# FOURTH STEP

*Miguel Lobo Martin, Borja Ruiz Amantegui, Francesca Sallicati, Antonio Martínez Payá*

*January 28th 2018*

## Contents

# 1 Introduction

The objective of the fourth step is to perform unsupervised classification. For this purpose, we are going to apply three different categories of clustering procedures: Partitional Clustering, Hierarchical Clustering and Model Based Clustering. We are going to check if the variable we want to classify, Position, follows any pattern within the clusters.

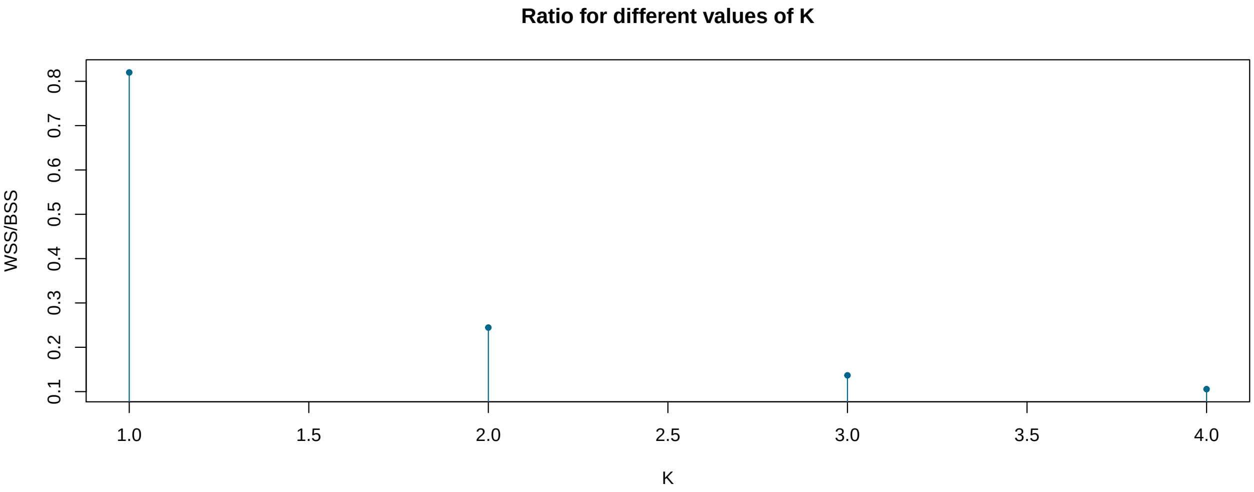
Because of the computational needs of hierarchical Clustering we decided to create a new dataset containing 1000 individuals of each class selected randomly in order to balance the classes. Nevertheless, we used in some methods the complete dataset.

# 2 Partitional Clustering

For these methods the number of clusters $k$ has to be decided *a priori*. What they do is to create k clusters that minimize a certain optimality criterion. Let's study what happens with the different methods.

## 2.1 K-means

K-means algorithm minimizes the ratio between the Within-cluster sums of squares and Between-cluster sums of squares. First of all we checked that this quantity is minimized for k=4.



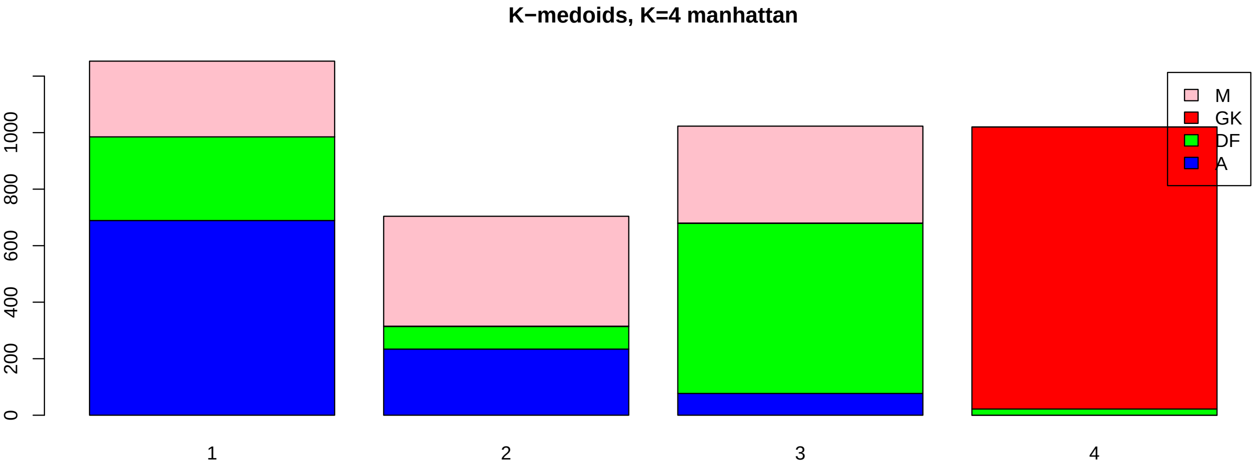
Indeed the optimal value is k=4.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 594 | 46 | 0 | 360 |
| DF | 548 | 59 | 2 | 391 |
| GK | 0 | 16 | 984 | 0 |
| M | 251 | 102 | 0 | 647 |

By applying kmeans with 4 centers the only category which is detected is goalkeeper.

## 2.2 K-medoids

K-medoids is quite similar to k-means but the sample mean vector is replaced with the medoid, which is the element of the cluster whose average distance to all the observations within the cluster is minimal. We tried with k=3 and k=4 and use both the euclidian and manhattan distances.



| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 689 | 234 | 77 | 0 |
| DF | 296 | 80 | 602 | 22 |
| GK | 0 | 1 | 1 | 998 |
| M | 268 | 389 | 343 | 0 |

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 681 | 69 | 250 | 0 |
| DF | 334 | 508 | 145 | 13 |
| GK | 0 | 0 | 28 | 972 |
| M | 295 | 315 | 390 | 0 |

Table 2: Manhattan vs euclidean with K=4

| | 1 | 2 | 3 |
|---|---|---|---|
| A | 740 | 260 | 0 |
| DF | 477 | 498 | 25 |
| GK | 0 | 3 | 997 |
| M | 348 | 652 | 0 |

| | 1 | 2 |
|---|---|---|
| A | 984 | 16 |
| DF | 947 | 53 |
| GK | 1 | 999 |
| M | 1000 | 0 |

Table 3: Manhattan K=3 vs K=2

With k-medoids as well the only almost fully detected category is still goalkeeper; however we can see that the manhattan distance seems to work better than the euclidian distance. Moreover we can notice that the third cluster contains mainly defenders compared to the other categories and clusters, and also the first cluster contains the majority of the attackers.

By reducing the number of clusters the algorithm classifies better goalkeepers, while it seems that the others categories have more similar attributes values which do not allow the method to distinguish between them.

## 2.3 Clara

This method is a generalization of k-medoids which allows to work with large datasets, therefore we decided to try to use our entire unbalanced dataset, obtaining the following clusterization:

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 19 | 70 | 34 | 497 |

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| DF | 66 | 212 | 127 | 1214 |
| GK | 202 | 491 | 263 | 2484 |
| M | 887 | 2194 | 1572 | 7354 |

We can deduce that this is the worst method among partitional clustering algorithms since it does not even detect goalkeepers.

Therefore within partitional clustering the best method is k-means with k=4 and using the manhattan distance.

# 3 Hierarchical Clustering

Hierarchical clustering procedures do not require to select the number of cluster to be made in advance. A distance matrix is first computed with different types of metrics available for the distance calculation, afterwards the tree can be cutted at the level desired.

## 3.1 Agglomerative

Agglomerative algorythims start with one clusters per single observation and continue merging clusters until all observations are clustered in the same group.

The metric we used for our distance matrix was the gower, since it permits categorical values in the dataframe.

Also every variable has been standarized to obtain distances.

Four clustering methods have been implemented:

1. Single linkage method.
2. Complete linkage method.
3. Average linkage method.
4. Ward linkage method.

### 3.1.1 Single Linkage Method

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 999 | 1 | 0 | 0 |
| DF | 999 | 0 | 1 | 0 |
| GK | 999 | 0 | 0 | 1 |
| M | 1000 | 0 | 0 | 0 |

We can observe that this method has created a really big cluster containing most of the observations.

### 3.1.2 Complete Linkage Method

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 474 | 493 | 33 | 0 |
| DF | 580 | 379 | 41 | 0 |
| GK | 16 | 0 | 0 | 984 |
| M | 212 | 704 | 84 | 0 |

Complete linkage method also categorizes well only goalkeepers.

### 3.1.3 Average Linkage Method

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 996 | 4 | 0 | 0 |
| DF | 993 | 7 | 0 | 0 |
| GK | 3 | 0 | 991 | 6 |
| M | 988 | 12 | 0 | 0 |

Though average linkage method is able to categorize goalkeeper, it doesn't improve complete linkage method because groups 2 and 4 are practically empty.

### 3.1.4 Ward Linkage Method

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 437 | 400 | 163 | 0 |
| DF | 394 | 279 | 327 | 0 |
| GK | 1 | 0 | 0 | 999 |
| M | 408 | 186 | 406 | 0 |

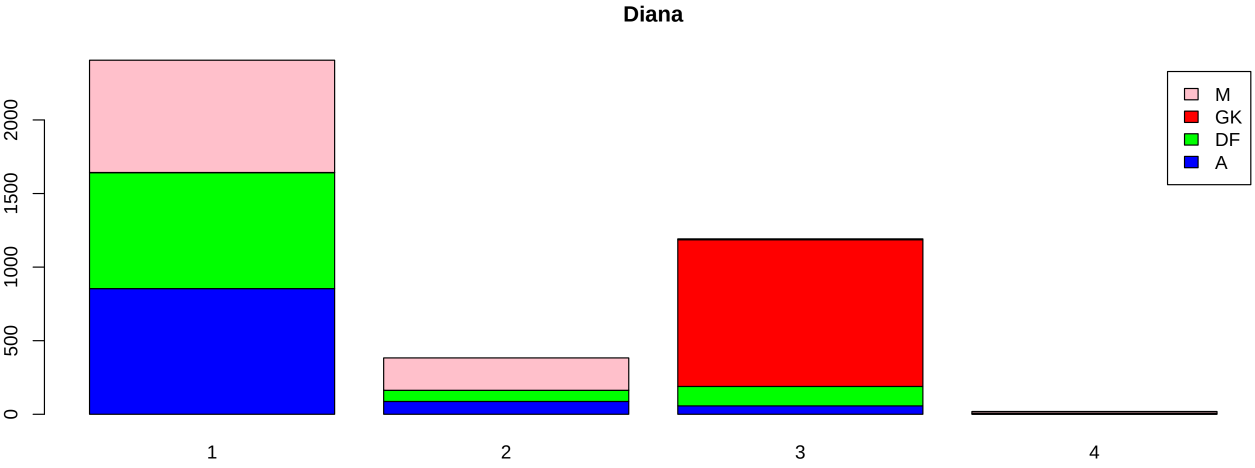
It only categorizes well goalkeepers.

Among all the methods tested the complete and ward linkage methods are the ones that provide better insights.

## 3.2 Divisive

This hierarchical clustering method works in the opposite way of the agglomerative algorithms. This approach starts with a single cluster iteratively splitting them until each observation is a cluster itself. The algorithm we will use is the Diana method, which is the most popular approach for divise hierarchical clustering.

### 3.2.1 Diana



| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 839 | 90 | 69 | 2 |
| DF | 768 | 164 | 63 | 5 |
| GK | 1 | 997 | 1 | 1 |
| M | 805 | 11 | 180 | 4 |

We can see that the Diana algorithm does not correctly classify the instances into the 4 groups that we previously had, but at least is able to classify goalkeepers in group 3.

# 4 Model based clustering

The main assumption behind this method is that observations are generated by different distributions with certain probabilities. Using a Maximun Likelihood Estimation of model parameters as well as the Bayesian Information Criterion the algorithm considered two clusters as the most appropiate number.

The table below shows the results which are unuseful for our classification problem:

| | 1 | 2 |
|---|---|---|
| A | 683 | 317 |
| DF | 633 | 367 |
| GK | 717 | 283 |
| M | 565 | 435 |

# 5 Conclusion

1. Goalkeepers are easily detected by most of the clustering algorithms.
2. Model-based clustering and Clara yield the worst performance among all the used methods.
3. K-medoids with k=4 and manhattan distance provides a useful insight for our classification purpose. Cluster 4 is mainly represented by goalkeepers. In cluster 3 we can see a lot of defenders as well as some midfielders that have similar characteristics according to k-medoids. The same happens in cluster 1 with attackers. Most of the observations in cluster 2 are midfielders though this category is spread out over clusters 1,2 and 3, which makes sense since midfielders can usually rotate freely across the field.