BIG DATA, NEURAL NETWORK AND PREDICTIVE ANALYTICS: APPLICATION IN THE FIELD OF SPORT

Mihail Konchev

National Sports Academy „Vassil Levski‘, Sofia, Bulgaria

ABSTRACT

The potential of big data and neural network to improve the prediction of sport outcomes is tremendous. In this study, the Recurrent Neural Network (RNN), like a class of artificial neural network, has been investigated for predicting the outcomes of football matches. The aim of this paper is to focus on the application of the neural network for predictive analysis purposes in the field of sport.

Classification is a task that is often encountered in our life. A classification process involves assigning objects into predefined groups or classes based on a number of observed attributes related to those objects. Although there are some more traditional tools for classification, such as certain statistical procedures, neural networks have shown to be an effective solution for this type of problems. Neural networks classify objects rather simply - they take data as input, derive rules based on those data, and make decisions.

The methodology of this study is based on data from the English Premier League for the period 1993-2017. The analyzed database includes: Full Time Home Team Goals, Full Time Away Team Goals, Half Time Home Team Goals, Half Time Away Team Goals, Home Team Shots on Target and Away Team Shots on Target. In this paper we have studied several different ways of forming up input data sequences, as well as different architectures of RNNs that may lead to effective prediction.

The test results have shown that neural networks may be used for successfully predicting the outcomes of football matches. For further increasing the performance of the prediction, prior information about each team would be desirable.

Key words: Neural Network, Data mining, Predictive Analytics

INTRODUCTION

Modeling complex processes and interactions is a common task in process management. Neural networks are of particular interest in modeling complex problems with multiple interactions. They offer effective tools both for describing data and for solving prediction tasks. In theory (Medvedev, Potiomkin, 2002), (Mladenov, Iordanova, 2006), (Venkov, Gilev, 2009) neural networks are considered as systems for parallel information processing. Their main feature is their ability to store and use knowledge.

Structurally, neural networks consist of simple information processing elements called neurons, which are organized into layers interconnected by weighted links. The weights of the links determine their strength. The input to each neuron is the weighted sum of the signals from the other neurons. The information received is accumulated and transmitted over the network through so-called activation functions.

The potential of big data and neural network to improve the prediction of sport outcomes is tremendous. In this study, the Recurrent Neural Network (RNN), like a class of artificial neural network, has been investigated for predicting the outcomes of football matches.

The beginning of the prediction of the time series was made in the 20's of the past century with the introduction of the autoregressive model for the prediction of the annual number of sunspots. Time series are samples of system's behaviour over discrete time values. The neural networks ability to cope with the nonlinearities, the speed of computation, the learning capacity and the accuracy made them valuable tools of prediction.

The most difficult systems to predict are: (Danilo, Mandic, Jonathon, 2001)

- Those with insufficient amount of data in the data series (for instance chaotic signals);
- The systems whose data series are obtained through observation and measurements, in this

case data being possibly affected by measurement errors;

- Systems whose behavior varies in time.

The aim of this paper is to focus on the application of the neural network for predictive analysis purposes in the field of sport.

AIM AND OBJECTIVES OF THE STUDY

The aim of the study is to attempt to answer the question: "To what extent can neural networks predict events of a random nature, such as the outcome of football matches (1X2)?" In order to achieve this objective, it is necessary to fulfill the following tasks, which also determine the structure of the study.

- Determination of the database on which the study will be performed using a neural network;
- Determination of input and output parameters;
- Presentation of the methodology and results of the study;
- Conclusions.

METHODS

The neural network topology determines how multiple neurons in the network are arranged and connected. It is the most important characteristic of neural networks because it defines their functional features and therefore their application areas.

The most common neural network architectures are:

- Single-layer networks with one-way signal transmission;
- Multilayer networks with one-way signal transmission;
- Recurrent networks - feedback architectures.

A recurrent neural network architecture (RNN) will be used for analysis purposes.

It is well known that neural networks only work steadily with limited activation functions. The most commonly used are:

- Linear functions - are unipolar, in which the output signal varies in the range $[0 \div 1]$ and bipolar, with the signal range $[-1 \div 1]$;
- Threshold Functions - Also called binary or sigma functions, they are unipolar, in which the output signal varies in the range $[0 \div 1]$ and bipolar, with the signal range $[-1 \div 1]$. This type of activation features is most commonly used to implement simple network nodes. Their main disadvantage is their limited training capabilities on the neural network;

- Sigmoid functions - are unipolar and bipolar. Most often the output signal is limited in the following intervals: $[0 \div 1]$, $[-1 \div 1]$ and $[-1/2 \div 1/2]$;
- Radial basis function - it obtains a maximum value only when all the input signals have values close to their assigned central value. The function is practically zero when the values of the input signals differ significantly from the set center value. The application of neural networks with radially basis functions is mainly related to the tasks of object identification and management.

A sigmoid function will be used for analysis purposes.

Neural damage training is a process in which, through a purposeful change in the weighting coefficients of the neural connections, the networks are given the desired properties and behavior. Training approaches can be grouped into three groups: teacher training, non-teacher training, and incentive training. *A teacher training will be used for analysis purposes.*

Teacher training is also known as "error minimization training". Its name is a consequence of the algorithm that follows the learning process.

1. The initial values of the weight coefficients are set;
2. Data is input at the input of the neural network;
3. The results obtained at the output are compared with the desired output values. The difference in values is accumulated as a cumulative error, which is taken into account in the subsequent tuning of the neural network;
4. The initial values of the weighting coefficients are varied according to the value of the cumulative error;
5. Repeat the process until the error is minimized to a preset value.

There are two stages in neural networks using teacher training - training and operating. During the training phase, the weights are adjusted to obtain the desired neural network behavior. Characteristic of the operation phase is that the weight coefficients found at the beginning are of constant values and the network works with them successfully with different input data.

In this study, 70.40% of the database volume will be used for neural network training and the remaining

20.60% for testing. The database covers 6225 football matches of the English Premier League. The input parameters of the neural network through which the analysis will be performed are:

- FTHG - goals scored by the home team (arithmetic mean);
- FTAG - goals scored by the away team (arithmetic mean);
- HTHG - Home team goals scored (arithmetic mean);
- HTAG - Away Team goals scored (arithmetic mean);
- HS - hits on the door by the home team (arithmetic mean);
- HA - Shots sent to the door by the guest team (arithmetic mean);
- HST - Accurate shots fired to the door by the home team (arithmetic mean);
- AST - sent accurate shots to the door by the guest team (arithmetic mean);

Output data

- FTR - final result (1x2)

The interpretation of the results will be interpreted using the „Cumulative gains chart”.

RESULTS

The experiment showed the following results:

The percentage of successfully made neural network forecasts is as follows:

51.60% of the training sample and 53.40% of the test sample. The results are presented in the Figure 1 below.

The cumulative gains chart shows the percentage of the overall number of cases in a given category “1X2” by targeting a percentage of the total number of cases. The diagonal line is the “baseline” curve; if we select 10% of the cases from the scored data-set at random, we would expect to “gain” approximately 10% of all of the cases that actually take the category 1 – 1, 2 – X or 3 – 2. The farther above the baseline a curve lies, the greater the gain. We can use the cumulative gains chart to help choose a classification cutoff by choosing a percentage that corresponds to a desirable gain, and then mapping that percentage to the appropriate cutoff value.

Figure 1. Cumulative gains chart of the predicted results

Importance of factors in the model

Figure 2 presents the factors that have the greatest impact on the forecasting model

Figure 2. Importance of the factors of the model

The optimum number of neurons in the architecture of the neural network is 7, as can be seen from figure 3.

Figure 3. Neural Network Architecture

The results are classified into 3 groups as follows: 1 - win for the home team, 2 - equal result and 3 - win for the away team.

DISCUSSION

As can be seen from Figure 1, the neural network predicts the most successful wins for the away team, followed by home wins and even results. The larger area locked between the model curves and the baseline, shows that the predictive model is better.

For example, the first point on the curve for the third category is at (10%, 21%), meaning that if you score a dataset with the network and sort all of the cases by predicted pseudo-probability of third category, you would expect the top 10% to contain approximately 21% of all of the cases that actually take the result „Win for the away team” (defaulters). Likewise, the top 20% would contain approximately 37% of the defaulters; the top 30% of cases would contain 50% of defaulters, and so on. If you select 100% of the scored dataset, you obtain all of the defaulters in the dataset.

The diagonal line is the “baseline” curve; if we select 10% of the cases from the scored *dataset at random*, you would expect to “gain” approximately 10% successfully forecasts

The results also shows that the most important factors in the forecasting model are the following 5 factors:

- *Home team goals scored;*
- *Half-time goals scored by the away team;*
- *Half-time goals scored by the home team;*
- *Away team goals scored;*

- *Shots on the door from the home team.*

In summary, the following conclusions can be drawn:

A number of factors have a significant impact on the architecture and forecasting efficiency. Depending on the purpose of the analysis, neural networks may be applicable to the predictive modeling of random events such as the results of football matches;

REFERENCES

- Medvedev, V. S., Potyomkin, V.G., (2002). Neironae seti. MATLAB 6, Moskva, 2002/ Медведев В.С., Потёмкин В.Г., (2002). Нейронные сети. MATLAB 6, Москва;
- Mladenov, V., Iordanova, S., (2006). Razmito upravlenie na nevronni mrezhi, Tehnicheski universitet – Sofia, 2006/ Младенов, В., Йорданова С. (2006). Размито управление и невронни мрежи, Технически университет – София;
- Venkov, G., Gilev, B., (2009). Prilozhenie na nevronniti mrezhi v kompiutarnoto modelirane, Tehnicheski universitet – Sofia, 2009/ Венков, Г., Гилев, Б. (2009). Приложение на невронните мрежи в компютърното моделиране, Технически университет – София, 2009;
- Danilo, P. Mandic, Jonathon, A. (2001). Chambers: Recurrent Neural Networks for Prediction, John Wiley and Sons Ltd.

Corresponding author:

Mihail Konchev

National Sports Academy „Vassil Levski”,
21 „Acad. Stefan Mladenov” Street,
1700 Sofia, Bulgaria
E-mail: mivailov@abv.bg