Using Item Response Theory to Refine Knowledge Tracing

Yanbo Xu
Carnegie Mellon University
RI-NSH 4105
5000 Forbes Ave, Pittsburgh, PA 15213
yanbox@cs.cmu.edu

Jack Mostow
Carnegie Mellon University
RI-NSH 4103
5000 Forbes Ave, Pittsburgh, PA 15213
mostow@cs.cmu.edu

ABSTRACT

Previous work on knowledge tracing has fit parameters per skill (ignoring differences between students), per student (ignoring differences between skills), or independently for each <student, skill> pair (risking sparse training data and overfitting, and under-generalizing by ignoring overlap of students or skills across pairs). To address these limitations, we first use a higher order Item Response Theory (IRT) model that approximates students' initial knowledge as their one-dimensional (or low-dimensional) overall proficiency, and combines it with the estimated difficulty and discrimination of each skill to estimate the probability *knew* of knowing a skill before practicing it. We then fit skill-specific knowledge tracing probabilities for *learn*, *guess*, and *slip*. Using synthetic data, we show that Markov Chain Monte Carlo (MCMC) can recover the parameters of this Higher-Order Knowledge Tracing (HO-KT) model. Using real data, we show that HO-KT predicts performance in an algebra tutors significantly better than fitting knowledge tracing parameters per student or per skill.

Keywords

Knowledge tracing, Item Response Theory, higher order models

1. Introduction

Traditional knowledge tracing (KT) [1] estimates the probability that a student knows a skill by observing attempted steps that require it, and applying a model with four parameters for each skill, assumed to be the same for all students: the probabilities *knew* of knowing the skill before practicing it, *learn* of acquiring the skill from one attempt, *guess* of succeeding at the attempt without knowing the skill, and *slip* of failing despite knowing the skill. Prior work shows that fitting such parameters for individual students can improve the model's accuracy in predicting student performance [2] or reduce unnecessary practice [3]. Such per-student parameters, however, ignore differences between skills. Fitting KT parameters separately instead for each <student, skill> pair risks sparse training data and overfitting, and under-generalizes by ignoring overlap of students or skills across pairs.

Item Response Theory (IRT) [4, 5] predicts a student's performance on an item based on the difficulty and discrimination of the skill(s) the item requires, and a one- (or low-) dimensional static estimate of the student's overall proficiency. Prior work adapted IRT to estimate the static probability of knowing a given skill [6], or dynamic changes in overall proficiency [7]. Here we *dynamically* estimate *individual skills* required in observed steps.

2. Approach

IRT's 2-Parameter Logistic model [4] estimates the probability $knew_{nj}$ of student n already knowing skill j as a logistic function of student proficiency θ_n, skill discrimination a_j, and difficulty b_j:

$$knew_{nj} = \frac{1}{1 + \exp\left(-1.7a_j(\theta_n - b_j)\right)}$$

Deriving $knew_{nj}$ instead of fitting it separately makes it a higher order model. We then fit each skill's KT parameters $learn_j$, $guess_j$, and $slip_j$. Figure 1 shows this hybrid Higher Order Knowledge Tracing (HO-KT) model's graphical representation. The observable state $Y^{(t)}$ tells if a skill is applied correctly at time t. The latent state $K^{(t)}$ models knowing it at time t; $\Pr(K^{(0)}) = knew$.

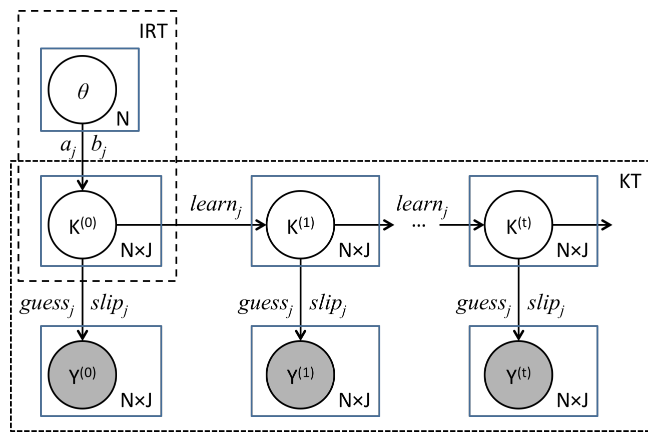

Figure 1. Graphical representation of Higher-Order Knowledge Tracing (HO-KT) model

For Markov Chain Monte Carlo (MCMC) estimation of HO-KT's parameters, we specify their prior distributions as follows:

$$\begin{aligned} \theta_n &\sim Normal(0,1) \\ b_j &\sim Normal(0,1) \\ a_j &\sim Uniform(0,2.5) \\ learn_j &\sim Beta(1,1) \\ guess_j &\sim Uniform(0,0.4) \\ slip_j &\sim Uniform(0,0.4) \end{aligned}$$

Given observations **Y**, MCMC finds vectors ***θ***, ***a***, ***b***, ***l (learn)***, ***g (guess)***, and ***s (slip)*** with maximum posterior probability, namely:

$$P(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{g}, \boldsymbol{s} | \boldsymbol{Y}) \propto L(\boldsymbol{Y} | \boldsymbol{g}, \boldsymbol{s}, \boldsymbol{K}) P\left(\boldsymbol{K}^{(0)} \middle| \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}\right) \times \prod_{t=1}^{T} P\left(\boldsymbol{K}^{(t)} \middle| \boldsymbol{K}^{(t-1)}, \boldsymbol{l}\right) P(\boldsymbol{\theta}) P(\boldsymbol{a}) P(\boldsymbol{b}) P(\boldsymbol{l}) P(\boldsymbol{g}) P(\boldsymbol{s})$$

HO-KT fits parameters to all data so far, in contrast to using IRT to fit ***θ***, ***a***, and ***b*** to early data and KT to fit ***l***, ***g***, and ***s*** to later data.

3. Experiment

We first generated synthetic data with N=100 students, each of whom practices J=4 skills required in a series of T=100 steps. We used OpenBUGS [8] to implement MCMC estimation for HO-KT in the BUGS language. We simultaneously ran the model in 5 chains for 10,000 iterations with a burn-in of 3000, each chain starting from randomly generated initial values, and considered MCMC to converge when all 5 chains overlapped in OpenBUGS' monitor window. Table 1 shows how well the estimated value of *learn* for each simulated skill recovered its true value; estimates of other parameters were similarly accurate but omitted here for lack of space. Moreover, MCMC correctly recovered 99.4% of the simulated students' 10,000 hidden binary knowledge states.

Table 1. Estimation of *learn* in synthetic data

Skill *j*	*learn*	Estimate (95% C.I.)	s.d.	MC_error
1	0.8	0.81 (0.48, 0.99)	0.13	0.006599
2	0.6	0.60 (0.52, 0.70)	0.05	0.002132
3	0.5	0.57 (0.38, 0.84)	0.11	0.005432
4	0.3	0.29 (0.25, 0.33)	0.02	7.79E-04

We then evaluated HO-KT on real data from the Algebra Cognitive Tutor® [9], containing a total number of 41,762 observations from 123 students performing 157 problem steps. Our model assumed each problem step required a single skill. We split the data evenly into training and test sets with no overlapping <student, skill> pairs. We limited the observed sequence length of each student to T=100, and still ran 5 chains starting from random initial values for 10,000 iterations with a burn-in of 3000.

For comparison, we also used BNT-SM [10] to fit knowledge tracing parameters per skill and per student to the algebra data. The data are unbalanced (85.10% are correct steps), so we also computed within-class and majority class accuracy. Table 2 compares the models' prediction accuracy and log-likelihood on the unseen test data. HO-KT is significantly higher in overall accuracy than KT per skill and per student, with $p<.0001$ in paired T-tests comparing HO-KT to the two KT models for each of 123 students. HO-KT also achieves the best log-likelihood.

Table 2. Evaluation on real data from algebra tutor

Model:	Accuracy: Overall	Accuracy: Correct	Accuracy: Incorrect	Log-likelihood
HO-KT	**87.13%**	97.76%	26.43%	**-5442.50**
KT per skill	85.92%	96.19%	27.28%	-5216.23
KT per student	85.15%	99.99%	0.92%	-5102.15
Majority class	85.10%	100.00%	0.00%	--

4. Discussion

HO-KT uses IRT to estimate students' initial knowledge of a skill based on its difficulty and discrimination and their overall proficiency, and KT to model learning over time. It outperforms per-student or per-skill KT by combining information about both. HO-KT estimates every probability *Knew*(student, skill) without requiring training data for every <student, skill> pair, because it can estimate student proficiency based on other skills, and skill difficulty and discrimination based on other students.

Future work should compare HO-KT to other methods and on data from other tutors. We should also test if *k*-dimensional student proficiency captures enough additional variance to justify fitting *k* times as many parameters. Finally, extending HO-KT to trace multiple subskills should use considerably fewer parameters than prior methods [11, 12], thanks to combining IRT and KT.

ACKNOWLEDGMENTS

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080157 and R305A080628 to Carnegie Mellon University, and by the National Science Foundation under Cyberlearning Grant IIS1124240. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute, the U.S. Department of Education, or the National Science Foundation. We thank Ken Koedinger for his algebra tutor data.

REFERENCES

[1] Corbett, A. and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 1995. *4*: p. 253-278.

[2] Pardos, Z. and N. Heffernan. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization*, 255-266. 2010. Big Island, Hawaii.

[3] Lee, J.I. and E. Brunskill. The Impact on Individualizing Student Models on Necessary Practice Opportunities *Proceeding of the 5th International Conference on Educational Data Mining (EDM)* 118-125. 2012. Chania, Greece.

[4] Hambleton, R.K., H. Swaminathan, and H.J. Rogers. *Fundamentals of Item Response Theory*. Measurement Methods for the Social Science. 1991, Newbury Park, CA: Sage Press.

[5] Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968: p. pp. 374 - 472.

[6] de la Torre, J. and J.A. Douglas. Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 2004. *69*(3): p. 333-353.

[7] Martin, A.D. and K.M. Quinn. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, 2002. *10*: p. 134-153.

[8] Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* 2009. *28*: p. 3049–306.

[9] Koedinger, K.R., R.S.J.d. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, et al., Editors, *Handbook of Educational Data Mining*, 43-55. CRC Press: Boca Raton, FL, 2010.

[10] Chang, K.-m., J.E. Beck, J. Mostow, and A. Corbett. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, K. Ashley and M. Ikeda, Editors. 2006: Jhongli, Taiwan, p. 104-113.

[11] Koedinger, K.R., P.I. Pavlik, J. Stamper, T. Nixon, and S. Ritter. Avoiding Problem Selection Thrashing with Conjunctive Knowledge Tracing. In *Proceedings of the 4th International Conference on Educational Data Mining*. 2011: Eindhoven, NL, p. 91-100.

[12] Xu, Y. and J. Mostow. Using Logistic Regression to Trace Multiple Subskills in a Dynamic Bayes Net. *Proceedings of the 4th International Conference on Educational Data Mining* 241-245. 2011. Eindhoven, Netherlands.