Estimating Student Proficiency Using an Item Response Theory Model

Jeff Johns, Sridhar Mahadevan, Beverly Woolf

Computer Science Department
University of Massachusetts Amherst
Amherst, MA 01003 U.S.A.
{johns, mahadeva, bev}@cs.umass.edu

Abstract. Item Response Theory (IRT) models were investigated as a tool for student modeling in an intelligent tutoring system (ITS). The models were tested using real data of high school students using the Wayang Outpost, a computer-based tutor for the mathematics portion of the Scholastic Aptitude Test (SAT). A cross-validation framework was developed and three metrics to measure prediction accuracy were compared. The trained models predicted with 72% accuracy whether a student would answer a multiple choice problem correctly.

1 Introduction

Student modeling is defined as the system's belief about a learner's state of knowledge. This is one of the most important aspects of an intelligent tutoring system. Any pedagogical strategy must rely on an accurate model to understand the effect of different tutorial actions on student performance. Student models can be categorized into two broad categories: expert-centric or data-centric [14]. The expert-centric approach, which includes cognitive modeling and knowledge tracing [1, 10], relies on an expert to identify the skills required to solve each problem. The expert provides the structure of the model and possibly the parameters. The data-centric approach relies on using the data to uncover the structure relating student ability to performance. Examples of data-centric student models are structure-learned dynamic Bayesian networks [14], models learned using the Q-Matrix method [6], and Item Response Theory [16, 17] models. Data-centric models typically have far fewer parameters compared to expert-centric models.

In this paper, we evaluate the predictive power of IRT models. A data-centric model was selected to contrast with our previous work [13] using an expert-centric model. From [13], we concluded that robust parameter estimation was difficult given the ratio of the amount of data available from student logs to the model complexity (i.e. number of parameters). IRT models are an attractive alternative because they have a relatively small number of parameters. To confirm this hypothesis, we developed a cross-validation framework to quantify a trained model's predictive accuracy.

2 Item Response Theory

IRT models and their corresponding parameter estimation techniques have a long history of development in the psychometrics literature. The purpose of these models is to probabilistically explain an examinee's responses to test items via a mathematical function based on his/her ability. Assessment of an examinee's ability is the first step of student modeling in an ITS because student state is a prerequisite for creating a pedagogical strategy.

The following two subsections describe the specific model and parameter estimation procedure used in our work.

2.1 Model

IRT posits a static, generative model that relates a student's ability, θ, to his/her performance on a given problem, u_i, via a characteristic curve, $f(u_i \mid \theta)$. A graphical view of this model is shown in Figure 1. Circles represent continuous variables, squares indicate discrete variables, and shaded items are observed variables.

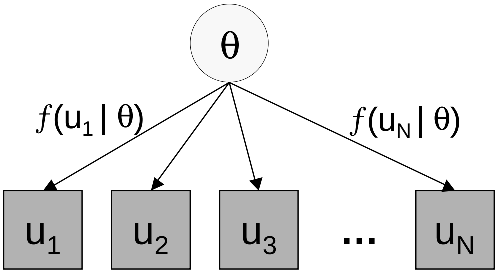

Fig. 1. Graphical depiction of an Item Response Theory model

In this work, we assume θ is drawn from a unidimensional normal distribution with mean 0 and variance 1. Experiments were also conducted with a multidimensional normal distribution, but those studies are not discussed in this paper. The random variables associated with each problem, u_i, come from a Bernoulli distribution with probability of a correct response (1) given by the following parameterized function.

$$P(u_i = \text{correct} \mid \theta) = f(u_i = 1 \mid \theta) = c_i + \frac{1-c_i}{1+\exp(-a_i(\theta - b_i))} \quad \textbf{(1)}$$

This is referred to as the three-parameter logistic equation, where a_i is the discrimination parameter that determines the slope, b_i is the difficulty parameter that determines the location, and c_i is the pseudo-guessing parameter that determines the lower asymptote. A plot of the function, with varying values of the discrimination parameter, is shown in Figure 2. Note that the two-parameter logistic equation is a special case of the three-parameter equation where c_i is set to 0. A more thorough description of the IRT model and the role of each of the parameters can be found in any text on the subject (i.e. [17]).

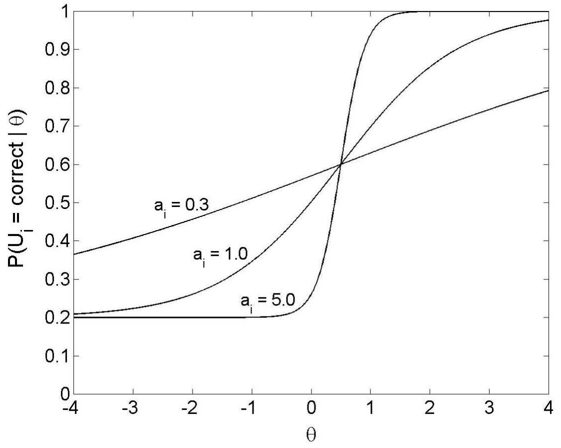

Fig. 2. Three parameter logistic function relating proficiency (θ) to the probability of a correct response. The three curves illustrate the discrimination parameter's effect, $a_i = \{0.3, 1.0, 5.0\}$, while keeping the other parameters constant at $b_i = 0.5$ and $c_i = 0.2$

2.2 Parameter Estimation

Marginal maximum likelihood estimation [8] is the most common technique used to learn the problem parameters (see [4] for a specific implementation of this algorithm). This is an instance of the expectation-maximization (EM) [11] algorithm where the hidden student variables (θ) as well as the parameters for each problem (a_i, b_i, c_i) are estimated simultaneously. The parameters are chosen to maximize the likelihood of the data.

In the most general case, the three parameters (a_i, b_i, c_i) are assumed to be constants that should be learned from the data. However, it is well known that jointly estimating parameters a_i and c_i can prove difficult. The estimates can be constrained however by assuming the parameters themselves have prior distributions. For example, the discrimination parameter, a_i, can be assumed to come from a lognormal distribution. The prior distribution assumption helps to avoid deviant parameter estimates by shrinking the values toward the specified mean of the distribution.

3 Design of Experiments

Experiments were designed to determine the effectiveness of IRT models at capturing a student's state of knowledge. Multiple experiments were conducted to find the most appropriate modeling assumptions given our dataset.

3.1 Domain and Data

The Wayang Outpost (`http://wayang.cs.umass.edu`) [2, 3] provides web-based tutoring on SAT mathematics problems. The tutor uses multimedia as a tool for engaging students and has been shown to be particularly beneficial for girls. Specifically, the tutor presents multiple choice geometry problems to students and offers them the option to seek help in solving the problems.

Data exists for 401 high school students and 70 multiple choice problems in the Wayang Tutor. Every student completed a minimum of ten problems and each problem was attempted at least thirty times. For each problem and each student, three pieces of information were recorded: number of mistakes made, number of hints requested, and the time spent. Furthermore, the order in which the students finished the problems was tracked. Problems were assigned randomly and a single problem was not given more than once to the same student (note that pairs of very similar problems do exist in the tutor). On average, a student completed 32 of the 70 available problems. The IRT assumption of static student proficiency is justified given this limited interaction with a student. Dynamic IRT models [12] or latent transition analysis models [9] that capture student learning could be used with longer data sequences.

The data was dichotomized because the relatively small sample size does not warrant using polytomous IRT models. A conservative dichotomization procedure was used: a response was labeled as correct only if the student's first action was to click on the correct answer. If the student answered incorrectly or asked for a hint, then the data point was labeled as incorrect.

3.2 Experiments

Four experiments were run with varying assumptions about the parameters in the logistic equation. The first two experiments use the two-parameter logistic equation while the last two experiments use the three-parameter equation. The first and third experiments assume a_i and b_i are constants to be estimated from the data. In the second and fourth experiments, the discrimination parameter, a_i, is assumed to come from a lognormal prior distribution with mean 1.1 and variance 0.6. The mean of 1.1 is a typical value for the discrimination parameter. These two experiments test whether constraints, in the form of prior distributions, help in estimating the parameters. Estimates for a_i that strongly deviate from the prior mean of 1.1 are penalized according to the lognormal distribution. This has the effect of shrinking a_i

estimates closer to the mean of the prior distribution. Table 1 summarizes these assumptions.

Table 1. Parameter assumptions for the four experiments

Experiment	a_i	b_i	c_i
1	constant	constant	N/A
2	~ lognormal(1.1, 0.6)	constant	N/A
3	constant	constant	0.2
4	~ lognormal(1.1, 0.6)	constant	0.2

The pseudo-guessing parameter, c_i, was not estimated during the parameter estimation process in Experiments 3 and 4. Given the small amount of data available, c_i was fixed at a value of 0.2 for each problem. This corresponds to an assumption of uniform guessing as there are five responses for each multiple choice problem.

3.3 Validation Framework

Five-fold cross validation was used to evaluate the IRT models learned in each of the four experiments. This means that ~320 students were used to train the model and ~80 students were used to test the model's predictive power. This process was repeated five times by rotating the training and testing populations such that each group of roughly 80 students was used once as the testing population.

Training the model involves running EM to learn the parameters a_i, b_i, and c_i for each problem. The testing procedure involves using the trained model to estimate a student's ability given performance on previous problems, and then to use the model again to predict how the student should fare on the next problem. The predicted response is compared with the actual student response. This is described in more detail in Figure 3.

```
Input:   $a_j$, $b_j$, $c_j$ for each problem
         Data (u) for each student in test population
Output:  ACC, MAE, MSE
for i = 1 to (# students in test population)
   // Assume $u_j^i$ refers to the i’th student’s response
   // (0 or 1) to the j’th problem he/she attempted
   for j = 2 to (max # problems student i performed)
      $\hat{\theta}_i$ = MLE of θ given $\left(u_1^i, a_1, b_1, c_1\right)$, … , $\left(u_{j-1}^i, a_{j-1}, b_{j-1}, c_{j-1}\right)$
      $p$ = $f\left(u = 1 \mid \hat{\theta}_i, a_j, b_j, c_j\right)$
      if ( $p$ ≥ 0.5)   then $\hat{u}$ = 1
                        else $\hat{u}$ = 0
      if $\left(u_j^i == \hat{u}\right)$       then correct += 1
                        else incorrect += 1
      MAE += $\left|u_j^i - p\right|$
      MSE += $\left(u_j^i - p\right)^2$
ACC = correct / (correct + incorrect)
MAE = MAE / (correct + incorrect)
MSE = MSE / (correct + incorrect)
```

Fig. 3. Pseudocode for the cross-validation framework. Note, MLE is short for maximum likelihood estimate

Three metrics were evaluated during the testing phase: accuracy, mean absolute error (MAE), and mean squared error (MSE). Accuracy compares the actual response with a predicted response, whereas MAE and MSE compare the actual response with the predicted probability of a correct response. A better model results in higher accuracy and lower MAE and MSE values. MAE and MSE are error metrics that provide a more granular explanation of the model’s accuracy than just the accuracy metric. To see this, consider an example where a student answers a problem correctly but the model predicted a 0.49 probability of a correct response. The accuracy metric will have an error of 1.0 whereas MAE will have an error of 0.51 (=1.0 – 0.49) and MSE will have an error of 0.26 (=0.51^2).

4. Results and Discussion

The results from the four experiments are shown in Table 2. Note that all three metrics track with one another; therefore, MAE and MSE do not provide additional insight compared to the accuracy metric for this dataset. However, it is still useful to track these metrics because they provide information on the sensitivity of the model (e.g. a MAE of 0.37 indicates the model is not predicting the probability of a correct response to be close to the extreme values of 0 and 1).

Table 2. Results averaged across the five cross-validation runs

Experiment	Accuracy	MAE	MSE
1	72%	0.37	0.19
2	72%	0.37	0.19
3	67%	0.40	0.23
4	71%	0.38	0.21

Experiments 1 and 2 produced the best average accuracy value of 72%. Both experiments used the two-parameter logistic equation. Experiment 1 assumed the parameter a_i was a constant whereas Experiment 2 assumed a_i came from a prior lognormal distribution. These two experiments were very robust to initial starting conditions for the parameters. Thus, the prior distribution (Experiment 2) did not provide additional lift over Experiment 1.

Experiments 3 and 4 resulted in accuracy values of 67% and 71% respectively. In this case, the prior distribution assumption on the discrimination parameter, a_i, had a significant effect. This occurred because several of the 70 problems had a_i estimates that either became very small (close to 0) or very large (close to the maximum allowable value of 30). The prior distribution helped to shrink some of those extreme values closer to the distribution's mean value.

The 72% accuracy from Experiments 1 and 2 can be compared with two simple baseline strategies:

1. Always predict the student answers incorrectly (i.e. the majority class label).
2. Predict based on a student's percentage of previous problems answered correctly (if percentage is ≥ 0.5, then predict a correct response).

The first strategy achieves 61% accuracy and the second strategy 67%. The 2-parameter IRT model significantly outperforms both baselines. To demonstrate significance, the Z-statistic was used assuming correct/incorrect predictions are modeled as binomial variables with an alpha value of 0.01 ($Z = 15.11 > Z_\alpha = Z_{0.01} = 2.33$ for strategy 1 and $Z = 6.89 > Z_\alpha = Z_{0.01} = 2.33$ for strategy 2). Accuracy values from 75% to 85% are reported in [15] for training IRT models with synthetic data. However, there are two significant differences between the synthetic datasets and the actual data used for this study. One, the sample size for the synthetic datasets is much larger. Two, there is presumably no off-task behavior (i.e. students "gaming" the system, [5]) in the synthetic datasets.

Given these differences, 72% accuracy is a good starting point for modeling the Wayang dataset.

Aside from the accuracy metric, we considered a more intuitive way to gauge the results of training the IRT model. The parameter b_i measures the difficulty level of a problem, where larger values correspond to a problem being more difficult. Another simple measure of difficulty that is not directly related to the IRT model is the percent of students who answered a problem incorrectly. Again, larger values of this metric indicate the problem is more difficult. The correlation between these two measures of problem difficulty across all 70 problems was $r = +0.68$ (using the b_i estimates from Experiment 1). This is a high correlation because the percentage incorrect metric does not account for the different students that did each problem, whereas the IRT model does. The EM algorithm appears to learn realistic values for the difficulty parameter b_i.

5. Conclusions

Dichotomous IRT models were used to estimate a student's proficiency in answering multiple choice questions. The results presented in this paper came from actual data of high school students using the Wayang Outpost, a SAT-style geometry tutoring system. A cross-validation framework was introduced to evaluate the predictive power of the student model.

The best results, which predicted a student's response with 72% accuracy, were achieved using the two-parameter logistic equation. Although the three-parameter equation is more expressive, there was not enough data to effectively learn the values of the parameters. The number of parameters (and thus complexity) of the student model should be determined through a cross-validation process. As more data is gathered over time, the complexity of the model can be incrementally increased. Longer sequences of data would also warrant use of dynamic IRT models that account for student learning.

Our prior research suggests that an expert-centric model must have a large amount of data to learn the parameters of a model with many hidden variables [13]. In contrast, IRT models posit a single hidden variable and a constrained function relating the hidden variable to performance. Based on this study, this data-centric model provided reliable and accurate estimates of a student's proficiency. In the future, we will investigate the relationship between expert-centric models and data-centric models given a finite amount of data from which to learn the model parameters.

We plan to implement the IRT model to estimate a student's proficiency while he/she interacts with the tutor. Different pedagogical strategies will be tested based on the student's proficiency to determine the impact on a student's gain from pretest to posttest. We are also extending the IRT model to capture a student's (unobserved and dynamic) motivation level. Intelligent tutors are in a unique position to measure engagement because they track the number of hints requested and the response time, both key variables in detecting "gaming" behavior. Several recent papers ([3], [7]) have proposed models of student engagement. However, student modeling as a whole will be enhanced by measuring proficiency and engagement in one unified model.

References

1. Anderson, J., Boyle, C., Corbett, A., and Lewis, M. Cognitive Modeling and Intelligent Tutoring. *Artificial Intelligence*, 42(1), 7-49 (1990).
2. Arroyo, I., Beal, C., Murray, T., Walles, R., and Woolf, B. Web-based Intelligent Multimedia Tutoring for High Stakes Achievement Tests. Proceedings of the Seventh International Conference on Intelligent Tutoring Systems, 468-477 (2004).
3. Arroyo, I., Murray, T., and Woolf, B. Inferring Unobservable Learning Variables from Students' Help Seeking Behavior. Proceedings of the Seventh International Conference on Intelligent Tutoring Systems (2004).
4. Baker, F. and Kim, S.-H. *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker, Inc. (2004).
5. Baker, R., Corbett, A., Koedinger, K., and Wagner, A. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System." Proceedings of the ACM CHI 2004 Conference on Human Factors in Computing Systems, 383-390 (2004).
6. Barnes, T. The Q-Matrix Method of Fault-Tolerant Teaching in Knowledge Assessment and Data Mining. Ph.D. Dissertation. North Carolina State University (2003).
7. Beck, J. Engagement Tracing: Using Response Times to Model Student Disengagement. Proceedings of the International Conference on Artificial Intelligent and Education (2005).
8. Bock, R. and Aitkin, M. Marginal Maximum Likelihood Estimation of Item Parameters: Applications of an EM Algorithm. *Psychometrika*, 46, 443-459 (1981).
9. Collins, L. and Wugalter, S. Latent Class Models for Stage-Sequential Dynamic Latent Variables. *Multivariate Behavioral Research*, 27(1), 131-157 (1992).
10. Corbett, A. and Anderson, J. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *Journal of User Modeling and User-Adapted Interaction*, 4, 253-278 (1995).
11. Dempster, A., Laird, N., and Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38 (1977).
12. Embretson, S. A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, 56, 495-515 (1991).
13. Jonsson, A., Johns, J., Mehranian, H., Arroyo, I., Woolf, B., Barto, A., Fisher, D., and Mahadevan, S. Evaluating the Feasibility of Learning Student Models from Data. *American Association for Artificial Intelligence Workshop on Educational Data Mining* (2005).
14. Mayo, M. and Mitrovic, A. Optimising ITS Behavior with Bayesian Networks and Decision Theory. *International Journal of Artificial Intelligence in Education*, 12, 124-153 (2001).
15. Rudner, L. An Evaluation of Measurement Decision Theory. `http://edres.org/mdt/home3.asp`.
16. Thissen, D. and Wainer, H. (Eds.). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates (2001).
17. van der Linden, W. and Hambleton, R. (Eds.). *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag (1997).