Fuzzy C-Means Clustering and Cluster Quality Evaluation

GNR602- ADVANCED METHODS IN SATELLITE IMAGE PROCESSING

Introduction

- Fuzzy C-Means (FCM) is a clustering algorithm that assigns each data point to a cluster with a degree of membership.
- The primary objective of this project is to implement the FCM clustering algorithm and evaluate the cluster quality for different values of C (number of clusters).
- m is a fuzziness parameter that controls the degree of overlap between the clusters.
- The minimum and maximum values of C will be specified by the user.

Introduction

- In this project, we will first discuss the theoretical background of the FCM algorithm and its various parameters.
- Next, we will implement the algorithm using Python programming language
- We will then evaluate the cluster quality for different values of C using Davies-Bouldin Index
- The results of this project will provide insights into the effect of the number of clusters on the quality of clusters obtained by the FCM algorithm.

Problem

- Label each pixel of the image into clusters based on a fuzzy membership of that pixel into the cluster

The objective of the FCM algorithm is to minimize the following objective function:

$$J = \sum_i \sum_j w_{ij}^m ||x_i - c_j||^2$$

J is the objective function to be minimized.
w_{ij} is the degree of membership of data point x_i in cluster c_j.
m is a fuzziness parameter that controls the degree of overlap between the clusters.
x_i is the i-th pixel's feature vector.
c_j is the centroid of the j-th cluster.

Algorithm

Fix	Initialize	Update	Evaluate	Repeat	Determine
Fix a C in the range of C specified by the user	Initialize the membership matrix via K-Means (K = C)	Iteratively update the membership and cluster centroids until convergence	Calculate the Davies Bouldin Index (DBI) for the clustering	Repeat these steps for all values of C specified by user	Determine the optimal value of C by the minimum DBI value

Initialize

Initialize the cluster memberships by running a few iterations (≈10) of the K-means algorithm (K = C).

Initialize the fuzzy memberships from the K Mean cluster vectors according to:

$$w_{ij} = \frac{\frac{1}{d(x_i, c_j)}}{\sum_{p=1}^{K} \frac{1}{d(x_i, c_p)}}$$

w_{ij} is the fuzzy membership of pixel i into cluster j
$d(x_i, c_j)$ is the Euclidean distance between feature vector of pixel i from mean vector of cluster j

$$\sum_{j=1}^{K} w_{ij} = 1$$

Update

Update memberships and cluster centres in each iteration as:

$$\forall_{i,j} \quad w_{ij} = \left[\sum_{k=1}^{C} \left(\frac{||x_i - c_j||}{||x_i - c_k||} \right)^{\frac{2}{m-1}} \right]^{-1}$$

$$\forall_j \quad c_j = \frac{\sum_i w_{ij}^m x_i}{\sum_i w_{ij}^m}$$

Evaluate Cluster Quality

- The Davies Bouldin Score (DBI) is a measure of cluster quality that evaluates the ratio of within-cluster similarity to between-cluster dissimilarity.
- The DBI is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters which are farther apart and less dispersed will result in a better score.
- A lower DBI score indicates better cluster quality, as it means that the clusters are more separated and have a higher degree of similarity within the clusters.

Silhouette Score

- The Silhouette Score is calculated for each data point in the dataset by computing the mean intra-cluster distance (a) and the mean nearest-cluster distance (b), and then subtracting them and dividing by the maximum of the two (max(a, b)).
- The Silhouette Score ranges from -1 to 1, where a score closer to 1 indicates that the data point is well-matched to its assigned cluster and that the clusters are well-separated from each other, and a score closer to -1 indicates that the data point may be better matched to a different cluster.
- But this has high computational time complexity (more than the clustering algorithm itself). So, we didn't use this quality metric

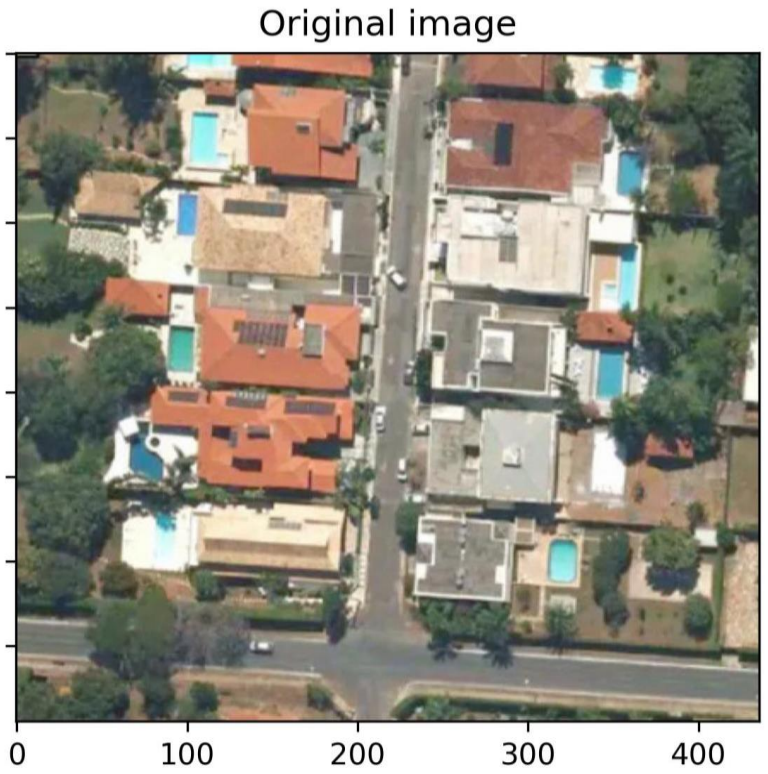
Repeat & Determine

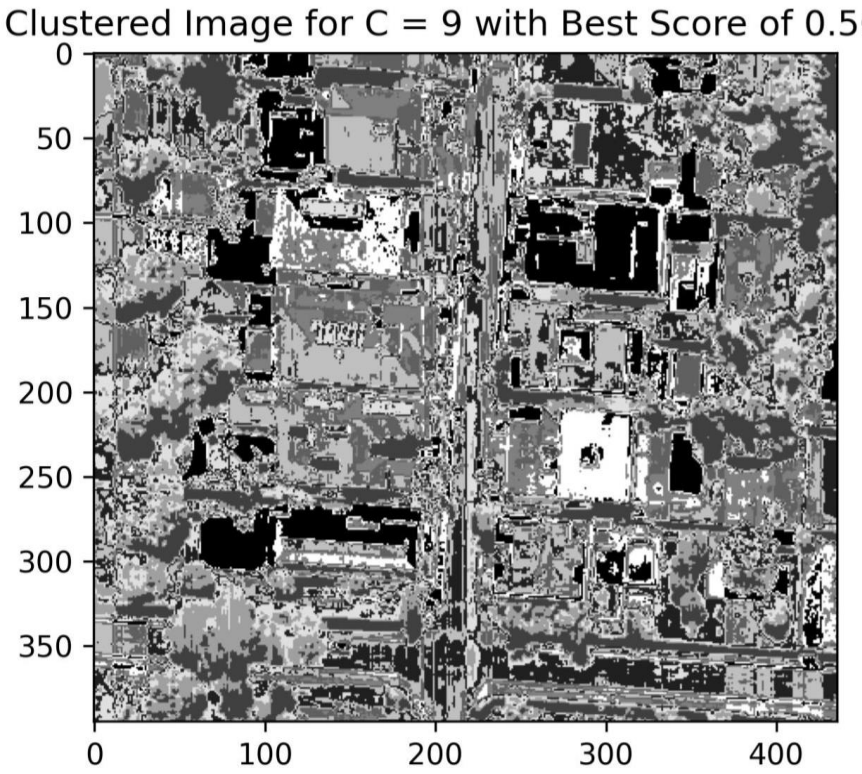
- Repeat the previous steps for all values in the range of C
- The optimal value of C is determined by the clustering which has the minimum Davies Bouldin Score

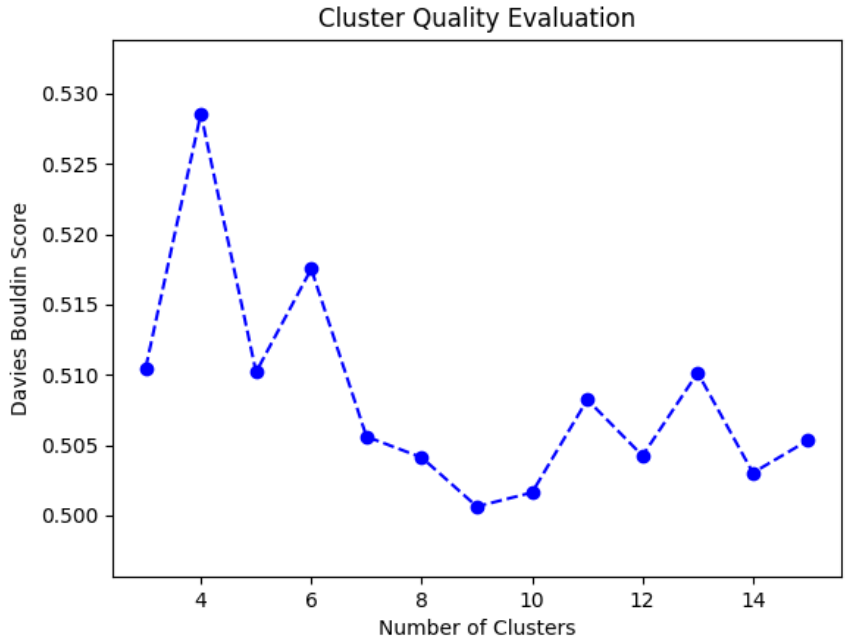
Dataset

Louisiana 2016 Flood Satellite Images
https://www.kaggle.com/datasets/rahultp97/louisiana-flood-2016

Results


Original image
0
100
200
300
400

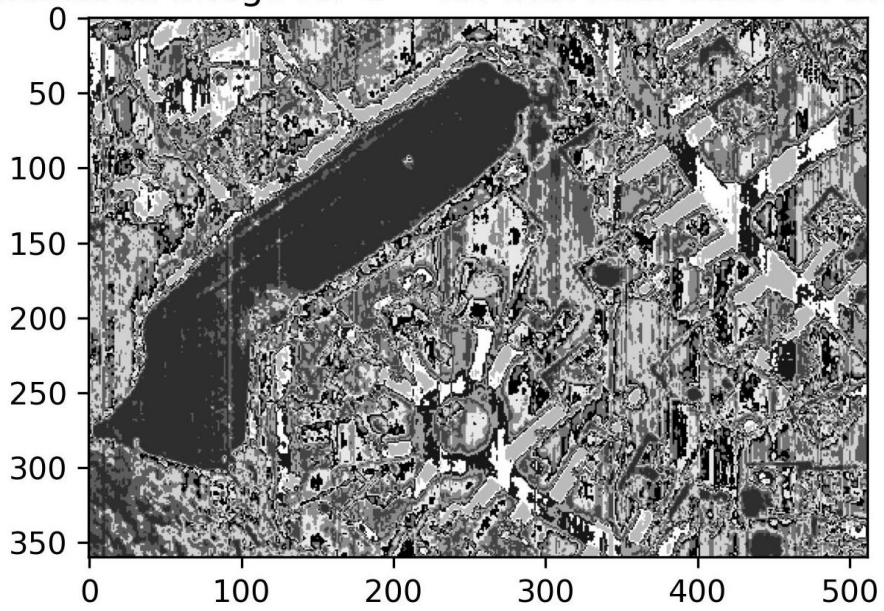

Clustered Image for C = 9 with Best Score of 0.5006
0
50
100
150
200
250
300
350
0
100
200
300
400

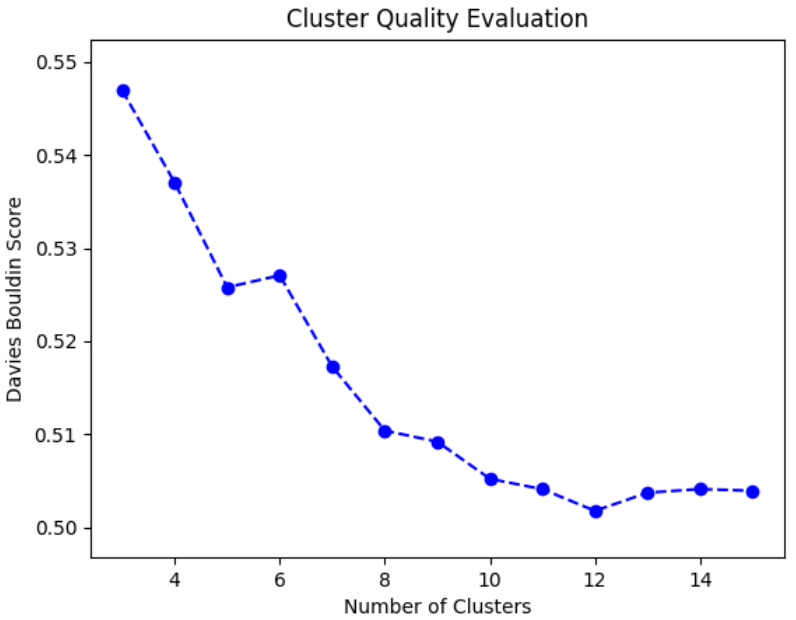

Cluster Quality Evaluation
Davies Bouldin Score
0.530
0.525
0.520
0.515
0.510
0.505
0.500
4
6
8
10
12
14
Number of Clusters

Results

Original image

Clustered Image for C = 12 with Best Score of 0.5018

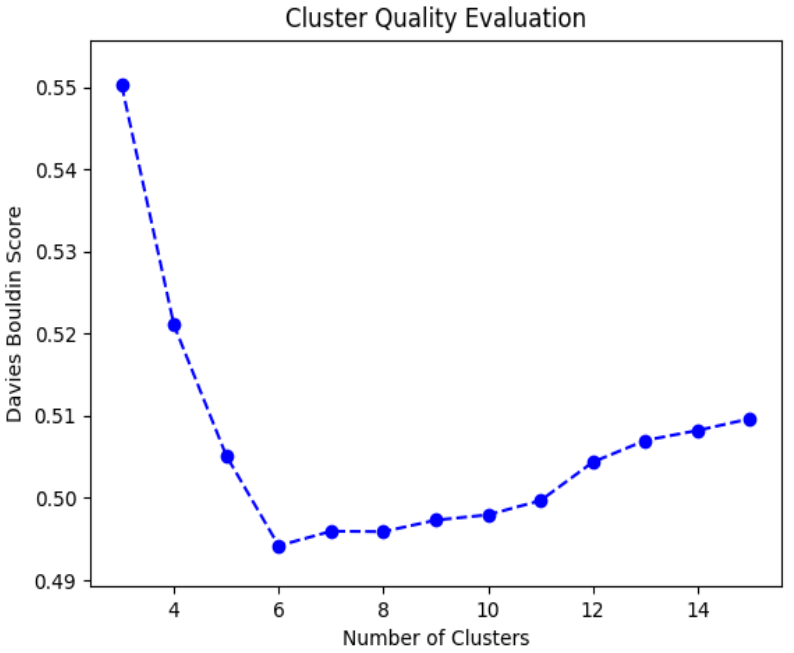




Results

Original image

Clustered Image for C = 6 with Best Score of 0.4942



Conclusion

- A low number of clusters can result in poor cluster quality because the clusters may be too large and heterogeneous, while a high number of clusters can result in poor cluster quality because the clusters may be too small and similar.
- Therefore, it is crucial to find the optimal number of clusters, which lies in the intermediate range and results in well-separated, homogeneous, and meaningful clusters.
- An optimal number of clusters are needed to model various aspects captured in a satellite image such as forests, roads, buildings, water bodies, vegetation, fields, etc.

Thankyou

Shantanu Welling (210010076)

Harshit Agarwal (210020054)

Arijit Saha (210050017)

Topic 3. Implement FCM clustering algorithm and evaluate the cluster quality for different values of ‘C’. (Minimum and maximum values of C to be specified by user)