The statistic explanation of the human prefrontal cortex development data

Yanming Lai, Bokai Yan, Ningyu Yan
Department of Mathematics
The Hong Kong University of Science and Technology
`{ylaiam,byanac,nyanac}@connect.ust.hk`

Abstract

The prefrontal cortex (PFC) in mammals is a complex network of specialized brain areas containing billions of cells. By identifying cell types and distinguishing their developmental features in the developing human prefrontal cortex, we can deepen our understanding of brain development, uncover the mechanisms of cognitive function, identify factors related to neurodevelopmental disorders, and provide guidance for stem cell research and treatment. This is of great significance for improving human health and treating neurological disorders. However, effectively handling high-dimensional cell data remains a major challenge in practical applications. In this study, we simulated a task of prefrontal cortex cell classification. For a human prefrontal cortex cell RNA dataset with a size of 24153×2394, collected from gestational weeks 8 to 26, we ultimately classified the cells into 6 classes. The specific approach involved using different dimensionality reduction methods (LLE, MLLE, MDS, t-SNE, Isomap, spectral embedding) to transform the high-dimensional data into low-dimensional data, and then using different clustering methods (K-means, spectral clustering, hierarchical clustering, Ward hierarchical clustering, BIRCH clustering) to partition the data into 6 clusters. In the experiment, we compared the effectiveness of different methods. We found that when dealing with sparse data, t-SNE (t-Distributed Stochastic Neighbor Embedding) and MDS (multidimensional scaling) performed much better than LLE (Locally Linear Embedding), which is consistent with theoretical results. Additionally, we observed that only when selecting appropriate dimensionality reduction methods to process the data, can clustering methods exert their maximum effectiveness.

1 Introduction

The prefrontal cortex in mammals is a complex network of specialized brain areas containing billions of cells. This region is responsible for the highest-order cognitive functions, such as memory, decision-making, cognitive ability, and social behavior [1]. During human embryonic development and early infancy, functional cells are generated and migrate to the appropriate locations to form neural circuits in the prefrontal cortex. Dysfunction of the prefrontal cortex can lead to cognitive deficits, as well as a range of neurodevelopmental disorders [2]. Therefore, there is a pressing need for detailed knowledge of the prefrontal cortex's development. By identifying cell types and distinguishing their developmental features in the developing human prefrontal cortex through biological dataset, we can deepen our understanding of brain development, uncover the mechanisms of cognitive function, identify factors related to neurodevelopmental disorders, and provide guidance for stem cell research and treatment. This is of great significance for improving human health and treating neurological disorders. However, identifying cell types and distinguishing their developmental features in practical applications is a high-dimensional problem[3].

The utilization of representation learning has yielded notable achievements across diverse domains such as computer vision, speech recognition, and gene expression analysis, with the latter being the primary focus of this report. Our central objective entails the representation of intricate gene expression data, characterized by high-dimensionality, into a low-dimensional latent space. This process effectively mitigates the analytical complexities associated with the problem. Furthermore, by operating within the reduced dimensionality of the latent space, it is anticipated that patterns or correlations between data points, which may be arduous or unfeasible to discern in the high-dimensional space, can be readily visualized and effectively leveraged [4].

The present report focuses on the analysis of a genetic (RNA) dataset from [5] concerning the developing human prefrontal cortex during gestational weeks 8 to 26. Each data is in the unit of TPM. TPM measures gene expression levels in RNA-sequencing experiments [6]. It adjusts for sequencing depth and gene length, allowing for accurate comparisons. The value represents transcripts per million mapped reads for a gene, giving a more accurate estimate of expression than other methods. TPM identifies differentially expressed genes and provides insights into biological processes and diseases.

In scenarios where dimensionality reduction becomes a necessity, it is postulated that the elimination of local directions of variation least represented in the training data should be prioritized [7]. Among the prevailing techniques for dimensionality reduction, Principal Component Analysis (PCA) and its robust outlier-resistant variants are frequently employed. Additionally, manifold learning methods are extensively utilized to tackle this challenge. Locally Linear Embeddings (LLE) focuses on preserving the local relationships among data points while reducing dimensionality. Isomap Embedding employs geodesic distances to capture the global structure of the data manifold. Multidimensional Scaling (MDS) seeks to preserve the pairwise distances between data points in the reduced space. Spectral embedding techniques are employed for non-linear dimensionality reduction, utilizing spectral properties to uncover latent patterns. T-distributed Stochastic Neighbor Embedding (t-SNE) is designed to preserve both local and global structure by modeling pairwise similarities. Lastly, Uniform Manifold Approximation and Projection (UMAP) aims to capture the topological structure of the data manifold while ensuring efficient computation. These methods collectively constitute a comprehensive toolkit for dimensionality reduction, catering to various data characteristics and analytical objectives.

Clustering is the grouping of similar data points based on their characteristics. In dimensionality reduction, clustering helps identify intrinsic patterns, enhances interpretability, validates reduction techniques, and improves preprocessing for other tasks [4].

The report is organized as follows. Section 2 provides an overview of the data processing steps. Section 3 elaborates on the chosen methods and the rationale behind their selection. Section 4 introduces clustering and highlights the relationship between dimensionality reduction and clustering. Finally, Section 5 presents the practical application of the methods on real data, accompanied by a comprehensive analysis of the results.

2 Data Preprocessing

The original dataset 'GSE104276_all_pfc_2394_UMI_TPM_NOERCC.xlsx' contains RNA information of human prefrontal cortex cells, with a matrix size of 24153×2394. The dataset includes data from donator 1 and donator 2. Each row represents genetic expression, and each column represents different cells from gestational weeks 8 to 26. Initially, we excluded the unidentified columns and rows since their content was unknown. Next, we removed cells (columns) with genetic expression below 1000, and genetic data (rows) with less than 3 cell expressions. This is a common practice in scRNAseq experiments to disregard such data.

After data preprocessing, we get a data matrix with the size of 19712 × 2345. The rows represent the different genes, the columns represent the different cells. Here are ten genes with the highest expression frequency (the number of cells with RNA value > 1).

Table 1: Ten genes with the highest expression frequency

genes	frequency	genes	frequency
MALAT1	2391	TUBA1A	2383
FTH1	2389	FTL	2381
TMSB4X	2388	STMN1	2381
EEF1A1	2386	RPLP1	2380
TMSB10	2385	ACTG1	2379

3 Methodology of data reduction

As we know, the underlying assumption of representation learning is the concentration of data around a low dimensional manifold in high dimensional spaces, frequently of nonlinear nature. This motivates the application of manifold learning methods to visualize and reduce the data dimension. The following are the manifold learning techniques utilized in this study:

- Locally linear embedding (LLE) [4, 8] is an approach aimed at reducing data dimensionality while maintaining the distances between data points within their local neighborhoods. It can be conceptualized as a collection of localized principal component analyses, which are subsequently compared globally to identify the optimal non-linear embedding. By preserving the local relationships and structure of the data, LLE facilitates effective visualization and analysis in a lower-dimensional space.

- Modified Locally Linear Embedding (MLLE) [4, 9] is a nonlinear dimensionality reduction technique that extends Locally Linear Embedding (LLE) by using a nonlinear function to reconstruct each data point and introducing regularization and scaling parameters to produce smoother and more robust embeddings. MLLE is effective in various applications but can be computationally expensive and requires parameter tuning.

- The Isomap algorithm [4, 10], derived from Isometric Mapping, represents one of the initial methodologies for manifold learning. Isomap can be perceived as an extension of Multi-dimensional Scaling (MDS) or Kernel PCA techniques. Its primary objective is to obtain a lower-dimensional embedding that accurately preserves the geodesic distances between all data points. By leveraging geodesic distances, Isomap captures the intrinsic structure of the data, enabling meaningful analysis and visualization in a reduced-dimensional space.

- Multidimensional scaling (MDS)[4, 11] is a technique used to find a low-dimensional representation of data where the distances between points accurately reflect the distances in the original high-dimensional space. It is commonly used for analyzing similarity or dissimilarity data by modeling them as distances in a geometric space. MDS allows for the visualization and interpretation of underlying patterns and relationships in the data.

- Spectral Embedding [4, 12] is a method for computing a non-linear embedding by leveraging the spectral decomposition of the graph Laplacian. It aims to capture the local relationships within the data by ensuring that points in close proximity on the underlying manifold are mapped to neighboring positions in the lower-dimensional space. By preserving local distances, Spectral Embedding effectively represents the data in a reduced-dimensional form while retaining its intrinsic structure.

- T-Distributed Stochastic Neighbor Embedding (t-SNE) [13] transforms the affinities between data points into probabilities. It assigns Gaussian joint probabilities to the affinities in the original space and represents the affinities in the embedded space using Student's t-distributions. While Isomap, LLE, and similar methods are better suited for unfolding a single continuous low-dimensional manifold, t-SNE focuses on the local structure of the data. It tends to identify and extract clustered local groups of samples, as exemplified by the S-curve example. This characteristic of t-SNE, grouping samples based on the local structure, can be advantageous in visually disentangling datasets that consist of multiple manifolds simultaneously.

4 Methodology of clustering

The connection between dimensionality reduction and clustering remains substantial, with dimensionality reduction often being employed as a valuable technique in the context of clustering. Dimensionality reduction serves two primary objectives:

Firstly, it aims to enhance model adaptability, particularly in datasets with limited samples where high-dimensional spaces and sparse sample distributions can lead to overfitting. In such cases, two approaches are commonly employed. The first involves utilizing regularization techniques, such as LASSO, to restrict model complexity. The second approach revolves around reducing dimensionality to alleviate the sparsity issue and modify the distribution of spatial dimensions. By reducing dimensionality, the adverse effects of sparse sample distributions can be mitigated, thereby facilitating improved adaptability of the models.

Secondly, dimensionality reduction addresses the challenge posed by the "curse of dimensionality." In high-dimensional spaces, conventional distance measures, like Euclidean distance, can result in a phenomenon known as the "dimensionality disaster." This occurs when most samples are compressed within a narrow range, rendering them indistinguishable. To overcome this challenge, two primary approaches are commonly employed. The first involves optimizing the distance measurement method and incorporating compression techniques during the calculation process. The second approach focuses on dimensionality reduction itself, effectively reducing the excessive dimensions and mitigating the dimensionality disaster by sacrificing a portion of the information.

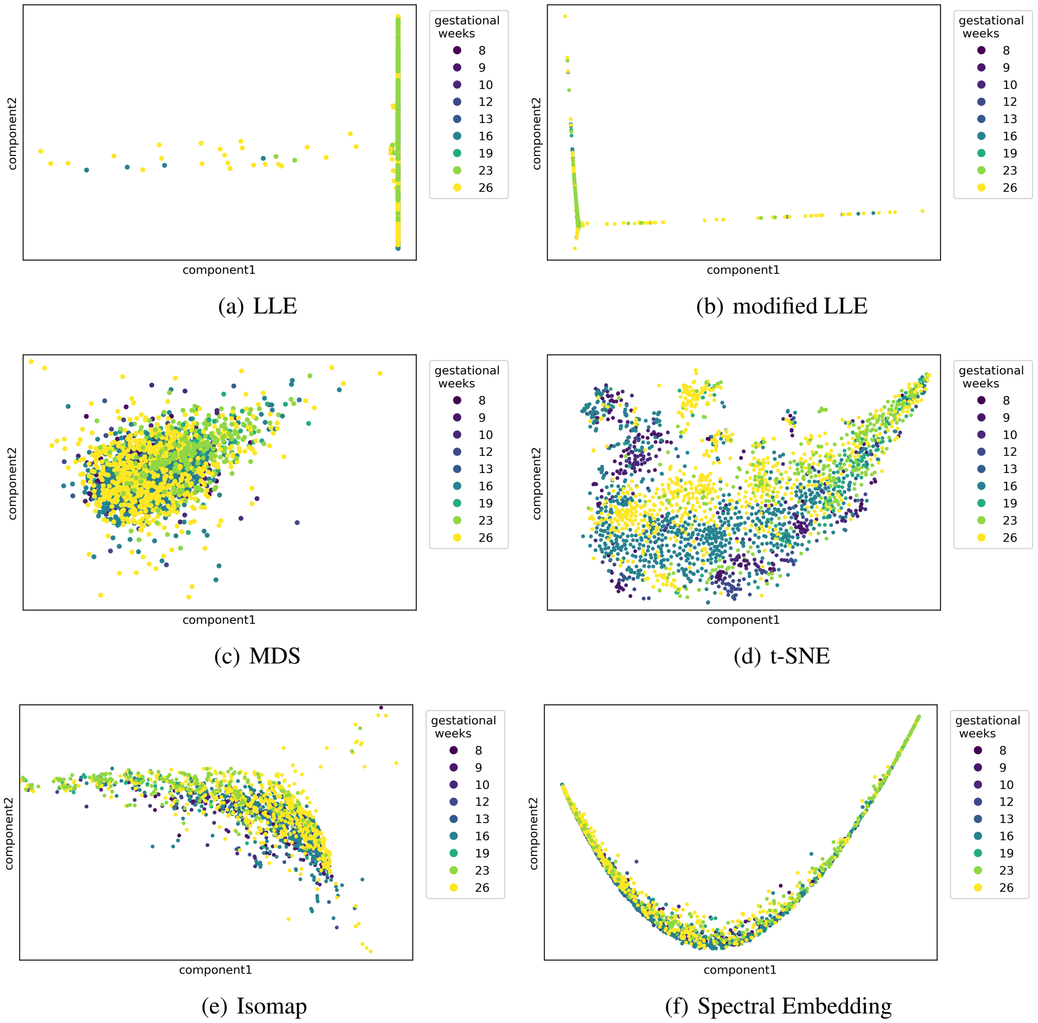
Clustering methods can be categorized into two main types: traditional methods, such as k-means, hierarchical clustering, and DBSCAN, which require explicit distance definitions, and graph-based clustering, which relies on the notion of connected relationships. Traditional clustering methods heavily rely on distance definitions, which can pose challenges when dealing with high-dimensional data. The generalizability of modified distance measurement algorithms is often limited and context-dependent. Therefore, dimensionality reduction methods are frequently employed as a means to address these challenges. The following are the clustering methods utilized in this study:

- K-means is a popular clustering algorithm in machine learning and data mining. It is used to partition a dataset into distinct groups or clusters based on their similarity. The algorithm aims to minimize the sum of squared distances between the data points and their assigned cluster centroids.
- Spectral clustering is a clustering algorithm that utilizes the spectral properties of the data to perform clustering. It is often used when the data has a non-linear or complex structure that cannot be effectively captured by traditional clustering algorithms like K-means. The main idea behind spectral clustering is to transform the data into a lower-dimensional space using the eigenvectors (spectral decomposition) of a similarity matrix derived from the data.
- Hierarchical clustering is a clustering algorithm used in data analysis and machine learning to group similar data points into nested clusters. It is a bottom-up (agglomerative) or top-down (divisive) approach that organizes data into a hierarchical structure, often represented as a dendrogram.
- Ward hierarchical clustering is an agglomerative algorithm used to create hierarchical clusters by minimizing within-cluster variance. It starts with each data point as a separate cluster and iteratively merges clusters that result in the smallest increase in within-cluster variance. The method prioritizes creating compact and balanced clusters, which can be advantageous in certain applications. However, Ward's method may be sensitive to outliers and non-convex cluster shapes.
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm designed to handle large-scale datasets efficiently. It constructs a tree-like structure, known as a CF Tree (Clustering Feature Tree), to cluster and summarize the data. The CF Tree allows for compact representations of clusters using Clustering Features (CFs), which include information such as the number of data points, centroid, and squared deviation.

5 Visualization and Discussion

We first use manifold learning to reduce the dimensionality of the data, and then apply clustering methods to divide the cells into several clusters.

Figure 1 shows the data reduction results of LLE, modified LLE, MDS, t-SNE, Isomap, Spectral Embedding respectively. From Figure 1, we find that due to the high sparsity of the data matrix, LLE and modified LLE fail to efficiently reduce the dimensionality, which is consistent with the theoretical result. In fact, LLE struggles with sparse data as it needs local neighborhoods around each data point, which may not exist in sparse data due to large distances and isolated points. Sparse data often has many zero values across dimensions, which makes estimating local linear relationships difficult due to a lack of information. In contrast, t-SNE and MDS can produce satisfactory results when dealing with sparse data, which is also consistent with theoretical results. t-SNE is effective in dealing with sparse data because it uses a probabilistic approach that can preserve the local structure and similarity of the data while effectively handling sparsity issues. It can map the high-dimensional information of sparse data to low-dimensional space while maintaining the relative distances between samples, making data visualization more intuitive.



(a) LLE

(b) modified LLE

(c) MDS

(d) t-SNE

(e) Isomap

(f) Spectral Embedding

Figure 1: Visulization of several reduction methods

The silhouette plot is a visualization tool used to assess the performance of clustering algorithms. It displays the clustering results for each sample in a dataset and calculates a metric called the silhouette coefficient, which measures the similarity of a sample to its own cluster compared to other clusters. In the silhouette plot, the width of each bar represents the silhouette coefficient of the corresponding data point, and the color represents the cluster to which it belongs. The bars for each cluster are normalized to the same height, and the clusters are arranged in the order of their labels. The average silhouette coefficient is indicated by a red dashed line. The range of the silhouette coefficient is [-1, 1], where positive values indicate that the samples are correctly assigned to appropriate clusters, negative values indicate that the samples may have been mistakenly assigned to neighboring clusters, and values close to 0 indicate that the samples are near the boundaries of two clusters. By observing the silhouette plot, one can evaluate the performance of the clustering algorithm and the quality of the clusters. A good clustering result is characterized by wide bars with high positive values, indicating high similarity among samples within the same cluster and significant differences from samples in neighboring clusters. Conversely, a poor clustering result may exhibit narrow bars or silhouette coefficients close to 0 or negative, indicating low similarity among samples or incorrect assignment to inappropriate clusters. Therefore, the silhouette plot can help in selecting an appropriate number of clusters and evaluating the performance of clustering algorithms.

In the experiment we tested the number of clusters ranging from 2 to 10 and observed from the silhouette plot that the clustering result appeared to be more reasonable when the number of clusters was set to 6. The silhouette plot for the number of clusters being 6 is shown in Figure 2. The average silhouette coefficient is 0.4.

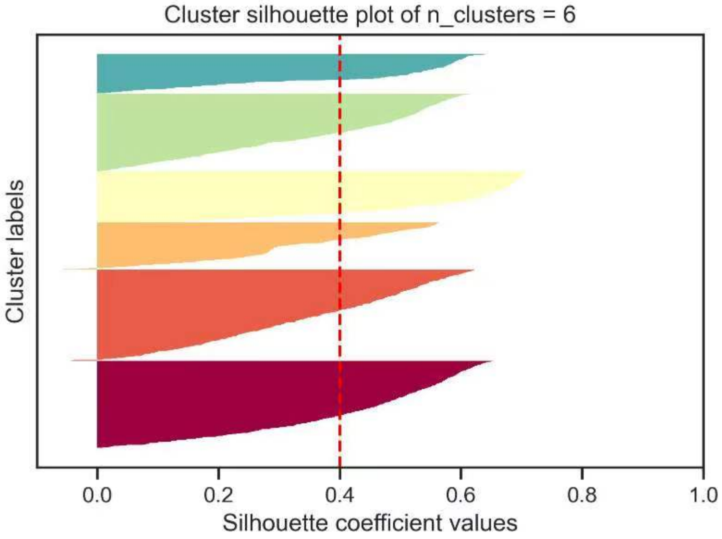
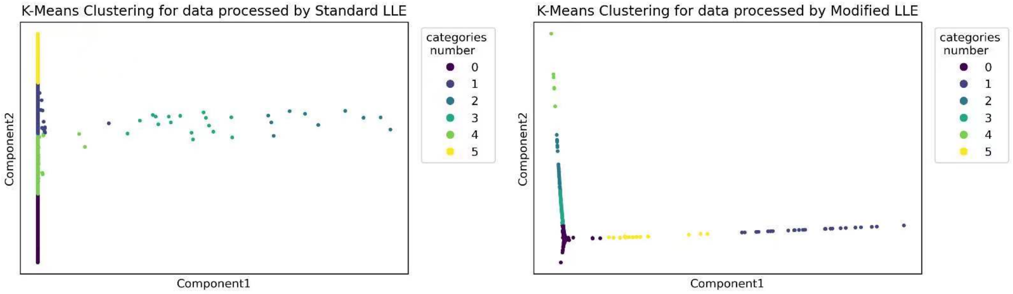


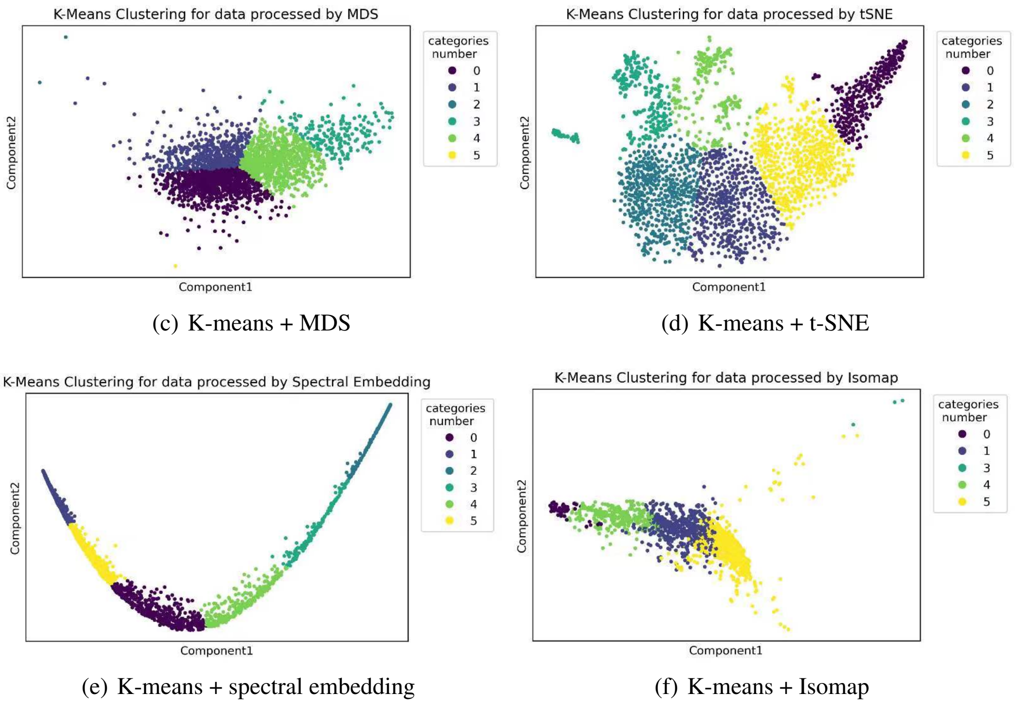
Figure 2: Cluster silhouette plot

Figure 3 displays the visualization outcome of applying K-means to various dimension reduction methods. Similar to Figure 1, when t-SNE and MDS are employed for reducing dimensionality, the results of K-means are satisfactory. However, the outcomes are unsatisfactory when LLE and modified LLE are utilized. This observation suggests that K-means can maximize its impact only when a suitable dimensionality reduction method is chosen.



(a) K-means + LLE

(b) K-means + MLLE



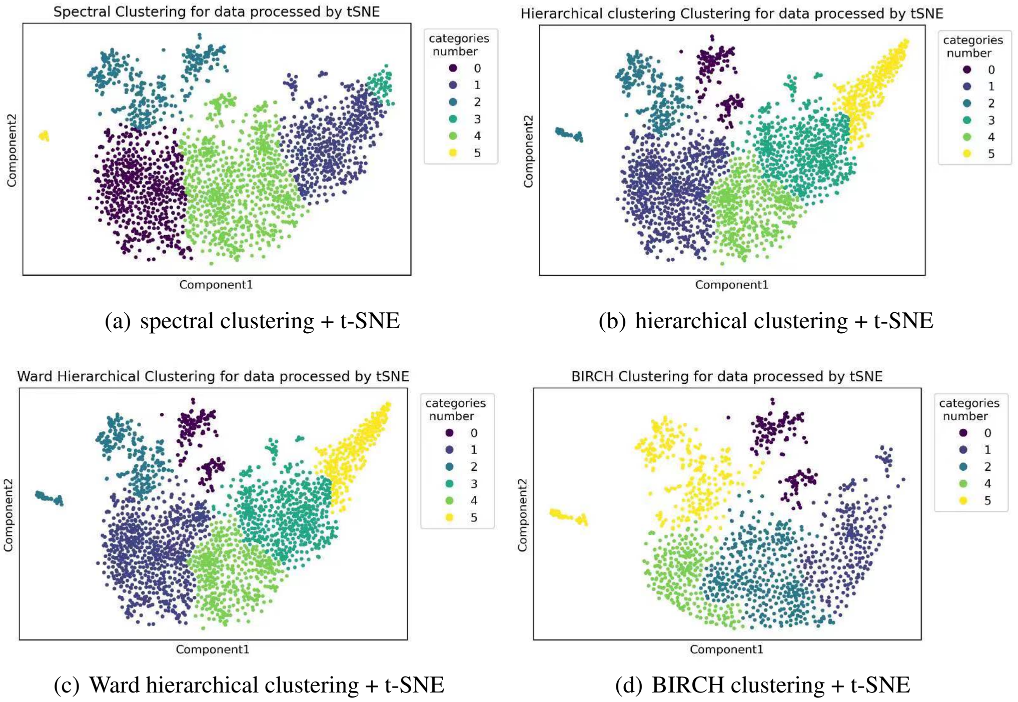
(c) K-means + MDS

(d) K-means + t-SNE

(e) K-means + spectral embedding

(f) K-means + Isomap

Figure 3: Visualization of K-means



(a) spectral clustering + t-SNE

(b) hierarchical clustering + t-SNE

(c) Ward hierarchical clustering + t-SNE

(d) BIRCH clustering + t-SNE

Figure 4: Visualization of several clustering methods

Figure 4 presents additional visualization outcomes for various clustering methods. It is evident that all these methods (spectral clustering, hierarchical clustering, Ward hierarchical clustering, and BIRCH clustering) yield satisfactory clustering results when combined with the t-SNE dimensionality reduction method. This is attributed to the fact that t-SNE effectively preserves the local structure and similarity of sparse data. Therefore, the results once again demonstrate that selecting a suitable dimensionality reduction method can ensure smooth subsequent work.

6 Conclusion

In this study, we simulated a task of prefrontal cortex cell classification. For a human prefrontal cortex cell RNA dataset with a size of 24153×2394, collected from gestational weeks 8 to 26, we ultimately classified the cells into 6 classes. The specific approach involved using different dimensionality reduction methods (LLE, MLLE, MDS, t-SNE, Isomap, spectral embedding) to transform the high-dimensional data into low-dimensional data, and then using different clustering methods (K-means, spectral clustering, hierarchical clustering, Ward hierarchical clustering, BIRCH clustering) to partition the data into 6 clusters. In the experiment, we compared the effectiveness of different methods. We found that when dealing with sparse data, t-SNE (t-Distributed Stochastic Neighbor Embedding) and MDS (multidimensional scaling) performed much better than LLE (Locally Linear Embedding), which is consistent with theoretical results. Additionally, we observed that only when selecting appropriate dimensionality reduction methods to process the data, can clustering methods exert their maximum effectiveness.

Contributions of Group Members

YAN Ningyu studied the dataset and conducted the data preprocessing. YAN Ningyu and LAI Yanming wrote the codes and performed the numerical simulations. YAN Bokai studied the background of data reduction methods. LAI Yanming and YAN Bokai wrote the report.

References

[1] Daniel Baldauf and Robert Desimone. Neural mechanisms of object-based attention. *Science*, 344(6182):424–427, 2014.

[2] Adele Diamond, Meredith B Prevor, Glenda Callender, and Donald P Druin. Prefrontal cortex cognitive deficits in children treated early and continuously for pku. *Monographs of the society for research in child development*, pages i–206, 1997.

[3] Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, Xiang Zhou, Xingjie Shi, and Jin Liu. Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic acids research*, 50(12):e72–e72, 2022.

[4] Yuan Yao. A mathematical introduction to data science. 2019.

[5] Suijuan Zhong, Shu Zhang, Xiaoying Fan, Qian Wu, Liying Yan, Ji Dong, Haofeng Zhang, Long Li, Le Sun, Na Pan, et al. A single-cell rna-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*, 555(7697):524–528, 2018.

[6] Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1201, 2020.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[8] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[9] Zhenyue Zhang and Jing Wang. Mlle: Modified locally linear embedding using multiple weights. *Advances in neural information processing systems*, 19, 2006.

[10] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[11] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[12] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.