Exploratory Analysis with MOOC Data used for Blended Learning

Rahul Rajesh, Silvana Miranda, Michael Chan
LEARN Center, EPFL, Switzerland

***Abstract*—In this study, we analyse regularity and clickstream activity of students in an introductory Linear Algebra course at EPFL. The course has started a blended learning approach from 2017 and the aim of this research is to generate useful insights for this course. For feature engineering, we make use of proven metrics, from existing research, to evaluate regularity/clickstream behaviour. The students were then clustered using K-Means/Hierarchical Clustering algorithms. The clustering has surprisingly shown that students who were least regular performed the best, indicating that other factors like background could play a role. The clustering has also shown that skipping/replaying for a longer duration is a possible sign of disengagement that leads to poor performance. In all, the exploratory analysis found some insights based on the MOOC data and can be further extended to pave the way for improved systems in blended learning models; systems that can offer personalised learning tracks for students and allow them to absorb the material at their own pace.**

I. Introduction

Technology is evolving at an enormous scale and it is revolutionizing our education system. A major part of this is due to Massive Open Online Courses (MOOCs). There are now many established MOOC platforms like Coursera, Udacity etc. that are able to offer quality content to millions of viewers worldwide. Even in universities, it is starting to pick up on usage. As an example, Georgia Tech University recently won Reimagine Education's gold medal for their online master's course using MOOCs [1].

Universities also incorporate MOOCs as part of a blended learning approach. One such approach is the flipped classroom. These involve 2 components: (1) individual instruction before classes, generally supported by technology; and (2) interactive group activities during class, focused on problem solving and brainstorming [2].

However, implementing a good blended system is not easy and involves a methodological change both for the teacher and a change of mindset for students [2]. It is thus, important to draw on patterns from available data and do feedback sessions with the students. In this paper, we look to understand more about MOOC data from a first year Linear Algebra course at EPFL. This specific course has employed a blended learning approach from 2017.

II. Related Work

Before we discuss our methodology, it is important to consider existing research done on MOOCs. MOOC data is complex and there are multiple ways in which one can go about analysing it.

In Sinha [3], information processing and attrition behaviour was inferred from MOOC clickstream data using a cognitive video watching model. Another study [4] used similar data to predict whether a user will be correct on first attempt when answering a question in a quiz that follows the video.

One interesting area in MOOC research has to do with regularity. In Sharma [5], different metrics were formulated to quantify regularity and this was correlated against grades. Regularity is important in blended learning models where students have to attend group activities during the week to complement their MOOC videos.

Another form of analysis done on MOOC has to do with analysing the individual clickstream events. In Nan Li [6], video interactions were explored and related to perceived video difficulty of the student. Nan Li came up with various metrics such as replay duration, pause frequency etc. to build various interaction profiles for a student.

III. Methodology

This study will seek to explore the regularity and clickstream behaviour for students in the Linear Algebra course.. The eventual goal would be to provide useful insights for improved implementations of flipped classroom methods.

A. Dataset

The dataset used in this research is from an introductory Linear Algebra course conducted at EPFL. We are given access to the clickstream activity and final grade of the students in the course. The students are from two batches in 2017 and 2018. The summary of the data is shown below in table I.

	2017-2018	2018-2019
Students	135	38
Events	171,040	41,458

Table I: Dataset Overview

Figure 1 shows the average grade for students based on how many videos they watched. The histogram shows that there are students in the dataset who scored good grades without watching too many videos and vice versa. This data will have to be processed first before we start our analysis.

B. Objectives of study

Before discussing further, it is important to have a clear set of objectives in how we want to explore the data. In order to do so, we draw upon two claims.

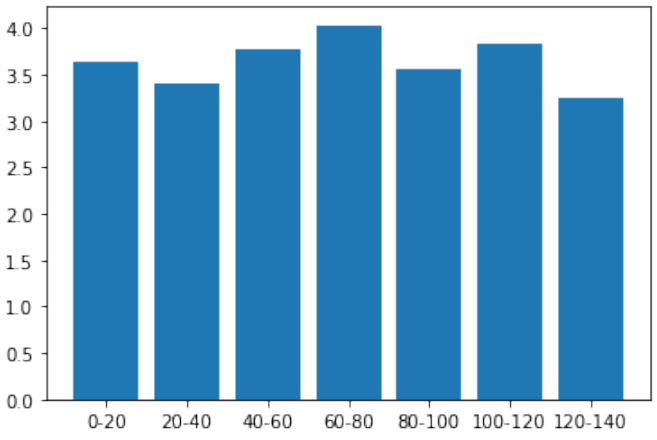

Fig. 1. Average Grade of students who watched a range of videos

Claim 1: Students who are more regular in watching the MOOC content tend to be more engaged with the material and do better in the course

This claim is drawn from Sharma's work on quantifying regularity. He has come up with a set of metrics and concluded that students who plan their learning activities in a regular manner have better chances of succeeding in the MOOC [5]. We will try to take up the same approach and see if it holds true even for blended models like ours.

Claim 2: Clickstream behaviour is linked to a student's learning ability and this would affect the final grade they achieve

This claim suggests that a student's activity (pause/seek forward etc.) will influence how he learns the content of the video. Nan Li has investigated frequency metrics such as replay count and inferred patterns e.g. students who replay more perceive the video to be more difficult [6]. We hope to do a similar study to try and see if there are any interesting patterns of behaviour that could influence how well a student does in Linear Algebra.

C. Feature Processing

As mentioned earlier, we have to be process the data first to ensure we are able to get good results. The main steps are outlined below:

Active Students: We considered only students who have watched more than 60 videos (refer to figure I). The highest number of videos watched is 139. This means all the students in our sample will have at least half of that amount. This would eliminate imbalance from students achieving good grades without engaging with the MOOCs.

Repeating Students: The dataset has 5 students who repeated the course. Their clickstream behaviour was only considered in their first year. For the repeating year, their background and prior knowledge would definitely skew the results (all of them passed in the second year).

New Events: Events such as "seek" were doubled to "seek back" and "seek forward" based on the timing information. Same was done for speed change. This allowed us to widen our features and do more complex analysis. Events that we did not use like hiding transcript or translations were dropped.

D. Feature Engineering

In order to analyse the claims we have done, we take inspiration from existing research and apply it on our dataset. This section will give an overview of that process.

1. Regularity

We make use of Sharma's features to analyse regularity. They are shown in table II below:

Feature	Description
PDH	Peak Activity on Day Hour
PWD	Peak Activity on Week Day
WS1/WS2/WS3	Weekly Similarity of Activity
FDH	Repetition of hourly pattern over days
FWH	Repetition of hourly pattern over weeks
FWD	Repetition of daily pattern over weeks

Table II: Regularity Features

The features were carefully implemented to encompass a few main aspects. It checks if a student has high activity on a certain day or a hour (every Tuesday for example). It checks if the student works on the same day every week. A Fourier transformation is also done on the activity patterns to infer if there is a certain frequency of behaviour (e.g. is a student active at 5h - 7h every day).

Some important considerations we did for our dataset involves the period. The flipped classroom was not all year round and happened only on a subset of weeks in the semester. We focused on student activity only in those periods.

2. Clickstream Analysis

Some notable techniques of analysing clickstreams including getting n-gram patterns or doing a frequency count of certain events. For this research, we chose to implement Nan Li's metrics as they are well thought out and makes sense in terms of our objectives. A summary of this is shown below:

Feature
Pause Duration Pause Frequency
Replayed Period Replay Count
Skipped Content Skip count
Speed Up Count Speed Down Count Average Speed Change

Table III: Derived Clickstream Features

As shown in table III, there are four main aspects considered here: Pause, Skip, Replay and Speed Change. Average Speed Change is the difference between the weighted arithmetic mean of video speeds at each second and the initial speed.

Some considerations outlined in Nan Li's research include:

- Pauses were only considered if they were between 2s to 5min. 1s pauses is probably noise and anything longer than 5min may be an actual break away from the video.

- Speed Changes were only considered in a 10s window. Students are likely to toggle between different speeds to get to their preferred speed and we want to disregard intermediate selection.

E. Model Choice and Strategy

We have now a reasonable set of features that would help us analyse the regularity and clickstream behaviour of the students. The main decision we now have to consider is the choice between using supervised learning and unsupervised learning.

For this research, we chose to do an unsupervised approach to analyse both our claims. The reason we did not do a supervised approach is three-fold. One is that we have high variance in our data due to the small sample size. This would mean we will not be to confidently assess our model's accuracy. Another reason, is that a student's background will undoubtedly affect his final grade. We do not have this data and thus, may not be able to make good predictions. Finally, the MOOCs only happened in a subset of the semester and may not accurately showcase the final exam performance.

A more reasonable approach to take is to do some form of clustering using the features we derive and inspect the patterns in each cluster. This would allow us to understand more about the data points and will pave the way for future research that can be done using MOOC data in a flipped classroom setting.

For our clustering we used 2 well-known algorithms, K-Means and Hierarchical Agglomerative clustering. We chose the best one based on the K-Means inertia and indexes like the silhouette score and the Davies Bouldin score. These are metrics that evaluate both intra-cluster and inter-cluster spread of the clusters generated by our algorithms.

IV. Results & Discussion

In this section, we showcase the results of our clustering and explain the patterns we found from analysing the distribution of features for each cluster. We clustered the students twice, one for regularity and one for click-stream activity. The reason for not doing one clustering for both these features is that they look at very different aspects. It is difficult to get good clusters by combining these feature sets.

A. Regularity: Is it necessary?

For regularity, both K-Means and Hierarchical Clustering achieved similar results. We achieved a Silhoutte score of 0.45 and Davies Bouldin Score of 0.81 for a cluster size of 3. The scores are not perfect but are indicative of a pretty reasonable

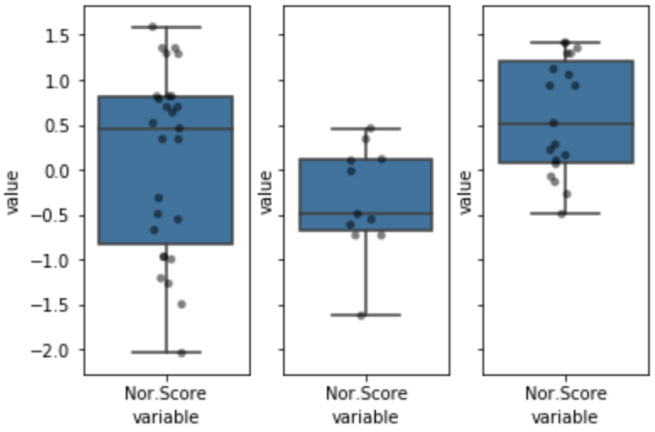

Fig. 2. Normalized Score of clusters formed from regularity

cluster. Figure 2 showcases a box-plot showing the distribution of normalized exam scores for each of the clusters.

The clusters can be broken down into three groups. The high performing group (one on the extreme right), the weak performers and one which has mixed performers. There is a fairly even distribution of students in each cluster.

By looking at the distribution of the regularity features in each cluster, we can infer behaviour of students who are high/low performers. Interestingly for our data set, students who are the best performers are the least regular. They have the lowest score for the regularity metrics. On the other hand, people who are the most regular and study the MOOCs at fixed intervals come from the low performing group.

This can be explained by considering the nature of our data. For an introductory Linear Algebra course, a student's background plays an important role. It is likely that students from a strong background form the high performing group and thus, do not need to engage with the material as closely to achieve good results. In a blended classroom setting, students with background can learn from the in-class sessions without having to always watch the videos in advance.

B. Clickstream Activity: Can it correlate with final performance?

For clickstream, the best features among the ones in table III were used to do the clustering. The strength of a feature was deduced by fitting a Generalized Additive Model(GAM) and plotting the partial dependency curve along with the confidence interval. Figure 3 shows an example of a GAM fit partial dependency curve for replay duration against normalized score. We see that GAMs are effective in modelling non-linear relations. It looks like longer replay duration leads to lower performance (although the confidence intervals are wide, indicating higher variance in the data set).

The final features chosen were skip duration, replay duration and average speed change. Hierarchical clustering performed better with a Silhoutte score of 0.45 and Davies Bouldin Score of 0.80 for a cluster size of 3. The cluster performance

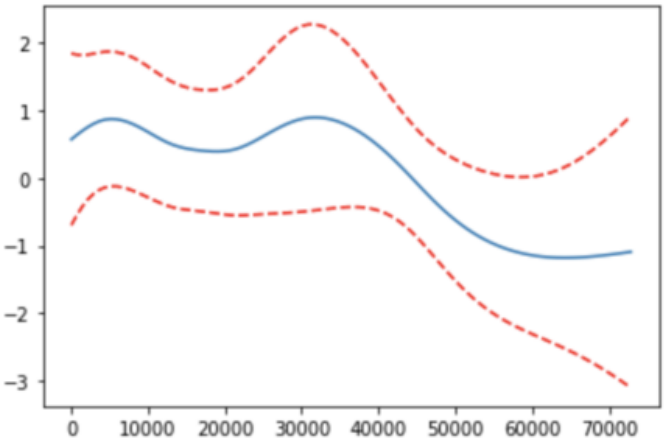

Fig. 3. GAM Partial Dependancy curve of normalized score against replay duration

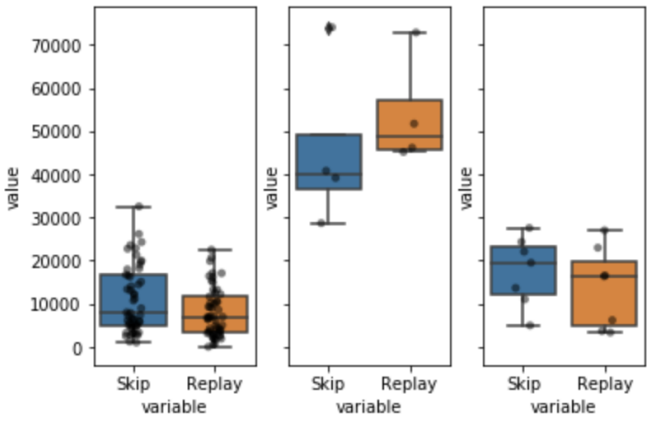

Fig. 4. Boxplot of skip duration and replay duration for the clusters formed from clickstream activity

is very similar to the regularity one shown in figure 2. There is one cluster with clear high performers, another with low performers and one that is fairly mixed. However, the distribution is rather uneven, we have 7 high performers, 4 low performers and 53 mixed performers. This is understandable considering the variance of our features.

An interesting observation we have can be shown on the box-plot in figure 4. This is a plot of the distribution of skip/replay duration for the three clusters. The high performing cluster is the one on the extreme right, the low one is in the centre and the left is the mixed cluster.

It is interesting to note that students who have performed poorly in the exams tend to have high replay and high skip duration. This is a sign of disengagement with the video content. It might be because these students did not understand the content they skipped and had to replay them again.

Apart from this, the higher performing tend to watch videos at a much higher average speed (always between 1.5 - 2.0). Again this could be a sign that due to their background they find the videos to be easier and could afford to watch them at a higher speed.

Both regularity and clickstream clustering has shown certain patterns of behaviour to separate out a group of high/low performers. This same approach taken on a larger sample could yield even better results and showcase powerful patterns that can give MOOC content creators more insights.

V. Conclusion

In conclusion, the exploratory analysis done on the MOOC data showcased some interesting patterns that could pave the way for recommendation systems in MOOC platforms.

One such system could be a personalised learning track for fast learners. In order to encourage students with background to be more regular, more videos could be shown to them or more difficult quizzes can be given. The benefit of blended learning is that there is more flexibility to cater to individual student's needs and let them learn at their own pace.

Another idea is to keep track of certain MOOC metrics to identify students who are less engaged with the material. If a student skips often, more attention could be given to him either through in-class sessions or through online forums.

Blended Learning models are becoming a highly effective way to teach large classrooms without the overhead of traditional methods. Exploratory MOOC research like the one employed in this paper can be extended in several ways to complement these models.

Acknowledgements

We thank the LEARN Center and our mentor Himanshu Verma for his advice and guidance throughout this project. We also thank our instructors Martin Jaggi and Rüdiger Urbanke for teaching the necessary concepts that we used in the research.

References

[1] T. Malone, "Omscs wins oscars of educational technology." [Online]. Available: https://www.news.gatech.edu/2019/12/12/omscs-wins-oscars-educational-technology

[2] M. P.-S. J. A. P. C. A.-H. María Fernanda Rodríguez, Josefina Hernández Correa, "A mooc-based flipped class: Lessons learned from the orchestration perspective," *EMOOCs 2017*, pp. 102–112, 2017.

[3] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg, "Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions," 2014.

[4] C. G. Brinton and M. Chiang, "Mooc performance prediction via clickstream data and social learning networks," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 2299–2307.

[5] M. Shirvani Boroujeni, K. Sharma, L. Kidzinskikidzinski, L. Lucignano, and P. Dillenbourg, "How to quantify student's regularity?" 09 2016.

[6] N. Li, Kidziński, P. Jermann, and P. Dillenbourg, "How do in-video interactions reflect perceived video difficulty?" 09 2015.