Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2306318

Puzzle Game, Easy Solution

Wordle, serialized in the New York Times, was well praised upon its launch. Players had to complete the solution word within 6 tries using 5 letters. This does not sound like a difficult task. But is it really so? For this reason, we have conducted the following research and analysis.

For the first part, we create a prediction interval by building an **ARIMA** model for the number of reported results on March 1, 2023. And the prediction interval is **[17737,25379]**. Based on the characteristics of the words, we extracted a total of **5 attributes**, which are word frequency, orthography neighbors, number of different letters included, same letter words, and the degree of dissonance between phonology and writing. Through **Spearman correlation analysis**, we concluded that attributes of the word **affect** the percentage of scores reported that were played in Hard Mode. Specifically, word frequency and number of different letters included will bring **positive impact** for players.

For the second part, we systematically interpreted the distribution of reported results by building a **GA-BP neural network**. The model was applied to predict the associated percentages of "EERIE" on March 1, 2023. The predicted results were **(0.197, 1.326, 10.712, 32.537, 33.203, 18.045, 3.981)**. Its uncertainty was measured quantitatively using **Monte Carlo dropout**. Moreover， We also evaluate the models and the mean relative error of GA-BP neural network is stable at 0.1298, which indicates that the model has high measurement accuracy and superior prediction percentage distribution performance.

For the third part, we innovatively used the **DT-kmeans** for cluster analysis. The solution words by difficulty were eventually classified into **three classes**, namely easy, medium and difficult. We consider that this division can basically summarize the difficulty of solution word under more limited data conditions. When classifying EERIE, we classified it in the difficulty. Actuality, EERIE is at the lower level of the difficulty. When we analyze the accuracy of the clustering, we take two approaches to evaluate the accuracy of our classification model, **internal evaluation(DBI)** and **external simulation(Monte Carlo simulation based on decision theory)**, respectively. The value of DBI is **0.272** which is lower than 0.8, indicating our model has **high classification accuracy**. And the stimulation results show that the classification accuracy can reach 89%. In identifying the attributes associated with each difficulty level, taking the difficulty level as an example, we were surprised to find that more orthography neighbors, same letter words, and the dissonance between phonology and writing in a high level will **increase the difficulty**. However, word frequency, number of different letters included in a high level will reduce the difficulty.

For the final part, the effect of orthography neighbors on player scores, however, may either reduce or enhance player performance. In other words, its effect on player performance is **difficult to control**.

Keywords: ARIMA model; GA-BP Neural Networks；DT-Kmeans ；DBI；Monte Carlo dropout ；

Content

1 Introduction

1.1 Problem Background

The puzzle game is an old game with a long history of playability and is a pioneer for combining intellectual and educational content with games. Not long ago, software engineer Josh Wardle has given new life to this ancient game in the New York Times, and this new life is the so-called Wordle. The novelty of Wordle lies in its unique rules: within the limit of 5 letters of the word, each guess will give a certain feedback (filled letters not included in the result are shown in gray; filled letters is yellow if it is included in the result but not in the correct position; green if the letter is correct). You have a maximum of 6 attempts. Fig. 1 shows an example where the authors of this paper found the correct result in 4 attempts, and the corresponding results are shown in Fig. 2.

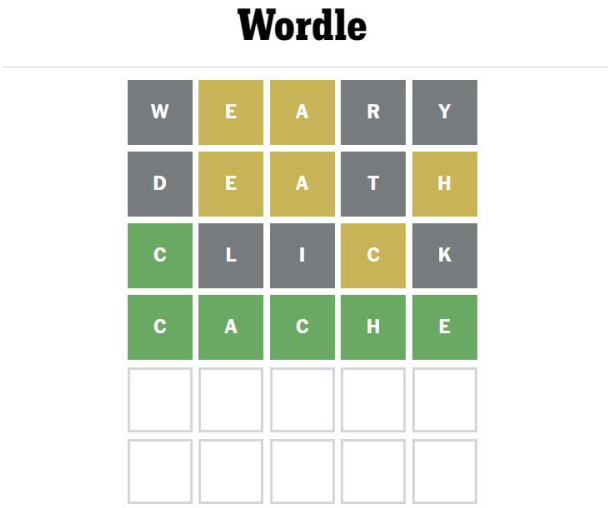

Fig. 1 the Game Page

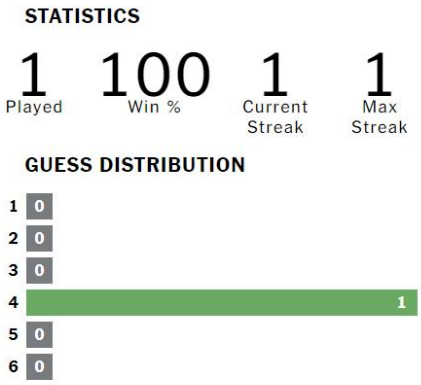

Fig. 2 the Output Page

1.2 Restatement and Our Work

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Explain the change in the number of results reported and create a prediction interval for the number of results reported on March 1, 2023. Explore whether any attributes of words affect the percentage of scores reported in hard mode.
- Predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date. Consider the uncertainty of the forecast results and Give a specific example of your prediction for the word EERIE on March 1, 2023. Evaluate the reliability of the predictions.
- Classify the solution words by difficulty and identify the attributes of the given word associated with each classification. Determine the difficulty of the word EERIE. Evaluate the accuracy of the classification.
- List and describe some other interesting features of this data set.

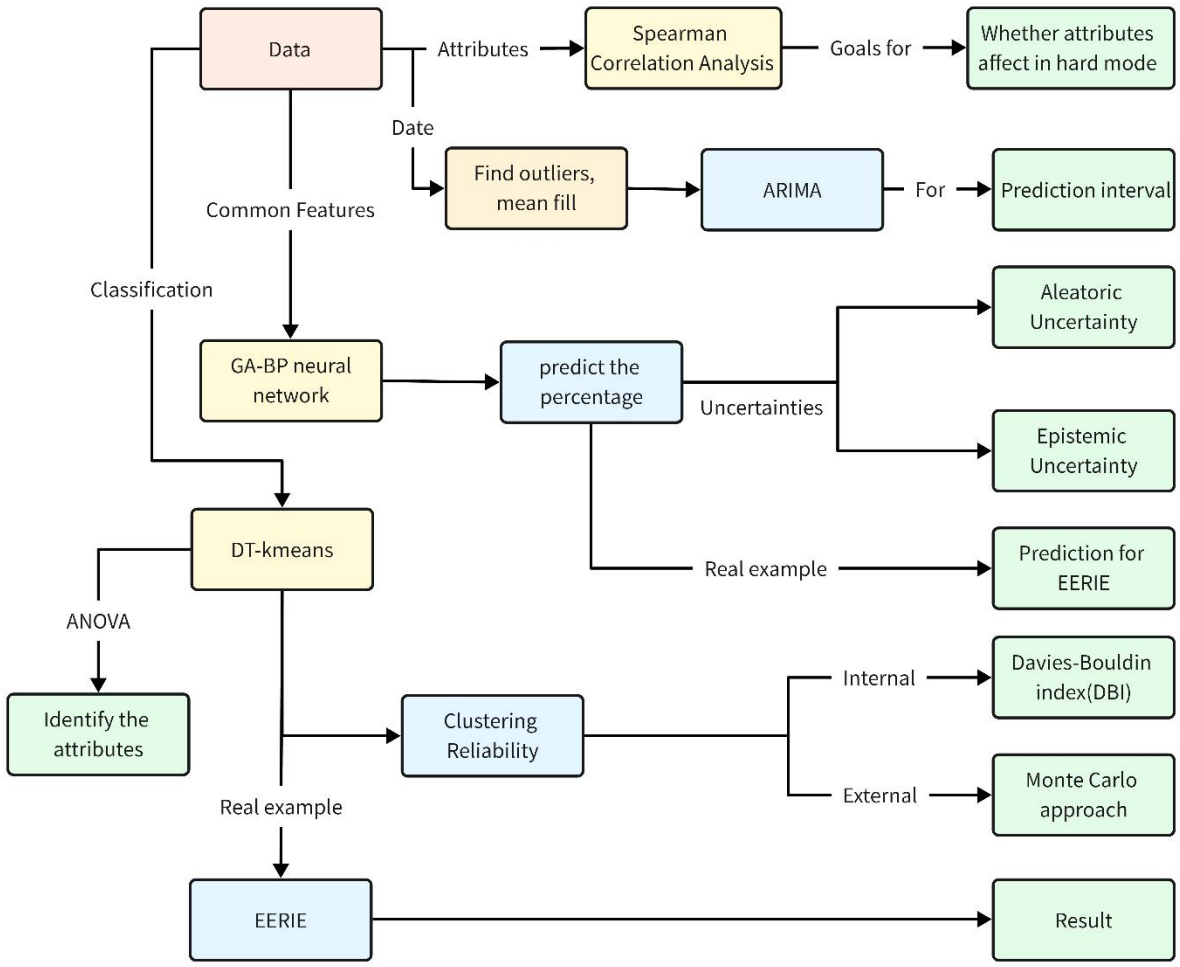

Fig.3 The flow chart in this paper

2 Assumptions and Notations

2.1 Assumptions

1. The game is stable in the short term and will not be taken off the market. The game's continuous operation is a prerequisite for this article to make predictions.
2. It is assumed that there is no cheating or prior knowledge of the answer.
3. We attributed the phenomenon of getting the correct answer on the first attempt to the simplicity of the word, eliminating the influence of the luck component as much as possible.

2.2 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description
D	Target data set
k	Number of clusters of classes to be clustered in the data set
Eps	neighborhood distance
η	neighborhood parameter
t	Data of objects in the set

3 Predicting Daily Results ARIMA Model

3.1 ARIMA Model

Box and Jenkins (1976) introduced the concept of ARIMA model [4]. That is, the model built by regressing the dependent variable on its lagged values only and the present and lagged values of the random error term in the process of transforming a non-stationary time series into a stationary time series. It is applicable to various fields of time series analysis, by considering the data series of the forecast object over time as a random series and approximating this series with a certain mathematical model. Once this model is identified it is possible to predict future values from the past values of the time series as well as the present values.

3.1.1 Preparation for Arima

Before performing data analysis on the ARIMA model, it should be ensured that the data are available and valid, so we perform data pre-processing.

The first step was to screen outliers. Data pre-processing of Problem C Data Wordle.xlsx using SPSS software revealed the existence of 2 additive outliers: STUDY (2022.11.30) and EXTRA (2022.12.25), respectively. To ensure the accuracy of the model, Their values are replaced by the mean values of the neighboring 6 samples.

3.1.2 The Establishment of Predicting Daily Results ARIMA Model

Converting dates into Serial Number starting from 1, i.e., 2022-1-7 is recorded as 1, 2022-1-8 is recorded as 2, and so on, the most recent date sample 2022-12-31 corresponds to 359, and 2023-3-1, which is to be predicted, corresponds to 419.

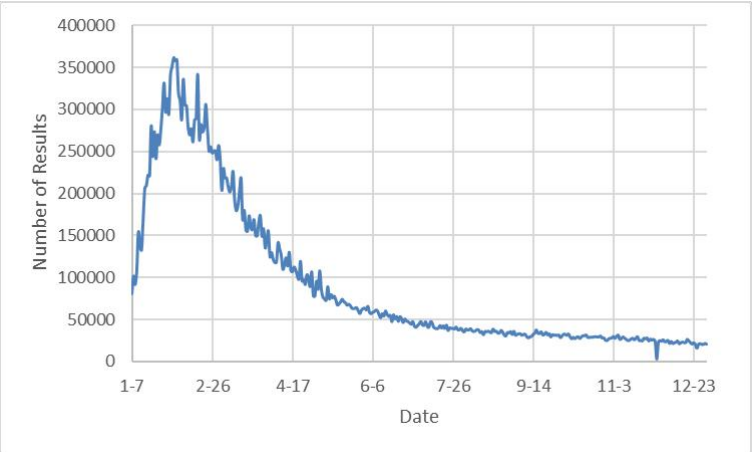

Fig.4 General trend graph of the number of results reported for the game

Based on the above graph we derive the trend of the number of reported results: in general, it shows an upward and then downward trend, and stabilizes after a period of time. Specifically, the number of reported results climbs sharply in the interval from 2022-1-7 to 2022-2-4, and thereafter begins to show a decreasing trend within a certain fluctuation range.

The optimal model is identified by SPSS as ARIMA (0,1,7) and the model is as follows：

$$\gamma_t = \gamma_{t-1} + 0.424\varepsilon_{t-1} - 0.135\varepsilon_{t-7} \tag{3.1}$$

t denotes the sequence number, and ε denotes a white noise sequence with variance σ^2. γ_t denotes the number of reports at number t.

Table 2： ARIMA(0,1,7) Model Assessment Results

Model	Model fit statistics			Ljung-Box Test		
	STATIONARY R^2	R^2	MAPE	Statistic	DF	Sig.
ARIMA(0,1,7)	0.813	0.982	4.329	24.567	16	0.089

The Ljung-Box significance in the model is greater than 0.05, and it reflected that this series residual fits the random distribution, while the Stationary R^2 and R^2 are greater than 0.8, indicating that the model has adequately identified our data and reflects a good fit of the data. The model has a high degree of feasibility. Mean absolute percentage error is 4.329 and the model predicts well.

Next, we use the ARIMA (0,1,7) model to predict the number of future reported results.

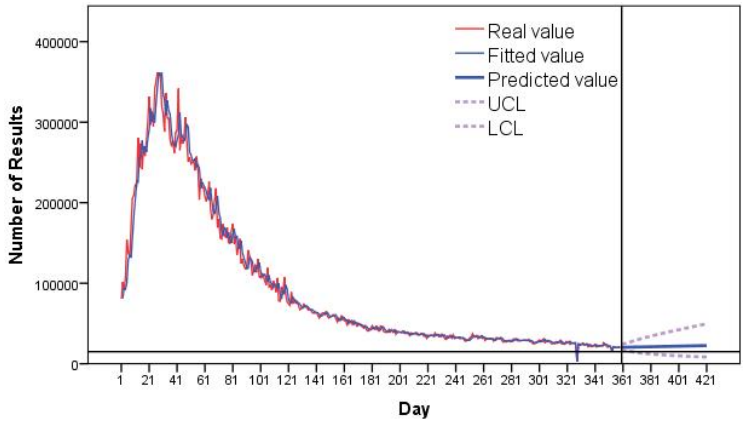

Fig 5 Predicted trend of the number of game report results

The results show that the number of reported results will stabilize at around 21,000 in the future, but there is still a possibility of renewed growth. The specific prediction of the number of results reported for future dates are as follows：

Table 3： Prediction of the number of results reported for future dates

Serial No.	Predicted value	LCL	UCL
…	…	…	…
416	21856	18000	25342
417	21787	17911	25355
418	21718	17823	25368
419	**21649**	**17737**	**25379**

…	…	…	…

The table intercepts the number of games reported with serial numbers from 416-419 . The prediction interval for March 1 (No. 419) is [17737,25379].

3.2 Spearman correlation analysis of the influence of attributes

Before performing Spearman correlation analysis, issues such as missing data and outliers should be fully taken into account, so data pre-processing was performed. We removed samples whose sum of percentages was outside [98,102], i.e., sample No. 281 generated on March 7, 2022; in addition, samples that did not meet the length of five letters, i.e., sample No. 314 generated on April 29, 2022, sample No. 525 on November 26, 2022, and sample No. 545 on December 16, 2022, were also removed. The remaining data were used as the data for the analysis in this paper.

3.2.1 Identification of Attributes of the Word

According to the characteristics of the word, we quantify word attributes from five perspectives: word frequency, orthography neighbors, number of different letters included, same letter words, and the degree of dissonance between phonology and writing. They are represented by X_1,X_2,X_3,X_4,X_5,respectively. The explanation of these attributes, the reasons for their selection, and the data sources are shown below.

Table 4 Identification of Attributes of the Word

Attributes	Description
Word Frequency (X_1)	● Explanation: The number of times the word is used. ● Reason: A higher word frequency tends to mean that the word is more commonly used and people are less likely to guess it incorrectly in Wordle games. ● Source of data: Based on how often the word appears in the Corpus of Contemporary American English (COCA)
Orthographic Neighbors (X_2)	● Explanation: Words that are the same length but differ by only one letter (e.g., SCARE and STARE)[2]. ● Reason: The more orthographic neighbors one has, the more words one has to choose from and therefore the more likely one is to guess incorrectly. ● Data source: Based on the number of orthographic neighbors of the word that appear in MCword.

Number of Different Letters Included (X_3)	● Explanation: A word contains several different letters. For example, the word every holds 4 letters, the word vivid contains 3 letters, etc. ● Reason: The fewer letters a word contains, the less useful Wordle's hinting mechanism is (e.g., If the player guesses the word VIVID, he can only get the information that the word contains the letters v, i, and d from the hinting mechanism ● Source: Python arithmetic recognition.
Same Letter Words (X_4)	● Explanation: i.e., different words made up of the same letters. For example, TEACH and CHEAT, BELOW and ELBOW, INAPT and PAINT, etc. ● Reason: The more words with the same letter, the more options there are to confuse people and the more likely people are to guess wrong. ● Source: Python arithmetic recognition.
The Degree of dissonance between Phonology and Writing (X_5)	● Explanation: Indicates the degree of consistency between pronunciation and writing of the word. For example, FAVOR is more consistent than AWFUL. ● Reason: The greater the degree of harmony between pronunciation and writing, the easier it will be for people to spell their desired outcome correctly. ● Calculation Formula： $$\mathrm{PR}=\frac{\mathrm{PN}}{\mathrm{VL}}$$ Where PR is the ratio of the number of letters in a word to the number of phonetic symbols, VL is the number of letters contained in a word, and PN represents the number of phonetic symbols contained in a word. According to the definition of PR, the PR value of all target words can be calculated[1].

Therefore we quantify the word attributes as follows:

Table 5：Word attribute feature extraction table

Word	X_1	X_2	X_3	X_4	X_5
gorge	1762	3	4	0	0.5
scare	8115	10	5	1	0.6
favor	14172	0	5	0	0.5
abbey	1335	1	4	0	0.5
vivid	4359	1	3	0	0.6
…	…	…	…	…	…

3.2.2 Spearman Correlation Analysis

To investigate whether the attributes of the words affect the reported percentage of

scores played in the difficult mode, we can perform a correlation analysis between the attributes and the associated percentages to explore the relationship. Common correlation coefficients are the Pearson correlation coefficient and the Spearman correlation coefficient. The Pearson correlation coefficient assesses the linear relationship between two continuous variables. It is more commonly used, but with more stringent conditions, as the data set has to meet specific conditions such as normal distribution, no outliers, and continuous variables. The Spearman correlation coefficient assesses the monotonic relationship between two continuous variables and is more relaxed than the Pearson correlation. The Spearman correlation coefficient is applied to data that are non-normal or non-continuous or are strongly influenced by outliers.

We first test the data for normality.

Table 6：Statistical description of the data and results of normality test

	Sample size	MD.	mean value	standard deviation	bias angle	kurtosis	S-W test
x_1	355	2280.5	26075.747	99265.055	7.815	72.473	0.255(0.000***)
x_2	355	2	2.802	2.548	1.126	1.214	0.889(0.000***)
x_3	355	5	4.706	0.481	-1.223	0.239	0.588(0.000***)
x_4	355	2	0.832	0.515	0.101	-0.979	0.894(0.000***)
x_5	355	0.56	0.5	0.242	-1.019	1.654	0.884(0.000***)
y_1	355	0	0.477	0.79	3.437	18.676	0.564(0.000***)
y_2	355	5	5.855	4.001	1.568	3.286	0.871(0.000***)
y_3	355	23	22.776	7.717	0.002	-0.475	0.992(0.060*)
y_4	355	34	32.951	5.269	-0.504	0.699	0.979(0.000***)
y_5	355	24	23.631	5.884	0.075	-0.193	0.994(0.227)
y_6	355	10	11.57	6.229	1.043	1.008	0.923(0.000***)
y_7	355	2	2.712	3.418	3.306	13.816	0.618(0.000***)

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

From this Table, we found that y_5 does not satisfy the normality prerequisite for conducting Pearson correlation, so Spearman correlation analysis is used.

Table 7：Spearman Correlation Analysis

variable	y_1	y_2	y_3	y_4	y_5	y_6	y_7
X_1	0.319 (0.000***)	0.352 (0.000***)	0.263 (0.000***)	-0.045 (0.403)	-0.283 (0.000***)	-0.228 (0.000***)	-0.236 (0.000***)
X_2	0.195 (0.000***)	0.153 (0.004***)	-0.03 (0.584)	-0.34 (0.000)	-0.121 (0.025*)	0.144 (0.007***)	0.276 (0.000***)
X_3	0.28 (0.000***)	0.464 (0.000***)	0.405 (0.000***)	0.038 (0.478)	-0.402 (0.000***)	-0.317 (0.000***)	-0.221 (0.000***)
X_4	-0.309 (0.000***)	-0.337 (0.000***)	-0.29 (0.000***)	0.002 (0.971)	0.293 (0.000***)	0.263 (0.000***)	0.271 (0.000***)

X_5	-0.213 (0.000***)	-0.237 (0.000***)	-0.212 (0.000***)	-0.013 (0.810)	0.213 (0.000***)	0.205 (0.000***)	0.226 (0.000***)

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively.

It was found that X1, X3, X4, and X5 had significant effects on y_1, y_2, y_3, y_5, y_6, and y_7, and X_2 had significant effects on y_1, y_2, y_5, y_6, and y_7. The results were visualized as above as follows.

variable	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	0.319	0.352	0.263	-0.045	-0.283	-0.228	-0.236
x_2	0.195	0.153	-0.030	-0.340	-0.121	0.144	0.276
x_3	0.280	0.464	0.405	0.038	-0.402	-0.317	-0.221
x_4	-0.309	-0.337	-0.290	0.002	0.293	0.263	0.271
x_5	-0.213	-0.237	-0.212	-0.013	0.213	0.205	0.226

Fig 4. Correlation coefficient color scale diagram

Based on the above Spearman correlation analysis we conclude the following:

① An increase in the values of X_1, X_3 will cause the values of y_1, y_2, y_3 to increase, while the values of y_5, y_6, y_7 will decrease, which will make the problem easier;

② The increase of X_4,X_5 will make the value of y_1,y_2,y_3 decrease, while the value of y_5,y_6,y_7 will increase, which will make the problem more difficult;

③ An increase in the value of X_2 will cause y_1, y_2, y_6, y_7.

For conclusions ① and ② we have given the explanation of the reasons when we selected the attributes, and the relevant analysis here further verifies our guesses, but a special phenomenon emerges, that is, the elevated value of X2 makes the topic appear polarized. The explanation of this conclusion will be presented in Part 6 (other interesting features of this data set).

4 GA-BP Neural Network for Predicting Percentage of Scores

4.1 GA-BP Neural Network

BP neural networks are characterized by self-learning, parallel distributed processing, nonlinear processing, and can avoid complex parameter estimation processes and reduce the influence of subjective factors [6,7]. Genetic algorithm is a stochastic search for optimal solutions based on the principle of genetic genetics, which is characterized by implicit parallelism, adaptive, optimized search space and global search for optimal solutions, and can reach the global optimal solution very quickly [8]. Therefore, in this paper, the two are combined to establish a customer mining model based on GA-BP neural network, as shown in Figure 6, and the modeling steps are as follows.

- Step 1: BP neural network initialization operation, that is, to determine the network input, output and the number of hidden layer nodes.
- Step 2: Set the population number and optimization objectives and calculate the fitness of the population.
- Step 3: Genetic manipulation of selection, crossover and variation.
- Step 4: Obtain the initial weight and threshold of the optimal neural network.
- Step 5: Train and test samples using optimal weights and threshold parameters.
- Step 6: Use the trained GA-BP neural network for prediction.

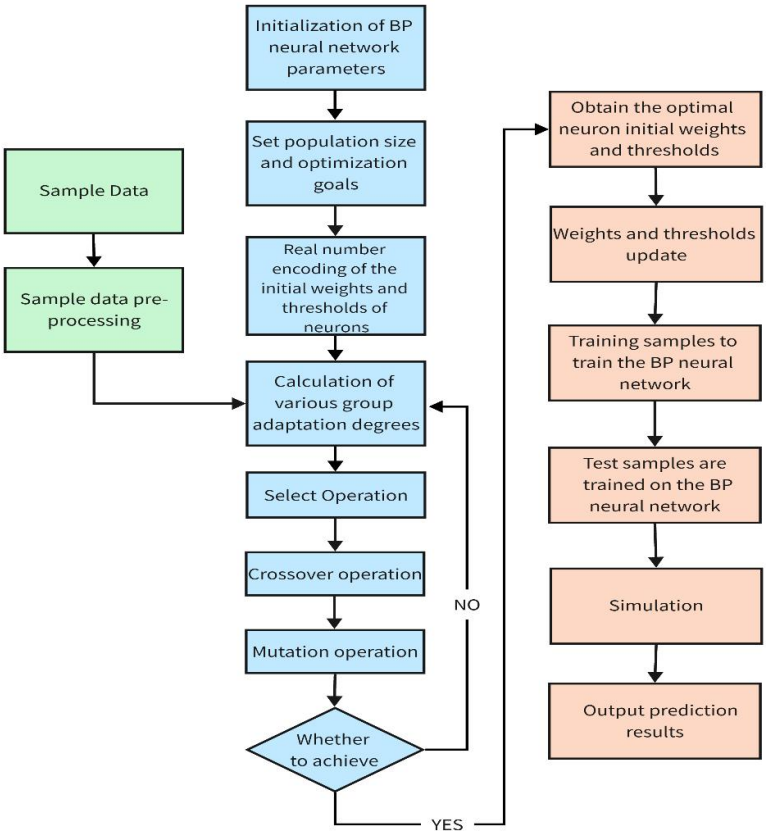

Fig.6 GA-BP neural network algorithm

4.2 Determine Training Samples, and Test Samples

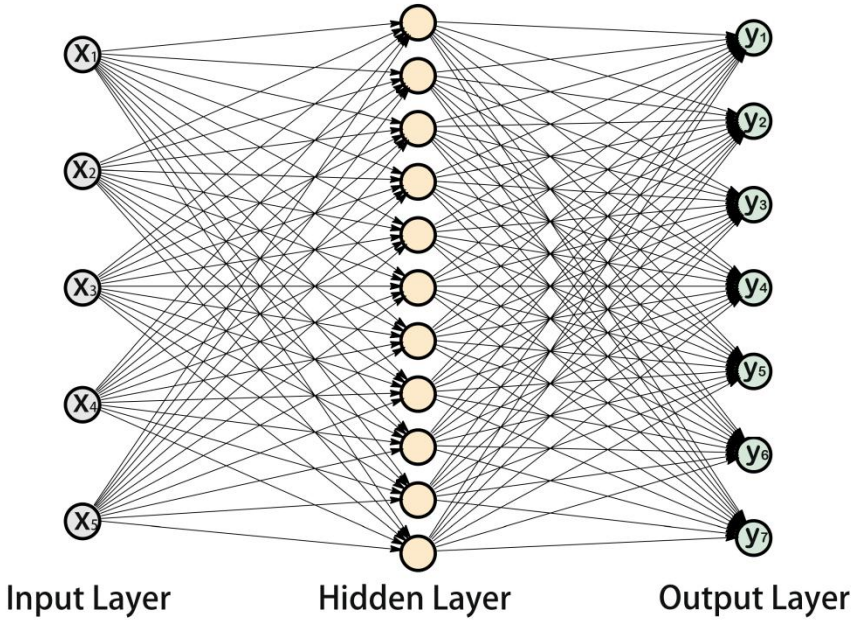

Fig.7 Neural network diagram

MATLAB software was applied to program the GA-BP neural network, and each of the five word attributes extracted according to Table 2 was used as a sample indicator; t The percentage of only one try in the report is defined as y_1, The percentage of only one try in the report is defined as y_2, and so on. y_7 indicated that the game participant did not arrive at a correct reported result in this round of the game. The GA-BP neural network model is thus constructed as shown in Fig. 7. Eighty percent of the 355 randomly selected samples were used as training samples (284 in total), and the remaining 20% were used as test samples (71 in total). Some of the samples are shown in Table 8.

Table 8：Input sample data of GA-BP neural network

Sets	X_1	X_2	X_3	X_4	X_5	y_1	y_2	y_3	y_4	y_5	y_6	y_7
	3012	7	5	1	0.6	1	7	19	22	19	18	15
	74861	0	5	0	0.5	1	5	18	30	26	16	3
Training	124739	3	4	3	0.6	0	5	21	32	26	14	3
Set	418	0	3	0	0.5	0	1	9	27	36	23	4
	5213	6	5	1	0.6	1	10	28	32	19	8	2
	…	…	…	…	…	…	…	…	…	…	…	…
	11575	4	5	0	0.5	1	9	29	33	19	7	1
Testing	473	7	5	2	0.6	1	4	17	30	27	17	4
Set	69783	2	5	1	0.6	1	7	29	35	20	8	1
	…	…	…	…	…	…	…	…	…	…	…	…

4.3 Associated Uncertainties

Regarding the uncertainty of models and forecasts, we believe that it is mainly reflected in two aspects, namely, aleatoric uncertainty and model uncertainty

1.**Aleatoric Uncertainty (AU)** is generated due to the noise in the observed data itself and usually stems from two aspects.

- Variability of the real-world situation: the real world is full of variability and randomness. The highly variable world makes it difficult for the data obtained from measurements to fully and accurately describe the reality. When the real-world situation changes compared to the training set, this can lead to large changes in the performance of the neural network.
- Inaccuracy and noise of the measurement system: Measurements are always based on certain premises and are susceptible to some noise. In this paper, for example, the data on the frequency of word usage is derived from the Corpus of Contemporary American English (COCA) and does not take into account the frequency of word usage by users in other countries.

2.**Model uncertainty (EU)** is generated due to uncertainty in model parameters, uncertainty in model structure.

Yarin and Zoubin proposed a method for approximating model uncertainty without changing the neural network structure or optimization techniques: **the Monte Carlo dropout**, which proceeds roughly as follows：

Step 1. Provide an input to the model.

Step 2. Multiple iterations can be predicted for a single input, each time randomly disabling a small portion of the neural network.

Step 3. Measuring the variance between iterations, which is the uncertainty of the model。

We set the loss rate to 0.5, i.e., each iteration of the neural network node has a 50% probability of being randomly disabled, and ultimately according to this theory we can express the model uncertainty in terms of the variance of the final output value：

$$\delta_i^2 = \frac{1}{N}\sum_{\mathrm{j}=1}^{N}\left(f^{\widehat{w}_\mathrm{j}}(X) - E(\widehat{y}_i)\right)^2, i = 1,2\ldots,7$$

δ_i^2 denotes the uncertainty of predicting y_i , X is the validation set, and $\widehat{w}_\mathrm{j}$ is the model parameter for each sampling . $E(\widehat{y}_i) = \frac{1}{N}\sum_{j=1}^{N}\left(f^{\widehat{w}_\mathrm{j}}(X)\right)$ refers to the mean of the predicted output.

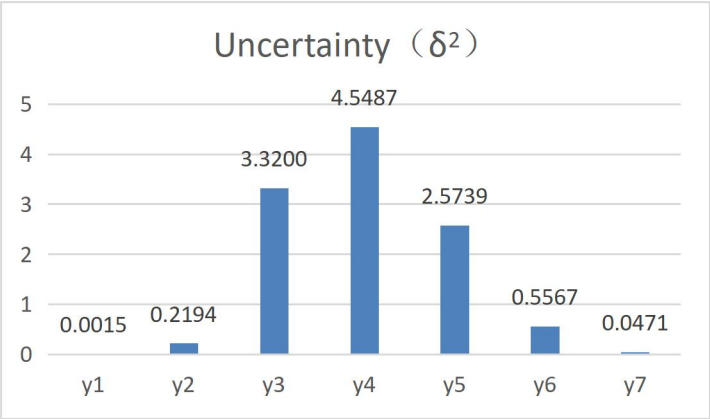

Fig.8 the uncertainty of Model

The results are shown in Fig. 8, where each bar height can be measured as the uncertainty of predicting y_i, where $y_4 > y_3 > y_5 > y_6 > y_2 > y_7 > y_1$。

4.4 Result prediction and reliability analysis

We use the trained GA-BP neural network model to predict associated percentages of (1, 2, 3, 4, 5, 6, X) of EERIE on March 1, 2023 In order to improve the representativeness of the prediction results, 20 predictions were made, and the average value was taken as the final result.

Table 9 Reported Results Predictions Chart

Number of predictions	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	0.3454	1.7351	9.9777	32.1192	32.3792	17.974	5.4694
2	0.0893	1.1319	11.159	32.4706	32.9297	18.4998	3.7196
3	0.0101	0.981	10.403	33.5028	33.3299	18.1928	3.5804
4	0.257	1.6215	11.2401	31.2463	34.478	18.3259	2.8312
5	0.281	1.1628	10.7787	33.3452	32.8988	17.2302	4.3033
…	…	…	…	…	…	…	…
Average	0.197	1.326	10.712	32.537	33.203	18.045	3.981

After 20 tests and taking their average, we see that the associated percentages of(1,2,3,4,5,6,X) for "EERIE" on March 1, 2023 are (0.197,1.326, 10.712, 32.537, 33.203, 18.045 and 3.981).

In order to judge whether EERIE has obtained reliable predictions, we analyzed the prediction effect using test sample data. The mean relative error(MRE) between the actual output and the ideal output of GA-BP neural network is calculated as the basis for reliability testing, and the results are shown in Figure 9 . From the calculation results, it can be obtained that after the 39th iteration is conducted, the MRE of GA-BP neural network is stable at 0.1298, which indicates that the model has high measurement accuracy, and the GA-BP neural network prediction percentage distribution model has superior performance.

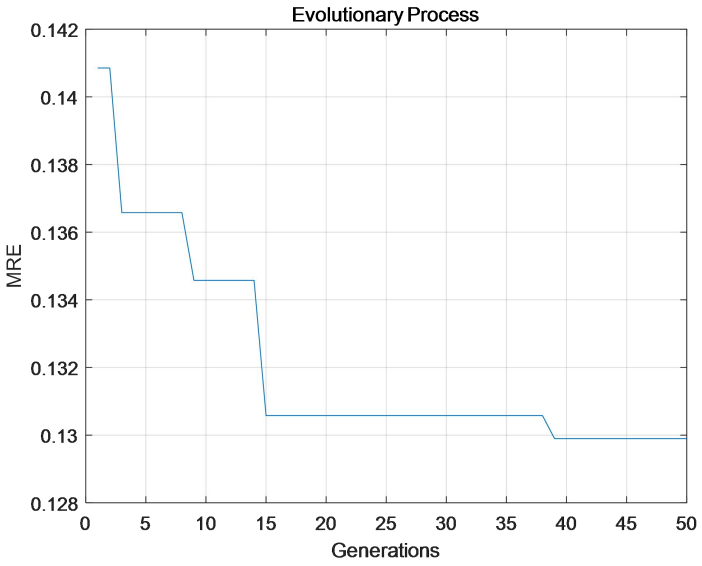

Fig.9 GA-BP Mean Relative Error

5 Difficulty classification based on DT-Kmeans

5.1 DT-Kmeans Algorithm

As a common clustering algorithm, K-means algorithm is fast and efficient. However, due to the random selection of the initial class cluster centers of the traditional K-means clustering algorithm, the clustering results of the algorithm are unstable and easily fall into local optimal solutions. To effectively solve this problem, this paper adopts an innovative K-means clustering algorithm that optimizes the selection of initial centers: the DT-Kmeans algorithm [5]. The computational steps are as follows：

1)Input a target data set D containing n data objects, the number of clusters to be clustered in the data set k;

2)The Euclidean distance between all data objects in the target data set D is calculated according to Equation (6.1) and stored in the distance distribution matrix D_{n*n};

$$d(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{i=1}^{m} (x_i - x_j)^2} \tag{6.1}$$

$x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}$, $x_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ denote the two data objects appearing in the Euclidean space R^m, respectively;

3)Based on the number of data objects n contained in the data set, the neighborhood parameter η of the neighborhood distance Eps is calculated according to Equation (6.2);

$$\eta = \left|\sqrt[4]{n}\right| + 1 \tag{6.2}$$

4)Based on the data set distance distribution matrix D_{n*n}, the distance parameter $d(xi\eta)$ of the ηth smallest is taken out in each row to obtain the distance array D_η;

5)Based on the distance array D_η, the neighborhood distance Eps is obtained by averaging the distance data in the array according to Equation (6.3);

$$\text{Eps} = \overline{D_\eta} = \frac{1}{n}\sum_{i=1}^{n} d(x_{i\eta}) \tag{6.3}$$

6)According to Equation (6.4) and Equation (6.5), the density information $\rho(x_i)$ of the statistical data object, i.e., the number of data objects in the data set whose Euclidean distance from this data object is less than or equal to the neighborhood parameter Eps;

$$\rho(x_i) = \sum_{i=1,j\neq i}^{n} u(Eps - d(x_i - x_j)) \tag{6.4}$$

$$u(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{6.5}$$

7) Define an empty set T, and put the information of data objects in the data set D with the dense information of corresponding data objects into the set T;
8) Define an empty set V to hold the class cluster centers;
9) Randomly select a data object from the set T and put it into V as the initial cluster center, and then remove the point from the set T;
10) The minimum value of the Euclidean distance between the data objects in the set T and the class cluster centers in the set V is counted according to Equation (6.6);

$$d(x_i, V) = \min\left(d(x_i, v_j)\right) \tag{6.6}$$

11) A data object from the set T is selected to be added to the set V of class cluster centers as a new class cluster center. For data object ti in set T, determine the weight $w(tj)$ of being selected as a class cluster center according to equation (3.9), the probability of data object ti being joined as $p = \frac{w(t_j)}{\sum_{t_{u\in T}} w(t_u)}$. Remove the data object being added to the set of class cluster centers V from set T;

12) Repeat steps 10 and 11 until the number of data objects in the set V is k;

13) Use the data objects in the set V obtained in step 12 as the initial class cluster centers of the K-means clustering algorithm to participate in K-means clustering;

14) Calculate the distance between each data object in data set D and the k class cluster centers, and assign the data object to the class cluster represented by the class cluster center with the closest Euclidean distance;

15) Count the information of data objects in each class cluster, take the mean value as the new class cluster center, and update the class cluster center information;

16) Iteratively execute steps 14 and 15 until the algorithm class cluster centers no longer change;

17) Output clustering results: k independent class clusters:$C = \{C_1, C_2, \cdots, C_k\}$.

5.2 Classify solution words by DT-Kmeans

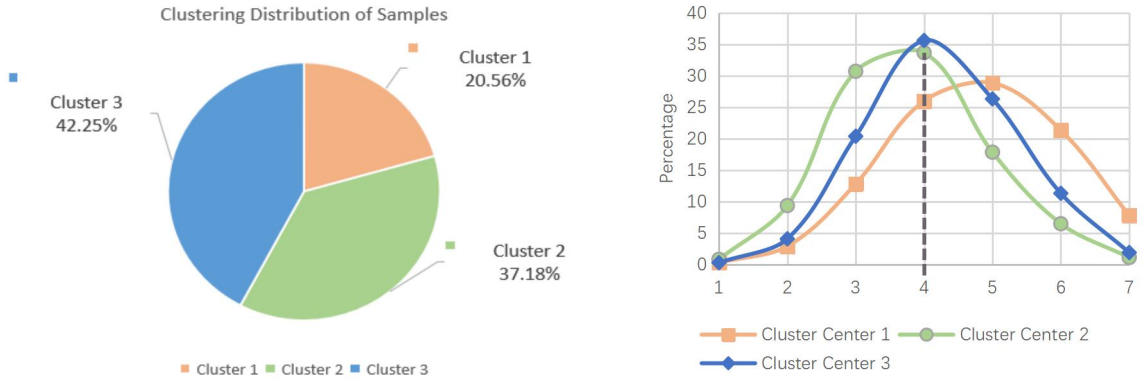

Fig.10 Distribution of Samples and Characteristics of clustering centers

According to the percentage distribution of the cluster centers above, when we take 4 as the dividing line between difficult and easy, cluster center 3 is characterized by 4 as the symmetry axis, showing an approximately normal distribution, so the difficulty of the words belonging to its class is medium; cluster center 1 obviously shows a negative skew, and most of the words belonging to its class need more than 4 times to be answered correctly, so the difficulty of the words belonging to its class is difficult; The clustering center 2 clearly shows a positive skew, and most of the words belonging to its class can be answered correctly within only 4 times, so the difficulty of the words belonging to its class is easy.

Table 10：solution words and classes they belong to

Category	Word	y_1	y_2	y_3	y_4	y_5	y_6	y_7
Difficult（Cluster 1）	gorge	1	3	13	27	30	22	4
	swill	1	1	8	19	31	30	10
	favor	1	4	15	26	29	21	4
	fewer	0	2	10	24	32	26	6
	proxy	1	2	11	24	31	26	6
	…	…	…	…	…	…	…	…
Easy（Cluster 2）	stair	2	21	36	26	11	4	1
	those	1	13	34	30	15	6	1
	dream	5	14	31	29	15	4	1
	moist	3	13	32	29	16	7	1
	train	6	26	32	22	10	3	0
	…	…	…	…	…	…	…	…
Moderate（Cluster 3）	crank	1	5	23	31	24	14	2
	month	1	5	26	37	22	8	1
	manly	0	2	17	37	29	12	2
	slump	1	3	23	39	24	9	1
	whack	1	4	22	37	24	10	2
	…	…	…	…	…	…	…	…

5.3 Relationship between Attributes and Classes

Since we have classified the given words into three classes according to difficulty and quantified the attributes of each given word, and we found in the previous analysis that each

attribute X_1, X_2,X_3, X_4, And the attributes of each given word were quantified. In the previous analysis we found that each attribute X_1, X_2, X_3, X_4, and X_5 passed the S-W test, all of which conformed to a normal distribution, and each group of samples came from a normal population, satisfying the conditions of a one-way ANOVA. Therefore, we use one-way ANOVA here to identify each classification with the attribute it is associated with.

Table 11：Statistical description and ANOVA results

Variable	Classification	N	Mean	Std Dev	F	Sig.	Multiple Comparisons
X_1	Difficult	73	8298.70	20169.94	4.837	0.008	2＞1,2＞3
	Easy	132	47558.84	153771.85			
	Moderate	150	16912.82	42386.85			
X_2	Difficult	73	3.69	2.77	7.449	0.001	1＞3,2＞3
	Easy	132	3.22	2.63			
	Moderate	150	2.30	2.23			
X_3	Difficult	73	4.49	0.56	20.963	0.013	2＞3＞1
	Easy	132	4.90	0.30			
	Moderate	150	4.64	0.51			
X_4	Difficult	73	1.92	0.54	11.886	0.007	1＞3＞2
	Easy	132	0.82	0.47			
	Moderate	150	1.32	0.49			
X_5	Difficult	73	2.93	0.83	6.842	0.001	1＞2,3＞2
	Easy	132	1.91	0.85			
	Moderate	150	2.77	0.80			

Note：1 represents Difficulty, 2 represents Easy, 3 represents Moderate.

Based on the results of the above one-way ANOVA, it can be seen that all five attributes of the words showed significant differences in different groups. In the case of X_1, for example, the different classes in the frequency of use, the level of the Easy class was significantly higher than the Difficult and Middle class ($p=0.008<0.05$), which shows that the words in the Easy category are generally used more frequently and are more common in everyday life.

We visualize the above analysis as Fig. 11, where each circle represents a category with attributes associated with that category inside the circle. dichotomizable attributes we define as High and Low (meaning High-level in some attributions or Low-level in some attributions), and similarly trivializable attributes are defined as High, Middle and Low. Taking the Difficult class in the figure as an example, the attributes of words in the Difficult class are usually shown as low X_1 (lower frequency of use), high X_2 (more Orthographic Neighbors), low X_3 (contains less number of different letters), high X_4 (more same letter words), high X_5 (high dissonance between phonology and writing)

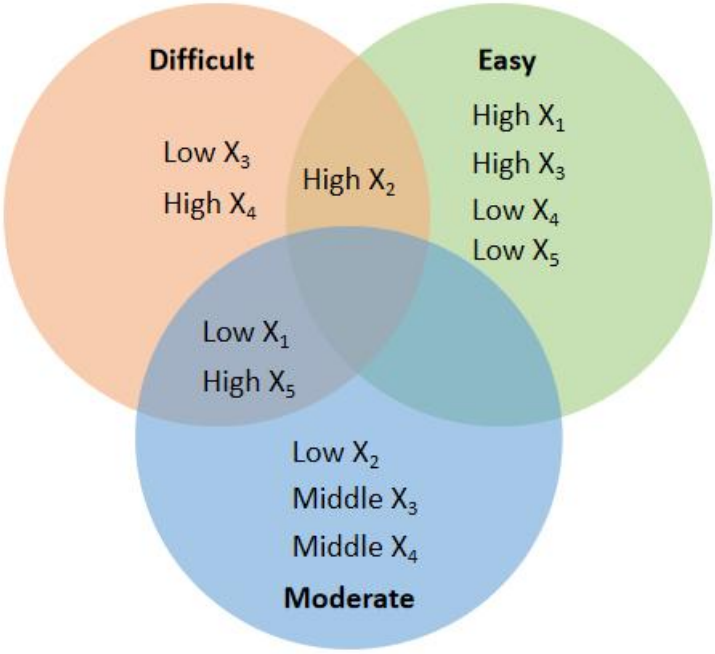

Fig.11 Characteristics of each difficulty class

5.4 Evaluation of reliability of classification effect of DT-Kmeans

We take two approaches to evaluate the accuracy of our classification model, internal evaluation and external simulation, respectively.

1. **Internal Evaluation——Davies-Bouldin index (DBI)**

The internal evaluation index is mainly based on the information of the set structure of the data set, and the clustering division is evaluated in terms of tightness and separability. Since there is no unique evaluation metric for the division of difficulty, we take the idea that excellent clustering should achieve high intra-class aggregation and low inter-class coupling as our criterion for determining the accuracy of the classification model.

The DBI is obtained by calculating the sum of the average intra-class distances of any two classes divided by the maximum of the distance between the centers of the two clusters, and then averaging.Smaller DBI means smaller intra-class distance and larger inter-class distance. When the algorithm generates clustering results that vary towards the minimum intra-cluster distance and the maximum inter-class distance, then the DBI will be smaller. Zero is the lowest possible value, and a value close to zero indicates better classification, and it is generally considered that a DBI below 0.8 results in good clustering.

The steps to calculate the DBI are shown below:

Step1：Calculate the Euclidean distance Si from the data within the class to the cluster center of mass.

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left(X_j - A_i\right)^2\right)^{1/2} \tag{5.1}$$

X_j represents the jth data in cluster class i, Ai is the center of mass of cluster class i, and T_i is the number of data in cluster class i.

Step2：Calculate the distance between the distance value class i and class j M_{ij}.

$$M_{ij} = (\sum_{k=1}^{N} (a_{ki} - a_{kj}))^{1/2} \quad (5.2)$$

a_{ki} denotes the Kth value of the center-of-mass point of the first class, and M_{ij} is the distance between the center-of-mass of the uth class and the jth class (the distance between the two points.

Step3：Calculate the similarity R_{ij} of class i and class j.

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \quad (5.3)$$

Step4：The maximum value of Rij is calculated for each cluster class i and is denoted as Di.

$$D_i = \max_{j \neq i} R_{i,j} \quad (5.4)$$

D_i is also the maximum similarity between cluster class i and other classes.

Step5：Then the maximum similarity of all classes is averaged to obtain DBI.

$$DBI = \overline{D} = \frac{1}{N} \sum_{i=1}^{N} D_i \quad (5.5)$$

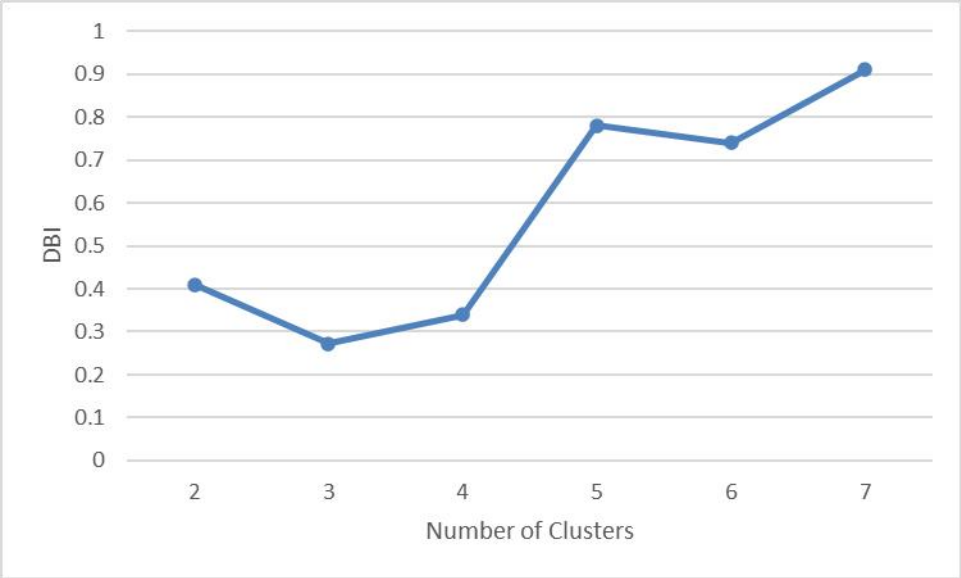

Fig.12 DBI values for different number of clusters

By selecting different number of clusters k and measuring their DBI against the reference value of 0.8, the results show that the clusters formed by the DT-Kmeans are excellent, and that our classification of words into three classes by difficulty is the optimal number of clusters，when the DBI index reaches an excellent level of 0.272.

2. **External Evaluation——Monte Carlo simulation based on decision theory**

Rand argues that clustering can be viewed as a series of decision processes, where we group two documents into the same cluster when and only when they are similar. Based on Rand's idea here, we take a Monte Carlo approach to simulate the case of data error to reflect the accuracy of the classification model.

We stipulate a correct decision (classification) is:

• TP:Classify two similar data into one cluster
• TN:Classifies two dissimilar data into different clusters

Wrong decision (classification) is:
• FP:Classifies two dissimilar data into the same cluster
• FN:classifies two similar data into different clusters

So the classification accuracy can be defined: Accuracy=(TP+TN)/(TP+TN+FP+FN).

Since the data were not validated and none of the data had class labels, we randomly selected 50 samples by Monte Carlo methods, increasing or decreasing random values of 3% to 5% of the original da-ta size, simulating similar data or disturbances due to various uncertainties in the component detection results, which we expected the classification model to assign to the same clusters as the original data. Similarly, a random sample of 50 data, increased or decreased by 30% to 50% of the value of the original data size (and normalized) as extremely dissimilar to the original data, which we expect the classification model to assign to a different cluster than the original data. Analysis of results is as follow：

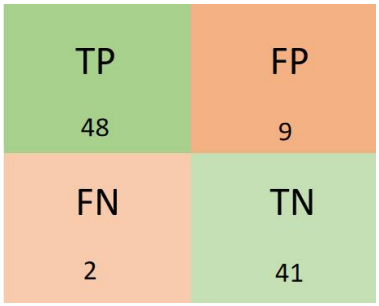

Fig.13 Confusion Matrix

The results measured 89 correctly classified and 11 incorrectly classified data, with a classification accuracy of 89%. This indicates that the model has a high classification accuracy.

Based on the above discussion, our classification model has high accuracy, and we next classify the word EERIE with difficulty. In Part IV of this paper, we predicted the percentage of relevance based on EERIE attributes on March 1 (1, 2, 3, 4, 5, 6, X) by GA-BP neural network, and the results were tested with high confidence. Therefore, we consider the correlation percentage of EERIE predicted in Part IV as its true value, calculate the Euclidean distance from each of the three clustering centers, and assign EERIE to the difficulty level corresponding to the clustering center with the smallest distance.

Table 12：Euclidean distance between EERIE and clustering center

	Cluster 1	Cluster 2	Cluster 3
EERIE	9.684	29.062	14.413

As can be seen from Table 12, EERIRE is closest to the center of cluster 1, so it is most appropriately classified as a difficulty class. Also to make the description more specific, we plotted the distribution of EERIE scores on the same graph as the distribution of scores for each cluster center.

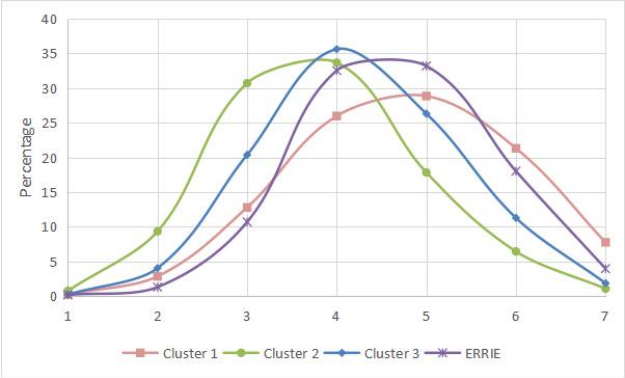

Fig.14 comparison chart of percentage of scores

From Fig.14: it can be seen that the position of the EERIE-related percentage distribution is between difficult and moderate difficulty, so that although it belongs to the difficult category, its difficulty is actually lower than the average difficulty of words in this category.

6 Some Other Interesting Features of This Data Set

By looking at the data, we found the following interesting features.

6.1 The five most frequently occurring letters are e, a, o, r, t

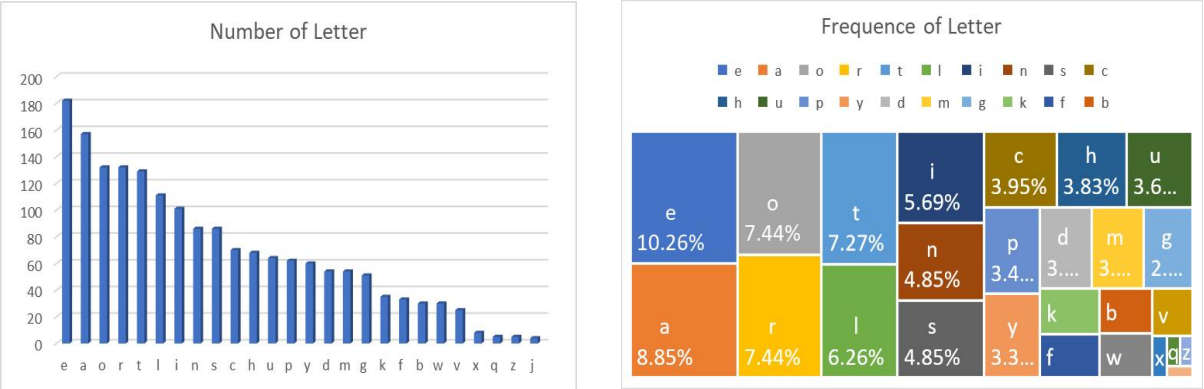

Fig.15 Number and frequency of letter occurrences

After the last 359 days of the word contains 1795 letters of the statistics, it was found that Wordle's five most frequently occurring words are e, a, o, r, t. Then users want to improve the winning rate, in the first guess words should cover the top-ranking words as much as possible, such as the first guess GREAT, you can cover the frequency of occurrence of the top five words in four, Then players are likely to be more likely to guess the correct answer.

6.2 Changes in the Number of Difficult Mode Participants Roughly Synchronized With Changes in the Number of Reported Results.

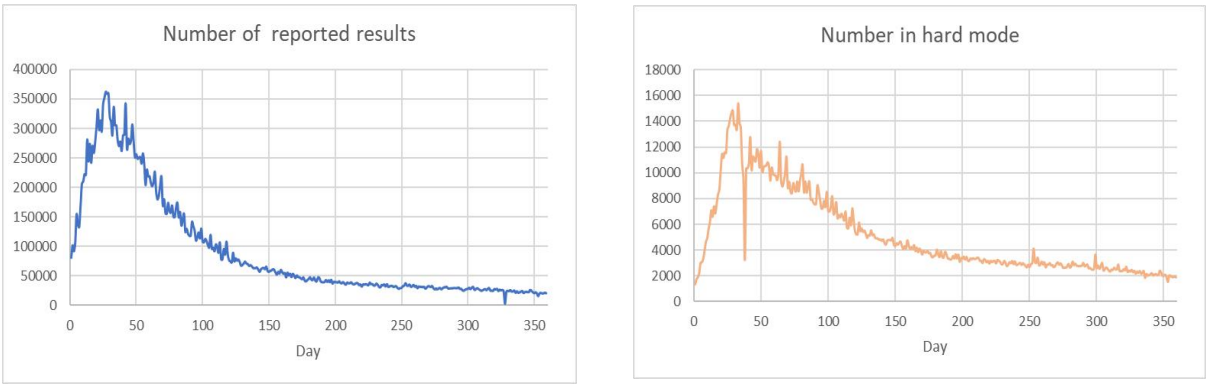

Fig16 The trend of the number of results and number in hard mode

Comparing the number of results reported from January 7 to December 31 with the number of people in difficult mode, we can see that the overall trend is roughly the same, with a sharp increase in the first period followed by a gradual decrease and finally a leveling off. On February 13 (marked 38 in the graph), there was a significant drop in the number of people in difficult mode, which returned to normal the next day.

The proportion of difficult patterns to the total showed a steady and slow increasing trend in general. It rose sharply to 93.62% on November 30 (marker 328 in the figure) and returned to normal on the next day; it was observed that the total number of simultaneously reported results also dropped sharply on November 30, only 10.82% (2569) of the previous day, and returned to normal on the next day. The number of difficult mode, on the other hand, remained stable. We can speculate that the players of hard mode are usually the loyal user group of the game.

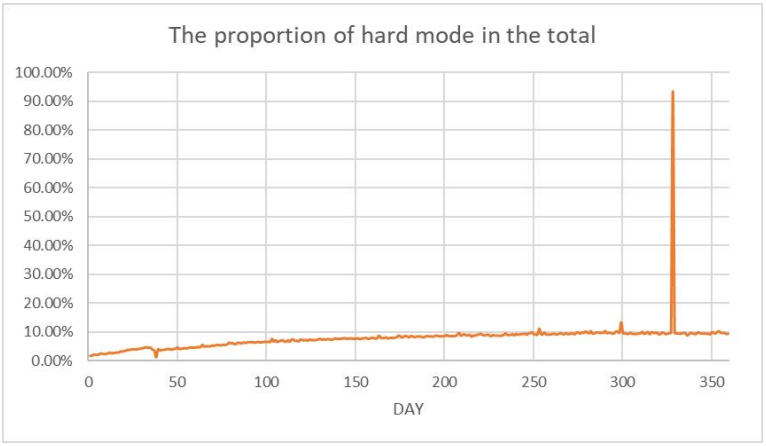

Fig.17 Trend of proportion change of hard model

6.3 The Number of Orthogonal Neighbors can Polarize Player Performance

In section 3.2.2 we found an interesting phenomenon: an increase in the number of orthogonal orders of a word makes the distribution of scores associated with it skew to the sides. This conclusion is surprising at first glance, but understandable. Based on our original

conjecture, we would normally assume that the number of orthogonal neighbors of a word would have a negative impact on player performance. Words with a high degree of similarity would lead to many wrong guesses about neighbors. It makes sense that the distribution of scores would be skewed toward 6 and 7. But how to explain the fact that the score distribution is skewed towards both 1 and 2? We give a possible reason: having more neighbors at the same time means that the structure of the word is easier to remember. Once its neighbors are guessed or a similar structure is guessed, then the player gets more hints and therefore is more likely to guess the correct answer.

7 Model Evaluation

7.1 Strengths

1. Optimization of BP networks by genetic algorithm can achieve the purpose of global optimization and fast and efficient.

2. The GA-BP network uncertainty is analyzed in terms of aleatoric uncertainty and model uncertainty, and the uncertainty is quantified by Monte Carlo dropout method, which is also applicable to the uncertainty measurement of various network prediction and classification models.

3. DT-Kmeans optimizes the traditional Kmeans model with reasonable identification of the selection of initial points, and the accuracy and credibility of Kmeans algorithm are improved.

7.2 Weaknesses

The DT-Kmeans algorithm has a high time complexity and may have problems with efficiency when facing large-scale data sets

References

[1]Lv Wenpeng. English difficult words mnemonics:A survey and analysis[J]. Foreign Language Teaching, 2001, 22(3):6.

[2]Li I. Analyzing difficulty of Wordle using linguistic characteristics to determine average success of Twitter players[J]. 2022.

[3] Hu Yiping, Gao Jiajia, Lu Hong. Determination of vocabulary difficulty in English vocabulary adaptive testing system[J]. Modern Educational Technology, 2016, 26(3):7.

[4] BoX G . Time series analysis, forecasting and control rev. ed[M]. Holden-Day, 1976.s

[5] Gaoxin. An improved K-means clustering algorithm with new clustering validity index [D]. Anhui University, 2020.

[6] Chen Fan, Xie Hongtao. Research on subway construction safety warning based on factor analysis and BP network[J]. Chinese Journal of Safety Science, 2012, (8): 85-91.

[7] Chen Wenbai. Principles and Practice of Artificial Neural Networks [M]. Xi'an: Xi'an University of Electronic Science and Technology Press, 2016.

[8] Bao ZY, Yu JZ, Yang SF. Intelligent optimization algorithms and their MATLAB examples (2nd edition) [M]. Beijing: Electronic Industry Publishing, 2018.

The New York Times & Wordle

Dear Editors of The New York Times：

I would like to thank you for giving us such a precious opportunity to participate as data analysts in solving the problem of Predicting Wordle Results. The New York Times, as a communication media with extraordinary influence worldwide, represents the top level of the profession and is the political and economic bellwether of contemporary America and the world. Moreover, the rigor with which your newspaper approached the issue appealed to our team. We are very enthusiastic about both the Wordle game itself and the analysis of the data related to it.

What we have done?

Firstly, we build ARIMA Model to create a prediction interval for the number of reported results on March 1, 2023. Determine if word attributes affect the player's score in hard mode. Secondly, by building a GA-BP neural network, the distribution of the reported results of the solution word for future dates is predicted. We used this model to predict the percentage of scores for the word ERRIE on March 1. Finally, DT-kmeans is used to classify the difficulty of solution word, identify the attributes associated with each difficulty, and predict the difficulty of EERIE.

Our results：

- ★ We use the ARIMA model to predict interval for the number of reported results on March 1, 2023， which is **[17737,25379]**；
- ★ We have divided 5 attributes for words, which are word frequency, orthography neighbors, number of different letters included, same letter words, and the degree of dissonance between phonology and writing. Combined with Spearman correlation analysis, we found high word frequency and number of different letters included will have a **positive impact** on player performance, well same letter words and the degree of dissonance between phonology and writing are the **opposite**.
- ★ We were astonished to find that the effect of orthography neighbors on player performance, however, may either reduce or enhance player performance. In other words, its effect on player performance is **difficult to control**. Therefore, we believe that all five attributes will affect scores in the difficult mode.
- ★ We build a GA-BP neural network for the prediction work. For EERIE, the predicted percentage of its score is（**0.197,1.326,10.712,32.537, 33.203,18.045, 3.981)**. We innovatively used DT-kmeans clustering method to classify **3 difficulty levels** for solution word, i.e., easy, moderate and difficult. And the difficulty of EERIE is in the difficult class but below average. In identifying the attributes associated with each difficulty level, taking the difficulty level as an example, we were surprised to find that more orthography neighbors, same letter words, and the dissonance between phonology and writing in a high level will **Increase the difficulty**. However, high word frequency, number of different letters included in a high level will reduce the difficulty.

It would be a great honor for us to help you in any way with the results of our analysis and predictions.

Yours sincerely,
Team 2306318