Automation and Machine Learning (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2024.050116
ISSN 2516-5003 Vol. 5 Num. 1

Research on Word Prediction Based on BP Neural Network

Yanqi Zhang, Xinhong Liu*, Xuanyi Jin, Ruisheng Zhang, Shuxia Wang

Beijing Institute of Petrochemical Technology, Beijing, 102617, China
**Corresponding author: liuxinhong@bipt.edu.cn*

Keywords: BP neural network, Time Series, K-Means clustering

Abstract: Today, the Wordle game is exploding all over the world. We study the mechanics of this game, and analyze the game. Smoothness and white noise tests were conducted on the data, and a time prediction model was established for prediction with a high degree of fit. Through analysis of variance, it was found that the properties of words have no effect on numbers in hard mode. BP neural network prediction model was established to predict. After that, the real values were compared with the predicted values for analysis. It is calculated that there is more than 85% confidence in the neural network prediction model. In the end, use K-means clustering algorithm to classify the difficulty of words into simple difficulty, moderate difficulty, and difficulty categories. The difficulty of the word EERIE belongs to the medium difficulty category. Finally, we summarized our results and making recommendations.

1. Introduction

Wordle is the first game to set the world on fire in 2022. Currently, Wordle is a popular daily puzzle offered by The New York Times. Players attempt to solve the puzzle by guessing a five-letter word no more than six times, with each guess receiving feedback. For this version, each guess must be a real English word. No guesses are allowed that are not recognized by the game as the result of a word. Wordle is growing in popularity, and versions of the game are now available in over 60 languages. The Wordle instructions on the New York Times website state that the color of the tile will change after you submit the text. A yellow flat tile indicates that the letter in that tile is in the word, but it is in the wrong place. A green sticker indicates that the letter in that sticker is in the word and in the correct location. The game has a regular mode and a hard mode, and players can play in either mode. Wordle's hard mode makes the game more difficult for the player because once the player finds the correct letters in a word (tiling is yellow or green), those letters must be used in subsequent guesses. Many users reported their scores on Twitter, and MCM collected this data to form a results file, which was analyzed to better understand the game's mechanics.

2. Modeling and analysis

In question C of the 2023 American College Student Mathematical Modeling Competition, there is a batch of relevant data on the Wordle game [1].

For the outliers of the data, they need to be removed first. In the data given in the question, the total proportion of nymph guessed reaches 126%, which is not in line with common sense; for the word clen, tash, are not in line with the rules of the Wordle game given in the question; for the word study, Number of reported results is much smaller than the other data, and there is a large deviation. The basic statistics after data pre-processing are shown in Table 1.

Table 1: Basic statistics

Item	value
Median	45812
Average value	91095
Maximum value	361908
Minimum value	15554

2.1. Time Series[2]

A time series is a set of data arranged at a certain time interval. Through the analysis of these time series, the laws of the development and change of phenomena are discovered and revealed from them and this knowledge and information is used to predict.

After testing with white noise and autocorrelation, a time series model can be established for data prediction. The predicted series plot is shown in Figure 1.

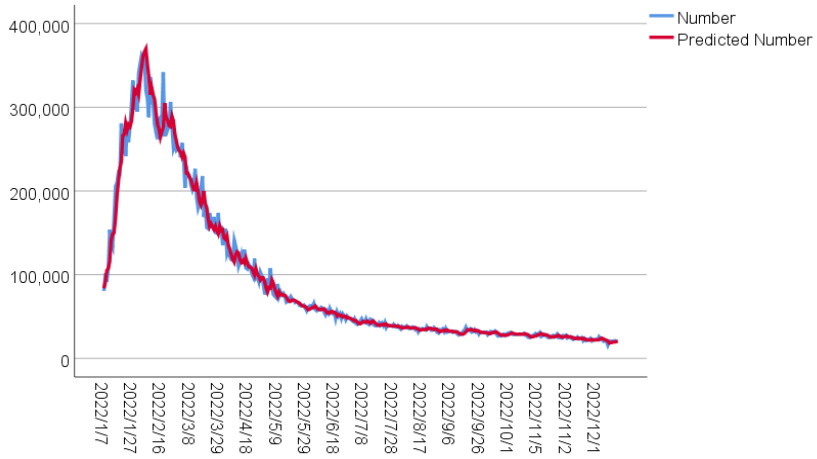

Figure 1: Sequence diagram

The R^2 of the model is 0.984, which means that the model fits well and the model is accurate; the significance is lower than 0.05, which means that the final coefficients obtained are relatively significant, thus indicating good forecasting results. The forecasting results for March 1, 2023 are within the range [19505, 20935], and the forecast value is specifically 20220.

2.2. Building a BP Neural Network Model

BP neural network [3-10] is a nonlinear dynamical system containing three layers of units: input layer, implicit layer, and output layer. It is a multilayer forward network based on back propagation algorithm, which contains one or more hidden layers in addition to the input and output layers. The processing unit of each layer only receives the signal from the output of the processing unit of the previous layer, and after processing the received signal, it is input to the next layer. The final model structure is a forward acyclic network topology with full connection between processing units of each layer and no connection between processing units of the same layer, and its network structure

is shown in Figure 2.

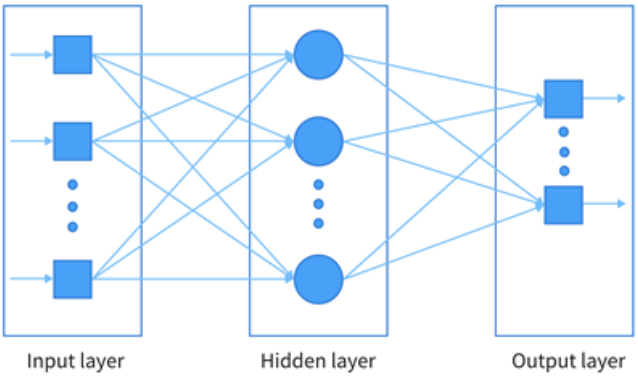

Figure 2: Layer forward network structure

Multi-parameter neural networks often require training samples to reduce errors, which generally require initializing the network, setting the excitation function, calculating the energy function and chain solving, as follows:

Step 1. Determination of the initial parameters of the network

The parameters given in the table can be chosen independently according to the actual prediction effect. We determined the initial parameters of the network and set. Maximum number of iterations is 1000, Number of neurons in the middle layer is 8, Network learning rate is 0.05, and error of training is 0.001.

Step 2. Initialize network weights and thresholds

When initializing the network, the weights and thresholds of the network are assigned to a matrix of random numbers. Since the network contains 6 factors y hidden layer neurons, the random number square of weights $w_{ij}(t)$ between the input layer and the first intermediate layer is the matrix of $y \times 6$, As the principle of BP neural network can be known, there are y neuron has y threshold, so the second layer of neuron threshold random number matrix $B_{ij}(t)$ is: $y \times 1$ of the matrix, and similarly the third layer of the weight matrix is $1 \times y$, threshold $B_{ij}(t)$ is 1×1 matrix.

Step 3. Calculate the input and output of each layer

● Layer 1 output:

Because the excitation function we place in the first layer is a linear function, the input and output of the first layer of the network used are the actual input samples, i.e.

$$O_1 = X .$$

● Layer 2 Inputs and Outputs:

The input of the latter layer comes from the sum of the values of the neurons in the previous layer and the threshold value, i.e.

$$I_2 = w_{ij} \times X + B_{ij} \times ones .$$

In the output of the second layer, assuming that the excitation function of the second layer is a unipolar S-type function, i.e.

$$f(x) = \frac{1}{1+e^{-x}} .$$

The output of the second layer is obtained by $O_2 = \frac{1}{1+e^{-I_2}}$.

● Layer 3 Inputs and Outputs

Similarly, the input to the latter layer comes from the sum of the values of the neurons in the previous layer and the threshold, so the input to the third layer is

$$I_3 = w_{jk} \times O_2 + B_{jk} \times ones$$

As a rule of thumb, the excitation function of the third layer is also generally linear, so the output matrix is $O_2 = I_3$.

Step 4. Calculating the energy function

The energy function is calculated to achieve the desired error on the completion of the training network.

$$E = \sum \left(Y - O_3\right)^2$$

Step5. Weight and threshold adjustment amount (between the second and third floors)

The amount of weight and threshold adjustment is the core of a multiparameter neural network, using a chained partial differential

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \times \left(Y - O_3\right) \times O_2$$

$$\Delta B_{jk} = -\eta \frac{\partial E}{\partial w_{jk}} = -\eta \times \left(Y - O_3\right) \times ones$$

Step 6. Weight and threshold adjustment amount (between the first and second floors)

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial B_{ij}} = -\eta \times w_{ij} \times \left(Y - O_3\right) \times O_2 \times \left(1 - O_2\right) \times X$$

$$\Delta B_{ij} = -\eta \frac{\partial E}{\partial B_{ij}} = -\eta \times w_{ij} \times \left(Y - O_3\right) \times O_2 \times \left(1 - O_2\right) \times ones$$

Step 7. Adjusted weights and thresholds

$$w_{jk}\left(t+1\right) = -\eta \frac{\partial E}{\partial w_{jk}} + w_{jk}\left(t\right) = \Delta w_{jk} + w_{jk}\left(t\right)$$

$$B_{jk}\left(t+1\right) = -\eta \frac{\partial E}{\partial B_{jk}} + B_{jk}\left(t\right) = \Delta B_{jk} + B_{jk}\left(t\right)$$

We encode words according to their composition properties, first coding the 26 letters of the alphabet, and then encoding the words, such as coding manly as 13, 1, 14, 12, 25; the rest of the words are treated according to such encoding. The real value and prediction results are plotted in Figure 3.

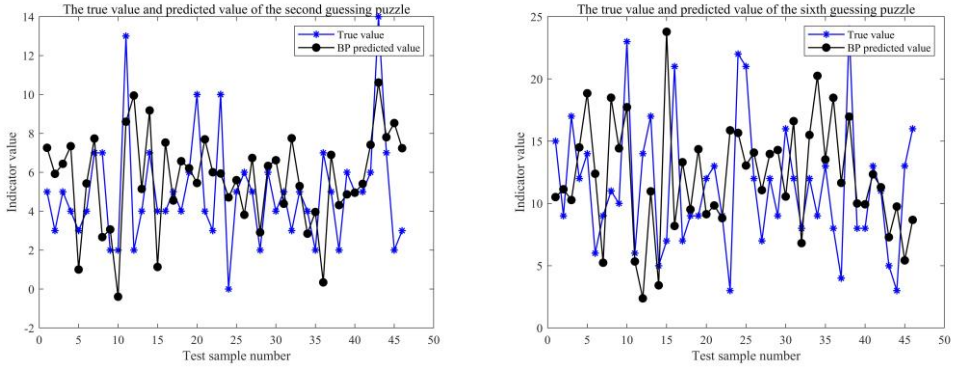

Figure 3: Partial comparison chart

According to Figure 3, it can be seen that the percentages corresponding to the number of attempts calculated by the model fit the sample better with less error and high model accuracy. Therefore, using the neural network model is fully capable of the data fitting task. We selected Percent in 3 tries and Percent in 4 tries data for error analysis. We can see that the error of most of the data is below 15%, so we can judge that we have more than 85% confidence in the neural network prediction model we have built. Based on the established BP neural network prediction model for the term EERIE on March 1, 2023, the prediction results are 3, 4, 5, 6, 7 or more tries are 10, 36, 32, 18, 4.

3. Clustering analysis[11]

Cluster analysis is the clustering of data sets based on similarity. The K-Means algorithm is a simple iterative clustering algorithm that uses distance as a similarity metric to discover k classes in a given dataset, and the center of each class is obtained from the mean of all values in the class, and the center of each class is described by the cluster center. For a given dataset x (containing n data points of one dimension and more than one dimension) and the number of class k to be obtained, the Euclidean distance is chosen as the similarity metric and the clustering objective implements the clustering vindication of the individual classes that minimizes.

$$J = \sum_{k=1}^{k} \sum_{i=1}^{n} \left\| x_i - u_k \right\|^2$$

Among them, k is the number of clustering groups, u_k is the mean vector of the kth cluster in the clustering, x_i is the i-th cluster in the clustering.

Table 2: Final clustering centers

Variables	Simple	Medium Difficulty	Difficulty
1 try	1	0	0
2 tries	9	4	3
3 tries	30	19	12
4 tries	34	35	25
5 tries	18	27	29
6 tries	7	12	22
7 or more tries	1	2	8

The percentage of user attempts has a strong relationship with the difficulty; the more attempts,

the greater the corresponding percentage the greater the difficulty. Therefore, we used SPSS as a modeling tool to perform K-mean clustering analysis on the data, using the percentage of attempts as a variable and the words as case markers for clustering analysis. Through analysis of variance, each variable is significant to the clustering. Table 2 shows the final clustering center, which is the mean value of each class.

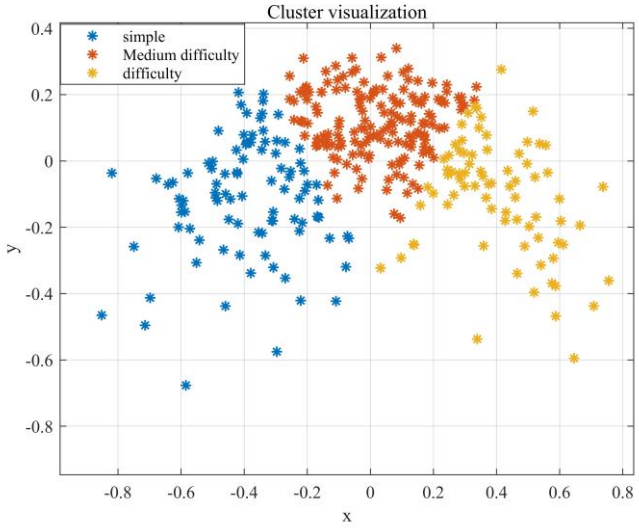

Figure 4: Clustering scatter plot

According to Figure 4 it can be seen simply and intuitively that the clustering gets 3 categories, each category is relatively more concentrated, especially the category Medium Difficulty is the most concentrated, which indicates the excellent results of the clustering analysis.

Table 3: Results of partial clustering of single words

Word	Degree of difficulty
manly	Medium Difficulty
extra	Medium Difficulty
aorta	Simple
taper	Simple
woken	Difficulty
happy	Simple
baker	Difficulty

Some results of the classification of the difficulty level of the words in the data are shown in Table 3. Based on the established cluster analysis model we predicted the EERIE and the results showed that the word belongs to the Medium Difficulty category.

There are many ways to evaluate the classification models, such as accuracy, precision, AUC, Gini, ROC, etc. The evaluation method of ROC curve is different from traditional evaluation methods, and there is no need for this limitation. Instead, based on the actual situation, intermediate states are allowed, and the experimental results can be divided into multiple ordered classifications. AUC is the area under the ROC curve, and the larger the AUC value, the higher the correct rate of the classifier. The ROC curve is shown in Figure 5. The calculated value of AUC is 0.86, the results show that the clustering model we developed is more accurate in predicting the results.

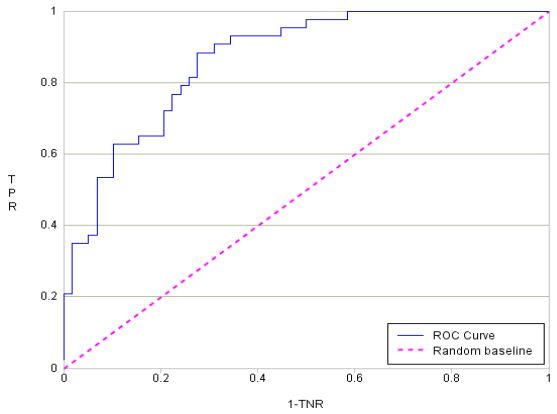

Figure 5: ROC graph

4. Conclusions

From our research, we can see that the Wordle game has a certain level of difficulty, but there is still room for improvement. Therefore, we put forward the following suggestions: (1) Reward mechanism can be implemented for users who can guess words once to increase the enthusiasm of users for guessing words once to maintain the popularity of the game. (2) 6-letter words can be added to improve the difficulty of the game and the appeal of users who are looking forward to challenges. (3) More language versions of Wordle games can be added to attract users from different countries and maintain the number of games. In a word, we hope that the Wordle game can develop better and attract more people to join the game.

Acknowledgements

The authors gratefully acknowledge the financial support from the 2023 Higher Education Science Research Planning Project of the China Association of Higher Education - Practice and Research on University Mathematics Curriculum Based on Project-Based Learning (23SX0411) ; 2023 Project of Beijing Higher Education Association-Practice and research on ideological and political education in probability and statistics courses from the perspective of critical thinking (MS2023024) ; The financial support from the Research on the Teaching Reform and Practice of Integrating Ideological and Political Elements into the Probability and Statistics A Course of the Key Teaching Reform Project of Beijing Institute of Petrochemical Technology in 2023 (ZD202306001).

References

[1] Question C of the 2023 American College Student Mathematical Modeling Competition [EB/OL].The official website of the American College Student Mathematical Modeling Competition, https://www.contest.comap. com/ undergraduate/contests/index.html.

[2] Wang Yan, Applied Time Series Analysis (6th Edition) [M]. Renmin University of China Press, 2022.

[3] Li D, Lv R, Si G, et al. Hybrid neural network-based prediction model for tribological properties of polyamide6-based friction materials[J]. Polymer Composites, 2017(8):38.

[4] Zhao Yame, Yang Jianguo, Li Beizhi. Implementation of a neural network-based prediction model input parameter configuration method [J]. Computer Measurement and Control, 2005, 13(9):3.

[5] Lv Xiaoling, Song Jie. Big data mining and statistical machine learning [M]. Beijing: People's University of China Press, 2016:114-124.

[6] Han Yuzheng, Li Chunliang, Wang Haitao. Prediction of yield based on BP neural network and fitting[J]. Metallurgical Engineering, 2019, 6(4):232-239.

[7] Hu Yixiang, Huang Yue, Wang Ting, et al. Trend analysis of house price based on neural network prediction model--Haikou and Sanya cities as examples[J]. Fujian Computer, 2018, 34(12):2.
[8] Chen Changsong, Duan Shanxu, Yin Jinjin. Design of a neural network-based prediction model for photovoltaic array power generation [J]. Journal of Electrotechnology, 2009(9):6.
[9] Chang C H, Hung Y H. A Neural Network-Based Prediction Model in Embedded Processes of Gold Wire Bonding Structure for Stacked Die Package [J]. Proceedings of the IEEE, 2009, 97(1):78-83.
[10] Yang T, Tsai T N, Yeh J. A neural network-based prediction model for fine pitch stencil-printing quality in surface mount assembly [J]. Engineering Applications of Artificial Intelligence, 2005, 18(3):335-341.
[11] He Xiaoqun, Multivariate Statistical Analysis (5th Edition) [M]. Renmin University of China Press, 2019.