Evaluating Course Grading Fairness in Comparison of Learning Activity Logs Before and After COVID-19

Hiroyuki Kuromiya[1] Majumdar Rwitajit[1] and Hiroaki Ogata[1]

[1] Kyoto University, Yoshida Nihonmatsu-cho, Kyoto, Japan
khiroyuki1993@gmail.com

Abstract. It is important to maintain course grading fairness over different years. For example, it is problematic if the percentage of students who receive the "A" grades suddenly increased because the instructors' grading criteria became easier, despite the same lectures. However, it is inevitable that the percentage of students who receive the "A" grades will fluctuate from year to year in courses that use achievement assessment. In such a situation, how can we guarantee that the variation is not due to a blurring of the grader's criteria? This study proposes an analysis procedure to ensure fairness in grading using a students' interaction logs and a prediction model. We used a dataset from a Japanese university collected from a learning analytics platform and examined whether the instructors' criteria became easier or not by setting a situation in which two different populations are evaluated by the same criteria. Our approach of using predictive models to examine the validity of human ratings is applicable for other datasets and can be used widely for evaluating grading fairness across different years and courses.

1 Context and Dataset

In this study, we compare and analyze the distribution of final course grades and digital material browsing behavior of learners taking the same courses offered in 2019 and 2020, respectively; due to the COVID-19 epidemic, data for 2019 are from face-to-face classes and data for 2020 are from online classes. A large difference was observed in the percentage of students who received the "A" grade between the face-to-face (16%) and online (40%) classes in this course, and this study examined the reasons for this difference by two hypotheses: (1) the instructors' grading criteria got more lenient, and (2) students were more likely to achieve higher levels of academic achievement. We aim to determine which of these two hypotheses is more plausible from an analysis of log data.

The log data were retrieved from a learning analytics platform [1] used in the target university in Japan. We focused on one course before and after COVID-19. The course in 2019 was conducted face-to-face over 8 weeks with 8 lecture materials. The course in 2020 was conducted online over 7 weeks with 9 lecture materials. We used the students' ebook interaction log during the two semesters. There were 352,467 logs

from 163 users in the face-to-face course (2019) and 197,593 logs from 90 users in the online course (2020) for each.

2   Methods and Results

In order to compare the two different data sets on the same grading criteria, we proposed an approach using a prediction model: training the model on the 2019 dataset and applying it to the 2020 dataset. In this way, it would be possible to evaluate 2020 students using the same criteria as the 2019 students. If the proportion of students who receive A grades does not change, then the hypothesis (1) is supported and vice versa. For the feature extraction process, we calculated 16 features that represent students' reading behavior based on the past study [2]. The OpenLA package [3] was used to calculate the features from ebook log data. The LightGBM model was selected as the prediction model which distinguishes students with "A" grade or not based on the data competition results [4].

The F1 scores, which represents the model accuracy, were 0.68 for the 2019 (face-to-face) dataset and 0.59 for the 2020 (online) dataset. An increase in the number of students obtaining "A" grade was observed in the online course (29%) compared to the face-to-face course (18%). This result suggested that the increase in the number of students obtaining "A" grades in online classes could be explained as the higher achievement levels of the student in the online course than the face-to-face course. Our approach of using predictive models to examine the validity of human ratings is applicable for other datasets and can be used widely for evaluating grading fairness across different years and courses.

References

1. Flanagan, B. & Ogata, H. Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal* 10, 469–484 (2018)
2. Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B. & Ogata, H. Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments* 6(4) (2019)
3. Murata, R., Minematsu, T. & Shimada, A. OpenLA: Library for Efficient E-book Log Analysis and Accelerating Learning Analytics. in *28th International Conference on Computers in Education, ICCE 2020* 301–306 (Asia-Pacific Society for Computers in Education, 2020).
4. EDE Data Challenge, https://sites.google.com/view/ede-datachallenge-22/Result, last accessed 2022/03/19