USING K-MEANS ALGORITHM FOR IMAGE CLUSTERING

Arshdeep Singh

Machine Learning intern

AI Technology and systems

171210014@nitdelhi.ac.in

www.ai-techsystems.com

Abstract—**In this paper we discuss about applying k means clustering on a high quality fruits image data set taken from kaggle. The goal is to find a mapping of the archive images into classes (clusters) such that the set of classes provide essentially the same information about the image archive as the entire image-set collection. The data set contains large no. of images which are clustered into 10 parts (k = 10) using k means clustering algorithm where features are reduced using principal component analysis.**

Keywords — K means clustering, principal Component analysis (PCA)

I. Introduction

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other clusters. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Here we go through the methodology and a brief working of k means clustering algorithm and principal component analysis (PCA).

II. K MEANS CLUSTERING

We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the k Means algorithm; an unsupervised learning algorithm.

The algorithm works as follows:

1. First we initialize k points, called means, randomly.

2. 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.

3. We repeat the process for a given number of iterations and at the end, we have our clusters.

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{\operatorname{argmin}} \, dist(c_i, x)^2$$

where dist is the standard (L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be Si.

2. Centroid update step:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

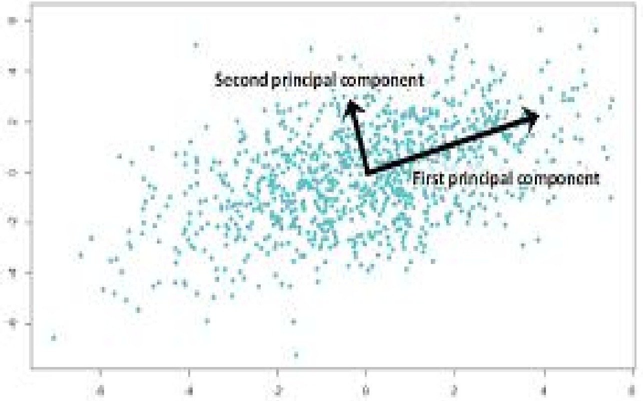
This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

III. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis, or PCA, is a dimension reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

How does PCA work -

1. Calculate the covariance matrix X of data points.
2. Calculate eigen vectors and corresponding eigen values.
3. Sort the eigen vectors according to their eigen values in decreasing order.
4. Choose first k eigen vectors and that will be the new k dimensions.
5. Transform the original n dimensional data points into k dimensions.



Advantages of Dimensionality Reduction

1. It helps in data compression, and hence reduced storage space.
2. It reduces computation time.
3. It also helps remove redundant features, if any.

IV. METHODOLOGY

A. *Dataset*

In this step we collect the data set from kaggle which contains high quality images of thousands of fruits

B. *Importing libraries*

In this step we import the following libraries :

NumPy: Base n-dimensional array package
SciPy: Fundamental library for scientific computing
Matplotlib: Comprehensive 2D/3D plotting
IPython: Enhanced interactive console
Sympy: Symbolic mathematics
Cv2: open Cv interface
Pandas: Data structures and analysis
Glob: Accessing Directories

C. *Principal component Analysis*

In this project we have a large dataset, when we are to apply k means clustering we have about 10000 features which when applied with PCA are reduced to just to 2 features. 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.

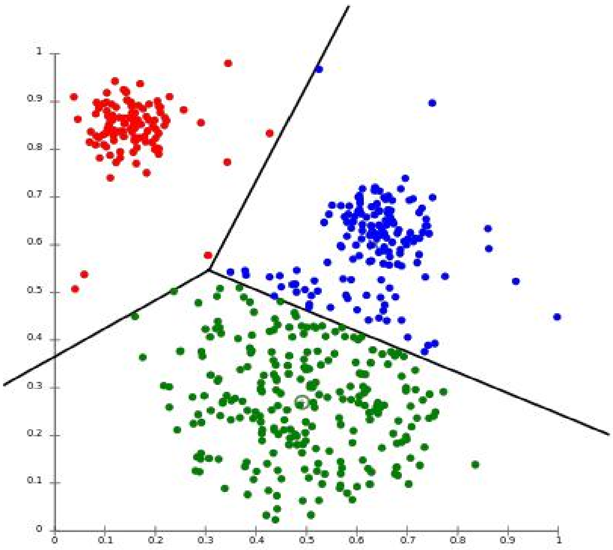
D. *K means clustering*

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data.

Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters.

A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of overfitting.

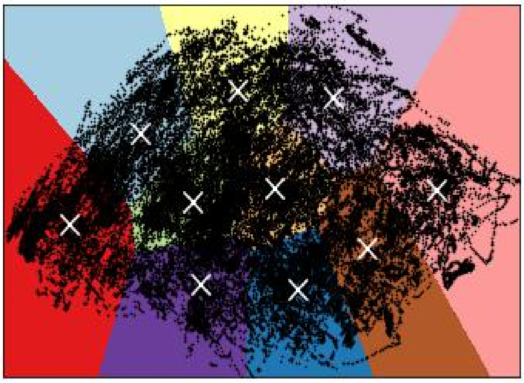
Here we choose k = 10 i.e we choose to make 10 clusters for this dataset.



V. EXPERIMENTAL RESULTS

We applied Principal component Analysis and then K-Means clustering on the image dataset. We got 10 clusters based on their shape and color. Thus k means clustering gives us a good result for the given fruits dataset.

The white cross represents the centroid and the different clusters are as shown in different colors. These clusters are made on the basis of their similarity.



REFRENCES

[1] J.M. Pena, J.A. Lozano, P. Larranaga, “An empirical comparison of four initialization methods for the K-Means algorithm”, Pattern Recognition Letters 20 (1999) 1020-1040.

[2] Sun Jigui, Liu Jie, Zhao Lianyu, “Clustering algorithms Research”,Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[3] Liwicki, Stephan, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. "Euler principal component analysis." International journal of computer vision 101, no. 3 (2013) : 498-518.