A Dynamic Mixture Model to Detect Student Motivation and Proficiency

Jeff Johns and **Beverly Woolf**
Computer Science Department
University of Massachusetts Amherst
Amherst, Massachusetts 01003
{johns, bev}@cs.umass.edu

Abstract

Unmotivated students do not reap the full rewards of using a computer-based intelligent tutoring system. Detection of improper behavior is thus an important component of an online student model. To meet this challenge, we present a dynamic mixture model based on Item Response Theory. This model, which simultaneously estimates a student's proficiency and changing motivation level, was tested with data of high school students using a geometry tutoring system. By accounting for student motivation, the dynamic mixture model can more accurately estimate proficiency and the probability of a correct response. The model's generality is an added benefit, making it applicable to many intelligent tutoring systems as well as other domains.

Introduction

An important aspect of any computer-based intelligent tutoring system (ITS) is the ability to determine a student's skill set and to tailor its pedagogical actions to address the student's deficiencies. Tutoring systems have demonstrated this ability in the classroom (VanLehn *et al*. 2005). However, even the most effective tutoring system will fail if the student is not receptive to the material being presented. Lack of motivation has been shown empirically to correlate with a decrease in learning rate (Baker, Corbett, & Koedinger 2004). While attempts to motivate a student by using multimedia and/or by couching the material as a game have proved partially successful, there is still significant room for improvement. In fact, these motivation tools can themselves cause undesirable behavior, where students uncover ways to game the system. This issue of motivation and performance is particularly relevant given the weight assigned to high stakes achievement tests, such as the Scholastic Aptitude Test (SAT), as well as other exams that can be required for graduation. Students use tutoring systems to practice for high-stakes tests but typically are not graded based on their performance, which leads to low effort. The concern of low motivation affecting performance is also being addressed by the educational assessment community (Wise & DeMars 2005).

Automated diagnosis is the first step in addressing a student's level of motivation. Several models have been proposed to infer a student's engagement using variables such as observed system use (time to respond, number of hints requested), general computer use (opening an Internet browser, mouse activity), and visual and auditory clues (talking to the person at the nearby computer). The purpose of this paper is not to point out new ways in which students display unmotivated behavior, but rather to provide a general, statistical framework for estimating both motivation and proficiency in a unified model provided that unmotivated behavior can be identified. We propose combining an Item Response Theory (IRT) model to gauge student proficiency and a hidden Markov model (HMM) to infer a student's motivation. The result is a very general, dynamic mixture model whose parameters can be estimated from student log data and can run online by an ITS. We validate the model using data from a tutoring system, but indicate the model can be applied to other types of data sets.

Background

The next two sections describe previous work in estimating motivation and provide a brief introduction to Item Response Theory.

Relevant Literature

Several models have been proposed to infer student motivation from behavioral measures. de Vicente and Pain (2000; 2002) designed a study where participants were asked to watch prerecorded screen interaction of students using their tutoring system. Participants in the study created over eighty inference rules linking motivation to variables that could be observed on the screen. The fact that so many rules could be derived purely from screen shots suggests that simple, inexpensive methods for estimating motivation should be useful for a tutoring system.

Conati (2002) presented a dynamic decision network to measure a student's emotional state based on variables such as heart rate, skin conductance, and eyebrow position (in contrast to the more easily attained data used by de Vicente and Pain). The structure and parameters of the model, in the form of prior and conditional probabilities, were set by hand and not estimated from data. The probabilistic model

applies decision theory to choose the optimal tutor action to balance motivation and the student's learning.

A latent response model (Baker, Corbett, & Koedinger 2004) was learned to classify student actions as either gaming or not gaming the system. Furthermore, instances of gaming the system were divided into two cases: gaming with no impact on pretest-posttest gain and gaming with a negative impact on pretest-posttest gain. The features used in the latent response model were a student's actions in the tutor, such as response time, and probabilistic information regarding a student's latent skills.

Arroyo and Woolf (2005) developed a Bayesian network using a student's observed problem-solving behavior and unobserved attitude toward the tutor. The unobserved variables were estimated from a survey that students filled out after using the tutor. Correlation between pairs of variables was used to determine the network's connectivity.

Beck (2005) proposed a function relating response time to the probability of a correct response to model student disengagement in a reading tutor. He adapted the item characteristic curve from IRT to include a student's speed, proficiency, response time, and other problem-specific parameters. The learned model showed that disengagement negatively correlated with performance gain.

These models embody different assumptions about the variables required to estimate student motivation (e.g. static versus dynamic models, complex versus simple features, user specified versus learned model parameters, generic versus domain specific models). The model proposed in this paper is different because it encompasses the following four principles which do not all exist in any one of the previous models. First, the model should estimate both student motivation and proficiency. These variables need to be jointly estimated because poor performance could be due to either low motivation or insufficient ability. Only one of the aforementioned models performs this function (Beck 2005). Second, the proposed model should run in real time. There exists a tradeoff between model complexity and expressiveness to ensure tutoring systems can take action at the appropriate time. Ideally, the model parameters should also be estimable from a reasonable amount of data. Third, the model should be flexible enough to easily include other forms of unmotivated behavior as researchers identify them. Fourth, motivation needs to be treated as a dynamic variable in the model. Empirical evidence suggests that a student's motivation level tends to go in spurts. For example, Table 1 shows actual performance data (initial response time, total time to click the correct answer, and number of incorrect guesses) of a single student doing multiple-choice geometry problems. The problems are not arranged according to difficulty; therefore, the obvious shift in the student's behavior after the seventh problem could be attributed to a change in motivation.

Problem	Initial Time (s)	Total Time (s)	Number Incorrect
1	40	40	0
2	44	44	0
3	13	13	0
4	19	19	0
5	7	9	4
6	22	22	0
7	35	35	0
8	2	3	2
9	2	2	0
10	3	4	1
11	2	4	4
12	2	3	3

Table 1: Data from a single student using the geometry tutor. Notice the change in behavior after the first seven problems

Item Response Theory

IRT models were developed by psychometricians to examine test behavior at the problem level (van der Linden & Hambleton 1997). This granularity is in contrast to previous work that examined behavior at the aggregate level of test scores. While IRT models encompass a wide variety of test formats, we focus in this paper on IRT models for dichotomous user responses (correct or incorrect).

Item Response Theory posits a static, generative model that relates a student's ability, θ, to his/her performance on a given problem, U_i, via a nonlinear characteristic curve, $f(U_i|\theta)$. IRT models are data-centric models (Mayo & Mitrovic 2001) because they do not presuppose a decomposition of problems into separate, required skills. Each problem in an IRT model is assumed independent of the other problems. The random variable θ is drawn from a normal distribution with a specified mean and variance. The random variables associated with each problem, U_i, come from a Bernoulli distribution with the probability of a correct response given by the following parameterized function (Equation 1, Figure 1).

$$P(U_i = correct|\theta) = c_i + \frac{1 - c_i}{1 + \exp(-a_i(\theta - b_i))} \quad (1)$$

This is referred to as the three-parameter logistic equation, where a_i is the discrimination parameter that affects the slope of the curve, b_i is the difficulty parameter that affects the location, and c_i is the pseudo-guessing parameter that affects the lower asymptote. Note that the two-parameter logistic equation is a special case of the three-parameter equation where c_i is set to zero. Consistent and efficient methods exist for estimating these parameters. A more thorough description of the IRT model, it's properties, and the role of each of the parameters can be found in any text on the subject (Baker & Kim 2004; van der Linden & Hambleton 1997).

Model

We propose a dynamic mixture model based on Item Response Theory (DMM-IRT). The probabilistic model consists of four types of random variables: student proficiency, motivation, evidence of motivation, and a student's response to a problem.

The latent variables in the student model correspond to proficiency (θ) and motivation (M_i). Proficiency is defined to be a static variable (note, if statistical estimates of proficiency are made online while a student uses the tutor, then each new data point causes the estimate to change, but the

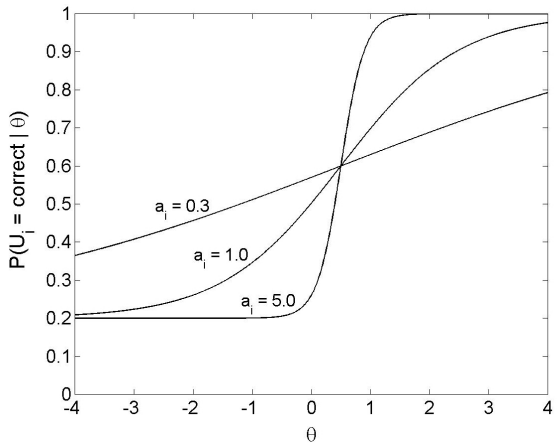

Figure 1: Three-parameter logistic function relating proficiency (θ) to the probability of a correct response. The three curves illustrate the discrimination parameter's effect while fixing the other parameters at $b_i = 0.5$ and $c_i = 0.2$

random variable is nonetheless assumed to be static). This assumption is also made in IRT modeling. In this paper, we assume θ has 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 because the small data set did not warrant more hidden dimensions. Student motivation is defined as a discrete, dynamic variable. The first-order Markov assumption is assumed to hold; therefore, motivation on the $(i+1)^{th}$ problem depends only on motivation on the i^{th} problem. The motivation variable can take on as many values as can be distinguished given the type of interaction allowed in the tutoring system. For the Wayang Tutor, we have identified three values for the motivation variable:

1. unmotivated and exhausting the hints to reach the final hint that gives the correct answer ('*unmotivated-hint*')
2. unmotivated and quickly guessing answers to find the correct answer ('*unmotivated-guess*')
3. motivated ('*motivated*')

The two unmotivated behaviors (1 and 2 above) are the most prevalent examples of how students game the Wayang Tutor. These are typical student behaviors and have been noted by other researchers (Baker *et al.* 2004). In Table 1, the student exhibits the *unmotivated-guess* behavior in the last five problems due to the short response times.

The student model uses two observed variables for each problem. The first variable is the student's initial response (U_i) which is either correct or incorrect. If a student's initial response is to ask for a hint, then U_i is labeled as incorrect. The second observed variable (H_i) is defined as the evidence corresponding to the hidden motivation variable, and thus takes on as many values as the motivation variable. For the Wayang Tutor, H_i has three values:

1. '*many-hints*' if the number of hints seen before responding correctly $> h_{max}$ (corresponds to *unmotivated-hint*)

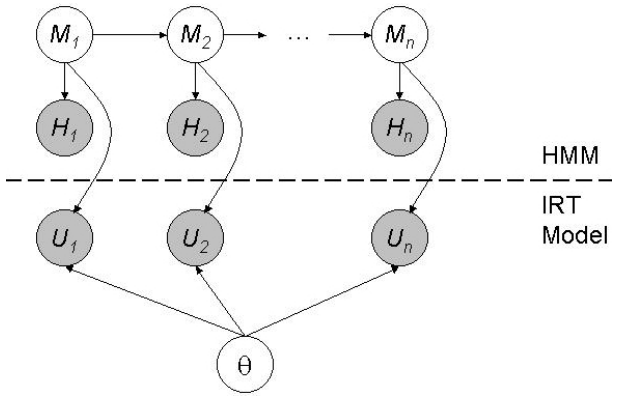

Figure 2: A graphical depiction of the dynamic mixture model based on Item Response Theory (DMM-IRT)

2. '*quick-guess*' if the number of hints seen before responding correctly $< h_{min}$ and if the time to first response $< t_{min}$ (corresponds to *unmotivated-guess*)
3. '*normal*' if neither of the other two cases apply (corresponds to *motivated*)

This variable is defined using the number of hint requests and time spent. Since the majority of intelligent tutoring systems already capture this data, the model has widespread use with minimal or no change to the system architecture. However, the random variables in this model would not change even if more sophisticated techniques or data sources were necessary to detect motivation. The conditions for the values of H_i would change to accommodate the new information.

The graphical model in Figure 2 describes how the observed and latent variables are connected. As the dashed line indicates, this model is a combination of a hidden Markov model (HMM) and an Item Response Theory (IRT) model. This is similar in structure to a switching state-space model (Ghahramani & Hinton 2000) with two exceptions: the continuous variable in the DMM-IRT is static instead of dynamic and the DMM-IRT uses a different function relating the continuous variable to the observed variables. The DMM-IRT is also similar to another class of models known as latent transition analysis (Collins & Wugalter 1992), or LTA. LTA employs dynamic, latent variables to capture changes in a student's ability over long periods of time. This is different from the DMM-IRT which assumes ability is static and motivation is dynamic.

The DMM-IRT is a mixture model where the mixtures are defined by the behavior the student exhibits. This is manifested in the model by the probability distribution $P(U_i|\theta, M_i)$. When the student is motivated, the distribution is described by the IRT item characteristic curve. We used the two-parameter logistic curve (Equation 2), but an alternate form can also be employed.

$$P(U_i = correct|\theta, M_i = motivated) = \frac{1}{1+\exp(-a_i(\theta - b_i))} \quad (2)$$

If the student is unmotivated, then the distribution takes

one of the following two forms.

$$P(U_i = \textit{correct}|\theta, M_i = \textit{unmotivated-guess}) = d_i \quad (3)$$

$$P(U_i = \textit{correct}|\theta, M_i = \textit{unmotivated-hint}) = e_i \quad (4)$$

The constant d_i corresponds to the probability of randomly guessing the answer correctly. This is equivalent to the pseudo-guessing parameter c_i in Equation 1. The constant e_i is the probability of a correct response given the *unmotivated-hint* behavior. The value of e_i should be close to zero since U_i is labeled as incorrect if the student's first response is to ask for a hint. A key model assumption is that both distributions described by the unmotivated behavior (Equations 3 and 4) do not depend on student proficiency (e.g. the two variables are uncorrelated). Hence, an action while in an unmotivated state will not alter the system's current estimate of a student's proficiency. If this independence assumption is correct, then the model does not underestimate student proficiency by accounting for motivation.

Parameter Estimation

Marginal maximum likelihood estimation (Bock & Aitkin 1981) is the most common technique used to learn the IRT problem parameters, a_i and b_i. For an implementation of this algorithm, see (Baker & Kim 2004). This is an instance of the expectation-maximization (EM) (Dempster, Laird, & Rubin 1977) algorithm where the hidden student variables as well as the parameters for each problem are estimated simultaneously. The parameters are chosen to maximize the likelihood of the data. We adapted the Bock and Aitkin procedure to include the latent motivation variables in the DMM-IRT.

The parameter estimation procedure iterates between the expectation step and the maximization step. In the E-Step, the probability distribution for the latent, continuous variable, $P(\theta)$, is integrated out of the likelihood equation. This integral is approximated using the Hermite-Gauss quadrature method which discretizes the distribution. For this paper, we used ten discrete points, or quadrature nodes, equally spaced over the interval $[-4, +4]$ to approximate the standard normal distribution. The E-Step results in estimates for the probability of a correct response at each of the quadrature nodes. These estimates are then used in the M-Step, which is itself an iterative process, to determine the problem parameters, a_i and b_i, that best fit the logistic curve.

Baker and Kim (2004) point out that this EM algorithm is not guaranteed to converge because IRT models using the two-parameter logistic equation are not members of the exponential family. We did not experience difficulty in getting the algorithm to converge, which is consistent with other reported findings.

Methodology

Domain

Experiments were conducted with data from high school students using the Wayang Outpost (Arroyo *et al.* 2004). The Wayang Outpost (`http://wayang.cs.umass.edu`) is an intelligent tutoring system for the mathematics section of the SAT. The tutor presents multiple-choice geometry problems to students and offers them the option to seek help in solving the problems. The data set consists of 401 students and 70 problems, where a student attempted, on average, 32 of the 70 total problems. For each problem that a student completed, the observed variables in the model (U_i and H_i) were recorded. The parameters used to determine H_i were set to $t_{min} = 5$ seconds, $h_{min} = 2$ hints, and $h_{max} = 2$ hints for all seventy problems. A slightly more accurate method would assign different values to different problems based on the number of available hints and problem difficulty.

Experiments

Three models were evaluated to determine their accuracy in predicting whether a student would correctly answer his/her next problem. We tested the DMM-IRT, an IRT model using the two-parameter logistic equation that does not measure student motivation, and a simple, default strategy of always guessing that the student answers incorrectly (the majority class label).

Given the relatively small size of the data set, not all the DMM-IRT parameters were estimated simultaneously. The discrete conditional probability tables associated with the HMM were estimated first using only the observed motivation variables, H. The learned values for these three distributions are shown below (where the order of values for M_i is $\textit{unmotivated-hint}$, $\textit{unmotivated-guess}$, and $\textit{motivated}$ and for H_i is $\textit{many-hints}$, $\textit{quick-guess}$, and $\textit{normal}$). The probability d_i (Equation 3) was set to 0.2 for all seventy problems, corresponding to an assumption of uniform guessing as there are five answers to each multiple-choice problem. The probability e_i (Equation 4) was set to 0.01 for all seventy problems. The item characteristic curve parameters a_i and b_i (Equation 2) were then estimated for each problem.

$$P(M_1) = \begin{pmatrix} 0.1 & 0.1 & 0.8 \end{pmatrix}^T$$

$$P(H_i|M_i) = \begin{pmatrix} 0.7 & 0.05 & 0.25 \\ 0.05 & 0.7 & 0.25 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$$

$$P(M_i|M_{i-1}) = \begin{pmatrix} 0.85 & 0.05 & 0.1 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}$$

Validation

Five-fold cross validation was used to evaluate the models. Thus, data from 320 students was used to train the models and 80 students to test the models' accuracy. The testing procedure involves using the trained model to estimate a student's ability and motivation given performance on previous problems (U_1, H_1, U_2, H_2, ..., U_{i-1}, H_{i-1}), predicting how the student will do on the next problem (U_i), and comparing the student's actual response with the predicted response. This process is repeated for each student in the test population and for each problem the student completed. The result is an accuracy metric (correct predictions divided by total predictions) that is averaged over the five cross validation runs. Pseudocode for this procedure is provided in Algorithm 1.

Algorithm 1 The cross validation framework

Input: a_j, b_j for each problem; U, H for each student
Output: accuracy
for $i = 1$ to (# students in test population) **do**
  // Assume U_j^i refers to the i'th student's response
  // (0 or 1) to the j'th problem he/she performed
  for $j = 2$ to (max # problems student i finished) **do**
    $\{\hat{\theta}, \hat{M}_j\} \leftarrow$ MLE given $(U_1^i, H_1^i, a_1, b_1), \ldots, (U_{j-1}^i, H_{j-1}^i, a_{j-1}, b_{j-1})$
    if $P(U_j = correct|\hat{\theta}, \hat{M}_j) \geq 0.5$ **then**
      $\hat{U} \leftarrow 1$
    else
      $\hat{U} \leftarrow 0$
    if $U_j^i == \hat{U}$ **then**
      correct $\leftarrow$ correct + 1
    else
      incorrect $\leftarrow$ incorrect + 1
accuracy $\leftarrow$ correct / (correct + incorrect)

Model	Cross Validation Accuracy		
	Average	Minimum	Maximum
Default	62.5%	58.2%	67.7%
IRT	72.0%	70.4%	73.6%
DMM-IRT	72.5%	71.0%	74.0%

Table 2: Average, minimum, and maximum accuracy values over the five cross validation runs

Results

The experimental results are shown in Table 2. The DMM-IRT achieved the best performance in terms of predicting the probability of a correct student response. It predicted with 72.5% accuracy whereas the IRT model had 72.0% and the baseline strategy of always predicting an incorrect response attained 62.5%. Accuracy values in the 70-80% range are generally good because both correct guesses and incorrect slips from knowledgeable students occur in multiple-choice tests.

The marginal improvement in performance by the DMM-IRT over the IRT model is not statistically significant. However, it is interesting that the DMM-IRT made different predictions. The model predicted more incorrect responses than the IRT model by accounting for student motivation. It is useful to consider how this occurs in the DMM-IRT. If a student is unmotivated and answers a problem incorrectly, then the model's estimate of proficiency does not decrease much. While this leads to larger estimates of student proficiency, this effect is offset by a decrease in the estimate of student motivation. The dynamics in the motivation variable allow the model's predictions (about the probability of a correct response to the next problem) to change more quickly than could be achieved via the static accuracy variable. While this lead to a modest improvement in prediction accuracy, we hypothesize that the difference could be greater with longer sequences where students perform more than 32 problems, which was the average for this data set.

The effect of motivation in the model can be explained by re-examining the student performance data from Table 1. This time, we show the smoothed estimates for the marginal probability of the motivation variable, $P(M_i)$. These values, shown in Table 3, are the probability of the student being in one of the three behavioral states: $motivated$, $unmotivated\text{-}guess$, and $unmotivated\text{-}hint$. After the seventh problem, the model believes the student is in an unmotivated state and can therefore adjust its belief about how the student will perform.

Problem	Initial Time (s)	$P(M_i = motivated)$	$P(M_i = unmotivated\text{-}guess)$
1	40	0.99	0.01
2	44	0.99	0.01
3	13	0.99	0.01
4	19	0.99	0.01
5	7	0.97	0.02
6	22	0.92	0.07
7	35	0.72	0.26
8	2	0.05	0.94
9	2	0.01	0.99
10	3	0.01	0.99
11	2	0.01	0.99
12	2	0.01	0.99

Table 3: Smoothed estimates of the student's motivation level, $P(M_i)$, for each of the twelve problems shown in Table 1. The remaining probability mass is associated with the $unmotivated\text{-}hint$ behavior

The data presented in Tables 1 and 3 is an ideal case of a student displaying two distinct, non-overlapping behaviors. This is easily captured by the model dynamics, $P(M_i|M_{i-1})$. Students do not always follow such ideal behavior. The top line in Figure 3 shows smoothed estimates of motivation for another student using the Wayang Tutor. The student's erratic behavior makes one-step predictions less reliable. Noisy dynamics may be hampering the DMM-IRT from achieving larger gains in cross validation accuracy. Figure 3 also demonstrates the difference in proficiency estimates between the IRT and DMM-IRT. The DMM-IRT model does not decrease its estimate of θ when the probability of the student being motivated is low (for example, during problems 26-31). The IRT model is potentially underestimating the student's ability.

Conclusions

We have presented and evaluated a dynamic mixture model based on Item Response Theory (DMM-IRT). This model uses a student's behavior to disambiguate between proficiency and motivation. Proficiency is modeled as a static, continuous variable and motivation as a dynamic, discrete variable. These assumptions are based on a student's tendency to exhibit different behavioral patterns over the course of a tutoring session. The DMM-IRT is a combination of a hidden Markov model and an IRT model. Given this generality, the model is easily tailored to many existing intelligent tutoring systems and can handle additional forms of unmotivated behavior. Users need only identify the evidence for the unmotivated behavior, $P(H|M)$, and the effect on the probability of a correct response, $P(U|\theta, M)$. The experiments,

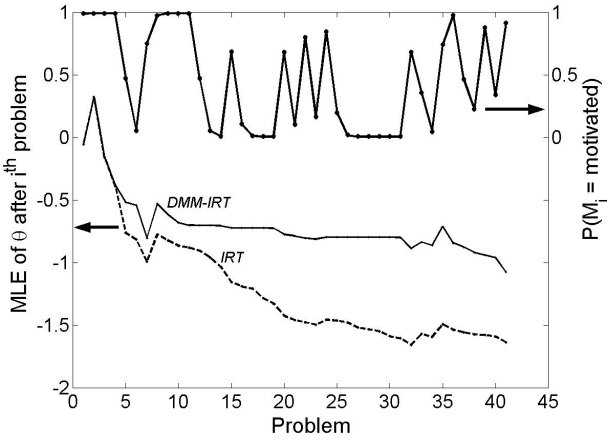

Figure 3: The top line (right y-axis) shows smoothed estimates of a student's motivation. The bottom two lines (left y-axis) show the maximum likelihood estimates of θ after i problems for the DMM-IRT (solid) and IRT (dashed)

though run offline in Matlab, were sufficiently fast that online inference in a tutoring system will occur in real time. These properties of the DMM-IRT satisfy the four principles discussed in the relevant literature section. We also point out that the model can be useful for other domains, such as object recognition, where a dynamic variable (e.g. lighting conditions) changes the context under which the static variable (e.g. shape of the object) is observed.

Experiments were conducted with real data of students using a geometry tutoring system. Results suggest that the DMM-IRT can better predict student responses compared to a model that does not account for motivation. The DMM-IRT produced increased estimates of student proficiency by assuming unmotivated behavior is independent of proficiency. This capability is important for preventing underestimation in low-stakes assessment, but further investigation is required to determine the correlation between these two variables.

In the future, we plan to implement the model to run online with the Wayang Tutor. Various strategies will be explored to determine (1) how quickly an unmotivated student can be re-engaged, and (2) the effect on the student's learning. We hope to improve cross-validation accuracy by updating the model's dynamics to better reflect student behavior.

References

Arroyo, I., and Woolf, B. 2005. Inferring Learning and Attitudes from a Bayesian Network of Log File Data. *In Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*, 33–40.

Arroyo, I.; Beal, C.; Murray, T.; Walles, R.; and Woolf, B. 2004. Web-based Intelligent Multimedia Tutoring for High Stakes Achievement Tests. *In Proceedings of the Seventh International Conference on Intelligent Tutoring Systems*, 468–477.

Baker, F., and Kim, S.-H. 2004. *Item Response Theory: Parameter Estimation Techniques*. New York, NY: Marcel Dekker, Inc.

Baker, R.; Corbett, A.; Koedinger, K.; and Wagner, A. 2004. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *In Proceedings of the ACM CHI 2004 Conference on Human Factors in Computing Systems*, 383–390.

Baker, R.; Corbett, A.; and Koedinger, K. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. *In Proceedings of the Seventh International Conference on Intelligent Tutoring Systems*, 531–540.

Beck, J. 2005. Engagement Tracing: Using Response Times to Model Student Disengagement. *In Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*, 88–95.

Bock, R., and Aitkin, M. 1981. Maximum Likelihood Estimation of Item Parameters: Applications of an EM Algorithm. *Psychometrika* 46:443–459.

Collins, L., and Wugalter, S. 1992. Latent Class Models for Stage-Sequential Dynamic Latent Variables. *Multivariate Behavioral Research* 27(1):131–157.

Conati, C. 2002. Probabilistic Assessment of User's Emotions in Educational Games. *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI"* 16(7-8):555–575.

de Vicente, A., and Pain, H. 2000. A Computational Model of Affective Educational Dialogues. *AAAI Fall Symposium: Building Dialogue Systems for Tutorial Applications*, 113–121.

de Vicente, A., and Pain, H. 2002. Informing the Detection of the Students' Motivational State: An Empirical Study. *In Proceedings of the Fifth International Conference on Intelligent Tutoring Systems*, 933–943.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society Series B* 39:1–38.

Ghahramani, Z., and Hinton, G. 2000. Variational Learning for Switching State-Space Models. *Neural Computation* 12(4):831–864.

Mayo, M., and Mitrovic, A. 2001. Optimising ITS Behavior with Bayesian Networks and Decision Theory. *International Journal of Artificial Intelligence in Education* 12:124–153.

van der Linden, W., and Hambleton, R., eds. 1997. *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

VanLehn, K.; Lynch, C.; Schulze, K.; Shapiro, J.; Shelby, R.; Taylor, L.; Treacy, D.; Weinstein, A.; and Wintersgill, M. 2005. The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence in Education* 15(3):147–204.

Wise, S., and DeMars, C. 2005. Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment* 10:1–17.