Temporal Analysis in Massive Open Online Courses – Towards Identifying at-Risk Students Through Analyzing Demographical Changes

Lei Shi[1](✉), Bokuan Yang[2], and Armando Toda[3]

[1] Durham University, Durham, UK
lei.shi@durham.ac.uk
[2] University of Liverpool, Liverpool, UK
b.yang12@student.liverpool.ac.uk
[3] University of Sao Paulo, Sao Carlos, Brazil
armando.toda@usp.br

Abstract. This chapter demonstrates a temporal analysis in Massive Open Online Courses (MOOCs), towards identifying at-risk students through analyzing their demographical changes. At-risk students are those who tend to drop out from the MOOCs. Previous studies have shown that how students interact in MOOCs could be used to identify at-risk students. Some studies considered student diversity by looking into subgroup behavior. However, most of them lack consideration of students' demographical changes. Towards bridging the gap, this study clusters students based on both their interaction with the MOOCs (activity logs) and their characteristics and explores their demographical changes along the MOOCs progress. The result shows students' demographical characteristics (membership of subgroups) changed significantly in the first half of the course and stabilized in the second half. Our findings provide insight into how students may be engaged in MOOCs and suggest the improvement of identifying at-risk students based on the temporal data.

Keywords: MOOCs · Clustering · Behavior patterns · Temporal analysis · Unsupervised machine learning · Learning analytics · Demographical characteristics

1 Introduction

Massive Open Online Courses (MOOCs) are a unique form of educational information systems offering free access to the intellectual holding of universities [35]. It has been spreading in both domestic and international education sectors. Many world-class universities have joined in the MOOC movement. A number of MOOC platforms have been launched across the globe in many subjects [21]. Despite the potential and hype

A prior version of this paper has been published in the ISD2019 Proceedings (http://aisel.aisnet.org/isd2014/proceedings2019).

A. Siarheyeva et al. (Eds.): ISD 2019, LNISO 39, pp. 146–163, 2020.
https://doi.org/10.1007/978-3-030-49644-9_9

associated with MOOCs, the persistence or completion rates overall are astonishingly low. Some studies reported that the completion rate could reach as low as 5% [34]. This challenge has catalyzed considerable studies on identifying dropout possibilities of MOOC students [1, 17, 27, 45], as well as how to increase persistence or completion [16, 36, 44]. The ultimate goal of this research is thus to identify the at-risk students as early as possible; such that early interventions can be injected to prevent them from dropping off from the MOOCs.

In comparison to traditional educational methods, MOOCs allow for prediction of whether a student may dropout off from a course using their prior voluntary actions logged in the database – so called "educational big data", since the dataset is normally diverse, complex and of a massive scale. Most existing studies of predicting or identifying at-risk students in MOOCs (those students who are likely to drop out from a MOOC) heavily rely on the "average/overall" analyses, lacking adequate examination of the potential differences amongst subgroups of students. This approach may produce result with potential pitfalls [5, 8, 18]. Thus, our study, presented in this chapter, aims at addressing this concern by exploring the diversity of students and their behavioral changes (the percentage of students falling into each subgroup and the subgroup transitional patterns) along the MOOCs progress.

In this study, we combine the previous study on identifying student subgroups, using both students' interaction data (behavioral) with the MOOCs and their characteristics (demographical) to allow for a more accurate clustering [11, 23, 38]. This chapter presents the student subgroups clustered from two MOOCs delivered on the FutureLearn[1] MOOC platform and visualizes demographical pattern changes of these subgroups along the courses progressed to help unmask these changes at different stages of the course. In particular, this study aims to answer the following three research questions:

RQ1. How can we subgroup students in MOOCs?
RQ2. How can demographical characteristics of each subgroup change by weeks?
RQ3. Are there transitional patterns amongst subgroups, on a weekly time scale?

2 Related Work

2.1 Learning Analytics

Learning Analytics (LA) is a rapidly expanding area, especially with the advent of "big data" era, more widely used data-driven analytics techniques, and new extensive educational media and platforms. It is defined as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs [46]. Using LA, many studies were conducted with the aim of understanding and predicting student behavior in educational information systems.

For example, [42] used machine learning and statistical modelling techniques to explore students' engagement in MOOCs. [33] investigated students' demographical

[1] https://www.futurelearn.com.

information in MOOCs, intended behavior and course interactions, to investigate variables indictive of MOOC completion. [41] examined how student demographic indicators might correlated to their activities in MOOCs. [12] used dimension reduction and clustering techniques with affinity propagation to identify clusters and determine students' profiles based on their help-seeking behavior. [9] explored the effects of common MOOC discussion forum activities on peer learning and learner performance. [38] identified three influential parameters to cluster students into subgroups and profiled them by comparing various behavioral and demographical patterns, in order to investigate their engagement in MOOCs.

Most of these studies grouped or clustered learners into subgroups and compared behavioral patterns amongst subgroups allowing for a deeper understanding of how MOOC learners engage and perform. In the currently study presented in this chapter, we also used the learning analytics approach, leveraging various techniques including unsupervised machine learning and statistical modeling.

2.2 Subgroup Clustering in MOOCs

Some previous studies attempted to cluster students based on their interaction with lecture videos and assignments using a variety of methods and approaches, including bottom-up approaches to identify potential subgroups [29, 31, 38] and top-down approaches to partition students into pre-defined groups [1, 32].

For example, [24] demonstrated a clustering technique based on a derivative single variable for engagement, where they labelled all the students either as "on track" (took the assessment on time), "behind" (turned in the assessment late), "auditing" (didn't do the assessment but engaged in watching videos), or "out" (didn't participate in the course at all). [19] extracted four types of engagement trajectories, including 1) "Completing" – the students who completed the majority of the assessments; 2) "Auditing" – the students who did assessment infrequently if at all and engaged instead of watching video lectures; 3) "Disengage" – the students who did assessment at the beginning of the course but then had a marked decrease in engagement; and 4) "Sampling" – the students who watched the lecture video(s) for only one or two assessment periods. While, in their research, the authors used the k-means clustering algorithm to categorical data, to a certain extent, since they simply assigned a numerical value to each of the labels ("on tack" = 3, "behind" = 2, "auditing" = 1, "out" = 0). However, converting categorical data into numeric values does not necessarily produce meaningful results in the case where categorical domains are not ordered [20]. Therefore, these approaches have potential problems with converting participation labels, although they still can provide a viable way to cluster students based on the log data from the MOOCs platforms. In our study, to mitigate this issue, we used the one-hot encoding [6] to convert categorical data, thus reducing the impact of the categorical data.

Other studies were focused on different approaches to identifying subgroups, but most of them did not consider behavioral changes over time from the clustering [18, 22, 26, 28]. It is important to explore behavior patterns of subgroups of the students on a specific time scale, since the characteristics of each subgroup, and the proportion of its total interaction, vary along a MOOC progresses. This can also help the platform adjust the content of the course, according to the progress of the course.

In our current study, we apply a bottom-up cluster approach using the k-means++ cluster algorithm with students' log data to identify distinct subgroups as well as observe their characteristics changes on a weekly time frame, thus offering a dynamic perspective for students' subgroups.

2.3 Learning Persistence in MOOCs

Considering the problem of the low completion rates in MOOCs, learning persistence was selected as a critical MOOC outcome, which can provide valuable insights into the interactions between the course design and students factors [13, 14, 19]. Several studies have demonstrated possible ways of using learning analytics on interaction and assessment to meaningfully classify student types or subgroups and visually represent patterns of student engagement in different phases of a MOOC. For example, Coffrin et al. [11] divided weekly participation into three mutually exclusive student subgroups: Auditors – those who watched videos in a particular week instead of participating assessments; Active learners – those who participated in an assessment in a week; and Qualified learners – those who watched a video or participated in an assessment. The study investigated students' temporal engagement along course progressed. It also showed a way of combining the State-Transition diagram with an analysis of student subgroups to illustrate the students' temporal engagement in courses. Their result indicated that different courses might show similar patterns, although they were different in terms of the curriculum and assessment design.

Similar studies have attempted to compute a description for individual students in terms of how they engaged in each assessment period of a course and then applied clustering techniques to find subgroups in these engagement descriptions [15, 18, 24]. While these studies have successfully concluded the proportion of students in different subgroups by week, they did not attempt to analyze the individual subgroup changes on a specific time scale. Student behavior may change along a MOOC progresses, where they may have been labelled into one subgroup and transit to another in subsequent weeks. It is meaningful to evaluate the transitional pattern for each subgroup on a certain time scale. Therefore, in this study, we measured the proportion of students falling into each subgroup and concluded the transitional pattern for each subgroup on a weekly time frame.

3 Method

3.1 MOOCs and Dataset

The two MOOCs under study included "Leadership for Healthcare Improvement and Innovation" and "Supply Chains in Practice: How Things Get to you", delivered on FutureLearn, a MOOC platform that is freely available for everyone. Each MOOC was structured in weekly learning units. A weekly learning unit was composed of a few learning blocks, each of which consisted of a number of steps. Steps were the basic learning items, which contained lecture streams that the students needed to access, during the learning process. Both MOOCs were *synchronous* – having an official starting week,

considered as Week 1 in this study, with a duration of six weeks, and an ending week, i.e. Week 6.

Both MOOCs attracted thousands of students. However, only around 7% of them finally completed the courses, reflecting one the of the biggest challenges in MOOC platforms – the low retention/completion rate [10]. According their completion, we categorized the students as the following:

- *Registered students* – have enrolled in the course
- *Participated students* – have attended at least one steps
- *Completed students* – have completed the courses by the end of Week 6
- *Purchased students* – have bought the certificate of the course.

Table 1 shows the statistics for these two courses.

Table 1. Course design and participants.

Course	"Leadership for healthcare improvement and innovation"	"Supply chain in practice: How things get to you"
Duration of the course	6 weeks	6 weeks
Total steps	73	109
Registered students	4,046	5,808
Participated students	2,397	2,924
Completed students	377	318
Purchased students	149	69

The dataset used in this study was from those two MOOCs and included:

- Step record – which student at what time visited which step; when they marked a step as complete.
- Comment record – which student at what time left what comment on which step; how many "likes" a comment received.
- Student record – students' demographical information such as gender, age group, country, highest educational level, employment status, as shown in Table 2.

Students' demographical information was collected using a pre-course survey asking optional questions about their gender, age group, country, and so on, as shown in Table 2, the column on the left. Only 9.5% of the students (506 out of 5,321) answered all the survey questions. As using incomplete student record would affect the result of the analysis, in this study we only used the records of students who answered all the survey questions.

Table 2. Demographic information in student record.

Variable	Description
User ID	The unique identifier for a student
Gender	The gender of the student
Age group	The age group where the student belongs to
Country	The country where the student belongs to
Highest educational level	Student's highest education level
Employment status	Students' employment status
Employment area	Students' employment area

3.2 Subgroup Clustering

In previous studies, watching lecture videos and submitting assignments were used for clustering students [18, 22]. Considering the conversational framework of FutureLearn and the course design, two interactive indicators were generated from the step record and the comment record:

- Steps visited – the proportional of all the steps available visited by a given student in a given week.
- Comments submitted – the number of comments submitted by a given student in a given week) and the gender (of a given student).

Other studies, e.g. [2, 38], used demographical indicators such as gender and age to predict student engagement; [37, 40] focused on the use of learning platform's features in order to analyze learning behavior patterns. Different from these previous studies, in this study, we selected both students' demographical data and their interaction (activity logs) data for the clustering process. We excluded the highly correlated variables with the numbers of steps visited or comments submitted, leaving gender as an extra variable for the clustering process.

The clustering process was based on the k-means++ algorithm [4], which could reduce the influence of randomly assigned initial centroids in the *k-means* algorithm [30]. Similar to previous studies, e.g. [42], we used the "Elbow method" to select the reference k value for the k-means++ algorithm [25]. We used a number of k values around the reference k to cluster subgroups of the students, and then we conducted Kruskal-Wallis H tests and Mann-Whitney U tests to examine whether the k value could differentiate subgroups on every clustering variable. Moreover, different from most existing studies, which used cumulative data from the entire course to cluster subgroups of the students, in this study, we used cumulative data from each week for the subgroups clustering.

3.3 Transitional Pattern for Subgroups

We clustered students into subgroups based on their temporary behavioral data (how they interact with the MOOCs including how they visited steps and submitted comments, from one week to another). We used State-Transition diagrams to visualize the weekly transitional patterns amongst the subgroups, where the dropped-out students were marked into a different subgroup. Similar to the subgroup clustering, two indicators were generated: 1) the number of steps a student visited, and 2) the number of comments a student submitted, as defined in Sect. 3.2. From the State-Transition diagram, we analyzed the proportion for students falling into each of the subgroups by week and generalized the transitional pattern for different subgroups each week.

4 Result

4.1 Subgroup Clustering

In this study, we selected the percentage (instead of the raw number) of the steps visited, and the number of comments submitted, by the students, as prime cluster variables, with additional demographical variables selected from the student record. From the correlation analysis, we excluded highly correlated variables. More specifically, we used the η(eta) statistics to measure the degree of association between categorical and numeric variables – the independent variable Y, i.e. Steps and Comments, and the dependent variable X, i.e. Gender, Country, Age range, Educational level, Employment area and Employment status, as Table 2 shows.

For the association between the categorical variables, we used the Chi-square test with the significant level $= 0.05$. The result suggested a strong association between the variables of Gender and Employment area ($\chi^2(23) = 39.9, p < 0.05$). Therefore, only one of these two variables might be selected as a clustering variable. Considering the fact that the MOOCs analyzed in this study were specialized in certain subjects thus maybe resulting in special employment distribution, the gender variable was selected for a general conclusion. Therefore, our absolute selection of variables included:

- Steps – the percentage of steps visited by a student.
- Comments – the number of comments submitted by a student.
- Gender – the gender of a student.

Although the FutureLearn MOOC platform provides multiple gender options in the pre-course survey, we only considered two options – *female* and *male*, as the other options were very underrepresented. Therefore, we considered the gender variable as a dummy variable and we used 0 to represent the option of *female* and 1 to represent the option of *male*.

Using the "elbow method", Mann-Whitney U tests and the K-means++ clustering algorithm, we successfully clustered those 506 students into three distinct subgroups based on the cumulative data. More specifically, we used the "elbow method" to estimate the optimal k value for the k-means++ algorithm processed in this study – the result can be seen below in Fig. 1.

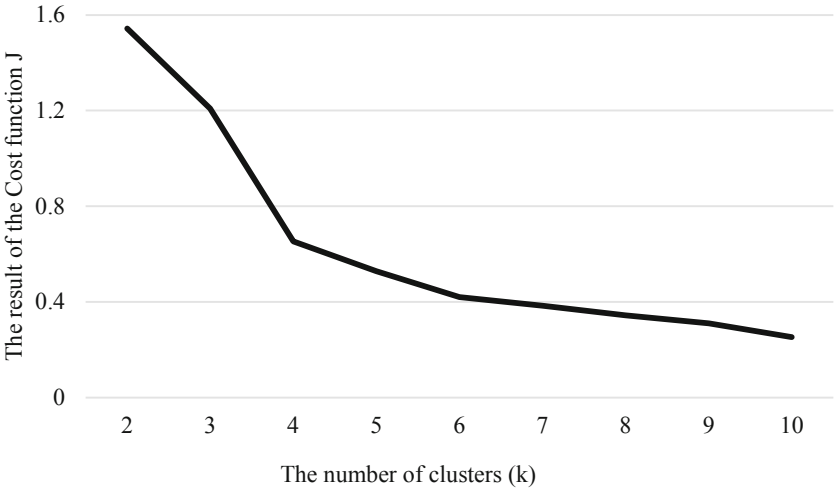

Fig. 1. Cost function J for the dataset

The "elbow method" believes that one should choose a number of clusters so that adding another cluster does not offer much better modelling of the data. As shown in Fig. 1, the result of ***cost function J*** experienced the most significant decrease in $k = 4$ (where the "elbow" appears). Therefore, the $k = 4$ was chosen as a reference k value candidate in the subsequent analysis. Based on this reference k value, we used several k values, ranging from 2 to 5, in order to cluster student into subgroups. In this case, the Mann-Whitney U test with significant level = 0.05 was chosen to validate whether there was a significant difference among these subgroups of the students, and the results suggested that neither $k = 4$ nor $k = 5$ could differentiate subgroups. Therefore, we chose $k = 3$ in this study and the cluster results can be seen below in Table 3 below.

Table 3. Subgroup cluster centroids.

	Steps	Comments	Gender	N
Cluster 1 – Samplers	.926	7.16	.360	113
Cluster 2 – Viewers	.107	.91	.353	369
Cluster 3 – All-rounders	.990	67.54	.550	24

Based on the previous work [3], where the authors labelled students into three subgroups, based on lecture video watching and assignment submission: Viewer (primary watching lecture videos, handing in few if any assignments), Solvers (primary handing in assignments, viewing few if any lecture videos) and All-rounders (balancing between watching lecture videos and handing in assignments). On the basis of this work, we further clustered students by their positivity. In this study, we did not choose assignment submission as one of the clustering variables, but we chose the number of comments

submitted, to replace assignment submission, as in the previous work. We labelled all those 506 students into the following subgroups:

- **Viewers** (Cluster 1; 22.33% of the total population): overall, they visited a very high percentage (92.6%) of the steps but submitted very few comments (Mean = 7.16).
- **Samplers** (Cluster 2; 72.92% of the whole population): they made up the largest student subgroup, but they were also the least engaged students – they visited only 10.7% of the steps and on average they left only 0.91 comments.
- **All-rounders** (Cluster 3; 4.74% of the total population). They made up the smallest student subgroup, yet they were the most engaged students – they visited 99.0% of the steps and on average they left 67.54 comments.

From this subgrouping method and its result, we can see that the least engaged students occupied the largest percentage of the total population. This is consistent with many previous studies, e.g., [7, 43], and has been one of the biggest challenged in the field of MOOCs.

4.2 Weekly Changes of Cluster Centroid

In order to explore the temporal changes of subgroup memberships, we further divided the students into two categories based on the number of steps they have visited and the number of comments they have submitted. The students who had partially participated (i.e. they have submitted at least one comment or visited at least one step) the courses in a given week were selected and clustered into 3 subgroups, based on the k-means++ algorithm. Steps, comments and gender were selected as the input variables for the clustering process. As shown in Table 4, the cluster centroids stabilized at a certain level across weeks, which suggests that the same subgroup had a similar behavior pattern at different stages of the MOOCs.

4.3 Weekly Changes of Subgroup

To investigate how the subgroups changed along the MOOCs, the percentage of the students labelled in each subgroup per week were also retrieved from the dataset. From Fig. 2 and Table 5 we can see that the first half of the MOOC and the second half of the MOOC had very different demographical characteristics, where the percentage of the students in each subgroup changed significantly in the first half of the courses (between Week 1 and Week 3). More specifically, the percentage of Samplers decreased from 50.4% to 17%, which may be caused by a large number of dropout students in the first two weeks. The proportion of Viewers increased significantly from 42.8% in Week 1 to 68.8% in Week 3 and kept stable at a certain level in the rest of the weeks. The proportion of All-rounders kept at a relatively stable level, i.e. around 10.0%, which suggests that these students were relatively stable, even in the beginning weeks when many students dropped out, and that this type of students had more chance to complete the MOOCs.

Here, we use the State-Transition Diagram to present in detail how the students shifted between subgroups, i.e. the changes of the students' memberships of the subgroups. We assumed possible student subgroups, i.e. Sampler, Viewer, All-rounder and

Table 4. Centroids for weekly subgroups.

		Steps	Comments	Gender
Viewer	Week 1	0.964	1.300	0.544
	Week 2	0.979	0.934	0.610
	Week 3	0.988	0.792	0.625
	Week 4	0.936	0.624	0.624
	Week 5	0.986	0.784	0.589
	Week 6	0.971	1.490	0.640
Sampler	Week 1	0.214	0.300	0.428
	Week 2	0.229	0.195	0.507
	Week 3	0.207	0.000	0.467
	Week 4	0.206	0.035	0.517
	Week 5	0.259	0.105	0.526
	Week 6	0.180	0.133	0.467
All-rounder	Week 1	0.986	11.886	0.571
	Week 2	0.998	12.138	0.483
	Week 3	0.990	10.880	0.560
	Week 4	1.000	10.583	0.625
	Week 5	0.998	12.320	0.640
	Week 6	0.952	14.875	0.687

Drop-out, as four possible states each week, and the transitions from one subgroup to another was indicated by the arcs between two states. Figure 3 provides a legend to understand the State-Transition Diagram used in the analysis. The legend shows two subgroups, A and B; the arcs between circles represent the students transited their subgroup from A to B in a subsequent week. In order to better visualize the number of students in each subgroup in each transition, the circle areas and arc's weight are linearly related to the number of students in the subgroups and the transitions respectively.

Figure 4 demonstrates the demographical changes for the Sampler subgroup – a very large proportion of the Samplers dropped out from the courses in the following weeks, while only a small percentage of them maintained their behavior or transited to become Viewers. This means that, the Samplers are definitely the "at-risk" students, who need immediate interventions to prevent them from drooping out from the MOOCs. Apart from the first week, no student had transited from the Sampler subgroup to the All-rounder subgroup (the most active and engaged group) in the following weeks, which suggests that, without any intervention, it is very unlikely for a highly inactive student to become highly active in a short period. Therefore, it is crucial that, early intervention is injected, once a student is detected or identified as being inactive or less engaged.

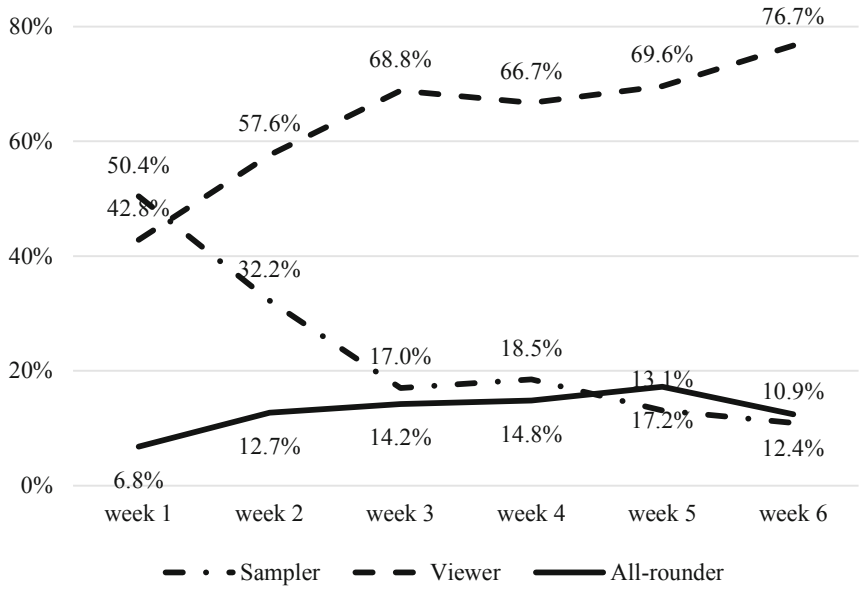

Fig. 2. The percentage of students each subgroup across weeks

Table 5. The number of students each subgroup across weeks.

	Sampler	Viewer	All-rounder
Week 1	252	214	34
Week 2	76	136	30
Week 3	30	121	25
Week 4	30	108	24
Week 5	19	101	25
Week 6	14	99	16

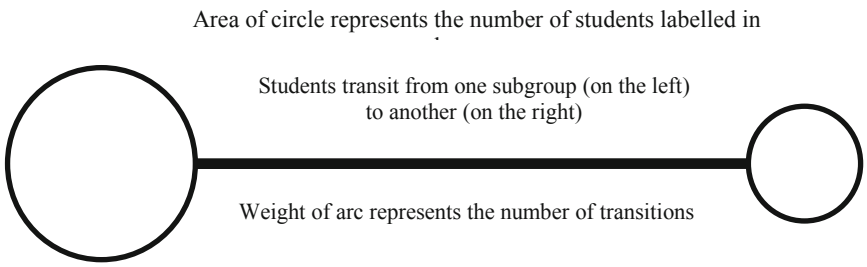

Fig. 3. State-transition diagram legend

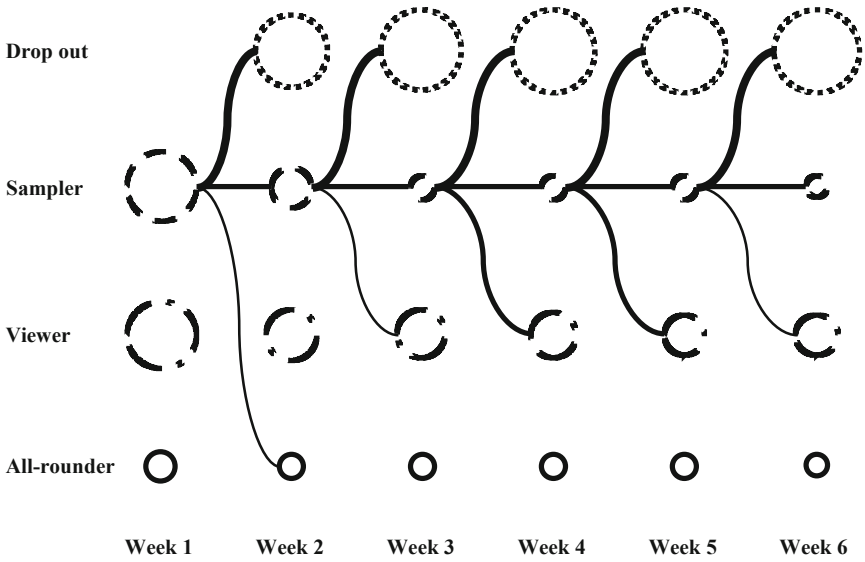

Fig. 4. Samplers' demographical changes across weeks

For example, a reminder email could be sent to them, emphasizing the importance of keeping up with the MOOC.

Figure 5 focuses on the demographical changes for the Viewer subgroup, which also indicates that each subgroup had a similar behavioral pattern transition each week. However, different from the Sampler subgroup, most students belong to the Viewer subgroup maintained their behavior patterns in the following week with only a very small percentage of them dropped out from the courses or transited to another subgroups. As it was unlikely that these students would drop out from the MOOCs, they were not clearly not as "at-risk" as those Sampler students. Nevertheless, according to the definition of Viewer (as per Sect. 4.2), although these students were focused on accessing learning materials, they did not tend to interact with peers. Previous studies, e.g. [39], have demonstrated that social interactions might be very helpful for the students to have better learning result in MOOCs. Therefore, some mild interventions, such as an email promoting participation in the discussion forum, may be very useful to be provided with.

Similarly, Fig. 6 shows that while All-rounders represented the smallest proportion of the students, they were the most stable subgroup – there was no significant demographical fluctuation event in the first half of the MOOCs, where the number of Samplers and the number of Viewers decreased from 250 to 30 and from 215 to 120, respectively. Students belong to this subgroup are clearly the least "at-risk" students. This means that it may be not necessary to provide them with any interventions; and on the contrary, unnecessary interventions may cause these students being interrupted thus becoming less active or engaged. In another word, when providing interventions, it is crucial to have a clear target group of students, as well as to avoid interrupt the students who do not need any intervention.

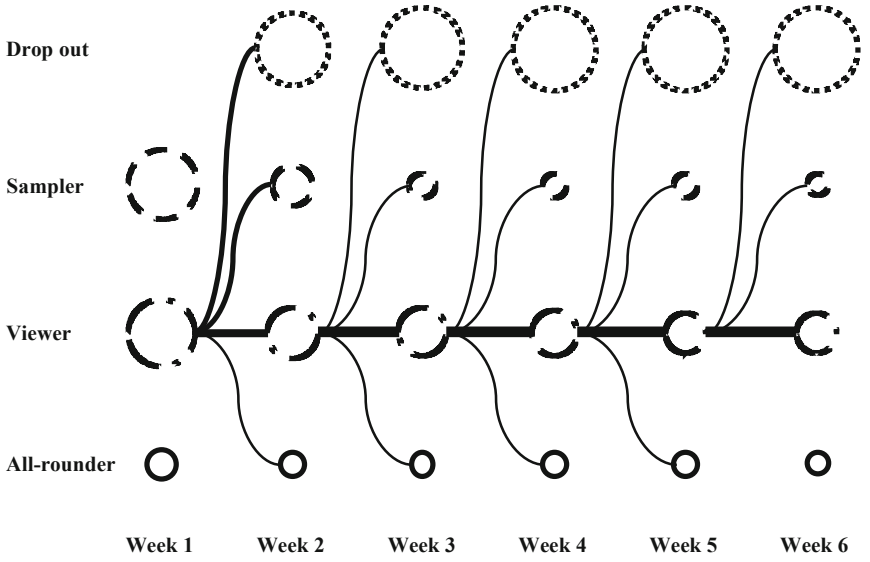

Fig. 5. Viewers' demographical changes across weeks

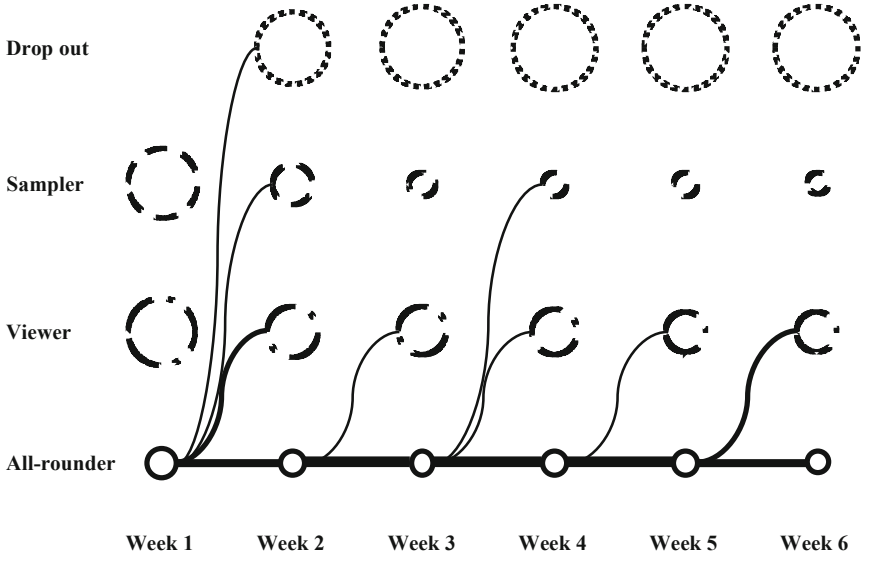

Fig. 6. All-rounders' demographical changes across weeks

5 Discussions

This chapter demonstrates a temporal analysis in Massive Open Online Courses (MOOCs), towards identifying at-risk students through analyzing their demographical changes. At-risk students are those who tend to drop out from the MOOCs. In this study,

we have examined how students' memberships of subgroups changed on a weekly time scale. Different from previous studies that used behavioral data to pre-define or cluster student subgroups, our study used both interaction log data and students' characteristics (gender, in particular).

In particular, to answer the first research question, **RQ1**, we clustered students into three distinct subgroups using the K-means++ algorithm and the "elbow method", as well as the Mann-Whitney test. Three subgroups, including *Sampler*, *Viewer* and *All-rounder*, were generalized. We have analyzed the differences amongst these subgroups and measured the proportion of students in different subgroups by week. To answer the second research question, **RQ2**, we examined the demographical changes for students labelled in each subgroup where we found that using similar cluster approaches on weekly accumulated data could generate similar subgroups as the overall clustering result. Most of the subgroup's centroid remained stable within a certain range except All-rounders with the number of comments continuously rising in the second half of the course. To answer the third research question, **RQ3**, we visualized the demographical changes of subgroups across weeks. Our result suggests that the first half of the course, i.e. Week 1 to Week 3, and the second half of the course, i.e. Week 4 to Week 6, had different demographical characteristic. The demographics of these subgroups changed significantly from the first half of the former and maintained a certain degree of stability in the second half. More specifically, our study suggests that the less active subgroups took up most of the participants in the early courses, and as the course progressed, the proportion of those subgroups continued shrinking to around 10% (see Fig. 2). This result is opposite to those from previous studies which assume proportion of participants falling into each category keep stable to some extent along courses progress.

For the transition of each subgroup, our result demonstrates that each of them had similar transitional pattern along the MOOCs progressed – most of the Samplers dropped out in the subsequent week with only a small percentage of them kept Sampler's behavior unchanged or transited into the Viewer subgroup. A large proportion of the Viewers maintained the same behavior pattern to a subsequent week, and a relatively small percentage of these students transited to the Sampler or All-rounder subgroups, or simply dropped out. The All-rounder was the most stable subgroup – the demographical characteristics stabled from the beginning to the end of the MOOCs, i.e. from Week 1 to Week 6. Interestingly, the result in Sect. 4.3 suggests that it was almost impossible for the students to switch from being highly inactive (Sampler, as in this study) to being highly active (All-rounder, as in this study) in a short period of time, and vice-versa. Therefore, once being detected or identified as inactive, these students should be strongly intervened, and as early as possible, in order to prevent them from dropping out from the MOOC; whereas for the active students, strong intervention may be not necessary, but mild interventions may be still useful to keep them active, as discussed in Sect. 4.3.

6 Conclusions

To conclude, in this study we have analyzed students' data from two MOOCs offered by the FutureLearn platform. The result suggests that the first half and second half of both MOOCs had different demographical characteristics and each student subgroup had

their unique behavior and transitional pattern along the MOOCs progressed. Given the fact that MOOC students have various study behavior, with a very different interaction patterns with the course materials and their peers, when designing MOOCs, there is a strong need for providing personalized support to students that can be labelled into different subgroup at different stages of the MOOC. This means that the MOOC platforms should personalize the way their users learn, such as adapting learning paths and supporting adaptive intervention for different subgroups of students. Moreover, the subgroups identified in this study and the weekly demographical changes of those clusters may help inform a range of strategies for the intervention and improvement of MOOCs and MOOC platforms. For example, providing more previews of learning materials allows Sampler students to make a more informed decision about whether to participate in the first place. Offering more reminders for students who labelled as Sampler on unfinished steps and reduce the incentives for their comment submissions.

This study contributes to the understanding of subgroup clustering and demographical changes in MOOCs. Empirical evidence from this study supports that students' characteristics can also be used as clustering variables/indicators, and the proportion of different subgroups in the total number of students each week may vary along the MOOCs progress. These results highlight the importance of examining subgroup to improve the effectiveness of the identification of at-risk students.

In future studies, the same research approach could be applied into MOOCs with more general content where there are more attributes with less association with students' interaction data (the number of steps that a student visited and the number of comments that a student submitted, as in current study). In this study, the course "Leadership for healthcare improvement and innovation" does not contain any assignment, hence the assessment factor was not considered in subgroup clustering. In a future study, the assignment submission and grade could also be considered as clustering variables/indicators.

In terms of limitations, first, the dataset available was limited – after removing students with incomplete information, only 506 students' data was retained, and those students might share different characteristics with eliminated students. Second, the filed involved in the MOOCs used in this study were highly targeted. Third, the MOOCs that we were focused on were unique in duration and structure in which students needed to access both a large number of steps and tools supporting reflection, comment and response. Therefore, the conclusion drawn from the analysis of the dataset may be not universally applicable to a MOOC in the other fields.

References

1. Alamri, A., et al.: Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In: Coy, A., et al. (eds.) Intelligent Tutoring Systems, pp. 163–173. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-22244-4_20
2. Alshehri, M., et al.: On the Need for fine-grained analysis of gender versus commenting behaviour in MOOCs. In: Proceedings of the 2018 the 3rd International Conference on Information and Education Innovations - ICIEI 2018, pp. 73–77. ACM Press, London (2018). https://doi.org/10.1145/3234825.3234833

3. Anderson, A., et al.: Engaging with massive online courses. In: Proceedings of the 23rd International Conference on World Wide Web - WWW 2014, Seoul, Korea, pp. 687–698. ACM Press (2014). https://doi.org/10.1145/2566486.2568042
4. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. http://ilpubs.stanford.edu:8090/778/. Accessed 05 Mar 2020
5. de Barba, P.G., et al.: The Role of students' motivation and participation in predicting performance in a MOOC: motivation and participation in MOOCs. J. Comput. Assist. Learn. **32**(3), 218–231 (2016). https://doi.org/10.1111/jcal.12130
6. Beck, J.E., Woolf, B.P.: High-level student modeling with machine learning. In: Gauthier, G., et al. (eds.) Intelligent Tutoring Systems, pp. 584–593. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45108-0_62
7. Bote-Lorenzo, M.L., Gómez-Sánchez, E.: Predicting the decrease of engagement indicators in a MOOC. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017, Vancouver, British Columbia, Canada, pp. 143–147. ACM Press (2017). https://doi.org/10.1145/3027385.3027387
8. Brinton, C.G., et al.: Mining MOOC clickstreams: on the relationship between learner behavior and performance. arXiv preprint arXiv:1503.06489 (2015)
9. Chiu, T.K.F., Hew, T.K.F.: Factors influencing peer learning and performance in MOOC asynchronous online discussion forum. Australas. J. Educ. Technol. (2017). https://doi.org/10.14742/ajet.3240
10. Clow, D.: MOOCs and the funnel of participation. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK 2013, Leuven, Belgium, p. 185. ACM Press (2013). https://doi.org/10.1145/2460296.2460332
11. Coffrin, C., et al.: Visualizing patterns of student engagement and performance in MOOCs. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge - LAK 2014, Indianapolis, Indiana, pp. 83–92. ACM Press (2014). https://doi.org/10.1145/2567574.2567586
12. Corrin, L., et al.: Using learning analytics to explore help-seeking learner profiles in MOOCs. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK 2017, Vancouver, British Columbia, Canada, pp. 424–428. ACM Press (2017). https://doi.org/10.1145/3027385.3027448
13. Cristea, A.I., et al.: Can learner characteristics predict their behaviour on MOOCs? In: Proceedings of the 10th International Conference on Education Technology and Computers - ICETC 2018, pp. 119–128, Tokyo, Japan. ACM Press (2018). https://doi.org/10.1145/3290511.3290568
14. Cristea, A.I., et al.: Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses. In: Presented at the 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden, 22 August 2018
15. Cristea, A.I., et al.: How is learning fluctuating? FutureLearn MOOCs fine-grained temporal analysis and feedback to teachers and designers. In: 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden. Association for Information Systems (2018)
16. Evans, B.J., et al.: Persistence patterns in massive open online courses (MOOCs). J. High. Educ. **87**(2), 206–242 (2016). https://doi.org/10.1353/jhe.2016.0006
17. Feng, W., et al.: Understanding dropouts in MOOCs. In: AAAI, vol. 33, pp. 517–524 (2019). https://doi.org/10.1609/aaai.v33i01.3301517
18. Ferguson, R., Clow, D.: Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK 2015, Poughkeepsie, New York, pp. 51–58 ACM Press (2015). https://doi.org/10.1145/2723576.2723606

19. Halawa, S., et al.: Dropout prediction in MOOCs using learner activity features. In: Proceedings of the Second European MOOC Stakeholder Summit, vol. 37, no. 1, pp. 58–65 (2014)
20. Huang, Z.: Extensions to the k-Means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. **2**(3), 283–304 (1998). https://doi.org/10.1023/A:1009769707641
21. Jung, Y., Lee, J.: Learning engagement and persistence in massive open online courses (MOOCS). Comput. Educ. **122**, 9–22 (2018). https://doi.org/10.1016/j.compedu.2018.02.013
22. Khalil, M., Ebner, M.: Clustering patterns of engagement in massive open online courses (MOOCs): the use of learning analytics to reveal student categories. J. Comput. High. Educ. **29**(1), 114–132 (2017). https://doi.org/10.1007/s12528-016-9126-9
23. Khalil, M., Ebner, M.: What massive open online course (MOOC) stakeholders can learn from learning analytics? arXiv:1606.02911 [cs], pp. 1–30 (2016). https://doi.org/10.1007/978-3-319-17727-4_3-1
24. Kizilcec, R.F., et al.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 170–179. ACM (2013)
25. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in K-Means clustering. Int. J. **1**(6), 90–95 (2013)
26. Kovanović, V., et al.: Profiling MOOC course returners: how does student behavior change between two course enrollments? In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S 2016, Edinburgh, Scotland, UK, pp. 269–272. ACM Press (2016). https://doi.org/10.1145/2876034.2893431
27. Li, B., et al.: What makes MOOC users persist in completing MOOCs? a perspective from network externalities and human factors. Comput. Hum. Behav. **85**, 385–395 (2018). https://doi.org/10.1016/j.chb.2018.04.028
28. Li, Q., Baker, R.: The different relationships between engagement and outcomes across participant subgroups in massive open online courses. Comput. Educ. **127**, 41–65 (2018). https://doi.org/10.1016/j.compedu.2018.08.005
29. Liao, J., et al.: Course drop-out prediction on MOOC platform via clustering and tensor completion. Tinshhua Sci. Technol. **24**(4), 412–422 (2019). https://doi.org/10.26599/TST.2018.9010110
30. Likas, A., et al.: The global k-means clustering algorithm. Pattern Recogn. **36**(2), 451–461 (2003). https://doi.org/10.1016/S0031-3203(02)00060-2
31. Maldonado-Mahauad, J., et al.: Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning. In: Pammer-Schindler, V., et al. (eds.) Lifelong Technology-Enhanced Learning, pp. 355–369. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-98572-5_27
32. Peng, X., Xu, Q.: Investigating learners' behaviors and discourse content in MOOC course reviews. Comput. Educ. **143**, 103673 (2020). https://doi.org/10.1016/j.compedu.2019.103673
33. Pursel, B.K., et al.: Understanding MOOC students: motivations and behaviours indicative of MOOC completion: MOOC student motivations and behaviors. J. Comput. Assist. Learn. **32**(3), 202–217 (2016). https://doi.org/10.1111/jcal.12131
34. Reich, J., Ruipérez-Valiente, J.A.: The MOOC pivot. Science **363**(6423), 130–131 (2019). https://doi.org/10.1126/science.aav7958
35. Rieber, L.P.: Participation patterns in a massive open online course (MOOC) about statistics: MOOC participation. Br. J. Educ. Technol. **48**(6), 1295–1304 (2017). https://doi.org/10.1111/bjet.12504
36. Salmon, G., et al.: Designing massive open online courses to take account of participant motivations and expectations: designing MOOCs. Br. J. Educ. Technol. **48**(6), 1284–1294 (2017). https://doi.org/10.1111/bjet.12497

37. Sanz-Martínez, L., et al.: Creating collaborative groups in a MOOC: a homogeneous engagement grouping approach. Behav. Inform. Technol. **38**(11), 1107–1121 (2019). https://doi.org/10.1080/0144929X.2019.1571109
38. Shi, L., et al.: Revealing the hidden patterns: a comparative study on profiling subpopulations of MOOC Students. In: The 28th International Conference on Information Systems Development (ISD2019), Toulon, France. Association for Information Systems (2019)
39. Shi, L., et al.: Social engagement versus learning engagement - an exploratory study of FutureLearn Learners. In: Presented at the 14th IEEE International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2019), Dalian, China November (2019)
40. Shi, L., et al.: Towards understanding learning behavior patterns in social adaptive personalized e-learning systems. In: The 19th Americas Conference on Information Systems, Chicago, Illinois, USA, pp. 1–10 Association for Information Systems (2013)
41. Shi, L., Cristea, A.I.: Demographic indicators influencing learning activities in MOOCs: learning analytics of FutureLearn courses. In: Presented at the 27th International Conference on Information Systems Development (ISD2018), Lund, Sweden, 22 August 2018
42. Shi, L., Cristea, A.I.: In-depth exploration of engagement patterns in MOOCs. In: Hacid, H., et al. (eds.) Web Information Systems Engineering – WISE 2018, pp. 395–409. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-02925-8_28
43. Sunar, A.S., et al.: How learners' interactions sustain engagement: a MOOC case study. IEEE Trans. Learning Technol. **10**(4), 475–487 (2017). https://doi.org/10.1109/TLT.2016.2633268
44. Tsai, Y., et al.: The effects of metacognition on online learning interest and continuance to learn with MOOCs. Comput. Educ. **121**, 18–29 (2018). https://doi.org/10.1016/j.compedu.2018.02.011
45. Xing, W., Du, D.: Dropout prediction in MOOCs: using deep learning for personalized intervention. J. Educ. Comput. Res. **57**(3), 547–570 (2019). https://doi.org/10.1177/0735633118757015
46. 1st International Conference on Learning Analytics and Knowledge 2011|Connecting the Technical, Pedagogical, and Social Dimensions of Learning Analytics. https://tekri.athabascau.ca/analytics/. Accessed 01 Mar 2020