Optimizing Student Models for Causality

Benjamin Shih [a] Kenneth Koedinger [a] Richard Scheines [a]
[a] *Carnegie Mellon University*

Abstract. Complex student models often include key parameters critical to their behavior and effectiveness. For example, one meta-cognitive model of student help-seeking in intelligent tutors includes 15 rules and 10 parameters. We explore whether or not this model can be improved both in accuracy and generalization by using a variety of techniques to select and tune parameters. We show that such techniques are important by demonstrating that the normal method of fitting parameters on an initial data set generalizes poorly to new test data sets. We then show that stepwise regression can improve generalization, but at a cost to initial performance. Finally, we show that causal search algorithms can yield simpler models that perform comparably on test data, but without the loss in training set performance. The resulting help-seeking model is easier to understand and classifies a more realistic number of student actions as help-seeking errors.

1. Introduction

Student models are a key component of intelligent tutoring systems. While student models are generally supported by prior research and domain expertise, they often involve parameters for which no such research or expertise exists. Incorrectly chosen parameter values can then lead to ineffective interventions. Automated learning techniques can optimize parameter values based on data from earlier studies, but the resulting models are often not human-readable and also may not generalize well[3]. Most algorithms also try to improve correlations, e.g. the correlation between tutor help and learning, but educational interventions require strong *causal* relationships.

Using a student model and log data from prior research[2], we show that while tuning parameters can improve the model's performance, the resulting model generalizes poorly. We then show that model reduction algorithms can improve the model's generalization, but that causal search algorithms can give similar generalization with simpler models, better performance, and more robustness.

We use two data sets from prior research with a help-seeking tutor. The first comes from an experiment where the experimental group was asked to explain their work by choosing a justifying theorem[1]. 40 students completed both the pre- and post-tests. The second data set comes from an experimental study using the same model to guide help-seeking interventions[6]. There were 30 students with pre-post scores. For the first data set, we include both the control and experimental groups. For the second data set, we only use the control group, as the experimental condition received a help-seeking intervention[6]. We will refer to these two datasets as "02" and "06", respectively.

The help-seeking model defined by Aleven et. al.[2] is a set of production rules. Each rule consists of an "AND" of thresholds. A transaction is a help-seeking bug if at least one rule is satisfied. See Figure 1 for an example. The model has 10 thresholds and 15 rules.

Rule: READ-HINT-QUICKLY
IF the student is engaged in a meta-cognitive problem
AND the current subgoal is to think about a hint
AND the student spent less than THRESHOLD seconds to read the hint
THEN Bug Message: "Slow down. Take some time to read the hint."

Figure 1. Example rule classifying a help-seeking error

Table 1. Original and Tuned Models

Model	Rules	02 Correlation	06 Correlation
Original	15	-0.60**	-0.07
Original + Tune	15	-0.62*	0.18

* $p < .0001$, ** $p < .00001$

2. Method and Results

Our basic approach is to consider four types of models: original, tuned, reduced, and causal. The original model comes from the Aleven et. al.[2]. The tuned model has parameters tuned based on log data. The reduced model has a subset of the original rules chosen by standard model reduction algorithms, while the causal model has a subset of rules chosen by causal algorithms. In general, causal algorithms search over equivalence classes of Bayesian networks to find graphs with strong causal relationships that fit the data. We use an algorithm called Greedy Equivalence Search (GES) that is effective in many situations[4] and that is provably correct in the asymptotic limit[5].

In the following tables, we list the number of rules in each model and the correlations on both the 02 and 06 data sets. The correlations are between the percentage of a student's actions classified as metacognitive errors and the student's pre-post learning gain, adjusted for pre-test scores. The models are labeled in terms of the procedures used in creating them, e.g., "Tune + Stepwise + Tune" corresponds to taking the original model, tuning its thresholds, performing a stepwise regression, and tuning the resulting model's thresholds again. The results for the original model and tuned model are shown in Table 1. The original model's performance on the 02 data set was promising ($p < .0001$), but the model does not generalize well to the 06 data set ($r = 0.18$). Tuning the model gave similar 02 performance, but reduced generalization to the 06 data set.

To improve generalizability, we used stepwise regression. The results are shown in Table 2. Performance on the test set improved immediately from model reduction, confirming that many of the rules were unnecessary or counterproductive. We then tried tuning the model before performing a stepwise regression, but the resulting 06 correlation was terrible. Overall, the stepwise regression results confirmed that there were too many rules in the original model, but did not offer a convincing solution.

While stepwise regression chooses rules based on statistical significance, we ideally want to choose rules based on their causal relationship with learning. Thus, we used the GES algorithm with $\alpha = .05$, resulting in a directed acyclic graph that represented the causal relationships between variables. We then chose only those rules that were directly causally linked to the learning gain. The results are shown in Table 3.

Running GES on the original model gave strong 02 performance ($p < .01$), but poor 06 performance. Tuning the resulting model gave a dramatic improvement in the 06 correlation, bringing the model up to par with the best previous model. The second

Table 2. Stepwise Reduced Models

Model	Num. Rules	02 Correlation	06 Correlation
Stepwise	5	-0.13	-0.22
Stepwise + Tune	5	-0.20	-0.23
Tune + Stepwise	5	-0.53**	0.09
Tune + Stepwise + Tune	5	-0.54**	0.03

* $p < .05$, ** $p < .001$

Table 3. GES Reduced Models

Model	Num. Rules	02 Correlation	06 Correlation
GES	3	-0.47*	-0.04
GES + Tune	3	-0.48*	-0.23
Tune + GES	2	-0.48**	-0.04
Tune + GES + Tune	2	-0.49**	-0.20

* $p < .01$, ** $p < .001$

GES + Tune	Tune + GES + Tune
READ-HINT-QUICKLY	READ-HINT-QUICKLY
TRY-STEP-AFTER-HINT-HI	TRY-STEP-AFTER-HINT-LOW
GLOSSARY-AFTER-HINT-LOW	

Figure 2. Best GES Models

approach, tuning the thresholds before running GES, gave very similar results. The two best GES models are shown in Figure 2. Both models are much smaller than the original model. While the original model has 15 rules, the two GES models have 3 rules and 2 rules. While the original model needs 10 thresholds, the two GES models need 4 thresholds and 3 thresholds. Both models also have very similar rules, suggesting a degree of robustness. Finally, the original model classified over 70% of all transactions as metacognitive errors[2], but intervening in 70% of all transactions would overwhelm students. The two GES reduced models classify only 37% and 42% of all transactions as metacognitive errors.

3. Conclusions

The original help-seeking model suffered from three major problems: the complexity of the model, its poor 06 performance, and the excessive classification of metacognitive errors. The two GES approaches improved the model on all three counts. For model complexity, they reduced the number of thresholds by a minimum of 60% and number of rules by a minimum of 80%. For 06 performance, they improved the correlation from -0.07 to between -0.20 and -0.23. And for classification, they reduced the percentage of metacognitive errors from over 70% to about 40%. In addition, the GES models all had similar rules and similar correlations across both data sets, indicating robustness.

Causal search can be applied to many types of student models, so long as earlier log data and learning gains are available. While the evidence is not complete on the efficacy of this approach, we demonstrated that the resulting models can be simpler and more

efffective than the original. More research is still needed on optimal model complexity, generalization to other contexts, and most importantly, on experimental validation of the resulting models.

References

[1] Aleven, V. Koedinger, K.R. (2002) "An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-based Cognitive Tutor." Cognitive Science, 26(2), 147-179.

[2] Aleven, V., McLaren, B.M., Roll, I., Koedinger, K.R. (2004) "Toward tutoring help seeking - Applying cognitive modeling to meta-cognitive skills." Proceedings of 7th Conference on Intelligent Tutoring Systems, 227-39.

[3] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) "Detecting Student Misuse of Intelligent Tutoring Systems." Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[4] Chickering, D. M. (2002) "Learning Equivalence Classes of Bayesian-Network Structures." Journal of Machine Learning Research, 2:445-498.

[5] Chickering, D. M. (2002) "Optimal Structure Identification with Greedy Search." Journal of Machine Learning Research, 3:507-554.

[6] Roll, I., Aleven, V., McLaren, B.M., Ryu, E., Baker, R.S., Koedinger, K.R. (2006) "The Help Tutor: Does Metacognitive Feedback Improves Students' Help-Seeking Actions, Skills and Learning?" Proceedings of 8th International Conference on Intelligent Tutoring Systems, 360-9.