MOOC Prediction Analysis and Pattern Discovery

Zhouxiang Cai
Co Tran
Rachel Witner

Mentor:
Prof. Huzefa Rangwala
Overview

- Motivation/background
- The data
- Co Tran’s work
- Rachel’s work
- Zhouxiang’s work
ENGAGE IN MORE THAN A COURSE:
ENGAGE IN A LEARNING EXPERIENCE.

Enroll in open, online courses from colleges, universities, and organizations worldwide.
Motivation

● How to define a successful MOOC?
● How to define course completion?
  ○ 38% of user-course pairs were active for at least one week
  ○ ~9% of students were “active” for at least half of the weeks of a course
  ○ 0.08% of student enrollment logs had a date of completion
  ○ 37% of computed final scores were missing
  ○ 45% of non-missing computed final scores were 0
● How to define engagement?
● Are there recurring patterns of interaction across courses, users, and time?
The Data

- Canvas Network open courses released by Harvard Dataverse
- January 2013 to July 2016
- ~380 courses
- ~400,000 students enrolled
- User page views (requests)

A sample of the star schema structure
Typology of learning behaviors of students in online courses - Co Tran

Motivation : - Previous studies of student learning pattern researched on the sample size of 1 or 2 courses.

Objective : - Studying the learning behaviors of students in the scale of multi-course using cluster analysis.
Method

Course-wise features computing normalized by the number of students

Courses cluster analysis (Hierarchical, K-mean, DBSCAN) - extract the characteristics of courses

Choose a cluster based on low variance features and distinctive characteristics

Extract the student ids in the cluster and computed page views by content (requests) features.
Divide the page views by quartile time intervals (0-25%, 25-50%, ..., And normalize by the length of course,

All time behaviors: What are the consistent learning behaviors through the courses?

Time dependent behaviors: What are the learning behaviors at each time period? How do they change through the courses?
All time behaviors approach - Characteristics and student outcomes explained

Cluster 1: high engagement in discussion and reading wiki pages and higher average score.

Cluster 2: low engagement in discussion and reading wiki pages, high engagement in assignment and low average score.

Cluster 3: has low engaging in every activity especially in assignment, discussion, and reading wiki pages.
All time behaviors approach - Characteristics and student outcomes explained
Results - Time dependent behaviors approach

Cluster 1: normal engagement in all activities.
Cluster 2: low engagement in all activities.
Cluster 3: high engagement in all activities
Changes in the memberships of clusters

**Jaccard similarity coefficient:** The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]  

(3)

<table>
<thead>
<tr>
<th>group</th>
<th>cluster_1</th>
<th>cluster_2</th>
<th>cluster_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>group_25_50</td>
<td>1.0</td>
<td>0.928501193988</td>
<td>0.0468164794007</td>
</tr>
<tr>
<td>group_50_75</td>
<td>0.999194414608</td>
<td>0.887533683166</td>
<td>0.1053227633070</td>
</tr>
<tr>
<td>group_75_100</td>
<td>0.999194414608</td>
<td>0.866238401142</td>
<td>0.1274418604651</td>
</tr>
<tr>
<td>group_100_25</td>
<td>1.0</td>
<td>0.892426367461</td>
<td>0.0179257362356</td>
</tr>
<tr>
<td>group_25</td>
<td>3721</td>
<td>6906</td>
<td>238</td>
</tr>
<tr>
<td>group_50</td>
<td>3721</td>
<td>6823</td>
<td>321</td>
</tr>
<tr>
<td>group_75</td>
<td>3724</td>
<td>6486</td>
<td>655</td>
</tr>
<tr>
<td>group_100</td>
<td>3721</td>
<td>6587</td>
<td>557</td>
</tr>
</tbody>
</table>

Table 7: Jaccard coefficient and number of students in each cluster
Interesting findings

- Cluster 1 in both approaches has the same memberships
  
  The learning behaviors of students in cluster 1 are mostly the same in each intervals of time as well as throughout the courses.

- The exchanges in memberships of time dependent behaviors approach mostly appear in cluster 2 (low engagement) and cluster 3 (high engagement).
Data Visualizations
Data Visualizations

Frequency of users per country

[Map showing user frequency]
Data Visualizations

Submissions (green) and due dates (red) for 2 randomly selected individuals

Course discipline
- Business and Management
- Education
- Life Sciences
- Medical Pre-Medical
- Professions and Applied Sciences
- Social Sciences

num submissions
- 1
- 2
- 3
- 4
- 5
- 6

Date (mm-yy)
Data Visualizations

Distribution of Courses per Student

Avg % of quizzes a student completed per course vs. courses per student
Data Visualizations

Quizzes per week per course

<table>
<thead>
<tr>
<th>Business and Management</th>
<th>Computer Science</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Mathematics &amp; Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Visualizations

Avg Quiz Completion Rate vs Quizzes Per Week

Avg % of students who completed each quiz

Avg quizzes per week

Duration (weeks)

100
75
50
25
## Weekly Interaction Clustering

<table>
<thead>
<tr>
<th>user id</th>
<th>course id</th>
<th>week</th>
<th>count social</th>
<th>count quiz subs</th>
<th>count other</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>876763763</td>
<td>3425142</td>
<td>23</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>876763763</td>
<td>3425142</td>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>876763763</td>
<td>9812343</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>892332345</td>
<td>1434241</td>
<td>57</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>?</td>
</tr>
</tbody>
</table>

... ... ... ... ... ...

\( n = 480,000 \)

---

**CLARA (Clustering Large Applications)**

- Draw a random sample D’ from the original dataset D
- Apply PAM (partitioning around medoids) algorithm to D’ to find the k medoids
- Use these k medoids and the dataset D to calculate the current dissimilarity
- If it is smaller than the one you get in the previous iteration, then these k medoids are kept as the best k medoids
- The whole process is performed a specified number of times
- In this case, I used 5,000 samples of size 10,000
Weekly Interaction Clustering

The “elbow” in the plot suggests an optimal number of clusters, as this is the point where each additional cluster only reduces SSE by a small amount.
## Clustered Weekly Course Interactions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Social Interactions</th>
<th>Quiz Submissions</th>
<th>Non-assessment Activities</th>
<th>Label + Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.99</td>
<td>0.217</td>
<td>3.72</td>
<td>M – moderate activity in all three features</td>
</tr>
<tr>
<td>2</td>
<td>1.13</td>
<td>0.440</td>
<td>1.00</td>
<td>L – low activity in all three features</td>
</tr>
<tr>
<td>3</td>
<td>4.15</td>
<td>1.94</td>
<td>0.00</td>
<td>S – mostly social interaction; no non-assessment activity</td>
</tr>
<tr>
<td>4</td>
<td>1.50</td>
<td>10.7</td>
<td>0.057</td>
<td>Q – most quiz submissions</td>
</tr>
<tr>
<td>5</td>
<td>0.286</td>
<td>0.252</td>
<td>2.00</td>
<td>A – moderate non-assessment; low quiz and social</td>
</tr>
<tr>
<td>6</td>
<td>0.0146</td>
<td>0.239</td>
<td>8.74</td>
<td>N – mostly non-assessment activity</td>
</tr>
<tr>
<td>user id</td>
<td>course id</td>
<td>week</td>
<td>count social</td>
<td>count quiz</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>876763763</td>
<td>3425142</td>
<td>23</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>876763763</td>
<td>3425142</td>
<td>25</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>876763763</td>
<td>9812343</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>892332345</td>
<td>1434241</td>
<td>57</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Creation of interaction string for each user-course. For weeks with no interaction, E represents ‘enrolled in course but didn’t interact with it’ and O represents weeks before or after the course’s official start/end dates.

<table>
<thead>
<tr>
<th>user id</th>
<th>course id</th>
<th>engagement string</th>
</tr>
</thead>
<tbody>
<tr>
<td>876763763</td>
<td>3425142</td>
<td>OOOEEEEEEEMLQSEAENOOOOO...</td>
</tr>
<tr>
<td>876763763</td>
<td>9812343</td>
<td>OOOQEEANEEOOOOOOOOOOOO....</td>
</tr>
<tr>
<td>892332345</td>
<td>1434241</td>
<td>OOOOSSEEEELLEEOOOOOO...</td>
</tr>
</tbody>
</table>

Creation of interaction string for each user-course. For weeks with no interaction, E represents ‘enrolled in course but didn’t interact with it’ and O represents weeks before or after the course’s official start/end dates.
Quiz completion rate, by number of active weeks

% of quizzes completed per student

Active weeks in course

0 1 2 3-4 3-12 12-53

257571 76868 29964 22806 29904 761
Early Warning Approach

1: Nationally, the average 6-year graduation rate is 60\%.

2: In universities or online courses with high enrollment, faculty and advisors are unaware of the challenges faced by students until the end of the semester.

3: Students without up-to-date help would fail in classes and can't graduate on time.

4: An early warning approach is a tool that can help instructors to identify students at-risk of receiving poor grades.
Feature Description (Course Feature)

CourseLen: How long a course is.

Type: There have 12 different discipline courses in database.

Size: denoted how many students register for this course.

#Q: The total number of quizzes of a course.

#A: The total number of assignment of a course
Feature Description (Student Feature)

QSubmission: How many quiz submissions of a student made before a specific timing.

QScore: How many scores student earned based on the submitted quiz and normalized the value by comparing the average quiz score of the class.

QAttempt: The average attempts times of the submitted quiz made by one student.

QTime: The average spending time of the submitted quiz made by one student.

ASubmission: Same with QSubmission

AScore: Same with QScore

Acperday: How many times a student access to course management system
Basic Framework
Figure 10: Average accuracy using course, student, hybrid features respectively for three different classification methods.
<table>
<thead>
<tr>
<th>Time-Stamp</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR.C</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>LR.S</td>
<td>0.48</td>
<td>0.527</td>
<td>0.576</td>
<td>0.619</td>
<td>0.657</td>
<td>0.685</td>
<td>0.724</td>
<td>0.771</td>
<td>0.809</td>
<td>0.853</td>
</tr>
<tr>
<td>LR.H</td>
<td><strong>0.485</strong></td>
<td>0.53</td>
<td>0.567</td>
<td>0.608</td>
<td>0.645</td>
<td>0.672</td>
<td>0.716</td>
<td>0.759</td>
<td>0.796</td>
<td>0.84</td>
</tr>
<tr>
<td>KNN.C</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
<td>0.384</td>
</tr>
<tr>
<td>KNN.S</td>
<td>0.391</td>
<td>0.396</td>
<td>0.398</td>
<td>0.401</td>
<td>0.403</td>
<td>0.411</td>
<td>0.423</td>
<td>0.433</td>
<td>0.444</td>
<td>0.45</td>
</tr>
<tr>
<td>KNN.H</td>
<td>0.39</td>
<td>0.393</td>
<td>0.396</td>
<td>0.397</td>
<td>0.399</td>
<td>0.405</td>
<td>0.418</td>
<td>0.425</td>
<td>0.434</td>
<td>0.438</td>
</tr>
<tr>
<td>RF.C</td>
<td>0.425</td>
<td>0.424</td>
<td>0.426</td>
<td>0.424</td>
<td>0.422</td>
<td>0.424</td>
<td>0.428</td>
<td>0.427</td>
<td>0.423</td>
<td>0.424</td>
</tr>
<tr>
<td>RF.S</td>
<td>0.456</td>
<td>0.485</td>
<td>0.508</td>
<td>0.538</td>
<td>0.565</td>
<td>0.592</td>
<td>0.624</td>
<td>0.656</td>
<td>0.68</td>
<td>0.707</td>
</tr>
<tr>
<td>RF.H</td>
<td>0.455</td>
<td>0.48</td>
<td>0.499</td>
<td>0.514</td>
<td>0.545</td>
<td>0.567</td>
<td>0.603</td>
<td>0.621</td>
<td>0.648</td>
<td>0.667</td>
</tr>
</tbody>
</table>

**Table 3:** Average F1 Score of 586 students

<table>
<thead>
<tr>
<th>Time-Stamp</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR.C</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
<td>0.551</td>
</tr>
<tr>
<td>LR.S</td>
<td>0.592</td>
<td>0.626</td>
<td><strong>0.665</strong></td>
<td><strong>0.701</strong></td>
<td><strong>0.731</strong></td>
<td><strong>0.755</strong></td>
<td><strong>0.786</strong></td>
<td><strong>0.824</strong></td>
<td><strong>0.856</strong></td>
<td><strong>0.891</strong></td>
</tr>
<tr>
<td>LR.H</td>
<td><strong>0.599</strong></td>
<td><strong>0.631</strong></td>
<td>0.66</td>
<td>0.694</td>
<td>0.724</td>
<td>0.745</td>
<td>0.779</td>
<td>0.815</td>
<td>0.846</td>
<td>0.88</td>
</tr>
<tr>
<td>KNN.C</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
</tr>
<tr>
<td>KNN.S</td>
<td>0.591</td>
<td>0.594</td>
<td>0.595</td>
<td>0.597</td>
<td>0.598</td>
<td>0.605</td>
<td>0.613</td>
<td>0.62</td>
<td>0.627</td>
<td>0.632</td>
</tr>
<tr>
<td>KNN.H</td>
<td>0.59</td>
<td>0.591</td>
<td>0.594</td>
<td>0.595</td>
<td>0.595</td>
<td>0.6</td>
<td>0.609</td>
<td>0.614</td>
<td>0.62</td>
<td>0.624</td>
</tr>
<tr>
<td>RF.C</td>
<td>0.572</td>
<td>0.572</td>
<td>0.572</td>
<td>0.57</td>
<td>0.569</td>
<td>0.572</td>
<td>0.574</td>
<td>0.574</td>
<td>0.568</td>
<td>0.569</td>
</tr>
<tr>
<td>RF.S</td>
<td>0.601</td>
<td>0.627</td>
<td>0.646</td>
<td>0.67</td>
<td>0.693</td>
<td>0.715</td>
<td>0.738</td>
<td>0.765</td>
<td>0.783</td>
<td>0.802</td>
</tr>
<tr>
<td>RF.H</td>
<td>0.599</td>
<td>0.622</td>
<td>0.64</td>
<td>0.652</td>
<td>0.676</td>
<td>0.697</td>
<td>0.725</td>
<td>0.739</td>
<td>0.76</td>
<td>0.772</td>
</tr>
</tbody>
</table>

**Table 4:** Average F1 Score of 586 students
90% student (training data)

\[ x\% \rightarrow TR_{S1} \]
\[ x\% \rightarrow TR_{S2} \]
\[ x\% \rightarrow TR_{Sn} \]

10% student (testing data)

\[ x\% \rightarrow \text{'Unknown'} \rightarrow \text{Baseline Method} \rightarrow TE_{S1} \ldots \ldots TE_{Si} \]
Figure 12: Average accuracy and F1 score result for Course-Specific-Approach
Designing Early Warning Approach using Student’s Early In-class Study Behavior

Zhouxiang Cai  
Georgia Mason University  
Fairfax, Virginia  
rng@wm.edu

Huzefa Rangwala  
Georgia Mason University  
Fairfax, Virginia  
rangwala@gmu.edu

1 INTRODUCTION

The environment of educational institutions today is more complicated. As student enrollment rates and class size along with the heterogeneity of students, such as virtual or face-to-face classes, grow, it becomes challenging to understand the learning outcomes of students. Therefore, in this study, we focus on student study behavior in the classroom. As an early warning approach is a tool that can help instructors to identify students who are struggling or at risk, we propose a model to identify students who are struggling due to miss or disruption of their grades. Early warning modeling techniques are used to identify students at risk by analyzing the students’ social behaviors such as gender and living habits, study habits, and the number of assignments and quizzes. Saha and Bradbury [18] took advantage of student’s social behaviors to predict the outcome. The results showed that taking the feature into account can reduce the error of student prediction.

In this study, we propose a new approach to enhance the class performance characteristic, an understanding of suitable approaches or theories of attended learning. As the number of students engaged in a learning behavior program in Fall 2009 failed to graduate within six years, Schneider and Yin [18] calculated the hidden cost for college dropouts from just a single cohort of entering students lost 8.3 billion dollars [19].

As previously mentioned, the learning experience of students was thoroughly analyzed by the learning behavior model. We also help identify essential features within a course found on the different stages of the course.

CCS CONCEPTS

• Computer systems organization → Embedded systems;  
• Security and privacy → Network security.

KEYWORDS

Early Warning, Learning Analytics, Regression, Classification, Early Student, Behavior

ACM Reference Format


2 LITERATURE REVIEW

Predicting students’ learning outcomes is more and more popular in EDM. Several papers have focused on the analysis and predict student’s in-class performance based on student’s social and learning features with specific factors. Lin et al. [17] evaluate different data mining approaches to classify students based on their EDM usage data. Lin et al. [17] presented student performance prediction using multi-approach models and it is proved that student’s interaction between Massive Open Online Courses (MOOCs) and students’ performance by analyzing the students’ social behaviors such as gender and living habits, study habits, and the number of assignments and quizzes. Saha and Bradbury [18] took advantage of student’s social behaviors to predict the outcome. The results showed that taking the feature into account can reduce the error of student prediction.

In this paper, we propose a new approach to enhance the class performance characteristic, an understanding of suitable approaches or theories of attended learning. As the number of students engaged in a learning behavior program in Fall 2009 failed to graduate within six years, Schneider and Yin [18] calculated the hidden cost for college dropouts from just a single cohort of entering students lost 8.3 billion dollars [19].

As previously mentioned, the learning experience of students was thoroughly analyzed by the learning behavior model. We also help identify essential features within a course found on the different stages of the course.
Thanks for this summer