Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 164 (2019) 131–136

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019

Predicting Sports Results with Artificial Intelligence – A Proposal Framework for Soccer Games

Gabriel Fialho[a,b], Aline Manhães[b], João Paulo Teixeira[a,c,*]

[a]*Instituto Politécnico de Bragança, Bragança 5300, Portugal*
[b]*Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Av. Maracanã, 229, Rio de Janeiro 20271-110, Brasil*
[c]*Research Centre in Digitalization and Intelligent Robotics (CEDRI), Applied Management Research Unit (UNIAG), Bragança 5300, Portugal*

Abstract

As the sports betting industry and technology have grown on a large scale, predicting the outcome of a sports match using technologies approach is now crucial. In fact, humans have a certain limitation when processing a large set of information. However, Artificial Intelligence techniques can overcome this issue. Furthermore, sports have a great amount of data to consider, thus, it is a great example of AI problem. A review of some research using different Artificial Intelligence techniques to predict a sport outcome is presented in this article. Different types of sports such as football, soccer, javelin throw, basketball, and horse race were analyzed, and showed distinct approaches to predict results. Finally, a framework to develop a system to predict the outcome of a soccer game based on AI is proposed, considering the present research review.

Keywords: Soccer Outcome Predition; Sport Outcome Prediction; AI in Sports Prediction; ANN in Sports; Betting with AI

* Corresponding author. Tel.: +351 273 30 3129; fax: +351 273 30 3051.
E-mail address: joaopt@ipb.pt

10.1016/j.procs.2019.12.164

1. Introduction

Sports have been part of human lives for millenniums, and the interest to be more than just a simple audience sends us back to more than 2000 years ago, when the Greeks went to the Colosseum to bet on a gladiator of their choosing [1]. Since that time, the interest of trying to know the future in sports matches has increased. For example, commentators on television shows guessing which team will win the Super Bowl or trying to figure out the result of a Premier League match, people on social medias discussing who will win the World Tennis Championship. It is part of the human lives. In addition, sports gambling industry has grown to such an extent that by 2017, the total gross income of the gambling industry totaled $17.8 billion [2]. The interest for knowing the result of sports matches before its conclusion is now clear, and indispensable to the sports gambling industry. For instance, the technology has expanded exponentially and has elevated us to another level of comprehension, and Artificial Intelligence is changing the way sports prediction is seen.

Artificial intelligent algorithms span several branches of computer science. Some examples are pattern recognition, predictive systems, inference and data analytics. The past few years were very important for machine learning technologies, they had an aggressive expansion of its accuracy, and now artificial neural networks can outperform humans in many areas [3]. One area that computers overcame humans is predictions, for instance many people are using this instrument to predict sports outcomes. As several sports have extremely large number of characteristics that are directly related to the result, is difficult for humans to consider all the features and predict with high accuracy a sports match. In these situations, a high-performance technique is needed to deal with all the data, and that is where Artificial Intelligence goes in. The enormous advances in this technology make it possible to process an exorbitant amount of data to draw extremely useful conclusions.

This article aims to study the use of Artificial Intelligence techniques to predict the results of sports matches. Every sport has particular rules, number of players, different styles, that is, a set of different features. For someone that is starting a predictive model in this area is challenging to obtain a considerable dataset, preprocess this data and create a design from scratch. At the end of this article, a discussion and propose to develop a predictive model for soccer will follow. As this subject is becoming popular, available documents, such as articles, academic papers, and research papers will be used as reference.

This article is divided into literature review of Artificial Intelligence techniques papers, discussion and conclusion. First research papers with Bayesian and Logistic Regression to predict a sport outcome will be reviewed, second, Artificial Neural Networks and then Support Vector Machine and Fuzzy Logic and Fuzzy Systems techniques. Furthermore, the discussion topic, where the papers will be analyzed and a framework to predict soccer outcome will be proposed. Finally, the conclusion closes the article.

2. Literature Review

2.1. Bayesian and Logistic Regression

Created in 18th century by Thomas Bayes, is a probabilistic prediction model that assumes all features to be conditionally independent from the target variable [4]. Farzin Owramipur et al. [5] considered season 2008-2009 of Spanish football league to predict Barcelona's results. They gathered 6 psychological features and 7 non psychological features and used a Bayesian Network for every match. The results of these operations where gathered and the final point of Barcelona team was determined. This result was compared with 2008-2009 season and the final result was correct in 92%.

Constantinou created a model that combines a rating system with a Hybrid Bayesian Network (BN) [6]. The rating system generates a rating score that captures the ability of a team relative to the residual teams within a league. The resulting ratings are used as input to the BN model for match prediction. The data consisted of a training dataset with 216.743 match instances from several football leagues throughout the world, and 206 match instances as a test dataset. His approach is that team ratings is based on recent historical match results, but match predictions are derived from historical observations, which include different teams. As a matter of fact, a match prediction between two teams was often based on historical results from other teams, throughout the world. With this model, the predictive accuracy was determined by the Ranked Probability Score (RPS) Function (Epstein, 1969), which

represents the difference between cumulative predicted and observed distributions. The result of this function represents how close the distribution is to the observed value. The score lies between 0 and 1, with lower values being better. Constantinou ranked 2nd in the international special issue competition *Machine Learning for Soccer* with 0.208256 RPS and 99.06% relative performance [6].

2.2. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) usually contains interconnected components (neurons) that transform a set of inputs into a desired output, trying to mimic a biological neural network. In 1996, Purucker [7] was one of the firsts to study predicting results in sport matches using ANNs. He collected data from the first eight rounds of the National Football League (NFL), and five features: yards gained, rushing yards gained, turnover margin, time of possession and betting line odds. A Multi-Layer-Perceptron (MLP) ANN trained with backward propagation algorithm was used, and he achieved 61% accuracy compared with 72% accuracy of the domain experts. Later, in 2003, Kahn [8] continued Purucker's work. He added new features, such as total yardage differential, rushing yardage differential, turnover differential, away team indicator and home team indicator. The problem was treated as a classification, with two classes: away team outcome and home team outcome (-1 for lost and +1 for win). Data on 208 matches in the 2003 season were used. He achieved 75% of accuracy. The results were compared to the predictions of eight sportscasters from ESPN. The domain experts predicted an average of 63% of matches correctly.

Igiri et al. [9] extracted data from 110 matches played in the 2014 and 2015 English Premier League season and used as input to a neural network system. The features used are: Home and Away goals (GHA), Home and Away shots (HAS), Home and Away corners (HAC), Home and Away Odds (HAOD), Home and Away attack strength (HAAT), Home and Away Players' performance index (HAPPI), Home and Away Managers' performance index (HAMPI), Home and Away streak (HASTK), Home and Away managers' win (HAMW). They used 20 matches played in 10th and 11th week of 2014/15 English Premier League to predict the outcomes. The result was 85% accuracy using Logistic Regression, optimizing features by weighting.

2.3. Support Vector Machine (SVM)

This method is based by the theory of statistical learning, developed by Vapnik [10]. SVM is a discriminative classifier defined by a separating hyperplane. Given labeled training data, the algorithm outputs an optimal hyperplane, which categorizes new examples [11]. In [12], Cao did an experiment with 4 different techniques to predict the outcome of NBA games. He used data of 5 regular NBA seasons for model training and 1 NBA season for testing. As results, he concluded that the accuracy to predict a NBA match using Bayes was 65.82%, 66.67% using Neural Networks, 67.22% using Support Vector Machine and 67.82% using Simple Logistics.

Tsakonas et al. in [13] demonstrated an example of how to predict football game winners by applying Support Vector Machines model. Data during 10 years in Ukrainian football championship was used to create and test the model. For this research, they used a common regression problem as a classification method: if forecasted value >=0 guest team will not win, if forecasted value<0 host team will not win. The algorithm was given 105 training data as input and 70 test data records. Before training and testing, all data were normalized in [-1, 1] range. After 1377 iterations, they achieved an accuracy of 61.4% correct prediction on test set. They concluded that further research using hybrid computational intelligent schemes [14] would offer an even high classification and prediction rate.

2.4. Fuzzy Logic and Fuzzy Systems

Fuzzy Logic is an extension of Boolean logic by Lotfi Zadeh [15]. It is based on the mathematical theory of fuzzy sets [16], and enables a condition to be in a state other than true or false, by introducing the notion of degree in the verification of a condition. Fuzzy logic provides a very valuable flexibility for reasoning, which makes it possible to consider inaccuracies and uncertainties [17].

Rotshtein et al. [18] proposed a model that makes possible to predict the result of a football match using the previous matches of both teams. The model is based on the method of identifying a past-future nonlinear

dependence by a fuzzy knowledge base. They used the tournament data for the championship of Finland for the research. 1056 matches from 1994 to 2001 as learning samples and 350 matches from 1991 to 1993 as testing samples. The outputs were a high score loss (d1), a low score loss (d2), a drawn game (d3), a low score win (d4) and a high score win (d5). The model with neural tuning had 91% accuracy predicting d1, 83% d2, 87% d3, 84% d4 and 94% d5. They conclude that the model can be used for creating commercial programs of predicting the results of football matches for bookmaker offices.

3. Discussion

The previous chapters illustrate different works on the same area of predicting sports outcome. Some of them are similar and some of them are very different. With only a set of 5 features, Purucker [7] tried to predict football matches and got an accuracy of 61%. Later, Kahn [8] added more features and achieved 75% accuracy. Using Bayesian, Farzin Owramipur [5] et al. got 92% of accuracy with 13 features, splitted into psychological and non-psychological. Igiri [9] added one more layer to the features. Splitting into home and away characteristics, he achieved an accuracy of 85% predicting soccer outcomes. Furthermore, some sports are more predictable than others. For example, sports that does not have a draw are usually more predictable than sports that have draw as a result. Using the literature review, a model to predict soccer matches is proposed during the next sections.

3.1. Dataset

Soccer has a relatively low number of data-driven studies, one reason for that might be the lack of publicly available databases. One of the most famous open databases for soccer analytics is the Open International Soccer Database for machine learning [19], which has essential information teams, from 216.743 league soccer matches. That is, the dataset is composed by: season, country and league, date on which the game was played, name of the home and away team, number of goals scored by home and away team, goal difference and outcome of the game in terms home win (W), draw (D) and away win (L). To demonstrate the use of this database, the creators [19] organized the 2017 Soccer Prediction Challenge, which the teams had to develop a predictive model from the database and then predict the outcome of 206 future soccer matches. Even though the database was composed of more than 200.000 matches, the winning team had an accuracy of 51.94%, which is relatively low comparing to the other studies above. Furthermore, the authors concluded that adding more data relevant to the outcome of a match (e.g., data about other match events, players, teams, etc.) might improve the predictability [19]. Later, in 2019, Runzuo Yang used data from players in the starting line-up besides that essential information to predict English Premier League outcomes [20]. In fact, he achieved a great accuracy of 81.8% based on the first 12 weeks. However, he extracted these features manually from websites, thus the input data couldn't be relatively large, and the model was only for the 2018/2019 season and couldn't be used to predict other seasons, according to the author.

A large amount of input data is good for a predictive model. In fact, it is important to have long historical records, as this is a key factor for reliable predictions. However, most datasets available doesn't have a considerable set of features from past years and the data may not be readily available for a wide range of countries and leagues. It is needed to balance between a relatively large input data and good amount of input features. Surely more detailed information about soccer matches are freely available on the internet such as in "https://www.whoscored.com" or "https://www.transfermarkt.co.uk", however it is rarely in a form directly usable by machine learning methods. Most websites present the results and information about soccer as HTML encoded. Furthermore, if someone want to work with as many features as possible and a considerable number of matches, it becomes impossible to extract manually the data such as in [20], so it is need to rely on techniques from computer science [21] to withdraw all that available data into a better format to manipulate such as CSV or XLS files.

The next step is to filter all the features into that ones that affect the result of a soccer match. Runzuo [20] proved that players attributes affect directly the predictability of a model and Igiri [9] demonstrated that features such as goals, shots, corners, attack strength, manager performance streak and managers win increase the accuracy. These are overall features, so for example, the corners average is the same if a team will play home or away. However, it is known the advantage of home field in soccer. In addition, Constantinou have proved that the pi-rating system is very useful to predict soccer outcomes in [22]. The pi rating system considers the team's performance in recent matches,

the well-known home advantage effect and, the fact that a win is more important than increasing the score difference. For instance, adding this feature to the dataset will be advantageous for the model.

A team can play one way in the home field, but it can be completely different playing in other places. Furthermore, it is proposed to split these features into home and away averages. For example, a team will have the corners average playing home and playing away, so if the team are going to play home, the home corners average will be on the place of corners overall average. A list of important features that should be added to the model is players attributes, goals, shots, corners, attack strength, managers' performance and pi-rating.

3.2. Model

In this section a model to predict soccer matches using the related work is proposed as a guide. From previous referred works, Igiri [9] has proved that with neural network it is possible to achieve a high accuracy predicting soccer outcomes. On the other hand, a common and crucial process to get a high-performance model is to test different types of machine learning algorithm to find the one that best fulfills the desired objective. Thus, a good approach is to use the models demonstrated in the literature review and compare the results.

The desired variable to predict the outcomes is the full-time result in the form of a vector composed by three binary values (the first one for the home team win, the second for the draw and the final one for the away team win). For example, for the home win result the vector will be [1, 0, 0]; for the draw result: [0, 1, 0]; for the away team win: [0, 0, 1]. This way, the vector will be equally distributed for all the result cases shown above. On the other hand, the response could be only one value and not a vector, ranging between two numbers, for example -1 that represents away win, 1 to represent home win and 0 for draw. This way is necessary to use mathematical equations to extract the result of a home win, draw and away win percentages becoming a more complex process, thus the first approach is preferred. All that data can be used as an input to a Neural Network to predict the outcomes. The response of the Neural Network will be in the same form as the vector, however, even though it was composed by binary values in the input dataset, the network can give us numerical outputs, and the percentages can be extracted from these values. For example, if the response is [0.49, 0.31, 0.20], that is, the home team would have 49% chances to win the match, 31% for draw, and the away team would have 20% to win. Furthermore, for this match, the home team is likely to win, and it can be used as support to a betting choice.

4. Conclusion and Future Works

Artificial Intelligence is becoming more and more popular as the technology evolves. With a properly data set and a specific technique for a sport's choice it is possible to achieve great accuracy predicting the outcome of a sport match, even better than the domain experts. Based on the analysis of the related works, it was proposed a model and feature selection for predicting soccer outcomes. Furthermore, these systems can also be useful to make profit in betting industries, using science on our side.

As a future work, the knowledge reported will serve as the base to create a predictive model for soccer matches. For this assignment, computer science methods will be used as described in discussion to get a relatively large amount of data input and features of several leagues and seasons throughout the world and Artificial Intelligence techniques such ANN to try to obtain a high accuracy model.

Acknowledgements

This work is supported by the Fundação para a Ciência e Tecnologia (FCT) under the project number UID/GES/4752/2019.

References

[1] Doeden, Matt (2010), "Legalized Gambling: Revenue Boom or Social Bust?" *Twenty-First Century Books*.

[2] Porreca, Rocco (2018), "Sports Gambling in Select Nations", *Aspen Institute*. Retrieved from: https://assets.aspeninstitute.org/content/uploads/2018/09/Foreign-sports-betting-research.pdf

[3] Steinberg, Roman (2017), "6 areas where artificial neural networks outperform humans" *UKIT AI*. Retrieved from: https://venturebeat.com/2017/12/08/6-areas-where-artificial-neural-networks-outperform-humans/

[4] Langaroudi, Milad and Yamaghani, Mohammad. (2019), "Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches: A Survey", *JACET J. ADV COMP ENG TECHNOL, Winter 2019,* 5(1).

[5] Owramipur, Farzin and Eskandarian, Parinaz and Mozneb, Faezeh (2013), "Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team", *International Journal of Computer Theory and Engineering*, 5 (5).

[6] Constantinou, Anthony (2018), "Dolores: a model that predicts football match outcomes from all over the world". *Springer US* 108: 49-75. Retrieved from: https://doi.org/10.1007/s10994-018-5703-7.

[7] Purucker, Michael (1996), "Neural network quarterbacking", *IEEE Potentials* 15 (3): 9–15.

[8] Kahn, Joshua (2003), "Neural network prediction of NFL football games", *World Wide Web Electronic Publication 2003*: 9–15.

[9] Igiri, Chinwe Peace and Nwachukwu, Enoch Okechukwu (2014), "An Improved Prediction System for Football a Match Result", *IOSR journal of Engineering* 4 (12)*:* 12-20.

[10] Vapnik, Vladimir (1995), "The nature of Statistical learning theory". *Springer-Verlag New York*.

[11] "Intoduction to Suport Vector Machines". Retrieved April 5, 2019, from: https://docs.opencv.org/2.4.13.7/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[12] Cao, Chenjie (2012), "Sports Data Mining Technology Used in Basketball Outcome Prediction", *Masters Dissertation of Dublin Institute of Technology*.

[13] Tsakonas, A. and Dounias, G. and Shtovba, S. (2003), "FORECASTING FOOTBALL MATCH OUTCOMES WITH SUPPORT VECTOR MACHINES", *Herald of Zhytomyr Engineering-Technological Institute*. 1: 181–186.

[14] Tsakonas A. and Dounias, G. (2003), "Hybrid Computational Intelligence Schemes in Complex Domains: An Extended Review". *Springer: Lecture Notes in Computer Science* 2308 (2002): 494-512.

[15] Zadeh, Lotfi (1965), "Fuzzy Sets". *Information and Control* 8 (3): 338-353 (1965).

[16] Zimmermann, Hans-Jurgen (2010), "Fuzzy set theory", *2010 John Wiley & Sons, Inc*.

[17] Dernoncourt, Franck (2013), "Introduction to fuzzy logic", *Massachusetts Institute of Technology*.

[18] Rotshtein, Alexander and Posner, Morton and A. B. Rakityanskaya (2005), "FOOTBALL PREDICTIONS BASED ON A FUZZY MODEL WITH GENETIC AND NEURAL TUNING". *Cybernetics and Systems Analysis*, 41 (4)

[19] Dubitzky, Wenet and Lopes, Philippe and Davis, Jesse and Berrar, Daniel (2018), "The Open International Soccer Database for machine learning". *Machine Learning January 2019,* 108 (1): 9-28

[20] Yang, Runzuo (2019), "Using Supervised Learning to Predict English Premier League Match Results From Starting Line-up Player Data". *School of Computing, Technological University Dublin Dissertations.* Retrieved 10 April 2019 from: https://arrow.dit.ie/cgi/viewcontent.cgi?article=1192&context=scschcomdis

[21] Saurkar, Anand and Pathare, Kedar and Gode, Shweta (2018), "An Overview On Web Scraping Techniques And Tools". *International Journal on Future Revolution in Computer Science & Communication Engineering*. 4 (4): 363-367

[22] Constantinou, Anthony and Fenton, Norman, (2013), "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries". *Journal of Quantitative Analysis in Sports*. 9(1): 37–50.