Journal of Education and Educational Research
ISSN: 2957-9465 | Vol. 7, No. 1, 2024

P&A: Make Wordle Game Better

Yuqi Zhang *

Mathematics and Applied Mathematics (Normal College), Hangzhou Normal University, Hangzhou, 311121, China
*** Corresponding author Email:** lunw20243243@126.com

Abstract: This study delves into enhancing the gaming experience by strategically predicting and analyzing Wordle outcomes. Rooted in prediction and analysis principles, our P&A model employs a systematic approach. Beginning with an exploration of statistical patterns in game results, we construct a predictive model for March 1 using methods such as trend analysis, median, variance, paired sample t-tests, and chi-square independence tests. Short-term outcome predictions incorporate time series ARIMA and support vector machine models for improved accuracy. The subsequent phase develops a model to predict future Wordle solutions, addressing uncertainties through practical examples and neural networks. Difficulty classification of Wordle solution words is tackled by associating attributes with each category, employing K-MEANS clustering, and optimizing it for better performance. Modified K-MEANS clustering assesses indicators' significance, and machine learning with XGBoost assigns importance scores. In conclusion, our predictive and analytical exploration of Wordle questions 1, 2, and 3 culminates in a dataset listing and description. The integration of the P&A Model, K-MEANS clustering, ARIMA, and various methodologies represents a significant advancement in unraveling the dynamics of Wordle.

Keywords: P&A Model; K-MEANS; ARIMA; Wordle.

1. Introduction

1.1. Problem Background

The Wordle description on the New York Times website explains that the color of the tile will change after you submit the word. A yellow square indicates that the letters in the square are in the word, but in the wrong position. A green square indicates that the letters in the square are in the word and in the correct position. A gray square indicates that the letters in the square are not included in the word at all.

(1) The number of reported results changes every day. Develop a model to explain this change and use our model to create a prediction interval for the number of results reported on March 1, 2023. Identify attribute of this word affecting the percentage of reported scores played in difficult mode.

Fig 1. The name of the crossword game

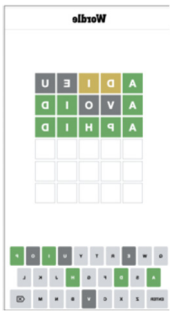

Fig 2. Crossword game running interface

(2) For a given future solution word for a future date, develop a model so that you can predict the distribution of report results. In other words, predict the relevant percentage of future dates (1, 2, 3, 4, 5, 6, X). Identify what uncertainties are associated with our models and predictions and take a concrete example of our prediction of the word EERIE on March 1,2023.

(3) Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word associated with each classification.

(4) List and describe some other interesting features of the dataset.

1.2. Our Work

In this problem, our work revolves around power curves. Our specific tasks are as follows:

Task 1: We are to build a model called P & A model in order to predict its results on March 1. Our model will analyze the statistical rules of the report results and predict their results on March 1 based on the recorded quantitative data. First of all, we need to pre-process the provided data, deal with the date in the attachment, the game number, the word of the day, the number of people reporting the score of the day, the number of players in the difficult mode, and the percentage of guessing the word once, twice, three times, four times, five attempts, six attempts or problems that cannot be solved. Studying the change rules between the indicators can be assisted by drawing visual charts. To analyze the statistical law of indicators, we must first use the most basic statistics. Trends, medians, variances, and so on are all factors we need to consider. Then we use the paired sample t test to process continuous data, and the chi-square independence test to process discrete data. The grey correlation method is also in our consideration solution strategy. We use time series ARIMA and support vector machine to predict the number of short-term, and combine multiple models to improve the prediction accuracy.

Task 2: We develop a model for a given future solution word for a future date, predict the distribution of report results, and find out which uncertainties are related to our model and prediction. The distribution results are predicted by practical examples, and the neural network is used for multi-value prediction to improve the training accuracy and dimension.

Task 3: We are to discuss the effect of environmental

factors on results. To identify the properties of a given word associated with each category, we develop and summarize an analysis model to classify solution words by difficulty. Think and adopt the commonly used K-MEANS clustering analysis problem algorithm, and improve it, innovate and improve it. To optimize the algorithm, we will use the modified K-MEANS clustering analysis after we use the good and bad indicators between different algorithms. This paper studies the relationship between multiple indicators. Firstly, the machine learning method uses XGBoost to assign importance scores to each indicator to construct a factor analysis-clustering model. Then, exploratory factor analysis is used to reduce the dimension and visualization of multiple variables, so as to more intuitively reveal the differences of each indicator. After the classification is completed, the rationality and sensitivity are analyzed, and the calculation results are put on. For the word attribute of EERIE, a discriminant function is established to analyze the classification accuracy.

Task 4: Based on our predictions and analysis of Questions 1, 2 and 3, we list and describe some other interesting features of the dataset.

We extend the P & A model separately to obtain different objective functions and control equations, and then find new optimal solutions.

2. Assumptions and Justifications

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

Assumption 1: Each player sample is independent of each other, and there is no interaction between the number of attempts, the number of reports, and the number of difficult patterns.

Justification: Ignore the interaction between the number of attempts, the number of reports, and the number of difficult modes of players on their independence.

Assumption 2: The data of the number of indicators and the number of attempts provided by the annex are true and credible.

Justification: Regard he data of the number of indicators, the number of attempts, and the type of attempts provided in the annex as the results of the real player situation.

3. Notations

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

4. The P&A Model

4.1. Data Pre-processing

Firstly, the data of Annex I and Annex II are preprocessed, we carry out equal precision measurement. get x_1 , x_2 ..., x_n ,calculate its arithmetic mean x and residual error $v_i = x_i$ - x(i =1,2, ..., n), The standard deviation δ is calculated according to Bessel formula. If the residual error of a measured value ($1 \le b \le n$) satisfies the following formula:

$$|v_b| = |x_b - x| > \pm 3\delta$$

We believe that it is a bad value with gross error value, which should be eliminated.

For some unreasonable data, we assume that this part of the data is due to manual writing errors or sampling errors, and properly handle these data, we decided to modify or eliminate the method. We assume that the background value range of a given index is limited to the x ± 3δ interval, so we consider the data in the interval (x- 3δ, x + 3δ) as a normal range value, and we correspondingly eliminate the abnormal data. Then we use the box plot to accurately determine the value of the extreme abnormal value. For this part of the data, we eliminate it and insert the corresponding value.

Table 1. Notations used in this paper

Description	Variable
Difficult mode or not	y
Number of players	x_1
Number of difficult choices	x2
Number of words	x3
Number of reports	x4
Try i times success number	w_{ij}
Spearman correlation coefficient	r_s
Discriminant coefficient	ρ
Grey correlation degree	r_i
Normalized data	x_j'
Threshold	θ_j
Expected output value	y_k
Skewness	S
f kurtosis	k

When there are outliers in the data, especially outliers with large deviations, it will bring errors to data analysis and modeling. Therefore, it is necessary to detect outliers. Commonly used outlier detection methods include 3σ rule or Z distribution method, which is based on the assumption of normal distribution. Because the distribution of some numerical features in this paper does not conform to the normal distribution, that is, the data distribution is not uniform. Therefore, this paper chooses to use the box line diagram that does not require data distribution to detect outliers in numerical characteristics.

The principle of using box plot to detect outliers in data is: by calculating the quartile plus or minus 1.5 times the quartile distance, that is, calculating the values of Q1-1.5IQR and Q3 + 1.5IQR, the data falling outside this interval is specified as outliers. In the box plot, the median, upper quartile, lower quartile, upper and lower edges and potential outliers of the variable data can be seen. In this paper, the upper quartile is used to replace the data whose value is greater than Q3 + 1.5IQR, and the lower quartile is used to replace the data whose value is less than Q1-1.5IQR, and the anomaly is drawn.

The predicted results subtract the error to obtain the left boundary of the prediction interval, and add the error to obtain the right boundary of the prediction interval. The final prediction interval is [18578-22562].

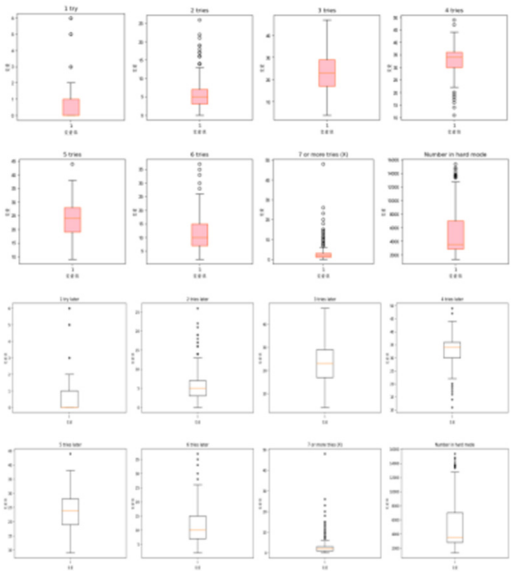

Fig 3. Box diagram operation results

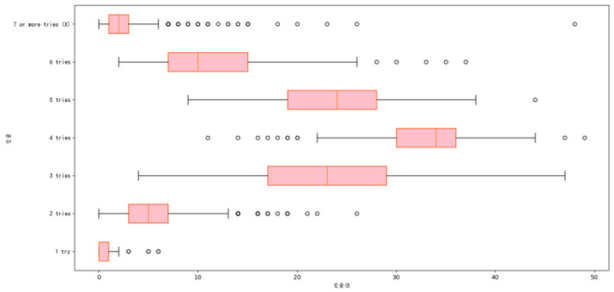

Fig 4. Box diagram of the operation results of the concentrated graph

4.2. Grey Relational Analysis

Using the theory of grey systems, grey relational degree analysis was conducted on various subsystems. Non-dimensionalization, difference calculation, and correlation calculation were used to overcome the dimensional differences of each factor, and to perform quantitative analysis on the numerical relationships of each subsystem (or element). Among them, grey relational degree analysis is based on the comparability and similarity of sequences, analyzing the main direction of system development and the correlation of each influencing factor, thus providing a quantitative measurement index for the evolution trend of a certain system.

In an objective system, grey system analysis can more truly and comprehensively reflect people's understanding of the objective system, and can achieve both qualitative and quantitative effects. The steps are as follows:

(a) Take the number of players as the mother sequence, and take the $x_0(k)$ as various index influencing factors, with adjacent player number indices as the subsequence.

(b) Calculate the difference sequence according to the following formulaΔi(k), and find the maximum and minimum absolute difference values in the calculation results.Δmin, $\Delta \mathrm{i}(\mathrm{k}) = x_i'(\mathrm{k}) - x_0'(\mathrm{k})$.

(c) Calculate the correlation coefficient and grey relational degree for each station location. $q_i(k), \mathrm{i} = 1,2,3 \ldots$and$\gamma_i, \mathrm{i} = 1,2,3\ldots \ \gamma_1, \gamma_2, \gamma_3$.

$$q_i(k) = \frac{\Delta \mathrm{min} + \rho \times \Delta \mathrm{max}}{\Delta \mathrm{i(k)} + \rho \times \Delta \mathrm{max}} \qquad \gamma_i = \frac{1}{n}\sum_{k=1}^{n} q_i(k)$$

where is the distinguishing coefficient, generally taken as 0.5. We believe that the closer the value is to 1, the better the correlation between this factor and the main direction of system development. Using thermodynamic diagrams and EXCEL software to perform linear analysis on the grey relational degree, the results are as follows.

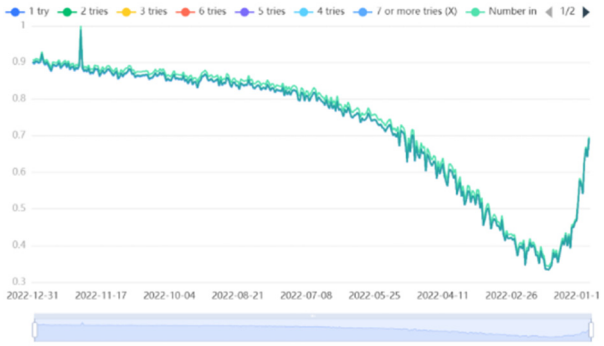

Fig 5. Thermodynamic diagram and EXCEL software linear analysis results of grey correlation degree

The visualization results of the player number are shown in the above figure.

Based on the dimension reduction of the twelve indicators in the problem to six factors, the relationship between the change in the number of players and each indicator can be obtained. Therefore, the differences in the number of difficulty patterns for words with different attributes can be analyzed and obtained. For example, the effect on Excel words has a positive coefficient of 3.55, while for EERIE, it is a negative correlation with a coefficient of -4.98.

Table 2. Excavation evaluation result

Evaluation item	correlation	ranking
1	0.724	1
2	0.718	2
3	0.697	3
4	0.671	4
5	0.668	5
6	0.662	6

The similarity between each evaluation item and the "reference value" (parent sequence) is calculated by the average of the correlation coefficients. By mining the discovered correlations, information about one attribute can be inferred from information about another attribute, and to some extent, this rule can be proven correct. The larger the correlation, the higher the correlation between the evaluation item and the "reference value", and the higher the correlation, the higher the correlation between the evaluation item and the "reference value". By combining the various evaluation indicators with their correlations, the weightings for each indicator can be obtained.

We still consider whether time should consider the difference between working days and non-working days, we use grey correlation analysis. And set the following variables: characteristic sequence variables: {weekday, 7 or more tries (X), 6 tries, 3 tries, 2 tries, 4 tries, Number in hard mode, 1 try, 5 tries}; parent sequence variable: {Number of reported results}; index item: {Date}. The dimensionless treatment: { mean } ; the resolution coefficient ρ : { 0.5 } is used as a parameter.

We have analysis steps: (Initialization: as the name suggests, it is to divide the data of this sequence by the initial

value). Because the order of magnitude of the sequence of the same factor is not much different, these values can be sorted to 1 by dividing the initial value.

Meaning: as the name implies, is to divide the data of this sequence by the mean, because the order of magnitude of the sequence mean is relatively large, so after removal can be normalized to the order of magnitude of 1.

Step 1: Dimensionless processing (mean value, initial value) is performed on the data.

Step 2: Solve the gray correlation coefficient between the parent sequence (comparison sequence) and the feature sequence.

Step 3: Solve the grey correlation value.

Step 4: The grey correlation value is sorted and the conclusion is drawn.

Initialization: as the name suggests, it is to divide the data of this sequence by the initial value. Because the order of magnitude of the sequence of the same factor is not much different, these values can be sorted to 1 by dividing the initial value.

Meaning: as the name implies, is to divide the data of this sequence by the mean, because the order of magnitude of the sequence mean is relatively large, so after removal can be normalized to the order of magnitude of 1. Our conclusions and analysis of the conclusions are as follows:

Table 3. Results of correlation coefficient

Results of correlation coefficient									
	weekday	7 or more tries (X)	6 tries	3 tries	2 tries	4 tries	Number in hard mode	1 try	5 tries
2022-01-07	0.9358951821882528	0.9403405806514696	0.9872563680789626	0.9852883525307801	0.9572453750293599	0.9656468363349628	0.9309735410529724	0.8709949218849603	0.9849039620339693
2022-01-08	0.8821281109450126	0.9539742670885542	0.9888471804516847	0.9877092591259384	0.9697673130318947	0.9795360247244486	0.9155696195109376	0.8923611231310168	0.9880958428860763
2022-01-09	0.8925240921931025	0.952182297770 93	0.9030577434965 56	0.9506374510186104	0.9443318815918503	0.9782504217108728	0.9298095458205319	0.8819689159103373	0.9695225139708973
2022-01-10	0.9675016922908487	0.9472725924933368	0.9662420734654031	0.9463064475649494	0.9442151734576499	0.9690448261275086	0.9187946888707791	0.8983059143244639	0.9893004437772719
2022-01-11	0.9728952892087255	0.8621322731265408	0.8689520491007342	0.9821246560191915	0.9821270783116267	0.9268161417533676	0.8835910636574016	0.9508939533659465	0.8916208007142551
2022-01-12	0.9936400815316169	0.9897151033350267	0.9650040908541789	0.9073470840894513	0.9097641756573387	0.9203059050823992	0.9017369785115993	0.9318785608220209	0.9668927384629082
2022-01-13	1	0.9554026714777 1	0.9687036943972 33	0.9039542147737373	0.8820176801475498	0.9352042122708936	0.9122572840022891	0.9263532287239585	0.9825991134584 79
2022-01-14	0.9538250154469847	0.9903799074527119	0.9365230086450 71	0.8988667054170062	0.8762739284976128	0.8976185023538 47	0.8852995108894 79	0.9698460385957293	0.9077881231831753
2022-01-15	0.7867283179956999	0.8140678193979748	0.8201456975920459	0.9202300843218755	0.9202322109234528	0.8715001652368909	0.8607636259257982	0.9835478881343881	0.8403099496372 48
2022-01-16	0.7837003353911451	0.8108261635815436	0.8309166548571674	0.9030146543558174	0.9160920700848177	0.8623434595916559	0.8623481860241936	0.9788198947669069	0.8440122254041631
2022-01-17	0.7736486996038761	0.8283628458733316	0.8266442008509397	0.8896954259230105	0.8860055143016636	0.8501889168718799	0.8619836217435213	0.9631899429712852	0.8323654501785946
2022-01-18	0.8959039236407617	0.9663643548287071	0.9788381014153721	0.8110437583967771	0.8000286494843771	0.8308003198960 68	0.8732256053489045	0.9647159763772482	0.8819118645284391
2022-01-19	0.8369747227261067	0.7536736760004766	0.7529615233665623	0.8513661229249543	0.9599768781060073	0.7888296287932239	0.8313160840914918	0.8967302033744851	0.7640967312576 47
2022-01-20	0.8722193568742562	0.7821326137980988	0.8075082371310434	0.8558323019042273	0.8640590480484232	0.8350009781713196	0.8572699338635256	0.9373090356620499	0.8197182772575707
2022-01-21	0.8433847300158411	0.7588672895070994	0.7764375293944202	0.8316821074067258	0.8357526842546695	0.8061670043161349	0.8428625622181735	0.9040921889558 85	0.7910173110215921

The diagram shows: the above table is the preview result; all data please click the download button to export. The correlation coefficient represents the degree of correlation between the corresponding dimension of the subsequence and the parent sequence (the larger the number, the stronger the correlation).

Intelligent analysis: it can be seen from the above table that the grey correlation analysis is carried out for 9 evaluation items (weekday, 7 or more tries X), 6 tries, 3 tries, 2 tries, 4 tries, Number in hard mode, 1 try, 5 tries and 359 items of data, and Number of reported results is used as the ' reference value ' (parent sequence). The correlation between 9 evaluation items (weekday, 7 or more tries X, 6 tries, 3 tries, 2 tries, 4 tries, Number in hard mode, 1 try, 5 tries and Number of reported results (correlation degree) is studied, and the analysis reference is provided based on correlation degree. When using grey correlation analysis, the resolution coefficient is 0.5, and the correlation coefficient value is calculated by combining the calculation formula of correlation coefficient. (PS : The resolution coefficient $\rho \in (0, \infty)$, the smaller the ρ, the greater the resolution. Generally, the value interval of ρ is (0,1), and the specific value can be determined according to the situation. When $\rho \leq 0.5463$, the resolution is the best, usually $\rho = 0.5$.)

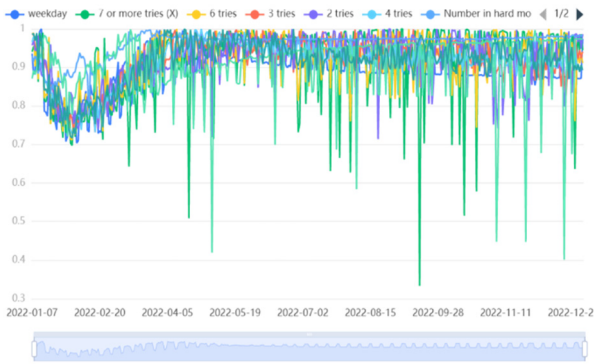

Fig 6. Correlation coefficient diagram

The diagram shows: the correlation coefficient represents the correlation degree of the subsequence weekday, 7 or more tries (X), 6 tries, 3 tries, 2 tries, 2 tries, 4 tries, Number in hard mode, 1 try, 5 tries corresponding dimension of the parent sequence (the larger the number, the stronger the correlation).

Table 4. Grey correlation degree

Correlation results		
appraisal items	degree of association	rank
Number in hard mode	0.962	1
6 tries	0.919	2
3 tries	0.916	3
5 tries	0.916	4
4 tries	0.915	5
2 tries	0.914	6
7 or more tries (X)	0.91	7
1 try	0.91	8
weekday	0.899	9

The diagram shows: the correlation degree indicates the degree of similarity between each evaluation item and the ' reference value ' (parent sequence), which is calculated by the average value of the correlation coefficient. The correlation value is between 0 and 1. The greater the value, the stronger the correlation between the evaluation item and the ' reference value ' (parent sequence), the higher the correlation degree, which means that the closer the relationship between the evaluation item and the ' reference value ' (parent sequence), the higher the evaluation. Combined with the correlation value, all evaluation items are sorted to obtain the ranking of each evaluation item.

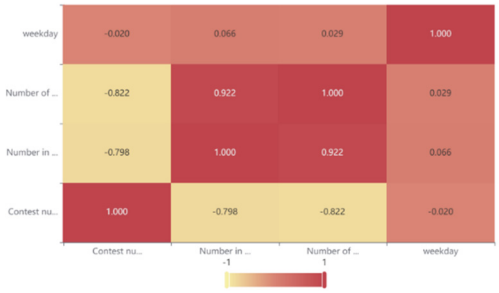

Fig 7. Correlation coefficient Correlation degree diagram

Intelligent analysis: Combined with the above correlation coefficient results, the correlation degree value is finally obtained, and the correlation degree value is used to evaluate and rank the nine evaluation objects. The correlation value is between 0 and 1, and the larger the value, the stronger the correlation between it and the ' reference value ' (parent sequence), that is, the higher the evaluation. From the above table, it can be seen that for the nine evaluation items, Number in hard mode has the highest evaluation (correlation degree : 0.962), followed by 6 tries (correlation degree : 0.919).

Analysis results: Grey correlation analysis is to calculate the correlation between the feature sequence and the parent sequence : the correlation between Number in hard mode and Number of reported results is 0.962, the correlation between 6 tries and Number of reported results is 0.919, and the correlation between 3 tries and Number of reported results is 0.916. The correlation between 5 tries and Number of reported results is 0.916, the correlation between 4 tries and Number of reported results is 0.915, the correlation between 2 tries and Number of reported results is 0.914, and the correlation between 7 or more tries (X) and Number of reported results is 0.91. The correlation degree between 1 try and Number of reported results is 0.91, and the correlation degree between weekday and Number of reported results is 0.899. The number in hard mode has the largest correlation degree with Number of reported results, and the weekday has the smallest correlation degree with Number of reported results.

The forecasting model based on weighted moving average is utilized to predict the change range of detection data on March 1st by analyzing historical records. As the number of historical data before March constitutes a time series in this problem, time series models are appropriate for predicting the quantity of various indicators on March 1st. Popular methods in time series modeling include moving average, exponential smoothing, and adaptive filtering. We choose the weighted moving average method for prediction.

Further analysis reveals that the difficulty level of players' gaming modes varies, resulting in a possibility of undetected data, leading to "0" values in the overall data. We handle the data by calculating the weighted average, and adopt the normal distribution curve function for weight distribution during the weight calculation. Suppose the 12 indicators, such as the number of players, detection results, and the quantity of difficult modes, are denoted as x1, x2, ..., x12, and their weights are denoted as w1, w2, ..., w12. After the change, the new 14 indicators are x1', x2', ..., w14' with weights w1', w2', ..., w14'. In assigning weights, standard normal distribution is utilized to assign weights to different quantities. The standard normal distribution formula is as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$$

By multiplying each column data with the standard normal distribution function, the weighted average is obtained as follows: (formula)

$$\bar{x} = \frac{X_1W_1 + X_2W_2 + \cdots + X_{14}W_{14}}{W_1W_2 + \cdots + W_{14}}$$

$$\bar{x'} = \frac{x'_1w'_1 + x'_2\omega'_2 + \cdots + x'_{14}w'_{14}}{w'_1 + w'_2 + \cdots + w'_4}$$

Let a_i represent the weighted proportion of the quantity for the i mode, then Matlab is used to program and solve this problem to forecast the range of the number of different difficult modes on March 1st.

$$\alpha_i = \frac{\bar{x_i}}{x_i}$$

5. BP Neural Network to Problem Two

5.1. Using BP Neural Network to Predict the Distribution of Results

BP neural network is a type of neural network whose learning process is based on error backpropagation. It is implemented using MATLAB. At each iteration, the weights and thresholds are modified based on the analysis of the errors between the obtained and expected results, until a model that can output results consistent with the expected ones is obtained.

Defining the input as the number of attempts in the game's

difficult mode, and the output as the occupancy percentage, the problem can be transformed into a complex function mapping problem, which is addressed using a five-layer hidden layer and simulation processing in a BP neural network. The function is treated as the activation function for the BP neural network simulation. The processed numerical values are transformed into the range of (0,1), by discretizing the continuous data into discrete data.

Step 1: The training set and testing set are obtained by randomly selecting 30 data samples with each indicator obtained after removing outliers and normalization. Five data samples are used as the validation set.

Step 2: The BP neural network is constructed to map the input of the neural units to the output. To establish the relationship between the input and output data, a function is used between the last hidden layer and the output layer, with the expression as follows.

Table 5. Correlation coefficient run result data

Contest number	report quantity	Number of difficulties	*x1*	.....	x2	x3	.....	*x6*
533	**32018**	**4733**	**49.30**	**。。**	**37.44**	**35.10**	**。。**	**1.05**
532	**31191**	**4835**	**55.76**	**。。**	**25.56**	**29.9**	**。。**	**1.8**
531	**35721**	**4809**	**57.46**	**。。**	**30.71**	**38.0**	**。。**	**2.01**
530	**31903**	**4906**	**56.75**	**。。**	**20.62**	**7.36**	**。。**	**2.05**
529	**35343**	**4928**	**17 52**	**。。**	**27 41**	**41 54**	**。。**	**2 63**
528	**33660**	**4973**	**13.76**	**。。**	**24.59**	**46.80**	**。。**	**2.71**
527	**34281**	**5238**	**37.24**	**。。**	**39.92**	**78.1**	**。。**	**2.83**
532	32018	4733	75.58805		30.22	50.3	。。	2.91

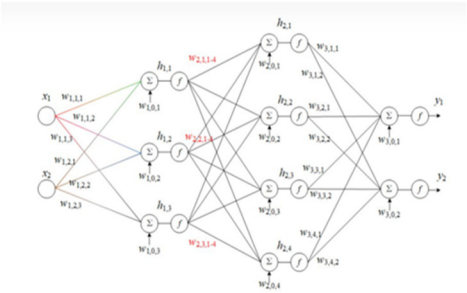

Fig 8. Correlation coefficient correlation result graph

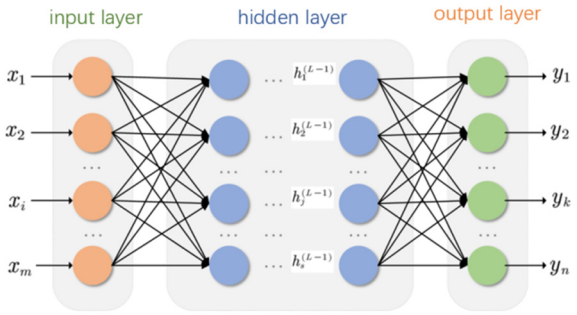

Fig 9. Correlation coefficient figure

The expression between network input data and output data is obtained.

$$y_k{}' = \sum_{j=1}^{r} v_j \cdot f[\sum_{i=1}^{m} w_{ij} \cdot p_i + \theta_j$$

(k = 1, 2, … N), It is the link weight, the threshold, the expected output value, and the actual output value of the network. The neural network structure diagram is as follows:

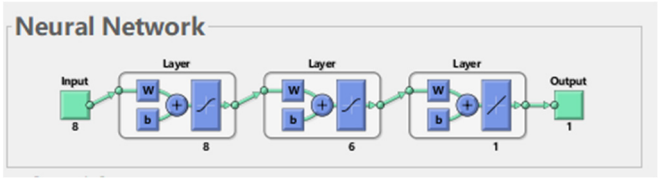

Fig 10. The neural network structure diagram

Step3 The parameters of the BP neural network the parameters set in this model are as follows:

Table 6. The parameters set in this model

Maximum training steps	training results interval steps	learning rate	training target error	training times
1000	1	0.0000001	0.000001	1000

The following table is the prediction results of BP network.

Stickiness level is a feature that simulates the imbalance of game time during feature extraction. It is calculated by calculating the difference of event counts between the first 30 % and the last 30 % of the time period. Negative values indicate that the user plays more at the end of the observed time interval, so it is unlikely to try too many times, contrary to players with a positive level of stickiness.

Sticky vectors provide a more detailed view of player activity distribution. It consists of seven elements, each representing 10 % of the total running time of the feature extraction cycle. The value of the element is the percentage of events recorded over the entire period over the corresponding time interval.

Various combinations of feature extraction and churn duration were extensively tested, plus several machine learning algorithms. For feature extraction, the intervals of 1, 2, 3 and 7 days are considered, and the period of 1 to 7 days is predicted, with a total of 28 combinations. The resulting

profiles are used as inputs for logistic regression, random forests, and k-nearest neighbors.

Table 7. EERIE prediction table of proportion of different attempts

number	proportion/%	variance	precision	accuracy
1	2.2	10.2	7.9	2.1
2	6.3	14.5	3.1	1.2
3	14.1	2.5	6.186	7.58
4	35.1	2.3	5.817	5.11
5	30.8	3.4	6.7	3.7
6	10.2	2.9	9.5	4.3
X	74.8	2.2	5.3	8.1

Models built in this way are evaluated using metrics previously described. Only based on accuracy, 72 % of the RF algorithm will prevail, while the k-nearest neighbor algorithm ranks the lowest at 69 %. In terms of F1 score, the highest result of logistic regression was 0.78.

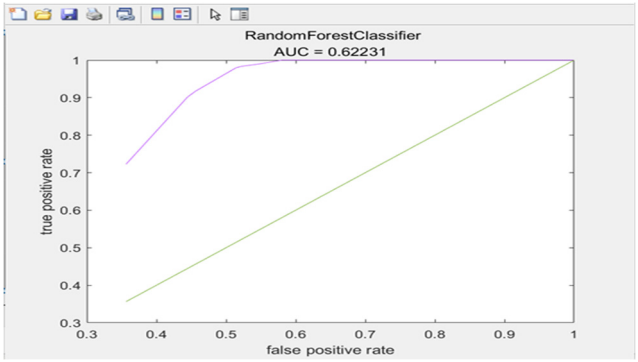

Fig 11. Logistic regression's highest outcome prediction graph

6. The Establishment and Solution of Problem Three

We classify different word attributes based on difficulty, and judge that the word EERIE belongs to a class. Further, **based on** classification, we establish a systematic clustering model of word types for the rationality and sensitivity analysis of the classification model. On this basis, according to the two divisions, the standard deviation problem is passed. The two overall modeling processes are as follows.

6.1. Data Preprocessing and XGBoost Algorithm

Table 8. Word feature type screening result table

type	>6	Total accuracy range
difficult	yes	99 81%- 100%
difficult	no	97.25%- 100%
sample	Yes	90. 17%-99.89%
sample	no	88.41%-99.98%

Fig 7 Use Excel software to remove the values that differ significantly between the two-word types and get the table 8.

Then the feature engineering based on machine learning method is used to test the importance of various features.

Machine learning strategies can be used for feature screening of different types of classification problems. The XGBoost algorithm is used for testing. The core idea of this method is to continuously add trees to fit the last residual to support parallelization and cross-validation to improve the partitioning effect.

$$IG(X,Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x) \cdot p(y|x) - \sum_{y} p(y) \cdot \log p(y)$$

Finally, using information entropy, according to the characteristics of the largest information gain, the information gain is sorted from small to large to obtain the most important characteristics.

6.1.1. Data Pre-processing Classification and Subclassification

The paired sample t test of continuous attributes in the data set and the chi-square independent test of discontinuous attributes are as follows:

Table 9. Table of pairing results for successive attributes in the dataset

Word degree	Chi-square test score threshold	The feature set after one screening of classification features
sample	1	1 ,3 ,6 ,12
difficult	10	3 ,9 ,11 ,12

The second screening is carried out on the basic classification, and the maximum information coefficient method is used to analyze the large sample data set, and the nonlinear variable relationship analysis is integrated. The formula is as follows:

$$I(X;Y) = \sum_{x \in y} \sum_{y \in x} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

6.1.2. Systematic Clustering Model of Word Types

System clustering (hierarchical clustering) algorithm:

Step1: Classify each function.

Step2: Calculate the distance matrix between classes. The last two classes are new classes.

Step3: Calculate the distance between the new class and each class. When the number of classes is 1, perform the next step, otherwise go to Step2.

Step 4: Drawing a cluster diagram.

Step 5: Determine the number and type of clusters.

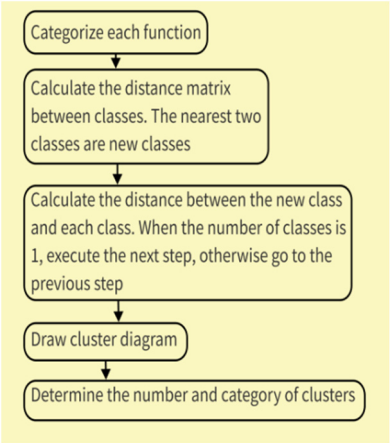

Fig 12. System clustering (hierarchical clustering) algorithm

6.2. Support Vector Machine Model-assisted Classification

6.2.1. Support Vector Machine Model

First, the training set and test set need to be extracted from the original data, and then preprocessed. Then use the training set to train the support vector machine, and finally use the obtained model to predict the classification labels of the test set. The process is as follows:

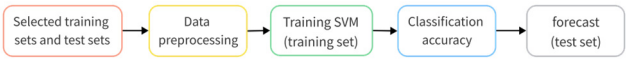

Fig 13. Model operation flow chart

a) Select training set and test set

Among the 35 samples, samples 1-8 belong to the first category (category label 1), samples 9-21 belong to the second category (category label 2), and samples 22-35 belong to the third category (category label 3). Each category is now divided into two groups and the data is recombined, one as a training set (train _ wine) and the other as a test set (test.wine).

b) Training and prediction

The training set train _ Wine is used to train the SVM classifier, and the obtained model is used to predict the labels of the test set. Finally, the classification accuracy is obtained.

Table 10. Model visualization result

Algorithm	type number	real number	accuracy
SVM	69	60	86.7%

The classification finally converges and the sensitivity is good.

6.2.2. Support vector Machine Modelhe Classification of Cluster Center Point and Each Word Component are Obtained

Calinski-harabasz criterion is sometimes called variance ratio criterion (VRC), which can be used to determine the optimal K value of clustering, draw scatter plots between variables, and mark cluster centers. The clustering effect is observed by the histogram on the diagonal. The smaller the cross section, the better the clustering effect, and vice versa.

Using Matlab programming to solve the clustering center point and the classification of each word component, visual analysis of ordinary words and difficult words before and after the comparison, as shown below:

The central point of the cluster is solved and the components of each word are classified, and the comparison graph between common words and difficult words is visually analyzed.

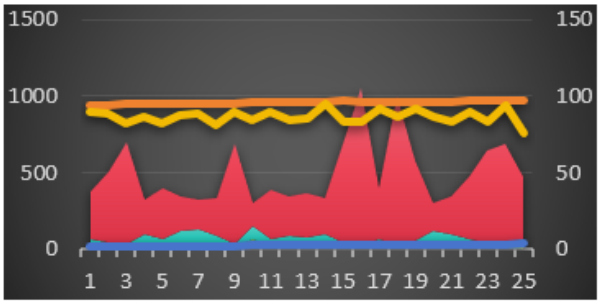

Fig 14. The central point of the cluster is solved and each word component is classified, and the result graph before and after common words and difficult words is visually analyzed

6.3. The Evaluation Index Selection and Solution of K-Means + + Model

The Calinski-Harabasz criterion can be used to determine the optimal K value of clustering, which corresponds to a larger inter-cluster variance and a smaller intra-cluster variance. The optimal number of clusters corresponds to the solution with the highest Calinski-Harabasz index value. The range is set to 2-5 in Matlab, and the larger the C-H value in this range, the better; draw a scatter plot between each variable and mark the cluster center. Observe the clustering effect through the histogram on the diagonal. The smaller the cross section, the better the clustering effect. Finally, Matlab is used to program and solve the clustering center point and the classification of each word component. Through the visual analysis of the comparison between ordinary words and difficult words before and after weathering, the results are viewed in the annex. The clustering analysis diagram of ordinary words is as follows:

Fig 15. The clustering analysis diagram of ordinary words

Analysis of clustering of common words the cluster center point data of each category is obtained. Due to the large amount of data, we put it in Appendix 3. The comparison of the categories of different word types before and after weathering is as Table 11.

6.4. Difference Judgment of Word Attribute Association between Different Categories

On the basis of K-Means + + clustering results, some appropriate indicators are selected to judge the difference of classification results. In this paper, we choose CHI, DBI and contour coefficient for comparative analysis. The description of the three indicators is as follows:

Table 11. Labels of categories of different types of words

Belonging Category	classification label			
	simple1	simple2	difficult1	difficult2
movie	2	2	2	3
cater	1	1	1	2
tease	3	1	1	2
smelt	3	4	1	2
focus	1	1	1	2
today	3	3	5	2
watch	1	1	1	2
lapse	1	4	1	2
month	1	1	3	1
sweet	1	1	4	4
hoard	1	1	1	2
cloth	1	1	1	2
brine	1	1	1	2
ahead	1	1	1	2

1) CHI index: The CHI index is the ratio of the separation degree to the closeness of the data set. The separation degree of the data set is measured by the sum of the square of the distance between the center point of each class and the center point of the data set, and the closeness of the data is measured by the sum of the square of the distance between each point in the class and its center. The better the clustering effect, the greater the gap between the classes, the smaller the gap within the class, that is, the closer the class itself, the more dispersed the classes, and the larger the CH index value, the better the clustering effect.

2) DBI index: The Davies-Bouldin index (DBI) (proposed by David L. Davis and Donald Bouldin) is a clustering algorithm for evaluating metrics.

3) Contour coefficient: Silhouette coefficient is an evaluation method of clustering effect. The best value is 1 and the worst value is − 1. Values close to 0 represent overlapping clusters. Negative values usually indicate that the sample has been assigned to the wrong cluster, because different clusters are more similar.

Using Matlab programming to solve the index changes before and after the weathering of different word types, the specific analysis is as follows:

Table 12. K-Means + + clustering evaluation index comparative analysis table

category	Optimal classification number	CHI	DB I	contour coefficient
simple1	3	603.9374	0.41187	0.83888
simple2	4	21478.346	0.40376	0.80504
difficul1	5	387.4223	0.13902	0.88246
difficult2	4	156.951	0.11074	0.93844

According to the analysis of the above table, the number of common word classifications increases and the number of difficult word classifications decreases. When the CHI index of common words increases significantly and the difficult words decrease relatively, the clustering effect of different word types can be obviously judged by CHI index. For DBI index, common words. The DBI index decreased after a thousand-fold increase; for the contour coefficient, the ordinary words are further away from + 1 after weathering, while the difficult words are closer to + 1 after weathering, indicating that the correlation of difficult words is stronger, while the correlation of ordinary words is relatively weakened.

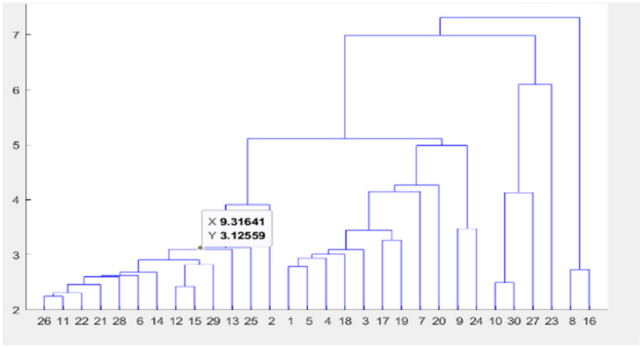

Fig 16. Pedigree diagram

The system cluster analysis model of word types:

Step 1: Take each sample as a class, each class has and only has one sample, and there are n classes.

Step 2: Calculate the distance between n samples, construct the distance matrix, and combine the two nearest classes into a new class.

Step 3: Calculate the distance between the new class and the current class. If the number of classes is equal to 1, proceed to the next step, otherwise go to step 2.

Step 4: Draw a clustering diagram.

Step 5: Determine the number of classes and the number of samples contained in each class, and explain the class accordingly.

Matlab is used for programming solution. In the programming solution, Euclidean distance is used to calculate the distance between two samples, and the shortest distance method is used to calculate.

Model based on statistical analysis and clustering method : factor analysis-clustering model .Considering the previous concepts, we can now define the problem complexity space. Given a standard training set D and the strategy of generating M new training subsets DSi, each subset contains only a percentage of the samples obtained from D. Through unified sampling and replacement, for each DSi, we can estimate the difficulty by complexity measure. Figure 2 shows the complexity space of a given classification problem, which is represented by two complexity measures (F1 and N2). Each element in this space corresponds to a subproblem (data subset, DSi) with its own difficulty level. It should be mentioned that the strategy for generating subsets of data from D plays an important role in the problem space representation. To explain more clearly, consider the projection neighborhood of a given test instance in the same space to find similar subproblems. In the example of Figure 2, the most similar subproblem is represented by DSi. We expect that the classifier trained on DSi can provide the necessary skills to handle the test instance.

We choose to distinguish clustering and portrait description to process features and understand the relationship between word attributes and various factors. We choose to compare K-Means algorithm and hierarchical clustering.

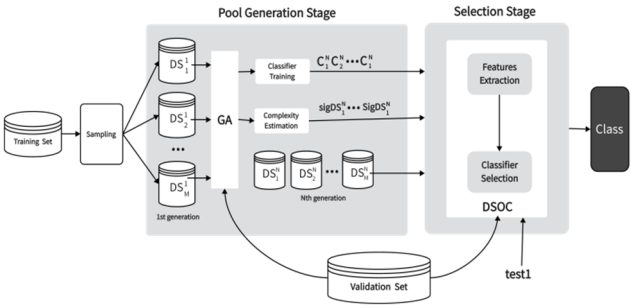

Fig 17. The most similar subproblem is represented by DSi.

The K-Means algorithm belongs to unsupervised learning, which classifies similar objects into the same cluster. Its central idea is to set a k value to randomly give K initial cluster center points, and allocate the data to the cluster contained in the nearest cluster center point. After all points have their own clusters, the new cluster center point is calculated again by averaging the values of all points in the cluster. Repeat the above operation until the center point is almost unchanged or repeat enough iterations. The algorithm is relatively fast and simple, and is suitable for mining large-scale data sets. It has high efficiency and scalability for large-scale data sets, and the overall application is very extensive.

Hierarchical clustering is a kind of very clustering algorithm, which is mainly based on the level, according to the different data categories but the similarity between each other to form a very hierarchical nested clustering tree. From this tree, we can obtain a view that can be directly observed, whether it is a bottom-up merge or a top-down classification, so that we can observe more vividly and directly.

The contour coefficient can also be called the Silhouette score. The main idea is to obtain the average value a by measuring the distance from the sample i to other points of the attribution degree of whether i belongs to the cluster, which can also be called similarity. At the same time, we measure the average distance between the sample and the sample in other cluster C as b, and we call b the non-attribution degree of the sample and cluster C also known as dissimilarity. When the obtained value is closer to 1, the clustering of the sample point is more appropriate.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Through the above manipulation, we can conclude that under these two methods, the best effect is when the number of clusters is 2. At the same time, the clustering results of the two methods are roughly the same. The following diagram is the hierarchical diagram obtained by using the hierarchical clustering method.

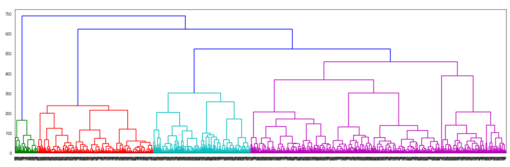

Fig 18. Hierarchical tree of hierarchical clustering

Hierarchical clustering method, also known as hierarchical clustering method, is a common method in cluster analysis. Its operation method is to take each sample itself as a class, and then the group with the smallest distance is first aggregated into a small class, and then the small classes aggregated together are merged according to the distance between the small classes, so as to continue, and finally merged into the large class. The advantage is that many classifications from coarse to fine can be obtained. The detailed operation process is shown in the following figure. The clustering map after dimension reduction is drawn with the three-dimensional coordinate system, and the factor analysis-clustering model is made by the projection of factor 0 and factor 1 on the two-dimensional plane.

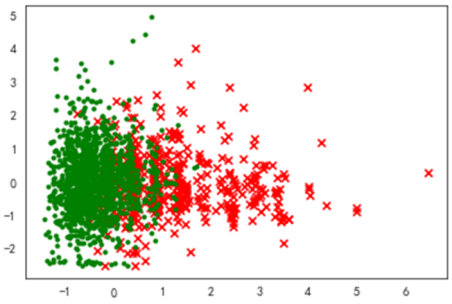

Fig 19. Factor analysis-clustering model

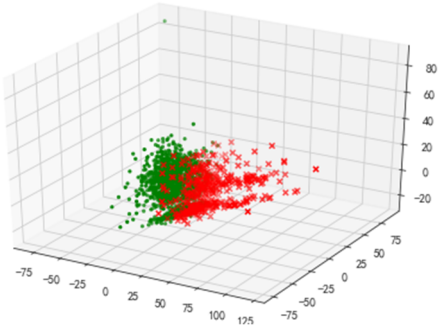

Fig 20. Factor analysis-clustering model

The following figure shows the corresponding results obtained through the clustering of SPSS. The elbow method is used to further obtain K (the number of clusters). The aggregation coefficient is the y-axis, and the number of clusters k is the x-axis. The elbow figure made by SPSS is shown below.

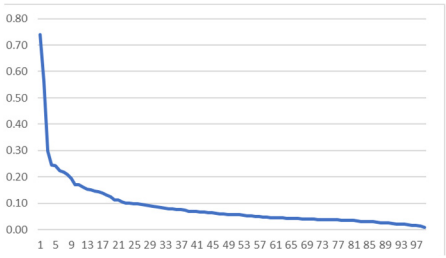

Fig 21. Elbow figure

According to the overall trend of the observation graph, it can be seen that when k is 5, the decreasing trend of the aggregation coefficient is obviously slowed down, so we think that the number of clusters is 5.

For the discriminant function method, we assume that ten words with different levels of difficulty are obtained according to six index factors. For a new word, we only need to substitute it into our discriminant equation to obtain the difficulty level it belongs to.

The sample mean of ten levels of word difficulty level in the d-dimensional feature space is:

$$M_i = \frac{1}{n_i} \sum y_k \in Y_i y_k \quad , i = 1,2, \ldots ,10$$

After mapping ω transform to one-dimensional feature space, the average values of various types are:

$$m_i = \frac{1}{n_i}\sum y_k \in Y_i y_k \quad , i = 1,2, \ldots ,10$$

After mapping, the ' within-class dispersion ' of various samples is defined as:

$$S_i^2 = \sum y_k \in Y_i (y_k - m_i)^2 \quad , i = 1,2, \ldots ,10$$

Obviously, we hope that after mapping, the larger the distance between the average values of the ten classes, the better, and the smaller the divergence within each class, the better. Therefore, the Fisher criterion function is:

$$J_F(\omega) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2 + S_3^2}$$

So the maximum solution is the best solution vector, namely Fisher linear discriminant method.

$$\omega = S_\omega^{-1}(M_1 - M_2 - M_3)$$

Where, w is the total dispersion matrix within the class. Ten kinds of discriminant functions are solved, and the determination coefficients and constants are obtained. The results are as follows:

Table 13. Coefficient result data table

	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	9.00
1	3.55 E-11	-4.98 E-12	-1.33 E-11	-1.46 E-11	-2.49 E-11	-4.93 E-13	-7.31 E-12	-5.35 E-11	-1.51 E-12
2	-8.51 E-11	2.26 E-11	7.09 E-11	5.10 E-11	8.96 E-11	1.92 E-11	3.36 E-11	1.04 E-10	8.69 E-11
3	0.00	0.00	0.00	0.00	-0.03	-5.37 E-5	0.00	0.00	-0.04
4	6.48 E-11	-9.39 E-11	-1.69 E-10	-1.69 E-10	-3.06 E-11	-8.58 E-11	-1.03 E-10	-1.54 E-10	-3.54 E-11
5	0.019	0.028	0.016	0.016	0.02	0.01	0.02	0.02	0.06
6	-2.66 E-11	4.10 E-11	3.59 E-11	4.49 E-11	7.52 E-11	1.01 E-11	2.20 E-11	1.03 E-10	3.45 E-11
constant	-3.33	-4.33	-3.47	-3.03	-4.41	-2.53	-2.78	-4.75	-11.2

$$F_1 = 3.55E^{-11}x_1 - 8.51E^{-11}x_2 + 6.48E^{-11}x_4 + 0.019x_5 - 2.66E^{-11}x_6 - 3.33$$

$$F_2 = -4.98E^{-12}x_1 + 2.26E^{-11}x_2 - 9.39E^{-11}x_4 + 0.028x_5 + 4.10E^{-11}x_6 - 14.33$$

.....

$$F_9 = -1.51E^{-12}x_1 + 8.69E^{-11}x_2 - 0.04x_2 - 3.54E^{-11}x_4 + 0.06x_5 + 3.45E^{-11}x_6$$

We bring the various indicators of the word EERIE predicted by the previous second question into the solution, and get the largest F6, indicating that its difficulty is above the medium level. This is through the classification method of discriminant function to quantify the qualitative difficulty, and the effect is very good. Model validation : manually label the data set, and sum the percentage of 1-4 attempts after normalization. If it is greater than 0.6, it is difficult to label, and less than 0.6 is simple. The clustering category results and manually labeled categories are calculated with an accuracy of 0.73.

7. Summarize Findings

Feature selection is a key part of data mining applications in the field of data analysis diversity. Most feature selection methods analyze the characteristics of classification, which is not enough to classify by feature selection technology. The traditional class involves classification, and the feature data is redundant, which is not conducive to further analysis of the database. The influence of using existing classes on feature analysis is negligible. Therefore, it proposes to obtain a new class or subclass by analyzing the small data of the feature, that is, the sub feature (SF) data of the corresponding instance in the traditional class. These data involve generating a limited number of important instances of new classes. Finding such instances with sub-features from any database is a challenging task. Therefore, this paper proposes an optimization model based on Lagrange multiplier to find such data and analyze new categories from traditional categories. Several algorithms have been used to interpret the sub-feature data. Theoretical methods such as domain-based and variance-based sub-features and sub-feature convergence are used to select sub-features with the effectiveness of the proposed model. In addition, several classifiers with search methods and statistical methods (i.e., local and global variance) are analyzed and classified by sub-feature data. Experimental results on different datasets show that the proposed model is beneficial to new classes based on the selected sub-feature data. The IEEE uses theoretical methods such as domain-based and variance-based sub-features and the convergence of sub-features to select sub-features that have the effectiveness of the proposed model. In addition,

several classifiers with search methods and statistical methods are analyzed and classified by sub-feature data. Experimental results on different datasets show that the proposed model is beneficial to new classes based on the selected sub-feature data.

8. Conclusion

In the initial phase of our mathematical modeling report, the team tackles the Wordle game challenge, presenting a holistic Prediction and Analysis (P&A) model. This model encompasses data pre-processing, grey relational analysis, and forecasting to predict results, analyze environmental effects, and classify solution words by difficulty. The team emphasizes assumptions, independence of player samples, and data credibility. Intelligent analysis techniques, including correlation coefficient diagrams and grey correlation degree analysis, unveil relationships in the dataset, culminating in the development of a weighted moving averages-based forecasting model for March 1.

Part 2 will delve into assumptions, notations, definitions, and further details on data pre-processing, grey relational analysis, and the forecasting model, contributing to a comprehensive understanding of Wordle dynamics. The second part introduces the Backpropagation (BP) neural network, implemented in MATLAB, addressing the problem as a complex function mapping issue. The report details training steps, parameter definitions, and feature extraction considerations, showcasing BP network prediction results and introducing the "Stickiness level" feature. Exploring various combinations of feature extraction and churn duration, the report tests machine learning algorithms, concluding with problem three's systematic clustering model using XGBoost and SVM.

Conclusively, the report demonstrates a thorough mathematical modeling approach involving feature extraction, machine learning, clustering, and systematic classification. The discriminant function quantifies the difficulty of the word "EERIE," proposing an optimization model based on Lagrange multipliers for enhanced sub-feature data understanding. Experimental results validate the model's effectiveness in classifying new categories, emphasizing the holistic approach to neural networks, clustering, and classification techniques.

References

[1] SPSSPRO. (2021). Scientific Platform Serving for Statistics Professional (Version 1.0.11), Volume 1, Pages 11.

[2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons, Volume 1, Pages 15-297.

[3] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts, Volume 1, Pages 1.

[4] Azzeh, M., Neagu, D., & Cowling, P. I. (2010). Fuzzy Grey Relational Analysis for Software Effort Estimation. Kluwer Academic Publishers, Volume 1, Pages 1-14.

[5] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, No. 14, pp. 281-297), Volume 1, Pages 281-297.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794), Volume 22, Pages 785-794.

[7] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley, Volume 1, Pages 1.

[8] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer, Volume 1, Pages 1.

[9] Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall, Volume 1, Pages 1.