ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

Prediction of Wordle Game based on RBF Neural Network Modeling

Biao Guo*, Jinyue Li, Xuetao Wang

Department of Automation, Hebei University, Baoding 071002, China

*Corresponding author: gb1917461801@163.com

Abstract

Wordle is a popular puzzle currently offered daily by the New York Times, where the number of times a player completes the game uses affects the sense of experience and thus the number of players. In this paper, we use an RBF neural network to predict the distribution of the percentage associated with the number of times a task is completed in terms of the number of times it is used, dividing each word into five letters and coding the pairs, and taking into account the effect of the date, we coded the date from in this paper as a way of investigating the effect of the word itself, as well as the date, on the distribution of the percentages. The prediction results show that the RMSE and R2 of the model are 0.2993 and 0.9021, respectively, the model has high accuracy, and the results are highly credible; this paper takes the word EERIE as an example, and predicts the percentage distribution of the word as 0.34, 3.92, 17.82, 30.99, 26.18, 15.09, and 5.39.

Keywords

RBF Neural Network Model; Pedictive Model; Wordle.

1. Introduction

For linear systems, the use of ARIMA model through a variety of parameter estimation methods can be a better solution to the system prediction problem [1]; for nonlinear systems, generally based on nonlinear autoregressive sliding average (KARMA) model for prediction [2]. In numerical prediction, statistical analysis, interpolation fitting and other methods are mostly used, which reflect their respective advantages in prediction [3]. Artificial Neural Networks (ANN) have obvious superiority in modeling and identification of nonlinear systems Artificial Neural Networks have inter-neuron nonlinearity [4], and its nonlinear mapping represents the percentage of this nonlinear system to improve the model accuracy. Considering that this prediction system is multi-input and multi-output, this paper chooses to use RBF neural network model [5].

2. Prediction Model based on Multiple-input Multiple-output RBF Neural Network

In order to obtain the optimal distribution of predicted results, the RBF neural network prediction model was chosen to predict the future distribution of relevant percentages. The model is described below:

2.1 RBF Neural Network Model Structure

The basic idea of the radial basis function neural network is to use the radial basis function (RBF) as the hidden node "base" to form the hidden space, if the width and center of the RBF are determined, the input layer vectors can be mapped directly to the hidden layer space, rather than connecting through the weights. The hidden layer space is mapped to the output layer space by connecting

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

linearly through the weights, and the final output of the network is the linearly weighted sum of the outputs of the neurons in the center of the hidden layer [6]. Figure 1 shows the structure of the RBF neural network model. In the following paper, the RBF neural network model will be described in detail with the topic problem:

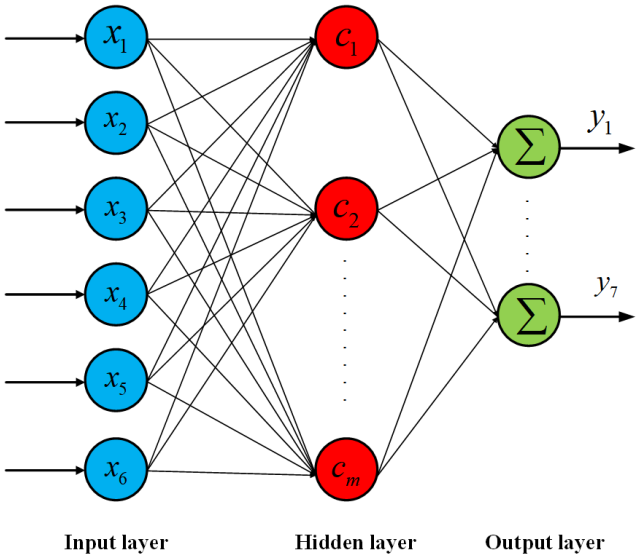

Figure 1. Structure of the RBF neural network model

(1) Determine the input and output layers:

In the topic of this paper, the input layer is the first letter of the word, the second letter, the coded value of the fifth letter in order, and the coded value of the date; so in this paper, we set the neural network to have six neurons, and the inputs are represented by equation (1):

$$x_i = \{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}\} \tag{1}$$

(2) Radial basis function

The hidden layer contains 12 neurons, the activation function is a radial basis function, when the input data and the central range of the basis function is closer, the hidden layer nodes can produce a greater response to the input vector, can produce a greater impact on the results of the output, the input vector and the central range of the distance from the basis function is farther away, the output of the hidden layer of the impact on the results of the results will be dramatically reduced.The hidden layer activation function of RBF neural network The activation functions of the hidden layer of the RBF neural network are generally taken as follows: the inverse multiquadratic function, the Sigmoidal function and the Gaussian function.

Since the Gaussian function has the advantages of simple form, good smoothness, good analyticity and so on, the Gaussian function is chosen as the radial basis kernel function of the RBF neural network in this paper, as shown in Fig. 2. Its formula (2) shows.

$$\varphi(\| x_j - c_j \|) = \exp(-\frac{\| x_j - c_j \|^2}{2\sigma_j^2}) \tag{2}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

where, x_j is the input vector, c_j is the center vector of the Gaussian function of the jth hidden node, $\| x_j - c_j \|^2$ is the 2-parameter of the input vector with respect to the center vector, and $2\sigma_j$ is the variance of the Gaussian function, which denotes the width of the Gaussian function.

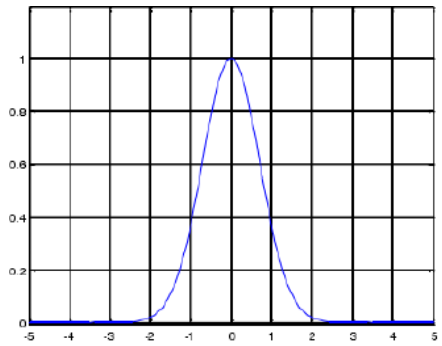

Figure 2. Gaussian function curc

(3) ROutput of the BF neural network:

Predict the percentage of times the result is guessed by the players for the seven outcomes, i.e., 1, 2, ... 6 times and when the player fails to guess the question (7 times). So the output layer of the neural network has two neurons, and the output is the sum of all neurons in the hidden layer and the product of the weights, which is expressed by equation (3):

$$y_{ki} = \sum_{j=1}^{m} w_{ji} \varphi(\| x_i - c_j \|), k = 1, 2 \cdots 7 \tag{3}$$

Where,k is the output neuron identification.

Output evaluation indicators:

$$E_j = \frac{1}{2}[y_i - y_{mi}]^2 \tag{4}$$

Where, y_i is the desired output value, and $y_i = [y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6}, y_{i7}]$; y_{mi} is the actual predicted value of the RBF neural network, and $y_{mi} = [y_{m1}, y_{m2}, y_{m3}, y_{m4}, y_{m5}, y_{m6}, y_{m7}]$.

(4) Weight adjustment for RBF neural networks:

Before the structure of RBF neural network is determined, it is necessary to use clustering algorithm and learning algorithm to find three parameters, which are: Center vectors of basis functions c_j, Width of Gaussian function σ_j, Connection weights between the implicit layer and the output layer w_j. The determination process is in two stages: In the first stage, which belongs to the unsupervised learning category of machine learning, clustering algorithms are used., Determine the center vector of the basis function c_j, Width of Gaussian function σ_j; In the second stage, which belongs to the supervised learning category of machine learning, the RBF neural network is trained from the samples by learning algorithms, Determine the weights from the implicit layer to the output layer w_j.

Since RBF needs to determine the center vector of the basis function in the unsupervised learning phase c_j, Width of Gaussian function σ_j,In this paper, this paper briefly introduces the flow of K-means clustering algorithm, the flowchart of the algorithm is shown in Figure 3.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

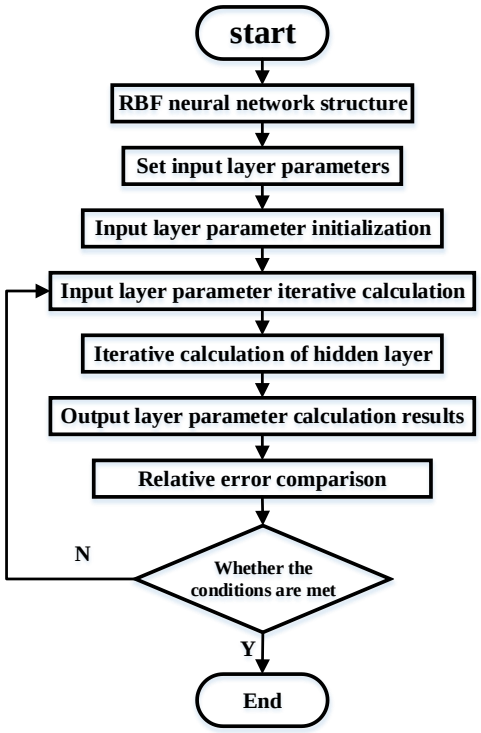

Figure 3. Flowchart of K-Means algorithm

2.2 Data Analysis

First of all this paper descriptive analysis statistical analysis of the percentage of players who scored from the 1st success to the 7th success, the total sample size is 357, the statistical results are shown in Table 1:

Table 1. Results of descriptive statistics

variable	average value	variance	standard deviation	skewness
1	0.465	0.609	0.78	0.469
2	5.776	15.871	3.984	1.453
3	22.658	60.029	7.748	-0.007
4	32.95	28.717	5.359	-0.647
5	23.7	34.862	5.904	0.074
6	11.605	38.38	6.195	1.032
7	2.818	17.054	4.13	1.453

As shown in Table 1, there are more players successfully scoring in the 3rd, 4th, 5th, and 6th attempts, and the number of players in the 1st, 2nd, and incomplete is less, indicating that there are a lot of players who can successfully score in 6 attempts, and the difficulty of the questions is moderate, and most of the players can successfully score in 6 attempts; fewer players can score very quickly or fail to complete the game; the variance of the 3rd, 4th, 5th, and 6th is large, indicating that the number of times players pass the game The variance of the 3rd,4th,5th,6th is large, indicating that the number of times the player passes is fluctuating greatly, i.e., seldom passes within a fixed number of times; the results of the skewness are all in the range of (-1.5,1.5), which indicates that the data distribution can satisfy the positive-target distribution well, and can improve the prediction accuracy of the subsequent RBF neural network model.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

Then, this paper pre-processes the percentage of attempts, through a reasonable allocation of less than 100% of the sample according to the proportion of each reasonable allocation of the proportion, more than 100% according to the proportion of the percentage of the size of a reasonable reduction, and then the words in EXCEL for the splitting process, each letter of the coding labeling, and at the same time, taking into account the effect of the date of the factor of the score of the game, this paper is also a week of the cycle of the quantitative treatment of the date. At the same time, considering the effect of date on the score of the game, this paper also takes a week as a cycle to quantify the date. The a-z are coded as 1-26: Monday-Sunday are coded as 1-7.

3. Results

3.1 Data Preparation

We chose 357 samples as input to the RBF neural network, and the partial data table for the input is shown in Table 2:

Table 2. Selected input data

Contest number	Words	X1	X2	X3	X4	X5	X6
203	crank	3	18	1	14	11	6
204	gorge	7	15	18	7	5	7
205	query	17	21	5	18	25	1
:	:	:	:	:	:	:	:
559	molar	13	15	12	1	18	5
560	manly	13	1	14	12	25	6

where,X1-X5 are the coded values from the first letter to the letter in order,X6 is the coded value of the date, which ranges from (1-7).

The input partial expectations are shown in Table 3:

Table 3. Selected Desired Outputs

Contest number	Word	Y1	Y2	Y3	Y4	Y5	Y6	Y7
203	crank	1	5	23	31	24	14	2
204	gorge	1	3	13	27	30	22	4
205	query	1	4	16	30	30	17	2
:	:	:	:	:	:	:	:	:
559	molar	0	4	21	38	26	9	1
560	manly	0	2	17	37	29	12	2

where,X1-X6 are the correlation percentages of scores in the first to sixth attempts, respectively, andY7 is the correlation percentage of unsuccessful scores.

The total sample size is 357,the number of training samples is 249,and the number of training and validation sets are 54 and 54 respectively.

3.2 Projected Results

The accuracy of the model will affect the accuracy of the prediction results. Therefore, this paper analyzes the sensitivity of the RBF neural network model. The two indexes, the number of neurons in the hidden layer and the expansion parameter, affect the prediction performance of the RBF neural

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

network, so this paper investigates the RMSE and R2 fitting effect of the prediction of the RBF neural network by varying the size of these two parameters[7].

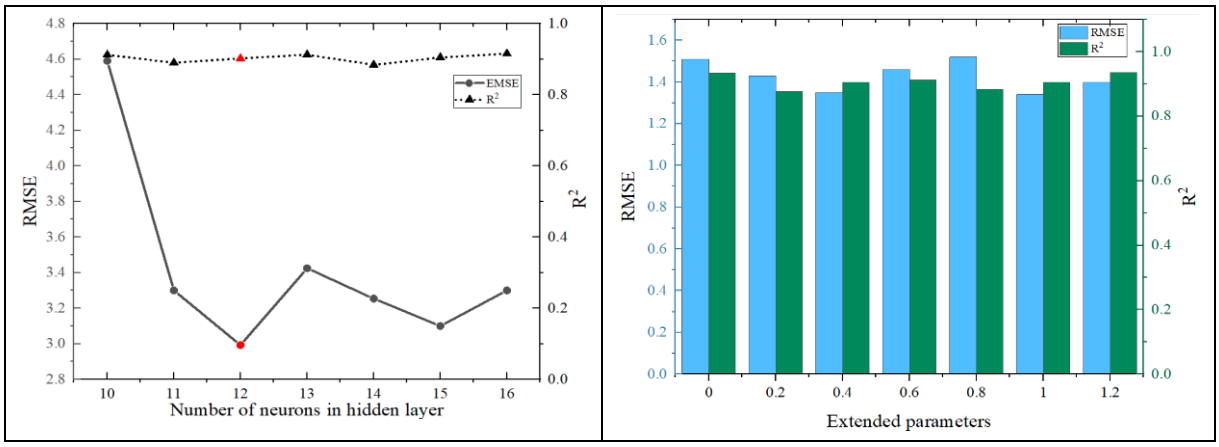

Figure 4. Effect of the number of neurons and expansion parameters of the hidden layer on R2 and RMSE

The above figure 4 shows the changes of RMSE and R2 with the number of neurons in the hidden layer, and it can be seen that: when the number of neurons in the hidden layer is around 12, the RMSE and R2 fluctuate greatly, and when the RMSE is more than 12, the number of neurons has no obvious effect on the fitting effect; Figure 4 shows the effects of the RMSE and R2 on the changes of the expansion parameters, and the EMSE and R2 have no obvious effect on the changes of the expansion parameters when the number of neurons in the hidden layer is 12. When the number of hidden layer neurons is 12, EMSE and R2 have no obvious effect on the variation of expansion parameters. In conclusion, 12 hidden neurons are the optimal state for this model RBF neural network. Based on the data preparation and model training in the above two steps, this paper finds that the prediction effect of the RBF-based neural network model is better, according to the requirements of the topic, this paper takes EERIE on March 1, 2023 as an example, and solves the correlation percentage of the day through the model, and the following are the results of the solution.

Based on the above two steps of data preparation and model training, we found that the neural network model based on RBF has better prediction effect. According to the requirements of the topic, we took EERIE on March 1, 2023 as an example and used the model to solve the correlation percentage on this day, and the solution results are shown in Table 4:

Table 4. Relevant Percentage Prediction Results

norm	Y1	Y2	Y3	Y4	Y5	Y6	Y7
projected value	0.34	3.92	17.82	30.99	26.18	15.09	5.39
average value	0.475	5.776	22.658	32.95	23.7	11.605	2.836

Where,Y1-Y7 are the percentage of correlation between the score of the 1st...6th attempt and the percentage of failures.

The results show that: the predicted value of word EERIE for the 1st.... .4 attempts at the word EERIE were predicted to be less than the mean, i.e., the number of people who passed the game with fewer attempts was less than the overall mean of the previous data, suggesting that the word was likely to be more difficult on this day; the percentage of predictions for the word EERIE that passed or failed on the 5th . .6 the predicted percentage of passing or failing the game on the 5th attempt was higher than the overall mean, indicating that more attempts were used to pass the game, suggesting that the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202310_9(10).0020

word was more difficult to guess correctly, which is related to factors such as the number of repeated letters in the word and the frequency of occurrence of the word.

4. Conclusion

From the results, it can be seen that: the predicted value of the 1st...4th times of the word EERIE on March 1 is smaller than the average value, that is, the number of people who passed the game with fewer times is smaller than the overall average value of the previous data, which indicates that the difficulty of the word on this day may be higher; the predicted percentage of passing or failing the game on the 5th...6th times is higher than the overall average, which indicates that the number of times of passing the game is higher, which indicates that the difficulty of guessing the right word is higher, and is related to the number of repetitive letters of a word, as well as the frequency of the word, and so on, so that the designer of the game can select the appropriate word according to the prediction results.

In practical applications, it is necessary to pay attention to issues such as the sensitivity of RBF to the amount of data and training time; in future research, some improvement strategies can be proposed, such as the introduction of a new feature selection method, the improvement of model parameter tuning algorithms, and the improvement of network structure design.

References

[1] Li Yucui,Huang Zhiming,Wei Jinling. Analysis of the development trend of cross-border e-commerce between China and Vietnam based on ARIMA model[J]. Foreign Trade and Economic Cooperation,2023 (07):6-10.

[2] LIU Zhuang,WANG Yongguo,DING Chengcheng et al. Prediction of chlorophyll a concentration in Changtan Reservoir based on ARIMA model[J]. Environmental Pollution and Prevention,2023,45(07): 895-902.

[3] SONG Yuhua,ZI Zixiao,LI Huanqun et al. A prediction method of fire supervision and inspection frequency based on random forest model[J]. Journal of China People's Police University,2023,39(02):51-56.

[4] MENG Lingxu,ZHANG Jing,ZHANG Hongtao et al. Optimization of ginseng vacuum freeze-drying process based on information entropy theory and back propagation artificial neural network[J]. Shizhen Guomao Guomao, 2023,34(01):91-95.

[5] Hualong,Qi Chong,Liu Xuejiao. Solar power generation prediction based on RBF neural network[J]. Science and Industry,2022,22(07):375-380.

[6] SUN Ji-Yun,WANG Qing-Hua,WANG Zhen-Yan et al. Prediction of adjustment parameters of inclined roll piercing machine based on PIO-RBF neural network[J]. Journal of Plasticity Engineering,2023, 30(03):197-203.

[7] GUO Chenxia,LI Da,YANG Ruifeng.RBF neural network in dynamic weighing of animals[J]. Electronic Design Engineering,2023,31(15):75-78+83.