Project 2 Report

Tram Le

In this project, I clustered RNA-Sqe data using KMeans clustering and reductional dimension as unsupervised learning techniques with two methods and evaluated the performance on each. Before apply the methods, I do data analysis with:

- Data processing and exploring: Initially, I load the two dataset **'data-1.csv'**, containing features without any labels and **'labels.csv'**, containing all the sample labels. The first columns in **'data-1.csv'** contain string datatype, so I removed it to retain only numerical gene values, and I do the same process on **'labels.csv'** to ensure it contained only labels itself. Next, I explored the dataset to confirm there is no noise values.

- Data standardization: I standardized the features to have mean 0 (centered around 0) and deviation of 1. In my opinion, this process is essential because some gene values are high and some are small

- Then I perform the two method:

Method 1: I applied KMeans directly on both raw data and scaled (normalized) data, then visualized the clusters using dimension reduction with PCA. I evaluate the performance using the Adjusted Rand Index and Normalized Mutual Information, metrics that measure the similarity between clusters rather than the distances within them. These metrics are suitable for evaluating unsupervised learning, providing an overall accuracy.

- I set k is 5 to match with the numbers of labels (type of tumor), and ran the algorithms 10 times with different centroid seeds (n_init=10), it helps to pick the best solution of inertia to be chosen and I use random_state as 50 consistently initialize the centroids

- This dataset is very high dimension, so I choose PCA to visualize KMeans of the entire features, I choose 2D PCA with n_components=2 to compare KMeans on both raw data and scaled data. This method show the PCA-transformed data where the colors and cluster labels are determined by the KMeans algorithm

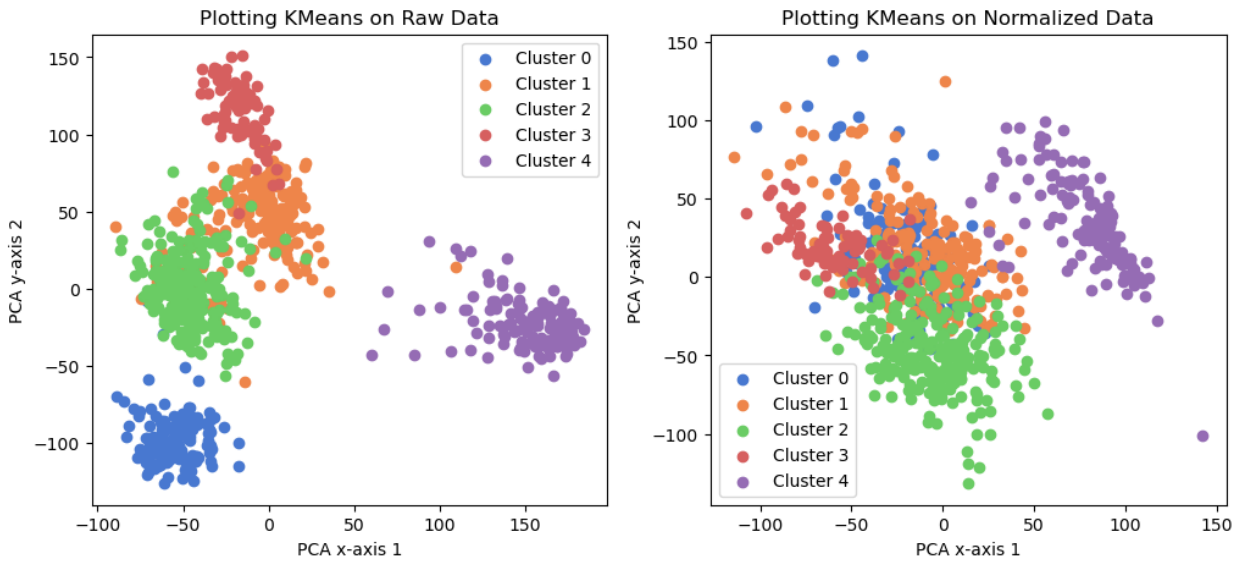

Figure 1: Compare Plotting Kmeans on Raw Data vs Normalized Data

Result: The plots showing that the normalized data is more overlap and less distant clusters, while the raw data is more distinct and separate from each other. As a result, the overall accuracy on raw data is 98.51% with adjusted rand index and 97.72% with normal mutual information. The standardized data achieved 80% with adjusted rand index and 85.62% with normal mutual information.

Method 2: I want to see if there is any improvement or differences in reduced dimensional space, I performed KMeans on PCA and T-SNE transformation of raw data, then evaluating KMeans on each reductional space. Since KMeans perform well on raw data, so for the second part, I transformed raw data to PCA and T-SNE. I visual comparisons of PCA with true labels and KMeans predict labels, as well as t-SNE with true labels and KMeans prediction.

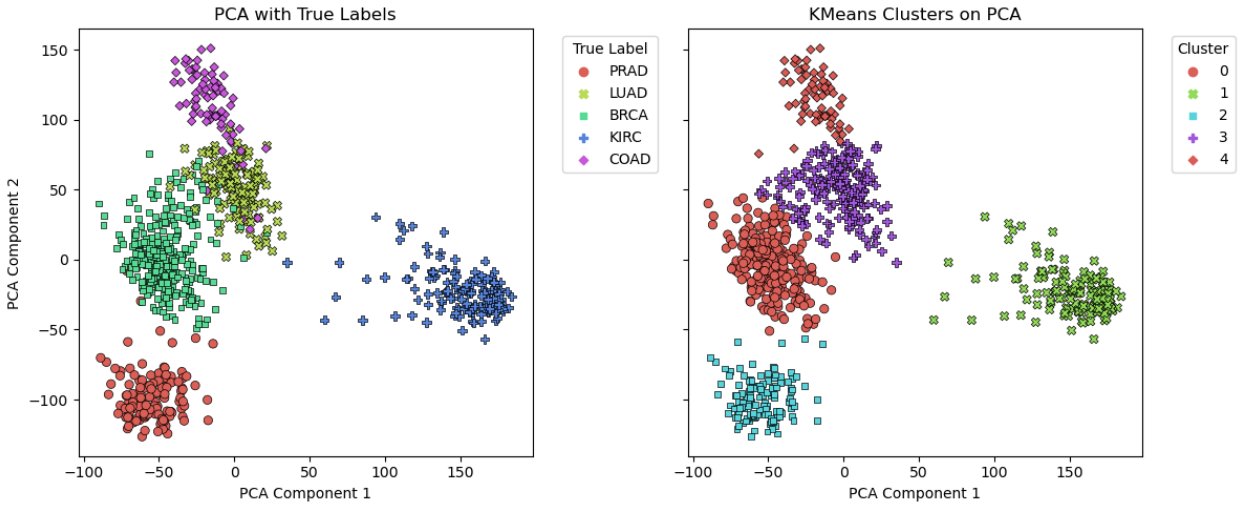

Figure 2: Showing PCA vs Kmeans on PCA

Result: KMeans Clusters on PCA Plot is distinct from each other in the last 2 labels *'KIRC'* and *'COAD'* (base on coloring). As a result, I achieved with 80.71% accuracy with adjusted rand index and 84.18% with normal mutual information, it's similar to standardized data result from method 1

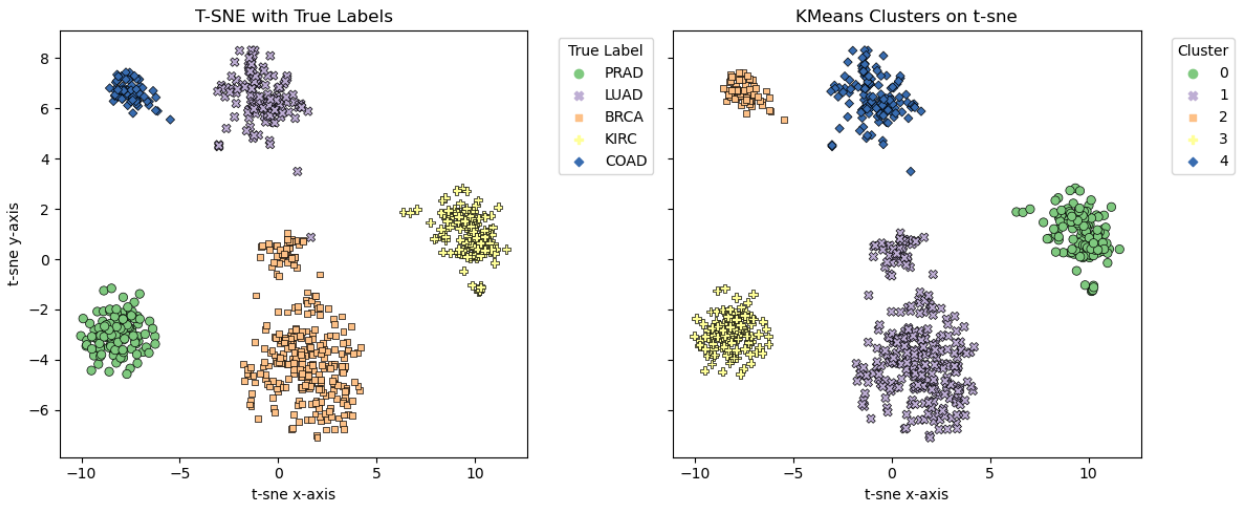

Figure 3: Showing T-SNE vs Kmeans on T-SNE

Result: For visualization, the KMeans Cluster on T-SNE visualize is identical between the predict label and true labels such as *'PRAD'* labels (green) in left plot also appear as *'cluster 0'* (green) in the right plot. The overall accuracy is 99.62% with adjusted rand index and 99.47% with normal mutual information. This result indicate a close similarity between the true labels in t-SNE reduced dimension and predict labels in the KMeans clusters.