# A Human-Machine Hybrid Peer Grading Framework for SPOCs

Yong Han
hanyong@nlsde.buaa.edu.cn

Wenjun Wu
wwj@nlsde.buaa.edu.cn

Suozhao Ji
jisuozhao@nlsde.buaa.edu.cn

Lijun Zhang
ljzhang@nlsde.buaa.edu.cn

Hui Zhang
hzhang@nlsde.buaa.edu.cn

State Key Laboratory of Software Development Environment,
School of Computer Science, Beihang University, China

## ABSTRACT

Peer-grading is commonly adopted by instructors as an effective assessment method for MOOCs (Massive Open Online Courses) and SPOCs(Small Private online course). For solving the problems brought by varied skill levels and attitudes of online students, statistical models have been proposed to improve the fairness and accuracy of peer-grading. However, these models fail to deliver accurate inference in the SPOCs scenario because affinity among students may seriously affect the objectivity and reliability of students in the peer-assessment process. To address this problem, this paper proposes a human-machine hybrid peer-grading framework, including an automatic grader to ensure reasonable peer grades before the Bayesian models are utilized to infer the true scores. This framework can significantly eliminate the severely biased grades by those undutiful students, and thus improve the accuracy of the true-score estimation in the Bayesian peer-grading models. Both simulated and real peer-grading datasets in our experiments demonstrate the effectiveness of this new framework for SPOCs.

## Keywords

peer grading, human-machine hybrid algorithm, Bayesian model, auto-grader, SPOCs

## 1. INTRODUCTION

SPOCs is a version of MOOCs used locally with on-campus students. Despite the difference between SPOCs and MOOCs that SPOCs has the relatively smaller number of students than a MOOCs course [8], a SPOC course needs the same peer-grading process as a MOOC course when the instructor has to evaluate hundreds of open-ended essays and exercises such as mathematical proofs and engineering design problems within a deadline.

Previous research efforts on peer-grading suggest that there is a great disparity between the observed scores presented by student graders and the true scores given by the instructor. This is because students sometimes can't perform grading tasks as a professional instructor with the right skill and dedication. In the process of peer grading of SPOCs, every student grader needs to submit his answer to the problems of home assignments, and evaluate other peer's submissions according to the rubrics provided by the course instructor. The previous models [7][6] mainly adopt a Bayesian-based approach by considering the major factors affecting the aggregation of peer graded scores including the bias and reliability of every student grader.

These peer grading algorithms mostly designed for MOOCs courses may have poor performance in the setting of SPOC courses because they ignore another important factor – student attitude toward their grading tasks. Due to affinity among students in a SPOC course, they trend to assign random scores to other submissions without seriously evaluating their peers' homework. Even worse, in our real experiment, we found that some students simply give a full score to every submission assigned to them. Therefore, such an undutiful grading behavior violates the basic assumption in those Bayesian statistical models and unavoidably generate data noises that severely degrade the performance of the models. Our simulation and real experiment confirm that the models produce inaccurate estimations for final scores in the process of per grading [3].

To address the problem, this paper proposes a novel human-machine hybrid framework that combines assessment effort of both human and machine for peer-grading. The framework adopts a document classifier as an auto-grader that evaluates students' submissions to estimates their scores, and compares the scores with the peer-graded scores. Then, it attempts to filter out the unreasonable peer-graded scores that are significantly different from its estimations, and retain these legitimate scores for the statistical models. In this way, it can alleviate the negative impact of student ran-

Yong Han, Wenjun Wu, Suozhao Ji, Lijun Zhang and Hui Zhang "A Human-Machine Hybrid Peer Grading Framework for SPOCs" In: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 300 - 305

dom grading behavior and improve the overall performance of peer-grading models. Experimental results on the actual and simulation datasets demonstrate that our hybrid framework outperform the original peer-grading models in terms of the true-score estimation accuracy without placing too much extra workload on course teaching assistants (TAs).

The rest of paper is organized as follows: Section 2 discusses the related work of our research. Section 3 elaborates the main problems of current models in the peer grading of our SPOCs and explains the motivation of combining the machine and the human effort in peer grading. Section 4 describes the design of the human-machine hybrid framework for peer-grading in detail. Section 5 presents our experimental results.

## 2. RELATED WORK

The focus of this paper is to combine the power of human graders and a machine grader to improve the predictive ability of the existing peer grading models. Numerous papers have been published on the field of peer-grading research. Most researchers attempt to tackle with the peer-grading problems from the two aspects: statistical methods for accurately inferring true scores and incentive mechanism to motivate and regulate student grading behaviors.

One of the major research topics in peer-grading is to build a Bayesian statistical model that can accurately infer the true-scores of student submissions. Such models were proposed in [7] and [6] for peer-grading in MOOC courses with bias and reliability of student graders as the major latent factors. In [10], Ueno utilize Item Response Theory to model the score estimation, difficulty of problem and a grader's capability as parameters in the IRT equation. The major limitation of these models is caused by their assumption that every student follows a statistical model in the peer-grading process. But in practice, especially in the scenario of SPOCs, students' grading behavior actually are heavily affected by their motivation and attitude towards peer-grading tasks. Some students grade homework in a dutiful manner whilst others simply assign scores randomly. Thus, a single statistical model cannot describe all the possible grading strategies among these students in a SPOC course.

The problem of student grading behavior has received attention from academic researchers in the field of game theory. Recently, peer prediction mechanism has been proposed to incentivize truthful reports from individual students in the process of peer-grading [3][1]. Without the ground truth scores for every submission to verify against, designers of peer prediction mechanism often introduce comparison algorithms that compare grading results among multiple student graders and enforce penalties on those whose evaluation outcomes are different from their peers. But peer-prediction has its inherent limitation because there are potentially multiple Nash equilibria where students might be able to coordinate to avoid penalty without revealing their informative signal truthfully. Even when the peer-prediction mechanisms do offer a truthfully equilibrium, they also always induce other uninformative equilibria [2]. In the settings of SPOCs, affinity among students make it highly possible for them to collude in the peer-grading process to cheat the peer-prediction mechanisms.

Our human-machine hybrid framework is complementary to the research efforts on the statistical peer-grading models and spot-checking mechanisms of peer-prediction. The auto-grader in our framework can help to eliminate unreliable assignment grades so as to ensure only quality grades are passed onto the statistical models such as the PG family model. In this way, the auto-grader can be adopted in spot-checking mechanisms and work as an online supervisor to perform checking tasks on behalf of TAs and update TAs with its screening results.

The development of reliable auto-grader is widely regarded as a challenging task. Many researchers such as [9][5] designed neural network-based auto-graders to evaluate open essays. The state-of-art automatic graders can't complete grading tasks in a full autonomous way, especially for science essays and technical reports in domain-oriented courses. Thus, our framework only assumes an automatic grader with limited classification capability and regard it as an intelligent assistant that can work with course instructors and TAs in the process of peer-grading.

## 3. PROBLEM ANALYSIS OF PEER GRADING MODELS

In the section, we first introduce the peer grading (PG) models, then discuss the problems of the PG models when they are applied in the SPOC settings. Through the simulation experiment, we analyze fault tolerance of the PG models with the increase in the number of undutiful students.

### 3.1 Peer Grading Models

We apply the PG models [7][6][4] in the SPOCs scenario, which are Bayesian graph models with the latent factors including the biases and reliabilities of the peer graders. These models of Eq (1)(2)(3)(4) define $z_u^v$ the observations grade which is affected by the latent factors including $b_v$, $\tau_v$, and the learner's true grade $s_u$. The parameter $\rho$ denotes factors that affect the reliability, and the remaining parameters $\beta$, $\eta$, $\mu$, $\gamma$, $\lambda$ in Eq (1)(2)(3)(4) are hyper-parameters.

$$\tau_v \sim \mathcal{N}(\rho, 1/\beta_0) \quad (1)$$

$$b_v \sim \mathcal{N}(0, 1/\eta_0) \quad (2)$$

$$s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0) \quad (3)$$

$$z_u^v \sim \mathcal{N}(s_u + b_v, \lambda/\tau_v) \quad (4)$$

### 3.2 Limitations of the PG models in SPOCs

There are two major factors that may prohibit SPOCs students from performing peer-grading tasks in a fair and accurate way. First, students without the right knowledge and dedication may regard peer-grading tasks as unnecessary burdens and decide to give the assignments random scores. Second, affinity among SPOCs students who often interact with each other in the same campus or even classroom may drive them to assign higher grades to her or his peers' submissions. Both factors can result in high deviation between the observation grades $z_u^v$ and the ground-truth grade. We run the simulation experiment to evaluate the impact of student's attitude of peer assessment and analyze the tolerance of the PG models against data error generated by student graders. Based on the configuration of the simulation, we extend the PG models as follows: Assume that each student becomes a dutiful or an undutiful students with a certain

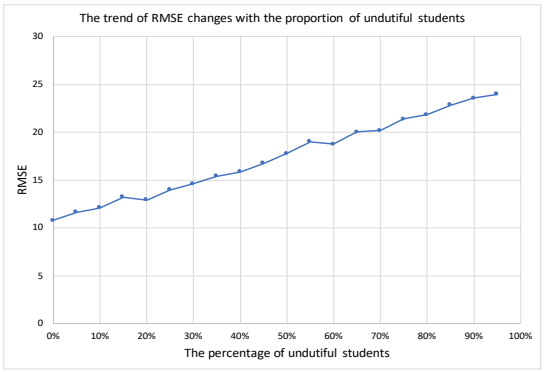


**Figure 1: The correlation between the RMSE of the PG model and the proportion of undutiful students in the simulation.**

probability each time they review. Define the number of students with undutiful grading attitude as $p \in [1, n]$. Although the value of p can also be 0, here we define the value of p starting from 1 for the convenience of calculation, and the $n$ is the total number of all students. Define a grading strategy set D where every element $d_i \in D$ denotes a particular distribution corresponding to the strategy to follow. The set $D$ contains the distributions (5) and (6):

$$z_u^v \sim \mathcal{N}(s_u + b_v, \lambda/\tau_v) \quad (5)$$

$$z_u^v = x \pm random(y) \quad (6)$$

The Eq (5) represents the strategy distribution in which the observed scores are presented by the good students with dutiful grading attitude, and the Eq (6) represents the other strategy distribution in which the observed scores are presented by the undutiful students with high deviation. In Eq (6), $x$ is the set to the average grading scores based on experiences and $y$ is set to an random value with the range $[0, 20]$. For simplicity, we assume that a student determine his/her choice of the grading strategy before he accepts the grading task and will not change in the middle of the grading process.

Figure 1 shows that the RMSE of the prediction grades has a linear correlation with the proportion of undutiful peer graders and its value ranges in $[10, 25]$. This result remains even when we change the parameters $(x, y)$ in the Eq (6). The expression of RMSE can be defined in Eq (7), where $X_{model,k}$ denotes the specific prediction grade prediction generated by the PG models for an exercise report k, and $X_{true,k}$ denotes the corresponding ground truth score of the exercise report k.

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}(X_{model,k} - X_{true,k})^2}{n}} \quad (7)$$

We can expand the Eq (7) by separating the errors generated by the dutiful group and undutiful group. First we define $e_k = X_{model,k} - X_{true,k} (k \in [1, n])$, then we define $p\bar{e} = \sum_{i=1}^{p} e_i (p \in [1, n])$ denotes the sum of the set $A = \{e_i | i \in [1, p]\}$ and $(n-p)\bar{f} = \sum_{j=p+1}^{n} e_j$ denotes the sum of the set $B = \{e_j | j \in [p+1, n]\}$. So, we transform the Eq (7) into Eq (8) on the condition that each element in A and B are equal,

$$RMSE \cong \sqrt{\frac{p(\bar{e}^2 - \bar{f}^2)}{n} + \bar{f}^2} \quad (8)$$

Because of the assumption $|\bar{e}| \geq |\bar{f}|$, the value of RMSE increases with p changing from 1 to n. Thus we can summarize that the grading attitude of the students can significantly affect the performance of the PG model.

## 3.3 Comparison among grading error distributions in the simulation and actual datasets

By comparing different inference performance of the PG models in both simulation experiments and the real dataset, we analyze the effect of the features of bias $b_v$ and reliability $\tau_v$ on the precision of inferring true score $s_u$. In the Gibbs sampling process for fitting the PG models, the Eq (9) updates $s_u$ in iterations. where the variable $z_u^v$ is a constant value, besides the $\tau_v$ and $b_v$, the others are hyperparameters. From Eq (9) we can infer that the main factors affecting the true grade $s_u$ include a grader's bias and reliability.

$$s_u \sim \mathcal{N}\left(\frac{\gamma_0\mu_0 + \beta_0\tau_{u_i} + \sum_{v:v\to u_i}\frac{\tau_v(z_u^v - b_v)}{\lambda}}{\gamma_0 + \beta_0 + \sum_{v:v\to u_i}\frac{\tau_v}{\lambda}}, \frac{1}{\gamma_0 + \beta_0 + \sum_{v:v\to u_i}\frac{\tau_v}{\lambda}}\right) \quad (9)$$

In order to verify the conclusion of our analysis, we compare the grading errors of the PG models in simulation experiments and the real dataset. The real dataset was collected in the SPOC course on Computer Network in our university. We build an online learning system to support the session of the course with the total enrollment of 724 students. Figure

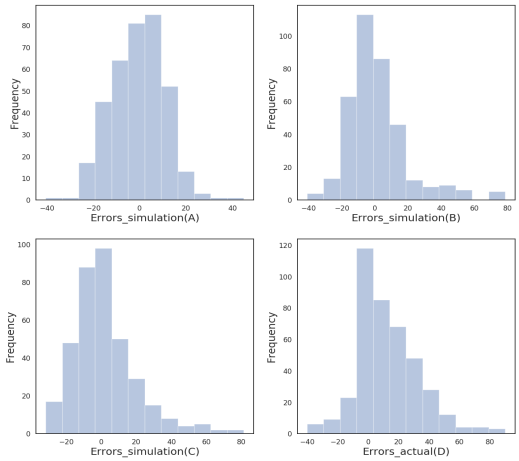


**Figure 2: The distributions of errors in three simulation datasets and the actual dataset. Fig A, B and C denote the histogram of grading errors generated by simulation experiments. Fig D denotes the distribution of the real dataset based on the Computer Network course.**

2 shows that the simulation and actual datasets have a very different error distribution. Fig 2A assumes that every student's grading behavior follows the gaussian model defined

in Eq (5). In contrast, the real dataset in Fig 2D indicates that many students' grading behavior doesn't satisfy the gaussian distribution. In order to further confirm the conclusion, we have conducted the other simulation experiments, in which we configure 40% undutiful students and 60% undutiful students to follow the random grading behavior defined in Eq (6), respectively. The results as shown in Fig 2B and Fig 2C demonstrates a similar error range to Figure 2D. These observations suggest that students in the SPOC experiment tend to exhibit random grading behavior. Clearly, such a high deviation of the peer grades in the real dataset from their ground truth is the reason why the PG models cannot achieve the low RMSE as we expect.

# 4. THE HUMAN-MACHINE HYBRID PEER GRADING FRAMEWORK

This section presents the design of our human-machine hybrid framework in detail, as shown Figure 3. The main idea of the framework is to use the auto-grader as an anomaly detector to screen the peer grades generated by undutiful students. The framework consists of three major components including a homework Auto-Grader, a Score-Filter and the PG models.

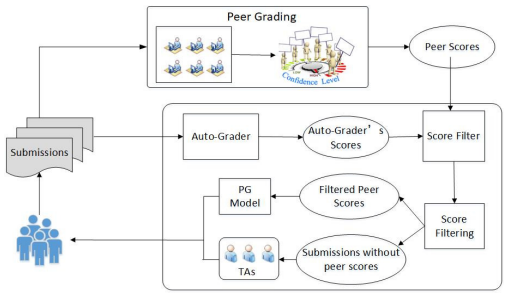


**Figure 3: The human and machine hybrid framework of peer grading.**

In the process of peer grading, the system first allocates the tasks for each student to perform their peer-grading tasks. After the Auto-Grader receives a score for a submission, it estimates a score for the same submission, and passes the estimation to the ScoreFilter. The ScoreFilter is responsible for comparing the Auto-Grader's estimation with the original peer score, and abandoning the peer score if the deviation between these two scores goes beyond the predefined threshold. With the co-ordination of the Auto-Grader and the ScoreFilter, the framework divides the student submissions into two groups: one group includes the submissions with legitimate peer grades that can be aggregated by the PG model for the grade inference, the other includes those without valid peer grades that have be sent to TAs for evaluation.

## 4.1 Naive Bayesian based Classifier as Auto-Grader Implementation

Based on Naive Bayesian method, we design a weak text classifier as the Auto-Grader in the hybrid peer grading framework. Each course assignment report often contains several problems. Thus the Auto-Grader's design consists of several classifiers, each of which classifies one problem in the assignment report. The grade classification results for all the problems of the assignment are mapped into scores based on its rubric and combined together as the total score of the assignment report.

## 4.2 Score Filtering and Postprocessing

The ScoreFilter in the hybrid human-machine grading framework adopts a simple filtering process. It computes the absolute value of the difference between grades estimated by the Auto-Grader and the peer-graded scores, sorts the scores in a descending order, and filter out the top 20% with the highest deviation values. The design of the score filter involves two major issues: The threshold for dropping unreasonable scores and the post-processing strategy for supplementing abandoned scores.

**The Error Threshold of Score Filtering**

Because our Auto-Grader is a weak classifier, we need to consider the classification error of each sub-problem of a homework report when we use the Auto-Grader to evaluate each sub-problem. We define the following equation to calculate the grading error.

$$Threshold_{error} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - a_i)^2}{n}} \quad (10)$$

In the Eq (10), $x_i \in x_1, x_2, \cdots, x_n$ denotes the score given by a student grader, $a_j \in a_1, a_2, \cdots, a_n$ denotes the score estimated by the Auto-Grader. The value of n presents the number of the problems in an assignment. We use Eq (10) to predict the error for each peer-graded score, and sort the list in a descending order according to the value of the prediction error, thus filtering out the peer grades with high errors values.

**The Post-Process Strategy of Score Fltering**

This simple filter algorithm above may cause potential problems for the PG models. After the ScoreFilter drops these unreasonable peer-graded scores, it can create extreme cases where most peer scores for a student assignment are eliminated. In such a case, a post-processing step is necessary in the ScoreFilter to supplement new scores for the downstream PG models. For the post-processing step, we propose

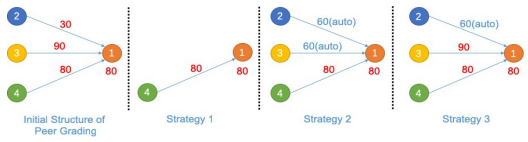


**Figure 4: Three Strategy to replace filtered grades.**

three strategies to handle the filtered-out scores. Dropping only: The ScoreFilter simply drops the scores identified by the auto-grader and does not supplement any new scores; Replacement by Auto-Grading: The Score-Filter directly uses the grades generated by the Auto-Grader to replace the peer scores that are identified as biased; Mixed Replacement: This strategy is only designed as a contrast strategy, which can choose the replacement score for a filtered peer score among the rest peer scores and the score predicted by the auto-grader based on their absolute difference value from the ground truth. Although it is impossible to implement

this strategy in the real system, it gives us an upper-bound for the strategy design when the ground-truth is available.

Figure 4 presents the example of all the strategies. In Figure 4, the leftmost graph is the relationship between the original peer score and the real score of the submission, from left to right, the second subgraph represents the score aggregation method using the first strategy, and the third subgraph Represents the score aggregation method using the second strategy, and the last subgraph represents the using of the third strategy for score aggregation.

# 5. EXPERIMENTS AND RESULTS

The peer-grading experiment was conducted in the course of Computer Network, which is offered to the senior college students of the computer science major. After class, students must design a networking plan and describe device configurations in their laboratory reports. These reports are evaluated through the peer-grading process. Our experimental dataset was collected from the class sessions in Year 2015-2017, including a total of 6 peer grading assignments and 724 students and 2354 assignment reports.

## 5.1 The prediction accuracy of the Auto-Grader

We choose the assignment reports on the sub-networking chapter of the course as the training and test data to develop the classifier of the Auto-Grader. This sub-networking assignment consists of six problems. For each problem in the assignment, there is a rubric specifying the grading category and score scheme. Table 1 displays the categories of rubric for each problem.

**Table 1: The categories of the each problem of the assignment.**

| Problem ID | Category 1 | Category 2 | Category 3 | Category 4 |
|---|---|---|---|---|
| 1, 2, 3 | 0 | 5 | 10 | — |
| 4 | 0 | 10 | 20 | — |
| 5 | 0 | 10 | 15 | 20 |
| 6 | 0 | 10 | — | — |

In the rubrics for Problem 1-3, there are three categories and the scoring values of each category are 0, 5 and 10 points. The rubric for Problem 4 also has three categories, including 0, 10 and 20 points. The rubric for Problem 5 has 4 categories, including 0, 10, 15, and 20 points. The rubric for Problem 6 only has two categories, including 0 and 10 points. Based on the above design of rubrics, the classifier of our Auto-Grader can achieve reasonable grading accuracy. The experimental results of the Auto-Grader are shown in Table 2. The grading accuracy of the Auto-Grader classifier within 5 points can achieve more than 60%, and the accuracy within 10 points becomes higher partly because of the design of the rubrics. This shows that the Auto-Grader can present reasonable score estimation as long as the threshold of the error is set to 10 points.

**Table 2: The prediction accuracy of Auto-Grader based on Naive Bayes.**

| Problem ID | $\leq 5$ | $\leq 10$ |
|---|---|---|
| 1 | 66.29% | 100% |
| 2 | 73.6% | 73.6% |
| 3 | 73.03% | 73.03% |
| 4 | 60.11% | 90.45% |
| 5 | 66.85% | 87.64% |
| 6 | 65.73% | 100% |

## 5.2 Choice of Post-processing Strategies for Score Filtering

We evaluate the performance of the ScoreFilter, especially the post-processing strategy. In addition to the three strategies described in Section 4.2, we also run the post-processing with the ground-truth strategy, in which the filtered top 20% peer scores are replaced by the ground-truth value. From Table 3, one can find that the Dropping-only strategy shows better performance than the Replacement by Auto-Grading strategy. The reason may be caused by the limited grading accuracy of the classifier in the Auto-Grader. Although the Mixed-Replacement strategy and the Ground-truth strategy achieve the lowest RMSE, their implementation is not feasible in the real scenario. Therefore, we have chosen the Dropping-Only strategy for post-processing in the ScoreFilter.

**Table 3: The value of RMSE of Adopting the three post-processing strategies.**

| Post-Processing Strategy | RMSE | Post-Processing Strategy | RMSE |
|---|---|---|---|
| Dropping only | 17.29 | Mixed Replacement | 16.45 |
| Replacement by Auto-Grading | 30.89 | Only Ground Truth | 15.96 |

## 5.3 Tuning the Threshold of the Score Filter

When the Naive Bayesian based auto-grader is used to each problem in an assignment submission, we need to consider the classification error of each sub-problem when we use auto-grader to evaluate the grade of each sub-problem by Eq (10).

**Tuning the error thresholds**

We investigated the impact of the error threshold by comparing the value of RMSE generated by the PG models and the Auto-Grader under different threshold values. The results are shown in Table 5. In Figure 5, we set the threshold

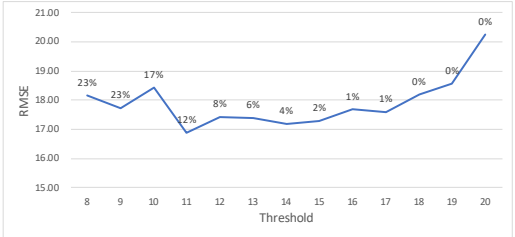


**Figure 5: The trend of RMSE changes with threshold. The online labels indicate the percentage of submissions that do not have a peer score as a percentage of the total of submissions.**

to filter the number of peer-graded scores with an interval of 1. We found that when the threshold is 11, the value of

RMSE drops to the lowest value, but the number of the submissions without peer grades accounts for 12% of the total submissions. In this case, the class TAs have to check these submissions and give their evaluations as the input for PG model. Therefore, when the number of assignment report is large, it will bring extra workload to the class TAs. Through our further experiments, we find optimal threshold should be 14.3, where the minimum RMSE can be calculated as 16.38, and only 3% submissions have to be assessed by the class TAs. In practice, the class instructors have to run a few rounds of peer-grading to determine the distribution of peer-grades scores and set the empirical value for the error threshold.

### 5.4 Overall Performance of the Hybrid Peer-Grading Framework

In order to evaluate the performance of the hybrid peer-grading framework, we run the PG models after the peer-graded scores are filtered by either the framework or random filtering respectively. In this way, we can generate three group of experimental data: the initial dataset without any score filtering, the dataset with Naive Bayesian-based Auto-Grader filtering, and the dataset with random filtering. The RMSE of the PG models based on the three data sets is shown in table 4.

After the peer-graded scores are sorted in a descending order of the estimated error, the top 20% of the scores are filtered out in each experiment. The filter process may eliminate all peer-graded scores for some submissions which have be re-evaluated by TAs. In the above experiment, when the error threshold is set to 15, 706 submissions are left with at least one peer-graded score. Only 18 submissions which account for 2% of all lose all the peer-graded scores. Thus, the task of re-evaluating these submissions does not bring too much burden to the course TAs. It can be seen from the Table 4, no matter which PG model is used, the human-machine hybrid framework can obtain the best performance, which averagely reduces the RMSE by 4. This outcome confirms that the hybrid human-machine peer-grading framework can improve prediction accuracy of the PG models with the presence of random grading behavior.

## 6. CONCLUSION

In this paper, we introduce a novel human-machine hybrid peer-grading framework to alleviate the problem of the random grading where student graders perform their peer-grading tasks in an undutiful manner. The most important component of the framework is the Auto-Grader that can classify students' submissions using machine learning and enable the framework to filter out the peer-graded scores with high errors. When filtering the peer grades, the framework calculates the error threshold according to the RMSE metric. Extensive experiments confirm that the hybrid framework can effectively eliminate the noise in peer-graded scores made by undutiful student graders and improve the prediction accuracy of the PG models.

**Table 4: RMSE comparison between the human-machine framework and the PG models.**

| | RMSE | | |
|---|---|---|---|
| Models | Without Auto grader Filtering | With Naive Bayes-based Auto-grader Filtering | Generated by filtering randomly |
| PG1 | 21.90 | 17.09 | 22.10 |
| PG3 | 20.40 | 17.30 | 21.36 |
| PG4 | 21.57 | 17.49 | 22.02 |
| PG5 | 20.26 | 16.71 | 21.86 |

## 7. ACKNOWLEDGMENTS

This work was supported by grant from the National Key Research and Development Program of China (Funding No. 2018YFB1004502) and supported by NSFC (Grant No. 61532004), and Grant from State Key Laboratory of Software Development Environment (Funding No. SKLSDE-2017ZX-04).

## 8. REFERENCES

[1] L. De Alfaro, M. Shavlovsky, and V. Polychronopoulos. Incentives for truthful peer grading. *arXiv preprint arXiv:1604.03178*, 2016.

[2] A. Gao, J. R. Wright, and K. Leyton-Brown. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. *arXiv preprint arXiv:1606.07042*, 2016.

[3] X. A. Gao, A. Mao, Y. Chen, and R. P. Adams. Trick or treat: putting peer prediction to the test. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 507–524. ACM, 2014.

[4] Y. Han, W. Wu, and X. Zhou. Improving models of peer grading in spoc. In *EDM*, 2017.

[5] N. Madnani and A. Cahill. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109, 2018.

[6] F. Mi and D.-Y. Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[7] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

[8] Y. Song, Z. Hu, and E. F. Gehringer. Collusion in educational peer assessment: How much do we need to worry about it? In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2017.

[9] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, 2016.

[10] M. Uto and M. Ueno. Item response theory for peer assessment. *IEEE transactions on learning technologies*, 9(2):157–170, 2016.