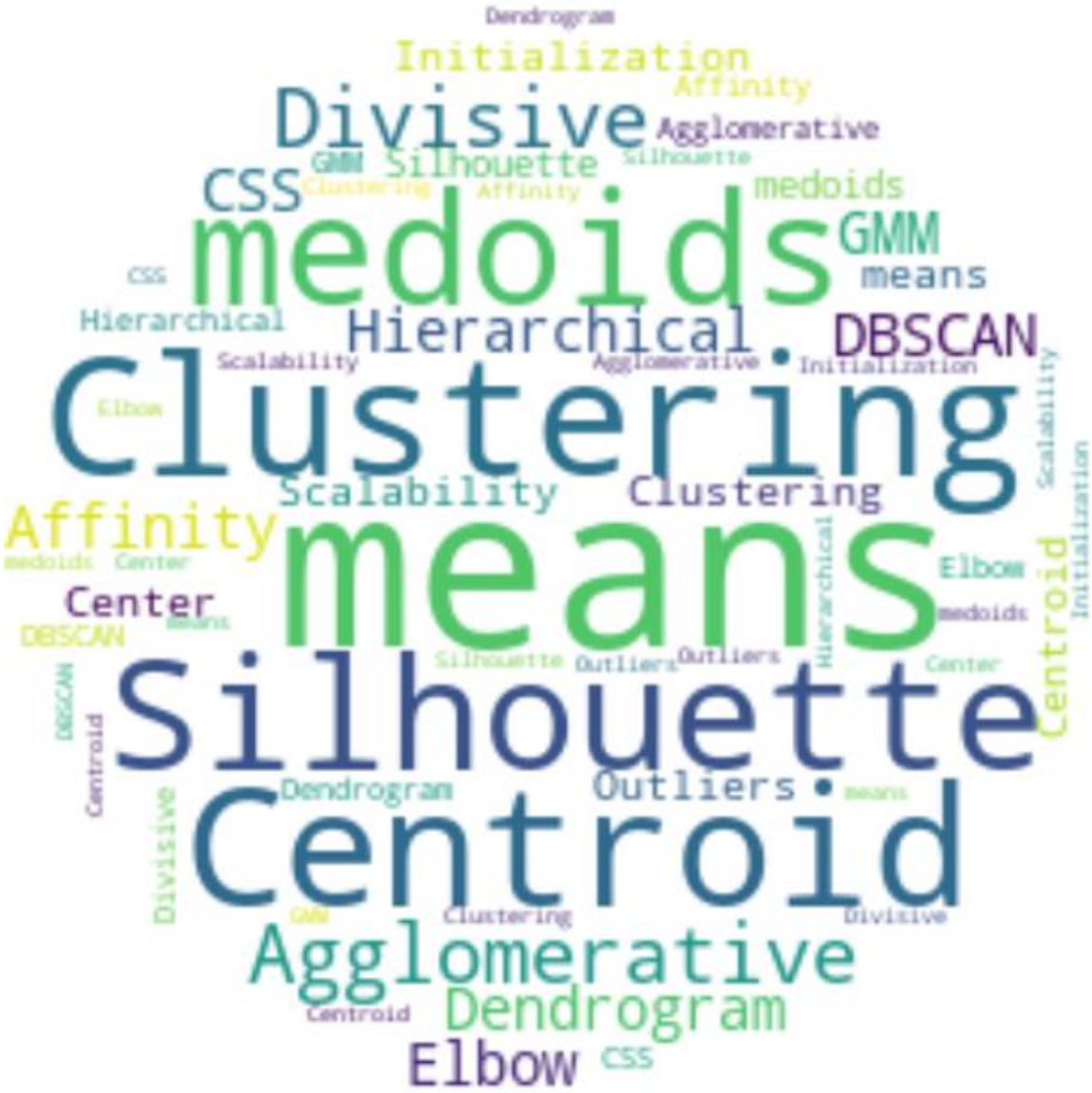

Dendrogram
Initialization
Affinity
Divisive
Agglomerative
GMM
Silhouette
CSS
medoids
GMM
means
Hierarchical
Hierarchical
DBSCAN
Scalability
Agglomerative
Initialization
Clustering
Scalability
Clustering
Affinity
means
Elbow
Center
Centroid
DBSCAN
Silhouette
Outliers
Outliers
Dendrogram
Outliers
Centroid
Agglomerative
Dendrogram
Elbow
CSS

2102470 Học máy

Bài giảng: Khái niệm về phân cụm, mô tả bài toán phân cụm, hàm mục tiêu

Chương 3: Phân cụm

VTPham@IUH, 2024

Ôn lại bài học trước

- Bạn có nhớ ? % ?

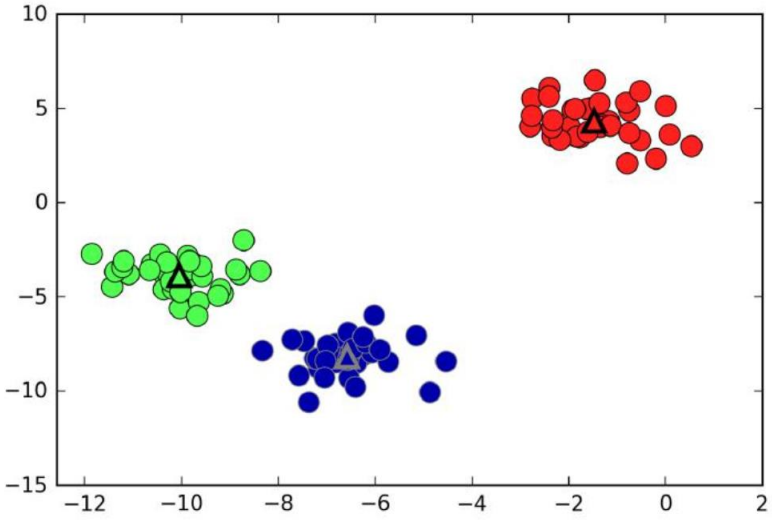
Nội dung chính

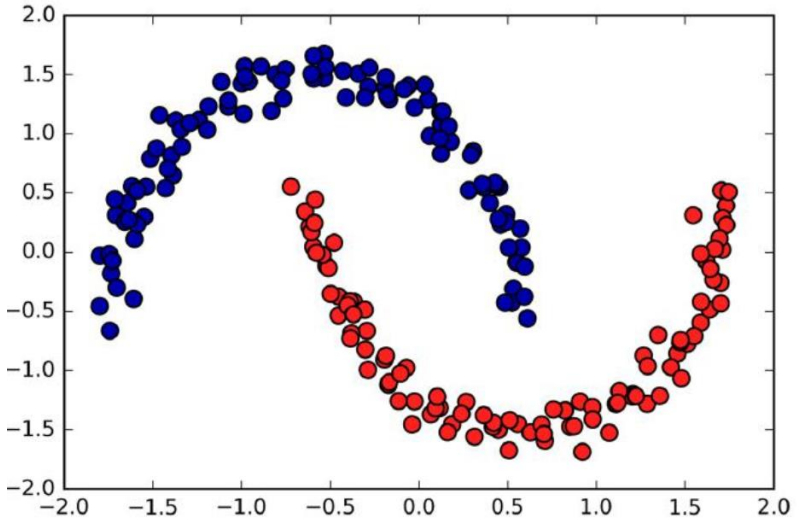
- 3.1 Khái niệm về phân cụm
- 3.2 Mô tả bài toán phân cụm
- 3.3 Hàm mục tiêu

3.1 Khái niệm về phân cụm

Cụm (cluster)

- Định nghĩa về cụm thường phụ thuộc vào ngữ cảnh
  - Các thuật toán khác nhau sẽ nắm bắt các kiểu cụm khác nhau





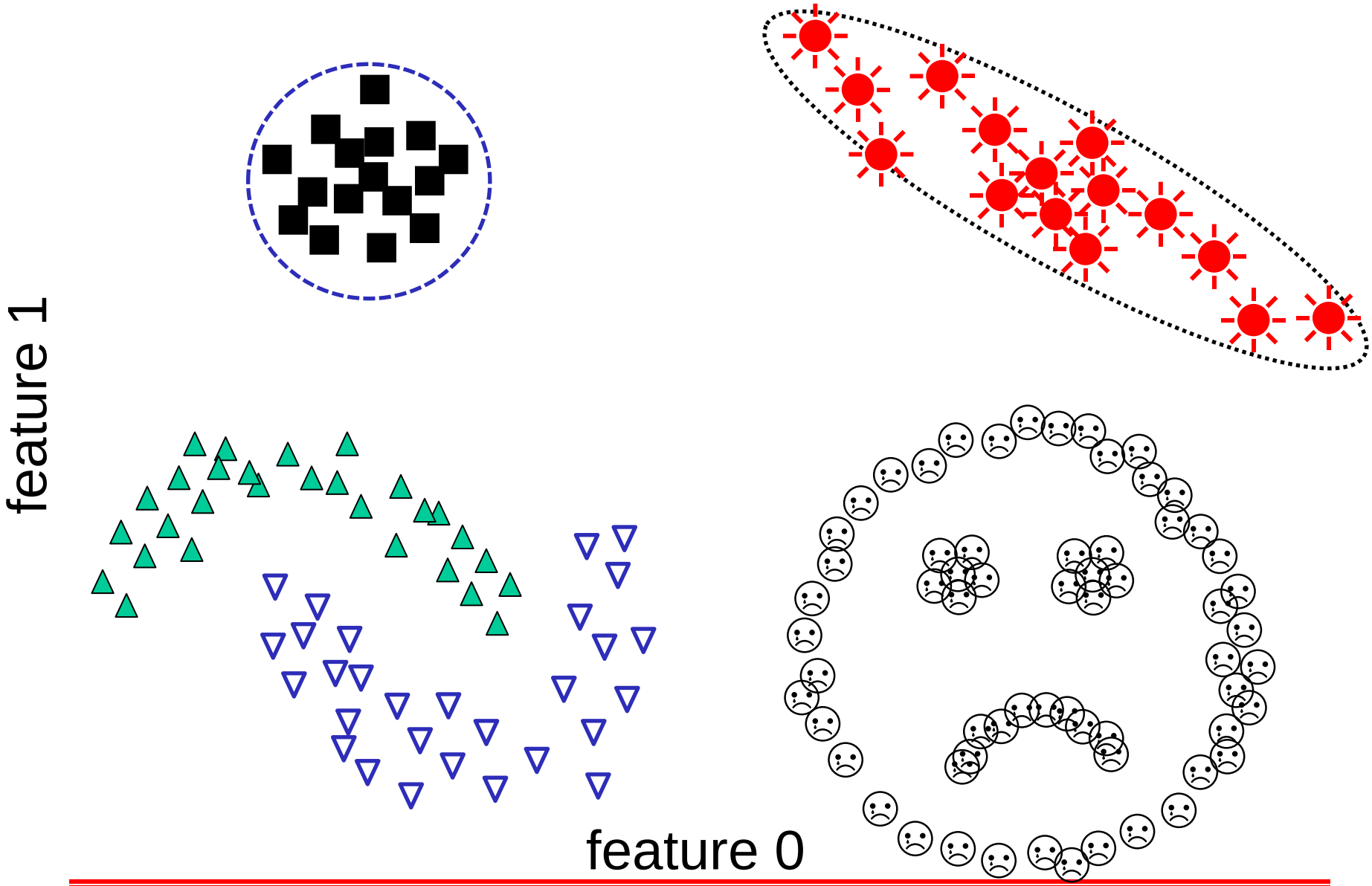
Phân cụm

- Học không giám sát
- Phân chia (phân vùng) tập dữ liệu thành các nhóm, gọi là các cụm (clusters)
  - Các điểm dữ liệu trong cùng một cụm tương đồng nhau hơn
    - So với các điểm dữ liệu trong các cụm khác

Phân cụm

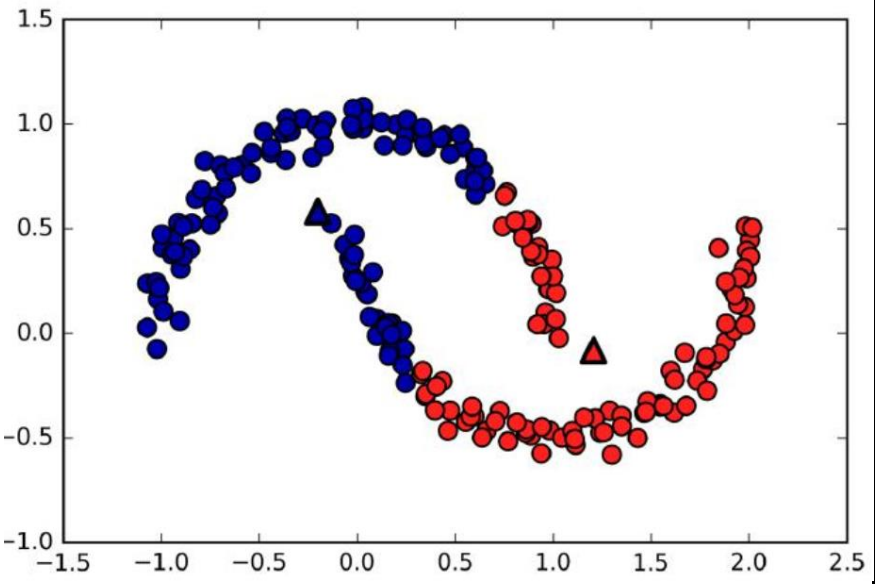
Học không giám sát	• Tập huấn luyện không có nhãn • Khám phá các mẫu, cấu trúc ẩn trong dữ liệu
Hình dạng (kích thước) cụm	• Hình dạng, kích thước, mật độ • Thuật toán để xử lý các loại cụm
Tương đồng (giống nhau)	• Thước đo sự tương đồng? • Ảnh hưởng đến kết quả phân cụm

Phân cụm



Phân cụm

- Các thuật toán thường dựa trên
  - Tìm kiếm các mẫu tập trung xung quanh một điểm cụ thể được gọi là tâm (centroid)
  - Tìm kiếm các vùng liên tục gồm các mẫu tập trung dày đặc
    - Các vùng liên tục này có thể có hình dạng bất kỳ
  - Phân cấp
    - Tìm kiếm đối với cụm trong cụm



Ứng dụng của phân cụm

- 5W1H

Who?

What?

Where?

When?

Why?

How?

Các hệ thống đề xuất

- Đề xuất nội dung cho những người sử dụng khác nhau trong cùng một phân khúc

Hỗ trợ phân tích dữ liệu

- Khi cần phân tích 1 tập dữ liệu mới
- Đầu tiên sẽ tiến hành phân cụm
- Việc khám phá các cụm độc lập thường dễ dàng hơn

Kỹ thuật giảm chiều dữ liệu

- Tập dữ liệu ban đầu được phân cụm
  - Số chiều của vector thuộc tính ban đầu: *N*
- Có thể đo lường sự giống nhau/tương đồng (affinity) của mỗi mẫu so với mỗi cụm
  - Vector thuộc tính của một mẫu có thể được thay thế bằng 1 vector mới gồm các affinity so với *k* cụm (thông thường $k << N$)
    - Vector thuộc tính mới có thể giữ đầy đủ thông tin cho quá trình xử lý tiếp theo

Phát hiện sự bất thường (ngoại lệ)

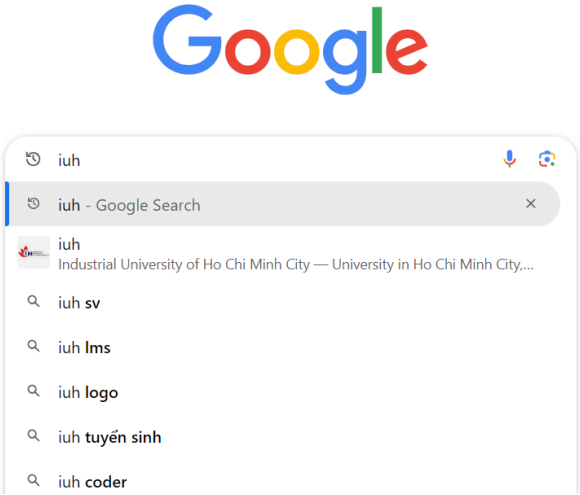
- 1 mẫu có sự tương đồng thấp đối với mọi cụm => có khả năng là một mẫu bất thường
- Đặc biệt hữu ích trong
  - Phát hiện khiếm khuyết trong quá trình sản xuất
  - Phát hiện sự gian lận

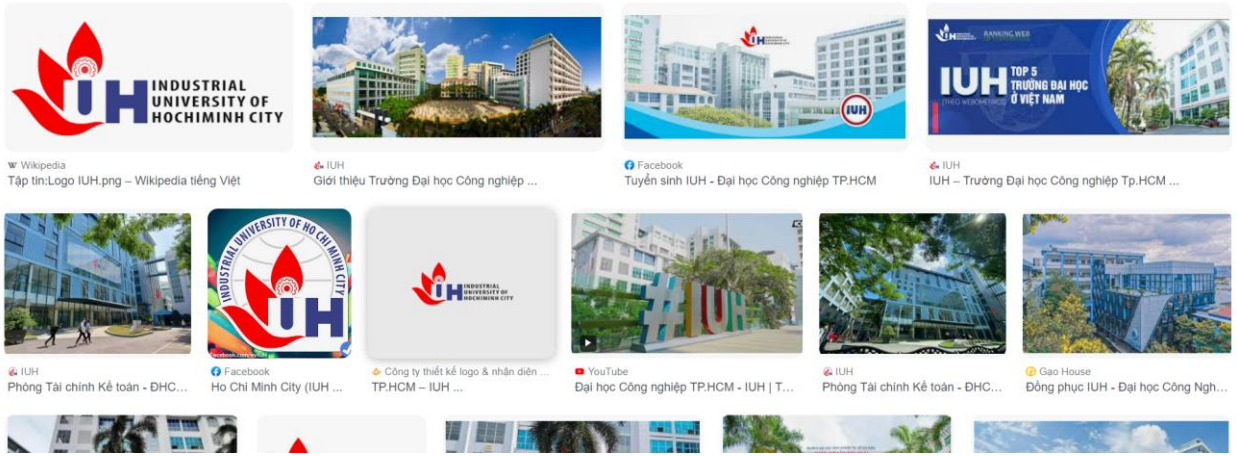
Học bán giám sát

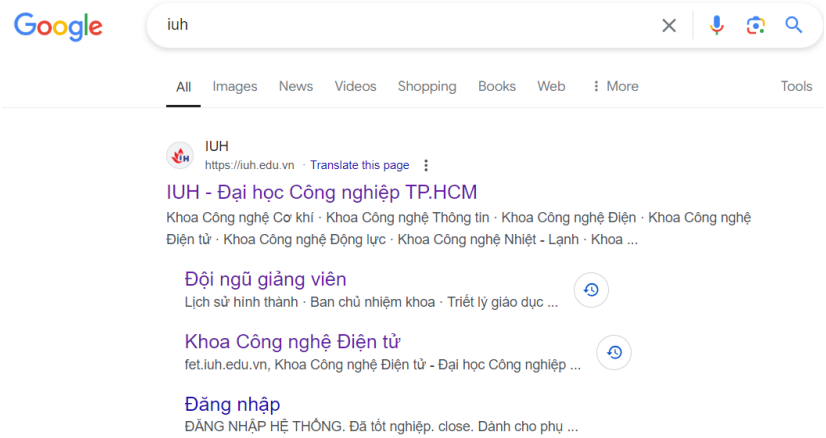
- Khi chỉ có 1 lượng nhỏ trong tập dữ liệu được dán nhãn
- Thực hiện phân cụm
  - Tất cả các mẫu trong cùng một cụm => dán cùng nhãn
- Làm tăng số lượng dữ liệu được dán nhãn => sử dụng trong thuật toán học có giám sát tiếp theo

Dùng cho các máy tìm kiếm

- Ví dụ: Tìm kiếm hình ảnh
  - Máy tìm kiếm cho phép tìm kiếm các hình ảnh tương tự

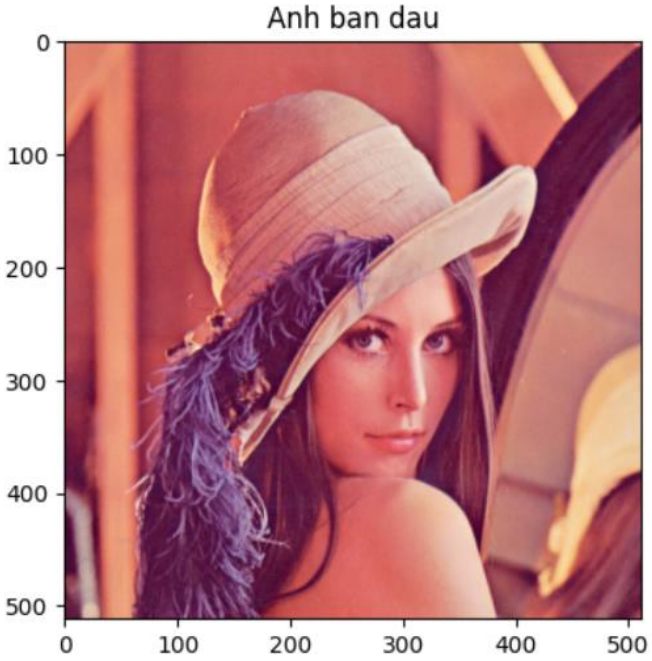


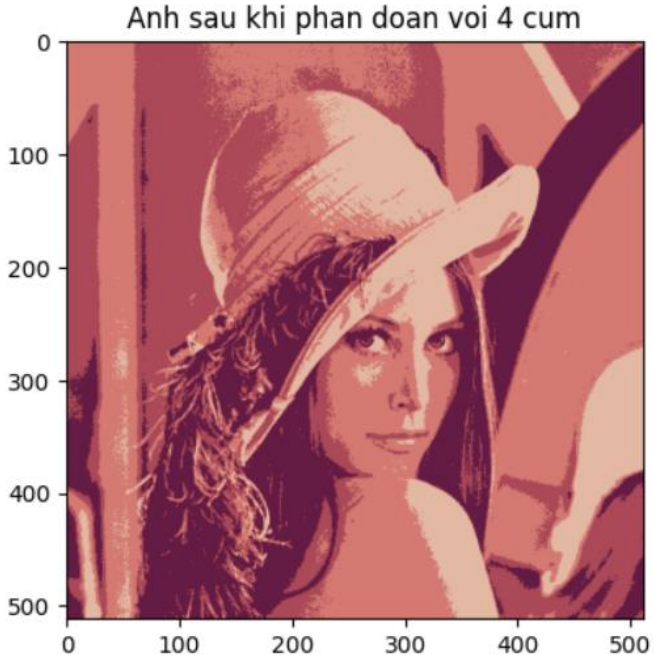




Phân đoạn một hình ảnh

- Phân cụm các pixel theo mầu sắc
  - Thay thế màu của mỗi pixel bằng màu trung bình của cụm mà pixel đó thuộc vào
- Dùng trong hệ thống phát hiện và theo dõi đối tượng

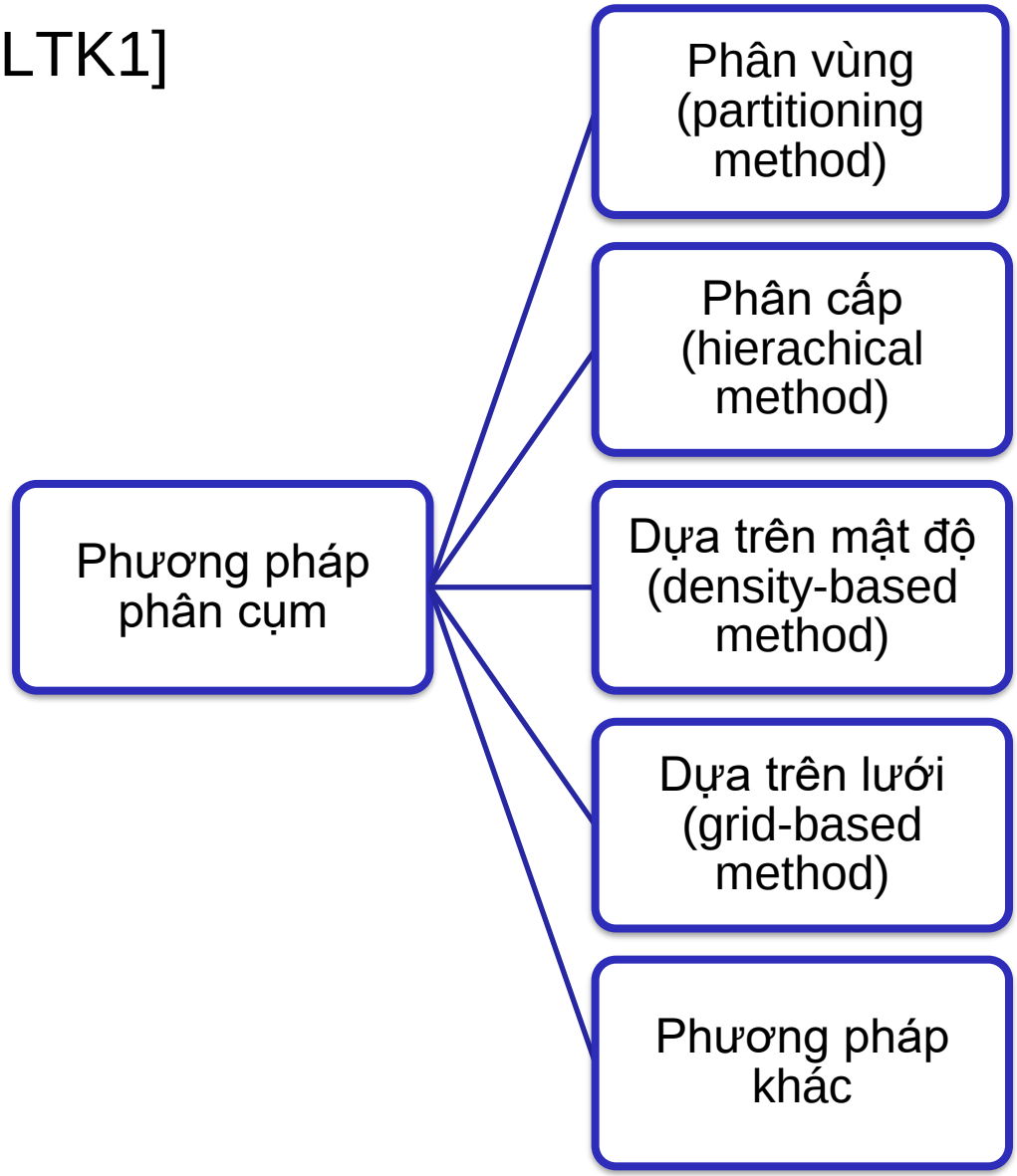




3.2 Mô tả bài toán phân cụm

Các phương pháp phân cụm

- [B6TLTK1]

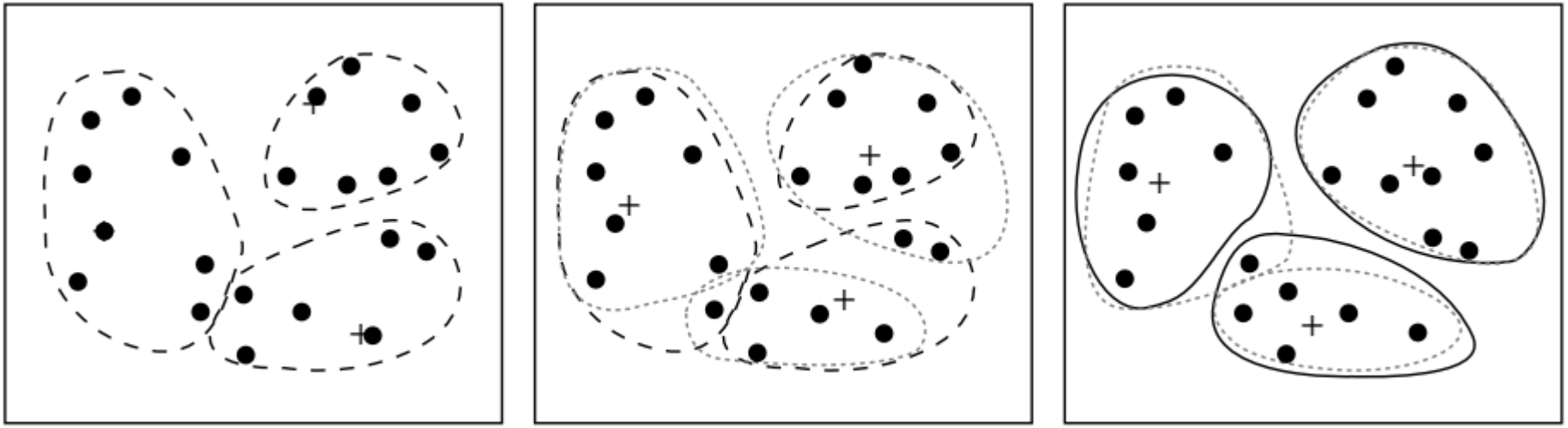


Phương pháp phân vùng [B6TLTK1]

Xét 1 tập dữ liệu: *N* điểm

Xác định *k* phân vùng (*k* =< *N*) (~ mỗi phân vùng đại diện cho 1 cụm)

- Mỗi điểm dữ liệu phải thuộc về một vùng
- Phần lớn các phương pháp phân vùng đều dựa trên “khoảng cách“



(a) Initial clustering **(b)** Iterate **(c)** Final clustering

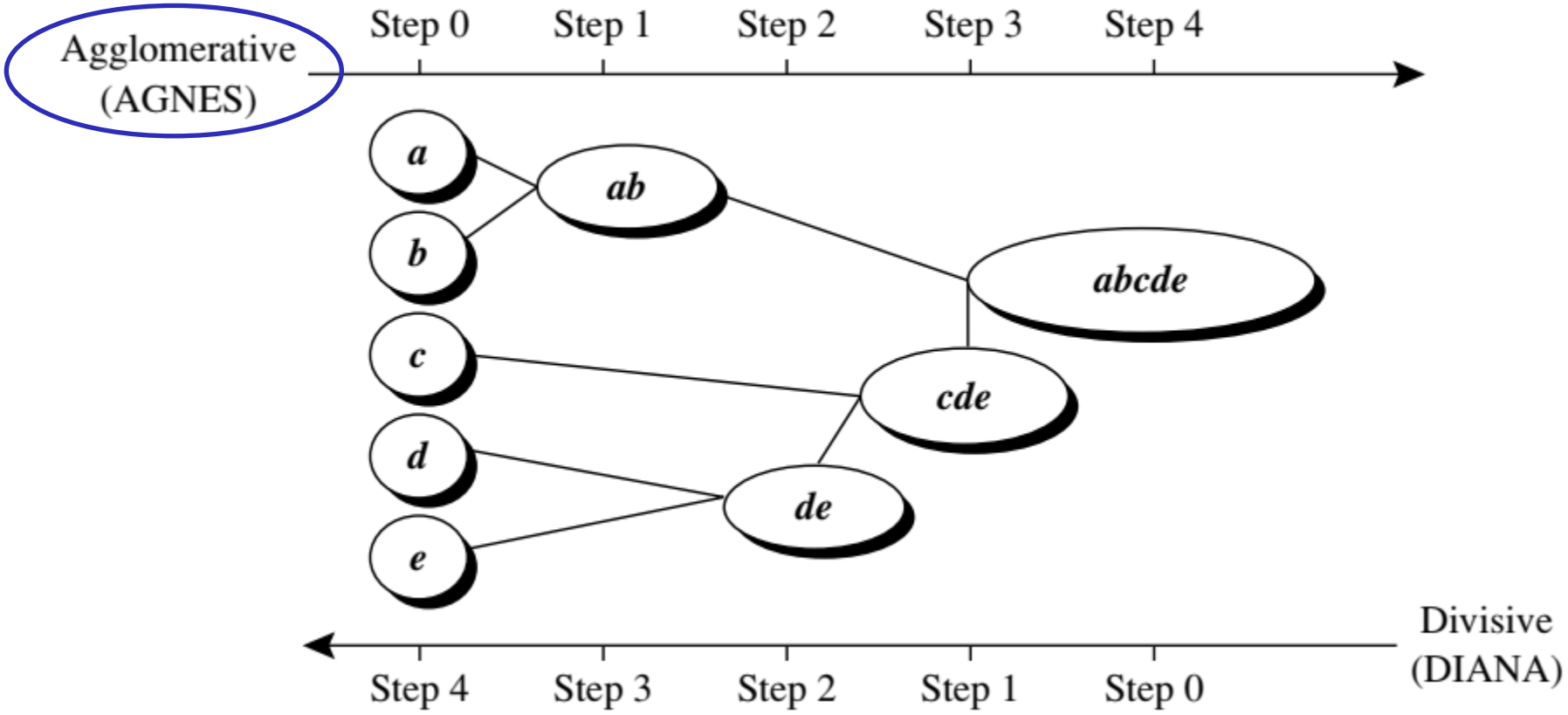
Phương pháp phân cấp [B6TLTK1]

- Không biết số lượng cụm *k* ?

Phương pháp phân cấp [B6TLTK1]

Agglomerative: tổng hợp

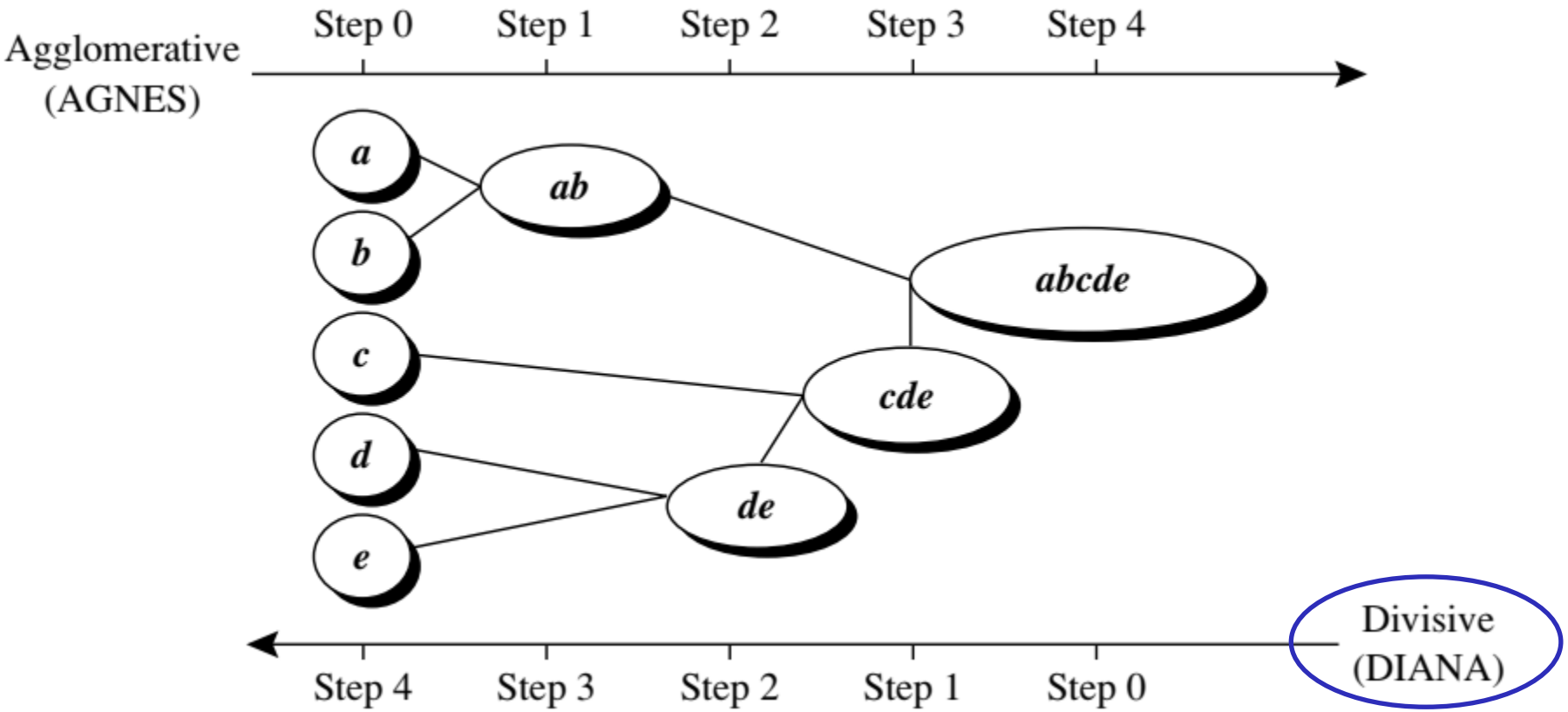
- Mỗi điểm dữ liệu là một cụm
- Tổng hợp các cụm dữ liệu tương đồng thành một cụm mới
- Lặp lại việc tổng hợp cho tới khi thỏa mãn điều kiện dừng hoặc không thể thực hiện được việc tổng hợp nữa



Phương pháp phân cấp [B6TLTK1]

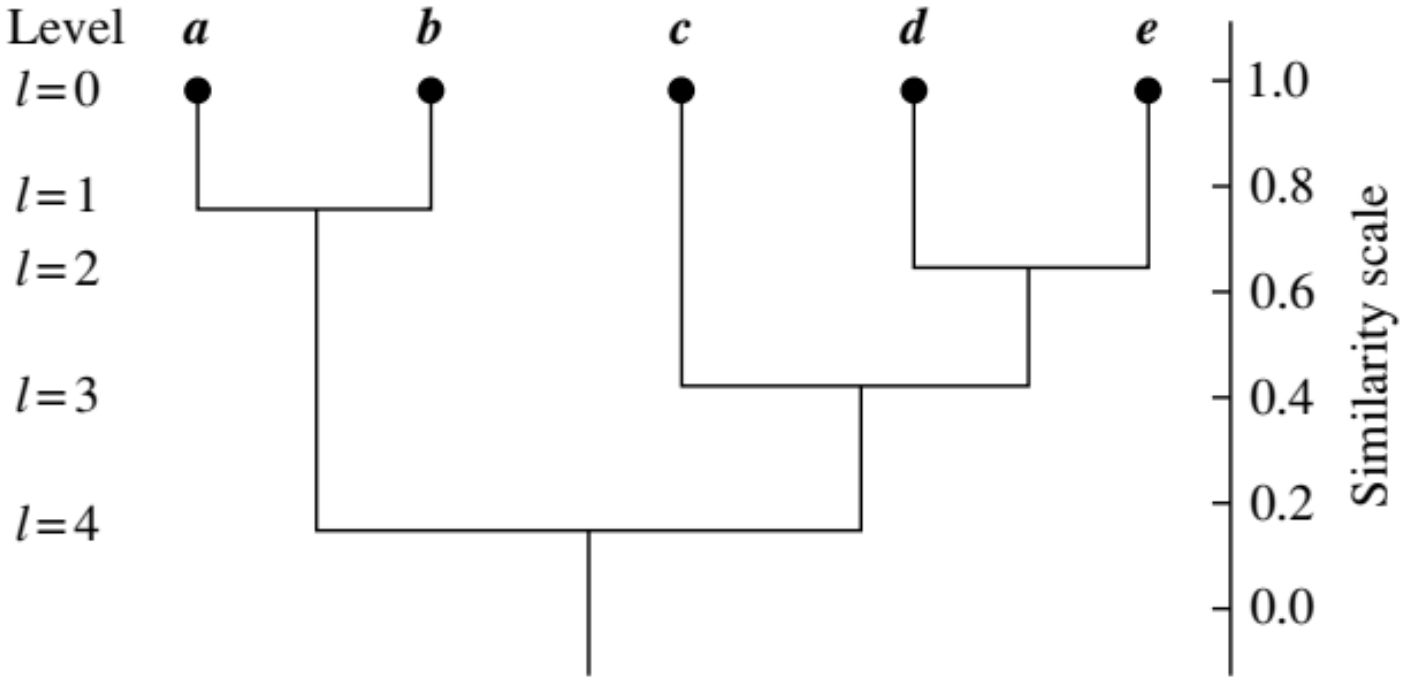
Divisive: phân chia

- Cụm ban đầu, $k = 1$
- Phân chia ra các cụm nhỏ hơn
- Lặp lại việc phân chia cho tới khi thỏa mãn điều kiện dừng hoặc không thể thực hiện việc phan chia được nữa



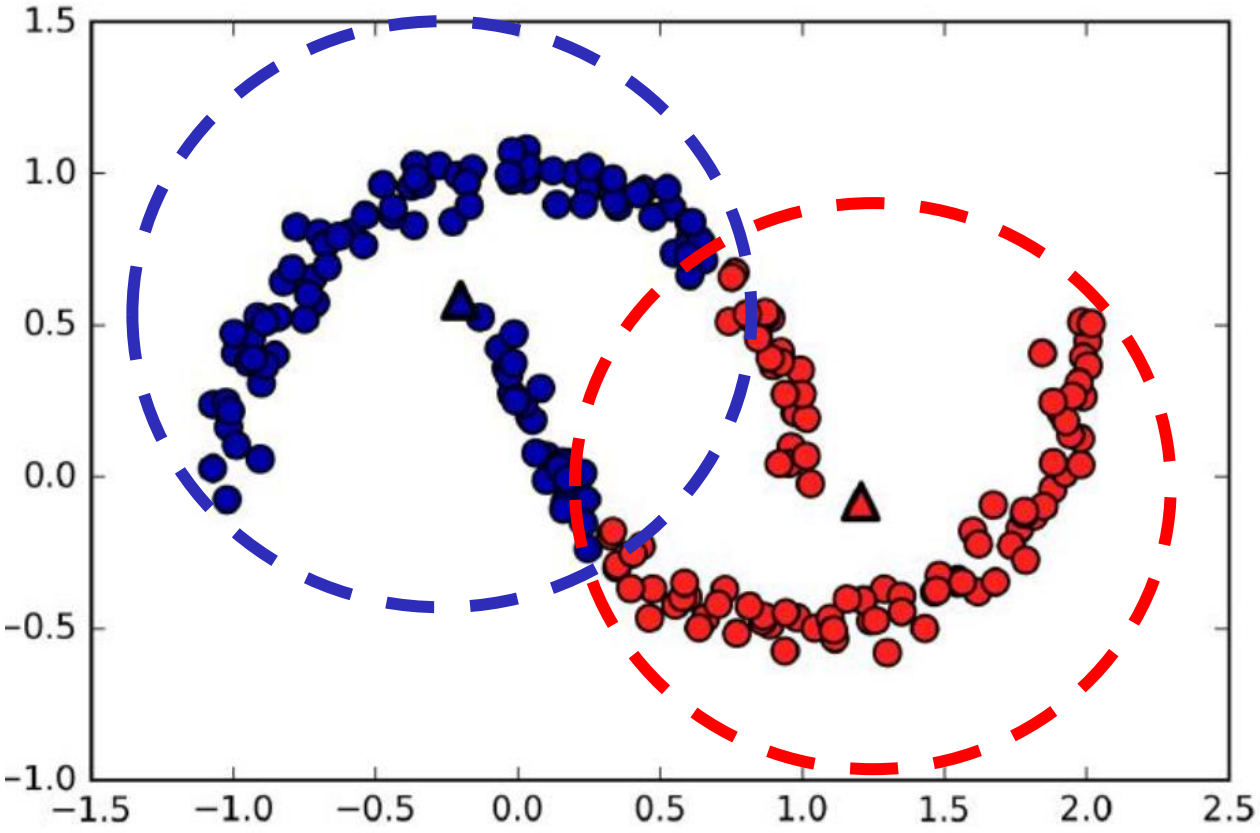
Phương pháp phân cấp [B6TLTK1]

- Có thể dựa trên khoảng cách hoặc mật độ và tính liên tục
- Có thể dùng dendrogram để trực quan hóa cho các tập dữ liệu nhiều chiều



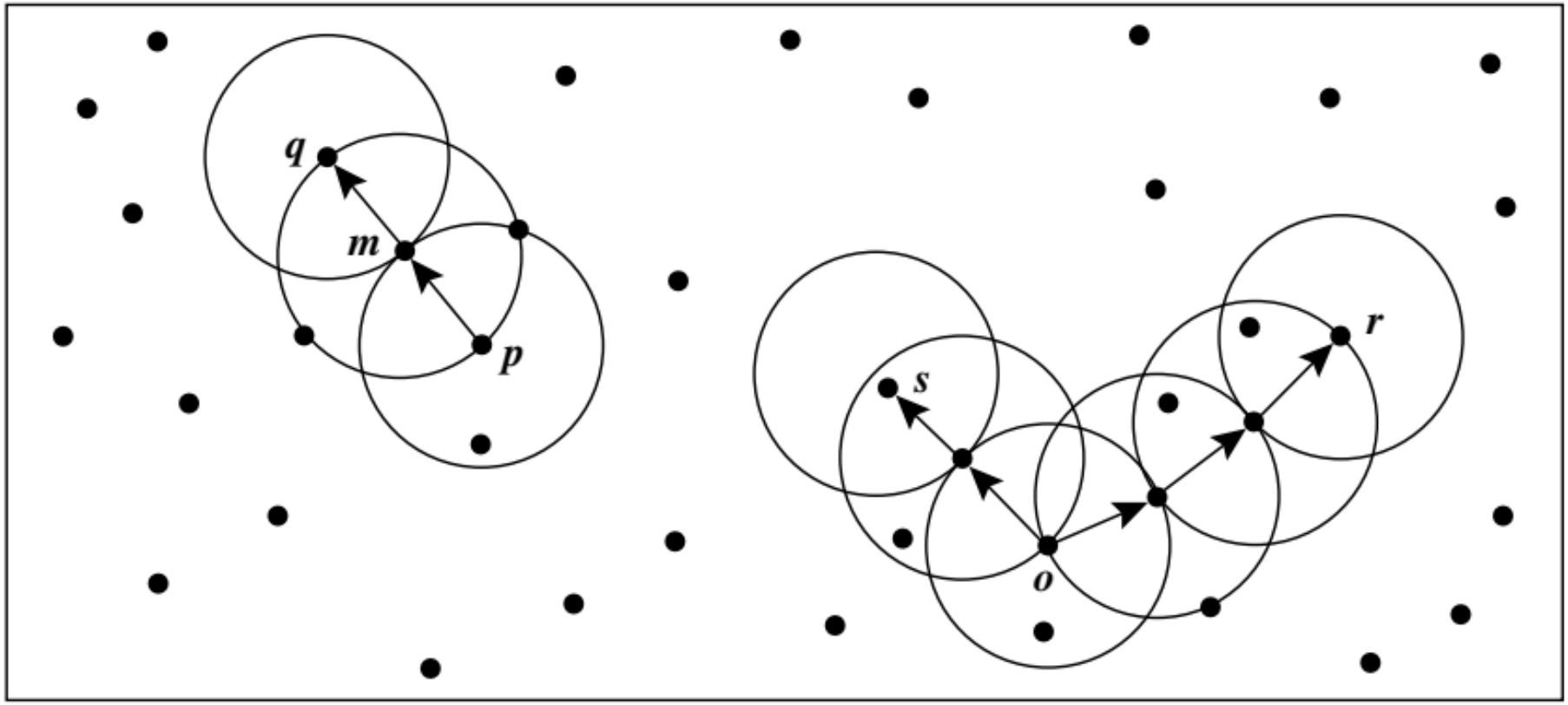
Phương pháp dựa trên mật độ [B6TLTK1]

- Dựa trên “khoảng cách“ có một số hạn chế
- Hình dạng của cụm: bất kỳ?



Phương pháp dựa trên mật độ [B6TLTK1]

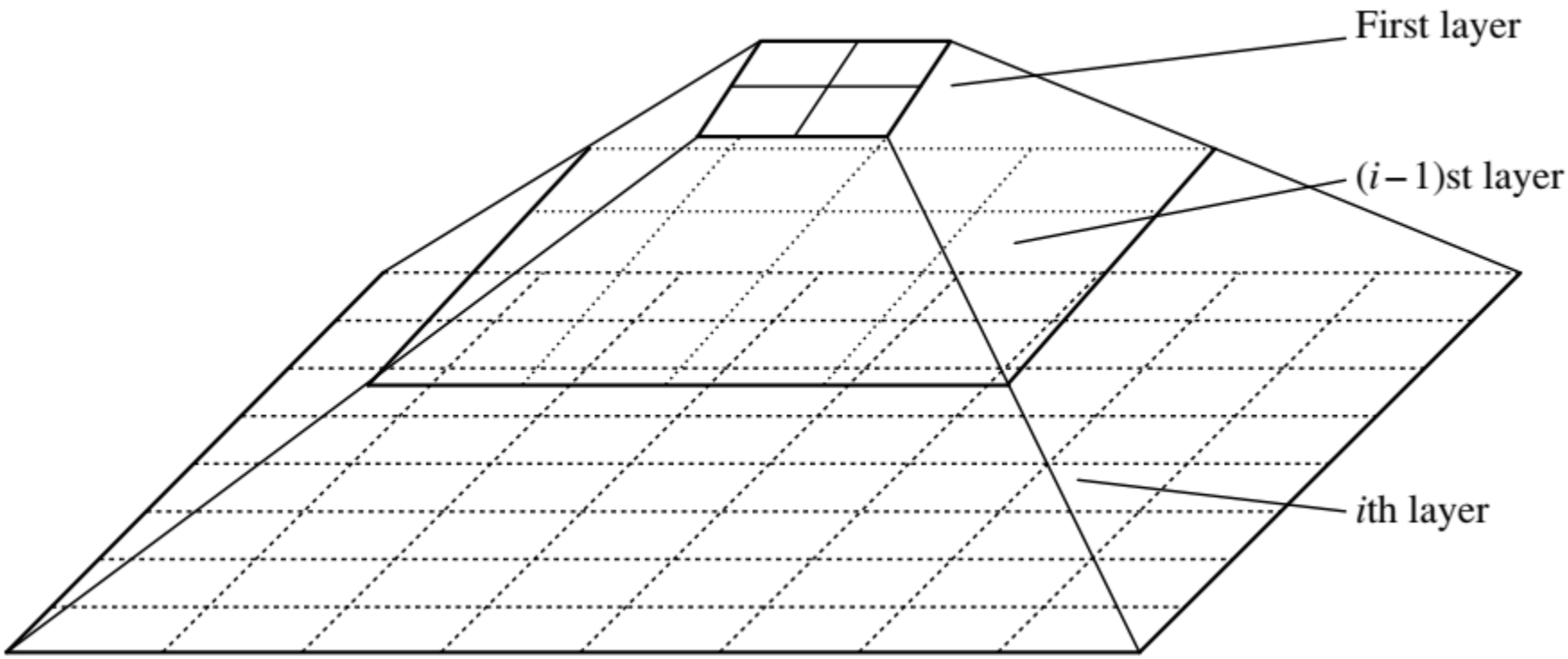
- Cụm: vùng dày đặc trong không gian dữ liệu, được phân tách bằng các vùng thưa thớt.
- Vùng lân cận của một điểm dữ liệu: có chứa một số lượng tối thiểu các điểm dữ liệu khác



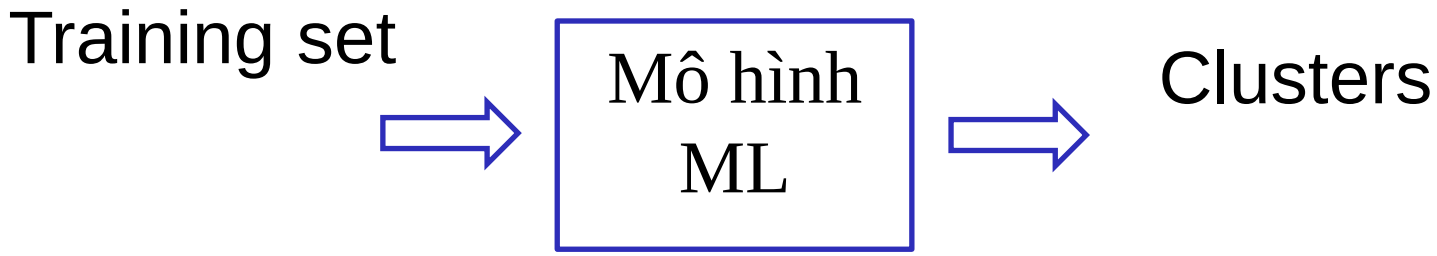
Phương pháp dựa trên lưới [B6TLTK1]

- Tiếp cận dựa trên dữ liệu => cách tiếp cận dựa trên không gian
- Phân vùng không gian nhúng thành các ô độc lập với sự phân bố của các đối tượng đầu vào
- Dựa trên lưới sử dụng cấu trúc dữ liệu lưới đa độ phân giải
  - lượng tử hóa không gian đối tượng thành một số lượng hữu hạn các ô tạo thành một cấu trúc lưới mà trên đó tất cả các hoạt động phân cụm được thực hiện

Phương pháp dựa trên lưới [B6TLTK1]



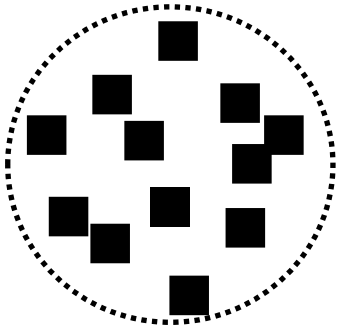
3.3 Hàm mục tiêu

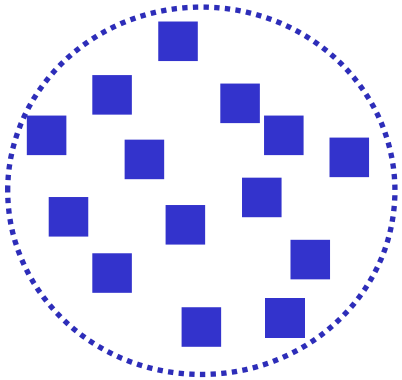


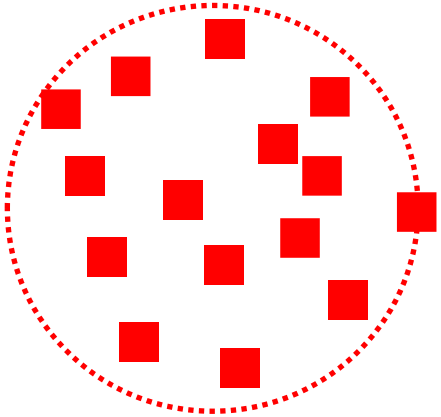
Sự tương đồng giữa các điểm trong cùng 1 cụm lớn nhất

Sự tương đồng giữa các điểm thuộc vào 2 cụm bất kỳ là nhỏ nhất

- Sự tương đồng giữa các điểm







Ví dụ tổng bình phương sai số

- SSE/WCSS (Sum of Squared Errors/ Within-Cluster Sum of Squared)

$$SSE = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d\left(\mathbf{x}, \mathbf{c}_i\right)$$

k: số lượng cụm

$\mathbf{c}_i$: tâm cụm C_i

$d\left(\mathbf{x}, \mathbf{c}_i\right)$: khoảng cách từ $\mathbf{x}$ đến $\mathbf{c}_i$

Tổng kết

- Sinh viên nắm được bài toán phân cụm và ứng dụng

Hoạt động sau buổi học

- Làm bài tập về nhà
- Các nhóm tập trung hoàn thành bài tập lớn
  - Nhóm nào đã hoàn thành có thể đăng ký cụ thể thời gian trình bày kết quả

Chuẩn bị cho buổi học tiếp theo

- Tìm hiểu về thuật toán phân cụm k-means và các biến thể của nó

Tài liệu tham khảo

- [B6TLTK1] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Morgan Kaupmann, 3rd Edition, 2011.