|  |  |  |
| --- | --- | --- |
| **Problem Chosen**  C | **2023**  **MCM/ICM**  **Summary Sheet** | **Team Control Number**  2314306 |

**Analysis and Prediction of Wordle Results Based on Two-layer BP Neural Network and TOPSIS-SVM Support Vector Machine Comprehensive Classification Model**

**Summary**

In question 1, after data preprocessing,we first uses the time series forecasting model **ARIMA**, and selects the parameters (0, 1, 13) with the best fitting effect based on the known data set. The test correlation coefficient is 0.999, and it calculates the number of reported people on March 8 is about **26540.**Then, after collecting and analyzing the open source corpus, four attributes of words that may affect the results of wordle are obtained: the number of vowels, the number of repeated letters, word frequency and the frequency of the letters contained.Then we creatively use **Multiple Linear Regression Model**,usingthe distribution of the proportion of the number of answering attempts to be the independent variables, and the four attributes of the word being used as the dependent variables respectively.And then the F-t tests are carried out-- we get the correlation coefficients between the above attributes.After standardization, they are **0.9389, 0.3143.0.3044, and 0.1403** respectively.

In question 2, we use a **two-layer BP neural network** to solve the problem in order to avoid the problem of to many quantities.Based on the existing data set, we conduct the improved deep-learning based on **Levenberg-Marquardt algorithm** . We have tested multiple nodes for the output data, and the successful rate is very stable.Finally, we filter out the results whose total sum is closest to 100, and the proportions of each trial number of a given word are **0.9832, 7.638, 20.2508, 27.492, 26.1353, 15.3697, 2.2298**.

In question 3, we innovatively use the **TOPSIS-SVM support vector machine classification comprehensive model** to realize the classification based on difficulty of the words and analyses of the attributes of the words. We use TOPSIS to gain the weight vector, the optimal vector and the worst vector, and then input them into the SVM model,which quantifies and sorts the difficulty coefficient of all words in the data set.We selects 0.1 as the accuracy of the difficulty coefficient, divides the words with the same difficulty coefficient into one category, and obtains **the difficulty coefficient of eeire as 1**, which is the most possibly simplest word in the game.

In question 4, we first horizontally analyze the distribution of the number of guesses regarding to each word, and conduct a JB test on the data, and find that **the normal distribution** is satisfied, that is, the most people answer the question with four or five attempts; then we vertically analyse the difficulty coefficients,conducting the JB test and found that the difficulty distribution of all words also satisfies **the normal distribution**, that is, the number of words with a difficulty coefficient of 0.4 or 0.5 is the most.

**Key words: ARIMA Multiple Linear Regression Model BP neural network TOPSIS-SVM support vector machine classification comprehensive model**

Contents

[I. Introduction 2](#_Toc24023)

[1.1 Background 2](#_Toc15864)

[1.2 Restatement of the problem 2](#_Toc30788)

[1.3 Our works 3](#_Toc15626)

[II. The Description of the Problem 4](#_Toc18606)

[III.Basic assumption 5](#_Toc14605)

[IV.Symbols 5](#_Toc2682)

[V. Models 5](#_Toc22851)

[5.1 Analysis and Solving of Question 1.1 6](#_Toc25514)

[5.1.1 Model Preparation 6](#_Toc24672)

[5.1.2 Model Establishment 8](#_Toc22002)

[5.1.3 Results 10](#_Toc13261)

[5.2 Analysis and Solving of Question 1.2 10](#_Toc6517)

[5.2.1 Model Preparation 10](#_Toc31233)

[5.2.2 Model Establishment 12](#_Toc13896)

[5.2.4 Analysis of the Result 15](#_Toc6959)

[5.3 Analysis and Solving of Question Two 15](#_Toc23969)

[5.3.1 Model Establishment 15](#_Toc9015)

[5.3.2 Results 17](#_Toc27489)

[5.4 Analysis and Solving of Question Three 19](#_Toc32351)

[5.4.1 Model Preparation 19](#_Toc24699)

[5.4.2 Model Establishment 20](#_Toc3665)

[5.4.3 Results 21](#_Toc28229)

[5.5 Analysis and Solving of Question Four 22](#_Toc27053)

[VI. A letter to the Puzzle Editor of the New York Times. 22](#_Toc13382)

[VII. Evaluation and Promotion of Model 23](#_Toc27253)

[6.1 Strength and Weakness 23](#_Toc28877)

[6.1.1 Strength and Weakness 23](#_Toc30280)

[VIII. References 24](#_Toc24003)

[X. Appendix 24](#_Toc15169)

1. **Introduction**

## Background

Recently, the New York Times launched Wordle, an English guessing game that is popular all over the world. The rules of the game are as follows:

1) The player makes at most six attempts

2) Words are actual English words of five letters

3) The player's guess must be an actual English word that can be recognized as a word by the game

4) After each guess, the color of the block will change to give the player a hint: yellow means the letter in the block is in the word, but the position is wrong, green means the letter in the block is in the word and in the correct position, gray Indicates that the letter in the square is not contained in the word at all

5) Players can choose to play the game in normal mode or hard mode. If the player chooses to play in hard mode, once the player finds the correct letter in the word (the square is yellow or green, the word must be used in subsequent guesses)

The New York Times can receive feedback from players after publishing the topic every day, and count the number of players who got the answer after trying different times.

After analysis, it is found that the results of this answer show some interesting rules, and the research of this paper is mainly based on this

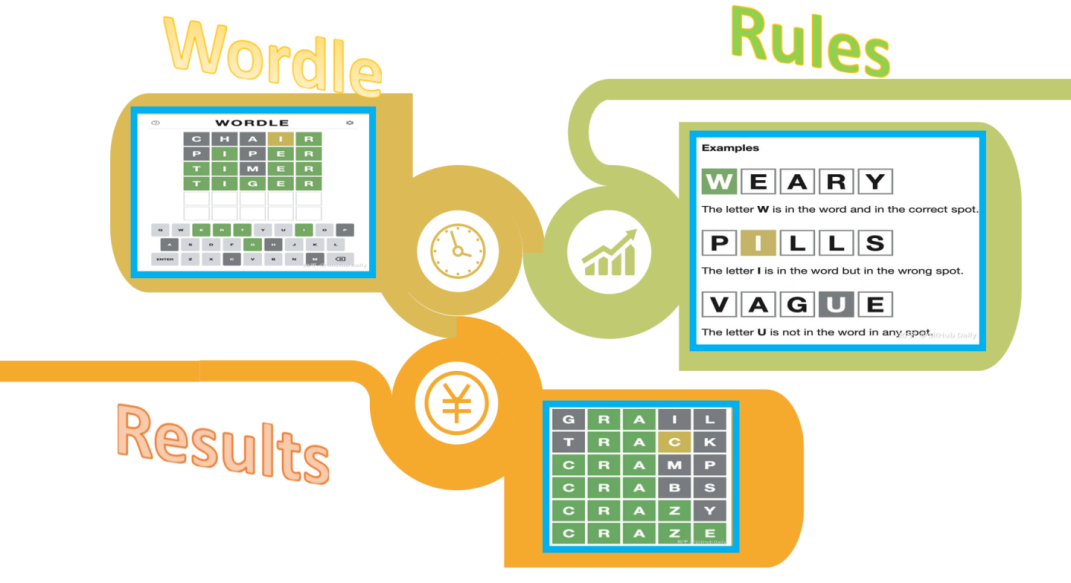


Figure 1:the introduction of Wordle

## **Restatement of the problem**

Based on the game results we received from players, we did a lot of research. After through in-depth analysis and research on the background of the problem, we can specify that our article should cover the following aspects:

1. Develop a model to explain the variation in the game's reported outcomes and predict the likely player's game outcome for March 1, 2023.

2. Analyze the percentage of word attributes for the distribution of results in difficult mode, and analyze the path of impact

3. Develop a model for predicting the distribution of reported outcomes given a word in the future (i.e. predict the relevant percentages of 1, 2, 3, 4, 5, 6, X word attempts for future dates)

4. Analyze the model, find out the uncertainty in the model, and test the feasibility of the model

5. Develop a model for classifying words by difficulty, identifying attributes of a given word associated with each class. Use this model to try to analyze the difficulty coefficient of the word EEPIE, and discuss the accuracy of the classification model

6. Discover other features of the dataset.

## **Our works**

* Task 1 Use Time Series Forecasting Model ARIMA to predict the number of people reported

Use the time series forecasting model ARIMA to predict the number of future reports, and fit it based on known data, and take the one with the highest fitting degree.

* Task 2 Using multiple linear regression to analyze the impact of different attributes of words on the difficulty
* Task 3 Using double-layer BP neural network to predict the proportion distribution of future answering attempts
* Carry out in-depth learning on the proportion of the number of people in the known data set, and then use the double-layer neural network to detect the set nodes. After getting better detection results, predict the distribution of the number of people who will answer the questions later, and filter out a set of data whose sum of the percentages closest to 100%.
* Task 4 Use the TOPSIS method to calculate the impact weight of each attribute on the word
* First process the data, sort out the attribute value corresponding to each attribute, then perform a weighted average attribute value for each attribute, and then sort
* Task 5 Use the support vector machine SVM to do the difficulty classification
* According to the sorted results, the difficulty of quantification is classified, and the difficulty coefficient is accurate to an appropriate level, and then the support vector machine is used to classify the words with the same value, and then the attribute similarities of such words are output.
* Task 6 Analysis of individual words and overall difficulty draws other interesting conclusions from the dataset

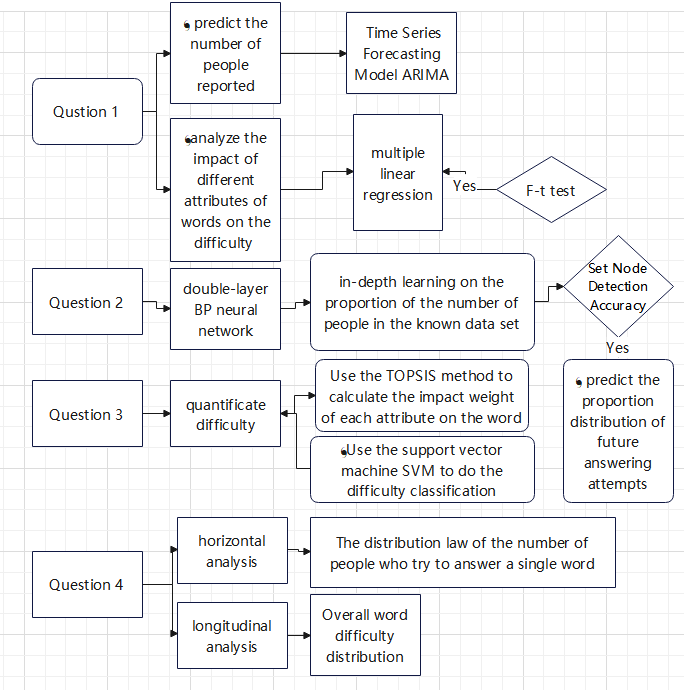


Figure 2:the mind map of our work

# II. The Description of the Problem



* For question 1,Predicting the number of people who will report in the future requires a time series forecasting model, and calculates the most appropriate parameter value using ARIMA based on multiple sets of data in the data set. After testing the model and getting a correlation coefficient close to 1, it can be used to obtain a given date in the future of reporting
* To detect the impact of word attributes on the prediction results, it is necessary to first determine which attributes may affect the guessed word. Then perform multiple linear regression analysis for each possible situation, and then perform F-t test to check whether this attribute can have a significant impact on the difficulty of guessing words
* For question 2,Carry out deep learning of the double-layer BP neural network for the known data set, and then detect the accuracy of the model by detecting the set multiple detection points. If the accuracy of the model is high, use the double-layer BP neural network for the future Prediction of the proportion distribution of the number of respondents
* For question 3,First use the TOPSIS method to obtain the weight of the above words, and then process each attribute of each word to obtain the attribute value corresponding to each attribute, and then perform a weighted average on the attribute value to obtain the difficulty coefficient, thereby completing the quantification of difficulty. Then sort the difficulty and select the appropriate precision. Use support vector machine SVM to classify words with the same annual coefficient and the characteristics of words in each category, and finally find out the category to which a given word belongs to determine its difficulty coefficient
* For question 4,(1) Observe the data set horizontally first, and find out the law of the distribution of the number of people who try to answer each word. Assuming that it satisfies the law of normal distribution, select enough sets of data for JB test.

1. Observe the data set vertically to find out the distribution law of the difficulty coefficient of all words. Assuming that it satisfies the law of normal distribution, select enough sets of data for JB test.

**III.Basic assumption**

* There is no obvious mutual influence relationship between the different attributes of words, that is to say, each attribute is independent of each other。
* Due to time and space constraints, only consider the impact of the three attributes proposed in this article on the analysis of this question

**IV.Symbols**

* For convenience,we introduce some important notations below.



|  |  |  |
| --- | --- | --- |
| ***Symbols*** | ***Definition*** | ***Units*** |
| **m** | Period length (take 12 for monthly data and 4 for quarterly data) | t |
|  | level smoothing parameter | \ |
|  | smoothing parameter of the trend | \ |
|  | Seasonal smoothing parameters | \ |
|  | Predicted value for period h | \K |
| **C** | Environment index of development | K |
| **F** | Impact index value | km3 |
| **S** | Reality index value | m |
|  | Influence coefficient | m2 |

Fugure 3:notations and symbols

**V. Models**



## Analysis and Solving of Question 1.1

* + 1. **Model Preparation**

**(1) Data Processing**

*  Remove six or four-letter statistics and keep only five-letter words
*  Due to missing percentages in the rounding process, only data with the sum of the percentages greater than 90% are considered
*  Too few data were removed from the statistics
*  For the missing values caused by the removal of the above flaw points, use the linear trend of the adjacent points to fill the missing values: the number of speaking periods is used as x, and the time series is used as y to perform regression, and the predicted value of the missing points is obtained

**(3)The Prepared elements of Model**

ACF autocorrelation coefficient and PACF partial autocorrelation coefficient

ACF: Under the condition that the time series is stationary

Sample autocorrelation coefficient:



({is the white noise sequence）

Check whether the residual is white noise:

{}is the white noise sequence，

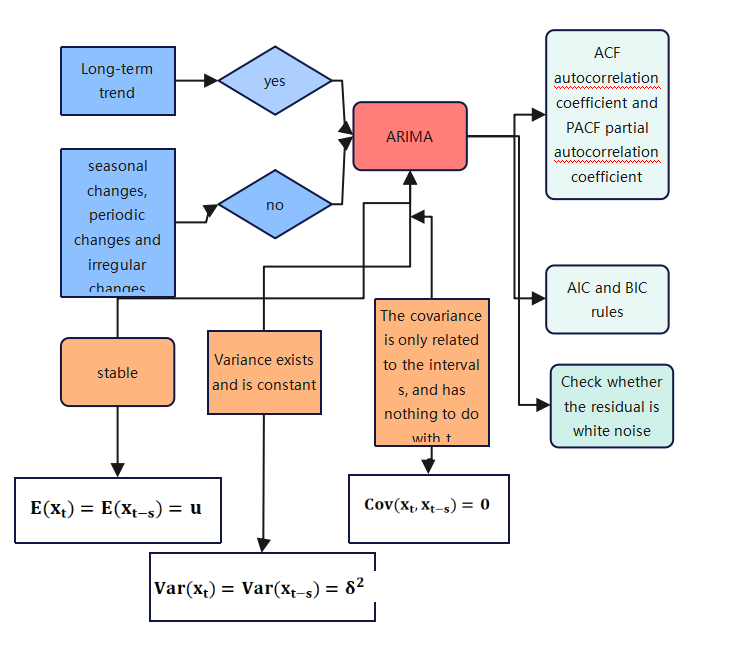
Under the condition that holds, the statistic

Q=T（T+2）

Based on data differential processing, it is transformed into a stationary sequence for modeling. We analyze the results based on the above principles, which conforms to the ARIMA (p, d, q) modelARIMA(p,d,q):

+

Build a differential autoregressive moving average model ARIMA.

Figure 4.the factors of ARIMA

### Model Establishment

**Step 1:**Model description-determine the value of p,q,k

Homogeneous part:

Transform the homogeneous part into a characteristic equation (algebraic equation)：let Simplify the homogeneous equation

|  |  |  |  |
| --- | --- | --- | --- |
| **Model description** | | | |
|  | | | model type |
| Model ID | V5 | model\_1 | ARIMA(0,1,13) |

Figure 5:the correct parameters of ARIMA

**Step 2:** use / BIC to measure the stationary of the model

Using a smooth measurement to compare the stationary part of the model with the simple average model, the result of the sum of the reported number of people at this time is 0.999, indicating that the fitting degree of the time series to the reported number of people is quite high.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **model fit** | | | | | | |
| Fit Statistics | average value | standard error | minimum value | maximum value | percentile | |
| 5 | 10 |
| Stationary | .931 | . | .931 | .931 | .931 | .931 |
|  | .999 | . | .999 | .999 | .999 | .999 |
| RMSE | 3612.004 | . | 3612.004 | 3612.004 | 3612.004 | 3612.004 |
| MAPE | 4.234 | . | 4.234 | 4.234 | 4.234 | 4.234 |
| MaxAPE | 27.954 | . | 27.954 | 27.954 | 27.954 | 27.954 |
| MAE | 2435.576 | . | 2435.576 | 2435.576 | 2435.576 | 2435.576 |
| MaxAE | 12619.430 | . | 12619.430 | 12619.430 | 12619.430 | 12619.430 |
| Normalized BIC | 17.196 | . | 17.196 | 17.196 | 17.196 | 17.196 |
|  |  |  |  |  |  |  |

Figure 6:the fitness of the model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Statistics** | | | | | | |
| Model | number of predictors | Model Fit Statistics | Young Box Q(18) | | | significant |
| Stationary | statistics | DF | significant |
| V5-model\_1 | 0 | .931 | 43.147 | 16 | .000 | 43 |

Figure 7:model statistic

**Step 3**: parameter estimates

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ARIMA model parameters** | | | | | | | | |
|  | | | | | estimate | standard error | t | significant |
| V5-model\_1 | V5 | do not convert | difference | | 1 |  |  |  |
| MA | delay 1 | .495 | .027 | 18.345 | .000 |
| delay 13 | -.338 | .028 | -11.943 | .000 |

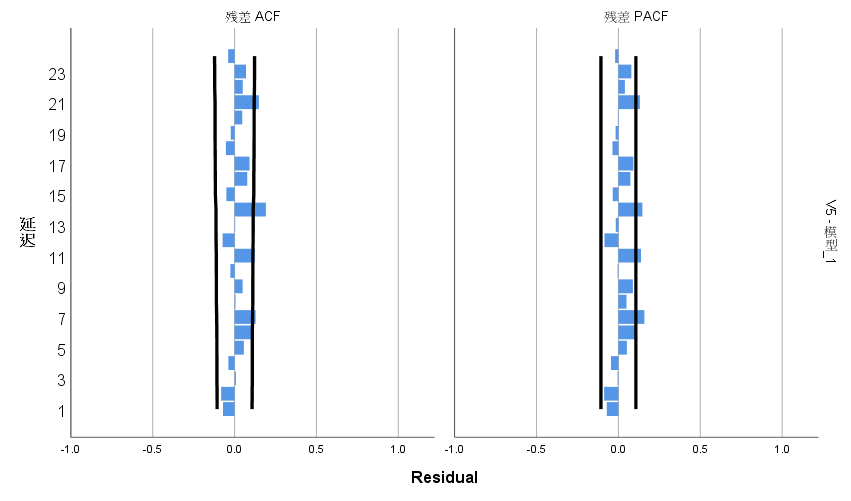
Figure 8:the model parameters

From the above equations, the value of the parameter can be obtained

**Step 3**：Residual testing using white noise

As can be seen from the ACF and PACF plots of the residuals below, the autocorrelation coefficients and partial autocorrelation coefficients of all lag orders are not significantly different from 0.

In addition, it can be seen from the above table that the p value obtained by the Q test on the residuals is 0.00, that is, we cannot reject the null hypothesis, and the thermal residuals are white noise sequences, so AMIRA can identify the reports in this question very well number of people.



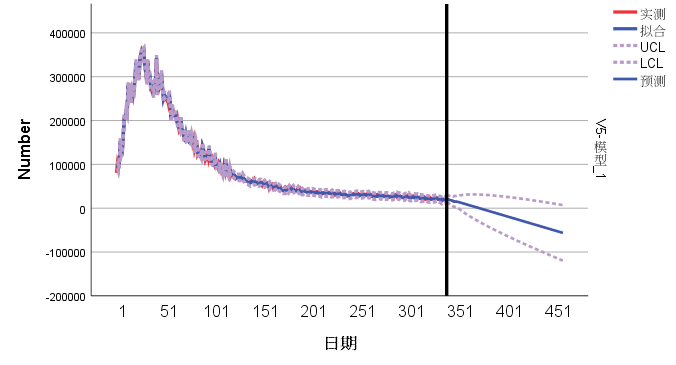
Fugure 9:the ACF and PACF tests

* + 1. **Results**

**Step 4:** output the result

It can be seen from the figure that the timing diagrams of the real data and the fitted data almost coincide, which shows that the AMIRA model has a good effect on the sum of the metadata, and it is in the confidence interval limited by the purple line (the confidence interval is 95%)

In addition, the predictive data independent variable is "day", if the 202-day question is used as the relative first question, it is necessary to predict the number of people who will report relative to the 398th day, a total of **26,540** people.

Figure 10.the predicted results of APIMA

## Analysis and Solving of Question 1.2

* + 1. **Model Preparation**

1. **Data Processing**

From the perspective of the participants' psychology and their game strategies, this paper finds that different words have different attributes that will lead to changes in the percentage distribution of the number of guesses. This article may consider that the main word attributes are: the number of vowels, the number of repeated letters in a word, and the probability of including letters. Now consider using multiple linear regression to analyze whether the above elements are related to the difficulty of the word, and then use the AHP to calculate the relevant weight to quantify the difficulty of the word

* The number of vowels: because most general counseling words have vowels, any experienced player will try words with more vowels. This article can guess the vowels contained in possible letters based on life experience The more letters there are, the less difficult the word is
* Number of repeated letters in a word: In addition, due to the limited number of guesses, the more repeated letters in a word, the easier it is to guess the word earlier, that is, the less difficult the word is
* Probability of containing letters: When the frequency of occurrence of different letters in all words is different, when the guessed word contains more high frequency letters, players are more likely to indirectly try out the correct answer through other words

**(2) Assumptions**

* There are only a few possible influencing factors of word difficulty
* For the measure of repetition of repeated elements in a word, it only depends on whether the letter can find the same letter in the word, not on what the exact same letter is

**(3) The Foundation of Model**

1. Use the test value (sample value) to make point estimates and hypothesis tests for the parameters and to establish a quantitative relationship between

2. Forecast and control the value of y, that is, make an interval estimate for y

3.the estimation of

Estimator by least squares method: do the sum of squared deviations

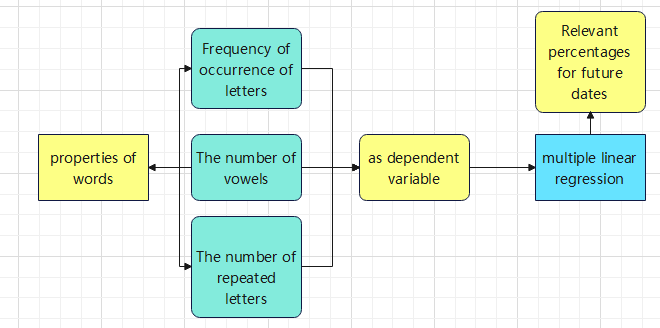
Choose to minimize Q

Solve for estimated value

The obtained value is put into the regression plane equation to get

0+1+……+k

At this time, y is the empirical regression plane equation, and i is the empirical regression coefficient



Fugure 11:the mind map of Multiple Linear Regression Model

* + 1. **Model Establishment**

**Step 1:**Taking the three attributes as the dependent variable and the percentage as the independent variable, respectively establish three quadratic multiple linear regression models

Suppose the data set satisfies the Gauss-Markov linear model (k-element linear regression model)，that is（Y，X，）

Y= , X=

get the regression plane equation  **Y=**

Suppose the observed value of the independent variable and the observed value corresponding to the dependent variable y satisfy the following relationship:

Among them， are random variables that are independent of each other and obey the normal distribution N(0,)

**Step 2：**Test the model goodness. From the results in the table below, we can see that the model can fit the data very well.

|  |  |  |  |
| --- | --- | --- | --- |
| **goodness of fit** | | | |
|  | Bangla | degrees of freedom | significant |
| pearson | 101555635153.956 | 576 | .000 |
| deviation | 447.896 | 576 | 1.000 |

Figure 12:the goodness of fit

**Step 3：**Bring the obtained results into the model building equations and solve the parameters:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **parameter estimates** | | | | | | | | | |
| V13a | | B | standard error | Wald | degrees of freedom | significant | Exp(B) | Exp(B) 95% confidence interval | |
| lower limit | upper limit |
| 0 | intercept | 274.264 | 274.915 | .995 | 1 | .318 |  |  |  |
| [Percent in =0] | 7.282 | 71.746 | .010 | 1 | .919 | 1454.354 | 1.237E-58 | 1.709E+64 |
| [Percent in =1] | 7.301 | 71.567 | .010 | 1 | .919 | 1481.201 | 1.789E-58 | 1.226E+64 |
| [Percent in =2] | 3.466 | 45.096 | .006 | 1 | .939 | 32.009 | 1.317E-37 | 7.781E+39 |
| [Percent in =3] | 9.835 | 74.199 | .018 | 1 | .895 | 18668.799 | 1.296E-59 | 2.690E+67 |
| [Percent in =5] | 15.707 | 78.806 | .040 | 1 | .842 | 6626034.824 | 5.518E-61 | 7.956E+73 |
| [Percent in =6] | 0b | . | . | 0 | . | . | . | . |

Fugure 13:the parameter estimates

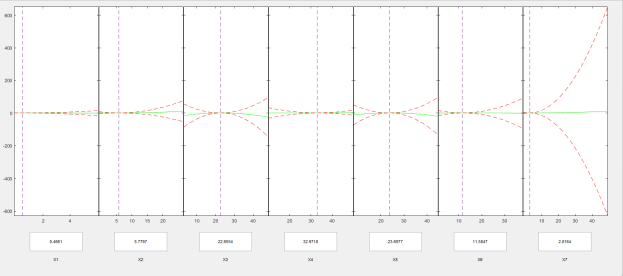
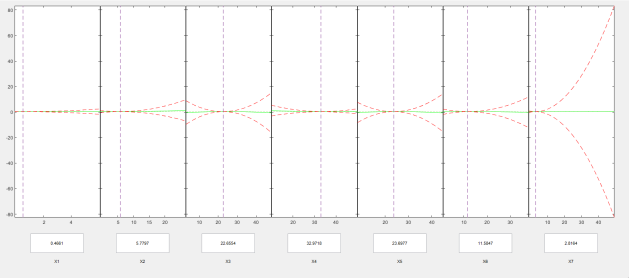
* + 1. **Results**

## 

Figure 14:the results of the Correlation of each of the four attributes

After entering the model according to the above data set, the correlation equation coefficient （betal) and error value (Rmse) can be obtained.

Visualize the data obtained above, and find out whether the three factors are related to the difficulty of word guessing by observing the image。Finally, it is only necessary to carry out the F-t test on the obtained results.



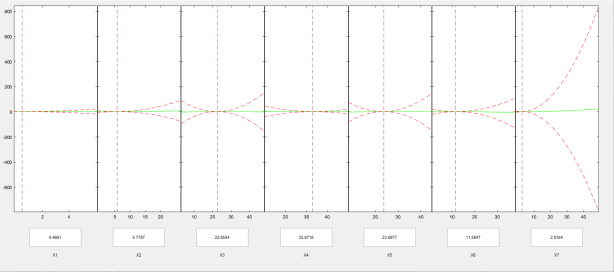
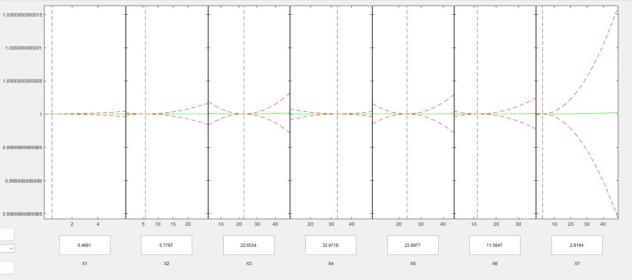


Figure 14：the F-t tests of the model

### Analysis of the Result

* **Linear model and regression coefficient test**

**Supposing**

**The F test was used to test the accuracy of the model.**

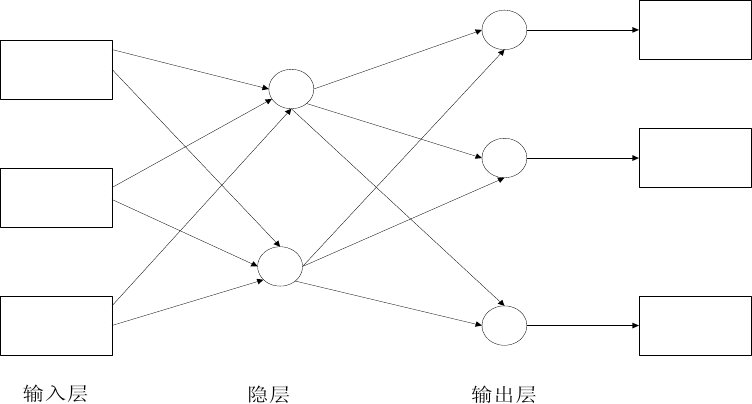
When established,

Then carry out F-test, and obtain F1=5.878742 F2=11.28712 F3=14.3308

When α=0.1, it is known from the t distribution table that the F-test is established, that is, the percentage is affected by the number of vowels in the word, the number of repeated letters in the word, and the frequency of letters in the word. These three factors, the model Evaluation is feasible

## Analysis and Solving of Question Two

* Introduction:Neural networks are widely used in classification and recognition methods. Through supervised learning during the training process, the weights of synapses are changed continuously through learning, and finally a simulation of the recognition system is completed. Due to the large number of recognition categories involved in this problem, it is necessary to choose a neural network with a multi-layer structure.

Figure 15:the illustration of the network

### Model Establishment

* The learning process of the two-layer BP neural network:

Step 1: Network initialization

The ontology makes a specific description of the model for the player's future answering situation, so the initial input node of the input layer is determined to be n=7; the number of hidden nodes is l; the number of output nodes is m=7, and the neural network between the hidden layer and the output layer is . The connection weight of the element is , hidden layer threshold, output layer threshold b=(b1,b2,……,)

Step 2: Calculation of implicit output

Among them, is the activation function and is the i-th output node variable

Step 3: Output layer input calculation of

=

Step 4: Weight update

=(t)+

Among them， is the efficiency of study，>0,;is Regulator，

Step 5: Threshold update

The model can be updated according to the error between the network output and the expected output

)

)

Step 6: Call related functions in MATLAB to solve the model until the end of the iteration, repeat Step 2 if it is not over

Based on the above steps and the established neural network model, predict the future answering situation.

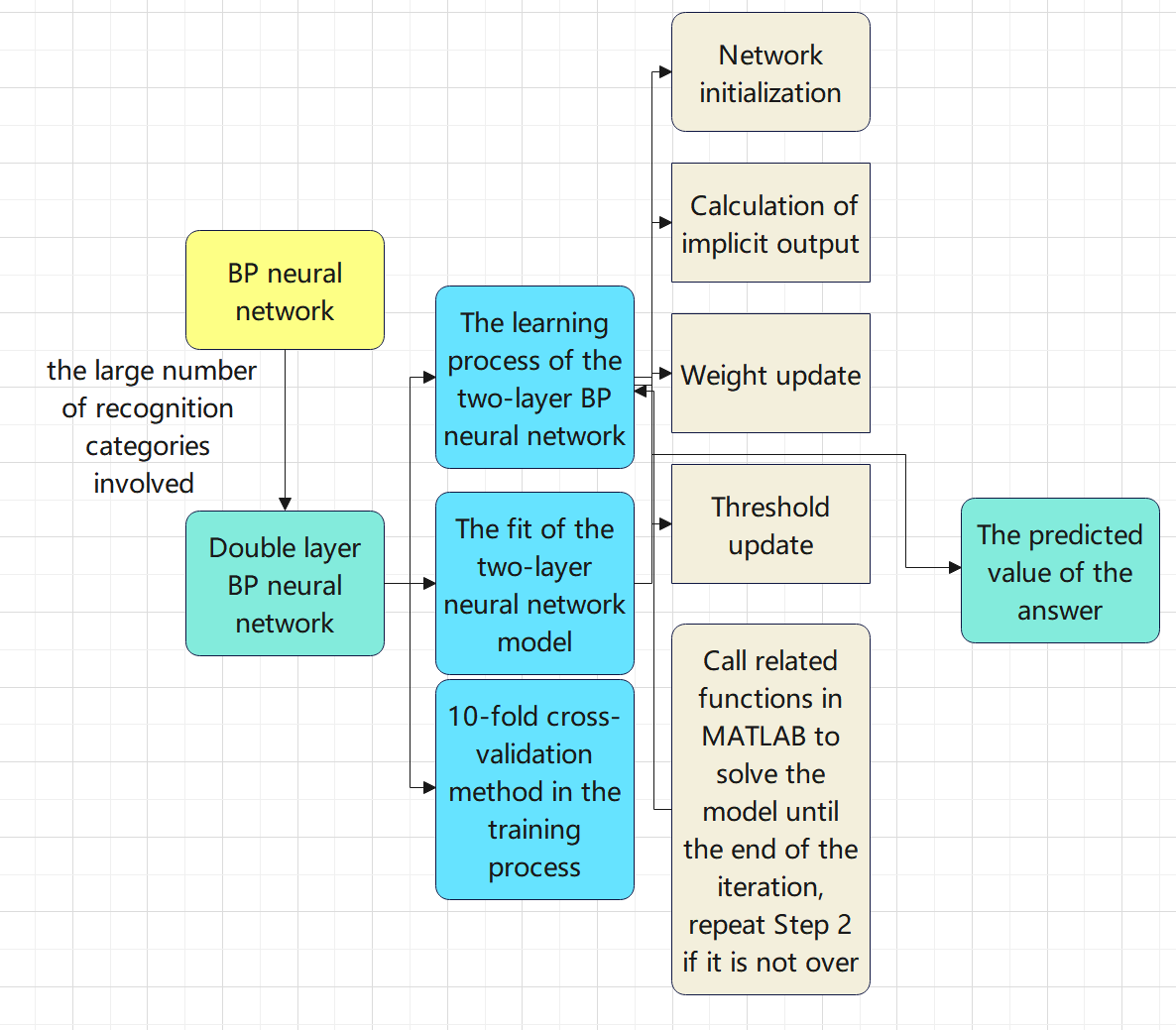


Figure 16 the illustration of the BP network

* The fit of the two-layer neural network model

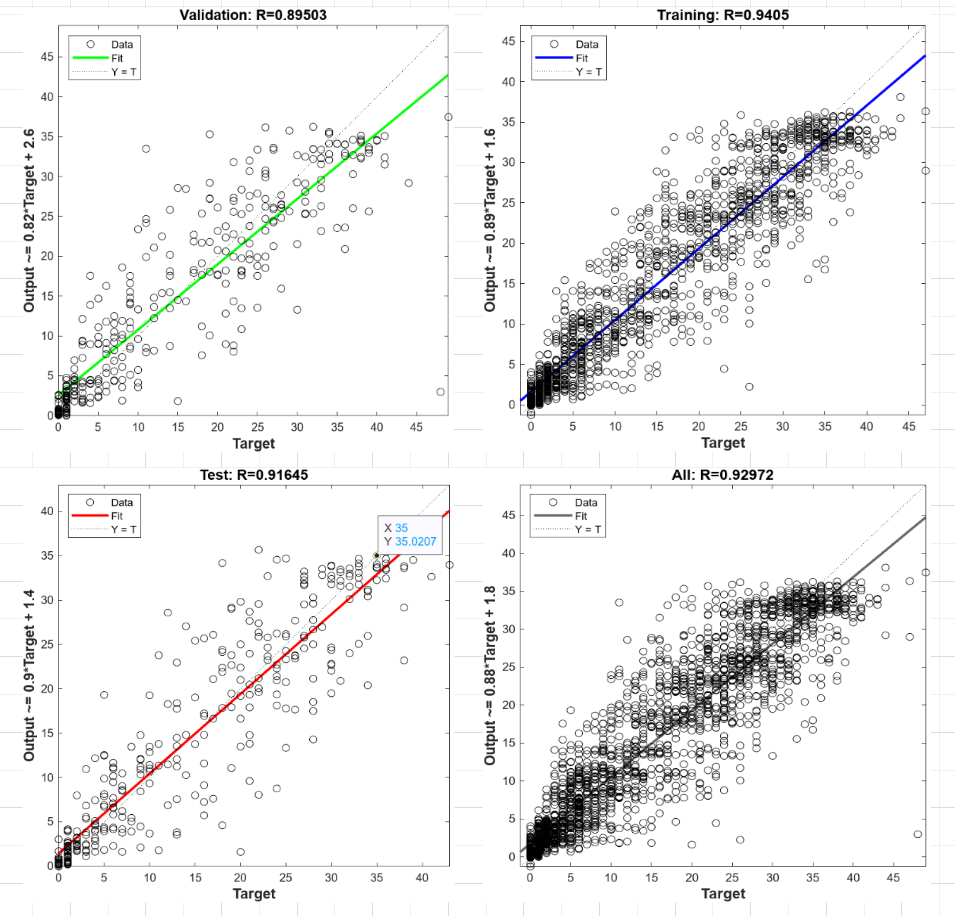
1. fitted value R：Calculate the value of R in the model to judge you and the situation. From the output 

Figure 17 the test of the BP network

results in the figure below, we can see that the fitting degree of the two-layer neural network model is very high (R1=0.89503,R2=0.9405,R3=0.91645,R4=0.92972)

### Results

We use a two-layer BP neural network to model and select 16 hidden nodes. In order to prevent overfitting, we use a 10-fold cross-validation method in the training process. For the original data set, we obtain its future The predicted value of the answer.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| data set | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 try | 0.0715 | 0.5213 | 1.4561 | 0.0196 | 0.6481 | 0.2779 | 0.1864 | -1.4377 | -0.3896 | 0.9832 |
| 2 tries | 11.0785 | 15.3004 | 8.0494 | 11.2602 | 10.4888 | 3.6522 | 13.5283 | 6.1447 | 6.8004 | 7.638 |
| 3 tries | 17.2932 | 30.6101 | 18.1985 | 24.316 | 22.4021 | 13.7152 | 17.097 | 15.589 | 21.2664 | 20.2508 |
| 4 tries | 18.882 | 28.4626 | 33.0186 | 35.3794 | 31.223 | 27.6123 | 39.4384 | 32.7723 | 32.2065 | 27.492 |
| 5 tries | 29.0333 | 17.6354 | 29.7734 | 25.6741 | 23.1789 | 29.4629 | 36.3278 | 31.9981 | 22.8964 | 26.1353 |
| 6 tries | 19.5795 | 6.7556 | 13.8814 | 14.9313 | 9.7193 | 18.1475 | 13.2448 | 13.3369 | 14.3088 | 15.3697 |
| X | 4.8695 | -2.076 | -0.5531 | 1.3362 | 3.252 | 7.1197 | -0.6134 | -0.8057 | 6.3839 | 2.2298 |
|  |  |  |  |  |  |  |  |  |  |  |
| sum | 100.8075 | 97.2094 | 103.8243 | 112.9168 | 100.9122 | 99.9877 | 119.2093 | 97.5976 | 103.4728 | 100.0988 |

## Figure 20:the output of BP network

The requested result is a percentage, which needs to satisfy the condition that the sum of the percentages is 100%, so after filtering, select:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 tries | 2 tries | 3 tries | 4 tries | 5 tries | 6 tries | X tries |
| rate | 0.9832 | 7.638 | 20.2508 | 27.492 | 26.1353 | 15.3697 | 2.2298 |

## Figure 21:the result of problem 3

## Analysis and Solving of Question Three

* + 1. **Model Preparation**

**I.Model Screening-TOPSIS Method**

**Step1**：Based on the evaluation index system established by question 1, a normalization matrix is established to standardize the data, that is

|  |  |
| --- | --- |
|  | (3) |

**Step2**：From the Z matrix obtained above, the optimal vector and the worst vector are obtained.

|  |  |
| --- | --- |
|  | (4) |

|  |  |
| --- | --- |
|  | (5) |

**Step3**：Compute the Euclidean distance from the optimal vector for each selected or constructed indicator for each country:

|  |  |
| --- | --- |
|  | (6) |

The Euclidean distance of each country's indicator from the worst vector:

|  |  |
| --- | --- |
|  | (7) |

**Step4：**Finally, the relative closeness to the optimal value is obtained:

|  |  |
| --- | --- |
|  | (8) |

**II.Support Vector Machine SVM Classification**

Support Vector Machine (SVM) [7] is a supervised method for pattern recognition and data analysis proposed by Vapnik et al. It constructs the optimal segmentation hyperplane in the feature space based on the structural risk minimization theory, so that the learner can be globally optimized, and the expected risk on the entire sample space meets a certain upper bound with a certain probability. Some studies have shown that SVM often has some advantages that other classifiers do not have, such as better performance in small samples and better generalization ability. Therefore, it is particularly suitable for the classification of words according to the difficulty coefficient in this paper.

SVM is often used to construct two-class classifiers, but it can also be used to deal with multiple classification problems, because here we only hope that SVM can output the class labels and results of classification. Therefore, this paper adopts the SVM classifier for multi-class classification.

### **5.4.2 Model Establishment**

**（1）weight determination**

For the triple index, the weight of the three is now measured and calculated. The idea of divide and conquer is now used, and the weight is calculated according to the results of pairwise comparison. Using the scale shown by the standard, compare two by two.

Since only a total of words are guessed every day, the attribute of the frequency of letters in words has a relatively small impact on the results due to the small sample size. In addition, from the perspective of game strategy, most players will choose a fixed word containing a large number of vowels for the first attempt, so the number of vowels is particularly important. Based on the above analysis, the constructed judgment matrix is:

|  |  |  |  |
| --- | --- | --- | --- |
|  | The number of vowels | Number of repeated letters in a word: | Probability of containing letters: |
| The number of vowels | **1** | **1/3** | **1/5** |
| Number of repeated letters in a word: | **3** | **1** | **1/4** |
| Probability of containing letters: | **5** | **4** | **1** |

## Figure 23:the weight of each attributes

1. e. get the matrix A =

**（2）The consistency check of the matrix is performed as follows:**

**Step 1**：Calculate the consistency index CL:

|  |  |
| --- | --- |
|  | () |

**Step 2**：Find the corresponding average random consistency index RI

**Step 3**：Calculate the consistency ratio CR

|  |  |
| --- | --- |
|  | () |

Running the code through MATLAB yields the following results:

|  |  |  |
| --- | --- | --- |
| The result of calculating the weight by the arithmetic mean method | The result of calculating the weight by the geometric mean method | The result of calculating the weight by the eigenvalue method |
| 0.9389 | 0.3143 | 0.1403 |

Figure 24:the three kinds of means of the model

It is concluded that the consistency index CI = -4.4409e-16, and the consistency ratio CR = -8.5402e-16.

Considering the robustness of the results, the consistency matrix can be obtained under the three weight analysis. Because CR<0.10, the consistency of the judgment matrix A is acceptable.

Finally, normalize the weights and simplify the data appropriately to get:

|  |  |  |  |
| --- | --- | --- | --- |
|  | The number of vowels | Number of repeated letters in a word: | Probability of containing letters: |
| **weight** | 0.9048 |  | 0.1552 |

Figure 25:the weight of each attributes

The specific process is as follows:

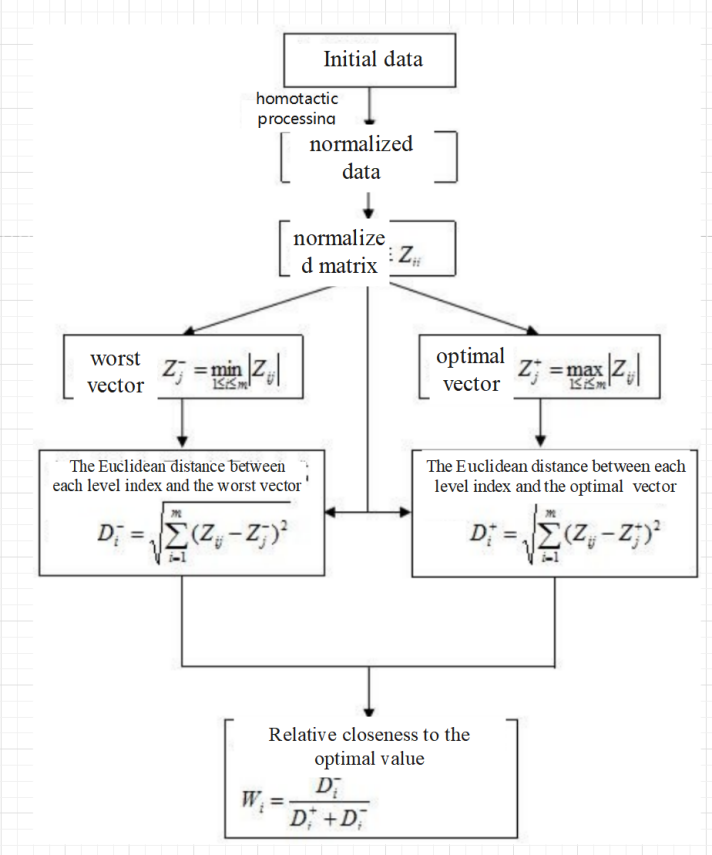


Figure 26:the mind map of the comprehensive model of TOPSIS and SVM

### Results

After obtaining the results of the support vector machine classification, proceed to the following steps

1) Use the weight to calculate the difficulty coefficient of quantifying each word

2) Difficulty sorting for each word

3) Take the precision as 0.1, and group for 0.1, 0.2..., 0.9, 1 respectively

4) Calculate the difficulty coefficient of eeire and classify

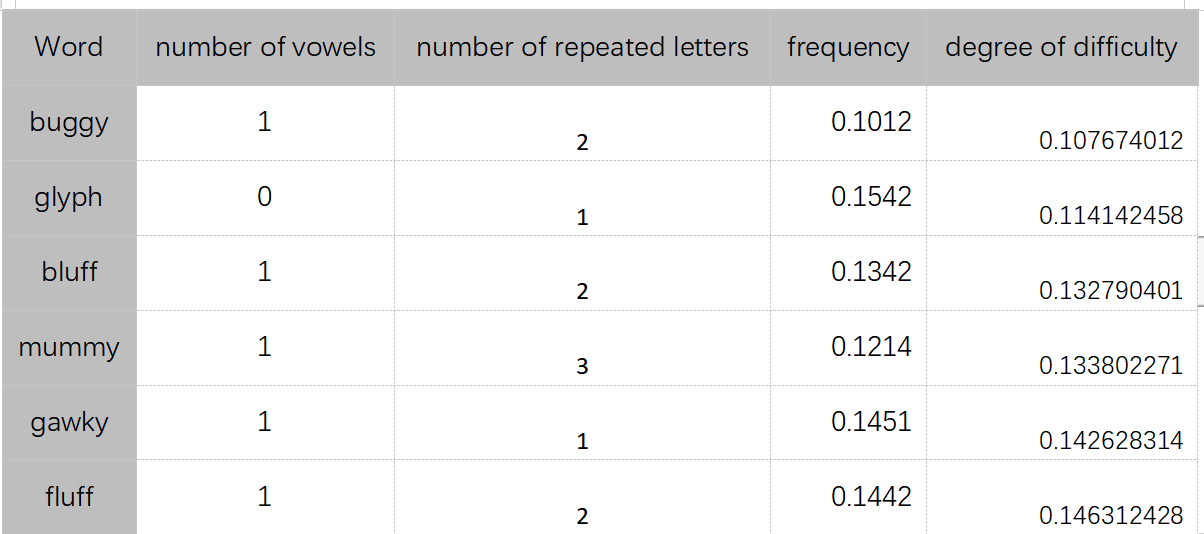
The following table takes the grouping with a difficulty coefficient of 0.1 as an example to show the grouping of the difficulty coefficient under this model. For the difficulty analysis and grouping of other words, see the attachment

Figure 27:Quantified difficulty for each word

Among them, after calculation, the difficulty coefficient of the word eeire under the weighted weight model is 1, which is the least difficult word in the game

## Analysis and Solving of Question Four

(1) Observe the data set horizontally first, and find out the law of the distribution of the number of people who try to answer each word.

(2) Observe the data set vertically to find out the distribution law of the difficulty coefficient of all words. Through the JB test described below, it is found that the above two laws satisfy the normal distribution.Normal distribution of difficulty coefficient: This paper conducts JB test on it after outputting the difficulty coefficient.For a random variable that is normally distributed, assuming its skewness is S and kurtosis is K, then we can construct the JB statistic:

JB=

It can be proved that if is a normal distribution, then in the case of n>30 (chi-square distribution with 2 degrees of freedom)

.**VI. A letter to the Puzzle Editor of the New York Times.**

Our team first started from the player's game strategy and sorted out the quantitative attributes of words. We learned on the Internet that most players have a fixed starting word, which generally contains multiple vowels, so the number of vowels is an important word attribute; in addition, considering the essence of this game It is a guess about the letters that make up a word. If there are many repeated letters in a word, the player has a higher probability of getting yellow and green letters. Therefore, the number of repeated letters is also a relatively important attribute; finally, due to different letters qualifying The probabilities are different, and the probability of being guessed is also different. Therefore, the influence of letter frequency on the game result should also be considered. Based on the above analysis, we get three attributes of analyzing game words.

For the first question, our team used the arima model to predict the number of players in the time series, and predicted the possible number of people participating in the game on March 1 as "". In addition, we tried to discuss the relationship between the three attributes and the percentage of difficulty through multiple linear correlation analysis, and finally came to the conclusion that the percentage is related to the three attributes concluded by our team, and all three multiple linear equations passed. f-test.

For the second question, our team established a bp neural network model, using the attribute data sorted out from the above attributes as the input of the model, and taking the 7 percentage data of each word as the output. after a lot of data

For question three, in order to facilitate subsequent expansion of the game word library and game development, we have established a word difficulty analysis model based on the Topsis model. The ranking position of the word is obtained from this to obtain the difficulty of the word.

# In order to verify the difficulty model, we substitute a very special word eerie. The word contains three e, and e is not only the word with the highest appearance in the word, but also a vowel letter, as long as one e is guessed, the data of three letters in the five letters can be obtained directly. Overall, the word is a very simple one.

# After bringing it into the model, we concluded that the difficulty of this word is also very low, so the model is reasonable.

**VII. Evaluation and Promotion of Model**



## Strength and Weakness

* + 1. **Strength and Weakness**
* ARIMA method+multiple linear regression:For the prediction of the number of future reports, time series prediction is mainly used. This method is accurate and effective, and the correlation coefficient of the fitted value is 0.999. The disadvantage is that the number of people forecast in this article is only based on the number of people in the past, and there are fewer factors considered
* Double layer BP neural network:Using neural networks to classify neurons according to spatial morphological features, on different data sets.Both can obtain quite high classification accuracy, and as the size of the data set increases, the performance of the classification can be improved. Therefore, the performance of the classifier can be improved by increasing the size of the dataset. However, this classification method also has certain defects. It does not have a clear classification standard, but the classification standard is hidden in the network structure, so it is difficult to intuitively display the key features of the input data from outside the network.
* Topsis method+SVM method:The advantage of the SVM method is that the physical meaning is relatively clear, and it has relatively good generalization ability. For this topic, it has similar pros and cons to neural networks, and only small-scale tests on the original and extended datasets show that SVMs perform slightly worse than neural network models. We suspect that this is due to the small and unbalanced dataset, and there may be some degree of overfitting in the neural network model. This requires validating the work on a larger test set.

**6.2 Promotion**

The main predictions made in this article are all based on algorithms that combine machine learning and text analysis. Therefore, if the ideas provided in this article are combined with NLP natural language recognition, it can be extended to more intelligent input methods. Based on word attributes After a variety of analysis, after entering a letter, predict what the next letter may be, and then predict the word the user wants to enter

# VIII. References

[1]Sanger, T., Probability density estimation for the interpretation of neural population codes, Journal of Neurophysiology, 1996.

[2]Ascoli, G.A., L-neuron:A modeling tool fro the efficient genetation and parsimonious description of dendritic morphology, Neurocoputing. **32**(33): p. 1003-1011, 2000.

[3]J. Suykens, J.V., The support vector method of function estimation, Nonlinear Modeling: Advanced Black-Box Techniques: p. 55-86, 1998.

[4]Wu, T.F.L., C.-J. Weng, Probability estimates for multi-class classification by pairwise coupling, JOURNAL OF MACHINE LEARNING RESEARCH. **5**(2): p. 975-1006, 2005.

[5]R.E. Fan, P.H.C., and C.J. Lin, Working set selection using the second order information for training SVM, Journal of Machine Learning Research. **6**: p. 1889-1918, 2005.

[6]P.Hamilton, A language to describe the growth of neurites, Biological Cybernetics. **68**: p. 559-565, 1

# X. Appendix

## 10.1 Code



%% Question 1

clear

clc

% 1.Multiple linear regression to determine correlation

% Get beta, rmse, residuals

% beta

% residuals

% rmse

% 2.1.y1

x = [x1, x2, x3, x4, x5, x6, x7];

rstool(x, y1, "quadratic")

beta

residuals

rmse

% 2.2.y2

rstool(x, y2, "quadratic")

beta

residuals

rmse

% 2.3.y3

rstool(x, y3, "quadratic")

beta

residuals

rmse

%% Question 2

% Get beta, rmse, residuals

% beta

% residuals

% rmse

% Clear the environment variables

clear;

clc;

% 1.Load input and output data

load("x1.mat")

load("x2.mat")

load("x3.mat")

load("x4.mat")

load("x5.mat")

load("x6.mat")

load("x7.mat")

load("y1.mat")

load("y2.mat")

load("y3.mat")

input = [y1 y2 y3];

output = [x1, x2, x3, x4, x5, x6, x7];

input\_train = input(1:330,:)';

output\_train =output(1:330,:)';

input\_test = input(331:354,:)';

output\_test =output(331:354,:)';

% 2.Solve an Input-Output Fitting problem with a Neural Network

% Script generated by Neural Fitting app

% Created 18-Feb-2023 20:02:25

%

% This script assumes these variables are defined:

%

% input\_train - input data.

% output\_train - target data.

x = input\_train;

t = output\_train;

% 3.1.Choose a Training Function

% For a list of all training functions type: help nntrain

% 'trainlm' is usually fastest.

% 'trainbr' takes longer but may be better for challenging problems.

% 'trainscg' uses less memory. Suitable in low memory situations.

trainFcn = 'trainlm'; % Levenberg-Marquardt backpropagation.

% 3.2.Create a Fitting Network

hiddenLayerSize = 10;

net = fitnet(hiddenLayerSize,trainFcn);

% 3.3.Setup Division of Data for Training, Validation, Testing

net.divideParam.trainRatio = 70/100;

net.divideParam.valRatio = 15/100;

net.divideParam.testRatio = 15/100;

% 3.4.Train the Network

[net,tr] = train(net,x,t);

% 3.5.Test the Network

y = net(x);

e = gsubtract(t,y);

performance = perform(net,t,y);

% 3.6.View the Network

view(net)

% 4.Plots

% Uncomment these lines to enable various plots.

%figure, plotperform(tr)

%figure, plottrainstate(tr)

%figure, ploterrhist(e)

%figure, plotregression(t,y)

%figure, plotfit(net,x,t)

%% Question 3

clear

clc

D = [1, 3, 5; 1/3, 1, 4; 1/5, 1/4, 1];

% 1.Check consistency

a\_max = max(eig(D));% a\_max is eigenvalue of maximum

disp(['λmax = ',num2str(a\_max)])

[~,n] = size(D);

CI = (a\_max-n)/(n-1);

disp('CI = ');disp(CI)

% RI is average random consistency index

RI = [0 0 0.52 0.89 1.12 1.26 1.26 1.36 1.41 1.49 1.52 1.54 1.56 1.58 1.59] ;

CR = CI/RI(n);

disp('CR = ');

disp(CR)

if CR<0.1

disp('\*\*\*\*\*\*\*\*\*\*CR<0.1,The judgment matrix passes the consistency test!\*\*\*\*\*\*\*\*\*\*')

else

error('The consistency test fails, please modify the judgment matrix')

end

% 2.Find the eigenvalues(E) and corresponding eigenvectors of the matrix D

% D is judgment matrix, E is eigenvalues, X is corresponding eigenvectors,

% A is diagonal matrix, w is weighted value

D = [1, 3, 5; 1/3, 1, 4; 1/5, 1/4, 1];

E = eig(D);

[X, A] = eig(D);

w = X(:, 1);

% 3.Standardize the data, for Z

% 3.1.Load data

load("Z.mat")

% 3.2.Standardize the data

S = zscore(Z);

tmp = S \* [w(1,:), 0, 0; 0, w(2,:), 0; 0, 0, w(3,:)];

cstar = max(tmp); % Positive ideal solution

c0 = min(tmp);

% 3.3.Get queue indicator value

for i=1:355

sstar(i)=norm(tmp(i,:)-cstar);% Find the distance to the positive ideal solution

s0(i)=norm(tmp(i,:)-c0);% Find the distance to the negitive ideal solution

end

f=s0./(sstar+s0);