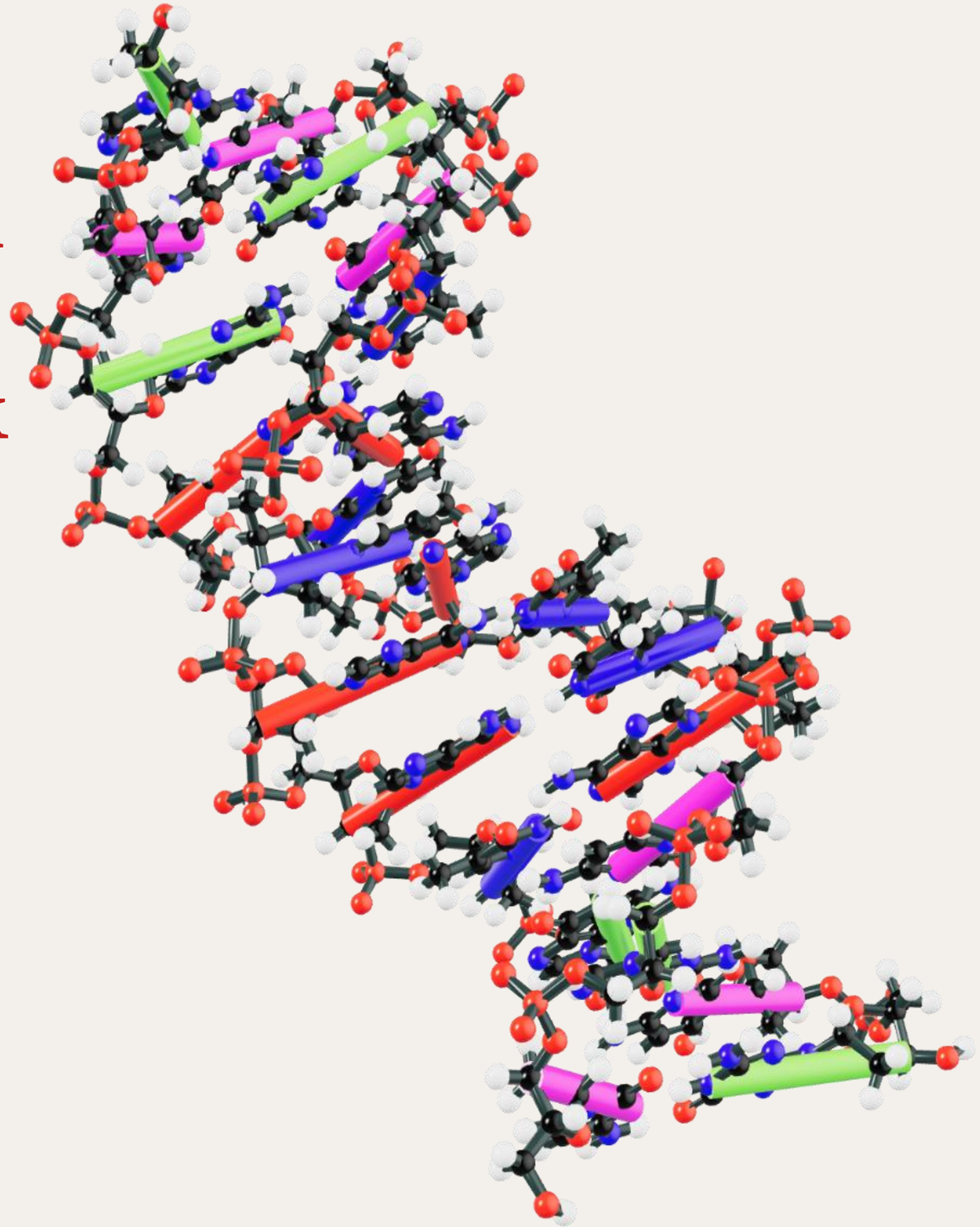
Nghiên cứu và triển khai thuật toán phân cụm dữ liệu được trình bày trong bài báo "Flexible Density Peak Clustering for Real-World Data"



GVHD: TH.S Nguyễn Thủy Đoan Trang

SV: Trương Trần Tâm - 65133127

Nội dung thực hiện

1.Giới thiệu về phân cụm dữ liệu

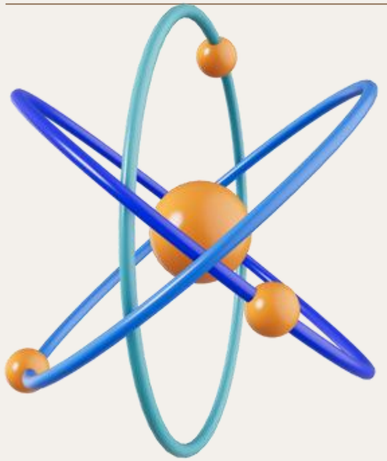
2.Thuật toán Clustering by Fast Search and Find of Density Peaks (DPC)

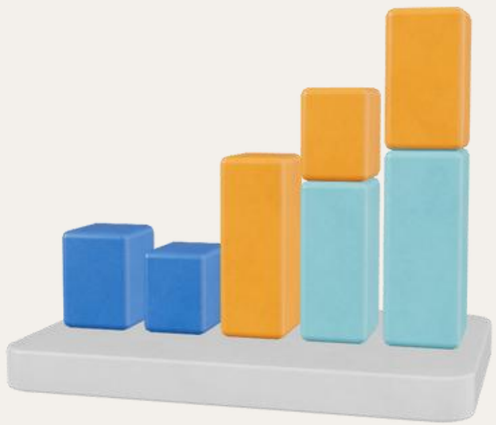
3.Thuật toán Sequential Density Peak Clustering (SDPC)

GIỚI THIỆU VỀ PHÂN CỤM DỮ LIỆU

Phân cụm dữ liệu là một phương pháp phân tích dữ liệu quan trọng, giúp phân loại các đối tượng thành các nhóm (cụm) sao cho các đối tượng trong cùng một cụm có mức độ tương đồng cao, đồng thời khác biệt so với các đối tượng thuộc cụm khác. Phân cụm thường được sử dụng trong các lĩnh vực như khai phá dữ liệu, học máy, nhận dạng mẫu và phân tích dữ liệu lớn. Các thuật toán phân cụm phổ biến bao gồm K-means, DBSCAN và các thuật toán dựa trên mật độ.

THUẬT TOÁN CLUSTERING BY FAST SEARCH AND FIND OF DENSITY PEAKS (DPC)

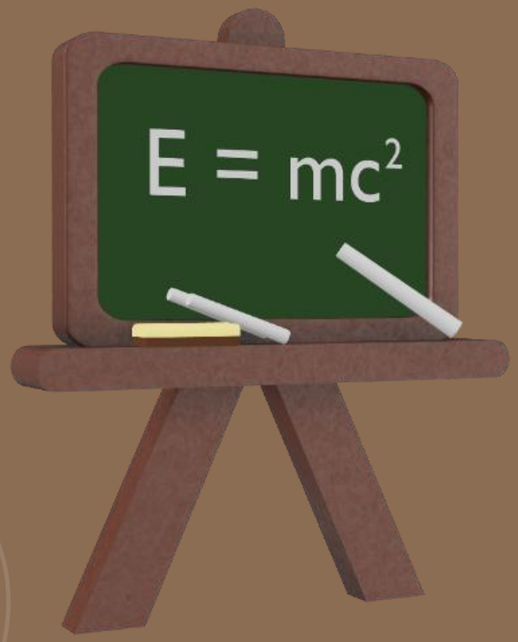




Giới thiệu:

Thuật toán Density Peak Clustering (DPC) được Rodriguez và Laio đề xuất dựa trên giả thuyết rằng tâm cụm là các điểm có mật độ cục bộ cao và nằm xa các điểm có mật độ lớn hơn. DPC sử dụng đồ thị quyết định (decision graph) để người dùng lựa chọn tâm cụm một cách trực quan

2.2 CÁC BƯỚC DPC



Bước 1: Tính mật độ cục bộ (ρ)

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c) .$$

Trong đó:

ρ_i: là mật độ cục bộ của điểm dữ liệu i

d_c: là tham số khoảng cách ngưỡng (cut-off distance)

d_ij : khoảng cách giữa hai điểm i và j

2.2 CÁC BƯỚC DPC

Bước 2: Tính δ (khoảng cách đến điểm có mật độ cao hơn)

$$\delta_i = \begin{cases} \max\limits_{j}(d_{ij}) & \rho_i \text{ is the max} \\ \min\limits_{j:\rho_j > \rho_i}(d_{ij}) & \text{else} \end{cases}.$$

2.2 CÁC BƯỚC DPC

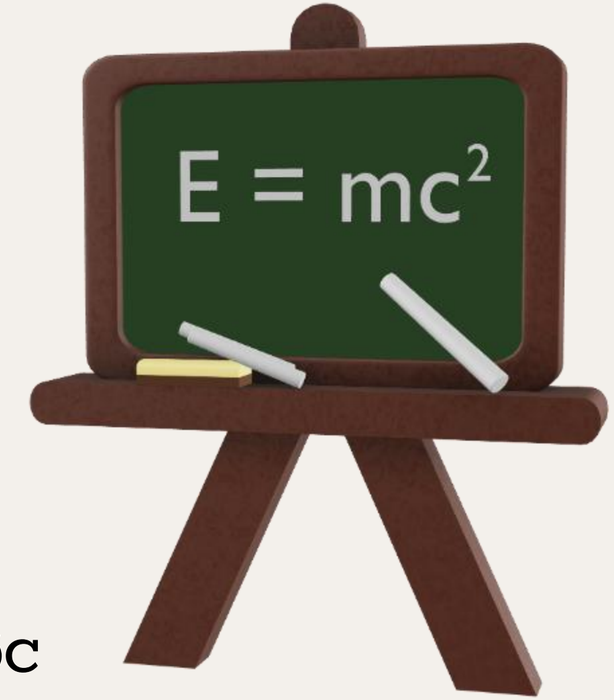
Bước 3: Xác định tâm cụm

Tâm cụm được xác định dựa trên hai giá trị: mật độ cục bộ ρ_i và khoảng cách δi.

+ Trục X: ρ_i
+ Trục Y: δ_i

Các điểm dữ liệu có mật độ cao và khoảng cách lớn đồng thời sẽ được chọn làm tâm cụm dựa trên biểu đồ quyết định (decision graph)

2.2 CÁC BƯỚC DPC



Bước 4: Gán điểm dữ liệu vào các cụm

-Trừ các điểm tâm cụm, những điểm dữ liệu còn lại được gán vào cụm của điểm gần nhất có mật độ cao hơn.

-Việc gán được thực hiện theo thứ tự giảm dần mật độ.

2.3 HẠN CHẾ

1. Phụ thuộc vào lựa chọn thủ công tâm cụm trên đồ thị (ρ, δ).

2. Khó tự động xác định số lượng cụm.

3. Với dữ liệu mật độ không đồng đều, việc chọn ngưỡng hoặc tâm cụm trở nên khó khăn và thiếu ổn định.

THUẬT TOÁN SEQUENTIAL DENSITY PEAK CLUSTERING (SDPC)

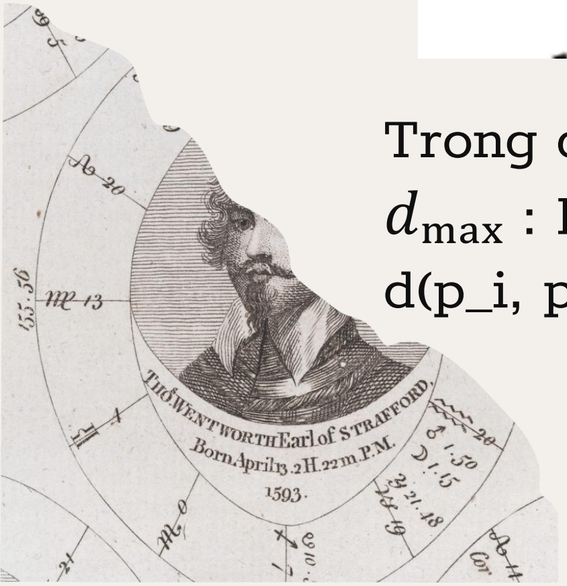
Giới thiệu:

- Sequential Density Peak Clustering (SDPC) là một thuật toán cải tiến từ DPC (Density Peak Clustering), được thiết kế để giải quyết các hạn chế của DPC.
- Thay vì xác định tất cả tâm cụm cùng lúc như DPC, SDPC trích xuất từng cụm một cách tuần tự (sequential) từ vùng có mật độ cao nhất, giúp tự động xác định số cụm.

3.2 CÁC BƯỚC SDPC

Bước 1: Tính mật độ linh hoạt (ρ)

$$\rho_i = \frac{d_{max}}{\frac{1}{k}\sum_{p_j \in S_{inn}} d(p_i, p_j)},$$



Trong đó:

$d_{\max}$: là là giá trị lớn nhất của tất cả các khoảng cách theo cặp của dữ liệu S

d(p_i, p_j): là khoảng cách giữa p_ivà p_j

3.2 CÁC BƯỚC SDPC

Bước 2:Xác định tâm cụm

Nguyên tắc: Chọn điểm chưa gán cụm có mật độ ρ lớn nhất làm tâm cụm.

Công thức:

$$p_c = \arg\max_{p_i \in S_{uc}} \rho_i$$

Suc là tập các điểm chưa gán nhãn.

3.2 CÁC BƯỚC SDPC

Bước 3:Tạo cụm ban đầu

-Dựa trên DBSCAN với các tham số:

+ MinPts=4 (cố định)

+ Eps được tính thích ứng theo từng cụm:

$$Eps = d(p_c, p_{\kappa nn}) \quad \text{với } \kappa > MinPts$$

Với pκnn là điểm thứ κ gần nhất của tâm cụm p_(c)

- Quy trình:

1. Nhóm các điểm trong bán kính Eps xung quanh tâm cụm thành cụm ban đầu

2.Kiểm tra từng điểm có phải core point (ít nhất MinPts điểm trong bán kính Eps) để mở rộng cụm

3.Lặp lại cho đến khi tất cả các điểm trong cụm khởi tạo được duyệt

3.2 CÁC BƯỚC SDPC

Bước 4: Xác định tập ứng viên (Candidate Set)

Khái niệm: Direct Neighbors — các điểm chưa phân cụm gần nhất với các điểm trong cụm ban đầu.

Công thức:

$$p_{dn}^{i} = \arg\min_{p_j \in S_{uc}} d(p_i, p_j), \quad \forall p_i \in S_{ic}$$

Tập dữ liệu ứng viên:

$$S_{candi} = \{p_{dn}^{i} \text{ cho tất cả } p_i \in S_{ic}\}$$

3.2 CÁC BƯỚC SDPC

Bước 5: Mở rộng cụm (Cluster Expansion)

Nguyên tắc: Mỗi điểm ứng viên p_i∈S_candisẽ được thêm vào cụm nếu superordinate của nó nằm trong ban đầu

$$p_i \in S_{ic} \quad \text{nếu } p_i^* \in S_{ic}$$

Superordinate (p)* : điểm cấp trên gần nhất có mật độ lớn hơn của p_i

Lặp lại việc tìm direct neighbors mới và gán điểm cho đến khi không còn ứng viên nào thỏa điều kiện

3.2 CÁC BƯỚC SDPC

Bước 6: Hoàn thiện cụm

Khi không còn điểm ứng viên nào có superordinate trong cụm, dừng mở rộng

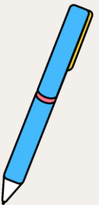
Cụm khởi tạo hiện tại trở thành cụm cuối cùng cho lần trích xuất này:

$$S_c = S_{ic}$$

3.2 CÁC BƯỚC SDPC

Bước 7: Trích xuất cụm tiếp theo

1. Loại bỏ các điểm vừa gán vào cụm khỏi S_uc
2. Lặp lại Bước 2 → Bước 6 cho đến khi tất cả các điểm được gán cụm
3. Số lượng cụm được xác định tự động.



3.3 VÍ DỤ

ID	X	Y
A	956203.5185	1357124.588
B	956203.5185	1357051.127
C	956203.5185	1357024.588
D	956237.3036	1357074.582
E	956254.2387	1357124.586
F	956263.7635	1357174.584
G	956303.531	1357124.584
H	956359.0144	1357224.588
I	956503.5339	1357147.438
J	956503.5385	1357124.576
K	956503.5532	1357051.129
L	956250.0059	1357224.588
M	956503.5185	1357224.588
N	956203.5185	1357174.588
O	956203.5185	1357224.588
P	956403.5348	1357124.58
Q	956303.5185	1357074.57
R	956403.5185	1357224.588

3.3 VÍ DỤ

Bước 1: Tính mật độ cục bộ ρ_i

$$\rho_i = \frac{d_{max}}{\frac{1}{k}\sum_{p_j \in S_{inn}} d(p_i, p_j)},$$

mật độ của điểm pi = max(matrankhoangcach) chia cho trung bình kc từ điểm pi đến k láng giềng gần nhất
với k là 1 tham số đầu vào

3.3 VÍ DỤ

Bảng khoảng cách Euclid giữa các điểm

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A	0	73.461	100	60.3493	50.7202	78.2883	100.0125	184.8756	300.8843	300.02	308.8965	110.2773	316.2278	50	100	200.0163	111.8114	223.6068
B	73.461	0	26.539	41.1287	89.2679	137.3721	124.0904	232.9543	315.0953	308.8798	300.0347	179.5823	346.5382	123.461	173.461	213.0771	102.7111	264.7427
C	100	26.539	0	60.3393	112.1255	161.6424	141.4274	253.3357	324.1934	316.2429	301.2063	205.3316	360.5551	150	200	223.6178	111.7954	282.8427
D	60.3493	41.1287	60.3393	0	52.7939	103.4433	82.9835	193.1717	276.0191	270.8882	267.2806	150.5428	305.5686	105.5587	153.7636	173.5875	66.2149	223.8955
E	50.7202	89.2679	112.1255	52.7939	0	50.8972	49.2923	144.839	250.3404	249.2998	259.9109	100.0915	268.5904	71.2232	112.1291	149.2961	70.2147	179.6799
F	78.2883	137.3721	161.6424	103.4433	50.8972	0	63.8863	107.5785	241.3022	244.9344	269.704	51.862	244.914	60.245	78.2934	148.4467	107.6256	148.4313
G	100.0125	124.0904	141.4274	82.9835	49.2923	63.8863	0	114.3644	201.3044	200.0075	213.0834	113.4272	223.5974	111.8164	141.433	100.0038	50.014	141.4153
H	184.8756	232.9543	253.3357	193.1717	144.839	107.5785	114.3644	0	163.8231	175.7544	225.7864	109.0085	144.5041	163.337	155.4959	109.4699	159.9537	44.5041
I	300.8843	315.0953	324.1934	276.0191	250.3404	241.3022	201.3044	163.8231	0	22.862	96.309	265.0067	77.15	301.2414	309.7763	102.5783	212.8753	126.3139
J	300.02	308.8798	316.2429	270.8882	249.2998	244.9344	200.0075	175.7544	22.862	0	73.447	272.5457	100.012	304.1598	316.2505	100.0037	206.1761	141.444
K	308.8965	300.0347	301.2063	267.2806	259.9109	269.704	213.0834	225.7864	96.309	73.447	0	307.2039	173.459	324.4425	346.5672	124.0916	201.4035	200.2373
L	110.2773	179.5823	205.3316	150.5428	100.0915	51.862	113.4272	109.0085	265.0067	272.5457	307.2039	0	253.5126	68.2721	46.4874	183.2286	159.2765	153.5126
M	316.2278	346.5382	360.5551	305.5686	268.5904	244.914	223.5974	144.5041	77.15	100.012	173.459	253.5126	0	304.1381	300	141.4155	250.0108	100
N	50	123.461	150	105.5587	71.2232	60.245	111.8164	163.337	301.2414	304.1598	324.4425	68.2721	304.1381	0	50	206.173	141.4341	206.1553
O	100	173.461	200	153.7636	112.1291	78.2934	141.433	155.4959	309.7763	316.2505	346.5672	46.4874	300	50	0	223.625	180.2925	200
P	200.0163	213.0771	223.6178	173.5875	149.2961	148.4467	100.0038	109.4699	102.5783	100.0037	124.0916	183.2286	141.4155	206.173	223.625	0	111.8225	100.008
Q	111.8114	102.7111	111.7954	66.2149	70.2147	107.6256	50.014	159.9537	212.8753	206.1761	201.4035	159.2765	250.0108	141.4341	180.2925	111.8225	0	180.2925
R	223.6068	264.7427	282.8427	223.8955	179.6799	148.4313	141.4153	44.5041	126.3139	141.444	200.2373	153.5126	100	206.1553	200	100.008	180.2925	0

3.3 VÍ DỤ

Cho k = 6

Điểm	6 láng giềng gần nhất	Trung bình (k=6)	pi = max/trung bình
A	N, E, D, B, F, C	68.8031	5.240388
B	C, D, A, E, Q, N	76.0948	4.738237
C	B, D, A, Q, E, G	92.0378	3.917469
D	B, E, C, A, Q, G	60.6349	5.946326
E	G, A, F, D, Q, N	57.5236	6.267953
F	E, L, N, G, A, O	63.912	5.641427
G	E, Q, F, D, P, A	74.3654	4.848425
H	R, F, L, P, G, M	104.9049	3.43697
I	J, M, K, P, R, H	98.1727	3.672661
J	I, K, P, M, R, H	102.2538	3.526078
K	J, I, P, M, R, Q	144.8246	2.489599
L	O, F, N, E, H, A	80.9998	4.451309
M	I, R, J, P, H, K	122.7568	2.937151
N	A, O, F, L, E, D	67.5498	5.337616
O	L, N, F, A, E, G	88.0571	4.094558
P	J, G, R, I, H, Q	103.981	3.467508
Q	G, D, E, B, F, C	84.7626	4.253704
R	H, M, P, I, G, J	108.9475	3.309437

3.3 VÍ DỤ

Bước 2:Xác định tâm cụm

Nguyên tắc: Chọn điểm chưa gán cụm có mật độ ρ lớn nhất làm tâm cụm.

Công thức:

$$p_c = \arg\max_{p_i \in S_{uc}} \rho_i$$

Suc là tập các điểm chưa phân cụm.

E có ρ_i lớn nhất → chọn làm tâm cụm đầu tiên

3.3 VÍ DỤ

Bước 3:Xây dựng ban đầu (Initial Cluster)

- MinPts=4 (cố định)
- Eps được tính thích ứng theo từng cụm:

$$Eps = d(p_c, p_{\kappa nn}) \quad \text{với } \kappa > MinPts$$

Lấy κ = 5 láng giềng gần E từ bảng khoảng cách:

⟶ Eps = 70.2147 (Q)

3.3 VÍ DỤ

Bước 3:Xây cụm ban đầu (Initial Cluster)

Chọn tất cả điểm nằm trong bán kính Eps = 70.2147 quanh E

Kiểm tra core points (≥ MinPts = 4 điểm trong bán kính Eps)

Điểm đang xét	Láng giềng trong eps	Core?	Cụm sau bước xét
E	A, D, G, F, Q	Yes	{E, A, D, G, F, Q}
A	N, E, D	No	Không đổi
D	B, E, C, A, Q	Yes	{E, A, D, G, F, Q, B, C}
G	E, Q, F	No	Không đổi
F	E, L, N, G	Yes	{E, A, D, G, F, Q, B, C, L, N}
Q	G, D, E	No	Không đổi
B	C, D	No	Không đổi
C	B, D	No	Không đổi
L	O, F, N	No	Không đổi
N	A, O, F, L	Yes	{E, A, D, G, F, Q, B, C, L, N, O}
O	L, N	No	Không đổi

→ Ta thu được cụm ban đầu : Sic = {E, A, D, G, F, Q, B, C, L, N, O}

3.3 VÍ DỤ

Bước 4: Xác định tập ứng viên (Candidate Set)

-Xung quanh Sic = {E, A, D, G, F, Q, B, C, L, N, O}

-Các điểm chưa được phân cụm:H, I, J, K, M, P, R

-Chọn các direct neighbors gần nhất:

Điểm Sic	Direct Neighbor
E	H
A	H
D	P
G	P
F	H
Q	P
B	P
C	P
L	H
N	H
O	H

Scandi = {H,P}

3.3 VÍ DỤ

Bước 5: Mở rộng cụm (Cluster Expansion)

- Xét từng điểm ứng viên nếu superordinate của nó nằm trong ban đầu sẽ được thêm vào cụm

Điểm	Superordinate	Thêm vào cụm
H	F	YES
P	J	NO

3.3 VÍ DỤ

Bước 5: Mở rộng cụm (Cluster Expansion)

- Sau bước mở rộng, Sic = {E, A, D, G, F, Q, B, C, L, N, O, H}
- Các điểm được phân cụm: I, J, K, M, P, R
- Chọn các direct neighbors gần nhất: Scandi = {P, R}

Điểm	Superordinate	Thêm vào cụm
P	J	NO
R	H	YES

3.3 VÍ DỤ

Bước 5: Mở rộng cụm (Cluster Expansion)

- Sau bước mở rộng, Sic = {E, A, D, G, F, Q, B, C, L, N, O, H,R}
- Các điểm được phân cụm:I, J, K, M, P
- Chọn các direct neighbors gần nhất: Scandi = {P,M}

Điểm	Superordinate	Thêm vào cụm
P	J	NO
M	I	NO

=>Dừng mở rộng

3.3 VÍ DỤ

Bước 6: Hoàn thiện cụm

•Cluster 1 = {E, A, D, G, F, Q, B, C, L, N, O, H, R}

Bước 7: Trích xuất cụm tiếp theo

•Loại bỏ các điểm vừa gán vào cụm khỏi S_uc

• S_uc={I, J, K, M, P}

•Lặp lại Bước 2 ⟶ Bước 6 cho đến khi tất cả các điểm được gán cụm

3.3 VÍ DỤ

Bước 7: Trích xuất cụm tiếp theo

•Loại bỏ các điểm vừa gán vào cụm khỏi S_uc

• S_uc={ I, J, K, M, P}

•Lặp lại Bước 2 ⟶ Bước 6 cho đến khi tất cả các điểm được gán cụm

3.3 VÍ DỤ

Bước 2:Xác định tâm cụm

- S_uc={ I, J, K, M, P}
- Công thức:

$$p_c = \arg\max_{p_i \in S_{uc}} \rho_i$$

I có ρ_i lớn nhất trong S_uc → chọn làm tâm cụm tiếp theo

3.3 VÍ DỤ

Bước 3:Xây dựng cụm khởi tạo (Initial Cluster)

-MinPts=4 (cố định)

-Eps được tính thích ứng theo từng cụm:

$$Eps = d(p_c, p_{\kappa nn}) \quad \text{với } \kappa > MinPts$$

Lấy κ = 5 láng giềng gần I từ bảng khoảng cách:

⟶ Eps = 126.3139 (R)

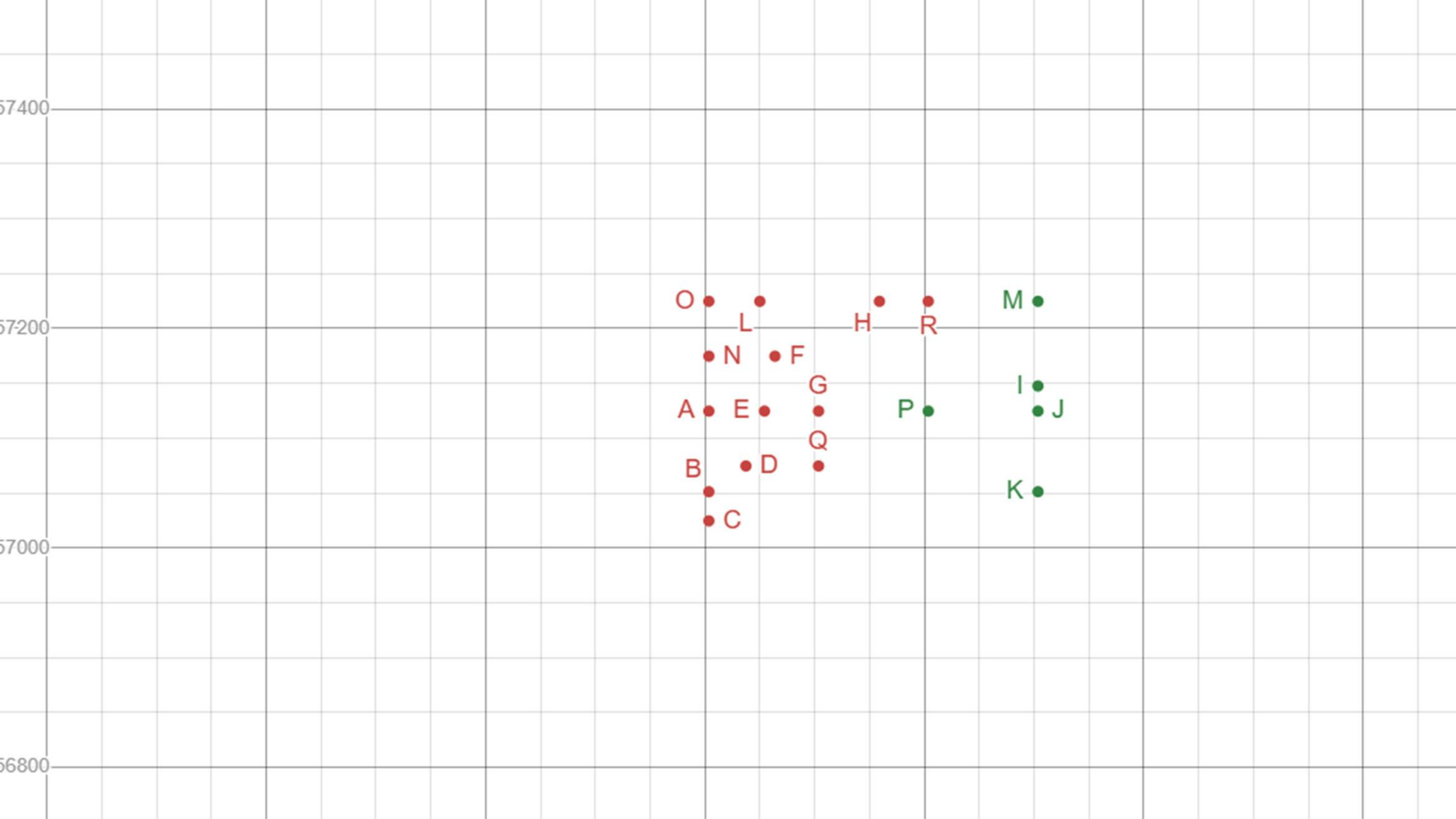
3.3 VÍ DỤ

Bước 3:Xây cụm khởi tạo (Initial Cluster)

Chọn tất cả điểm nằm trong bán kính Eps = 126.3139 quanh I:

Khoảng cách ≤ **126.3139** : {J,K,M,P,R}

⟶ **Cluster 2** = {I,J,K,M,P}


57400
57200
57000
56800
O
L
H
R
M
N
F
G
I
A
E
P
J
Q
B
D
K
C

Hạn chế

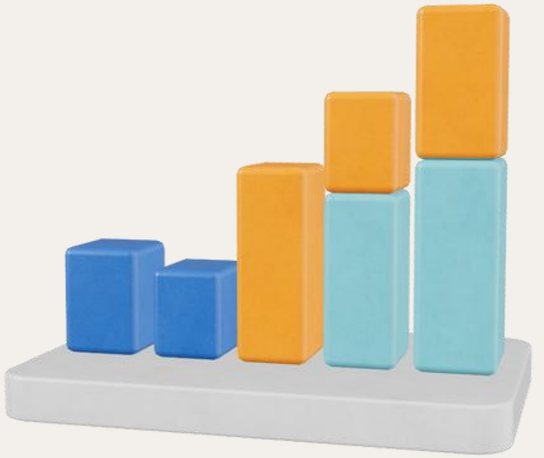
1. Kích thước cụm: Hoạt động chưa tốt với các cụm có kích thước rất nhỏ
2. Sự chồng lấn: Gặp khó khăn khi các cụm có sự chồng lấn (overlapping) đáng kể.
3. Phân bố dữ liệu: Hiệu quả giảm sút đối với các cụm có phân bố không theo dạng Gaussian.
4. Dữ liệu nhiều chiều: Việc sử dụng t-SNE đơn giản để giảm chiều dữ liệu xuống 2 chiều gây mất mát thông tin đáng kể đối với dữ liệu có số chiều cao (high-dimensional data).

HƯỚNG PHÁT TRIỂN

- Cải thiện xử lý cụm kích thước nhỏ: Nghiên cứu các phương pháp thích nghi mới trong bước trích xuất cụm ban đầu, nhằm kiểm soát kích thước vùng lõi, tránh việc tạo ra các cụm ban đầu quá lớn nuốt chửng các cụm nhỏ.

- Giải quyết vấn đề chồng lấn cụm (Overlapping): Kế thừa và tích hợp các lý thuyết từ mô hình hỗn hợp Gaussian (Gaussian Mixture Models) để xử lý hiệu quả hơn các trường hợp dữ liệu giữa các cụm bị chồng lấn nghiêm trọng.

HƯỚNG PHÁT TRIỂN



- Xử lý phân bố phi Gaussian: Khai thác sâu hơn thuật toán DBSCAN, đặc biệt là việc sử dụng linh hoạt các ngưỡng mật độ để khắc phục vấn đề phân bố mật độ dữ liệu không đều và không tuân theo quy luật Gaussian.

- Nâng cao hiệu quả trên dữ liệu chiều cao: Thay thế phương pháp t-SNE đơn giản bằng các kỹ thuật giảm chiều dữ liệu hiệu quả hơn để bảo toàn thông tin. Đồng thời, nghiên cứu áp dụng các phương pháp học độ tương đồng (similarity learning), ví dụ như thuật toán SPRG, để cải thiện độ chính xác trong không gian nhiều chiều.