Jacob Krol Nov 6, 2023 BIOS 7747 Unsupervised Learning

Motivation

To understand clustering techniques and use them to solve a biomedical problem

Study summary

The department of developmental biology is interested in exploring if expression data from a number of genes can be used to identify specific cell types. They used a publicly available dataset to run a pilot study and explore if there are specific patterns in the data supporting their hypothesis (to be used in question number 1). Assuming that their hypothesis would be supported by their analysis, they also initiated a parallel study sequencing their own data, for which the information about each cell type was available (to be used in question 2).

Goals:

Q1:

- To evaluate the performance of different clustering techniques identifying cell types based on gene expression (use only the pilot dataset).
- Compare quantitatively the performance of K-means, DBSCAN and OPTICS identifying clusters of gene expression that may be associated with cell types. Justify parameter selection independently for each algorithm. How similar are the results obtained from the three algorithms?
- What is the most likely number of cell types based on previous results? Justify the answer using only the pilot dataset.

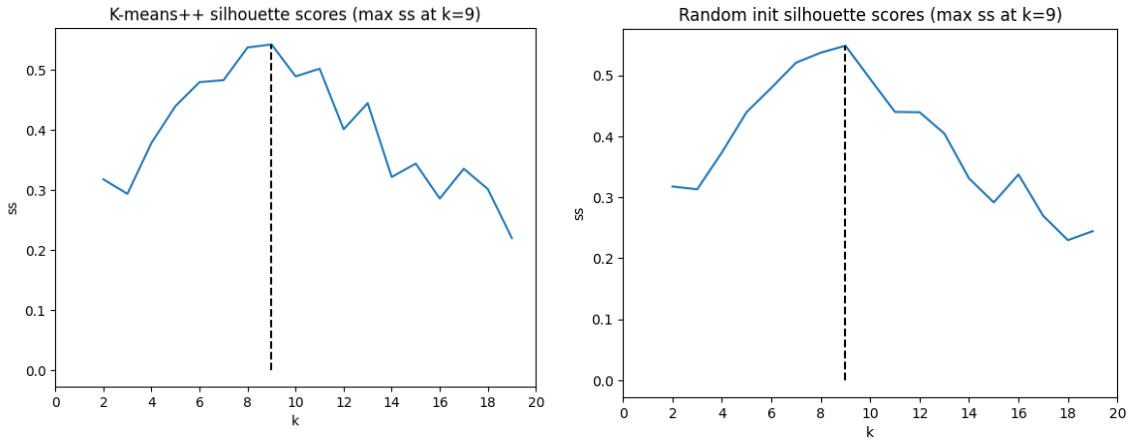

Fig 1. PILOT K-means clustering results measured with silhouette scores (ss) at various K. **A** (left): K-means++ centroid initialization method results. **B** (right): Randomized centroid initialization results.

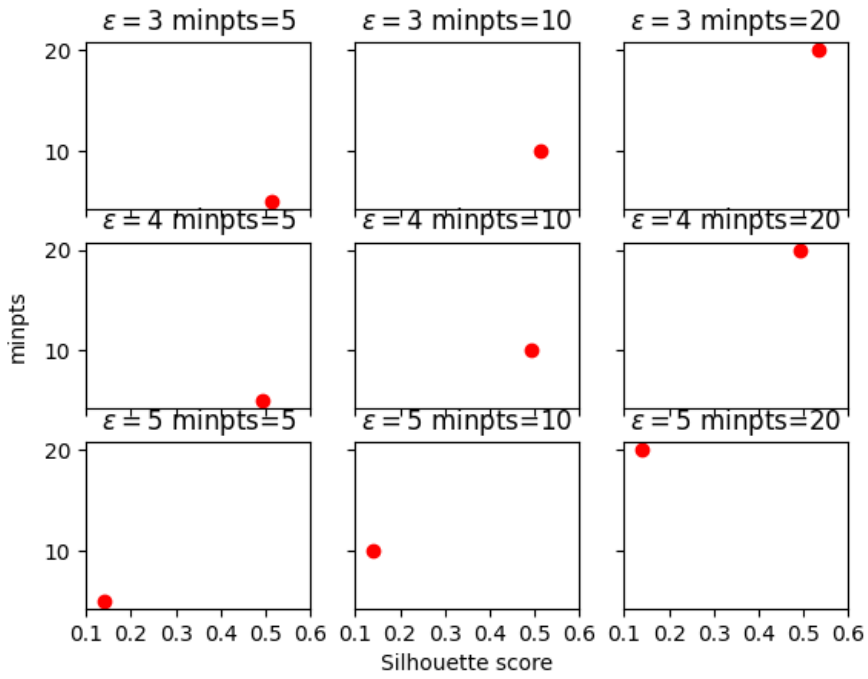

Fig 2. PILOT DBSCAN clustering hyperparameter selection. The subplots measure the silhouette score (**x-axis**) with various hyperparameter combinations (epsilon and minpts [**y-axis**]).

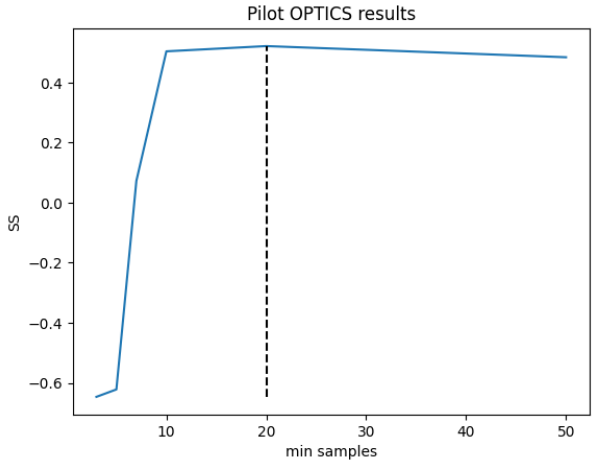

Fig 3. PILOT OPTICS silhouette score across various min samples values.

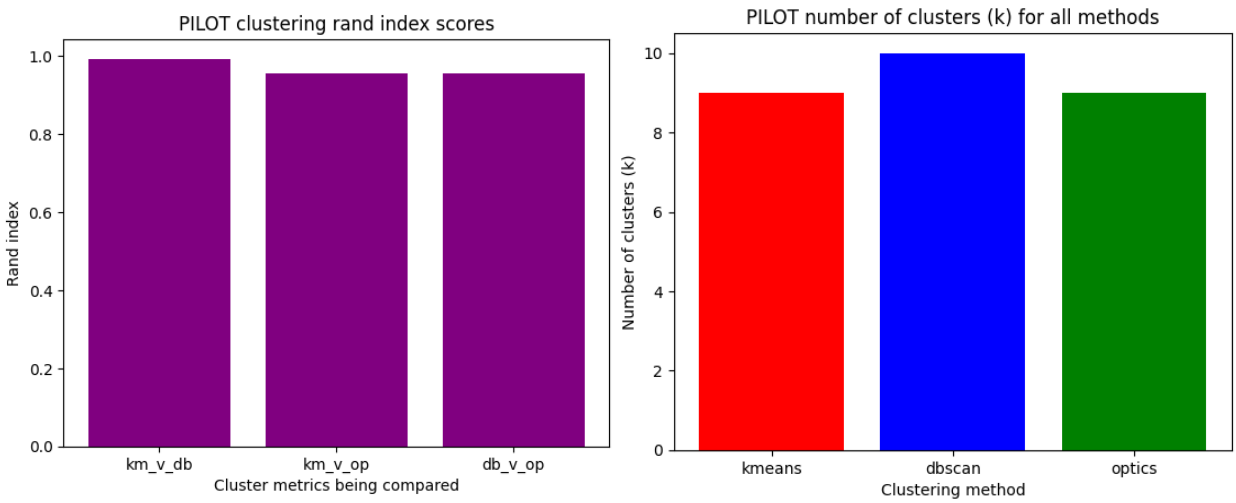

Fig 4. PILOT rand index scores between clustering methods and number of clusters for each method. **A** (left): The rand index score between each method. **B (right**): The number of clusters for each clustering method using the hyperparameters that maximized silhouette scores from **Figs 1-3**.

Using the **PILOT** dataset, the silhouette score was used to find optimal hyperparameters for each clustering method (**Figs 1-3**), and the rand index was used to measure the agreement of samples being clustered together across methods (**Fig 4A**) with an additional plot showing the number of clusters allocated for each method (**Fig 4B**). **Figs 1-3** demonstrated hyperparameter selection of three different clustering methods K-means, DBSCAN, and OPTICS, and the optimal hyperparameters were selected when the silhouette score (SS) was maximized. For each method, the maximum silhouette score in the grid search was around 0.55. The evaluation of cluster agreement using the rand index in **Fig 4A** indicates that the agreement is very high (i.e., samples are placed in the same relative cluster across methods) between all 3 methods with K-means and DBSCAN's rand index value approaching 1 (1 indicates perfect cluster agreement). The number of clusters for each method (based on calculating silhouette score across various hyperparameters [**Figs 1-3**]) was 9 for K-means and OPTICS and 10 for DBSCAN (11 if the noise cluster is considered) shown in **Fig 6**. Based on number of clusters at maximum SS for each clustering method (**Fig 4B**) on the pilot dataset, the predicted number of cell types is likely 9 or 10 groups; however, the K-means SS plot shows that 8 or 9 might be most optimal.

Q2

- To evaluate the performance of clustering techniques on a separate dataset (use the test dataset).
- Evaluate quantitatively the performance of the previous methods in the new dataset for which the cell types are known.
- Compare the performance of previous clustering techniques with linear discriminant analysis. Justify any difference in the obtained results.

Note: support your discussion with plots and graphic representation when possible.

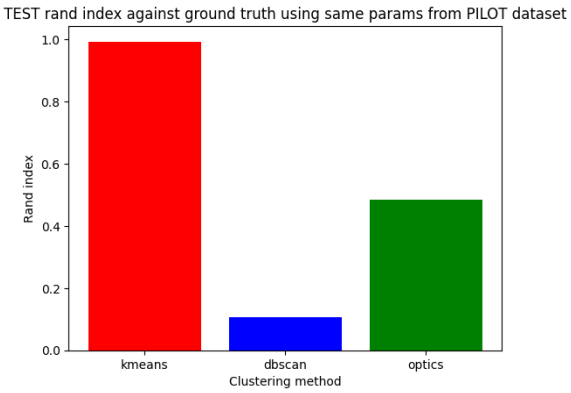

Fig 5. Evaluate clustering methods (Rand index against ground truth) fit on the TEST datasets using hyperparameters from PILOT dataset results.

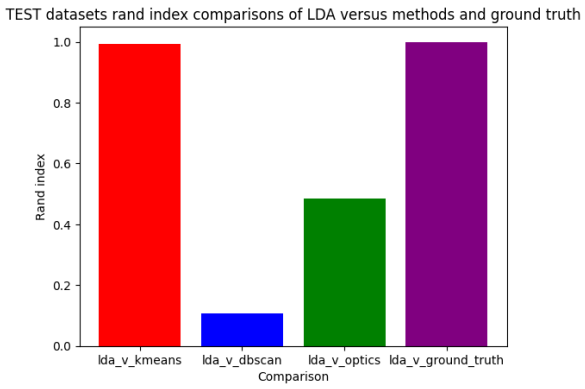

Fig 6. TEST dataset Linear Discriminant Analysis (LDA) results against clustering methods and ground truth.

Using the same hyperparameters from the PILOT dataset, the previous clustering methods (Kmeans, DBSCAN, and OPTICS) were performed on the TEST dataset and the accuracy was measured by calculating the rand index against the ground truth (**Fig 5).** Kmeans clustering best

matched the ground truth cell types (RI=0.99), DBSCAN performed the worst (RI=0.11), and OPTICS clustered moderately well (RI=0.48).

LDA was performed using the TEST data and the ground truth cell types. The results were compared against Kmean, DBSCAN, OPTICS, and the ground truth (**Fig 6**). LDA's cluster labels had full agreement with the ground truth (RI=1.0), near full agreement with Kmeans (RI=0.99), low cluster agreement with DBSCAN (RI=0.11), and moderate agreement with OPTICS (RI=0.48). Finally, when using LDA to re-predict on the TEST dataset it performed (RI=1) which suggests the assumptions of linearity, uniform variance, and normal distributions hold for the gene expression data. Along with the assumptions of LDA, LDA also made use of the ground truth clusters in the model fitting which likely explains its dominant performance.

In conclusion, Kmeans clustering performed well on the test dataset using the same hyperparameters found from the pilot dataset **Fig 5** and mimicked the performance of LDA clustering on the test dataset **Fig 6**. Interestingly, despite the cluster agreement between Kmeans, DBSCAN, and OPTICS in the pilot dataset, shown by high rand index scores in **Fig 4A**, DBSCAN and OPTICS did not perform well on the test dataset using the hyperparameters that were optimal in the pilot dataset (**Fig 5**).