



CSE - 4255 Data Mining and Warehousing  
Lab

*Comparison Between the Performance of K - Means  
and K - Medoids Algorithm in Clustering*

Saif Mahmud  
Roll: SH - 54

M. Tanjid Hasan Tonmoy  
Roll: SH - 09

**Submitted To:**

Dr. Chowdhury Farhan Ahmed  
Professor

&

Abu Ahmed Ferdaus  
Associate Professor

Department of Computer Science and Engineering  
University of Dhaka

October 28, 2019

# 1 Problem Definition

In this experiment, we have implemented K-Means and K-Medoids Algorithm for Clustering. We have evaluated cluster quality in terms of purity measure and included analysis for determining number of clusters using elbow method. Execution time for varied number of clusters for both algorithms have been compared.

## 2 Dataset Description

Since the task at hand is targeted towards unsupervised segmentation of data, the datasets that have been selected do not have any class label annotations. We chose a number of datasets which have been described in the following subsections.

### 2.1 Credit Card Unsupervised

The credit card dataset contains the usage record of 9000 credit card users over a period of 6 months. This data is available at [kaggle.com](https://www.kaggle.com/arjunbhasin2013/ccdata)<sup>1</sup> and contains 18 variables related to the particular customers behavior relating to finance. The attributes include account balance, frequency of balance updates, amount of purchases made from account etc. These categorical attributes do not include any class labels.

This data may be used to segment the customers into different groups so that a company may design effective marketing and advertising strategy. Such strategies may help to provide customized services and aid internal decision making.

### 2.2 Weather Madrid 1997 - 2015

This dataset<sup>2</sup> contains the daily weather records of the city of Madrid from 1997 to 2015. There are 23 attributes in the that contains information such as minimum and maximum temperature, dew point, humidity etc. We exclude one column containing major weather event since majority of the data are missing.

---

<sup>1</sup><https://www.kaggle.com/arjunbhasin2013/ccdata>

<sup>2</sup>[https://www.kaggle.com/juliansimon/weather\\_madrid\\_1emd\\_1997\\_2015.csv](https://www.kaggle.com/juliansimon/weather_madrid_1emd_1997_2015.csv)

Such dataset may help segment the days into different categories which may correspond to the seasons for example summer and winter days or more granular like hot and humid etc days depending on the number of clusters.

### **2.3 Google Review Ratings**

This dataset includes reviews of different types of places from a number of visitors. Available in the UCI repository, this dataset may also help segment the visitors into groups similar to credit card dataset in subsection 2.1. The attributes contain ratings for different type of places.

### **2.4 London bike sharing dataset**

This unsupervised dataset provides count of new bike shares grouped by hour of the day in London during the years 2015-2016. In addition, the remaining attributes contain information regarding weather(temperature, humidity), season and information like whether the day is a holiday or weekend. This may be used to cluster the different times of the day which may demonstrate a wide range of distinct usage patterns of bike sharing depending on the granularity of the segmentation.

### **2.5 BuddyMove Data Set**

Containing user interest information extracted from user reviews published in the website holidayiq.com, this data may also be used for user segmentation similar to the google review dataset. The attributes contain the number of reviews a user has posted about some particular place type e.g. theatres or parks.

### **2.6 Summary**

## **3 Theory and Implementation**

### **3.1 K - Means**

K - Means is centroid-based partitioning algorithm where the centroid of a cluster, represented by  $C_i$  is used to represent the respective cluster. The distance between an object and the centroid of the cluster is given by the

Table 1: Dataset Statistics

Dataset	Number of Samples	Number of Attributes
Credit Card Unsupervised	8950	18
Weather Madrid 1997 - 2015	6812	23
Google Review Ratings	5456	24
London bike sharing dataset	17414	9
BuddyMove Data Set	250	6

Euclidean distance denoted as  $\text{dist}(p, c_i)$ . The sum of squared error between all objects in a cluster and centroid are minimized to improve the cluster quality.

$$E = \sum_{i=1}^k \sum_{p \in C_i} \text{dist}(p, c_i)^2 \quad (1)$$

The centroid of a cluster is defined as the mean value of the points within the cluster for K-means algorithm. K representative objects are chosen randomly at the beginning. All of the remaining objects are assigned to the cluster for which the distance between the centroid and the object is minimum. The within-cluster variation is improved in an iterative manner by calculating new centroid and updating the cluster assignments.

### 3.2 K - Medoids

K-Medoids algorithm uses actual object from the data as representative for the clusters. Compared K-Means, this approach makes K-Medoids less sensitive to outliers. Each remaining object in the dataset is assigned to the cluster of the closest representative object.

K-medoids algorithm groups  $n$  objects into  $k$  clusters by minimizing the absolute error defined in (2)

$$E = \sum_{i=1}^k \sum_{p \in C_i} \text{dist}(p, o_i) \quad (2)$$

The Partitioning Around Medoids (PAM) is an iterative, greedy algorithm to implement  $k$  medoids since finding exact median every time incurs

quadratic cost. Similar to k-means algorithm, the initial representative objects are randomly chosen. It is considered whether replacing a representative object by a nonrepresentative would improve the clustering quality. All the possible replacements are tried out. This process continues until the quality of the resulting clustering cannot be improved by performing any replacement. In our implementation, a sampling-based method called CLARA (Clustering LARge Applications) has been used to deal with larger data sets. Instead of taking the whole data set into consideration, CLARA uses a random sample of the data set.

## 4 Evaluation of Clustering

### 4.1 Elbow Method

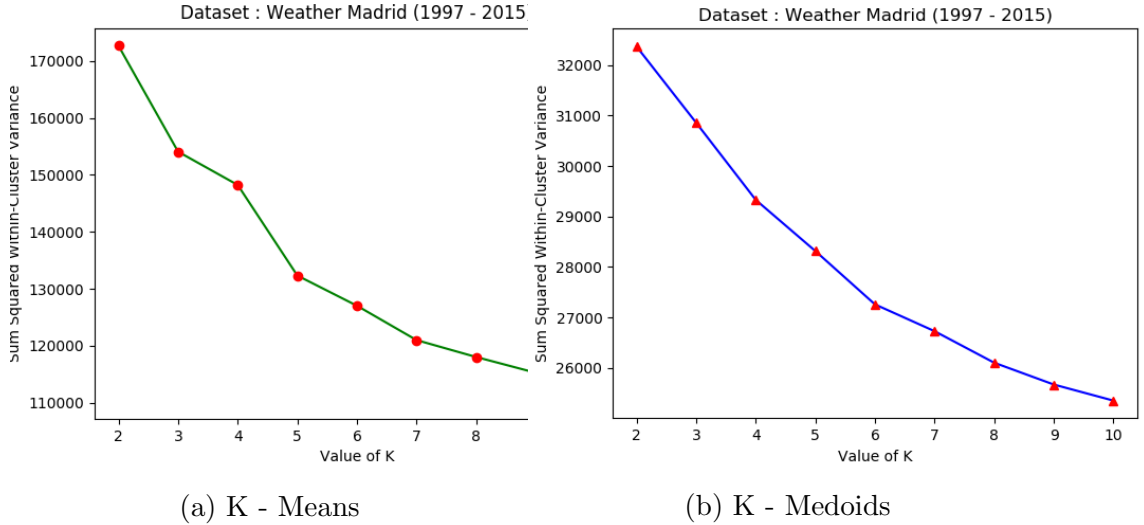


Figure 1: Determining Value of K through Elbow Method



Figure 2: Determining Value of K through Elbow Method (K - Means)

## 4.2 Visualization

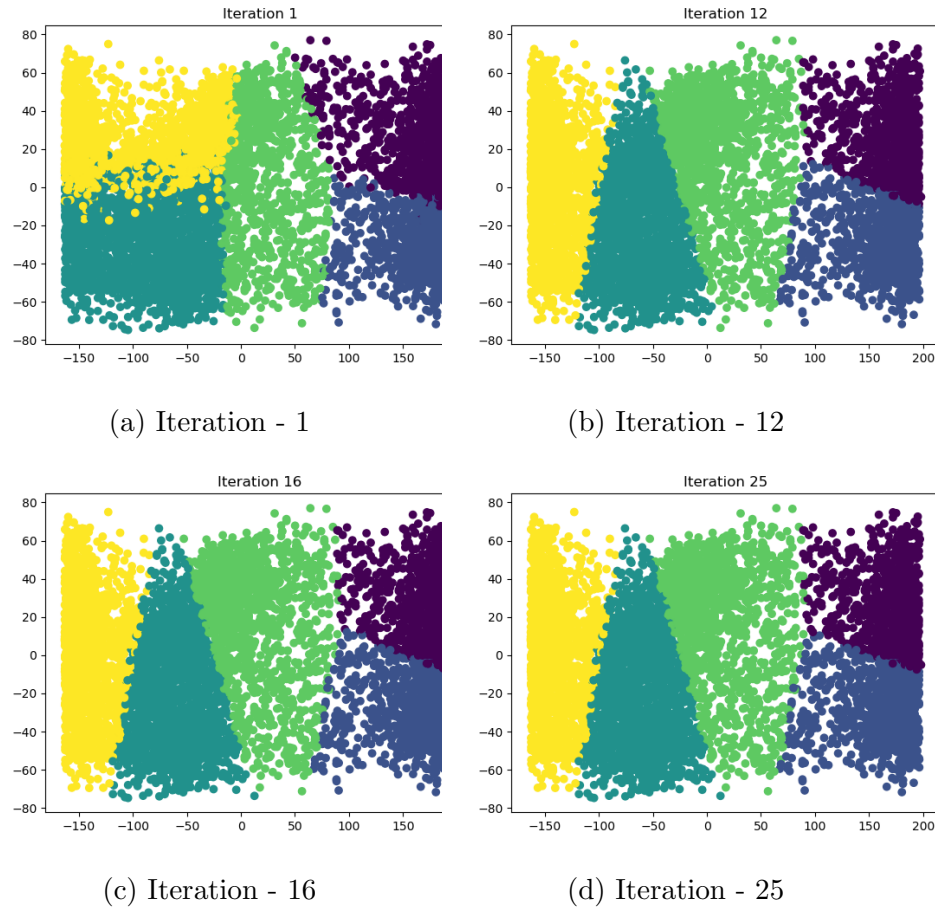


Figure 3: Visualization of K Means using PCA on Weather Madrid Dataset

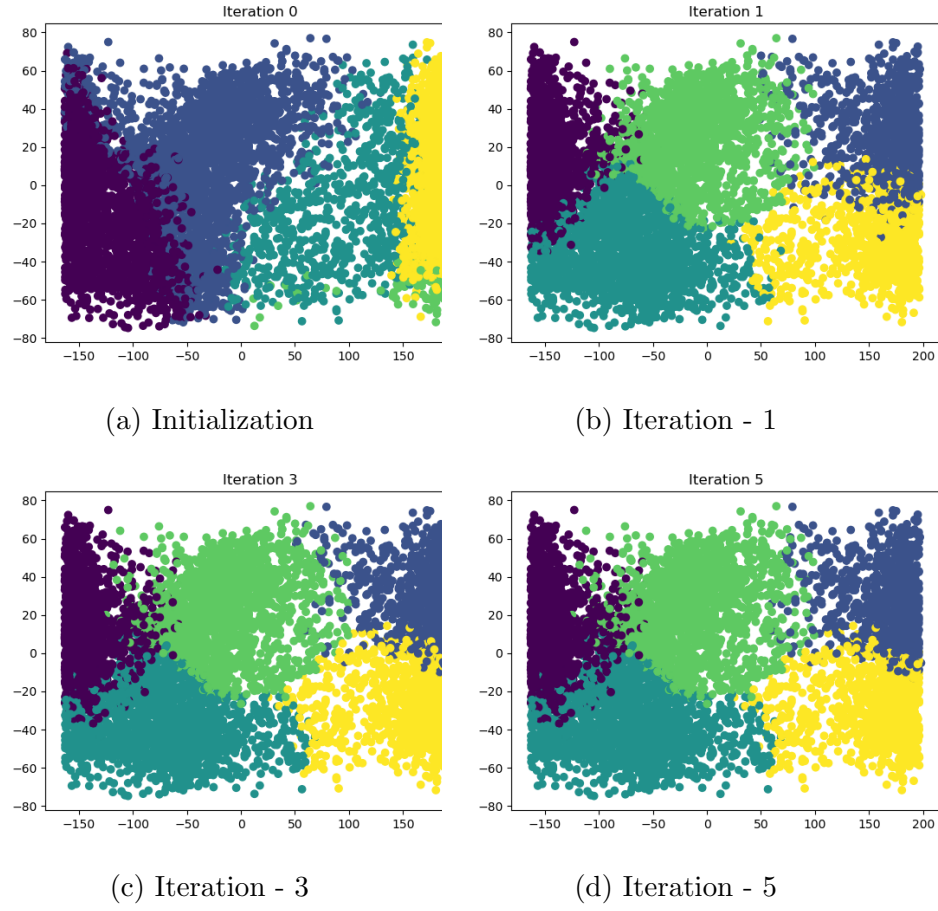


Figure 4: Visualization of K-Medoids using PCA on Weather Madrid Dataset



### 4.3 Time Complexity Comparison

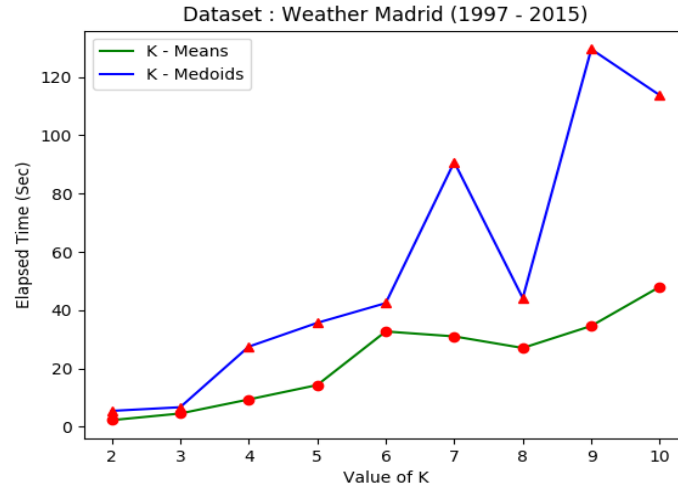


Figure 5: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

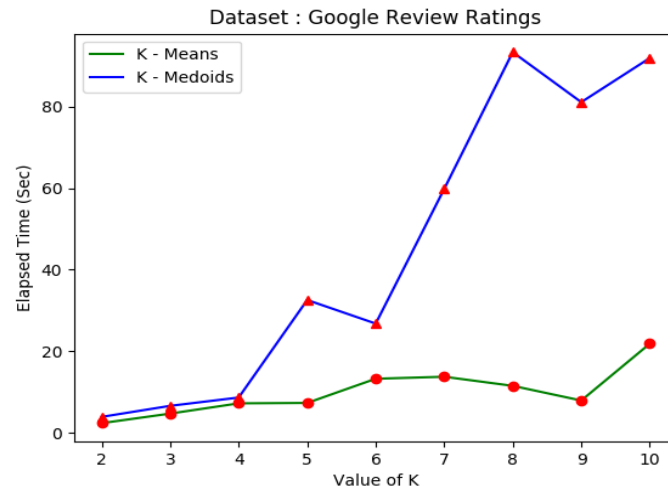


Figure 6: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

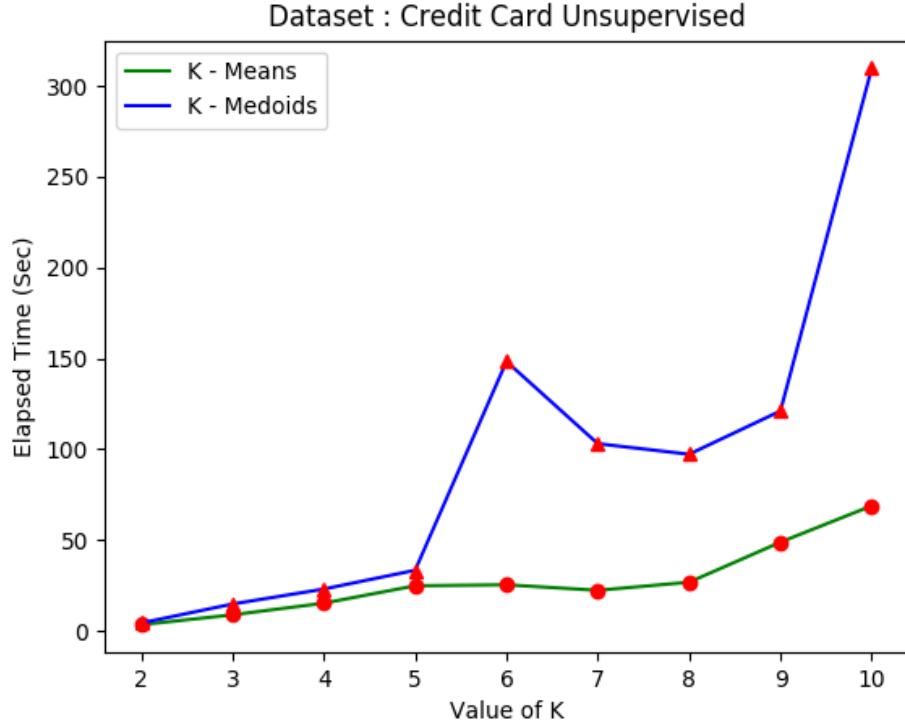


Figure 7: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

## 5 Conclusion

It can be concluded based on the experimental results that, K-medoids algorithm is more robust than K-means since it is less affected by outliers and noise. However, K-medoids is much costlier in terms of time complexity, especially if PAM implementation is used. In the case of using CLARA, the size of the random sample is also a factor. Both of these algorithms have the limitation that the number of clusters needs to be defined prior to running the algorithms.