# Analysis and Countermeasures of College Students' English Online Learning Behavior Data Based on Data Mining Technology

Mei Zhao

20811276@qq.com

Shandong Institute of Commerce and Technology, Jinan, Shandong, 250014, China

**Abstract.** With the deepening of educational informatization, online learning methods are rapidly emerging and gradually applied in the field of higher education. Although online learning has practical advantages such as convenience, interactivity and timeliness, it also has some shortcomings such as separation of teachers and students and lagging teaching supervision and management. In this regard, based on the current teaching situation of English majors in colleges and universities, this paper will build an online learning behavior index system and analysis framework, summarize the characteristics of students' behavior, and put forward suggestions for improvement on the subsequent curriculum teaching plan. The analysis framework takes data mining technology as the core, and focuses on using Pearson coefficient to calculate the correlation between learning behavior and English learning performance. At the same time, it combines Logistic regression model to classify and predict students' learning performance, which is convenient for teachers to analyze and deal with students' learning behavior and improve the effectiveness of online learning. Practice has proved that students' operational behavior, cooperative behavior and problem-solving behavior are significantly correlated with their English performance, and learning intervals, communication times and the completion of learning processes are the main learning behavior characteristics that affect their English performance. Teachers need to take corresponding measures to intervene and provide students with necessary guidance and help.

**Keywords:** data mining; English; online learning; learning behavior data; analysis model

## 1 Introduction

With the in-depth implementation of the education informatization 2.0 action plan, the new educational technology represented by online learning is having a far-reaching impact on the traditional teaching mode and has become the key path to promote the transformation and upgrading of the education field [1]. With the blessing of network information technology, the teaching environment and teaching process are reconstructed, teaching methods and teaching forms are integrated and innovated, and educational resources are fully shared and optimized, which greatly promotes students' independent learning and personalized development. However, at the same time, the online learning mode is also facing some challenges, such as the change of the roles of students and teachers, the lack of students' self-control ability and the lag of teaching supervision and management, which leads to the low learning status of students and the unsatisfactory learning effect [2]. In view of this, in order to deeply explore

ITEI 2023, November 24-26, Zhengzhou, People's Republic of China

DOI 10.4108/eai.24-11-2023.2343635

the characteristics of students' online learning behavior under the current English major courses in colleges and universities, and strengthen the supervision and intervention of students' bad learning behavior, this paper will analyze the research status of data mining of online learning behavior of college students at home and abroad through literature review. Based on the analysis method of online learning behavior characteristics proposed in reference [3], combined with the analysis results of the application of educational data mining in online learning by Chwen Jen Chen[4] and the experimental conclusions of M Munshi [5] to enhance the online learning effect through data mining algorithms and models, this paper puts forward a set of online learning behavior index system and analysis framework for college students, and summarizes many characteristics of different students' online learning behaviors. The overall scheme is driven by network data, involving data collection, data preprocessing, data mining, result analysis and other links and steps. It focuses on the correlation between learning behavior and learning performance and the prediction of learning performance to help teachers complete the supervision and management of English online learning process.

## 2 Data mining and model building

### 2.1 Data mining

Data mining is an interactive process between users or knowledge bases. It automatically analyzes a large number of data with the help of computer, machine learning and other technologies, extracts useful information, and reveals its inherent laws and knowledge, so as to better explain phenomena and solve problems. The overall process of data mining depends on the characteristics of data and the specific needs of business projects, mainly involving defining the purpose of research, research data collection, research data preprocessing, determining data mining algorithms, and interpreting and evaluating according to mining results. The specific process structure is shown in Figure 1 [6].

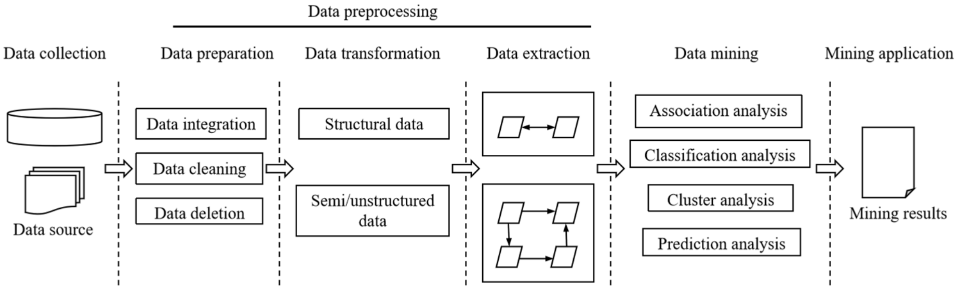

**Fig. 1.** Data mining process

In this study, college students' online English learning behavior data belongs to the category of educational big data. The application of data mining technology can help teachers distinguish the deep-seated valuable information behind it, and it is one of the important tools to solve the educational problems in the era of network big data. After analyzing and comparing the relevant literature, we choose Pearson coefficient as an important basis for the correlation analysis between learning behaviors, and combine Logistic regression model to classify and predict students' learning performance.

## 2.2 Pearson coefficient

Correlation analysis is an important statistical analysis method, which can be used to measure the strength and direction of linear relationship between two or more variables. Through correlation analysis, we can find the degree of correlation between variables, so as to better understand the relationship between objective things. Pearson correlation coefficient is one of the most commonly used statistical indicators in correlation analysis. It can reflect the degree of correlation between two variables by calculating the linear correlation coefficient between them.

Pearson correlation coefficient is usually expressed as *r*, and its value range is between -1 and 1. When *r* =1, it represents two variables perfect positive correlation; When *r* =-1, it means that the two variables are completely negatively correlated; When *r* =0, there is no linear relationship between two variables [7]. In the definition of the algorithm, $X=\{x_1, x_2, x_3, \ldots, x_n\}$, $Y=\{y_1, y_2, y_3, \ldots, y_n\}$ respectively represent two samples, where *X* is the average value of *X* sample and *Y* is the average value of *Y* sample. Pearson correlation coefficient is calculated as Formula 1.

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

Generally speaking, when describing the degree of correlation between two variables, the *r* value can be subdivided into multiple intervals to further clarify the division of the degree of correlation. Table 1 shows the common rules for judging the relationship between correlation coefficient and correlation.

**Table 1.** Correlation coefficient and correlation judgment rules

| Correlation judgment result | Negative correlation coefficient | Positive correlation coefficient |
|---|---|---|
| Irrelvance | -0.2 ~ 0.0 | 0.0 ~ 0.2 |
| Low correlation | -0.5 ~ -0.2 | 0.2 ~ 0.5 |
| Moderate correlation | -0.8 ~ -0.5 | 0.5 ~ 0.8 |
| High correlation | -1.0 ~ -0.8 | 0.8 ~ 1.0 |

## 2.3 Logistic regression algorithm

Logistic regression algorithm is a generalized linear regression analysis model, which is common in the field of data mining and belongs to supervised learning in machine learning. Logistic regression algorithm can solve binary classification problem, regression problem and feature selection problem, and can be fused with other classifiers or models to improve model performance and calculation accuracy [8]. In the definition of the algorithm, the independent variable $X=\{x_1, x_2, x_3, \ldots, x_m\}$, the dependent variable *Y*, and the value of *Y* is 0 or 1, which makes it obey binomial distribution and constitutes a standard binary classification problem. Logistic regression model can be expressed as Formula 2:

$$W\left(Y=1\right)=\frac{\exp^{\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_mx_m\right)}}{1+\exp^{\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_mx_m\right)}} \tag{2}$$

$\beta_0$ is the intercept, $\beta_i$ is the partial regression coefficient corresponding to $x_i$, and Formula 3 is obtained after taking the logarithm on both sides of Formula 2:

$$\ln\frac{W\left(Y=1\right)}{1-W\left(Y=1\right)}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_mx_m \tag{3}$$

Make $z = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+...+\beta_mx_m$, that is:

$$\ln\left(\frac{p}{1-p}\right)=z \qquad p=\frac{1}{1+\exp^{(-z)}}, -\infty < z < +\infty \tag{4}$$

Formula 4 is called logarithmic unit, and p /(1-p) is called opportunity ratio. According to the mathematical model, the Logistic regression algorithm is described as pseudocode as follows:

```
Algorithm: Logistic regression algorithm
Input: sample set$D$={($x_1$,$y_1$), ($x_2$,$y_2$), ($x_3$,$y_3$), …, ($x_n$,$y_n$)}, $x$ is the independent variable and $y$ is the dependent variable. $y \in \{-1, +1\}$ Output: parameters $w$ and $b$
    1. Initialization parameters $w$ = [0, 0, ..., 0], $b$ = 0
    2. Define the learning rates $\eta$
    3. Repeat the following steps until the stop condition is met:
      A. For each sample in the training set ($x_i$, $y_i$):
        i. Calculate the predicted value $z = w \cdot x_i + b$
        ii. Calculating the activation function sigmoid($z$) = 1 / (1 + $\exp^{(-z)}$)
        iii. Update parameters $w = w + \eta \cdot (y_i$ - sigmoid($z$))$\cdot x_i$
            Update the intercept parameter $b = b + \eta \cdot (y_i$ - sigmoid($z$))
      b. Calculate the loss function value of the current model on the training set $J$ = -1/n ∑($y_i \cdot$log(sigmoid($w \cdot x_i + b$)) + (1 - $y_i$)$\cdot$log(1 - sigmoid($w \cdot x_i + b$)))
      c. Determine whether the stopping conditions are met,
        - Reach the maximum number of iterations
        - Loss function convergence
    4. Return the parameter $w$ and $b$
```

# 3 Application instances

## 3.1 Characteristics of learning behavior indicators

The online learning behavior of students in college English courses studied in this paper is the general name of the actual operation process of students and users in the network environment. Under the network environment, the core of students' learning behavior will be transformed into data information and stored in the online platform of colleges and universities, and the characteristics of learning behavior indicators will also be selected from three data tables:

course information, user information and learning behavior. According to the data information of learning behavior and the actual needs of learning behavior mining and analysis, combined with the advice of many experts, this paper constructs an online learning behavior index system, as shown in Table 2 [9].

**Table 2.** Online learning behavior index system

| Learning behavior dimension | Characteristic index | Description |
|---|---|---|
| Student user information | Student ID | Student coding |
| | Age | One of the identification elements |
| | Recorded information | 0 or 1, 1 means that all records remain intact |
| Operation behavior | Number of learning events | Summary of learning process interaction |
| | Learning interval | Interval time of different learning events |
| | Browsing learning ratio | Percentage of course content module browsing learning |
| Collaborative behavior | Communication times | Number of times to participate in and initiate communication after class |
| | Number of interactions | Number of exchanges and discussions in class |
| Problem-solving behaviour | Learning degree | 0 or 1, 1 means that the interactive or browsing ratio of course content modules exceeds 50% |
| | Learning process completion degree | Percentage of completed learning progress |

### 3.2 Data acquisition and preprocessing

The original data used in this study comes from the online learning system of English courses for college students in our school from 2021 to 2022. Because of the large amount of data, it is necessary to transform the original data set into the data set of students' learning behavior analysis through preprocessing, and keep all the features in a unified numerical form through normalization and discretization, and summarize them in the same data table to provide convenience for subsequent mining processing.

The subsequent data mining running environment configuration includes two parts: hardware equipment and software program. In terms of hardware equipment selection, the CPU selects Intel Core i5-13600KF 3.5GHz, 8 cores, 32GB of memory and 500GB of hard disk space. In terms of software program, Windows 10.0 x86-64bit is selected as the bottom operating system, Python 3.8 as the development language, Pycharm 2019 as the integrated development tool and Mysql 5.7.31 as the database server. In addition, the deployment and import of algorithms such as Numpy and Pandas are directly completed in the PyCharm tool, and each algorithm model is also realized by scripting code.

### 3.3 Analysis mining

According to the online learning behavior index system proposed above, we first need to explore the correlation between various online learning behaviors and learning performance. There are three basic learning behaviors in the dimension of operational behavior: the number of learning events, learning interval and browsing learning ratio. After data preprocessing, 11,732 valid data were obtained, and the correlation analysis results were obtained by Perason correlation coefficient, as shown in Table 3.

**Table 3.** Analysis results of correlation between operational behavior and learning performance

| | | Number of learning events | Learning interval | Browsing learning ratio |
|---|---|---|---|---|
| Learning performance | Pearson correlation | 0.157 | 0.492 | 0.244 |
| | Significance | 0.000 | 0.000 | 0.000 |
| | Total sample number | 11732 | 11732 | 11732 |

The results show that the correlation coefficient between operational behavior and learning performance is greater than 0, and the significance is less than 0.001, which shows that operational behavior is positively correlated with learning performance. At the same time, the correlation index between learning interval and learning performance is the highest, which belongs to the main influencing factor in the dimension of operational behavior.

Similarly, the results of correlation analysis between cooperative behavior and learning performance, problem-solving behavior and learning performance are shown in tables 4 and 5. The results show that collaborative behavior and problem-solving behavior are significantly positively correlated with learning performance, and communication times and learning process completion are the main influencing factors in the two dimensions respectively.

**Table 4.** Analysis results of the correlation between cooperative behavior and learning performance

| | | Communication times | The number of interactions |
|---|---|---|---|
| Learning performance | Pearson correlation | 0.322 | 0.135 |
| | Significance | 0.000 | 0.000 |
| | Total sample number | 10311 | 10311 |

**Table 5.** Analysis results of correlation between problem-solving behavior and academic performance

| | | Learning degree | Learning process completion degree |
|---|---|---|---|
| Learning performance | Pearson correlation | 0.238 | 0.539 |
| | Significance | 0.000 | 0.000 |
| | Total sample number | 11245 | 11245 |

In addition, based on the above correlation analysis results between learning behavior and learning performance, this paper will focus on using Logistic regression algorithm to predict students' learning performance and form a regression model of learning performance. According to the requirements of English learning performance in colleges and universities, students are classified into two categories, that is, when their learning performance is $S \geqq 60$, their grades are judged as passing, and the strain is assigned to 1, which means there is no risk; When the learning performance is $S < 60$, it is judged as failing, and the value of the strain is 0, indicating that there is risk [10]. After inputting the Logistic regression model, the calculation results of each variable are shown in Table 6.

**Table 6.** Logistic regression model variable calculation results

| | b | SE | w | df | p | Exp(B) | 95% confidence interval | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower limit | Upper limit |
| Learning interval | 0.128 | 0.005 | 451.555 | 1 | 0.000 | 1.124 | 1.127 | 1.158 |
| Communication times | 0.011 | 0.004 | 5.154 | 1 | 0.021 | 1.015 | 1.002 | 1.021 |
| Browsing learning ratio | 0.022 | 0.002 | 132.446 | 1 | 0.000 | 1.023 | 1.016 | 1.024 |
| Learning process completion degree | 1.366 | 0.115 | 148.553 | 1 | 0.000 | 3.906 | 3.134 | 4.877 |
| constant | -4.121 | 0.168 | 623.115 | 1 | 0.000 | 0.018 | | |

According to the parameter results, the final fitting equation is shown in Formula 5. Among them, $x_1$, $x_2$, $x_3$, $x_4$ respectively represent four characteristics of learning behavior.

$$Logistic\left(S\right)=-4.121+0.128x_1+0.011x_2+0.022x_3+1.366x_4 \tag{5}$$

In order to further verify the prediction accuracy of Logistic regression model, this study compares the prediction results with the real data results, and the comparison results are shown in Table 7. The results show that the prediction accuracy of students' learning performance by Logistic regression model is 83.6%, which meets the research expectation and meets the practical application requirements.

**Table 7.** Horizontal comparison results between predicted results and real data results

| Real data results | Predicted results | | |
|---|---|---|---|
| | Learning performance | | Correct percentage |
| | Fail | Pass | |
| Pass | 537 | 2349 | 81.4% |
| Fail | 2944 | 332 | 89.9% |
| Total percentage | | | 83.6% |

## 4 Conclusions

In order to strengthen the supervision and management of online English teaching in colleges and universities, this paper puts forward a set of online learning behavior index system and analysis framework in view of the current situation of online English teaching in colleges and universities. The analysis framework takes data mining technology as the core, and focuses on using Pearson coefficient to calculate the correlation between learning behavior and English learning performance, and combining with Logistic regression model to classify and predict students' learning performance. The results show that students' operational behavior, cooperative behavior and problem-solving behavior are significantly correlated with their English performances, and learning intervals, communication times and the completion of learning processes are the key factors affecting their English performances. According to this result, teachers can adjust teaching methods in the follow-up English teaching to strengthen students' performance in this respect, and give corresponding guidance to help students correct bad learning behavior and improve the learning effect of the course. In the follow-up research, we will use deep learning algorithm to strengthen the accuracy of the prediction model, realize the intelligent analysis and processing of students' behavior, and make contributions to the implementation and development of the teaching reform process in colleges and universities.

**Acknowledgement:** The Project of Shandong Institute of Commerce and Technology “Analysis of Vocational College Students' Learning Characteristics and Research on Teaching Countermeasures under Blended Learning Background ” (Project No.23A3021)

## References

[1] Li Xinhao. Research on the Construction of Educational Informatization in Smart Campus Environment[J]. Office Informatization.10: 59-61 (2023)

[2] Nikolina Pleša Puljić Damir Ribić. Students' Perception of Online Teaching and Face to Face Teaching[J]. Drustvena istrazivanja. 10.517-536 (2023)

[3] Wei Ming et al. Analysis of Online Learning Behavior Characteristics Based on Data Mining[J]. Journal of Mianyang Teachers' College. 08: 97-104 (2023)

[4] Chwen Jen Chen Chee Siong Teh. Global Trends of Educational Data Mining in Online Learning[J]. International Journal of Technology in Education. 10: 656-680 (2023)

[5] M Munshi Tarun Shrimali. Data Model and Algorithm for Analysis of Data to Enhance Online Learning Using Graph Mining Techniques[J].Wireless Personal Communications. 04: 1-20 (2023)

[6] Yu Wenyu. Data Mining Technology and Application in Big Data Era[J]. Information & Computer. 03: 1-3 (2023)

[7] Wang Jing, Zhang Huan. Research on Implicit Behavior of College Students' Online Learning Based on Pearson Analysis[J]. China Management Informationization. 07: 159-161 (2018)

[8] Wei Qiang et al. Research on Influencing Factors of College Students' Online Learning Behavior Based on Binary Logistic Model[J]. Journal of Adult Education College of Hebei University. 09: 108-113 (2020)

[9] Yang Yuanqi. Study on Learning Behavior Analysis and Academic Prediction Based on Online Course Data[D]. Tianjin University of Commerce. 05 (2020)

[10] Fu Xiaobing. Study on Online Learning Performance Prediction Modeling Based on Learning Analysis[D]. Yunnan Normal University. 05 (2019)