# Towards Data Driven Model Improvement

**Yumeng Qiu, Zachary A. Pardos, Neil T. Heffernan**
Department of Computer Science
Worcester Polytechnic Institute
ymqiu@wpi.edu, zpardos@wpi.edu, nth@wpi.edu

**Abstract**

In the area of student knowledge assessment, knowledge tracing is a model that has been used for over a decade to predict student knowledge and performance. Many modifications to this model have been proposed and evaluated, however, the modifications are often based on a combination of intuition and experience in the domain. This method of model improvement can be difficult for researchers without high level of domain experience and furthermore, the best improvements to the model could be unintuitive ones. Therefore, we propose a completely data driven approach to model improvement. This alternative allows for researchers to evaluate which aspects of a model are most likely to result in model performance improvement. Our results suggest a variety of different improvements to knowledge tracing many of which have not been explored.

## Introduction

The Knowledge tracing model (KT) [1] has been use for over a decade to predict student knowledge and performance in the area of student knowledge assessment. As one of the most proven and accepted methods in the Intelligence Tutoring Systems field (ITS), KT uses a Dynamic Bayesian Network to track student knowledge. It has a set of four parameters, which are typically learned from data for each skill in the tutor. These parameters dictate the model's inferred probability that a student knows a skill given that student's chronological sequence of incorrect and correct responses to questions of that skill thus far. The two parameters that determine a student's performance on a question given their current inferred knowledge are the guess and slip parameters. KT provides both the ability to predict future student response values, as well as providing an addition parameter: the probability of student knowledge. For this reason, KT provides insight that makes it useful beyond the scope of simple response prediction. The standard Knowledge Tracing model is shown in Figure 1.

Numerous past researchers have shown that KT has its limitations. Many modifications to the KT model have been proposed and evaluated, however these modifications are often based on a combination of intuition and experience in the domain. This method of model improvement can be difficult for researchers without high-level of domain experience and the best improvements to the model could be unintuitive ones. Furthermore, KT can be computationally expensive [2][3]. Model fitting procedures, which are used to train KT, can take hours or days to run on large datasets. Therefore, we propose a completely data driven approach to model improvement. This alternative allows for researchers to evaluate which aspects of a model are most likely to result in model performance improvements based purely on the attributes of the dataset.

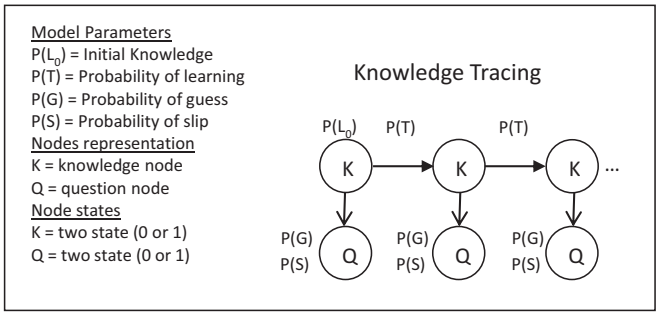


**Figure 1** The standard Knowledge Tracing model

## Dataset

We analyzed the KT model with a dataset from a real world tutor called the Cognitive Tutor. Our Cognitive Tutor dataset comes from the 2006-2007 "Bridge to Algebra" system. This data was provided as a development dataset in the 2010 KDD Cup competition [4].

In the Cognitive Tutor, students answer algebra problems from their math curriculum, which is split into sections. The problems consist of many steps (associate with skills) that students must answer to go to the next problem. The Cognitive Tutor uses the Knowledge Tracing model to determine when a student has mastered a skill. A problem in the tutor can also consist of questions of various skills. However, once a student has mastered a

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

skill, as determined by KT, the student no longer needs to answer questions of that skill within a problem. When a student mastered all the skills in their current section they are allowed to move on to the next. The time for students using this system is determined by their teachers.

### Selected Attributes

The Cognitive Tutor consists of many attributes such as student ID; step name, problem name; sub-skill name; step start time; hints and many more. Based on previous work, in this paper our primary goal is to discover how time information would impact model improvement.

To make the dataset more interpretable five attributes were computed from the original dataset that is most related to student performance time to test its individual impact on model improvement. The chosen attributes were listed as below:

- Time interval between responses
- Count of the number of days spent trying to master a skill
- Opportunity count (number of steps answered of a skill)
- Percent correct of a student
- Percent correct of a skill

The attributes of percent correctness of student and skill were calculated base on the number of correct responses for one student and for that skill, it is a continues number in the range of 0 ~ 1. The time interval between responses was separated into four bins. 1 represents the response that were answer in one day, 2 represent the time interval between the consecutive responses is one day, 3 represents the time interval is within a week and 4 represents the time interval between consecutive responses is more than a week. The third attribute was calculated based on the number of days the student work per skill. As for Opportunity count, it represents how many responses a student made per skill.

The original dataset was divided by sub-skills. Each sub-skill such as “identify number as common multiple”, ”list consecutive multiple of a number” and “calculate the product of two numbers” were all counted as skills in this analysis. Each skill individually is counted as a dataset. Here, eleven skills were randomly chosen from the pool of math skills that the original dataset provided for analysis, which exclude the action steps such as “press enter” that do not represents math skills. The skills had an average of 900 student responses per skill.

## Methodology

A two-fold cross-validation was done in order to acquire the KT model prediction on the datasets. The two-fold cross-validation involved randomly splitting each dataset into two bins, one for training and one for testing. A KT model was trained for each skill. The training phase involved learning the parameters of each model from the training set data. The parameter learning was accomplished by using the Expectation Maximization (EM) algorithm [5]. EM attempts to find the maximum log likelihood fit to the data and stops its search when either the max number of iterations specified has been reached or the log likelihood improvement is smaller than a specified threshold.

Since we wanted to learn more about exactly how KT was performing we combined all the prediction results together in order to track residuals on a per opportunity basis. Figure 2 show the graph for the first 10 student responses. It should be noted that the majority of our student response sequences are about 10 responses long. The behavior of the graphs from 11-15 is based on fewer data points than the rest of the graph. The residual graph showed that KT is under-predicting early in the response sequence. In Wang et al. [6], their intuition for this phenomenon is that KT takes too long to assess that a student knows a skill and once it believes a student knows a skill, KT over predicts correctness late into a student’s response sequence. Essentially the authors point out that KT has systemic patterns of errors. We believe these errors can be corrected for by looking to unutilized features of the data.

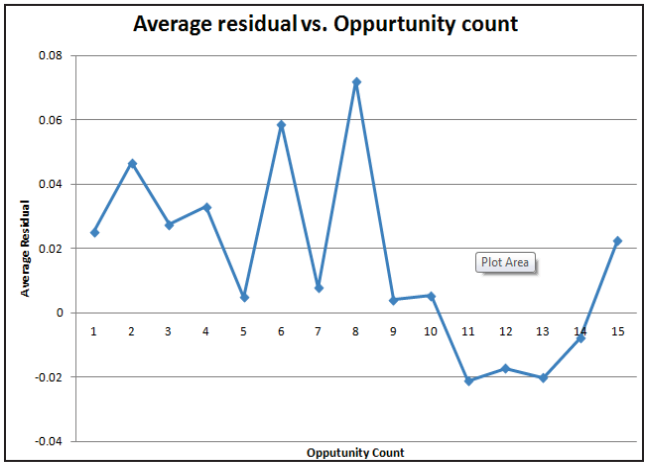


**Figure 2** The residual graph of the KT model

With this graph we were able to convince ourselves that some simple correction could exist that could smoothen the residual curve in order to improve the model. In this paper we conducted three experiments to evaluate the selected 5 attribute of the dataset.

### Experiment 1

In the first experiment, the selected 11 skills were completely combined together to make a one large dataset, each row represents a student response to a given problem

and are all treated equally disregard which skill it was from. A five-fold cross-validation was used to make predictions on the dataset, which means randomly splitting the dataset into five bins at the response level.

In order to know which attributes of the dataset could lead to a better improvement of the KT model, a regression analysis was conducted for each attribute. Regression analysis is used to understand which among the independent variables are related to the dependent variable. The regression function is shown below. In this function the unknown variable is denoted as $\boldsymbol{\beta}$, the dependent variable is denoted as Y and the independent variable is denoted as X.

$$\boldsymbol{Y \approx f(X, \beta)} \quad (1)$$

All analysis takes the residual result of the KT model as the dependent variable and the selected attributes of the dataset are treated as the single independent variable of each analysis. After a regression function was trained for each attribute, the estimated value of the unknown variable was treated as a correction to the prediction of the KT model. Therefore we gain the corrected prediction of that attribute. By doing this we are able to predict patterns in KT's error (residual) based on various dataset features (independent variable). If the error can be predicted with high accuracy then this tells us that the KT model can benefit from inclusion of that variable information.

## Experiment 2

The second experiment was done at the skill level. Similar to experiment 1 a five-fold cross-validation was also used to make prediction on the dataset. There will be five rounds of training and testing where at each round a different bin served as the test set, and the data from the remaining four bins served as the training set. To note that the skills in the training set will not appear in the testing set in order to avoid over fitting. The cross-validation approach has more reliable statistical properties than simply separating the data in to a single training and testing set and should provide added confidence in the results.

The regression analysis was also conducted similar to experiment 1 and the new model prediction was corrected based on the given attributes. Because this correction is done at the skill level the analysis for the attributes "% correct by skill" was omitted here.

## Experiment 3

The third experiment was done at the student level. Similar to experiment 1 and 2 a five-fold cross-validation was also used to make prediction on the dataset. There will still be five rounds of training and testing where at each round a different bin served as the test set, and the data from the remaining four bins served as the training set. To note that the student in the training set will not appear in the testing set in order to avoid over fitting.

The regression analysis was also conducted similar to experiment 1 and 2. The new model prediction was corrected based on the given attributes, also here in experiment 3 the attribute "% correct by student" was also omitted because the correction is done at the student level.

## Results

Predictions made by each model were tabulated and the accuracy was evaluated in terms of root-mean-square error (RMSE). RMSE is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated. Here we use the Knowledge Tracing model prediction residual as the observed value. Therefore the correction is apply to the residual it would minimize it's distance to the ground truth.

The cross-validated model prediction results for experiment 1 are shown in Table 1; the cross-validated model prediction results for experiment 2 are shown in Table 2 and the cross-validated model prediction results for experiment 1 are shown in Table 3. The p values of paired t-test comparing the correction models and the standard KT model are included in addition to the RMSE for each model in each table.

**Table 1.** RMSE results of KT vs. Correction models at opportunity level

| Attributes | RMSE | T-test |
|---|---|---|
| **KT** | 0.3934 | |
| **Time interval** | 0.3891 | **<< 0.01** |
| **Day count** | 0.3912 | **<< 0.01** |
| **% correct by student** | 0.4050 | > 0.05 |
| **% correct by skill** | 0.3931 | **0.0128** |
| **Opportunity count** | 0.3930 | **0.0347** |

**Table 2.** RMSE results of KT vs. Correction models at skill level

| Attributes | RMSE | T-test |
|---|---|---|
| **KT** | 0.3934 | |
| **Time interval** | 0.3898 | **<< 0.01** |
| **Day count** | 0.3928 | 0.2639 |
| **% correct by student** | 0.4047 | > 0.05 |
| **Opportunity count** | 0.3937 | 0.0686 |

**Table 3.** RMSE results of KT vs. Correction models at student level

| Attributes | RMSE | T-test |
|---|---|---|
| **KT** | 0.3934 | |

| | | |
|---|---|---|
| **Time interval** | 0.3892 | **<< 0.05** |
| **Day count** | 0.3913 | **<< 0.05** |
| **% correct by skill** | 0.3929 | **<< 0.05** |
| **Opportunity count** | 0.3932 | 0.2451 |

The results from evaluating the models with the cognitive tutor datasets are strongly in favor of the time interval correction model in all three experiments. With the time interval correction model beating KT in RMSE The average RMSE for KT was 0.3934 while the average RMSE for the time interval correction model was 0.3892, 0.3898 and 0.3892. These differences were all statistically significantly reliable p =1.92E-10, p= 1.27E-08 and p=2.61E-10 using a two tailed paired t-test.

As for the other correction models the three experiments seem all agree that the "% correct by student" attributes is not useful in improving the KT model and opportunity count is also not a very good correction model, we can assume that it is not very likely to see a very large improvement if this attribute is considered as the medication to the KT model. According to Table 1 and Table 3, the day count correction model's average RMSE were 0.3912 and 0.3892 which are both better than the KT RMSE 0.3934 and the difference are all statistically significantly reliable p =1.05E-06 and p=4.74E-06. Yet the evaluation at the skill level seems not to agree to the other evaluations, even though the error is still smaller but it is not significantly reliable.

## Discussion and Future Work

From the experiments above, we assume that taken "time interval" attribute into account as a modification will lead to a significant improvement to the standard Knowledge Tracing model. The model proposed in the Qui et al. [7] is a very nice proof of the feasibility of this method. Currently with the result from this paper we can eliminate the work of trying to improve student assessment with this dataset by using the attribute of "% correct by students" but the other two attributes "% correct by skill" and "day count" still need further evaluation. Especially the "day count" attribute seem to suggest a great possibility to impact the Knowledge Tracing model.

Many more experiments like the ones in this paper could also be done for other attributes and generated features of datasets. There are several interesting inferences that could be made about the impact or lack of impact of the various features on the knowledge tracing model's predictions.

## Contribution

We have described a methodology for identifying areas within a model that can be improved upon. The residual corrections of our different features gave a strong indication that time between responses would be of significant benefit to the knowledge tracing model. The general student feature of % correct across the system was not beneficial to model prediction, indicating that it may not be worth the effort to implement individualized student priors into the knowledge tracing model with this dataset due to the high variability in performance across skills.

The idea of data driven user modeling is a powerful one. While domain expert derived user models are valuable, they are also prone to expert blind spots. We believe that educational researchers and researchers outside this field can benefit substantially from employing data driven techniques to help build accurate and generalizable user models.

## Acknowledgments

This research was supported by the National Science foundation via grant "Graduates in K-12 Education" (GK-12) Fellowship, award number DGE0742503, and Neil Heffernan's NSF CAREER grant, award number REC0448319. We also acknowledge the many additional funders of the ASSISTments Platform found here: http://www.webcitation.org/5ym157Yfr

## References

[1] Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User Adapted Interaction, 4, 253 278. (1995)

[2] Bahador, N., Pardos, Z., Heffernan, N. T. AND Baker, R. 2011. Less is More: Improving the Speed and Prediction Power of Knowledge Tracing by Using Less Data. Educational Data Mining 2011 Conference

[3] Ritter, S., Harris, T., Nixon, T., Dickison, D., Mirray, C., Towel, B. 2009. Reducing the knowledge tracing space. In Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain, 151 160

[4] Pardos, Z.A., Heffernan, N. T.: Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. To appear in Journal Of Machine Learning Research W & CP (In Press)

[5] Chang, K., Beck, J., Mostow, J., and Corbett, A. T.: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. Intelligent Tutoring Systems, 8th International Conference, ITS 2006, 104 113

[6] Wang, Q., Pardos, Z.A, Herffernan, N. T.: Response Tabling A simple and practical complement to Knowledge Tracing. KDD workshop (2011)

[7] Qiu, Y., Qi, Y., Lv, H., Pardos, Z.A, Herffernan, N. T.: Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing, Educational Data Mining 2011 Conference