Lonari, Nikit
Final Project
Machine Learning

Hierarchical Clustering for Seed Categorisation

In seed dataset, we have 210 examples. I have divided this dataset in two parts training and test data. Since, the available dataset has three class labels as 1,2,3 so, my training data has examples from 1 to 50, 71 to 120, 141 to 190 making the set of 150 training examples. Rest all the samples are in test dataset.

First, I have created a distance metrics (using Euclidean distance) for all training data set which makes a matrix of 150*150; Next, to begin hierarchy I have done using Single linkage. Below, fig. show first few iterations of clustering, i.e which cluster goes with other, and also has the distance between them.

8	29	0.17901064
35	50	0.20601721
72	73	0.212035
8	22	0.21647873
14	15	0.21748621
108	124	0.22458183
77	84	0.22472659
45	48	0.22613511
123	143	0.23799935
129	133	0.25306292
114	130	0.25682774
113	118	0.2583332
72	85	0.25916267
105	119	0.2691517
127	142	0.27162909
113	127	0.28461801
3	8	0.28493018

Second, I have used a library named as cluster data which creates cluster for the given metrics. This library helped me to generate automatic threshold value for clustering. Later, I got matrix having clusterID's from clusterdata.

Then I took each example from training dataset having same classID's, calculated there average value for each cluster. Then, I applied Euclidean distance between each clusters averaged value with each testing sample; for that particular test sample I checked the minimum distance, this minimum distance gave me the classID for the same test example.

Prediction shows the predicted clusterID for the 60 test samples, which depends upon number of clusters user wants to create. Moreover, the UCL seed dataset has 3 classID's so, we can check its prediction.
Don't go with the exact number but each number suggest that they are in same cluster. So number really doesn't matter.

So output for only three cluster has,

* 14 test examples predicted correctly, which belongs to same cluster. i.e cluster number 2 in my o/ p and classID 1 in UCL dataset.
  (14/20)*100=0.7. 70% accuracy.

* 20 test examples predicted correctly, which belongs to same cluster. i.e cluster number 3 in my o/p and classID 2 in UCL dataset.
  (20/20)*100=1. 100% accuracy.

* 17 test examples predicted correctly, which belongs to same cluster. i.e cluster number 1 in my o/p and classID 1 in UCL dataset.
  (17/20)*100=0.85. 85% accuracy.

Overall accuracy for just three clusters is 85%.