)
- Xử lý ảnh (image processing)
- Phân mảnh thị trường (market segmentation)
- Gom cụm tài liệu ((WWW) document clustering)
- …

Phân cụm - làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - Phân tích cụm (cluster analysis)



Phân cụm dữ liệu trên web



Phân cụm dữ liệu ảnh

http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeatures.data.html