Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities

William J. Hawkins[1], Neil T. Heffernan[1], Ryan S.J.d. Baker[2]

[1]Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA
{bhawk90, nth}@wpi.edu
[2]Department of Human Development, Teachers College, Columbia University, New York, NY
baker2@exchange.tc.columbia.edu

Abstract. Student modeling is an important component of ITS research because it can help guide the behavior of a running tutor and help researchers understand how students learn. Due to its predictive accuracy, interpretability and ability to infer student knowledge, Corbett & Anderson's Bayesian Knowledge Tracing is one of the most popular student models. However, researchers have discovered problems with some of the most popular methods of fitting it. These problems include: multiple sets of highly dissimilar parameters predicting the data equally well (identifiability), local minima, degenerate parameters, and computational cost during fitting. Some researchers have proposed new fitting procedures to combat these problems, but are more complex and not completely successful at eliminating the problems they set out to prevent. We instead fit parameters by estimating the mostly likely point that each student learned the skill, developing a new method that avoids the above problems while achieving similar predictive accuracy.

Keywords: Bayesian Knowledge Tracing · Expectation Maximization · Student Modeling

1 Introduction

Within the field of Intelligent Tutoring Systems (ITSs), student modeling is important because it can help guide interaction between a student and an ITS. By having a model of student knowledge, an ITS can estimate how knowledgeable a student is of various knowledge components (or "skills") over time and use that to determine what the student needs to practice.

However, student modeling is also important to researchers. The parameters learned from BKT can be used to characterize how students learn and to evaluate ITS content. Examples of this include studying the effects of "gaming the system" on learning [8] and evaluating hint helpfulness [4], among many other studies.

While BKT is popular and useful, researchers have found problems with fitting BKT models. One such problem is identifiability: there may be multiple sets of parameters that fit the data equally well [3], making interpretation difficult. Additionally, the learned parameters may produce what is called a degenerate model, or a model

adfa, p. 1, 2011.

that fits the data well but violates the assumptions of the approach, generally leading to inappropriate pedagogical decisions if used in a real system [1].

Two popular fitting methods in the literature, Expectation-Maximization (EM) [9] and brute force grid search, both suffer from identifiability. Additionally, EM can get stuck on local minima, and brute force comes with a high computational cost.

Researchers have attempted to deal with these issues through strategies like limiting the values brute force searching can explore [2], determining which starting values lead to degenerate parameters in EM [12], computing Dirichlet priors for each parameter and using these to bias the search [13], clustering parameters across similar skills [14], and using machine-learned models to detect two of the parameters [1].

This work introduces a simple method of estimating BKT parameters that sacrifices the precision of optimization techniques for the efficiency and interpretability of empirical estimation. Briefly, we estimate when students learn skills heuristically, and then use these estimates to help compute the four BKT parameters. Our goal is to efficiently produce accurate, non-degenerate BKT models.

2 Data

For this work, we used data from ASSISTments [7], an ITS used primarily by middle- and high-school students. In this dataset taken from the 2009-10 school year, 1,579 students worked on 61,522 problems from 67 skill-builder problem sets. The skill-builders used had data from at least 10 students, used default mastery settings (three consecutive correct answers to achieve mastery, ending the assignment), and had at least one student achieve mastery. A student's data was only included for a specific skill-builder if they answered at least three questions.

3 Methods

In this work, we developed and analyzed a new fitting procedure for BKT. We begin this section by describing BKT and then introduce our empirical approach to fitting BKT models. Finally, we describe the analyses we performed.

3.1 Bayesian Knowledge Tracing

Bayesian Knowledge Tracing [5] is a student model used in ITS research that infers a student's knowledge given their history of responses to problems, which it can use to predict future performance. Typically, a separate BKT model is fit for each skill. It assumes that a given student is always either in the known state or the unknown state for a given skill, with a certain probability of being in each. To calculate the probability that a student knows the skill given their performance history, BKT needs to know four probabilities: $P(L_0)$, the probability a student knows the skill before attempting the first problem; P(T), the probability a student who does not currently know the skill will know it after the next practice opportunity; P(G), the probability a student will answer a question correctly despite not knowing the skill; and P(S), the probability a student will answer a question incorrectly despite knowing the skill.

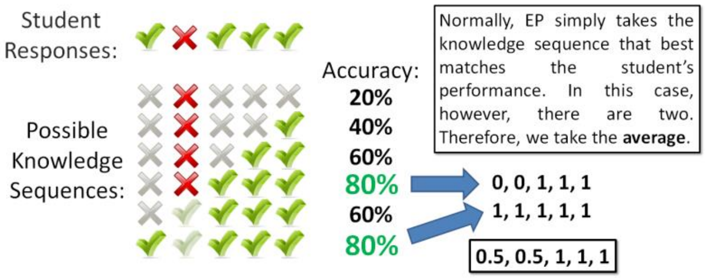
According to this model, knowledge affects performance (mediated by the guess and slip rates), and knowledge at one time step affects knowledge at the next time step: if a student is in the unknown state at time *t*, then the probability they will be in the known state at time *t+1* is P(T). Additionally, BKT models typically assume that forgetting does not occur: once a student is in the known state, they stay there.

3.2 Computing Knowledge Tracing Using Empirical Probabilities

In this section, we present a new approach to fitting BKT models we call Empirical Probabilities (EP). EP is a two-step process that involves annotating performance data with knowledge, and then using this information to compute the BKT parameters.

Annotating Knowledge. The first step in EP is to annotate performance data for each student within each skill with an estimate of when the student learned the skill. We assume there are only two knowledge states: known (1) and unknown (0), and do not allow for forgetting (a known state can never be followed by an unknown state).

In this work, we use a simple heuristic for determining when a student learns a skill: we choose the knowledge sequence that best matches their performance. This is illustrated by Figure 1. A full description of this heuristic can be found online [6].



Normally, EP simply takes the knowledge sequence that best matches the student's performance. In this case, however, there are two. Therefore, we take the **average**.

Fig. 1. Each of the six possible knowledge sequences are tried for a student's performance history, and in this case, the best two are averaged together to get the final sequence.

Computing the Probabilities. Using the knowledge estimates, we were able to compute each of the four BKT parameters for each skill empirically from the data.

The first of these parameters is P(L_0), the probability that the student knew the skill before interacting with the system. We can empirically estimate this by taking the average value of student knowledge on the first practice opportunity:

$$P(L_0) = \frac{\sum K_0}{|K_0|} \tag{1}$$

Equation (1) is similar to a heuristic in [11] for estimating individual student prior knowledge. While that paper used performance to compute a prior for each student as opposed to using knowledge to compute a prior for each skill as we do here, the idea that prior knowledge can be estimated mathematically in this way is similar.

Using K_i and C_i as knowledge and correctness at problem *i*, respectively, the following equations are used to compute the other three BKT parameters:

$$P(T) = \frac{\sum_{i \neq 0}(1-K_{i-1})K_i}{\sum_{i \neq 0}(1-K_{i-1})} \tag{2}$$

$$P(G) = \frac{\sum_i C_i(1-K_i)}{\sum_i(1-K_i)} \tag{3}$$

$$P(S) = \frac{\sum_i(1-C_i)K_i}{\sum_i K_i} \tag{4}$$

3.3 Experiments

In this paper, we compare BKT models fit with EM and EP in terms of predictive accuracy, model degeneracy, and training time. Due to space constraints, only the predictive accuracy results are reported here. Results for the other experiments as well as the code and data used in all the experiments are available online [6].

To fit EM, we used Murphy's Bayes Net Toolbox for MATLAB (BNT) [10]. For EM, it is necessary to specify a starting point. We chose an initial $P(L_0)$ of 0.5, and set the other three parameters to 0.1. Additionally, we set a maximum of 100 iterations and used the default BNT improvement threshold value of 0.001.

To compute the parameters using EP, we implemented the equations in the previous section in MATLAB using basic functionality. Then, we entered these values into the conditional probability tables of a BKT model constructed with BNT.

4 Results

First, we examine how predictive each method is of student performance under five-fold student-level cross-validation. We evaluated the methods using mean absolute error (MAE), root mean squared error (RMSE), and A'. These metrics were computed for each student and then used in two-tailed paired t-tests to determine the significance of the differences between the overall means of the two models. The degrees of freedom for the MAE and RMSE significance tests was one less than the number of students, whereas that of the A' significance test was lower due to some students being excluded (students who gave all correct or all incorrect answers for all skills were excluded since A' is undefined in such cases). The values below represent the average of the student metrics. Lower values of MAE and RMSE indicate better performance, whereas the opposite is true of A'. The results are shown in Table 1.

Table 1. Prediction results for the two methods of learning BKT parameters: Expectation Maximization and Empirical Probabilities

Learning Method	MAE	RMSE	A'
EM (BNT)	0.3830	0.4240	0.5909
EP	0.3742	0.4284	0.6145

Although the differences between these metrics are all statistically significant according to two-tailed paired t-tests (MAE: t(1,578) = 10.88, RMSE: t(1,578) = -6.74, A': t(1,314) = -7.01, p < 0.00001), the differences are small. Therefore, we believe the two methods are comparable in terms of predicting performance.

We also tested EM and EP in terms of model degeneracy and fitting time. In summary, we found that only EM learned degenerate parameters, and that EP runs significantly faster than EM. The full results are available online [6].

5 Conclusions and Future Work

From this work, it appears that a simple estimation of knowledge followed by computing empirical probabilities may be a reasonable approach to estimating BKT parameters. We found that EP had comparable predictive accuracy to that of EM. Additionally, it is mathematically impossible for EP to learn theoretically degenerate guess and slip rates (i.e. above 0.5) [6], and it is at least as good as EM at avoiding empirically degenerate parameters, based on tests suggested and used in [1]. We also found it was considerably faster than EM [6].

An improvement to EP would be to annotate knowledge more probabilistically. EP makes only binary inferences of knowledge based on predictive performance. For example, EP always considers incorrect responses on the first problem to be made in the unknown state, even though some of these are slips. Therefore, a more probabilistic approach may be able to produce better parameter estimates.

EP could be used as a tractable way to help improve accuracy by incrementally incorporating data into models as it becomes available during a school year. This would improve models for skills with little or no previous data and make use of student and class information. If a skill has little or no previous data, using current school year data may improve estimates of its parameters. Also, it has been shown that incorporating student [11] and class [15] information can improve predictive performance, which cannot be done before the start of a school year.

While EP achieves similar accuracy to EM and appears not to learn degenerate parameters, we did not perform any external validations of the learned parameters for either approach. Such an analysis would help determine how much we can trust EP parameters, especially when they differ from those learned by EM.

Acknowledgements. We acknowledge funding from NSF (#1316736, 1252297, 1109483, 1031398, 0742503), ONR's 'STEM Grand Challenges' and IES (# R305A120125 & R305C100024).

References

1. Baker, R.S.J.d., Corbett, A. T., Aleven, V.: More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In:

Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.) ITS 2008. LNCS, vol. 5091/2008, pp. 406-415. Springer, Berlin Heidelberg (2008)

2. Baker, R.S.J.d., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. M., Kauffman, L. R., Mitchell, A. P., Giguere, S.: Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. In: De Bra, P., Kobsa, A., Chin, D. (Eds.) UMAP 2010. LNCS, vol. 6075/2010, pp. 52-63. Springer-Verlag, Berlin Heidelberg (2010)
3. Beck, J. E., Chang, K. M.: Identifiability: A fundamental problem of student modeling. In: Conati, C., McCoy, K., Paliouras, G. (Eds.) UM 2007. LNCS, vol. 4511/2007, pp. 137-146. Springer, Berlin (2007)
4. Beck, J.E., Chang, K., Mostow, J., Corbett, A.: Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In: Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.) ITS 2008. LNCS, vol. 5091/2008, pp. 383-394. Springer, Berlin Heidelberg (2008)
5. Corbett, A., Anderson, J.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278 (1995)
6. Empirical Probabilities. https://sites.google.com/site/whawkins90/publications/ep
7. Feng, M., Heffernan, N. T., Koedinger, K. R.: Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. User Modeling and User-Adapted Interaction, 19, 243-266 (2009)
8. Gong, Y., Beck, J., Heffernan, N., Forbes-Summers, E.: The impact of gaming (?) on learning at the fine-grained level. In: Aleven, V., Kay, J., Mostow, J. (Eds.) ITS 2010. LNCS, vol. 6094/2010, pp. 194-203. Springer (2010)
9. Moon, T. K.: The expectation–maximization algorithm. IEEE Signal Process. Mag., 13, 47–60 (1996)
10. Murphy, K.: The bayes net toolbox for matlab. Computing science and statistics, 33, 1024-1034 (2001)
11. Pardos, Z. A., Heffernan, N. T.: Modeling individualization in a bayesian networks implementation of knowledge tracing. In: De Bra, P., Kobsa, A., Chin, D. (Eds.) UMAP 2010. LNCS, vol. 6075/2010, pp. 255-266. Springer, Berlin Heidelberg (2010)
12. Pardos, Z. A., Heffernan, N. T.: Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In: Baker, R.S.J.d., Merceron, A., Pavlik, P.I. (Eds.) Proceedings of the 3rd International Conference on Educational Data Mining, pp. 161-170 (2010)
13. Rai, D., Gong, Y., Beck, J.: Using Dirichlet priors to improve model parameter plausibility. In: Barnes, T., Desmarais, M., Romero, C., Ventura, S. (Eds.) Proceedings of the 2nd International Conference on Educational Data Mining, pp. 141-150 (2009)
14. Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R.C.: Reducing the Knowledge Tracing Space. In: Barnes, T., Desmarais, M., Romero, C., Ventura, S. (Eds.) Proceedings of the 2nd International Conference on Educational Data Mining, pp. 151-160 (2009)
15. Wang, Y., Beck, J.: Class vs. Student in a Bayesian Network Student Model. In: Lane, H. C., Yacef, K., Mostow, J., Pavlik, P. (Eds.) AIED 2013. LNCS, vol. 7926/2013, pp. 151-160. Springer (2013)