International Journal of Engineering & Technology, 7 (4) (2018) 4766-4768

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet. v7i4.21472
Research paper

A Comparison of K-Means Clustering Algorithm and CLARA Clustering Algorithm on Iris Dataset

Tanvi Gupta[1]*, Supriya P. Panda[2]

[1]Research Scholar Manav Rachna International Institute of Research& Studies
[2]ProfessorManavRachna International Institute of Research&Studies
**Corresponding author E-mail:Tanvigupta.fet@mriu.edu.in*

Abstract

K-Means Clustering is the clustering technique, which is used to make a number of clusters of the observations. Here the cluster's center point is the 'mean' of that cluster and the others points are the observations that are nearest to the mean value. However, in Clustering Large Applications (CLARA) clustering, medoids are used as their center points for cluster, and rest of the observations in a cluster are near to that center point .Actually in this, clustering will work on large datasets as compared to K-Medoids and K-Means clustering algorithm, as it will select the random observations from the dataset and perform Partitioning Around Medoids (PAM) algorithm on it. This paper will state that out of the two algorithms; K-Means and CLARA, CLARA Clustering gives better result.

Keywords:*K-Means Clustering; CLARA Clustering; K-Medoids Clustering; PAM Algorithm; Iris Dataset.*

1. Introduction

Clustering is the technique to partition the data according to the characteristics .It says that the data which are similar in nature are in one cluster. Actually clustering is the unsupervised learning algorithm and is mainly used for data analysis and data mining.

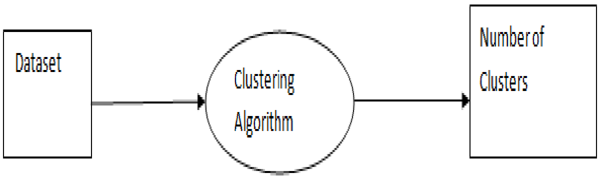


Fig. 1: Clustering Concept.

Fig.1 depicts the concept of clustering, which reflects that the dataset when passes through the clustering algorithm gives the final result as the number of clusters.

There are four categories of clustering that are:

1) Partition based algorithm
2) Hierarchal based algorithm
3) Density based algorithm
4) Grid based algorithm

The variations in these categories are as following:

i) The procedure that is used for partitioning the dataset,
ii) In constructing the clusters use of the thresholds value, and
iii) Manner of clustering.

1.1. Partition Based Algorithm

This algorithm partitions the dataset according to the center point which can be mean, medoid, mode etc. into number of clusters. But, the drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor outcome due to the overlapping of data points [1]. Also, it uses a number of greedy heuristic schemes of iterative optimization.

Some of the partition based algorithm is K-Means algorithm, K-Medoids algorithm, Partitioning around Medoids (PAM) algorithm, and Clustering Large Applications (CLARA).

1.1.1. K-Means Clustering

K-means Clustering is an unsupervised learning algorithm. The basic idea of this algorithm is to define k clusters using k centers i.e., one for each, respectively. There are certain steps for this algorithm that are as follows:

Let X = $\{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$\boldsymbol{v}_i = (\boldsymbol{1} / \boldsymbol{c}_i) \sum_{j=1}^{c_i} \boldsymbol{x}_i \quad ,$$

where, 'c_i' represents the number of data points in the ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages:

1) It is easier to understand, fast and robust,
2) It works just on numeric values,
3) The cluster has convex shapes [2], and

4) It gives the best result when dataset are distinct or well separated from each other[3].

Disadvantages:

1) The learning algorithm needs the prior information of number of cluster to be formed.
2) If there is an overlapping of data, K-means will not be able to solve as there are clusters formed at the same space.
3) This algorithm uses a Euclidean distance whose measures can unequally weight underlying factors.
4) It is also not able to handle outliers and noisy data,
5) This learning algorithm is not invariant to non-linear transformation i.e., with different representation of data we get different results,
6) This algorithm will not work for non-linear data,
7) If we choose the cluster center randomly, this will not lead to the actual result.

1.1.2. K-Medoid Clustering Algorithm

K-Medoid Clustering algorithm is like the K-means clustering algorithm but the difference lies in the center point .Medoids is the center point in K-Medoids Clustering instead of means. Also, K-Medoids use Manhattan as a distance metric instead of Euclidean, which is used by K-Means. Due to this, K-Medoids is more robust to outliers and noises.

K-Medoid Clustering Algorithm [3] is as follows:

Input: K_y: the number of clusters, D_y: a data set containing n objects.

Output: A set of K_y clusters.

Algorithm:

1) Randomly select K_y as the Medoids for ‘n’ data points.
2) Find the closest Medoids by calculating the distance between data points n and Medoids k and map data objects to that.
3) For each Medoids ‘m’ and each data point ‘o’ associated to ‘m’, do the following:

a) Swap ‘m’ and ‘o’ to compute the total cost of the configuration, next
b) Select the Medoids ‘o’s with the lowest cost of the configuration.

4) If there is no change in the assignments, repeat steps 2 and 3 alternatively.

Advantages:

1) It is more robust to outliers and noises.
2) It uses Manhattan distance for calculating the dissimilarity among the nodes, which is more robust than Euclidean distance.

1.1.3. CLARA Clustering

Clustering Large Applications is anticipated (Kaufman and Rousseeuw, 1990). It is an extension of K-Medoids Clustering algorithm, which uses the sampling approach to handle the large datasets.

2. Literature Survey

In [4], the author named Tagaram Soni Madhulatha, explains the concept of clustering with the difference between K-Means and K-Medoids clustering .Author says that clustering is an unsupervised form of learning, which helps to partition the data in the clusters using distance measures without any background knowledge.

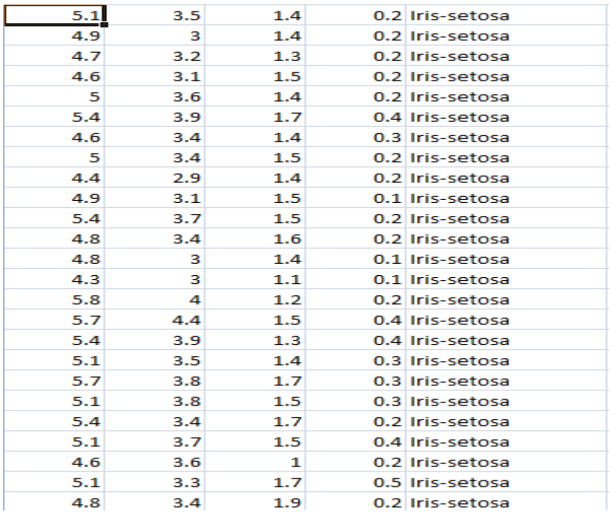
In [5], the authors named T.Velmurugan, T.Santhanam, explained the clustering by saying that it is an unsupervised learning. They say that the clustering is decided based on the type of available data and the purpose for which the clustering is to be done. They say from the experiments they perform K-Medoids is more robust than K-Means clustering in terms of noises and outliers but K-Medoids is good for only small datasets.

In [6], the authors named K. Chitra, D.Maheswari explain the different types of clustering. They say that the clustering is one of the data mining processes. It is an unsupervised learning and is the arrangement of the set of identical objects in one cluster. The authors in this paper compare the different types of clustering like partition-based, hierarchal, grid-based, density –based.

In [7], the author named T.Velmurugan, analyses the performance of the two clustering algorithms using the calculation of distance between the two data points. Also, the computational time is calculated in order to measure the performance of the algorithm. In this paper, K-Means is better than the K-Medoid clustering in terms of efficiency for the application they have chosen.

3. Experimental Setup and Dataset

To show the comparison between K-Means clustering algorithm and CLARA clustering algorithm, R programming is used to cluster plot the graph for both clustering techniques on Iris Dataset.



5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setosa
5.1	3.8	1.5	0.3	Iris-setosa
5.4	3.4	1.7	0.2	Iris-setosa
5.1	3.7	1.5	0.4	Iris-setosa
4.6	3.6	1	0.2	Iris-setosa
5.1	3.3	1.7	0.5	Iris-setosa
4.8	3.4	1.9	0.2	Iris-setosa

Fig. 2: Sample of Iris Dataset.

In Fig. 2 above first attribute is Sepal .Length, second attribute is Sepal.Width, Third attribute is Petal.Length, And Last attribute is Petal.Width. So, according to the attributes defined Iris dataset is the dataset of a flower .This dataset represents the three species of flower that are setosa, versicolor, virginica having different attributes specified above.

There are three types of graphs Fig. 3, Fig. 4, Fig. 5 as shown below:

1) Simple plotting of graph without clustering

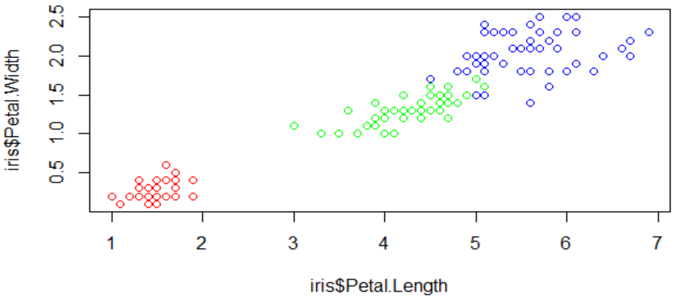


Fig. 3: Graph of Iris Dataset without Clustering Having Three Types of Species (Setosa, Versicolor, Virginica) with Two Attributes of PetalLength and PetalWidth.

2) K-Means cluster plot of Iris Dataset

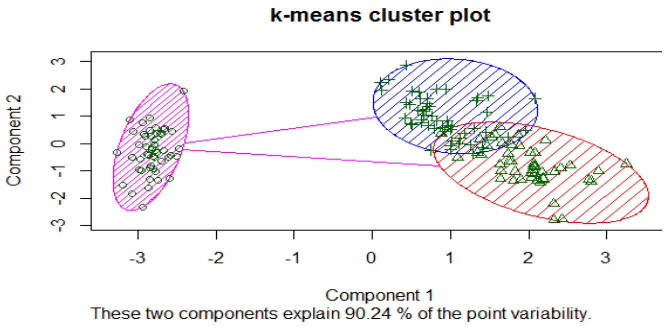


Fig. 4:K-Means Cluster Plot of Iris Dataset Using Two Components: PetalLength and Petal Width.

3) Cluster plot of CLARA clustering on Iris Dataset.

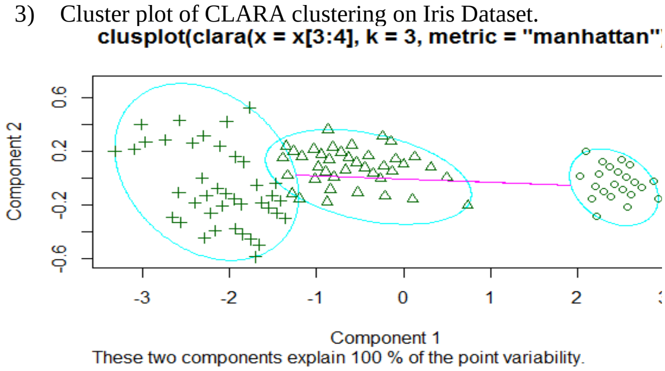


Fig. 5: Clusplot of Clara Clustering on Iris Dataset Using Two Components Petal.Length and Petal.Width.

The above 2 figures show the cluster plot of both the clustering techniques. Fig. 4 uses K-means algorithm with Euclidian distance and Fig.5 uses CLARA clustering with Manhattan distances. From both the figures we conclude that CLARA clustering has more power to detect outliers and noise than K-Means clustering as the point of variability is 100% for CLARA and approx. 90% for K-Means. This proves the CLARA clustering to be more robust as compared to K-means algorithm.

Code and output in R for finding Euclidean distance and Manhattan distance on the Iris Dataset.

```
P<-1:10
> Q<-11:20
> x<-rbind(P,Q)
>distance(x,method="euclidean")
euclidean
31.62278
>distance(x,method="manhattan")
manhattan
100
```

The distance is finding between 1 to 10 rows and 11 to 20 rows.

As the distance here means dissimilarity, so from the above output we observe that the Euclidean distance will detect less dissimilarity than Manhattan.

So, Manhattan distance measure is better than the Euclidean distance measure.

4. Conclusion

This paper is regarding the comparison of K-Means Clustering and CLARA Clustering on Iris Dataset, which are using Euclidean distance and Manhattan Distance as a dissimilarity measure, respectively. After plotting of graphs using the two attributes of dataset that are “Petal. Length” and “Petal. Width”, we can conclude that CLARA Clustering using Manhattan distance is better than K-Means Clustering with Euclidean distance.

References

[1] S. AnithaElavarasi and J. Akilandeswari ,A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems, 2011.

[2] Navneet Kaur, Survey Paper on Clustering Techniques, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, 2013.

[3] Preeti Arora, Deepali , ShipraVarshney,” Analysis of K-Means and K-Medoids Algorithm For Big Data”, International Conference on Information Security & Privacy (ICISP2015), 2015, Nagpur, INDIA.

[4] TagaramSoniMadhulatha, “Comparison between K-Means and K-Medoids Clustering Algorithms”, advancing in computing and information technology, volume 198, pp. 472-481.

[5] T.Velmurugan, T. Santhanum, “Performance Analysis of K-Means and K-Medoids Clustering Algorithms for A Randomly Generated Data Set”, International Conference on Systemics, Cybernetics and Informatics, pp.578-583, 2008.

[6] K. Chitra, Dr. D.Maheswari, “A Comparative Study of Various Clustering Algorithms in Data Mining”, IJCSMC, Vol. 6, Issue 8, pp.109 – 115, 2017.

[7] Dr. T. Velmurugan ,”Efficiency of K-Means and K-Medoids algorithms for clustering arbitrary data points”, International Journal Computer Technology & Applications, Volume 3(5),pp1758-1764, 2012.

[8] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.