**Project 2**

**Evan Lydon**

The test and train datasets were loaded into separate pandas dataframes. The cellular location labels were stored in a separate series for clustering purposes. Kmeans, GMM and agglomerative hierarchical clustering were applied to the training dataset. The n\_cluster parameter in each clustering method was varied and scored based on average silhouette score, the Calinski-Harabasz score, the DBI score and the elbow method. For kmeans, the average silhouette score and CH score had local maximums around 8 and 9 and the DBI score had a local minimum at 9; the elbow method showed a slight change around 8 to 9, aswell. Agglomerative hierarchical clustering with complete and average linkage showed local maximums for the above mentioned scoring methods in the same area. Given that the authors of the paper identify 8 cellular regions, this information supports further clustering classification using n\_clusters = 8(1).

Kmeans, GMM and AHC were run again on the training dataset using n\_clusters = 8 and their agreement with the known labels was assessed using the Munkres package installed from PyPI. The AHC with average linkage performed best in terms of classifying the datapoints with a classification success rate of 76% on the training dataset. With complete linkage, AHC fell to 72%. GMM correctly classified at 66% and Kmeans at 60%.

The AHC with average linkage was then applied to the full dataset (combined train and test datasets) in order to obtain labels for the 15 unknown cellular locations. This yielded the following classification for the unknowns: ['cp', 'cp', 'cp', 'cp', 'cp', 'im', 'im', 'im', 'im', 'im', 'pp', 'pp', 'pp', 'pp', 'pp'].

For comparison, the random forest algorithm was trained on the training dataset and achieved an 86% out of bag classification success rate. The trained forest was then used to classify the unknown data in the test set and arrived at the following classification for the unknowns: ['cp' 'cp' 'cp' 'cp' 'cp' 'im' 'im' 'im' 'im' 'im' 'pp' 'cp' 'pp' 'pp' 'pp']. There is a classification disagreement at the 12th datapoint, but otherwise the classification agrees well with that from the AHC method.

**Citations:**

|  |  |
| --- | --- |
| (1) | Nakai, K.; Kanehisa, M. Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria. *Proteins* **1991**, *11* (2), 95–110. |





