Using Machine Learning to Predict Soccer Match Results with Scikit-learn

Minh Vo

Department of Computer Science, Earlham College

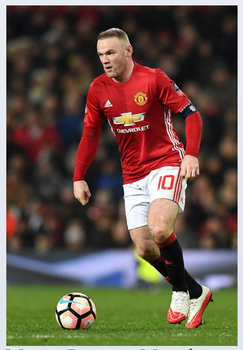
ABSTRACT

With the non-stop improvements in technology, more and more fields are trying to apply computer science to achieve their goals in a more efficient and less time consuming way. Sports are no outsiders to this group of fields. In sports, especially in soccer, technology has become an essential part. Soccer experts now make use of technology to evaluate a player's or a team's performances. Other than using their experience and their management abilities after many years being parts of the game, the soccer coaches also use statistics from data providers to improve their knowledge of their own players and teams so that they can come up with different strategies/tactics that bring them closer to the wins. Besides coaches, soccer analysts also make use of the data to predict results in the future as well as evaluate new talents emerging from the scene. This is where Machine Learning techniques can become useful. Machine Learning is one of the intelligent methodologies that have shown promising results in the domains of classification and prediction [8]. Therefore, Machine Learning can be used as the learning strategy to provide a sport prediction framework for experts.

BACKGROUND

1. What Is Soccer?

Soccer (or football) is a sport that involves two teams, each of which has eleven players on the pitch, including one goalkeeper. The basic idea of soccer is that to win a match, a team needs to score more than the other team.



Wayne Rooney. Manchester United. (Image by Getty)

2. English Premier League:

The English Premier League is the English soccer league that is at the highest order in the English soccer league system. It has been around for more than twenty years. The Premier League is often regarded the most exciting soccer league on the planet.

3. Data:

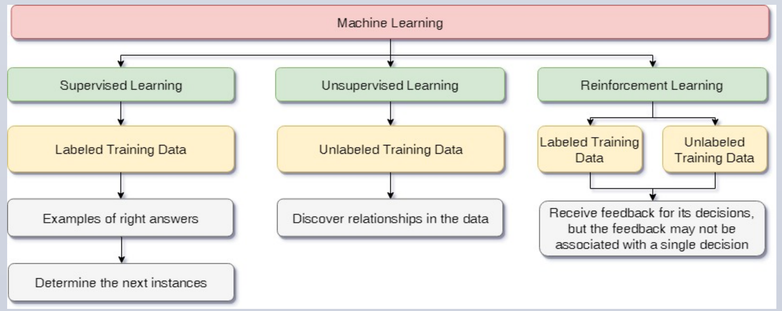
Buursma in [2], Constantinou et al. in [3], Timmaraju et al. in [4], Ulmer et al. in [5], and Tax and Joustra in [7] all did their researches/works using the data from one source, the Football-Data site, http://www.football-data.co.uk [9]. This is also where I am getting the data from. Below is an example of what the data looks like:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Div	Date	HomeTear	AwayTean	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS
2	E0	13/08/16	Burnley	Swansea	0	1	A	0	0	D	J Moss	10	17
3	E0	13/08/16	Crystal Pal	West Bron	0	1	A	0	0	D	C Pawson	14	13
4	E0	13/08/16	Everton	Tottenhan	1	1	D	1	0	H	M Atkinso	12	13
5	E0	13/08/16	Hull	Leicester	2	1	H	1	0	H	M Dean	14	18
6	E0	13/08/16	Man City	Sunderlan	2	1	H	1	0	H	R Madley	16	7
7	E0	13/08/16	Middlesbr	Stoke	1	1	D	1	0	H	K Friend	12	12
8	E0	13/08/16	Southamp	Watford	1	1	D	0	1	A	R East	24	5
9	E0	14/08/16	Arsenal	Liverpool	3	4	A	1	1	D	M Oliver	9	16
10	E0	14/08/16	Bournemo	Man Unite	1	3	A	0	1	A	A Marriner	9	11
11	E0	15/08/16	Chelsea	West Ham	2	1	H	0	0	D	A Taylor	16	7
12	E0	19/08/16	Man Unite	Southamp	2	0	H	1	0	H	A Taylor	12	13
13	E0	20/08/16	Burnley	Liverpool	2	0	H	2	0	H	L Mason	3	26
14	E0	20/08/16	Leicester	Arsenal	0	0	D	0	0	D	M Clattenl	8	13
15	E0	20/08/16	Stoke	Man City	1	4	A	0	2	A	M Dean	8	12

Football-Data Site. English Premier League Matches' Statistics of Season 2016 – 2017.

4. Machine Learning:

Machine Learning is the concept of learning from the data provided/given. It gives the computers the capability to learn without having to be set up manually, which means that the computer will determine the next steps to take from the past events/data.

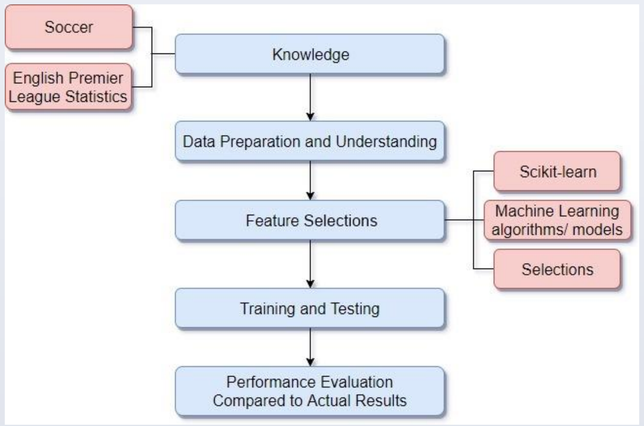


Three Types of Machine Learning.

In sports in general, and in soccer in particular, data is usually carefully labeled. This means that data in soccer has some kind of meaningful tag, or label, such as number of shots, fouls, points, etc. Therefore, the most common machine learning techniques in sports/soccer belong to the category of supervised learning. This is also the case for the data which I chose for this project.

FRAMEWORK AND METHODOLOGY

The framework for predicting soccer match results in the English Premier League include the following components: knowledge about the sport and the English Premier League, data understanding and preparation, feature extraction, training and testing, and finally, performance evaluation. This framework slightly follows the one proposed by Bunker and Thabtah [8].



Using Machine Learning for Soccer Prediction Framework.

Scikit-learn [10] – The Tool:

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [1]. The most important characteristic of the Scikit-learn is that it provides the users with the algorithms for machine learning tasks including classification, regression, dimensionality reduction, and clustering. It also provides modules for extracting features, processing data, and evaluating models [6].

Currently used Machine Learning algorithms (also used by researchers who have worked on soccer prediction):

- Decision Trees; Bayesian Networks; Logistic Regression; Support Vector Machine.

Selections:

One way to select the features or statistics that can have positive influence on the accuracy of the predictions in this problem is to follow the basic understandings of the data in soccer. For examples, home advantage, the number of goals scored or goals conceded can be deciding factors in the results of soccer matches. However, the selection process must not stop here. The features needs to be checked every time it is observed and based on the results, the list of selections can be changed.

Training and Testing:

After having attained an adequate set of selections, we go into training and testing. The process during which we let the machine learning algorithm calculate the probabilities for the matches is called training [2]. It is important to preserve the order of the training data for the sport prediction problem, so that upcoming matches are predicted based on past matches only [8]. With training and testing, we can see not only the accuracy of the algorithms but also their efficiencies based on the selections. This allows us to determine the best algorithm for the set of features selected.

PRELIMINARY EXPERIMENT RESULTS

After having gathered the data, I tested the 4 algorithms, Decision Trees, Gaussian Naïve Bayes, Logistic Regression and Support Vector Machine. The features I selected for this preliminary experiment were Home Team, Away Team, Half-time Home Goals, Half-time Away Goals, Home Shots, Away Shots, Home Shots on Target, and Away Shots on Target. Moreover, I only used the data from season 2013 – 2014 to this season, 2017 – 2018.

Algorithms	Accuracy
Decision Trees	56.98%
Gaussian Naïve Bayes	63.79%
Logistic Regression	64.78%
Support Vector Machine	54.15%

Preliminary Experiment Results.

From the results above, we can see that Gaussian Naïve Bayes and Logistic Regression performed better than Decision Trees and Support Vector Machine. However, the results may change drastically with the addition of more data and the alteration of the feature selections.

CONCLUSION AND FUTURE WORK

Machine Learning has been considered as one of the most efficient approaches in the problem of classification and prediction. From the preliminary experiment, we can confirm Machine Learning algorithms also have positive outcomes in predicting soccer match results (with accuracies of over 50%).

I plan to keep working with these four algorithms with more data (from season 1999 – 2000 instead of just from season 2013 – 2014) as giving the algorithms more training data can greatly improve their performances. Another or a few other algorithms can be explored as well, such as Gradient Boosting, Linear Regression, and Artificial Neural Networks.

After I have got the algorithm with its best dataset, one direction to follow in the future is to create a simple interactive program/script that allows the users to choose the home team and the away team to for prediction. Another option can be letting the users choose to predict the matches of one team for the whole season.

REFERENCES

[1] Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, Oct. 2011, p. 28252830.

[2] 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, vol. 21, 2011.

[3] Constantinou, Anthony C., et al. "Pi-Football: A Bayesian Network Model for Forecasting Association Football Match Outcomes." Knowledge-Based Systems, vol. 36, 2012, pp. 322-339.

[4] Timmaraju, Aditya Srinivas, et al. Game ON! Predicting English Premier League Match Outcomes. 2013.

[5] Ulmer, Ben, et al. Predicting Soccer Match Results in the English Premier League. Ph. D. dissertation, 2013.

[6] Hackeling, Gavin. Mastering Machine Learning With ScikitLearn. Packt Publishing, 2014.

[7] Tax, Niek, and Yme Joustra. "Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach." Transactions on Knowledge and Data Engineering, vol. 10, no. 10, 2015, pp. 1-13.

[8] Bunker, Rory P., and Fadi Thabtah. "A Machine Learning Framework for Sport Result Prediction." Applied Computing and Informatics, 2017.

[9] Football Betting — Football Results — Free Bets — Betting Odds. http://www.football-data.co.uk/. Accessed 9 Nov. 2017.

[10] Scikit-learn: Machine Learning in Python. http://scikitlearn.org/stable/. Accessed 16 Nov. 2017.

ACKNOWLEDGEMENT

This work was supported by the Earlham College Computer Science Department as part of the Senior Capstone Experience Course. Special thanks go to Charlie Peck who, as my Methods of Research and Dissemination Course instructor, led me in the right direction to determine the best topic for me as well as the necessary tools for the project. I would also like to thank David Barbella for his guidance and instructions as my topic advisor, which is keeping me on track to complete my Senior Capstone project.