PHÂN CỤM DỮ LIỆU

Giảng viên: Đặng Thị Thu Hiền, Nguyễn Tu Trung
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2021

Nội dung

- Khái niệm về phân cụm dữ liệu (PCDL)
- Ứng dụng của PCDL
- Các vấn đề cần giải quyết đối với PCDL
- Biểu diễn dữ liệu
- Kỹ thuật phân hoạch dữ liệu

Khái niệm về PCDL

- PCDL:
  - Là một kĩ thuật trong KPDL
  - Nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên, tiềm ẩn, quan trọng trong tập dữ liệu lớn => cung cấp thông tin, tri thức hữu ích cho việc ra quyết định
  - Là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" với nhau
- Trong học máy, PCDL là vấn đề học không có giám sát
- Trong nhiều trường hợp, PCDL là một bước trong phân lớp dữ liệu: Có thể khởi tạo các lớp cho phân lớp bằng cách xác định các nhãn cho các nhóm dữ liệu

Ứng dụng của PCDL

- Có thể ứng dụng PCDL trong nhiều lĩnh vực, ví dụ:
  - Thương mại: Tìm kiếm nhóm các khách hàng quan trọng có đặc trưng tương đồng và những đặc tả họ từ các bản ghi mua bán trong CSDL khác hàng
  - Sinh học: Phân loại các gen với các chức năng tương đồng
  - Thư viên: Phân loại sách
  - Bảo hiểm: Nhận dang nhóm thao gia bảo hiểm có chi phí bồi thường cao, nhận dạng gian lận
  - Quy hoạch đô thị: Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý …nhằm cung cấp thông tin cho quy hoạch đô thị
  - Nghiên cứu trái đất: Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm
  - WWW: Có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web => Các lớp tài liệu này trợ giúp cho việc KPDL từ dữ liệu

Các vấn đề cần giải quyết đối với PCDL

- Có khả năng mở rộng:
  - Các tập dữ liệu lớn có thể lên tới hàng triệu đối tượng
- Khả năng thích nghi với các kiểu thuộc tính khác nhau:
  - Nhiều ứng dụng có thể đòi hỏi việc phân cụm với nhiều kiểu dữ liệu khác nhau: kiểu số, nhị phân, kiểu tường minh (định danh – không thứ tự), và dữ liệu có thứ tự hay dạng hỗn hợp của những kiểu dữ liệu này
- Khám phá các cụm với hình dạng bất kỳ:
  - Nhiều thuật toán phân cụm các phép đo khoảng cách Euclidean và Manhattan => hướng tới việc tìm kiếm các cụm hình cầu với mật độ và kích cỡ tương tự nhau
  - Thực tế, một cụm có thể có bất cứ một hình dạng nào => phát triển các thuật toán có thể khám phá ra các cụm có hình dạng bất kỳ là rất quan trọng

Các vấn đề cần giải quyết đối với PCDL

- Tối thiểu lượng tri thức cần cho xác định các tham số đầu vào:
  - Nhiều thuật toán phân cụm yêu cầu những tham số đầu vào nhất định trong phân tích phân cụm (như số lượng các cụm mong muốn)
  - Kết quả của phân cụm thường khá nhạy cảm với các tham số đầu vào
  - Nhiều tham số rất khó để xác định, nhất là với các tập dữ liệu có lượng các đối tượng lớn => người dùng khó điều chỉnh để đạt chất lượng phân cụm tốt
- Khả năng thích nghi với dữ liệu nhiễu:
  - Dữ liệu ngoại lai, dữ liệu lỗi, dữ liệu chưa biết hoặc dữ liệu sai
  - Nhạy cảm với dữ liệu => dẫn đến chất lượng phân cụm

Các vấn đề cần giải quyết đối với PCDL

- Số chiều lớn
- Ít nhạy cảm với thứ tự của các dữ liệu vào:
  - Cùng một tập dữ liệu, thứ tự khác nhau, cùng một thuật toán có thể sinh ra các cụm rất khác nhau
  - Phát triển các thuật toán ít nhạy cảm với thứ tự vào của dữ liệu rất quan trọng
- Phân cụm ràng buộc:
  - Yêu cầu: phân cụm tốt và thỏa mãn thêm các yêu cầu ràng buộc
- Dễ hiểu dễ sử dụng:

Biểu diễn dữ liệu

- Độ đo khoảng cách
- Biến trị khoảng
- Biến nhị phân đối xứng và bất đối xứng
- Biến định danh
- Biến thứ tự
- Biến tỷ lệ theo khoảng
- Biến có kiểu hỗn hợp

Độ đo khoảng cách

- Sử dụng độ đo khoảng cách để đánh giá độ tương tự giữa các điểm dữ liệu
- Không có một độ đo nào có thể dùng chung cho mọi trường hợp
- Tùy theo mục tiêu khảo sát và bản chất dữ liệu => chọn độ đo khoảng cách phù hợp
- Gọi K là không gian dữ liệu; x, y, z là các điểm dữ liệu tùy ý trong K. Độ đo D là hàm số d: $KxK -> R$ thỏa:
  - d(x,y) >= 0 (tính chất không âm)
  - d(x,y) = 0 nếu x = y (tính chất điểm)
  - d(x, y) = d(y,x) (tính chất đối xứng)
  - d(x, y) ≤ d(x, z) + d(z, y) (tính chất bất đẳng thức tam giác)
- Giá trị của độ đo d(x, y) càng nhỏ thì x và y càng gần nhau (càng tương tự nhau)

Biến trị khoảng

- Là độ đo liên tục của các đại lượng tuyến tính đơn giản như: trọng lượng, chiều cao, nhiệt độ, tuổi…
- Ảnh hưởng rất nhiều đến kết quả gom cụm
- Tuỳ vào lĩnh vực ứng dụng và tiêu chí của phương pháp tiếp cận mà chẩn hoá dữ liệu
- Xem xét:
  - Phương pháp chuẩn hoá các độ đo
  - Các độ đo khoảng cách thông dụng

Phương pháp chuẩn hoá các độ đo

- Tính sai số tuyệt đối trung bình:

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|)$$

- Với m_f là giá trị trung bình của các x_{if}, i = 1..n

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

- Tính độ chuẩn (z-score)

$$z_{if} = \frac{x_{if} - m_f}{S_f}$$

- Nhận xét: Số tuyệt đối trung bình càng lớn thì hiện tượng cá biệt càng giảm

Các độ đo khoảng cách thông dụng

- Khoảng cách Minkowski

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q}$$

- Khoảng cách Manhattan là khoảng cách Minkowski khi q = 1

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- Khoảng cách Euclide là khoảng cách Minkowski khi q = 2

$$d(i,j) = \sqrt[2]{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

Biến nhị phân đối xứng và bất đối xứng

- Biến nhị phân: Chỉ có 2 trạng thái 0 hoặc 1
- Biến nhị phân đối xứng: Cả hai trạng thái là tương đương (về mặt ý nghĩa của ứng dụng, có thể đổi vị trí, vai trò cho nhau)
- Biến nhị phân bất đối xứng:
  - Nếu có một trạng thái có ý nghĩa quan trọng hơn, thường có xu hướng thiên vị trạng thái ưu tiên nó
  - Ví dụ: Trong chẩn đoán y khoa:
    - Có dương tính, âm tính
    - Những trạng thái chưa rõ ràng (như triệu chứng bệnh chưa rõ ràng) thì cũng có thể kết luận là 1 để ưu tiên cho bước chẩn đoán chuyên sâu hoặc cách lý theo dõi

Biến định danh

- Là biến có thể nhận nhiều hơn hai trạng thái
- Ví dụ biến màu sắc có thể nhận đỏ, vàng, xanh, lục
- Có hai phương pháp để xác định khoảng cách theo biến định danh:
  - Hệ số đối sánh đơn giản: $d(i,j) = \frac{p-m}{p}$
    - m là số thuộc tính có giá trị trùng khớp giữa hai đối tượng i và j
    - p là tổng số thuộc tính
  - Chuyển về biến nhị phân bằng cách thay mỗi trạng thái định danh bằng một biến nhị phân mới:
    - Ví dụ biến màu sắc (đỏ, vàng, xanh, lục) có thể chuyển thành bốn biến nhị phân: đỏ (có/không), vàng (có/không), xanh (có/không), lục (có/không)

Biến thứ tự

- Là biến trên một tập giá trị có xác định quan hệ thứ tự trên đó
- Có thể rời rạc hoặc liên tục
- Ví dụ hạng xếp loại huy chương vàng, bạc, đồng
- Xây dựng độ đo cho biến thứ tự x
- Thay thế x bởi hạng của $x \in \{1,2, \dots, M\}$
- Ánh xạ hạng của từng biến vào [0, 1] bởi: $z = \frac{x-1}{M-1}$
- Tính độ phân biệt theo các phương pháp đã biết đối với biến z: ví dụ khoảng cách Euclide

Biến tỷ lệ theo khoảng

- Là độ đo dương trên các tỉ lệ phi tuyến
- Ví dụ:
  - Các đại lượng biểu diễn theo hàm mũ chẳng hạn Ae^{Bt}
- Trong đa số trường hợp không thể áp dụng trực tiếp phương pháp độ đo cho các biến trị khoảng cho loại biến này vì có thể gây sai số lớn
- Phương pháp tốt hơn:
  - B1: Tiền xử lý bằng cách chuyển sang logarit y = log(x)
  - B2: Áp dụng trực tiếp theo phương pháp độ đo cho các biến trị khoảng hay thứ tự mới nhận được từ B1

Biến có kiểu hỗn hợp

- CSDL có thể chứa cả sáu loại biến đơn nêu trên
- Ta có thể dùng công thức được gán trọng để kết hợp các hiệu quả của các biến thành phần:

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

- δ_{ij} tính như sau:
  - $\delta_{ij} = 0$ nếu x_i hoặc x_j không tồn tại hoặc $x_i = x_j = 0$
  - $\delta_{ij} = 1$ trong các trường hợp khác
- d_{ij} tính như sau:
  - Đối với các biến trị khoảng hoặc thứ tự: d_{ij} là khoảng cách đã chẩn hoá
  - Đối với các biến nhị phân hoặc định danh
    - $d_{ij} = 0$ khi $x_i = x_j = 0$
    - $d_{ij} = 1$ trong các trường hợp khác

Kỹ thuật phân hoạch dữ liệu

- Khái niệm phân hoạch dữ liệu
- Ý tưởng của kỹ thuật phân hoạch
- Một số thuật toán phân cụm:
  - Thuật toán K-Means
  - Thuật toán AGNES
- Một số tiếp cận phân cụm khác:
  - Tiếp cận dựa trên mật độ
  - Tiếp cận dựa trên mô hình
  - Tiếp cận dựa trên lưới

Khái niệm phân hoạch dữ liệu

- Phân cụm dữ liệu chia thành hai loại chính là: phân hoạch và phân cấp
- Phân hoạch: Cho một CSDL có n đối tượng
  - Thực hiện phân hoạch tạo ra k vùng dữ liệu, mỗi vùng được tối ưu theo một tiêu chuẩn nào đó
  - Thường xuất phát từ việc khởi tạo ngẫu nhiên, sau đó tìm cách tiến đến cực trị của hàm mục tiêu
- Phân cấp:
  - Thuật toán phân cấp tạo ra các vùng có đối tượng đươợc phân cấp
  - Có hai kiểu phân cấp: từ dưới lên và phân cấp từ trên xuống
  - Phân cấp từ dưới lên:
    - Khởi tạo mỗi đối tượng là một lớp; Tiếp tục ghép các nhóm theo một phép đo khoảng cách nào đó
    - Thuật toán dừng khi tất cả các đối tượng thuộc một nhóm thỏa điều kiện "đồng nhất" nào đó
  - Phân cấp từ trên xuống thì ngược lại: Phân chia nhóm nếu không thỏa mãn điều kiện "đồng nhất"

Ý tưởng của kỹ thuật phân hoạch

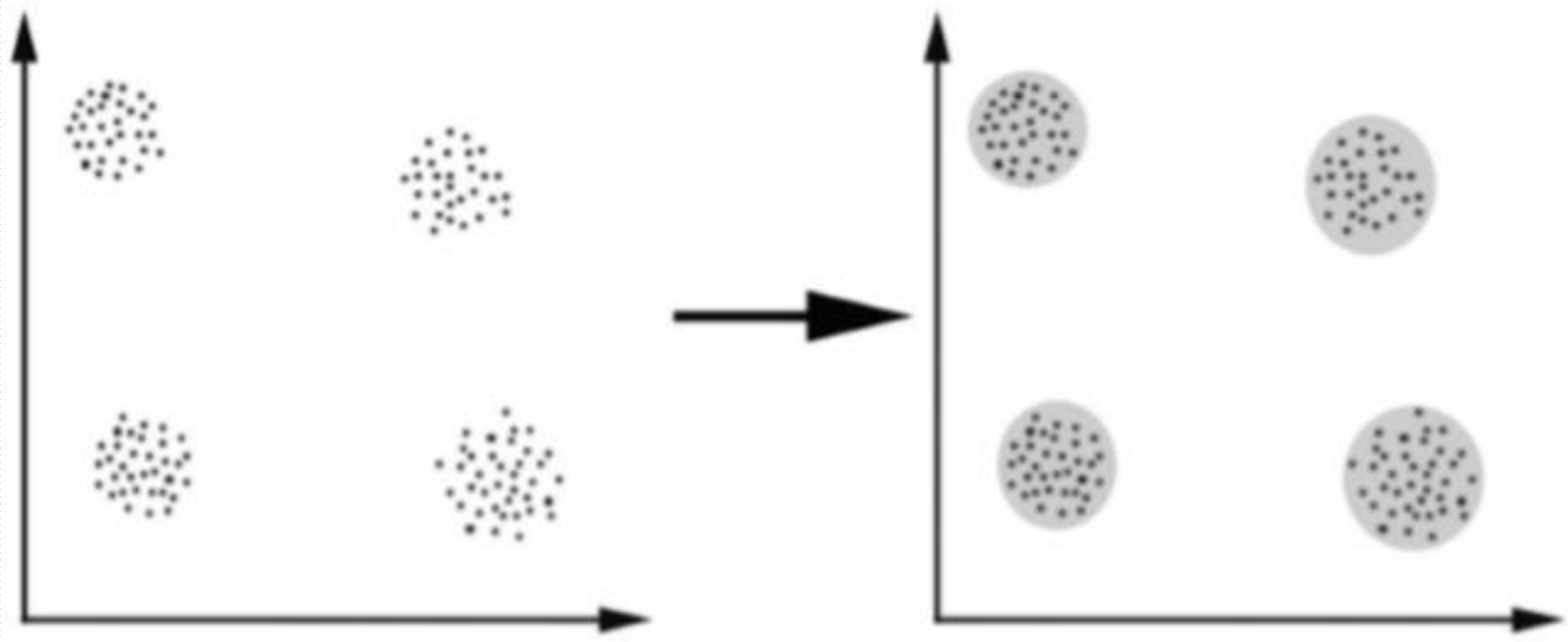
- Mục tiêu:
  - Phân hoạch cơ sở dữ liệu D có n đối tượng thành k cụm
- Ý tưởng phân hoạch truyền thống:
  - i) Mỗi cụm chứa ít nhất một đối tượng
  - ii) Mỗi đối tượng thuộc về một cụm duy nhất
  - iii) K là số cụm đã được cho trước
- Gần đây xuất hiện phương pháp phân hoạch dựa trên lý thuyết tập mờ:
  - Tiêu chuẩn (ii) là không quan trọng => thay bằng mức độ thuộc về (membership) của đối tượng vào cụm
  - Mức độ này có thể có giá trị liên tục từ 0 đến 1

Thuật toán K-Means

- Giới thiệu thuật toán K-Means
- Phát biểu bài toán phân cụm
- Các bước thuật toán K-Means
- Điều kiện dừng và chất lượng phân cụm
- Nhận xét thuật toán K-Means

Phát biểu bài toán phân cụm

- Input: n đối tượng và số các cụm k
- Output: Các cụm C_i (i=1... k) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu



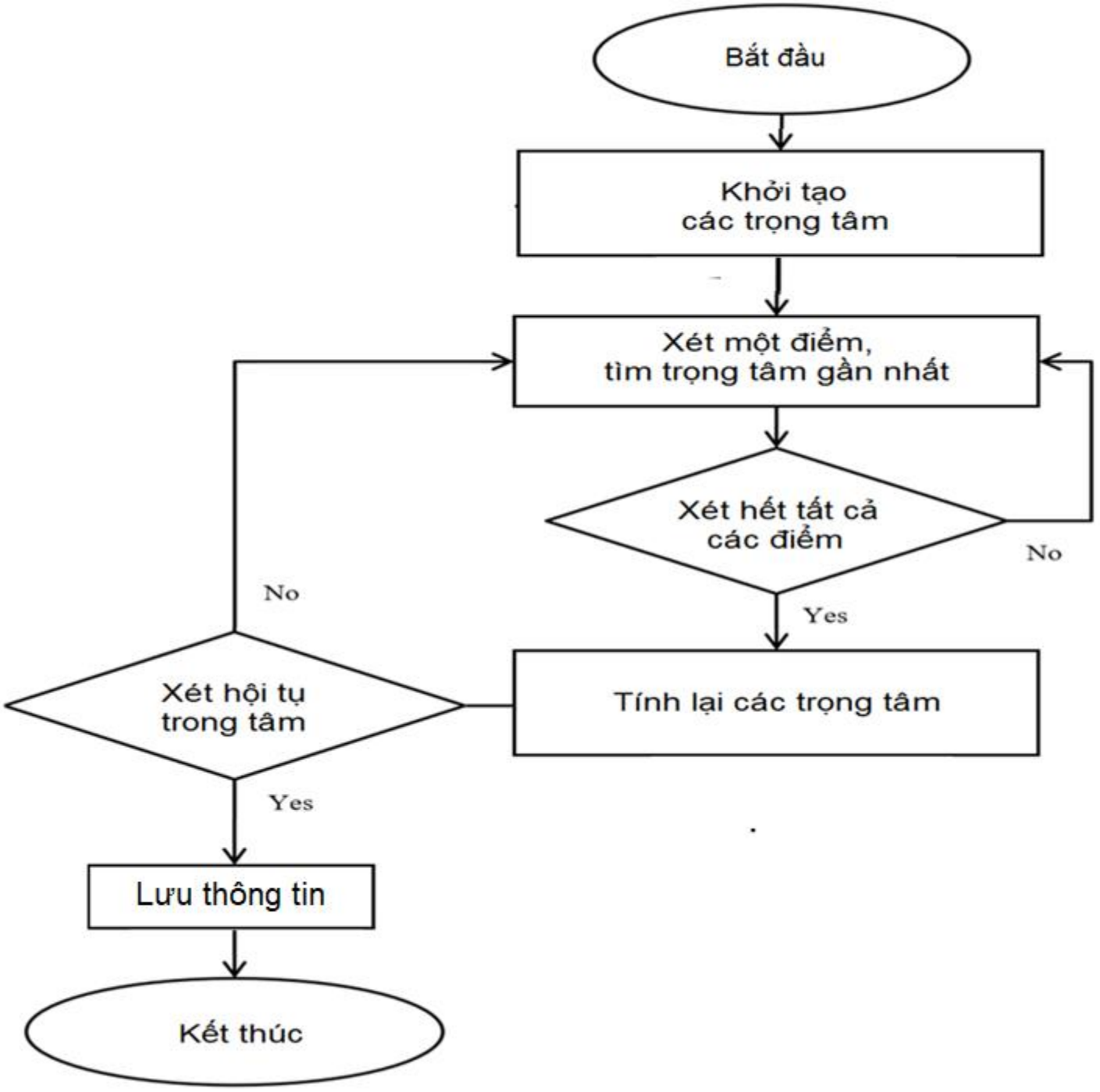
Giới thiệu thuật toán K-Means

- Là một trong các thuật toán phân cụm đơn giản và điển hình nhất
- Do MacQueen đề xuất trong lĩnh vực thống kê năm 1967
- Mục đích:
  - Sinh ra k cụm dữ liệu từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian p chiều $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, i = 1..n, sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu
    - $E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i)^2$
  - m_i là vector trọng tâm của cụm C_i, giá trị của mỗi phần tử là trung bình cộng các thành phần tương ứng của các đối tượng vector dữ liệu trong cụm đang xét
  - d là khoảng cách Euclide giữa hai đối tượng

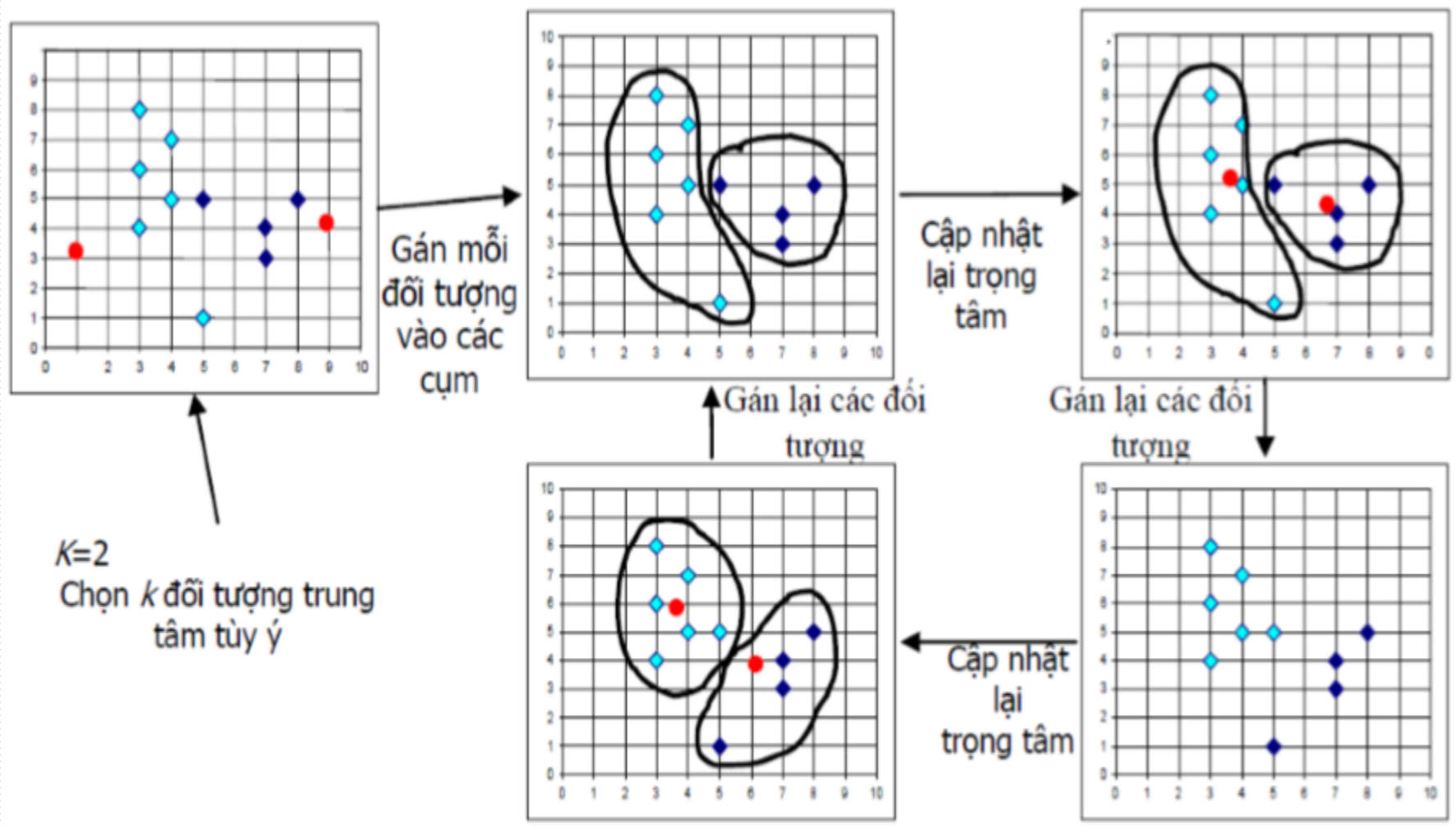
Các bước thuật toán K-Means

- Bước 1: Khởi tạo tâm cụm
  - Chọn k đối tượng m_j (j=1...k) là trọng tâm ban đầu của k cụm từ tập dữ liệu (Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm)
- Bước 2: Tính toán khoảng cách và gán cụm
  - Với mỗi đối tượng Xi (1 i n), tính toán khoảng cách từ nó tới mỗi trọng tâm mj với j=1,..,k, sau đó tìm trọng tâm gần nhất đối với mỗi đối tượng
- Bước 3: Cập nhật lại trọng tâm
  - Với mỗi j=1,..,k, cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu
- Bước 4: Kiểm tra điều kiện dừng
  - Lặp các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi

Lưu đồ thuật toán K-Means



Minh họa quá trình phân cụm



Điều kiện dừng và chất lượng phân cụm

- Điều kiện dừng
  - Không có (hoặc có không đáng kể) việc gán lại các ví dụ vào các cụm khác
  - Không có (hoặc có không đáng kể) thay đổi về các điểm trung tâm (centroids) của các cụm
  - Không giảm hoặc giảm không đáng kể về tổng lỗi phân cụm E
- Chất lượng phân cụm
  - Phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k trọng tâm khởi tạo ban đầu
  - Nếu các trọng tâm khởi tạo ban đầu quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của k-means là rất thấp => các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế

Nhận xét thuật toán K-Means

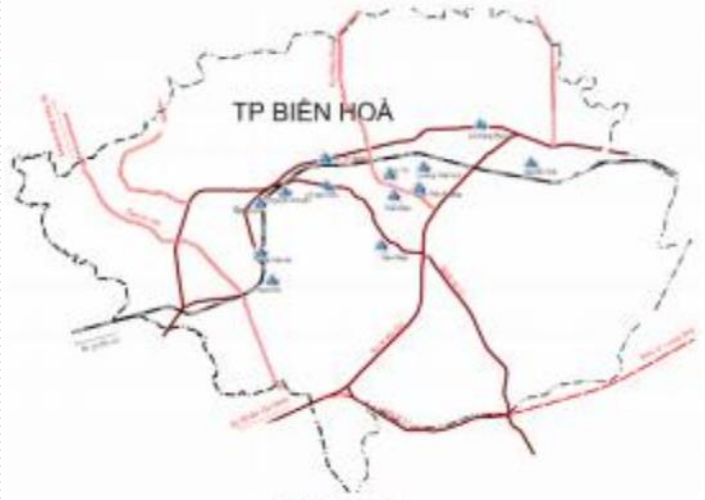
- Phần lớn khối lượng tính toán tập trung ở bước 2: tính khoảng cách từ mỗi điểm (đối tượng) tới các tâm cụm
- Số lượng đối tượng trong tập dữ liệu càng lớn, thời gian cần cho bước này càng nhiều
- Việc tính toán khoảng cách từ một điểm tới tâm cụm là độc lập, không phụ thuộc vào điểm khác
- => Việc tính khoảng cách từ các điểm đến các tâm cụm có thể thực hiện song song, đồng thời với nhau

Thuật toán AGNES

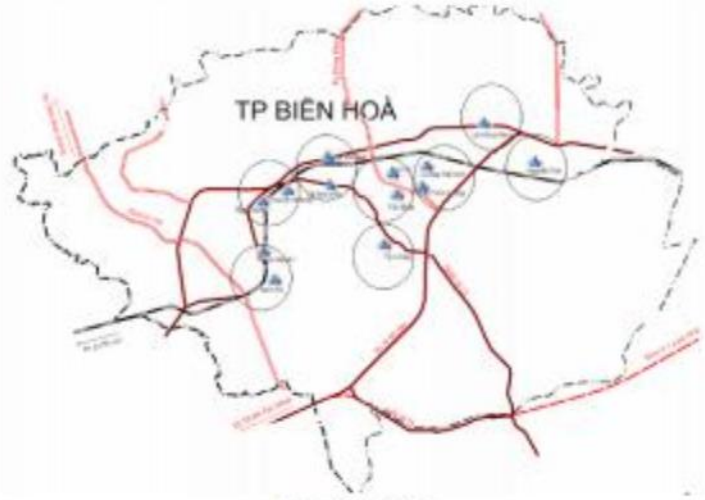
- Là kỹ thuật kiểu tích tụ
- Bắt đầu với mỗi đối tượng dữ liệu trong các cụm riêng lẻ
- Các cụm được nhập theo một số loại của cơ sở luật, cho đến khi chỉ có một cụm ở đỉnh của phân cấp, hoặc gặp điều kiện dừng
- Thuộc loại phân cụm phân cấp, theo tiếp cận Bottom – up:
  - Bắt đầu ở dưới với các nút lá trong các cụm riêng lẻ
  - Tiếp tục duyệt lên trên phân cấp tới nút gốc - là cụm đơn cuối cùng với tất cả các đối tượng dữ liệu chứa trong cụm đó
- Các bước thực hiện:
  - B1: Mỗi đối tượng là một nhóm
  - B2: Hợp nhất các nhóm có khoảng cách giữa các nhóm là nhỏ nhất
  - B3: Nếu thu được nhóm "toàn bộ" thì dừng, ngược lại quay về B2

Thuật toán AGNES

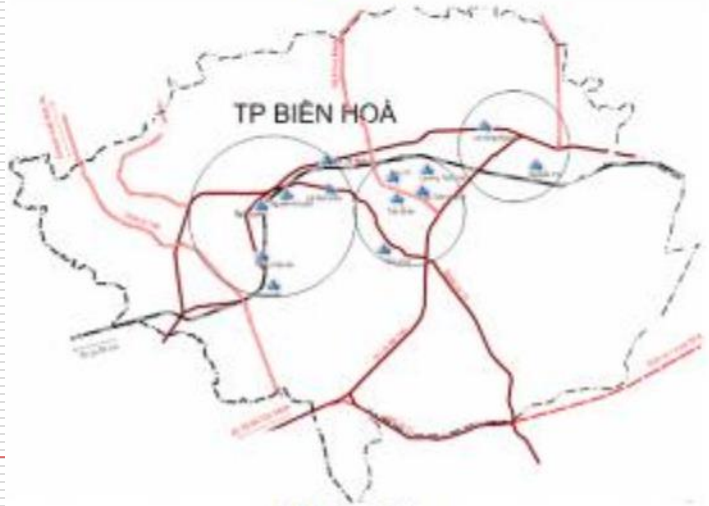
- Minh họa thuật toán:



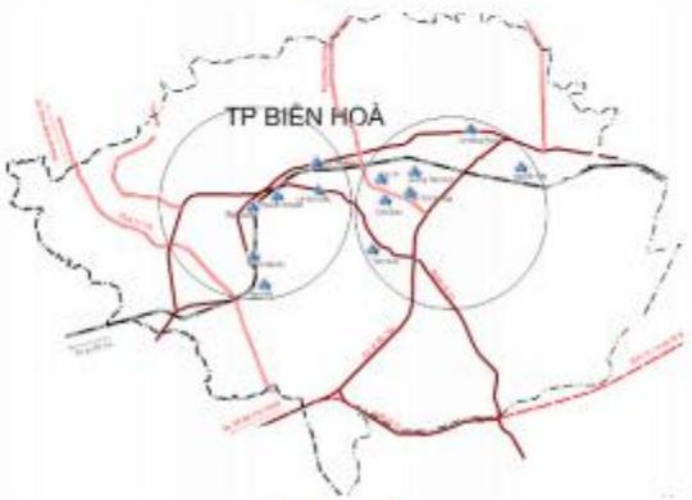
Bước 1



Bước 2a



Bước 2b



Bước 3

Tiếp cận dựa trên mật độ

- Các kí hiệu và khái niệm
- Ví dụ minh họa
- Thuật toán DBSCAN
- Nhận xét thuật toán DBSCAN

Các kí hiệu và khái niệm

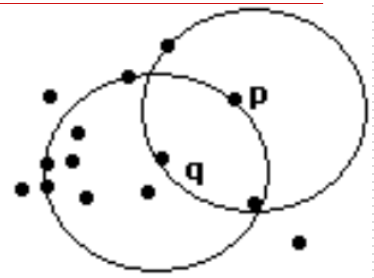
- p, q, o là các điểm dữ liệu bất kỳ (các đối tượng)
- Với Eps dương cho trước, tập hợp NEps(p) = {q|d(q,p) ≤ Eps} gọi là lân cận bán kính Eps của p, d(q,p) là khoảng cách Euclide
- p gọi là điểm hạt nhân nếu thoả: |NEps(p)| ≥ minPts
  - minPts: số nguyên dương cho trước, là ngưỡng tối thiểu để coi một điểm là trù mật (Số điểm tối thiểu)
  - => khi nói một điểm là hạt nhân => nó gắn với một bán kính Eps và một ngưỡng trù mật minPts
- p gọi là điểm biên nếu nó không phải là điểm nhân
- q gọi là *đi tới được trực tiếp theo mật độ* từ p nếu p là một điểm nhân và q thuộc lân cận của p
- p_n gọi là *đi tới được theo mật độ* từ p_1 nếu tồn tại một dãy các điểm p_i (i = 2, …n) sao cho p_i liên thông mật độ trực tiếp từ p_{i+1}
- p và q gọi là có *kết nối theo mật độ* nều tồn tại điểm o sao cho cả p và q đều liên thông mật độ từ o

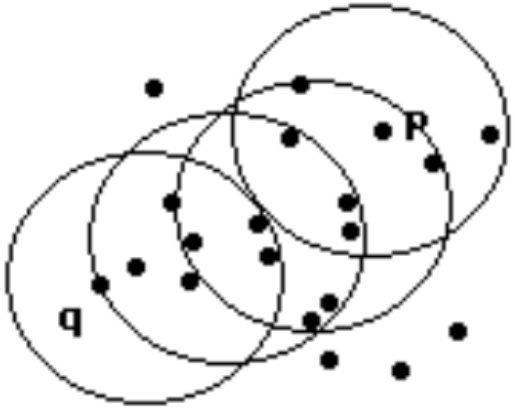
Ví dụ minh họa

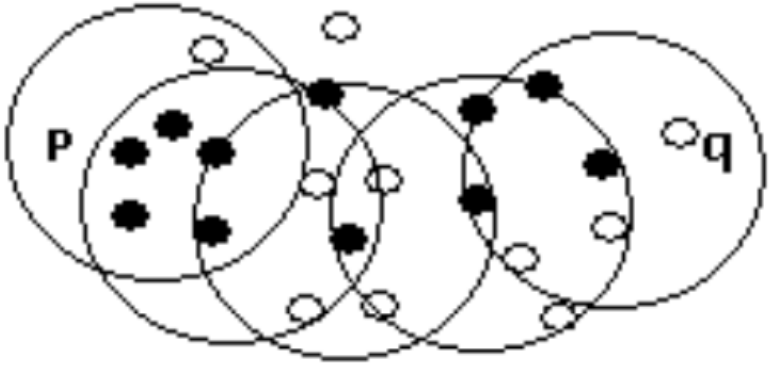
- Ví dụ 1:
  - p là một điểm hạt nhân với bán kính Eps 1cm và ngưỡng trù mật là min Pts là 3
  - q là một điểm liên thông mật độ trực tiếp từ p
- Ví dụ 2: q là một điểm liên thông mật độ từ p
- Ví dụ 3: p và q là hai điểm có kết nối mật độ
- Ý tưởng của các thuật toán dựa trên mật độ: Một cụm là một tập tối đa các điểm có kết nối mật độ

min Pts = 3

Eps = 1cm





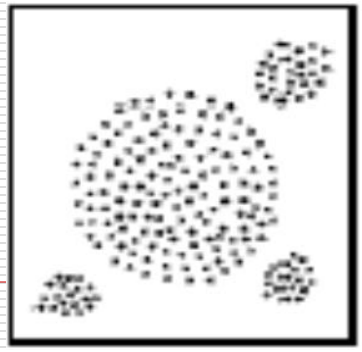
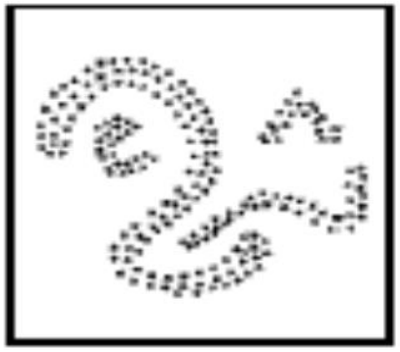
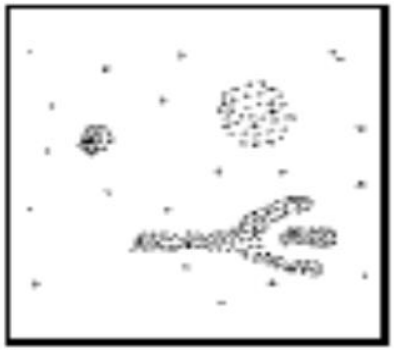


Thuật toán DBSCAN

- B1: Chọn một điểm p bất kỳ thuộc không gian dữ liệu D
- B2: Tìm tập P gồm tất cả các điểm liên thông mật độ từ p với ngưỡng bán kính Eps và ngưỡng mật độ Pts
- B3: Nếu p là một điểm hạt nhân thì
  - P chính là một cụm cần tìm
  - D = D \ P (loại P ra khỏi D)
- B4: Quay lại B1 cho đến khi tất cả các điểm trong D đều đã xét
- B5: Các điểm đã xét mà không thuộc cụm nào thì chính là các mẫu cá biệt

Nhận xét thuật toán DBSCAN

- Bán kính lân cận và ngưỡng trù mật là các tham số quyết định đến kết quả gom cụm
- Ưu điểm: Tìm được các cụm từ có hình dạng bất kỳ do nhiễu hoặc mẫu khác biệt gây ra
- Hạn chế:
  - Khó chọn được các ngưỡng EPs và minPts tốt => kết quả gom cụm không tốt khi mật độ trong các cụm tự nhiên là chênh lệch nhau nhiều
  - Không phù hợp cho yêu cầu phân cấp cụm mà chỉ đáp ứng nhu cầu phân hoạch
- Các thuật toán khác: OPTICS, DENCLUE
- Một số hình dạng cụm dữ liệu khám phá được theo tiếp cận dựa vào mật độ

Tiếp cận dựa trên mô hình

- Là tiếp cận dựa trên sự phù hợp giữa dữ liệu và các mô hình toán học
- Ý tưởng:
  - Dự liệu phát sinh từ một sự kết hợp nào đó của các phân phối xác suất ẩn
- Có hai phương pháp tiếp cận chính:
  - Tiếp cận thống kê (phương pháp COBWEB, CLASSIT, AutoClass)
  - Tiếp cận mạng nơron học cạnh tranh, bản đồ tự cấu trúc SOM

Tiếp cận dựa trên lưới

- Ý tưởng:
  - Dùng các cấu trúc dữ liệu dạng lưới với nhiều cấp độ phân giải
  - Những ô lưới có mật độ cao sẽ tạo thành những cụm
  - Tiếp cận này rất phù hợp với các phân tích trong gom cụm ứng dụng trong không gian (phân loại sao, thiên hà…)
  - Một số thuật toán khác: Sting, WaveCluster, CLIQUE


THE END