A dropout prediction method based on time series model in MOOCs

Min Chen¹, Lifa Wu¹

¹Nanjing University of Posts and Telecommunications, Nanjing, China

Corresponding author and e-mail: Lifa Wu, wulifa@njupt.edu.cn

Abstract. In recent years, MOOCs has enjoyed great popularity due to its convenience and openness. However, with the development of MOOCs, the high dropout rate has aroused extensive attention. By analyzing the data of students’ behavior and then predicting whether students are at risk of dropout, it can improve the course completion rate. Most of the existing methods relying on feature engineering and the sequential characteristic of data is not effectively utilized. In this paper, we propose a time series model named CNN-LSTM-ATT, which focuses more on local valid information and temporal information of the data. Through extensive experiments on a public dataset, it shows that the proposed model can effectively predict students’ dropout behavior.

1. Introduction

As a new type of online education platform, Massive Open Online Courses (MOOCs) become popular all over the world. Different from traditional teaching models, MOOCs break the limitations of time and space on the learning environment. It makes it possible for everyone to learn at all times and places. Not only can MOOCs provide adequate learning resources, but also meet the learning needs of different student groups to a certain extent.

Despite the convenience and popularity, more and more problems have been exposed during the development of MOOCs. The problem of student loss has become the inhibitor restricting its further development. The most prominent problem is that the completion rate of courses is very low, and it is common for students to drop out. Since MOOCs lack an effective supervision mechanism like in traditional classrooms, students need to control the learning progress by themselves. Coupled with the freedom of the platform, students can withdraw from the course at any time, so the dropout rates are invariably high.

There are many interactions between students and the MOOC platform, such as watching video, forum communication, homework exercises and etc. These behaviors generate a lot of valuable information, so we can explore student behavior from the perspective of "dropout prediction". Further, we propose a prediction model named as CNN-LSTM-ATT, which combines Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Attention. Through extensive experiments on a public dataset, we show that the proposed model can achieve results better than or comparable to existing methods.

2. Related work
The current research on students’ dropout prediction in MOOCs is mainly based on machine learning algorithms. These types of methods extract features from user data and build a prediction model based on the features. Kloft M et al. [1] and Burgos C et al. [2] respectively used Support Vector Machine (SVM) and Logistic Regression (LR) as classification models to predict whether students are at risk of dropout. Al-Shabandar R et al. [3] used Decision Tree algorithm which mainly based on features classify the student instances. LIU T et al. [4] used K-Means clustering algorithm to cluster students based on student behavior and then predict dropout behavior. At the same time, researchers try to combine multiple algorithms to make predictions. For example, Qiu J et al. [5] and Liang J et al. [6] both combine LR and SVM to build an effective prediction model. These methods have the following two shortcomings. One is that the process of feature extraction is quite complicated. The other is that the temporal information is not effectively used, so the obtained prediction accuracy is not ideal.

Considering the temporal characteristic of the behavior data, some researchers regard the prediction of dropouts as a time series prediction problem. Hidden Markov Model (HMM) is a time series probability model. Balakrishnan G et al. [7] marked all states (hidden states) as discrete states in their research, and then used EM algorithm to optimize parameters. But the data state is not all discrete, Wang W et al. [8] used the Recurrent Neural Network (RNN) to build model. Because RNN faces the problem of gradient disappearance, Tang C et al. [9] used the Long Short-Term Memory (LSTM), which structure adds three special gates to control the input and output. However, as the length of the input data sequence increases, data loss will occur in the process of model training.

Therefore, based on the analysis of student behavior data in MOOC, we propose a MOOC dropout prediction method based on the time series model so that it can maintain temporal relationships between data and achieve a better prediction result. The model uses CNN to extract local abstract features, uses LSTM to extract temporal features, and integrates with Attention mechanisms to capture key features.

3. Prediction model

3.1. Problem description

The MOOC platform provides a variety of online courses, which often last for several weeks. Therefore, students will leave a lot of learning records in the process of interacting with the platform. These behavior records are saved in a time-stamped form, as shown in Table 1.

| enrollment_id | time          | source | event |
|---------------|---------------|--------|-------|
| 1             | 2014-06-17T15:19:28 | browser | problem |

We describe the dropout prediction problem as predicting whether a student will drop out of the course at a future time, given a student's behavior record over a period of time (as shown in Figure 1). Here, we define this period as a week. In other word, our purpose is to predict whether the student will drop out next week based on the student’s learning activity in the previous week. If a student has no learning activities in the next week, we then regard him or her as an instance of dropout. Specifically, for a certain student $x$, the log of the student's kth week in a certain course is $x_k$ and the dropout label is $y_k$, the goal task is to predict the behavior label $y_{k+1}$ of the student in the k+1 week (as shown in Figure 2).

![Figure 1. Definition of dropout problem.](image-url)
3.2. CNN-LSTM-ATT

We construct a dropout prediction model based on the time series model named CNN-LSTM-ATT, as shown in Figure 3. First, we build a behavior feature matrix based on the student's log data and take it as input to the Convolutional Neural Networks (CNN). Through the processing of the convolutional layer and the max pooling layer, it can extract abstract features and reduce data dimensions. After receiving the feature sequence output by CNN, use Long Short-Term Memory (LSTM) to process the temporal information of the data. Then we integrate with Attention mechanism to recalculate the hidden state output by the LSTM layer. It makes the model more focused on information useful to the prediction result. Finally, the prediction result will be obtained through continuous training of the model.

3.2.1. Extract abstract features and reduce feature dimensions.

The CNN receives the student's behavior feature matrix, and outputs the results to the LSTM after the processing of convolutional layer and pooling layer. The CNN is characterized by local connectivity, shared weights and pooling. The combination of these reduces the number of parameters and improves the training efficiency of the model. And compared with machine learning algorithms, CNN has a stronger ability to extract abstract features.

The convolutional layer is used to extract local abstract features of students' behavior. It includes two steps. (1) Filtering input student behavior feature matrix with the convolution kernel and add the offset to output a new feature vector. (2) Using the ReLU activation function to process the filtered output result of the convolution kernel. The ReLU function calculation formula is as follows.

$$ReLU(x) = \begin{cases} 
  x, & x > 0 \\
  0, & x \leq 0 
\end{cases}$$  

(1)

Use max pooling to take the point with the largest value in the local feature, so as to retain the most significant feature. That is, to filter out the characteristic information that has the greatest impact on students' dropout behavior. Through the processing of max pooling layer, the number of parameters and the data dimension can be reduced. At the same time, in order to further prevent the occurrence of overfit, a Dropout layer is added after the CNN. For parameter updates during training, a certain percentage of input neurons are randomly disconnected.

3.2.2. Process temporal information. The LSTM receives the feature sequence processed by the CNN, and uses its special gate structure to keep the students' historical behavior information persistent. It
conditionally changes the state of the unit through the combination of "forget gate", "input gate" and "output gate" so that it can control the flow of information through the storage unit.

First, select the information to be discarded from the unit through the forget gate. It makes a nonlinear mapping \( \sigma \) to the previous output \( h_{t-1} \) and the current input \( x_t \), and then outputs the vector \( f_t \). The value of each dimension of the vector is between 0 and 1. "1" means retain, while "0" means abandon. The calculation formula is:

\[
f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)
\]

where \( W \) represents weight and \( b \) represents deviation.

Then, determine the information that needs to be stored in the unit through the input gate. It mainly updates the unit \( C_t \). The structure consisting of the sigmoid layer that calculates the updated value, the output \( i_t \), and the tanh layer that creates a new candidate value vector \( C_t^\prime \) is added to the unit. The calculation formula is:

\[
C_t = f_t \cdot C_{t-1} + i_t \cdot C_t^\prime
\]

where \( f_t \) is the forgotten information at time \( t \), \( C_{t-1} \) is the unit information at the previous time, and \( C_t \) is the unit information at the current time. The calculation formulas of \( i_t \) and \( C_t^\prime \) are as follows:

\[
i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)
\]

\[
C_t^\prime = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)
\]

Finally, the output information is determined through the output gate. Multiply the output of the sigmoid layer with the unit state processed by tanh, and output \( h_t \). The calculation formula is:

\[
h_t = o_t \cdot \tanh(C_t)
\]

where

\[
o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)
\]

3.2.3. Strengthen important information. Attention mechanism is used to calculate the attention score from the hidden state output by the LSTM layer so that it enables the model to focus on more useful information for predicting student dropout results. For input information, the attention mechanism calculates its attention distribution and performs a weighted average of the value. In this way, different weights can be assigned to the input information, and the accuracy of the prediction model will also be improved to a certain extent.

For the \( t \)-th input information, its attention distribution (probability distribution) \( \alpha_t \) is:

\[
\alpha_t = p(z = t|X,q) = \text{soft}\max(s(X_t, q)) = \frac{\exp(s(X_t, q))}{\sum_{1 \leq t \leq T} \exp(s(X_t, q))}
\]

Where \( q \) is a query vector related to the task, and \( X_t \) represents the \( t \)-th input information. \( s(X_t, q) \) calculates as follow:

\[
s(X_t, q) = X_t^T q
\]

At the same time, a "soft" information selection mechanism is adopted for the encoding of the input information, and the calculation formula is:

\[
\text{att}(X, q) = \sum_{t=1}^{N} \alpha_t X_t
\]

Finally, use the sigmoid function to output the model prediction results. The output value is between 0 and 1, indicating the probability of students’ dropout. The corresponding calculation formula is:

\[
sigmoid(x) = \frac{1}{1 + e^{-x}}
\]

4. Experimental results

4.1. Dataset
The dataset used in this paper comes from KDD Cup2015\(^1\). This dataset is selected from the real data records of students' behavior in 39 courses in the "XuetangX", with a total of 79186 registered users and 120542 activity records. Among them, there are 95,581 dropout records, and the average dropout rate reached 79.29%. The course period of each course is approximately five weeks. There are seven different types of student learning behaviors, namely as “navigate”, “access”, “problem”, “close page”, “video”, “discussion” and “wiki”. This paper will build a dropout prediction model based on these seven different types of behavior data.

4.2. Evaluation metric
In traditional classification models, precision is often used as an evaluation metric. Considering the dataset used in our experiment, the number of dropouts is 95,581, and the number of students who do not drop out is 24,961. The imbalance of the data is very obvious, so the AUC is a more appropriate choice.

The receiver operating characteristic curve (ROC) takes "true positive rate" (TPR) as the vertical axis and "false positive rate" (FPR) as the horizontal axis. The definitions of TPR and FPR are as follows:

\[
TPR = \frac{TP}{TP + FN} \quad (12)
\]
\[
FPR = \frac{FP}{TN + FP} \quad (13)
\]

where TP (true positive), FP (false positive), TN (true negative), and FN (false negative) are represented by the confusion matrix of the dropout prediction results in Table 2.

| True label | Prediction label |
|------------|------------------|
| Dropout    | TP               |
| Not Dropout| FN               |

4.3. Results and analysis
In order to validate the performance of the CNN-LSTM-ATT prediction model proposed in this paper, three machine learning models were selected as the baseline model, namely Logistic Regression (LR) [1], Support Vector Machine (SVM) [2] and Decision Tree (DT) [10]. At the same time, compared with the single time series model LSTM [9] and the CNN-LSTM model without attention mechanism, the experimental results are shown in Figure 4 and Figure 5.

\[\text{Figure 4. Precision values of different models.}\]

\[\text{Figure 5. AUC value of different models.}\]

\(^1\) http://www.kddcup2015.com
From the above figures, we can see that:

1. Compared with the existing machine learning model, the proposed CNN-LSTM-ATT model performs better, increasing AUC by about 4%. Due to the small number of registered participants in the first two weeks of the course, the performance of LR, SVM and DT are not particularly good. However, as the course progresses, the number of students is also increasing, and the prediction result is also improved. In the fifth week we see the performance degradation again because the course is about to end and the students' behavior logs were also reduced. Therefore, for LR, SVM and DT, sufficient data is necessary to ensure the prediction performance of the model.

2. Compared with the single temporal model LSTM, the CNN-LSTM-ATT model takes the local information and temporal information in the student behavior data into account, and increases AUC by about 3%.

3. Compared with the CNN-LSTM model, the CNN-LSTM-ATT model with Attention focuses more on the information useful for the prediction results, strengthens its connection with the original data, and increases the AUC by about 1%.

5. Conclusions
In order to solve the problem of student loss in MOOCs, a dropout prediction model, CNN-LSTM-ATT based on time series, is proposed. Compared with the existing model, this model simplifies the feature selection process, maintains the temporal information, and strengthens the connection with the original data, thereby further improving the prediction performance of the model. Future research will comprehensively consider student information (such as gender, education level) and course information that affects students' dropping out, and then build a more accurate prediction model, so as to provide more reasonable suggestions for the implementation of personalized interventions on the MOOC platform to further improve students' completion rate of courses.

Acknowledgement
This research is supported by National Natural Science Foundation of China (61571238).

References
[1] Kloft M, Stiehler F, Zheng Z, et al. Predicting MOOC dropout over weeks using machine learning methods[C]//Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs. 2014: 60-65.

[2] Burgos C, Campanario M L, de la Pena D, et al. Data mining for modeling students’ performance: A tutoring action plan to prevent academic dropout [J]. Computers & Electrical Engineering, 2018, 66: 541-556.

[3] Al-Shabandar R, Hussain A, Laws A, et al. Machine learning approaches to predict learning outcomes in Massive open online courses[C]//2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017: 713-720.

[4] LIU T, Xiu L I. Finding out reasons for low completion in MOOC environment: an explicable approach using hybrid data mining methods [J]. DEStech Transactions on Social Science, Education and Human Science, 2017 (meit).

[5] Qiu J, Tang J, Liu T X, et al. Modeling and predicting learning behavior in MOOCs[C]//Proceedings of the ninth ACM international conference on web search and data mining. ACM, 2016: 93-102.

[6] Liang J, Li C, Zheng L. Machine learning application in MOOCs: Dropout prediction[C]//2016 11th International Conference on Computer Science & Education (ICCSE). IEEE, 2016: 52-57.

[7] Balakrishnan G, Coetzee D. Predicting student retention in massive open online courses using hidden markov models [J]. Electrical Engineering and Computer Sciences University of California at Berkeley, 2013, 53: 57-58.

[8] Wang W, Yu H, Miao C. Deep model for dropout prediction in MOOCs[C]//Proceedings of the
2nd International Conference on Crowd Science and Engineering. 2017: 26-32.

[9] Tang C, Ouyang Y, Rong W, et al. Time series model for predicting dropout in massive open online courses[C]/International Conference on Artificial Intelligence in Education. Springer, Cham, 2018: 353-357.

[10] Chen J, Feng J, Sun X, et al. MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine [J]. Mathematical Problems in Engineering, 2019, 2019.