Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset

Tanvi Gupta, Supriya P Panda
Department of Computer Science & Engineering
Manav Rachna International Institute of Research and Studies, Faridabad, India
tanvigupta.fet@mriu.edu.in, supriya.fet@mriu.edu.in

Abstract: **This paper is regarding the comparison of two techniques; Clustering Large Applications (CLARA) clustering and K-Means clustering using popular Iris dataset. CLARA clustering and K-Means clustering are the two techniques of "partitioning based" clustering. One considers medoids using random sample data to form a cluster whereas the other considers centroid (means) of the dataset to form a cluster. In this paper, Cluster plot, Silhouette plot and Dunn Index on Iris dataset are shown for both the techniques. These all are used for "cluster validation". The "Silhouette Analysis" is the measurement of an approximated average distance among the clusters. The "Silhouette plot" is the measurement of the closeness of the points in one cluster to the neighboring clusters, whereas the other internal clustering validation measure is the DUNN Index; higher the "Dunn Index" better is the clustering. All these statistical analysis is done in R programming. The final outcome attains that the CLARA clustering stands better than the K-Means clustering**

Keywords: *Cluster plot, Silhouette plot, Dunn Index, CLARA Clustering, K-Means Clustering*

I. INTRODUCTION

Cluster validation is the technique to evaluate the different clustering algorithms' goodness. This helps to avoid in finding the patterns in a random dataset. There are three classes of cluster validation [1, 2, 3]:

1) **Internal cluster validation**: In this validation technique [4] the "goodness" of the clustering algorithm is calculated without any "external" information.
2) **External cluster validation**: This validation technique [4] is used to compare the two clustering techniques for example externally defined class labels. This approach is primarily useful in choosing the right "clustering algorithm" for the right "dataset" using the knowledge of true clustering number in advance.
3) **Relative cluster validation**: This cluster validation [4] is used to evaluate the structure of cluster by altering the values of the unlike parameters for the same algorithm

Here, we are considering internal cluster validation, which consists of two measures: Silhouette coefficient and Dunn Index.

Silhouette coefficient is accustomed to measure the mean distance amongst the clusters. The "Silhouette plot" is the measurement of the closeness of the points in one cluster to the neighboring clusters, and "Dunn Index" is the lowest of this pair-wise distance.

$$intra\,cluster\,compactness = \frac{inter\,cluster\,separ\,ation}{maximumdiameter}$$

Maximum diameter is also referred to as intra-cluster distance.

II. LITERATURE SURVEY

The authors in [7] discusses three partitioning clustering algorithms are stated naming K-Means, K-Medoids and CLARA clustering. These algorithms are explained using R programming and conclude the optimal number of clusters for the partitioning algorithm. This article helps to understand the partitioning clustering in detail using R programming.

Authors[5] gives the key idea that categorizes the methods on the basis of different aspects as partitioning, hierarchical etc. so that it helps in selecting algorithms for any further enhancement and optimization. The author says "unsupervised clustering" methods of data mining are the most active research algorithms.

In [6], different clustering calculations are verified for their clustering superiority using recognized "partitioning algorithms". Here, the comparative study of the three clustering algorithms such as "centroid based K-Means", "representative objects based K-Medoids" and "representative sample based CLARA" are described and observed according to the distance between two objects. The algorithms are associated with respect to their clustering superiority and respective performances based on the practical results. The "total lapsed time" to cluster all the data sets and memory used for each cluster are also determined in milliseconds and kilobytes.

Authors [4] discuss about validation techniques, Dunn Index and Silhouette analysis using graph plot. All these techniques are used to explain which clustering is better. This paper helps to validate the clusters using the above mention techniques in R programming.

In [8], the authors had done the assessment of cluster's tendency of pharmacological dataset to get the Dunn's index value. In this pharmacological data sets were represented by machine-learning-selected molecular descriptors. This paper shows the relationship between Dunn index which is a measure of cluster separability and classification accuracy.

[9] uses the silhouette and sum of squared errors to validate clusters in K-Means clustering. These two techniques are compared for the results on the number of clusters from 2 to 10.

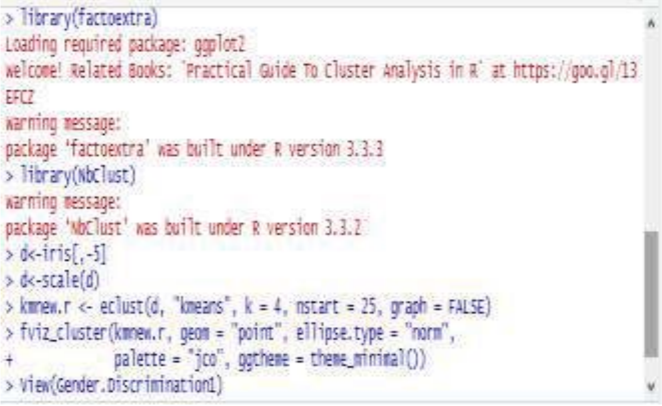
978-1-7281-0211-5/19/$31.00 2019 ©**IEEE**

III. PROPOSED WORK

The proposed work is showing that CLARA clustering is better than the K-Means clustering. R programming is used along with the two validation techniques as Dunn Index, Silhouette Plot.

Cluster plot of K-Means and CLARA clustering

Below is the code for generating cluster plot of K-Means Clustering through R shown in Fig1 and Fig 2 shows the plot.



```
> library(factoextra)
Loading required package: ggplot2
Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13
EFCZ
Warning message:
package 'factoextra' was built under R version 3.3.3
> library(NbClust)
Warning message:
package 'NbClust' was built under R version 3.3.2
> d<-iris[,-5]
> d<-scale(d)
> kmnew.r <- eclust(d, "kmeans", k = 4, nstart = 25, graph = FALSE)
> fviz_cluster(kmnew.r, geom = "point", ellipse.type = "norm",
+              palette = "jco", ggtheme = theme_minimal())
> View(Gender.Discrimination1)
```

Fig. 1. R Code for the K –Means Clustering Cluster plot

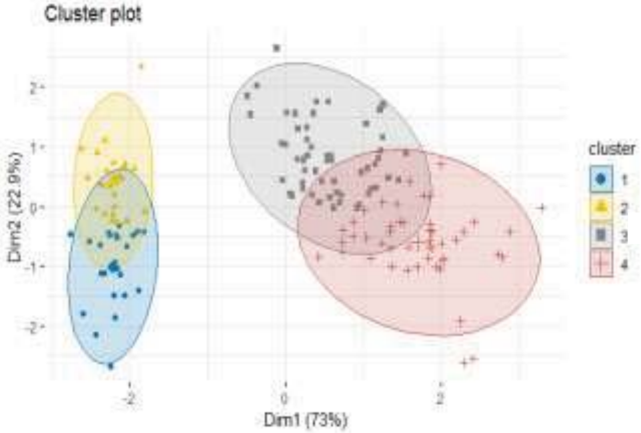
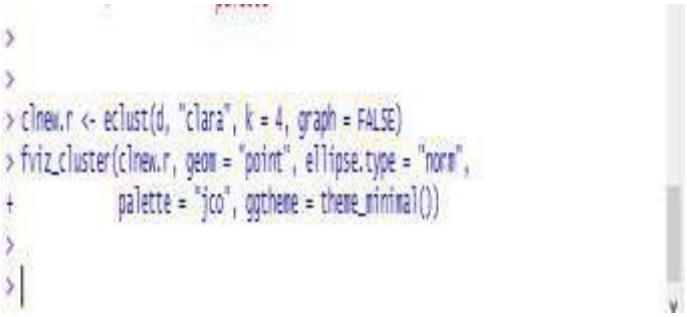


Fig. 2. Cluster plot for K-Means Clustering (Source: Kassambara et.al (2017))

Below is the code for generating cluster plot of CLARA clustering shown in Fig 3and Fig 4 shows the plot.



```
>
>
> clnew.r <- eclust(d, "clara", k = 4, graph = FALSE)
> fviz_cluster(clnew.r, geom = "point", ellipse.type = "norm",
+              palette = "jco", ggtheme = theme_minimal())
>
>|
```

Fig. 3. R Code for the CLARA Clustering Cluster plot

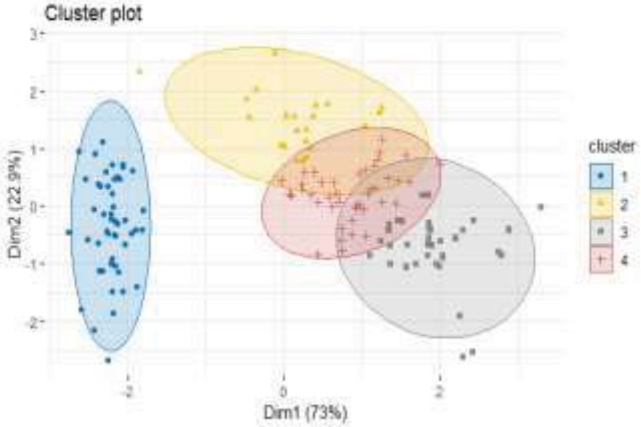
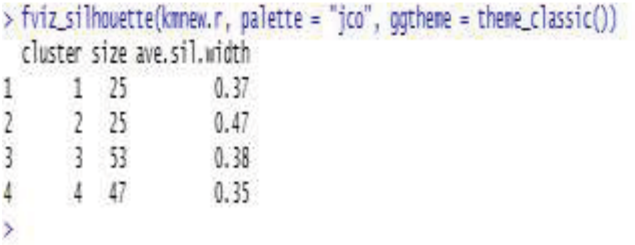


Fig. 4. Cluster plot for CLARA clustering

Silhouette plot of K-Means and CLARA clustering

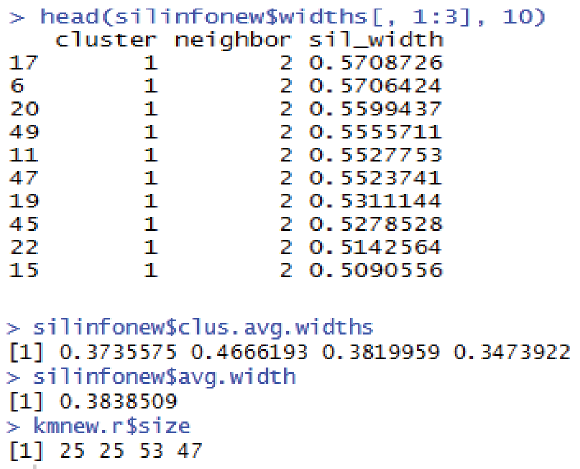
The range of silhouette plot lies between (−1, +1), in which higher is the value ,greater is the matching to its own cluster, else is poorly matched to the nearby clusters. In case most of the objects have a high value, then the clustering configuration turns out to be appropriate.

Below is the code in R for plotting Silhouette of K-Means Clustering and Fig 5 shows the same.



```
> fviz_silhouette(kmnew.r, palette = "jco", ggtheme = theme_classic())
  cluster size ave.sil.width
1       1   25          0.37
2       2   25          0.47
3       3   53          0.38
4       4   47          0.35
>
```

```
> silinfonew <- kmnew.r$silinfo
> names(silinfonew)
```



```
> head(silinfonew$widths[, 1:3], 10)
   cluster neighbor sil_width
17       1        2 0.5708726
6        1        2 0.5706424
20       1        2 0.5599437
49       1        2 0.5555711
11       1        2 0.5527753
47       1        2 0.5523741
19       1        2 0.5311144
45       1        2 0.5278528
22       1        2 0.5142564
15       1        2 0.5090556

> silinfonew$clus.avg.widths
[1] 0.3735575 0.4666193 0.3819959 0.3473922
> silinfonew$avg.width
[1] 0.3838509
> kmnew.r$size
[1] 25 25 53 47
```

Fig. 5. code in R for plotting Silhouette of K-Means Clustering

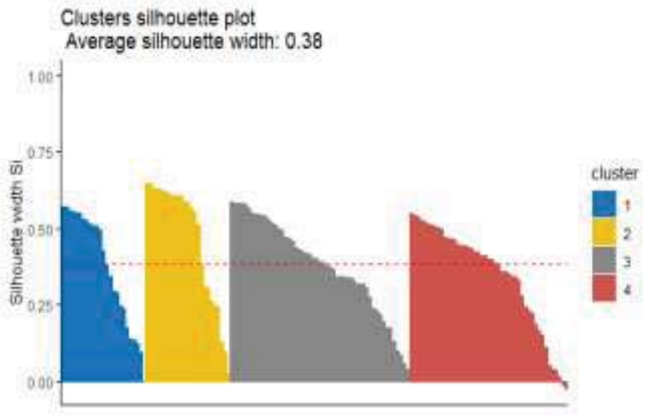
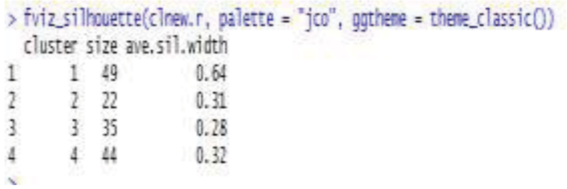
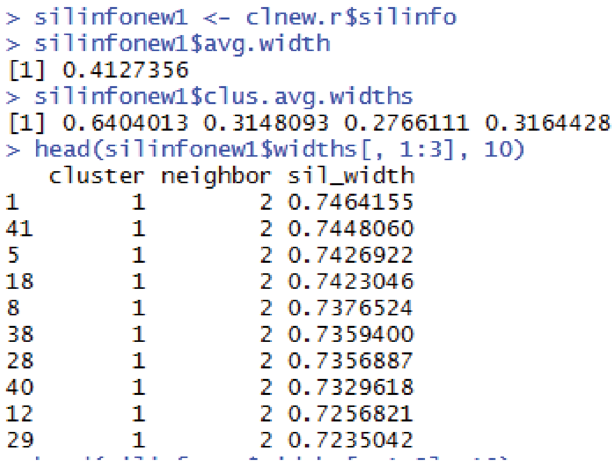


Fig. 6. Silhouette plot of K-Means Clustering (Source: Kassambara et.al (2017))

Below is the code in R, which will help to plot Silhouette of CLARA Clustering and Fig 7 shown below represents the same.



```
> fviz_silhouette(clnew.r, palette = "jco", ggtheme = theme_classic())
  cluster size ave.sil.width
1       1   49          0.64
2       2   22          0.31
3       3   35          0.28
4       4   44          0.32
```



```
> silinfonew1 <- clnew.r$silinfo
> silinfonew1$avg.width
[1] 0.4127356
> silinfonew1$clus.avg.widths
[1] 0.6404013 0.3148093 0.2766111 0.3164428
> head(silinfonew1$widths[, 1:3], 10)
   cluster neighbor sil_width
1        1        2 0.7464155
41       1        2 0.7448060
5        1        2 0.7426922
18       1        2 0.7423046
8        1        2 0.7376524
38       1        2 0.7359400
28       1        2 0.7356887
40       1        2 0.7329618
12       1        2 0.7256821
29       1        2 0.7235042
```

Fig. 7. Code in R to plot Silhouette of CLARA Clustering

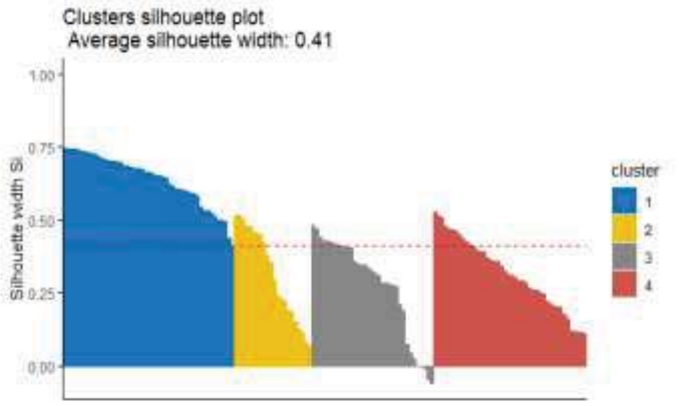
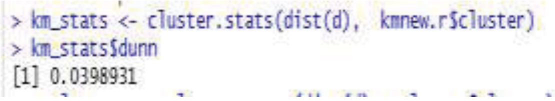


Fig. 8. Silhouette plot of CLARA Clustering

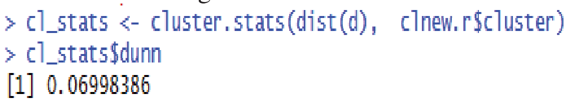
Dunn Index of K-Means and CLARA Clustering

K-Means Clustering DUNN index



```
> km_stats <- cluster.stats(dist(d),  kmnew.r$cluster)
> km_stats$dunn
[1] 0.0398931
```

CLARA Clustering DUNN index



```
> cl_stats <- cluster.stats(dist(d),  clnew.r$cluster)
> cl_stats$dunn
[1] 0.06998386
```

Fig 9: R Code for calculating the DUNN index

For a given assignment of clusters, higher the Dunn index better is the clustering.

Table I: Analysis on Iris Dataset for Two Clustering Techniques

CLUSTERING TECHNIQUES	SILHOUETTE WIDTH	DUNN INDEX
CLARA CLUSTERING	0.41	0.06998386
K-MEANS CLUSTERING	0.38	0.0398931

Table I above shows the results for the two clustering techniques that are CLARA and K-Means. As, we know that, the values Silhouette width and Dunn Index for the techniques are greater than the other respectively then that technique is better than the other. So, the CLARA clustering is better than K-Means Clustering.

IV. CONCLUSIONS

From the graphs and survey, it is concluded that Silhouette measure and Dunn Index can validate the clustering, which tells CLARA clustering is better than the K-Means Clustering. Silhouette analysis says that the graph of CLARA clustering is tightly coupled than the K-Means Clustering graph, which shows that how well an observation is clustered and shows the dissimilarity clearly among the clusters. Furthermore, Dunn Index also states that CLARA is better than K-Means as higher Dunn Index indicates better clustering. In future, other measures can be finding to compare the two clustering algorithms.

REFERENCES

[1] Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta (2008), “ClValid: An R Package for Cluster Validation.” *Journal of Statistical Software,volume* 25 ,issue no. 4,pp-1–22.

[2] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, Azam Niknafs (2014), “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* , volume 61, issue 6,pp-136.

[3] Sergios Theodoridis ,Konstantinos ,Koutroumbas (2008), *Pattern Recognition*. 2nd ed., Academic Press.

[4] Kassambara(2017),Articles- “Cluster Validation Essentials, Cluster Validation Statistics :Must know Methods”, Statistical tools for high-throughput data analysis. Available: http://www.sthda.com/english/articles/29-cluster-validation-essentials/97-cluster-validation-statistics-must-know-methods/.

[5] Kaur Sonamdeep, Chaudhary Sarika, Bishnoi Neha,(2015) “A Survey: Clustering Algorithms in Data Mining”, International Journal of Computer Applications, Innovations in Computing and Information Technology,pp:12-14.

[6] Kelde Dharmendra , Nair Pramod (2016), “A Framework for Comparative Study of K-Mean, K-Medoid and Clara” International Journal of Engineering Science and Computing, Volume 6 Issue No. 7.

[7] Kassambara(2018),“Partitional Clustering in R: The Essentials”,[online].Available:https://www.datanovia.com/en/courses/partitional-clustering-in-r-the-essentials/.

[8] Oscar Miguel Rivera-Borroto ,Monica Rabbasa ,Ricardo Del(2012) ,”Dunn’s index for cluster tendency assessment of pharmacological datasets”, Canadian Journal of Physiology and Pharmacology 90(4):425-33.

[9] Tippaya Thinsungnoena*, Nuntawut Kaoungkub , Pongsakorn Durongdumronchaib , Kittisak Kerdprasopb , Nittaya Kerdprasopb(2015), “The Clustering Validity with Silhouette and Sum of Squared Errors”, International Conference on Industrial Application Engineering,pp-44-51.