CS690A: Assignment - Clustering of Spatial Transcriptomics Data

Hamim Zafar
hamim@iitk.ac.in

Indian Institute of Technology Kanpur — September 12, 2022

Introduction

Spatial transcriptomics methods (such as Spatial Transcriptomics (ST/Visium), Slide-seq, and High Definition Spatial Trascriptomics) spatially barcode entire transcriptomes with a spatial resolution larger than a single cell, ranging from 50 m to 100 m for ST to 10 m for Slide-seq. Such datasets can be thought of consisting of a spot-gene matrix, where the each spot corresponds to a tissue location from where the mRNA molecules have been captured and the data corresponding to a spot represents a gene expression vector. With such datasets, an important task is to **identify spatial domains** defined as regions that are spatially coherent in both gene expression and histology. Traditional clustering methods such as K-means and Louvain's method may not perform the best in such scenario. To account for spatial dependency of gene expression, new methods have been developed. Some of these methods are listed in here (`https://docs.google.com/spreadsheets/d/1u_aT3l6RfLfSTBSkB4C-BmSV_eaLYTKx83vF9YtESnQ/edit?usp=sharing`).

The goal of this assignment is to evaluate the existing algorithms for clustering spatial transcriptomics datasets and also come up with a clustering algorithm that improves the clustering of spots into spatial domains.

Info: To learn more details on the problem, read the following papers

- Hu, Jian, et al. "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network." Nature methods 18.11 (2021): 1342-1351.
- Fu, H., Xu, H., Chong, K., Li, M., Ang, K. S., Lee, H. K., ... Chen, J. (2021). Unsupervised spatially embedded deep representation of spatial transcriptomics. Biorxiv.

1 Dataset Description

The assignment contains one training and one test datasets as described in the following

1.1 Training and Testing Dataset

For this assignment, you are given spatial transcriptomics data generated with the 10x Genomics Visium platform (`https://www.10xgenomics.com/products/spatial-gene-expression`). For both the training and testing datasets, both h5 and RData files have been provided.

For the training data, following files are present.

- **training_filtered_feature_bc_matrix.h5** - contains the spot-gene matrix
- **training_metadata.csv** - Ground Truth clustering is present here - Cluster information for each spot (each row is a spot Barcode: identifier for each spot)
- **training_tissue_positions_list.csv** - Tissue positions (x & y coordinates for each spot), this may be utilized by some of the clustering methods.

- **training.RData** - an R data file consisting of the gene expression matrix, ground truth clustering and x-y coordinates information.

For the testing data, following files are present.

- **testing_filtered_feature_bc_matrix.h5** - contains the spot-gene matrix
- **testing_tissue_positions_list.csv** - Tissue positions (x & y coordinates for each spot), this may be utilized by some of the clustering methods.
- **testing.RData** - an R data file consisting of the gene expression matrix, ground truth clustering and x-y coordinates information.
- **barcodes_to_avoid.csv** - This file contains the barcodes that need to be removed before clustering. These barcodes (i.e., spots) will not be utilized for evaluating your clustering performance.

All files can be accessed from the kaggle competition page.

1.2 Assignment of methods to teams

For the first task in the assignment, each team is assigned an existing clustering method for spatial transcriptomic data. The method assignment can be found here (`https://docs.google.com/spreadsheets/d/1IgGyD0DbXXFLO8Dy1qcoVsYQ3UK72_4lV4RcE8TSgSE/edit?usp=sharing`).

2 Tasks

2.1 Task 1 (40 points)

The first task is to run the clustering method assigned to your team on the training and testing datasets. For training data, you will be able to compute ARI as a measure of your clustering accuracy. For the testing data, you will submit the solution in the format as given below

Listing 1: Format of output csv file

```
Id,Expected
AAACAAGTATCTCCCA-1,1
AAACAATCTACTAGCA-1,2
AAACACCAATAACTGC-1,3
AAACAGAGCGACTCCT-1,2
AAACAGCTTTCAGAAG-1,1
```

The solution will be evaluated for clustering accuracy. You can number your clusters in the range $[1, K]$ where K is the number of clusters inferred by your method. You also need to generate a spatial feature plot of the clusters and provide it as an output (see the deliverables section).

2.2 Task 2 (60 points)

You need to come up with a clustering algorithm for clustering the spatial transcriptomic data. You can evaluate your clustering algorithm on the training dataset and then submit the predicted clustering on the test data for evaluation. The leader board for the challenge will be maintained based on the performance on the test data. Consider the points below when preparing your solution.

- Spatial transcriptomic datasets are high-dimensional and sparse. A simple clustering like k-means/hierarchical will not work well for such a complex dataset. You need to be creative in designing your clustering algorithm. You can adopt clustering algorithms recently published in conferences such as Neurips and ICML. Specifically, you should consider clustering methods developed for high-dimensional, sparse datasets.
- Dimension reduction is an important step for clustering such datasets. You can experiment with deep learning-based nonlinear dimension reduction techniques for better clustering of the dataset.

- Some clustering methods can have certain parameters which can affect the clustering results. Make sure to experiment with such parameters to improve clustering accuracy.
- Try to be as creative as possible in solving the problem. It involves a research question and if the solutions you submit are better than the state-of-the-art, we will write a research paper in which you will be a co-author.

Notice: In case we require a change in the format of the csv file, we will notify you. Keep an eye on the announcements.

Kaggle Leader board: You can submit the csv files multiple times and check your performance on the test data. In Kaggle we will maintain two competitions for the two tasks.

3 Deliverables

The deliverables for the assignment are the following

1. Clustering prediction on the test data. These results will be evaluated and the leader board will be maintained based on the scores in evaluation. For task 1, you will be graded based on the successful execution and experimentation of the method assigned to you. For task 2, you will be graded based on your position on the leaderboard.
2. Runnable code (in Jupyter Notebook) for the method assigned to you and the clustering algorithm you have developed.
3. Scripts for running your code to generate the predictions on test data. TAs will run these scripts to reproduce the csv files you submit for the challenge
4. A short report describing the steps taken to solve the challenge. Describe in brief the algorithms you have used, any dimension reduction you have performed, training process, training accuracy, etc. The report should also contain the spatial feature plots (based on the obtained clusters) of both the training and testing datasets for both the method assigned to you and the clustering method that you develop for task 2. The writeup should also contain a section describing the contribution of each member in the team. The writeup should mention the names and roll numbers of the team members.

For submission, all the deliverables should be zipped in a single file and the zip file should be named as `Group_i_CS690_ST_clustering_assignment.zip`, `i` should be replaced by your group number. Also, each file in the zip folder should start with the phrase `Group_i_` (`i` replaced by your group number). The file should be emailed to the instructor with the associated TA copied in the email. You can find the allotted TA for your group from the excelsheet (`https://docs.google.com/spreadsheets/d/1IgGyD0DbXXFLO8Dy1qcoVsYQ3UK724lV4RcE8TSgSE/edit?usp=sharing`) that contains the group info. The subject line of the email should mention the group number, [CS690A] and the phrase 'ST Clustering assignment'.

4 Submission Deadline

October 2^{nd} 11:59 PM.

5 Kaggle Competition links

Task1 - `https://www.kaggle.com/t/53371d213159432bb0d21018fd2277fd`
Task2 - `https://www.kaggle.com/t/4fa63ec2cd164c0bbe22273887fddb53`