Modeling Student Learning Outcomes in MOOCs

Shuai Guo, Wenjun Wu
State Key Lab of Software Development Environment
Department of Computer Science and Engineering, Beihang University
Beijing, China
{guoshuai, wwj}@nlsde.buaa.edu.cn

Abstract—As an innovative teaching and learning platform, Massive Open Online Courses (MOOCs) have attracted widespread attention in recent years. Learning outcome assessment is of great significance in the field of traditional on-campus teaching and emerging online education. In this work, we take advantage of the large-scale data offered by MOOCs to propose a two-stage analytic model for estimation of learning outcome: first, based on learning behavior, we preliminarily predict student mastery level of knowledge; second, using homework performance in combination with the prediction results obtained in the first stage, we build an individualized knowledge tracing model to accurately assess student knowledge status. Our research not only gains an insight into student learning patterns in MOOCs, but also achieves high accuracy in assessment of learning outcomes.

Keywords—MOOCs; Learning Outcome; Learning Behavior Analysis; Individualization; Knowledge Tracing

I. INTRODUCTION

Massive Open Online Courses (MOOCs) have enjoyed significant limelight in recent years, both in academia and industry. MOOC providers such as edX, Coursera and XuetangX boast hundreds of courses developed by top-tier universities, which consist of various learning resources, like videos, lectures, forum, homework assessments and online exams. MOOCs provide several substantial advantages to educational researchers, most notably the detailed digital trail left by students in the form of log data and the size of the student cohorts, which are often several orders of magnitude larger than typical on-campus offerings [1].

Whether in the field of traditional campus teaching or open online education, learning outcome assessment is of great significance. Accurate evaluation of students learning outcomes not only gives students equitable course scores, assists instructors in finding difficulties and confusion in study to enhance the learning experience, but also provides an opportunity to personalize instruction and exercises for different students.

In the same way as the traditional education, MOOCs can take the midterm and final exams to evaluate students' overall learning outcomes of a course. However, with large-scale learner communities and rich behavior data, MOOCs are able to provide more detailed and accurate assessment of student performance. In this paper, we segment course content into chapters according to the goal of mastering knowledge skills, thus developing a correspondence between a chapter and a specific knowledge skill. Then we establish a two-stage model to assess student mastery level of each knowledge skill, in other words, to estimate the probability that a student has learned the specific knowledge of a chapter. Specifically, we first analyze student learning activities within a chapter, such as video-watching clickstream, page-view records and forum interactions, extract interpretable quantities to predict the probability that a student has mastered the knowledge of that certain chapter; second, we combine the students' performance data on homework problems with the results obtained in the first stage, build sequential models to accurately assess student learning outcomes. To our best knowledge, this is the first comprehensive effort of modeling student outcomes, which makes full use of student learning and practicing data and produces knowledge-specific outcome assessment in the context of MOOCs.

The rest of this paper is organized as follows: We begin with the brief review of related work. Section 3 describes the dataset used in our research. In Section 4 and 5, we detail the two-stage analytic model for assessment of learning outcome. Section 6 presents the prediction results and experimental findings. Finally, we conclude our work in this paper.

II. RELATED WORKS

Various research works have studied student outcome modeling in traditional and online education. Especially the knowledge tracing model [2], which was introduced by Corbett and Anderson in 1995, has become the dominant method of modeling student outcomes. The model utilizes practice performances to maintain an estimation of the probability that the student has learned the certain skill. It is currently employed by the cognitive tutor, used by hundreds of thousands of students, and many other intelligent tutoring systems to predict performance and determine when a student has mastered a particular knowledge skill. Reye [3] showed that the formulas used by Corbett and Anderson in their knowledge tracing work could be derived from a Hidden Markov Model or Dynamic Bayesian Network (DBN). Many other works were introduced to improve the prediction accuracy of the standard knowledge tracing model by setting model parameters manually [4], using the contextualization of parameters [5] or individualizing the parameters [6]. But extremely few investigators exploit other sources of data, like students’ behaviors on learning resources, except for practice performances in terms of knowledge tracing.
There has been a lack of work studying outcome assessment for MOOCs. Most of MOOCs studies have used students’ behaviors to understand issues of high dropout rate in MOOCs. Kloft [7] predict dropout from click-stream data using a Support Vector Machine (SVM). Balakrishnan [8] extracts features mainly from discussion forums and video lectures, and employs Hidden Markov Models (HMMs) to predict student retention. Halawa, Greene, and Mitchell [9] study accurate and early dropout prediction using student activity features capturing lack of ability or interest. Various other factors have also been investigated, such as demographics [10], social positioning in forums [11], and sentiment in forums [12]. Jiang and his colleagues [13] uses a combination of students’ Week 1 assignment performance and social interaction within the MOOC to predict their final performance in the course. However, none of above work utilizes knowledge tracing methodology to study outcome assessment for MOOCs.

In our work, we combine students’ learning behaviors and practice performances to evaluate student outcomes. Specifically, we use a modified knowledge tracing model, which is individualized by learning behaviors, to achieve higher estimate accuracy.

III. DATASETS

The data for our study comes from our own MOOCs platform Beihang Xuetang (http://mooc.buaa.edu.cn), which is set up upon Open edX. Open edX is the open-source release of the edX platform developed by founding partners Harvard and MIT. Nowadays, there are nearly one hundred universities and institutions hosting their own instances of Open edX to offer online classes.

We have already released 12 MOOCs on Beihang Xuetang. The specific layout of each MOOC varies, but most follow a similar format. Course content is sectioned into chapters, usually using weeks as intervals. Each chapter is composed of lecture videos, homework assignment and a peer grading activity consisting of ten chapters and a final exam. Each chapter is composed of lecture videos, homework assignment and a peer grading activity consisting of ten chapters and a final exam. Every MOOC includes exams, a forum and a Wiki as well.

In this work, we analyze a MOOC titled "The Experiment of Computer Network" hosted on Beihang Xuetang from March 2015. The duration of the course was eleven weeks, and comprised of ten chapters and a final exam. Each chapter includes a number of videos: the majority (6/10) had 7, while the most one has 11 and the least has 4, for a total of 68 videos. Over these 68 video clips, the average length is 9.61 min (standard deviation = 4.25). To supplement the videos, we created in-chapter homework assignment to test student understanding with the specific knowledge of a chapter. The size of the homework assignment ranges from 5 problems to 13, for a total of 91 problems. It is worth noting that the assignment not only played a role of testing, but also taught students again, because when students answered wrong, there would come up immediate and detailed feedback.

The course attracted more than 2000 student enrolled, of which there were almost 600 active users. Finally, 512 students passed the exam and received a certificate of completion for the course. Since every user interaction with the website (click, view, submit and post) was recorded by the logging system, the session of the course created a rich dataset, containing 1220446 log entries. Each log specifies the user, URL, event type, UNIX time and other identifiers like video ID. The original dataset could be generally classified into two types: clickstream data and forum data, of which the clickstream data has the page-log and the video-watching log, the forum data included the content of the forum post and comment. According to statistics, each active student watched 28 lecture videos and submitted 31 problems on average: the course forum accumulated 417 posts and 1855 comments. The abundant data presented an opportunity to investigate how to assess student outcome accurately and finely grained.

IV. PREDICTIVE MODELING USING LEARNING BEHAVIOR

This is the first stage of our two-stage model. In this stage, we attempt to preliminarily predict student outcome using learning behavior. As previously mentioned, our model provides knowledge-specific outcome assessment after we treat each chapter as a complete knowledge skill. Thus we analyze student learning activities within a chapter comprehensively to predict the probability of knowledge master level in that certain chapter. We cast the problem as a supervised binary classification task where class labels, namely student outcomes, are 1 if students get more than 60% of the homework problems correctly and 0 otherwise. Then, the following essential task is to construct predictive features and extract the feature values per student in each chapter.

A. Feature Construction

Based on the student abundant learning behavior data, we aim to design sophisticated interpretive features hypothesized to be predictive of outcome. Since video-watching is the most common and extensive behavior among student learning activities, six relevant features are designed as follows:

1) Total playing time: Total time a student spends playing the videos within a chapter.
2) Number of video played: The number of distinct videos that a student plays within a chapter, not counting repeated ones.
3) Number of pauses: The number of times a student jumps backward in the videos within a chapter.
4) Number of fast forwards: The number of times a student jumps forward in the videos within a chapter.
5) Number of slow play rate use: The number of times a student slows down the play rate. The player allows rates between 0.75x and 1.5x the default speed.

These characteristics can represent students’ learning attitudes and motivation to some extent. The total playing time and number of video played reflect students’ engagement in courses, while the number of rewinds reflects students’ learning initiative. Eventually, both attitude and motivation of every student can affect his learning outcomes.
Besides video-watching, website-browsing and forum-interaction are also important learning activities. So we extract four more feature as follows:

7) Number of requests: Total number of requests for a student to interact with courseware including page views and video clicks within a chapter.

8) Number of sessions: The frequency of students logging into the learning platform within a chapter.

9) Number of forum posts: The number of forum posts that a student makes within a chapter.

10) Content length of forum posts: The total word length of forum posts within a chapter.

Note that some of these quantities are affected by video length and count, which are varied across chapters, so it is necessary to conduct normalization: feature total playing time (id 1) is normalized by the total video length of the corresponding chapter; feature id 2–8 are normalized by the total video count. Furthermore, these features are not independent of each other, but each can tell us a unique aspect about learner behaviors.

B. Feature Analysis

For each feature, we perform two groups of analysis. First, we depict the student distribution over feature values to reveal how students learn in MOOCs. Second, we calculate the correlation coefficients between homework correctness rate and feature value to show how strongly predictive a feature can be and filter those features that contribute little to prediction.

Fig. 1 depicts the user distribution over different features. For total playing time (normalized by total video length), the students show a bimodal-like distribution. The users who play less than 20% of lecture video present a spike, representing the cohort with rarely or no participation in course after registration. This reflects the characteristics of freedom, self-restraint and high dropout rate in MOOCs. Except those students, the remaining active students spread similarly to normal distribution: most students complete 80% of lecture video; minority students study particularly seriously, whose cumulative video playing time exceeds the total video length for repeated video playing. For the number of requests and the number of forum posts (normalized by total video count), the distributions are both similar to the one over total playing time that except inactive users, the remainders are approximate normal distribution. The case for the number of forum posts is a bit different, which suggests that a few forum-active users contribute a majority of forum posts.

Fig. 2 depicts the relation between homework correctness rate and feature values. For total playing time, the most important learning activity, homework correctness rate presents a linear trend with it, which indicates the capability of predicting outcome. Moreover, the correctness rate presents an even stronger linear trend with number of rewinds, as it is a more important measure of students' learning initiative and seriousness. For other features, the correlation coefficients with homework correctness rate are show in table 1. Note that the coefficient value of number of request and number of forum posts are so small that we decide to remove those two from our feature set. So the final feature set contains 8 members.

V. KNOWLEDGE TRACING USING HOMEWORK PERFORMANCE

Online learning behaviors have an indirect effect on student outcomes such that we can preliminarily estimate their mastery level of knowledge based on learning characteristics. On the other hand, the student performances in homework assignment directly reflect their learning outcome, which allows us to accurately determine student knowledge status through student responses to homework problems.

One straightforward way to utilize response data is to take homework correctness rate as the metrics of mastery level. However, the overall correctness rate loses the sequential information of responses. For example, there are two students A and B whose response sequences are \{1,1,0,0\} and \{0,0,1,1\},...
respectively. Their correctness rates are both 50%, but their learning effects are not the same, especially in MOOCs context where homework practice not only plays a role of testing, but also teaches students once again with feedback when they get wrong answers. Considering the sequential information, it is more likely that student B gradually acquires the knowledge through practice, while student A may have guessed the first two questions but still fails the last two problems, therefore the learning outcome of student A should be not as good as student B.

The above qualitative analysis shows the response sequential information has a certain influence on the evaluation of learning effects. The knowledge tracing model, popularized in Intelligent Tutoring System (ITS), does leverage the response sequence to assess student learning outcomes. Our work adopts the knowledge tracing model and ameliorates it by combining the prediction results obtained in the first stage of our two-stage-model

A. Standard Knowledge Tracing Model

Knowledge tracing is the predominant method for modelling student knowledge and learning over time. It is based on a 2-state dynamic Bayesian network where student responses are the observed variable and student knowledge is the latent one. The model takes student response sequences and uses them to estimate the student’s level of knowledge. There are two performance parameters: guess and slip, which mediate student knowledge and student performance respectively. The guess parameter $P(G)$ represents the fact that the student may sometimes generate a correct response in spite of not knowing the correct skill. The slip parameter $P(S)$ acknowledges that even students who understand a skill can make an occasional careless mistake. There are also two learning parameters. The first is initial knowledge $P(L_0)$, the likelihood the student knows the skill at his first answer. The second is the learning rate $P(T)$, the probability a student acquires a skill as a result of an opportunity to practice it.

The use of knowledge tracing has two steps, one step in which the four parameters are learned often through the expectation maximization method [14], conjugate gradient search [2], or discretized brute-force search [5], and the other step where an individual student’s knowledge is inferred from his responses. The knowledge status update algorithm is defined as follows:

\[
\text{if } \text{performance}_n = \text{correct} \\
P(L_n|\text{correct}_n) = P(L_n) * P(S) + P(L_n) * P(S) + P(L_{n-1}) * P(G) \\
\text{else} \\
P(L_n|\text{incorrect}_n) = \frac{P(L_n) * P(S) + P(L_{n-1}) * P(S)}{1 - P(L_{n-1})} + P(L_{n-1} | \text{evidence}_n) * P(T)
\]

The $P(L_n)$ is the prior probability of knowledge at that time, while $P(L_n | \text{evidence}_n)$ is the posterior probability of knowledge calculated after taking an observation at that time into account. Since the student will be presented with feedback, there is a chance to learn, which is formulated by the last formula.

B. Individualized Knowledge Tracing Model

The four parameters of standard knowledge tracing model are skill-specific, that is, all students share the same probabilities of guess, slip, initial knowledge and learning rate under the same skill. Even though it seems that students differ in this regard. A number of research results strongly suggest that student-specific variability could enhance the model accuracy. Therefore, we take full advantage of MOOCs characters to individualize the standard knowledge tracing model by means of assigning every student an initial knowledge level which is predicted based on his learning behavior.

Though the four parameters could all be individualized, the model we present in this work focuses only on individualizing the prior knowledge parameter. Prior knowledge and initial knowledge are synonymous. The difference with the standard knowledge tracing model is the ability to represent a different prior knowledge parameter per student. Knowledge tracing model is a special case of this prior per student model and can be derived by fixing all the priors to the same values.

Fig. 3 shows the topologies of both models. The individualization of the prior is achieved by adding a student node before the first knowledge state node. The student node can take on values ranging from one to the number of students being considered. The conditional probability table of the initial knowledge node is therefore conditioned upon the student node value, which has been estimated for each student based on his unique learning behavior in the first stage of our two-stage model. The student node itself also has a conditional probability table associated with it which determines the probability that a student will be of a particular ID. The
parameters for this node are set to be 1/N where N is the number of student. Since the student node is an observed node in the individualized model, the parameter values for this node are fixed and never need to be inferred.

VI. EXPERIMENTAL RESULT

A. Accuracy of Predictive Modeling in Stage 1

In this section, we conduct experiments to evaluate the effectiveness of learning outcome prediction based on learning behavior. As discussed in Section 4, we treat this problem as a supervised binary classification task. The class labels are determined by homework correctness rate, specifically, labels are 1 if students correctly finish more than 60% of the homework problems and 0 otherwise. After being filtered by correlation analysis, the feature set is left with eight major attributes including total playing time, number of video played, number of rewinds, number of pauses, number of fast forwards, number of slow play rate use, number of sessions and content length of forum posts. Three different classification models are adopted in our experiments, Linear Discriminant Analysis(LDA), logistic regression(LR), support vector machine(SVM), which are all popular models both in academia and industry.

We extract the eight feature quantities per student within a chapter. Therefore, for each student-chapter pair, it is an 8-dimensional feature vector representing the student learning behavior within that certain chapter. In addition to the feature vector, the homework correctness rate is also calculated for each student-chapter pair in order to get the class label. In total, there were 1220446 logs, with 12596 student-chapter pairs. But we removed the ones that did not submit any homework problem at all, bringing the total to 6126. From the remainder, we eliminated 476 entries that were either abnormal or error. In the end, we were left with 5418 pairs, of which 1642 pairs are positive instances and 3776 pairs are negative instances.

The whole dataset is split into a training set and a test set according to the ratio of 4:1. During the model training phase, we used k-fold cross validation, which randomly divides the training dataset into k folds: some are used in constructing a model, and others are used to evaluate the model. Cross validation tests a predictive model without the test dataset, thus helping to gauge whether a model might be over-fitting. We set k = 5 in our evaluation. During the model evaluation phase, because of data imbalance, not only the prediction accuracy was utilized to evaluate models, but also the metrics of recall, precision and AUC (area under the ROC curve), which are broadly used to deal with imbalance. The experimental result is presented in table 2.

The highest AUC achieves 0.783 with accuracy of 0.762. As the first stage of our two-stage model, it achieves the expectation of preliminarily predicting student outcome and producing the initial knowledge level estimation for each student. What's more, while we constructed models with different techniques, we found the consistent accuracy arising across techniques which was dependent on the features we used. Generally speaking, using more informative features can yield superior accuracy that was consistent across modeling techniques. It illustrates that taking the extra effort to extract complex predictive features rather than trying multiple modeling techniques is the more important contributor to successful MOOC data science.

B. Validity of Knowledge Tracing in Stage 2

After preliminarily predicting the knowledge mastery probabilities for each student within a chapter, we incorporated them into the general knowledge tracing model to individualize the initial knowledge parameter. In this section, we design experiments to test the real world utility of the individualized knowledge tracing model, and compare to the baseline model and the standard knowledge tracing model.

The experimental dataset consists of student responses to homework problems of each chapter. We had 10 chapters in total, that is 10 homework assignments. Only the problems within the homework assignment with the exactly same skill tag were used. Each homework assignment had an average of 427 students and each student completed an average of six homework assignments.

We used the students' responses to the last problem of each homework as test data, while the other responses as training data. For each homework, we trained three separate models: the baseline model, the standard knowledge tracing model and the individualized knowledge tracing model. The baseline model directly utilized the correctness rate of the training data to predict the test data. For knowledge tracing models, we first constructed the appropriate sized dynamic Bayesian network, where the size of student nodes corresponded to the number of students completing homework. Then the parameters were learned using Expectation Maximization (EM) algorithm. Note that the initial values for learning rate, guess and slip parameters of both models were set randomly, while the difference lay in the prior knowledge parameter: the standard model set the single parameter randomly and allowed it to be adjusted by EM, but the individualized model set a prior parameter per student based on the prediction results obtained by learning behavior and fixed them during EM.

After the parameters were learned, we estimated the

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Models</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.758</td>
<td>0.513</td>
<td>0.832</td>
<td>0.769</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.762</td>
<td>0.598</td>
<td>0.791</td>
<td>0.783</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.760</td>
<td>0.554</td>
<td>0.814</td>
<td>0.778</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 4 Prediction accuracy of three different models](image)

TABLE 2 PREDICTION RESULT USING LEARNING BEHAVIOR
performance of every student by entering the responses as evidence to the Bayesian network, except the last one. Then the probability of the student answering the last question correctly was computed and binarized by a certain threshold so that it could be compared to the actual response later. Fig. 4 shows the prediction accuracy of three different models.

As is shown in Fig. 4, both knowledge tracing models outperform the baseline model at least slightly for all 10 homework assignments, which was due to the use of sequential information in student responses. The individualized knowledge tracing model predicts more accurately than standard model in 9 out of 10 homework assignments. To dig deeper into this result, we found that the improvement was more significant when there were less problems in the homework, like the cases of chapter one with 6 problems and chapter nine with 5, while the improvement was marginal for homework with relatively more problems, like chapter five with 11 problems and chapter seven with 13 problems. That was because the effect of the prior knowledge parameter gradually diminished in the process of iteration. In summary, the average accuracy of the individualized knowledge tracing model reaches as high as 87%, which demonstrates that the proposed two-stage model is of high effectiveness in learning outcome assessment.

VII. CONCLUSION AND FUTURE WORK

Assessment of student learning outcome is an intriguing research area, and especially so in MOOCs context because of its potential benefits. In this paper, using data from one of our own MOOCs, we proposed a two-stage model: preliminary prediction based on student learning behavior features and accurate evaluation based on homework response sequences. Through experiments, we found that our analytic model outperformed the baseline model and standard knowledge tracing model under almost all test datasets. Also, we obtained a certain understanding of student learning patterns in MOOCs by analyzing behavioral characteristics.

There remain a number of issues to be investigated in future work. Besides summary quantitative features, there are other types of learning features worth considering, such as semantic information of forum posts and high-level click patterns of video-watching behavior. For knowledge tracing modeling, we only focused on the individualization of the prior knowledge parameter, while the other three parameter of knowledge tracing could be individualized as well. It would be interesting to explore these parameters and examine their effects on further improvement of prediction accuracy. Moreover, we plan to incorporate our analytic schemes into our own MOOC platform to provide personalized practice according to student learning outcomes in the future.

ACKNOWLEDGMENT

This work was supported in part by grant from State Key Laboratory of Software Development Environment (Funding No. SKLSDE-2013ZX-03) and National High-Tech R&D Program of China (Grant No. 2013AA01A210).

REFERENCES