Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230823

Predictive Modeling and Analysis of Wordle Player Engagement: Insights and Recommendations for Future Operations

Changze WU[1], Yuyang LI[2], Shanling LIN[3], Xuhui JIANG[4]
School of Advanced Manufacturing, Fuzhou University Quanzhou, China

Abstract—To further improve the stickiness of puzzle game players, we developed a prediction model based on the BP neural network. Firstly, we preprocessed the existing data to remove some useless information. Then, we introduced the WOA algorithm to optimize the BP neural network model and improve its prediction accuracy. To predict the distribution percentage of attempts to guess a given word in the future, we improved the BP neural network using the multi-input multi-output CIWOA algorithm (Cubic map improvement whale optimization algorithm). The experimental results showed that the WOA-BP prediction model predicted player data for March 1, 2023, to be between 20,000 and 25,000, while the actual data was 19,655. Our prediction model had a small margin of error, indicating good prediction accuracy. Based on this model, we provided reasonable adjustment suggestions for the game company's later operations.

Keywords-component: BP Neural Network; WOA; CIWOA

1. Introduction

We take the game Wordle as an example. Wordle is a popular daily puzzle currently offered by The New York Times. Players have to guess a five-letter word six or fewer times, and each guess gets feedback. Each guess must be a real word in English. Guesses that are not recognized as words by the contest are not allowed. The Wordle instructions on The Times website state that the color of the squares will change after you submit your text. Yellow indicates that the letter in the cell is in the word, but it is in the wrong place. Green means that the letters in the post are in the word and in the correct position. Gray indicates that the letters in the cell are not included in the word at all. With many users reporting their scores on Twitter, we have made a file of daily results from January 7, 2022 to December 31, 2022. The file lists dates, game numbers, daily words, score reports, player count, and percentage of players unable to solve the

[1] Changze WU, School of Advanced Manufacturing, Fuzhou University Quanzhou; e-mail: 252005222@fzu.edu.cn
[2] Yuyang LI, School of Advanced Manufacturing, Fuzhou University Quanzhou; e-mail: 852203324@fzu.edu.cn
[3] Corresponding author: Shanling LIN, School of Advanced Manufacturing, Fuzhou University Quanzhou; e-mail: sllin@fzu.edu.cn
[4] Xuhui JIANG, School of Advanced Manufacturing, Fuzhou University Quanzhou; e-mail: 251902138@fzu.edu.cn

puzzle on their first to seventh attempt. Based on the above data, we use document data to build a mathematical model for better predicting and analyzing the daily reports in the future period, and through this model and the results, improve the player stickiness of the Wordle game.

2. Data Acquisition and Preprocess

2.1 Data Description

MCM report on January 7,2022 to December 31,2022 report to the real value after the blue line chart, we can intuitively see the players on twitter report the number of the results of the Wordle game first from January 7, 80630, and then experienced a month of a rapid rise phase, in February reached the user report the number of the highest value, about 361908. After the highest value, it experienced several violent fluctuations in the short term, followed by a fluctuating decline. Three months later, the number of user reports dropped from the highest 361908 to 44212, and then entered a long and slow decline moment.

The data of the specific MCM report is shown in Figure 1.

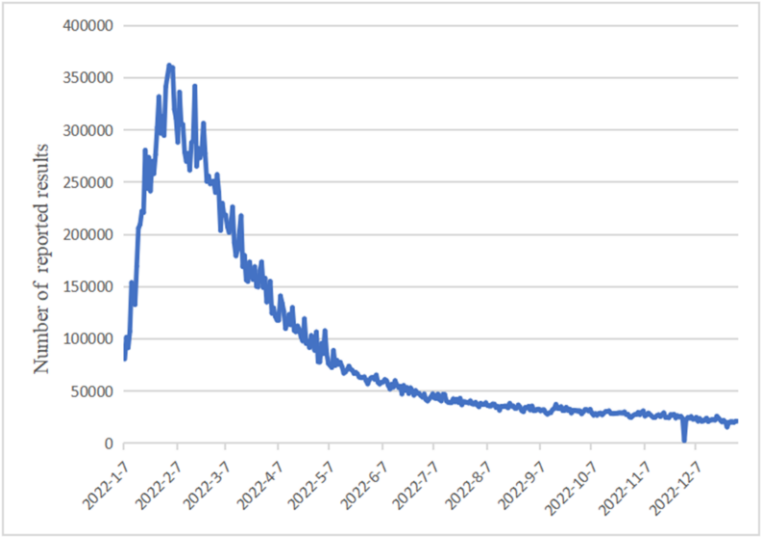

Figure 1: Report quantitative trends to fit the prediction

2.2 Prepare Data

Load report files into MATLAB and convert them into an acceptable format for the BP neural network. Specifically, we need to convert the date, the game number, the words of the day into numerical or categorical variables, and convert the number of people reporting scores on the day into the target variable.We digitally encode the individual letters of each word: in the order of the 26 letters, such as a replaced by 1, b with 2, y with 25, and z with 26.

2.3 Data Pre-Processing

Because the data is obtained from Twitter, there is some anomalous data in the data, and we need to remove these anomalous data values so as not to interfere with the predictive model. Our method is to first remove words that are not 5 letters in the word, because we are studying 5 letter words. Second, sum the number of guessing attempts for each word separately, and delete words with a sum greater than 100% and less than 98%.

2.4 Separation of Data

Dividing the data set into training set and test set. The training set is mainly used to train the neural network, and the test set is mainly used to verify the prediction accuracy of the model.

3. Construction and Analysis of Neural Network Prediction Models

3.1 Establishment of BP Neural Network

BP neural network is a widely used feedforward neural network applied in pattern recognition, classification, and prediction[1]. In Wordle prediction, we use BP neural network to predict the number of players and distributed percentage of game guessing attempts. We represent each word as a sequence of encoded numbers and split the data into training and testing. Then, we define the network structure with input layer for encoded word sequence, output layer for category labels, and hidden layer with appropriate activation functions and nodes.

We initialize connection weights, perform forward propagation and backpropagation algorithms to update node error and adjust connection weights. Optimization algorithms like batch gradient descent and momentum improve the training speed and accuracy of the network. To evaluate network performance, we test its generalization ability and accuracy on an independent dataset. However, in practice, using only the BP neural network model to predict future results is not accurate enough. So, we introduce two optimization algorithms to improve it further.

3.2 WOA Optimization Algorithm

The WOA algorithm is an abbreviation for the Whale Optimization Algorithm, which is a swarm intelligence-based optimization algorithm. The algorithm imitates the hunting behavior of whales, dividing individuals into leader and follower whales, guiding the search direction of follower whales through the search behavior of leader whales to achieve global optimal solution search[2]. In this study, we used the WOA algorithm to optimize the BP neural network model to predict the number of people playing on March 1,2023(We conducted research and reached a conclusion before March 1st, 2023).

When using the WOA algorithm to optimize the BP neural network model, we first take the weight matrix of the BP neural network as the search space of the WOA algorithm and use the WOA algorithm to search for the optimal solution of the weight matrix[3]. In the search process of the WOA algorithm, we use three parameters to control the search direction: the position of the whale individual (i.e., weight matrix), the

search step size, and the search direction. By iteratively searching, we can gradually optimize the weight matrix of the BP neural network to improve the prediction accuracy and the generalization ability of the model.

When using the WOA algorithm to optimize the BP neural network model, we need to set some parameters, such as the population size, the maximum number of iterations, the search space range, and initialize the WOA algorithm. Through multiple experiments and parameter adjustments, we finally obtained an optimized BP neural network model. The network parameter Settings are shown in Table 1.

Table 1: Network parameter configuration

Network parameter configuration	
frequency of training	1000
learning rate	0.01
Training objective minimum error	0.00001
Display frequency	25
Momentum factor	0.01
Minimum performance gradient	1e-6
Maximum number of failures	6

3.3 CIWOA Optimization Algorithm

Earlier, it was mentioned that due to the problem of local optima, the BP neural network often gets stuck, and to address this issue, an improved Whale Optimization Algorithm (WOA) called Cubic map improve whale optimization algorithm (CIWOA) was proposed. This algorithm uses the Cubic map chaotic map to initialize the positions of the whales, assuming that the current best candidate solution is the target prey or close to the optimal solution, and then continuously optimizes the initial weights and thresholds to quickly and effectively solve the problem of BP neural network getting stuck in local optima[4].

Therefore, we use this algorithm to predict the distribution of player guesses and attempts for a future date, i.e., predicting the related percentages of (1,2,3,4,5,6,X) for a future date.On March 1,2023, the word "EERIT" was used as the research object.

As a chaotic map, CIWOA has been proven to be very effective in generating pseudo-random numbers for use in optimization algorithms[5]. Therefore, in the WOA, we update the positions of the whales using Eq. (1).

$$\begin{aligned} &\text{Xnew}=\text{Xrand}+AD \\ &D = abs(CX[i] - \text{Xrand}) \end{aligned} \tag{1}$$

Where Xrand is a randomly generated position, A is the search agent, C is a random number between 0 and 1, and D is the distance between the current position of the whale and the randomly generated position.

To incorporate the Cubic map into the WOA, we take the following steps:

1) Generate a sequence of pseudo-random numbers using the Cubic map.
2) Use these numbers to modify the value of the search agent A.
3) Update the position of the whales using the modified value of A.

4. Experimental Results and Analysis

4.1 Research Results Display

4.1.1. A range of the Number of Visitors Predicted One Day

As shown in Figure 2, on this day, the official Wordle game Twitter account announced that the number of participants was 19655. As shown in Figure 3, the WOA-BP curve indicates that the predicted number of participants was roughly around 20000 to 25000. This indicates that the model we established is accurate to some extent.

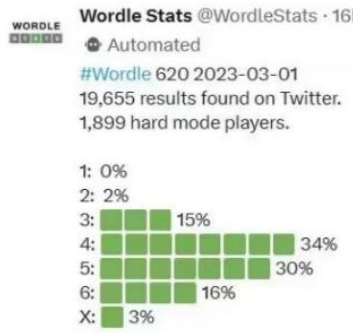

Figure 2: The official number of visitors is released on March 1,2023

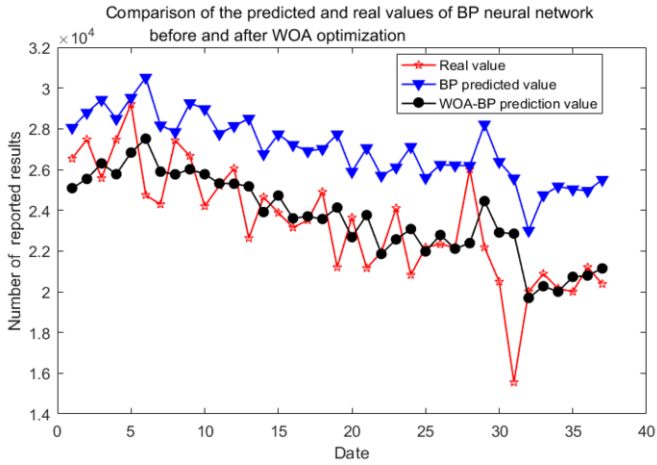

Figure 3: Comparison of the predicted and real values of BP neural network before and after WOA optimization

4.1.2. Predict the Distribution Percentage of Guess Attempts for a Given Word on a Future Day

After the prediction of our model, the percentage of the word EERIE on March 1,2023 is shown in figure 4. This result has a strong fit with the official data released in Figure 2, indicating that this improved whale algorithm has obvious advantages in dealing with multi-input multi-output prediction problems.

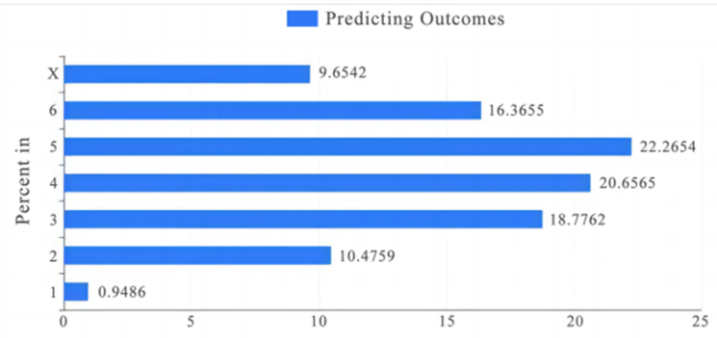

Figure 4: Percentage of the predicted number of attempts

4.2 Error Index Analysis

As shown in Figure 5, the inner circle represents the standard BP neural network, while the outer circle represents the BP neural network optimized by WOA (CIWOA). It can be seen that our prediction data is more accurate. We optimized the weight and threshold parameters of the BP neural network and used the mean square error between the training set and the entire test set as the fitness function. The smaller the fitness function value, the more accurate the balance model training and the better the prediction accuracy.

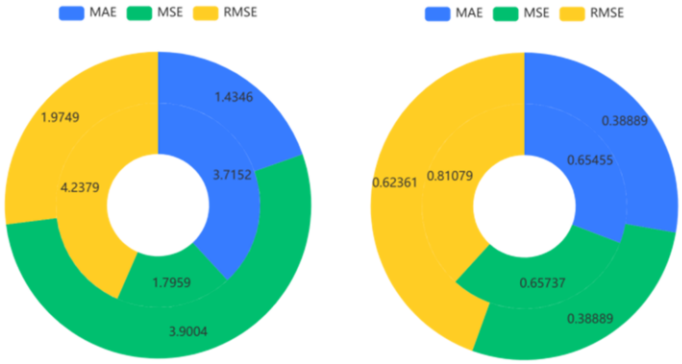

Figure 5: Comparison of error indicators

(Left: standard BP neural network and WOA-BP neural network

Right: standard BP neural network and CIWOA-BP neural network)

In addition, we calculated that for the prediction interval problem, the mean square error of the training set is 0.0054074 when the number of hidden layer nodes is 9. Based on analysis, we found that the optimal number of hidden layer nodes is also 9, with a corresponding mean square error of 0.0054074. For the distribution percentage problem, the mean square error of the training set is 0.090467 when the number of hidden layer nodes is 11, and based on analysis, we found that the optimal number of hidden layer nodes is also 11, with a corresponding mean square error of 0.090467[6].

As shown in Figures 6 to 12, we compare and analyze the predicted values and actual values of the seven attempts.

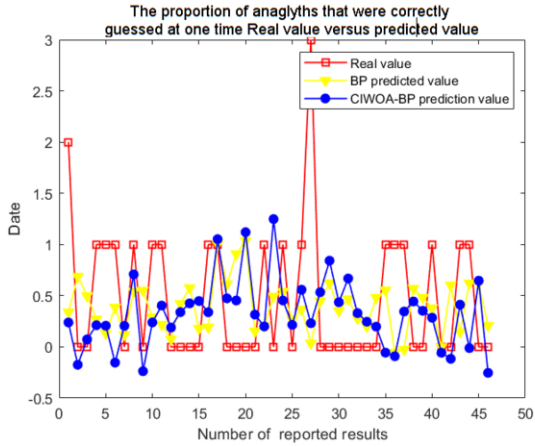

Figure 6: Degree of fit of the first attempt

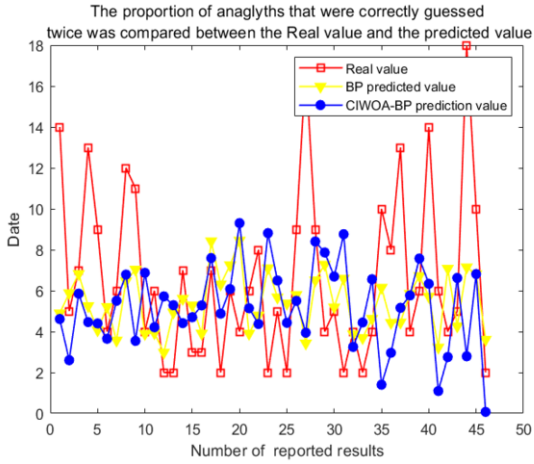

Figure 7: Degree of fit of the Second attempt

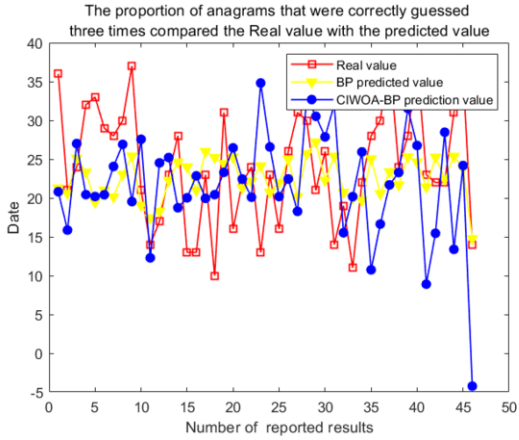

Figure 8: Degree of fit of the Third attempt

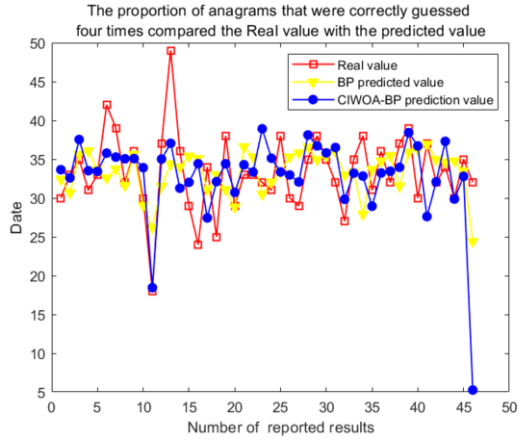

Figure 9: Degree of fit of the Fourth attempt

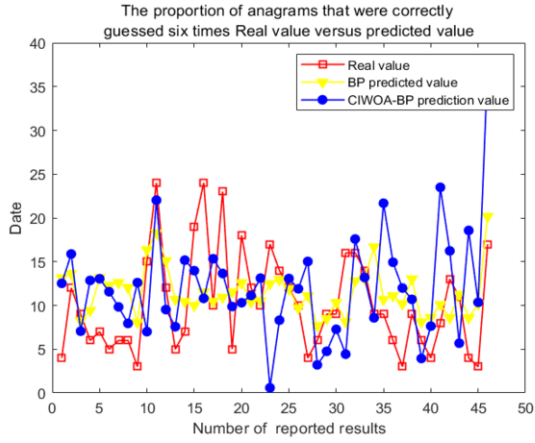

Figure 10: Degree of fit of the fifth attempt

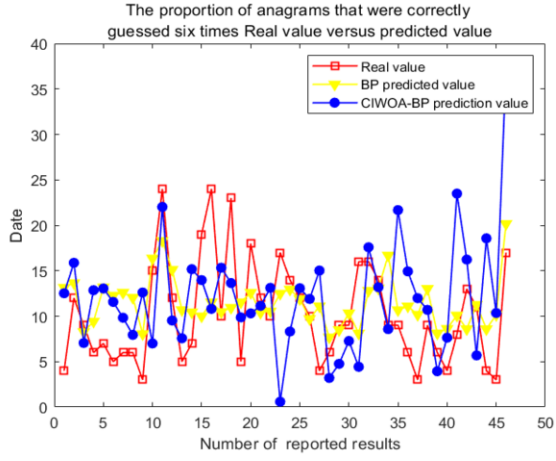

Figure 11: Degree of fit of the sixth attempt

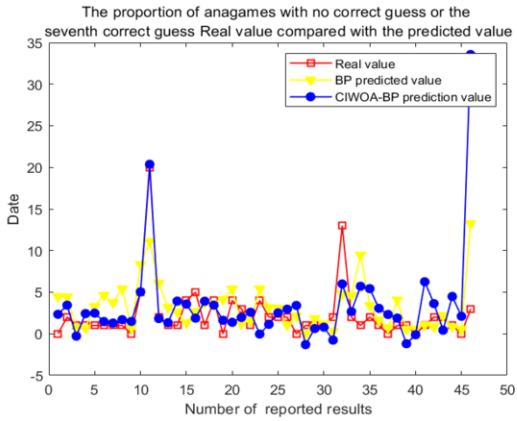

Figure 12: Degree of fit of the seventh attempt

We compare the accuracy of BP neural network model and CIWOA-BP neural network algorithm in solving multi-input multi-output problems. It is found that the adaptive ability of CIWOA-BP neural network algorithm to data is significantly better than that of BP neural network algorithm. This is also an important reason why CIWOA-BP can better solve the multi-input multi-output problem. However, some of the data are still far from the actual value. After repeated experiments, we inferred that the reason for this phenomenon was that the data volume of the original data set was not large enough, and it was easy to produce overfitting in this intelligent algorithm. We just need to get enough data sets to train it, and we can get more accurate prediction values.

5. Conclusions

Using heuristic algorithms that simulate biological behavior to analyze data can provide more sensitive and real-time feedback on changes in the data, and serve as a powerful basis for predicting future data changes, which can help optimize various policies and measures. In solving the single-input single-output problem, by comparing the effectiveness of the BP neural network algorithm and the WOA-BP neural network algorithm in fitting and predicting the number of players, it was found that the WOA-BP neural network had significantly better performance than the BP neural network model. The study also compared the accuracy of the BP neural network model and the CIWOA-BP neural network algorithm in solving the multi-input multi-output problem, and found that the CIWOA-BP neural network algorithm had better adaptability to data than the BP neural network algorithm, which is an important reason why it can better solve the multi-input multi-output problem. The study also found that both WOA-BP and CIWOA-BP have high requirements for the size of the training set and exhibit a certain degree of "lag", leading to conservative predictions for data that may produce large fluctuations. Therefore, future research should appropriately incorporate judgments on data trend changes before training begins, and further process the predicted results. How to further optimize the algorithm to reduce the "lag" and improve the prediction accuracy is an important research direction in the future. We also hope that this algorithm can help game companies to improve the engagement of puzzle game players.

References

[1] Zhao Fang, Li Weide. A Combined Model Based on Feature Selection and WOA for PM2.5 Concentration Forecasting[J]. Atmosphere, 2019, 10(4): 223.
[2] Hussein Alahmer, Ali Alahmer, Razan Alkhazaleh, Mohammad Alrbai. Exhaust Emission Reduction of a SI Engine Using Acetone–Gasoline Fuel Blends: Modeling, Prediction, and Whale Optimization Algorithm[J]. Energy Reports, 2023, 9(S1): 77-86.
[3] Gaganpreet Kaur, Sankalap Arora. Chaotic Whale Optimization Algorithm[J]. Journal of Computational Design and Engineering, 2018, 5(3): 275-284.
[4] Chen Ge, Ji Jianqiang, Huang Chaofeng. Student Classroom behavior Recognition based on OpenPose and Deep Learning[C]. 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022.
[5] Xiao Rongge, Jin Shuaishuai. Corrosion Rate Prediction of Submarine Pipeline Based on WOA-BP Algorithm[J]. Marine Science, 2022, 46 (6): 116-123.