Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Pairwise learner model for collaborative learning and its application in genetic group formation

Changhao LIANG[a*], Kensuke TAKII[a] & Hiroaki OGATA[a]
[a] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
* liang.changhao.8h@kyoto-u.ac.jp

Abstract: This paper proposes a pairwise learner model for collaborative learning and its application in genetic algorithm-based group formation. To overcome the limitations of traditional learner models in capturing complex group learning dynamics, the pairwise model treats learner pairs as the fundamental unit of analysis, quantifying pair relationships through measures such as knowledge structure similarity. Employing genetic algorithms, this study explores different group formation strategies using knowledge graph distances, adopting Wasserstein distance to measure relational disparities. The results exhibit the model's effectiveness in forming both homogeneous and heterogeneous groups while reducing cross-group deviations. It can also extend beyond group formation to support various aspects of the group learning process and its outcomes. Apart from knowledge graph data, the model has the potential to accommodate a broader range of data from different modalities in future studies.

Keywords: learner model, group formation, knowledge graph, genetic algorithm, group learning

1. Evolution of learner models

Learner modeling has evolved from relying on scores and surveys to utilizing digital behavior logs and, more recently, capturing multimodal real-world learning dynamics (Giannakos & Cukurova, 2023). Traditional models focused on paper-based tests and self-reported surveys, using descriptive statistics such as averages and variances to represent learning performance. With the rise of digital platforms, indicator-based models, captured as triplets of ("learner, indicator, value"), have gained prominence, including metrics such as reading time, the number of annotations, and behavior sequences. Such models provide a deeper understanding of learning engagement and support adaptive learning through dashboards that inform material recommendations and performance predictions (Chen et al., 2021).

However, such models often fall short of capturing the complex, contextual nature of collaborative learning. Multimodal Learning Analytics (MMLA) introduces rich data types, such as trajectories and proximity (Blikstein & Worsley, 2016), but these are challenging to reduce to single, interpretable indicators at the individual level. To address this, we propose a pairwise learner model, which characterizes the relationship between learner pairs rather than individuals in isolation, viewing pairs as the fundamental unit of analysis. Drawing on transactional-level analysis within the multimodal analytics schema (Oviatt, 2018), pairwise features allow the use of distance or similarity measures to generate lower-dimensional, pedagogically interpretable metrics.

This paper examines the potential of the pairwise model for facilitating group formation, a key challenge in orchestrating collaborative learning. We present a pilot implementation and evaluation of this model for optimizing group composition using knowledge graph distance.

2. Data-driven genetic group formation with pairwise knowledge graph distance

Genetic algorithms offer a flexible solution for group formation (Moreno et al., 2012), allowing for multiple input variables, balanced group sizes, and the inclusion of all students. The intra-

group similarity is quantified by a fitness value (F_g) of group g, where lower values indicate homogeneous groups and higher values reflect heterogeneity. In practice, grouping is often guided by learners' knowledge proficiency. Following Flanagan et al. (2021), mathematical domains with structured nodes allow for squared difference measures across **independent** knowledge components C within a cohort of students S, as shown in Equation (1):

$$F_g = \sum_{s=1}^{S} \sum_{j=1}^{C} \left(c_{j,s} - \overline{x}_{j,g}\right)^2 \quad (1)$$

However, in domains such as language learning, where knowledge is semantically complex and interrelated, measures of isolated proficiency can reduce interpretability. To address this, we employ a pairwise learner model that captures the relational knowledge gap between learner pairs using the Wasserstein Distance (Panaretos & Zemel, 2019). This approach accounts for both node proficiency and structural differences in knowledge graphs by quantifying the effort required to transform one distribution into another.

This model represents the learner model as quadruplets ("learner 1, learner 2, indicator, value"), enabling finer-grained analysis of knowledge imbalance. F_g is computed as the average pairwise distance P within each group g (Equation 2). To avoid high variation across groups, a standard deviation penalty is introduced (Equation 3) when determining the final F for homogeneous (F_{homo}) or heterogeneous (F_{hetero}) grouping (Liang et al., 2024) from all F_g, inspired by Konert et al. (2016) and Hui et al. (2025).

$$F_g = \frac{1}{|g|(|g|-1)} \sum_{(i,j)\in g} P(i,j) \quad (2)$$

$$F_{homo} = \frac{\left(1+\sigma_{F_g}\right)}{|F_g|} \sum F_g \, , F_{hetero} = \frac{1}{\left(1+\sigma_{F_g}\right)|F_g|} \sum F_g \quad (3)$$

We conducted four group formation trials using knowledge graph data from 32 learners in a Japanese university academic reading course in Spring 2024. The knowledge graph, constructed via word co-occurrence in OKLM (Takii et al., 2024), comprised 208 vocabulary nodes. Node-level proficiency was inferred from annotation behaviors such as highlighting and difficulty marking. Pairwise distances followed a normal distribution (Shapiro-Wilk $p < .001$).

Figure 1 compares fitness results across grouping strategies. Without coordination, heterogeneous and homogeneous groups showed significant differences ($t = 2.305$, $p = .037$, see the right two boxes), but large cross-group deviations were observed. With coordination under Equation (3), deviations were reduced, and a bigger statistical difference appeared ($t = 3.203$, $p = .006$, see the left two boxes), though the difference in mean values got less prominent than the original algorithm.

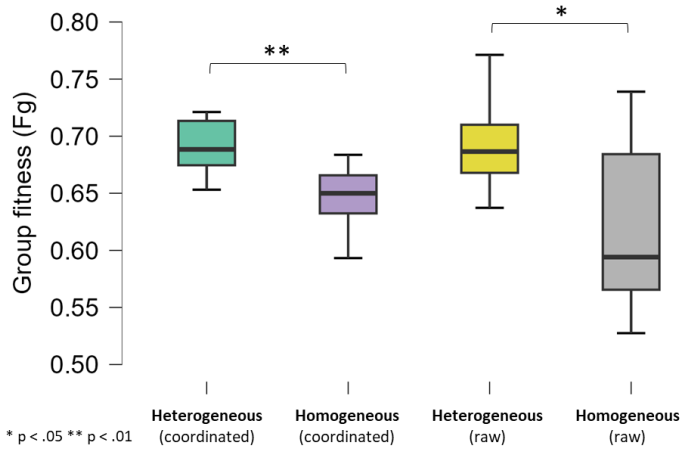

Figure 1. Group fitness results under different group formation strategies.

While results may vary with different settings, this pilot demonstrates that genetic algorithms using pairwise indicators can effectively optimize group composition. Despite

analyzing one run per condition in Figure 1, repeated trials showed stable outcomes (F_g variations within ±0.01). Further testing across more rounds and varying parameters for group and population sizes is necessary to assess the robustness and scalability of the algorithm.

3. Implication and Conclusion

This study introduced the concept of pairwise indicators in collaborative learning and demonstrated their application in knowledge graph-based group formation. According to Janssen & Kirschner (2020), beyond their role in group formation as antecedents, these indicators can also support various aspects of the group learning process and outcomes, such as group awareness dashboards, adaptive agents, and real-time learning interventions.

While this study initially illustrates pairwise indicators using knowledge graph data, it focuses on pairs as fundamental analytical units in collaborative learning. Since pairs serve this foundational role, more complex data sources such as multimodal interactions captured during group orchestration can also be systematically mapped to each pair with quantified values. These include textual artifact analysis, synchronized eye-tracking data, and virtual space proximity metrics.

Acknowledgements

This work was partially supported by JST CSTI SIP Program Grant Number JPJ012347 and JSPS KAKENHI Grant Number 25K21357.

References

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of learning analytics*, *3*(2), 220-238.

Chen, C. H., Yang, S. J., Weng, J. X., Ogata, H., & Su, C. Y. (2021). Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. *Australasian Journal of Educational Technology*, *37*(4), 130-144.

Flanagan, B., Liang, C., Majumdar, R., and Ogata, H. (2021). Towards explainable group formation by knowledge map-based genetic algorithm. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp.370–372).

Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology*, *54*(5), 1246-1267.

Hui, B., Adeyemi, O., Phan, K., Schoenit, J., Akins, S., & Khademi, K. (2025, March). Diversity Considerations in Team Formation Design, Algorithm, and Measurement. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 36-46).

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research and Development*, *68*(2), 783-805.

Konert, J., Burlak, D., & Steinmetz, R. (2014). The group formation problem: an algorithmic approach to learning group formation. In *EC-TEL 2014,* Proceedings *9* (pp. 221-234).

Liang, C., Horikoshi, I., & Ogata, H. (2024). Enabling Mixed Genetic Algorithm for Automatic Group Formation System. In *International Conference on Collaboration Technologies and Social Computing* (pp. 220-228).

Moreno, J., Ovalle, D. A., & Vicari, R. M. (2012). A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. *Computers & Education*, *58*(1), 560-569.

Oviatt, S. (2018, October). Ten opportunities and challenges for advancing student-centered multimodal learning analytics. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 87-94).

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, *6*(1), 405-431.

Takii K., Liang C., & Ogata H. (2024). Open Knowledge and Learner Model: Mathematical Representation and Applications as Learning Support Foundation in EFL. *In 32nd International Conference on Computers in Education* (Vol. 1, pp. 595-604).