Game Performance - Forwards - Elite European Football leagues

Gustav Sternelöv

28 april 2016

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 181 players from these leagues had playing time as forwards corresponding to at least six full games. Data for 47 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 or 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 k-means. 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.

Introduction

Among all the football leagues in the world are the top european football leagues the most competitive. To these exclusive group of leagues belongs the top four leagues on the UEFA ranking [*Source*], the english Premier League, german Bundesliga, italian Serie A and spanish La Liga. The most expensive and wanted players in all fotball teams, but especially in the top teams, are the forwards. Football is all about winning matches and scoring goals, and the main role of a forward is to score goals. It has therefore always been a bit of a mystery around the forwards, why are just them so good at scoring goals?

Compared to other sports is not Football the most complex sports in terms of rules and the basics of the sport is easy to understand. However, regarding the amount of possible situations and events that can occur during a game is football a sport with high complexity. A player can take many different roles (central defender, winger, forward etc) and each role can be executed in several different ways depending on formation, tactics and so forth. For example the forward role is widely thought to be executed very different depending on both the tactics/formation used by the team, but also depending on the player. A tall player is often thougt to be best used as a target man and a small and fast player more lika a runner and goalscorer.

Traditionally is the main statistic used for assessing a forwards performance's the number of goals he scores. During recent years has the role of the forward changed a bit and its not longer that obvious what the main indicator should be for judging the performance of the forwards. Perhaps goals still is the most important contribution for a forward to the team, but as the new forward types as the *false nine* has been introduced the defintion of what a striker is definitely has become broader.

The aim with this paper is to examine if game statistics can be used for labeling the strikers in the european top leagues. If different types of strikers can be found or if it is difference between good and bad strikers that will form the different clusters.

Of interest is also the assess what that is the profile of each cluster. How can a forward in each cluster be defined, regardless if the clustering groups players after how good they are or after which type of forward they are.

The objective with the paper then is to, using clustering methods, investigate how many different types of forwards that can be found. Apart from goals and number of shots is also variables regarding key passes, dribbles, headers, lost balls and many more included. This because what is stated above, the role of the forward in the modern football is much more than just scoring goals. For different types of forwards it might be that case that completely different contributions to the team is important.

[**Sources!**]

Background

Former study

Basketball players article

Uses actions per game for clustering basketball players into different groups of player performance.
Finds seven different profiles and presents the mean value and standard deviation for each profile and variable.
The all-star players in kind of the same clusters. All of them in three of the seven clusters

Statsbomb, the per 90 minutes metric

That study is the inspiration to the aim of this study, if football players can be categorized into different groups. More narrowly, how and if forwards can be grouped into different types of forwards.
Will the clustering find different types of forwards or will it cluster the forwards in good or bad after how good they are.
Do have a lot of variables, which of them will prove to be important?

Types of strikers

Data and variables

In this study is game statistics uesd for 47 different variables. No positional tracking data or such. The metrics per 90 is used because it is more interesting to cluster the forwards after what they contribute with on average to the team during 90 minutes rather than the total statistics for the last season, not taking into respect the number of minutes played.

Excludes the number of 90s, is not a characteristic that says anything relevant about the players profile.

Four different types of statistics, you could say. Three big categories which are shots, goals and key passes, and a fourth containing other statistics(aerials, dribbles, fouls, offsieds, lost balls)

Methodolgy

Dataset

Algorithms

K-means, number of strikers decided after what that can be reasonable and something more? Is the goal to minimize the sum of squares within the clusters? Should also test a number of different seeds?

Hierarchical clustering
Motivate the choice of hierarchical clustering and the choice of link. Test a few different links perhaps, in that case hae to motivate the choice. Where to cut the tree and how to choose link?

Results

Discussion