Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics

Mohammad KHALIL, Christian KASTL & Martin EBNER

Graz University of Technology, Educational Technology,
{mohammad.khalil, martin.ebner}@tugraz.at, kastl@sbox.tugraz.at

Abstract

Massive Open Online Courses are remote courses that excel in their students' heterogeneity and quantity. Due to the peculiarity of being massiveness, the large datasets generated by MOOCs platforms require advance tools to reveal hidden patterns for enhancing learning and educational environments. This paper offers an interesting study on using one of these tools, clustering, to portray learners' engagement in MOOCs. The research study analyse a university mandatory MOOC, and also opened to the public, in order to classify students into appropriate profiles based on their engagement. We compared the clustering results across MOOC variables and finally, we evaluated our results with an eighties students' motivation scheme to examine the contrast between classical classes and MOOCs classes. Our research pointed out that MOOC participants are strongly following the Cryer's scheme of ELTON (1996).

Keywords

MOOCs, Learning Analytics, Clustering, Engagements, Patterns

1 Introduction

In the last years, Technology Enhanced Learning (TEL) has been developed rapidly so that now is including modern online classes in which they are called MOOCs (MCAULEY et al., 2010). The word MOOCs is an abbreviation of four letters, 'M' which is Massive, and it means massive in the number of enrollees than what is in regular classes. 'O' and this is Open, and that is an implication of a field that has no accessibility limitations. Furthermore, openness also means that these massive courses should be open to anyone. The second 'O' stands for Online where all courses are held on the Internet without any borders. In the end, 'C' means courses, this represented a structured learning material and is mostly embodied as filmed lectures, documents and interactive social media such as discussion forums or even social media channels.

The first version of MOOCs was named cMOOCs, which were developed by George Siemens and Stephan Downes back in 2008, and it adopted the connectivism theory that is based on the role of social and networks of information (HOLLAND & TIRTHALI, 2014). After that, other versions of MOOCs become available, but it was noticeable that the *extended MOOCs* or so-called *xMOOCs* attracted the eyes of today's online courses learners.

One of the prominent and most successful activities of xMOOCs has been done by Sebastian Thrun in 2011. He and his colleagues launched an online course called "Introduction to Artificial Intelligence" which attracted over 160,000 users from all over the world (YUAN et al., 2013). xMOOCs follow theories that are based on guided learning and the classical information transmission (RODRIGUEZ, 2012). FERGUSON & CLOW (2015) argued that xMOOC is an extended version of cMOOC with additional elements of content and assessment as well as a larger-scale role of educators to be part of the content; in other words, an online course for hundreds of learners simultaneously (CARSON & SCHMIDT, 2012).

The benefits of MOOCs are crystallized to be welfare in improving educational outcomes, extending accessibility, and reducing costs. In addition, Ebner and his colleagues addressed the advantages that MOOCs can add to the Open Educational Resources (OER) movement as well as lifelong learning experiences in TEL contexts (Ebner, et al., 2014). Despite their advantages, MOOCs suffered from students who register and afterwards do not complete the courses. This has been cited in several scientific researches and is now commonly named as "the dropout rate" (MEYER, 2012; JORDAN, 2013). Various investigations have been done to identify the reasons behind the low completion rates, such as the research studies by KHALIL & EBNER (2014; 2016), LACKNER et al. (2015). Furthermore, lack of interaction between learners and instructor(s), and the controversy argument about MOOCs pedagogical approach, are the negative factors that obstruct the positive advancement of MOOCs. In addition to all this, recent research publications discussed the patterns of engagement and the debates about categorizing students in MOOCs (KIZILCEC et al., 2013; FERGUSON & CLOW, 2015; KHALIL & EBNER, 2015a).

Since MOOCs include a large quantity of data that is generated by students who reside in an online crucible, the heed toward what is so-called Learning Analytics steered the wheel into an integration of both sectors (KHALIL & EBNER, 2016). KNOX (2014) discussed the high promises behind Learning Analytics when it is applied to MOOCs datasets for the principles of overcoming their constraints. The needs for Learning Analytics emerged to optimize learning, and for a better students' commitment in distance education applications (KHALIL & EBNER, 2015b).

In this research study, we employ Learning Analytics, using a clustering methodology, on a dataset from one of the courses offered by the leading Austrian MOOC platform, iMooX[1]. The sought objectives behind clustering are to portray the engagement and behaviour of learners in MOOC platforms and to support decisions of following up the students for purposes of increasing retention and improving interventions for a specific subpopulation of students. In addition, this research study will contribute with an additional value to ease the grouping of MOOCs participants.

The publication is organized as follow: Section 2 covers the research methodology of this research study. Section 3 gives an overview about the MOOC platform itself as well as the demographics of the course. Section 4 covers in details the clustering methodology and data analysis. Section 5 is the discussion and the comparison with the Cryer's scheme, while section 6 concludes the findings.

[1] http://www.imoox.at (last visited October 2015)

2 Research Methodology

This research study is based on data collected by a formal Learning Analytics application of the iMooX MOOC-platform. By tracking their traces, the application records learners actions within the divergent MOOC indicators such as videos, files downloads, reading in forums, posting in forums and the quizzes performance. In the present study, a MOOC named "Social Aspects of Information Technology", shortly *GADI* (abbreviated from the original German title), was chosen for further analysis and research.

The collected information after that, which takes the form of log files, was parsed to filter the duplicated and unstructured data format. The data analysis was carried out using the R software, and the clustering methodology was performed using an additional package called NbClust (CHARRAD et al., 2014). We followed content analysis in which units of analysis get measured and benchmarked based on qualitative decisions (NEUENDORF, 2002). These decisions are founded on sustained observations on a weekly basis and examination of surveys at the end of the course by one of the researchers.

3 Stats and Overview

3.1 The MOOC-Platform

iMooX is the leading Austrian MOOC platform founded by the cooperation of Graz University of Technology and University of Graz (NEUBÖCK et al., 2015). The offered courses vary in topics between social science, engineering and technology topics and cope with lifelong learning and OER tracks. The target groups are assorted among school children, high-school students and university degree holders. Additionally, iMooX offers certificates and badges to successful students who fulfilled courses requirements at no cost.

3.2 Course Overview and Demographics

Our analysis of portraying learners is based on a summer course provided by Graz University of Technology in 2015 called “Social Aspects of Information Technology” abbreviated in German and in this research study as GADI. This course was selected because it is specialized of being mandatory for the university students of Information and Computer Engineering (Bachelor-6th semester), Computer Science (Bachelor-2nd Semester), Software Development, Business Management (Bachelor-6th semester) and for the Teacher’s Training Certificate of computer science degree (2nd Semester). Furthermore, the course was also opened for external participants and not only restricted to university students. The main content of the course is based on discussions about the implications of information technology on society.

The course lasted 10 weeks. Every week includes 2 or 3 video lectures, a discussion forum, further readings and a multiple choice quiz. Each quiz could be repeated up to five times. The system is programmed to record the highest grade of these trials. MOOC’s workload was predefined with about 3 hours/week, and the passing grade for each quiz was set to be 75%. Students of Graz University of Technology gain 2.5 ECTS (credits) for completing the MOOC but they have also to do an additional essential practical work.

Finally, there were in summary 838 participants in the course, 459 of them were university students, while 379 were voluntary external participants. Because this MOOC is obligatory to pass the university class, the completion ratio was much higher compared to other MOOCs. The general certification rate of this particular MOOC is 49%. The certification ratio of the university students was 80%, and 11.35% of the external participants.

Candidates, who successfully completed all quizzes, were asked to submit answers for a predefined evaluation form. The collected data showed that most of the external participants are from Austria and Germany. University students’ average age was 23.1 years old, while the average age of the external participants was 46.9 years old. Table 1 reports the course demographics based on the evaluation results.

Table 1: The GADI MOOC Demographics of completed participants

	Gender (M/F)	High School	Bachelors	MSc & PhD	Others
Students	327/40	357	4	4	2
Others	23/20	13	3	3	4

4 Clustering and Analysis

The main goal behind clustering is to assign each participant in the MOOC to a suitable group with common behaviours. Each group should be as distinct as possible to prevent overlaps. The elements in these groups should fit tied to the defined group parameters. Therefore, clustering using the k-mean algorithm with the Euclidean distance was selected as our tool of choice. In order to begin clustering, we labeled the variables that will be referenced in the algorithm. The expected results should be clustered with activities and characteristics that distinguish the MOOC participants.

Due to the relations between certain variables, we excluded the high correlated indicators as this will not affect the grouping sequence. As a consequence, the used variables in clustering were:

1. Reading Frequency: This indicates the number of times a user clicked on particular posts in the forum.
2. Writing Frequency: This variable determines the number of written posts in the discussion forum.
3. Videos Watched: This variable contains the total number of videos a user clicked.
4. Quiz Attempts: It calculates the sum of attempts that have been spent on all ten quizzes.

Because of the structure of the examined MOOC, which is obligatory for university students and opened for external participants, the clustering was done independently in both groups. The intention of each group could vary. For example, are the university students attending the MOOC for learning purposes or are they *only seeking* for the grade?

4.1 Case 1: University Students

In this case, the k value was assigned with a value from 3 to 6, as long as we do not really want more than 6 groups. The suggested cluster, based on the variables value and the NbClust package, resulted to four clusters. Figure 1 illustrates the four clusters of the MOOC university students.

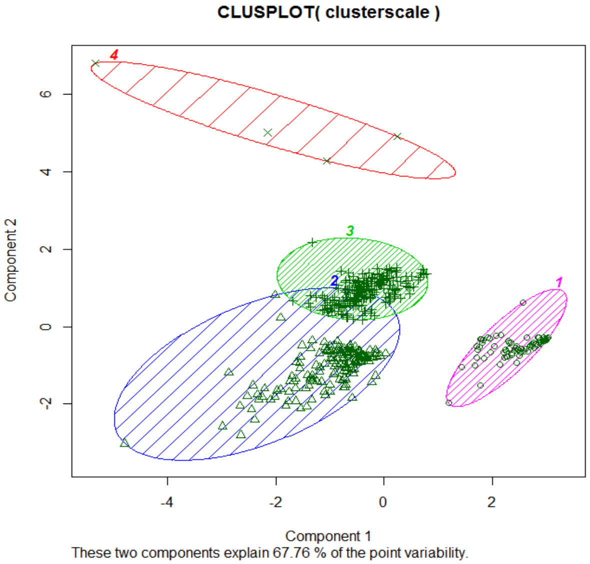

Figure 1: MOOC's University Students Clusters

Figure 1 shows a cluster amount of four classes. Two of the groups, the blue and the green are overlapping. The relation between components in x-axis and y-axis is valued at 67.76%. This percentage means that we have nearly 70% of unhidden information based on this clustering value[2]. The clusters are characterized as the following:

Cluster (1) with the pink oval shape contains 95 students. This group has low activity among the four variables. Only 10 students are certified, and the dropout rate is high.

Cluster (2) with the blue oval shape contains 154 students. Most of the participants in this group completed the course successfully. This cluster is distinguishable by their videos' watching.

Cluster (3) with the green oval shape has 206 participants. The certification rate was 94%. Both of cluster 2 and cluster 3 share a high certification rate, but differ in watching the videos.

Cluster (4) is the smallest cluster, containing 4 students. By observing the variables, we noticed that the students in this cluster are the only ones that had been writing on the forums. The amount of certified students in cluster 4 totals to 50%.

4.2 Case 2: External Learners

Figure 2 shows the proposed cluster solution of the external participants who do not belong to the university class. Again, k value was set to be from 3 to 6. The point variability shows a competitive rate of 88.89%, which indicates a steep seclusion among the three groups. The clusters of this case are characterized as the following:

Cluster (1) with the blue oval shape contains 42 participants. The certification rate of this group is 76.20%. The social activity and specifically reading in forums are moderate compared to the other clusters. Whiles the number of quiz trials is high.

Cluster (2) with the red oval shape holds only 8 participants. The certification rate in this group is 100%. Participants from cluster 2 showed the highest number of written contributions and the highest reading frequency in the forum.

[2] Explanation: http://stats.stackexchange.com/questions/141280/understanding-cluster-plot-and-component-variability (Last accessed, 15th October 2015).

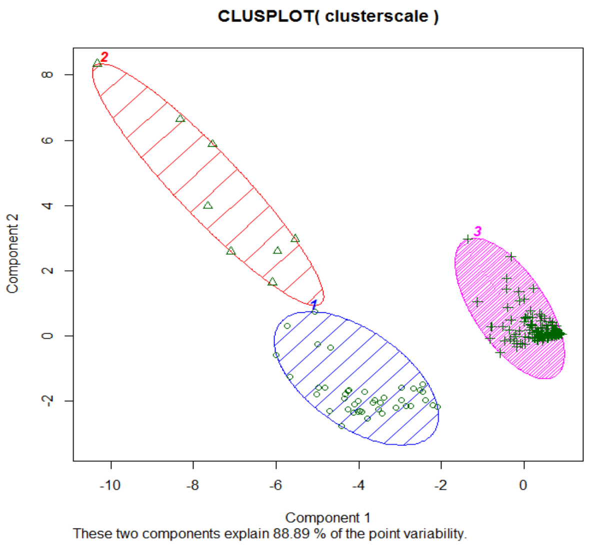

Figure: 2 MOOC's External Learners Clusters

Cluster (3) with the pink oval shape includes all the other participants. This group showed a high dropout rate and a completion rate of only 1%.

5 Discussion

Within the previous clustering results in both cases, we studied the values of each variable in each cluster. The next step was to make a classification scale of "low", "moderate" and "high" that describes characteristics and the activity level of each group. Table 2 shows them for both of the cases, university and external participants.

Table 2: Characteristics of each cluster of both MOOC cases

Case: University Students					
	Reading Freq.	Writing Freq.	Watching Videos	Quiz Attempts	Certification Ratio
Cluster 1	Low	Low	Low	Low	10.53%
Cluster 2	High	Low	High	High	96.10%
Cluster 3	Moderate	Low	Low	High	94.36%
Cluster 4	High	High	Low	Moderate	50%
Case: External Participants					
Cluster 1	Moderate	Low	Moderate	High	76.19%
Cluster 2	High	High	High	High	100%
Cluster 3	Low	Low	Low	Low	1%

By analyzing the clusters, we think the opportunity to portray students' behaviours in the MOOC becomes possible nearby. However, a study by ELTON in (1996), which examined the general strategies to motivate learners in the classes, meets a similar scheme of our clustering results. Figure 3 illustrates the so-called Cryer's scheme, which shows student behavior within a course. The x-axis represents intrinsic factors, which are achievements and subject. The y-axis includes the examination preparation, which is named as the extrinsic factor. It must be stated that this scheme does not only include the shown specific profiles, but it also contains other learners who reside between these four profiles.

The students, on the bottom left of the Cryer's scheme, describe the ones who are not interested in the course subject nor score positive results.

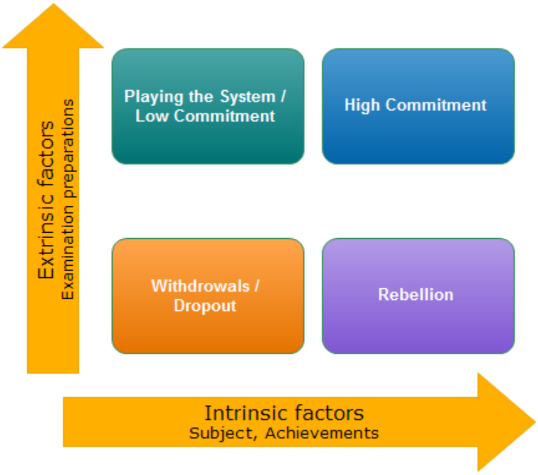

Figure 3: Cryer’s Scheme Based on Levels of Student Commitment

This class represents Cluster (1) of our university students’ case, and Cluster (3) of external participants’ case. An appropriate profile name of this cluster would be simply “Dropout”. This profile shares common patterns of being inactive among all the MOOC variables. The certification rate in this profile is low.

The class, on the top left in the scheme, describes learners who *play the system*. This term comes from a case when students are treated and just doing what instructors want to do for getting a grade. Using Learning Analytics, some students were determined watching the learning videos with various skips, or even they start a quiz without watching the weekly video. Such students were named as “Gamblers”. In spite of certain questions that are hard to answer without watching a video, some of them could pass the exams. It should also be considered that the MOOC platform offers up to five trials per week quiz, which might be the reasons behind a high percentage of *gamblers* among university students.

Rebellions are those who show interest in the course, but fail because of bad exam preparations. In the Cluster Analysis, this group was available in the university students’ group, which is represented by Cluster (4). However, it was hard to detect in the external participants' group. Cluster (4) was distinct for being very active with the social activities in the forums. We named them as the "Sociable Students".

The last class is the students whom their commitment is high. "Perfect Students" might be the appropriate name for them. Every MOOC platform looks to have such students. With their high certification rate, Cluster (2) in both cases embodied this profile.

6 Conclusion

This research study examined learners' behavior in a mandatory xMOOC offered by iMooX. Because the course was also opened to the public, we studied patterns of the involved students and separated them into two cases, internal and external participants. Within our research study, we performed a cluster analysis, which pointed out participants in MOOCs, whether they did the course on a voluntary basis or not. Furthermore, we found that the clusters can be applied on the Cryer's scheme of ELTON (1996). This leads to the assumption that tomorrow's instructors have to think about the increase of the intrinsic motivation by those students who are only "playing the system". Our research study also pointed out that online courses behave very similar to traditional face-to-face courses. Therefore, we strongly recommend researching on how MOOCs can be more engaging and creating new didactical concepts to increase motivational factors.

References

Carson, S., & Schmidt, J. (2012). The Massive Open Online Professor Academic Matter. *Journal of higher education*.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2013). *NbClust: An examination of indices for determining the number of clusters: NbClust Package*.

Ebner, M., Kopp, M., Wittke, A., & Schön, S. (2014). Das O in MOOCs – über die Bedeutung freier Bildungsressourcen in frei zugänglichen Online-Kursen. *HMD Praxis der Wirtschaftsinformatik, 52*(1), 68-80. Springer, December 2014.

Elton, L. (1996). Strategies to enhance student motivation: a conceptual analysis. *Studies in Higher Education*, *21*(1), 57-68.

Ferguson, R., & Clow, D. (2015). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 51-58). ACM.

Hollands, F. M., & Tirthali, D. (2014). *MOOCs: Expectations and reality*. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Jordan, K. (2013). *MOOC Completion Rates: The Data*. Retrieved October 2015, from http://www.katyjordan.com/MOOCproject.html

Khalil, H., & Ebner, M. (2014). Moocs completion rates and possible methods to improve retention-a literature review. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications* (Vol. 2014, No. 1, pp. 1305-1313).

Khalil, M., & Ebner, M. (2015a). A STEM MOOC for School Children – What Does Learning Analytics Tell us? In *Proceedings of 2015 International Conference on Interactive Collaborative Learning, Florence, Italy.* IEEE.

Khalil, M., & Ebner, M. (2015b). Learning Analytics: Principles and Constraints. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1326-1336).

Khalil, M., & Ebner, M. (2016). What can Massive Open Online Course (MOOC) Stakeholders Learn from Learning Analytics? *Learning, Design, and Technology. An International Compendium of Theory, Research, Practice, and Policy*. Springer. Accepted, in print.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). 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.

Lackner, E., Ebner, M., & Khalil, M. (2015). MOOCs as granular systems: design patterns to foster participant activity. *eLearning Papers*, *42*, 28-37.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). Massive Open Online Courses Digital ways of knowing and learning. *The MOOC model For Digital Practice*. Retrieved October 2015, from http://www.elearnspace.org/Articles/MOOC_Final.pdf

Meyer, R. (2012). *What it's like to teach a MOOC (and what the heck's a MOOC?)*. Retrieved October 2015, from http://tinyurl.com/cdfvvqy

Neuböck, K., Kopp, M., & Ebner, M. (2015). What do we know about typical MOOC participants? First insights from the field. In M. Lebrun, I. de Waard, M. Ebner, & M. Gaebel (Eds.), *Proceedings of eMOOCs 2015 conference* (pp. 183-190). Mons, Belgium.

Neuendorf, K. A. (2002). *The content analysis guidebook* (Vol. 300). Thousand Oaks, CA: Sage Publications.

Rodriguez, C. O. (2012). MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses. *European Journal of Open, Distance and E-Learning*.

Yuan, L., Powell, S., & Cetis, J. (2013). *MOOCs and open education: Implications for higher education*.