Premier League data analysis

Richard Yang
gyang79@wisc.edu

Zhoujingpeng Wei
zwei74@wisc.edu

Jiaxin Hu
jhu267@wisc.edu

Abstract

This project aims at investigating the connection between soccer strategies and the result of the game in Premier league data. The research helps us understand how strategies would work considering teams' own style and their opponents' style. The model uses PCA and clustering on information extraction in assessing the different aspects(attack, defence, domination) of a team. Then we use softmax or logistic regression and their statistical meaning to explain the relationship between game results and different features, in which we also take the both teams' style in a game to explain our model. Finally, we use the most intuitive way–decision tree to classify the model, and try to find corresponding rule to explain our idea.

In conclusion, we put much emphasis on the model explanation in this project, and get good and interpretable results.

1. Introduction

1.1. Background

Soccer is the most charming sports worldwide, not just because it's a sport showing good skill, reasonable strategies, but also the hot-blood atmosphere as well as the impressive team culture. And the World Cup is the most typical representative, which excites the world every 4 years.

In the World Cup of 2018, France won the final champion, and it's their strategy that really impresses the world: sometimes France even gives up the ball possession and use counter-attack as their only way to attack. Conversely, it seems that the team with high general performance, like good ball possession or pass accuracy, could not usually win the game.

This world cup also let us recall some very common phenomenons in modern professional league: the high game domination or shot attempts usually leads to a bad outcome, which is kind of counter-intuitive.

1.2. Motivation

There are so many unexpected results in the World Cup, does that mean the data for general performance is counter-intuitive in modern game? or it's just some coincidence? Is choosing the conservative strategy always a good idea? These questions stimulate our curiosity to figure out the true inner logic(which would be extremely tricky) of the soccer strategies and the game result. Moreover, instead of assuming all the games happen in the same background, we also want to find the ideal strategies under different scenarios including varied opponents and the special styles of the team.

In a nutshell, we hope our research can reveal an understandable and reasonable inner relationship between the soccer strategies, gaming scenarios and the result of the game. Therefore, our research would provide some useful guidance for teams to better choose strategies to attain ideal result facing different opponents.

In addition, the accuracy of classification would not be the crucial part of our project, for purely pursuing the accuracy would lead to overfitting and the randomness of soccer game is too strong, which isn't quite predictable.

2. Related Work

2.1. Problem Statement

As individual features cannot reflect the style of a team intuitively, we would firstly find some features to describe that. Based on the experience, we would like to describe a team style from three aspects: attack, defence and dominating. It's not necessary to construct precise model to assess a team in each aspect, but to find teams having similar performance in each aspect and assign the group a level. Then we can use categorical level features to distinguish the style of teams.

Then we want to find the connection between the result and features. Because the response variable, the result of a game including win, lose and tie, is a categorical variable, we should find a proper model connect the probability of the result with features. Intuitively, the probability of these three results would also have different relationship with features.

Last, we also would like to apply our research to the reality scenario and show a more understandable and interpreted result.

We have read some introduction paper to get some intuition and make sure the statistical meaning of our model is proper for our goal.

2.2. Implement Steps

Based on our problem statement and goals, we separate our project into three steps.

- Extract ”style” features from original data through classification from three aspects.
- Statistically analyze the relationship between game result and features.
- Try other method to show the result more directly and understandable.

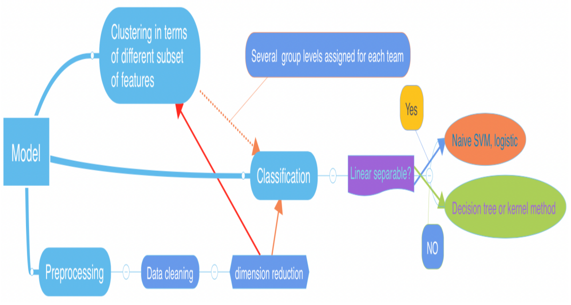
Model
Clustering in terms of different subset of features
Several group levels assigned for each team
Classification
Linear separable?
Yes
NO
Naive SVM, logistic
Decision tree or kernel method
Preprocessing
Data cleaning
dimension reduction

Figure 1. The flow chart for the whole process.

3. Proposed Method

Figure 1 draws the general picture for research process of our project.

3.1. Feature Extraction

To determine which model to use, we should analyze the data structure first. Then decide how to extract the information.

Data Collection In this research, we use the Premier league data. To extract the style of a team, we need the average performance of the team. So we collect the whole year data of each team’s games . Then we get an overall data for each team in one row which represent the average performance of this team this year. We take Arsenal as an example:

Sub-group features Because original features will represent different aspects of the performance, we divide features to sub-groups (Table 2) :

PCA Because there is a strong multicolinearity in each sub-group, so we apply PCA(Principal Components Analysis) to these features and extract suitable components.

	Shots	ShotsOT	Dribbles	Fouled
Arsenal	15.6	6.2	10.1	9.9
	Goals	Offsides	ShotsCon	Tackles
Arsenal	74	2.3	11.1	16.4
	Intercept	Fouls	Possess	Pass Acc
Arsenal	11.3	10.1	58.5	84.3
	AerWon			
Arsenal	17			

Table 1. Average data of the example team

Attack	Shots,ShotsOT,Dribbles, Fouled,Goals,Offsides
Defence	Shots Conceded,Tackles, Interceptions,Fouls
Dominating	Possesssion,Pass Accuracy, Aeriel Won

Table 2. Features divided in sub-groups

Clustering PCA is a process that projects the features to another space, thus the meaning of components would be hard to explain, and thus hard to give a score to each team. Therefore, we use clustering to let the teams with similar performance in one aspect in the same group. The specific method we use for clustering is K-means.Then add the categorical result of clustering to the original data, which to some extent add ”style or level” information for each team.

3.2. Statistical Analysis

Obviously, game result should be a multinomial distribution. The response in the model should be the probability of win,lose and tie. Therefore, we choose **softmax** to implement this multinomial model, which makes it easy to explain.

Softmax regression [3] The probability of different results varies according to features which includes the information of different game statistics and the style of both teams in a game. In softmax model, assumption equation is that:

$$P_r(y = j|X) = \frac{exp(X^T W_j)}{\sum_k exp(X^T W_k)}$$

we can analyze the coefficient to determine the relationship between each feature and game results (of course we scale the data first to make coefficients comparable).

For example, if W_{j1} is positive, the probability of j th class is positively correlated with feature X_1.

Improvement

- To fit and assess a model more precisely, we use 10-folds stratified cross-validation, and choose the L_2 penalty to implement our algorithm, and average the accuracy to get the score.

- First we want to add the label assigned previously into original features in the model. Then we would like to consider interaction term(That is to consider different team-opponents pairs respectively)

3.3. Decision Tree explanation [2]

In reality, sometime we cannot calculate the specific probability for three kinds of result, so we need to get intuitive rule.

Decision Tree Decision tree has the most intuitive way to do binary classification, and its leaf nodes correspond to rules, which can give us most clear explanation. In addition, decision tree can help us to better understand the statistical result from softmax regression and directly show the useful strategies facing different situations.

And we don't use the three types of game results as our response, because it's too specific. Instead, we combine two of three results and generate a binary response, and the reason is quite intuitive—The strong team need wins, and the weak team just need points.

4. Experiments

4.1. Dataset

As mentioned in the previous part **3.1 Feature Extraction**, we use the Premier league data.

Structure The structure of the data is also shown in that part.

Source We find the data from official website of Premier League and the website of Whoscore. the game data has less feature than average performance data. They are from different dataset.

Software Python

Hardware PC

4.2. Some choice of our model

In order to make our experiments repeatable,, we need to make sure the following issue.

4.2.1 General

- Although we use regression model, we choose accuracy as measurement about how model fits our data. And we use 10-fold cross-validation to give a score of our model. And because our goal is to extract the information from data, we finally use the score on the whole training set to make analysis.

- We sometimes combine two of the response variables (win, lose, draw) to form a binary response , and make inference.

4.2.2 Softmax

- We may consider the interaction between softmax coefficients with teams' levels, that is to consider different team with different levels respectively.

- We use l_2 penalty, and use the Newton-cg Solver to optimize the softmax likelihood.

4.2.3 Decision tree

- In the decision tree, we just choose the most informative model to explain.

- Because the data isn't balanced, we use the balance option in sklearn.

5. Results and Discussion

5.1. Feature Extraction

As mentioned in the previous part, there are 3 steps to implement our goal to describe the style of a team from attack,defence and dominating using categorical labels(variables): 1.Averaging the whole year data of each team;2.Using PCA to find the main components to represent the main information on each aspects;3.Clustering the team from each aspects and assigning the category labels.

Averaging data The example of averaging data is also show in the previous part.We take the average data of Arsenal as example.

PCA The result of PCA [4] can be show in figures:

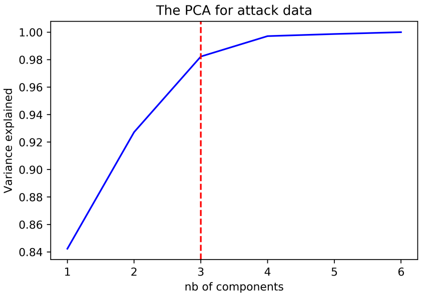

Figure 2. PCA result on Attack

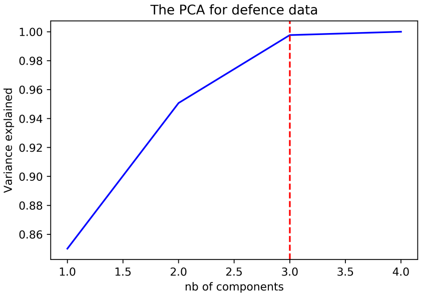

Figure 3. PCA result on Defence

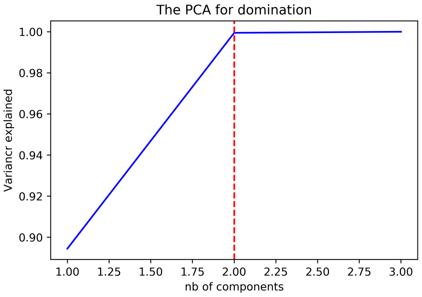

Figure 4. PCA result on Domination

In each figure above, the y-axis refers to the percentage of variance explanation by components while x-axis refers to the number of components. To guarantee the components can represent the most information of data, we choose 3 components on attack with 98% variance explanation ratio, 3 components on defence with nearly 100% variance explanation ratio and 2 components on dominating(control) with nearly 100% variance explanation ratio. Then we plug these hard-to-explained components to implement clustering.

Clustering [1] We use the components from PCA to do KNN clustering. And after trying some times, we choose the parameter of group number $k = 4$ for when the value of k is too large the cluster won't make sense. The result can also show in figures:

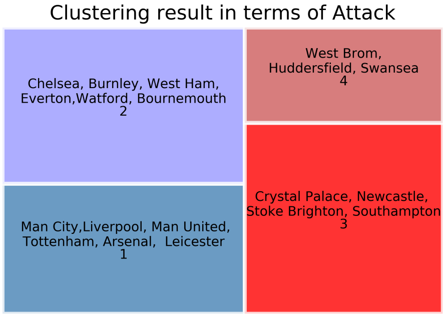

Figure 5. Cluster result on Attack

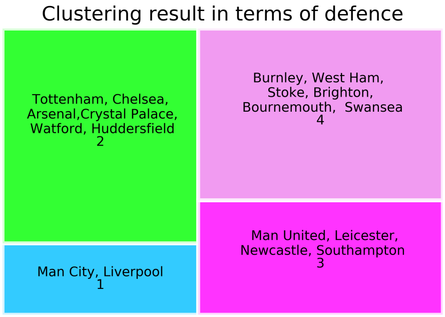

Figure 6. Cluster result on Defence

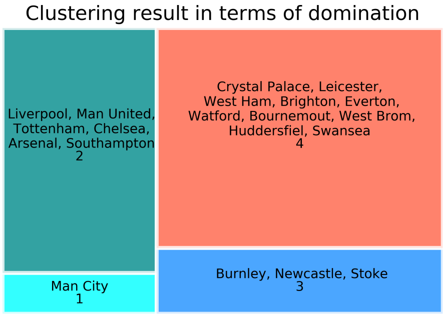

Figure 7. Cluster result on Domination

According to the figures, we can see that in each aspect, the teams are divided into four groups. Then we can use the clustering result to assign labels to each team and add the labels to the original data. Take one game data of Arsenal when the opponents is Brighton as a example:

Table 3. Game data for Arsenal when vs Brighton

Shots	ShotsOT	Fouls	corners
25	12	7	6
passacc	aerialsuc	possession	tackle
88	70	65	15
Attack	Defence	Domination	
2	2	2	
OpAttack	OpDefence	OpDominate	
3	1	1	

Note When we see the result of clustering, in fact, we can not say that the performance level is totally correspond to the index of the group. In the other word, we can not say that all teams in Attack group 1 attack better than all teams in Attack group 2. The criterion we judge the general performance of a group is tricky – using experience. The reason why we do that is that clustering itself is hard to given a specific meaning of reality when we use PCA components, which means different groups probably represent different style of attack rather than purely attack level. However, for interpret, when we say something like ”opponents are good at attack”, it means we consider the opponent team is in an Attack group contains other teams that are good at attack considered by most people.

5.2. Statistical Analysis

We first just consider plugging the label assigned in the step of clustering back in the corresponding game statistics (We treat the labels as nominal features), and fit the general model, it achieve the accuracy of 61% in training data ,and 58% in cross validation. Because the success for random guess is just $\frac{1}{3}$, and the games result is actually really tricky to predict, it’s not a bad result.

In softmax, coefficient differs when team-opponent varies, and more precisely, we believe that the style of team and opponent would also have effect on coefficients. Therefore, we would like to analyze different team-opponent pairs respectively.

We just analyze some common scenarios, in this report, we show the softmax model coefficients which can imply the connection between strategies and the result in following three scenarios, opponents are good at attack, good at dominating and when a general stronger team facing a weaker team.

In the figures, y-axis should be the value of coefficients and x-axis should be the name of each features.And in one plot, we can see the difference of the coefficient of the same features when we consider various game result.

The games with high attack level opponents:(Figure 8) When confronting teams with strong attack level (group1), the probability of win or tie is negatively correlated with shots, shots on target and ball possession, which means that teams need a conservative strategies to gain a ideal result.

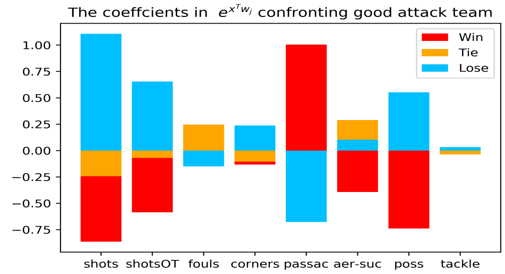

Figure 8. Coefficients facing high attack level opponents

The games with high domination level opponents:(Figure 9) When confronting teams with strong dominating level (group1,2),the results are similar as the strong attack one, because teams in these two kinds of groups have a strong overlap. And we guess the proper strategy is counter-attack.

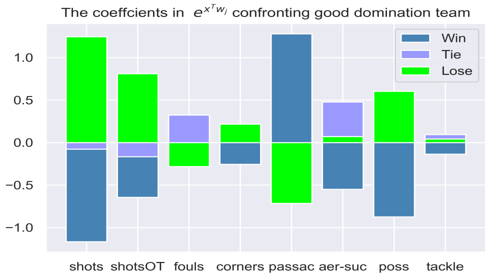

Figure 9. Coefficients facing high dominating level opponents

The game stronger team vs weaker team:(Figure 10) For a high ranking team, having a tie with a weak team has the same effect as losing a game, so we just consider the probability to win (using regular logistic regression). According to the graph, we know the probability to win is positively correlated with shots, shots on target, aerial-ball success, which means the high ranking team should use an aggressive strategy to win the game.

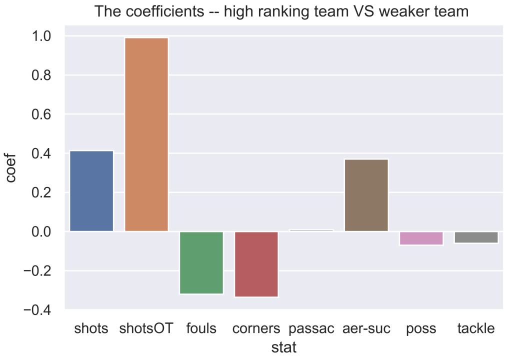

Figure 10. Coefficients of win when stronger team facing weaker team

5.3. Decision Tree

In the decision tree, we know every leaf node represents a rule, so we can find some useful information from it. As mentioned in the previous part, sometime we just consider the result of win or lose, so we can construct two binary decision tree.

We try different models, and plot following two graphs that best explain the logic, and the accuracy is 0.74 and 0.76 respectively, which can well explain the potential information.

The result of decision tree can be extracted using following graphics:

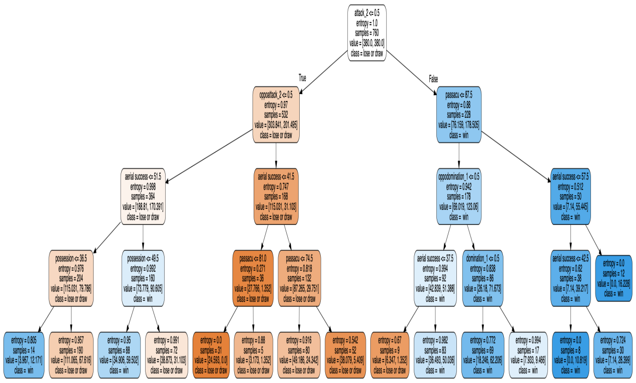

Figure 11. Decision Tree:Win vs lose&draw

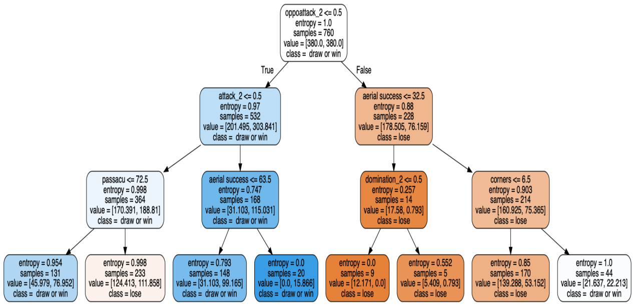

Figure 12. Decision Tree:Win &draw vs lose

5.3.1 Some explanation

Because we treat the level of team as nominal data, the attack_2 in the graph means the dummy variable indicating whether the team's level is 2(The strongest level in our coding), it corresponds to Attack1 group in our clustering graph, the same thing holds in other level features.

Additionally, we can find the first factor of this tree is about the style of team, then we can find most informative leaf nodes and their corresponding rules to explain the phenomenons happening during a game:

In these graphics, we do binary classification, the blue box means the most team in the box win the game while the orange box the most team in the box lose or draw the game. The darker the color is, the purity is higher in the box.

5.3.2 Analysis

- Strong attack team(The right part of the root in the first tree) would lose or draw the games only when (1) its pass accuracy is not extremely high (2) but its domination level is high !! (3) aerial success rate is low. **That means these team don't play games in their own style and is forced to play in an unfamilar way.**
- When confronting strong attack (The right part of root in the second tree) level teams, the only predictable way to gain points (draw or win) is to (1) have a high aerial-success (2) have many corners .

6. Conclusions

According the analysis above, we have different strategies for different level of teams to improve their wining chances

6.1. For a weak team

Our strategies are aimed to draw or win since we can at least gain some points.

1. We should make more chances to get corners.
2. We should guarantee the aerial-success.
3. Take a conservative attack and be patient to flooding attack from the opponents

6.2. For a strong team

1. Use an aggressive way to attack and avoid serious errors in the game.
2. Guarantee the attack or pass accuracy.
3. Guarantee aerial success rate in defence.

7. Acknowledgements

Thanks to our class instructor: Sebastian Raschka. His guidance and suggestions help us better finish the project because our thought is not very clear at first.

8. Contributions

During the project, our group members have done our own job and eventually finished the project together.

- Richard Yang is mainly response for the idea the implement of the model;
- Zhoujingpeng Wei is mainly response for data collection and pre-processing;
- Jiaxin Hu is mainly response for some graphics and the writing of the report.

Every members' work is indispensable. The project won't be finished smoothly without our cooperation and anyone's contribution.

References

[1] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):881–892, 2002.

[2] D. M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 276–283. Association for Computational Linguistics, 1995.

[3] S. Tao, T. Zhang, J. Yang, X. Wang, and W. Lu. Bearing fault diagnosis method based on stacked autoencoder and softmax regression. In *Control Conference (CCC), 2015 34th Chinese*, pages 6331–6335. IEEE, 2015.

[4] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.