Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

TÀI LIỆU LÝ THUYẾT KTDL & UD

Gom nhóm dữ liệu (P1)
Cluster Analysis

Giảng viên: ThS. Lê Ngọc Thành
Email: lnthanh@fit.hcmus.edu.vn

Summer 2012

Nội dung

- **Khái niệm cơ sở về gom nhóm**
  - Gom nhóm là gì?
  - Ứng dụng của gom nhóm
  - Thế nào là một nhóm tốt?
  - Yêu cầu đối với phương pháp gom nhóm
  - Đo độ tương tự
  - Một số phương pháp gom nhóm
- Phương pháp phân hoạch
- Phương pháp phân cấp

Tình huống

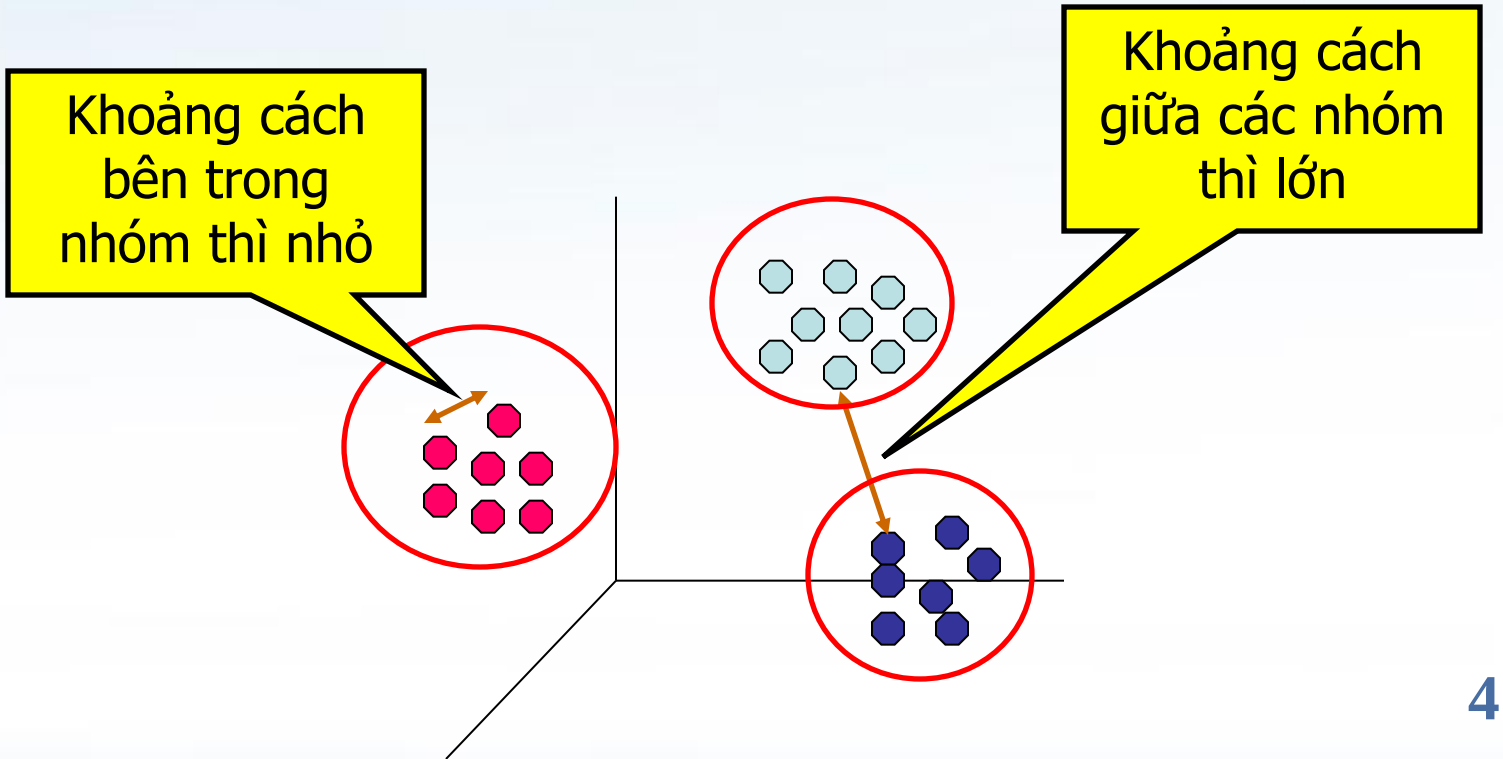
- Bạn là giám đốc một cửa hàng bán máy tính và bạn có 5 nhà quản lý giúp bạn.
- Bạn muốn chia khách hàng công ty thành 5 nhóm để phân công cho 5 nhà quản lý.
- Cách chia: bạn muốn các khách hàng trong mỗi nhóm tương tự nhau về các đặc trưng nào đó.

- Mục tiêu là bạn sẽ có các chiến lược kinh doanh khác nhau đối với từng nhóm

Phương pháp nào giúp bạn giải quyết bài toán này?

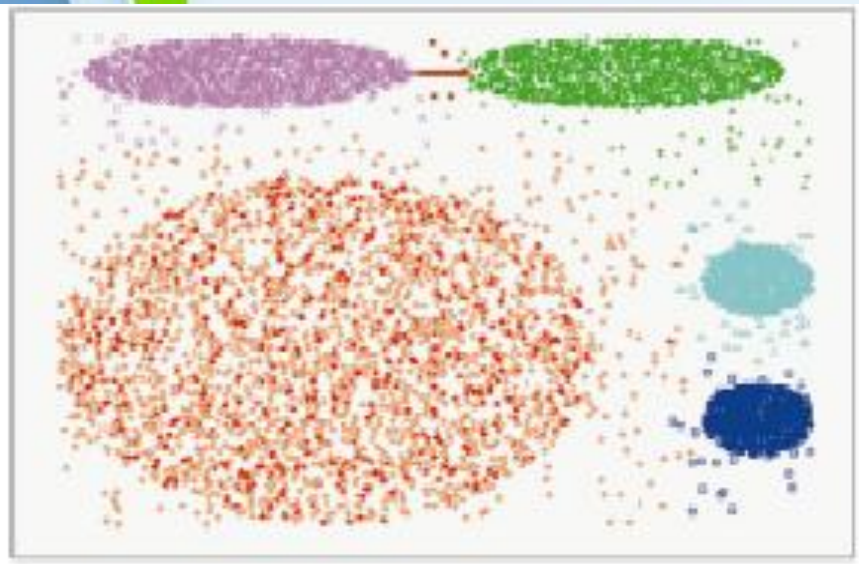
Gom nhóm (1/2)

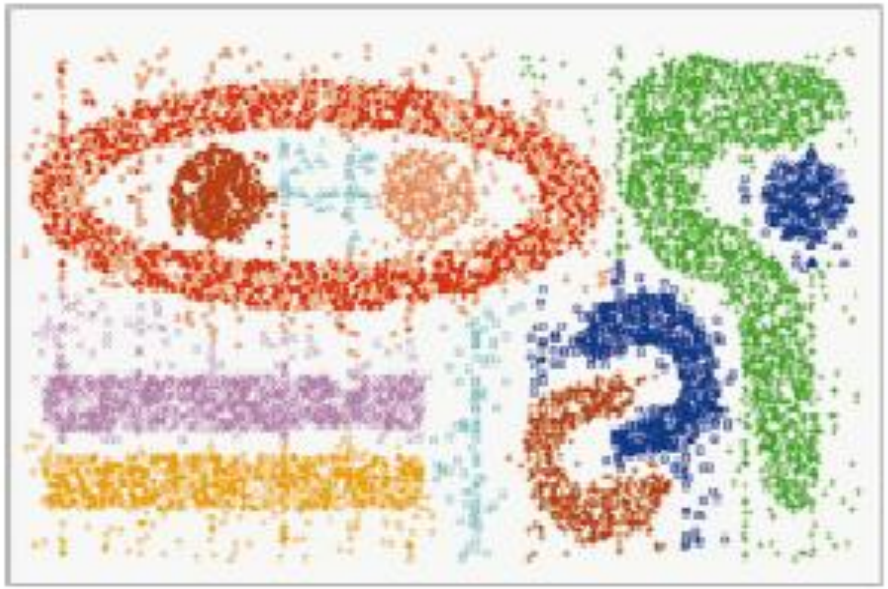
- Gom nhóm (clustering) là quá trình nhóm các đối tượng thành những cụm sao cho:
  - Các đối tượng cùng nhóm có độ tương tự cao.
  - Và rất khác với đối tượng ở các nhóm còn lại.

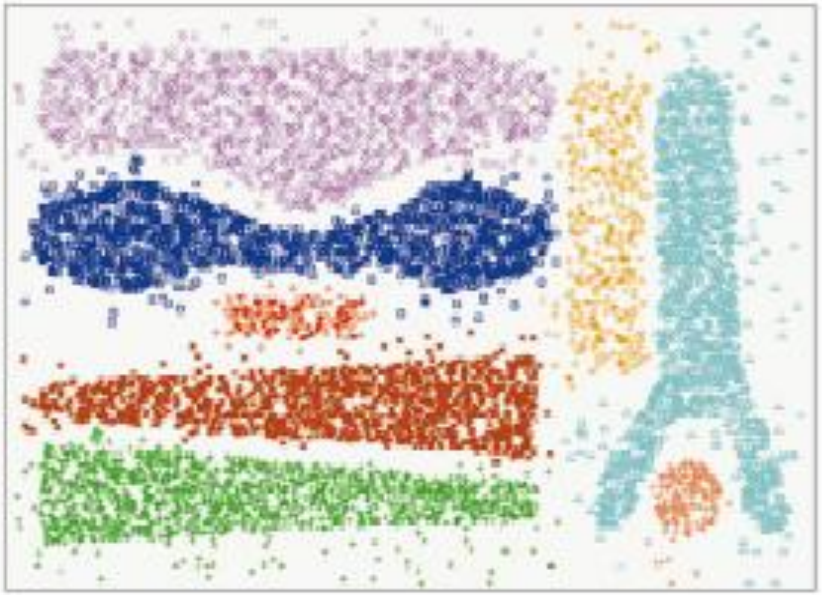


Gom nhóm (2/2)

- Gom nhóm là dạng *học không giám sát* (unsupervised learning) bởi vì nhãn/lớp không được định trước
- Vì vậy, gom nhóm là dạng của *học dựa trên quan sát* (learning by observation) hơn là *học dựa trên mẫu* (learning by examples)



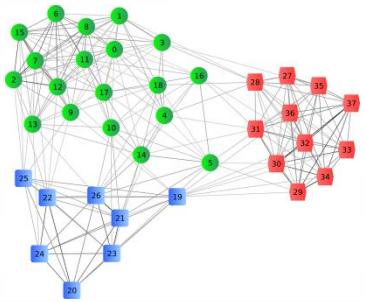




CHAMELEON

Một số ứng dụng gom nhóm

- Nhóm các tài liệu liên quan để duyệt web
- Nhóm các gien và protein có cùng chức năng
- Nhóm các cổ phiếu có cùng biến động
- Nhóm các khu vực có loại đất giống nhau trong địa lý
- Xác định nhóm nhà theo loại nhà, giá trị và vị trí địa lý
- Xác định nhóm đối tượng chơi game
- …

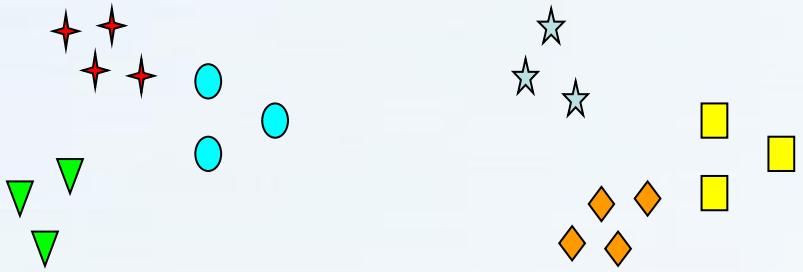


Phân loại ứng dụng

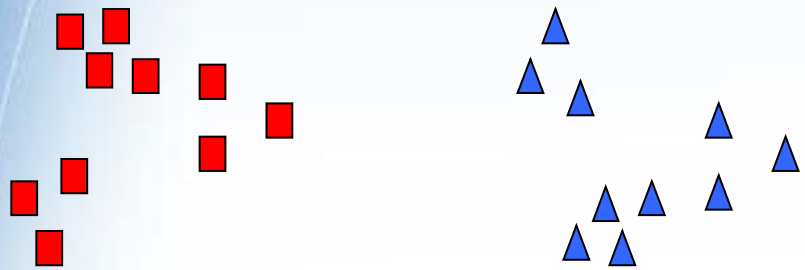
- Loại ứng dụng đặc thù:
  - Gom nhóm đóng vai trò như là *công cụ độc lập* để tìm hiểu sự phân bố dữ liệu (các ví dụ trước)
  - Gom nhóm đóng vai trò như là bước *tiền xử lý* cho các thuật toán khác
    - Ví dụ: đặc trưng hóa dữ liệu, chọn lựa tập con của thuộc tính, phát hiện outlier,…

Thế nào là một nhóm?

Có bao nhiêu nhóm?



6 nhóm



2 nhóm

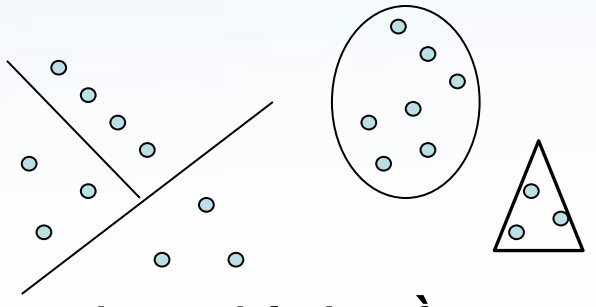
4 nhóm

Một gom nhóm tốt?

- Một phương pháp gom nhóm tốt sẽ phải tạo ra các nhóm có chất lượng cao:
  - Độ tương tự trong nhóm cao.
  - Độ tương tự với các nhóm khác thấp.
- Chất lượng của việc gom nhóm phụ thuộc vào:
  - Độ đo sự tương tự
  - Sự thực thi của nó
  - Khả năng khám phá ra một số hay tất cả mẫu tiềm ẩn

Yêu cầu đối với phương pháp gom nhóm (1/3)

- *Tính mở rộng*:
  - Làm việc trên cơ sở dữ liệu lớn
- *Khả năng xử lý các loại khác nhau của dữ liệu:*
  - Xử lý dữ liệu số, nhị phân, rời rạc hay đồ thị, luồng, hình ảnh, văn bản…
- *Khám phá ra nhóm dưới bất kì hình dạng nào:*



  - Thường các gom nhóm tìm ra dạng hình cầu với kích thước và mật độ giống nhau. Nó còn đòi hỏi tìm ra nhóm dựa trên hình dạng đường biên bất kì

Yêu cầu đối với phương pháp gom nhóm (2/3)

- *Mức đòi hỏi kiến thức để xác định tham số đầu vào:*
  - Một số thuật toán đòi phải biết số nhóm được gom. Trong một số trường hợp việc xác định này rất khó như dữ liệu đa chiều, người dùng không nắm rõ. Kết quả của việc gom nhóm sẽ tùy thuộc rất nhiều vào các tham số này.

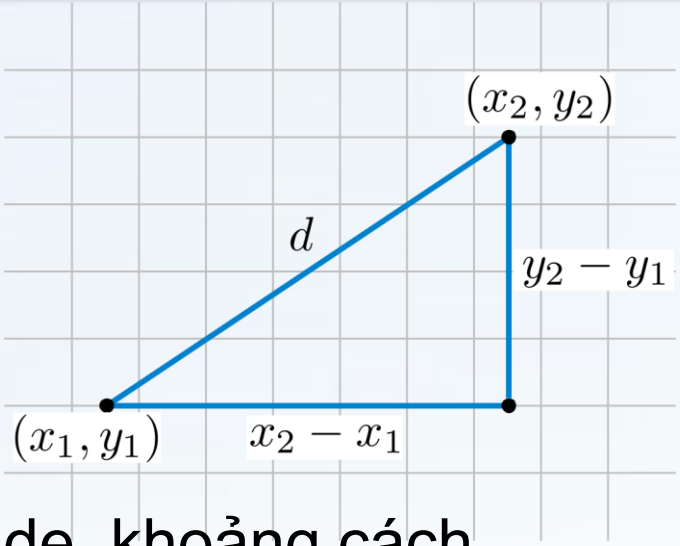
- *Khả năng vượt qua dữ liệu nhiễu*
- *Gom nhóm tăng cường và độc lập với thứ tự đầu vào*:
  - Dữ liệu mới vào sẽ được xử lý tiếp hay phải làm lại từ đầu? Thứ tự đưa vào có ảnh hưởng đến kết quả không?

Yêu cầu đối với phương pháp gom nhóm (3/3)

- *Khả năng gom nhóm dữ liệu đa chiều:*
  - Ví dụ: khi gom nhóm văn bản, mỗi từ trong văn bản đóng vai trò như một thuộc tính (chiều)
- *Gom nhóm dựa trên một số ràng buộc:*
  - Ví dụ: bạn được yêu cầu lắp các máy ATM ở một khu vực, bạn tiến hành gom nhóm dựa trên số hộ gia đình. Đi kèm với đó là các ràng buộc như mạng lưới giao thông, số lượng cũng như loại khách hàng trong nhóm.
- *Dễ hiểu và sử dụng được:*
  - PP cần giải thích ngữ nghĩa và ứng dụng. Chỉ ra mục tiêu của ứng dụng ảnh hưởng đến việc chọn các đặc trưng và pp gom nhóm như thế nào?

Đo độ tương tự

- Khoảng cách được sử dụng chủ yếu để đo độ tương tự hay không tương tự giữa hai đối tượng.
  - Ví dụ: khoảng cách Euclide, khoảng cách Cosin, Minkowski, Mahattan…

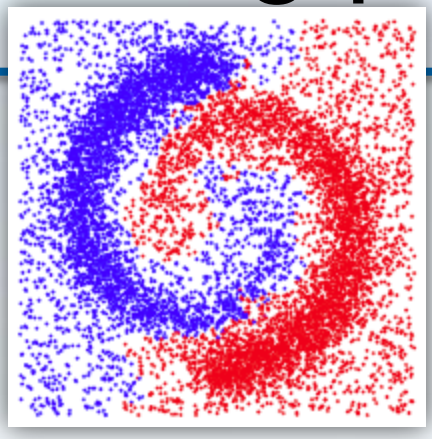


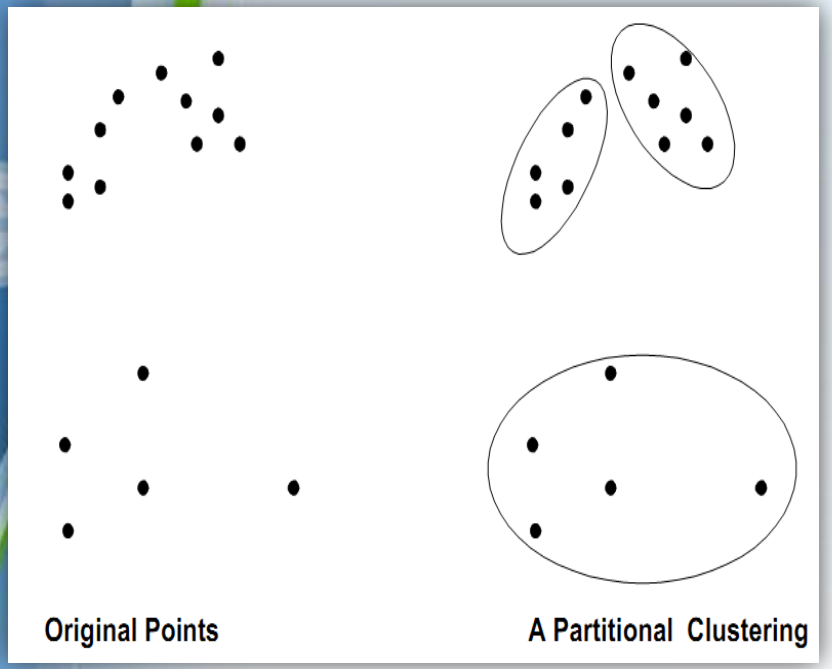
- Các hàm khoảng cách thường khác nhau về phạm vi giá trị, loại, bậc và thành phần biến
- Trọng số các biến phụ thuộc trên ứng dụng và ý nghĩa dữ liệu.

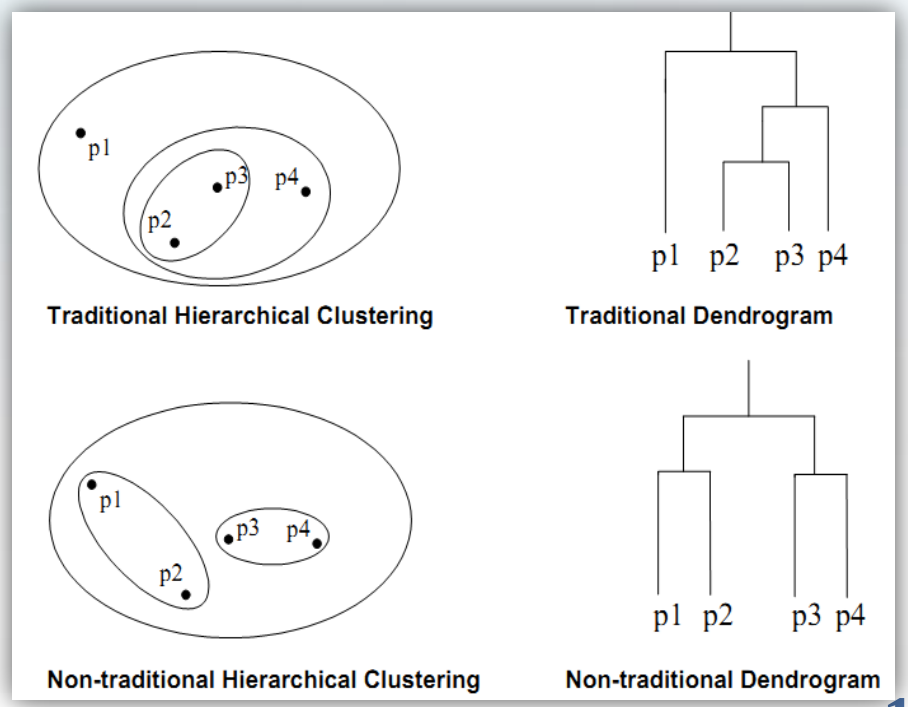
Một số pp gom nhóm (1/2)

- **Phương pháp phân hoạch**
  - Hình thành các phân hoạch và đánh giá chúng dựa trên một số tiêu chí
  - Các thuật toán: k-means, k-medoids, CLARANS
- **Phương pháp phân tầng**
  - Tạo ra sự phân chia các tầng
  - Các thuật toán: Diana, Agnes, BIRCH, CAMELEON
- **Phương pháp dựa trên mật độ**
  - Dựa trên hàm kế và hàm mật độ
  - Các thuật toán: DBSACN, OPTICS, DenClue

Ví dụ về các phương pháp







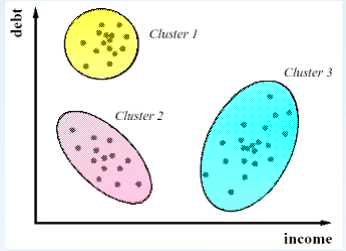
Một số pp gom nhóm (2/2)

- **Phương pháp dựa trên lưới**
  - Dựa trên cấu trúc hạt đa mức
  - Các thuật toán: STING, WaveCluster, CLIQUE
- **Phương pháp dựa trên mô hình**
- **Phương pháp dựa trên mẫu phổ biến**
- **Phương pháp dựa trên ràng buộc hay hướng dẫn của người dùng**
- **Phương pháp dựa trên liên kết**
- **…**

Nội dung

- Khái niệm cơ sở về gom nhóm
- **Phương pháp phân hoạch**
  - Khái niệm phân hoạch
  - Thuật toán k-means
  - Thuật toán k-medoids
- Phương pháp phân cấp

Gom nhóm phân hoạch



- *Partitioning Clustering* là pp đơn giản và nền tảng nhất trong số các pp gom nhóm
- Ý tưởng: phân hoạch một cơ sở dữ liệu D gồm n đối tượng thành tập *k nhóm* (k cho trước) sao cho tối ưu tiêu chí phân hoạch.
- Tối ưu toàn cục: thể hiện đầy đủ tất cả các nhóm
- Thuật toán heuristic:
  - k-means: mỗi nhóm được thể hiện bởi *giá trị trung tâm* của nhóm
  - k-medoids, PAM: mỗi nhóm được thể hiện bởi *một trong các đối tượng* của nhóm

Thuật toán k-means (1/2)

Cho trước số k, mỗi nhóm được biểu diễn bằng giá trị trung tâm (centroid) của nhóm.

- ***B1:*** Chọn ngẫu nhiên k đối tượng làm trung tâm của các nhóm
- ***B2:*** Gán từng đối tượng còn lại vào nhóm gần nhất dựa trên độ đo khoảng cách như Euclide, độ tương tự Cosin, correlation, …
- ***B3:*** Tính lại giá trị trung tâm của từng nhóm dựa trên các đối tượng mới gia nhập.
- ***B4:*** Nếu các trung tâm nhóm không có gì thay đổi hay chỉ còn một ít điểm thay đổi nhóm thì dừng, ngược lại quay lại B2.

Thuật toán k-means (2/2)

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

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

Output: A set of *k* clusters.

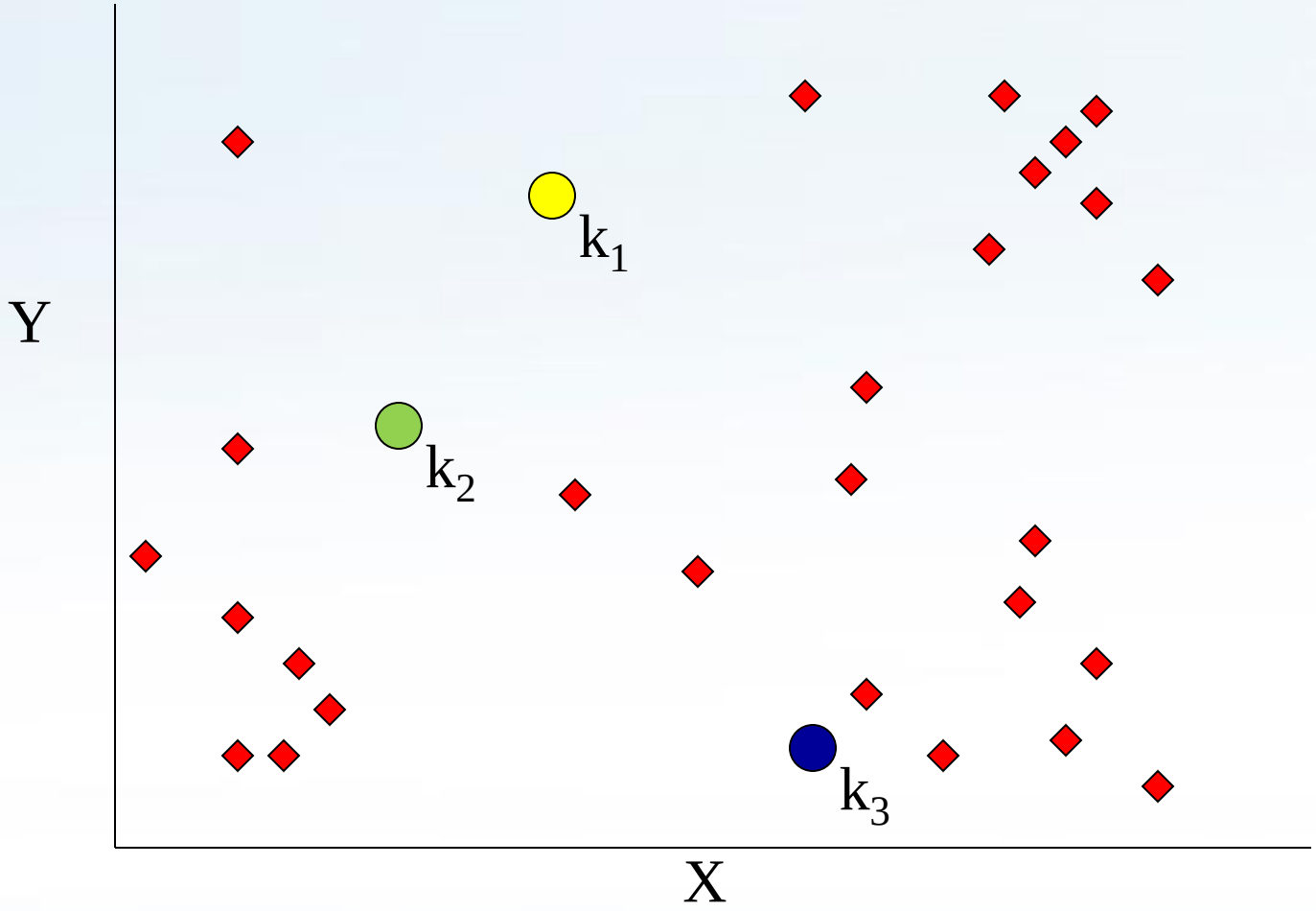
Method:

(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;

Ví dụ K-means (1/4)

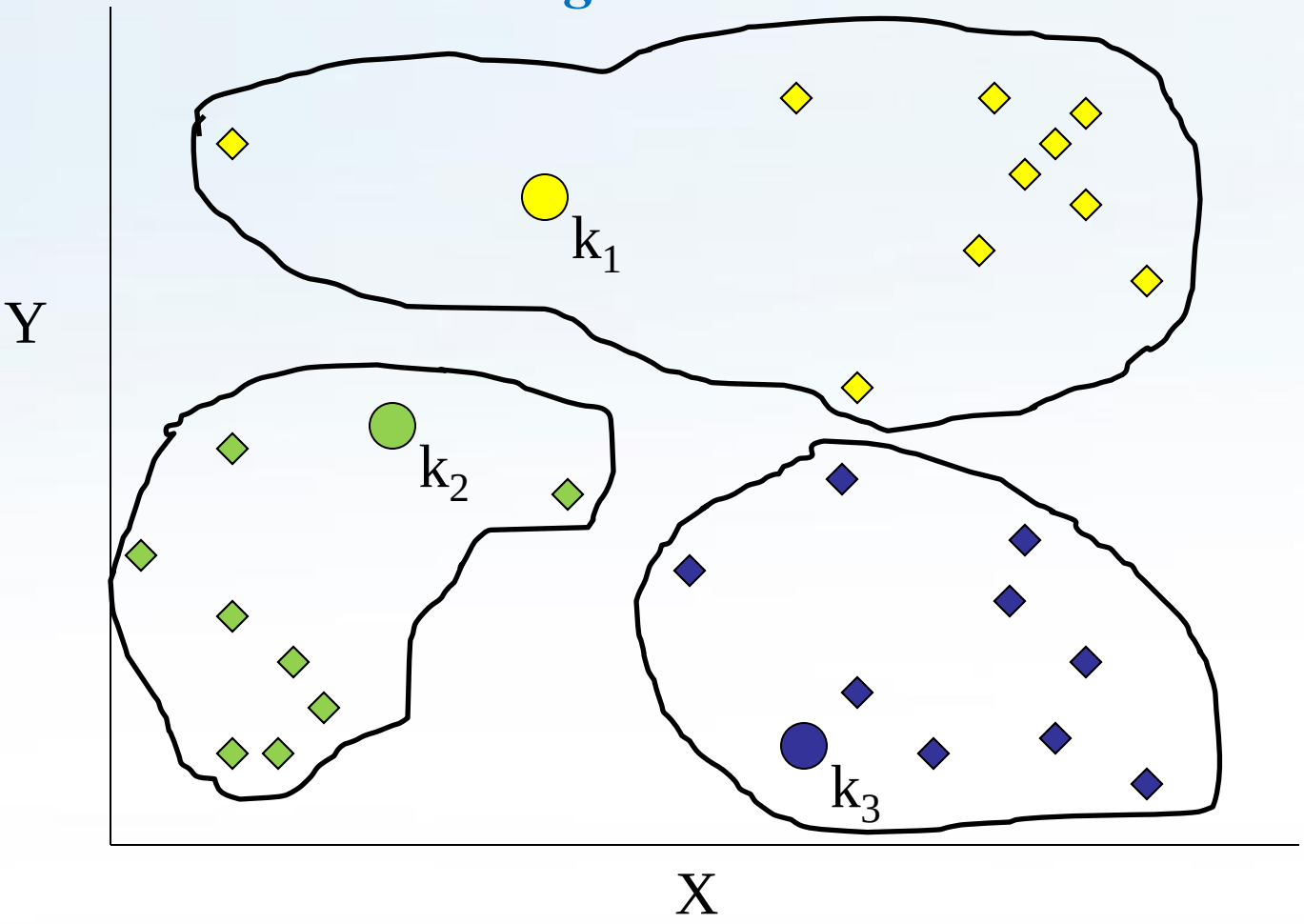
k = 3

B1: Chọn 3 trung tâm nhóm bất kỳ: k_1, k_2, k_3



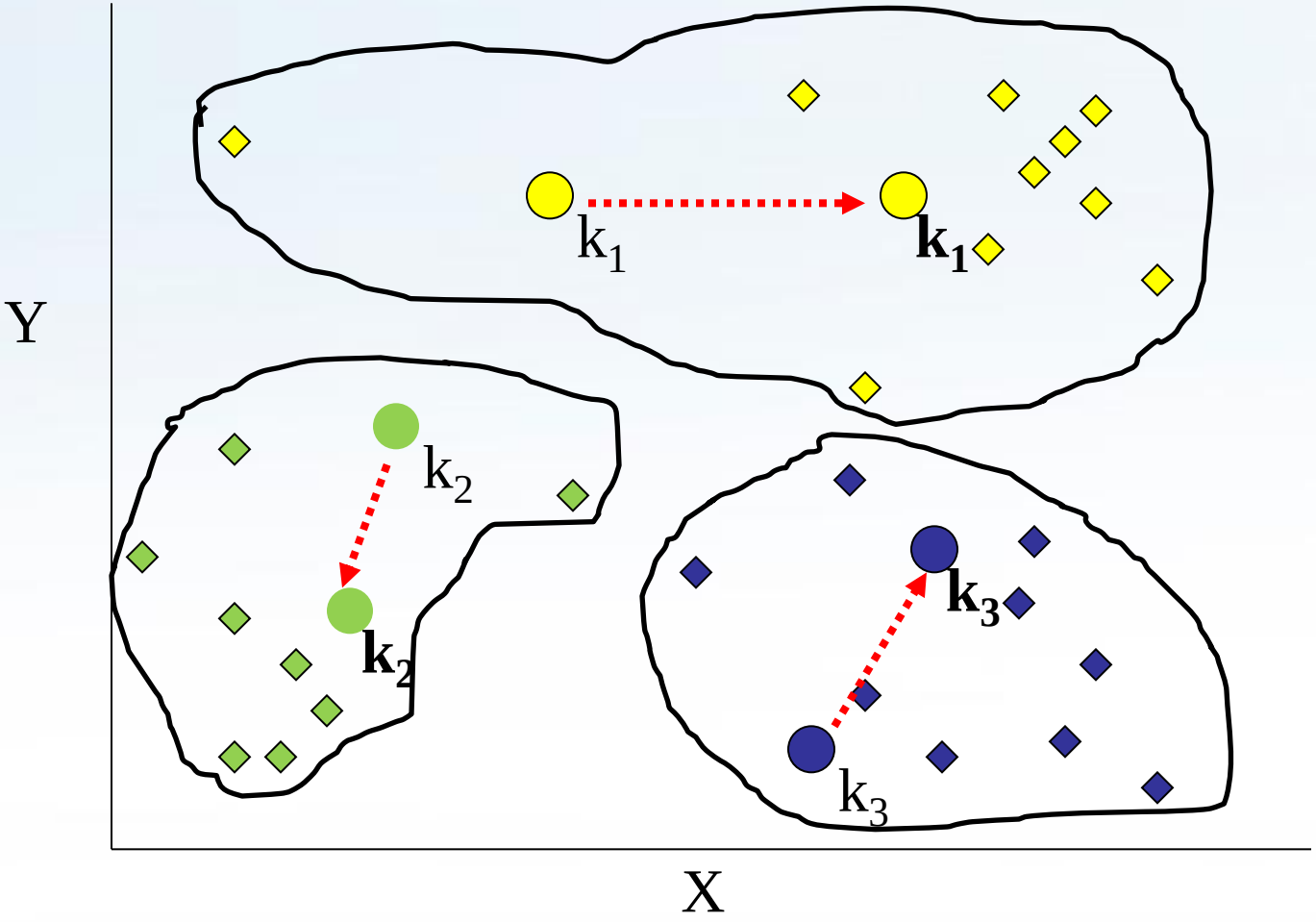
Ví dụ K-means (2/4)

B2: Gán từng điểm vào nhóm có trung tâm nhóm gần nhất



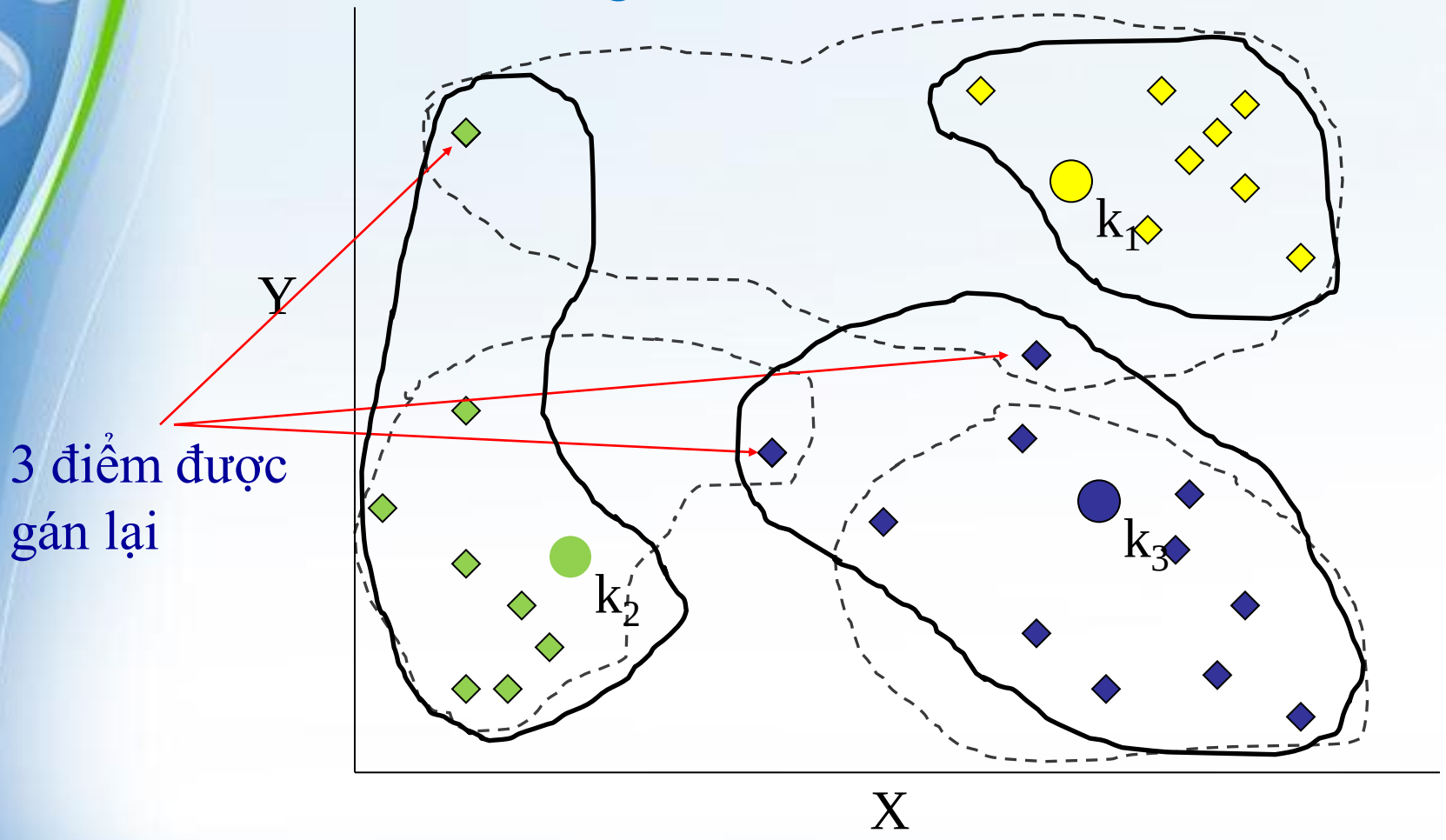
Ví dụ K-means (3/4)

B3: Di chuyển trung tâm từng nhóm về điểm trung bình mới của nhóm



Ví dụ K-means (4/4)

Lặp lại: Gán lại các điểm cho gần với các trung tâm nhóm mới …



Chất lượng của nhóm

- Chất lượng của nhóm C_i có thể được đo bằng độ biến đổi bên trong nhóm hay chính bằng *tổng bình phương sai* khoảng cách của tất cả đối tượng trong nhóm đến trung tâm:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} (dist(p, c_i))^2$$

c_i là trung tâm (centroid) của nhóm C_i

p là đối tượng trong C_i

k là số nhóm

dist là hàm tính khoảng cách

Ví dụ chất lượng nhóm

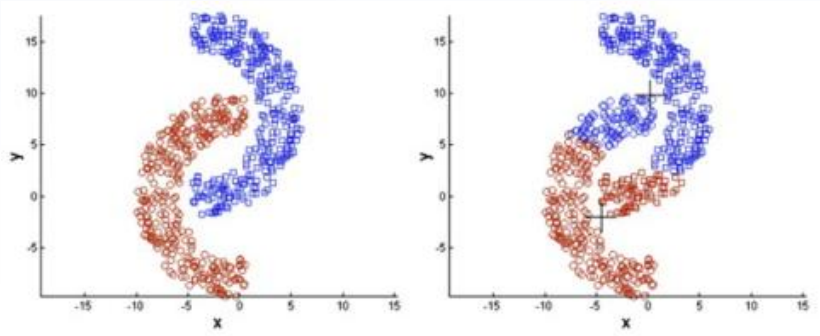
Cho 2 nhóm với các trung tâm tương ứng m_1= 3, m_2=4. Giả sử khoảng cách được tính bằng cách lấy hiệu giá trị hai đối tượng.

- K_1= {2,3};
- K_2= {4,10,12,20,30,11,25}
- SSE = ?

→ $SSE = 1^2+0+0+6^2+8^2+16^2+26^2+7^2+21^2 = 1523$

Nhận xét k-means

- Ưu điểm:
  - Đơn giản, hiệu quả. Độ phức tạp O(tkn); t,k << n.
  - Đạt được tối ưu cục bộ.
- Khuyết điểm:
  - Chỉ áp dụng được với đối tượng trong không gian n-chiều liên tục
  - Cần xác định số nhóm k trước
  - Nhạy cảm với dữ liệu nhiễu và cá biệt
  - Không thích hợp để khám phá các nhóm với hình tròn/cầu



Bài tập 1

Sử dụng thuật toán k-means và khoảng cách Euclide để gom nhóm 8 mẫu vào 3 nhóm:

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Ma trận khoảng cách dựa trên Euclide được cho ở slide sau. Giả sử hạt giống khởi tạo (trung tâm) là k1= A1, k2 = A4 và k3 = A7. Chạy k-means 1 lần

a) Xác định các nhóm được hình thành
b) Trung tâm mới của từng nhóm ở đâu
c) Vẽ trong không gian 10 x 10 các mẫu đã cho kèm với nhóm qua 1 lần chạy (vẽ khung bao) và trung tâm của mỗi nhóm.
d) Bao nhiêu vòng lặp k-means sẽ hội tụ (dừng)? Minh họa kết quả ở mỗi vòng lặp.

Bài tập 1 (tt)

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Bảng khoảng cách Euclide giữa các đối tượng

Bài tập 1 – Đáp án (1/5)

**** Xét A1:***

d(A1, k1)= 0 vì A1 là k1

d(A1, k2)= $\sqrt{13}$

d(A1, k3)= $\sqrt{65}$

→ A1 ∈ nhóm 1

**** Xét A2:***

d(A2,k1)= $\sqrt{25}$ = 5

d(A2, k2)= $\sqrt{18}$ = 4.24

d(A2, k3)= $\sqrt{10}$ = 3.16

→ A2 ∈ nhóm 3

**** Xét A3:***

d(A3, k1)= $\sqrt{36}$ = 6

d(A3, k2)= $\sqrt{25}$ = 5

d(A3, k3)= $\sqrt{53}$ = 7.28

→ A3 ∈ nhóm 2

**** Xét A4:***

d(A4, k1)= $\sqrt{13}$

d(A4, k2)= 0 vì A4 là k2

d(A4, k3)= $\sqrt{52}$

→ A4 ∈ nhóm 2

**** Xét A5:***

d(A5, k1)= $\sqrt{50}$ = 7.07

d(A5, k2)= $\sqrt{13}$ = 3.60

d(A5, k3)= $\sqrt{45}$ = 6.70

→ A5 ∈ nhóm 2

**** Xét A6:***

d(A6, k1)= $\sqrt{52}$ = 7.21

d(A6, k2)= $\sqrt{17}$ = 4.12

d(A6, k3)= $\sqrt{29}$ = 5.38

→ A6 ∈ nhóm 2

Bài tập 1 – Đáp án (2/5)

**** Xét A7:***

d(A7, seed1)= $\sqrt{65}$

d(A7, seed2)= $\sqrt{52}$

d(A7, seed3)= 0 vì A7 là k3

→ A7 ∈ nhóm 3

**** Xét A8:***

d(A8, seed1)= $\sqrt{5}$

d(A8, seed2)= $\sqrt{2}$

d(A8, seed3)= $\sqrt{58}$

→ A8 ∈ nhóm 2

Các nhóm sau 1 lần chạy k-means:

1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

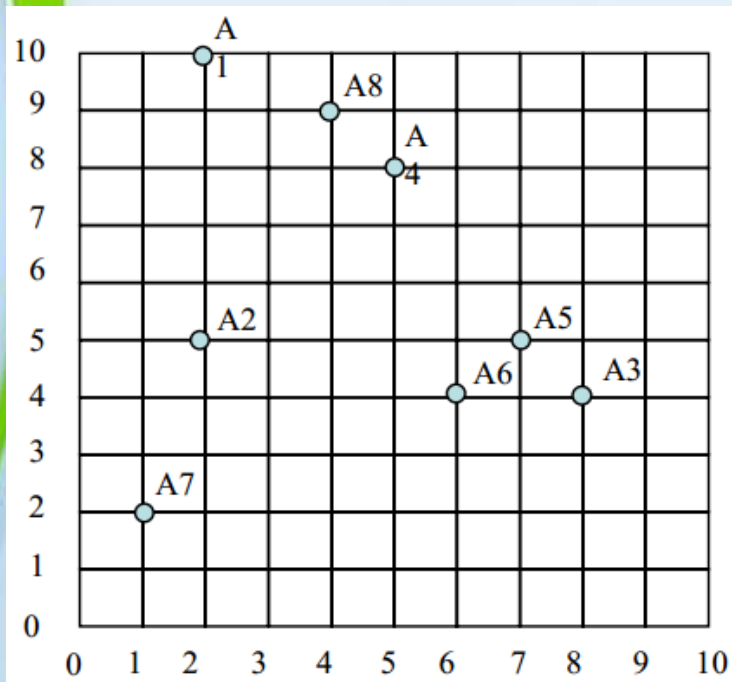
b) Trung tâm các nhóm mới

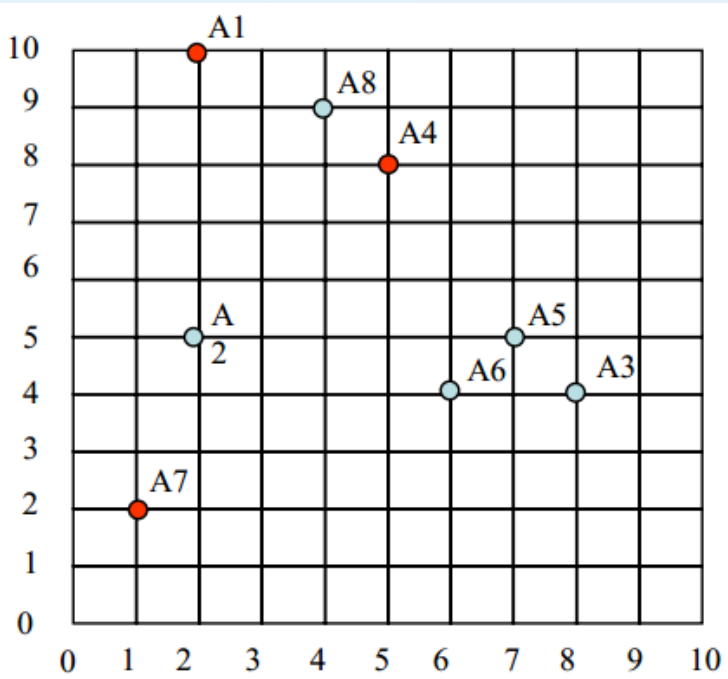
C1= (2, 10)

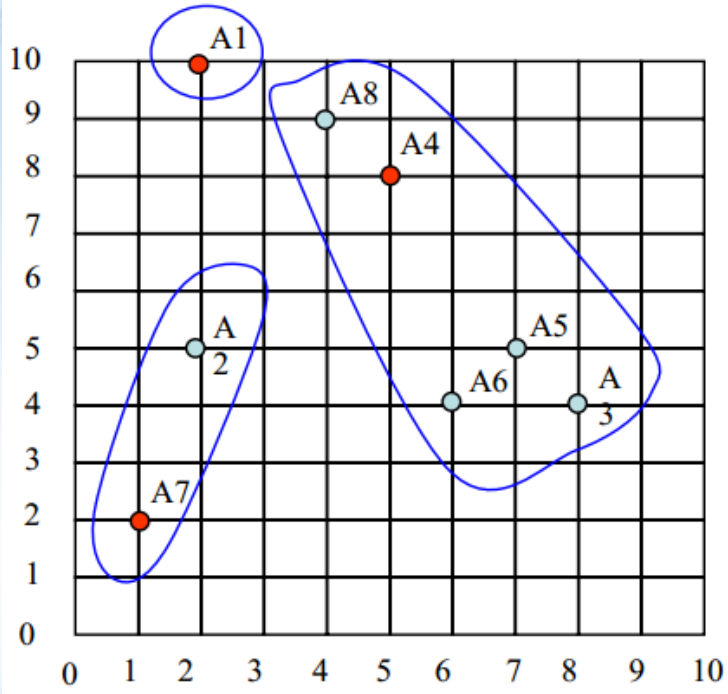
C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5)= (6, 6)

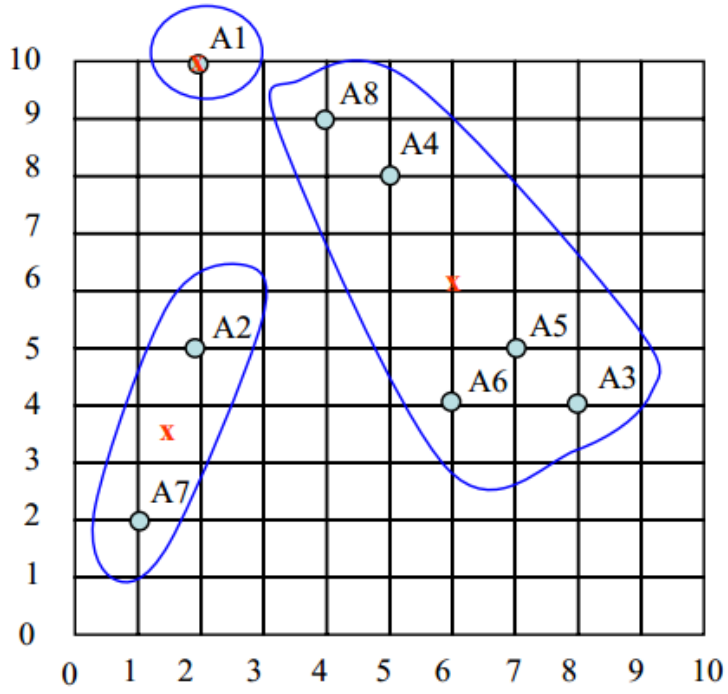
C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

c)


A
1
A8
A
4
A2
A5
A6
A3
A7


A1
A8
A4
A
2
A5
A6
A3
A7


A1
A8
A4
A
2
A5
A6
A
3
A7


A1
A8
A4
A2
A5
A6
A3
A7

Bài tập 1 – Đáp án (4/5)

d) Số vòng lặp cần chạy nữa là 2.

**** Vòng lặp thứ 2:***

1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}

Với các trung tâm:

C1=(3, 9.5), C2=(6.5, 5.25) và C3=(1.5, 3.5).

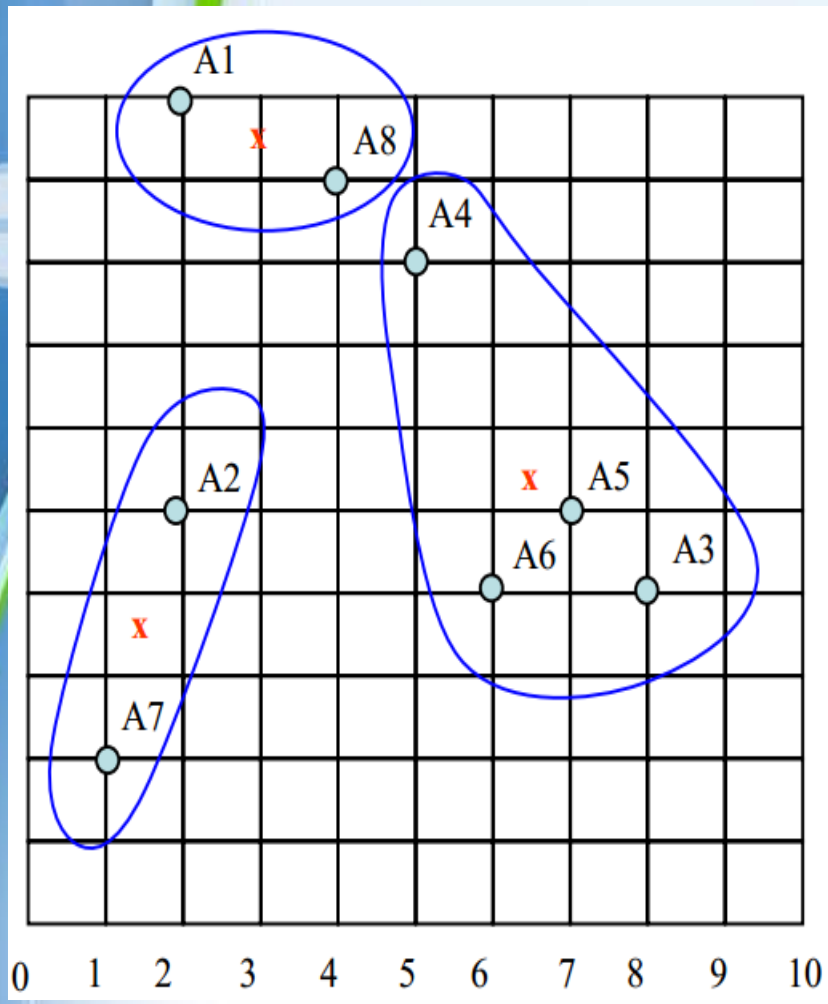
**** Vòng lặp thứ 3:***

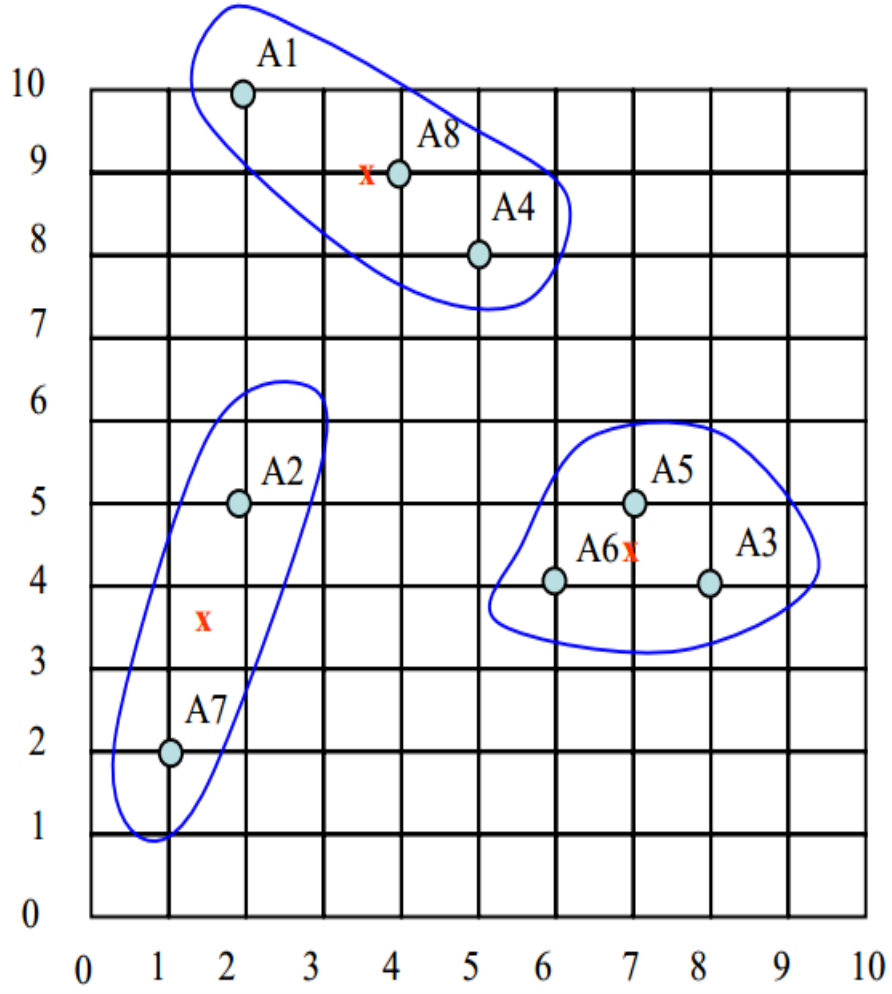
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}

Với các trung tâm:

C1=(3.66, 9), C2=(7, 4.33) và C3=(1.5, 3.5).

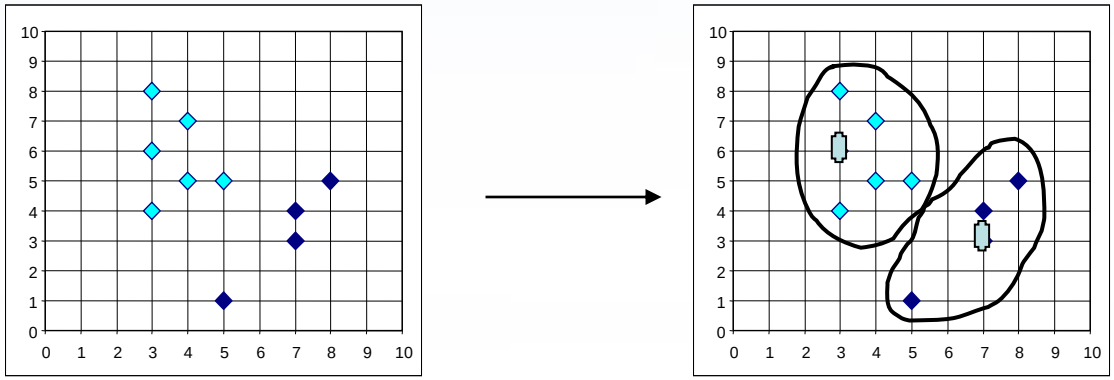
Bài tập 1 – Đáp án (5/5)





k-medoids

- Thuật toán k-means nhạy cảm với nhiễu hay giá trị cá biệt
  - Một đối tượng có giá trị cực lớn có thể làm sai lệch đáng kể phân bố của dữ liệu
- k-medoids: thay vì lấy giá trị trung tâm làm điểm đối chiếu, k-medoids lấy đối tượng nằm ở *vị trí trung tâm* của nhóm



Thuật toán PAM

Cho trước số *k*, mỗi nhóm được thể hiện bằng một trong các đối tượng của nhóm

- ***B1:*** Chọn tùy ý *k* đối tượng làm đại diện cho các nhóm.
- ***B2:*** Gán từng đối tượng còn lại vào nhóm mà nó gần đối tượng đại diện nhất.
- ***B3:*** Chọn ngẫu nhiên một đối tượng không phải là đối tượng đại diện.
- ***B5:*** Tính toán chi phí tổng cộng khi hoán đổi đối tượng ngẫu nhiên với đối tượng đại diện.
- ***B6:*** Nếu chi phí < 0 (nghĩa là độ lỗi giảm hay chất lượng nhóm tăng lên) thì thực hiện chọn đối tượng ngẫu nhiên là đối tượng đại diện mới của nhóm. Tiếp tục thực hiện B2 cho đến khi không còn có thay đổi.

Thuật toán PAM (tt)

Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

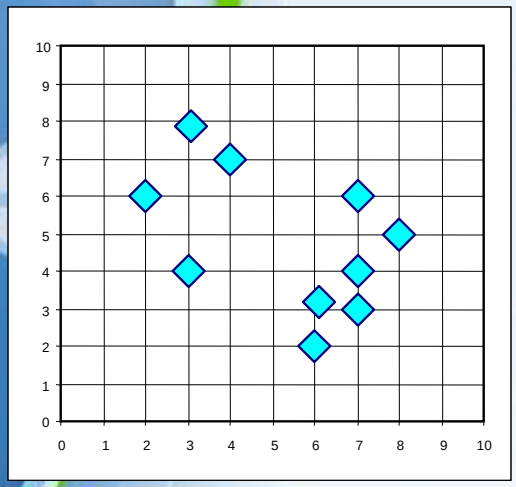
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

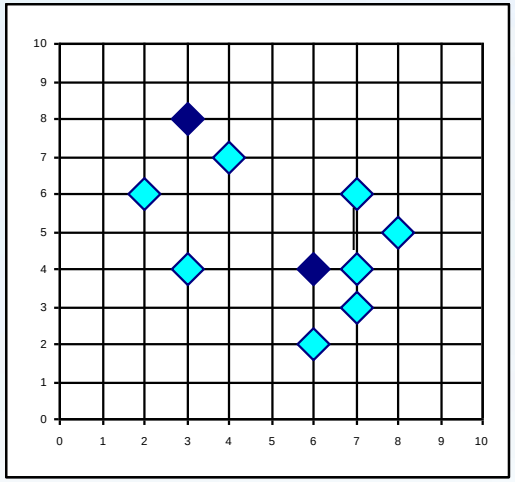
(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

Ví dụ PAM



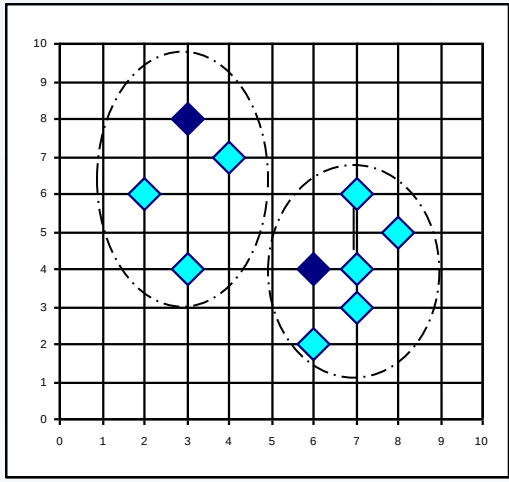
k =2

Chọn tùy ý k đối tượng làm trung vị

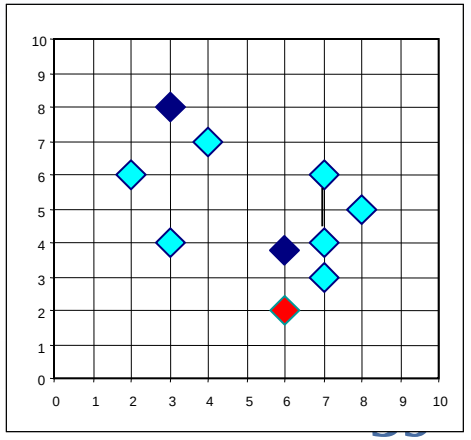


Tổng chi phí = 26

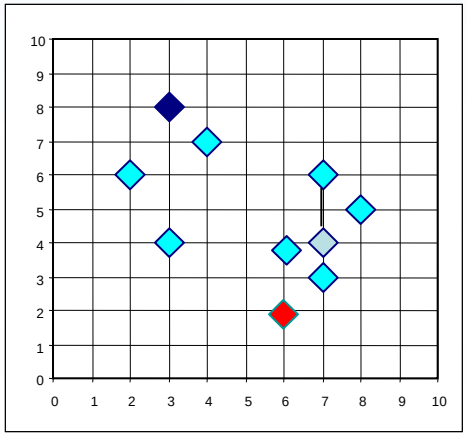
Gán mỗi đối tượng còn lại đến đại diện gần nhất



Chọn ngẫu nhiên đối tượng khác đại diện, O_{ramdom}



Tính toán tổng chi phí khi hoán đổi



Thay thế đối tượng đại diện nhóm bằng O_{ramdom} *nếu chất lượng được cải thiện*

Tổng chi phí = 20

Chất lượng của nhóm

- Chất lượng của nhóm dựa trên tổng độ dị biệt giữa các đối tượng của nhóm hay tổng lỗi tuyệt đối:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)$$

o_i là đối tượng đại diện cho nhóm C_i

p là đối tượng trong C_i

k là số nhóm

dist là hàm tính khoảng cách

Nhận xét k-medoids

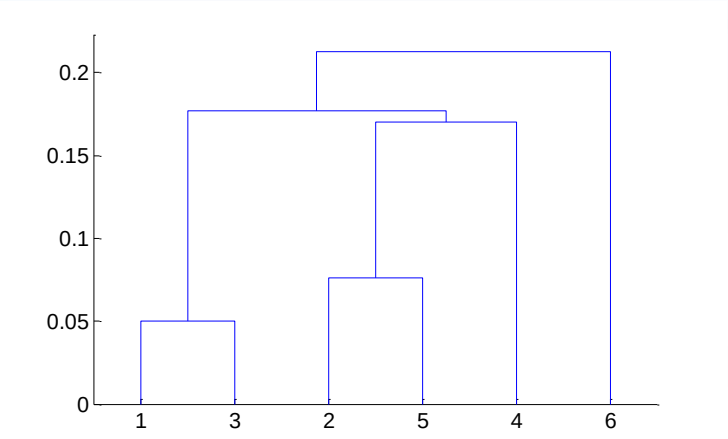
- k-medoids tốt hơn k-means trong việc xử lý nhiễu hay cá biệt.
- Làm việc hiệu quả đối với dữ liệu nhỏ nhưng không hiệu quả trong tập dữ liệu lớn vì độ phức tạp $O(k(n-k)^2)$ (bài toán NP-Hard)
- Một số thuật toán cải tiến PAM
  - CLARA (Kaufmann & Rousseeuw, 1990)
  - *CLARANS* (Ng & Han, 1994)

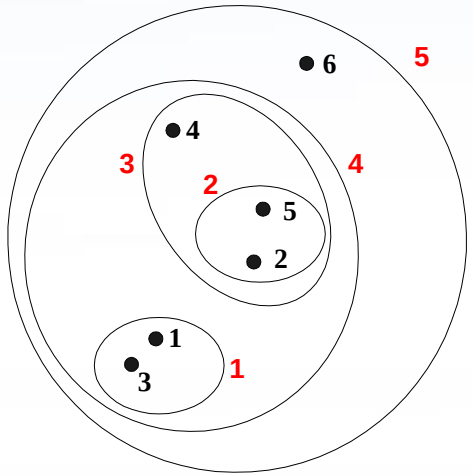
Nội dung

- Khái niệm cơ sở về gom nhóm
- Phương pháp phân hoạch
- **Phương pháp phân cấp**
  - Khái niệm gom nhóm phân cấp
  - Phân loại
  - Thuật toán AGNES
  - Thuật toán DIANA
  - Độ đo khoảng cách nhóm

Gom nhóm phân cấp (1/2)

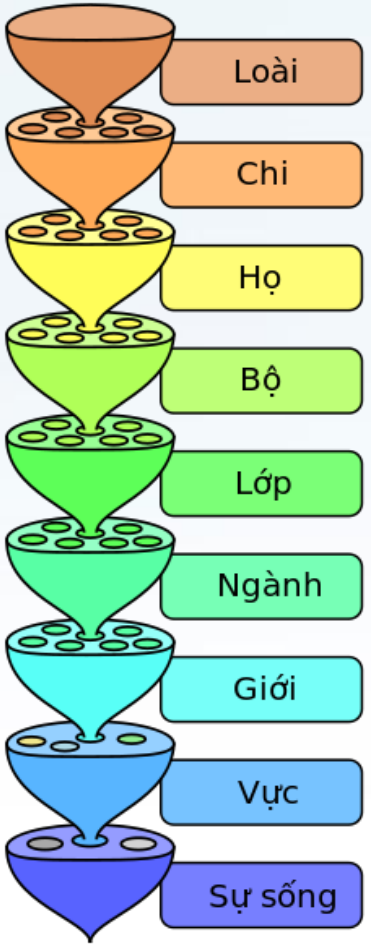
- *Hierachical Clustering* là phương pháp sinh ra tập các nhóm lồng nhau được tổ chức như một cây phân cấp
- Có thể biểu diễn trực quan bằng lược đồ *dendrogram*
  - Là dạng cây giống lược đồ thể hiện quá trình sát nhập hay phân rã





Gom nhóm phân cấp (2/2)

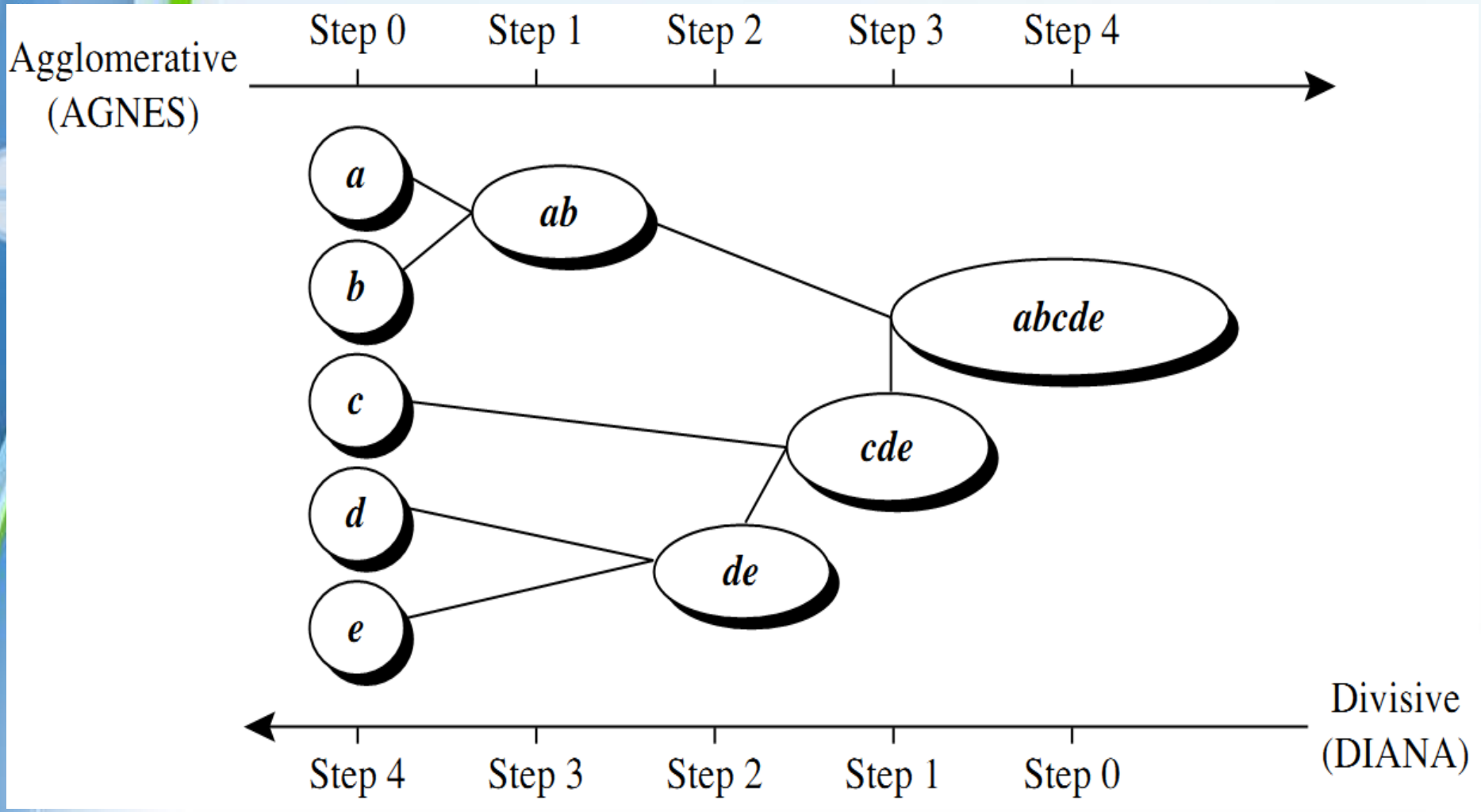
- Gom nhóm phân cấp không cần xác định trước số nhóm *k*.
  - Xác định số nhóm cần thiết bằng việc cắt ngang sơ đồ hình cây tại mức thích hợp.
- Thích hợp với các ứng dụng cần phân tầng ngữ nghĩa
  - Ví dụ trong sinh học có cây về loài.
  - Trong tin tức, phân cấp thể thao>>bóng đá, cầu lông, điền kinh>>…



Loại gom nhóm phân cấp (1/2)

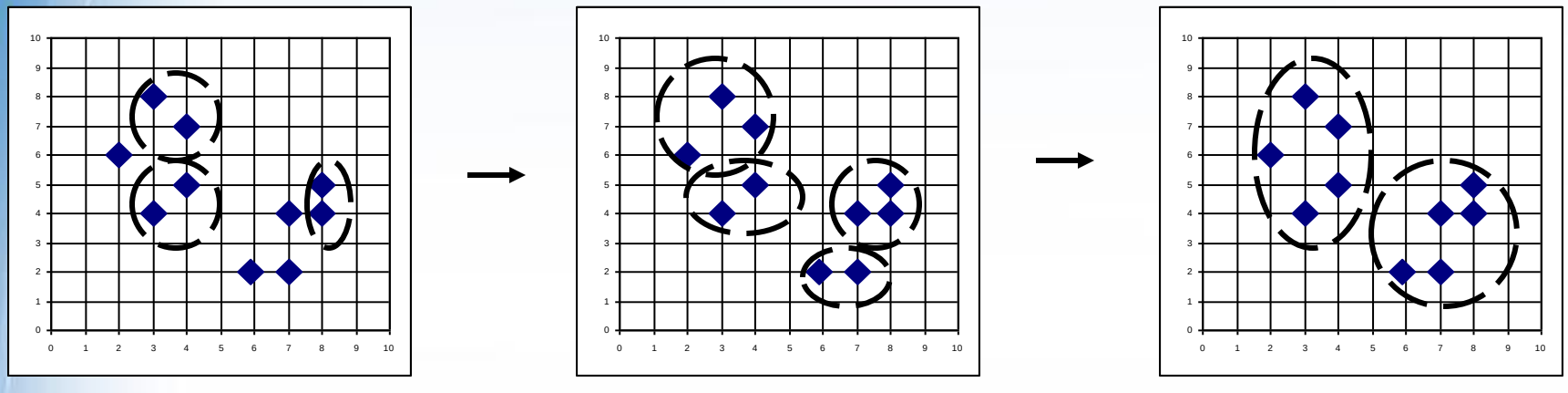
- Có 2 loại gom nhóm phân cấp chính:
  - ***Tích tụ*** (agglomerative): thuật toán AGNES
    - Bắt đầu với mỗi đối tượng như là nhóm riêng biệt.
    - Trong mỗi bước, tiến hành gom cặp nhóm gần nhất cho đến khi chỉ còn 1 nhóm (hay k nhóm) còn lại.
  - ***Phân rã*** (divisive): thuật toán DIANA
    - Nhóm ban đầu bao gồm tất cả đối tượng
    - Trong mỗi bước, tiến hành chia một nhóm cho đến khi mỗi nhóm chứa một đối tượng (hay có k nhóm)

Loại gom nhóm phân cấp (2/2)



Thuật toán AGNES

- AGNES (**AG**glomerative **NES**ting) sử dụng độ đo Single-link và ma trận dị biệt (dissimilarity matrix)
- Bắt đầu từ việc xem mỗi đối tượng là một nhóm và tích tụ các nhóm có độ tương tự cao/dị biệt ít

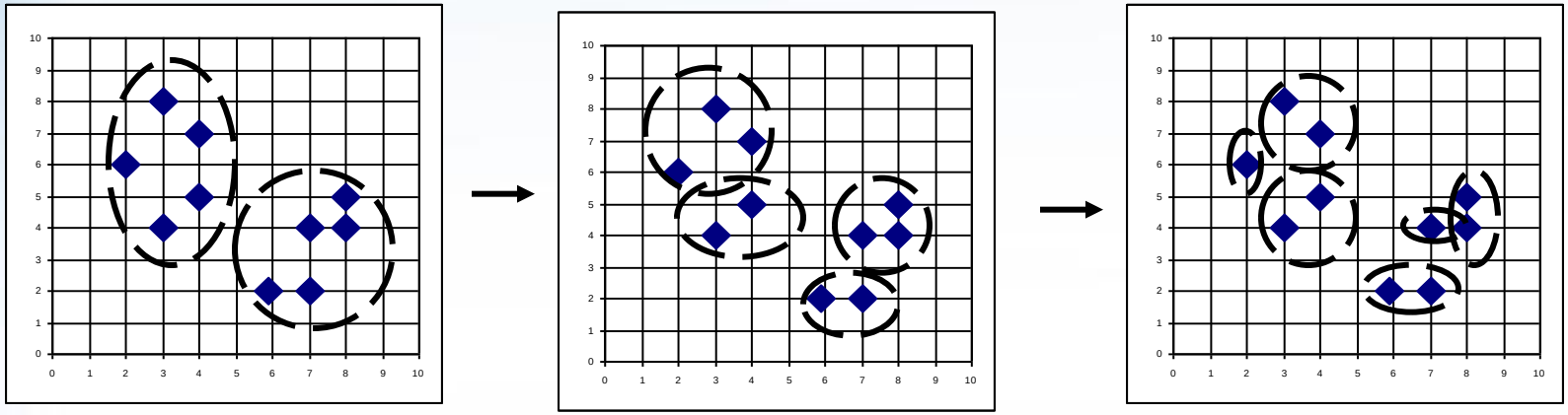


Nhận xét AGNES

- Điểm mạnh:
  - Dễ thực thi
- Điểm yếu:
  - Khó có thể quay lui lại bước trước
  - Không thích hợp cho dữ liệu lớn: độ phức tạp thời gian ít nhất là $O(n^2)$ với n là số đối tượng
- Một số thuật toán cải tiến: BIRCH, CHAMELEON

Thuật toán DIANA

- DIANA (**DI**visive **ANA**lysis) chia các nhóm dựa trên tiêu chí như *khoảng cách Euclide cực đại* giữa các đối tượng láng giềng gần nhất trong nhóm.
- Quá trình ngược lại với AGNES.

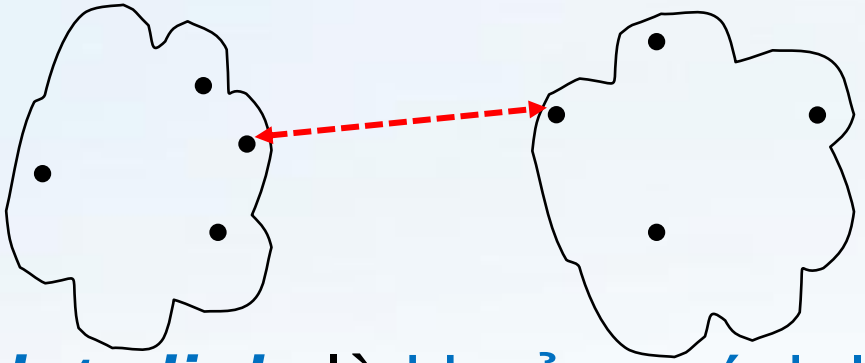


Nhận xét DIANA

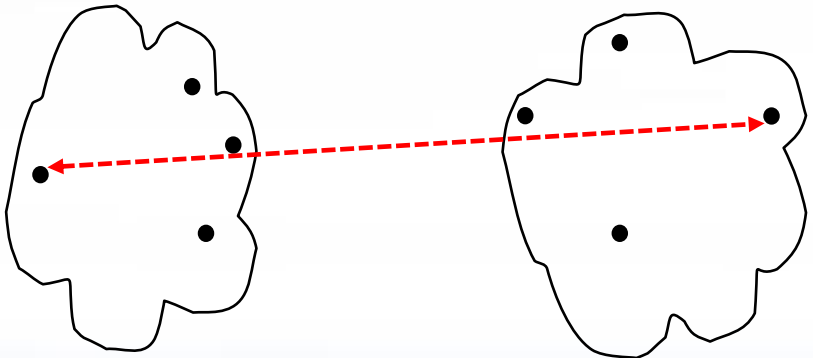
- Điểm mạnh:
  - Có thể nhanh đạt đến số nhóm mong muốn nếu số nhóm nhỏ.
- Điểm yếu:
  - Khó khăn trong việc chia nhóm thành các nhóm nhỏ hơn.
    - Ví dụ: có $2^{n-1} - 1$ khả năng để chia tập n đối tượng thành 2 tập không giao nhau.
  - Sử dụng heuristic để chia dễ dẫn đến kết quả không chính xác
  - Khó có thể quay lại bước trước.

Độ đo khoảng cách

- ***Single link***: là khoảng cách nhỏ nhất giữa một đối tượng trong nhóm này đến một đối tượng trong nhóm khác.

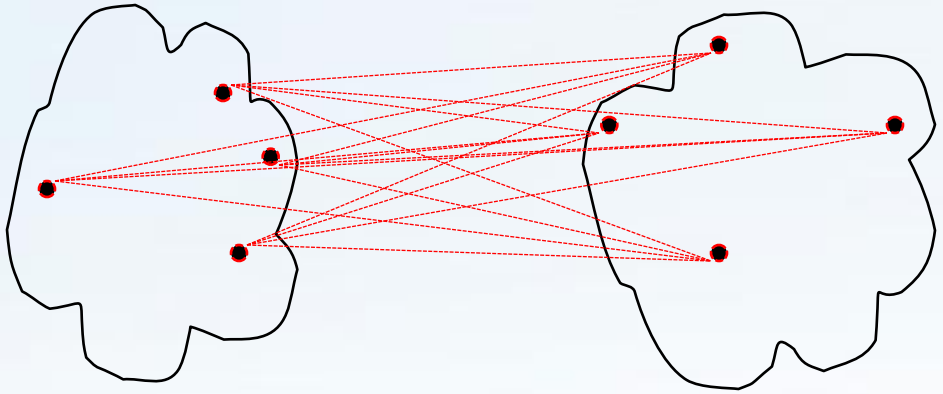


- ***Complete link***: là khoảng cách lớn nhất giữa một đối tượng trong nhóm này đến một đối tượng trong nhóm khác.

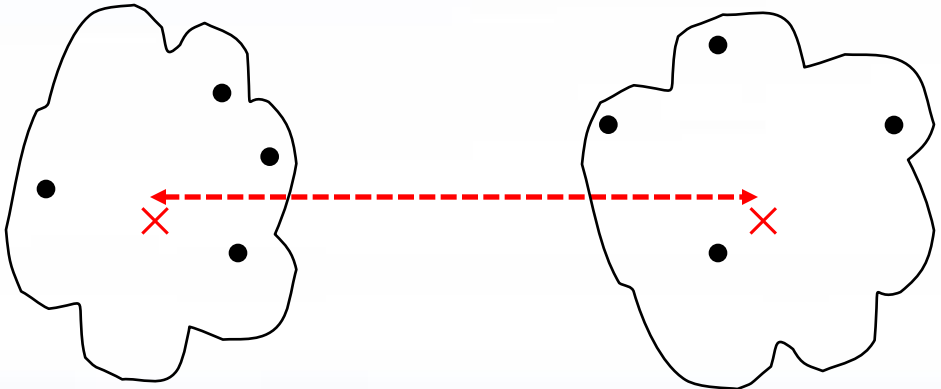


Độ đo khoảng cách (tt)

- ***Average***: là *khoảng cách trung bình* giữa các đối tượng trong nhóm này đến các đối tượng trong nhóm khác.

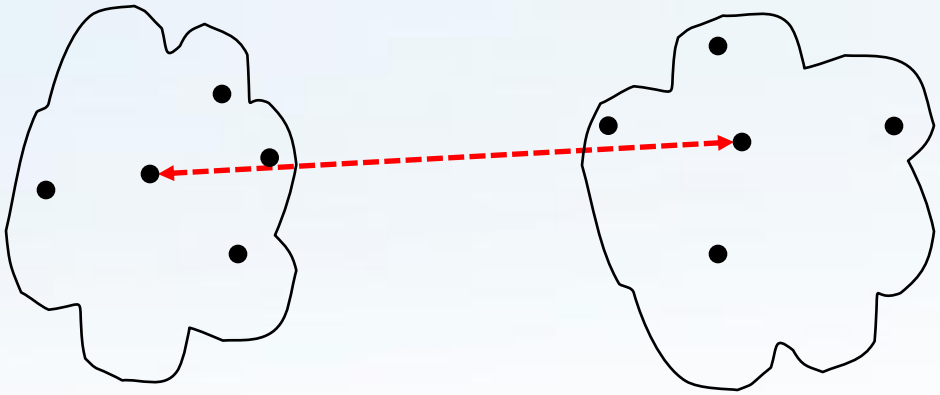


- ***Centroid***: là *khoảng cách giữa 2 centroid* của hai nhóm



Độ đo khoảng cách (tt)

- ***Medoid***: là *khoảng cách giữa 2 trung vị* của 2 nhóm.



- Một số phương pháp khác như phương pháp của Ward sử dụng bình phương sai

Các phương pháp gom nhóm khác

- Đọc thêm trong phần 10.3.3, 10.3.4, 10.3.5 cuốn Data Mining: Concepts and Techniques (3rd Edition, J.Han) để nắm thêm một số thuật toán gom nhóm như BIRCH, Chameleon, gom nhóm phân cấp xác suất.

Bài tập 2

Cho tập DL gồm 6 điểm trong không gian 2 chiều. Sử dụng thuật toán AGNES với Single link để gom nhóm

Điểm	Tọa độ x	Tọa độ y
P1	0.40	0.53
P2	0.22	0.38
P3	0.353	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Bài tập 2 – Đáp án

Xây dựng ma trận khoảng cách (độ đo Euclide) giữa các điểm

	P1	P2	P3	P4	P5	P6
P1	0.00	0.23	0.22	0.37	0.34	0.24
P2	0.23	0.00	0.15	0.19	**0.14**	0.24
P3	0.22	0.15	0.00	0.16	0.29	**0.10**
P4	0.37	0.19	0.16	0.00	0.28	0.22
P5	0.34	**0.14**	0.29	0.28	0.00	0.39
P6	0.24	0.24	**0.10**	0.22	0.39	0.00

Bài tập 2 – Đáp án (tt)

B1: mỗi điểm là một nhóm

B2:

Trong số các nhóm gồm một điểm thì dist(3,6) là nhỏ nhất nên gộp điểm P3 và P6 với nhau thành một nhóm

Thu được các nhóm : {1}, {4}, {2}, {5}, {3,6}.

* Quay lại B2 do chưa thu được nhóm “toàn bộ”

Tính khoảng cách giữa các nhóm . Ví dụ :

Dist({3,6},{1})=min(dist(3,1),dist(6,1))

=min(0.22, 0.24)

= 0.22

Bài tập 2 – Đáp án (tt)

dist(2,5) là nhỏ nhất nên gộp P2 và P5.

→ Ta có các nhóm sau : {1}, {4}, {3,6}, {2,5}

* Tính khoảng cách giữa các nhóm. Ví dụ :

dist({3,6},{2,5}) =

= min(dist(3,2),dist(6,2),dist(3,5),dist(6,5))

= min(0.15, 0.24, 0.28, 0.39)

= 0.15 …

dist({3,6},{2,5}) nhỏ nhất nên gộp các nhóm {3,6}, {2,5} thành một nhóm.

→ Ta thu được các nhóm : {1},{4},{2,3,5,6}

Bài tập 2 – Đáp án (tt)

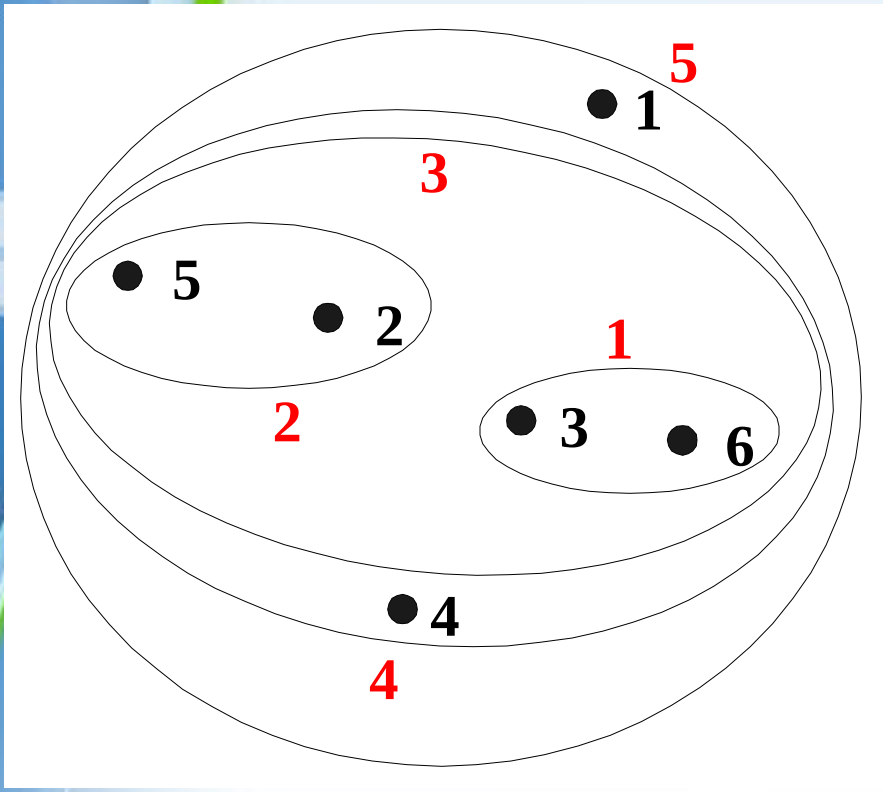
* Tiếp tục tính khoảng cách giữa các nhóm.

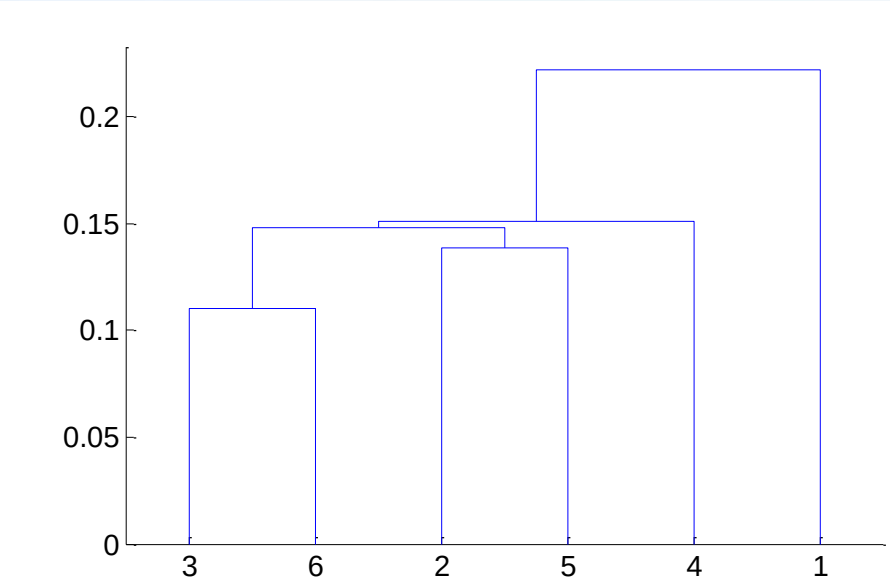
….

→ Gộp {4} với {2,3,5,6} thu được các nhóm {1}, {2,3,4,5,6}

→ Gộp 2 nhóm này ta thu được nhóm “toàn bộ” và thuật toán dừng

Bài tập 2 – Đáp án (tt)





Các nhóm

(Single Link)

Sơ đồ hình cây

Bài tập 3

Cho tập DL gồm 6 điểm trong không gian 2 chiều với ma trận khoảng cách như sau.

	P1	P2	P3	P4	P5	P6
P1	0.00	0.23	0.22	0.37	0.34	0.24
P2	0.23	0.00	0.15	0.19	0.14	0.24
P3	0.22	0.15	0.00	0.16	0.29	0.10
P4	0.37	0.19	0.16	0.00	0.28	0.22
P5	0.34	0.14	0.29	0.28	0.00	0.39
P6	0.24	0.24	0.10	0.22	0.39	0.00

Sử dụng thuật toán AGNES với Complete link để gom nhóm. Vẽ sơ đồ hình cây.

Bài tập 4

Cho 2 đối tượng: (22,1,42,10) và (20,0,36,8)

a) Tính khoảng cách Euclide giữa 2 đối tượng
b) Tính khoảng cách Mahanttan giữa 2 đối tượng
c) Tính khoảng cách Minkowski giữa 2 đối tượng với q=3

Tham khảo công thức trong slide phụ lục

Tóm tắt

- Gom nhóm là quá trình nhóm các đối tượng thành những cụm sao cho các đối tượng cùng nhóm có độ tương tự cao và rất khác với đối tượng ở các nhóm còn lại.
- Gom nhóm phân hoạch là pp đơn giản và nền tảng nhất trong số các pp gom nhóm dựa trên việc phân hoạch một cơ sở dữ liệu D thành k nhóm (với k cho trước) sao cho tối ưu tiêu chí phân hoạch. Hai thuật toán điển hình là k-means và k-medoids.
- Gom nhóm phân cấp là phương pháp sinh ra tập các nhóm lồng nhau được tổ chức như một cây phân cấp. Có hai cách để phân cấp: phân cấp kiểu tích tụ điển hình là thuật toán AGNES và phân cấp kiểu phân rã điển hình là thuật toán DIANA.

Tài liệu tham khảo

1. J.Han, M.Kamber, Chương 10 – Cluster Analysis:Basic Concepts and Methods và Chương 11 – Advanced Cluster Analysis, cuốn "*Data mining: Basic Concepts and Methods*", 3rd edition
2. J.Han, M.Kamber, J.Pei, Chapter 10, www.cs.uiuc.edu/homes/hanj/cs412/bk3_slides/10ClusBasic.ppt
3. Tan, Steinbach, Kumar, Lecture Notes for Chapter 8 Introduction to Data Mining, http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap8_basic_cluster_analysis.pdf

Hỏi & Đáp

Distance Measurements

Euclidean Distance

$$d_{ij} = \left[\sum_{k} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

This is usually applied to standardized data to give the same weight to all the metrics (except when the input is the Principal Components)

Ref http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/i290_280I_Lecture_6a.ppt

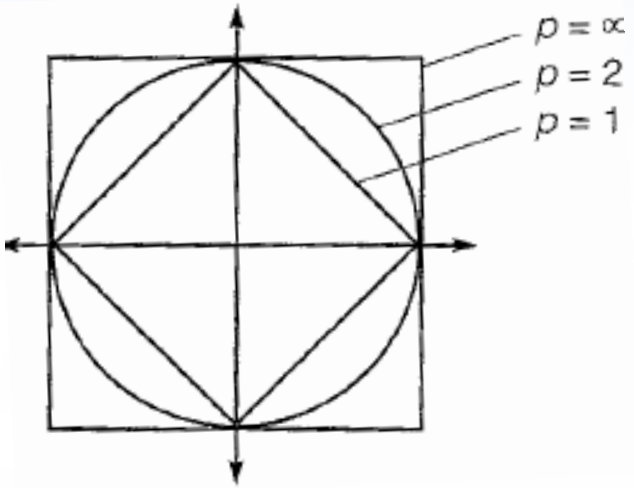
Distance Measurements

Ref http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/i290_280I_Lecture_6a.ppt

Minkowski Metric

$$d_{ij} = \left[\sum_{k} \left| x_{ik} - x_{jk} \right|^{p} \right]$$

- The Euclidean distance is a especial case when p=2
- When p=1 it is called city-block metric because it is like walking from point A to point B in a city laid out with a grid of streets



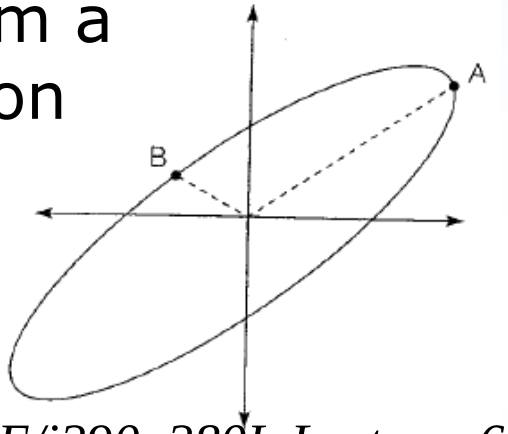
Mahalanobis Distance

This distance takes into account the covariance patterns of the data

$$D_{ij}^2 = (x_i - x_j)'\Sigma^{-1}(x_i - x_j)$$

Where Σ is the population covariance matrix of the data matrix X

The Mahalanobis distance captures the fact that a point A and a point B are equally likely to have been drawn from a multivariate normal distribution



Ref http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/i290_280I_Lecture_6a.ppt