Supporting Group Formation in Ongoing MOOCs using Actionable Predictive Models

Erkan Er
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
erkan@gsic.uva.es

Eduardo Gómez-Sánchez
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
edugom@tel.uva.es

Miguel L. Bote-Lorenzo
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
migbot@tel.uva.es

Juan I. Asensio-Pérez
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
juaase@tel.uva.es

Yannis Dimitriadis
School of Telecommunications Engineering
Universidad de Valladolid
Valladolid, Spain
yannis@tel.uva.es

Abstract— **Although the massive and open nature of MOOCs necessitates more technology support for instructors, existing prediction research has been barely capable of offering real-world solutions. One critical case where MOOC instructors could benefit from real-time technological support is the design of collaborative activities, in particular group formation. In this regard, this research work investigated the use of in-situ learning technique to produce useful and actionable information that could assist instructors in group formation while the course continues. Focusing on a particular MOOC context, a predictive model was created to compute the probability that students would participate in group discussions or not. Using these probability scores, actual group behavior was also predicted for three cases: at least 2, 3 or 4 different students would post in group discussions. According to the results, the model was able to accurately predict individual student behavior as well as group behavior before the actual collaborative activity had taken place, suggesting its potential for real-time use. Future research involves the exploration of other approaches for creating actionable predictions and the application of the predictions in practice in an ongoing MOOC.**

***Keywords*— moocs, transfer learning, predictive models, collaborative learning**

I. Introduction

In the last years, predictive analytics have been a popular research area in the MOOC literature. There has been an abundance of studies demonstrating the relevancy of particular prediction techniques for the estimation of various learning behaviors in MOOCs[1]–[3]. However, the research findings are generally limited to theoretical contributions with minor impact in advancing the MOOC practice [4]. That is, most existing predictive models are not feasible for real-time use in ongoing courses as they are built using posterior data (which is available only after the target action is realized) [4]. For example, these models can predict whether a student will drop out (or not) only after the course is completed and students actually drop out the course. The use of transfer learning techniques such as transferring across courses or in-situ learning have been investigated as a possible solution to overcome this limitation [5], [6]. The models trained using transfer learning are found to perform as accurate as those optimistic models trained using post-hoc approaches such as cross-validation [7]. Despite the evidence regarding their capacity for creating actionable and accurate predictive models, transfer learning techniques have been rarely applied in real-world MOOC contexts for improving teaching and learning activities, and thus having limited impact in advancing the MOOC pedagogy.

On the other hand, in MOOCs there is more need for supporting instructors with timely technological solutions since addressing a massive learner population demands a teaching workload that is not feasible to supply manually by the instructors [8], [9]. Accordingly, MOOC platforms embody various features to allow instructors to perform various tasks automatically. A particular case that obligates the use of such features is the formation of student groups for conducting collaborative activities. Many MOOC platforms assists instructors in forming groups by randomly matching students, which otherwise would not be possible using manual methods considering the massive number of learners [10]. However, such automated features may have limited capacity in enacting the pedagogy intended by the instructor. Given that groups are created randomly, the performance of the resulting groups would be random as well: for example, while in some groups all group members would participate in the collaborative activity, in some other groups students might suffer from peers who even barely visit the course. Therefore, with random grouping approaches, many MOOC learners are likely to miss what collaborative learning could offer such as developing higher-order thinking skills and improving conceptual understanding [11]–[13].

Given the gaps in the literature, this research work offers a novel approach to support group formation in ongoing MOOCs in a way that can potentially promote collaboration among peers. We focused on a particular MOOC context and investigated: (1) the use of in-situ learning as a transfer learning approach to create an actionable model for predicting student engagement in collaborative activities, and (2) the potentials of this predictive model in real-world practice for assisting instructors in forming engaging collaborative groups. This research is guided by the following research question: *How can in-situ learning be used to inform group formation in ongoing MOOCs?*

The paper is structured as follows. After the related work is discussed in Section II, the current research work is presented in Section III, where the approach, context, and method are explained. The results are presented and discussed in Section IV. The paper concludes with Section V where future research opportunities are presented.

II. RELATED WORK

Collaborative learning is a learner-centered and team-based pedagogical strategy that promotes active learning of students [12]. In the MOOC literature, collaborative learning has been considered as one of the approaches to be adapted to promote student learning and decrease the dropout rates [9]. Several efforts [10], [14]–[17] have been devoted to formation of student groups within massive and diverse learner population in MOOCs in a way that can enhance collaboration between peers, and therefore supporting their learning. Besides several theoretical works [14], [15], in most of these efforts, group formation task has been considered as a optimization problem, and various algorithmic approaches has been investigated to address it [16], [17]. That is, the focus has been on the research methods regarding group formation rather than supporting instructors with useful information that can be used to form groups that meet pedagogical intentions of the instructor. As described in the following section, differently, the current research work focuses on the generation of actionable information that can be utilized to inform the group formation and the collaborative activity as desired by the instructor.

III. THE CURRENT RESEARCH WORK

A. The Approach

The proposed approach involves the early predictions of students' possible engagement in an upcoming collaborative activity, and the use of these predictions to support group formation for enhanced collaboration between students. To achieve predictions that are actionable in a real-world context, the approach uses techniques such as transferring across courses or in-situ learning [4], [6] for training the predictive models. Such techniques take advantage of previous course data or proxy labels and allow training models only using the information available at the time of the prediction. Following this approach, in the present research, a predictive model was built to estimate student participation in a collaborative activity. In particular, in-situ learning technique was used to train a classification model that estimates how likely each student would post in their own group discussions. The use of transferring across courses was not feasible because there were no previous runs of the course.

Once an accurate model is built, the individual probability scores (i.e., the likelihood of each student's participation behavior) generated by the model can be utilized in several ways to support MOOC instructors during the formation of collaborative groups. First, individual probabilities can be used to form groups according to the instructors' pedagogical design. For example, groups can be formed homogenously or heterogeneously according students' estimated level of participation. Additionally, the prediction scores could be combined with other variables (e.g., learning styles) to design a more complex grouping criterion. Second, the probability scores can be used to optimize the group formation task in an effort to satisfy the instructors' desired level of participation (e.g., at least 2 students will participate in groups of size 4). Last, the proposed predictive model can inform back the instructors of how many groups would meet their demanded level of group activity. Accordingly, the instructor can use this information to update the design of the collaborative activity (e.g., increasing or decreasing the group size or the requirements for group submission).

B. The Context

The context is a MOOC[1] that teaches translation of business terms between English and Spanish languages. Total number of enrolled students was 1031. The course was composed of 7 blocks (or modules or weeks), and it had two collaborative activities (in the third and fifth weeks), which involved the extraction of relevant terms from given documents dealing with finance. Groups of six participants were formed for carrying out these two collaborative activities. Within the scope of this research, only the first collaborative activity is studied.

C. Method

1) Feature generation. In order to build the classification model, 12 features regarding students' engagement in the core course components (i.e., discussion forums, quizzes, assignments, introduction to modules, review videos, lecture content pages) were generated. These features were generated based on MOOC prediction literature [18] and our previous work [19], [20]. The features and their descriptions are given in Table I.

TABLE I. FEATURES GENERATED TO BUILD THE PREDICTIVE MODEL

Feature	**Description**
TTL_DISCPOSTCOUNT	Total number of discussion posts
TTL_QUIZATTEMPT	Total number of quiz attempts
TTL_QUIZSCORE	Total quiz score
TTL_QUIZTIMESPENT	Total time spent on quizzes
TTL_ASSIGN_SBM	Total number of assignment submissions
TTL_PAGEVIEW	Total page views
TTL_INTRO_VIEW	Total views on introduction pages
TTL_LECCONT_VIEW	Total views on lecture pages
TTL_DISC_VIEW	Total views on discussion forums
TTL_QUIZ_VIEW	Total views on quiz pages
TTL_ASSIGN_VIEW	Total views on assignment pages
TTL_VIDEO_VIEW	Total views on review-video pages

2) Building the prediction model. Logistic regression was used as the classifier as it has been effective in various classification tasks in the MOOC literature [5], [21]. Feature selection was performed using L1 regularization. The Scikit-Learn [22] implementations of logistic regression and L1 regularization were used. The predictive model was trained using both cross validation and in-situ learning approach. 10-fold cross validation was used in which the data that contains the real class labels (e.g., whether a student posted in group discussions or not) were split into 10 folds, and then 9 folds were used as training set while the remaining fold was reserved for testing (this process is repeated for each combination of folds). For training with in-situ learning, a

[1] https://learn.canvas.net/courses/1343

proxy label was identified, which was students' submission statuses of the assignment in the second week (i.e., one week before the collaborative activity). A model was trained using only past data available until the end of the second week and tested using the all data available until the end of the third week. All data were standardized before performing feature selection and training the model.

3) Assessing the performance of the the prediction model. The performance of the model was first assessed in terms of predicting individual behavior (i.e., posting at least once in group discussions) using area under the curve (AUC) score as the measure. AUC was particularly chosen because the class distribution was unbalanced in the current dataset, and AUC is robust to the prediction bias caused by unbalanced class distributions [23]. The categorization of model performance based on AUC scores is: .9-1: excellent, .8-.9: very good, .7-.8: good, .6-.7: fair, and .5-.6: bad [24].

In order to assess the model performance in terms of predicting group behavior, the following participation levels in groups were identified: whether at least 2, or at least 3, or at least 4 different students would post in their group discussions. This criteria of participation level within groups would be determined by the instructor later when intervening group formation while the course continues. The estimation of group behavior was computed based on individual probability scores. The accuracy of the predictions was assessed again using AUC. A high accordance between the predicted and the actual group behavior would ensure that the actionable information obtained with in-situ learning can accurately predict level of participation within groups, thus supporting its potential use in an ongoing course.

IV. Results and Discussion

The approach described above is applied in the current MOOC context to generate actionable information for forming student groups in the collaborative activity. First, using in-situ learning, the predictive model was built to predict if students would post in their group discussions or not, where students' submissions to the individual assignment in week 2 (last assignment before the collaborative activity starts in week 3) were used as proxy labels for training.

First, the model performance was assessed in terms of predicting individual student behaviors. As seen in Table II, AUC scores were 0.851 and 0.843 using cross validation and in-situ learning as training approaches respectively. That is, in-situ learning has resulted in predictions as accurate as those obtained with cross validation, which cannot be used in real time and produces optimistic estimations [5]. Therefore, these results suggest that using the ins-situ learning technique, the predictions that students would post in their discussion groups were quite accurate and actionable, thus having great potential for real-world use.

TABLE II. The AUC scores of prediction accuracies using CV and in-situ learning

Cross validation	**In-situ learning**
0.851	0. 843

The feature importance was assessed based on the coefficient calculated by L1 regularization as provided in Table III. According to the results, 6 features were identified to have a certain level of predictive capacity. Among them, assignment-related features were the most predictive, followed by quizzes and discussions.

TABLE III. Feature Importances Based on L1 Regularization

Feature	**Coefficient**
TTL_ASSIGN_VIEW	1.207
TTL_ASSIGN_SBM	0.761
TTL_QUIZSCORE	0.455
TTL_DISC_VIEW	0.205
TTL_QUIZTIMESPENT	-0.053
TTL_INTRO_VIEW	-.0.069

Next, the model performance was assessed in terms of predicting group behavior (i.e., at least 2, 3, or 4 students would participate in group discussions). Before the analysis, participants with no activity in the course were excluded, which resulted in varying number of students in each group (although initially it was set to six by the instructor). This was necessary to avoid bias in the predictions that would be caused by 0 probability scores for students with no course activity. The prediction accuracies at each participation level are provided in Table IV. According to the AUC scores (ranging from 0.823 to 0.884), it was possible to predict the actual group behavior accurately using the individual probability scores obtained before the collaborative activity had started. Therefore, these results suggest the potential use of the individual probability scores in assisting collaborative group formation while the course continues.

TABLE IV. The AUC scores of prediction accuracies Regarding Group Behavior

Number of participants		
At least 2	***At least 3***	***At least 4***
0.823	0.877	0.884

V. Conclusions and Future Work

This research work investigated the use of a specific transfer learning technique called in-situ learning for training a predictive model that can produce actionable yet accurate information. The results provided enough evidences regarding the true potential of the produced actionable information in predicting actual group behavior in the same context for which the model was intended. Thus, using in-situ learning technique, it was possible to predict individual and group level participations in group discussions as accurately as batch models (such as cross validation) that are non-realistic for real-world use in ongoing MOOCs.

In the current research, only in-situ learning was used for training the predictive model. However, transferring across courses, another common approach for obtaining actionable predictions in MOOCs [4], [5], should be also investigated in the future runs of the same course. Once the most effective training approach is decided, future work should focus on the real-world use of the proposed approach for supporting instructors during the design and implementation of collaborative learning activities. In the future run of the same MOOC, we plan to apply the same approach to produce actionable information for forming groups in a way that maximizes the interaction among students within each group (e.g., in each group at least 2 students would participate in discussions). This criteria in group formation would be determined by the instructor depending on the design of the collaborative activity. Additionally, future research is needed to replicate this work in MOOCs on different areas (e.g., engineering, philosophia) to verify the applicability of the proposed research in various contexts.

ACKNOWLEDGEMENTS

This work has been partially funded by research projects SmartLET (TIN2017-85179-C3-2-R) and RESET (TIN2014-53199-C3-2-R) funded by the Spanish State Research Agency and the European Regional Development Fund; JCyL2016-2018 (VA082U16) funded by the Regional Government of Castilla y León and the European Regional Development Fund; ColMOOC (588438-EPP-1-2017-1-EL-EPPKA2-KA) funded by the European Commission; and by the Spanish network of excellence SNOLA (TIN2015-71669-REDT). Access to the data used in this paper was granted by Canvas Network.

REFERENCES

[1] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, “Predicting MOOC performance with week 1 behavior,” in Proceedings of the Seventh International Conference on Educational Data Mining (EDM), 2014, pp. 273–275.

[2] M. Romero and M. Usart, “The time factor in MOOCS: Time-on-task, interaction temporal patterns, and time perspectives in a MOOC,” 6th Int. Conf. Comput. Support. Educ. CSEDU 2014, vol. 1, pp. 53–62, 2014.

[3] M. L. Bote-Lorenzo and E. Gómez-Sánchez, “Predicting the decrease of engagement indicators in a MOOC,” in Proceedings of Seventh International Conference on Learning Analytics and Knowledge, 2017, pp. 143–147.

[4] J. Gardner and C. Brooks, “Student success prediction in MOOCs,” User Model. User-adapt. Interact., vol. 28, no. 2, pp. 127–203, 2018.

[5] S. Boyer and K. Veeramachaneni, “Transfer learning for predictive models in Massive Open Online Courses,” in Proceedings of the 17th Conference on Artificial Intelligence in Education, Madrid, Spain, 2015, pp. 54–63.

[6] S. Boyer and K. Veeramachaneni, “Robust predictive models on MOOCs: Transferring knowledge across courses,” in Proceedings of the Ninth International Conference on Educational Data Mining, 2016, pp. 298–305.

[7] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, “MOOC dropout prediction: How to measure accuracy?,” in Proceedings of the Fourth ACM Conference on Learning@Scale, 2017, pp. 161–164.

[8] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, “Learning about social learning in MOOCs: From statistical analysis to generative model,” IEEE Trans. Learn. Technol., vol. 7, no. 4, pp. 346–359, 2014.

[9] P. Dillenbourg, A. Fox, C. Kirchner, and M. Wirsing, “Massive Open Online Courses: Current State and Perspectives,” Dagstuhl Manifestos, vol. 4, no. 1, pp. 1–27, 2014.

[10] L. Sanz-Martínez, J. A. Muñoz-Cristobal, M. L. Bote-Lorenzo, A. Martínez-Monés, and Y. Dimitriadis, “Toward criteria-based automatic group formation in MOOCs,” in Proceedings of the Fifth European MOOCs Stakeholders Summit, 2017.

[11] D. Jonassen, M. Davidson, M. Collins, J. Campbell, and B. Haag, “Constructivism and computer mediated communication in distance education,” Am. J. Distance Educ., vol. 9, no. 2, pp. 7–25, 1995.

[12] P. Dillenbourg, “What do you mean by collaborative learning?,” in Collaborative learning: Cognitive approaches, & L. C. J. R. Setchi, I. Jordanov, R. J. Howlett, Ed. Oxford: Elsevier, 1999, pp. 1–19.

[13] P. Dillenbourg, Collaborative Learning: Cognitive and Computational Approaches. New York, NY: Elsevier Science, Inc., 1999.

[14] T. Sinha, “Together we stand, Together we fall, Together we win: Dynamic team formation in massive open online courses,” Fifth Int. Conf. Appl. Digit. Inf. Web Technol. (ICADIWT 2014), pp. 107–112, 2014.

[15] H. Spoelstra, P. van Rosmalen, and P. Sloep, “Toward Project-based Learning and Team Formation in Open Learning Environments,” J. Univers. Comput. Sci., vol. 20, no. 1, 2014.

[16] I. Srba, M. Bielikova, and S. Member, “Dynamic Group Formation as an Approach to Collaborative Learning Support,” no. JUNE, pp. 173–186, 2015.

[17] C. Tucker, B. Pursel, and A. Divinsky, “The impact of small learning group composition on student engagement and success in a MOOC,” in Proceedings of the 8th International Conference on Educational Data Mining, 2014.

[18] K. Veeramachaneni, U.-M. O’Reilly, and C. Taylor, “Towards feature engineering at scale for data from massive open online courses,” arXiv Prepr. arXiv1407.5238., 2014.

[19] E. Er, E. Gómez-Sánchez, M. L. Bote-Lorenzo, Y. Dimitriadis, and J. I. Asensio-Pérez, “Predicting peer-review participation at large scale using an ensemble learning method,” in Learning Analytics Summer Institute, 2017, pp. 55–62.

[20] E. Er, M. L. Bote-Lorenzo, E. Gómez-Sánchez, Y. Dimitriadis, and J. I. Asensio-Pérez, “Predicting student participation in peer reviews in MOOCs,” in Proceedings of the Second European MOOCs Stakeholder Summit 2017, 2017.

[21] D. B. Kurka, A. Godoy, and F. J. Von Zuben, “Delving deeper into MOOC student dropout prediction,” CEUR Workshop Proc., vol. 1691, pp. 21–27, 2016.

[22] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” vol. 12, pp. 2825–2830, 2012.

[23] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data recommendations for the use of performance metrics.,” in Proceedings of the Fifth Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245–251.

[24] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, “Evaluation measures for models assessment over imbalanced data sets,” J. Inf. Eng. Appl., vol. 3, no. 10, pp. 27–39, 2013.