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 outliers 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.

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.

Data:

Buursma in [2], Constantinoiu et al. in [3], Timmaraju et al. in [4], Ulmer et al. in [5], and Tax and Joutsa in [7] all did their research 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:

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

Minh Vo

Department of Computer Science, Earlham College

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].

PRELIMINARY EXPERIMENT RESULTS

After having gathered the data, I tested the 4 algorithms, Decision Trees, Gaussian Naive 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.


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.

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


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 adviser, which is keeping me on track to complete my Senior Capstone project.