Research on Learner Profiles for Predicting Online Learning Behavior

Jiacheng ZHU, Zhangang WANG

Abstract— . **With the rapid development of online education, how to efficiently analyze and enhance learners' learning outcomes has become an important topic. Learner profiling, as one of the significant research directions of big data technology in the field of education, provides strong support for personalized teaching and learning alerts. This research employs the K-means clustering algorithm to meticulously classify and construct profiles of the learning records of 2,059 students in an online learning system at a certain university. Additionally, a predictive model is established using the gradient boosting decision tree algorithm to assess learning outcomes, aiming to provide specific improvement suggestions for online education, thereby more effectively enhancing learning quality.**

Index Terms—**Learner profiles, online learning behavior, K-means clustering algorithm, machine learning.**

I. Introduction

In 2022, the Ministry of Education mentioned at the press conference on the construction and application effectiveness of the National Smart Education Platform: "China ranks first in the world in both the number of MOOCs and the number of learners, and the proportion of college teachers using blended teaching has increased from 34.8% before the pandemic to 84.2%" [1], indicating the widespread application of online education. Learner profiling, as an application of this technology in the field of education, analyzes the data generated during learners' online learning processes to characterize group labels, helping learners and educators improve learning behaviors[2].

II. RELATED WORKS

The concept of "user persona" was first proposed by Alan Cooper, a pioneer in interaction design in the United States[3], viewing it as a target user model constructed based on real data. The essence of "user persona" is to generate digital labels and knowledge systems by mining user data, comprehensively presenting user characteristics, and providing management decision-making references for enterprises[4]. "Learner persona," on the other hand, is a visual learning analysis technique that applies "user persona" in the field of education, commonly referred to as "learner persona." The main research directions of learner personas include modeling, algorithm design, and real-world application scenarios. In the modeling dimension, Feng Xiaoying, Zheng Qinhua, and others explored indicators reflecting online learning quality, identifying 13 learning behavior indicators significantly related to online cognitive levels[5]. In the direction of algorithm design, Francisco J extracted learner personas based on the Shapelet time series classification algorithm using association rules from big data[6]. In the direction of real-world application scenarios, Wu Hanlin completed the recommendation of students' learning paths in online learning systems[7], and Han et al. developed a proxy detection mechanism based on user propagation behavior profiles[8].

Manuscript received August 28, 2024

Foundation item: Supported by Research Project of Continuing Education Teaching Reform and Quality Improvement in Tianjin in 2023 (J2023014)

Jiacheng ZHU, School of Software, Tiangong University, Tianjin, 300387,China.

Zhangang WANG, School of Software, Tiangong University, Tianjin, 300387,China.

III. Building Learner Profile Models Based on Online Learning Behavior

This paper will construct learner profiles through five steps: data acquisition, data preprocessing, label extraction, modeling, and visualization. Subsequently, based on the profile results, predictions of student performance will be made, as shown in Figure 1.

A. Data Acquisition

Data is the core component of learner profiles, and its quality directly affects the accuracy and comprehensiveness of the learner profile results. This research utilizes an online teaching system from a certain university to analyze the learning records of 2,059 students in the course "*Introduction to MAO Zedong Thought and the theoretical system of socialism with Chinese characteristics*" The main data collected includes students' basic information and online learning records, as detailed in Table 1.

B. Data Preprocessing

In the process of constructing learner profiles, data preprocessing is a crucial step. The data preprocessing process in this research is illustrated in Figure 2.

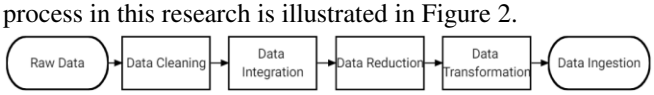

Figure 2 Data Preprocessing Process

To improve the predictive performance of the model, this research employed the min-max normalization method to normalize data such as attendance, scaling the data to the range of [0, 1]. The normalization formula is as follows:

$$x' = \frac{x - X_{min}}{X_{max} - X_{min}}$$

Through the above data preprocessing steps, the original data has been cleaned and organized, providing a reliable data foundation for the subsequent generation and analysis of learner profiles.

C. Profile Label Extraction

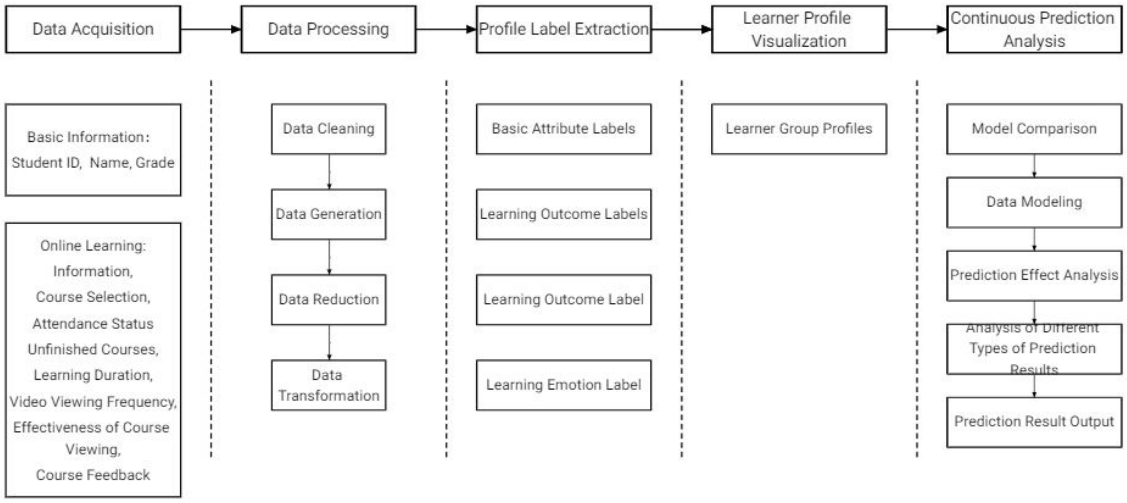

Figure 1. Flowchart of This research

Table 1 Data Information

Data Name	Data Content	Value
Basic Information	Student ID, Name, Class	Basic information of the student, string
Learning Records	Attendance	Value ranging from 0 to 10
	Quiz Performance	Value ranging from 0 to 20
	Final Score	Value ranging from 0 to 100
	Course Access Duration, Video Watching Duration	Value in numbers (unit: seconds)
	Video Viewing Count, Course Access Count	Value in numbers (unit: times)
	Grade	A ≥ 90, B ≥ 80 and < 90, C ≥ 60 and < 80, D < 60

By identifying learner characteristics, this research reflects the common traits of the learning group, thereby forming learner profiles [9]. In this research, after preprocessing the learners' learning data, it can be categorized into four labels: basic attributes, learning outcomes, learning behaviors, and learning emotions, as shown in Table 2.

Table 2. Establishment of Profile Labels

Profile Label	Data Indicators
Basic Attribute Label	Student ID, Name, Class
Learning Behavior Label	Attendance, Course Access Duration, Video Watching Duration
Learning Outcome Label	Quiz Performance, Final Score, Grade Level
Learning Emotion Label	Course Access Frequency, Video Viewing Frequency

D. Data Analysis

Based on relevant research, this research primarily uses Excel and SPSS for data processing, employing the K-means clustering algorithm to classify learners.

When applying the K-means clustering algorithm, this research utilized the elbow method and found a significant inflection point at the number of clusters (k = 3), indicating that the clustering effect is optimal when (k = 3). At this point, three learning groups are formed.

Table 3. Clustering Results

Group	Number of People	Learning Behavior	Learning Outcome	Learning Emotion
1	208	0.353	0.457	0.124
2	451	0.388	0.933	0.466
3	1400	0.370	0.927	0.235

By clustering the three labels reflecting learning outcomes and conducting mean analysis, different types of learners can be distinguished. The specific mean data is shown in Table 3. In Table 3, it can be observed that the mean values of all labels for learner group 1 are the lowest, indicating that there is still significant room for improvement in learning behavior, learning outcomes, and learning emotions. Conversely, learner group 2 has the highest mean values for all labels, demonstrating the most positive learning behavior and emotions, as well as the best learning outcomes. The mean values for learner group 3 fall between the two.

E. Visualization Output of Profiles

After completing data preprocessing, profile label extraction, and data analysis, it is essential to visualize the group profiles of learners to more intuitively reflect their learning outcomes.

First, principal component analysis (PCA) is used in SPSS to reduce the dimensionality of the learners' learning data. The output results indicate that the cumulative percentage of variance explained by PC1 and PC2 reaches 78.26%, demonstrating that these two principal components are highly representative. By using these two principal components as the x-axis and y-axis, the results of the K-means clustering algorithm are displayed in a scatter plot, as shown in Figure 3.

Then, by conducting a visual analysis of the label values in Table 3, three learner groups' profiles are formed. The group profiles are presented in the form of radar charts, as shown in Figure 4.

Finally, the overall mean of each label is calculated and compared with the label values of the three groups to conduct a specific profile analysis, resulting in three types of profiles:

Marginal Learners, Diligent Learners, and Balanced Learners.

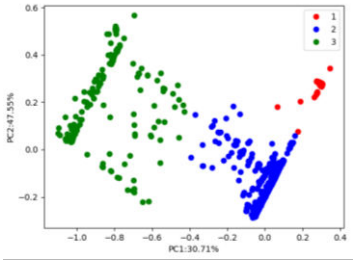

Figure 3. K-means Clustering Diagram

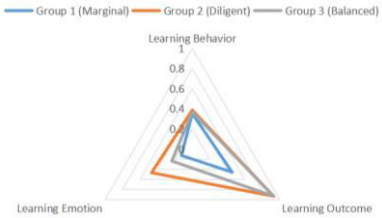

Figure 4. Group Profile Radar Chart

The profile of Marginal Learners is illustrated in Figures 5 and 6. This group has the smallest number of individuals, and all label values are below the overall mean. In terms of learning behavior, this group generally has shorter research durations and lower attendance rates. Their learning emotions are also low, with fewer accesses to teaching resources such as learning videos. Regarding learning outcomes, this group's performance in regular tests and final scores is below the mean, indicating significant room for improvement in test scores.

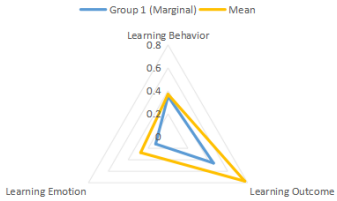

Figure 5. Comparison of Label Mean Values for Marginal Learners

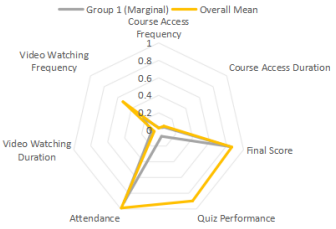

Figure 6. Comparison of Group Attributes Mean Values for Marginal Learners

The profile of Diligent Learners is illustrated in Figures 7 and 8. This group has label values that are significantly higher than the overall mean. In terms of learning behavior, This group generally exhibits longer study durations and higher attendance rates, indicating a more serious attitude towards their courses. Their learning emotions are also high, with the highest access frequency to teaching resources, allowing them to complete almost all learning tasks. Regarding learning outcomes, this group's performance in regular tests and final scores is above the overall mean, reflecting a relatively excellent level.

The profile of Balanced Learners is illustrated in Figures 9 and 10. This group has the largest number of individuals, and their label values are nearly on par with the overall mean. In terms of learning behavior, this group generally has moderate study durations and high attendance rates. Their learning emotions are relatively high, with frequent access to teaching resources, allowing them to complete most learning tasks. Regarding learning outcomes, this group's performance in regular tests and final scores is in line with the overall mean, achieving a good level.

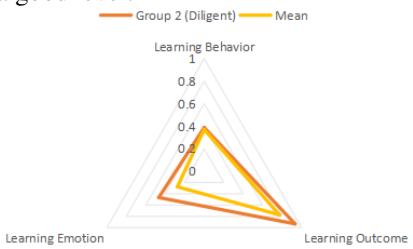

Figure 7. Comparison of Label Mean Values for Diligent Learners

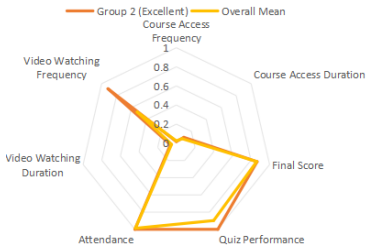

Figure 8. Comparison of Group Attributes Mean Values for Diligent Learners

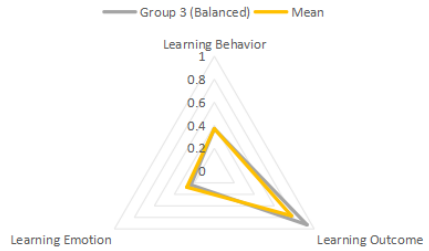

Figure 9. Comparison of Label Mean Values for Balanced Learners

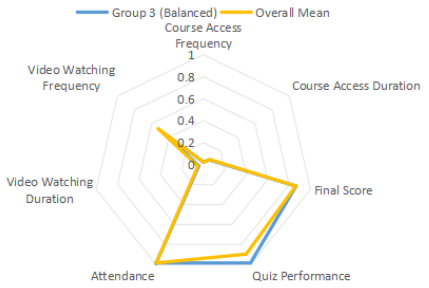

Figure 10. Comparison of Group Attributes Mean Values for Balanced Learners

Through the analysis of the profiles of Marginal Learners, Diligent Learners, and Balanced Learners, and after comparing the mean values of learner attributes, it is found that the frequency of video views, attendance rates, and test scores are highly correlated with final exam results. Therefore, it can be concluded that video viewing frequency, attendance rates, and test scores are key factors reflecting learners' learning outcomes.

IV. Learning Outcome Prediction Analysis Based on Learner Profiles

A. Performance Evaluation Criteria

For the training set

$$T = \{(x_1, y_1), \ldots, (x_m, y_m)\}$$

Samples can be categorized into four situations based on the true classes and the predicted classes of the learners: True Positives (TP) refer to samples that are predicted as positive and are actually positive; False Positives (FP) refer to samples that are predicted as positive but are actually negative; True Negatives (TN) refer to samples that are predicted as negative and are actually negative; False Negatives (FN) refer to samples that are predicted as negative but are actually positive.

Precision (P), defined as the proportion of correctly classified samples for each class out of the total number of samples.

$$P = \frac{TP}{TP + FP}$$

Accuracy (T), defined as the proportion of all correctly predicted samples out of the total number of samples.

$$\text{accuracy}(T) = \frac{1}{m}\sum_{i=1}^{m} |f(x_i) = y_i|$$

Recall (R), defined as the proportion of correctly classified positive samples among all positive samples.

$$R = \frac{TP}{TP + FN}$$

F1 Score is calculated using the following formula:

$$F = \frac{2 \times PR}{P + R}$$

The F1 Score takes into account both precision and recall, providing a more comprehensive assessment of the algorithm's performance.

B. Experimental Results Analysis

This research is based on learner profiles, using indicators such as the number of course learning sessions, course learning duration, exam scores, test scores, attendance scores, video viewing time, and video viewing frequency as independent variables, while the students' grade levels are treated as the dependent variable. Various models, including logistic regression, decision tree classifier, random forest classifier, linear support vector machine, polynomial kernel support vector machine, radial basis function kernel support vector machine, gradient boosting decision tree, and AdaBoost classifier, are used to predict student grades and compare the prediction performance of each model, as shown in Figure 11. From the figure, it can be seen that the gradient boosting decision tree algorithm performs the best overall, achieving an accuracy of 82.4% and an F1 score of 82.1%. Therefore, this research selects the gradient boosting decision tree algorithm for predictive analysis of the three groups.

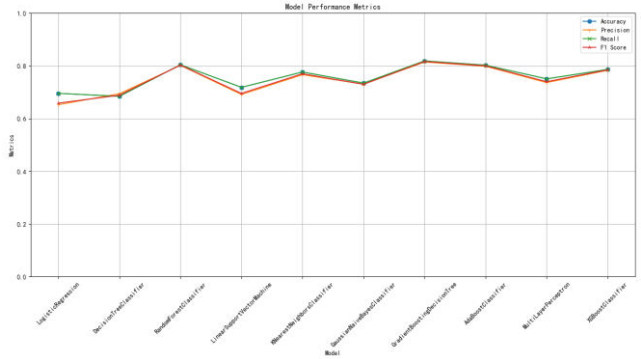

Figure 11 Model Performance Metrics

For different types of student groups, the model's prediction accuracy shows significant differences. The classification prediction results for the groups are shown in Table 4.

For marginal students, the algorithm's accuracy in predicting grade A is 18.2%, while the prediction accuracies for grades C and D are 81.4% and 50.0%, respectively. This indicates that the algorithm has a weak ability to predict high scores for marginal students but demonstrates better predictive capability in the middle and low score ranges (grades C and D).

In the diligent student group, the algorithm achieves a prediction accuracy of 87.6% for grade A, showing good predictive performance. However, the prediction accuracies for grades B and C are 51.5% and 41.7%, respectively. This may be due to a lack of sample data at this level, resulting in poor prediction performance.

For balanced students, the algorithm's prediction accuracy for grade A is 89.7%, indicating strong predictive capability, especially in the high score range. However, the prediction accuracies for grades B, C decline to 68.8% and 66.7%,respectively. This suggests that while the prediction performance for balanced students is relatively stable, improvements are needed in predicting low scores.

Table 4: Group Classification Prediction Results

	Marginal	Diligent	Balanced
A	18.2%	87.6%	89.7%
B	33.3%	51.5%	68.8%
C	81.4%	41.7%	66.7%
D	50.0%	0.0%	0.0%
Accuracy by Group	61.9%	74.3%	81.4%

Note: A 0% prediction accuracy indicates a lack of relevant dataset, making effective prediction impossible for the model.

In summary, the gradient boosting algorithm performs best in predicting the grades of high-achieving students, providing important data support for educational managers in formulating teaching strategies tailored to different student groups. Although the algorithm has certain limitations in prediction accuracy for marginal students, this does not indicate a deficiency in the algorithm itself but rather reflects the characteristics of the dataset and the scarcity of samples. Future research could consider collecting more data on marginal students to enhance the algorithm's predictive capability or explore the combination of other algorithms with the gradient boosting algorithm to improve prediction accuracy for this group.

V. Summary and Outlook

This research analyzed the learning records of 2,059 students enrolled in the "Introduction to Systems" course at a certain university's online teaching system. Using the K-means clustering algorithm, three groups were formed: marginal learners, diligent learners, and balanced learners. The visualization of learner profiles was completed, and the gradient boosting decision tree algorithm was used to validate the test set, revealing that this algorithm accurately predicts grade A for diligent and balanced learners.

Learning alerts are one of the important applications of learner profiling. By analyzing the learning records generated during the learning process, it predicts risks to achieve the purpose of early warning. In future research, learner profiles can be utilized to implement learning alerts during the learning process, proposing corresponding learning strategies based on different learner types to reduce learning risks and enhance online learning outcomes.

REFERENCES

[1] Kong, X. (2022). Analysis of Network Security and Management of Online Open Courses in Higher Education in the Post-Pandemic Era. China Information Security, (06), 87-89.
[2] Sun, F., & Dong, W. (2020). Research on User Profiling in Online Learning Based on Learning Analytics. Modern Educational Technology, 30(4), 5-11.
[3] Cooper, A. (2004). The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity. Indianapolis: SAMS.
[4] Mo, W. (2021). Research on the Construction and Application of Learner Profiles. Journal of Hunan University of Science and Technology (Natural Science Edition), (3), 64-69.
[5] Feng, X., Zheng, Q., & Chen, P. (2016). Research on Evaluation Models of Online Cognitive Levels from the Perspective of Learning Analytics. Distance Education Journal, 34(06), 39-45.
[6] Baldán, F. J., & Benítez, J. M. (2019). Distributed FastShapelet Transform: A Big Data Time Series Classification Algorithm. Information Sciences, 496.
[7] Wu, H. (2021). Research and Design of Learning Path Recommendations Based on Learner Profiles (Master's thesis). Zhejiang Normal University.
[8] Han, Z. H., Chen, X. S., Zeng, X. M., et al. (2019). Detecting Proxy Users Based on Communication Behavior Portrait. The Computer Journal, 62(12), 1777-1792.
[9] Tian, Y. (2020). Research on MOOC Learning Situation Early Warning Based on Learner Profiles (Master's thesis). East China Normal University. DOI: 10.27149/d.cnki.ghdsu.2020.001377.