PR3: Clustering

Description:

****
This is an individual assignment.
****

Overview and Assignment Goals:

The objectives of this assignment are the following:

- Use an existing clustering algorithm (e.g., K-means, DBSCAN, Agglomerative, etc).
- Extract features from time series data.
- Think about best metrics for evaluating clustering solutions.

Detailed Description:

For the purposes of this assignment, you will cluster EKG signals from 23 subjects captured while performing 5 different activities. Each sample represents 1 second of EKG signal and there are 500 samples for each activity performed by each subject, for a total of 11500 samples. Each sample has 178 EKG values.

You will need to first decide whether you will use the samples as-is or try to extract or add additional features from the given features. You may also think about ways you might want to normalize the data. Does it make sense to standardize the features, or the samples, or neither? After processing the samples, cluster them into 115 clusters and submit the result to CLP. For each sample in the dataset, in order, record a number from 1 to 115 denoting the cluster ID that the sample belongs in, one per line.

The train.dat file is a simple CSV file containing the 11,500 samples. The format.txt file shows you an example submission file with random assignment of cluster IDs for each sample.

For evaluation purposes (leaderboard ranking), we will use the Normalized Mutual Information Score (NMI), which is an external index metric for evaluating clustering solutions. Essentially, your task is to assign each of the instances in the input data to K clusters identified from 1 to K. All objects in the training data set must be assigned to a cluster. Thus, if using DBSCAN, you can either assign all noise points to cluster K+1 or apply post-processing after DBSCAN and assign noise points to the closest cluster.

Some things to note:

- The public leaderboard shows results for 50% of randomly chosen test instances only. This is a standard practice in data mining challenges to avoid gaming of the system. The private leaderboard will be released after the deadline and evaluates all the entries in the data set.
- Each day, you can submit a prediction file up to 5 times.

- There are no test.dat files in this assignment.

Rules:

- This is an individual assignment. Discussion of broad level strategies is allowed but any copying of submission files and source codes will result in honor code violation.

Deliverables:

- Valid submissions to the Leader Board website: https://clp.engr.scu.edu/ (username is your SCU username, password is your SCU password).
- **Canvas Submission of report:**
  - Create a 2-4 page, single-spaced report describing details regarding the steps you followed for developing the clustering solution for text data. The report should be in PDF format and the file should be called **<SCU_ID>.pdf**. Be sure to include the following in the report:
    1. Name and SCU ID.
    2. Rank & NMI for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
    3. Your approach.
    4. Implement/Use your choice of internal evaluation metric and plot this metric on the y-axis for the clusterings you obtained using the different clustering methods you tried. Make sure you label each method in the plot.
    5. Describe, any feature selection/reduction or custom proximity measure you used in this study.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (30%).

Files:

- On Canvas