Student Modeling Method Integrating Knowledge Tracing and IRT with Decay Effect

Shinichi Oeda[1] and Kouta Asai[2⋆]

[1] Department of Information and Computer Engineering,
National Institute of Technology, Kisarazu College
11-1, Kiyomidaihigashi 2-chome Kisarazu City, Chiba, Japan
oeda@j.kisarazu.ac.jp
[2] Advanced Control and Information Engineering Course,
National Institute of Technology, Kisarazu College

Abstract. Educational data mining (EDM) involves the application of data mining, machine learning, and statistics to information generated from educational settings. Modeling students' knowledge is a fundamental part of intelligent tutoring systems. One of the most popular methods for estimating students' knowledge is *knowledge tracing*. It is the de-facto standard for inferring students' knowledge from performance data. The goal of this study is to estimate future student performance from massive amounts of examination results. We propose a novel method to improve the precision of student modeling using knowledge tracing with *item response theory*, including the decay theory of forgetting.

Keywords: Educational data mining, knowledge tracing, item response theory, hidden Markov model, decay theory

1 Introduction

Intelligent tutoring systems (ITS) and learning management systems (LMS) have been widely used in the fields of education, and have allowed us to collect log data from learners, such as students. Educational data mining (EDM) aims at discovering useful information from the massive amounts of electronic data collected by these educational systems. EDM is an emerging multi-disciplinary research area where methods and techniques for exploring data originating from various educational information systems have been developed [1].

One of the goals of EDM is student modeling. It is one of the key factors affecting automated tutoring systems in making instructional decisions. The purpose of student modeling is the estimation of students' skills and The prediction whether a student solve an item or not from log data such as examination results. One of the most popular methods for estimating student knowledge is *knowledge*

⋆ Currently NIFTY Corporation, Human Resources Department, Shinjuku Front Tower 21-1, Kita-shinjuku 2-chome, Shinjuku-ku, Tokyo, Japan, asai.kota@nifty.co.jp

tracing [2]. It is the de-facto standard for inferring students' knowledge from performance data. An ITS provides efficient learning environments for students by assigned a suitable item for a student's skill level. The ITS employs a student model. In order to create a high-performance ITS, a student model is needed that can predict students' answers and estimate the state of their skills.

However, knowledge tracing did not consider the process of the decay theory of forgetting, whereby human memory fades over time. Conventional methods for knowledge tracing cannot handle the decay effect because it is difficult to estimate the parameters of model using the forgetting process. In order to comprehend the learning effects in the educational process, it is significant to study how the distribution of students' latent skills changes over time. We address the issue by incorporating *item response theory* into the decay effect. In this paper, we propose a novel method to improve the precision of student modeling using knowledge tracing with item response theory, including the decay theory of forgetting.

2 Knowledge Tracing

Knowledge Tracing was developed in 1995, and has since established its position as a well-known method of student modeling. Figure 1 uses the plate notation to show a graphical model of knowledge tracing. A question item in an examination requires several skills to solve.

The diagram shows that t is a learning opportunity, k_t is a latent variable as a skill state (master or not master) of the student, y_t is an observation variable as a result (correct or incorrect) of the student's response. Knowledge tracing is represented *hidden Markov model*, since student's skill states are not observed while student's results are observed.

In knowledge tracing, four parameters $P(L_0), P(T), P(G), P(S)$ for each skill are defined as follows:

already know

$$P(L_0) \stackrel{\text{def}}{=} P(k_0 = \mathit{true}), \tag{1}$$

learn

$$P(T) \stackrel{\text{def}}{=} P(k_t = \mathit{true} | k_{t-1} = \mathit{false}), \tag{2}$$

guess

$$P(G) \stackrel{\text{def}}{=} P(y_t = \mathit{true} | k_t = \mathit{false}), \tag{3}$$

slip

$$P(S) \stackrel{\text{def}}{=} P(y_t = \mathit{false} | k_t = \mathit{true}). \tag{4}$$

There are four types of model parameters used in knowledge tracing as the initial probability of knowing a skill a priori. $P(L_0)$ is the probability that a student has learned how to apply a knowledge component prior to the first opportunity to apply it in the ITS. $P(T)$ is the probability of a student's knowledge of a skill transitioning from the *not known* to the *known* state after an opportunity to apply it. Here, knowledge tracing assumes that a student does not forget

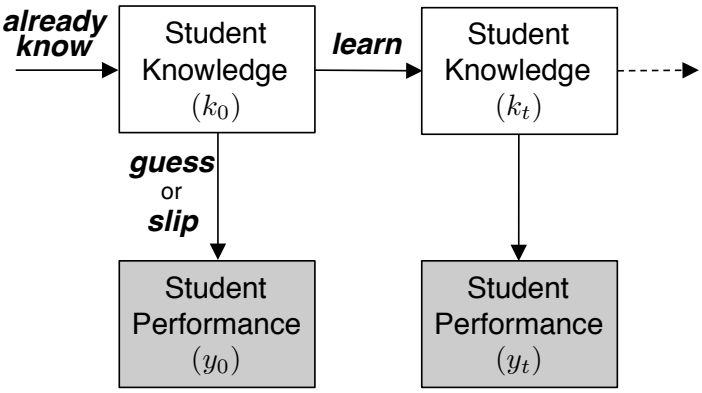


Fig. 1. Knowledge tracing.

a mastered skill if even once. Accordingly, the probability of skill transition from *master* to *not master* is zero. $P(G)$ is the probability of correctly applying an unknown skill, and $P(S)$ is the probability of making a mistake when applying a known skill.

Given that parameters $P(L_0), P(T), P(G), P(S)$ are set for all skills, the formulae used to update student knowledge of skills are as follows, from Equation (5) to (8), from the results of students' answers until opportunity t:

$$P(L_t = true|y_t = true) = \frac{P(L_t)(1 - P(S))}{P(L_t)(1 - P(S)) + (1 - P(L_t))P(G)}, \tag{5}$$

$$P(L_t = true|y_t = false) = \frac{P(L_t)P(S)}{P(L_t)P(S) + (1 - P(L_t))(1 - P(G))}, \tag{6}$$

$$P(L_{t+1} = true) = P(L_t|y_t) + (1 - P(L_t|y_t))P(T), \tag{7}$$

$$P(y_{t+1} = true) = P(L_{t+1})(1 - P(S)) + (1 - P(L_{t+1}))P(G). \tag{8}$$

Equations (5) and (6) update a skill state from the answer to opportunity t. The skill state of future opportunity $t + 1$ is calculated by Equation (7) by the updated value of Equation (5) and (6). Moreover, the probability that a student can answer an assigned item at $t + 1$ is calculated using Equation (8) by the derived value of Equation (7).

2.1 Estimation of parameters

In the knowledge tracing model, the four parameters $P(L_0), P(T), P(G), P(S)$ per skill are unknown. Although these parameters are defined by an expert, they are estimated by results from past data in general. We can estimate these parameters by the Baum–Welch algorithm [3], since knowledge tracing is a hidden Markov model.

3 Item Response Theory

3.1 Overview of model

IRT (item response theory) [4] is the study of examination and item scores based on assumptions concerning the mathematical relationship between a latent ability and item responses. The IRT model predicts the probability that a certain student will give a certain response to a certain item. Students can have different levels of ability, and items can differ in many respects. In IRT models, *Rasch model* like a logistic function is used on the ability variable to explain examinees' item responses as follows:

$$P_{ij}(y = true) = \frac{1}{1 + \exp(-1.7(\theta_i - \beta_j))}, \tag{9}$$

where index i indicates a student, j indicates an item, θ_i is the student's ability parameter for item j, and β_j is the difficulty parameter of item j.

Variable θ_i is considered the ability required to perform well on question items. The item response function gives the probability that a student with a given ability level will answer a question correctly. Students with lower ability have less of a chance, whereas those with higher ability are more likely to answer correctly.

3.2 Estimation of parameters

The common estimation methods for IRT are joint maximum likelihood estimation, marginal maximum likelihood estimation, and Bayesian estimation. However, it is difficult to calculate the joint maximum likelihood if the number of students increases. Marginal maximum likelihood estimation overcomes this issue by reducing the number of students through marginalization. On the other hand, it does not work when results are all correct or all incorrect. In this paper, we use Bayesian estimation in order to estimate parameters because it solves above the problems.

Although Bayesian estimation can analytically solve for a simple model like the Rasch model through Equation (9), it cannot solve the following complex model. In this paper, we use the *Markov Chain Monte Carlo* method, which can estimate the parameters of a complex model.

4 Related Work

4.1 Rasch model with forgetting

Lindsey et al. have developed the Rasch model using a theory of forgetting [5] through Equation (10), which is based on Equation (9), as follows:

$$P_{ij}(y = true) = \frac{(1 + ht_{ij})^{-\exp(\tilde{\theta}_i - \tilde{\beta}_j)}}{1 + \exp(-1.7(\theta_i - \beta_j))}, \tag{10}$$

where t_{ij} indicates the elapsed time between the initial presentation of item j to student i and a later recall test, $\tilde{\theta}_i$ indicates a forgetting parameter for student i, $\tilde{\beta}_j$ indicates a forgetting parameter for item j, and h is a scaling parameter.

The Rasch model with forgetting takes into account the elapsed time and the forgetting parameter. It is believed that human memory decays over time. The proposed model Equation (10) incorporates elapsed time, because of which the probability of a correct response decreases with time.

4.2 Combination of knowledge tracing and IRT

Khajah et al. have developed a method that combines knowledge tracing and the Rasch model in Equation (9), and yielded a higher prediction accuracy than previous methods [6].

We describe the method of combining two models. Equation (8) for knowledge tracing is rearranged as Equation (11) as follows:

$$P(y_t|\boldsymbol{y}^{(t-1)}) = \sum_{\substack{l\in\{\text{mastered,}\\ \text{not masterd}\}}} P(y_t|k_t=l)\cdot P(k_t=l|\boldsymbol{y}^{(t-1)}), \tag{11}$$

where $\boldsymbol{y}^{(t-1)} = y_0...y_{t-1}$. $P(y_t|k_t=l)$ which, appears on the right-hand side of Equation (11), and represents *slip* and *guess*. This part is replaced with the Rasch model as follows:

$$P(y_t|\boldsymbol{y}^{(t-1)}) = \sum_{\substack{l\in\{\text{mastered,}\\ \text{not masterd}\}}} \text{Rasch}(\theta_{i_t},\beta_{j_t},c_l)\cdot P(k_t=l|\boldsymbol{y}^{(t-1)}). \tag{12}$$

The Rasch model, as Equation (12), is added as a parameter of c_l. Although the IRT does not have the two parameters of *slip* and *guess*, c_l is added to the model. The model adds parameter c_l to Equation (9) of the Rasch model to Equation (13) as follows:

$$\text{Rasch}(\cdot) = c_l + \frac{1-c_l}{1+\exp(-1.7(\theta_i-\beta_j))}. \tag{13}$$

5 Proposed Method

In this paper, we propose a method that combines knowledge tracing and the Rasch model with forgetting in order to improve prediction accuracy. In the proposed model, we replace the Rasch function in equation (12) with the Rasch model with forgetting in equation (10). We similarly adds parameter c_l to Equation (10). The combined model can be represented as follows:

$$P(y_t|\boldsymbol{y}^{(t-1)}) = \sum_{\substack{l\in\{\text{mastered,}\\ \text{not masterd}\}}} \text{RF}(\theta_{i_t},\beta_{j_t},c_l)\cdot P(k_t=l|\boldsymbol{y}^{(t-1)}), \tag{14}$$

$$\mathrm{RF}(\cdot) = c_l + \frac{(1+ht_{ij})^{-\exp(\tilde{\theta}_i - \tilde{\beta}_j)} - c_l}{1+\exp(-1.7(\theta_i - \beta_j))}. \tag{15}$$

We employed Bayesian estimation to estimate the parameters of the model as in Section 3.2. We did not use simple a Bayesian model, but applied a Bayesian hierarchical model because it has hyperprior distributions.

6 Experiments

6.1 Overview of experiments

We conducted two experiments to evaluate the proposed model. A dataset was divided into training and test data. The training data was used to fit the parameters of the model and the test data to assess its generalization error. We verified that the proposed method could predict whether a given answer by a student was correct. We compared our method with two others: (i) original knowledge tracing, and (ii) the method represented in Equation (12). The proposed method is as in Equations (14) and (15).

We employed AUC (Area Under the Curve) and RMSE (Root Mean-squared Error) as measures for evaluation. AUC is a metric for a two-class prediction problem; the value of the AUC is 1 if the prediction is completely correct and 0.5 if the prediction is random. RMSE is a metric for numerical predictions, where its value represents the difference between the values predicted by a model and those observed. In short, a high-performance model indicates a value close to 1 on the AUC and close to 0 in terms of the RMSE.

6.2 Dataset

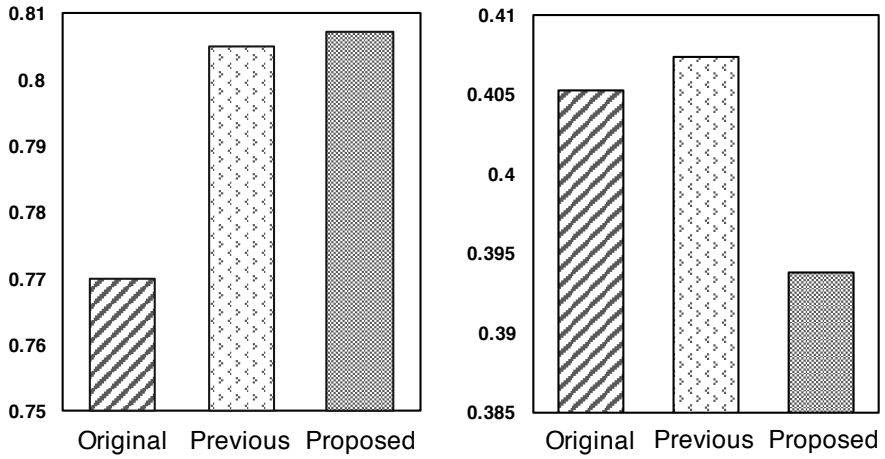
In this experiment, we applied three methods to two datasets of synthetic data and the *Bridge to Algebra 2006-2007* [7]. Table 1 presents an overview of each dataset.

Table 1. Details of datasets.

	Records	Students	Items	Skills
Synthetic	200,000	1,000	25	5
Algebra	225,880	1,127	612	114

(1) Synthetic data We employed IRT to generate the synthetic data. We assumed that if an item was assigned to a student once, the student's skill to solve the item increased. In order to add a decay effect, we calculated the retention interval between the initial presentation of an item to a student and a later recall assignment. If the elapsed time was long, the student's skill to solve the item decreased.

(2) Bridge to Algebra 2006-2007 This dataset was used at the KDD Cup 2010 Educational Data mining Challenge as actual data from an e-Learning system. We omitted items that have less than 200 records and items requiring a defined skill to be solved.



(a) AUC (b) RMSE

Fig. 2. Results with synthetic data.

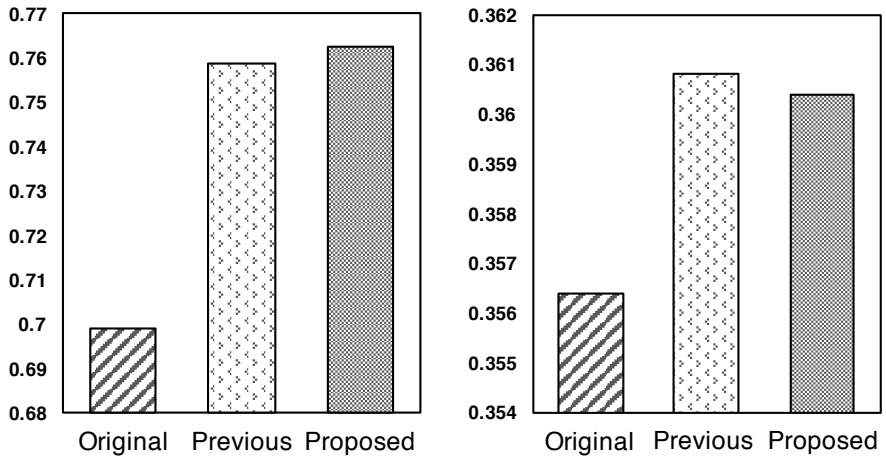
6.3 Results

(1) Synthetic data Figure 2 shows the prediction results for each method for synthetic data. The graphs show (i) original method (knowledge tracing), (ii) previous method (knowledge tracing and IRT), and (iii) the proposed method (knowledge tracing and IRT with forgetting) from the left in Figure 2. The values of the AUC of the previous method and the proposed method were greater than that for original knowledge tracing in Figure 2(a). There was no significant difference between the previous method and the proposed method. However, the value of RMSE in Figure 2(b) shows that the proposed method has superior prediction ability than the previous methods. Therefor, the results indicated that the proposed method was the most effective.

(2) Bridge to Algebra 2006-2007 Figure 3(a) shows the prediction results for each method on actual data. Our proposed methods yielded the best performance, whereas there was slight difference between the results for the proposed method and the previous method. However, the value of RMSE of the proposed method indicated lower than previous method in Figure 3(b).

7 Conclusion

In this paper, we proposed a novel combination of knowledge tracing and IRT with a decay effect in order to improve the previous method. The proposed approach showed promising effectiveness on real-world datasets.



(a) AUC (b) RMSE

Fig. 3. Results of Bridge to Algebra 2006-2007.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16K01095.

References

1. T. Calders, M. Pechenizkiy, Introduction to The Special Section on Educational Data Mining, SIGKDD, Vol. 13, Issue. 2, pp. 3–5, 2011.
2. A. T. Corbett, J. R. Anderson, Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge, User Modeling and User-Adapted Interaction, 4(4), pp. 253–278, 1995.
3. S.E.Levinson, L.R. Rabiner, M.M. Sondhi, An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, Bell System Technical Journal, Vol. 62, Issue. 4, pp. 1035–1074, 1983.
4. Wim J. van der Linden, Ronald K. Hambleton, Handbook of Modern Item Response Theory, Springer, 1996.
5. R.V. Lindsey, M.C. Mozer, Predicting Individual Differences in Student Learning via Collaborative Filtering, Submitted, 2014.
6. M. Khajah, Y. Huang, J. P. González-Brenes, M. C. Mozer, and P. Brusilovsk, Integrating Knowledge Tracing and Item Response Theory: A Tale of Two Frameworks, Proceedings of Workshop on Personalization Approaches in Learning Environments (PALE2014) at the 22th International Conference on User Modeling, Adaptation, and Personalization, pp. 7–12, 2014.
7. J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, K. R. Koedinger, Bridge to Algebra 2006-2007, Development data set from KDD Cup 2010 Educational Data Mining Challenge, (`http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp`).