Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Word data prediction based on statistical method

Lingwen Jiang *

School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China

* Corresponding author: 1372626528@qq.com

Abstract. Puzzle game wordle has been wildly popular since its launch, its simple and unique game rules have generated a lot of data while gaining a large number of users in a short period of time, making wordle prediction a hot topic. In this paper, we construct four models based on the properties of words: interpretive prediction report counting model, correlation judgment model, percentage prediction model and word difficulty category classification model to provide a solution for predicting words. In order to measure the difficulty of words, we evaluated the words from two aspects of commonality and structure, and defined sub-indicators such as the harmony of word pronunciation (both American and British), the number of vowels, the number of repeated letters, and the word frequency of words to measure the difficulty of words comprehensively. We built a BP neural network model for multivalue prediction and predicted for a given future puzzle word its correlation to the percentage distribution of each attempt (1, 2, 3, 4, 5, 6, X). We developed a K-means clustering model to classify the difficulty of words. We found that words fall into two categories. The closer a word is to the center of the category; the word belongs to that category. We suggest that in addition to players taking the initiative to advertise and attract traffic, newspapers should also take the initiative to promote the game (such as advertising), firstly warming up the publicity before the game goes live. According to the popularity trend curve fitted by the differential equation model, mass interest in emerging games rises rapidly and, after peaking for a short time, slowly declines due to the loss of novelty. Therefore, in order to maintain public enthusiasm as much as possible, publicity should be stepped up after the peak.

Keywords: ARIMA; Central-center log-ratio transformation; BP neural network; K-means clustering; Spearman correlation coefficient.

1. Background introduction

Wordle is a popular guessing game published in Time magazine, and now Wordle has been translated into 60 national languages, making it a worldwide hit.

The rules of the game itself are very simple, that is, players can play the game by giving a word 5 character length, if the guess is correct, 5 small squares will become the corresponding green, if the letter of the guess word does not appear in the word, the grey box will change, if the letter of the guess word appears in the word but not in the correct order, the yellow square will change and the player can be excluded by prompting , make multiple guesses. The website's data background, which counts the percentage of the total number of guesses in different rounds of guessing and not guessing, as well as the number of people playing each day in normal mode and hard mode, guessing the words provided by wordle with fewer guesses and predicting wordle's results, is becoming a very hot topic.

2. Model building

To better model word prediction, we first need to visualize the known data and complete the data cleaning. The steps are as follows:

2.1. Data Interpretation

In the table given, we get the word count of 2,022 wordle answers per day, the number of participants in the easy mode and the difficult mode, and the proportion of different correct answers to the total.

Through matlab, we can visualize the original data given by these questions, laying a good foundation for the subsequent data analysis and data processing, as shown in Figure 1.

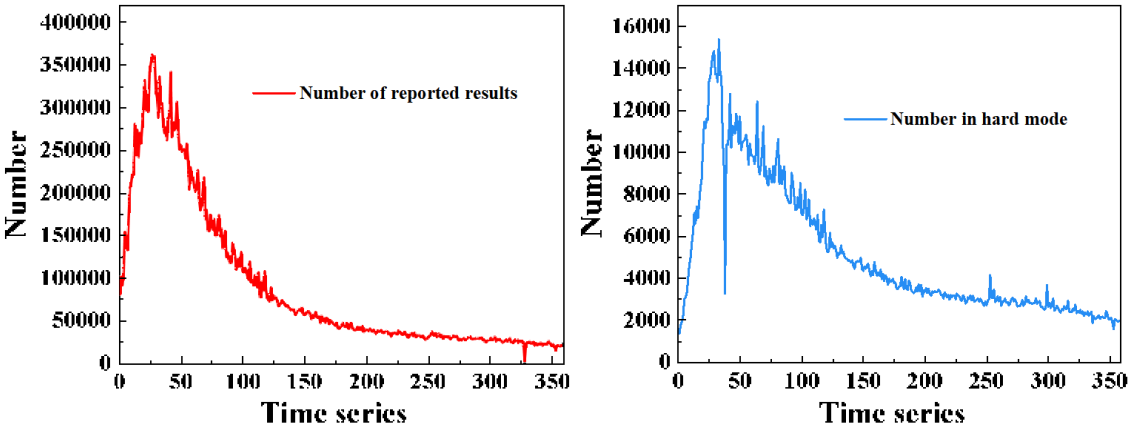

Figure 1. Represents the Number of reported results and the Number in hard mode.

They also show the distribution of correct answers for different keywords over the course of a year, from once to all. In the case of Slump and Macho, we plotted the percentage distribution of correct guesses relative to their bar chart, as shown in Figure 2:

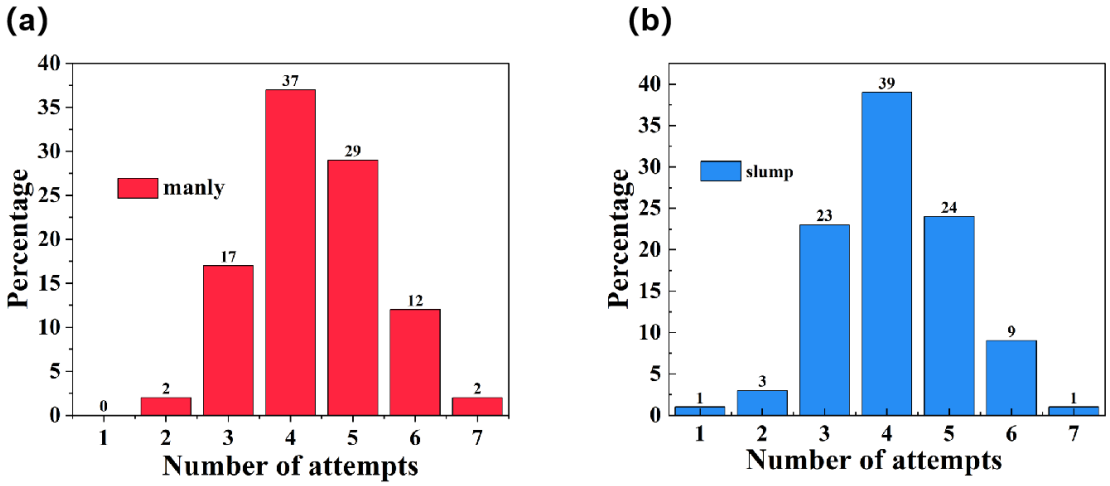

Figure 2. Histograms of tries for different words in wordle, (a) for "manly" and (b) for "slump".

2.2. Data Cleansing

In order to make the prediction results more accurate, we first clean and process the data. The data affecting the accuracy can be divided into two categories. The first type is abnormal data, including data with spelling mistakes and data with excessive fluctuation, which will affect our training model. Here, we remove abnormal digital and modular data. The second type is the scale error caused by rounding. Here, we use the central logarithmic ratio transformation to process the data, and the sum of the scale distribution is 100%, which is convenient for subsequent data processing and analysis.

For the first type, abnormal data is eliminated directly.

For the second type of data, we use the Central log-ratio transform to process. The specific processing details can be described by the following equation.

2.2.1. Intuitive understanding of the CLR transformation.

The index after the CLR transform takes into account not only the absolute size of a component, but also the relative size of that component with respect to other components, i.e. the phase importance. In order to better analyze the data and statistical regularities, perform the CLR transformation on the given data.

3. Differential equation model

First, through data analysis, we can easily see that, due to the rounding process, the proportion of the distribution of statistics is less than 100%, or the sum of the proportion of the statistical distribution is greater than 100%, and this error will have an impact on the subsequent modeling. Therefore, in order to mitigate the error, we clean up the data before building the model. Through CLR transformation, wordle's sum of success times distribution ratio is 100%, which is conducive to the establishment of subsequent models.

3.1. Prediction and interpretation of wordle data structure interval

It is easy to conclude from known data that the number of participants is closely related to the release time of the game. The trend in user data reflects the fact that Scrabble has finally gone from hot to normal. Through previous data collation, we can see that the number of people playing Scrabble increased rapidly over time, then slowly declined. Eventually it leveled off or gradually died out. Differential equation models and ARIMA time prediction models can be used to predict wordle's future data regions.

3.1.1. Mathematical descriptions of differential equation models and ARIMA models.

▲ Differential Equation Model:

First we analyze 4 variables: new users, playing users, playing users and tweets, which can be represented by the vector [xt1,x, xt,xt3t2]. At the same time, we can define the twitter rate, which represents the number of people who tweet as a percentage of total players.

$$x_{t3} = \alpha_{twitter} x_t$$

We can assume that the sum of the following three variables remains constant:

$$x_{t1} + x_t + x_{t2} = 1500000$$

In order to better describe the number conversion rules of the above three types of users, the change rates of these three types of players are discussed respectively. We can describe it in mathematical language as follows:

At some point (very small) after time t, new users will see new tweets, and some of them will try to play the game, so new users will enter the playing user group, and the coefficient for new users to be influenced by twitter is; $\alpha_{twitter-spread}$ And there is a subset of players who will not continue to play the game, and these players will naturally become part of the playing group. Let's say the ratio of the total number of old players to the total number of new users at time t is, then we can represent the change in the number of users playing the game amongΔt. $\alpha_{leaving-rate}$

$$\Delta x_t = (\alpha_{twitter-spread} x_{t3} x_{t1} - \alpha_{leaving-rate} x_t) \Delta t$$

When both sides are of the same genus and approach infinitesimal value, the expression of the change rate of the number of game users is:

$$\frac{dx_t}{dt} = (\alpha_{twitter-spread} x_{t3} x_{t1} - \alpha_{leaving-rate} x_t)$$

At some point after time t (very small), the users who played the game want to play the game again, and this group is the same as the new users, so this group of people will enter the new user group. Let's say the conversion factor from the played user to the new user is; As can be seen from the figure above, there will be some users who will not continue to play the game. $\alpha_{refrashed}$ So the rate of change in the number of players can be expressed in the same way:

$$\frac{dx_{t2}}{dt} = (-\alpha_{refrashed} x_{t2} + \alpha_{leaving-rate} x_t)$$

Similarly, from the above derivation, we can use the same method to find the expression of the rate of change of new users:

$$\frac{dx_{t1}}{dt} = (\alpha_{refrashed} x_{t2} - \alpha_{twitter-spread} x_{t3} x_{t1})$$

Finally, the system of differential equations listed is as follows:

$$x_{t3} = \alpha_{twitter} x_t$$

$$\frac{dx_{t1}}{dt} = (\alpha_{refrashed} x_{t2} - \alpha_{twitter-spread} x_{t3} x_{t1})$$

$$\frac{dx_t}{dt} = (\alpha_{twitter-spread} x_{t3} x_{t1} - \alpha_{leaving-rate} x_t)$$

$$\frac{dx_{t2}}{dt} = (-\alpha_{refrashed} x_{t2} + \alpha_{leaving-rate} x_t)$$

$$x_{t1} + x_t + x_{t2} = 1500000$$

The specific solution steps are as follows:

Step p1: Set the initial values of all variables:

$$[x_{t1}, x_t, x_{t2} x_{t3}] = [150000, 0, 0, 0]$$

Step2: Set the upper and lower limits for all coefficients, then use python's dichotomy program to find the optimal coefficients.

Step3: At the end of the iteration, print the curve of the change in the number of tweets.

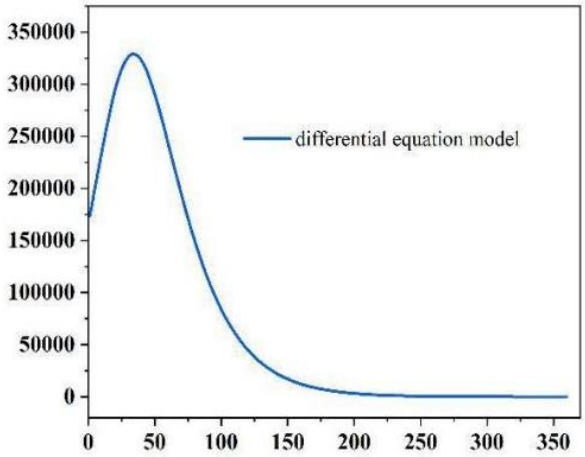

Figure 3. Tweet number change

As can be seen from the figure 3, the above inference can qualitatively explain the change rule of the number of reports, but there is a certain gap in the quantitative analysis, so the ARIMA model should be used to predict the change range of the number of reports.

4. ARIMA prediction model for time series

▲ARIMA model:

In order to choose the most suitable mathematical model for time series prediction, we try to use several different models for prediction analysis, such as exponential smoothing model, simple model, linear trend model and damped trend model. However, since none of these models can pass the significance test, and the ARIMA model can maintain good performance under the premise of passing the significance test, we finally choose to use the ARIMA model for time series prediction.

ARIMA(p,d,q) is an autoregressive moving average model, and its mathematical expression is as follows:

$$\phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$$

$$\nabla^d = (1-B)^d,$$

$$\phi(B) = 1 - \phi_1 B - \cdots\cdots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \cdots\cdots - \theta_q B^q$$

$\nabla^d = (1-B)^d$ Represents the difference set of order d in the ARIMA model, represents the content AR(p) of the set in the ARIMA model, and represents the MA(q) set in the ARIMA model.$\phi(B)\Theta(B)$

AR(p) represents the P-order autoregressive model. Autoregression can only reflect the impact of historical factors of game software on the number of reports, but the market of game software is heavily influenced by social factors, so we use ARIMA to avoid the limitations of autoregression.

However, since the time series of word data can be an ordered unit-root process, the time series of word data can be unstable. Therefore, it is necessary to differentiate the data, transform it into a stable time series, and then build a prediction model.

$$E(\varepsilon_t) = 0$$

$$Var(\varepsilon_t) = \sigma_E^2$$

Where, represents the white noise of the time series at time t. For the ARIMA model, the mathematics of the white noise must be 0 and the variance as small as possible.$\varepsilon_t E(\varepsilon_t)\varepsilon_t Var(\varepsilon_t)$

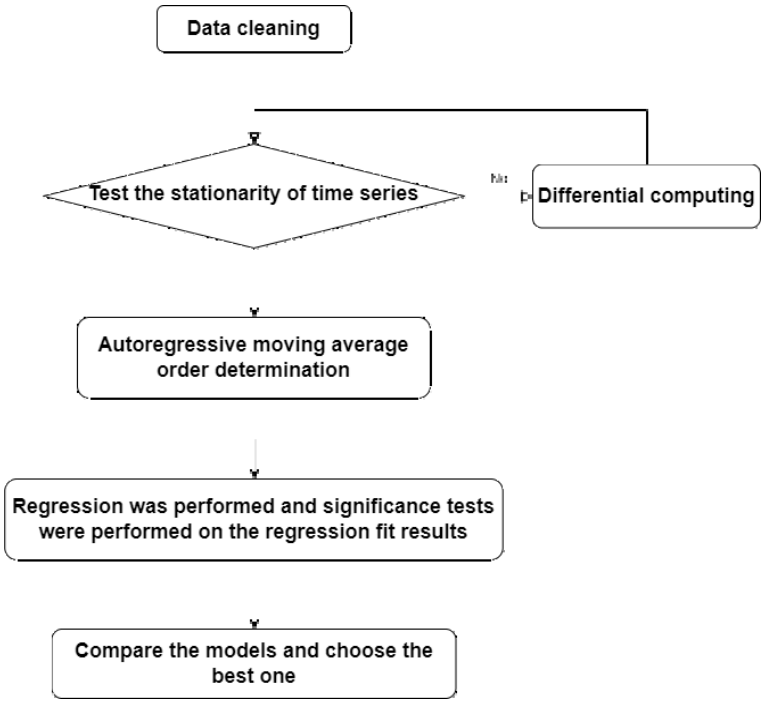

Figure 4. Depicts the process of the ARIMA model and shows how the data is processed at each step.

By comparing the predictions and comparing the results of the significance test, ARIMA(0,1,6) is finally determined to predict the time series, as shown in Figure 4 and Figure 5.

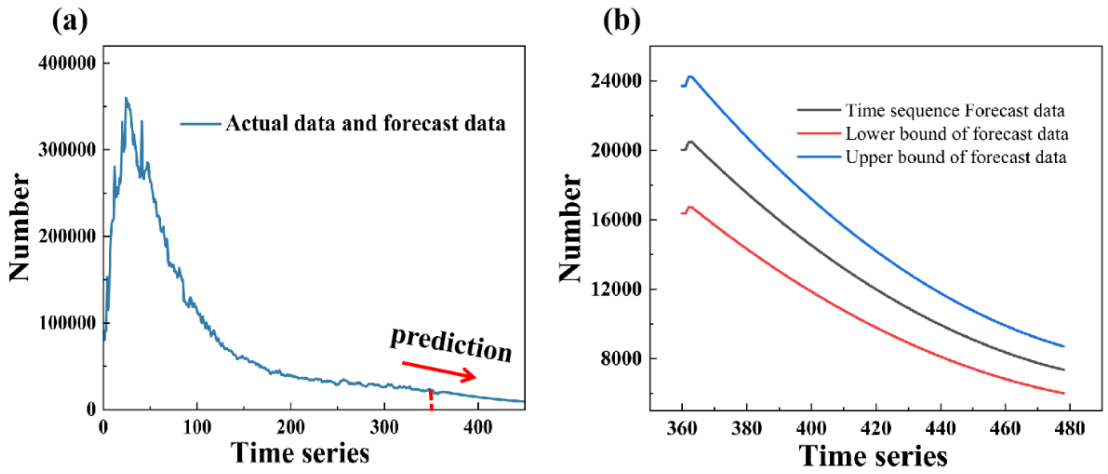

Figure 5. (a) depicts the data given in the form of combined prediction data, and (b) the results of time series prediction data with lower and upper boundaries.

5. BP neural network

5.1. Model description and mathematical principal analysis

BP neural network is a nonlinear algorithm, which is essentially a multivariate composite function with good nonlinear mapping ability, composed of input layer, output layer and hidden layer. The specific topology is shown in the figure 6 below.

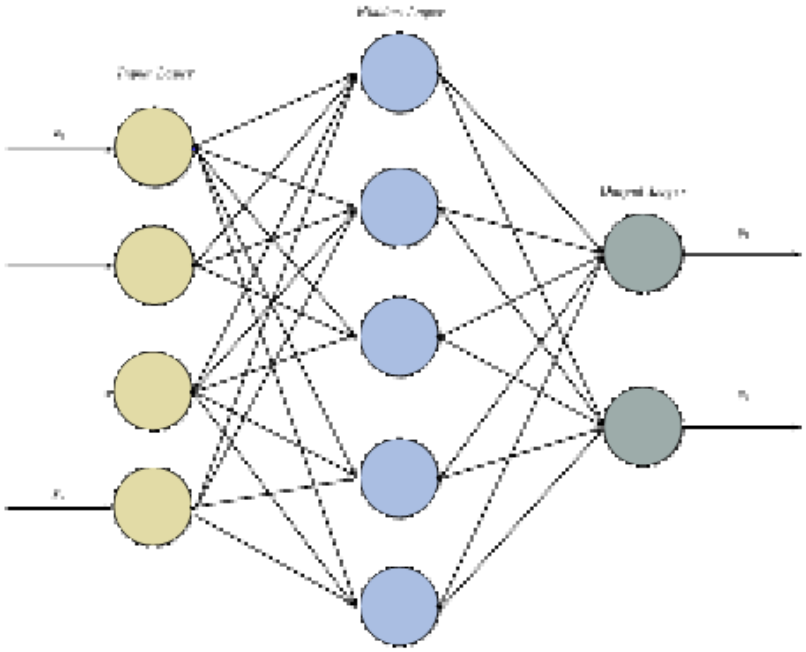

Figure 6. Represents the internal mechanism of the BP neural network, respectively representing the input layer, hidden layer and output layer from left to right.

The core steps of BP neural network algorithm are divided into two parts: forward propagation and back propagation. Forward propagation is used for prediction, and backward propagation is used for adjusting network parameters. After the BP neural network is established, the word frequency, two harmonic parameters, the number of repeated letters and the number of vowels of the word EERIE can be input into the BP neural network. Then the percentage of its associated (1, 2, 3, 4, 5, 6, X) is inverted through the central logarithm to obtain the predicted percentage formula(1) of the word EERIE as shown below.

$$w_j = v_j - v_7, j = 1,2\cdots 6$$
$$\xi_j = \frac{e^{w_j}}{1+\sum_{i=1}^{6} e^{w_j}}$$
$$\xi_j = \frac{1}{1+\sum_{i=1}^{6} e^{w_j}} \quad (1)$$

5.2. Analysis of prediction results:

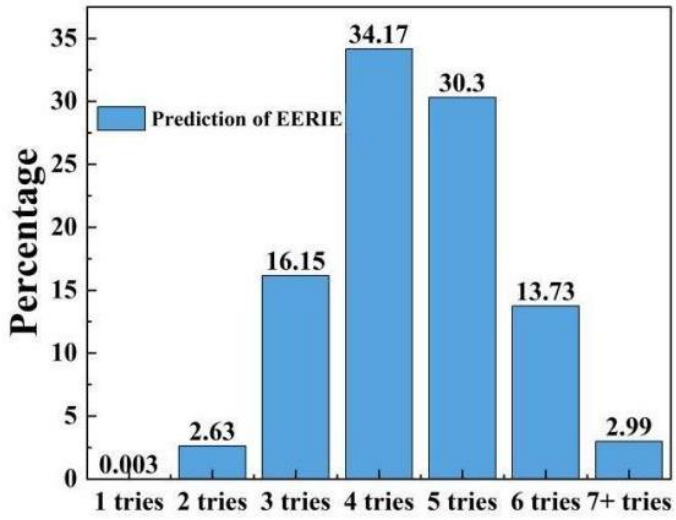

Figure 7. The histogram represents the predicted results of the word EERIE, the X-axis represents the Game Round won by the player, and the Y-axis represents the percentage of players.

By comparing the predicted percentage data to the actual measured percentage data, this paper finds that there are similarities between the two. As this paper uses the data after the central log-ratio transformation for neural network calculation and analysis, when predicting the data on March 1, 2023, it can ensure that the cumulative sum of the percentages in each case of the predicted results is 100%, which is consistent with the objective facts. To sum up, the prediction method proposed in this paper is applicable to the prediction of percentage data of other words , as shown in Figure 7.

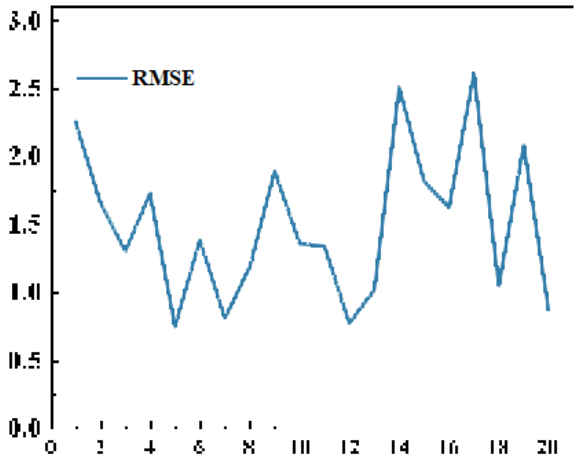

Figure 8. The RMSE line

As can be seen from the figure 8, the RMSE line chart is close to the abscissa, indicating that the error is within the allowable range.

6. K_means clustering method

K-means mathematical model description , as shown in Figure 9.

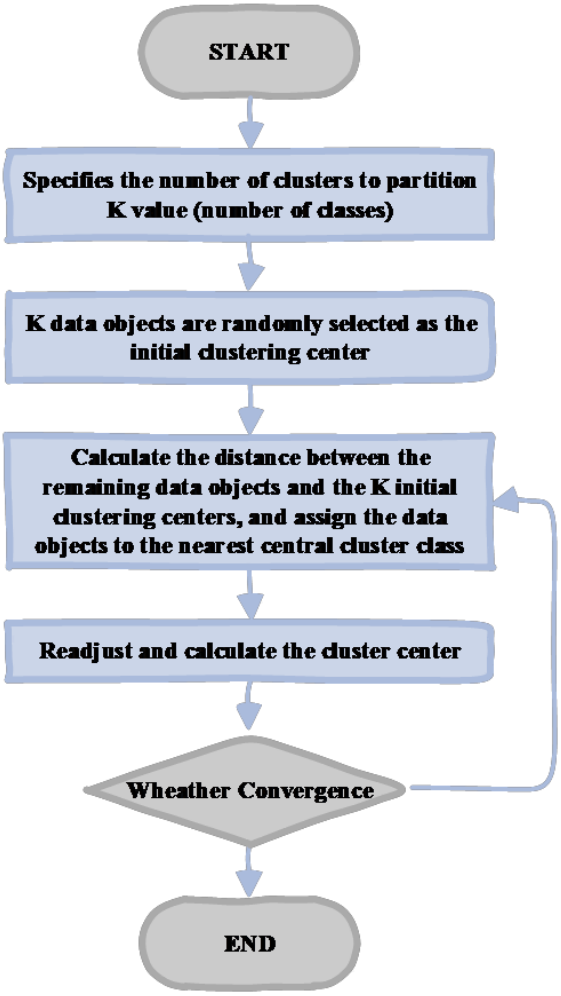

Figure 9. K-Flow chart of mean clustering algorithm

7. Model evaluation

7.1. Advantages

1. This paper uses the central logarithm conversion to process the percentage data, which can avoid the interference of the fixed sum value of the percentage to the statistical method and make the data more accurate.

2. The percentage data after the central logarithm transformation is used to ensure that the sum of the predicted percentages is 100%, which makes the predicted data more real and reliable.

3. The sensitivity analysis of the classification model by using stochastic perturbation method can prove the robustness of the model.

4. In this paper, we can only do qualitative analysis, but not quantitative analysis, when we use differential equations to explain changes in report numbers.

7.2. Weaknesses

1. BP neural network and K_means clustering method have been around for decades, and their improved algorithms have emerged endlessly. For example, the improved neural network algorithm has faster training speed and higher stability than the traditional neural network algorithm. This is a problem we did not take into account.

7.3. Further Discussion

By reading the literature, we can improve BP neural network and K-means clustering algorithm to achieve better prediction effect. At the same time, we can search more data to further optimize the model, so that it has better prediction effect and better adaptability at the same time.

References

[1] Guo, Wu, L., & Wang, L. (2024). Prediction and Analysis of Non-linear Data based on BP Neural Network.

[2] IEEE Xplore. This article provides insights into the application of BP neural networks for prediction in nonlinear data analysis, which is relevant to the BP neural network model built for Wordle prediction.2. K-Means Clustering for Word Difficulty Classification in Wordle. Although a specific paper is not referenced, you can cite various resources on K-means clustering applied toword difficulty classification in the context of Wordle, for example, you can reference tutorials or technical reports that explain the implementation and application ofK-means clustering in similar scenarios.

[3] Wordle Strategies and User Behavior Analysis, you can reference online resources, forums, or blogs that discuss Wordle strategies and user behavior patterns. These resources can provide insights into player strategies, which can inform your prediction models.

[4] Measuring Word Difficulty in Word Games. While there may not be a specific paper directly related to Wordle, you can reference studies that explore methods for measuring word difficulty in word games. These studies often consider factors like word frequencypronunclation, and letter patterns.

[5] Differential Equation Models for Predicting Popularity Trends of Online Games. You can cite studies that use diferential equation models to predict popularity trends, especially in the contextiofonline games. These models can provide insights into maintaining public enthusiasm and timing publicity efforts.

[6] Advertising and Promotion Strategies for Online Games. Reference articles or case studies that discuss advertising and promotion strategies for online games, focusing on how to maintain player interest and attract new users.