Bài 5: Phân tích gom cụm

Cluster Analysis

PGS. TS. Đỗ Phúc
Khoa Hệ thống thông tin
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Phân tích bằng gom cụm

Khái quát

- **Phân tích bằng gom cụm là gì ?**
- **Đối tượng tương tự và không tương tự**
- **Các loại dữ liệu trong phân tích bằng gom cụm**
- **Các phương pháp gom cụm chính**
- **Các phương pháp phân hoạch**
- **Các phương pháp phân cấp**
- **Tổng kết**

Phân tích bằng gom cụm là gì ?

- **Gom cụm:** gom các đối tượng dữ liệu
  - Các đối tượng có độ tương đồng cao sẽ nằm trong cùng cụm
  - Các đối tượng có độ tương đồng thấp sẽ nằm trong các cụm khác nhau.
- **Mục tiêu của gom cụm:** để gom tập các đối tượng thành các nhóm có độ tương đồng cao.

Các ứng dụng tiêu biểu của gom cụm

- Công cụ để xem xét phân bố dữ liệu. Dữ liệu tập trung ở các nơi nào
- Bước tiền xử lý cho các giải thuật. Ví dụ: ta có 1000 điểm dữ liệu. Sau khi gom cụm, ta phát hiện được 30 cụm. Nếu mỗi cụm có 1 phần tử đại diện thì ta chỉ xét 30 phần tử đại diện thay vì phải xét 1000 điểm dữ liệu.

Các ứng dụng của gom cụm

- **Tiếp thị:** khám phá các nhóm khách hàng phân biệt trong CSDL mua hàng. Từ đó có chiến lược tiếp thị riêng.
- **Sử dụng đất:** nhận dạng các vùng đất sử dụng giống nhau khi khảo sát CSDL quả đất
- **Bảo hiểm,kinh doanh:** nhận dạng các nhóm công ty có đặc tính giống nhau. Ví dụ cổ phiếu, báo cáo tài chính giống nhau.
- **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

- Phương pháp gom cụm tốt sẽ tạo ra các cụm có chất lượng cao với:
  - Độ tương tự cao cho các đối tượng cùng lớp (**intra-class)**
  - Độ tương tự thấp cho các đối tượng khác lớp (**inter-class)**
- Chất lượng của kết quả gom cụm phụ thuộc vào:
  - Độ đo tương tự sử dụng

Các yêu cầu của gom cụm trong KPDL (1)

- Có thể xử lý khối dữ liệu lớn (scalability)
- Khả năng làm việc các loại thuộc tính khác nhau.
- Khám phá các cụm có hình dáng bất kỳ
- Các yêu cầu cung cấp tri thức lĩnh vực nhằm xác định các tham biến nhập

Các yêu cầu về gom cụm trong KPDL (2)

- Khả năng làm việc với nhiễu và phần tử dị biệt (outliers)
- Không nhạy cảm với thư tự các bản ghi nhập vào giải thuật.
- Có số chiều (số thuộc tính) cao, ví dụ gom cụm văn bản
- Phối hợp với các ràng buộc do người dùng chỉ định
- Có thể giải thích kết quả 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 tượng tự giữa các đối tượng tùy thuộc vào
  - Loại dữ liệu khảo sát
  - Loại tương tự cần thiết

Loại dữ liệu trong phân tích cụm

Ma trận khoảng cách Data matrix and Dissimilarity matrix

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

n đối tượng, p biến

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

d(i, j) khoảng cách)
d(i, j) = d(j, i)

Sự tương tự và bất tương tự (2)

- Tương tự /Bất tượng tự giữa đối tượng thường được biểu diễn qua độ đo khoảng cách *d(x,y)*
- Lý tưởng, mọi độ đo khoảng cách phải thỏa các điều kiện sau:
  1. $d(x, y) \geq 0$
  2. $d(x, y) = 0 \text{ iff } x = y$
  3. $d(x, y) = d(y, x)$
  4. $d(x, z) \leq d(x, y) + d(y, z)$

Loại dữ liệu trong phân tích cụm

- Các biến khoảng tỉ lệ
- Biến nhị phân
- Các biến định danh, thứ tự, tỉ lệ
- Các biến có kiểu hổn hợp
- Các kiểu dữ liệu phức tạp

Các biến trị khoảng (1)

- **Các đại lượng liên tục**
- Ví dụ: trọng lượng, chiều cao, tuổi
- Đơn vị đo (thứ nguyên) có thể ảnh hưởng đến phân tích cụm
- Để tránh sự phụ thuộc vào đơn vi đo (kg, mét, năm), cần chuẩn hoá dữ liệu

Chuẩn hóa theo miền trị theo thứ nguyên (2)

Chuẩn hoá các đơn vị đo :

- Tính sai biệt tuyệt đối trung bình

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

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

- Tính độ đo chuẩn (*z-score*)

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

Độ đo được dùng trên các thuộc tính số (3)

- Độ đo theo khoảng phổ biến là 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 $j = (x_{j1}, x_{j2}, ..., x_{jp})$ là các đối tượng dữ liệu p-chiều và q là số nguyên dương

Độ đo khoảng cách (4)

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

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

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

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

Các biến nhị phân (1)

- Biến nhị phân chỉ có hai trạng thái là 0 hay 1, 0 có nghĩa là có biến đó và 1 là không có,
- Bảng **contingency table** cho dữ liệu nhị phân:

Trong đó a, b,c ,d là số vị trị có giá trị ở object i và object j như trong bảng.

		Object *j*		
		1	0	*sum*
	1	*a*	*b*	*a+b*
Object *i*	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

Các biến nhị phân (2)

- **Hệ số đối sánh đơn giản** (tương tự bất biến, nếu biến nhị phân là đối xứng-hai trạng thái là tương đương):

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- **Hệ số Jaccard** (tương tự không bất biến, nếu biến nhị phân là bất đối xứng-có thiên vị một trạng thái):

$$d(i,j) = \frac{b+c}{a+b+c}$$

Các biến nhị phân (3)

Ví dụ: sự bất tương tự giữa các biến nhị phân: Bảng record bệnh nhân

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Tám thuộc tính trong đó
  - gender là thuộc tính đối xứng
  - Các thuộc tính còn lại là bất đối xứng nhị phân

Các biến nhị phân (4)

- Gọi các trị **Y** và **P** được gán trị 1, và trị **N** được gán trị 0
- Tính khoảng cách giữa các bệnh nhân dựa vào các bất đối xứng dùng hệ số Jaccard:

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Các biến định danh (nominal variables)

- Mở rộng biến nhị phân để biến có thể nhận nhiều hơn hai trạng thái chẳng hạn đỏ, vàng, xanh, lục
- Phương pháp 1: đối sánh đơn giản
  - **m:** số lần trùng khớp, p: tổng số biến

$$d(i, j) = \frac{p - m}{p}$$

- Phương pháp 2: **dùng một số lượng lớn các biến nhị phân**
  - Tạo biến nhị phân mới cho từng trang thái định danh

Các biến thứ tự

- Các biến thứ tự có thể là liên tục hay rời rạc
- Thứ tự của các trị là quan trọng, ví dụ hạng
- Có thể xử lý như tỷ lệ khoảng
  - Thay thế x_{if} bởi hạng của chúng
  - Ánh xạ phạm vi của từng biến vào đoạn [0, 1] bằng cách thay thế đối tượng thứ *i* trong biến thứ *f bởi*

$$r_{if} \in \{1, \ldots, M_f\}$$

  - Tính sự khác nhau dùng các phương pháp cho biến tỉ lệ theo khoảng

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Các biến tỉ lệ tỉ số

- **Độ đo dương trên thang phi tuyến**, xấp xỉ thang đo mũ
  - Ví dụ $\mathbf{Ae^{Bt}}$ hay $\mathbf{Ae^{-Bt}}$
- **Các phương pháp:**
  - xử lý chúng như các biến thang đo khoảng — không phải là lựa chọn tốt *! (why?)*
  - áp dụng biến đổi logarithmic **yif = log(xif)**
  - xử lý chúng như dữ liệu thứ tự liên tục và xử lý chúng theo hạng như thang đo khoảng

Các biến có kiểu hỗn hợp (1)

- CSDL có thể có 6 loại biến
- Có thể dùng công thức trọng để kết hợp chúng:

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

với

$\delta_{ij}^{(f)} = 0$ if x_{if} or x_{jf} is missing, or $x_{if} = x_{jf} = 0$;
otherwise $\delta_{ij}^{(f)} = 1$

Các phương pháp gom cụm (clustering) chính yếu

- **Phương pháp phân hoạch**
- **Phương pháp kiến trúc (hierarchical)**
- **Phương pháp dựa trên mật độ (Density-based methods)-> Seminar**

Các phương pháp phân hoạch

- **Phương pháp phân hoạch:** tạo phân hoạch CSDL *D* chứa *n* đối tượng thành tập có *k* cụm sao cho:
  - Mỗi cụm chức ít nhất một đối tượng
  - Mỗi đối tượng thuộc về một cụm duy nhất
- Cho trước *k*, tìm phân hoạch có ***k* cụm** nhằm tối ưu tiêu chuẩn chọn để phân nhóm.

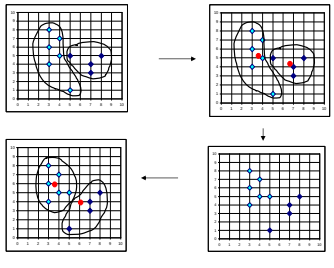
Tiêu chuẩn suy đoán chất lượng phân hoạch

- **Tối ưu toàn cục**: liệt kê vét cạn tất cả các phân hoạch
- **Phương pháp heuristic**:
  - **k-means** (MacQueen'67): mỗi cụm được biểu diễn bằng tâm của cụm(**centroid**)
  - **k-medoids** (Kaufman & Rousseeuw'87): mỗi cụm được biểu diễn bằng một trong các đối tượng của cụm (**medoid**)

Phương pháp gom cụm k-means(1)

- **Đầu vào thuật toán**: số cụm *k* và CSDL có *n* đối tượng.
- **Thuật toán gồm 4 bước**:
  1. Phân hoạch đối tượng thành *k* tập con/cụm khác rỗng.
  2. Tìm điểm hạt giống (seed) được xem là **centroid** (tính trị trung bình của các đối tượng trong cụm) cho từng cụm trong phân hoạch hiện hành
  3. Gán từng đối tượng vào cụm có centroid gần với đối tượng nhất
  4. Quay về bước 2, dùng khi không còn gán nào mới cả.

Ví dụ: phương pháp gom cụm k-means



Điểm mạnh của phương pháp gom cụm k-means

- **Xử lý khối lượng dữ liệu lớn**
- **Xử lý nhanh**: $O(tkn)$, với n là số đối tượng , k là số cụm và t là số lần lặp. Thông thường, $k, t << n$.
- Thuật toán thường kết thúc ở điểm tối ưu cục bộ; có thể dùng thuật toán GA để tìm tối ưu toàn cục

Điểm yếu của phương pháp gom cụm k-means

- Chỉ có thể áp dụng khi định nghĩa được trị trung bình của đối tượng
- Cần chỉ định trước *k (số cụm cần gom)*
- Không thể xử lý nhiễu và phần tử dị biệt
- Không thích hợp để tìm các cụm có dạng lồi hay các cụm có kích thước khác nhau.

Phương pháp phân cấp

- **Phương pháp phân cấp (hierarchical method):** tạo cây phân cấp các cụm
- Không cần chỉ định số k.
- Dùng ma trận khoảng cách làm tiểu chuẩn gom cụm
- Điều kiện dừng (ví dụ số cụm)

Cây phân cấp

- Cây phân cấp **dendrogram**
  - Nút lá biểu diễn từng đối tượng
  - Các nút bên trong biểu diễn các cụm

Hai loại phương pháp tạo kiến trúc cụm (1)

Two main types of hierarchical clustering techniques: agglomerative , divisive

Từ dưới lên

agglomerative (bottom-up):

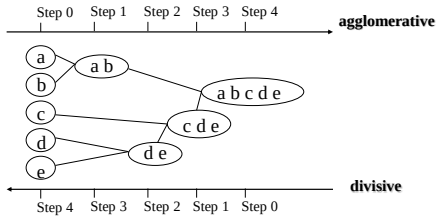
- Gán từng đối tượng vào cụm của nó (singleton-cụm chứa 1 đối tượng)
- Trộn theo từng bước hai cụm có độ tương tự cao nhất cho đến khi chỉ còn một cụm hay thỏa điều kiện kết thúc

Từ trên xuống

divisive (top-down):

- Bắt đầu bằng một cụm lớn chứa tất cả đối tượng
- Chia cụm phân biệt nhất thành các cụm nhỏ hơn và tiếp tục cho đến khi có n cụm hay thỏa điều kiện kết thúc

Hai loại phương pháp tạo kiến trúc phân cấp cụm (2)



Ví dụ về gom cụm phân cấp

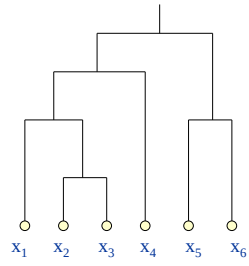
$\{x_1, x_2, x_3, x_4, x_5, x_6\}$

$\{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}\}$

$\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6\}\}$

$\{\{x_1\}, \{x_2, x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$

$\{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$



Từ dưới lên (Agglomerative Hierarchical clustering)

Starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.

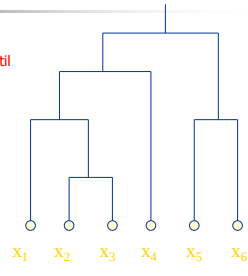
$\{x_1, x_2, x_3, x_4, x_5, x_6\}$

$\{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}\}$

$\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6\}\}$

$\{\{x_1\}, \{x_2, x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$

$\{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$



Từ trên xuống (Divisive Hierarchical Clustering)

Starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions.

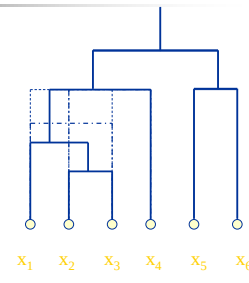
$\{x_1, x_2, x_3, x_4, x_5, x_6\}$

$\{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}\}$

$\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6\}\}$

$\{\{x_1\}, \{x_2, x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$

$\{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$



Tổng kết (1)

- Phân tích gom cụm các đối 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

Tổng kết (2)

- Có thể chia các thuật toán gom cụm thành các loại partitioning methods,
  - Các phương pháp phân cấp
  - Các phương pháp dựa trên mật độ,
  - Các phương pháp dựa trên lưới
  - Các phương pháp dựa trên mô hình
- Có nhiều vấn đề nghiên cứu về phân tích gom cụm