K-means

THUẬT TOÁN K-MEAN VÀ ỨNG DỤNG

NỘI DUNG CHÍNH

I. Phân cụm

II. Thuật toán K-Mean

1. Khái quát về thuật toán
2. Các bước của thuật toán
3. Ví dụ minh họa – Demo thuật toán
4. Đánh giá thuật toán
5. Tổng quát hóa và Các biến thể

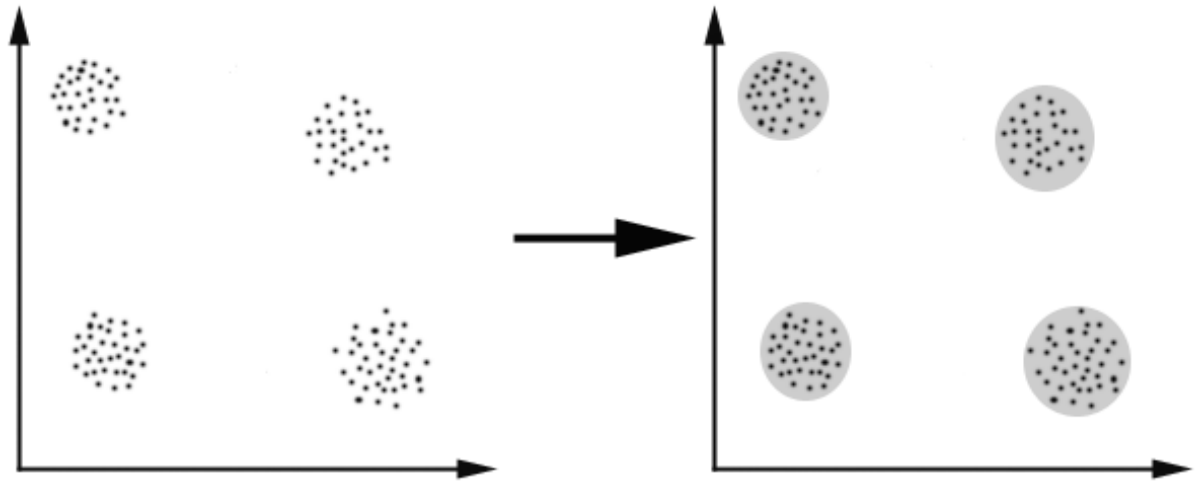
III. Ứng dụng của thuật toán K-Mean

I. PHÂN CụM

1. *Phân cụm là gì?*

- Quá trình phân chia 1 tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn:
  - Các đối tượng trong 1 cụm "tương tự" nhau.
  - Các đối tượng khác cụm thì "không tương tự" nhau.
- Giải quyết vấn đề tìm kiếm, phát hiện các cụm, các mẫu dữ liệu trong 1 tập hợp ban đầu các dữ liệu không có nhãn.

I. PHÂN CụM



Nếu X : 1 tập các điêm dữ liệu

C_i : cụm thứ i

$X = C_1 \cup \ldots \cup C_k \cup \ldots \cup C_{\text{ngoại lai}}$

$Ci \cap Cj = \varnothing$

I. PHÂN CụM

2. *<u>Một số độ đo trong phân cụm</u>*
   - Minkowski

$$\sum_{i=1}^{n} (|| x_i - y_i ||^p)^{\frac{1}{p}}$$

     - Với xi,yi là 2 vector
   - Euclidean: p = 2
   - Độ đo tương tự (gần nhau): cosin hai vectơ

$$\cos\mu = \frac{v.w}{\| v \| . \| w \|}$$

I. PHÂN CụM

3. *Mục đích của phân cụm*

- Xác định được bản chất của việc nhóm các đối tượng trong 1 tập dữ liệu không có nhãn.

- Phân cụm không dựa trên 1 tiêu chuẩn chung nào, mà dựa vào tiêu chí mà người dùng cung cấp trong từng trường hợp.

I. PHÂN CụM

5. *Một số phương pháp phân cụm điển hình*
   - Phân cụm phân hoạch
   - Phân cụm phân cấp
   - Phân cụm dựa trên mật độ
   - Phân cụm dựa trên lưới
   - Phân cụm dựa trên mô hình
   - Phân cụm có ràng buộc

II.PHÂN CụM PHÂN HOạCH

- Phân 1 tập dữ liệu có n phần tử cho trước thành k tập con dữ liệu ($k \leq n$), mỗi tập con biểu diễn 1 cụm.
- Các cụm hình thành trên cơ sở làm tối ưu giá trị hàm đo độ tương tự sao cho:
  - Các đối tượng trong 1 cụm là tương tự.
  - Các đối tượng trong các cụm khác nhau là không tương tự nhau.
- Đặc điểm:
  - Mỗi đối tượng chỉ thuộc về 1 cụm.
  - Mỗi cụm có tối thiểu 1 đối tượng.
- Một số thuật toán điển hình : K-mean, PAM, CLARA,…

II.2. Thuật toán K-Means

Phát biểu bài toán:

- Input

➢ Tập các đối tượng $X = \{x_i | i = 1, 2, \ldots, N\}$, $x_i \in R^d$

➢ Số cụm: K

- Output

➢ Các cụm C_i ($i = 1 \div K$) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.

II.1. KHÁI QUÁT Về THUậT TOÁN

- Thuật toán hoạt động trên 1 tập vectơ d chiều, tập dữ liệu X gồm N phần tử:

$$X = \{x_i \mid i = 1, 2, \ldots, N\}$$

- K-Mean lặp lại nhiều lần quá trình:
  - Gán dữ liệu.
  - Cập nhật lại vị trí trọng tâm.

- Quá trình lặp dừng lại khi trọng tâm hội tụ và mỗi đối tượng là 1 bộ phận của 1 cụm.

II.1. KHÁI QUÁT Về THUậT TOÁN

- Hàm đo độ tương tự sử dụng khoảng cách Euclidean

$$E = \sum_{i=1}^{N} \sum_{x_i \in C_j} (\| x_i - c_j \|^2)$$

trong đó c_j là trọng tâm của cụm C_j

- Hàm trên không âm, giảm khi có 1 sự thay đổi trong 1 trong 2 bước: gán dữ liệu và định lại vị trí tâm.

II.2. CÁC BƯỚC CủA THUậT TOÁN

- Bước 1 - Khởi tạo

  Chọn K trọng tâm $\{c_i\}$ (i = 1÷K).

- Bước 2 - Tính toán khoảng cách

$$S_i^{(t)} = \{ x_j : \| x_j - c_i^{(t)} \| \leq \| x_j - c_{i^*}^{(t)} \| \text{ for all } i^* = 1, \ldots, k\}$$

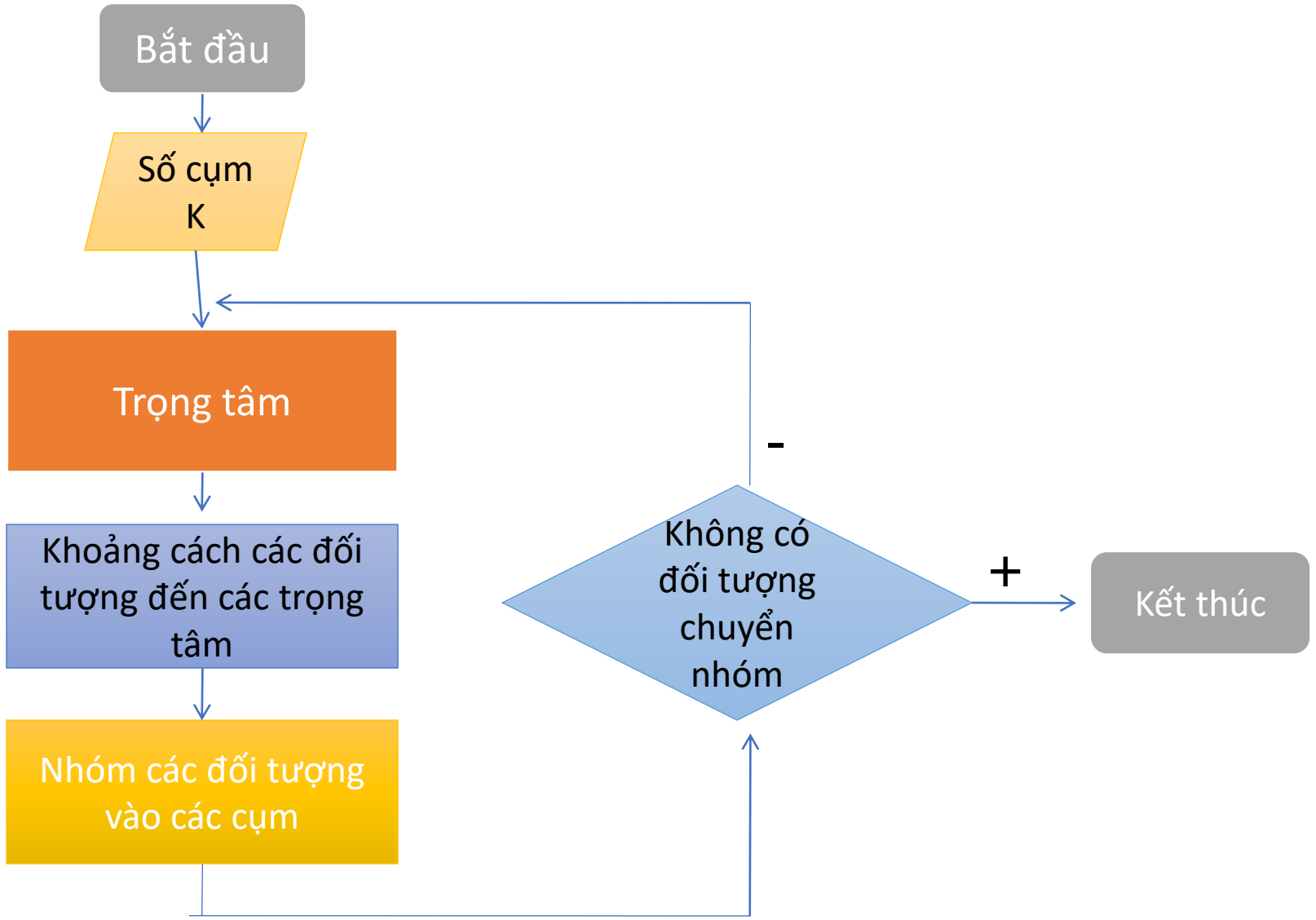
- Bước 3 - Cập nhật lại trọng tâm

$$c_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

- Bước 4 – Điều kiện dừng

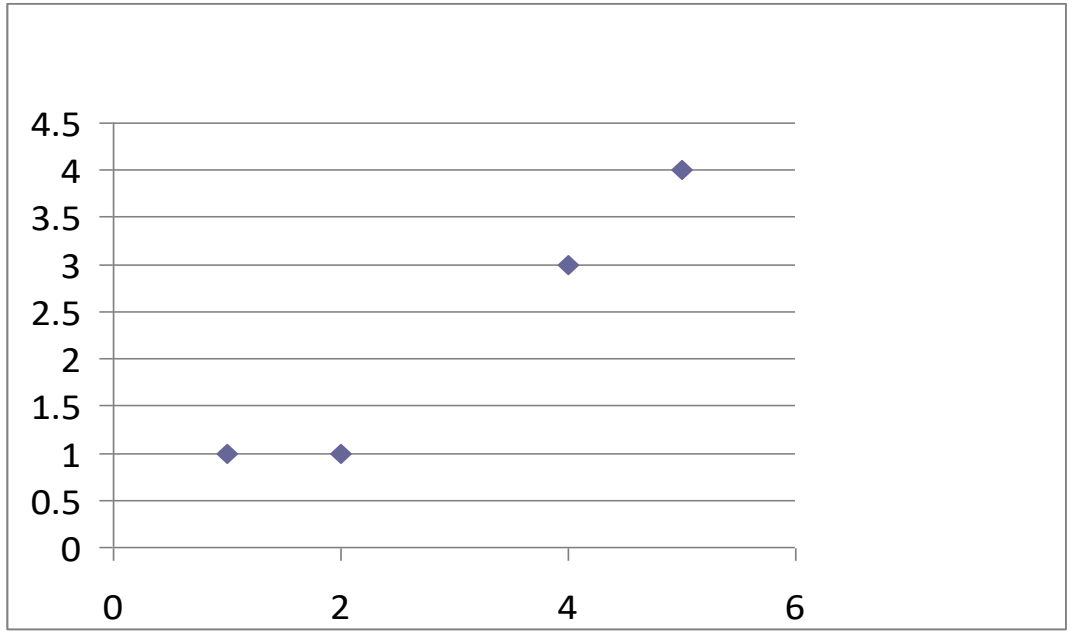
  Lặp lại các bước 2 và 3 cho tới khi không có sự thay đổi trọng tâm của cụm.

II.2. CÁC BƯỚC CủA THUậT TOÁN



II.3 VÍ Dụ MINH HọA

Đối tượng	Thuộc tính 1 (X)	Thuộc tính 2 (Y)
A	1	1
B	2	1
C	4	3
D	5	4

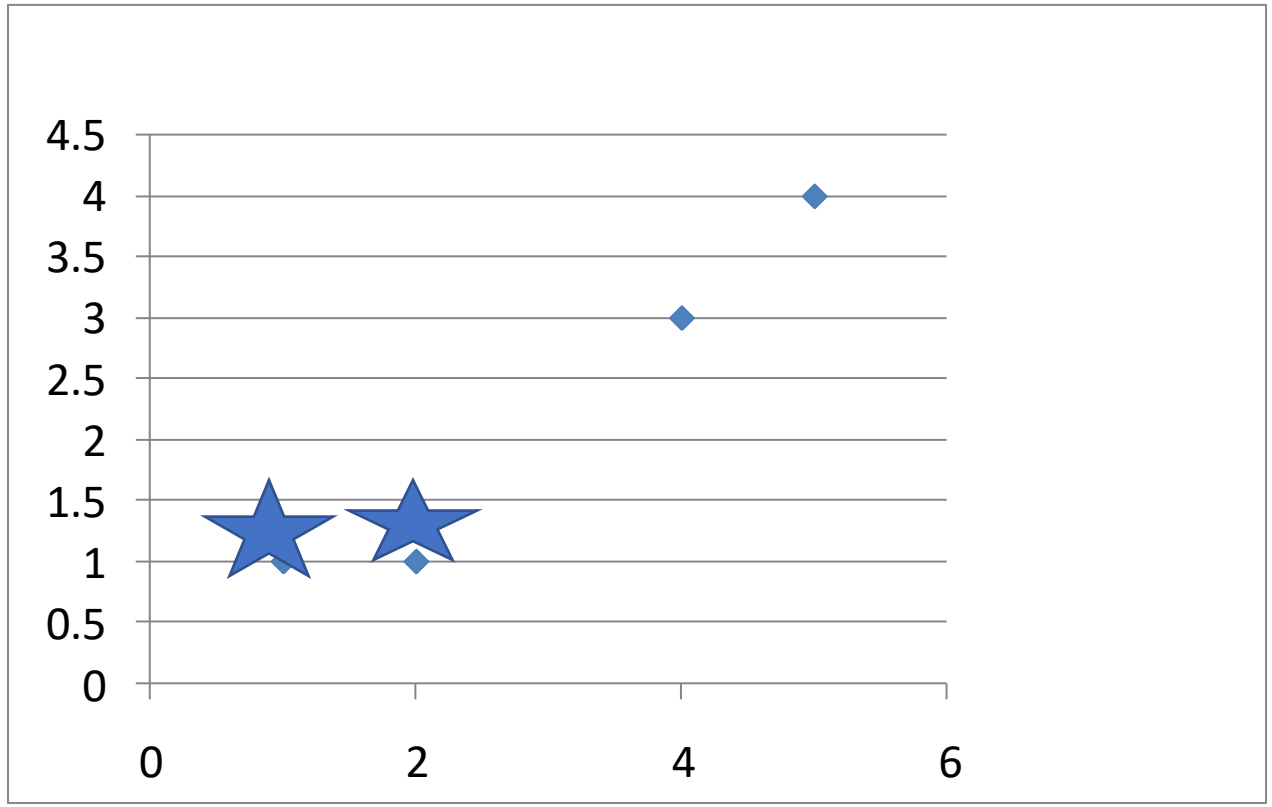


II.3 VÍ Dụ MINH HọA

- **Bước 1:** Khởi tạo

  Chọn 2 trọng tâm ban đầu:

  $c_1(1,1) \equiv A$ và $c_2(2,1) \equiv B$, thuộc 2 cụm 1 và 2



II.3 Ví Dụ MINH HọA

- **Bước 2:** Tính toán khoảng cách

➢ $d(C, c_1) = \sqrt{(4-1)^2 + (3-1)^2}$

$= 13$

$d(C, c_2) = \sqrt{(4-2)^2 + (3-1)^2}$

$= 8$

$d(C, c_1) > d(C, c_2)$ ⇒ C thuộc cụm 2

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

$= 25$

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

$= 18$

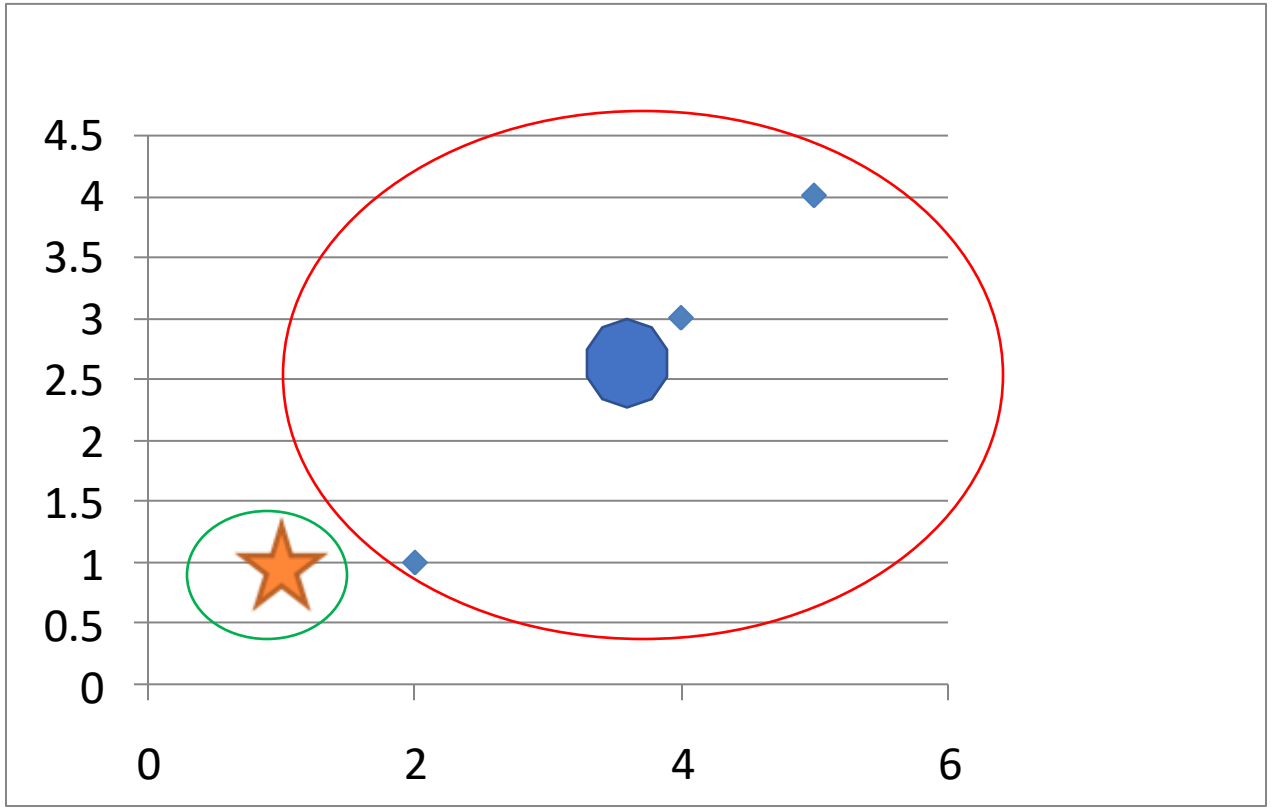
$d(D,c_1) > d(D, c_2)$ ⇒ D thuộc cụm 2

II.3 Ví Dụ MINH HọA

- **Bước 3:** Cập nhật lại vị trí trọng tâm

➢ Trọng tâm cụm 1 c_1 ≡ A (1, 1)

➢ Trọng tâm cụm 2 c_2 (x,y) = $(\frac{2+4+5}{3}, \frac{1+3+4}{3})$



II.3 VÍ Dụ MINH HọA

- **Bước 4-1:** Lặp lại bước 2 – Tính toán khoảng cách

➢ $d(A, c_1) = 0 < d(A, c_2) = 9.89$

A thuộc cụm 1

➢ $d(B, c_1) = 1 < d(B, c_2) = 5.56$

B thuộc cụm 1

➢ $d(C, c_1) = 13 > d(C, c_2) = 0.22$

C thuộc cụm 2

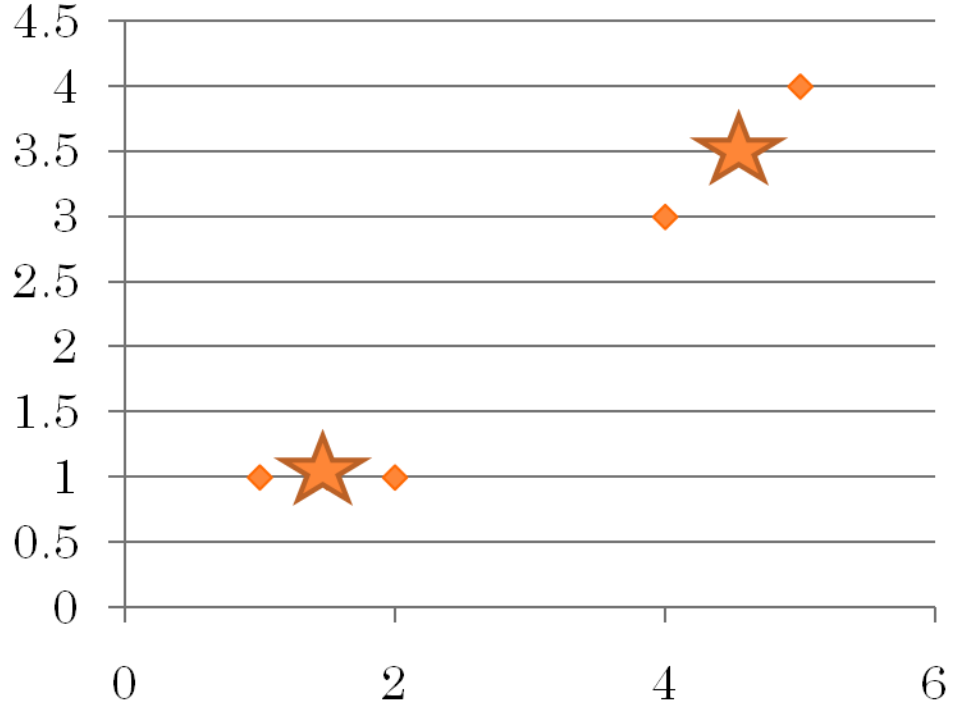
➢ $d(D, c_1) = 25 > d(D, c_2) = 3.56$

D thuộc cụm 2

II.3 VÍ Dụ MINH HọA

- **Bước 4-2:** Lặp lại bước 3-Cập nhật trọng tâm

  $c_1 = (3/2, 1)$ và $c_2 = (9/2, 7/2)$



II.3 VÍ Dụ MINH HọA

- **Bước 4-3:** Lặp lại bước 2

➢ $d(A, c_1) = 0.25 < d(A, c_2) = 18.5$

A thuộc cụm 1

➢ $d(B, c_1) = 0.25 < d(B, c_2) = 12.5$

B thuộc cụm 1

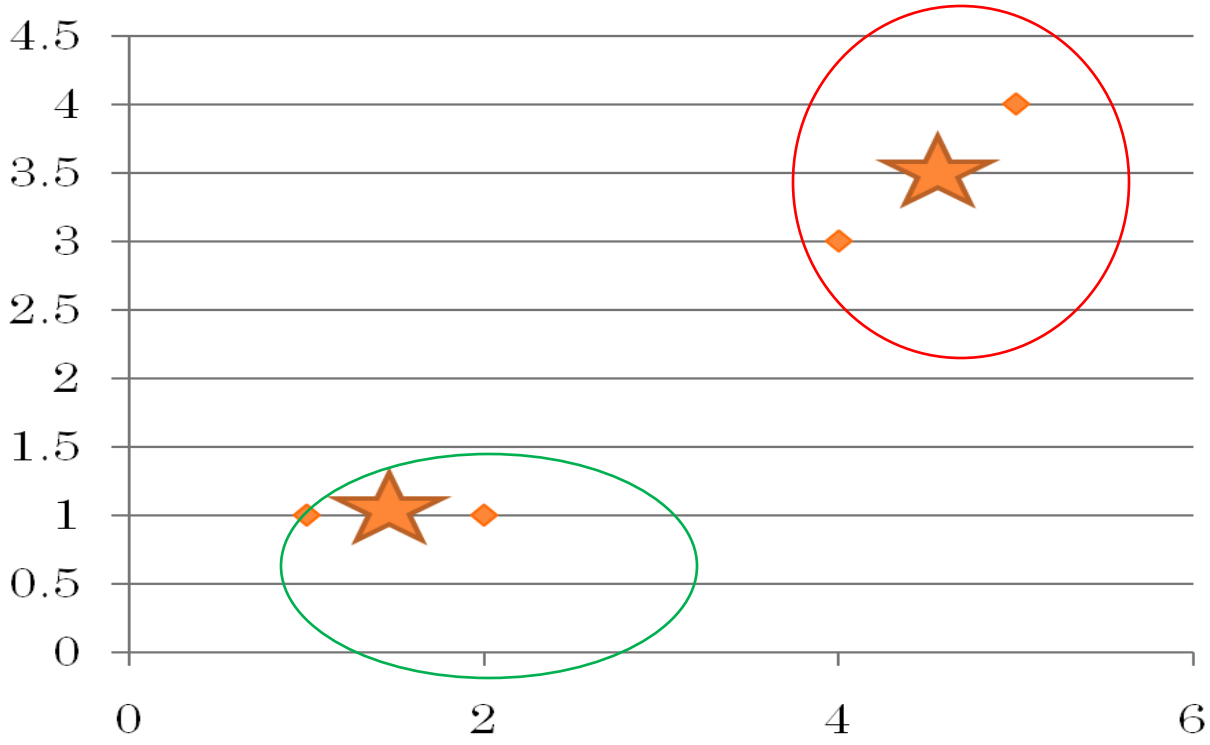
➢ $d(C, c_1) = 10.25 < d(C, c_2) = 0.5$

C thuộc cụm 2

➢ $d(D, c_1) = 21.25 > d(D, c_2) = 0.5$

D thuộc cụm 2

II.3 Ví Dụ MINH HọA



Bài tập

- Có 1 mẩu khoáng sản được dự đoán có vàng, bạc và đồng, ta chia làm 5 điểm tọa độ để xét phân vùng các kim loại: chia làm 3 nhóm

Point	X	Y
A	1	2
B	3	1
C	4	2
D	3	3
E	2	2

II.4 ĐÁNH GIÁ THUậT TOÁN – ƯU ĐIểM

1. Độ phức tạp: O($K N l$) Với l: số lần lặp
2. Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.
3. Bảo đảm hội tụ sau 1 số bước lặp hữu hạn.
4. Luôn có K cụm dữ liệu
5. Luôn có ít nhất 1 điểm dữ liệu trong 1 cụm dữ liệu.
6. Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau.
7. Mọi thành viên của 1 cụm là gần với chính cụm đó hơn bất cứ 1 cụm nào khác.

II.4 ĐÁNH GIÁ THUẬT TOÁN – NHƯỢC ĐIỂM

1. Không có khả năng tìm ra các cụm không lồi hoặc các cụm có hình dạng phức tạp.
2. Khó khăn trong việc xác định các trọng tâm cụm ban đầu
   - Chọn ngẫu nhiên các trung tâm cụm lúc khởi tạo
   - Độ hội tụ của thuật toán phụ thuộc vào việc khởi tạo các vector trung tâm cụm
3. Khó để chọn ra được số lượng cụm tối ưu ngay từ đầu, mà phải qua nhiều lần thử để tìm ra được số lượng cụm tối ưu.
4. Rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu.
5. Không phải lúc nào mỗi đối tượng cũng chỉ thuộc về 1 cụm, chỉ phù hợp với đường biên giữa các cụm rõ.

II.5 TỔNG QUÁT HÓA VÀ CÁC BIẾN THỂ

B. Các biến thể

1. Thuật toán K-medoid:
   - Tương tự thuật toán K-mean
   - Mỗi cụm được đại diện bởi một trong các đối tượng của cụm.
   - Chọn đối tượng ở gần tâm cụm nhất làm đại diện cho cụm đó.
   - K-medoid khắc phục được nhiễu, nhưng độ phức tạp lớn hơn.

II.5 TỔNG QUÁT HÓA VÀ CÁC BIẾN THỂ

2. Thuật toán Fuzzy c-mean (FCM):

- Chung chiến lược phân cụm với K-mean.
- Nếu K-mean là phân cụm dữ liệu cứng (1 điểm dữ liệu chỉ thuộc về 1 cụm) thì FCM là phân cụm dữ liệu mờ (1 điểm dữ liệu có thể thuộc về nhiều hơn 1 cụm với 1 xác suất nhất định).
- Thêm yếu tố quan hệ giữa các phần tử và các cụm dữ liệu thông qua các trọng số trong ma trận biểu biễn bậc của các thành viên với 1 cụm.
- FCM khắc phục được các cụm dữ liệu chồng nhau trên các tập dữ liệu có kích thước lớn hơn, nhiều chiều và nhiều nhiễu, song vẫn nhạy cảm với nhiễu và các phần tử ngoại lai.

III. ứNG DụNG CủA THUậT TOÁN

- Phân cụm tài liệu web.
  1. Tìm kiếm và trích rút tài liệu
  2. Tiền xử lý tài liệu: Quá trình tách từ và vecto hóa tài liệu: tìm kiếm và thay thế các từ bởi chỉ số của từ đó trong từ điển.Biểu diễn dữ liệu dưới dạng vectơ.
  3. Áp dụng K-Mean

Kết quả trả về là các cụm tài liệu và các trọng tâm tương ứng.

- Phân vùng ảnh

TÀI LIệU THAM KHảO

- ***Tài liệu chính***: [WKQ08] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu , Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg (2008). Top 10 algorithms in data mining, *Knowl Inf Syst (2008)* 14:1–37
- Pavel Berkhin (). Survey of Clustering Data Mining Techniques
- http://en.wikipedia.org/wiki/K-means_clustering
- http://en.wikipedia.org/wiki/Segmentation_(image_processing)
- Slide KI2 – 7 Clustering Algorithms - Johan Everts
- http://vi.wikipedia.org/wiki/Học_không_có_giám_sát
- http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm