**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**
KHOA HỆ THỐNG THÔNG TIN

**TÀI LIỆU HƯỚNG DẪN THỰC HÀNH**
Môn: Khai thác dữ liệu

# Bài thực hành 5
# GOM CỤM DỮ LIỆU

## I. Mục tiêu

1. Hiểu và vận dụng giải thuật *k*-Means để gom cụm dữ liệu.
2. Hiểu và vận dụng giải thuật mạng Kohonen để gom cụm dữ liệu.

## II. Thời gian

1. Thực hành: 5 tiết
2. Bài tập làm thêm: 8 tiết

## III. Hướng dẫn chung

Cho bảng dữ liệu do máy tính thu thập được sau khi phân tích 06 bức tranh như sau:

| Tranh | Số màu | Số đường nét | Số hình khối |
|---|---|---|---|
| **1** | 16 | 124 | 19 |
| **2** | 6 | 13 | 70 |
| **3** | 10 | 22 | 59 |
| **4** | 5 | 81 | 92 |
| **5** | 21 | 97 | 23 |
| **6** | 7 | 94 | 88 |

**Yêu cầu:**

1. Dựa vào dữ liệu ở trên sinh viên sử dụng thuật toán *k*-Means để gom thành 3 nhóm tranh.
2. Hãy áp dụng thuật toán mạng Kohonen để gom các bức tranh thành 3 nhóm với các thông số như sau: số lần lặp $epochs = 5$, bán kính $R = 0$ và tốc độ học (learning rate) $\alpha = 0{,}4$, chu kỳ cập nhật bán kính không đề cập.
3. So sánh kết quả thu được từ thuật toán *k*-Means và mạng Kohonen với nhau.
4. Một bức tranh chưa biết có kết quả phân tích như sau:

| Tranh | Số màu | Số đường nét | Số hình khối |
|---|---|---|---|
| **7** | 13 | 95 | 73 |

Sinh viên hãy giúp máy tính tìm ra những bức tranh có đặc điểm tương đồng với bức tranh trên.

**Hướng dẫn:**

1. Sinh viên tham khảo các bước thực hiện thuật toán k-Means trong tài liệu[1, 2]. Áp dụng trên dữ liệu đề bài ta có:

   Khởi tạo ma trận phân hoạch

| | **Tranh 1** | **Tranh 2** | **Tranh 3** | **Tranh 4** | **Tranh 5** | **Tranh 6** |
|---|---|---|---|---|---|---|
| *Cụm 1* | *1* | *0* | *0* | *0* | *0* | *0* |
| *Cụm 2* | *0* | *1* | *0* | *0* | *0* | *0* |
| *Cụm 3* | *0* | *0* | *1* | *1* | *1* | *1* |

   Vector trọng tâm của các cụm lúc này là:

   - Cụm 1: $V1 = (16, 124, 19)$
   - Cụm 2: $V2 = (6, 13, 70)$
   - Cụm 3: $V3 = (10.75, 73.5, 65.5)$

   Khoảng cách **Euclide** từ các bức tranh đến lần lượt các cụm là:

| | **Cụm 1** | **Cụm 2** | **Cụm 3** |
|---|---|---|---|
| *Tranh 1* | *0* | 122.5643 | 68.84811 |
| *Tranh 2* | 122.5643 | *0* | 60.85279 |
| *Tranh 3* | 109.7269 | *14.76482* | 51.91399 |
| *Tranh 4* | 85.43419 | 71.47727 | *28.13472* |
| *Tranh 5* | *27.74887* | 97.41663 | 49.63429 |
| *Tranh 6* | 75.77599 | 82.98193 | *30.66859* |

   Ma trận phân hoạch các điểm thuộc cụm

| | **Tranh 1** | **Tranh 2** | **Tranh 3** | **Tranh 4** | **Tranh 5** | **Tranh 6** |
|---|---|---|---|---|---|---|
| *Cụm 1* | *1* | *0* | *0* | *0* | *1* | *0* |
| *Cụm 2* | *0* | *1* | *1* | *0* | *0* | *0* |
| *Cụm 3* | *0* | *0* | *0* | *1* | *0* | *1* |

   Vector trọng tâm của các cụm lúc này là:

   - **Cụm 1**: $V1 = (18.5, 110.5, 21)$
   - **Cụm 2**: $V2 = (8, 17.5, 64.5)$
   - **Cụm 3**: $V3 = (6, 87.5, 90)$

Khoảng cách Euclide từ các bức tranh đến lần lượt các cụm là:

| | **Cụm 1** | **Cụm 2** | **Cụm 3** |
|---|---|---|---|
| *Tranh 1* | *13.87444* | 116.0883 | 80.45651 |
| *Tranh 2* | 109.834 | *7.382412* | 77.13786 |
| *Tranh 3* | 96.68764 | *7.382412* | 72.57582 |
| *Tranh 4* | 78.06087 | 69.26399 | *6.873864* |
| *Tranh 5* | *13.87444* | 90.61733 | 69.3127 |
| *Tranh 6* | 69.95356 | 80.03437 | *6.873864* |

| | **Tranh 1** | **Tranh 2** | **Tranh 3** | **Tranh 4** | **Tranh 5** | **Tranh 6** |
|---|---|---|---|---|---|---|
| *Cụm 1* | *1* | *0* | *0* | *0* | *1* | *0* |
| *Cụm 2* | *0* | *1* | *1* | *0* | *0* | *0* |
| *Cụm 3* | *0* | *0* | *0* | *1* | *0* | *1* |

Như vậy các điểm thuộc cụm không thay đổi, thuật toán dừng lại, với kết quả sau:

- Tranh 1 & 5 thuộc cụm 1
- Tranh 2 & 3 thuộc cụm 2
- Tranh 4 & 6 thuộc cụm 3

2. Mạng Kohonen (hay còn gọi là SOM) là một mạng nơron truyền thẳng sử dụng phương pháp học không giám sát, áp dụng trong việc ánh xạ để giảm kích thước dữ liệu đầu vào. Từ tập các đối tượng trong không gian nhiều chiều ở đầu vào, mạng Kohonen có số chiều nhỏ hơn (thường là 2 chiều) được sử dụng để đặc trưng cho chúng ở đầu ra của thuật toán.

   Mạng Kohonen chứa các Nơron. Mỗi nơron chứa một vector trọng số có số chiều bằng số chiều của vector dữ liệu đầu vào. Cụ thể như sau:

   - Đầu vào thuật toán: đối tượng cần gom cụm là tập các vector trong không gian n chiều, số lần lặp của thuật toán *epochs*, bán kính vùng lân cận $R$, chu kỳ cập hật bán kính và tốc độ học $\alpha$.
   - Đầu ra thuật toán: bản đồ mạng Kohonen với mỗi nơron trên mạng đặc trưng cho một cụm.

Thuật toán trải qua các bước như sau:

| ***Bước*** | **Thao tác** |
|---|---|
| 0 | Khởi tạo giá trị của các vector trọng số. Gán giá trị cho biến $R$ và $\alpha$ |
| 1 | Nếu chưa thỏa điều kiện dừng thì lặp lại từ bước 2 đến bước 8 |
| 2 | Với mỗi vector đầu vào $x$ thực hiện bước 3 đến bước 5 |
| *3* | Với mỗi nơron trên mạng $j$, tính khoảng cách Euclide đến $x$ theo công thức: $$D(j) = \sqrt{\sum_{i=1}^{n}(x_i - w_{ij})^2}$$ |
| *4* | Tìm nơron $J$ sao cho khoảng cách từ điểm dữ liệu đang xét $x$ đến $J$ là ngắn nhất. |
| *5* | Duyệt qua tất cả những vector trọng số thuộc nơron láng giềng $j$ của $J$ và xét tất cả các chiều $i$, tiến hành cập nhật $$w_{ij}(new) = w_{ij}(old) + \alpha(x_i - w_{ij}(old))$$ |
| *6* | Cập nhật lại tốc độ học $\alpha$. Điều này giúp làm giảm số lần lặp. |
| *7* | Giảm bán kính vùng lân cận khi đến chu kỳ |
| *8* | Kiểm tra điều kiện dừng. Thông thường thì đó là lúc giá trị của tốc độ học đã rất nhỏ hoặc giá trị của vector trọng số hội tụ nên việc cập nhật không còn thay đổi nhiều. |

Với bài tập trên, trước hết cần xác định các tham số của thuật toán:

- Dữ liệu đầu vào là thông tin của 06 bức tranh, mỗi thông tin có thể xem như là một vector trong không gian $n = 3$ chiều.
- Bán kính vùng lân cận là $R = 0$, điều này có nghĩa là khi thay đổi trọng số của một nơron thì những nơron lân cận không bị ảnh hưởng.
- Số lần lặp của thuật toán là $epochs = 5$
- Tốc độ học là $\alpha = 0.4$
- Không xét chu kỳ cập nhật bán kính vì bán kính $= 0$
- Giả sử mạng Kohonen là mảng 2 chiều có 1 dòng 3 cột. Mỗi phần tử trên mạng là 1 nơron có số chiều bằng số chiều của các mẫu học. Mỗi nơron chứa 1 trọng số là vector $w_i$ có số chiều bằng $n$.

Khởi tạo giá trị của các vector trọng số:

$$w_1 = (19, 111, 21.5)$$
$$w_2 = (6.5, 88, 90.5)$$
$$w_3 = (8.5, 18, 65)$$

Ở lần lặp thứ nhất:

- Xét vector đầu tiên (Tranh 1) $x_1$

  Khoảng cách từ $x_1$ đến $w_1$ là:

  $$D_{11} = \sqrt{(16-19)^2 + (124-111)^2 + (19-21.5)^2} \approx 13.57$$

  Tương tự, khoảng cách từ $x_1$ đến $w_2$, $w_3$ lần lượt là: $D_{12} \approx 80.61$, $D_{13} \approx 115.79$

  Như vậy, nơron có trọng số $w_1$ là nơron có khoảng cách đến $x_1$ ngắn nhất. Tiến hành cập nhật lại trọng số của $w_1$

  $$w_{11}(new) = 19 + 0.4 \times (16 - 19) = 17.8$$
  $$w_{12}(new) = 111 + 0.4 \times (124 - 111) = 116.2$$
  $$w_{13}(new) = 21.5 + 0.4 \times (19 - 21.5) = 20.5$$

  Lúc này, do $R = 0$ nên không cần cập nhật lại vùng lân cận, $w_2, w_3$ sẽ giữ nguyên giá trị, ta có:

  $$w_1 = (17.8, 116.2, 20.5)$$
  $$w_2 = (6.5, 88, 90.5)$$
  $$w_3 = (8.5, 18, 65)$$

- Lần lượt xét các vector còn lại

  Tương tự như trên, sau khi xét lần lượt các vector còn lại, ta có trọng số lúc này là:

  $$w_1 = (18.72, 110.08, 21.2)$$
  $$w_2 = (6.34, 88.72, 89.86)$$
  $$w_3 = (8.5, 18.4, 63.8)$$

- Trước khi kết thúc lần lặp đầu tiên, ta giảm tốc độ học đi một nửa:

  $$\alpha = \frac{0.4}{2} = 0.2$$

Tiếp tục lặp lại các thao tác trên thêm 4 lần, ta được:

$$w_1 = (18.7, 109.78, 21.17)$$
$$w_2 = (6.18, 88.19, 89.9)$$
$$w_3 = (8.32, 18.1, 63.95)$$

Tính khoảng cách từ mỗi điểm đến các vector ta kết luận được:

- Tranh **1 & 5** thuộc cụm đại diện bởi **vector 1 = w1**

- Tranh **4 & 6** thuộc cụm đại diện bởi **vector 2 = w2**
- Tranh **2 & 3** thuộc cụm đại diện bởi **vector 3 = w3**

3. Từ kết quả thu được ở câu 1 và 2, có thể thấy ở cả hai thuật toán các bức tranh trong một cụm đều giống nhau. Các cụm tương ứng ở hai thuật toán là:

| *k*-Means | Mạng Kohonen |
|---|---|
| Cụm 1 | *Cụm 1* |
| Cụm 2 | *Cụm 3* |
| Cụm 3 | *Cụm 2* |

4. Sinh viên có thể tìm ra sự tương đồng bằng cách áp dụng kết quả của thuật toán *k*-Means hoặc mạng Kohonen. Nếu áp dụng kết quả gom cụm từ thuật toán *k*-Means, chúng ta sẽ phải tính khoảng cách Euclide từ điểm đại diện cho Tranh 7 đến tâm các cụm. Sau đó chọn ra cụm có khoảng cách gần nhất, đó chính là cụm chứa các bức tranh có sự tương đồng nhiều nhất với Tranh 7. Tương tự như trên, nếu áp dụng kết quả từ mạng Kohonen, chúng ta sẽ phải tính khoảng cách Euclide từ điểm đại diện cho Tranh 7 đến vector trọng số của các nơron.

   Đối với thuật toán *k*-Means:

   - Khoảng cách đến cụm 1

   $$D_{71} = \sqrt{(13 - 18.7)^2 + (95 - 110.5)^2 + (73 - 21)^2} \approx 54.54$$

   - Tương tự, khoảng cách đến cụm 2 và 3 lần lượt là $D_{72} \approx 78.12$; $D_{73} \approx 19.86$
   - Vì khoảng cách từ Tranh 7 tới tâm cụm 3 là ngắn nhất nên có thể kết luận được rằng Tranh 7 có sự tương đồng với các bức tranh 4 và 6.

   Đối với thuật toán mạng Kohonen:

   - Khoảng cách đến nơron 1

   $$D_{71} = \sqrt{(13 - 18.5)^2 + (95 - 109.78)^2 + (73 - 21.17)^2} \approx 54.2$$

   - Tương tự, khoảng cách đến nơron 2 và 3 lần lượt là $D_{72} \approx 19.46$; $D_{73} \approx 77.57$
   - Vì khoảng cách từ Tranh 7 tới nơron 2 là ngắn nhất nên có thể kết luận được rằng Tranh 7 có sự tương đồng với các bức tranh 4 và 6.

   Như vậy, dù với phương pháp nào thì chúng ta cũng có cùng một kết luận về sự tương đồng của Tranh 7 với tranh 4 và 6.

## IV. Thực hành

1. Trong quá trình thống kê doanh thu, một công ty kinh doanh chuỗi cửa hàng pizza phát hiện những vị trí có nhu cầu cao nhưng chưa có cửa hàng trong khu vực. Tọa độ tương đối của những vị trí tiềm năng đó được cho trong bảng sau:

| Vị trí | *Tọa độ x* | *Tọa độ y* | *Số đơn đặt gần vị trí trong năm* |
|---|---|---|---|
| **1** | 8 | 4 | 200 |
| **2** | 8 | 6 | 350 |
| **3** | 9 | 7 | 650 |
| **4** | 10 | 5 | 400 |
| **5** | 11 | 4 | 320 |
| **6** | 11 | 8 | 250 |
| **7** | 12 | 6 | 600 |
| **8** | 12 | 7 | 300 |
| **9** | 14 | 5 | 200 |

Công ty này muốn xây dựng 3 cửa hàng pizza mới trong những khu vực trên nhằm mở rộng chuỗi cửa hàng và phục vụ tốt hơn cho những khu vực này. Sinh viên hãy giúp công ty bằng cách tìm ra vị trí đặt 3 cửa hàng sao cho thuận lợi nhất và chỉ ra cụm khách hàng của từng cửa hàng với những yêu cầu cụ thể sau:

a) Sử dụng thuật toán $k$-Means.

b) Sử dụng mạng Kohonen với các thông số: $epochs = 10$, $R = 0$ và $\alpha = 0{,}8$ (tốc độ học)

c) So sánh kết quả thu được từ thuật toán $k$-Means và mạng Kohonen với nhau.

d) Một vị trí tiềm năng mới xuất hiện tại tọa độ:

| Vị trí | *Tọa độ x* | *Tọa độ y* | *Số đơn đặt gần vị trí trong năm* |
|---|---|---|---|
| 10 | 11 | 6 | 450 |

Sinh viên hãy giúp công ty xác định cửa hàng nào trong 3 cửa hàng trên sẽ phục vụ cho vị trí này tốt nhất.

2. Một website thương mại điện tử chuyên kinh doanh thiết bị gia dụng thu thập được những đơn hàng như sau:

| ***Đơn hàng*** | **Máy lạnh** | **Máy giặt** | **Tủ lạnh** | **Tivi** | **Bếp điện** |
|---|---|---|---|---|---|
| *1* | 1 | 0 | 1 | 1 | 1 |
| *2* | 1 | 0 | 1 | 0 | 1 |
| *3* | 0 | 0 | 1 | 1 | 1 |
| *4* | 1 | 0 | 0 | 1 | 0 |
| *5* | 0 | 1 | 0 | 1 | 0 |
| *6* | 0 | 1 | 0 | 0 | 1 |
| *7* | 0 | 0 | 0 | 0 | 0 |
| *8* | 1 | 1 | 1 | 0 | 1 |
| *9* | 0 | 1 | 0 | 1 | 0 |

Công ty dự định xây dựng một hệ thống khuyến nghị dựa trên việc gom cụm những đơn hàng trên thành 3 nhóm có độ tương đồng với nhau. Từ đó gợi ý cho việc mua hàng của khách hàng dựa trên sự tương đồng này. Sinh viên hãy tham gia xây dựng hệ thống khuyến nghị bằng cách thực hiện những yêu cầu sau:

a) Sử dụng thuật toán $k$-Means.

b) Sử dụng mạng Kohonen với các thông số: $epochs = 10, R = 0$ và $\alpha = 0{,}8$.

c) So sánh kết quả thu được từ thuật toán $k$-Means và mạng Kohonen với nhau.

d) Một khách đang thực hiện đặt hàng trên hệ thống đã mua Máy lạnh, tủ lạnh. Theo em những mặt hàng nào nên được khuyến nghị dựa theo kết quả đã tính được từ các câu trên.

## V. Bài tập thêm

1. Chọn một ngôn ngữ lập trình, sinh viên hãy cài đặt:
   a) Thuật toán $k$-Means
   b) Thuật toán mạng Kohonen
2. Có nhiều thuật toán gom cụm tự động xác định được số lượng cụm mà không cần người dùng phải cho trước như $k$-Means. Nhiều người cho rằng đây là điểm vượt trội của chúng so với $k$-Means. Em hãy liệt kê ít nhất 2 trường hợp để phản biện lại quan điểm này.
3. Bạn được cho một tập dữ liệu gồm 100 dòng và yêu cầu gom cụm chúng. Bạn sử dụng thuật toán $k$-Means để giải quyết bài toán, tuy nhiên với tất cả các giá trị $k$, $1 \leq k \leq 100$, thuật toán $k$-Means đều cho ra kết quả là một cụm không rỗng duy nhất. Bạn lại tìm cách áp dụng tất cả các thuật toán cải tiến của $k$-Means và đều nhận được kết quả tương tự. Bạn hãy giải thích vì sao?

4. Tìm hiểu thêm các thuật toán gom cụm phổ biến khác:
   a) *k*-Medoids, CLARANS
   b) Chameleon, BIRCH
   c) DBSCAN, OPTICS
   d) STING, CLIQUE
5. Để kiểm tra xem khối u, tổn thương trong ngực bệnh nhân có phải là ung thư hay không, người ta thực hiện phương pháp chọc hút tế bào bằng kim nhỏ (FNA). Tế bào lấy được sau đó được phân tích dưới kính hiển vi. Bảng dữ liệu … trong mục tài liệu tham khảo được tính từ hình ảnh dưới kính hiển vi, các thuộc tính trong bảng mô tả các đặc tính của tế bào được phân tích. Sinh viên hãy dùng thuật toán gom cụm để gom nhóm các khối u lành tính (benign) hoặc ác tính (malignant)

## VI. Tài liệu tham khảo


1. Slide bài giảng lý thuyết môn Khai thác dữ liệu.
2. Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, third edition Morgan Kaufmann Publishers
3. *Breast Cancer Wisconsin (Diagnostic) Data Set: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)*
4. *Link Google Drive tới các tập dữ liệu: https://drive.google.com/drive/folders/1rJEK-nBu7VuaJvd7l-iuDm5sumWGdHYD?usp=sharing*