Modeling student engagement with a tutoring system using Hidden Markov Models

Carole Beal, Sinjini Mitra, Paul Cohen
USC Information Sciences Institute
Marina del Rey, California

DRAFT

Abstract

High school students' actions with a mathematics tutoring system were modeled with Hidden Markov Models. The results indicated that including a hidden state estimate of learner engagement increased the accuracy and predictive power of the models, both within and across tutoring sessions. Groups of students with distinct engagement trajectories were identified, and findings were replicated in two independent samples. Engagement trajectories were not predicted by prior math achievement. Results suggest that modeling learner engagement may help to increase the effectiveness of intelligent tutoring systems.

1 Introduction

Intelligent tutoring systems (ITS) are recognized as one of the "success stories" of artificial intelligence. A number of studies have demonstrated that a detailed model of the learner's knowledge of the domain can be used to provide individualized instruction, and to accelerate student learning well above what would be expected in a whole-class context. Traditionally, tutoring systems researchers have focused primarily on modeling the learner's cognitive processes while solving problems, i.e., the "model tracing" approach. However, there is growing recognition that learning involves more than cognition, and that students' actions with an ITS also reflect "engagement," meaning the transient shifts in attention and the emotions that are often associated with learning. For example, a student may become bored or fatigued over the course of a session and deliberately enter incorrect answers in order to elicit the correct answer from the system ("help abuse"). Another student may avoid using help features in the ITS even when he would benefit from doing so, because he prefers to attribute success in problem solving to his own efforts rather than to the resources available in the tutoring system.

Including estimates of engagement in learner models could be used to improve the effectiveness of tutoring systems, for example, by shifting the type of problem being delivered if boredom is detected, or by providing supportive feedback if the student appears to be frustration. However, relatively little is known about modeling learners' engagement while using an ITS. One approach has been to use sensors to capture physiological indices of emotional states such as interest, boredom, and frustration. Although very promising in the laboratory, this approach is difficult to scale for use by large numbers of students in real classroom situations in which intrusive and fragile equipment cannot be used. Here, we investigate an alternative approach: the use of students' action traces to model their engagement as they work with an ITS.

1.1 Hidden Markov Models

We focus on the use of Hidden Markov Models (HMM; [Rabiner, 1989]) because such models are designed to represent explicitly the influence of unobservable processes. We assume that engagement is a hidden state that influences student actions with the ITS. We fit HMMs to the action patterns of individual students using the student's hypothesized "engagement level" as the hidden variable with 3 possible states – "low", "medium" and "high". The output of an HMM can be used to make inferences as follows:

- The transition matrices help us determine the probabilities of students moving from one level of engagement to another, as well as of remaining in a particular engagement state while attempting one problem after another. For example, if a student's engagement level is known at time point t, the transition matrix will tell us his/her likely engagement level at time point (t+1).
- The emission matrices provide estimates of the probabilities of occurrence of a particular action when the student is in a particular engagement state.
- The Viterbi algorithm yields the maximum likelihood estimate of the most likely path of a student's engagement level throughout the duration of the problem-solving session.

1.2 Data sources

Our analyses focused on records of student actions with a tutoring system for high school mathematics. Data from two independent samples were available: The "MA" sample included 122 students who completed an average of 30 math problems. The "CA" sample included 91 students who com-

completed an average of 20 problems. The MA sample was more heavily weighted towards lower-achieving students, whereas the CA sample included a broader range of achievement levels. All students included in the analyses completed at least five math problems. The initial analyses focused on the action patterns for students in the first session.

In both samples, the data were automatically logged as students worked with the mathematics tutoring system during their regular school instruction. Students worked on a series of math problems; each problem included a figure, graph or table; the problem text and equation; and five answer options. On each problem, the student could choose answers and receive feedback on accuracy (e.g., a click on an incorrect answer elicited a red "X" whereas the correct answer elicited a green checkmark); request a series of multimedia hints leading to the solution; or move on to the next problem without answering.

Prior work had shown that a student's actions associated with a math problem could be machine-classified into five categories: Guessing/Help Abuse; Independent-Accurate problem solving; Independent-Inaccurate problem solving; Learning with Help; and Skipping, with 95% accuracy (Authors, 2006). For example, if the student clicks on an answer in less than 10 seconds after the problem loads on the screen, and then clicks on answers with inter-click intervals of less than 5 seconds, the action trace would be classified as Guessing. The latency intervals defined in the classification rules were selected on the basis of the performance of high-achieving students, e.g., if a proficient student requires at least 10 seconds to read the math problem before answering, a student who responds in a shorter time is unlikely to have read the problem and is likely to be guessing.

In the present study, the action and latency data recorded for each student on each math problem were machine classified into action patterns. Thus, each student's data record included an ordered sequence of action patterns representing her or his behavior on the math problems over the course of the tutoring session. For example, Student 2's data might be viewed as:

Problem	Action Pattern	Code
1	Guess	A
2	Guess	A
3	Skip	E
4	Independent-Accurate	C
5	Learn	B
6	Independent-Inaccurate	D
…	…	…

The action pattern sequence for this student would thus consist of a record [AAECBD…].

2 Results: Modeling hidden states

2.1 Fitting HMMs to individual student action pattern sequences

Here, we focused on how well the action patterns observed for the students could be modeled with an HMM. The average transition matrices for the two samples are shown in Tables X (CA) and X (MA). The rows represent time t, and the corresponding column values represent time t + 1.

Hidden state	Low	Medium	High
Low	0.3014	0.2423	0.4563
Medium	0.1982	0.4721	0.4176
High	0.1517	0.3312	0.6050

Table X. Average transition matrix for CA sample.

Hidden State	Low	Medium	High
Low	0.4727	0.1966	0.3307
Medium	0.1720	0.5583	0.2696
High	0.1470	0.2772	0.5758

Table X. Average transition matrix for MA sample.

Both transition matrices suggest that there is a degree of "inertia" in students' actions, i.e., the transition probabilities for persisting in the same state are generally higher than the probabilities of shifting to another state. For example, in both samples, the highest probabilities are observed for students who are in a high engagement state and are likely to persist in that state. However, the standard deviations for the transition matrices (not shown due to space restrictions) are quite high, suggesting that there is considerable individual variation. We address this issue in Section 2.3, below.

To assess the utility of the HMM for an individual student, we tested the model's ability at one point in time to predict future actions by that student. We used the transition and emission matrices for each student at each time step to predict the action pattern at the next step for problems M=16,…,S (S is taken as the mean number of problems per session for the two datasets – 20 for CA students and 30 for MA students). The results showed that the prediction accuracy was 42.13% for the CA sample, and 48.35% for the MA sample. The prediction accuracy is above chance (20%) for both samples, indicating that students' current actions do have value in predicting their future behavior with the tutoring system.

2.2 Comparison of HMM and MC models

We next compared the HMM models to simple Markov Chain models, which do not assume that the observed actions are influenced by a hidden state, i.e., "engagement". The individual transition and emission matrices for individual students at each time step were again used to predict the subsequent action pattern. As may be viewed in Table X, the individual MCs performed less well than the individual HMMs, although still somewhat better than chance. Thus, we conclude that the HMMs with the hidden variable (engagement level) is capable of modeling the variability in students' behavior during a problem-solving session more

effectively than the MCs, which do not involve such a hidden variable.

	Ind. HMM	Ind. MC	Group HMM	Group MC
CA	42.13%	33.43%	34.40%	18.73%
MA	48.35%	32.48%	26.00%	15.52%

Table X: Prediction accuracies using the HMMs and the Markov chain models for the two datasets.

2.3 Clustering students based on HMMs

As noted above, the mean transition matrices for the individual HMMs showed high standard deviations, suggesting that there is considerable variation in students' engagement trajectories. We next investigated if the variance was at the level of individual students, or at the level of groups of students with similar engagement trajectories. As a first step, based on the individual HMMs, we clustered students belonging to each dataset using the Kullback-Leibler (KL) distance between the individual transition matrices as the metric. Such groups represent students who follow similar trajectories through the different levels of engagement during an ITS session. We obtained 3 groups for the students from the CA sample, and 4 groups for the students from the MA sample, as shown in Table X.

Each group can now be characterized by a single transition matrix computed by taking the average of the individual transition matrices of students belonging to that group, and also by a single estimated optimal path through the engagement level. The group transition matrices for the CA and MA students appear in Tables X and X respectively.

We compared the prediction accuracy of individual HMMs to the group HMMs, using the group transition and group emission matrices corresponding to each student. Results are shown in Table X. Not surprisingly, the group HMM accuracy is not as good as the individual HMMs, but knowing the student's group does offer some predictive power, especially for the CA sample which includes a broader range of students. Note also that, by comparison, the predictive value of group MC models is very poor, again suggesting that including an estimate of a hidden state increases our ability to predict the student's actions with the ITS.

Groups	CA	MA
1	51	81
2	34	20
3	6	7
4	--	14
Total	91	122

Table X: Groups based on the HMM transition matrices.

Group	Hidden state	Low	Medium	High
1	Low	0.3641	0.1008	0.5351
N = 51	Medium	0.2105	0.5801	0.2094
56%	High	0.1957	0.2928	0.5115
2	Low	0.2606	0.4899	0.2495
N = 34	Medium	0.0384	0.3933	0.8036
37%	High	0.1124	0.4086	0.7143
3	Low	0.000	0.0417	0.9583
N = 6	Medium	1.000	0.0000	0.0000
7%	High	0.000	0.2189	0.7811

Table X: Group transition matrices for the CA students.

Group	Hidden state	Low	Medium	High
1	Low	0.5056	0.1089	0.3855
N = 81	Medium	0.1392	0.6637	0.1971
66%	High	0.0970	0.2201	0.6830
2	Low	0.2864	0.6809	0.0327
N = 20	Medium	0.0260	0.2839	0.6901
16%	High	0.2970	0.1455	0.5574
3	Low	0.1735	0.1429	0.6837
N = 7	Medium	0.9429	0.0000	0.0571
6%	High	0.0000	0.1805	0.8105
4	Low	0.7105	0.0344	0.2551
N = 14	Medium	0.1648	0.6152	0.2201
11%	High	0.3064	0.6936	0.0000

Table X: Group transition matrices for the MA students.

The estimated probabilities indicate that there are distinct groups of students that show different levels of engagement in a problem solving session. There is also considerable similarity in the groups across the two independent samples, as indicated in Table X. Both samples include one fairly large cluster of students. More specifically, the students tend to persist in medium and high engagement states (50-65%) ("steady state"), and the probability of moving to the low engagement state is about 20% or less for both samples.

A second group with a distinctly different profile is also observed in both samples: Students who tend to become more engaged ("increasing engagement"), and who have a very low probability of moving towards the low engagement state. For example, these students are most likely to move from the medium engagement state to high engagement (80% for CA, 69% for MA).

Both samples include a small third group characterized by strong shifts in engagement ("fluctuating engagement"). One shift is from low to high engagement (95% for CA, 68% for MA). However, these students also have a high probability of moving from medium to low engagement (100% for CA, 94% for MA).

Finally, the MA sample includes a fourth group ("declining engagement") characterized by persisting in a low engagement state (71%) and moving towards a less-engaged state, e.g., from high to medium, 69%. Recall that the MA sample included more students with low math achievement.

Group	Engagement	CA	MA

1	"steady state"	56%	66%
2	"increasing"	37%	16%
3	"fluctuating"	7%	6%
4	"declining"	--	11%

Table X. Percent of students in each engagement cluster

.

To evaluate the similarity of the groups across the two samples, we computed the KL distances between the 12 pairs, representing the 3 groups for CA and 4 groups for MA. The symmetrized KL distances for each pair of transition matrices are shown in Table X. The results show that the majority clusters ("steady state") are most similar (smallest KL distances) to each other, as are the other two clusters observed in both samples ("increasing" and "fluctuating").

	MA1	MA2	MA3	MA4
CA1	0.0667	0.9259	3.1276	1.9843
CA2	0.4902	0.2533	3.1584	3.2520
CA3	5.4051	6.9469	0.7876	9.7838

Table X. KL distances between HMM-based clusters across CA and MA samples.

2.3.2 Behavior of students in HMM groups

We next looked in more detail at differences in how students in the HMM clusters behaved with the ITS. The results were again fairly consistent across the two independent samples:

We first looked at whether there were differences in the amount of guessing behavior for students in the different HMM groups. Mean proportion scores are shown in Table X. (Rates for Problem Skipping were very low and are not analyzed here.) For the CA sample, an analysis of variance with HMM group as the grouping factor (Group 1, 2, 3) and proportion of actions classified as Guessing as the dependent measure showed a main effect of HMM Group, $F(2,97) = 4.346$, $p < .05$. Tukeys' HSD tests indicated that Groups 1 and 3 had a significantly higher proportion of Guessing than Group 2.

Similar results were observed for the MA sample. An analysis of variance showed a main effect of HMM Group, $F(3,109) = 9.436$, $p < .01$. Groups 1 and 3 again had significantly higher Guessing scores than Group 2. In addition, Group 4 showed significantly higher Guessing scores than the other three groups.

With regard to the effective use of multimedia help (Learn), there were no significant differences between the HMM clusters for the CA sample. For the MA sample, an analysis of variance showed a significant effect of HMM group, $F(3,109) = 4.968$, $p < .01$. Tukeys' HSD tests indicated that the means for Groups 1, 2 and 3 were similar to each other, and significantly higher than the mean for Group 4.

HMM Group:	Guessing	Learn	Ind.-Accurate	Ind.-Inaccurate
CA 1	.21	.20	.27	.24
CA 2	.10	.23	.39	.18
CA 3	.19	.30	.39	.09
MA 1	.22	.25	.19	.20
MA 2	.08	.39	.20	.21
MA 3	.09	.42	.22	.16
MA 4	.46	.13	.10	.18

Table X. Mean proportion scores for HMM Groups

With regard to Independent-Inaccurate problem solving, the samples differed somewhat. In CA, there was a main effect of Group, $F(2,97) = 3.784$, $p < .05$, although Tukeys' HSD tests showed no significant differences in the group means (perhaps due to the small number of students in Group 3). There were no significant differences on this measure for the MA sample.

Students' ability to solve math problems accurately without viewing multimedia help may be viewed as an indirect indication of their math proficiency. Analyses of variance conducted on the proportion scores for Independent-Accurate problem solving showed no difference for the HMM groups for either sample. However, as may be viewed in Table X, mean scores for the CA sample as a whole were generally higher than for the MA sample (0.35 and 0.24, respectively), consistent with the observation that the MA sample included more low-achieving students.

2.3.3 HMM clusters and math achievement

One possibility is that the HMM groups simply capture aspects of students' behavior with the ITS which can be traced to their mathematics proficiency. For example, students who do not have very good math skills might be more likely to view the multimedia help features; students with strong math ability might be more likely to solve problems accurately and independently, e.g., to enter the correct answer to a problem without viewing any of the help features. This plausible interpretation would be consistent with the model-tracing view of student learning, i.e., that students' actions reflect their cognitive understanding of the domain. However, in the present study we did not find a relation between students' math achievement and their HMM group membership.

In the CA sample, the students' classroom teachers had provided independent ratings of each student's math proficiency before the start of the study. Each student was rated by his or her teacher as above-average; average (student will pass the class); or below-average (student is in danger of failing the class). A chi-square analysis indicated that there was no significant association between the students' HMM cluster membership and their math achievement as rated by their teachers.

In the MA sample, students had completed a 42-item pre-test of math skills before starting to work with the ITS. A logistic regression analysis showed that there was no association between students' pre-test scores and their HMM

group membership. On the one hand, our conclusions must be somewhat tentative because the available information about the students' prior math skill was limited. On the other hand, the same result was observed in two independent samples, with different measures of achievement. This suggests that math proficiency does not completely explain students' behavior with the ITS, consistent with the view that ITS student models may be enhanced with estimates of engagement.

3 Results: Predicting students' engagement

The previous analyses indicate that we can model the student's engagement by using the student's action sequence. However, from the perspective of an ITS designer, this conclusion is somewhat limited in that it requires that we have the action pattern sequence already in hand in order to diagnose the learner's likely engagement transitions. Here, we describe additional work designed to predict a student's likely engagement trajectory, with the goal of diagnosing the student's behavior early enough that we could deploy interventions to sustain engagement.

3.1 Predicting session 2 from session 1

A second approach to the goal of predicting a student's behavior is based on students who have multiple sessions with the ITS. Recall that only the first sessions of these students were used to train the HMMs, and so we now use the second sessions of these students for validating our models using the fitted HMMs from the first sessions. There are 10 CA students and 25 MA students with 2 sessions that we use to assess the predictive abilities of our models using both the individual HMMs and the group HMMs. We compare our prediction results from the HMMs to those from Markov chain models (MCs) that do not take into account any hidden variable information. Table X shows these prediction accuracies.

	Individual HMM	Group HMM	Individual MC
CA	45.93%	30.12%	10.15%
MA	36.50%	25.22%	12.53%

Table X: Prediction accuracies on multiple sessions using the HMMs and the Markov chain models for the two datasets.

The highest accuracies are obtained with the individual HMMs, followed by the group HMMs. All these results are significantly better than pure random guessing (20% for 5 possible actions). The accuracies of the MCs are much worse than chance. This suggests that the fitted HMMs have good predictive power for future problem-solving sessions for both datasets, whereas the MCs fail completely in this regard.

3.2 Qualitative assessment of prediction accuracies

We now investigate whether there exists any association between the predictive ability of the HMMs on the first session and subsequent sessions; in other words, we would like to see if the prediction accuracy on the first session is able to predict the prediction accuracy on a future session. This can thus be used as a qualitative assessment tool for the prediction capacity of an HMM for future problem-solving sessions. For this, we consider individual HMMs only since they had the best accuracies for both sessions. Figure X shows a scatter-plot of the prediction accuracies in the two sessions for each individual student in each of the two datasets. The plots for both schools show a very high positive correlation between the prediction accuracies for the two sessions (significant Pearson's correlation coefficients of 0.8851 for CA and 0.7984 for MA). On the basis of this, we thus conclude that if the HMMs are good at predicting the first session, they are likely to have high prediction accuracies on subsequent sessions as well. On the other hand, if an individual's HMM is not good at predicting the first session, it is unlikely to be very useful for predicting performance in future sessions.

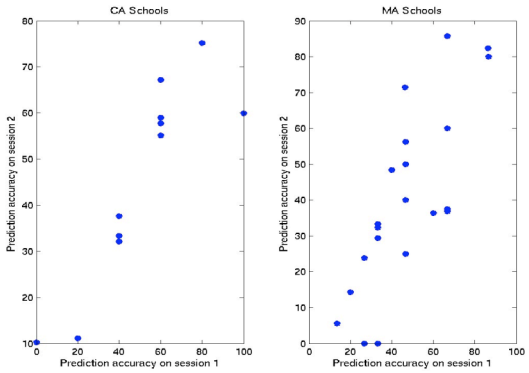

Figure X: Prediction accuracies (%) of the 2 sessions for the CA and the MA students. We used the sequential HMMs for session 1.

Comparing these individual prediction accuracies with the distributions of the action patterns on the 2 sessions, we find that students with high prediction accuracies for both sessions typically have low entropy between the two distributions (measured by KL distance), and vice versa for students who have poor accuracies on both sessions. There are, however, a few students with low accuracies on session 2 despite having medium accuracies on session 1, and these are students who miss certain action patterns on session 1 that appear on session 2, thus leading to poor prediction results. For example, consider the MA student having 75% accuracy on session 1 but only about 38% accuracy on session 2 (we call him "Student A" for easy reference). A look at the action data tells us that Student A only has actions 1, 2 and 3 on session 1, so based on the HMM fitted to such a session it will not be possible to estimate emission probabilities for actions 4 and 5 for future prediction. Now, Student A's session 2 contains frequent occurrences of actions 4 and 5, and therefore the HMM will never be able to pre-

dict these correctly and hence cause many errors. Although the prediction accuracies are not perfect for every student, the Session 1 data can at least be used to evaluate which students we can model prospectively and which ones we cannot.

4. Conclusions and future work

Our primary goal in this study was to use a student's actions with a tutoring system to estimate his or her engagement: an unobservable state that is hypothesized to influence the actions emitted by the student. Hidden Markov Models are an ideal analytic approach because these models include hidden states. Here, we fit HMMs to students' actions with a math tutoring ITS, with a three-level hidden state corresponding to "engagement". The results indicated that students' action sequences could be modeled well with HMMs, that their actions at one point in time could be successfully used to predict the subsequent action, and that the prediction accuracy of models with the hidden state were superior to simple Markov Chain models that did not include a hidden state.

We also found that, in some cases, the HMM models from session 1 could be used to predict future student engagement trajectories on session 2. Our conclusions were limited by the small number of students with intact multiple sessions, and the predictions were not strong for every student. However, the observed correlations were quite high, and we can identify those students for whom the predictions are likely to be valid. Thus, one conclusion from the study is that students' actions with the tutoring system can be modeled and that models that include estimates of engagement – here, defined as a hidden state parameter – are likely to be more useful for prediction than simple descriptive models.

By clustering the HMMs, we also identified groups of students with distinct trajectories through the hidden state space. Similar groups were observed in two independent samples from studies conducted in different school systems and geographical locations, suggesting that the patterns may be "real". The largest group included students who were generally engaged with the tutoring system, and whose engagement tended to remain stable over the course of the session. The second group was also generally engaged but also showed a stronger tendency to become more highly engaged over time. One interpretation is that these students do not really require any intervention from the ITS to support their learning. However, it should be noted that their actions, although indicating *engagement*, were not necessarily *effective*. In particular, these students guessed or attempted to solve the problems on their own, often with many errors, more than they used the multimedia help available in the ITS. In fact, overall, students used the help features on only about 1 in four of the math problems. This suggests that the next step will be to refine the models to distinguish effective and ineffective engagement states, and to encourage students to learn rather than to guess. This will be particularly critical for students whose actions indicate consistently low engagement with the ITS. We can then evaluate the impact of interventions by looking for shifts in a student's HMM cluster, e.g., moving from a low-engagement to a higher-engagement group, as well as through the proportion of different actions observed, e.g., increases in help-seeking.

The primary limitation of the study is that we do not have direct evidence that the hidden state included in the HMMs actually corresponds to any of the processes that have been suggested to reflect "engagement" in the learner, including transient shifts in the learner's attention, emotions, and cognitive effort. It is quite possible that the hidden state models reported here may reflect some other process or mechanism. We are currently investigating the relation of action patterns to gaze-tracking, self-reports of mood, and EEG estimates of cognitive workload and distractibility observed for the student on a math problem. However, we argue that the HMM models have utility regardless of the true nature of the hidden state, in that they can be used to identify students' patterns of actions, to group students into similar clusters and, in many cases, to predict students' actions in the future.

References

Authors, 2006. Masked for blind review.

Beck, J. E. (2005). Engagement tracing: Using response times to model student disengagement. In C-K. Looi et al. (Eds.), Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, pp. 88-95. Amsterdam: IOS Press.

Qu, L., & Johnson, W. L. (2005). Detecting the learner's motivational states in an interactive learning environment. In . In C-K. Looi et al. (Eds.), Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, pp. 547-554. Amsterdam: IOS Press.

Ramoni, M., Sebastiani, P., & Cohen, P. R., (2001). Bayesian clustering by dynamics. Machine Learning, 47, 91-121.

Rabiner. (1989, February). A tutorial on HMM and selected applications in speech recognition. In Proceedings of IEEE, Vol. 77, No. 2, pp. 257-286.

Vicente, A., & Pain, H. (2002). Informing the detection of the student's motivational state: An empirical study. In 6th International Conference on Intelligent Tutoring Systems, pp. 933-943. Biarritz, France.

Acknowledgements

We would like to thank project members < masked > for their assistance, as well as teachers and staff members in the following school districts: < masked >. Support for the research is provided by grants < masked for blind review >.