Individualized Bayesian Knowledge Tracing Models

Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon

Carnegie Mellon University,
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{yudelson,koedinger}@cmu.edu,ggordon@cs.cmu.edu

Abstract. Bayesian Knowledge Tracing (BKT)[1] is a user modeling method extensively used in the area of Intelligent Tutoring Systems. In the standard BKT implementation, there are only skill-specific parameters. However, a large body of research strongly suggests that student-specific variability in the data, when accounted for, could enhance model accuracy [5, 6, 8]. In this work, we revisit the problem of introducing student-specific parameters into BKT on a larger scale. We show that student-specific parameters lead to a tangible improvement when predicting the data of unseen students, and that parameterizing students' speed of learning is more beneficial than parameterizing a priori knowledge.

Keywords: Bayesian knowledge tracing, model fitting, model selection, student-specific model parameters

1 Introduction

Modeling student knowledge as a latent variable is a popular approach. The latent variable is updated based on the correctness of the observed student opportunities to apply the skill in question. In general case, this modeling approach is called a Hidden Markov Model. A special case of the approach is known as Bayesian Knowledge Tracing (BKT) [1]. BKT assumes that student knowledge is represented as a set of binary variables – one per skill (the skill is either mastered by the student or not). Observations in BKT are also binary: a student gets a problem [step] either right or wrong.

BKT has a long history of being actively used in Intelligent Tutoring Systems (ITS) in the context of mastery learning and problem sequencing. In its standard implementation that is still in predominant use today, BKT only has skill-specific parameters. Starting with the original publication on BKT [1] and including more recent works (e.g. [5]), there exist strong indicators that BKT models (often called individualized BKT models) that somehow account for student variance are superior to the standard BKT model.

Prior work on individualized BKT models (e.g. [1], [5]), and [8]) describes quite different approaches to defining and learning student-specific parameters as well as report radically different performance measures. In this paper, we

approach the problem of introducing student-specific parameters in a more systematic manner. We build several individualized BKT models in an incremental manner (adding student-specific parameters in batches) and examine the effect each addition has on the model's cross-validation performance.

We find that BKT parameters corresponding to the a priori student knowledge give BKT models only a marginal cross-validation performance improvement. At the same time, student-specific speed of learning parameters result in a considerable boost in the model prediction accuracy.

2 Related Work

2.1 Bayesian Knowledge Tracing

There are four types of model parameters used in Bayesian Knowledge Tracing: initial probability of knowing the skill a priori – $p(L_0)$ (or *p-init*), probability of student's knowledge of a skill transitioning from *not known* to *known* state after an opportunity to apply it – $p(T)$ (or *p-transit*), probability to make a mistake when applying a known skill – $p(S)$ (or *p-slip*), and probability of correctly applying a not-known skill – $p(G)$ (or *p-guess*). Given that parameters are set for all skills, the formulae used to update student knowledge of skills are as follows. The initial probability of student u mastering skill k is set to the p-init parameter for that skill Equation (1a). Depending on whether the student u applied skill k correctly or incorrectly, the conditional probability is computed either using Equation (1b) or Equation (1c). The conditional probability is used to update the probability of skill mastery according to Equation (1d). To compute the probability of student u applying the skill k correctly on an upcoming practice opportunity one uses Equation (1e).

$$p(L_1)_u^k = p(L_0)^k, \tag{1a}$$

$$p(L_{t+1}|obs = correct)_u^k = \frac{p(L_t)_u^k \cdot (1 - p(S)^k)}{p(L_t)_u^k \cdot (1 - p(S)^k) + (1 - p(L_t)_u^k) \cdot p(G)^k}, \tag{1b}$$

$$p(L_{t+1}|obs = wrong)_u^k = \frac{p(L_t)_u^k \cdot p(S)^k}{p(L_t)_u^k \cdot p(S)^k + (1 - p(L_t)_u^k) \cdot (1 - p(G)^k)}, \tag{1c}$$

$$p(L_{t+1})_u^k = p(L_{t+1}|obs)_u^k + (1 - p(L_{t+1}|obs)_u^k) \cdot p(T)^k, \tag{1d}$$

$$p(C_{t+1})_u^k = p(L_t)_u^k \cdot (1 - p(S)^k) + (1 - p(L_t)_u^k) \cdot p(G)^k \tag{1e}$$

In the standard BKT model, we use one copy of each of the above four parameters $\langle\, p(L_0), p(T), p(S), p(G) \,\rangle$ per skill. BKT models are usually fit using the expectation maximization method (EM) [2], Conjugate Gradient Search [1], or discretized brute-force search [7].

2.2 Student-specific Parameters in Bayesian Knowledge Tracing

In the area of building cognitive models of practice, student-specific parameters have been used for quite some time. The logistic regression based Rasch model [3]

(also known as 1PL IRT) and its descendant the Additive Factors Model [6] both include a 'student proficiency' parameter to account for variability in student a priori abilities. In our prior work, we found that the inclusion of student-specific parameters has a significant positive effect on prediction accuracy and interpretability, as well as reduces over-fitting [4].

Prior work introducing student-specific parameters to BKT is limited. Corbett and Anderson, in the original BKT paper [1], discussed fitting all four BKT parameters for students (e.g. $p(T)_u$) as well as skills (e.g. $p(T)^k$). Namely, data of all students practicing skill k would be used to fit four BKT parameters for that skill, and all data of student u would be used to fit four BKT parameters for that student. The student and skill parameters would then be combined using a special function to yield a value (here $p(T)^k_u$) to be used for updating the probability of skill mastery. The individualized BKT model led to better correlation between actual and expected accuracy across students when compared to the same correlation for the non-individualized BKT model. However, accuracy of predicting student test scores (after a period of working with a tutoring system) did not improve tangibly.

Pardos and Heffernan [5] individualized the initial probability of mastery $p(L_0)^k$ by assigning according to a set of heuristics: randomly, by selecting from two pre-set values based on first student response correctness, by using overall percent correct. The 'prior per-student' models fit better than traditional BKT on a significant fraction of the problem sets authors considered.

Lee and Brunskill [8] investigated individualizing all four BKT parameters. However, in contrast to [1], the student-specific parameters were fit differently. Instead of fitting per skill and per-student BKT parameters to be combined later, they only fit per-student parameters for each student (assuming there is one skill all students have to learn). Lee and Brunskill did not discuss goodness of fit of their individualized models. Their focus was whether the individualized model, when used in an intelligent tutoring system, would schedule fewer or more practice opportunities than the traditional BKT skill-specific model (or *population* model as authors referred to it). The results showed that a considerable fraction of students, as judged by individualized model, would have received too few or too many practice opportunities (although no confidence intervals were given).

Although the [potential] benefits of individualized BKT models are visible, the results are unclear about the ideal configuration of student-specific parameters (4 per student [1], 1 heuristic value per student [5], 4 per student [8]), are limited in the evidence for improved mode prediction and are hard to operationalize for the purpose of implementing in an ITS. The original work on BKT [1] pointed out that operationalization of the discussed individualized BKT model could be problematic. Work by Pardos and Heffernan [5] showed that their prior-per-student BKT does not always win over traditional BKT. Lee and Brunskill [8] made a practically important derivation that using individualized model parameters could save time for stronger students and could allocate more time for struggling ones. However, this derivation assumed that individualized BKT models predict student data better which was not tested.

Table 1: BKT parameters in matrix form

(a) Priors (Π)

known	$p(L_0)$
unknown	1-$p(L_0)$

(b) Transitions (A)

	to known	to unknown
from known	1	0
from unknown	$p(T)$	$1-p(T)$

(c) Observations (B)

	right	wrong
known	1-$p(S)$	$p(S)$
unknown	$p(G)$	1-$p(G)$

3 Methods

Our goal is to unify and extend prior work on individualized BKT models. We construct four variants of individualized BKT models varying the number of student-specific parameters. and we rank the constructed models with respect to predictive accuracy on unseen data.

3.1 Bayesian Knowledge Tracing with Student-specific Parameters

Instead of a traditional Expectation Maximization (EM) method for learning BKT parameters, we base our method on the so-called *optimization techniques* approach described in [2] for the following reasons. First, EM does not directly optimize a likelihood of the student observations given BKT parameters (a standard metric for HMM). As a result, the EM algorithm could make adjustments to BKT parameters that would actually worsen the fit. Second, using the gradient-based optimization techniques allows us to introduce student-specific parameters to BKT without expanding the structure of the underlying HMM (cf. [5]). Keeping the structure of the underlying HMM unchanged permits us to lower the computational cost of fitting.

Table 1 shows BKT parameters defined in matrix format, as they are normally represented in HMM. A priori probability of mastery $p(L_0)$ belongs in the *Priors* matrix $\Pi = \{\pi_i\}$ in an HMM, $i \in [1, N]$ (N is the number of hidden states, in our case two), learning probability $p(T)$ is in the *Transitions* matrix $A = \{a_{ij}\}$, $i, j \in [1, N]$ (note that there is no forgetting – transition from known to unknown is zero), probabilities of slipping and guessing belong to the *Observations* matrix $B = \{b_j(m)\}$, $j \in [1, N]$, $m \in [1, M]$ (M is the number of observations, in our case two). These matrices follow two constraints: all of the elements should be non-negative, and the priors vector and the rows of transitions and observations matrices should sum to one.

To successfully implement our BKT models, we need to solve two problems. First, the *evaluation problem*: given BKT parameters $\lambda = \{\Pi, A, B\}$ and a sequence of observations (practice attempts) $O = \{o_t\}$, $t \in [1, T]$, what is the probability that the observations are generated given BKT model, or formally $p\{O|\lambda\}$. Second, the *learning problem*: given BKT parameters λ and a sequence of observations O, how should λ be adjusted to maximize $p\{O|\lambda\}$.

The objective function we use in our method is negative log likelihood, or $J = -log(L_{tot})$, where L_{tot} is the sum of all likelihoods $p\{O|\lambda\}$ for all student-skill

practice sequences in our data. We will define our search for better λ parameters of the BKT as gradient search (cf. Equation 2a, where η is the search step size). Here, gradients with respect to our matrices from Table 1 are defined in terms of the so-called *forward* variables α (cf. Equation 2b and 2c) and *backward* variables β (cf. Equation 2d and 2e). Gradients with respect to BKT parameters are given in Equation 2f, 2g, and 2h. For detailed discussion of forward and backward variables as well as derivations of the gradients see [2].

$$\lambda^{new} = \lambda^{old} - \eta \left[\frac{\partial J}{\partial \lambda}\right]_{\lambda=\lambda old} \tag{2a}$$

$$\alpha_1(j) = \pi_j b_j(o_1), j \in [1, N] \tag{2b}$$

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}, j \in [1, N], t \in [1, T] \tag{2c}$$

$$\beta_T(i) = 1, i \in [1, N] \tag{2d}$$

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}), i \in [1, N], t \in [1, T-1] \tag{2e}$$

$$\frac{\partial J}{\partial \pi_i} = -\frac{1}{L_{tot}} \beta_1(i) b_i(o_1) \tag{2f}$$

$$\frac{\partial J}{\partial a_{ij}} = -\frac{1}{L_{tot}} \sum_{t=2}^{T} \beta_t(j) b_j(o_t) \alpha_{t-1}(i) \tag{2g}$$

$$\frac{\partial J}{\partial b_j(o_t)} = -\frac{1}{L_{tot}} \frac{\alpha_t(j)\beta_t(j)}{b_j(o_t)} \tag{2h}$$

$$\tag{2i}$$

We have defined how to compute gradients with respect to traditional BKT parameters. To introduce student-specific parameters we *split* the skill-specific BKT parameters into two components the following way. Using w to substitute for each of the corresponding skill-specific BKT parameters (π_i, a_{ij}, or $b_j(m)$), we define it in terms of both student- and skill-specific parameters as shown in Equation 3a. Here, w^k is the skill-specific component of the parameter, w^u is the student-specific component, $l(p) = log[p/(1-p)]$ is a logit function, and $\sigma(x) = 1/(1+e^{-x})$ is a sigmoid function (inverse of logit). Not that in summing logistic functions in Equation 3b to combine student and skill parameters we are incorporating the compensatory logic behind the IRT and AFM family of models [3, 6]. Updating parameter gradients is possible using the chain rule (illustrated in Equation 3b for the student-specific component of the parameter w), since both the sigmoid and logit functions are differentiable.

$$w = \sigma(l(w^k) + l(w^u)) \tag{3a}$$

$$\frac{\partial J}{\partial w^u} = \frac{\partial J}{\partial w}\frac{\partial w}{\partial w^u} \tag{3b}$$

The importance of having all the gradients' derivations in Equations 2f to 2h is two-fold. First of all, freely available specialized HMM toolkits usually target general purpose Bayesian inference algorithms (most often EM) that are more computationally intensive. Second, without computing the gradients explicitly, a general-purpose optimization packages (part of tools like Matlab and R) would have to make computationally inefficient approximations.

3.2 Data

We used the datasets from the KDD Cup 2010 Educational Datamining Challenge (`http://pslcdatashop.web.cmu.edu/KDDCup`). The data was donated by Carnegie Learning Inc., a publisher of math curricula and a producer of intelligent tutoring systems for middle school and high school. There are two datasets, Algebra I, and Bridge to Algebra, both collected in 2008-2009 school year. Each dataset is a log of students' step-by-step performance (correctness and timing) during problem solving and was tagged with two alternative skill models.

The Algebra I dataset has 8,918,054 rows covering practice attempts of 3,310 students. 4,419,705 rows of the Algebra I dataset are tagged with 515 distinct skills from skill model 1 (used for problem sequencing in an ITS) and 6,442,137 rows are tagged with 541 distinct skills from an alternative skill model 2. The Bridge to Algebra dataset contains data of 6,043 students comprised of 20,012,498 rows, 11,239,188 and 12,350,449 of which are tagged with skills from skill model 1 (807 distinct skills) and model 2 (933 distinct skills) respectively. It is worth underlining the sheer size of each of the datasets. Except for the prior-per-student model reported in [5], none of the BKT models were ever tried on the dataset of this size, and prior-per-student has been individualized by using simple heuristics including random, correctness of first response defines the choice of one of two pre-set priors, and overall per-student percent correct.

3.3 Fitting Procedures

We created a tool capable of fitting and cross-validating standard and individualized BKT models using the derivations discussed in Section 3.1. To facilitate efficiency, it was implemented in C/C++. The tool is capable of fitting classical BKT models using the EM method, as well as fitting classical and individualized BKT models using the gradient descent method (using linear step size search) and a set of versions of conjugate gradient descent method.

We tested four different model variants on four different dataset-skill model combinations. We chose gradient descent method, since, although conjugate gradient methods are expected to yield better fits, the actual advantage was minimal to non-existent. When fitting individualized models, the coordinate descent

method was used: two blocks of parameters – skill-specific and student-specific – by interleaving fits if one block at a time. The BKT model variants we fit were:

1. Standard BKT model,
2. Individualized BKT with student-specific $p(L_0)$,
3. Individualized BKT with student-specific $p(T)$,
4. Individualized BKT with student-specific $p(L_0)$ and $p(T)$.

While constructing the models, we constrained model values for all guess and slip parameters to prevent the occurrence of a phenomenon called model degeneracy (cf. [7]). All of the models were cross-validated using 10 randomly assigned user-stratified folds. For each of the cross-validation results we computed root mean squared error (RMSE) and accuracy (number of correctly predicted student successes and failures).

Our tool is implemented to handle large datasets in an efficient manner. For example, 10-fold cross-validation of the simplest standard BKT model on Algebra I dataset with skill model 1 takes under 2.5 minutes, for the most complex model 4 in the list above on the larger Bridge to Algebra dataset and skill model 2 the running time is under 70 minutes.

4 Results

Table 2 is a summary of cross-validation results for the standard BKT and the three individualized BKT models. For each dataset - skill model pair, in addition to RMSE and Accuracy, the contrasts to other BKT model variants are given in terms of fewer/more correct predictions. The correctness is computed using model's prediction (rounded toward 0 or 1 using 0.5 as threshold) and the actual correctness of student step in the data.

Across both datasets and both skill models, student-specific a priori probability of mastery ($p(L_0)$) in model 2 has no effect on model performance. On the other hand, introduction of student specific speed of learning ($p(T)$) in model 3 results in a consistent and more pronounced advantage over models 1 and 2. Moreover, the improvement in model accuracy resulting from adding individualized $p(L_0)$ on top of individualized $p(T)$ (going from model 3 to model 4) is even smaller than when adding individualized $p(L_0)$ to the standard BKT model (going from model 1 to model 2), despite the fact that model 3 has half as many student specific parameters as model 4. Given that, model 3 with individualized $p(T)$ can be considered superior to the standard BKT and other individualized models.

Bear in mind that results in Table 2 are for student-stratified validation. Namely, individualized BKT models are making predictions on data from unseen students unable to use their learnt student-specific parameters. Considering a potential operationalization of our findings, this shows a valuable property of model 3 (and model 4): producing *cleaner* skill-specific parameters (read, devoid of student-specific noise/variability). In an incremental ITS design cycle it would mean that, even if the core system only has a standard BKT implemented, it is

Table 2: Model cross-validation statistics for datasets Algebra I (A) and Bridge to Algebra (B) and skill models 1 and 2. Subscripts next to RMSE and Accuracy denote respective rank. The correct predictions difference tables show how many more correct predictions a model in the row makes over the model in the column header (a negative number means a model makes fewer correct predictions).

(a) Dataset A, skill model 1

model	RMSE	Accuracy	Correct rows	Correct predictions difference model 1	model 2	model 3	model 4
1	0.36273^{4}	0.827550^{3}	3,657,527	0	348	-6232	-5972
2	0.36265^{3}	0.827471^{4}	3,657,179	-348	0	-6580	-6320
3	0.36116^{1}	0.828960^{1}	3,663,759	6232	6580	0	260
4	0.36119^{2}	0.828901^{2}	3,663,499	5972	6320	-260	0

(b) Dataset A, skill model 2

model	RMSE	Accuracy	Correct rows	Correct predictions difference model 1	model 2	model 3	model 4
1	0.34187^{4}	0.84914^{3}	5,470,279	0	783	-6390	-6594
2	0.34180^{3}	0.84902^{4}	5,469,496	-783	0	-7173	-7377
3	0.34065^{2}	0.85013^{2}	5,476,669	6390	7173	0	-204
4	0.34060^{1}	0.85016^{1}	5,476,873	6594	7377	204	0

(c) Dataset A, skill model 1

model	RMSE	Accuracy	Correct rows	Correct predictions difference model 1	model 2	model 3	model 4
1	0.36294^{4}	0.82261^{4}	9,245,493	0	-6638	-78249	-76805
2	0.36255^{3}	0.82320^{3}	9,252,131	6638	0	-71611	-70167
3	0.35851^{1}	0.82957^{1}	9,323,742	78249	71611	0	1444
4	0.35854^{2}	0.82945^{2}	9,322,298	76805	70167	-1444	0

(d) Dataset A, skill model 2

model	RMSE	Accuracy	Correct rows	Correct predictions difference model 1	model 2	model 3	model 4
1	0.35895^{4}	0.82757^{4}	10,220,891	0	-7122	-78339	-77993
2	0.35857^{3}	0.82815^{3}	10,228,013	7122	0	-71217	-70871
3	0.35484^{2}	0.83392^{2}	10,299,230	78339	71217	0	346
4	0.35482^{1}	0.83389^{1}	10,298,884	77993	70871	-346	0

possible to improve overall student model accuracy by incrementally updating the skill-specific weights once a new group of students finishes a course or a course unit.

5 Conclusions

In this paper we presented an approach to building individualized Bayesian Knowledge Tracing models that are capable of accounting for student differences with respect to initial mastery probabilities and skill learning probabilities. Our approach does not require the underlying Hidden Markov Model to be changed. It is based on gradients of prior (Π), transition (A), and observation (B) parameter matrices and can be used together with a wide range of existing gradient descent algorithms. Our own implementation includes a conjugate gradient method with a variety of kernel formulas for computing the direction of parameter updates.

As we were able to show, our implementation of individualized BKT models is capable of tangibly improving the accuracy of predicting the success of student work in an intelligent tutoring system. An interesting finding was that adding student-specific probability of learning (*pLearn*) is more beneficial for the model accuracy than adding student-specific probability of initial mastery (*pInit*). In an alternative realm of models of learning practice that are based on logistic regression (for example, Item Response Theory), the analog of initial probability of mastery is student proficiency, which is thought to be critical for the model performance. Could it be in those models that individualizing learning rate is better than individualizing proficiency.

It is out intent to continue developing the instrumental framework for fitting standard and individualized BKT models as well as to persist with its empirical evaluation on real-world and synthetic datasets. As part of this work we intend to include item-stratified and unstratified cross-validation to the currently implemented student-stratified and to extend individualization features to currently not covered observation matrix parameters – *pSlip* and *pGuess*.

Acknowledgments This research was supported by the Learnlab DataShop team, Carnegie Learning Inc., and National Science Foundation (NSF award #SBE-0836012).

References

1. Corbett, A. T. and Anderson, J. R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278. (1995)
2. Levinson, S. E., Rabiner, L. R., and Sondhi, M. M.: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. Bell System Technical Journal, 62(4): 1035-1074. (1983)
3. van der Linden, W. J. and Hambleton, R. K.: Handbook of Modern Item Response Theory. New York, NY: Springer. (1997)
4. Yudelson, M., Pavlik, P. I., and Koedinger, K. R.: User Modeling - A Notoriously Black Art. In: Konstan, J.A., Conejo, R., Marzo, J.-L., Olive, N. (eds.) Proceedings of the 19th International Conference on User Modeling, Adaption and Personalization (UMAP 2011), LNCS vol. 6787, pp. 317-328. Springer (2011)

5. Pardos, Z. A. and Heffernan, N. T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In: Paul De Bra, Alfred Kobsa, David N. Chin (eds.) Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2010), LNCS vol. 6075 pp. 255-266. Springer (2010)
6. Cen, H., Koedinger, K. R., and Junker, B.: Comparing Two IRT Models for Conjunctive Skills. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S.P. (eds.) Proceedings of the 9th International Conference On Intelligent Tutoring Systems (ITS 2008), LNCS vol. 5091, pp. 796-798. Springer (2008).
7. Baker, R. S. J., Corbett, A. T., and Aleven, V.: More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S.P. (eds.)Proceedings of the 9th International Conference On Intelligent Tutoring Systems (ITS 2008), LNCS vol. 5091, pp. 406-415. Springer (2008)
8. Lee, J. I. and Brunskill, E.: The Impact on Individualizing Student Models on Necessary Practice Opportunities. In: Yacef, K., Zaïane, O.R., Hershkovitz, A., Yudelson, M., Stamper, J.C. (eds.) Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012), pp. 118-125. (2012)