Game ON! Predicting English Premier League Match Outcomes

Aditya Srinivas Timmaraju
Stanford University
adityast@stanford.edu

Aditya Palnitkar
Stanford University
aditpal@stanford.edu

Vikesh Khanna
Stanford University
vikesh@stanford.edu

Abstract

Among the different club-based soccer leagues in the world, the English Premier League (EPL), broadcast in 212 territories to 643 million homes, is the most followed. In this report, we attempt to predict match outcomes in EPL. One of the key challenges of this problem is the high incidence of drawn games in EPL. We identify desirable characteristics for features that are relevant to this problem. We draw parallels between our choice of features and those in state-of-the-art video search and retrieval algorithms. We demonstrate that our methods offer superior performance compared to existing methods, soccer pundits and the betting markets. We also share a few insights we gained from this interesting exploration.

1 Introduction

The most popular soccer league in the world is the English Premier League (EPL), which is based in England. It was watched by an estimated figure of 4.7 billion people in the 2010-11 season. So, it doesn't come as a surprise that the TV rights are valued at £1 billion per season. In EPL, there are 20 teams that contest for the first place. The bottom three teams are relegated and replaced by other teams from lower leagues who perform better. Each team plays every other team twice, once at home and once away. So, there are a total of 380 ($= 2 *^{20} C_2$) games per season. A season runs from August to May of the following year.

1.1 Challenges

One of the things that makes predicting outcomes tricky is the significant incidence of draws (neither team wins, if they both score the same number of goals) as compared to other sports. Most popular games viz., tennis, cricket, baseball and American football either have no draws or very few draw occurrences. Consider this - out of the 380 games in the 2012-13 season, 166 games were won by the home team, 106 games by the away team and there were 108 draws! To put these numbers in perspective, we calculate the season's entropy. We regard home wins(1), away wins(2) and drawn games(3) as three separate classes and compute the fraction of each of these outcomes.

$p_1 = 166/380\ p_2 = 106/380\ p_3 = 108/380$

Since, we have a 3-way classification, we use the base-3 logarithm.
$Entropy = -(p_1 * log_3(p_1) + p_2 * log_3(p_2) + p_3 * log_3(p_3)) = 0.9789$

An entropy of 1 would correspond to a perfectly random setting ($p_1 = 1/3, p_2 = 1/3, p_3 = 1/3$). So, considering this, a random prediction would give an expected accuracy of 33.33%. Another naive, but a marginally better approach would be to predict a home win always, which would deliver

44-45% accuracy (about the average fraction of home wins in EPL). These are however, not very good numbers.

1.2 Prior Work

In [1], a Bayesian dynamic model is built, but there are many parameters and the authors do not justify the choice of their values, which may not work well with other testing sets. Also, out of the 48 EPL games they bet on, they won on only 15 games [1]. Their primary focus is towards computing pseudo-likelihood values for betting on match outcomes. In [2], the authors specifically choose to focus on predicting match outcomes of a single team - Tottenham Hotspur. Also, their model is dependent on the presence of particular players and is thus, by their own admission, not scalable. There is no standard way of reporting results for this problem. Most authors do not mention prediction accuracy directly, some specify precision/recall and some others specify geometric mean of predicted probabilities of correct classes, called Pseudo-likelihood Statistic (PLS). We report the accuracy of predictions, precision/recall as well as PLS. We compare the performance of our algorithms on all these fronts and show a marked improvement.

The remainder of the report is organized as follows. Section 2 discusses the choice of features, which we feel is a key step. In Section 3, we give a description of the approaches we have taken towards addressing the problem. Section 4 describes our results. In the remainder of the report, we use the words *match* and *game* interchangeably. By convention, in "A vs. B", A is the home team and B is the away team.

2 Selection of Features

We first identify the most relevant performance metrics: Goals, Corners and Shots on Target. Goals are an obvious choice as they determine which team wins. As for corners, a higher number of corners indicate a team playing well, and thus enforcing a corner kick from the opponent. Shots on Target convey the number of times a shot taken was on target (including those stopped by the keeper). We consciously did not pick possession percentage, because some teams just have a more possessive style of play than others, and moreover, higher possession percentage does not necessarily imply better performance. So, our performance metric vector for a game is a 3-element vector of the form (Goals scored, Corners, Shots on Target).

We recognize three important characteristics for a feature. It should:

- Incorporate a notion of the competing nature (between teams) of the problem
- Be reflective of the recent form of a particular team
- Manifest the home advantage factor
- Capture the improving/declining trend in the relative performance of competing teams (i.e., has team A been on the rise more than team B? Has team B slumped in form more than team A?)

Keeping the above characteristics in mind, we propose the following choice of features (we progressively arrived at the above list and so, the first choice below does not capture all these properties).

2.1 KPP (k-Past Performances)

In KPP, we use the performance of a team in the last k games to determine our prediction. In the game A vs. B, we first take an average of each performance metric of team A in the last k games to arrive at P_A. If $G = [g_1, g_2, g_3, ..., g_k]$ is a vector containing the number of goals scored in each of the last k games, we have $g_{avg} = \frac{(g_1+g_2+...+g_k)}{k}$. Similarly, we arrive at c_{avg} and st_{avg} for corners and shots on target respectively. Thus, we have $P_A = [g_{avg}; c_{avg}; st_{avg}]$.

Similarly, we arrive at P_B. We then take the ordered difference P_A-P_B, a 3-element vector, as our feature. The ordered difference is chosen to inherently include information about which team is playing home and which team away.

2.2 TGKPP (Temporal Gradient k-Past Performances)

In TGKPP, as in KPP, we first arrive at the 3-element vector P_A-P_B. Now, consider the performance metric (say Goals scored) vector for team A in the past "k" games. Denote it as $G = [g_1, g_2, g_3, ...g_k]$ We now apply a temporal differencing operator on this vector, which is essentially a convolution with a filter of the form "$diff = [1; -1]$". We now have

$$g_{dA} = convolution(G, diff) = (g_2 - g_1, g_3 - g_2, ..., g_k - g_{k-1})$$

We do this for the other performance metrics and compute c_{dA} and st_{dA}. We repeat this process for team B and arrive at g_{dB}, c_{dB} and st_{dB}. Then, we compute

$$P_{diff} = [\mu(g_{dA}) - \mu(g_{dB}); \mu(c_{dA}) - \mu(c_{dB}); \mu(st_{dA}) - \mu(st_{dB})]$$

where $\mu(*)$ indicates the standard mean of elements in a vector. Thus, our final feature vector is a 6-element vector of the form $[P_A - P_B; P_{diff}]$. This feature is similar to the video fingerprint in [7], where spatial and temporal gradient are applied to blocks in a frame of video. We have used the website `http://www.football-data.co.uk/` for obtaining our data.

3 Approaches taken

We first toyed with the idea of using Naive Bayes, but found that independence is a very strong assumption in this problem. Also, we did not try using Gaussian Discriminant Analysis because our data was nowhere close to being normally distributed. Owing to space constraints, we have omitted including figures that showed this. We used multiple variations on Multinomial Logistic Regression and Support Vector Machines.

3.1 Approach 1

The first approach we took was using Multinomial Logistic Regression (since there were more than 2 possible outcomes). In this approach, during the training phase, we only considered the performance metrics derived from the current match, rather than taking the average over the last "k" matches. During testing, suppose we are required to predict the match outcome of team A vs team B, we arrived at the feature vector using KPP.

3.2 Approach 2

In the second approach, we trained in the same way we would later test the data. More precisely, in the training phase too, instead of using the feature vector as the performance metric vector corresponding to the current match, we used KPP. This meant that the trained parameters now inform our beliefs about the result of a match based on the performance in last "k" matches. In this approach, we also used TGKPP instead of KPP and evaluate our approach.

3.3 Approach 3 : Using teamwise models

So far, all approaches we tried would find a global set of parameters, which were independent of the competing teams. So, given the past "k" performances of the team playing home and the team playing away, our model was agnostic to the identity of the actual teams playing. However, we felt we maybe missing out on some team-specific trends using this method. So, in this approach, we trained different models for different teams. However, this approach placed a limitation on the data

Table 1: MLR with Approach 1

Season	k value	Accuracy
2013-14	3	34.43
2013-14	4	32.79

Table 2: Prediction Accuracy: KPP & TGKPP

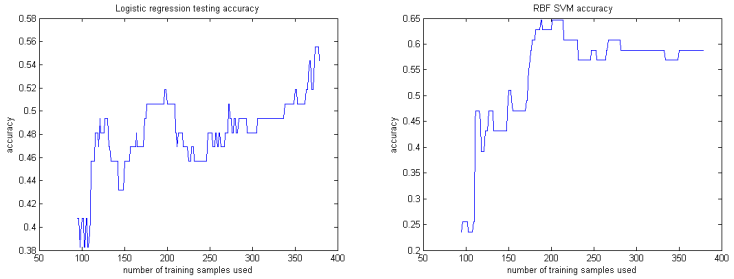
Season	k	MLR(KPP)	RBF-SVM(KPP)	MLR(TGKPP)	RBF-SVM(TGKPP)	2 Class (RBF-SVM)
2013-14	4	58.21	46.91	54.32	55.56	76.79
2013-14	5	56.86	59.15	59.15	54.93	76.17
2013-14	6	57.38	54.10	59.02	54.10	76.74
2013-14	7	**60.78**	**58.82**	**56.86**	**66.67**	**83.78**

we could use to test/train our model. We could no longer combine data from two different seasons, due to the form of the teams varying between seasons, and major players being traded between teams. So, due to the limited data, and the increased noise induced by increasing the granularity in the model, we ended up getting a lower accuracy (an average of 47%).

4 Results and Discussion

We train all our models on the 2012-13 season and test on the 2013-14 (current) season. Table 1 contains the prediction accuracy (in %) obtained using MLR with Approach 1. Figure 1 contains the learning curves for the two algorithms, MLR and RBF-SVM. Our definition of "k" imposes a restriction on starting point for testing. For k=4, 5, 6, 7 we test on 81, 71, 61, 51 number of games respectively. Table 2 contains the prediction accuracy (in %) obtained using Approach 2 with MLR and RBF-SVM. The last column of Table 2 contains accuracy obtained using only Home win and Away win classes, disregarding the draws. Basically, if C is the confusion matrix, we truncate the 3rd row and 3rd column to obtain C_{trunc}. We then compute $2class accuracy = trace(C_{trunc})/\sum_{i,j} C_{trunc,ij}$. We computed this because we felt it would offer a fair metric for comparing against accuracy in other sports which do not have the concept of "drawn games".

As our problem is a 3-way classification, there is no positive and negative class demarcation. So, while computing them for the "Home wins" class, we regard both the other classes as negatives. Table 3 lists the precision and recall values. While the precision of "Drawn games" is fairly okay, it is evident our method misses out majorly on recall value for draws. Under-estimating draws is a problem with the existing methods too.



(a) (b)

Figure 1: Accuracy vs training set size (for k=7) (a)MLR (b)RBF-SVM

Table 3: Precision and Recall with RBF-SVM (TGKPP and k=7)

	Home win	Away win	Drawn game
Precision	71.43	50.00	54.10
Recall	86.21	66.67	23.08

Table 4: Confusion Matrix (TGKPP and k=7)

	Predicted Home wins	Predicted Home losses	Predicted draws
Actual Home wins	25	3	1
Actual Home losses	3	6	0
Actual draws	7	3	3

In [1], the authors propose a metric which is computed as the geometric mean of the predicted probabilites of actual outcomes, they call it "PLS" (Pseudo-likelihood Statistic). They obtain a PLS of 0.357 for EPL. In [3], the best value of PLS obtained is 0.36007. In comparison, our average PLS value is 0.4015, with a maximum of 0.4613 with k=4 using MLR and TGKPP. In [4], the authors report their results in the form of a confusion matrix. But they do not predict drawn games at all. Our method offer superior performance than [4] too. We also considerably outdo the accuracy of the probabilistic model in [5], which reaches an accuracy of 52.1% for 2011-12 and 48.15% 2012-11.

In comparison to our best accuracy of 66.67%, the accuracy of expert pundits like Mark Lawrenson of BBC is 52.6% [6]. Also, we outperformed the accuracy of betting markets which averaged 55.3% last season [6], suggesting possible avenues for making money!

The results we obtained, while interesting and significant in their own right, also offered interesting insights into the match data we analysed. Though our results using approach 3 were not encouraging, we found a striking trend. Since we have two models, we are faced with a choice: If A & B are the teams competing, should we use θ_A or θ_B for prediction? Or should we always use the home team's parameters? We observed that using θ_{away} for prediction always does better as compared to using θ_{home}. In hindsight, we may be tempted to interpret that using θ_{away} does better because there is more variability in an away teams performance, and so that gives us more information. But *a priori*, we might as well have thought that home performances are more consistent and so a better predictor would use θ_{home}.

5 References

[1] 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.

[2] Joseph, A., Norman E. Fenton, & Martin Neil, "Predicting football results using Bayesian nets and other machine learning techniques." *Knowledge-Based Systems* 19.7 (2006): 544-553.

[3] Goddard, John, "Regression models for forecasting goals and match results in association football." International Journal of forecasting 21.2 (2005): 331-340.

[4] Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002), "Dynamic modelling and prediction of English Football League matches for betting". Journal of the Royal Statistical Society: Series D (The Statistician), 51: 157168. doi: 10.1111/1467-9884.00308.

[5] Constantinou, Anthony C., Norman E. Fenton, and Martin Neil, "pi-football: A Bayesian network model for forecasting Association Football match outcomes." Knowledge-Based Systems (2012).

[6] www.pinnaclesports.com/online-betting-articles/09-2013/lawrenson-vs-pinnacle-sports.aspx

[7] Oostveen, Job, Ton Kalker, and Jaap Haitsma, "Feature extraction and a database strategy for video fingerprinting," Recent Advances in Visual Information Systems (2002): 67-81.