Improving Models of Peer Grading in SPOC *

Yong Han, Wenjun Wu, Xuan Zhou
State Key Laboratory of Software Development Environment,
School of Computer Science, Beihang University, China
{hanyong, wwj, zhouxuan}@nlsde.buaa.edu.cn

ABSTRACT

Peer-grading is commonly used to allow students to work as graders to evaluate their peer's open-ended assignments in MOOC courses. As a variant of MOOCs, SPOC (Small Private online course) adopt the peer-grading method to grade a number of student submissions. We propose a new ability-aware peer-grading model for SPOC courses by introducing prior knowledge level of each student grader as their grading ability in the process of calculating grading score.

1. INTRODUCTION

Small Private online course (SPOC) is a version of MOOCs used locally with on-campus students. It often has the relatively smaller number of students than a MOOCs course. SPOC students may come from the same classroom and know each other. Previous research efforts on peer-grading suggest that there is great disparity between the observed scores presented by student graders and the true scores (the instructor-given scores). Therefore, it is a major challenge on how to correctly aggregate peer assessment results to generate a fair score for every homework submission.

To solve the problem, we propose a group of new peer-grading models by considering the student mastery of knowledge level as a major factor for estimating final scores. Throughout the paper, we call the mastery of knowledge level as the students' grading ability. Based on every student's learning behavior and quiz-answering outcomes, we design a two-stage individualized knowledge tracing model to accurately assess their grading ability. Moreover, we introduce the new peer-grading models by integrating every student's grading ability into the factor of reliability. Experimental results in our SPOC course verify the effectiveness of our new models.

2. RELATED WORK

Many research efforts have been made to investigate the factors that can affect the grader bias and reliability.

*The accompanying appendix at: http://admire.nlsde.buaa.edu.cn/paper/2017-3.pdf

Goldin et al. [1] used the Bayesian models for peer grading in the setting of traditional classrooms. They explored the major factors including grader bias, and the rubric biases in their models. Walsh introduced a new algorithm named by PeerRank[4] based on the assumptions that the ability of student graders can be measured by the grades they received in the process of peer grading. Our models are inspired from the previous research work done in [3, 2]. We introduce the grading ability of students in their models and develop an individualized knowledge tracing model to estimate such ability.

3. DATASETS

The data sets in our experiments were collected in a SPOC course named by "The Experiment of Computer Network" that is hosted on our MOOC platform. The course is designed to teach both 4th grade CS undergraduate and the first-year graduate students about basic knowledge and skills on designing networking plans and configuring networking devices at the multiple levels of link protocol, TCP/IP protocol and network applications.

The course comprises of 10 chapters, each of which has 8-14 problems as homework assignment for students. The course also includes two open-ended assignments in graduate courses and three open-ended assignments in undergraduate courses. Preliminary statistical analysis of the dataset reveals that most peer-graded score tend to be higher than instructor-given scores for the same submissions.

4. PROBABILISTIC MODELS OF PEER GRADING IN SPOC

In this paper, we first establish a two-stage model to assess student mastery level of each knowledge skill, which can be used for estimating the graders' reliability. And then, we present three probabilistic graph models for peer grading by extending the models PG4 and PG5 of [3].

4.1 Individualized Knowledge-Tracing model for Ability Estimation

At the first stage, we extract interpretive quantities to predict the probability that a student has mastered the knowledge of that certain chapter in which the logistic regression method is used to fit these features and predict the engagement level of every student[5]. At the second stage, our work adopts the knowledge tracing model and ameliorates it by combining the prediction results obtained in the first stage. The sequence of the exercises in each unit is modeled by H-

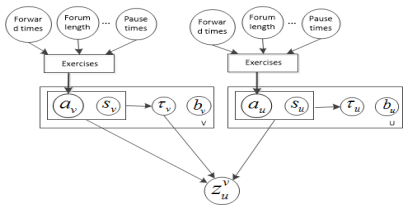

Figure 1: The relationship of the factors used in our models.

MM named as PPS (the Prior Per Student Model). We refer to the results that the HMM generated as a_v, which denotes the ability of graders prior to the peer-grading tasks. We train the model of HMM by using a_v as the initial element of the sequence and then introduce it and the true score as the parameters to model the reliability of a grader by a distribution of Gamma or Gaussian.
Our Experiments show that our estimated ability has relevance with the true score and can be used to estimate the grader reliability. Thus it is reasonable to use grader ability to estimate the reliability.

4.2 Peer-Grading models

We represent a_v as the prior distribution of estimating every grader's mastery of preparatory knowledge, τ_v as the reliability of the student grader v, b_v as the bias of the student grader v, s_u as the true score of a submission, and z_u^v as observed score for the submission.
Model PG6

$$\tau_v \sim \mathcal{G}(a_v, \beta_v)$$
$$b_v \sim \mathcal{N}(0, 1/\eta)$$
$$s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0)$$
$$z_u^v \sim \mathcal{N}(s_u + b_v, 1/\tau_v)$$

We refer to our first model as PG6: the reliability variable τ_v follows the Gamma distribution with a_v as the shape parameter instead of the true score in PG4 in [2] and utilize the student's performance on multiple-choice exercises to estimate his reliability in the process of peer-grading tasks.
Based on Model PG6, we introduce the Model PG7 by remodeling the reliability variable τ_v ($\tau_v \sim \mathcal{N}(a_v, \beta_v)$) with the Gaussian distribution instead of the Gamma distribution. The mean value of the Gaussian distribution in PG7 is still a_v. We also make further extension on Model PG7 by adding the true score s_v with the a_v to calculate the mean of the reliability variable τ_v ($\tau_v \sim \mathcal{N}(\theta_1 a_v + \theta_2 s_v, 1/\beta_v)$) and introduce the parameter λ to re-model the observed variable z_u^v ($z_u^v \sim \mathcal{N}(s_u + b_v, \lambda/\tau_v)$). This extended model is named as Model PG8.
In the above three models (PG6-PG8), we assume the overall bias random variable b_v follows the Gaussian distribution with the mean value at zero. The true score s_u follows the Gaussian distribution with the mean value at μ_0. Moreover, the hyper-parameters β_0, η_0, μ_0, γ_0, θ_1, θ_2, λ are the priors. For the observed scores z_u^v in the PG8, the parameter λ is similar to β_0 in PG6 and PG7, whose function is to scale the variance of its Gaussian.

4.3 Inference and evaluation

The details of the model inference procedures for PG6, PG7 and PG8 are described in the appendix. Our experiments are all based on Gibbs sampling. At the beginning of the Gibbs sampling process, the values of these parameters β_0, η_0, μ_0, γ_0 and λ are initialized to empirical values. We run our experiments by running for 400 iterations with the first 50 burn-in samples eliminated.

5. EXPERIMENTAL RESULTS

We compare our models PG6-PG8 with the baseline model based on simple median value, the models of PG1-PG3 proposed in [3], and the models of PG4-PG5 defined in [2]. The evaluation metric is the root-mean-score-error (RMSE), which is computed as the deviation between the estimated score and the true score assigned by the course staff.
Compared to PG1-3 and PG4-5, our models PG6 and PG7 demonstrate the same level of RMSE in most cases. The model PG8 has more obvious improvement than PG6-7, achieving the lowest RMSE. Therefore, it confirms that PG8 demonstrates the best performance among all the models on average. By combining the grader ability and the true score, the model PG8 is the best approach among all the models for estimating the peer-grading scores in SPOC courses.

6. CONCLUSIONS

In this paper, we first introduce a two-stage individualized knowledge tracing model to estimate each grader's level of knowledge mastery as their grading ability. And then, we propose three new probability graph models by introducing the grading ability as the major parameter for the latent variable of grader reliability. The experiments based on the dataset of our SPOC course demonstrate that our models can be effectively applied to aggregate the peer grades in SPOC courses.

7. ACKNOWLEDGMENTS

This work was supported by grant from State Key Laboratory of Software Development Environment of Beihang university of China (Funding No. SKLSDE-2015ZX-03) and NSFC (Grant No. 61532004).

8. REFERENCES

[1] Ilya M Goldin. Accounting for peer reviewer bias with bayesian models. In *Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems*, 2012.
[2] Fei Mi and Dit-Yan Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs. In *AAAI*, pages 454–460, 2015.
[3] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.
[4] Toby Walsh. The peerrank method for peer assessment. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, pages 909–914. IOS Press, 2014.
[5] Hsiang-Fu Yu, Hung-Yi Lo, Hsun-Ping Hsieh, Jing-Kai Lou, Todd G McKenzie, Jung-Wei Chou, Po-Han Chung, Chia-Hua Ho, Chun-Fu Chang, Yin-Hsuan Wei, et al. Feature engineering and classifier ensemble for kdd cup 2010. In *KDD Cup*, 2010.