International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE **Open Access**

Dropout prediction and decision feedback supported by multi temporal sequences of learning behavior in MOOCs

Xiaona Xia[1,2*] and Wanxue Qi[1,2]

*Correspondence:
xiaxn@qfnu.edu.cn; xiaxn@sina.com

[1] Faculty of Education, Qufu Normal University, Qufu 273165, Shandong, China
[2] Chinese Academy of Education Big Data, Qufu Normal University, Qufu 273165, Shandong, China

Abstract

The temporal sequence of learning behavior is multidimensional and continuous in MOOCs. On the one hand, it supports personalized learning methods, achieves flexible time and space. On the other hand, it also makes MOOCs produce a large number of dropouts and incomplete learning behaviors. Dropout prediction and decision feedback have become an important issue of MOOCs. This study carries out sufficient method design and decision analysis on the dropout trend. Based on a large number of learning behavior instances, we construct a multi behavior type association framework, design dropout prediction model to analyze the temporal sequence of learning behavior, then discuss the corresponding intervention measures, in order to provide adaptive monitoring mechanism for long-term tracking and short-term learning method selection, and enable adaptive decision feedback. the full experiment shows that the designed model might improve the performance of the dropout prediction, which achieves the reliability and feasibility. The whole research can provide key technical solution and decision, which has important theoretical and practical value for dropout research of MOOCs.

Keywords: MOOCs, Learning behavior, Multi temporal sequence, Dropout trend, Dropout prediction model, Learning analytics

Introduction

The continuous improvement and full application of MOOCs (Massive Open Online Courses, MOOC) has fundamentally changed the learning process. Learners can flexibly access online learning resources according to their own learning needs, and are no longer restricted by time, space and regional teachers. MOOCs support the personalization and autonomy of the learning process to a certain extent (Borrella et al., 2022; Xia & Wang, 2022). MOOCs, as the platforms that primarily manage learning behavior, will not fully support learners' social behavior. However, in order to facilitate knowledge exchange and problem feedback between learners, or between learners and instructors, some MOOCs platforms have set up corresponding forums, Q&A and other means to help learners build relevant interactive behaviors. Learners can submit their own requests in real-time and wait for relevant feedback. Based on this communication, identify learners' confusion and risks. In addition, a large number of learning behavior instances have been generated, and the corresponding attributes are becoming more and more abundant, and the characteristics will become more diverse and complex. The online learning mode supported by big data technology makes it possible to track and predict the learning behavior (Anghel et al., 2022). How to realize the continuous and complete description from massive data, how to give full play to the data value and effectively enable education and teaching, and how to realize the associated calculation and analysis have become hot topics (Xia, 2021a). This issue raises a key problem, that is needed to define learning behavior as a continuous temporal sequence, which contains both the long-term evolution trend of learning behavior and the short-term fluctuation characteristics, that are constantly changing throughout the temporal sequence (Xia & Qi, 2022a).

At present, the dropout prediction method is only applicable to a single time node and interval (Khoushehgir & Sulaimani, 2023). It cannot achieve continuous tracking and feedback of temporal sequence, and it is difficult to effectively integrate the long-term evolution trend and short-term fluctuation characteristics. The scale of computable attributes is limited, and it is easy to lose key data, It has seriously affected the reliability and credibility of dropout prediction (Mourdi et al., 2022). At the same time, learners are vulnerable to interference from a variety of factors, and the interference makes the learning behavior contain a lot of noise and increases the difficulty of dropout prediction. Because it is necessary to mine key data, it is important to implement effective cleaning and standardization (Ashenafi et al., 2022). In order to accurately and efficiently track and predict the dropout trend, we might consider the characteristics of temporal sequence at different time nodes and intervals, analyze the impact of noise on the associated data, derive the key time nodes and intervention intervals (Xia, 2020a), improve the effectiveness of learning behavior and optimize learning results.

This study will carry out in-depth analysis on the dropout trend supported by multi temporal sequences. Based on the massive learning behavior instances of MOOCs, a multi learning behavior feature association framework is constructed. Multiple parallel full-convolution neural networks are used to mine multi-type features from the multi temporal sequences, and the variational information bottleneck is used to filter all the noises irrelevant to the dropout trend, effectively reduce the impact of noise or interference on the dropout prediction. It is needed to realize the accuracy of dropout prediction and the reliability of decision feedback, and provide feasible strategies for the self-organization and self-regulation of MOOCs.

Related work

Different from the traditional education and teaching mode, MOOCs realize the support of the complete learning process and the collection of relevant data. Through the analysis and modeling of data and relationships (Xia, 2021b), it can effectively restore the history teaching process and learning behavior. The data involves the click rate of learning resources, homework submission information, project collaboration information, performance evaluation results, forum data, collaboration relationship, etc., which plays an important role in the study of dropout prediction.

Through the mining, analysis and prediction of historical learning behavior, we can understand the relationships between learning state and learning behavior, and obtain the key factors that affect learning motivation and interest preferences (Kim et al., 2017). So we can predict and evaluate whether learners may have negative emotions or even give up learning in a certain period. So the accurate prediction of learning behavior is the key issue in exploring the dropout trend. The effective machine learning models and algorithms have been gradually adopted by researchers (Anttila et al., 2022), mainly including the following three aspects:

(1) Prediction method of dropout probability. This method needs to transform the dropout problem into the learning behavior classification, and construct the classification model by logical regression, support vector machine and decision tree (Ghada et al, 2016). it is needed to track the registration characteristics and learning behavior characteristics, complete the collaborative analysis, build the dropout prediction model by the improved decision tree, so as to predict the dropout probability of learners, and divide it into different levels to formulate different intervention measures (Rodr í guez et al., 2023).
(2) Method for locating dropout factors. This method also needs to track the continuous temporal sequence of learning behavior, and mine the key factors that may cause dropout (Xia, 2021c). It is needed to collect the individual characteristics, interactive and collaborative activities, learning contents and survey data related to learning behavior, explore the key factors that lead to the dropout trend by machine learning methods (such as random forest), count and test the learning behavior differences between dropouts and non-dropouts, and explore the key time nodes or intervals (Gubbels et al., 2019).
(3) Dropout tracking method supported by feature engineering. It is needed to mine the effective features of learning behavior, that will have a significant impact on the data analysis (Chanaa & Faddouli, 2022). Some researchers have proposed some dropout tracking methods supported by feature engineering to identify effective multi features, then achieve correlation analysis between features. Researchers might identify the problems related to the dropout prediction, mine the effective input feature sets through different feature engineering methods, realize adaptive fusion of comprehensive features (Xia, 2020b), take the temporal sequence of learning behavior as the timeline, track the learning behavior change process, find out the key behavior types and implement corresponding intervention.

In fact, the above three aspects are interrelated. The dropout trend is not a sudden decision of learners. It is related to the dynamic change of learning behavior and the continuous evolution of temporal sequence, as well as many factors, attributes and characteristics, and has become an important behavioral profile of multi data and complex relationships (Rodríguez et al., 2023). Traditional machine learning methods promote small-scale static data research, but they are not applicable to large-scale learning behavior instances. When temporal sequence is defined as the dynamic characteristics, and the correlation analysis of multi features, attributes and relationships is realized, it has become the key issue of dropout trend. So we might explore full innovative analysis and demonstration of relevant models and algorithms (Hsu, 2022).

However, the research on dropout prediction enabled by massive data is constrained by the structure, characteristics, attributes, and relationships of data itself. Relevant research strategies need to design appropriate methods and techniques suitable for data analysis. In the process of explicit or implicit data analysis, it is necessary to identify dropout risks in the learning process, which come from learners themselves, as well as the management model of learning resources, the interaction mode of the learning process, the presentation form of the learning content, and the learning tasks etc., that are all directly related. This conclusion not only provides timely intervention and early warning for learning behavior, but also provides key decision for the online learning process organization, learning resource recommendation, and learning behavior optimization of MOOCs.

Data description and key problems

Of course, it is possible to classify learners based on their learning background and demographic information, and divide them into potential dropouts or non dropouts in advance. However, this can only be a qualitative study, and cannot represent the results and effects of learners' current learning process. At the same time, the learning process is complex, and it is not possible to directly mark learners as dropouts or non dropouts based on historical data. Moreover, the learning process is dynamic, and the learning intent can also jump out of the original data profile, which can also affect learning enthusiasm due to the organizational model of online learning resources. Therefore, for massive data driven dropout prediction, it should be an unsupervised learning process, tracking the entire learning process through complete temporal sequence, rather than directly completing the classification of learners in advance.

To effectively achieve the dropout prediction and decision feedback of MOOCs, we might select the complete data generated by MOOCs in a certain period. At the same time, in order to compare different methods and realize the application and optimization of relevant research conclusions, the selected data needs to be desensitized. This study focuses on one MOOC platform "XuetangX", we mine a large number of related learning behavior instances, and form "MOOC Dataset", which includes sufficient course selection logs and learning behavior records.

MOOC Dateset contains 79186 learners. The interaction between learners and 39 courses forms selection logs and learning behavior instances with temporal characteristics. The selection logs describe the learning trend, and also construct mutual mapping and correlation with learning behavior. Each record mainly includes learners' ID "LID", course ID "CID", start time "StartTime", end time "EndTime", URL, behavior type "BehaviorType", behavior trend "BehaviorTrend" and others, that have generated 8157277 records in total. These records involve seven learning behavior types, as shown in Table 1. These seven learning behavior types summarize the learning process, that have undergone sufficient statistics and analysis of all learning behavior instances, are also the main way for learners to participate in this learning platform, as well as can reflect the different behavioral characteristics of learners. Among them, it relates to the main behavioral trends of learners throughout the learning process, but there are

Table 1 Learning behavior type

Learning behavior type	Description
Question and Answer(QA)	Learners answer relevant questions set in the learning process.
Watching Video(WV)	Learners watch the videos related to some course.
Access Course(AC)	Learners access some course.
Access Wiki(AW)	Learners access the corresponding wiki related some course.
Forum Discussion(FD)	Learners discuss some topics in Forum.
View Contents(VC)	Learners view the associated contents of some course.
Close Pages(CP)	Learners close the pages that show the contents of some course.

differences in the association between different learners' behavioral types, some resulting in dropouts and some not. This is extremely meaningful for data driven dropout prediction.

In this study ,we will divide the entire learners into two groups based on the predictive result of learning behavior: non dropouts and dropouts. When a learner has a learning behavior record within 30 consecutive days and has a learning behavior record in any subsequent 10 consecutive days, it is defined as not dropout and labeled as 0, otherwise labeled as 1. However, there are still some data without labels, which can not effectively distinguish whether there are dropouts, and can only analyze the records in the first 30 days. In order to make full use of the characteristics of learning behavior in different intervals, the data generated by the seven learning behavior types every day is counted. The nonexistent behavior types are labeled as 0, so that the learning behavior can be transformed into a matrix, and the temporal sequence of a learner for 30 days can be expressed as $lh = \begin{bmatrix} lh_{1,1} & lh_{1,2} & \cdots & lh_{1,7} \\ lh_{2,1} & lh_{2,2} & \cdots & lh_{2,7} \\ \cdots & \cdots & \cdots & \cdots \\ lh_{30,1} & lh_{30,2} & \cdots & lh_{30,7} \end{bmatrix}$, so that the 30-day learning behaviors can be described as a vector $LH = \left[lh^1, lh^2, \cdots, lh^N\right]$, N represents the number of learners. The complete sequence can be described by the rows represented by the extended matrix lh. But not all learners have the same days in a complete learning period, the length of temporal sequence is inconsistent, so it is necessary to classify the temporal sequence of learning behavior.

We take the course as the classification condition, take the assessment time as the maximum sequence length, that is defined as T_{CID}, then the complete learning behavior of each learner of a course is described as a matrix $flh = \begin{bmatrix} lh_{1,1} & lh_{1,2} & \cdots & lh_{1,7} \\ lh_{2,1} & lh_{2,2} & \cdots & lh_{2,7} \\ \cdots & \cdots & \cdots & \cdots \\ lh_{T_{CID},1} & lh_{T_{CID},2} & \cdots & lh_{T_{CID},7} \end{bmatrix}$, and all the complete learning behavior involved in the whole data set can be expressed as $FLH = \left[flh^1, flh^2, \cdots, flh^M\right]$, M is the number of courses, LH can be used to distinguish whether learners have dropout trend, and then divide learning behavior into two parts: dropout behavior and non-dropout behavior. FLH can achieve a complete description and compare the learning trends based on the course, mine the relevant influencing factors, and explore the causes of dropout trend.

Based on the seven learning behavior types in Table 1, as well as the method of learning behavior representation, this study puts forward corresponding problems for dropout prediction and decision feedback supported by multi temporal sequences. Through preliminary data statistics and analysis, it is found that 98.950% of learners regard accessing courses (AC) as the first type of learning behavior during the entire learning process. That is, in the temporal occurrence ranking of each learner, AC is usually the earliest learning behavior type, and each learner also indicates the end of a certain learning behavior type with a closing page (CP). However, 1.050% of learners do not form statistically significant temporal sequence of learning behavior, which cannot be used as a test sample for continuous learning behavior instances. meanwhile, during the analysis of learning behavior path, it was also observed that the routing strategies for the other five learning behavior types are constrained by one learning behavior type that is directly associated with AC. Therefore, in the design of related test problems, the path of the entire learning behavior is defined that AC is as the starting node, CP is as the ending node, and AC is associated with one of the other learning behavior types, the intrinsic relationships of the remaining four learning behavior types as key factors in routing analysis. Whether they are the dropouts or not, the path of their learning behavior will be directly related to these seven types and the characteristics of their relationships, as shown in Fig. 1.

Figure 1 describes the potential relationships between the seven learning behavior types. so the seven learning behavior types can realize mutual correlation and cooperation, thus forming the related description of learning behavior. With AC as the starting node and CP as the ending node, the test results are whether the path has produced the dropout trend, and the factors that cause the dropout trend are explored, so there are six problems are defined, namely:

P1. The learning behavior path formed by AC→CP can lead to dropout tend. It mainly tests the impact of course content on the dropout trend.

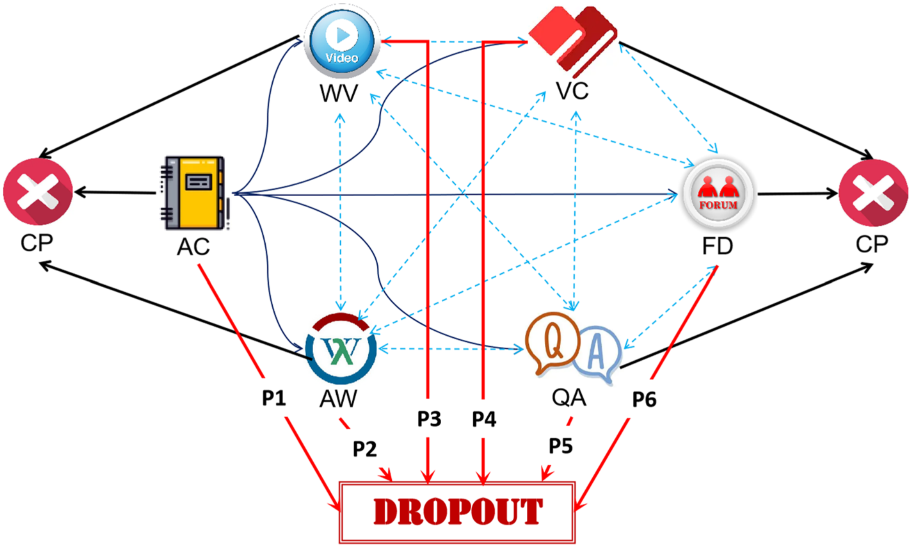

Fig. 1 Seven learning behavior types and six problems

P2. The learning behavior path formed by AC→AW→(WV, VC, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of four learning behavior types (WV, VC, QA, FD).

P3. The learning behavior path formed by AC→WV→(AW, VC, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (AW, VC, QA, FD).

P4. The learning behavior path formed by AC→VC→(AW, WV, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (AW, WV, QA, FD).

P5. The learning behavior path formed by AC→QA→(WV, AW, WV, FD) →CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (WV, AW, WV, FD).

P6. The learning behavior path formed by AC→FD→(WV, AW, WV, QA)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (WV, AW, WV, QA).

Methodology

Research design

Based on the multiple learning behavior types and related test problems of MOOCs dropout prediction, the construction of the entire research design is divided into three steps:

Firstly, locate a complete learning period according to seven learning behavior types, mining relevant instances along a temporal order, using learners as clues, define relevant features and relationships, and forming an initial dataset;

Secondly, analyze the attributes and distribution characteristics of the data, define suitable learning behavior routing methods, and derive a dropout prediction algorithm based on whether learners have a dropout trend as the observation variable. Then select corresponding evaluation indicators to test the effectiveness and reliability of the algorithm;

Thirdly, based on the analysis results of dropout prediction, verify the corresponding six questions, summarize the conclusions of the problem testing, and explore the relevant paths of dropout and non dropout learning behaviors.

Because the temporal sequence of learning behavior in MOOCs is a multi data with complex relationships, which contains both the long-term evolution trend of learning behavior and the short-term fluctuation characteristics. Taking temporal sequence as the timeline, learning behavior will show different characteristics in different intervals. The current dropout prediction methods mainly analyze the learning behavior in a single interval, which is difficult to integrate the long-term evolution trend and short-term fluctuation characteristics, and easy to lose important temporal data, that might affect the effect of dropout prediction (Gupta et al., 2022).

Data analysis

This study proposes the dropout prediction method of MOOCs, in order to achieve effective decision feedback, we design the dropout prediction model based on the multi temporal sequences of learning behavior, namely DPM-MTS, which comprehensively considers the multi temporal sequences and seven learning behavior types, improves the multi full convolution neural network, fuses the multi iterative networks, realizes the parallel analysis and fusion calculation, mines the relevant characteristics of learning behavior types in different intervals, and adopts the global average pooling layer to ensure the integrity, reliability and efficiency of the data analysis. DPM-MTS needs to complete two tasks, first, we analyze the feature changes during the transformation of learning behavior types in different intervals, and complete the splicing to obtain the multi features; The second is to use the change information bottleneck to clean multi features, remove the interference data. The analysis process is shown in Fig. 2.

For the first task, DPM-MTS includes three sets of parallel double-convolution neural network, each of which can realize the iterative analysis of two full-convolution neural networks, in order to mine sufficient multi features comprehensively. Each double-full convolution neural network is composed of two convolution layers and two associated global pooling layers, which can correspond to a complete learning process. Convolutional kernels with different iterations have different domains of learning behavior. Convolutional kernels with larger iterations can capture the long-term evolution trend, and convolution kernels with smaller iterations can effectively mine the short-term fluctuations in real time.

The data analysis process of double full convolution neural network is described as follows.

Step 1. The multi temporal sequences of learning behavior are defined as input data x, which is expressed as a two-dimensional tensor. The pth convolution kernel of the ith full convolution neural network gradually moves along the temporal sequence, and generates a new feature sequence. The calculation formula is $s_{i,\mathrm{p}} = f(W \otimes x + b)$ (Formula 1), W is the parameter tensor of the convolution neural network. For the construction of the double full convolution neural network, another feature sequence $s_{j,q}$ of the full convolution neural network is the iterative result.

Step 2. It is needed to predict the dropout trend and obtain the continuous change of learning behavior. The global pooling layer is used to connect the convolution layer to process the feature sequence, then get the high-dimensional feature matrix. This process can analyze any length of temporal sequence, which can be used as

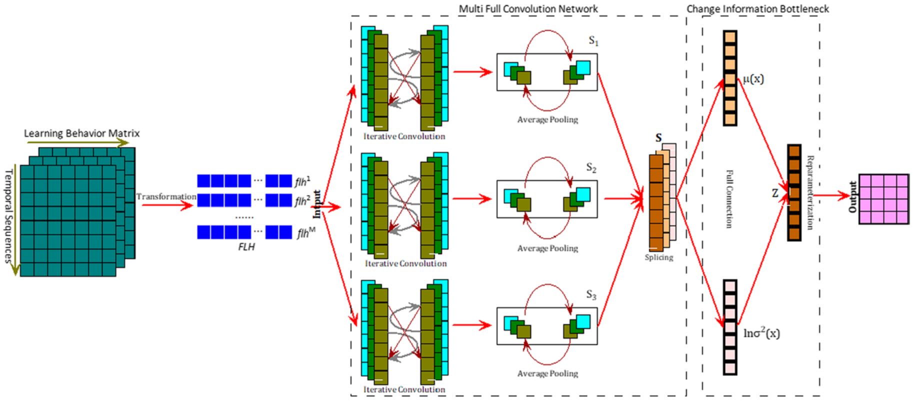

Fig. 2 Analysis Process of DPM-MTS

input. The calculation formula of global average pooling is described as $\bar{s}_{i,\mathrm{p}}=\frac{1}{\mathrm{m}}\sum_{k=1}^{m}s_{i,p}^{k}$ (Formula 2), $\bar{s}_{j,\mathrm{q}}=\frac{1}{\mathrm{n}}\sum_{l=1}^{n}s_{j,p}^{l}$ (Formula 3), $s_{i,p}^{k}$ respectively represent the kth and lth eigenvalues, mrespectively represent the feature number of the temporal sequence processed by the double-convolution neural network. Formulas 2-3 get the local mean of the feature sequence, and then compare the global mean corresponding to the two full-convolution neural networks, in order to select the optimal value, namely $s_{optimal}=\left[opt\left(\bar{s}_{i,1},\bar{s}_{j,1}\right),\cdots,opt\left(\bar{s}_{i,\mathrm{m}},\bar{s}_{j,m}\right),\cdots opt\left(\bar{s}_{i,n},\bar{s}_{j,m}\right)\cdots\right]$ is the selection function of the optimal feature value.

Step 3. We connect the features of three groups of double convolution neural networks to obtain the multi features of temporal sequence, that is $S=[s_1,s_2,s_3]$ (Formula 5). The rich and relatively optimal eigenvalues on multi intervals are realized, and the iterative analysis of the double convolution neural network complements the eigenvalue differences\, that makes the dropout prediction results more accurate.

For the second task, the main function of Change Information Bottleneck is to compress the multi temporal sequences of learning behavior, retain the characteristics related to dropout trend, and remove the irrelevant noise. The calculation process is described as follows.

Step 1. Calculate the logarithm $\ln\sigma^2(x)$ and the mean of $p(Z|x)$. The calculation formula is $\mathrm{p}(Z|x)=\mathrm{Neural}lNet\left(Z|\mu(x),\sigma^2(x)\right)$ (Formula 6), $\mathrm{Neural}lNet(\cdot)$ represents the encoder network composed of the full connection layer.

Step 2. The hidden variable Z is obtained by Formula 6, and the calculation process is expressed as $Z=\sigma(x)\cdot\varepsilon+\mu(x)$ (Formula 7). ε is the standard Gaussian function. Because the sampling process is discrete and non-differentiable, it is necessary to use reparameterization to ensure the continuity of the sampling process, in order to achieve effective training.

Step 3. Z is inputted into a softmax network to get the feasible dropout probability of learners.

Step 4. The loss function of the DPM-MTS is $\mathrm{L}=E_{x\sim p(x)}\left[E_{Z\sim p(Z|x)}[-\ln q(d|z)]\right.$ $\left.+\beta KL[p(Z|x)\|q(Z)]\right]$ (Formula 8), $E_{Z\sim p(Z|x)}[-\ln q(\mathrm{d}|z)]$ is to encode the description data x of the temporal sequence as hidden variable Z, and Z classifies the characteristics of learning behavior to obtain the dropout probability. d is the discrete variable to describe the dropout trend, that follows Bernoulli distribution. $\beta KL[p(Z|x)\|q(Z)]$ is a regular term, that makes Z tend to random distribution, and inhibit the information flow between input data and hidden variables. β is the Lagrange multiplier, which can control the balance of the two optimization conditions in Formula 8, K is a constant parameter.

Through the relevant steps of these two tasks, DPM-MTS might realize the continuous tracking of multi temporal sequences and the effective prediction of dropout trend.

Experiment

Based on the description in Section 3, DPM-MTS is fully tested and verified by experiment. The whole experiment is divided into three parts, (1) preprocessing and standardization of multi temporal sequences of learning behavior, (2) prediction of dropout trend, (3) evaluation of model performance. For the first part, the learning behavior instances are effectively processed, and the multi temporal sequences are converted into a matrix as the input of DPM-MTS; For the second part, DPM-MTS is used to adaptively mine the multi features of learning behavior, and the iterative calculation realizes the effective dropout prediction; For the third part, the performance of DPM-MTS is measured by four indicators, Precision, Recall, F1 and AUC. The similar models are selected to complete the comparative experiment. The first two parts have been completed in the process of data analysis and model design. This section focuses on the experimental process of the third part.

First, five basic models are selected, namely Classification and Regression Tree (CART), Naive Bayes (NB), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Convolution Neural Network (CNN). The input data of the first four models need to convert the matrix into a linear vector. Like DPM-MTS, CNN can directly use the matrix of temporal sequence as the input data.

As for the parameter configuration of DPM-MTS, the convolution kernel of each full convolution neural network is 9, each group is 18, the scale of each convolution kernel is 7, and the number of neurons in the full connection layer of change information bottleneck is 28. The learning rate is 0.001, and the batch size is 64. The essence of predicting dropout trend is to classify learners twice. Therefore, cross entropy is used as the loss function, and the whole process is trained 300 times. Experiments show that AUC of DPM-MTS reaches the optimal value when $\beta = 0.001$. During DPM-MTS analyzes the temporal sequence, sufficient experiments are carried out based on the above parameters.

In order to accurately describe the generalization ability of DPM-MTS and the other five comparative models, a tenfold cross-validation is applied. Before each cross-validation, the learning behavior instances are randomly mined. The test results are shown in Table 2. DPM-MTS has relatively optimal test results on four indicators compared with the others. Because the learning process is highly autonomous and personalized, it is easy to form a sparse matrix because of the uneven distribution of temporal sequence, which has a great impact on models that can only process vectors, and is easy to produce over-fitting phenomenon. At the same time, the correlation is formed between the characteristics of temporal sequence. Linear prediction analysis without relationships is inefficient and easy to produce undesirable prediction results.

Based on AUC of DPM-MTS and other five models, ROC is drawn, as shown in Fig. 3. The abscissa is False Positive Rate (FPR), and the ordinate is True Positive Rate

Table 2 Indicators of DPM-MTS and five basic models

Model	Precision	Recall	F1 Value	AUC
DPM-MTS	0.855	0.951	0.912	0.882
CART	0.813	0.841	0.856	0.619
NB	0.832	0.903	0.879	0.704
LDA	0.794	0.947	0.884	0.818
SVM	0.839	0.929	0.895	0.836
CNN	0.835	0.930	0.890	0.857

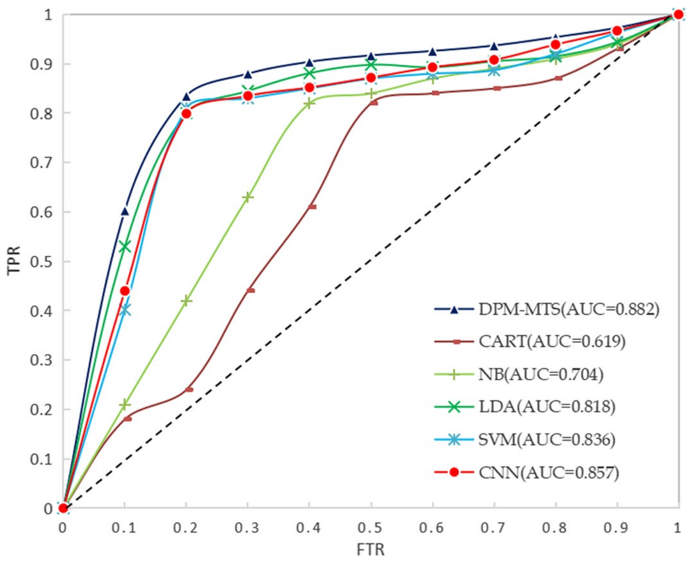

Fig. 3 ROC of different models

Table 3 Indicators of DPM-MTS and other approximate deep learning models

Model	Precision	Recall	F1 Value	AUC
DPM-MTS	0.945	0.977	0.934	0.904
CLSA	0.862	0.854	0.851	0.861
CONV-LSTM	0.927	0.883	0.892	0.870
DT-ELM	0.891	0.965	0.919	0.884

(TPR). About the area under ROC, DPM-MTS is superior and has stronger robustness and accuracy.

Compared with CART, NB, LDA and SVM, DPM-MTS can directly process the temporal sequence of learning behavior represented by matrix, fully track the temporal sequence, mine more features, and achieve more reliable dropout prediction. Compared with CNN, DPM-MTS can automatically mine and fuse the features of multi temporal sequences, realize the long-term evolution trend and short-term fluctuation characteristics. With the help of the change information bottleneck, it realizes the multi-interval feature screening, directly locate the features most relevant to the dropout trend, and improve the convolution analysis effect. Therefore, the performance is better than CNN.

Secondly, we also complete the comparative experiment with the latest in-depth learning models related to dropout prediction in recent years, which are: (1) CLSA (Fu et al., 2021), which integrates CNN, LSTM and static attention mechanism; (2) CONV-LSTM (Mubarak et al., 2021), which combines CNN and LSTM; (3) DT-ELM (Chen et al., 2019), which combines decision tree and limit learning machine. The test results are shown in Table 3. It can be seen that DPM-MTS is superior to others, and the prediction of dropout trend is more accurate, and the misgrading probability is less. When the multi temporal sequences of learning behavior are used as the input data, the prediction accuracy of dropout trend can be submitted more effectively.

Table 4 Test results of different temporal intervals

Temporal Interval	Precision	Recall	F1 Value	AUC
10 days	0.850	0.922	0.879	0.804
20 days	0.893	0.957	0.913	0.858
30 days	0.904	0.972	0.936	0.899

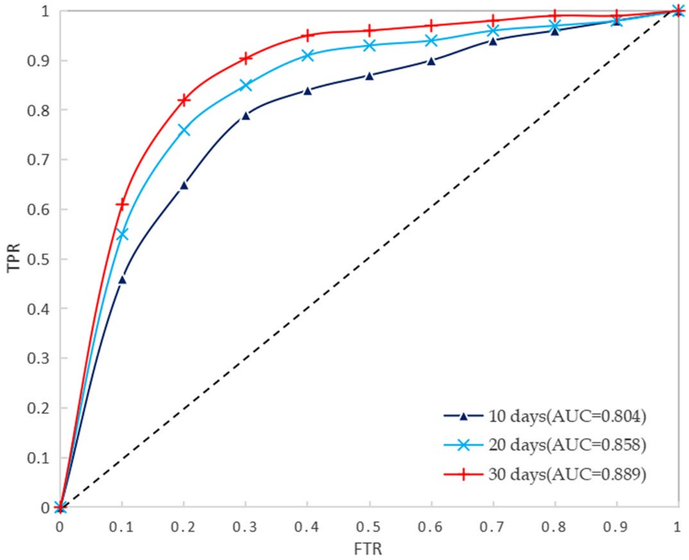

Fig. 4 ROC of input data in different temporal sequences

Based on the above experimental results, DPM-MTS in different temporal intervals is further tested, the early prediction at the beginning of the course is achieved. After the learners start course learning, the prediction results of dropout trend in the previous 10, 20 and 30 days is shown in Table 4, and ROC corresponding to AUC is shown in Fig. 4. The effect is the best when the temporal matrix of the previous 30 days is the input data, but the effect of the previous 10 days is the worst. At the beginning of course learning, there are fewer learning tasks and relatively few useful features, it can maintain high Recall. With the progress of the temporal sequence, there are more and more useful features, which also improves the performances.

Compared with the benchmark model and the approximate optimal deep learning model, the full test shows that DPM-MTS has improved the performance of the dropout prediction, which achieves the reliability and feasibility. The prediction accuracy and completeness of dropout trend are high, and the experimental results are effective.

Results

MOOCs have realized the creation and tracking of online complete learning behavior, and also completed the collection and management of learning process. Corresponding data describes the learning behavior, and also records the problems and risks (Xia & Qi, 2022b). Data-driven dropout prediction and decision feedback have become an important topic of empirical research. Based on the six problems raised in Section 3, this section analyzes the results of the dropout prediction, and combines the relationships described in Fig. 1, in order to fully demonstrate the dropout prediction results of MOOC platform "XuetangX", and classifies the multi temporal sequences and completes labeling dropouts. In the process of testing each problem, the potential association characteristics between different learning behavior types are mined, the corresponding massive learning behavior instances are obtained, and then the internal risk of learning behavior is deduced, so as to improve the quality of dropout trend prediction.

Since not all test results are statistically significant, this is directly related to the independent variables. In order to support the experimental analysis and explore the key conditions for the dropout trend, the variables are defined, the lower quartile and upper quartile are located for the corresponding values, if the value less than the lower quartile is low, and the value higher than the upper quartile is high, the value between the low quartile and upper quartile is moderate.

At the same time, in the process of testing the learning behavior path of P2–P6, it was clearly found that non dropout and dropout learners form two significantly different routes, Non dropout learners form positive route, that is, learners are able to complete the entire learning process with a positive and optimistic attitude, and achieved relatively ideal learning achievements, The driving and implementation of positive route can avoid the dropout trend; The negative route formed by dropout learners is that the learning process is discontinuous, the learning motivation is not high, the interaction between learners and learning behavior types is unstable, and the learning behavior instances have strong randomness. The formation of negative routes is prone to learning falling into the dropout trend.

P1. The learning behavior path formed by AC→CP can lead to dropout tend. It mainly tests the impact of course content on the dropout trend.

AC involves pages related to course contents, and learners will be involved in the corresponding CP during the browsing process, thus generating learning behavior and interest trend between AC and CP. Whether the learning path based on AC→CP can lead to dropout is directly related to the attractiveness of course content and the attention of learners. It also depends on the knowledge structure, the organization and presentation of the content, and the learning background of learners. Based on the analysis results of DPM-MTS, according to the descriptive rules of the characteristics, three independent variables are defined in the test process, which are: CC_Frequency (the frequency of the learner closing the page within 300s) , AC_Duration (the time from opening one page to closing this page), CT_Probability (the probability of transferring to other pages when the current page is closed), whether learners have dropout labels is as the observation variable. After data analysis, the statistical results in Table 5 are obtained.

In the experiment, CC_Frequency, AC_Duration and CT_Probability are tested with the dropout tend of learners. Due to the obvious differences in the acceptance of online learning among learners of different majors, the data analysis shows that the interest of STEM learners is significantly stronger than that of Social Science learners, and STEM learners have more diverse learning organization methods, such as

Table 5 AC→CP test results of dropout trend

Variable	Dropout (P-value)	Variable	Dropout (P-value)	Variable	Dropout (P-value)
CC_Frequency	STEM: 0.509 SS: 0.034*	CC_Frequency↑ AC_Duration↓	STEM: 0.035* SS: 0.005**	CC_Frequency↑	STEM: 0.000*** SS: 0.000***
AC_Duration	STEM:-7.311 SS: -0.025*	CC_Frequency↑ CT_Probability↓	STEM: 0.004** SS:0.018*	AC_Duration↓	
CT_Probability	STEM:-13.205 SS:-0.009**	AC_Duration↓ CT_Probability↓	STEM: 0.009** SS: 0.002**	CT_Probability↓	

***p<0.001, **p<0.01, *p<0.05; *SS* Social Science, *STEM* Science, Technology, Engineering & Mathematics

online communication, collaboration and practice. Therefore, the multi temporal sequences of learning behavior is classified by major, then the correlation analysis and training are completed respectively. These three independent variables have different effects on the corresponding majors of dropout trend, and Social Science might be significant. With high CC_Frequency, low AC_Duration and low CT_Probability, the combination of any two independent variables has a significant impact on the dropout trend. Whether the course is STEM or Social Science, Social Science is more significant; With high CC_Frequency, low AC_Duration and low CT_Probability, STEM and Social Science, there is a strong significance. The experimental results show that the learning behavior path formed by AC →CP will have key impacts on the dropout trend within the value range of key characteristics.

P2. The learning behavior path formed by AC→AW→(WV, VC, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (WV, VC, QA, FD).

When learners have a certain behavioral trend of AW in AC, it will lead to the association of other learning behavior types and the formation of paths. Each learning behavior type is associated with CP, thus forming a learning behavior sequence and interest preferences about AC→AW→(WV, VC, QA, FD)→CP, which also means that it will have a negative impact and forms dropout trend. The key solution to this problem is to take AC→AW as the basic route of learning path, locate other learning behavior types associated in the temporal sequence. Based on the analysis results of DPM-MTS on the multi temporal sequences of learning behavior, according to the descriptive rules of characteristics, the independent variable of the relevant learning behavior type is defined as the participation frequency (click rate), whether the learner has dropout labels as the observation variable, and the structural relationship between the independent variable and the observation variable is constructed. After data analysis, the learning path in Fig. 5 is obtained. There are two routes, positive route is the behavior path without dropout trend, which forms strong correlation, while negative route is the behavior path with dropout trend, which also produces certain correlation. However, the two routes are different, the route that does not produce strong correlation does not mean that there will be a dropout trend, but the behavior topology and correlation of dropout trend constitute the risk route of learning behavior.

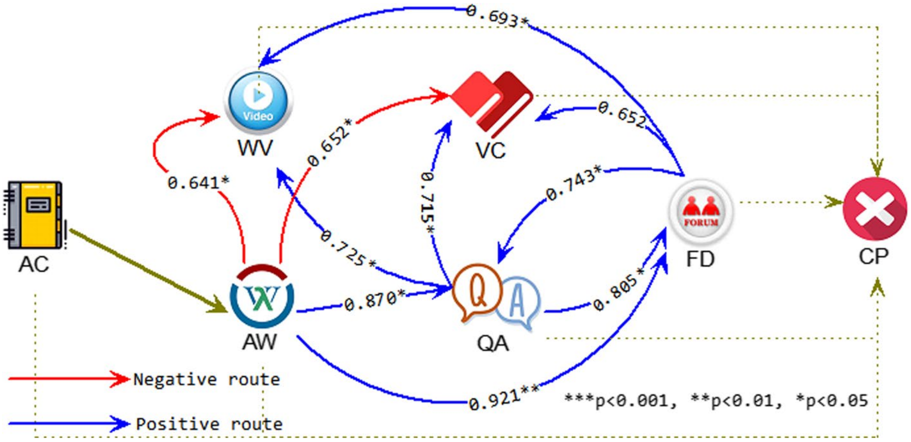

Fig. 5 Learning Path of AC→AW→(WV,VC,QA,FD)→CP

It can be seen from Fig. 5 that the positive route is based on the strong correlation between AC→AW, QA and FD have been fully applied, and AW→QA and AW→FD have correlation and significant. When QA and FD are named as the starting nodes, learners have built strong links with WV and VC, which shows that AW enables learners' attention for knowledge, realizes the communication between learners and teachers during the participation of QA and FD, forms good learning behavior, and constructs a lasting and continuous temporal sequence. However, negative route obviously ignores the interaction between QA and FD. there is no correlation. Learners directly realize a relatively significant learning path between WV and VC. Learning behavior shows strong personalizationn, and has an obvious dropout trend in the earlier temporal sequence, that might lead learners to end the learning process.

P3. The learning behavior path formed by AC→WV→(AW, VC, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (AW, VC, QA, FD).

When learners take part in WV during AC, it will lead to the association and path formation of other learning behavior types, and the participation of each learning behavior type is also associated with the corresponding CP, thus forming a learning behavior sequence and interest preference about AC→WV→(AW, VC, QA, FD)→CP, which also means that there will be a reverse intention, there is the dropout trend. The key solution is to define AC→WV as the basic participating route of the learning path, and mine other learning behavior types associated in the temporal sequence, as well as the potential relationships. Therefore, based on the analysis results of DPM-MTS for the multi temporal sequences of learning behavior, according to the description rules of characteristics, the independent variable of the relevant behavior type is defined as the participation frequency (click rate), whether the learner has dropout labels as the observation variable, and the structured relationship between the independent variable and the observation variable is constructed. According to the dropout labels of the learning behavior sequence, the learning behavior trend is also divided into two different processes to describe the positiveness and negativity in the learning process. The dropout path in Fig. 6 is obviously diverse and more discrete.

As can be seen from Fig. 6, the construction of positive route is based on AC→AW as a strong correlation, and learners have a very strong and significant correlation with VC. Helped by VC, the retrieval trend of WIKI has been triggered. There is a significant correlation between VC and AW, which further drives the more significant QA. Some learners directly take part in FD, thus forming a benign and rational learning behavior sequence. This process shows that learners achieve the conversion from self-answering to collaborative solving, and achieved positive learning behavior, then a persistent and continuous learning behavior sequence is constructed. During the construction of negative route, WV also enables learners' enthusiasm to participate in VC, forming a significant correlation. However, the choice and participation of learning behavior types are discrete and random. Many learners have the communication trend of FD, which guides the interest of learners to AW, but it has little relevance to the learning content. At the same time, learners obviously ignore the communication role of QA. For the interaction and cooperation related to learning tasks, there is an obvious dropout trend in the earlier temporal sequence, that might lead learners to end the learning process.

P4. The learning behavior path formed by AC→VC→(AW, WV, QA, FD)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (AW, WV, QA, FD).

When learners take part in AC, they will also have certain behavioral trend about VC. Even though the occurrence of this probability is not absolutely universal, it also has a strong correlation, but it is not necessarily directly related to the course content, it will also lead to the formation of other learning behavior types and paths. The participation of each learning behavior type is also associated with the corresponding CP, which forms a learning behavior sequence and interest preference about AC→VC→(AW, WV, QA, FD)→CP, and also has the potential dropout trend. The key solution to this problem is to define AC→VC as the basic route of the learning path, and mine other learning behavior types associated in the temporal sequence, as well as the potential relationships.

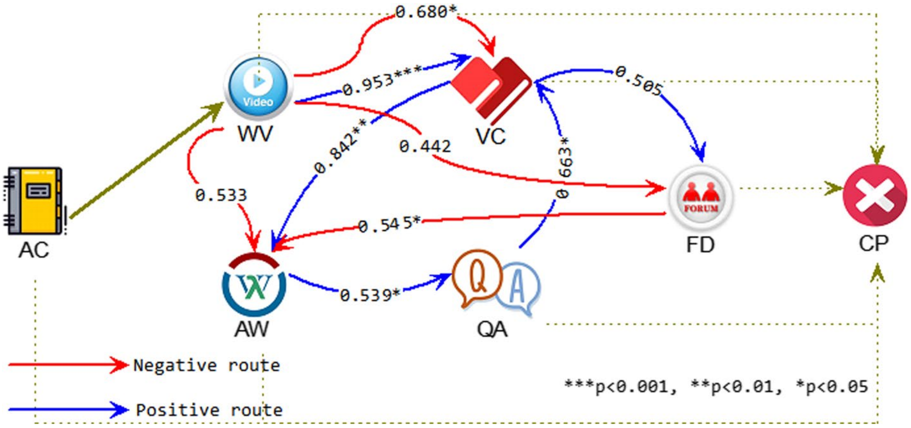

Fig. 6 Learning path of AC→WV→(AW,VC,QA,FD)→CP

Therefore, based on the analysis results of DPM-MTS for the multi temporal sequences of learning behavior, according to the description rules of characteristics, the independent variable of the relevant behavior type is defined as the participation frequency (click rate), whether the learner has dropout labels as the observation variable, and the structured relationship between the independent variable and the observation variable is constructed. According to the dropout labels of the learning behavior sequence, there are two obvious branches in the behavior trend, namely, positive route and negative route. Positive route is an effective learning behavior, and negative route produces a high probability of dropout trend, as shown in Fig. 7.

As can be seen from Fig. 7, the construction of positive route is based on AC→VC as a strong correlation. Learners have a strong and significant correlation with AW. Driven by AW, learners achieve WV and QA. QA drives FD, but there is no strong correlation or significance, It shows that learners do not form groupness and preferences with AC→VC as the basic path. The effective learning process of learners reflects strong personalization, also achieves better learning objectives, and forms a lasting learning behavior sequence. In general, learners generally have a good knowledge base, insight and self-learning ability. In the construction process of negative route, VC is associated with WV, AW and FD, but there is no effective correlation and significance. Learners' participation in learning behavior types has generated highly discrete data, which shows that learning behavior is random, and there is no direct relationship with actual learning needs. There are obvious problems in learning attitude, which can not produce stable learning interest, and negative emotions are obvious, Therefore, there is an obvious dropout trend in the earlier temporal sequence, that no lasting and reliable learning behavior has been constructed.

P5. The learning behavior path formed by AC→QA→(WV, AW, WV, FD) →CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (WV, AW, WV, FD).

When learners directly produce the behavior trend of QA in AC, it is mainly due to the direct demand of some learners, such as when exploring the learning task of the

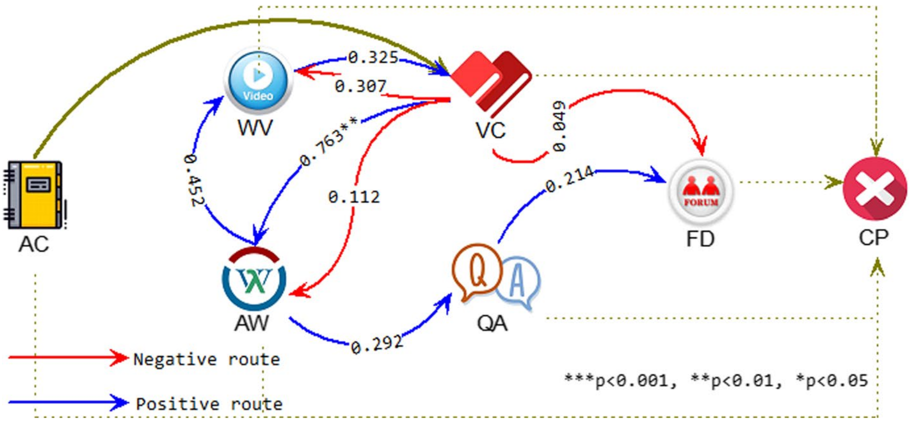

Fig. 7 Learning path of AC→VC→(AW,WV,QA,FD)→CP

course, whether they are suitable for the course, how difficult the course is to learn, which will drive some learners to generate the behavioral trend of online communication. QA is generally open, and will also attract the attention of other learners, so as to build some paths with other relevant learning behavior types. The participation of each learning behavior type is also associated with the corresponding CP, thus forming a learning behavior sequence and interest preference about AC→QA→(WV, AW, WV, FD)→CP, which also has the dropout trend. The key solution to this problem is to define AC→QA as the basic route of the learning path, and mine other learning behavior types associated in the temporal sequence, as well as the potential relationships. Therefore, based on the analysis results of DPM-MTS, according to the description rules of characteristics, the independent variable of the relevant learning behavior type is defined as the participation frequency (click rate), whether the learner has dropout labels as the observation variable, and the structured relationship between the independent variable and the observation variable is constructed. According to the dropout labels of the temporal sequence, the learners with positive route construct long-term and stable learning behavior, while the learners with negative route use QA inefficiently, and result in a high probability of dropout trend, as shown in Fig. 8.

As can be seen from Fig. 8, the positive route is based on the strong correlation of AC→QA, and driven by QA, learners have strong and significant correlation with WV, AW and VC at the same time. With AW as the enabled node, learners can also strengthen their attention to WV and produce extremely significant correlation. Learners also generate the interaction and cooperation with FD during WV, FD has a direct relationship with learning content. It shows that learners have strong groupness and preferences for the learning behavior, when AC→QA is viewed as the basic path. Effective learning process is transitive and relevant, and learners also complete better learning objectives, forming a lasting learning behavior sequence. QA has an important enabling role for the positive learning behavior. In the construction process of negative route, QA does not trigger statistically significant correlation and significance. QA drives some learners to participate in AW and WV, but the relationship is loose . This participation is not stable, learning behavior types are highly random, and there is no direct correspondence with actual learning needs. Learners do not have effective subjective intention to

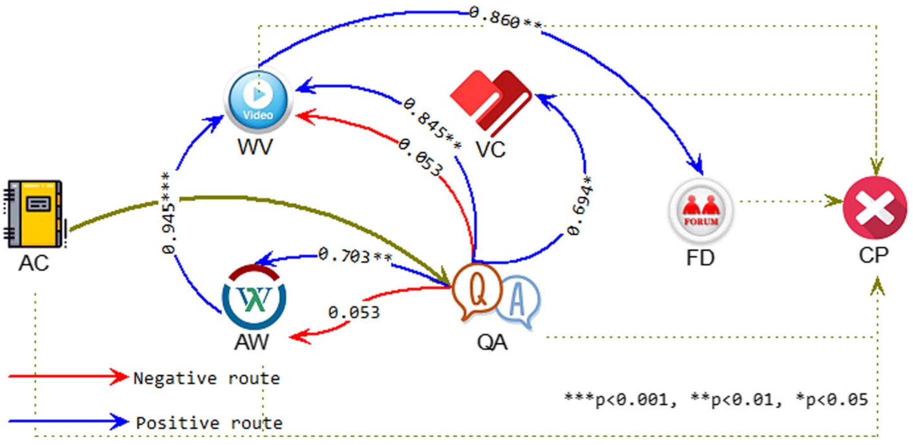

Fig. 8 Learning path of AC→QA→(WV,AW,WV,FD)→CP

participate in the whole learning process, so it is impossible to generate stable learning interest. Therefore, there might be the dropout trend in the earlier temporal sequence, and the lasting and reliable learning behaviors is not constructed.

P6. The learning behavior path formed by AC→FD→(WV, AW, WV, QA)→CP can lead to dropout. It is necessary to explore and build the internal risk path of learning behavior types (WV, AW, WV, QA).

When learners directly produce the learning behavior trend of FD in AC, they directly visit the forum before learning some course, and there are many such learners. The demand of STEM learners is obviously stronger. Helped by FD, they can directly locate effective relevant learning resources. Anyone who becomes a participant in FD will usually pay attention to the associated learners of online learning. FD enables a large number of learners to participate in the learning process. Of course, there are also learners who give up directly, or still choose to dropout after a period. The information in FD can also be read by registered users. Driven by FD, the participation and association of other learning behavior types are activated, and the participation of each learning behavior type is also associated with CP, thus forming a learning behavior sequence and interest preference about AC→FD→(WV, AW, WV, QA)→CP, which also has the dropout risk. The key solution to this problem is to define AC→FD as the basic route of the learning path, and mine other learning behavior types associated in the temporal sequence, as well as the potential relationships. Based on the analysis results of DPM-MTS, according to the descriptive rules of characteristics, the independent variable of the relevant learning behavior type is defined as the participation frequency (click rate), and the structured relationships between the independent variable and the observation variable are constructed, whether the learner has dropout labels as the observation variable. According to the dropout labels of the temporal sequence, the learners with positive route construct long-term and stable learning behavior, while the learners with negative route use FD inefficiently, resulting in a high probability of dropout trend, as shown in Fig. 9.

It can be seen from Fig. 9 that the construction of positive route is based on the strong correlation of AC→FD. FD and WV have strong correlation and significance, as well as

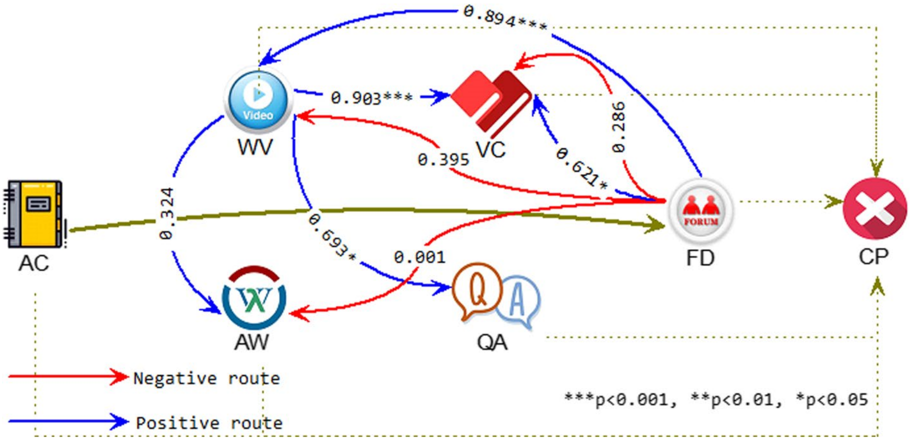

Fig. 9 Learning path of AC→FD→(WV,AW,WV,QA)→CP

FD and VC. WV enables the learning path, WV and VC have a strong correlation and significance, and QA can attract more learners, but AW is significantly lower than the test results of the above five problems. It shows that learners form strong groupness and preferences when AC→FD is defined as the basic path. Before entering the actual learning process, learners have generated a relatively strong trend for knowledge and interest preferences. They have a strong awareness of solving problems independently, and completing tasks interactively and cooperatively. They build a lasting learning behavior sequence, and FD is fully implemented at the beginning of the course, It plays a key role in the continuity and transmission of positive learning behavior. In constructing the negative route, the intention to participate in FD is obvious, and the subjective initiative about the course learning is not strong. Although some instances of AW, WV and VC are generated, the learning target is not strong, and the construction of learning behavior types are highly discrete and random, and have no obvious statistical significance. There is no direct correspondence between learning behavior participation and actual learning needs. Learners have no confidence to participate in the whole learning process, and their attitude is negative, so it is impossible to generate stable learning interest. Therefore, there might be dropout trend in the earlier temporal sequence, and the lasting and reliable learning behaviors is not constructed.

So the dropout trend in MOOCs is directly related to the selection order of learning behavior types, subjective intentions, awareness of interaction and cooperation, and the early warning mechanism and intervention measures. It is found that the positive and negative route of temporal sequence are not complementary, so the learning behavior is not a simple mathematical problem. It needs not only to meet the educational laws, but also to consider the learning background and intention of learners, then implement corresponding adaptive tracking, data analysis and decision feedback in the whole temporal sequence (Xia, 2022a), which will play an important role in building effective learning behavior, and has strong practical significance for the research and positive guidance of dropout trend.

Discussion

MOOCs realize the autonomy and personalization of the online learning process, which provides great convenience for course selection. However, it is also separated from the direct tracking and monitoring of the instructors, making MOOCs produce a large number of dropouts. There are many factors for learners to give up the learning process, the key problem is that learners do not form strong and lasting learning interest, lack of goal and loose attention. How to make more learners make full use of online resources, build the sustainable learning behavior sequence, develop positive learning attitude, reduce the negative emotion of dropouts, and improve the learning effect and course passing rate? that needs MOOCs to give full play to the role of data analysis and decision feedback, and timely discover learning trends and preferences, rather than just a data collector.

We hope that more learners can have a good experience in the learning process of MOOCs, and successfully complete the entire learning tasks. In this process, it is primarily up to learners to make their own choices of relevant learning resources and form their own learning behavior paths. This is beneficial for learners with MOOCs learning experience, but for new learners, appropriate learning behavior guidance or recommendations of suitable learning resources are needed to help them better and faster find suitable learning methods and learning interests. Therefore, this study believes that the construction process of learning behavior should not only consider the subjective wishes of learners, but also have a certain analysis and prediction mechanism to provide guidance and intervention to learners at key stages.

The research purpose of this study is to design a massive data driven learning behavior tracking and dropout prediction method, that the massive dropout problems are caused by the online learning process of MOOCs. We introduce continuous temporal sequence of the learning process, analyze the change rules, and timely locate the routes that might cause dropout trends. Therefore, based on the interaction and participation characteristics of learning behavior, it is divided into seven types, with the observation target of whether learners are dropouts, two types of learning behavior routes are formed, namely, positive route and negative route. Through the analysis and prediction of massive learning behavior, it is found that positive route can effectively avoid the dropout trend, and negative routing should be targeted for early warning and intervention in the learning period, learners might be guided to avoid the dropout risk. Such research results can provide assistance for the benign construction of similar learning behaviors in MOOCs, and also facilitate the introduction of appropriate learning behavior guidance strategies and recommendation mechanisms on the other learning platforms, which can improve subsequent learning achievements.

The research process of this study considers the dynamic nature of the learning process and the personalization of learners, fully calculates learners' learning needs and the learning organization methods of MOOCs, introduces the temporal characteristics of learning behavior, and achieves effective correlation and fusion prediction throughout the learning process. On this basis, through improving and designing suitable neural network, Completed the actual data testing and comparison of the dropout prediction effect are completed. Experimental analysis shows that the research process and experimental scheme of this study are reliable and feasible. Moreover, through the division and construction of positive and negative routes for learning behavior, we train the comparative data on non dropout guidance and dropout intervention, provide a key basis for decision feedback on learning behavior, which traditional qualitative research on dropout cannot do.

Furthermore, this study fully considers the temporal sequence of learning behavior in MOOCs, designs a dropout prediction method to achieve temporal sequence tracking and decision feedback. According to the data analysis results, the routes of positive and negative learning behavior are visually constructed, and the learning behavior types and relationship characteristics corresponding to dropout trend are deduced. On this basis, it further summarizes the effective conclusions of the dropout trend tracking and decision feedback, and demonstrates effective intervention measures and strategies, which are mainly reflected in the following aspects.

(1) it is needed to build the analysis mechanism of learning background and knowledge structure, and mine the key influencing factors of potential dropout trend.

At present, MOOCs give learners full flexible choice. As long as learners can become registered users, they can find the resources they need. This learning process depends entirely on the learners' choice, that is highly unstable and random. The learners are extremely vulnerable to the influence of the information at the first time, or the learners can't get the applicable course contents. The learning preferences will also change irregularly due to the knowledge difficulty and teaching methods. The learning behavior is unstable, which leads to the dropout trend.

Therefore, a large number of dropouts in MOOCs explain that it is not reliable for learners to construct their own learning behaviors, MOOCs need to implement early and effective data analysis of learning background and knowledge structure, and give relatively objective judgments, locate and predict the key elements of potential dropouts, evaluate learning interests and knowledge preferences, and implement the correlation analysis of existing online resources, as well as adaptive selection and recommendation, that might help learners build effective candidate learning contents, and give the corresponding strategies of temporal sequence of learning behavior. On this basis, the tracking mechanism of learning behavior is integrated to achieve data monitoring and decision feedback.

(2) It is needed to deploy timely communication media for interaction and collaboration between learners, and understand learning needs and preferences as soon as possible.

It is an objective conclusion to get learning needs and preferences through data analysis, but it does not represent the current subjective intention. After all, thinking is a complex process. Learners can also break through the previous learning background and knowledge, and carry out their researches in new fields. It is necessary to objectively analyze the learning background and knowledge structure, and also to conduct relatively sufficient researches on learners through timely communication.

Therefore, a large number of dropouts in MOOCs explain that the learning process does not understand the real needs, that is not stable. MOOCs need to deploy appropriate timely communication media, such as QA, Forum, Chat, etc., which defines learning needs and preferences as a series of questions, so that learners can complete the collection of results before entering the learning process. Through certain data analysis means, we realize the comparison and correlation between the analysis results and the current needs, explore the obstacles and problems that learners may encounter in the learning process, improve the strategies of the temporal sequence of learning behavior. Then we integrate the tracking mechanism, realize the data monitoring and decision feedback, and reduce or avoid the dropout trend.

MOOCs need to continuously optimize the management of learning resources, improve the pass rate and learning enthusiasm, and achieve effective prediction and adaptive decision feedback of dropout trend, and build reliable learning behavior tracking strategies (Anghel et al., 2022). The effective implementation of (1) and (2) requires MOOCs to integrate and track the learning analytics and decision recommendation mechanism.

Conclusion

By analyzing and predicting the dropout trend with a large number of learning behavior instances in MOOCs, we can find out the learners' emotion and intention in advance, and help MOOCs or instructors to take effective early warning and intervention in time, reduce the dropout rate, and improve the learning achievement and pass rate. After all, the learning process is very complex. We need to fully consider the knowledge structure and relationship, but also the learning ability and needs of learners. This is not a completely autonomous, random and personalized process (Xia, 2022b). MOOCs need to timely monitor the learning behavior, and should make relatively accurate prediction, For learners who have the potential to give up learning, we might timely answer questions and guide them.

This study designs corresponding method for the analysis and prediction of dropout trend in MOOCs, in order to provide adaptive monitoring mechanism for long-term learning behavior tracking and short-term learning method selection, and provide appropriate decision feedback for learners. First of all, the key learning behavior types are mined, the descriptive framework is constructed, and the corresponding problems to achieve the effective prediction of dropout trend are deduced; Secondly, DPM-MTS is designed. The improved change information bottleneck is used to locate the key features and attributes of dropout trend; Thirdly, we analyze and demonstrate the problems of dropout prediction, discuss the effective strategies to improve the tracking and intervention of temporal sequence of learning behavior, and analyze the learning behavior types and related needs that should be strengthened in the decision feedback. The whole research might have important practical significance, and will also provide a feasible argument for the online learning process and the solution of dropouts.

There are also some limitations, such as the entire analysis is based on a complete set of learning behavior instances. MOOCs' definitions and organizational standards for the learning process determine the attributes and characteristics of data, which will also affect the algorithm design process (Xia & Qi, 2023). However, the differences in the research and design process for other similar issues mainly lie in different parameter definitions and different feature scale setting. Since the purpose of this study is to provide technical support and decision feedback for MOOCs' dropout prediction, the design process of the method also considers openness and compatibility, and considers the sufficient feature scale as much as possible. Therefore, relevant researchers can appropriately expand or limit the scale of input data based on this study, and can also complete dropout prediction on other MOOCs platforms. This is the significance of this study.

At the same time, MOOCs can achieve data collection, but it might not be able to collect complete learning process data. The learning behavior dataset we obtain may not be complete, that is, there may be some data loss, and MOOCs may not necessarily need to collect all the data. It is also full of a lot of noise, and we need more valuable data, while minimizing the interference of meaningless data, this can improve our dropout prediction results, which requires the joint progress of MOOCs and data analytics.

In the follow-up research, we will further improve the dropout prediction model, solve the dropouts at the beginning of the new course, and implement the adaptive recommendation strategies for new learners when facing massive learning resources, so as to improve the effectiveness of dropout prediction and the adaptability of decision feedback.

Acknowledgements

Thanks for the technical support provided by of National Taiwan University of Science &Technology and Tsinghua University, as well as the theoretical guidance and practical reference of University of Agder and Qufu Normal University.

Author contributions

All authors contributed to the writing of this manuscript. The work of Xiaona Xia included methodology, validation, investigation, Resources, data analytics, writing-original draft, writing-review & editing, visualization and project administration. The work of Wanxue Qi included conceptualization and writing-review & editing. The names are in order of the amount of contribution given. All authors read and approved the final manuscript.

Authors' information

Xiaona Xia is an associate professor and PhD of Qufu Normal University. She is also a member of IEEE computer society and CCF. Her research interests include learning analytics, interactive learning environments, collaborative learning, education big data, educational statistics, data mining, service computing, etc.

Wanxue Qi is a PhD supervisor of Qufu Normal University. He is a famous education expert and has made remarkable achievements in higher education and moral education theory. His research interests include education big data, moral education, etc.

Funding

This study is supported by National Office for Philosophy and Social Sciences (Grant NO. BEA190107) and Social Science Planning Project of Shandong Province (Grant NO. 22CJYJ20).

Availability of data and materials

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Received: 2 March 2023 Accepted: 27 April 2023
Published online: 30 May 2023

References

Anghel, E., Tobias-Littenberg, J., & Reich, J. (2022). Location in the multiverse of methods: Measuring online users' contexts. *International Journal of Social Research Methodology., 2022*(9), 1–20. https://doi.org/10.1080/13645579.2022.2125648

Anttila, S., Lindfors, H., Hirvonen, R., Määttä, S., & Kiuru, N. (2022). Dropout intentions in secondary education: Student temperament and achievement motivation as antecedents. *Journal of Adolescence*. https://doi.org/10.1002/jad.12110

Ashenafi, M. M., Andres-Bray, J. M., Hutt, S., Baker, R. S., & Brooks, C. (2022). Controlled outputs, full data: a privacy-protecting infrastructure for mooc data. *British Journal of Educational Technology., 53*(4), 756–775. https://doi.org/10.1111/bjet.13231

Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2022). Taking action to reduce dropout in MOOCs: Tested interventions. *Computers & Education*. https://doi.org/10.1016/j.compedu.2021.104412

Chanaa, A., & Faddouli, N. (2022). An analysis of learners' affective and cognitive traits in context-aware recommender systems (CARS) using feature interactions and factorization machines (FMS). *Journal of King Saud University-Computer and Information Sciences., 34*(8), 4796–4809. https://doi.org/10.1016/j.jksuci.2021.06.008

Chen, J., Feng, J., Sun, X., Wu, N., & Chen, S. (2019). Mooc dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering, 2019*(1), 1–11. https://doi.org/10.1155/2019/8404653

Fu, Q., Gao, Z., Zhou, J., & Zheng, Y. (2021). CLSA: a novel deep learning model for MOOC dropout prediction. *Computers & Electrical Engineering, 94*(4), 107315. https://doi.org/10.1016/j.compeleceng.2021.107315

Ghada, Refaat, El, & Said. (2016). Understanding how learners use massive open online courses and why they drop out. Journal of Educational Computing Research, 55(5), 724-752. https://doi.org/10.1177/0735633116681302

Gubbels, J., van der Put, C.E. & Assink, M. (2019). Risk factors for school absenteeism and dropout: A meta-analytic review. Journal of Youth and Adolescence. 48(1), 1637–1667. https://doi.org/10.1007/s10964-019-01072-5

Gupta, A., Garg, D., & Kumar, P. (2022). Mining sequential learning trajectories with hidden markov models for early prediction of at-risk students in e-learning environments. *IEEE Transactions on Learning Technologies., 15*(6), 783–797. https://doi.org/10.1109/TLT.2022.3197486

Hsu, L. (2022). EFL learners' self-determination and acceptance of LMOOCs: The UTAUT model. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2021.1976210

Khoushehgir, F., & Sulaimany, S. (2023). Negative link prediction to reduce dropout in Massive Open Online Courses. *Education and Information Technologies., 2023*(1), 1–20. https://doi.org/10.1007/s10639-023-11597-9

Kim, T. D., Yang, M. Y., Bae, J., Min, B. A., Lee, I., & Kim, J. (2017). Escape from infinite freedom: effects of constraining user freedom on the prevention of dropout in an online learning context. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2016.09.019

Mourdi, Y., Sadgal, M., Elalaoui Elabdallaoui, H., El Kabtane, H., & Allioui, H.(2022). A recurrent neural networks based framework for at-risk learners' early prediction and MOOC tutor's decision support. Computer Applications in Engineering Education. 2022(11), 1061-3773. https://doi.org/10.1002/cae.22582

Mubarak, A. A., Han, C., & Hezam, I. M. (2021). Deep analytic model for student dropout prediction in massive open online courses. *Computers & Electrical Engineering, 93*(1), 107271. https://doi.org/10.1016/j.compeleceng.2021.107271

Rodríguez, P., Villanueva, A., Dombrovskaia, L., & Valenzuela, J. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies., 2023*(1), 1–47. https://doi.org/10.1007/s10639-022-11515-5

Xia, X. (2020a). Random field design and collaborative inference strategies for learning interaction activities. *Interactive Learning Environments., 2020*(12), 1–25. https://doi.org/10.1080/10494820.2020.1863236

Xia, X. (2020b). Learning behavior mining and decision recommendation based on association rules in interactive learning environment. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2020.1799028

Xia, X. (2021a). Sparse learning strategy and key feature selection in interactive learning environment. *Interactive Learning Environments., 2021*(11), 1–25. https://doi.org/10.1080/10494820.2021.1998913

Xia, X. (2021b). Decision application mechanism of regression analysis of multi-category learning behaviors in interactive learning environment. *Interactive Learning Environments., 2021*(4), 1–14. https://doi.org/10.1080/10494820.2021.1916767

Xia, X. (2021c). Interaction recognition and intervention based on context feature fusion of learning behaviors in interactive learning environments. *Interactive Learning Environments., 2021*(1), 1–19. https://doi.org/10.1080/10494820.2021.1871632

Xia, X. (2022a). Application technology on collaborative training of interactive learning activities and trend preference diversion. *SAGE Open, 12*(2), 1–15. https://doi.org/10.1177/21582440221093368

Xia, X. (2022b). Diversion inference model of learning effectiveness supported by differential evolution strategy. *Computers and Education: Artificial Intelligence., 3*(1), 100071. https://doi.org/10.1016/j.caeai.2022.100071

Xia, X., & Qi, W. (2022a). Early warning mechanism of interactive learning process based on temporal memory enhancement model. *Education and Information Technologies., 2022*(7), 1–22. https://doi.org/10.1007/s10639-022-11206-1

Xia, X., & Qi, W. (2022b). Temporal tracking and early warning of multi semantic features of learning behavior. *Computers and Education: Artificial Intelligence., 3*(1), 100045. https://doi.org/10.1016/j.caeai.2021.100045

Xia, X., & Qi, W. (2023). learning behavior interest propagation strategy of MOOCs based on multi entity knowledge graph. *Education and Information Technologies., 2023*(3), 1–29. https://doi.org/10.1007/s10639-023-11719-3

Xia, X., & Wang, T. (2022). Multi objective evaluation between learning behavior and learning achievement. *Asia-Pacific Education Researcher., 2022*(12), 1–15. https://doi.org/10.1007/s40299-022-00703-z

Publisher's Note