TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KÌ

ĐỀ TÀI: PHÂN CỤM DỮ LIỆU BIỂU HIỆN GEN VỚI ENSEMBLE LEARNING

Môn học: Học máy

GVHD: TS. Phan Thị Huyền Trang

Mã môn học: MALE431984_09

Nhóm: 1

Sinh viên thực hiện:

Đỗ Kiến Hưng	23133030
Huỳnh Ngọc Thạch	23133072
Huỳnh Hữu Huy	23133027

TP Hồ Chí Minh, ngày 06 tháng 01 năm 2026

ĐỀ TÀI: PHÂN CỤM DỮ LIỆU BIỂU HIỆN GEN VỚI ENSEMBLE LEARNING

DANH SÁCH THÀNH VIÊN THỰC HIỆN

Họ và tên	**MSSV**
Đỗ Kiến Hưng	23133030
Huỳnh Ngọc Thạch	23133072
Huỳnh Hữu Huy	23133027

Nhận xét của giáo viên:

Ngày 06 tháng 01 năm 2026

Giảng viên chấm điểm

TP.HCM, ngày 06 tháng 01 năm 2026

MỤC LỤC

BẢNG PHÂN CÔNG NHIỆM VỤ

Nhóm: 1

Đề tài: PHÂN CỤM DỮ LIỆU BIỂU HIỆN GEN VỚI ENSEMBLE LEARNING

STT	MSSV	Họ và tên	Nội dung thực hiện	Mức độ hoàn thành
1	23133030	Đỗ Kiến Hưng	- Xây dựng model K-Means++ - Xây dựng hàm Ensemble Weighted CSPA và SCENA - Xây dựng app demo sử dụng Streamlit	100%
2	23133072	Huỳnh Ngọc Thạch	- Xây dựng model DBSCAN - Tiền xử lý dữ liệu và giảm chiều dữ liệu dùng PCA - Trực quan hóa dữ liệu và kết quả đạt được - Hỗ trợ phát triển SCENA trong Ensemble	100%
3	23133027	Huỳnh Hữu Huy	- Xây dựng model Hierarchical - Chạy thực nghiệm và tinh chỉnh tham số và đánh giá kết quả đạt được. - Hỗ trợ phát triển Weighted CSPA trong Ensemble	100%

Ghi chú: % mức độ hoàn thành công việc

DANH MỤC HÌNH ẢNH

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt / Diễn giải
ARI	Adjusted Rand Index	Chỉ số Rand điều chỉnh (Đo độ tương đồng phân cụm)
CSPA	Cluster-based Similarity Partitioning Algorithm	Thuật toán phân hoạch dựa trên độ tương đồng cụm
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Thuật toán phân cụm dựa trên mật độ có nhiễu
KNN	K-Nearest Neighbors	K-Láng giềng gần nhất
ML	Machine Learning	Học máy
NMI	Normalized Mutual Information	Thông tin tương hỗ chuẩn hóa
PCA	Principal Component Analysis	Phân tích thành phần chính (Giảm chiều dữ liệu)
RNA-Seq	RNA Sequencing	Kỹ thuật giải trình tự RNA
SCENA	Single-cell Clustering by Enhancing Network Affinity	Phân cụm đơn bào bằng tăng cường ái lực mạng lưới
TCGA	The Cancer Genome Atlas	Bộ dữ liệu bản đồ gen ung thư

PHỤ LỤC TÀI LIỆU

Tài liệu	Liên kết
Dataset (UCI)	Gene Expression Cancer RNA-Seq
Demo App	geneExEnCluG1.streamlit.app
GitHub Repo	github.com/darktheDE/gene-expression-ensemble-clustering
Google Drive	Process, Colab, Báo cáo
Slide trình bày	Canva Presentation

TÓM TẮT

Việc phân loại các phân nhóm ung thư dựa trên dữ liệu biểu hiện gen đóng vai trò quan trọng trong y học chính xác nhưng thường gặp khó khăn do đặc thù dữ liệu có số chiều lớn và sự tồn tại của nhiễu sinh học. Các thuật toán phân cụm đơn lẻ thông thường, thường có những nhược điểm về tính kém ổn định với các dữ liệu ngoại lai. Để giải quyết vấn đề nhóm em đề xuất mô hình lai Weighted SCENA-based Ensemble nhằm nâng cao chất lượng phân cụm thông qua kỹ thuật học máy tổ hợp. Phương pháp này kết hợp thuật toán phân hoạch dựa trên sự tương đồng CSPA có gán trọng số với mạng lưới KNN để tận dụng tối đa cấu trúc dữ liệu cục bộ. Kết quả thực nghiệm trên bộ dữ liệu Gen Expression Cancer RNA-Seq cho thấy mô hình đề xuất đã cải thiện đáng kể chỉ số Silhouette và Adjusted Rand Index (ARI) so với các mô hình cơ sở, nên có thể xác định các nhóm bệnh tiềm ẩn.

1. GIỚI THIỆU

1.1. Vấn đề cần giải quyết

Y học bây giờ, ung thư không được xem là một bệnh duy nhất mà gồm nhiều dạng khác nhau. Các khối u xuất phát từ cùng một cơ quan có thể mang đặc điểm phân tử riêng, dẫn tới sự khác biệt về tiến triển bệnh và phản ứng điều trị. Vì lý do đó, việc phân nhóm ung thư dựa trên dữ liệu biểu hiện gen được quan tâm, vì nó hỗ trợ lựa chọn phác đồ phù hợp cho từng bệnh nhân.

Khi áp dụng học máy cho dữ liệu biểu hiện gen, nhiều khó khăn xuất hiện. Đặc điểm nổi bật là số lượng gen rất lớn, trong khi số mẫu quan sát lại ít. Tình huống này khiến khoảng cách giữa các điểm dữ liệu khó diễn giải và làm giảm hiệu quả của các phương pháp phân cụm dựa trên khoảng cách.

Ngoài ra, dữ liệu sinh học thường có nhiễu do kỹ thuật đo đạc và sự biến thiên tự nhiên. Các nhóm dữ liệu không tách biệt rõ, mật độ không đồng đều và có thể chồng lấn. Các phương pháp phân cụm đơn lẻ như K-Means++ hay Hierarchical Clustering dễ bị ảnh hưởng bởi nhiễu và kết quả không ổn định.

1.2. Các phương pháp Machine Learning hiện có

K-Means++ hoạt động dựa trên nguyên lý phân hoạch dữ liệu thành các nhóm rời rạc bằng cách tối thiểu hóa tổng bình phương khoảng cách từ các điểm dữ liệu đến tâm cụm. Hierarchical Clustering xây dựng cấu trúc phân cấp dưới dạng biểu đồ cây, cho phép quan sát mối quan hệ lồng ghép giữa các nhóm dữ liệu mà không cần xác định trước số lượng cụm cụ thể. DBSCAN tiếp cận vấn đề dựa trên mật độ, tập trung vào việc gom nhóm các điểm dữ liệu nằm sát nhau và tách biệt các điểm nhiễu nằm ở vùng thưa thớt, giúp nhận diện tốt các cụm có hình dạng bất kỳ.

Các thuật toán này có ưu điểm về tốc độ tính toán và khả năng diễn giải, nhưng chúng vẫn tồn tại nhiều hạn chế khi áp dụng đơn lẻ trên dữ liệu sinh học đa chiều. K-Means++ gặp khó khăn với việc khởi tạo tâm cụm ngẫu nhiên và giả định các cụm hình cầu, dẫn đến kết quả thiếu ổn định. DBSCAN tuy xử lý nhiễu tốt nhưng lại rất nhạy cảm với việc lựa chọn tham số bán kính trong không gian dữ liệu gen thưa thớt và nhiều chiều. Nhìn chung, các phương pháp đơn lẻ này dễ bị ảnh hưởng bởi nhiễu và có nguy

cơ bị quá khớp khi số lượng đặc trưng Gen quá lớn so với số lượng mẫu bệnh phẩm hạn chế.

1.3. Lý do đề xuất giải pháp Ensemble

Việc kết hợp ba thuật toán K-Means++, Hierarchical Clustering và DBSCAN là sự đa dạng về cơ chế học của từng phương pháp. K-Means++ hoạt động hiệu quả với các cụm hình cầu lồi, Hierarchical Clustering mô tả tốt cấu trúc phân cấp sinh học, thì DBSCAN lại có ưu thế vượt trội trong việc phát hiện các cụm dựa trên mật độ và xử lý nhiễu.

Nhóm đề xuất áp dụng kỹ thuật Weighted CSPA và SCENA nhằm giải quyết vấn đề chất lượng không đồng đều giữa các mô hình thành phần. Cơ chế gán trọng số dựa trên chỉ số Silhouette cho phép giảm thiểu tác động của các mô hình có kết quả phân cụm kém, giúp hệ thống hoạt động ổn định hơn. Chỉ số Silhouette (Rousseeuw, 1987) được sử dụng để đánh giá độ tách biệt giữa các cụm mà không cần nhãn thực tế [9]. Kỹ thuật SCENA còn được tích hợp để khai thác lại thông tin láng giềng từ dữ liệu gốc thông qua mạng lưới KNN.

1.4. Giới thiệu sơ lược Workflow đề xuất

Quy trình xử lý của mô hình đề xuất được thiết kế thành một luồng khép kín, bắt đầu từ việc tiếp nhận dữ liệu biểu hiện gen thô. Dữ liệu này trước hết trải qua bước tiền xử lý chuẩn hóa và giảm chiều không gian bằng kỹ thuật PCA nhằm trích xuất các thành phần đặc trưng quan trọng nhất. Kết quả sau khi giảm chiều được sử dụng làm đầu vào song song cho ba thuật toán phân cụm cơ sở gồm K-Means++, Hierarchical Clustering và DBSCAN để tạo ra các phương án phân chia đa dạng. Các kết quả này sau đó được hợp nhất thông qua cơ chế tính trọng số dựa trên chất lượng phân cụm, hình thành nên ma trận đồng thuận. Để tăng cường độ chính xác, ma trận đồng thuận tiếp tục được tinh chỉnh bằng cách kết hợp với thông tin láng giềng gần nhất (KNN) trích xuất từ dữ liệu gốc. Cuối cùng, kết quả phân cụm tối ưu được xác định dựa trên ma trận tương đồng đã qua xử lý nâng cao.

2. CÔNG VIỆC LIÊN QUAN

2.1. Các mô hình Machine Learning cơ sở

2.1.1 Mô hình K-Means++

K-Means++ tập trung giải quyết vấn đề khởi tạo tâm cụm ngẫu nhiên. Trong đề tài, nhóm đã cài đặt thuật toán từ đầu để kiểm soát chi tiết quá trình tính toán, với quy trình thực hiện cụ thể qua 5 bước sau:

Bước 1: Khởi tạo tâm cụm (Initialization Strategy):

1. Chọn ngẫu nhiên một điểm dữ liệu đầu tiên làm tâm c_1
2. Với mỗi điểm dữ liệu x còn lại, tính khoảng cách D(x) đến tâm cụm gần nhất đã được chọn.
3. Chọn tâm cụm tiếp theo ci+ từ các điểm dữ liệu với xác suất tỷ lệ thuận với bình phương khoảng cách $D(x)^2$ quy trình này đảm bảo các tâm cụm khởi tạo được phân bố rải rác trong không gian dữ liệu, giúp thuật toán hội tụ nhanh hơn và tránh rơi vào cực trị địa phương.

Bước 2: Vòng lặp gán nhãn và cập nhật (Assign-Update Loop)

- Gán nhãn: Tính khoảng cách Euclidean từ mỗi điểm dữ liệu đến tất cả các tâm cụm. Mỗi điểm sẽ được gán vào cụm có tâm gần nhất. Quá trình này được thực hiện bằng phép tính vector hóa trên toàn bộ ma trận dữ liệu để tối ưu tốc độ.
- Cập nhật: Tính lại vị trí tâm cụm mới bằng cách lấy trung bình cộng tọa độ của tất cả các điểm thuộc về cụm đó

Bước 3: Kiểm tra điều kiện hội tụ

Thuật toán dừng lại khi sự thay đổi vị trí của các tâm cụm giữa hai vòng lặp liên tiếp nhỏ hơn một ngưỡng dung sai $\epsilon = 1e{-}4$ hoặc khi đạt số vòng lặp tối đa.

Bước 4: Xác định số cụm tối ưu:

- Phương pháp Elbow: Theo dõi sự thay đổi của giá trị Inertia. Đồ thị Inertia cho thấy tốc độ giảm bắt đầu chững lại tại k=5.
- Chỉ số Silhouette: Kết quả tính toán cho thấy giá trị Silhouette Score đạt cực đại tại k=5, đồng thuận với kết quả từ phương pháp Elbow và phù hợp với số lượng nhãn thực tế của bộ dữ liệu (5 loại ung thư).

Bước 5: Huấn luyện và Đánh giá mô hình

Mô hình cuối cùng được huấn luyện với k=5 trên toàn bộ tập dữ liệu. Kết quả phân cụm được trực quan hóa trên không gian 2 chiều (PC1, PC2) cùng với vị trí các tâm cụm. Hiệu năng của mô hình được đánh giá định lượng thông qua chỉ số Adjusted Rand Index (ARI) và Normalized Mutual Information (NMI) bằng cách đối chiếu với nhãn phân loại bệnh thực tế.

K-Means là một trong những thuật toán đầu tiên được áp dụng thành công để phân tích dữ liệu biểu hiện gen, điển hình là nghiên cứu của Tavazoie và cộng sự (1999) trong việc xác định mạng lưới di truyền [1]. Tuy nhiên, các nghiên cứu tổng quan sau này chỉ ra rằng K-Means thường gặp hạn chế khi xử lý các dữ liệu có nhiễu cao và kích thước cụm không đồng đều, dẫn đến kết quả thiếu ổn định trên các bộ dữ liệu ung thư phức tạp [2].

2.1.2 Mô hình Hierarchical Clustering

Hierarchical Clustering tiếp cận bài toán theo hướng xây dựng một cấu trúc lồng ghép (bottom-up). Phương pháp này phù hợp do bản chất sinh học của các loại bệnh ung thư thường tồn tại mối quan hệ phân cấp. Quy trình cụ thể qua 5 bước sau:

Bước 1: Tính toán ma trận khoảng cách (Distance Matrix)

Đầu tiên, dữ liệu đầu vào (đã qua giảm chiều PCA còn 30 thành phần) được sử dụng để tính toán độ tương đồng giữa các cặp mẫu. Nhóm sử dụng khoảng cách Euclidean để đo lường sự khác biệt giữa hai điểm dữ liệu xi và xj. Kết quả là một ma trận đối xứng N×N lưu trữ khoảng cách giữa tất cả các cặp điểm, làm cơ sở cho việc gom nhóm.

Bước 2: Xác định phương pháp liên kết (Linkage Criterion)

Nhóm lựa chọn phương pháp Ward Linkage. Ward Linkage hướng tới việc tối thiểu hóa sự tăng lên của phương sai (variance) trong cụm khi gộp hai nhóm lại. Công thức tính khoảng cách Ward giữa cụm A và cụm B được cài đặt như sau:

$$d(A,B) = \sqrt{\frac{2n_A n_B}{n_A + n_B}} ||\bar{x}_A - \bar{x}_B||_2$$

Bước 3: Thuật toán gom cụm (Agglomerative Algorithm)

Quy trình gom cụm được thực hiện theo vòng lặp:

1. Khởi tạo mỗi điểm dữ liệu là một cụm riêng biệt (801 cụm ban đầu).

2. Trong mỗi bước lặp, thuật toán duyệt qua tất cả các cặp cụm hiện có để tính khoảng cách Ward.
3. Cặp cụm có khoảng cách nhỏ nhất sẽ được hợp nhất thành một cụm mới. Trọng tâm của cụm mới được cập nhật lại.
4. Quá trình lặp lại cho đến khi số lượng cụm giảm xuống còn k cụm mong muốn.

Bước 4: Xác định số cụm tối ưu (k)

Do thuật toán có độ phức tạp tính toán lớn ($O(N^3)$), nhóm đã áp dụng kỹ thuật lấy mẫu (sampling) ngẫu nhiên 300 điểm dữ liệu để tìm tham số tối ưu. Hai phương pháp được sử dụng song song:

Phương pháp Elbow: Quan sát tổng bình phương khoảng cách nội bộ (WCSS). Đồ thị cho thấy điểm uốn (elbow point) xuất hiện rõ tại k=5

1. Hệ số Silhouette: Đo lường độ tách biệt giữa các cụm.
2. Kết hợp kết quả phân tích định lượng với kiến thức miền (bộ dữ liệu gồm 5 loại ung thư), nhóm quyết định chọn k=5 cho mô hình cuối cùng.

Bước 5: Trực quan hóa và Đánh giá

Để kiểm chứng cấu trúc phân cấp, nhóm sử dụng thư viện scipy để vẽ biểu đồ Dendrogram dựa trên ma trận Ward Linkage, qua đó quan sát được các nhánh tách biệt tại ngưỡng cắt tương ứng với 5 cụm.

Cuối cùng, mô hình được huấn luyện trên toàn bộ tập dữ liệu với k=5. Kết quả phân cụm được trực quan hóa trên không gian 2 chiều (PC1, PC2) và đánh giá bằng các chỉ số Adjusted Rand Index (ARI) và Normalized Mutual Information (NMI) so với nhãn thực tế.

Trong tin sinh học, Hierarchical Clustering được xem là 'tiêu chuẩn vàng' (gold standard) kể từ công bố kinh điển của Eisen và cộng sự (1998) [3]. Phương pháp này cho phép trực quan hóa dữ liệu dưới dạng bản đồ nhiệt (heatmap) và cây phả hệ, giúp các nhà nghiên cứu dễ dàng nhận diện các phân nhóm gen đồng biểu hiện. Mặc dù vậy, độ phức tạp tính toán cao và tính chất không thể đảo ngược (irreversible) của các bước gộp cụm là rào cản lớn khi áp dụng trên các tập dữ liệu quy mô lớn.

2.1.3 Mô hình DBSCAN

DBSCAN là thuật toán phân cụm dựa trên mật độ, có khả năng phát hiện các cụm hình dạng bất kỳ và xử lý nhiễu hiệu quả. Quy trình thực hiện bao gồm các bước sau:

- **Bước 1: Xác định tham số bán kính tối ưu (ϵ)**
    1. Sử dụng thuật toán NearestNeighbors để tính khoảng cách từ mỗi điểm dữ liệu đến láng giềng gần nhất thứ k của nó (với k=5).
    2. Sắp xếp các khoảng cách này theo thứ tự tăng dần và vẽ đồ thị.
    3. Áp dụng thuật toán KneeLocator để tìm "điểm khuỷu tay" (elbow/knee point). Giá trị khoảng cách tại điểm này được xác định là ϵ tối ưu, giúp cân bằng giữa việc tách biệt nhiễu và duy trì cấu trúc cụm.
- **Bước 2: Xây dựng hàm truy vấn vùng (Region Query)**

Thành phần cốt lõi đầu tiên của thuật toán là hàm region_query. Hàm này nhận đầu vào là một điểm dữ liệu và bán kính ϵ, sau đó tính toán khoảng cách Euclidean từ điểm đó đến tất cả các điểm còn lại trong tập dữ liệu. Những điểm có khoảng cách nhỏ hơn hoặc bằng ϵ được xác định là "hàng xóm" trực tiếp, đóng vai trò quan trọng trong việc xác định mật độ cục bộ.

- **Bước 3: Xây dựng hàm mở rộng cụm (Expand Cluster)**
    1. Gán nhãn cụm hiện tại cho điểm đó.
    2. Thu thập tất cả hàng xóm của nó vào một hàng đợi.
    3. Duyệt qua từng hàng xóm trong hàng đợi: nếu hàng xóm đó cũng là một điểm lõi, các hàng xóm của nó sẽ tiếp tục được thêm vào hàng đợi. Quá trình này lặp lại cho đến khi không còn điểm nào có thể kết nối mật độ được nữa, tạo thành một cụm hoàn chỉnh.
- **Bước 4: Vòng lặp chính và Xử lý nhiễu**

  Thuật toán duyệt tuần tự qua từng điểm dữ liệu trong tập mẫu N:
    - Nếu điểm đó đã được duyệt, thuật toán bỏ qua.
    - Nếu điểm đó chưa được duyệt, thực hiện truy vấn vùng. Nếu số lượng hàng xóm nhỏ hơn min_samples, điểm đó tạm thời được gán nhãn là Nhiễu (Noise, nhãn -1).
    - Ngược lại, nếu đủ điều kiện mật độ, hàm mở rộng cụm (Bước 3) được kích hoạt để xây dựng một cụm mới.
- **Bước 5: Đánh giá kết quả**

Kết quả cuối cùng của thuật toán bao gồm các cụm dữ liệu được phân tách và một tập hợp các điểm nhiễu. Nhóm tiến hành đánh giá hiệu năng mô hình thông qua chỉ số ARI và NMI so với nhãn gốc, đồng thời kiểm tra tỷ lệ điểm nhiễu để đảm bảo tham số ϵ đã được lựa chọn hợp lý, không loại bỏ quá nhiều thông tin quan trọng của bộ dữ liệu gen.

DBSCAN được giới thiệu bởi Ester và cộng sự (1996) với ưu điểm vượt trội trong việc phát hiện các cụm có hình dạng bất kỳ và xử lý nhiễu [4]. Trong phân tích gen, DBSCAN thường được sử dụng như một bộ lọc để loại bỏ các mẫu ngoại lai (outliers) trước khi thực hiện các phân tích chuyên sâu hơn. Tuy nhiên, hiệu suất của DBSCAN rất nhạy cảm với việc lựa chọn tham số trong không gian dữ liệu nhiều chiều (high-dimensional space) [5].

2.2. Các kỹ thuật Ensemble Clustering

2.2.1 Phương pháp CSPA (Cluster-based Similarity Partitioning Algorithm).

Phương pháp CSPA được Strehl và Ghosh đề xuất như một khung làm việc (framework) để tái sử dụng tri thức từ nhiều phân hoạch khác nhau [6]. CSPA xây dựng một ma trận tương đồng kích thước N×N, thường được gọi là ma trận đồng thuận. Giá trị tại hàng i, cột j của ma trận này đại diện cho tỷ lệ phần trăm số lần mà hai điểm dữ liệu xi và xj được gán cùng một nhãn trong tập hợp các mô hình cơ sở.

Ưu điểm của CSPA là về tính đơn giản và khả năng trực quan hóa mối quan hệ giữa các điểm dữ liệu mà không cần truy cập vào dữ liệu gốc. Phương pháp này đặc biệt hiệu quả khi các mô hình cơ sở có chất lượng tương đồng nhau.

CSPA cũng có hạn chế là độ phức tạp tính toán và chi phí lưu trữ. Việc xây dựng và xử lý ma trận kích thước N×N tiêu tốn tài nguyên bộ nhớ lớn, đạt độ phức tạp $O(N^2)$, gây khó khăn khi áp dụng cho các bộ dữ liệu có số lượng mẫu quá lớn. Ngoài ra, trong phiên bản CSPA tiêu chuẩn, vai trò của các mô hình cơ sở là ngang nhau, nên có thể làm giảm độ chính xác nếu tồn tại những mô hình kém chất lượng trong tập hợp.

2.2.2 Phương pháp SCENA và vai trò của thông tin láng giềng (KNN).

Kỹ thuật tăng cường ái lực mạng lưới (Network Affinity Enhancement) được phát triển dựa trên nghiên cứu của Zhang và cộng sự (2021) về phân cụm dữ liệu đơn bào [7]. SCENA tích hợp cơ chế tăng cường mạng lưới để tinh chỉnh kết quả hợp nhất.

Cụ thể, thuật toán thực hiện phép nhân giữa ma trận đồng thuận toàn cục (từ các mô hình cơ sở) với ma trận láng giềng cục bộ được xây dựng từ dữ liệu gốc, giúp lan truyền độ tương đồng giữa các mẫu gen thông qua các điểm trung gian.

Ưu điểm của SCENA là việc tận dụng thông tin láng giềng gần nhất (KNN) để bảo tồn cấu trúc hình học (manifold) của dữ liệu. Mạng lưới KNN đóng vai trò như một bộ lọc nhiễu tự nhiên, giúp loại bỏ các liên kết ngẫu nhiên sai lệch xuất hiện trong quá trình chạy các mô hình đơn lẻ, giúp nâng cao đáng kể độ chính xác khi phân loại các nhóm tế bào hoặc bệnh phẩm có đặc điểm sinh học gần giống nhau.

Tuy nhiên, cũng tồn tại hạn chế về chi phí tính toán do yêu cầu phải xây dựng và thao tác trên đồ thị láng giềng từ dữ liệu gốc, thay vì chỉ làm việc trên nhãn như CSPA. Độ phức tạp của thuật toán sẽ tăng lên khi số chiều dữ liệu hoặc số lượng mẫu quá lớn. Hiệu suất của SCENA phụ thuộc vào việc lựa chọn tham số K (số lượng láng giềng) phù hợp; việc chọn K không tối ưu có thể dẫn đến việc gộp sai các cụm nhỏ hoặc không đủ sức mạnh để kết nối các phần của cùng một cụm lại với nhau.

2.3. Tổng kết lý do chọn hướng tiếp cận

Qua quá trình tìm hiểu và phân tích, nhóm chúng em nhận thấy có một khoảng trống đáng kể giữa các phương pháp phân cụm đơn lẻ và các kỹ thuật tổ hợp truyền thống khi áp dụng trên dữ liệu biểu hiện gen. Các thuật toán cơ bản như K-Means++ hay Hierarchical Clustering dù phổ biến nhưng thường thiếu ổn định và nhạy cảm với đặc thù nhiều nhiễu của dữ liệu sinh học. Trong khi đó, kỹ thuật Ensemble cơ bản như CSPA dù cải thiện được độ ổn định nhưng lại bộc lộ hai hạn chế lớn chưa được giải quyết triệt để.

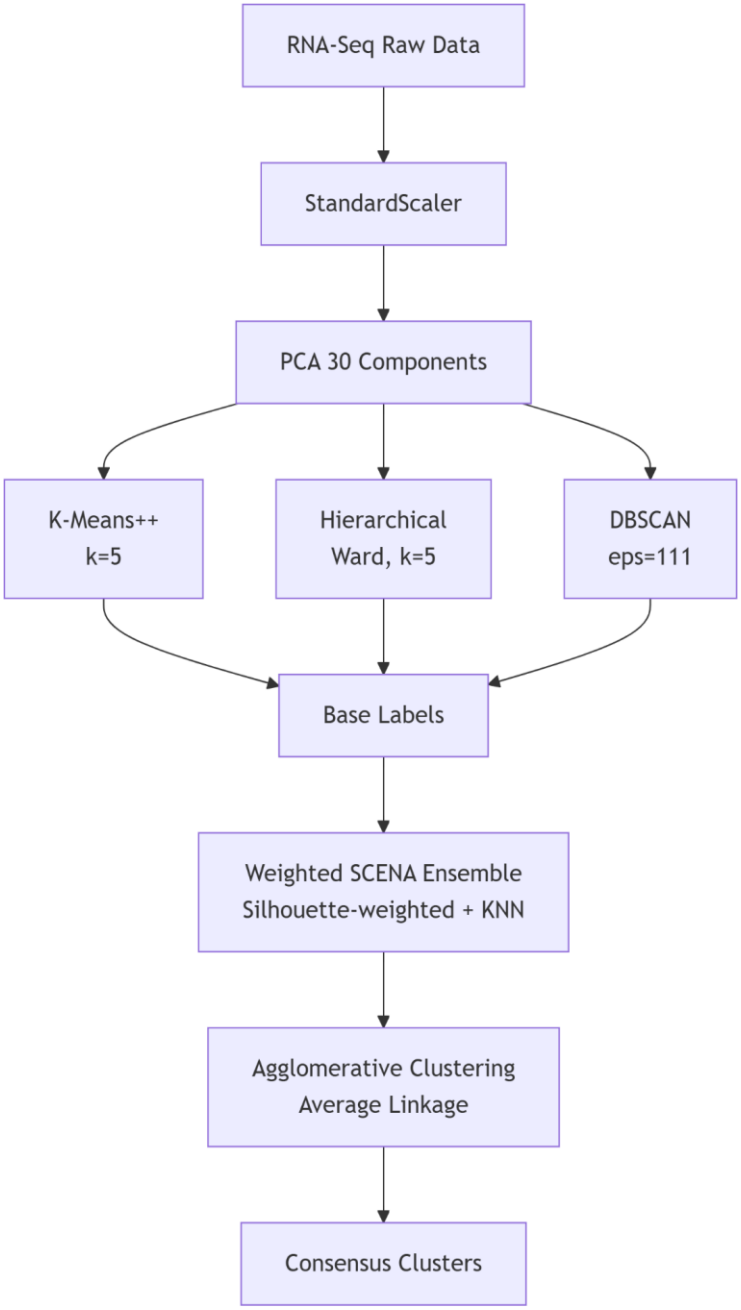
Hạn chế đầu tiên nằm ở cơ chế đồng thuận ngang hàng của CSPA truyền thống. Phương pháp này mặc định coi trọng số của tất cả các mô hình cơ sở là như nhau, dẫn đến việc kết quả cuối cùng có thể bị kéo lệch bởi các mô hình có hiệu suất kém.

Hạn chế thứ hai là sự mất mát thông tin cấu trúc địa phương. Khi chuyển đổi từ dữ liệu gốc sang ma trận nhãn để thực hiện Ensemble, CSPA vô tình bỏ qua các thông tin hình học chi tiết giữa các điểm dữ liệu. Điều này đặc biệt lãng phí đối với dữ liệu gen vốn chứa đựng nhiều thông tin quan trọng trong cấu trúc láng giềng. Đây chính là lý do nhóm đề xuất giải pháp lai ghép Weighted SCENA. Việc kết hợp cơ chế trọng số Weighted CSPA giúp lọc bỏ tác động của các mô hình kém, trong khi kỹ thuật SCENA

với mạng lưới KNN giúp khôi phục lại cấu trúc hình học từ dữ liệu gốc. Việc gán trọng số cho các mô hình thành phần dựa trên độ đo nội bộ (như Silhouette) đã được chứng minh là giúp cải thiện độ chính xác so với cơ chế đồng thuận ngang hàng truyền thống [8].

3. PHƯƠNG PHÁP ĐỀ XUẤT

Dữ liệu biểu hiện gen RNA-Seq mà nhóm chọn có đặc trưng về số chiều rất lớn, nhiễu cao và số lượng mẫu hạn chế, khiến các thuật toán phân cụm đơn lẻ dễ cho kết quả thiếu ổn định. Từ những vấn đề đó thì nhóm đã nghiên cứu đề xuất một phương pháp PCA-Enhanced Weighted SCENA Heterogenous Ensemble Clustering. Mục tiêu chính của phương pháp là cải thiện tính ổn định và nâng cao chất lượng kết quả phân cụm thông qua việc khai thác ưu điểm của nhiều thuật toán khác nhau.



Hình 1. Kiến trúc tổng thể của mô hình phân cụm Ensemble.

3.1. Bước 1: Tiền xử lý và giảm chiều dữ liệu (Preprocessing & PCA)

- Mục tiêu: Chuẩn hóa dữ liệu về cùng một miền giá trị và trích xuất các thành phần đặc trưng cốt lõi, tạo dữ liệu đầu vào phù hợp cho các thuật toán phân cụm.
- Đầu vào: Bộ dữ liệu RNA-Seq thô gồm 801 mẫu và khoảng 20.000 gen.
- Cách tiếp cận:
    - Chuẩn hóa dữ liệu (StandardScaler): Đưa dữ liệu về phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1, nhằm tránh hiện tượng các gen có biên độ lớn lấn át các gen khác.
    - Giảm chiều bằng PCA: Áp dụng Principal Component Analysis (PCA) để giảm dữ liệu xuống 30 thành phần chính.
- Lý do lựa chọn: Dữ liệu RNA-Seq có số lượng gen rất lớn trong khi số lượng mẫu tương đối hạn chế, dẫn đến không gian đặc trưng có độ phức tạp cao và chứa nhiều nhiễu. Trong không gian nhiều chiều, việc tính toán khoảng cách giữa các mẫu trở nên kém hiệu quả, khiến các thuật toán phân cụm khó xác định ranh giới cụm rõ ràng và dễ bị ảnh hưởng bởi nhiễu. Việc giảm chiều dữ liệu xuống 30 thành phần chính giúp cô đọng các thông tin quan trọng, loại bỏ các chiều ít mang ý nghĩa phân biệt, đồng thời tạo ra không gian đặc trưng gọn hơn, phù hợp cho các thuật toán phân cụm hoạt động ổn định, đặc biệt là DBSCAN.

3.2. Bước 2: Tạo sinh các phân cụm cơ sở (Base Clustering Genration)

- Mục tiêu: Tạo ra sự đa dạng trong các kết quả phân cụm, đó yếu tố quan trọng của một hệ thống ensemble.
- Cách tiếp cận là ba mô hình phân cụm cơ sở được triển khai song song trên dữ liệu PCA 30 chiều:

    + K-Means++: Số cụm k được lựa chọn dựa trên thực nghiệm bằng phương pháp Elbow nhằm đảm bảo khả năng phân tách cụm hợp lý.

    + Hierarchical Clustering: Sử dụng phương pháp liên kết Ward Linkage nhằm tối thiểu hóa phương sai trong cụm, giúp tạo ra các cụm chặt chẽ và đồng đều.

    + DBSCAN: Các tham số eps và min_samples được tinh chỉnh nhằm phát hiện cấu trúc mật độ của dữ liệu và tách các điểm nhiễu.

- Lý do lựa chọn: Việc kết hợp các thuật toán có cách tiếp cận khác nhau giúp hệ thống ensemble tận dụng được ưu điểm của từng mô hình và tăng tính ổn định của kết quả phân cụm.

3.3. Bước 3: Cơ chế đồng thuận có xử lý nhiễu (Noise-Aware Consensus)

- Mục tiêu: Tổng hợp kết quả từ các mô hình cơ sở có tính đến chất lượng từng mô hình và cấu trúc cục bộ của dữ liệu.
- Cách tiếp cận:

1. Gán trọng số cho từng mô hình: Trọng số được tính dựa trên chỉ số Silhouette Score của mỗi mô hình trên dữ liệu PCA30. Mô hình có chất lượng cao được trao trọng số lớn hơn.
2. Xây dựng ma trận đồng thuận có trọng số (Weighted Co-association Matrix):
    - Mỗi cặp mẫu (i, j) nhận giá trị bằng tổng có trọng số của các lần chúng được xếp chung cụm.
    - Với DBSCAN, các mẫu bị gán nhãn nhiễu (label = -1) không tham gia đóng góp vào ma trận đồng thuận (các liên kết tương ứng bị gán bằng 0).
3. Tăng cường bằng đồ thị K-NN (SCENA Enhancement):
    - Xây dựng ma trận kề K-NN (k=20) dựa trên khoảng cách Euclid trong không gian PCA30.
    - Nhân ma trận đồng thuận có trọng số với ma trận K-NN để lan truyền sự đồng thuận qua các láng giềng gần, tạo ra ma trận SCENA chuẩn hóa về [0, 1].
    - Bước này giúp làm mượt kết quả đồng thuận và tăng tính ổn định, đặc biệt hữu ích với dữ liệu nhiễu cao như RNA-Seq.

3.4. Bước 4: Phân cụm cuối cùng (Final Clustering)

- Ma trận tương đồng M được chuyển đổi thành ma trận khoảng cách theo công thức:

$$D=1-M$$

- Sau đó, thuật toán Agglomerative Clustering với liên kết trung bình (Average Linkage) được áp dụng trên ma trận D để thu được các cụm cuối cùng, phản ánh tốt nhất sự đồng thuận của toàn bộ hệ thống ensemble.

3.5. Thuật toán tổng quát

Để tổng hợp quy trình đề xuất, nhóm trình bày giải thuật Weighted SCENA-based Ensemble dưới dạng mã giả hình thức như sau:

Weighted SCENA-based Ensemble Clustering

Input:

X: Tập dữ liệu Gen Expression thô (N mẫu)

M: Tập hợp các mô hình cơ sở {K-Means++, Hierarchical, DBSCAN}

k_knn: Số lượng láng giềng cho SCENA (mặc định 20)

Output:

Y: Vector nhãn phân cụm cuối cùng (N × 1)

Process

1. Preprocessing

X_scaled ← StandardScaler(X)

X_pca ← PCA(X_scaled, n_components = 30)

2. Base Clustering & Weighting

H ← ∅ (Tập hợp các vector nhãn)

W ← ∅ (Tập hợp trọng số)

For each model m_i in M:

$labels_i$ ← m_i.fit_predict(X_pca)

$score_i$ ← SilhouetteScore(X_pca, $labels_i$)

Add $labels_i$ to H

Add $score_i$ to W

3. Construct Weighted Consensus Matrix (C)

Khởi tạo C là ma trận 0 kích thước N × N

For each cặp mẫu (u, v):

C[u, v] =

$(\Sigma (W_i \times I(labels_i[u] == labels_i[v]))) / (\Sigma W_i)$

Ghi chú: Bỏ qua đóng góp của mô hình nếu nhãn là Noise (-1)

4. SCENA Enhancement

A ← KNeighborsGraph(X_pca, k = k_knn)

S ← C × A (Dot product để lan truyền ái lực)

S_norm ← Normalize(S)

5. Final Clustering

D ← 1 − S_norm (Chuyển sang ma trận khoảng cách)

Y ← AgglomerativeClustering(D, linkage = "complete")

6. Return

Trả về Y

4. KẾT QUẢ THỰC NGHIỆM

4.1. Chuẩn bị tập dữ liệu

Nhóm đã nghiên cứu, tìm hiểu và quyết định sử dụng bộ dữ liệu Gen Expression Cancer RNA-Seq, được thu thập từ các nguồn công khai như UCI Machine Learning Repository.

- Link UCI Machine Learning Repository:
https://archive.ics.uci.edu/dataset/401/gen+expression+cancer+rna+seq

- Bộ dữ liệu bao gồm:

- 801 mẫu tương ứng với 801 bệnh nhân.
- 20.531 đặc trưng ban đầu, mỗi đặc trưng đại diện cho mức độ biểu hiện của một gen.
- 5 lớp nhãn thực tế (ground truth), tương ứng với các loại ung thư:
    - BRCA: Ung thư biểu mô xâm lấn vú (Breast Invasive Carcinoma).
    - KIRC: Ung thư biểu mô tế bào thận (Kidney Renal Clear Cell Carcinoma).
    - COAD: Ung thư kết tràng (Colon Adenocarcinoma).
    - LUAD: Ung thư phổi (Lung Adenocarcinoma).
    - PRAD: Ung thư tuyến tiền liệt (Prostate Adenocarcinoma).

Toàn bộ quá trình phân cụm được thực hiện theo hướng học không giám sát, trong đó nhãn thực tế không được sử dụng trong quá trình huấn luyện mô hình. Các nhãn này chỉ được dùng ở bước đánh giá cuối cùng nhằm kiểm chứng chất lượng phân cụm thông qua các chỉ số Adjusted Rand Index (ARI) và Normalized Mutual Information (NMI). Để đánh giá độ tương đồng giữa kết quả phân cụm và nhãn bệnh học, nhóm sử dụng chỉ số Adjusted Rand Index (Hubert & Arabie, 1985). Bên cạnh đó, chỉ số Silhouette Score được sử dụng để đánh giá mức độ gắn kết và phân tách của các cụm dựa trên chính cấu trúc nội tại của dữ liệu, không phụ thuộc vào nhãn thật.

4.2. Thiết lập thực nghiệm

4.2.1. Môi trường triển khai

- Nền tảng tính toán: Google Colab (sử dụng GPU T4).
- Ngôn ngữ lập trình: Python.
- Các thư viện chính:

- NumPy & Pandas: Đóng vai trò nòng cốt trong việc xử lý dữ liệu số, thao tác ma trận và đặc biệt là tự cài đặt (implement from scratch) các thuật toán phân cụm cơ sở (K-Means++, DBSCAN, Hierarchical) mà không phụ thuộc vào thư viện có sẵn.
- Scikit-learn: Được sử dụng để thực hiện các bước tiền xử lý (StandardScaler, PCA), tính toán các chỉ số đánh giá hiệu năng (Silhouette Score, ARI, NMI) và hỗ trợ thuật toán phân cụm phân cấp ở bước hợp nhất cuối cùng.
- Matplotlib & Seaborn: Trực quan hóa dữ liệu và kết quả phân cụm (heatmap, t-SNE scatter plot, biểu đồ phân bố).

4.2.2. Cấu hình tham số mô hình

Các tham số của mô hình không được lựa chọn ngẫu nhiên mà được xác định dựa trên các phương pháp định lượng, thực hiện trên tập dữ liệu đã được giảm chiều bằng PCA.

- Mô hình K-Means++:
  - Số cụm (k): 5.
  - Cơ sở lựa chọn: Sử dụng phương pháp khuỷu tay dựa trên tổng bình phương khoảng cách nội bộ và kiểm chứng chéo với chỉ số Silhouette. Đồ thị cho thấy điểm gãy rõ rệt tại $k = 5$, giá trị này cũng trùng khớp với số lượng lớp nhãn bệnh thực tế trong bộ dữ liệu.
  - Phương pháp khởi tạo: K-Means++, nhằm tối ưu hóa vị trí tâm cụm ban đầu và giảm nguy cơ rơi vào nghiệm tối ưu cục bộ.
  - Kiểm soát ngẫu nhiên: Thiết lập random_state cố định để đảm bảo tính nhất quán và khả năng tái lập của kết quả thực nghiệm.
- Mô hình DBSCAN:
  - Bán kính lân cận (eps): 111.0.
  - Số điểm tối thiểu (min_samples): 5.
  - Cơ sở lựa chọn: Tham số eps được xác định thông qua phân tích biểu đồ k-distance với k = 5 trên dữ liệu đã được giảm chiều bằng PCA. Quan sát đồ thị khoảng cách k-lân cận được sắp xếp tăng dần, điểm uốn xuất hiện rõ rệt tại giá trị xấp xỉ 111, do đó eps = 111 được lựa chọn làm ngưỡng phù hợp. Thiết lập này

giúp DBSCAN phân biệt hiệu quả giữa các vùng có mật độ cao hay còn gọi là cụm và các điểm nằm rải rác được xem là nhiễu trong không gian PCA 30 chiều.

- Mô hình Hierarchical Clustering:

- Số cụm: k = 5 Giá trị này được thống nhất lựa chọn để đảm bảo tính công bằng khi so sánh hiệu năng với K-Means++ và kết quả tìm được từ DBSCAN.
- Phương pháp liên kết: Sử dụng Ward Linkage.
- Cơ sở lý thuyết và Cài đặt: Trong thuật toán tự xây dựng, nhóm áp dụng công thức khoảng cách Ward nhằm mục tiêu tối thiểu hóa sự tăng lên của tổng phương sai (within-cluster variance) khi gộp hai cụm lại với nhau. Phương pháp này có ưu điểm vượt trội so với các phương pháp liên kết đơn (Single Linkage) hay liên kết đầy đủ (Complete Linkage) ở chỗ nó thường tạo ra các cụm có kích thước tương đối đồng đều, cấu trúc chặt chẽ và ít bị ảnh hưởng bởi nhiễu cục bộ.

4.2.3. Các kịch bản so sánh

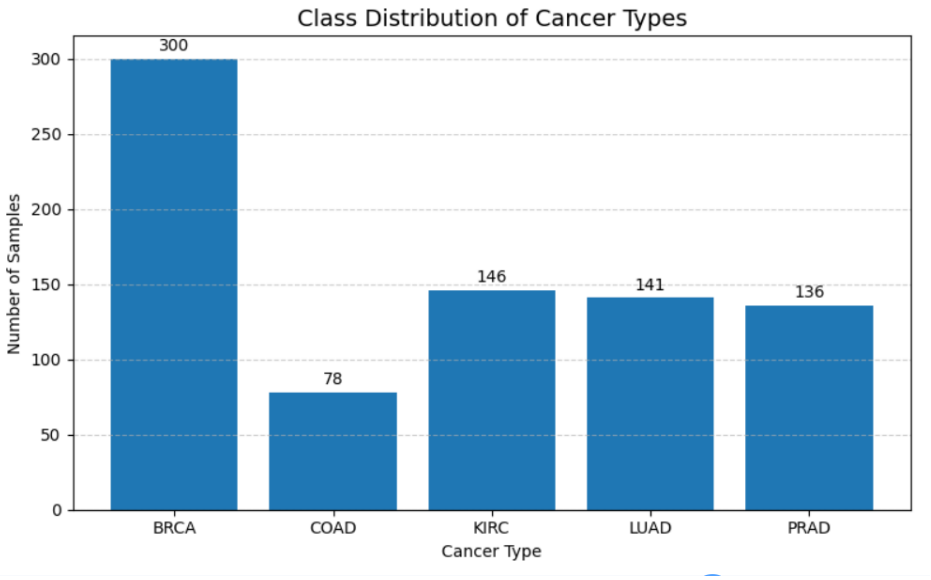
Để đánh giá hiệu quả của phương pháp đề xuất, nhóm tiến hành thiết lập các kịch bản so sánh giữa mô hình ensemble và các mô hình phân cụm cơ sở. Cụ thể:

- Các mô hình cơ sở (Single Models): Bao gồm K-Means++, DBSCAN và Hierarchical Clustering, được triển khai độc lập trên tập dữ liệu đã giảm chiều bằng PCA. Các mô hình này đóng vai trò làm đường cơ sở (baseline) để đánh giá chất lượng phân cụm khi sử dụng từng thuật toán riêng lẻ.
- Mô hình đề xuất (Proposed Ensemble): Kết quả nhãn từ ba mô hình cơ sở trên được tổng hợp thông qua Ma trận đồng thuận có xử lý nhiễu. Tại bước này, các mẫu bị DBSCAN xác định là nhiễu (Label -1) sẽ bị loại bỏ liên kết để làm sạch ma trận. Trên cơ sở ma trận đồng thuận đã làm sạch, thuật toán Agglomerative Clustering được áp dụng để xác định cấu trúc phân nhóm cuối cùng.

Việc so sánh giữa các kịch bản trên nhằm làm rõ mức độ cải thiện của mô hình ensemble về chất lượng phân cụm và tính ổn định so với từng thuật toán phân cụm đơn lẻ.

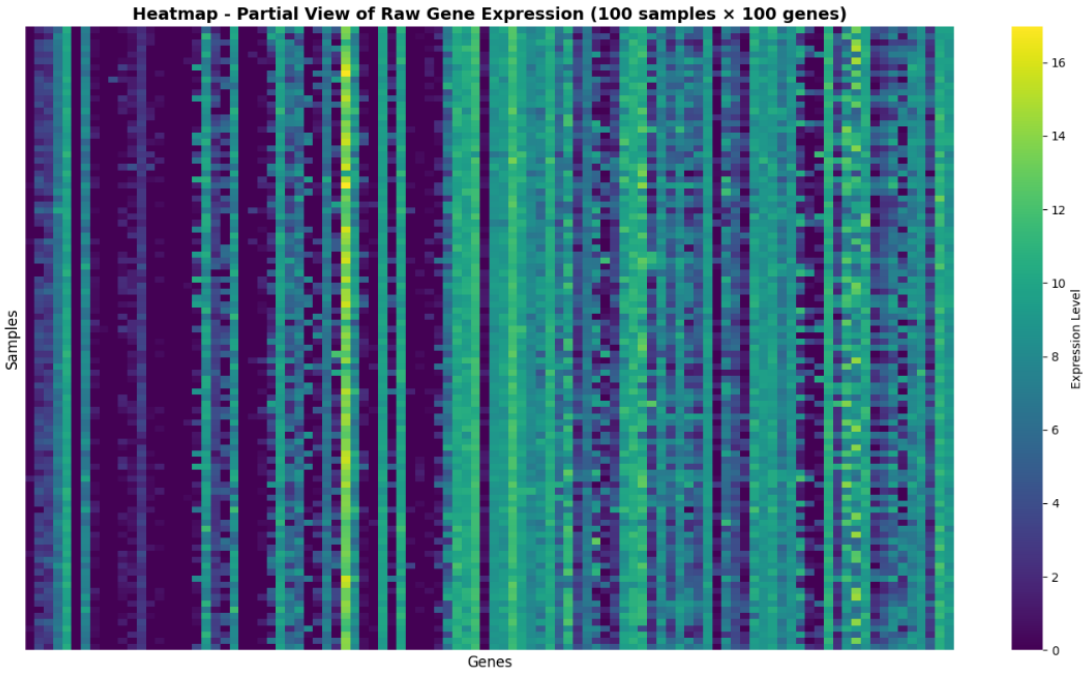
4.3. Phân tích trực quan dữ liệu

4.3.1. Trực quan dữ liệu gốc (EDA trên tập RNA-Seq)



Hình 2. Phân bố số lượng mẫu theo từng loại ung thư

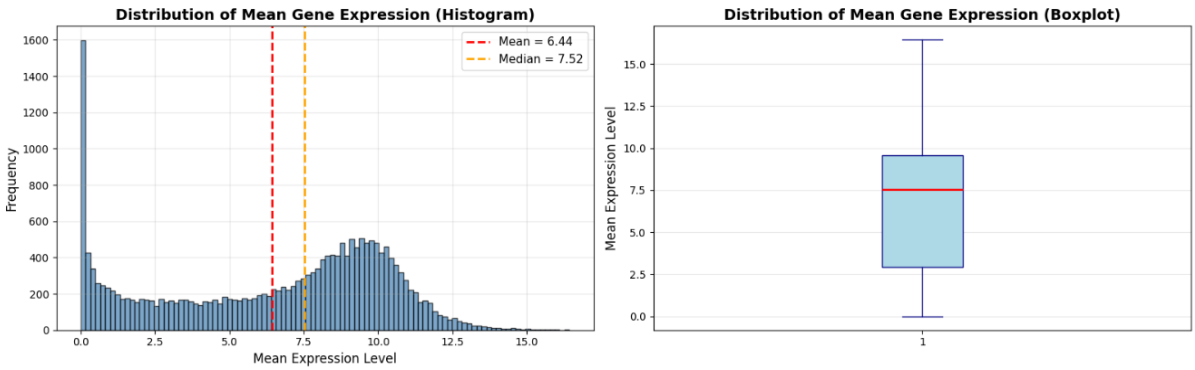
Biểu đồ cho thấy dữ liệu bị mất cân bằng giữa các lớp ung thư, trong đó BRCA chiếm tỷ lệ lớn nhất và COAD có số mẫu ít nhất. Đây là đặc thù thường gặp trong dữ liệu y sinh thực tế. Thông tin này được sử dụng để diễn giải kết quả đánh giá phân cụm bằng các chỉ số ARI và NMI ở các bước sau, nhằm đảm bảo việc so sánh được đặt trong đúng bối cảnh phân bố dữ liệu.



Hình 3. Heatmap biểu hiện gen trên tập dữ liệu gốc

Biểu đồ heatmap cho thấy mức độ biểu hiện gen thay đổi mạnh giữa các gen và các mẫu, không xuất hiện ranh giới cụm rõ ràng trong không gian đặc trưng gốc. Điều này phản ánh dữ liệu có tính nhiễu cao và số chiều lớn, gây khó khăn cho việc phân cụm trực tiếp, từ đó cho thấy sự cần thiết của bước giảm chiều trước khi áp dụng các thuật

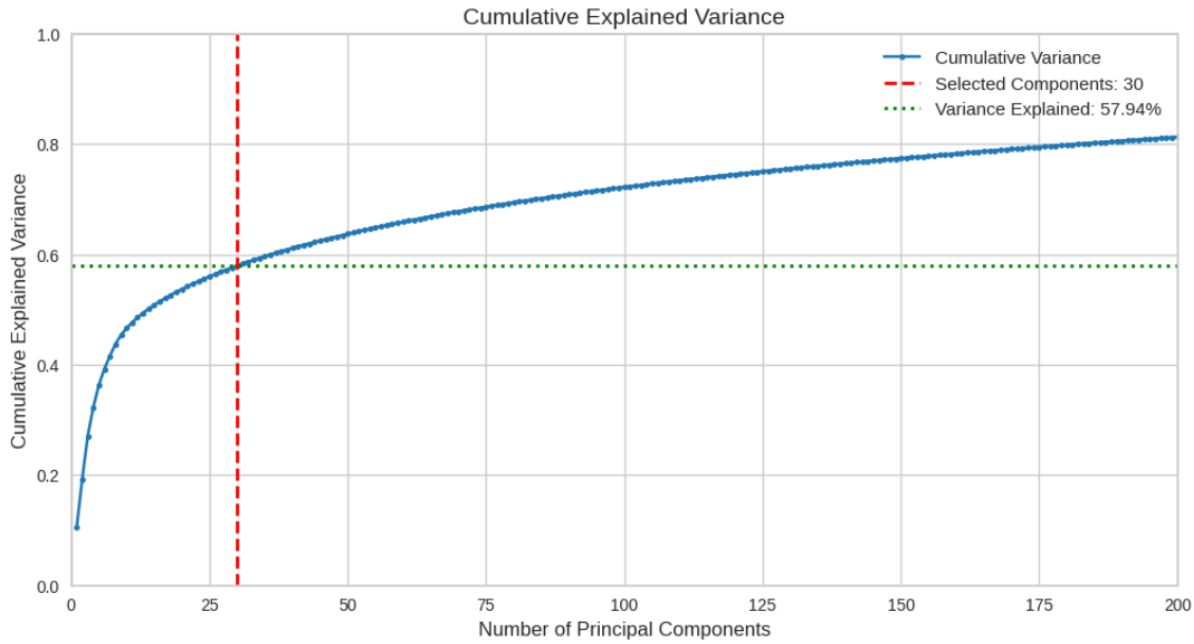
toán phân cụm.



Hình 4. Phân bố mức độ biểu hiện gen trung bình

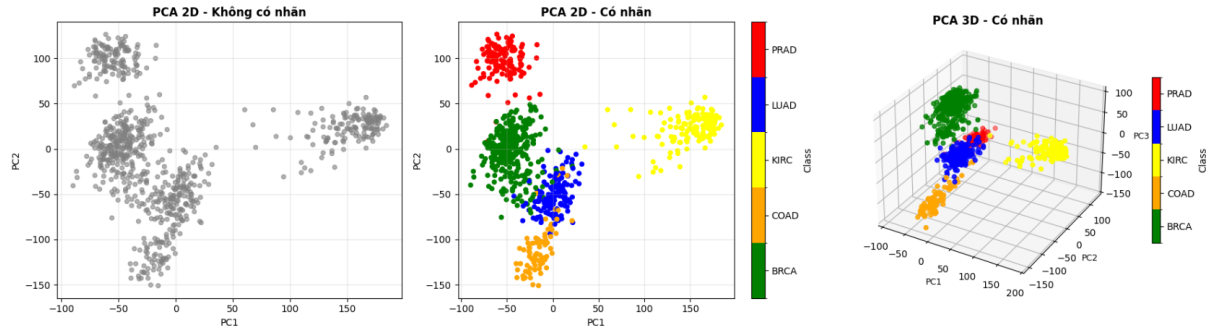
Hai biểu đồ cho thấy mức độ biểu hiện gen phân bố không đều và có độ biến thiên lớn, phản ánh dữ liệu chứa nhiều nhiễu và giá trị ngoại lai. Do đó cần chuẩn hóa và giảm chiều dữ liệu trước khi thực hiện phân cụm giúp tránh việc các gen có biên độ lớn lấn át các tín hiệu sinh học quan trọng khác.

4.3.2. Trực quan dữ liệu sau giảm chiều PCA



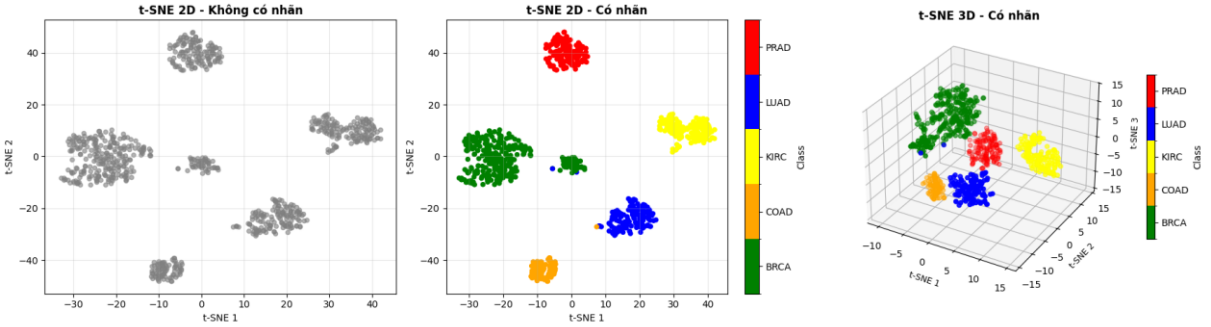
Hình 5. Biểu đồ Scree Plot và Tỷ lệ phương sai tích lũy

Biểu đồ cho thấy phương sai tích lũy tăng nhanh ở các thành phần đầu và tăng chậm dần sau khoảng 30 thành phần chính. Tại mốc 30 PCs, dữ liệu giữ lại khoảng 57.94% tổng phương sai. Việc lựa chọn 30 chiều giúp giảm đáng kể số chiều dữ liệu, loại bỏ các biến thiên kém ổn định và tạo không gian đặc trưng phù hợp hơn cho các thuật toán phân cụm, đặc biệt là DBSCAN.



Hình 6. Trực quan dữ liệu sau giảm chiều bằng PCA (2D và 3D)

Sau khi chiếu dữ liệu xuống không gian PCA thấp chiều, các mẫu bắt đầu hình thành những cụm tương đối rõ ràng, trong đó một số lớp như KIRC và PRAD có mức độ tách biệt tốt, dù vẫn còn hiện tượng chồng lấn giữa các nhóm ung thư có đặc điểm gần nhau. Kết quả này cho thấy việc giữ lại 30 thành phần chính đã bảo toàn được cấu trúc phân nhóm tổng thể của dữ liệu, tạo không gian đặc trưng phù hợp cho các thuật toán phân cụm ở các bước tiếp theo.



Hình 7. Biểu diễn t-SNE 2D và 3D của dữ liệu RNA-Seq

Biểu đồ t-SNE 2D và 3D cho thấy các mẫu dữ liệu được phân tách thành những cụm khá rõ ràng, với mức độ chồng lấn thấp giữa các nhóm. Kết quả này cho thấy dữ liệu RNA-Seq tồn tại cấu trúc phân cụm tự nhiên ở mức cục bộ, đồng thời khẳng định dữ liệu phù hợp để áp dụng các thuật toán phân cụm và mô hình ensemble trong các bước tiếp theo.

4.4. Đánh giá kết quả đạt được

4.4.1 Mô hình baseline K-MEANS++

Kết quả thuật toán K-Means++ đạt hiệu suất phân loại cao. Với chỉ số ARI đạt 0.9832 và NMI đạt 0.9756, mô hình đã xác định các tâm cụm cực kỳ chính xác, tạo ra kết quả phân hoạch gần như trùng khớp hoàn toàn với 5 nhóm bệnh ung thư thực tế. Chỉ số Silhouette Score của K-Means++ ở mức 0.3701. Hiện tượng này phản ánh đặc thù

của dữ liệu biểu hiện gen sau khi giảm chiều: mặc dù các điểm dữ liệu được gán nhãn rất đúng về mặt sinh học (ARI cao), nhưng về mặt hình học, khoảng cách giữa các cụm bệnh lý này trong không gian vector không quá lớn hoặc có sự tiếp giáp nhất định, khiến chỉ số độ tách biệt (Silhouette) không đạt mức tuyệt đối.

4.4.2 Mô hình baseline HIERARCHICAL

Kết quả thực nghiệm cho thấy thuật toán Hierarchical Clustering hoạt động rất hiệu quả trên tập dữ liệu đã qua xử lý. Với chỉ số ARI đạt 0.9907 và NMI đạt 0.9860, mô hình cho thấy sự tương đồng gần như tuyệt đối giữa các cụm dự đoán và nhãn bệnh thực tế. Ward Linkage giúp tối thiểu hóa phương sai, giúp các nhóm dữ liệu được gom gọn và tách biệt rõ ràng.

4.4.3 Mô hình baseline DBSCAN

Kết quả cho thấy thuật toán DBSCAN hoạt động rất hiệu quả trong việc nhận diện cấu trúc dữ liệu dựa trên mật độ. Với chỉ số ARI đạt 0.9577 và NMI đạt 0.9400, mô hình đã phân tách các nhóm bệnh phẩm gần như trùng khớp hoàn toàn với nhãn thực tế. Mặc dù chỉ số Silhouette Score ở mức 0.3663 do đặc thù của thuật toán trong việc xử lý các điểm nhiễu và các cụm không có hình cầu, nhưng độ chính xác phân loại cao đã giúp DBSCAN trong việc lọc bỏ các dữ liệu ngoại lai.

4.4.4 Mô hình Ensemble

Kết quả cuối cùng trên mô hình Ensemble Weighted SCENA cho thấy sự cải thiện đáng kể về độ ổn định và chính xác so với các mô hình đơn lẻ. Chỉ số Adjusted Rand Index (ARI) đạt mức 0.9907 và Normalized Mutual Information (NMI) đạt 0.9860. Đây là những giá trị rất cao, tiệm cận mức tuyệt đối, khẳng định rằng mô hình đề xuất đã phân tách thành công các nhóm bệnh phẩm ung thư với tỷ lệ sai sót cực thấp. Điều này có tính hiệu quả của cơ chế Weighted CSPA trong việc gán trọng số thấp cho các mô hình có hiệu suất kém hơn (như DBSCAN trong trường hợp này), từ đó lọc bỏ các nhiễu loạn và giữ lại những đặc trưng phân cụm tốt nhất. Đồng thời, việc tích hợp thông tin láng giềng (KNN) thông qua kỹ thuật SCENA đã giúp củng cố cấu trúc hình học của dữ liệu, đảm bảo kết quả cuối cùng không bị suy giảm chất lượng ngay cả khi một trong các mô hình con hoạt động không tối ưu.

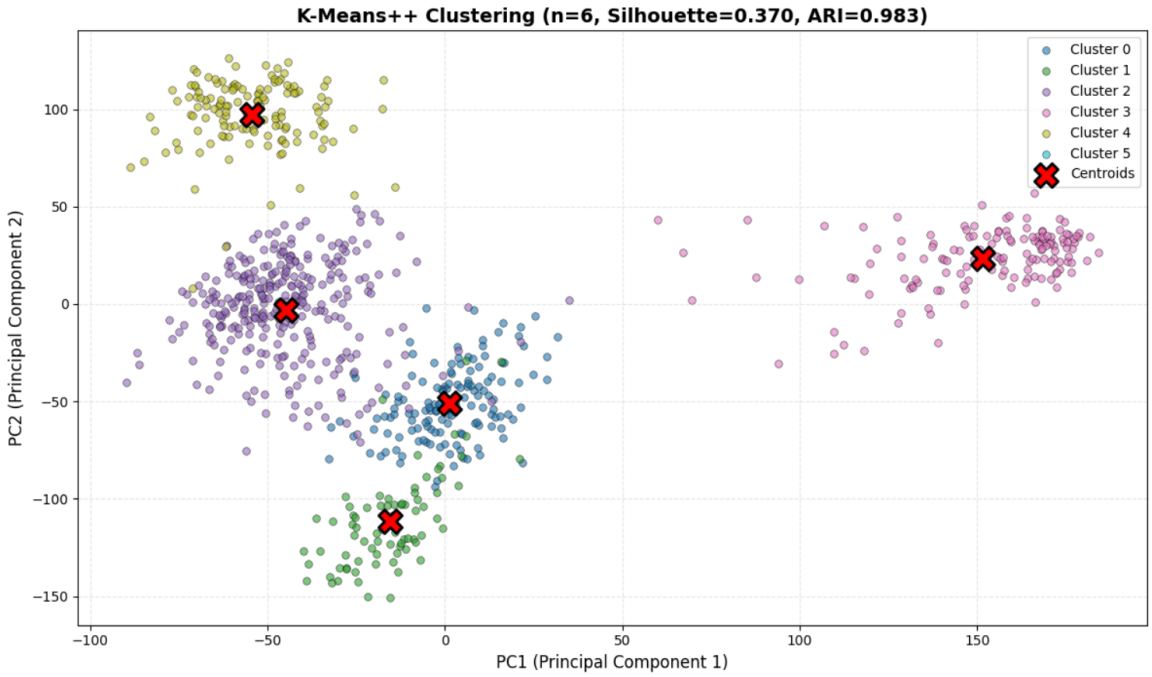
4.5. Kết quả định lượng

So sánh hiệu quả của giải pháp đề xuất, nhóm đã thực hiện so sánh trực tiếp các chỉ số đánh giá giữa mô hình lai ghép Weighted SCENA và ba thuật toán phân cụm cơ sở độc lập. Quá trình thực nghiệm cho thấy sự chênh lệch nhất định về hiệu năng giữa các mô hình đơn lẻ. Cụ thể, thuật toán Hierarchical Clustering hoạt động ổn định nhất trên tập dữ liệu này với chỉ số ARI đạt mức 0.9907. K-Means++ xếp ở vị trí thứ hai với kết quả 0.9851, trong khi DBSCAN cho thấy hiệu năng thấp nhất với ARI chỉ đạt 0.9577. Sự sụt giảm của DBSCAN là do thuật toán này khá kém với việc lựa chọn tham số bán kính trong không gian dữ liệu gen thưa thớt.

Đối với mô hình Ensemble Weighted SCENA được đề xuất, kết quả ghi nhận cuối cùng đạt chỉ số ARI là 0.9907 và NMI là 0.9860. Có thể thấy, kết quả này hoàn toàn ngang bằng với mô hình đơn lẻ tốt nhất (Hierarchical Clustering) và vượt trội hơn hẳn so với K-Means++ cũng như DBSCAN.

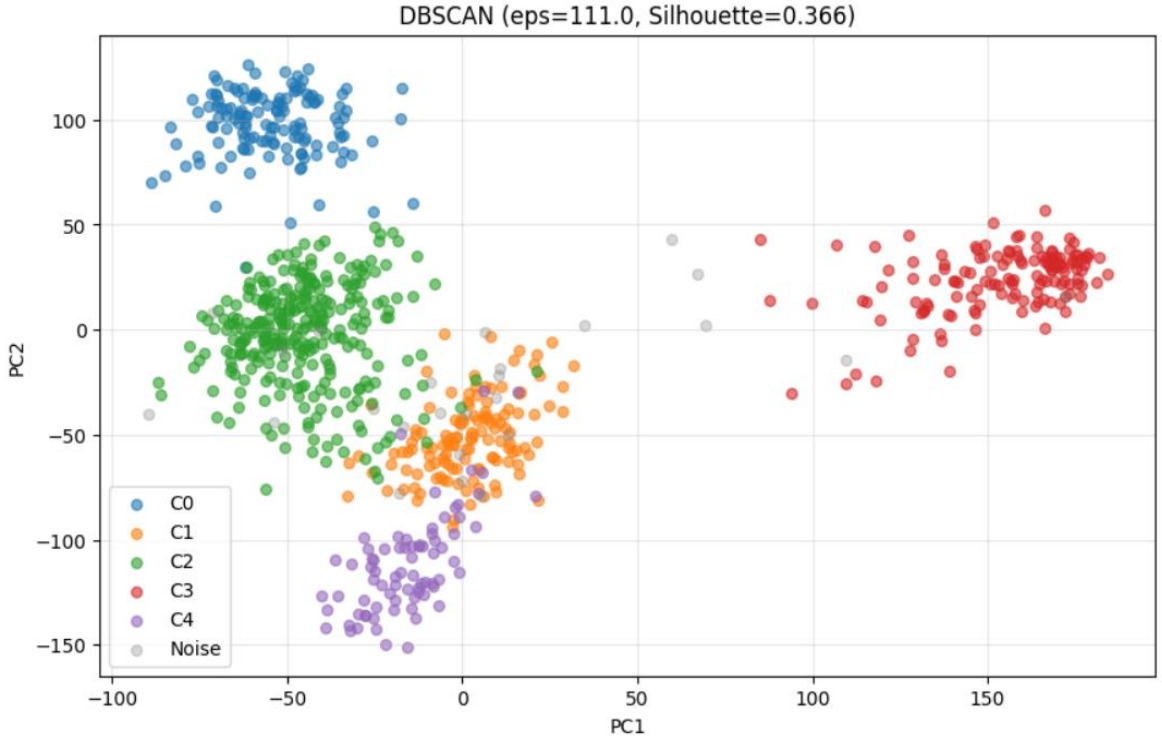
Mặc dù trong tập hợp các mô hình thành phần có DBSCAN với độ chính xác thấp hơn, nhưng hệ thống đã không bị ảnh hưởng. Cơ chế Weighted CSPA đã nhận diện và hạn chế đóng góp của DBSCAN, đồng thời ưu tiên các đặc trưng tốt từ Hierarchical và K-Means++. Nhờ vậy, mô hình tổng hợp vẫn duy trì được mức độ chính xác cao nhất (tiệm cận mức tuyệt đối 1.0), đảm bảo tính bền vững (robustness) cho hệ thống phân loại ngay cả khi một số thành phần con hoạt động không như kỳ vọng.

4.5.1. Trực quan kết quả phân cụm



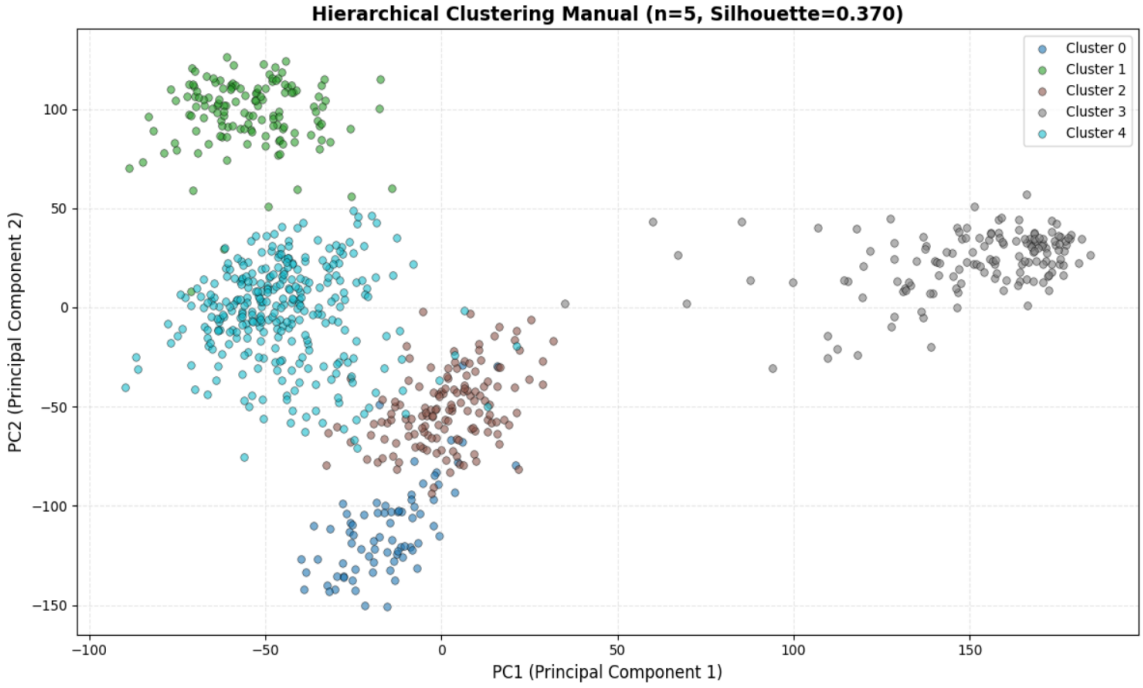
Hình 8. Kết quả phân cụm trên không gian PCA 2D sử dụng thuật toán K-Means++

Biểu đồ cho thấy K-Means++ phân hoạch dữ liệu thành các cụm có độ tập trung cao và xu hướng hình cầu khá rõ rệt. Tuy nhiên, ranh giới giữa một số cụm vẫn còn sự tiếp giáp sát sao, phản ánh hạn chế của thuật toán khi cố gắng áp đặt cấu trúc hình học lồi lên dữ liệu sinh học phức tạp. Mặc dù vậy, kết quả này vẫn cung cấp một cái nhìn tổng quan tốt về sự phân tách của các nhóm bệnh chính.



Hình 9. Kết quả phân cụm và phát hiện nhiễu trên không gian PCA 2D sử dụng thuật toán DBSCAN.

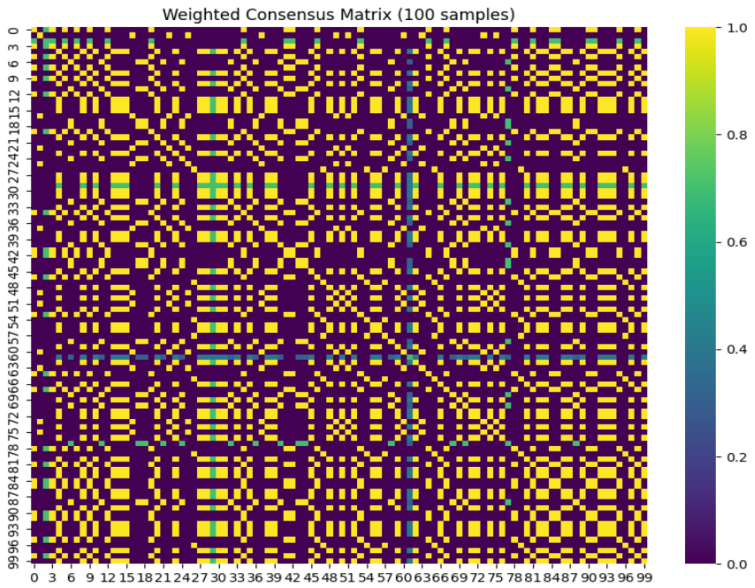
Quan sát biểu đồ, ta thấy sự xuất hiện của các điểm màu xám đại diện cho nhiễu (noise) nằm rải rác ở các vùng thưa thớt hoặc vùng giao thoa giữa các cụm. Điều này khẳng định khả năng vượt trội của DBSCAN trong việc làm sạch dữ liệu và nhận diện các cấu trúc cụm tự nhiên mà không bị ép buộc gán nhãn cho các điểm bất thường. Các cụm chính (Core Clusters) được định hình rõ ràng, tách biệt khỏi các yếu tố gây nhiễu.



Hình 10. Kết quả phân cụm trên không gian PCA 2D sử dụng thuật toán Hierarchical Clustering.

Kết quả phân cụm từ Hierarchical Clustering thể hiện sự tương đồng đáng kể với K-Means++ nhưng có sự tinh tế hơn trong việc xử lý các điểm dữ liệu ở vùng biên. Cấu trúc phân cấp giúp các nhóm dữ liệu được gom tụ dựa trên sự gần gũi về không gian đặc trưng, tạo ra sự phân tách mạch lạc giữa các phân nhóm ung thư, đồng thời duy trì được tính liên kết nội tại của từng nhóm.

4.6. Phân tích cơ chế Ensemble



Hình 11. Trực quan hóa Ma trận đồng thuận của mô hình Ensemble.

Biểu đồ nhiệt hiển thị các khối cấu trúc (block structures) màu vàng sáng nổi bật trên nền tối, minh chứng cho mức độ đồng thuận cao giữa ba mô hình cơ sở. Các ô vuông rực rỡ dọc theo đường chéo chính cho thấy khi một cặp mẫu được xếp chung nhóm bởi một mô hình, xác suất rất cao chúng cũng được đồng thuận bởi các mô hình còn lại. Đặc biệt, nhờ cơ chế "Noise-Aware", các tín hiệu nhiễu đã bị triệt tiêu (vùng nền tối), giúp ma trận đồng thuận trở nên sắc nét, tạo tiền đề vững chắc cho thuật toán phân cụm cuối cùng đạt độ chính xác và ổn định cao nhất.

4.7. Thảo luận

Các thuật toán như K-Means++ hay Hierarchical luôn ép mọi điểm dữ liệu vào một cụm nào đó, kể cả những điểm nhiễu. Điều này dễ làm sai lệch tâm cụm. Ngược lại, DBSCAN hoạt động dựa trên mật độ nên nó sẽ gán nhãn -1 cho các điểm nằm ở vùng thưa thớt thay vì gom bừa. Trong mô hình Ensemble của nhóm, các điểm mang nhãn -1 này sẽ bị loại bỏ khỏi quá trình tính toán ma trận đồng thuận. Nhờ đó, DBSCAN đóng vai trò như một bộ lọc, giúp loại bỏ các dữ liệu sai lệch để kết quả cuối cùng sạch và chính xác hơn.

Thực nghiệm cho thấy chất lượng các mô hình không đồng đều: Hierarchical đạt ARI 0.99 trong khi DBSCAN đạt 0.95. Nếu dùng CSPA thường (trọng số ngang nhau), mô hình kém sẽ kéo tụt kết quả chung xuống mức trung bình. Weighted CSPA giải quyết vấn đề này bằng cách gán trọng số cao cho mô hình tốt (Hierarchical) và giảm ảnh hưởng của mô hình yếu (DBSCAN). Cơ chế này giúp Ensemble đúng đắn hơn, đảm bảo kết quả cuối cùng luôn tiệm cận với mô hình tốt nhất.

4.8. Phân tích cắt bỏ

Để làm rõ vai trò của kỹ thuật SCENA (tăng cường ái lực mạng lưới KNN) trong kiến trúc tổng thể, nhóm đã thực hiện cắt bỏ (Ablation Test). Nhóm đã chạy mô hình chỉ với lõi Weighted CSPA (bỏ qua bước nhân ma trận KNN) và so sánh với mô hình đầy đủ Weighted SCENA.

Kết quả thực nghiệm được ghi nhận như sau:

- Weighted CSPA (Chưa có SCENA)**:** ARI = 0.9907
- Weighted SCENA (Full Pipeline)**:** ARI = 0.9907

Việc áp dụng thêm bước SCENA không làm thay đổi các chỉ số định lượng trên bộ dữ liệu này. Nhóm phân tích hiện tượng này dựa trên hai lý do chính:

Thứ nhất là sự bão hòa về hiệu suất. Cơ chế gán trọng số đã hoạt động quá hiệu quả, giúp mô hình tận dụng chức năng của thuật toán Hierarchical Clustering (vốn đã đạt ARI ~0.99). Khi ma trận đồng thuận tạo ra từ bước CSPA đã tiệm cận mức hoàn hảo và trùng khớp với cấu trúc tự nhiên của dữ liệu, mạng lưới láng giềng (KNN) sẽ không tìm thấy các liên kết sai lệch nào để cải thiện thêm.

Thứ hai, kết quả này đóng vai trò xác thực tính nhất quán. Việc chỉ số không bị sụt giảm khi tích hợp thêm thông tin láng giềng từ dữ liệu gốc chứng tỏ rằng kết quả phân cụm tìm được hoàn toàn phù hợp với cấu trúc hình học của dữ liệu.

Với bộ dữ liệu này, Weighted CSPA đã đủ để đạt hiệu suất tối đa. Tuy nhiên, nhóm vẫn đề xuất giữ nguyên module SCENA trong kiến trúc tổng quát. Lý do là trong các bài toán thực tế khác hoặc với các bộ dữ liệu có độ nhiễu cao hơn (nơi bước đồng thuận ban đầu chưa tối ưu), lớp bảo vệ từ mạng lưới KNN sẽ là yếu tố dự phòng quan trọng để đảm bảo tính ổn định cho hệ thống.

5. KẾT LUẬN

Qua đồ án, nhóm đã xây dựng thành công quy trình phân cụm tổ hợp lai ghép (Hybrid Ensemble Clustering) dành cho dữ liệu biểu hiện gen ung thư. Mô hình được thiết kế dựa trên sự kết hợp giữa thuật toán Weighted CSPA (phân hoạch dựa trên độ tương đồng có trọng số) và kỹ thuật SCENA (tăng cường ái lực mạng lưới). Kết quả thực nghiệm cho thấy hệ thống hoạt động ổn định và đạt độ chính xác cao với chỉ số ARI lên tới 0.9907, hoàn thành tốt mục tiêu phân tách 5 nhóm bệnh lý từ dữ liệu RNA-Seq đa chiều.

Điểm mạnh lớn nhất của mô hình đề xuất nằm ở khả năng "tự vệ" trước các mô hình thành phần. Dù mô hình DBSCAN có hiệu năng thấp hơn (ARI ~0.95) so với các mô hình khác, cơ chế trọng số đã tự động giảm thiểu tác động tiêu cực của nó, giúp kết quả cuối cùng không bị kéo tụt xuống mà vẫn duy trì ở mức ngang bằng với mô hình tốt nhất. Việc tích hợp mạng lưới láng giềng (KNN) từ dữ liệu gốc giúp khôi phục lại các đặc trưng hình học cục bộ mà phương pháp Ensemble truyền thống thường bỏ sót.

Phương pháp tiếp cận này vẫn tồn tại nhược điểm về chi phí tài nguyên. Do phải vận hành song song ba mô hình cơ sở (K-Means++, Hierarchical, DBSCAN) và thực hiện các phép tính trên ma trận tương đồng kích thước N×N, độ phức tạp tính toán và thời gian thực thi của mô hình Ensemble cao hơn đáng kể so với việc chạy một thuật toán đơn lẻ. Điều này có thể trở thành trở ngại khi áp dụng trên các bộ dữ liệu có số lượng mẫu quá lớn.

Trong thời gian tới, nhóm đề xuất hướng mở rộng nghiên cứu sang các bộ dữ liệu giải trình tự gen đơn bào (Single-cell RNA-seq) với kích thước mẫu lớn hơn để kiểm tra khả năng mở rộng (scalability) của thuật toán. Ngoài ra, để cải thiện bước tiền xử lý, nhóm dự kiến thay thế phương pháp PCA truyền thống bằng các kỹ thuật Deep Learning như Autoencoder. Việc này sẽ trích xuất được các đặc trưng phi tuyến tính phức tạp của dữ liệu gen, từ đó nâng cao hơn nữa chất lượng đầu vào cho hệ thống phân cụm.

TÀI LIỆU THAM KHẢO

[1] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," Nature genetics, vol. 22, no. 3, pp. 281-285, 1999.

[2] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," IEEE Transactions on knowledge and data engineering, vol. 16, no. 11, pp. 1370-1386, 2004.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proceedings of the National Academy of Sciences, vol. 95, no. 25, pp. 14863-14868, 1998.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in KDD, vol. 96, pp. 226-231, 1996.

[5] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis," Bioinformatics, vol. 22, no. 19, pp. 2405-2412, 2006.

[6] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," Journal of machine learning research, vol. 3, pp. 583-617, 2002.

[7] S. Zhang, X. Li, Q. Lin, and Y. Lin, "Consensus clustering of single-cell RNA-seq data by enhancing network affinity," Briefings in Bioinformatics, vol. 22, no. 6, p. bbab236, 2021.

[8] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering with hybrid weighting," IEEE Transactions on Cybernetics, vol. 48, no. 6, pp. 1833-1844, 2018.

[9] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics, vol. 20, pp. 53-65, 1987.

[10] L. Hubert and P. Arabie, "Comparing partitions," Journal of classification, vol. 2, no. 1, pp. 193-218, 1985.

[11] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2019.