Investigate Performance of Expected Maximization on the Knowledge Tracing Model

Junjie Gu, Hang Cai, and Joseph E. Beck

Department of Computer Science, Worcester Polytechnic Institute
Worcester, MA, USA
{jgu2,hcai,josephbeck}@wpi.edu

Abstract. The Knowledge Tracing model is broadly used in various intelligent tutoring systems. As it estimates the knowledge of the student, it is important to get an accurate estimate. The most common approach for fitting the model is Expected Maximization (EM), which normally stops iterating when there is minimal model improvement as measured by log-likelihood. Even though the model's predictive accuracy has converged, EM may not have come up with the right parameters when it stops, because the convergence of the log-likelihood value does not necessarily mean the convergence of the parameters. In this work, we examine the model fitting process in more depth and answer the research question: when should EM stop, specifically for the Knowledge Tracing model. While typically EM runs for approximately 7 iterations, in this work we forced EM to run for 50 iterations for a simulated dataset and a real dataset. By recording the parameter values and convergence states at each iteration, we found that stopping EM earlier leads to problems, as the parameter estimates continue to noticeably change after the convergence of the log-likelihood scores.

Keywords: Knowledge Tracing, Bayesian Networks, Intelligent Tutoring Systems, Expected Maximization.

1 Introduction

The Knowledge Tracing (KT) model is widely used in various intelligent tutoring systems. KT is based on two knowledge parameters: learning rate and prior knowledge, and two performance parameters: guess rate and slip rate. Prior knowledge is the initial probability that the student knows a particular skill, guess is the probability of guessing correctly given the student does not know the skill, slip is the probability of making a slip given the student does know the skill, and learning is the probability of learning the skill given the student does not know the skill. The goal of KT is to infer the knowledge state of students from their observed performances.

The most common model fitting procedure for KT is Expected Maximization (EM). EM is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models [6]. This method is guaranteed to improve the likelihood function at each iteration. In [3], the authors also claimed

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 156–161, 2014.

that KT + EM results in more accurate models than KT + BF (Brute Force). To sum up, EM has the following distinct attributes:

1. Convergence of likelihood does not equal to convergence of parameters.
2. Initial values for the parameters are critical.
3. Parameter values sometimes exhibit extremely sharp changes after convergence of log likelihood.

As the parameters of KT represent the knowledge (the prior knowledge node), intelligence (the learning node) and attitude (the guess and slip rates) of a student, obviously, an incorrect estimate of the parameters may result in a wrong evaluation of a student, possibly causing the tutor or teachers to give additional assignments. Also, researchers interpreting the models to draw scientific conclusions will reach inaccurate conclusions if the parameters are incorrect. Thus, the acquisition of the right parameter is essential, as it will give the researchers the true knowledge of how students learn. Regarding that the values of the parameters may vary at different EM iteration, it is valuable to know when to stop running EM in order to get the right parameters.

2 Methodology

There were two components to our study. The first involved simulated data. For the simulation we used 5,000 students giving 10 responses to a skill, for 50,000 total sample data points. We set up the KT parameter values for the simulated data to: prior: 0.5, learn: 0.4, guess: 0.15, slip: 0.2, based on our knowledge of student learning. The real data we considered came from the 2009-2010 school year of ASSISTments. We select those student-skill sequences with less than or equal to 10 attempted opportunities. The final dataset contains 1,775 distinct students, 123 distinct skills and 695,732 data points. The BNT toolbox [4] is used to implement EM on the KT model, and EM stops when either of the two conditions is met:

1. The slope of the log-likelihood function falls below the threshold, which is set to 10^{-3} by default.
2. The number of iterations reaches the maximum number of iterations (max_iter), 100 by default.

The first condition indicates the process should stop when the model's accuracy ceases to noticeably improve. The threshold for improvement is normally set up to a default value 10^{-3}. The second condition, typically not encountered fitting KT models, represents a model that is not behaving well, and is possibly stuck in an infinite loop. Thus, EM typically stops when the slope of the log-likelihood score reaches the threshold, which we suspect is not equivalent to convergence of parameters. As models like the Student Skill model [5] have complicated Bayesian Network structures and massive number of parameters, it is important for researchers to decide when to stop running EM, more specifically, how to set the proper max_iter and threshold for EM to search for the right parameters.

In order to know the best time for EM to stop, we set max_iter to 50 and modified the code to stop iterating only when the current number of iteration reaches max_iter. Consequently, EM will always run 50 times before its termination. At each iteration, we also recorded the parameter values, the log-likelihood scores, and checked if the log-likelihood converged using EM's default threshold to see when EM would stop normally. Based on the fact that we know in advance the real parameters of the simulated data, it is more straightforward to observe how the parameters of the KT model change over time. Meanwhile, it is still worthwhile to see how EM performs on the real dataset, since the results may have some common features with those from the simulated data. For our experiments, we used the same set of initial parameters for both the simulated and the real data: prior: 0.3, learning: 0.5, guess: 0.15, slip: 0.05.

Our hypothesis is that the parameters may still change later on after the convergence of the log-likelihood. Whether this change represents change overfitting or a better estimate of the parameter is the question we will now explore.

3 Results

We first ran our experiment on the simulated dataset. Fig. 1 shows the values of the four KT parameters for each iteration. In order to observe how the parameters change over time and how close they are to the true parameters used to generated the data, we set the initial values as the starting points and the real values as the ending points in the graph. The vertical dashed line indicates when EM would have stopped using its default stopping criteria. As we can observe, all four parameters converged by the 35th iteration, and they converged at different points. The slip rate converged comparably quickly; on the contrary, the other three parameters converged slowly, but almost at the same time. Note that the parameters still changed considerably after the dashed line, meaning we would get an inaccurate estimate of the parameters using EM's default threshold. Therefore, we argue that it is necessary to wait for all the parameters to converge before stopping EM. Finally, the parameters at the 50th iteration are very close to the true parameters, which confirms the additional iterations of EM are not causing overfitting but are actually causing the parameters to become more accurate.

We also inspected the log-likelihood values at each iteration, which was in accordance with our hypothesis that the log-likelihood value converged quickly at early iterations and only changed slightly after that. We confirm that the convergence of log-likelihood indeed does not equal to the convergence of parameters, especially for the KT model. We believe this is also the reason why EM outperforms BF, considering BF searches for the best set of parameters based only on the predictive accuracies on the test data. And there exist multiple global and local maximums for the KT model [7], and EM always push the values of parameters closer to the real values at each iteration.

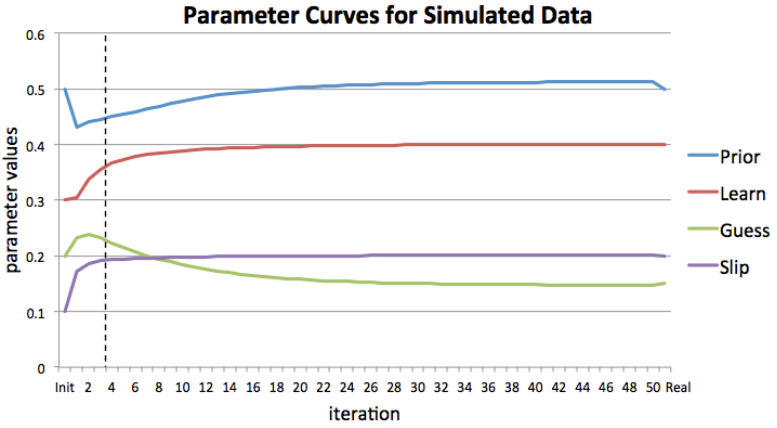

Fig. 1. Parameter curves for simulated data

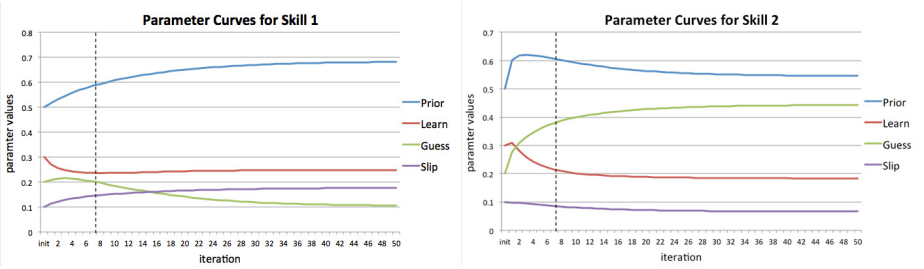

Fig. 2. Parameter curves for real data

For the simulated data, we then tested the predictive accuracies of the KT models on unseen test data. For AUC, EM's default setting gave an accuracy of 0.643, which was unchanged by running it for additional iterations. MAE slightly improved from 0.292 to 0.285, while RMSE was slightly worsened going from 0.459 for the default to 0.464 for running for 50 iterations. Based on our experience, the normal values for a KT model to predict a real dataset are around 0.35 for MAE, 0.46 for RMSE and 0.65 for AUC. Therefore, our simulated data behaves similarly to real data, and the model fit is similar to that on actual data. Although there is not much difference in predictive accuracy, Fig. 1 demonstrates that the parameter values are much closer to the real values when we let EM keep running.

We did a similar approach to examining model fit on real data. We performed a five-fold cross-validation, separating groups by skill. We compared EM under its default settings vs. the models we obtained after 50 iterations. Fig. 2 shows the parameter values at each iteration for two skills randomly selected. As these are real data, we cannot know their true parameter values. However, in agreement with the results from the simulated data (Fig. 1), the parameters continue to change after EM's default settings would cause it to stop. If the trend from Fig. 1 holds true, these

parameters are also more accurate. However, at a minimum we know there is no particular reason to believe the parameter estimates obtained after the default stopping criteria. Therefore, anyone using KT's parameter estimates for science, rather than for prediction, will encounter problems. A model stopping after 7 iterations and one stopping after 30 iterations have similar accuracy when making predictions, but in this case they make rather different claims about how quickly the skills are learned, and what students know when they begin working with the tutor. Besides the two skills showed here, we also inspected the graphs generated by all the other skills we tested, and found that generally parameter estimates were not stable at the point when EM in its default settings stopped its estimation process.

4 Conclusion and Future Work

In this work, we examined the popular model fitting process -- Expected Maximization for the Knowledge Tracing model in more depth and intended to answer the basic research question: when should EM stop. As the parameters represent the knowledge state of a student, it is crucial for the researchers to get an accurate estimate of the parameters. Although we cannot say when is the best time for EM to stop (which needs further exploration), we did find some valuable results, and most importantly, we found that stopping EM by its default threshold is definitely flawed, as the parameters still exhibit considerable changes after the convergence of the log-likelihood. From the predictive accuracy perspective, there is not much difference between the performances after 50 iterations and after default stopping, but simulation studies indicate that the parameter values are much closer to the real values when you let EM keep running. Although different datasets and parameters converge at different rates, our simulations indicate that 50 iterations are sufficient for parameter to converge. To sum up, we claim that EM definitely needs to run for more iteration to get the right values of the parameters. Overall, the results for all the datasets using both sets of initial parameters hold the following statements for the KT model:

1. Initial values (in large) do not affect the convergence of the parameters.
2. For different datasets, EM needs different number of iterations to make the parameters converging.
3. The parameters converge at different iteration and do not exhibit extremely sharp changes across one iteration after convergence of log likelihood.
4. Most importantly: *convergence in log likelihood space does not mean the convergence in the parameter space*.

The largest limitation of this work is that we only tested Expected Maximization on one particular Bayesian network model – Knowledge Tracing. However, the results may differ for other models. We intend to test EM on more invariants of the KT model like the Student Skill model, to check if the same results hold. For example, if the other models also don't show extremely sharp change after the convergence of the log-likelihood? Furthermore, although different initial values didn't make a difference in our experiments, they did affect the time for convergence. Thus we may integrate

the work with parameter plausibility such as Dirichlet priors in the future. We also wish to understand better rules for when to terminate search, and propose using the parameter curve graphs generated by the simulated data to assist searching for the parameters for the real dataset, because we believe, how the parameters change over time is also a factor in determining their true values.

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

References

1. Cen, H., Koedinger, K.R., Junker, B.: Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 164–175. Springer, Heidelberg (2006)
2. Corbett, A., Anderson, J.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction 4, 253–278 (1995)
3. Gong, Y., Beck, J.E., Heffernan, N.T.: Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 35–44. Springer, Heidelberg (2010)
4. Murphy, K.P.: The Bayes Net Toolbox for Matlab. Computing Science and Statistics (2007) DOI= http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html
5. Wang, Y., Heffernan, N.T.: The Student Skill Model. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 399–404. Springer, Heidelberg (2012)
6. Wikipedia, http://en.wikipedia.org/wiki/Expectation-maximization_algorithm
7. Beck, J.E., Chang, K.-m.: Identifiability: A Fundamental Problem of Student Modeling. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 137–146. Springer, Heidelberg (2007)
8. Cen, H., Koedinger, K.R., Junker, B.: Comparing two IRT models for conjunctive skills. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 796–798. Springer, Heidelberg (2008)