Prediction of Wordle Puzzles Based on Neural Networks and K-Means Improved Models

Zihong Qiu *, Chengbin Duan

Department of Automation, China University of Mining and Technology, Xuzhou, China, 221116

* Corresponding Author Email: 05204402@cumt.edu.cn

Abstract. Wordle is an online word puzzle game that gained viral popularity in January 2022. The goal is to guess a hidden five letter word. In order to study the distribution of the reported results and the difficulty of the words, based on BP neural network theory, combined with the advantages of Genetic Algorithm in Stability and speed, a neural network prediction model is constructed and the model can be used to predict the percentage distribution of the number of player guesses for a given word on a given number of days. In addition, a Canopy-Kmeans-based word difficulty classification model was further developed. It can be combined with the previous neural network model to determine whether the difficulty of guessing a given word is easy, normal or hard.

Keywords: Neural Network, Genetic Algorithm, Canopy-Kmeans.

1. Introduction

WORDLE is a phenomenal product that is widely used in the fields of entertainment, education and research. Studying the daily data of WORDLE puzzles is important for the long-term development of this product [1-2]. Using traditional neural networks for prediction of relevant data is prone to problems such as getting trapped in local optima and taking too long [3]. The traditional K-means algorithm can also lead to poor clustering due to improper selection of initial clustering centers and k-values [4]. Therefore, this paper tries to optimize and improve the traditional algorithm with innovative methods.

2. The basic funamental of modles

2.1. Building a BP neural network model

BP neural network is a multilayer feed-forward artificial neural network (ANN), which emerged thanks to the BP algorithm and is based on the gradient descent method. The loss function is generally the root mean square error, and the goal of training is to minimize the root mean square error. The training process is the weights and thresholds of the network are continuously corrected by the mean square error of each training until the mean square error meets the requirements (model convergence) or the specified number of training sessions is reached. It is one of the most important tools of machine learning, which has the advantages of self-adaption, self-learning, nonlinear mapping and fault tolerance [5].

The BP neural network has no requirements on the distribution characteristics of the data and is not sensitive to problems such as multicollinearity and outliers among the data and is very suitable for the regression of nonlinear relationships. Therefore, we take the feature of the word as the independent variable and the percentage of the number of times answered as the dependent variable to establish a BP neural network model. The network consists of an input layer, a hidden layer, and an output layer, and each layer consists of one or more neurons. The performance of the BP neural network model depends largely on the model parameters, so the choice of parameters is very important. The number of neurons in the input layer and output layer can be determined according to the input characteristics and the target; the number of hidden layers is set according to empirical methods. Since the features and labels in the data are not simply linearly separable, the activation function from the input layer to the hidden layer is selected as sigmoid; for regression problems, the

activation function from the hidden layer to the output layer is selected as purelin. The algorithm flow of BP neural network is shown in Figure 1.

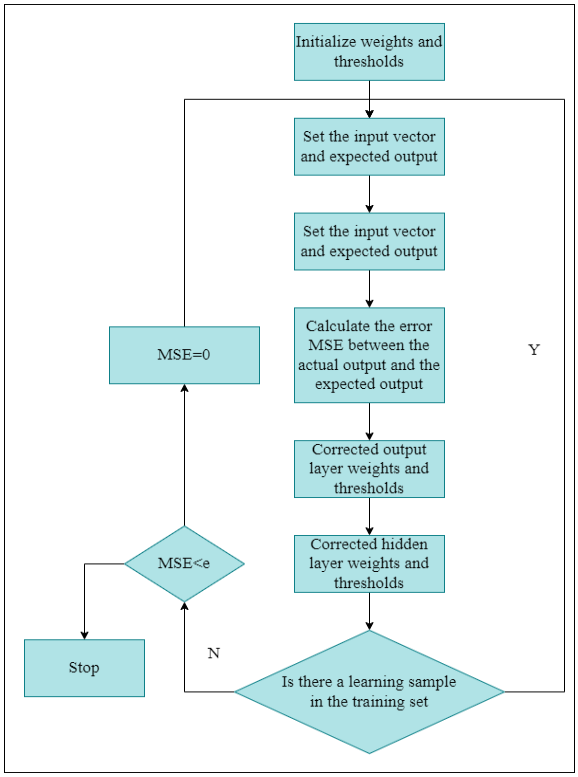

Figure 1. Algorithm flowchart of BP neural network

After continuous debugging, we finally determined more suitable parameters and established the BP neural network model shown in Figure 2.

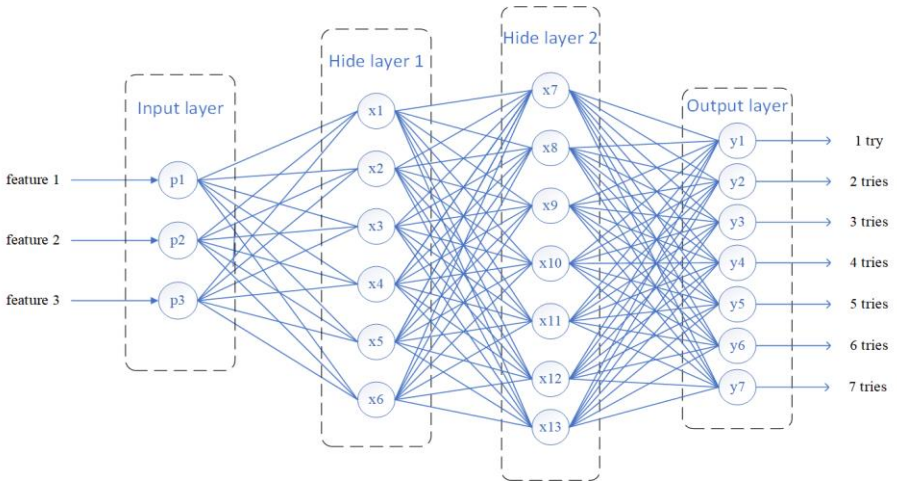

Figure 2. BP neural network

2.2. Principles of genetic algorithms for optimization

BP neural network has many advantages, but the initial weights and thresholds of the network are randomly generated, and the initial weights and thresholds of BP neural network largely affect the effect of data fitting, and unsuitable initial weights often increase the number of training, and the model falls into local optimum. The traditional method is empirical or randomly generated initial values, so we consider the genetic algorithm to perform the initial weights and thresholds of the BP neural network to find the optimum so that the model can converge as soon as possible and reduce the possibility of falling into the local optimum. The flowchart of GA-BP algorithm is shown in Figure 3.

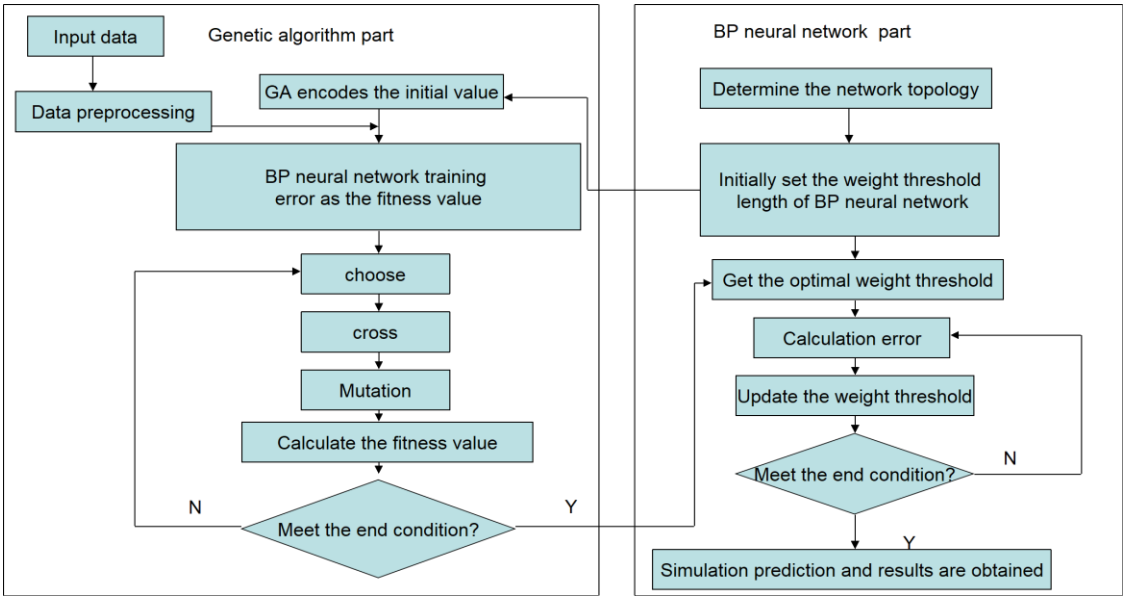

Figure 3. GA-BP algorithm flow chart

2.3. Canopy algorithm to determine K values and initial values

K-means algorithm is the most classical unsupervised partitioning algorithm, which has the features of simple algorithm idea, fast convergence, and high efficiency in processing large-scale data sets, and is widely used in business, e-commerce, big data mining and other fields [6]. To further differentiate the difficulty of a given word, we clustered the words in seven dimensions using the percentage of the number of players who solved the puzzle in multiple guesses given in the question. Due to the large amount of data, k-values and initial clustering centers in the K-MEANS algorithm are not easy to choose, so in this paper, we use the Canopy algorithm, which has lower accuracy but faster execution, to cluster first, initially determine the difficulty level of the words, and derive the clustering centers for each word difficulty category, and then use these clustering centers as the initial clustering centers of the K-means algorithm to obtain the difficulty level classification using the K-means algorithm [7-8]. Finally, combining with the GA-BP model established in the previous question we can get the difficulty classification model for a given word.

All the relevant percentages of the pre-processed word list in the previous section are listed, and a set of thresholds T1 and T2(T1>T2) is obtained by repeated debugging, so that the final number of formed canopies is moderate and the number of words in a canopy is moderate. After that, the following operation is performed as shown in Figure 4.

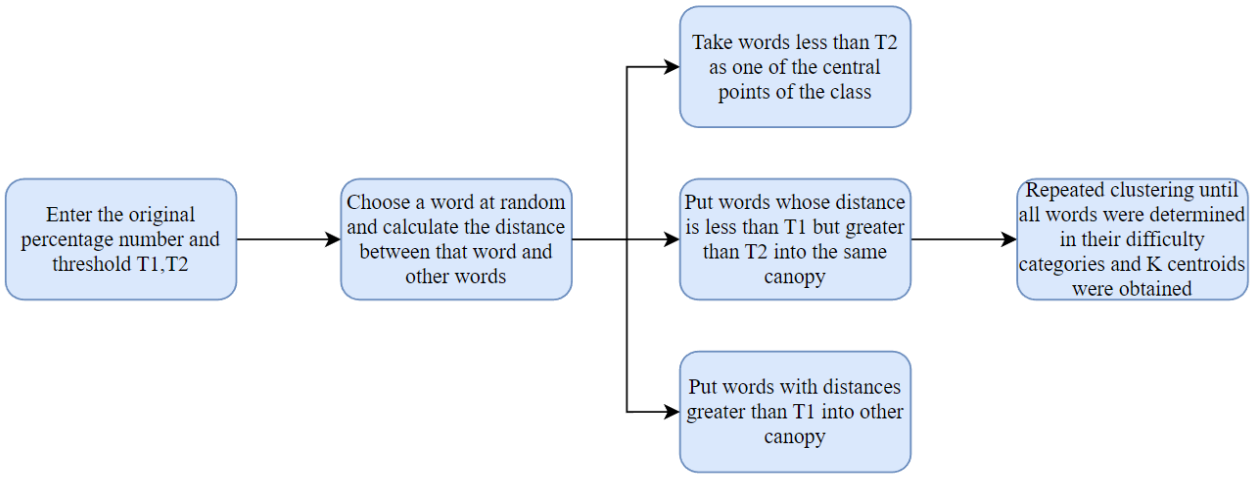

Figure 4. Canopy's flow chart

3. Results

3.1. The establishment of GA-BP model

The BP neural network described above was built using matlab, as shown in Figure 5.

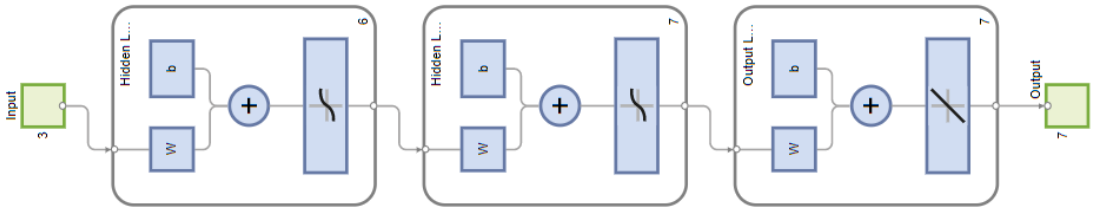

Figure 5. BP neural network

The learning rate is set to 0.01 and the number of iterations is set to 1000. 80% of the data is used as the training set and 20% as the test set to train the model, and after adjustment, the model is able to converge and has a good fitting effect.

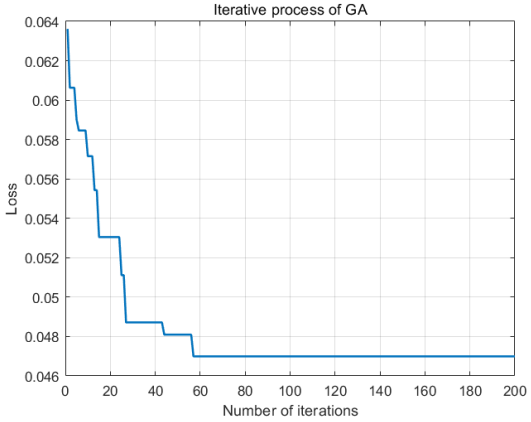

Figure 6. Iterative process of GA

The iterative process of GA search is shown in Figure 6. From the figure, it can be seen that the loss function converges approximately to 0 after about 60 iterations. The fitting effect of the model is shown in Figure 7, and the fitting effect on the test and validation sets is acceptable.

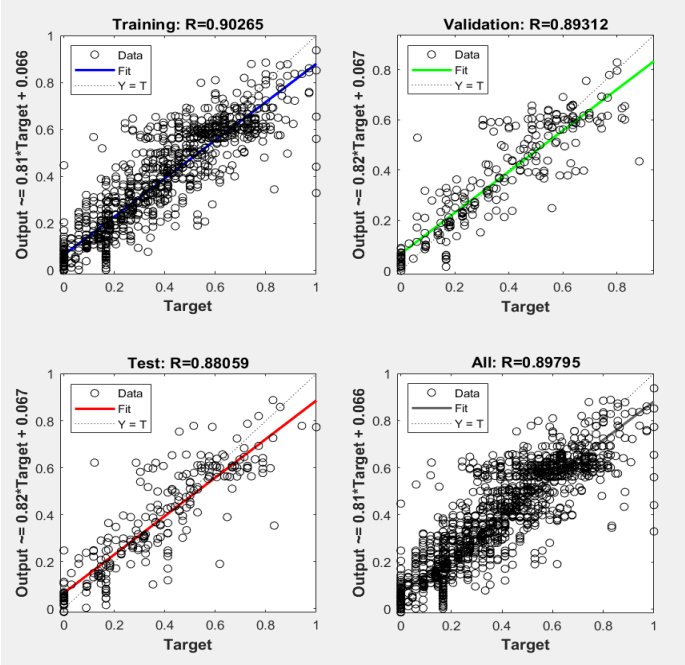

Figure 7. Fitting effect

According to the model, the word "EERIE" is predicted, and the proportion of the number of players guessing each time in the difficult mode is obtained, as shown in Table 1. It can be seen from Figure 7 that the accuracy of the GA-BP model is high, and the training effect is good, so we have full confidence in the model.

Table.1. Add caption.

Word	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 tries
EEIRE	0.039	2.838	16.13	32.31	29.319	15.55	3.8314

3.2. Analysis of Uncertainty Factors

The uncertainty factors of the model and forecast in this question mainly come from the uncertainty of the data and the uncertainty of the model.

The uncertainty of the data is mainly due to the possibility of statistical incompleteness of the second type of error in the statistics on Twitter and the misestimation of the common degree of words, resulting in large fluctuations in the data. This problem is unavoidable, and the government can only take measures to improve the stability of data collection or improve the accuracy of measurement indicators to include various objective influencing factors. We also performed data preprocessing during modeling to remove outliers and try our best to improve the accuracy of the model.

The uncertainty of the model is mainly due to the fact that the estimation of the input data by the bp neural network itself may be inaccurate due to poor training, insufficient number of days of game operation, etc., and has nothing to do with a single word attribute. So cognitive uncertainty measures the uncertainty of the model parameters estimated during the bp neural network training process itself. When we actually build the model, we adjust the parameters repeatedly and find the initial weight and threshold in advance through the genetic algorithm to alleviate it [9].

3.3. Canopy-Kmeans model for difficulty classification

Firstly, we need to perform coarse clustering. To present the effect of the initial clustering, we conducted a principal component analysis on the data of the percentage of people who tried once to seven times and extracted two principal components measuring difficulty to draw a clustered scatter plot, with the final result in Figure 8.

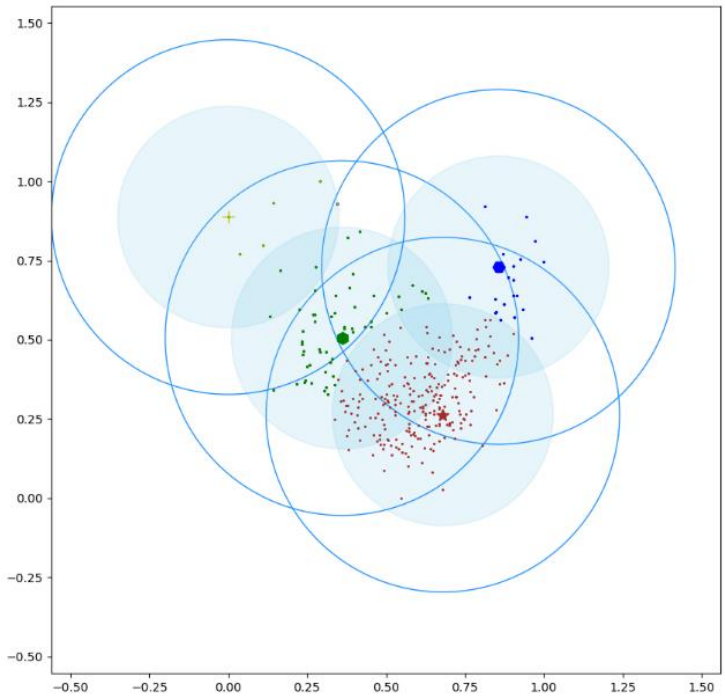

Figure 8. Initial Clustering Center

Next, we input the above results into the K-means model. The elbow rule can determine the final best way to classify the difficulty [10]. The horizontal coordinate of the figure below is the class of word difficulty, and the vertical coordinate is the loss function of K-means clustering, defined as the sum of the squared distances of all words to the center of the class, which is also the sum of the squared errors. From the Figure 9, it can be seen that the slope of the curve tends to be flat when choosing to classify words into three levels: Easy, Normal, and Hard, and better clustering results can be obtained.

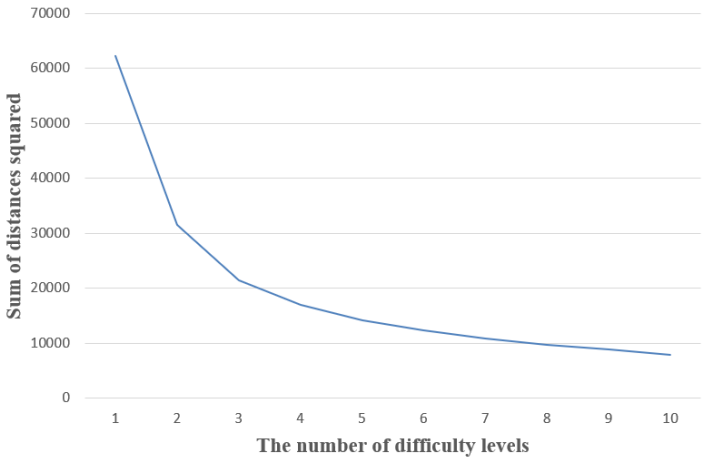

Figure 9. Elbow rule

Then variance analysis was performed for each percentage stratum to confirm that all seven data sets would be significantly different when the difficulty of the words was determined, and finally the cluster summary Figure 10 and cluster scatter Figure 11 and cluster centers were obtained as in Table 2.

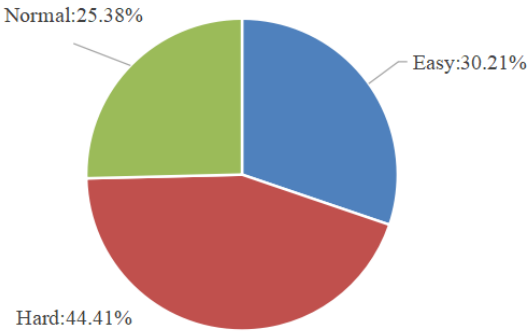

Figure 10. Clustering Summary Chart

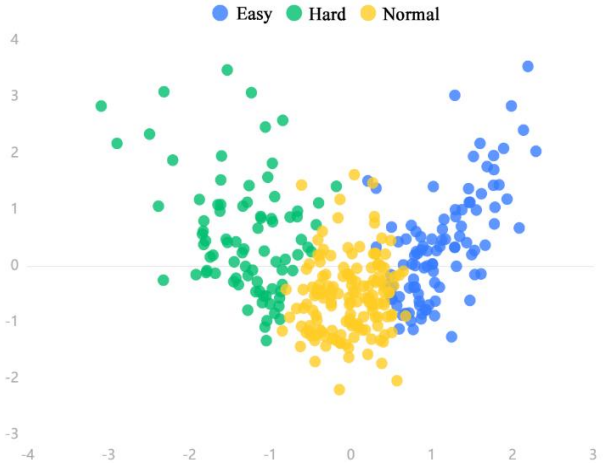

Figure 11. The final clustered scatter plot

Table.2. Clustering Center

Difficulty	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 tries
Easy	0.89	9.77	31.56	33.54	17.2	6.03	1.03
Normal	0.279	4.592	21.73	35.605	25.08	10.71	1.864
Hard	0.286	2.905	13.12	27.179	29.56	20.44	6.452

Based on the previous prediction results for the word 'ERRIE', the Euclidean distances d for a given word and different difficulty category centers were calculated as in Table 3.The difficulty of the given word 'EERIE' can be determined as 'Hard', mainly due to its word properties, i.e. it is not common in everyday life and has three repeated letters.

Table.3. Distance of 'ERRIE' from the cluster center

Difficulty	Distance
Easy	23.09956811
Normal	9.517433469
Hard	8.148271131

4. Conclusions

During the study of Wordle, there was no simple linear relationship between the attributes of these words since they do not exist. Therefore, in this paper, a multiple-input, multiple-output regression model is built using BP neural network with the commonness of words, the frequency of initial letters and the number of repeated letters of words as features, and a genetic algorithm is used to find their initial weights and thresholds in advance before the neural network is built. In a subsequent study of word difficulty, this paper performed a cluster analysis of the data on the percentage of successful player challenges over a year. We obtained K cluster centers after the initial clustering by the Canopy algorithm, and then performed the exact clustering by the K-means algorithm to classify all words into Easy, Normal, and Hard difficulties. Then we combined the prediction of the relevant percentage of the word EERIE in the previous model to get the difficulty of the word EERIE as Hard. The experimental results show that the model has good prediction and robustness and has some practical application value.

References

[1] Anderson B J, Meyer J G. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning [J]. arXiv preprint arXiv:2202.00557, 2022.

[2] Brown K A. MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research [J]. npj Computational Materials, 2022, 8(1): 97.

[3] Han J X, Ma M Y, Wang K. Product modeling design based on genetic algorithm and BP neural network [J]. Neural Computing and Applications, 2021, 33: 4111-4117.

[4] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation [J]. Electronics, 2020, 9(8): 1295.

[5] Cui Y, Liu H, Wang Q, et al. Investigation on the ignition delay prediction model of multi-component surrogates based on back propagation (BP) neural network [J]. Combustion and Flame, 2022, 237: 111852.

[6] FANG Shiqiao, HU Peiling, HUANG Yingying, ZHANG Xin. Optimization and Application of K-means Algorithm [J]. Modern Information Technology,2023,7(06):111-115.

[7] Zhang G, Zhang C, Zhang H. Improved K-means algorithm based on density Canopy [J]. Knowledge-based systems, 2018, 145: 289-297.

[8] Kumar A, Ingle Y S, Pande A, et al. Canopy clustering: a review on pre-clustering approach to k-means clustering[J]. Int. J. Innov. Adv. Comput. Sci. (IJIACS), 2014, 3(5): 22-29.

[9] BAI Guozheng. Application of improved genetic neural network algorithm in automatic book borrowing system fault detection [J]. Automation & Instrumentation,2023(05):65-69.

[10] Humaira H, Rasyidah R. Determining the appropiate cluster number using elbow method for k-means algorithm [C]//Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia. 2020.