Psychological Analysis and Xgboost Algorithm Based Modeling of Students' Cognitive Rules of English Vocabulary

Chen Hongxia[1, 2], Jiang Lili[1, 2*]

1.Cas Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China

2.Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049 China

*Corresponding Author

Keywords: Psychological analysis, Cognitive rule, English education, Xgboost, Loss function

Abstract: Different groups of students have different levels of mastery of different English vocabulary. Based on the real teaching scenario of an Internet education company, we collected a large amount of student data for a certain English vocabulary learning course in the fifth grade of elementary school. After data preprocessing and correlation analysis remove weakly related attributes, a predictive model of the difficulty of mastering English vocabulary is constructed based on the Xgboost algorithm. By real data test, the predictive performance of the model is analyzed and discussed. Finally, the loss/profit ratio of the model based on the profit function and loss function are given and discussed.

1. Introduction

In elementary school, mastering vocabulary is the most direct learning method to improve English ability, and it is also the most useful learning method. First of all, when students learn English, they first face vocabulary, then sentences composed of vocabulary, and then articles composed of sentences. The famous scholar Diller's (1978) research shows [1] that if we know 25 vocabulary and randomly select a page of a reading, we will know 23% of the vocabulary on this page; if we know 135 vocabulary, the percentage will reach 50%; if we know 2500 words, the percentage will reach 78%. Obviously, understanding vocabulary is the basis for understanding sentences and understanding articles, and mastering the core common vocabulary plays an important role in our understanding of articles.

Second, vocabulary learning helps to increase the interest of younger students in learning English. For example, the sentence “I like candy”, even if the student does not have any grammatical knowledge, as long as he understands the meaning of I, like and candy, he can understand the meaning of this sentence. Conversely, if the student is proficient in grammar but does not know the vocabulary candy, then he still cannot understand the meaning of this sentence. In the actual teaching process, we found that many students have weak oral communication skills. The essence is also because the students' English vocabulary is not sufficient. If they dare not communicate because of insufficient vocabulary, students may eventually lose interest in English learning, and even become afraid of difficulties [2].

In the process of vocabulary teaching, we found that there are differences in students' cognition of different vocabulary. That is to say, for some vocabulary, the students have a good grasp; for other vocabulary, the students are not good enough. How to describe and measure the differences in vocabulary cognition laws, many literatures have given research results from the perspective of cognitive psychology. Literature [3] pointed out that usually students are not good enough for longer vocabulary, but better for shorter vocabulary. The literature [4] pointed out that students have a good grasp of vocabulary that can be seen and heard at a high frequency in life, but are not good enough for vocabulary that does not often appear in life. Literature [5] pointed out that younger students have a better grasp of quantifiable nouns than emotional vocabulary.

The above-mentioned research results do have some value, but there are still some problems: First, most of the literature focuses on the general population of students, rather than specifically for

DOI: 10.25236/iccem.2021.066

students in a certain grade. And they are not enough to study group differences among students of the same age. Second, with the rapid development of the times, especially with the popularization of Internet education models, the psychological characteristics of students have undergone major changes. The above research conclusions may no longer be valid or need to be adjusted appropriately. Third, most of the literature is based on the perspectives of pedagogy and psychology, and there are few methods using big data, statistics and machine learning. Therefore, most of the qualitative research results are given, while the quantitative research results are relatively few. At the same time, the above-mentioned research results pay more attention to summarizing the facts that have occurred in the past, but lack of prediction research on future uncertain events.

We will focus on the topic of English vocabulary learning for fifth-grade students. First, give a measurement method of the difficulty of vocabulary recognition, model the prediction of the difficulty of vocabulary recognition as a binary classification problem, and give a prediction model of the gain function and loss function. Second, collect student behavior data based on the real Internet teaching environment, and extract feature fields from the original data based on the laws of educational psychology and Pearson correlation analysis. Third, use machine learning algorithms to construct the Xgboost model and logistic regression model of English vocabulary cognition rules. Finally, the model is applied to actual data to test the validity of the model.

2. Problem Analysis

2.1 The Measurement of English Vocabulary Mastery

Generally, people have the following consensus on the quantitative description of the level of English vocabulary mastery Y:

1) Y should usually be modeled as a continuous variable, $Y \in [0,1)$.

2) The angle of measurement is different, and there are different measurement methods for Y.

3) It is very difficult to try to find a universal and accurate formula to quantitatively describe Y.

However, starting from the actual needs of English teaching, we do not need to make an accurate measurement of Y. Usually, vocabulary-related test questions can be used to examine students' mastery of a certain vocabulary. For the sake of simplicity, the vocabulary is usually divided into two levels according to the students’ mastery: easier to master (it takes less learning time to master) and harder to master (it takes more learning time to master). Therefore, Y can be modeled as a two-dimensional variable, that is, Y=1 means that the vocabulary is difficult to master (referred to as “*hard*”), and Y=0 means that the vocabulary is easier to master (referred to as “*easy*”).

2.2 A Binary Classification Model for Predicting the Difficulty of English Vocabulary Recognition

Suppose that all the vocabulary that the student has learned form the set A, and all the vocabulary that the student will learn in the recent period form the set C. Assuming that the set A contains N words in total, it can be divided into two categories: M (*hard*) and N-M (*easy*). The predictive modeling of vocabulary difficulty level is to mine the characteristics of the two types of vocabulary in the set A, and construct a two-category model F. This model will be used to predict which words in set C belong to the *hard* category and which belong to the *easy* category, so as to provide a basis for teachers to carry out differentiated teaching in future vocabulary teaching.

Suppose that the two classification model F predicts that K words belong to the *hard* category. The perfect prediction result is that these K words belong to the *hard* category. However, this is only a theoretical possibility. The actual situation will have the following four prediction results, as shown in Table 1.

Table 1 a Confusion Matrix for Customer Churn Forecasts

	The model predicts as *hard*	The model predicts as *easy*	Sum
Actually belong to *hard*	*a*	*b*	*M*
Actually belong to *easy*	*c*	*d*	*N-M*
Sum	*K*	*N-K*	

In Table 1, *a* is the *hard* vocabulary which is correctly predicted by the model (abbreviated as hit). *b* is the *hard* vocabulary which is incorrectly predicted as *easy* (abbreviated as missed). *c* is the *easy* vocabulary which is incorrectly predicted as *hard* (abbreviated as misjudgment). And *d* is the *easy* vocabulary which is correctly predicted.

After the model is built, it will be used in differentiated teaching. For the easy (*b*+*d*) vocabulary, the teacher will spend less teaching resources (including time, energy, etc.) in teaching, set as q. For the (*a*+*c*) vocabulary predicted to be *hard*, the teacher will spend more teaching resources in teaching and set it as $\alpha q(\alpha > 1)$. Suppose the income of students mastering one word is *p*. If the model is a perfect prediction, the total income Q can be expressed as

$$Q=N\times p \qquad (1)$$

Obviously, the omission and misjudgment of the model will have a negative impact on the teaching process, which is embodied in:

1) The misjudged c words, each misjudged vocabulary originally only needs resources q, which wastes resources (α-1)q.

2) The missed d vocabularies cost insufficient teaching resources, resulting in students not being able to master them well, and each missed vocabulary brings loss p.

The overall loss S of the prediction model F can be expressed as

$$S=c\times(\alpha-1)\times q + d\times p \qquad (2)$$

The loss/gain ratio β can be expressed as

$$\beta=\frac{S}{Q}=\frac{c\times(\alpha-1)\times q+d\times p}{N\times p}=\frac{c\times(\alpha-1)\times\frac{q}{p}+d}{N} \qquad (3)$$

Optimizing the ratio of misjudgment and miss-judgment of the prediction model F can make S and β obtain the minimum value.

3. Data Analysis and Modeling Preparation

3.1 Data Collection

The data comes from the fifth grade vocabulary learning module of a Chinese Internet English education company from January to February 2021. 10,000 students were randomly selected from all student users, of which 9,000 were classified as training set A, and 1,000 were classified as test set B. The vocabulary learning module has a total of 1000 vocabularies, of which 233 are marked as *hard* (Y=1) and 767 are marked as *easy* (Y=0).

3.2 Data Preprocessing

A total of 180 static attributes and dynamic attributes generated during the learning process were collected. Many attributes are difficult to directly use for modeling, so data preprocessing is required first.

1) Duplicate data audit: It is to conduct data audit on the overall key indicators, for example: check whether each user has a unique record, check the accuracy of the data, and delete duplicate data.

2) Singular value data: Singular values are reflected in the data in the form of outliers, that is, they deviate greatly from most normal values, and are identified and deleted by the histogram or scatter diagram of the variables.

3) Severe missing features: For missing data, evaluate the difficulty and value of complementing, and identify and delete the severely missing features.

4) Re-encoding of strings: Such as “school”, “gender”, “parents' education”, etc., re-encode the strings into different variables in advance.

3.3 Pearson's Correlation Analysis and Feature Elimination

There are a total of 180 data set attributes. If all are used for modeling, on the one hand, the modeling time is long, on the other hand, some interference fields will reduce the accuracy of the model. Therefore, it is necessary to perform a correlation analysis on the data attributes first, and eliminate the weaker correlation attributes, and only retain the highly correlated attributes for subsequent modeling. Table 2 shows the strong correlation attributes retained after Pearson correlation analysis.

Table 2 Attributes retained after Pearson correlation analysis

Attributes I / Student		Attributes II / Vocabulary	
Student number	Father occupation	Vocabulary length	Noun ?
Gender	Mother occupation	Closeness to life,	Adverb ?
Age	School name	Frequency in learning materials	Preposition ?
Degree of diligence	Last teaching module test total score	Verb ?	Number of syllables
Years of learning English	Last vocabulary test score	Adjective ?	Pure vowel ?
…	…	…	…

4. Algorithm Modeling and Evaluation

4.1 Principle of Xgboost Algorithm

Boosting is a commonly used statistical learning method. In the training process, by changing the weight of the training sample, learning multiple classifiers, and finally obtaining the optimal classifier. After each round of training, reduce the weight of the correctly classified training sample and increase the weight of the incorrectly classified sample. After multiple trainings, some of the incorrectly classified training samples will get more attention, and the correct training sample weight tends to 0. Multiple simple classifiers are obtained, and a final model is obtained by combining these classifiers. Xgboost algorithm [6] is based on traditional Boosting, using CPU multi-threading, introducing regularization items, adding pruning, and controlling the complexity of the model.

4.2 Training Model

The prediction problem of cognition difficulty is a typical binary classification problem, which belongs to the application category of XGBOOST algorithm. The training data set A we constructed has a total of about 80,000 users, a vocabulary of about 1,000, and the difficulty of vocabulary recognition Y as a dependent variable. The ratio of positive and negative samples is about 1:3, which is obtained after data preprocessing and correlation analysis are eliminated. The 30 variables of are used as independent variables X. Through repeated tuning, finally output the vocabulary recognition difficulty degree prediction model F.

4.3 Performance Comparison between Xgboost Model and Gbdt Model

Figure 1 shows the ROC curve (green) of the prediction model F constructed by the Xgboost algorithm on the test set B. In order to make a comparison of model performance, the ROC curve (blue) of the prediction model constructed using the GBDT algorithm is also given. Comparing the ROC curves of the two models, it can be seen that the prediction performance of the Xgboost model is significantly better than that of the GBDT model.

4.4 Loss Function Evaluation

The profit function is given by (1), the loss function is given by (2), and its parameter settings are given in Table 3.

Table 3 Parameters of Profit Function and Loss Function

Parameter	α	p	q
Value	4	250	100

The minimum value of the loss function S is solved in the Xgboost model and the GBDT model, and the optimal decision threshold is shown in point B and point A in Fig. 1. Point A (93.2% hit rate, 51.1% false positive rate), point B (85.3% hit rate, 61.4% false positive rate). The loss/profit ratio β is given by (3), and the settlement result is shown in Fig. 2. The loss/gain ratio of the Xgboost model is less than that of the GBDT model, which is only 6.54%, and the effect is relatively satisfactory.

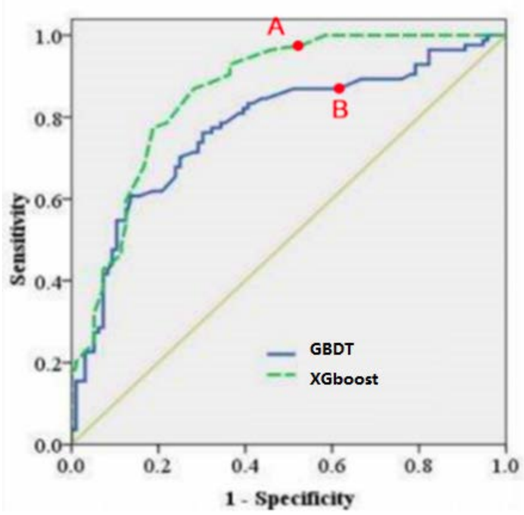

Fig.2 Mparison of Roc Curves

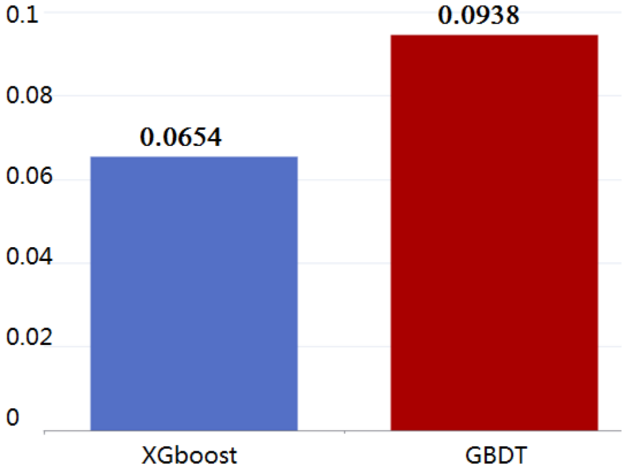

Fig. 3 Comparision of Loss/Profit Ratio

5. Summary

Test results based on actual data show that the Xgboost predictive model can effectively predict the difficulty of students' cognition of the vocabulary they are about to learn. The overall loss/benefit ratio caused by the prediction error is only 6.54%, which can provide a more reliable basis for teachers to carry out differentiated vocabulary teaching. The Xgboost prediction model points out that there are certain differences in the cognitive laws of student groups of different genders, different learning bases, and different family backgrounds. Taking all student groups together, on average, the top three factors that determine whether it is difficult to master vocabulary are: vocabulary length, closeness of vocabulary to life, and frequency of vocabulary in learning materials.

References

[1] Karl Conrad Diller. (1978). The language teaching controversy. Newbury House Publishers.

[2] Chen Xiaotang. (2006). English Vocabulary Learning Strategies. Shandong: Shandong University Press.

[3] Gu Yongqi, Hu Guangwei. (2003). Empirical Study on English Learning Strategies. Xi'an: Shaanxi normal University Press.

[4] Liu Shaolong. (2003). On the Acquisition and Development of Second Language Vocabulary. Foreign language Teaching, 2,102-105.

[5] Yao Meilin, Wu Jianmin, Pang Hui. (2000). English Vocabulary Memory Strategies of Junior High School Students. Psychological Science, 5, 26-30.

[6] T. Chen, S. Singh, B. Taskar, and C. Guestrin. Efficient second-order gradient boosting for conditional random fields. In Proceeding of 18th Artificial Intelligence and

[7] Statistics Conference (AISTATS’15), volume 1, 2015. Shandong normal University. 3, 11-14.