IMAGE CLUSTERING USING K MEANS ALGORITHM

Anmol Gulati
AITS
7 th August, 2019
Delhi, India
gulati.anmol10@gmail.com

***Abstract*- The objective of this report is to cluster images.Image clustering is the task of clustering same type of images in a single cluster .K means clustering algorithm along with Incremental PCA is used to cluster the large dataset of images.The dataset is the fruits dataset that is taken from kaggle.The dataset consists of large amount of fruit images.The images of the dataset are clustered into 10 different clusters.**

***Keywords*—K means algorithm, Incremental PCA, Normalisation.**

I. Introduction

Image clustering has become an important part in today's time because of the availability of large number of untitled image from various places.The main aim of image clustering is to cluster similar types of unlabelled images and provide them with labels which can be used for future.

K means is the technique for clustering similar data in one cluster. K means is an unsupervised learning technique in which the data is unlabelled. In this technique data is clustered together based on the similarity bw them.in this technique each datapoint is assigned to a particular cluster .The data point is assigned to that cluster to which its properties match significantly.Euclidean distance is used as the measure for assigning a data point to the cluster .K means technique can be used to provide label to the unlabelled data by labelling the data point corresponding to the cluster number it belongs to.

There are two steps involved in image clustering 1.) data preprocessing and 2.) feeding preprocessed data to algoithm.In first step the data is cleaned i.e. normalised or augmented(in case of images).
Data cleaning step is followed by the second step i.e. passing the cleaned data to the algorithm for clustering purpose.

In this report we are performing image clustering on fruits dataset.In this we have around 58,000 images
And we tried to group them i.e cluster them in 10 groups using K means algorithm.

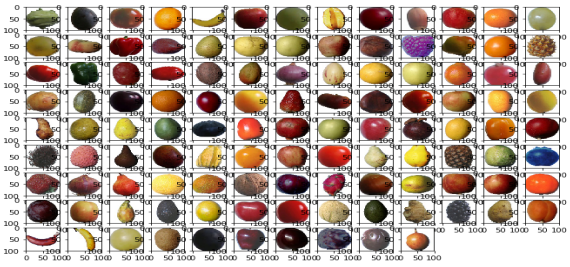

Fig 1. Each type of fruit in the dataset

II. Methodology

Image collection,normalisation of image dataset and application of pca are discussed in this methodology whereas the algorithm used is discussed in the next section.

A. Image collection

This is the initial step of the project where we load the images of the Fruits from the kaggle site on which we have to perform the clustering.

B. Normalisation

Normalisation is the image pre processing step in which we change the range of the values of pixels. This is important because some of the images are not very good because of the glayer. This helps in improving the accuracy of the model.

Inew=(I-I.min)(newmax-newmin)/(I.max-I.min)+ne wmin*

C. Incremental PCA

PCA is the unsupervised technique. PCA is used to reduce the dimensions of the data.It finds the maximum variance along the axis and ignore the dimensions which has very low variances.It uses the concept of eigenvalues and eigen vectors.Eigen vector points in the direction of maximum variance and eigen value represents the value of variance in that direction. It helps in reducing the dimensions of the data with large margins with very little loss in the information of the data.PCA helps in reducing the training time by reducing the number of features to be processed.

When the dataset is large the normal PCA cannot be used as it requires large memory.For this purpose we use Incremental PCA .In incremental PCA we do not send the entire dataset at one go ,we send the data in batches ,this approach helps in saving memory and fitting the data to desired components.

III. Dataset Sources

The fruits dataset is taken from the kaggle site.This dataset contains around 58000 images of fruits.The clustering is done on these images to arrange the images in 10 different clusters.

IV .CLUSTERING ALGORITHM

A. K Means

K means is the technique for clustering similar data in one cluster. K means is an unsupervised learning technique in which the data is unlabelled. In this technique data is clustered together based on the similarity bw them.in this technique each datapoint is assigned to a particular cluster .The data point is assigned to that cluster to which its properties match significantly.Euclidean distance is used as the measure for assigning a data point to the cluster .K means technique can be used to provide label to the unlabelled data by labelling the data point corresponding to the cluster number it belongs to.

There are two steps involved in image clustering

1) image preprocessing

In image preprocessing the image is first normalised in a particular range.After normalisation PCA is applied on the image to reduce its dimensions.

2) Model Fitting

After the image is preprocessed the image is passed to the model for fitting.In this fitting process each data point i.e.image is assigned to the cluster whose centroid is close to the image.

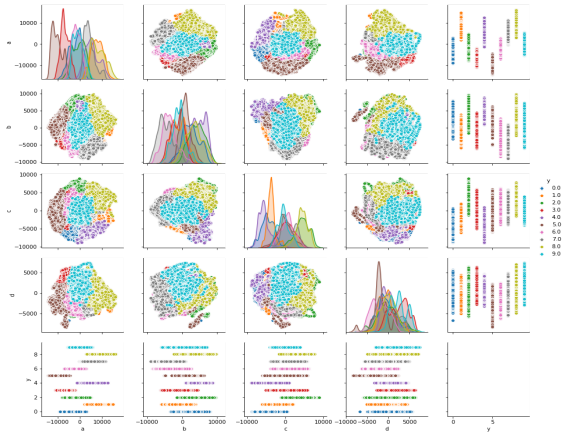

Fig 2. Cluster Formation

Fig 2 shows the formation of clusters through pairplot of the values 4 pixels.Here each point in cluster

represents the image belonging to that cluster,all the points of the same color belong to the same cluster.

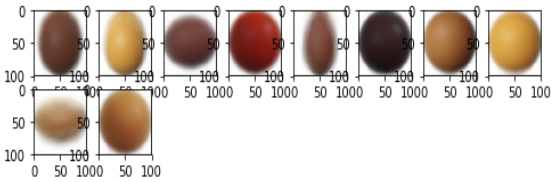

Fig 3 image of centroid of clusters

Fig 3 shows the images that corresponds to the cluster centers.It shows that what type of image is clustered in which cluster based on the shape and color of the fruit in it.The clusters contain images of fruit of specific shape and color only.

CONCLUSION

All the images are clustered in the 10 clusters on the basis of their shape and color.Each cluster contains images of fruit of specific shape only.

the number of images in 0 cluster is = 5918
the number of images in 1 cluster is = 4571
the number of images in 2 cluster is = 6994
the number of images in 3 cluster is = 5509
the number of images in 4 cluster is = 3354
the number of images in 5 cluster is = 5604
the number of images in 6 cluster is = 7076
the number of images in 7 cluster is = 8245
the number of images in 8 cluster is = 4341
the number of images in 9 cluster is = 5664