Creative Education, 2023, 14, 1527-1535
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Study on Online Learning Behavior Analysis and Performance Prediction Based on Improved Random Forest Algorithm

Dongxu Liu

College of Information Science, Zhejiang Open University, Hangzhou, China
Email: 1184010040@qq.com

How to cite this paper: Liu, D. X. (2023). Study on Online Learning Behavior Analysis and Performance Prediction Based on Improved Random Forest Algorithm. *Creative Education, 14,* 1527-1535.
https://doi.org/10.4236/ce.2023.148097

Received: July 17, 2023
Accepted: August 7, 2023
Published: August 10, 2023

Abstract

With the popularity of online learning, the analysis of student learning behavior and the prediction of academic performance have become a hot research topic in the field of education. This study aims to analyze the factors influencing learning outcomes on online platforms, extract the feature variables from online learning behavior, and construct a weighted random forest model to predict students' online learning performance. Additionally, a method for computing feature importance is designed to analyze online learning behavior based on the significance of different behavioral features. This contributes to gaining early insights into students' learning progress, proactively identifying learning issues and strategically allocating educational resources, thus providing students with enhanced guidance and support.

Keywords

Online Learning, Behavioral Features, Personalized Learning, Random Forest

1. Introduction

With the rapid development of information technology and the widespread availability of the Internet, online learning has gained extensive application and promotion in the field of education (Li et al., 2014). The online learning mode supports flexible learning, enabling students to autonomously arrange their studies according to their own time and location. Furthermore, online learning platforms offer a diverse range of learning resources, allowing students to select resources that best suit their needs, thereby enhancing learning effectiveness. As a result, the popularization of online learning in university education provides students with more learning options and opportunities, while also promoting

innovation and development of university education (Liu et al., 2020).

The online learning platform records students' learning behavior data, which contains valuable information that can be used to analyze students' learning progress, preferences, and pace, among other factors. This analysis can provide personalized learning support to students and improve learning outcomes (Xiao et al., 2019). Additionally, the analysis of students' learning behavior can also be utilized to predict their academic performance, in order to identify learning problems early and provide targeted tutoring and interventions (Liu, 2020). Therefore, the analysis of students' learning behavior and performance prediction is of great significance in achieving personalized learning, improving teaching strategies, implementing early interventions, and advancing educational research.

Currently, numerous scholars have conducted research on learning behavior analysis and performance prediction. Chen et al. (2019) focused on short-term courses on online platforms and used classifiers to make early predictions of student learning outcomes, but the sample selection had certain limitations. Wang (2021) used the PSO-BP algorithm to predict the learning performance of target courses and demonstrated higher prediction accuracy compared to traditional BP algorithms. However, the data was not sufficiently collected and preprocessed. Fan (2019) predicted student performance using neural network algorithms based on data association principles. Zheng et al. (2021) used deep learning-based CNN and LSTM models to explore factors that affect learning outcomes and achieve prediction of online learning performance. Lee et al. (2021) created two deep neural network models to predict student performance based on their learning behaviors and exercise completion. But they all did not solve the problems of imbalanced data categories and parameter updates. Ren and Shang (2019) extracted behavior indicator data closely related to performance and used neural networks to predict online course grades. Mi et al. (2022) used a multivariate linear regression model for feature extraction and deep neural networks for prediction. But they overlooked the impact of manual feature filtering on model performance and failed to identify key factors for performance prediction.

The random forest (RF), proposed by Breiman (2001), is an ensemble learning method with excellent performance and scalability. It constructs multiple decision trees and makes predictions by voting or averaging the results, which enables it to handle high-dimensional data with good generalization ability and robustness. Yu and Zhou (2020) proposed an early warning model based on the random forest algorithm to predict the risk level of student performance. Arafiyah et al. (2020) used the similarity features of learners obtained from previous formative assessments to predict learners' progress based on the random forest model. Song et al. (2017) employed the random forest algorithm to predict students' scores and used the Gini index to rank the importance of courses affecting student performance. However, these algorithms do not comprehensively extract learning

behavior features and do not perform personalized analyses for different subjects and learning environments.

The aim of this study is to analyze online learning behaviors and predict students' academic performance based on an improved random forest algorithm. Specifically, first, key features are extracted from large-scale online learning data, and then data quality control methods are proposed to ensure the quality of training sample data. Subsequently, a weighted random forest model is constructed to accurately predict students' learning achievements, enabling early identification of learning problems and the provision of targeted guidance and intervention. Lastly, the importance of each feature is analyzed to provide strong evidence for personalized learning support and improvement of teaching strategies.

2. Data Collection and Preprocessing

2.1. Feature Selection

Prior to utilizing big data techniques for predicting any given entity, the foremost task is to collect sample data associated with the target of prediction. The same applies to predicting student performance, where we need to analyze and extract behavioral features related to student performance on online learning platforms.

1) Learning Duration: It refers to the total amount of time spent by students on the learning platform. It reflects the level of engagement in learning. A longer online learning duration indicates that the student has invested more time in studying, which may have a positive impact on their final grades.

2) Learning Progress: It refers to the ratio between a student's current completion level of learning materials and the expected completion level. Delays or advancements in the learning progress may potentially impact the final grades. Failure to adhere to the correct learning sequence could result in knowledge gaps or confusion in understanding, consequently influencing the ultimate academic outcome.

3) Click Records: It refers to the number of times students click on different resources such as videos, texts, and exercises on the learning platform. These records reflect students' learning preferences, learning strategies, and level of activity. A vector can be used to represent the click records for different resources.

4) Resource Viewing Count: It refers to the number of times teaching resources are viewed repeatedly, which to some extent reflects the difficulty level of the teaching content.

5) Assignment Submission: It refers to the ratio of students who submit assignments on time, which can be indicative of their learning diligence.

6) Assignment Quality Grading: It refers to the scoring of student assignments, portraying their interim learning achievements.

7) Interaction Frequency: It refers to the interaction between students and others in the discussion area of the learning platform, such as the number of

replies, the frequency of questions asked, etc. Interactive behavior provides a glimpse into students' learning attitude and collaborative skills.

8) Previous Semester Grades: Students' performance in the previous semester typically reflects their learning abilities and study habits, which to some extent impact their grades in the current semester.

Therefore, the input vector of the dataset used for predicting student grades consists of eight feature variables: Learning duration, Learning progress, Click records, Resource viewing count, Assignment submission, Assignment quality grading, Interaction frequency, and the previous semester grades.

2.2. Data Quality Assurance

Before using a dataset composed of feature sample data, it is necessary to filter out some abnormal data to ensure the quality of the sample data. In general, online learning behavior feature data may exhibit the following types of anomalies:

1) Missing Data: Students may not engage in online learning activities during certain time periods, resulting in the absence of data for those time intervals.

2) Outliers: There may be abnormal situations in students' online learning behavior, such as extreme study durations, unusually frequent or infrequent clicking behaviors, and the like.

3) Data Errors: Errors may occur during the data collection process, including incorrect timestamps, erroneous click counts, and so forth.

4) Data Duplication: Problems during the data collection process may lead to duplicate recording of data, resulting in data redundancy.

5) Data Bias: Data collection may suffer from biases, such as the inclusion of only specific types of students or data from specific time periods, thus leading to an incomplete dataset.

To address these anomalous data, the following strategies can be employed to ensure the quality and accuracy of the data:

1) Missing Data: Missing data can be filled by employing interpolation techniques to fill in the gaps or by predicting the missing values based on correlated data.

2) Outliers: Outliers can be identified and handled by setting thresholds or utilizing outlier detection algorithms.

3) Data Errors: Data cleaning and validation can be conducted to eliminate erroneous data.

4) Data Duplication: Duplicate records can be eliminated by applying deduplication techniques, ensuring the preservation of unique data entries.

5) Data Bias: Data synthesis methods can be employed to generate new samples to increase data diversity.

Ultimately, the data is partitioned into a training set and a testing set. The training set is employed to train and optimize the random forest model, while the testing set is employed to assess the model's performance. Subsequently, predictions for the students' academic performance are made.

3. Weighted Random Forest Model

Random forest is an algorithm that integrates multiple decision trees through ensemble learning, effectively handling high-dimensional data and complex relationships. In this study, the weighted random forest algorithm is employed to model and analyze students' learning behavior data, with the aim of predicting student performance.

3.1. Student Performance Prediction

The essence of random forest modeling lies in selecting feature samples to build multiple regression decision trees. In the prediction of academic performance, the eight influencing factors of academic performance are considered as the input feature vector, denoted as $x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(8)}\right)$. The corresponding performance for this feature vector is defined as the target value y_i. Consequently, we establish the training set for predicting student performance, designated as $\left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \cdots, \left(x_N, y_N\right)\right\}$, where N represents the total sample capacity of the training set.

Utilizing a resampling technique, m data sets are randomly extracted from the training data. Within the input space of each training dataset, the process recursively divides each region into two sub-regions and determines the output value for each sub-region, thereby constructing a binary regression tree. The specific steps are as follows:

1) Randomly select a few feature variables from the eight available variables and iterate through all the values of each variable in the dataset to solve for the optimal split variable j and split point *c*.

2) Utilize the selected optimal split (*j*, *c*) to partition the input space into two regions and determine the corresponding output values.

3) Repeat Steps (1) and (2) for each partitioned sub-dataset until the termination condition is met.

4) Divide the input space into datasets $R_1, R_2, \cdots, R_K$, with corresponding output values $G_1, G_2, \cdots, G_K$. The regression tree is generated as shown in Formula (1):

$$f_i\left(x\right) = \sum_{k=1}^{K} G_k I\left(x \in R_k\right) \tag{1}$$

where $f_i\left(x\right)$ is the regression tree trained on the i^{th} dataset, I is the indicator function. By repetitively applying the above steps in each dataset, a collection of m regression trees can be trained.

Generally, the final prediction result based on classical random forest is the average output value of m regression trees. In this paper, considering the different fitting accuracy of each regression tree, a weighted random forest (WRF) model can be used to predict student performance. The weight coefficients of respective regression trees are set based on their fitting capabilities. The higher the fitting accuracy, the greater the weight value of the regression tree. By combining the m regression trees with weighted integration, the overall accuracy and

generalization performance of the model can be improved.

Given the weighted coefficients $v_i\left(i=1,2,\cdots,m\right)$ of the weighted random forest algorithm, the calculation of the predicted student performance value $F(x)$ based on the weighted random forest algorithm can be derived using Formula (2).

$$F\left(x\right)=\sum_{i=1}^{m}v_i\,f_i\left(x\right) \tag{2}$$

3.2. Feature Importance Evaluation

One significant feature of the prediction model based on random forest is its ability to output feature importance. The feature importance helps in feature selection to a certain extent, thereby enhancing the robustness of the model.

Feature importance can be obtained by assessing the impact of each feature on the model's accuracy. First, the order of the eigenvalues of a feature s in the sample is disturbed to generate a new sample, and then the new sample is put into the established random forest model to calculate the accuracy. If feature "s" is highly important, the error of the modified samples will undergo significant changes. Conversely, if the error remains largely unchanged after modifying feature "s," it implies that feature "s" is not important. Therefore, the importance of feature "s" can be quantified using the difference in error values before and after the variation of feature "s", denoted as Z_s. The formula is as follows:

$$Z_s=\frac{\sum_{i=1}^{m}\left(w_i-ws_i\right)}{m} \tag{3}$$

where Z_s represents the average error value of m regression trees. w_i denotes the error obtained from the regression tree $f_i\left(x\right)$, while ws_i represents the error obtained after modifying the value of feature "s" in the sample.

4. Learning Behavior Analysis Based on Feature Importance

Feature importance reflects the extent to which a feature influences the predictive outcome of a model. Therefore, the feature importance in online learning performance prediction can provide valuable information for online learning behavior analysis. Specifically, it includes:

1) Understanding Factors Influencing Student Behavior: By analyzing feature importance, we can comprehend which student behaviors have a significant impact on learning outcomes. For instance, behaviors such as study time, participation in discussions, and submission of assignments may have a substantial influence on academic performance. It can assist teachers in understanding the contribution of student behavior to learning outcomes and subsequently devising appropriate instructional strategies and counseling measures.

2) Personalized Learning Support: By considering feature importance, we can gain insights into the discrepancies in significance among different students. This facilitates the creation of personalized learning support systems that offer targeted learning recommendations and resources in accordance with the im-

portance of specific features for each student. For instance, for a student with a high importance of learning time, guidance and resources related to time management can be provided.

3) Uncovering Learning Challenges: By comprehending the features that hold greater significance in predicting learning outcomes, corresponding alert models can be established to assist educators in preemptively identifying students' learning difficulties and implementing appropriate intervention measures.

Attention should be paid to the fact that the significance of features is merely a relative metric, derived from the training results of a model. In practical applications, it is crucial to integrate domain knowledge and real-life circumstances to effectively determine the actual impact of features on learning behavior.

5. Conclusion

This study employs an analysis of factors such as student study duration, learning progress, and assignment submission to assess their influence on academic performance. By extracting feature variables from online learning behaviors and improving the quality of feature sample data, a weighted random forest model is constructed to predict students' online learning performance. Furthermore, the model is utilized to conduct feature importance analysis for the purpose of isolating the key features that significantly influence student achievement. This facilitates the analysis of students' online learning behaviors in a more comprehensive manner.

The analysis of online learning behaviors and the prediction of academic performance contribute to an early understanding of students' learning situations and promptly identifying of learning issues. This allows for the rational allocation of educational resources to provide students with better guidance and support. By considering the varying impact of different behavioral characteristics on individual students, personalized learning support and recommendations can be designed, offering targeted tutoring and interventions to help students grasp knowledge and skills better. Moreover, performance prediction assists teachers in evaluating their teaching methods and strategies, fostering instructional improvements and formulating more scientific and effective educational policies. Ultimately, this approach aims to elevate students' learning achievements and experiences.

With the widespread adoption of online learning platforms, learning behavioral analysis and performance prediction have become a focal point of research in the field of education. The future prospects for research on online learning behavioral analysis are as follows:

1) Multimodal Data Analysis: In addition to traditional learning behavior data, future research can integrate multimodal data, such as students' voices, eye movements, facial expressions, and other physiological data, as well as data from social media, to conduct a more comprehensive analysis of learning behaviors. This will contribute to a better understanding of students' learning processes

and difficulties, and provide more accurate predictions and interventions.

2) Application of Deep Learning Algorithms: Deep learning algorithms have achieved remarkable breakthroughs in fields like image and speech processing. In the future, these algorithms can be employed for analyzing online learning behaviors and predicting academic performance. By harnessing the power of deep learning algorithms, it becomes possible to uncover more intricate patterns and principles of learning, thereby enhancing the accuracy and efficacy of predictions.

3) Issues of Educational Data Privacy and Ethics: With the accumulation and utilization of online learning data, concerns surrounding educational data privacy and ethics have become increasingly significant. Future research should focus on safeguarding students' privacy rights, while thoroughly leveraging the data to improve educational practices.

In summary, future research on online learning behavior analysis and performance prediction will prioritize personalized learning and the resolution of educational data privacy and ethical concerns. These studies will provide vital support and guidance for the advancement of online education and educational reform.

Acknowledgements

The research is funded by Zhejiang Open University 2023 First Class Curriculum Construction Project (Grant No. XYLKC202309).

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Arafiyah, R., Hasibuan, Z. A., & Santoso, H. B. (2020). Monitoring Online Learners Performance Based on Learning Progress Prediction. *AIP Conference Proceedings, 2331,* Article ID: 060012.

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/A:1010933404324

Chen, W. Y., Brinton, C. G., Cao, D. et al. (2019). Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors. *IEEE Transactions on Learning Technologies, 12,* 44-58. https://doi.org/10.1109/TLT.2018.2793193

Fan, L. (2019). Student Achievement Prediction Model Based on Big Data in the Context of Intelligent Education. *China Computer & Communication, 24,* 223-225.

Lee, C. A., Tzeng, J. W., Huang, N. F. et al. (2021). Prediction of Student Performance in Massive Open Online Courses Using Deep Learning System Based on Learning Behaviors. *Educational Technology & Society, 24,* 130-146.

Li, Y., Powell, S., Ma, H. et al. (2014). The Impact of MOOCs on Higher Education: A Perspective from Theory of Disruptive Innovation. *Modern Distance Education Research*, 55-59.

Liu, Y. G. (2020). Analysis and Research on Early Warning of College Students' Achieve-

ments Based on Big Data. *Digital Technology & Application, 38,* 95-96.

Liu, Z. Y., Lomovtseva, N., & Korobeynikova, E. (2020). Online Learning Platforms: Reconstructing Modern Higher Education. *International Journal of Emerging Technologies in Learning, 15,* 4-21. https://doi.org/10.3991/ijet.v15i13.14645

Mi, H. C., Gao, Z. H., Zhang, Q. R. et al. (2022). Research on Constructing Online Learning Performance Prediction Model Combining Feature Selection and Neural Network. *International Journal of Emerging Technologies in Learning, 17,* 94-111. https://doi.org/10.3991/ijet.v17i07.25587

Ren, Z. G., & Shang, F. H. (2019). Online Course Grade Prediction Model Based on Behavior Analysis. *Computer Technology and Development, 29,* 139-143.

Song, Y., Zhu, L. Q., & Cheng, Z. K. (2017). Research on Student Performance Evaluation Based on Random Forest. *Journal of Qiqihar University (Natural Science Edition), 33,* 1-5.

Wang, R. (2021). Research on Learning Achievement Prediction Based on PSO-BP Network. *Journal of Anyang Normal University, 2,* 41-45.

Xiao, Y. Z., Wang, P., Huang, C. Y. et al. (2019). Personalized Accurate Recommendation Analysis of Online Learning Content Based on Big Data. *Microcomputer Applications, 35,* 41-43.

Yu, X., & Zhou, Y. F. (2020). Research on Early Warning of College Students' Achievements Based on Random forest Algorithm in the Context of Big Data. *Jiangsu Science & Technology Information, 20,* 50-53.

Zheng, A. Q., Wang, Y. Q., & Hao, C. Y. (2021). Research on Online Academic Achievement Prediction with Deep Learning. *Computer Era, 12,* 69-72.