English Premier League Soccer Matches Prediction with Scikit-learn

Minh Vo
Computer Science
Earlham College
801 National Road West
Richmond, Indiana 47374
mbvo14@earlham.edu

I. Introduction

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 [7]. Therefore, Machine Learning can be used as the learning strategy to provide a sport prediction framework for experts.

The end goal of this project is to produce a program/script that will automatically execute the complete procedure of results prediction.

II. Related Prior Work

In this section, a few attempts at predicting soccer results will be briefly reviewed to provide the basic ideas and design for this project. The prior researches have provided me a wide range of knowledge (discussed in the sub sections below). However, the researches which I have come across did not automate, or at least did not mention the automation of, their process of prediction.

A. Data Source

There are a few sites that actually provide data publicly although the data provided are only on the statistics of matches, not players. Buursma in [1], Constantinou et al. in [2], Timmaraju et al. in [4], Ulmer et al. in [5], and Tax and Joustra in [6] all did their researches/works using the data from one source, the Football-Data site, http://www.football-data.co.uk [8]. The database contains all full-time and half-time score lines and the bookmaker odds for most of the matches [1]. The site provides data for matches in the English Premier League since season 1993 - 1994. However, from season 1993 - 1994 to season 2000 - 2001, the data only contain half-time results and full-time results, which are inadequate for the usage of Machine Learning. From season 2000 - 2001 until the current season, the site provides data with much more diversity than before. The data contains the number of shots, shots on target, fouls, corner kicks, cards, etc.

B. The Tool

When it comes to the task of using Machine Learning for prediction purpose, there are several tools that are highly recommended. Buursma in [1], Kurma in [3], and Tax and Joustra in [6], etc. all used Waikato Environment for Knowledge Analysis (WEKA). WEKA provides users with several graphical user interfaces that make it easier for the users to use. At first, WEKA seems like the perfect tool for sports prediction projects. However, the main (and probably the only) way to use WEKA is by clicking and dragging instead of purely coding. Thus, the users would have to choose everything and examine the results manually. This makes it hard to include WEKA in the automatic process of data mining and predicting.

After a few discussion with my instructor, Charlie Peck, we both agreed that a brilliant solution for this problem is Scikit-learn [10], which can allow the researchers to actually create a script that automates the whole process. 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 [11]. It also provides modules for extracting features, processing data, and evaluating models [9].

C. The Popular Machine Learning Algorithms for The Problem

In many researches about soccer results prediction, the authors mainly focused on testing the following Machine Learning Algorithms:

- Decisions Tree;
- Bayesian Networks;
- Artificial Neural Networks;
- Linear Regression;
- Logistic Regression;
- Support Vector Machine.

In each of these general algorithms, there can be many specific methods to choose from. For example, we have Naive Bayes [3] (for Bayesian Networks), Classification via Regression [1] (for Linear Regression), LogitBoost [6] (for Logistic Regression), etc.

After I have chosen the features to be used as well as executed the training & testing process by using the Ten-fold Cross-Validation method [7], I will determine which algorithms will be used for the final purpose of the project. Here, the results from previous works can also be considered to make up my final decisions.

III. Proposal of The Basic Design

As I have said, my project goal was to make a program/script that can carry out every step needed in the process of predicting the future results. This starts from writing code that retrieves the data from the source stated above and preprocesses the data in an usable format. During the course of the project and after training and testing have been performed, the selections of features and certain Machine Learning algorithms (provided and tested by using Scikit-learn) will be determined to predict the future results. I would then set those features and algorithms to be used in the program.

In my current idea of the design, the program will first ask the user for the home team and the away team. Then, it will use the decided Machine Learning algorithms to predict the result between the specified teams and print that result out to the screen. However, the user can also choose to predict the 38 matches of one team for the whole season. The program will then write the predictions to an output file which can be accessed by the user.

IV. Timeline

The tasks that I will be completing for the project are listed in the proposed timeline below:

- Jan 10 - Jan 17: Successfully install Scikit-learn and start working.
- Jan 18 - Feb 3: Finish writing code that retrieves the data provided by the source.
- Feb 4 - Feb 14: Finish writing code that preprocesses the data in a usable format.
- Feb 15 - Feb 18: Early Semester Break
- Feb 23: Submit the first draft of the paper.
- Feb 24 - March 9: Get a better idea of how the Machine Learning algorithms work and how to use them in the tool, Scikit-learn.
- March 10 - March 18: Spring break.
- March 23: Submit the second draft of the paper.
- March 24 - April 14: Finish the process of feature selection and determining the best Machine Learning algorithms for the problem after training and testing.
- April 15 - April 22: Finish writing the program (script/code) that can be used by the user to get the prediction for a match between a certain pair of teams or for all 38 matches per season of one team.
- April 23 - 26: Demo the program/script for instructor/mentor.
- April 28: Submit the final draft of the paper.
- April 30 - May 4: Finals week - Presentation.

References

[1] 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.

[2] 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.

[3] Kumar, Gunjan. *Machine Learning for Soccer Analytics.* Cambridge University Press, MSc thesis, KU Leuven, 2013.

[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] 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.
[7] Bunker, Rory P., and Fadi Thabtah. “A Machine Learning Framework for Sport Result Prediction.” Applied Computing and Informatics, 2017.
[8] Football Betting — Football Results — Free Bets — Betting Odds. http://www.football-data.co.uk/. Accessed 9 Nov. 2017.
[9] Hackeling, Gavin. *Mastering Machine Learning With Scikit-Learn*. Packt Publishing, 2014.
[10] *Scikit-learn: Machine Learning in Python*. http://scikit-learn.org/stable/. Accessed 16 Nov. 2017.
[11] Pedregosa, Fabian, et al. “Scikit-Learn: Machine Learning in Python.” Journal of Machine Learning Research, vol. 12, Oct. 2011, p. 28252830.