The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Maximizing accuracy of group peer assessment using item response theory and integer programming

Masaki Uto[*1] Duc-Thien Nguyen[*1] Maomi Ueno[*1]

[*1] University of Electro-Communications

With the wide spread of large-scale e-learning environments, peer assessment has been widely used to measure learner ability. When the number of learners increases, peer assessment is often conducted by dividing learners into multiple groups. However, in such cases, the peer assessment accuracy depends on the method of forming groups. To resolve that difficulty, this study proposes a group formation method to maximize peer assessment accuracy using item response theory and integer programming. Experimental results, however, have demonstrated that the method does not present sufficiently higher accuracy than a random group formation method does. Therefore, this study further proposes an external rater assignment method that assigns a few outside-group raters to each learner after groups are formed using the proposed group formation method. Through results of simulation and actual data experiments, this study demonstrates that the method can substantially improve peer assessment accuracy.

1. Introduction

As an assessment method based on a social constructivist approach, peer assessment, which is mutual assessment among learners, has become popular in recent years. One common use of peer assessment is for summative assessment. The importance of this usage has been increasing concomitantly with the wider use of large-scale e-learning environments [Suen 14, Shah 14]. Peer assessment, however, entails the difficulty that the assessment accuracy of learner ability depends on rater characteristics such as severity and consistency. To resolve that difficulty, item response theory (IRT) models incorporating rater parameters have been proposed [Eckes 11, Uto 18]. The IRT models are known to provide more accurate ability assessment than average or total scores do because they can estimate the ability considering rater characteristics [Uto 16, Uto 18].

In learning contexts, peer assessment has often been adopted for group learning situations such as collaborative learning and active learning [Staubitz 16, Suen 14, Nguyen 15]. Specifically, learners are divided into multiple groups in which they work together, and peer assessment is conducted within the groups. However, in such cases, the ability assessment accuracy depends also on a way to form groups. For example, if a group consists of learners who tend to assess others randomly, their abilities are difficult to be estimated accurately. Therefore, group optimization is important to maximize the accuracy of peer assessment. However, no studies have focused on this issue.

For the reason, this study proposes a new group formation method that maximizes peer assessment accuracy based on IRT. Specifically, the method is formulated as an integer programming (IP) problem that maximizes the lower bound of the Fisher information (FI) measure: a widely used index of ability assessment accuracy in IRT. The method is expected to improve the ability assessment accuracy because groups are formed so that the learners in the same group can assess one another accurately. However, experimental results demonstrated that the method did not present sufficiently higher accuracy than that of a random group formation method. The result suggests that it is generally difficult to assign raters with high FI to all learners when peer assessment is conducted only within groups.

To alleviate that shortcoming, this study further proposes an external rater assignment method that assigns a few optimal outside-group raters to each learner after forming groups using the method presented above. We formulate the method as an IP problem that maximizes the lower bound of the FI for each learner given by assigned outside-group raters. Simulations and actual data experiments demonstrate that assigning a few optimal external raters using the proposed method can improve the peer assessment accuracy considerably.

Contact: Masaki Uto, 1-5-1, Choufugaoka, Choufu-shi, Tokyo, Japan, 042-443-5627, uto@ai.lab.uec.ac.jp

2. Peer assessment data

This study assumes that peer assessment data $\boldsymbol{U}$ consists of rating categories $k \in \mathcal{K} = \{1, \cdots, K\}$ given by each peer-rater $r \in \mathcal{J} = \{1, \cdots, J\}$ to each learning outcome of learner $j \in \mathcal{J}$ for each task $t \in \mathcal{T} = \{1, \cdots, T\}$. Letting u_{tjr} be a response of rater r to learner j's outcome for task t, the data $\boldsymbol{U}$ are described as $\boldsymbol{U} = \{u_{tjr} \mid u_{tjr} \in \mathcal{K} \cup \{-1\}, t \in \mathcal{T}, j \in \mathcal{J}, r \in \mathcal{J}\}$, where $u_{tjr} = -1$ denotes missing data.

Furthermore, this study assumes that peer assessment is conducted by dividing learners into multiple groups for each task $t \in \mathcal{T}$. Here, let x_{tgjr} be a dummy variable that takes 1 if learner j and peer r are included in the same group $g \in \mathcal{G} = \{1, \cdots, G\}$ for task t, and which takes 0 otherwise. Then peer assessment groups for task t can be described as $\boldsymbol{X}_t = \{x_{tgjr} \mid x_{tgjr} \in \{0, 1\}, g \in \mathcal{G}, j \in \mathcal{J}, r \in \mathcal{J}\}$. Consequently, when peer assessment is conducted among group members, the rating data u_{tjr} become missing data if learners j and r are not in the same group ($\sum_{g=1}^{G} x_{tgjr} = 0$).

The purpose of this study is to estimate the learner ability accurately using IRT for peer assessment [Uto 16] from the data $\boldsymbol{U}$ by optimizing the groups $\boldsymbol{X} = \{\boldsymbol{X}_t \mid t \in \mathcal{T}\}$.

3. IRT for peer assessment

The IRT for peer assessment [Uto 16] has been formulated as a graded response model that incorporates rater parameters. The model defines the probability that rater r responds in category k to learner j's outcome for task t as

$$P_{tjrk} = P^*_{tjrk-1} - P^*_{tjrk}, \quad (1)$$
$$P^*_{tjrk} = [1 + \exp(-\alpha_t \gamma_r (\theta_j - \beta_{tk} - \varepsilon_r))]^{-1}$$

Here, θ_j denotes the ability of learner j; γ_r reflects the consistency of rater r; ε_r represents the severity of rater r; α_t is a discrimination parameter of task t; and β_{tk} denotes the difficulty in obtaining category k for task t ($\beta_{t1} < \cdots < \beta_{tK-1}$); $P^*_{tjr0} = 1$, and $P^*_{tjrK} = 0$.

In IRT, the standard error estimate of ability assessment is defined as the inverse square root of the FI. More information implies less error of the assessment. Therefore, FI can be regarded as an index of the ability assessment accuracy. For the above model, FI of rater r in task t for a learner with ability θ_j is calculable as

$$I_{tr}(\theta_j) = \alpha_t^2 \gamma_r^2 \sum_{k=1}^{K} \frac{\left(P^*_{tjrk-1} Q^*_{tjrk-1} - P^*_{tjrk} Q^*_{tjrk}\right)^2}{P^*_{tjrk-1} - P^*_{tjrk}}, \quad (2)$$

where $Q^*_{tjrk} = 1 - P^*_{tjrk}$.

The FI of multiple raters for learner j in task t is definable by the sum of the information of each rater. Therefore, when peer assessment is conducted within group members, the FI for learner j in task t is calculable as shown below.

$$I_t(\theta_j) = \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \quad (3)$$

A high value of FI $I_t(\theta_j)$ signifies that the group members can assess learner j accurately. Therefore, if we form groups to provide great amounts of FI for each learner, then the ability assessment accuracy can be maximized.

4. Group formation method

Based on this idea presented above, we formulate the group formation optimization method (designated as *PropG*) as an IP problem that maximizes the lower bound of FI for each learner. Specifically, *PropG* for task t is formulated as the following IP problem.

$$\begin{aligned}
&\text{maximize} && y_t && && (4)\\
&\text{subject to} && \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \geq y_t, && \forall j, && (5)\\
& && \sum_{g=1}^{G} x_{tgjj} = 1, && \forall j, && (6)\\
& && n_l \leq \sum_{j=1}^{J} x_{tgjj} \leq n_u, && \forall g, && (7)\\
& && x_{tgjr} = x_{tgrj}, && \forall g, j, r && (8)
\end{aligned}$$

The first constraint requires that FI for each learner j be larger than a lower bound y_t. The second constraint restricts each learner as belonging to one group. The third constraint controls the number of learners in a group. Here, n_l and n_u represent the lower and upper bounds of the number of learners in group g. In this study, $n_l = \lfloor J/G \rfloor$ and $n_u = \lceil J/G \rceil$ are used so that the numbers of learners in respective groups become as equal as possible. This IP maximizes the lower bound of FI for learners. Therefore, by solving the problem, one can obtain groups that provide as much FI as possible to each learner.

4.1 Evaluation of group formation methods

To evaluate the effectiveness of *PropG*, we conducted the following simulation experiment. 1) For $J = 30$ and $T = 5$, the true IRT model parameters were generated randomly. 2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG* and a random group formation method (designated as *RndG*). For *PropG*, the FI values were calculated using the true parameter values. 3) Given the created groups and the true model parameters, peer assessment data were sampled randomly for the current task t based on the IRT model. 4) Given the true rater and task parameters, the learner ability was estimated from the data generated to date. 5) RMSE between the estimated ability and the true ability were calculated. 6) Procedures 2) – 5) were repeated for the remaining tasks. 7) After 10 repetitions of the procedures described above, the average values of RMSE were calculated.

Fig. 1 presents the results. Results demonstrate that RMSE decreases with the decreasing number of groups G or with increasing numbers of tasks or learners because the number of data for each learner increases. Generally, the increase of data per learner is known to engender improvement of the ability assessment accuracy [Uto 16]. Comparing the group formation methods, however, *PropG* does not decrease RMSE sufficiently. The results indicate that it is difficult to form groups to sufficiently increase the peer assessment accuracy. To overcome this shortcoming, we further propose the assignment of outside-group raters to each learner, given the groups created using *PropG*.

5. External rater assignment

The proposed external rater assignment method (designated as *PropE*) is formulated as an IP problem that maximizes the lower bound of information for learners given by the assigned outside-group raters. Specifically, given a group formation $\boldsymbol{X}_t$, *PropE* for task t is defined as folllows.

$$\begin{aligned}
&\text{maximize :} && y'_t && && (9)\\
&\text{subject to :} && \sum_{r \in \boldsymbol{C}_{tj}} I_{tr}(\theta_j) z_{tjr} \geq y'_t, && \forall j && (10)\\
& && \sum_{r \in \boldsymbol{C}_{tj}} z_{tjr} = n^e, && \forall j && (11)\\
& && \sum_{j=1}^{J} z_{tjr} \leq n^J, && \forall r && (12)\\
& && z_{tjj} = 0, && \forall j && (13)
\end{aligned}$$

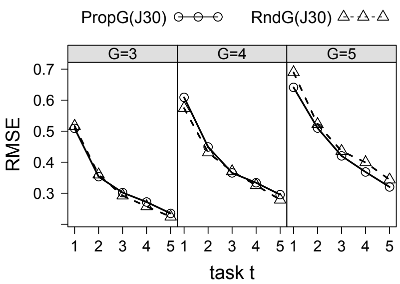

Figure 1: RMSE values of group formation methods in the simulation experiment.

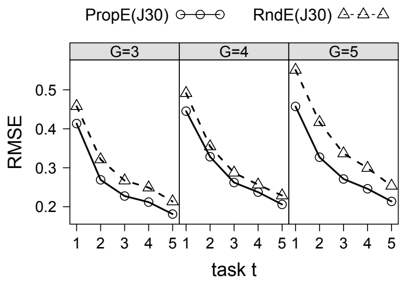

Figure 2: RMSE values of external rater assignment methods for each G and t in the simulation experiment.

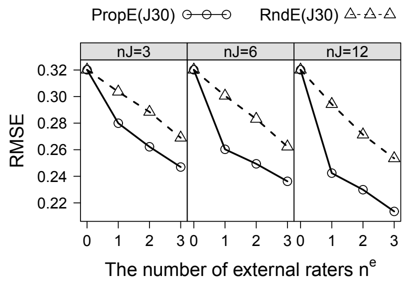

Figure 3: RMSE values of external rater assignment methods for each n^J and n^e in the simulation experiment.

Here, $\boldsymbol{C}_{tj} = \{r \mid \sum_{g=1}^{G} x_{tgjr} = 0\}$ is the set of outside-group raters for learner j in task t given a group formation $\boldsymbol{X}_t$. In addition, z_{tjr} is a variable that takes 1 if external rater r is assigned to learner j in task t; it takes 0 otherwise. Furthermore, n^e denotes the number of external raters assigned to each learner; n^J is the upper limit number of outside-group learners assignable to each rater. Here, n^e and n^J must satisfy $n^J \geq n^e$. The increase of n^J makes it easier to assign optimal raters to each learner, although differences in the workload among the learners increases.

The first constraint in the IP restricts that the FI for each learner given by the assigned outside-group raters must exceed a lower bound y'_t. The second constraint requires that n^e number of outside-group raters must be assigned to each learner. The third constraint restricts that each learner can assess at most n^J number of outside-group learners. The objective function is defined as the maximization of the lower bound of the FI for learners given by assigned external raters. Therefore, by solving the proposed method, an external rater assignment z_{tjr} is obtainable so that n^e outside-group raters with high FI are assigned to each learner.

5.1 Evaluation of external rater assignment

To evaluate the performance of the proposed method, we conducted the following simulation experiment, which is similar to that conducted in 4.1. 1) For $J = 30$ and $T = 5$, the true model parameters were generated randomly. 2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG*. Then, given the created groups, $n^e \in \{1, 2, 3\}$ outside-group raters were assigned to each learner using *PropE* and a random assignment method (designated as *RndE*). Here, we changed the value of n^J for $\{3, 6, 12\}$ to evaluate its effects. In *PropG* and *PropE*, FI was calculated using the true parameter values. 3) Peer assessment data were sampled randomly for current task t following the IRT model, given the true model parameters, the formed groups and the rater assignment. 4) The following procedures were identical to procedures 4) – 7) of the previous experiment.

Fig. 2 shows the RMSE for each t and G when $n^J = 12$ and $n^e = 3$, and Fig. 3 shows the RMSE for each n^e and n^J when $G = 5$ and $t = 5$. In Fig. 3, the results for $n^e = 0$ correspond to *PropG*. Results show that the accuracy of the external rater assignment methods tends to increase concomitantly with decreasing number of groups and increasing number of tasks and assigned external raters n^e because the number of rating data for each learner increases. Furthermore, Fig. 3 shows that both external rater assignment methods reveal the lower RMSE than *PropG* in all cases, which suggests that the addition of the external raters is effective to improve the ability assessment accuracy. Comparison of the external rater assignment methods reveals that *PropE* presented higher accuracy than *RndE* in all cases. Furthermore, the RMSE difference between *PropE* and *RndE* tends to increase with increasing n^J value because the increase of n^J makes it easier to assign optimal raters to each learner.

From those results, we infer that the proposed method can improve the peer assessment accuracy efficiently when a large value of n^J and a small value of n^e are given.

6. Usage in actual e-learning situations

PropG and *PropE* require IRT parameter estimates to calculate FI. Although the experiments described above used the true parameter values, they are practically unknown. Therefore, this section presents a description of how to use *PropG* and *PropE* when the IRT parameters are unknown in actual e-learning situations. We consider the following two assumptions for using *PropG* and *PropE* in an e-learning course. 1) More than one task is offered in the course. 2) All tasks were used in past e-learning courses at least once, and past learners' peer assessment data corresponding to the tasks were collected. Although the second assumption might not necessarily be satisfied in practice, it is necessary to estimate the task parameters.

Under the second assumption, we can estimate the task parameters. Given task parameter estimates, we can use *PropG* and *PropE* through the following procedures under the first assumption. 1) For the first task, peer assessment is conducted using randomly formed groups. 2) The rater parameters and learner ability are estimated from the obtained peer assessment data. 3) For the next task, group formation and external rater assignment are conducted using *PropG* and *PropE* given the parameter estimates. 4) Repeat procedures 2) and 3) for remaining tasks.

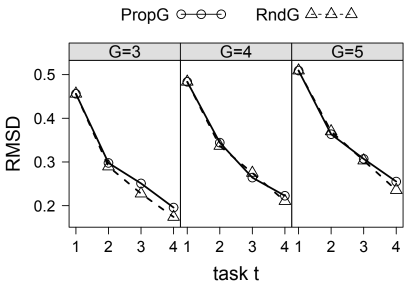

Figure 4: RMSDs of group formation methods in the actual data experiment.

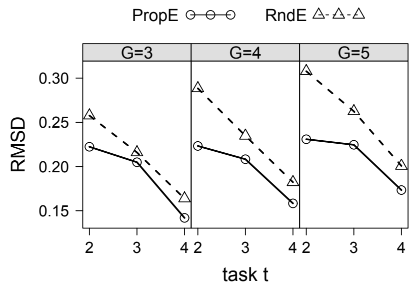

Figure 5: RMSDs of external rater assignment methods for each G and t in the actual data experiment.

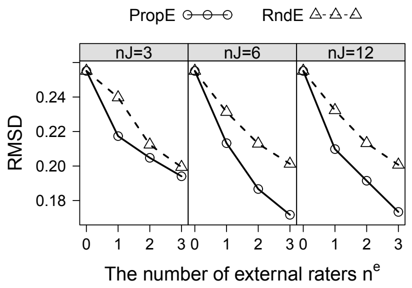

Figure 6: RMSDs of external rater assignment methods for each n^J and n^e in the actual data experiment.

7. Actual data experiment

This section evaluates the effectiveness of *PropG* and *PropE* using actual peer assessment data based on the above usage. We gathered actual data using the following procedures. 1) As subjects for this study, 34 university students were recruited. 2) They were asked to complete four essay writing tasks offered in NAEP. 3) After the participants completed all tasks, they were asked to evaluate the essays of all other participants for all four tasks using a rubric with five rating categories. Furthermore, we collected additional rating data (designated as *five raters' data*) for task parameter estimation. The data consist of ratings assigned by 5 graduate school students to the essays gathered in the experiment above.

Using the actual data, we conducted the following experiments. 1) The task parameters in the IRT model were estimated using the five raters' data. 2) Given the task parameter estimates, the rater parameters and learner ability were estimated using the full peer assessment data. 3) For the first task, $G \in \{3, 4, 5\}$ groups were created randomly. 4) The peer assessment data without peer-rater assignment were changed to missing data. 5) From the peer assessment data up to the current task, the rater parameters and learner ability were estimated given the task parameters estimated in Procedure 1). 6) RMSD between the ability estimates and that estimated from the complete data in Procedure 2) was calculated. 7) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Then, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI. 8) For the remaining tasks, procedures 4) – 7) were repeated. 9) After repeating the procedures described above 10 times, the average values of the RMSD were calculated.

Fig. 4 presents results of each group formation method. Figs. 5 and 6 show those of the external rater assignment methods. Fig. 5 presents results for each $t \geq 2$ and $G \in \{3, 4, 5\}$ when $n^J = 12$ and $n^e = 3$. Fig. 6 shows those for each n^e and n^J when $G = 5$ and $t = 4$. Results show similar tendencies to those obtained from the simulation experiments. Specifically, comparing the group formation methods, *PropG* does not improve the accuracy much, while the assessment accuracy is improved drastically by introducing external raters. Furthermore, the proposed external rater assignment method realizes the higher accuracy than the random assignment method when n^J is large and n^e is small.

8. Conclusion

This study proposed the group formation method and external rater assignment method to improve peer assessment accuracy using IRT and IP. The experimentally obtained results showed that the external rater assignment method, which assigns a few optimal outside-group raters to each learner, improved the accuracy dynamically, although the proposed group formation method did not improve the accuracy sufficiently.

References

[Eckes 11] Eckes, T.: *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, Peter Lang (2011)

[Nguyen 15] Nguyen, T., Uto, M., Abe, Y., and Ueno, M.: Reliable Peer Assessment for Team-project-based Learning using Item Response Theory, pp. 144–153 (2015)

[Shah 14] Shah, N. B., Bradley, J., Balakrishnan, S., Parekh, A., Ramchandran, K., and Wainwright, M. J.: Some Scaling Laws for MOOC Assessments, in *KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)* (2014)

[Staubitz 16] Staubitz, T., Petrick, D., Bauer, M., Renz, J., and Meinel, C.: Improving the peer assessment experience on MOOC platforms, in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pp. 389–398ACM (2016)

[Suen 14] Suen, H. K.: Peer assessment for massive open online courses (MOOCs), *The International Review of Research in Open and Distributed Learning*, Vol. 15, No. 3, pp. 312–327 (2014)

[Uto 16] Uto, M. and Ueno, M.: Item Response Theory for Peer Assessment, Vol. 9, No. 2, pp. 157–170 (2016)

[Uto 18] Uto, M. and Ueno, M.: Empirical comparison of item response theory models with rater's parameters, *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32 (2018)