Chương 5: Gom cụm dữ liệu

Câu 1. Định nghĩa và nêu ví dụ minh họa cho các thuật ngữ sau:

a. Mô hình học không có giám sát (unsupervised learning)
b. Gom cụm dữ liệu (data clustering)
c. Gom cụm dữ liệu dựa trên phân hoạch (partitioning)
d. Gom cụm dữ liệu phân cấp (hierarchical)
e. Gom cụm dữ liệu dựa trên mô hình (model)
f. Gom cụm dữ liệu dựa trên mật độ (density)
g. Gom cụm dữ liệu dựa trên lưới (grid)
h. Gom cụm dữ liệu cứng (hard clustering)
i. Gom cụm dữ liệu mờ (fuzzy clustering)
j. Độ đo tương tự
k. Độ đo sai biệt (khoảng cách)
l. Ma trận sai biệt (dissimilarity matrix)
m. Đánh giá chất lượng gom cụm ngoại
n. Đánh giá chất lượng gom cụm nội

Câu 2. Liệt kê ít nhất 3 ứng dụng minh họa của bài toán gom cụm dữ liệu trong thực tiễn.

Câu 3 – 8. Cho tập dữ liệu huấn luyện trong **Bảng 1**. Thực hiện gom cụm 12 đối tượng OID từ 1 đến 12 bằng các giải thuật như sau. Nhận xét kết quả trả về từ mỗi giải thuật và so sánh với kết quả trả về từ các giải thuật khác.

Bảng 1 - Tập dữ liệu huấn luyện trong không gian 2 chiều

OID	x	y	Cụm thật
1	8	8	1
2	6	5	2
3	5	5	2
4	10	9	1
5	5	4	2
6	1	2	2
7	6	7	1
8	7	7	1
9	7	6	1
10	3	3	2
11	9	9	1
12	4	5	2

Câu 3. Bằng giải thuật ***k-means*** với điều kiện dừng là mean của mỗi cụm không đổi:

a. k = 2, mean của cụm 1 được khởi trị là OID 1 và mean của cụm 2 là OID 2.
b. k = 2, mean của cụm 1 được khởi trị là OID 1 và mean của cụm 2 là OID 11.
c. Với k = 2 và số lần lặp là 4, đánh giá chất lượng gom cụm của giải pháp (a) và (b) dùng các độ đo sau:

i. Entropy
ii. F-measure
iii. Hàm mục tiêu (tham khảo [1], pp. 451)
iv. Dunn index

Kết quả:

a. Số lần lặp = 4; cụm 1 = {1, 4, 7, 8, 9, 11}; cụm 2 = {2, 3, 5, 6, 10, 12}; mean của cụm 1 = (7.833333, 7.666667); mean của cụm 2 = (4, 4)
b. Số lần lặp = 6; cụm 1 = { 2, 3, 5, 6, 10, 12)}; cụm 2 = {1, 4, 7, 8, 9, 11}; mean của cụm 1 = (4, 4); mean của cụm 2 = (7.833333, 7.666667)

Câu 4. Bằng giải thuật ***PAM*** với điều kiện dừng là medoid của mỗi cụm không đổi; k = 2; medoid của cụm 1 được khởi trị là OID 1 và medoid của cụm 2 là OID 2.

Kết quả: số lần lặp = 4; cụm 1 = {1, 4, 7, 8, 9, 11}, cụm 2 = {2, 3, 5, 6, 10, 12}, medoid của cụm 1 = OID 1 = (8, 8); medoid của cụm 2 = OID 5 = (5, 4)

Câu 5. Bằng giải thuật ***AGNES*** với điều kiện là chỉ thực hiện trộn 2 cụm có khoảng cách đơn ngắn nhất (*single linkage*) ở mỗi mức.

Kết quả:

- Mức 1: trộn {2} và {3} với min distance = 1.
- Mức 2: trộn {4} và {11} với min distance = 1.
- Mức 3: trộn {8} và {9} với min distance = 1.
- Mức 4: trộn {8, 9} và {7} với min distance = 1.
- Mức 5: trộn {2, 3} và {12} với min distance = 1.
- Mức 6: trộn {2, 3, 12} và {5} với min distance = 1.
- Mức 7: trộn {1} và {8, 9, 7} với min distance = 1.414214.
- Mức 8: trộn {1, 8, 9, 7} và {4, 11} với min distance = 1.414214.
- Mức 9: trộn {1, 8, 9, 7, 4, 11} và {2, 3, 12, 5} với min distance = 1.414214.
- Mức 10: trộn {6} và {10} với min distance = 2.236068.
- Mức 11: trộn {1, 8, 9, 7, 4, 11, 2, 3, 12, 5} và {6, 10} với min distance = 2.236068.

Câu 6. Bằng giải thuật ***DBSCAN*** với bán kính vùng láng giềng (ε) là 1.414214 và số lượng láng giềng (minPts) ít nhất được yêu cầu là 2.

Kết quả:

- Core objects: 1, 2, 3, 5, 7, 8, 9, 11, 12
- 1 cụm gồm: 1, 2, 3, 4, 5, 7, 8, 9, 11, 12
- Noise, outliers: 6, 10

Câu 7. Bằng giải thuật ***fuzzy c-means*** với c = 2, m = 2, và điều kiện dừng là sự sai biệt giữa các c1 và c2 so với giá trị của chúng trước đó là nhỏ hơn hay bằng 0.001. Ma trận thành viên (membership matrix) giả sử được khởi trị ngẫu nhiên như sau:

OID	1	2	3	4	5	6	7	8	9	10	11	12

c1	0.5	0.4	0.3	0.2	0.1	0.5	0.9	0.8	0.7	0.6	0.5	0.4
c2	0.5	0.6	0.7	0.8	0.9	0.5	0.1	0.2	0.3	0.4	0.5	0.6

Kết quả:

- Số lần lặp = 8
- Delta_c = 0.000524
- Means của các cụm: c1 = (8, 7.749684); c2 = (3.99776, 4.048949)
- Membership matrix:

OID	1	2	3	4	5	6	7	8	9	10	11	12
c1	0.998161	0.29344	0.100324	0.915768	0.040438	0.139628	0.733227	0.917562	0.755744	0.04327	0.950974	0.035995
c2	0.001839	0.70656	0.899676	0.084232	0.959562	0.860372	0.266773	0.082438	0.244256	0.95673	0.049026	0.964005

Câu 8. Bằng giải thuật ***SOM*** với số neurons ở tầng xuất và tầng nhập đều là 2 và kích thước vùng láng giềng là 0 và điều kiện dừng là các weight không thay đổi. Learning rate α được cho biến thiên thời gian (thể hiện qua mỗi lần lặp t) như sau:

- $\alpha(t) = 0.6$ với $1 <= t <= 4$
- $\alpha(t) = 0.5$ với $5 <= t <= 12$
- $\alpha(t) = 0.4$ với $t > 12$

a. Weight ở neuron #1 ở tầng xuất được khởi trị là (0.5, 0.8) và ở neuron #2 là (0.7, 0.4).
b. Weight ở neuron #1 ở tầng xuất được khởi trị là (5, 8) và ở neuron #2 là (7, 4).

Kết quả:

a. t = 9; weight #1 = (7.68254, 7.650794); weight #2 = (3.492063, 4.047619); c1 = {1, 4, 7, 8, 9, 11}; c2 = {2, 3, 5, 6, 10, 12}.
b. t = 9; weight #1 = (7.68254, 7.650794); weight #2 = (3.492063, 4.047619); c1 = {1, 4, 7, 8, 9, 11}; c2 = {2, 3, 5, 6, 10, 12}.

Câu 9. Giải thuật gom cụm nào có thể trả về các cụm có hình dạng tuỳ ý?

a. k-means
b. PAM
c. DBSCAN
d. AGNES
e. Ý kiến khác.

Câu 10. Giải thuật EM (Expectation – Maximization) có đặc điểm nào sau đây?

a. Gom cụm dựa trên khoảng cách giữa đối tượng và phần tử đại diện của cụm
b. Gom cụm dựa trên mật độ phân bố của các đối tượng đối với mỗi cụm
c. Gom cụm dựa trên kết quả phân hoạch của giải thuật k-means
d. Gom cụm dựa trên mô hình bằng cách xem xét trọng số của đối tượng với mỗi cụm và sau đó là ước lượng trị thông số
e. Ý kiến khác.

Câu 11. Giải thuật gom cụm nào bị ảnh hưởng bởi các phần tử kỳ dị (outliers)?

a. k-means
b. PAM
c. DBSCAN

d. AGNES
e. Ý kiến khác.

Câu 12. Giải thuật PAM có yếu điểm gì so với giải thuật k-means?
a. Khó khăn trong việc xác định giá trị k
b. Chi phí giải thuật cao
c. Ảnh hưởng bởi nhiễu
d. Cần phải xác định số lần lặp
e. Ý kiến khác.

Câu 13. Giải thuật nào sau đây có thể tạo các cụm mà trong đó, một đối tượng có thể thuộc về nhiều hơn 1 cụm với một mức độ nào đó ?
a. k-means
b. fuzzy c-means
c. DBSCAN
d. EM
e. Ý kiến khác.

Câu 14. So sánh 2 bài toán khai phá dữ liệu: phân loại (classification) và gom cụm (clustering).
a. Hai bài toán này về bản chất là giống nhau và chỉ khác ở cách tiếp cận xử lý dữ liệu huấn luyện.
b. Kết quả trả về của mỗi bài toán đều là mô hình khai phá nhưng gom cụm thực hiện nhanh hơn do không có giai đoạn sử dụng mô hình; trong khi đó, phân loại cần thực hiện thêm giai đoạn sử dụng mô hình để phân loại đối tượng.
c. Quá trình học để xây dựng mô hình của bài toán phân loại có giám sát; trong khi đó, quá trình học để xây dựng mô hình của bài toán gom cụm là không giám sát.
d. Bài toán phân loại phức tạp hơn do cần phải xem xét thông tin lớp của mỗi đối tượng trong quá trình học; trong khi đó, bài toán gom cụm chỉ cần xem xét bản thân các đối tượng trong quá trình học.
e. Ý kiến khác.