Less Is More: Improving the Speed and Prediction Power of Knowledge Tracing by Using Less Data

BAHADOR NOORAEI
ZACHARY A. PARDOS
NEIL T. HEFFERNAN
RYAN S.J.D. BAKER
Worcester Polytechnic Institute, USA

Knowledge Tracing is perhaps the most widely used student model in the field of educational data mining. In this paper we report on the effects of using only a subset of data in training the Bayesian Network that represents this student model. The standard practice is to use all of the students' data for a given skill to fit the model. We analyze two datasets; one from the Algebra Cognitive tutor and the other from the Genetics Cognitive tutor. We found that in both datasets, the difference in accuracy between using all the students' data versus only the most recent 15 data points of each student was not significantly different. Using only 15 responses however, resulted in an EM training time which was 15 times faster than using all data. This result suggests that the Knowledge Tracing model needs only a small range of data in order to learn reliable parameters. The implications of this result is a substantial savings in model training time that allows for more complex models to be fit or individualized models to be trained online.

Keywords and Phrases: Knowledge Tracing, Expectation Maximization, Prediction, Cognitive Tutor, Data filtering

1. INTRODUCTION

Knowledge Tracing (KT) [Corbett & Anderson 1995] is perhaps the most widely used student model in the field of educational data mining and has been used in many cognitive tutors [Koedinger, Anderson, Hadley & Mark 1997]. The standard practice is to use all of the students' data for a given skill to fit the model; and a model trained for each skill in the system.

In [Ritter, Harris, Nixon, Dickison, Murray & Towle 2009] it is discussed that reducing the parameter space of KT by means of clustering gives us models of student performance that are as good as the standard approach that gives us a different fit for each skill. So instead of 9600 parameters for the 2400 skills in the dataset, each fit differently, they settle on a set of 92 parameters, without changing the behavior of the system. In a similar vein, we aim to reduce the Knowledge Tracing training time by reducing the training data while retaining predictive performance.

In this paper we explore another question: how sensitive is the KT model to the amount of data used in its training. We train the model with different limits imposed on the maximum length of interaction instance sequences that is allowed for each student, and see their effect on prediction power of the system. To our knowledge this work is the first

Authors' addresses: Bahador Nooraei B., Zachary A. Pardos, Neil T. Heffernan. Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609 USA
Emails: bahador@wpi.edu, zpardos@wpi.edu, nth@wpi.edu

to explore using less data to do better when training a student model. As it is later shown, limiting the amount of data can reduce the training time of KT model using Expectation Maximization (EM) substantially. We analyze two datasets; one from the Algebra Cognitive tutor and the other from the Genetics Cognitive tutor.

1.1 KNOWLEDGE TRACING

Corbett & Anderson's Bayesian Knowledge Tracing model is one of the most popular methods for estimating students' knowledge. It underlies the Cognitive Mastery Learning algorithm used in Cognitive Tutors for Algebra, Geometry, Genetics, and other domains [Koedinger & Corbett 2006].

The canonical Bayesian Knowledge Tracing (BKT) model assumes a two-state learning model: for each skill/knowledge component the student is either in the learned state or the unlearned state. At each opportunity to apply that skill, regardless of their performance, the student may make the transition from the unlearned to the learned state with *learning* probability $P(T)$. The probability of a student going from the learned state to the unlearned state (i.e. forgetting a skill) is fixed at zero. A student who knows a skill can either give a correct performance, or *slip* and give an incorrect answer with probability $P(S)$. Similarly, a student who does not know the skill may *guess* the correct response with probability $P(G)$. The model has another parameter, $P(L_0)$, which is the probability of a student knowing the skill from the start. After each opportunity to apply the rule, the system updates its estimate of student's knowledge state, $P(L_n)$, using the evidence from the current action's correctness and the probability of learning:

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * \big(1 - P(S)\big) + \ (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + \ (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_n|Action_n) = \ P(L_{n-1}|Action_n) + \big((1 - P(L_{n-1}|Action_n)\big) * P(T))$$

The four parameters of BKT, $(P(L_0), P(T), P(S),$ and $P(G)$, are learned from existing data, historically using curve-fitting [7], but more recently using expectation maximization (*EM*). For *EM* the parameters were unbounded and initial parameters were set to a $P(G)$ of 0.14, $P(S)$of 0.09, $P(L_0)$of 0.50, and $P(T)$ of 0.14. These initial values were the average parameter values across all skills in prior modeling work conducted on a different algebra tutor [Pardos, Heffernan, Ruiz and Beck, 2008].

2. DATASETS

2.1 KDD DATASET (BRIDGE TO ALGEBRA)

This dataset comes from the Carnegie Learning Bridge to Algebra Tutor, which is an Intelligent Tutoring System (ITS) used by many students over the course of the 2007-2008 school year. This was the dataset that was one of the KDD 2010's "development" datasets [Pardos & Heffernan, In Press].

This dataset contains 1323 unit-skills (from now on, we call each unit-skill in this dataset simply a skill), and 1,817,476 data points (student actions). In order to demonstrate the effects of using less data, we limited our experiment only to those skills that have a

median of student response sequence of 40 or more. So we ended up with 33 skills with 663,491 data points (36% of all data points in the original dataset).
In this paper we refer to this dataset as KDD dataset.

2.2 GENETICS 2009 DATASET

This data was taken from a Cognitive Tutor for Genetics [Corbett, et al. 2010]. The dataset contains the results of in-tutor performance data of 76 students on 9 different skills, with data from a total of 11,581 student actions.
Six of the skills included in this dataset have average interaction sequence lengths of around 10. The 3 remaining skills have 20, 30, and 41 as average length of interaction sequence. But these 3 larger skills cover 60% of data points in the dataset.
For the students in this dataset, we also have the results of a problem solving post-test, covering some of the skills that were exercised in the tutor. Having this data, we could correlate student knowledge estimates of different configurations with the post-test data, as it is discussed in section 4.3.

3. METHODOLOGY

In our work, we group the dataset based on *skills* and fit a separate set of KT parameters for each skill. So an *instance* of interaction with the tutor in the dataset is a specific user's performance record when encountered with problems that exercise a specific skill. Here's where the idea using less data comes into the picture: should we use the full length of data in each instance, or should we limit the length of the data we feed into the EM? In order to explore the effects of imposing this limitation on the prediction-ability of KT model, we ran our experiments with different limitations on the number of data points used: from using all the data in each interaction sequence to using only the *most recent* 5 items.
For example, suppose we have two student with interaction instance sequences of $A = 0110011101111$ and $B = 1010110$ for a skill in the dataset (0 here denotes an incorrect response (wrong answer or request for help), and 1 denotes a correct response). Now if we limit the interaction sequence length to 5, the following two sequences are presented to the EM as interaction instance data related to the skill: $L_5(A) = 01111$ and $L_3(B) = 10110$. But if the limit is set at 10, we will have the following sequences: $L_{10}(A) = 0011101111$ and $L_{10}(B) = 1010110$. Notice that in the second case, whole sequence of B is used (length = 7).
For KDD dataset we tried fitting parameters with these interaction sequence length limits: 5, 10, 15, 20, 25, 35, 40, 75, 100, 150, 200, 400, and no-limit. The maximum interaction sequence in this dataset was 679, but as it is shown in the results section, not so many instances of interactions with that length are present in the dataset.
For Genetics dataset, the range of limits tried for this study is shorter because this dataset is considerably smaller and the lengthiest interaction sequence contains 88 data points. The limits we've tried for this dataset are: 5, 10, 15, 20, 25, 30, 40, 50, 60, and no-limit.

3.1 TRAINING AND TESTING OF THE KNOWLEDGE TRACING MODELS

For both datasets, we used trained KT models with different levels of student interaction data cut-off. Then all those models are used to predict student actions with cross-validation. For this prediction (or *trace*) phase, we used two different version of the

model: one that even limits the input to the *trained model* when it demands a prediction from it, and one that feeds the whole available history of student performance to the model, regardless of the limitation imposed at training stage.

Then we calculated the Root Mean Square Errors (RMSEs) at the student level and averaged them to get the number that is reported here. This way we can be sure that the reported accuracy is not biased towards students who have more those points for that skill. It also enables us to calculate a measure of statistical significance for the results.

We use Kevin Murphy's Bayes Net Toolbox for Matlab[1] for this experiment. 5-fold cross-validation was used for the Genetics dataset and a 2-fold for KDD to evaluate KT prediction performance. We used 2-fold cross-validation for KDD dataset solely to reduce the total amount of time needed to run all the experiments on that large dataset. The folds were created by randomly assigning students (and their associated responses) to folds. In each run of the cross-validation, one fold served as the test set and the other folds served as a training set.

4. RESULTS

4.1 KDD DATASET RESULTS

Root Mean Square Errors (RMSE) of cross-validated prediction of in-tutor performance of students are shown in figure 1, as the red trend line, and table 1. The x-axis in the graph (figure 1) shows the number of data points across all skills included in the EM training. As it is evident in the graph, increasing the amount of the data in training does not contribute to the model's accuracy, past a certain point (around a max of 75 responses per student). The prediction difference between five and 75 data points per student is not significant; however, the increase in runtime is substantial, shown by the blue line.

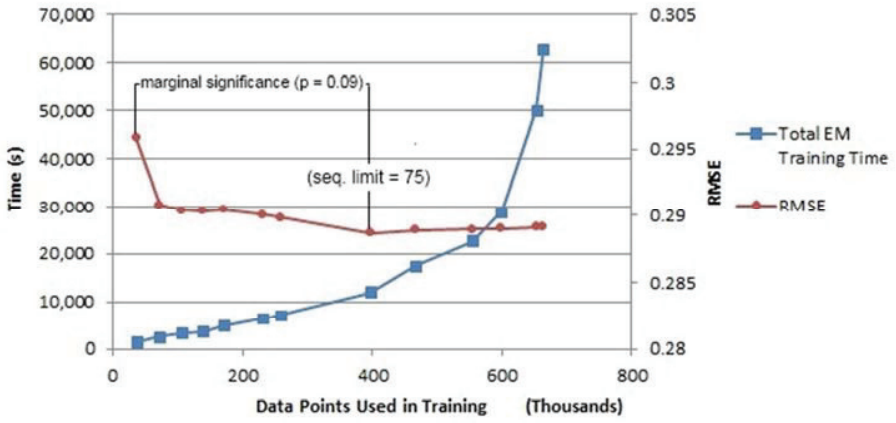

Figure 1. KDD In-Tutor Prediction Results

Figure 1 also shows the time it takes for EM to fit parameters for different amounts of data. It shows the potential exponential time complexity of the BN toolbox EM algorithm. We speculate that the change from linear to exponential time increase may have been attributed to the dataset exceeding the machine's 8 GIG memory capacity and disk swapping occurring.

[1] http://code.google.com/p/bnt/

Table I KDD In-Tutor Prediction Results (sorted by RMSE)

Sequence Limit	data points	RMSE
75	397,921	0.288648
100	467,471	0.288871
150	554,999	0.288948
200	598,774	0.288977
400	653,664	0.289098
No limit	663,491	0.289123
40	257,395	0.289811
35	230,149	0.290057
20	138,571	0.290328
15	105,442	0.290411
25	170,511	0.290428
10	71,276	0.290854
5	36,066	0.295847

When we put a maximum sequence limit of 10 on the training data, the trained model only became 0.6% less accurate than the fully trained one. The best accuracy was achieved by training with a limit of 75 on the each student's sequence length (represented 60% of all data points). Note that this was more accurate than using all the data yet it takes one-fifth of the time required to run the full training.

As mentioned in section 2, the KT model is a Bayesian Network with four parameters. So all differences in prediction ability of models, or lack thereof, is a consequence of the four parameters that is fit by EM. Figure 2 shows a graph of these four parameters as they are fit by using different amounts of data. The parameter values are an average of the parameter values in the 33 different KT skill models. The most dramatic change occurs in the prior parameter, which decreases monotonically. One explanation for the decrease in prior with longer sequences is that the longer the sequence, the more likely the data is produced by a student who is off task. Since students stop answering questions of a given skill when they master it, the students still answering questions after 75 opportunities are likely low achieving students with a low prior.

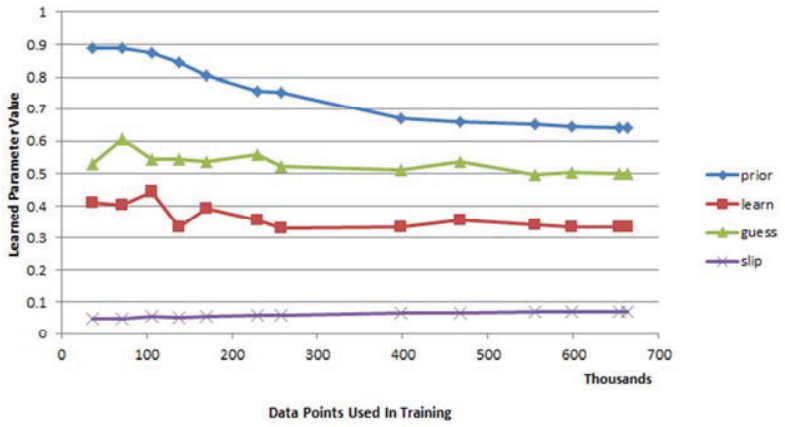

Figure 2 Avereage learned parameters with varrying amounts of data

The slip rate, however, was extremely stable; remaining around 0.08 for almost any amount of data. One interesting implication of this, at least when fitting models to Cognitive Tutor data, is that KT could be reduced to a three parameter model by learning the slip with very little data and then fixing that parameter to the value learned while the other three parameters are trained on more data.

4.2 Genetics Dataset Results

In the case of the genetics dataset, we are dealing with much less data than the KDD (Bridge to Algebra) dataset. Figure 3 shows that error does decrease steadily with more data, however, the decrease is very small and none of the errors are statistically significant. However, while the RMSE axis is zoomed to a scale that demonstrates the small change in error (the errors fall between 0.31 and 0.32 RMSE), the time axis (on the left) ranges between 10 minutes with 5 data point cut-off and 100 minutes with full data. This is a 10x training time increase to achieve no significant increase in prediction.

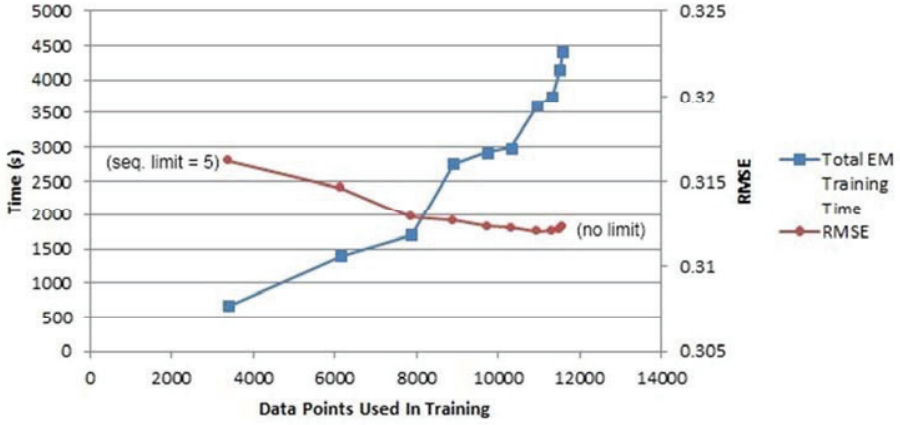

Figure 3 Genetics In-Tutor Prediction Results

In other words, limiting response sequence lengths to 5 (denoted by the first notch in the trend lines), which results in using only 29% of data points available, does not affect the prediction ability of the model at all. The best accuracy for in-tutor prediction is attained when using a sequence limit of 40, which includes 95% of data points; this is an increase in average RMSE of 0.00393 or 1% as shown in Table II.

Table II Genetics In-Tutor Prediction Errors (sorted by RMSE)

Sequence Length Limit	**Number of Data Points Included**	**RMSE**
40	10959	0.31198
50	11336	0.31203
60	11506	0.31206
30	10322	0.31223
No Limit	11581	0.31228
25	9734	0.31230
20	8898	0.31263
15	7875	0.31287
10	6156	0.31462
5	3386	0.31621

Figure 4 shows the average KT learned parameters for the Genetics dataset. A similar trend can be observed here as in the Cognitive Tutor dataset. The prior drops with more data and the slip remains nearly constant throughout. Unlike the Cognitive tutor, the guess rate decreases and the learn rate increases with more data. More investigation is necessary to explain these trends.

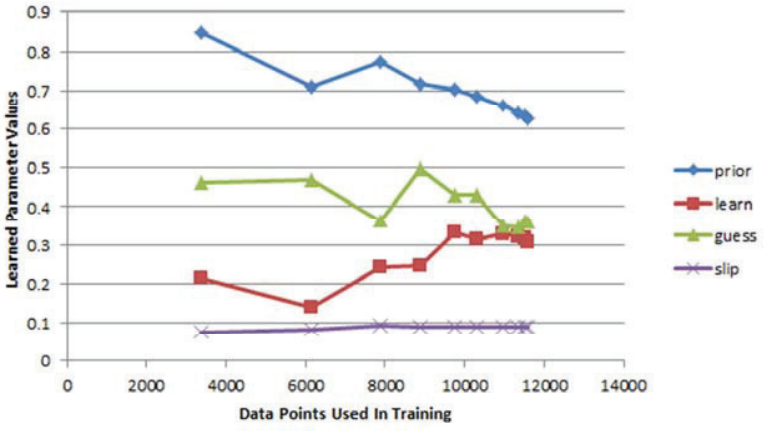

Figure 4 Average of Learned Parameters for Genetics Dataset

4.3 Predicting Post-Test Results for Genetics Dataset

In predicting the post-test, we account for the number of times each skill will be utilized on the test. Of the nine skills in the dataset, one is not exercised on the test, and is eliminated from the model predicting the post-test. Of the remaining seven skills, four are exercised once, two are exercised twice and one is exercised three times, in each of the two posttest problems. These first two skills are each counted twice and the latter skill three times in our attempts to predict the post-test. We use Pearson's correlation as the goodness metric since the model estimates and the post-test scores are both numerical. Correlation between each model and the post-test is given in table III.

The best correlation happens when we use a sequence limit of 20 in training (77% of data). The fact that using less data gives us better predictions for the Genetics Tutor students post-test was mentioned [2]in a recent work focusing on ensemble methods by the same authors [Baker, Pardos, Gowda, Nooraei, Heffernan In Press].

In all our experiments at predicting post-test we tried limiting the data in the tracing step as well: when using the trained model to predict student performance (we call this action *tracing)* we limited the amount past information we feed to the Bayesian Network. In other words, the same limit was imposed in tracing phase too. The results were no different from the normal *full* trace, so we eliminated any mention of them in this paper. But here, when predicting post-test results related to genetics dataset, we see an interesting phenomenon that a trace limited to only 5 most recent student data, yields a much better prediction of post-test results (table III and figure 5).

[2] This current article in submission to EDM2011 is cited in the prior Baker work.

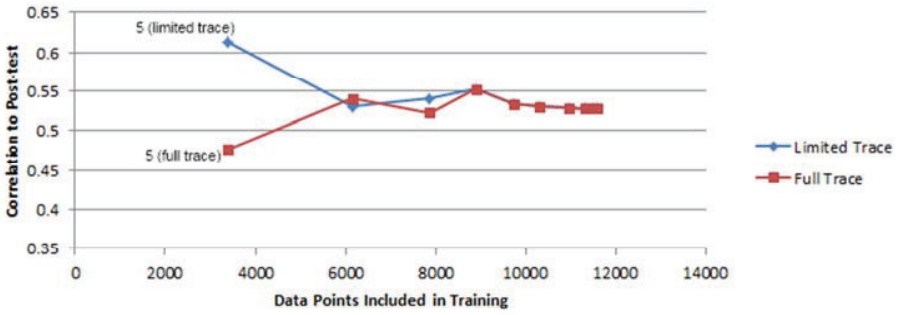

Figure 5 Genetics post-test correlation

Table III Genetics Post-Test Prediction Correlations

Sequence Length Limit	Data Points	Correlation with Post-test
5 (limited trace)	3386	0.61356
20	8898	0.55217
10	6156	0.54004
25	9734	0.53303
30	10322	0.52965
40	10959	0.52793
60	11506	0.52766
50	11336	0.52762
No limit	11581	0.52758
15	7875	0.52207
5 (normal trace)	3386	0.4761

5. CONCLUSION AND FUTURE WORKS

There are many practical reasons why one might be interested in decreasing the time it takes to fit/refit a model. In previous research [Pardos & Heffernan In Press], when we wanted to work with different variations of KT, long EM training runs were a huge impediment to rapid research cycles, so it motivated us to explore more in this area. In this paper we showed that fitting KT using EM with only a small subset of data gives us a model practically the same as a model fit with the whole available data. We also show that using only the most recent 5 data points to trace on provided the best correlation to post-test. This suggests that student's past history can be severely discounted when predicting their future performance. Tractability of individualized student models have been limited in part by the resources and time required to fit models. With our result that a good fit model can be achieved with very few data points, individualized models trained on the client can now be considered.

Our findings were largely unexpected; using 10% of the data in the case of the KDD dataset and 29% of the data in the case of the Genetics dataset lead to the same predictive power as using all the data. Given these results, ITS administrators can more wisely train their models, knowing the potential low benefit and high cost of using a student's entire response sequence to train their models. Researchers interested in predicting post-test measures from tutor data should also benefit from this finding that severely discounting the past can not only save model training time but also produce improve prediction results.

ACKNOWLEDGMENT

This research was supported by the National Science foundation via grant "Graduates in K-12 Education" (GK-12) Fellowship, award number DGE0742503 and the Department of Education IES Math center for Mathematics and Cognition grant. We would like to thank the Pittsburg Science of Learning Center for the Cognitive Tutor KDD dataset and Sujith Gowda for Genetics dataset preparation help. We would also like to thank Joseph Beck for his advice early on in the research.

We also acknowledge the many additional funders of ASSISTments Platform found here: http://www.webcitation.org/5ym157Yfr

REFERENCES

BAKER, R.S.J.D., PARDOS, Z.A., GOWDA, S.M., NOORAEI, B.B., HEFFERNAN, N.T. 2011 (in press.) Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems. To appear in *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization*.

CORBETT, A.T., ANDERSON, J.R. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

CORBETT, A., KAUFFMAN, L., MACLAREN, B., WAGNER, A., JONES, E. 2010. A Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. *Journal of Educational Computing Research*, 42, 219-239.

KOEDINGER, K. R., ANDERSON, J. R., HADLEY, W. H., & MARK, M. A. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30–43.

KOEDINGER, K. R., CORBETT, A. T. 2006. Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61-78). New York: Cambridge University Press.

PARDOS, Z.A., HEFFERNAN, N. T. 2001 (in press.) Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. To appear in the *Journal of Machine Learning Research* W & CP, In Press

PARDOS, Z. A., HEFFERNAN, N. T., RUIZ, C. & BECK, J.E. 2008. Effective Skill Assessment Using Expectation Maximization in a Multi Network Temporal Bayesian Network. The Young Researchers Track at the 20th International Conference on Intelligent Tutoring Systems. Montreal, Canada.

RITTER, S., HARRIS, T.K., NIXON, T., DICKISON, D., MURRAY, R.C., AND TOWLE, B. 2009. Reducing the Knowledge Tracing Space. In *Proceedings of EDM 2009*, 151-160.