Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh

Chương 5
Gom Cụm Dữ Liệu – Data Clustering

TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://www.cse.hcmut.edu.vn/staff/Staff/quangtran
http://researchmap.jp/quang

Nội Dung

1. Tổng quan về gom cụm dữ liệu
2. Gom cụm dữ liệu bằng phân hoạch
3. Gom cụm dữ liệu bằng phân cấp
4. Gom cụm dữ liệu dựa trên mật độ
5. Gom cụm dữ liệu dựa trên mô hình
6. Các phương pháp gom cụm dữ liệu khác
7. Tóm tắt

TÀI LIỆU THAM KHẢO

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

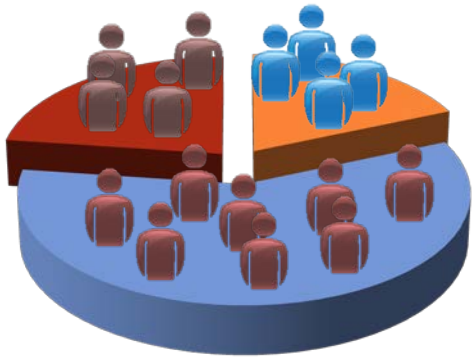
[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

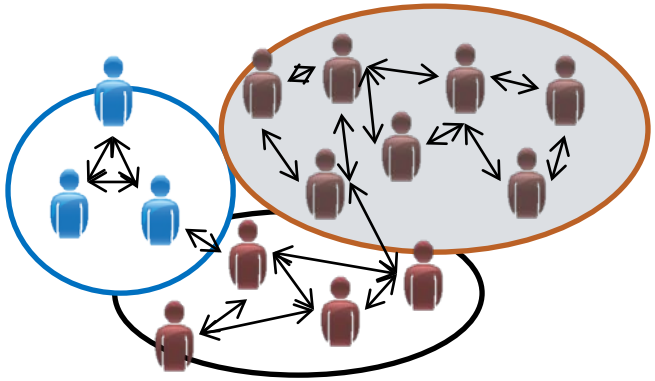
[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

1. Tổng Quan về Gom Cụm DL

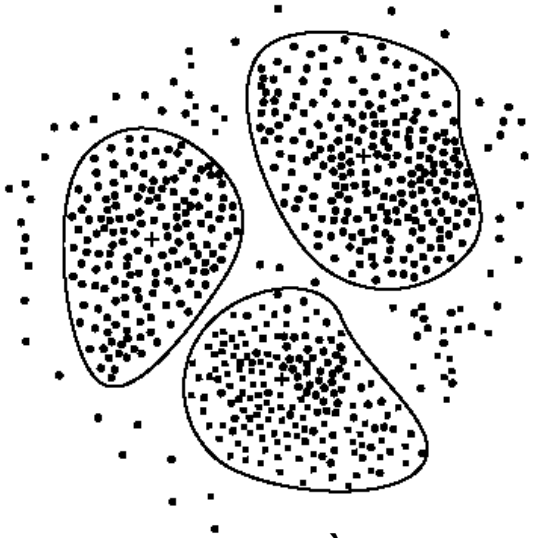
- Các tình huống



Gom nhóm khách hàng



Phân nhóm các mối quan hệ trong mạng xã hội

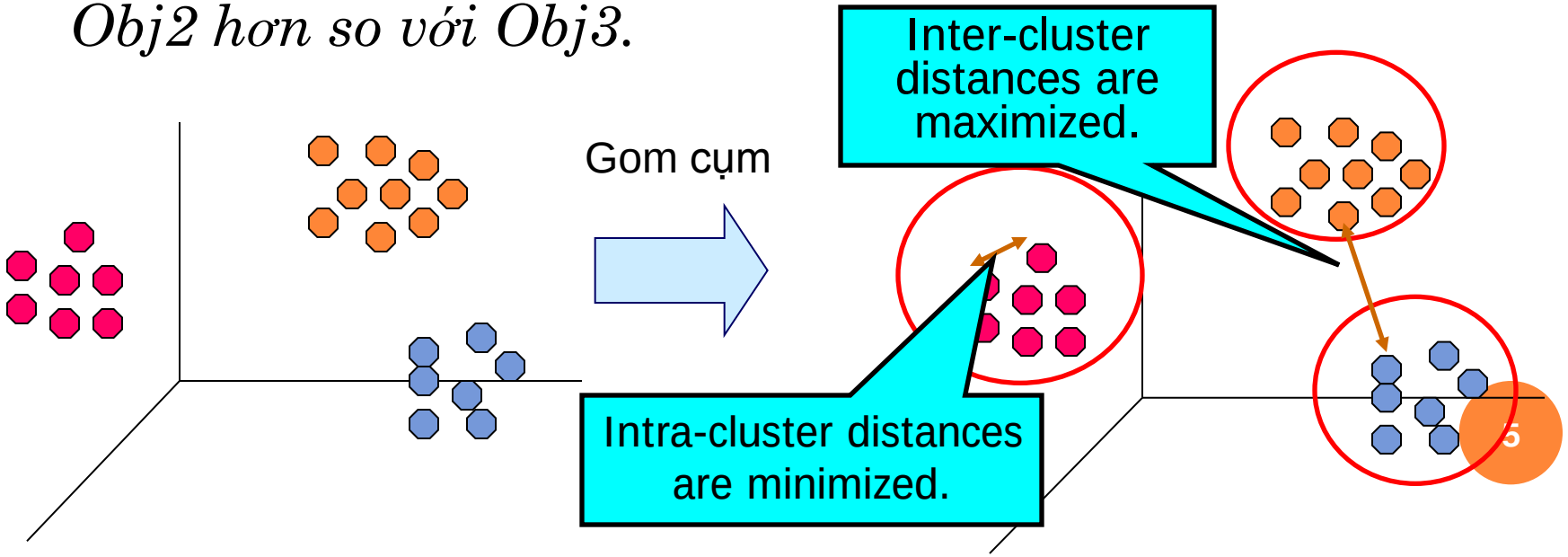


Tìm các phần tử biên, giảm thiểu nhiễu

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

- Gom cụm là quá trình gom nhóm/cụm dữ liệu/đối tượng
- Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác
  - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với Obj3.*



1. Tổng Quan về Gom Cụm DL

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm

Ma trận dữ liệu (Data matrix)

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

-n đối tượng (objects)

-p biến/thuộc tính (variables/attributes)

Ma trận sai biệt (Dissimilarity matrix)

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

-d(i,j) khoảng cách giữa đối tượng i và j, được tính tùy thuộc vào kiểu của thuộc tính (biến)

Note: $d(i,i)=0$; $d(i,j)= d(j,i) \geq 0$; $d(i,j)= d(i,k)+ d(k,j)$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

- Vector objects: i và j được biểu diễn dưới dạng vectors x, y
- Độ tương tự giữa i, j được tính bởi độ đo cosin

$$s(x, y) = \frac{x^T . y}{|x||y|}$$ Với $$x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix}; y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{bmatrix}$$

Hay

$s(x, y) = (x1*y1 + \dots + xp*yp)/((x1^2 + \dots + xp^2)^{1/2}*(y1^2 + \dots + yp^2)^{1/2})$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

- Cách tính khoảng cách tùy thuộc vào kiểu thuộc tính

✓Thuộc tính có giá trị theo khoảng (Interval-scaled variables/attributes)

✓Thuộc tính nhị phân (Binary variables/attributes)

✓Thuộc tính phân loại (Categorical variables/attributes)

✓Thuộc tính có thứ tự (Ordinal variables/attributes)

✓Thuộc tính có giá trị theo hệ số (Ratio-scaled variables/attributes)

✓Các thuộc tính phức hợp (Variables/attributes of mixed types)

1. Tổng Quan về Gom Cụm DL

- Interval-scaled variables/attributes

Mean absolute deviation $$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

Mean $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

Z-score measurement $$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Note: dùng z_{if} thay cho x_{if}; i=1..n, f=1..p

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

- Độ đo Euclidean

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

- Độ đo Minkowski

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

- Độ đo Manhattan

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

- Binary variables/attributes

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p (= a + b + c + d)

Hệ số so trùng đơn giản (nếu symmetric): $$d(i,j) = \frac{b+c}{a+b+c+d}$$

Hệ số so trùng Jaccard (nếu asymmetric): $$d(i,j) = \frac{b+c}{a+b+c}$$

1. Tổng Quan về Gom Cụm DL

- Binary variables/attributes
  - Ví dụ

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - gender: symmetric (Xác suất dl nhận giá trị “M”, “F” là như nhau)
  - Binary attributes còn lại: asymmetric
  - Y, P → 1, N → 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

1. Tổng Quan về Gom Cụm DL

- Variables/attributes of mixed types

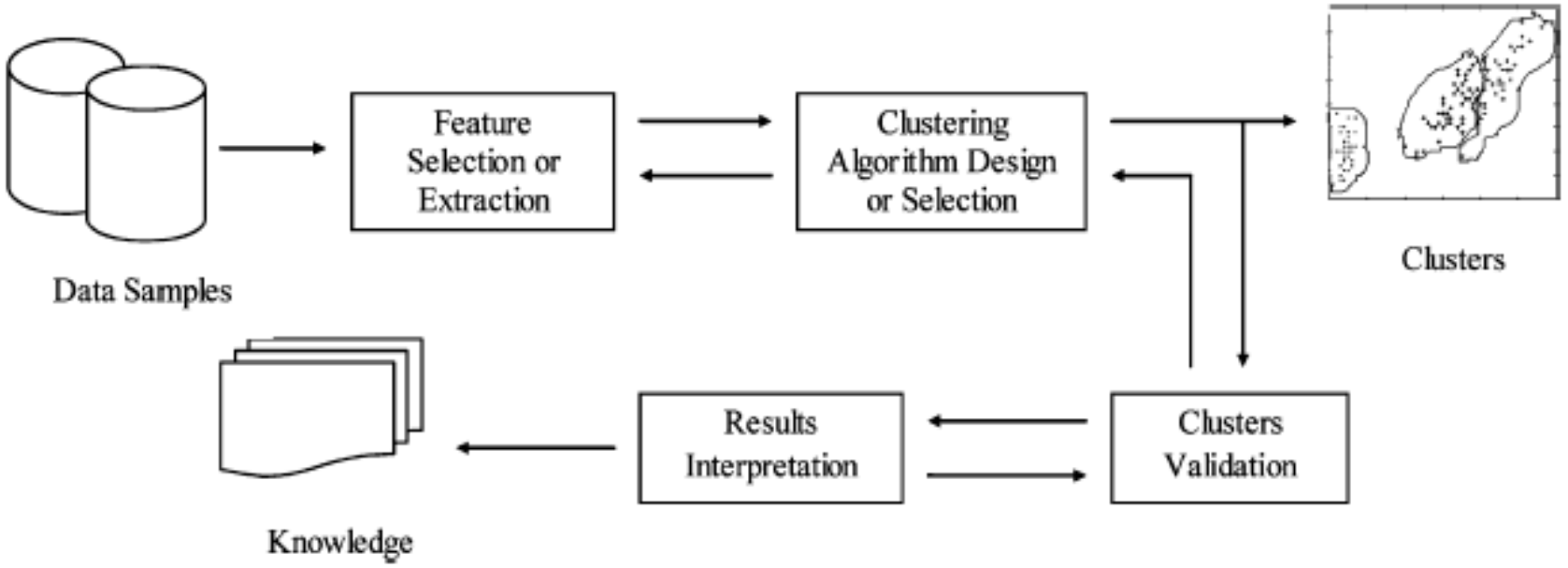
$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - Nếu x_{if} hoặc x_{jf} bị thiếu (missing) thì $\delta_{ij}^{(f)} = 0$
  - *f (variable/attribute)*: binary (nominal): $d_{ij}^{(f)} = 0$ nếu $x_{if} = x_{jf}$; $d_{ij}^{(f)} = 1$ trong các trường hợp khác
  - *f* : interval-scaled (Minkowski, Manhattan, Euclidean)
  - *f* : ordinal (có thứ tự, e.g. huy chương vàng, bạc, đồng) hoặc ratio-scaled: chuyển các x_{ij} thành $r_{if} = \{1,..,M_f\}$, sau đó thay x_{ij} bởi z_{ij}

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

1. Tổng Quan về Gom Cụm DL

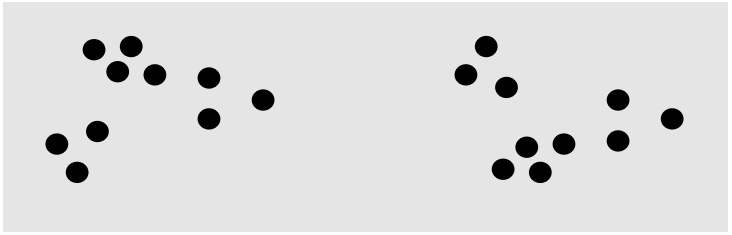
- Quá trình gom cụm dữ liệu



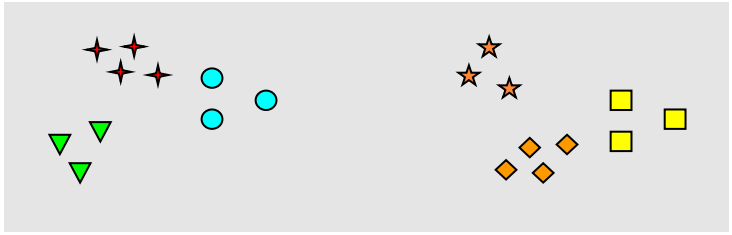
R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), May 2005, pp. 645-678.

1. Tổng Quan về Gom Cụm DL

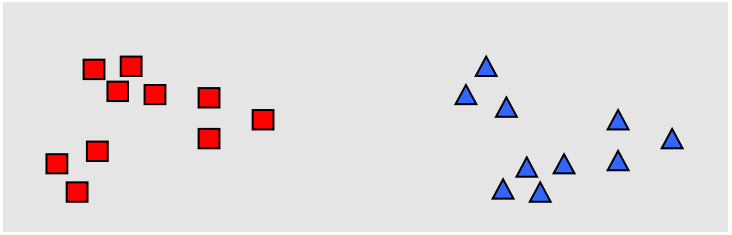
- Mỗi cụm nên có bao nhiêu phần tử?
- Các phân tử nên được gom vào bao nhiêu cụm?
- Bao nhiêu cụm nên được tạo ra?



Bao nhiêu cụm?



6 cụm?



2 cụm?

4 cụm?

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

1. Tổng Quan về Gom Cụm DL

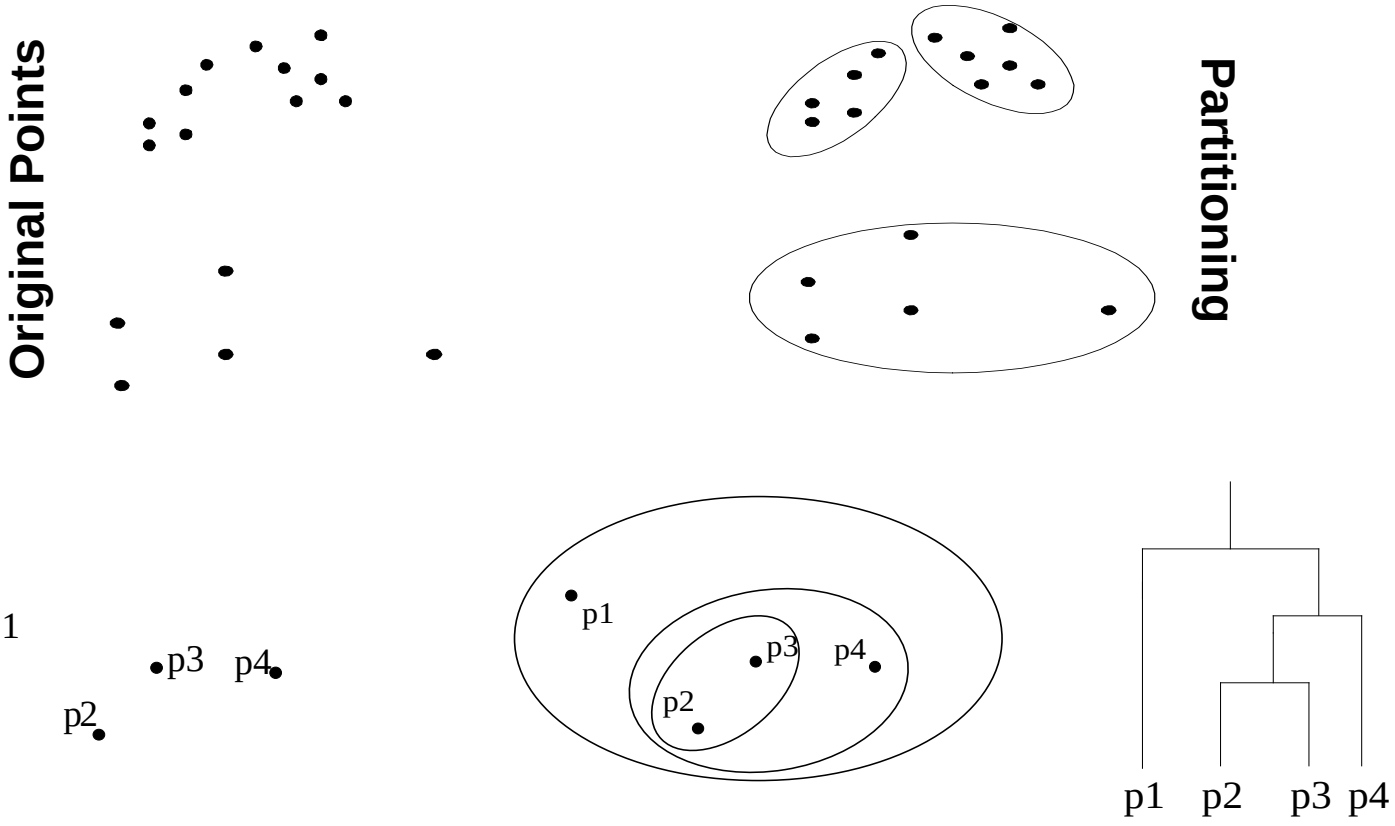
- Các yêu cầu tiêu biểu về việc gom cụm dữ liệu
  - Khả năng co giãn với tập dữ liệu (scalability)
  - Khả năng xử lý nhiều kiểu thuộc tính khác nhau
  - Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape)
  - Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập
  - Khả năng xử lý dữ liệu có nhiễu (noisy data)
  - Khả năng gom cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records)
  - Khả năng xử lý dữ liệu đa chiều (high dimensionality)
  - Khả diễn và khả dụng (interpretability and usability)

1. Tổng Quan về Gom Cụm DL

- Các phương pháp gom cụm dữ liệu tiêu biểu
  - Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó
  - Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó
  - Dựa trên mật độ (density-based): dựa trên độ kết nối (connectivity) và mật độ (density)
  - Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu nhất
  - …

1. Tổng Quan về Gom Cụm DL

- Các phương pháp gom cụm dữ liệu tiêu biểu



Original Points **Hierarchical**

1. Tổng Quan về Gom Cụm DL

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Đánh giá ngoại (external validation): Đ. giá kết quả gom cụm dựa vào cấu trúc được chỉ định trước cho tập dữ liệu
  - Đánh giá nội (internal validation): Đ. giá kết quả gom cụm theo số lượng các vector của chính tập dữ liệu (ma trận gần – proximity matrix)
  - Đánh giá tương đối (relative validation): Đ. giá kết quả gom cụm bằng việc so sánh các kết quả gom cụm khác ứng với các bộ trị thông số khác nhau

→ Tiêu chí đánh giá và chọn kết quả gom cụm tối ưu

- *Độ nén (compactness): các đối tượng trong cụm nên gần nhau*
- *Độ phân tách (separation): các cụm nên xa nhau*

1. Tổng Quan về Gom Cụm DL

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Đánh giá ngoại (external validation)
    - Độ đo: Rand statistic, Jaccard coefficient, Folkes and Mallows index, …
  - Đánh giá nội (internal validation)
    - Độ đo: Silhouette index, Dunn's index, …
  - Đánh giá tương đối (relative validation)

1. Tổng Quan về Gom Cụm DL

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Các độ đo đánh giá ngoại (external validation measures – contingency matrix)

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i}\log\frac{p_{ij}}{p_i})$	$[0, \log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i[2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$	$[0, 2\log\max(K, K')]$
5	Mutual Information	MI	$\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$	$(0, \log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i\binom{n_{i\cdot}}{2} - \sum_j\binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}{\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}[\binom{n}{2}-\sum_i\binom{n_{i\cdot}}{2}][\binom{n}{2}-\sum_j\binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i\binom{n_{i\cdot}}{2} - 2\sum_j\binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j\binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n}\max_\sigma\sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i\max_j n_{ij} - \sum_j\max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n.$

J. Wu and et al. Adapting the Right Measures for K-means Clustering. KDD'09, Paris, France, July 2009.

2. Gom Cụm DL Bằng Phân Hoạch

- Đánh giá kết quả gom cụm

Partition C

Partition P	C_1	C_2	$\cdots$	$C_{K'}$	$\sum$
P_1	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1.}$
P_2	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2.}$
.	.	.	$\cdots$	.	.
P_K	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K.}$
$\sum$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.K'}$	n

Contingency matrix

-Partition P: kết quả gom cụm trên n đối tượng

-Partition C: các cụm thật sự của n đối tượng

-$n_{ij} = |P_i \cap C_j|$: số đối tượng trong P_i từ C_j

2. Gom Cụm DL Bằng Phân Hoạch

- Đánh giá kết quả gom cụm

I	C_1	C_2	C_3	$\sum$
P_1	3	4	12	19
P_2	8	3	12	23
P_3	12	12	0	24
$\sum$	23	19	24	66

II	C_1	C_2	C_3	$\sum$
P_1	0	7	12	19
P_2	11	0	12	23
P_3	12	12	0	24
$\sum$	23	19	24	66

Kết quả gom cụm theo phương án I và II

-Partition P: kết quả gom cụm trên n (=66) đối tượng

-Partition C: các cụm thật sự của n (=66) đối tượng

-$n_{ij} = |P_i \cap C_j|$: số đối tượng trong P_i từ C_j

2. Gom Cụm DL Bằng Phân Hoạch

- Đánh giá kết quả gom cụm
  - Entropy (trị nhỏ khi chất lượng gom cụm tốt)

$$Entropy(I) = -\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i}\right)$$

$$= -\sum_i \frac{n_i}{n} \left(\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i}\right)$$

$$= -\frac{19}{66}\left(\frac{3}{19}\log\frac{3}{19} + \frac{4}{19}\log\frac{4}{19} + \frac{12}{19}\log\frac{12}{19}\right)$$

$$-\frac{23}{66}\left(\frac{8}{23}\log\frac{8}{23} + \frac{3}{23}\log\frac{3}{23} + \frac{12}{23}\log\frac{12}{23}\right)$$

$$-\frac{24}{66}\left(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24}\right)$$

$$= ???$$

$$Entropy(II) = -\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i}\right)$$

$$= -\sum_i \frac{n_i}{n} \left(\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i}\right)$$

$$= -\frac{19}{66}\left(\frac{0}{19}\log\frac{0}{19} + \frac{7}{19}\log\frac{7}{19} + \frac{12}{19}\log\frac{12}{19}\right)$$

$$-\frac{23}{66}\left(\frac{11}{23}\log\frac{11}{23} + \frac{0}{23}\log\frac{0}{23} + \frac{12}{23}\log\frac{12}{23}\right)$$

$$-\frac{24}{66}\left(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24}\right)$$

$$= ???$$

→ Gom cụm theo phương án I hay phương án II tốt???

2. Gom Cụm DL Bằng Phân Hoạch: K-Means

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

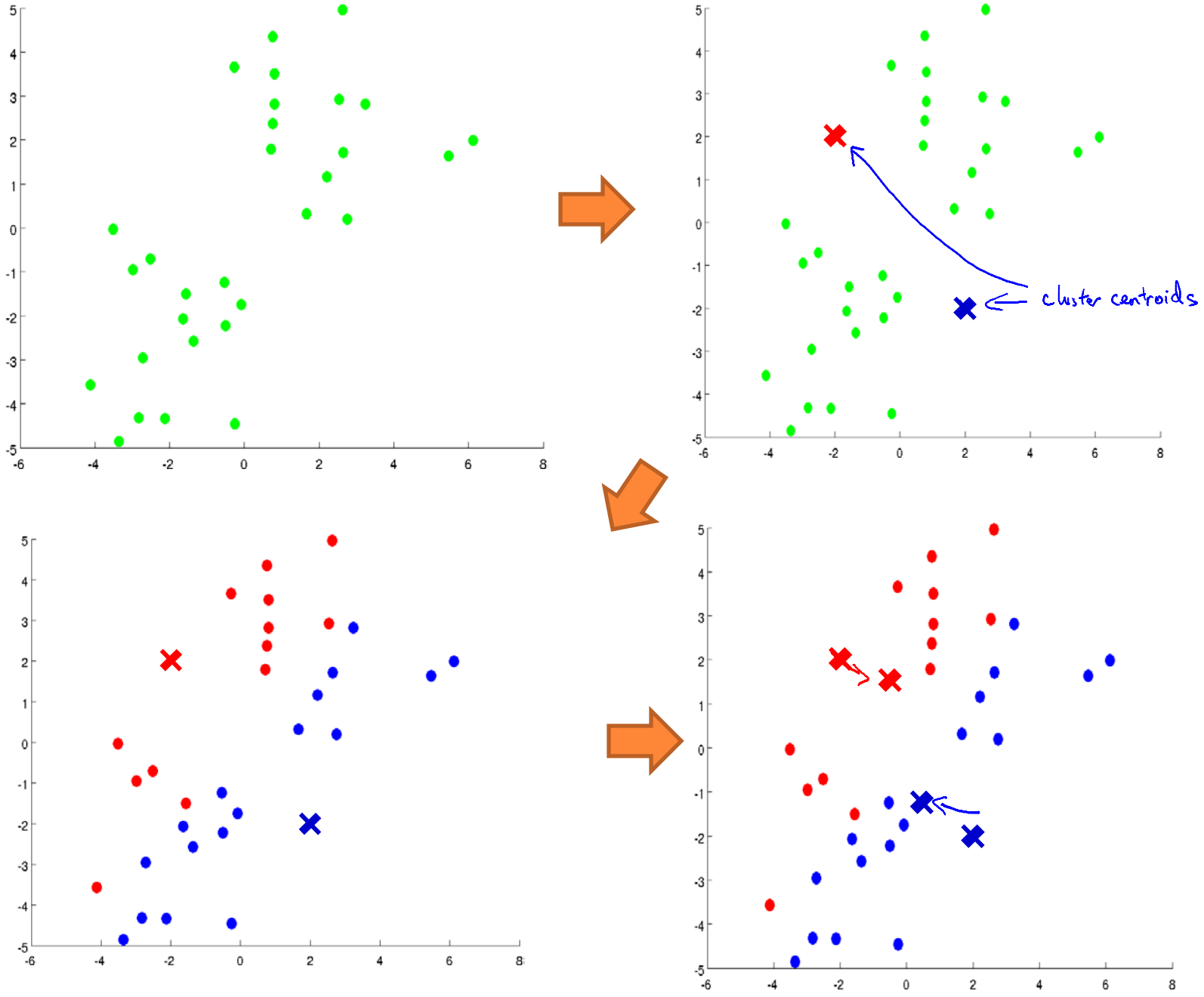
Input:

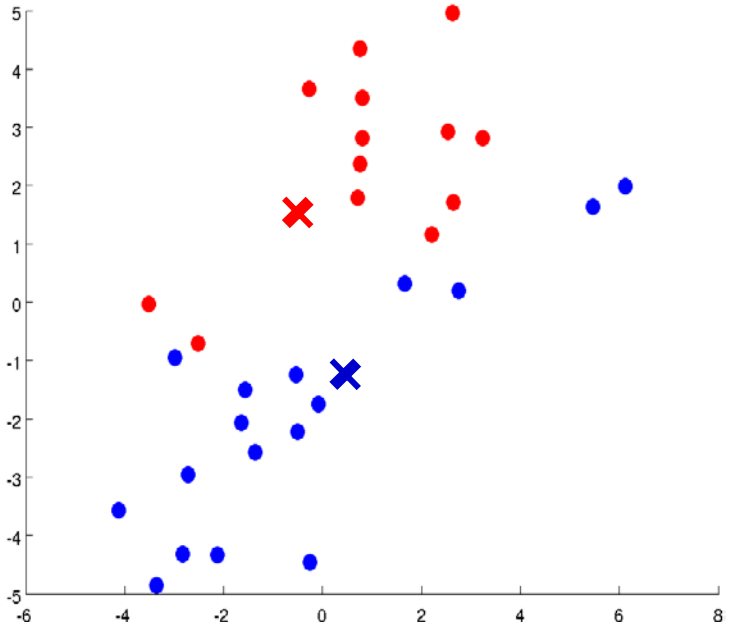
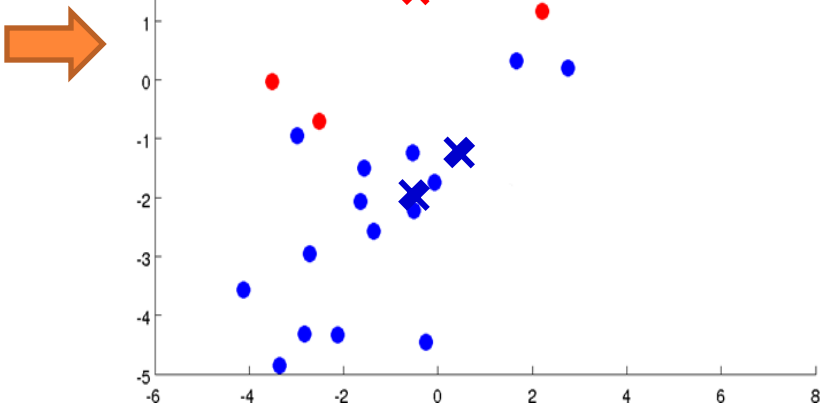
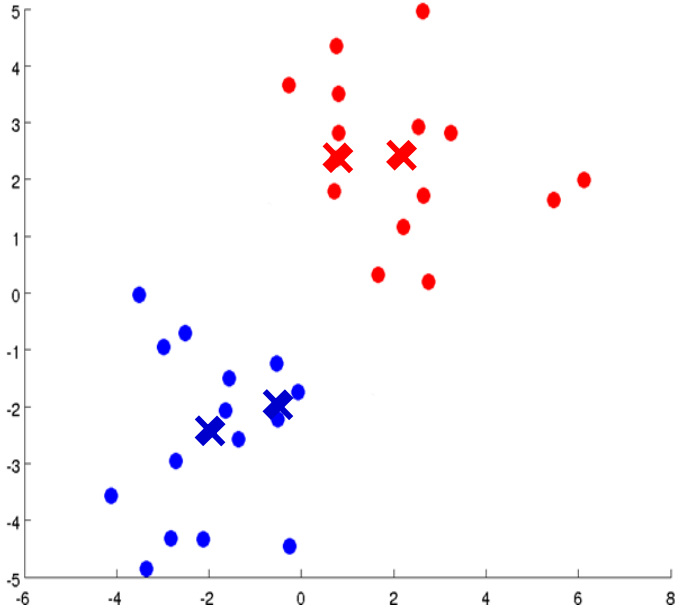
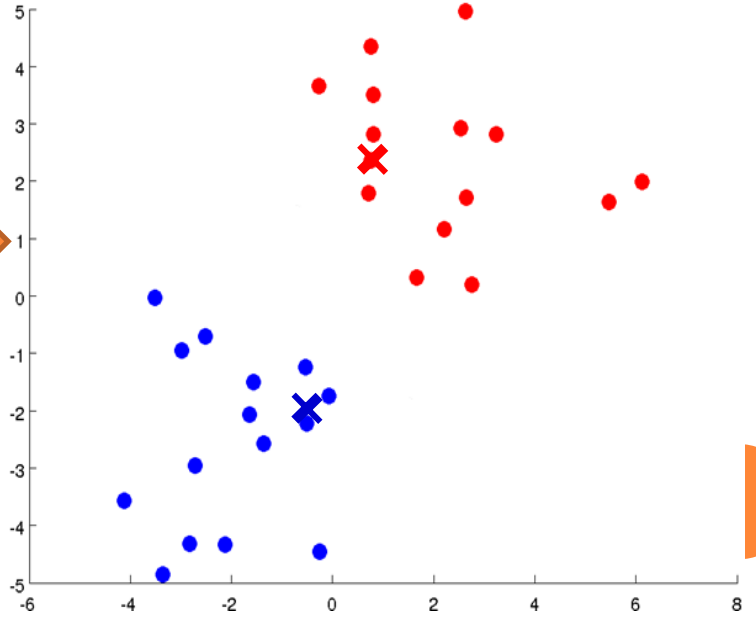
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

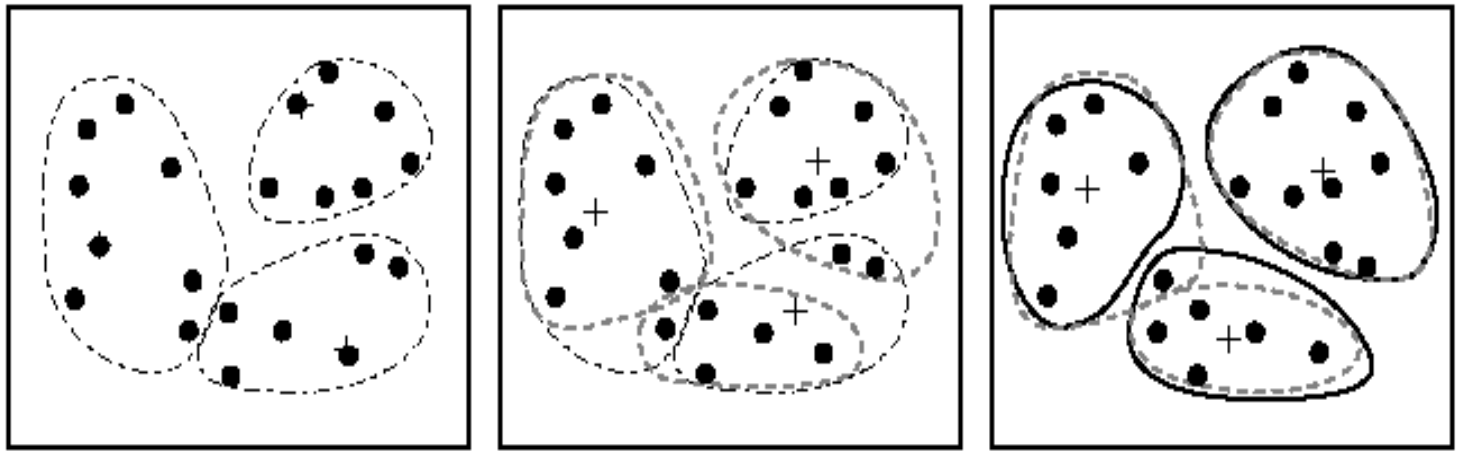
Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;


cluster centroids

2. Gom Cụm DL Bằng Phân Hoạch



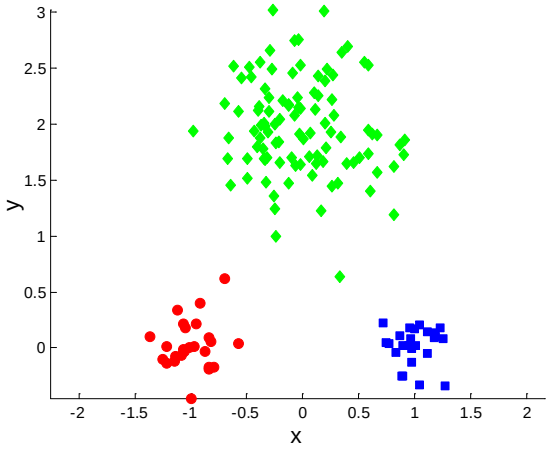
(a) (b) (c)

Clustering of a set of objects based on the *k*-means method. (The mean of each cluster is marked by a "+".)

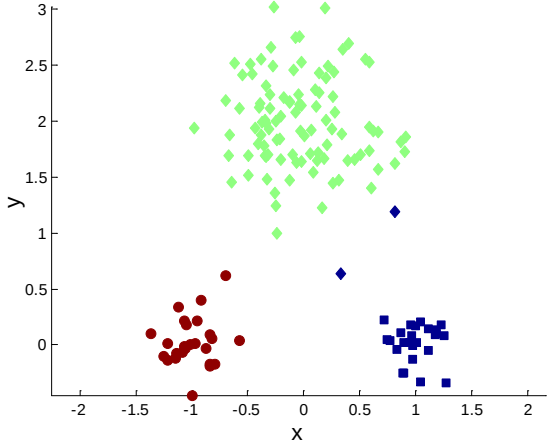
2. Gom Cụm DL Bằng Phân Hoạch

- Đặc điểm của giải thuật k-means
  - Bài toán tối ưu hóa với "Cực trị cục bộ"
  - Mỗi cụm được đặc trưng hóa bởi trung tâm của cụm (i.e. đối tượng trung bình (mean))
    - ✓ Số cụm k nên là bao nhiêu?
    - ✓ Độ phức tạp: $O(nkt)$, với n là số đối tượng, k là số cụm, t là số lần lặp ($k << n$, $t << n$)
  - Ảnh hưởng bởi nhiễu (các phần tử kì dị/biên)
  - Không phù hợp cho việc khai phá ra các cụm có dạng không lồi (nonconvex) hay các cụm có kích thước rất khác nhau
    - ✓ Kết quả gom cụm có dạng siêu cầu (hyperspherial)
    - ✓ Kích thước các cụm kết quả thường đồng đều (relatively uniform sizes)

2. Gom Cụm DL Bằng Phân Hoạch

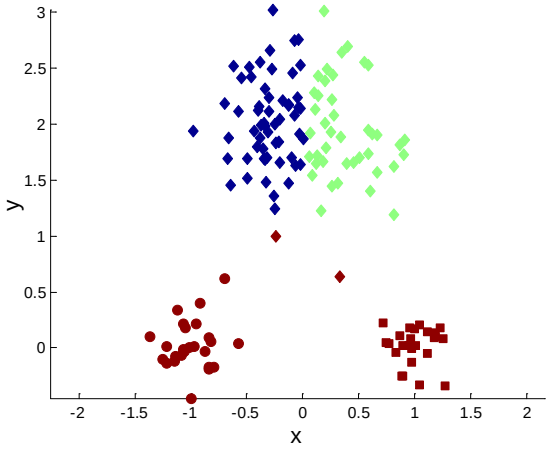


Original Points



Optimal Clustering

Note: Đọc thêm các giải thuật gom cụm bằng phân hoạch khác như giải thuật PAM(k-medoids)



Sub-optimal Clustering

3. Gom Cụm DL Bằng Phân Cấp

- Gom cụm dữ liệu bằng phân cấp (hierarchical clustering): nhóm các đối tượng vào cây phân cấp của các cụm
  - Agglomerative: bottom-up (trộn các cụm)
  - Divisive: top-down (phân tách các cụm)

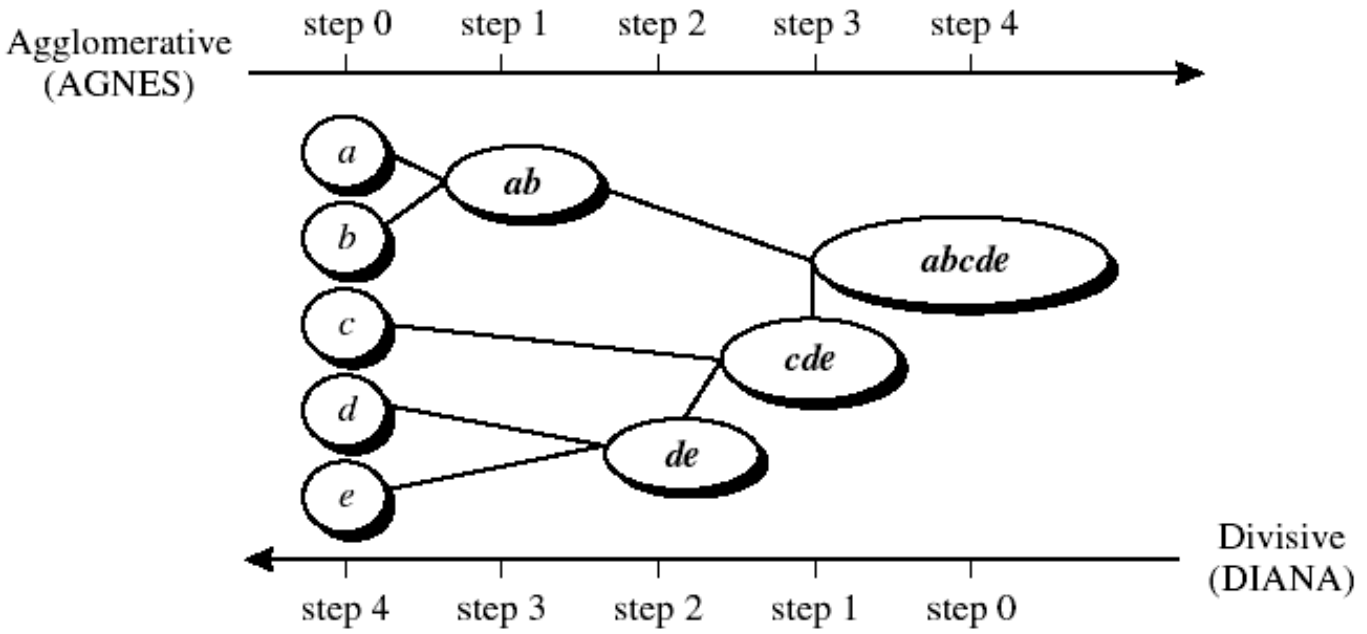
→ Không yêu cầu thông số nhập k (số cụm)

→ Yêu cầu điều kiện dừng

→ Không thể quay lui ở mỗi bước trộn/phân tách

3. Gom Cụm DL Bằng Phân Cấp

- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting) → bottom-up
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis) → top-down



Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

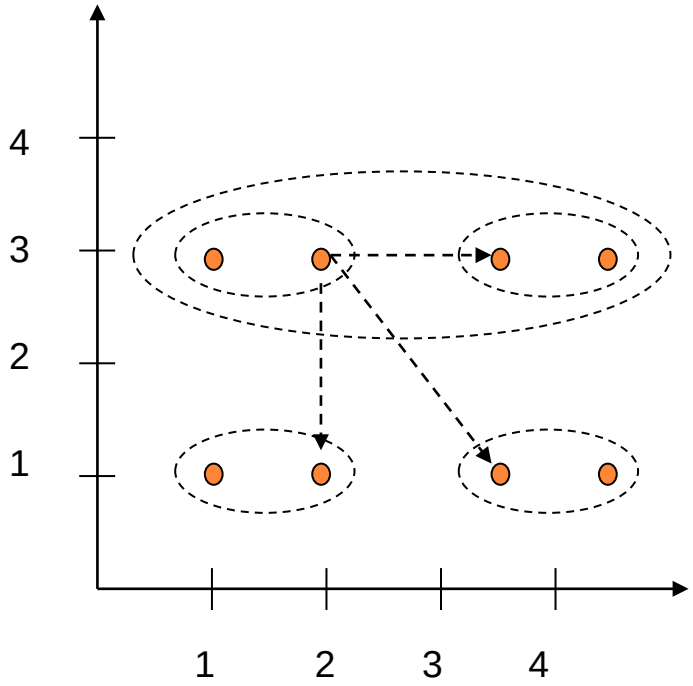
3. Gom Cụm DL Bằng Phân Cấp

- AGNES (Agglomerative NESting)
  - Khởi đầu, mỗi đối tượng tạo thành một cụm (n cụm)
  - Các cụm sau đó được trộn lại theo một tiêu chí nào đó
    - Cách tiếp cận single-linkage: cụm C1 và C2 được trộn lại nếu khoảng cách giữa 2 đối tượng từ C1 và C2 là ngắn nhất
  - Quá trình trộn các cụm được lặp lại đến khi tất cả các đối tượng tạo thành một cụm duy nhất

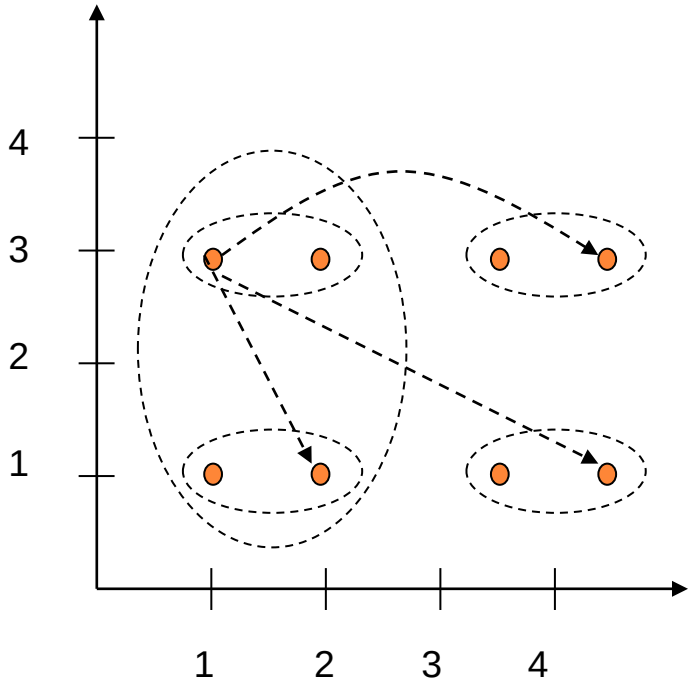
- DIANA (Divisive ANAlysis)
  - Khởi đầu, tất cả các đối tượng tạo thành một cụm duy nhất
  - Một cụm được phân tách theo một tiêu chí nào đó đến khi mỗi cụm chỉ có một đối tượng

3. Gom Cụm DL Bằng Phân Cấp

Single-linkage



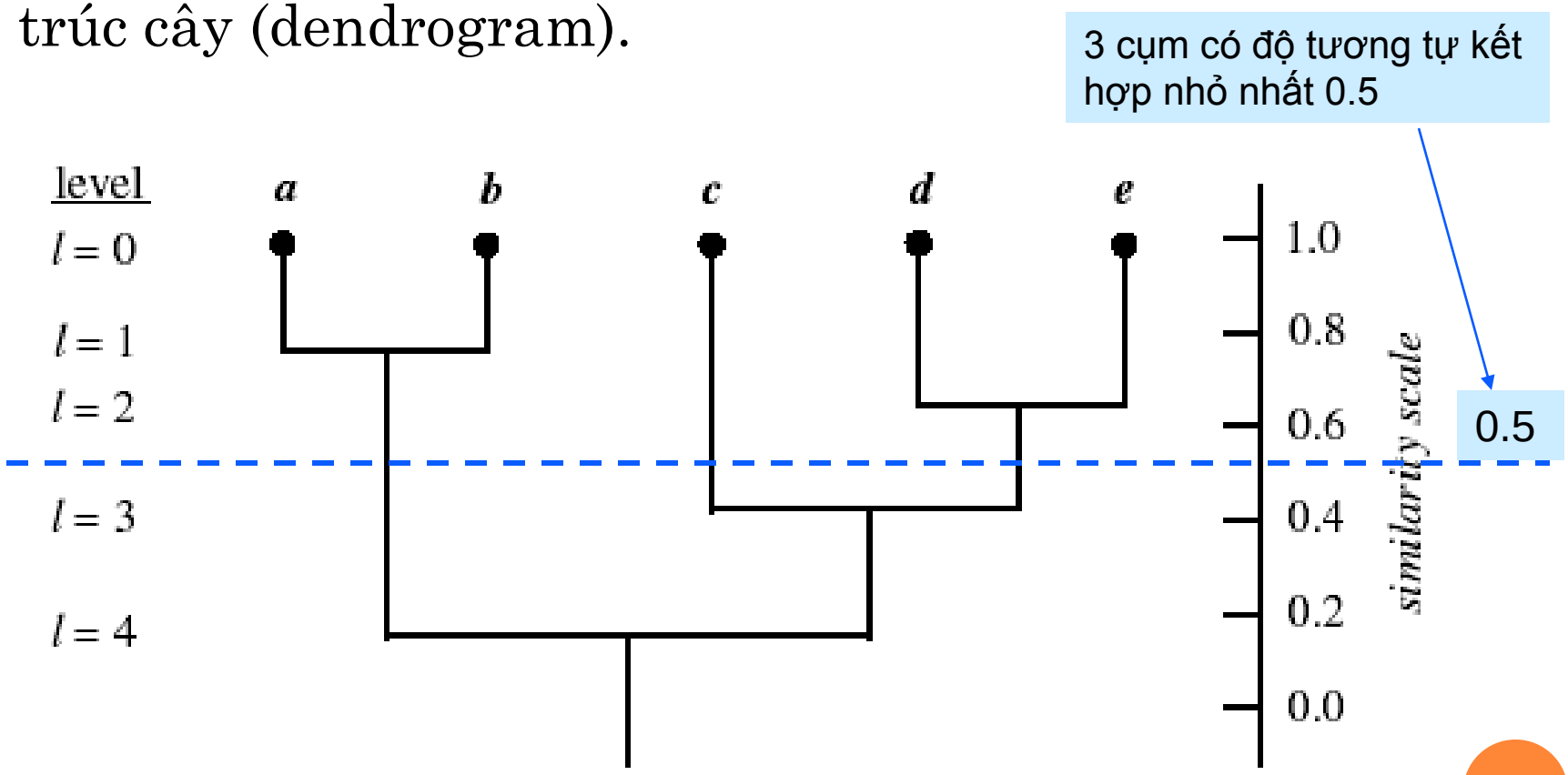
Complete-linkage



Tiêu chí trộn các cụm: single-linkage và complete-linkage

3. Gom Cụm DL Bằng Phân Cấp

- Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).



3. Gom Cụm DL Bằng Phân Cấp

- Các độ đo dùng đo khoảng cách giữa các cụm C_i và C_j

$$\text{Minimum distance}: \quad d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

$$\text{Maximum distance}: \quad d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

$$\text{Mean distance}: \quad d_{mean}(C_i, C_j) = |m_i - m_j|$$

$$\text{Average distance}: \quad d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

p, p': các đối tượng

|p-p'|: khoảng cách giữa p và p'

m_i, m_j: đối tượng trung bình của C_i, C_j, tương ứng

n_i, n_j: số lượng đối tượng của C_i, C_j, tương ứng

3. Gom Cụm DL Bằng Phân Cấp

- Một số giải thuật gom cụm dữ liệu bằng phân cấp
  - BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): phân hoạch các đối tượng dùng cấu trúc cây theo độ co giãn của phân giải (scale of resolution)
  - ROCK (Robust Clustering using linKs): gom cụm dành cho các thuộc tính rời rạc (categorical/discrete attributes), trộn các cụm dựa vào sự kết nối lẫn nhau giữa các cụm
  - Chameleon: mô hình động để xác định sự tương tự giữa các cặp cụm

3. Gom Cụm DL Bằng Phân Cấp

- Một số vấn đề với gom cụm dữ liệu bằng phân cấp
  - Chọn điểm trộn/phân tách phù hợp
  - Khả năng co giãn (scalability)
    - Mỗi quyết định trộn/phân tách yêu cầu kiểm tra/đánh giá nhiều đối tượng/cụm

→ Tích hợp gom cụm dữ liệu bằng phân cấp với các kỹ thuật gom cụm khác

→ Gom cụm nhiều giai đoạn (multiple-phase clustering)

4. Gom Cụm DL Dựa vào Mật Độ

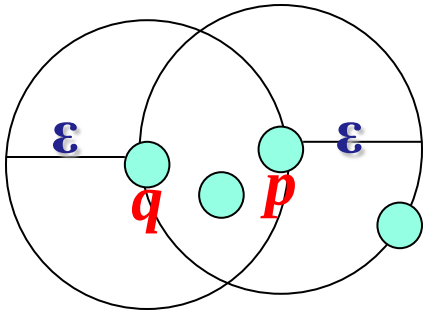
- Gom cụm dl dựa trên mật độ (Density-based clustering)
  - Mỗi cụm là một vùng dày đặc (dense region) các đối tượng
  - Các đối tượng trong vùng thưa hơn được xem là nhiễu
  - Mỗi cụm có dạng tùy ý

- Một số giải thuật tiêu biểu
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise): *Phân tích các điểm kết nối nhau dựa vào mật độ*
  - OPTICS (Ordering Points To Identify the Clustering Structure): *Tạo ra thứ tự các điểm dữ liệu tùy vào cấu trúc gom cụm dựa vào mật độ của tập dữ liệu*
  - DENCLUE (DENsity-based CLUstEring): *Gom cụm dựa vào các hàm phân bố mật độ*

4. Gom Cụm DL Dựa vào Mật Độ

- Các khái niệm
  - ε: bán kính vùng láng giềng của một đối tượng
  - ε-neighborhood: Số đối tượng trong vùng láng giềng
  - Đối tượng lõi (core object) là đối tượng có

    ε-neighborhood $\geq$ ***MinPts***
  - ***Directly density-reachable*** (khả năng đạt được trực tiếp): *q* có thể đạt được trực tiếp từ *p* nếu *q* trong vùng láng giềng ε-neighborhood của *p* và *p* phải là core object.



p: core object (***MinPts*** = 3)

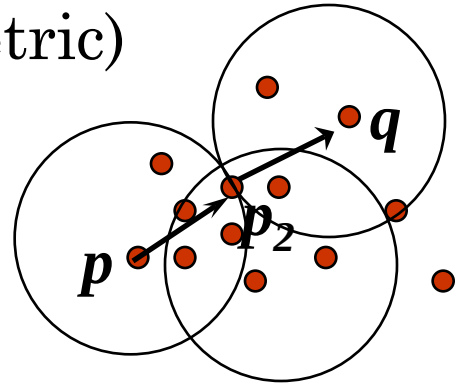
q: không là core object

p: directly density-reachable đối với *q*? **X**

q: directly density-reachable đối với *p*? √

4. Gom Cụm DL Dựa vào Mật Độ

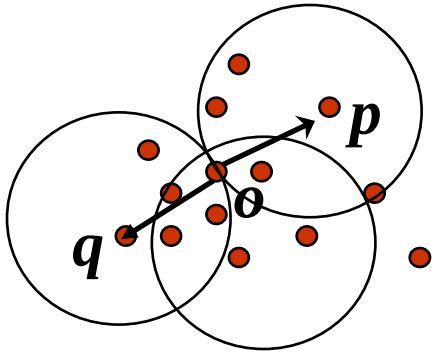
- ***Density-reachable*** (khả năng đạt được):
  - Cho trước tập đối tượng ***D***, ***ε*** và ***MinPts***
  - q ***density-reachable*** từ p nếu ∃ chuỗi các đối tượng $p_1, ..., p_n \in \boldsymbol{D}$ với $p_1 = p$ và $p_n = q$ sao cho p_{i+1} ***directly density-reachable*** từ p_i theo các thông số ***ε*** và ***MinPts***, $1 \le i \le n$.
  - Bao đóng truyền (transitive closure) của directly density-reachable
  - Quan hệ bất đối xứng (asymmetric)

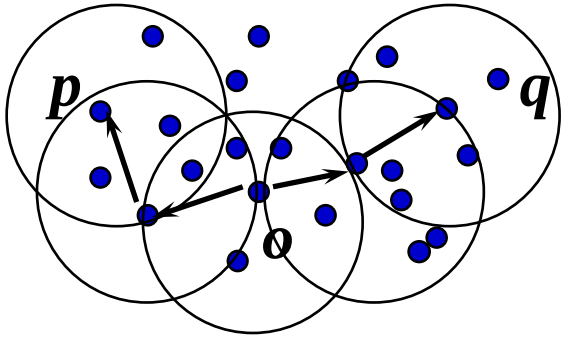


MinPts = 5

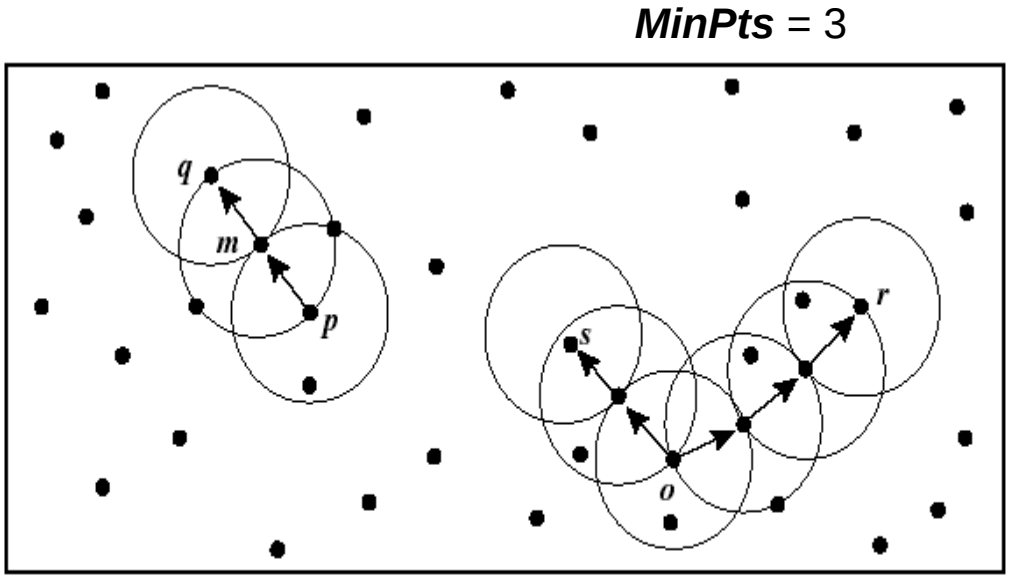
4. Gom Cụm DL Dựa vào Mật Độ

- ***Density-connected*** (nối kết dựa trên mật độ):
  - Cho trước tập các đối tượng ***D***, ***ε*** và ***MinPts***
  - $p, q \in \boldsymbol{D}$
  - q ***density-connected*** với p nếu $\exists\, o \in \boldsymbol{D}$ sao cho cả q và p đều ***density-reachable*** từ o theo các thông số ***ε*** và ***MinPts***.
  - Quan hệ đối xứng

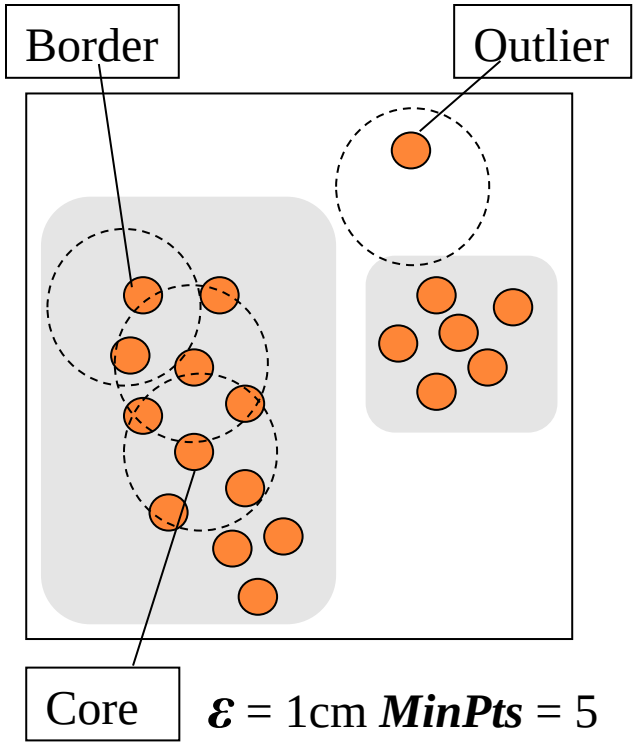




4. Gom Cụm DL Dựa vào Mật Độ



Density reachability and density connectivity in density-based clustering



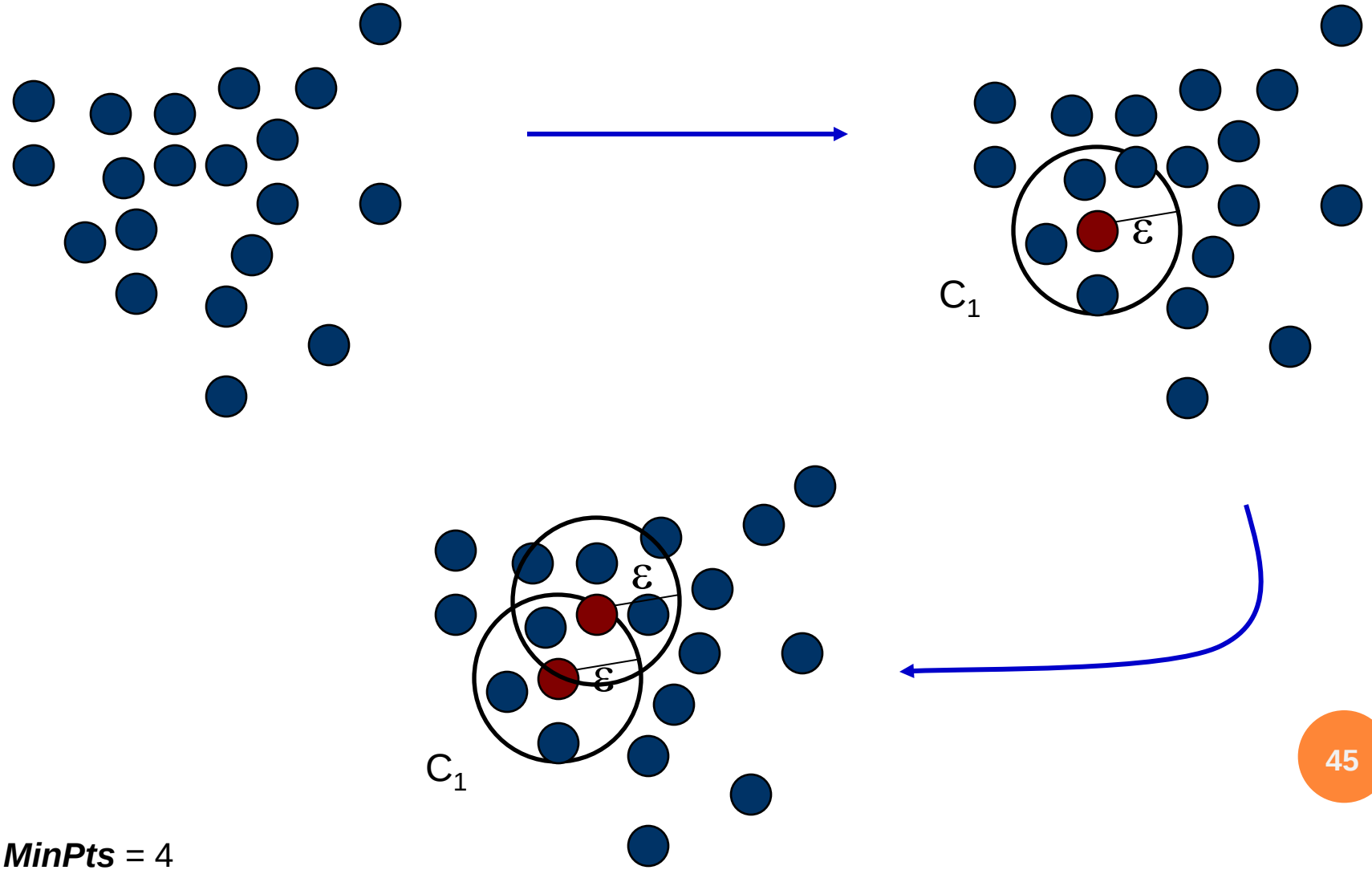
- Cụm dựa trên mật độ (density based cluster): tập tất cả các đối tượng được nối kết với nhau dựa trên mật độ gồm: core objects và border objects
- Đối tượng không thuộc về cụm nào được xem là nhiễu (noise/outlier)

4. Gom Cụm DL Dựa vào Mật Độ

- DBSCAN
  - Input: tập đối tượng ***D***, ***ε***, ***MinPts***
  - Output: density-based clusters (và noises/outliers)

  - Giải thuật
    - 1. Xác định ***ε***–neighborhood của mỗi đối tượng $p \in \boldsymbol{D}$
    - 2. If p là core object, tạo được một cluster
    - 3. Từ bất kì core object p, tìm tất cả các đối tượng ***density-reachable*** và đưa các đối tượng này (hoặc các cluster) vào cùng cluster ứng với p
      - 3.1. Các cluster đạt được (density-reachable cluster) có thể được trộn lại với nhau
      - 3.2. Dừng khi không có đối tượng mới nào được thêm vào

4. Gom Cụm DL Dựa vào Mật Độ



MinPts = 4

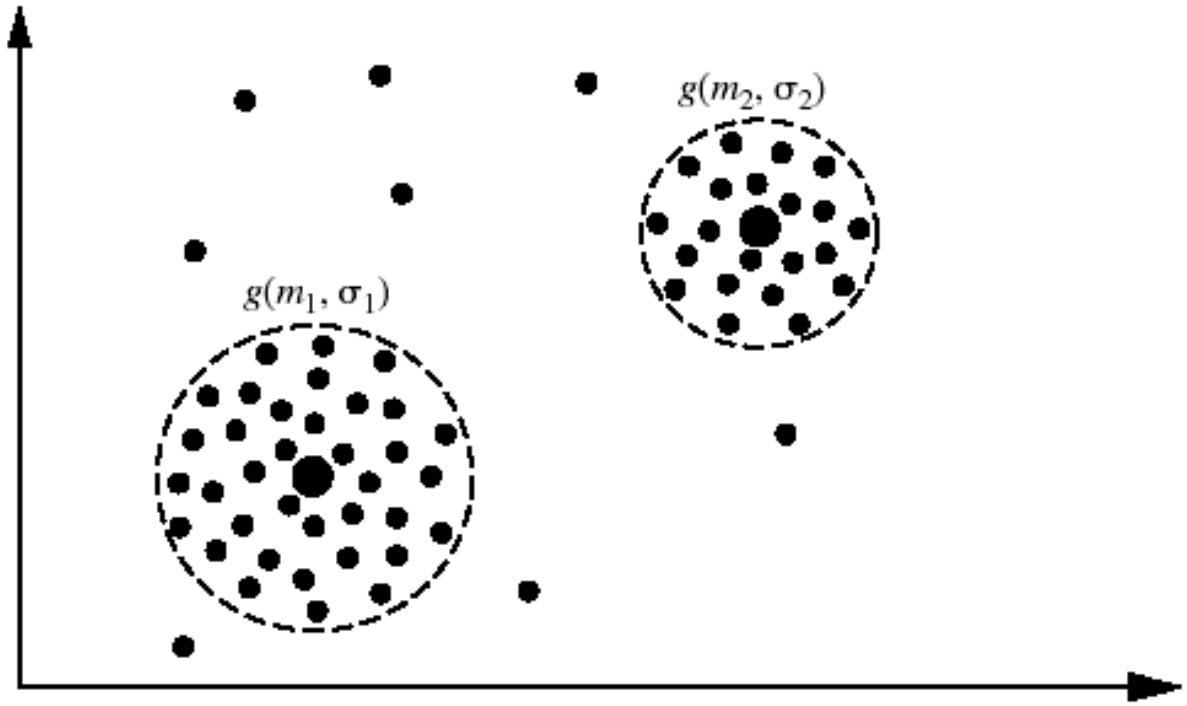
4. Gom Cụm DL Dựa vào Mật Độ

- Đặc điểm của DBSCAN
  - Các cụm có dạng và kích thước khác nhau.
    - Không có giả định về phân bố của các đối tượng dữ liệu
    - Không yêu cầu về số cụm
    - Không phụ thuộc vào cách khởi động (initialization)
    - Yêu cầu định nghĩa của mật độ (density), ε và ***MinPts***
  - Xử lý nhiễu (noise) và các phần tử biên (outliers)
  - Độ phức tạp: O(nlogn) → $O(n^2)$

5. Gom Cụm DL Dựa trên Mô Hình

- Tối ưu hóa sự phù hợp giữa dl và mô hình toán nào đó
  - Giả định: Dl được tạo ra với nhiều sự phân bố xác suất khác nhau
- Các phương pháp
  - Tiếp cận thống kê: Mở rộng của giải thuật gom cụm dựa trên phân hoạch k-means: Expectation-Maximization (EM)
  - Tiếp cận học máy: gom cụm ý niệm (conceptual clustering)
  - Tiếp cận mạng neural: Self-Organizing Feature Map (SOM)

5. Gom Cụm DL Dựa trên Mô Hình



Each cluster can be represented by a probability distribution, centered at a mean, and with a standard deviation. Here, we have two clusters, corresponding to the Gaussian distributions $g(m_1, \sigma_1)$ and $g(m_2, \sigma_2)$, respectively, where the dashed circles represent the first standard deviation of the distributions.

5. Gom Cụm DL Dựa trên Mô Hình

- Giả sử dữ liệu được tạo ra từ nhiều phân bối Gaussian
- Mỗi phân bố Gaussian có bộ thông số Θ (μ_i, Σ_i)
  - Center: μ_i
  - Variance: Σ_i (ignore)
- Tìm cụm hợp lý (k cụm) cho x_i

z_{ij} : if x_i belongs to j-th cluster

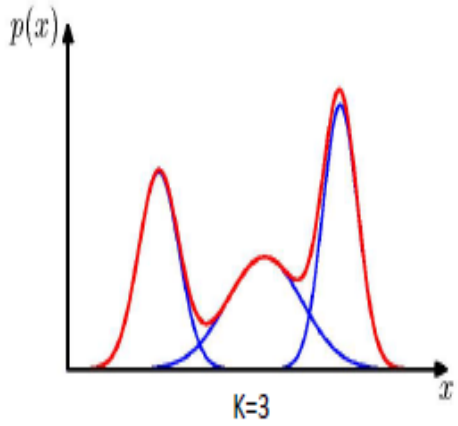
Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



5. Gom Cụm DL Dựa trên Mô Hình

- Probability

$$p(x = x_i)$$

$$p(x = x_i) = \sum_{\mu_j} p(x = x_i, \mu = \mu_j) = \sum_{\mu_j} p(\mu = \mu_j)\, p(x = x_i \mid \mu = \mu_j)$$

$$= \sum_{\mu_j} p(\mu = \mu_j) \frac{1}{\left(2\pi\sigma^2\right)^{d/2}} \exp\left(-\frac{\left\| x_i - \mu_j \right\|_2}{2\sigma^2} \right)$$

- Log-likelihood of data

$$\sum_i \log p(x = x_i) = \sum_i \log \left[\sum_{\mu_j} p(\mu = \mu_j) \frac{1}{\left(2\pi\sigma^2\right)^{d/2}} \exp\left(-\frac{\left\| x_i - \mu_j \right\|_2}{2\sigma^2} \right) \right]$$

  - Tìm giải thuật để đạt Maximize Log-likelihood

5. Gom Cụm DL Dựa trên Mô Hình

- Giải thuật Expectation-Maximization (EM)
  - Là giải thuật lặp để tìm ***Maximum Likelihood (ML)*** – dùng được ngay cả trường hợp 1 số dữ liệu bị khuyết
  - **EM** gồm hai bước:
    - **Expectation step:** the (missing) data are estimated given the observed data and current estimates of model parameters
    - **Maximization step:** The likelihood function is maximized under the assumption that the (missing) data are known

5. Gom Cụm DL Dựa trên Mô Hình

E-Step

$$E[z_{ij}] = p(\mu = \mu_j \mid x = x_i)$$

$$= \frac{p(x = x_i \mid \mu = \mu_j)\, p(\mu = \mu_j)}{\sum_{n=1}^{k} p(x = x_i \mid \mu = \mu_n)\, p(\mu = \mu_j)}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}\, p(\mu = \mu_j)}{\sum_{n=1}^{k} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}\, p(\mu = \mu_n)}$$

M-Step

$$\mu_j \leftarrow \frac{1}{\sum_{i=1}^{m} E[z_{ij}]} \sum_{i=1}^{m} E[z_{ij}]\, x_i$$

$$p(\mu = \mu_j) \leftarrow \frac{1}{m} \sum_{i=1}^{m} E[z_{ij}]$$

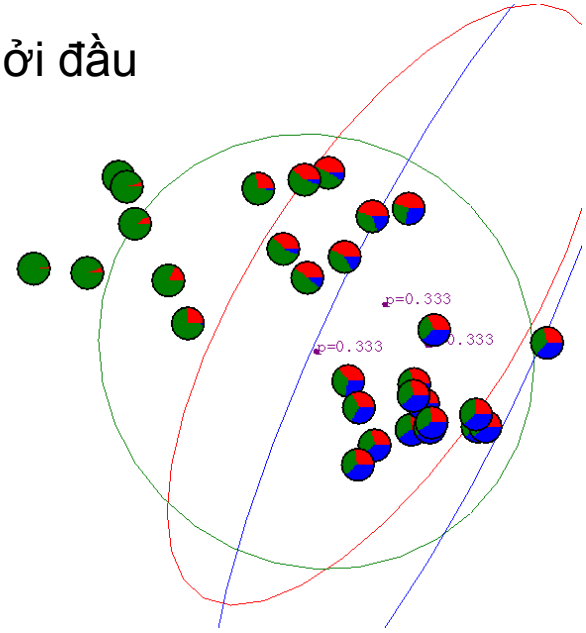
5. Gom Cụm DL Dựa trên Mô Hình

- Tóm tắt giải thuật (EM)
  - Input: tập **D** gồm **n** đối tượng, **K** cụm
  - Output: trị tối ưu cho các thông số của mô hình Θ (μ_i, Σ_i)
  - Giải thuật:

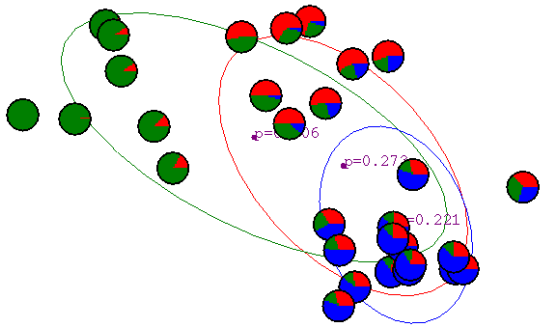
1. Khởi trị
   1.1. Chọn ngẫu nhiên **K** đối tượng làm trung tâm cụm
   1.2. Ước lượng trị ban đầu cho các thông số (nếu cần)
2. Lặp lại quá trình tinh chỉnh các thông số (cụm):
   2.1. Bước kỳ vọng (E-step): gán mỗi đối tượng x_i vào cụm C_k với xác suất $P(x_i \in C_k)$ với k=1..**K**
   2.2. Bước cực đại hóa (M-step): ước lượng trị các thông số
   2.3. Dừng khi thỏa điều kiện định trước (e.g. ML)

5. Gom Cụm DL Dựa trên Mô Hình

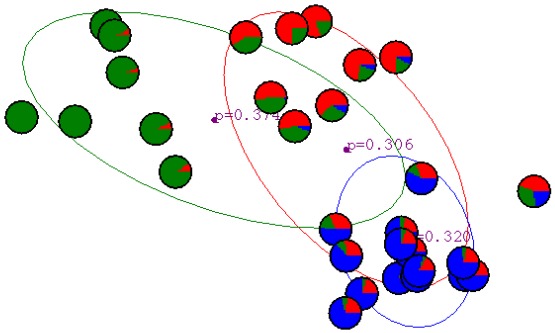
Khởi đầu



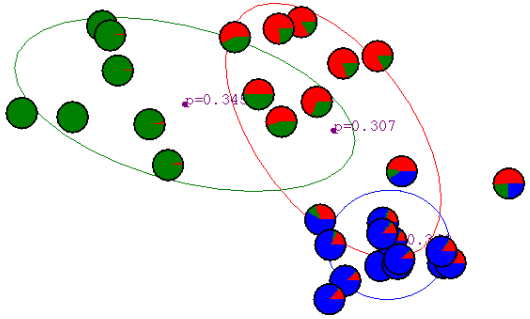
Sau bước lặp thứ 1



Sau bước lặp thứ 2

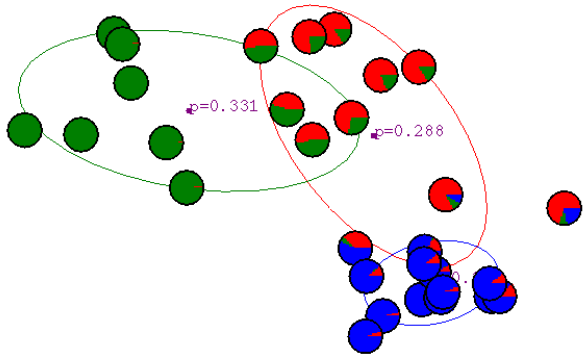


Sau bước lặp thứ 3

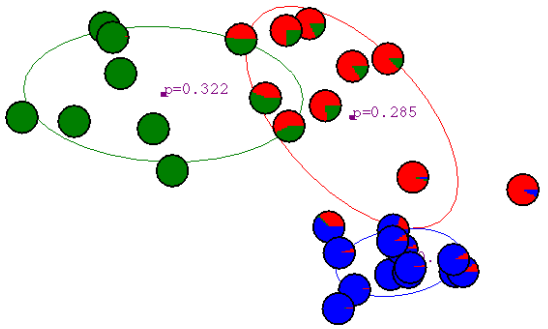


5. Gom Cụm DL Dựa trên Mô Hình

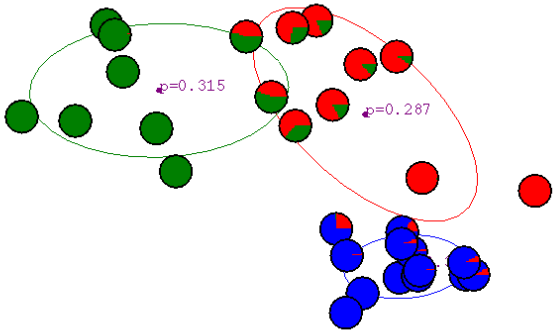
Sau bước lặp thứ 4



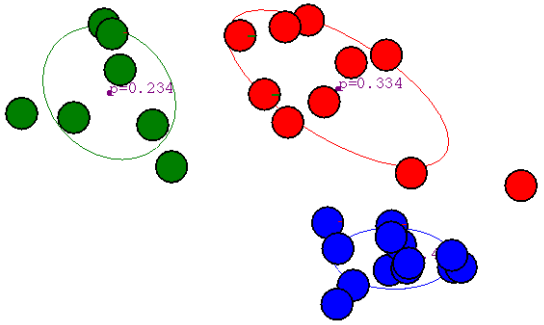
Sau bước lặp thứ 5



Sau bước lặp thứ 6

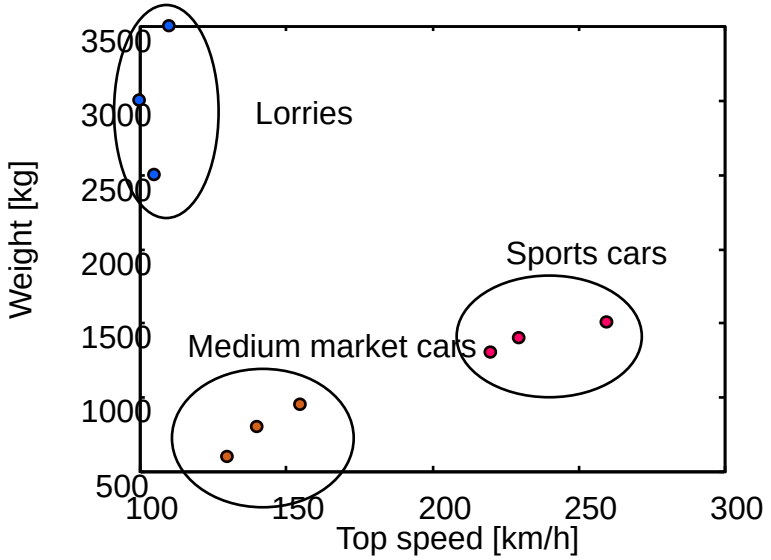


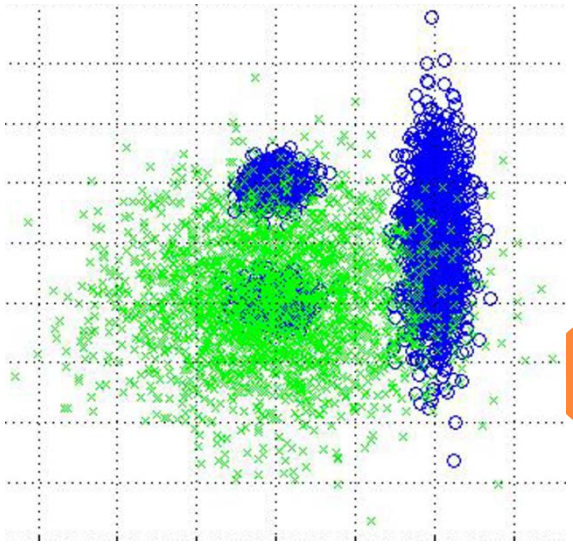
Sau bước lặp thứ 20



6. Các PP Gom Cụm DL Khác

- Gom cụm cứng (hard clustering)
  - Mỗi đối tượng chỉ thuộc về một cụm
  - Mức thành viên (degree of membership) của mỗi đối tượng với một cụm hoặc là 0 hoặc là 1
  - Ranh giới (boundary) giữa các cụm rõ ràng
- Gom cụm mờ (fuzzy clustering)
  - Mỗi đối tượng thuộc về nhiều hơn một cụm với mức thành viên nào đó từ 0 đến 1
  - Ranh giới giữa các cụm không rõ ràng (mờ - vague/fuzzy)





7. Tóm Tắt

- Gom cụm: nhóm các đối tượng vào các cụm dựa trên sự tương tự giữa chúng
- Độ đo đo sự tương tự tùy thuộc vào kiểu dữ liệu/đối tượng cụ thể
- Các giải thuật gom cụm được phân loại thành: phân hoạch, phân cấp, dựa trên mật độ, dựa trên mô hình, …

7. TÓM TẮT

Cluster algorithm	Complexity	Capability of tackling high dimensional data
K-means	$O(NKd)$ (time) $O(N+K)$ (space)	No
Fuzzy c-means	Near $O(N)$	No
Hierarchical clustering*	$O(N^2)$ (time) $O(N^2)$ (space)	No
CLARA	$O(K(40+K)^2+K(N-K))^+$ (time)	No
CLARANS	Quadratic in total performance	No
BIRCH	$O(N)$ (time)	No
DBSCAN	$O(N \log N)$ (time)	No
CURE	$O(N_{sample}^2 \log N_{sample})$ (time) $O(N_{sample})$ (space)	Yes
WaveCluster	$O(N)$ (time)	No
DENCLUE	$O(N \log N)$ (time)	Yes
FC	$O(N)$ (time)	Yes
CLIQUE	Linear with the number of objects, Quadratic with the number of dimensions	Yes
OptiGrid	Between $O(Nd)$ and $O(Nd \log N)$	Yes
ORCLUS	$O(K_0^3+K_0Nd+K_0^2d^3)$ (time) $O(K_0d^2)$ (space)	Yes

R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), May 20 pp. 645-678.

Q&A

quangtran@hcmut.edu.vn