A Comparative Analysis of clustering Gene Expression Data for Cancer Subtypes

Project Lead: Leila Michal (Andrew ID: lmichal)

Project Idea

Gene expression data offers insights into biological processes and diseases, but analyzing high-dimensional datasets to uncover meaningful patterns can be challenging. This project explores clustering algorithms—K-means, hierarchical clustering, and Leiden clustering—to identify gene or sample groups with similar expression profiles. The goal is to evaluate these algorithms in detecting biologically relevant patterns using external validation metrics like the Jaccard Index and Adjusted Rand Index (ARI). Graph-based Leiden clustering will be emphasized to uncover complex patterns, and internal validation metrics like stabilization and coherence will assess result consistency and quality. The project will use the Gene Expression Cancer RNA-Seq and Arcene datasets for evaluation.

Software Implementation

Data Preprocessing

Gene expression data will be preprocessed through normalization, scaling, and transformation. Dimensionality reduction methods like PCA, UMAP, and t-SNE will simplify the data while preserving structure, aiding in both clustering and visualization.

Clustering Algorithms

Three algorithms will be applied: K-means, hierarchical clustering, and Leiden clustering. K-means partitions data into fixed clusters, hierarchical clustering reveals relationships between clusters, and Leiden clustering uses modularity optimization for complex pattern detection. Distance measures, such as Euclidean and Pearson correlation, will evaluate similarity between genes or samples.

Internal Validation Metrics

Stabilization will assess the consistency of clustering results across multiple runs, while coherence will measure internal similarity within clusters. High stabilization and coherence suggest reliable and meaningful clusters.

Evaluation Metrics

The algorithms will be evaluated using the Jaccard Index and Adjusted Rand Index (ARI), which measure the similarity between predicted and true clusters, providing a basis for algorithm comparison.

Visualization

Dendrograms, cluster centers, and network graphs will visualize hierarchical, K-means, and Leiden clustering results. t-SNE and UMAP will be used to visualize high-dimensional data in 2D or 3D for clearer interpretation.

Resources:

Eisen, M. B., et al. "Cluster analysis and display of genome-wide expression patterns." *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, 1998, pp. 14863–14868.

MacQueen, J. "Some Methods for Classification and Analysis of Multivariate Observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281-297.

van der Maaten, L., & Hinton, G. "Visualizing Data using t-SNE." *Journal of Machine Learning Research*, vol. 9, 2008, pp. 2579-2605.