Integration of PCA and K-Means to Cluster Soccer Players into Similar Groups

Michael Akinola, *makinola@my.apsu.edu*

Abstract

This project uses European Soccer Database which has more than 25,000 matches and more than 10,000 players for European professional soccer seasons from 2008 to 2016. The exploratory data analysis includes some steps for exploring and cleaning our dataset, some steps for feature engineering using Pearson's correlation coefficient with key attributes and domain knowledge, and some steps for grouping similar clusters using unsupervised machine learning algorithm, K-Means aided by Principal Components Analysis

Keywords - Machine learning, Soccer, Clustering, Principal Components Analysis, K-Means, Pearson's correlation coefficient

1. INTRODUCTION

Data has now become an integral of modern sport administration and management. Sports analytics - data analytics as applied to sports, is the processes that identify and acquire the knowledge and insight about potential players' performances based on the use of a variety of data sources such as game data and individual player performance data. Sports analytics is utilized in various ways like predicting the outcome of a game, predicting the performances of teams or individual players, building new strategies for upcoming competitions, deciding the price of a player if a club was to loan/sell/buy him/her.

1.1. Brief history of soccer

From the Americas to Europe, Asia to Africa, soccer is a sport, enjoyed and played by people from various countries all over the world. According to the world governing body Fédération Internationale de Football Association (FIFA), soccer is one of the world's oldest sports. In the game, players could use any of their body parts save from the hands to get the ball in motion and eventually send the ball through an opening into the net to get a goal scored. The game is equipped with players of various styles and specialties and has evolved over the years with guiding rules being adjusted from year to year to improve transparency and fair play. Soccer matches may be played on natural or artificial surfaces, according to the rules of the competition. The color of artificial surfaces must be green. The field of play must be rectangular and marked with lines.

1.2. Motivation

There are a huge number of factors that define every soccer match, making the sport both exciting and unpredictable. In their quest to soccer glories, soccer coaches have had to change the positions of play of the players available to them for various reasons of which the need of the team and efficient delivery on the field of play are key factors. Another reason could be that a close fit for the role if the role players are not available mostly due to injuries sustained. We have seen coaches both from national sides and club sides change the position of players from defensive midfield to the heart of defense, attacking midfield to the heart of attack, back wings to forward wings. These roles transformations have become a norm in the game. It is important to point out that this is never done without underlying statistics. I seek to deploy machine learning approach to unravel these prevalent actions taken by soccer coaches to maximize the potential of players in the quest to victories and glories.

2. RELATED WORK/APPLICATION

Some football clubs in Europe like S.L. Benfica of Portugal, Arsenal of England and Ajax Amsterdam of Netherlands make as much money from carefully nurturing, training, and selling players as actually playing football. They maximize the knowledge of expert data scientists and the learned experience of the trainer. The data scientists use advanced

technology, machine learning, and predictive analytics, on data collected. The result of these give coaches access to information such as how fast players are running, the distance they are covering in a game and various accuracies in real-time and after a match. Moving forward on this, and by studying patterns of play and player movements, coaches can reconfigure play strategy to make use of each player's strengths and offset their weaknesses to improve overall team performance. Over time, coaches can study the impact of and use data-driven decisions and strategies on overall player and team performance by analyzing the change in player data.

In sports betting, prediction models are based on detailed data and indicators such as player performance, player location stats, expected goals, expected assists, sequence and possession and defensive coverage, contribute to the game's prediction process. With the expansion of detailed data, sports betting operators focus a significant part of their investments in machine learning methods that have shown promising results in prediction.

3. DATASET AND FEATURES

3.1. Data

The data used in this project came from an open dataset from the popular site, Kaggle. The dataset has more than 25,000 matches and more than 10,000 players for European professional soccer seasons from 2008 to 2016. The dataset contains other details like match result details and countries. Only the players attribute data was utilized in this project.

3.2. Exploratory data analysis

This step involves a statistical approach and domain knowledge to know the attributes that most contribute to a player's performance rating. The ratings are low, medium or high attacking work rate and defensive work rate in order to get distinct clusters of players in the Goalkeeping, Defense, Midfield and Forward positions.

3.2.1. Cleaning of data

The data was in .sqlite format and I was able to interact with it using Python. There were 183,978 records and 42 attributes in the dataset. First, I checked for and removed duplicates in the data and this reduced the data to about 10,410 records. After this stage, there was no null value in the data. I proceeded to carry out correlation analysis using Pearson's correlation coefficient which gave an insight into the statistical relationship between the overall rating of the players and 22 other attributes. Figure 1 below shows the values of the correlation coefficients of overall rating of players with other features.

Table 1: Pearson's correlation coefficient of selected attributes with overall rating of players.

Attribute	Correlation coefficient
overall_rating	1.000000
potential	0.817766
crossing	0.292637
finishing	0.266259
heading_accuracy	0.236227
short_passing	0.411985
volleys	0.303953
dribbling	0.295225
curve	0.325220
long_passing	0.391255
ball_control	0.379493
acceleration	0.175934
sprint_speed	0.189308
reactions	0.790788
shot_power	0.348248
long_shots	0.325896
aggression	0.267939
interceptions	0.201302
positioning	0.284508
vision	0.410448
marking	0.119554
standing_tackle	0.149165
sliding_tackle	0.131721

3.2.2. Feature selection

The result of the correlation analysis proved to be good but not perfect. I further applied my domain

knowledge to drop some features. The initial features were 42 out of which 3 are for indexing, 3 are non-numerical features and 27 others were dropped because they are either not distinguishing features or not needed for further analysis. The features that were dropped are; 'id', 'player_fifa_api_id', 'player_api_id', 'date', 'overall_rating', 'potential', 'preferred_foot', 'attacking_work_rate', 'defensive_work_rate', 'crossing', 'heading_accuracy', 'dribbling', 'curve', 'free_kick_accuracy', 'ball_control', 'acceleration', 'sprint_speed', 'agility', 'reactions', 'balance', 'jumping', 'stamina', 'strength', 'long_shots', 'aggression', 'interceptions', 'positioning', 'penalties', 'gk_diving', 'gk_handling', 'gk_kicking', 'gk_positioning', 'gk_reflexes'. Nine attributes; 'finishing', 'volleys', 'shot_power', 'short_passing', 'vision', 'long_passing', 'marking', 'standing_tackle' and 'sliding_tackle', were used for further analysis.
A sample of 974 players with overall rating and potential value of more than 77 and 75 respectively was selected for further analysis. This is because players with this rating can be said to be competent to a large extent in their position of play.

3.2.3. Components Analysis

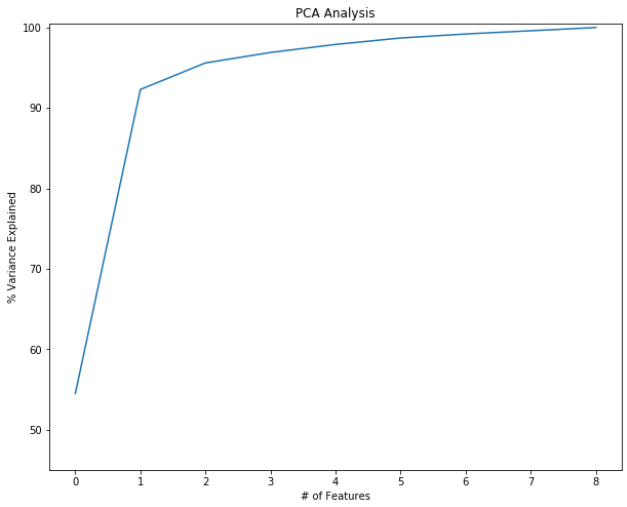

Figure 2: Visualization of the result of Principal Components Analysis

Principal components analysis (PCA) was used for the identification of a smaller number of uncorrelated variables known as principal components from the dataset in order to capture the features having lesser variances and strong patterns in the data set. The result showed that six of the features contributes 98.7% to the patterns formed by the data. This was done to reduce the stress of having to iterate through the nine features in pairs to discover patterns during clustering. Figure 2 shows a visual version of the result of the PCA. This result was improved using the Seaborn library to view scatter charts of the nine features at once. From that, I was able to figure out the feature pairs that had distinguishable clusters.

4. METHOD

The unsupervised machine learning algorithm, K-Means was implemented in this work using the Scikit learn library. K-means is a form of unsupervised learning that looks to find naturally occurring clusters in the data by minimizing the distance between some cluster representative and members of that cluster for each cluster. Most unsupervised learning techniques are a form of cluster analysis. The algorithm iterates between selecting cluster representatives from the clustered points and assigning data points to the nearest cluster, looping until the cluster assignments no longer change.
For this project, I used four as the number of clusters (derived from the plot shown in figure 3) and the 'elkan' variation. The 'elkan' variation is more efficient and uses the triangle inequality. It supports dense data and is a good fit for my type of data.

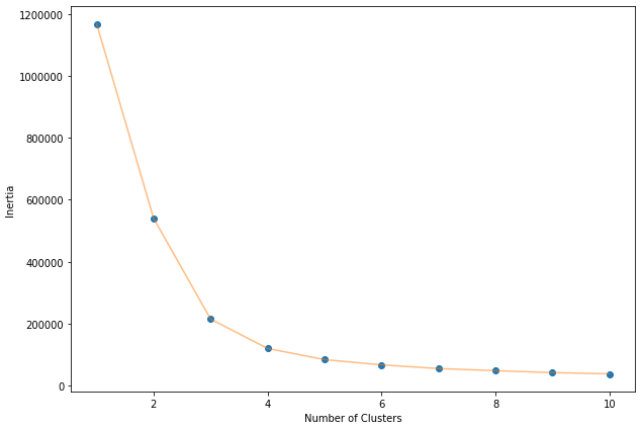

Figure 3: Plot showing possible choice of number of clusters (4 being the best)

5. RESULTS AND DISCUSSION

I plotted six of the features whose pair plots showed the most distinguishable clusters from the seaborn's pair plots. I settled for five pair plots of clusters. The plotted pairs are; finishing versus sliding tackle, volleys versus marking, shot power versus standing tackle, finishing versus marking and finishing versus standing tackle.

5.1. Clustering results

Figures 4 – 8 show the clusters that were obtained from the plots of the K-Means model.

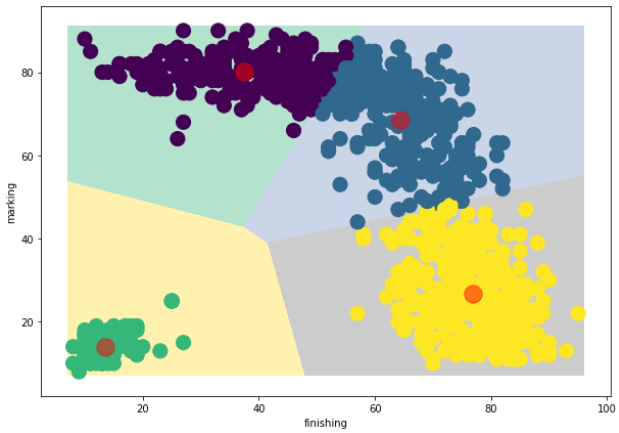

Figure 4: Clustering result of the plot of marking versus finishing

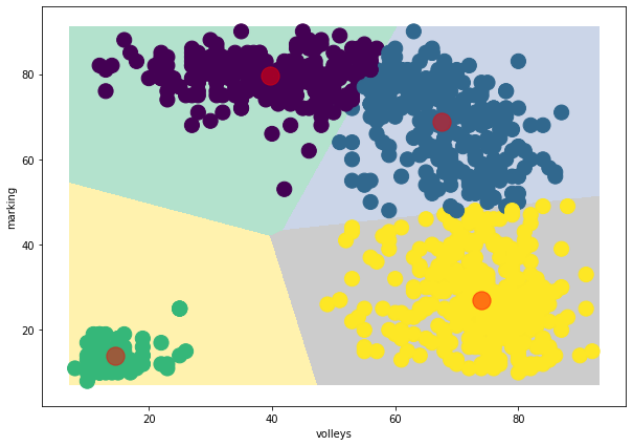

Figure 5: Clustering result of the plot of marking versus volleys

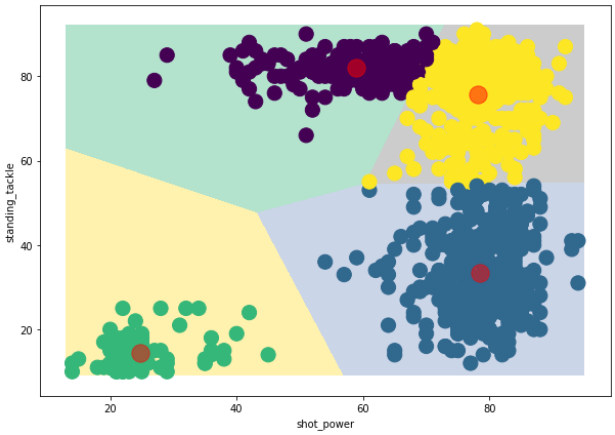

Figure 6: Clustering result of the plot of standing tackle versus shot power

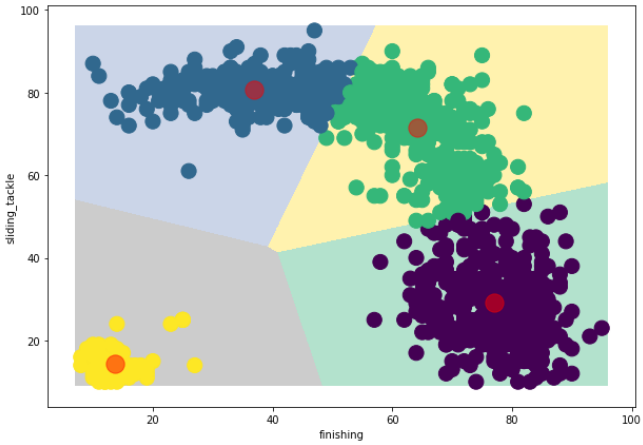

Figure 7: Clustering result of the plot of sliding tackle versus finishing

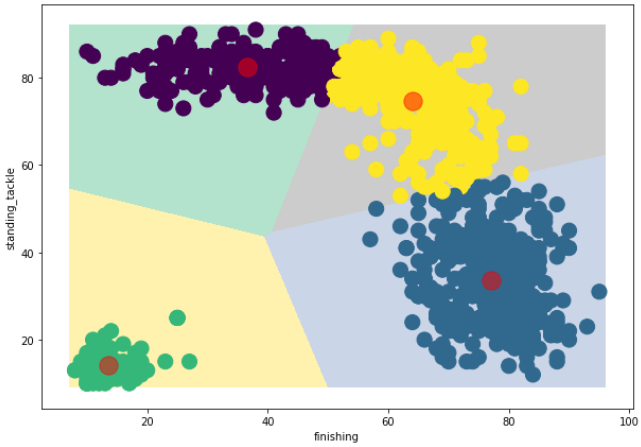

Figure 8: Clustering result of the plot of standing tackle versus finishing

5.2. Discussion

From the plots above, the smallest cluster has a value of 95 counts all through the five pair plots. This cluster represents players in the goalkeeping position. They do not possess high rating values for the attributes I plotted. The other three clusters have counts in the range 174 and 372. This is accounted for by the exceptional abilities of some players in the defense, midfield or forward position having unusual rating values for some attributes that are not peculiar to their position of play. This is a good result because the coach can maximize these exceptional abilities when there is an emergency, for instance, injuries to a player in the position.

ACKNOWLEDGMENTS

Thank you to Dr Daniel Mayo for teaching the methods used in this project and the guidance provided throughout this project.

REFERENCES

https://www.kaggle.com/hugomathien/soccer

https://www.techopedia.com/definition/32509/principal-component-analysis-pca

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

https://www.wagerbop.com/machine-learning-in-sports-betting/

https://medium.com/@nabil_lathif/the-number-games-how-machine-learning-is-changing-sports-4f4673792c8e