Journal of Informatics and Web Engineering
Vol. 2 No. 2 (September 2023) eISSN: 2821-370X

Dropout Prediction Model for College Students in MOOCs Based on Weighted Multi-feature and SVM

Zhang Yujiao[1*], Ang Ling Weay[2], Shi Shaomin[3], Sellappan Palaniappan[4]
[1,2,4] School of Information Technology, Malaysia University of Science & Technology, Petaling Jaya, Malaysia.
[1,3] 4WGP+JQP, Shenheer Rd, Chang'An, Xi'An, 710100 Shaanxi, China.
**corresponding author:(zhang.yujiao@phd.must.edu.my; ORCiD: 0009-0006-9191-8942)*

Abstract - Due to the COVID -19 pandemic, MOOCs have become a popular form of learning for college students. However, unlike traditional face-to-face courses, MOOCs offer little faculty supervision, which may result in students being insufficiently motivated to continue learning, ultimately leading to a high dropout rate. Consequently, the problem of high dropout rates in MOOCs requires urgent attention in MOOC research. Predicting dropout rates is the first step to address this problem, and MOOCs have a large amount of behavioral data that can be used for such predictions. Most existing models for predicting MOOC dropout based on behavioral data assign equal weights to each behavioral characteristic, despite the fact that each behavioral characteristic has a different effect on predicting dropout. To address this problem, this paper proposes a dropout prediction model based on the fusion of behavioral data and Support Vector Machine (SVM). This innovative model assigns different weights to different behavior features based on Pearson principle and integrates them as data inputs to the model. Dropout prediction is essentially a binary problem, Support Vector Machine Classifier is then trained using the training dataset 1 and dataset 2. Experimental results on both datasets show that this predictive model outperforms previous models that assign the same weights to the behavior features.

Keywords—MOOCs, Dropout Prediction, Weighted Feature, SVM, Machine Learning

Received: 30 March 2023; Accepted: 19 June 2023; Published: 16 September 2023

I. INTRODUCTION

MOOC is an acronym for Massive Open Online Course [1], an online course that grew out of MIT's Open Educational Resources (OER) movement. The first open course was organized in 2007 on the Wikiversity platform, which was founded in 2006 [2]. In 2011, Andrew Ng, a renowned professor of Artificial Intelligence (AI) at Stanford College, launched an online course that attracted nearly 100,000 students worldwide, and since then MOOCs have spread around the world. The most popular MOOC platforms include Coursera, Udacity, and edX [3].

There are two types of MOOCs: “cMOOC” and “xMOOC” [4], coined by Stephen Downes in 2008. cMOOCs emphasize communication, sharing, creation, and acquisition of knowledge about a particular field through networks among diverse learners, based on the theory of connectivism, and provide a small amount of course content. In contrast, xMOOCs are more similar to traditional instructional models that focus on knowledge transfer by instructors rather than knowledge sharing among students. Originally, MOOCs were considered an

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2023.2.2.3

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

additional form of teaching, but since the outbreak of the COVID -19 pandemic, they have become one of the main teaching methods at universities.

However, MOOCs have a high dropout rate because they are online courses with relatively little student support. Reportedly, dropout rates for some MOOC courses range from 91% to 93% [5]. Therefore, predicting the dropout rate is necessary to prevent students from dropping out.

Online learning behavioral data, such as login data, visits to the course wiki, duration of video study, and posts and comments in discussion forums, are typically stored on MOOC platforms. As Abdelali [6] noted, the abundance of behavioral data in online education presents a valuable opportunity for Data Mining (DM). Researchers have found a strong relationship between online learning behaviors and dropout rates by studying behavioral characteristics in MOOCs, which has led to the development of dropout prediction models based on multi-behavioral data.

However, existing dropout prediction models based on multiple online behavioral data may not be equally accurate. Different behavioral characteristics have different effects on the prediction of dropouts according to the Pearson principle. To address this problem, this research aims to predict whether a college student will be absent from a MOOC based on behavioral data stored in MOOC platforms in the previous days. First, the raw data are collected and processed. Then, appropriate behavioral features are selected according to the magnitude of the Pearson correlation coefficient and an appropriate weight is assigned to each feature. Finally, a Support Vector Machine (SVM) classifier is used to predict dropout since the integrated features are the training data. In addition, the main contributions of this study are as follows:

(1) We propose a novel weighted muti-feature fusion algorithm for behavior data based on Pearson' principle.
(2) We propose a model for predicting the dropout of college students in MOOCs based on behavior features and a SVM classifier.
(3) We investigate the feasibility of using the proposed model to predict college student dropout at the initial stage and over time in a MOOC platform.

II. LITERATURE REVIEW

The field of Learning Analytics (LA) [7] in online learning mainly relies on recorded data about students' learning processes on web-based teaching platforms. LA aims to generate statistics about students' behavior, behavior patterns (login, browsing resources, online communication and so on), behavior objects (different resources, online course modules and so on), and when behavior occurs. The data are then visualized and analyzed to identify factors that influence online learning. Given the importance of predicting learning dropouts, several models have been developed based on online behavioral data.

Chen et al. [8] proposed a novel DT-ELM model that combines Decision Tree (DT) and Extreme Learning Machines (ELM) to map DT to ELM, based on entropy theory. This model achieves Accuracy, AUC, and F1 score of 0.941, 0.8596, and 0.9558, respectively, over 5 weeks. The outcomes show higher Accuracy, AUC, and F1 score in the KDD Cup 2015 dataset (extracted 23 kinds of behavior features) than benchmark algorithms. Muthukumar and Bhalaji [9] developed a dropout prediction system using Deep Neural Networks (DNN) Algorithm, using five weeks of behavioral data (9 types) from a MOOC course co-hosted by Harvard and MIT to construct the predictive model and provide appropriate interventions for students at-risk. This DNN based model achieves Precision and Recall of 0.9868 and 0.8468, respectively, the results show that the deep learning algorithm is more accurate in predicting dropout than the baseline model.

Wen et al. [10] found that students in MOOCs often have similar learning status on several consecutive days, indicating a local correlation of learning behavior that should not be ignored. Therefore, they proposed an innovative Convolutional Neural Network (CNN) model that utilizes a simple feature matrix to retain pertinent information about the local correlation of learning behaviors. This model uses seven kinds of behavior features in the KDD Cup 2015 dataset and aims to predict learning dropout. To validate the proposed model, Wen et al. adopted seven other classification models such as Classification And Regression Tree (CART), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Gradient Boosted Decision Tree (GBDT), Logical Regression (LR), Random Forest (RF), and SVM for comparison. The experimental results show that their proposed CNN model outperforms the other baseline algorithms in terms of $F-measure$ and Accuracy, and ranks second in Precision among the seven baseline algorithms.

Fu et al. [11] focused their study on predicting whether students in a MOOC will drop the course within the next 10 days. They developed a CLSA-based (Composite LSH-Sensitive Approximation) model for predicting course dropout, which is a type of machine learning model that is particularly suitable for high-dimensional data.

The model incorporates a static attention mechanism that allows it to focus on the most important dimensions of the data by assigning an attention weight to each dimension. To this end, the model uses a vector that is integrated into the time series and captures the temporal behavior of students. This vector is then used to obtain an attention weight for each dimension. To train the model, Fu et al. tracked students' behavioral data for five weeks, using seven features such as the number of videos viewed, pages closed, course wikis viewed, number of accessing other course items, navigation to another part of the course, number of forum posts and quizzes attempted. Using these data, they trained the model to predict whether students would be absent from the MOOC within the next 10 consecutive days. The model achieves an Accuracy of 87.6% and an F1 score of 86.9%, both of which are measures of how well the model predicts dropout. This suggests that the model is good at predicting student behavior and could be used to identify and target at-risk students for intervention.

In a separate study, Nitta et al. [12] presented a dropout prediction model using graph-based Machine Learning (ML) algorithms on OULAD using 20 types of features. Their approach is based on tensor decomposition and transformation, a technique for analyzing and processing high-dimensional data. The model they developed is designed to identify patterns in student behavior that are associated with dropping out of school and use those patterns to predict which students are at risk of dropping out. This model achieves a higher Precision of 0.745 compared to graph convolutional networks (GCN) and is overall almost as effective as the comparison model.

Nithya and Umarani [13] developed a MOOC dropout prediction model based on FIAR-ANN and features of participants' learning behavior (7 types). The model used an association rule FP growth approach for feature generation, and the neural network was implemented from Frequent Itemset-3, achieving higher Accuracy, Precision, Recall and F1-score of 0.92, 0.93, 0.99 and 0.91, respectively. This model is better than the baseline values in every respect. Sultan et al. [14] proposed a MOOC dropout prediction model using Artificial Neural Networks (ANNs) in KDD Cup 2015 (111 kinds of behavior data) that achieves a Precision of 91% and an Accuracy of 90%. This dropout algorithm is better than the baselines in terms of Precision and Accuracy. The summary of dropout predictive model in MOOC is shown in Table 1.

Table 1. Summary of Dropout Predictive Model in MOOC

Authors	Algorithm	Behavior Features	Dataset	Results
Chen et al. [8] (2019)	DT-ELM	23 types	KDD Cup 2015 5 weeks	DT-ELM with entropy-based feature selection, has higher Accuracy, AUC, and F1 score of 0.941, 0.8596, and 0.9558, respectively, over 5 weeks.
Muthukumar and Bhalaji [9] (2020)	DNN	9 types	A course jointly launched by MIT and Harvard 5 weeks	DNN based algorithm achieves Precision and Recall of 0.9868 and 0.8468, respectively, and is more accurate in predicting dropout than the baseline model.
Wen et al. [10] (2020)	CNN, CART, LDA, NB, GBDT, LR, RF, SVM	7 types	KDD Cup 2015 30 days	CNN algorithm outperforms the other baseline algorithms in terms of $F-measure$ (0.9247) and Accuracy (0.8764), and ranks second in Precision (0.8938) among the seven baseline algorithms.
Fu et al. [11] (2021)	CLSA	7 types	KDD Cup 2015 5 weeks	The algorithm achieves an Accuracy of 87.6% and an F1 score of 86.9%.
Nitta et al. [12] (2021)	Graph-based ML	20 types	OULAD 30 days	The algorithm achieves a higher Precision of 0.745 compared to graph convolutional networks (GCN) and is overall almost as effective as the comparison model.

Nithya and Umarani [13] (2022)	FIAR-ANN	7 types	KDD Cup 2015 30 days	FIAR-ANN achieves Accuracy, Precision Recall and F1-score of 0.92, 0.93, 0.99 and 0.91, respectively, which is better than the baseline values in every respect.
Sultan et al. [14] (2022)	ANNs	111 types	KDD Cup 2015 30 days	The algorithm achieves a Precision of 91% and an Accuracy of 90%, which is higher than that of baselines

III. RESEARCH METHODOLOGY

A. Support Vector Machine

The Support Vector Machine (SVM) is a binary classification algorithm developed by Vapnik and Cortes based on small sample data [15]. It is based on statistical learning theory and structural risk minimization [16], which makes it more suitable for limited training data compared to other machine learning classifiers [17]. Dropout prediction is essentially a problem of identifying two types of data, and SVM is primarily designed to find the optimal hyperplane for two types of data and maximize the separation between them [18]. The algorithm of SVM is shown in Figure 1.

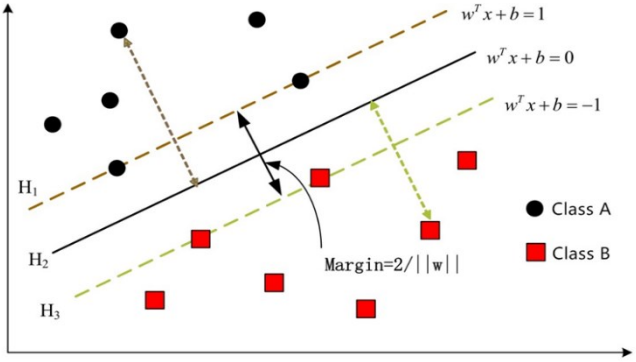

Figure 1. Classification Hyperplane of Support Vector Machine

The aim of SVM is to seek a hyperplane to optimally separate two classes of samples. For a space R, its dimension is $d=2$ and a linearly separable training dataset of n points: $(\vec{x}_i, y_i)_{1\le i\le n}, x\in R^d, d=2$, where $(\vec{x}_i, y_i)$ is the training sample, $y\in\{-1,+1\}$ is a label representing the class to which $\vec{x}_i$ belongs, the decision function of SVM classifier can be described as the following algorithm:

The optimal hyperplane is defined in formula (1):

$$\vec{w}\cdot\vec{x}+b=0 \tag{1}$$

Where $\vec{w}$ is the normal vector to the hyperplane and b is the bias value.

The constraint condition of the sample data $(\vec{x}_i, y_i)$ to the hyperplane is given in (2):

$$y_i\cdot(\vec{w}\cdot\vec{x}_i+b)\ge 1,\ i=1,2,...,n \tag{2}$$

In SVM classifier, in order to separate the two classes of data perfectly, we need to maximize the margin between them and obtain the best hyperplane [19]. Therefore, the original classification problem is converted to solve the following constrained problem (3):

$$\begin{aligned} &\textit{maximize} && \frac{2}{\|\vec{w}\|} \\ &\textit{subject to} && y_i\cdot(\vec{w}\cdot\vec{x}_i+b)\ge 1, i=1,2,...,n \end{aligned} \tag{3}$$

Where $\|\vec{w}\|$ is the norm of $\vec{w}$, $\frac{2}{\|\vec{w}\|}$ is the margin of two classes of data.

Then the constrained problem (3) is transformed into the solution of the constrained problem (4) according to the duality theorem.

$$\begin{aligned} &\textit{minimize} \quad \Phi(\vec{w}) = \frac{1}{2}\|\vec{w}\|^2 \\ &\textit{subject to} \quad 1 - y_i \cdot (\vec{w} \cdot \vec{x}_i + b) \le 0,\ i = 1,2,...,n \end{aligned} \tag{4}$$

Lagrangian function [20] is used to optimize the problem (4), and the original problem is then transformed into solve the following optimization problem (5):

$$\begin{aligned} \textit{minimize}_{w,b}\textit{maximize}_{\alpha} \quad L(w,b,\alpha) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\alpha_i\left[1 - y_i(\vec{w}\cdot\vec{x}_i + b)\right] \\ &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i(\vec{w}\cdot\vec{x}_i + b) + \sum_{i=1}^{n}\alpha_i \\ \textit{subject to} \quad \alpha_i &\ge 0,\ i = 1,2,...,n. \end{aligned} \tag{5}$$

Where $L(w,b,\alpha)$ represents the Lagrangian function, and α is the corresponding Lagrange multiplier. After processed by Lagrange multiplier, the optimal solution (6) is obtained as follows:

$$\begin{cases} \vec{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \vec{x}_i \\ b^* = -\dfrac{\max y_{i=-1}(\vec{w}\cdot\vec{x}_i) + \min y_{i=1}(\vec{w}\cdot\vec{x}_i)}{2} \end{cases} \tag{6}$$

Where $\vec{a}^*, \vec{w}^*, b^*$ represent the corresponding optimal solution of $\vec{a}, \vec{w}, b$, and they must satisfy the condition in formula (7):

$$\alpha_i^*\left[y_i(\vec{w}\cdot\vec{x}_i + b^*) - 1\right] = 0,\ i = 1,2,\cdots,n \tag{7}$$

Finally, the optimal hyperplane function is obtained in formula (8):

$$f(\vec{x}, \vec{\alpha}^*, b^*) = \sum_{i,j=1}^{n} y_i \alpha_i^* (\vec{x}_i \cdot \vec{x}_j) + b^* \tag{8}$$

Where $(\vec{x}_i \cdot \vec{x}_j)$ is the dot product of $\vec{x}$. And the decision function of SVM is given in formula (9) according to the Kuhn–Tucker theorem [21].

$$f(\vec{x}) = sign\left\{(\vec{w}^* \cdot \vec{x}) + b^*\right\} = sign\left(\sum_{i,j=1}^{n} \alpha_i^* y_i (\vec{x}_i \cdot \vec{x}_j) + b^*\right) \tag{9}$$

Where $sign$ is the function in formula (10):

$$\mathrm{sign}\, x = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{10}$$

For linear inseparable data sets, a relaxation factor ξ and a penalty factor C are included in the decision function to find the optimal hyperplane. Meanwhile, for nonlinear dataset, it can be mapped into a high-dimensional space using a suitable kernel function. The kernel functions most widely used in SVM are shown in formula (11):

$$\begin{cases} \textit{Polynomial kernel}: & K\left(x_i, x_j\right) = \left[\left(x_i \cdot x_j\right) + 1\right]^q \\ \textit{Gaussian kernel}: & K\left(x_i, x_j\right) = \exp\left\{-\frac{\left\| x_i - x_j \right\|^2}{\sigma^2}\right\} \\ \textit{Sigmoid kernel}: & K\left(x_i, x_j\right) = \tanh\left(v \cdot \left(x_i \cdot x_j\right) + c\right) \end{cases} \tag{11}$$

Where q is the degree of the polynomial in the Polynomial kernel, σ is the width of the Gaussian filter in Gaussian kernel, v, c, and tanh are slope, intercept, and hyperbolic tangent function, respectively, in the Sigmoid kernel.

B. The Design of the Dropout Prediction Model

In this work, we approach the dropout prediction as a binary problem, taking into account that during the COVID -19 pandemic, many e-classes at colleges have at most a few hundred students. For this reason, Support Vector Machines are used, which have been shown to be particularly powerful for small sample sizes. In addition, the study incorporates online behavioral data with the fusion of multiple features to identify the importance of multi-features to the predictive model. The basic structure of the proposed dropout prediction model is shown in Figure 2, and its main steps are described below.

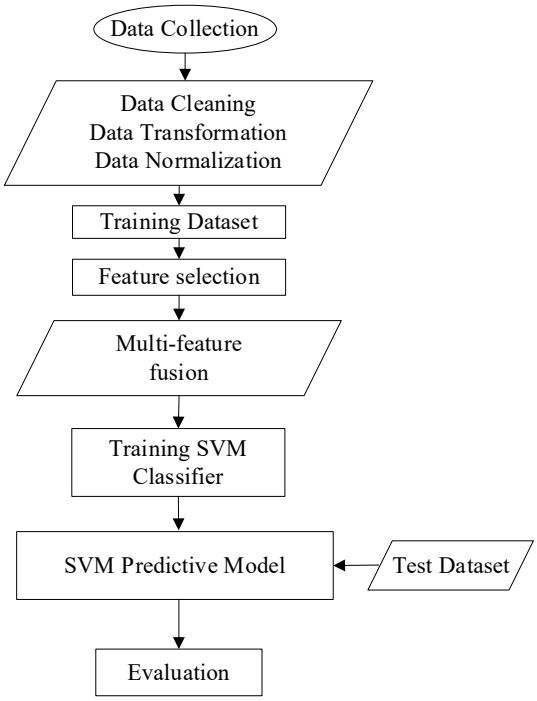

Figure 2. The Proposed Predictive Model of College Students Based on Weighted Multi-feature and SVM

The proposed model for predicting dropout involves several steps, which are described below:

Step 1: Collecting the raw data.

In this step, the raw data of students' online learning behaviors are collected from the Chinese university MOOC.

Step 2: Preprocessing the raw data collected in Step 1 by cleaning, transforming and normalizing.

The raw data collected in Step 1 may contain irrelevant, inconsistent, or incomplete information, which could negatively affect the prediction accuracy of the model. Therefore, in this step, the data is preprocessed by cleaning, transformation and normalization techniques to improve the quality of the data. "Max-Min" normalization [22] is used in this work, it is described as in formula (12):

$$y_i = \frac{x_i - \min_{1 \le i < n}(x_i)}{\max_{1 \le i < n}(x_i) - \min_{1 \le i < n}(x_i)} \tag{12}$$

Where x_i stands for any feature of a kind, y_i stands for the normalized values of x_i, n is the number of x_i.

Step 3: Selection of the appropriate online behavior features based on the correlation coefficient

In this step, the online behavior features that have significant correlation with dropout behavior are selected from the preprocessed dataset. For this purpose, the correlation coefficient of each feature with the dropout behavior is analyzed, and the features that have a higher Pearson correlation coefficient are selected from the training dataset.

Step 4: Weighting each feature according to the ratio of correlation coefficients and merging all selected behavior features.

In this step, each selected feature is assigned a weight based on its correlation coefficient with dropout behavior. The features are then merged to obtain a dataset weighted by multiple features.

Step 5: Training the SVM classifier with training dataset.

In this step, the multi-feature weighted dataset is divided into a training dataset and a test dataset randomly. The training data accounted for 80% of the entire dataset. The SVM classifier is then trained with 80% of the training dataset.

Step 6: Evaluating the proposed predictive model with the test dataset. Obtain the trained SVM classifier.

After the SVM classifier is trained, it can be used to predict the dropout of college students. In this step, the performance of the proposed predictive model is evaluated with the test dataset. To evaluate the performance of the model, the Accuracy, Precision, Recall and $F-measure$ are calculated. If the performance of the model is satisfactory, it can be used to predict the dropout behavior of students.

C. Data Collection

There are several types of behavioral data that are collected in MOOCs. These numerous data not only capture students' behavioral patterns, but also serve as a valuable source for analyzing and predicting dropout. It is worth noting that different types of behavioral data have different effects on predicting dropout, as stated by Pearson's principle [23]. This principle shown in formula (13) states that the correlation between two variables is stronger when the two variables are related by a linear relationship. Therefore, the selection of appropriate behavioral data is critical to the development of an accurate model for predicting dropout:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{13}$$

where $x_1, x_2, ..., x_i$ and $y_1, y_2, ..., y_i$ are the measured values of both variables, $\bar{x}$ and $\bar{y}$ are the arithmetic means of both variables. Pearson correlation coefficient r is used to measure the linear correlation between two variables x and y, and its value is between -1 and 1, the closer it is to 1, the greater the correlation between the two variables.

In this work, we collected 11 different types of online learning behavior characteristics to analyze their correlation with dropout. These characteristics include duration of online learning, number of views of announcements, number of views of assessment criteria, number of views of videos, number of views of instructional texts, number of comments on topic posts, number of publications of topic posts, number of views of topic posts, number of participations in quizzes, number of submissions of homework, and number of visits to the MOOC. To evaluate the correlation between each behavioral characteristic and dropout, we calculated the correlation coefficient and created a heatmap using MATLAB. The heatmap is shown in Figure 3.

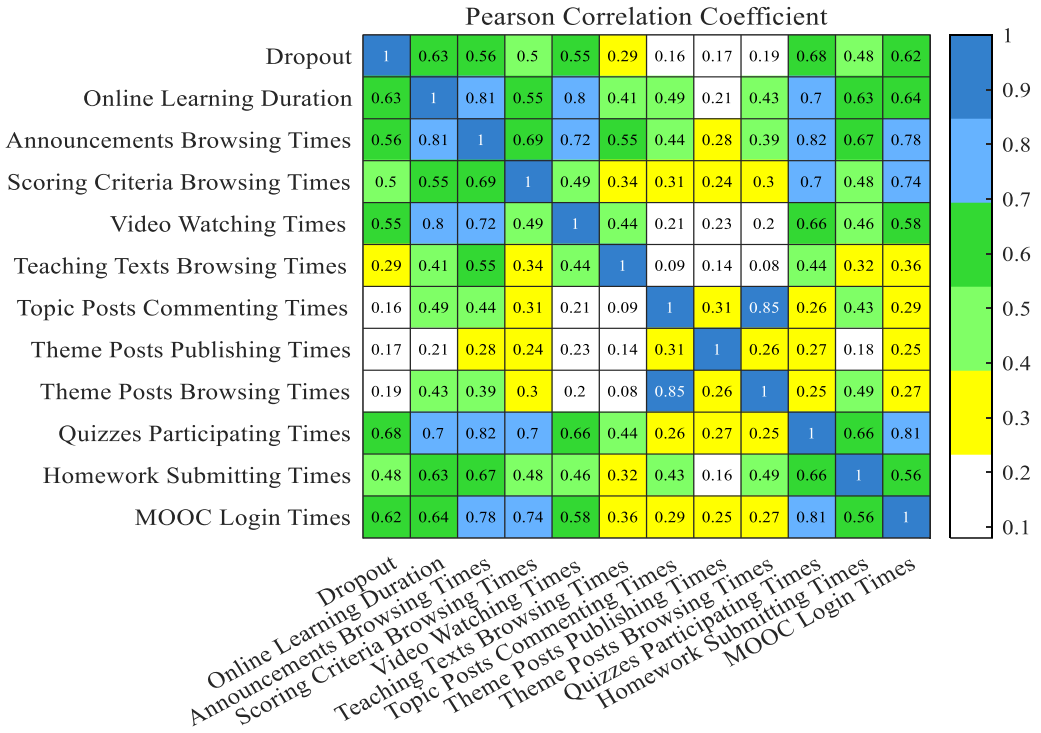

Figure 3. Correlation Coefficient Heatmap

A correlation coefficient greater than 0.6 generally indicates a strong relationship between two variables. Based on the heat map analysis, the duration of online learning, quiz participation times, and MOOC registration times were found to have the strongest influence on MOOC dropout. The 11 characteristics of online learning behavior were collected from two datasets respectively.

Dataset 1 was collected from the platform of Chinese University MOOC. These 11 kinds of characteristics were extracted from eight MOOCs offered by the platform in 2022. Table 2 provides details of the courses and class sizes.

Table 2. Courses and Class Size of Dataset 1

Serial Number	Course Name	Class Size
1	Introduction of College Computer Science	211
2	Python Programming	180
3	The Basis of Computer Network	172
4	Data Structures	135
5	Color Science of Art Design	189
6	Introducing Child Psychology	177
7	Ancient Chinese Literary History	155
8	Ancient Chinese Philosophy of Music	163

From Table 2, it can be seen that the courses selected for this database are from different disciplines, and the class sizes are small and medium. In reality, the class sizes of online courses in many universities are not always large-scale, therefore, such sample selection can make the dropout prediction model proposed in this study accurately fit to the ordinary online courses of colleges.

Dataset 2 is KDD CUP 2015 [24], a publicly available database from XuetangX (the first MOOC platform in China). KDD CUP 2015 is widely used for research on predicting course dropout in MOOCs. It provides 30-day behavioral data and dropout labels in the fourth 10-day period for 39 courses with about 1,000 participants of each course. It is a dataset with more courses and larger class size than dataset 1, which is of great value for studying dropout prediction in MOOCs with large class size.

Dataset 1 and dataset 2 are derived from two different famous MOOCs platforms in China, and have different sample sizes. In order to comprehensively test the performance of the proposed dropout prediction model with different sample sizes and better compare other dropout models, two different types of datasets are used in this research.

D. Evaluation Criterion

In this work, we adopt the Accuracy shown in formula (14), the Precision shown in formula (15), the Recall shown in formula (16) and the $F-measure$ shown in formula (17) as the components of the evaluation matrix, which are the common evaluation criteria for supervised machine learning classifiers [19, 25].

$$A = \frac{TP + TN}{TP + FN + FP + TN} \tag{14}$$

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$F = 2 * \frac{P * R}{P + R} \tag{17}$$

Where A stands for Accuracy, P stands for Precision, R stands for Recall and F stands for $F-measure$.

TP represents True Positive, which refers to the number of samples that are accurately predicted as positive by the model; TN represents True Negative, which represents the number of samples that the model accurately predicts as negative; FP represents False Positive, which indicates the number of samples that are incorrectly predicted as positive by the model; FN represents False Negative, which signifies the number of samples that are incorrectly predicted as negative by the model.

Normally, the greater the value of Accuracy, Precision, Recall and *F*-measure are, the better the predictive model performs. In this work, the dropout prediction models are evaluated based on these metrics.

IV. RESULTS AND DISCUSSIONS

To improve the accuracy of models predicting school dropout based on multiple characteristics, it is important to consider the differential contributions of each behavioral characteristic. Conventional models often assume that each characteristic is equally weighted, but heat map analysis shows otherwise. Therefore, in this research, we propose a novel approach to integrate multiple features by assigning different weights based on the Pearson correlation coefficient. Specifically, we introduce a dropout prediction model based on multiple features fusion and SVM to improve predictive performance.

A. Proposed Behavior Weighted Muti-feature Fusion Algorithm

The dropout prediction model proposed in this research weights each behavioral trait according to its contribution to predictive outcomes, improving upon existing relevant models that assign equal weight to each behavioral trait by default. This differentiation in weighting results in higher predictive accuracy. To achieve this, the model uses a special algorithm to fuse the weights of behavior features, which is described below:

Step 1: Collect the raw data and process the data, obtain all kinds of behavioral data that has a likelihood to be used in the dropout prediction model. Define any one of the behavior features as a_i, $i \in N^+$, Set the obtained n

kinds of behavior feature as an *n*-dimensional vector $A=[a_1,\ a_2,\ ...,\ a_n],\ n\in N^+$. Where is N^+ a positive integer.

Step 2: Calculate the Pearson correlation coefficient between each behavior feature $a_i \in A,\ i\in n,\ n\in N^+$ and dropout. Define the correlation coefficient of behavior feature a_i as $r_i,\ i\in N^+$, then get the *n*-dimensional correlation coefficient vector $R=[r_1,\ r_2,\ ...,\ r_n],\ n\in N^+$.

Step 3: Select the appropriate behavior feature for the predictive model from R. For any one of R, if $r_i \geq \lambda,\ i\in n,\ n\in N^+,\ |\lambda|\leq 1$. it is believed that the corresponding behavior feature a_i have a strong correlation with dropout and this behavior feature a_i is the chosen one for the predictive model. Then, define any one of the behavior feature conforming to the formula as $a_i',\ i\in N^+$, the number of all the selected behavior features as $m,\ m\leq n, n\in N^+$. And then, combine all the chosen behavior features into an m-dimensional vector $A'=[a_1',\ a_2',\ ...,\ a_m'],\ m\leq n,\ m,n\in N^+$.

Step 4: Calculate the weight of each selected behavior feature a_i'. Define $w_i,\ i\leq m,\ i,\ m\in N^+$ as the weight of the chosen behavior feature $a_i'\in A',\ i\leq m,\ i,\ m\in N^+$. w_i can be defined in formula (18):

$$\begin{aligned} w_i &= \frac{r_i'}{\sum_{i=1}^{m} r_i'},\ i\leq m,\ i,m\in N^+ \\ \sum_{i=1}^{m} w_i &= 1,\ i\leq m,\ i,m\in N^+ \end{aligned} \tag{18}$$

Where r_i' is the Pearson correlation coefficient of a_i'.

Step 5: Feature fusion. Defined B as the integrated behavior feature. It is explained in formula 19:

$$\begin{aligned} B &= \sum_{i=1}^{m} w_i a_i',\ i\leq m,\ i,m\in N^+ \\ \sum_{i=1}^{m} w_i &= 1,\ i\leq m,\ i,m\in N^+ \end{aligned} \tag{19}$$

Where w_i is the corresponding weight of the selected feature a_i'

B. Analysis of Experimental Results

In order to verify the performance of the proposed model (described in *B. The Design of the Dropout Prediction Model* from III RESERCH METHODOLOGY), 8 courses (shown in Table 2) and three behavior features (online learning duration, quizzes participating times and MOOC login times.) with a correlation coefficient greater than $\lambda=0.6$ were used as data sources. Dataset 1 was collected from Chinese University MOOC. 88972 pieces of behavioral data obtained after raw data with cleaning, transformation and normalization. In an experiment, the dataset was randomly divided into a training set and a test set with the ratio 4 to 1.

The behavioral logs of MOOCs of college students are defined as $D=\{D_1,\ D_2\}$, where D_1 is the subset for the logs of 30 days, D_2 is the logs of the fourth 10 days. A student is considered as dropout if he or she does not have any behavioral record on MOOCs for the fourth 10 consecutive days.

The predictive status of college students is either retention or dropout, the coding of the status is shown as Table 3.

Table 3. Status of The Predictive Model and Coding

Predictive results	Classification samples	Coding
Retention	Positive sample	1
Dropout	Negative sample	0

MATLAB was used to simulate this predictive model. In the experiment 1, the proposed dropout prediction model was compared with the model based on behavior feature given the same weight and SVM in dataset 1. The results of experiment 1 are shown in Figure 4.

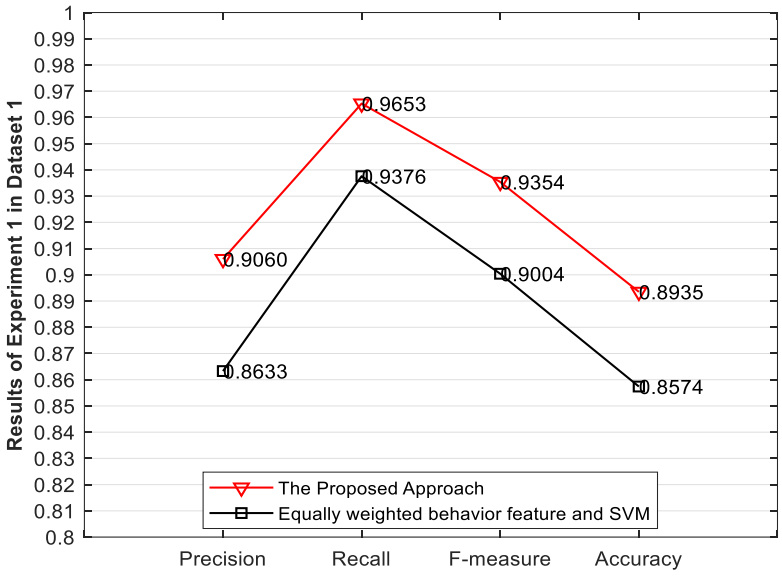

Figure 4. Experiment 1 Results

Generally, the closer the value of the evaluation index is to 1, the better the classification result will be, the ideal value is 1. From Figure 4, it can be seen the proposed model is superior to the comparison model in every evaluation index. The proposed approach achieves Precision, Recall, $F-measure$, and Accuracy of 0.9060, 0.9653, 0.9354, and 0.8935, respectively, in dataset 1. It can be concluded that the weights of the behavior features have a significant impact on the dropout prediction results of college students based on the same dataset and the same behavior characteristics.

In the experiment 2, the performance of the proposed prediction model based on multi-feature fusion and SVM was tested using dataset 2. The results of the comparison between the proposed model and the baseline (SVM model in [10]) for dataset 2 are shown in Figure 5.

Figure 5 shows that the proposed approach achieves Precision, Recall, $F-measure$, and Accuracy of 0.8990, 0.9627, 0.9311, and 0.8885, respectively, in dataset 2. It can be seen that the proposed model performs better than the baseline SVM models in terms of Precision, F-measure, and Accuracy, except for the slightly higher Recall value compared to the benchmark model. The SVM-based dropout predictive model proposed by Wen et al. [10] use the same dataset and the same time range of data extraction as in this study. This SVM model [10] adopts seven types of behavior feature, and the weights of each feature are not considered, i.e., the weights of each feature are set to the same by default. This is the main difference from the SVM dropout predictive model proposed in this study.

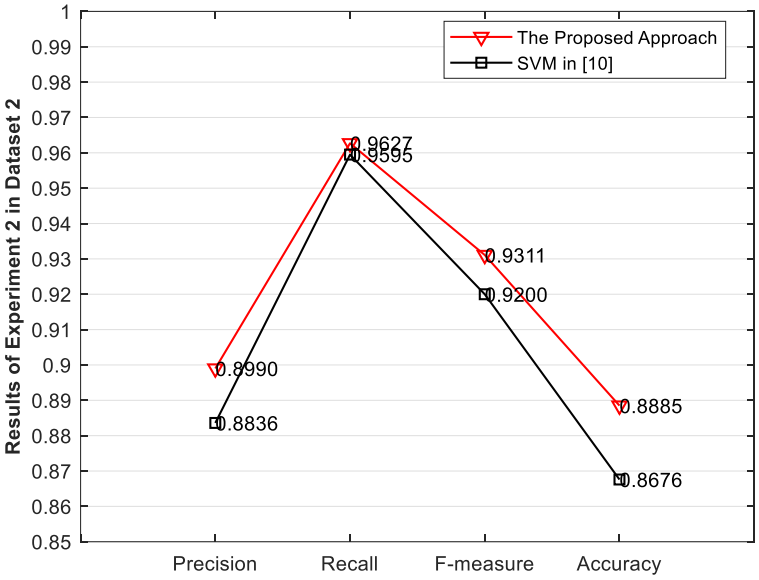

Figure 5. Experiment 2 results

In the experiment 3, to test whether the proposed dropout prediction model outperforms other predictive models with behavior features of same weights, we compared other predictive models [10] based on the same 30-day data in the same dataset KDD CUP 2015. The results are shown in Table 4.

Table 4. Experiment 3 Results

Predictive Model	Precision	Recall	F-measure	Accuracy
Regression Tree (CART)	0.8807	0.8868	0.8837	0.8150
Naive Bayes (NB)	0.8813	0.9247	0.9024	0.8414
Gradient Boosted Decision Tree (GBDT)	0.8892	0.9626	0.9244	0.8751
Linear Discriminant Analysis (LDA)	0.8597	**0.9771**	0.9147	0.8555
Logical Regression (LR)	0.8736	0.9675	0.9181	0.8632
Random Forest (RF)	0.8985	0.9446	0.9209	0.8714
Convolutional Neural Network (CNN)	0.8938	0.9579	0.9247	0.8764
The Proposed Approach	**0.8990**	0.9627	**0.9311**	**0.8885**

Table 4 lists the evaluation results of the various predictive models, including their Precision, Recall, F-measure, and Accuracy. These metrics are commonly used to evaluate the performance of classification models in machine learning. According to the experimental data, the proposed model is effective in predicting MOOC dropout among college students in a large database. The dropout prediction model proposed in this research, which combines weighted behavior features and employs the SVM classifier, achieves Precision, F-measure, and Accuracy of 0.8990, 0.9311, and 0.8885 respectively. LDA algorithm achieves the highest Recall, closely followed by LR algorithm and ours (SVM). These models have a high ability to capture actual positive samples (retention status) correctly. From the evaluation matrix, it can be seen although the value of Recall comes in the second, the other evaluation indexes of this predictive model outperform several other machine learning models proposed by Wen et al. [10].

V. CONCLUSION

In this paper, we briefly discussed the dropout problem of MOOCs, reviewed the existing predictive models of MOOCs and pointed out their common shortcoming. Then, we proposed a model for predicting MOOC abandonment based on the fusion of behavior features and SVM algorithm. This model addresses the limitation of the existing prediction model that assigns equal weight to all behavior features. Empirical results on both datasets demonstrate the effectiveness of the proposed model.

In the first experiment, the proposed dropout prediction model was conducted on dataset 1 which is a small and medium-sized dataset. The proposed approach achieves Precision, Recall, $F-measure$, and Accuracy of 0.9060, 0.9653, 0.9354, and 0.8935, respectively, which is the best performance achieved among the three experiments. It shows that the weighted multi-feature fusion and SVM-based method proposed in this study is more effective for predicting dropout of college students in MOOCs for data with a small and medium sample size.

In the second experiment, the proposed dropout prediction model was run on dataset 2 (KDD Cup 2015) which is a widely used dataset with a large sample size. The model for comparison is also SVM-based but with equal weights of all features. Finally, in the same dataset, the proposed model achieves Precision, Recall, $F-measure$, and Accuracy of 0.8990, 0.9627, 0.9311, and 0.8885, respectively, which is higher in all respects. It is proved that behavior weighted muti-feature fusion algorithm proposed in this study has a significant influence on predicting dropout in MOOCs.

In the third experiment, the proposed dropout prediction model was compared with other seven basic models proposed by Wen et al. in dataset 2. Our model achieves Precision, F-measure, and Accuracy of 0.8990, 0.9311, and 0.8885 respectively. All metrics outperform other models except the value of Recall (0.9627), which ranks second. The models used for comparison use equal weights in processing different behavior features. According to the benchmarking, the proposed model improves the accuracy of dropout prediction models and provides a more comprehensive understanding of the factors contributing to dropout rates. By including multiple behavioral characteristics and assigning appropriate weights, the proposed model provides a more nuanced understanding of the complex relationship between online learning behaviors and dropout rates. Meanwhile, the experimental results also show that the proposed model can perform well even with large-scale samples.

In this paper, we use data from the first 30 days of courses to build a dropout prediction model, which is important for early dropout prediction in MOOCs. Instructors can thus identify early if students tend to drop out and intervene to prevent them from dropping out. The future research will focus more on predicting academic performance of college students in MOOCs.

ACKNOWLEDGEMENT

We would like to gratefully acknowledge all those who have made valuable comments on and support to this article. This work received no funding from any party for the research and publication of this article.

REFERENCES

[1] H. Haron, A. R. M. Yusof, H. Samad, N. Ismail, A. Juanita and H. Yusof, "The platform of MOOC (Massive Open Online Course) on open learning: Issues and challenges," International Journal of Modern Education, vol. 1, pp. 1-9, 2019.

[2] C. M. Stracke, S. Downes, G. Conole, D. Burgos and F. Nascimbeni, "Are MOOCs Open Educational Resources? A Literature Review on History, Definitions and Typologies of OER and MOOCs", Open Praxis, vol. 11, pp. 331-341, 2019.

[3] N. Alhazzani, "MOOC's impact on higher education", Social Sciences & Humanities Open, vol. 2, 100030, 2020.

[4] J. Kennedy, "Characteristics of massive open online courses (MOOCs): A research review," Journal of Interactive Online Learning, vol. 13, pp. 1-16, 2014.

[5] Y. Goel and R. Goyal, "On the effectiveness of self-training in MOOC dropout prediction", Open Computer Science, vol. 10, pp. 246-258, 2020.

[6] E. H. Othman, S. Abdelali and E.B. Jaber, "Education data mining: mining MOOCs videos using metadata based approach", IEEE International Colloquium on Information Science and Technology (CiSt), pp. 531-534, 2016.

[7] A. Ezen-Can, K. E. Boyer, S. Kellogg and S. Booth, "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach", International Conference on Learning Analytics and Knowledge, pp. 146-150, 2015.

[8] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang and S. Chen, "MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine", Mathematical Problems in Engineering, vol. 2019, pp. 1-11, 2019.

[9] V. Muthukumar and N. Bhalaji, "MOOCVERSITY-deep learning based dropout prediction in MOOCs over weeks," Journal of Soft Computing Paradigm (JSCP), vol. 2, pp. 140-152, 2020.

[10] Y. Wen, Y. Tian, B. Wen, Q. Zhou, G. Cai and S. Liu, "Consideration of the Local Correlation of Learning Behaviors to Predict Dropouts from MOOCs", Tsinghua Science and Technology, vol. 25, pp. 336-347, 2020.

[11] Q. Fu, Z. Gao, J. Zhou and Y. Zheng, "CLSA: A novel deep learning model for MOOC dropout prediction", Computers & Electrical Engineering, vol. 94, pp. 1-12, 2021.

[12] I. Nitta, R. Ishizaki, M. Shingu, S. Nakashima, K. Maruhashi, A. Tolmachev and M. Todoriki, "Graph-based massive open online course (MOOC) dropout prediction using clickstream data in virtual learning environment", International Conference on Computer Science & Education (ICCSE), 2021, pp. 48-52.

[13] S. Nithya and S. Umarani, "MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features," International Journal of Advanced Computer Science and Applications, vol.13, no. 9, pp. 607-617, 2022.

[14] M. T. Sultan, H. El Sayed, M. A. Khan and M. Abduljabar, "A Deep Learning Model for MOOC Dropout Prediction Using Learner's Course-relevant Activities", IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT), 2022, pp. 13-18.

[15] S. Sun, X. Qian, L. Mu, H. Zan and Q. Zhang, "Performance Prediction Based on Analysis of Learning Behavior", Data Science: International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE), 2018, pp. 632-644.

[16] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression", Neural networks, vol. 17, pp. 113-126, 2004.

[17] Y. Lim, K. W. Ng, P. Naveen and S. C. Haw, "Emotion Recognition by Facial Expression and Voice: Review and Analysis," Journal of Informatics and Web Engineering, vol. 1, no. 2, pp. 45-54. 2022.

[18] F. F. Chua, T. Y. Lim, B. Tajuddin and A. P. Yanuarifiani, "Incorporating Semi-Automated Approach for Effective Software Requirements Prioritization: A Framework Design," Journal of Informatics and Web Engineering, vol. 1, pp. 1-15, 2022.

[19] M. Shafiq, H. Ng, T. T. V. Yap and V. T. Goh, "Performance of Sentiment Classifiers on Tweets of Different Clothing Brands," Journal of Informatics and Web Engineering, vol. 1, pp. 16-22, 2022.

[20] J. Zhao, "The development and application of support vector machine," Journal of Physics: Conference Series, vol. 1748, no. 5, pp. 052006, 2021.

[21] R. S. Chauhan and D. Ghosh, "An erratum to 'Extended Karush-Kuhn-Tucker condition for constrained interval optimization problems and its application in support vector machines'", Information Sciences, vol. 559, pp. 309-313, 2021.

[22] B. Hong, Z. Wei and Y. Yang, "Discovering learning behavior patterns to predict dropout in MOOC", 2017 12th International Conference on Computer Science and Education (ICCSE), 2017, pp. 700-704, doi: 10.1109/ICCSE.2017.8085583.

[23] B. Wu, "Influence of MOOC learners discussion forum social interactions on online reviews of MOOC", Education and Information Technologies, vol. 26, pp. 3483-3496, 2021.

[24] S. Nithya and S. Umarani, "Comparative Analysis of the Learning on KDD Cup 2015 Dataset", Webology, vol. 19, pp. 705-717, 2022.

[25] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, pp. 37-63, 2011.