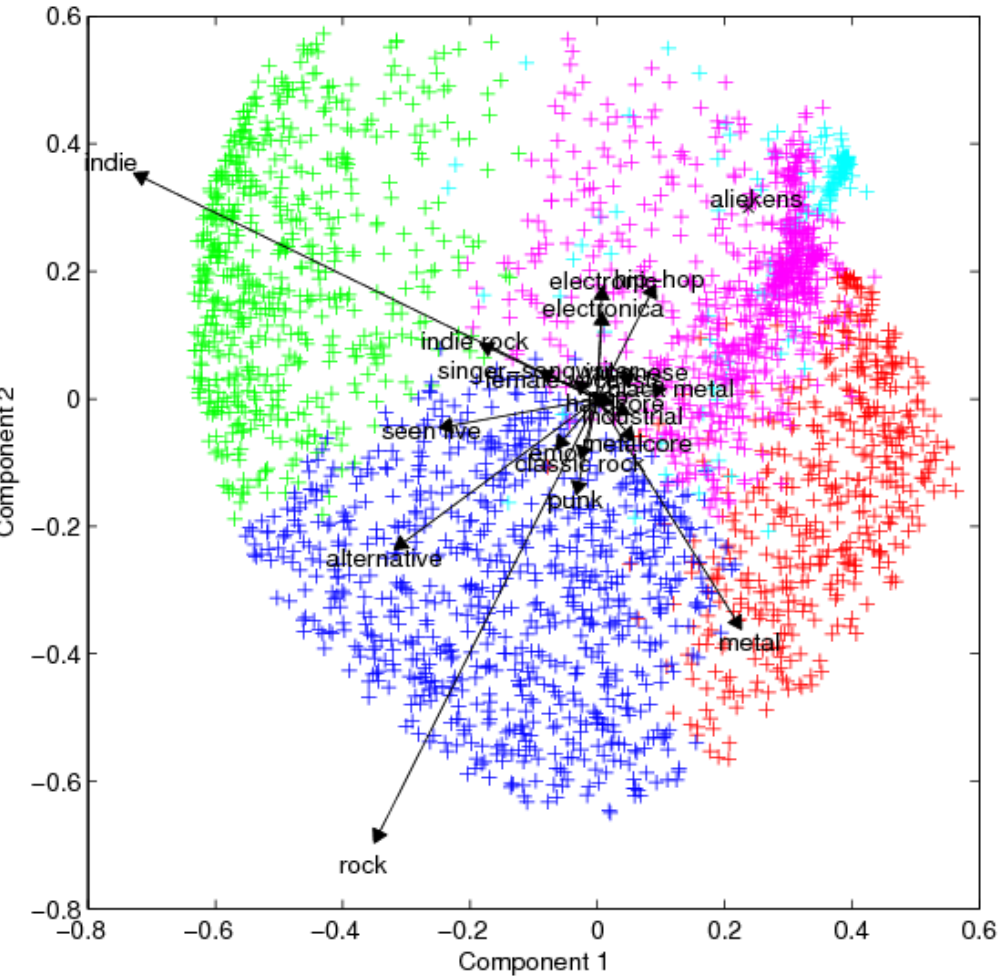
Chương 5

Gom cụm dữ liệu
Data Clustering



Giới thiệu

- Sự bùng nổ thông tin hiện nay do tác động của các siêu phương tiện và WWW.
- Các hệ thống truy vấn thông tin dựa trên việc phân nhóm, gom cụm (clustering) ra đời để làm tăng tốc độ tìm kiếm thông tin.
- Do sự biến động thường xuyên của thông tin nên các thuật toán clustering đang tồn tại không thể duy trì tốt các nhóm, cụm (cluster) trong một môi trường như thế.
- Vấn đề đặt ra là làm thế nào để cập nhật các cluster trong hệ thống mỗi khi thông tin được cập nhật thay vì phải thường xuyên clustering lại toàn bộ dữ liệu?

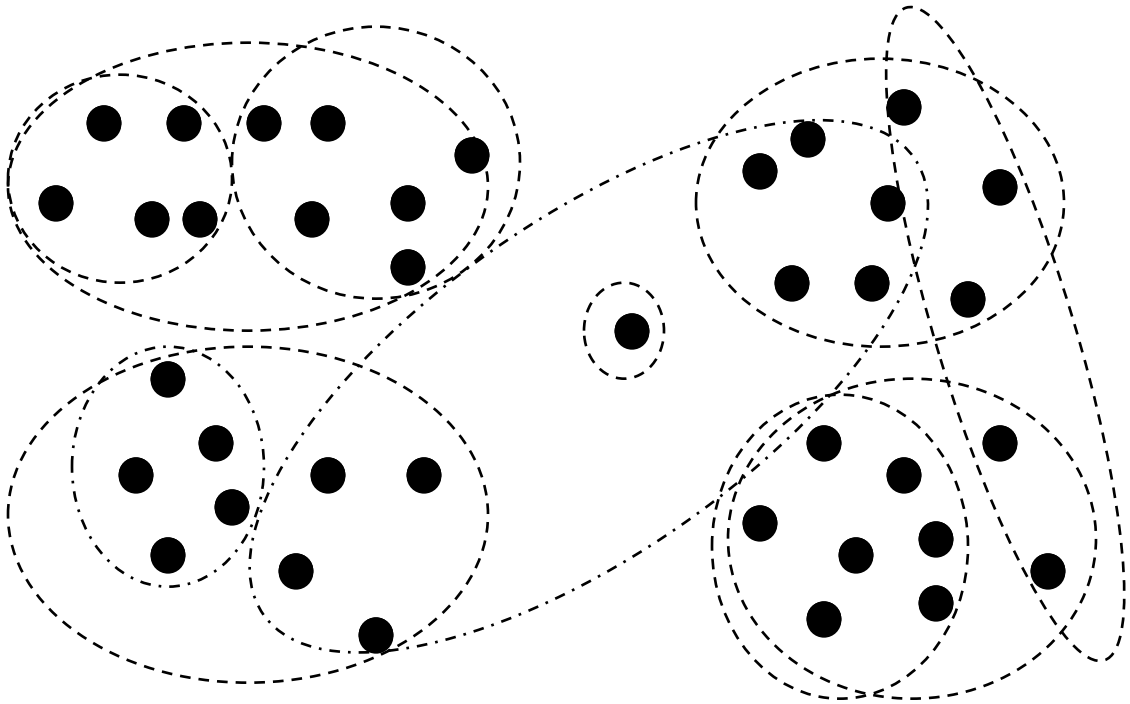
Giới thiệu

Gom cụm (clustering) là quá trình nhóm tập đối tượng thành các cụm (cluster) có các đối tượng giống nhau.

Cho CSDL D={t1,t2,…,tn} và số nguyên k, gom cụm là bài toán xác định ánh xạ f: Dg{1,…,k} sao cho mỗi ti được gán vào một cụm (lớp) Kj, 1 <= j <= k .

Không giống bài toán phân lớp, các cụm không được biết trước.

Giới thiệu



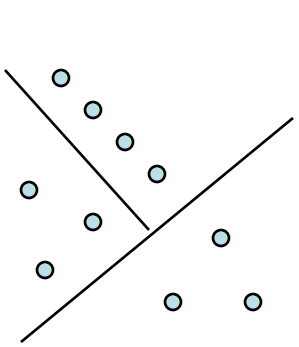
Dựa trên khoảng cách

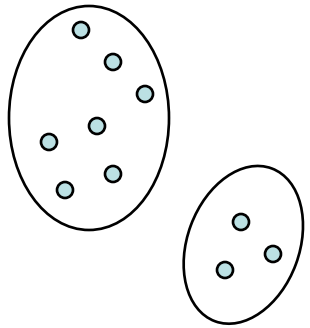
4

Giới thiệu

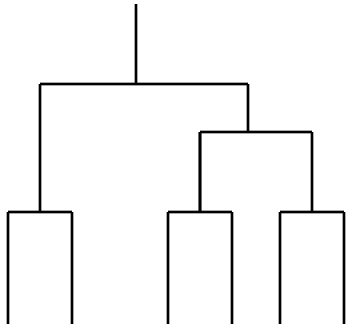
Cách biểu diễn các cụm

- Phân chia bằng các đường ranh giới
- Các khối cầu
- Theo xác suất
- Hình cây
- …





	1	2	3
I1	0.5	0.2	0.3
I2			
…			
In			



5

Mở đầu

Gom cụm dữ liệu là hình thức học không giám sát, trong đó các mẫu học chưa được gán nhãn.

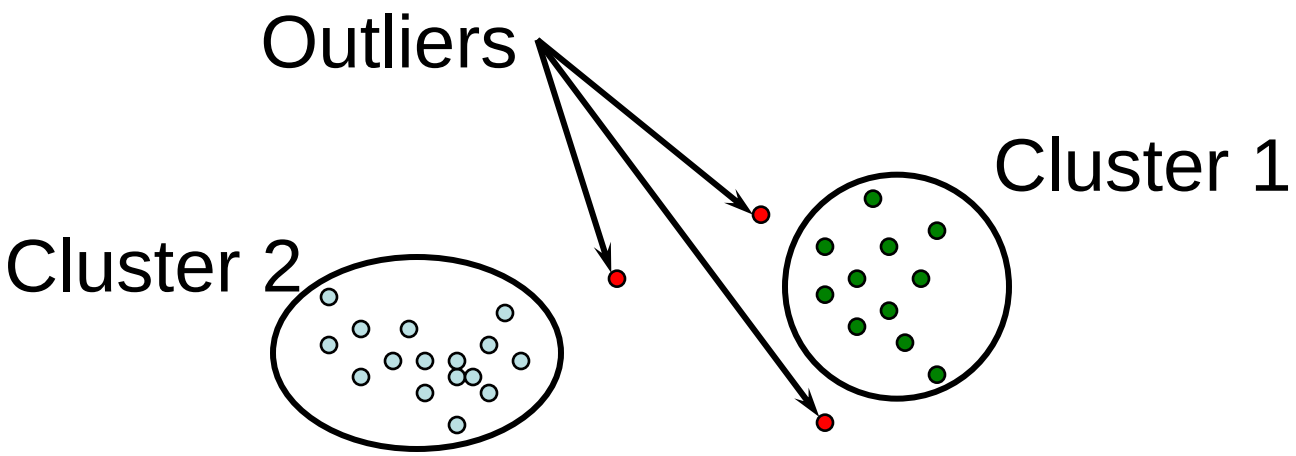
Mục đích của gom cụm dữ liệu là tìm những mẫu đại diện hoặc gom cụm tương tự nhau (theo một tiêu chuẩn nào đó) thành các cụm

Định nghĩa: Gom cụm là quá trình xây dựng một tập hợp từ một tập dữ liệu mẫu, các phần tử trong tập đã gom cụm tương tự nhau về một vài thuộc tính chọn trước.

What Is Clustering?

Group data into clusters

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters
- Unsupervised learning: no predefined classes



Application Examples

A stand-alone tool: explore data distribution

A preprocessing step for other algorithms

Pattern recognition, spatial data analysis, image processing, market research, WWW, …

- Cluster documents
- Cluster web log data to discover groups of similar access patterns

Thế nào là PP gom cụm tốt?

- Có độ tương tự cao trong cùng cụm (intra-class)
- Có độ tương tự thấp giữa các cụm (inter-class)
- Khả năng phát hiện mẫu ẩn (hidden patterns)
- Có khả năng làm việc hiệu quả với mẫu lớn (scalability)
- Khả năng làm việc với nhiều loại dữ liệu khác nhau
- ….

Ma trận dữ liệu (Data Matrix)

- Dùng để mô hình hóa bài toán gom cụm
- Ma trận biểu diễn không gian dữ liệu gồm n đối tượng theo p thuộc tính
- Ma trận biểu diễn mối quan hệ đối tượng theo thuộc tính:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Ma trận phân biệt (Dissimilarity Matrix)

Biểu diễn khoảng cách giữa 2 điểm (đối tượng) trong không gian dữ liệu gồm n đối tượng theo p thuộc tính

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

3. Độ đo khoảng cách

Distances are normally used measures

Minkowski distance: a generalization

$$d(i,j) = \sqrt[q]{|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q} \quad (q > 0)$$

If q = 2, d is Euclidean distance

If q = 1, d is Manhattan distance

Weighed distance

$$d(i,j) = \sqrt[q]{w_1 |x_{i_1} - x_{j_1}|^q + w_2 |x_{i_2} - x_{j_2}|^q + \ldots + w_p |x_{i_p} - x_{j_p}|^q) \quad (q > 0)}$$

Properties of Minkowski Distance

Nonnegative: $d(i,j) \geq 0$

The distance of an object to itself is 0

- $d(i,i) = 0$

Symmetric: $d(i,j) = d(j,i)$

Triangular inequality

- $d(i,j) \leq d(i,k) + d(k,j)$

Các phương pháp phân cụm (Categories of Clustering Approaches)

Thuật toán phân hoạch (Partitioning algorithms)

Phân hoạch cơ sở dữ liệu D có n đối tượng thành k cụm:

- Mỗi cụm có ít nhất 1 đối tượng
- Mỗi đối tượng thuộc về 1 cụm duy nhất
- K là sô 1cụm cho trước

Thuật toán phân cấp (Hierarchy algorithms)

- Gộp:
  - Xuất phát mỗi đối tượng và tạo một cụm chứa nó.
  - Nếu 2 cum gần nhau thì gộp thành 1 cụm
  - Lặp lại bước 2 cho đến khi còn 1 cụm duy nhất là toàn bộ không gian
- Tách:
  - Xuất phát từ 1 cụm duy nhất là toàn bộ không gian
  - Chọn cụm có độ phân biệt cao nhất (ma trận phân biệt có phần tử lớn nhất hoặc giá trị trung bình lớn nhất) để tách đôi
  - Lặp lại bước 2 cho đến khi mỗi đối tượng thuộc 1 cụm hoặc đạt điều kiện dừng (đủ số cụm hoặc khoảng cách giữa các cụm đủ nhỏ)

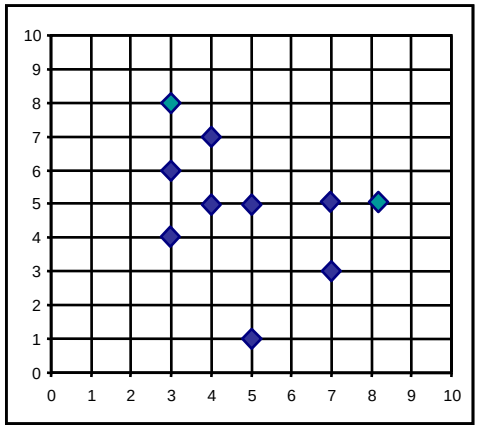
Các phương pháp phân cụm (tiếp)

- Phương pháp dựa trên mật độ (Density-based methods)
- Phương pháp dựa trên lưới (Grid-based methods)
- Phương pháp dựa trên mô hình (Model-based)

4 Thuật toán K-means

- Phân hoạch n đối tượng thành k cụm
- Thuật toán K-means gồm 4 bước:
  - Chọn ngẫu nhiên k điểm làm trọng tâm ban đầu
  - Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần điểm đang xét.
  - Nếu không có phép gán nào thì dừng (các cụm đã ổn định và thuật toán không cải thiện thêm được nữa)
  - Tính trọng tâm cho từng cụm
  - Quay lại bước 2.

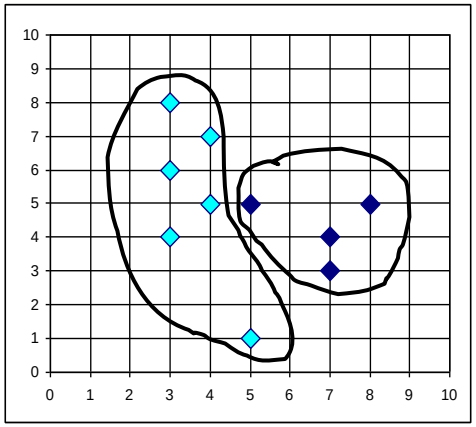
K-Means: Example



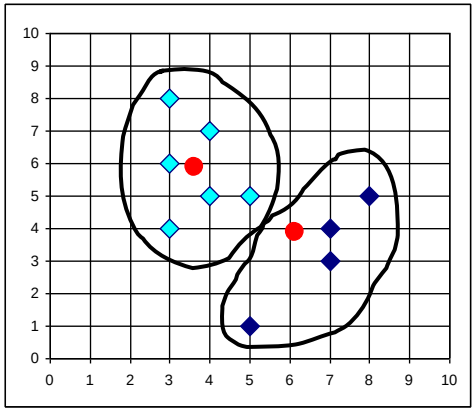
K=2

Arbitrarily choose K object as initial cluster center

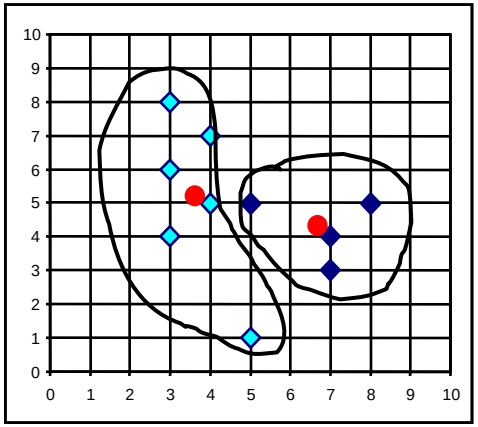
Assign each objects to most similar center



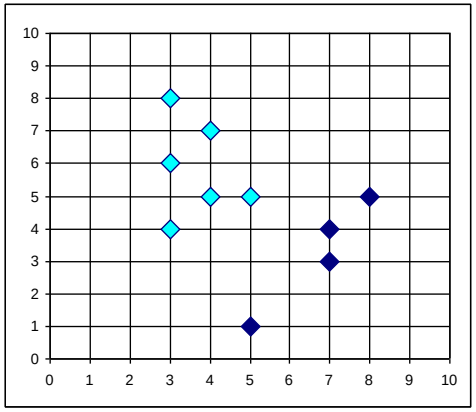
reassign



Update the cluster means



reassign



Update the cluster means

Example

Cho tập điểm:

X_1={1,3} = {$x_{11,}$, x_{12}}

X_2={1.5,3.2} = {$x_{21,}$, x_{22}}

X_3={1.3,2.8} = {$x_{31,}$, x_{32}}

X_4={3,1} = {$x_{41,}$, x_{42}}

Dùng K-mean phân cụm với k=2

Bước 1: Khởi tạo ma trận phân hoạch U (2 rows and 4 columns)

Bước 2: U=(m_{ij}) 1<=i<=2; 1<=j<=4

Cho n=0 (số lần lặp) tạo U_0

Every column has 1 bit number 1

		x1	x2	x3	x4
U0=	C1	1	0	0	0
	C2	0	1	1	1

Bước 3: Tính các vector trọng tâm

Do có 2 cụm C1, C2 do đó có 2 vector trọng tậm

Vector v_1 for cluster C_1:

$$v11 = \frac{m11 * x11 + m12 * x21 + m13 * x31 + m14 * x41}{m11 + m12 + m13 + m14}$$

$$= \frac{1*1 + 0*0.5 + 0*1.3 + 0*3}{1 + 0 + 0 + 0} = 1$$

$$v12 = \frac{m11 * x12 + m12 * x22 + m13 * x32 + m14 * x42}{m11 + m12 + m13 + m14}$$

$$= \frac{1*3 + 0*3.2 + 0*2.8 + 0*1}{1 + 0 + 0 + 0} = 3$$

$$v1 = (1,3)$$

Vector v_2 for cluster C_2:

$$v21 = \frac{m21 * x11 + m22 * x21 + m23 * x31 + m24 * x41}{m21 + m22 + m23 + m24}$$

$$= \frac{0 * 1 + 1 * 1.5 + 1 * 1.3 + 1 * 3}{0 + 1 + 1 + 1} = \frac{5.8}{3} = 1.93$$

$$v22 = \frac{m21 * x12 + m22 * x22 + m23 * x32 + m24 * x42}{m21 + m22 + m23 + m24}$$

$$= \frac{0 * 3 + 1 * 3.2 + 1 * 2.8 + 1 * 1}{0 + 1 + 1 + 1} = \frac{7}{3} = 2.33$$

$$v1 = (1.93 \ .2.33)$$

Computing the Euclide distance for all points to clustering:

$$d(x1, v1) = \sqrt{(x11 - v11)^2 + (x12 - v12)^2}$$

$$= \sqrt{(1-1)^2 + (3-3)^2} = 0$$

$$d(x1, v2) = \sqrt{(x11 - v21)^2 + (x12 - v22)^2}$$

$$= \sqrt{(1-1.93)^2 + (3-2.33)^2} = 1.14$$

Gộp x1 vào cụm C1 vì d(x1,v1) < d(x1,v2)

Tương tự:

d(x2,v1)=0.54 < 0.97 = d(x2,v2) gộp x2 vào C1

d(x3,v1)=0.36 < 0.78 = d(x3,v2) gộp x3 vào C1

d(x4,v1)=2.83 > 1.70 = d(x4,v2) gộp x4 vào C1

Tăng n lên 1

Ma trận phân hoạch U sẽ là:

		x1	x2	x3	x4
U0=	C1	1	1	1	0
	C2	0	0	0	1

Lặp lại cho đến khi

Không có phép gán nào thì dừng, nếu sai quay lại bước 3

2) Fuzzy C-means

Thuật toán K-means phân hoạch tập dữ liệu thành các cụm là các tập rõ.

Phân hoạch mờ xem các cụm là các tập mờ và 2 điểm dữ liệu sẽ có mức đội thuộc về một cụm với giá trị trong [0,1].

Thuật toán Fuzzy C-means cực tiểu hàm mục tiêu:

$$J = \sum_{j=1}^{c} \sum_{i=1}^{n} (\mu_{ij})^m \, d(x_i^{(j)}, C_i)^2$$

2) Fuzzy C-means

Thuật toán Fuzzy C-means cực tiểu hàm mục tiêu:

$$J = \sum_{j=1}^{c} \sum_{i=1}^{n} (\mu_{ij})^m \, d(x_i^{(j)}, C_i)^2$$

Trong đó:

- μ_{ij} là phần tử hàng i cột j của ma trận thành viên U, biểu diễn độ thuộc của x_j vào cụm j (có C_j là trọng tâm)
- m>1 là tham số mờ hóa (m điều chỉnh độ thuộc về của 1 điểm vào cụm tương ứng, thông thường chọn m=2)

Thuật toán Fuzzy C-means

Không gian dữ liệu gồm n điểm x_i i=1,n. Cần phân hoạch thành c cụm (2<= c <n) Thuật toán Fuzzy C-means gồm các bước:

1. Chon tham số mờ hóa m>1
2. Khởi tạo ma trận thành viên U_{nxc} với 0<= μ_{ij} <=1 sao cho

$$\sum_{j=1}^{n} \mu_{ij} = 1$$

3. Tính trọng tâm C_j của cum j (j=1..c) $C_j = \dfrac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}$

Thuật toán Fuzzy C-means(tiếp)

4. Cập nhật ma trận khoảng cách D_{nxc} theo độ đo khoảng cách đã chọn (d_{ij} là khoảng cách từ x_i đến C_j)
5. Cập nhật ma trận thành viên U:

   Nếu $d_{ij} > 0$ thì

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}} \right]^{-1}$$

   Ngược lại, nếu $d_{ij}=0$ thì x_{ij} trùng với trọng tâm C_j của cụm j, $\mu_{ij} = 1$
6. Nếu U thay đổi là đủ nhỏ so với bước trước thì dừng, ngược lại lặp lại từ bước 3.

Example

Cho tập điểm:

X_1={1,3} = {$x_{11,}$, x_{12}}

X_2={1.5,3.2} = {$x_{21,}$, x_{22}}

X_3={1.3,2.8} = {$x_{31,}$, x_{32}}

X_4={3,1} = {$x_{41,}$, x_{42}}

Dùng Fuzzy C-mean phân cụm với k=2

Phân hoạch mờ ban đầu $U^{(0)}$, giả sử m=2 và tiêu chuẩn hội tụ ε=0.01.

Phân hoạch mờ ban đầu là:

U0=	1	0	0	0
	0	1	1	1

Tính trọng tâm ban đầu bắng công thức sau với m=2

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{ki}}{\sum_{k=1}^{n} (\mu_{ik})^m}$$

a. Với cụm 1(c=1)

Vector v_1 cho cụm 1

$$v_{1j} = \frac{\mu_1^2 * x_{1j} + \mu_2^2 * x_{2j} + \mu_3^2 * x_{3j} + \mu_4^2 * x_{4j}}{\mu_1^2 + \mu_2^2 + \mu_3^2 + \mu_4^2}$$

$$= \frac{1 * x_{1j} + 1 * x_{2j} + 1 * x_{3j} + 0 * x_{4j}}{1 + 1 + 1 + 0}$$

$$v_{11} = \frac{x_{1j} + x_{2j} + x_{3j}}{3} = \frac{1 + 1.5 + 1.3}{3} = 1.26$$

$$v_{12} = \frac{3 + 3.2 + 3.8}{3} = 3.0$$

b. Với cụm 2(c=2)

Vector v_1 cho cụm 1

$$v_{2j} = \frac{x_{4j}}{0^2 + 0^2 + 0^2 + 1^2}$$

$$v_{21} = 3,\ v_{22} = 1$$

$$v_2 = \{3,1\}$$

Tiếp theo tính khoảng cách từng điểm đến trọng tâm của nó. Tính khoảng cách từng điểm đến các cụm C1,C2 và chọn cụm có khoảng cách nhỏ nhất để đưa đối tượng vào cụm

b. Với cụm 2(c=2) (tiếp)

$$d_{11} = \sqrt{(1-1.26)^2 + (3-3)^2} = 0.26$$

$$d_{12} = \sqrt{(1.5-1.26)^2 + (3.2-3)^2} = 0.31$$

$$d_{13} = \sqrt{(1.3-1.26)^2 + (2.8-3)^2} = 0.20$$

$$d_{14} = \sqrt{(3-1.26)^2 + (1-3)^2} = 2.65$$

$$d_{21} = \sqrt{(1-3)^2 + (3-1)^2} = 2.82$$

$$d_{22} = \sqrt{(1.5-3)^2 + (3.2-1)^2} = 2.66$$

$$d_{23} = \sqrt{(1.3-3)^2 + (2.8-1)^2} = 2.47$$

$$d_{24} = \sqrt{(3-3)^2 + (1-1)^2} = 0.0$$

Với độ đo khoảng cách, cập nhật U

$$\mu_{ik}^{(r+1)} = \left[\sum_{k=1}^{c} \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right) \right]^{-1} \quad (PT \quad 6.1)$$

$$\mu_{11} = \left[\sum_{j=1}^{c} \left(\frac{d_{11}}{d_{j1}} \right)^2 \right]^{-1} = \left[\left(\frac{d_{11}}{d_{11}} \right)^2 + \left(\frac{d_{11}}{d_{21}} \right)^2 \right]^{-1} = \left[\left(\frac{0.26}{0.36} \right)^2 + \left(\frac{0.26}{2.82} \right)^2 \right]^{-1} = 0.991$$

$$\mu_{12} = \left[\left(\frac{d_{12}}{d_{12}} \right)^2 + \left(\frac{d_{12}}{d_{22}} \right)^2 \right]^{-1} = \left[1 + \left(\frac{0.31}{2.66} \right)^2 \right]^{-1} = 0.968$$

$$\mu_{13} = \left[\left(\frac{d_{13}}{d_{13}} \right)^2 + \left(\frac{d_{13}}{d_{23}} \right)^2 \right]^{-1} = \left[1 + \left(\frac{0.20}{2.47} \right)^2 \right]^{-1} = 0.993$$

$$\mu_{14} = \left[\left(\frac{d_{14}}{d_{14}} \right)^2 + \left(\frac{d_{14}}{d_{24}} \right)^2 \right]^{-1} = \left[1 + \left(\frac{2.65}{0} \right)^2 \right]^{-1} = 0 \quad Cho \quad I_4 = 0$$

Dùng PT 6.1 chi các giá trị phân hoạch μ_{2j} (j=1..4). Các hàm thành viên mới được tạo ra từ việc cập nhât phân hoạch mờ cho bởi:

U0=	0.991	0.986	0.993	0
	0.009	0.014	0.007	1

Xét đạt tiêu chuẩn hội tụ chưa, xét chuẩn của ma trận, chẳng hạn giá trị tuyệt đối lớn nhất của từng cặp trong ma trận:

$$\max_{i,k} | \mu_{ik}^{1} - \mu_{ik}^{1} | = 0.0134 > 0.01$$

Do kết quả chưa hội tụ, tiến hành lặp lại: Tính trọng tâm với các giá trị trong ma trận phân hoạch mới

Với cụm 1(c=1)

Vector v_1 cho cụm 1

$$v_{1j} = \frac{(0.991)^2 * x_{1j} + (0.986)^2 * x_{2j} + (0.993)^2 * x_{3j} + (0)^2 * x_{4j}}{(0.991)^2 + (0.986)^2 + (0.993)^2 + (0)^2}$$

$$v_{11} = \frac{(0.991)^2 * 1 + (0.986)^2 * 1.5 + (0.993)^2 * 1.3}{2.94} = \frac{3.17}{2.94} = 1.26$$

$$v_{11} = \frac{(0.991)^2 * 3 + (0.986)^2 * 3.2 + (0.993)^2 * 2.8}{2.94} = \frac{8.816}{2.94} = 3.0$$

$$V_1 = (1.26, 3)$$

Với cụm 2(c=2)

$$v_{2j} = \frac{(0.009)^2 * x_{1j} + (0.014)^2 * x_{2j} + (0.007)^2 * x_{3j} + (1)^2 * x_{4j}}{(0.009)^2 + (0.014)^2 + (0.007)^2 + (1)^2}$$

$$v_{21} = \frac{(0.009)^2 * 1 + (0.014)^2 * 1.5 + (0.007)^2 * 1.3 + (1)^2 * 3}{1.00} = 3.0$$

$$v_{22} = \frac{(0.009)^2 * 3 + (0.014)^2 * 3.2 + (0.007)^2 * 2.8}{2.94} = 1.0$$

$$V_2 = (3.0, 1.0)$$

Tiếp tuc cho đến khi thỏa tiêu chuẩn hội tu. Khi đó, ma trận phân hoạch sẽ xác định mội phân cụm mời các đối tượng.