ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

Wordle data analysis based on time series analysis model

Xuyi Shi[*], Jiachen Guang[#], Liangsu Shao[#]

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, China, 100048
[*]Corresponding author: uccshixuyi@163.com
[#]These authors contributed equally.

Abstract: *Using LSTM time series analysis and forecasting is an important guide for Wordle's game development direction planning and economic revenue visualization. Accurate game report data prediction is of great significance for game development, economic investment, post-game planning, and improving player experience. As Wordle's game becomes more and more popular, it is essential to make predictions and projections about the future of the game as well as collate the data. In order to accurately predict the data reported by Wordle players in the future, based on the theory of time series analysis, combined with the extensive collection and screening of retrieval data, and the advantages of LSTM model and linear regression equation in the direction of prediction, a multi-dimensional prediction model for big data was established. With this prediction model, the development of Wordle games can be predicted according to a variety of prediction dimensions. After the accurate prediction of big data, the influential factors behind the data can be analyzed, which can simplify people's understanding of data to a certain extent, and successfully realize the transition from sophisticated technology to service-oriented demand.*

Keywords: *Machine learning, Time series analysis, Linear regression*

1. Introduction

Time series prediction methods mainly include statistical method and machine learning based method[1]. Time Series (TS), also known as dynamic data, refers to the sequence of the values of the same statistical index in order of their occurrence time, which reflects the changing trend of random variables over time.[2] In order to accurately predict time series data, machine learning, as the basis of artificial intelligence research, has absolute advantages in complex time series analysis [3]. As a related technology, time series prediction has become a hot field pursued by experts in academic research. Since time series analysis method and prediction are developed together, they play an important role in many fields[4]. However, for multivariable time series, there is often a mutual influence relationship between variables, which makes the analysis of time series data more complicated[5].Similarly, it is important to predict the number of reports on game data.

2. Basic model of LSTM

2.1 Document Description

There are actually varying numbers of letters in the word itself among the large amount of available data. Suppose the Wordle puzzle has five letters to fill in. First choose the word with five letters. On the basis that the number of letters meets the requirements of charades, the data of these letters are analyzed, and the LSTM time series prediction model is used for modeling and calculation.

Collection and preparation of data sets: The article study needs to collect historical time series data on variables related to outcome prediction. Then, it is necessary to preprocess the data, such as removing outliers, missing values, standardization, etc.

Separate training and testing: The data set needs to be divided into training and testing sets in the early stages of testing. Typically, the first 80% of the data in an article is chosen as the training set and the last 20% as the test set.

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

Preparing training data: The training of data in this article needs to divide the training set data according to the time step, and take the single variable value of each time step as the input of the LSTM model. For example, if you want to predict a variable and each time step has five values, the model operation can take the first four values as input and the fifth value as output.

2.2 Establishment of LSTM

In the process of model establishment, LSTM was chosen to construct the model because of the characteristics of large amount of data, high similarity of data and large number of permutations and combinations. Long Short-Term Memory (LSTM) is a machine learning algorithm based on recursive neural network (RNN). The recursive neural network transmits information through a unique gated design, and the information transmission mechanism is more sophisticated. [6] Because LSTM has memory function, it can use long sequence of information to build a learning model, which can reflect and speculate the future trend of data and the representation of hidden layer through LSTM.

Then, Combined with the multivariate regression model, the percentages of 1, 2, 3, 4, 5, 6 and X related to the future date are predicted through the form of multiple input and multiple output. Dividing the letters of each word into one to five letters (as shown in Figure 1) means that the characteristics of "w1, w2, w3, w4, w5" should be extracted from the data set.

Percent in

	Date	Contest number	Word	Number of reported	Number in hard mode	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)	Words	w1	w2	w3	w4	w5
0	2022/12/31	560	manly	20380	1899	0	2	17	37	29	12	2	m,a,n,l,y	m	a	n	l	y
1	2022/12/30	559	molar	21204	1973	0	4	21	38	26	9	1	m,o,l,a,r	m	o	l	a	r
2	2022/12/29	558	havoc	20001	1919	0	2	16	38	30	12	2	h,a,v,o,c	h	a	v	o	c
3	2022/12/28	557	impel	20160	1937	0	3	21	40	25	9	1	i,m,p,e,l	i	m	p	e	l
4	2022/12/27	556	condo	20879	2012	0	2	17	35	29	14	3	c,o,n,d,o	c	o	n	d	o
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
354	2022/1/11	206	drink	153880	3017	1	9	35	34	16	5	1	d,r,i,n,k	d	r	i	n	k
355	2022/1/10	205	query	107134	2242	1	4	16	30	30	17	2	q,u,e,r,y	q	u	e	r	y
356	2022/1/9	204	gorge	91477	1913	1	3	13	27	30	22	4	g,o,r,g,e	g	o	r	g	e
357	2022/1/8	203	crank	101503	1763	1	5	23	31	24	14	2	c,r,a,n,k	c	r	a	n	k
358	2022/1/7	202	slump	80630	1362	1	3	23	39	24	9	1	s,l,u,m,p	s	l	u	m	p

Figure 1. The first time to deform the words

At the same time, a serial number is assigned to the time and date during data processing. The linear regression model was used to further analyze and develop the data. In the second step, quantize the letters a to z within reason, such as a as 1, b as 2, and so on (as shown in Figure 2). This allows individual words to be quantified in order to build models. In the third step, the article uses known resources to build a new data set by analyzing the frequency of word searches and the frequency of meta-consonants in each word, so that the LSTM model can well cover the words in the data table to achieve better results. This also makes the model established in this article more suitable for Wordle's specific characteristics, so as to reflect the data more accurately and get accurate results.

Percent in

	Date	Contest number	Word	Number of reported	Number in hard mode	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)	Words	w1	w2	w3	w4	w5
0	2022/12/31	560	manly	20380	1899	0	2	17	37	29	12	2	m,a,n,l,y	13	1	14.0	12	25.0
1	2022/12/30	559	molar	21204	1973	0	4	21	38	26	9	1	m,o,l,a,r	13	15	12.0	1	18.0
2	2022/12/29	558	havoc	20001	1919	0	2	16	38	30	12	2	h,a,v,o,c	8	1	22.0	15	3.0
3	2022/12/28	557	impel	20160	1937	0	3	21	40	25	9	1	i,m,p,e,l	9	13	16.0	5	12.0
4	2022/12/27	556	condo	20879	2012	0	2	17	35	29	14	3	c,o,n,d,o	3	15	14.0	4	15.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
354	2022/1/11	206	drink	153880	3017	1	9	35	34	16	5	1	d,r,i,n,k	4	18	9.0	14	11.0
355	2022/1/10	205	query	107134	2242	1	4	16	30	30	17	2	q,u,e,r,y	17	21	5.0	18	25.0
356	2022/1/9	204	gorge	91477	1913	1	3	13	27	30	22	4	g,o,r,g,e	7	15	18.0	7	5.0
357	2022/1/8	203	crank	101503	1763	1	5	23	31	24	14	2	c,r,a,n,k	3	18	1.0	14	11.0
358	2022/1/7	202	slump	80630	1362	1	3	23	39	24	9	1	s,l,u,m,p	19	12	21.0	13	16.0

Figure 2. Quantify the words

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

2.3 Solution of LSTM

The results of this article are calculated using the established LSTM model to reasonably predict the number of results reported in the future and to predict the number of results reported on March 1, 2023.

Using the LSTM model analysis, the following graph is computed, which shows the trend in the number of results reported from March 2022 to January 2023.

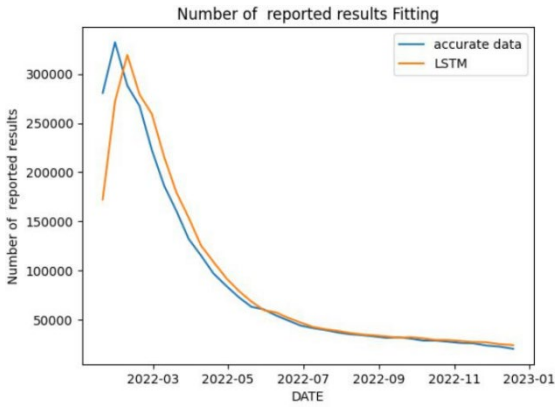

Figure 3. Number of reported results Fitting

It is not difficult to see from Figure 3 that the number of reported results was about 170,000 before March 2022, but was still on an upward trend and peaked at 300,000 in March 2022. But then there was a downward trend, with fewer than 50,000 results published around July 2022.

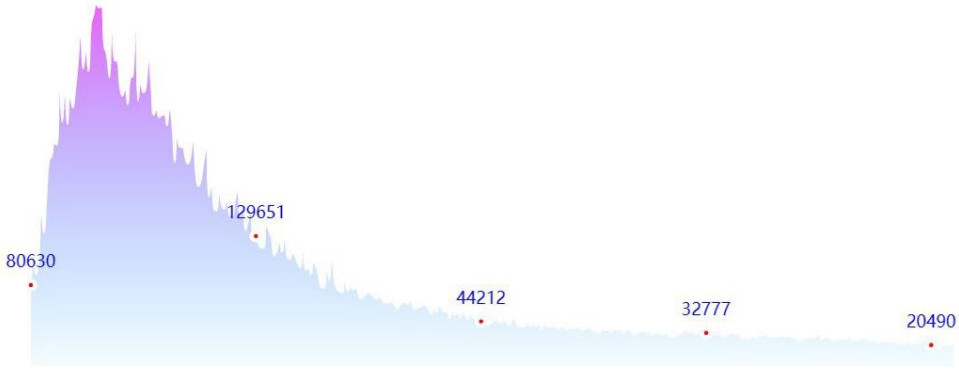

Figure 4. Report the number of results over time trend chart

In Figure 4, you can observe that the curve shown in the graph is not smooth and the data is not perfect. Therefore, on this basis, the article optimizes the accuracy and visibility of the graphics. On the premise that horizontal axis and vertical axis represent the same meaning, the article optimizes the graph to better reflect the changes in the data. The values of special points are shown in Figure 4, which makes the changes in values more intuitive and the curve smoother.

At the same time, the article also uses LSTM to build a model to predict the number of results reported on March 1, 2023. Results from the LSTM model operation show that the number of reported results on March 1, 2023 was 33,233. The mean absolute percentage error is 8.98%.

The root mean square error of the model is 0.901. Although time series correlation is taken into account, traditional statistical methods are not suitable for complex curves, and it is difficult to achieve the required prediction accuracy when dealing with non-stationary series. [7] In addition, the article estimates the probability of each letter in the word by calculating the frequency of consonant letters in the word, and uses this probability to predict the difficulty of the word "EERIE". Linear predictions are made by machine learning and other models. The percentage of "EERIR" obtained is

[0.46327684][5.77683616][22.67231638][32.97457627][23.68361582][11.5819209][2.81920904].

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

3. Basic linear regression model

3.1 Data Description

In many words, the article tries to solve this problem by extracting and distinguishing the different properties of words. The first word attribute chosen for analysis is frequency of use. During data collection, the article tries to use corpus query tools: there are some corpus query tools that can search for words in large corpora and see their frequency of use. COCA, for example, developed by Professor Mark Davis of Brigham Young University in the US, provides free access to a large and well-balanced corpus of American English, containing more than a billion words, including 20 million words per year from 1990 to 2019. The corpus is divided into eight genres, including spoken language, fiction, popular magazines, newsarticles, academic articles, TV and film subtitles, blogs and other web pages [8]. Using this huge lexicon, the frequency data of 219,000 words were downloaded from COCA, and these data were modeled and analyzed. Because of the large number of words, filtering starts with five-letter words from the large amount of data, and then searches and matches in Excel using the VLOOKUP function. Words with no frequency data were screened. Finally, the scoring rate combination in hard mode is calculated.

3.2 Establishment of linear regression model

The number of words selected by the data is too large, so a simple analysis is not feasible. However, in linear regression, the data is modeled using linear prediction functions, and unknown model parameters are estimated from the data. Linear regression is the first regression analysis that has been rigorously studied and widely used in practical applications, which can be used to study the correlation between two or more variables [9]. This is because a model with linear dependence on its unknown parameters is easier to fit than a model with nonlinear dependence on its unknown parameters. The statistical properties of the result estimation are easier to determine. In many practical problems, the traditional linear regression method will encounter many difficulties, which will lead to inaccurate prediction results, so we conducted two model training to reduce the error [10-11]. Using linear fitting method, this article explores the law behind these data, and makes an important analysis on whether lexical attributes affect the difficulty percentage of applicants through linear regression.

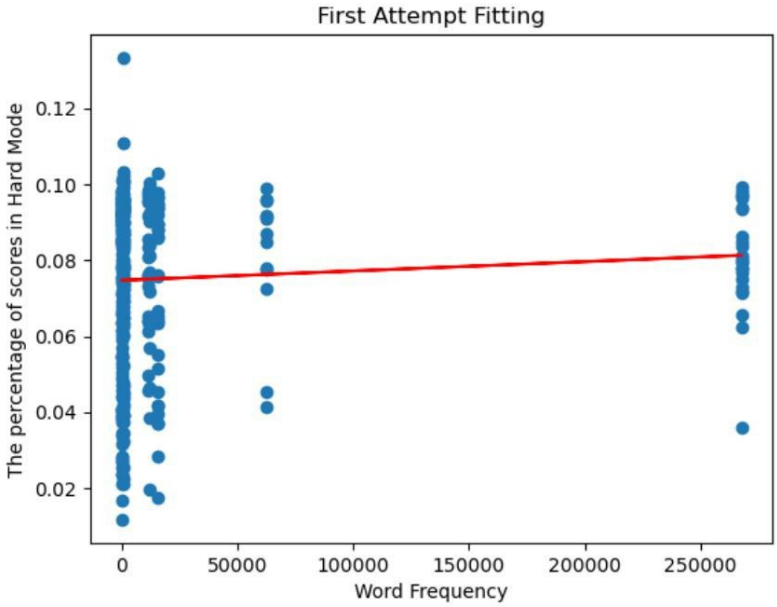

Figure 5. First Attempt Fitting

In this article, linear regression equation is used to fit the results. From the fitting method (as shown in Figure 5), first order fitting results are selected, and the equation fitting results with coefficient of 2.45317846e-08 and intercept of 0.07472318351816165 are obtained.

This first-order fitting sets the equation as

$$y = 2.45317846e-08x + 0.07472318351816165 \quad (1)$$

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

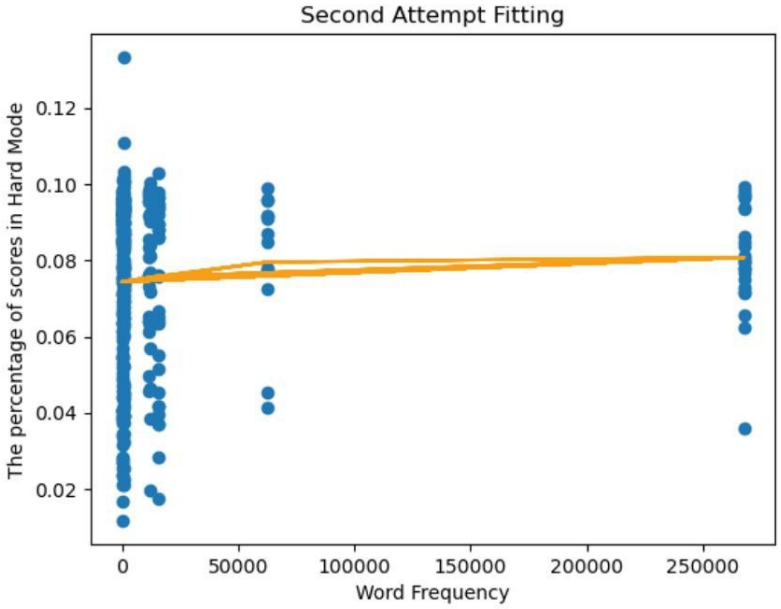

Figure 6. Second Attempt Fitting

As shown in Figure 6, in order to optimize the fitting results, the article attempted to conduct quadratic fitting of the data. The coefficients in the figure are 0.00000000 e + 00, 1.00692249e-07, -2.86653504 e-13, 0.0743783964607707. The first digit 0 corresponds to the coefficient of the constant term in X _; The second number corresponds to the coefficient of the first term (x); The third number corresponds to the coefficient of the quadratic term (X^2) in X _, which is the coefficient a. And then the last term is the intercept, the constant term c.

3.3 Solution of linear regression model

For a highly fitting linear regression model, the original intention of this article is to expect its actual value to fall on the fitting curve as much as possible, that is to say, the residual squares and RSS should be as small as possible. According to the calculation formula for R^2

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

R^2 should be as large as possible. Where Yi is the actual value, Y is the predicted value, the average Y is the average of all the scattered points, and R^2 is the square of R.

In order to limit the excessive characteristic variables, the concept of Adj.R-squared is introduced in the process of data fitting, which considers the number of characteristic variables on the basis of R^2. The formula is as follows

$$\overline{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{3}$$

Where n is the number of samples and k is the number of characteristic variables. When considering the number of characteristic variables, Adj. R-squared can reflect the fitting degree of linear model more accurately.

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

OLS Regression Results

Dep. Variable:	success rate	**R-squared (uncentered):**	0.117
Model:	OLS	**Adj. R-squared (uncentered):**	0.114
Method:	Least Squares	**F-statistic:**	44.88
Date:	Sat, 18 Feb 2023	**Prob (F-statistic):**	8.70e-11
Time:	17:58:52	**Log-Likelihood:**	405.05
No. Observations:	341	**AIC:**	-808.1
Df Residuals:	340	**BIC:**	-804.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
frequency	3.704e-07	5.53e-08	6.699	0.000	2.62e-07	4.79e-07

Omnibus:	79.797	**Durbin-Watson:**	0.244
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	136.229
Skew:	-1.351	**Prob(JB):**	2.62e-30
Kurtosis:	4.513	**Cond. No.**	1.00

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 7. OLS Regression Results

As shown in Figure 7, R^2 and Adj.R-squared were chosen to evaluate the model to measure the quality of the linear fit. Here R^2 is 0.117 and Adj.R-squared is 0.114, indicating that the linear fitting degree of the model is very low.

According to the fitting results obtained from the model, there is no significant correlation between the attribute of the word and the percentage of the score in difficult mode. However, if the answer to the puzzle is a word within the range of words used in everyday life, and the player is more familiar with it and uses it more frequently in everyday life, the puzzle will be solved more quickly. It's not so much about the attributes of the word as it is about the frequency and familiarity with which the word is used in the public eye. Conversely, words that have simple attributes but are not ordinary words decrease the reporting percentage. For example, the word "amino" translated into an amino group is not a difficult word attribute, but its cognitive range is not high, so it may account for a low percentage of reported.

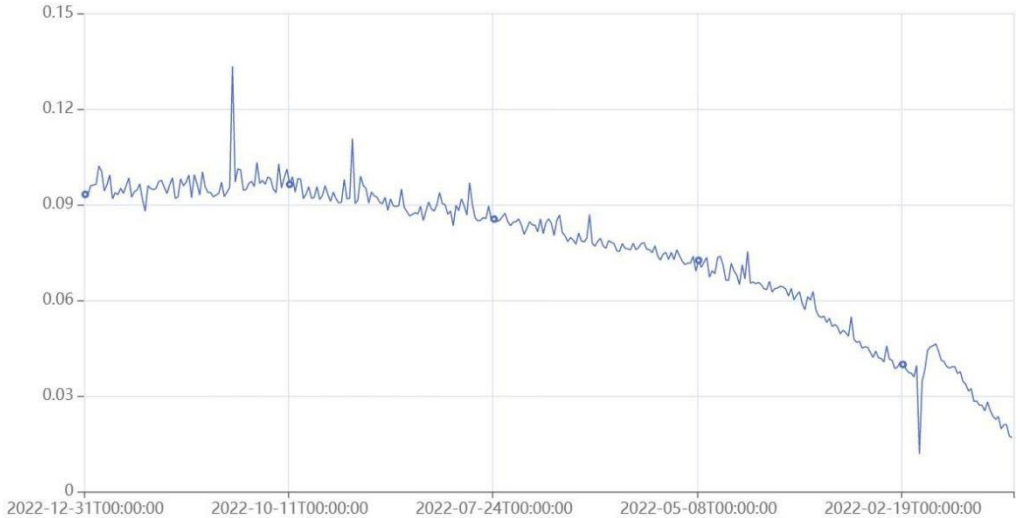

Figure 8. Score ratio over time in difficulty mode

Figure 8 is a chart of score ratios over time in difficulty mode. The article also guesses the reason for the appearance of such a chart, in the early background research found that this online game has not been developed for a long time, and is still new. It came out popular for a while, had a period of rapid growth, but then continued to decline in popularity. In hard mode, however, the game's scoring rate has increased, a trend the article attributes to older players or related puzzle enthusiasts continuing to play the game, trying to break personal records, seeking faster and more accurate solutions to problems, and challenging

ISSN 2616-5805 Vol. 4, Issue 2: 53-59, DOI: 10.25236/AJMS.2023.040208

themselves in hard mode. Therefore, there is no correlation between word attributes and score rate.

4. Sensitivity analysis and error analysis

(1) When using LSTM model, the average absolute error is 8.98%. To this end, the article summarizes the following reasons:

(2) In the research on the attributes of words, only a single attribute is considered in the data screening process, without considering the multidimensional attributes of words.

(3) The reported raw numbers and the actual number of people may not be equal (not everyone tweeted after the game), which is not ideal for the accuracy of the results.

(4) The data frequency has not been further processed and should be more related to the frequency of word use in 2022. However, it is hard to find data. More websites and databases have been visited during the data collection process to find as many relevant data clues as possible.

5. Conclusions

According to the game report of Wordle puzzle game and other data trends, it provides the basis for the establishment of the prediction model. Both the LSTM model and the linear regression model used in this article have the advantage of accurately describing the problem. Among them, LSTM model can effectively realize multi-input and multi-output of data, while linear regression model can help find the law of data in the fitting process, so as to better solve the problem.

Through the research findings: the article predicted the number of results reported by March 1, 2023 is 33,233, with an error of 8.98%. At the same time, the distribution of EERIE's prediction results is about [0.46] [5.78] [22.67] [32.97] [23.68] [11.58] [2.81] and the RMSE error is determined to be within 0.901.

Combining various attributes of vocabulary (among which the difficulty attribute is the most significant), this article further analyzes the word attribute problem, quantifies the number of words, and uses the linear regression model for training to calculate the fitting results of the correlation coefficient between the word attribute and the percentage of the score report in the difficult mode. The first time is 0.117, and the second time is 0.114. It is concluded that lexical attributes do not affect the percentage of scores obtained in difficult mode, that is, they do not constitute a correlation.

References

[1] Zhou Yehan. Research on Time Series Analysis Based on Deep Learning and its Application in Data Center [D]. Nanjing University of Posts and Telecommunications, 2022:1-65.
[2] Brochwell P.J, Davis R.A, Berger J.O, et al. Time Series: Theory and Methods [M]. Berlin, Springer-Verlag, 2015: 2-35.
[3] Jia Mingzhu. Research and Application of Time Series Analysis Method based on Machine Learning [D]. Xi 'an University of Science and Technology, 2020:1-77.
[4] Yang Yujun. Research on Time Series Model Based on Machine Learning and its application [D]. University of Electronic Science and Technology of China,2018:1-116
[5] Jiao Zinan, Chen Nian, Jin Tao, Wang Jianmin. Anomaly detection of Industrial Internet Time Series based on Spectral Residual method [J/OL]. Computer Integrated Manufacturing Systems, 2023: 1-21.
[6] Singh P, Dwivedi P, Kant V. A hybrid method based on neural network and improved environmental adaptation method using Controlled Gaussian Mutation with real parameter for short-term load forecasting [J]. Energy, 2019, 174(1): 460-477.
[7] Yuan Ximin, Huang Yuqi, Tian Fuchang, Cao Lugan. Prediction method of storm surge Water Increase based on LSTM-GM neural network model [J/OL]. Water Resources Conservation,2023:1-12
[8] Wu Jianping, Hou Ke. A comparative study of English Synonyms based on COCA Corpus [J]. Expo of Chinese Nationalities,2020,188(16):93-94
[9] Li Xiaohan. Analysis of the relationship between Marine economy development and transportation in China based on multiple linear regression analysis model. Transportation Energy Conservation and Environmental Protection,2023:1-7
[10] Shuai Wang, Yufu Ning, Hongmei shi. A new uncertain linear regression model based on equation deformation [J]. Soft Comput, 2021, 25(20): 12817-12824.
[11] Zhang Hanxia. Scenario analysis applicable to linear regression and logistic regression [J]. Automation & Instrumentation, 2022, 276(10):1-4+8.