Lumiere Research Paper – Neural Networks in Sports Predictions

Kabir Chawla

The Cathedral and John Connon School

Abstract

The following research paper is a literature review that I have pursued based on my interest in machine learning and the football. The paper explores the different types and variations of neural network systems used in the prediction of football matches and analyses the data available. Considering different factors such as input encoding, data types, training time and overall accuracy of the system, the paper tries to answer the question of 'What neural network design factors influence the successful prediction of football match outcomes?' The paper goes through and analyses specifically created data tables using a number of different papers in order to achieve this goal. The form of this paper is a literature review and provides an insight to what the future of predicting matches may be using machine learning and artificial neural networks.

The Research Question

What neural network design factors influence the successful prediction of football match outcomes?

Introduction

General introduction

Recently, with the increasing popularity of machine learning, an interesting opportunity has been created where machine learning could be used to analyse the large sets of data of sports. With new neural network architectures being used, the successful rate of predictions has been increasing. These networks have successfully predicted various outcomes ranging from predicting the score at the interval of a match, to the team's performance over the entire season. Machine learning provides a novel approach to make sense of the large sets of data that are available for a number of sports.

This paper will primarily be focusing on the usage of machine learning and artificial neural networks to predict matches in the sport of football. The paper will analyse the different types of neural networks used to predict the outcome of football matches. I will compare the accuracy of different design factors of neural networks among a number of papers and review the components that could lead to a greater rate of successful predictions. The paper will be a literature review and I hope to provide an interesting insight on the use of machine learning in football and sports to predict successful outcomes.

Motivation behind the paper

Predicting the outcome of a football match can have numerous benefits and purposes to a lot of people. For instance, people may use these stats as a reliable way of placing bets. Secondly, it also aids managers and clubs in understanding their likelihood of winning, analysing their opponents and figuring out how to prove the stats and predictions wrong. This eventually helps the team win tournaments and leagues due to the helpful aid of predicting matches.

Presently, very few methods exist which focus on understanding and creating the most efficient method of predicting football matches. Innovative methods of combining different types of data input and data encoding, along with newer neural network technology may provide a pathway to a highly successful prediction rate.

My interest in this topic

There are a number of different factors which made me interested in exploring this area of research. Firstly, I have been an avid football fan ever since I can remember and also have a love for math and numbers. While seeing matches, the stats or predictions that would pop up would always grab my attention, sometimes even more than the match itself. I have always been intrigued by the question of how these predictions were made and how accurate they really were. Was it a football pundit that was known for his accuracy in predicting things? Or was it someone using another Paul the Octopus? (Paul the Octopus was a very famous octopus that successfully predicted every one of Germany's football matches at the 2008 FIFA World Cup. By choosing between two tanks of identical food, one marked Germany and one the opponent, Paul was able to successfully predict their entire tournament!)

After understanding that it was machine learning and artificial neural networks that were being used for these interesting stats and prediction of matches, I started to read and learn about how exactly they worked. Although I have a lot to learn in the fields of machine learning and artificial intelligence, I hope to convey some interesting things about it in this paper.

Gap in the field

While there exist a vast number of papers which make use of different artificial neural network types, different data inputs, data encoding to predict various sport outcomes, the number of papers which compare these models and possibly suggest the best factors are very limited. Exploring the differences between models and their accuracy has not often been considered in the subfield of machine learning for sports, and I believe this important gap needs to be bridged.

In this paper, I will be presenting, summarising and comparing a number of research papers which have made use of different data sets and different neural network systems to predict the outcome of football matches and some papers which have considered usually abnormal factors into their predictions, such as the weather, home fans after COVID-19, and a few more interesting ones.

In this paper, the following will be presented:

a) Explaining what exactly a neural network is.
b) Papers and methods of comparisons.

c) Summary and analysis of neural networks and design factors.
d) My proposal for the ideal set of design factors.

What is an Artificial Neural Network (ANN)?

As the term suggests, a neural network is inspired by the brain. A neural network consists of multiple layers made out of neurons. Each successive layer has fewer neurons than the previous one. Essentially this is a neural network. In this case, imagine a neuron to be a thing which can hold a number only between 0 and 1. This number which the neuron contains is called its activation.

The layers between the first and last layer of a neural network are often called, "hidden layers." The way this basic network would work is that the activations in one layer would determine the activations in the following layer. While there might be thousands of different combinations in certain layers, which would influence the activations in the following layers, "the learning" process of the neural network essentially refers to this. Over time, as the network goes through huge data tables, it is able to learn and predict the same results much faster. The speed at which neural networks are able to learn is often one of the key factors or benefits when choosing to use a particular network.

Hidden layers exist between the input and output of the algorithm of a neural network in which a certain function applies a weight to the input and works towards an output. Each hidden layer corresponds to the next and the output of one hidden layers acts as the input of the next. This process continues until the input reaches the output layer and the neural network may provide the task it is set out to accomplish.

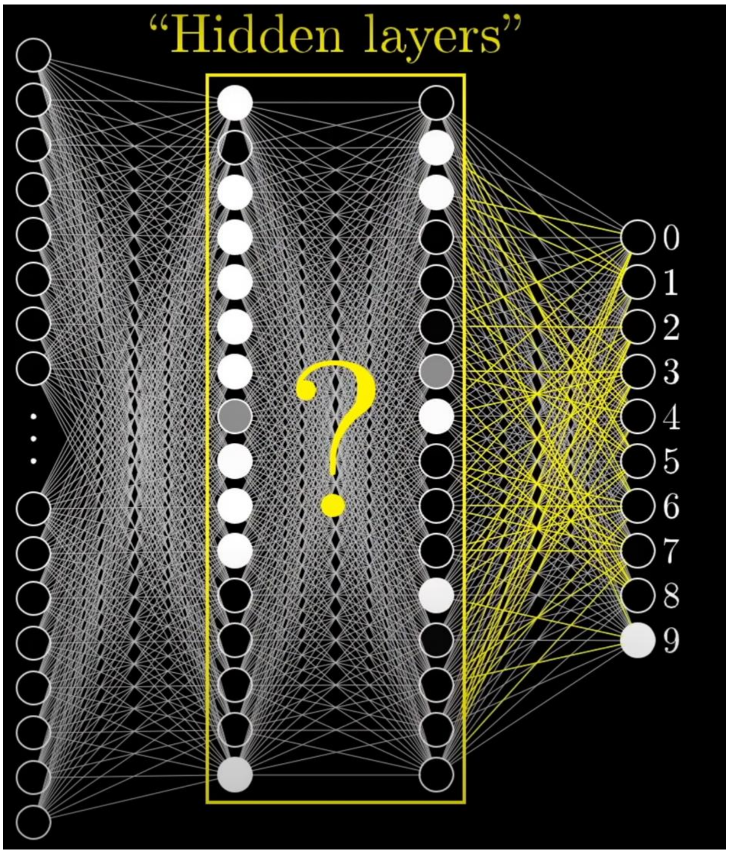

Figure 1 A simple diagram showing the basic structure of a neural network with "hidden layers." [8]

Methodology

This paper considers a total of 6 papers from 2010-2019 which are related to predicting football matches using different types of algorithms and also bookmaker odds. The papers were compared based on their systems and models used for predictions, amount of input data, time to process and the overall accuracy of the algorithm.

The papers used for analysis in this literature review make use of the application of Machine Learning, particularly in the team sport of football. The review is not completely restricted to just Machine Learning networks but also includes papers which have made use of numerous algorithms such as CART, C4.5, and WEKA's J48 algorithms. Some common and general themes that this paper will explore, but are not limited to are:

- Understanding the application of Machine Learning in sports prediction
- Optimizing training times for ANN's
- Improving input encoding
- Suggesting the best data set to use for greater accuracy
- Using these ANN's in different team sports

As the papers all have different types of data and other factors unique to them, establishing parameters which may be used to compare them is important. This paper will compare the different papers through the type of neural network, training time, data size and type and overall successful rate of prediction.

Summary and Analysis of Existing Papers

Although there are 6 papers used for the data tables, only 4 have been extensively summarised and analysed as these provided the required answers for this paper and were written in more concise manners.

1) Odachowski and Grekow (2012)

This paper investigates the possibility of how upcoming methods of predicting football matches may change or end due to a change in bookmaker odds. This paper makes use of the classic 1-X-2 Type bets wherein a prediction is either made for the home team to win, '1' a draw between the teams 'X' or for the away team to win, '2'.

The input data used for this paper was from a website in the form of an XML file. While importing an XML file, the process consisted of tracking its contents for 10 hours preceding the football game and then recording the data in the file on the changing odds. The data for this paper was collected from 2615 matches over a period of 6 months.

The system worked by constantly updating the data based on the fluctuating bookmaker odds. This constant change based on live factors such as injuries of players, condition of the pitch and weather predictions during the match led to an effectiveness of nearly 70%. These results make it possible to suggest the possibility of implementing the use of constantly changing bookmaker odds in a system for more accurate results. A decision-making system may further be developed on this basis and by using specific classifiers and algorithms, a unique and successful system may be created.

The results of this paper are as follows:

Type of classifiers	Algorithm	Accuracy
Standard data set	DecisionTable	46.51%
Win for home team	Bagging	70.56%
Win for away team	NaiveBayes	65.46%
Draw	EnsembleSelction	56.99%

2) Prasetio (2016)

This paper uses the form of logistic regression to predict football matches. In the method of logistic regression, the dependent variable only has two possible values. The model is generally used to predict the probability of the binary response based on one or more variables.

This paper uses official data from the Barclays' Premier League and takes into consideration 4 variables; home offence, home defence, away offence and away defence. Essentially, these are key factors regarding a football match and are usually the most criticised by experts and fans. The paper makes use of a training process flow and a testing process flow in the following manner:

Training process flow:
Match Records → Pre-processing → Estimate Regression Coefficients → Save Regression Coefficients

Testing process flow:
Match Data → Pre-processing → Load Model → Predict Result → Show Result

An equation was formed for the logistic regression which required certain coefficients. To produce the coefficients used in the logistic regression, the paper made use of four experiments, which used different training data to estimate the coefficients.

The accuracy of the testing results with the coefficients obtained from the experiments is as follows:

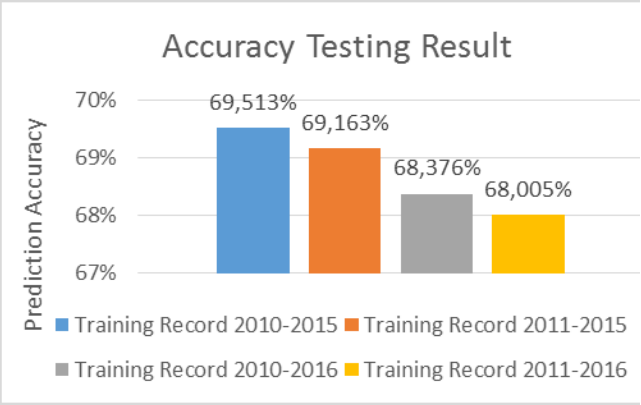

Figure 2 Accuracy comparison between input data for logistic regression [2]

The paper was successful in understanding the fact that the results from the 2015/16 season were unusual which led to a decrease in the predictive accuracy of matches. Furthermore, the main variables out of the four initially considered were home defence and away defence. However, when repeated using just these two variables the results did not show an increase in accuracy, thus establishing the fact that more variables balance out the prediction in a logistic regression system.

In conclusion, although the method of logistic regression is relatively easy to implement and understand, it is often easily influenced by factors such as anomalies or in the case of football matches, random results and should not be considered as the prime approach to predicting matches.

3) Tax and Joustra (2015)

This paper considers data collected from 13 years of the Dutch Premier League, from 2000-2013. The paper's goal was to determine the best combination of dimensionality reduction and machine learning techniques best suited for the Dutch football competition.

This paper considers a large number of factors and their influence on the final result after referring to existing papers which have made use of these factors as the primary factor. These factors include: Previous performances in current season, Performance in earlier encounters, Streaks, Managerial change, Home advantage, matches with special importance, Fatigue, National team players, Promotion to higher league, Expert predictions, Football skills, Strategy, Travel distance, Betting odds, Club budgets and Availability of key players. The data type for these factors include scalar, nominal and Boolean.

As one of the paper's objectives is to identify a key method of dimensionality reduction, the following approaches are considered for this, Principle Component Analysis, Sequential Forward Selection, ReliefF. After considering a number of machine learning algorithms to use, the paper decided to use the following for the experiment: Naïve Bayes, LogitBoost, Neural Network, Random Forest, CHIRP, FURIA, DTNB Decision tree (J48), Hyper Pipes.

Using public data, the paper was able to establish the fact that this highest accuracy data model was seen when using the NaiveBayes system or the Multiplier Perceptron Classifier in combination with a Principle Components analysis. The overall accuracy, however, was 54.702%, which is relatively less compared with the other networks and thus may not be the most efficient way to predict a football match.

4) Danisik et al. (2018)

This paper makes use of neural networks which depend on factors such as player attributes. After analysing the key concepts of football, this paper decided to make use of two particular categories for their dataset, Player Stats and Match History. By creating a unique way to measure player attributes such as physicality, skill and psychological state on a scale of 0-100, with 100 being the strongest, this neural network system had a new style of data input. For the Match History, the data included was match information from the best five leagues in Europe from the 2010/11 season to the 2015/16 season, 5 years of matches.
The neural network structure used for this paper was based on the design of LSTM neural network, a Recurrent Neural Network developed for the vanishing gradient problem. The difference between a RNN and LSTM according to the authors is as follows:

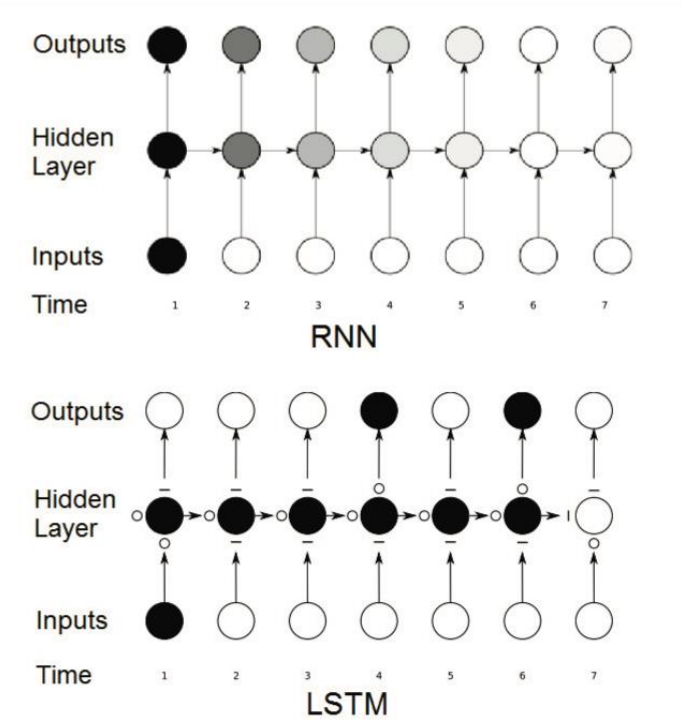

Figure 3 RNN v/s LSTM [4]

In conclusion, this paper provided a unique approach to predicting matches by introducing a new type of data input inspired by video game statistics. The results of this paper provide support that the LSTM network has a promising future for predicting sport matches.

Table of Data

Paper	**ANN's**	**Input Data Type (Number of matches)**	**Accuracy**
Odachowski and Grekow (2012)	BayesNet, SVM, LWL, Ensemble Selection, CART, Decision Table	1,116	70.3%
Prasetio (2016)	Logistic Regression	2,280	69.5%
Tax and Joustra (2015)	WEKA: NaiveBayes, LogitBoost, ANN, RandomForest, CHIRP, FURIA, DTNB, J48, HyperPipes	4,284	56.1%

Danisik et al. (2018)	LSTM NN classification, LSTM NN regression, Dense Model	1,520	52.5%
Huang and Chang (2010)	ANN trained with BP	64	62.5%
Constantinou (2019)	Hybrid Bayesian Network	216,743	51.5%

These are the papers that I will be summarising and analysing in this paper and the table above gives a brief insight about the papers, with regards to the type of Artificial Neural Network Used, the number of matches it was used to predict or test and the overall accuracy of the entire network.

Analysing the ANN’s

Name	**Structure (Primary)**	**Input encoding**	**Training Time (Approximation)**
Odachowski and Grekow (2012)	BayesNet	6 months of matches	Unkown
Prasetio (2016)	Logistic Regression	5 years of matches	n(O(d)) = O(nd) Very low training time
Tax and Joustra (2015)	WEKA: NaiveBayes	Unknown	8 months of data
Danisik et al. (2018)	LSTM NN classification	5 years of top 5 league matches	TBA
Huang and Chang (2010)	ANN trained with BP	-/	-/
Constantinou (2019)	Hybrid Bayesian Network	-/	-/

The table above is a brief representation of how some differences between the papers has been made and the general parameters and design factors that I had considered.

Conclusion

In conclusion, this paper analysed the systems used in a variety of papers and picked out key statistics, data and unique features. The future of predicting sport matches has a promising scope as the number and type of systems being developed are continuously improving. As analysed in this paper, the most successful system was based on bookmaker odds (Paper 1) but I believe that the greatest potential lies in Logistic Regression.

Paper 2 provided a solid foundation of the benefits and usefulness of Logistic Regression in successfully predicting matches. According to the factors of the other paper, I believe that considering key factors such as: Previous performances in current season, Performance in earlier encounters, Streaks, Managerial change, Home advantage, matches with special importance, Fatigue, National team players, Promotion to higher league, Expert predictions, Football skills, Strategy, Travel distance, Betting odds, Club budgets and Availability of key players as well as implementing the unique idea of developing scores from 0-100 for each player for different attributes would provide the Neural Network with adequate and important data. Although the training time of the system might be increased, the fact that it is based on Logistic Regression would ensure it is kept relatively low.

The future of predicting sports matches is promising and with the constantly evolving study of Machine Learning, there is no doubt that newer technologies may soon surpass neural networks in predicting matches. As for now, Logistic Regression with unique data input and data encoding seems to be the most ideal system.

Citations

1. Odachowski, K., & Grekow, J. (2012). Using bookmaker odds to predict the final result of football matches. In International Conference on Knowledge- Based and Intelligent Information and Engineering Systems (pp. 196–205). Springer.
2. Prasetio, D. (2016). Predicting football match results with logistic regression. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) (pp. 1–5). IEEE.
3. Tax, N., & Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. Transactions on Knowledge and Data Engineering, 10, 1–13.
4. Danisik, N., Lacko, P., & Farkas, M. (2018). Football match prediction using players attributes. In 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA) (pp. 201–206). IEEE.
5. Constantinou, A. C. (2019). Dolores: a model that predicts football match outcomes from all over the world. Machine Learning, 108, 49–75.
6. Huang, K.-Y., & Chang, W.-L. (2010). A neural network method for pre- diction of 2006 world cup football game. In The 2010 international joint conference on neural networks (IJCNN) (pp. 1–8). IEEE.
7. Rory Bunker, Teo Susnjak (2019). The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review, 15-19.
8. 3Blue1Brown. (2018). YouTube: But what is a neural network? | Chapter 1, Deep Learning