Predicting Peer-Review Participation at Large Scale Using an Ensemble Learning Method

Erkan Er, Eduardo Gómez-Sánchez, Miguel L. Bote-Lorenzo,
Yannis Dimitriadis, Juan I. Asensio-Pérez

GSIC/EMIC, Universidad de Valladolid, Valladolid, Spain.
erkan@gsic.uva.es, {edugom|migbot|yannis|juaase}@tel.uva.es

Abstract. Peer review has been an effective approach for the assessment of massive numbers of student artefacts in MOOCs. However, low student participation is a barrier that can result in inefficiencies in the implementation of peer reviews, disrupting student learning. In this regard, knowing earlier the estimate number of peer works that students will review may bring numerous pedagogical utilities in MOOCs. Previously, we have attempted to predict student participation in peer review in a MOOC context. Building on our previous work, in this study we propose an ensemble learning approach with a refined set of features. Results show that the prediction performance improves when a preceding classification model is trained to identify students with no peer-review participation and that the refined features were effective with more transferability to other contexts.

Keywords: MOOC · Peer review · Engagement prediction · Ensemble learning

1 Introduction

Peer review (or peer assessment), in which an equal-status student assesses a peer's work [1], has been a solution to the evaluation of thousands of student artefacts (e.g., an essay) in MOOCs [2]. However, this solution itself brings some practical challenges at large scale, one of which is the low student participation [3]. Given that MOOC participants have different goals and come from diverse backgrounds, their participation in peer reviews might not be persistent [4]. With low participation rates, a peer review activity might yield various issues. For example, submissions of striving students may receive neither feedback nor a grade, which may lead to a decrease in their motivation to continue the course. Nevertheless, not many researchers have focused on student participation in peer review at large scales [3]. More research is needed to develop practical solutions for effective peer-review activities at large scale. One research line could involve the prediction of students' participation in peer reviews. An accurate estimation of peer-review participation can be utilized in various practical ways. For example, instructors can use this information to tune peer-review activities (e.g., incorporating an adaptive time schedule for completing peer reviews based on students' expected level of participation). This information can be also used to inform the design of

other collaborative activities (e.g., forming groups that are inter-homogenous in terms of students' desire to review teammates' work).

The work presented in [5] was our first attempt to predict the number of peer works a student will review by using regression methods with a large feature set. The results were promising with a reasonably low error that decreases as the course progresses and more data reflecting the student behaviour becomes available. However, the model was built with a large feature set, which may result in overfitting in MOOC contexts with fairly less students participating in peer reviews. Further, a large part of the error was accumulated on those students who submitted their assignment but did not review any peer submission. This paper addresses these limitations by building a new feature set with less yet more informative variables, and by proposing an ensemble learning model. In the following section, we describe the course data at hand and provide the details of our feature-generation approach. Next, we present the experimental study by describing the feature selection approach and the details of the ensemble method. Then, the prediction performance of each prediction model employed are shared. We conclude by discussing follow-up research ideas.

2 Previous Findings

In our previous work [5], we obtained promising results by using regression methods to predict student participation in peer reviews in a MOOC (with 3620 enrollments) published by Canvas Network[1]. The feature set contained more than 80 items, including weekly cumulative features (e.g., number of discussion activities in total during whole week) as well as daily features (e.g., number of content visits per each day before the peer-review activity). There were four assignments involving submission of a learning artefact, and they were evaluated using peer reviews. Figure 1 provides the histograms along with descriptive statistics regarding the number of peer works reviewed by each student. The recommended (or required) number of peer reviews appears to be three as most students performed three peer reviews at each session.

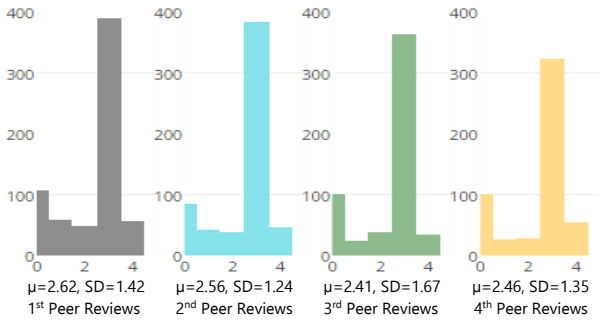

Fig. 1. Peer review participation with mean and standard deviation scores.

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XB2TLU The id of the course is 770000832960949.

The prediction models included one of three regression methods (LASSO –least absolute shrinkage and selection operator, ridge, and elastic net) and the performance of each method was tested. In Table 1 shows the results of the prediction performance with the LASSO regression (which was chosen as it was the best performing method). The total mean absolute error (MAE) scores were reasonably low in general, and the performance improved considerable with the inclusion of past peer-review activities starting from the 2nd peer-review session. However, the prediction of the participation of students with no actual peer-review participation was inaccurate. This finding has a non-negligible impact on the overall error (note that around 1/6 of students who submitted their assignment did not review any of their peers), suggesting a need for reducing the error resulted from the disengaged students to improve the overall prediction performance. Furthermore, we found that many features were redundant, particularly those derived based on student activities on a specific day (e.g., quiz activity 2 days before the peer reviews). Therefore, the predictive model obtained was complex with many features that were particular to the context, limiting the transferability of the model to other MOOCs. Another possible problem could be the overfitting as this complex model were trained and tested on a small sample. The current study addresses the limitations of the previous work by studying more deeply the feature space and proposing an ensemble learning approach, as described in the following sections.

Table 1. The MAE scores per each actual value of the peer-review participation.

	0	**1**	**2**	**3**	**4**	**TOTAL**
Peer reviews 1	2.24	1.25	0.39	0.64	1.60	1.02
Peer reviews 2	1.59	0.82	0.67	0.40	1.08	0.66
Peer reviews 3	1.18	0.90	0.77	0.32	0.93	0.56
Peer reviews 4	1.12	1.04	0.71	0.31	0.97	0.58

3 Improvements

3.1 Feature Generation

Given the limitations of the features used previously, we have revised them to obtain a reduced yet predictive set that can be transferable over different peer-review sessions within the same course and that can also apply to other MOOC contexts. For this purpose, we mainly adopted the features proposed in [6], which are based on edX MOOCs. Given that Canvas Network MOOCs have a different database structure than edX MOOCs, we have either adopted similar features or extracted the same ones when possible. The effectiveness of such features in predicting student engagement in MOOCs has been shown [7]. These features could be effective in predicting students' peer-review participation as their overall course engagement is likely to be associated with their peer-review engagement [8]. Each feature was computed using the data between consecutive peer-review sessions (e.g., features for the 3rd peer reviews were calculated using the data obtained after the 2nd peer reviews) since students' recent activities could be more relevant to their subsequent peer-review participation.

Furthermore, features about learners' activity sequences (e.g., taking a quiz followed by reading) can be powerful predictors of engagement in MOOC contexts [10]. The sequence features are about the order of student activities and can help to identify different student profiles. Sequence features can easily scale up to thousands as activities could follow many different orders [10]. To obtain a small yet relevant set, we decided to focus on assignment, discussion and content activities and generated 2-activity length features. The complete list of features generated (n=41) is provided in Table 2.

Table 2. Features extracted for the prediction of participation in peer reviews

{*a*}_count	Number of *a*-type requests.
days_with_{*a*}	Number of days with at least one *a*-type request.
avgt_btw_{*a*} [1]	Average time in minutes between *a*-type requests
{*a*}_within1h [1]	Number of *a*-type requests within a one-hour interval.
uncomp_qs	Number of uncomplete quiz submissions.
comp_qs	Number of successful quiz submissions.
ttl_quizattempts	Number of quiz attempts
avg_quizattempts	Average number of quiz attempts
ttl_quiz_time	Total time spent in quizzes (in minutes).
avg_quiz_time	Average time spent in quizzes (in minutes).
avg_qs_score	Average quiz scores.
de_count	Total number of discussion entries.
de_msg_cc	Average character-length of the discussion entries posted.
days_with_de	Number of days with at least one discussion entry.
assign_score	Past assignment score.
pr_subms_count	Number of student submissions reviewed.
pr_count [2]	Number of past peer reviews performed.
reviews_received	Number of reviews received for the previous assignment of a student.
da_count [3]	Number of discussion-assignment activity sequences.
qa_count [3]	Number of quiz-assignment activity sequences.
ca_count [3]	Number of content-assignment activity sequences.
ad_count [3]	Number of assignment-discussion activity sequences.
qd_count [3]	Number of quiz-discussion activity sequences.
cd_count [3]	Number of content-discussion activity sequences.
ac_count [3]	Number of assignment-content activity sequences.
qc_count [3]	Number of quiz-content activity sequences.
dc_count [3]	Number of discussion-content sequences.

a denotes the type of the request (content, quiz, assignment, or discussion); [1] is also calculated combining all requests; [2] is different than *pr_subms_count* if students reviewed the same submission multiple times; and [3] are divided by the total number of requests.

3.2 Ensemble Learning Method

Ensemble learning method is a type of machine learning technique that involves the use of multiple learning algorithms to achieve higher predictive performance than what could be achieved using a single learning algorithm. Ensemble methods are found to improve predictive models in the MOOC literature [11]. The motivation for using an ensemble learning method for the current prediction task has emerged from our previous work, in which we found that overall prediction performance suffers largely from

poorly predicting the participation of students who have zero actual peer-review participation. Identifying such students beforehand using classification methods (i.e., non-participants vs participants) and running the regression models for only participants of peer reviews might potentially lead to higher accuracy. Therefore, to improve the prediction accuracy, we propose a sequential ensemble approach [12], in which a classification step is integrated prior to regression to identify those with no peer reviews ahead of time and exclude them from the regression analysis. Later, those classified as having no participation were combined with the regression predictions to evaluate the overall performance. Figure 2 depicts the ensemble method proposed.

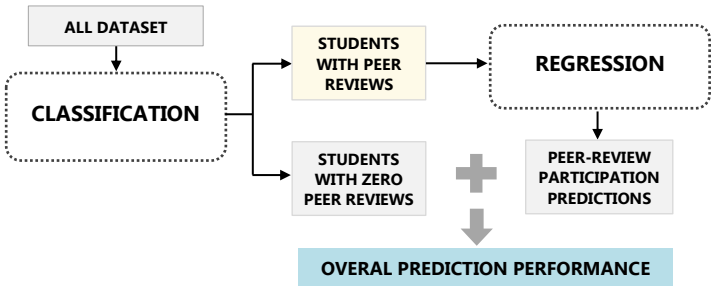

Fig. 2. The components of the ensemble method.

4 Experimental Study

4.1 Method

First, we replicated our previous study with revised feature set. Two regression methods were tested. The first one is LASSO, which has an internal-feature selection mechanism based on L1 regularization. LASSO has been effective in previous MOOC research [13]. However, LASSO may have performance issues when features are correlated [14], which might be the case in the current study as some features were extracted from similar data. Therefore, we also used a correlation-based feature-selection (CFS) [15] to train a linear regression (LR) model. CFS focuses on the predictive ability of each feature while maintaining a low correlation among them to minimize the redundancy.

In the ensemble learning model, logistic regression (LGR) was chosen as the classifier as it was found to be more accurate compared to the others that were pilot-tested (e.g., stochastic gradient descent and decision trees). L1 regularization and CFS were also used to perform feature-selection for the classification model. While whole dataset was used to train the classification model, only data about students with at least one peer review was used to train the regression model. Only students who submitted the corresponding assignment were included in predictions since only those students could review others' submissions. Beginning with the 2nd assignment, features of previous assignment score and peer-review participation were included in the predictions. Since the sample size was small, 10-fold cross validation method was used, and the performance was evaluated using MAE [16]. MAE was used as the metric since it provides

plain interpretation of performance when target variable has a narrow range (i.e., 0-4). Also, please note that prediction scores were rounded to the closest integer value (as decimal numbers would not be practical in a real course). We used the scikit-learn implementations of LASSO, LGR, and LR, and WEKA implementation of CFS.

4.2 Results and Discussion

The MAE scores at each actual participation level, which is 0 to 4, as well as the total MAE scores of each prediction model are provided in Table 3 and Table 4. When compared to the previous results (see Table 1), the performance of the regression model (see Table 3) seemed to remain almost the same with the refined list of features, with a similar trend of increasing accuracy at each subsequent prediction. The error rates were the highest at the 0-participation level. Given the likelihood of overfitting with complex models, we favour the use of the refined feature set to minimize this possibility. Also, the current feature set has the capacity to be transferred to any other week involving a peer-review prediction as well as to other MOOCs.

Table 3. The MAE scores per each actual value at each peer-review session when L1 regularization is used for the feature-selection.

		0	**1**	**2**	**3**	**4**	**TOTAL**
1st Peer	Regression	2.06	1.08	0.24	0.75	1.68	1.04 (Std. = 0.40)
Reviews	Ensemble	2.06	1.08	0.24	0.76	1.68	1.04 (Std. = 0.41)
2nd Peer	Regression	1.73	0.71	0.76	0.23	1.31	0.60 (Std. = 0.75)
Reviews	Ensemble	1.59	0.83	0.79	0.24	1.30	0.59 (Std. = 0.84)
3rd Peer	Regression	1.19	0.78	0.82	0.20	0.88	0.49 (Std. = 0.94)
Reviews	Ensemble	0.74	1.08	1.05	0.20	1.06	0.45 (Std. = 1.12)
4th Peer	Regression	1.06	1.03	0.73	0.21	0.98	0.52 (Std. = 0.99)
Reviews	Ensemble	0.73	1.28	0.97	0.23	0.98	0.50 (Std. = 1.16)

Table 4. The MAE scores per each actual value at each peer-review session when CFS is used for the feature-selection.

		0	**1**	**2**	**3**	**4**	**TOTAL**
1st Peer	Regression	2.05	1.13	0.33	0.70	1.63	1.01 (Std. = 0.46)
Reviews	Ensemble	2.05	1.13	0.33	0.70	1.63	1.01 (Std. = 0.43)
2nd Peer	Regression	1.68	0.66	0.71	0.28	1.17	0.60 (Std. = 0.79)
Reviews	Ensemble	1.45	0.97	0.74	0.25	1.32	0.58 (Std. = 0.92)
3rd Peer	Regression	1.10	0.78	0.85	0.24	0.91	0.50 (Std. = 0.99)
Reviews	Ensemble	0.75	1.17	0.89	0.22	1.00	0.45 (Std. = 1.14)
4th Peer	Regression	0.96	1.03	0.73	0.22	0.98	0.51 (Std. = 1.03)
Reviews	Ensemble	0.73	1.28	0.93	0.23	0.98	0.50 (Std. = 1.16)

According to the results of the ensemble model in Table 3, the prediction performance has slightly increased (except the 1st peer reviews) when a classification phase is incorporated before running the regression model, compared to the performance of regression alone. That is, the classification model helped reduce the error introduced by students with zero peer-review participation. However, at the same time, it seems

that the error increased in the prediction of other levels of participation due to poor classification performance. Also, no improvement was noted for the predictions at the 1st peer-review session probably because students who do and who do not contribute to peer reviews seem to have very similar profiles at this stage of the course based on the current feature set used. Further, the feature-selection methods did not appear to have different effects on the prediction performance.

The results showed that the proposed ensemble method produced better predictions than that obtained using the regression method alone. This was because students with no peer-review participation were undermining the performance of the regression model, which was addressed by incorporating a classification phase to identify and exclude those with no participation when training the regression model. However, the overall performance did not improve considerably as the students with no peer-review participation were not classified perfectly, therefore yielding a mediocre performance at certain levels of participation. Nonetheless, given that the standard deviation of actual peer-review participation has a range of 2.41-2.62, the MAE scores achieved with the ensemble method seem to be promising, ranging from 0.45 to 1.04. Thus, the proposed predictive model holds potential to be utilized in a real MOOC context.

5 Conclusion and Future Work

In this study, building on our previous work we proposed a sequential ensemble learning method with a refined set of features to obtain an accurate prediction of students' peer-review participation. The results showed that proposed ensemble model holds a potential to be further explored in future research. First, the classification model needs further attention. The reasons for its moderate performance needs to be explored and addressed accordingly using different classification approaches and more relevant features. For example, a nested ensemble approach could be utilized. Second, the ensemble method failed to improve the prediction performance at the 1st peer reviews. Possibly student profiles as identified with the current feature set was not distinctive early in the course, and therefore they offered no benefits for the classification. More distinctive features need to be identified to improve the classification performance. Nonetheless, the challenge of identifying students who will not participate in peer reviews early in the semester constitute an interesting research opportunity. Moreover, although the approach used in this study demonstrates the validity of the prediction model, it is not applicable to an ongoing MOOC as the values of the target variable (which is the number of peer work reviewed) would be needed to train the models. Therefore, other relevant training paradigms (e.g., in-situ learning) should be used to build accurate yet practical models that can be useful in continuing MOOCs [17].

6 Acknowledgements

Access to the data used in this paper was granted by Canvas Network. This work has been partially funded by research projects TIN2014-53199-C3-2-R and VA082U16, and by the Spanish network of excellence SNOLA (TIN2015-71669-REDT).

References

1. Topping, K.: Peer assessment between students in colleges and universities. Rev. Educ. Res. 68, 249–276 (1998).
2. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned Models of Peer Assessment in MOOCs. In: International Conference on Educational Data Mining. pp. 153–160 (2013).
3. Estevez-Ayres, I., Crespo-García, R.M., Fisteus, J.A., Delgado-Kloos, C.: An algorithm for peer review matching in Massive courses for minimising students' frustration. J. Univers. Comput. Sci. 19, 2173–2197 (2013).
4. Suen, H.: Peer assessment for massive open online courses (MOOCs). Int. Rev. Res. Open Distrib. Learn. 15, (2014).
5. Er, E., Bote-Lorenzo, M.L., Gómez-Sánchez, E., Dimitriadis, Y., Asensio-Pérez, J.I.: Predicting Student Participation in Peer Reviews in MOOCs. In: Proceedings of the Second European MOOCs Stakeholder Summit 2017. , Madrid (2017).
6. Veeramachaneni, K., O'Reilly, U.-M., Taylor, C.: Towards Feature Engineering at Scale for Data from Massive Open Online Courses. arXiv:1407.5238v1. 6, (2014).
7. Jayaprasad, S., Jayaprasad, S.: Transfer Learning for Predictive Models in Massive Open Online Courses. Artif. Intell. 1–12 (2015).
8. Tseng, S.-F., Tsao, Y.-W., Yu, L.-C., Chan, C.-L., Lai, K.R.: Who will pass? Analyzing learner behaviors in MOOCs. Res. Pract. Technol. Enhanc. Learn. 11, 1–11 (2016).
9. Crossley, S., Paquette, L., Dascalu, M., McNamara, D.S., Baker, R.S.: Combining click-stream data with NLP tools to better understand MOOC completion. Proc. Sixth Int. Conf. Learn. Anal. Knowl. - LAK '16. 6–14 (2016).
10. Brooks, C., Thompson, C., Teasley, S.: A time series interaction analysis method for building predictive models of learners using log data. Proc. Fifth Int. Conf. Learn. Anal. Knowl. - LAK '15. 126–135 (2015).
11. Boyer, S., Veeramachaneni, K.: Robust Predictive Models on MOOCs: Transferring Knowledge across Courses. Proc. 9th Int. Conf. Educ. Data Min. 298–305 (2016).
12. Zhou, Z.-H.: Ensemble Learning. In: Li, S.Z. (ed.) Encyclopedia of Biometrics. pp. 2170–273 (2009).
13. Robinson, C., Yeomans, M., Reich, J., Hulleman, C., Gehlbach, H.: Forecasting student achievement in MOOCs with natural language processing. Proc. Sixth Int. Conf. Learn. Anal. Knowl. - LAK '16. 383–387 (2016).
14. Zou, H., Hastie, T.: Regularization and variable selection via the elastic-net. J. R. Stat. Soc. 67, 301–320 (2005).
15. Hall, M.: Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 359–366 (2000).
16. Sawyer, R.: Sample size and the accuracy of predictions made from multiple regression equations. J. Educ. Stat. 7, 91–104 (1982).
17. Bote-Lorenzo, M.L., Gómez-Sánchez, E.: Predicting the decrease of engagement indicators in a MOOC. In: Seventh International Conference on Learning Analytics and Knowledge. pp. 143–147 (2017).