The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks

Zachary A. Pardos, Neil T. Heffernan, Brigham Anderson, and Cristina L. Heffernan

Worcester Polytechnic Institute, Carnegie Mellon University, Worcester Public Schools
{zpardos,nth}@wpi.edu

Abstract. A standing question in the field of Intelligent Tutoring Systems and User Modeling in general is what is the appropriate level of model granularity (how many skills to model) and how is that granularity derived? In this paper we will explore models with varying levels of skill generality (1, 5, 39 and 106 skill models) and measure the accuracy of these models by predicting student performance within our tutoring system called ASSISTment as well as their performance on a state standardized test. We employ the use of Bayes nets to model user knowledge and to use for prediction of student responses. Our results show that the finer the granularity of the skill model, the better we can predict student performance for our online data. However, for the standardized test data we received, it was the 39 skill model that performed the best. We view this as support for fine-grained skill models despite the finest grain model not predicting the state test scores the best.

1 Introduction

There are many researches in the user modeling community working with Intelligent Tutoring Systems and using Bayesian networks to model user knowledge [3, 4, 7]. Greer and colleagues [6] have investigated methods for using different levels of granularity and ways to conceptualize student knowledge. We seek to address the question of what is the right level of granularly to track student knowledge. Essentially this means how many skills should we attempt to track? We will call a set of skills (and their tagging to questions) a skill model. We will compare different skill models that differ in the number of skills and see how well the different models can fit a data set of student responses collected via the ASSISTment system [8].

1.1 Background on the MCAS State Test and ASSISTment Project

We will be evaluating our models by using the 8th grade 2005 Massachusetts Comprehensive Assessment System (MCAS) mathematics test which was taken after the online data being used was collected. The ASSISTment system is an e-learning and e-assessing system [8]. In the 2004-2005 school year, 600+ students used the system about once every two weeks. Eight math teachers from two schools would bring their students to the computer lab, at which time students would be presented

C. Conati, K. McCoy, and G. Paliouras (Eds.): UM 2007, LNAI 4511, pp. 435–439, 2007.

with randomly selected MCAS test items. Each tutoring item, which we call an ASSISTment, is based upon a publicly released MCAS item which we have added "tutoring" to. We believe that the ASSISTment system has a better chance of showing the utility of fine-grained skill modeling due to the fact that we can ask scaffolding questions that break the problem down in to parts and allow us to tell if the student got the item wrong because they did not know one skill versus another. As a matter of logging, the student is only marked as getting the item correct if they answer the question correctly on the first attempt without assistance from the system.

2 Model Creation and Prediction

In April of 2005, a 7 hour "coding session" was staged where our subject-matter expert, Cristina Heffernan, with the assistance of the 2nd author, set out to make up skills and tag all of the 300 existing 8th grade MCAS items with these skills. Because we wanted to be able to track learning between items, we wanted to come up with a number of skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We imposed upon our subject-matter expert that no one item would be tagged with more than 3 skills. She gave the skills names, but the real essence of a skill was what items it was tagged to. This model is referred to as the 'April' model or the WPI-106. The National Council of Teachers of Mathematics and the Massachusetts Department of Education use broad classifications of 5 and 39 skill sets. The 39 and 5 skill classifications were not tagged to the questions. Instead, the skills in the coarse-grained models were mapped to the finer-grained models in a "is a part of" type of hierarchy, as opposed to a prerequisite hierarchy [3]. The appropriate question-skill tagging for the WPI-5 and WPI-39 models could therefore be derived from this hierarchy.

2.1 How the Skill Mapping Is Used to Create a Bayes Net

Our Bayes nets consist of 3 layers of binomial random variable nodes. The top layer nodes represent knowledge of a skill set with a background probability of 0.50, while the bottom layer nodes are the actual question nodes with conditional probabilities set to 0.10 for guess and 0.05 for slip. The intermediary 2nd layer consists of ALL[1] gates that, in part, allow us to only specify a guess and slip parameter for the question nodes regardless of how many skills are tagged to it. The guess and slip parameters were not learned but instead set ad hoc. When we later try to predict MCAS questions, a guess value of 0.25 will be used to reflect the fact that the MCAS items being predicted are all multiple choice, while the online ASSISTment items have mostly been converted from multiple-choice to text-input fields. Future research will explore learning the parameters from data.

[1] An 'ALL' gate is equivalent to a logical AND. The Bayes Net Toolkit (BNT) we use evaluates Matlab's ALL function to represent the Boolean node. This function takes a vector of values as opposed to only 2 values if using the AND function. Since a question node may have more than 2 skills tagged to it, the ALL function is used.

2.2 Model Prediction Procedure

A prediction evaluation is run for each model one student at a time. The student's responses are presented to the Bayes net as evidence and inference (exact join-tree) is made on the skills to attain knowledge probabilities. To predict each of the 29 questions we used the inferred skill probabilities to ask the Bayes Net what the probability is that the student will get the question correct. We get a *predicted score* by taking the sum of the probabilities for all questions. Finally, we find the percent error by taking the absolute value of the difference between predicted and actual score and dividing that by 29. The *Average Error* of a skill model is the average error across the 600 students.

3 Results

An early version of the results in this section (using approximate inference instead of exact inference and without Section 3.1) appears in a workshop paper [8]. The MAD score is the mean absolute difference between predicted and actual score. The under/over prediction is our predicted average score minus the actual average score on the test. The centering is a result of offsetting every user's predicted score by the average under/over prediction amount for that model and recalculating MAD and error percentage.

Table 1. Model prediction performance results for the MCAS test. All models' non-centered error rates are statistically significantly different at the p<.05 level.

Model	Error	MAD	Under/Over	Cent. Error	Cent. MAD
WPI-39	13.40%	3.89	↓ 1.9	12.28%	3.56
WPI-106	14.88%	4.31	↓ 1.7	14.12%	4.10
WPI-5	18.60%	5.39	↓ 4.2	13.98%	4.06
WPI-1	23.77%	6.90	↓ 5.0	18.70%	5.42

3.1 Internal/Online Data Prediction Results

To answer the research question of how well these skill sets model student performance *within the system* we measure the internal fit. The internal fit is how accurately we can predict student answers to our online question items. If we are able to accurately predict a student's response to a given question, this brings us closer to a computer adaptive tutoring application of being able to intelligently select the appropriate next questions for learning and or assessing purposes. Results are shown bellow.

Table 2. Model prediction performance results for internal fit

Model	Error	MAD	Under/Over	Cent. Error	Cent. MAD
WPI-106	5.50%	15.25	↓ 12.31	4.74%	12.70
WPI-39	9.56%	26.70	↓ 20.14	8.01%	22.10
WPI-5	17.04%	45.15	↓ 31.60	12.94%	34.64
WPI-1	26.86%	69.92	↓ 42.17	19.57%	51.50

The internal fit prediction was run similar to an N-fold cross validation where N is the number of question responses for that student. The network was presented with evidence minus the question being predicted. One point was added to the internal total score if the probability of correct was greater than 0.50 for that question. This was repeated for each question answered by the student. The mean absolute difference between predicted total and actual total score was tabulated in the same fashion as the section above. All the differences between the models in Table 2 were statistically significantly different at the $p < .05$ level.

4 Discussion and Conclusions

The results we present seem mixed on first blush. The internal fit showed that the finer grained the model, the better the fit to the data collected from the ASSISTment system. This result is in accord with some other work we have done using mixed-effect-modeling rather than Bayes nets [5]. Somewhat surprising, at least to us, is that this same trend did not continue as we expected in the result shown in Table 1. In hindsight, we think we have an explanation. When we try to predict the MCAS test, we are predicting only 29 questions, but they represent a subset of the 109 skills that we are tracking. So the WPI-106, which tries to track all 106 skills, is left at a disadvantage since only 27% of the skills it is tracking appear on the 2005 MCAS test. Essentially ¾ of the data that the WPI-106 collects is practically thrown out and never used. Whereas the WPI-39 can benefit from its fine-grained tracking and 46% of its skills are sampled on the 29 item MCAS test.

As a field we want to be able to build good fitting models that track many skills. Interestingly, item response theory, the dominate methodology used in assessing student performance on most state tests, tends to model knowledge as a unidimensional construct by allowing the items themselves to vary in difficulty (and other properties of items like discrimination and the probability of guessing). Some of our colleagues are pursuing item response models for this very dataset [1, 2] with considerable success, but we think that item response models don't help teachers identify what skills a students should work on, so even though it might be very good predictor of students, it seems to suffer in other ways.

5 Future Work

Our results suggest the 106 skill model as being best for internal fit while the 39 skill model is best for the MCAS test, however, a combination of models may be optimal. Building a hierarchy in an aggregate or prerequisite way [3] will likely best represent the various granularities of student understanding and comprehension. These levels of understanding may change over time, so a dynamic Bayes approach will be needed to model these changes as well as model the important variable of learning.

Acknowledgments. This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are those of the authors, and not those of any of the funders. This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team that helped make possible this dataset.

References

1. Anozie, N., Junker, B.W.: Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In: Beck, J., Aimeur, E., Barnes, T. (eds.) Educational Data Mining: Papers from the AAAI Workshop, Technical Report WS-06-05, Menlo Park, pp. 1–6. AAAI Press, Stanford, California, USA (2006)
2. Ayers, E., Junker, B.W.: Do skills combine additively to predict task difficulty in eighth grade mathematics? In: Beck, J., Aimeur, E., Barnes, T. (eds.) Educational Data Mining: Papers from the AAAI Workshop, Technical Report WS-06-05, Menlo Park, pp. 14–20. AAAI Press, Stanford, California, USA (2006)
3. Carmona1, C., Millán, E., Pérez-de-la-Cruz, J.L., Trella1, M., Conejo, R.: Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model. In: Ardissono, B., Mitrovic (eds.) User Modeling 2005. 10th International Conference, pp. 347–356. Springer, Heidelberg (2005)
4. Conati, C., Gertner, A., VanLehn, K.: Using bayesian networks to manage uncertainty in student modeling. User. Modeling and User.-Adapted Interaction 12(4), 371–417 (2002)
5. Feng, M., Heffernan, N.T., Mani, M., Heffernan, C.: Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models. In: Beck, J., Aimeur, E., & Barnes, T (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 57-66. Technical Report WS-06-05. ISBN 978-1-57735-287-7 (2006)
6. McCalla, G.I., Greer, J.E.: Granularity– based reasoning and belief revision in student models. In: Greer, J.E., McCalla, G.I. (eds.) Student Modelling: The Key to Individualized Knowledge–Based Instruction, pages, pp. 39–62. Springer, Berlin (1994)
7. Mislevy, R.J., Gitomer, D.H.: The role of probability-based inference in an intelligent tutoring system. User.-Modeling and User. Adapted Interaction 5, 253–282 (1996)
8. Pardos, Z.A., Heffernan, N.T., Anderson, B., Heffernan, C.L.: Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. Workshop in Educational Data Mining held at the Eight International Conference on Intelligent Tutoring Systems. Taiwan (2006)