Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Giải thuật gom cụm
Clustering algorithms

Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

Nội dung

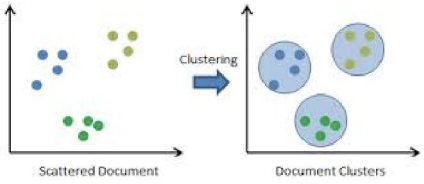
- **Giới thiệu về clustering**
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

3

Clustering

- **Giới thiệu về clustering**
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- Gom nhóm-cụm/clustering
  - Gom nhóm: mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất **tương tự nhau** và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau



  - Phương pháp học không giám sát
  - Dữ liệu thường không có nhiều thông tin sẵn có như lớp (nhãn)

4

Một số ứng dụng của phương pháp clustering

Phương pháp Clustering được sử dụng rộng rãi trong nhiều ứng dụng như nghiên cứu thị trường, tìm kiếm thông tin, phân tích dữ liệu, và xử lý hình ảnh

- Có thể giúp các nhà tiếp thị khám phá các nhóm khách hàng riêng biệt. Và họ có thể đặc trưng nhóm khách hàng của họ dựa trên các lịch sử mua hàng.
- Trong lĩnh vực sinh học, clustering được sử dụng để phân loại thực vật và động vật, phân loại gen có chức năng tương tự
- Clustering cũng giúp trong việc phân loại tài liệu trên web để phát hiện thông tin.

Một số ứng dụng của phương pháp clustering

- Clustering cũng được sử dụng trong các ứng dụng phát hiện outlier như phát hiện các gian lận thẻ tín dụng.
- Bảo hiểm: Xác định các nhóm chính sách bảo hiểm xe máy. Chủ sở hữu được bồi thường chi phí trung bình, cao, thấp khác nhau tuỳ đối tượng.
- Clustering cũng giúp trong việc xác định các khu vực sử dụng đất tương tự trong một cơ sở dữ liệu quan sát trái đất. Nó cũng giúp trong việc xác định các nhóm nhà ở một thành phố theo kiểu nhà, giá trị, và vị trí địa lý.

Clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- có nhiều nhóm giải thuật khác nhau
  - **hierarchical clustering,**
  - **K-Means (Partitional clustering),**
  - Dendrogram,
  - SOM, EM,…

7

Clustering

- gom nhóm
  - thường dựa trên cơ sở **khoảng cách**
  - nên chuẩn hóa dữ liệu
  - khoảng cách được tính theo từng kiểu của dữ liệu
    - Kiểu số,
    - Kiểu nhị phân
    - Kiểu rời rạc (nominal type),

Gom nhóm: mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất **tương tự nhau** và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau

8

Các độ đo khoảng cách - Kiểu số

- Khoảng cách *Minkowski*

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

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ là 2 phần tử dữ liệu trong p-dimensional, q là số nguyên dương

- nếu $q = 1$, d là khoảng cách Manhattan

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

- nếu $q = 2$, d là khoảng cách Euclid

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

9

Kiểu rời rạc (nominal type)

- VD: thuộc tính color có giá trị là red, green, blue, etc.
  - phương pháp matching đơn giản,
    - m là số lượng matches và
    - p là tổng số biến (thuộc tính),
    - khoảng cách được định nghĩa :

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

10

Kiểu rời rạc (nominal type)

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

- m là số lượng matches và
- p là tổng số biến (thuộc tính),

	Màu tóc	Màu mắt	Chiều cao	Cân nặng	Trình độ
Nam	Đen	Đen	Cao	Trung bình	Cao đẳng
Lan	Nâu	Đen	Thấp	Trung bình	Đại học

d(Nam, Lan) =?

11

Các độ đo khoảng cách - Kiểu nhị phân

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

- khoảng cách đối xứng : $d(i,j)=\frac{b+c}{a+b+c+d}$
- khoảng cách bất đối xứng : $d(i,j)=\frac{b+c}{a+b+c}$
- hệ số Jaccard bất đối xứng : $sim_{Jaccard}(i,j)=\frac{a}{a+b+c}$

12

Kiểu nhị phân

- **Binary variables/attributes**
  - **Ví dụ**

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

  - **gender: symmetric**
  - **Binary attributes còn lại: asymmetric**
  - **Y, P → 1, N → 0**

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

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

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

13

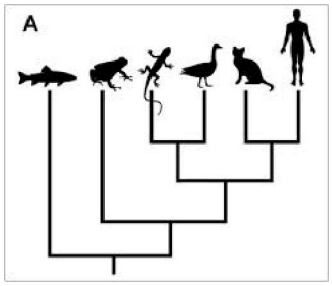
Nội dung

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

14

Hierarchical Clustering

- Xây dựng một cây phân cấp dựa trên sự phân loại theo cấp bậc từ một tập hợp các dữ liệu

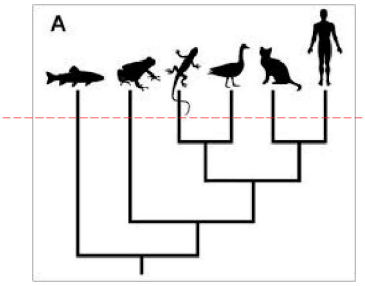


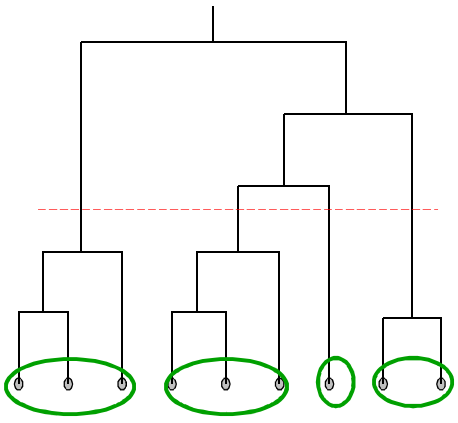
- Dựa trên điểm cắt ở đâu mà ta thu được các cụm tương ứng

15

Hierarchical Clustering

- Dựa trên điểm cắt ở đâu mà ta thu được các cụm tương ứng





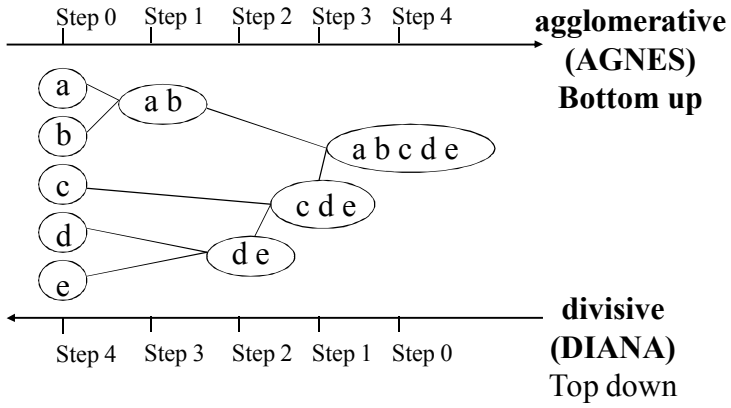
16

Hierarchical clustering

- bottom up
  - bắt đầu với những clusters chỉ là 1 phần tử
  - ở mỗi bước, merge 2 clusters gần nhau thành 1
  - khoảng cách giữa 2 clusters : 2 điểm gần nhất từ 2 clusters, hoặc khoảng cách trung bình, etc.
- top down
  - bắt đầu với 1 cluster là tất cả dữ liệu
  - tìm 2 clusters con
  - tiếp tục đệ quy trên 2 clusters con
- kết quả sinh ra dendrogram

17

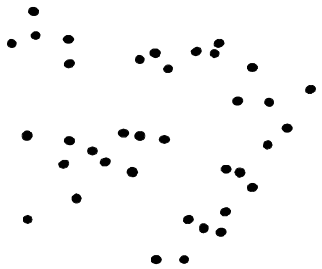
Hierarchical clustering



18

Hierarchical clustering (Single link)

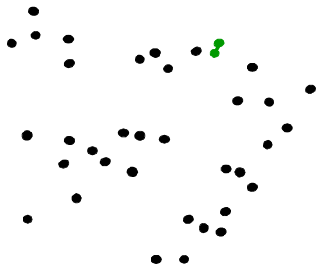
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

Hierarchical clustering (Single link)

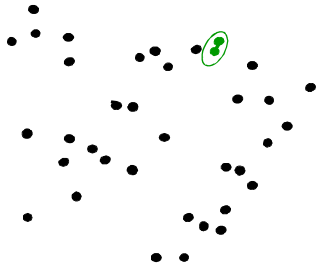
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

② Tìm "khoảng cách" tương tự nhất giữa các cặp cụm

Hierarchical clustering (Single link)



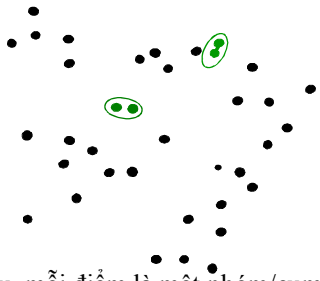
① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

② Tìm "khoảng cách" tương tự nhất giữa các cặp cụm

③ Kết hợp từng 2 cặp điểm thành một cụm mẹ/cụm lớn hơn

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)



① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

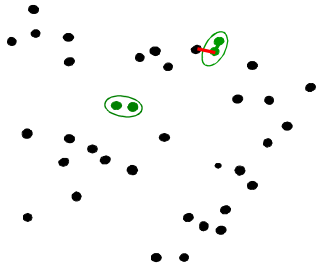
② Tìm "khoảng cách" tương tự nhất giữa các cặp cụm

③ Kết hợp từng 2 cặp điểm thành một cụm mẹ/cụm lớn hơn

④ **Lặp lại…**

Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

24

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)

25

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)

26

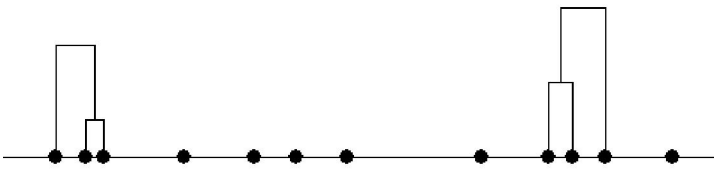
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)

27

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

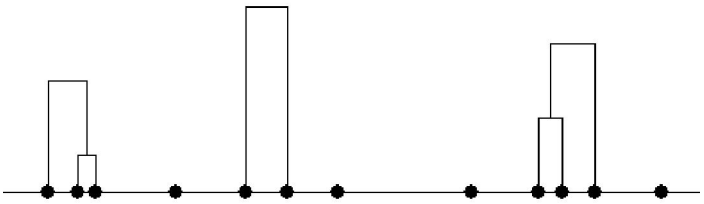
Hierarchical clustering (Single link)



28

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

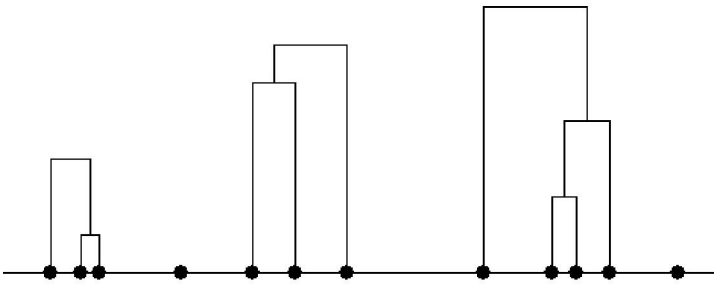
Hierarchical clustering (Single link)



29

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

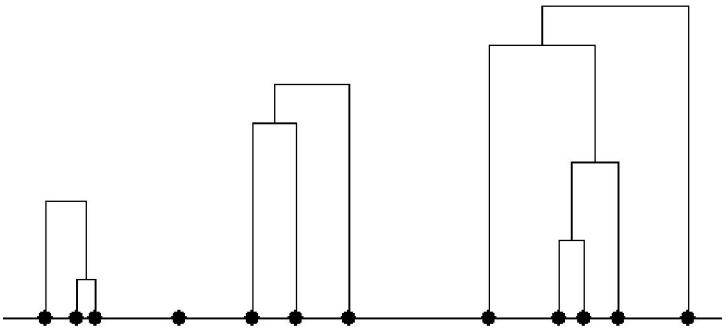
Hierarchical clustering (Single link)



30

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

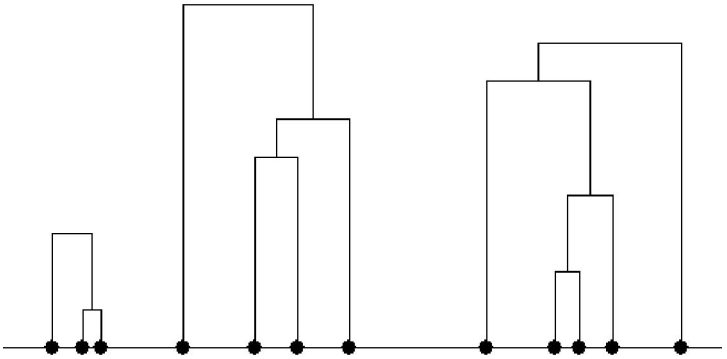
Hierarchical clustering (Single link)



31

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

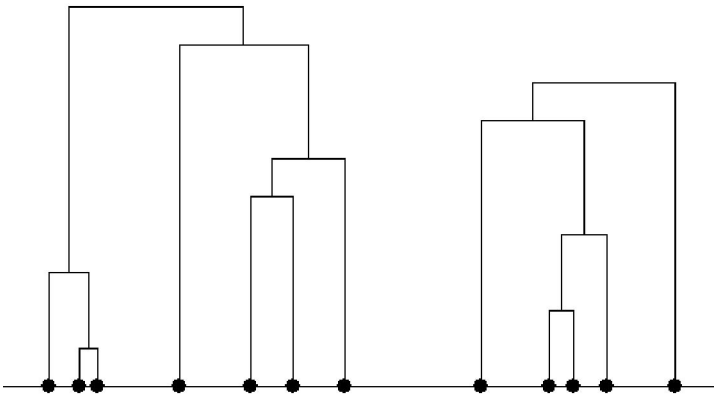
Hierarchical clustering (Single link)



32

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

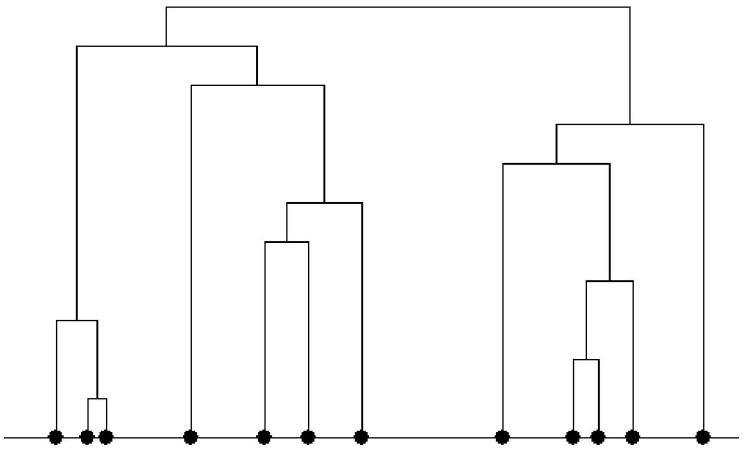
Hierarchical clustering (Single link)



33

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

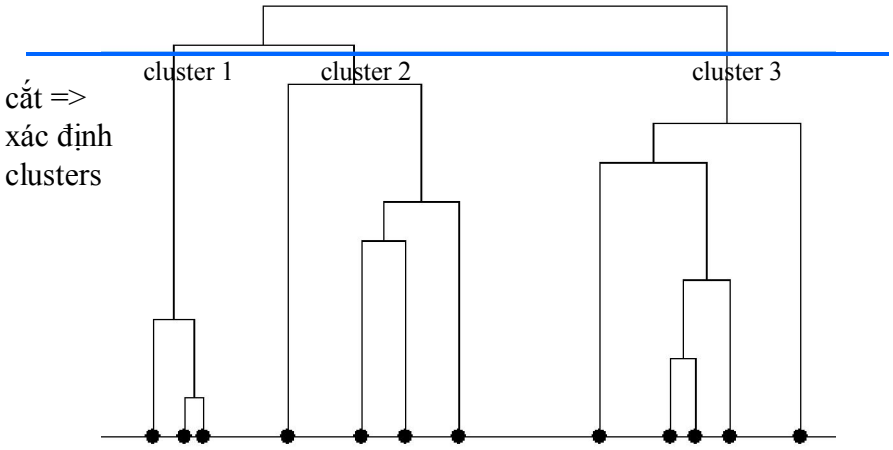
Hierarchical clustering (Single link)



34

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

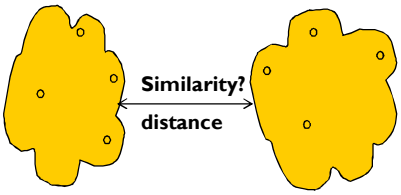
Hierarchical clustering (Single link)

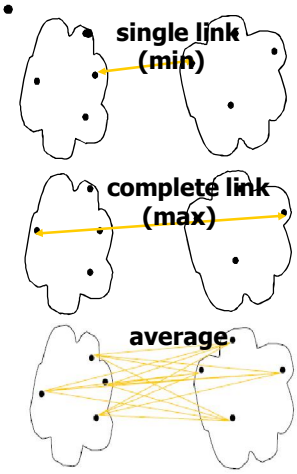


cắt => xác định clusters

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

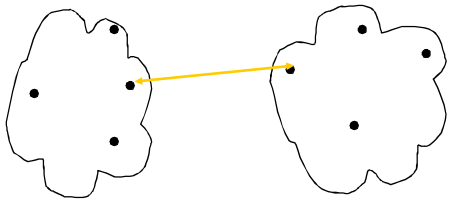
Hierarchical clustering





- Định nghĩa khoảng cách, độ tương tự của 2 nhóm
  - MIN
  - MAX
  - Group Average

Hierarchical clustering

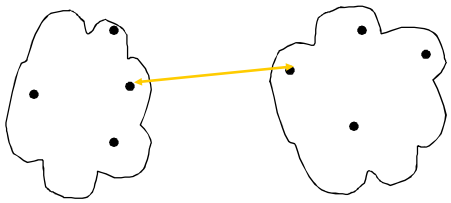


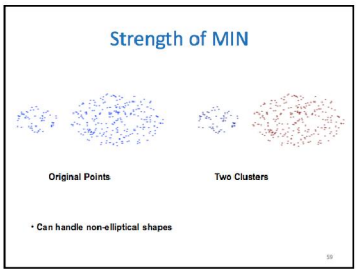
- **distance = shortest distance, nearest neighbor clustering algorithm**
  - **MIN Linkage**
  - **MAX Linkage**
  - **Group Average**

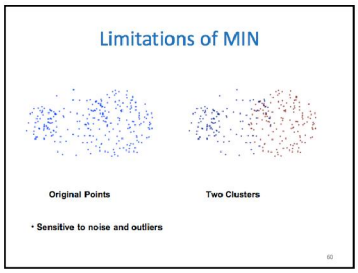
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering

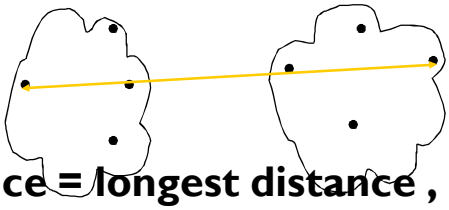
- **MIN Linkage**
- **distance = shortest distance, nearest neighbor clustering algorithm**







Hierarchical clustering

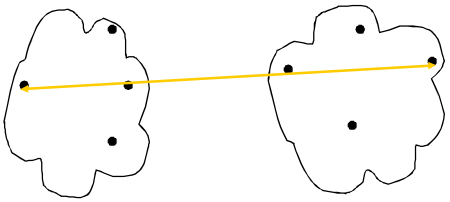


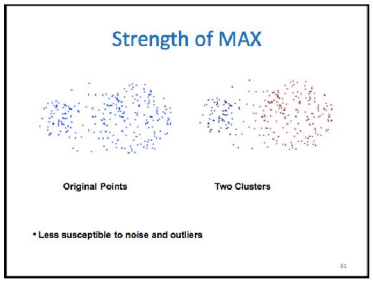
- **distance = longest distance , farthest neighbor clustering algorithm**
  - **MIN Linkage**
  - **MAX Linkage**
  - **Group Average Linkage**

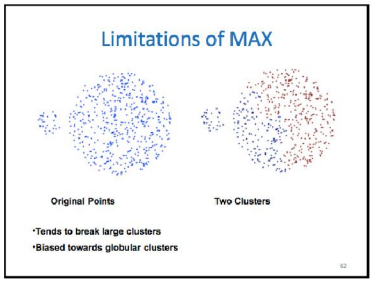
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering

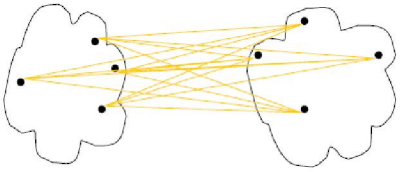
- **MAX Linkage**
- **distance = longest distance , farthest neighbor clustering algorithm**







Hierarchical clustering



***average-link* clustering, distance = average distance**

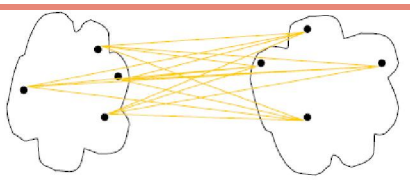
- **MIN Linkage**
- **MAX Linkage**
- **Group Average Linkage**

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering

Group Average Linkage

***average-link* clustering, distance = average distance**



- **Ít nhạy cảm với nhiễu và outliers**

Hierarchical clustering (Single link)

- **Bài tập ví dụ**

Sử dụng phương pháp Hierarchical clustering (Single link) để gom nhóm một số thành phố của Ý dựa vào khoảng cách giữa các thành phố này

Hierarchical clustering (Single link)

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	**138**
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Bước 1:

Cặp vực thành phố gần nhau nhất là MI và TO, ở khoảng cách 138 Chúng được sáp nhập vào một cụm duy nhất được gọi là "MI / TO".

Mức độ cluster mới là L (MI / TO) = 138 và số thứ tự mới là m = 1 thì ta tính khoảng cách từ đối tượng hợp chất mới này cho tất cả các đối tượng khác.

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

Bước 1

Nguyên tắc trong Hierarchical clustering (Single link): khoảng cách từ cụm/nhóm đối tượng mới tạo đến các đối tượng khác bằng với khoảng cách ngắn nhất từ các thành viên của cụm/nhóm đến các đối tượng bên ngoài. Vì vậy, khoảng cách từ "MI / TO" đến RM được chọn là 564, đó là khoảng cách từ MI đến RM, vv

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

Bước 2

min d(i,j) = d(NA,RM) = 219
=> Trộn NA và RM thành nhóm mới gọi là NA/RM
Khoảng cách của nhóm mới là L(NA/RM) = 219
m = 2

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

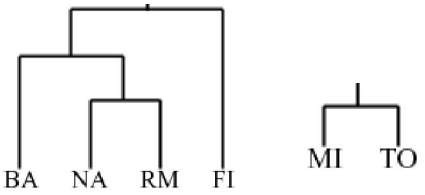
Bước 3

min d(i,j) = d(BA,NA/RM) = 255
=> Gom BA vàNA/RM vào nhóm mới gọi là BA/NA/RM
L(BA/NA/RM) = 255
m = 3

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

Bước 4

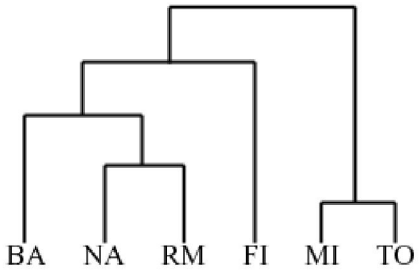
min d(i,j) = d(BA/NA/RM,FI) = 268
=> Gom cụm BA/NA/RM vào FI tạo thành nhóm mới gọi là BA/FI/NA/RM
L(BA/FI/NA/RM) = 268
m = 4



	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

Bước cuối cùng

Trộn 2 nhóm có giá trị khoảng cách 295 với nhau, tao được cây kết quả



Hierarchical clustering

- nhận xét
  1. giải thuật đơn giản
  2. cho kết quả dễ hiểu
  3. không cần tham số
  4. chạy chậm
  5. BIRCH (Zhang et al., 1996) sử dụng cấu trúc index để xử lý dữ liệu lớn

Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

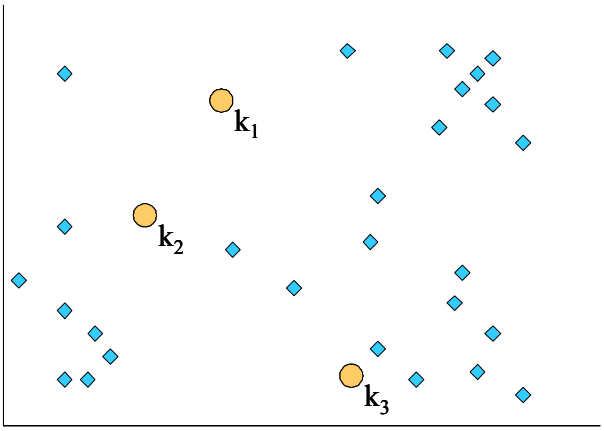
- giải thuật
  1. khởi động ngẫu nhiên **K tâm** (center) của **K clusters**
  2. mỗi phần tử được gán cho tâm gần nhất với phần tử dựa vào khoảng cách (e.g. khoảng cách Euclid)
  3. **cập nhật lại các tâm của K clusters**, mỗi tâm là giá trị trung bình (mean) của các phần tử trong cluster của nó
  4. lặp lại bước 2,3 cho đến khi hội tụ

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

Y

khởi động ngẫu nhiên 3 tâm của 3 clusters



X

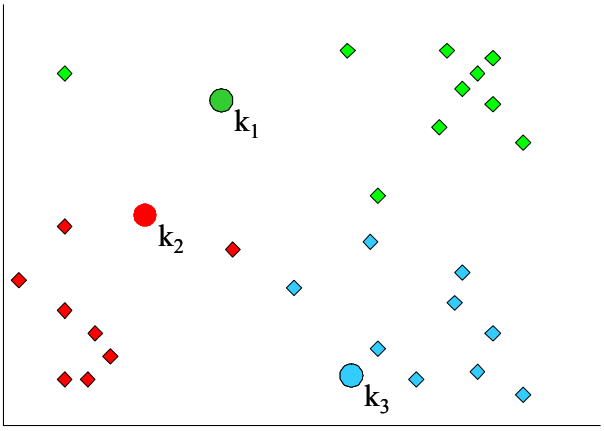
55

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

Y

mỗi phần tử được gán cho tâm cluster gần nhất của nó

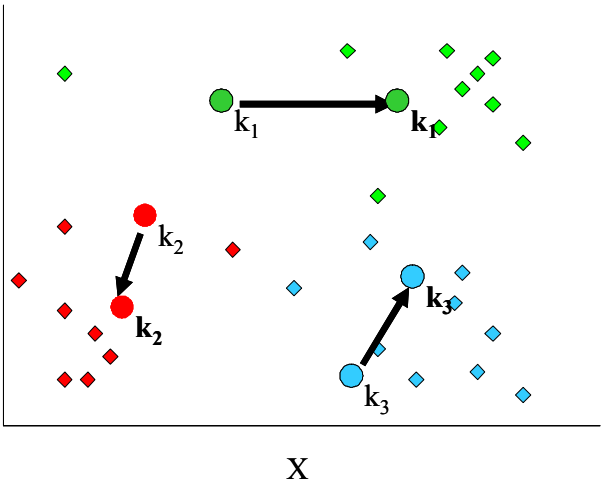


X

56

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

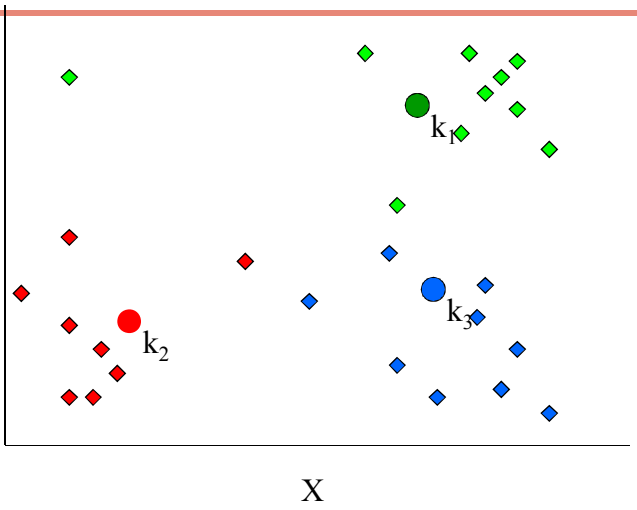
Giải thuật K-Means



cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means



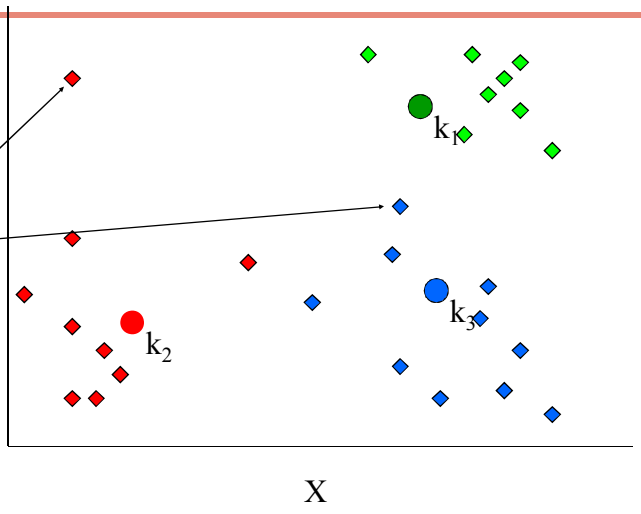
cấu hình mới của lần lặp tiếp theo

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

mỗi phần tử được gán lại cho tâm cluster gần nhất của nó

có 2 phần tử thay đổi nhóm

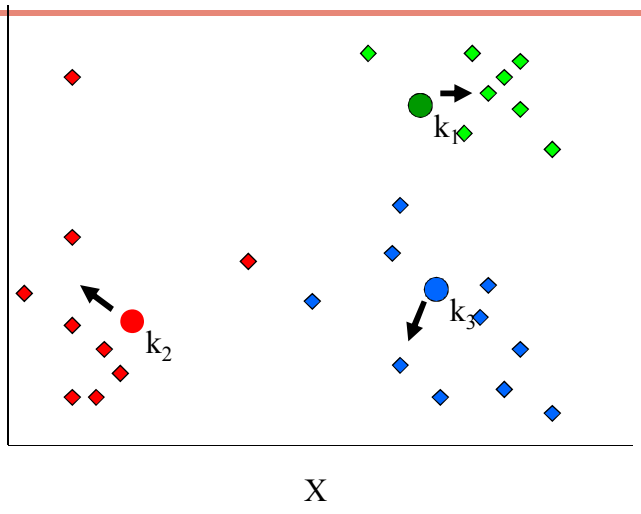


59

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)

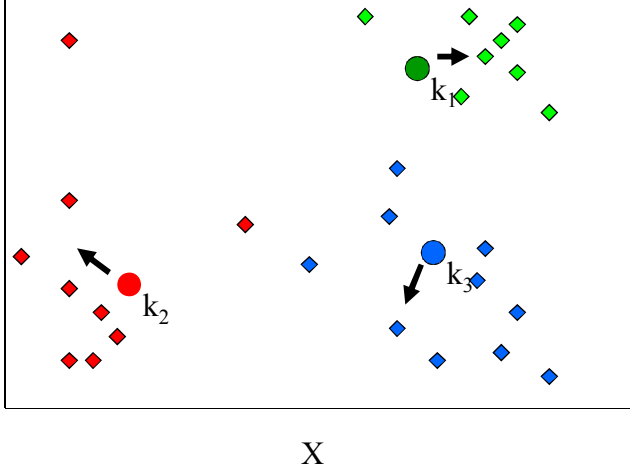


60

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

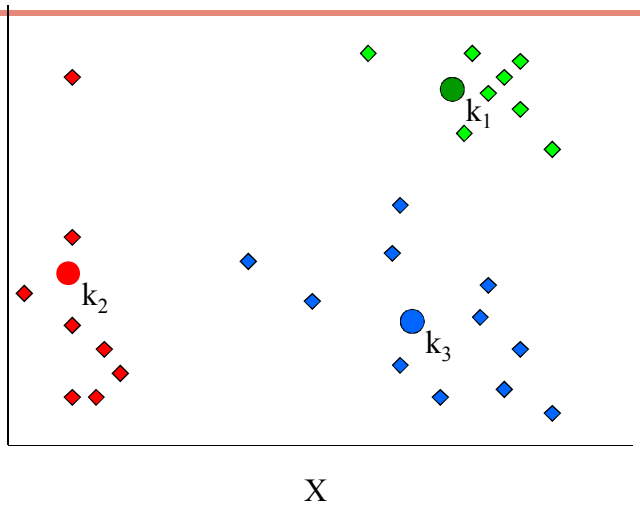
cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)



61

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means



62

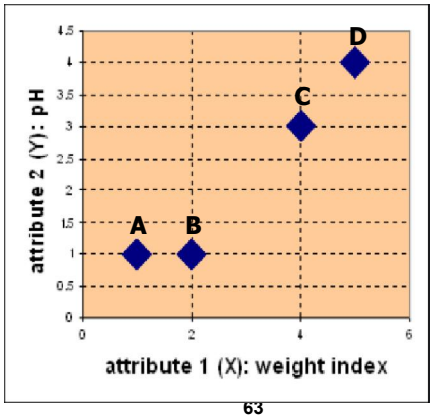
Bài tập

Bài tập 1:

Cho 4 loại thuốc mỗi loại có 2 thuộc tính pH và Weight

Yêu cầu nhóm những loại thuốc này thành 2 nhóm sử dụng khoảng cách Euclidea với 2 điểm khởi tạo là A và B

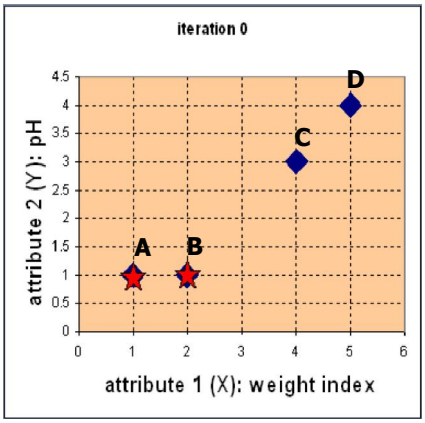
Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



63

Bài tập 1

Bước 1: khởi tạo các trọng tâm: A,B: trọng tâm nhóm 1/2 ; tính khoảng cách của các điểm còn lại đến 2 trọng tâm này



$c_1 = A, c_2 = B$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

Euclidean distance

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & & & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

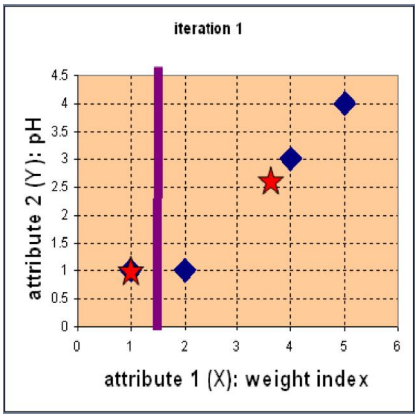
$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluster with the nearest seed point

64

Bài tập 1

Bước 2: tính lại 2 trọng tâm mới dựa vào các thành viên của nhóm vừa tạo ra ở bước 1



$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

$$c_1 = (1,\ 1)$$

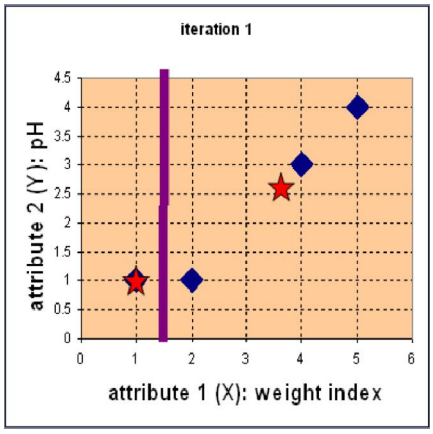
$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right)$$

$$= \left(\frac{11}{3},\ \frac{8}{3}\right)$$

65

Bài tập 1

- **Bước 2: Tính lại các thành viên theo 2 trọng tâm mới**



Compute the distance of all objects to the new centroids

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = \left(\frac{11}{3}, \frac{8}{3}\right) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

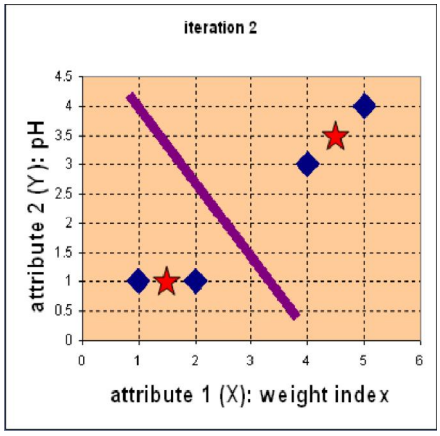
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Assign the membership to objects

66

Bài tập 1

- **Bước 3: Lặp lai 2 bước đầu tiên cho đến khi hội tụ**



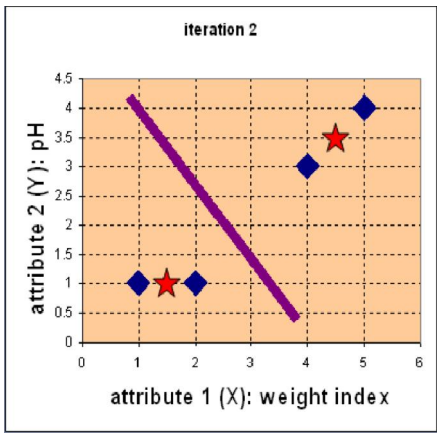
Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\frac{1}{2}, 3\frac{1}{2})$$

Bài tập 1

Bước 3: Lặp lai 2 bước đầu tiên cho đến khi hội tụ



Compute the distance of all objects to the new centroids

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} c_1 = (1\frac{1}{2}, 1) & group-1 \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Stop due to no new assignment
Membership in each cluster no longer change

Bài tập

- Cho tập dữ liệu gồm 10 phần tử có 2 thuộc tính x_1, x_2 được mô tả trong bảng bên cạnh. Anh, chị hãy thực hiện gom dữ liệu trên thành 2 nhóm bằng giải thuật Kmeans, với các thông tin sau:
- Các tâm khởi động ngẫu nhiên : c1(3,2) ; c2(5,3).
- Khoảng cách sử dụng: khoảng cách Euclid

STT	x_1	x_2
1.	1	2
2.	2	1
3.	3	2
4.	3	3
5.	5	2
6.	7	4
7.	5	3
8.	7	1
9.	6	3
10.	7	2

Lần lặp 1:
Tính khoảng cách từ các phần tử đến 2 tâm c1(3,2) ; c2(5,3)

I	x_1	x_2	$d^2(I,x1)$			$d^2(I,x2)$	Nhóm gần nhất
1.	1	2	4	2	4.123105626	17	1
2.	2	1	2	1.414213562	3.605551275	13	1
3.	3	2	0	0	2.236067977	5	1
4.	3	3	1	1	2	4	1
5.	5	2	4	2	1	1	2
6.	7	4	20	4.472135955	2.236067977	5	2
7.	5	3	5	2.236067977	0	0	2
8.	7	1	17	4.123105626	2.828427125	8	2
9.	6	3	10	3.16227766	1	1	2
10.	7	2	16	4	2.236067977	5	2

Cập nhật lại tâm: c1((1+2+3+3)/4,(2+1+2+3)/4)= (2.25;2)
C2((5+7+5+7+6+7)/6, (2+4+3+1+3+2)/6= (6.17; 2.5)

Lần lặp 2:
Tính khoảng cách từ các phần tử đến 2 tâm c1 (2.25;2) và c2 (6.17; 2.5)

I	x_1	x_2	$d^2(I,x1)$			$d^2(I,x2)$	Nhóm gần nhất
1.	1	2	1.5625	1.25	5.190837012	26.94478889	1
2.	2	1	1.0625	1.030776406	4.428474781	19.61138889	1
3.	3	2	0.5625	0.75	3.205930269	10.27798889	1
4.	3	3	1.5625	1.25	3.205930269	10.27798889	1
5.	5	2	7.5625	2.75	1.269326156	1.61118889	2
6.	7	4	26.5625	5.153882032	1.715922169	2.94438889	2
7.	5	3	8.5625	2.926174978	1.269326156	1.61118889	2
8.	7	1	23.5625	4.85412196	1.715922169	2.94438889	2
9.	6	3	15.0625	3.881043674	0.527056819	0.27778889	2
10.	7	2	22.5625	4.75	0.971796733	0.94438889	2

Các phần tử trong nhóm không thay đổi, giải thuật dừng lại ta có thông tin nhóm 1 gồm 4 phần tử đầu tiên, nhóm 2 gồm 6 phần tử còn lại

Bài tập 2: k=2

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Bước 1:

Khởi tạo k=2 trọng tâm: m1=(1.0,1.0) và m2=(5.0,7.0).

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Bước 2:

- **Sau bước 1 ta được 2 nhóm: {1,2,3} và {4,5,6,7}.**
- **Their new centroids are:**

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1,2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2,2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Step 3:

Nhóm mới: {1,2} and {3,4,5,6,7}

- **Trọng tâm mới: m1=(1.25,1.5) và m2 = (3.9,5.1)**

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- **Bước 4:**

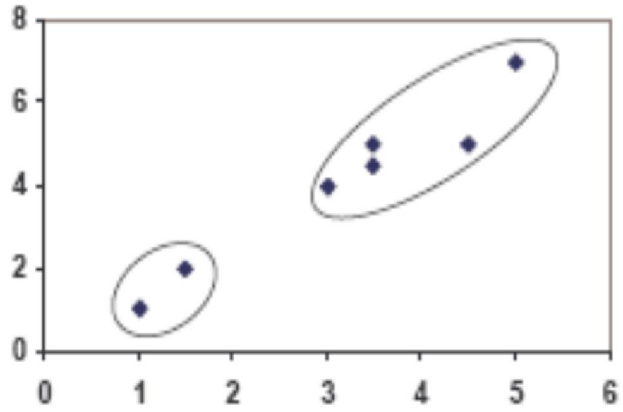
  Nhóm:

  {1,2} và {3,4,5,6,7}

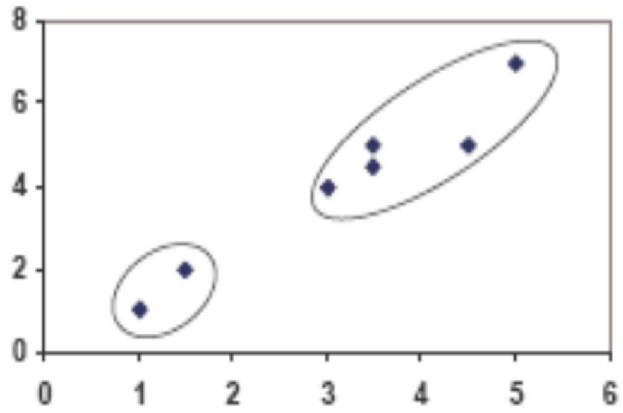
- **=> các thành viên trong nhóm không thay đổi => giải thuật dừng, ta có 2 nhóm {1,2} và {3,4,5,6,7}.**

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

PLOT



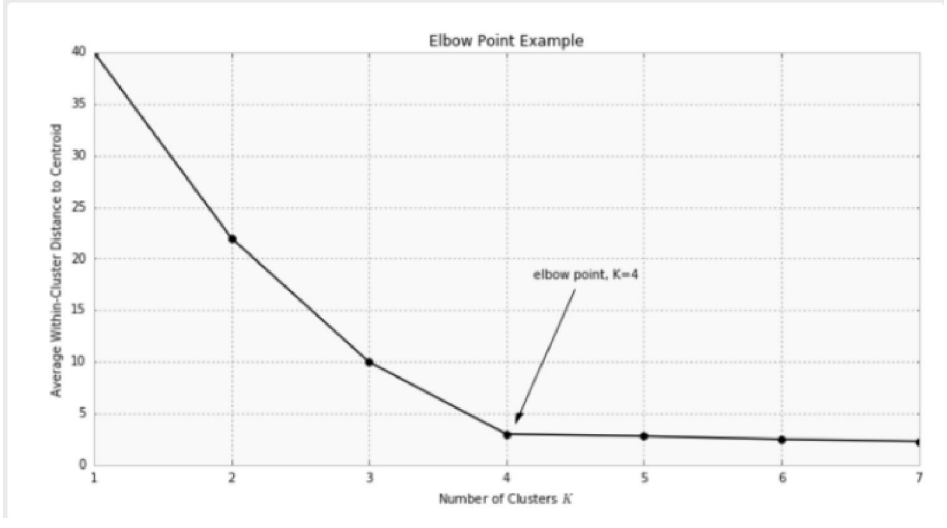
PLOT



Giải thuật K-Means

- nhận xét
  1. giải thuật đơn giản
  2. cho kết quả dễ hiểu
  3. cần cho tham số K (số lượng clusters)
  4. kết quả phụ thuộc vào việc khởi động ngẫu nhiên K tâm (center) của K clusters : có thể khắc phục bằng cách khởi động lại nhiều lần.
  5. khả năng chịu đựng nhiễu không tốt (ảnh hưởng bởi các phần tử outliers) : có thể khắc phục bằng K-Medoids, không sử dụng giá trị trung bình, nhưng sử dụng phần tử ngay giữa

79



Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

81

Giải thuật clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

- còn nhiều phương pháp khác
  - density-based : DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DENCLUE (Hinneburg & Keim, 1998)
  - model-based : EM (Expected maximization), SOM (Kohonen, 1995)

82

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

Hướng phát triển

- các kiểu dữ liệu phức tạp
- tăng tốc độ xử lý
- các tham số đầu vào của giải thuật
- diễn dịch kết quả sinh ra
- phương pháp kiểm chứng chất lượng mô hình

Cám ơn !

Kiểu nhị phân

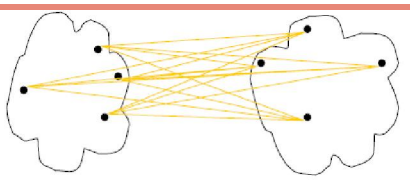
		Object *j* 1	0	*sum*
Object *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

- khoảng cách đối xứng : $d(i,j) = \frac{b+c}{a+b+c+d}$
- khoảng cách bất đối xứng : $d(i,j) = \frac{b+c}{a+b+c}$
- hệ số Jaccard bất đối xứng : $sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$

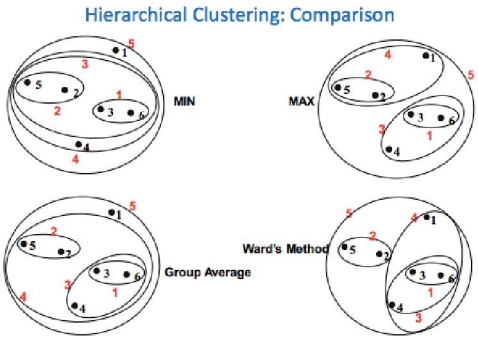
Hierarchical clustering

Group Average Linkage

***average-link* clustering, distance = average distance**



• Strengths– Less susceptible to noise and outliers• Limitations– Biased towards globular clusters



What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.