Application of neural network methods for crossword prediction

Nie Wenxuan, Sun Ruotong, Wu Siyu
Sun Yat-sen University
Guangzhou, China
niewx3@mail2.sysu.edu.cn

2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA) | 979-8-3503-1467-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIPCA59209.2023.10257716

***Abstract*—With the widespread application of the machine learning, this paper is aimed to analyze a game named 'Wordle puzzles' launched by [New York Times], using the machine learning big data analysis and neural network models to get better puzzle design and promotion. These models will analysis the relationship between word attributes and result distribution and then predict the results. Firstly,this paper constructed a time series model and used quadratic exponential smoothing method to predict the total number of player reports.Then it quantifies the difficulty attributes of each words after classifying them by word repetition and commonness as indicators.And it uses correlation analysis to explore the effect of word attributes on the percentage of people reporting using the difficulty mode. After that, a BP neural network model is constructed, learned, trained and it can predict the results of a given word on a certain date, as well as predicting the difficulty level and score of that word. Combined with machine learning, neural network and some research, It is recommended that the game companies adjust word difficulty attributes such as commonness and repetition according to the number of players, generally keep word difficulty smooth, and then adjust word difficulty attributes according to the average number of correct submissions of reported results.**

Keywords—Machine learning, Neural Network Model, Difficulty attributes, Time series models, Correlation analysis, Outcome prediction

I. Introduction

The machine learning method addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields,lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science.Also the adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce,leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.And this article will use machine learning to solve a game marketing implementation problem.This game is named Wordle puzzles.

The Wordle puzzles launched by [New York Times] have been very popular since its release and now are available in more than 60 languages. In Wordle puzzles, players have 5 chances per day to guess a word with 5 letters according to certain rules,of which they can choose between normal or hard mode.Over time,the percentage of players using the difficult mode and the percentage of attempts per result are constantly changing, which shows certain patterns. This paper will use time series models and BP neural network models to explore such patterns and connections and make predictions to suggest better settings for the game. Regarding the data analysis of this game, this paper is going to solve the following problems:building a time series model to fit and predict the change in the number of reported results over time, and analyze the correlation between the difficulty attributes of the words on the certain day and the percentage of players using the difficulty model; building a BP neural network model to predict the distribution of the reported results of the day based on the words of a future certain day;to determine the difficulty attributes of a given word and analysis the sensitivity of this model. Finally,propose the advantages, disadvantages and improvements of the modified model.

II. Time series model

The fundamental concept of time series forecasting involves studying the trend of changes in a target variable's time series by processing the time series itself. Commonly used forecasting methods include the moving average method, the exponential smoothing method, and the differenced exponential smoothing method, among others.

In this model, we aim to summarize the pattern of changes in the number of reports over time and predict the future quantity of reports on a specific day based on this pattern. Generally, the impact of historical data on future values decreases as the time interval grows. Therefore, a more practical approach is to calculate the weighted average of each period's observed values based on the order of time as the forecast value. The exponential smoothing method satisfies this requirement and has a simple recursive form.

Upon obtaining the 2022 report data from the official website of Wordle, we conducted preliminary processing procedures to eliminate any data with irregularities in the number of reports or word structures. Subsequently, we utilized the organized data to generate a scatter plot displaying the number of reports, as presented in Figure 1.

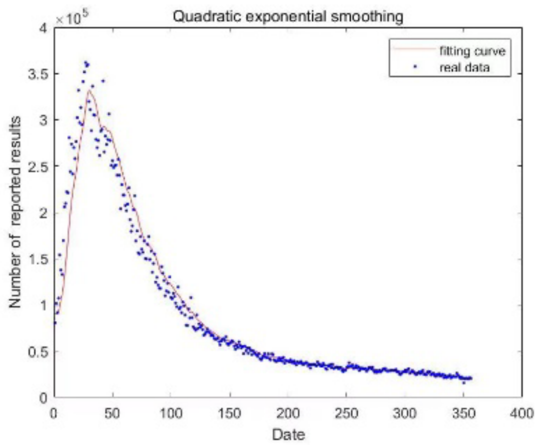


Fig. 1. Fitted scatter plot of report number changes

Based on the results above, we first consider using a single exponential smoothing method. The time series are $y_1, y_2, \cdots, y_t, \cdots$, N is the number of terms in the moving average, and $\alpha\left(0<\alpha<1\right)$ is the weighting coefficient. The single exponential smoothing formula improved from the average moving formula is

$$S_t^{(1)} = \alpha y_t + \left(1-\alpha\right) S_{t-1}^{(1)} = S_{t-1}^{(1)} + \alpha\left(y_t - S_{t-1}^{(1)}\right)$$

Using this smooth value to predict, we derive the prediction model

$$\hat{y}_{t+1} = \alpha y_t + \left(1-\alpha\right)\hat{y}_t$$

The standard deviation of the prediction is

$$S = \sqrt{\frac{\sum_{t=N+1}^{T}\left(\hat{y}_t - y_t\right)^2}{T-N}}$$

where $y_i\left(i=1,2,3,\cdots,t\right)$ is the number of reported results collected daily.

For the weighting coefficient α, we notice that the number of reported results collected daily has not fluctuated much recently and it is relatively stable, so we choose a smaller α. After several tests, we finally choose 0.3 with the smallest standard deviation (relatively good output) as the weighting coefficient of the model.

Given the large number of dates included in the collected reports, the effect of the initial value on subsequent predicted values is minimal. Therefore, we choose the initial value to be

$$S_0^{(1)} = \frac{y_1 + y_2}{2}$$

Due to the linear trend in the time series, there is a noticeable lag bias when using single exponential smoothing for prediction. Therefore, we apply double exponential smoothing on top of single exponential smoothing to correct for the lag bias and establish a linear trend model based on the pattern of the lag bias. The calculation formula for this model is

$$\begin{cases} S_t^{(1)} = \alpha y_t + \left(1-\alpha\right) S_{t-1}^{(1)} \\ S_t^{(2)} = \alpha S_t^{(1)} + \left(1-\alpha\right) S_{t-1}^{(2)} \end{cases}$$

The prediction model is

$$\hat{y}_{t+T} = a_t + b_t T, T = 1, 2, \cdots$$

where

$$a_t = 2S_t^{(1)} - S_t^{(2)}, b_t = \frac{\alpha}{1-\alpha}\left(S_t^{(1)} - S_t^{(2)}\right)$$

The predicted results obtained from this model are shown in Figure 2.

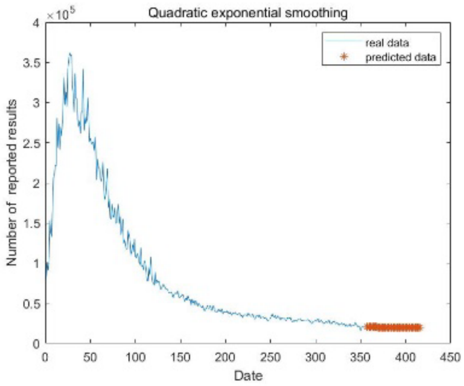


Fig. 2. Forecast results obtained by quadratic exponential smoothing

When puzzle games were initially released, people showed a higher level of enthusiasm for this novel activity. Over time, this enthusiasm may gradually decline, eventually stabilizing into a relatively consistent group of players. The changes observed in the model are in accordance with the aforementioned pattern.

III. CORRELATION ANALYSIS

Correlation analysis is typically used to analyze two or more variables or elements that correlate, to measure the degree of correlation between the two variables or factors. [2]

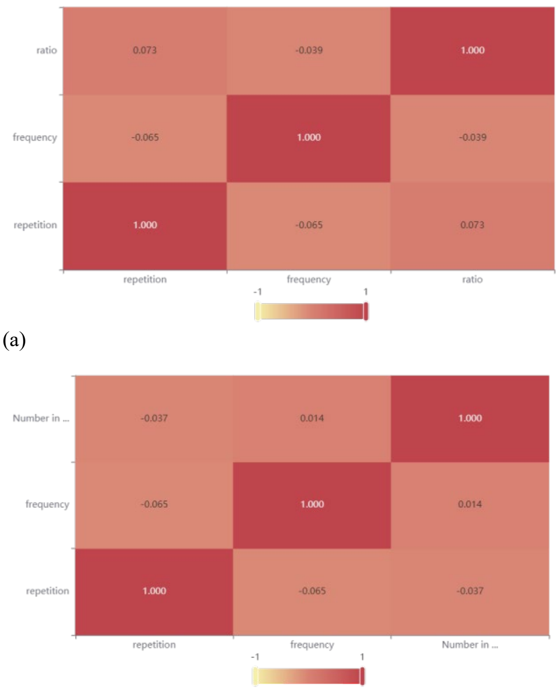
The report data includes another indicator, which is the percentage of difficulty mode selection. Certain properties of words may have a certain influence on this indicator, and they are variables that can be controlled by the question setter. Through exploring the properties of words and language habits, we have identified two properties that are the most significant and likely to affect the percentage of scores in reports under difficult mode: the maximum number of letter repetitions in a word and the frequency of occurrence of the word.

In this study, word properties were quantified by counting the maximum number of repeated letters in a word and obtaining the word frequency star rating from a corpus [1]. Next, the percentage score of the report in the difficult mode needs to be determined. First, the percentage is quantified into difficulty level, which is the average number of attempts for a question on a certain day. Thus, the difficulty level of question i is determined by the formula

$$S_i = X_i = \frac{1}{100n}\sum_{i=1}^{n}\sum_{j=1}^{7} a_{ij}$$

In the above equation, the results with more than 6 attempts are combined and regarded as the 7th attempt. i is the date sequence number corresponding to the word; j is the number of attempts; a_{ij} is the frequency of the attempt on the word. n is the total number of words in the sample.

We consider using the proportion of difficult mode selection as the dependent variable and word attributes as the independent variable for correlation analysis. In this study, we conducted a statistical test to determine whether there is a significant relationship between the highest repetition frequency of letters and the commonness level of words, and the proportion of difficult mode selection. The results are shown in Figure 3(a). In addition, we also conducted a correlation test between the above word properties and the number of submissions for the difficult mode, and the results are shown in Figure 3(b).



(a)

(b)

Fig. 3. Heat map of correlation coefficient

Based on the results of the two analyses, the correlation coefficients are relatively small, indicating that the properties of words are not significantly correlated with the percentage of scores in the difficult mode. (This may be because the mode selection is made before each game, and players cannot predict the difficulty of the daily puzzle in advance, so there is no significant correlation between the word properties and the score in the difficult mode.)

IV. Neural Network Model

BP neural network is a multi-layer feedforward neural network that can analyze and master the underlying rules between a set of corresponding input-output data provided in advance, and ultimately, according to these rules, use new input data to calculate the output result.

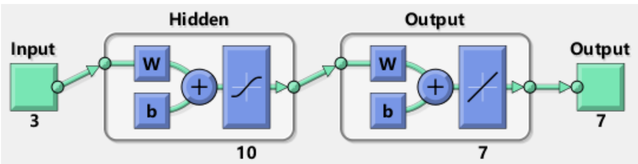


Fig. 4. Three-layer BP neural network

We have established a three-layer neural network as shown in Figure 4, consisting of an input layer, a hidden layer, and an output layer. The sample data was divided into a training set, a validation set, and a test set. The learning process of this neural network can be divided into two parts: the forward propagation of signals and the backward propagation of errors.

Let $X=(x_0,x_1,\cdots,x_n)$ be the input vector, $Y=(y_0,y_1,\cdots,y_n)$ be the output vector and $W=(wi_0,wi_1,\cdots,wi_n)$ be the weight vector. The model of the neural network is

$$\begin{cases} net_i=\sum_{j=1}^{n} w_{ij}x_j-\theta=\sum_{j=0}^{n} w_{ij}x_j \\ y_i=f(net_i) \end{cases}$$

where net_i is the net amount of activation, $x_1 \sim x_n$ is the input signal from other neurons, w_{ij} represents the connection weight from neuron j to neuron i, and θ represents a threshold or bias. y_i is the activation function, including linear function, S-shape function, threshold function and bipolar S-shape function.

The error function is

$$e=\frac{1}{2}\sum_{o=1}^{q}\left(d_o(k)-yo_o(k)\right)^2$$

where $d_o=(d_1,d_2,\cdots,d_q)$ is the expected output vector, $yo=(yo_1,yo_2,\cdots,yo_q)$ is the output vector of the output layer, and q is the number of neurons at the output layer.

To make the error between the actual output and the standard output reach the minimum value, the derivative is carried out, and the weight is calculated as

$$w_{oh}^{N+1}=w_{oh}^{N}+\mu\delta_o(k)ho_h(k)$$

where μ is the set learning rate.

The global error is

$$E=\frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{q}\left(d_o(k)-yo_o(k)\right)^2$$

Finally, it is judged whether the network error meets the requirements. The algorithm ends when the error reaches the preset accuracy or the number of learning is greater than the set maximum number. Otherwise, the next learning sample and the corresponding desired output are selected and returned to the next round of learning.-The neural network was used to train and predict the score cases of word attribute influence and the difficulty classification cases of words, respectively.

A. Model of The Score Situation Influenced by Word Attributes

A BP neural network model with 3 inputs and 7 outputs was developed. The highest number of repetitions of letters in a word, word commonness, and date are used as signals acting on the input layer, and the percentage of success of each attempt is output as the predicted score. Figure 5 shows the validation of the neural network and the regression of each set, from which it can be seen that the training of the neural network is very reliable.

Fig. 5. Neural network verification and regression analysis diagram of training

The word EERIE on March 1, 2023 was quantified for metrics as input and the trained neural network model was used to predict the answers for that day. The output of the neural network is shown in the table I below.

TAB. I PREDICTED ANSWERS

Date	Word	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries
2023/3/1	EERIE	1	7	25	29	23	14	1

B. Difficulty Classification Model of Words

Directly using the two-word attributes extracted in the previous paper as indicators for difficulty classification suffers from too few criteria and too subjective conclusions. Based on this, we decided to use the average number of attempts by users to classify word difficulty. This method is characterized by high objectivity and is capable of quantitative analysis. Based on the correlation analysis, the difficulty scores were classified into four levels regarding the extreme difference and sample size Figure 6 shows the number of reports under each difficulty classification.

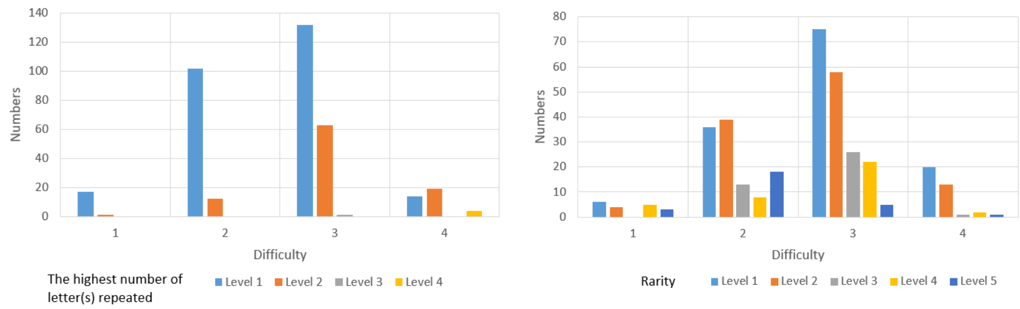


Fig. 6. Word attributes under each difficulty category

A BP neural network model with 2 inputs and 2 outputs is established. The highest number of repetitions of letters in a word and the word commonness are the inputs of the neural network, and the difficulty scores and difficulty levels of words are the outputs of the neural network. Figure 7 shows the validation of the neural network and the regression of each set, from which it can be seen that the training of the neural network is very reliable.

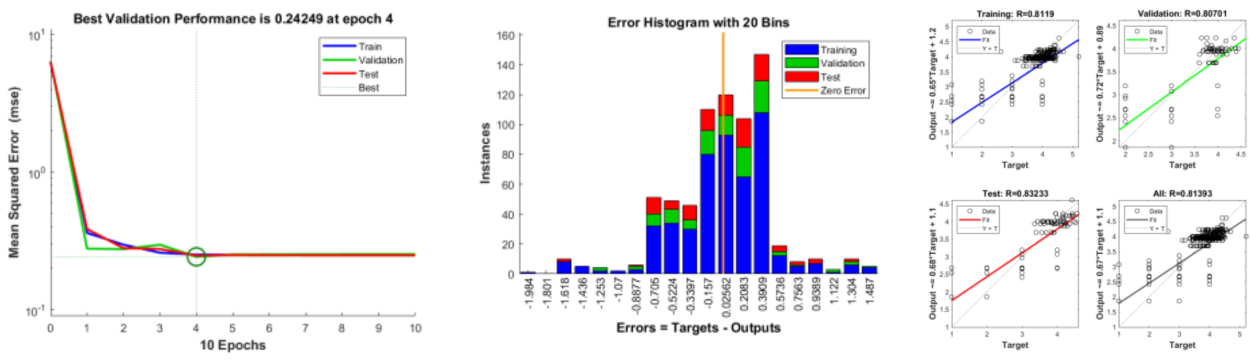


Fig. 7. Neural network verification and regression analysis diagram of training

The word EERIE was quantified for metrics as input, and the trained neural network model was used to predict its difficulty score and difficulty level. The results of the neural network output are shown in the table II below.

TAB. II DIFFICULTY PREDICTION

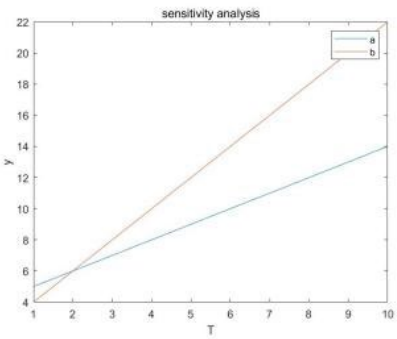
Word	Difficulty scores	Grade of difficulty
EERIE	4.5091	4

V. SENSITIVITY ANALYSIS AND MODEL ADVANTAGES AND DISADVANTAGES

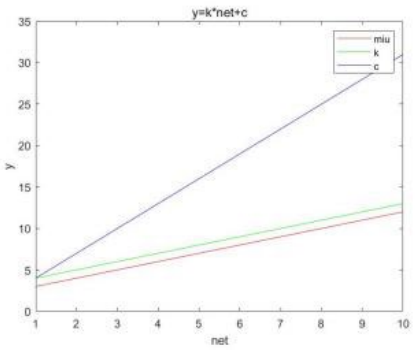
The sensitivity analysis about the time series model is shown in Figure 8(a), that is, the predicted value increases with the increase of a_t and b_t, and is linear. At the same time, the predicted value is affected by b_t grows faster that is more sensitive. Regarding the neural network model, if the activation function is $y = k \times net + c$, as shown in Figure 8(b), the prediction value increases with the increase of k, c and the learning rate μ, and it is linear, while the prediction value is more influenced by c grows faster that is more sensitive; if the activation function is

$$y = \frac{1}{1 + e^{-\alpha \times net}}$$

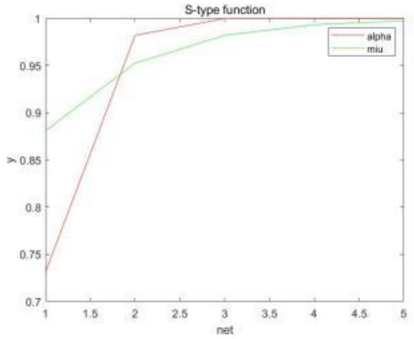
As shown in Figure 8(c), the prediction value increases with the increase of alpha and μ increases with the increase of α and μ, but it is a non-linear relationship, which is more influenced by α and grows faster until net is less than 2. However, α reaches a plateau more quickly and μ continues to grow.



(a)



(b)



(c)

Fig. 8. Sensitivity analysis curves

The advantages of the model are two, firstly, the method used in the time series model, the simple moving average, trend moving average, primary exponential smoothing and secondary exponential smoothing methods were tried in turn during the test, and the secondary exponential smoothing method was determined only after combining the advantages of each and the applicable situations, and the optimal prediction curve was finally obtained after debugging the alpha parameter several times. Secondly, the attributes of words are clearly classified, such as letter repetition and commonness, and the words are classified into five categories after consulting the commonness of each word in the dictionary, and the difficulty value of each word is classified quantitatively, which is more accurate when performing correlation analysis and neural network analysis. However, the model also has shortcomings, such as the time series model in the prediction when the external world changes significantly, there is often a large deviation, time series prediction method for short and medium-term prediction than long-term prediction, and BP neural network can be optimized for genetic algorithm-based neural network better results.

VI. CONCLUSION

Using machine learning, this paper builds a time series model and a neural network model that can extract useful information from a large amount of player data and predict the answers to specific words and grade the difficulty, which provides a good reference and guidance for puzzle design and promotion. Therefore, it is recommended that game companies strengthen the application of data analysis and machine learning to better understand players' needs and preferences in order to design more appealing puzzles and promote them to a wider audience. From the perspective of word difficulty game companies can set more patterns of different difficulties to attract players of different levels by adjusting the commonness and repetition of words, from the perspective of freshness of game duration game companies can update and add new puzzles regularly to keep the game fresh and challenging, from the perspective of user experience and interactivity they can provide more social features, leaderboards and reward mechanisms to stimulate players' competition psychology and sense of achievement, which is more conducive to the promotion of the game.

REFERENCES

[1] Collins Publishing Company, UK. Collins COBUILD Advanced English Learning Dictionary, 8th Edition [M]. Beijing: Foreign Language Teaching and Research Press, 2017.

[2] Xu Weichao. Overview of research on correlation coefficient [J].

Journal of Guangdong University of Technology, 2012,29 (3): 12-17.

[3] Jordan, M. I.; Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. doi:10.1126/science.aaa8415

[4] Carleo G, Cirac I, Cranmer K, et al. Machine learning and the physical sciences[J]. Reviews of Modern Physics, 2019, 91(4): 045002.

[5] Bishop C M. Neural networks and their applications[J]. Review of scientific instruments, 1994, 65(6): 1803-1832.