University of Massachusetts Dartmouth

CIS 602-01 Data Mining Project

Tu Luong, Hai Ha, Minh Nguyen

Predicting Results of English Premier League Matches Using Data Mining

May 10th, 2017

Contents

1. Introduction

In this project, we use methods of machine learning in order to predict result of soccer matches. Although it is impossible to consider all factors affecting to a match's result, we will try to extract most significant features. We test using many methods of classifiers to solve this problem. The aim is using data mining algorithms in combination with machine learning to predict results of soccer matches in English Premier League (EPL). The basic implementation is by gathering and processing of EPL data, analyzing soccer matches, classification using various algorithms. Although it is difficult to consider all features that effect the results of the matches, we will find the most significant features which are the most effective to the results (e.g. team current form, team players index, head-to-head history, etc.) and different classifiers are tested to attempt solving the problem.

Using sports statistics to make accurate predictions is a very valuable application of data mining. As the problem is one of classification problems, instead of the Bayesian dynamic model is built [1], but there are more than its parameters we think need to be considered in order to get better results and furthermore we can apply the model in others data set. Therefore, selecting features to analyze the data is the key point. Firstly, we must choose features which have the most significant in the results of the match, for example, in a particular match, features such as team's players' index, team's current formation, playing in home or away stadium will be significant impact to the final results. Secondly, extracting from data all the features we chose, then using machine learning methods, such as K-nearest neighbors (KNN), Support Vector Machine (SVM), and others to produce the output representative of the probable outcome of the match. Thus, the project is comsidered to be successful if it predicts the outcome with sufficient accuracy approximately 60%.

2. Related Work

Most of the work have been done by famous gambling companies for the benefit of oddmakers. However, there are also several groups have taken to the predicting game as well. We have explored two papers from Stanford University by Ben Ulmer, and Matthew Fernandez [2], and another from Saint-Petersburg State University by Albina Yezus [3].

The first paper focused on improving hyperparameters and class imbalance, thus their approach is using grid searches and ROC curve analysis respectively. The dataset includes gameday data and current team performance. This group used five different classifiers: Linear, Naive Bayes, Hidden Markov Model, SVM, and Random Forest to predict the outcome and get the best accuracy. Their best accuracy with Linear classifier 52%, Random Forest 50%, and SVM 50%.

In the second paper, the researcher used KNN, Random Forest, Logistic Regression, and SVM. Their algorithm implemented classifier from Python to get the best result. They reached a 55.8% success rate in KNN and another 63.4% in Random Forest.

Both papers explored several methods and found neural networks to generate the most successful.

3. Problem Definition

To create a model that predicts outcome of premier matches with highest accuracy. Highest accuracy means as follows:

- ~60% accuracy of match results.
- Benefit entertainment areas.
- Raising fund that concentrates from sport betting.

Our basic framework should be as following steps:

1. Choosing matches set data, normalizing and plotting data points;
2. Analyzing and selecting key features;
3. Data and features extraction;
4. Testing on various mining algorithms;
5. Implementing the most accurate algorithm;

4. Methodology

4.1 Deciding Match Set

We have many competitive soccer leagues in Europe to get data from, however, we choose English Premier League (EPL) from England and Wales. We choose this because although English Premier League is no longer the best soccer league in Europe, this is still considered the most competitive and attractive league with many surprises and unpredictable results. That's key reason we want to predict its results. EPL consists of 20 best soccer teams in England and Wales, each team will play against each other team twice (at home and when away) in a season, team winning a match will gain 3 points, drawing gains 1 point for each team, and losing is certainly 0 point, there are total 380 matches in a single season, after finishing all matches, team with most points win the league. The first 4 positions will attend UEFA Champions League (the highest-class league of all clubs in Europe) next season, the teams in 5^{th} and 6^{th} position will play

at Europa League (the second-highest league of all clubs in Europe) next season, 3 teams at the bottom of the table standing will be relegated to Championship next season.

There are some issues we have discussed why we collect data:

1. We want to avoid betting odds from famous betting companies. This is because each betting company has its own algorithm to predict the outcome and let people place bet, we want our features not relating any results of any betting algorithm, by that we ensure the final prediction is purely based on data.
2. There are some discussed features which have certain effect on match's result and may improve the accuracy of final prediction, such as team's squad in each match, the weather of each match, the style of play from each team, etc. However, due the limitation of data collection online, and also inside this project's constraint, we do not have enough time to collect, clean-up and extract all of the desired features.
3. When discussing, we want to avoid human errors in each match as much as possible, these human errors mostly come from the referees, such as false offside decision (incidents where players were/weren't offside but referee decided otherwise, this issue can have a massive effect to the match when that player scored), false penalty decision, biased referees (referees who are bised to a certain teams), match-fixing, etc. However, this is nearly impossible since human errors are essential part of a match.

4.2 Deciding on Key Features

In order to select key features for the dataset, and not only applicable to this particular soccer league, it can be extended to apply for other leagues, other seasons, too. To achieve that, our features should meet following conditions:

1. The features must be legit.
2. The features must be possible to extract without too much complicated algorithms.
3. The features must have significant effect to the match's results.
4. The features must not be computed based on any biased data or another algorithm.

It is noted that it's not always the case where more features produce more accurated results. Sometimes, more features decreases accuracy. That's why deciding which key features is crucial. Data extraction is based on internet source, which is widely available since soccer is very popular among Europe, the raw data file is in CSV format, and is parsed using JavaScript.

4.3 Testing Various Machine Learning Algorithms

The problem we are trying to solve is samples classification problem, in which each match is a sample, each factor effecting match's outcome is a feature, and labels are 3 values {0, 1, 2}, 0 means home team wins, 1 means draw result, and 2 means away team wins. For classification, several methods can be applied, in this project's constrainst, we applied following methods:

1. K-nearest neighbors
2. Support Vector Machine (Linear, Polynomial, or Gaussian kernel)

5. Algorithms

5.1 Key Features

5.1.1 Team form

This feature indicates the latest 10 games from each team. We simply calculate the mean of possible results in those previous 10 games, the result is in range [0, 1]

$$\frac{1}{10} \sum_{k=1}^{10} result_k$$

Where: $result_k$ is the outcome from k^{th} match, and its value lies in {0, 0.5, 1}

5.1.2 History head-to-head

This feature indicates one previous result between two teams, this is particularly useful in the last half of the season when each team has played against each other once. This feature simply returns result in {0, 0.5, 1}

5.1.3 Score difference

This feature indicates the score of two teams in the previous head-to-head match.

$$\frac{1}{2} + \frac{(|homeGoal - awayGoal|)}{2 * max(homeGoal, awayGoal)}$$

Where:

- $homeGoal$: number of goals scored by home team
- $awayGoal$: number of goals scored by away team

5.1.4 Motivation

Motivation features each team's motivation in each particular match. This can be hard to calculate since it involves human emotion, however we sum up into this formula:

$$min(max(1 - \frac{dist}{3 * left},\ derby,\ tour),\ 1)$$

Where:

- *dist* : distance to the nearest key position (position in {1, 2, 3, 4, 5, 6, 17, 18})
- *derby* : if this match is derby match, return 1, otherwise, return 0
- *tour* : return 1 if there are less than 6 matches left, 0 otherwise
- *left* : matches left in the season of this team

5.1.5 Concentration

This feature indicates the level of concentration for each team in a particular match. Each match will have a weaker and a stronger team. Thus, the level of concentration is different for each team (stronger team tends not to concentrate too much when playing against much weaker team). The algorithm is as follows:

- if played matches are less than 6
  return 1 for both weak and strong team
- if played matches > 6 and difference between teams position is < 7
  return 1 for weak team, and 1 / diff for stronger team
- if played matches > 6 and difference between teams position is > 7
  return 1 for weak team, and 1 / 7 for stronger team

By this way, the closer the two teams are, the more concentrated the stronger team is.

5.1.6 Home team winning ratio

This feature returns the winning ratio of home team in all previous matches playing at home, the return value in range [0, 1]

5.1.7 Away team winning ratio

This feature returns the winning ratio of away team in all previous matches playing away, the return value in range [0, 1]

5.2 Data Extraction

Then data was collected directly from https://github.com/footballcsv/eng-england, the format of return data is CSV format, and it is straightforward to understand with date of match, name of two teams, half-time and full-time score:

Date	Team 1	Team 2	FT	HT
2013-08-17	Arsenal	Aston Villa	1-3	1-1
2013-08-17	Liverpool	Stoke	1-0	1-0
2013-08-17	Norwich	Everton	2-2	0-0
2013-08-17	Sunderland	Fulham	0-1	0-0
2013-08-17	Swansea	Man United	1-4	0-2
2013-08-17	West Brom	Southampton	0-1	0-0
2013-08-17	West Ham	Cardiff	2-0	1-0
2013-08-18	Chelsea	Hull	2-0	2-0
2013-08-18	Crystal Palace	Tottenham	0-1	0-0
2013-08-19	Man City	Newcastle	4-0	2-0

Original data

When we finish extracting features from data, we export to 2 CSV files which are **fea.csv** (features) and **gnd.csv** (labels), a sample of **fea.csv** file has following representation:

1	history	diff	motivation1	movtivation2	form1	form2	concentration1	concentration2	homeRatio	awayRatio
92	1	0.25	1	0.988505747	0.5	0.45	0.428571429	1	0.75	0.4
93	0	0	0.988505747	0.965517241	0.3	0.3	1	0.571428571	0.25	0
94	0.5	0	0.966666667	0.977011494	0.3	0.25	0.571428571	1	0.333333333	0
95	0.5	0.5	0.954022989	0.988505747	0.3	0.45	1	0.142857143	0.4	0.75
96	0	0.166666667	0.988505747	0.977011494	0.5	0.4	0.142857143	1	0.4	0.25
97	0	0	0.988505747	1	0.3	0.35	1	0.142857143	0.6	0.5
98	0	0	0.988505747	0.977011494	0.5	0.3	1	0.571428571	0.2	0
99	0	0	0.988505747	0.944444444	0.35	0.3	1	0.714285714	0	0
100	0	0	0.988505747	0.988505747	0.2	0.35	0.142857143	1	0.8	0
101	0	0.25	0.976190476	0.976190476	0.45	0.35	0.857142857	1	0.5	0.2
102	0	0	0.952380952	0.988095238	0.35	0.5	1	0.142857143	0.25	0.8
103	0	0	1	1	0.45	0.4	0.142857143	1	0.5	0
104	0	0	0.954022989	0.988095238	0.35	0.3	1	0.428571429	0	0
105	0	0	0.988095238	0.964285714	0.35	0.35	1	0.142857143	0.2	0.25
106	0	0	1	0.988095238	0.3	0.4	0.428571429	1	0.25	0
107	0.5	0.5	0.976190476	1	0.5	0.2	1	0.142857143	0.2	0
108	0	0.125	0.988095238	0.976190476	0.5	0.55	0.142857143	1	0.8	0.25
109	0	0.333333333	0.988095238	0.964285714	0.4	0.6	0.571428571	1	0.8	0.75
110	0	0	0.964285714	0.976190476	0.35	0.35	1	0.857142857	0.6	0.25
111	0	0.2	0.975308642	0.987654321	0.45	0.6	1	0.857142857	0.4	0.6
112	1	0.25	0.962962963	0.975308642	0.6	0.5	0.857142857	1	0.166666667	0.4
113	1	0	0.987654321	0.962962963	0.4	0.45	0.142857143	1	0.833333333	0.166666667

Data after extraction

6. Technology

6.1 Programming languages

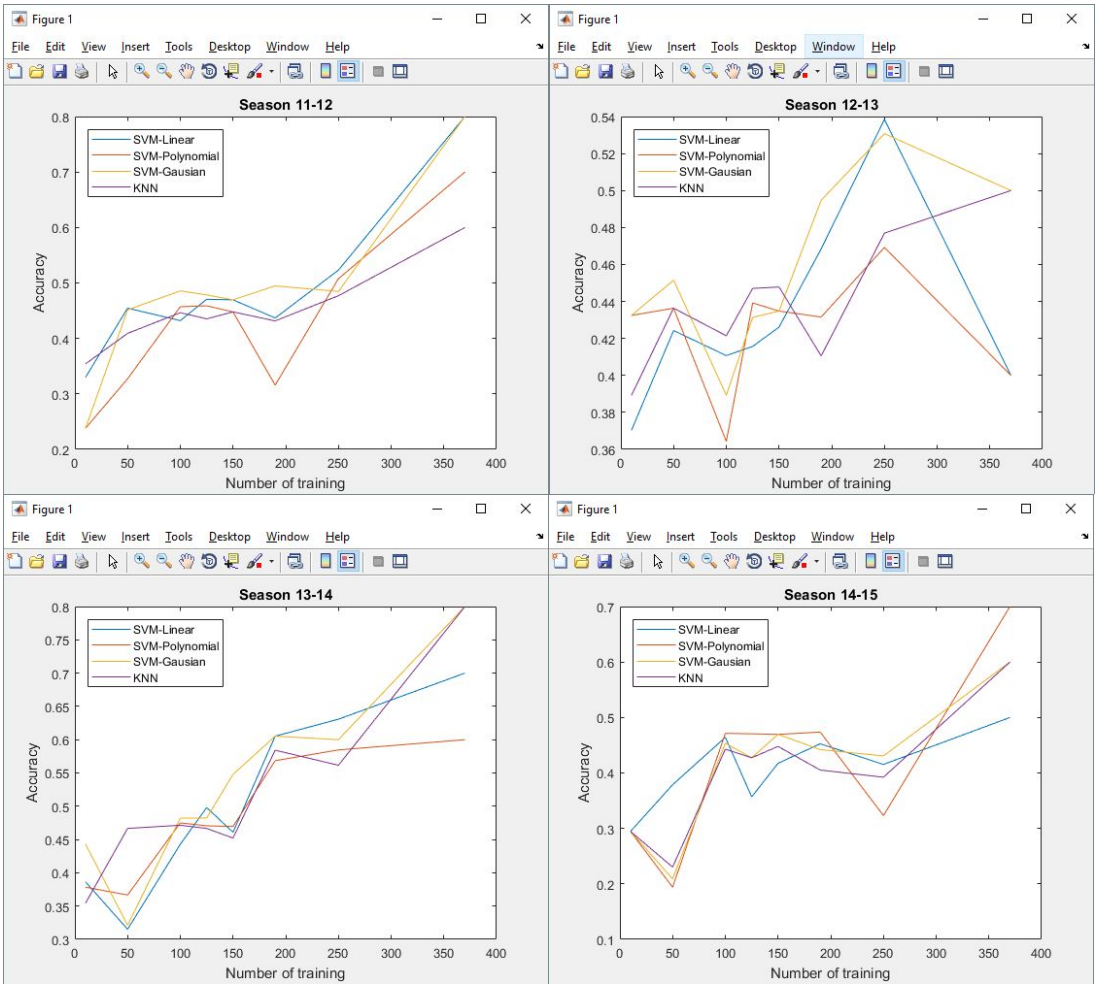
We use JavaScript as our main language to process, compute feature's formula and export to CSV file. We choose JavaScript because it is dynamic and not much strict to syntax, moreover, our team are much more familiar with JavaScript than other languages.

6.2 Libraries

In order to process data, we use d3.js (https://d3js.org/) from Mike Bostock (https://github.com/mbostock). D3 is great when processing and visualizing data. When classifying data in Matlab, we use Libsvm (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) to support our classification methods and compute accuracy.

7. Experimental Result

Predicting using SVM with 3 different kernels (5-folds, “gird search” with different slack C, gamma, and degree) and K-nn with k = 3 (number of classes):



Experimental results

We predicted in four seasons as above, in the total of 400 samples for each season, the number of training sample increased from 10 to 390 and the rest for testing. It is clearly to see that the accuracy will increase when we use more sample for training, and accuracy with SVM-Gaussian is usually highest and the most stable.

Even though, there is a season (season 12-13) has pretty low accuracy: highest is 0.54 and that is even worse when we use more samples to train. It might be caused by the lack of feature and the result of soccer matches is quite unpredictable in real life.

8. Conclusion & Further Work

In the future, we want to increase the number of features which have significant effect to a match to improve outcome accuracy. To do that, we need collecting more data and improve each feature's formula. One more thought we are discussing is that, if we can eliminate one possible outcome for a specific match (e.g. a match has possibility of drawing, home winning, home losing is very low), then we can improve the accuracy of that match.

This approach is to the prediction of soccer matches but also expect that it may have a broader applicability for many different kinds of sports. The latter refers to the potential of the proposed project to benefit entertainment area. Some people aim to give the public investors access to the traditional private world of sports prediction, since there is a fund that concentrates on sports betting should therefore be an attractive investment as some investors like sport, and like to have a bet. Gambling is illigal but it could generate more tax. If the bets were legal, the tax revenue would be enormous.

By using the prediction, we could turn soccer community into more interesting field. Including players who can look at the estimated result for a better preparation as well as their leader or managers to recognize the power of sport, not only by their knowledge, but by their ability to bring out the best in every member of their teams. Moreover, critic can have a better review before and after the match. For the recruitment, the project can be used as a simulation to check if the player is suitable for the current team. Cheating in a game is also detected if a player play too good as it could be a sign of drug usage.

9. References

[1] Rue, Harvard, 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] Ulmer, Ben and Matthew Fernandez, “Predicting Soccer Match Resutls in the English Premier League.” (2014).

[3] Yezus, Albina, “Predicting Outcome of Soccer Match using Machine Learning”. (2014), Saint-Petersburg State University, Mathematics and Mechanics Faculty.

[4] Predicting the outcome of NFL games using machine learning, Babak Hamadani, cs229 - Stanford University (2006)

[5] Numerical Algorithms for Predicting Sports Results by Jack David Blundell, School of Computing, Faculty of Engineering (2009)

[6] Predicting football results using Bayesian nets and others, machine learning techniques, A. Joseph ¤, N.E. Fenton, M. Neil (2006)

[7] Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line, Jim Warner (2010)

[8] A Review of Data Mining Techniques for Result Prediction in Sports, Maral Haghighat, Hamid Rastegari, and Nasim Nourafza, ACSIJ Advances in Computer Science: an International Journal, Vol. 2, Issue 5, No.6 , November 2013 ISSN: 2322-5157, www.ACSIJ.org

[9] Cao, C., “Sports data mining technology used in basketball outcome prediction”, Master dissertation, Dublin Institute of technology, Ireland, 2012.

[10] Buursma, D., “Predicting sports events from past results Towards effective betting on football matches”, Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, 21 January 2011.

[11] Zdravevski, E., Kulakov, A., “System for Prediction of the Winner in a Sports Game”, In: ICT Innovations 2009, Part 2, 2010.

[12] ”The World’s Most Watched League.” Web. 05 Apr. 2017. <http://www.premierleague.com/en-gb/about/the-worlds-most-watched- league.html>.

[13] A. S. Timmaraju, A. Palnitkar, & V. Khanna, Game ON! Predicting English Premier League Match Outcomes, 2013.

[14] ”England Football Results Betting Odds - Premiership Results & Betting Odds.” Web. 05 Apr. 2017. <http://football-data.co.uk/englandm.php>.

[15] “Team Stats Database - FIFA Index”, Web. 05 Apr, 2017. <https://www.fifaindex.com/teams/>

[16] “Football data for England (and Wales) incl. English Premier League, The Football League (Championship, League One, League Two), Football Conference etc.”, Web. 05 Apr, 2017. <https://github.com/footballcsv/eng-england>.

[17] “What Is the Best Method for Predicting Football Matches?”, Web. 15 Apr, 2017. <http://cartilagefreecaptain.sbnation.com/2014/3/5/5473358/what-is-the-best-method-for-predicting-football-matches>.

[18] “Can a formula predict the outcome of a soccer match?”, Web. 15 Apr, 2017. <https://phys.org/news/2010-03-formula-outcome-soccer.html>.

[19] “Machine Learning for Soccer Analytics”, Gunjan Kumar, Thesis, September 2013, University College Dublin, <https://www.researchgate.net/publication/257048220_Machine_Learning_for_Soccer_Analytics>.

[20] A. Tenga. Reliability and validity of match performance analysis in soccer : a multidimensional qualitative evaluation of opponent interaction. PhD thesis, Norwegian School of Sport Sciences, 2010.