K-Means Algorithm and Application in Data Compression using Pascal and SwinGame API

Dung T. Lai
Faculty of Science, Engineering and Technology
Swinburne University of Technology
Hawthorn, Victoria 3122
Email: tuandunglai@gmail.com

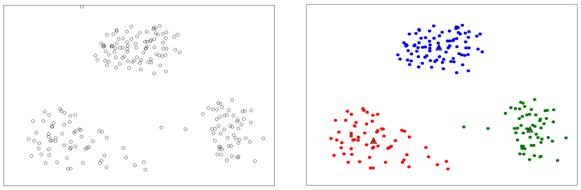
***Abstract*—Clustering problem is a task of dividing a set of objects (also called members) into different groups (called clusters) based on object's characteristics. Members of a group will have more similarities in comparison with those in other group. This report discusses a traditional clustering method called K-Means algorithm from mathematical perspective. Additionally, an experiment is provided to examine the algorithm in two dimensional space then an application in image compressing.**

I. Introduction

Clustering problems arise in many different applications: machine learning data mining and knowledge discovery, data compression and vector quantization, pattern recognition and pattern classification. [1]

The goal of K-Means Algorithm is to correctly separating objects in a dataset into groups based on object's properties. For instance, objects could be house and their properties are size, number of floor, location, power consumption per year, etc. The goal is to classify house dataset into groups which are luxury, average, poor. In that case, all properties of houses have to be processed to turn into number to create a vector, this process is called vectorization.

Another example, take each points in a panel as a objects and each object has two properties which are x-axis and y-axis location. And input $K = 3$. The algorithm correctly finds the cluster. (Fig. 1.a)



(a) Input: $N = 200$, $K = 3$

(b) Output

Fig. 1: K-Means on 2-dimensional Points

II. Mathematical Analysis

A. Input and Output

The K-Means Algorithm takes a set of observations $X = [x_1, x_2, ..., x_N] \in R^{d \times N}$ where each observation is a d-dimensional vector, N is the number of observations (members) and the number of group $(K, K < N)$ as two input. The algorithm outputs the center of K group $[m_1, m_2, ..., m_K] \in R^{d \times K}$ and the index or name of group that each member belonged to (label).

B. Lost Function and Optimization Problem

Suppose x_i $(i \in [1, N])$ belong to cluster k $(k \in [1, K]$, the lost value of observation x_i is the distance from observation x_i to center m_k in euclidean space, defined by $(x_i - m_k)$. Let's $y_i = [y_{i1}, y_{i2}, ..., y_{iK}]$ be the label vector of each observation x_i, $y_{ik} = 1$ if x_i belongs to group k and $y_{ij} = 0$ $\forall j \neq k$.

Label vector of each observation contains only one digit 1 because each observation belongs to only one group which leads to the following equation.

$$\sum_{k=1}^{K} y_{ik} = 1 \tag{1}$$

The objective is to minimize the within-cluster sum of squares (variance), also known as square errors of, where each square error of an observation x_i from group m_k is defined by:

$$||x_i - m_k||^2 = y_{ik}||x_i - m_k||^2 \tag{2}$$

From the equation 1, sum of all elements in a label vector is equal 1. The square error of an observation is:

$$y_{ik}||x_i - m_k||^2 = \sum_{j=1}^{K} y_{ij}||x_i - m_j||^2 \tag{3}$$

The square error of all observation is the sum of every square error of in the given set of observation. The goal is minimize the lost function, equation 4 where $Y = [y_1, y_2, ..., y_N]$ be the matrix contains all label vector of N observation and $M = [m_1, m_2, ..., m_K]$ be the center of K groups (clusters).

$$f(Y, M) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij}||x_i - m_j||^2 \tag{4}$$

The objective is also to find the center and label vector of each observation which are Y and M, the two outputs that are mentioned in II-A.

$$Y, M = argmin_{Y,M} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij}||x_i - m_j||^2 \tag{5}$$

C. Solving Optimization Problems

There are two variable in equation 5 which are center of each group of observation and label vector of each observation. The problem could be solved by fixed each variable then minimize the other variable.

1) Fixed M, center of observation group: Because all centers (M) are constant, the objective is to correctly identify label vector which is identifying the group that each observation belonged to so that the square error in equation 4 is minimized.

$$y_i = argmin_{y_i} \sum_{j=1}^{K} y_{ij}||x_i - m_j||^2 \tag{6}$$

Retrieving from equation 1. Because only one element in vector y_i $i \in [1, K] = 1$. Equation 6 could be rewritten as:

$$j = argmin_j||x_i - m_j]]^2 \tag{7}$$

The value of $||x_i - m_j||^2$ is the square of distance from observation to center of group in euclidean space. Concretely, when M is constant, equation 7 shows that minimizing the sum of square error could be achieved by choosing label vector so that the center are closest to observation.

2) Fixed Y, label vector of each observation: When label vector (Y) are constant, the objective is to correctly identify the center so that the square error in equation 4 is minimized. In this case, the optimization problem in equation 5 could be rewritten by the following equation.

$$m_j = argmin_{m_j} \sum_{i=1}^{N} y_{ij}||x_i - m_j||^2 \tag{8}$$

The equation 8 is a convex function and differentiable for each $i \in [1, N]$. Hence equation 8 could be solved by finding the root of the partial derivative function. This approach will make sure that the root is the the value that make the function reach a optimum.

Let's $g(m_j) = \sum_{i=1}^{N} y_{ij}||x_i - m_j||^2$ (retrieving from equation 8 and take the partial derivative of $g(m_j)$:

$$\frac{\partial g(m_j)}{\partial m_j} = 2\sum_{i=1}^{N} y_{ij}(m_j - x_i) \tag{9}$$

The equation 9 is equal 0 is equivalent to:

$$m_j \sum_{i=1}^{N} y_{ij} = \sum_{i=1}^{N} y_{ij}x_i \tag{10}$$

$$\Leftrightarrow m_j = \frac{\sum_{i=1}^{N} y_{ij}x_i}{\sum_{i=1}^{N} y_{ij}} \tag{11}$$

The value of $y_{ij} = 1$ when observation x_i belongs to group m_j. Hence, the denominator of equation 11 $\sum_{i=1}^{N} y_{ij}$ is the number of observations that belonging to group m_j and the nominator $\sum_{i=1}^{N} y_{ij}x_i$ is the sum of all observations belonging to group m_j.

In other word, when Y is constant, the square errors could be minimize by assigning the centers to the means of observations in the groups that the observations belonging to.

D. Algorithm summary and Flowchart

1) Summary: The algorithm can be done by continuously constantize Y and M, one each a time as discussed in II-C1 and II-C2.
Step 1.Clusters the data into k groups where k is predefined.
Step 2.Select k points at random as cluster centers.
Step 3.Assign objects to their closest cluster center according to the Euclidean distance function.
Step 4.Calculate the centroid or mean of all objects in each cluster.
Step 5.Repeat steps 2.

2) Flowchart: The following chart describe K-Means Algorithm

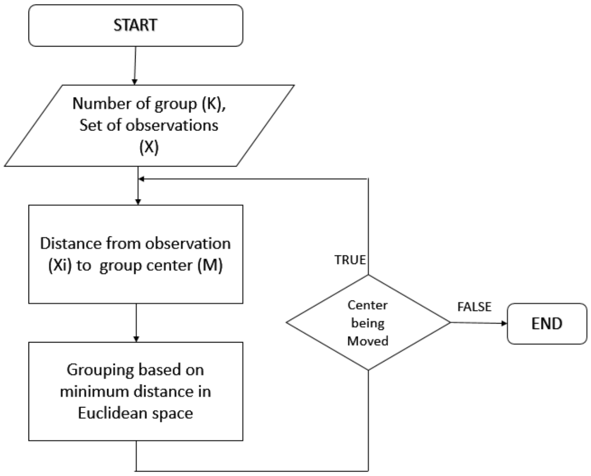

START
Number of group (K), Set of observations (X)
Distance from observation (Xi) to group center (M)
Grouping based on minimum distance in Euclidean space
TRUE
Center being Moved
FALSE
END

Fig. 2: K-Means Algorithm Flowchart

E. Discussion

1) Convergence: The algorithm will stop after a certain number of iteration because the square error function is a strictly decreasing sequence and the square error is always greater than 0. But this algorithm will not make sure that it will find a global optimum because solving the equation 8 by finding the root when the partial derivative is equal 0 will only return the value for local optima but not make sure that local optima will be a global minimum.

The following figure describe a case where poorly seeding leads to a local optimum.

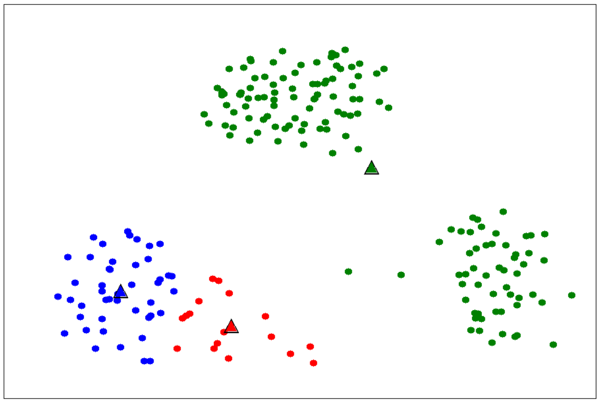


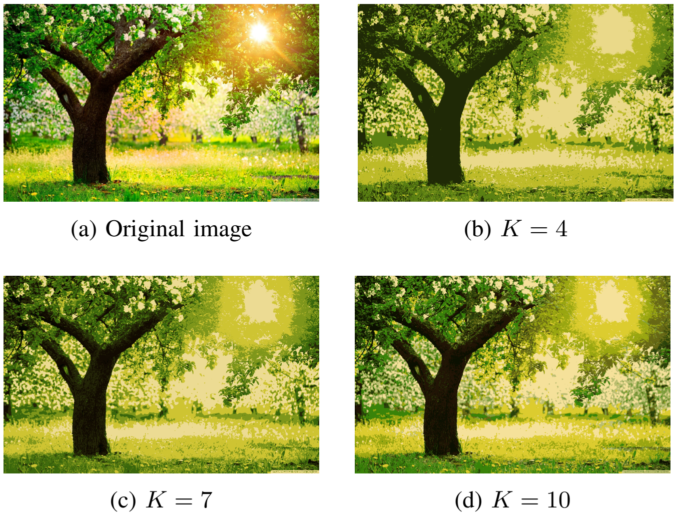
Fig. 3: Poorly Seeding K-Means

In this case, the square error is 6769747 which is about 4 times greater than the square error produce by figure 1b (1614826).

2) Sensitiveness to initial cluster: K-Means algorithm requires careful seeding, which means the final result is very sensitive to the initial value of cluster. Numerous efforts have been made to improving K-Means clustering algorithm due to its drawbacks [2]

III. Application in Data Compression

An experiment will be reperformed where K-Means algorithm is applied to reduce the size of image and outputs a new image without the smaller number of color as compared to the original one. This experiment is carried using Pascal programming language and SwinGame API. Each pixel of a image contains three elements which are red, green, blue (RGB) value. Let's each pixel be the observation (X) then the number of pixel in an image be the number of observations. Each observation has three properties which are RGB value. In this case, K-Means algorithm is applied to identify K main colors in that image.



(a) Original image

(b) $K = 4$

(c) $K = 7$

(d) $K = 10$

Fig. 4: Image Segmentation on 2560x1600 image

The table below shows how file size of the original image is reduced.

K (Color)	File Size (KB)
Original Image	1572
10	1063
7	742
4	507

IV. Conclusion

K-Means Algorithm could be very simple and quick to be implemented, the clustering problems where all clusters are centroids and separated can be solved by the algorithms. However, it will not be effective when the dataset and clusters are more complex.

This report doesn't come with new idea to improve the effectiveness of the algorithm, the aim of the report is to introduce the reader to a basic, entry level clustering methods with some visual example on 2-dimensional and 3-dimensional dataset.

V. Further research

The algorithm is simple to implement, however, the sensitivity to initial centroids, and strict structure of dataset, etc are those drawbacks of the algorithm which are undeniable . Further research could be made to improve the value of initial centroids. The traditional algorithm is depended on randomness, research could also be made to discover a way to make a fixed initial centroids. Furthermore, on a large dataset, the algorithm could be very slow to converge. Research could be spent to make the iteration stops earlier.

Acknowledgment

The authors would like to thank Tiep V. Huu for useful machine learning blog and support, and also Prof. Andrew Ng for inspirational machine learning course on Coursera.

References

[1] Joaqun Prez Ortega, Ma. Del, Roco Boone Rojas, and Mara J. Somodevilla *”Research issues on k-means algorithm: An experimental trial using matlab”*.

[2] Arthur, D., Vassilvitskii, S., 2016. *k-Means++: The Advantages of Careful Seeding.*Technical Report, Stanford.