International Journal of Pure and Applied Mathematics
Volume 118 No. 22 2018, 533-536
ISSN: 1314-3395 (on-line version)
url: http://acadpubl.eu/hub
Special Issue

ijpam.eu

ENGLISH FOOTBALL PREDICTION USING MACHINE LEARNING CLASSIFIERS

[1]Anand Ganesan, [2]Harini M
[1]Student, [2]Assistant Professor,
Department of Computer Science and Engineering, SRM IST
[1]anand.ganesan8@gmail.com, [2]harini.m@ktr.srmuniv.ac.in

Abstract: Sports Analysis and Betting have been on the rise lately with the ever increasing ease of Internet accessibility and popularity of Machine Learning. This is an interesting area of research for football, as football is regarded as much more complex and dynamic when compared to a few other sports. It is also the world's most popular sport, played in over 200 countries. Several methodologies and approaches are being taken to develop prediction systems. In this paper, we predict the match outcomes of the English Premier League, by performing a detailed study of past football matches and observing the most important attributes that are likely to decide the conclusion. We use algorithms such as Support Vector Machines, XGBoost and Logistic Regression and then select the best one to give us the target label. This model is applied on real team data and fixture results gathered from http://www.football-data.co.uk/ for the past few seasons.

Keywords: Football, Prediction, Machine Learning, Data Mining.

1. Introduction

Prediction systems have made their way into a variety of fields such as stock markets, sports, online shopping, and so on. In sports, these systems could be used for betting, for coaches to analyze the performance of the squad, enhance their game plan, etc. Sports' betting also has been growing in double digit rates over the past few years. As a result, Machine Learning is currently a highly trending approach. Microsoft's search engine, Bing, accurately predicted each and every knockout stage match outcome of FIFA's 2014 World cup. The model used by them for making predictions however is extremely confidential and it is surprising as to how predictions were made with 100% accuracy.

With love for the sport and inspiration from Bing's model, we attempt to predict the results of football matches in the English Premier League, which is hailed to be the most popular and exhilarating league of football in the world. The League operates on a promotion and relegation basis with twenty teams challenging each other to accomplish their ambitions as a club. Before proceeding to the major section of the paper, we will review a few previous works in this field.

A. Literature Survey

In the paper [1], the authors are describing their approaches of developing a software system that can predict the result of matches in the UEFA Champions League with about 60% accuracy. They have used algorithms such as Naive Bayes, k-nearest neighbors, Random forest and Bayesian networks in order to obtain a suitable combination of features and classifiers required to make predictions. The software system was implemented in the Java programming language with use of Weka API. As mentioned by the authors themselves, there is scope for further improvement, mainly in terms of selection of features. Monitoring and taking into consideration, various other performance related features of teams and larger data sets for learning would prove beneficial to predict future outcomes with better accuracy.

In the paper [2], the authors study multiple techniques in data mining and their prediction results are correlated to devise a good model for predicting matches of the Dutch football team. They use three major models namely Generalized Boosted Models (GBM), K-nearest neighbor and Naive Bayes classification. Using GBM, they attained 60.22% accuracy on average, while the other models were not as accurate. The results of the paper were based on a data-set that only included information about the Dutch team but no data regarding the opponent except for their FIFA ranking. To further improve this research, more data and statistics could be taken into account such as the opponent team's overall form in that season, and other factors such as head-to-head results or information about each team's previous games.

In paper [3], authors have proposed a logistic regression model to estimate 2015/2016 Barclays' Premier League match results with an accuracy of around 69.5%. They develop this model with the help of data from Barclays Premier League and sofifa.com using four significant variables: Home Offense, Home Defense, Away Offense, and Away Defense. They implement this method in software called *Football Predictor.* Their work predicts who is going to win a match (home/away), and list out details regarding the odds and probability,

and the coefficients of regression. This model consists of just four variables but gives strong prediction accuracy.

In paper [4], the authors applied the Bayesian Network Model to predict the results of football fixtures involving the FC Barcelona team in the Spanish La Liga during the 2008-2009 season. They split the data-set into Non-physiological factors such as record of five previous matches, weather conditions, results for or against team; and physiological factors such as the players' mean age, the number of key players injured in the team, mean number of goals scored in all home and away fixtures. NETICA software was used to design the model, which produced values for players' average age as a medium, history of the last few games as win, injured main players, psychological state of the players and weather forecast during the match. They accomplished a prediction accuracy of 92% when applied to predict the 2008-2009 season fixtures. The limitation in this paper, is that only one team was taken into consideration for predicting outcomes.

In the paper [5], the authors present an approach to estimate football match results with Neural Networks. Initially, they have classified a match into three categories using a Learning Vector Quantization neural network to determine the strength contrast between the two opponents. Then they use specific Back Propagation networks on the data they have designed according to the classifying result. They have trained and tested their model on actual football match outcomes from the famous Italian league, Serie A and have achieved a great performing model with good accuracy.

B. Proposed Model

In this paper, we propose a model to predict the outcomes of football matches in the English Premier League. We train the final data-set on various machine learning classifiers. We compare the performances of each classifier and choose the one that returns the best result. Then, we optimize the classifier that yields the best result to further enhance the model accuracy in making predictions. Our final output or target label is the Full time Match Result (FTR). This label will indicate a Home team win (H), an Away team win (A), or a Draw (D).

2. Architecture/Design

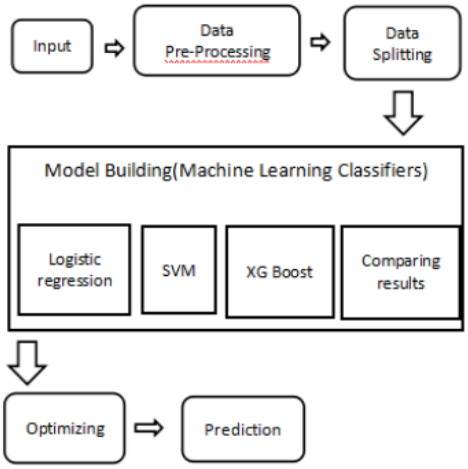

Figure 1. Architecture Diagram

A. Dataset Description

We are going to be predicting match outcomes using data from past games for a few seasons. We obtained this data-set from football.co.uk. The data-set contains around 65 attributes per season regarding the home team, the away team, the venue, scores, to name a few. We filtered these attributes into a final list of 10-15 attributes which proved to be the most influential for predicting the outcome.

B. Pre-Processing

The data-set we obtained from football.co.uk consists of several attributes from each season. A lot of these features are pretty much unnecessary for making outcome predictions.

Hence, our primary task is to clean the data-set and to only retain the features or attributes that we require the most. We calculate the Scatter Matrix to observe how much one attribute affects another attribute in the data-set and their correlations. This will help us pick the most influential features that we want to use to build our new data-set.

C. Data Splitting:

Once we finish building our new set of crucial attributes, we split the data into training and testing data.

D. Model Building

In this module, we apply the machine learning classifiers required for making our prediction.

Logistic Regression:

Logistic regression seeks to design the possibility of an event occurring based on values of independent variables, that could be categorical or numerical. It is a statistical method that operates on data-sets having one or more independent variables which decide an outcome. The objective of this method is to compute the perfect fitting model that describes the interrelationship amidst the dependent variable and a list of independent (predictory) variables. [5]
Our problem is a multi-class classification problem as there exists more than two possible outcomes i.e, Home Win, Draw and Away Win. Hence we are going to be using Multinomial Logistic Regression.

Support Vector Machines:

Support Vector Machines are models in Machine Learning that is useful for regression analysis and classification tasks. We map each data item as a point in a space of n-dimensions (n being the number of features) in which each feature-value corresponds to a particular co-ordinate. The target is to obtain a hyper-plane that classifies all training vectors into two classes. The finest choice is the hyper-plane that leaves the maximum margin from both the classes. [6].

XGBoost:

We use XGBoost algorithm to develop a predictive model that is based on an ensemble of decision trees. XGBoost is an algorithm that is on the rise recently dominating applied machine learning for tabular or structured data. This algorithm is an application of gradient-boosted decision trees designed for performance and speed. [7]

3. Comparison of Results

We contrast the results of the Machine Learning algorithms and pick the best performing one to give our result.

4. Optimizing

Once we select the best performing classifier, we optimize its hyper-parameters, to further enhance the performance and accuracy of the model in making our prediction. Finally, we obtain our target variable FTR for predicting the outcome of the match.

5. Conclusion and Future Work

The model we devised is based on statistical analysis of past football games. We will be able to make fairly accurate predictions. Although the accuracy of this model is pretty good, it's not guaranteed to be always right and there is a lot of scope for future work in this regard.

We could bring in sentiment analysis, features such as individual player and team performance metrics, studying the trending hash-tags on twitter on match day, the posts from fans on social media, etc to further enhance the accuracy of the model.

References

[1] Hucaljuk, J & Rakipović, A. (2011). Predicting football scores using machine learning techniques. 2011 Proceedings of the 34th International Convention MIPRO,, 1623-1627.

[2] Abel Hijmans. Dutch football prediction using machine learning classifiers (unpublished)

[3] Darwin, P & Dra, H (2016). Predicting Football Match Results with Logistic Regression. 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA).

[4] Farzin, O., Parinaz, E., &Faezeh, S. M. Football result prediction with Bayesian network in Spanish league-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 2013, 812-815

[5] T. Cheng, D. Cui, Z. Fan, J. Zhou, and S. Lu, “A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system,” Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003.

[6] https://www.medcalc.org/manual/logistic_regression.php

[7] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[8] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[9] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on 26-27 Feb. 2015,Publisher: IEEE,DOI: 10.1109/ECS.2015.7124833.

[10] T. Padmapriya and V. Saminadan, “Inter-cell Load Balancing Technique for Multi- class Traffic in MIMO - LTE - A Networks”, International Conference on Advanced Computer Science and Information Technology , Singapore, vol.3, no.8, July 2015