Chương 5.
PHÂN CỤM

5.1. Giới thiệu

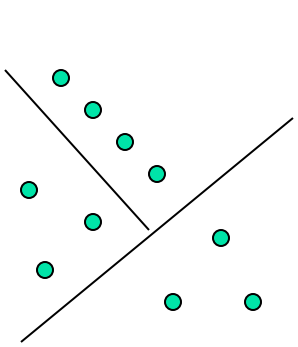
- Phân cụm là quá trình nhóm các đối tượng thành những cụm có ý nghĩa. Các dữ liệu trong cùng một cụm có nhiều tính chất chung và khác với dữ liệu trong các cụm khác

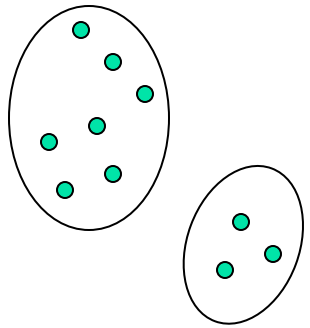
- Cho CSDL D={t_1,t_2,…,t_n} và số nguyên k, phân cụm là bài toán xác định ánh xạ f : D->{1,…,k} sao cho mỗi t_i được gán vào một cụm (lớp) Kj, 1<= j <= k .

- Không giống bài toán phân lớp, các nhóm không được biết trước.

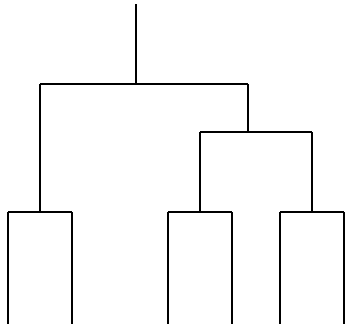
5.1. Giới thiệu

- **Cách biểu diễn các cụm**
  - **Phân chia bằng các đường ranh giới**
  - **Các khối cầu**
  - **Theo xác suất**
  - **Hình cây**
  - **…**





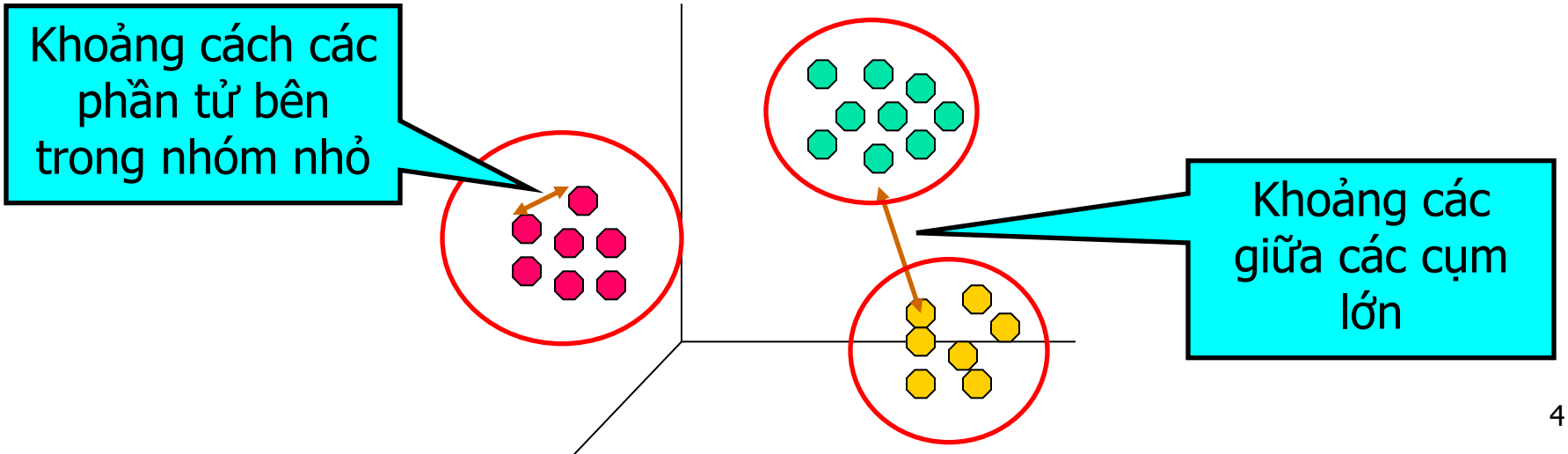
	1	2	3
I1	0.5	0.2	0.3
I2			
…			
In			



5.1. Giới thiệu

Tiêu chuẩn phân cụm

- Phương pháp phân cụm tốt là phương pháp sẽ tạo các cụm có chất lượng :
  - ***Sự giống nhau giữa đối tượng trong cùng một cụm cao.***
  - ***Giữa các cụm thì sự giống nhau thấp.***
  - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 giống Obj2 hơn so với sObj3.*



5.1. Giới thiệu

Tiêu chuẩn gom cụm

- Chất lượng của kết quả phân cụm dựa trên 2 yếu tố
  - ***Độ đo sự giống nhau dùng trong phương pháp phân cụm và sự thi hành nó.***
- Chất lượng của phương pháp phân cụm còn được đo bằng khả năng phát hiện một số hay tất cả các mẫu bị ẩn, bị dấu.

5.1. Giới thiệu

Ứng dụng của phân cụm

- Nhận dạng
- Phân tích dữ liệu không gian
- Xử lý ảnh
- www
  - Phân cụm tài liệu liên quan để dễ tìm kiếm
  - Phân dữ liệu Weblog thành cụm để tìm các cụm có cùng kiểu truy cập
- Giảm kích thước dữ liệu lớn

5.1. Giới thiệu

Ứng dụng của phân cụm

- Nhóm gen và protein có cùng chức năng
- Nhóm các cổ phiếu có xu hướng giá dao động giống nhau
- **Tiếp thị:** khám phá các nhóm khách hàng phân biệt trong CSDL mua hàng để xây dựng chương trình tiếp thị mục tiêu
- **Sử dụng đất:** nhận dạng các vùng đất sử dụng giống nhau khi khảo sát CSDL quả đất
- **Bảo hiểm:** nhận dạng các nhóm công ty có chính sách bảo hiểm mô tô với chi phí đền bù trung bình cao
- **Hoạch định thành phố:** nhận dạng các nhóm nhà cửa theo loại nhà, giá trị và vị trí địa lý.
- **Dự báo động đất**: dựa trên các kết quả phân cụm các vết đứt gãy của địa tầng
- …

5.2. Các phương pháp phân cụm

Các phương pháp phổ biến

- Phân cụm phân vùng
  - Xây dựng từng bước phân hoạch các cụm và đánh giá chúng theo các tiêu chí tương ứng
  - Độ đo tương tự / khoảng cách
  - K-mean, k-mediod
  - CLARANS, …
- Phân cụm phân cấp
  - Xây dựng hợp (tách) dần các cụm tạo cấu trúc phân cấp và đánh giá theo các tiêu chí tương ứng
  - Độ đo tương tự / khoảng cách
  - HAC: Hierarchical agglomerative clustering
  - CHAMELEON, BIRRCH và CURE, …

5.2.Các phương pháp phân cụm

- Phân cụm dựa theo mật độ
  - Hàm mật độ: Tìm các phần tử chính tại nơi có mật độ cao
  - Hàm liên kết: Xác định cụm là lân cận phần tử chính
  - DBSCAN, OPTICS…
- Phân cụm dựa theo lưới
  - Sử dụng lưới các ô cùng cỡ
  - Tạo phân cấp ô lưới theo một số tiêu chí: số lượng đối tượng trong ô
  - STING, CLIQUE, WaweCluster…
- Phân cụm dựa theo mô hình
  - Sử dụng một số mô hình giả thiết được phân cụm
  - Xác định mô hình tốt nhất phù hợp với dữ liệu
  - MCLUST…

5.3.Độ đo khoảng cách

- **Độ đo khoảng cách thường dùng để xác định sự khác nhau hay giống nhau giữa hai đối tượng.**
- **Khoảng cách Minkowski :**

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

với $i=(x_{i1}, x_{i2}, ..., x_{ip})$ *và* $j=(x_{j1}, x_{j2}, ..., x_{jp})$*:*
hai đối tượng p-chiều và q là số nguyên dương

- **Nếu q=1, d là khoảng cách Manhattan :**

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

5.3. Độ đo khoảng cách

- **Nếu *q*=2, *d* là khoảng cách Euclid :**

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

- **Tính chất của độ đo khoảng cách**
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$

5.4.Thuật toán K-mean

Cho số k, mỗi nhóm được biểu diễn bằng giá trị trung bình của DL trong nhóm

B1: Chọn ngẫu nhiên K đối tượng làm tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

B2 : Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách **Euclid**)

B3. Nhóm các đối tượng vào nhóm gần nhất

B4 : Tính lại giá trị trung tâm của từng nhóm

Cho nhóm $K_i=\{t_i1,t_i2,\ldots,t_im\}$, giá trị trung bình của nhóm là: $m_i= (1/m)(t_i1+ \ldots + t_im)$

Di chuyển trung tâm nhóm về giá trị TB mới của nhóm.

B5 : Nếu các trung tâm nhóm không có gì thay đổi thì dừng, ngược lại quay lại B2.

Ví dụ 1

Cho dữ liệu 1 chiều X sau và k = 2 :

X={2,4,10,12,3,20,30,11,25}

Gán ngẫu nhiên tâm cho 2 nhóm : $m_1=3$, $m_2=4$

Tính khoảng cách từ tâm $m_1$cho đến các phần tử

$d(x_1,m_1)=|3-2|=1$ $d(x_1,m_2)=|4-2|=2$

$d(x_2,m_1)=|3-4|=1$ $d(x_2,m_2)=|4-4|=0$

$d(x_3,m_1)=|3-10|=7$ $d(x_3,m_2)=|4-10|=6$

$d(x_4,m_1)=|3-12|=9$ $d(x_4,m_2)=|4-12|=8$

……..

Ta thấy $d(x_1,m_1)<d(x_1,m_2)$ nên phân x1 vào k_1,

$d(x_2,m_1)>d(x_2,m_2)$ nên phân x2 vào k_2 …

Ta được: $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$,

Ví dụ 1

Tính lại trọng tâm

$m_1=(2+3)/2=2.5$, $m_2=16$

$K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$,

$m_1=3$, $m_2=18$

$K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$,

$m_1=4.75$, $m_2=19.6$

$K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$,

$m_1=7$, $m_2=25$

Dừng khi trung tâm cụm không thay đổi

Thực hiện phân cụm với

$m_1=5$, $m_2=10$

Ví dụ 2

- Giả sử ta có 4 loại thuốc A,B,C,D, mỗi loại thuốc được biểu diễn bởi 2 đặc trưng X và Y như sau.
- A(1,1); B(2,1); C(4,3); D(5,4)
- Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng.
- Giải:
- **Bước 1.** Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất c1(1,1)) và B là tâm của nhóm thứ 2 (tọa độ tâm nhóm thứ hai c2 (2,1)).

Ví dụ 2

- **Bước 2.** Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean)
- Tính khoảng cách từ c1, c2 đến C(4,3) như sau

$$\mathbf{c}_1 = (1,1) \quad \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$\mathbf{c}_2 = (2,1) \quad \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

- Tương tự tính cho các phần tử khác ta được kết quả như sau:

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Ví dụ 2

- **Bước 3.** Nhóm các đối tượng vào nhóm gần nhất
  - Ta có k1={A}; k2={B, C,D}
- **Bước 4.** Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, c1(1,1). Tâm nhóm 2 được tính như sau:

$$\mathbf{c}_2 = (\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (\frac{11}{3}, \frac{8}{3}).$$

Ví dụ 2

- **Bước 5.** Tính lại khoảng cách từ các đối tượng đến tâm mới

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

- **Bước 6.** Nhóm các đối tượng vào nhóm
- Ta có k1={A, B}; k2={C,D}

Ví dụ 2

- **Bước 7**. Tính lại tâm cho nhóm mới

$$\mathbf{c}_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\frac{1}{2}, 1) \qquad \mathbf{c}_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\frac{1}{2}, 3\frac{1}{2})$$

- **Bước 8.** Tính lại khoảng cách từ các đối tượng đến tâm mới

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) & group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

- **Bước 9.** Nhóm các đối tượng vào nhóm
- Ta có k1={A, B}; k2={C,D}
- Như vậy không có gì thay đổi trong các nhóm nên thuật toán dừng tại đây

Bài tập

1. Cho tập điểm

X1(1,3)

X2(1.5 , 3.2)

x3 (1.3 ,2.8)

X4(3, 1)

Dùng k-means để phân cụm với k = 2

2: Cho tập điểm

X1(4,1) ; X2(5,1) ; X3(5,2) ; X4(1,4) ;
X5(1,5) ; X6(2,4) ; X7(2,5)

Dùng K-Mean để phân cụm (K=2)

Ưu điểm của K-means

- Tương đối nhanh
  - Độ phức tạp của thuật toán là O(tkn)
    - n: số điểm trong không gian dữ liệu
    - k: số cụm cần phân hoạch
    - t: số lần lặp (t << n)
- K-Means phù hợp với các cụm có dạng hình cầu

Nhược điểm của K-means

- Không đảm bảo đạt được tối ưu toàn cục
  - kết quả đầu ra phụ thuộc vào việc chọn k điểm khởi đầu
- Cần phải xác định trước số cụm k
- Khó xác định số cụm thực sự mà không gian dữ liệu có thể có
- Khó phát hiện các loại cụm có hình dạng phức tạp và nhất là các dạng cụm không lồi
- Không thể xử lý nhiễu và biệt lệ
- Chỉ có thể áp dụng khi tính được trọng tâm