Báo cáo Thực hành Phân cụm Dữ liệu

1. Giới thiệu

Phân cụm là một phương pháp quan trọng trong khai phá dữ liệu và học máy không giám sát, cho phép khám phá cấu trúc tiềm ẩn trong tập dữ liệu mà không cần nhãn. Trong bài thực hành này, tôi tiến hành phân tích và so sánh hai phương pháp phân cụm phổ biến: K-means và Phân cụm phân cấp (Agglomerative Hierarchical Clustering - AHC). Mục tiêu của thực hành là tìm hiểu tính hiệu quả của mỗi phương pháp trong việc phát hiện cấu trúc tự nhiên của dữ liệu và đánh giá hiệu suất của các thuật toán trong điều kiện cụ thể của dữ liệu.

2. Dữ liệu và Phương pháp

Dữ liệu sử dụng cho bài thực hành bao gồm các vector đặc trưng embedding của các đối tượng thuộc bốn nhóm: động vật (animals), quốc gia (countries), trái cây (fruits), và rau củ (veggies). Các đặc trưng này được biểu diễn dưới dạng vector số, nhằm phục vụ quá trình tính toán khoảng cách giữa các đối tượng.

tôi thực hiện phân cụm dữ liệu sử dụng hai phương pháp:

- **K-means**: Đây là thuật toán phân cụm dựa trên centroid, hoạt động dựa trên giả định rằng các cụm có hình cầu và đồng đều về kích thước.
- **AHC**: Phương pháp phân cụm phân cấp với hai chiến lược liên kết là GAAC (Group Average Agglomerative Clustering) và liên kết hoàn toàn (Complete Linkage), cho phép khám phá cấu trúc phân cấp của dữ liệu.

3. Thực hiện và Kết quả

3.1. Kết quả Phân cụm K-means

K-means được thực hiện với các giá trị k từ 1 đến 10. Các chỉ số Precision, Recall và F1 Score được sử dụng để đánh giá hiệu quả phân cụm. Biểu đồ kết quả cho thấy các chỉ số đều có giá trị rất thấp và không thay đổi nhiều khi k tăng, ngoại trừ giá trị Recall đạt đỉnh tại k=7. Thời gian chạy của K-means dao động từ 0.0279 giây đến 0.3591 giây, tăng dần khi số cụm k tăng.

Kết quả này cho thấy K-means không đạt hiệu quả tốt trên dữ liệu, có thể do cấu trúc phức tạp của dữ liệu không phù hợp với giả định cụm hình cầu của K-means. Kết quả phân cụm của K-means không ổn định và không thể hiện rõ ràng các nhóm tự nhiên trong dữ liệu.

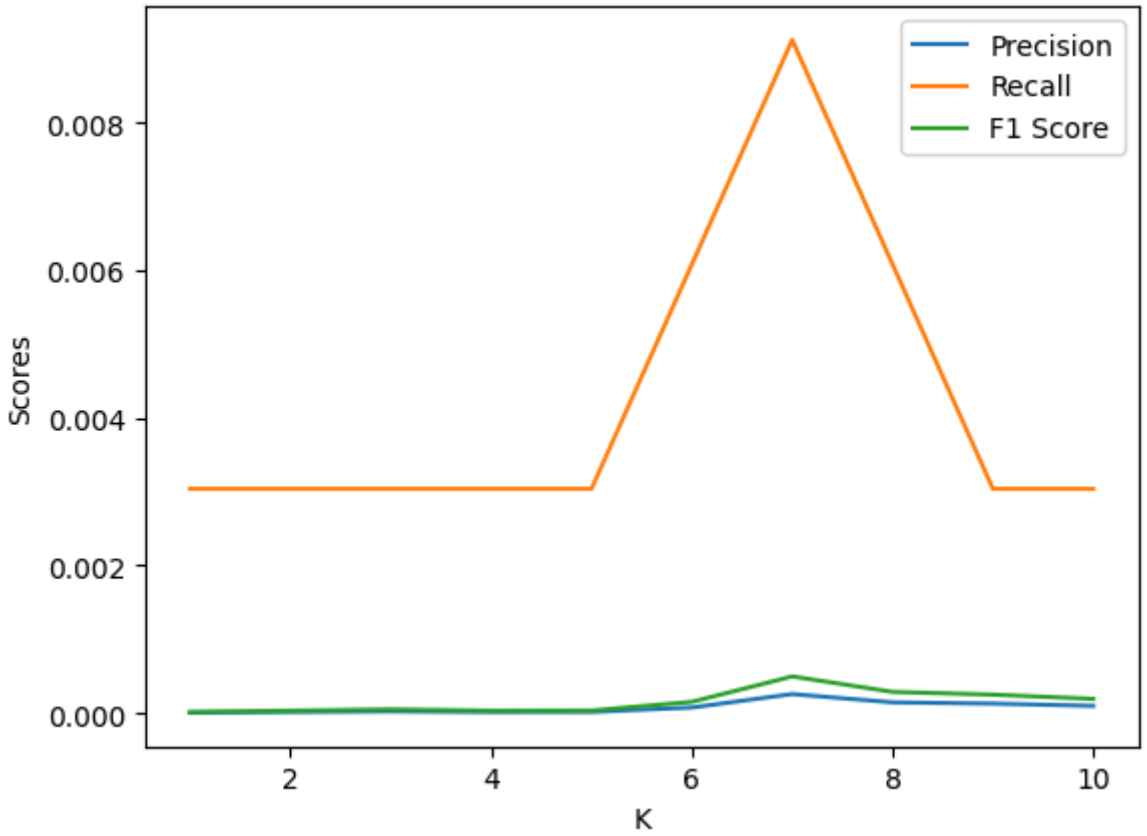

Figure 1. K-mean result

3.2. Kết quả Phân cụm Phân cấp (AHC)

Đối với AHC, tôi thực hiện phân cụm với hai chiến lược liên kết GAAC và Complete Linkage, sau đó hiển thị kết quả dưới dạng biểu đồ cây phân cấp (dendrogram). Cả hai phương pháp đều hiển thị cấu trúc phân cấp phức tạp, với nhiều nhánh nhỏ thể hiện các cụm con trong dữ liệu. Thời gian chạy của GAAC và Complete Linkage là 0.0130 giây và 0.0123 giây, rất nhanh do dữ liệu không quá lớn.

Dendrogram từ AHC cho phép nhìn thấy rõ ràng hơn cấu trúc phân cấp trong dữ liệu. Người dùng có thể chọn mức cắt khác nhau để tạo các cụm với số lượng tùy ý, phù hợp cho các bài toán yêu cầu phân tích đa cấp độ.

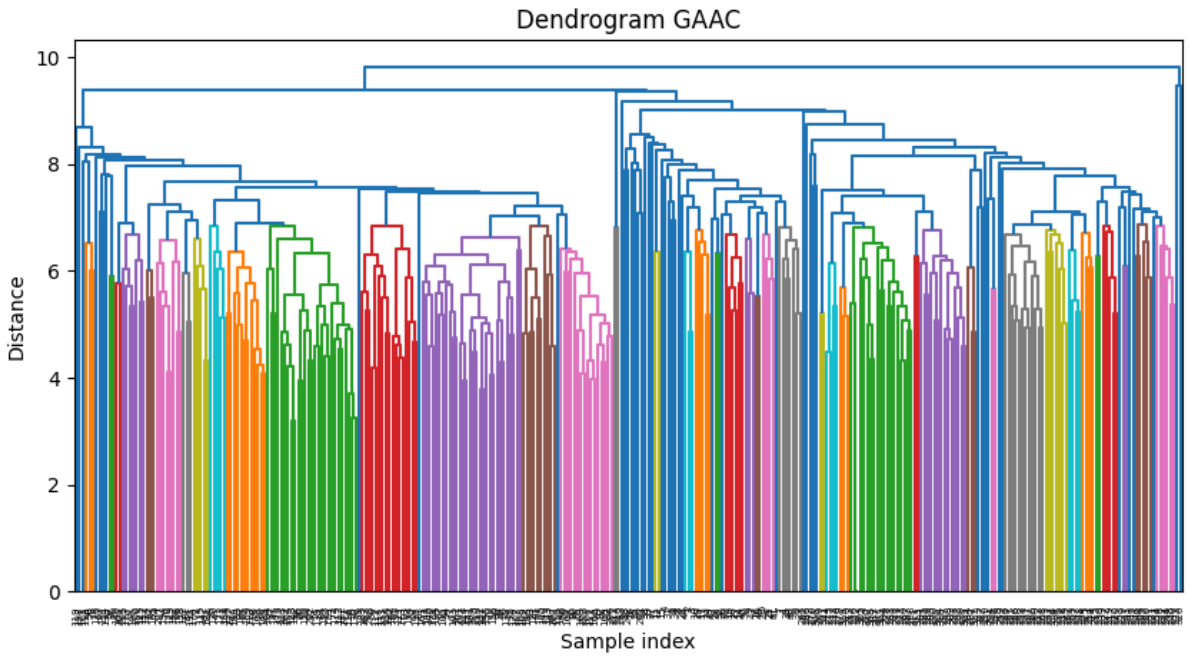

Figure 2. AHC results - Dendrogram GAAC

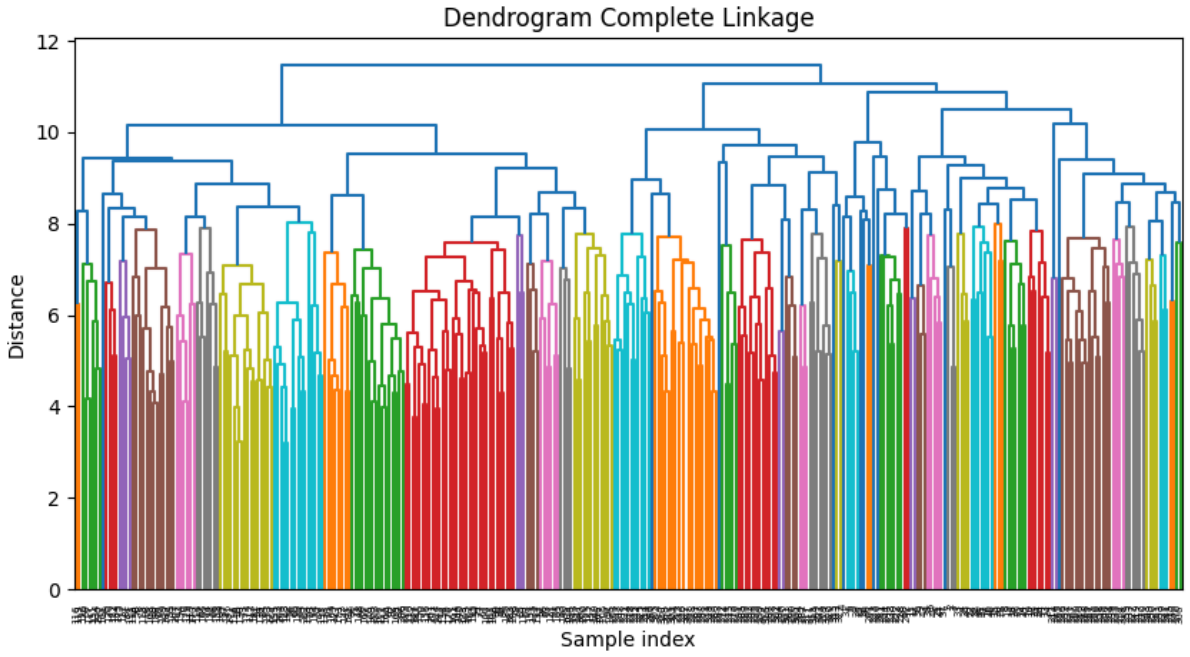

Figure 3. AHC results - Dendrogram Complete Linkage

4. So sánh K-means và AHC

Kết quả phân cụm cho thấy sự khác biệt rõ rệt giữa hai phương pháp:

- **K-means** có thời gian chạy nhanh và ổn định, tuy nhiên không phù hợp với dữ liệu có cấu trúc phức tạp hoặc không đồng đều. Các chỉ số đánh giá rất thấp, cho thấy K-means khó phân tách chính xác các nhóm trong dữ liệu này.

- **AHC** cung cấp một cái nhìn chi tiết hơn về cấu trúc phân cấp của dữ liệu, đặc biệt là khả năng hiển thị các cụm con qua dendrogram. Thời gian chạy của AHC tương đối ngắn, và phương pháp này có khả năng thích nghi tốt hơn với dữ liệu có nhiều lớp phân cụm.

Với tập dữ liệu này, AHC cho thấy hiệu quả vượt trội hơn so với K-means do khả năng mô tả chi tiết mối quan hệ phân cấp và tính linh hoạt trong việc lựa chọn số lượng cụm.

5. Kết luận

Qua bài thực hành, tôi rút ra kết luận rằng AHC là phương pháp phù hợp hơn so với K-means cho bài toán phân cụm dữ liệu có cấu trúc phức tạp và không đồng nhất. K-means, mặc dù có ưu điểm về tốc độ, nhưng không đạt hiệu quả phân cụm cao khi dữ liệu không phù hợp với giả định cụm hình cầu. Ngược lại, AHC cho phép phân tích đa cấp và dễ dàng phát hiện các cụm con, phù hợp cho các bài toán yêu cầu sự phân cấp trong dữ liệu.

Để cải thiện chất lượng phân cụm trong tương lai, tôi đề xuất thử nghiệm các phương pháp phân cụm khác như DBSCAN hoặc Gaussian Mixture Model (GMM), đặc biệt khi dữ liệu có thể có cấu trúc không gian phức tạp hoặc không đồng đều về kích thước cụm.