Image Dataset clustering using K-Means Clustering (Unsupervised Learning)

Somesh Sunariwal
Someshsunariwal@gmail.com
Rajasthan, India
AI Tech System
www.ai-techsystems.com

Abstract **– K-Means clustering is an unsupervised machine learning algorithm which is used in image segmentation or clustering the data which have the similar properties. In this K-Means clustering unsupervised machine learning project I clustered the images based on their shapes and colors. To start this project I took the dataset from kaggle.com which has 114 classes of images and each class contain approx 400 images. This dataset have 100x100x3 pixel images. K-means clustering algorithm helped to cluster images. I used the n_clusters = 10 to achieve 10 clusters in this image dataset.**

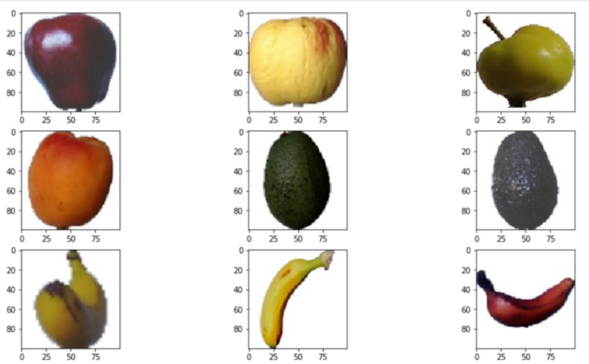
***Keywords*—Machine Learning, K-Means**

I. INTRODUCTION

K-Means clustering is a popular algorithm for unlabelled data. In order to achieve the great result from K-Mean select the value of k precisely. It generates clusters of the similar data points in datasets. In this K-Means project I set the k value to 10 in order to get 10 different clusters which hold the same properties. To perform the clustering on images I got the dataset from the Kaggle.com[1]. This dataset consist of 57,276 images of fruits and these images are divided into 114 classes of fruits and each class contain approx. 400 images. Each image shape is 100x100 pixels with 3 colour channel. For this project I used this dataset without label because it is unsupervised learning. K-Means was applied to entire dataset with 10 clusters which cluster the entire dataset images into 10 clusters. Result shows that the entire dataset images were clustered based on their shape of fruits and the colour of fruits that will I show you later in report.

II. DATA PRE-PROCESSING

Images are the representation of array of pixels which have the values between 0-1. Colour images have 3 dimensions and each dimension have the array of pixel which holds the value between 0-255. The dataset which I got from kaggle have the colour images which have 3 dimensions.



(Fig 1:- Images of dataset)

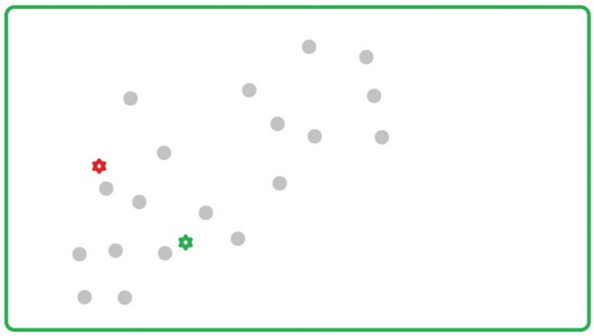
K-Means works on 2 dimension dataset. In order to apply K-Means I converted this 3 dimension data into 2 dimensions by multiplying the image height, weight and shape that gave me the shape of (19121, 30000). To reduce the time of computing I took the 1/3rd images from each class and these extracted images have all type of possible rotation. So model was able to produce precise result.

III. MODEL

K-Means:- This is unsupervised algorithm which process on n values and divide into k clusters in which each cluster is belong to observed nearest mean values. Cluster is a collection of similar features points. The goal of K-Means algorithm is to find groups in data points. This algorithm creates the clusters based on the features that are provided into dataset. K-Means algorithm work on following steps which are:

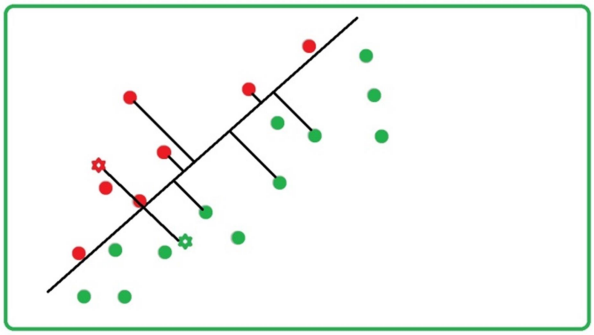
Step 1: First step is choosing the number of k clusters which is important step of K-Means. First find the no of clusters in a dataset by either randomly or using the elbow method and find best value of k.

Step 2: Second step is to select random k points that become a centroid of dataset



(**Fig 2**:- Random k points)

Step 3: At this point assign each data point to the



(**Fig 2:-** Assign each data point to cluster)

closest centroids that form k clusters by placing a perpendicular line between two centroids so the Euclidean distance from every point on line is same from the centroid.

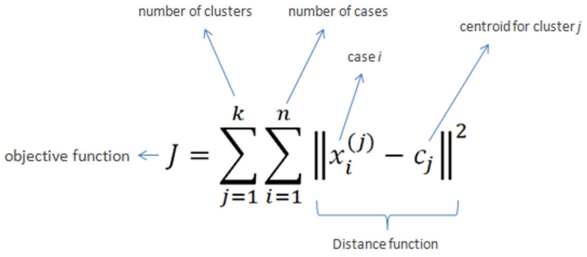
Euclidean Distance: Euclidean Distance formula is given by:

$$(|X_1-Y_1|^2 + |X_2-Y_2|^2 + \ldots + |X_{N-1}-Y_{N-1}|^2 + |X_N-Y_N|^2)^{1/2}$$

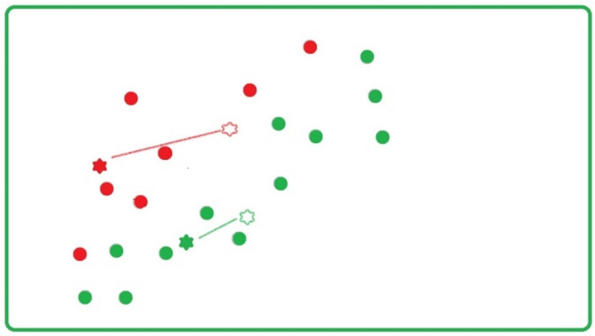
Where X represent the first data point and Y represent the second data point. N is the number of attributes.

Starting from first randomly generated cluster centers in data set, each data point of the dataset is assigned to the nearest cluster, after which each center is updated.

Step 3: Compute and place the new centroid of each cluster [2].

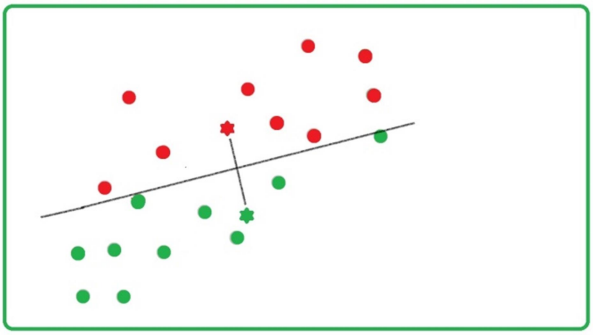


(**Fig 3**:- Computing new centroid using formula)



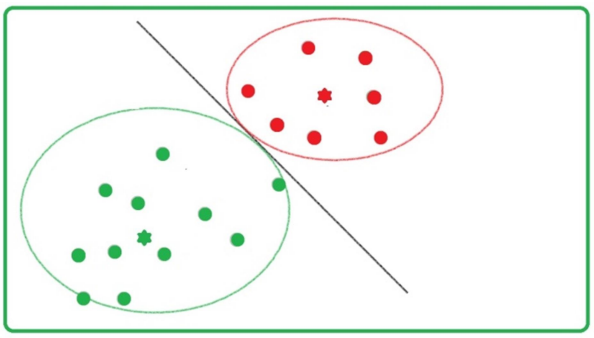
(**Fig 4**:- New Centroids)

Step 4: Reassign each data point to new clusters using Euclidean distance.



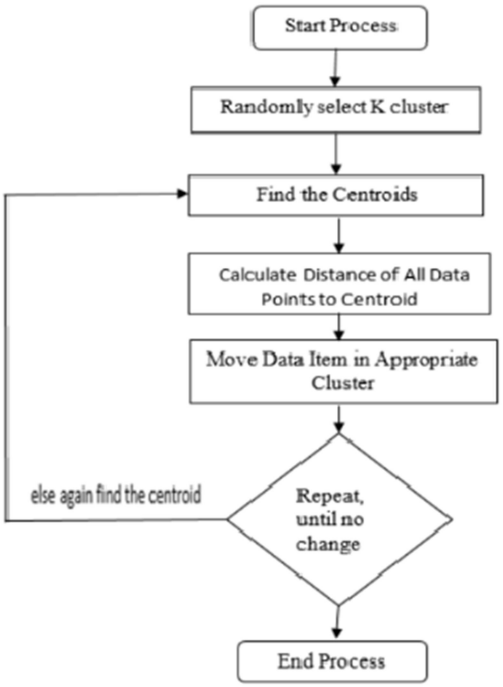
(**Fig 5**:- Reassigning new data points to new centroids)

Step 5: After assign data points to new centroid again compute new centroids if no new centroid take place means clustering process is done.



(**Fig 6**: End clustered result)

Flow Chart: This is a full work process of K-Means Clustering unsupervised machine learning algorithm [3].



(Fig 7: K-Means flow chart**)**

In case of images K-Means work differently. K-Means generate the number of colour patches according to the number of k.

(**Fig 7**: K-Means clustering on images with different k values)

K-Means model is used from sklearn library. Entire dataset was passed through the K-Means model with value of k=10. Value of k creates 10 clusters finding

similarity in dataset. This K-Means model produces the clusters centroid and labels for the image dataset. Clusters centroid shape is (10, 30000) and labels shape is (19121,) respectively. Clusters and labels are help to reproduce images which are clustered on the basis on their shapes and colours.

IV. LIMITATION OF CLUSTERING

K-Means clustering algorithm has some limitation and these are:

1. In this clustering method number of clusters (k) needs to be determined beforehand.
2. This algorithm is sensitive to an initial centroid selection. Due to which initial centroid points it is susceptible to a local optimum and may miss the global optimum. It may converge to suboptimal solution [4].
3. It can model such a way that it can cluster the data point in spherical shape. Thus the non-convex shape of clusters cannot be modelled in center based clustering.
4. It has memory issues. It cannot be used where clustering problem does not fit in main memory.

V. CLUSTERS PROPERTIES

Produced images are clustered on the basis of their shapes and colours are shown below:

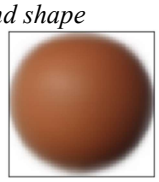
1. *Brown Oval Shape*

(**Fig 8**: - image 1)

2. *Brown Rotated Oval Shape*

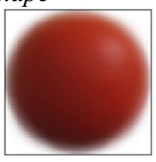
(**Fig 9**: - image 2)

3. *Brown Round shape*



(**Fig 10**: - image 3)

4. *Red Round shape*

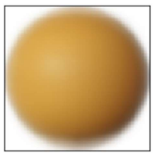


(**Fig 11**: - image 4)

5. *Yellow Oval shape*

(**Fig 12**: - image 5)

6. *Yellow Round shape*

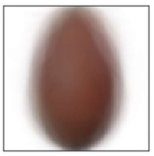


(**Fig 13**: - image 6)

7. *Darkest Red round*

(**Fig 14**: - image 7)

8. *Brown Long Oval*

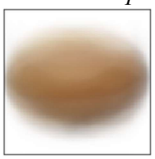


(**Fig 15**: - image 8)

9. *Dark Yellow Round shape*

(**Fig 16**: - image 9)

10. *Yellow Rotated Oval shape*



(**Fig 17**: - image 10)

Above result clearly shows that K-Means algorithm clusters the dataset images on the basis of shapes and colours.

VI. CONCLUSION

This paper shows the result when K-Means clustering applied to entire dataset. And as I discussed earlier the K-Means clustering algorithm is unsupervised learning algorithm. It separates the data according to the similarity in data points.

(**Fig 18**:- Result Image after applying K-Means)

And the above resulted image shows that the dataset images are clustered based on their shapes and colours.

VII. REFERENCES

[1]. Kaggle:https://www.kaggle.com/moltean/fruits

[2].https://www.researchgate.net/publication/271616608_A_Clustering_Method_Based_on_K-Means_Algorithm

[3].https://www.researchgate.net/publication/321133921_Image_Segmentation_using_K-means_clustering_and_Thresholding

[4]. International Journal of Engineering Research & Technology (IJERT)
Vol. 2 Issue 7, July - 2013