ISSN 2616-5775 Vol. 6, Issue 6: 140-144, DOI: 10.25236/AJCIS.2023.060622

English Word Difficulty Classifier Based on Random Forest Model

Miao Peng[1], **Yujie Wu**[1,*], **Yannan Qiu**[2]

[1]School of Finance, Guangdong University of Foreign Studies, Guangzhou, China, 510006
[2]School of Accounting, Guangdong University of Foreign Studies, Guangzhou, China, 510006
**Corresponding author: 13421006293@163.com*

Abstract: *Recently, Wordle has become popular worldwide as a daily puzzle game launched by the New York Times. Players try to solve the puzzle by guessing a five-letter word in six tries or less. According to Wordle's statistical data, this paper first uses the K-means algorithm to cluster the difficulty of solution words to quantify the difficulty of English words and analyzes the accuracy and scientificity of the clustering results. Then, the paper uses the Random Forest model to classify the difficulty of words into three categories: 'easy', 'normal' and 'hard'. The results show that the classification accuracy on the training set and the test set reaches 0.972 and 0.978 respectively.*

Keywords: *English Word Difficulty Classifier, Random Forest model, K-means algorithm*

1. Introduction

Reading English texts is one of the most important ways to improve English proficiency, and with the rapid development of Internet technology and educational informatization, there are more and more English reading materials available online. For English learners, it is important to have a way to quickly assess the difficulty of reading materials to match the reading ability of readers [1]. The most important indicator to evaluate the difficulty of reading materials is the difficulty of English words in the materials. Therefore, this paper takes the classification of the difficulty of English vocabulary as the starting point to reflect the difficulty of English reading materials. Traditional level assessment methods rely on the subjective judgment of experts and the use of linear relationships to quantify text difficulty, while artificial intelligence algorithms such as machine learning are rarely used to classify the difficulty of words [2-3]. This paper innovates research methods. Firstly, the K-means algorithm is used to cluster English words to quantify their difficulty, and then the Random Forest ensemble learning algorithm is used to classify the difficulty of words.

2. Model Construction

2.1 Data sources and Variable selection

The data in this paper are derived from Wordle, an English word-guessing game provided daily by The New York Times, which is now available in more than 60 languages. In this paper, 12 variables are extracted from the statistics of Wordle and the attributes of words, as shown in Table 1 [4].

Table 1. Notion

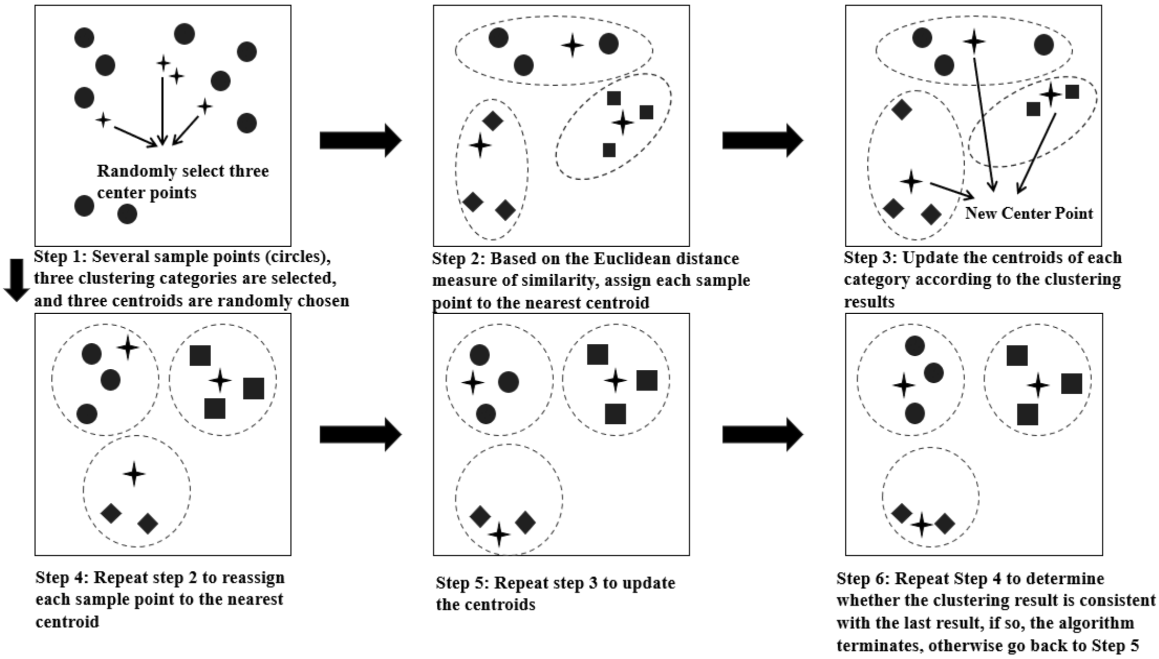
Symbol	Definition
$f_{frequency}$	The frequency of word
f_{repeat}	The proportion of repeated letter
f_{letter}	The proportion of high-frequency letter
$f_{initial}$	The proportion of high- frequency initial
f_{affix}	The proportion of high- frequency affix
try_1	The percentage of players solving the puzzle in one guess.
try_2	The percentage of players solving the puzzle in two guesses.
try_3	The percentage of players solving the puzzle in three guesses.
try_4	The percentage of players solving the puzzle in four guesses.
try_5	The percentage of players solving the puzzle in five guesses
try_6	The percentage of players solving the puzzle in six guesses.
try_x	The percentage of players that could not solve the puzzle in six or fewer tries.

2.2 K-means Algorithm

The core idea of the K-means algorithm is to calculate the distance from each sample point to each center point and assign the sample point to the class represented by the nearest center point. After one iteration is completed, the center point of each class is updated according to the clustering results, and then the previous operation is repeated for another iteration until there is no difference between the two classification results. The formula for calculating the distance between samples is (Using Euclidean distance) :

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

The principle of the K-means algorithm is shown in Figure 1:



Step 1: Several sample points (circles), three clustering categories are selected, and three centroids are randomly chosen

Step 2: Based on the Euclidean distance measure of similarity, assign each sample point to the nearest centroid

Step 3: Update the centroids of each category according to the clustering results

Step 4: Repeat step 2 to reassign each sample point to the nearest centroid

Step 5: Repeat step 3 to update the centroids

Step 6: Repeat Step 4 to determine whether the clustering result is consistent with the last result, if so, the algorithm terminates, otherwise go back to Step 5

Figure 1: Principle of the K-means algorithm

Since the distribution of word guesses best reflects the difficulty of words, this paper takes the seven variables $try_1 \sim try_x$ as the clustering criteria, and clusters the difficulty of all words into three categories, namely ‘easy’, ‘normal’ and ‘hard’.

2.3 Random Forest Model

The Random Forest model uses an ensemble learning method, which is essentially an improvement of the decision tree algorithm by combining multiple decision trees, with the building of each tree relying on independently drawn samples. As shown in Figure 2, the Random Forest model randomly samples n different sample datasets in the original dataset, then builds in different decision tree models based on these datasets, and finally obtains the final results based on the voting of these decision tree models [5].

In order to ensure the generalization ability of the model, the Random Forest model tends to follow the two basic principles of ‘Sample randomness’ and ‘Feature randomness’ when building each tree. ‘Sample randomness’ means that when drawing samples for the training set, the bootstrap method is used to randomly draw N training samples k times, and the k data sets are independently distributed from each other. ‘Feature randomness’ is to select m features from M features (m<<M), and M is the feature dimension of the sample.

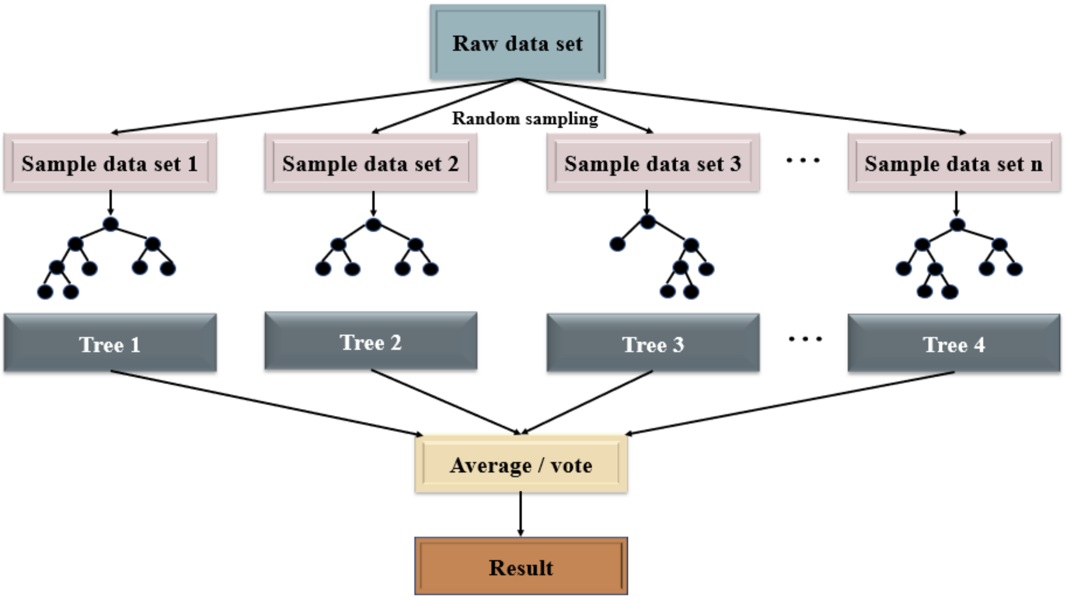


Figure 2. Principle of the Random Forest model

The process of classification using the Random Forest model is as follows.

1) Selecting feature vectors

This paper selects a total of 12 variables (Table 1) of word attributes and word guesses as the classification criteria of words, which constitute the feature vectors of the model to classify the difficulty of words classified by the K-means algorithm.

2) Sample Balance

Since there is an imbalance in the categories of words, the paper uses the smote method to equalize the sample categories.

3) Dividing the training set and test set

To ensure the accuracy of the training model, the paper takes 319 sets of the 456 data after sample equalization as the training set and 137 sets as the test set to modify the model and improve its accuracy.

4) Training and testing the model

After optimizing the parameters of the model, the paper selects 50 base decision trees and sets the minimum number of samples required for node splitting to 5 to train the training set, and test the model on the test set.

3. Results

3.1 Clustering Results

The clustering results of the K-means algorithm are shown in Figure 3, 'class1' means 'easy', 'class2' means 'normal', and 'class3' means 'hard'. This paper finds that there are more 'normal' and 'easy' words in the sample and fewer 'hard' words.

To analyze the accuracy of the clustering results, the paper analyzes the difficulty of each category. As shown in Figure 4, the figure shows that as the difficulty of solution words increases, the total number of people who can solve the problem fewer times decreases. Therefore, the clustering results are scientific and accurately reflect the difficulty of words in different categories.

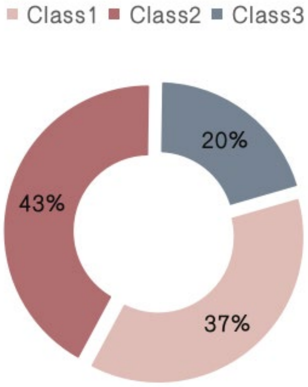


Figure 3: Clustering result

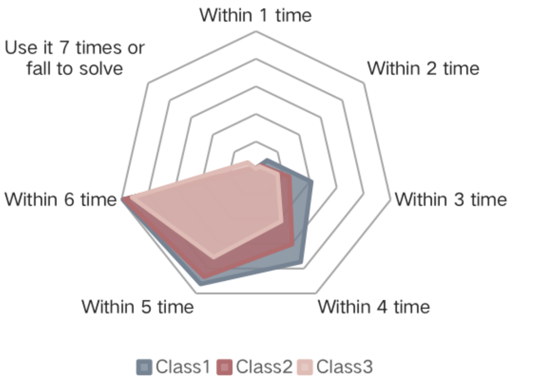


Figure 4: Analysis of the difficulty of solution words in different categories

3.2 Classification Results

The classification results of the Random Forest model are shown in Figure 5 and Figure 6, where Figure 5 denotes the confusion matrix of the training set and Figure 6 denotes the confusion matrix of the test set, the horizontal coordinates denote the predicted categories and the vertical coordinates denote the true clustering categories, so the squares on the main diagonal are the number of correct predictions. By calculating these values, the paper finds that the model achieves a prediction accuracy of over 97% in both the test and training sets (Table 2).

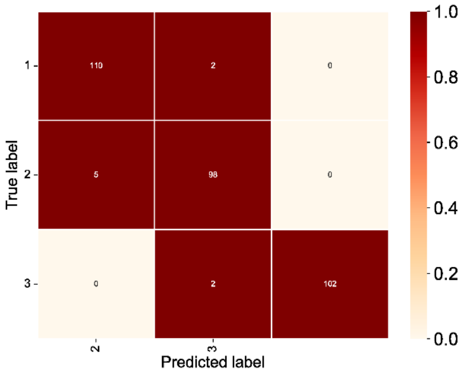


Figure 5. The training set confusion matrix

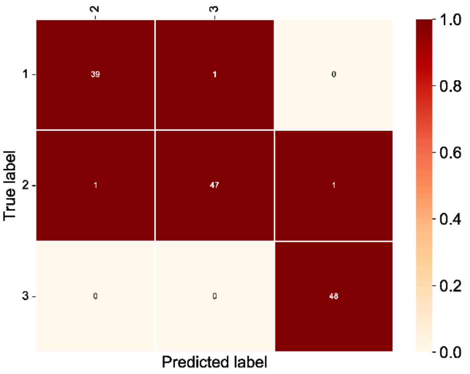


Figure 6. The test set confusion matrix

Table 2. Model Classification Accuracy

	Training set	Test set
Accuracy rate	0.9717868338557993	0.9781021897810219

4. Conclusion

The English word difficulty classifier can quickly identify the difficulty of English words and provide readers with a basis for judging the difficulty of reading materials, which has some practical application value. The experimental results show that the classification accuracy of the model reaches over 97% on both the training and test sets, and the model has good generalization ability and prediction accuracy. Subsequent studies can consider applying this model to solve the classification problems of other language words.

References

[1] Yubo Fu. Research on English Text Difficulty Evaluation Based on Decision Tree [D]. Central China Normal University, 2018: 1-6.

[2] Kang An, Yongbo Zhang, Ze Huang. English text difficulty estimation model based on multiple linear regression [J]. Modern Information Technology, 2022, 11(6): 30-33.

[3] Curto P, Mamede N, Baptista J. Automatic text difficulty classifier [C]. Proceedings of the 7th International Conference on Computer Supported Education. 2015: 36-44.

[4] Hashimoto B J, Egbert J. More than frequency? Exploring predictors of word difficulty for second language learners[J]. Language Learning, 2019, 69(4): 839-872.

[5] Parmar A, Katariya R, Patel V. A review on random forest: An ensemble classifier [C]. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Springer International Publishing, 2019: 758-763.