Universidad de Navarra

EA FIFA PLAYER POSITION IDENTIFICATION USING ANALYTICAL CLUSTERING

Villanueva Jove, Javier E.

Machine Learning

Introduction

For the purpose of this project we will be analyzing a data set containing detailed statics of the world's best football players. Using this information, we have implemented an algorithm capable of identifying players based on their attributes and playing styles. This way we are able to sort through the entire player database and identify player positions in the most efficient way possible.

The Data Set

The data set used contained the information of the best 3,500 football players on the world, from each player we have 3 qualitative characteristics and 38 quantitative characteristics that describe all the important skills and attributes a football player should posses. (The numerical values of the variables are ranked from 1 to 99)

The variables are:

Ranking	Name	Age	Nationality	Overall
Club	Acceleration	Aggression	Agility	Balance
Ball control	Composure	Crossing	Curve	Dribbling
Finishing	Free kick accuracy	Diving	Handling	Kicking
Goal positioning	Reflexes	Heading accuracy	Interceptions	Jumping
Long passing	Long shots	Marking	Penalties	Positioning
Reactions	Short passing	Shot power	Sliding tackle	Sprint speed
Stamina	Standing tackle	Strength	Vision	Volleys

Objective

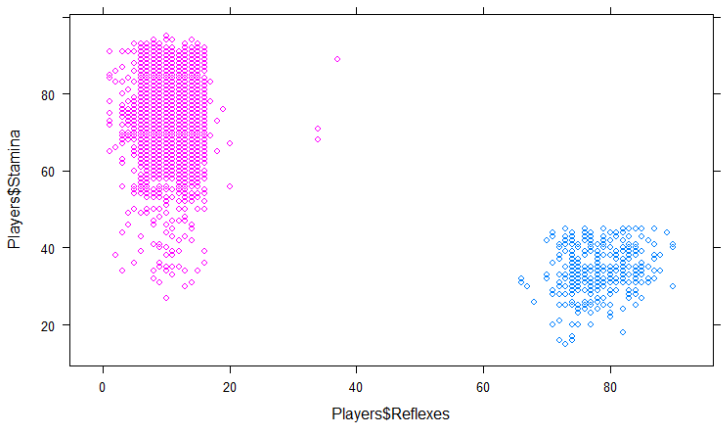
Our purpose is to identify players that could potentially be a better fit, and therefore increase their performance, in a different position that the one they are currently in. For example, it could be the case that certain players that play in an attacking position have all the attributes necessary to be a world-class goalkeeper but instead they are playing in a position that is not their natural one based on their skillset. Such cases are not that far from reality.

Take for example Gianluigi Buffon, considered to be amongst the best goalkeepers of all time. Buffon started his career playing as a striker for a third division Italian team, it wasn't only until the injuries of the first and second choice goalkeepers in his team that he got the chance to play in that position. Cases like this happen more often than not and through our player identification model we are able to properly recognize players natural position to what could prove to change a game changing decision.

Analysis

After examining and analyzing the data set, we found that the values of certain skills were very similar among the same positions but varied considerably between players and goalkeepers, for example:

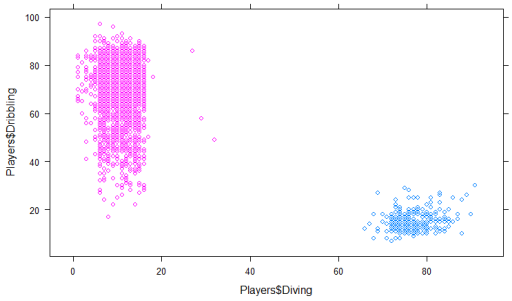
```
> xyplot(Players$Stamina ~ Players$Reflexes, group=Position, data=Players))
```

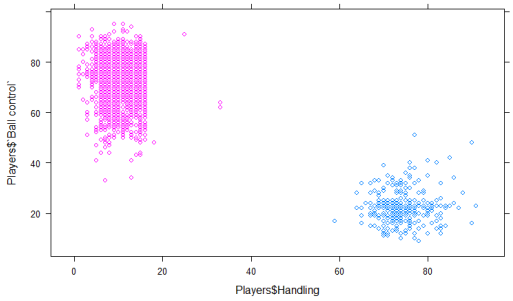


As we can see, comparing attributes like reflexes and stamina revealed that most field players (pink) possessed high values of stamina but low values of reflexes, as opposed to goalkeepers (blue) that showed values completely opposite.

These results appeared to become a pattern among several other skills such as dribbling, diving, ball control and handling:

```
> xyplot(Players$Dribbling ~ Players$Diving, group=Position, data=Players)
> xyplot(Players$`Ball control` ~ Players$Handling, group=Position, data=Players)
```





Analysis

To fulfill our objective, we proceeded with the first step which consisted in creating a copy of our data set and eliminating all variables with non numerical values and any other irrelevant to our study:

```
> DF <- Players
> DF$Ranking = NULL
> DF$Name = NULL
> DF$Age = NULL
> DF$Nationality = NULL
> DF$Overall = NULL
> DF$Potential = NULL
> DF$Club = NULL
```

Next, we apply the ***kmeans*** algorithm to divide our new data set into 2 clusters using all the skill variables as its division criteria:

```
> Results <- kmeans(DF[,1:34], 2, nstart=20)
```

And create a table that compares and displays the cluster results with the positions of each of the 3500 players in the data set:

```
> table(Players$Position, Results$cluster)
```

	1	2
GK	344	0
PLAYER	0	3155

Based on the results we arrived when running our experiment, we have found that amongst the top 3,500 players in the world there are no players who could perform better in a position that differs from one the one they are currently in, there are no field players that could perform better as goalkeepers and vice versa.

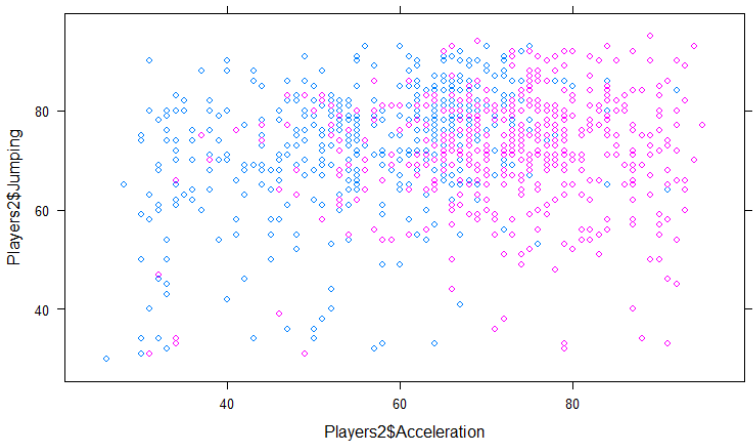
Analysis

Because of these results we have decided to run the experiment again with the difference that this time we will be comparing two positions that were previously inside the “Player” category used before.

This time we will be comparing two types of field players:

- Central Back (CB)
- Striker (ST)

We will run our algorithm under the same conditions as our previous experiment in order to see if we get different results for these particular positions. We are certain that because of the similarity in attributes of these two particular positions we will be able to find players that could perform better in a category different to the one they are currently in. As we can see below by comparing the variables jumping and acceleration, this two positions indeed posses more similar values:



First, we imported a new data set with the same content as the one previously used, but having only the information of the players who’s current positions are central back and striker, and as done before, eliminated the irrelevant variables:

```
> DF2 <- Players2
> DF2$Ranking = NULL
> DF2$Name = NULL
> DF2$Age = NULL
> DF2$Nationality = NULL
> DF2$Overall = NULL
> DF2$Potential = NULL
> DF2$Club = NULL
```

Analysis

Secondly, we proceed to use the ***kmeans*** algorithm again and create the table to display the results:

```
> Results2 <- kmeans(DF2[,1:34], 2, nstart=20)
> table(Players2$Position, Results2$cluster)
```

	1	2
CB	557	1
ST	1	509

After running our new experiment, we have found that there are effectively two players that would perform better in the other category. We have a Central Back (CB) who has the necessary attributes to perform as a Striker and vice versa.

Now, we finish our experiment by inserting the cluster results into the data set to reveal the player's identities:

```
Players2$cluster = Results2$cluster
```

```
which(Players2$Position=="CB" & Players2$cluster =="2")
[1] 410
print(x=Players2$Name[410])
[1] "A. Ekdal"
print(x=Players2$Nationality[410])
[1]  "Sweden"
print(x=Players2$Club[410])
[1]  "Hamburger SV"
which(Players2$Position=="ST" & Players2$cluster =="1")
[1]  584
print(x=Players2$Name[584])
[1] "G. Shibasaki"
print(x=Players2$Nationality[584])
[1] "Japan"
print(x=Players2$Club[584])
[1]  "Getafe CF"
```

Conclusion

In conclusion, we have successfully implemented an algorithm into our dataset capable of identifying players based on their attributes and grouping them into what could be considered as a “better fit” for the players skills. This Player Identification protocol could prove useful for real life teams to assess the viability of changing player positions and signing players for a position different to the one they currently play on, while at the same time spending less money and time on scouting programs that rely on conventional wisdom (subject to biass) rather than data and statistics.