Educational Technology International
2014, Vol. 15, No. 2, 71-88

Analysis of Online Behavior and Prediction of Learning Performance in Blended Learning Environments[*]

Il-Hyun JO Yeonjeong PARK[**] Jeonghyun KIM Jongwoo SONG
Ewha Womans University
South Korea

A variety of studies to predict students' performance have been conducted since educational data such as web-log files traced from Learning Management System (LMS) are increasingly used to analyze students' learning behaviors. However, it is still challenging to predict students' learning achievement in blended learning environment where online and offline learning are combined. In higher education, diverse cases of blended learning can be formed from simple use of LMS for administrative purposes to full usages of functions in LMS for online distance learning class. As a result, a generalized model to predict students' academic success does not fulfill diverse cases of blended learning. This study compares two blended learning classes with each prediction model. The first blended class which involves online discussion-based learning revealed a linear regression model, which explained 70% of the variance in total score through six variables including total log-in time, log-in frequencies, log-in regularities, visits on boards, visits on repositories, and the number of postings. However, the second case, a lecture-based class providing regular basis online lecture notes in Moodle show weaker results from the same linear regression model mainly due to non-linearity of variables. To investigate the non-linear relations between online activities and total score, RF (Random Forest) was utilized. The results indicate that there are different set of important variables for the two distinctive types of blended learning cases. Results suggest that the prediction models and data-mining technique should be based on the considerations of diverse pedagogical characteristics of blended learning classes.

Keywords : Learning Management System, Educational Data Mining, Blended Learning, Prediction, Multiple Regression, Random Forest

[*] This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5A2A0304410).

[**] ypark78@ewha.ac.kr

Introduction

The use of learning management system (LMS) has grown exponentially. It is becoming ubiquitous in current higher education. LMS offers a great variety of channels and workspaces to facilitate information sharing and communication among participants in a course, to let educators distribute information to students, produce content materials, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. (Romero, Ventura, & García, 2008). Further, the large amount of students' behavioral data left in LMS can be accumulated as web-log files, extracted as valuable information, and finally utilized to improve students' learning achievement (Jo & Kim, 2013). Compared to human observation of live or videotaped tutoring, such logs can be more extensive in the number of students, more comprehensive in the number of sessions, and more exquisite in the level of details (Mostow et al., 2005). As a result, a variety of studies using web-log data to predict students' performance have been conducted.

However, it is still challenging to predict student's learning achievement in blended learning class which is commonly defined as an integration of traditional face-to-face and online approaches to instruction (Garrison & Vaughan, 2013; Graham, Woodfield, & Harrison, 2013). In higher education, there is a considerable complexity in its implementation with the challenge of virtually limitless design possibilities and applicabilities (Garrison & Kanuka, 2004). This makes it difficult to predict student's achievement by using online learning patterns with a one-for-all prediction model. Therefore, we intended to develop multiple prediction models to predict students' academic achievements according to the pedagogical types of blended learning.

As an exploratory study, this paper started with the description of blended learning, and then examined previous achievement prediction research on online learning environment. After that, this paper developed two different prediction models for two different types of blended learnings. Finally, we summarized our research and made some conclusions.

Theoretical Backgrounds

Blended Learning

Liebowitz and Frank (2010) defined blended learning as a hybrid of traditional face-to-face and online learning instruction occurring both in classrooms and online and where the online component becomes a natural extension of traditional learning. As a new traditional instructional model in higher education (Graham et al., 2013), previous studies have indicated that such a blended learning contributes to increased interactions between students and faculty. The strengths and weaknesses of the online learning are complemented by the weaknesses and strengths of the traditional face-to-face learning and vice versa.

However, there is a complexity in its implementation with the challenge of virtually limitless design possibilities and applicabilities to so many contexts. It is not clear as to how much, or how little, online learning is inherent in blended learning (Garrison & Kanuka, 2004). Hence, diverse cases of blended learning have been formed from the simple use of LMS to the full usages.

Singh (2003) introduced dimensions of the blend. The first dimension is blending offline and online learning at the simplest level; the second dimension refers to blending self-paced and live, collaborative learning; the third one indicates blending structured and unstructured learning; the forth one is blending custom content with off-the-shelf content; and the last dimension refers to blending learning, practice, and performance support. Meanwhile, Francis and Raftery (2005) introduced the complexity and level of LMS within three e-learning modes of engagement. The first mode indicates baseline course administration and learner support; the second mode allows blended learning leading to significant enhancements to learning and teaching process; and the third one extends all of two modes to the level of personalized instruction through diverse online courses and modules.

As Voos (2003) explained, the "blendedness" does not make significant difference in the comparative outcomes. Rather it makes the fundamental re-consideration of the content in light of new instructional and media choices. There have been diverse approaches to make blended learning more dynamic. As a broader perspective, Margulieux, Bujak, McCracken, and Majerich (n.d) distinguished several terms such as hybrid, blended, flipped, and inverted into a framework based on two dimensions including 1)information transmission vs. praxis, and 2) delivery via instructor vs. delivery via technology. The learning experience taxonomy that they adopted suggested major four types of learning F2F mixed (e.g., course with lab), Lecture hybrid (e.g., part f2f, part online lecture), Practice hybrid (e.g., part f2f, part online praxis), and online mixed (e.g., MOOC).

Recently, a flipped classroom gained an attention as a new pedagogical method, which employs asynchronous video lectures and problem-solving practices as students' homework, and active and diverse group-based activities in the classroom. It represents not only the combination of instructional methods in online and offline but also combination of learning theories such as problem-based approaches based on constructive ideology vs. traditional lectures derived from direct instruction method which is founded upon behaviorist principles (Bishop & Verleger, 2013).

The previous studies introduced above have classified blended learning in diverse frameworks, or throughout *theory-driven approach*. However, it is also expected to classify them throughout the *data-driven approach* by analyzing the usage patterns of online learning activities in LMS and several major attributes of classes such as type of contents, affiliation of students, and the size of classes. Kim, Yoo, Park, and Jo (2014) analyzed 4,416 classes which used virtual classrooms during one semester in a private university located in Seoul, Korea, by using big data tracked from learning management system. While the descriptive statistics presented relatively low usages of diverse functions in comparing to what virtual classroom provides, the top five most frequently employed activities were 1) resources, 2) notice, 3) questions & answers, 4) lecture notes, and 5) assignment submission. They also highlighted that very few classes used online activities such as asynchronous discussion very actively and large portion of classes utilized the function of virtual classrooms as a supplementary tool for traditional classroom learning.

Based on previous studies which classify diverse blended learning theoretically and empirically, this study selected two typical types of blended learning. One is a formal classroom course but incorporated online discussion as its major learning activity, the other is a typical lecture-based class but utilized virtual classroom as a supplementary tool for downloading learning resources and submitting assignments. The purpose of this study is to analyze online behavior patterns and develop prediction models customized to the distinctive patterns. Throughout this study, it is further expected that diverse prediction models are developed based on different blended learning models classified by pedagogical frameworks as well as verified by a data-driven approach.

Predication of Students' Performance

LMS data is considered as a useful representation of students' learning behaviors that are difficult or impossible to apprehend. A variety of studies, using web-log files, to analyze learners' online behaviors and predict their future achievements has been conducted and several applications were applied in practice in order to solve the educational issues and problems.

Arnold & Pistilli (2012) reported the impact of Course Signals(CS) at Purdue University. CS utilizing large data sets is a student success system that allows faculty to provide meaningful feedback to students based on predictive models. Their research indicated that the use of learning analytics through the application of Course Signals has shown great promise with regard to the success of students, as well as their overall retention to the University (Arnold & Pistilli, 2012).

Lykourentzou, Giannoukos, Nikolopoulos, Mpardis & Loumos (2009) conducted a research on dropout prediction method to accurately identify dropout-prone students in early stages of the e-learning course. They drew on the detailed student weblogs, extracted from the LMS, and used three machine learning techniques, namely feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP. Experimental results indicated that combining the outputs of the three machine learning techniques enables a more accurate and prompt identification of dropout students. Macfadyen & Dawson (2010) extracted and analyzed relevant student tracking variables from LMS and developed a regression model of student success. Beck and Woolf (2000) construct a learning agent for high-level student modeling with machine learning in intelligent tutoring systems.

Despite the abundant amount of research analyzing a massive amount of data and controlling student's academic achievement, introduced above, the pedagogical types of classes in their diverse research context have not been seriously dealt with. For example, diverse dimensions of blended learning or e-learning modes of engagement have not been considered in their prediction models. In spite of applying the highly complicated and advanced data-mining technique, it is found that a single algorithm with the best classification and accuracy in all cases are not possible (Romero, Espejo, Zafra, Romero, & Ventura, 2013). Including not only online learning activity of students but also offline information such as classroom attendance, punctuality, participation, attention and predisposition were suggested to increase the prediction power (Romero et al., 2013). However, this approach requires collecting the data which depends on the instructors' extra effort to provide such information (Romero et al., 2013).

To overcome this issue and to still utilize useful LMS log-files for the educational purpose, some of previous studies have attempted to use online log variables to predict learning outcome. Jo and his colleagues explored the variables such as students' total log-in time, log-in frequency, log-in regularity, visits on board, time spent on board, and visits on repository were meaningful predictors of students learning outcome. Table 1 shows a series of studies that were in different contexts (corporate environment and higher education) and different learning models (online 100% and blended learning), where diverse online behavior variables were inserted to predict final learning achievement. The extents of prediction, indicated by adjusted R-squared values, were all diverse but the variable such as log-in regularity was consistently significant. Therefore this study continued investigating such online behavior variables but further exploring them in higher education and blended learning context. In two different types of blended learning, we attempted to confirm the possibility to predict student's achievement by analyzing such online log variables.

Table 1. Previous studies using online log variables to predict learning outcome

	Studies	Contexts	Learning Type	N	Inserted Variables for Prediction model	R^2
1	Jo and Y. Kim (2013)	Corporate	Online	632	total log-in time, log-in frequency, log-in regularity	.09
2	Jo and J. Kim (2013)	Higher Ed	Online	23	total log-in time, log-in frequency, log-in regularity	.26
3	Yu and Jo (2014)	Higher Ed	Blended (F2F lecture and online discussion)	84	total log-in time, log-in frequency, log-in regularity, download of materials, peer interaction, interaction between students and instructor	.34
4	Jo, Kim, and Yoon (2014)	Corporate	Online	200	total log-in time, log-in frequency, log-in regularity	.21

Method

Research Context

For this study, we collected and analyzed the web log data of 43 college students of 'Class A', entitled "Administration and Politics", and 29 college students of 'Class B', entitled "Corporate Education", opened in the regular fall semester in a large higher educational institution in 2013. Although they were the courses for different majors and department, both are social science in nature. These were considered as blended learning classes since major and regular classroom meetings were held during fifteen weeks but the classes utilized Moodle-based LMS for teaching and learning purposes. While the major online activity in 'Class A' was discussion forum, the second class involved a supplemental tool for submitting assignments and downloading learning materials. In this study, we call 'Class A' an 'online discussion-based learning' and 'Class B' an 'offline lecture-based learning course. The final total score of each course was calculated with the suggested weight in Table 2.

Table 2. Items of grading in both classes

Items	Class A	Class B
Offline attendance	5%	-
Group presentation	-	10%
Individual tasks	20%	-
Team tasks	-	30%
Mid-term exam	30%	30%
Final exam	30%	30%
Online discussion	15%	-
Total	100%	100%

Data Collection

In both cases, the data source (web-log data) was the Moodle database, and the independent variables for this study were computed by automatic data collection module embedded in the LMS. First, the total log-in time in LMS was computed by calculating the total amount of time spending between log-in and logout. Second, the total log-in frequency in LMS was calculated by adding up the number of each student's log-in time into the LMS. Third, regularity of learning interval in LMS was computed by calculating the standard deviation of visit intervals into the LMS. Fourth, visit on repository and board was calculated by counting the total number of each student's visit time on repository and board, respectively. Last, the number of postings was calculated by counting postings and replies uploaded by each learner, but that variable was used only for 'Class A', because there was no 'number of postings' variable for B class (an offline lecture-based learning course). And this study used the Total Score (TS) as a dependent variable for each course. Because TS involved most topics covered in each course, it was considered as the proper factor to examine the student's achievement generally.

Results

Descriptive statistics

Boxplots were used to provide an overview of the descriptive characteristics of the variables we utilized in both classes. As shown in Figure 1, Class A showed higher mean values than class B in Total Log-in Time (TLT), Total Log-in Frequency (TLF) and Visit on Board (VOB). In those variables, the variance of Class A was also greater than that of class B. But Visit on Repository (VOR) values were generally higher in Class B. Also, Log-in Regularity (LIR) values, calculated by the standard deviation of visit intervals into the LMS, were lower overall in Class A. In other words, students in Class A progressed their learning more regularly than students in Class B. The mean of the total score was 43.38 (SD=11.42) in Class A and 79.84 (SD=9.78) in Class B, so it was found that the total scores of Class B became upward leveling when it compared to Class A.

Multiple regression analysis

Case 1: Discussion-Based Learning

In Class A, a blended learning which involves online discussion-based learning, linear multiple regression analysis was conducted, in order to develop a predictive model. As shown in Table 3, this process generated a 'predictive model' of the student's total score (F=12.551, *p*=.000). The multiple adjusted squared correlation coefficient for this model is .646, but only two variables, log-in regularity and the number of postings in online forum, were statistically significant contributors (*p* < .05).

Table 3. The result of multiple linear regression analysis in Class A

Model	Unstandardized		Standardized	t	Sig.
	B	Std. Error	Beta		
(constant)	49.107	5.888		8.34	0
Total Log-in Time	-0.03	0.056	-0.074	-0.529	0.6
Total Log-in Frequency	-0.026	0.03	-0.145	-0.881	0.385
Log-in Regularity	-0.231	0.066	-0.504	-3.511	0.001
Visits on Board	0.009	0.008	0.287	1.214	0.234
Visits on Repository	-0.029	0.09	-0.04	-0.324	0.748
Number of postings	0.15	0.07	0.389	2.156	0.039

a. N=43

b. Dependent Variable: Total Score

c. R^2(adj.R^2)=.702(.646), F=12.551, *p*=.000

Case 2: Lecture-Based Learning

In Class B, a blended learning which involves offline lecture-based learning and online supplemental tool, we tried to find a model with a linear multiple regression analysis. However, as shown in Table 4, only the total log-in frequency was significant, and the R^2 value (0.274) was relatively small. Moreover, the overall model test (F-test) was not significant. Therefore, it was not appropriate to use a linear model to analyze this data.

Table 4. The result of multiple linear regression analysis in Class B

Model	Unstandardized		Standardized	t	Sig.
	B	Std. Error	Beta		
(constant)	67.064	12.985		5.165	.000
Total Log-in Time	-.100	.192	-.182	-.522	.607
Total Log-in Frequency	.151	.065	.791	2.334	.029
Log-in Regularity	.143	.142	.365	1.006	.325
Visits on Board	-.042	.045	-.318	-.933	.360
Visits on Repository	.002	.083	.006	.030	.976

a. N=29

b. Dependent Variable: Total Score

c. $R^2(adj.R^2)$= .274(.116), F=1.735, *P=.167*

Random Forest Analysis

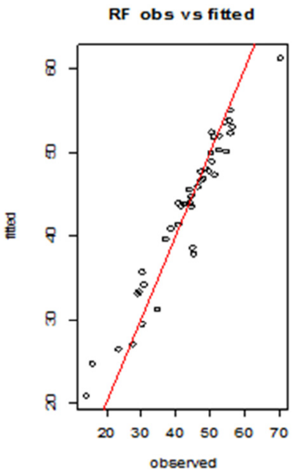
We attempted to confirm whether there is a difference in variables associated with academic performance for each course through utilizing Random Forest (RF) analysis, which is non-linear variation of decision tree method (Breiman, 2001). Decision trees utilized in RF are highly nonlinear models, so it cannot be guaranteed that the important variables in RF are equal to the ones in linear model. Nevertheless, RF is a recommended approach in verifying important variables (Bayazit, Askar, & Cosgun, 2014), so we tried it and the two courses were compared in terms of the importance of variables.

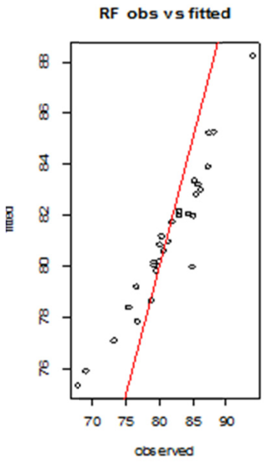
While the discussion-based learning indicated the important variables as visits on board, the total log-in time, and the number of posting in forum (Pseudo R^2= 0.91), the lecture-based learning indicated log-in regularity, the total log-in frequency, visits on board, and the total log-in time (Pseudo R^2=0.70). Here Pseudo R^2 is defined as follows.

- Pseudo R^2 = 1 – RSS/SST
- RSS = Residual sum of squares
- SST = Sum of squares of total

Table 5. Comparison of important variables in different blended learning cases

Important Variable	Case 1 (Discussion-Based BL)	Case2 (Lecture-Based BL)
	N=43, Pseudo R^2= 0.91	N=29, Pseudo R^2=0.70
1	Visits on Board	Log-in Regularity
2	Total log-in time	Total log-in frequency
3	Number of Posting in forum	Visits on Board
4	Log-in Regularity	Total log-in time





Conclusion

The large proportion of residuals occurred by off-line behaviors that are not explained by student's online activity makes it difficult to predict model for blended learning classes. Furthermore, there is a variety of blended learning classes in universities and they are assumed to show different prediction models with a wide range of R^2 value. In this study, we presented that two different types of blended learning class show different models: linear and non-linear.

In case of the discussion-based blended learning course, which involves active learner's participations in online forum, a linear multiple regression analysis model explains the student's achievement. The prediction model showed the increased adjusted R^2 value (.646), in comparing with previous study in higher education (Yu & Jo, 2014) where R^2 value was .335. Also, not only log-in regularity but also the number of postings turned out to be statistically significant.

However, in case of the lecture-based blended learning course, which involves submitting tasks or downloading materials as main online activities, linear multiple regression analysis model was not appropriate for prediction, which was evidenced by relatively small value of R^2 and insignificance of most variables and overall model test. The different results from two cases suggest that prediction models should be based on the considerations of diverse blended learning types, more specifically what extent and what kinds of online activities were involved in the class.

Additionally, in using a Random Forest approach, we found that the two blended learning cases indicated different important variables which reflect the attributes of discussion-based learning class and lecture-based learning class, respectively. This result suggests that a future study needs to be conducted by clustering the types of blended learning classes throughout the students' online learning behavior data and predicting their learning achievement according to the clustered models. We conclude that the prediction models and data-mining technique should be based on the considerations of diverse pedagogical characteristics in blended learning.

References

Arnold, K. E., & Pistilli, M. D. (2012). *Course signals at Purdue: Using learning analytics to increase student success*. Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.

Bayazit, A., Askar, P., & Cosgun, E. (2014). Predicting learner answers correctness through eye movements with random forest. In A. Pena-Ayala (Ed.), *Educational data mining: Applications and trends*. Switzerland: Springer.

Beck, J. E., & Woolf, B. P. (2000). *High-level student modeling with machine learning*. Paper presented at the Intelligent tutoring systems.

Bishop, J. L., & Verleger, M. A. (2013). *The flipped classroom: A survey of the research*. Paper presented at the ASEE National Conference Proceedings, Atlanta, GA.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Francis, R., & Raftery, J. (2005). Blended learning landscapes. *Brookes eJournal of learning and teaching, 1*(3), 1-5.

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The internet and higher education, 7*(2), 95-105.

Garrison, D. R., & Vaughan, N. D. (2013). Institutional change and leadership associated with blended learning innovation: Two case studies. *The internet and higher education, 18*, 24-28.

Graham, C. R., Woodfield, W., & Harrison, J. B. (2013). A framework for institutional adoption and implementation of blended learning in higher education. *The internet and higher education, 18*, 4-14.

Jo, I., Kim, D., & Yoon, M. (2014). *Analyzing the log patterns of adult learners in LMS using learning analytics* Paper presented at the The 4th International Conferecne on Learning Analytics Knowledge, Indianapolis, Indiana, U.S.A.

Jo, I., & Kim, J. (2013). Investigation of Statistically Significant Period for Achievement Prediction Model in e-Learning. *Journal of Educational Technology, 29*(2), 285-306.

Jo, I., & Kim, Y. (2013). Impact of learner's time management strategies on achievement in an e-learning environment: A learning analytics approach. *The Journal of Educational Information and Media, 19*(1), 83-107.

Kim, Y., Yoo, Y., Park, Y., & Jo, I. (2014). *Big data analytics in higher education: Usage status of virutal classroom and activities in LMS*. Paper presented at the e-learning Korea 2014, Seoul, COEX.

Liebowitz, J., & Frank, M. (2010). *Knowledge Management and E-learning*: CRC Press.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers* & *Education, 53*(3), 950-965.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers* & *Education, 54*(2), 588-599.

Margulieux, L. E., Bujak, K. R., McCracken, W. M., & Majerich, D. (n.d). Hybrid, Blended, Flipped, and Inverted: Defining Terms in a Two Dimensional Taxonomy.

Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., & Heiner, C. (2005). *An educational data mining tool to browse tutor-student interactions: Time will tell.* Paper presented at the Proceedings of the Workshop on Educational Data Mining, National Conference on Artificial Intelligence.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education, 21*(1), 135-146.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers* & *Education, 51*(1), 368-384.

Singh, H. (2003). Building effective blended learning programs. *Educational Technology, 43*(6), 51-54.

Yu, T., & Jo, I. (2014). *Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education.* Paper presented at the The 4th International Conferecne on Learning Analytics Knowledge, Indianapolis, Indiana, U.S.A.

Il-Hyun JO

Professor, Dept. of Educational Technology, College of Education, Ewha Womans University.

Interests: Learning Analytics, Social Network Analysis, Knowledge Management, Human Resource Development, Mobile-based Informal Learning in Workplace

E-mail: ijo@ewha.ac.kr

Yeonjeong PARK

Research Professor, Dept. of Educational Technology, College of Education, Ewha Womans University.

Interests: Mobile and Smart Learning, Socio-cultural Aspects of Learning, Human Resource Development, Program Evaluation

E-mail: ypark@ewha.ac.kr

Jeonghyun KIM

Ph.D. Candidate, Dept. of Educational Technology, College of Education, Ewha Womans University.

Interests: Learning Analytics, Social Network Analysis, Human Resource Development, Mobile and Smart Learning

E-mail: naralight@naver.com

Jongwoo SONG

Associate Professor, Department of Statistics, Ewha Womans University

Interests: Regression, Classification, Extreme value theory, Datamining, Computational Statistics

E-mail: josong@ewha.ac.kr

Received: September 18, 2014 / Peer review completed: October 06, 2014 / Accepted: October 14, 2014