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 27, 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 3 datasets which we describe 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[1] 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[2] 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.

[1] `https://www.kaggle.com/arjunbhasin2013/ccdata`

[2] `https://www.kaggle.com/juliansimon/weather_madrid_lemd_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 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.

2.5 Summary

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
BuddyMove Data Set	250	6

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.

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 (1)

$$\sum_{i=1}^{k} \sum_{p \epsilon C_i} dist(p, o_i) \tag{1}$$

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

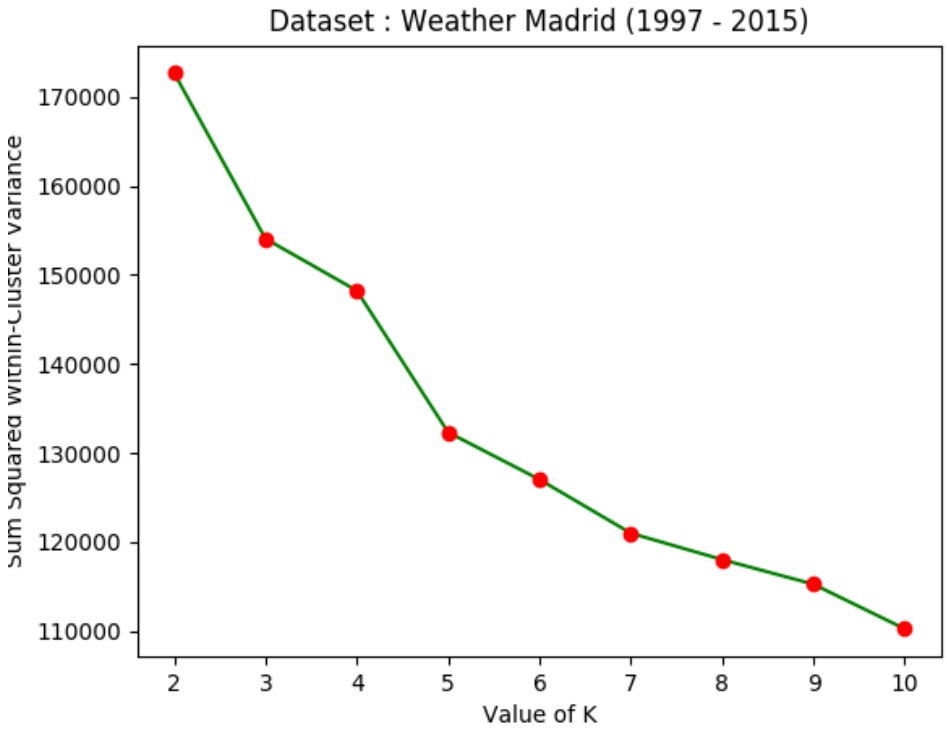

Figure 1: Determining Value of K through Elbow Method (K - Means)

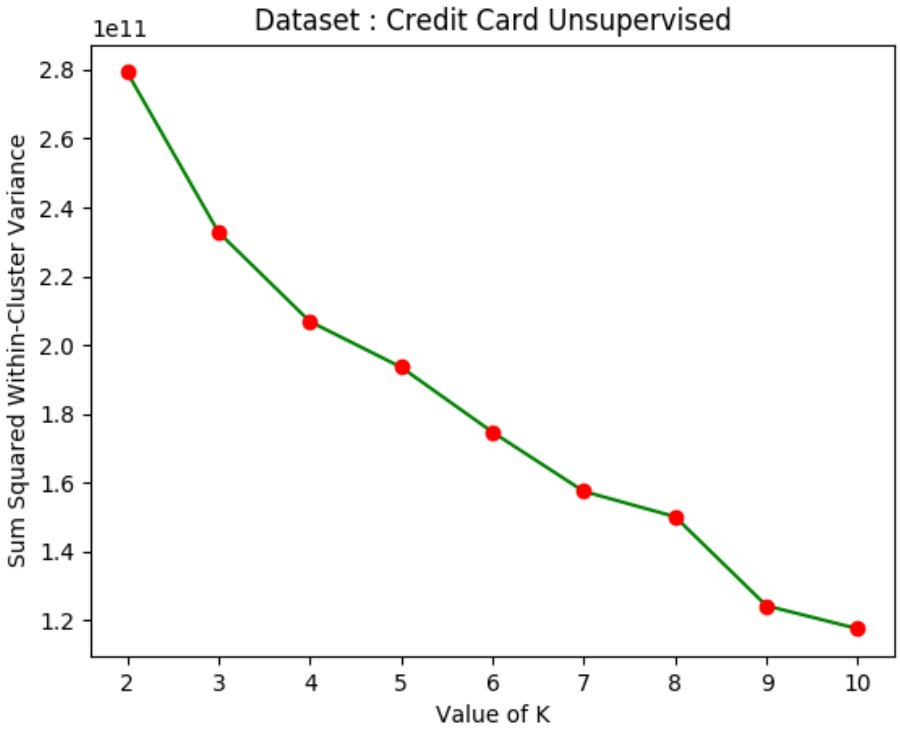

Figure 2: Determining Value of K through Elbow Method (K - Means)

4.2 Time Complexity Comparison

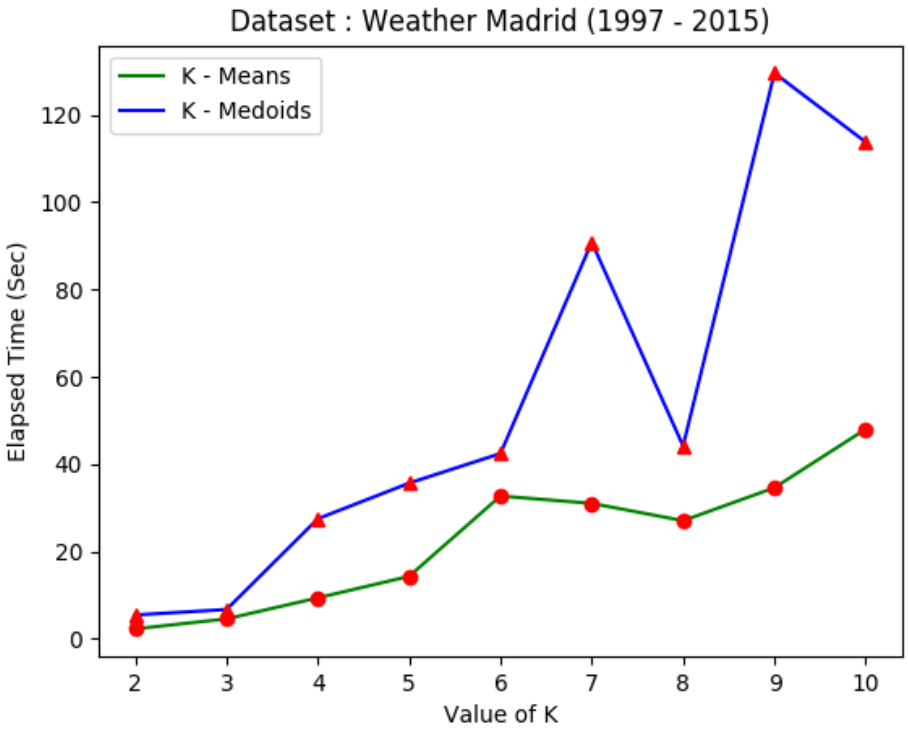

Figure 3: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

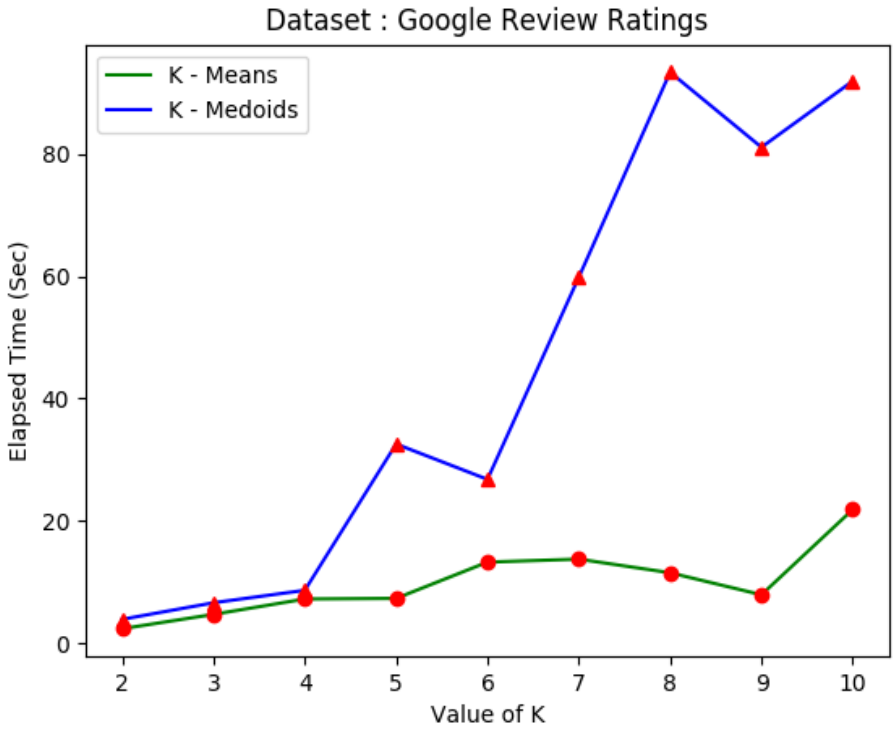

Figure 4: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

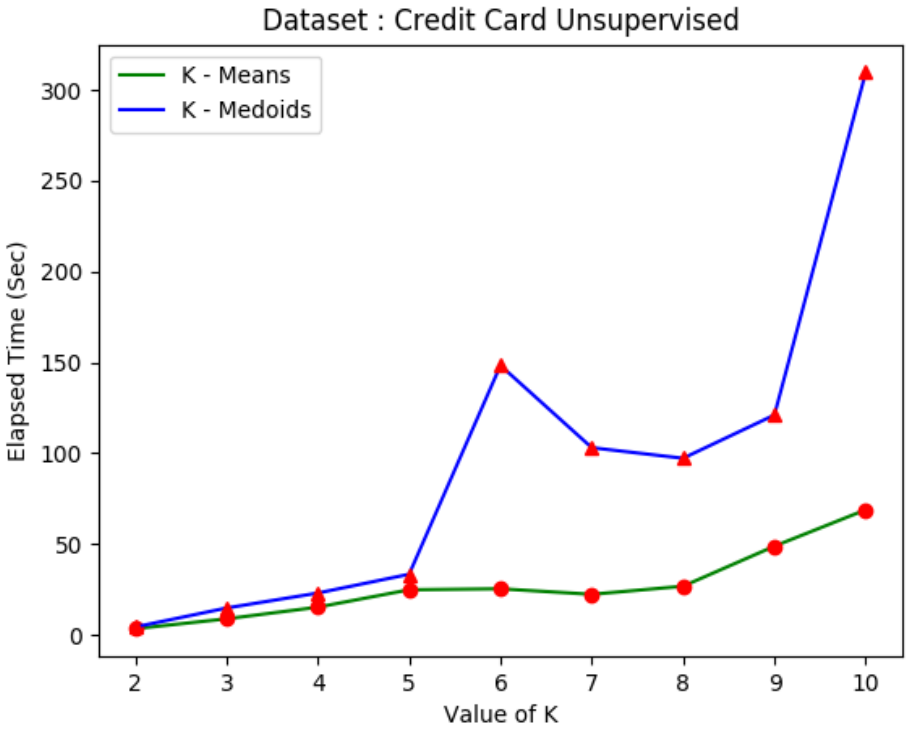

Figure 5: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

5 Conclusion