Predicting Students' Learning Performance by Using Online Behavior Patterns in Blended Learning Environments: Comparison of Two Cases on Linear and Non-linear Model

Jeong Hyun Kim
Ewha Womans University
College of Education Bldg. #533, Daehyun-dong,, Seodaemun-gu, Seoul, Korea, 120-750
+82-2-3277-3201
naralight@naver.com

Yeonjeong Park
Ewha Womans University
College of Education Bldg. #533, 52, Daehyun-dong, Seodaemun-gu, Seoul, Korea, 120-750
+82-2-3277-3201
ypark78@ewha.ac.kr

Jongwoo Song
Ewha Womans University
Science Bldg. B #561, Daehyun-dong, Seodaemun-gu, Seoul, Korea, 120-750
+82-2-3277-2299
josong@ewha.ac.kr

Il-Hyun Jo
Ewha Womans University
College of Education Bldg. #533, Daehyun-dong,, Seodaemun-gu, Seoul, Korea, 120-750
+82-2-3277-3201
ijo@ewha.ac.kr

ABSTRACT

A variety of studies using educational data traced from LMS has been conducted to predict students' performance. However, because of the complexity in its implementation, it is still challenging to predict students' learning achievement in blended learning environment. As an exploratory study, we selected two types of blended learning classes and compared their prediction models. While the first blended learning class which involves online discussion-based learning revealed a linear regression model, the second case, which was a lecture based blended learning class providing regular base online lecture notes in Moodle, did not present a linear regression model. After that, to examine the important variables of each class, RF (random forest) method was utilized. The results indicated different important variables in two cases. We concluded that the prediction models and data-mining technique should be based on the considerations of diverse pedagogical characteristics in blended learning.

Keywords

Educational Data Mining, Blended Learning, Prediction, Multiple Regression, Random Forest

1. INTRODUCTION

The use of learning management system (LMS) has grown exponentially. LMS offers a great variety of channels and workspaces to facilitate information sharing and communication among participants, 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. [1]. 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. As a result, a variety of studies using web-log data to predict students' performance has been conducted.

However, despite the abundant amount of research analyzing a massive amount of data and controlling student's academic achievement, 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 [2, 3]. 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 [4]. In higher education, there is considerable complexity in its implementation with the challenge of virtually limitless design possibilities and applicability [5]. This makes it difficult to predict student's achievement by using online learning patterns with a one-for-all prediction model.

Although there is no single framework for blended learning, it is generally assumed there are several types of blending in practice as the previous studies have attempted [6, 7]. 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, we implemented a different approach for two different types of blended learning, and tried to confirm the possibility to predict student's achievement in blended learning environments.

2. METHOD

2.1 Research Context

We analyzed the web log data of 43 college students of 'class A' and 30 college students of 'class B' opened in the regular fall semester in a large higher educational institution in 2013. 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.

2.2 Data Collection

In both cases, the data source (web-log data) was tracked from the Moodle. The independent variables for this study were computed by automatic data collection module embedded in the LMS. Total log-in time, log-in frequencies, regularities of log-in interval, visits on boards, visits on repositories were used as independent variables for both courses. Because there was no 'number of postings' variable for B class, the number of postings was used only for 'class A', This work used the Total Score as a dependent variable for each course.

2.3 Data Analysis

The procedure of data analysis consists of two phases. In phase 1, we conducted multiple linear regression analysis for both class A

and B. While the class A showed a linear regression model, for the class B the linear model was neither proper nor statistically significant to predict student's achievement. Hence, as the second phase, we implemented Random Forest (RF) algorithm to increase the prediction accuracy.

3. RESULTS

3.1 Case 1: Discussion-based Learning

In class A, a blended learning which involves online discussion-based learning, linear multiple regression analysis was conducted, and this process generated a 'predictive model' of the student's final score (R^2 =.646, F=12.551, p =.000). Only two variables, log-in regularity and the number of postings in online forum, were statistically significant contributors.

3.2 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, only the total log-in frequency was significant, and F-test was insignificant with pretty low R^2 value (R^2 =.116, F=1.735, p =.167).

3.3 Random Forest Analysis

RF (random forest) method was tired in both cases to find important variables. As shown in Table 1, 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), but 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 was defined as follows.

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

4. CONCLUSION

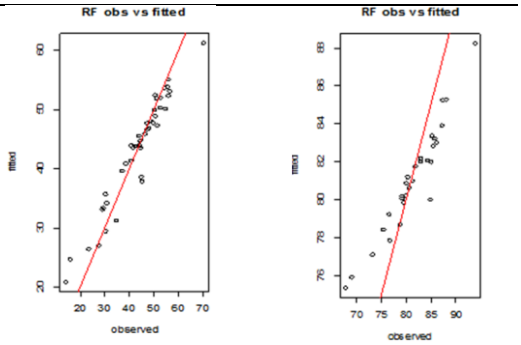
There is a variety of blended learning classes in universities and they are assumed to show different prediction models with a wide range of R^2 value. In this study, we presented that two different types of blended learning class show different models: linear and non-linear.

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. But 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 proper for prediction.

Additionally, in using a Random Forest approach, we found that two 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.

Table 1. Comparison of important variables in two 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



5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, pp. 368-384, 2008.

[2] D. R. Garrison and N. D. Vaughan, "Institutional change and leadership associated with blended learning innovation: Two case studies," The Internet and Higher Education, vol. 18, pp. 24-28, 2013.

[3] C. R. Graham, W. Woodfield, and J. B. Harrison, "A framework for institutional adoption and implementation of blended learning in higher education," The Internet and Higher Education, vol. 18, pp. 4-14, 2013.

[4] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use Moodle courses," Computer Applications in Engineering Education, vol. 21, pp. 135-146, 2013.

[5] D. R. Garrison and H. Kanuka, "Blended learning: Uncovering its transformative potential in higher education," The internet and higher education, vol. 7, pp. 95-105, 2004.

[6] H. Singh, "Building effective blended learning programs," Educational Technology, vol. 43, pp. 51-54, 2003.

[7] R. Francis and J. Raftery, "Blended learning landscapes," Brookes eJournal of learning and teaching, vol. 1, pp. 1-5, 2005.