2025-05-08 13:40

Tags: [[data mining]],

# So sánh sự khác nhau giữa Kmeans, K\_medoids, Kernel K-Means

### **Giới thiệu**

Phân cụm (clustering) là một kỹ thuật quan trọng trong học máy không giám sát (unsupervised learning), với mục tiêu phân chia tập dữ liệu thành các nhóm sao cho các điểm trong cùng một nhóm có tính chất tương đồng. Ba thuật toán phân cụm phổ biến được sử dụng rộng rãi là **K-Means**, **K-Medoids** và **Kernel K-Means**. Mỗi thuật toán có những ưu điểm, hạn chế và phù hợp với các loại dữ liệu khác nhau.

### **1. Điểm giống nhau**

Cả ba thuật toán đều chia sẻ những đặc điểm chung cơ bản như:

* **Là thuật toán phân cụm không giám sát**: Không yêu cầu nhãn (label) của dữ liệu.
* **Yêu cầu xác định trước số lượng cụm K**: Người dùng cần cung cấp giá trị K — số cụm mong muốn.
* **Mục tiêu tối ưu tương tự nhau**: Cả ba đều cố gắng giảm thiểu sự khác biệt trong cùng một cụm, tức là làm cho các điểm dữ liệu trong mỗi cụm gần nhau nhất có thể (theo định nghĩa riêng của từng thuật toán).
* **Quy trình lặp**: Sử dụng quá trình lặp gồm bước gán cụm và cập nhật tâm cụm cho đến khi hội tụ.

### **2. Điểm khác nhau chi tiết**

| **Tiêu chí** | **K-Means** | **K-Medoids** | **Kernel K-Means** |
| --- | --- | --- | --- |
| **Tâm cụm (centroid)** | Là trung bình (mean) của tất cả điểm trong cụm | Là một điểm thực tế trong dữ liệu | Là trung bình trong không gian ánh xạ (kernel space) |
| **Khoảng cách sử dụng** | Euclidean (mặc định) | Bất kỳ khoảng cách (thường dùng Manhattan) | Khoảng cách tính bằng hàm kernel (phi tuyến) |
| **Độ nhạy với ngoại lệ** | Cao — dễ bị ảnh hưởng bởi điểm nhiễu hoặc giá trị cực trị | Thấp hơn vì tâm là một điểm thực tế trong dữ liệu | Có thể giảm đáng kể nếu chọn kernel phù hợp |
| **Khả năng xử lý dữ liệu phi tuyến** | Kém — chỉ phù hợp dữ liệu tuyến tính | Kém | Tốt — ánh xạ dữ liệu vào không gian đặc trưng phi tuyến |
| **Tính ổn định** | Thấp nếu dữ liệu nhiều nhiễu hoặc phân bố không đều | Ổn định hơn vì tâm là điểm cụ thể | Ổn định nếu kernel và tham số được lựa chọn hợp lý |
| **Độ phức tạp tính toán** | Thấp — O(nkt), nhanh với dữ liệu lớn | Cao hơn — do phải tính tổng khoảng cách tới mọi điểm | Rất cao — phải tính toán ma trận kernel K(n×n) và cập nhật phức tạp |
| **Tính khả diễn giải (interpretability)** | Dễ hiểu và dễ minh họa | Dễ hiểu hơn vì tâm là một điểm cụ thể | Khó diễn giải vì không gian ánh xạ thường không trực quan |
| **Ứng dụng phù hợp** | Phân tích khách hàng, ảnh số, dữ liệu tuyến tính đơn giản | Bài toán yêu cầu tính ổn định cao, chống nhiễu tốt | Dữ liệu phức tạp có cấu trúc phi tuyến, ví dụ: hình ảnh, âm thanh, sinh học |

# References