German University in Cairo
Faculty of Media Engineering and Technology
Spring 2019

CSEN 1083 – Data Mining

Assignment #3

(Due on: May 1 at mid-night)

(This assignment can be done in groups of 2 students each)

In this assignment, you will use the Weka data mining tool to cluster the data of Assignment 2 into 26 clusters. The Weka tool can be downloaded from:
https://www.cs.waikato.ac.nz/~ml/weka/downloading.html

The images that you used in Assignment 2 (training images) are stored in the file Images.csv. Each row in the file represents one image (total 182 images) and each column represents the brightness of one pixel (total 144 pixels). The last column represents the character that is present in each image.

Use Weka Explorer to examine the performance of each of the clustering algorithms available in Weka on the data. Ideally, the clustering algorithm should cluster the 7 images corresponding to each character in a separate cluster. Use the ignore attribute button in the clustering tab to remove the last column before clustering the data.

In the Hierarchical clustering, examine 3 different linkage methods (Single, Complete and Average).

Deliverables:

- A PDF report file showing snapshots of the output obtained using Weka. The snapshots should represent a figure for each clustering algorithm examined in which the x-axis represents the image number (0 to 181) and the y-axis represents the label assigned to the image by the clustering algorithm.
- In the submitted PDF file, state the conclusion of examining all these algorithms on the data. Which method appears to be the best?
- A ".model" file saved from Weka for each of the following algorithms:
  - Hierarchical Clustering with Single Linkage
  - Hierarchical Clustering with Complete Linkage
  - K-means clustering