Khoa Công Nghệ Thông Tin

Trường Đại học Công Nghệ - ĐHQGHN

PHÂN CỤM DỮ LIỆU

Trần Trọng Hiếu

Khoa CNTT – Trường ĐH Công nghệ - ĐHQGHN

Nội dung

- Các khái niệm cơ bản
- Các phương pháp:
  - Dựa trên phân hoạch
  - Dựa trên phân cấp
  - Dựa trên mật độ
  - Dựa trên lưới
- Các kỹ thuật đánh giá
- Tổng kết

Các khái niệm cơ bản

Các khái niệm

- ***Cụm:*** Một tập/nhóm các đối tượng dữ liệu
  - Giống nhau(hoặc liên quan đến nhau) sẽ nằm trong một nhóm.
  - Khác nhau(hoặc không liên quan đến nhau) sẽ nằm trong các nhóm khác nhau.
- ***Phân tích cụm*** (hay *phân cụm*, *phân mảnh dữ liệu*,…)
  - Tìm những điểm giống nhau trong dữ liệu dựa vào các đặc điểm đã tìm được trong dữ liệu này và nhóm những đối tượng dữ liệu giống nhau vào các cụm.
- Tương đương với khái niệm *Học không giám sát* trong *Học máy*
- Các nhóm ứng dụng chính:
  - Là một công cụ độc lập để hiểu rõ hơn về sự phân phối dữ liệu.
  - Là một bước tiền xử lý cho các thuật toán khác.

Công cụ để hiểu dữ liệu

- Sinh học: phân loại các sinh vật: giới, ngành, lớp, bậc, họ, chi và loài
- Truy hồi thông tin: phân cụm tài liệu
- Sử dụng đất: Xác định các khu vực sử dụng đất tương tự trong cơ sở dữ liệu quan trắc trái đất
- Tiếp thị: Giúp các nhà tiếp thị khám phá các nhóm khác biệt trong cơ sở khách hàng của họ và sau đó sử dụng kiến thức này để phát triển các chương trình tiếp thị mục tiêu
- Quy hoạch thành phố: Xác định các nhóm các ngôi nhà theo loại nhà, giá trị và vị trí địa lý
- Nghiên cứu động đất: Các tâm chấn động đất quan sát được nên được phân cụm dọc theo các đứt gãy lục địa
- Khí hậu: hiểu khí hậu trái đất, tìm các mô hình khí quyển và đại dương
- Khoa học kinh tế: nghiên cứu thị trường

Công cụ để tiền xử lý dữ liệu

- Tóm tắt:
  - Tiền xử lý để phân tích hồi quy, PCA, phân lớp và luật kết hợp
- Nén:
  - Xử lý hình ảnh: lượng tử hóa vector
- Tìm K-láng giềng gần nhất:
  - Địa phương hóa tìm kiếm thành một hoặc một số ít các cụm
- Phát hiện ngoại lai
  - Những đối tượng ngoại lai thường được coi là những đối tượng "ở xa" tất cả các cụm.

Đánh giá chất lượng phân cụm

- Một phương pháp phân cụm tốt sẽ tạo ra các cụm chất lượng cao
  - Sự tương trong cùng lớp cao: sự gắn kết trong mỗi cụm
  - Sự tương đồng giữa các lớp thấp: phân biệt giữa các cụm
- Chất lượng của phương pháp phân nhóm phụ thuộc vào
  - Độ đo sự tương đồng được sử dụng bởi phương pháp
  - Các thực hiện/cài đặt
  - Khả năng khám phá ra các mẫu ẩn

Đo chất lượng phân cụm

- Chỉ số về độ giống nhau / không giống nhau
  - Sự giống nhau được thể hiện dưới dạng hàm khoảng cách, ***d(i, j)***
  - Các định nghĩa của các hàm khoảng cách thường khá khác nhau đối với các biến tỷ lệ theo khoảng, boolean, phân loại, tỷ lệ thứ tự và vectơ
  - Trọng số nên được liên kết với các biến khác nhau dựa trên ứng dụng và ngữ nghĩa dữ liệu
- Chất lượng phân cụm:
  - Thường có một hàm “chất lượng” riêng để đo “độ tốt” của một cụm.
  - Định nghĩa độ "đủ tương tự" hoặc "đủ tốt“ là khó
    - Câu trả lời thường rất chủ quan

Các vấn đề cần quan tâm

- Các tiêu chuẩn phân vùng
  - Phân vùng đơn cấp so với phân vùng
- Sự phân tách của các cụm
  - Tách bạch (exclusive) hay chồng lấn (overlaping) nhau.
- Độ đo tương tự
  - Dựa trên khoảng cách (ví dụ: Euclidian, mạng đường, vectơ) hay dựa trên sự kết nối (ví dụ: mật độ hoặc độ tiếp giáp)
- Không gian phân cụm
  - Không gian đầy đủ (khi có số chiều thấp) hay không gian con (số chiều cao).

Các yêu cầu và thách thức

- Khả năng mở rộng
  - Phân cụm tất cả dữ liệu thay vì chỉ trên các mẫu
- Khả năng đối phó với các loại thuộc tính khác nhau
  - Dạng số, dạng nhị phân, dạng phân loại, dạng thứ tự, dạng liên kết và dạng hỗn hợp của những dạng này
- Phân cụm dựa trên ràng buộc
  - Người dùng có thể cung cấp đầu vào cho các ràng buộc
  - Sử dụng kiến thức miền để xác định các tham số đầu vào
- Khả năng diễn giải và khả năng sử dụng
- Khác:
  - Khám phá các cụm có hình dạng tùy ý
  - Khả năng đối phó với dữ liệu nhiễu
  - Phân cụm gia tăng và không nhạy cảm với thứ tự đầu vào
  - Số chiều dữ liệu lớn

Các phương pháp tiếp cận (1)

- Cách tiếp cận phân hoạch:
  - Xây dựng các phân hoạch khác nhau và sau đó đánh giá chúng theo một số tiêu chí, ví dụ: giảm thiểu tổng sai số bình phương
  - Các phương pháp điển hình: k-means, k-medoids, CLARANS
- Cách tiếp cận phân cấp:
  - Tạo phân cấp theo thứ bậc của tập dữ liệu (hoặc đối tượng) bằng cách sử dụng một số tiêu chí
  - Phương pháp tiêu biểu: Diana, Agnes, BIRCH, CAMELEON

Các phương pháp tiếp cận (2)

- Phương pháp dựa trên mật độ:
  - Dựa trên các hàm kết nối và mật độ
  - Các phương pháp điển hình: DBSACN, OPTICS, DenClue
- Phương pháp dựa trên lưới:
  - Dựa trên một cấu trúc chi tiết nhiều cấp
  - Các phương pháp điển hình: STING, WaveCluster, CLIQUE
- Dựa trên mô hình:
  - Một mô hình được giả thuyết cho mỗi cụm và sau đó cố gắng tìm sự phù hợp nhất của mô hình đó với các cụm khác
  - Phương pháp điển hình: EM, SOM, COBWEB

Các phương pháp tiếp cận (3)

- Dựa trên các mẫu phổ biến:
  - Dựa trên phân tích các mẫu phổ biến
  - Các phương pháp điển hình: p-Cluster
- Do người dùng hướng dẫn hoặc dựa trên ràng buộc:
  - Phân cụm bằng cách xem xét các ràng buộc do người dùng chỉ định hoặc ứng dụng cụ thể
  - Các phương pháp điển hình: COD (chướng ngại vật), phân cụm có ràng buộc
- Phân cụm dựa trên liên kết:
  - Các đối tượng thường được liên kết với nhau theo nhiều cách khác nhau
  - Số lượng lớn các liên kết lớn có thể được sử dụng để phân cụm các đối tượng: SimRank, LinkClus

Phương pháp dựa trên phân hoạch

Khái niệm cơ bản

- Ý tưởng : Phân hoạch CSDL ***D*** gồm ***n*** đối tượng vào một tập gồm ***k*** cụm sao cho tổng bình phương các khoảng cách là nhỏ nhất.

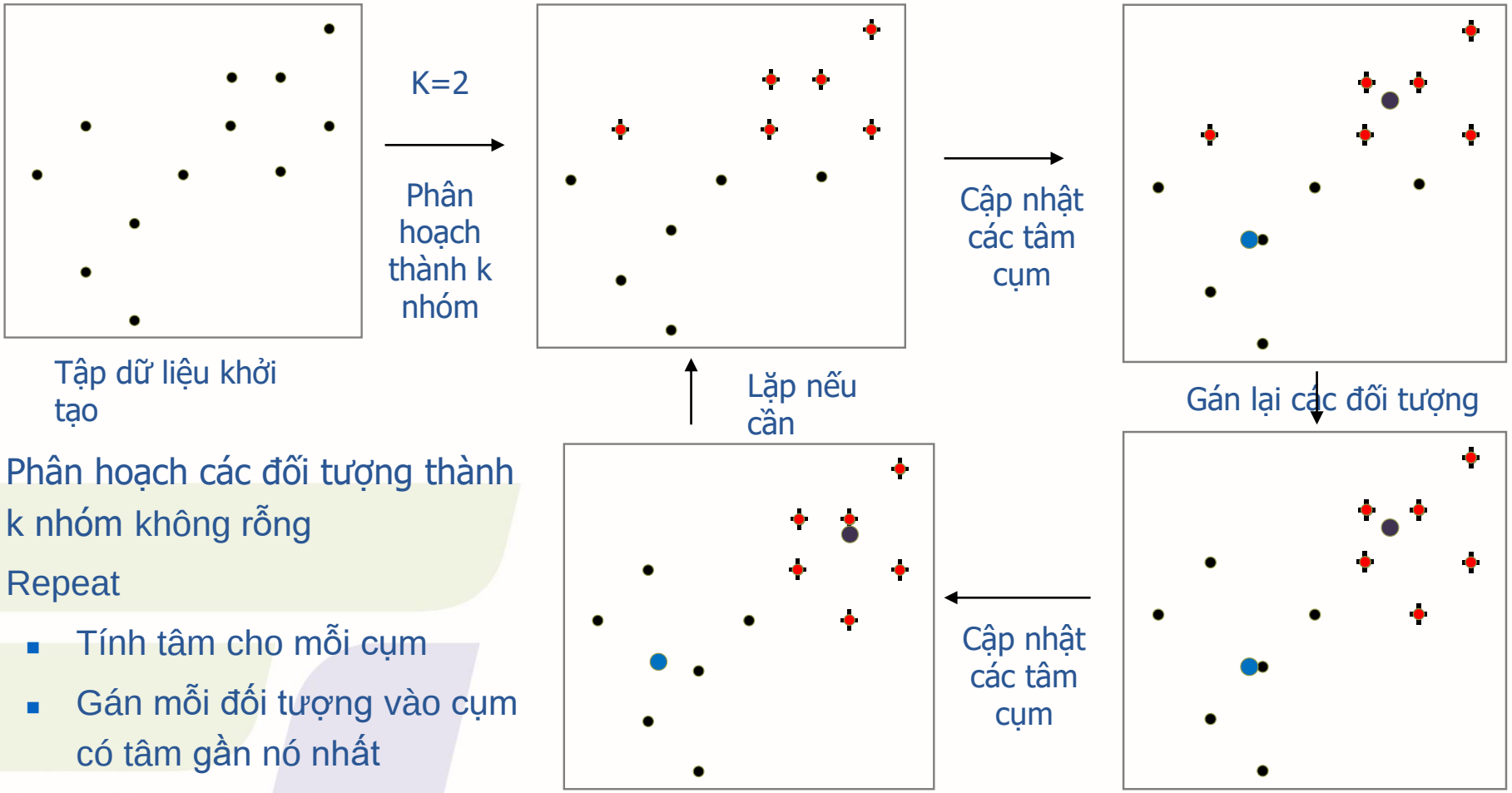
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$ (c_i là tâm của cụm C_i)

- Cho ***k***, tìm một phân hoạch gồm ***k*** cụm sao cho tối ưu hóa tiêu chuẩn phân hoạch đã chọn
- Tối ưu toàn cục: liệt kê đầy đủ tất cả các phân hoạch
- Phương pháp heuristic:
  - Thuật toán k-means (MacQueen'67, Lloyd'57 / '82): Mỗi cụm được đại diện bởi trung tâm của cụm
  - Thuật toán k-medoids hoặc PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Mỗi cụm được đại diện bởi một trong các đối tượng trong cụm

Phương pháp phân cụm K-Means

- Cho trước ***k***, thuật toán ***k***-mean được thực hiện qua bốn bước:
    1. Phân hoạch các đối tượng vào k tập hợp con không rỗng.
    2. Tính toán các điểm hạt giống là tâm cụm của mỗi tập con.
    3. Gán từng đối tượng vào cụm với điểm hạt giống gần nhất
    4. Quay lại Bước 2, dừng lại khi việc gán không còn thay đổi

Ví dụ về phân cụm K-Means



- Phân hoạch các đối tượng thành k nhóm không rỗng
- Repeat
  - Tính tâm cho mỗi cụm
  - Gán mỗi đối tượng vào cụm có tâm gần nó nhất
- Until không còn thay đổi

Nhận xét về phương pháp K-Means

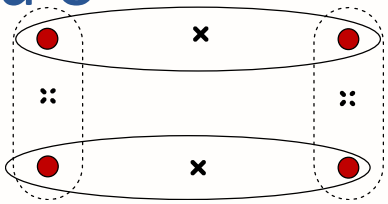
- **Điểm mạnh:** Hiệu quả: ***O(tkn)***, trong đó ***n*** là số đối tượng, ***k*** là số cụm, và ***t*** là số vòng lặp. Thông thường ***k, t*** << ***n***.
  - So sánh: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$

Thường kết thúc tại một điểm tối ưu cục bộ.

- **Điểm yếu:**
  - Chỉ áp dụng cho các đối tượng trong không gian ***n*** chiều liên tục
    - Sử dụng phương pháp k-mode cho dữ liệu phân loại
    - Trong khi đó, k-medoids có thể được áp dụng cho nhiều loại dữ liệu
  - Cần xác định trước ***k***, số lượng cụm (có nhiều cách để tự động xác định ***k*** tốt nhất (xem Hastie và cộng sự, 2009)
  - Nhạy cảm với dữ liệu nhiễu và ngoại lệ
  - Không thích hợp để khám phá các cụm có hình dạng không lồi

Các biến thể của K-means

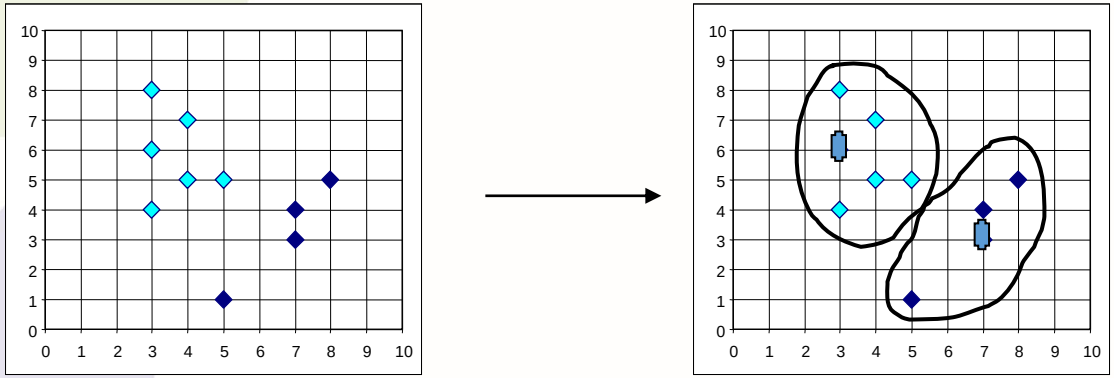
- Hầu hết các biến thể của k-means khác nhau ở
  - Lựa chọn k điểm trung bình ban đầu
  - Cách tính toán sự không giống nhau
  - Các chiến lược để tính toán trung bình của cụm



- Xử lý dữ liệu dạng phân loại: k-modes
  - Thay thế các điểm trung bình của các cụm bằng các chế độ (mode).
  - Dùng các độ đo sự không giống nhau để đối phó với các đối tượng phân loại.
  - Dùng một phương pháp dựa trên sự phổ biến để cập nhật chế độ của các cụm.
  - Hỗn hợp dữ liệu phân loại và dữ liệu số: phương pháp k-prototype.

Vấn đề đối với K-means

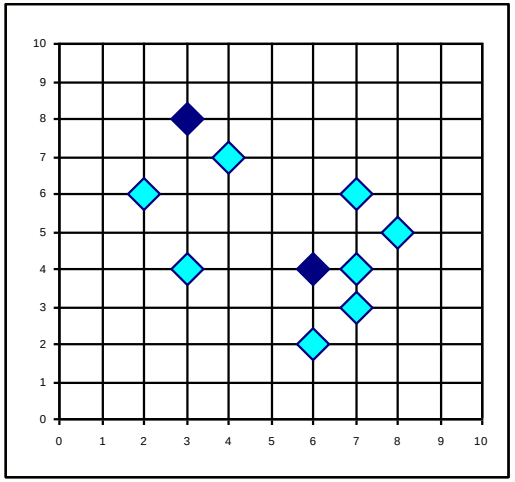
- Thuật toán k-mean rất nhạy cảm với các ngoại lệ!
- Vì một đối tượng có giá trị cực lớn có thể làm sai lệch đáng kể việc phân phối dữ liệu
- K-Medoids: Thay vì lấy giá trị trung bình của đối tượng trong một cụm làm điểm tham chiếu, có thể sử dụng medoid là đối tượng nằm ở vị trí trung tâm nhất trong một cụm.



PAM: Một thuật toán K-Medoids

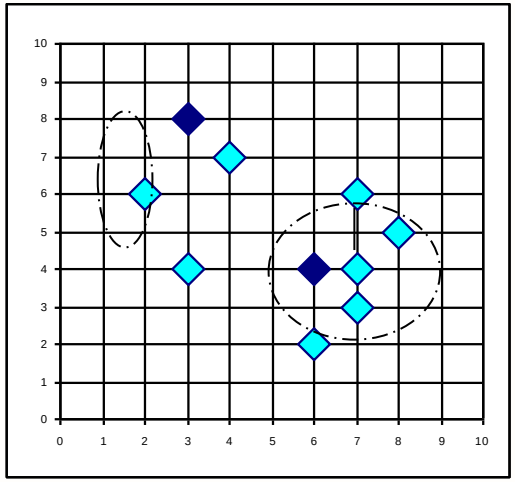
K=2

Chọn k đối tượng làm các tâm cụm

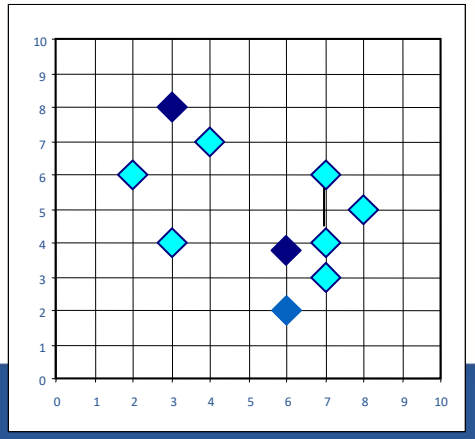


Gán mỗi đối tượng còn lại vào cụm với tâm gần nhất

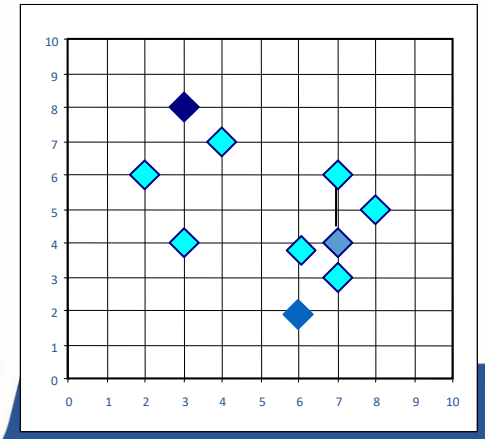
Tổng chi phí = 20



Chọn một đối tượng không là tâm O_{ramdom}



Tính tổng chi phí để hoán đổi



Tổng chi phí = 26

Hoán đổi O và O_{ramdom}

Nếu chất lượng được cải thiện.

Lặp cho đến khi không thay đổi

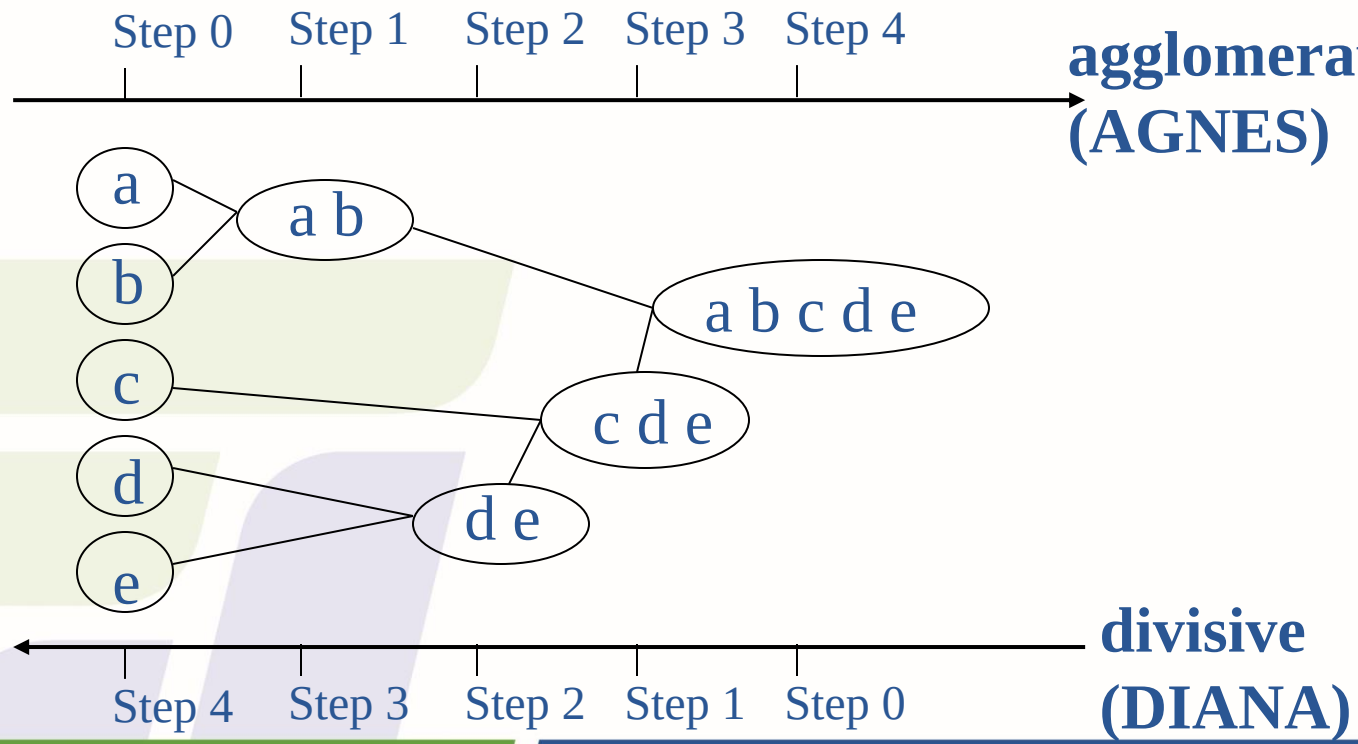
Phương pháp phân cụm K-Medoid

- Phân cụm K-Medoids: Tìm các đối tượng đại diện (medoid) trong các cụm
  - PAM (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987)
    - Bắt đầu từ một tập hợp các medoid ban đầu và thay thế lặp đi lặp lại một trong các medoid bằng một trong các đối tượng không phải medoid nếu nó cải thiện tổng khoảng cách của cụm kết quả
    - PAM hoạt động hiệu quả cho các tập dữ liệu nhỏ, nhưng không mở rộng quy mô tốt cho các tập dữ liệu lớn (do tính toán phức tạp)
- Cải thiện hiệu quả của PAM
  - CLARA (Kaufmann & Rousseeuw, 1990): PAM trên các mẫu (samples)
  - CLARANS (Ng & Han, 1994): Ngẫu nhiên hóa quá trình lấy mẫu lại

Phương pháp dựa trên phân cấp

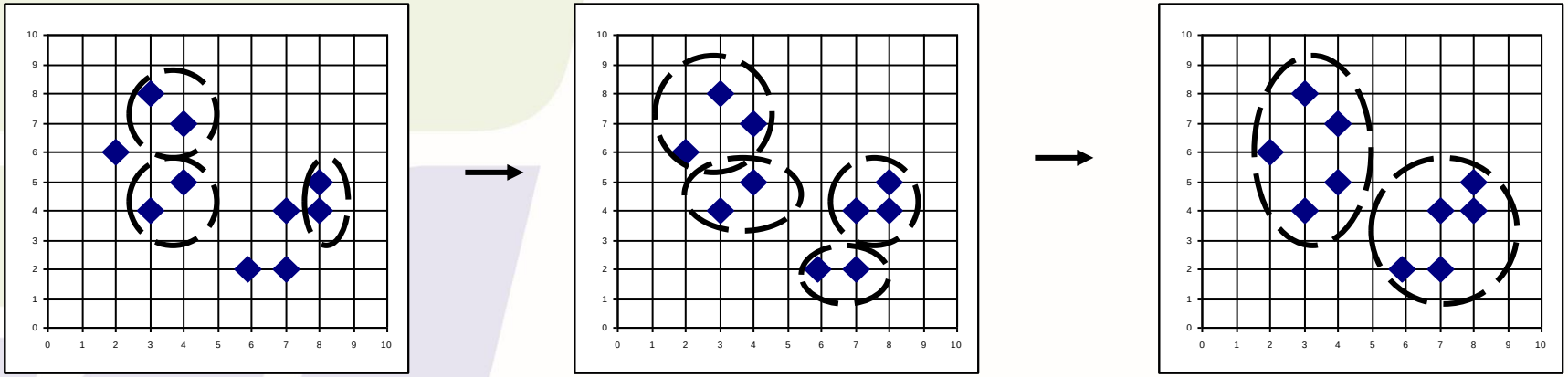
Phân cụm dựa trên phân cấp

- Dùng ma trận khoảng cách làm tiêu chí phân cụm.
- Không yêu cầu số lượng cụm ***k*** làm đầu vào, nhưng cần điều kiện kết thúc.



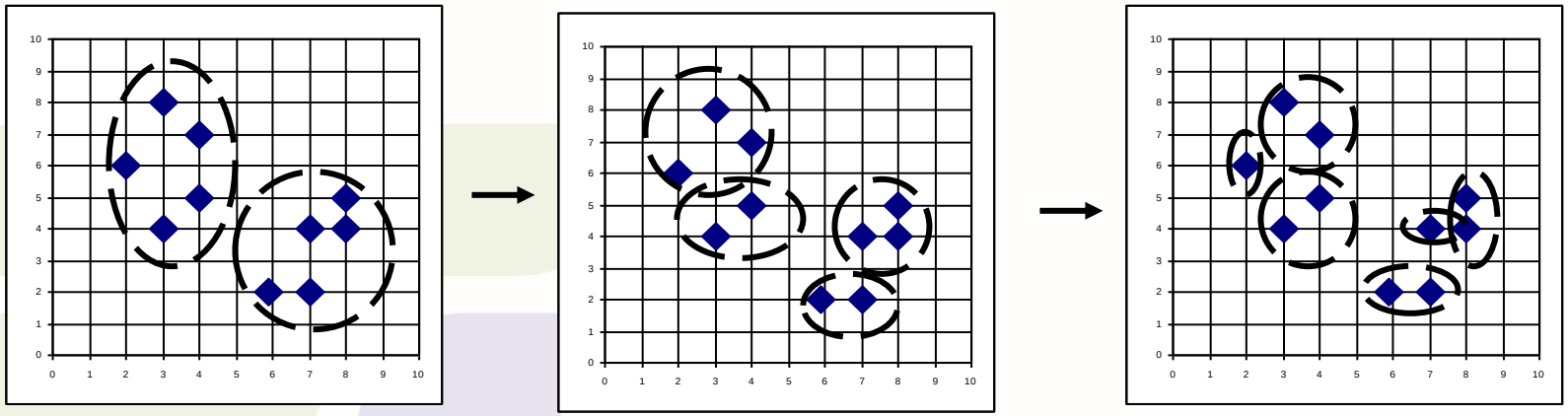
AGNES (Agglomerative Nesting)

- Được giới thiệu trong Kaufmann và Rousseeuw (1990)
- Được triển khai trong các gói thống kê, ví dụ: Splus
- Sử dụng phương pháp liên kết đơn và ma trận khác biệt
- Hợp nhất các nút có ít khác biệt nhất
- Tiếp tục theo kiểu không giảm dần
- Cuối cùng tất cả các nút thuộc cùng một cụm



DIANA

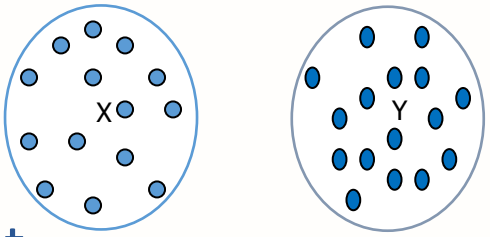
- Được giới thiệu trong Kaufmann và Rousseeuw (1990)
- Được triển khai trong các gói thống kê, ví dụ: Splus
- Đảo ngược thứ tự của AGNES
- Cuối cùng mỗi nút tạo thành một cụm



Khoảng các giữa các cụm

- ***Liên kết đơn:*** khoảng cách nhỏ nhất giữa một phần tử trong một cụm và một phần tử trong cụm khác:

  $$dist(K_i, K_j) = min(t_{ip}, t_{jq})$$



- ***Liên kết hoàn chỉnh:*** khoảng cách lớn nhất giữa một phần tử trong một cụm và một phần tử trong cụm khác: $dist(K_i, K_j) = max(t_{ip}, t_{jq})$
- ***Trung bình:*** khoảng cách trung bình giữa một phần tử trong một cụm và một phần tử trong cụm khác:

  $$dist(K_i, K_j) = avg(t_{ip}, t_{jq})$$

- ***Centroid:*** khoảng cách giữa các tâm của hai cụm:

  $$dist(K_i, K_j) = dist(C_i, C_j)$$

- ***Medoid:*** khoảng cách giữa các medoid của hai cụm:

  $$dist(K_i, K_j) = dist(M_i, M_j)$$

Tâm, bán kính và đường kính

- ***Tâm:*** "điểm giữa" của một cụm

$$C_m = \frac{\Sigma_{i=1}^{N}(t_{ip})}{N}$$

- ***Bán kính:*** Căn bậc hai của trung bình bình phương khoảng cách từ tất cả các điểm đến tâm.

$$R_m = \sqrt{\frac{\Sigma_{i=1}^{N}(t_{ip} - c_m)^2}{N}}$$

- ***Đường kính:*** Căn bậc hai của trung bình bình phương khoảng cách của mọi cặp điểm trong cụm.

$$D_m = \sqrt{\frac{\Sigma_{i=1}^{N}\Sigma_{i=1}^{N}(t_{ip} - t_{iq})^2}{N(N-1)}}$$

Các mở rộng

- Điểm yếu chính của các phương pháp phân cụm tổ hợp
- Không bao giờ có thể hoàn tác những gì đã làm trước đó
- Không mở rộng quy mô tốt: độ phức tạp về thời gian ít nhất là $O(n^2)$, trong đó n là tổng số đối tượng.
- Tích hợp phân cụm dựa trên khoảng cách và phân cấp (Phần tự đọc thêm)
  - BIRCH (1996): sử dụng cây CF và từng bước điều chỉnh chất lượng của các cụm con
  - CHAMELEON (1999): phân cụm phân cấp sử dụng mô hình động

Phương pháp dựa trên mật độ

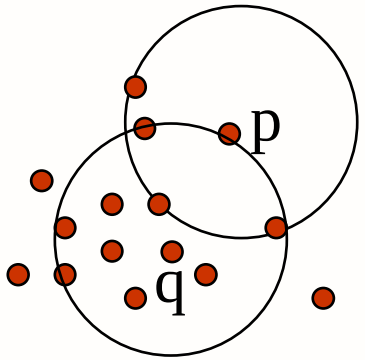
Giới thiệu

- Phân cụm dựa trên mật độ (tiêu chí cụm cục bộ), chẳng hạn như các điểm được kết nối với mật độ
- Các tính năng chính:
  - Khám phá các cụm có hình dạng tùy ý
  - Xử lý được nhiễu
  - Một lần quét
  - Cần các thông số mật độ làm điều kiện kết thúc
- Một số nghiên cứu:
  - *DBSCAN: Ester, et al. (KDD'96)*
  - *OPTICS: Ankerst, et al (SIGMOD'99).*
  - *DENCLUE: Hinneburg & D. Keim (KDD'98)*
  - *CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)*

Các khái niệm

- Hai tham số:
  - ***Eps:*** Bán kính lớn nhất của vùng lân cận
  - ***MinPts***: Số điểm tối thiểu trong vùng lân cận xác định bởi ***Eps*** của điểm đó
- ***NEps(p)***: {***q*** thuộc ***D*** | ***dist (p, q) ≤ Eps***}
- ***Tiếp cận trực tiếp theo mật độ:*** Điểm ***p*** có thể tiếp cận theo mật độ một cách trực tiếp từ điểm ***q*** đối với ***Eps, MinPts*** nếu
  - p thuộc NEps (q)
  - điều kiện cốt lõi:

    #|NEps (q)| ≥ MinPts

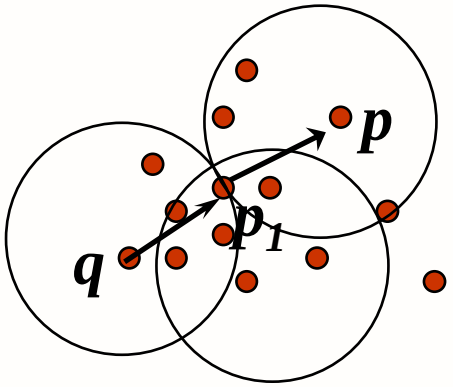


MinPts = 5

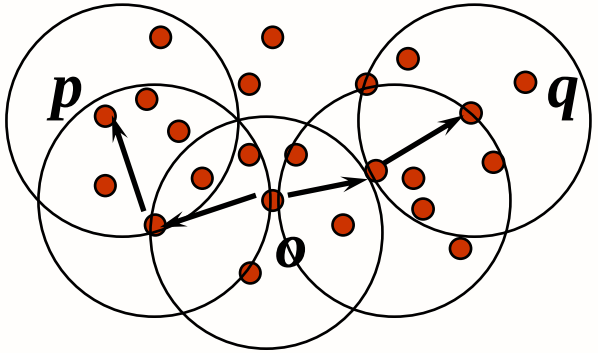
Eps = 1 cm

Tiếp cận và kết nối theo mật độ

- **Tiếp cận theo mật độ:**
  - Điểm ***p*** có thể *tiếp cận theo mật độ* từ điểm ***q*** đối với ***Eps, MinPts*** nếu có một chuỗi các điểm $p_1,\ldots, p_n$, $p_1 = q$, $p_n = p$ sao cho $p_i + 1$ có thể tiếp cận trực tiếp theo mật độ từ p_i.

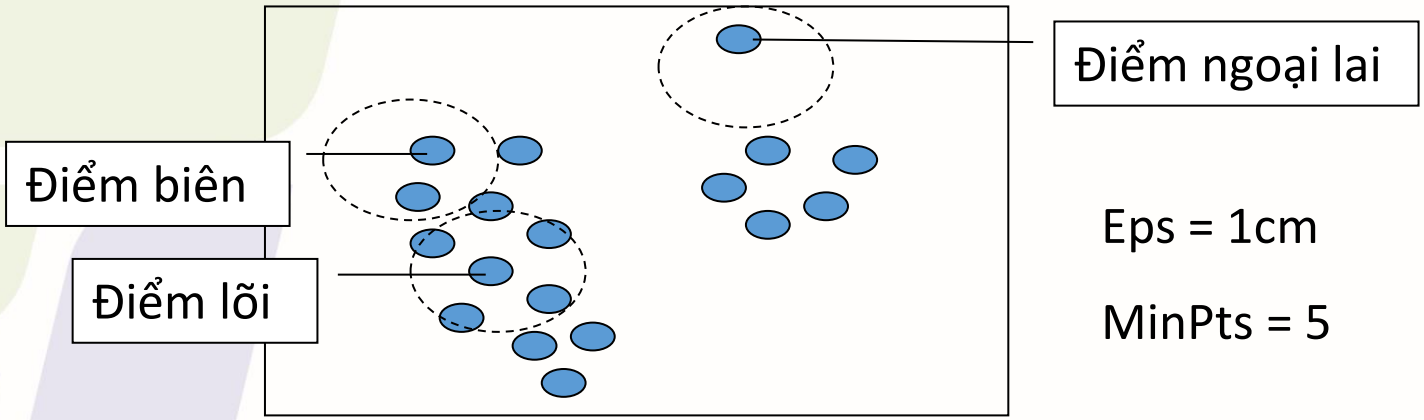


- **Kết nối theo mật độ:**
  - Một điểm ***p*** có thể kết nối theo mật độ tới một điểm ***q*** đối với ***Eps, MinPts*** nếu có một điểm ***o*** sao cho cả ***p*** và ***q*** đều có thể tiếp cận theo mật độ từ ***o*** đối với ***Eps*** và ***MinPts***



DBSCAN

- ***DBSCAN** (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise)* - Phân cụm không gian dựa trên mật độ của các ứng dụng có nhiễu.
  - Dựa trên khái niệm mật độ của cụm:
    - Một cụm được định nghĩa là một tập hợp cực đại các điểm được kết nối theo mật độ
  - Khám phá các cụm *có hình dạng tùy ý* trong cơ sở dữ liệu không gian có nhiễu.

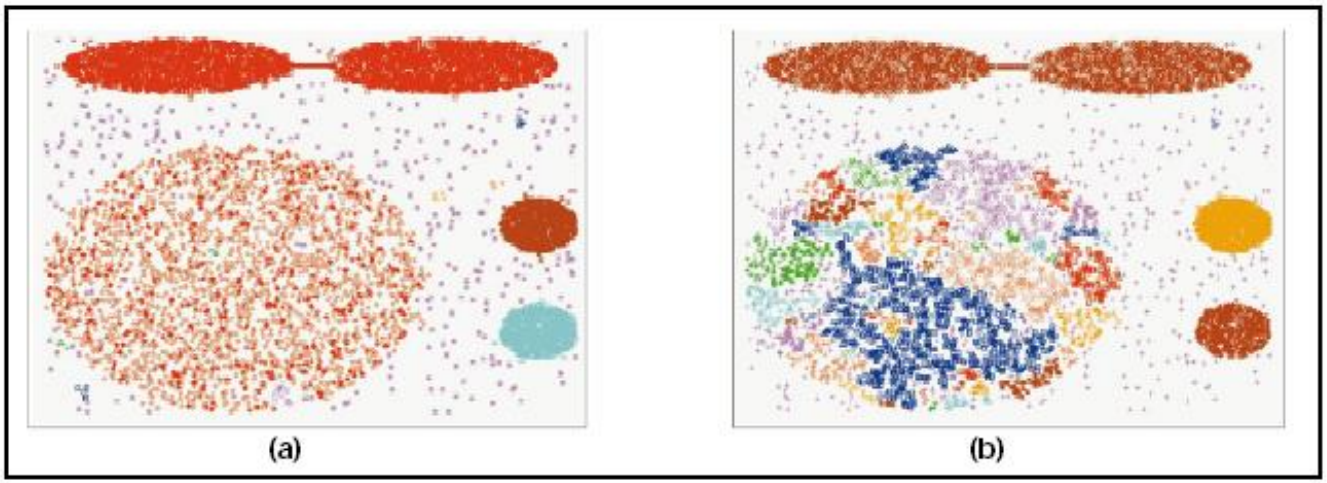


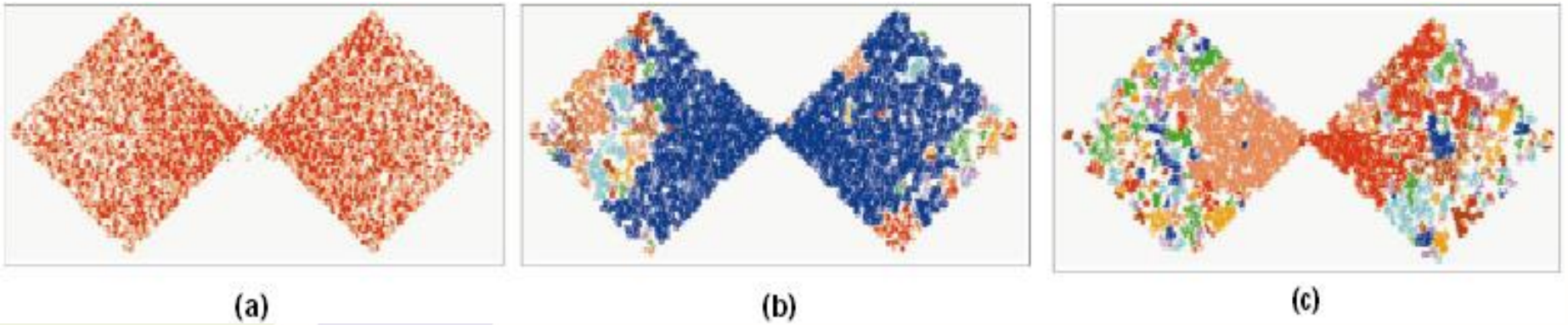
DBSCAN: Thuật toán

1. Chọn tùy ý một điểm ***p***
2. Truy xuất tất cả điểm có thể tiếp cận theo mật độ từ ***p*** đối với ***Eps*** và ***MinPts***
3. Nếu ***p*** là một điểm lõi, một cụm được hình thành
4. Nếu ***p*** là điểm biên, không có điểm nào có thể tiếp cận theo mật độ từ ***p*** và DBSCAN sẽ truy cập điểm tiếp theo của cơ sở dữ liệu.
5. Tiếp tục quá trình cho đến khi tất cả các điểm đã được xử lý

DBSCAN: Nhạy cảm theo tham số

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.





Các phương pháp mở rộng

(Tự đọc)

- OPTICS: Ordering Points To Identify the Clustering Structure
- Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
- DENCLUE: Using Statistical Density Functions
- DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)

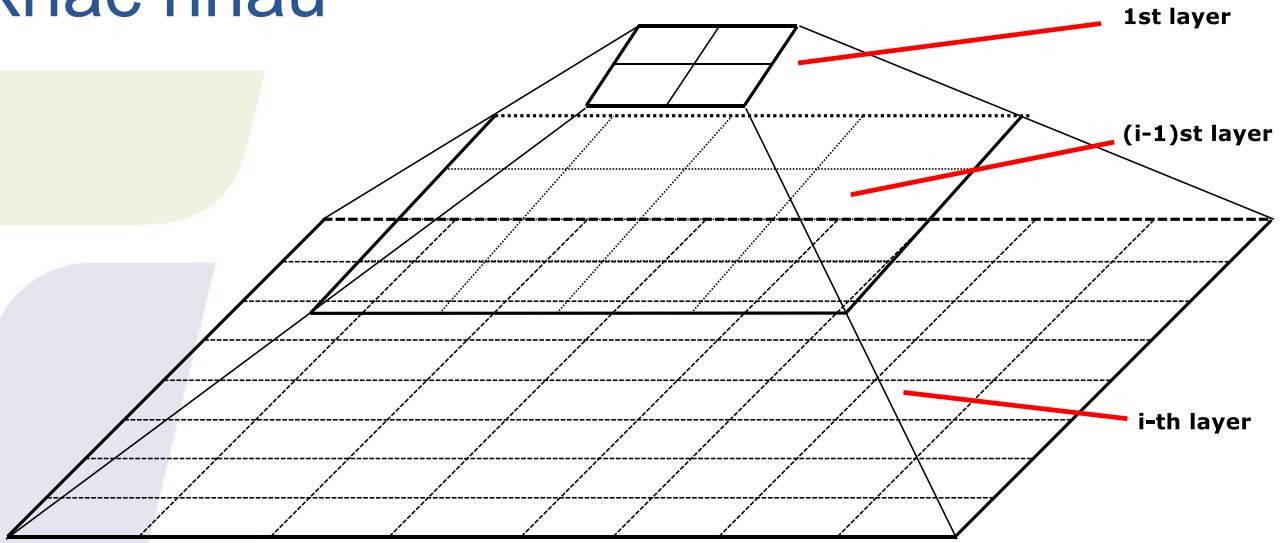
Phương pháp dựa trên lưới

Phân cụm dựa trên lưới

- Sử dụng cấu trúc dữ liệu lưới đa độ phân giải
- Một số phương pháp:
  - STING (phương pháp tiếp cận Lưới thông tin thống kê) của Wang, Yang và Muntz (1997)
  - WaveCluster của Sheikholeslami, Chatterjee và Zhang (VLDB’98)
    - Cách tiếp cận phân cụm đa độ phân giải sử dụng phương pháp wavelet
  - CLIQUE: Agrawal, et al. (SIGMOD’98)
    - Cả phân cụm dựa trên lưới và không gian con

STING

- STING: *A **S**tatistical **In**formation **G**rid Approach*
- Wang, Yang và Muntz (VLDB'97)
- Một vùng không gian được chia thành các ô hình chữ nhật
- Có một số cấp độ của các ô tương ứng với các độ phân giải khác nhau



STING

- Mỗi ô ở cấp cao được phân chia thành một số ô nhỏ hơn ở cấp thấp hơn tiếp theo
- Thông tin thống kê của mỗi ô được tính toán và lưu trữ trước và được sử dụng để trả lời các truy vấn
- Các tham số của ô cấp cao hơn có thể dễ dàng tính toán từ các tham số của ô cấp thấp hơn
  - ***count, mean, s, min, max***
  - kiểu phân phối — ***normal, uniform***, v.v.
- Sử dụng phương pháp từ trên xuống để trả lời các truy vấn dữ liệu không gian
- Bắt đầu từ một lớp được chọn trước — thường có số lượng nhỏ ô
- Đối với mỗi ô ở cấp hiện tại, tính toán khoảng tin cậy.

Thuật toán STING

- Loại bỏ các ô không liên quan để xem xét thêm
- Khi hoàn tất việc kiểm tra lớp hiện tại thì chuyển sang lớp có cấp thấp hơn tiếp theo
- Lặp lại quá trình này cho đến khi đạt đến lớp dưới cùng
- Ưu điểm:
  - Không phụ thuộc vào truy vấn, dễ dàng song song hóa , cập nhật gia tăng
  - O(K), trong đó K là số ô lưới ở mức thấp nhất
- Nhược điểm:
  - Tất cả các ranh giới cụm đều nằm ngang hoặc dọc và không có ranh giới đường chéo nào được phát hiện

Phương pháp CLIQUE

(Tự đọc)

Các kỹ thuật đánh giá

Đánh giá xu hướng phân cụm

- Đánh giá xem có tồn tại cấu trúc không ngẫu nhiên trong dữ liệu hay không bằng cách đo xác suất dữ liệu được tạo ra bởi một phân phối dữ liệu đồng nhất
- Kiểm tra tính ngẫu nhiên trong không gian bằng thử nghiệm thống kê: Hopkins Static
  - Cho một tập dữ liệu ***D*** được coi là mẫu của một biến ngẫu nhiên ***o***, hãy xác định xem ***o*** còn bao xa để được phân phối đồng nhất trong không gian dữ liệu
  - Lấy mẫu ***n*** điểm, $p_1, \ldots, p_n$, đồng nhất từ ***D***. Với mỗi số p_i, hãy tìm lân cận gần nhất của nó trong ***D***: $x_i = min\{dist(p_i, v)\}$ trong đó ***v*** trong ***D***
  - Lấy mẫu ***n*** điểm, $q_1, \ldots, q_n$, đồng nhất từ ***D***. Với mỗi q_i, tìm lân cận gần nhất của nó trong D - $\{q_i\}$: $y_i = min\{dist(q_i, v)\}$ trong đó ***v*** thuộc ***D*** và $v \neq q_i$
  - Tính toán Thống kê Hopkins: $$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$
  - Nếu ***D*** phân bố đều, $\sum x_i$ và $\sum y_i$ sẽ gần nhau và ***H*** gần bằng 0,5. Nếu ***D*** lệch nhiều thì ***H*** gần bằng 0

Xác định số các cụm

- Phương pháp thực nghiệm
  - Số cụm ≈ $\sqrt{n}$ / 2 cho tập dữ liệu n điểm
- Phương pháp khuỷu tay
  - Sử dụng bước ngoặt trong đường cong của tổng trong phương sai cụm đối với số cụm
- Phương pháp kiểm tra chéo
  - Chia một tập dữ liệu đã cho thành m phần
  - Sử dụng $m - 1$ phần để có được mô hình phân cụm
  - Sử dụng phần còn lại để kiểm tra chất lượng của phân cụm

Đo chất lượng phân cụm

Hai nhóm phương pháp: bên ngoài và bên trong

- Bên ngoài: có giám sát, tức là có sẵn ground truth
  - So sánh phân cụm với ground truth bằng cách sử dụng độ đo chất lượng phân cụm
  - Ví dụ: Các độ đo hồi tưởng và chính xác BCubed
- Bên trong: không giám sát, tức là không có ground truth
  - Đánh giá mức độ tốt của một cách phân cụm bằng cách xem xét các cụm được phân tách tốt như thế nào và mức độ gọn nhẹ của các cụm
  - Ví dụ. Hệ số Silhouette

Đo chất lượng phân cụm: bên ngoài

- Phép đo chất lượng phân cụm: $Q(C, C_g)$, đối với phân nhóm C với giá trị ground truth là C_g.
- Q là tốt nếu nó thỏa mãn 4 tiêu chí cơ bản sau:
  - Tính đồng nhất của cụm: càng tinh khiết, càng tốt
  - Tính đầy đủ của cụm: nên gán các đối tượng thuộc cùng một loại trong ground truth vào cùng một cụm
  - Mớ hỗn độn: đặt một đối tượng không đồng nhất vào một cụm thuần túy sẽ bị phạt nhiều hơn so với việc đặt nó vào một mớ hỗn độn (tức là danh mục "linh tinh" hoặc "khác")
  - Duy trì cụm nhỏ: chia một danh mục nhỏ thành nhiều phần có hại hơn là chia một danh mục lớn

Tổng kết

- Phân tích cụm nhóm các đối tượng dựa trên sự giống nhau của chúng và có các ứng dụng rộng rãi
- Phép đo mức độ tương đồng có thể được tính toán cho nhiều loại dữ liệu khác nhau
- Các thuật toán phân cụm có thể được phân loại thành các phương pháp dựa trên phân hoạch, các phương pháp dựa trên phân cấp, phương pháp dựa trên mật độ, các phương pháp dựa trên lưới và các phương pháp dựa trên mô hình
- Các thuật toán K-mean và K-medoids là các thuật toán phân cụm dựa trên phân hoạch
- Birch và Chameleon là các thuật toán phân cụm dựa trên phân cấp theo xác suất
- DBSCAN, OPTICS và DENCLU là các thuật toán dựa trên mật độ
- STING và CLIQUE là các phương pháp dựa trên lưới, trong đó CLIQUE cũng là một thuật toán phân cụm trên không gian con
- Chất lượng của kết quả phân nhóm có thể được đánh giá theo nhiều cách khác nhau

Tài liệu tham khảo (1)

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD’99.
- Beil F., Ester M., Xu X.: "Frequent Term-Based Text Clustering", KDD'02
- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. VLDB’98.
- V. Ganti, J. Gehrke, R. Ramakrishan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.

Tài liệu tham khảo (2)

- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB’98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In ICDE'99, pp. 512-521, Sydney, Australia, March 1999.
- A. Hinneburg, D.l A. Keim: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD’98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
- G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. COMPUTER, 32(8): 68-75, 1999.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB’98.

Tài liệu tham khảo (3)

- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- L. Parsons, E. Haque and H. Liu, Subspace Clustering for High Dimensional Data: A Review, SIGKDD Explorations, 6(1), June 2004
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB’98.
- A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-Based Clustering in Large Databases, ICDT'01.
- A. K. H. Tung, J. Hou, and J. Han. Spatial Clustering in the Presence of Obstacles, ICDE'01
- H. Wang, W. Wang, J. Yang, and P.S. Yu. Clustering by pattern similarity in large data sets, SIGMOD’02
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB’97
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : An efficient data clustering method for very large databases. SIGMOD'96
- X. Yin, J. Han, and P. S. Yu, “LinkClus: Efficient Clustering via Heterogeneous Semantic Links”, VLDB'06