Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

TÀI LIỆU LÝ THUYẾT KTDL & UD

Gom nhóm dữ liệu (P2)
Cluster Analysis

Giảng viên: ThS. Lê Ngọc Thành
Email: lnthanh@fit.hcmus.edu.vn

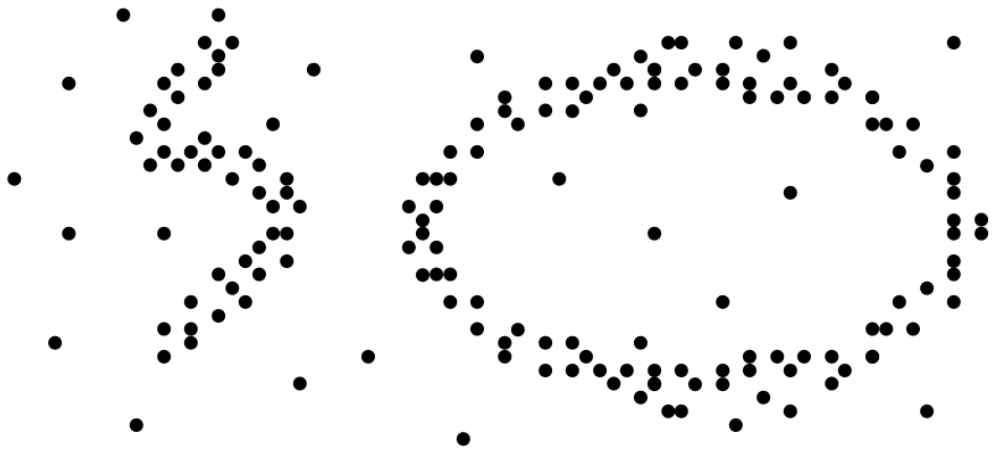
Summer 2012

Nội dung

- **Phương pháp dựa trên mật độ**
  - Định nghĩa gom nhóm dựa trên mật độ
  - Một số khái niệm cơ sở
  - Thuật toán DBSCAN
  - Sự phụ thuộc DBSCAN vào tham số
  - Nhận xét DBSCAN
- Phương pháp dựa trên lưới
- Đánh giá gom nhóm

Về pp phân hoạch và phân cấp

- Đa số các phương pháp phân hoạch và phân cấp được thiết kế để tìm ra các nhóm có dạng hình cầu.
- Rất khó trong việc tìm ra các nhóm hình dạng tùy ý như chữ “S” hay hình bầu dục.
- Nếu dữ liệu nhiễu hay cá biệt, hầu hết các thuật toán đều xác định không chính xác miền bao

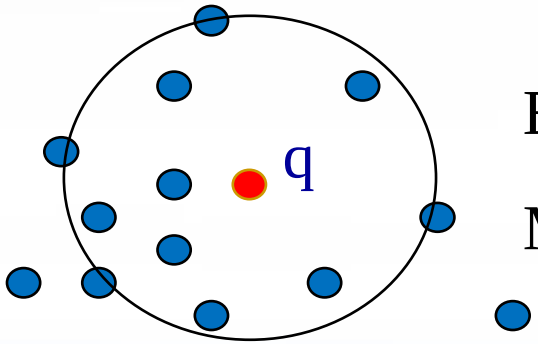


PP dựa trên mật độ

- *Mở rộng các nhóm cho đến khi mật độ của đối tượng dữ liệu trong vùng lân cận vượt qua ngưỡng.*
- Đặc điểm chính:
  - Khám phá nhóm có hình dạng bất kì
  - Kiểm soát nhiễu
  - Quét một lần
  - Cần xác định các tham số như là điều kiện dừng
- Một số thuật toán:
  - DBSCAN: Ester và đồng nghiệp (KDD'96)
  - OPTICS: Ankert và đồng nghiệp (SIGMOD'99)
  - DENCLUE: Hinneburg và D.Keim (KDD'98)

Khái niệm cơ sở (1/5)

- ***Eps***: bán kính cực đại của vùng lân cận
- ***MinPts***: số đối tượng/điểm ít nhất trong lân cận Eps của một đối tượng
- $N_{Eps}(q)$: tập hợp các đối tượng/điểm nằm trong lân cận Eps của q
  - {p thuộc D| dist(p,q) ≤ Eps}

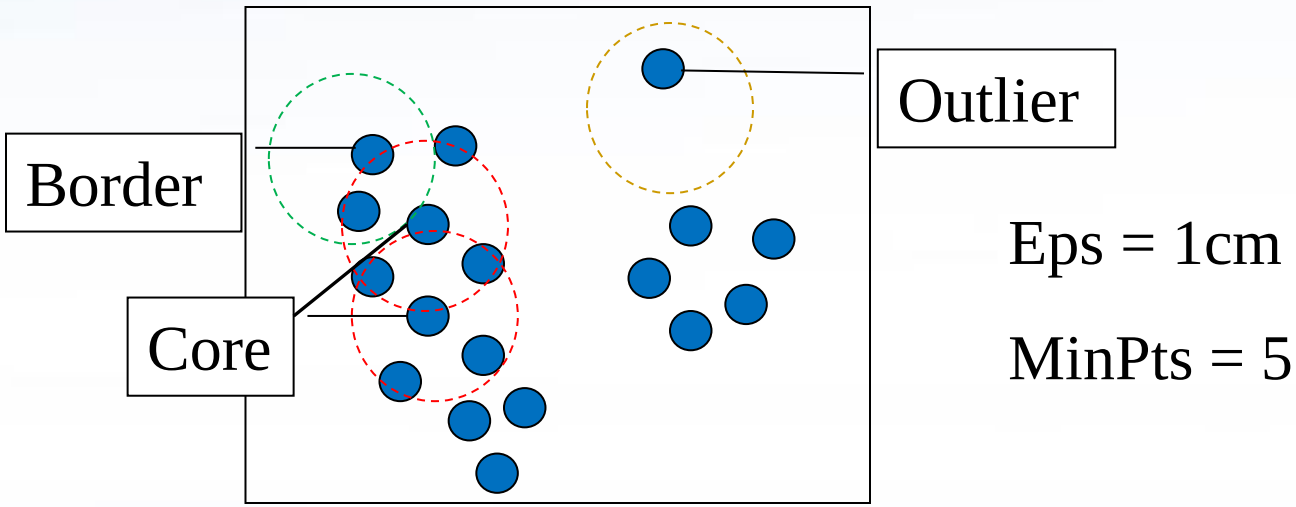


Eps = 1 cm

MinPts = 5

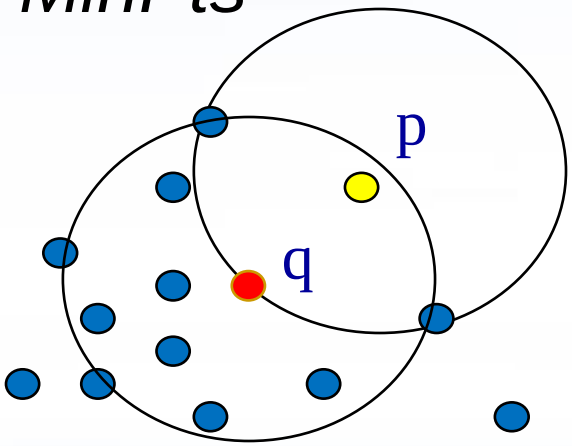
Khái niệm cơ sở (2/5)

- ***Đối tượng lõi*** (core object) là đối tượng thỏa Eps và MinPts
- ***Đối tượng biên*** (border object) là đối tượng có số điểm lân cận ít hơn MinPts trong Eps nhưng là lân cận của đối tượng lõi
- ***Đối tượng nhiễu*** (noise object) là bất kì điểm nào không phải là lõi hay biên



Khái niệm cơ sở (3/5)

- ***Đạt được mật độ trực tiếp*** (directly density-reachable): một điểm **p** gọi là đạt được mật độ trực tiếp từ **q** nếu:
  - **p** nằm trong lân cận Eps của **q**
  - $N_{Eps}(\mathbf{q})$ phải thỏa MinPts hay $|N_{Eps}(\boldsymbol{q})| \geq MinPts$

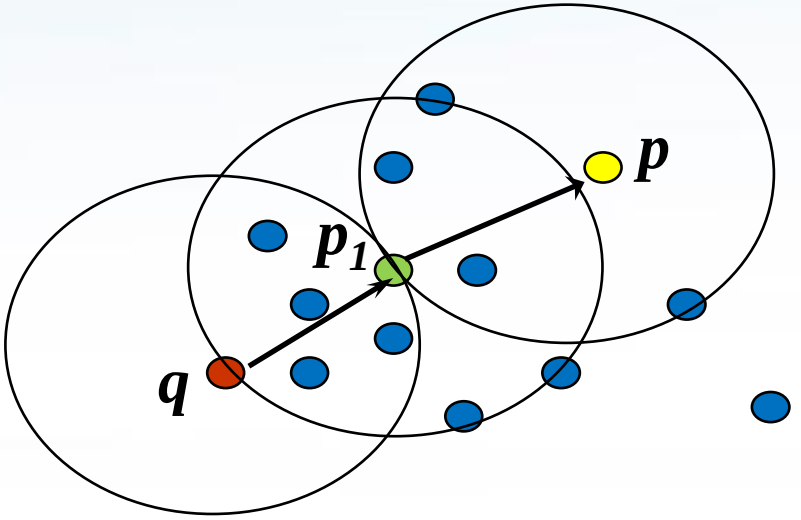


Eps = 1 cm

MinPts = 5

Khái niệm cơ sở (4/5)

- ***Đạt được mật độ*** (density-reachable): Một điểm **p** gọi là đạt được mật độ từ điểm **q** (thỏa Eps, MinPts) nếu tồn tại một chuỗi các điểm $\mathbf{p_1}$, $\mathbf{p_2}$,…,$\mathbf{p_n}$ với $\mathbf{p_1}$ là **q** và $\mathbf{p_n}$ là **p** để mà $\mathbf{p_{i+1}}$ là *đạt được mật độ trực tiếp* từ $\mathbf{p_i}$

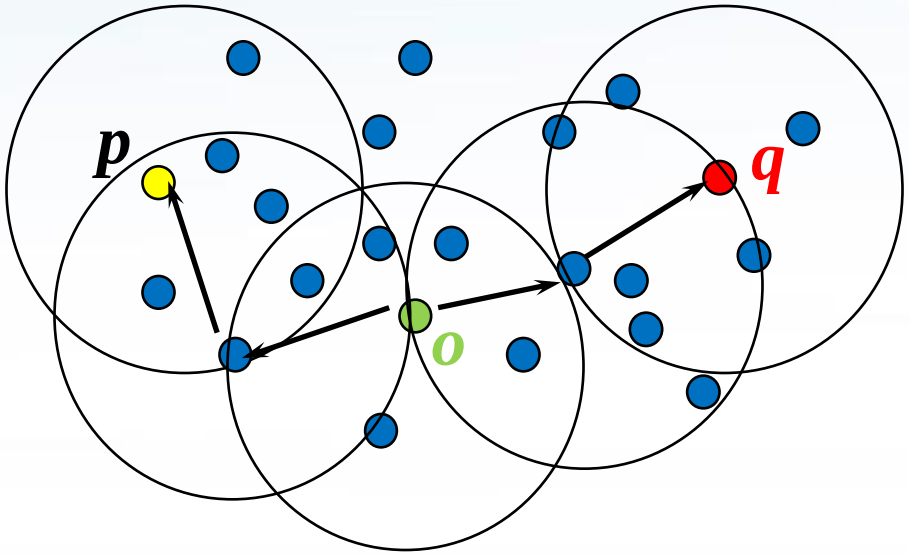


Eps = 1 cm

MinPts = 5

Khái niệm cơ sở (5/5)

- ***Liên thông mật độ*** (density-connected): một điểm **p** gọi là liên thông mật độ đến điểm **q** (thỏa Eps, MinPts) nếu tồn tại một điểm **o** (cũng thỏa Eps, MinPts) mà cả hai điểm **p** và **q** đều là *đạt được mật độ* từ **o**

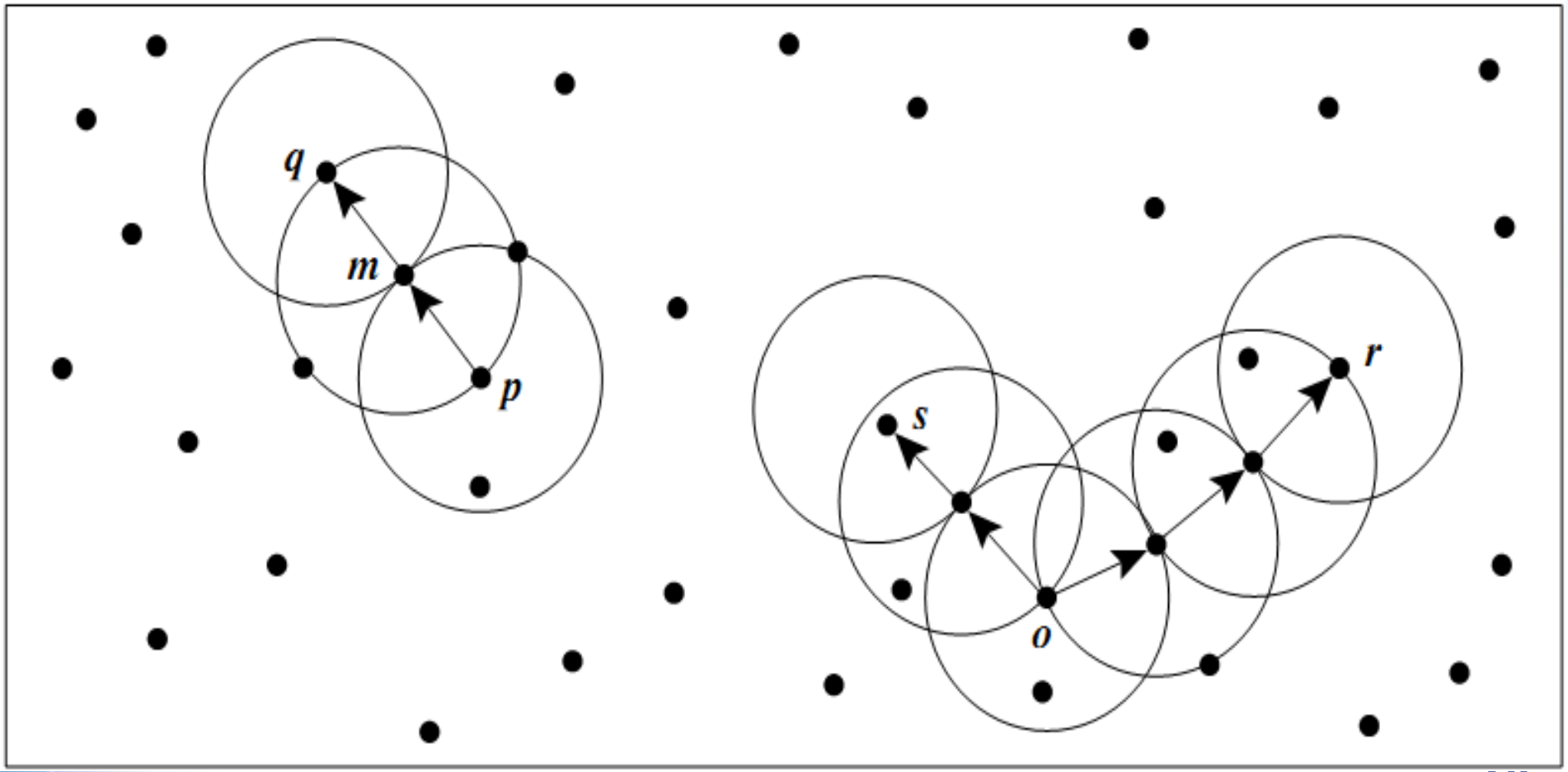


Eps = 1 cm

MinPts = 5

Bài tập 1

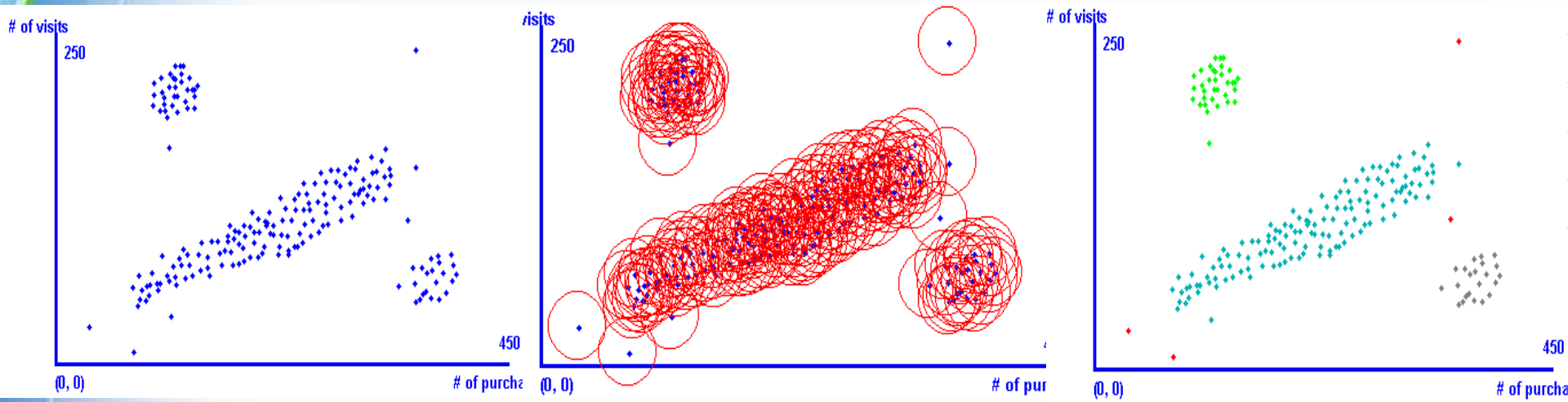
- Xác định mối quan hệ giữa các điểm trong hình sau:



Đáp án trong [1] trang 472

DBSCAN

- Một ***nhóm dựa trên mật độ*** (density-based cluster) là một nhóm có số lượng điểm *liên thông mật độ* tối đại.
- Hay, một tập con $\mathbf{C} \subseteq \mathbf{D}$ là một nhóm nếu:
  - Với bất kì 2 điểm $\mathbf{o_1}$,$\mathbf{o_2}$ thuộc **C**, $\mathbf{o_1}$ và $\mathbf{o_2}$ là các điểm liên thông mật độ
  - Không tồn tại một điểm **o** thuộc **C** và một điểm **o'** khác thuộc (**D-C**) mà **o** và **o'** liên thông mật độ với nhau.
- Dựa trên đó, DBSCAN có thể tìm ra các nhóm có hình dạng bất kì trong không gian dữ liệu nhiễu



Thuật toán DBSCAN (1/2)

B1. Khởi tạo các điểm với nhãn chưa viếng thăm ("***unvisited***").

B2. Chọn bất một điểm **p** chưa viếng thăm và đánh nhãn "***visited***".

B3. Nếu **p** không thỏa *Eps* và *MinPts*, **p** được đánh nhãn là ***điểm nhiễu***

B4. Nếu **p** thỏa *Eps* và *MinPts*

B5. Một nhóm **C** mới được tạo từ **p**.

B6. Các điểm lân cận *Eps* của **p** được đưa vào *tập ứng viên* **N**.

B7. Xét từng ứng viên **p'** trong **N**

B8. Nếu **p'** chưa là thành viên của nhóm nào, thêm **p'** vào **C**.

B9. Nếu **p'** mang nhãn "***unvisited***" sẽ được đánh nhãn "***visited***".

B10. Nếu **p'** thỏa *Eps* và *MinPts* thì các điểm trong lân cận *Eps* của **p'** được thêm vào tập ứng viên **N**.

B11. Lặp lại ***B7***.

B12. Lặp lại ***B2***.

Thuật toán DBSCAN (2/2)

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                     add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

Bài tập 2

Xét tập dữ liệu sau: P1={1,3}, P2={2,3}, P3={4,1}, P4={4,4}, P5={5,2}, P6={5,5}, P7={A,5,6}, P8={6,1}, P9={5,1}, P10={6,3}, P11={6,2}, P12={5,3}, P13={5,2}, P14={4,2}, P15{4,5}

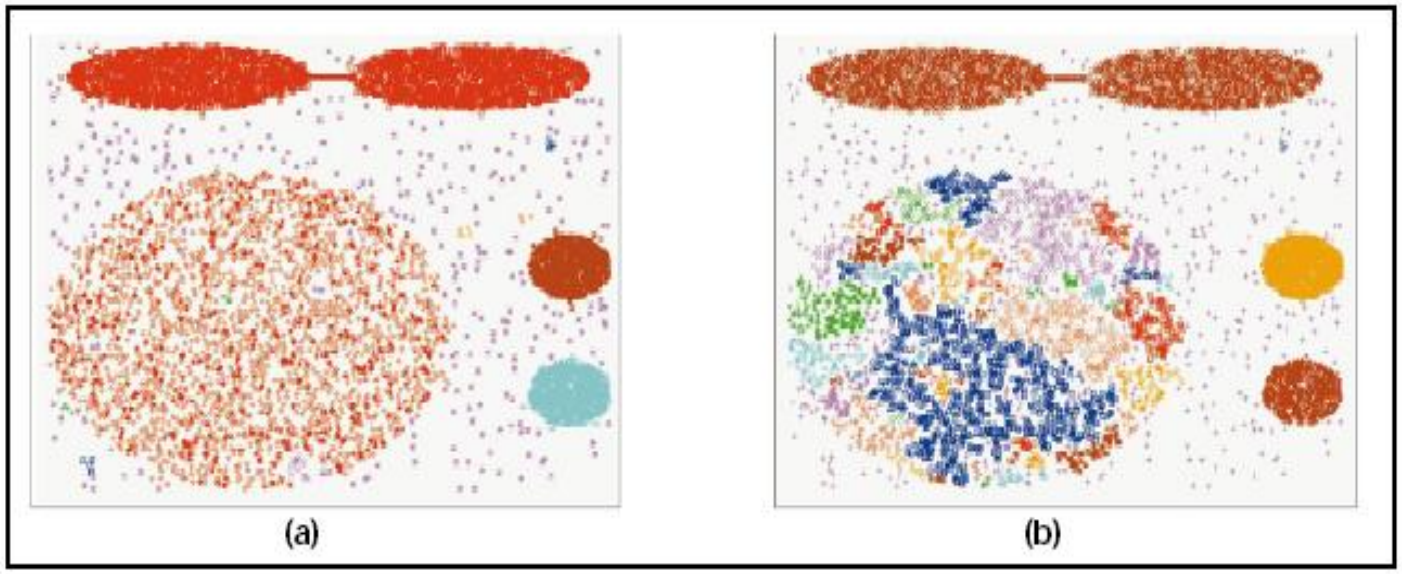
Gom nhóm dữ liệu sử dụng DBSCAN. Xác định đối tượng lõi, biên của từng nhóm, liệt kết tất cả đối tượng nhiễu

Sử dụng khoảng cách Euclide với Eps = 1.5 theo thông số MinPts:

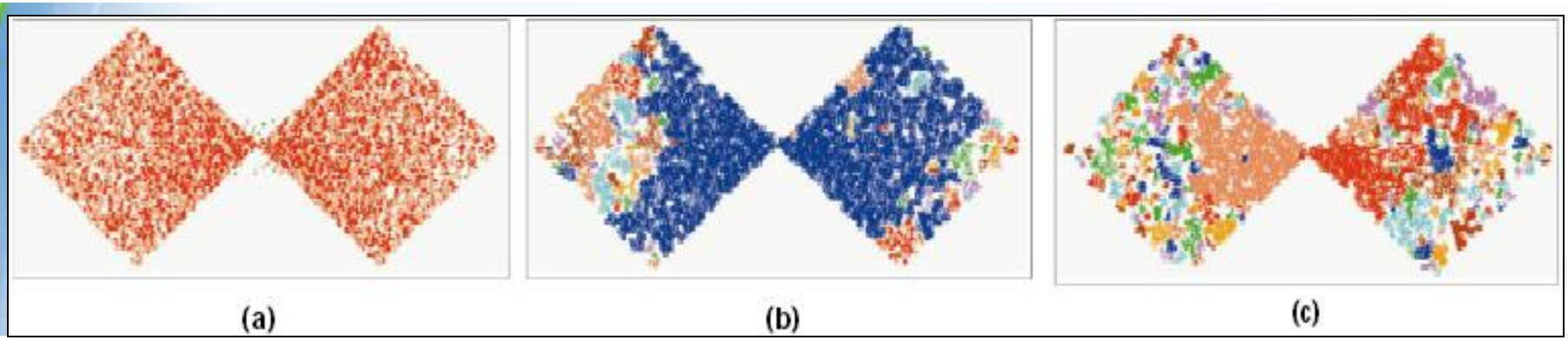
a) MinPts = 3

b) MinPts = 4

DBSCAN phụ thuộc tham số



Kết quả DBSCAN cho tập dữ liệu DS1 với MinPts = 4 và Eps là (a) 0.5 và (b) 0.4



DBSCAN cho tập dữ liệu DS2 với MinPts = 4 và Eps là (a) 5.0, (b) 3.5 và (c) 3.0

Online Demo http://webdocs.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

Nhận xét DBSCAN

- ***Ưu điểm:***
  - Làm việc tốt với dữ liệu nhiễu
  - Có thể giải quyết các trường hợp các nhóm có hình dáng và kính thước khác nhau
- ***Nhược điểm:***
  - Gặp vấn đề khi các nhóm có mật độ khác nhau
  - Độ phức tạp cao đối dữ liệu nhiều chiều
  - Phụ thuộc vào giá trị Eps, MinPts
- Một số thuật toán cải tiến:
  - OPTICS (Ordering Points to Identify the Clustering Structure)
  - DENCLUE(Clustering Based on Density Distribution Functions)

(Đọc thêm trong [1] phần 10.4.2 và 10.4.3)

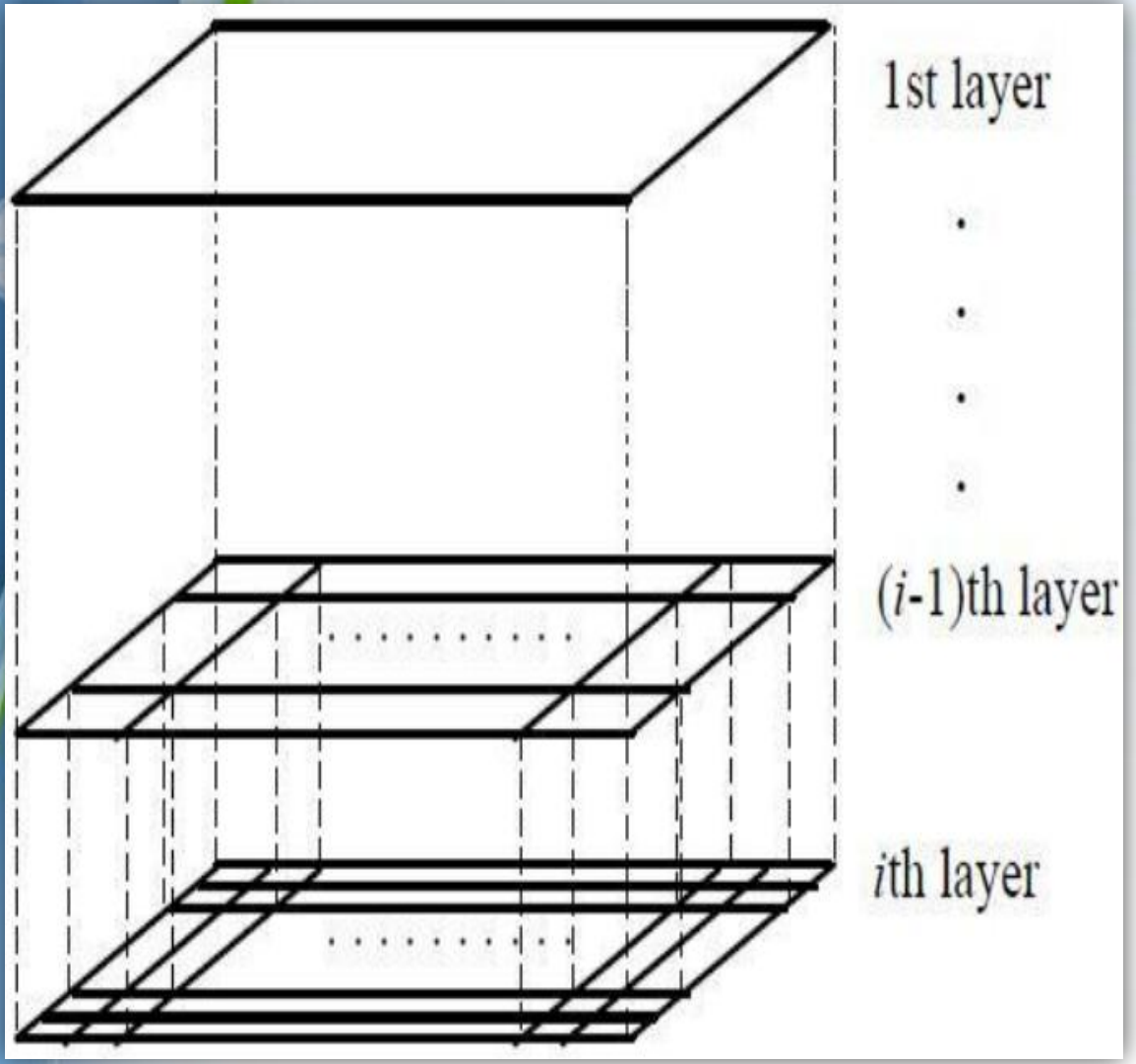
Nội dung

- Phương pháp dựa trên mật độ
- **Phương pháp dựa trên lưới**
  - Gom nhóm dựa trên lưới
  - STING
  - Tham số thống kê
  - Loại truy vấn
  - Thuật toán STING
  - Nhận xét STING
- Đánh giá gom nhóm

Gom nhóm dựa trên lưới

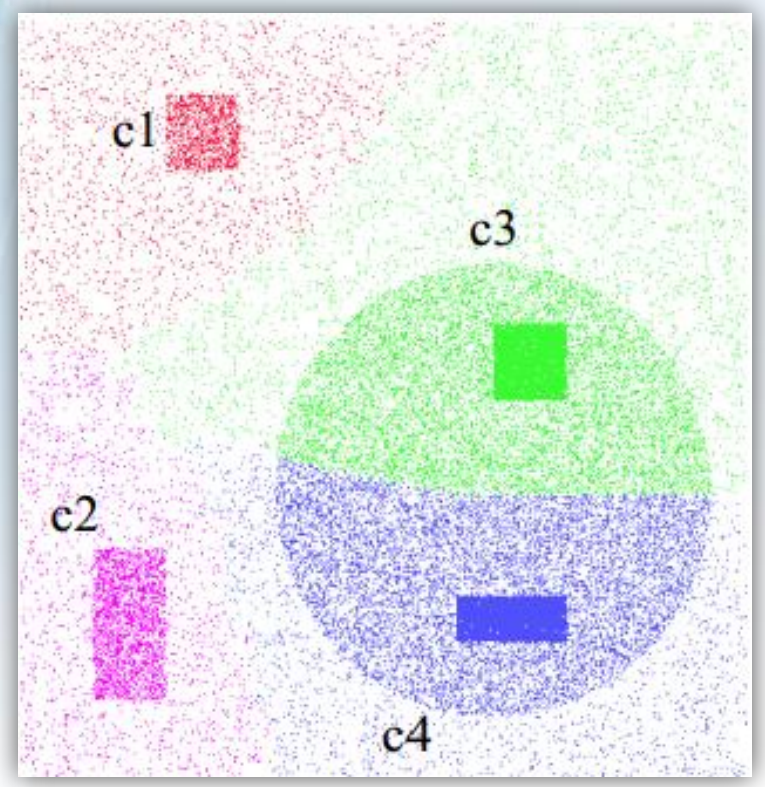
- Các phương pháp gom nhóm đã đề cập đều hướng dữ liệu (data-driven)
  - Chia tập đối tượng và thích ứng với phân bố đối tượng trong không gian.
- ***Gom nhóm dựa trên lưới*** (grid-based clustering) áp dụng phương pháp hướng không gian (space-driven)
  - Chia không gian thành các ô (cell), độc lập với phân bố đối tượng đầu vào.

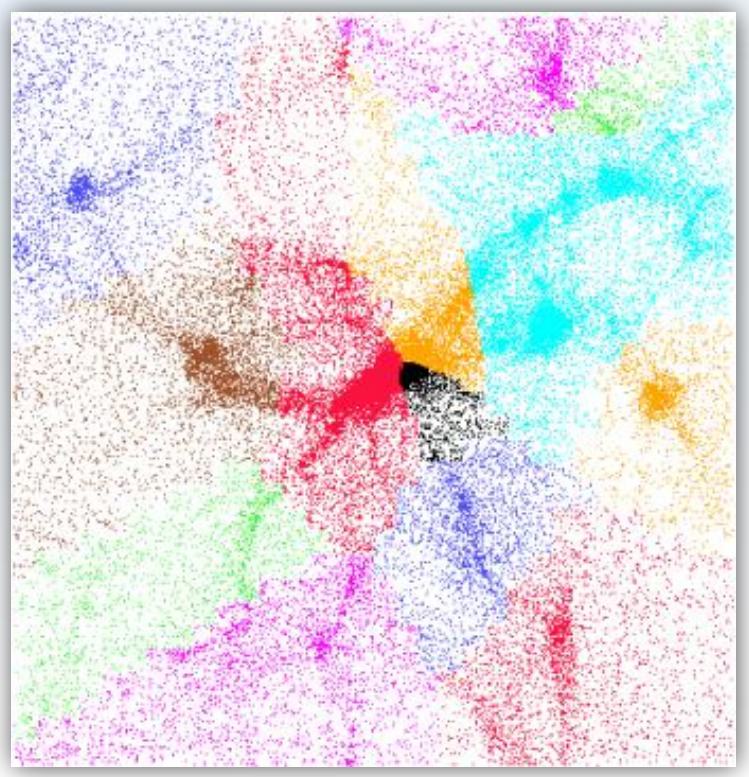
Gom nhóm dựa trên lưới (tt)



- Sử dụng cấu trúc dữ liệu *lưới đa phân giải* (multi-resolution grid data structure).
  - Chia không gian đối tượng thành một tập hữu hạn các ô (cell) hình thành nên cấu trúc lưới.
  - Thời gian xử lý nhanh, độc lập với số lượng đối tượng, chỉ phụ thuộc vào số lượng ô trong không gian

Ví dụ gom nhóm dựa trên lưới

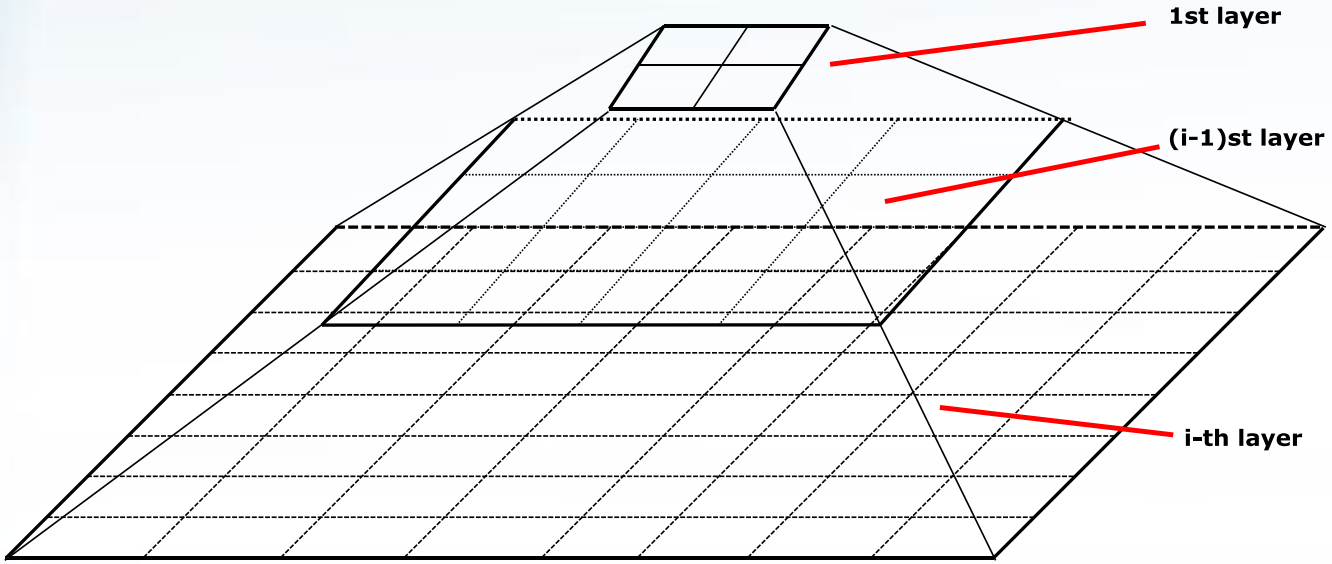




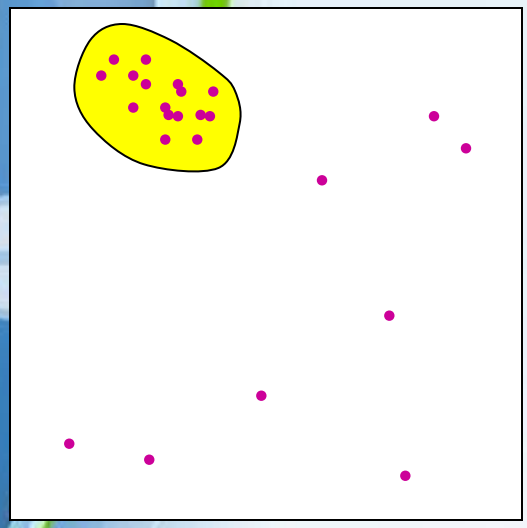
Trích từ: "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement" (http://users.eecs.northwestern.edu/~choudhar/Publications/LiaLiu04A.pdf)

STING

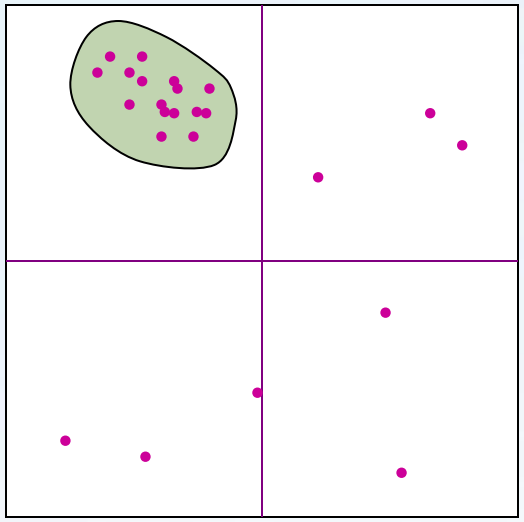
- STING (**ST**atistical **IN**formation **G**rid) đề xuất bởi Wang, Yang và Muntz (VLDB'97)
- Không gian được chia thành các ô hình chữ nhật.
- Mỗi ô có thể được chia thành nhiều mức tương ứng với các mức khác nhau của độ phân giải và hình thành cấu trúc phân tầng



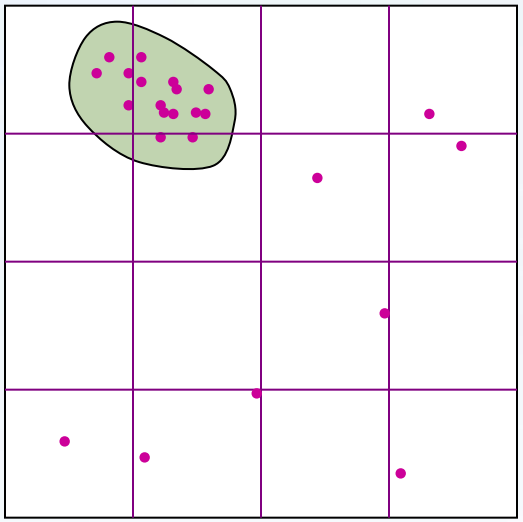
Ví dụ STING



Level 0

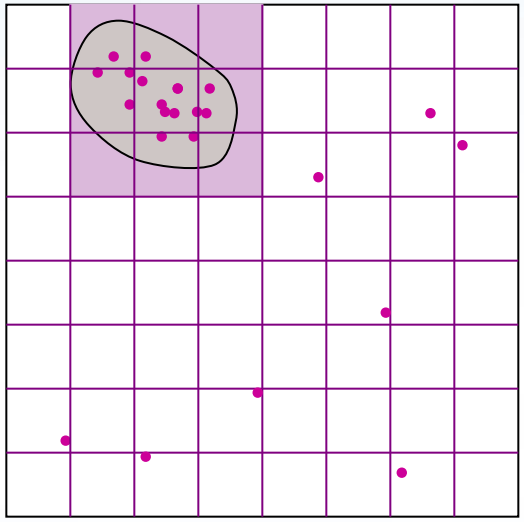


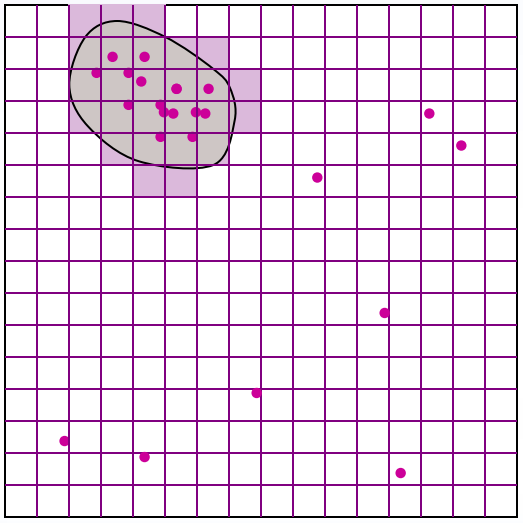
Level 1



Level 2

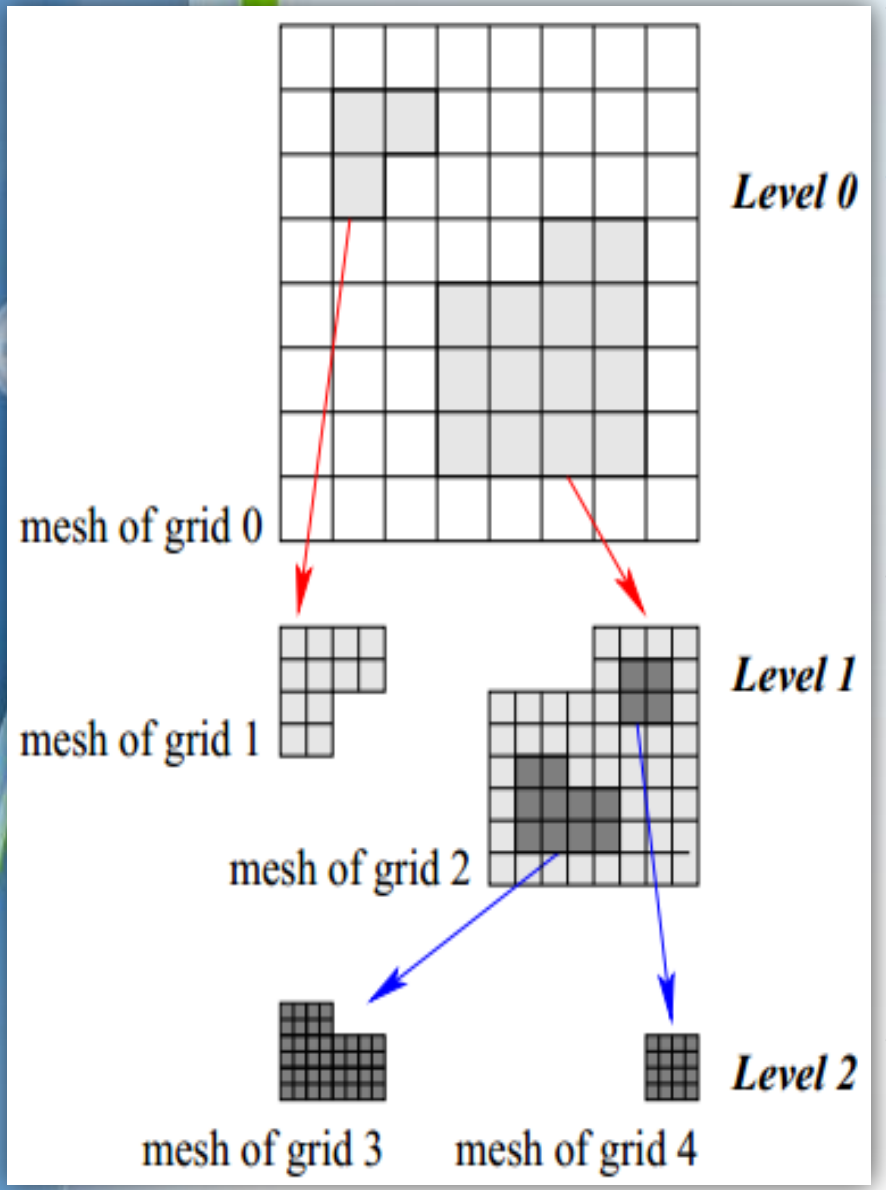
Level 3





Level 4

STING (tt)



- Mỗi ô ở mức cao được phân thành một số ô ở mức thấp hơn.
- *Thông tin thống kê* như giá trị trung bình, cực đại, cực tiểu trong mỗi ô được tính toán lại và lưu trữ sẵn phục vụ cho *nhiệm vụ truy vấn* và gom nhóm sau này.
- Các *tham số thống kê* của ô mức cao hơn được tính từ các tham số của ô mức thấp hơn:
  - *count, mean, stdev, min, max*
  - Loại phân phối: *phân phối chuẩn*, *phân phối đều*, *phân phối mũ,…*
- Các tham số thống kê ở *mức đáy* sẽ được tính trực tiếp từ dữ liệu

Tham số thống kê

- Tính tham số thống kê:

$$n = \sum_i n_i \qquad m = \frac{\sum_i m_i n_i}{n}$$

$$s = \sqrt{\frac{\sum_i (s_i{}^2 + m_i{}^2) n_i}{n} - m^2}$$

$$min = \min_i(min_i) \qquad max = \max_i(max_i)$$

Trong đó:

n_i: *là (count) số lượng đối tượng ở mức thấp hơn*

m_i:*là (mean) giá trị trung bình ở mức thấp hơn*

s_i: *là (stdev) độ lệch chuẩn ở mức thấp hơn*

min_i, max_i : *là giá trị nhỏ nhất/lớn nhất ở mức thấp hơn*

Phân phối dữ liệu

- Tham khảo thêm trong phần phụ lục 1 để biết cách tính phân phối ở mỗi lớp dựa trên phân phối lớp dưới

Ví dụ tham số thống kê

i	1	2	3	4
n_i	100	50	60	10
m_i	20.1	19.7	21.0	20.5
s_i	2.3	2.2	2.4	2.1
min_i	4.5	5.5	3.8	7
max_i	36	34	37	40
$dist_i$	NORMAL	NORMAL	NORMAL	NONE

- Các tham số ở lớp cha: *n,m,s,min,max, dist*?

  n = 220; m = 20.27; s=2.37; min=3.8; max = 40;

  dist = NORMAL

Loại truy vấn

- Cấu trúc STING cho phép trả lời cho rất nhiều loại truy vấn.
- Thậm chí nếu thông tin thống kê không thích hợp để trả lời truy vấn, chúng ta vẫn có thể phát sinh ra một tập các trả lời có thể có.
- Ví dụ: *"Chọn các miền cực đại mà có ít nhất 100 ngôi nhà/mỗi khu vực đơn vị và ít nhất 70% giá nhà trên 400K với tổng khu vực ít nhất 100 đơn vị và độ tinh cậy 90%"*

```
SELECT REGION
FROM house-map
WHERE DENSITY IN (100, ∞)
AND price RANGE (400000, ∞) WITH PERCENT (0.7, 1)
AND AREA (100, ∞)
AND WITH CONFIDENCE 0.9
```

Thuật toán STING

Các tham số thống kê được sử dụng theo kiểu top-down.

B1. Một lớp được chọn để bắt đầu tiến trình (chứa một lượng nhỏ các ô)

B2. Với mỗi ô lớp hiện tại

B3. Tính toán khoảng tin cậy thể hiện sự liên quan của ô với câu truy vấn

B4. Các ô không liên quan bị bỏ đi.

B5. Xử lý các ô liên quan ở mức thấp hơn

B6. Lặp lại tiến trình cho đi khi chạm lớp đáy

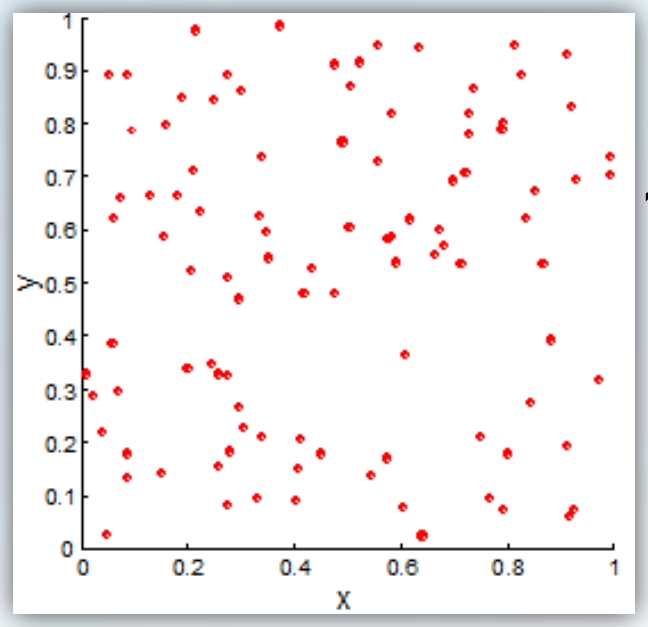
Nhận xét STING

- Ưu điểm:
  - Độc lập truy vấn, có thể thực hiện song song và cập nhật tăng cường.
  - Độ phức tạp, O(g) với g là số ô lưới ở lớp thấp nhất (thường < n)
- Khuyết điểm:
  - Chất lượng phụ thuộc vào cách chia ô lưới ở mức thấp nhất.
  - Đường biên của tất cả các nhóm hoặc theo chiều ngang hoặc chiều đứng, không có dạng chéo.
- Hướng tiếp cận khác:
  - CLIQUE (Agrawal – SIGMOD'98): đọc thêm trong [1] phần 10.5.2

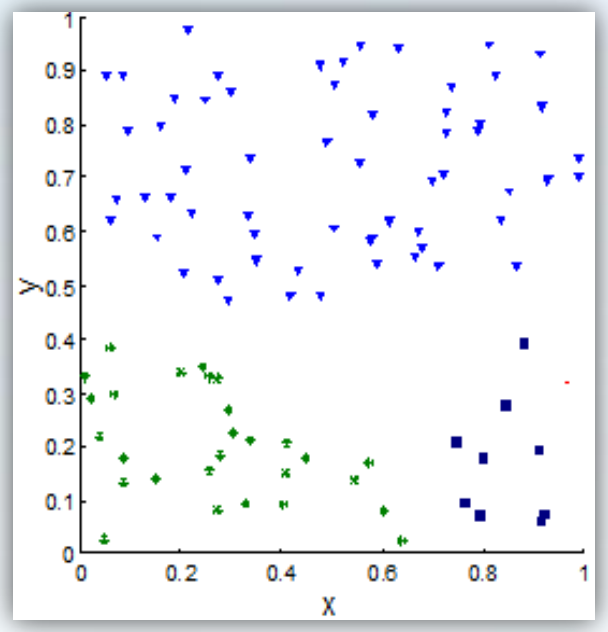
Nội dung

- Phương pháp dựa trên mật độ
- Phương pháp dựa trên lưới
- **Đánh giá gom nhóm**
  - Đánh giá xu hướng gom nhóm
  - Xác định số nhóm
  - Đánh giá chất lượng nhóm
    - Chỉ số trong
    - Chỉ số ngoài
    - Chỉ số tương quan

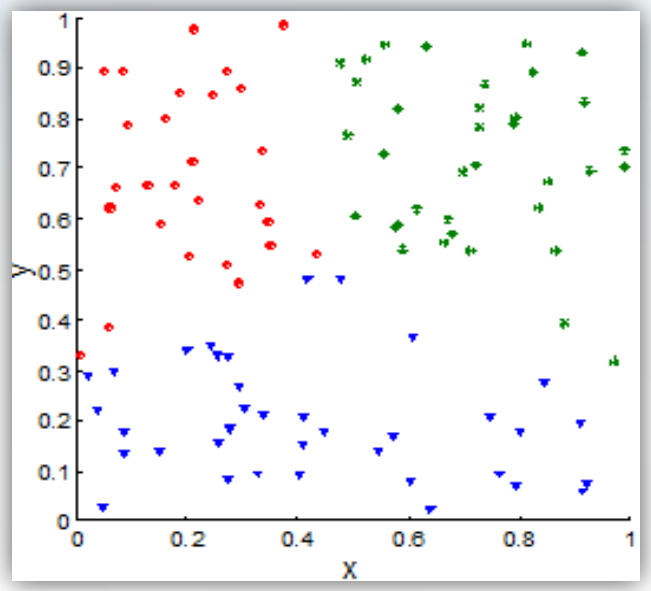
So sánh kết quả gom nhóm



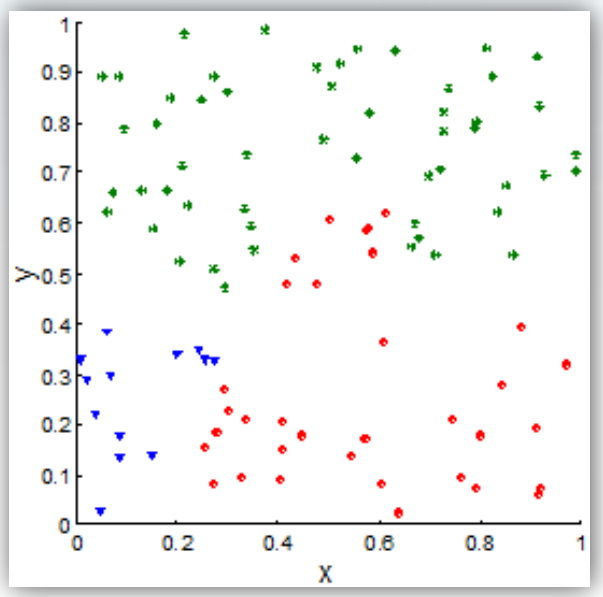
DL ban đầu



DBSCAN



K-means



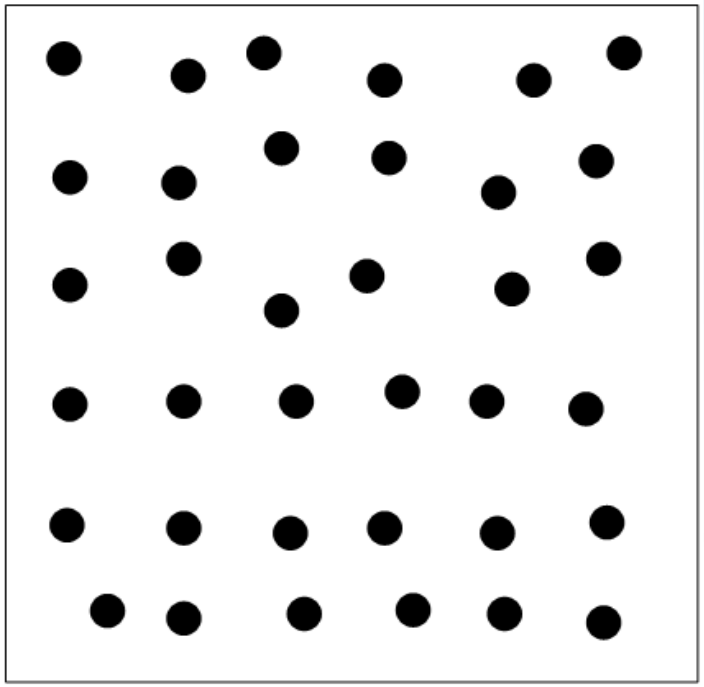
Complete Link

Đánh giá gom nhóm

- Đánh giá gom nhóm là quá trình đánh giá *tính khả thi* của phương pháp gom nhóm trên tập dữ liệu và *chất lượng* kết quả đạt được.
- Nhiệm vụ chính bao gồm:
  - Đánh giá xu hướng gom nhóm
  - Xác định số lượng nhóm trong tập dữ liệu
  - Đo lường chất lượng gom nhóm

Đánh giá xu hướng gom nhóm

- Với tập dữ liệu cho trước, đánh giá xu hướng gom nhóm là đánh giá tập dữ liệu có *cấu trúc không ngẫu nhiên* và có thể sinh ra các nhóm có nghĩa hay không?
  - Áp dụng tùy tiện lên tập dữ liệu ngẫu nhiên có thể sinh ra các nhóm nhưng chúng không nói lên điều gì.
- Việc gom nhóm đòi hỏi phân bố phải không đều (*nonuniform)*



Tập dữ liệu phân bố quá đều

Đánh giá xu hướng gom nhóm (tt)

- Kiểm tra tính ngẫu nhiên không gian bằng phương pháp: **Hopkins Statistic**
  - Cho một tập dữ liệu D, được xem như mẫu của biến ngẫu nhiên ***o***, cần xác định ***o*** cách phân bố đều bao xa trong không gian dữ liệu.
  - Lấy mẫu n điểm, $p_1,\ldots,p_n$ đồng đều từ D. Tìm láng giếng gần p_i nhất trong D: $x_i = min\{dist\ (p_i, v)\}$ với v thuộc D
  - Lấy mẫu n điểm, $q_1,\ldots,q_n$ đồng đều từ D. Tìm láng giếng gần q_i nhất trong D-{q_i}: $y_i = min\{dist\ (q_i, v)\}$ với c thuộc D và $v \neq q_i$.
  - Tính Hopkins Statistic: $H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$
  - Nếu D là phân bố đều, $\sum x_i$ và $\sum y_i$ sẽ gần bằng nhau và H gần bằng 0.5. Nếu D được gom nhóm, H sẽ gần bằng 1

Xác định số nhóm

- Số lượng nhóm "đúng" thường phụ thuộc vào hình dạng phân bố, độ lớn tập dữ liệu cũng như độ phân giải nhóm mà người dùng yêu cầu.
- *Phương pháp thực nghiệm:*
  - Số nhóm ≈ $\sqrt{n/2}$ với n là số điểm trong dữ liệu
- *Phương pháp Elbow:*
  - Sử dụng điểm chuyển trong đường cong biểu diễn tổng biến đổi trong nhóm như là số nhóm
- *Phương pháp cross validation:*
  - Chia tập dữ liệu thành m phần.
  - Sử dụng m-1 phần để tìm mô hình gom nhóm.
  - Sử dụng phần còn lại để kiểm thử chất lượng của gom nhóm.
  - Với bất kì k>0, lặp lại quá trình m lần, so sánh chất lượng với các k khác nhau từ đó rút ra được số nhóm phù hợp.

Đo lường chất lượng nhóm

- Có 3 loại chỉ số:
  - ***Chỉ số ngoài*** (external index): đo chất lượng nhóm dựa trên nhãn các nhóm đã được đánh sẵn (supervised)
    - Entropy, Purity, độ chính xác và độ phủ BCubed
  - ***Chỉ số trong*** (internal index): đo chất lượng nhóm mà không cần thông tin từ bên ngoài (unsupervised)
    - Sum of squared error, Silhouete coefficient
  - ***Chỉ số liên quan*** (relative index): sử dụng để so sánh các nhóm hay các phương pháp gom nhóm khác nhau
    - Thường chỉ số trong/ngoài có thể được sử dụng.
- Ngoài ra có thể đánh giá chất lượng nhóm thông qua *độ tương quan* (correlation)

Chỉ số ngoài - Entropy

- ***Entropy của một nhóm j:***

$$e_j = -\sum_{i=1}^{m} p_i log_2(p_i)$$

m: số lượng lớp

p_i: xác suất mẫu trong nhóm j thuộc về lớp i ($\frac{\#_{ij}}{\#_j}$)

- ***Entropy tổng cộng***: tập hợp tỉ lệ của nhóm nhân với entropy của nhóm

$$e = \sum_{j=1}^{k} \frac{\#_j}{\#_D} \times e_j$$

k: số nhóm

$\#_j$: số mẫu trong nhóm j

$\#_D$: số mẫu trong dữ liệu D

Chỉ số ngoài – Purity

- ***Purity (độ thuần) của một nhóm j:***

$$purity_j = \max(p_i)$$

p_i: xác suất mẫu trong nhóm j thuộc về lớp i ($\frac{\#_{ij}}{\#_j}$)

- ***Độ purity tổng cộng:***

$$purity = \sum_{j=1}^{k} \frac{\#_j}{\#_D} \times purity_j$$

k: số nhóm

$\#_j$: số mẫu trong nhóm j

$\#_D$: số mẫu trong dữ liệu D

Bài tập 3

Kết quả của gom nhóm k-means cho tập dữ liệu văn bản LA

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	?	?
4	10	162	3	119	73	2	?	?
5	331	22	5	70	13	23	?	?
6	5	358	12	212	48	13	?	?
Total	354	555	341	943	273	738	?	?

Tính độ Entropy và Purity cho các nhóm còn lại và chi phí tổng cổng?

Bài tập 3 – Đáp án

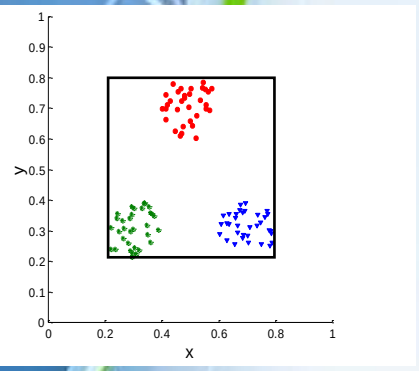
Cluster/ Class	Entert	Financial	Foreign	Metro	National	Sports	Tổng	Entropy	Purity
1	3	5	40	506	96	27	**677**	**1.2270**	**0.7474**
2	4	7	280	29	39	2	**361**	**1.1472**	**0.7756**
3	1	1	1	7	4	671	**685**	**0.1813**	**0.9796**
4	10	162	3	119	73	2	**369**	**1.7487**	**0.4390**
5	331	22	5	70	13	23	**464**	**1.3976**	**0.7134**
6	5	358	12	212	48	13	**648**	**1.5523**	**0.5525**
Tổng	**354**	**555**	**341**	**943**	**273**	**738**	**3204**	**1.1450**	**0.7203**

Kết quả của gom nhóm k-means cho tập dữ liệu văn bản LA

Chỉ số ngoài - BCubed

- Đọc thêm trong [1] trang 489 để biết cách tính độ chính xác và độ phủ BCubed

Chỉ số trong - SSE



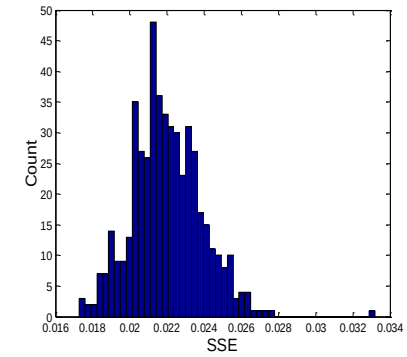
- ***Sự gắn kết nhóm*** (cluster cohesion): đo độ liên quan giữa các đối tượng trong nhóm sử dụng tổng bình phương sai bên trong nhóm (within)

$$\boldsymbol{WSS} = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

- ***Sự phân chia nhóm*** (cluster separation): đo độ phân biệt giữa nhóm này với nhóm khác sử dụng tổng bình phương sai giữa các nhóm (between)

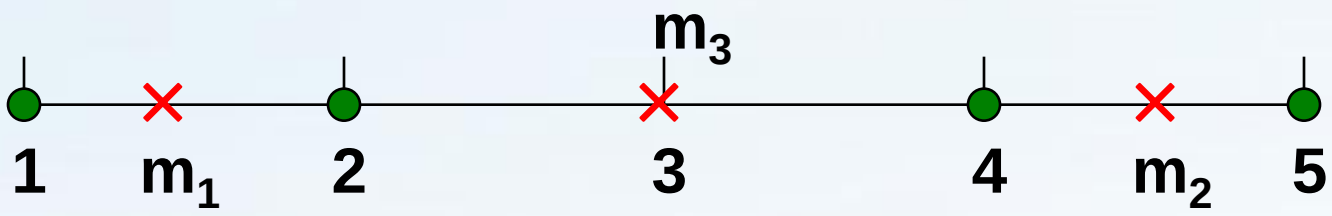
$$\boldsymbol{BSS} = \sum_i |C_i| (m - m_i)^2$$

Trong đó, $|C_i|$ là kích thước nhóm i



Bài tập 4

- BSS + WSS = hằng số



* *k= 1 nhóm* (1,2,4,5):

WSS = $(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$

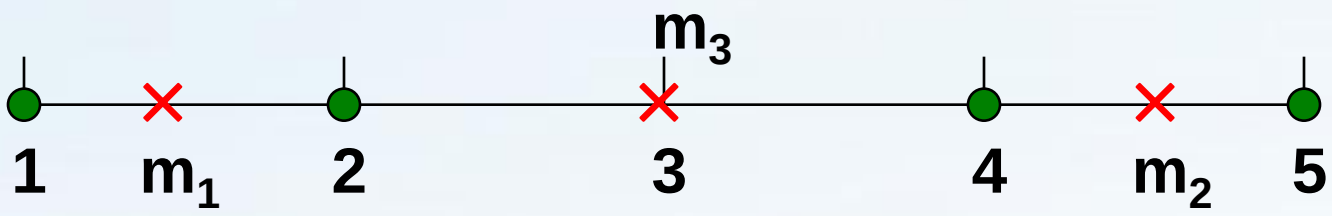
BSS = $4 \times (3-3)^2 = 0$

Total = 10 + 0 = 10

* *k = 2 nhóm* {(1,2); (4,5)}: tính WSS, BSS?

Bài tập 4 – Đáp án

- BSS + WSS = hằng số



* *k = 2 nhóm* {(1,2); (4,5)}:

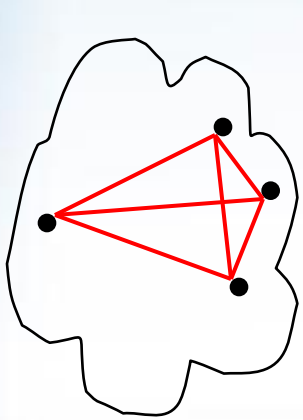
WSS = $(1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$

BSS = $2 \times (3-1.5)^2 + 2 \times (4.5 - 3)^2 = 9$

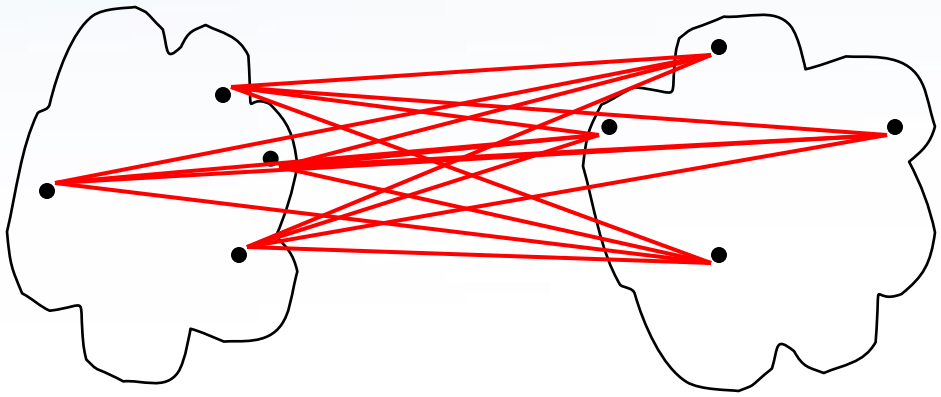
Total = 1 + 9 = 10

Chỉ số trong – Đồ thị xấp xĩ

- Phương pháp dựa trên ***đồ thị kề*** (proximity graph) cũng được sử dụng để đánh giá độ gắn kết và độ phân chia nhóm:
  - *Độ gắn kết* là tổng trọng số của tất cả các liên kết bên trong nhóm
  - *Độ phân chia* là tổng trọng số giữa các node trong nhóm và node ngoài nhóm.



Độ gắn kết (cohesion)



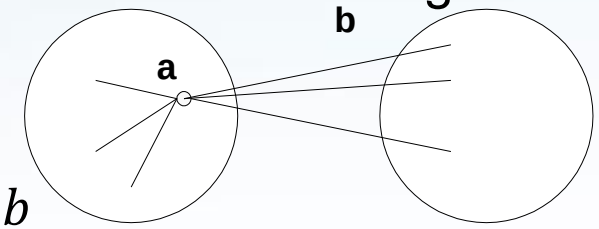
Độ phân chia (separation)

Chỉ số trong – Hệ số Silhouette

- ***Hệ số Silhouette*** dựa trên ý tưởng của độ gắn kết và độ phân chia nhưng chỉ cho bản thân từng điểm
- Xét điểm i:

  a = khoảng cách trung bình của i đến các điểm trong nhóm của nó

  b = min (khoảng cách trung bình của i đến các điểm trong nhóm khác)

  Hệ số Silhouette cho điểm này:

$$s = \begin{cases} 1 - \frac{a}{b} & nếu\ a < b \\ \frac{b}{a} - 1 & nếu\ a \geq b\ (hiếm) \end{cases}$$

  Càng gần 1 thì càng tốt

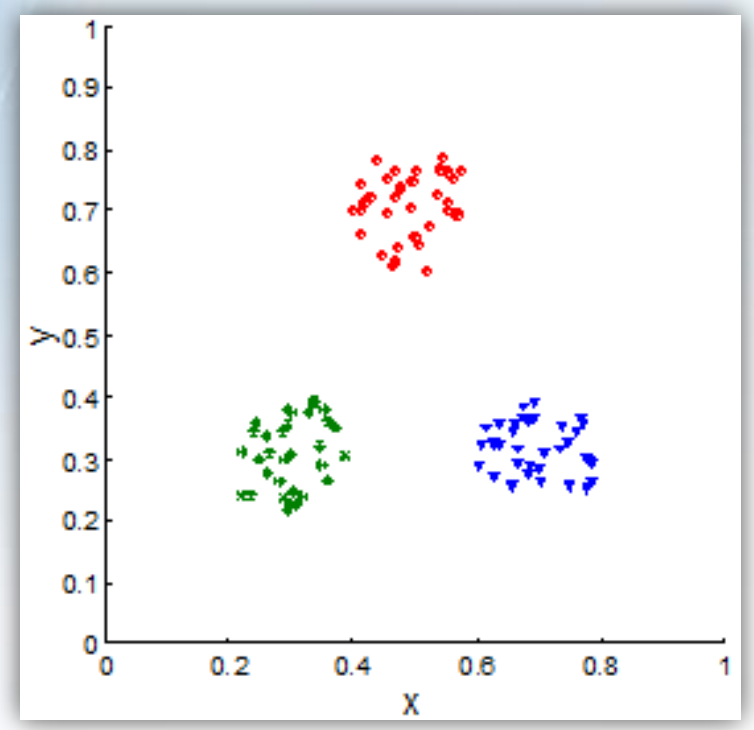
- Tính toán hệ số Silhouette trung bình cho nhóm hay cho toàn bộ quá trình gom nhóm

Độ tương quan

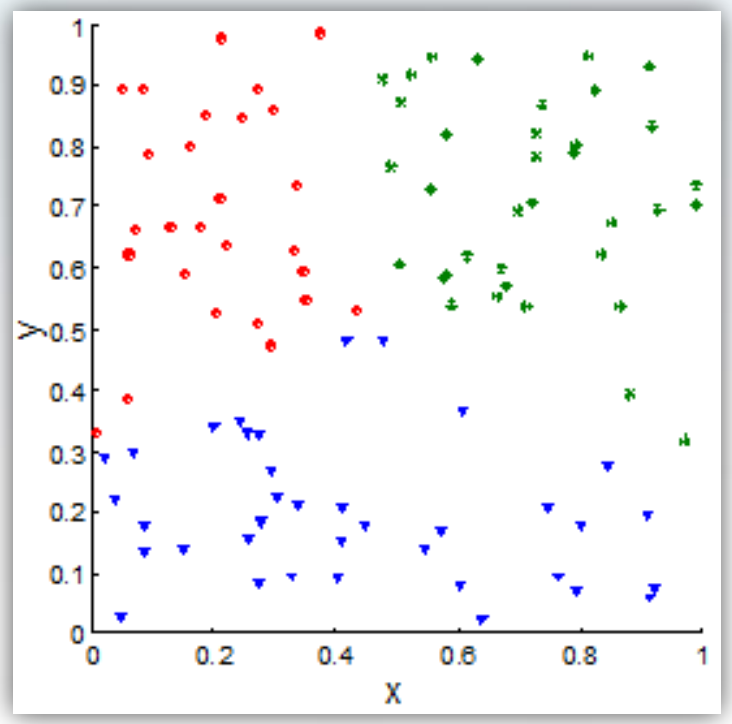
- Xét 2 ma trận:
  - Ma trận kề (ma trận khoảng cách)
  - Ma trận "incidence":
    - Hàng và cột là các điểm dữ liệu.
    - Giá trị tại (i,j) = 1 nếu cặp (i,j) thuộc về cùng một nhóm
    - Ngược lại = 0
- Tính toán độ tương quan giữa hai ma trận (xem phụ lục 2).
- Độ tương quan cao cho thấy các điểm thuộc về cùng một nhóm khá gần nhau.
- Tuy nhiên đây không phải là độ đo tốt cho gom nhóm dựa trên mật độ.

Ví dụ độ tương quan

- Độ tương quan của ma trận incidence và ma trận kề cho gom nhóm k-means ứng với 2 tập dữ liệu.



Corr = -0.9235



Corr = -0.5810

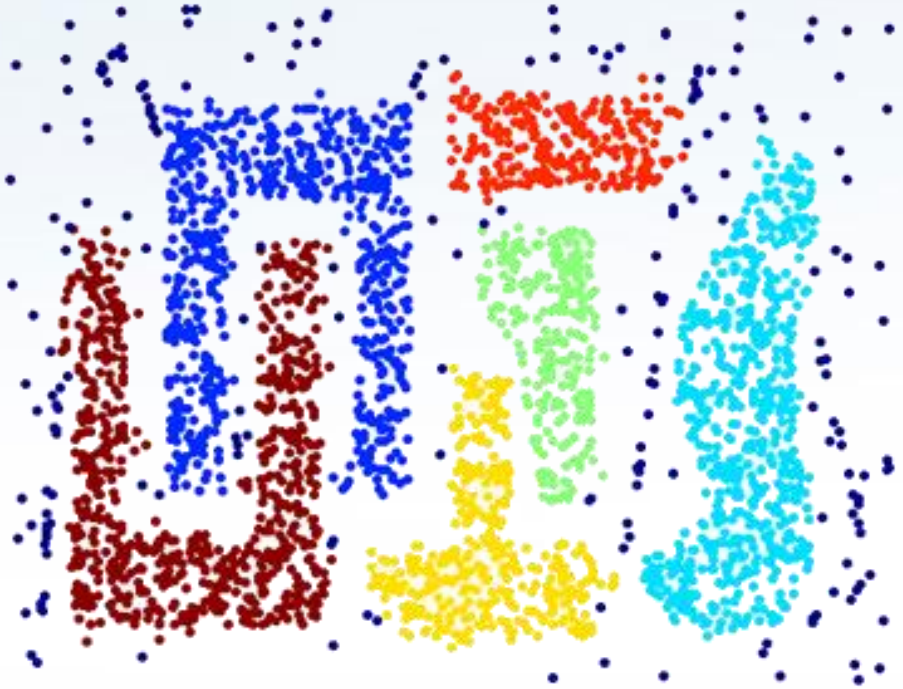
Nhận xét đánh giá gom nhóm

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Bài tập tổng hợp phần gom nhóm



Bài tập tổng hợp 1

1. Cho tập DL một chiều: {6, 12, 18, 24, 30, 42, 48}
   a) Với mỗi tập trung tâm nhóm sau, hình thành 2 nhóm đầu tiên dựa k-means (k=2). Tính tổng bình phương lỗi (SSE) cho từng tập 2 nhóm . So sánh kết quả
      - m1 = 18, m2 = 45
      - m1 = 15, m2 = 40
   b) Nếu tiếp tục chạy thuật toán k-mean(k=2) trên tập DL trên với các trung tâm nhóm đã cho, có sự thay đổi như thế nào?
   c) Sử dụng thuật toán Agnes với Single Link để xác định 2 nhóm từ DL trên. So sánh kết quả với k-mean (k=2 và chọn kết quả cho SSE tốt nhất)

Bài tập tổng hợp 2

2. Cho ma trận sai số sau. Hãy tính độ đo entropy và purity.

Cluster/ Class	Entertainment	Financial	Foreign	Metro	National	Sports	Tổng cluster
1	1	1	0	11	4	676	**693**
2	27	89	333	827	253	33	**1562**
3	326	465	8	105	16	29	**949**
Tổng class	**354**	**555**	**341**	**943**	**273**	**738**	**3204**

Tóm tắt

- Phương pháp dựa trên mật độ có khả năng khám phá nhóm có hình dạng bất kì, kiểm soát nhiễu, số lần đọc dữ liệu ít nhưng cần xác định các tham số như Eps, MintPts. DBSCAN dựa trên khái niệm liên thông mật độ để tìm các nhóm.
- Phương pháp dựa trên lưới có thể trả ra kết quả truy vấn với nhiều mức độ khác nhau dựa trên cấu trúc lưới đa phân giải, tham số thống kê của các ô ở mỗi lớp được tính toán trước và dựa trên các ô ở lớp dưới.
- Việc đánh giá gom nhóm là một quá trình khó khăn và phụ thuộc vào ngữ cảnh. Việc đánh giá được chia làm ba loại chính: đánh giá xu hướng gom nhóm, xác định số nhóm và đánh giá chất lượng nhóm dựa trên chỉ số trong, ngoài và tương quan.

Tài liệu tham khảo

1. J.Han, M.Kamber, Chương 10 – Cluster Analysis:Basic Concepts and Methods và Chương 11 – Advanced Cluster Analysis, cuốn "*Data mining: Basic Concepts and Methods*", 3rd edition
2. J.Han, M.Kamber, J.Pei, Chapter 10, www.cs.uiuc.edu/homes/hanj/cs412/bk3_slides/10ClusBasic.ppt
3. Tan, Steinbach, Kumar, Lecture Notes for Chapter 8 Introduction to Data Mining, http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap8_basic_cluster_analysis.pdf
4. Saurav.K.S, STING:STatistical INformation Grid, *http://www.cse.iitd.ernet.in/~cs5080225/file/presentation.pdf*

Hỏi & Đáp

Phụ lục 1: Phân phối trong STING

Trích từ http://www.cse.iitd.ernet.in/~cs5080225/file/presentation.pdf

- Set **dist** as the distribution type followed by most points in this cell
- Now check for conflicting points in the child cells call it *confl*.

1. If $dist_i \neq dist$, $m_i \approx m$ and $s_i \approx s$, then *confl* is increased by an amount of n_i;

2. If $dist_i \neq dist$, but either $m_i \approx m$ or $s_i \approx s$ is not satisfied, then set *confl* to n

3. If $dist_i = dist$, $m_i \approx m$ and $s_i \approx s$, then *confl* is increased by 0;

4. If $dist_i = dist$, but either $m_i \approx m$ or $s_i \approx s$ is not satisfied, then *confl* is set to n.

If *confl/n* is greater than a threshold *t* set **dist** as ONE.

Other wise keep the original type.

Phụ lục 2: Cách tính độ đo tương quan

Trích từ http://www.cs.kent.edu/~jin/DM11/ClusterValidation.ppt

Correlation Measure

- Hubert's Tau Statistics:

$$\Gamma = \frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_P(i,j)X_C(i,j)$$

- Normalized Tau Statistics:

$$\hat{\Gamma} = \frac{\frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(X_P(i,j)-\mu_P)(X_C(i,j)\mu_C)}{\sigma_P\sigma_C}$$

where μ_P and μ_C are the means and σ_P and σ_C are the variances of the matrices X_C and X_P.