

Phân cụm dữ liệu

(Data Clustering)

Phan Xuân Hiếu

Khoa Công nghệ Thông tin
Trường DH Công nghệ (UET), ĐHQG Hà Nội (VNU)
hieupx@vnu.edu.vn

(last updated: 02–11–2015)

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K-means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Nội dung

1 Các khái niệm cơ bản

- **Vấn đề phân cụm và ứng dụng**
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Phân cụm là gì

Phân cụm dữ liệu là bài toán gom các đối tượng dữ liệu vào thành từng nhóm/cụm (group/cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó.

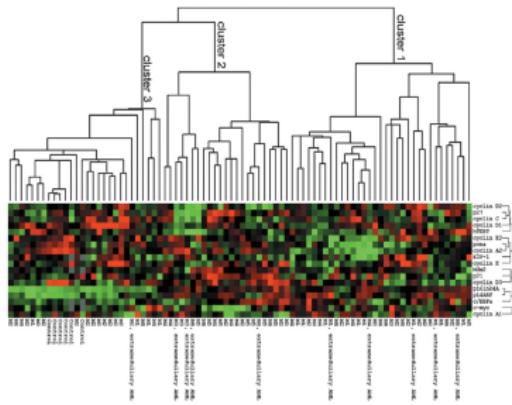
- Tên gọi
 - ▶ Phân cụm (clustering)
 - ▶ Học không giám sát (unsupervised learning)
 - ▶ Phân lớp không giám sát (unsupervised classification)
- Một số đặc điểm
 - ▶ Số cụm dữ liệu thường không được biết trước một cách chính xác
 - ▶ Nhiều cách tiếp cận để giải quyết, mỗi cách có vài kỹ thuật cụ thể
 - ▶ Các kỹ thuật khác nhau thường mang lại những kết quả khác nhau
- Các kiểu phân cụm
 - ▶ Phân cụm cứng (hard clustering)
 - ▶ Phân cụm mềm (soft clustering)

Ứng dụng của phân cụm

- Phân cụm tài liệu/văn bản.
- Phân cụm dữ liệu giao dịch mua bán.
- Phân cụm/nhóm người dùng.
- Phân đoạn/vùng ảnh (image segmentation).
- Phân cụm dữ liệu gen di truyền (gene expression data).
- Phân cụm tìm các cộng đồng (communities) trong mạng xã hội.
- ...



[nguồn: buffy.eecs.berkeley.edu]



[nguồn: www.nature.com]

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- **Điểm (points) và không gian (spaces)**
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

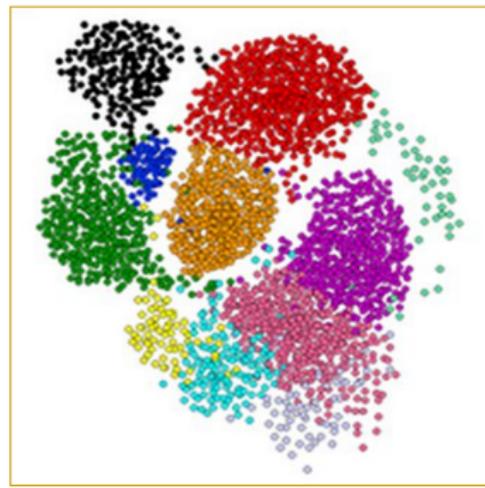
2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

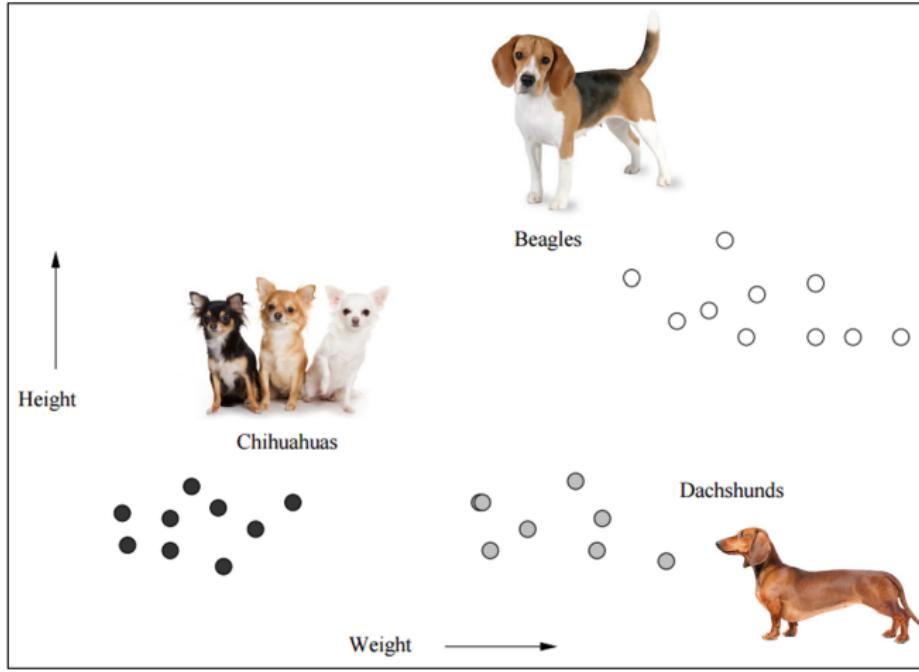
4 Tổng kết bài giảng

Điểm dữ liệu (data points) và không gian (spaces)



- Tập dữ liệu cần phân cụm $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ gồm n đối tượng.
- Mỗi đối tượng $\mathbf{x} \in \mathbf{D}$ là một điểm hay một véc tơ $\mathbf{x} = (x_1, x_2, \dots, x_m)$ trong không gian m chiều \mathbf{X} .

Ví dụ về điểm và cụm trong không gian hai chiều



Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- **Các độ đo khoảng cách (distances)**
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Các độ đo khoảng cách (distance measures)

- Các tính chất tiên đề của độ đo khoảng cách:
 - ▶ Tính không âm (non-negative): $d(\mathbf{x}, \mathbf{y}) \geq 0$ và $d(\mathbf{x}, \mathbf{y}) = 0$ khi và chỉ khi $\mathbf{x} = \mathbf{y}$.
 - ▶ Tính đối xứng (symmetric): $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
 - ▶ Tính tam giác (triangle inequality): $d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$.
- Các độ đo khoảng cách quan trọng (một số không thỏa mãn hết 3 tính chất tiên đề):
 - ▶ Khoảng cách Euclid (Euclidean distance)
 - ▶ Khoảng cách Manhattan (Manhattan distance)
 - ▶ Khoảng cách Cosine (Cosine distance)
 - ▶ Khoảng cách Hamming (Hamming distance)
 - ▶ Khoảng cách Jaccard (Jaccard distance)
 - ▶ Khoảng cách Kullback–Leibler (Kullback–Leibler divergence)

Độ đo Euclid chuẩn và độ đo Manhattan

- Cho hai điểm $\mathbf{x} = (x_1, x_2, \dots, x_m)$ và $\mathbf{y} = (y_1, y_2, \dots, y_m)$.
- Độ đo khoảng cách Euclid được xác định theo công thức:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^m |x_i - y_i|^r \right)^{\frac{1}{r}} \quad (1)$$

- Độ đo Euclid chuẩn ($r = 2$ tức L_2 -norm):

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (2)$$

- Độ đo khoảng cách Manhattan ($r = 1$, tức L_1 -norm):

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^m |x_i - y_i| \quad (3)$$

Độ đo Cosine

- Cho hai véc tơ $\mathbf{x} = (x_1, x_2, \dots, x_m)$ và $\mathbf{y} = (y_1, y_2, \dots, y_m)$.
- Độ đo cosine giữa hai véc tơ \mathbf{x} và \mathbf{y} được xác định:

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^m x_i y_i}{\sqrt{\sum_{i=1}^m x_i^2} \sqrt{\sum_{i=1}^m y_i^2}} \quad (4)$$

- Trong không gian dương (positive space):
 - Độ đo cosine thỏa mãn 3 tính chất tiên đề.
 - Giá trị nằm trong khoảng $[0, 1]$.

Độ đo Hamming

- Khoảng cách Hamming giữa hai véc tơ được xác định là số chiều mà ở đó các giá trị tương ứng của hai véc tơ là khác nhau.
- Thỏa mãn 3 tính chất tiên đề của độ đo.
- Thường được sử dụng khi các véc tơ ở dạng logic (true/false hay 0/1).
- Ví dụ, khoảng cách Hamming giữa hai véc tơ $(1, 0, 1, 0, 1)$ và $(1, 1, 1, 1, 0)$ là 3.

Độ đo Jaccard

- Cho x và y là hai tập hợp.
- Chỉ số Jaccard (Jaccard index), ký hiệu là $J(x, y)$, được xác định như sau:

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|} \quad (5)$$

- Khi đó, độ đo khoảng cách Jaccard được định nghĩa:

$$d(x, y) = 1 - J(x, y) \quad (6)$$

- Độ đo Jaccard thỏa mãn 3 tính chất tiên đề về độ đo.

Khoảng cách Kullback–Leibler (KL divergence)

- Còn được gọi là *information divergence* hoặc *relative entropy*.
- Cho $\mathbf{x} = (x_1, x_2, \dots, x_m)$ và $\mathbf{y} = (y_1, y_2, \dots, y_m)$ là hai phân phối xác suất rời rạc (discrete probabilistic distributions).
- Kullback–Leibler divergence giữa hai phân phối \mathbf{x} và \mathbf{y} được định nghĩa:

$$D_{KL}(\mathbf{x}||\mathbf{y}) = \sum_{i=1}^m x_i \log \frac{x_i}{y_i} \quad (7)$$

trong đó không xét đến những vị trí có $x_i = 0$ hoặc $y_i = 0$.

- Kullback–Leibler divergence không thỏa mãn tính chất đối xứng, tức $D_{KL}(\mathbf{x}||\mathbf{y})$ có thể khác $D_{KL}(\mathbf{y}||\mathbf{x})$.
- Do đó, có thể sử dụng độ đo dựa trên Kullback–Leibler divergence như sau:

$$d(\mathbf{x}, \mathbf{y}) = \frac{D_{KL}(\mathbf{x}||\mathbf{y}) + D_{KL}(\mathbf{y}||\mathbf{x})}{2} \quad (8)$$

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- **Các tiếp cận phân cụm (clustering approaches)**
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Các cách tiếp cận phân cụm (clustering approaches)

- **Phân cụm phân cấp (hierarchical methods):**
 - ▶ Còn gọi là phân cụm dựa vào kết nối (connectivity-based).
 - ▶ Phương pháp: top-down (divisive) hoặc bottom-up (agglomerative).
- **Phân cụm phân hoạch (partitioning/centroid-based methods):**
 - ▶ Khởi tạo với k cụm. Lặp đi lặp lại việc gán mỗi đối tượng vào cụm có tâm gần nhất, sau đó điều chỉnh lại các tâm cụm.
 - ▶ Phương pháp: K -means, K -medoids.
- **Phân cụm dựa trên phân bố (distribution-based methods):**
 - ▶ Giả định mỗi cụm dữ liệu được sinh ra từ một phân bố xác suất.
 - ▶ Phương pháp: mô hình trộn Gaussian với thuật toán EM.
- **Phân cụm dựa trên mật độ (density-based methods):**
 - ▶ Các cụm là các vùng có mật độ cao. Phương pháp DBSCAN, OPTICS.
- **Phân cụm dựa trên lưới (grid-based methods):**
 - ▶ Phân chia dữ liệu thành các ô lưới, phù hợp với dữ liệu không gian.
- **Phân cụm dựa trên đồ thị (graph-based methods):**
 - ▶ Các cụm là các vùng đồ thị con dày đặc (đồ thị dày đủ hoặc gần dày đủ – cliques/quasi-cliques). Phương pháp HCS.

Tính chất của các cách tiếp cận phân cụm

- **Phân cụm phân cấp (hierarchical methods):**
 - ▶ Phân cụm cứng (hard clustering); cây phân cụm (dendrogram); nhiều cách đo khoảng cách giữa các cụm; ...
- **Phân cụm phân hoạch (partitioning/centroid-based methods):**
 - ▶ Phân cụm cứng; cần xác định số cụm; dựa trên khoảng cách (distance-based); cụm dạng hình khối cầu (spherical shape); phù hợp với dữ liệu vừa và nhỏ; ...
- **Phân cụm dựa trên phân bố (distribution-based methods):**
 - ▶ Phân cụm mềm; cần xác định số cụm; cần giả định phân bố dữ liệu (distributional assumption); hình dạng cụm theo phân bố giả định; ...
- **Phân cụm dựa trên mật độ (density-based methods):**
 - ▶ Hình dạng cụm bất kỳ (arbitrary shapes); các cụm được phân biệt bởi các vùng mật độ thấp; có thể tìm được ngoại lai (outliers); ...
- **Phân cụm dựa trên lưới (grid-based methods):**
 - ▶ Nhiều kích thước ô lưới; thời gian tính toán nhanh; ...
- **Phân cụm dựa trên đồ thị (graph-based methods):**
 - ▶ Khai phá đồ thị con dày đặc (dense sub-graph mining); không cần giả định nào về tính chất các cụm; ...

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K-means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Phân cụm trong không gian số chiều (rất) lớn

(The Curse of Dimensionality)

- Trong không gian số chiều rất lớn, hầu hết các cặp điểm bất kỳ đều xa nhau. Xét công thức tính khoảng cách Euclid sau:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (9)$$

Giả sử hạn chế giá trị tại các chiều trong khoảng $[0, 1]$ (unit cube), thì cận dưới của $d(\mathbf{x}, \mathbf{y})$ là 1 và cận trên là \sqrt{m} . Hầu hết các cặp điểm có khoảng cách là trung bình là $\frac{1+\sqrt{m}}{2}$ với m rất lớn.

- Trong không gian số chiều rất lớn, hai vec tơ bất kỳ thường gần như trực giao. Xét công thức tính khoảng cách Cosine sau:

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^m x_i y_i}{\sqrt{\sum_{i=1}^m x_i^2} \sqrt{\sum_{i=1}^m y_i^2}} \quad (10)$$

Khi m rất lớn, tử số thường nhỏ và mẫu số luôn rất lớn.

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K-means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Phân cụm phân cấp (hierarchical clustering)

- Phân cụm phân cấp bottom-up trong không gian Euclid.
- Các cách thức (tiêu chí) chọn hai cụm để sát nhập.
- Hiệu quả của phân cụm phân cấp.
- Phân cụm phân cấp trong không gian khác Euclid.

Phân cụm phân cấp bottom-up (agglomerative)

Ý tưởng:

Khởi đầu mỗi điểm (đôi tượng) là một cụm riêng. Thuật toán sẽ tạo các cụm lớn hơn bằng cách sát nhập các cụm nhỏ hơn *gần nhau nhất* tại mỗi vòng lặp.

Cần xác định trước:

- Các cụm được biểu diễn như thế nào.
- Tiêu chí chọn hai cụm *gần nhau nhất* để sát nhập.
- Khi nào thuật toán ngừng việc sát nhập các cụm.

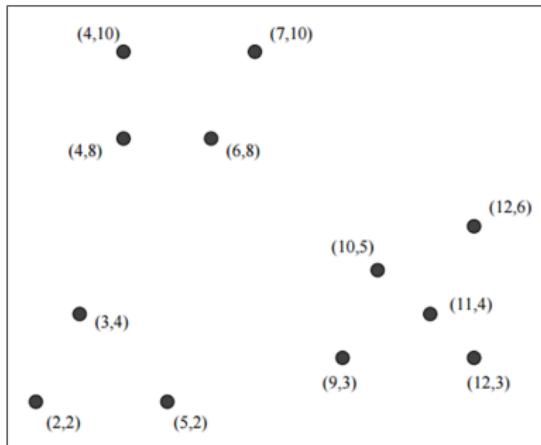
Thuật toán:

WHILE (chưa thỏa mãn điều kiện dừng) DO

- Chọn hai cụm thích hợp nhất để sát nhập;
- Sát nhập hai cụm đó thành một cụm lớn hơn;

END WHILE

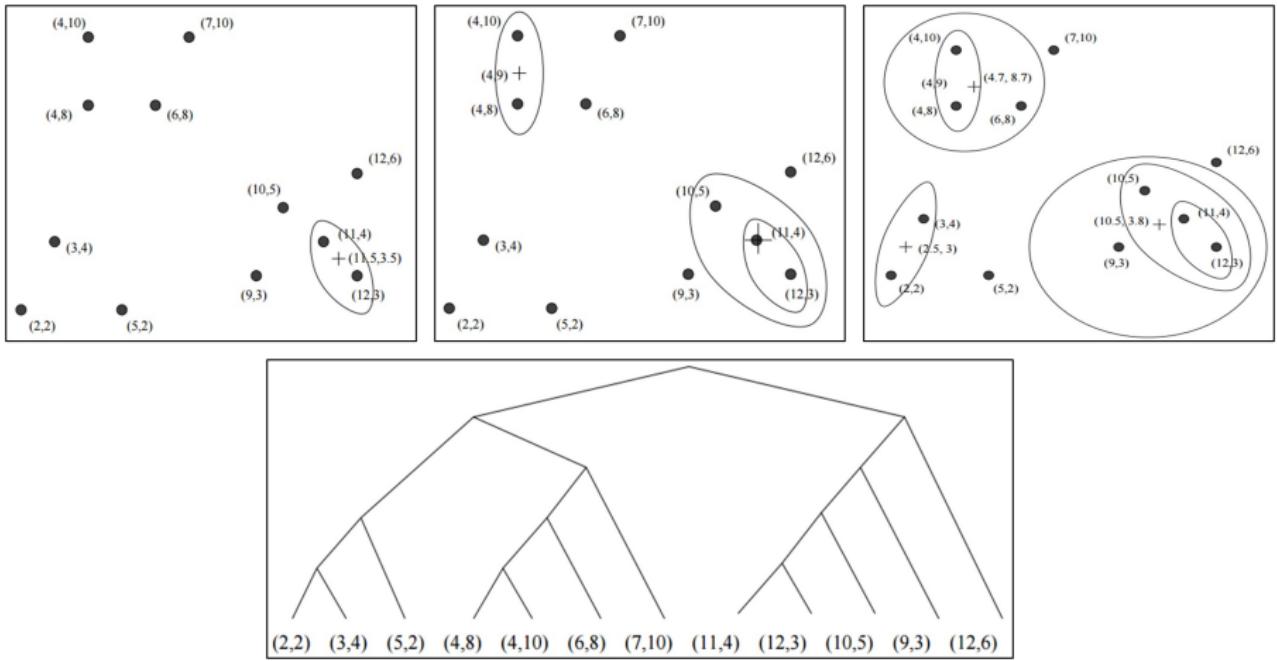
Ví dụ minh họa phân cụm phân cấp



- Tập dữ liệu gồm 12 đôi tượng tương ứng với 12 điểm trong không gian hai chiều.
- Khởi tạo ban đầu: mỗi đôi tượng là một cụm riêng rẽ.
- Các cặp điểm có khoảng cách Euclid nhỏ nhất ban đầu:
 $d((10, 5), (11, 4)) = d((11, 4), (12, 3)) = \sqrt{2}$.

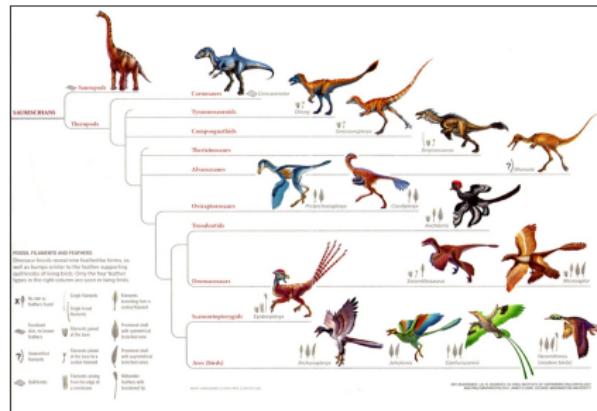
Ví dụ minh họa phân cụm phân cấp: sát nhập cụm

[(2, 2)]	[(2, 2)]	[(2, 2)]	[(2, 2)]	[(2.5, 3)]	[(2.5, 3)]	[(2.5, 3)]
[(3, 4)]	[(3, 4)]	[(3, 4)]	[(3, 4)]			
[(5, 2)]	[(5, 2)]	[(5, 2)]	[(5, 2)]	[(5, 2)]	[(5, 2)]	[(5, 2)]
[(4, 8)]	[(4, 8)]	[(4, 9)]	[(4, 9)]	[(4, 9)]	[(4.7, 8.7)]	[(4.7, 8.7)]
[(4, 10)]	[(4, 10)]					
[(6, 8)]	[(6, 8)]	[(6, 8)]	[(6, 8)]	[(6, 8)]		
[(7, 10)]	[(7, 10)]	[(7, 10)]	[(7, 10)]	[(7, 10)]	[(7, 10)]	[(7, 10)]
[(11, 4)]	[(11.5, 3.5)]	[(11.5, 3.5)]	[(11, 4)]	[(11, 4)]	[(11, 4)]	[(10.5, 3.8)]
[(12, 3)]						
[(10, 5)]	[(10, 5)]	[(10, 5)]				
[(9, 3)]	[(9, 3)]	[(9, 3)]	[(9, 3)]	[(9, 3)]	[(9, 3)]	
[(12, 6)]	[(12, 6)]	[(12, 6)]	[(12, 6)]	[(12, 6)]	[(12, 6)]	[(12, 6)]



Khi nào nên dừng việc sát nhập cụm

- Có hiểu biết hoặc phỏng đoán về số cụm trong tập dữ liệu.
 - Khi việc sáp nhập hai cụm nào đó tạo ra một cụm *kém chất lượng* (ví dụ: khi khoảng cách trung bình từ các điểm đến tâm cụm lớn hơn một ngưỡng nào đó).
 - Có thể dừng khi tạo ra cụm cuối cùng bao gồm tất cả các đối tượng. Kết quả là một cây phân cấp cụm (dendrogram). Có ý nghĩa trong một số trường hợp, ví dụ cây tiến hóa của các loài.



Các cách thức (tiêu chí) chọn hai cụm để sát nhập

- **Centroid-linkage:** khoảng cách giữa hai tâm của hai cụm. Sát nhập hai cụm nào có khoảng cách này nhỏ nhất.
- **Single-linkage:** khoảng cách giữa hai điểm gần nhau nhất thuộc hai cụm. Sát nhập hai cụm nào có khoảng cách này nhỏ nhất.
- **Average-linkage:** trung bình các khoảng cách giữa hai cặp điểm bất kỳ thuộc hai cụm. Sát nhập hai cụm nào có khoảng cách này nhỏ nhất.
- **Complete-linkage:** khoảng cách giữa hai điểm xa nhất của hai cụm. Sát nhập hai cụm nào có khoảng cách này nhỏ nhất.
- **Radius:** bán kính (radius) của một cụm là khoảng cách từ điểm xa nhất của cụm tới tâm cụm. Sát nhập hai cụm nào tạo nên một cụm có bán kính nhỏ nhất.
- **Diameter:** đường kính (diameter) của một cụm là khoảng cách của hai điểm xa nhất trong cụm. Sát nhập hai cụm nào tạo nên một cụm có đường kính bé nhất.

Hai kiểu phân cụm phân cấp

- **Agglomerative hierarchical clustering:** xuất phát mỗi điểm là một cụm, việc phân cụm là thực hiện sát nhập các cụm nhỏ thành cụm to hơn (bottom-up).
- **Divisive hierarchical clustering:** tất cả các đôi tượng/điểm là một cụm, việc phân cụm là thực hiện chia tách cụm lớn thành các cụm nhỏ hơn (top-down).

Hiệu quả của phân cụm phân cấp

- Độ phức tạp tính toán:
 - ▶ Bước đầu tiên: $O(n^2)$.
 - ▶ Các bước tiếp theo tỉ lệ với $(n - 1)^2, (n - 2)^2, \dots$
 - ▶ Tổng cộng: $O(n^3)$.
- Cải tiến:
 - ▶ Khởi đầu: tính toán khoảng cách giữa tất cả các cặp điểm cần $O(n^2)$.
 - ▶ Lưu tất cả các cặp điểm cùng khoảng cách của chúng trong một hàng đợi ưu tiên cần $O(n^2)$.
 - ▶ Khi sát nhập hai cụm, ví dụ C_i và C_j , loại bỏ tất cả những thành phần trong hàng đợi ưu tiên liên quan đến một trong hai cụm cần $O(n\log n)$.
 - ▶ Tính toán khoảng cách của tất cả các cụm và cụm vừa sinh ra rồi lưu trong hàng đợi ưu tiên cần $O(n\log n)$.
 - ▶ Tổng cộng cần $O(n^2\log n)$.

Phân cụm phân cấp trong không gian khác Euclid

- Không dựa vào tọa độ các điểm (đôi tượng).
- Áp dụng các độ đo khác như:
 - ▶ Jaccard.
 - ▶ Kullback–Leibler divergence.
 - ▶ String edit distance.
 - ▶ ...
- Không tính được tâm (centroid) của cụm từ các đôi tượng trong cụm. Sử dụng một trong những đôi tượng trong cụm làm đối tượng đại diện của cụm (clustroid). Đối tượng này thường gần với tất cả các đối tượng trong cụm. Một số cách chọn:
 - ▶ Tổng khoảng cách từ clustroid đến các đối tượng khác trong cụm là nhỏ nhất.
 - ▶ Khoảng cách từ clustroid đến điểm xa nhất trong cụm là nhỏ nhất.
 - ▶ Trung bình khoảng cách từ clustroid đến các đối tượng khác trong cụm là nhỏ nhất.

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- **Thuật toán K-means**
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Phân cụm với K -means

- Giới thiệu K -means
- Mục tiêu tối ưu
- Thuật toán
- Ví dụ minh họa K -means
- Khởi tạo tâm các cụm ban đầu
- Chọn số lượng cụm k phù hợp
- Ưu và nhược điểm của K -means

Giới thiệu K-means

Ý tưởng:

Khởi đầu với k tâm điểm của k cụm, lặp đi lặp lại việc gán các điểm vào các cụm có tâm cụm gần nhất và tính toán lại tâm cụm cho đến khi không hoặc ít có sự thay đổi cụm của các điểm hay sự dịch chuyển các tâm cụm.

- Một trong những thuật toán phân cụm đơn giản nhất, nhưng được biết đến và sử dụng nhiều nhất.
- Cần xác định trước số cụm (k cụm). Tuy nhiên, có thể suy diễn giá trị k phù hợp bằng phương pháp thử sai hoặc bằng các phương pháp điều chỉnh số cụm.
- Cần khởi tạo trước tâm cụm (centroid) cho k cụm này.

Mục tiêu tối ưu

- Tập dữ liệu cần phân cụm $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ gồm n đối tượng.
- Số lượng cụm k ($\leq n$).
- Gọi $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ là tập hợp k cụm tương ứng với một phân hoạch trên tập dữ liệu \mathbf{D} , khi đó tổng bình phương khoảng cách từ các đối tượng đến tâm mỗi cụm trên toàn bộ phân hoạch \mathbf{S} là

$$\sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \quad (11)$$

Trong đó $\boldsymbol{\mu}_i$ là tâm (centroid) của cụm S_i .

- Mục tiêu: tìm phân hoạch $\mathbf{S}^* = \{S_1^*, S_2^*, \dots, S_k^*\}$ tối ưu sao cho giá trị biểu thức trên đạt cực tiểu. Nói cách khác:

$$\mathbf{S}^* = \operatorname{argmin}_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \quad (12)$$

- Tìm phân hoạch tối ưu \mathbf{S}^* : NP-hard.
- Giải pháp: nghiệm xấp xỉ, tối ưu cục bộ.

Thuật toán K-means

- ① Khởi tạo $\mu_1, \mu_2, \dots, \mu_k$ là tâm của k cụm S_1, S_2, \dots, S_k .
- ② **Assignment step:** với mỗi đối tượng $\mathbf{x}_p \in \mathbf{D}$, gán \mathbf{x}_p cho cụm gần nhất S_i nào đó. Khoảng cách từ \mathbf{x}_p tới tâm cụm μ_i của S_i nhỏ hơn so với khoảng cách tới tâm của các cụm còn lại thể hiện qua công thức:

$$S_i^{(t)} = \{\mathbf{x}_p : \|\mathbf{x}_p - \mu_i^{(t)}\|^2 \leq \|\mathbf{x}_p - \mu_j^{(t)}\|^2 \forall j \neq i, 1 \leq j \leq k\} \quad (13)$$

Mỗi $\mathbf{x}_p \in \mathbf{D}$ được gán cho một cụm duy nhất. t là vòng lặp thứ t^{th}

- ③ **Update step:** Tính toán lại tâm của các cụm (cho vòng lặp tiếp theo):

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j, \quad 1 \leq i \leq k \quad (14)$$

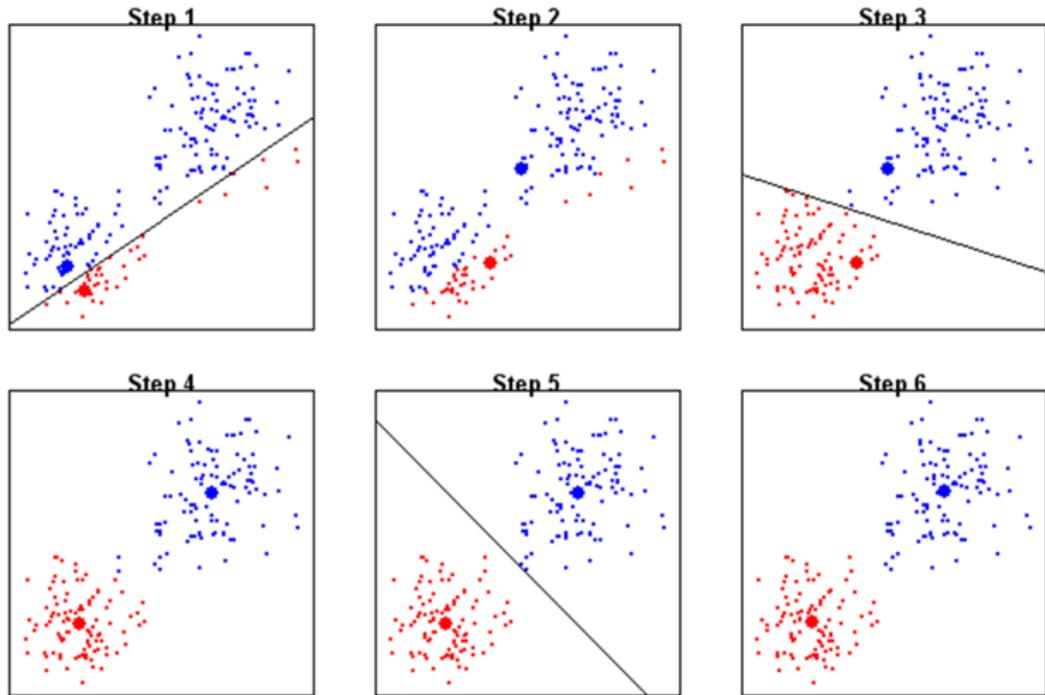
- ④ Nếu thỏa mãn *điều kiện dừng* (xem slide tiếp theo), thuật toán sẽ dừng. Ngược lại, thực hiện vòng lặp tiếp theo từ bước 2.

Điều kiện dừng của K-means

- Khi không có đối tượng x_p nào cần phải gán lại cụm, hoặc
- Tổng độ sai khác của các tâm cụm giữa hai vòng lặp kế tiếp $t - 1$ và t nhỏ hơn một ngưỡng $\epsilon (> 0)$ cho trước nào đó:

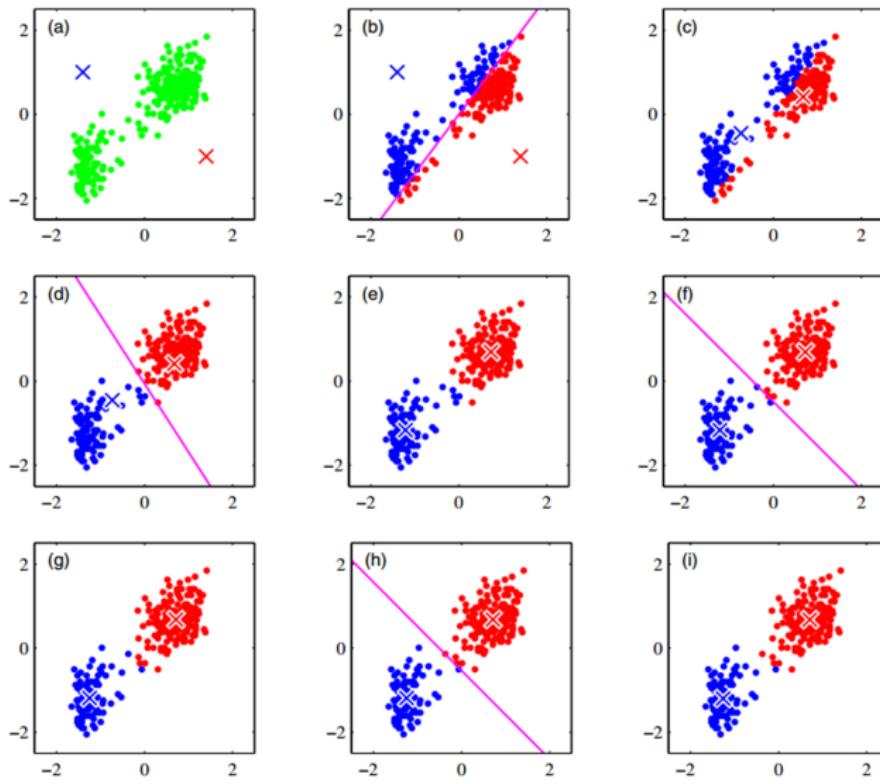
$$\sum_{i=1}^k \|\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_i^{(t-1)}\| \leq \epsilon \quad (15)$$

Ví dụ minh họa K-means



[nguồn: <http://sherrytowers.com/2013/10/24/k-means-clustering>]

Ví dụ minh họa K-means (2)



[nguồn: Pattern Recognition and Machine Learning by Christopher M. Bishop]

Khởi tạo tâm các cụm ban đầu

Mục đích

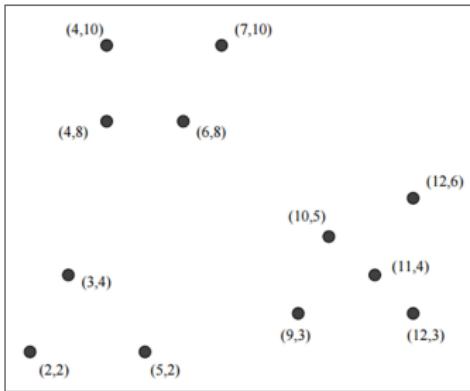
Lựa chọn các điểm làm tâm các cụm sao cho các tâm này nằm ở các cụm khác nhau thực sự trong dữ liệu. Có hai cách tiếp cận:

- Lựa chọn các điểm làm tâm các cụm từng đôi một càng xa nhau càng tốt.
- Thực hiện clustering trên một mẫu nhỏ (small sample) của tập D sử dụng một thuật toán clustering nào đó, ví dụ phân cụm phân cấp. Dừng ở mức k cụm, sau đó lấy mỗi điểm gần tâm mỗi cụm làm tâm cụm đầu vào cho thuật toán K -means.

Với cách tiếp cận đầu:

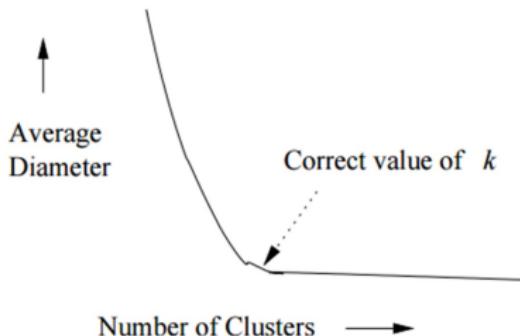
- Chọn tâm cụm đầu tiên là một điểm ngẫu nhiên.
- Với mỗi tâm cụm còn lại, chọn điểm có khoảng cách nhỏ nhất tới các tâm điểm đã chọn lớn nhất có thể.

Ví dụ minh họa việc khởi tạo các tâm cụm ban đầu



- Nếu tâm cụm số một được lựa chọn là $(6, 8)$, tâm cụm tiếp theo được lựa chọn là $(12, 3)$, trong số các điểm còn lại, điểm $(2, 2)$ được chọn làm tâm của cụm tiếp theo vì
$$\min\{d((2, 2), (6, 8)), d((2, 2), (12, 3))\} = \min\{\sqrt{52}, \sqrt{101}\} = \min\{7.21, 10.05\} = 7.21$$
 là lớn nhất so với các lựa chọn còn lại.
- Nếu $(10, 5)$ được chọn cho tâm của cụm đầu tiên thì các tâm cụm tiếp theo sẽ là $(2, 2)$ và $(4, 10)$.

Lựa chọn số lượng cụm k phù hợp



- Có nhiều cách thức để đo mức độ gắn kết (cohesion) các đối tượng trong một cụm. Một trong những cách đó là ước lượng giá trị trung bình của đường kính (average diameter) của các cụm.
- Khi số cụm k nhỏ hơn số cụm có thật trong dữ liệu, đường kính trung bình sẽ tăng cao.
- Có thể lựa chọn k bằng cách thực hiện k-means với các giá trị $k = 1, 2, 4, 8, \dots$ và tìm khoảng giá trị $[v, 2v]$ của k mà ở đó đường kính trung bình giảm rất ít.

Ưu điểm của K-means

- Dễ thi hành và hiệu quả trong nhiều trường hợp.
- Tường minh và dễ hiểu.
- Khá hiệu quả: độ phức tạp tính toán là $O(tknm)$, trong đó:
 - ▶ t là số vòng lặp của thuật toán.
 - ▶ k là số cụm.
 - ▶ n là số đối tượng dữ liệu cần phân cụm.
 - ▶ m là số chiều trong không gian biểu diễn các đối tượng.
- Mang lại kết quả tốt khi các cụm trong dữ liệu đứng khá riêng rẽ và có hình dạng giống khối cầu.

Nhược điểm của K -means

- Cần xác định trước số cụm k .
- Khởi tạo tâm cụm ngẫu nhiên ban đầu thường không đảm bảo kết quả tốt.
- Không áp dụng cho các thuộc tính dữ liệu dạng thể loại (categorical).
- Nếu có những vùng dữ liệu bị chồng chéo, K -means rất khó phân biệt các cụm một cách chính xác.
- Tối ưu tâm cụm chứ chưa tối ưu phần biên giữa các cụm.
- Nghiêm túc cục bộ.
- Không hiệu quả trong việc xử lý nhiễu và ngoại lai (outliers).
- Không mang lại kết quả tốt khi các cụm trong dữ liệu có hình dạng bất kỳ (thay vì dạng khói cầu) hoặc dữ liệu không tuyến tính (non-linear data).

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

4 Tổng kết bài giảng

Phân cụm dựa trên mật độ (density-based clustering)



[Nguồn: Data Mining: Concepts and Techniques by J. Han và cộng sự]

- Các phương pháp phân cụm phân cấp hay phân hoạch (k -means) chỉ cho kết quả tốt khi các cụm có dạng khói cầu (spherically-shaped clusters).
- Các phương pháp phân cụm dựa trên mật độ thường được áp dụng khi các cụm dữ liệu có hình dạng bất kỳ (arbitrarily shaped clusters).
- Hai thuật toán chính: DBSCAN và OPTICS.

Thuật toán DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)

Ý tưởng:

Mật độ của một đối tượng x được đo là số đối tượng láng giềng lân cận (neighborhood) của x . DBSCAN thực hiện phân cụm bằng cách chọn các đối tượng *cốt lõi* (core objects) có mật độ cao và sau đó kết nối các đối tượng đó với láng giềng của chúng để tạo lập các vùng có mật độ cao (dense regions) chính là các cụm dữ liệu.

Một số khái niệm:

- ϵ -neighborhood(x) là khối cầu có bán kính ϵ (> 0) với tâm tại đối tượng x .
- ϵ -density(x) đo mật độ của x là tổng số đối tượng nằm trong khối cầu ϵ -neighborhood(x).
- Một đối tượng x là đối tượng *cốt lõi* nếu ϵ -density(x) $\geq MinPts$

Đối tượng cốt lõi và vùng có mật độ cao (dense region)

- Các giá trị ngưỡng ϵ và $MinPts$ do người dùng định nghĩa.
- Các đối tượng cốt lõi chính là các điểm trụ (pillars) của các vùng mật độ cao hay đậm đặc (dense regions).
- Ký hiệu $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ là tập dữ liệu gồm n đối tượng cần phân cụm:
 - ▶ Có thể xác định được tập tất cả các điểm cốt lõi trong \mathbf{D} với các ngưỡng ϵ và $MinPts$.
 - ▶ Cho \mathbf{q} là một điểm cốt lõi trong \mathbf{D} , một đối tượng $\mathbf{p} \in \mathbf{D}$ được gọi là **có-thể-với-tới-trực-tiếp-theo-mậtđộ** (directly density-reachable) từ \mathbf{q} nếu \mathbf{p} nằm trong ϵ -neighborhood(\mathbf{q}).
 - ▶ Mỗi đối tượng cốt lõi có thể gom các đối tượng **có-thể-với-tới-trực-tiếp-theo-mậtđộ** từ nó để tạo lập một vùng mật độ cao (dense region).
- Làm thế nào để tạo lập một vùng có mật độ cao rộng (large dense region) từ các vùng có mật độ cao nhỏ hơn (small dense regions) đại diện bởi các đối tượng cốt lõi?

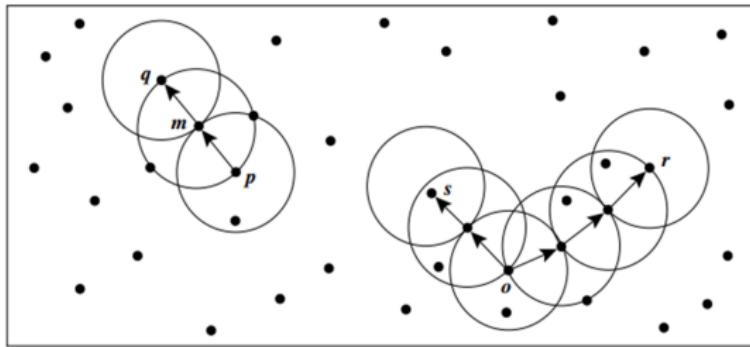
Có-thể-với-tới-theo-mật-degree (density-reachable)

- Trong DBSCAN, cho hai đối tượng $p, q \in D$, p có-thể-với-tới-theo-mật-degree từ q (với các ngưỡng ϵ và $MinPts$) nếu:
 - Tồn tại một chuỗi các đối tượng p_1, p_2, \dots, p_h với $p_1 = q$ và $p_h = p$ và p_{i+1} có-thể-với-tới-trực tiếp-theo-mật-degree từ p_i ($1 \leq i < h$ và $p_i \in D$).
- Có-thể-với-tới-theo-mật-degree không phải là một quan hệ tương đương (equivalence relation) vì nó không có tính đối xứng:
 - Nếu q và p là hai đối tượng cốt lõi và nếu p có-thể-với-tới-theo-mật-degree từ q thì ngược lại q cũng có-thể-với-tới-theo-mật-degree từ p .
 - Còn nếu q là đối tượng cốt lõi và p không phải là đối tượng cốt lõi thì p có-thể-với-tới-theo-mật-degree từ q nhưng ngược lại thì không.

Kết-nối-theo-mật-độ (density-connectedness)

- Hai đối tượng $\mathbf{p}_1, \mathbf{p}_2 \in \mathbf{D}$ được gọi là kết-nối-theo-mật-độ (density-connected) với các ngưỡng ϵ và $MinPts$ nếu:
 - ▶ Tồn tại đối tượng $\mathbf{q} \in \mathbf{D}$ sao cho cả \mathbf{p}_1 và \mathbf{p}_2 có-thể-với-tới-theo-mật-độ từ \mathbf{q} .
- Kết-nối-theo-mật-độ là mối quan hệ tương đương (equivalence relation):
 - ▶ Tính đối xứng (symmetry): \mathbf{p}_1 kết-nối-theo-mật-độ với \mathbf{p}_2 thì \mathbf{p}_2 cũng kết-nối-theo-mật-độ với \mathbf{p}_1 .
 - ▶ Tính bắc cầu (transitivity): với $\mathbf{p}_1, \mathbf{p}_2$, và \mathbf{p}_3 , nếu \mathbf{p}_1 kết-nối-theo-mật-độ với \mathbf{p}_2 và \mathbf{p}_2 kết-nối-theo-mật-độ với \mathbf{p}_3 thì \mathbf{p}_1 cũng kết-nối-theo-mật-độ với \mathbf{p}_3 .

Ví dụ minh họa (với ngưỡng ϵ nào đó và $MinPts = 3$)



[nguồn: Data Mining: Concepts and Techniques by J. Han và cộng sự]

- **m, p, o** là các đối tượng (điểm) cốt lõi bởi trong ϵ -neighborhood của chúng có ít nhất 3 đối tượng khác.
- Đối tượng **q** có-thể-với-tới-trực-tiếp-theo-mật-degree từ **m** và **m** có-thể-với-tới-trực-tiếp-theo-mật-degree từ **p**.
- **q** có-thể-với-tới-theo-mật-degree (gián tiếp) từ **p**. Nhưng **p** không-thể-với-tới-theo-mật-degree từ **q** vì **q** không phải là đối tượng cốt lõi.
- **r** và **s** có-thể-kết-nối-theo-mật-degree nhờ **o**.

Cụm dữ liệu trong DBSCAN

Một tập con $S \in \mathbf{D}$ là một cụm nếu:

- Với bất kỳ hai đối tượng $x_1, x_2 \in S$, thì x_1 và x_2 có-thể-kết-nối-theo-mật-độ, và
- Không tồn tại đối tượng $x \in S$ và đối tượng $x' \in (\mathbf{D} \setminus S)$ mà x và x' có-thể-kết-nối-theo-mật-độ.

Cách thức DBSCAN phân cụm

- DBSCAN tìm các cụm theo cách:

- ▶ Khởi đầu, các đối tượng $x \in D$ được đánh dấu “unvisited”.
- ▶ DBSCAN chọn ngẫu nhiên một đối tượng “unvisited” x , đánh dấu x “visisted” và kiểm tra ϵ -neighborhood(x) có chứa ít nhất $MinPts$ đối tượng hay không.
 - ★ Nếu không: x được đánh dấu là “noise”.
 - ★ Nếu có: một cụm S được tạo ra cho x và tất cả các đối tượng trong ϵ -neighborhood(x) được thêm vào trong tập ứng viên C . DBSCAN lần lượt thêm các đối tượng x' trong C mà chưa thuộc cụm nào vào S . Với mỗi x' có nhãn “unvisited”, DBSCAN đánh dấu “visited” và kiểm tra ϵ -neighborhood(x') có chứa ít nhất $MinPts$ hay không. Nếu có, DBSCAN thêm tất cả các đối tượng thuộc ϵ -neighborhood(x') vào tập ứng viên C . DBSCAN tiếp tục kiểm tra các đối tượng trong C và thêm vào S cho đến khi S không thể mở rộng và C rỗng. Khi đó cụm S đã đầy đủ.
- ▶ Để tìm cụm tiếp theo, DBSCAN chọn ngẫu nhiên một đối tượng có nhãn “unvisited” và thực hiện như bước trên.
- ▶ Quá trình phân cụm kết thúc khi tất cả các đối tượng đều được xem xét (“visited”).

Mã thuật toán DBSCAN

```
1: procedure DBSCAN( $D = \{x^1, x^2, \dots, x^n\}$ ,  $\epsilon$ ,  $MinPts$ )
2:   Output: Tập các cụm dữ liệu  $S = \{S_1, S_2, \dots, S_k\}$ ; khởi tạo  $S = \emptyset$ ;
3:   đánh dấu tất cả đối tượng trong  $D$  là "unvisited";
4:   while (còn đối tượng có nhãn "unvisited") do
5:     chọn ngẫu nhiên một đối tượng "unvisited"  $x$  và gắn nhãn "visited" cho  $x$ ;
6:     if ( $\epsilon$ -neighborhood( $x$ ) có ít nhất  $MinPts$  đối tượng) then
7:       tạo cụm  $S$  và thêm  $x$  vào  $S$ ;
8:       khởi tạo  $C$  là tập tất cả các đối tượng thuộc  $\epsilon$ -neighborhood( $x$ );
9:       for (với mỗi đối tượng  $x' \in C$ ) do
10:         if (nếu  $x'$  có nhãn "unvisited") then
11:           đánh dấu  $x'$  là "visited";
12:           if ( $\epsilon$ -neighborhood( $x'$ ) có ít nhất  $MinPts$  đối tượng) then
13:             thêm các đối tượng trong  $\epsilon$ -neighborhood( $x'$ ) vào  $C$ ;
14:           end if
15:         end if
16:         if (nếu  $x'$  chưa thuộc cụm nào) then
17:           thêm  $x'$  vào  $S$ ;
18:         end if
19:       end for
20:       thêm cụm  $S$  vào  $S$ ;
21:     else
22:       gắn nhãn "noise" cho  $x$ ;
23:     end if
24:   end while
25: end procedure
```

Hiệu quả của DBSCAN

- Với n là số lượng đối tượng trong \mathbf{D} cần phân cụm.
- Nếu đánh chỉ mục các đối tượng theo không gian thì độ phức tạp tính toán của DBSCAN là $O(n \log n)$.
- Nếu không đánh chỉ mục, độ phức tạp tính toán là $O(n^2)$.
- Với các tham số ϵ và $MinPts$ phù hợp, DBSCAN thực sự hiệu quả trong việc tìm ra các cụm dữ liệu với bất cứ hình dạng nào.

Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đặc trưng phân cụm

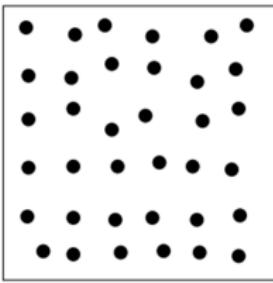
4 Tổng kết bài giảng

Đánh giá và đo đạc trong phân cụm

(Evaluation of Clustering)

- Đánh giá và hiểu xu hướng phân bố của cụm (assessing clustering tendency).
- Xác định số cụm dữ liệu (determining the number of clusters).
- Đánh giá và đo đạc chất lượng phân cụm (measuring clustering quality).

Đánh giá và hiểu xu hướng phân bố của cụm



[nguồn: Data Mining: Concepts and Techniques by J. Han và cộng sự]

- Với dữ liệu có phân bố hoàn toàn ngẫu nhiên:
 - ▶ Dữ liệu được phân phối đều (uniformly distributed) trong không gian.
 - ▶ Thiếu cấu trúc cụm.
 - ▶ Phân cụm với phương pháp nào thì kết quả phân cụm đều ít ý nghĩa.
- Với dữ liệu có phân bố phi ngẫu nhiên (non-random):
 - ▶ Các đối tượng không tuân theo phân phối đều.
 - ▶ Dữ liệu có cấu trúc cụm rõ ràng.
 - ▶ Kết quả phân cụm có ý nghĩa.
- Làm sao để đánh giá một tập dữ liệu có cấu trúc cụm rõ ràng?

Xác định xu hướng phân bố dữ liệu với Hopkins Statistic

- $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ là tập n đôi tượng dữ liệu trong không gian \mathbb{R}^m .
- Chọn ngẫu nhiên h ($< n$) đôi tượng $\mathbf{x}_1, \dots, \mathbf{x}_h$ từ \mathbf{D} . Với mỗi điểm \mathbf{x}_i , tìm khoảng cách từ nó tới lảng giềng gần nhất:

$$a_i = \min_{\mathbf{u} \in \mathbf{D}} d(\mathbf{x}_i, \mathbf{u})$$

- Sinh ngẫu nhiên h đôi tượng dữ liệu giả $\mathbf{y}_1, \dots, \mathbf{y}_h$ trong không gian \mathbb{R}^m với giá trị tại các chiều ngẫu nhiên trong miền xác định của chúng. Với mỗi đối tượng \mathbf{y}_i , tìm khoảng cách tối thiểu từ nó tới lảng giềng gần nhất thuộc \mathbf{D} :

$$b_i = \min_{\mathbf{v} \in \mathbf{D}} d(\mathbf{y}_i, \mathbf{v})$$

- Tính Hopkins Statistic, H , như sau:

$$H = \frac{\sum_{i=1}^h b_i}{\sum_{i=1}^h b_i + \sum_{i=1}^h a_i} \quad (16)$$

- Nếu \mathbf{D} có phân phối ngẫu nhiên thì H xấp xỉ 0.5, nếu H càng gần về 1.0 thì \mathbf{D} có cấu trúc cụm rõ ràng (non-random structure).

Xác định số cụm dữ liệu

Xem lại slide 44 (Lựa chọn số lượng cụm k phù hợp).

Đánh giá và đo đặc chất lượng phân cụm

- Giả sử đã đánh giá xu hướng phân bố dữ liệu (clustering tendency).
- Giả sử đã ước lượng số lượng cụm dữ liệu (number of clusters).
- Áp dụng nhiều phương pháp phân cụm trên tập dữ liệu.
- Câu hỏi:
 - ▶ Đánh giá chất lượng phân cụm của một phương pháp/thuật toán cụ thể.
 - ▶ So sánh chất lượng phân cụm của các phương pháp/thuật toán khác nhau.
- Cách thức đánh giá:
 - ▶ Nếu có dữ liệu đánh giá (ground truth) xây dựng bởi chuyên gia (human experts), sử dụng các phương pháp bên ngoài (extrinsic/supervised methods).
 - ▶ Ngược lại, sử dụng các phương pháp nội tại (intrinsic/unsupervised methods).

Các phương pháp bên ngoài (extrinsic methods)

Cách thức đánh giá:

- Gọi $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ là kết quả phân cụm của một thuật toán.
- Gọi $\mathbf{S}^g = \{S_1^g, S_2^g, \dots, S_h^g\}$ là phương án phân cụm tối ưu (ground truth) bởi các chuyên gia (human experts).
- Ước lượng điểm so sánh chất lượng phân cụm $Q(\mathbf{S}, \mathbf{S}^g)$.

So sánh chất lượng $Q(\mathbf{S}, \mathbf{S}^g)$ cần thỏa mãn các tính chất:

- Cluster homogeneity
- Cluster completeness
- Rag bag
- Small cluster preservation

Cluster homogeneity

- Các cụm càng thuần khiết (pure), chất lượng phân cụm càng cao.
- Giả sử các đối tượng trong \mathbf{D} thuộc về các lớp $S_1^g, S_2^g, \dots, S_h^g$, so sánh hai phương án phân cụm \mathbf{S}_1 và \mathbf{S}_2 :
 - ▶ Nếu một cụm $S \in \mathbf{S}_1$ bao gồm các đối tượng thuộc hai lớp S_i^g và S_j^g .
 - ▶ Nếu phương án phân cụm \mathbf{S}_2 y giống hệt phương án \mathbf{S}_1 ngoại trừ S trong \mathbf{S}_2 được chia thành hai cụm mà mỗi cụm chỉ chứa các phần tử thuộc lớp S_i^g hoặc lớp S_j^g .
 - ▶ Khi đó: $Q(\mathbf{S}_2, \mathbf{S}^g) > Q(\mathbf{S}_1, \mathbf{S}^g)$ vì mức độ thuần khiết của phương án \mathbf{S}_2 cao hơn phương án \mathbf{S}_1 .

Cluster completeness

- Các đối tượng thuộc cùng một lớp trong phương án phân cụm tối ưu (ground truth) thì nên được phân vào cùng một cụm bởi các thuật toán.
- Trong phương án phân cụm \mathbf{S}_1 , hai cụm S_1 và S_2 có đối tượng thuộc cùng một lớp nào đó của phương án tối ưu \mathbf{S}^g .
- Phương án phân cụm \mathbf{S}_2 giống hết phương án \mathbf{S}_1 ngoại trừ hai cụm S_1 và S_2 đã được gộp lại thành một cụm.
- Khi đó: $Q(\mathbf{S}_2, \mathbf{S}^g) > Q(\mathbf{S}_1, \mathbf{S}^g)$.

Rag bag

- Trong hầu hết các bài toán thực tế, luôn có lớp “rag bag” bao gồm các đối tượng không sát nhập được với các đối tượng hay cụm khác (thường gọi là “other” hoặc “miscellaneous”).
- Tiêu chí “rag bag” cho thấy: việc đưa một đối tượng vào một cụm thuần khiết nên bị “phạt” (penalized) nhiều hơn so với việc đưa đối tượng đó vào cụm “rag bag”.

Small cluster preservation

- Nếu một lớp nhỏ (small category) được chia thành các phần nhỏ trong một phương án phân cụm, những phần nhỏ đó có thể trở thành nhiễu (noise) và lớp nhỏ đó khó được khôi phục từ phương án phân cụm đó.
- Tiêu chí này phát biểu rằng: chia một lớp nhỏ thành các thành phần nhỏ sẽ gây hậu quả nghiêm trọng hơn so với việc chia một lớp lớn (large category) thành các thành phần.
- Ví dụ:
 - ▶ Tập dữ liệu \mathbf{D} có $n + 2$ đối tượng, trong đó $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n$ thuộc về một lớp trong \mathbf{S}^g , còn \mathbf{x}^{n+1} và \mathbf{x}^{n+2} thuộc về một lớp nhỏ khác.
 - ▶ Phương án phân cụm thứ nhất:
$$\mathbf{S}_1 = \{\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}, \{\mathbf{x}^{n+1}\}, \{\mathbf{x}^{n+2}\}\}.$$
 - ▶ Phương án phân cụm thứ hai:
$$\mathbf{S}_2 = \{\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{n-1}\}, \{\mathbf{x}^n\}, \{\mathbf{x}^{n+1}, \mathbf{x}^{n+2}\}\}.$$
 - ▶ Khi đó: $Q(\mathbf{S}_2, \mathbf{S}^g) > Q(\mathbf{S}_1, \mathbf{S}^g)$.

BCubed precision và BCubed recall

- Hầu hết các phép đo chất lượng thỏa mãn một số trong bốn tiêu chí trên.
- *BCubed precision* và *recall* thỏa mãn cả bốn tiêu chí trên.
 - ▶ BCubed precision của một đối tượng cho biết có bao nhiêu đối tượng khác trong cùng một cụm dữ liệu thuộc về cùng lớp của nó.
 - ▶ BCubed recall của một đối tượng phản ánh số đối tượng thuộc cùng lớp được phân cùng cụm.
- Quy ước các ký hiệu:
 - ▶ Cho $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ là tập n đối tượng cần phân cụm.
 - ▶ Gọi \mathbf{S} là một phương án phân cụm trên \mathbf{D} .
 - ▶ Gọi $S^g(\mathbf{x}^i)$ là lớp của đối tượng \mathbf{x}^i (ground truth), và $S(\mathbf{x}^i)$ là định danh cụm của đối tượng \mathbf{x}^i theo phương án phân cụm \mathbf{S} .
 - ▶ Khi đó, mức độ chính xác (correctness) của hai đối tượng \mathbf{x}^i và \mathbf{x}^j ($1 \leq i, j \leq n, i \neq j$) được xác định:
 - ★ $Correctness(\mathbf{x}^i, \mathbf{x}^j) = 1$ nếu $S^g(\mathbf{x}^i) = S^g(\mathbf{x}^j)$ và $S(\mathbf{x}^i) = S(\mathbf{x}^j)$.
 - ★ $Correctness(\mathbf{x}^i, \mathbf{x}^j) = 0$ nếu ngược lại.

BCubed precision và BCubed recall (2)

- BCubed precision được định nghĩa:

$$BCubedPrecision = \frac{\sum_{i=1}^n \frac{\sum_{x^j: i \neq j, S(x^i) = S(x^j)} \text{Correctness}(x^i, x^j)}{|\{x^j | i \neq j, S(x^i) = S(x^j)\}|}}{n} \quad (17)$$

- BCubed recall được định nghĩa:

$$BCubedRecall = \frac{\sum_{i=1}^n \frac{\sum_{x^j: i \neq j, S^g(x^i) = S^g(x^j)} \text{Correctness}(x^i, x^j)}{|\{x^j | i \neq j, S^g(x^i) = S^g(x^j)\}|}}{n} \quad (18)$$

Các phương pháp nội tại (intrinsic methods)

- Được sử dụng khi không có dữ liệu chuyên gia (ground truth).
- Đánh giá chất lượng phân chia cụm và mức độ gắn kết của các cụm.
- Phương pháp: hệ số silhouette (silhouette coefficient).

Hệ số silhouette (silhouette coefficient)

- Giả sử tập $\mathbf{D} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ được phân hoạch thành k cụm S_1, S_2, \dots, S_k . Với mỗi $\mathbf{x} \in \mathbf{D}$, tính:
 - $a(\mathbf{x})$ là khoảng cách trung bình giữa \mathbf{x} và các đối tượng khác trong cụm chứa \mathbf{x} (giả sử đó là cụm S_i):

$$a(\mathbf{x}) = \frac{\sum_{\mathbf{x}' \in S_i, \mathbf{x}' \neq \mathbf{x}} d(\mathbf{x}, \mathbf{x}')}{|S_i| - 1} \quad (19)$$

- $b(\mathbf{x})$ là giá trị nhỏ nhất trong các giá trị trung bình khoảng cách từ \mathbf{x} đến các đối tượng thuộc các cụm không chứa \mathbf{x} :

$$b(\mathbf{x}) = \min_{S_j: 1 \leq j \leq k, j \neq i} \frac{\sum_{\mathbf{x}' \in S_j} d(\mathbf{x}, \mathbf{x}')}{|S_j|} \quad (20)$$

Hệ số silhouette (silhouette coefficient) (2)

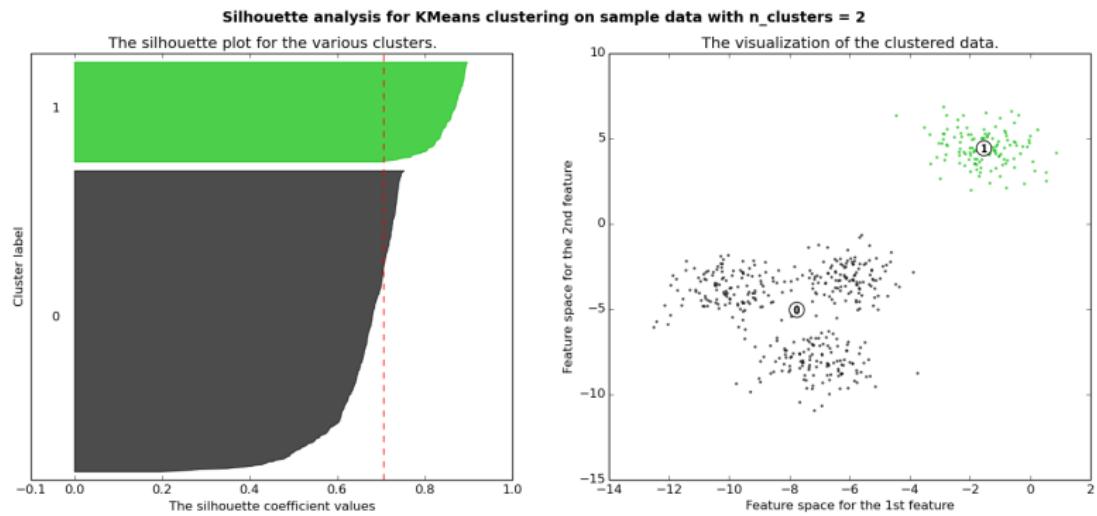
- Hệ số silhouette của đối tượng x được xác định như sau:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \quad (21)$$

- Giá trị của $s(x)$ biến thiên từ -1 đến 1 .
 - ▶ Giá trị $a(x)$ nhỏ cho biết mức độ gắn kết trong cụm chứa x cao.
 - ▶ Giá trị $b(x)$ lớn cho biết x càng xa các cụm còn lại.
 - ▶ Khi $s(x)$ tiệm cận 1 thì cụm chứa x rất gắn kết và x xa các cụm còn lại.
 - ▶ Khi $s(x)$ có giá trị âm thì x gần các đối tượng trong các cụm khác hơn là các đối tượng trong cụm chứa nó.
- Để đo chất lượng của phương án phân cụm, tính hệ số silhouette trung bình trên toàn bộ các đối tượng.
- Hệ số silhouette cũng có thể được sử dụng như một thước đo để xác định số cụm trong một tập dữ liệu.

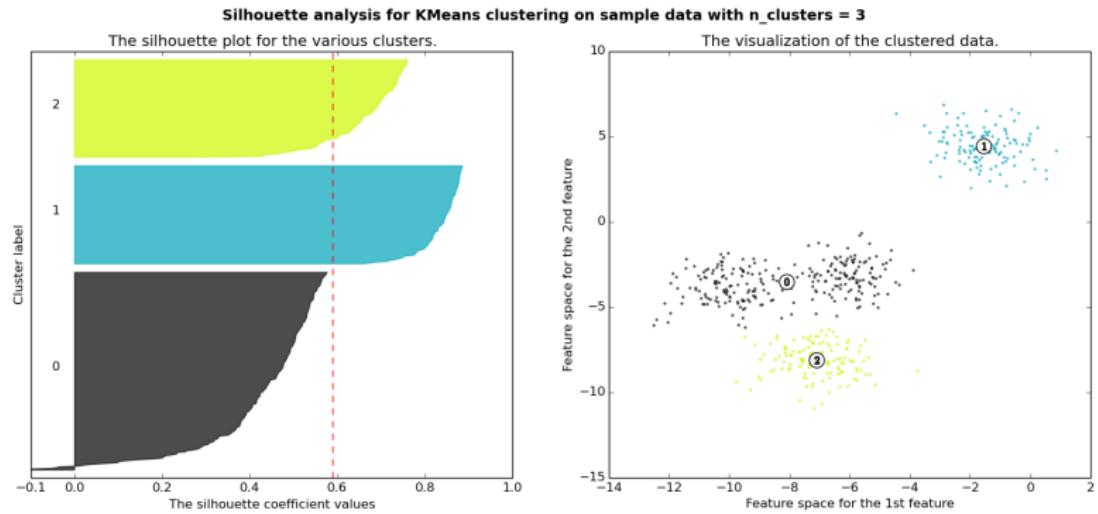
Trực quan hóa hệ số silhouette cho phân cụm K-means

[nguồn: scikit-learn.org]



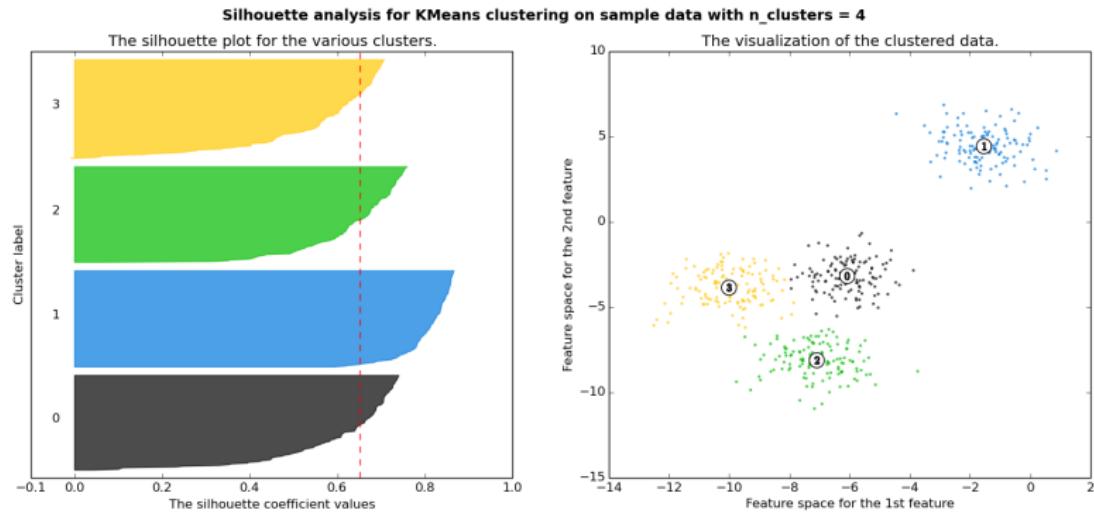
Trực quan hóa hệ số silhouette cho phân cụm K-means

[nguồn: scikit-learn.org]



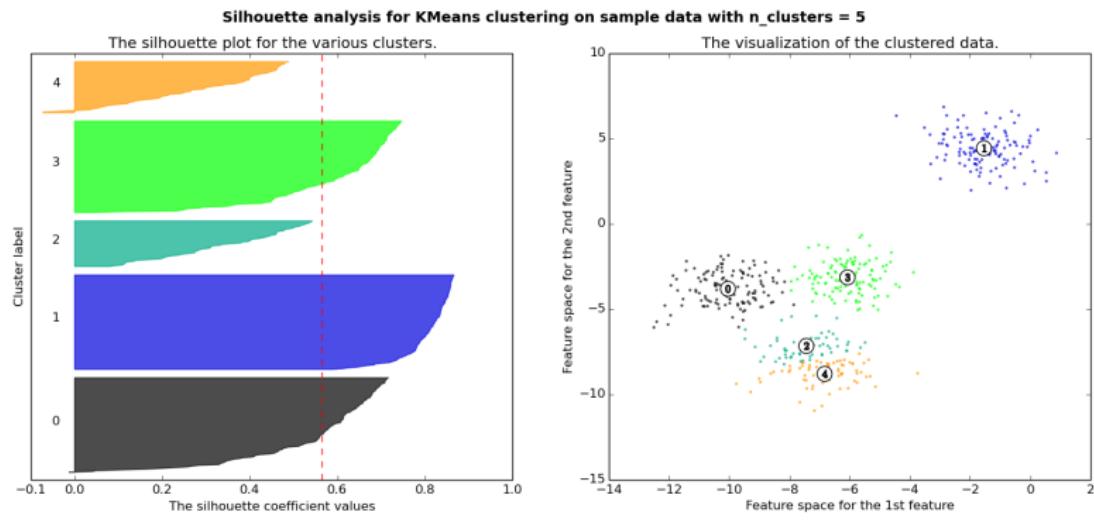
Trực quan hóa hệ số silhouette cho phân cụm K-means

[nguồn: scikit-learn.org]



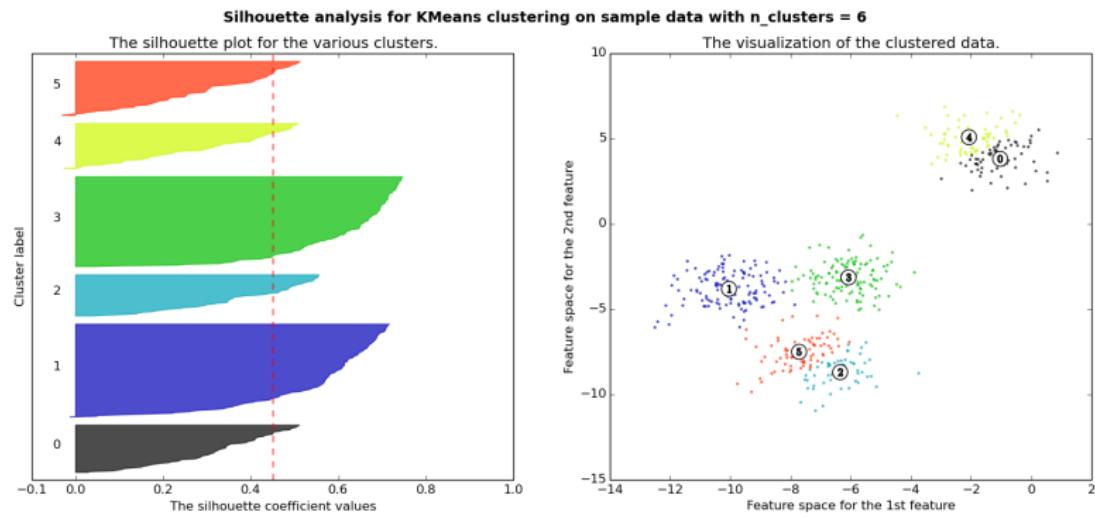
Trực quan hóa hệ số silhouette cho phân cụm K-means

[nguồn: scikit-learn.org]



Trực quan hóa hệ số silhouette cho phân cụm K-means

[nguồn: scikit-learn.org]



Nội dung

1 Các khái niệm cơ bản

- Vấn đề phân cụm và ứng dụng
- Điểm (points) và không gian (spaces)
- Các độ đo khoảng cách (distances)
- Các tiếp cận phân cụm (clustering approaches)
- Phân cụm trong không gian số chiều lớn (high-dimensional space)

2 Các phương pháp phân cụm quan trọng

- Phân cụm phân cấp (hierarchical clustering)
- Thuật toán K -means
- Phân cụm dựa trên mật độ: thuật toán DBSCAN

3 Đánh giá và đo đạc trong phân cụm

4 Tổng kết bài giảng

Tổng kết bài giảng

- Định nghĩa và ý nghĩa của bài toán phân cụm.
- Khái niệm đối tượng, điểm, véc tơ, không gian.
- Các độ đo khoảng cách: Euclid, Cosine, Manhattan, Jaccard, ...
- Sáu cách tiếp cận phân cụm: phân cấp, phân hoạch, phân bô, mật độ, ...
- Khó khăn khi phân cụm trong không gian số chiều (rất) lớn.
- Ba phương pháp/thuật toán quan trọng: phân cấp, K-means, DBSCAN.
- Xác định số cụm trong tập dữ liệu.
- Đánh giá và đo đặc chất lượng phân cụm.

Tài liệu tham khảo



C. M. Bishop.

Pattern Recognition and Machine Learning (9. Mixture Models and EM).

Springer, 2006.



J. Han, M. Kamber, and J. Pei.

Data Mining: Concepts and Techniques (Chapter 10. Cluster Analysis: Basic Concepts and Methods).

The 3rd Edition, Morgan Kaufmann, Elsevier, 2012.



A. Rajaraman, J. Leskovec, and J. D. Ullman.

Mining of Massive Datasets (7. Clustering).

The 2nd Edition, Cambridge University Press, 2013.



M. J. Zaki and W. M. Jr.

Data Mining and Analysis: Fundamental Concepts and Algorithms (III. Clustering).

Cambridge University Press, 2013.