Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Giải thuật gom cụm
Clustering algorithms

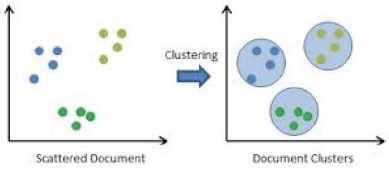
Bộ môn KHMT

Nội dung

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

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)

3

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.

4

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 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.

5

Clustering

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

6

Top 10 DM algorithms (2015)

Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %

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
Điệp	Nâu	Đen	Cao	Trung bình	Cao đẳng

d(Nam, Lan) =? **d(Nam, Điệp) =?**

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

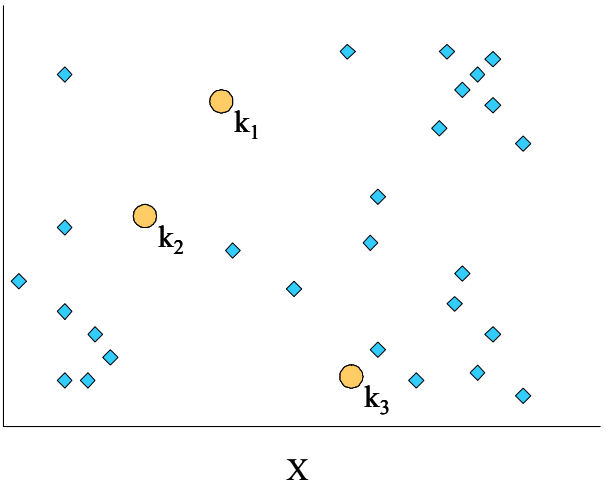
14

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

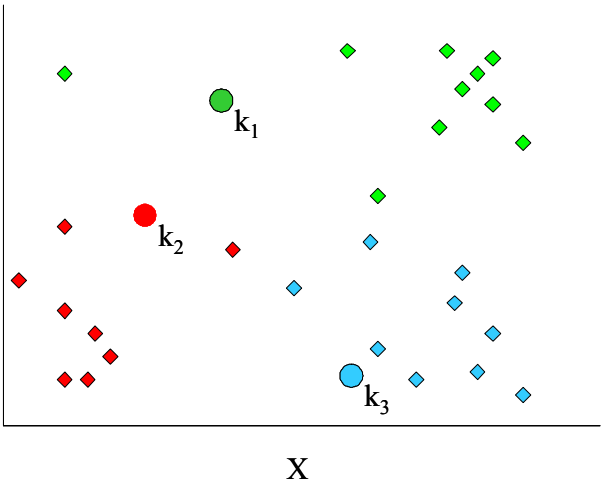
15

Giải thuật K-Means



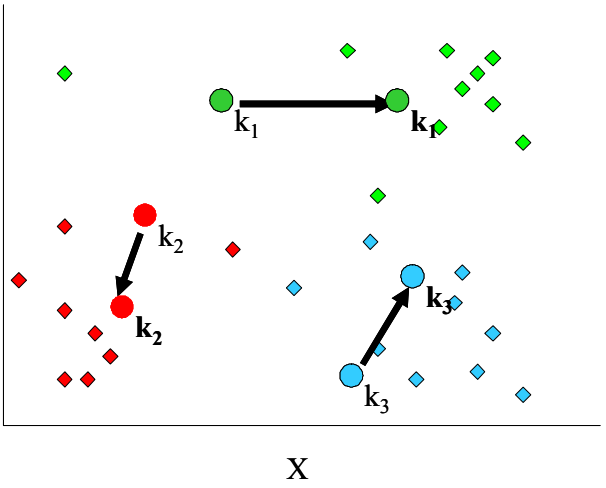
16

Giải thuật K-Means



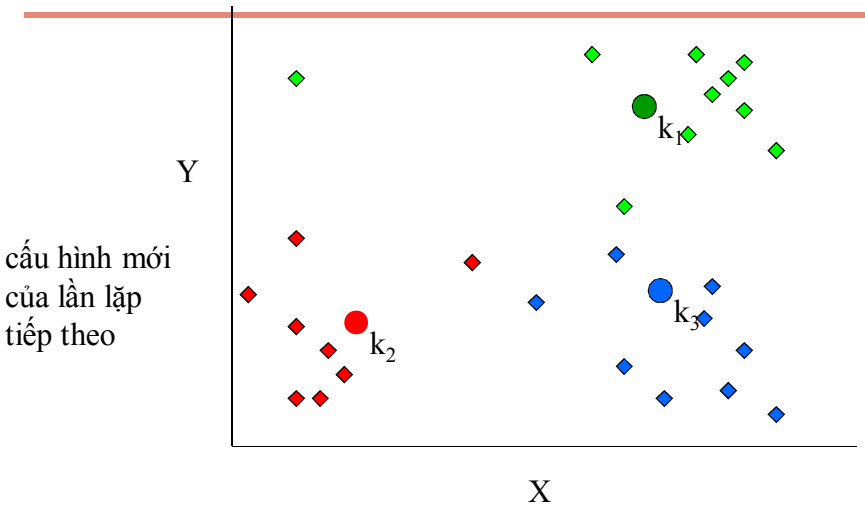
17

Giải thuật K-Means



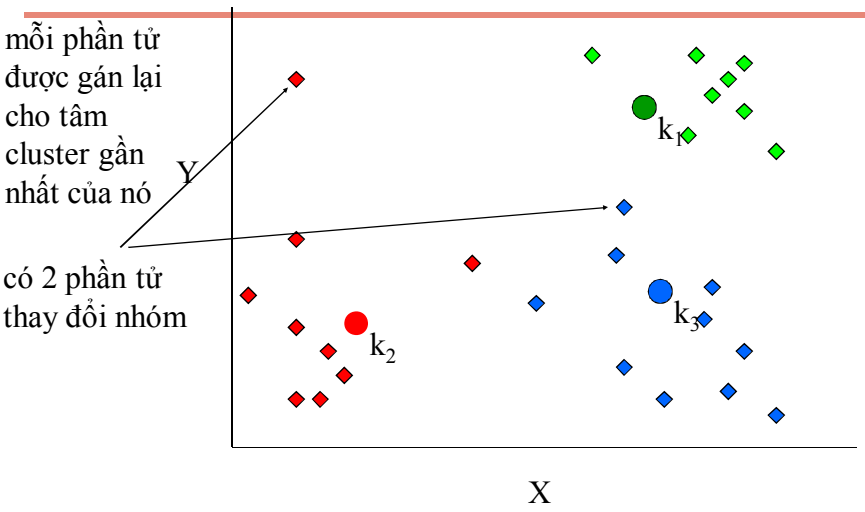
18

Giải thuật K-Means



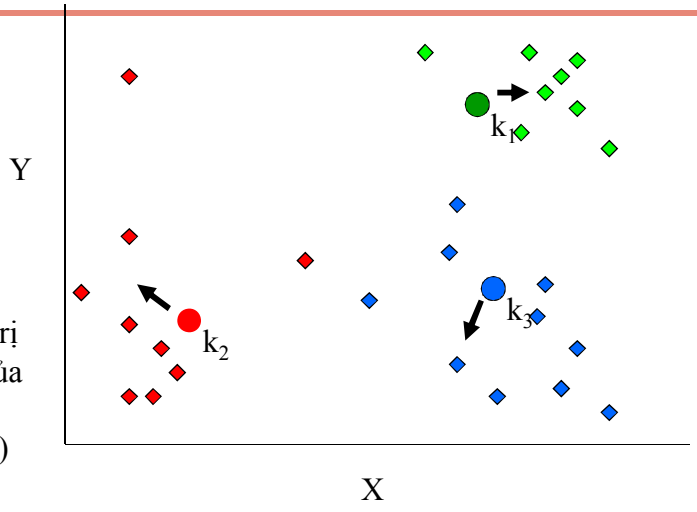
19

Giải thuật K-Means



20

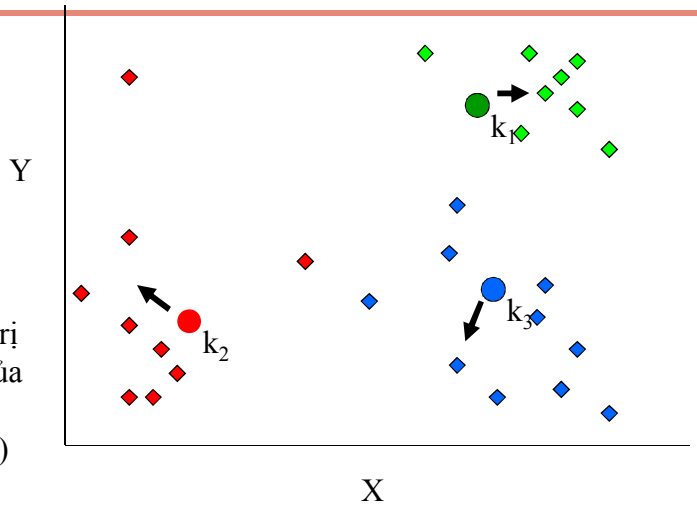
Giải thuật K-Means



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

21

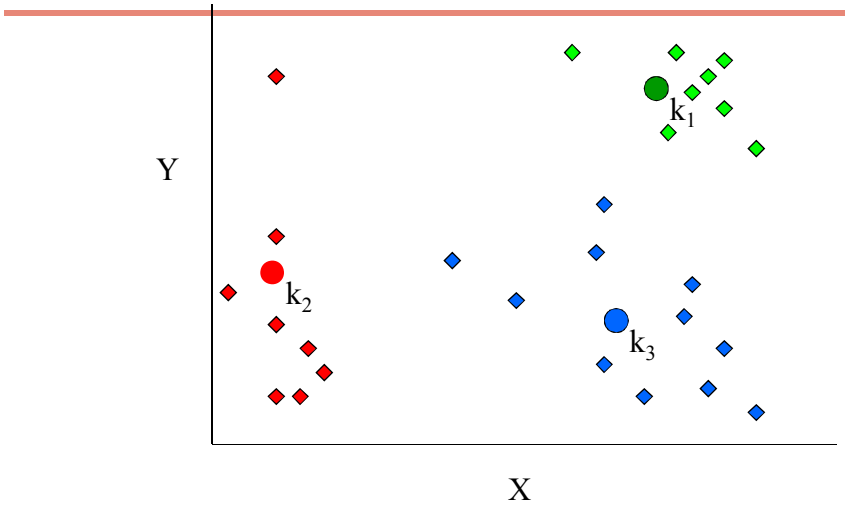
Giải thuật K-Means



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

22

Giải thuật K-Means



23

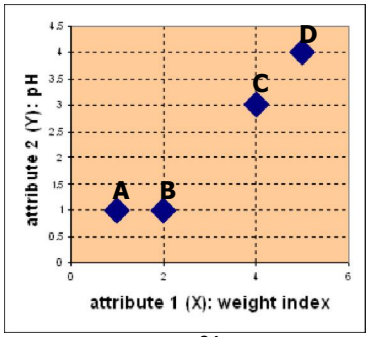
Bài tập

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

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



24

Bài tập

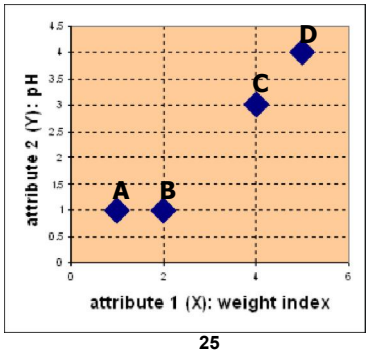
$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

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 Euclidea 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

Bài tập

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

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 Euclidea 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),

26

Bài tập

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

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 Euclidea 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)

27

Bài tập

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

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)

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC}\mathbf{1} \equiv \mathrm{A}) = \sqrt{((\mathbf{1}-\mathbf{1})^{\mathbf{2}}+(\mathbf{1}-\mathbf{1})^{\mathbf{2}})} = \mathbf{0}$$

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC}\mathbf{2} \equiv \mathrm{B}) = \sqrt{((\mathbf{1}-\mathbf{2})^{\mathbf{2}}+(\mathbf{1}-\mathbf{1})^{\mathbf{2}})} = \mathbf{1}$$

$$\boldsymbol{d(B, tamC1 \equiv A) = ?}$$

$$\boldsymbol{d(B, tamC2 \equiv B) = ?}$$

28

Bài tập

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

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

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

$$\boldsymbol{d}(\mathrm{B}, \mathrm{tamC}\mathbf{1} \equiv \mathrm{A}) = \sqrt{((\mathbf{1}-\mathbf{2})^2+(\mathbf{1}-\mathbf{1})^2)} = \mathbf{1}$$

$$\boldsymbol{d}(\mathrm{B}, \mathrm{tamC}\mathbf{2} \equiv \mathrm{B}) = \sqrt{((\mathbf{1}-\mathbf{1})^2+(\mathbf{1}-\mathbf{1})^2)} = \mathbf{0}$$

29

Bài tập

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

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

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

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

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

30

Bài tập

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

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

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

31

Bài tập

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

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

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

32

Bài tập

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

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

	Tâm c1 (A)	Tâm c2 (B)	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

33

Bài tập

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

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

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

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

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

34

Bài tập

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

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

	Tâm c1 (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)

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC1} \equiv \mathrm{A}) = \sqrt{((1-1)^2+(1-1)^2)} = 0$$

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC2} \equiv \mathrm{B}) = \sqrt{((1-11/3)^2+(1-8/3)^2)} = ?$$

$$d(B, tamC1 \equiv A) = ?$$

$$d(B, tamC2 \equiv B) = ?$$

35

Bài tập

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

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

	Tâm c1 (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)

$$d(B, tamC1 \equiv A) = ?$$

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC1} \equiv \mathrm{A}) = \sqrt{((1-1)^2+(1-1)^2)} = 0$$

$$d(B, tamC2 \equiv B) = ?$$

$$\boldsymbol{d}(\mathrm{A}, \mathrm{tamC2} \equiv \mathrm{B}) = \sqrt{((1-11/3)^2+(1-8/3)^2)} = 3.14$$

36

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

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

37

Bài tập

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

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	?

Bài tập

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

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

	Tâm c1 (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

39

Bài tập

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

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

40

Bài tập

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

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

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

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

41

Bài tập

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

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

	Tâm c1 (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	?

42

Bài tập

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

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

	Tâm c1 (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

=> Tính lại trọng tâm c1 và c2???

43

Bài tập

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

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

	Tâm c1 (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

=> Tính lại trọng tâm c1 và c2???

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

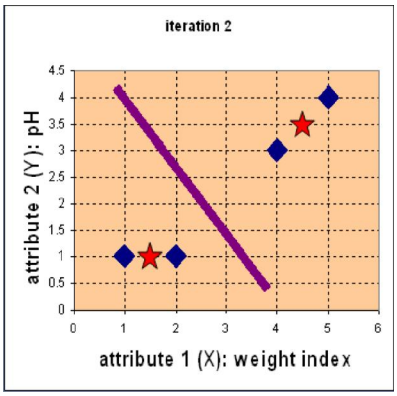
$$c_2=\left(\frac{4+5}{2},\ \frac{3+4}{2}\right)=(4\frac{1}{2},\ 3\frac{1}{2})$$

44

Bài tập

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

	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



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

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

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

45

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:**

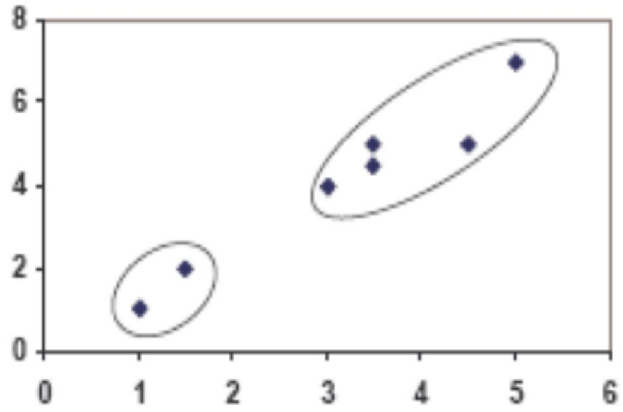
  Nhóm:

  {1,2} và {3,4,5,6,7}

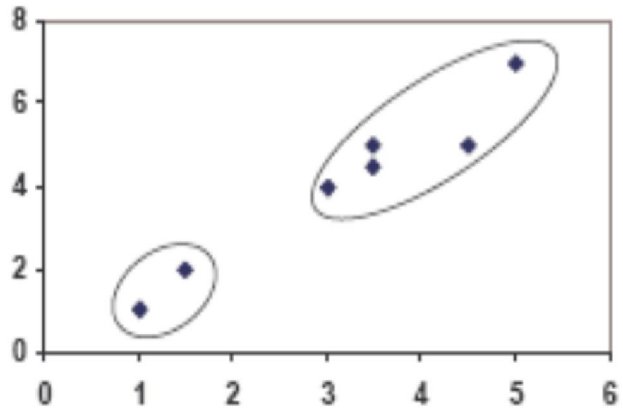
- **=> các thành viên trong nhóm không thay đổi => giải thuật dừng, ta có 2 nhóm {1,2} và {3,4,5,6,7}.**

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

53

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

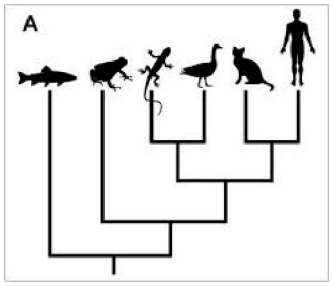
Nội dung

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

55

Hierarchical Clustering

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

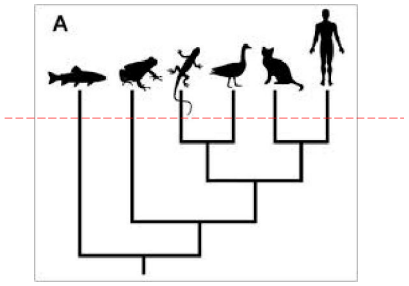


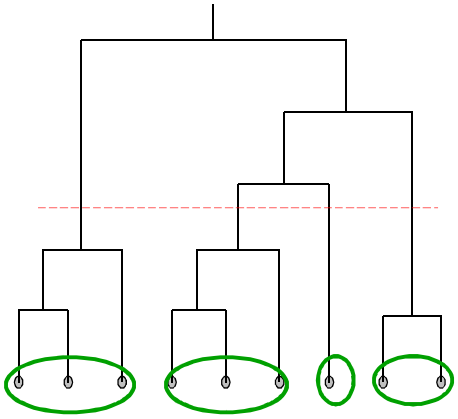
- Dựa trên điểm cắt ở đâu mà ta thu được các cụm tương ứng

56

Hierarchical Clustering

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





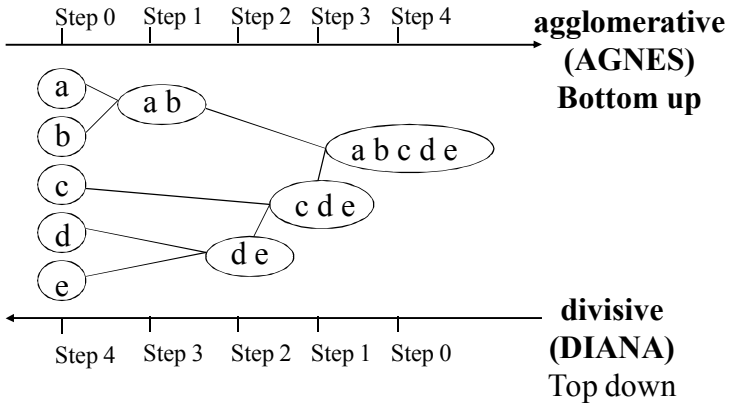
57

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

58

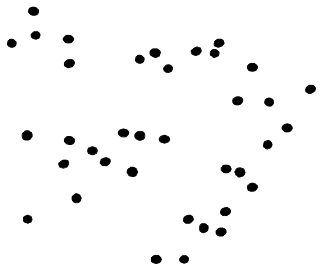
Hierarchical clustering



59

Hierarchical clustering (Single link)

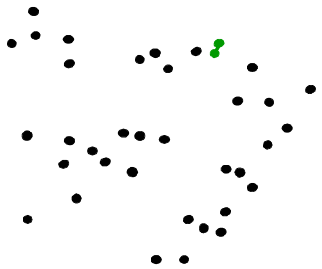
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



①Khởi 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

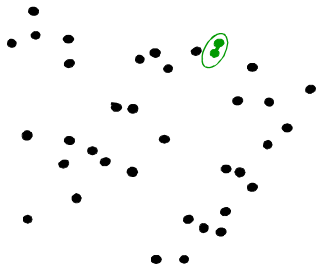


① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

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

Hierarchical clustering (Single link)

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

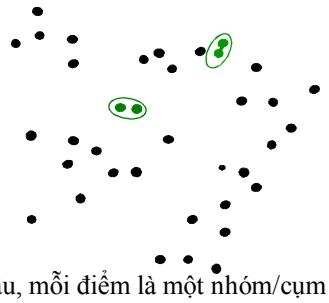


① Khởi đầu, mỗi điểm là một nhóm/cụm riêng biệt

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

Hierarchical clustering (Single link)



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

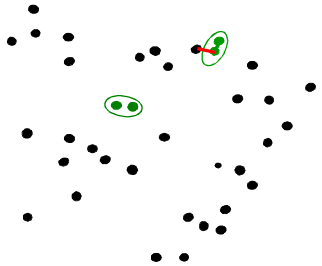
② Tìm "khoảng cách" tương tự nhất giữa các cặp cụm

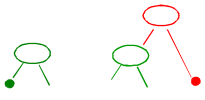
③ Kết hợp từng 2 cặp điểm thành một cụm mẹ/cụm lớn hơn

④ Lặp lại…

- 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

Hierarchical clustering (Single link)

65

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

Hierarchical clustering (Single link)

66

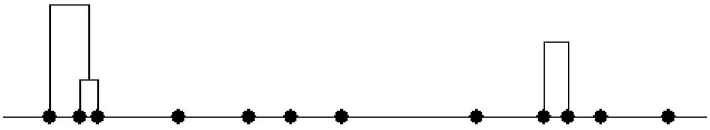
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)

67

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

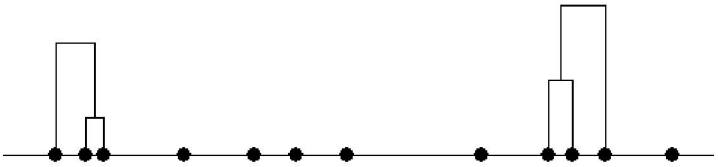
Hierarchical clustering (Single link)



68

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

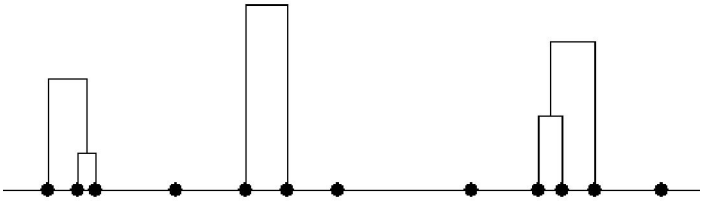
Hierarchical clustering (Single link)



69

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

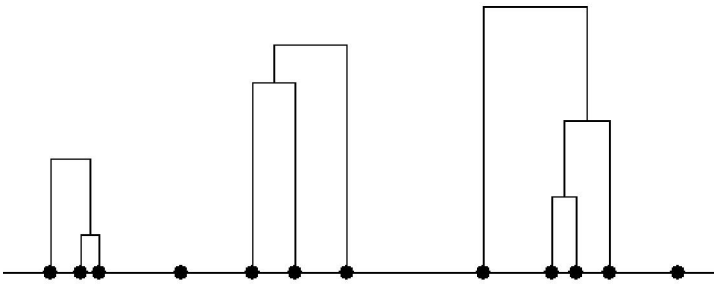
Hierarchical clustering (Single link)



70

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

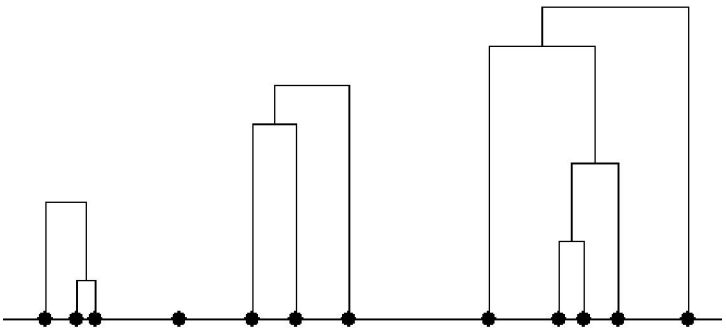
Hierarchical clustering (Single link)



71

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

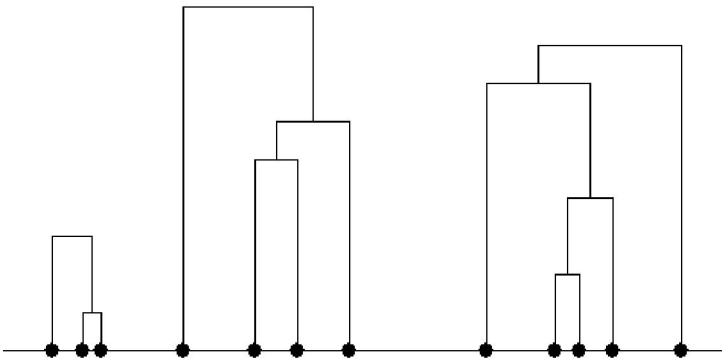
Hierarchical clustering (Single link)



72

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

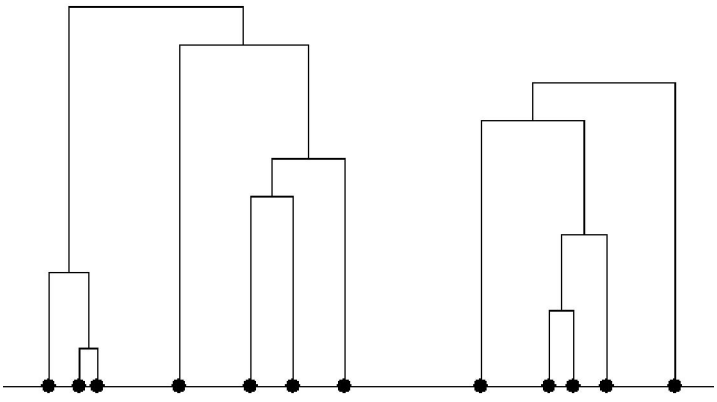
Hierarchical clustering (Single link)



73

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

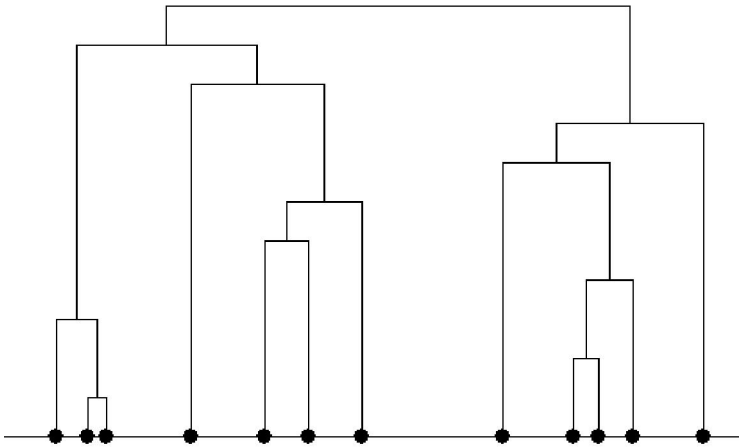
Hierarchical clustering (Single link)



74

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

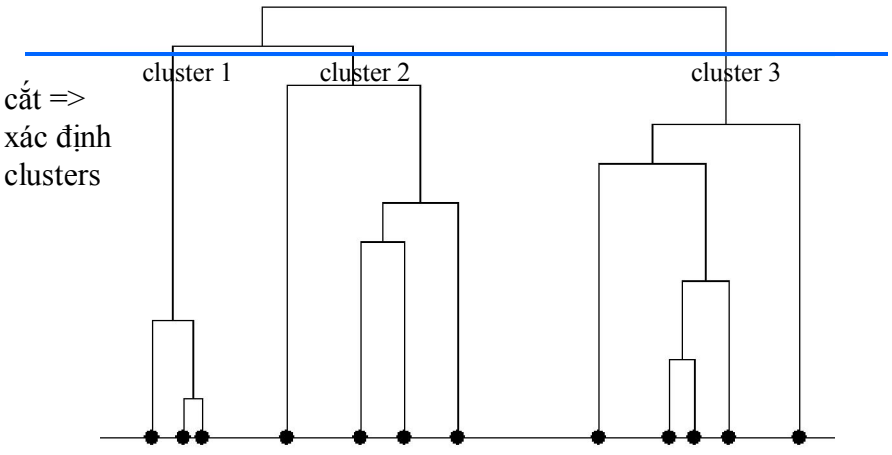
Hierarchical clustering (Single link)



75

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

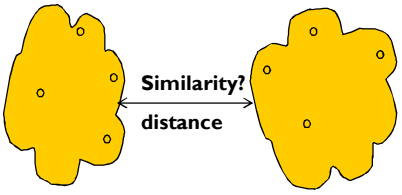
Hierarchical clustering (Single link)

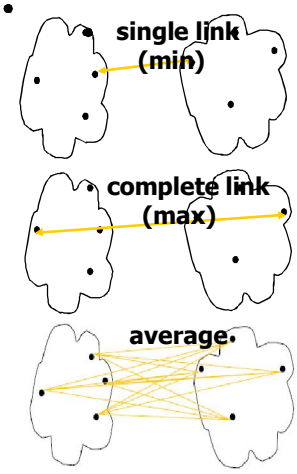


cắt => xác định clusters

76

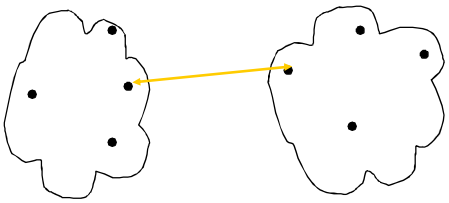
Hierarchical clustering





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

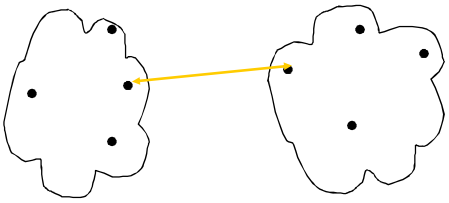
Hierarchical clustering

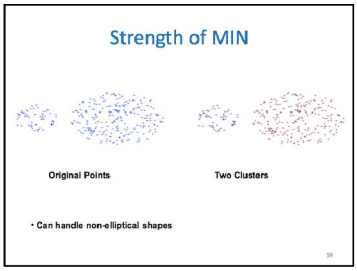


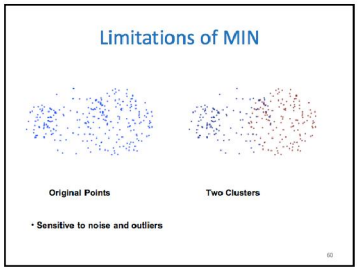
- **distance = shortest distance, nearest neighbor clustering algorithm**
- **MIN Linkage**
- **MAX Linkage**
- **Group Average**

Hierarchical clustering

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

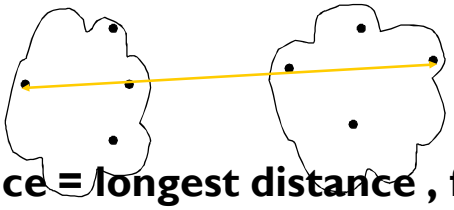






- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

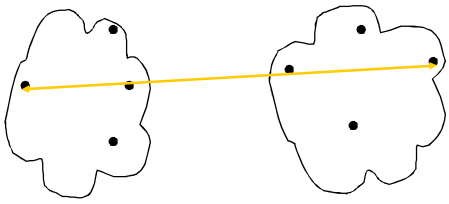
Hierarchical clustering

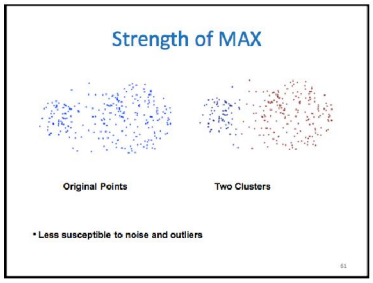


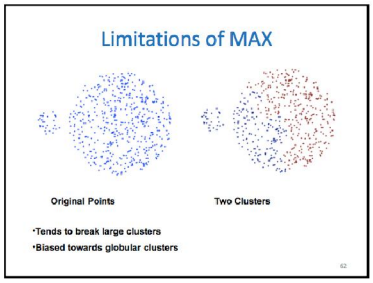
- **distance = longest distance , farthest neighbor clustering algorithm**
- **MIN Linkage**
- **MAX Linkage**
- **Group Average Linkage**

Hierarchical clustering

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

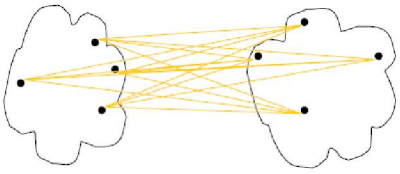






- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering



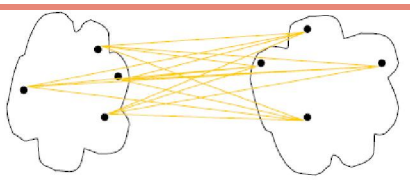
***average-link* clustering, distance = average distance**

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

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

84

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

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

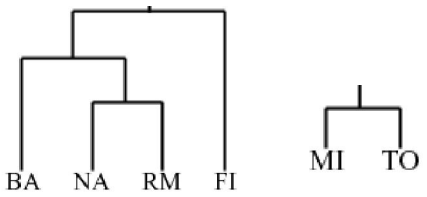
Bước 3

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

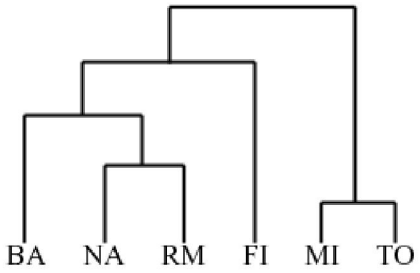
min d(i,j) = d(BA/NA/RM,FI) = 268
=> Gom cụm BA/NA/RM vào FI tạo thành nhóm mới gọi là BA/FI/NA/RM
L(BA/FI/NA/RM) = 268
m = 4



	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

Hierarchical clustering

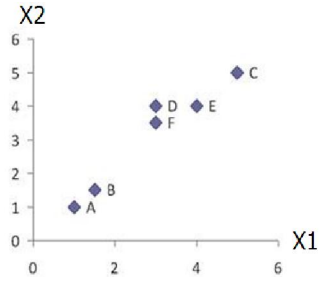
Bài tập

Sử dụng phương pháp Hierarchical clustering (Single link) để gom nhóm các phần tử A,B,C,D,E,F với thông tin được cho bên dưới sử dụng khoảng cách Euclidean

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Example

- **Problem: clustering analysis with agglomerative algorithm**



	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

data matrix

$$d_{AB} = \left((1-1.5)^2 + (1-1.5)^2\right)^{\frac{1}{2}} = \sqrt{\tfrac{1}{2}} = 0.7071$$

$$d_{DF} = \left((3-3)^2 + (4-3.5)^2\right)^{\frac{1}{2}} = 0.5$$

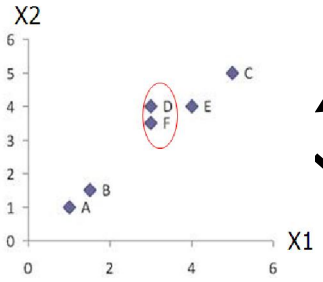
Euclidean distance

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

distance matrix

Example

- **Merge two closest clusters (iteration 1)**



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Example

- **Update distance matrix (iteration 1)**

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

$d_{(D,F)\to A} = \min\{d_{DA}, d_{FA}\} = \min\{3.61, 3.20\} = 3.20$

$d_{(D,F)\to B} = \min\{d_{DB}, d_{FB}\} = \min\{2.92, 2.50\} = 2.50$

$d_{(D,F)\to C} = \min\{d_{DC}, d_{FC}\} = \min\{2.24, 2.50\} = 2.24$

$d_{E\to(D,F)} = \min\{d_{ED}, d_{EF}\} = \min\{1.00, 1.12\} = 1.00$

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Min Distance (Single Linkage)

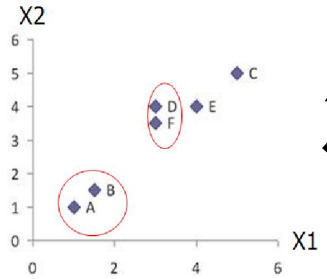
Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	**3.20**	**2.50**	**2.24**	0.00	1.00
E	4.24	3.54	1.41	**1.00**	0.00

Example

- **Merge two closest clusters (iteration 2)**

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	**3.20**	**2.50**	**2.24**	0.00	1.00
E	4.24	3.54	1.41	**1.00**	0.00



Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0

Example

- **Update distance matrix (iteration 2)**

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	**3.20**	**2.50**	**2.24**	0.00	1.00
E	4.24	3.54	1.41	**1.00**	0.00

$d_{C\to(A,B)} = \min\left(d_{CA}, d_{CB}\right) = \min\left(5.66, 4.95\right) = 4.95$

$d_{(D,F)\to(A,B)} = \min\left(d_{DA}, d_{DB}, d_{FA}, d_{FB}\right) = \min\left(3.61, 2.92, 3.20, 2.50\right) = 2.50$

$d_{E\to(A,B)} = \min\left(d_{EA}, d_{EB}\right) = \min\left(4.24, 3.54\right) = 3.54$

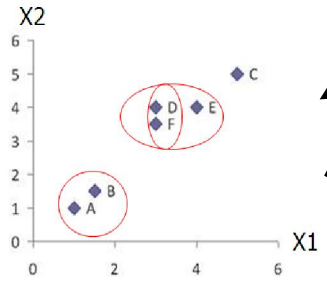
Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0

Min Distance (Single Linkage)

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	**4.95**	0	2.24	1.41
(D, F)	**2.50**	2.24	0	1.00
E	**3.54**	1.41	1.00	0

Example

- **Merge two closest clusters/update distance matrix (iteration 3)**



Min Distance (Single Linkage)

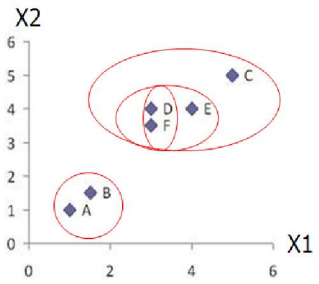
Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	**4.95**	0	2.24	1.41
(D, F)	**2.50**	2.24	0	1.00
E	**3.54**	1.41	1.00	0

Min Distance (Single Linkage)

Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	**2.50**	**1.41**	0.00

Example

- **Merge two closest clusters/update distance matrix (iteration 4)**



Min Distance (Single Linkage)

Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00

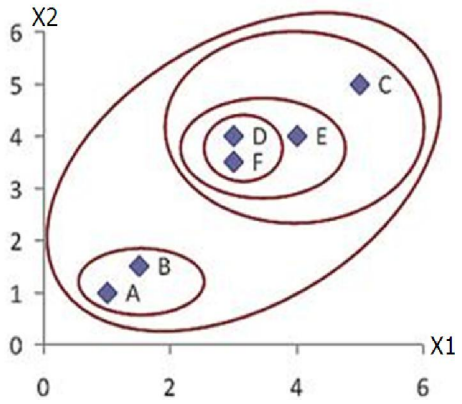
Min Distance (Single Linkage)

Dist	(A,B)	((D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

Example

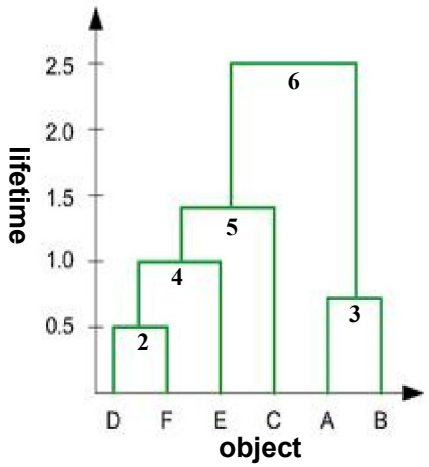
- **Final result (meeting termination condition)**

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



Example

- **Dendrogram tree representation**



1. **In the beginning we have 6 clusters: A, B, C, D, E and F**
2. **We merge clusters D and F into cluster (D, F) at distance 0.50**
3. **We merge cluster A and cluster B into (A, B) at distance 0.71**
4. **We merge clusters E and (D, F) into ((D, F), E) at distance 1.00**
5. **We merge clusters ((D, F), E) and C into (((D, F), E), C) at distance 1.41**
6. **We merge clusters (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B) at distance 2.50**
7. **The last cluster contain all the object thus conclude the computation**

Nội dung

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

Giải thuật clustering

- 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)

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 !