Prediction of Wordle report distribution results based on PSO-LBGM prediction model

Yuqing Zhang *

School of Information and Control, Keyi College of Zhejiang Sci-Tech University, Shaoxin, China, 312369

* Corresponding author: 19957544097@163.com

Abstract. Accurate prediction of Wordle report headcount distribution is an important reference for Wordle's later word difficulty setting to expand the number of players. In order to increase the number of purchased products, the prediction model of the number of people distribution results is constructed by using Lightweight Gradient Boosting Machine (LightGBM), and PSO is used to optimize the hyperparameters of the LightGBM model. The historical data were preprocessed and the word attributes were extracted using unique thermal coding to build prediction models based on PSO-LightGBM, LightGBM and LSTM. The results show that the mean absolute percentage (MAPE) of the training and test sets predicted by PSO-LightGBM for (1, 2, 3, 4, 5, 6, X) is 0.531%, 0.410%, respectively. and the model was more accurate in predicting the number distribution results than LightGBM and LSTM models.

Keywords: Number of people distribution results, Prediction Model, PSO, LightGBM.

1. Introduction

Wordle Accurate prediction of the results of the reported number distribution can help Wordle operators to optimize the setting [1] of word association rules and increase the number of products purchased. The proportion of the number of people in different difficulty models is the key to determining Wordle revenue. The prediction error directly affects the difficulty of Wordle subsequent word setting and the number of products purchased, which is of great significance to the setting of Wordle revenue strategy. Traditional prediction methods are based on linear regression, such as time series method [2-4] and gray prediction method [5] respectively have some defects.

2. Data pre-processing

2.1. Wordle word property analysis - one hot code

We must first extract the attributes of the words [6]and address the impact analysis of their relevance to the number of enrolments. When referring to the attributes of the word itself, we can think of the length of the word, the commonness of the word, the word's morphological structure, the semantic range of the word, the phonemes of the word, and so on. However, since the information and logic involved in the extraction of these ways are more complicated, we use a very widely recognized way of attribute extraction in machine learning applications - unique hot coding Dok-Heat coding is a common technique for converting categorical variables into binary vectors, which is illustrated in Table 1 below: assume there are four samples (rows), each with three features (columns)

Table 1. The sample of eigenvalues

	Feature_1	Feature_2	Feature_3
Case_1	1	4	3
Case_2	2	3	2
Case_3	1	3	2
Case_4	2	1	1

Feature_1 has two values, 0 or 1. Feature_2 and Feature_3 each have four values. one-hot coding is to ensure that only one bit of a single feature of each sample is in state 1 and the others are 0. The above states are coded in one-hot as shown in Table 2 below.

Table 2. Eigenvalues after one-hot processing

	Feature_1	Feature_2	Feature_3
Case_1	0 1	1 0 0 0	1 0 0
Case_2	1 0	0 1 0 0	0 1 0
Case_3	0 1	0 1 0 0	0 1 0
Case_4	1 0	0 0 0 1	0 0 1

The specific steps for its application to the word vector domain are.

Step1: Count all the words in the corpus, and then number each word.

Step2: For each word to create a V-dimensional vector, each dimension of the vector represents a word, that is, the value of the dimension on the corresponding numbered position is 1, and all other dimensions are 0. For example, there is the text ["US", "Europe"] will be replaced by a unique heat encoding, as: [010, O01]

Overall, unique thermal coding is a simple and effective technique for converting categorical variables to numbers, which helps to improve the performance and accuracy of the model, and for extracting letters in the same way as words are extracted.

Based on the above method, we have extracted all wordle words with one hot code, and the extracted results are shown in Figure 1below.

Word	a	b	c	d	e	f	..	v	w	x	y	z
manly	1	0	0	0	0	0	...	0	0	0	1	0
molar	1	0	0	0	0	0	...	0	0	0	0	0
havoc	1	0	1	0	0	0	...	1	0	0	0	0
impel	0	0	0	0	1	0	...	0	0	0	0	0
condo	0	0	1	1	0	0	...	0	0	0	0	0
judge	0	0	0	1	1	0	...	0	0	0	0	0
extra	1	0	0	0	1	0	...	0	0	1	0	0
poise	0	0	0	0	1	0	...	0	0	0	0	0
---	---	---	---	---	---	---	...	---	---	---	---	---
tangy	1	0	0	0	0	0	...	0	0	0	1	0
abbey	1	2	0	0	1	0	...	0	0	0	1	0
favor	1	0	0	0	0	1	...	1	0	0	0	0
drink	0	0	0	1	0	0	...	0	0	0	0	0
query	0	0	0	0	1	0	...	0	0	0	1	0
gorge	0	0	0	0	1	0	...	0	0	0	0	0
crank	1	0	1	0	0	0	...	0	0	0	0	0
slump	0	0	0	0	0	0	...	0	0	0	0	0

Figure 1. Results of word attributes extracted by one hot

3. The basic funamental of LightGBM

3.1. The structure of LGBM

(1) LightGBM algorithm

LightGBM algorithm is an improved XGB efficient algorithm proposed by Microsoft, which is proposed mainly to overcome the problem of slow running speed and large memory consumption of GBDT in massive data. LGBM optimization mainly contains Histogram-based decision tree

algorithm, with depth limitation of growth-by-leaf strategy, histogram to do differential acceleration, direct support for category features, and multi-threaded optimization in five areas.

The basic idea of Histogram's decision tree algorithm is to first discretize the continuous floating point feature values into k integers and construct a histogram of width k. When traversing the data, the statistics are accumulated in the histogram based on the discretized values as indexes. After traversing the data once, the histogram accumulates the required statistics, and then traverses the histogram to find the optimal splitting point based on the discretized values of the histogram, as shown in Figure 2.

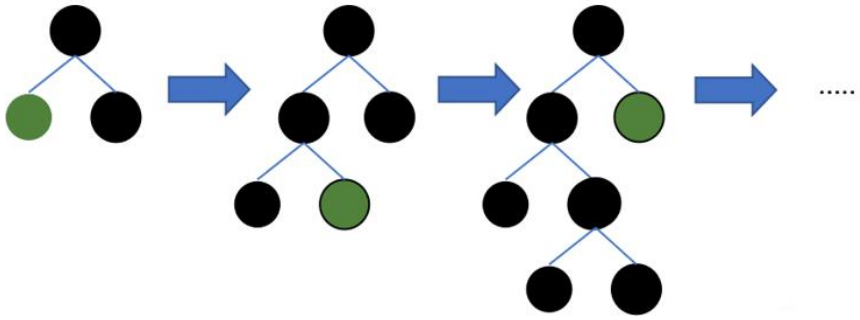

Figure 2. Decision tree histogram

The basic idea of histogram differencing acceleration is that the histogram of a leaf can be obtained by differencing the histogram of its father node from the histogram of its brother. Usually to construct a histogram, it is necessary to traverse all the data on that leaf, but histogram differencing only requires traversing K buckets of the histogram. The principle of the algorithm is shown in Figure 3.

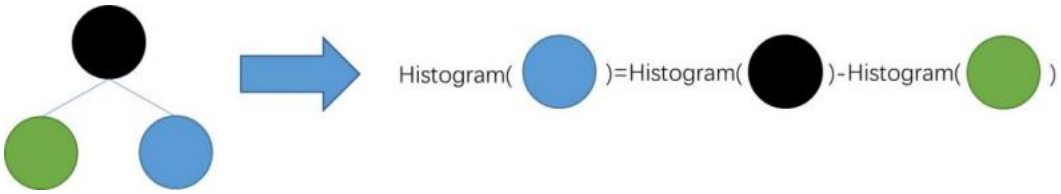

Figure 3. Histogram for difference principle

The LGBM algorithm directly supports category features. Most machine learning algorithms cannot directly support category features and need to do further transformation, LGBM optimizes the support of category features and can directly input category features, which improves the efficiency of space and time.

The LGBM algorithm also has the feature of supporting efficient parallelism, which supports both feature parallelism and data parallelism. The main idea of feature parallelism is that different machines find the optimal segmentation points in different feature sets separately, and then synchronize the optimal segmentation points among machines. Data parallelism, on the other hand, allows different machines to first construct histograms locally, then perform a global merge, and finally find the optimal segmentation points on top of the merged histograms.

(2) Particle swarm algorithm (PSO) parameter tuning

By constructing good machine learning and deep learning models for problems such as regression, good results are achieved, but usually there are often a large number of hyperparameters in the models, which refer to parameters that cannot be learned from the data during training. The setting of hyperparameters will have a direct impact on the performance and performance of the model. In determining the model hyperparameters, the optimization search is often performed in the manner of Figure 4.

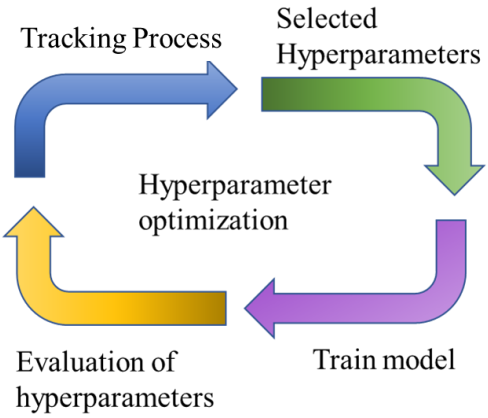

Figure 4. Hyperparameter optimization

The standard PS0 algorithm is a global search algorithm that concentrates on the principles of "population and "optimization", and finds the optimal value through the adaptation of the particles. space at a certain speed, which is dynamically adjusted by the individual flight experience and the flight experience of the population. In each iteration, each particle adjusts its flight speed and position according to the following.

$$v_{ij}(t+1) = wv_{ij}(t) + c_1 r_{ij}\left[p_{vj}(t) + x_{ij}(t)\right] + c_2 r_{2j}\left[p_{gj}(t) - x_{ij}(t)\right] \quad (1)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (2)$$

j denotes the j-th dimension of the particle; i denotes the i-th particle; t denotes the t-th generation; c_1, c_2 denote the acceleration constants; r_1~U (0,1), r_2~U (0,1), are two mutually independent random functions.

The steps to optimize the parameters based on PSO are shown below.

The PSO algorithm has shown good robustness and effectiveness in solving function-valued optimization problems in real number space. A particle swarm algorithm-based parameter search process is performed according to the above process to obtain the optimal model and perform text classification [7-8].

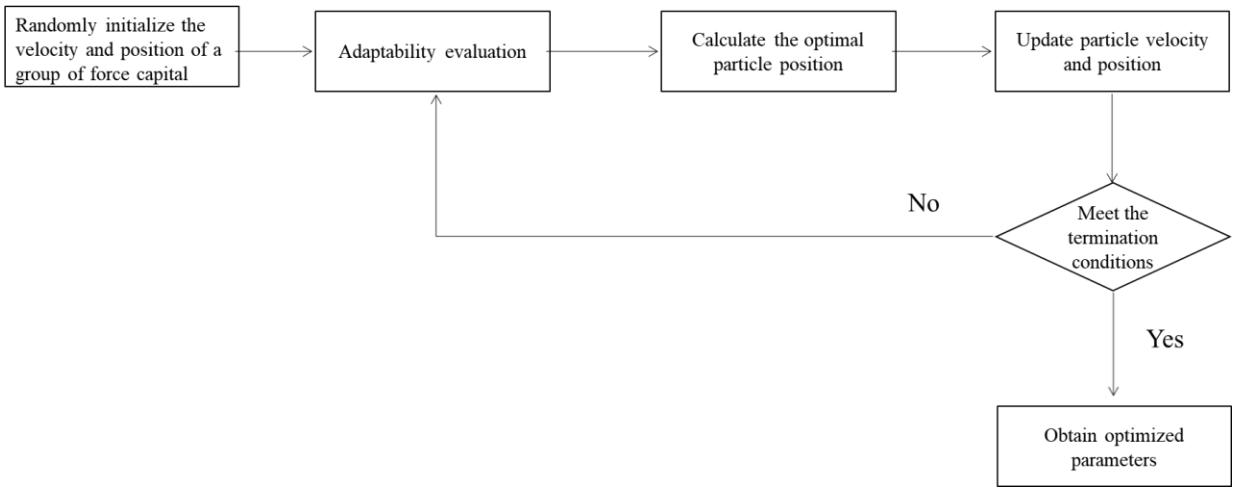

Figure 5. PSO algorithm tuning process

4. Results

(1) For LGBM, no parametric processing is performed during training, all default parameters of the model are used, and the LGBM parameters are listed in the following Table3:

Table 3. LGBM Parameter

Parameter Name	Parameter Value	Parameter Name	Parameter Value
bagging_fraction	1.0	feature_fraction	1.0
bagging_freq	0	num_leaves	31
max_bin	20	learning_rate	0.1
min_sum_hessian_in_leaf	0.001	max depth	5
n_estimators	10		

Figure 6 below shows the distribution comparison of the prediction results of the LGBM model, where green is the MAPE value of the test set and yellow is the MAPE value of the training set.

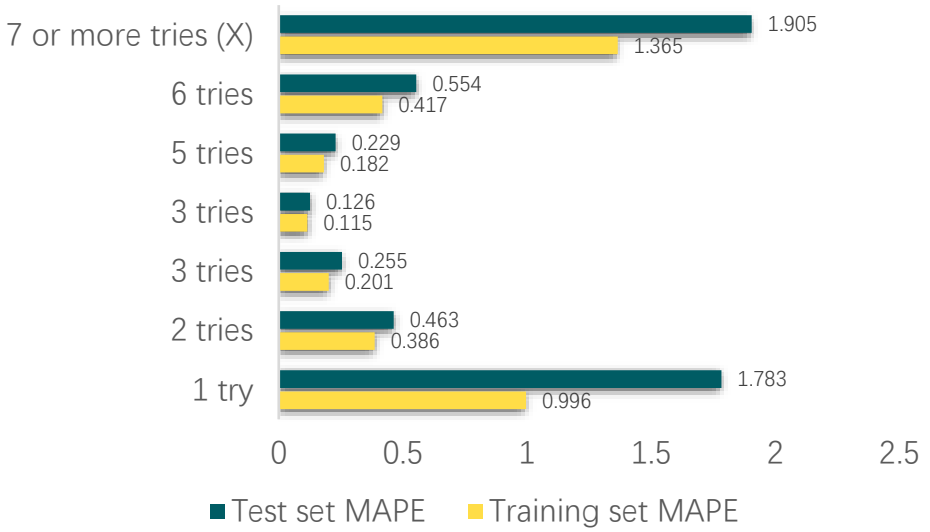

Figure 6. Comparison of LGBM model predictions

(2) The previous LGBM performed well, so we performed hyperparameter tuning

based on PSO, and the results of the analysis were also analyzed using a 5-level cross-test with the following steps.

Step 1: Initialize the parameters of the PSO of the LGBM based on experience.

Step 2: According to Bootstrap algorithm, k samples from the sample data are randomly selected to generate decision trees.

Step 3: Calculate the output of the model.

Step 4: The above classification results are used as the fitness values, and the PSO algorithm is used to continuously iterate, perform parameter optimization, and compare with the historical results to finally output the optimal model parameters.

Step 5: The LGBM is trained according to the obtained model parameters, and finally the hyperparameter with the lowest MSE, the best effect, is derived.

As shown in Figure 7, the histograms are used to visually compare the changes in the results before and after applying PSO optimization, and it can be clearly seen that the results of the model have significantly improved after applying the PSO algorithm [9-10] for optimization.

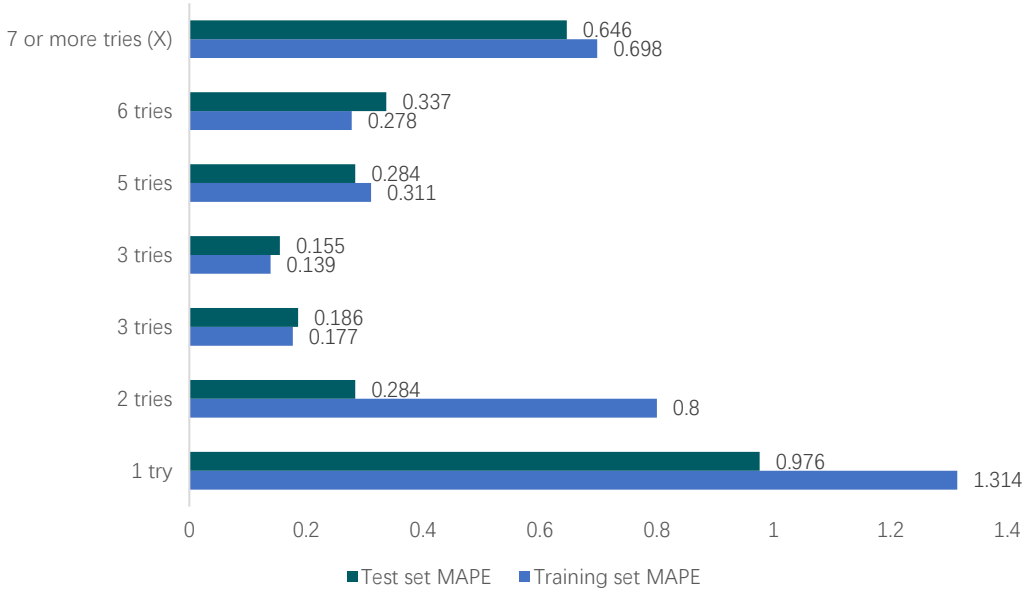

Figure 7. Comparison of PSO-LGBM model predictions

Therefore, the optimal solution is obtained when choosing the optimal parameters of the LGBM model Table 4:

Tabel 4. PSO-LGBM Parameter

Parameter Name	Optimum value	Parameter Name	Optimum value
bagging_fraction	0.87	feature_fraction	0.476
bagging_freq	2	num_leaves	18
max_bin	131	learning_rate	0.14
min_sum_hessian_in_leaf	10	min_data in leaf	26
n_estimators	391		

Figure 8 below shows a summary of the evaluation parameters of these three models, and we can see that the PSO-LGBM model has the best model representation.

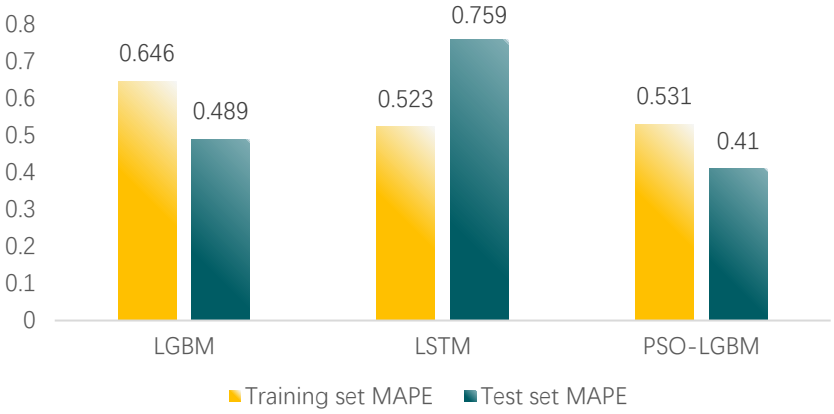

Figure 8. Comparison of three types of model evaluation

Taking the ERRIE word of March 1, 2023 as an example, here first need to encode ERRIE for unique heat, the specific approach refer to 5.1.2 to get its word vector: [0,0,0,0,0,3.0,0,0100,0,0,0,0,0,0,0. 0,1,0,0,0,0,0,0,0,0,0,0], and the predicted results of the correlation percentages of ERRIE words 1, 2, 3, 4, 5, 6, and X were obtained by substituting into the model as shown in Table 5 below:

Tabel 5. PSO-LGBM prediction results

Distribution	Predicted results
1 try	0.230
2 tries	3.513
3 tries	18.326
3 tries	31.263
5 tries	27.500
6 tries	15.751
7 or more tries (X)	2.861

5. Conclusion

In this paper, a prediction model of headcount distribution based on PSO-LightGBM is established, and the parameters of LightGBM are optimized using particle swarm algorithm for improving the accuracy of headcount distribution prediction, and the results of ERRIE word prediction for March 1, 2023 show that the training set and test set of (1, 2, 3, 4, 5, 6, X) predicted by PSO-LightGBM The mean absolute percentage (MAPE) mean values are 0.531% and 0.410%, respectively, and their prediction accuracy and stability are higher than those of LightGBM model and LSTM model. It is known that using PSO to optimize the hyperparameters of LIghtGBM model can improve the prediction effect of this model on the number distribution. There are many factors affecting the number distribution, such as the intensity of media publicity and the common use of words, etc. More influencing factors should be considered in the future to combine with larger-scale data to predict the number distribution.

References

[1] Li, Renyuan,Zhu,Shenglong. Playing Mastermind with Wordle's Feedback [D]. Mathematics, Nanjing University, China, 2022.

[2] Yin XY, Wang XY, Shi A, et al. Feasibility study on predicting cotton yield based on gray theory and time series model [J]. Cotton Science, 2021.

[3] Luo J-P, Zhang Y-Z, Yang S-B. Bus journey time prediction based on PSO-LightGBM [J]. Transportation Engineering, 2023.

[4] Wang, Meixia. Research and application of time series forecasting model based on conjugate gradient method and optimization theory [D]. Yanshan University, 2017.

[5] Wang S.F., Bao C.C. Research on the application of intelligent algorithms in grid load forecasting [J]. Journal of Anhui University of Engineering, 2021.

[6] Liu, ChaoLin. Using Wordle for Learning to Design and Compare Strategies [D]. National Chengchi University, Taiwan, 2022.

[7] Lokshtanov, Daniel. Wordle Is NP-Hard [D]. University of California, Santa Barbara, United States, 2022.

[8] De Silva, Nisansa. Selecting Optimum Seed Words for Wordle using Character Statistics [D]. Moratuwa University, 2022.

[9] Wang Mingfeng. Research on text clustering algorithm based on particle swarm optimization algorithm (PSO) [D]. Guangdong University of Technology, 2020.

[10] Zong Min, Yang Yuqun, Xu Gang. A diversity-driven adaptive particle swarm optimization algorithm [J]. Journal of Nanchang University (Science Edition), 2022.

[11] Jiang Qirong, Wei Y, Gao Xian Song et al. Short-term electric load forecasting by combined LSTM-LightGBM model [J]. China Plant Engineering, 2023.

[12] hang Wei, Yu Chiongbian Shibin et al. multi-feature short-term power load forecasting based on VMD-LSTM-LightGBM [J]. Southern Power Grid Technology, 2023.