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

May 18, 2023

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Outline

1. Introduction

2. Data Preprocessing

3. Methodology

4. Clustering

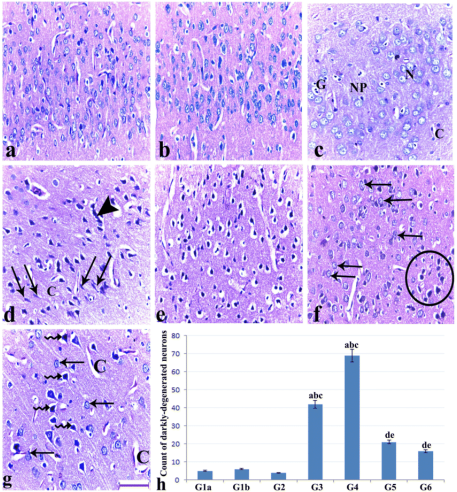
5. Visualization and Discussion

6. Conclusion

Outline

1. Introduction
2. Data Preprocessing
3. Methodology
4. Clustering
5. Visualization and Discussion
6. Conclusion

Introduction



Introduction

The prefrontal cortex in mammals is a complex network of specialized brain areas containing billions of cells. Studying prefrontal cortex in mammals 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

Outline

1. Introduction
2. Data Preprocessing
3. Methodology
4. Clustering
5. Visualization and Discussion
6. Conclusion

Data Preprocessing

The original dataset contains RNA information of human prefrontal cortex cells, with a matrix size of 24153×2394. 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.

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

Outline

1 Introduction

2 Data Preprocessing

3 Methodology

4 Clustering

5 Visualization and Discussion

6 Conclusion

Methodology

In our report, 6 manifold learning techniques are utilized:

- Locally linear embedding (LLE)
- Modified Locally Linear Embedding (MLLE)
- Isomap
- Multidimensional scaling (MDS)
- Spectral Embedding
- T-Distributed Stochastic Neighbor Embedding (t-SNE)

Outline

1 Introduction

2 Data Preprocessing

3 Methodology

4 Clustering

5 Visualization and Discussion

6 Conclusion

Clustering

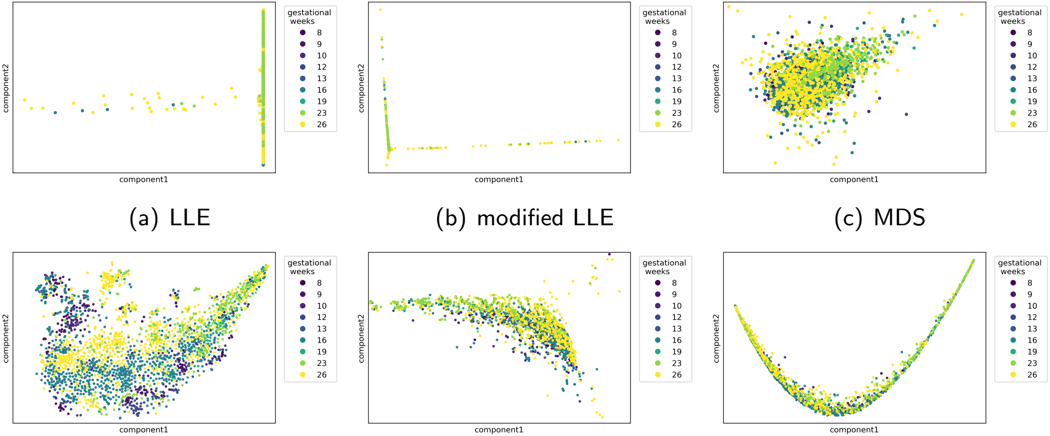
The following are the clustering methods used in this study:

- K-means
- Spectral clustering
- Hierarchical clustering
- Ward hierarchical clustering
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Outline

1. Introduction
2. Data Preprocessing
3. Methodology
4. Clustering
5. Visualization and Discussion
6. Conclusion

Visualization and Discussion



(a) LLE

(b) modified LLE

(c) MDS

(d) t-SNE

(e) Isomap

(f) Spectral Embedding

Figure 1: Visualization of several reduction methods

Visualization and Discussion

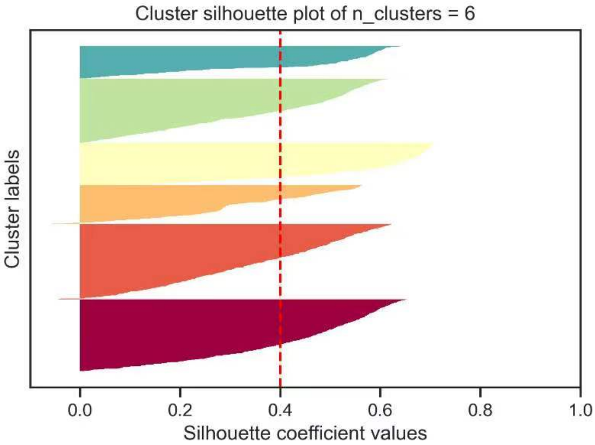
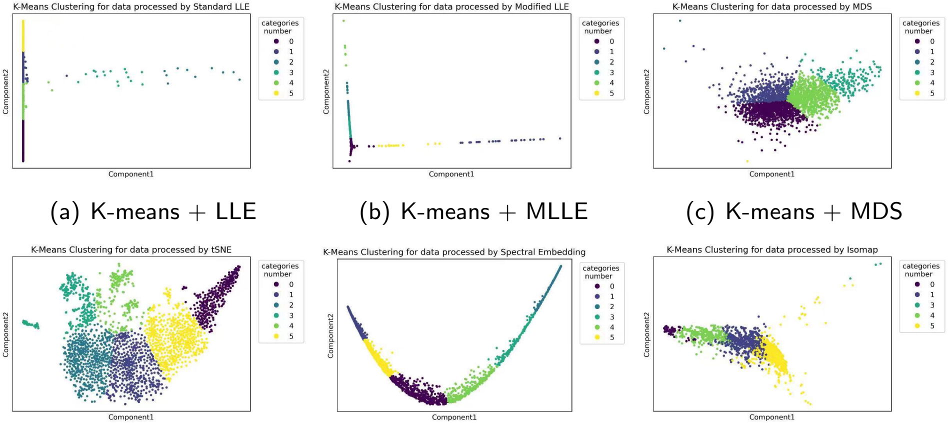


Figure 2: Cluster silhouette plot

Visualization and Discussion



(a) K-means + LLE

(b) K-means + MLLE

(c) K-means + MDS

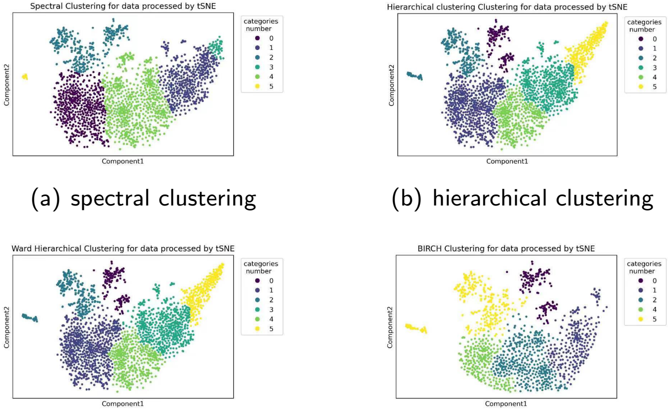
(d) K-means + t-SNE

(e) K-means + SE

(f) K-means + Isomap

Figure 3: Visualization of K-means

Visualization and Discussion



(a) spectral clustering

(b) hierarchical clustering

(c) Ward hierarchical clustering

(d) BIRCH clustering

Figure 4: Visualization of several clustering methods

Outline

1. Introduction
2. Data Preprocessing
3. Methodology
4. Clustering
5. Visualization and Discussion
6. Conclusion

Conclusion

In this study, we simulated a task of prefrontal cortex cell classification.

- We changed the human prefrontal cortex cell RNA dataset with a size of 24153×2394 to 19712 × 2345.
- using different dimensionality reduction methods (LLE, MLLE, MDS, t-SNE, Isomap, spectral embedding) to transform the high-dimensional data into low-dimensional data.
- using different clustering methods (K-means, spectral clustering, hierarchical clustering, Ward hierarchical clustering, BIRCH clustering) to partition the data into 6 clusters.

Thank you!