



Giải thuật gom cụm Clustering algorithms

Đỗ Thanh Nghi

dtngchi@cit.ctu.edu.vn

Trần Nguyễn Minh Thư

tnmthu@cit.ctu.edu.vn

2

Nội dung

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

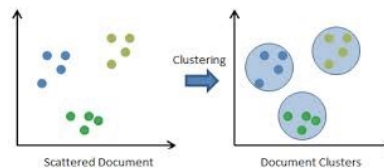
3

3

Clustering

■ 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

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.

5

5

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 chi phí bồi thường trung bình, cao, thấp khác nhau tùy 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.

6

6

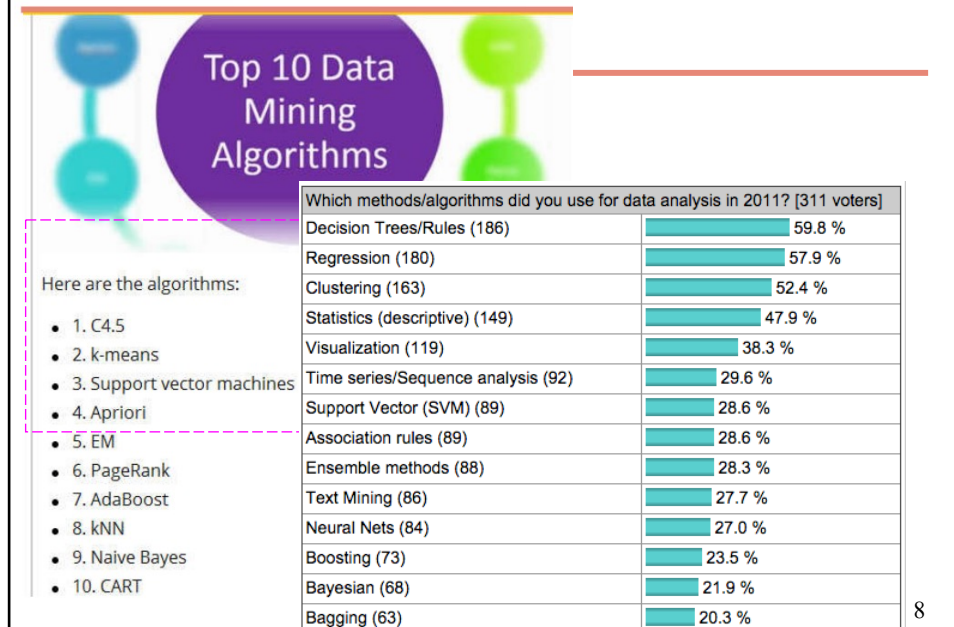
Clustering

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

7

7

Top 10 DM algorithms (2015)



8

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

9

9

Các độ đo khoảng cách - Kiểu số

■ Khoảng cách *Minkowski*

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \dots + |x_{i_p} - x_{j_p}|^q)}$$

$i = (x_{i_1}, x_{i_2}, \dots, x_{i_p})$ và $j = (x_{j_1}, x_{j_2}, \dots, x_{j_p})$ là 2 phần tử dữ liệu trong p -dimensional, q là số nguyên dương

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

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

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

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

10

10

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 :

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

11

11

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
Điệp	Nâu	Đen	Cao	Trung bình	Cao đẳng

d(Nam, Lan) = ?

d(Nam, Điệp) = ?

12

12

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 : $sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$

13

13

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 \rightarrow 1, N \rightarrow 0

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

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

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

14

14

Nội dung

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

15

15

Giải thuật K-Means

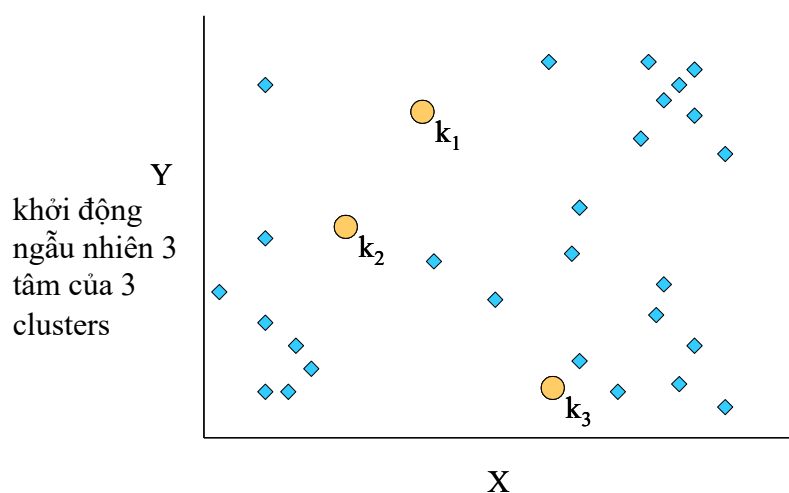
■ giải thuật

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

16

16

Giải thuật K-Means

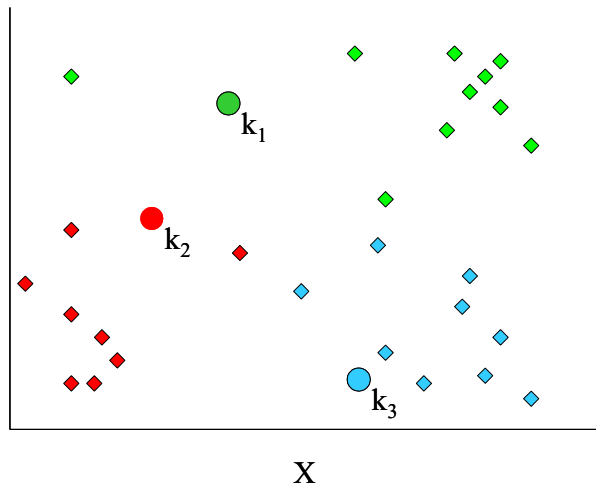


17

17

Giải thuật K-Means

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

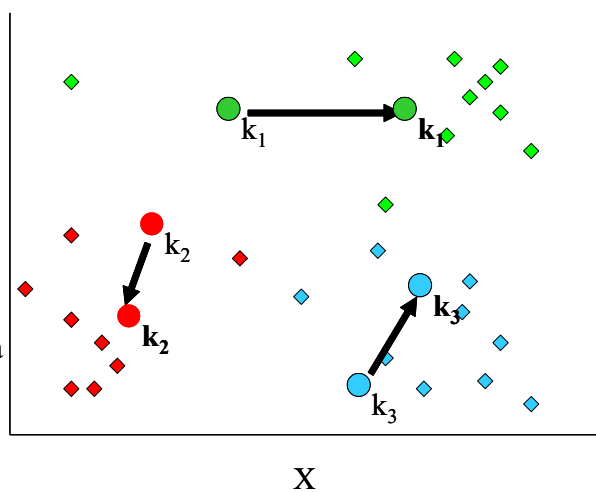


18

18

Giải thuật K-Means

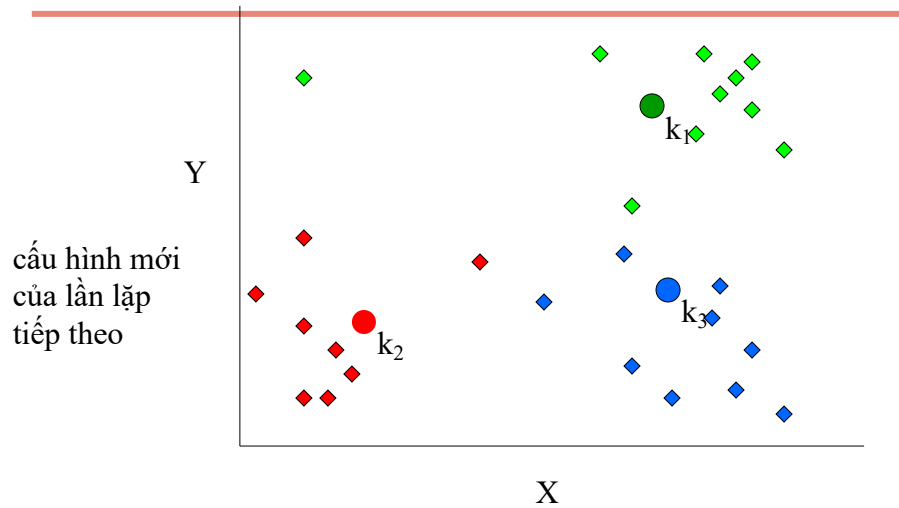
Y
cập nhật lại
tâm của các
cluster (giá trị
trung bình của
các phần tử
trong cluster)



19

19

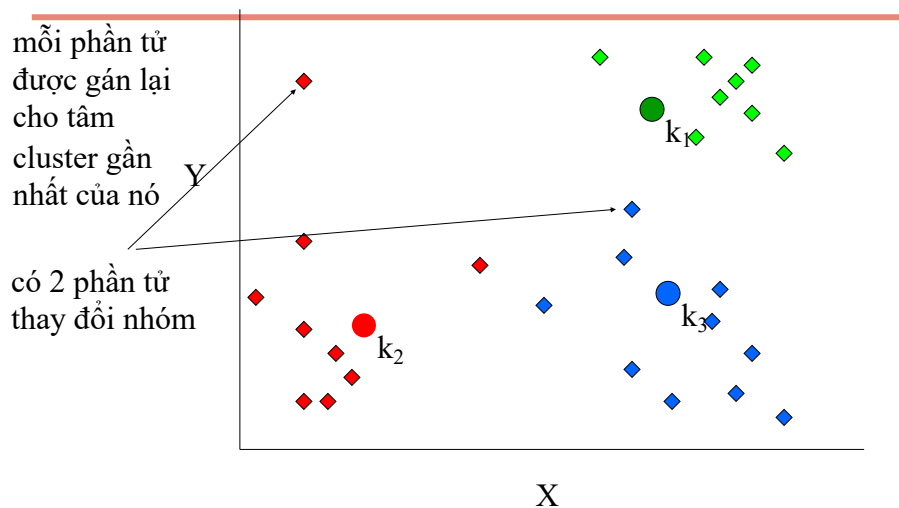
Giải thuật K-Means



20

20

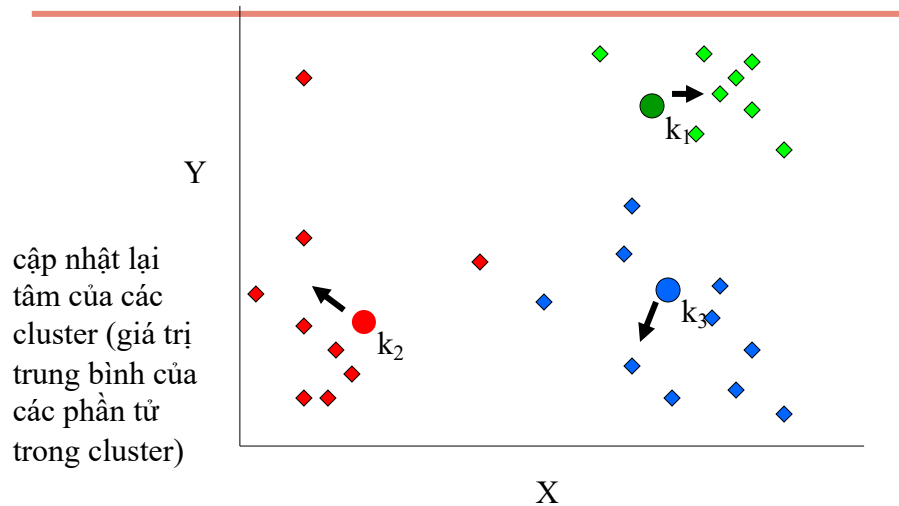
Giải thuật K-Means



21

21

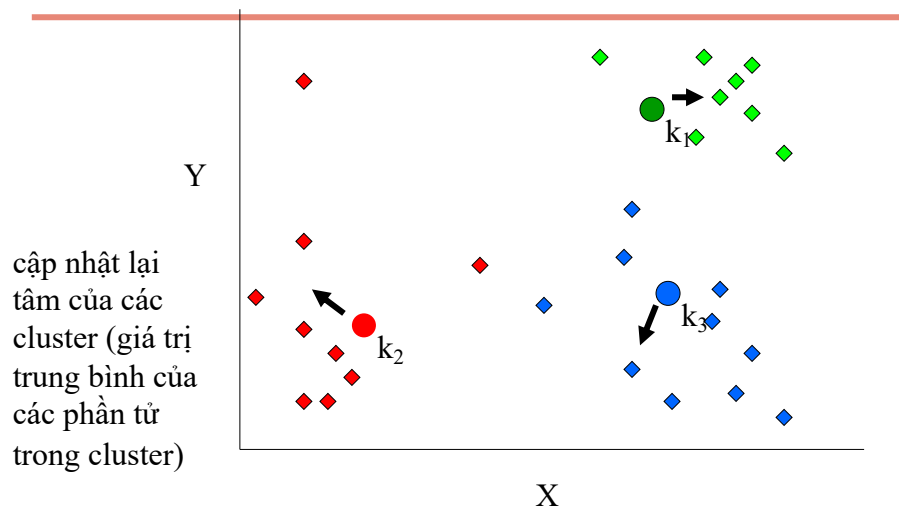
Giải thuật K-Means



22

22

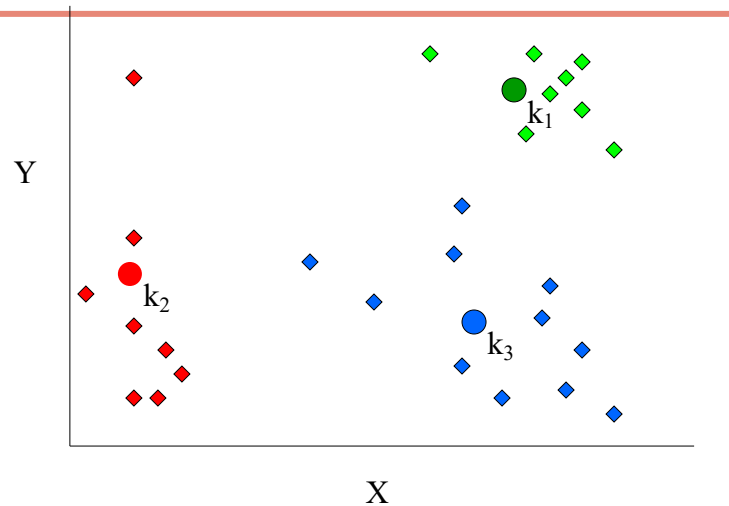
Giải thuật K-Means



23

23

Giải thuật K-Means



24

24

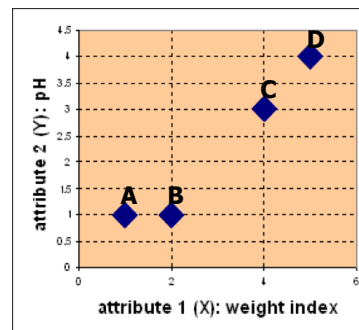
Bài tập

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 Euclidean với 2 điểm khởi tạo là **A và B**

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



25

25

Bài tập

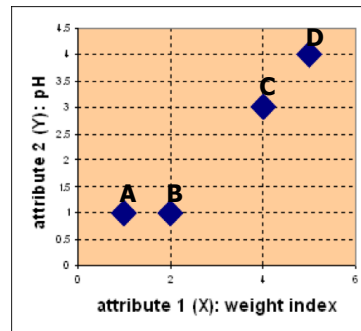
$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

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 Euclidean với 2 điểm khởi tạo là **A và B**

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



26

26

Bài tập

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

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 Euclidean với 2 điểm khởi tạo là **A và B**

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

d(A, tâm 1), d(B, tâm 1),
d(C, tâm 1), d(D, tâm 1),

d(A, tâm 2), d(B, tâm 2),
d(C, tâm 2), d(D, tâm 2),

27

27

Bài tập

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

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 Euclidean với 2 điểm khởi tạo là A và B

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

	Tâm c1	Tâm c2
A	d(A, tâm 1)	d(A, tâm 2)
B	d(B, tâm 1)	d(B, tâm 2)
C	d(C, tâm 1)	d(C, tâm 2)
D	d(D, tâm 1)	d(D, tâm 2)

28

28

Bài tập

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

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

	Tâm c1 (A)	Tâm c2 (B)
A	d(A, tâm 1)	d(A, tâm 2)
B	d(B, tâm 1)	d(B, tâm 2)
C	d(C, tâm 1)	d(C, tâm 2)
D	d(D, tâm 1)	d(D, tâm 2)

$$d(A, \text{tâmC1} \equiv A) = \sqrt{((1-1)^2 + (1-1)^2)} = 0$$

$$d(B, \text{tâmC1} \equiv A) = ?$$

$$d(A, \text{tâmC2} \equiv B) = \sqrt{((1-2)^2 + (1-1)^2)} = 1$$

$$d(B, \text{tâmC2} \equiv B) = ?$$

29

29

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)
A	0	1
B	1	0
C	d(C, tâm 1)	d(C, tâm 2)
D	d(D, tâm 1)	d(D, tâm 2)

$$d(B, \text{tâm C1} \equiv A) = \sqrt{((1-2)^2 + (1-1)^2)} = 1$$

$$d(B, \text{tâm C2} \equiv B) = \sqrt{((1-1)^2 + (1-1)^2)} = 0$$

30

30

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)
A	0	1
B	1	0
C	3.61	2.83
D	5	4.24

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

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

31

31

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)	Nhóm
A	0	1	?
B	1	0	?
C	3.61	2.83	?
D	5	4.24	?

32

32

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)	Nhóm
A	0	1	1
B	1	0	2
C	3.61	2.83	2
D	5	4.24	2

33

33

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)	Nhóm
A	0	1	1
B	1	0	2
C	3.61	2.83	2
D	5	4.24	2

Bước 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

=> Tính lại trọng tâm c1 và c2

34

34

Bài tập

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

Bước 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

=> Tính lại trọng tâm c1 và c2

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

$$c_1 = (1, 1)$$

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left(\frac{11}{3}, \frac{8}{3} \right)$$

35

35

Bài tập

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

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

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)
A	d(A, tâm 1)	d(A, tâm 2)
B	d(B, tâm 1)	d(B, tâm 2)
C	d(C, tâm 1)	d(C, tâm 2)
D	d(D, tâm 1)	d(D, tâm 2)

$$d(A, \text{tâm C1} \equiv A) = \sqrt{((1-1)^2 + (1-1)^2)} = 0 \quad d(B, \text{tâm C1} \equiv A) = ?$$

$$d(A, \text{tâm C2} \equiv B) = \sqrt{((1-11/3)^2 + (1-8/3)^2)} = ? \quad d(B, \text{tâm C2} \equiv B) = ?$$

36

36

Bài tập

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

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

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)
A	0	3.14
B	d(B, tâm 1)	d(B, tâm 2)
C	d(C, tâm 1)	d(C, tâm 2)
D	d(D, tâm 1)	d(D, tâm 2)

$$d(B, \text{tâm C1} \equiv A) = ?$$

$$d(A, \text{tâm C1} \equiv A) = \sqrt{((1-1)^2 + (1-1)^2)} = 0 \quad d(B, \text{tâm C2} \equiv B) = ?$$

$$d(A, \text{tâm C2} \equiv B) = \sqrt{((1-11/3)^2 + (1-8/3)^2)} = 3.14$$

37

37

Bài tập

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

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)
A	0	3.14
B	1	2.36
C	3.61	0.47
D	5	1.89

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

38

38

Bài tập

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

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

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)	Nhóm
A	0	3.14	?
B	1	2.36	?
C	3.61	0.47	?
D	5	1.89	?

39

Bài tập

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

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

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)	Nhóm
A	0	3.14	1
B	1	2.36	1
C	3.61	0.47	2
D	5	1.89	2

40

40

Bài tập

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

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

=> Tính lại trọng tâm c1 và c2

	Tâm c1 (1,1)	Tâm c2 (11/3,8/3)	Nhóm
A	0	3.14	1
B	1	2.36	1
C	3.61	0.47	2
D	5	1.89	2

41

41

Bài tập

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

Bước 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

=> Tính lại trọng tâm c1 và c2

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

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

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

42

42

Bài tập

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

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

	Tâm c1 (3/2;1)	Tâm c2 (9/2;7/2)	Nhóm
A	0.5	4.3	?
B	0.5	3.54	?
C	3.2	0.71	?
D	4.61	0.71	?

43

43

Bài tập

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

Thuốc	Weight	pH-Index	Tâm c1 (3/2;1)	Tâm c2 (9/2;7/2)	Nhóm
A	1	1	0.5	4.3	1
B	2	1	0.5	3.54	1
C	4	3	3.2	0.71	2
D	5	4	4.61	0.71	2

=> Tính lại trọng tâm c1 và c2???

44

44

Bài tập

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

Thuốc	Weight	pH-Index	Tâm c1 (3/2;1)	Tâm c2 (9/2;7/2)	Nhóm
A	1	1	0.5	4.3	1
B	2	1	0.5	3.54	1
C	4	3	3.2	0.71	2
D	5	4	4.61	0.71	2

=> Tính lại trọng tâm c1 và c2???

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

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

45

45

Bài tập

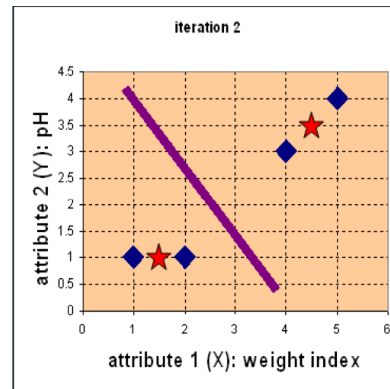
$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

	Tâm c1 (3/2;1)	Tâm c2 (9/2;7/2)	Nhóm
A	0.5	4.3	1
B	0.5	3.54	1
C	3.2	0.71	2
D	4.61	0.71	2

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

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

Trọng tâm không thay đổi, quá trình gom nhóm đã hội tụ
=> tìm được nhóm 1 (A,B), nhóm 2(C,D)



46

46

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

47

Sử dụng khoảng cách Euclid

Khởi tạo $k=2$ trọng tâm: $m_1=(1.0,1.0)$ và $m_2=(5.0,7.0)$.

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)

48

Nội dung

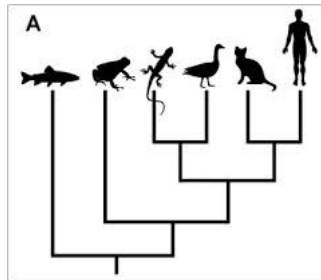
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

49

49

Hierarchical Clustering

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



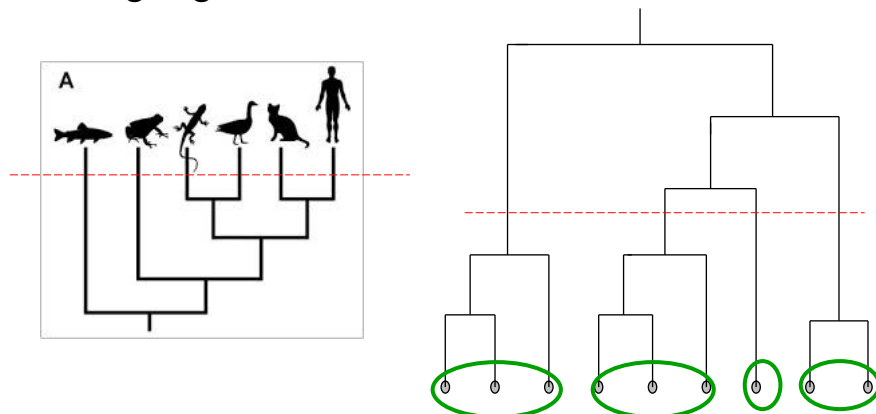
- Dựa trên điểm cắt ở đâu mà ta thu được các cụm tương ứng

50

50

Hierarchical Clustering

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



51

51

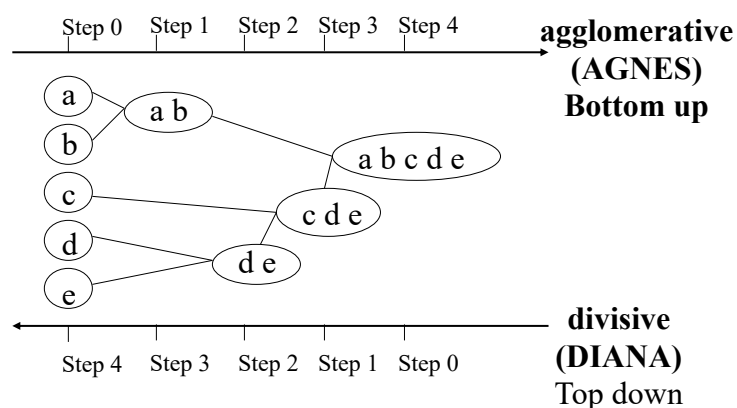
Hierarchical clustering

- 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

52

52

Hierarchical clustering

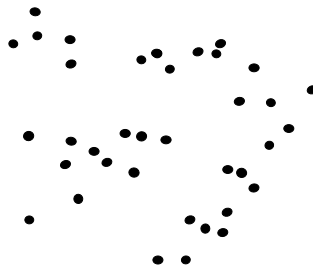


53

53

Hierarchical clustering (Single link)

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

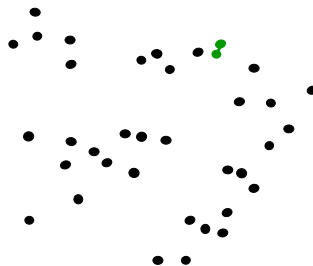


① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

54

Hierarchical clustering (Single link)

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



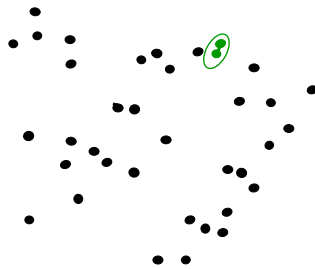
① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

② Tìm “khoảng cách” tương tự nhất giữa các cặp cụm

55

Hierarchical clustering (Single link)

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



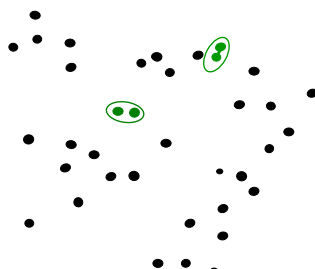
- ① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt
- ② Tìm “khoảng cách” tương tự nhất giữa các cặp cụm
- ③ Kết hợp từng 2 cặp điểm thành một cụm mẹ/cụm lớn hơn



56

Hierarchical clustering (Single link)

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



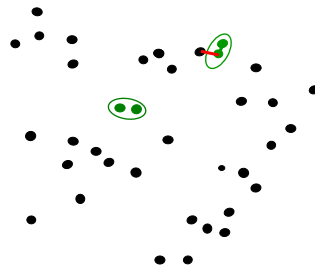
- ① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt
- ② Tìm “khoảng cách” tương tự nhất giữa các cặp cụm
- ③ Kết hợp từng 2 cặp điểm thành một cụm mẹ/cụm lớn hơn
- ④ **Lặp lại...**



57

Hierarchical clustering (Single link)

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



58

Hierarchical clustering (Single link)

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



59

59

Hierarchical clustering (Single link)

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



60

60

Hierarchical clustering (Single link)

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

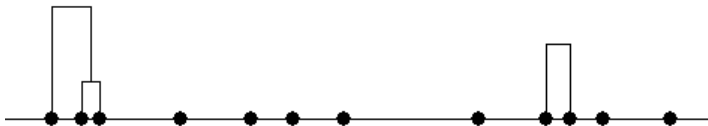


61

61

Hierarchical clustering (Single link)

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

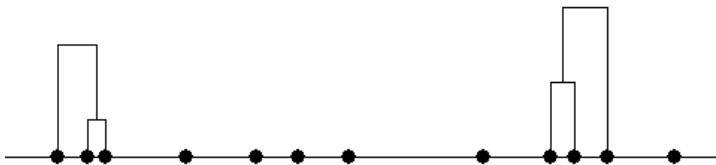


62

62

Hierarchical clustering (Single link)

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

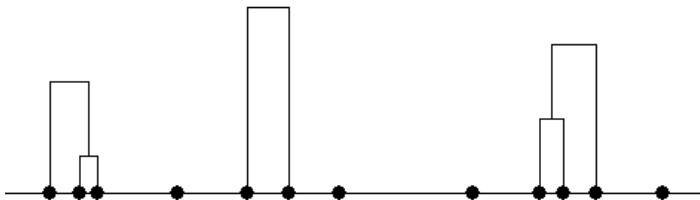


63

63

Hierarchical clustering (Single link)

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

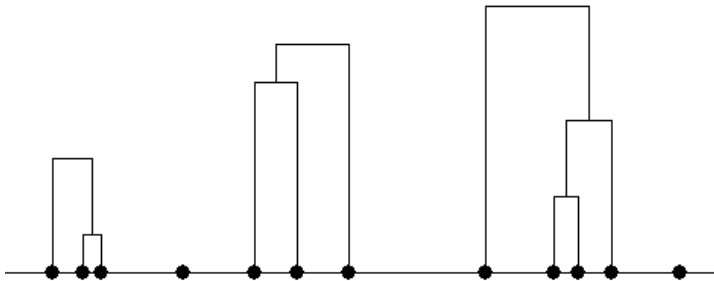


64

64

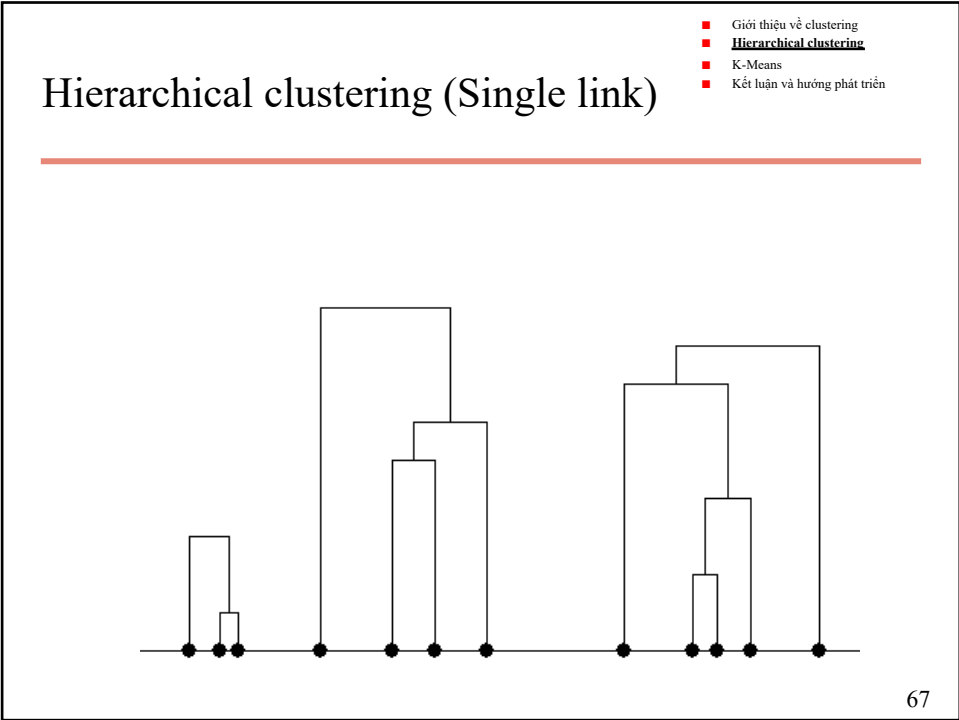
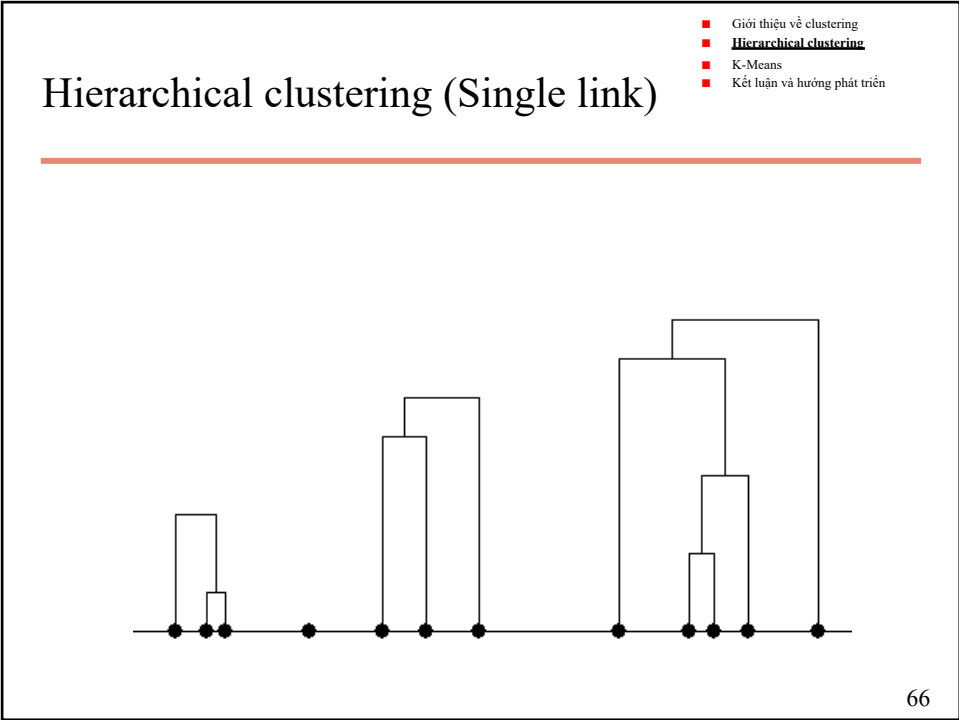
Hierarchical clustering (Single link)

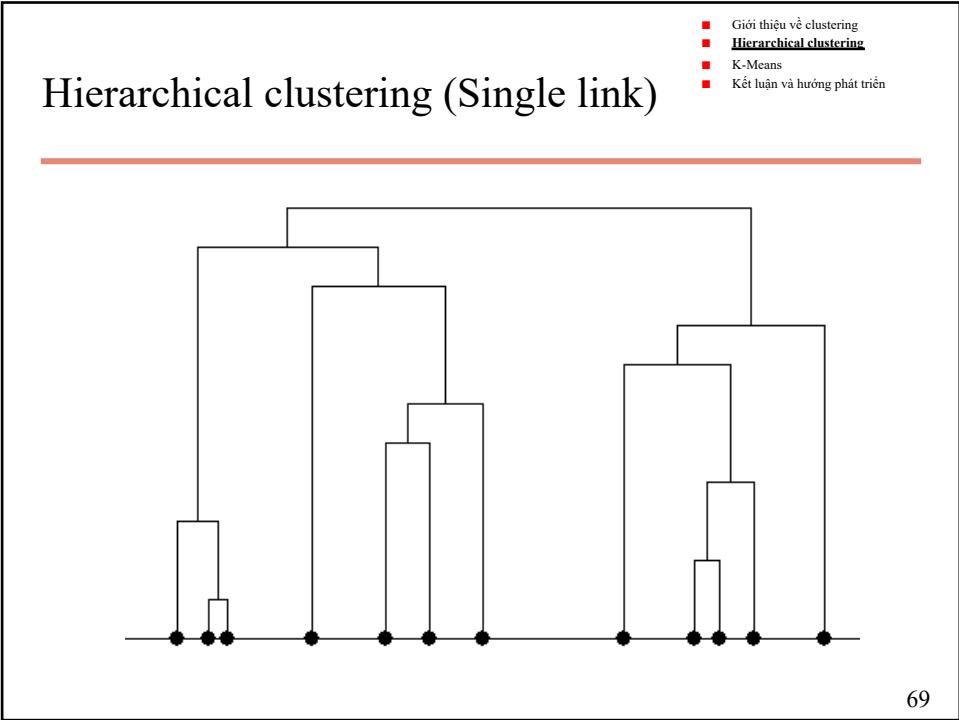
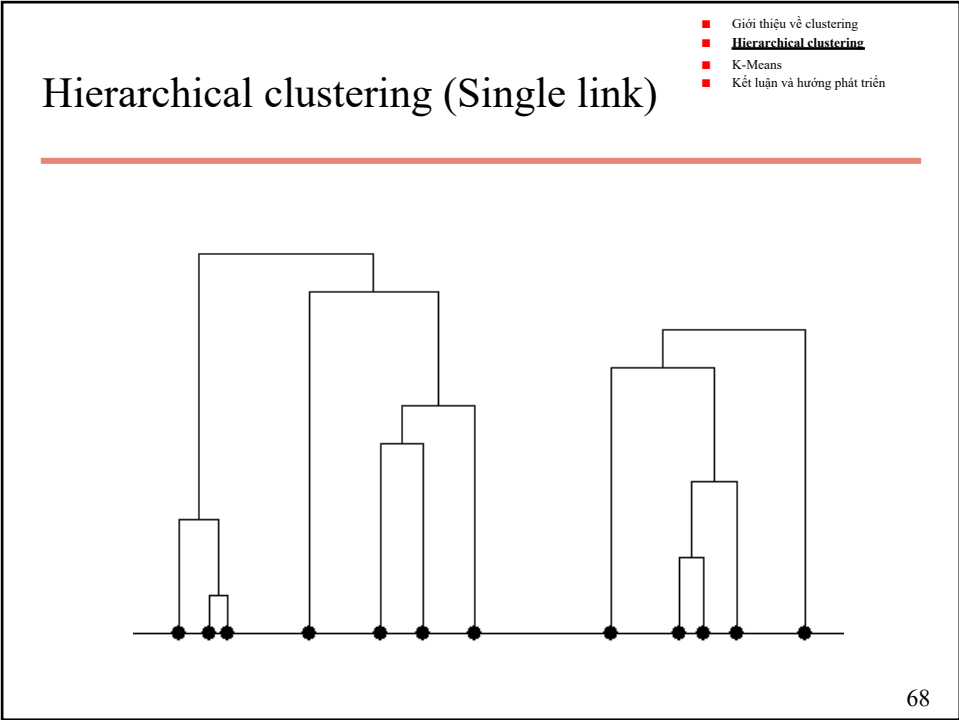
- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển



65

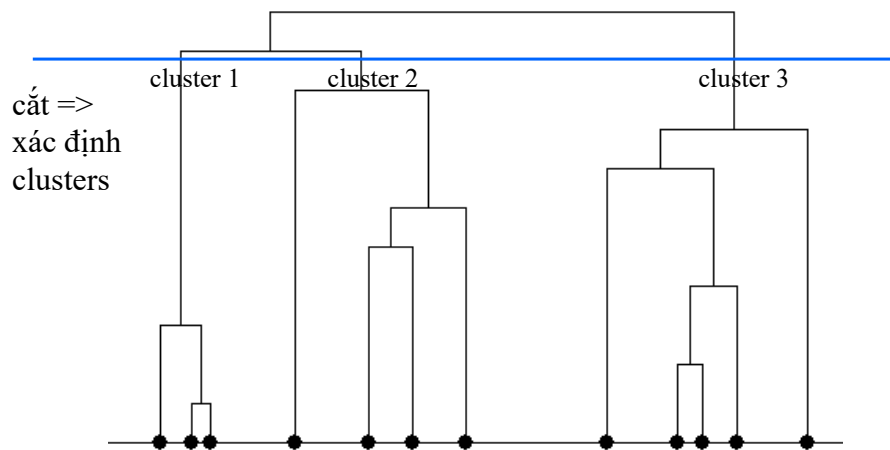
65





Hierarchical clustering (Single link)

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

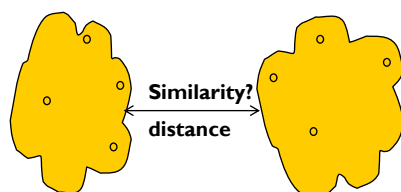


70

70

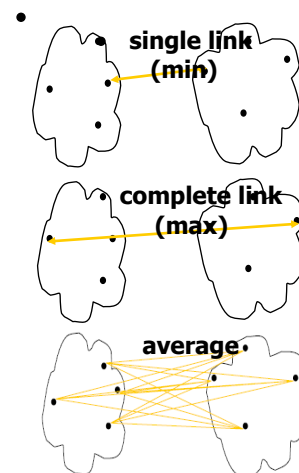
Hierarchical clustering

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



● Định nghĩa khoảng cách,
độ tương tự của 2 nhóm

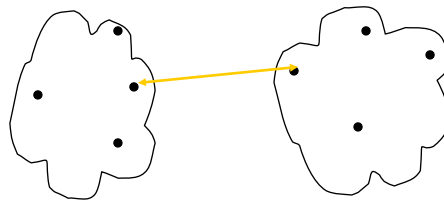
- MIN
- MAX
- Group Average



71

Hierarchical clustering

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



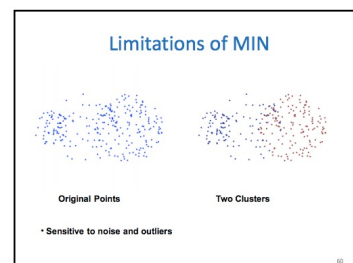
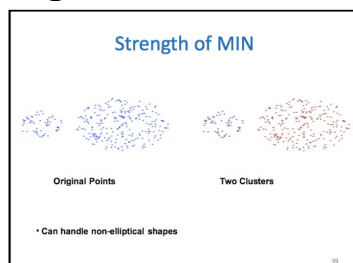
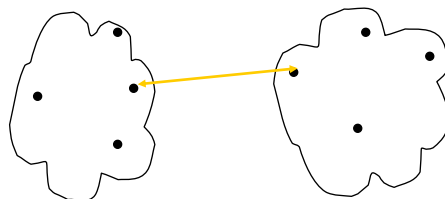
- **distance = shortest distance, nearest neighbor clustering algorithm**
- **MIN Linkage**
- **MAX Linkage**
- **Group Average**

72

Hierarchical clustering

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

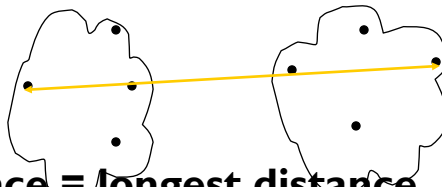
- **MIN Linkage**
- **distance = shortest distance, nearest neighbor clustering algorithm**



73

Hierarchical clustering

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



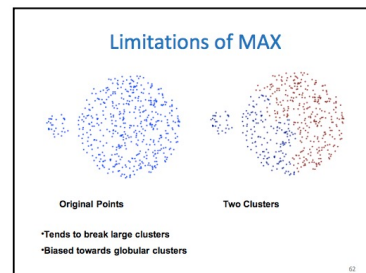
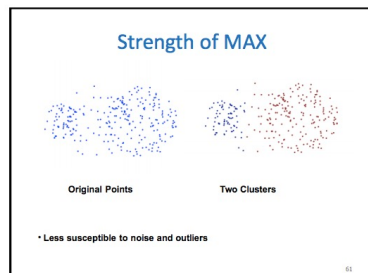
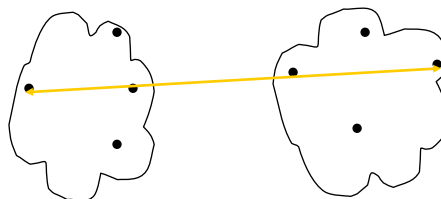
- distance = longest distance , farthest neighbor clustering algorithm
- MIN Linkage
- **MAX Linkage**
- Group Average Linkage

74

Hierarchical clustering

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

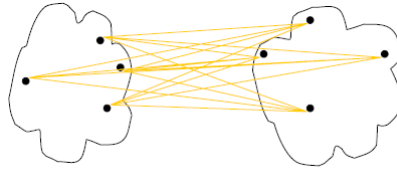
- **MAX Linkage**
- distance = longest distance , farthest neighbor clustering algorithm



75

Hierarchical clustering

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



average-link clustering, distance = average distance

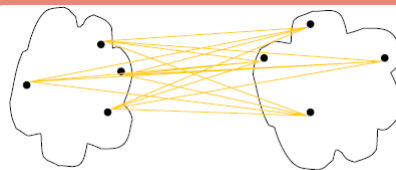
- MIN Linkage
- MAX Linkage
- **Group Average Linkage**

76

Hierarchical clustering

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

Group Average Linkage
average-link clustering, distance = average distance



- Ít nhạy cảm với nhiễu và outliers

77

Hierarchical clustering (Single link)

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

■ 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



78

78

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

79

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

80

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

81

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

82

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

83

Bước 3

$$\min d(i,j) = d(BA, NA/RM) = 255$$

=> Gom BA và NA/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

84

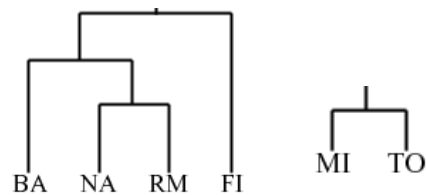
Bước 4

$$\min d(i,j) = d(BA/NA/RM, FI) = 268$$

=> Gom cụm BA/NA/RM vào FI tạo thành nhóm mới gọi là BA/FI/NA/RM

$$L(BA/FI/NA/RM) = 268$$

$$m = 4$$

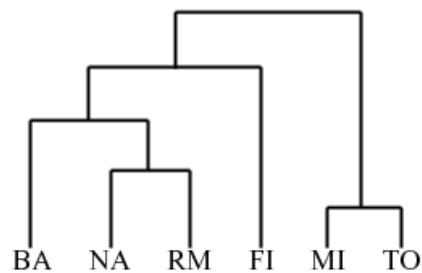


	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

85

Bước cuối cùng

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



86

Hierarchical clustering

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

■ 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

87

87

Nội dung

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

88

88

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)

89

89

Hướng phát triển

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và 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

90

90



Cám ơn !

91