CENG 499 Homework 2 Report

Çağdaş Fil, 2093839

K-Nearest Neighbors Algorithm

What are the shortcomings of the Euclidean distance?

The problem with Euclidean distance is that it doesn't consider the similarity between attributes and it considers each attribute different from other attributes. Therefore, attributes with different scales may be problematic in Euclidean distance. Another drawback of the Euclidean distance is that it is not suitable for categorical values.

Why does the dataset trigger the shortcomings of the Euclidean distance?

In the given dataset, attribute values are scaled diversely. This situation triggers the shortcomings of the Euclidean distance.

How can the dataset be preprocessed so that the test set accuracy improves?

As stated above, attribute values are scaled diversely. Reducing the values in a common scale may improve the test set accuracy. I normalized the data with this normalization formula:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

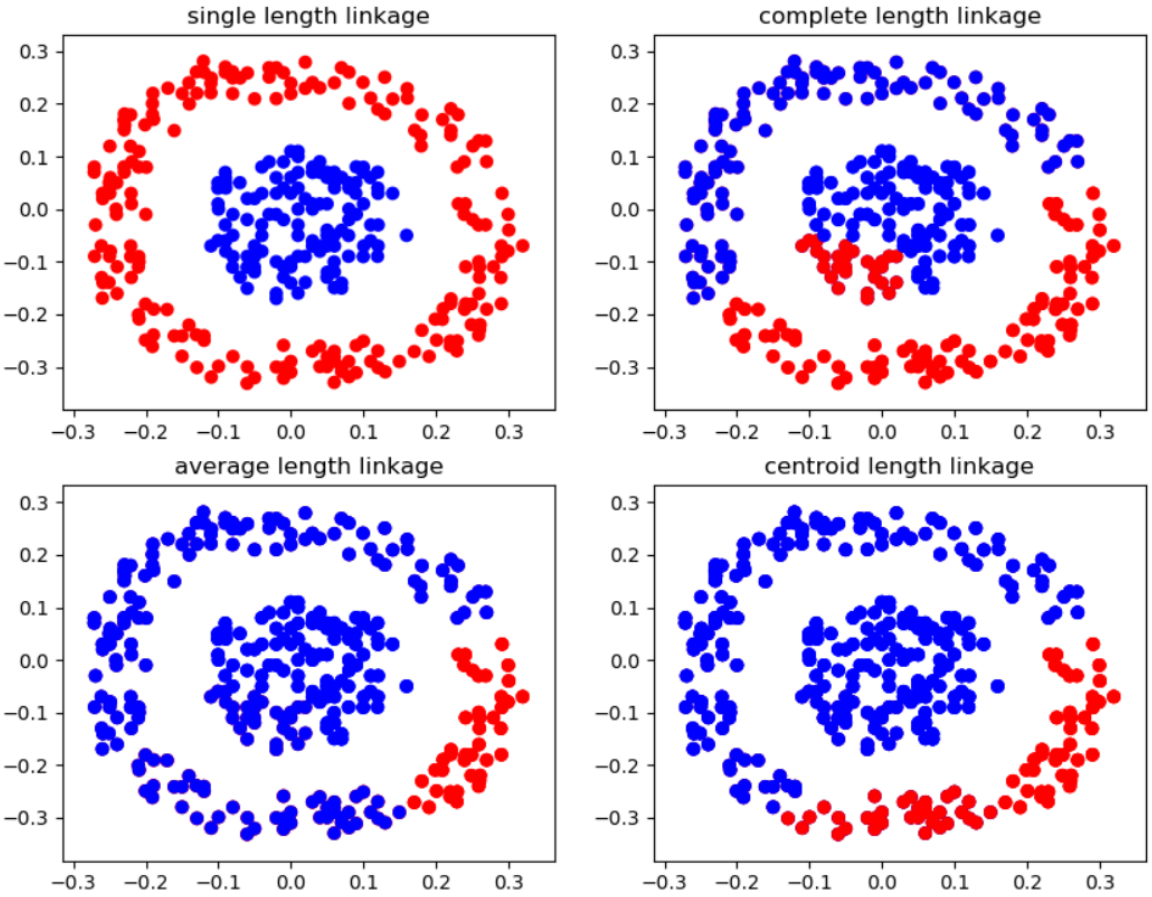
As a result, the test accuracy is improved with this preprocessing operation.

Note: *Implementation of normalization function and sample script for running provided as comment in* ***task1.py*** *file.*

Hierarchical Agglomerative Clustering

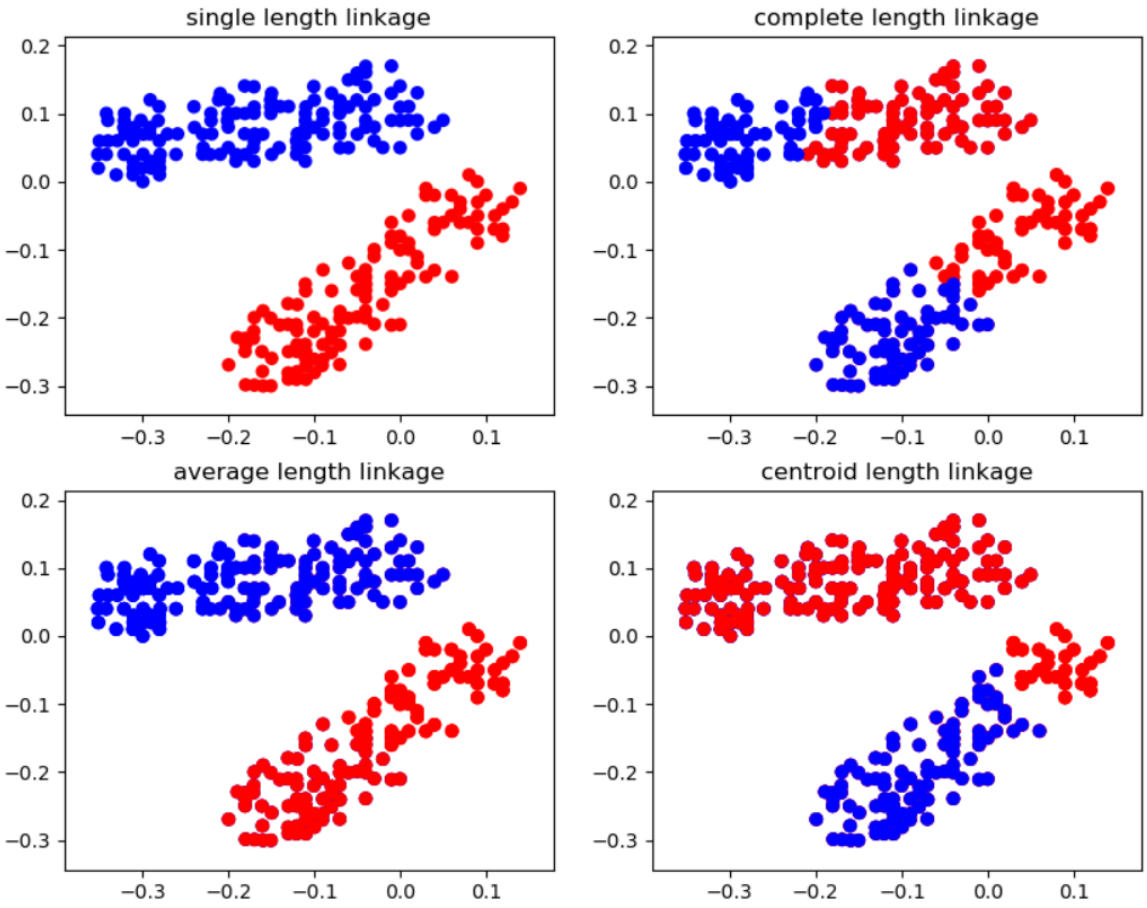
Dataset 1

For this dataset, simple length linkage criterion is the best option. Other criterions clusters the outer data with the inner data, but simple length linkage criterion separates them by considering the minimum distance between clusters.



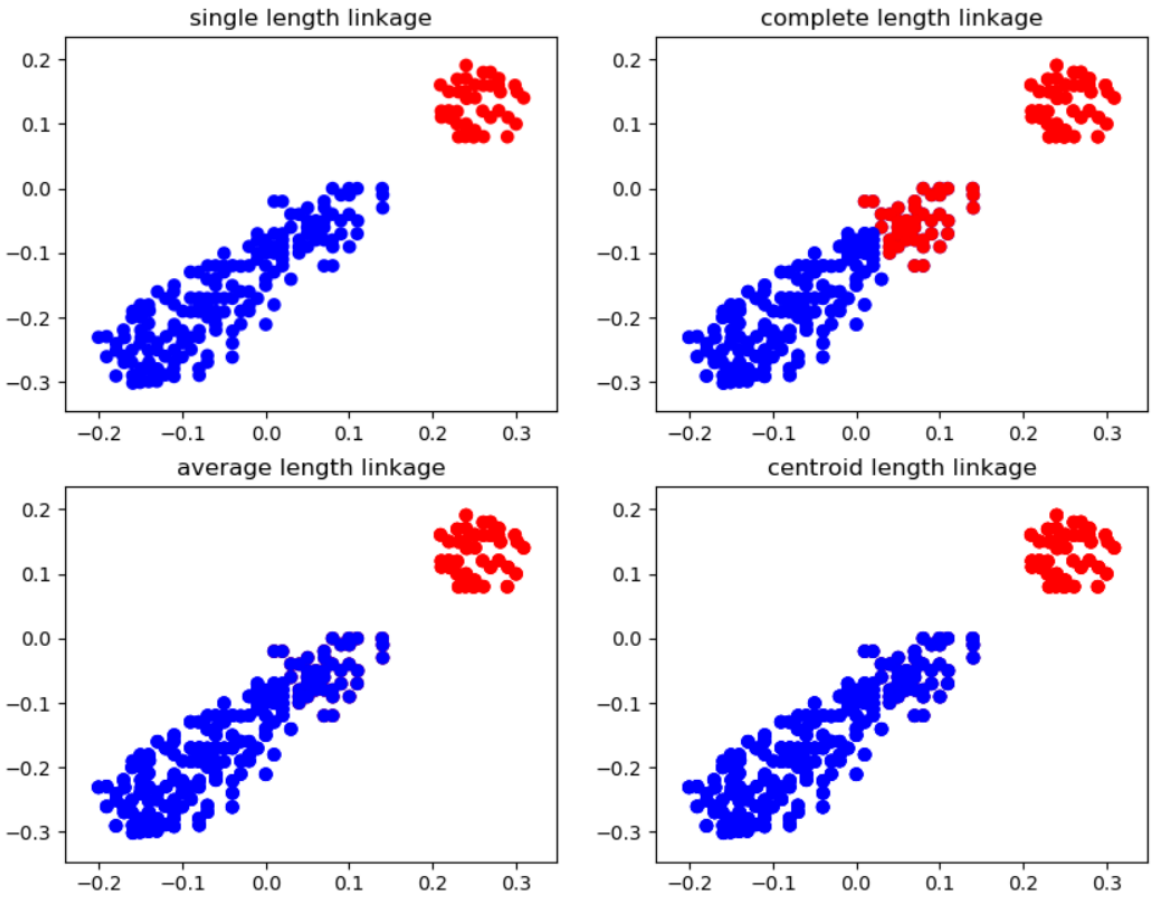
Dataset 2

For this dataset, simple length linkage and average length linkage criterions are the best choices. Centroid length linkage criterion behaves bad by taking some part of the other cluster. Complete length linkage criterion is the worst criterion by far. It took half of the both clusters to create cluster.



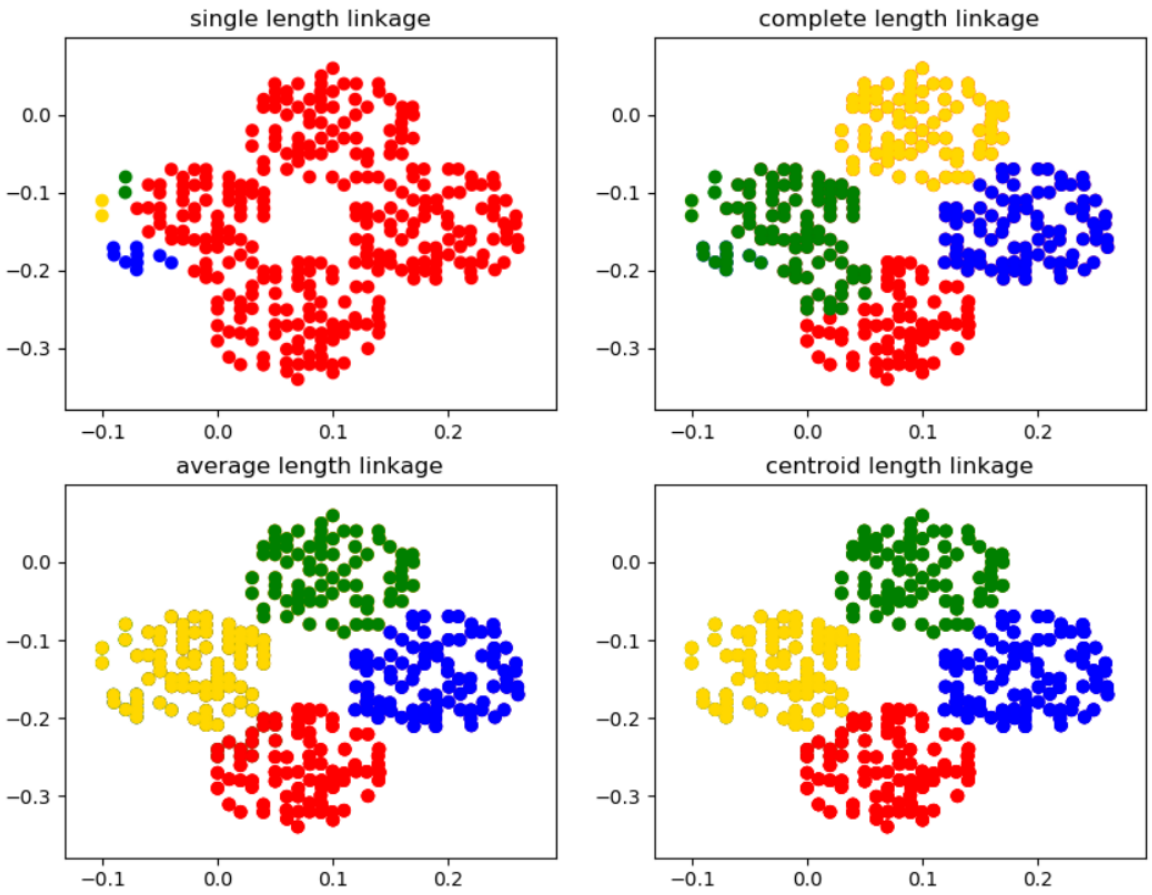
Dataset 3

For this dataset, simple length linkage, average length linkage and centroid length linkage criterions gives the best results because clusters are separated clearly. However, Complete length linkage criterion couldn't cluster the given dataset successfully.



Dataset 4

For this dataset, clusters are very close to the each other. Therefore, simple length linkage criterion gives the worst result. Complete length linkage criterion has some mistakes on clustering but still shows admissible performance. Average length linkage and centroid length linkage criterions are the best criterions on this dataset.



Decision Trees

Information Gain

This selection strategy worked well on the given dataset. It has high accuracy, but the created decision tree has a bit much nodes. Results are given below:

Accuracy: 93.12

Number of Nodes: 579

Gain Ratio

This selection strategy also worked well. Results are very similar to the "Information Gain" strategy, accuracy is high but number of nodes is a bit much. Results are given below:

Accuracy: 93.28

Number of Nodes: 586

Average Gini Index

This selection strategy ended up with worse results. Besides that, it has low accuracy, it also generates excessively big decision tree. Results are given below:

Accuracy: 50.88

Number of Nodes: 20509

Average Gini Index with Chi-squared Pre-pruning

I couldn't implement this part.

Average Gini Index with Reduced Error Post-pruning

I couldn't implement this part.