PREDICTING PERFORMANCE AND ANALYZING STUDENTS' BEHAVIOR FROM COMPUTER-GENERATED LOG FILES IN MOODLE

Lan Umek, Damijana Keržič, Nina Tomaževič, and Aleksander Aristovnik
University of Ljubljana
Slovenia

Abstract

This paper presents the results of the investigation into whether the student's behavior in the e-classroom in terms of action logs correlates with their academic performance. The empirical research was based on logs obtained from Moodle, which is used to manage blended learning at the faculty. The main finding from this research is that there is some relationship between some types of interactions and academic performance in selected online courses. The findings of the paper confirm that by monitoring Moodle activity data, lecturers can identify weak students and promptly adjust and individually support their pedagogical activities during the semester.

Introduction

Blended learning (BL), a combination of traditional face-to-face learning and technology-mediated instruction, is becoming common at all levels of education, and higher education is no exception. E-learning systems supplemented by face-to-face courses are known as learning management systems (LMS), learning platform (LP), course management systems (CMS), learning content management systems (LCMS) or managed learning systems (MLS) (Romero, Ventura, & García, 2008). Although known by different names, they are all used to manage online learning and teaching. One of the most popular LMSs worldwide is Moodle, which is flexible, open source and free. On the other side, it is user-friendly and well supported (Cabero-Almenara, Arancibia, & del Prete, 2019). In online courses, teachers: deliver to students information, content and learning materials; prepare assignments or quizzes; and manage collaborative learning with discussions in forums, workshops or wikis. Online courses offer different opportunities for adapting learning processes suited to the individual student's needs, abilities and learning style.

Each time a user accesses an LMS with his/her user account, a digital trace is saved in log files. A user's behavior in online course, i.e., activities and interactions with the system or other users during learning process, is therefore recorded. Each day the LMS collects huge amounts of data in digital format and accumulates the learning history of each user in log files. But we are not limited just to data from the LMS; we can also include demographic data and other student data from the information student systems (Romero & Ventura, 2013) The challenge is how to capture, process, present and use this data to make better decisions for tomorrow (Daniel, 2017). Accordingly, a new field of research rises with big data and analytics in higher education (Baker & Inventado, 2014; Ferguson, 2012). In recent years, learning analytics (LA) has become an important research trend and, as Elias (2011) said is “the measurement, collection, analyzing and reporting of data about learners and their context for purpose of understanding and optimizing learning and the environment in which it occurs”. LA explores learning log files and other educational data, as well as learners’ profiles, to provide proposals for improving learning processes and educational outcomes (Ferguson, 2012; Conijn, Snijders, Kleingeld, & Matzat, 2017). Different data mining techniques and tools are used to analyze accumulated data to discover useful information and patterns, commonly named educational data mining (EDM), such as prediction, clustering, classification, outlier detection, relationship mining, and visualization techniques (Chatti, Dyckhoff, Schroeder, & Thüs, 2012). These methods are used to get a more objective view on students’ behavior in online courses and evaluate this behavior to help improve the teaching and learning processes in the system (Romero et al., 2008).

The primary goal of educational institutions is to maximize the success of learners. Therefore, predicting learner performance from LMS data now presents a challenge. Many researchers investigate the possible correlations between students’ involvement in online course and their performance, usually based on the final grade for the course. Due to a wide range of online sources and activities, such as announcements, links, lecture notes, files, resources, questions and answers forums, discussion forums, quizzes, group works, wikis or assignment submissions, several studies have already explored impacts of various online events in different educational environments and courses. However, there is no general conclusion about the single best way yet to predict the performance or the online behaviour of potential students at risk (Conijn et al., 2017).

In the current study we analyze if there exists a relationship between students’ behavior in the e-classroom (number of logs in the e-classroom, number of visited activities, etc.) and their academic performance as measured with the final grade. The paper is arranged in the following way. First, there is a literature review or related works, followed by the empirical study, including the description of data, methodology and results. The paper concludes with key findings.

Related Works

BL educational data comes from different resources: traditional face-to-face classroom and online course environment, where the second source of data is a richer source of information about the learning process. In recent years, many studies analyzed LMS data in order to predict a student’s academic performance and usually take into account the final grade or simply whether the student has passed the final exam or not (Conijn et al., 2017; Romero, López, Luna, & Ventura, 2013; Romero, Espejo, Zafra, Romero, & Ventura, 2010; Zacharis, 2015). Studies have addressed different types of courses and various selected variables taken from LMS data; therefore, comparison of results is difficult. The general conclusions about the predictors could not be readily deduced (Conijn et al., 2017). Some studies attempt to detect patterns of students' behavior in online courses and discover students who are at risk more than likely will not complete the course (Félix, Ambrósio, Neves, Siqueira, & Brancher, 2017). Such predictions could help teachers to take steps towards bringing the students back to the course before is too late.

Determining the factors affecting academic performance is the focus of many studies. Many researchers seek associations between learners who passed or failed a course and student learning through collaboration or interaction in the online course, namely forum posts, quiz attempts and assignment submissions. This is certainly not surprising, as in those activities, a student actively participates and demonstrates the acquired knowledge.

In studying Moodle log files, Zacharis (2015) investigated online activities in a BL course trying to predict final grades. The focus was on the usage (or participation) time on the activities. Findings reveal that a high level of communication in online activities, such as posting messages on forums and content creating (wiki, blog), strongly correlates with final course success. Similar conclusions were made by Romero et al. (2013) when investigating different data mining technique to predict university students' final performance based on their participation in an online discussion forum in a first-year computer science course. They were able to make an early prediction if a student will pass or fail at the end of the course, considering only students' posts with a subject's content. A very different conclusion was made from a survey that included 17 different courses, namely, a discussion forum and wiki usage had the lowest percentage of significant correlations with final exam grade (Conijn et al., 2017).

Quiz activity in LMS Moodle can be configured in various forms with automatically generated feedback and score marks, showing the correct answer or not; therefore, it allows learners to check understanding of the study materials and acquired knowledge almost immediately. Due to a wide range of different learning purposes for which quizzes can be used, they are an almost indispensable part of the e-course. Possible variables observed are number of quizzes viewed, number of attempts per quiz, number of quizzes failed/passed, (total) time used in a specific quiz or all quizzes etc. (e.g., Conijn et al., 2017; Rebucas Estacio & Callanta Raga Jr., 2017; Kadoić & Oreski, 2018; Romero et al., 2008; Zacharis, 2015). However, except for the case of Zacharis (2015), those preliminary surveys take into account compulsory quizzes, so all the students had to attempt them. In the case of the survey of Zacharias (2015), quizzes were optional with unlimited access with no impact on final score. His results revealed that the higher final course scores of students were associated with a higher number of attempts in online quizzes. Therefore, it can be deduced that only motivated students used quizzes to revise learned material. Using classification techniques with an if-then-else rule, Romero et al. (2008, 2010) classified students into four categories. Classification into fail or excellent categories (determined by final grades) is mainly based on the number of passed quizzes; therefore, a teacher can detect a student with learning problems to give him additional support and motivation in a timely manner.

In BL courses, self-regulation strategies in learning are important and critical to success. From LMS log data, self-regulated factors are usually measured with frequency (You, 2016): number of logins, number of content views, time spent reading pages, etc. However, the time a student spends and the number of logins and hits in an online course were found to be insignificant factors in predicting student's performance (Rebucas, Estacio, & Callanta Raga Jr., 2017). On the contrary, Kadoić and Oreski (2018) studied Moodle log data of one course and reported that the students' final grades correlate with the number of logs in the e-course. Moreover, the results revealed that students with the highest grade completed the e-course activities just before the deadline. A different conclusion was revealed by You (2016) in his study, namely, regular study has the strongest impact on the course achievement, and late submissions (negative correlation) and login sessions are in the second and third places.

In line with these studies, hereinafter, we analyzed how the students' behavior in the selected e-classroom (Basic Statistics) is related to their performance to see if the results are in line with the literature reviewed.

Research

Data and methodology

Our data sample consisted of learners from the 1st year of a professional study programme at the Faculty of Public Administration (FPA), University of Ljubljana. In each academic year, this group of students is the largest homogeneous group of students at the FPA. For our analysis we selected the course, Basic Statistics, which was held in the first semester and has plenty of activities in its e-classroom. Every week, students have three hours of face-to-face lecture, and for the remaining one hour, study materials and activities are prepared in the e-classroom. For the tutorial, three extensive assignments are prepared in e-course during the semester and the teacher gives feedback on the correctness of the solutions; in the 12 weeks, the tutorial is held in the traditional way. In the academic year 2018/19, the total number of the students enrolled in the e-classroom was 244; 52 of them never entered the e-classroom so it was impossible to analyze their behavior. Therefore, we limited our analysis to the 192 (79%) "active" students.

The e-course Basic Statistics contained 90 activities (available to all students):

- 25 quizzes which replace 15 hours of traditional face-to-face learning, i.e. their content is not held in face-to-face form;
- 21 folders with content used in the teaching/learning process – most of the folders (16 out of 21) contain slides and files that are used in face-to-face lectures; 5 folders contain files that are needed for three tutorials held in e-classroom;
- 20 links to sites in the e-course that contain hints and explanations for solving quizzes;
- 15 links to files that are used for students' self-preparation for the lectures;

- three assignments that replace six hours of traditional face-to-face tutorials;
- two forums: an announcement forum and a forum for student discussion;
- four activities that are not directly related to the course but provide interesting content related to the subject matter.

Quizzes were the most visited activities in the e-course Basic Statistics. The primary reason is that students' participation in quizzes is evaluated and represents 20% of the final grade mark. Each student has at least three attempts per quiz and the final score from a quiz is the best score out of three attempts. That stimulates the students to use multiple attempts to achieve better scores. The other activities were not visited so frequently. Although many folders contain slides and files that are used in face-to-face lectures, students typically visit these folders once and download the study material. The study materials for self-preparation are not visited so frequently since self-preparation is not obligatory and does not contribute to the final grade score. Similarly, the assignments that replace face-to-face tutorial for individual work at home and for which the teacher should review and give feedback to students about the correctness of their submitted solutions, have no influence on the final grade. The e-classroom contains two forums – the forum for students' discussion had no entries while the lecturers posted 12 topics in the announcement forum. The basic information about the course (contact information about the lecturers, students' obligations, etc.) do not belong to specific activities. They are found on the course's home page.

For the purpose of our analysis we counted how many times each student visited each activity. We collected the records from October 1st, 2018 (the beginning of the semester) to February 15th, 2019 (the end of exam period). Altogether we collected 98,213 records, which means that on average each student had 512 activities in the e-classrooms. On average, each activity has been visited 5.94 times per student.

Most of the activities belong to the logs to the e-classroom (40,865). That means that, on average, students visited the e-course Basic Statistics 213 times in the semester (i.e., an average 1.55 logs per day). The second most visited activity is the quiz intended as preparation for the first mid-term exam. It was visited 4,059-times, i.e., 21 times on average per student. Most of the quizzes appear at the list of top visited activities (Table 1). For our analysis we chose 25 activities with more than 6 visits per student on average (above the overall average which equals 5.94). Table 1 summarizes these activities with the average number of visits by student. The activities are sorted in descending order in terms of the average number of visits per student.

Table 1

25 activities in the e-classroom with the highest number of average visits per student (n=192)

Activity	Avg. visits per student
visits to the e-classroom	212.84
quiz - Definitions	21.14
quiz - Gapminder	12.27
quiz - Statistical Office (1st part)	12.21
quiz - Indices	11.99
assignment 1	10.97
quiz - Statistical Office (2nd part)	10.19
quiz - Ranking	9.08
quiz - Frequency distributions	8.71
quiz - Time series - forecasting	8.49
quiz - Review quiz (1st part)	8.43
quiz - Ranking practical	8.33
quiz - Indices practical	8.31
quiz - Excel functions (1st part)	8.23
quiz - Sampling	8.19
quiz - Correlation and regression	8.05
quiz - Hypotheses testing	7.99
quiz - Probability	7.91
assignment 2	7.89
quiz - Measures of central tendency and variability	7.41
quiz - Frequency distributions - practical	6.87
quiz - Excel functions (2nd part)	6.66
quiz - Time series - practical	6.54
quiz - Review quiz (2nd part)	6.32
quiz – Definitions of statistical terms	6.20

From the Table 1 we can see that the most-visited activities are quizzes. The exceptions are the number of visits to the e-course and the first two assignments. Most of the quizzes with the highest number of visits belong to the topics that are covered at the beginning of the course schedule. Notice that some of the quizzes have two parts (1st part and 2nd part). In contrast to the ordinary quizzes, they did not cover a specific topic but combined several. Their purpose was to prepare students for the two mid-term exams.

In the paper, we investigate how the students' behavior in the e-classroom Basic Statistics is related to their performance. For this purpose, we added a 0/1 variable describing if a student passed or failed the exam. Out of 192 active students, 126 attended the exam; 102 (81%) passed the exam, and 24 (19%) failed. We limited our survey to the 126 students who attended the exam. We compared mean number of visits for activities from Table 1 using Student's t-test for independent samples (two groups: passing and failing the exam). Due to the large number of tested hypotheses we applied the Bonferroni correction of p-values.

Results

Table 2 represents the comparison of two groups of students (passed/failed) in terms of mean values of 25 most visited activities. For each activity we calculated the average number of visits and its standard deviation for both groups. We then computed p-values using the t-test for independent samples. Significant differences (after the Bonferroni correction) are marked with stars. Table 2 is sorted in ascending order by p-values. That means that the most interesting findings appear at the top of the table.

Table 2

Comparison of mean number of visits of 25 activities between students who passed and who failed the exam. The Student's t-test was used for computation of p-values

Variable	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Significance**	**Prob.**
	pass (n=102)		fail (n=24)			
quiz - Ranking practical	14.30	8.65	3.13	4.54	6.316E-13	***
quiz - Review quiz (2nd part)	11.29	8.55	2.29	3.61	3.897E-12	***
quiz - Excel functions (2nd part)	11.86	8.04	2.75	3.85	5.607E-12	***
quiz – Definitions (2nd part)	36.16	26.22	8.38	11.92	1.763E-11	***
quiz - Hypotheses testing	14.25	10.71	3.17	4.90	4.639E-11	***
quiz - Correlation and regression	14.29	9.79	3.42	4.97	4.868E-11	***
quiz - Time series - forecasting	11.67	9.65	2.63	4.37	9.489E-10	***
quiz - Frequency distributions – practical	11.55	7.46	3.13	4.42	1.184E-09	***
quiz - Probability	13.77	10.39	3.75	5.97	4.013E-08	***
quiz - Frequency distributions	14.23	9.49	4.42	6.38	1.474E-07	***
quiz - Time series – practical	14.24	8.48	4.33	7.32	5.804E-07	***
visits	324.11	144.66	163.25	151.06	3.466E-06	***
quiz - Review quiz (1st part)	13.53	8.44	4.79	6.49	5.653E-06	***
quiz – Definitions (1st part)	10.33	7.16	3.13	4.77	7.241E-06	***
quiz - indices practical	13.45	8.16	4.88	7.85	7.901E-06	***
assignment 1	15.55	8.66	7.29	6.71	2.616E-05	***
assignment 2	12.40	10.60	4.67	7.25	9.122E-05	**
quiz - Statistical Office (2nd part)	15.58	9.90	7.79	8.04	4.881E-04	*
quiz - Sampling	14.05	11.34	5.33	8.58	5.820E-04	*

Variable	Mean	Standard Deviation	Mean	Standard Deviation	Significance	Prob.
quiz - Excel functions (1st part)	12.56	8.83	6.29	8.41	2.000E-03	*
quiz - Gapminder	17.12	11.87	9.00	9.60	2.274E-03	
quiz - Indices	17.10	11.98	10.08	10.96	9.844E-03	
quiz - Ranking	12.80	8.79	8.00	8.86	1.764E-02	
quiz - Measures of central tendency and variability	10.31	7.21	7.17	7.30	5.725E-02	
quiz - Statistical Office (1st part)	14.87	7.90	11.96	7.50	1.032E-01	
Probability level of significance: *** < 0.001, ** < 0.01, * < 0.05						

Table 2 indicates that we discovered significant differences between the two groups of students in terms of most activities. The mean number of visits is always higher in the group of students who passed the exam. This indicates that students who are more active in the e-classroom achieve better results on the final exam.

Conclusions

In the paper, we investigated how the students' behavior in the selected e-classroom (Basic Statistics) is related to their performance. The empirical findings imply that level of activity in the e-classroom is strongly related to the final success on the exam. We found that students who passed the exam had on average more visits to activities in the e-classroom compared to those who failed. We also identified which activities were the most discriminatory between the two groups. At the top of the list appear the quizzes which were designed for the preparation for the second mid-term exam (review quiz, Excel functions, definitions of statistical terms) and more advanced statistical topics (hypotheses testing, correlation and regression). The single exception is the quiz that covers topic of "Ranking". According to previous literature and our empirical results, regular monitoring of the Moodle activity data can help lecturers of the courses to identify weak students and promptly adjust and individually support their pedagogical activities during the semester.

References

Baker, R. S. J. D. & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). New York: Springer.

Cabero-Almenara, J., Arancibia, M., & del Prete, A. (2019). Technical and Didactic Knowledge of the Moodle LMS in Higher Education. Beyond Functional Use. *Journal of New Approaches in Educational Research (NAER Journal), 8*(1), 25–33. Retrieved from https://www.learntechlib.org/p/207147/

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5–6). doi:10.1504/IJTEL.2012.051815

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on learning technologies*, *10*(1), 17–29. doi: 10.1109/TLT.2016.2616312

Daniel, B. K. (2017). Overview of big data and analytics in higher education. In B. K. Daniel (Ed.), *Big data and learning analytics in higher education: Current theory and practice* (pp. 1–4). Springer.

Elias, T. (2011). *Learning Analytics: Definitions, Processes and Potential*. Retrieved from https://pdfs.semanticscholar.org/732e/452659685fe3950b0e515a28ce89d9c5592a.pdf

Félix, I., Ambrósio, A., Neves, P., Siqueira, J., & Brancher, J. (2017). Moodle Predicta: A data mining tool for student follow up. In P. Escudeiro, G. Costagliola, S. Zvacek, J. Uhomoibhi, & B. M. McLaren (Eds.), *Proceedings of the 9th International Conference on Computer Supported Education.* Vol. 1: CSEDU (pp. 339–346). Porto, Portugal. doi: 10.5220/0006318403390346

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317. doi: 10.1504/IJTEL.2012.051816

Kadoić, N. & Oreški, D. (2018). Analysis of student behavior and success based on logs in Moodle. In K. Skala (Ed.), *Proceedings of the 41st International Convention on ICT, Electronics and Microelectronics MIPRO 2018* (pp. 730–725). Retrieved from https://urn.nsk.hr/urn:nbn:hr:211:302374

Rebucas Estacio, R. & Callanta Raga Jr, R. (2017). Analyzing students online learning behavior in blended courses using Moodle. *Asian Association of Open Universities Journal*, *12*(1), 52–68. doi:10.1108/AAOUJ-01-2017-0016

Romero, C. & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery, 3,* 12–27. doi: 10.1002/widm.1075

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384. doi: 10.1016/j.compedu.2007.05.016

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458–472. doi: 10.1016/j.compedu.2013.06.009

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2010). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135–146. doi: 10.1002/cae.20456

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *Internet and Higher Education*, *27*, 44–53. doi: 10.1016/j.iheduc.2015.05.002

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and higher education*, *29*, 23–30. Doi: 10.1016/j.iheduc.2015.11.003

Author Details

Lan Umek
lan.umek@fu.uni-lj.si

Damijana Keržič
damijana.kerzic@fu.uni-lj.si

Nina Tomaževič
nina.tomazevic@fu.uni-lj.si

Aleksander Aristovnik
aleksander.aristovnik@fu.uni-lj.si