ISSN 2616-5775 Vol. 6, Issue 7: 59-63, DOI: 10.25236/AJCIS.2023.060709

Integration Learning-Based User Activity Prediction —the Case of Wordle

Yipeng Miao[1], Junhe Hou[2,*], Yenan Xu[3]

[1]School of Mathematics and Computer Science, Jilin Normal University, Siping, China
[2]High School Attached to Northeast Normal University, Changchun, China
[3]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China
[*]Corresponding author: houjunhe2023@126.com

Abstract: *This paper introduces how to use machine learning methods to solve the user activity prediction problem of Wordle game. By analyzing time series forecasting methods, this paper introduces regression models and integrated learning models, and focuses on the AdaBoost regression algorithm, which is an integrated learning algorithm that can combine multiple weak models to build a strong model. In time series forecasting, the AdaBoost regression algorithm has high prediction accuracy and stability, so it is widely used in this field. In terms of research progress, there have been many applied studies in the field of time series forecasting, but this problem is still challenging. In the future, with the continuous development of machine learning and artificial intelligence algorithms, the time series prediction problem is expected to be studied and applied more deeply and extensively. This paper provides an overview of CART decision tree regression and AdaBoost regression algorithms as well as an introduction to application scenarios, which can provide useful reference suggestions for Wordle game developers.*

Keywords: *Regression Decision Tree, AdaBoost, Integration Learning*

1. Introduction

Wordle is a popular crossword puzzle game in recent years, and its unique rules and simple interface design are very popular among players. However, developers of the game need to understand its user activity behavior to improve game strategy and enhance user experience. This paper aims to explore how to use machine learning methods to solve the user activity prediction problem of Wordle game.

Time series prediction, which uses past time series data to predict future data trends, has important applications in both understanding user activity behavior and predicting users' future behavior. In this field, regression models and integrated learning models have been widely used. In particular, the AdaBoost regression algorithm is an integrated learning algorithm that has been used in several fields for its high accuracy and stability in the field of time series prediction.

In this paper, we will discuss regression models and integrated learning models and the application of AdaBoost regression algorithm in time series prediction and analyze its advantages in solving the user activity prediction problem of Wordle games. In addition, we will discuss research advances and future trends to better understand and apply machine learning algorithms to solve Wordle game user activity prediction problems.

2. Literature Review

Time series prediction has been a research hotspot in the field of data analysis and machine learning, and with the rapid development and popularity of machine learning algorithms and computational techniques, more and more researchers have started to explore how to use machine learning methods to solve time series prediction problems. Literature1 reviews the commonly used machine learning algorithms in the field of time series forecasting, including neural networks, support vector machines, regression models and integrated learning models, and discusses their respective strengths and weaknesses [1].

Among the regression models, ridge regression and Lasso regression, etc. are widely used in time series forecasting. A hybrid time series forecasting algorithm based on Ridge regression and Lasso

ISSN 2616-5775 Vol. 6, Issue 7: 59-63, DOI: 10.25236/AJCIS.2023.060709

regression is proposed in literature 5, which combines the results of Ridge regression and Lasso regression to produce more accurate forecasting results. Besides, the XGBoost algorithm is also a commonly used regression model, which is based on an adaptive tree enhancement algorithm and trained using a gradient boosting method with strong generalization performance and accuracy.

Among the integrated learning models, Boosting algorithms including AdaBoost and Gradient Boosting are proved to have high accuracy and stability in time series prediction. Literature 2 introduces the AdaBoost algorithm and its application in the dichotomous classification problem, while literature 4 explores the application of Gradient Boosting algorithm in stock price prediction. In addition, the Stacking model has also yielded some results in time series forecasting, and it has achieved excellent forecasting results on certain data sets by building multilevel models for forecasting [2-5].

In summary, regression models and integrated learning models have a wide range of applications in the field of time series forecasting, each with its own strengths and weaknesses. When choosing a model, a comprehensive consideration is needed based on task requirements, datasets and practical situations.

3. Data pre-processing

3.1. Overview of the data set

The dataset used in this paper consists of Date, Contest number, Word, Number of reported results, Number in hard mode, 1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries, 7ormore tries (X) The data consists of 359 detailed attempts by Wordle players to solve the puzzle in both modes from 2022/1/7 to 2022/12/31, with days as the interval.

The data were obtained from the official MCM tournament (or the New York Times).

3.2. Move the window mean extraction

In order to solve the problem, for the original data, the basic trend of the data is first observed through the visualization technology. Where the independent variable is Date, and the vertical axis data is Number of reported results, i.e. Figure 1.

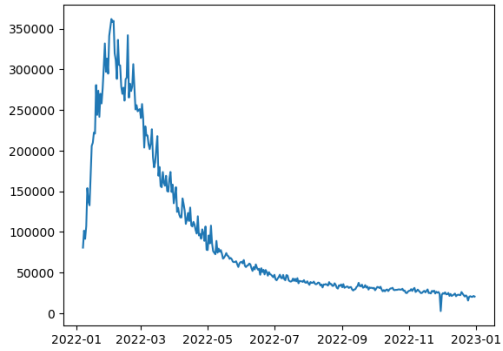

Figure 1: Sequence description diagram

It is not difficult to find that this sequence has no long-term trend or seasonal effect, so the main purpose of predicting this sequence at this time is to eliminate the effects of random fluctuations. At this time, we all need to use the simple moving average [1]. The simple moving average is the weighted weighted arithmetic mean of the past n period as the predicted value of the sequence, which can be expressed as:

$$\hat{x}_{t+1} = \frac{x_t + x_{t-1} + \cdots + x_{t-n+1}}{n} \tag{1}$$

Because $x_t = \mu + \varepsilon_t$, and $\varepsilon_t \sim N(0, \sigma^2)$, so

$$\hat{x}_{t+1} = \frac{x_t + x_{t-1} + \cdots + x_{t-n+1}}{n} = \mu + \frac{\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_{t-n+1}}{n} \tag{2}$$

Easy to launch:

$$E(\hat{x}_{t+1}) = \mu \tag{3}$$

ISSN 2616-5775 Vol. 6, Issue 7: 59-63, DOI: 10.25236/AJCIS.2023.060709

$$Var(\hat{x}_{t+1}) = \frac{\sigma^2}{n} \tag{4}$$

This indicates that the predicted values obtained using simple moving averages are unbiased estimates of the true values of the sequence, and that the larger the moving average period, the smaller the error in prediction.

In conclusion, it can be seen that the design of the simple exponential smoothing method is a better method by considering the influence of the time interval without affecting the unbiased nature of the predicted value. Therefore, in this problem, we take the method of moving window mean extraction, and the control window width is 10.

4. Model Overview

4.1. CART decision tree

CART (Classification and Regression Tree) is a decision tree model that performs classification and regression analysis and is a very popular machine learning algorithm. CART models represent a collection of data in a tree structure, with each leaf node representing a decision or outcome of the model. The splitting of the decision tree is sorted based on the relevance of an attribute or a set of attributes to the target variable. CART is a growing decision tree that builds a decision tree by recursively dividing the data set and the components until the stopping conditions are met.

CART decision tree regression can be used for prediction and modeling of continuous variables, such as house prices, stock prices, etc. For the prediction of continuous variables, the target variable is considered as numerical, and the value in each leaf node is the average of all instances in that leaf node.

The two most important steps in the CART algorithm are splitting and pruning, which together affect the accuracy and generalization ability of the model.

The splitting rule is usually based on the Least Squared Error (MSE) criterion, which is formulated as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

where N is the sample set size,y_i is the true target variable value, and $\hat{y}_i$ is the target variable value predicted by the model. the CART algorithm selects the best split for the data by comparing the MSEs obtained from different attribute splits.

Pruning is used to prevent overfitting of the model, where the number of leaf nodes is limited to avoid overcomplicating the model. Pruning uses cross-validation to determine the most suitable pruning scheme.

Specifically, the CART modeling steps are as follows:

First, starting from the root node, the node is divided into two child nodes by selecting the best attributes based on the MSE.

For each child node, step 1 is repeated until the number of leaf nodes reaches a preset target or the MSE of all instances in the leaf nodes reaches a certain threshold to stop recursion and generate a decision tree.

Pruning is performed next. Starting from the bottom leaf nodes, the function used for pruning is calculated upward layer by layer (since this paper focuses on regression, the specific cross-validation method and algorithm implementation will not be discussed here). When the pruned result is better than the unpruned result, the branch is replaced with a leaf node whose value is the average of all instances in the branch.

The advantages of the CART model are that it is easy to understand and interpret, and that it can handle discrete or continuous input variables. However, it also has some disadvantages, such as sensitivity to data set noise and the tendency to fall into local optimal solutions. Therefore, in practical applications, it would be more effective to use integrated learning methods to reduce model errors.

4.2. AdaBoost

AdaBoost regression is an integrated learning method based on decision trees that constructs a strong

ISSN 2616-5775 Vol. 6, Issue 7: 59-63, DOI: 10.25236/AJCIS.2023.060709

regressor by combining multiple weak regressors. Unlike AdaBoost classification, it aims to build a model that predicts real-valued targets rather than classifying samples into different categories.

The following is a formal definition of AdaBoost regression:

Let the given training data set be $\{(x_i, y_i)\}, i = 1,2,\cdots,N, of\ which, x_i \in \chi \subseteq \mathbb{R}^d, y_i \in \mathbb{R}, i = 1,2,\cdots,N$.The goal of AdaBoost regression is to learn a regressor that can accurately predict the real-valued target.

The algorithmic process of AdaBoost regression:

Initialize the sample weights:

$$D_i(i) = \frac{1}{N}, i = 1,2,\cdots,N \tag{6}$$

For $t = 1,2,\cdots,T$: train a weak regressor $h_t: \chi \to \mathbb{R}$ under the weight distribution D_t.

Calculate the error rate of regressor h_t on the training set:

$$\epsilon_t = \sum_{i=1}^{N} D_t(i)|y_i - h_t(x_i)| \tag{7}$$

Calculate the weights of regressor h_t:

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \tag{8}$$

Update the sample weights:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \exp(-\alpha_t|y_i - h_t(x_i)|) \tag{9}$$

Where Z_t is the normalization factor that makes $D_{t+1}(i)$ a probability distribution.

Output the final regressor:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{10}$$

In the AdaBoost regression algorithm, each weak regressor is based on a decision tree, and each decision tree has only one non-leaf node. Again, the sample weights are updated by changing the weights of those samples that deviate far from the true target value. The final strong regressor is composed of the weighted sum of all weak regressors.

In the AdaBoost regression algorithm, the weighted sum of regressors is a weight calculated based on the error rate of each regressor. In addition, the sample weights are updated by changing the weights of those samples whose predicted values are farther away from the true values.

In summary, AdaBoost regression is an integrated learning method based on decision trees, which constructs a strong regressor by combining multiple weak regressors. It is modeled by adjusting the sample weights according to the errors and increasing the weights of correctly predicted samples. Commonly used weak regressors are decision tree regression, linear regression, etc [6-8].

5. Analysis of Results

The r-square of the regression tree was significantly improved after AdaBoost integration. The final trained model was scored to obtain its r_square value of 0.9768312796594524. Obviously this is an excellent score, so this model was used to predict the number of visitors on March 1, 2023, and the predicted number of visitors for that date was 42,355.

In summary, a good time series prediction model can be fitted using the above algorithm, and the predicted value of the number of visitors to the game on March 1, 2023, is given as 42,355.

6. Conclusion

Through the research in this paper, we can conclude that integrated learning methods can help solve the user activity prediction problem of Wordle games. Regression models and integrated learning models can improve the accuracy and stability of prediction. In particular, the AdaBoost regression algorithm, an integrated learning algorithm with high prediction accuracy and stability, is an effective method to solve the time series prediction problem.

ISSN 2616-5775 Vol. 6, Issue 7: 59-63, DOI: 10.25236/AJCIS.2023.060709

In future research, we suggest that developers developing game applications such as Wordle games should incorporate machine learning algorithms to try to predict user activity behaviors more accurately and effectively, so as to improve game strategies and increase user retention time. Meanwhile, in the field of time series prediction, we also hope that researchers can improve the prediction accuracy and stability in more in-depth research of integrated learning algorithms, so that these algorithms can play a greater role in a wider range of application scenarios.

References

[1] Yi Danhui, Wang Yan. Applied a time-series analysis [M]. Beijing: China Renmin University Press, 2019:173.

[2] Patel P, Hill A, Nayee U, et al. Observing Prefrontal Cortex Activity During Rule-Based and Information-Integration Category Learning[J]. Journal of Vision, 2016, 16(12):261. DOI:10.1167/16.12.261.

[3] Wu Gang, Xu Guoyu, Liu Guangtao, et al. Application of decision tree model and logistic regression model in the prognosis analysis of cerebral hemorrhage [J]. Journal of PLA Medicine 2015, 40 (12): 1003- -1006.

[4] Xiao Hui, Hao Yuantao, Xu Xiao, etc. Diabetes risk factors study based on the random forest algorithm and the Logistic regression model [J]. Digital Medicine in China, 2018, 13 (1): 33-35.

[5] Cui Dongwen, Jin Bo. Comprehensive evaluation of water ecological civilization based on the random forest regression algorithm [J]. Progress of Science and Technology of Water Conservancy and Hydropower, 2014, 34 (05): 56-60 + 79.

[6] Li Hua, Rong Jieyu. Modeling analysis of NIR spectroscopy based on the AdaBoost method [J]. Journal of Jilin Normal University (Natural Science Edition), 2022, 43 (03): 83-89.

[7] Cao Jie, Deng Lujuan. Python Data mining and application: Micro-course version [M]. Beijing: Tsinghua University Press, 2021:119.

[8] He Xiaoqun, Liu Wenqing. Applied the regression analysis [M]. Beijing: China Renmin University Press, 2019:17.