**Bài 9: Gom cụm dữ liệu (tiếp)**

# Nội dung

5.1. Tổng quan về gom cụm dữ liệu

5.2. Gom cụm dữ liệu bằng phân hoạch

5.3. Gom cụm dữ liệu bằng phân cấp

5.4. Gom cụm dữ liệu dựa trên mật độ

5.5. Gom cụm dữ liệu dựa trên mô hình

5.6. Các phương pháp gom cụm dữ liệu khác

5.7. Tóm tắt

# 5.2. Gom cụm dữ liệu bằng phân hoạch

## Đặc điểm của giải thuật k-means

* Bài toán tối ưu hóa
* Cực trị cục bộ
* Mỗi cụm được đặc trưng hóa bởi trung tâm của cụm (i.e. đối tượng trung bình (mean)).
* Không thể xác định được đối tượng trung bình???
* Số cụm k nên là bao nhiêu?
* Độ phức tạp: O(nkt)
* n là số đối tượng, k là số cụm, t là số lần lặp
* k << n, t << n
* Ảnh hưởng bởi nhiễu (các phần tử kì dị/biên)
* Không phù hợp cho việc khai phá ra các cụm có dạng không lồi (nonconvex) hay các cụm có kích thước rất khác nhau
* Kết quả gom cụm có dạng siêu cầu (hyperspherial)
* Kích thước các cụm kết quả thường đồng đều (relatively uniform sizes)
* Đánh giá kết quả gom cụm của giải thuật k-means với hai trị k1 (phương án I) và k2 (phương án II) khác nhau trên cùng tập dữ liệu mẫu cho trước
* F-measure (trị lớn khi chất lượng gom cụm tốt)
* F-measure (I) = ???
* F-measure (II) = ???

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

🡪 Kết quả đánh giá trùng với kết quả đánh giá dựa trên độ đo Entropy?

* Tính “total cost S of swapping Oj và Orandom” = ΣpCp/OiOrandom
* Tính “total cost S of swapping Oj và Orandom” = ΣpCp/OiOrandom

# 5.2. Gom cụm dữ liệu bằng phân hoạch

## Đặc điểm của giải thuật PAM (k-medoids)

* Mỗi cụm được đại diện bởi phần tử chính giữa cụm (centroid).
* Giảm thiểu sự ảnh hưởng của nhiễu (phần tử biên/kì dị/cực trị).
* Số cụm k cần được xác định trước.
* Độ phức tạp cho mỗi vòng lặp O(k(n-k)2)
* Giải thuật bị ảnh hưởng bởi kích thước tập dữ liệu.

# 5.3. Gom cụm dữ liệu bằng phân cấp

* Gom cụm dữ liệu bằng phân cấp (hierarchical clustering): nhóm các đối tượng vào cây phân cấp của các cụm
* Agglomerative: bottom-up (trộn các cụm)
* Divisive: top-down (phân tách các cụm)
* Không yêu cầu thông số nhập k (số cụm)
* Yêu cầu điều kiện dừng
* Không thể quay lui ở mỗi bước trộn/phân tách
* 5.3. Gom cụm dữ liệu bằng phân cấp
* An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting) 🡪 bottom-up
* A divisive hierarchical clustering method: DIANA (Divisive ANAlysis) 🡪 top-down
* 5.3. Gom cụm dữ liệu bằng phân cấp
* An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting)
* Khởi đầu, mỗi đối tượng tạo thành một cụm.
* Các cụm sau đó được trộn lại theo một tiêu chí nào đó.
* Cách tiếp cận single-linkage: cụm C1 và C2 được trộn lại nếu khoảng cách giữa 2 đối tượng từ C1 và C2 là ngắn nhất.
* Quá trình trộn các cụm được lặp lại đến khi tất cả các đối tượng tạo thành một cụm duy nhất.
* A divisive hierarchical clustering method: DIANA (Divisive ANAlysis)
* Khởi đầu, tất cả các đối tượng tạo thành một cụm duy nhất.
* Một cụm được phân tách theo một tiêu chí nào đó đến khi mỗi cụm chỉ có một đối tượng.
* Khoảng cách lớn nhất giữa các đối tượng cận nhau nhất.
* 5.3. Gom cụm dữ liệu bằng phân cấp
* 5.3. Gom cụm dữ liệu bằng phân cấp
* Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).
* 5.3. Gom cụm dữ liệu bằng phân cấp
* Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).
* 5.3. Gom cụm dữ liệu bằng phân cấp
* Các độ đo dùng đo khoảng cách giữa các cụm Ci và Cj
* 5.3. Gom cụm dữ liệu bằng phân cấp
* Một số giải thuật gom cụm dữ liệu bằng phân cấp
* BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): phân hoạch các đối tượng dùng cấu trúc cây theo độ co giãn của phân giải (scale of resolution)
* ROCK (Robust Clustering using linKs): gom cụm dành cho các thuộc tính rời rạc (categorical/discrete attributes), trộn các cụm dựa vào sự kết nối lẫn nhau giữa các cụm
* Chameleon: mô hình động để xác định sự tương tự giữa các cặp cụm
* 5.3. Gom cụm dữ liệu bằng phân cấp
* Một số vấn đề với gom cụm dữ liệu bằng phân cấp
* Chọn điểm trộn/phân tách phù hợp
* Khả năng co giãn (scalability)
* Mỗi quyết định trộn/phân tách yêu cầu kiểm tra/đánh giá nhiều đối tượng/cụm.

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

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

* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Gom cụm dữ liệu dựa trên mật độ
* Mỗi cụm là một vùng dày đặc (dense region) gồm các đối tượng.
* Các đối tượng trong vùng thưa hơn được xem là nhiễu.
* Mỗi cụm có dạng tùy ý.
* Giải thuật
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
* OPTICS (Ordering Points To Identify the Clustering Structure)
* DENCLUE (DENsity-based CLUstEring)
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
* Phân tích các điểm kết nối nhau dựa vào mật độ
* OPTICS (Ordering Points To Identify the Clustering Structure)
* Tạo ra thứ tự các điểm dữ liệu tùy vào cấu trúc gom cụm dựa vào mật độ của tập dữ liệu
* DENCLUE (DENsity-based CLUstEring)
* Gom cụm dựa vào các hàm phân bố mật độ
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* ***ε***: bán kính của vùng láng giềng của một đối tượng, gọi là ***ε***-neighborhood.
* ***MinPts***: số lượng đối tượng ít nhất được yêu cầu trong ***ε***-neighborhood của một đối tượng.
* Nếu đối tượng có ***ε***-neighborhood với ***MinPts*** thì đối tượng này được gọi là đối tượng lõi (core object).
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* ***Directly density-reachable*** (khả năng đạt được trực tiếp): *q* có thể đạt được trực tiếp từ *p* nếu *q* trong vùng láng giềng ***ε***-neighborhood của *p* và *p* phải là core object.
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* ***Density-reachable*** (khả năng đạt được):
* Cho trước tập đối tượng ***D***, ***ε*** và ***MinPts***
* *q* ***density-reachable*** từ *p* nếu ∃ chuỗi các đối tượng *p1*, ..., *pn*  ***D*** với *p1* = *p* và *pn* = *q* sao cho *pi+1* ***directly density-reachable*** từ *pi* theo các thông số ***ε*** và ***MinPts***, 1 ≤ *i* ≤ *n*.
* Bao đóng truyền (transitive closure) của directly density-reachable
* Quan hệ bất đối xứng (asymmetric relation)
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* ***Density-connected*** (nối kết dựa trên mật độ):
* Cho trước tập các đối tượng ***D***, ***ε*** và ***MinPts***
* *p*, *q* ∈ ***D***
* *q* ***density-connected*** với *p* nếu  *o*  ***D*** sao cho cả *q* và *p* đều ***density-reachable*** từ *o* theo các thông số ***ε*** và ***MinPts***.
* Quan hệ đối xứng
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
* Cụm dựa trên mật độ (density based cluster): tập tất cả các đối tượng được nối kết với nhau dựa trên mật độ.
* Đối tượng thuộc về cụm có thể là core object.
* Nếu đối tượng đó không là core object thì gọi là đối tượng ranh giới (border object).
* Đối tượng không thuộc về cụm nào được xem là nhiễu (noise/outlier).
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
* Input: tập đối tượng ***D***, ***ε***, ***MinPts***
* Output: density-based clusters (và noise/outliers)
* Giải thuật
* 1. Xác định ***ε***–neighborhood của mỗi đối tượng *p* ∈ ***D***.
* 2. If *p* là core object, tạo được một cluster.
* 3. Từ bất kì core object *p*, tìm tất cả các đối tượng ***density-reachable*** và đưa các đối tượng này (hoặc các cluster) vào cùng cluster ứng với *p*.
* 3.1. Các cluster đạt được (density-reachable cluster) có thể được trộn lại với nhau.
* 3.2. Dừng khi không có đối tượng mới nào được thêm vào.
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* 5.4. Gom cụm dữ liệu dựa trên mật độ
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
* Đặc điểm ???
* Các cụm có dạng và kích thước khác nhau.
* Không có giả định về phân bố của các đối tượng dữ liệu
* Không yêu cầu về số cụm
* Không phụ thuộc vào cách khởi động (initialization)
* Xử lý nhiễu (noise) và các phần tử biên (outliers)
* Yêu cầu trị cho thông số nhập
* Yêu cầu định nghĩa của mật độ (density)
* ***ε*** và ***MinPts***
* Độ phức tạp
* O(nlogn) 🡪 O(n2)
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* Tối ưu hóa sự phù hợp giữa dữ liệu và mô hình toán nào đó
* Giả định về quá trình tạo dữ liệu
* Dữ liệu được tạo ra với nhiều sự phân bố xác suất khác nhau.
* Các phương pháp
* Tiếp cận thống kê
* Mở rộng của giải thuật gom cụm dựa trên phân hoạch k-means: Expectation-Maximization (EM)
* Tiếp cận học máy: gom cụm ý niệm (conceptual clustering)
* Tiếp cận mạng neural: Self-Organizing Feature Map (SOM)
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* Gom cụm Expectation-Maximization (EM)
* Giải thuật tinh chỉnh lặp để gán các đối tượng vào các cụm (bước kỳ vọng) và ước lượng trị thông số (bước cực đại hoá).
* Gom cụm ý niệm (conceptual clustering)
* Tạo ra cách phân lớp các đối tượng chưa được gán nhãn dựa vào các mô tả đặc trưng cho mỗi nhóm đối tượng ứng với mỗi khái niệm (concept).
* Gom cụm với mạng neural
* Biểu diễn mỗi cụm là một ví dụ tiêu biểu (exemplar).
* Exemplar đóng vai trò của một prototype của cụm.
* Các đối tượng mới được phân bố vào một cụm nếu tương tự với exemplar của cụm đó nhất dựa trên độ đo khoảng cách.
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* Giải thuật Expectation-Maximization (EM)
* Gán một đối tượng vào một cụm nếu tương tự trung tâm (mean) của cụm đó nhất
* Dựa vào trọng số (weight) của đối tượng đối với mỗi cụm
* Xác suất thành viên (probability of membership)
* Không có ranh giới giữa các cụm
* Trung tâm của mỗi cụm được tính dựa vào các độ đo có trọng số (weighted measures).
* Hội tụ nhanh nhưng có thể tối ưu cục bộ
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* Giải thuật Expectation-Maximization (EM)
* Input: tập n đối tượng, **K** (số cụm)
* Output: trị tối ưu cho các thông số của mô hình
* Giải thuật:
* 1. Khởi trị
* 1.1. Chọn ngẫu nhiên **K** đối tượng làm trung tâm của **K** cụm
* 1.2. Ước lượng trị ban đầu cho các thông số (nếu cần)
* 2. Lặp tinh chỉnh các thông số (cụm):
* 2.1. Bước kỳ vọng (expectation step): gán mỗi đối tượng xi đến cụm Ck với xác suất P(xi ∈ Ck) với k=1..**K**
* 2.2. Bước cực đại hóa (maximization step): ước lượng trị các thông số
* 2.3. Dừng khi thỏa điều kiện định trước.
* 5.5. Gom cụm dữ liệu dựa trên mô hình
* Giải thuật Expectation-Maximization (EM)
* Giải thuật:
* 1. Khởi trị
* 2. Lặp tinh chỉnh các thông số (cụm):
* 2.1. Bước kỳ vọng (expectation step): gán mỗi đối tượng xi đến cụm Ck với xác suất P(xi ∈ Ck)
* 2.2. Bước cực đại hóa (maximization step): ước lượng trị các thông số
* 5.6. Các phương pháp gom cụm dữ liệu khác
* Gom cụm cứng (hard clustering)
* Mỗi đối tượng chỉ thuộc về một cụm.
* Mức thành viên (degree of membership) của mỗi đối tượng với một cụm hoặc là 0 hoặc là 1.
* Ranh giới (boundary) giữa các cụm rõ ràng.
* Gom cụm mờ (fuzzy clustering)
* Mỗi đối tượng thuộc về nhiều hơn một cụm với mức thành viên nào đó từ 0 đến 1.
* Ranh giới giữa các cụm không rõ ràng (mờ - vague/fuzzy).
* 5.7. Tóm tắt
* Gom cụm nhóm các đối tượng vào các cụm dựa trên sự tương tự giữa các đối tượng.
* Độ đo đo sự tương tự tùy thuộc vào kiểu dữ liệu/đối tượng cụ thể.
* Các giải thuật gom cụm được phân loại thành: nhóm phân hoạch, nhóm phân cấp, nhóm dựa trên mật độ, nhóm dựa trên lưới, nhóm dựa trên mô hình, …
* 5.7. Tóm tắt
* Hỏi & Đáp …
* Giới thiệu môn học