Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting

Yue Gong , Joseph E. Beck, Neil T. Heffernan

Computer Science Department, Worcester Polytechnic Institute
100 Institute Road, Worcester, MA, 01609, USA
{ygong, josephbeck, nth}@wpi.edu

Abstract. Student modeling is very important for ITS due to its ability to make inferences about latent student attributes. Although knowledge tracing (KT) is a well-established technique, the approach used to fit the model is still a major issue as different model-fitting approaches lead to different parameter estimates. Performance Factor Analysis, a competing approach, predicts student performance based on the item difficulty and student historical performances. In this study, we compared these two models in terms of their predictive accuracy and parameter plausibility. For the knowledge tracing model, we also examined different model fitting algorithms: Expectation Maximization (EM) and Brute Force (BF). Our results showed KT+EM is better than KT+BF and comparable with PFA in predictive accuracy. We also examined whether the models' estimated parameter values were plausible. We found that by tweaking PFA, we were able to obtain more plausible parameters than with KT.

Keywords: Student modeling, Knowledge tracing, Performance Factors Analysis, Expectation Maximization, Machine learning, Model fitting approaches

1 Introduction

Student modeling is one of the major issues for Intelligent Tutoring System as it has been widely used for making inferences about the student's latent attributes. Its working mechanism is to take observations of a student's performance (e.g. the correctness of the student response in a practice opportunity) or a student's actions (e.g. the time he stayed for a question), and then use those to estimate the student's underlying hidden attributes, such as knowledge, goals, preferences, and motivational state, etc. Those attributes are unable to be determined directly, thus student modeling techniques have always attracted a great deal of attention.

In ITS, student modeling has two common usages. The first, and most frequently used one, is to predict student behaviors, such as student performance in the next practice opportunity. The second one is to obtain plausible and explainable parameter estimates, where plausibility concerns how believable the parameters are, often tested by comparing them to some external gold standards. Being explainable indicates the parameter estimates produced by the model have practical meanings, which can help researchers know more about learning. Consequently, student models are evaluated by how well they predict student's behaviors, as well as by parameter plausibility [1].

1.1 Knowledge tracing model

There are a variety of student models. The knowledge tracing model [2] shown in Fig. **1**, has been broadly used. It is based on a 2-state dynamic Bayesian network where student performance is the observed variable and student knowledge is the latent. The model takes student performances and uses them to estimate the student's level of knowledge. There are two performance parameters: slip and guess, which mediate student knowledge and student performance. The guess parameter represents the fact that the student may sometimes generate a correct response in spite of not knowing the correct skill. The slip parameter acknowledges that even students who understand a skill can make an occasional careless mistake. There are also two learning parameters. The first is initial knowledge (K0), the likelihood the student knows the skill when he first uses the tutor. The second is the learning rate, the probability a student will acquire a skill as a result of an opportunity to practice it.

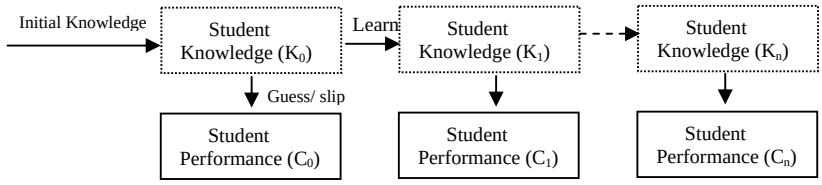

Fig. 1. Knowledge tracing model

As pointed out in [3, 4], KT suffers two major problems with trying to estimate parameters: local maxima and multiple global maxima. The first one is common to many error surfaces and has known solutions such as multiple restart. The second difficulty is known as identifiability and means that for the same model structure, given the same data, there are multiple (differing) sets of parameter values that fit the data equally well. Based on statistical methods, there is no way to differentiate which set of parameters is preferable to the others. Consequently, for the KT model, different model fitting approaches lead to different parameter estimation outcomes.

1.2 Performance Factor Analysis

Recently, a new alternative student modeling approach was presented by Pavlik et al. [5], Performance Factor Analysis (PFA). PFA is a variant of learning decomposition [6], and is based on reconfiguring Learning Factor Analysis (LFA) [7]. Briefly speaking, it takes the form of standard logistic regression model with the student performance as dependent variable. It reconfigures LFA on its independent variables, by dropping the student variable and replacing the skill variable with the question identity (i.e. one parameter per question). The model estimates a parameter for each item representing the item's difficulty, and also two parameters for each skill reflecting the effects of the prior successes and prior failures achieved for that skill.

Previous work compared KT and PFA models, and found PFA to be superior. In this study, we ran replication study, but also focusing on the impacts when using different model fitting approaches for KT. In addition, we attempted and tested

different methods of handling multiple-skill problems. We present our comparison results based on both predictive accuracy and parameter plausibility.

2 Methodology

2.1 Model fitting approaches for Knowledge tracing model

Aside from student models, there are also a variety of model fitting approaches. Different approaches have various criteria to fit the data, and thus produce different parameter estimates and further lead to different predictive abilities. Therefore, we explored the impact of modeling fitting approach on model accuracy and plausibility.

The Expectation Maximization (EM) algorithm is a model fitting approach for KT. It finds a set of parameters that maximize the data likelihood (i.e. the probability of observing the student performance data). EM processes student performance as a piece of evidence with time order, and uses this evidence for the expectation step where the expected likelihood is calculated. The model then computes the parameters which maximize that expected likelihood. The algorithm, by accessing more evidence, iteratively runs these two steps until it finds the final best fitting parameters. There is no guarantee of finding a global, rather than a local, maxima.

Recently, the brute force approach has been proposed for estimating parameters for KT. The algorithm uses exhaustive search for finding the best set of parameters over a reasonable sampling of the entire parameter space. Contrary to EM that maximizes the data likelihood, it attempts to minimize the sum of squared error (SSE). Originally, KT's parameters are continuous, so that there is no way to compose a finite search space, which, however, is a must for an exhaustive search. We used source code provided by Ryan Baker, which resolves the issue by only considering two decimal places of precision. In this way, the parameter space is reduced from infinity to 99^4 possible parameter sets for each skill (i.e. there are four parameters for each skill and each of them has 99 possible values ranging from 0.01 to 0.99). Initially, every parameter starts from the value of 0.01 and is incremented by 0.01 on every iteration. Ultimately, for each skill, it finds the lowest SSE, and the corresponding set of parameters for that skill. The major drawback is that the method suffers from a severe computational cost problem due to the large search space, so most of the time the search space is cut down even smaller by setting searching boundaries. In order to make a careful comparison, we followed the same protocol as Pavlik et al. followed [5]. Specifically, we used the same set of bounded ceiling values for the four parameters, so that the maximum probabilities of initial knowledge, guess, slip and learning are 0.85, 0.3, 0.1 and 0.3, respectively.

Conjugate Gradient Descent, an optimization method used to solve systems of equations, is used to estimate parameters in the CMU cognitive tutors. Chang et al. [8] found that EM produces models with higher predictive accuracy than Conjugate Gradient Descent.

Unlike the KT model, the family of learning decomposition models is based on the form of standard logistic regression, so that the model fitting procedure is ensured to

reach global maxima; thus resulting in unique best fitting parameters. Consequently, for PFA, the model fitting approach is not an issue.

2.2 Problem with multiple skill questions in KT model

The KT framework has a major problem: when there is more than one skill involved in a question, the model lacks the ability to handle all the skills simultaneously, because KT works by looking at historical observations on *a* skill. However, in some tutors, a question is usually designed to require multiple skills to achieve a correct answer. If a student model cannot take this common phenomenon well, its ability of making plausible parameter estimates and accurate prediction is likely to be weakened.

A common solution is to attach performance on the problem with all skills needed to solve it, by listing the performance in all of those skills' historical observations [e.g., 10]. Thus, a multiple skill question is split into multiple single skill questions. This strategy enables parameter estimation to proceed, but increases the probability of over fitting and also results in an accompanying problem: multiple predicted performances. Each split performance is associated with a particular skill that has its own set of parameters. We then are able to use those parameters to calculate the predicted performance. However, those calculated values are probably not equivalent across all of the skills, which means for the same student, on the same practice opportunity, our models make different claims about how likely he is to produce a correct response. Given the conflicting predictions, some means of making a prediction is needed.

In this study, we attempted two approaches to address the problem. The first is similar to [9] and inspired by the joint probability in Probability Theory. The probability a student generates a correct answer in a multi-skill question is dependent on his ability to achieve correctness in all required skills. Therefore, we multiplied each skill's predicted performance together and assign the product as the new predicted performance for all corresponding observations. Yet, the reasonableness of this method relies on a strong assumption, which is how likely a student can answer correctly for one skill must be independent of the probability that he responds correctly in another skill.

The second approach, in the other hand, takes the minimum probability of the predicted performances as the final value. The intuition behind this strategy is the likelihood a student gets a correct answer is dominated by his knowledge of his weakest skill.

The above strategies are necessary options for KT due to its lack of ability to handle multi-skill performances. However, it is not the case for PFA, which does have the ability to handle with multi-skill performances. Therefore in this study, in addition to repeating multiple skill questions like we do for KT, we also examined PFA using the data that still keeps the original multi-skill performances.

2.3 Data set

For this study, we used data from ASSISTment, a web-based math tutoring system. The data are from 343 twelve- through fourteen- year old 8th grade students in urban school districts of the Northeast United States. They were from four classes. These data consisted of 193,259 problems completes in ASSISTment during Nov. 2008 to Feb. 2009. Performance records of each student were logged across time slices for 104 skills (e.g. area of polygons, Venn diagram, division, etc).

3 Results

We used BNT-SM [8] to perform the EM algorithm for the knowledge tracing model to estimate the model parameters. We used Ryan Baker's unpublished java code to accomplish the brute force model fitting method. We also replicated the PFA model using the same model settings as in [5], except where noted below. We fit the three models with the data that contains split performances (i.e. problems that require multiple skills). For PFA, we also examined it by fitting the original data which keeps multi-skill questions as a single unit. We referred the PFA model handling multi-skill performances as PFA_M and the other as PFA_S. We did 4-fold crossvalidation and tested our models on the unseen students, which is different from what Pavlik, et al. did. They conducted 7-fold crossvalidation and tested their models on seen students' unseen performances. We prefer to hold out at the student level since that results in a more independent test set. Another aspect in which our approach differed from Pavlik's is that we did not restrict the impact of a correct response to be non-negative, that is, skill could have negative "learning rates."

3.1 Main model comparisons: Predictive Accuracy

Predictive accuracy is the measure of how well the instantiated model fits the data. We used three metrics to examine the model predictive performance on the unseen test set: Mean Squared Error (MSE), R^2 and AUC (Area Under Curve) of ROC curve. We also reported the number of parameters produced by each model.

Table 1. Crossvalidated predictive accuracy comparison among three main models

	MSE	R2	AUC	# of parameters
KT + EM	0.215	0.072	0.661	416
KT + BF	0.223	0.036	0.656	416
PFA_S	0.220	0.048	0.673	1013

Table 1 shows the results of the comparison for the three metrics. The values are calculated by averaging corresponding numbers obtained in the 4-fold crossvalidation. Although most numbers seem very close, KT+EM outperforms KT+BF in all three metrics, and PFA_S seems to beat KT+BF as well. To examine whether the

differences are statistically reliable, for every two models, we did a 2 tailed paired t-test based on the results from the crossvalidation. Only between KT+BF and KT+EM, we found the differences are significant in all three metrics ($p<0.01$ in MSE and R2; $p=0.02$ in AUC). We failed to find any reliable differences between PFA_S and KT+BF, even though the mean values appear a trend suggesting PFA_S is probably better than KT+BF. Compared to PFA_S, KT+EM wins in the first two metrics, but does worse on the third one. Besides, none of the statistical test results suggests there are any significant differences between these two models. We noticed that PFA produced more parameters than KT, which seems inconsistent with the results reported in [5].But, given that the number of parameters in PFA = 2*# of skills + # of items, while the number of parameters in KT is 4*# of skills. Consequently having fewer items favors PFA, while fewer skills benefits KT.

One reason for the lack of reliable difference could be the relatively low statistical power of the t-tests. Four independent observations (one for each fold of the cross validation) provide low power to differentiate samples. Although further research needs to perform the comparisons based on larger samples, for now we maintain a conservative view of KT+EM is comparable with PFA_S.

KT+EM provides more accurate predictions than KT+BF. We think this consequence is caused by the range restrictions used on the search space. In contrast, EM derives its estimations without such range restrictions, so it is more likely for EM to produce more plausible parameter estimates which further are used to yield more accurate prediction.

One aspect hindering the performance of PFA_S is that the PFA model is designed to handle problems that require multiple skills. Hence, it is more reasonable to inspect PFA's performance when it works in its natural way. To make a fair comparison, we trained and tested PFA using the same data sets as we used for the other models, but without splitting multi-skill performances.

Table 2. Crossvalidated comparisons among PFA with different model settings

	R2	AUC
PFA_S (base line)	0.048	0.673
PFA_M	0.047	0.680
PFA_S_bounded	0.066	0.681
PFA_M_bounded	0.074	0.690

As seen in the first two rows of Table 2, generally PFA_M results in comparable performances, although the difference in AUC is statistically reliable at $p<0.01$. This implies that PFA works better when used in its original spirit for handling multiple skill questions. However, compared to KT+EM, shown in Table 1, we only found little evidence to support it is good at multi-skill questions (its AUC value is higher, but not reliably so). Therefore, again we failed to be able to show that the true PFA can reliably outperform this version of KT+EM that also attempts to deal with multi-skill questions. We do not present MSE in Table 2 since PFA_S and PFA_M use slightly different datasets, it is not appropriate to compare MSE.

One drawback of PFA models is they could produce negative learning rates due to over fitting, so in the original work [3], the researchers set 0 as the lower bound. We

were unable to get our model fitting software (SPSS 17.0) to replicate this procedure[1]. Since PFA's lack of better predictive performance could result from loosing this constraint, we next manually checked which skills had negative learning rates, substituted 0 in, and then re-ran the model predicting procedure on the test data using the altered parameters. We found almost half of the skills had negative learning rates. The results shown in the last two rows of Table 2 indicate both two bounded models indeed achieved higher predictive accuracy ($p<0.01$ in all two metrics, compared to PFA_S), suggesting that the negative learning rates were not accurate and the result of overfitting. Considering the number of negative learning rates produced at first, it seems that setting bounded value is necessary for PFA.

3.2 Comparing approaches to the problem of multi-skill questions

Given the problem of multi-skill questions in KT, we compared the two proposed approaches for predicting performance (multiplication and min()) with the default model (making multiple, different predictions, on student performance on a problem, one for each skill).

Table 3. Crossvalidated **c**omparisons of the min models and the default models

	MSE	R2	AUC
KT_BF	0.223	0.036	0.656
KT_BF_min	0.220	0.046	0.670
KT_EM	0.215	0.072	0.661
KT_EM_min	0.214	0.073	0.676

We found the approach of calculating the product results in worse predictive accuracy in all attempted models, and do not report it here. However, taking the minimum value of the predicted performances provides more accurate models. As shown in Table 3, the min models are generally better than the default models, with the AUC values are reliably different in every pair of the comparisons.

The problem of multi-predicted performances also influences PFA, when it takes the data with manually split performances. Therefore, we applied the two approaches on PFA_S and PFA_S_bounded as well. We found that min model didn't consistently work well for PFA.

3.3 Parameter Plausibility

Predictive accuracy is a desired property, but ITS researchers are also interested in interpreting models to make scientific claims. Therefore, we prefer models with more plausible parameters when we want to use those for scientific study. We followed the technique in [4]: using external measurement to evaluate parameter plausibility.

[1] If any readers know how to coerce SPSS's logistic regression function to do so, they are invited to contact us.

The students in our study had taken a 33-item algebra pre-test before using ASSISTment. Taking the pre-test as external measure of incoming knowledge, we calculated the correlation between the students' initial knowledge estimated by the models and their pretest scores.

In other to acquire student's initial knowledge parameter, we used KT to model the students instead of skills (see [4] for details). Since PFA has no student parameter by default, we tweaked it to include student as an additional independent variable. In Table 4, we see that the PFA model that fit by the data keeping multi-skill performances produces the strongest correlation. Even, PFA_S, modified to behave like KT with respect to multiple skill questions, the number still remains the largest (0.886) compared to the rest. KT+BF surprisingly shows a higher ability to estimate plausible parameters than KT+EM. One thing to notice is this correlation is produced by KT+BF with bounded parameter values, thus if the search space is enlarged, it might be able to derive potentially better parameter estimates. The KT+EM is reliably different ($P<0.05$) from PFA_S and PFA_M; none of the other differences is reliable.

Table 4. Comparison of parameter plausibility

	KT+BF	KT+EM	PFA_S	PFA_M
Correlation	0.865	0.827	0. 886	0.906

4 Contributions

This paper examines and compares the different model fitting approaches of estimating parameters for the knowledge tracing model. We are able to extend the result that EM produces more predictive models than Conjugate Gradient Descent. We are now able to say, for our dataset, that EM also outperformed the brute force algorithm in predictive accuracy. Others [11] have found brute force outperforms EM, so more work is needed here. Furthermore, we inspect the parameter plausibility produced by the models with these two fitting methods and show brute force is good at estimating more plausible parameters, and has the potential to perform even better.

This work also replicates the testing between Performance Factor Analysis model and the knowledge tracing model. This replication is non-trivial, as there is increasing concern that a research finding is less likely to be true when the studies conducted in a field are small [10], and independent replication should be given great importance as it extends the original work upon different environments and subjects, outside the original research teams, based on different measurement metrics. We conclude that PFA is a comparable model with KT. We also examine the PFA's predictive performances given different model settings. PFA with bounded learning rates that directly handles multi-skill questions outperforms the models with the other settings. In addition, by tweaking PFA to endow it a powerful ability to capture individual differences, the model produces highly plausible parameters.

We also attempt two methods to solve the problem of multi-skill questions. Since the regular KT has no ability to deal with such questions, we assigned the minimum value and the product value as the predicted performance for all relevant split

performances. We found some evidence to support that min models are somewhat better than the default models in both KT+EM and KT+BF.

5 Future work

There are several interesting open questions. First, brute force is a new proposed model fitting approach for KT, so there are many research questions worth exploring. Setting bounded values for parameter estimation is important and somewhat necessary for this method, as removing bounds can seriously reduce the algorithm's performance, but how to select reasonable ceiling values is an important and difficult issue, especially when the approach is applied to a new tutor environment. Second, it might be a productive step that we speed the algorithm up by first performing coarse-grained search, and after locating the promising regions refine the search by examining more digital decimals. Meanwhile, similar to the beam search, maintaining multiple promising regions simultaneously is beneficial to finding the optimal solutions.

Although this study failed to find PFA outperforms KT, one of our hypotheses is that perhaps PFA works better in the circumstance where questions for a particular skill vary greatly in difficulty. In this case, the question difficulty parameters in PFA might be able to differentiate student performance better and further achieve high predictive accuracy. One line of research is to consider integrating this concept with KT. Since it makes sense to be aware of the question difficulty when using a model to fit student performances, it potentially helps KT model capture more variance of the data, further leads to more plausible parameters and more accurate prediction.

6 Conclusions

PFA is an alternative approach to KT. In this study, we failed to show there are any real differences in predictive accuracy between PFA and a version of KT that attempts to deal with multi-skill questions. We were able to show that for fitting KT, EM achieves significantly higher predictive accuracy than brute force. We also found that, for multi-skill problems, considering the skill with the lowest proficiency was the superior approach for predictive accuracy.

Parameter plausibility is another comparison object in this study. We showed that PFA is the best method for estimating student knowledge parameters, but the regular PFA without any bounded values could result in negative learning rates in even half of the cases. KT+BF estimated more plausible parameters than KT+EM, and meanwhile it has the potential ability to achieve even higher plausibility as the current results were obtained based on the limited search space.

In conclusion, researchers should feel free to use either PFA or KT. PFA works well with a minor tweak of restricting learning rates to be non-negative. However, the use of KT requires a careful consideration of model fitting approaches for parameter estimation and the methods for the handling of multi-skill problems, as performance varies.

References

1. Beck, J. E.: Difficulties in inferring student knowledge from observations (and why you should care). Proceedings of the AIED2007 Workshop on Educational Data Mining, Marina del Rey, CA, pp.21-30.
2. Corbett, A., Anderson, J.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 1995. 4: pp. 253-278.
3. Beck, J. E., Chang, K.-m.: Identifiability: A Fundamental Problem of Student Modeling. Proceedings of the 11th International Conference on User Modeling, Greece
4. Rai, D, Gong, Y, Beck, J. E. : Using Dirichlet priors to improve model parameter plausibility. Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain, pp141-148.
5. Pavlik, P. I., Cen, H., Koedinger, K.: Performance Factors Analysis - A New Alternative to Knowledge. Proceedings the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, pp. 531-538
6. Learning decomposition
7. Cen, H., Koedinger, K., Junker, B.: Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. Proceedings the 8th International Conference on Intelligent Tutoring Systems, Jhongli, pp. 164-175.
8. Chang, K.-m., Beck, J., Mostow, J., Corbett, A.: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems: Intelligent Tutoring Systems: Intelligent Tutoring Systems Volume 4053/2006
9. Pardos, Z. A., Beck, J., Ruiz, C., Heffernan, N. T.: The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS. In Baker & Beck (Eds.) Proceedings of the First International Conference on Educational Data Mining. Montreal, Canada. pp. 147-156.
10. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124v
11. Baker, R. *Personal communication.* 2010.