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 [1]. 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 41 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 [2]. 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 [3]. This summer a similar case is on our hands as Zlatan Ibrahimović will leave PSG at the end of the season. How the Frenchman’s are going to replace Ibrahimović will be one of the hottest topics this summer [4].

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 Ibrahimović 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 [5]. 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*” by Sampaio et al. are basketball players clustered to create game performance profiles based on different game roles [6]. 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.

<http://www.uefa.com/memberassociations/uefarankings/country/> [1], 20/5

<http://www.liverpoolecho.co.uk/sport/football/football-news/luis-suarez-barcelona-how-liverpool-7383131> [2], 20/5

<http://www.sportskeeda.com/football/brendan-rodgers-explains-liverpool-failed-signing-mario-balotelli-instead-alexis-sanchez> [3], 20/5

<http://www.espnfc.com/club/paris-saint-germain/160/blog/post/2871975/zlatan-ibrahimovic-replacements-ronaldo-neymar-hazard-griezmann-sanchez> [4], 20/5

<http://pena.lt/y/2014/02/10/comparing-players-using-cluster-analysis/> [5], 20/5

[6], “*Exploring Game Performance in the National Basketball Association Using Player Tracking Data*” by Sampaio et al.

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 [7]. 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 |  |  |

Example of player, Roberto Firmino:

<https://www.whoscored.com/Glossary> [7], 24/5

# Method

## K-means

Z-score normalization of data because of the different magnitudes for data. So that it will fit the computations of the Euclidean distance

### Elbow method

## (Principal components)

# Software

All the computations and visualizations in the report are done with R. To perform the K-means algorithm is the package *Rweka* used and for visualizations *ggplot2*.