So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Estimating Student Learning Ability from Massive Open Online Courses

Lanling HAN[a*] & Yuqin LIU[b] & Qi CHEN[c] &Hua SHEN[d,e]

[a,b]*School of International Information and Software, Dalian University of Technology, China*
[c]*School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan*
[d]*School of Computer Science and Technology, Dalian University of Technology, China*
[e]*College of Mathematics and Information Science, Anshan Normal University, China*
*lanling@dlut.edu.cn

Abstract: Massive open online courses (MOOC) provide a possible way for students to learn knowledge by themselves. Since the number of enrollments for each course is much larger than traditional in-person courses, it is hard for teaching faculty to master the learning ability of each student. However, estimating student learning ability is of great importance for delivering course content. Towards this goal, in this paper, we provide a novel way to estimate personalized learning ability for each student from their answering records in an exam by applying machine learning techniques, called truth discovery, which can automatically estimate a weight for each student and infer answers of questions. The weight can be considered as the student learning ability. The experimental results demonstrate the effectiveness of the utilized truth discovery approach for estimating student learning ability.

Keywords: Student learning ability, massive open online courses, truth discovery

1. Introduction

With the rapid development of science and technologies, more and more people want to seek the cutting-edge knowledge from massive open online courses (MOOC) and learn the knowledge by themselves. This is a new resolution compared with traditional in-person courses. In traditional courses, faculty can adjust the teaching speed and content based on the reaction of students. On the other hand, since the number of students is small, faculty is able to master the learning ability of each student accurately. However, for online courses, it is hard to estimate the learning ability for each student as the number of enrollments is much larger than traditional classes (Hwang et al., 2017, Yin et al., 2017; Yin et al., 2015).

However, estimating student learning ability is of great importance for delivering course content. If students master all the content, then teaching faculty may introduce more advanced knowledge to enhance the quality of the course. Otherwise, teaching faculty may introduce more details and provide more examples to help students understand the course content well. Thus, how to design a new approach to automatically estimate student learning ability is a key challenging task in massive open online courses.

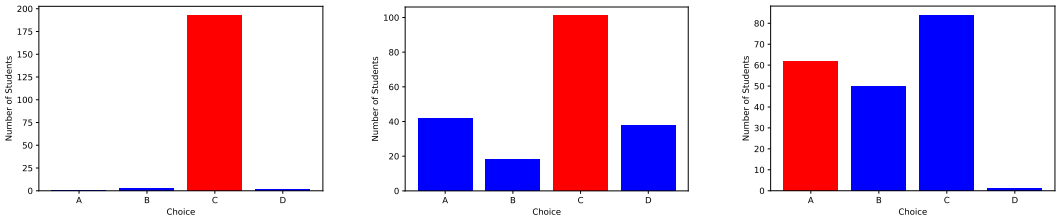
Existing work on student learning ability estimation mainly focuses on supervised learning (Lincke et al., 2019a; Lincke et al., 2019b; Wang et al., 2019). These approaches use machine learning and data mining methods to learn student learning ability based on the answering history data. For each question in the training data, the ground truth data are available. However, for online courses, especially for the essay questions, it is hard to obtain the ground truth data. In this scenario, supervised machine learning approaches cannot work. Thus, the new challenge is how to estimate student learning ability even without using ground truth data.

To address this challenge, in this paper, we provide a novel way to estimate personalized learning ability for each student from their answering records in an exam by applying state-of-the-art machine learning techniques, called truth discovery, which can automatically estimate a weight for each student and infer answers of questions (Li, et al., 2014; Ma et al., 2015; Xiao et al., 2016, Wang, et al., 2017, Zhang et al., 2016, Ma, et al., 2017; Zhang, et al., 2018; Yao, et al., 2018). The weight

can be considered as the student learning ability. We conduct experiments on an exam dataset collected from an online Japanese course at Dalian University of Technology. The experimental results demonstrate the effectiveness of the utilized truth discovery approach for estimating student learning ability.

2. Exam Data Analysis

The Exam dataset contains 43 questions, 199 students, and 8,544 answering records. Each question has four choices but only one correct answer. We provide distributions of student answers on three questions with different difficulty levels in Figure 1. The red bar represents the number of the students who provided the correct answers.



(a) Question ID: 11 (b) Question ID: 13 (c) Question ID: 43

Figure 1. Examples of questions with different difficulty levels.

In Figure 1, Question 11 is easy, and 97.0% students can answer it correctly. However, for Question 43, only 31.2% students can provide the correct answers. Therefore, this is a hard question. Intuitively, if students always answer correctly on hard questions, the learning ability should be higher than others, i.e., the answers provided by these students are more reliable. In other words, the more reliable a student is, the more likely this student would provide trustworthy answers, and vice versa. This principle is in accord with the idea of truth discovery approaches.

3. Truth Discovery

The goal of truth discovery approaches is to learn a weight for each crowd worker and estimate the true answer for each question or task. In our scenario, the weight of each crowd worker can be considered as the learning ability of each student, and we apply a commonly-used truth discovery model, called CRH (Li, et al., 2014), to estimate the learning ability.

Let N represent the number of students and M denote the number of questions. The answering record provided by the n-th student on the m-th question is denoted as a_m^n. The goal of this task is to estimate the learning ability w_n for the n-th student and infer the true answer t_m for the m-th question following the above-mentioned principle, which can be mathematically defined as follows:

$$\min_{T,W} \sum_{n=1}^{N} \sum_{m=1}^{M} w_n d(a_m^n, t_m) \quad s.t. \sum_{n=1}^{N} \exp(-w_n) = 1,$$

where T is the set of inferred answers, W is the set of estimated learning ability of all the students, and $d(\cdot,\cdot)$ is the distance function. In this paper, we use 0-1 loss as the distance function, i.e., if the answer provided by a student is the same as the estimated truth, the loss is 0; otherwise, the loss is 1. The loss function is formally defined as follows:

$$d(a_m^n, t_m) = \begin{cases} 1 & if\ a_m^n \neq t_m \\ 0 & otherwise \end{cases}$$

We can apply iterative procedures to solve the above optimization problem as (Li, et al., 2014). First, we use majority voting approach to infer the true answers for questions. We then fix these answers

to learn the learning ability of students using $w_n = -\log \frac{\sum_{m=1}^{M} d(a_m^n, t_m)}{\sum_{n=1}^{N} \sum_{i=1}^{M} d(a_i^n, t_i)}$, and in turn, we fix the learning ability to estimate the true answers, i.e., $t_m \leftarrow arg\min \sum_{n=1}^{M} w_n d(a_m^n, t_m)$. These two steps are iteratively updated until these two parameters converge.

4. Experiments

We use *error rate* as the evaluation metric, which is defined as the number of incorrect estimated questions divided by the total number of questions *M*, to evaluate the proposed approach for estimating student learning ability. The lower the error rate, the better the performance. We use majority voting (MV), TruthFinder (Yin et al., 2008), and Investment (Pasternack et al., 2010) as baselines, and the results are listed in Table 1. We can observe that the applied CRH can achieve the best performance, which illustrates the effectiveness of the used approach.

Table 1. *Performance Comparison.*

Method	Error Rate
MV	0.1163
TruthFinder	0.2326
Investment	0.2326
CRH	**0.0698**

Besides, to validate the reasonability of applying CRH framework, we conduct the following experiment. For this exam, the lecturer assigned scores for each question, and the full scores are 100 points. We plot the comparison graph between the real scores and the estimated student learning ability in Figure 2. We can observe that the learned ability values (Y-axis) are positively correlated to the final scores given by the lecturer (X-axis), which clearly shows that the estimated weights by the CRH framework are reasonable and accurate to reflect the learning ability of students.

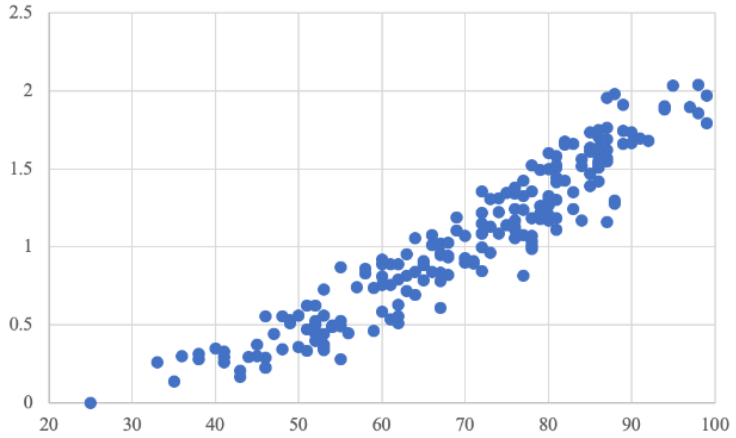


Figure 2. Real obtained scores given by the lecturer v.s. the estimated learning ability by CRH.

5. Conclusions

In this paper, we investigate a new machine learning approach to estimate student learning ability. Through analyzing the results in a real online Japanese course exam, we validate the effectiveness and reasonableness of the applied approach. In the future, we aim to design more advanced and novel approaches for estimating student learning ability.

Acknowledgements

This research was financially supported by the Foundation of Education Department of Dalian University of Technology (Grant No. ZD2020017) and the National Social Science Foundation of China under Grant No. 17BYY192.

References

Hwang, G.J., Chu, H.C., Yin, C. (2017) Objectives, Methodologies and Research Issues of Learning Analytics. In Interactive Learning Environments, Volume 25, Issue 2, pp. 143-146.

Yin, C. J., Uosaki, N., Chu, H. C., Hwang, G. J., Hwang, J. J., Hatono, I., Kumamoto, E., Tabata, Y. (2017) Learning Behavioral Pattern Analysis based on Students' Logs in Reading Digital Books. In Proceedings of the International Conference on Computers in Education, pp. 549-557.

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., Yamada, M., Kojima K., Ogata, H. (2015) Analyzing the Features of Learning Behaviors of Students using e-Books. In Workshop Proc. of International Conference on Computers in Education, pp.617-626.

Lincke, A., Fellman, D., Jansen, M., Milrad, M., Berg, E., Jonsson, B. (2019a) Correlating Working Memory Capacity with Learners' Study Behavior in a Web-Based Learning Platform. In Proceedings of the 27th International Conference on Computers in Education, pp. 90-92.

Lincke, A., Jansen, M., Milrad, M., Berge, E. (2019b) Using Data Mining Techniques to Assess Students' Answer Predictions. In Proceedings of the 27th International Conference on Computers in Education, pp. pp. 42-50.

Wang, T., Ma, F., Gao, J. (2019) Deep hierarchical knowledge tracing. In Proceedings of the 12th International Conference on Educational Data Mining.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J. (2014) Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In Proceedings of SIGMOD, pp. 1187-1198.

Ma, F., Li, Y., Li, Q., Qui, M., Gao, J., Zhi, S., Su, L., Zhao, B., Ji, H., Han, J. (2015) FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In Proceedings of SIGKDD, pp. 675-684.

Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., Zhang, A. (2016) Towards Confidence in the Truth: A Bootstrapping based Truth Discovery Approach. In Proceedings of SIGKDD, pp. 1935-1944.

Wang, Y., Ma, F., Gao, J. (2017) Discovering Truths from Distributed Data. Proceedings of ICDM, pp. 505-514.

Zhang, H., Li, Q., Ma, F., Xiao, H., Li, Y., Gao, J., Su, L. (2016) Influence-aware truth discovery. In Proceedings of CIKM, pp. 851-860.

Ma, F., Meng, C., Xiao, H., Li, Q., Gao, J., Su, L., Zhang, A. (2017) Unsupervised discovery of drug side-effects from heterogeneous data sources. In Proceedings of SIGKDD, pp. 967-976.

Zhang, H., Li, Y., Ma, F., Gao, J., Su, L. (2018) Texttruth: an unsupervised approach to discover trust-worthy information from multi-sourced text data. In Proceedings SIGKDD, pp. 2729-2737.

Yao, L., Su, L., Li, Q., Li, Y., Ma, F., Gao, J., Zhang, A. (2018) Online truth discovery on time series data. In Proceedings of SDM, pp. 162-170.

Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., Zhang, A. (2018) Towards confidence interval estimation in truth discovery. IEEE TKDE, 31(3), pp. 575-588.

Yin, X., Han, J., Yu, P. S. (2008) Truth discovery with multiple conflicting information providers on the web. In IEEE Transactions on Knowledge and Data Engineering, 20(6):796–808.

Pasternack, J., Roth, D. (2010) Knowing what to believe (when you already know something). In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 877–885.