KHAI THÁC DỮ LIỆU
(Data Mining)

Lê Văn Hạnh
levanhanhvn@gmail.com

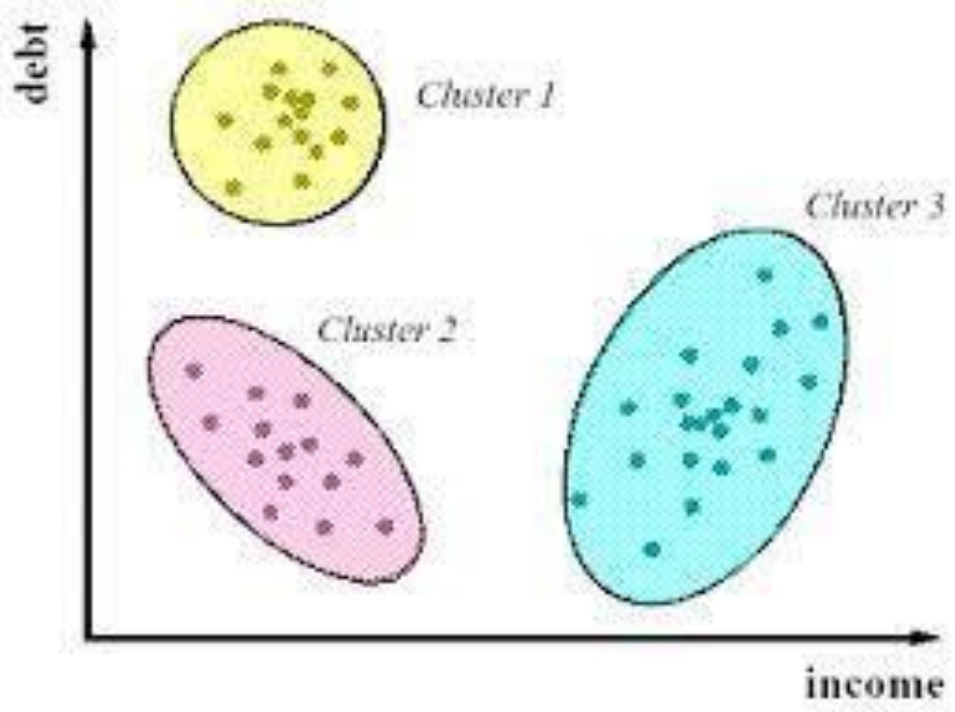
NỘI DUNG MÔN HỌC

1. Tổng quan về Data Science
2. Tìm hiểu dữ liệu
3. Tiền xử lý dữ liệu
4. Khai thác các mẫu phổ biến, mối kết hợp và mối tương quan
5. Phương pháp Phân loại (*Classification analysis method*)
6. Phương pháp phân tích cụm (*Cluster analysis method*)

Chương 6

PHƯƠNG PHÁP PHÂN TÍCH CỤM

(Cluster analysis method)



Lê Văn Hạnh
levanhanhvn@gmail.com

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

1. GIỚI THIỆU

Hãy tưởng tượng rằng bạn là Giám đốc Quan hệ Khách hàng tại công ty và bạn có năm người quản lý làm việc cho bạn. Bạn muốn tổ chức tất cả khách hàng của công ty thành năm nhóm để mỗi nhóm có thể được giao cho một người quản lý khác nhau.

Về mặt chiến lược, bạn muốn khách hàng trong mỗi nhóm càng giống nhau càng tốt. Hơn nữa, không nên xếp hai khách hàng có mô hình kinh doanh rất khác nhau vào cùng một nhóm. Mục đích của bạn đằng sau chiến lược kinh doanh này là phát triển các chiến dịch quan hệ khách hàng nhắm mục tiêu cụ thể đến từng nhóm, dựa trên các đặc điểm chung được chia sẻ bởi khách hàng trong mỗi nhóm.

⇒ Loại kỹ thuật khai thác dữ liệu nào có thể giúp bạn hoàn thành nhiệm vụ này?

1. Giới thiệu

- Phân cụm là quá trình nhóm một tập hợp các đối tượng dữ liệu thành nhiều nhóm hoặc cụm sao cho các đối tượng trong một cụm có độ tương tự cao nhưng rất khác với các đối tượng trong các cụm khác.
- Sự khác biệt và tương đồng được đánh giá dựa trên các giá trị thuộc tính mô tả đối tượng và thường liên quan đến thước đo khoảng cách.
- Phân cụm như một công cụ khai thác dữ liệu đã và đang được ứng dụng vào hầu hết các lĩnh vực/ngành nghề như marketing, sinh học, bảo hiểm, tài chính, bảo mật, kinh doanh thông minh, tìm kiếm trên Web, ...

1. Giới thiệu

- Một số ví dụ

i. ***Tiếp thị bán lẻ***: Các công ty bán lẻ thường sử dụng phương pháp phân cụm để xác định các nhóm hộ gia đình có điểm tương đồng với nhau. Ví dụ: một công ty bán lẻ có thể thu thập thông tin sau về các hộ gia đình:

- Thu nhập hộ gia đình
- Quy mô hộ gia đình
- Chủ hộ Nghề nghiệp
- Khoảng cách từ khu đô thị gần nhất

Sau đó, họ đưa các biến này vào một thuật toán phân cụm để có thể xác định các cụm sau:

- *Cluster 1*: Gia đình nhỏ, chi tiêu nhiều
- *Cluster 2*: Gia đình đông người hơn, chi tiêu nhiều hơn
- *Cluster 3*: Gia đình nhỏ, chi tiêu thấp
- *Cluster 4*: Gia đình đông người, chi tiêu thấp

Dựa vào kết quả phân cụm, công ty có thể gửi quảng cáo được cá nhân hóa hoặc thư bán hàng đến từng hộ gia đình dựa trên khả năng họ phản hồi với các loại quảng cáo cụ thể.

1. Giới thiệu

- Một số ví dụ

ii. Dịch vụ phát trực tuyến: Các dịch vụ phát trực tuyến thường sử dụng phân tích phân cụm để xác định những người xem có hành vi tương tự. Ví dụ: dịch vụ phát trực tuyến có thể thu thập dữ liệu sau về các cá nhân khi họ sử dụng dịch vụ:

- Số phút xem mỗi ngày
- Tổng số phiên xem mỗi tuần
- Số lượng chương trình độc đáo được xem mỗi tháng

Bằng cách sử dụng các số liệu này, dịch vụ phát trực tuyến có thể thực hiện phân tích cụm để xác định người dùng thuộc loại nào trong 2 loại sau để họ có thể biết họ nên chi phần lớn số tiền quảng cáo cho ai.

- Mức sử dụng cao
- Mức sử dụng thấp

1. Giới thiệu

- Một số ví dụ

iii. Trong Khoa học thể thao: Các nhà khoa học dữ liệu cho các đội thể thao thường sử dụng phương pháp phân cụm để xác định những vận động viên thi đấu giống nhau. Ví dụ: các đội bóng rổ chuyên nghiệp có thể thu thập thông tin sau về từng vận động Điểm ghi được.

- Tỷ lệ ném rỗ thành công
- Mức độ phối hợp của VDV với các đồng đội trong thi đầu
- Khả năng tranh cướp bóng
- Sự phục hồi sau mỗi trận

Sau đó, họ có thể đưa các biến này vào một thuật toán phân cụm để xác định những người chơi giống nhau để có thể yêu cầu những người chơi này luyện tập với nhau và thực hiện các bài tập cụ thể dựa trên điểm mạnh và điểm yếu của họ.

1. Giới thiệu

- Một số ví dụ

iv. ***Tiếp thị qua Email:*** Nhiều doanh nghiệp sử dụng phân tích cụm để xác định những người tiêu dùng tương tự nhau để họ có thể điều chỉnh email gửi đến người tiêu dùng theo cách tối đa hóa doanh thu của họ. Ví dụ: một doanh nghiệp có thể thu thập thông tin sau về người tiêu dùng:

- Tỷ lệ email được mở
- Số lần nhấp chuột trên mỗi email
- Thời gian xem email

Bằng cách sử dụng các số liệu này, doanh nghiệp có thể thực hiện phân tích cụm để xác định người tiêu dùng sử dụng email theo những cách tương tự và điều chỉnh loại email cũng như tần suất email họ gửi cho các cụm khách hàng khác nhau.

1. Giới thiệu

- Một số ví dụ

v. ***Bảo hiểm y tế:*** Các chuyên gia tính toán tại các công ty bảo hiểm y tế thường sử dụng phân tích cụm để xác định "cụm" người tiêu dùng sử dụng bảo hiểm y tế của họ theo những cách cụ thể. Ví dụ: chuyên gia tính toán có thể thu thập thông tin sau về hộ gia đình:

- Tổng số lượt khám bác sĩ mỗi năm
- Tổng quy mô hộ gia đình
- Tổng số bệnh mãn tính trong mỗi hộ gia đình
- Tuổi trung bình của các thành viên trong hộ

Sau đó, chuyên gia tính toán có thể đưa các biến này vào thuật toán phân cụm để xác định các hộ gia đình tương tự nhau. Sau đó, công ty bảo hiểm y tế có thể ấn định phí bảo hiểm hàng tháng dựa trên tần suất họ mong đợi các hộ gia đình trong các cụm cụ thể sẽ sử dụng bảo hiểm của mình.

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

2. PHÂN TÍCH CỤM (Cluster Analysis)

2.1. Phân tích cụm là gì?

- Phân tích cụm (*Cluster analysis*) hoặc đơn giản là phân cụm (*clustering*) là quá trình phân vùng một tập hợp các đối tượng dữ liệu (hoặc quan sát) thành các tập hợp con. Mỗi tập hợp con là một cụm (cluster), sao cho các đối tượng trong một cụm tương tự với nhau nhưng không giống với các đối tượng trong các cụm khác.
- Tập hợp các cụm thu được từ phân tích cụm có thể được gọi là phân cụm.
- Các phương pháp phân cụm khác nhau có thể tạo ra các phân cụm khác nhau trên cùng một tập dữ liệu.
- Việc phân vùng không phải do con người thực hiện mà do thuật toán phân cụm. Do đó, phân cụm rất hữu ích ở chỗ nó có thể dẫn đến việc phát hiện ra các nhóm chưa biết trước đó trong dữ liệu.

2.2. Yêu cầu đối với phân tích cụm?

2.2.1. Các yêu cầu điển hình của phân cụm trong khai thác dữ liệu

i. ***Khả năng mở rộng*** (*Scalability*): Do có nhiều thuật toán phân cụm hoạt động tốt trên các tập dữ liệu nhỏ nhưng có thể chưa tốt trên CSDL lớn (chứa hàng triệu hoặc hàng tỷ đối tượng).

ii. ***Khả năng xử lý các kiểu thuộc tính khác nhau*** (*Ability to deal with different types of attributes*): Nhiều thuật toán được thiết kế để phân cụm dữ liệu số. Tuy nhiên, các ứng dụng có thể yêu cầu phân cụm các loại dữ liệu khác như dữ liệu nhị phân, danh nghĩa (*nominal*) và thứ tự (*ordinal*) hoặc hỗn hợp các loại dữ liệu này. Gần đây, ngày càng có nhiều ứng dụng cần kỹ thuật phân cụm cho các kiểu dữ liệu phức tạp như biểu đồ, chuỗi, hình ảnh và tài liệu.

2.2.1. Các yêu cầu điển hình của phân cụm trong khai thác dữ liệu

iii. Khám phá các cụm có hình dạng tùy ý

- Nhiều thuật toán phân cụm xác định các cụm dựa trên thước đo khoảng cách *Euclide* hoặc *Manhattan*. Các thuật toán dựa trên các thước đo khoảng cách như vậy có xu hướng tìm các cụm hình cầu có kích thước và mật độ tương tự nhau.
- Tuy nhiên, trong thực tế một cụm có thể có hình dạng bất kỳ. Ví dụ, muốn sử dụng phương pháp phân cụm để tìm ra ranh giới của một đám cháy rừng đang diễn ra.

iv. Yêu cầu về kiến thức lĩnh vực để xác định tham số đầu vào

- Nhiều thuật toán phân cụm yêu cầu người dùng cung cấp kiến thức lĩnh vực dưới dạng tham số đầu vào như số cụm mong muốn.
- Do đó, kết quả phân cụm có thể nhạy cảm với các tham số như vậy. Các tham số thường khó xác định, đặc biệt đối với các tập dữ liệu có nhiều chiều và khi người dùng chưa hiểu rõ về dữ liệu của họ.

  ⇒ Việc yêu cầu đặc tả kiến thức lĩnh vực không chỉ tạo gánh nặng cho người dùng mà còn khiến chất lượng phân cụm khó kiểm soát.

2.2.1. Các yêu cầu điển hình của phân cụm trong khai thác dữ liệu

- ***Khả năng xử lý dữ liệu nhiễu***
  - Hầu hết các bộ dữ liệu trong thế giới thực đều chứa các dữ liệu ngoại lệ và/hoặc dữ liệu bị thiếu, không xác định hoặc sai. Ví dụ, các kết quả đọc từ cảm biến thường nhiễu do cơ chế của cảm biến và một số kết quả đo có thể sai do sự can thiệp từ các vật thể thoáng qua xung quanh.
  - Các thuật toán phân cụm có thể nhạy cảm với nhiễu như vậy và có thể tạo ra các cụm có chất lượng kém.

    ⇒ Cần các phương pháp phân cụm có khả năng chống nhiễu tốt.
- ***Phân cụm gia tăng và không nhạy cảm với thứ tự đầu vào***
  - Trong thực tế nhiều ứng dụng sẽ cần thường xuyên cập nhật thêm mới dữ liệu. Một số thuật toán phân cụm không thể kết hợp các cập nhật gia tăng vào các cấu trúc phân cụm hiện có mà phải tính toán lại một phân cụm mới từ đầu.
  - Các thuật toán phân cụm cũng có thể nhạy cảm với thứ tự dữ liệu đầu vào. Nghĩa là, các thuật toán phân cụm có thể trả về các phân cụm khác nhau đáng kể tùy thuộc vào thứ tự các đối tượng được đưa vào.

    ⇒ Cần các thuật toán phân cụm tăng dần và các thuật toán không nhạy cảm với thứ tự đầu vào.

2.2.1. Các yêu cầu điển hình của phân cụm trong khai thác dữ liệu

vii. *Khả năng phân cụm dữ liệu nhiều chiều*

- Một tập dữ liệu có thể chứa nhiều chiều (dimensions hoặc thuộc tính). Ví dụ: khi phân cụm tài liệu, mỗi từ khóa có thể được coi là một chiều và thường có hàng nghìn từ khóa.
- Hầu hết các thuật toán phân cụm đều xử lý tốt dữ liệu ít chiều. Việc tìm kiếm các cụm đối tượng dữ liệu trong một không gian nhiều chiều là một thách thức, đặc biệt khi dữ liệu đó có thể rất thưa thớt và có độ lệch cao.

viii. *Phân cụm dựa trên ràng buộc*

- Các ứng dụng trong thế giới thực có thể cần thực hiện phân cụm theo nhiều loại ràng buộc khác nhau. Ví dụ: việc chọn địa điểm để lắp đặt mới cho một số máy rút tiền tự động (ATM) trong một thành phố sẽ chịu 1 số ràng buộc về:
  - Mạng lưới sông ngòi và đường cao tốc của thành phố
  - Loại hình và số lượng khách hàng trong mỗi cụm.

$\Rightarrow$ Cần tìm các nhóm dữ liệu có khả năng phân cụm tốt, thỏa mãn các ràng buộc đã chỉ định.

2.2.1. Các yêu cầu điển hình của phân cụm trong khai thác dữ liệu

ix. Khả năng diễn giải và khả năng sử dụng

- Người dùng muốn kết quả phân cụm có thể diễn giải được, dễ hiểu và có thể sử dụng được. Nghĩa là, việc phân cụm có thể cần phải gắn liền với các cách diễn giải và ứng dụng ngữ nghĩa cụ thể.
- Điều quan trọng là nghiên cứu xem mục tiêu ứng dụng có thể ảnh hưởng như thế nào đến việc lựa chọn các tính năng phân cụm và phương pháp phân cụm.

2.2. Yêu cầu đối với phân tích cụm?

2.2.2. Các khía cạnh mà các phương pháp phân cụm có thể được so sánh

i. *Tiêu chí phân vùng*

- ***Không** có thứ bậc trong việc phân vùng*: tất cả các cụm đều ở cùng cấp độ về mặt khái niệm. Ví dụ phân chia khách hàng thành các nhóm để mỗi nhóm có người quản lý riêng.
- ***Có** thứ bậc trong việc phân vùng:* các cụm có thể được hình thành ở các cấp độ ngữ nghĩa khác nhau. Ví dụ: trong khai thác văn bản, có thể sắp xếp một kho tài liệu thành nhiều chủ đề (như "chính trị", "thể thao", …). Mỗi chủ đề có thể có chủ đề phụ. Ví dụ: chủ đề thể thao gồm các chủ đề phụ "bóng đá", "bóng rổ", " bóng chày" và "khúc côn cầu".

ii. *Tách các cụm*

- ***1 đối tượng chỉ thuộc duy nhất 1 cụm*** => các cụm loại trừ lẫn nhau (*mutually exclusive clusters*). Khi phân nhóm khách hàng thành các nhóm sao cho mỗi nhóm do một người quản lý chăm sóc, mỗi khách hàng chỉ có thể thuộc một nhóm.
- ***1 đối tượng có thể thuộc nhiều cụm khác nhau*** => các cụm không độc quyền (*not be exclusive*). Ví dụ: khi phân cụm tài liệu thành các chủ đề, một tài liệu có thể liên quan đến nhiều chủ đề. Vì vậy, các chủ đề dưới dạng cụm có thể không độc quyền.

2.2. Yêu cầu đối với phân tích cụm?

2.2.2. Các khía cạnh mà các phương pháp phân cụm có thể được so sánh

iii. Độ đo độ tương tự

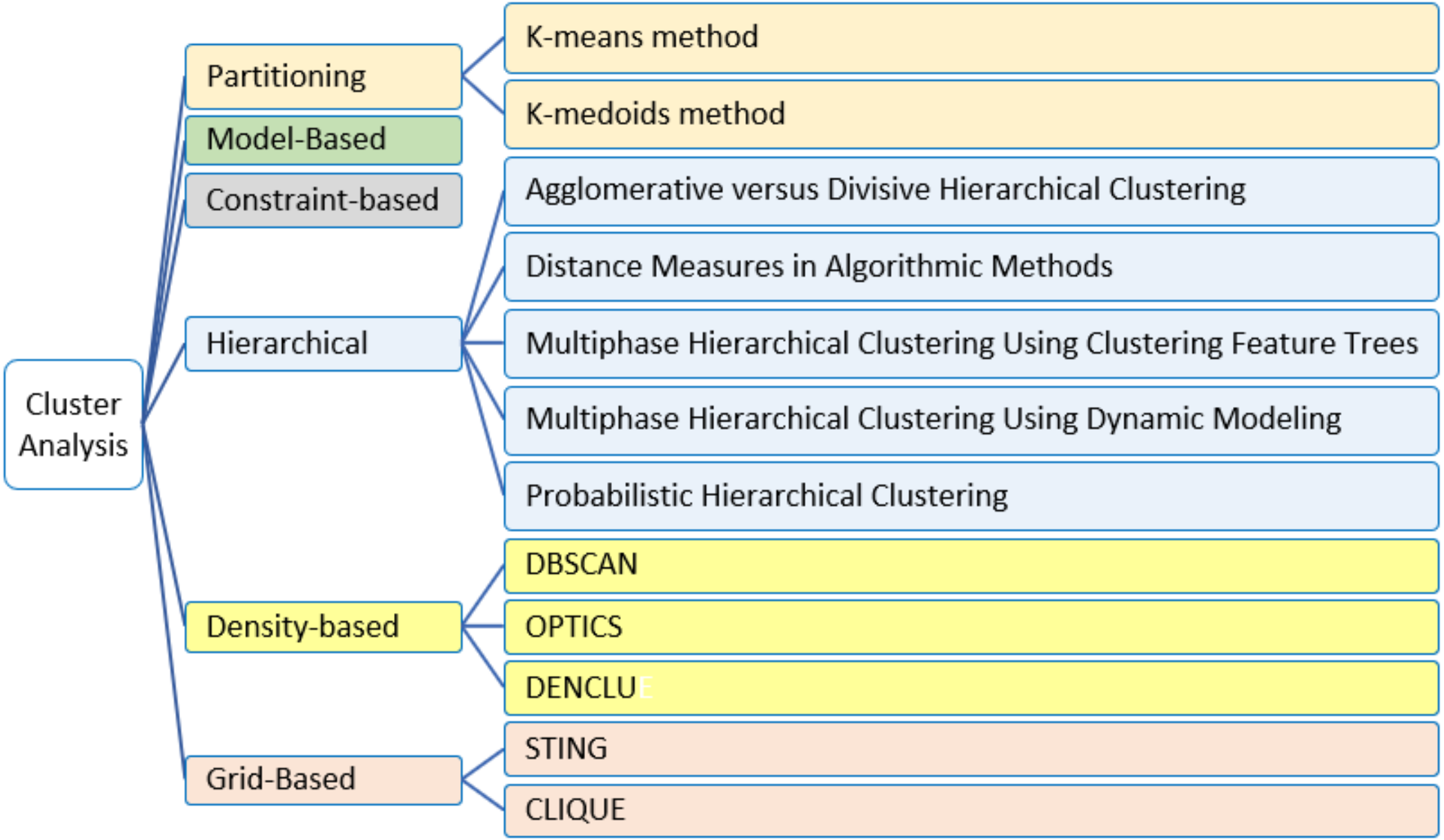
Việc xác định độ tương tự giữa hai đối tượng dựa trên:

- Khoảng cách giữa chúng: xác định trên không gian Euclide, không gian vectơ, … với đặc điểm là có thể tận dụng các kỹ thuật tối ưu hóa.
- Dựa trên mật độ hoặc sự liền kề. Với đặc điểm thường có thể tìm thấy các cụm có hình dạng tùy ý.

iv. Không gian phân cụm

- Khi thực hiện phân cụm trên toàn bộ không gian dữ liệu, có thể có nhiều thuộc tính không liên quan, có thể làm cho các phép đo độ tương tự không đáng tin cậy. Do đó, các cụm được tìm thấy trong không gian đầy đủ thường vô nghĩa.
- nên tìm kiếm các cụm trong các không gian con khác nhau của cùng một tập dữ liệu, qua đó giúp phát hiện các cụm và không gian con thể hiện sự tương đồng tốt hơn giữa các đối tượng.

2.3. Các phương pháp phân tích cụm



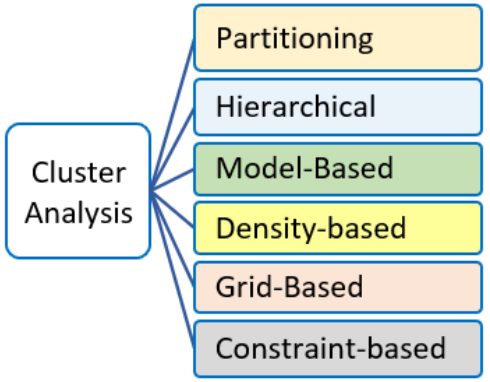
2.3. Các phương pháp phân tích cụm

Phương pháp	*Đặc điểm chung*
Partitioning methods	– Tìm các cụm hình cầu loại trừ lẫn nhau – Dựa trên khoảng cách – Có thể sử dụng giá trị trung bình hoặc trung gian (medoid) để biểu diễn trung tâm cụm – Hiệu quả đối với các tập dữ liệu có kích thước vừa và nhỏ
Hierarchical methods	– Phân cụm là sự phân rã theo thứ bậc (tức là nhiều cấp độ) – Không thể sửa lỗi sáp nhập hoặc chia tách – Có thể kết hợp các kỹ thuật khác như phân cụm vi mô (*micro clustering*) hoặc xem xét các "liên kết" đối tượng (*object "linkages"*)
Density based methods	– Có thể tìm các cụm có hình dạng tùy ý – Cụm là các vùng dày đặc của các vật thể trong không gian được ngăn cách bởi các vùng có mật độ thấp – Mật độ cụm: Mỗi điểm phải có số lượng điểm tối thiểu trong "lân cận" của nó – Có thể lọc ra các ngoại lệ
Grid based methods	– Sử dụng cấu trúc dữ liệu lưới đa độ phân giải (*multi resolution grid data structure*) – Thời gian xử lý nhanh (thường không phụ thuộc vào số lượng đối tượng dữ liệu nhưng vẫn phụ thuộc vào kích thước lưới)

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

3. CÁC PHƯƠNG PHÁP PHÂN VÙNG (*Partitioning Methods*)



3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.1. Phân vùng dựa trên trung tâm

- Giả sử tập dữ liệu D chứa n đối tượng trong không gian Euclide. Các phương pháp phân vùng phân bố các đối tượng trong D thành k cụm, C_1, ..., C_k, tức là $C_i \subset D$ và $C_i \cap C_j = \emptyset$ với $(1 \leq i,j \leq k)$.
- Hàm mục tiêu được sử dụng để đánh giá chất lượng phân vùng sao cho các đối tượng trong một cụm giống nhau nhưng không giống với các đối tượng trong các cụm khác. Nghĩa là, hàm mục tiêu hướng các đối tượng trong cùng cụm có độ tương tự cao và độ tương tự giữa 2 đối tượng khác cụm là thấp.
- Kỹ thuật phân vùng dựa trên trung tâm (*centroid-based*) sử dụng trung tâm C_i của một cụm để biểu diễn cụm đó. Về mặt khái niệm, tâm của một cụm là điểm trung tâm của nó.
- Trọng tâm có thể được xác định theo nhiều cách khác nhau như theo giá trị trung bình (mean) hoặc trung gian (*medoid* như *median, mode*, ...) của các đối tượng.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.1. Phân vùng dựa trên trung tâm

- Sự khác biệt giữa một đối tượng $\boldsymbol{p} \in C_i$ và c_i (là đại diện của cụm), được đo bằng $dist(\boldsymbol{p}, c_i)$, trong đó $dist(\boldsymbol{x}, \boldsymbol{y})$ là khoảng cách Euclide giữa hai điểm $\boldsymbol{x}$ và $\boldsymbol{y}$.
- Tổng sai số bình phương giữa tất cả các đối tượng trong C_i và tâm c_i, được định nghĩa là

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

trong đó:

- E là tổng sai số bình phương của tất cả các đối tượng trong tập dữ liệu;
- $\boldsymbol{p}$ là điểm trong không gian đại diện cho một đối tượng nhất định;
- $\mathbf{c_i}$ là trọng tâm của cụm C_i (cả $\boldsymbol{p}$ và $\boldsymbol{c_i}$ đều đa chiều).

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.2. Thuật toán K-means

- Là phiên bản đơn giản và cơ bản nhất của phân tích cụm là phân vùng, tổ chức các đối tượng của một tập hợp thành một số nhóm hoặc cụm độc quyền (*exclusive groups or clusters groups*).
- Về mặt hình thức, với một tập dữ liệu D gồm n đối tượng và k là số cụm cần hình thành, thuật toán phân vùng sẽ tổ chức các đối tượng thành k phân vùng ($k \leq n$), trong đó mỗi phân vùng đại diện cho một cụm. Các cụm được hình thành để tối ưu hóa một tiêu chí phân vùng một cách khách quan, chẳng hạn như hàm khác biệt dựa trên khoảng cách, sao cho các đối tượng trong một cụm là "tương tự" với nhau và "không giống" với các đối tượng trong các cụm khác về các thuộc tính tập dữ liệu.

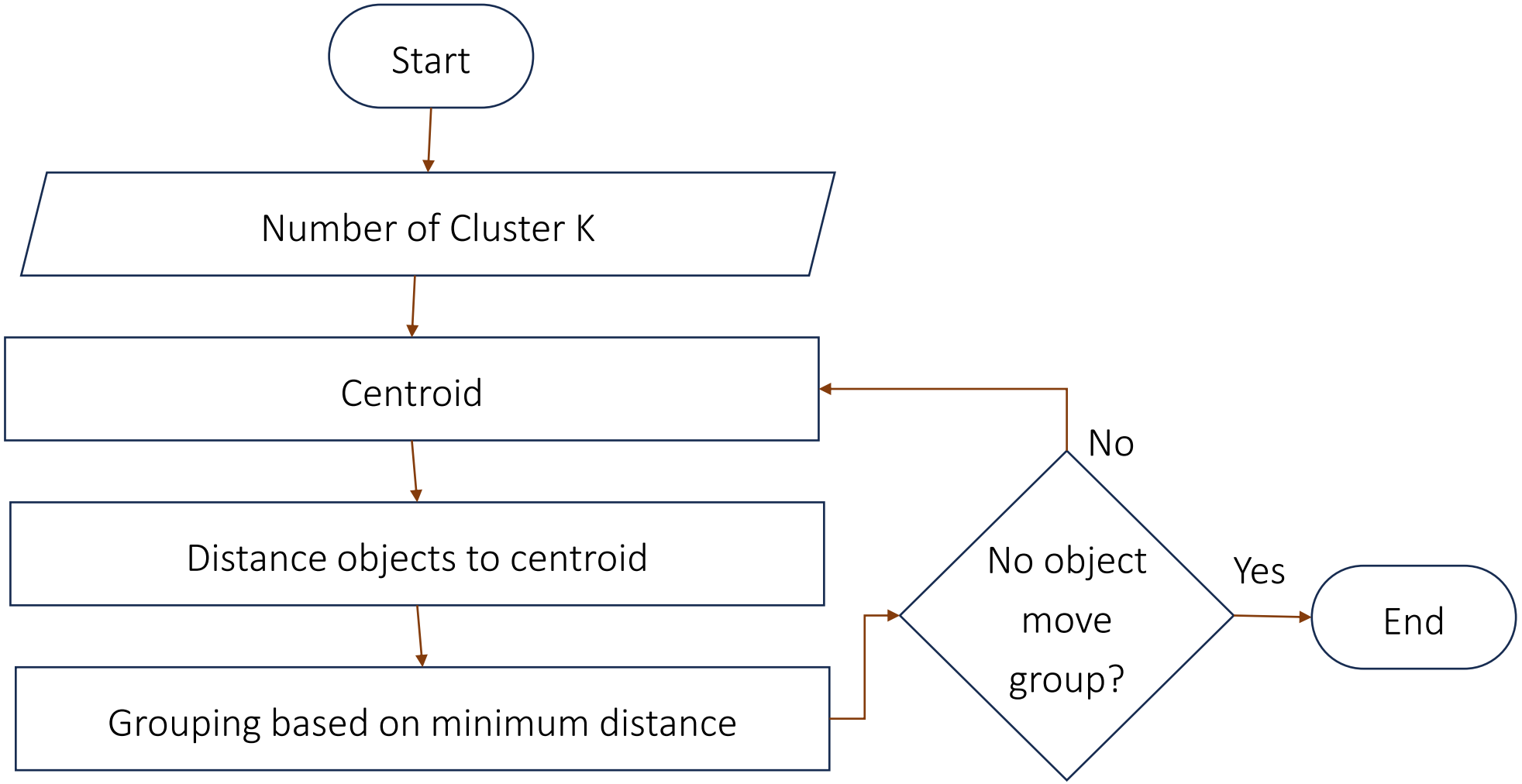
3.1.2. Thuật toán K-means

- Thuật toán k-means xác định trọng tâm của cụm là giá trị trung bình của các điểm trong cụm. Nó tiến hành như sau:
  - Đầu tiên, chọn ngẫu nhiên k đối tượng trong D làm tâm của k cụm.
  - Mỗi đối tượng còn lại sẽ được gán cho cụm mà nó giống nhất, dựa trên khoảng cách Euclide giữa đối tượng và giá trị trung bình của cụm.
  - Lặp lại nhiều lần các công việc sau cho đến khi tất cả các cụm không còn xuất hiện sự thay đổi nào (các cụm được hình thành ở vòng hiện tại giống với các cụm được hình thành ở vòng trước) bằng cách:
    - Đối với mỗi cụm, tính toán giá trị tâm mới bằng cách lấy trung bình trong cụm () bằng cách sử dụng các đối tượng được gán cho cụm trong lần lặp trước đó.
    - Dựa trên khoảng cách từ đối tượng đến tâm mới, đối tượng sẽ thuộc về cụm nào mà khoảng cách giữa đối tượng và tâm mới đó là nhỏ nhất.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.2. Thuật toán K-means

- Lưu đồ Thuật toán k-means:



3.1.3. *Ví dụ minh họa thuật toán K-Means*

- Giả sử ta có 4 loại thuốc A, B, C, D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau:

Medicine	*Feature 1 (X)*: weight index	*Feature 2 (Y)*: pH
A	1	1
B	2	1
C	4	3
D	5	4

- Yêu cầu: nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 1:

- ***Bước 1.*** *Khởi tạo tâm (centroid) cho 2 nhóm*. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất $C_1(1,1)$) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai C_2 (2,1)).
- ***Bước 2.*** Tính khoảng cách từ các đối tượng đến tâm của các nhóm (khoảng cách Euclidean)

		A	B	C	D		
	X	1	2	4	5		
	Y	1	1	3	4		
D^1=	[	**0**	1 (i)	3.61 (ii)	5 (iii)	]	Group1: C_1=(1,1)
		1 (iv)	**0**	**2.83** (v)	**4.24** (vi)		Group2: C_2=(2,1)

- Tính khoảng cách từ điểm B(2,1), C(4,3) và D(5,4) đến tâm A(1,1) của Group1:

(i) $d(A,B) = \sqrt{(1-2)^2 + (1-1)^2} = \sqrt{1} = 1$

(ii) $d(A,C) = \sqrt{(1-4)^2 + (1-3)^2} = \sqrt{13} = 3.61$

(iii) $d(A,D) = \sqrt{(1-5)^2 + (1-4)^2} = \sqrt{25} = 5$

- Tính khoảng cách từ điểm A(1,1), C(4,3) và D(5,4) đến tâm B(2,1) của Group2:

(iv) $d(B,A) = \sqrt{(2-1)^2 + (1-1)^2} = \sqrt{1} = 1$

(v) $d(B,C) = \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{8} = 2.83$

(vi) $d(B,D) = \sqrt{(5-2)^2 + (4-1)^2} = \sqrt{18} = 4.24$

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 1:

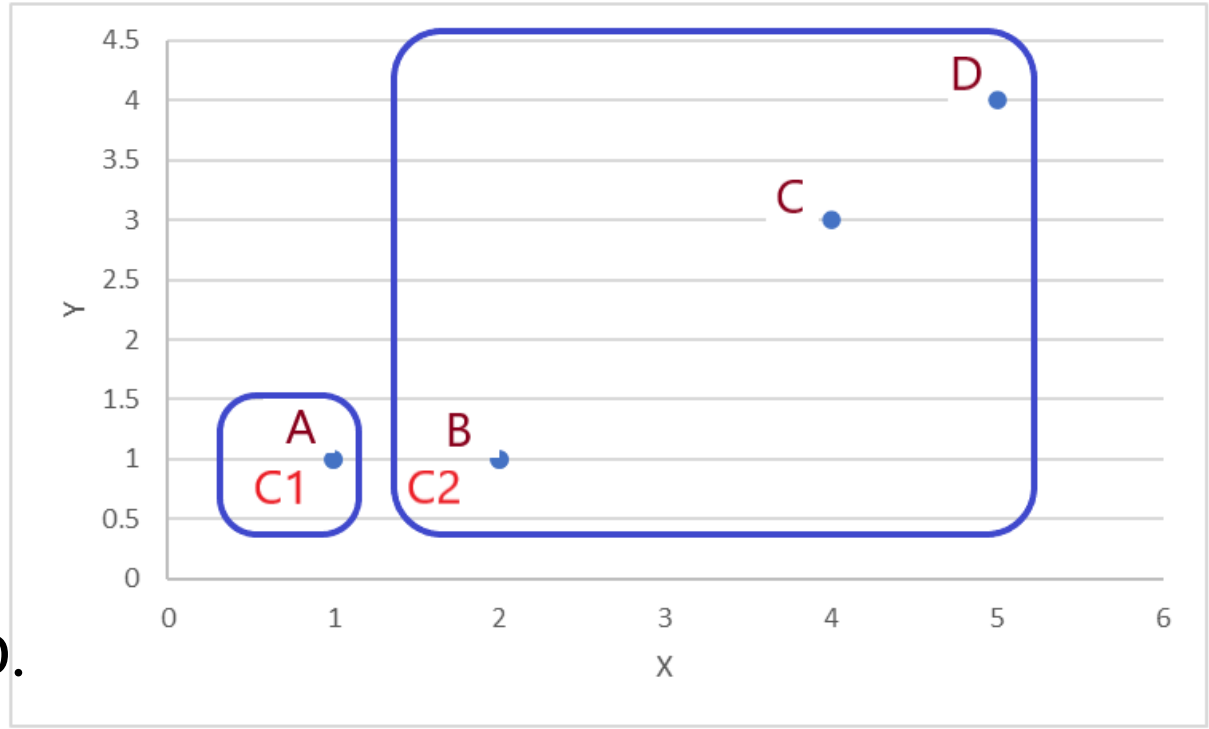
- ***Bước 3.*** Nhóm các đối tượng vào nhóm có tâm gần nhất

$G^1=$

A	B	C	D
1	0	0	0
0	1	1	1

Group1: $C_1=(1,1)$

Group2: $C_2=(2,1)$



⇒ Sau là lặp thứ 1,

nhóm 1 gồm có 1 đối tượng A

nhóm 2 gồm các đối tượng còn lại B, C, D.

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 2:

- ***Bước 1.*** Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, $C_1(1, 1)$. Tâm nhóm 2 được tính như sau:

$$C_2=\left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right)=\left(\frac{11}{3}, \frac{8}{3}\right)$$

- ***Bước 2.*** Tính lại khoảng cách từ các đối tượng đến tâm của các nhóm mới (khoảng cách Euclidean)

		A	B	C	D		
	X	1	2	4	5		
	Y	1	1	3	4		
$D^1=$	[	**0**	**1**(vii)	3.61 (viii)	5 (ix)	]	Group1: $C_1=(1,1)$
		3.14 (x)	1.83 (xi)	**0.48** (xii)	**1.89** (xiii)		Group2: $C_2=(11/3, 8/3)$

- Tính khoảng cách từ điểm B(2,1), C(4,3) và D(5,4) đến tâm $C_1(1,1)$ của Group1:

(vii) $d(C_1,B)=\sqrt{(1-2)^2+(1-1)^2}=\sqrt{1}=1$

(viii) $d(C_1,C)=\sqrt{(1-4)^2+(1-3)^2}=\sqrt{13}=3.61$

(ix) $d(C_1,D)=\sqrt{(1-5)^2+(1-4)^2}=\sqrt{25}=5$

- Tính khoảng cách từ điểm A(1,1), B(2,1), C(4,3) và D(5,4) đến tâm mới $C_2(11/3, 8/3)$ của Group2:

(x) $d(C_2,A)=\sqrt{(\frac{11}{3}-1)^2+(\frac{8}{3}-1)^2}=\sqrt{7.111+2.777}=3.14$

(xi) $d(C_2,B)=\sqrt{(\frac{11}{3}-2)^2+(\frac{8}{3}-1)^2}=\sqrt{1.666+1.666}=1.83$

(xii) $d(C_2,C)=\sqrt{(\frac{11}{3}-4)^2+(\frac{8}{3}-3)^2}=\sqrt{0.111+0.444}=0.48$

(xiii) $d(C_2,D)=\sqrt{(\frac{11}{3}-5)^2+(\frac{8}{3}-4)^2}=\sqrt{3.554}=1.89$

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 2:

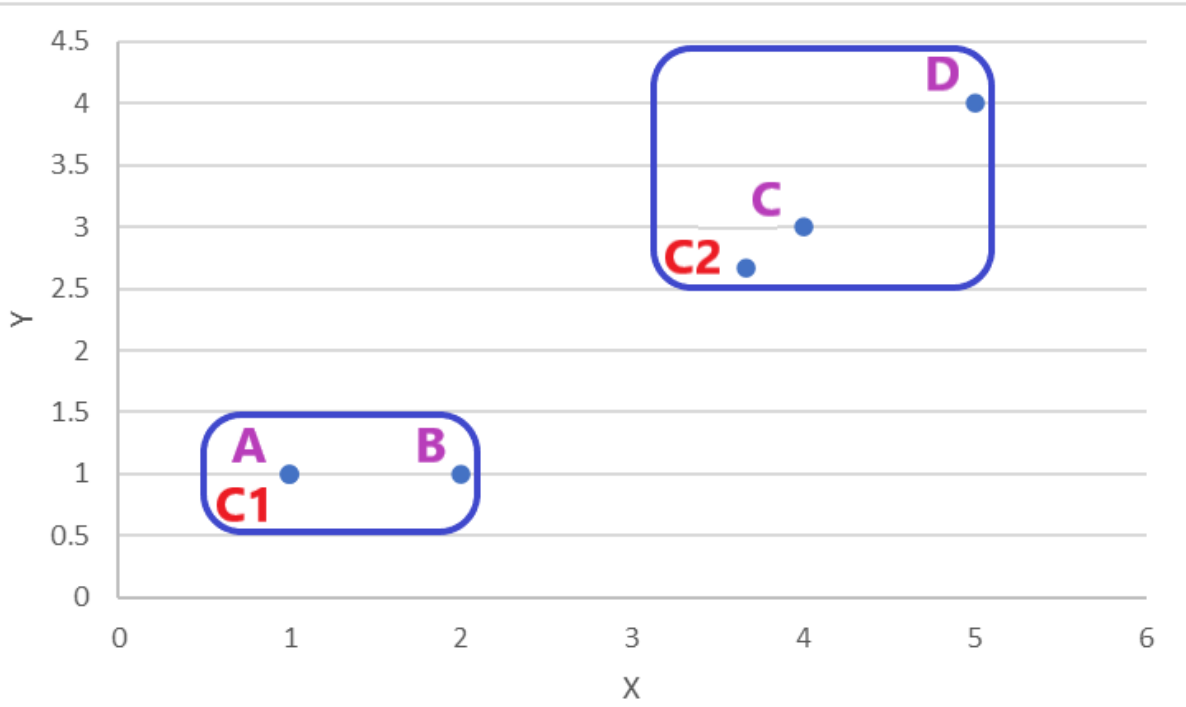
- ***Bước 3.*** Nhóm các đối tượng vào nhóm có tâm gần nhất

$G^2=$

A	B	C	D
1	1	0	0
0	0	1	1

Group1: $C_1=(1,1)$

Group2: $C_2=(2,1)$



⇒ Sau lần lặp thứ 2,

nhóm 1 gồm có 2 đối tượng A, B

nhóm 2 gồm các đối tượng còn lại C, D.

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 3:

- ***Bước 1.*** Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm. Nhóm 1 chỉ có 1 đối tượng A nên tâm nhóm 1 vẫn không đổi, $C_1(1,1)$. Tâm nhóm 2 được tính như sau:

$$C_1=\left(\frac{1+2}{2}, \frac{1+1}{2}\right)=\left(\frac{3}{2}, 1\right)$$

$$C2=\left(\frac{4+5}{2}, \frac{3+4}{2}\right)=\left(\frac{9}{2}, \frac{7}{2}\right)$$

- ***Bước 2.*** Tính lại khoảng cách từ các đối tượng đến tâm của các nhóm mới (khoảng cách Euclidean)

	A	*B*	*C*	*D*
X	1	2	4	5
Y	1	1	3	4

$D^1=$

[	**0.05** (xiv)	**0.05**(xv)	3.2 (xvi)	4.61 (xvii)
	4.71 (xviii)	3.53 (xix)	**0.71** (xx)	**0.71** (xxi)

]

Group1: $C_1=(3/2,1)$

Group2: $C_2=(9/2, 7/2)$

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 3:

- ***Bước 2.*** Tính lại khoảng cách từ các đối tượng đến tâm của các nhóm mới (khoảng cách Euclidean)

	A	B	C	D
X	1	2	4	5
Y	1	1	3	4

$D^1 = \begin{bmatrix} \mathbf{0.05}\ (xiv) & \mathbf{0.05}(xv) & 3.2\ (xvi) & 4.61\ (xvii) \\ 4.71\ (xviii) & 3.53\ (xix) & \mathbf{0.71}\ (xx) & \mathbf{0.71}\ (xxi) \end{bmatrix}$

Group1: $C_1=(3/2,1)$
Group2: $C_2=(9/2, 7/2)$

- Tính khoảng cách từ điểm A(1,1), B(2,1), C(4,3) và D(5,4) đến tâm $C_1(3/2, 1)$ của Group1:

$$(xiv)\ d(C_1,A) = \sqrt{(3/2-1)^2+(1-1)^2} = \sqrt{0.25} = 0.05$$

$$(xv)\ d(C_1,B) = \sqrt{(3/2-2)^2+(1-1)^2} = \sqrt{0.25} = 0.05$$

$$(xvi)\ d(C_1,C) = \sqrt{(3/2-4)^2+(1-3)^2} = \sqrt{6.25+4} = 3.2$$

$$(xvii)\ d(C_1,D) = \sqrt{(3/2-5)^2+(1-4)^2} = \sqrt{12.25+9} = 4.61$$

- Tính khoảng cách từ điểm A(1,1), B(2,1), C(4,3) và D(5,4) đến tâm mới $C_2(11/3, 8/3)$ của Group2:

$$(xviii)\ d(C_2,A) = \sqrt{(\tfrac{9}{2}-1)^2+(\tfrac{7}{2}-1)^2} = \sqrt{16+6.25} = 4.71$$

$$(xix)\ d(C_2,B) = \sqrt{(\tfrac{9}{2}-2)^2+(\tfrac{7}{2}-1)^2} = \sqrt{6.25+6.25} = 3.53$$

$$(xx)\ d(C_2,C) = \sqrt{(\tfrac{9}{2}-4)^2+(\tfrac{7}{2}-3)^2} = \sqrt{0.25+0.25} = 0.71$$

$$(xxi)\ d(C_2,D) = \sqrt{(\tfrac{9}{2}-5)^2+(\tfrac{7}{2}-4)^2} = \sqrt{0.25+0.25} = 0.71$$

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.3. *Ví dụ minh họa thuật toán K-Means*

Lặp lần 3:

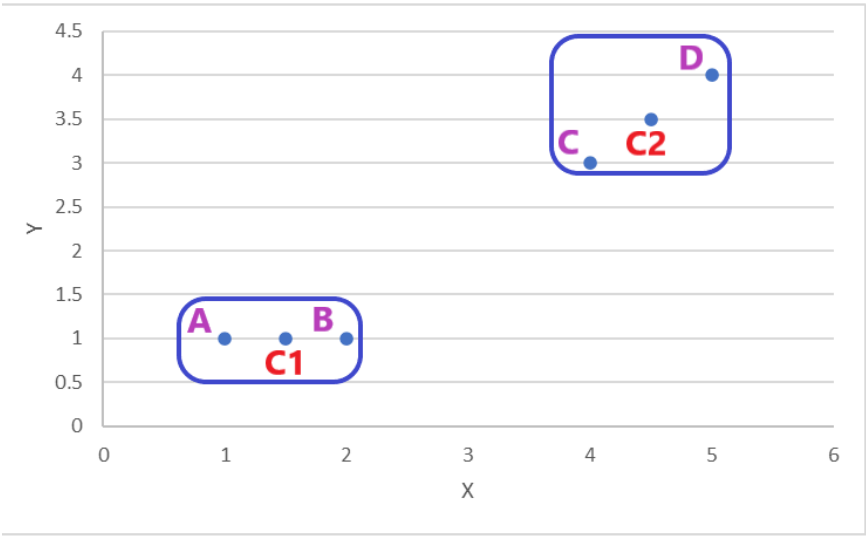
- ***Bước 3.*** Nhóm các đối tượng vào nhóm có tâm gần nhất

$G^1 =$

A	B	C	D
1	1	0	0
0	0	1	1

Group1: $C_1=(1,1)$

Group2: $C_2=(2,1)$



⇒ Sau là lặp thứ 3, không có sự thay đổi thành viên của các nhóm nên thuật toán dừng. Kết quả phân nhóm như sau:

Medicine	*Feature 1 (X)*: weight index	*Feature 2 (Y)*: pH	*Group*
A	1	1	1
B	2	1	1
C	4	3	2
D	5	4	2

3.1.4. *Nhận xét*

- Thuật toán k-means **không đảm bảo hội tụ về mức tối ưu toàn cục** và thường kết thúc ở mức tối ưu cục bộ.
- Kết quả có thể **phụ thuộc vào việc lựa chọn** ngẫu nhiên ban đầu các **trung tâm cụm**. Do đó, để đạt được kết quả tốt trong thực tế, người ta thường chạy thuật toán k-means nhiều lần với các tâm cụm ban đầu khác nhau.
- Độ phức tạp về thời gian của thuật toán k-means là **O(nkt)**, trong đó n là tổng số đối tượng, k là số cụm và t là số lần lặp. Thông thường, k << n và t << n. Do đó, phương pháp này có khả năng mở rộng tương đối và hiệu quả trong việc xử lý các tập dữ liệu lớn.
- Thuật toán k-means chỉ có thể được áp dụng khi **giá trị trung bình của một tập hợp các đối tượng được xác định**. Điều này có thể không xảy ra trong một số ứng dụng, chẳng hạn như khi liên quan đến dữ liệu có thuộc tính danh nghĩa.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.4. *Nhận xét*

- Việc người dùng **phải chỉ định trước số lượng cụm k**, có thể được coi là một bất lợi. Tuy nhiên, đã có những nghiên cứu về cách khắc phục khó khăn này, chẳng hạn như bằng cách cung cấp một phạm vi gần đúng của các giá trị k và sau đó sử dụng kỹ thuật phân tích để xác định k tốt nhất bằng cách so sánh kết quả phân cụm thu được cho các giá trị k khác nhau.
- Thuật toán k-means **không phù hợp** để khám phá các cụm có **hình dạng không lồi** (*nonconvex shapes*) hoặc các cụm có kích thước rất khác nhau.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.4. *Nhận xét*

- Thuật toán k-means rất nhạy cảm với các điểm dữ liệu nhiễu (*noise*) và ngoại lệ (*outlier*) vì các đối tượng như vậy nằm cách xa phần lớn dữ liệu và do đó, khi được gán cho một cụm, chúng có thể làm sai lệch đáng kể giá trị trung bình của cụm. Ví dụ: Xét sáu điểm trong không gian 1-D có các giá trị lần lượt là 1, 2, 3, 8, 9, 10 và 25.
  - Bằng cách kiểm tra trực quan, có thể tưởng tượng các điểm được phân chia thành các cụm {1, 2, 3} và {8, 9, 10}, trong đó điểm 25 bị loại trừ vì nó có vẻ là một điểm ngoại lệ. Do đó tổng sai số bình phương (E) trong trường hợp này sẽ là:

$$(1-2)^2+(2-2)^2+(3-2)^2+(8-13)^2+(9-13)^2+(10-13)^2+(25-13)^2=196$$

  - Nếu áp dụng *k-means* với k = 2 và công thức 1, sẽ tạo thành 2 phân vùng {{1, 2, 3, 8},{9, 10, 25}}. Lưu ý rằng lúc này, do có thêm giá trị 25 ở trong cụm đi sau nên giá trị 8 đã được tách khỏi cụm cũ và được đưa vào cụm mới. Do đó ***3,5*** là giá trị trung bình của cụm {1, 2, 3, 8} và **14,67** là giá trị trung bình của cụm {9, 10, 25}. Vì vậy, *k-means* sẽ tính tổng sai số bình phương (E) là:

$$(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(8-3.5)^2+(9-14.67)^2+(10-14.67)^2+(25-14.67)^2=189.67$$

  - Nhận xét: với giá trị của tâm cụm thứ hai là 14,67, về cơ bản là rất xa so với tất cả các thành viên trong cụm.

3.1. K-means: Kỹ thuật phân vùng dựa trên trung tâm

3.1.5. *Một số biến thể của Thuật toán k-means*

- Có một số biến thể của Thuật toán k-means. Chúng có thể khác nhau ở việc lựa chọn k-means ban đầu, tính toán độ khác nhau và chiến lược tính toán cluster.
- Một số cách tiếp cận để có thể làm cho thuật toán k-means có khả năng mở rộng, hiệu quả hơn là:
  i. *Cách tiếp cận 1*: sử dụng một tập hợp mẫu có kích thước phù hợp để phân cụm.
  ii. *Cách tiếp cận 2*: sử dụng phương pháp lọc sử dụng chỉ mục dữ liệu phân cấp theo không gian để tiết kiệm chi phí khi tính toán.
  iii. *Cách tiếp cận 3*: khám phá ý tưởng phân cụm vi mô, trong đó đầu tiên nhóm các đối tượng lân cận thành "cụm vi mô" (*"micro clusters"*) và sau đó thực hiện phân cụm k-means trên các cụm vi mô.

3.2. k-Medoids: kỹ thuật dựa trên đối tượng tiêu biểu

- Để giảm độ nhạy cảm đối với các giá trị ngoại lệ như trong thuật toán k-mean: Thay vì lấy giá trị trung bình của các đối tượng trong một cụm làm điểm tham chiếu, có thể chọn các đối tượng thực tế để đại diện cho các cụm, sử dụng một đối tượng đại diện cho mỗi cụm. Mỗi đối tượng còn lại được gán vào cụm mà đối tượng đại diện giống nhất.
- Phương pháp phân vùng sau đó được thực hiện dựa trên nguyên tắc giảm thiểu tổng các khác biệt giữa từng đối tượng p và đối tượng đại diện tương ứng của nó. Tiêu chí sai số tuyệt đối được định nghĩa là

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{o}_i)^2$$

trong đó

- $\boldsymbol{E}$ là tổng sai số tuyệt đối của tất cả các đối tượng p trong tập dữ liệu
- $\boldsymbol{o_i}$ là đối tượng đại diện của $\boldsymbol{C_i}$.
- Khi $\boldsymbol{k = 1}$, có thể tìm được trung vị chính xác trong thời gian $\boldsymbol{O(n^2)}$. Tuy nhiên, khi $\boldsymbol{k}$ là số dương tổng quát thì bài toán *k-medoid* là *NP-hard*

3.2. k-Medoids: kỹ thuật dựa trên đối tượng tiêu biểu

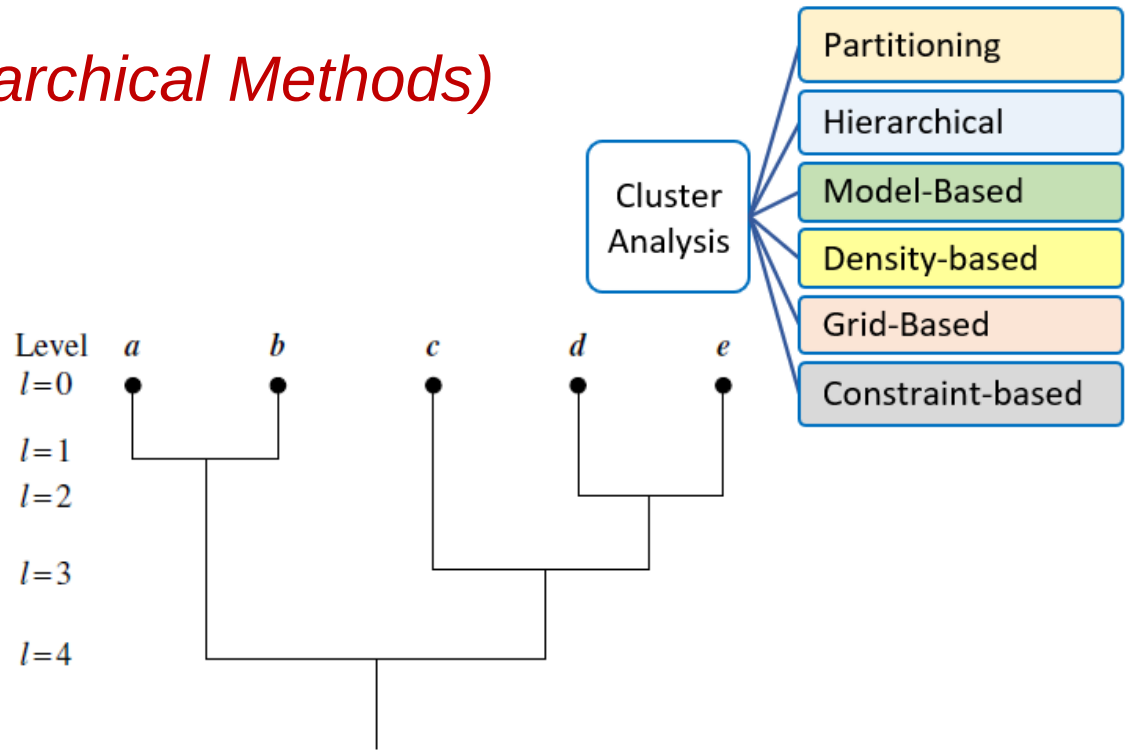
- *So sánh giữa k-means và k-medoids*:
  - Cả hai phương pháp đều yêu cầu người dùng chỉ định *k*, số lượng cụm.
  - Phương pháp *k-medoids* mạnh hơn *k-means* khi có nhiễu và giá trị ngoại lệ vì giá trị đại diện ít bị ảnh hưởng bởi giá trị ngoại lệ (*outliers values*).
  - Độ phức tạp của mỗi lần lặp trong thuật toán *k-medoids* là $O(k(n-k)^2)$. Đối với các giá trị lớn của *n* và *k*, việc tính toán như vậy trở nên rất tốn kém và tốn kém hơn nhiều so với phương pháp *k-mean*.
  - Phương pháp *k-modes* là một biến thể của *k-means,* mở rộng mô hình *k-means* thành cụm dữ liệu danh nghĩa bằng cách thay thế *means* của cụm bằng các ***modes***. Nó sử dụng các biện pháp phân biệt mới để xử lý các đối tượng danh nghĩa và phương pháp dựa trên tần suất để cập nhật các *modes* của cụm. Phương pháp *k-means* và *k-modes* có thể được tích hợp vào cụm dữ liệu với các giá trị số và danh nghĩa hỗn hợp.

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

4. CÁC PHƯƠNG PHÁP PHÂN CẤP *(Hierarchical Methods)*

- Phương pháp phân cụm theo cấp bậc hoạt động bằng cách nhóm các đối tượng dữ liệu thành một hệ thống phân cấp hoặc "cây" của các cụm.
- Việc biểu diễn các đối tượng dữ liệu dưới dạng phân cấp rất hữu ích cho việc tóm tắt và trực quan hóa dữ liệu.



Ví dụ: với tư cách là người quản lý nhân sự tại Công ty, bạn có thể tổ chức nhân viên của mình thành các nhóm chính như giám đốc điều hành, người quản lý và nhân viên. Bạn có thể phân chia thêm các nhóm này thành các nhóm nhỏ hơn. Ví dụ, nhóm nhân viên nói chung có thể được chia thành các nhóm nhỏ gồm các nhân viên cao cấp, nhân viên và người tập sự. Tất cả các nhóm này tạo thành một hệ thống phân cấp. Có thể dễ dàng tóm tắt hoặc mô tả đặc điểm của dữ liệu được tổ chức thành một hệ thống phân cấp, có thể được sử dụng để tìm mức lương trung bình của người quản lý và nhân viên.

4. Các phương pháp phân cấp *(Hierarchical Methods)*

- **Một số ứng dụng của phương pháp phân cụm theo cấp bậc**:
  - *Trong nhận dạng ký tự viết tay*: tập hợp các mẫu chữ viết tay trước tiên có thể được phân chia thành các nhóm chung trong đó mỗi nhóm tương ứng với một ký tự duy nhất. Một số nhóm có thể được chia thành các nhóm con vì một ký tự có thể được viết theo nhiều cách khác nhau. Nếu cần thiết, việc phân vùng theo cấp bậc có thể được tiếp tục đệ quy cho đến khi đạt được mức độ chi tiết mong muốn.
  - *Trong nghiên cứu về quá trình tiến hóa*: việc phân cụm theo thứ bậc có thể nhóm các loài động vật theo đặc điểm sinh học của chúng để khám phá các con đường tiến hóa, đó là hệ thống phân cấp của các loài.
  - *Trong nghiên cứu các trò chơi mang tính chiến lược* (ví dụ: cờ vua - chess - hoặc cờ đam - checkers) theo cách phân cấp có thể giúp phát triển các chiến lược trò chơi có thể được sử dụng để đào tạo người chơi.
  - …

4.1. Phân cụm phân cấp kết tụ và phân chia *(Agglomerative versus Divisive Hierarchical Clustering)*

4.1.1. Phân loại

- *Phương pháp phân cụm theo cấp bậc kết tụ* (*agglomerative hierarchical clustering method*)
  - Sử dụng chiến lược từ dưới lên. Thường bắt đầu bằng cách cho phép mỗi đối tượng tạo thành cụm riêng của nó và lặp đi lặp lại việc hợp nhất các cụm thành các cụm lớn hơn cho đến khi tất cả các đối tượng nằm trong một cụm duy nhất hoặc các điều kiện kết thúc nhất định được thỏa mãn.
  - Cụm đơn cuối cùng sẽ được xem là gốc của hệ thống phân cấp.
  - Đối với bước hợp nhất, nó tìm hai cụm gần nhau nhất (theo một số thước đo tương tự) và kết hợp cả hai để tạo thành một cụm.

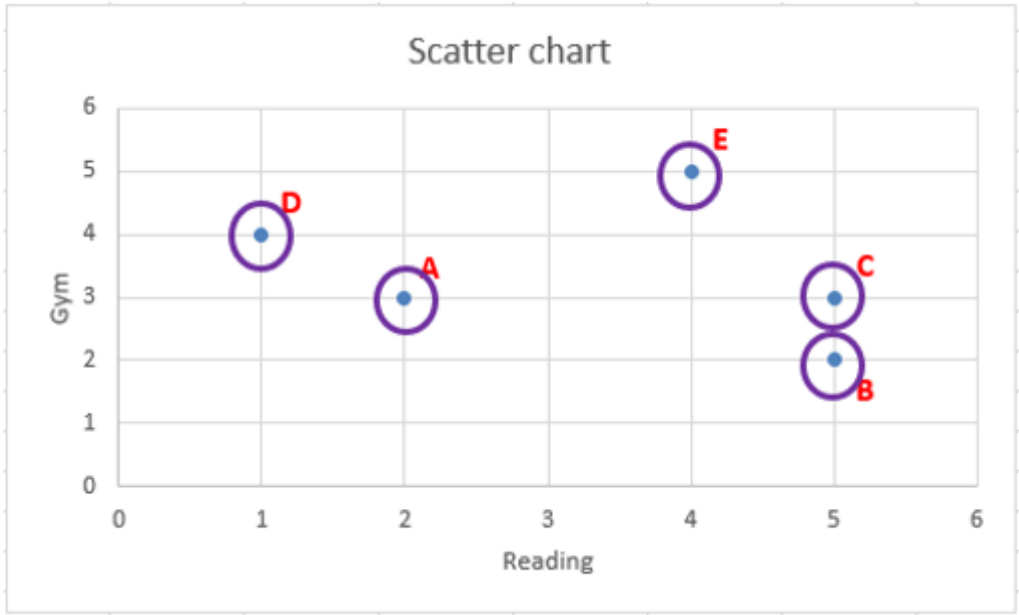
4.1. Phân cụm phân cấp kết tụ và phân chia

- *Phương pháp phân cụm theo phân chia cấp bậc* (*divisive hierarchical clustering method*)
  - Sử dụng chiến lược từ trên xuống.
  - Bắt đầu bằng cách đặt tất cả các đối tượng vào một cụm, cụm này là gốc của hệ thống phân cấp. Sau đó, nó chia cụm gốc thành nhiều cụm con nhỏ hơn và phân vùng đệ quy các cụm đó thành các cụm nhỏ hơn.
  - Quá trình phân chia tiếp tục cho đến khi mỗi cụm ở mức thấp nhất đủ mạch lạc hoặc chỉ chứa một đối tượng hoặc các đối tượng trong một cụm đủ giống nhau.

4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ (agglomerative hierarchical clustering method)

Cho số liệu về việc sử dụng thời gian thư giãn của 5 người qua số giờ đọc sách và số giờ tập gym như sau:

Name	*Reading*	*Gym*
A	2	3
B	5	2
C	5	3
D	1	4
E	4	5



- ***Bước 1***: gán mỗi điểm vào 1 cụm

4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ *(agglomerative hierarchical clustering method)*

- ***Bước 2***: tính khoảng cách giữa các điểm (sử dụng phép đo Euclidean để tính ma trận khoảng cách giữa các điểm và Single-linkage để tính liên kết giữa các cụm)
  - Sử dụng phép đo Euclidean: $d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
    - $d(A,B) = \sqrt{(2-5)^2 + (3-2)^2} = \sqrt{10} = 3.16$
    - $d(A,C) = \sqrt{(2-5)^2 + (3-3)^2} = \sqrt{9} = 3.00$
    - $d(A,D) = \sqrt{(2-1)^2 + (3-4)^2} = \sqrt{2} = 1.41$
    - $d(A,E) = \sqrt{(2-4)^2 + (3-5)^2} = \sqrt{8} = 2.82$
    - $d(B,C) = \sqrt{(5-5)^2 + (2-3)^2} = \sqrt{1} = 1.00$
    - $d(B,D) = \sqrt{(5-1)^2 + (2-4)^2} = \sqrt{20} = 4.47$
    - $d(B,E) = \sqrt{(5-4)^2 + (2-5)^2} = \sqrt{13} = 3.61$
    - $d(C,D) = \sqrt{(5-1)^2 + (3-4)^2} = \sqrt{17} = 4.12$
    - $d(C,E) = \sqrt{(5-4)^2 + (3-5)^2} = \sqrt{5} = 2.24$
    - $d(D,E) = \sqrt{(1-4)^2 + (4-5)^2} = \sqrt{10} = 3.16$
  - Lập ma trận khoảng cách giữa các điểm

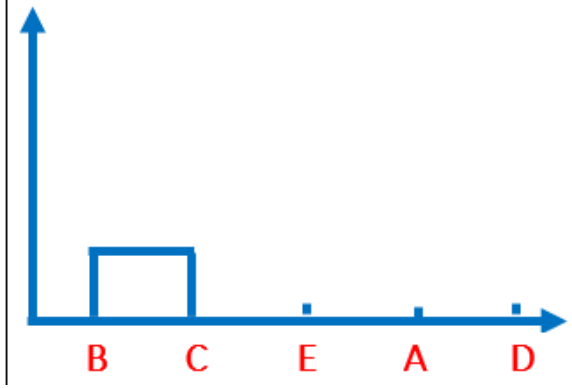
Name	*Reading*	*Gym*
A	2	3
B	5	2
C	5	3
D	1	4
E	4	5

	A	B	C	D	E
A	0				
B	3.16	0			
C	3.00	1.00	0		
D	1.41	4.47	4.12	0	
E	2.82	3.61	2.24	3.16	0

4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ *(agglomerative hierarchical clustering method)*

- ***Bước 3***: tính khoảng cách giữa các điểm (sử dụng phép đo *Euclidean* để tính ma trận khoảng cách giữa các điểm và Single-linkage để tính liên kết giữa các cụm)
  - *Lần 1*: gom các cụm có khoảng cách gần nhau và vẽ liên kết đầu tiên trong *Dendrogram*
    - Xác định giá trị nhỏ nhất, từ đó gom 2 điểm tương ứng vào 1 cụm
    - Gom các cụm có khoảng cách gần nhau và vẽ liên kết đầu tiên trong Dendrogram

	A	B	C	D	E
A	0				
B	3.16	0			
C	3.00	*1.00*	0		
D	1.41	4.47	4.12	0	
E	2.82	3.61	2.24	3.16	0



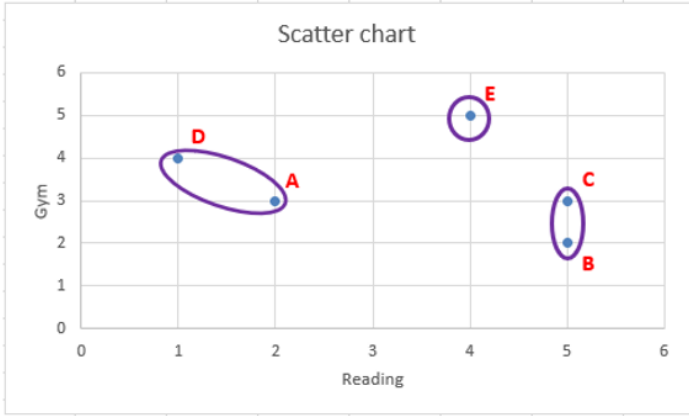
4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ *(agglomerative hierarchical clustering method)*

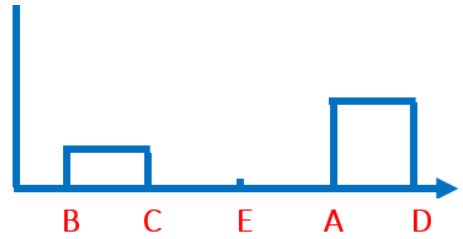
- ***Bước 3***:
  - *Lần 2*: Sau khi gộp B và C, ma trận khoảng cách có sự thay đổi, ở đây sẽ sử dụng phương pháp *Single-linkage* (khoảng cách giữa hai cụm được xác định bởi khoảng cách nhỏ nhất giữa 2 phần tử trong mỗi cụm). Do C có khoảng cách gần nhất từ cụm BC đến các cụm A, D và E, vì vậy, ta cần tính khoảng cách C-A, C-D, C-E.

    Sau khi tính, do khoảng cách giữa A và D gần nhất nên gộp A và D vào 1 cụm và có liên kết thứ 2 trong biểu đồ Dendrogram

	A	B	C	D	E
A	0				
B	3.16	0			
C	3.00	*1.00*	0		
D	1.41	4.47	4.12	0	
E	2.82	3.61	2.24	3.16	0

	A	B,C	D	E
A	0			
B,C	3.00	0		
D	1.41	4.12	0	
E	2.82	2.24	3.16	0





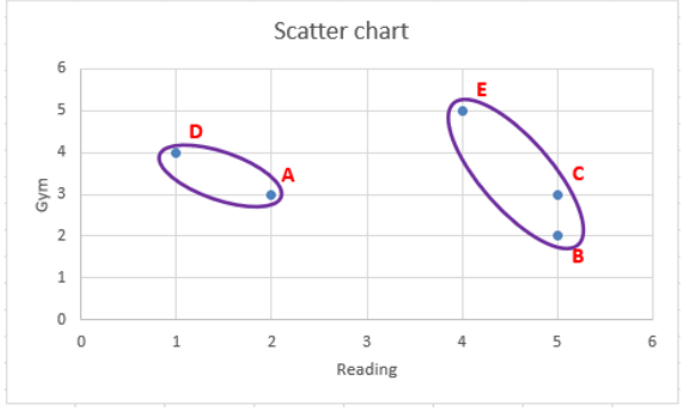
4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ *(agglomerative hierarchical clustering method)*

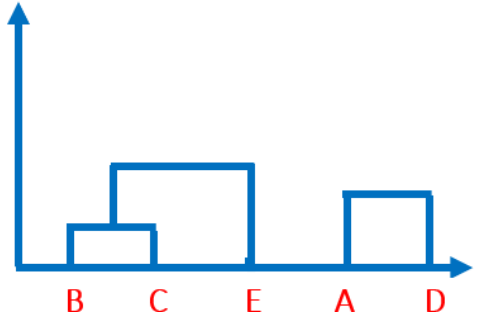
- ***Bước 3***:
  - *Lần 3*: Sau khi gộp A và D, ma trận khoảng cách có sự thay đổi, ở đây vẫn sử dụng phương pháp Single-linkage. Vì vậy, khoảng cách giữa hai cụm được xác định bởi các phần tử gần nhất với nhau. Do A có khoảng cách gần nhất từ cụm AD đến các cụm BC và E, vì vậy, ta cần tính khoảng cách A-C, A-E, C-E.

    Sau khi tính, khoảng cách giữa E và cụm BC gần nhất nên gộp E và BC vào 1 cụm và có liên kết thứ 3 trong biểu đồ Dendrogram

	A	B,C	D	E
A	0			
B,C	3.00	0		
D	1.41	4.12	0	
E	2.82	2.24	3.16	0

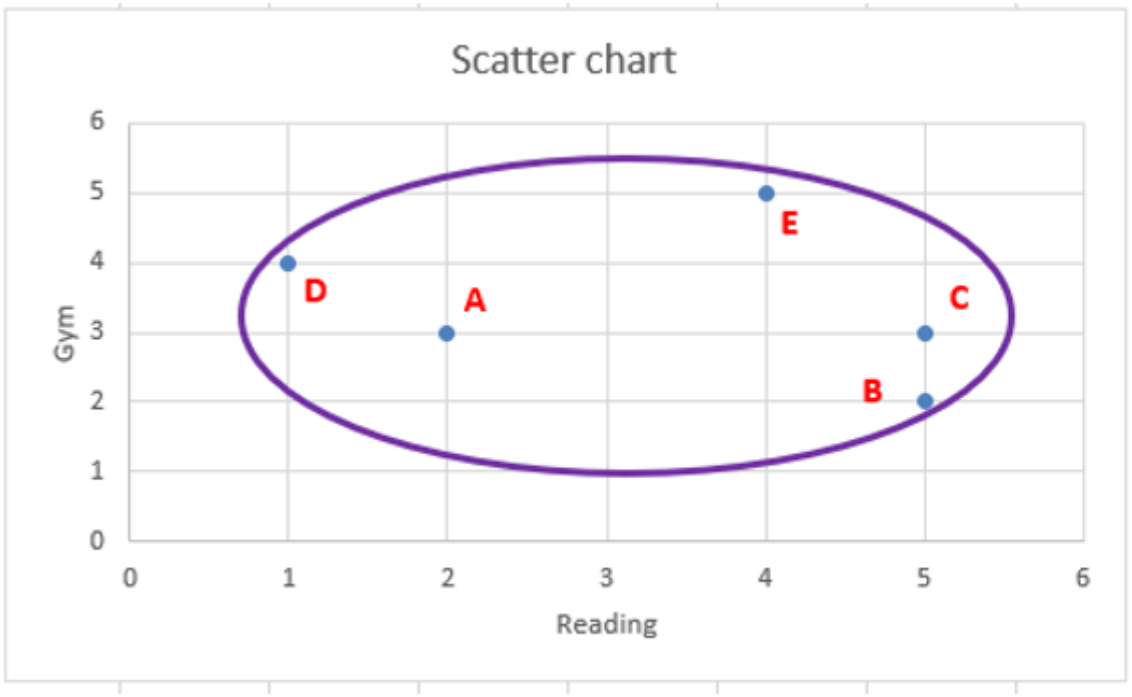
	A,D	B,C	E
A,D	0		
B,C	3.00	0	
E	2.82	2.24	0

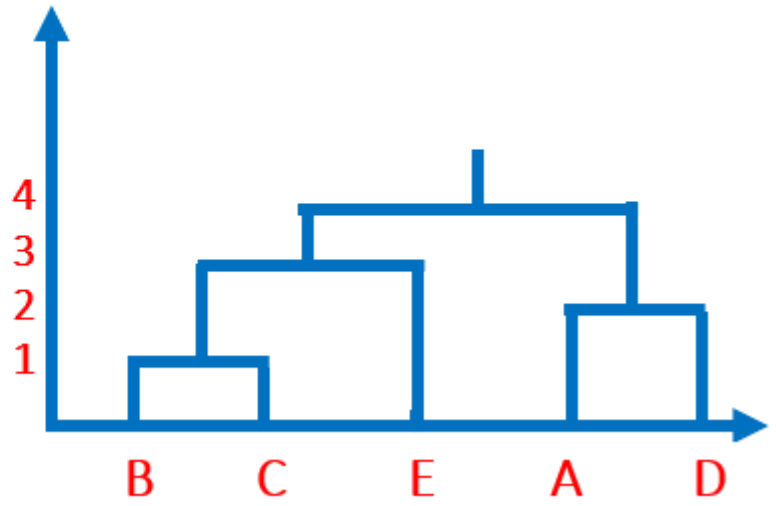




4.1.2. Ví dụ về Phương pháp phân cụm theo cấp bậc kết tụ *(agglomerative hierarchical clustering method)*

- ***Bước 3***:
  - *Lần 4*: Sau khi gộp E vào BC, còn lại hai cụm, gộp chúng lại vào 1 cụm tổng thể và có biểu đồ Dendrogram hoàn chỉnh như sau.





4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

Cho bảng toạ độ của các điểm sau trong một cụm dữ liệu tổng thể, hãy tiến hành phân chia theo cấp bậc.

Trong ví dụ này, sẽ dùng công thức tính khoảng cách Manhattan

$$(distance(p,q) = \sum_{i=1}^{n}|p_i - q_i|)$$

để tính ma trận khoảng cách giữa các điểm

– *Bước 1*: ma trận khoảng cách giữa các điểm

	A	B	C	D	E
x_1	1	3	3	4	5
x_2	6	7	5	8	8
x_3	-1	0	-2	-1	0

	A	B	C	D	E
A	0	|1-3|+|6-7|+|-1-0|	|1-3|+|6-5|+|-1-(-2)|	|1-4|+|6-8|+|-1-(-1)|	|1-5|+|6-8|+|-1-0|
B		0	|3-3|+|7-5|+|0-(-2)|	|3-4|+|7-8|+|0-(-1)|	|3-5|+|7-8|+|0-0|
C			0	|3-4|+|5-8|+|-2-(-1)|	|3-5|+|5-8|+|-2-0|
D				0	|4-5|+|8-8|+|-1-0|
E					0

	A	B	C	D	E
A	0	4	4	5	7
B	4	0	4	3	3
C	4	4	0	5	7
D	5	3	5	0	2
E	7	3	7	2	0

4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

Như vậy, cụm tổng thể có 5 điểm (A, B, C, D, E)

– *Bước 2*:

- *B2.1 lần 1*: Chọn điểm có khoảng cách trung bình lớn nhất (so với các phần tử trong cùng cụm) để tách khỏi cụm
  - *Tính khoảng cách trung bình giữa 1 điểm với các điểm còn lại trong cùng cụm*:

	A	B	C	D	E
A	0	4	4	5	7
B	4	0	4	3	3
C	4	4	0	5	7
D	5	3	5	0	2
E	7	3	7	2	0

Với điểm	*Khoảng cách trung bình giữa các điểm trong cùng cụm*
A	$\frac{d(A,B)+d(A,C)+d(A,D)+d(A,E)}{4}=\frac{4+4+5+7}{4}=5$
B	$\frac{d(B,A)+d(B,C)+d(B,D)+d(B,E)}{4}=\frac{4+4+3+3}{4}=2.8$
C	$\frac{d(C,A)+d(C,B)+d(C,D)+d(C,E)}{4}=\frac{4+4+5+7}{4}=5$
D	$\frac{d(D,A)+d(D,B)+d(D,C)+d(D,E)}{4}=\frac{5+3+5+2}{4}=3.75$
E	$\frac{d(E,A)+d(E,B)+d(E,C)+d(E,D)}{4}=\frac{7+3+7+2}{4}=4.75$

  - *Tách 1 điểm ra khỏi cụm*: Vì điểm A và C có khoảng cách trung bình đến các điểm khác cùng lớn nhất (giống nhau) nên có thể chọn 1 trong 2 điểm để tách thành 1 cụm mới. Giả sử điểm A được chọn để tách ra cụm mới, như vậy có hai cụm mới là cụm I(A) và II(B, C, D, E)

4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

– *Bước 2*:

- *B2.1 lần 1*:
  - *Tách điểm "gần" với cụm vừa tách (cụm I) để đưa vào cụm vừa có:* Tính khoảng cách trung bình giữa các điểm trong tập II để xem điểm nào gần tập I (chứa A) hơn. Điểm được chọn sẽ là điểm có khoảng cách dương (>0):

	A	B	C	D	E
A	0	4	4	5	7
B	4	0	4	3	3
C	4	4	0	5	7
D	5	3	5	0	2
E	7	3	7	2	0

Với điểm	*Khoảng cách trung bình giữa các điểm với cụm I*
B	$\frac{d(B,C)+d(B,D)+d(B,E)}{3} - d(B,A) = \frac{4+3+3}{3} = \frac{10}{3} - 4 = -\frac{2}{3}$
C	$\frac{d(C,B)+d(C,D)+d(C,E)}{3} - d(C,A) = \frac{4+5+7}{3} = \frac{16}{3} - 4 = \frac{4}{3}$
D	$\frac{d(D,B)+d(D,C)+d(D,E)}{3} - d(D,A) = \frac{3+5+2}{3} = \frac{10}{3} - 5 = -\frac{5}{3}$
E	$\frac{d(E,B)+d(E,C)+d(E,D)}{3} - d(E,A) = \frac{3+7+2}{3} = \frac{12}{3} - 7 = -3$

$\Rightarrow$ Chỉ có điểm C có khoảng cách lớn hơn 0 nên gộp điểm C vào tập I với điểm A và ta có tập I(A, C) và tập II(B, D, E)

4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

- *Bước 3*:
  - *B2.1 lần 2*:
    - Tính khoảng cách trung bình giữa các điểm trong tập II để xem điểm nào gần tập I hơn. Nếu điểm nào có khoảng cách này là số dương sẽ được tách:

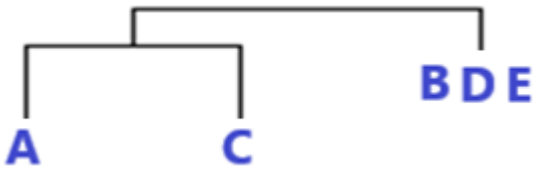
	A	B	C	D	E
A	0	4	4	5	7
B	4	0	4	3	3
C	4	4	0	5	7
D	5	3	5	0	2
E	7	3	7	2	0

Với điểm	*Khoảng cách trung bình giữa các điểm trong cụm II với cụm I*
B	$\frac{d(B,D)+d(B,E)}{2}-\frac{d(B,A)+d(B,C)}{2}=\frac{3+3}{2}-\frac{4+4}{2}=\frac{2}{2}=-1$
D	$\frac{d(D,B)+d(D,E)}{2}-\frac{d(D,A)+d(D,C)}{2}=\frac{3+2}{2}-\frac{5+5}{2}=\frac{-5}{2}=-2.5$
E	$\frac{d(E,B)+d(E,D)}{2}-\frac{d(E,A)+d(E,C)}{2}=\frac{3+2}{2}-\frac{7+7}{2}=\frac{-9}{2}=-4.5$

$\Rightarrow$ Do các khoảng cách đều âm tức là các điểm trong tập II gần tập II hơn tập I nên dừng quá trình phân chia và giữ nguyên 2 tập I(A, C) và tập II(B, D, E).

4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

- *Bước 2*:
  - *B2.1 lần 3*:
    - Chọn cụm có khoảng cách giữa hai điểm bất kỳ trong cụm là lớn nhất để tách cụm này:
      - Cụm I. d(A,C) = 4
      - Cụm II: max[d(B,D); d(B,E); d(D,E)] = max(3, 3, 2) = 3
    - Tách cụm I(A,C): do tập I có khoảng cách lớn nhất nên tách cụm I(A,C) thành 2 cụm mới là III(A) và IV(C). Việc tách cụm I đã hoàn tất do mỗi cụm vừa tách chỉ còn 1 phần tử.



4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

- *Bước 2*:
  - *B2.1 lần 3*:
    - Tách cụm II(B,D,E)
      - Tương tự như trên, đối với mỗi điểm trong cụm này, cần tìm điểm có khoảng cách trung bình lớn nhất đến các điểm còn

	A	B	C	D	E
A	0	4	4	5	7
B	4	0	4	3	3
C	4	4	0	5	7
D	5	3	5	0	2
E	7	3	7	2	0

Với điểm	*Khoảng cách trung bình giữa các điểm trong cùng cụm II*
B	$\frac{d(B,D)+d(B,E)}{2}=\frac{3+3}{2}=3$
D	$\frac{d(D,B)+d(D,E)}{2}=\frac{3+2}{2}=2.5$
E	$\frac{d(E,B)+d(E,D)}{2}=\frac{3+2}{2}=2.5$

⇒ Do B có khoảng cách trung bình lớn nhất nên tách B khỏi cụm II. Lúc này cụm II được tách thành cụm V(B) và VI(D,E).

Đến đây, cần xem xét xem có điểm nào trong tập VI(D,E) có thể di chuyển vào cụm V(B) nữa không:

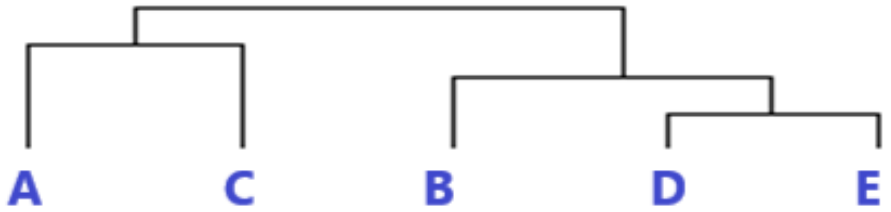
Với điểm	*Khoảng cách trung bình giữa các điểm trong cụm VI với cụm V*
D	$d(D,E)-d(D,B)=2-3=-1$
E	$d(E,D)-d(E,B)=2-3=-1$

⇒ Do các khoảng cách đều âm nên không thực hiện tách cụm VI(D,E)

4.1.3. Ví dụ về Phương pháp phân cụm theo phân chia cấp bậc (divisive hierarchical clustering method)

- *Bước 2*:
  - *B2.1 lần 4*:
    - Chọn cụm có khoảng cách giữa hai điểm bất kỳ trong cụm là lớn nhất để tách cụm này:
      - Cụm III: d(A) chỉ còn 1 phần tử nên không cần xét
      - Cụm IV: d(C) chỉ còn 1 phần tử nên không cần xét
      - Cụm V: d(C) chỉ còn 1 phần tử nên không cần xét
      - Cụm VI: max[d(D,E)] = max(2) = 2
    - Tách cụm VI(D,E): do tập VI có khoảng cách lớn nhất nên tách cụm này thành 2 cụm mới là VII(D) và VIII(E).

Việc tách cụm đã hoàn tất do mỗi cụm vừa tách chỉ còn 1 phần tử. Kết quả gồm các cụm nhỏ sau {A}, {C}, {B}, {D} và {E}. Biểu đồ Dendrogram hoàn chỉnh sau:



4.2. Đo khoảng cách trong phương pháp phân cấp

- Bốn thước đo khoảng cách (hay đo lường sự liên kết) giữa các cụm.

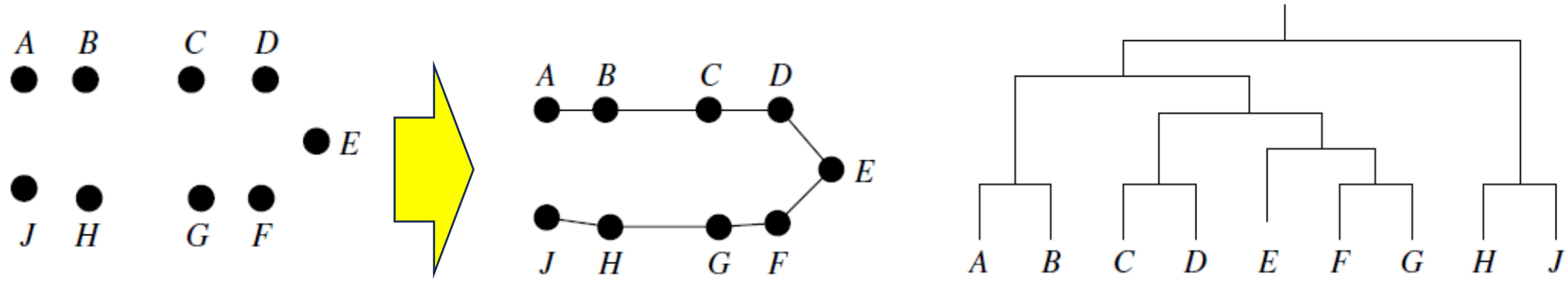
  - **Minimum distance**: $dist_{min} = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$
  - **Maximum distance**: $dist_{max} = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$
  - **Mean distance**: $dist_{mean} = |m_i - m_j|$
  - **Average distance**: $dist_{avg} = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} [p - p'|$

trong đó

- $|p-p'|$ là khoảng cách giữa hai đối tượng hoặc điểm p và p';
- m_i là giá trị trung bình của cụm C_i;
- và n_i là số lượng đối tượng trong C_i.

4.2. Đo khoảng cách trong phương pháp phân cấp

- *Khi thuật toán sử dụng khoảng cách tối thiểu* $d_{min}(C_i, C_j)$ để đo khoảng cách giữa các cụm:
  - Thuật toán còn được gọi là thuật toán phân cụm lân cận gần nhất (*nearest-neighbor clustering algorithm*). Riêng thuật toán phân cụm theo cấp bậc kết tụ còn được gọi là thuật toán cây bao trùm tối thiểu (*minimal spanning tree algorithm*), trong đó cây bao trùm của đồ thị là cây kết nối tất cả các đỉnh và cây bao trùm tối thiểu là cây có tổng trọng số là nhỏ nhất.
  - Nếu quá trình phân cụm bị chấm dứt khi khoảng cách giữa các cụm gần nhất vượt quá ngưỡng do người dùng xác định thì nó được gọi là thuật toán liên kết đơn (*single-linkage algorithm*).

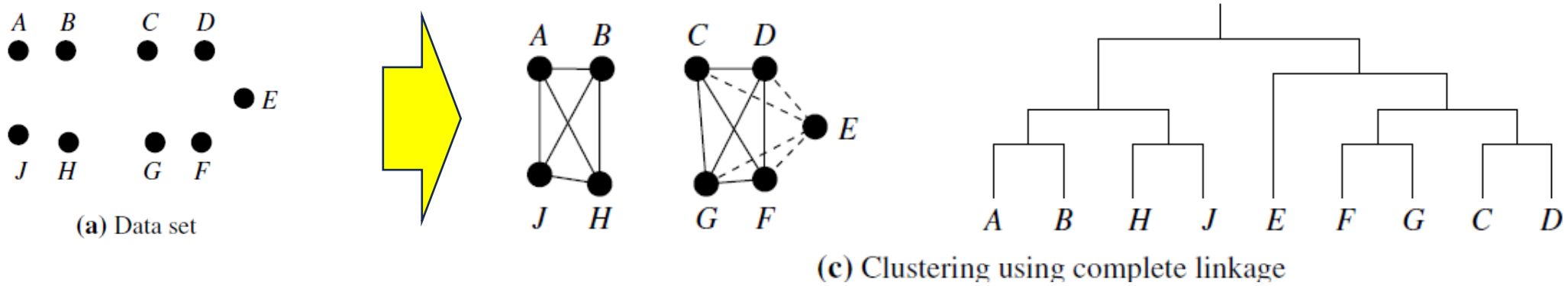


(a) Data set

(b) Clustering using single linkage

4.2. Đo khoảng cách trong phương pháp phân cấp

- *Khi một thuật toán sử dụng khoảng cách tối đa* dmax(Ci,Cj), để đo khoảng cách giữa các cụm:
  - Thuật toán còn được gọi là thuật toán phân cụm lân cận xa nhất (*farthest-neighbor clustering algorithm*).
  - Nếu quá trình phân cụm bị chấm dứt khi khoảng cách tối đa giữa các cụm gần nhất vượt quá ngưỡng do người dùng xác định thì nó được gọi là thuật toán liên kết hoàn chỉnh (*complete-linkage algorithm*).
  - Bằng cách xem các điểm dữ liệu dưới dạng các nút của biểu đồ, với các cạnh liên kết các nút, có thể coi mỗi cụm là một sơ đồ con hoàn chỉnh, nghĩa là với các cạnh kết nối tất cả các nút trong cụm.
  - **Khoảng cách giữa hai cụm được xác định bởi các nút ở xa nhất trong hai cụm**. Các thuật toán lân cận xa nhất có xu hướng giảm thiểu sự gia tăng đường kính của các cụm tại mỗi lần lặp.
  - Nếu các cụm thực khá nhỏ gọn và có kích thước xấp xỉ bằng nhau thì phương pháp này sẽ tạo ra các cụm có chất lượng cao. Nếu không, các cụm được tạo ra có thể vô nghĩa.



(a) Data set

(c) Clustering using complete linkage

4.3. Phân cụm theo cấp bậc nhiều pha bằng cách sử dụng cây tính năng phân cụm *(BIRCH: Multiphase Hierarchical Clustering Using Clustering Feature Trees)*

- *BIRCH* được thiết kế để phân cụm một lượng lớn dữ liệu số bằng cách tích hợp các phương pháp phân cụm:
  - Phân cụm theo cấp bậc (ở giai đoạn phân cụm vi mô ban đầu).
  - Các phương pháp phân cụm khác như phân vùng lặp (ở giai đoạn phân cụm vĩ mô sau này).
- *BIRCH* khắc phục được hai khó khăn trong các phương pháp phân cụm kết tụ:
  i. Khả năng mở rộng
  ii. Không thể hoàn tác những gì đã được thực hiện ở bước trước.
- Để thể hiện hệ thống phân cấp cụm, *BIRCH* sử dụng các khái niệm về tính năng phân cụm nhằm tóm tắt một cụm và cây tính năng phân cụm (*clustering feature tree – CF tree*). Các cấu trúc này giúp phương pháp phân cụm đạt được tốc độ và khả năng mở rộng tốt trong cơ sở dữ liệu lớn, đồng thời cũng giúp phương pháp này hiệu quả đối với việc phân cụm gia tăng và phân cụm động đến các đối tượng.

4.3. Phân cụm theo cấp bậc nhiều pha bằng cách sử dụng cây tính năng phân cụm *(BIRCH: Multiphase Hierarchical Clustering Using Clustering Feature Trees)*

- Xét một cụm gồm n các đối tượng hoặc điểm dữ liệu *d-chiều* (*d-dimensional*).
- Đặc điểm phân cụm (*clustering feature - CF*) của cụm là một vectơ 3-D tóm tắt thông tin về các cụm đối tượng được định nghĩa là:

$$CF = (n, LS, SS)$$

trong đó: + *LS* là tổng tuyến tính của *n* điểm (tức là $\sum_{i=1}^{n} x_i$)

+ *SS* là tổng bình phương của các điểm dữ liệu (tức là $\sum_{i=1}^{n} {x_i}^2$).

4.3. Phân cụm theo cấp bậc nhiều pha bằng cách sử dụng cây tính năng phân cụm *(BIRCH: Multiphase Hierarchical Clustering Using Clustering Feature Trees)*

- Tính năng phân cụm về cơ bản là bản tóm tắt số liệu thống kê cho cụm nhất định. Sử dụng tính năng phân cụm, có thể dễ dàng rút ra nhiều số liệu thống kê hữu ích về một cụm. Ví dụ: trọng tâm của cụm x_0, bán kính R và đường kính D là:

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{LS}{n}$$

$$R = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_0)^2}{n}} = \sqrt{\frac{nSS - 2LS^2 + nLS}{n^2}}$$

$$D = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2}{n(n-1)}} = \sqrt{\frac{2nSS - 2LS^2}{n(n-1)}}$$

trong đó: + R: khoảng cách trung bình từ các đối tượng thành viên đến tâm.

+ D: khoảng cách trung bình theo cặp trong một cụm.

+ Cả R và D đều phản ánh độ chặt của cụm xung quanh tâm.

4.3. Phân cụm theo cấp bậc nhiều pha bằng cách sử dụng cây tính năng phân cụm

- Đối với hai cụm rời nhau, C_1 và C_2, với các đặc điểm phân cụm lần lượt là $CF_1 = (n_1, LS_1, SS_1)$ và $CF_2 = (n_2, LS_2, SS_2)$, đặc điểm phân cụm cho cụm được hình thành bằng cách hợp nhất C_1 và C_2 chỉ đơn giản là

$$CF_1 + CF_2 = (n_1 + n_2, LS_1 + LS_2, SS_1 + SS_2)$$

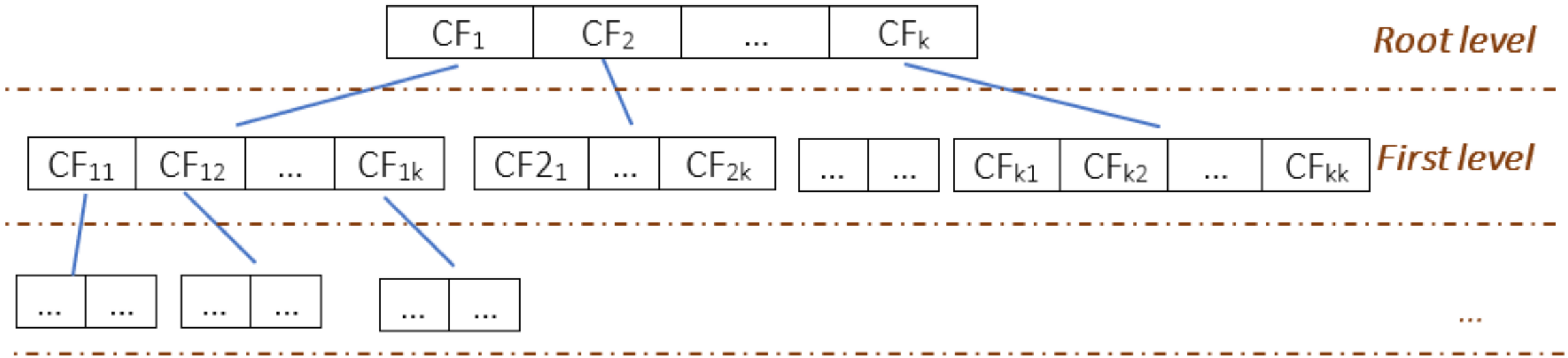
- Ví dụ: Giả sử có ba điểm (2, 5), (3, 2) và (4, 3) trong một cụm C_1. Tính năng phân cụm của C_1 là:

$$CF_1 = (3, (2+3+4, 5+2+3), (2^2 + 3^2 + 4^2, 5^2 + 2^2 + 3^2)) = (3, (9,10), (29,38))$$

- Giả sử rằng C_1 tách rời khỏi cụm thứ hai (C_2), trong đó $CF_2 = (3, (35,36), (417,440))$. Tính năng phân cụm của cụm mới (C_3) được hình thành bằng cách hợp nhất C_1 và C_2, được rút ra bằng cách thêm CF_1 và CF_2. Đó là,

$$CF_3 = (3 + 3, (9 + 35, 10 + 36), (29 + 417, 38 + 440)) = (6, (44, 46), (446, 478))$$

4.3. Phân cụm theo cấp bậc nhiều pha bằng cách sử dụng cây tính năng phân cụm

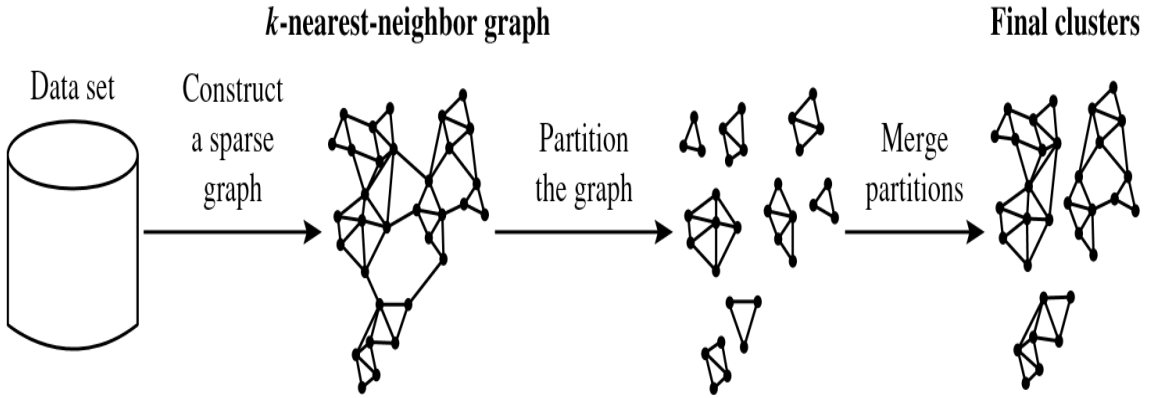


Hình 5-9 Cấu trúc CF-tree

4.4. Phân cụm phân cấp nhiều pha bằng mô hình động *(Chameleon: Multiphase Hierarchical Clustering Using Dynamic Modeling)*

- *Chameleon* là một thuật toán phân cụm theo cấp bậc sử dụng mô hình động để xác định sự giống nhau giữa các cặp cụm.
- Trong *Chameleon*, độ tương tự của cụm được đánh giá dựa trên:
  - (i) Mức độ kết nối của các đối tượng trong một cụm.
  - (ii) Mức độ gần nhau của các cụm. Nghĩa là, hai cụm được hợp nhất nếu khả năng kết nối của chúng cao và chúng ở gần nhau.
- Do đó, *Chameleon* không phụ thuộc vào mô hình tĩnh do người dùng cung cấp và có thể tự động thích ứng với các đặc điểm bên trong của các cụm được hợp nhất. Quá trình hợp nhất tạo điều kiện thuận lợi cho việc khám phá các cụm tự nhiên và đồng nhất và áp dụng cho tất cả các loại dữ liệu miễn là có thể chỉ định chức năng tương tự.

4.4. Phân cụm phân cấp nhiều pha bằng mô hình động



- Cách thức hoạt động của *Chameleon*.
  - *Chameleon* sử dụng cách tiếp cận đồ thị k-láng giềng gần nhất để xây dựng một đồ thị thưa thớt, trong đó mỗi đỉnh của đồ thị đại diện cho một đối tượng dữ liệu và tồn tại một cạnh giữa hai đỉnh (đối tượng) nếu một đối tượng nằm trong số k đối tượng giống nhau nhất so với các đối tượng khác.
  - Các cạnh được tăng trọng số để phản ánh sự giống nhau giữa các đối tượng. *Chameleon* sử dụng thuật toán phân vùng đồ thị để phân vùng đồ thị thành một số lượng lớn các phân cụm tương đối nhỏ sao cho giảm thiểu việc cắt cạnh (*edge cut*). Nghĩa là, cụm C được phân chia thành các cụm con C_i và C_j để giảm thiểu trọng số của các cạnh sẽ bị cắt nếu C được chia đôi thành C_i và C_j. Nó đánh giá khả năng liên kết tuyệt đối giữa cụm C_i và C_j.
  - Sau đó, *Chameleon* sử dụng thuật toán phân cụm theo cấp bậc kết tụ để hợp nhất lặp đi lặp lại các cụm con dựa trên sự giống nhau của chúng. Để xác định các cặp của hầu hết các cụm con giống nhau, *Chameleon* phải tính đến cả khả năng kết nối và độ gần nhau của các cụm.

4.4. Phân cụm phân cấp nhiều pha bằng mô hình động

- Cách thức hoạt động của *Chameleon*.
  - Sau đó, *Chameleon* sử dụng thuật toán phân cụm theo cấp bậc kết tụ để hợp nhất lặp đi lặp lại các cụm con dựa trên sự giống nhau của chúng.
  - Để xác định các cặp của hầu hết các cụm con giống nhau, *Chameleon* phải tính đến cả khả năng kết nối và độ gần nhau của các cụm.
    - *Khả năng kết nối tương đối* (*relative interconnectivity*), $RI(C_i, C_j)$, giữa hai cụm C_i và C_j, được định nghĩa là khả năng kết nối tuyệt đối giữa C_i và C_j, được chuẩn hóa theo khả năng liên kết nội bộ của hai cụm C_i và C_j. Đó là,

$$RI(Ci,Cj) = \frac{|EC_{\{Ci,Cj\}}|}{\frac{1}{2}(|EC_{Ci}| + |EC_{Cj}|)}$$

trong đó:

- $EC_{\{Ci,Cj\}}$ là cạnh cắt như được xác định trước đó cho cụm chứa cả C_i và C_j.
- EC_{Ci} (hoặc EC_{Cj}) là tổng nhỏ nhất của các cạnh cắt để phân chia C_i (hoặc C_j) thành hai phần gần bằng nhau.

4.4. Phân cụm phân cấp nhiều pha bằng mô hình động

- Cách thức hoạt động của *Chameleon*.
  - *Độ gần tương đối* (*relative closeness*) $RC(C_i, C_j)$ giữa một cặp cụm C_i và C_j, là độ gần tuyệt đối giữa C_i và C_j, được chuẩn hóa theo độ gần bên trong của hai cụm C_i và C_j. Độ gần tương đối được định nghĩa là

$$\boldsymbol{RC(C_i, C_j)} = \frac{\bar{\boldsymbol{S}}_{\boldsymbol{EC}_{\{\boldsymbol{C_i}, \boldsymbol{C_j}\}}}}{\frac{|\boldsymbol{C_i}|}{|\boldsymbol{C_i}| + |\boldsymbol{C_j}|}\bar{\boldsymbol{S}}_{\boldsymbol{EC}_{\boldsymbol{C_i}}} + \frac{|\boldsymbol{C_j}|}{|\boldsymbol{C_i}| + |\boldsymbol{C_j}|}\bar{\boldsymbol{S}}_{\boldsymbol{EC}_{\boldsymbol{C_j}}}}$$

trong đó

- $\bar{S}_{EC_{\{C_i, C_j\}}}$ là trọng số trung bình của các cạnh nối các đỉnh trong C_i với các đỉnh trong C_j
- $\bar{S}_{EC_{C_i}}$, (hoặc $\bar{S}_{EC_{C_j}}$) là trọng số trung bình của các cạnh thuộc đường phân giác cắt nhỏ (*mincut bisector*) của cụm C_i (hoặc C_j) .

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- *Các phương pháp phân cụm theo cấp bậc thuật toán* sử dụng các biện pháp liên kết có xu hướng dễ hiểu và thường hiệu quả trong việc phân cụm. Chúng thường được sử dụng trong nhiều ứng dụng phân tích phân cụm. Tuy nhiên, các phương pháp phân cụm theo cấp bậc thuật toán có thể gặp phải một số hạn chế:

i. Việc chọn một thước đo khoảng cách tốt cho phân cụm theo cấp bậc thường không hề đơn giản.

ii. Để áp dụng một phương pháp thuật toán, các đối tượng dữ liệu không thể thiếu bất kỳ giá trị thuộc tính nào. Trong trường hợp dữ liệu được quan sát một phần (tức là thiếu một số giá trị thuộc tính của một số đối tượng), việc áp dụng phương pháp phân cụm theo cấp bậc thuật toán không thể tiến hành tính toán khoảng cách.

iii. Hầu hết đều là phương pháp phỏng đoán và ở mỗi bước tìm kiếm cục bộ một quyết định hợp nhất/tách tốt.

$\Rightarrow$ Do đó, mục tiêu tối ưu hóa của hệ thống phân cấp cụm kết quả có thể không rõ ràng.

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- *Phân cụm theo cấp bậc xác suất*
  - Nhằm mục đích khắc phục một số nhược điểm của *phân cụm theo cấp bậc thuật toán* bằng cách sử dụng các mô hình xác suất để đo khoảng cách giữa các cụm.
  - Ví dụ: khi tiến hành phân tích phân cụm trên một tập hợp các khảo sát tiếp thị, giả định rằng các khảo sát được thu thập là mẫu ý kiến của tất cả các khách hàng có thể có. Ở đây, cơ chế tạo dữ liệu là sự phân bổ xác suất của các ý kiến đối với các khách hàng khác nhau, không thể lấy được một cách trực tiếp và đầy đủ. Nhiệm vụ của việc phân cụm là ước tính mô hình tổng quát càng chính xác càng tốt bằng cách sử dụng các đối tượng dữ liệu được quan sát để phân cụm.
  - Trong thực tế, có thể giả định rằng các mô hình tạo dữ liệu áp dụng các hàm phân phối chung, chẳng hạn như phân phối *Gaussian* hoặc phân phối *Bernoulli*, được điều chỉnh bởi các tham số. Sau đó, nhiệm vụ học một mô hình tổng quát được giảm xuống còn việc tìm các giá trị tham số mà mô hình phù hợp nhất với tập dữ liệu được quan sát.

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- Ví dụ:Giả sử cho một tập hợp các điểm 1-D X, với X = {x_1, ..., x_n} để phân tích phân cụm. Giả sử rằng các điểm dữ liệu được tạo bởi thuật toán phân phối Gaussian,

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  trong đó:
  - μ: giá trị trung bình (*mean*)
  - σ^2: phương sai (*variance*)
- Xác suất để một điểm $x_i \in X$ được tạo ra bởi mô hình là:

$$P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Do đó, khả năng X được tạo ra bởi mô hình là

$$L(\mathcal{N}(\mu, \sigma^2): X) = P(X|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- Ví dụ:Giả sử cho một tập hợp các điểm 1-D X, với X = $\{x_1, ..., x_n\}$ để phân tích phân cụm. Giả sử rằng các điểm dữ liệu được tạo bởi thuật toán phân phối Gaussian,

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  trong đó:
  - μ: giá trị trung bình (*mean*)
  - σ^2: phương sai (*variance*)
- Xác suất để một điểm $x_i \in X$ được tạo ra bởi mô hình là:

$$P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Do đó, khả năng X được tạo ra bởi mô hình là

$$L(\mathcal{N}(\mu, \sigma^2): X) = P(X|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Nhiệm vụ của việc học mô hình tổng quát là tìm các tham số μ và σ^2 sao cho khả năng $L(N(\mu,\sigma^2) : X)$ là lớn nhất, tức là tìm

$$\mathcal{N}(\mu, \sigma^2) = argmax\{L(\mathcal{N}(\mu, \sigma^2): X)\}$$

  trong đó $max\{L(\mathcal{N}(\mu, \sigma^2): X)\}$ được gọi là khả năng tối đa.

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- Với một tập hợp các đối tượng, chất lượng của một cụm được hình thành bởi tất cả các đối tượng có thể được đo bằng khả năng tối đa. Đối với tập đối tượng được phân chia thành m cụm $C_1, \ldots, C_m$ thì chất lượng có thể được đo bằng

$$Q(\{C_1, \ldots, C_m\}) = \prod_{i=1}^{m} P(C_i)$$

trong đó:

- *P()* là khả năng tối đa.
- Nếu hợp nhất hai cụm C_{j1} và C_{j2} thành một cụm $C_{j1} \cup C_{j2}$ thì sự thay đổi về chất lượng của toàn bộ phân cụm là

$$Q\left(\left(\{C_1, \ldots, C_m\} - \{C_{j1}, C_{j2}\}\right) \cup \{C_{j1} \cup C_{j2}\}\right) - Q(\{C_1, \ldots, C_m\}) = \frac{\prod_{i=1}^{m} P(C_i) \,.\, P(C_{j1} \cup C_{j2})}{P(C_{j1}).P(C_{j2})} - \prod_{i=1}^{m} P(C_i)$$

$$= \prod_{i=1}^{m} P(C_i) \left(\frac{P(C_{j1} \cup C_{j2})}{P(C_{j1}).P(C_{j2})} - 1\right)$$

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- Khi chọn hợp nhất hai cụm trong phân cụm theo cấp bậc, $\prod_{i=1}^{m} P(C_i)$là không đổi đối với bất kỳ cặp cụm nào. Do đó, với cụm C_1 và C_2, khoảng cách giữa chúng có thể được đo bằng

$$dist(C_i, C_j\}) = -log\frac{P(C_1 \cup C_2)}{P(C_1).P(C_2)}$$

- Phương pháp phân cụm theo cấp bậc xác suất có thể áp dụng khung phân cụm kết tụ nhưng sử dụng các mô hình xác suất của công thức trên để đo khoảng cách giữa các cụm.
- Quan sát lại công thức

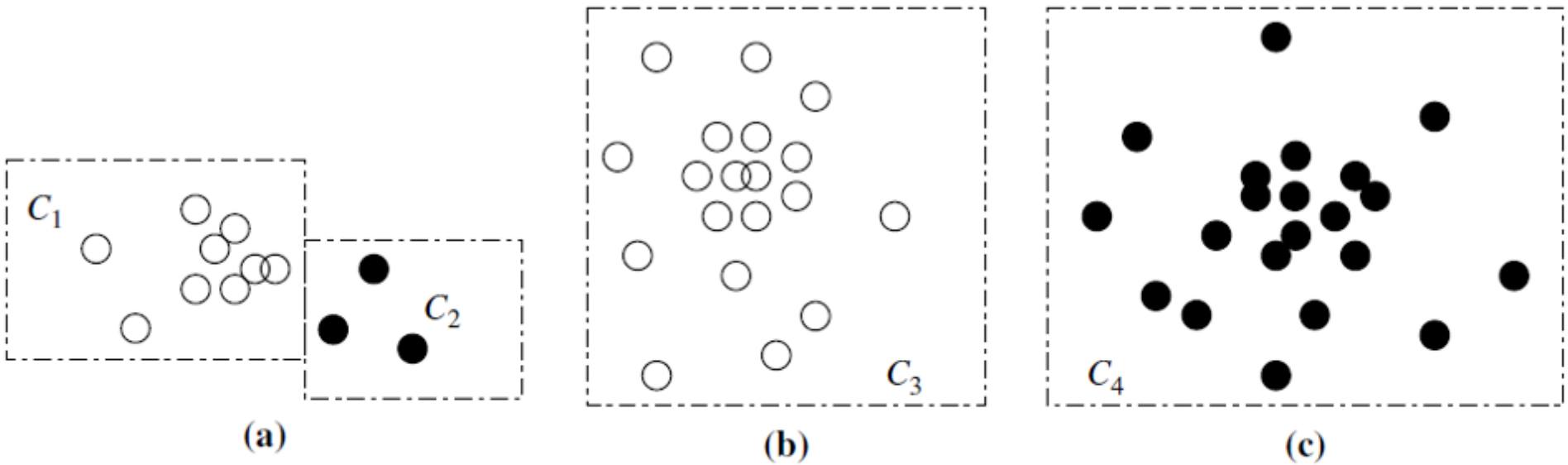
$$Q\left(\left(\{C_1, \dots, C_m\} - \{C_{j1}, C_{j2}\}\right) \cup \{C_{j1} \cup C_{j2}\}\right) - Q(\{C_1, \dots, C_m\}) = \frac{\prod_{i=1}^{m} P(C_i) \,.\, P(C_{j1} \cup C_{j2})}{P(C_{j1}).P(C_{j2})} - \prod_{i=1}^{m} P(C_i)$$

$$= \prod_{i=1}^{m} P(C_i)\left(\frac{P(C_{j1} \cup C_{j2})}{P(C_{j1}).P(C_{j2})} - 1\right)$$

ta thấy rằng việc hợp nhất hai cụm có thể không phải lúc nào cũng dẫn đến sự cải thiện về chất lượng phân cụm, nghĩa là $\frac{P(C_{j1} \cup C_{j2})}{P(C_{j1}).P(C_{j2})}$ có thể nhỏ hơn 1.

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- Ví dụ: giả sử rằng các hàm phân phối *Gaussian* là được sử dụng trong mô hình của Hình 11. Mặc dù việc hợp nhất các cụm C_1 và C_2 tạo ra một cụm phù hợp hơn với phân bố *Gaussian*, nhưng việc hợp nhất các cụm *C3* và C_4 sẽ làm giảm chất lượng phân cụm vì không có hàm *Gaussian* nào có thể phù hợp tốt với cụm được hợp nhất.



Hình 5-11 Việc hợp nhất các cụm trong phân cụm theo cấp bậc xác suất: (a) Việc hợp nhất các cụm C1 và C2 dẫn đến sự gia tăng chất lượng cụm tổng thể, nhưng việc hợp nhất các cụm (b) C3 và (c) C4 thì không.

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

- **Thuật toán**: Một thuật toán phân cụm theo cấp bậc xác suất.

Input:

$D = \{o_1, \ldots, o_n\}$: tập dữ liệu chứa *n* đối tượng;

Output: Một hệ thống phân cấp của các cụm.

Mã giả:

(1) **tạo** cụm cho từng đối tượng $C_i = \{o_i\}$, $1 \le i \le n$;

(2) **for** i = 1 to n

(3) Tìm cặp cụm C_i và C_j sao cho $C_i, C_j = argmax_{i \neq j} log \frac{P(C_i \cup C_j)}{P(C_i).P(C_j)}$

(4) if $log \frac{P(C_i \cup C_j)}{P(C_i).P(C_j)} > 0$ thì hợp nhất C_i và C_j;

(5) else stop;

4.5. Phân cụm phân cấp xác suất *(Probabilistic Hierarchical Clustering)*

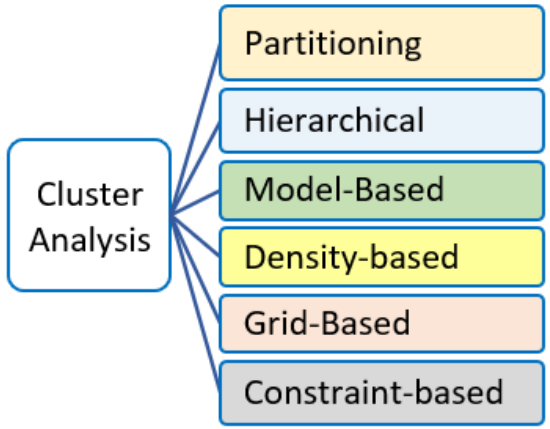
- ***So sánh giữa phân cụm theo cấp bậc xác suất và phân cụm theo cấp bậc kết tụ*:**
  - Nhìn chung cả 2 phương pháp có hiệu quả tương tự như nhau.
  - Các mô hình xác suất dễ hiểu hơn nhưng đôi khi kém linh hoạt hơn.
  - Các mô hình xác suất có thể xử lý dữ liệu được quan sát một phần. Ví dụ: với một tập dữ liệu đa chiều (*multidimensional*) trong đó một số đối tượng thiếu giá trị trên một số chiều (*dimensions*), có thể tìm hiểu mô hình *Gaussian* trên từng chiều một cách độc lập bằng cách sử dụng các giá trị được quan sát trên chiều đó.
  - Một nhược điểm của việc sử dụng phân cụm phân cấp theo xác suất là nó chỉ đưa ra một phân cấp đối với mô hình xác suất đã chọn. Nó không thể xử lý sự không chắc chắn của hệ thống phân cấp cụm.
  - Với một tập dữ liệu, có thể tồn tại nhiều hệ thống phân cấp phù hợp với dữ liệu được quan sát. Cả phương pháp tiếp cận thuật toán lẫn phương pháp xác suất đều không thể tìm thấy sự phân bố của các hệ thống phân cấp như vậy.

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

5. CÁC PHƯƠNG PHÁP DỰA TRÊN MẬT ĐỘ *(Density-Based Methods)*

- Các phương pháp phân vùng và phân cấp được thiết kế để tìm các cụm có dạng hình cầu. Hai phương pháp đó gặp khó khăn trong việc tìm kiếm các cụm có hình dạng tùy ý như hình chữ "S" và cụm hình bầu dục. Với những dữ liệu như vậy, 2 phương pháp trên có thể sẽ xác định không chính xác các vùng lồi, nơi nhiễu hoặc các phần tử ngoại lệ được đưa vào các cụm.



Các cụm có hình dạng tùy ý.

- Ngoài ra, để tìm các cụm có hình dạng tùy ý, có thể mô hình hóa các cụm dưới dạng các vùng dày đặc trong không gian dữ liệu, được phân tách bằng các vùng thưa thớt. Đây là chiến lược chính đằng sau các phương pháp phân cụm dựa trên mật độ, có thể khám phá các cụm có hình dạng không hình cầu.

5.1. Phân cụm mật độ dựa trên các khu vực được kết nối với mật độ cao

(DBSCAN: Density-Based Clustering Based on Connected Regions with High Density)

- Mật độ của một vật thể ***o*** có thể được đo bằng số lượng vật thể gần với ***o***. DBSCAN tìm các đối tượng cốt lõi, tức là các đối tượng có vùng lân cận dày đặc. Nó kết nối các đối tượng cốt lõi và các vùng lân cận của chúng để tạo thành các vùng dày đặc dưới dạng các cụm.
- Tham số do người dùng chỉ định > 0 được sử dụng để chỉ định bán kính của vùng lân cận mà *DBSCAN* sẽ xem xét cho mọi đối tượng. Lân cận của một vật ***o*** là không gian trong bán kính có tâm ở ***o***.
- Do kích thước vùng lân cận cố định được tham số hóa bằng **ε**, nên mật độ của vùng lân cận (*density of a neighborhood*) có thể được đo đơn giản bằng số lượng đối tượng trong vùng lân cận. Để xác định xem một vùng lân cận có mật độ dày đặc hay không, DBSCAN sử dụng một tham số khác do người dùng chỉ định là *MinPts*, chỉ định ngưỡng mật độ của các vùng dày đặc. Một đối tượng là đối tượng cốt lõi (*core object*) nếu vùng lân cận của đối tượng chứa ít nhất các đối tượng *MinPts*. Các đối tượng cốt lõi là trụ cột của các vùng dày đặc.

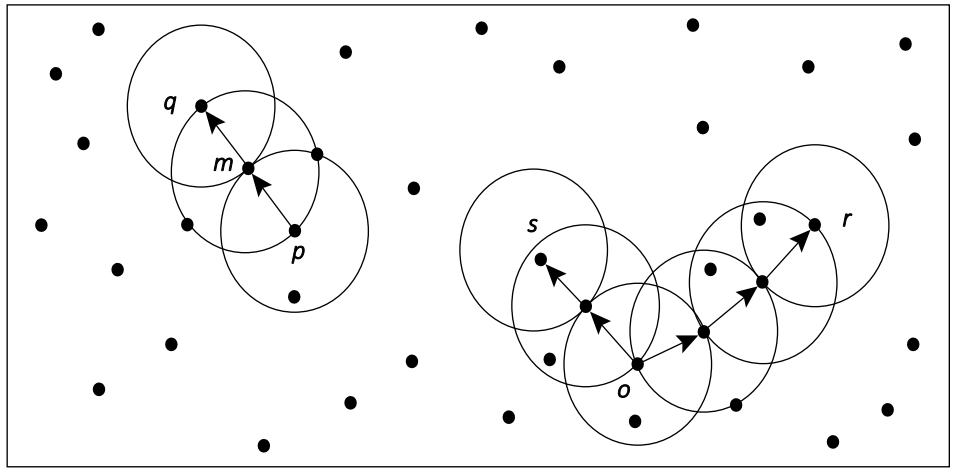
5.1. Phân cụm mật độ dựa trên các khu vực được kết nối với mật độ cao *(DBSCAN)*

- Trong DBSCAN, p có thể truy cập được theo mật độ từ q (liên quan đến ε và *MinPts* trong D có thể truy cập được từ mật độ) nếu có một chuỗi các đối tượng p_i tương ứng với từ $p_1, ..., p_n$, sao cho $p_1=q$, p_n=p và pi+1 có thể truy cập trực tiếp mật độ từ pi thông qua phép $\in$ và *MinPts* với $1 <= I <= n, p_i \in D$.
- Lưu ý rằng khả năng tiếp cận trực tiếp mật độ không phải là một mối quan hệ tương đương vì nó không đối xứng.
  - Nếu cả o_1 và o_2 đều là đối tượng cốt lõi và o_1 có thể truy cập mật độ từ o_2, thì o_2 có thể truy cập mật độ từ o_1.
  - Tuy nhiên, nếu o_2 là đối tượng cốt lõi còn o_1 thì không, thì o_1 có thể truy cập được mật độ từ o_2, nhưng không phải ngược lại.

5.1. Phân cụm mật độ dựa trên các khu vực được kết nối với mật độ cao *(DBSCAN)*

- Để kết nối các đối tượng cốt lõi cũng như các đối tượng lân cận của chúng trong một khu vực dày đặc, DBSCAN sử dụng khái niệm kết nối mật độ (*density-connectedness*). Hai đối tượng $p_1, p_2 \in D$ được liên thông mật độ liên quan đến ε và *MinPts* nếu có một đối tượng $q \in D$ sao cho cả p_1 và p_2 đều có thể truy cập được mật độ từ q đối với và *MinPts*. Không giống như khả năng tiếp cận mật độ (*density-reachability*), tính liên kết mật độ (*density connectedness*) là một mối quan hệ tương đương. Dễ dàng chỉ ra rằng, đối với các đối tượng o_1, o_2 và o_3, nếu o_1 và o_2 được liên kết mật độ, và o_2 và o_3 được liên kết mật độ, thì o_1 và o_3 cũng vậy.

5.1. Phân cụm mật độ dựa trên các khu vực được kết nối với mật độ cao *(DBSCAN)*



- Ví dụ 5-7 về Khả năng tiếp cận mật độ và khả năng liên kết mật độ. Xem Hình 14 với một giá trị đã cho được biểu thị bằng bán kính của các đường tròn, và giả sử, đặt *MinPts* = *3*.

  - Trong số các điểm được gắn nhãn, *m, p, o, r* là các đối tượng cốt lõi vì mỗi điểm nằm trong vùng lân cận chứa ít nhất ba điểm. Đối tượng *q* có thể truy cập mật độ trực tiếp từ *m*. Đối tượng *m* có thể truy cập mật độ trực tiếp từ *p* và ngược lại.
  - Đối tượng *q* có thể truy cập mật độ (gián tiếp) từ *p* vì *q* có thể truy cập mật độ trực tiếp từ *m* và *m* có thể truy cập mật độ trực tiếp từ *p*. Tuy nhiên, *p* không thể tiếp cận được mật độ từ *q* vì *q* không phải là đối tượng cốt lõi. Tương tự, *r* và *s* có thể truy cập mật độ từ *o* và *o* có thể truy cập mật độ từ *r*. Do đó *o*, *r* và *s* đều được kết nối mật độ.

5.1. Phân cụm mật độ dựa trên các khu vực được kết nối với mật độ cao *(DBSCAN)*

- ***Thuật toán***: DBSCAN: thuật toán phân cụm dựa trên mật độ.

Input:

- *D* : tập dữ liệu chứa n đối tượng,
- ε: tham số bán kính, và
- *MinPts*: ngưỡng mật độ vùng lân cận.

Output: Một tập hợp các cụm dựa trên mật độ.

Mã giả:

```
(1)  đánh dấu tất cả các đối tượng là "unvisited";
(2)  do
(3)     chọn ngẫu nhiên một đối tượng p trong nhóm "unvisited";
(4)     đánh dấu p là "visited";
(5)     if ε-neighborhood của p có ít nhất đối tượng MinPts
(6)        Tạo cụm C mới và thêm p vào C;
(7)        Gọi N là tập các đối tượng trong lân cận của p;
(8)        for mỗi điểm p' in N
(9)           if p' thuộc "unvisited"
(10)             Đánh dấu p' là "visited";
(11)             if  ε-neighborhood của p' có ít nhất điểm MinPts thì thêm
                                                  các điểm đó vào N;
(12)          if p' chưa phải là thành viên của bất kỳ cụm nào, thêm p' vào C;
(13)       end for
(14)       output C;
(15)    else đánh dấu p là tiếng ồn;
(16) until không có đối tượng nào còn gán nhãn "unvisited";
```

5.2. Thứ tự các điểm để xác định cấu trúc phân cụm *(OPTICS: Ordering Points to Identify the Clustering Structure)*

- ***Những khó khăn của DBSCAN***: Mặc dù DBSCAN có thể phân cụm các đối tượng với các tham số đầu vào như (bán kính tối đa của vùng lân cận) và MinPts (số điểm tối thiểu cần thiết trong vùng lân cận của đối tượng lõi), nhưng cũng dẫn đến những khó khăn:
  - Người dùng phải chịu trách nhiệm chọn các giá trị tham số. Các cài đặt tham số như vậy thường được đặt theo kinh nghiệm và khó xác định, đặc biệt đối với các tập dữ liệu trong thế giới thực.
  - Hầu hết các thuật toán đều nhạy cảm với các giá trị tham số này: Cài đặt hơi khác nhau có thể dẫn đến việc phân cụm dữ liệu rất khác nhau.
  - Các tập dữ liệu nhiều chiều trong thế giới thực thường có sự phân bố rất sai lệch sao cho cấu trúc phân cụm nội tại của chúng có thể không được đặc trưng rõ ràng bởi một tập hợp các tham số mật độ toàn cục.

5.2. Thứ tự các điểm để xác định cấu trúc phân cụm *(OPTICS: Ordering Points to Identify the Clustering Structure)*

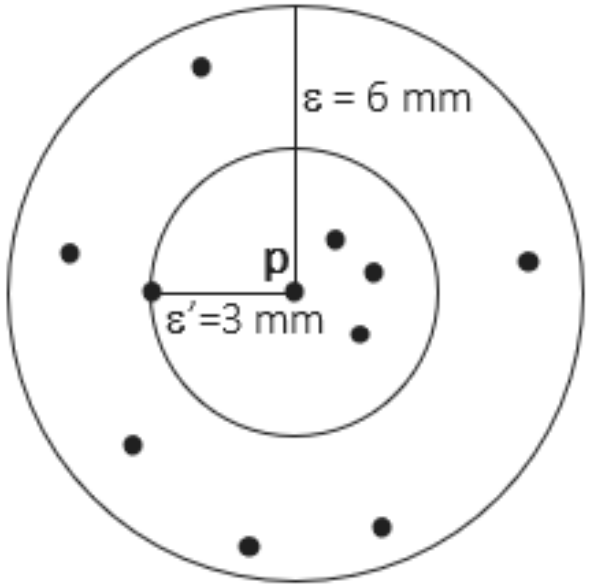
- Để khắc phục khó khăn trong việc sử dụng một bộ tham số chung trong phân tích phân cụm, phương pháp phân tích cụm OPTICS đã được đề xuất.
- OPTICS không tạo ra một cách phân cụm tập dữ liệu một cách rõ ràng. Thay vào đó, nó đưa ra thứ tự cụm.
  - Thứ tự cụm là danh sách tuyến tính của tất cả các đối tượng được phân tích và thể hiện cấu trúc phân cụm **dựa trên mật độ** của dữ liệu.
  - Các đối tượng trong cụm dày đặc hơn được liệt kê gần nhau hơn theo thứ tự cụm. Thứ tự này tương đương với phân cụm dựa trên mật độ thu được từ một loạt các cài đặt tham số. Do đó, **OPTICS không yêu cầu người dùng cung cấp ngưỡng mật độ cụ thể**.
  - Thứ tự cụm có thể được sử dụng để trích xuất thông tin phân cụm cơ bản (ví dụ: tâm cụm hoặc cụm có hình dạng tùy ý), rút ra cấu trúc phân cụm nội tại, cũng như cung cấp hình ảnh trực quan của phân cụm.

5.2. Thứ tự các điểm để xác định cấu trúc phân cụm *(OPTICS: Ordering Points to Identify the Clustering Structure)*

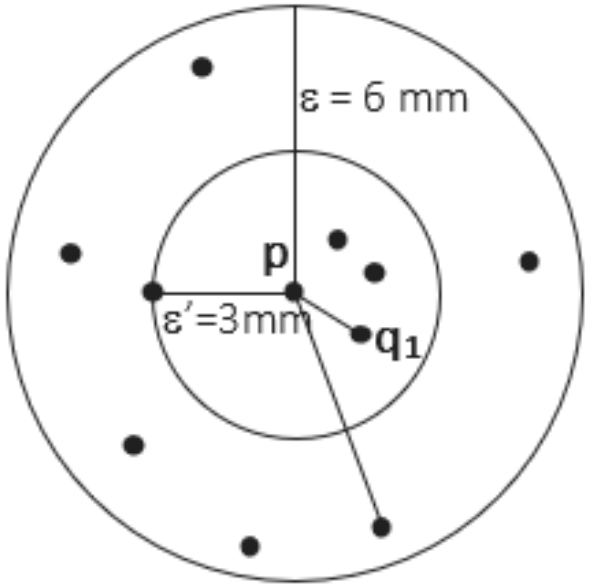
- Để xây dựng các cụm khác nhau cùng một lúc, các đối tượng được xử lý theo một thứ tự cụ thể. Thứ tự này chọn một đối tượng có thể truy cập được theo mật độ tương ứng với giá trị thấp nhất để các cụm có mật độ cao hơn (thấp hơn) sẽ được hoàn thành trước. Dựa trên ý tưởng này, OPTICS cần hai thông tin quan trọng cho mỗi đối tượng:
  - Khoảng cách lõi của đối tượng p là giá trị 0 nhỏ nhất sao cho vùng lân cận 0 của p có ít nhất *MinPts* đối tượng. Nghĩa là, 0 là ngưỡng khoảng cách tối thiểu làm cho p trở thành đối tượng cốt lõi. Nếu p không phải là đối tượng cốt lõi đối với và *MinPts*, thì khoảng cách lõi của p không được xác định.
  - Khoảng cách có thể tiếp cận tới đối tượng p từ q là giá trị bán kính tối thiểu làm cho mật độ p có thể tiếp cận được từ q. Theo định nghĩa về khả năng tiếp cận mật độ, q phải là đối tượng cốt lõi và p phải nằm trong vùng lân cận của q. Do đó, khoảng cách có thể tiếp cận từ q đến p là *max{core-distance(q), dist(p, q)}*. Nếu q không phải là đối tượng cốt lõi đối với *MinPts*, thì khoảng cách có thể tiếp cận tới p từ q là không xác định.
- Một đối tượng p có thể được truy cập trực tiếp từ nhiều đối tượng cốt lõi. Do đó, p có thể có nhiều khoảng cách tiếp cận đối với các đối tượng cốt lõi khác nhau. Khoảng cách có thể tiếp cận nhỏ nhất của p được đặc biệt quan tâm vì nó đưa ra đường đi ngắn nhất mà p được kết nối với một cụm dày đặc.

5.2. Thứ tự các điểm để xác định cấu trúc phân cụm *(OPTICS: Ordering Points to Identify the Clustering Structure)*

- Ví dụ: Hình sau minh họa về Khoảng cách lõi (*coredistance*) và khoảng cách có thể tiếp cận (*reachability-distance*). Giả sử rằng ε=6 mm và *MinPts* = 5. Khoảng cách lõi của p là khoảng cách, 0, giữa p và đối tượng dữ liệu gần nhất (thứ tự tính từ p). Khoảng cách có thể tiếp cận của q_1 từ p là khoảng cách lõi của p (tức là ε' = 3 mm) vì khoảng cách này lớn hơn khoảng cách Euclide từ p đến q_1. Khoảng cách có thể tiếp cận của q_2 đối với p là khoảng cách Euclide từ p đến q_2 vì khoảng cách này lớn hơn khoảng cách lõi của p.



Core-distance of **p**

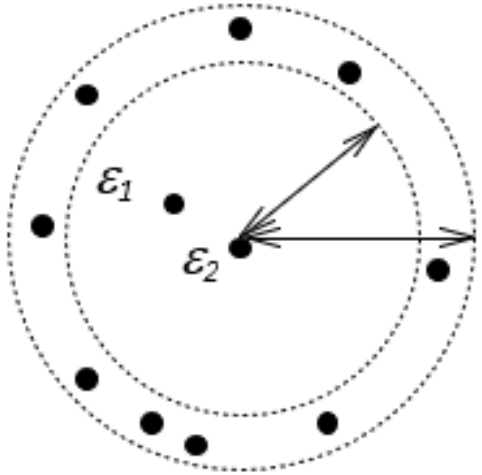


Reachability-distance (**p**, $\mathbf{q_1}$) = ε' = 3 mm
Reachability-distance (**p**, $\mathbf{q_2}$) = dist(**p**,$\mathbf{q_2}$)

- Cấu trúc của thuật toán *OPTICS* rất giống với thuật toán *DBSCAN*. Do đó, hai thuật toán có độ phức tạp về thời gian như nhau. Độ phức tạp là $O(n\log n)$ nếu sử dụng chỉ mục không gian và $O(n^2)$ nếu không sử dụng, trong đó n là số lượng đối tượng.

5.3. Phân cụm dựa trên chức năng phân phối mật độ *(DENCLUE: Clustering Based on Density Distribution Functions)*

- Ước tính mật độ (***density estimation***) là vấn đề cốt lõi trong các phương pháp phân cụm dựa trên mật độ. DENCLUE (*DENsity-based CLUstEring*) là một phương pháp phân cụm dựa trên một tập hợp các hàm phân phối mật độ.
- Trong xác suất và thống kê, ước tính mật độ là ước tính của hàm mật độ xác suất cơ bản không thể quan sát được dựa trên một tập hợp dữ liệu được quan sát. Trong bối cảnh phân cụm dựa trên mật độ, hàm mật độ xác suất cơ bản không thể quan sát được là sự phân bố thực sự của tổng thể của tất cả các đối tượng có thể được phân tích. Tập dữ liệu được quan sát được coi là một mẫu ngẫu nhiên từ quần thể đó.
- Trong *DBSCAN* và *OPTICS*, mật độ được tính bằng cách đếm số lượng đối tượng trong vùng lân cận được xác định bởi tham số bán kính. Những ước tính mật độ như vậy có thể rất nhạy cảm với giá trị bán kính được sử dụng.



5.3. Phân cụm dựa trên chức năng phân phối mật độ

- Để khắc phục vấn đề này, có thể sử dụng ước tính mật độ hạt nhân (*kernel density estimation*), đây là phương pháp ước tính mật độ phi tham số từ số liệu thống kê. Ý tưởng chung đằng sau việc ước tính mật độ hạt nhân rất đơn giản. Coi một đối tượng được quan sát như một chỉ báo về mật độ có xác suất cao ở khu vực xung quanh. Mật độ xác suất tại một điểm phụ thuộc vào khoảng cách từ điểm này đến vật thể được quan sát.
- Xem $x_1, \ldots, x_n$ là một mẫu độc lập và được phân phối giống hệt nhau của một biến ngẫu nhiên f . Xấp xỉ mật độ hạt nhân của hàm mật độ xác suất (*the probability density function*) là

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

trong đó

- *K()* là hạt nhân.
- *h* là băng thông đóng vai trò làm tham số làm mịn (*smoothing parameter*).

5.3. Phân cụm dựa trên chức năng phân phối mật độ

- Hạt nhân có thể được coi là một hàm mô hình hóa ảnh hưởng của một điểm mẫu trong vùng lân cận của nó.
- Về mặt kỹ thuật, hạt nhân *K()* là hàm tích hợp có giá trị thực không âm, phải đáp ứng hai yêu cầu: $\int_{-\infty}^{+\infty} K(u)du = 1$ và *K(-u) = K(u)* với tất cả các giá trị của u.
- Hạt nhân được sử dụng thường xuyên là hàm Gaussian tiêu chuẩn có giá trị trung bình (*mean*) bằng *0* và phương sai (*variance*) là *1*:

$$K(\frac{x-x_i}{h}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

- *DENCLUE* sử dụng nhân *Gaussian* để ước tính mật độ dựa trên tập hợp các đối tượng đã cho được phân cụm. Điểm ***x*** được gọi là điểm thu hút mật độ (*density attractor*) nếu nó là điểm cực đại cục bộ của hàm mật độ ước lượng. Để tránh các điểm cực đại cục bộ tầm thường, *DENCLUE* sử dụng ngưỡng nhiễu ξ và chỉ xem xét các bộ thu hút mật độ ***x*** sao cho $\hat{f}(x) \geq \xi$. Những điểm thu hút mật độ không cần thiết này là trung tâm của các cụm.

5.3. Phân cụm dựa trên chức năng phân phối mật độ

- Hạt nhân có thể được coi là một hàm mô hình hóa ảnh hưởng của một điểm mẫu trong vùng lân cận của nó.
- Về mặt kỹ thuật, hạt nhân *K()* là hàm tích hợp có giá trị thực không âm, phải đáp ứng hai yêu cầu: $\int_{-\infty}^{+\infty} K(u)du = 1$ và *K(-u) = K(u)* với tất cả các giá trị của u.
- Hạt nhân được sử dụng thường xuyên là hàm Gaussian tiêu chuẩn có giá trị trung bình (*mean*) bằng *0* và phương sai (*variance*) là *1*:

$$K(\frac{x-x_i}{h}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

- *DENCLUE* sử dụng nhân *Gaussian* để ước tính mật độ dựa trên tập hợp các đối tượng đã cho được phân cụm. Điểm ***x*** được gọi là điểm thu hút mật độ (*density attractor*) nếu nó là điểm cực đại cục bộ của hàm mật độ ước lượng. Để tránh các điểm cực đại cục bộ tầm thường, *DENCLUE* sử dụng ngưỡng nhiễu ξ và chỉ xem xét các bộ thu hút mật độ ***x*** sao cho $\hat{f}(x) \geq \xi$. Những điểm thu hút mật độ không cần thiết này là trung tâm của các cụm.

5.3. Phân cụm dựa trên chức năng phân phối mật độ

- Các đối tượng được phân tích được gán vào các cụm thông qua các bộ thu hút mật độ bằng cách sử dụng quy trình leo đồi từng bước (*stepwise hill-climbing procedure*). Đối với một đối tượng ***x***, quy trình leo đồi bắt đầu từ ***x*** và được hướng dẫn bởi độ dốc của hàm ước tính mật độ (*the estimated density function*). Nghĩa là, bộ thu hút mật độ (*the density attractor*) cho ***x*** được tính là

$$x^0 = x$$

$$x^{j+1} = x^j + \delta \frac{\nabla \hat{f}(x^j)}{|\nabla \hat{f}(x^j)|}$$

trong đó

- δ là tham số để kiểm soát tốc độ hội tụ (*the speed of convergence*)
- $\nabla \hat{f}(x) = \frac{1}{h^{d+2} n \sum_{i=1}^{n} K(\frac{x-x_i}{h})(x_i - x)}$

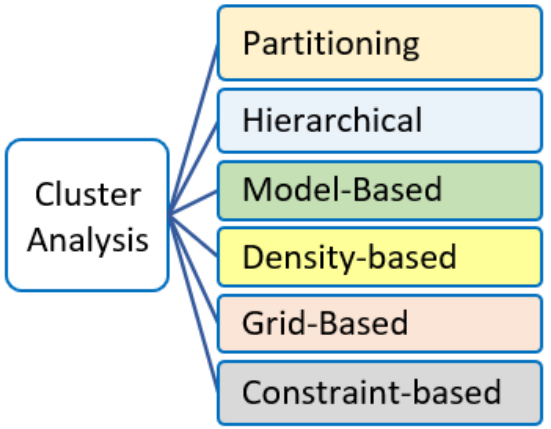
- Quy trình leo đồi dừng ở bước $k > 0 \; if \; \hat{f}(x^{k+1}) < \hat{f}(x^k)$ và gán ***x*** cho bộ thu hút mật độ $x^* = x^k$. Một đối tượng ***x*** là một ngoại lệ hoặc nhiễu nếu nó hội tụ (*converges*) trong thủ tục leo đồi đến mức x^* cực đại cục bộ với $\hat{f}(x^*) < \xi$.

NỘI DUNG CHƯƠNG 5

1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

6. CÁC PHƯƠNG PHÁP DỰA TRÊN LƯỚI *(Grid-Based Methods)*

- Các phương pháp phân cụm trước đều dựa trên dữ liệu - chúng phân vùng tập hợp các đối tượng và thích ứng với sự phân bố của các đối tượng trong không gian nhúng (*the embedding space*).

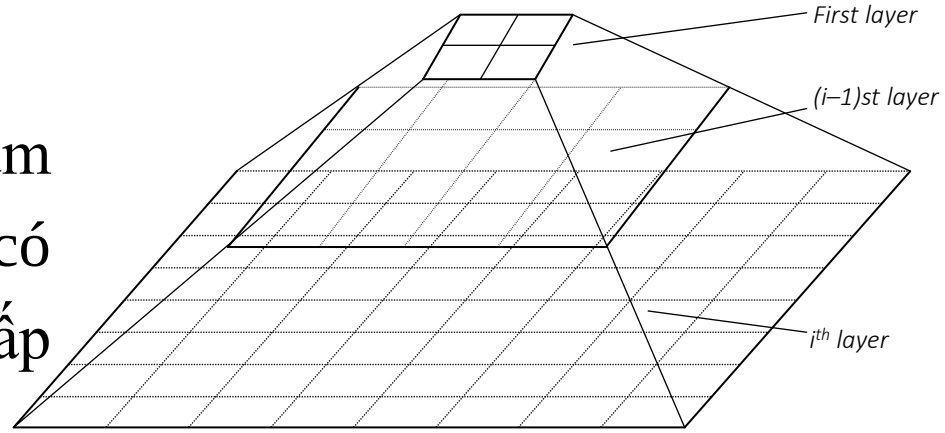


- Phương pháp phân cụm dựa trên lưới thực hiện cách tiếp cận theo hướng không gian bằng cách phân vùng không gian nhúng vào các ô độc lập với sự phân bố của các đối tượng đầu vào.
- Phương pháp phân cụm dựa trên lưới sử dụng cấu trúc dữ liệu lưới đa độ phân giải (*multiresolution grid data structure*). Nó lượng tử hóa không gian đối tượng thành một số lượng hữu hạn các ô tạo thành cấu trúc lưới trên đó tất cả các hoạt động phân cụm được thực hiện.
- Ưu điểm chính của phương pháp này là thời gian xử lý nhanh, thường không phụ thuộc vào số lượng đối tượng dữ liệu, nhưng chỉ phụ thuộc vào số lượng ô trong mỗi chiều trong không gian lượng tử hóa.

6.1 STING: Lưới thông tin thống kê *(STatistical INformation Grid)*

- *STING* là một kỹ thuật phân cụm đa độ phân giải dựa trên lưới, trong đó vùng không gian nhúng của các đối tượng đầu vào được chia thành các ô hình chữ nhật.
- Không gian có thể được phân chia theo cách phân cấp và đệ quy. Một số cấp độ của các ô hình chữ nhật như vậy tương ứng với các mức độ phân giải khác nhau và tạo thành cấu trúc phân cấp: Mỗi ô ở cấp độ cao được phân vùng để tạo thành một số ô ở cấp độ thấp hơn tiếp theo. Thông tin thống kê liên quan đến các thuộc tính trong mỗi ô lưới, chẳng hạn như giá trị trung bình, giá trị tối đa và giá trị tối thiểu, được tính toán trước và lưu trữ dưới dạng tham số thống kê. Các tham số thống kê này rất hữu ích cho việc xử lý truy vấn và cho các nhiệm vụ phân tích dữ liệu khác.

6.1 STING: Lưới thông tin thống kê *(STatistical INformation Grid)*



- Hình bên cho thấy cấu trúc phân cấp cho phân cụm *STING*. Các tham số thống kê của các ô cấp cao hơn có thể được tính toán dễ dàng từ các tham số của các ô cấp thấp hơn.

- Các tham số này bao gồm:
  - Tham số độc lập với thuộc tính (*the attribute-independent parameter*): số lượng;
  - Tham số phụ thuộc thuộc tính (*the attribute-dependent parameters*):
    - *min*, *max* , *mean* (trung bình)
    - stdev (*stdev - standard deviation* - độ lệch chuẩn)
    - Loại phân phối mà giá trị thuộc tính trong ô tuân theo như: *normal*, *uniform*, *exponential* hoặc none (nếu sự phân bố chưa được biết).
- Khi dữ liệu được tải vào cơ sở dữ liệu, các tham số đếm, trung bình, stdev, min và max của các ô cấp dưới cùng được tính trực tiếp từ dữ liệu. Giá trị của phân phối có thể được người dùng chỉ định nếu loại phân phối được biết trước hoặc thu được bằng các thử

6.1 STING: Lưới thông tin thống kê *(STatistical INformation Grid)*

- Ưu điểm của *STING* :
  i. Tính toán dựa trên lưới không phụ thuộc vào truy vấn vì thông tin thống kê được lưu trữ trong mỗi ô thể hiện thông tin tóm tắt của dữ liệu trong ô lưới, độc lập với truy vấn;
  ii. Cấu trúc lưới hỗ trợ xử lý song song và cập nhật gia tăng;
  iii. Thời gian xử lý nhanh: *STING* đi qua cơ sở dữ liệu một lần để tính toán các tham số thống kê của các ô, do đó độ phức tạp về thời gian của việc tạo cụm là *O(n)*, trong đó ***n*** là tổng số đối tượng. Sau khi tạo cấu trúc phân cấp, thời gian xử lý truy vấn là *O(g)*, trong đó *g* là tổng số ô lưới ở mức thấp nhất (thường nhỏ hơn nhiều so với *n)*.
- Chất lượng của phân cụm *STING* phụ thuộc vào mức độ chi tiết của mức thấp nhất của cấu trúc lưới. Nếu độ chi tiết rất mịn thì chi phí xử lý sẽ tăng lên đáng kể; ngược lại, nếu mức dưới cùng của cấu trúc lưới quá thô, có thể làm giảm chất lượng phân tích cụm.
- *STING* không xem xét mối quan hệ không gian giữa các ô con và các ô lân cận để xây dựng ô cha. Do đó các ranh giới của cụm là ngang hoặc dọc và không phát hiện được ranh giới đường chéo. Điều này có thể làm giảm chất lượng và độ chính xác của các cụm mặc dù thời gian xử lý kỹ thuật nhanh.

6.2. CLIQUE: Phương pháp phân cụm không gian con giống như Apriori

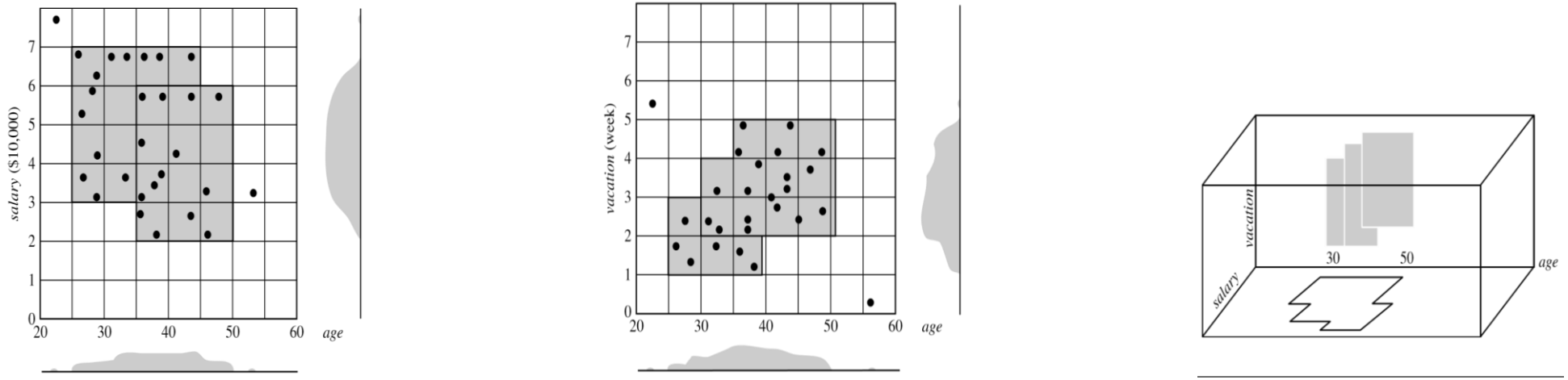
(An Apriori-like Subspace Clustering Method)

- Một đối tượng dữ liệu thường có hàng chục thuộc tính, trong đó có nhiều thuộc tính không liên quan. Giá trị của các thuộc tính có thể thay đổi đáng kể. Những yếu tố này có thể gây khó khăn cho việc xác định các cụm trải rộng trên toàn bộ không gian dữ liệu. Thay vào đó, có thể có ý nghĩa hơn nếu tìm kiếm các cụm trong các không gian con khác nhau của dữ liệu.
- Ví dụ: Khó có thể tìm được một nhóm bệnh nhân mà tất cả hoặc thậm chí hầu hết các đặc điểm đều giống nhau hoàn toàn. Ví dụ, ở bệnh nhân cúm gia cầm, độ tuổi, giới tính và đặc tính công việc có thể thay đổi đáng kể trong một phạm vi giá trị rộng. Vì vậy, có thể khó tìm được cụm như vậy trong toàn bộ không gian dữ liệu. Thay vào đó, bằng cách tìm kiếm trong không gian con, có thể tìm thấy một nhóm bệnh nhân tương tự trong không gian có chiều thấp hơn (ví dụ: những bệnh nhân giống nhau về các triệu chứng như sốt cao, ho nhưng không sổ mũi và từ 3 đến 16 tuổi).

6.1 CLIQUE: Phương pháp phân cụm không gian con giống như Apriori

- *CLIQUE* (*CLustering In QUEst*) là một phương pháp dựa trên lưới đơn giản để tìm các cụm dựa trên mật độ trong không gian con. *CLIQUE* phân vùng mỗi chiều thành các khoảng không chồng chéo, từ đó phân vùng toàn bộ không gian nhúng của đối tượng dữ liệu vào các ô.
- *CLIQUE* sử dụng ngưỡng mật độ để xác định các ô dày đặc và các ô thưa thớt. Một ô dày đặc nếu số lượng đối tượng được ánh xạ tới nó vượt quá ngưỡng mật độ.
- Chiến lược chính đằng sau CLIQUE để xác định không gian tìm kiếm ứng viên sử dụng tính đơn điệu của các ô dày đặc liên quan đến chiều. Điều này dựa trên thuộc tính Apriori được sử dụng trong khai phá luật kết hợp và mẫu thường xuyên. Trong bối cảnh các cụm trong không gian con, tính đơn điệu nói như sau. Một ô (cell) ***c***, *k*-chiều (*k-dimensions*) (với $k > 1$) chỉ có thể có ít nhất ***l*** điểm nếu mọi hình chiếu hai chiều $(k-1)$ của ***c***, là một ô trong không gian con $a(k-1)$-chiều ($a(k-1)$-*dimensions*), có ít nhất ***l*** điểm.

6.1 CLIQUE: Phương pháp phân cụm không gian con giống như Apriori



- Xét hình trên, trong đó không gian dữ liệu nhúng chứa ba chiều: tuổi (*age*), tiền lương (*salary*) và kỳ nghỉ (*vacation*). Một ô 2-D, chẳng hạn như trong không gian con được hình thành bởi tuổi và mức lương, chỉ chứa l điểm nếu hình chiếu của ô này theo mọi chiều, nghĩa là tuổi và mức lương tương ứng, chứa ít nhất l điểm.

6.1 CLIQUE: Phương pháp phân cụm không gian con giống như Apriori

- Hiệu quả của *CLIQUE*:
  - *CLIQUE* tự động tìm các không gian con có số chiều cao nhất sao cho các cụm mật độ cao tồn tại trong các không gian con đó.
  - Nó không nhạy cảm với thứ tự của các đối tượng đầu vào và không giả định bất kỳ phân phối dữ liệu chuẩn nào. Nó chia tỷ lệ tuyến tính theo kích thước của đầu vào và có khả năng mở rộng tốt khi số lượng chiều trong dữ liệu tăng lên.
  - Việc thu được một phân cụm có ý nghĩa phụ thuộc vào việc điều chỉnh kích thước lưới thích hợp (đây là cấu trúc ổn định ở đây) và ngưỡng mật độ.
    - Điều này có thể khó khăn trong thực tế vì kích thước lưới và ngưỡng mật độ được sử dụng trên tất cả các kết hợp chiều trong tập dữ liệu. Do đó, độ chính xác của kết quả phân cụm có thể bị suy giảm do tính đơn giản của phương pháp.
    - Đối với một vùng dày đặc nhất định, tất cả các hình chiếu của vùng đó lên các không gian con có chiều thấp hơn cũng sẽ dày đặc. Điều này có thể dẫn đến sự chồng chéo lớn giữa các khu vực dày đặc.
    - Rất khó tìm được các cụm có mật độ khá khác nhau trong các không gian con có chiều khác nhau.

NỘI DUNG CHƯƠNG 5

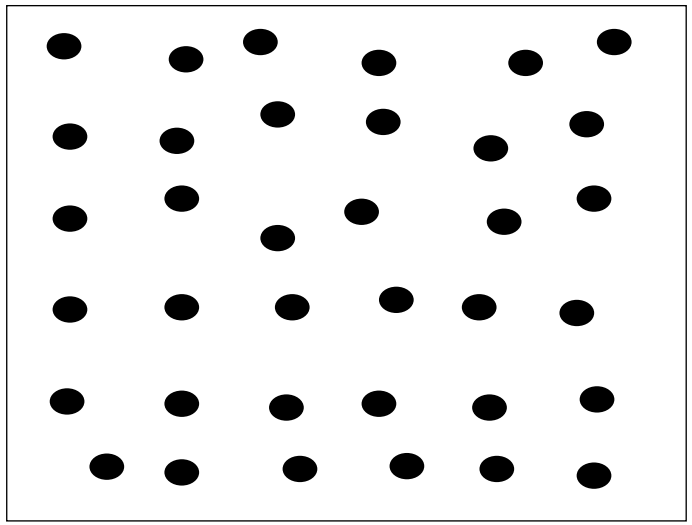
1. Giới thiệu
2. Phân tích cụm (*Cluster Analysis*)
3. Các phương pháp phân vùng (*Partitioning Methods*)
4. Các phương pháp phân cấp (*Hierarchical Methods*)
5. Các phương pháp dựa trên mật độ (*Density-Based Methods*)
6. Các phương pháp phân cụm dựa trên lưới (*Grid-Based Methods*)
7. Đánh giá phân cụm (*Evaluation of Clustering*)
8. Bài tập

7. ĐÁNH GIÁ PHÂN CỤM (*Evaluation of Clustering*)

7.1. Đánh giá xu hướng phân cụm *(Assessing Clustering Tendency)*

- Đánh giá xu hướng phân cụm xác định xem một tập dữ liệu nhất định có cấu trúc không ngẫu nhiên hay không, điều này có thể dẫn đến các cụm có ý nghĩa.

Hình sau cho thấy một tập dữ liệu được phân bố đồng đều trong không gian dữ liệu 2-D. Mặc dù thuật toán phân cụm vẫn có thể phân chia các điểm thành các nhóm một cách giả tạo, nhưng các nhóm này khó có thể có ý nghĩa quan trọng đối với ứng dụng do sự phân bố dữ liệu đồng đều.



7. Đánh giá phân cụm (*Evaluation of Clustering*)

7.1. Đánh giá xu hướng phân cụm *(Assessing Clustering Tendency)*

- Thống kê Hopkins là một thống kê kiểm tra tính ngẫu nhiên về mặt không gian của một biến được phân bố trong một không gian.
- Cho tập dữ liệu *D*, được coi là mẫu của một biến ngẫu nhiên *o*, để xác định xem o còn cách phân bố đồng đều trong không gian dữ liệu bao xa. Tính toán thống kê Hopkins như sau:
  i. Mẫu *n* điểm, $p_1, ..., p_n$, đều từ *D*. Nghĩa là mỗi điểm thuộc *D* có cùng xác suất lọt vào mẫu này. Với mỗi điểm p_i, tìm lân cận gần nhất của *pi* ($1 \le i \le n$) trong *D* và gọi x_i là khoảng cách giữa p_i và lân cận gần nhất của nó trong *D*. Nghĩa là, $x_i = \min_{v \in D}\{dist(p_i, v)\}$
  ii. Lấy mẫu *n* điểm, $q_1, ..., q_n$, đều từ *D*. Với mỗi q_i ($1 \le i \le n$), ta tìm điểm lân cận gần nhất của q_i trong $D - \{q_i\}$, và gọi y_i là khoảng cách giữa q_i và láng giềng gần nhất của nó trong $D - \{q_i\}$. Đó là,

$$y_i = \min_{v \in D, v \neq q}\{dist(q_i, v)\}$$

iii. Tính thống kê Hopkins (H) $H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$

Thống kê Hopkins cho biết:

- Nếu D được phân bố đồng đều, thì $\sum_{i=1}^{n} y_i$ và $\sum_{i=1}^{n} x_i$ sẽ ở gần nhau và do đó H sẽ vào khoảng 0,5. Nếu H > 0,5 thì khó có khả năng D có các cụm có ý nghĩa thống kê.
- Ngược lại $\sum_{i=1}^{n} y_i$ sẽ nhỏ hơn đáng kể so với $\sum_{i=1}^{n} x_i$ trong kỳ vọng, và do đó H sẽ gần bằng 0.

7.2. Xác định số lượng cụm *(Determining the Number of Clusters)*

- Việc xác định số cụm "đúng" trong một tập dữ liệu là quan trọng, không chỉ vì một số thuật toán phân cụm như k-means yêu cầu tham số như vậy mà còn vì số lượng cụm thích hợp sẽ kiểm soát mức độ chi tiết thích hợp của phân tích cụm. Nó có thể được coi là tìm ra sự cân bằng tốt giữa khả năng nén và độ chính xác trong phân tích cụm. Hãy xem xét hai trường hợp cực đoan sau:

i. Toàn bộ tập dữ liệu là một cụm:

- Điều này sẽ tối đa hóa khả năng nén dữ liệu.
- Nhưng phân tích cụm như vậy không có giá trị.

ii. Chỉ có một đối tượng trên mỗi cụm: do đó không cho phép tóm tắt dữ liệu.

7.2. Xác định số lượng cụm *(Determining the Number of Clusters)*

- Việc xác định số lượng cụm không hề dễ dàng vì để xác định số cụm "đúng" là không rõ rang vì phụ thuộc vào:
  - Hình dạng của cụm,
  - Tỷ lệ của phân phối trong tập dữ liệu,
  - Độ phân giải phân cụm mà người dùng yêu cầu.
- *Một số phương pháp đơn giản nhưng phổ biến và hiệu quả*

i. Một phương pháp đơn giản là đặt số lượng cụm thành khoảng $\sqrt{\frac{n}{2}}$ cho tập dữ liệu gồm $\boldsymbol{n}$ điểm. Với kỳ vọng, mỗi cụm có $\sqrt{2n}$ điểm.

7.2. Xác định số lượng cụm *(Determining the Number of Clusters)*

- Một số phương pháp đơn giản nhưng phổ biến và hiệu quả:

  ii. Phương pháp khuỷu tay (*elbow method*) dựa trên quan sát rằng việc tăng số lượng cụm có thể giúp giảm tổng phương sai trong cụm của mỗi cụm. Điều này là do việc có nhiều cụm hơn cho phép người ta nắm bắt được các nhóm đối tượng dữ liệu tốt hơn, giống nhau hơn. Tuy nhiên, tác động cận biên của việc giảm tổng phương sai trong cụm có thể giảm nếu có quá nhiều cụm được hình thành, bởi vì việc chia một cụm gắn kết thành hai chỉ mang lại mức giảm nhỏ.

Do đó, một phương pháp heuristic để chọn số cụm phù hợp là sử dụng điểm ngoặt trên đường cong của tổng phương sai trong cụm đối với số lượng cụm. Về mặt kỹ thuật, với một số $k>0$, có thể tạo k cụm trên tập dữ liệu được đề cập bằng cách sử dụng thuật toán phân cụm như k-means và tính tổng phương sai bên trong cụm, $var(k)$. Sau đó có thể vẽ đường cong của var đối với k. Điểm rẽ đầu tiên (hoặc quan trọng nhất) của đường cong gợi ý số "đúng".

7.2. Xác định số lượng cụm *(Determining the Number of Clusters)*

- Một số phương pháp đơn giản nhưng phổ biến và hiệu quả:

iii. ***Xác thực chéo*** (*cross validation*), một kỹ thuật thường được sử dụng trong phân loại.

- Đầu tiên, chia tập dữ liệu đã cho D thành m phần.
- Tiếp theo, sử dụng $m-1$ phần để xây dựng mô hình phân cụm và sử dụng phần còn lại để kiểm tra chất lượng của phân cụm. Ví dụ: đối với mỗi điểm trong tập kiểm tra, có thể tìm được trọng tâm gần nhất. Do đó, có thể sử dụng tổng bình phương khoảng cách giữa tất cả các điểm trong tập kiểm tra và các trọng tâm gần nhất để đo mức độ phù hợp của mô hình phân cụm với tập kiểm tra.
- Với bất kỳ số nguyên $k>0$ nào, lặp lại quá trình này m lần để rút ra các cụm gồm k cụm bằng cách sử dụng lần lượt từng phần làm tập kiểm tra.
- Giá trị trung bình của thước đo chất lượng được lấy làm thước đo chất lượng tổng thể. Sau đó, có thể so sánh thước đo chất lượng tổng thể với các giá trị khác nhau của k và tìm số cụm phù hợp nhất với dữ liệu.

7.3. Đo lường chất lượng phân cụm *(Measuring Clustering Quality)*

- Các phương pháp đo lường có thể được phân loại thành hai nhóm tùy theo sự thật cơ bản (*ground truth*) có sẵn hay không. Ở đây, sự thật cơ bản là phân cụm lý tưởng thường được xây dựng bằng cách sử dụng các chuyên gia con người.
  - ***Phương pháp giám sát*** (*supervised methods*): Nếu sự thật cơ bản có sẵn, nó có thể được sử dụng bằng các phương pháp bên ngoài, so sánh việc phân cụm với sự thật và thước đo của nhóm.
  - ***Phương pháp không giám sát*** (*unsupervised methods*): Nếu không có thông tin cơ bản, có thể sử dụng các phương pháp nội tại để đánh giá mức độ tốt của việc phân cụm bằng cách xem xét các cụm được phân tách tốt như thế nào. Sự thật cơ bản có thể được coi là sự giám sát dưới dạng "nhãn cụm".

7.3.1. Phương pháp giám sát *(hay phương pháp bên ngoài - Extrinsic Methods)*

7.3.1.1. Thước đo chất lượng phân cụm Q: gồm bốn tiêu chí thiết yếu sau:

i. *Tính đồng nhất của cụm* (*Cluster homogeneity*).

- Điều này đòi hỏi các cụm trong một cụm càng thuần túy (*pure*) thì phân cụm càng tốt.
- Giả sử sự thật cơ bản nói rằng các đối tượng trong tập dữ liệu D có thể thuộc các loại $L_1, ..., L_n$. Hãy xem xét việc phân cụm $C1$, trong đó cụm $C \in C_1$ chứa các đối tượng thuộc hai loại L_i, L_j $(1 \le i < j \le n)$. Cũng xem xét phân cụm C_2, giống hệt với C_1 ngoại trừ C_2 được chia thành hai cụm chứa các đối tượng tương ứng trong L_i và L_j. Thước đo chất lượng phân cụm, Q, tôn trọng tính đồng nhất của cụm sẽ cho điểm C_2 cao hơn C_1, nghĩa là $Q(C_2, C_g) > Q(C_1, C_g)$.

7.3.1.1. Thước đo chất lượng phân cụm Q: gồm bốn tiêu chí thiết yếu sau:

ii. *Tính đầy đủ của cụm* (*Cluster completeness*)

- Đây là bản sao của tính đồng nhất của cụm. Tính đầy đủ của cụm yêu cầu rằng để phân cụm, nếu bất kỳ hai đối tượng nào thuộc cùng một loại theo sự thật cơ bản thì chúng phải được gán vào cùng một cụm.
- Xét phân cụm C_1, trong đó chứa các cụm C_1 và C_2, trong đó các thành viên thuộc cùng loại theo sự thật cơ bản. Giả sử cụm C_2 giống hệt C_1 ngoại trừ C_1 và C_2 được hợp nhất thành một cụm trong C_2. Sau đó, thước đo chất lượng phân cụm, Q, tôn trọng tính đầy đủ của cụm sẽ cho điểm C_2 cao hơn, nghĩa là $Q(C_2, C_g) > Q(C_1, C_g)$.

7.3.1.1. Thước đo chất lượng phân cụm Q: gồm bốn tiêu chí thiết yếu sau:

iii. ***Túi giẻ rách*** (*Rag bag*)

- Trong nhiều tình huống thực tế, thường có một danh mục "*rag bag*" chứa các đối tượng không thể hợp nhất với các đối tượng khác. Danh mục như vậy thường được gọi là "*linh tinh*" (*miscellaneous*), "*khác*" (*other*), v.v.
- Tiêu chí "*rag bag*" nói rằng việc đặt một vật thể không đồng nhất vào một cụm đã đồng nhất sẽ gây "thiệt hại" nhiều hơn là khi bỏ nó vào một túi giẻ rách.
- Xét một cụm C_1 và một cụm $C \in C_1$ sao cho tất cả các đối tượng trong C ngoại trừ một đối tượng, ký hiệu là o, thuộc cùng một loại theo *sự thật cơ bản*. Hãy xem xét một cụm C_2 giống hệt với C_1 ngoại trừ o được gán cho cụm $C' \neq C$ trong C_2 sao cho C' chứa các đối tượng thuộc nhiều danh mục khác nhau tùy theo sự thật cơ bản và do đó nhiễu. Nói cách khác, C' trong C_2 là một cái túi đựng giẻ rách. Sau đó, thước đo chất lượng phân cụm Q tôn trọng tiêu chí *Rag bag* sẽ cho điểm C_2 cao hơn, nghĩa là $Q(C_2, C_g) > Q(C_1, C_g)$.

7.3.1.1. Thước đo chất lượng phân cụm Q: gồm bốn tiêu chí thiết yếu sau:

iv. *Bảo quản cụm nhỏ* (*Small cluster preservation*)

- Nếu một danh mục nhỏ được chia thành các phần nhỏ trong một cụm thì những phần nhỏ đó có thể trở thành nhiễu và do đó danh mục nhỏ không thể được phát hiện từ cụm.
- Tiêu chí bảo tồn cụm nhỏ nêu rõ rằng việc chia một danh mục nhỏ thành nhiều phần sẽ có hại hơn so với việc chia một danh mục lớn thành nhiều phần.
- Xét một trường hợp cực đoan. Cho D là tập dữ liệu gồm $n+2$ đối tượng sao cho, theo thực tế cơ bản, n đối tượng, ký hiệu là $o_1, ..., o_n$, thuộc một loại và hai đối tượng còn lại, ký hiệu là o_{n+1}, o_{n+2}, thuộc về một loại khác. Giả sử phân cụm C_1 có ba cụm, $C_1 = \{o_1, ..., o_n\}$, $C_2 = \{o_{n+1}\}$ và $C3 = \{o_{n+2}\}$. Giả sử phân cụm C_2 cũng có ba cụm, cụ thể là $C_1 = \{o_1, ..., o_{n-1}\}$, $C_2 = \{o_n\}$, và $C_3 = \{o_{n+1}, o_{n+2}\}$. Nói cách khác, C_1 chia danh mục nhỏ và C_2 chia danh mục lớn. Thước đo chất lượng phân cụm Q bảo toàn các cụm nhỏ sẽ cho điểm C_2 cao hơn, nghĩa là $Q(C_2, C_g) > Q(C_1, C_g)$.

7.3.1.2. Thước đo chất lượng phân cụm BCubed

- Có nhiều thước đo chất lượng phân cụm đáp ứng một số trong bốn tiêu chí này. Ở đây, giới thiệu các số liệu về độ chính xác và thu hồi của *BCubed*, đáp ứng cả bốn tiêu chí.
- *BCubed* đánh giá độ chính xác và thu hồi của mọi đối tượng trong một cụm trên một tập dữ liệu nhất định theo *sự thật cơ bản*. Độ chính xác của một đối tượng cho biết có bao nhiêu đối tượng khác trong cùng một cụm thuộc cùng loại với đối tượng đó. Việc thu hồi một đối tượng phản ánh có bao nhiêu đối tượng cùng loại được gán vào cùng một cụm.

7.3.1.2. Thước đo chất lượng phân cụm BCubed

– Về mặt hình thức, đặt $D=\{o_1, ..., o_n\}$ là một tập các đối tượng và C là một cụm trên D. Đặt $L(o_i)$ $(1 \le i \le n)$ là phạm trù của o_i được cho bởi *sự thật cơ bản*, và $C(o_i)$ là cluster_ID của o_i trong C. Khi đó, với hai đối tượng o_i và o_j, $(1 \le i, j \le n, i \ne j)$, tính đúng đắn của quan hệ giữa o_i và o_j trong phân cụm C được đưa ra bởi:

$$\text{Correctness}(o_i, o_j) = \begin{cases} 1 \; if \; L(o_i) = L(o_j) \leftrightarrow C(o_i) = C(o_j) \\ 0 \; otherwise \end{cases}$$

- Độ chính xác của BCubed được định nghĩa là

$$\text{Precision BCubed} = \frac{\sum_{i=1}^{n} \frac{\sum_{o_{j:i \ne j, C(o_i)=C(o_j)}} \text{Correctness}(o_i, o_j)}{\|\{o_j | i \ne j, C(o_i) = C(o_j)\}\|}}{n}$$

- Độ thu hồi BCubed được định nghĩa là

$$\text{Precision BCubed} = \frac{\sum_{i=1}^{n} \frac{\sum_{o_{j:i \ne j, L(o_i)=L(o_j)} \text{Correctness}(o_i, o_j)}}{\|\{o_j | i \ne j, L(o_i) = L(o_j)\}\|}}{n}$$

7.3.2. Phương pháp không giám sát *(hay phương pháp nội tại - Intrinsic Methods)*

- Khi không có sẵn thông tin cơ bản về tập dữ liệu, phải sử dụng phương pháp nội tại (*intrinsic method*) để đánh giá chất lượng phân cụm. Nói chung, các phương pháp nội tại đánh giá một phân cụm bằng cách kiểm tra xem các cụm được phân tách tốt như thế nào và các cụm nhỏ gọn đến mức nào. Nhiều phương pháp nội tại có ưu điểm là số liệu tương tự giữa các đối tượng trong tập dữ liệu.
- Hệ số hình bóng (*silhouette coefficient*) là một thước đo như vậy. Với tập dữ liệu D gồm n đối tượng, giả sử D được phân chia thành k cụm $C_1, ..., C_k$. Đối với mỗi đối tượng $o \in D$, tính $a(o)$ là khoảng cách trung bình giữa o và tất cả các đối tượng khác trong cụm mà o thuộc về. Tương tự, $b(o)$ là khoảng cách trung bình tối thiểu từ o tới tất cả các cụm mà o không thuộc về.

7.3.2. Phương pháp không giám sát *(hay phương pháp nội tại - Intrinsic Methods)*

– Về mặt hình thức, giả sử $o \in Ci\ (1 \le i \le k)$; sau đó

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o, o')}{|C_i| - 1}$$

và

$$b(o) = \min_{C_j : 1 \le j \le k, j \neq i} \left\{ \frac{\sum_{o' \in C_i} dist(o, o')}{|C_j|} \right\}$$

Hệ số hình bóng của o được định nghĩa:

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

7.3.2. Phương pháp không giám sát *(hay phương pháp nội tại - Intrinsic Methods)*

- Các giá trị:
  - Giá trị của hệ số hình bóng $s(o)$ nằm trong khoảng từ -1 đến 1.
    - Khi $s(o)$ tiến tới 1, cụm chứa o nhỏ gọn và o ở xa các cụm khác, đây là tình huống tốt.
    - Khi $s(o)$ là âm (tức là $b(o) < a(o)$), điều này có nghĩa là o gần các đối tượng trong cụm khác hơn so với các đối tượng trong cùng cụm với o. Trong nhiều trường hợp, đây là một tình huống xấu và nên tránh.
  - Giá trị của $a(o)$ phản ánh mức độ nén của cụm mà o thuộc về. Giá trị càng nhỏ thì cụm càng nhỏ gọn.
  - Giá trị của $b(o)$ thể hiện mức độ tách biệt của o với các cụm khác. $b(o)$ càng lớn thì o càng tách biệt với các cụm khác.
- Để đo lường mức độ phù hợp của cụm trong cụm, có thể tính giá trị hệ số hình bóng trung bình của tất cả các đối tượng trong cụm. Để đo chất lượng phân cụm, có thể sử dụng giá trị hệ số hình bóng trung bình của tất cả các đối tượng trong tập dữ liệu. Hệ số hình bóng và các thước đo nội tại khác cũng có thể được sử dụng trong phương pháp khuỷu tay để suy ra số lượng cụm trong một tập dữ liệu bằng cách thay thế tổng phương sai bên trong cụm.

7. BÀI TẬP

i. Cho tập 7 điểm trong mặt phẳng xOy như sau:

Giả sử cần phân các điểm này thành 3 cụm, với các điểm được chọn làm tâm ban đầu là: D, E, G. Thực hiện:

a. Phân cụm bằng K-means.

b. Phân cụm bằng K-Medoids (xác định trọng tâm của cụm với median)

c. So sánh kết quả thực hiện của 2 câu a và b.

point	*X*	*Y*
A	2	3
B	1	2
C	4	4
D	2	2
E	1	1
F	4	5
G	3	3

Thực hiện lại yêu cầu trên nhưng với số cụm là 2 với 2 tâm là A và B.

7. BÀI TẬP

ii. Cho tập 10 điểm trong mặt phẳng xOy như sau:

Yêu cầu thực hiện: lần lượt thực hiện việc phân cụm theo các phương pháp sau:

a. *Phương pháp phân cụm theo cấp bậc kết tụ* (*agglomerative hierarchical clustering method*)

b. *Phương pháp phân cụm theo phân chia cấp bậc* (*divisive hierarchical clustering method*)

i	x	y
A	2	6
B	3	4
C	3	8
D	4	7
E	6	2
F	6	4
G	7	3
H	7	4
I	8	5
J	7	6


Q&A
How
Why
When
Where
What
Who