# KHAI PHÁ DỮ LIỆU

## Bài 4. Phân cụm dữ liệu

**Giáo viên: TS. Trần Mạnh Tuấn**

**Bộ môn: Hệ thống thông tin**

**Khoa: Công nghệ thông tin**

**Email: tmtuan@tlu.edu.vn**

**Điện thoai: 0983.668.841**

# Nội dung

- Tổng quan
- Các tiếp cận trong phân cụm
- Các thuật toán phân cụm

# Tổng quan

## Bài toán tình huống – ngoại lai

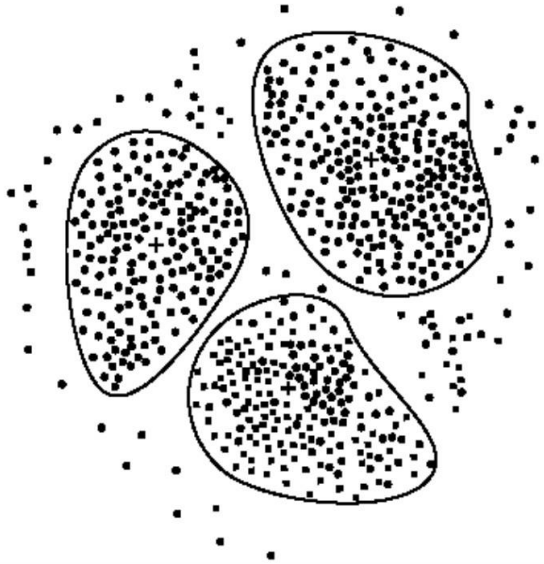



Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?

# Bài toán tình huống – biên và nhiễu

Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)

- Giải pháp giảm thiểu nhiễu
  - Phân tích cụm (cluster analysis)



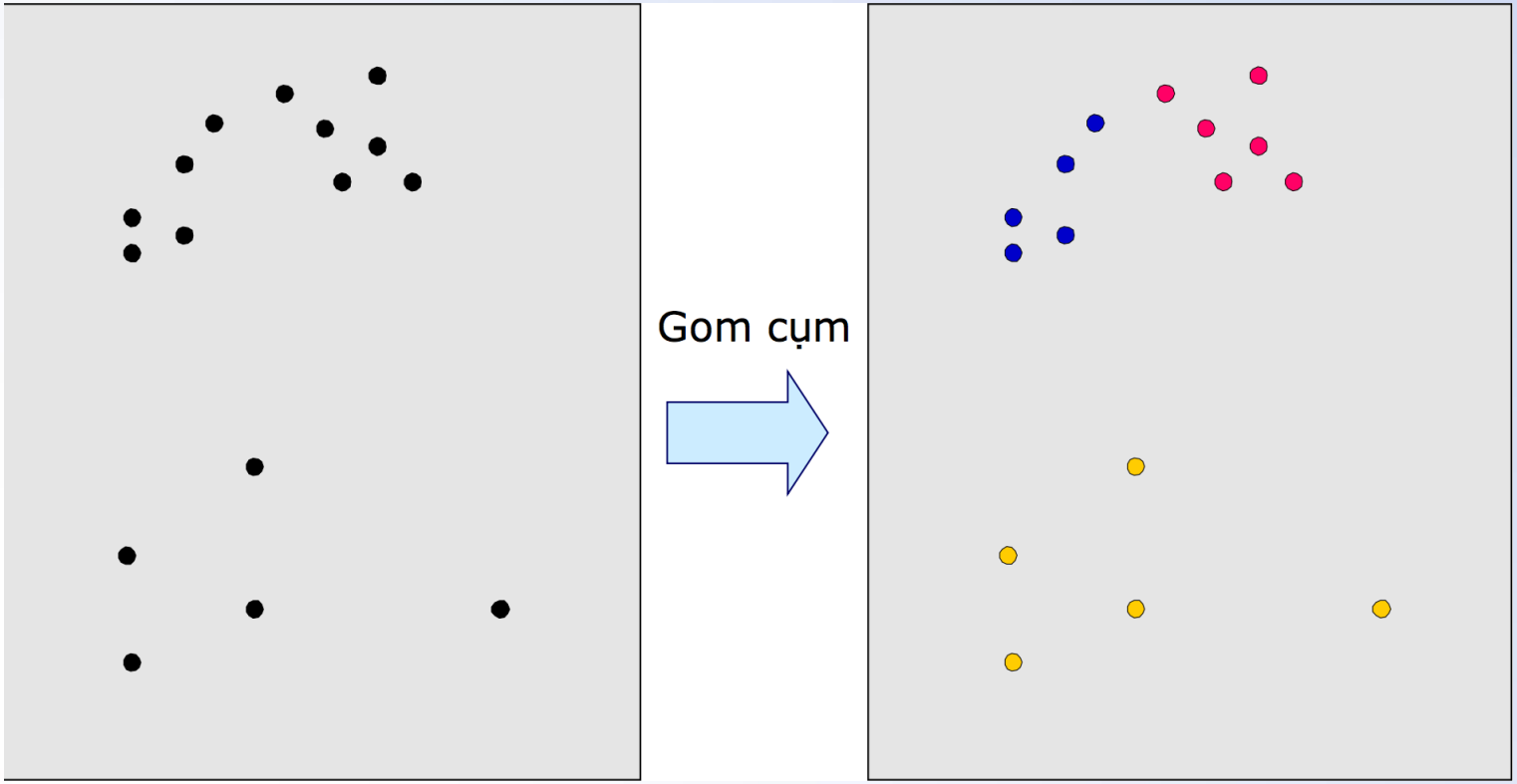
# Tổng quan

## Tình huống – phân cụm ảnh

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

# Tổng quan

## Tình huống



# Tổng quan

- Hỗ trợ giai đoạn tiền xử lý dữ liệu (data preprocessing)
- 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)
- …

# Tổng quan

- PCDL là một lĩnh vực liên ngành đang được phát triển mạnh mẽ. Ở một mức cơ bản nhất, đưa ra định nghĩa PCDL như sau [10][11]:

"PCDL là một kỹ thuật trong DATA MINING, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định"

# Tổng quan

- Như vậy, PCDL là quá trình phân chia một tập DL ban đầu thành các cụm DL sao cho:
  - Các phần tử trong một cụm "tương tự" (Similar) nhau.
  - Các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) nhau.
  - Số các cụm được xác định trước theo kinh nghiệm hoặc tự động.

# Tổng quan

## Các hướng tiếp cận trong phân cụm

- Trong học máy, PCDL được xem là vấn đề học không có giám sát.
  - Nó phải đi giải quyết vấn đề tìm một cấu trúc trong tập hợp các DL chưa biết trước các thông tin về lớp/tập VDHL.
- Nhiều trường hợp, khi phân lớp(Classification) được xem là học có giám sát thì PCDL là một bước trong phân lớp DL.
  - Trong đó PCDL sẽ khởi tạo các lớp cho phân lớp bằng cách xác định các nhãn cho các nhóm dl.

# Tổng quan

## Các hướng tiếp cận trong phân cụm

- Vấn đề thường gặp trong PCDL là hầu hết các DL cần phân cụm đều có DL "nhiễu" (noise) do quá trình thu thập thiếu chính xác, không đầy đủ.
- Cần phải xây dựng chiến lược cho bước tiền xử lý DL để loại bỏ "nhiễu" trước khi bước vào giai đoạn phân tích PCDL.
- Kỹ thuật xử lý nhiễu phổ biến là thay thế giá trị các thuộc tính của đối tượng "nhiễu" bằng giá trị thuộc tính tương ứng của đối tượng DL gần nhất.

# Tổng quan

## Các hướng tiếp cận trong phân cụm

- Tìm phần tử ngoại lai (Outlier) là hướng nghiên cứu quan trọng trong PCDL cũng như trong Data Mining.
- Xác định một nhóm nhỏ các đối tượng DL "khác thường" so với các DL trong để tránh sự ảnh hưởng của chúng tới quá trình và kết quả của PCDL.
- Khám phá các phần tử ngoại lai đã được phát triển và ứng dụng trong viễn thông, dò tìm gian lận thương mại và trong làm sạch dữ liệu,…

# Tổng quan

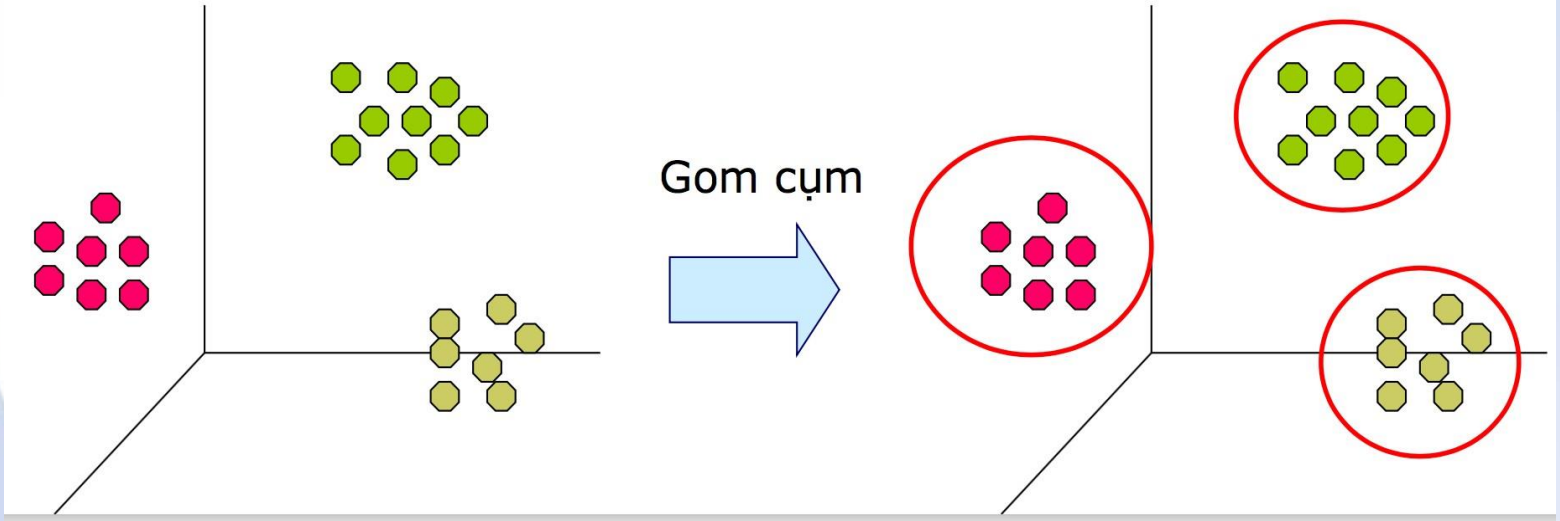
- PCDL là một vấn đề khó, phải giải quyết các vấn đề con cơ bản sau:
  - Xây dựng hàm tính độ tương tự.
  - Xây dựng các tiêu chuẩn phân cụm.
  - Xây dựng mô hình cho cấu trúc cụm dữ liệu.
  - Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo.
  - Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm.

# Tổng quan

- Đến nay chưa có một phương pháp phân cụm tổng quát nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cụm DL.
- Các phương pháp PC cần có cách thức biểu diễn cấu trúc của các cụm DL, với mỗi cách thức biểu diễn sẽ tương ứng một thuật toán PC phù hợp.
- PCDL đang là vấn đề mở và khó, cần giải quyết những vấn đề phù hợp với nhiều dạng DL khác nhau, đặc biệt là DL hỗn hợp, đây cũng là một thách thức lớn trong lĩnh vực Data Mining.

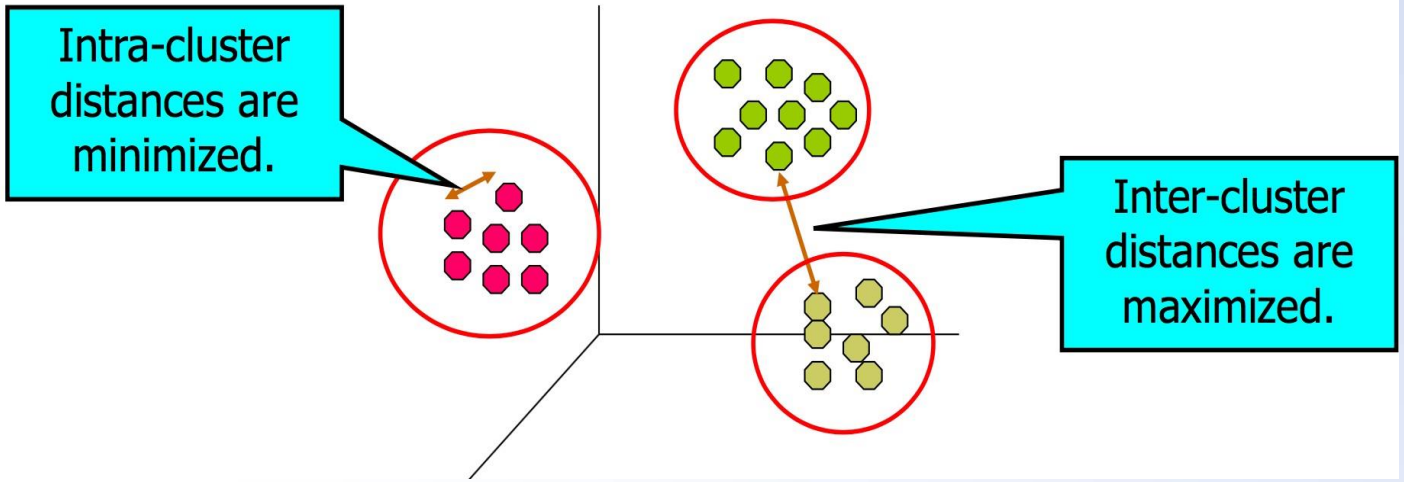
# Tổng quan

- Gom 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.*



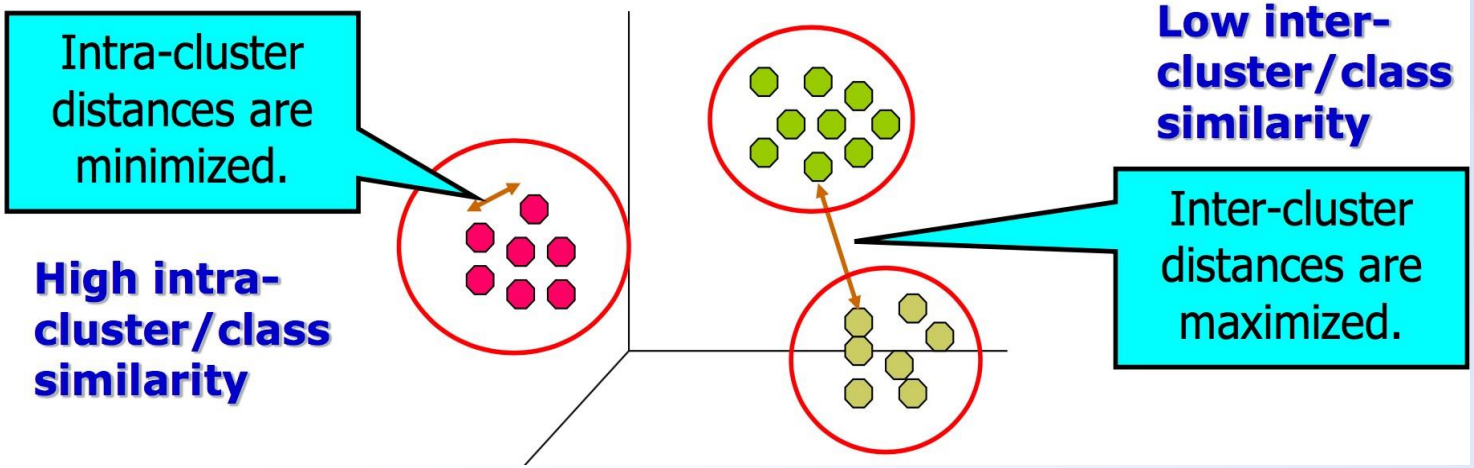
# Tổng quan

- Gom 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.*



# Tổng quan

- Gom 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.*



# Tổng quan

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm
  - Ma trận dữ liệu (data matrix)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

-n đối tượng (objects)

-p biến/thuộc tính (variables/attributes)

# Tổng quan

## Vấn đề kiểu dữ liệu/đối tượng được gom cụm

- Ma trận sai biệt (dissimilarity matrix)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

d(i, j) là khoảng cách giữa đối tượng i và j; thể hiện sự khác biệt giữa đối tượng i và j; được tính tuỳ thuộc vào kiểu của các biến/thuộc tính.

# Tổng quan

## Vấn đề kiểu dữ liệu/đối tượng được gom cụm

d(i, j) là khoảng cách giữa đối tượng i và j; thể hiện sự khác biệt giữa đối tượng i và j; được tính tuỳ thuộc vào kiểu của các biến/thuộc tính.

$$d(i,j) \geq 0$$

$$d(i,i) = 0$$

$$d(i,j) = d(j,i)$$

$$d(i,j) \leq d(i,k) + d(k,j)$$

# Tổng quan

- Độ đo khoảng cách Minkowski

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

- Độ đo khoảng cách Manhattan

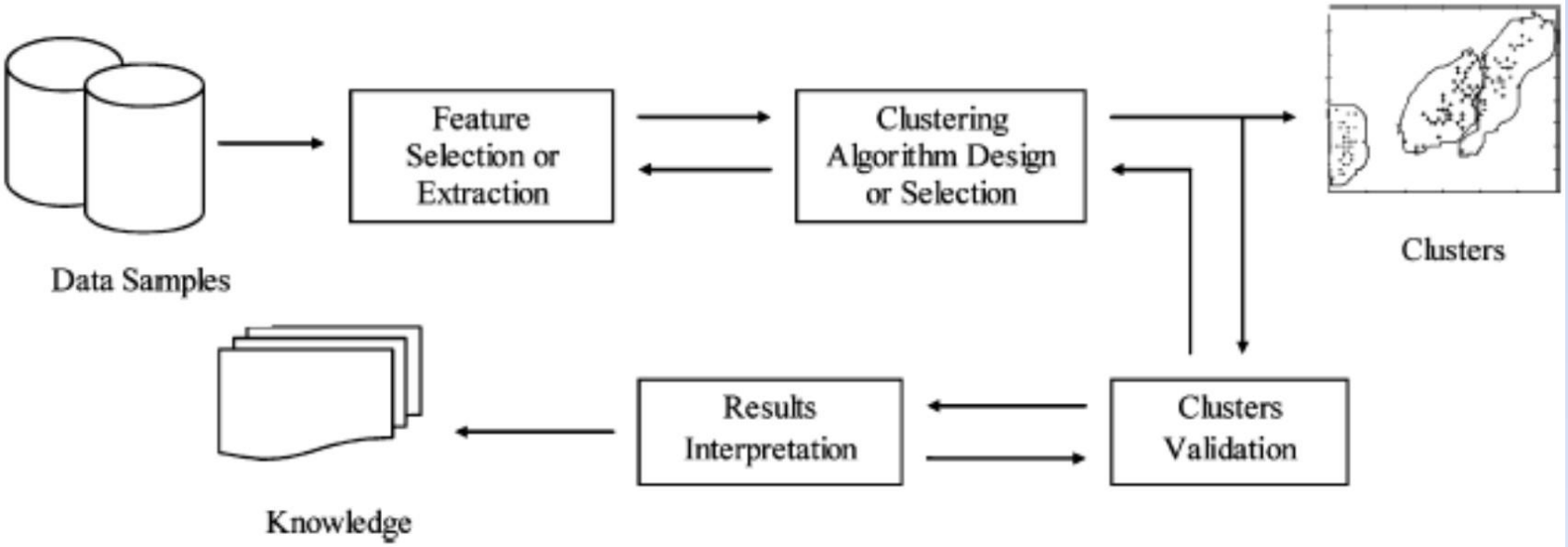
$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

- Độ đo khoảng cách Euclidean

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

# Tổng quan

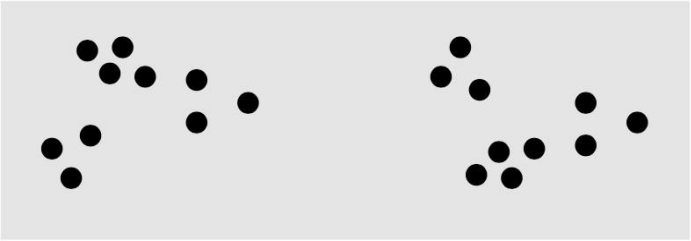
- Quá trình gom cụm dữ liệu



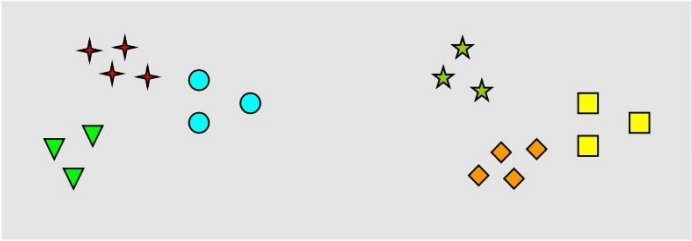

R. Xu, D. Wunsch II. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, 16(3), May 2005, pp. 645-678.

# Tổng quan

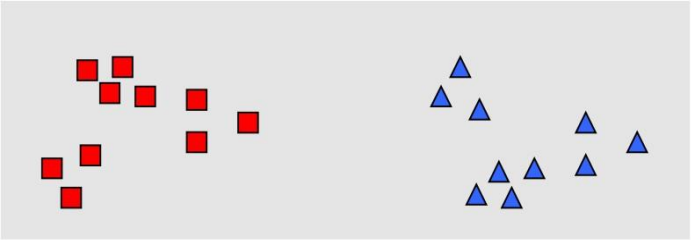
- Mỗi phần tử nên được gom vào bao nhiêu cụm?
- Mỗi cụm nên có bao nhiêu phần tử?
- Bao nhiêu cụm nên được hình thành?



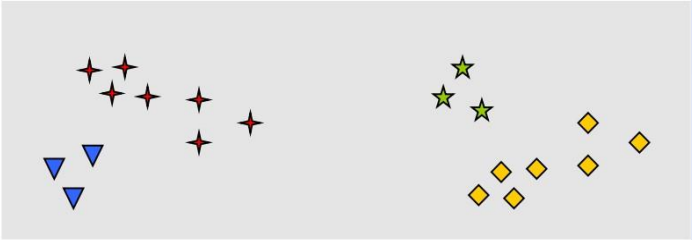
Bao nhiêu cụm?



6 cụm?



2 cụm?



4 cụm?

# Tổng quan

Các yêu cầu tiêu biểu về việc gom cụm dữ liệu

- Khả năng co giãn về tập dữ liệu (scalability)
- Khả năng xử lý nhiều kiểu thuộc tính khác nhau (different types of attributes)
- Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape)
- Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập (domain knowledge for input parameters)
- Khả năng xử lý dữ liệu có nhiễu (noisy data)

3

# Tổng quan

- Khả năng gom cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records)
- Khả năng xử lý dữ liệu đa chiều (high dimensionality)
- Khả năng gom cụm dựa trên ràng buộc (constraint-based clustering)
- Khả diễn và khả dụng (interpretability and usability)

# Tổng quan

Phân loại các phương pháp gom cụm dữ liệu tiêu biểu

- Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
- Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
- Dựa trên mật độ (density-based): dựa trên connectivity and density functions.
- Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.
- Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

# Tổng quan

## Các phương pháp đánh giá việc gom cụm dữ liệu

- Đánh giá ngoại (external validation)
  - Đánh giá kết quả gom cụm dựa vào cấu trúc được chỉ định trước cho tập dữ liệu
- Đánh giá nội (internal validation)
  - Đánh giá kết quả gom cụm theo số lượng các vector của chính tập dữ liệu (ma trận gần – proximity matrix)
- Đánh giá tương đối (relative validation)
  - Đánh giá kết quả gom cụm bằng việc so sánh các kết quả gom cụm khác ứng với các bộ trị thông số khác nhau

→ Tiêu chí cho việc đánh giá và chọn kết quả gom cụm tối ưu

- Độ nén (compactness): các đối tượng trong cụm nên gần nhau.
- Độ phân tách (separation): các cụm nên xa nhau.

## Một số ứng dung

- ✓ **Marketing:** Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, 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;
- ✓ **Biology:** Phân nhóm động vật và thực vật dựa vào các thuộc tính của chúng;

## Một số ứng dung

- ✓ **Libraries:** Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…;
- ✓ **Insurance, Finance:** 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);
- ✓ **WWW:** Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog);…

# Cách tiếp cận phân cụm

- *Phân cụm (clustering):* là tập các phương pháp nhằm tìm ra các nhóm con trong dữ liệu
  - Các mẫu có đặc điểm chung trong cùng 1 nhóm nhưng khác với các mẫu ở ngoài nhóm
  - Việc gom nhóm là phân tích cấu trúc dữ liệu nội tại, điều này khác với phân lớp

# Cách tiếp cận phân cụm

## Phân cụm là gì?

➢ Là quá trình phân chia 1 tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn:

- Các đối tượng trong 1 cụm "tương tự" nhau.
- Các đối tượng khác cụm thì "không tương tự" nhau.

➢ Mục đích: giải quyết vấn đề tìm kiếm, phát hiện các cụm, các mẫu dữ liệu trong 1 tập hợp ban đầu các dữ liệu không có nhãn.

# Mục đích của phân cụm

- Xác định được bản chất của việc nhóm các đối tượng trong 1 tập dữ liệu không có nhãn.
- Phân cụm không dựa trên 1 tiêu chuẩn chung nào, mà dựa vào tiêu chí mà người dùng cung cấp trong từng trường hợp.

# Các phương pháp phân cụm

- Các kỹ thuật đều hướng tới:
  - Chất lượng của các cụm
  - Tốc độ thực hiện của thuật toán

1. Phân cụm phân hoạch
2. Phân cụm phân cấp
3. Phân cụm dựa trên mật độ
4. Phân cụm dựa trên lưới
5. Phân cụm dựa trên mô hình
6. Phân cụm có ràng buộc

# Các thuật toán phân cụm

## Phân cụm K-means

- Các tâm cụm cực tiểu sự biến đổi giữa các cụm

$$J = \frac{1}{n} \sum_{i=1}^{K} \sum_{x \in c_i} |x - \mu_i|^2 \longrightarrow \text{MIN}$$

  - Các tâm cụm (trung tâm của cụm): $\mu_i = \frac{1}{|c_i|} \sum_{x \in c_i} x$

- Bài toán cực tiểu hóa này là tối ưu tổ hợp

  Giải pháp cho cực tiểu hóa địa phương ta sử dụng phương pháp lặp

# Các thuật toán phân cụm

## Thuật toán K–means

### *Input*

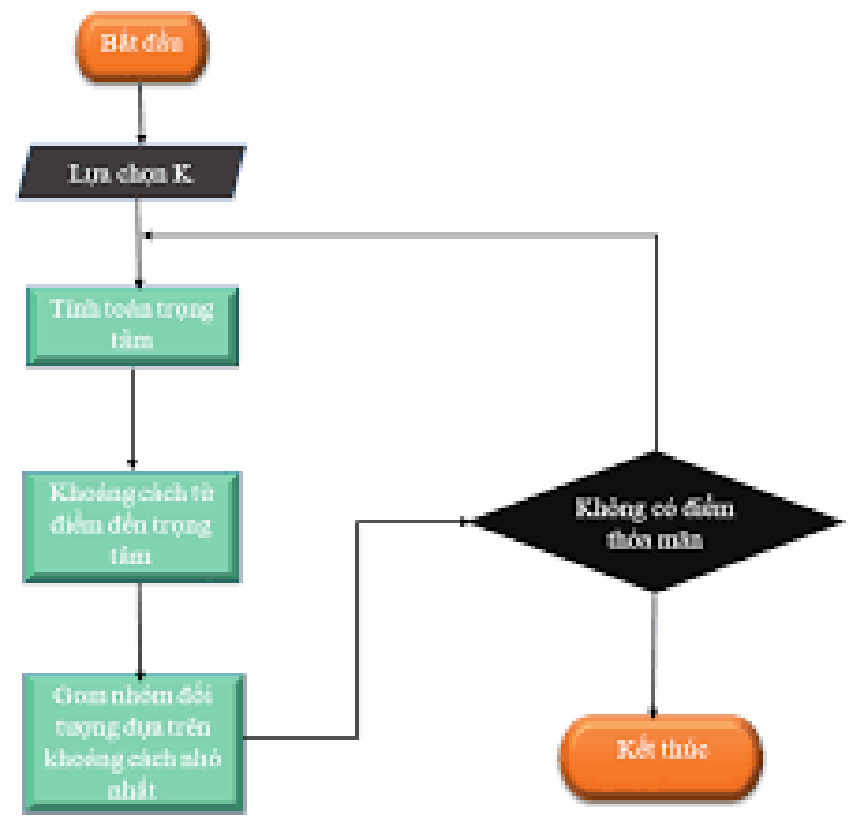
- Tập mẫu $X = \{x_i | i = 1, 2, \ldots, N\}, x_i \in R^d$
- Số cụm: K

### *Output*

Các cụm $C_k$ ( $k = 1 \div K$) tách rời và hàm mục tiêu J đạt cực tiểu

# Các thuật toán phân cụm

## Thuật toán K-means



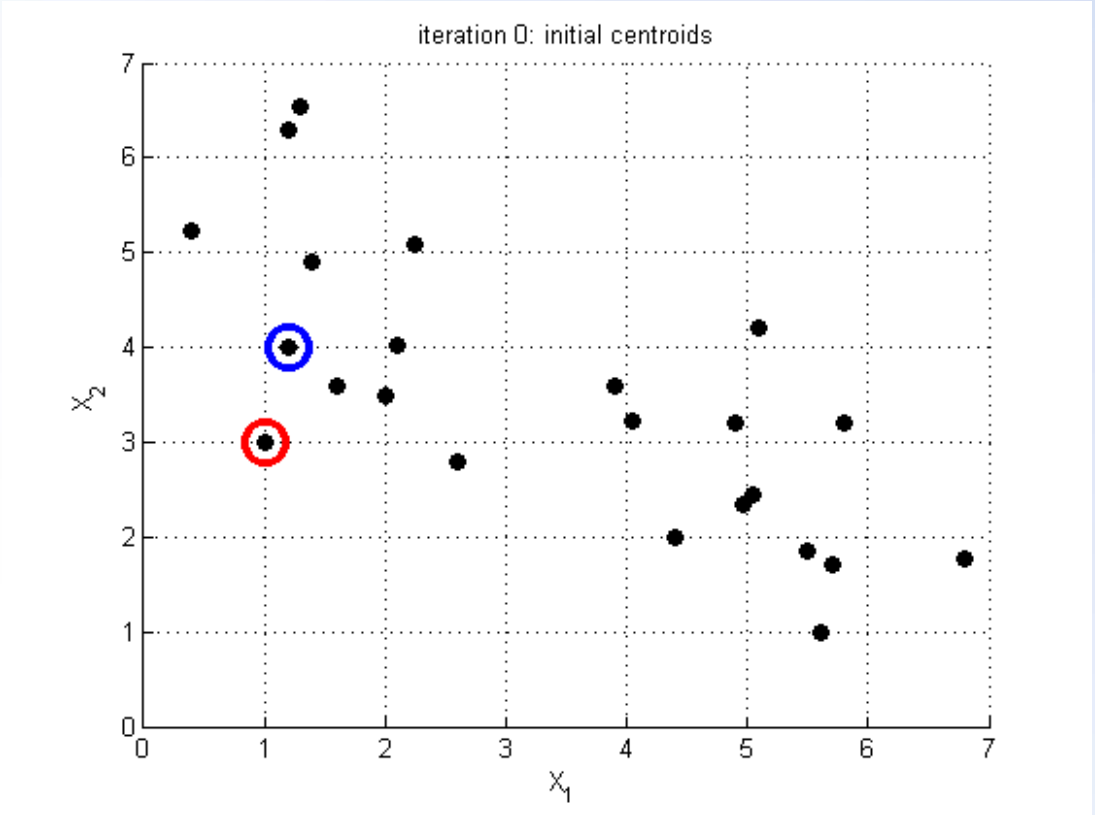
# Các thuật toán phân cụm

## Thuật toán K-means

1) Khởi tạo: Chọn **ngẫu nhiên** K tâm cụm
2) Tính toán khoảng cách từ các đối tượng đến các tâm để phân hoạch dữ liệu (bằng cách gán mỗi đối tượng vào cụm mà nó gần tâm nhất)
3) Tính lại các tâm cụm mới trong mỗi cụm
4) Lặp lại 2 và 3 cho đến khi "thỏa mãn điều kiện" (khi các tâm cụm ổn định và các đối tượng không dịch chuyển giữa các cụm)

# Các thuật toán phân cụm

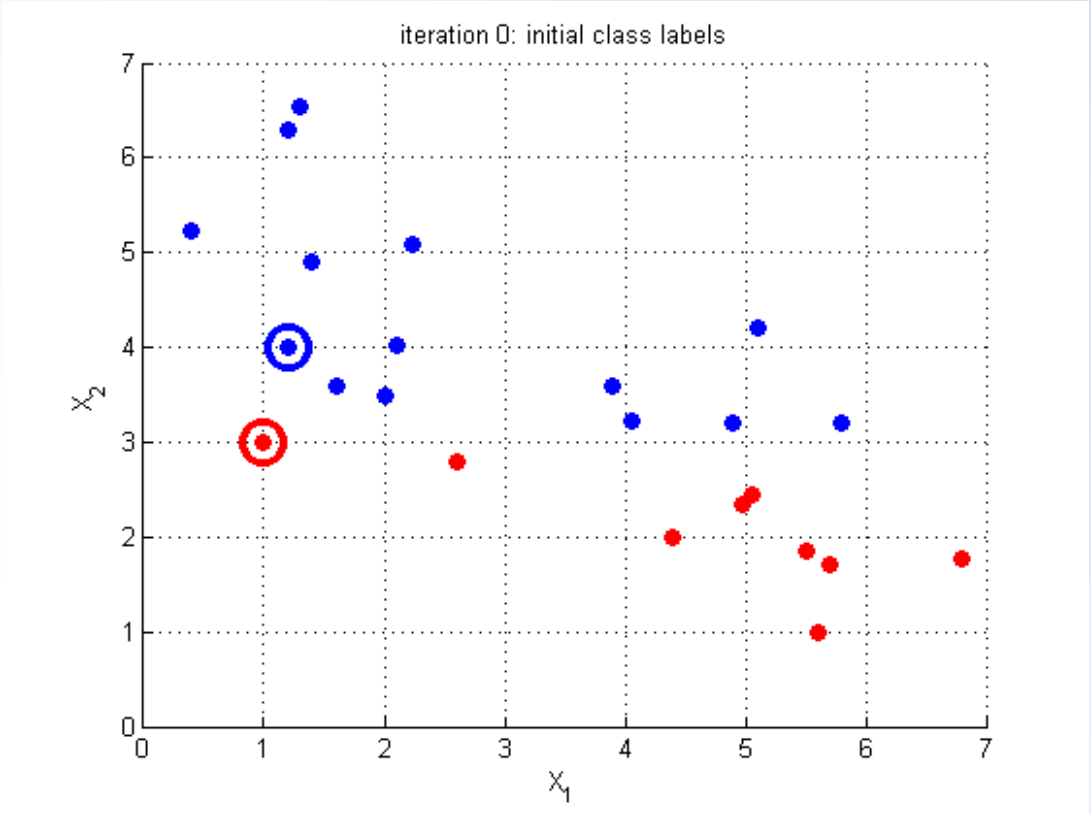
## Thuật toán K–means




Khởi tạo tâm cụm

# Các thuật toán phân cụm

## Thuật toán K–means

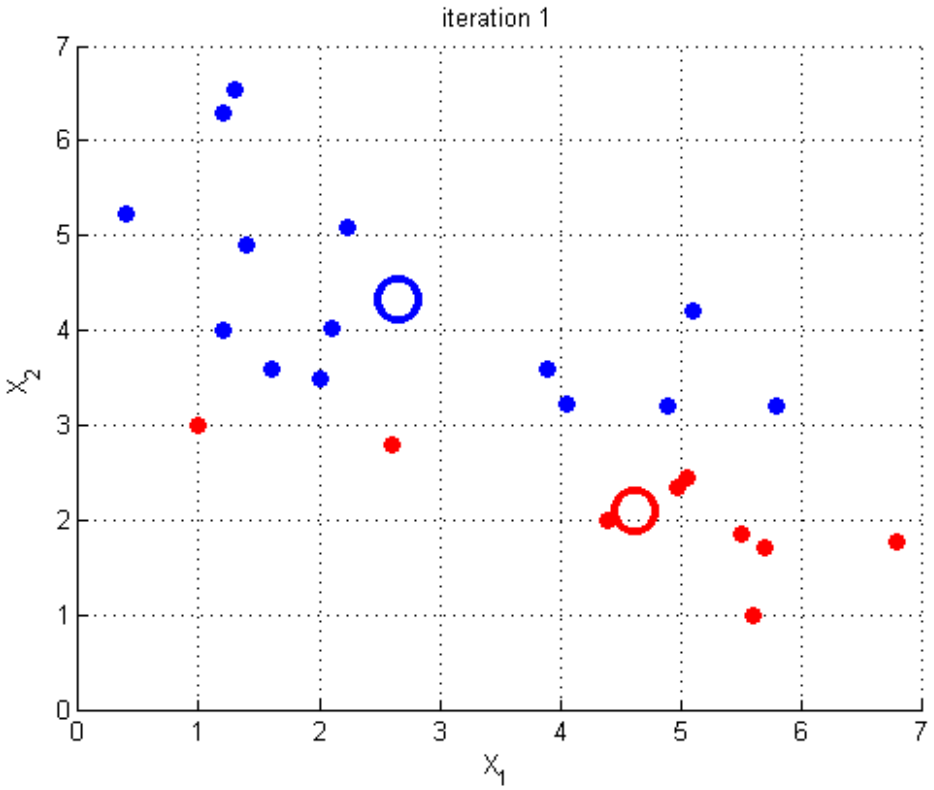



Khởi tạo tâm cụm

**Gán các cụm ban đầu**

# Các thuật toán phân cụm

## Thuật toán K-means



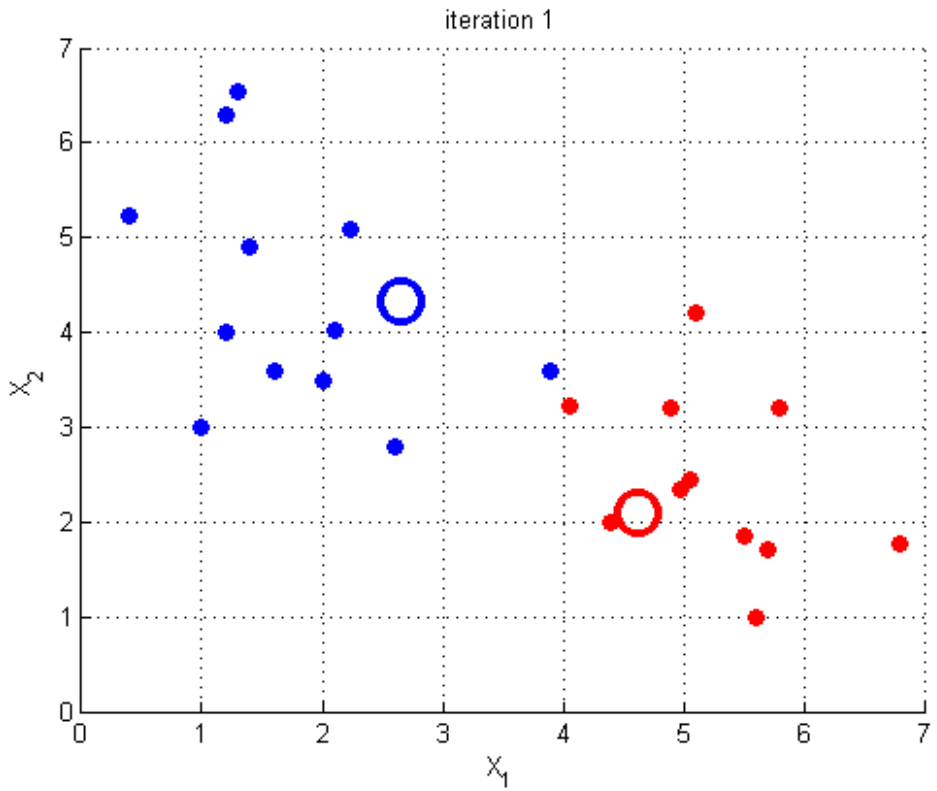

Khởi tạo tâm cụm

Gán các cụm ban đầu

**Cập nhật các tâm cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

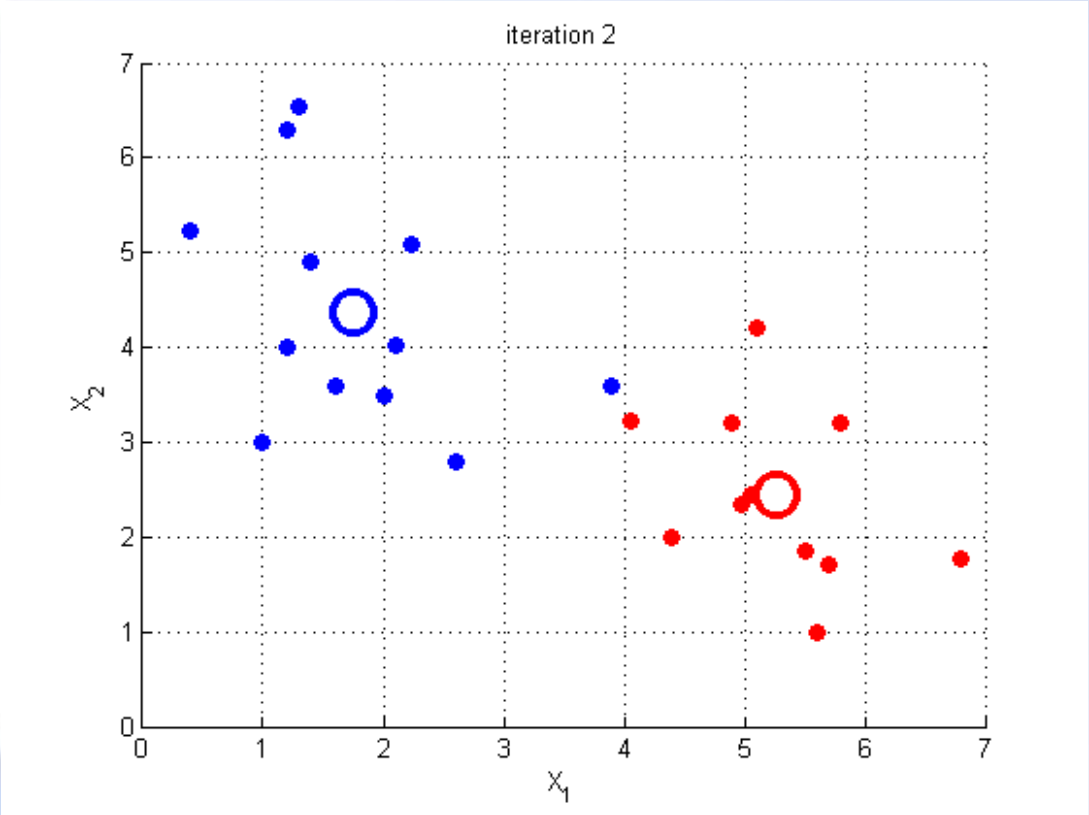
Gán các cụm ban đầu

Cập nhật các tâm cụm

**Gán lại các cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

Gán các cụm ban đầu

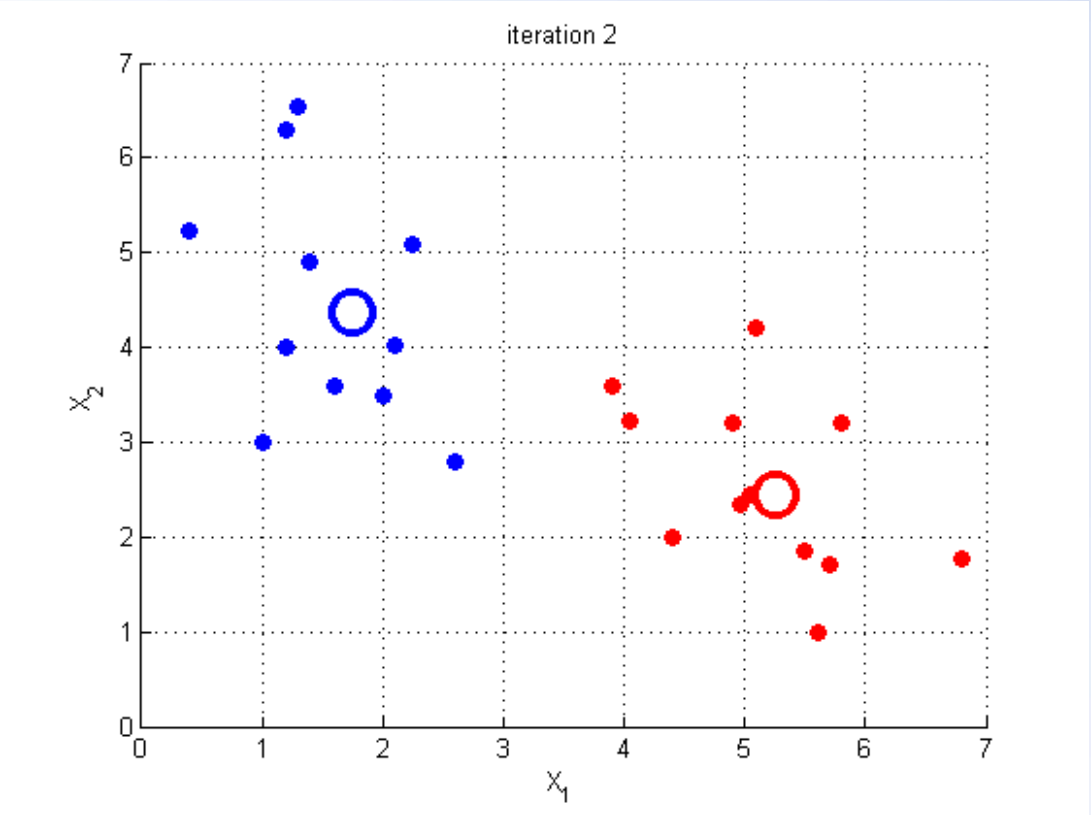
Cập nhật các tâm cụm

Gán lại các cụm

**Cập nhật tâm cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

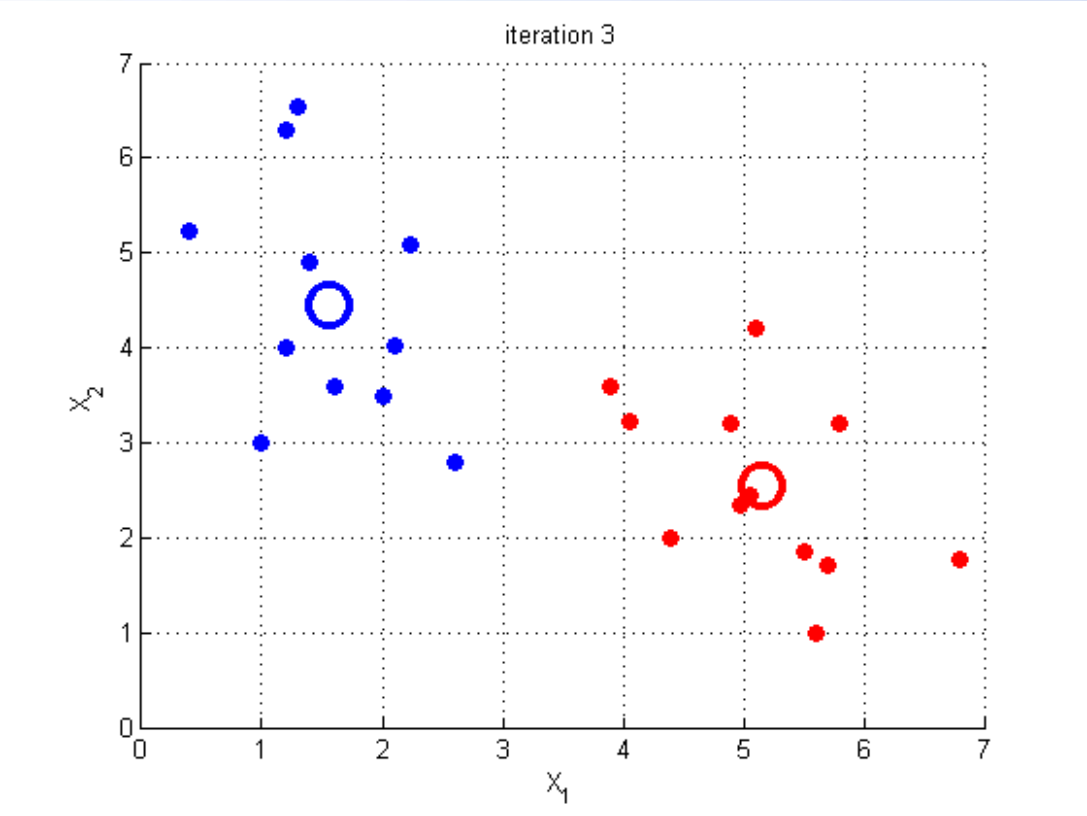
Gán lại các cụm

Cập nhật tâm cụm

**Gán lại các cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

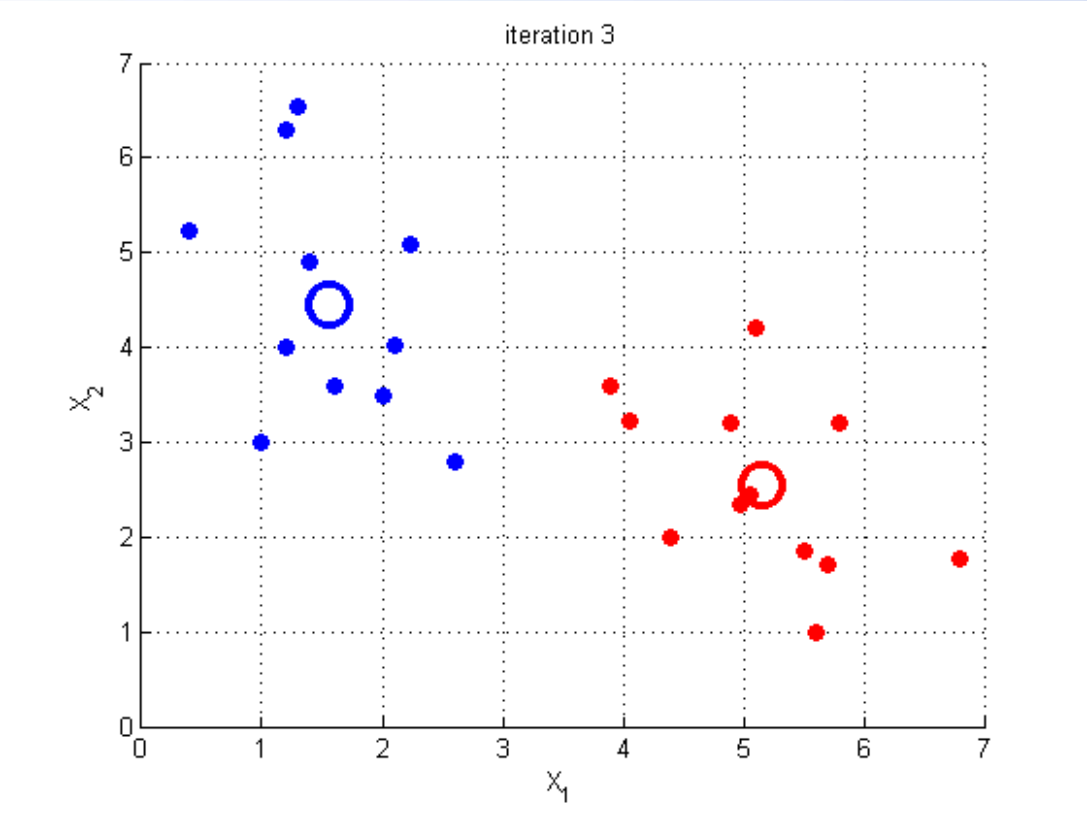
Cập nhật tâm cụm

Gán lại các cụm

**Cập nhật tâm cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

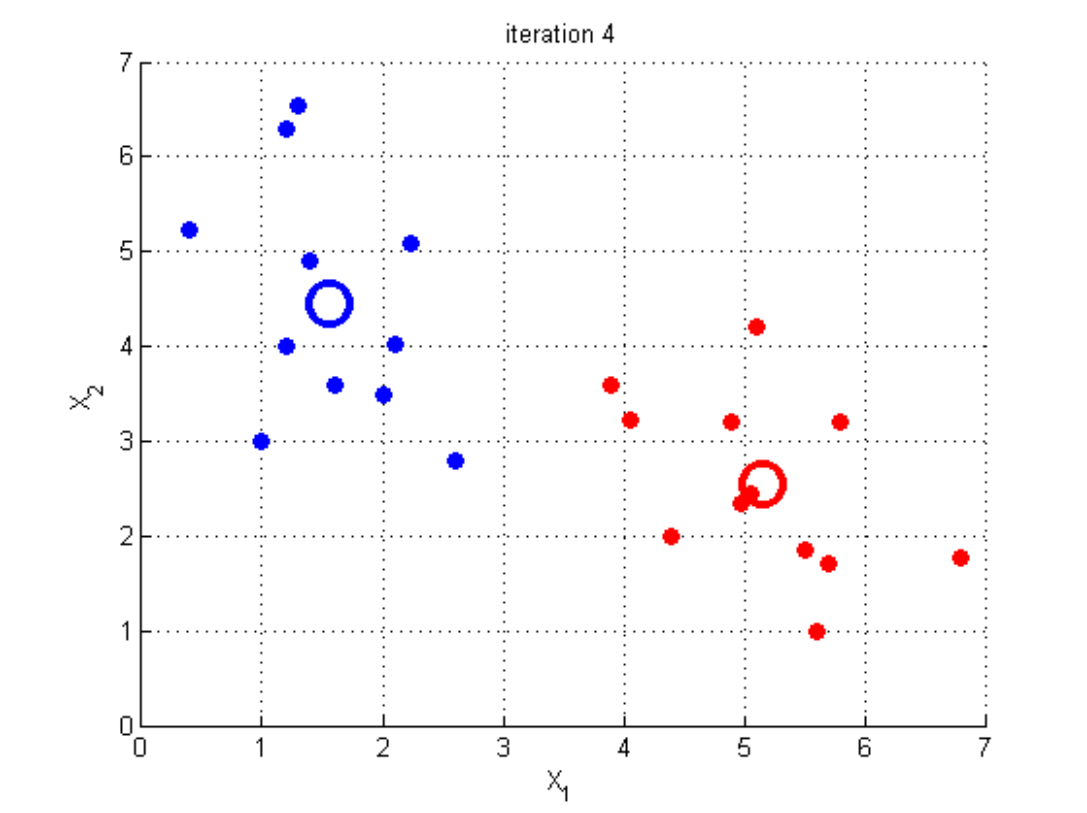
Gán lại các cụm

Cập nhật tâm cụm

**Gán lại các cụm**

# Các thuật toán phân cụm

## Thuật toán K–means




Khởi tạo tâm cụm

Gán các cụm ban đầu

Cập nhật các tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

Gán lại các cụm

Cập nhật tâm cụm

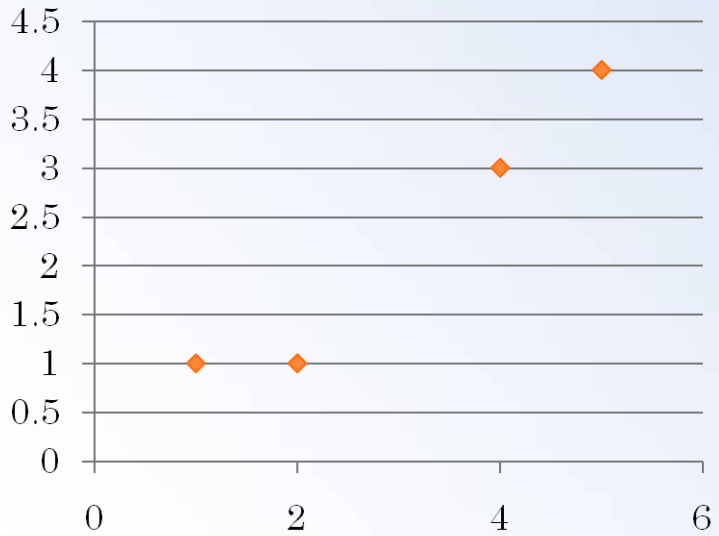
Gán lại các cụm

**Thỏa mãn điều kiện**

# Các thuật toán phân cụm

## VÍ DỤ: KHỞI TẠO TÂM C1 = A, C2 = B. ÁP DỤNG K-means CHO DỮ LIỆU SAU

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



# Các thuật toán phân cụm

## ví dụ minh họa

- **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**

- **Bước 2:** Tính toán khoảng cách

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

    $= 3.61$

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

    $= 2.83$

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

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

    $= 5$

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

    $= 4.24$

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

# Các thuật toán phân cụm

## ví dụ minh họa

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

$$= \left(\frac{11}{3}, \frac{8}{3}\right)$$

# Các thuật toán phân cụm

## ví dụ minh họa

- **Bước 4-1: Lặp lại bước 2 – Tính toán khoảng cách**
- **$d(A, c_1) = 0 < d(A, c_2) = 3.14$**
  **A thuộc cụm 1**
- **$d(B, c_1) = 1 < d(B, c_2) = 2.36$**
  **B thuộc cụm 1**
- **$d(C, c_1) = 3.61 > d(C, c_2) = 0.47$**
  **C thuộc cụm 2**
- **$d(D, c_1) = 5 > d(D, c_2) = 1.89$**
  **D thuộc cụm 2**

# Các thuật toán phân cụm

## ví dụ minh họa

- Bước 4-2: Lặp lại bước 2
- $d(A, c_1) = 0.5 < d(A, c_2) = 4.3$

  A thuộc cụm 1
- $d(B, c_1) = 0.5 < d(B, c_2) = 3.54$

  B thuộc cụm 1
- $d(C, c_1) = 3.2 > d(C, c_2) = 0.71$

  C thuộc cụm 2
- $d(D, c_1) = 4.61 > d(D, c_2) = 0.71$

  D thuộc cụm 2

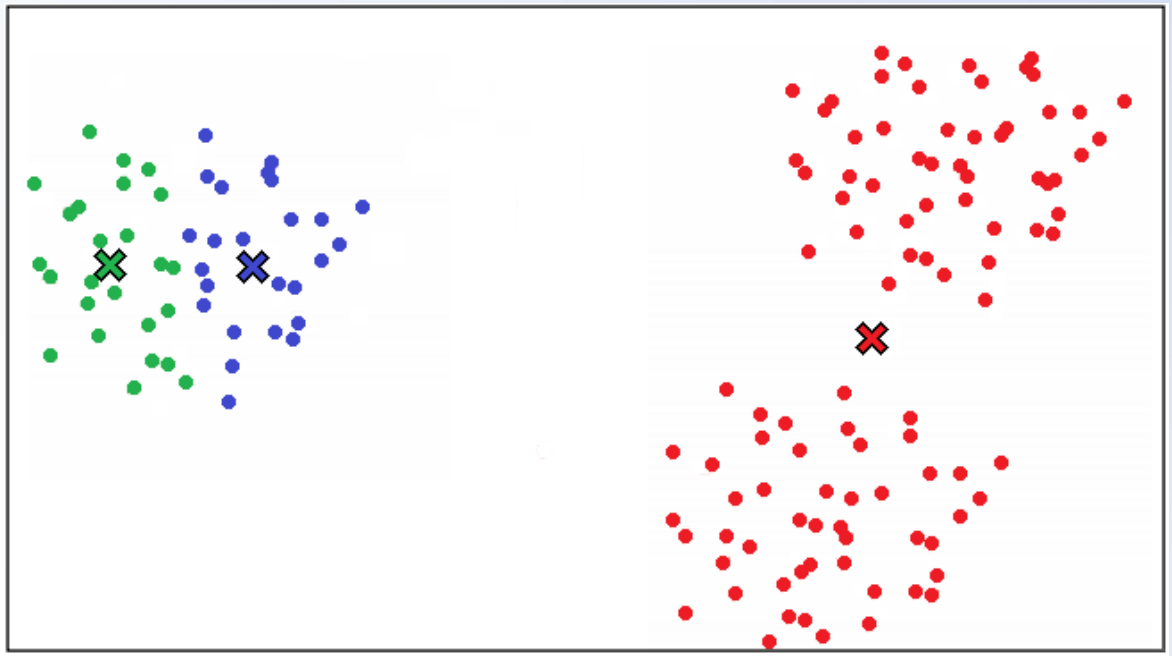
=> Vì không có sự thay đổi trọng tâm của cụm nên thuật toán dừng.

- Với: cụm 1 gồm: A,B cụm 2 gồm: C,D

# Các thuật toán phân cụm

## Thuật toán K–means

- Khởi tạo không tốt dẫn đến kết quả phân cụm kém



# Phân cụm FCM

**Phương pháp phân cụm**

- Phân cụm rõ: dữ liệu được chia vào các cụm, trong đó mỗi điểm dữ liệu thuộc vào chính xác một cụm.
- Phân cụm mờ: các điểm dữ liệu có thể thuộc vào nhiều hơn một cụm và tương ứng với các điểm dữ liệu là ma trận độ thuộc.
- Phân cụm mờ bán giám sát: là phân cụm mờ kết hợp với các thông tin bổ trợ hình thành lên nhóm các thuật toán gọi là phân cụm mờ bán giám sát.

# Các thuật toán phân cụm

## ❖ Thuật toán Fuzzy C-means

- **Hàm mục tiêu**

$$J = \sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^{m} \left\| X_k - V_j \right\|^2 \rightarrow \min$$

- **Điều kiện ràng buộc**

$$\sum_{j=1}^{C} u_{kj} = 1; \qquad u_{kj} \in [0,1]; \qquad \forall k = \overline{1, N}$$

- **Tính tâm cụm**

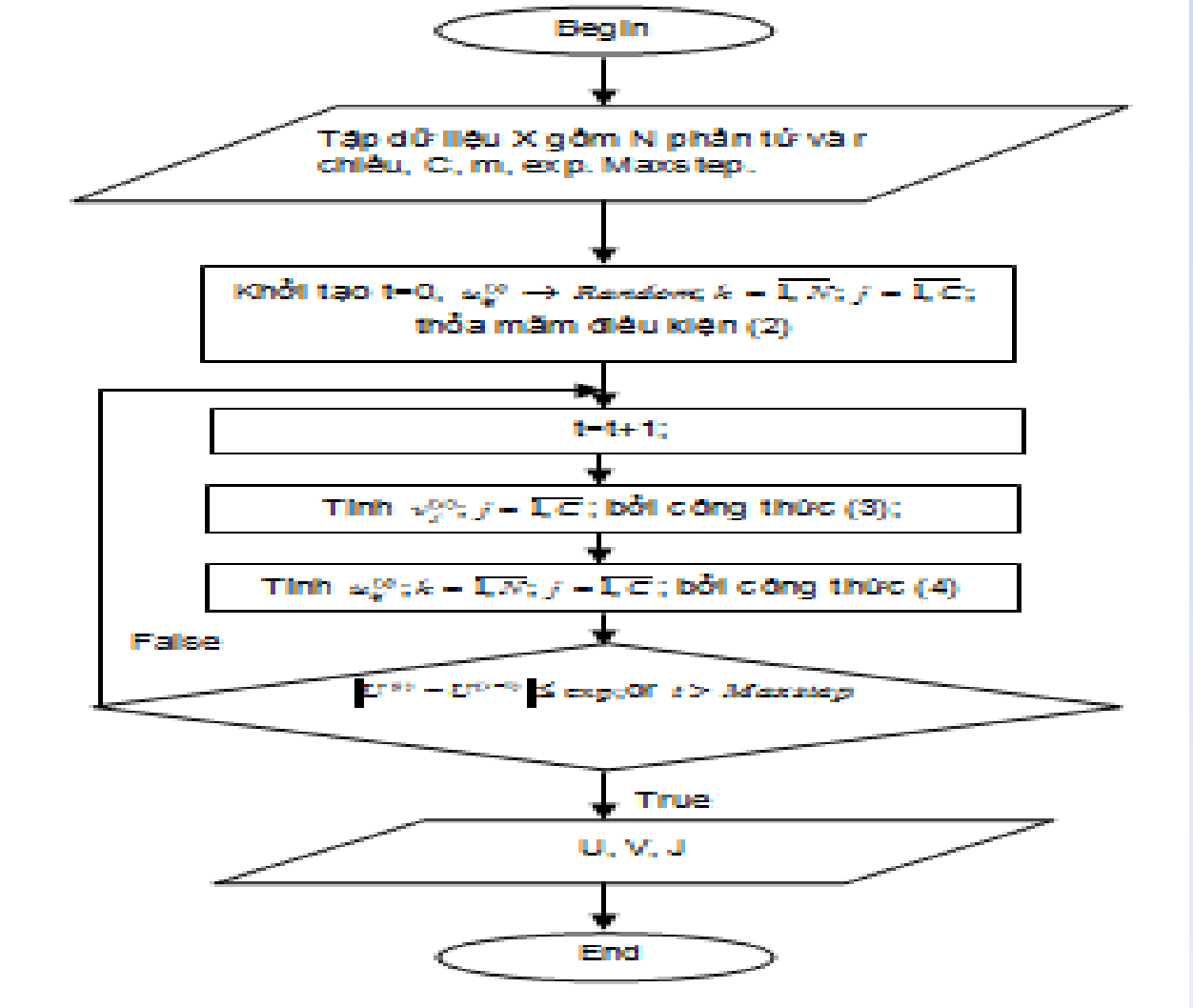
$$V_j = \frac{\sum_{k=1}^{C} u_{kj}^{m} X_k}{\sum_{k=1}^{C} u_{kj}^{m}}$$

- **Tính hàm mức độ thành viên**

$$u_{kj} = \frac{1}{\sum_{i=1}^{C} \left( \frac{\left\| X_k - V_j \right\|}{\left\| X_k - V_i \right\|} \right)^{\frac{1}{m-1}}}$$

# Các thuật toán phân cụm

## ❖ Thuật toán Fuzzy C-means



# Các thuật toán phân cụm

## ❖ Thuật toán Fuzzy C-means

| | |
|---|---|
| Input | Tập dữ liệu X gồm N phần tử trong không gian r chiều; số cụm C; mờ hóa m; ngưỡng ε; số lần lặp lớn nhất MaxStep>0. |
| Output | Ma trận U và tâm cụm V. |
| FCM | |
| 1 | t=0 |
| 2 | $u_{kj}^{(t)} \leftarrow random \quad \left(k=\overline{1,N}; j=\overline{1,C}\right)$ thỏa mãn điều kiện (2.27) |
| 3 | Lặp lại |
| 4 | t=t+1 |
| 5 | Tính $V_j^{(t)}; \left(j=\overline{1,C}\right)$ bởi công thức (2.28) |
| 6 | Tính $u_{kj}^{(t)}; \left(k=\overline{1,N}; j=\overline{1,C}\right)$ bởi công thức (2.29) |
| 7 | Cho đến khi: $\left\|U^{(t)}-U^{(t-1)}\right\| \le \varepsilon$ hoặc t > MaxStep |

# Các thuật toán phân cụm

## ❖ Thuật toán Fuzzy C-means

| | |
|---|---|
| Input | Tập dữ liệu X gồm N phần tử trong không gian r chiều; số cụm C; mờ hóa m; ngưỡng ε; số lần lặp lớn nhất MaxStep>0. |
| Output | Ma trận U và tâm cụm V. |
| FCM | |
| 1 | t=0 |
| 2 | $u_{kj}^{(t)} \leftarrow random \quad \left(k=\overline{1,N}; j=\overline{1,C}\right)$ thỏa mãn điều kiện (2.27) |
| 3 | Lặp lại |
| 4 | t=t+1 |
| 5 | Tính $V_j^{(t)}; \left(j=\overline{1,C}\right)$ bởi công thức (2.28) |
| 6 | Tính $u_{kj}^{(t)}; \left(k=\overline{1,N}; j=\overline{1,C}\right)$ bởi công thức (2.29) |
| 7 | Cho đến khi: $\left\|U^{(t)}-U^{(t-1)}\right\| \le \varepsilon$ hoặc t > MaxStep |

# Các thuật toán phân cụm

## Tổng quan về phân cụm mờ bán giám sát

Thông tin bổ trợ trong phân cụm mờ bán giám sát, có 3 loại cơ bản[31]:

- Các ràng buộc Must-link và Cannot-link;
- Các nhãn lớp của một phần dữ liệu;
- Độ thuộc được xác định trước.

Trong bài báo này nhóm nghiên cứu sử dụng thông tin là giá trị hàm độ thuộc nhận được sau khi sử dụng thuật toán phân cụm FCM.

# Các thuật toán phân cụm

## ❖ SEMI-SUPERVISED STANDARD FUZZY CLUSTERING[29] (SSSFC)

- **Hàm mục tiêu**

$$J(U,V)=\sum_{k=1}^{N}\sum_{j=1}^{C}|u_{kj}-\bar{u}_{kj}|^{m}\,\|X_k-V_j\|^2\rightarrow \min \qquad (5)$$

- **Thông tin bổ trợ**

$$\overline{U}=\left\{\overline{u_{kj}}\,|\,\overline{u_{kj}}\in[0,1],k=\overline{1,N},j=\overline{1,C}\right\}\quad \sum_{j=1}^{C}\overline{u_{kj}}\le 1\quad \left(\forall k=\overline{1,N}\right)$$

- **Tính tâm cụm**

$$V_j=\frac{\sum_{k=1}^{N}\left|u_{kj}-\bar{u}_{kj}\right|^{m}X_k}{\sum_{k=1}^{N}\left|u_{kj}-\bar{u}_{kj}\right|^{m}},\quad j=\overline{1,\mathrm{C}} \qquad (6)$$

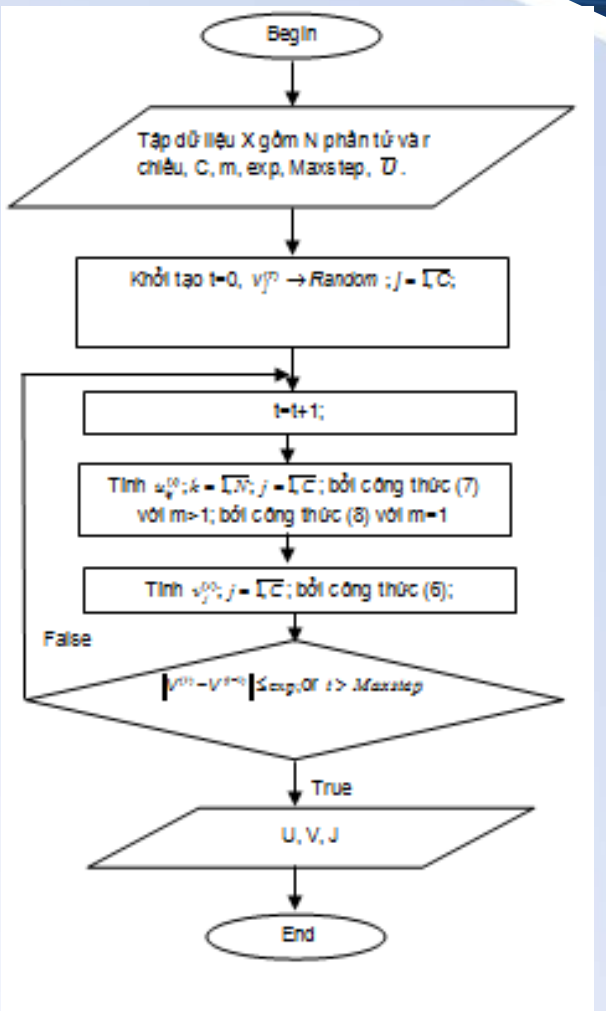
- **Tính hàm mức độ thành viên**

m>1

$$u_{kj}=\overline{u_{kj}}+\left(1-\sum_{i=1}^{C}\overline{u_{kj}}\right)\frac{\left(\frac{1}{\|X_k-V_j\|}\right)^{\frac{2}{m-1}}}{\sum_{i=1}^{C}\left(\frac{1}{\|X_k-V_i\|}\right)^{\frac{2}{m-1}}} \qquad (7)$$

m=1

$$u_{kj}=\begin{cases}\overline{u_{kj}}+1-\sum_{j=1}^{C}\overline{u_{kj}}, & k=\arg\min_i\|X_k-V_i\|^2\\ \overline{u_{kj}}, & \text{otherwise.}\end{cases} \qquad (8)$$



# Các thuật toán phân cụm

## SEMI-SUPERVISED ENTROPY REGULARIZED FUZZY CLUSTERING[29] (eFCM)

- **Hàm mục tiêu**

$$J(U,V)=\sum_{k=1}^{N}\sum_{j=1}^{C}u_{kj}\left\|X_k-V_j\right\|_A^2+\lambda^{-1}\sum_{k=1}^{N}\sum_{j=1}^{C}\left(\left|u_{kj}-\overline{u_{kj}}\right|\ln\left|u_{kj}-\overline{u_{kj}}\right|\right)\rightarrow\min \qquad (13)$$

- **Độ đo Mahalanobis**

$$P=\frac{1}{N}\sum_{j=1}^{C}\sum_{k=1}^{N}u_{kj}^2\left(x_k-\bar{v}_j\right)\left(x_k-\bar{v}_j\right)^T;\ A=P^{-1}$$

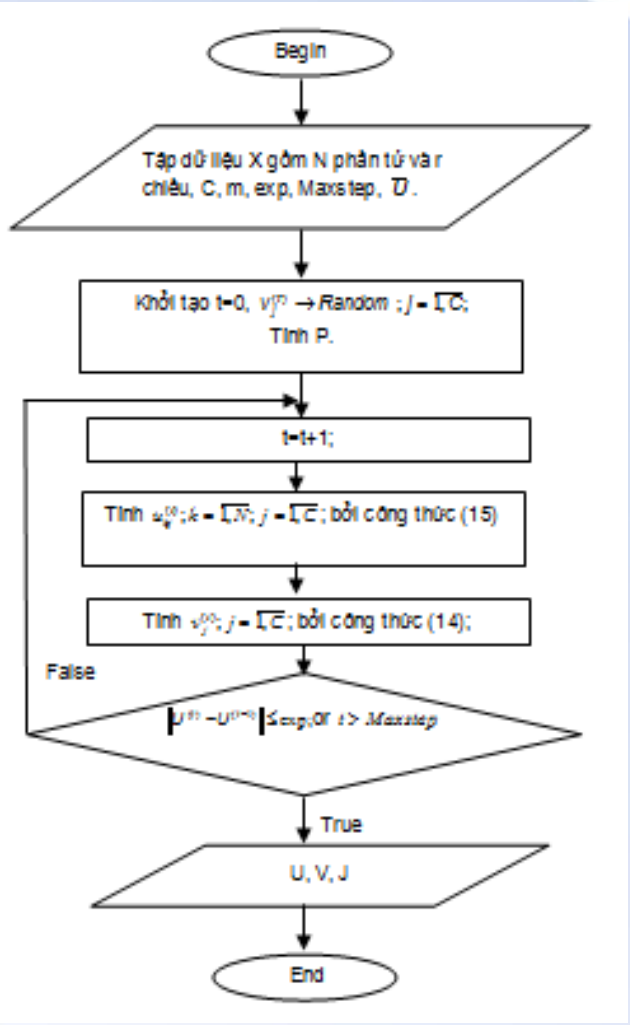
$$d_A^2(x_1,x_2)=\left(x_1-x_2\right)^T A\left(x_1-x_2\right)$$

- **Tính tâm cụm**

$$V_j=\frac{\sum_{k=1}^{N}u_{kj}X_k}{\sum_{k=1}^{N}u_{kj}}\ ;j=\overline{1,C} \qquad (14)$$

- **Tính hàm mức độ thành viên**

$$u_{kj}=\overline{u_{kj}}+\frac{e^{-\lambda\left\|X_k-V_j\right\|_A^2}}{\sum_{i=1}^{C}e^{-\lambda\left\|X_k-V_i\right\|_A^2}}\left(1-\sum_{i=1}^{C}\overline{u_{ki}}\right) \qquad (15)$$



# Các thuật toán phân cụm

## ❖ Thuật toán Semi-Supervised Fuzzy C-Mean của Bouchachia và Pedrycz [3] (SSFCMBP)

- **Hàm mục tiêu**

$$J(U,V,\lambda)=\sum_{i=1}^{C}\sum_{k=1}^{N}u_{ik}^{2}d_{ik}^{2}+\alpha\sum_{i=1}^{C}\sum_{k=1}^{L}(u_{ik}-\overline{u_{ik}})^{2}d_{ik}^{2}-\lambda\sum_{i=1}^{C}(u_{ik}-1) \quad (16)$$

- **Tính M**

$$M=(m_{hi})_{H\times C} \qquad m_{hi}=\begin{cases}1; & i\in h\\ 0; & i\notin h\end{cases} \quad (17)$$

- **Tính hàm mức độ thành viên**

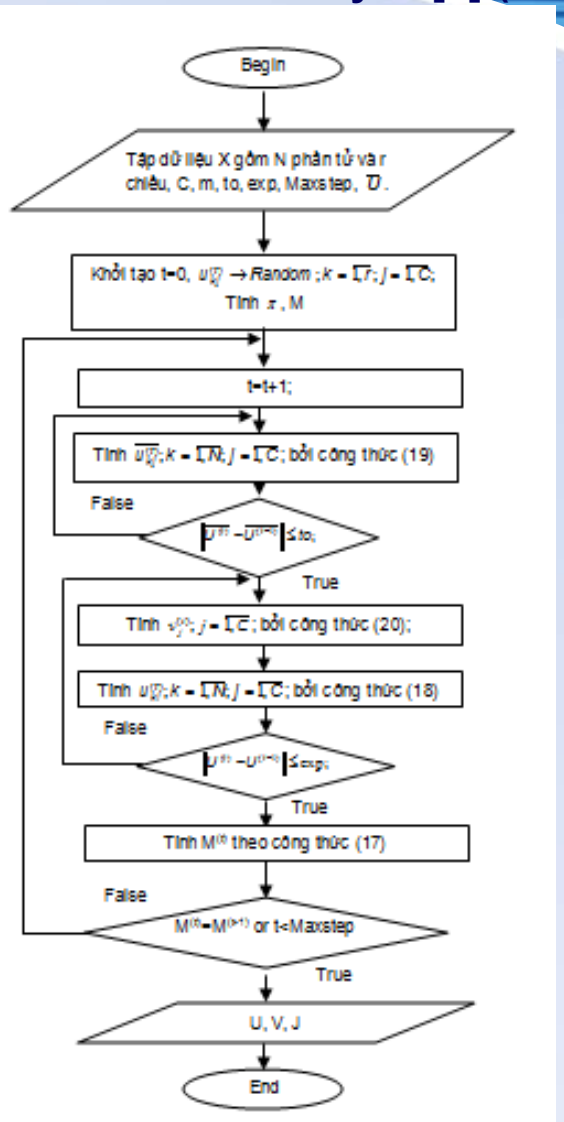
$$u_{ik}=\frac{\alpha\overline{u_{ik}}}{1+\alpha}+\frac{1-\frac{\alpha}{1+\alpha}\sum_{l=1}^{C}\overline{u_{ik}}}{\sum_{l=1}^{C}\frac{d_{ik}}{d_{lk}}} \qquad m_{hi}=1 \quad (18)$$

- **Thông tin bổ trợ**

$$\overline{u_{ik}}^{(t)}=\overline{u_{ik}}^{(t-1)}+2\beta\delta_k\sum_{h=1}^{H}\left(f_{hk}-\sum_{i\in\pi_h}\overline{u_{ik}^{(t-1)}}\right)*\begin{cases}1, & k\in\pi_h\\ 0, & k\notin\pi_h\end{cases} \quad (19)$$

- **Tính tâm cụm**

$$v_i=\frac{\sum_{j=1}^{N}\left(u_{ij}^{2}+\alpha(u_{ij}-\overline{u_{ik}})^{2}\right)x_j}{\sum_{j=1}^{N}\left(u_{ij}^{2}+\alpha(u_{ij}-\overline{u_{ik}})^{2}\right)} \quad (20)$$



# Trao đổi, câu hỏi?