2017 6th IIAI International Congress on Advanced Applied Informatics

Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data

Nobuhiko Kondo
University Education Center
Tokyo Metropolitan University
Tokyo, Japan
kondo@tmu.ac.jp

Midori Okubo
School of Engineering
Osaka University
Osaka, Japan
midori.okubo@ist.osaka-u.ac.jp

Toshiharu Hatanaka
Department of Information Science
and Technology
Osaka University
Osaka, Japan
hatanaka@ist.osaka-u.ac.jp

***Abstract*— Analytics in education has been received much attention over the past decade. It is necessary to maintain high retention rate in any institutions of higher education, therefore several attempts on the application of analytics have been done for this problem. To detect students at high drop-out risk early and intervene them effectively, utilizing the educational big data can be useful. In this paper, an automatic detection method of academically at-risk students by using log data of learning management systems is considered. Some well-known machine learning methods are used to build a predictive model of student performance evaluated by GPA. By using actual data set, we investigate an availability of the proposed method and discuss its ability to early detection of off-task behavior. The experimental results indicated that some characteristics of behavior about learning which affect the learning outcomes can be detected with only the online log data. Furthermore, comparative importance of explanatory variables obtained by the approach would help to estimate which variable affects comparatively to the learning outcome and it can be used in institutional research.**

Keywords—learning analytics; enrollment management; institutional research; detection of at-risk students; machine learning; LMS log data

I. Introduction

Over the last decade, the word of “big data” has emerged and it has been used widely in many areas, such as marketing research, health care services, business, social media analysis, medicine development, and so on. Using “big data”, making an adequate decision, finding useful patterns or rule set, fault or change detection and chance discovery are tried with some statistical techniques.

Analytics of big data in education has been studied actively from approximately 2000. Particularly, learning analytics has been one of the major field of the research on analytics in education. Learning analytics was defined as follows [1]: *Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.*

A concept of academic analytics has been also proposed in late 2000s from the viewpoint of an application of analytics in the level of institution or politics [2]. In academic analytics, quality assurance of education or accountability of university are rather focused than the learning or teaching perspective, thus learning outcomes or retention of students are targets of analytics. Academic analytics has a close relationship with the viewpoint of enrollment management in institutional research (IR). For educational institutions, especially colleges or universities, it is necessary that high retention rate is maintained, therefore enrollment management is important.

In recent studies, for example, early detection of at-risk students with learning analytics has been considered in this context [3]. In order to detect students who have high drop-out risk early and intervene them effectively, utilizing the educational big data can be useful. Several studies also investigated the correlation between learning outcome and usage of learning management system (LMS) [4], and the log data of LMS have turned to be useful to analyze students’ learning behavior.

In this study, an approach to detection of academically at-risk students by using machine learning methods based on log data of learning management system is proposed. Then results of some numerical experiments with actual data implemented to investigate the performance of the approach will be shown.

II. Experimental Data

In this study, students’ data of a university, called “university X” in this paper, are used. The university X is a private liberal arts university in Japan. The used data are records of 202 students admitted to the department Y of the university X in 2015.

In the university X, a LMS is used on the whole university. Students should use the LMS to manage their own learning in several classes, to check some information from the university, to use an e-portfolio system, to learn with some self-learning contents, and so on. They are expected to use the LMS enough throughout their school life. The LMS records a logfile whenever any students operate it. One record contains the student ID, the operating date and the type of operation. Although the LMS is not very often used in some classes, a level of usage is expected to reflect a level of commitment to

978-1-5386-0621-6/17 $31.00 © 2017 IEEE
DOI 10.1109/IIAI-AAI.2017.51

learning in the university, because the students need to use it to spend their school life smoothly.

In the experiment, all logfiles of a period between April 1st and August 5th in 2015 were used. The number of record was 200,979. Each record contains the type of operation. For example, the operation types are to log-in, or log-out, to boot the player of e-learning, to start or end the lesson, to submit an assignment, and so on. In this study, some features were extracted from the LMS log data. These features and the rate of attendance of offline compulsory classes in the first semester were used as the explanatory variables, and then the problem of predicting a GPA of the first semester was considered.

The explanatory variables used in this study are shown in Table I. "(1) GPA" is a binary variable. Let μ be the mean and let σ be the standard deviation of GPA of all students, "(1) GPA" of a certain student is to be 1 when the GPA of him/her is greater than $\mu - \sigma$, or it is to be 0 otherwise. "(1) GPA" = 0 means the student has been off-task and he/she is academically at-risk. Seven types of explanatory variables were considered as shown in Table I. "(2) Attendance rate" is the rate of attendance of offline compulsory classes in the first semester. (3)-(8) are the variables extracted from the LMS log data. These variables express several perspectives of student's action on the LMS. "(3) # of booting the player" refers to the total number of booting the player of any e-learning contents. "(4) Night Act" is the total number of operation during the night (0 a.m. to 5 a.m.). "(5) # of logging-in" is the total number of logging in the LMS. "(6) # of starting a lesson" is the total number of booting a function for an output of learning, and "(7) # of submission completion" refers to complete such an output activity. "(8) Duration of logging-in time" is the estimated value of the total duration of logging in the LMS.

III. Numerical Experiments

A. Predictive modeling with machine learning

Machine learning is the approach to give computers the ability to learn automatically like human beings. The machine learning methods have some sort of algorithms that discover patterns or rules from actual data, and the model learned appropriately can predict unseen data properly. The methods are often used in several fields such as pattern recognition, medical diagnostics, search engine, robotics, and so on. In the fields of analytics in education, the machine learning methods are frequently used for the construction of predictive model of learners. In this study, we used well-known machine learning methods [5], which are logistic regression, support vector machine and random forest, to predict the GPA by the explanatory variables shown in Table I.

Logistic regression is a kind of generalized linear model and used often as two class classifier. Due to its easiness to handle and applicability, it has been used in several fields. Support vector machine (SVM) is a kernel machine widely used for pattern classification and regression problem. As it is said that SVM has high generalization ability, it has been used widely as well as the logistic regression. Random forest is one of the ensemble learning model. The random forest model contains some simple decision trees as weak learners and output the value as the average or majority vote of outputs of the decision trees. It is known that the random forest model has some advantage such as robustness against the noise, quickness of learning, easiness of setting the hyper parameter, and so on.

TABLE I. Variables Used in This Study

Type	Variables	Data source
Response variables	(1) GPA	Grade data
Explanatory variables	(2) Attendance rate	Attendance data
	(3) # of booting the player	LMS log data
	(4) Night Act	
	(5) # of logging-in	
	(6) # of starting a lesson	
	(7) # of submission completion	
	(8) Duration of logging-in time	

The numerical experiments were implemented by Python 3.6.0 and scikit-learn package were used for the construction of the predictive models with machine learning methods.

B. Early detection of at-risk students

From the purpose of this study, an at-risk student detecting method should have an ability to detect such students as early as possible. Therefore, we performed an experiment to investigate how the detection ability changes for each week in the first semester. The period of classes per a semester of the target university is 15 weeks. In this experiment, the classification metrics for each week were calculated with all data obtained by each week. We used common metrics such as precision, recall, and F-measure. 10-fold cross validation was implemented for 10 times and each metric was averaged.

A weekly change of the classification metric values for three machine learning methods are shown in Fig. 1 to Fig. 6. Week "0" is the week when a series of orientation for freshmen is implemented. Fig. 1 to Fig. 3 show the case of the "(2) Attendance rate" included, and Fig. 4 to Fig. 6 show the results without the "(2) Attendance rate".

In latter part of the period, the detection ability of the logistic regression model was relatively higher than the other models. On the other hand, the random forest model can detect more at-risk students at the early stage, especially until the third week. Moreover, the random forest model seems to have the most stable behavior and good valance of precision and recall. The SVM model had high precision values, but had low recall values comparatively.

Although the detection performances in the case of not using "(2) Attendance rate" were lower than in the case of using it, the random forest model relatively performed well with only the LMS log data. As shown in Fig. 6, the recall at week 0 was about 0.3 and the recall at week 3 was about 0.4. Therefore, the random forest model can detect about 30 % of at-risk students until the first week, and can detect about 40 % of at-risk students until the fourth week.

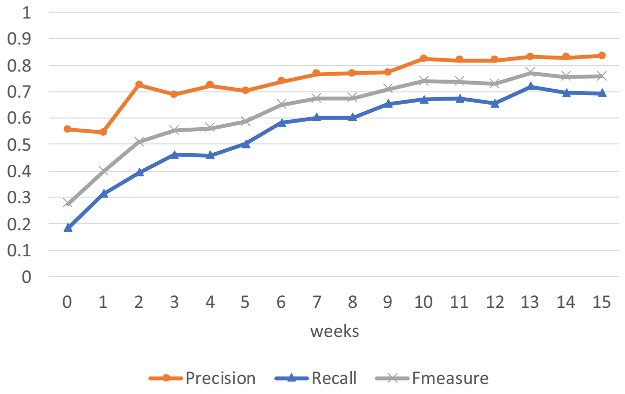

Fig. 1. Classification metrics for logistic regression with attendance data.

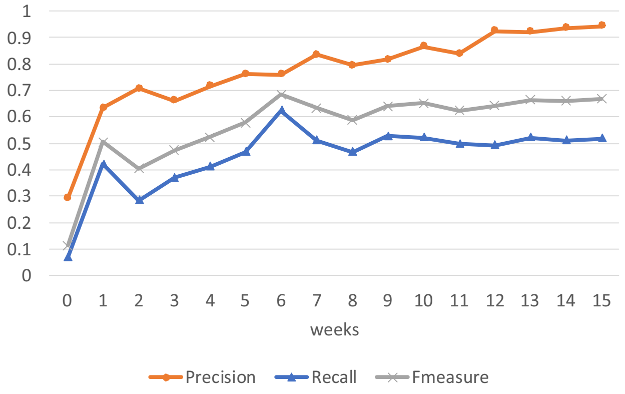

Fig. 2. Clssification metrics for SVM with attendance data..

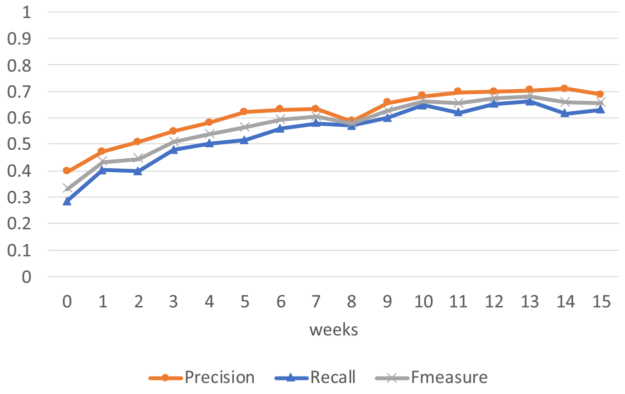

Fig. 3. Clssification metrics for random forest with attendance data..

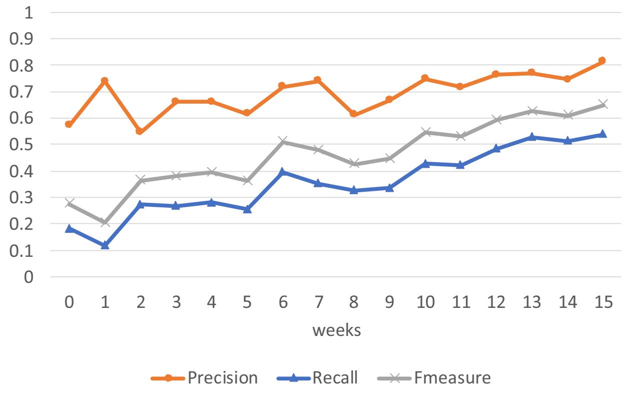

Fig. 4. Classification metrics for logistic regression without attendance data.

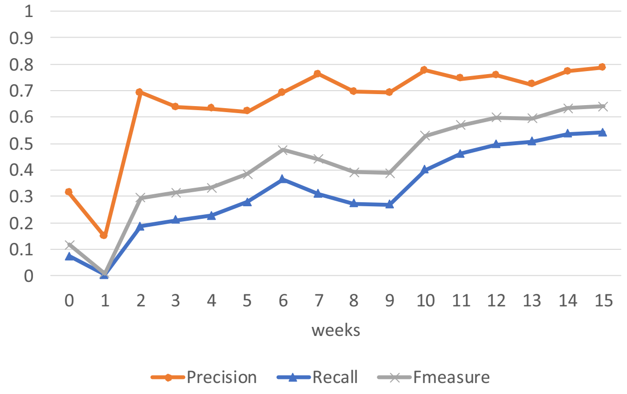

Fig. 5. Classification metrics for SVM without attendance data.

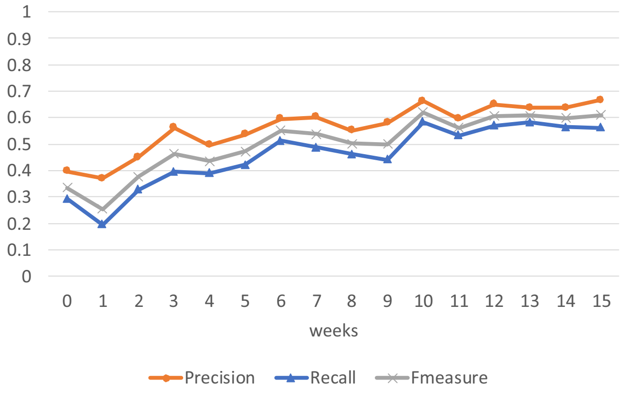

Fig. 6. Classification metrics for random forest without attendance data.

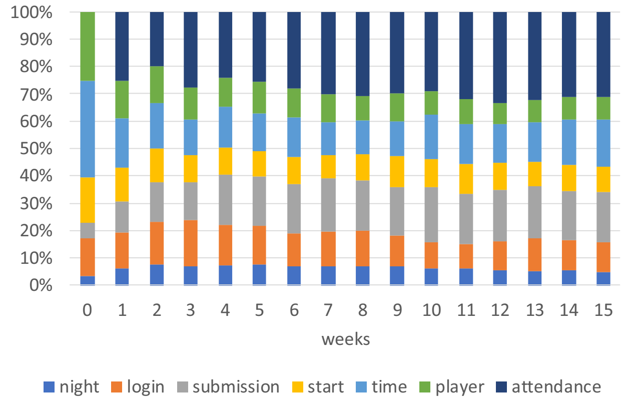

Fig. 7. Weekly change of the comparative importance of explanatory variables with the attendance data.

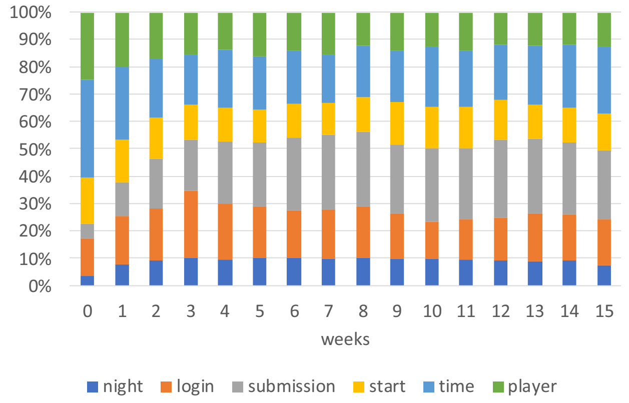

Fig. 8. Weekly change of the comparative importance of explanatory variables without the attendance data.

C. *Weekly change of comparative importance of variables*

It would be preferable to investigate which variable affect the classification ability strongly. As the random forest model can calculate comparative importance of variables based on Gini index, we investigated weekly change of the importance of variables as an approach to deal with such a problem.

The comparative importance of variables for each week are shown in Fig. 7 and Fig.8. These figures are corresponding to Fig. 3 and Fig. 6 respectively. Though a ratio of importance changed in any given week, the "(2) Attendance rate" variable was the most important for every weeks excluding the orientation week in the case of using the "(2) Attendance rate". In another case of not using the "(2) Attendance rate", it seems that some timings of relatively meaningful change corresponded to the timings of change of the classification metric values. For example, during weeks 0 to 3, the number of logging-in was becoming more important, and the classification metric values were also becoming higher. At week 10, the classification metric values sharply became higher, and the importance of the logging time also became higher and the duration of logging in became lower. This phenomenon indicates that the important activities can be inferred by carefully watching the weekly change of the comparative importance of variables, and it helps us assess the curriculum and student support strategy from the perspective of institutional research.

IV. Conclusion

In this study, we considered automatic detecting method for at-risk students. We examined the typical machine learning techniques to such students based on the actual log data of LMS and investigated their performance.

The random forest model showed the most stable behavior and good valance of precision and recall. The model can detect about 40 % of at-risk students at the end of third week of first semester with only the LMS log data. As the approach can detect a sign of off-task behavior of students with only the log data which are to be stored automatically to the LMS, a certain level of applicability of the approach is shown. It is indicated that some characteristics of behavior about learning which affect the learning outcomes can be detected with only the online log data.

Furthermore, ranking of comparative importance of explanatory variables obtained by the approach would help to estimate which variable affects comparatively to the learning outcome at any given point of time. By watching the importance of variable constantly, it is expected that an intervention strategy will be more adaptively and planning of classes, curriculum and student support can be considered based on the information.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16K16331 and JP16H03082.

References

[1] The 1st International Conference on Learning Analytics and Knowledge, Call For Papers, July 22, 2010. [Online]. Available: https://tekri.athabascau.ca/analytics/call-papers. [Accessed: Apr.7, 2017].

[2] J. P. Campbell and D. G. Oblinger, "Academic Analytics," EDUCAUSE Review, 2007.

[3] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron, "Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative," *Journal of Learning Analytics*, vol. 1, no. 1, pp.6–47, 2014.

[4] M. Andergassen, F. Mödritscher, and G. Neumann, "Practice and Repetition during Exam Preparation in Blended Learning Courses: Correlations with Learning Results," *Journal of Learning Analytics*, vol. 1, no. 1, pp.48–74, 2014.

[5] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.