Proceedings of the International Conference on Collaboration Technologies and Social Computing (CollabTech2020), Tartu, Estonia, 8-11 September 2020. DOI: https://doi.org/10.1007/978-3-030-58157-2_11

Quantifying Collaboration Quality in Face-to-Face Classroom Settings Using MMLA

Pankaj Chejara[1], Luis P. Prieto[1], Adolfo Ruiz-Calleja[2], María Jesús Rodríguez-Triana[1], Shashi Kant Shankar[1], and Reet Kasepalu[1]

[1] Tallinn University, Tallinn, Estonia
{pankajch, lprisan, mjrt, shashik, reetkase }@tlu.ee
[2] GSIC-EMIC Group, University of Valladolid, Spain
adolfo@gsic.uva.es

Abstract. The estimation of collaboration quality using manual observation and coding is a tedious and difficult task. Researchers have proposed the automation of this process by estimation into few categories (e.g., high vs. low collaboration). However, such categorical estimation lacks in depth and actionability, which can be critical for practitioners. We present a case study that evaluates the feasibility of quantifying collaboration quality and its multiple sub-dimensions (e.g., collaboration flow) in an authentic classroom setting. We collected multimodal data (audio and logs) from two groups collaborating face-to-face and in a collaborative writing task. The paper describes our exploration of different machine learning models and compares their performance with that of human coders, in the task of estimating collaboration quality along a continuum. Our results show that it is feasible to quantitatively estimate collaboration quality and its sub-dimensions, even from simple features of audio and log data, using machine learning. These findings open possibilities for in-depth automated quantification of collaboration quality, and the use of more advanced features and algorithms to get their performance closer to that of human coders.

Keywords: Computer-Supported Collaborative Learning · Multimodal Learning Analytics · Collaboration Quality

1 Introduction

Collaboration has been traditionally studied using observation, interviews and ethnographic methods [10]. Although these methods offer in-detailed information, they also demand a lot of human effort and time, which are difficult to scale up [10]. The use of technology to mediate collaboration has provided researchers with large amounts of learner activity data (in the form of logs), offering an alternative to traditional analyses of collaboration. Researchers have used a variety of data (e.g., system logs, chats, discussion forums) to understand the underlying process of collaboration using Learning Analytics (LA) methods like content analysis, and interaction analysis [3]. The results of these analyses have

been employed to develop various kinds of feedback systems, from mirroring to guiding support [6]. While Computer-Supported Collaborative Learning (CSCL) often involves face-to-face and computer-mediated interactions, collaborative LA support often relies on just digital logs, thus offering only a partial picture of the interactions. Aware of this limitation, the field of Multimodal Learning Analytics (MMLA) [2] emerged with the goal of understanding learning through multimodal data from digital and physical spaces. Recent MMLA studies showed that it is feasible to estimate collaboration aspects categorically (e.g., high vs. low collaboration) in face-to-face settings by combining physical and digital activity traces [12, 13]. In addition, researchers have found verbal interactions and speaking activity features as an important indicator for collaboration behavior [1, 7]. However, most of these studies are conducted in laboratory settings, so their results might not hold under authentic classroom constraints (e.g., noisy data). Moreover, the qualitative estimation of collaboration quality into a few classes provides end users (e.g., a teacher) with little information about what might be the underlying problem or reason why collaboration quality is high/low.

In order to estimate quality of collaboration in a more fine-grained fashion, this paper explores regression analysis models to quantitatively estimate collaboration quality in a classroom setting from audio and log data. To reach that goal, we carried out a case-study where we collected data from two groups (each with four participants) in an authentic classroom setting. The learning activity involved face-to-face discussion and collaborative writing using digital means. We applied various regression models and compared their performance with that of human coders, in the task of coding collaboration quality and its sub-dimensions along a continuum.

2 Related Work

Researchers have investigated the problem of estimating collaboration into a limited set of categories, in various settings: pair-programming [4], project-based learning [12], and tabletop-based collaborative learning [8]. These studies collected data through different means (audio [1, 13], Kinect sensors [4], system logs [7, 13], and video [13]) and extracted a wide variety of features from them, e.g.: non-verbal features like MFCC features, energy [1]; or spatial and dynamic features like hand movement or distance between learners [12]. These features were in turn used to estimate different aspects of the collaboration process: collaboration quality [1, 13, 8], or success in collaboration [12]. Certain studies [4, 13, 8] have included data from *both* physical and digital spaces to investigate collaboration behavior.

While most of these studies devised their own coding schemes to annotate or classify collaboration quality, others [8] have used collaboration rating schemes that are widely used in the collaborative learning sciences (e.g., [9]). Although these rating schemes often output quantitative scores (e.g., collaboration quality [8], grading of collaboration work [12]), such scores have often been mapped into two or three categories (e.g., high vs. low collaboration), as binary clas-

sification is an easier problem (from an information theory point of view) and often results in better performance when using machine learning models [13, 12]. However, this "flattening" of the scores also takes away much of the nuance and the different aspects that contribute to high-quality collaboration. In terms of performance, classification accuracy has been reported from above average (48% [4]) to moderate (69% [7]) and high level (80% [12], 96% [13]).

A number of gaps emerge from the aforementioned state of the art. First, that MMLA researchers have mostly built models to estimate collaboration quality categorically, which offers limited information about the reasons or underlying structure of that judgement (i.e., limited explainability and actionability). In consequence, there is still a lack of understanding regarding whether we can estimate collaboration quality along a continuum (or to what extent). Third, that MMLA studies often report their results without frames of reference that can help the community understand how far (or how close) we are to developing solutions of practical relevance to our classrooms (e.g., how they compare with human-level classification or quantification of collaboration quality).

3 Methodology

To address the gaps identified in previous section, we setup a study to explore the following research questions: **RQ1.** How well can we estimate collaboration quality using machine learning, using audio and log data from an authentic classroom setting in upper secondary school? **RQ2.** How well can we estimate the various sub-dimensions of collaboration quality with machine learning, using audio and log data from an authentic classroom setting?

To start addressing these questions, we have conducted a first case study [14] in an authentic classroom setting, where learners performed collaborative discussion and writing tasks, as part of their normal classes. Such case study methodology allowed us to understand the situation in depth, and explore multiple aspects of the research questions (e.g., data fusion and regression models).

A 30 minute collaboration activity was co-designed by a researcher and a teacher in which the students had to discuss and fill in a worksheet regarding genetic mutations[3]. The activity was enacted in a secondary education biology course with 10 students in autumn 2019. During the enactment, two researchers were present in the classroom for data collection purposes and technical support. A brief introduction was given to the students about the aim of study and their consent for data collection was taken in written form before the activity. In the case study, the data from two groups (four students each) are analyzed.

3.1 Data Collection

The students used an audio-capturing prototype -CoTrack- with an omni-directional microphone placed in the center of the group's table, and Etherpad[4] for the collaborative writing. CoTrack detects presence of voice and provides the direction

[3] Given learning activity is available at:https://bit.ly/collabtech-LD

[4] An open source real-time collaborative text editor, see https://etherpad.org

from which voice is detected. CoTrack then maps the direction to a particular learner and extracts various features (e.g., speaking time, number of characters added or deleted) from the audio and Etherpad logs. Our analyses below use a total of 12 features: three features (speaking time, number of characters added, and number of characters deleted) for each of the four students in the groups, for every 30-second window of time (see below).

Table 1: Inter-rater agreement of human coders in each collaboration quality sub-dimension (Cohen's kappa)

SMU	CF	KE	ARG	SPST	CO	ITO-1	ITO-2	ITO-3	ITO-4
0.71	0.91	0.74	0.80	0.65	0.68	0.72	0.76	0.75	0.78

3.2 Data Annotation

We used Rummel et al.'s [11] collaboration quality rating scheme (itself adapted from [9]), assigning a collaboration quality score along seven dimensions[5]. We decided to use the adaptable version instead of the original scheme due to its applicability to a variety of CSCL settings. Two raters coded the dimensions at the group level (except the ITO, which is coded at an individual level and averaged to get the group-level feature). Following the recommendations by Martínez et al. [7], we used time windows of 30 seconds, in which each of the aforementioned sub-dimensions was assigned a score between -2 (very bad) and $+2$ (very good). The sub-dimension scores at the group level were then added up to get the overall collaboration quality score of the group for that time window (which can theoretically range from -14 to $+14$). This overall score was used as dependent variable in the regression analysis. The annotation phase resulted in a dataset with 121 data points from the collaboration of two learner groups. Two raters went through four iterations of coding before reaching substantial agreement on each sub-dimension in terms of Cohen's kappa (Table 1).

3.3 Data Analysis

To map the individual student audio and log features to group-level features, we explored three different approaches[6]: simple averaging of individual scores, using dimensionality reduction, and entropy-based fusion. In the dimensionality reduction approach we applied principal component analysis (PCA) on all

[5] Sustaining Mutual Understanding (SMU), Collaboration Flow (CF), Knowledge Exchange (KE), Cooperative Orientation (CO), Argumentation (ARG), Structuring Problem Solving Process and Time Management (SPST), and Individual Task Orientation (ITO)

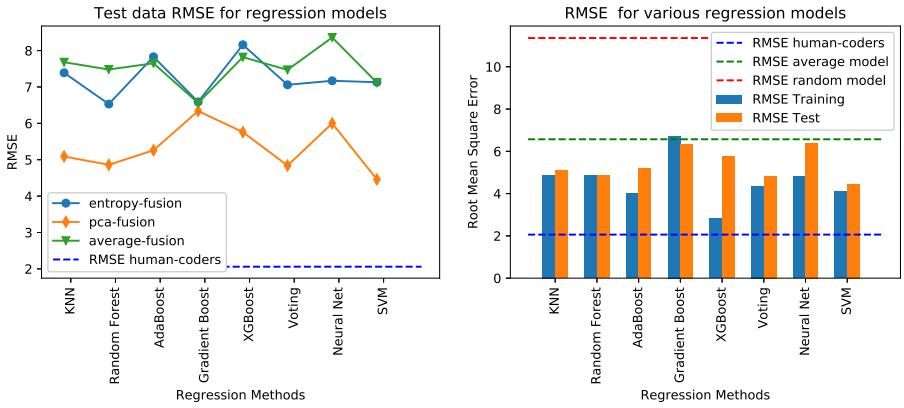
[6] Data analysis source code available at: https://github.com/pankajchejara23/collab-analysis

individual-level features and extracted the four components that explained most variance. The entropy-based approach has been used to map individual features to group-level features in previous research [1] using Shannon's Entropy.

For the training and evaluation of the machine learning models, we used Python's Scikit-learn library[7]. We randomly divided our dataset into training and test sets, using a ratio of 70:30. We trained regression models of different kinds on our training set and investigated their performance on the test set. Concretely, machine learning model families explored included K-Nearest Neighbors, Random forest, Adaboost, Gradient boost, XGboost, Support vector regressors (SVR), Neural networks, and ensemble (voting) regression models (using SVR, Random forest and Adaboost). We used GridSearchCV (from Scikit-learn) with 3-fold cross validation to tune the model's parameters.

3.4 Results

We used RMSE (Root Mean Square Error) as the performance metric to compare the different regression models. As frames of reference, we computed the RMSE that the human coders had achieved in their last round of manual collaboration quality scoring. We also computed the RMSE of two "no-information" regressors, one that just estimates random values within the range of possible quality scores, and one that provides an estimation equal to the average value of the collaboration quality (quality=1.93, for this dataset).



(a) RMSE for different Fusion

(b) RMSE of regression models

Fig. 1: Performance scores of regression models

From our analysis of the three fusion approaches, we found PCA-based fusion as a better option than entropy and average, in terms of the performance of the

[7] https://scikit-learn.org/stable/

different regression models on test data. Figure 1(a) shows that PCA-fusion based regression models achieved lower RMSE on test data than entropy-fusion and average-fusion based regression models.

For the comparative analysis among regression models, we used PCA-based fusion and trained different kinds of regression models, computing RMSE for both the training and test data. All regression models (except Gradient Boost) performed better than the average estimation model (figure 1(b)). XG Boost and Neural Network regression models reported the highest variation between training and testing errors, which can probably be explained by the models overfitting the small dataset available. The support vector regression (SVR) model performed better than the other models, both in terms of lower RMSE, and lower difference between training and test error. Comparing the performance of this SVR model (which, let's remember, used only very basic audio and log features) with that of the no-information models (average and random) and the human coders' own RMSE values, we find that SVR covered about 50% of the gap between the best no-information predictors and human-level performance.

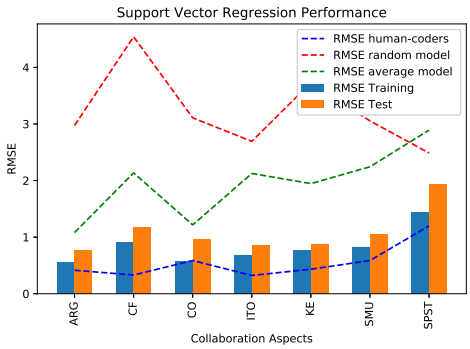


Fig. 2: SVR performance on various sub-dimensions of collaboration quality

We also applied similar regression models to estimate the seven sub-dimensions of collaboration quality. Again, support vector regression models performed better than other models in estimating the majority of the sub-dimensions. Figure 2 shows the RMSE scores of support vector regression model, compared with the no-information and human-level frames of reference. For these dimensions, SVR covered 50% or more of the gap between no-information and human-level performance.

4 Conclusions and future work

This paper investigated the feasibility of estimating the quality of collaboration in face-to-face classroom setting using simple features from audio and log

data, and machine learning regression models. These results suggest that it *is* feasible to quantitatively estimate collaboration quality along a continuum, and even open the door to more in-depth estimation of the different collaboration sub-dimensions (e.g., collaboration flow, knowledge exchange), which can be of greater value to practitioners. We also provided three frames of reference (average and random no-information estimators, as well as human coders' performance) with the aim to offer a more interpretable view on their performance. We suggest future MMLA researchers to analyze their models' performance using such frames of reference, to help our research community in better understanding how far our models and solutions have to go to achieve human-level performance .

This work is not without limitations. The small size of our dataset is probably the main weakness of our results so far, limiting greatly the generalizability of the particular models and performance claims made. This issue can explain the discrepancy between training and test errors of some of the regression models (e.g., Gradient Boost, AdaBoost, Neural Networks) due to over-fitting. The expansion of this dataset with data from groupwork performed in different kinds of authentic classroom settings, is one of our most important avenues of future work. Thus, we plan to assess the generalizability of the results in terms of effectiveness of the approach.

Moreover, in the current case study we only used simple audio and log features, and a limited set of machine learning models, considering all dataset samples independently (i.e., not looking at their sequence). The use of more complex features (e.g., intensity, pitch, MFCC for audio data, or conversion of voice to text and subsequent analyses of content), consideration of final document quality (in terms of matrices e.g., error rate, redundancy, keywords [5]), different data fusion models, impact of task duration, and the exploration of time-dependent machine learning models (e.g., Hidden Markov Models, sequence analysis) will be considered as strategies to expand our work towards automated estimation of collaboration quality that is close to human-level performance.

Acknowledgement

This research has been partially funded by the European Union via the European Regional Development Fund and in the context of CEITER and Next-Lab (Horizon 2020 Research and Innovation Programme, grant agreements no. 669074 and 731685). It was also partially funded by the European Regional Development Fund and the Regional Council of Education of Castile and León under grant VA257P18 and the National Research Agency of the Spanish Ministry of Science, Innovation and Universities, under project grant TIN2017-85179-C3-2-R.

References

1. Bassiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., D'Angelo, C., Alozie, N.: Privacy-preserving speech analytics for automatic assessment of student collaboration. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. pp. 888–892 (2016)

2. Blikstein, P., Worsley, M.: Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. Journal of Learning Analytics **3**(2), 220–238 (2016)
3. Dillenbourg, P., Järvelä, S., Fischer, F.: The evolution of research on computer-supported collaborative learning. In: Balacheff, N., Ludvigsen, S., de Jong, T., Lazonder, A., Barnes, S. (eds.) Technology-Enhanced Learning: Principles and Products, pp. 3–19. Springer Netherlands, Dordrecht (2009)
4. Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D., Divakaran, A.: Multimodal analytics to study collaborative problem solving in pair programming. In: Proceedings of the Sixth International Conference on Learning Analytics Knowledge. p. 516–517. LAK '16, ACM, NY, USA (2016)
5. Ignat, C., Oster, G., Fox, O., Shalin, V.L., Charoy, F.: How do user groups cope with delay in real-time collaborative note taking. In: Boulus-Rødje, N., Ellingsen, G., Bratteteig, T., Aanestad, M., Bjørn, P. (eds.) ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway. pp. 223–242. Springer (2015)
6. Jermann, P., Soller, A., Muehlenbrock, M.: From mirroring to guiding: A review of the state of art technology for supporting collaborative learning. International Journalof Artificial Intelligence in Education (IJAIED) **15**, 261–290 (2005)
7. Martinez, R., Wallace, J.R., Kay, J., Yacef, K.: Modelling and identifying collaborative situations in a collocated multi-display groupware setting. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) Artificial Intelligence in Education. pp. 196–204. Springer, Berlin, Heidelberg (2011)
8. Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., Yacef, K.: Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. International Journal of Computer-Supported Collaborative Learning **8**(4), 455–485 (2013)
9. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. International Journal of Computer-Supported Collaborative Learning **2**(1), 63–86 (2007)
10. Mercer, N., Littleton, K., Wegerif, R.: Methods for studying the processes of interaction and collaborative activity in computer-based educational activities. Technology, Pedagogy and Education **13**(2), 195–212 (2004)
11. Rummel, N., Deiglmayr, A., Spada, H., Kahrimanis, G., Avouris, N.: Analyzing collaborative interactions across domains and settings: an adaptable rating scheme. In: Puntambekar, S., Erkens, G., Hmelo-Silver, C. (eds.) Analyzing Interactions in CSCL: Methods, Approaches and Issues. pp. 367–390. Springer US, Boston, MA (2011)
12. Spikol, D., Ruffaldi, E., Dabisias, G., Cukurova, M.: Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. Journal of Computer Assisted Learning **34**(4), 366–377 (2018)
13. Viswanathan, S.A., VanLehn, K.: Using the tablet gestures and speech of pairs of students to classify their collaboration. IEEE Transactions on Learning Technologies **11**(2), 230–242 (2018)
14. Yin, R.K.: Case study methods. APA handbooks in psychology®., American Psychological Association, Washington, DC, US (2012)