Department of Computer Science and Engineering

University of Dhaka

4^{th} *year* 2^{nd} *Semester*, 2021

CSE4255: Introduction to Data Mining and Warehousing Lab

Assignment: 03

Implementation and Comparison of k-Means and k-Medoids Clustering Algorithms.

Submitted by:

Md. Ahasanul Alam, Roll: 10

Submission Date:

14 March 2021

Submitted to:

Dr. Chowdhury Farhan Ahmed.

Professor, Department of Computer Science and Engineering,

University of Dhaka.

1 Introduction

Clustering is a Data Mining technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm. In this assignment, we have implemented two clustering algorithms. The algorithms are:

- **k-Means Clustering Algorithm**
- **k-Medoids Clustering Algorithm**

1.1 k-Means Clustering Algorithm

k-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of theclusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

1.2 k-Medoids Clustering Algorithm

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. The basic idea is as follows: Select K representative points to form initial clusters, and then repeatedly moves to better cluster representatives. All possible combinations of representative and non-representative points are analyzed, and the quality of the resulting clustering is calculated for each pair. An original representative point is replaced with the new point which causes the greatest reduction in distortion function. At each iteration, the set of best points for each cluster form the new respective medoids.

The time complexity of the k-Medoids algorithm is $O(K(NK)^2I)$. Naive k-Medoids is not scalable for large dataset, and some algorithms have been proposed to improve the efficiency, such as Clustering LARge Applications (CLARA) [2] and Clustering Large Applications based upon RANdomized Search (CLARANS) [3]. In case of CLARANS, First, it randomly selects k objects in the data set as the current medoids. It then randomly selects a current medoid x and an object y that is not one of the current medoids. Can replacing x by y improve the absolute-error criterion? If yes, the replace- ment is made. CLARANS conducts such a randomized search l times. The set of the current medoids after the l steps is considered a local optimum. CLARANS repeats this randomized process m times and returns the best local optimal as the final result. In this experiment we have used $l = 20\%of\#dataset$ and $m = 30$.

Dataset Name	# Attribute	# cluster	# Instances	Attr Type	Dataset Area
2D points	2	4	100	Real	Synthetic Data
Iris	4	3	150	Real	Life
Wine	13	3	178	Integer, Real	Physical
ILPD [1]	9	2	579	Integer, Real	Medical
seeds	7	5	210	Real	Life
Buddy Move	7	6	249	Real	N/A
HCV	10	3	615	Real	Life
Travel Review	10	7	980	Real	Life
Weekly Sales Transactions	52	3	831	Integer	Market
Wholesale Customers	6	8	440	Real	Business
User Knowledge Modeling	4	6	403	Real	Computer

Table 1: Information about experimental datasets

2 Implementation Detail

In this assignment, we have implemented the k-Means and k-Medoids clustering algorithms and experimented on 10 popular real life data-sets to compare the performances. To visualize the clusters we have also generated a synthetic datasets of 100 2D points and compared our clustering methods in it. The programs are written in Python programming language (version 3.8.5). All the experiments are executed on a laptop with Intel Core i5 2.2GHz CPU and 12GB RAM. To determine the number of cluster k, we have run the elbow method by varying the k from [2-10]. To avoid statistical variations we have run each experiment 5 times and took the mean of the results. Performance were evaluated based on 2 standard measures, such as (i) Variance (ii) Silhouette Coefficient Score. For the datasets where the ground truths are available we have also calculated the Bcubed Precision and Bcubed Recall to analysis the clustering quality. For Silhouette Coefficient score, we have used python sklearn API [4]. We have used the following 8 datasets to compare the performances of the two algorithms. The characteristics of the datasets are shown in Table 1

3 Determining K

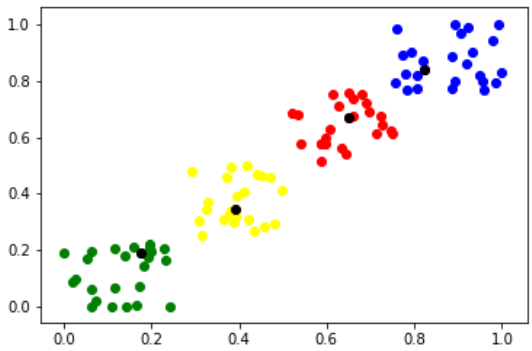

Figure 1: 2D points dataset

[1] ILPD stands for Indian Liver Patient Dataset

To determine what the actual number of the clusters can be we have used the elbow method by varying the K in range [2-10].

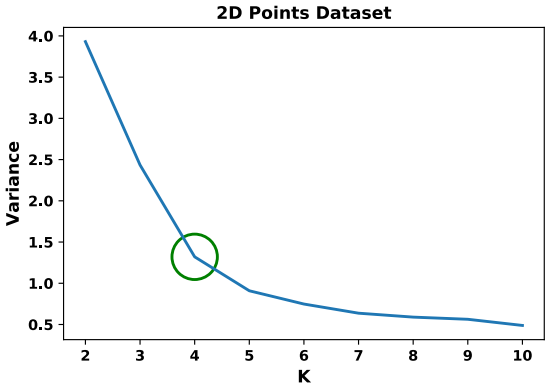

Figure 2: Elbow method for 2D points dataset

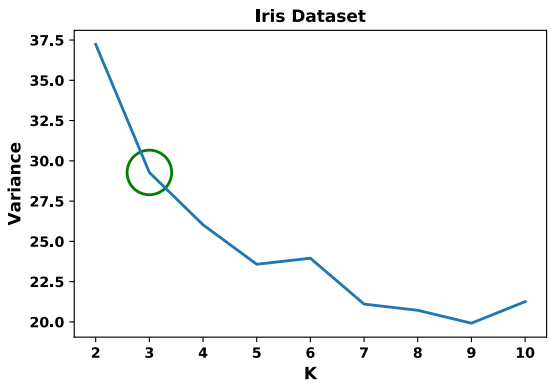

Figure 3: Elbow method for Iris dataset

3.1 Dataset: 2D pints

This is a self-made sythetic dataset with two attributes (x-axis value and y-axis value). This dataset was generated with four deterministic yet close clusters to test the algorithms, shown in Figure 1. The k vs variance graph is shown in Figure 2. From the graph we can see the ideal k for the dataset is 4.

3.2 Dataset: Iris

This is one of the the best known databases to be found in the pattern recognition literature. Fisher's paper [1] is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. The k vs variance graph is shown in Figure 3. From the graph we can see the ideal k for the Iris dataset is 3.

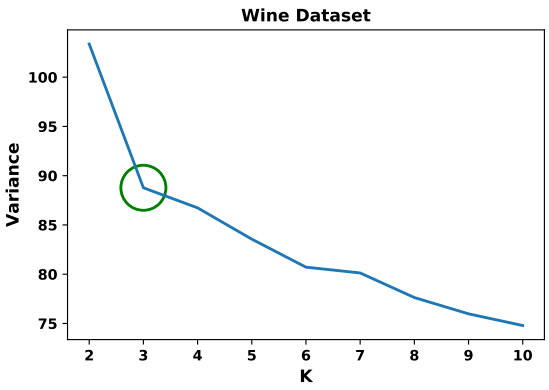

Figure 4: Elbow method for Wine

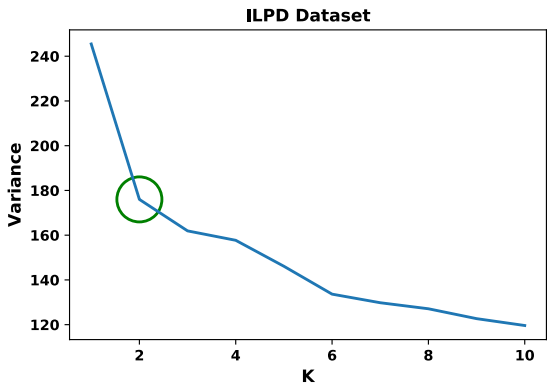

Figure 5: Elbow method for ILPD dataset

3.3 Dataset: Wine

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The k vs variance graph is shown in Figure 4. From the graph we can see the ideal k for the Iris dataset is 3.

3.4 Dataset: ILPD

This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90". The k vs variance graph is shown in Figure 5. From the graph we can see the ideal k for the Iris dataset is 3.

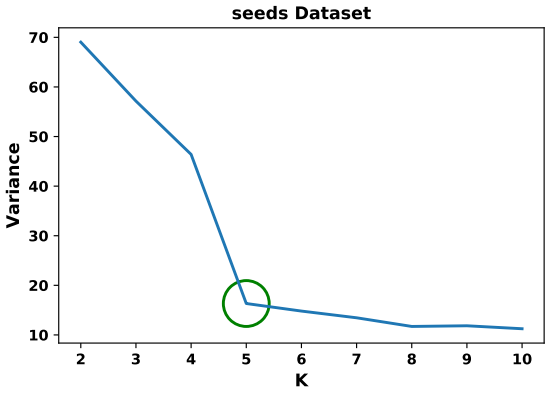

Figure 6: Elbow method for Seeds dataset

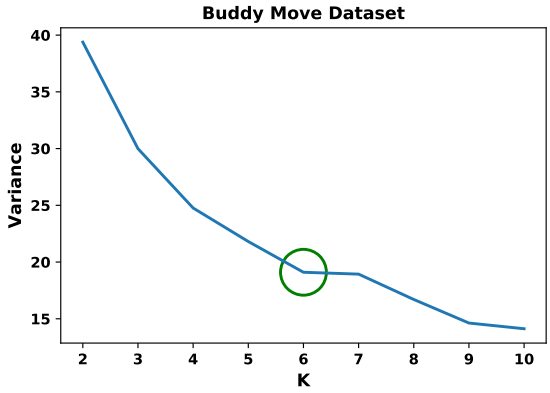

Figure 7: Elbow method for Buddy Move dataset

3.5 Dataset: Seeds

The examined group comprised kernels belonging to different varieties of wheat, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set can be used for the tasks of classification and cluster analysis. The k vs variance graph is shown in Figure 6. From the graph we can see the ideal k for the Iris dataset is 5.

3.6 Dataset: Buddy Move

This dataset was populated from destination reviews published by 249 reviewers of holidayiq.com till October 2014. Reviews falling in 6 categories among destinations across South India were considered and the count of reviews in each category for every reviewer (traveler) is captured. The k vs variance graph is shown in Figure 6. From the graph we can see the ideal k for the Iris dataset is 6.

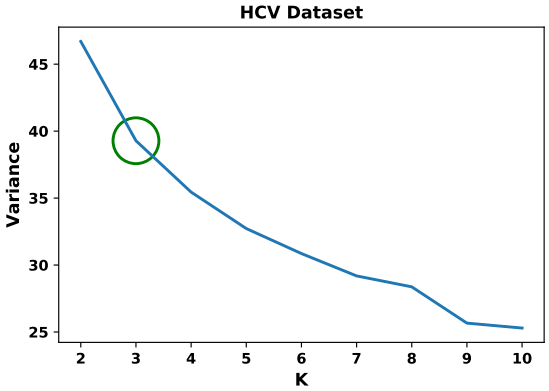

Figure 8: Elbow method for HCV dataset

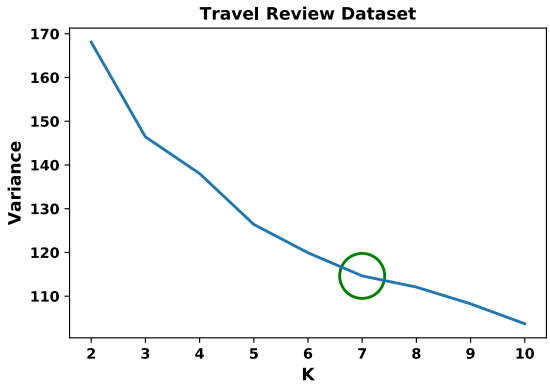

Figure 9: Elbow method for Travel Review dataset

3.7 Dataset: HCV

The data set contains laboratory values of blood donors and Hepatitis C patients and demographic values like age. The dataset was donated by Medical University Hannover (MHH) on 2010. The k vs variance graph is shown in Figure 8. From the graph we can see the ideal k for the Iris dataset is 3.

3.8 Dataset: Travel Review

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. The k vs variance graph is shown in Figure 9. From the graph we can see the ideal k for the Iris dataset is 7.

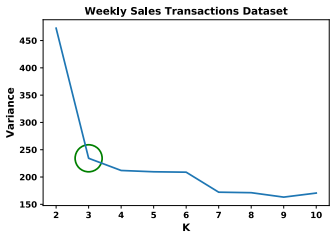

Figure 10: Weekly Sales Transactions dataset

Figure 11: Wholesale Customer dataset

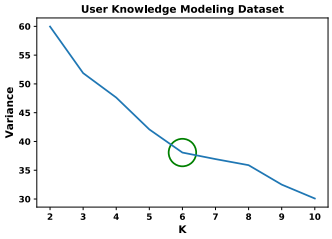

Figure 12: User Knowledge Modeling dataset

3.9 Dataset: Weekly Sales Transactions

The weekly sales transaction dataset contains weekly sales of over 800 items across a year of 52 weeks. Each row represent the information for a product and the countities that were sold in each 52 weeks. The k vs variance graph is shown in Figure 10. From the graph we can see the ideal k for the Iris dataset is 3.

3.10 Dataset: Wholesale Customers

The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories. The k vs variance graph is shown in Figure 11. From the graph we can see the ideal k for the Iris dataset is 8.

3.11 Dataset: User Knowledge Modeling

The dataset is about the users' learning activities and knowledge levels on subjects of DC Electrical Machines. The institution was Faculty of Technology, Department of Software Engineering, Karadeniz Technical University, Trabzon, Turkiye. The dataset was created on October, 2009. The k vs variance graph is shown in Figure 12. From the graph we can see the ideal k for the Iris dataset is 6.

Dataset	**Variance**		**Silhouette Coeff.**		**BCubed Precision**		**BCubed Recall**	
	K-Mean	K-Med.	K-Mean	K-Med.	K-Mean	K-Med.	K-Mean	K-Med.
2D points	10.872	**9.644**	0.538	**0.581**	NA	NA	NA	NA
Iris	**29.321**	30.48	**0.5**	0.489	0.816	**0.834**	0.823	**0.839**
Wine	**88.73**	99.626	**0.3**	0.291	**0.906**	0.877	**0.907**	0.881
ILPD	**175.949**	180.148	**0.249**	0.238	0.591	0.591	0.498	**0.502**
Seeds	**53.026**	53.381	**0.299**	0.286	NA	NA	NA	NA
Buddy Move	**66.945**	69.098	0.262	**0.273**	NA	NA	NA	NA
HCV	**128.776**	131.382	**0.205**	0.136	NA	NA	NA	NA
Travel Review	**318.804**	333.284	**0.143**	0.129	NA	NA	NA	NA
Weekly Sales	**367.785**	416.43	**0.615**	0.609	NA	NA	NA	NA
Wholesale	42.465	**40.934**	**0.287**	0.249	NA	NA	NA	NA
User Knowledge	**93.843**	97.489	**0.189**	0.184	NA	NA	NA	NA

Table 2: Result comparison between k-mean and k-medoid algorithms on different datasets.

4 Result Analysis and Discussion

From the Result Comparison in table 2, we can see that both K-mean and k-medoid algorithms perform almost similar to all the data set. Between this two algorithms, K-means performs slightly better in most of the datasets. This is because of the characteristics of these two classifier. K-mean algorithm takes the mean of the cluster dataset as its new position. So if the clusters are well-shaped and there is very few outliers then k-mean can divided the clusters more perfectly. On the other hand k-medoids randomly takes one of the data points So, if there is no datapoint exist that will cover uniformly all the datapoints within the cluster then it will hamper the performance of k-medoid algorithm. Moreover, as we have used CLARANS for k-medoid technique, in each iteration it randomly checks on 20% of the non-medoid data to replace the current medoid. So the performance of the algorithm has a dependency on randomness too. Secondly, K-mean randomly choose the initail centroids. If the choice of the initial centroids are bad, the the clusters formed can be hampered.

As the 2D point dataset has a high clustering tendency (see Figure 1) it has low variance and high silhouette coefficient value. For Iris, Wine and ILPD datasets, the ground truths are available. Using these ground truth we have calculated the BCubed precision and recall. From table 2 we can see both iris and wine dataset has high precision and recall while ILPD has slightly lower. But still the silhouette coefficient of ILPD is better than most of the dataset. For the rest of the datasets, the ground truth was not available, so we have only calculated the variance and silhouette coefficient as performance analysis and showed them in table 2.

5 Conclusion

The k-means clustering algorithm has several drawbacks, such as reliance on Euclidean distance, susceptibility to outliers, and obtaining centroids that are not representative of real data points. These are resolved using K-medoids and its variations. However, K-medoids is harder to implement and runs slower than the k-means method. This has prompted researchers to develop methods like CLARA, and CLARANS, which increases the speed of clustering at the cost of optimal assignment of clusters.

References

[1] Ronald A Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2 (1936), pp. 179–188.

[2] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[3] Raymond T. Ng and Jiawei Han. "CLARANS: A method for clustering objects for spatial data mining". In: *IEEE transactions on knowledge and data engineering* 14.5 (2002), pp. 1003–1016.

[4] API Reference. *Scikit Learn*. URL: `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics`. (accessed: 22.02.2021).