Stopout Research Based on Behavior in Massive Open Online Courses
Wen YIN, Ming-hua JIANG*, Chang-long ZHOU and Li-ling ZHOU
Wuhan Textile University, Wuhan, China
*Corresponding author

Keywords: MOOC, Stopout, Behavior, Predicted, Logistic regression.

Abstract. In recent years, the loss of a large number of learners has attracted widespread attention in Massive Open Online Courses (MOOCs). A great deal of research has been modeling analysis and predicted based on the learning behavior. In this paper, according to the modules provided by moocroom, extract some related behavior, select the learners registered after the first three weeks of the data as the data set, using logistic regression modeling analysis, forecast for the next few weeks whether learners stopout, final assessment of the model performance. Experimental results indicate that the model established achieves high prediction accuracy, and the models attained the average AUC is 0.76.

Introduction
With the recent boom in development of Massive Open Online Courses (MOOCs), as a new form of online teaching has gained enormous popularity and attention. However, against a background of thriving enrollment, the completion rate of MOOCs is staggeringly low [1],[6]. So more concern is that the significant decrease in participation. Previous research indicates that fewer than 7% of learners who enroll in a MOOC actually complete it [2]. Learner stopout usually takes place by the second week of the course [3]. There are relevant research have analyzed the cause of stopout [4],[5]. One reason is that the online learning teachers can't see the learning situation of the learners in time, and the teachers can't get the learning feedback as the face-to-face teaching.

By analyzing the behavior of learners [7], it can well reflect the purpose, status and habits of learners, and provide effective feedback for teachers. This feedback allows the teacher to adjust the course in a timely manner to suit the learners’ rhythm. A potential interaction between teachers and students is achieved. It can not only improve the experience of MOOC, but also improve the learning efficiency.

In this paper, We put the moocroom[10] (A MOOC platform) learners as the research object, according to the specific function module extraction behavior of moocroom, using logistic regression to learn the first three weeks of the conduct of the modeling and analysis, predict whether learners stopout in the next 7 weeks and the average weight of various feature.

The remainder of this paper is organized as follows: A brief description of four types of behavior in moocroom is in Section 2. Training and evaluation models are presented by logistic regression in Section 3. The experiment results and evaluation through training is illustrated in Section 4. Finally, Section 5 concludes the paper.

Behavior Features in Moocroom
According to the activity frequency of each module in the moocroom, the learner activity can be divided into four categories: explore behavior, course behavior, interactive behavior and other behaviors.

1. exploring behavior: Look for courses you want to study or browse through a circle of interest in moocroom.
2. course behavior: the learner and the specific course learning related behavior.
3. interaction behavior: The interaction between a learner and a teacher or platform
4. other behavior: personal settings, account security, learning centers and other learning operations unrelated to the behavior.
Figure 1 shows the proportion of four types of behavior. Course behavior accounts for the largest proportion. It shows that the core of MOOC is the part of the course that best reflects the learner's motivation, purpose, status and habits. Therefore, this paper will extract the learner's behavioral features from the course behavior as model input.

![Figure 1. The proportion of four types of behavior.](image)

**Predicting Stopout Based on Course Behavior**

We made several assumptions to more precisely define the stopout prediction problem and interpret the data. These assumptions include time-slice delineation and defining persistence (stopout) as the event we attempt to predict.

1. **Time-slice and Stopout definition**
   Temporal prediction of a future event requires us to assemble explanatory variables along a time axis. This axis is subdivided to express the time-varying behavior of variables so they can be used for explanatory purposes. We decided to define weekly units as time slices. The duration of observation was 10 weeks after registration.

   The next question we had to address was our definition of stopout. We considered defining it by the learner’s last behavior in the course, regardless of the nature of the course. In any given week, the learner displays any course behavior and judges that the learner has persistence. Determining a learner's stopout means not showing any learning behavior during the week, such as logging in, but without any course behavior or no login.

2. **Features per learner**
   We extracted 15 interpretive features on a per-learner basis. These are the features we use to build a model. In this paper for the sake of brevity, we only list the features and their brief descriptions in the Table 1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Describe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Course behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of independent courses</td>
<td>Number of independent courses</td>
<td></td>
</tr>
<tr>
<td>Number of certification courses</td>
<td>Number of certification courses</td>
<td></td>
</tr>
<tr>
<td>Number of Live broadcast</td>
<td>Number of Live broadcast</td>
<td></td>
</tr>
<tr>
<td>Total course duration</td>
<td>Total time spent on the course video</td>
<td></td>
</tr>
<tr>
<td>Total Live broadcast duration</td>
<td>Total time spent on live broadcast</td>
<td></td>
</tr>
<tr>
<td>Number of course data browsing</td>
<td>the total number of course resources</td>
<td></td>
</tr>
<tr>
<td>Number of course data download</td>
<td>The total number of downloaded courses</td>
<td></td>
</tr>
<tr>
<td>Number of course notes</td>
<td>The total number of notes written</td>
<td></td>
</tr>
<tr>
<td>View note frequency</td>
<td>Browse the frequency of view notes</td>
<td></td>
</tr>
<tr>
<td>Number of click entry</td>
<td>The total number of click entry</td>
<td></td>
</tr>
<tr>
<td>Number of knowledge access</td>
<td>The total number of knowledge access</td>
<td></td>
</tr>
<tr>
<td>Number of finish the video</td>
<td>Number of finish Course video</td>
<td></td>
</tr>
<tr>
<td>Number of fast forward / rewind</td>
<td>Number of fast forward / rewind</td>
<td></td>
</tr>
<tr>
<td>Number of pause</td>
<td>Number of pause</td>
<td></td>
</tr>
<tr>
<td>Average playback speed</td>
<td>Average playback speed</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Self-extracted covariates.
The course behavior is represented by the C, so the explanatory variables are represented by C1, C2, ..., and Cn.

Training

Logistic regression is a commonly used binary predictive model. It calculates a weighted average of a set of variables, submitted as covariates, as an input to the logit function. Thus, the input to the logit function, Z, takes the following form:

\[ Z = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n \]  

(1)

Here, \( a_1 \) to \( a_n \) are the coefficients for the feature values, \( X_1 \) to \( X_n \). \( a_0 \) is a constant. The logit function, given by,

\[ g(z) = y_i = \frac{1}{1 + e^{-z}} \]  

(2)

takes the shape as shown in figure 2. Note that the function’s range is between 0 and 1, which is optimal for probability. Also note that it tends to ‘smooth out’ at extreme input value, as the range is capped.

![Figure 2. The logit (aka logistic or sigmoid) function.](image)

For a binary classification problem, such as ours, the output of the logit function becomes the estimated probability of a positive training example. These feature weights, or coefficients, are similar to the coefficients in linear regression. The difference is that the output ranges between 0 and 1 due to the logit function, rather than an arbitrary range for linear regression.

Corresponding binary label associated with the covariates. After training a model, the predicted probability, or the output of the logit function, should predict higher probabilities for the positive ‘+1’ class examples in the training data and a lower probability for the negative ‘0’ class examples.

There is no closed form solution to find the optimal coefficients to best fit the training data. As a result, training is usually done iteratively through a technique called maximum likelihood estimation. First, a random set of coefficients are chosen. At each iteration, an algorithm such as Newton’s method is used to find the gradient between what the coefficients predict and what they should predict, and updates the weights accordingly. The process repeats until the change in the coefficients is sufficiently small. This is called convergence. After running this iterative process over all of the training examples, the coefficients represent the final trained model.

Evaluation

With training in place, the next step is evaluating the classifier’s performance. A testing set comprised of untrained covariates and labels evaluates the performance of the model on the test data following the steps below:

1. The logistic function learned and presented in (2) is applied to each data point and the estimated probability of a positive label \( y_i \) is produced for each data point in test set.

2. A decision rule is applied to determine the class label for each probability estimate \( y_i \). The decision rule is given by:
Given the estimated labels for each data point \( g(z) \) and the true labels, we can calculate the confusion matrix, true positives and false positives and thus obtain an operating point on the ROC curve.

3. By varying the threshold \( \lambda \) in (3) the decision rule above we can evaluate multiple points on the ROC curve. We then evaluate the area under the curve and report that as the performance of the classifier on the test data.

**Prediction Result and Evaluation**

We applied logistic regression to student persistence prediction. We used the 15 interpretive features we described earlier in this paper to form the feature vectors, and maintained the stopout value as the label.

According to figure 3, take \( m=3 \), the flow is as follows:

1. Performed 10 fold cross validation on the training set. This involved training the model on 9 folds of the train dataset and testing on the last fold.
2. Trained a logistic regression model on the entire train dataset.
3. Applied the model to the test dataset by putting each data point through the model then applying the decision rule in (3) and following the steps to determine the AUC under the ROC.
4. Evaluating the model using mean cross validation ROC AUC and test set ROC AUC.
Conclusion

In Figure 4, these three factors (C4/Total course duration, C5/Total Live broadcast duration, C12/Number of finish the video) have a greater impact on whether learners have stopout in the next few weeks. In Figure 5, the maximum AUC is 0.88, and the average AUC is 0.76, which shows that the model is feasible for prediction classification. With the increase of the predicted week number, AUC will decrease. It shows that the model is better for short-term prediction. Of course, there are many parts of this paper that have not been investigated. Will it be better to use more historical data modeling? Or different models such as Hidden Markov model [8]? Whether the model applies to a different discipline, a different context, or a different MOOC platform [9]?

References


[8] www.moocroom.com