Classifying Soccer Matches in the English Premier League

Justin Wang, Peter Shi

Abstract

Soccer is a very exciting sport mainly because of its unpredictability. Teams often counted as favorites do not always prevail against lesser teams as they pull off miraculous upsets. With the outcome of each match being unexpected, we wanted to see if outcomes of soccer games (wins, draws and loss) can be classified using features relevant to current game form (shots, shots on target, corners, home advantage). We also decided to incorporate features that detail past performances or "streakiness" to see if such can further improve our prediction. Using machine learning classifiers such as Naïve Bayes, Multinomial Bayes, Random Forests, Linear SVM (support vector machines) and RBF SVM, we found that soccer matches despite the high uncertainty of outcomes can be predicted with accuracy of 0.68. We found that football matches are better classified when we incorporate past game performances or a streakiness factor.

1. Introduction

Soccer is one of the most popular sports in the world, with the English Premier League being the most watched league with 4.7 billion viewers during the 2011/12 season. Every season, 20 teams strive to attain the best record while the worst 3 teams get relegated to a lower division. With such large backing from many fans across the world, teams are pressured to do their best and win the championship. With this in mind, our study strives to address the following questions: (1) do performance measures characterize a winning team and predict a match result? (2) does the addition of "streakiness" or past performances improve our model?

1.1 Data

Because we wanted to see if classification can be effective in the most unpredictable league, we looked into the English Premier League, where the championship team changes more often than other leagues such as the Bundesliga and La Liga. Our data contained the matches of 20 teams from the 2010/11 season to the 2014/2015 season. Our data was taken from the following website: http://www.football-data.co.uk/englandm.php [1]

Given that every team plays each other twice throughout the season, we have results from 380 games per season – totaling to 1900 games. Since we are training our model and validating the results, we set our test set to be the 380 games from the 2014/15 season while the training set consisted of 1520 games from the remaining seasons.

Our data consists of the date, the teams participating, the result, performance measures (goals, shots, corners, shots on target), and fouls of the corresponding game. It also contains the odds of the home team winning which are calculated by many online betting services.

1.2 Initial Analysis and Challenges

One challenge is the lack of data on injuries, roster availability data, and the psychological factor of facing an extremely dominant or easy team. For instance, we would like to investigate the effect of individual star players who potentially have a huge role on the team's performance and observe if their absence can alter a match's results. Since it is obvious that wealthy teams such as Manchester United, Arsenal, and Liverpool are more successful due to their abundance of star players, it is unfortunate that we cannot factor in this variable. Moreover [2] uses a Bayesian dynamic linear model to predict the outcome of matches while incorporating a psychological feature measuring the difference of strength between teams that we cannot access with our data.

Lastly, classification of match results may be difficult due to the unpredictability of the game in particular with the presence of draws where neither the home team wins or loses. We calculated the measure of unpredictability for the match results where uncertainty for each season is measured by the ratio of inverse odds.

$$U = \frac{odds\ of\ winning^{-1}}{odds\ of\ winning^{-1} + odds\ of\ losing^{-1} + odds\ of\ drawing^{-1}}$$

Therefore, in a perfect random scenario where the odds of winning losing and drawing are equal, this measure should be 0.666.

Figure 1 shows that the uncertainty for each season is extremely high for every season available in the data signifying that it may be extremely difficult to classify draws.

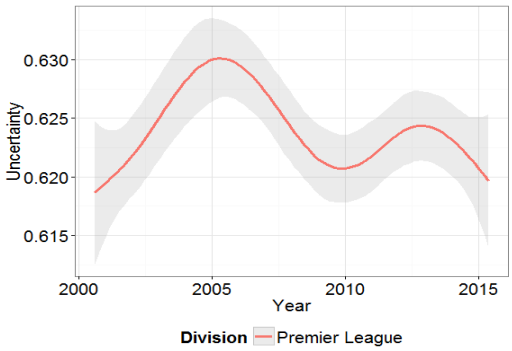

Figure 1: Uncertainty of prediction over the 2000-2015 seasons

2. Feature Selection

Evidently the performance features that affect the game results (win, draw and loss) are goals, shots, shots on target, corners. If the home team has more shots and shots on target, this suggests that their offense is creating more chances for goals and ultimately the win. In contrary letting the away team take more shots and shots on target implies that the home team's defense is giving chances for the opponent thus losing the game. The same logic should apply to corners as well. Although the number of goals is obvious in determining the winner of the game, we decided to exclude this variable as a performance metric. We excluded the variable of goals for home and away teams because it is obvious that the more goals a team scores they will win the game, thus there is no point of classifying wins based on goals scored.

Other features provided in the dataset such as yellow cards, fouls and red cards were not included as performance features because they reflect behavioral implications. Oftentimes teams end up fouling more since they take risky defensive schemes.

2.1 Past Performances

We also select features that show the past performances of a team. This is important because we want to access whether we can better classify wins, draws, and losses if we include these features. Since this study aims to find a discrete categorizations of game results, rather than sentiment analysis as in [3], we hypothesized that Machine Learning techniques such as Naïve Bayes and SVMs would be effective in predicting the game outcomes. This would include the goals, shots, shots on target, and corners in the past n games for both the away and the home team. Here, we can include past goals because these goals do not necessarily determine the result of winning losing or drawing the current game. Rather it shows whether a team is consistently good at scoring prior to the match. The following formula shows the scheme that we take to calculate these features where a_i represents the number of shots, shots on target or corners taken by a specific team i games prior.

$$\boldsymbol{Past\ Performance} = \sum_{i=1}^{n} \boldsymbol{a_i/n}$$

Although our training set included data from 4 seasons, we only calculated the past performances of teams for each season due to (1) 3 teams are relegated at the end of each season (2) the trade market opens during the offseason thus teams can acquire key players (3) current players can become incrementally better or worse during an offseason due to many factors that we cannot account for.

Ultimately, we selected the past performance measure as the mean of the past 5 games. Doing so has led to other problems that we had to consider, specifically in regards to the

beginning of the season in which a team has not played 5 games yet. Although we initially decided to take the mean of the past m games such that $m < 5$, we realized that our model suffered an increase in error. Thus we excluded these games from our data whenever we modeled using these features. This makes sense since teams need adjustment in the beginning of the season.

2.2 Recent Record

Lastly, we wanted a feature to determine the recent record of a team for the past 5 games. This would measure the form of a team prior to a match. Although it is possible to calculate the percentage of wins or losses before the game, this ignores the instances of draws. By assigning a win to be worth 3 points, a loss to be worth 0 and a draw to be 1 similarly to soccer standings, we took the sum of the points for the past 5 games to determine the team's recent performance. This is preferential because it does not ignore the outcome of a draw.

3. Procedure

3.2 Analysis on current performance

We ran our multiclass classifiers on training data (2010/11, 2011/12, 2012/13 and 2013/14 seasons) containing the current match features (home shots, away shots, home shots on target, away shots on target, home corners and away corners), to predict if the home team wins loses or draws. We assigned the outcome variable of win, draw and loss to be 1, 0 and -1. Then we will compare our prediction on the test data (2014/15 season) to find the error rate of our classifiers in which we will conclude if the outcome of matches can be accurately predicted by these features.

3.3 Analysis on current performance and past performances

We ran our multiclass classifiers on training data (2010/11, 2011/12, 2012/13 and 2013/14 seasons) containing the current match features (home shots, away shots, home shots on target, away shots on target, home corners and away corners) as well was the features of recent record and past performances. Similarly, we assigned the outcome variable of win, draw and loss to be 1, 0 and -1. Then we will compare our prediction on the test data to find the error rate of our classifiers.

3.4 Comparison

We compare the error rates of both models in their respective classifiers and then access if adding features on past performances would better contribute to predicting match outcomes. Furthermore we can observe which classifier predicts the most accurate. We will also look at the accuracy of predicting wins losses and draws and see if adding a past performance metric will increase this accuracy through the use of confusion matrices.

4. Results and Conclusion

The classifiers we used are Gaussian Naïve Bayes, Multinomial Naïve Bayes, Random Forest, Linear SVM (one vs rest), RBF (Gaussian kernel) SVM and Linear Discriminant Analysis. All of our models are multiclass as we are classifying a three class model.

	Gaussian Naïve Bayes	Multinomial Naïve Bayes	Random Forest
Error	0.547	0.524	0.516
	Linear SVM	RBF SVM	LDA
Error	0.466	0.526	0.465

Table 2: Analysis on current performance

	Gaussian Naïve Bayes	Multinomial Naïve Bayes	Random Forest
Error	0.409	0.390	0.329
	Linear SVM	RBF SVM	LDA
Error	0.327	0.439	0.339

Table 3: Analysis on current performance and past performance

4.1 Accuracy of the best classifier for both models

	Predicted Loss	Predicted Draw	Predicted Win
Actual Loss	78	2	17
Actual Draw	28	4	43
Actual Win	16	3	139

Table 4: confusion matrix for linear SVM

	Predicted Loss	Predicted Draw	Predicted Win
Actual Loss	75	12	10
Actual Draw	25	21	18
Actual Win	14	18	126

Table 5: confusion matrix for Random Forest

From the full model, we found that the Linear SVM predicts with the best accuracy (0.673), followed by the random forest classifier (0.671). We find that the model with the past performance and recent record feature surpasses the previous model with a better accuracy of 0.14 for linear SVM, 0.13 for LDA and 0.18 for random forests, clearly showing that past performances enhances our prediction of match results. These results are consistent with [4] that found SVM to be more accurate than naïve bayes and other methods. Interestingly, the accuracy ratings are comparable with past studies on predicting sports outcomes, such as with basketball in [5]

Table 4 and 5 shows the confusion matrices for the classifiers Linear SVM and random forests on the full model with past records and performance metrics. Since our problem solves a 3-way classification in which the features only reflect on whether a team has performed better or worse than its opponents, it is reasonable why our model fails to predict draws accurately. However, it certainly does predict losses with an accuracy of .77 and .80 for Random Forest and SVM respectively and wins with .79 and .88 for random forest and linear SVM respectively. The accuracy of draw predictions are only 0.32 for the random forest and .05 for the SVM which is significantly worse.

4.2 Conclusion

By adding features on past performances we can easily establish a better model for classifying whether the home team wins, lose or draws. Accuracy of full model (past performances included) increases significantly for each respective classifier compared to the model with only current performances. Thus, we can establish that teams on a streak who perform better in the past can replicate this success moving forwards.

Sources

[1] "England Football Results Betting Odds | Premiership Results & Betting Odds." *England Football Results Betting Odds | Premiership Results & Betting Odds*. N.p., n.d. Web.

[2] 06 Dec Rue, Havard, and Oyvind Salvesen, “Prediction and retrospective analysis of soccer matches in a league” Journal of the Royal Statistical Society: Series D (The Statistician) 49.3 (2000): 399-418

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79-86. Association for Computational Linguistics, 2002.

[4] Faria, Brigida Mónica, Luis Paulo Reis, Nuno Lau, and Gladys Castillo. "Machine Learning algorithms applied to the classification of robotic soccer formations and opponent teams." In *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conference on*, pp. 344-349. IEEE, 2010.

[5] Miljkovic, Dragan, Ljubiša Gajić, Aleksandar Kovacevic, and Zora Konjovic. "The use of data mining for basketball matches outcomes prediction." In*Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*, pp. 309-312. IEEE, 2010.