



Data Mining

KHAI PHÁ DỮ LIỆU

Bài 5 Clustering (gom cụm)

Mai Xuân Hùng

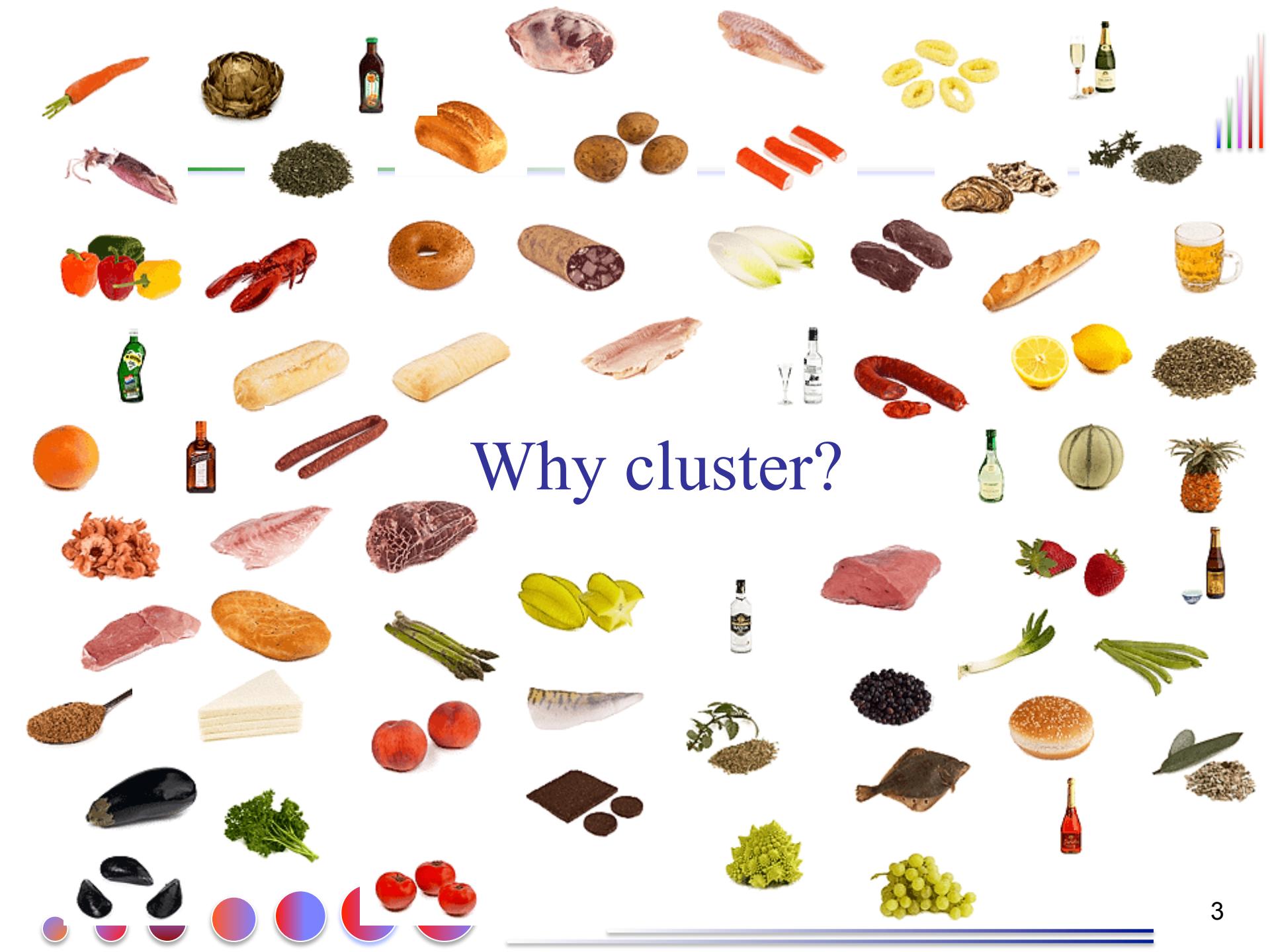


Phân tích bằng gom cụm là gì ?



- **Gom cụm:** gom các đối tượng dữ liệu
 - Tương tự với một đối tượng khác trong cùng cụm
 - Không tương tự với các đối tượng trong các cụm khác
- **Mục tiêu của gom cụm:** để gom tập các đối tượng thành các nhóm





Why cluster?



Các ứng dụng tiêu biểu của gom cụm

- Một công cụ độc lập để xem xét phân bố dữ liệu
- Làm bước tiền xử lý cho các thuật toán khác



Các ứng dụng của gom cụm



- **Tiếp thị:** khám phá các nhóm khách hàng phân biệt trong CSDL mua hàng
- **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.



Thế nào là gom cụm tốt



- Một phương pháp tốt sẽ tạo ra các cụm có chất lượng cao với:
 - Tương tự cao cho trong lớp (**intra-class**)
 - Tương tự thấp giữa các lớp (**inter-class**)
- Chất lượng của kết quả gom cụm phụ thuộc vào:
 - Độ đo tương tự sử dụng
 - Cài đặt độ đo tương tự
- Chất lượng của phương pháp gom cụm cũng được đo bởi khả năng phát hiện vài hay tất cả các mẫu bị che (**hidden patterns**)



Các yêu cầu của gom cụm trong KPD(1)

- Có thể thay đổi quy mô (scalability)
- Khả năng làm việc các loại thuộc tính khác nhau
- Khám phá các cụm có hình dáng bất kỳ
- Không nhạy cảm với thứ tự các bản ghi nhập vào
- Có số chiều cao
- Hợp tác với các ràng buộc do người dùng chỉ định
- Có thể diễn dịch và khả dụng



Tương tự và bất tương tự giữa hai đối tượng (1)



- Không có định nghĩa duy nhất về sự tương tự và bất tương tự giữa các đối tượng dữ liệu
- Định nghĩa về tương tự và bất tương tự giữa các đối tượng tùy thuộc vào

◦ Loại dữ liệu khảo sát

◦ Loại tương tự cần thiết



Sự tương tự và bất tương tự (2)



- Tương tự /Bất tương tự giữa đối tượng thường được biểu diễn qua độ đo khoảng cách $d(x,y)$
- Lý tưởng, mọi độ đo khoảng cách phải là một và phải thỏa các điều kiện sau:
 - $d(x,y) \geq 0$
 - $d(x,y) = 0$ if $x=y$
 - $d(x,y)=d(y,x)$
 - $d(x,z) \leq d(x,y) + d(y,z)$



Các biến trị khoảng (1)



- **Các độ đo liên tục của các thang đo tuyến tính, thô**
- Ví dụ: trọng lượng, chiều cao, tuổi
- Đơn vị đo có thể ảnh hưởng đến phân tích cụm
- Để tránh sự phụ thuộc vào đơn vị đo, cần chuẩn hóa dữ liệu



Các biến thang đo theo khoảng (2)

Chuẩn hoá các độ đo :

- Tính sai biệt tuyệt đối trung bình

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

với $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$, và

- Tính độ đo chuẩn (z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$



Các biến thang đo khoảng (3)

- Một nhóm các độ đo khoảng cách phổ biến cho biến tỉ lệ theo khoảng là khoảng cách **Minkowski**.

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

với $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ là các đối tượng dữ liệu p -chiều và q là số nguyên dương



Các biến thang đo khoảng (4)



- Nếu $q = 1$, độ đo khoảng cách là Manhattan (or city block)

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

- Nếu $q = 2$, độ đo khoảng cách là khoảng cách Euclidean

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



Thuật toán gom cụm k-mean



- Cho k là số cụm sau khi phân hoạch, với n là số điểm (đối tượng) trong không gian dữ liệu)
- Thuật toán k-means gồm 4 bước:



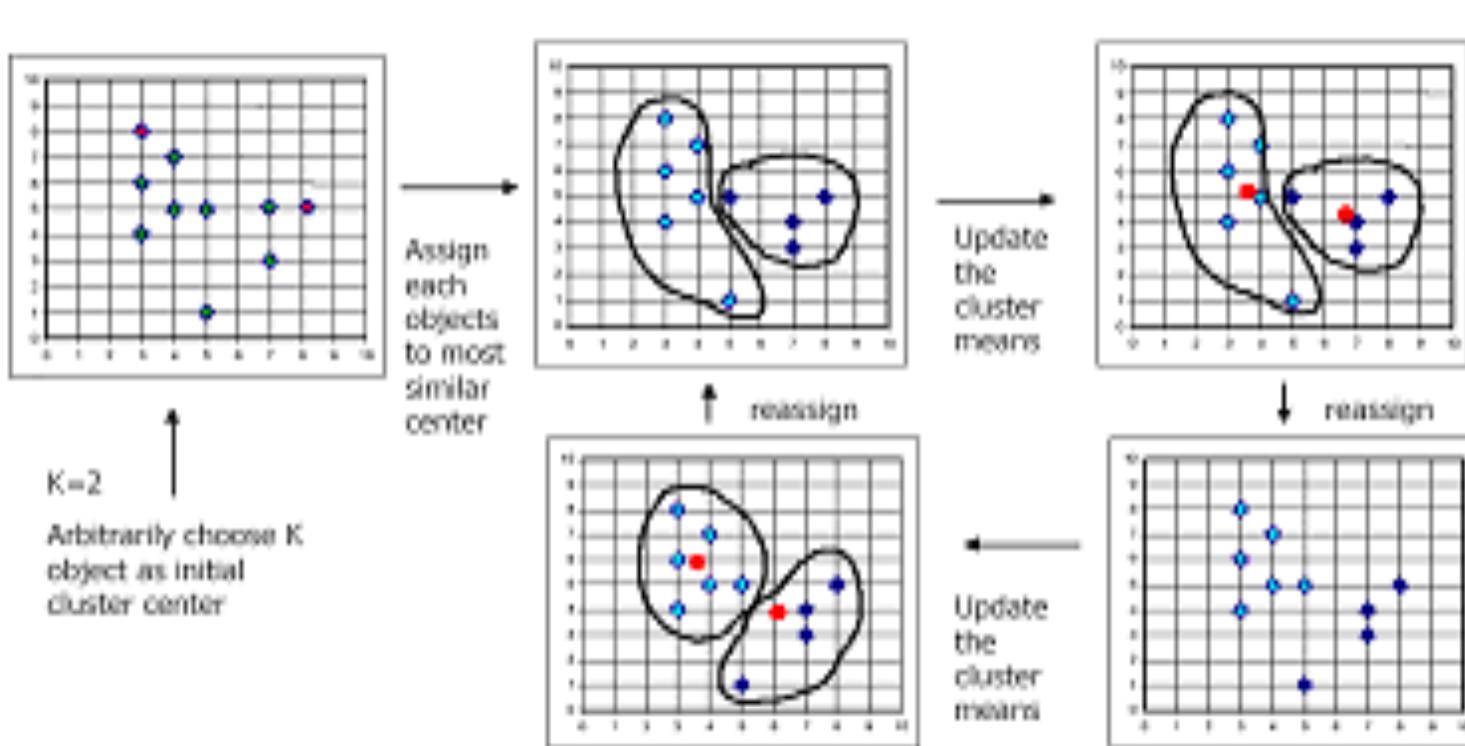
Các bước của thuật toán gom cụm



- Chọn ngẫu nhiên k điểm làm trọng tâm ban đầu của k cụm.
- Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần điểm đang xét nhất.
- Nếu không có phép gán lại nào thì dừng. Vì không có phép gán lại nào có nghĩa là các cụm đã ổn định và thuật toán không thể cải thiện làm giảm độ phân biệt hơn được nữa.
- Tính lại trọng tâm cho từng cụm.
- Quay lại bước 2.



Minh họa thuật toán với k= 2



Ví dụ



Cho tập điểm

$$x_1 = \{1, 3\} = \{x_{11}, x_{12}\}$$

$$x_2 = \{1.5, 3.2\} = \{x_{21}, x_{22}\}$$

$$x_3 = \{1.3, 2.8\} = \{x_{31}, x_{32}\}$$

$$x_4 = \{3, 1\} = \{x_{41}, x_{42}\}$$

Dùng k-means để gom cụm với $k = 2$



Ví dụ (tt)

- Dùng **k-means** để gom cụm với $k = 2$
- **Bước 1 :** Khởi tạo ma trận phân hoạch U có 4 cột ứng với 4 điểm và 2 dòng ứng với 2 cụm,
- **Bước 2:** $U=(m_{ij})$, $1 \leq i \leq 2$ và $1 \leq j \leq 4$
 - Cho $n=0$ (số lần lặp), tạo U_0

		x1	x2	x3	x4
U0=	c1	1	0	0	0
	c2	0	1	1	1



Ví dụ (tt)



• **Bước 3: Tính vector trọng tâm:**

Do có hai cụm C1,C2 nên có hai vector trọng tâm v1,v2

Các tính vector trọng tâm:

Với vector v1 cho cụm 1:

$$v_{11} = \frac{m_{11} * x_{11} + m_{12} * x_{21} + m_{13} * x_{31} + m_{14} * x_{41}}{m_{11} + m_{12} + m_{13} + m_{14}}$$

$$= \frac{1 * 1 + 0 * 1.5 + 0 * 1.3 + 0 * 3}{1 + 0 + 0 + 0} = 1$$



Ví dụ (tt)

$$v_{12} = \frac{m_{11} * x_{12} + m_{12} * x_{22} + m_{13} * x_{32} + m_{14} * x_{42}}{m_{11} + m_{12} + m_{13} + m_{14}}$$

$$= \frac{1 * 3 + 0 * 3.2 + 0 * 2.8 + 0 * 1}{1 + 0 + 0 + 0} = 3$$

- Vậy $v_1 = (1, 3)$



Ví dụ (tt)



- Với vector v2 cho cụm 2:

$$v2_1 = \frac{m21 * x11 + m22 * x21 + m23 * x31 + m24 * x41}{m21 + m22 + m23 + m24}$$
$$= \frac{0 * 1 + 1 * 1.5 + 1 * 1.3 + 1 * 3}{0 + 1 + 1 + 1} = \frac{5.8}{3} = 1.93$$

Ví dụ (tt)

$$v_{22} = \frac{m_{21} * x_{12} + m_{22} * x_{22} + m_{23} * x_{32} + m_{24} * x_{42}}{m_{21} + m_{22} + m_{23} + m_{24}}$$

$$= \frac{0 * 3 + 1 * 3.2 + 1 * 2.8 + 1 * 1}{0 + 1 + 1 + 1} = \frac{7}{3} = 2.33$$

- Vậy $v_2 = (1.93, 2.33)$



Gom các đối tượng vào cụm



- Tính khoảng cách Euclidean từ từng điểm đến cụm c1, c2 chọn cụm có khoảng cách gần nhất để đưa đối tượng vào cụm

$$d(x1, v1) = \sqrt{(x11 - v11)^2 + (x12 - v12)^2}$$

$$= \sqrt{(1 - 1)^2 + (3 - 3)^2} = 0$$



Ví dụ (tt)



$$d(x1, v2) = \sqrt{(x11 - v21)^2 + (x12 - v22)^2}$$

$$\sqrt{(1 - 1.93)^2 + (3 - 2.33)^2} = 1.14$$

- Gộp $x1$ vào cụm $c1$ vì $d(x1, v1) < d(x1, v2)$



Ví dụ (tt)



- Tính toán tương tự ta có:
 - $d(x_2, v_1) = 0.54 < d(x_2, v_2) = 0.97$ xếp x_2 vào cụm c_1
 - $d(x_3, v_1) = 0.36 < d(x_3, v_2) = 0.78$ xếp x_3 vào cụm c_1
 - $d(x_4, v_1) = 2.83 > d(x_4, v_2) = 1.70$ xếp x_4 vào cụm c_2



Ví dụ (tt)



- Tăng n lên 1
- Ma trận phân hoạch

Un sẽ là :

		x1	x2	x3	X4
U1=	c1	1	1	1	0
	c2	0	0	0	1

Lặp cho đến khi $| U_n - U_{n-1} | < \text{epsilon}$ thì dừng,
nếu sai thì quay về bước 3.



Tính lại trọng tâm



- V1(1.13,3)
- V2(3,1)
- $d(x_1, v_1) = 0.27, d(x_1, v_2) = 2.8 \rightarrow x_1 \in c_1$
- $d(x_2, v_1) = 0.52, d(x_2, v_2) = 2.67 \rightarrow x_2 \in c_1$
- $d(x_3, v_1) = 0.2, d(x_3, v_2) = 2.48 \rightarrow x_3 \in c_1$
- $d(x_4, v_1) = 2.64, d(x_4, v_2) = 0$



Ví dụ (tt)



Ma trận phân hoạch U2 là:

		x1	x2	x3	X4
U₁₂=	c₁	1	1	1	0
	c₂	0	0	0	1

$V_i |U_2 - U_1| = 0 \rightarrow$ thuật toán hội tụ \rightarrow dừng



Tổng kết (1)



- Phân tích gom cụm các đôi tượng dựa trên sự tương tự
- Phân tích gom cụm có phạm vi ứng dụng to lớn
- Có thể tính độ đo tương tự cho nhiều loại dữ liệu khác nhau
- Việc lựa chọn độ đo tương tự tùy thuộc vào dữ liệu được dùng và loại tương tự cần tìm



Bài tập



- Cho tập điểm
 - $x_1=\{0.7, 0.45\}$, $x_2=\{2.8, 1\}$, $x_3 =\{2.6, 1\}$, $x_4=\{1, 0.8\}$, $x_5=\{2.5, 1.2\}$, $x_6=\{1.3, 1.4\}$ $x_7=\{0.4, 0.7\}$,
 $x_8=\{1.7, 1.8\}$, $x_9=\{2, 2\}$
- Dùng k-means để gom cụm với $k = 3$

