Journal of Soft Computing and Artificial Intelligence
JSCAI, 1(2): pp:92-99 (2020), www.dergipark.org.tr/en/pub/jscai
JSCAI
ISSN: 2717-8226
Received: November 27 2020 Accepted: December 19 2020

Categorization of Qualifying Football Clubs for European Cups with Backpropagating Artificial Neural Networks

[1]Bünyamin Fuat Yıldız

[1]M.Sc. (Econ.), 24 Crooks Ave Apt 229, Clifton, NJ 7011-1614 United States
Corresponding author: Bünyamin Fuat Yıldız e-mail: bunyaminfuatyildiz@yahoo.com

ABSTRACT European cups are the most popular and most profitable football organization in the world. The participation of football clubs in the Champions League and the Europa League is, therefore, a matter of interest to all parts of society. In this respect, this paper uses backpropagating ANNs to understand the capability of categorizing football clubs from Italy, England, and Spain. The sample consists of 10 years of data from Seri A, English Premier League, and La Liga — and teams categorized as qualified and unqualified. As a result of the test, backpropagating ANN classifies the clubs with 92.7 percent accuracy. Our model correctly categorized 40 of 51 qualified teams in our test dataset—that is approximately 78 percent accuracy. However, our backpropagating ANN provides more significant accuracy while predicting unqualified teams, that is approximately 98.5 percent. The probable reason for lower accuracy in the categorization of qualified teams might be underrepresentation in the dataset and lack of variable diversity. The success of ANNs implies that it could be interesting to integrate ANNs into an online betting platform to develop solutions for more complex events by introducing more data. The application of other machine learning approaches will contribute to the literature and provide an opportunity to compare methods.

KEYWORDS: Machine Learning, Backpropagated Artificial Neural Networks, Football

1. INTRODUCTION

Although oil was an influential source in the last hundred years, data seems to take its place. Nevertheless, it's not enough to merely have masses of data. It is essential to extract information from this meaningless bulk of data by applying different techniques. In this context, one of the most effective methods to turn data into information based on machine learning. Especially, artificial neural network algorithms are skilled at extracting meaningful information and efficient predictions from cumbersome amounts of data. Therefore, it has become a workhorse of researchers from different branches that produce effective solutions from customer loss analysis to cancer detection. However, there are fewer implementations of machine learning techniques in football—which creates high economic benefit for the economies.

It is the goal of football clubs to earn millions of euros in broadcasting and sponsorship revenues, as well as prestige, by participating in European Cups. Therefore, high commercialization has made football no longer an ordinary sport and makes it the focus of scientific research. Besides, various empirical studies have been carried out to achieve optimal performance. Nonetheless, machine learning applications do not have a very long history in football. One well-known early work that is often cited in research used Bayesian nets to investigate the outcome of the match result of a single team [1]. There is a great deal of recent football literature that focuses on tactical knowledge [2-4]. The number of studies on betting in football, which will have widespread popularity in the future, are relatively few

and mostly based on offline data [5-8]. [9] conducted a big data analysis to detect hooliganism in stadiums which will probably reduce the vandalism in stadiums. Studies involve various approaches for evaluating footballers' values on the transfer market [10-11]. [12] categorized European football teams into three sub-groups and investigated the categorization power of decision trees to understand certain qualities of each classified club.

In the context of the work of [12], it is possible to classify football clubs depending on their specific characteristics. If the classification process is carried out successfully, common characteristics of football clubs can be revealed. This study was conducted to evaluate the categorization performance of backpropagation ANNs —which aimed to infer common characteristics of football clubs for the qualifying teams in European Cups. The efficiency of artificial neural networks (ANNs) was tested for the first time to predict whether football clubs would be in the category to participate in European cups. In this respect, the classification made in the output layer into two groups: a) the qualified football clubs which finished the season well to qualify European cups b) the remaining football clubs denoted as unqualified. The dataset consists of 200 observations from each league. The number of ANN input layer nodes: goal per game (gperg), conceded goals per game (gapg), successful pass percentage (pass), and game possession (poss). Apart from being the first in the literature, this study might initiate new studies on the categorization of football clubs in this field. The following parts of this paper presented as follows. Section 2 contains information regarding the method and the data. In the first half of section 3, preliminary statistics shared, and then the result of the application of artificial neural networks was evaluated. The conclusion of the study is provided in the end.

2. METHOD AND DATA

2.1. Artificial Neural Networks

The technique of ANNs was developed by inspiring the relationship between the neuron and nervous system in the human brain and became popular in various fields of studies[13-14]. Information acquisition is achieved by creating various configurations with several mathematically structured neurons. By combining mathematical neurons with a variety of input vectors, a "neural network" could be configured. In all of the studies that have been carried out so far, neurons form parallel layers. Although the neuron structure has the same type of transfer function, each neuron has a weight matrix and a bias vector with a diversified set of values corresponding to the input fit of a threshold value [15]. Given the equivalent arrangement of the neurons and neural layer, the essential knowledge process and update the weights and biases. Typically, each weight matrix and vector of biases commences with arbitrary values, so as not to generate any bias in the neural network. In this situation, learning rules are necessary to get rid of randomness systematically and achieve meaningful patterns. This need is met by the learning rules that emerge as a result of training the network with data that updates weights and biases [16]. Thus, the smallest structure that makes calculations to classify the properties of the inputs called perceptron is formed [17].

The method utilized in this research is the backpropagating ANN algorithm. In the current literature, backpropagation algorithms have remained popular because of their ease of application and their multi-functionality. The main benefit of using this algorithm is to propagate the error between the output value and the actual output value obtained during the training of the network back along with their weights. When the output produced falls below a certain error value, the update process is terminated by assuming that the algorithm has learned. The special issue in the working principle of the backpropagating algorithm is that the network evaluates the output it produces with real outputs and updates each iteration until it reaches the desired error rate level. To implement the backpropagating ANNS, the neuralnet package was used which is developed by [18] in the R programming language. Since the number of hidden layers not determined by precise rules, The number of layers with the highest predictive performance will be reported. (Basheer & Hajmeer, 2000). Application of backpropagating ANN algorithm with the processor 2.60GHz i5-4300 CPU with 8192MB RAM. Accordingly, it is possible to obtain various performances with a different number of layers and nodes on computers with different features. To summarize the process of the ANN algorithm structure, inferences made using functional relationships in three stages: (i) it takes the inputs, (ii) sends it to the hidden layer (iii) and generates outputs. Besides, 70 percent of the data set was randomly allocated for training, and the remaining part for performance testing.

2.2. Dataset

The dataset was gathered from statistics sections of whoscored.com covering 600 observations from Premier Leauge, Serie A, and La Liga between the seasons 2009-2010 and 2018-2019. The dependent variable is denoted as "class". The football clubs playing in the mentioned leagues are divided into two categories. The clubs that have finished the league well enough to qualify for European Trophies classified as qualified. The rest of the clubs were categorized as unqualified. Four different independent variables were used in the learning process. These are the goals per game (gperg), the average percentage of successful passes per game (pass), game dominance (poss), and an average conceded goal per game (gapg).

3. EMPIRICAL RESULTS

3.1. Preliminary Statistical Information

In this sub-section, all analyses and graphs were carried out using Stata software. Descriptive statistics for all variables were presented in Table 1. Table 1 includes the number of observations (obs) in the second column. It is clear from the number of observations that there are no missing values. To show mean, minimum (min.), Maximum (max.), and standard deviation (SD). To reflect the properties of the data set, means, standard deviation (SD); and minimum (min) and maximum (max) values were shared.

Goal per game has a mean of 1.35 with 0.45 standard deviation. Minimum and maximum values are 0.57 and 3.18. The possession has a mean of 0.49 by 0.04 standard deviation ranging from 0.39 to 0.67. The successful pass percentage ranges from 0.62 to 0.89, with a mean of 0.77 with a 0.05 standard deviation. The number of average goals conceded per match has a mean 1.35 minimum 0.47 maximum of 2.47 and 0.33 standard deviations on average. Moreover, there are no extremities in the data set. So, there is no need for additional operation for normalization.

Table 1. Preliminary statistics for each class of teams

Variables	Obs	Mean	SD	Min	Max
gperg	600	1.35	0.45	0.57	3.18
poss	600	0.49	0.04	0.39	0.67
pass	600	0.77	0.05	0.62	0.89
gapg	600	1.35	0.33	0.47	2.47

Table 2 provides pairwise correlation coefficients for subjected variables to assess the strength of the relationship. The first column is aimed to demonstrate correlations among goals per game with other variables. As shown in Table 2, the goal per game is most highly correlated with average ball possession per game. The correlation coefficient of a goal per game with the successful pass percentage is 0.64. However, goal per game has only a negative correlation with conceded goals per game. Interestingly, the correlation between ball possession and a successful pass percentage has the strongest correlation. The negative relationship between possession percentage and conceded goals is -0.58 which is the second-highest negative relationship presented on the table. Lastly, the correlation coefficient between conceded goals and successful pass percentage is -0.51.

Table 2. Pairwise Correlations

Variables	gperg	poss	pass	gapg
gperg	1.00	—	—	—
poss	0.75	1.00	—	—
pass	0.64	0.83	1.00	—
gapg	-0.67	-0.58	-0.51	1.00

Since the most important measure of success in football is to score, the figure below is also shared. In Figure 1, a graph of the quadratic ordinary least squares regression estimates was shared to reflect the relationship of the goal scored per game with other variables. In the regression determined quadratically between the variables, the outlier numbers outside the confidence interval were found to be relatively few. When we consider the left pane of figure 1, the average number of goals scored per game has a negative linear relationship with the average number of goals conceded per game. Also, in

the middle of Figure 1, there is a positive linear relationship between the percentage of successful passes and goals scored per game. On the right side of Figure 1, possession appears to be a significant factor in average goals scored per game.

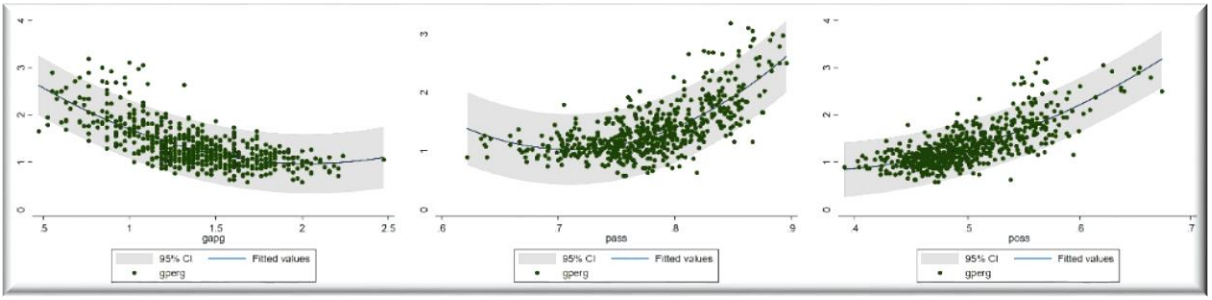

Figure 1 *The quadratic OLS plot between goal per game and other variables*

3.1. Result of Artificial Neural Networks Application

Figure 2 displays the result of ANN giving the optimum output. As seen in Figure 2, two hidden layers are consisting of 3 and 1 neurons respectively. The weight is calculated per synapse showing the influence of the related neuron, and the relevant data is sent as signals to the neural network. Since the second hidden layer with the single neuron helps in the performance of the algorithm two hidden layers were used.

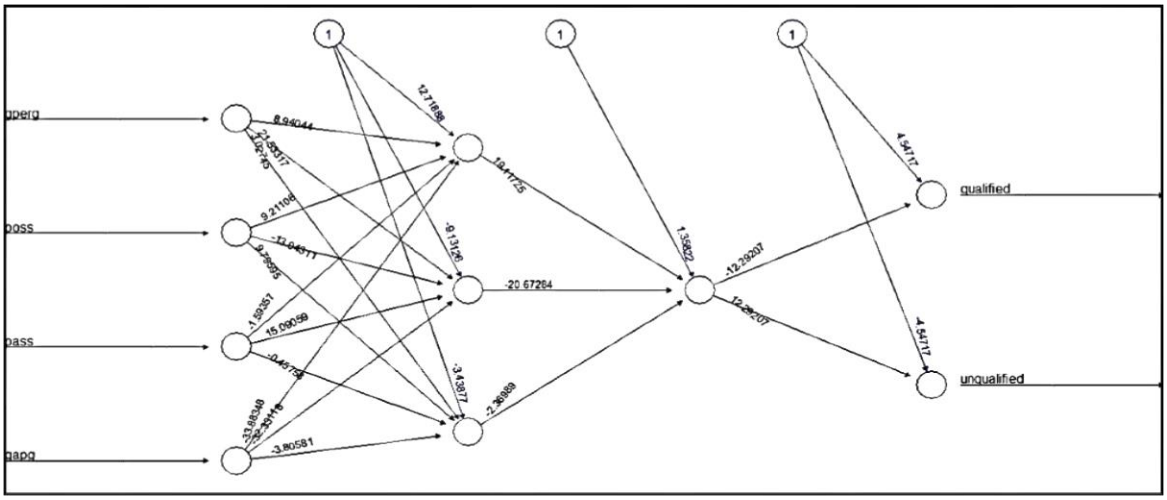

Figure 2 *The ANN plots.*

Table 3 gives the results of the confusion matrix to show the classification performance. Accuracy is the rate of correctly categorized cases compared with the number of whole cases. The accuracy of the ANN application is 92.7 percent. Out of the 51 qualified teams in the test data set, 40 were correctly predicted; however, 11 of the teams were unqualified. This constitutes a success percentage of about 78 percent. Our ANN model has achieved very successful results in the categorization of unsuccessful teams. Out of the 128 unsuccessful teams, only 2 were wrongly predicted. This corresponds to an accuracy rate of about 98.5 percent. The No Information Rate (NIR) reflects the accuracy achieved

when the prediction option is used in the direction of the majority-forming category. The NIR is 0.71, as unqualified teams make up 71 percent of the data. The "95% CI" indicates the range of values of the accuracy level within the 95 percent confidence interval. The smallness of the value obtained in P-Val [Acc> NIR] makes it easier to decide on the statistical reliability of the applied method. The p-value is extremely small shows that there is certainly no possibility that the accuracy of NIR (71%) is higher than the accuracy of the ANN (92%).

Table 3. Confusion Matrix and Statistics

CONFUSION MATRIX		**Reference Values**	
		qualified	unqualified
Predicted Values	qualified	40	2
	unqualified	11	126
Accuracy	0.927	**NIR**	0.71
95% CI	0.87, 0.96	**P-Val [Acc>NIR]**	1.337e-12

4. CONCLUSION

This study set out to evaluate how effective the backpropagating ANN algorithm classifies football clubs in the three major European leagues. The application aims to determine the status of teams in the English Premier League, Seri A, and La Liga whether to participate in European Cups. Seventy percent of the available data were selected to train the ANN. The accuracy of the model was checked with the test data, following the training process. The results of the test show that the accuracy of the ANN was 92.7%. Note that while the accuracy rate is lower for qualified teams, the prediction success of unqualified teams is higher. There could be two probable reasons for this. The success of clubs constitutes a more complex form. Therefore, it might difficult to categorize qualified teams with the existing variable set. The second possible reason may be due to the underrepresentation of qualified European football clubs in our data set.

These results contribute to the rapidly expanding field of machine learning applications in the literature of football. However, this research has thrown up many topics in need of further investigation. The methods used for this analysis may be applied to other leagues elsewhere in the world. Besides, it would be interesting to repeat the experiments described here with different techniques such as support vector machines. Hence, it is possible to highlight the difference between the accuracy of techniques. All in all, techniques like artificial neural networks are used effectively in all disciplines; they are relatively new in football. The limitations of this study predictions are made from historical information. Determining the result by live performance indicators might probably be more interesting. Also, the performance of the teams at the end of the league could be predicted from data such as player performances at the start of the league.

Concerning the implications of the new technological tools for football fans, closing the gap between football clubs can be achieved as a consequence of coworking sports scientists and data analysts.

REFERENCES

[1] Joseph A, Fenton NE, Neil M. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems. 2006;19(7):544-53.

[2] Rein R, Memmert D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus. 2016 Dec;5(1):1-3.

[3] Berrar D, Lopes P, Dubitzky W. Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine learning. 2019 Jan 15;108(1):97-126. https://doi.org/10.1007/s10994-018-5747-8.

[4] Herold M, Goes F, Nopp S, Bauer P, Thompson C, Meyer T. Machine learning in men's professional football: Current applications and future directions for improving attacking play. International Journal of Sports Science & Coaching. 2019 Dec;14(6):798-817. https://doi.org/10.1177/1747954119879350.

[5] Bunker RP, Thabtah F. A machine learning framework for sport result prediction. Applied computing and informatics. 2019 Jan 1;15(1):27-33.

[6] Hubáček O, Šourek G, Železný F. Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning. 2019;108(1):29-47. https://doi.org/10.1007/s10994-018-5704-6.

[7] Stübinger J, Mangold B, Knoll J. Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. Applied Sciences. 2020;10(1):46. https://doi.org/10.3390/app10010046.

[8] Rudrapal D, Boro S, Srivastava J, Singh S. A Deep Learning Approach to Predict Football Match Result. InComputational Intelligence in Data Mining 2020 (pp. 93-99). Springer, Singapore.

[9] Fenil E, Manogaran G, Vivekananda GN, Thanjaivadivel T, Jeeva S, Ahilan A. Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Networks. 2019;151:191-200. https://doi.org/10.1016/j.comnet.2019.01.028.

[10] Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S. L., & Ortega, G. J. (2018). Transfer market activities and sportive performance in European first football leagues: A dynamic network approach. PloS one, 13(12), e0209362.

[11] Singh P, Lamba PS. Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. Journal of Discrete Mathematical Sciences and Cryptography. 2019 17;22(2):113-26. https://doi.org/10.1080/09720529.2019.1576333.

[12] Yıldız BF. Applying Decision Tree Techniques to Classify European Football Teams. Journal of Soft Computing and Artificial Intelligence. 2020; 1(2): 29-35.

[13] Kaplan K, Kuncan M, Ertunc HM. Prediction of bearing fault size by using model of adaptive neuro-fuzzy inference system. In2015 23nd Signal Processing and Communications Applications Conference (SIU) 2015 May 16 (pp. 1925-1928). IEEE.

[14] Bayram, S., Kaplan, K., Kuncan, M., & Ertunç, H. M. (2014, April). The effect of bearings faults to coefficients obtaned by using wavelet transform. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)* (pp. 991-994). IEEE.

[15] Cavuto DJ. An exploration and development of current artificial neural network theory and applications with emphasis on artificial life. Unpublished Master of Engineering Thesis. The Cooper Union, Albert Nerken School of Engineering, New York, NY. 1997.

[16] Dreyfus SE. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. Journal of guidance, control, and dynamics. 1990;13(5):926-8. https://doi.org/10.2514/3.25422.

[17] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods. 2000 1;43(1):3-1.

[18] Günther F, Fritsch S. neuralnet: Training of neural networks. The R journal. 2010 1;2(1):30-8.