Predictive Analysis: Player Transfers in Soccer

Aditya Bhardwaj
abhardw2@ncsu.edu

Varun Jayathirtha
vjayath@ncsu.edu

Satvik Shetty
smshetty@ncsu.edu

Abstract

Player transfers in Soccer is an important aspect to a team's success. In this project, we analyze a European Soccer Database with important information spanning across the top 11 leagues. After sampling and handling missing data, we use t-SNE for dimensionality reduction which is then fed to K-Means algorithm to classify the players into 4 distinct clusters. After calculating the weak link in a team, we would then use KNN on the overall ratings as a naive (baseline) algorithm to predict players at a similar level to that of the team. A more refined version involves analyzing each player skill relevant to his own field position and comparing that with the skills needed by the team to find a more suitable candidate for replacement.

1 Introduction

Soccer is a highly popular sport across the globe with full-fledged leagues and increased competition with every passing year. The teams in these league are not only professional sport teams, but also a business venture for the owners who expect the best possible return(performance) on investment. To remain competitive, the team dynamics have to be continually improved upon. One of the most common ways to do this is through player transfers.

Player Transfers in Soccer is a highly complex task of identifying the weak areas in a team (defense, attack, midfield) that need to be strengthened, then identifying specific player(s) to be replaced (who are not up to the team standards) and finally finding a suitable replacement who can prove to be an improvement - all of this while considering a player who would fit the budget and are within the reach of the team in question.

While most Data Analysis techniques applied in the field of Soccer tend to be around predicting match winners and league winners (used by the betting industry), this particular project concentrates on building a model that suggests appropriate player transfers (intended to be used by the team management) that would help improve the team. This project uses a comprehensive soccer database which documents information across the top 11 European soccer leagues from 2008-2016. The data from all those years is used for classification of players. But, for analyzing the weakness of the teams, we use 2013-15 seasons' data so as to have a more pertinent information about players and the teams in order to predict the player transfers for the 2015-16 season. The data from this season can be used to verify/validate our prediction and hence analyze the effectiveness of our model.

1.1 Data set

1. Country

Rows = 11, Columns = 2

2. League: Contains the league name and country in which it exists.

Rows = 11, Columns = 3

3. Match_info: Contains match statistics, starting co-ordinates of all 11 players in each team.

Rows = 24158, Columns = 85

4. Player: Contains name and birthday information.

Rows = 11060, Columns = 4

5. PlayerTeam: Season-wise mapping of each player to a team.

Rows = 13686, Columns = 5

6. Player_Attributes: Season-wise statistics of player skills.

Rows = 183978, Columns = 41

7. Player_xy: Most frequent starting position (XY co-ordinates) for each player.

Rows = 11060, Columns = 3

2 Methods

2.1 Data Cleaning

The original data set contained possibly every aspect of the matches that were recorded. This means a lot of information within that was not relevant to our purpose. For example, we removed the 'height' and 'weight' fields from the 'player' table, because that is not necessary to achieve to our goal.

Another aspect of data cleaning involved removal of those tuples from the 'match' table which had null entries for any of the player id and playing positions. Below presented is the data set which is obtained after the data cleaning process.

2.2 Sampling

As mentioned previously in the introduction, we use the complete data set to classify players into different field positions. But we sample data only from 2013-16 and map the players to their respective teams for these three seasons. From this we find weaknesses in teams in 2013-15, in order to predict player transfers for the 2015-16 season. This sampling is made to keep have the latest team details and existing weaknesses in the teams.

2.3 Classification

First, we identify the field positions (defense, attack, midfield, goal keeper) of each player in a team. This is being approached in two ways:

* Based on field positions: Applying common football knowledge, we map playing coordinates to different positions. We then build a decision tree, based on this knowledge.

* Based on player attributes(skills): The data is reduced using t-SNE dimensionality reduction technique into 2 dimensions.

We then use K-Means clustering algorithm with K = 4 (one centroid for each field position) to classify the players into their fielding positions. We applied it on players with overall skill >= 80 (fewer skilled players), >= 70 and >= 50(more players). As the number of data points increases the polarity of the player skills reduces, therefore leading to more misclassification. To avoid this, we plan to provide the starting centroid locations for the K-means algorithm rather than letting the algorithm choose random centroids. This may help form better clusters and minimize the number of misclassification.

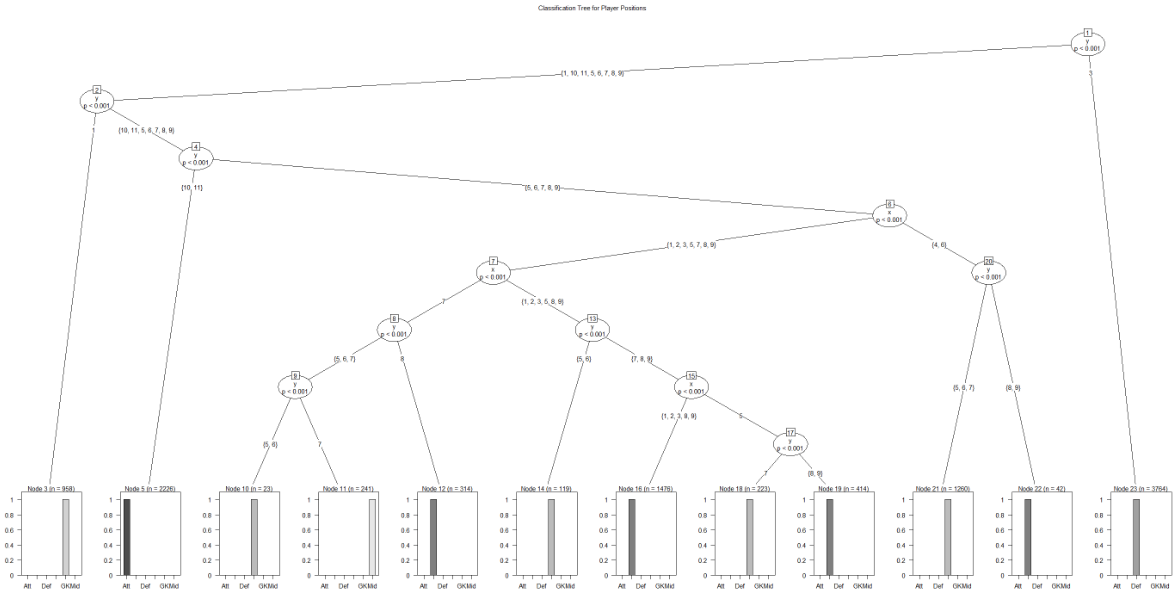

Figure 1 Decision Tree to classify Players based on coordinates

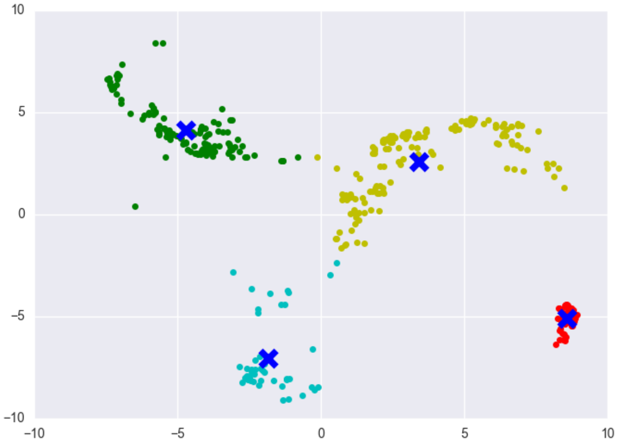

Figure 2 K-Means Clustering when Overall Skill >= 80

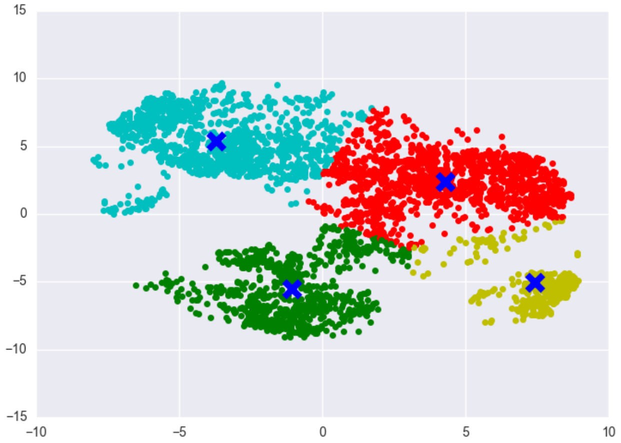

Figure 3 K-Means Clustering when Overall Skill >= 70

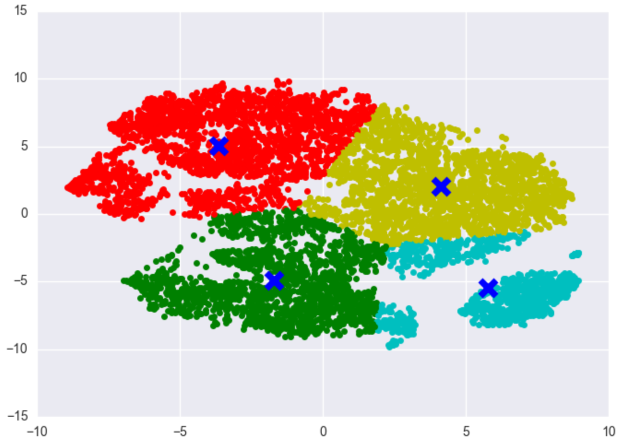

Figure 4 K-Means Clustering when Overall Skill >= 50

3 Identifying Teams' Weakness

In order to identify weakness of a team, each player attribute(skill) is categorized as important to one or more field positions. Then to estimate the overall team skill for each of the player attributes, all the players of that field position(s), are considered. For example, to find the average 'finishing' skill of the team we consider all the players in the team who are classified as 'attacker'. This is because 'finishing' is an important trait mostly for an 'attacker'.

Once this is complete, we group the skills into the field positions and compare them. This gives us an idea about the weak area of the team.

4 Identifying a weak player (work in progress)

Within the weak area of a team, one can look for the player who is having the least overall skill. Although younger players (age < 21) are ignored for this, because they are expected to improve with time, but may not have the best attribute values as of now.

5 Suggesting a replacement (Work in progress)

A suggested player who improves the weak area to bring it up to the levels of its stronger areas, while not going beyond its reach of buying skilled players. To achieve this, we first plot 4 separate graphs (one for each field position). These graphs have player attributes as the axes and players as data points with each player being plotted only on the graph relevant to him, i.e. attackers are plotted on the graph meant for them.

Now, to replace a player from one of the weak areas, say 'attack' we plot the team's average skill in the 'attack' graph. Then apply KNN to get the players who match the overall skill level of the team, while being good the field that needs improvement. Here the best value of 'k' needs to be figured out by trial and error method, to get a good set of suggestions. Other alternative approaches also need to be applied.

References

[1] Linde, J.B. & Løkketangen, M. (2014) Predicting Outcomes of Association Football Matches Based on Individual Players' Performance, NTNU – Trondheim.

[2] Dataset: https://www.kaggle.com/hugomathien/soccer

[3] Kaggle Project for plotting KMeans clustering the Players on basis of their stats: https://www.kaggle.com/ericcouto/d/hugomathien/soccer/exploring-player-stats/notebook

[4] Kaggle Project for assigning players to team positions: https://www.kaggle.com/forums/f/1357/european-soccer-database/t/24503/assign-players-to-team-and-position