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[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 Travel Reviews Data Set

This data set is populated by crawling TripAdvisor.com. Reviews on destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0) and average rating is used against each category per user.

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

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
Travel Reviews Data Set	980	11
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. The distance between an object and the centroid of the cluster is given by the Euclidean distance denoted as dist(p, c_i). The sum of squared error between all objects in a cluster and centroid are minimized to improve the cluster quality.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} dist(p, c_i)^2 \tag{1}$$

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)

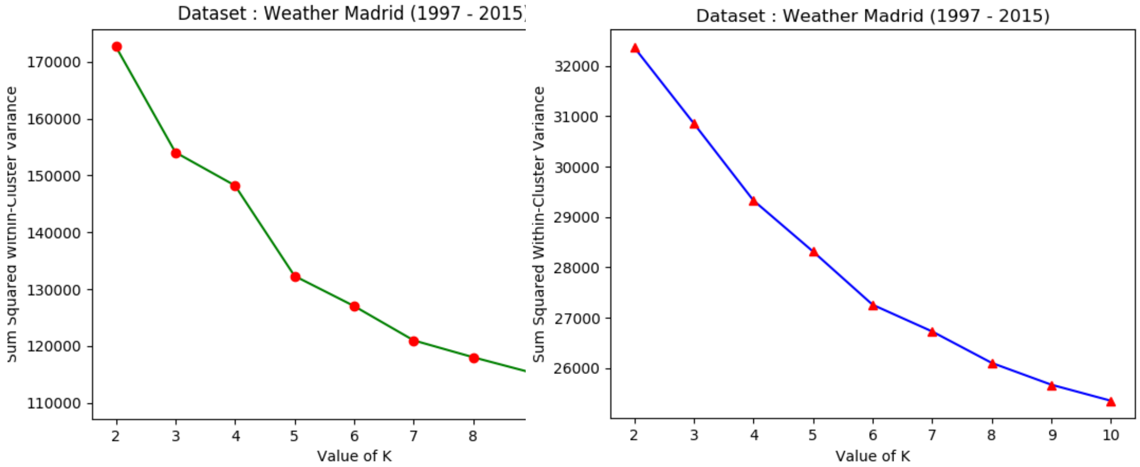
$$E = \sum_{i=1}^{k} \sum_{p \epsilon C_i} dist(p, o_i) \tag{2}$$

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

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. However, the marginal effect of reducing the sum of within-cluster variances may drop if too many clusters are formed, because splitting a cohesive cluster into two gives only a small reduction. Consequently, a heuristic for selecting the right number of clusters is to use the turning point in the curve of the sum of within-cluster variances with respect to the number of clusters.



(a) K - Means

(b) K - Medoids

Figure 1: Determining Value of K through Elbow Method

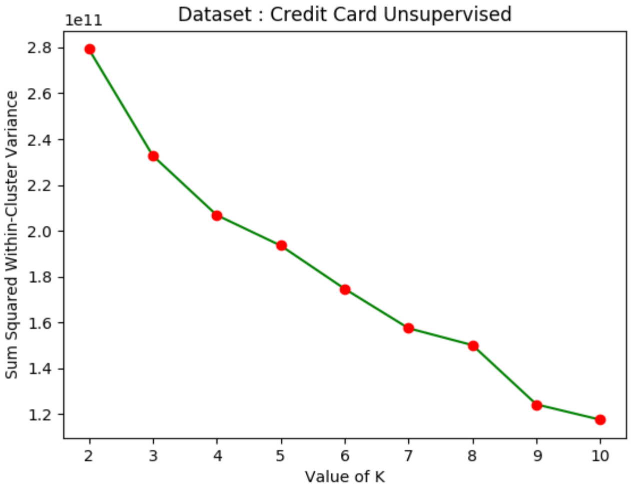
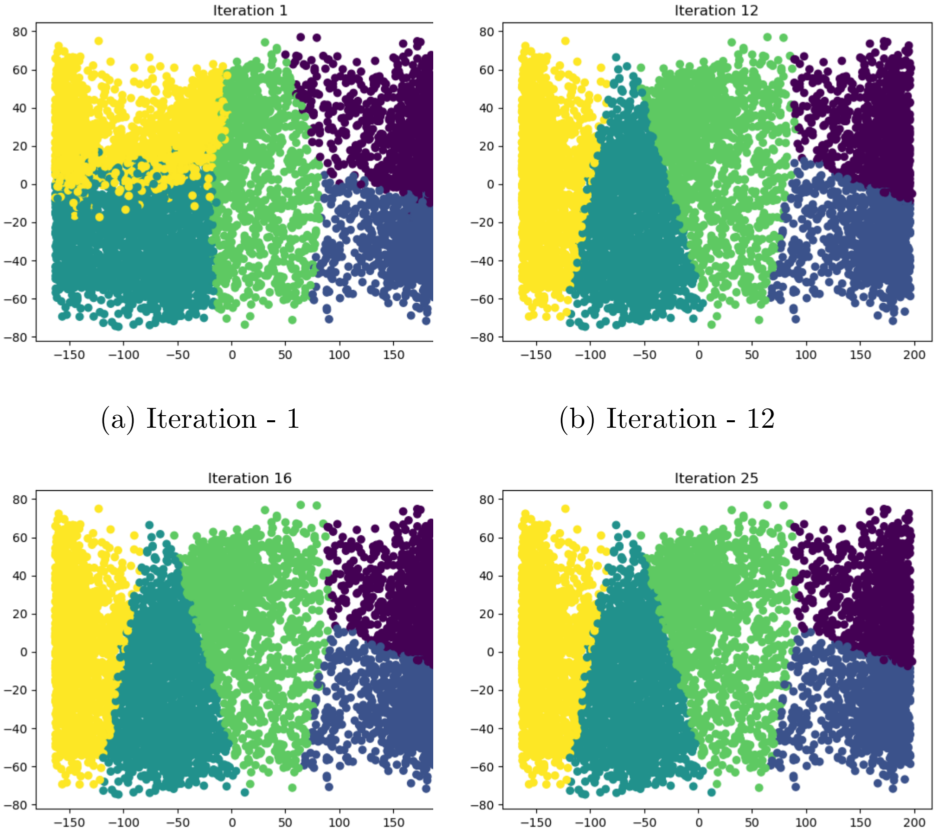


Figure 2: Determining Value of K through Elbow Method (K - Means)

4.2 Visualization



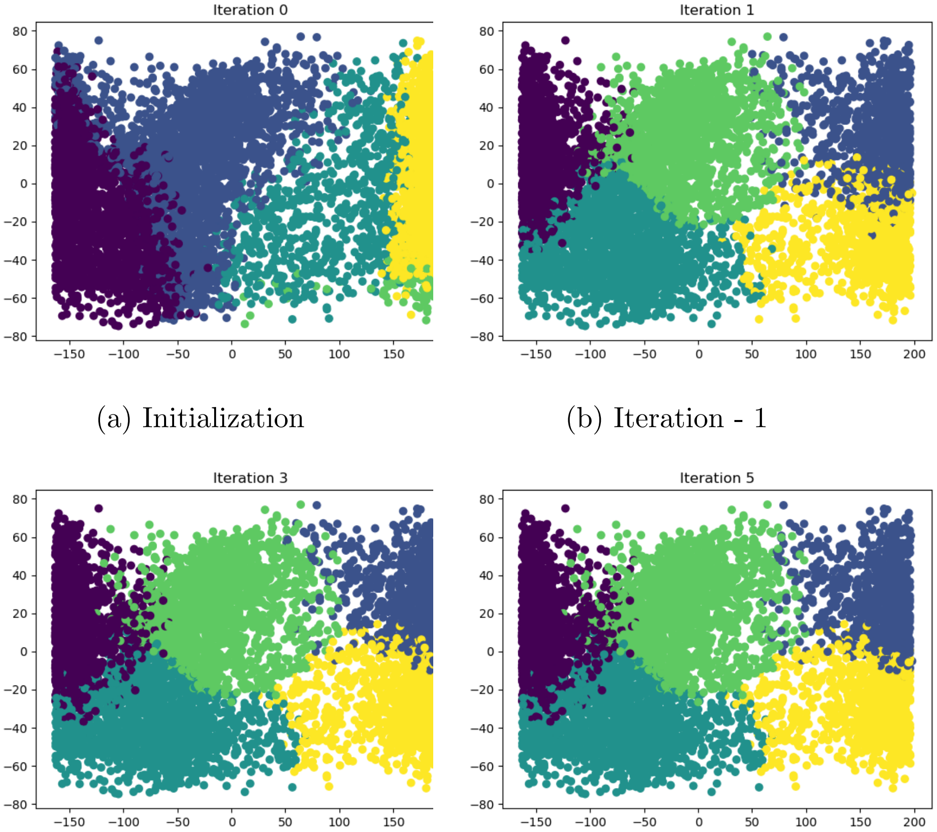
(a) Iteration - 1

(b) Iteration - 12

(c) Iteration - 16

(d) Iteration - 25

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



(a) Initialization

(b) Iteration - 1

(c) Iteration - 3

(d) Iteration - 5

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

4.3 Time Complexity Comparison

4.4 Cluster Evaluation

When the ground truth of a data set is not available, we have to use an intrinsic method to assess the clustering quality. In general, intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are. Many intrinsic methods have the advantage of a similarity metric between objects in the data set. The silhouette coefficient is such a measure.

The range of Silhouette score is $[-1, 1]$. Its analysis is as follows:

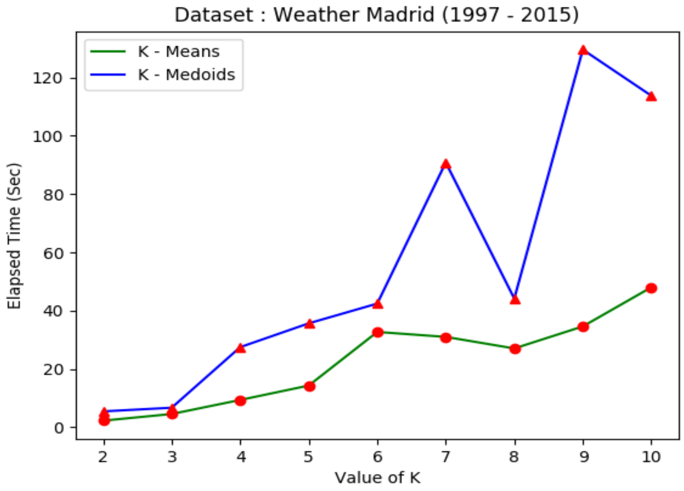


Figure 5: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

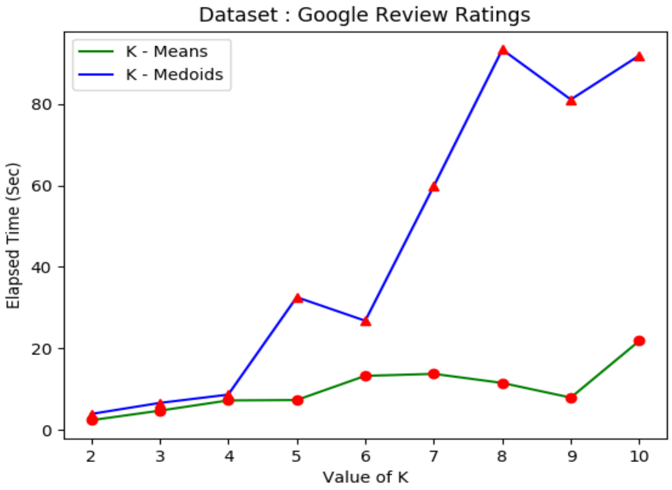


Figure 6: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

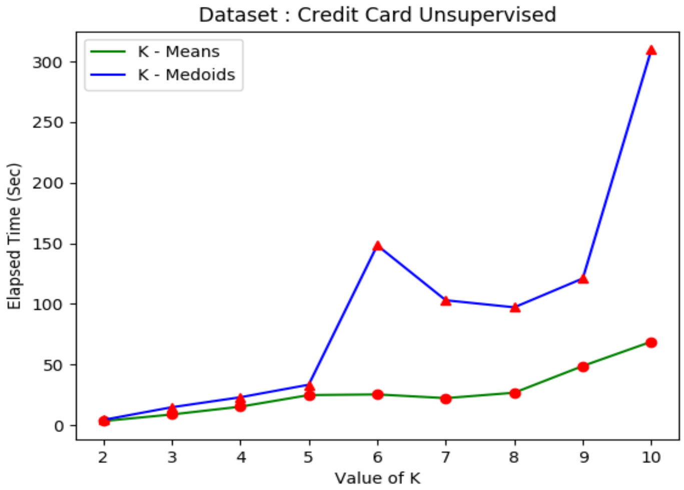


Figure 7: Comparison of Elapsed Time between K - Means and K - Medoids Algorithm

- +1 Score : Near +1 Silhouette score indicates that the sample is far away from its neighboring cluster.
- 0 Score : 0 Silhouette score indicates that the sample is on or very close to the decision boundary separating two neighboring clusters.
- -1 Score : 1 Silhouette score indicates that the samples have been assigned to the wrong clusters.

The calculation of Silhouette score can be done by using the following formula:

$$SilhouetteScore = \frac{(p-q)}{max(p,q)} \tag{3}$$

Where, p = mean distance to the points in the nearest cluster, q = mean intra-cluster distance to all the points.

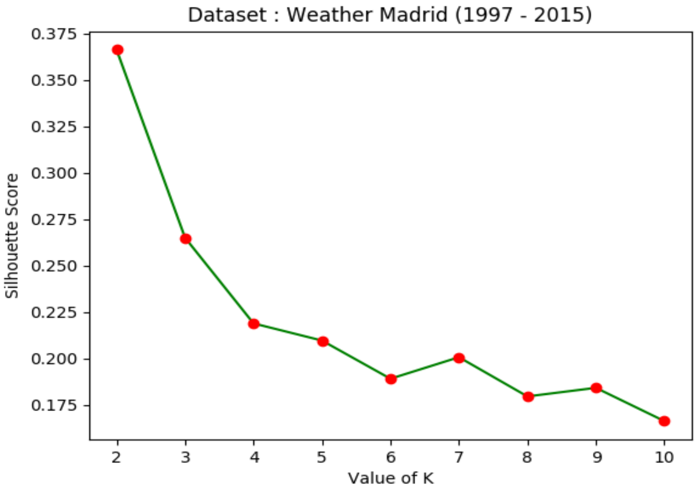


Figure 8: Intrinsic Evaluation of Cluster : Silhouette Score (K - Means)

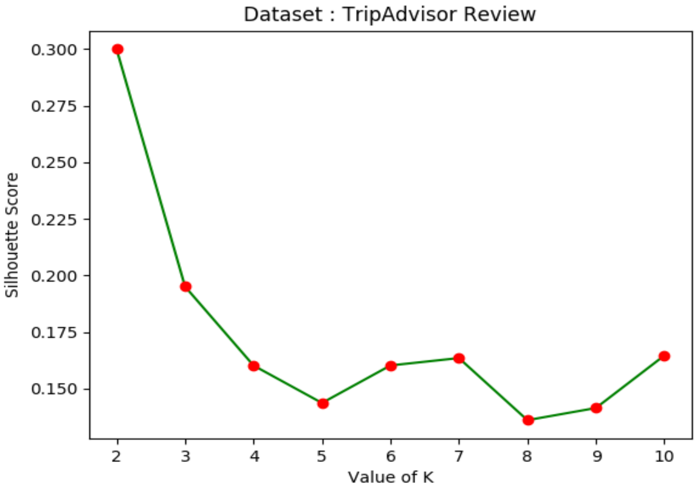


Figure 9: Intrinsic Evaluation of Cluster : Silhouette Score (K - Means)

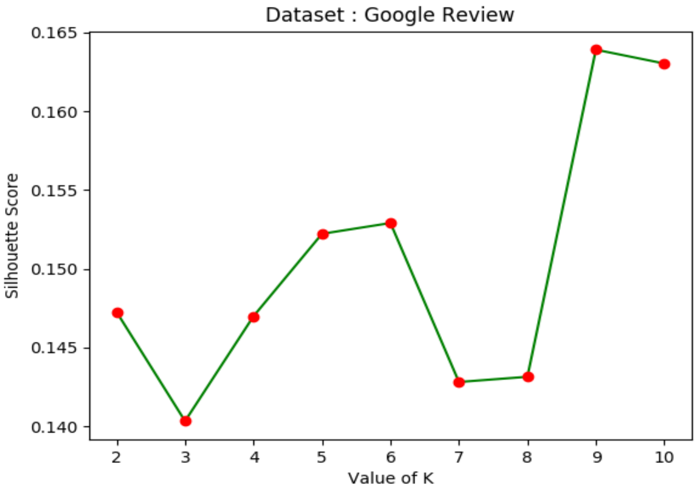


Figure 10: Intrinsic Evaluation of Cluster : Silhouette Score (K - Means)

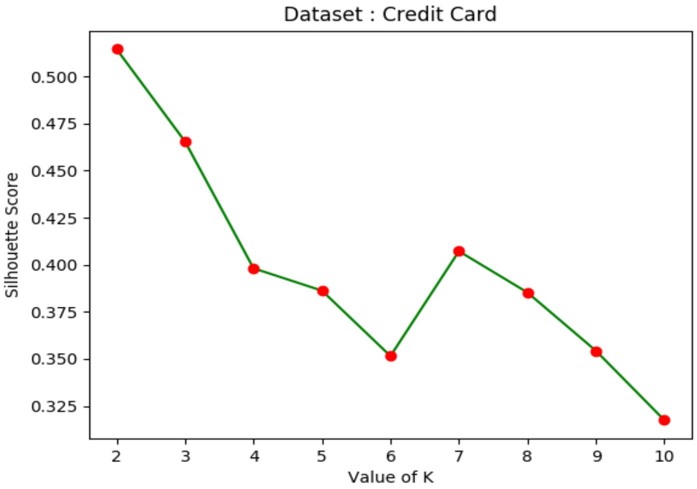


Figure 11: Intrinsic Evaluation of Cluster : Silhouette Score (K - Means)

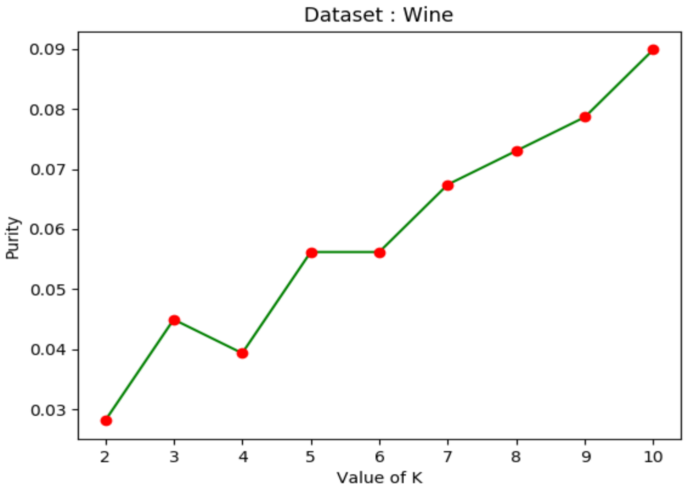


Figure 12: Extrinsic Evaluation of Cluster : Purity (K - Means)

Within the context of cluster analysis, Purity is an external evaluation criterion of cluster quality. It is the percent of the total number of objects(data points) that were classified correctly, in the unit range $[0, 1]$.

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j|c_i \cap t_j| \tag{4}$$

where N = number of objects (data points), k = number of clusters, c_i is a cluster in C, and t_j is the classification which has the max count for cluster c_i.

5 Conclusion

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