Report for Clustering of Spatial Transcriptomics Data Using MENDER (Assignment – 2)

Team Number: 2
Course: CS690A
Instructor: Hamim Zafar

Name – Yash Suryavanshi (Roll No. 211196)

---

1. Introduction

In this assignment, we explored clustering algorithms to identify spatial domains from two imaging-based spatial transcriptomics datasets. The datasets were acquired using MERFISH and OSM-FISH techniques, each having unique characteristics for clustering and spatial domain identification.

Our goal was to cluster the datasets and evaluate clustering performance using the Adjusted Rand Index (ARI). The ARI scores were calculated for both datasets (using MENDER) and are as follows:

- **Dataset 1 (MERFISH):** ARI = 0.48041
- **Dataset 2 (osmFISH):** ARI = 0.72529

We applied the following methods to the datasets and generated clustering predictions.

---

2. Dataset Descriptions

Dataset 1: MERFISH Mouse Brain Preoptic Hypothalamus Region

- This dataset contains spatial gene expression data from the preoptic hypothalamus region of the mouse brain using the MERFISH technique.
- The data includes a spot-gene matrix and corresponding spatial (x, y) coordinates.
- **Number of clusters:** 8

Dataset 2: osmFISH Mouse Brain Somatosensory Cortex Region

- This dataset represents the somatosensory cortex region of the mouse brain acquired with OSM-FISH.
- Similar to the MERFISH dataset, it contains a spot-gene matrix and spatial coordinates.
- **Number of clusters:** 11

---

3. Methods Used

I implemented the following steps in Jupyter Notebooks (Team_2_Part1.ipynb and Team_2_Part2.ipynb) to process and cluster both datasets:

1. **Preprocessing:**
    - **Quality Control (QC):** QC measures were applied to filter out poor-quality cells/spots and genes with low expression across the datasets.
    - **Normalization:** Data normalization was performed to account for differences in sequencing depth.
    - **Feature Selection:** Highly variable genes were selected to reduce noise and improve clustering accuracy.
2. **Dimensionality Reduction:**
    - I used **PCA (Principal Component Analysis)** to reduce the dimensionality of the dataset where needed while retaining significant variation for clustering.
3. **Clustering Algorithms:**
    - **Leiden Algorithm:** The Leiden clustering algorithm was used to identify clusters in the data. It is effective for detecting communities in spatial transcriptomics data.
    - **MENDER:** This method was applied for the main clustering task, particularly to account for spatial coherence in the gene expression data.
    - **Spatial Feature Plots:** Spatial feature plots were generated to visualize the clusters in relation to the tissue regions.
    - **Experimentation:** I experimented with different n_scales values (the default 6 and 10) and different nn_mode parameters ('radius' vs 'knn'). We observed that using a scale of 6 and the 'radius' mode produced the most coherent spatial domains, while increasing the scale to 10 made the clusters too small and fragmented.
4. **Post-Clustering:**
    - After clustering, the cluster labels were renumbered from 1 to the total number of clusters for consistency.
    - Saved the results in the specified CSV format.

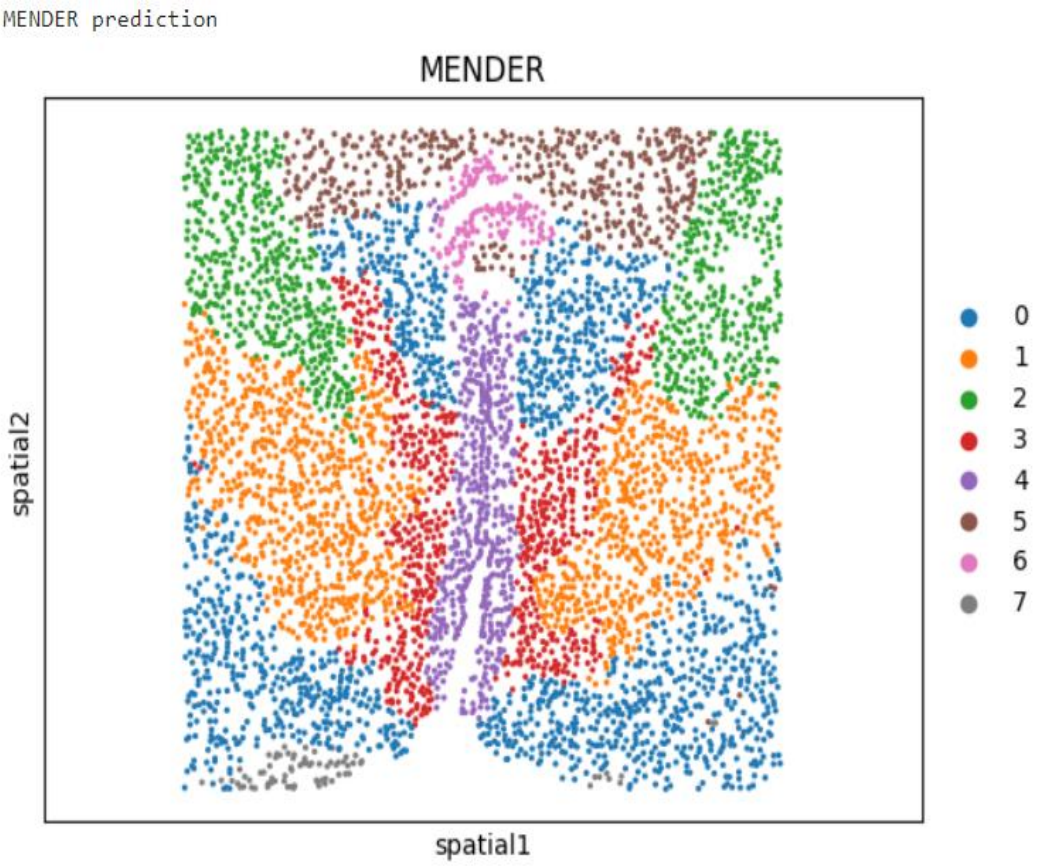
---

4. Results

Dataset 1 (MERFISH)

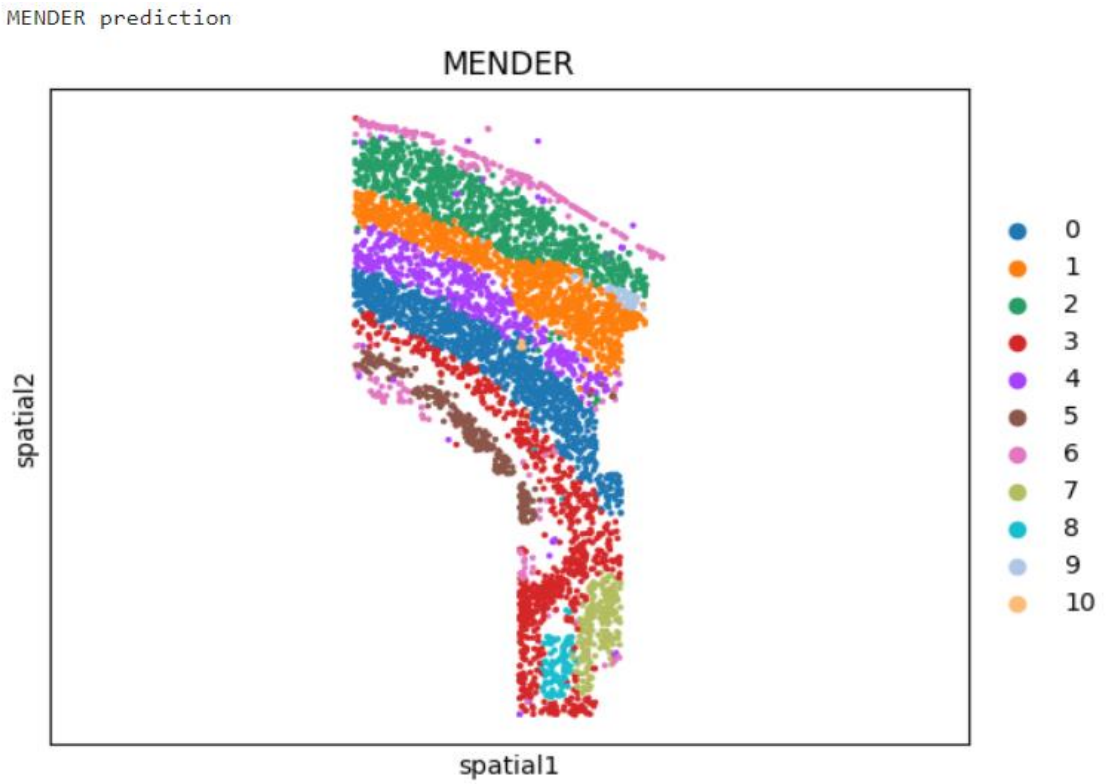
- **Clustering Method:** MENDER

- **Number of Clusters:** 8
- **Adjusted Rand Index (ARI):** 0.48041
- **Spatial Feature Plot:**



Dataset 2 (osmFISH)

- **Clustering Method:** MENDER
- **Number of Clusters:** 11
- **Adjusted Rand Index (ARI):** 0.72529
- **Spatial Feature Plot:**



---

5. Conclusion

The clustering methods applied on the MERFISH and osmFISH datasets successfully identified spatial domains. The ARI scores indicate moderate to high clustering accuracy, with the osmFISH dataset achieving a better ARI score than MERFISH. The spatial feature plots illustrate the spatial coherence of the identified clusters.

The use of Leiden and MENDER algorithms was effective for clustering spatial transcriptomics data, and further exploration and experimentation was done using other clustering algorithms like k-map and Lovain and they were also performing near to the performance of Leiden.