**Bài 01 Buổi 04 Bài tập thêm**

<https://drive.google.com/file/d/1knowt6qNhawLjPSr1Af1APPRfDdCEojE/view?usp=sharing>

**Bài 02 Buổi 04 Bài tập thêm**

Có nhiều thuật toán gom cụm tự động xác định được số lượng cụm mà không cần

người dùng phải cho trước như k-Means. Nhiều người cho rằng đây là điểm vượt

trội của chúng so với k-Means. Em hãy liệt kê ít nhất 2 trường hợp để phản biện lại

quan điểm này.

* Chúng ta không thể quyết định số cụm mà chúng ta mong muốn, số lượng cụm mà chúng ta muốn phân loại cũng là một thông số quan trọng trong tập dataset. Việc không thể quyết định cụm cũng làm cho chúng ta khó có thể tùy biến cho phù hợp với từng tình huống.
* Một số phương pháp gom cụm không cần xác định số lượng cụm như là: “hierarchical clustering”. Điểm mạnh là nó không cần phải xác định số cụm nhưng nó có điểm yếu là hiếm khi đưa ra được kết quả tốt nhất, không chạy được với dữ liệu thiếu hoặc dữ liệu hỗn hợp. Khi tập dữ liệu lớn thì tốc độ chạy không được tốt.

***Bài 03 Buổi 04 Bài tập thêm***

Bạn được cho một tập dữ liệu gồm 100 dòng và yêu cầu gom cụm chúng. Bạn sử

dụng thuật toán k-Means để giải quyết bài toán, tuy nhiên với tất cả các giá trị k, 1 ≤

k ≤ 100, thuật toán k-Means đều cho ra kết quả là một cụm không rỗng duy nhất.

Bạn lại tìm cách áp dụng tất cả các thuật toán cải tiến của k-Means và đều nhận được

kết quả tương tự. Bạn hãy giải thích vì sao?

***Bài 04 Buổi 04 Bài tập thêm***

Tìm hiểu thêm các thuật toán gom cụm phổ biến khác:

a) k-Modes, CLARANS

b) Chameleon, BIRCH

c) DBSCAN, OPTICS

d) STING, CLIQUE

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ý nghĩa | Ưu điểm | Nhược điểm |
| K-modes | Khắc phục khuyết điểm của K-means là chỉ chạy với dữ liệu định lượng, K-modes dùng để chạy với tập dữ liệu định tính. | Tính được dữ liệu định tính | Không phù hợp với dữ liệu lớn  Khó chọn số lượng nhóm(k) |
| CLARANS | Khắc phục điểm yếu thuật toán k-medoid là tập dữ liệu lớn. | Cải tiến hơn so với PAM và CLARA  Có thể xử lí ngoại biên | Khó cài đặt |
| Chameleon | Thuật toán sử dụng phương pháp tiếp cận là từ dưới lên. Mỗi điểm là một nhóm riêng và khi càng đi lên thì các điểm giống nhau được gộp lại. | Không phải nhập tham số là số lượng cụm như K-means  Có thể áp dựng với tập dữ liệu lớn | Khó áp dụng với tập dữ liệu nhiều chiều |
| BIRCH | Thuật toán này được xây dựng dựa trên cây CF (clustering features). | Chạy nhanh với tập dữ liệu lớn vì chỉ cần 1 lần duyệt dataset. | Không xử lí được dữ liệu dạng định tính. |
| DBSCAN | Thuật toán được xây dựng dựa trên Phương pháp phân cụm dựa trên mật độ (Density-Based Clustering). | Khó bị ảnh hưởng bởi phần tử ngoại biên nên thường được dùng để phát hiện các phần tử ngoại biên  Có thể hoạt động hiệu quả với mọi kích thước dữ liệu | Khó có thể áp dụng với dữ liệu có nhiều chiều.  Chỉ hoạt động với dữ liệu định lượng. |
| OPTICS | Thuật toán được cải tiến từ DBSCAN, ít bị ảnh hưởng bởi các tham số cài đặt hơn DBSCAN. (eps và minPts) | Có thể tìm được các cụm có mật độ khác nhau.  Giảm thời gian chạy hơn so với DBSCAN vì tham số eps không bắt buộc. | Vẫn còn bị ảnh hưởng bởi tham số minPts |
| STING | Thuật toán áp dụng phương pháp tiếp cận dạng lưới (Grid). | Tốc độ nhanh vì chỉ bị ảnh hưởng bới số lượng ô cần chia, không phải số lượng dữ liệu. Thời gian chạy là O(K) với k là số ô cần chia ở mức thấp nhất.  Không cần tham số đầu vào nhiều.  Có thể xử lí song song (parallel processing) | Thường thì biên sẽ theo chiều dọc hoặc ngang, khó có thể thấy phát hiện biên theo đường chéo. |
| CLIQUE | Thuật toán cũng áp dụng phương pháp tiếp cận dạng lưới (Grid). | Không bị ảnh hưởng bởi thứ tự của đầu vào (order of the records in input)  Dễ dàng mở rộng cũng như tăng tuyến tính (linear) với kích cỡ của tập dữ liệu. | Nếu kích thước của ô (cell) không phù hợp với dữ liệu thì khó có thể tìm được cụm chính xác. |