MACHINE LEARNING - CSE 6363

Sai Santhosh Bhattaram – 1001874167

Project Report - Hierarchical Clustering for Seed Categorization

Description:

To Implement Hierarchical Clustering on the UCI seed dataset to divide it into groups and use the cluster IDs as features for a subsequent K nearest neighbor classifier to identify the target. Should use multiple clustering and have to determine what a good number of clusters would be based on determining similarity between clusters and a data point.

Implementation Details:

The entire project has been implemented in Python from scratch without using any machine learning libraries. The project has two goals: to determine the clusters and then perform KNN classification based on the newly generated data with cluster ID as an additional feature.

1. Clustering:

The project has been implemented - with Single, Average, and Complete linkages based on the user choice. The clustering would happen based on the linkage selected and the optimal number of clusters would be identified based on the maximum Dunn Index value (runs for different clusters 2, 3,4,5,6,7,8,10 and get the maximum Dunn index of these). Dunn Index is the ratio of the lowest intercluster distance to the highest intracluster distance. The higher the Dunn Index value more similar the clusters would be. Based on the optimal cluster found (based on max Dunn index) the clustered index would be added.

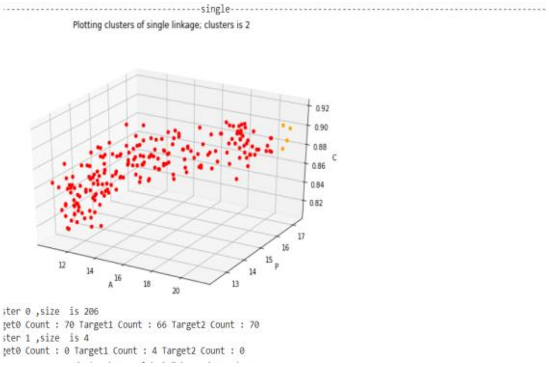
2. Classification KNN:

KNN classifier is run to predict the target labels of the new feature data (training data along with new feature cluster membership). Leave-one out algorithm is used to find the accuracy of the model. The model is run for different values of k like 1,3,5,7,9 and accuracies were noted.

Results:

1.First Experiment:

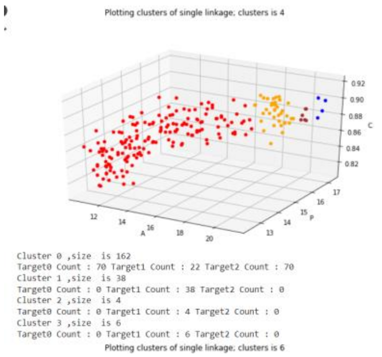
Formed clusters for all linkages (single, complete, average) and for the number of clusters 2,4, 6,8. Attaching a sample result screenshot for reference.



```
------------------------------------single------------------------------------------------------------
```

Plotting clusters of single linkage; clusters is 2

```
ster 0 ,size  is 206
get0 Count : 70 Target1 Count : 66 Target2 Count : 70
ster 1 ,size  is 4
get0 Count : 0 Target1 Count : 4 Target2 Count : 0
```



Plotting clusters of single linkage; clusters is 4

```
Cluster 0 ,size  is 162
Target0 Count : 70 Target1 Count : 22 Target2 Count : 70
Cluster 1 ,size  is 38
Target0 Count : 0 Target1 Count : 38 Target2 Count : 0
Cluster 2 ,size  is 4
Target0 Count : 0 Target1 Count : 4 Target2 Count : 0
Cluster 3 ,size  is 6
Target0 Count : 0 Target1 Count : 6 Target2 Count : 0
```

Plotting clusters of single linkage; clusters is 6

Experiment 2:

Based on user choice the linkage is being selected, the optimal number of clusters is being found and the KNN algorithm would be run

i) **Single**: The max optimal cluster found is 8 and leave- one out algorithm accuracy for KNN algorithm and comparing with KNN algorithm with actual training data without cluster Index membership is also provided below.

Linkage	Optimal cluster based on max Dunn Index	KNN with cluster Index accuracy -Leave One out	KNN without cluster Index accuracy -Leave One out

Single	8	90.95238095238095(k=1)
		89.04761904761904(k=3)
		88.57142857142857(k=5)
		90.47619047619048(k=7)
		90.95238095238095(k=9)
			90.47619047619048(k=1)
			88.57142857142857(k=3)
			87.61904761904762(k=5)
			89.52380952380953(k=7)
			90.47619047619048(k=9)

ii) **Complete**: The max optimal cluster found is 7 and leave- one out algorithm accuracy for KNN algorithm and comparing with KNN algorithm with actual training data without cluster Index membership is also provided below.

Linkage	Optimal cluster based on max Dunn Index	KNN with cluster Index accuracy --Leave One out	KNN without cluster Index accuracy -Leave One out

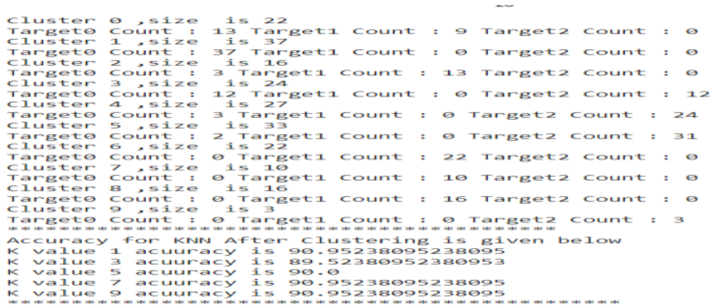
Complete	7	90.95238095238095 (k=1) 90.0 (k=3) 89.04761904761904 (k=5) 90.95238095238095 (k=7) 90.0 (k=9)	90.47619047619048 (k=1) 88.57142857142857 (k=3) 87.61904761904762 (k=5) 89.52380952380953 (k=7) 90.47619047619048 (k=9)

iii)Average: The max optimal cluster found is 9 and leave- one out algorithm accuracy for KNN algorithm and comparing with KNN algorithm with actual training data without cluster Index membership is also provided below.

Linkage	Optimal cluster based on max Dunn Index	KNN with cluster Index accuracy --Leave One out	KNN without cluster Index accuracy -Leave One out
Average	9	90.47619047619048 (k=1) 88.09523809523809 (k=3) 87.14285714285714 (k=5) 88.09523809523809 (k=7) 87.14285714285714 (k=9)	90.47619047619048 (k=1) 88.57142857142857 (k=3) 87.61904761904762 (k=5) 89.52380952380953 (k=7) 90.47619047619048 (k=9)

Experiment 3 :

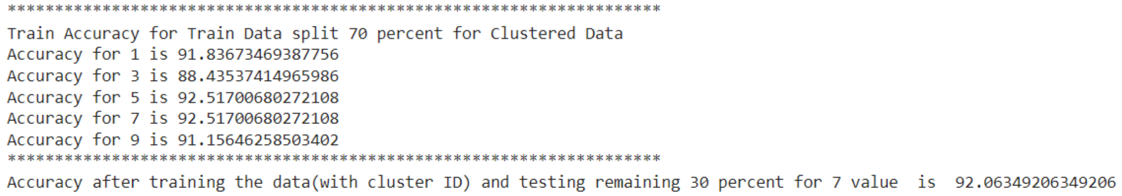
The number of clusters and linkage based on user choice max of 10 clusters(should give more colors for plotting if more clusters are needed only 10 colors are defined at present). Providing results for the same.



```
Cluster 0 ,size  is 22
Target0 Count : 13 Target1 Count : 9 Target2 Count : 0
Cluster 1 ,size  is 37
Target0 Count : 37 Target1 Count : 0 Target2 Count : 0
Cluster 2 ,size  is 16
Target0 Count : 3 Target1 Count : 13 Target2 Count : 0
Cluster 3 ,size  is 24
Target0 Count : 12 Target1 Count : 0 Target2 Count : 12
Cluster 4 ,size  is 27
Target0 Count : 3 Target1 Count : 0 Target2 Count : 24
Cluster 5 ,size  is 33
Target0 Count : 2 Target1 Count : 0 Target2 Count : 31
Cluster 6 ,size  is 22
Target0 Count : 0 Target1 Count : 22 Target2 Count : 0
Cluster 7 ,size  is 10
Target0 Count : 0 Target1 Count : 10 Target2 Count : 0
Cluster 8 ,size  is 16
Target0 Count : 0 Target1 Count : 16 Target2 Count : 0
Cluster 9 ,size  is 3
Target0 Count : 0 Target1 Count : 0 Target2 Count : 3
**************************************************
Accuracy for KNN After Clustering is given below
K value 1 acuuracy is 90.95238095238095
K value 3 acuuracy is 89.52380952380953
K value 5 acuuracy is 90.0
K value 7 acuuracy is 90.95238095238095
K value 9 acuuracy is 90.95238095238095
************************************************************
```

Experiment 4:

Splitting 70 percent train and 30 percent test data and comparing accuracies are provided below.



```
**********************************************************************
Train Accuracy for Train Data split 70 percent for Clustered Data
Accuracy for 1 is 91.83673469387756
Accuracy for 3 is 88.43537414965986
Accuracy for 5 is 92.51700680272108
Accuracy for 7 is 92.51700680272108
Accuracy for 9 is 91.15646258503402
**********************************************************************
Accuracy after training the data(with cluster ID) and testing remaining 30 percent for 7 value  is  92.06349206349206
```