Clustering of forwards based on game performance in the elite European football leagues

Introduction

The area of study that the report will cover is how clustering of football players can be conducted. To narrow down the problem a little bit further only the football players who plays as forwards has been included in the study. Hence, the main aim with the report is to investigate what types of clusters forwards from the top European football leagues can be clustered into.

The top European football leagues are the English Premier League, German Bundesliga, Italian Serie A and Spanish La Liga. In the world of football it is a well-known fact that these leagues are the strongest and it is also stated in a more official way by the UEFA ranking of the European leagues (Uefa, 2016). During the 2015-16 season a total of 178 players from these leagues had playing time as forwards corresponding to at least six full games. Data for 44 different variables (more about the data set in the background chapter) is collected for each player and the amount of variables makes it hard to just by the eye detect and group similar players together.

This difficulty, to find players which are similar, is a constantly ongoing problem for football clubs all over the world. How to replace your star player when he leaves? In the summer of 2014 Luis Suarez joined FC Barcelona from Liverpool leaving the latter club with the hard task of replacing their forward star (Walsh, K. 2014). The British club signed the Italian striker Mario Balotelli to cover up for the loss of Suarez, but he failed miserably and the signing of Balotelli has been heavily criticized (Ran, E. 2016). This summer a similar case might take place as Robert Lewandowski is rumored to be leaving Bayern Munchen for Real Madrid (Aarons, E. 2016). How the German's are going to replace Lewandowski would in that case be one of the hottest topics this summer.

In many cases it is reasonable to think that the clubs want to replace the forward that leaves with a similar forward. To find a forward that is similar to Suarez or Lewandowski is of course always going to be very difficult, but perhaps you at least want to find someone who takes a similar number of shots per game or creates goal opportunities’ for his teammates at a similar rate. Here is where the use of Data Mining techniques, and especially clustering techniques, becomes interesting.

Earlier studies in which football players has been clustered is quite rare. However, the article published on the blog pena.lt/y is one example of this (pena.lt/y, no date). In this article the author clusters players playing on all possible positions by using principle components and the k-means clustering algorithm. Principle components is used for reducing the dimensionality since the amount of variables is high. The k-means algorithm splits the players into five different groups and the given results are not very surprising. Goalkeepers are in one cluster, defenders in another and so on. The article does not examine any further if, for example, the group of midfielders can be divided into any subgroups of midfielders.

To find more examples of similar studies it is necessary to look at sports other than football. In the article *Exploring Game Performance in the National Basketball Association Using Player Tracking Data* (Sampaio et al., 2015) are basketball players clustered to create game performance profiles based on different game roles. The authors used k-means to create the profiles and presented seven different types of game performance profiles which according to the authors agrees well with the existing roles that a basketball player can take.

Background

# Data

The data used for the study is collected by Opta and made available via the website WhoScored.com. In total the dataset consists of 178 players and for each player there are 44 variables. The collected data is for the full season 2015/16 in the English Premier League, German Bundesliga, Italian Serie A and Spanish La Liga. To be included in the dataset a player must have playing time as a forward that corresponds to at least six full games. It is not enough to normally play as a forward for being included. Instead, since the players positions are logged by Opta, only the data for the time during the season the players has played as forwards is of interest. Otherwise it would not be sure that what is evaluated actually is what a player contributes with when used as a forward.

On the website WhoScored.com there are more data from Opta then the 44 variables that I picked. The reason why some variables are rejected is that they were considered to be uninteresting or meaningless for the study. For example measures of defensive actions a forward player very rarely execute, like blocking shots or crosses, does not add any info since they so seldom occur during a game. Other variables, like number of appearances or minutes played, are also discarded since they not contribute with any interesting information about a players actions on the pitch.

Every numeric variable in the dataset that is used in the report is converted to be the average per 90 minutes played instead of the total number or per game. This to see what a player on average contributes with during 90 minutes of football. Therefore, the number of 90’s for a player is used instead of the number of appearances or minutes.

All of the variables in the dataset are count variables which measures how often per 90 minutes a certain action is executed. In short, the variables can be thought to be divided into four different top-levels: Goals, Shots, Key Passes or Others. For instance the top-level Goals has lower levels like number of goals in the penalty area, in open play and so forth. A key pass is defined as the final pass which leads to a shot at goal (WhoScored, no date). The full list of variables is presented below.

|  |  |  |  |
| --- | --- | --- | --- |
| Goals | Shots | Key Passes | Other |
| Six-yard-box | Six-yard-box | Long | Fouled |
| Penalty area | Penalty area | Short | Fouls |
| Out-of-box | Out-of-box | Cross | Dribbles |
| Open play | Open play | Corner | Unsuccessful dribbles |
| Counter | Counter | Through ball | Successful dribbles |
| Set Piece | Set Piece | Free kick | Dispossessed |
| Penalty | Penalty | Other | Aerials total |
| Normal | Footed |  | Aerials won |
| Footed | Headed |  | Aerials lost |
| Headed | Off target |  | Caught offside |
|  | On post |  |  |
|  | On target |  |  |
|  | Blocked |  |  |

# Method

## K-means

The clustering algorithm that will be used in the report to obtain the different groups of forward players is k-means. This is a well-known clustering method and the standard version of the algorithm, see for example Han, J., Kamber, M., & Pei, J. (2011). , will be used. Since the values for the metrics in the dataset have different magnitude a z-score normalization of the variables is conducted before the algorithm is applied on the dataset. The decision to normalize data is based on the algorithms use of the Euclidean distance and the aim to give the variables the same weight.

In the earlier studies where the aim has been to cluster players, in either football or basketball, k-means has been used. For both cases the obtained results indicates that this clustering technique seem to fit this kind of problems quite well. However, for the football case it has only been used for clustering players into general positions like goalkeepers, defenders, midfielders and so on. That the method will work equally well, or that it is the most suitable method, in this case is not self-evident but given the available information in form of earlier research it is concluded to be a good choice.

### The elbow method

A challenge with the k-means algorithm is to set *k*, the number of clusters, to a fitting number. This can be done in a various number of ways, but there is no general methodology. Neither among football people is there any mutual agreement on how many different types of forward that can be defined. Most common is to specify four (Obisesan, A. 2014) or five (football-bible, no date) different types but there are no general agreement on which these types are and it can both be more and fewer groups.

## An alternative approach for solving this problem is the elbow method (Wikipedia, 2016) (sthda, vet ej år). The aim with this method is to examine how the percentage of variance explained or total within cluster sum of squares changes as the number of clusters exceeds. When the increase of percentage of variance explained or decrease of total within cluster sum of squares no longer add so much information the number of *k* is found. This drop in information given by adding another cluster often looks like an elbow when it is visualized and is the reason till why it is called the elbow method.

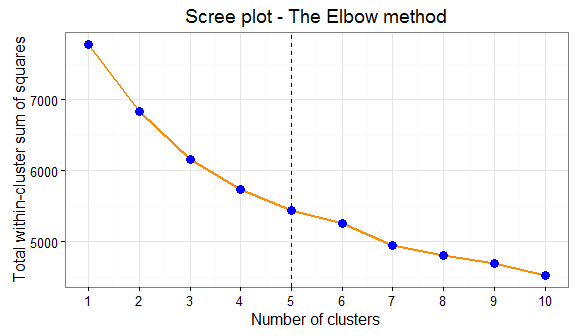
# Software

All the computations and visualizations in the report are done with R. To perform the K-means algorithm is the package *stats* (R Core Team, 2015) used and for visualizations *ggplot2* (Wickham, 2009).

Results

# Number of clusters

The *k* in the k-means algorithm is decided by looking at how the total within cluster sum of squares decreases at the number of clusters increases.



No evident elbow can be seen in the plot but that the improvement lowers as the number of clusters increases is clear. A *k* equal to either four or five is interpreted as being reasonable as the difference between five and six clusters is quite small. The difference between four and five clusters on the other hand is slightly clearer so a *k* is set to 5.

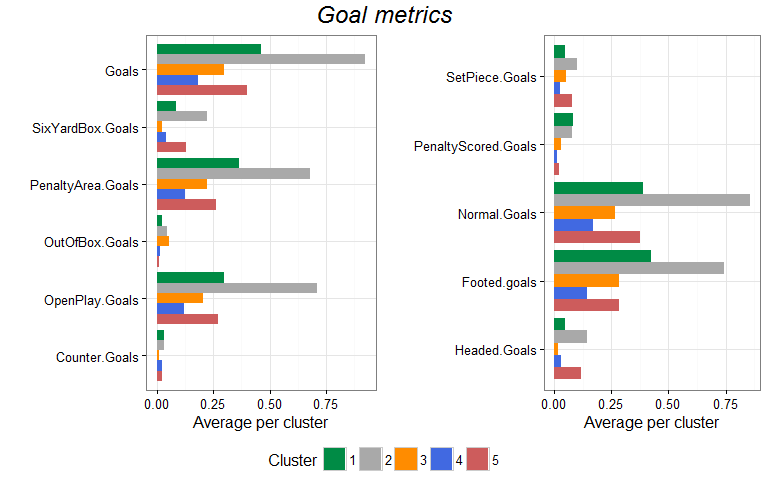
# The clusters

The size of the respective clusters are relatively even except for cluster two which only has nine players.  
## 1 2 3 4 5   
## 49 9 28 51 41

The five different clusters are compared by one category at the time. First, how the groups differs for the goal variables is presented and then they are compared given the variables in the categories shots, key passes and other.

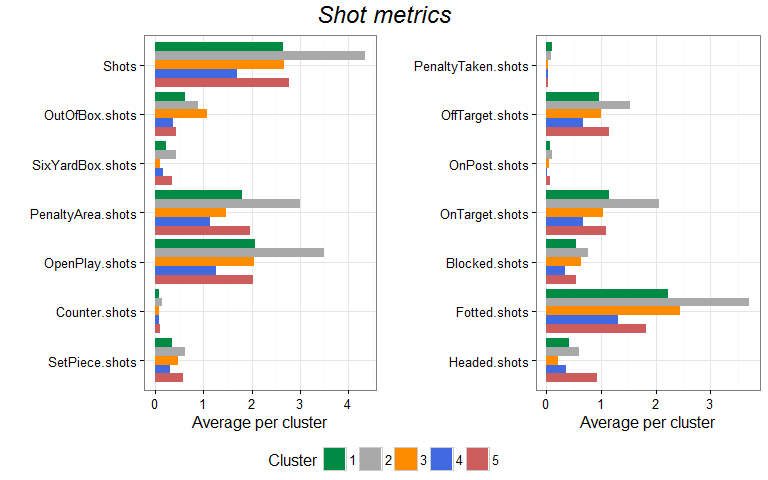
## Goals

The average contribution from the players in the respective clusters is compared with the visualizations below. For each shot metric is the average number computed and represented with a bar in the graph.



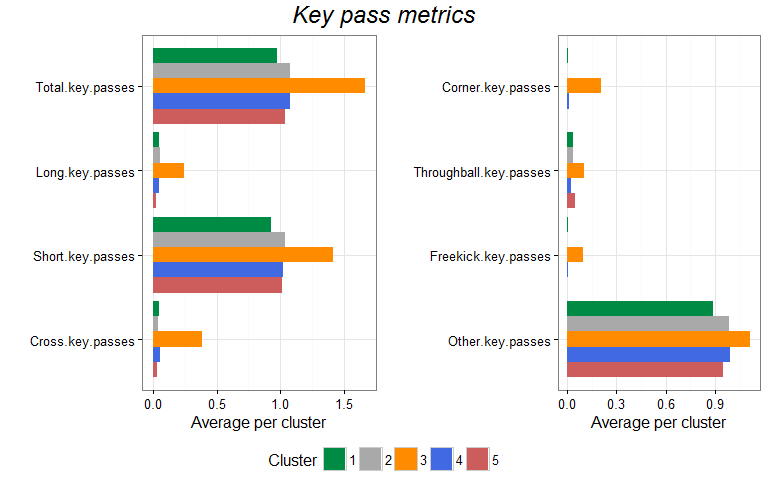
Cluster number 2, the grey one, includes the players that scores the most goals per 90 minutes and the average number is just below one. Cluster 1, the green, comes next and thereafter follows cluster 5, the red, and cluster 3, the orange. The lowest number of goals per 90 minutes has the players in cluster 4, the blue one. For most of the metrics the same order and relative difference between the clusters is observed. However, deviations from that pattern can be seen for a few metrics. Players that are clustered into the orange cluster very seldom scores in the six-yard box. Instead, they score relatively often with shots from outside the box. Another exception can be seen for the players in the red cluster since a notably big part of their goals are headed.

## Shots



The forwards in the grey cluster has the highest amount of shots per 90 minutes played. Thereafter follows again the green, orange and red cluster and it is really close between the averages of these clusters. The blue cluster has the lowest average number of shots. The metrics for which this pattern cannot be seen is Out-of box shots and headed shots. In average does the players in the orange cluster take a lot of shots from outside the box and the players in the red cluster has quite many headed shots per 90 minutes. The general pattern for the green and blue cluster is similar to the pattern for the grey cluster, with the exception that their average number is lower.

## Key passes



The average number of key passes is by margin highest for the orange cluster. Between the other clusters is it very even. None of the averages for the other metrics give any especially notable information apart from that the highest average is found for the orange cluster.

## Other

## 

# Lewandowski´s replacement

Lewandowski is in cluster two, the cluster with only nine players in total. Exclusive group of players who on average scores around one goal per game and takes 4-5 shots per game. A bit more spread out for the key passes. The big common factor is the number of chances these players creates for themselves and how well they converts the chances into goals.

## Players Goals90 Shots90 KeyPasses90  
## 1 Harry Kane 0.7 4.2 1.2  
## 2 Sergio Agüero 0.9 4.5 1.0  
## 3 Daniel Sturridge 0.7 4.2 0.6  
## 4 Robert Lewandowski 0.8 4.8 0.7  
## 5 Pierre-Emerick Aubameyang 0.9 4.2 1.0  
## 6 Gonzalo Higuaín 1.1 5.5 1.5  
## 7 Luis Suárez 1.1 3.9 1.6  
## 8 Karim Benzema 1.1 4.4 1.9  
## 9 Imanol Agirretxe 1.0 3.4 0.2

Discussion

References

Aarons, E. The Guardian (2015-05-27), <https://www.theguardian.com/football/2016/may/27/bayern-munich-robert-lewandowski-real-madrid-talks> , Accessed 31/5-16

Football-bible (No date), <http://www.football-bible.com/soccer-info/soccer-positions-explained.html>, Accessed7/6-16

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Obisesan, A. Businessdayonline (2014-12-14), <http://businessdayonline.com/2014/12/five-dynamics-of-modern-soccer-strikers/>, Accessed 7/6-16

Pena.lt/y (No date), <http://pena.lt/y/2014/02/10/comparing-players-using-cluster-analysis/> , Accessed 20/5-16

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ran, E. Sportskeeda (2016-01-18), <http://www.sportskeeda.com/football/brendan-rodgers-explains-liverpool-failed-signing-mario-balotelli-instead-alexis-sanchez>, Accessed 26/5-16

Sampaio, J., McGarry, T., Calleja-González, J., Sáiz, S. J., i del Alcázar, X. S., & Balciunas, M. (2015). Exploring game performance in the national basketball association using player tracking data. *PloS one*, *10*(7), e0132894.

Sthda (No date), <http://www.sthda.com/english/wiki/determining-the-optimal-number-of-clusters-3-must-known-methods-unsupervised-machine-learning>, Accessed 7/6-16

UEFA (2016-06-01), <http://www.uefa.com/memberassociations/uefarankings/country/> , Accessed 20/5-16

Walsh, K. Liverpool Echo (2014-07-11),<http://www.liverpoolecho.co.uk/sport/football/football-news/luis-suarez-barcelona-how-liverpool-7383131> , Accessed 26/5-16

WhoScored (No date), <https://www.whoscored.com/Glossary> , Accessed 24/5-16

Wikipedia (2016-05-29), <https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set>, Accessed 7/6-16