Soccer Match Prediction with Data Mining

Jaekyum Kim

April 27 2020

1 Collaboration Statement

THIS CODE IS MY OWN WORK, IT WAS WRITTEN WITHOUT CONSULTING A TUTOR OR CODE WRITTEN BY OTHER STUDENTS - Jaekyum Kim

2 Problem description

Sports video games series "FIFA" has the most recent and accurate players' data so that players' in-game abilities reflect their real-world performance. In order to provide the most realistic gaming experience to users, each of player's data is represented as a combination of 33+ features, which includes both physical(power, acceleration, etc) and technical(dribbling, shooting, etc) skills. In this project, we use each player's data to analyze the team's strategy and to predict the outcome of the game. We acknowledge that sports prediction project has numerously been implemented by others in the past; the idea itself may lack originality but we hereby state that all work including algorithm implementation and data pre-processing was done without any help or assistance from outside sources.

3 Data description

The original data were collected from www.kaggle.com/hugomathien/soccer. The data contained +25,000 matches +10,000 players in 11 European Countries with their lead championship Seasons 2008 to 2016 Players and Teams' attributes sourced from EA Sports' FIFA video game series, including the weekly updates Team line up with squad formation (X, Y coordinates) Betting odds from up to 10 providers Detailed match events (goal types, possession, corner, cross, fouls, cards, etc.) for +10,000 matches.

4 Data pre-processing

There were two parts in this project's pre-data processing, each of them corresponding to the different roles. However, both data types were prepossessed using sqlite library as the downloaded data were already in a .sqlite format.

First part of data was used to analyze the types of team strategies. Second part of the data pre-processing involved massive calculation, filtering, and adjustments majorly due to incomplete segments of data. Below is the small chunks of the source code, "function compiler", that presents the general guideline of the data processing.

```
def compiler():
        sql = "SELECT * From MATCH WHERE id > 10000 AND id <13000"
        src = "MATCH"
        sql1 = "SELECT * From Player_attributes"
        src1 = "Player_attributes"
        data_1 = connection(sql, src) # MATCH
        data_2 = connection(sql1, src1) # Player_attributes

        role = ["GK","DF","MF","ATK"]
        attack = [4,10,11]
        midfielder = [4,10,15]
        defense = [4,15,18]
        goalkeeper = [4,37,41]
        total_ = [goalkeeper,defense,midfielder,attack]
        kvalue = 3

        ############ Function begins here ############
        matchAttributes_ = matchAttributes(data_1)//
        abilityExtract = playerAbilityExtract(matchAttributes, data_2,
        role, total)
        criteria = criteriaMaker(abilityExtract, role, kvalue)
        prediction = playerComparison(criteria, matchAttributes, data_2,
        role, total)
        returnVal = winloseChecked(matchAttributes, prediction)
        return(returnVal)
```

"connection" method was used to load data from .sqlite file and convert it into enumerable array for further computation. No further filtering or customization was done before fetching the data since calculating the outcome of match required interaction of different features of players.

Initially, "Match" data were raw and chaotically vectorized, but based on our inputs, "matchAttributes" method was used to extract essential information such as teams that played against each other, api of players involved in the Match,

XY squad formation, and goals scored by each team. Then, based on the squad formation, players are distributed into different positions, for example, into attack, defense, midfielder, and goalkeeper.

"playerAbilityExtract" was used to further extract the stats of each players on both teams involved in a match with a data of players' apikey.

"criteriaMaker" method was used to make a tier of stats based on players' positions, and "playerComparison" method begins calculating the prediction for the outcome.
These functions compile in order as provided above on "compile" method.

5 Data mining

Given the data of 380 matches that occurred in 15/16 season in the English Premier League, each of two squads (Home vs. Away) in every match in the league is represented as a combination of the features of the players. To define this aggregation function, on-field player is represented as a set of 33 features which are split into 6 main categories like Table I.

TABLE I: Game data feature list

Attacking	Skill	Movement	Power	Mentality	Defending	Goalkeeping
Crossing	Dribbling	Acceleration	Shot Power	Aggression	Marking	GK Diving
Finishing	Curve	Sprint Speed	Jumping	Interceptions	Standing Tackle	GK Handling
Heading Accuracy	Free Kick Accuracy	Agility	Stamina	Positioning	Sliding Tackle	GK Kicking
Short Passing	Long Passing	Reactions	Strength	Vision		GK Positioning
Volleys	Ball Control	Balance	Long Shots	Penalties		GK Reflexes

Then for each team of 11 players, top four Attacking, top five Skill, Movement, Power, Mentality, top four Defending, and top 1 Goalkeeping attributes are calculated and averaged according to its 6 main categories. Since there are usually 4 defenders and attackers, only top 4 Defending/Attacking attributes are chosen.
Before running K-means algorithm on the aggregate data, normalization is required since top teams are likely to have higher scores than the ones for weaker teams in every attributes. To purely identify which of 6 main features have been emphasized, each of aggregate vectors are normalized by making sure all the scores add up to 1. Thus the number for each category for its Strategy will be represented in proportion to all the other attributes in the combination.

To first see the validity of K-means algorithm, we visualized and identified the clusters with 2 component PCA with the acquired squad data of 760 unique team vectors (380 matches * two teams(home vs. away) = 760). We could see that there are some overlaps with yellow and red clusters since it's in two dimensions, and, in the reduced dimensional space, it is easy to see which strategies

seems to be similar to each other and which seems to be not the case.

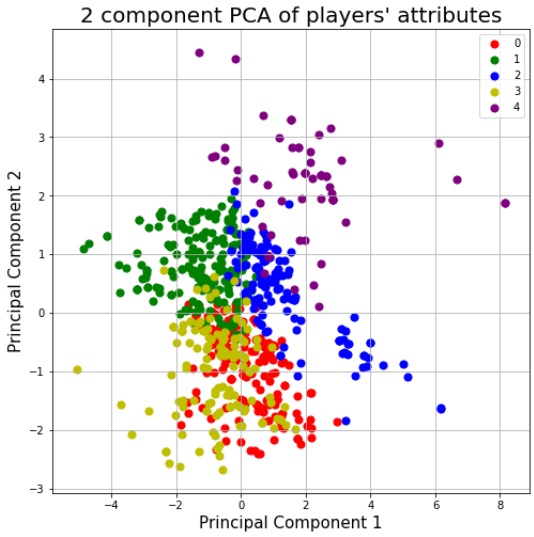

Figure 1: PCA of squad Strategies for each squad played in the 380 matches of the season.

After generating K-means algorithm and identifying 5 unique clusters that teams in the league seems to be generally using, winning odds were analyzed to see which strategy is more plausible in certain situations. We calculated which clusters are applicable for home team and winning team and calculated the odds by going through all of 380 matches. Notice on the "Results" section that the odds were calculated through the home team's perspective, and we took the account that are numbers could be very different depending on the opponent and the stadium (including the fans).

Further more, We calculated kmeans cluster with players' stats to evaluate them for further prediction of outcome. However, it would not be a biased comparison if players whose role is goalkeeper are evaluated based on their skills on crossing or finishing. In order to resolve this bias, we designed clusters of each positions based on the strength and aggregation shown on the extracted data by collecting players' focused stats and running kmeans algorithm. This formulated well distributed tier, which then worked as a criteria to relatively evaluate players against others in similar positions After identifying distinctive features of players, we created three clusters on each positions to classify players into corresponding tier. We first averaged the sum of squared of featured stats on each positions and matched them to the nearest cluster. Then we evaluated the difference between two teams' matches cluster points. We then summed them and compared to predict which team will win.

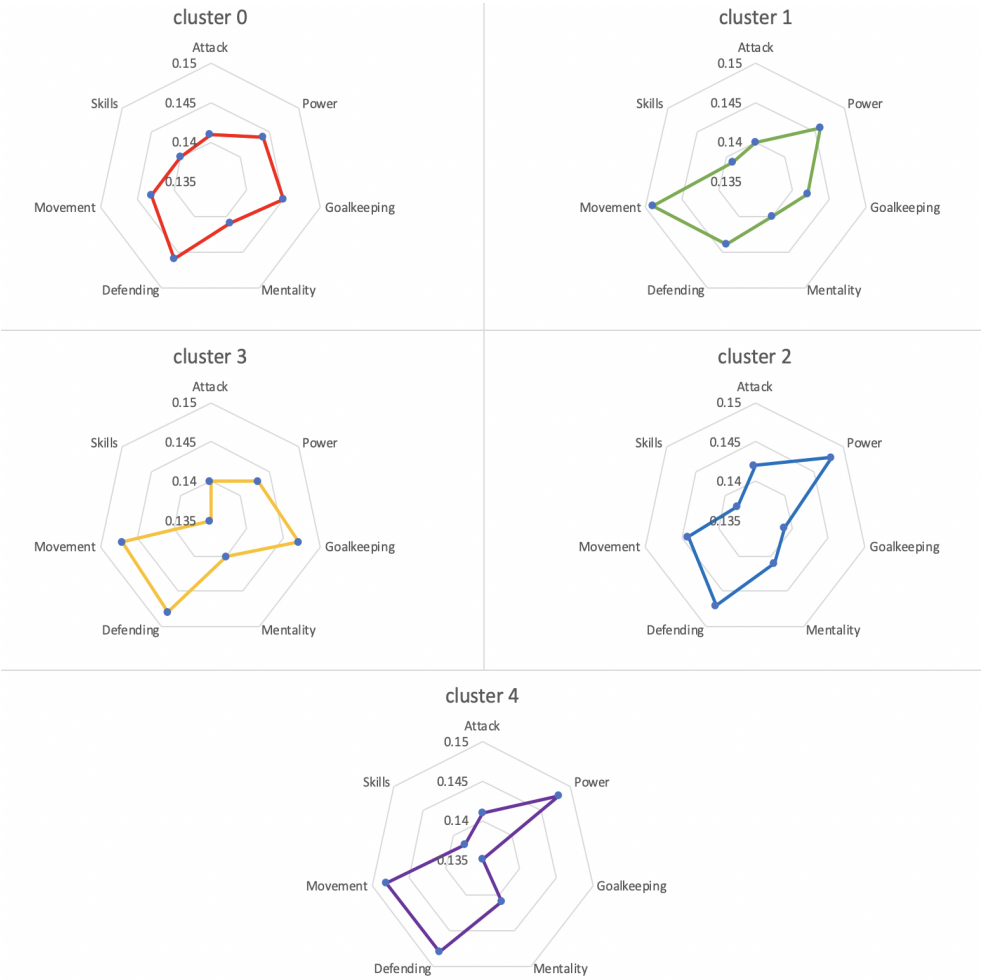

Figure 2: Cluster 0 to 4 represents a unique strategy. Radar graph shows how different skills are emphasized in each of strategies.

6 Results

home/away	cluster0	cluster1	cluster2	cluster3	cluster 4
cluster0	.24 (5/21)	.52 (14/27)	.61 (14/23)	.29 (5/17)	.71 (5/7)
cluster1	.25 (6/24)	.39 (9/23)	.54 (14/26)	.43 (6/14)	.55 (6/11)
cluster2	.19 (4/21)	.21 (3/14)	.22 (2/9)	.22 (4/18)	.88 (7/8)
cluster3	.24 (6/25)	.63 (15/24)	.46 (6/13)	.53 (9/17)	1. (4/4)
cluster4	.45 (5/11)	.57 (4/7)	.6 (3/5)	.1 (1/10)	0 (0/1)

The actual final league table in 15/16 season seemed to be stunning, Le-

icester City won the championship for the first time in their 132-year history. Many commentators consider the miracle to be one of the most shocking news in history, especially considering that the team spent half of the last season at the bottom of the table before finishing 14th. For our calculated winning odds, the cluster/strategy 0 and 3 seems to be the most effective to face other teams' tactics. Considering that Leicester City used strategy 3 for 20 out of 38 possible matches in a season, their winning of the season seems to be reasonable.

Showing the difference of impacts between focusing on featured stats versus regular stats was the purpose of the project. Speaking of the prediction of results, when players were distributed into tiers based on clusters created by general stats, accuracy rate of prediction ranged between 35 and 38 percent. However, when we created clusters using the focused stats differentiated by positions, the accuracy rose about 4 5 percents, ending up in mid 40s area. Although the it showed poor prediction accuracy, it has been proven that the using clustered stats and focused attributes tends to help the aggregation.

Soccer is a sports that have countless factors to consider when predicting an outcome. Conditions of players, humidity of weather, loudness of field, and even the subtle factors such as quality of foods consumed by players in the morning all could function as critical factor affecting the outcome. However in this project, we only considered player's physical and technical attributes, ignoring all the external sources such as environment, which I believe was our huge limitation.

7 Future work

Predicting the outcome of sports game is not an easy task. Many unexpected happenings can happen from a match such as player getting sent off or scoring an own goal. Also, strategic trends of soccer tactics change so fast each year because Leicester City finished off 12th place after the following year. This series provided accurate and realistic data of soccer players, but if the data of each player had more than 33 features, we believe that we can come up with better outcome of the data.
For further work on top of anaylzing and keeping up with the latest trends of professional soccer strategies , I believe this prediction accuracy can also be vastly improved simly by collecting more data with more centralized and focused key attributes. Thus, we will explore more to discover crucial factors to add on to our prediction model.