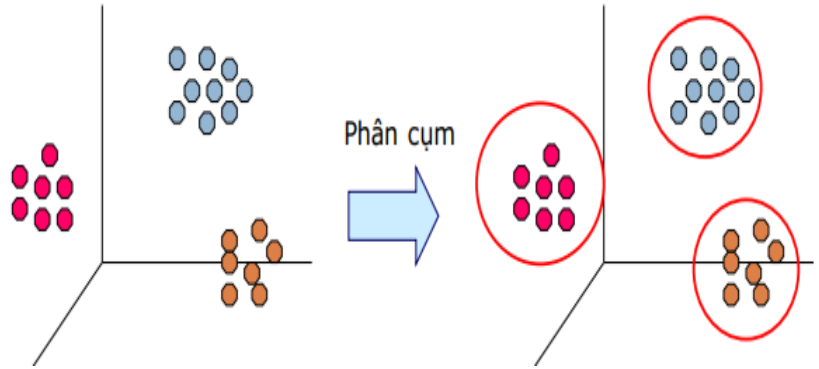
Phân cụm dữ liệu

Nội dung

- Tổng quan về phân cụm dữ liệu
- Thuật toán K-Means
- Bài tập

Tổng quan về phân cụm dữ liệu

- Phân cụm: là quá trình phân chia tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm thì "tương tự" (Similar) với nhau các phần tử trong các cụm khác nhau sẽ "không tương tự" (dissimilar) với nhau.



- Mục tiêu phân lớp: trích rút các đặc trưng từ dữ liệu cho phép phân loại các phần tử mới vào các lớp đã xác định
- Mục tiêu phân cụm: nhóm các đối tượng tương tự, nhờ đó phát hiện cấu trúc ẩn của dữ liệu

Tổng quan về phân cụm dữ liệu

- Một số độ đo được sử dụng:
  - Khoảng cách Euclidean, Mahattan, Minkowski
  - Độ tương tự cosine

Tổng quan về phân cụm dữ liệu

- Lĩnh vực ứng dụng
  - Nghiên cứu thị trường (Marketing): Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng lớn, phân loại và dự đoán hành vi khách hàng, …)
  - Sinh học (Biology): Phân nhóm động vật, thực vật, ...
  - Tài chính, Bảo hiểm (Finance and Insurance): Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng,
  - …

Tổng quan về phân cụm dữ liệu

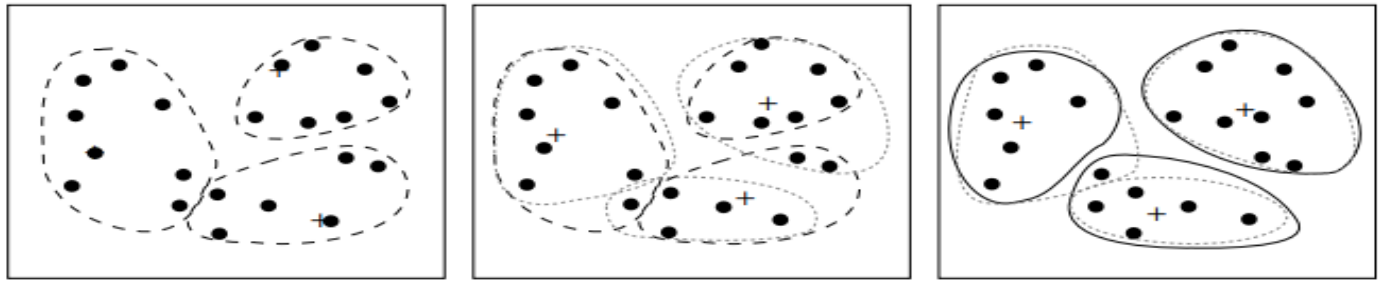
- Một số phương pháp phân cụm
  - **Phân hoạch** (partitioning): phân hoạch một tập hợp dữ liệu có n phần tử thành k nhóm (k <= n) và sau đó đánh giá chúng dựa trên các tiêu chí xác định.
    Thuật toán điển hình: K-means, K-medoids, CLARANS,…
  - **Phân cấp** (hierarchical): xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét.
    Thuật toán điển hình: BIRCH, Chameleon,…
  - **Phân cụm dựa trên mật độ** (Density-Based): nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định.
    Thuật toán điển hình: DBSCAN, OPTICS,…
  - **Phân cụm dựa trên lưới** (Grid-Based): thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm.
    Thuật toán điển hình: STING, CLIQUE,…

Nội dung

- Tổng quan về phân cụm dữ liệu
- Thuật toán K-Means
- Bài tập

Thuật toán K-Means

- Ý tưởng:
  - Mỗi cụm được đại diện bởi trọng tâm.
  - Một đối tượng được phân vào một cụm nếu khoảng cách từ đối tượng đó đến trọng tâm của cụm đang xét là nhỏ nhất.
  - Sau đó trọng tâm của cụm được cập nhật lại.
  - Quá trình lặp đi lặp lại cho đến khi các trọng tâm không đổi.



(a) Initial clustering (b) Iterate (c) Final clustering

Thuật toán K-Means

- Thuật toán

Input:

- k: the number of clusters,
- D: a data set containing n objects.

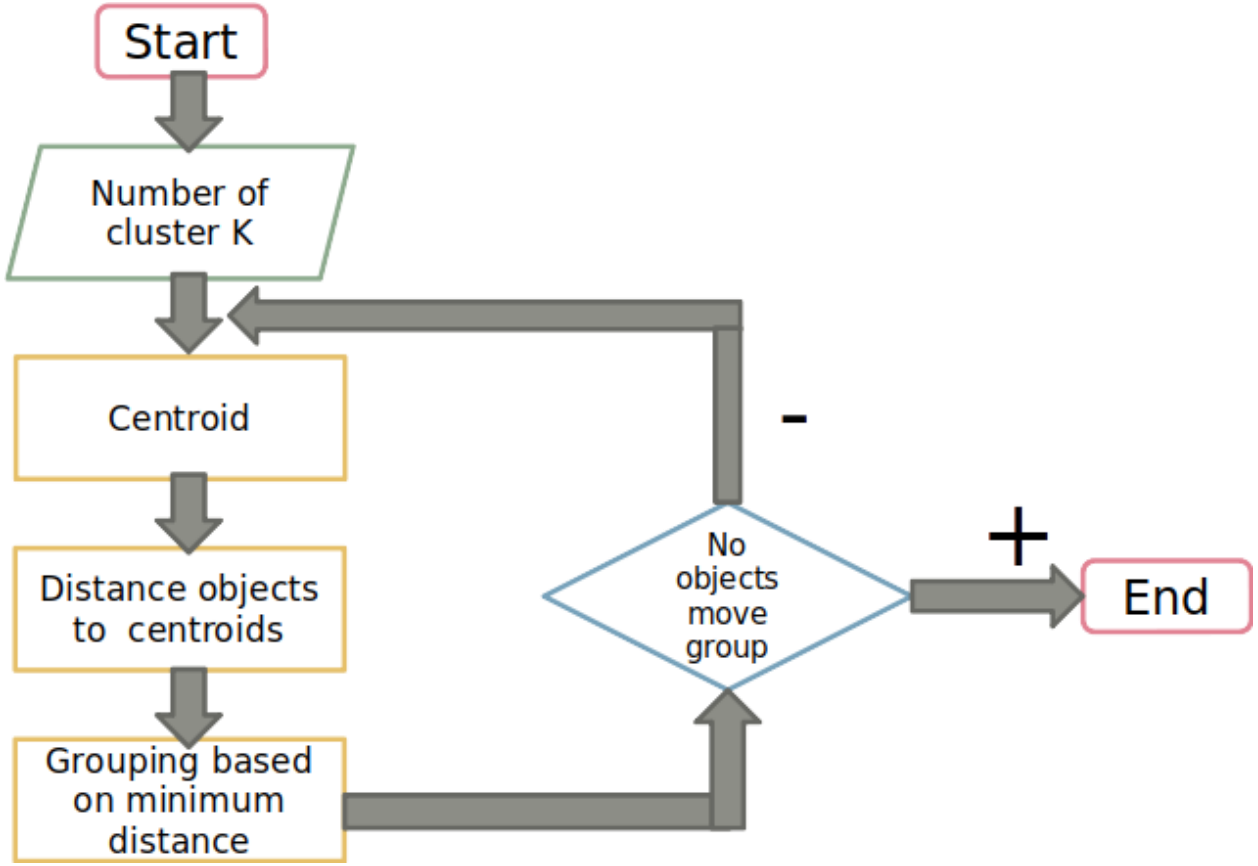
Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Thuật toán K-Means

- Sơ đồ thuật toán



Thuật toán K-Means

- Ví dụ
  - Phân cụm dữ liệu sau với k=2

Đối tượng	Thuộc tính 1(x)	Thuộc tính 2(y)
A	1	2
B	2	2
C	4	4
D	6	5

Thuật toán K-Means

- Bước 1: Khởi tạo
  - Chọn 2 trọng tâm ban đầu:

    $C_1(1,2) \equiv A$ và $C_2(2,2) \equiv B$, thuộc 2 cụm 1 và 2
- Lặp 1:
  - Tính toán khoảng cách

	$C_1(1, 2)$	$C_2(2, 2)$	Cụm
A(1, 2)	**0**	1	1
B(2, 2)	1	**0**	2
C(4, 4)	3.606	**2.828**	2
D(6, 5)	5.381	**5**	2

  - Cụm 1 : A(1,2); Cụm 2: B(2,2), C(4,4), D(6,5)
  - Cập nhật lại trọng tâm: C_1 = A(1,2); $C_2(x,y)$= ((2+4+6)/3, (2+4+5/3)=(4,11/3)

Thuật toán K-Means

- Lặp 2:
  - Tính toán khoảng cách

	c1(1, 2)	c2(4, 11/3)	Cụm
A(1, 2)	**0**	3.442	1
B(2, 2)	**1**	2.603	**1**
C(4, 4)	3.606	**0.333**	2
D(6, 5)	5.381	**2.404**	2

  - Cụm 1 : A(1,2), B(2,2); Cụm 2: C(4,4), D(6,5)
  - Cập nhật lại trọng tâm: C_1 (x,y) = ((1+2)/2,(2+2)/2) = (3/2,2)

    C_2(x,y)= ((4+6)/2, (4+5/2)=(5,9/2)

Thuật toán K-Means

- Lặp 3:
  - Tính lại khoảng cách

	C_1(3/2, 2)	C_2(5, 9/2)	Cụm
A(1, 2)	**0.5**	4.717	1
B(2, 2)	**0.5**	3.905	1
C(4, 4)	3.202	**1.118**	2
D(6, 5)	5.408	**1.118**	2

  - Cụm 1 : A(1,2), B(2,2); Cụm 2: C(4,4), D(6,5)
  - Cập nhật lại trọng tâm: C_1 (x,y) = ((1+2)/2,(2+2)/2) = (3/2,2)

    C_2(x,y)= ((4+6)/2, (4+5/2)=(5,9/2)
  - Không có sự thay đổi tâm cụm => Dừng

Nội dung

- Tổng quan về phân cụm dữ liệu
- Thuật toán K-Means
- Bài tập

Bài tập

Bài 1. Cho tập dữ liệu sau

	X1	X_2
A	1	2
B	2	2
C	2	3
D	3	3
E	3	4
F	2	4

Tiến hành phân cụm tập dữ liệu trên với k=2 với tâm khởi tạo cho 2 cụm là A và F

Bài tập

Bài 2(exercise 10.2 page 492) Suppose that the data mining task is to cluster points (with *(x,y)* representing location)
into three clusters, where the points are

$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9)$

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) The three cluster centers after the first round of execution

(b) The final three clusters