

Research Article

Neural Network Based Technical Analysis of Football Games

Zhaojun Li* 

School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China

Abstract

Football is the most famous sports in the world, and English Premier League is the number one league in the world for three consecutive years (Fédération internationale de football association, FIFA). It is always interesting to apply technical analysis to understand what makes the best football player and football team. In this article, we try to answer the question "what kind of tactical play is the most advanced" through statistical analysis on the game data of English Premier League in the 21-22 season. To be more specific, firstly, we applied descriptive statistics to analyze the technical and tactical play of each team, and then screen out the technical and tactical indicators that significantly affect the outcome of the game through one-way analysis of variance (ANOVA) and discriminate analysis, and the preliminary target conclusions were obtained. BP neural network was then carried out to predict the rankings of the Premier League teams by using the indicators selected by ANOVA and discriminant analysis, as input value. BP neural network prediction model is then established to predict the ranking of each team in the 22-23 season. A general conclusion and make suggestions on the planning of the technical and tactical playing methods of our country's youth soccer sports.

Keywords

English Premier League, One-way ANOVA, Discriminant Analysis, BP Neural Network

1. Introduction

The history of game data analysis is not long. However, since the arrival of the information age, with the acceleration of network communication speed and iterative upgrading of intelligent analysis system, the sports data become more available, and the number of game data analysis has grown explosively. In this article, we choose to analyze the data of football games. As one of the most famous sports in the world, football games have large number of participants, diversified tactical play, and changeable formations, and the relationship between the technical data and the final result is intricate. Through statistical analyzing tools, we hope to examine the

relationship between technical performance and the final result (win, draw, or defeat), what type of tactical styles will lead to better results, and the trends in the technical and tactical play of today's top football leagues [1, 2].

With the popularization of the Internet, the upgrading of broadcasting technology, and the progress of science and technology, the audience of football around the world has increased significantly, and more and more people are concerned about the factors affecting the outcome of football matches. Formation defines the spatial position and division of responsibilities of players on the football field. Previous

*Corresponding author: lzj407@zufe.edu.cn (Zhaojun Li)

Received: 29 February 2024; **Accepted:** 21 March 2024; **Published:** 25 March 2024



Copyright: © The Author(s), 2023. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

researchers proposed a series of mathematical models and mining algorithms to study formations, but these were usually focused on the local area and could not be analyzed globally. Xie, X. proposed a visual analysis system to visually observe the movement of players when the formation changes, and users can understand the reasons for the formation change and the impact of the formation change based on the contextual background information, and then understand the impact of the formation on the outcome of the game [3]. Based on the generalized linear model and the data series inference method, Liu, H. and Peng, Z. concluded that every 2 standard deviation shots can increase the team's winning rate by 16.3%, the number of shots on target with 2 standard deviations can increase by 33.8%, the increase of 2 standard deviation passes, pass success rate and through plug can bring 21.6%, 27.3% and 16.9% to the team's probability of winning, and the number of fouls with 2 standard deviations will reduce the probability of winning by 25.4% [4]. When it comes to the football games in Europe, Hu, Z. studied the technical and tactical application of Euro 2008, and showed that the mid-range pass was the most effective pass to launch an attack, the cross was the most effective way to assist the goal, the midfield steal was the focus of cutting off the opponent's rapid attack, and the goalkeeper's save was the key to winning or losing the game [5]. Liu, H. analyzed Leicester City in the 15-16 championship season and found that Leicester City is a team with distinct overall offensive tactics, mainly in a 4-4-2 formation, focusing on wing attacks, and combining effective attacks in the middle [6]. Although Leicester City that season was known as the biggest dark horse in football in the past decade, judging from its subsequent player development, the Leicester City player talent was also top-notch. N'Golo Kante, a midfielder with historic coverage and defensive skills, Riyad Mahrez, who has been an African footballer for many years, and Leicester City's all-time goalscorer Vardy in the center of the field. From the perspective of statistics, it can also be seen that the championship is not accidental, Leicester City in terms of interceptions, tackles, accurate crosses and other statistics are significantly higher than other teams that season, it is a team that relies on excellent midfield defensive interceptions and efficient counter-attacks to win the championship. Chen, Z. analyzed the 19-20 championship team Liverpool and found that Liverpool is a team that pursues fast attack, and sports scoring is the main way to score, and its pressure in the front court is very strong [7]. In the fast attack, stealing the opponent's possession in the front court is Liverpool's main way to get offensive opportunities, usually Liverpool's players will form a chance to hit the goal through short-distance passing, and complete the shot through the striker in the way of receiving the ball. When attacking from the side, the two full-backs are different from other teams.

With the advent of the era of big data, statistical analysis methods and artificial intelligence technology have been developed unprecedentedly, and have been widely used in various disciplines. For example, the artificial neural network

a computation system that attempts to mimic (or inspired by) the neural connections in 63 animals' nervous system [8-10].

However, the classic BP neural network has two major drawbacks. On the one hand, from a mathematical point of view, the algorithm can only look for local maxima or minima, and so might not be suitable for problems with several extreme values; on the other hand, the structure of a neural network is usually chosen by previous experience, and so the prediction results might not be desired.

As a result, in this article, we propose to analyze the data through a three-stage procedure: first, select variables that are statistically significant to the final result of the game; apply discriminant analysis to further screen out important variables; and finally, apply BP neural network to predict the result of the next season.

2. Data and Methodology

2.1. Methodology

One-way ANOVA is a method used to detect whether factors have a significant difference in test results. It is a hypothesis-based test, that is, the test is designed to evaluate our material on multiple mutually exclusive principles. In this paper, we examine whether the technical and tactical indicators have significant results on the team results (win, lose, or tie). The response variable then has three levels, and the hypotheses are:

H_0 : the effect of the indicator on the game results are the same;

H_1 : the effect of the indicator on the game results are different.

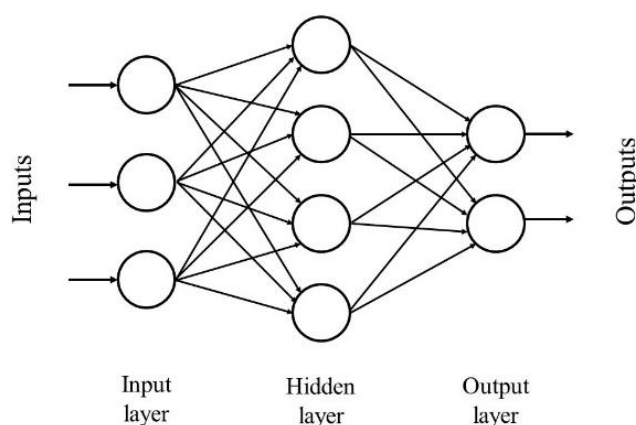


Figure 1. An example of a neural network.

Discriminant Analysis is a statistical method that emerged in the 1930's, which identifies samples of unknown types by identifying known types of samples. At present, discriminant analysis has been widely used in medicine, architecture, journalism, economics and management. The characteristics

of discrimination analysis are that, in the past history, the classification rules of objective objects were summarized by summarizing the data and information of a certain type of multiple samples, and the identification equations and identification criteria were constructed accordingly. For new samples, the samples can be classified through the inductive judgment formulas and judgment criteria.

With the advent of the era of big data, statistical analysis methods and artificial intelligence technology have been developed unprecedentedly, and have been widely used in various disciplines. For example, the artificial neural network is a computation system that attempts to mimic (or inspired by) the neural connections in animals' nervous system [3], and among the popular learning algorithms, the BP feed-forward neural networks (Figure 1) would probably be the most successful one [11-13]. A feed-forward neural network consists of the following four parts:

- 1). Input layer: the first layer that receives input and passes onto the next layer.
- 2). Output layer: the last layer that contains the predicted values.
- 3). Hidden layer: the layers that are between the input layer

and the output layer. They should contain a huge number of neurons, which perform transformation on the inputted values and pass them onto the next layer.

- 4). Neuron weights: similar to the regression coefficients in the regression analysis, implying the strength or amplitude of a connection between two neurons.

2.2. Indicators

In this study, game results (win, lose or tie) is treated as dependent variable, and technical and tactical indicators are independent variables. To be more specific, 22 technical indicators are considered, namely: goals, shots on target, shots on goal, shots on target, expected goals, offside, passes, precise passes, precise long passes, precise short balls, possession, corners, free kicks, serves, successful tackles, interceptions, blocks, clearances, goalkeeper saves, red cards and yellow cards, and based on previous literature, we group them into three categories, namely, goal-scoring related variables, offensive organization-related variables, and defense-related variables. The grouping results are summarized in the following Table 1.

Table 1. Groups of independent variables.

Group	Indicators
goal-scoring related indicators	goals, shots on target, shots, shots on target, shots in the penalty area, expected goals
offensive organization-related variables	offside, passing, precise passing, precise long pass, precise crossing, possession, corners, free kicks, serve
defense-related variables	successful tackles, interceptions, blocks, clearances, goalkeeper saves, yellow cards, red cards

2.3. Data

The study focuses on the technical and tactical performance indicators of 380 matches in the main stage of the English Premier League in the 2021-2022 season, and the data is collected through Tencent Sports and FotMob.

Due to the large differences between the training data variables, the collected raw data cannot be directly used as inputs to the BP neural network, and so normalization must be carried out. Since the sample data are all non-negative, we applied the following linear transformation method:

$$y = (x - \min x) / (\max x - \min x)$$

3. Analyzing Result

3.1. Selection of Indicators

First of all, the one-way variance (ANOVA) was used to analyze the differences in technical and tactical performance between the victorious, tie, and losing teams, and the significance of the impact of these technical and tactical indicators on the game results was observed [14]. The results are shown in Table 2 below.

Table 2. Result of one-way ANOVA.

Group	Indicators	Win	Tie	Lose	F	P-value
1	goal	2.4 ±1.2	1.1 ±0.8	0.5 ±0.7	333.6	0.000
1	Shooting accuracy	40.3 ±13.2	32.2 ±15.2	30.4 ±16.0	38.2	0.000
1	Shot	15.2 ±5.4	13.0 ±5.4	10.4 ±4.8	62.8	0.000
1	Shot on target	6.0 ±2.6	4.0 ±2.2	3.1 ±2.0	121.2	0.000
1	Shot from the penalty area	10.1 ±4.2	8.2 ±3.7	6.4 ±3.5	66.9	0.000
1	Expected goals	2.1 ±0.9	1.4 ±0.7	1.0 ±0.6	131.9	0.000
2	offside	1.6 ±1.4	1.4 ±1.3	1.8 ±1.8	4.2	0.015
2	Pass	492.0 ±145.2	443.1 ±129.3	415.6 ±112.1	25.1	0.000
2	Precise Pass	408.6 ±152.3	353.3 ±133.0	331.8 ±112.9	25.2	0.000
2	Precise long passes	25.6 ±7.4	25.2 ±6.9	23.8 ±7.1	5.0	0.000
2	Precise lateral passing	4.0 ±2.3	4.4 ±2.8	3.8 ±2.4	3.2	0.041
2	Possession	53.6 ±12.3	50.0 ±13.0	46.4 ±12.4	22.1	0.000
2	corner kick	5.7 ±3.0	5.3 ±3.0	4.7 ±2.7	8.6	0.000
2	free kick	9.6 ±3.6	10.0 ±3.5	9.7 ±3.6	0.4	0.652
2	serve	19.3 ±5.6	20.5 ±5.8	19.9 ±6.3	2.6	0.079
3	successful intercept	9.3 ±3.6	9.4 ±3.6	9.5 ±3.3	0.3	0.778
3	intercept	9.3 ±4.0	10.2 ±4.7	9.6 ±3.8	2.8	0.059
3	Block	3.0 ±2.2	3.8 ±2.6	4.1 ±2.6	13.9	0.000
3	Clearance	18.4 ±8.7	19.7 ±8.5	17.8 ±8.0	2.8	0.060
3	Goalkeeper saves	2.5 ±1.8	2.9 ±1.9	3.4 ±2.1	15.5	0.000
3	Yellow card	1.5 ±1.2	2.0 ±1.2	1.7 ±1.3	8.4	0.000
3	Red card	0.0 ±0.1	0.1 ±0.2	0.1 ±0.3	10.0	0.000

From the above table, it can be observed that, except for interceptions and free kicks, which are not significantly prominent, all other technical and tactical indicators are significant. However, such results still cannot clearly indicate which tactical approach has a more pronounced impact on the outcomes of the matches. Further variable selection is needed for a clearer understanding.

The next step involves the use of discriminant analysis to analyze the technical and tactical indicators that make a significant contribution to defeating the opposing teams. Discriminant analysis refers to a method based on observed or

measured values of certain variables to classify the subjects of the study. In this case, discriminant analysis establishes discriminant functions for the technical and tactical performance indicators of teams that won, drew, or lost. By observing the standardized canonical discriminant function coefficients (SCDFC) of each technical and tactical performance indicator in the discriminant function, their relative contribution to the discriminant function can be determined. After standardizing the collected data, this study conducted linear discriminant analysis, and the coefficients obtained are presented in Table 3 [15, 16].

Table 3. Result of discriminant analysis..

Indicators	LD1	LD2
Goal	1.144	0.111
Possession rate	-0.109	0.441
Shot	0.508	-1.222
Shot on target	-0.345	0.968
Shoot accuracy	0.279	-0.530
Shot from the penalty area	0.054	-0.096
Expected goals	0.070	0.266
Pass	0.619	-3.813
Precise pass	-0.397	3.302
Precise long pass	0.0667	-0.135
Precise lateral pass	-0.255	-0.170
Serve	0.005	0.026
Corner kick	0.111	0.034
Free kick	0.051	-0.141
Offside	-0.004	0.330
Successful interception	0.003	0.085
Interception	0.072	-0.261
Block	-0.084	-0.057
Clearance	0.397	-0.480
Goalkeeper	-0.175	0.106
Yellow card	0.059	-0.522
Red card	-0.064	0.043

As the previous study pointed out, the indicators with absolute values of SCDFC greater than or equal to 0.3 are considered to significantly contribute to the composition of the discriminant function [7]. Therefore, we can conclude that six technical and tactical indicators, namely: goals, shots, shots on target, passes, accurate passes, and clearances are prominent and significant contributors. These six indicators precisely cover three different groups of variables, with three variables related to goals and shots, two related to attacking organization, and one related to defense.

Let us now take the significant indicators and incorporate them into the initial win-loss table for examination. An interesting result emerges: among the top 6 teams in the league, in specific match victories, the majority have a lower number of clearances than their opposing teams. However, for mid-table and lower-ranked teams in the league, when achieving victories, the number of clearances tends to be greater than that of their opposing teams. To confirm its discriminating accuracy, the predicted group data obtained after

discrimination are brought into the initial group data for predictive accuracy testing. The resulting contingency table is shown in Table 4 below.

From Table 4, it can be seen that the false positive rate is $(70+29+38+31+19+46)/760=0.3$, which passes the accuracy test.

Table 4. Contingency table of LDA discriminant coefficient predict..

Initial group	Prediction group		
	win	lose	tie
Win	235	38	19
Lose	70	60	46
tie	29	31	232

3.2. Neural Network Based Prediction Model

BP neural networks include an input layer, a hidden layer, and an output layer, and the hidden layer may have only one or more neurons. In the BP network, the recognition rate of nonlinear functions can be improved by increasing the number of hidden layers of the network, while the recognition rate of the network will be decreased if the number of hidden layers is too high. Finally, the most basic three-layer BP neural network architecture is determined, and it is theoretically confirmed that the three-layer BP neural network can approximate any kind of nonlinear continuous equation with any form with any precision.

Step 1: determine the number of nerve cells at each level.

Input layer: The number of input layer neurons is determined according to the number of input variables of the problem to be solved, this paper takes the 21-22 Premier League games as the main analysis content, and tries to construct and use the BP neural network model to predict the win, draw and loss of each team in the 22-23 season of the Premier League and the final ranking of the season. In the previous paper, we screened out 20 significant technical and tactical indicators by one-way ANOVA, and then performed discriminant analysis to screen out 6 prominent significant variables. Therefore, the input variables we finally determined were 6 variables of goal, shot, shot on target, pass, precise pass, and clearance, so there are 6 neurons in the input layer.

Output layer: In the prediction model, the solution to the problem is to find a team's scoring expectation, that is, the output is a value of a (0,3) interval, so the number of neurons in the output layer is determined to be 1.

Hidden layer: This hidden layer plays a crucial role in whether the BP neural network is suitable for its model. If the number of neurons in the hidden layer is too small, the prediction accuracy of the model will be reduced, and if the number of neurons in the hidden layer is too large, it will

cause overfitting of the network. Rather than overfitting, we prioritize avoiding the problem of low prediction accuracy. However, at present, there is no definitive scientific method for determining the exact number of neurons in the hidden layer. There are three empirical formulas available to calculate the number of neurons, namely.

$$s = \alpha + \sqrt{n+1},$$

$$s = \log_2 n,$$

$$s = \sqrt{nl}$$

where s is the number of neurons in the hidden layer, n represents the number of neurons in the input layer, l represents the number of output layers, and α is a natural number between 1 and 10. Based on these three formulas, the number of neurons in the hidden layer is decided to be 3 in our study.

Step 2: determine the excitation function.

It is used to assist the neural network in learning and understanding the nonlinear functional features, which is the soul of the model. The mathematical model of each neuron is shaped by a different excitation function, which allows the neural network to process a wide variety of types of data with different information processing modes. The information processing mode is the key factor affecting the overall performance of the model, and the selection of appropriate and excellent excitation functions can greatly improve the prediction accuracy and operational efficiency of the network. In this case, in order to better predict wins and losses in the league, we used three levels of BP neural networks, and selected three different incentive functions. {em Input layer:} Since the training samples of the 6 neurons are discrete, there is no need to perform any functional transformation, so the input layer can directly use the general linear excitation function:

$$f(x) = x.$$

Hidden layer: In order to enhance the ability of the neural network to deal with nonlinear relationships, the excitation function selected in the hidden layer in this paper is the Sigmoid function, also known as the logistic function. The Sigmoid function is defined as:

$$S(x) = 1 / (1 + e^{-x}).$$

Output layer: The range of the Sigmoid function is [0,1], and the desired scoring expectation is a probability value in the interval of [0,3], and therefore, we decide to use the Tanh function as the stimulus function in the output layer, defined as:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

The topology diagram of the built BP neural network is shown in Figure 2.

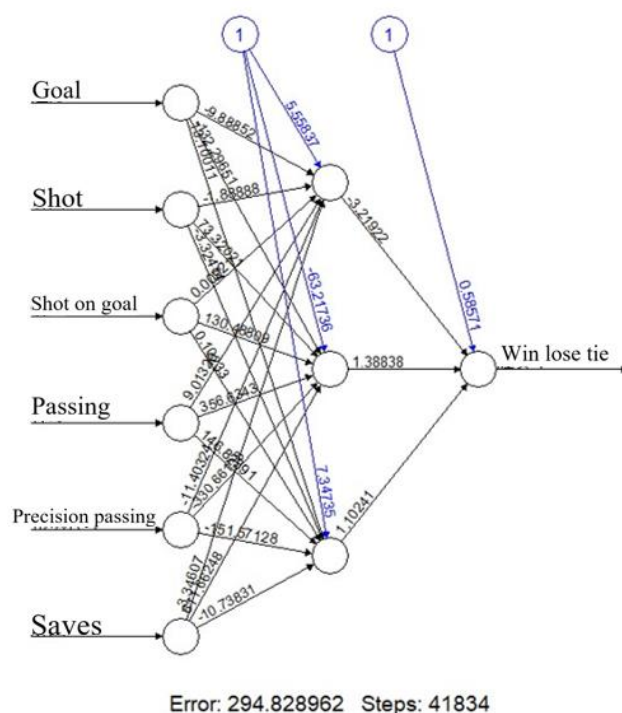


Figure 2. The topology diagram of the BP neural network.

Finally, based on this prediction model, we predict the performance of each team in the next season, and the result are summarized in Table 5.

Table 5. Coefficient and predicted ranking of each team.

Team	Predicted season points coefficient	Rank
Liverpool	80.216	1
Arsenal	78.941	2
Everton	73.019	3
Manchester City	71.264	4
Southampton	71.068	5
Aston Villa	70.248	6
West Ham United	66.881	7
Wolverhampton	65.193	8
Watford	63.565	9
Brighton	63.528	10
Manchester United	60.438	11
Leeds United	60.300	12
Brentford	58.709	13
Newcastle United	57.680	14
Chelsea	51.765	15

Team	Predicted season points coefficient	Rank
Burnley	45.624	16
Leicester City	43.769	17
Tottenham	34.741	18
Norwich City	33.981	19
Crystal Palace	29.901	20

4. Conclusion

In this article, we apply ANOVA and DA to analyze the technical data of football players. From the perspective of descriptive statistical indicators, the degree of technical and tactical variables that affect the outcome of the team's win, loss and draw, the goal and shooting variables are greater than the offensive organization-related variables and the defense-related variables. According to the one-way analysis of variance and discriminant analysis of the technical and tactical indicators of the team's win, loss and draw in the 21-22 season of the Premier League. And the BP neural network prediction model screened out after discriminant analysis can be used as a reference for the skills and tactics of high-level football, and put forward opinions, so as to improve the performance for training. Whether it is one-way ANOVA or discriminant analysis and BP neural network prediction, the conclusion is that attack is the most important part of the current Premier League, whether you are a team competing for the title or a team that wants to succeed in relegation, attack is the key to survival in the Premier League. At the spiritual level, it is necessary to cultivate the awareness of offensive and defensive conversion of young people from an early age, which is often improved from the daily game, the faster the rhythm of offensive and defensive conversion in the daily game, and the awareness can only be maintained in the subsequent games, which cannot be trained by simple training, and only in high-intensity games can we see their own technical level and awareness, so in the training of youth football, the intensity of the game is very important.

Abbreviations

ANOVA: An Alys of Variance
 BP: Back Propagation
 DA: Discriminant Analysis
 SCDFC: Standardized Canonical Discriminant Function Coefficients

Author Contributions

Zhaojun Li is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Gardasevic, J.; Bjelica, D. Body composition differences between football players of the three top football clubs. *International Journal of Morphology*, 2020, 38(1), 153-158. <https://doi.org/10.4067/S0717-95022020000100153>
- [2] Yildizparlak, A. An application of contest success functions for draws on European soccer. *Journal of Sports Economics*, 2018, 19(8), 1191-1212. <https://doi.org/10.1177/1527002517716973>
- [3] Xie, X. Visual analytics for football tactics. Master Thesis. Zhejiang University, Zhejiang, 2020. <https://doi.org/10.27461/d.cnki.gzjdx.2020.003589>
- [4] Liu, H.; Peng, Z. Big data analysis of football technical and tactical performance: Based on generalized linear model and data series inference method. *Journal of Physical Education*, 2017, 24, 109-114. <https://doi.org/10.16237/j.cnki.cn44-1404/g8.2017.02.01>
- [5] Hu, Z. A study on the use of techniques and tactics in the 2008 Euro Cup. Hunan University, Hunan, 2008. <https://doi.org/10.19715/j.tiyukexueyanjiu.2023.06.001>
- [6] Liu, H. Analysis of the attacking characteristics of Leicester City in the 2015-2016 Premier League season. Beijing Sport University, Beijing, 2018. <https://doi.org/10.27315/d.cnki.gstyx.2023.000182>
- [7] Chen, Z. Analysis of the offensive tactical characteristics of the Liverpool team in the 2019-2020 Premier League season. Guangxi University for Nationalities, Guangxi, 2021. <https://doi.org/10.27035/d.cnki.ggxmc.2021.000212>
- [8] Noughabi, R.A.; Mohammadpour, A. Multivariate regression with stable errors using order statistics. *Fluctuation and Noise Letters*, 2001, 21(03), 1348-1363. <https://doi.org/10.1142/S0219477522500298>
- [9] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6), 1137-1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- [10] Chen, X.; Zhang, M.; Yang, S. A ranging model based on BP neural network. *Intelligent Automation and Soft Computing*, 2016, 22(2), 325-329. <https://doi.org/10.1080/10798587.2015.109548>
- [11] Cui, K.; Jing, X. Research on prediction model of geotechnical parameters based on BP neural network. *Neural Computing & Applications*, 31(12), 8205-8215. <https://doi.org/10.1007/s00521-018-3902-6>
- [12] Wang, X.; Wu, Y.; Gui, Y. Gray BP neural network based on prediction of rice protein interaction network. *Cluster Computing- The Journal of Networks Software Tools and Applications*, 2019, 22(2), 4165-4171. <https://doi.org/10.1007/s10586-017-1663-0>

- [13] Li, J.; Zhao, D.; Chen, Y. A link prediction method for heterogeneous networks based on BP neural network. *Physica a Statistical Mechanics and its Applications*, 2018, 495, 1-17. <https://doi.org/10.1016/j.physa.2017.12.018>
- [14] Kim, M. Application of functional ANOVA and functional MANOVA. *Korean Journal of Applied Statistics*, 2022, 35(5), 579-591. <https://doi.org/10.5351/KJAS.2022.35.5.579>
- [15] Wang, J.; Liu, Z.; Zhang, M. Robust sparse manifold discriminant analysis. *Multimedia Tools and Applications*, 2022, 81(15), 20781-20796. <https://doi.org/10.1007/s11042-022-12708-3>
- [16] Mai, Q. and Zou, H. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 2015, 135, 175-188. <https://doi.org/10.1016/j.jmva.2014.12.009>