Classifying Soccer Match Results

Justin Wang, Peter Shi

STA561: Machine Learning-Professor Sayan Mukherjee

ABSTRACT

Soccer is an exciting sport because of its unpredictability. Less favored teams often miraculously upset more favored teams. Since the outcome of each match is unexpected, we wanted to see if outcomes of soccer games (wins, draws, and loss) can be classified using features relevant to the current game form (shots, shots on target, corners, home advantage). To improve our classifiers, we decided to incorporate features that detail past performances (i.e. "streakiness") to see if such can 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 despite the high entropy of outcomes, soccer match outcomes can be predicted with accuracy of 0.68. We found that soccer matches are better classified when we incorporate past results – also known as the "hot-hand" factor.

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.

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

INITIAL ANALYSIS

Any classification attempt may be ineffective due to the high entropy of the data. For instance in the 2014/15 season 172 games (45%) ended up with the home team winning, 115 (30%) with the home team losing and 93 (25%) with a draw.

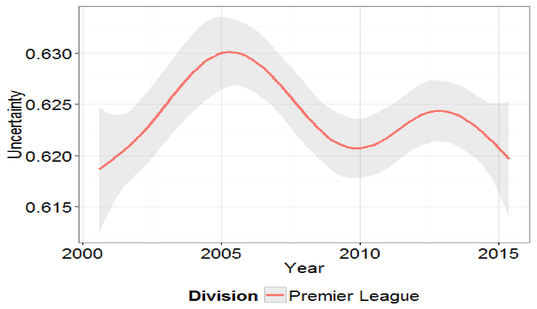
$$entropy = \sum_{1}^{3} \log(p_i)p_i$$

Which for the above case is 0.97 which shows near randomness for this season.

Moreover a look at the uncertainty of wins for seasons from 2010/11 to the 2014/15 in which the uncertainty value is measured by:

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

Shows a downward trend during the recent years however it does remain high showing that the games are nonetheless hard to predict



Furthermore, another challenge is that we do not have data consisting of individual players and their performances. For instance major trades, injuries on star players etc. during seasons can result in drastic changes in the team's record.

FEATURE SELECTION

We select features that are indicative of a team's performance. This includes goals, shots, shots on target, corners for both the home and the away team, but not fouls, yellow cards and red cards, since these are behavioral factors rather than performance factors.

We excluded the factors of goals because it is obvious that team with the high number of goals will win the game, thus there is no point of classifying wins based on goals scored.

We also select features that show the past performances of a team. This is important because we want to determine whether we can better classify wins, draws and losses if we include these features. This would include the goals, shots, shots on target, and corners in the past *n* games. At this point, we can include past goals because these goals do not necessarily determine the result of winning losing or drawing the current game. Instead, past goals shows whether a team is consistently good at scoring prior to coming into the current game.

We chose these features as the mean of the past 5 games in the same season in the following categories: goals, corners, shots and shots on targets. We avoid looking at teams from a prior season primarily because teams may get long breaks in between seasons or be relegated. For instance, QPR played in the 2011/12 season but failed to qualify for the 2012/13 season.

In addition, choosing the past 5 games creates the issue that the first *m* games (such that *m*<=*n*) played by a team for a season does not have enough past results. We considered taking the mean of the past *m* games; however, that model turned out to be less accurate. Ultimately, we removed all the data from these games because it makes sense that the earlier games of the season are more unpredictable as teams are adjusting to the new season.

Lastly we wanted to keep track of the recent record of a team up to the past 5 games. By assigning a win to be worth 3 points, a loss to be worth 0, and a draw to be 1 similarly to official soccer standings, we took the sum of the points for the past 5 games to determine the team's record. We thought this would be a better approach than simply taking the percentages of games won or lost because that approach includes the results of a draw.

.

PROCEDURE

We divided the data into a training set and a test set. The training set consisted of games of the **2010/11, 2011/12, 2012/13 and 2013/14** seasons. The test set consisted of data of the **2014/15** season. The results of the classification was home wins, home loses and draws in which we assigned values of 1, -1 and 0 respectively.

STEP I: Analysis on current performance

One model will include only features detailing performance factors of the current games. This includes the shots, shots on target, corners for home and away teams.

STEP II: Analysis on past performance

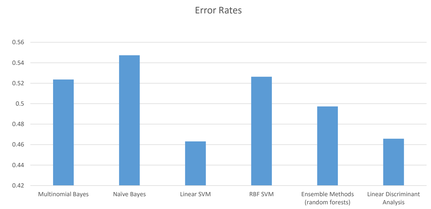
The other model will include the same features as the prior approach in addition to features of past performances and recent record as detailed in the Feature Selection.

STEP III: Comparison

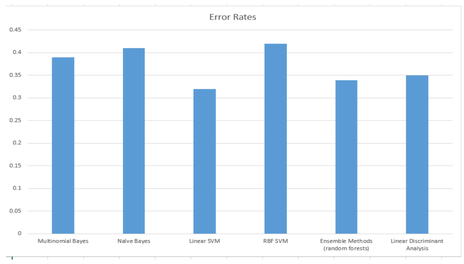
We compare the error rates of both models in their respective classifiers and then access if adding features on past performances would better contribute to predicting wins. Furthermore, we can observe which classifier leads to the best prediction and compare our results with existing methods.

RESULTS

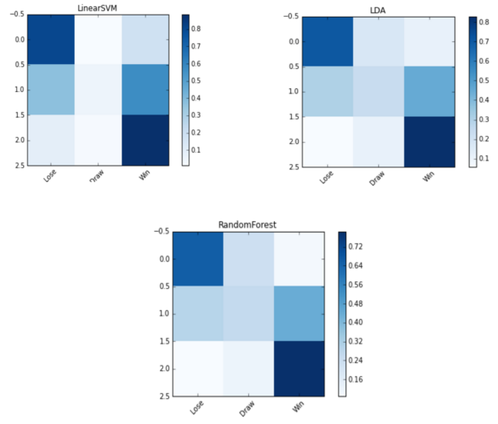
STEP I: Analysis on current performance



STEP II: Analysis on past performance



Confusion Matrix for Linear SVM, LDA and Random Forests for full dataset



CONCLUSION

–By adding features on past performances we can establish a better model for classifying whether the home team wins, lose or draws. Accuracy of full model (past performances included) increases about 0.10 for each respective classifier compared to the model with only current performances.

–Our best model (Linear SVM) has an accuracy of 0.68.

–The confusion matrix shows that there is high accuracy in predicting wins and losses correctly and a lower accuracy in predicting draws. This indicates that the features we use are effective at establishing non-parity between away and home teams.

FUTURE WORK

While we have obtained high accuracy in our models, we could further improve our models with industry information that includes performance variables such as active players, a factor on home advantage, and the influence on star players.

Furthermore, we would like to observe the influence of having a star player on the field and see whether his presence can lead to a team to win.