Clustering: Measuring Cluster Goodness

Lecture/Practical-20

20/09/2021

Measuring Cluster Goodness

- Clustering models need to be evaluated.
- There are two methods for measuring cluster goodness.

➢Silhouette Method

➢ Psuedo-F statistic

- Sum of Squares Error (SSE) is also a good measure of cluster quality.

- Any measure of cluster goodness , or cluster quality should address the concept of cluster separation as well as cluster cohesion. Cluster separation represents how distant the clusters are from each other. Cluster cohesion refers to how tightly related the records within the individual clusters.
- Sum of Squares Error (SSE) is a good measure of cluster quality. However, by measuring the distance between each record and its cluster center, SSE accounts only for cluster cohesion and does not account for cluster separation.
- SSE will monotonically decrease as the number of clusters increases which is not a desired property of a valid measure of cluster goodness.

Silhouette Method

- Silhouette is a characteristic of each data value and is defined as follows.
- For each data value i,

$$\text{Silhouette}_i = (b_i - a_i)/\max(b_i, a_i)$$

where a_i is the distance between the data value and its cluster center and b_i is the distance between the data value and the next closest cluster center

Silhouette Value

- It is used to assess how good the cluster assignment is for that particular point.
- A positive value indicates that the assignment is good with higher values being better than lower values.
- A value which is close to 0 is considered to be a weak assignment as the observation could have been assigned to the next closest cluster with limited consequence.
- A negative silhouette value is considered to be misclassified, as assignment to the next closest cluster would have been better.

Average Silhouette Value

- Taking the average silhouette value over all records yields a useful measure of how well the cluster solution fits the data.
- Interpretation

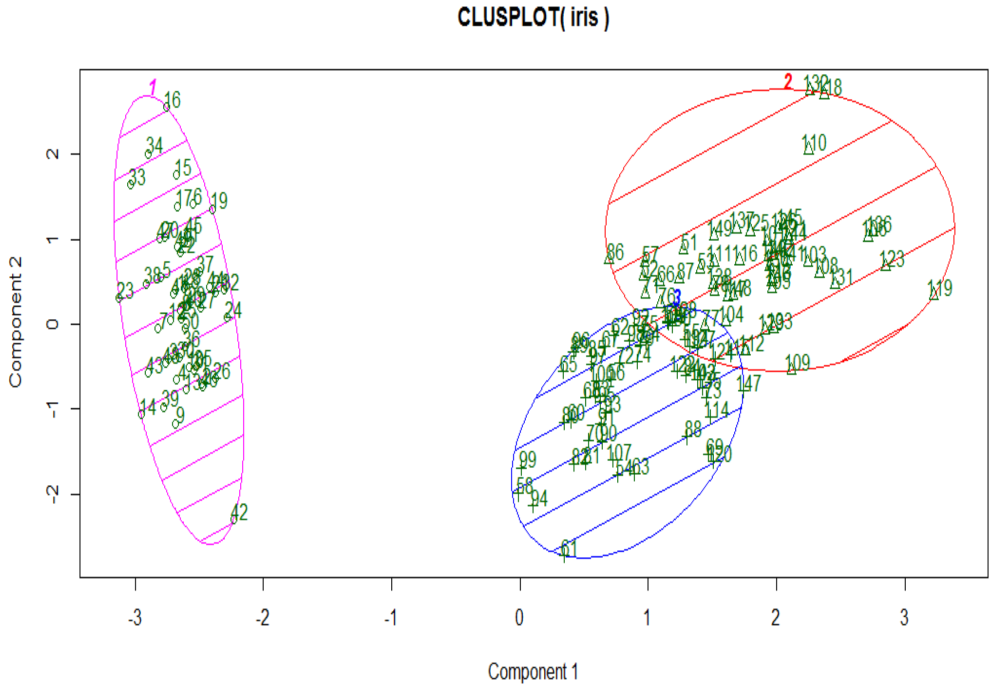
➢0.5 or better: Good evidence of reality of the clusters in the data.

➢0.25-0.5: some evidence of reality of clusters in the data.

➢Less than 0.25: Scant evidence of cluster reality

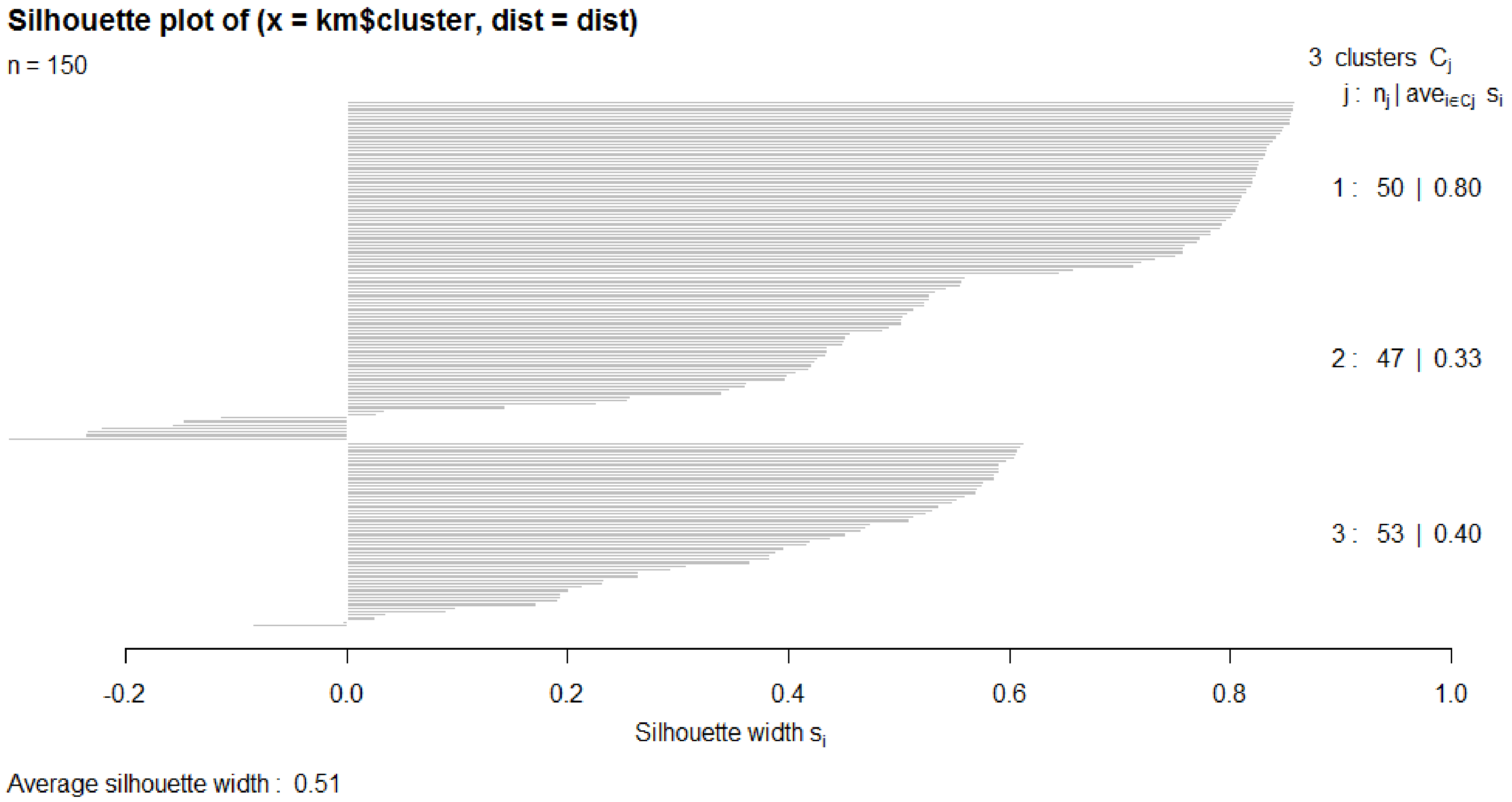
Plotting cluster plot (k=3)

- clusplot(iris, km$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)

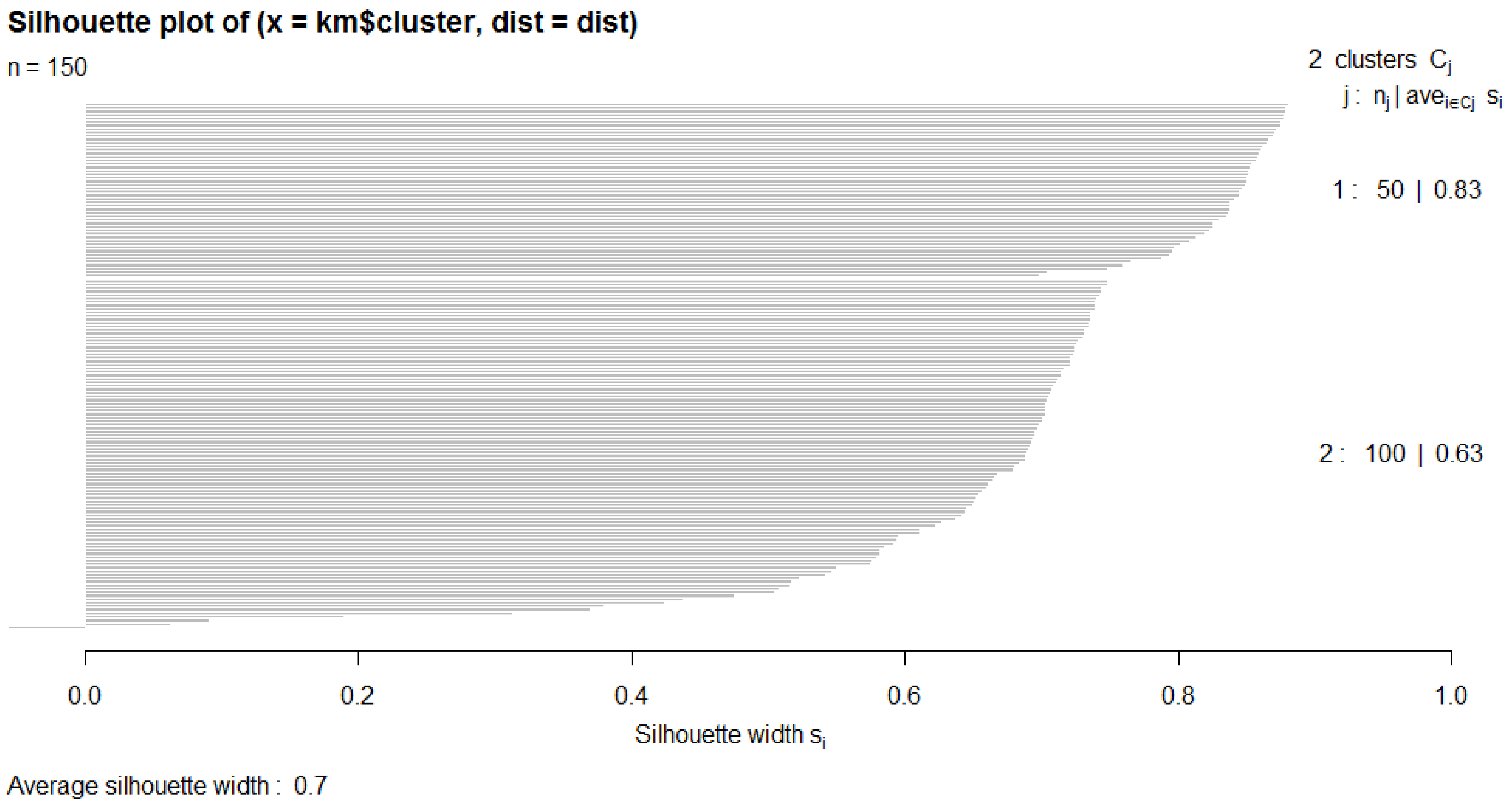


These two components explain 95.08 % of the point variability.

Silhouette Plot with average values



Silhouette Analysis(k=2)



Pseudo-F statistic

- Pseudo-F statistic can be considered as one of the main method for determining the number of clusters.
- It compares the between-cluster to the within-cluster sum-of-squares.
- $Pseudo - F = \frac{\left[\text{between–cluster–sum–of–squares}\big/_{(k-1)}\right]}{\left[\text{within–cluster–sum–of–squares}\big/_{(n-k)}\right]}$
- where k the number of clusters and n the number of observations.
- Large Pseudo-F statistic indicates distinct clusters or peaks in the pseudo F statistic are indicators of greater cluster separation.

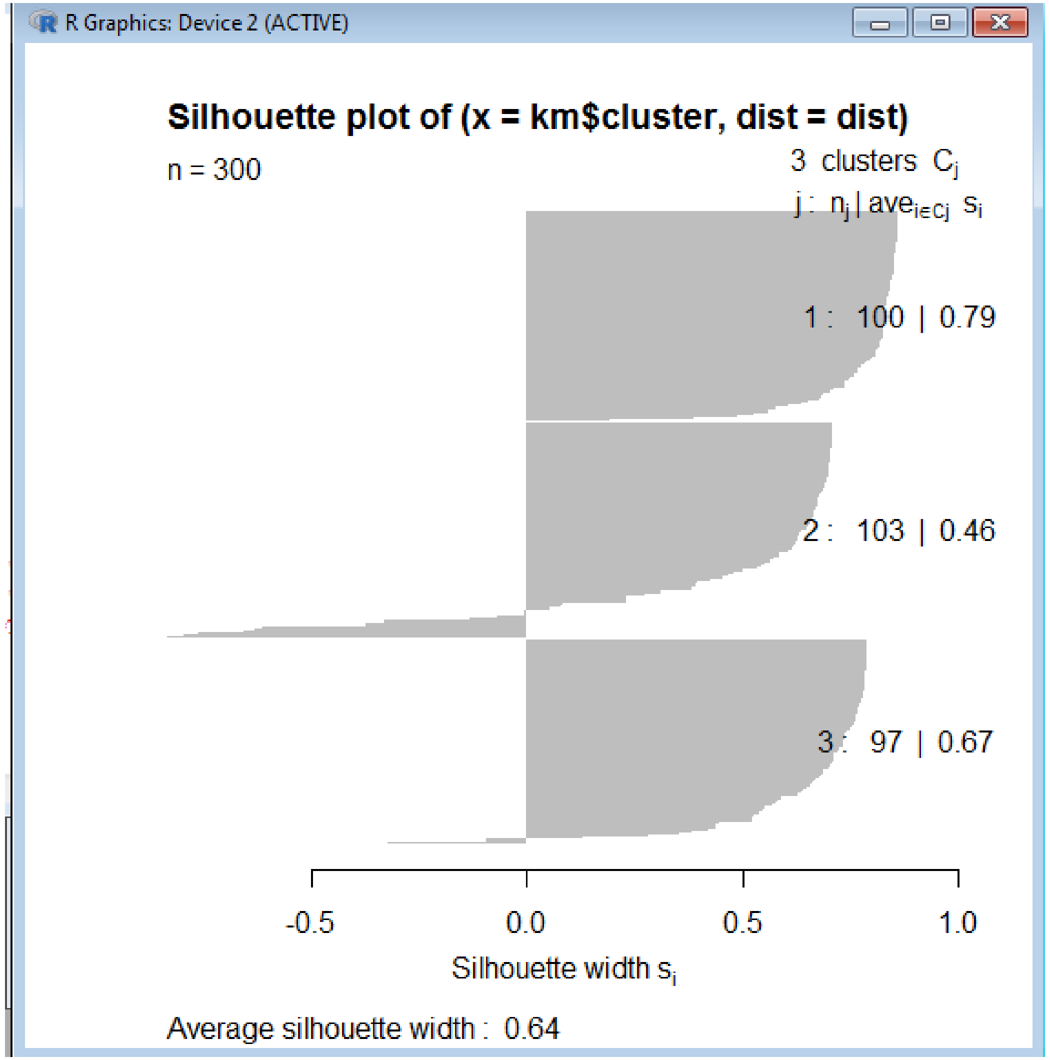
Using pseudo-F to select optimal number of clusters

- Use a clustering algorithm to develop a clustering solution for a variety of values of k
- Calculate the pseudo-F statistic and p-value for each candidate, and select the candidate with smallest p-value as the best clustering solution.
- Note : It has been written that the best clustering model is the one with largest value of pseudo-F. This is Not always correct. One must account for different d.f. (k-1) and (n-k) for each model.

Silhouette values and plot: R zone(income data)

- dist<- dist(income, method ="euclidean")
- sil<-silhouette(km$cluster, dist)
- plot(sil)

Silhouette values and plot



R zone: pseudo-F and p-value(income data)

- library(clusterSim)
- n<-dim(income)[1]
- psF1<-index.G1(income,cl=km$cluster)
- psF1
- #The hypothesis being tested are the following
- #H0:There are no clusters in the data
- #H1:There are k clusters in the data.
- pf(psF1,3,n-3)
- #p value is not rejecting the Null Hypothesis
- psF2<-index.G1(income,cl=km1$cluster)
- psF2
- pf(psF2,2,n-2)