CS-550 - Spring 2023 HW 3: Clustering Algorithms

Salih Deniz Uzel (deniz.uzel@bilkent.edu.tr)
22201382

I. Introduction

In this homework 2 different type of clustering algorithms are implemented and analyzed. Computational time, clustering errors are computed. Corresponding clustered images are ploted for k = 2, 3, 4, 5, 6 values. Best number of cluster value picked by using elbow method. For the implementation python programming language and for the array operations scientific computing library numpy is used. Time measured by using time library of pyhton language.

A. Error Function and Evaluation

For both methods, Red Greeb Blue (RGB) values are divided by 255 and scaled between 0 and 1. Clustering error is calculated with Squared Error (SE) between the centroid values which is assigned to pixel and the pixel rgb values. Summation of the squared errors for all pixels divided by the number of pixel values to obtain average clustering error (CError). In the following formula P is pixel, C is assigned Centroid and R, G, and B is the color Values of the pixel and the centroid.

$$CError = \frac{1}{m}\sum_{i=1}^{m}((P_R - C_R)^2 + (P_G - C_G)^2 + (P_B - C_B)^2)$$

II. Section 1, k-means

In the k-means algorithm, K number of centroid initalized by randomly picking R, G and B values of the base image pixels. If the selected random centroid exists it is discarded and re-selected. For the initial centroid values each pixel of the image is assigned to its nearest centroids. Similarities are calculated by calculating the Euclidean distance between the pixel and the assigned centroid values. These centroid values are updated by the average of the R, G, and B values of each cluster. This process is called 1 epoch and continues until the maximum number of epochs is reached or the clustering error improvement is less than 1 %.

In the Table I, the error rate decreases as the number of clusters increases. This is the expected output because the base image contains more colors than the number of clusters given in the task. However, the Calculation Time is increased. As the number of clusters increases, the total distance calculations, center updates, label calculations increase. In the Fig. 1, although it is not very clear, around K=4 can be chosen as elbow point. As the K value increases, the improvement in the clustering performance can be observed in Fig. 2.

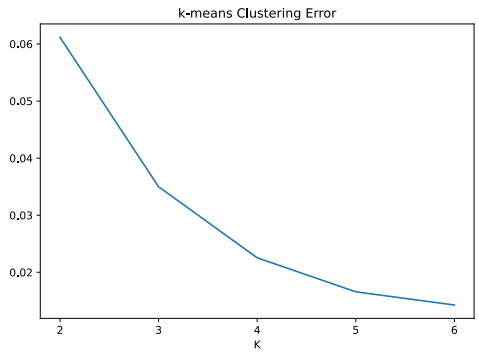

Fig. 1: HW Section 1. k-means Clustering Elbow Curve of Average Error for following values k = 2, 3, 4, 5, 6.

TABLE I: HW Section 1. k-means Clustering Error

K	Clustering Error $\times 10^{-2}$
2	6.12
3	3.5
4	2.25
5	1.66
6	1.43

TABLE II: HW Section 1. k-means Computational Time

K	Computational Time (s)
2	33.37
3	69.22
4	53.56
5	81.59
6	115.47

III. Section 2, Agglomerative Hierarchical Clustering

In the Agglomerative Hierarchical Clustering (AHC) method, number of clusters equal to the number of pixels at the initial state. The shortest distance between all clusters is calculated and the cluster pair with the shortest distance is combined. This process continues until K clusters remain. In this process, the method to be used in calculating the distance between clusters should be chosen. The **Centroid Method** method was chosen for this assignment in order to reduce the total computational cost. In this method, the mean of the R, G and B values of the pixels in the cluster is calculated and the distance between the clusters is found by calculating

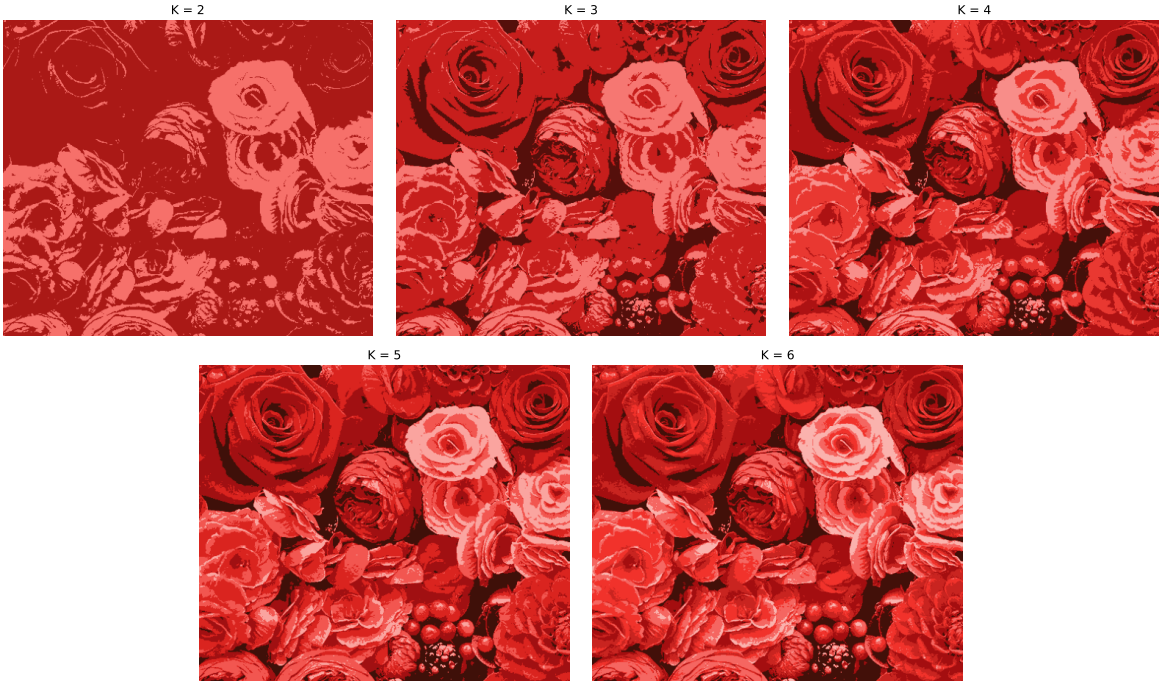

Fig. 2: HW Section 1. k-means clustering algorithm results for following values k = 2, 3, 4, 5, 6

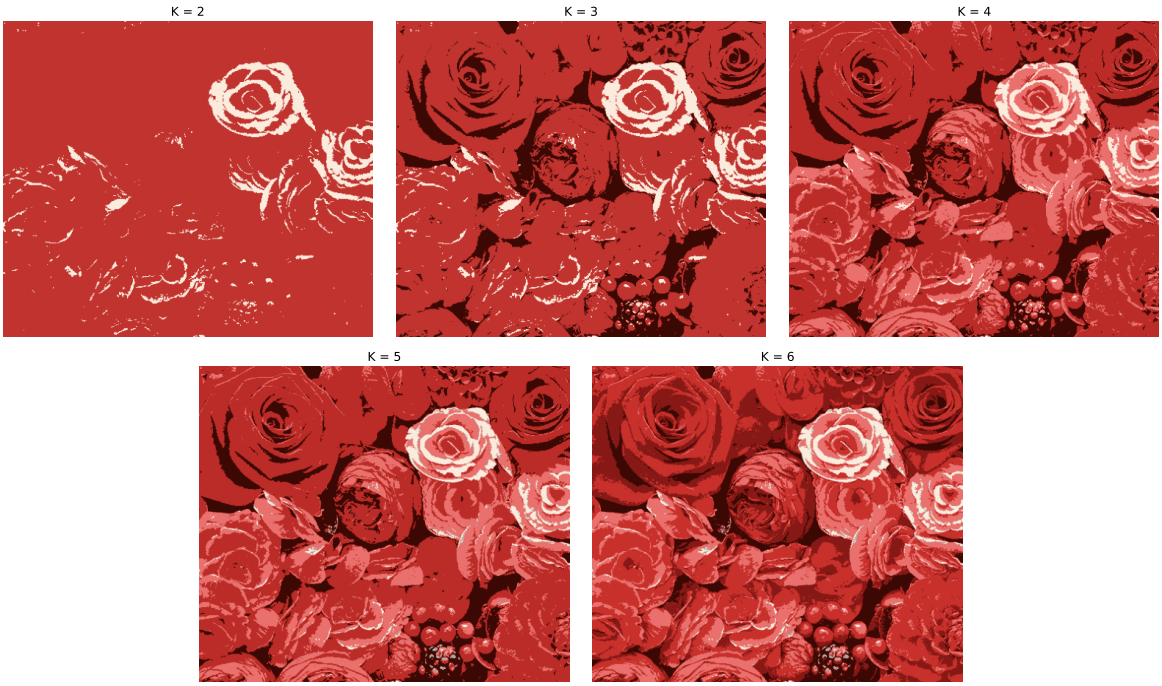

Fig. 3: HW Section 2. Agglomerative Hierarchical Clustering algorithm results for following values k = 2, 3, 4, 5, 6

the Euclidean distance between these mean values. The mean values of the clusters are re-calculated only for the clusters that change. This reduces the overall distance calculation cost. It is a computationally cheaper method than calculating all the distances between the pixels in the clusters, as in the Average Linkage method.

However, even this method is very costly for 435x510 pixels alone. As a solution to this, the **Sub-Sampling Method** was chosen. In the homework, 1000 pixels were randomly selected from the image and the AHC algorithm was run on these pixels. The centroids obtained as the output of the algorithm are used to label all the pixels of the base image. Each pixel is assigned to the centroid at the nearest Euclidean distance.

In this method, the same sub-samples were used for all K values. Clustering performance depends on how well the selected sub-sample represents the color distribution in the

image. Since the AHC method decreases from the number of sub-samples to the K dimension, the computation time is almost the same for all K values Table IV, unlike k-means algorithm. In addition, the amount of color in the base image is more than the K value, therefore the clustering error decreases with the increase of the K value III. In the Fig. 4, K=4 can be chosen as elbow point.

In the Fig. 3as the K value increases, the change in clustering performance can be observed in Fig. 2. Compared to Fig. 2, the clustering performance is lower in the AHC results are lower. This can be explained by how well the random sub-sampling performed at the beginning of the AHC algorithm represents the color distribution in the main image.

TABLE III: HW Section 1. Agglomerative Hierarchical Clustering Clustering Error

K	Clustering Error $\times 10^{-2}$
2	9.51
3	5.97
4	3.38
5	3.37
6	2.49

TABLE IV: HW Section 1. Agglomerative Hierarchical Clustering Computational Time

K	Computational Time (s)
2	38.17
3	37.79
4	37.94
5	37.66
6	37.84

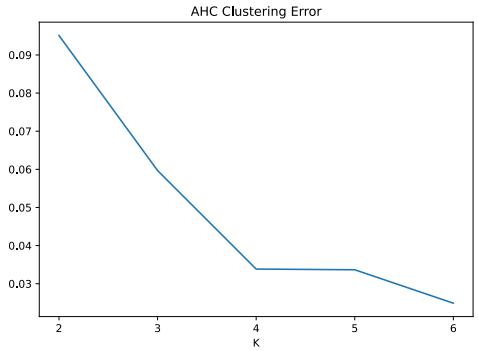

Fig. 4: HW Section 1. Agglomerative Hierarchical Clustering Elbow Curve of Average Error for following values k = 2, 3, 4, 5, 6.