PR3: Image Clustering

Published Date:
Feb 24, 2020, 9:10 a.m.

Deadline Date:
Mar 13, 2020, 11:59 pm

Description:

****
This is an individual assignment.
****
Overview and Assignment Goals:

The objectives of this assignment are the following:

- Experiment with different image feature extraction techniques.
- Use an existing clustering algorithm (e.g., K-means, DBSCAN, etc).
- Use dimensionality reduction techniques (must try at least one, even if you do not use it in the final solution).
- Design a proximity function for clusters of image data.
- Think about best metrics for evaluating clustering solutions.

Detailed Description:

For the purposes of this assignment, you will extract features from images and cluster them. *You may use any libraries both for the feature extraction and for the clustering.* Additionally, you will gain experience with internal cluster evaluation metrics.

Traffic congestion seems to be at an all-time high. Data Mining methods must be developed to help solve traffic problems. In this assignment, you will analyze features extracted from tiny traffic images depicting different traffic-related objects to determine groups of like-objects. This is an important task in many traffic analysis tasks, including vehicle tracking, counting, and outlier detection. You can experiment with any kind of image feature extraction techniques you would like, as long as the output is a vector representation of the image. Note that, as in the real world, images have different sizes, so that will be something you will need to deal with. A few classic examples of image features are Histogram of Oriented Gradients (HOG) features, Normalized Color Histogram (Hist) features, Local Binary Pattern (LBP) features, Color gradient (RGB) features, Depth of Field (DF) features, etc.

Once you obtain feature vectors for all images, cluster the data such that you obtain 14 clusters using an off-the-shelf clustering method (e.g., K-means, DBSCAN, etc.). You can use existing libraries for this step.

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. Additionally, you will need to compute an internal clustering effectiveness measure for each cluster in your clustering, which you will include in the report.

The output of your clustering should be a file with 100,000 lines, each line containing the cluster label (a number in [1-14]) for the associated sample image. The leaderboard will report the NMI on 50% of the samples from the dataset.

Data Description:

The dataset consists of 100,000 images. We have included a small subset of the data in the *traffic-small* dataset which you may use to practice your codes locally. We have also included a sample clustering output for the traffic-small dataset in the file clusters.txt within that dataset directory. However, we recommend you use the HPC for the final computation on the *traffic* dataset. Note that the dataset file is very large (**236 MB**, expands to **524 MB**). Ensure you have enough space on your drive before expanding the file. Moreover, you will need to think carefully about how you will organize computations so you do not run out of RAM during training. Finally, remember this in an unsupervised problem and you should not attempt to use the example clustering output provided to you for the traffic-small dataset in finding your overall solution for the traffic dataset.

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 cluster assignment file up to 5 times.
- The final ranking will always be based on the last submission.

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.
- Feel free to use the programming language of your choice for this assignment.
- You can use any libraries for dealing with input data or to obtain a clustering result.

Deliverables:

- Valid submissions to the Leader Board website: https://clp.engr.scu.edu (username is your SCU username and your password is your SCU password).
- **Canvas Submission for the report:**
- Include a 2-page, single-spaced report describing details regarding the steps you followed for feature extraction, feature selection/reduction, and clustering. 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:
    - Name and SCU ID.

        - Rank & NMI-score for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
        - Your approach.
        - Your methodology of choosing the approach and associated parameters.
        - Implement/Use your choice of internal evaluation metric and plot this metric on the y-axis for the 14 clusters in your final result. Sort the values in decreasing order, so as to show their distribution. Label each point/bar with the number of points in that cluster, or, alternatively, plot the sizes of the clusters with a separate line in the same plot.
    - Ensure you submitted the correct code on CLP that matches your output. Code does not need to be submitted on Canvas.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (20%) and ranking results (10%).

Files:

- Due to the size of the dataset, files are available through the CLP server, at: https://clp.engr.scu.edu/static/datasets/traffic-cl.tar.gz
- Additionally, a copy of the dataset can be found on the HPC, under: /WAVE/projects/COEN-281-Wi20/data/traffic and /WAVE/projects/COEN-281-Wi20/data/traffic-small. You should access the data from there and not make copies of the dataset to your own directories.