Word Data Research and Prediction Based on Wordle Game

Qi Yang[1], Yifan Fang[1,*], Yu Zheng[1]

[1]College of Electronic Engineering and Artificial Intelligence, South China Agricultural University, Guangzhou, China
[*]Corresponding author

Abstract: *As a very popular game in recent years, Wordle contains many interesting rules. This paper hopes to get a mathematical model that can measure the difficulty of words by analyzing the words that appear in Wordle, and use this model to predict the difficulty of the word "EEIRE". First of all, we process the data of response times. The data with 1-6 times of answer is used as negative indicator, and the data with more than 6 times of answer is used as positive indicator.Then, we using the comprehensive rank-sum ratio evaluation method, the RSR values of each sample are calculated, and then sorted and divided into three difficulty levels: 1, 2 and 3, where "1" means simple and "3" means difficult. Finally, we code the words and train them with CNN. Finally, we predict the word EEIRE to be of medium difficulty.*

Keywords: *RSR, Convolutional Neural Network*

1. Introduction

1.1. Problem Background

The word game wordle has taken the world by storm recently. Many readers may be familiar with this interface, which looks a lot like a crossword puzzle. Both wordle and crossword are classified as crossword puzzles. Crossword puzzles have a long history and a wide audience, and they vary widely. Wordle, by contrast, is much simpler in both difficulty and form. Offered by the New York Times, the simple rules combined with the ease of opening and playing have made many people obsessed with this game. The game is updated daily and the player's sole goal is to guess a five-letter word in six tries. You can even clock in every day and post it in your community if you get it done first. On the other hand, it greatly increases people's communication, as shown in Figure 1.

Figure 1: Wordle game.

2. Assumptions and Explanations

Considering that practical problems always contain many complex factors, first of all, we need to make reasonable assumptions to simplify the model, and each hypothesis is closely followed by its corresponding explanation:

Asumption: Assume that the percentage of (1,2,3,4,5,6, X) can objectively reflect the difficulty of words.

Explanation: Since the RSR method is used to rank the difficulty of a word in Model 3, only the percentage (1,2,3,4,5,6,X) correctly reflects the difficulty of the word can the result of the ranking be objective and persuasive.

3. Model

3.1. Word feature of data

This paper uses the number of each letter in the word to represent the word features, that is, 26 features are extracted from a word. If the word apple contains two letters p, one letter a, one letter l and one letter e, the characteristic value corresponding to the letter p is 2, the characteristic value corresponding to the letter a is 1, the characteristic value corresponding to the letter l is 1, the characteristic value corresponding to the letter e is 1, and the characteristic value corresponding to the other letters is 0, as shown in Table 1.

Table 1: Characteristic data of the word apple.

a	b	c	d	e	f	g	h	i	j	k	l	m	o	p	q	r	s	t	u	v	w	x	y	z
1	0	0	0	1	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0

3.2. Rank-sum Ratio Comprehensive Evaluation Method

The general process of RSR[1] is to rank the benefit-type indicators from small to large, rank the cost-type indicators from large to small, calculate the rank sum ratio, and finally make statistical regression and rank by grades. The dimensionless statistic RSR is obtained by rank transformation; On this basis, the distribution of RSR is studied by using the concept and method of parameter statistical analysis; The RSR value is used to directly rank or rank the quality of the evaluation object. The advantage of this is that the algorithm is based on non-parametric method, has no special requirements for the selection of indicators, and is applicable to various evaluation objects. Since the numerical value used in the calculation is rank, it can eliminate the interference of outliers, as shown in Figure 2.

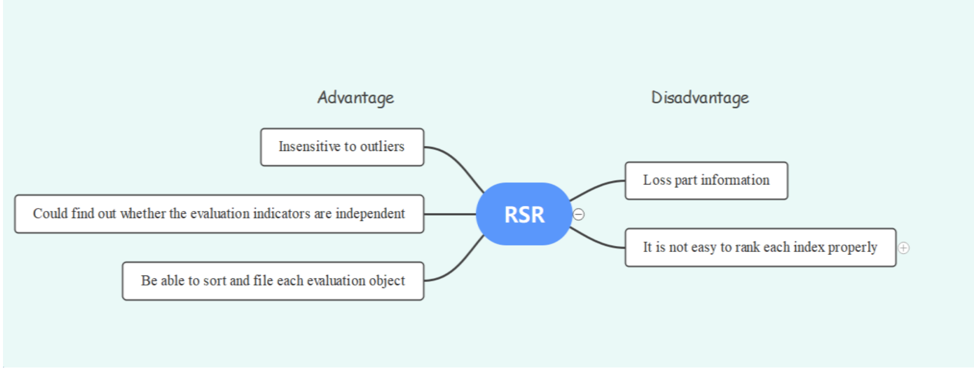
Advantage
Disadvantage
Insensitive to outliers
Could find out whether the evaluation indicators are independent
Be able to sort and file each evaluation object
RSR
Loss part information
It is not easy to rank each index properly

Figure 2: Advantages and disadvantages of RSR.

At the same time, the index value is ranked in a way similar to linear interpolation (also similar to the distance method of superior and inferior solutions) to improve the shortcomings of RSR method. There is a quantitative linear correspondence between the rank and the original index value, thus overcoming the disadvantage that the quantitative information of the original index value is easily lost when RSR method is ranked.

3.3. Convolution Neural Network

These words has been coded. At this time, the encoding form is a 26-bit "bit stream" (the value of each bit is not necessarily 0 or 1). This is a typical "grid" data structure, which is similar to the matrix after image digitization. It can be considered as a "matrix" with one row and 26 columns. CNN has excellent performance in the classification task of processing matrix structure (such as image)[2]. In order to make a comparison, we also applied several classical machine learning classification algorithms to explore the performance differences between machine learning and deep learning algorithms in classification tasks, as shown in Figure 3.

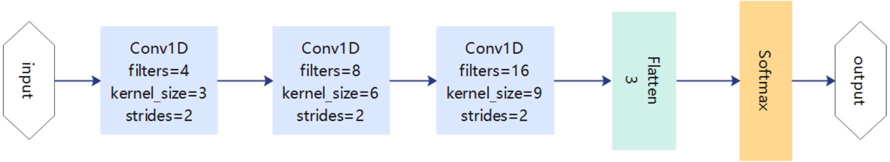

Figure 3: CNN Structure Diagram.

The matrix of 1 * 26 is transposed and input into the constructed and well-trained CNN. After the feature is extracted through three layers of convolution operation, a full-link layer is passed, and then an array of three elements is output through the Softmax activation function. These three elements are the probabilities of the three grades of the input word.

In the training preparation, Adam is selected as the optimization solver, the learning rate is 0.0001, and the loss function is cross entropy; Each batch is fed with 64 samples (batch_size=64), with a total of 300 rounds of training (epoch=200).

4. Conclusions

4.1. Result

First, calculate the RSR value, then rank the RSR value according to the calculation results, then calculate the RSR distribution, and finally carry out regression fitting.Regression fitting results are shown in Figure 4:

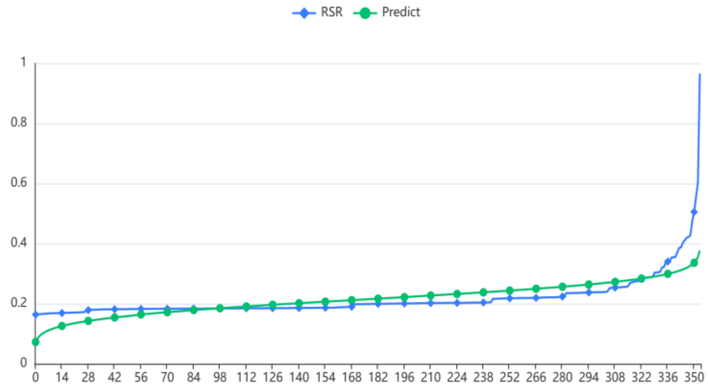

Figure 4: Regression Fitting Results.

Finally, the table of critical values for sorting by grades is obtained. According to this critical value, all 359 sample words are divided into three grades, namely, 0, 1, and 2, representing "simple", "medium", and "difficult", as shown in Table 2.

Table 2: Critical value of grading.

grade	Percentile threshold	Probit threshold	RSR threshold (Fitting)
1	< 15.866	< 4	< 0.430
2	15.866 ~	4 ~	0.430 ~
3	84.134 ~	6 ~	0.575 ~

In recent years, on the basis of decision tree and random forest algorithm, the machine learning field has combined some bagging or boosting methods and proposed algorithms such as XGboost[3] and LightGBM[4] with excellent performance. Test the performance of these two algorithms, traditional BP neural network and logistic regression on the same training set and test set.

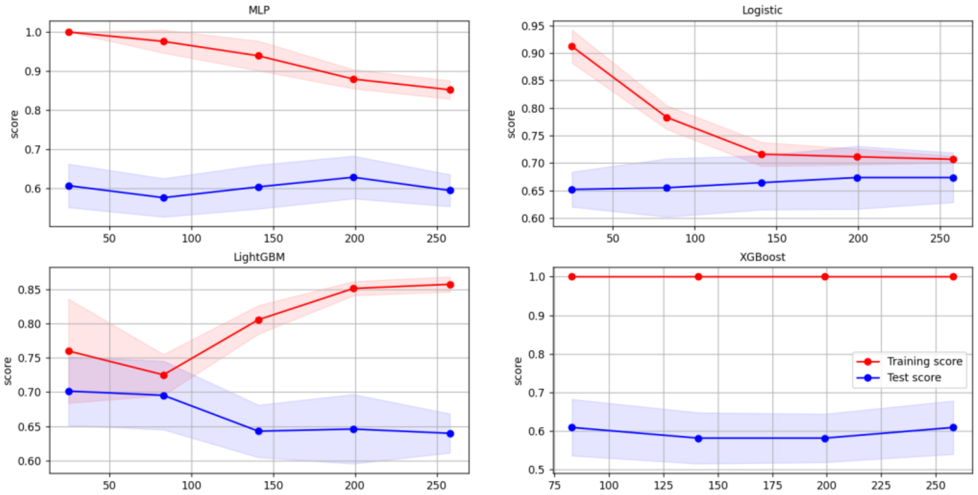

Figure 5: Machine Learning Result Chart.

Observe the above figure 5 and find that the performance of machine learning algorithms is generally poor. LightGBM's score in the test set is only about 0.65, while XGboost fits well in the training set, but its effect in the test set is poor, which may be due to over-fitting. On this basis, using the same training set and test set, using CNN to learn, it is found that its F1 score in the training set is 0.807, and the F1 score in the test set is 0.839. The effect is relatively good.

Finally, the word "eeire" is coded and predicted using the trained CNN model. The result is that the difficulty level of the word "eeire" is 1, which means that the word is a "Medium" word.

4.2. Strength

- Innovatively proposed the encoding of words, thus quantifying text data into numerical data.
- When processing the encoded words, CNN is used to extract the features of the gridded data, thus completing the classification task excellently.

References

[1] Zhou Q., Shao Z., & Lin H. (2018). Grade evaluation of user harmonic hazards based on rank sum ratio comprehensive evaluation method. Power Capacitor & Reactive Power Compensation.
[2] Chua L., Roska T., Kozek T., & Á Zarándy. (1993). The CNN paradigm—a short tutorial.
[3] Chen T., Tong H., & Benesty M. (2016). Xgboost: extreme gradient boosting.
[4] Qi M. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems. Curran Associates Inc.