

Analyze the clusters and common properties found for each cluster

Applying K-means clustering (k=10) to fruit dataset.

Abhishek Kumar
Machine Learning Engineer Intern
21 August, 2019
abhishekpro47@gmail.com
Bihar, India
AI Tech System
www.ai-techsystems.com

Abstract—The relationship among the large amount of data has become a hot research topic. It is desirable to have clustering methods to group similar data together so that, when a lot of data is needed, all data are easily found in close proximity to some search result. Here we have done preprocessing of fruits and vegetables images and used a popular method, k-means clustering, to create 10 clusters. Then we analyze the clusters and common properties found for each cluster.

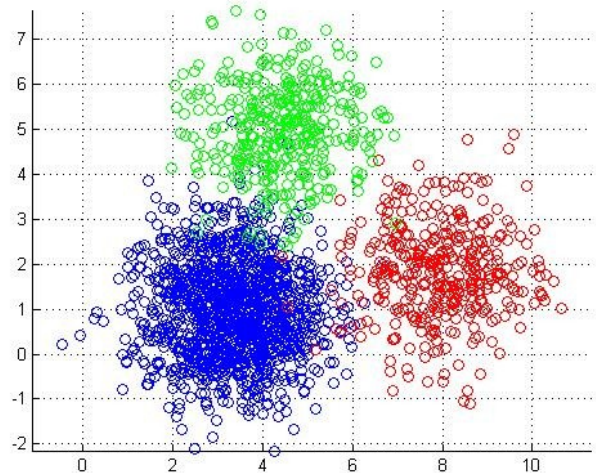
I. INTRODUCTION

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

An easy abstraction for clustering data is based on a proximity relationship. Data that are close to each other tend to share some external relationship. This relationship can be established to group the data into clusters.

II. ALGORITHM DESCRIPTIONS

The k-means clustering algorithm is one of the popular data clustering approaches. The k-means clustering algorithm receives as input a set of points and the number k of desired centers or cluster representatives. With this input, the algorithm then gives as output a set of point sets such that each set of points have a defined center that they “belong to” that minimizes the distance to a center for all the possible choices of each set.



The algorithm works as follows:

First we initialize k points, called means, randomly. We categorize each item to its closest mean, and we update the mean's coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and at the end, we have our clusters. The K-means algorithm defined above aims at minimizing an objective function, which in this case is the squared error function.

The objective function for the K-means clustering algorithm is the squared error function:

$$J = \sum_{i=1}^k \sum_{j=1}^n (||x_i - v_j||)^2 = 1$$

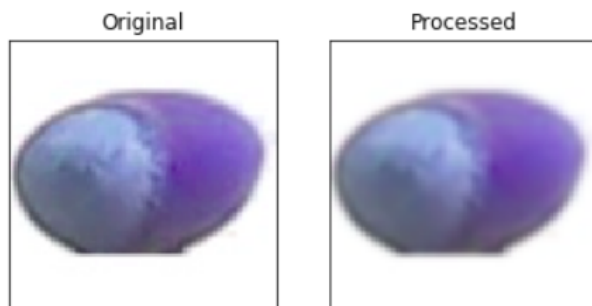
where,
 $||x_i - v_j||$ the distance between a point, x_i and the j th cluster center v_j over all k points in the i th cluster.

III. DATA PREPROCESSING

Before applying K-means to our data set we will do some pre-processing so that K-means will give best result. Our data set is consist of images of 114 varieties of fruits and vegetables.

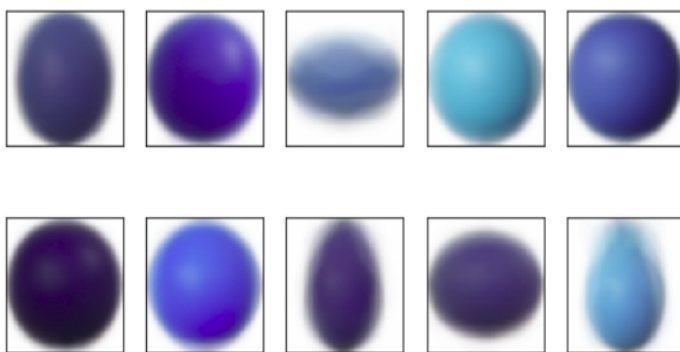
A. Removing Noise

For removing noise from the image we used Gaussian blur. Gaussian blur (also known as Gaussian smoothing) is the result of blurring an image by a Gaussian function. It enhances image structures at different scales.



IV. RESULT

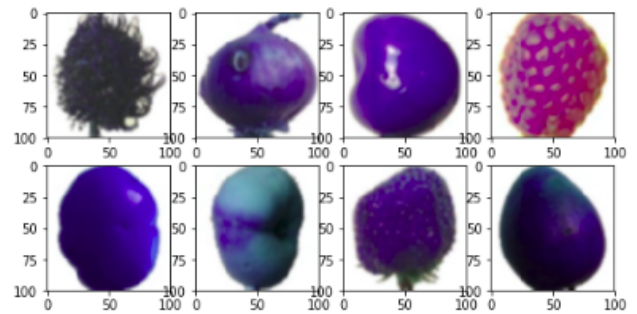
After applying Kmeans with $k=10$ to our data set, we get 10 centers. Images at the centers were.



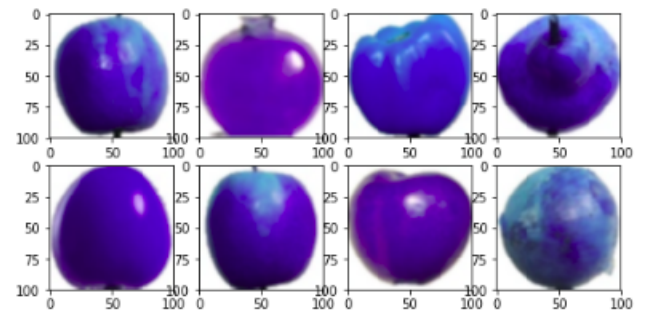
It can be clearly seen that, it has done the clustering based on the shape and colors.

Let's visualize the some images of each clusters.

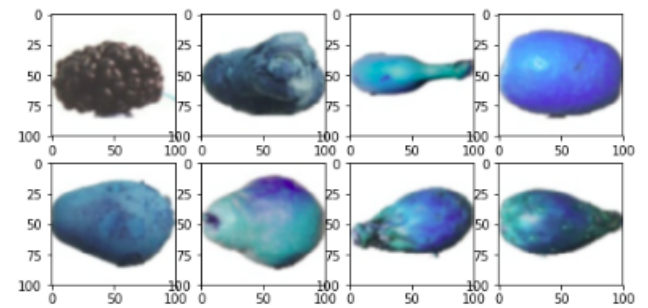
There are 1787 members in cluster 1



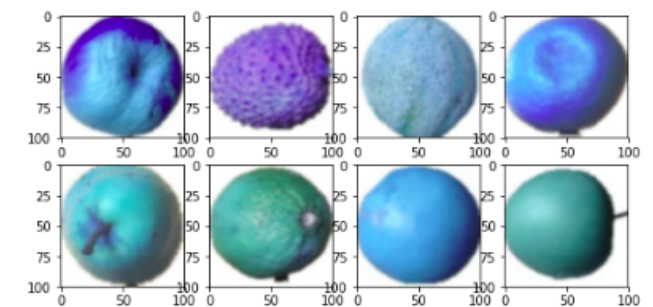
There are 1939 members in cluster 2



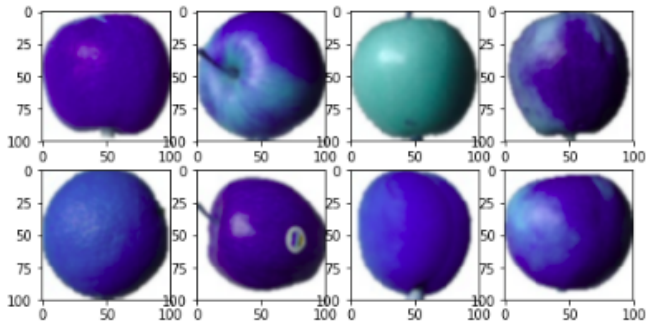
There are 1303 members in cluster 3



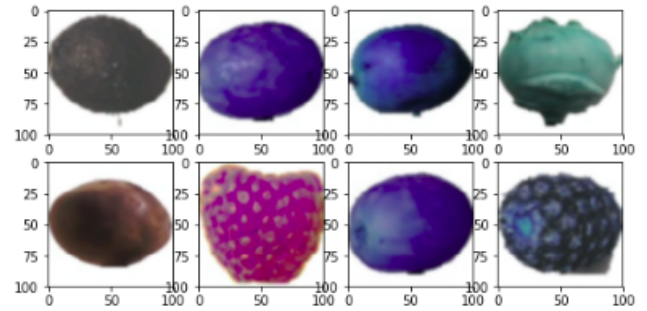
There are 2211 members in cluster 4



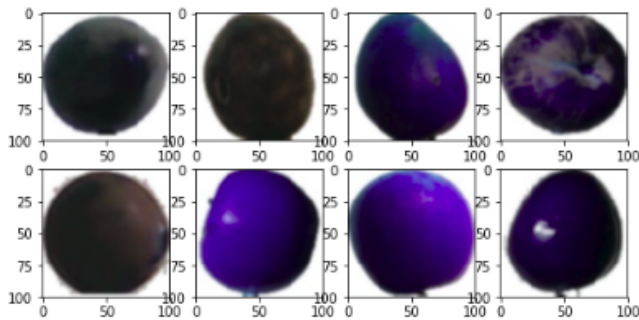
There are 1973 members in cluster 5



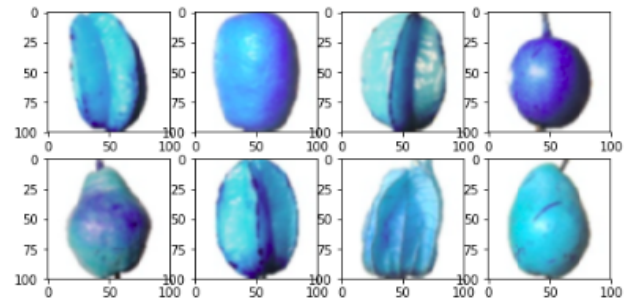
There are 2211 members in cluster 9



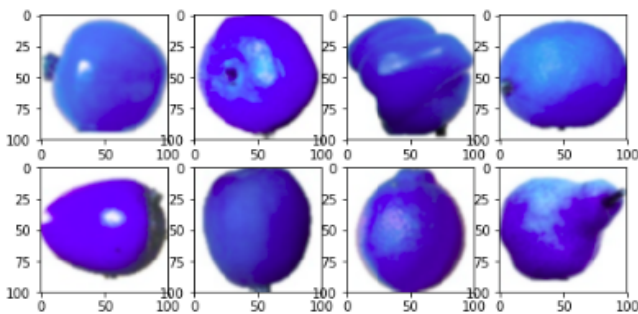
There are 2052 members in cluster 6



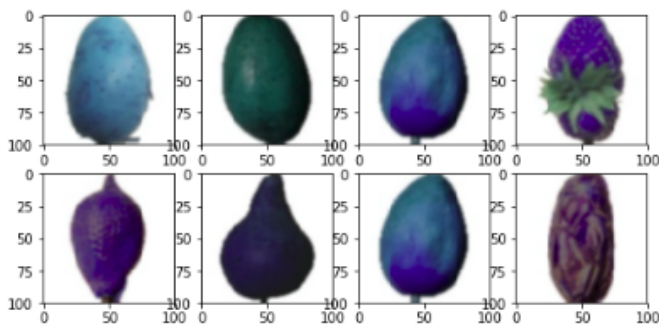
There are 844 members in cluster 10



There are 1789 members in cluster 7



There are 1140 members in cluster 8



V. CONCLUSION

We can further improve the clustering:

- Choosing optimal value of K(number of clusters).++a
- Merging neighboring clusters if the resulting cluster's variance is below the threshold
- Running the algorithm using different initializations of centroids.
- Standardize the data