Hierarchical and K-Means

Belias Michael

30 March 2016

Contents

1 Introduction

Cluster Analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). This definition is not so precise that is why *clustering* can be approached by a variety of methods :

- Centroid Models :
  - K-means
  - K-Medians
- Connectivity models :
  - Hierarchical Models
- Distribution models:
  - Expectation Maximization
- Density models:
- Subspace models:
- Group models:
- Graph-based models:

2 The Data

Our data came from the UCI Machine Learning repository, the name of the Data-Set is seed and it contains values of 3 kinds of Wheat Seeds Kama, Rosa and Canadian.

Table 1: 10 first observations in our Data-Set (continued below)

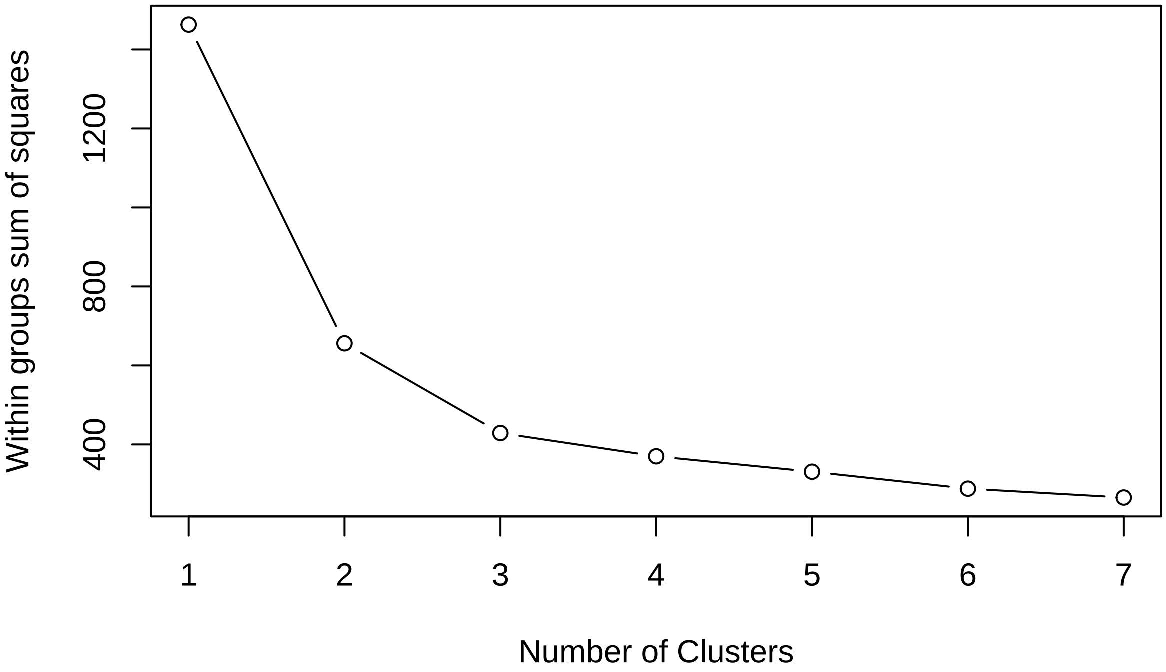
area	perimeter	compactness	kernel.length
15.26	14.84	0.871	5.763
14.88	14.57	0.8811	5.554
14.29	14.09	0.905	5.291
13.84	13.94	0.8955	5.324
16.14	14.99	0.9034	5.658
14.38	14.21	0.8951	5.386
14.69	14.49	0.8799	5.563
14.11	14.1	0.8911	5.42
16.63	15.46	0.8747	6.053
16.44	15.25	0.888	5.884

kernel.width	asymmetry	groove.length	species
3.312	2.221	5.22	Kama
3.333	1.018	4.956	Kama
3.337	2.699	4.825	Kama
3.379	2.259	4.805	Kama
3.562	1.355	5.175	Kama
3.312	2.462	4.956	Kama
3.259	3.586	5.219	Kama
3.302	2.7	5	Kama
3.465	2.04	5.877	Kama
3.505	1.969	5.533	Kama

A more detailed explanation of the variables is reported as follow:

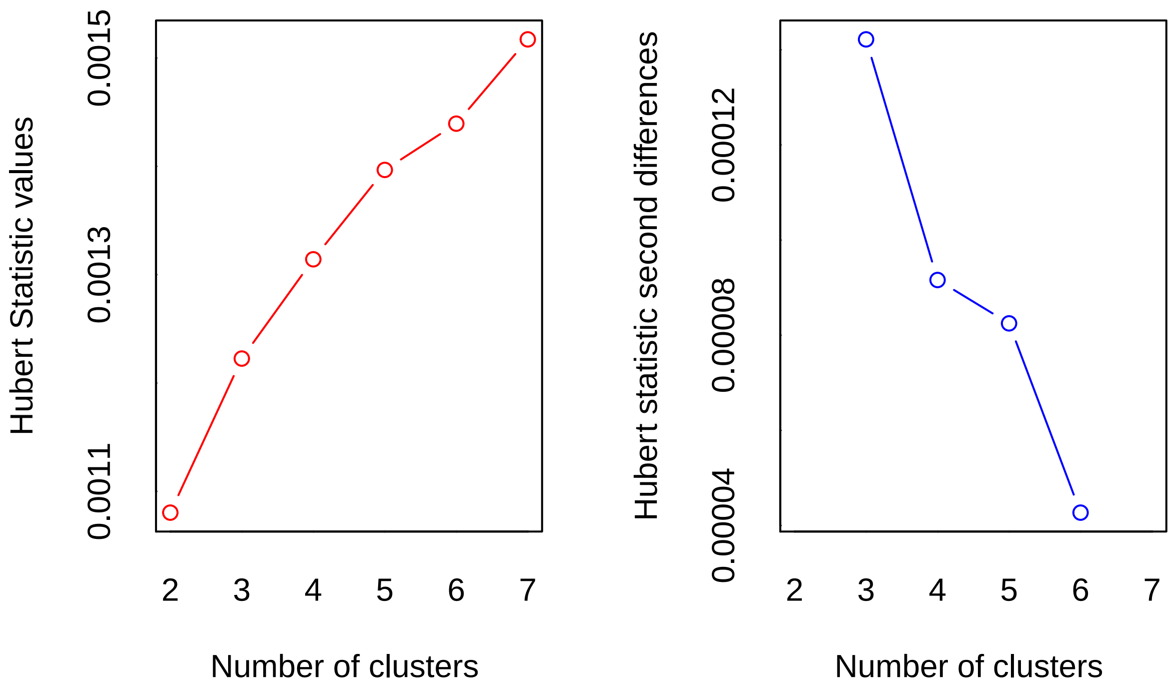
1. area ,
2. perimeter ,
3. compactness ($C = 4\pi A/P^2$),
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.
8. wheat species

The first question we want to answer is how many clusters shall we keep? A nice approach is to check the within group sum of squares for each k (number of Centroids), plot them and with the *elbow method* spot the point where the sum of squares is not decreasing quickly. We can easily build a function for that.

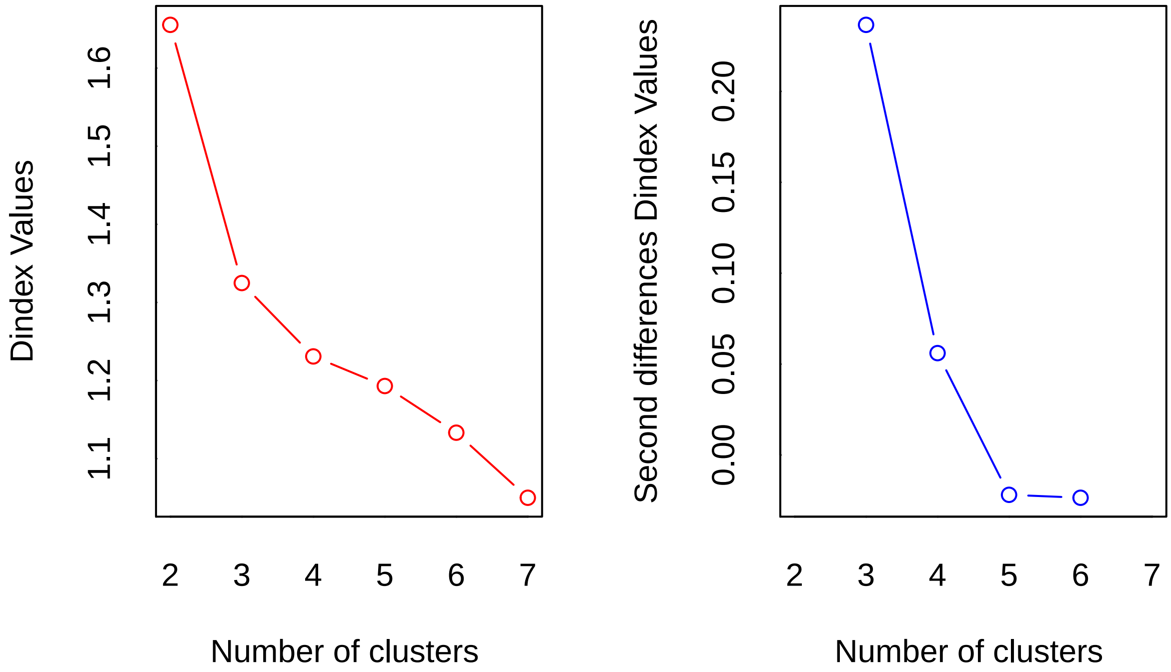


With the elbow method we can easily suggest the use of 3 centroids.

There is also the **NbClust** function in the **NbClust** library.

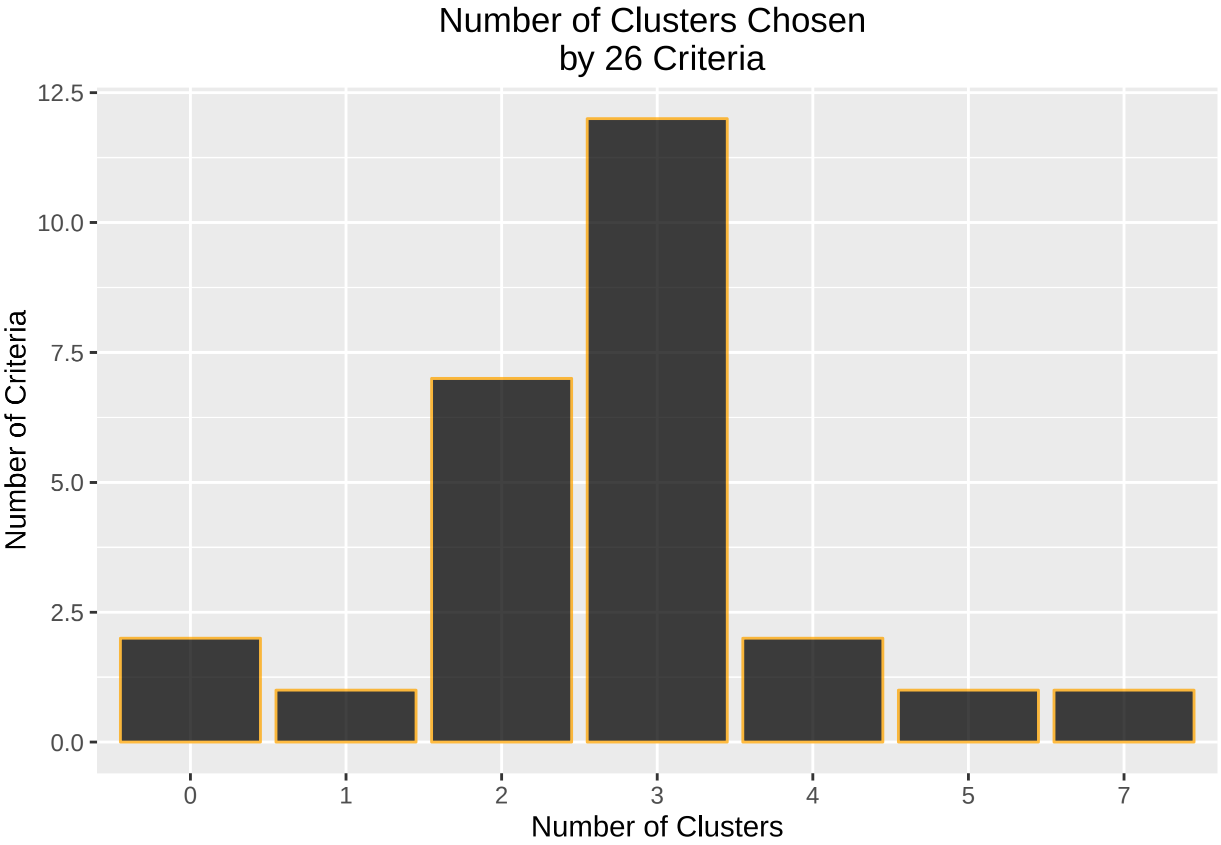


```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds
                significant increase of the value of the measure i.e the significant pea
                index second differences plot.
```



```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak
                second differences plot) that corresponds to a significant increase of t
                the measure.

*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 12 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

0	1	2	3	4	5	7
2	1	7	12	2	1	1

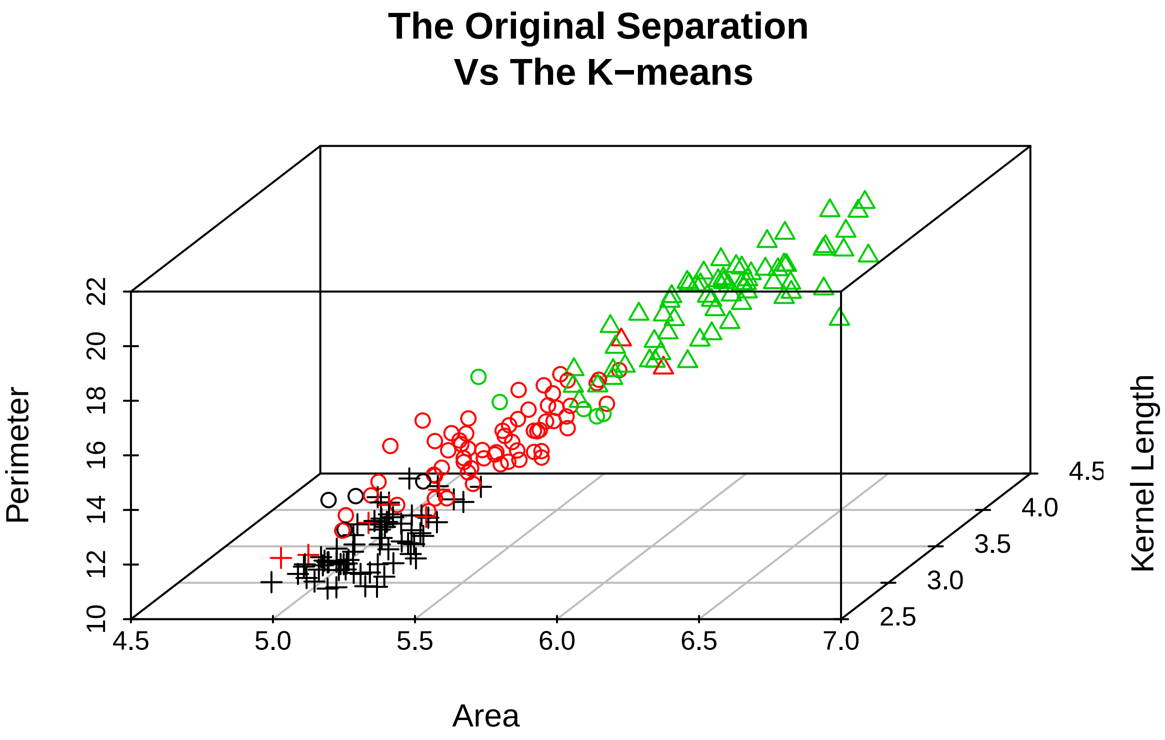


So if we take into consideration the 26 Criteria and pick the one with the most hits we conclude aka. 3 means (of course in our case we knew that the species were 3).

Last step is to scale (Standardise the Data) the Data-Set in order to get better prediction.

The function for the k-means in R is *kmeans*.

```
           1  2  3
Canadian   4  0 66
Kama      62  2  6
Rosa       5 65  0
```



```
       ARI
0.7732937
```

3 K-means Cluster Analysis

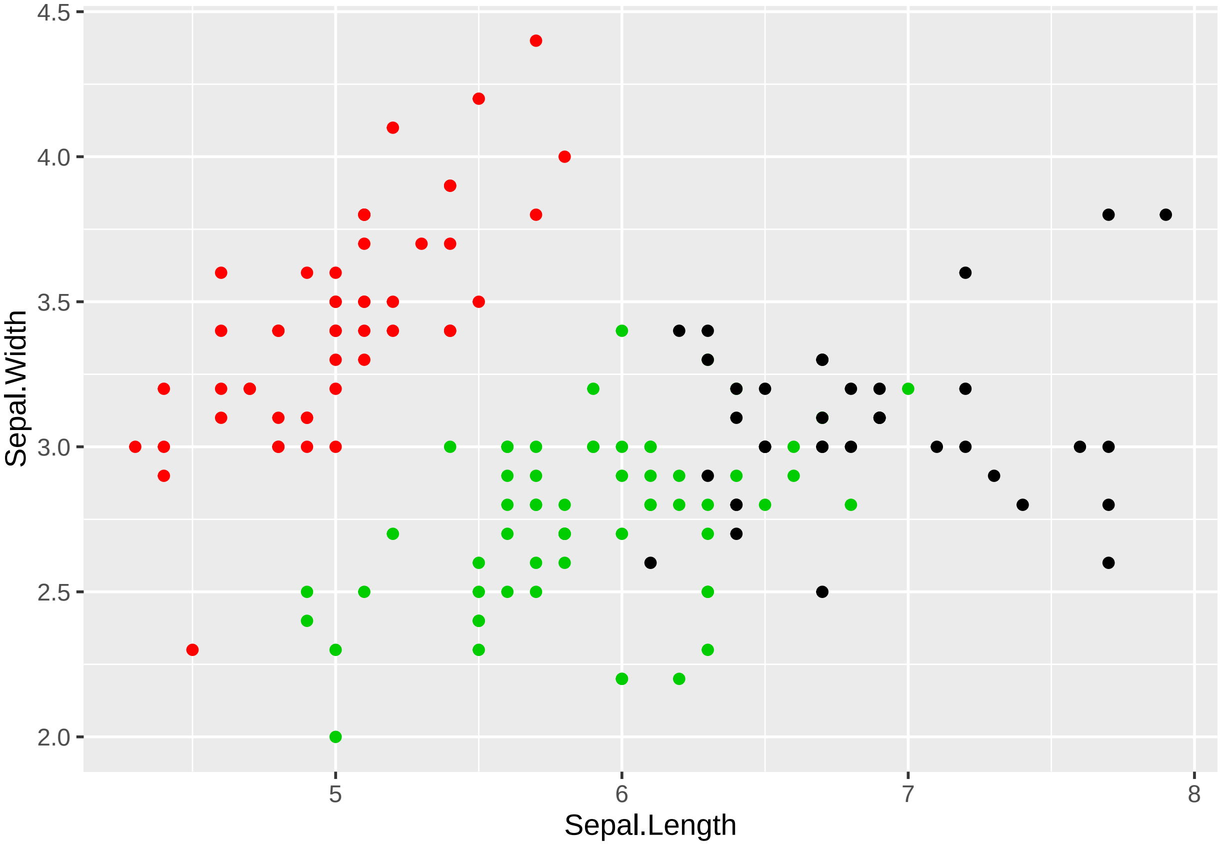
Distance matrix for the means of the analysis variables.

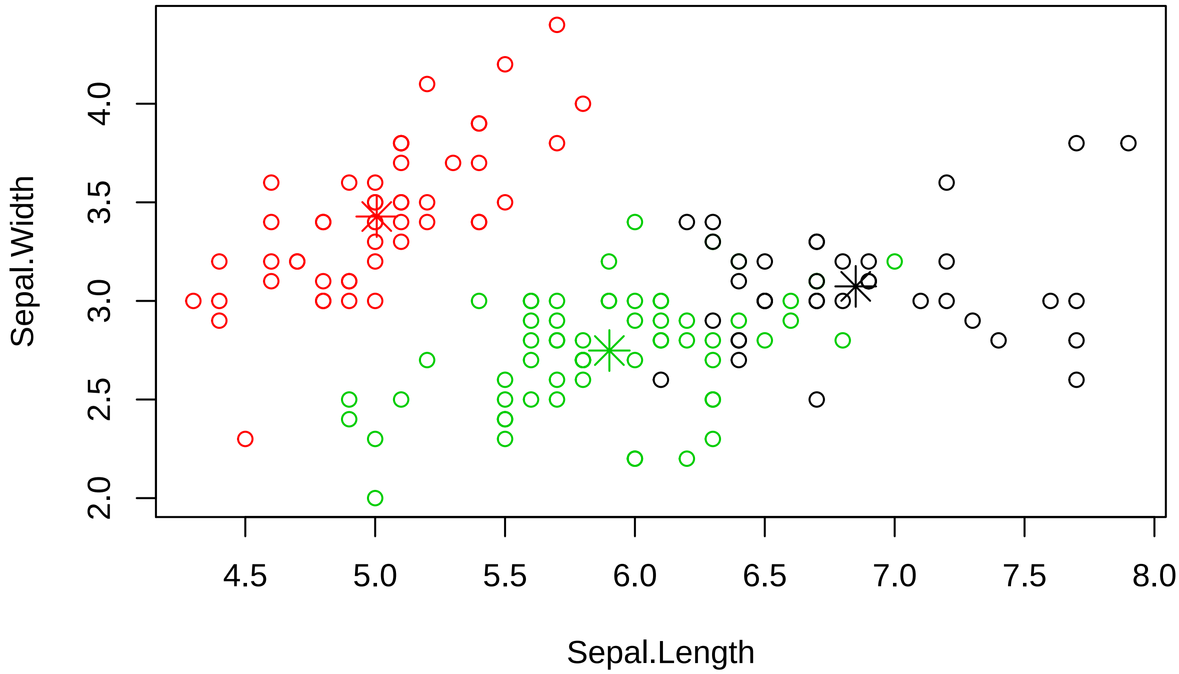
4 The k-Medoids Clustering

```
set.seed(8953)
library(pander)
iris2 <- iris

iris2$Species <- NULL

kmeans.result <- kmeans(x = iris[,-5],centers =  3)

table(iris$Species, kmeans.result$cluster)
```

```
##
##                1  2  3
##   setosa       0 50  0
##   versicolor   2  0 48
##   virginica   36  0 14
```

You can also embed plots, for example:





Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.