CSCI 5523 Project 3 - Clustering Due: May 6 2019

*Note: Please read the entire project description, especially the **INSTRUCTIONS** and **SUBMISSION GUIDELINES** before you begin.*

Instructions

- Please use the same version of Weka that you used for Project 1. It can be downloaded from the Project 1 files.

- You will need to install extra package to perform DBSCAN. The instruction can be found at the end of this file.

- Under Cluster tab set the "**Cluster Mode**" to "**Classes to clusters evaluation**". Also, wherever running clustering algorithms use the default value for the 'Seed' parameter.

- You will need the following clustering algorithms for this project:

Clustering algorithm	Weka name	Settings
K-Means	Simple KMeans	
Hierarchical clustering	HierarchicalClusterer	Link type = AVERAGE
DBSCAN	DBSCAN	

- Visually explore the data sets in all experiments by considering the following characteristics of the data. This exploration will help you understand the data and answer the questions.
  - Types of attributes
  - Class distribution
  - Which attributes, if any, appear to be good predictors of the cluster structure
  - Possible correlation between attributes
  - Any special structure in the attributes that you might observe

- You will need to use an online entropy calculator that takes the confusion matrix returned by Weka as an input and outputs entropy value. The tool can be accessed here: http://www-users.cs.umn.edu/~lixx1166/entropy.php

Submission Guidelines

- Prepare a report (no more than 6 pages, no smaller than 11pt font, 1-column format) addressing the questions asked in the following problems. **Submit the hard copy of your**

report in class on May 6, 2019.

- Your project will be evaluated based **only** on what you write in your report.
- Do not simply copy and paste program output into your report, unless specifically asked for it. **Program Output should be submitted electronically** as a single zipped file using the Canvas submission tool. We will only look at your output if something is ambiguous in your report.
- Please **provide the answer to each specific question we asked separately**; do not compose a big paragraph that includes everything.
- If you are a ***UNITE*** student, you should upload your Weka output on Canvas like other students and submit your project report via UNITE as homework.

Problem 1

Dataset to be used in this experiment: Ecoli_v2

This is a modification of the e-coli UCI dataset that contains 8 attributes and 4 classes. The classes are protein localization sites in the E-coli bacteria: site1, site2, site3, and site4, which have to be discovered using clustering. The objective of this problem is to become familiar with various clustering techniques. Load the dataset into the Weka Explorer and visualize the data.

1. Perform the following set of experiments.
    a. Perform K-means clustering with K = 4. Report the confusion matrix and the entropy. Comment on the quality of this clustering based on your analysis of the confusion matrix and entropy. Which classes are easily separable from others and which classes are mostly confused?

    b. Perform K-means clustering with K = 3. What changes did you notice in the confusion matrix and entropy of the overall clustering? Based on these observations, what can you comment on site2, site3, and site4 classes?

    c. Plot the SSE as the number of clusters varies from 2 to 10. Describe the behavior of this plot in a sentence or two.

2. Use the 'HierarchicalClusterer' clustering approach to find 3 and 4 clusters. Make sure to set the **'linkType'** property of this method to **'AVERAGE'**. Report the confusion matrix and the entropy. Also, answer the following questions:

   a. Comment on the quality of this clustering based on your analysis of the confusion matrix and entropy. Which classes are confused?

   b. Based on these observations, how does 'HierarchicalClusterer' differ from K-means? Which property of the data may cause this effect?

Problem 2

Datasets to be used in this experiment:

1) *p2data1*
2) *p2data2*
3) *p2data3*

In all these dataset, there are 5 classes of points, three of which form clusters and two of which are noise. Load the p2data1, p2data2 and p2data3 data sets into the Weka Explorer and visualize them. You will be using the 'HierarchicalClusterer' method for this problem. Make sure to set the '**linkType**' property of this method to '**AVERAGE**'.

1) Use the HierarchicalClusterer clustering approach to cluster the points in all the three datasets for K = 1, 2, 4, 6, 8 clusters. Report the confusion matrix and the corresponding entropy for each run. Plot the trend followed by entropy for each of the three data sets as the number of clusters is varied (plot all the three curves in the same figure). Comment on how entropy varies as the number of clusters increases, and also how the entropy compares among the three data sets as the number of clusters is varied. What characteristic difference(s) between the data sets lead to this variation in entropy? Explain in detail.

2) What are the general problems with these data sets that may lead to poor clustering results? Is there anything that can be done to improve the clustering results obtained with the HierarchicalClusterer technique?

Problem 3

Datasets to be used in this experiment:

1) *p2data1*
2) *p2data2*
3) *p2data3*

In all these dataset, there are 5 classes of points, three of which form clusters and two of which are noise. Load the p2data1, p2data2 and p2data3 data sets into the Weka Explorer and visualize them.

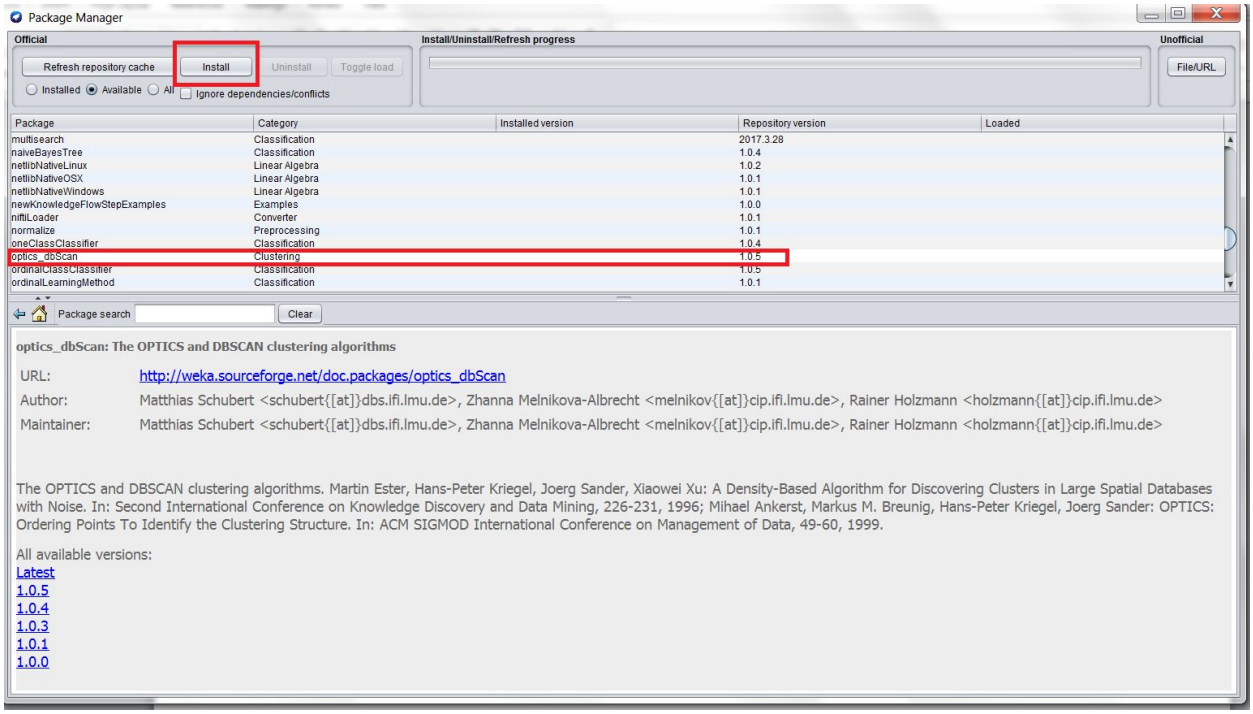
1) Use the **DBSCAN** clustering approach to cluster the points in all the three data sets with the following settings: **epsilon = 0.05**, **minPoints = 4**. Report the confusion matrix and the corresponding entropy for each run. Comment on how the entropy compares among the three datasets. What characteristic difference(s) between the data sets lead to this variation in entropy? Explain in detail.

2) Use the **DBSCAN** clustering approach to cluster the points in all the three data sets with the following settings: **epsilon = 0.1**, **minPoints = 4**. Report the confusion matrix and the corresponding entropy for each run. Comment on how the entropy varies comparing with the results from part 3 as Epsilon increases, and briefly explain.

3) Based on your observations in dealing with datasets above which involved noise, and varying density, what conclusions can you make about the behavior of hierarchical clustering and DBSCAN? Please limit your answer to this question to no more than a page.

Instruction: How to install DBSCAN

Step 1: Open Weka Package manager



Step 2: Select package: ***optics_dbScan***, and hit “Install”.

Step 3: Restart your Weka, and open Explorer. You should be able to find DBSCAN under “Cluster” tab.