Analysis of stLearn normalization effectiveness for obtaining better tissue *classifications*, compared to normalization free clustering.

Jeff LaFrence

Lafre020@umn.edu

4735554

Introduction

In, the stLearn paper by *Pham et al.*[1], they show that SMEclust outperforms SpatialLIBD at classifying tissues on human dorsolateral prefrontal cortex (HDPFC) tissue data. This clustering happens post-normalization in the presented Stlearn pipeline, and only data following this workflow is presented. This made me curious if the improvements in clustering demonstrated in the paper were due to superior clustering abilities, or if the additional disk smoothing by the stSMe module was the driver of improved fit. I performed stLearn analysis on the same tissue set with and without normalization to determine if the normalization step is making a significant difference in the clustering results.

Data Preparation

As with the stLearn paper, the data here was drawn from the *Pardo et al* paper.[2] Specifically, data was taken from the Lieber Institute Github database for the project.[1] The ground truth data represents tissue classifications annotated by a clinician, and is considered the result if ideal clustering was achieved.[2]

For each tissue sample, its numbered folder was drawn from the 10x file. The important files from these folders were: scalefactors_json.json , tissue_hires_image.png, tissue_lowres_image.png, and tissue_positions_list. Another necessary file, cluster_labels_XXXXXX.csv was drawn from the outputs/SpatialDE_clustering folder and added to the downloaded folder of its respective sample. These contain the ground truth tissue annotations. The additional file from the Lieber Instiurte, XXXXXX_filtered_feature_bc_matrix.h5 was drawn from the numbered folder in the jhpce#HumnaPilot10X globus database for each sample and added to its respective folder.[3]

Next, I formatted the data to that expected by stLearn. First, the file tissue_positions_list was converted into a .csv file. The sample ID was also removed from eacg samples XXXXXX_filtered_feature_bc_matrix.h5 file so each .h5 file was named filtered_feature_bc_matrix.h5. In each tissue samples numbered folder, an additional folder named spatial was made, and the files scalefactors_json.json , tissue_hires_image.png, tissue_lowres_image.png, and tissue_positions_list.csv were moved into it. An example of one of samples folder can be seen attached to the homework submission.

Methods

To make tissue classifications in the standard workflow, we began by performing a principal component analysis (PCA), comparing gene expression across all scRNAseq data sets for the given tissue sample. The tissue images were then broken down into thousands of tiles and deep learning is used to classify tiles by tile location and contents compared against the other tiles. Using the deep learning tissue classifications, the PCA simplified scRNAseq data was normalized by disk smoothing to the surrounding like tissues. PCA analysis is then re-performed on the normalized data. This normalized PCA data is then run through kmeans clustering to determine final tissue groupings.

The analytical outputs of the program are the adjusted rand score, normalized mutual info score, purity score, PCA homogeneity, PCA completeness, and PCA v measure. The adjusted rand score measures how similar the computer clustering's are to that of the ground truth data, adjusted by the fitting of a random placement grouping as a 0 baseline. 1 is perfectly matching the ground truth data and 0 would be matching equivalent to a random distribution. The normalized mutual info score is a measure of how often labels occur together compared to

how often they occur separately.[4] These are then normalized to a score of 0 to 1.[5] A score of 1 would mean that the computed points always group with the respective ground truth designation, and 0 would indicate they never group with their respective ground truth designation. The purity score is an internal criterion for clustering, it measures how much like samples are grouped together, with random sorting being close to 0 and a clustering that only occurs local to like clustering indicated as a 1.[6] The PCA homogeneity score indicates how much samples belong to a single class, indicated against the ground truth data. Scores close to 0 would indicate nearly random sampling and a 1 would indicate that the computational clustering perfectly overlaps the ground truth data.[7] The PCS completeness score is a check if all data points of a given class are also members of the same cluster.[8] A perfect overlap between computed clustering and ground truth would be equal to one, with random distributions being closer to 0. The PCA V measure is the harmonic mean of the PCA completeness and homogeny score.[9] For all of these indicators 1 indicates the best clustering for its respective measure and scores closer to 0 indicating poorer clustering, thus these can all be plotted on the same axes.

Experiment Design

The normalized data analysis was done using the pipeline outlined by the stLearn Human Brain dorsolateral prefrontal cortex stSME clustering tutorial.[10] StLearn overwrites the non-normalized PCA data (data_.obsm["X_pca"]), when performing the post normalization PCA analysis. Since the output fields are the same, the process of removing the normalization step was as simple as removing the commands st.spatial.SME.SME_normalize() and data_.X = data_.obsm['raw_SME_normalized'] from the workflow. The second round of PCA analysis was also removed, as the normalized field it is running off is missing. This constitutes removing the

commands, st.pp.scale(data_) and st.em.run_pca(data_,n_comps=15). This allows the data to enter kmeans clustering with raw PCA data rather than the normalized PCA data.

There is an option to read in the clustering results from other methods and studies by merging the output data frame with that of the ongoing project, this part was removed in this study as the analysis is limited to stLearn. The seaborn plotting exemplified in the tutorial was left unmodified, with the exception that I uncommented out the saving of the clustering performance figure as clustering_performance.png.

Results

Data was successfully formatted and entered into stLearn for Spatial Morphological gene Expression normalized (SME Normalized) followed by stLearn spatial clustering. This produced a clustering structure visually similar to that of ground truth data (Figure 1a,b). After validating that similar performance to that of the Pham et al paper could be obtained, the program was run again with normalization functions removed. The resulting cluster map was notably less organized and strayed far further from the ground truth data than the normalized clustering did (Figure 1). However, several of the clusters such as 0,1, and 5 appear largely intact, indicating that the clustering algorithm still has a low level of ability to effectively perform clustering of tissues without SME normalization first.

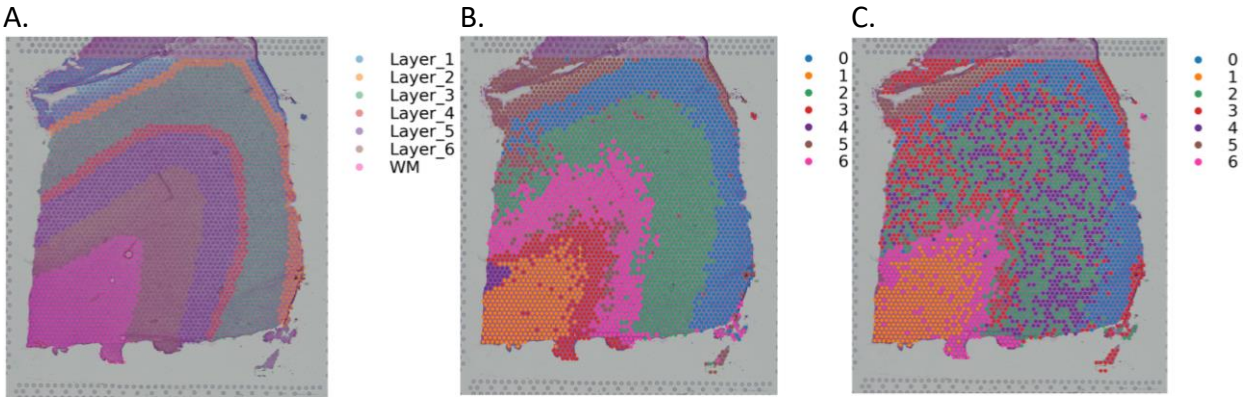

Figure 1: Comparative clustering visualization of HDLPFC tissue slide. A. Input ground truth data **B.** stLearn output when stSME spatial normalization applied **C.** stLearn output when stSME spatial normalization is not applied

Following visualization, the clustering classification with and without was analyzed on several indices for clustering precision, the first of which is a Rand score. The normalized data produced an average rand score of approximately 0.15 higher than its non-SME normalized counterpart (Figure 2, blue). This indicates that the SME normalized clustering was much more like the ground truth data than the non-normalized clustering was. The normalized data produced a normalized mutual information score of approximately 0.175 higher than its non-SME normalized counterpart (Figure 2, orange). This indicates that the SME normalized data matched ground truth in its classification %17 more than the non-normalized data set, a nearly two-fold improvement. The normalized data produced an average purity score of approximately 0.15 higher than its non-SME normalized counterpart (Figure 2, green). This indicates that the SME clustering put far more like clustered points near each other than non-normalized clustering did.

The normalized data produced an average homogeneity score of approximately 0.2 higher than its non-SME normalized counterpart (Figure 2, red). This indicates that the SME clustering put nearly 50% more like clustered points in the same class compared to the ground truth than non-normalized clustering did. The normalized data produced an average completeness score of approximately 0.175 higher than its non-SME normalized counterpart (Figure 2, purple). This indicates that the SME clustering, clustered more points in their ground truth designated class than non-normalized clustering did. The harmonic means of the homogeneity scores and completeness scores also differed by approximately 0.175, furthering that SME normalized clustering is more true to class than its non-normalized counterpart (Figure 2, brown).

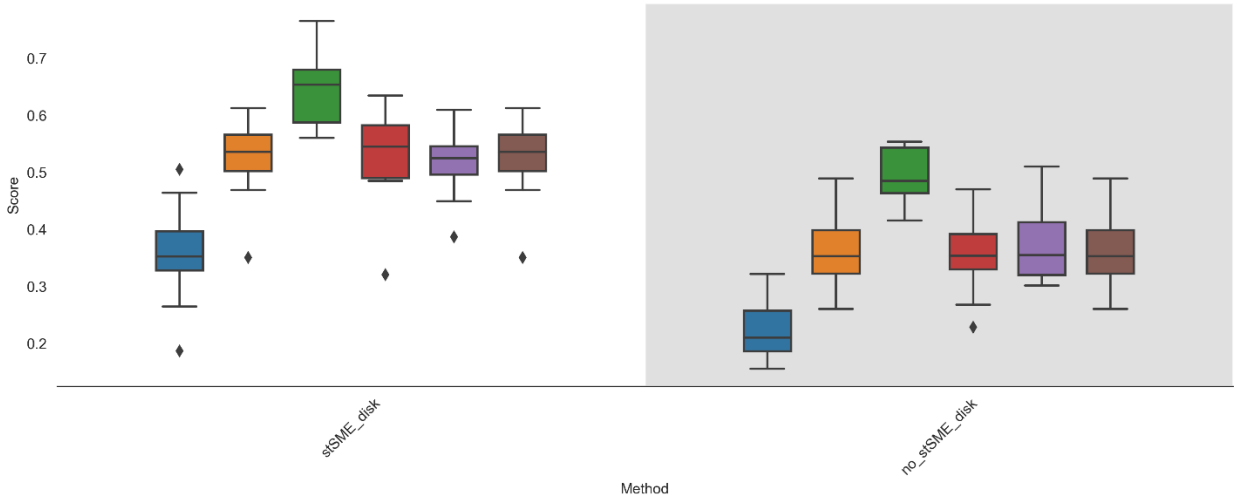

Figure 2: Statistics comparing normalized and non-normalized tissue clustering. Disk smoothing was applied to the data set on the left for each of the HDLPFC tissue slides, and not applied on the right. The scores for each slide sample were gathered and their distribution is represented by a box and whisker plot. **Blue:** Adjusted Rand Score **Orange:** Normalized Mutual Information Score **Green:** Purity Score **Red:** Homogeneity score **Purple:** Completeness score **Brown:** V measure score

Discussion and Conclusions

Altogether, the SME normalized data set appears to cluster significantly better that it's non-normalized counterpart. there is notably less scattering of clusters in the normalized data set than in the non-normalized data set (Figure 1), without measures indicating a nearly 50% improvement in tissue classification (figure 2). With this I can conclude that normalization of scRNAseq data with spatial data using stLearn produces far better tissue classifications than the stLearn clustering algorithm alone. This also furthers the theory that single cell sequencing data is far more reliable following spatial normalization than it is without it.

If this project was furthered, I would have also liked to have included the data from analysis with SpatialLIBD as well. The rand scores for that algorithm appear far more like that of the non-normalized analysis than to that of the SME normalized analysis.[1] This could also have helped determine of there were any particular measures of classification that SpatialLIBD scored better in than stLearn.

Sources

1. Pham, D., Tan, X., Xu, J., Grice, L. F., Lam, P. Y., Raghubar, A., … Nguyen, Q. (2020). stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. https://doi.org/10.1101/2020.05.31.125658

2. Pardo, B., Spangler, A., Weber, L. M., Hicks, S. C., Jaffe, A. E., Martinowich, K., … Collado-Torres, L. (2021). spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. https://doi.org/10.1101/2021.04.29.440149

3. *spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data*. an R/Bioconductor package to visualize spatially-resolved transcriptomics data http://research.libd.org/spatialLIBD/,
Additional data: https://github.com/LieberInstitute/HumanPilot

4. Matuška, O. (2019, April 3). *MI Score*. Sketch Engine. https://www.sketchengine.eu/my_keywords/mi-score/#:~:text=The%20Mutual%20Information%20score%20expresses,score%20which%20may%20be%20misleading.

5. *sklearn.metrics.normalized_mutual_info_score*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html.

6. Evaluation of clustering. https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html.

7. Bonaccorso, G. Mastering Machine Learning Algorithms. O'Reilly Online Learning. https://www.oreilly.com/library/view/mastering-machine-learning/9781788621113/2830f738-c6a5-460a-b518-23ecd3745c2d.xhtml.

8. *sklearn.metrics.completeness_score*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html.

9. *sklearn.metrics.v_measure_score*. scikit. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.v_measure_score.html#sklearn.metrics.v_measure_score.

10. *stSME clustering tutorial*. stSME clustering tutorial - stLearn 0.3.1 documentation. (n.d.). https://stlearn.readthedocs.io/en/latest/tutorials/stSME_clustering.html.