A Diagnostic Question Analysis Model based on a Modified Item Response Theory

Haonan Gao(✉)

University of Toronto, University College, Toronto M5S1A1, Canada
haonan.gao@mail.utoronto.ca

Abstract. Digital technologies are being more widely used in education, allowing students all around the world to access individualized, high-quality educational resources. This paper analyzes a massive amounts of data derived from students' interactions with these diagnostic questions can help us more accurately understand the students' learning status and thus allow us to automate learning curriculum recommendations, as evidenced by thousands of examples of students' answers to mathematics questions provided by The NeurIPS 2020 Education Challenge [1, 2]. In this paper, a new generated model based on Item Response Theory (IRT) is put forward. Additionally, with discrimination parameter added and classification by groups, the model on real-world dataset is verified.

Keywords: Item Response Theory · Missing Value Prediction · Education Study

1 Introduction

Online education programmes like Khan Academy and Coursera make high-quality education available to a wider audience. Students can learn new content by attending a lecture, reading course materials, and conversing with instructors in a forum on these platforms. However, one downside of the online platform is that assessing students' comprehension of course material is difficult. Many online education systems incorporate an evaluation component to ensure that students comprehend the essential themes in order to address this issue. Diagnostic questions, each of which is a multiple choice question with just one right answer, are frequently used in the assessment component. The diagnostic question is constructed in such a way that each erroneous answer exposes a common misunderstanding. When students answer the diagnostic question wrong, it indicates the nature of their misperception, and the platform may provide extra information to help them address it by understanding these beliefs. We may, in principle, extract relevant educational information such as how students learn from the data gathered by these exams and offer appropriate learning interventions to enhance learning outcomes by mining the data provided by these assessments. The quality of the insights gained, on the other hand, is determined by the quality of the evaluation questions.

The item response theory (IRT), also known as the latent response theory, is a set of mathematical models that attempts to explain the link between latent traits (unobservable characteristics or attributes) and manifestations (i.e. observed outcomes, responses or

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 3164–3169, 2022.
https://doi.org/10.2991/978-2-494069-31-2_371

performance). They make a connection between the qualities of things on an instrument, how people respond to them, and the underlying attribute being assessed. The latent construct (e.g., stress, knowledge, attitudes) and measure items are assumed to be structured on an unobservable continuum in IRT. As a result, its primary goal is to determine an individual's place on that continuum. In this task, besides from the common methods people usually apply, such as k-Nearest Neighbour, Matrix Factorization or Ensemble [3], we care about the unique characteristic of IRT, which is adding discriminative parameters and constants, as well as giving specific groups [4].

This paper introduces a modified Item Response Theory (IRT) model to predict whether a student can correctly answer a specific diagnostic question based on the student's answers and other students' responses to. The paper has a number of real-world implications, including the ability to recommend questions of appropriate difficulty to a given student based on their background and learning status, the ability to discover potential common misconceptions among students by clustering question-answer pairs that may indicate the same or related misconceptions, the ability to provide feedback to authors of diagnostic questions so that they can revise poor quality questions, and the ability to guide teachers to choose appropriate diagnostic questions.

2 Methodology

2.1 Data Collection

A dataset given by Eedi is used, an online education platform that is already being used in many schools, to subsample 542 students' responses to 1774 diagnostic questions. The programme provides students in grades K-12 with crowd-sourced mathematics diagnostic problems (between 7 and 18 years old).

2.2 Model Description

The IRT assigns each student an ability value and each question a difficulty value to formulate a probability distribution. In the one-parameter IRT model, β_j represents the difficulty of the question j, and θ_i represents the i-th students' ability. Then, the probability that the question j is correctly answered by student i is formulated as:

$$p(c_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \tag{1}$$

Derive the log-likelihood $\log(p(C|\theta, \beta))$ for all students and questions. Here C is the sparse matrix. The derivative of the log-likelihood with respect to θ_i and β_j can be calculated by the derivative of the logistic model with respect to the parameters [5]:

$$p(c_{ij} = 1|\theta_i, \beta_j) = \prod_i \prod_j \left(\frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \right)^{c_{ij}} \cdot \left(1 - \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \right)^{1-c_{ij}} \tag{2}$$

$$\frac{\partial l}{\partial \theta_i} = \sum_j \left(c_{ij} - \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \right) \tag{3}$$

$$\frac{\partial l}{\partial \beta_i} = \sum_i \left(-c_{ij} + \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \right) \tag{4}$$

We then introduce the discrimination parameter k_i into the one-parameter IRT model to show how steep the sigmoid looks, or to say how discriminative the question is. In the simple one-parameter model, all the curves are expected to have the same shape, which is somehow not reasonable in practice. In such way, the probability of a correct answer is given by:

$$\begin{aligned} p(c_{ij} = 1 | \theta_i, \beta_j) &= sigmoid\left(k_j \cdot (\theta_i - \beta_j)\right) \\ &= \frac{\exp(k_j \cdot (\theta_i - \beta_j))}{1 + \exp(\exp(k_j \cdot (\theta_i - \beta_j))} \end{aligned} \tag{5}$$

In one-parameter model of above, we initialized β_j to be 0 for every single question, which is not wrong but not precise. We noticed that there is a variable called subject_id in the question metadata, where each question is related to several subjects. Therefore, we decided to make some changes at the beginning of optimization, by calculating one minus the weighted mean of correctness rate of all the subjects this question is related to, and apply to β_j (difficulty of j-th question) [6]. For example, the difficulty of subject *3D Plane Symmetry* should be obviously higher than subject *Line Symmetry*. In such way, the initial β_j will be influenced by the subject difficulty of this question, which will help the model to be more fitted. Note that since the subject 0 is *Maths*, which is in all the questions' subject list, so we ignore this subject for non-informative reasons [7].

Lastly, the constant c is introduced, which represents the probability of getting a question right via. Random guess to improve optimization. At first, we set the value of c to be 0.25, which we believe is the probability when a student randomly selects an answer from 4 choices. However, we then recalled the real situation of ourselves when we are stuck at a question during examination, we might usually struggle with two or three potential answers, while the other one or two choices are labeled false surely. Thus, there are three potential situations, which are picking from 4 unsure choices, 3 unsure choices and 2 unsure choices. i.e., the probabilities of correctly answering the question under those situations are 0.25, 0.333…, and 0.5. Therefore, we decided to let c be the mean value of all possible probabilities, which is:

$$c = \frac{\left(\frac{1}{4} + \frac{1}{3} + \frac{1}{2}\right)}{3} \approx 0.3611 \tag{6}$$

The probability equation of our final model will be.

3 Results

The final model did improve on overall performance. By plotting the accuracy curve for both the baseline model and final model w.r.t. different iterations, we can see that the accuracy of the final model exceeds the baseline model when iteration time is larger than approximately 40. However, we still see a pretty clear improvement on the final model

Table 1. The experiment result on the baseline model and the final model

Iteration	Baseline Model	Final Model	Difference
50	0.70604007	0.70730073	+ 0.0012603
60	0.70660457	0.70797052	+ 0.0013659
70	0.70646344	0.70730073	+ 0.0008372
80	0.70519333	0.71827361	+ 0.0130802
90	0.70589895	0.71554253	+ 0.0096435
100	0.70632232	0.71827361	+ 0.0119512

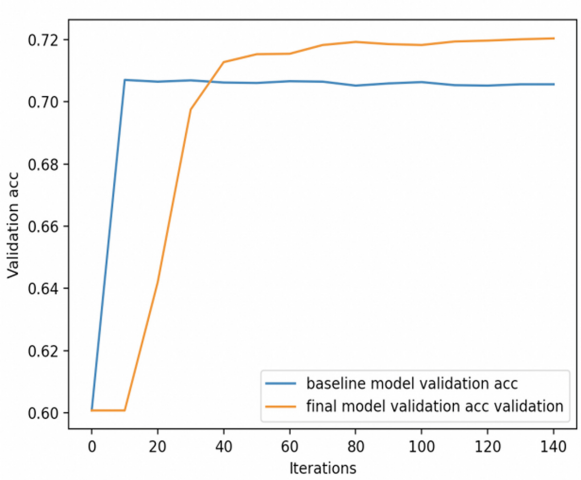

Fig. 1. Final model shows a better performance after 40 iterations

accuracy. Moreover, if we are still not sure if this is by chance or our model really helped the whole process, we can try some more bunches of smaller experiments [8, 9] (Table 1 and Fig. 1).

4 Conclusion

In this work, we introduced a modified Item Response Theory (IRT) model to predict whether a student can correctly answer a specific diagnostic question based on the student's previous answers to other questions and other students' responses. Based on the accuracies gained by this model compared to baseline models, such as the k-Nearest Neighbour (kNN) algorithm, matrix factorization approaches, and bagging ensemble, the IRT-based model did improve on the accuracy of successfully predicting students' answers to the diagnostic question. As a result, we'll be able to better predict a student's skill level on a customized learning platform.

Some other factors are not taken into account, for example the background of students. Student_meta.csv in the extension is not used, which contains factors like the age, gender of the student which may have influence on.

There are two limitations regarding the discriminative term. Firstly, the way of calculating the discriminative variable may not be thorough enough. This paper tries to take into account the weight of subjects. Since a question may belong to many subjects and some subjects are obviously bigger. But we do not know the deeper relation of subjects. For instance, subjects may have containment relation, or overlap [10]. So the weight might be biased. Secondly, this paper does not include the impact of the subject to students. Some students may have better performance on particular subjects, some may have worse performance. But the discriminative term does not give much influence to students' ability due to subjects. So maybe they should also be initialized based on how discriminative the question is.

This study tried to make variables more reasonable, but there are still aspects that can be improved. We assumed that there are three possible scenarios when students are guessing the answer. But the fact might be more complicated. A possible improvement might be initializing for every student based on their ability. Students with lower ability might have a higher possibility to guess the answer.

For some questions and subjects, we are suffering from the lack of training data [11]. In our extension, we are calculating the accuracy of each subject in the process of getting the discriminative variable. But some subjects are so small that they only have a handful of data, which might be biased. Bagging might be a good solution for this.

Acknowledgments. I would first like to thank my essay advisor Dr.Ta'asan of the Department of Mathematical Sciences at Carnegie Mellon University. He consistently allowed this paper to be my own work but steered me in the right the direction whenever he thought I needed it.

I must express my very profound gratitude to my parents and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Authors' Contributions. This paper si independently completed by Haonan Gao.

References

1. Wang Z, Lamb A, Saveliev E, et al. Diagnostic questions: The neurips 2020 education challenge [J]. arXiv preprint arXiv:2007.12061, 2020.
2. Wang Z, Tschiatschek S, Woodhead S, et al. Educational Question Mining At Scale: Prediction, Analysis and Personalization [C]//Symposium on Educational Advances in Artificial Intelligence (AAAIEAAI), 2021. https://www.microsoft.com/en-us/research/publication/educationalquestion-mining-at-scale-prediction-analysis-and-personalization.
3. Shah T, Olson L, Sharma A, et al. Explainable Knowledge Tracing Models for Big Data: Is Ensembling an Answer? [J] arXiv preprint, 2020, arXiv:2011.05285.

4. Vermunt J. K. Multilevel mixture item response theory models: an application in education testing [J] Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal, 2007, 2228.
5. Chen C. M, Lee H. M, Chen Y. H. Personalized e-learning system using item response theory [J] Computers & Education, 2005, 44(3): 237–255.
6. Chen C. M, Liu C. Y, Chang M. H. Personalized curriculum sequencing utilizing modified item response theory for web-based instruction [J] Expert Systems with applications, 2006, 30(2): 378–396.
7. C. Wylie and D. Wiliam. Diagnostic questions: Is there value in just one? In Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 6 to 12, 2006, in San Francisco, CA, 2006.
8. Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. arXiv preprint arXiv:1806.02920, 2018.
9. Loken E, Rulison K. L. Estimation of a four-parameter item response theory model[J]. British Journal of Mathematical and Statistical Psychology, 2010, 63(3): 509-525.
10. Wu M, Davis R L, Domingue B W, et al. Variational item response theory: Fast, accurate, and expressive [J] arXiv preprint arXiv:2002.00276, 2020.
11. Item Response Theory | Columbia Public Health. (2022). Retrieved 9 February 2022, from https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory.