2025-05-22 00:08

Tags: [[data mining]],

# Data\_Mining\_Lab7\_Report

## Tóm tắt phần Code "Implementation From Scratch"

Phần code "Implementation From Scratch" tự xây dựng thuật toán Phân cụm Phân cấp Tập hợp (Agglomerative Hierarchical Clustering). Các thành phần chính bao gồm:

1. **Hàm tính khoảng cách:**
   * euclidean\_distance(point1, point2): Tính khoảng cách Euclidean giữa hai điểm dữ liệu.
2. **Hàm hợp nhất cụm:**
   * merge\_clusters(cluster1, cluster2): Gộp hai cụm (danh sách các điểm) thành một cụm mới.
3. **Hàm tính khoảng cách liên cụm:**
   * \_calculate\_inter\_cluster\_distance(cluster1, cluster2, linkage\_method): Tính khoảng cách giữa hai cụm dựa trên phương pháp liên kết (linkage\_method) được chọn:
     + 'single': Khoảng cách nhỏ nhất giữa các điểm của hai cụm.
     + 'complete': Khoảng cách lớn nhất giữa các điểm của hai cụm.
     + 'average': Khoảng cách trung bình giữa tất cả các cặp điểm từ hai cụm.
4. **Hàm phân cụm chính:**
   * agglomerative\_clustering(dataset, num\_clusters, linkage\_method='average'): Thực hiện thuật toán chính.
     + **Khởi tạo**: Mỗi điểm dữ liệu ban đầu là một cụm riêng.
     + **Vòng lặp gộp cụm**
       - Lặp cho đến khi số cụm hiện tại bằng num\_clusters.
       - Trong mỗi vòng lặp:
         1. min\_inter\_cluster\_dist = float('inf'): Khởi tạo khoảng cách nhỏ nhất giữa các cụm là vô cực.
         2. closest\_pair\_indices = None: Khởi tạo chỉ số của cặp cụm gần nhất.
         3. **Tìm cặp cụm gần nhất**: Duyệt qua tất cả các cặp cụm hiện có (for i ... for j ...).

Gọi \_calculate\_inter\_cluster\_distance để lấy khoảng cách giữa cặp cụm hiện tại theo linkage\_method.

Nếu khoảng cách này nhỏ hơn min\_inter\_cluster\_dist, cập nhật min\_inter\_cluster\_dist và closest\_pair\_indices.

* + - * 1. **Kiểm tra điều kiện dừng**: Nếu không tìm thấy cặp nào để hợp nhất (closest\_pair\_indices is None), thoát vòng lặp.
        2. **Xác định chỉ số cụm**: Lấy idx1 (nhỏ hơn) và idx2 (lớn hơn) từ closest\_pair\_indices.
        3. **Hợp nhất**: Gọi merge\_clusters để tạo merged\_cluster từ clusters[idx1] và clusters[idx2].
        4. **Cập nhật danh sách cụm**:

clusters.pop(idx2): Xóa cụm có chỉ số lớn hơn trước.

clusters.pop(idx1): Xóa cụm có chỉ số nhỏ hơn.

clusters.append(merged\_cluster): Thêm cụm mới đã hợp nhất.

* + - **Kết quả**: Trả về danh sách các cụm cuối cùng.

## So sánh với Hàm Có Sẵn trong Thư viện (ví dụ: sklearn.cluster.AgglomerativeClustering, scipy.cluster.hierarchy.linkage)

| Tính năng/Khía cạnh | Implementation From Scratch | Hàm Thư Viện (sklearn/scipy) |
| --- | --- | --- |
| **Mục đích chính** | Hiểu rõ cơ chế, mục đích học tập, tùy chỉnh sâu. | Sử dụng thực tế, hiệu suất cao, độ tin cậy, nhiều tính năng. |
| **Phương pháp Liên kết** | Hỗ trợ: single, complete, average (triển khai thủ công). | Hỗ trợ nhiều hơn: single, complete, average, ward, centroid, median, v.v. |
| **Thước đo Khoảng cách** | Chỉ euclidean\_distance. | Nhiều lựa chọn: euclidean, manhattan, cosine, precomputed (ma trận khoảng cách), v.v. |
| **Hiệu suất** | Thấp hơn đáng kể, đặc biệt với dữ liệu lớn (có thể O(N³)). | Tối ưu hóa cao (thường O(N²) hoặc tốt hơn), sử dụng thuật toán hiệu quả. |
| **Đầu ra** | Danh sách các cụm (danh sách các điểm trong mỗi cụm). | sklearn: Nhãn cụm (labels\_), cây phân cấp (children\_).scipy: Ma trận liên kết (Z) để vẽ dendrogram. |
| **Vẽ Dendrogram** | Không trực tiếp. Cần xử lý thêm để tạo ma trận liên kết. | scipy.cluster.hierarchy.dendrogram dùng trực tiếp ma trận Z. sklearn ít trực tiếp hơn. |
| **Tính linh hoạt** | Cao nếu muốn sửa đổi logic cốt lõi. | Cao về tham số, khó can thiệp sâu vào thuật toán cơ bản. |
| **Xử lý lỗi & Độ mạnh mẽ** | Cơ bản, phụ thuộc người triển khai. | Rất tốt, được kiểm thử rộng rãi. |
| **Dễ sử dụng** | Cần hiểu rõ các bước. | API đơn giản, dễ sử dụng. |

# References