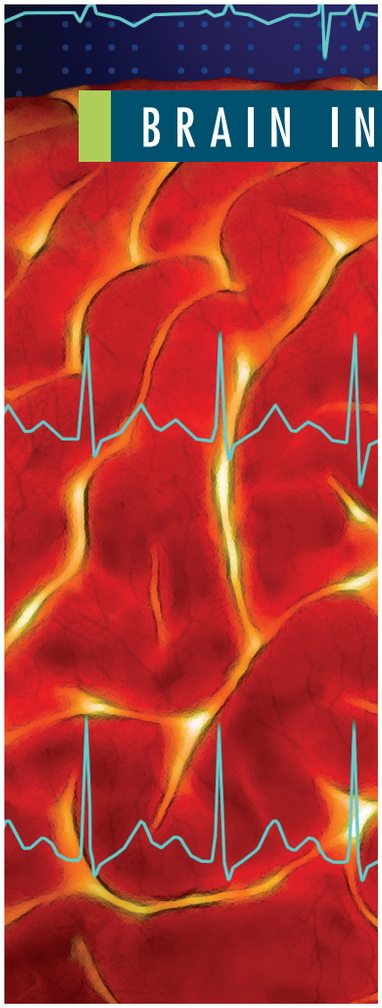

Using Brain Imaging to Interpret Student Problem Solving

John R. Anderson, Shawn Betts, Jennifer L. Ferris, and Jon M. Fincham,
Carnegie Mellon University

Jian Yang, *Beijing University of Technology*

Hidden Markov models can be used to combine behavioral and brain-imaging data from an intelligent tutoring system to track mental states during students' problem-solving episodes.

At Carnegie Mellon University, we have developed a successful approach to computerized instruction called Cognitive Tutors.[1] These widely used tutors focus on mathematics instruction. For instance, the Algebra Tutor,[2] which is currently deployed in more than 2,600 schools throughout the US, interacts with approximately 500,000 students each year. Cognitive Tutors are built on cognitive models that solve problems in the same way that students do. They individualize instruction using two processes. The first, *model tracing*, uses a model of students' problem solving to interpret their actions by finding a path of cognitive decisions that matches the observed actions. Given such an interpretation, the tutoring system provides real-time instruction individualized to where a student is in the problem. The second process, *knowledge tracing*, involves inferring which skills the student has mastered and then selecting new problems and instruction suited to that student's knowledge state.

Although the principle of individualizing instruction to a particular student holds great promise, the practice has been considerably limited by the ability to diagnose exactly what the student is thinking. The only information available to a typical tutoring system comes from the students' actions in the computer interface. The inference from such surface behavior to underlying thought is perilous.

We have been exploring whether multivoxel pattern analysis (MVPA)[3–8] of functional magnet resonance imaging (fMRI) data can be used to infer the mental states of students learning mathematics. This approach has shown considerable success in tracking static mental states such as whether a person is thinking about a location or an animal. Applying this to our case involves significant challenges not faced in many MVPA applications because it is necessary to track changing student states over time. The paths of states that students take in solving problems can be quite variable. Nevertheless, we have achieved relatively high accuracy in determining what step a student is on when solving a sequence of problems and whether that step is being performed correctly.

1541-1672/11/$26.00 © 2011 IEEE
Published by the IEEE Computer Society

Tracking Students' Mental States

Our approach involves combining two sources of information:

- A *behavioral model* generates sequences of problem-solving states in response to a problem. For instance, given the equation in Figure 1, the model would generate a sequence of transformations of the equation resulting in a solution. Reflecting the alternative correct and error transformations as in Figure 1, the model must be able to produce many possible sequences for an equation. In addition to representing the choices that students might make, the model represents the range of times students take in performing different steps when problem solving.
- *Brain imaging* involves sequences of fMRI scans that reflect brain activity occurring as students solve problems. Each scan involves whole images of brain activity, and we acquire these at intervals of about 2 seconds. This frequency of sampling is adequate because of the sluggish nature of the hemodynamic response that is tracked in fMRI scanning. fMRI's high spatial resolution compensates for this rather coarse temporal tracking by letting us track different activity in small regions of the brain.

Interpreting the student problem solving involves taking the brain-imaging data stream and identifying the mental states in the behavioral model.

Our research has used an experimental tutoring system that teaches a complete curriculum for solving linear equations based on Paul Foerster's classic algebra text.[9] (See previous work for more details on the experimental tutoring system.[10,11]) The tutoring system has a minimalist design to facilitate experimental control and detailed data collection. Nonetheless, it has the basic components of a cognitive tutor: instruction when new material is introduced, help upon request, and error flagging during problem solving. We are concerned with tracking students' mental states as they work through problems after the initial instruction in a section.

Students use a mouse for all tutor interactions—to select parts of the problem, select operations from a menu, and enter values from a numeric keypad. They cycle through four steps when solving an algebra problem. As a simple example, Figure 2 illustrates the following four steps in solving the equation $x - 10 = 17$:

1. *Selecting a transformation* involves selecting an operation called "unwind" and indicating that it applies to the whole equation.
2. *Executing the transformation* involves entering $x = 10 + 17$ by clicking on a keypad.
3. *Selecting evaluation* involves $10 + 17$ as the term to be evaluated.
4. *Executing the evaluation* involves entering 27 as the result.

In the example in Figure 2, solving the problem involves just one cycle of these four steps. More complex problems could involve many cycles of these four steps.

A Methodology for fMRI State Tracing

We developed a novel synthesis of three well-developed methodologies for following students' mental states.[12] Our approach consists of three components: hidden Markov models (HMMs), MVPA, and student modeling. We illustrate them here with respect to an example study, which

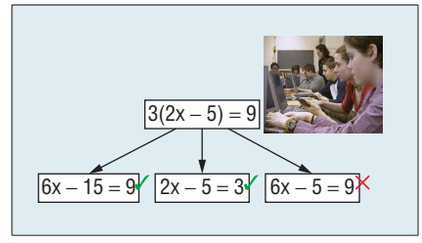

Figure 1. Tracking student progress. The Cognitive Tutor tracks students as they follow various correct and incorrect paths.

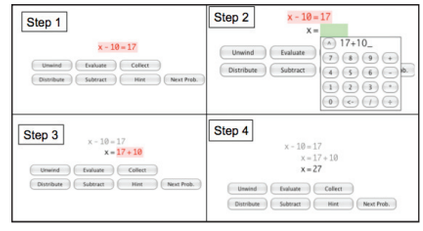

Figure 2. Algebraic problem solving. Each panel illustrates the state of the Algebra Tutor during one of the four steps of a problem solving cycle.

we described in detail in a previous ~~study~~.[13] That study followed 16 children going through a sequence of algebra problems with the tutor's assistance. They worked with the tutor over six days (days 0 to 5) and were scanned on days 1 and 5. This article focuses on interpreting the students' problem solving on day 5. To interpret a particular student's behavior on day 5, we combined information from other students on day 5 and data from that student on day 1. This is similar to the development and application of Cognitive Tutors, which are deployed with statistics based on pilot students. As a particular student progresses through the curriculum, Cognitive Tutors build a model of that student.

Our brain-imaging data come from blocks, which are sequences of about six problems. We used the brain-imaging data to determine what problem a student is working on and what step the student is performing within a problem. Determining what step the student is on is referred

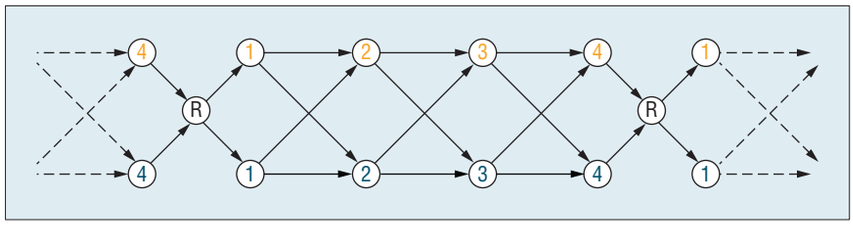

Figure 3. The behavioral model as a semi-Markov process. States correspond to steps (green correct, red incorrect) and rest periods (R).

to as the *segmentation goal*. We also used the imaging data to determine whether that step is being performed correctly. We refer to this as the *diagnosis goal*. We chose these goals because there is a hard definition of ground truth here—namely, the computer logs of the students' progress through these problems.

Component 1

We used HMMs to represent the students' nondeterministic progress through the problems. Figure 3 illustrates part of the HMM state structure used in the application under discussion. The figure shows the state structure for a block fragment that involves finishing a prior problem, transitioning to a rest state, stepping through one cycle of four steps to solve a problem, and returning to a rest state before the next problem. Each of the steps in solving the problem can be performed correctly or incorrectly. A block of problems was represented with a sequence of separate states for each problem in a block. A typical block might consist of 50 such states.

HMMs offer a powerful way to represent a student model because of the efficient algorithms for assigning probabilities to different possible state sequences. The critical feature of HMMs is their Markov property that the future course of problem solving only depends on the current state and not past history. Cognitive science models are not typically cast in a way that satisfies this Markov property, but nonetheless we have found ways to convert typical cognitive models into Markov state structures.

In the current example, we set the probabilities of state transitions for a particular student on the basis of other students' behavioral data. In addition to these transition probabilities, the HMM algorithms use the conditional probabilities of different observed data given different states to help determine the sequence of states. In our case, these are the conditional probabilities of different brain-imaging patterns. Different HMM algorithms[14] can find the probability that a student is in a particular state during any scan (suitable for model tracing), the most probable interpretation of a block of scans (suitable for knowledge tracing), or the probability of a set of data given a particular model (suitable for model evaluation).

In many applications (including the current one), we do not have access to when states change and must infer the boundaries between states as well as the states themselves. This requires using a HMM variant called *semi-Markov models* because the duration in a state is variable.[15] The same basic HMM algorithms apply with semi-Markov models except that we must, in effect, treat each different number of scans in a state as its own state. We also use other students' behavioral data to infer a probability distribution of durations in the different states.

Component 2

In performing our MVPA, we can use the tens of thousands of whole-brain images from different students going through different problems. We train multivoxel classifiers to associate different brain patterns with different mental states. These classifiers deliver the conditional probabilities that a given brain pattern comes from particular states, which is what is needed for the HMM logic. Several features distinguish our approach to this pattern classification problem:

- *State abstraction*. The number of states in the model can be large. The current example involves approximately 50 states for each block of problems. Even though there are tens of thousands of images, there is not enough data to recognize each pattern without running into severe overfitting problems. Therefore, it is necessary to find some abstraction of the specific states into a smaller number of states. This application used nine abstract states; one corresponded to the rest period between problems and the other eight were the four basic steps (see Figure 2) performed correctly or incorrectly.
- *Coarse, whole brain activation patterns*. Although we have enjoyed some success using specific predefined regions in some of our other work,[16] we throw away information in such a complex task if we do not use the activation over the full brain. We take the brain activity over approximately 400 megaregions, each a little more than a centimeter cubed. Using smaller regions does not yield much more information, and the sheer number of such regions leads to serious overfitting problems. Figure 4 illustrates the regions and their weights of association with being in an error state.
- *Linear discriminant analysis*. We use efficient LDA methods. We examined several methods sometimes associated with improved performance in the literature such as support vector machines (SVMs) with radial basis functions and other kernels.

These methods did not help, perhaps because of the large number of regions and scans in our dataset. Chih-Wei Hsu and colleagues also noted that SVMs do not give better results than linear classifiers when the number of features and instances are large.[17] LDA is much more efficient and produces the conditional probabilities that HMMs require.

- *Scan lag.* We train the classifier to associate states with the brain activity that follows 4 to 5 seconds after the student is in that state. This delay gives us optimal performance, which is not surprising given the lag of the hemodynamic response. The current example uses the activity two scans (4 seconds) later to classify the state the student was in during a particular scan. We have tried using multiple scans rather than a single scan to classify a target scan, but this typically results in overfitting the data.
- *Merging group and individual data.* The best performance comes from combining imaging patterns both from other students and from the particular student (in this case from day 1) to train the classifier. The data from specific students are useful because each student's activation patterns contain his or her own idiosyncrasies. However, there is not enough data from individual students to reliably train student-specific classifiers. In the current example, we equally weighted the imaging data from a particular student and the other 15 students.

Component 3

The probabilities of various state transitions and duration of residences within a state can be estimated from other students and a particular student's past behavior. In contrast to MVPA, where abstract states are needed to avoid overfitting, the best results come from using estimates of durations in specific states and transition probabilities between these states. This lets us capture the large differences among problems.

For instance, in the current example we obtained from other students statistics on state durations and error probabilities for each step of each problem. We tuned these to reflect the overall speed and accuracy of the particular student on day 1 to predict that student on day 5. Figure 5 shows the distribution of times and errors rates aggregated into three categories of problem difficulty. The continuous latency distributions fit these data, which let us calculate the probability of any duration for any problem.

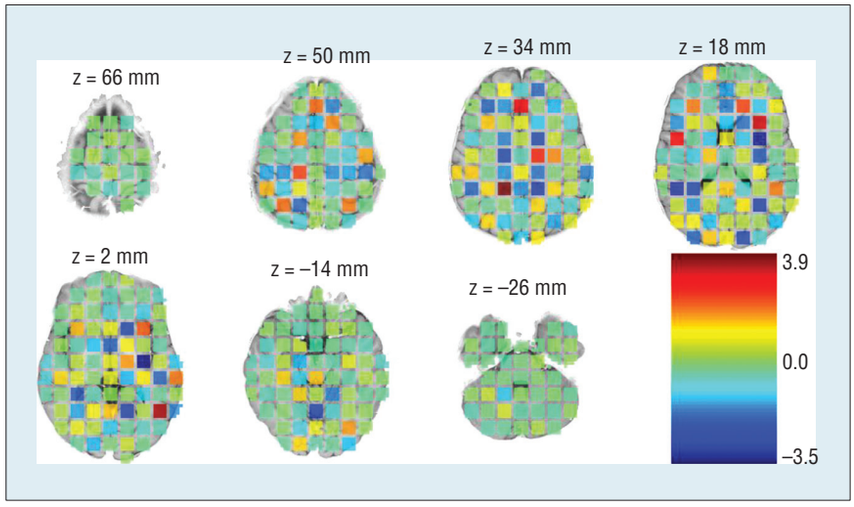

Figure 4. Output of multivoxel pattern analysis. Standardized differences between weights associated with error states and correct states.

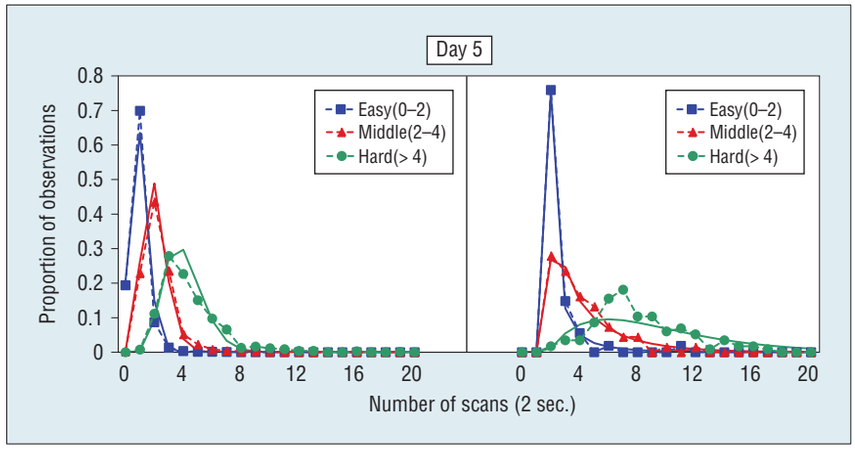

Figure 5. Distribution of (a) correct and (b) error step times for a student on day 5 as a function of the difficulty other students experienced with that step. The points connected by dotted lines are the proportions of observations with different number of scans. The smooth lines are fitted gamma functions.

Combining the Components

With this framework, we can calculate the probability of any interpretation of a sequence of scans. For example, consider a situation where a student is solving a sequence of problems that involves m scans, going through r steps, which are a subset of s states. In the current

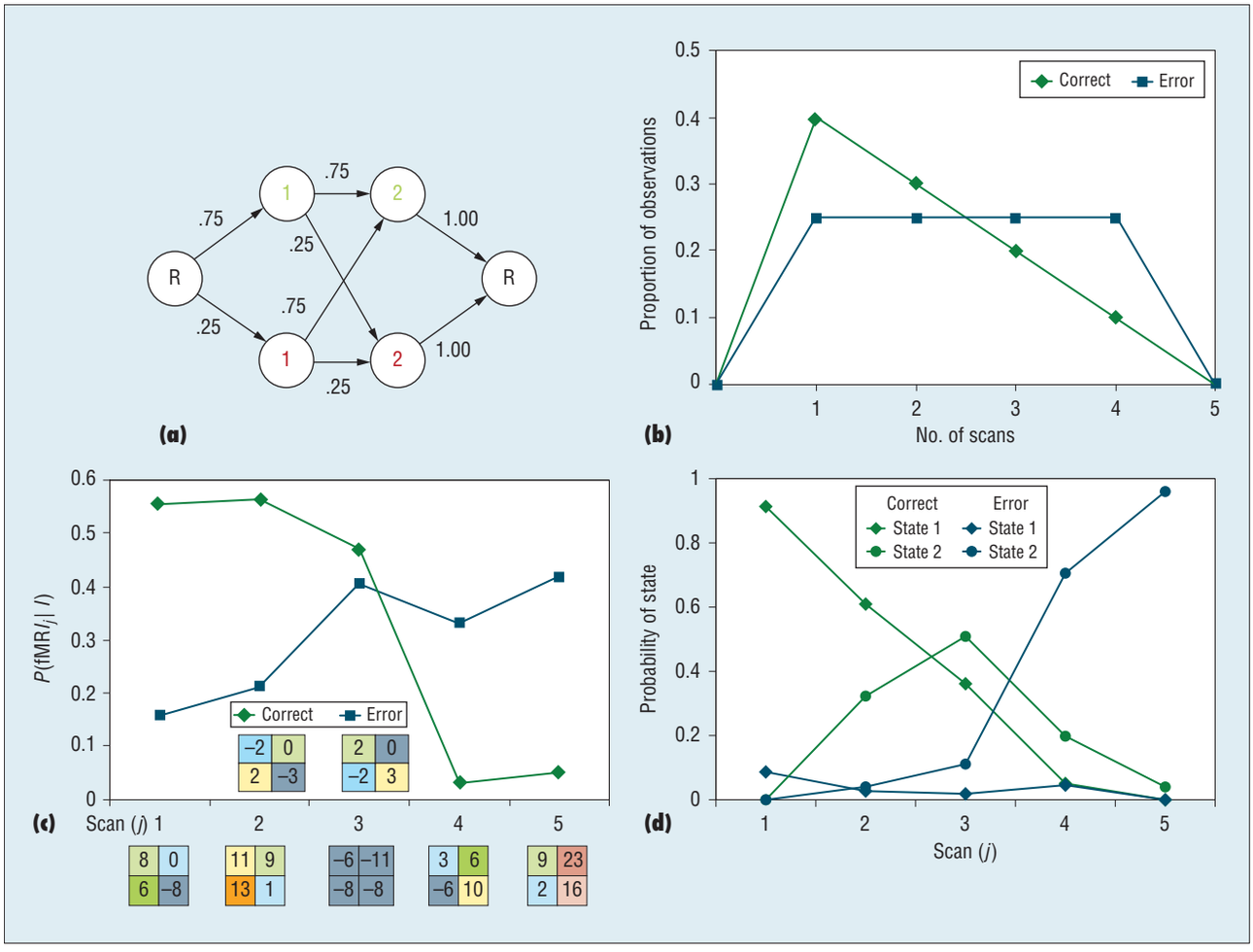

Figure 6. A simplified illustration of the components involved in the algorithm and their combination. The example involves a two-step problem, where each step can either be from an abstract correct or incorrect state. (a) A semi-Markov model containing the transition probabilities $t_{k,k+1}$ between steps (green correct, red incorrect). (b) The probabilities $p_k(a_k)$ that correct and incorrect steps will last different numbers of scans. (c) The conditional densities $p(\text{fMRI}_j|I)$ of the fMRI images given interpretations as coming from correct or incorrect states. Average image patterns are illustrated for correct and incorrect steps. Images for each scan are below the scan number. (d) The posterior probabilities that a scan comes from each of the four states given the images up to and including the image for that scan.

experiment, for a specific block of problems and a particular student, m would be on the order of 150 scans, r would be on the order of 30 steps, and s on the order of 50 states. An interpretation I of the m scans is an assignment of the scans to a subset of the s states.

Using a naive Bayes rule, we can calculate the probability of any interpretation I as the product of prior probability determined by the behavioral model and the conditional probabilities of the fMRI signals given the assignment of scans to states:

$$p(I \mid \text{fMRI}) \propto \left[S_r(a_r) \prod_{k=1}^{r-1} p_k(a_k) * t_{k,k+1} \right] * \left[\prod_{j=3}^{m+2} p(\text{fMRI}_j \mid I) \right]$$

The first term in the product is the prior probability (based on the behavioral model) and the second term involves the conditional probabilities (based on the imaging data's LDA). The term $p_k(a_k)$ in the prior probability is the probability that the kth interval is of length a_k, and $S_r(a_r)$ is the probability of the rth interval surviving at least as long as a_r. The term $t_{k,k+1}$ is the probability of transitioning from state k to $k + 1$. The second term contains $p(\text{fMRI}j|I)$, which are the probability density values for the fMRI signal on scan $j + 2$ given I's assignment of scan j to a state.

Figure 6 illustrates this computation for a hypothetical problem that

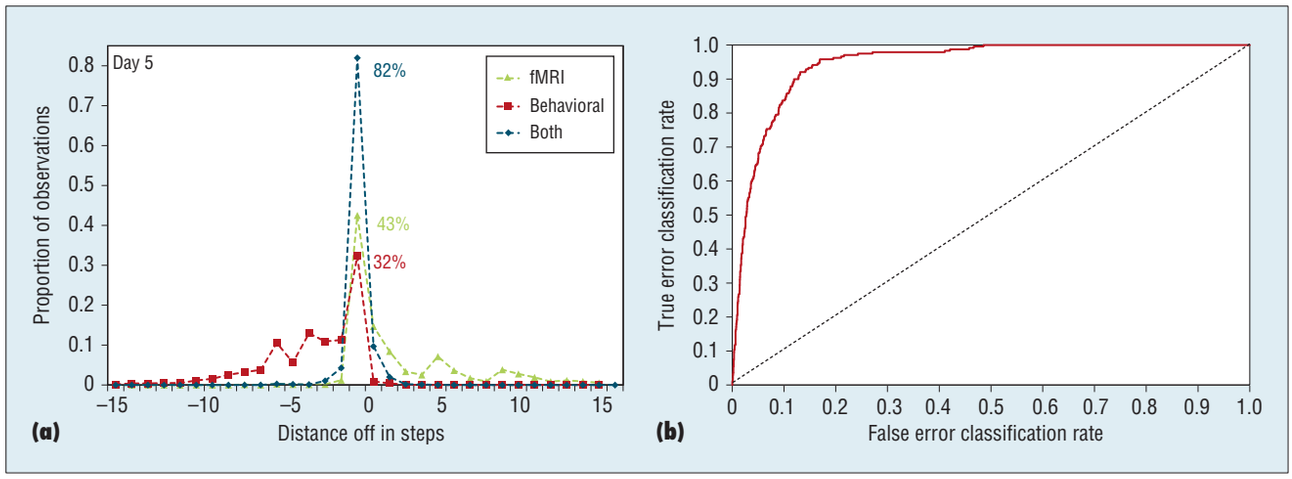

Figure 7. Algorithmic performance. We report our algorithm's performance on the (a) segmentation goal and (b) diagnosis goal.

takes five scans. Figure 6a shows a Markov model, simpler than Figure 3, with only two steps. It provides on its links the $t_{k,k+1}$ probabilities of transitioning from state k to $k + 1$. Figure 6b shows a distribution of q number of scans, simpler than Figure 4, with only two distributions, $p_k(a_k)$, for correct and incorrect steps. The durations of the steps range from one to four scans, with a mean of two for correct steps and 2.5 for errors. Figure 6c illustrates the calculation of $p(\text{fMRI}j|I)$, assuming four-voxel images rather than the 408-voxel images in Figure 3. This simplified version is still representative of the noise in images. It illustrates the images associated with the five scans for this problem, plus the mean image patterns in correct and error states. It also gives the densities for each scan under the hypotheses that the scan is correct and under the hypothesis this it is an error. (The calculation of these conditional probabilities by LDA depends on the covariance matrix of the training scans as well as the displayed means. The weights for corrects are –0.61, 0.26, 0.80, and –0.77; for errors the sign is inversed.) The forward HMM algorithm can combine the information in Figures 6a through 6c to assign a probability that each scan comes from each state (see Figure 6d). As we discussed in a previous work,[12] the probabilities in Figure 6d for scan j reflect the sum of all interpretations $p(I|\text{fMRI})$ from the beginning to scan j. The forward algorithm can be used to assign an interpretation, updated as each scan comes in.

We can also use the Viterbi algorithm to identify the most probable interpretation of the entire sequence. In this simple example, there are just 16 possible interpretations (four possible break points between the two steps and each step can either be correct or incorrect: $4 \times 2 \times 2$). The most probable interpretation assigns the two scans to a correct first step and the remaining three to an incorrect second step. The posterior probability of this interpretation is 0.408. The closest competitor to this interpretation, with a posterior probability of 0.315, assigns the first three scans to a correct first step and the remaining two to an incorrect second step. Although the third scan more likely comes from a correct step (Figure 6c) in line with this alternative interpretation, there is a higher probability of a two-scan correct step than a three-scan correct step (Figure 6b), in line with the more probable interpretation. Thus, both interpretations agree that the first step is correct and the second is in error, although there is this slight disagreement about segmentation.

Evaluation and Discussion

Although the simple example involves just 16 possible interpretations, the real data comes in blocks with an astronomical number of possible interpretations. The Viterbi algorithm can efficiently find the most probable. We have computer logs giving us exactly what the student did. We held this information back from the algorithm as a definition of ground truth and investigated whether the algorithm could predict the data in those logs. (Visit http://act-r.psy.cmu.edu/actrnews/index.php?id=34 to see a demonstration of our system's performance on one block of algebra problems, predicting the actual mouse clicks in the problem.) We performed separate evaluations of the segmentation goal (identifying what step a student is performing) and the diagnosis goal (determining whether that step is being performed correctly).

Figure 7 illustrates the algorithm's performance on these two dimensions. Figure 7a illustrates the algorithm's success at identifying where the student is in a sequence of problems that might take between 3 and 6 minutes to complete. The algorithm assigned each scan to a step of some problem, and the figure plots the mean difference between the assigned step and the true step. It shows how well the algorithm can do using just

THE AUTHORS

John R. Anderson is the Richard King Mellon Professor of Psychology and Computer Science at Carnegie Mellon University. He is known for developing ACT-R, which is the most widely used cognitive architecture in cognitive science, and as an early leader in research on intelligent tutoring systems. Anderson has a PhD in psychology from Stanford University. He is a past-president of the Cognitive Science Society and has been elected to the American Academy of Arts and Sciences, the National Academy of Sciences, and the American Philosophical Society. Contact him at ja@cmu.edu.

Shawn Betts is a research programmer at Carnegie Mellon University. His research interests include the design of instructional software. Betts has a BS in computer science from Simon Fraser University. Contact him at sabetts@andrew.cmu.edu.

Jennifer L. Ferris is a research associate at Carnegie Mellon University. Her research interests include the learning of mathematics. Ferris has a MA in psychology from Bucknell University. Contact her at jlferris@andrew.cmu.edu.

Jon M. Fincham is a research psychologist at Carnegie Mellon University. His research interests include the neural imaging and simulation modeling of complex cognitive processes. Fincham has PhD in psychology from Carnegie Mellon University, Contact him at fincham@cmu.edu.

Jian Yang is a lector in the International WIC Institute at the Beijing University of Technology. His research interests include manifold learning, machine learning, data mining, fMRI data analysis, and Web intelligence. Yang has a PhD in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences. Contact him at jian1yang@gmail.com.

statistics based on fMRI data, just behavioral data, and both. It illustrates the powerful multiplicative effect we get by combining behavioral and fMRI data.

Figure 7b illustrates the success of the approach on the diagnosis goal. We were able to vary the criterion for classifying a step as an error and so generate a curve giving the probability of a hit (classifying an error step as an error) as a function of the probability of a false alarm (classifying a correct step as an error). A measure of classification accuracy is the area under the curve, which is 0.5 for chance classification and 1.0 for perfect classification. The figure shows the level of performance is as high as 0.946.

The combination of MVPA and a behavioral model of the student can yield a fairly accurate diagnosis of where a student is in problem-solving episodes lasting many minutes. Moreover, prediction accuracy using both information sources was substantially greater than using either source alone.

The performance in Figure 7 should not be taken as the limit of what can be achieved. We could improve performance by enhancing the imaging data, adding additional data sources, or improving the behavioral model.

Achieving improved performance will be critical for tutoring applications of this methodology. Critical to instructional decisions are diagnosing mental states such as whether the student is confused, is doing critical problem solving, or has reached a point of routine procedural execution. This could help the system determine whether to offer help, let the students work on their own, or advance the students to a new curriculum objective. A challenge in making such discriminations is defining ground truth. The current example used computer records to decide what step the student was on and whether the step was in error. There are not similarly hard definitions of things such as whether a student is confused, but researchers are making progress on defining such states within the context of intelligent tutoring systems.[18]

Acknowledgments

This research was supported by a James S. McDonnell Scholar Award. We thank Julie Fiez for her comments on this research.

References

1. J.R. Anderson et al., "Cognitive Tutors: Lessons Learned," *J. Learning Sciences*, vol. 4, no. 2, 1995, pp. 167–207.
2. S. Ritter et al., "Cognitive Tutor: Applied Research in Mathematics Education," *Psychonomic Bulletin and Rev.*, vol. 14, no. 2, 2007, pp. 249–255.
3. C. Davatzikos et al., "Classifying Spatial Patterns of Brain Activity with Machine Learning Methods: Application to Lie Detection," *NeuroImage*, vol. 28, no. 3, 2005, pp. 663–668.
4. J.D. Haynes and G. Rees, "Predicting the Stream of Consciousness from Activity in Human Visual Cortex," *Current Trends in Biology*, vol. 15, no. 14, 2005, pp. 1301–1307.
5. J.D. Haynes et al., "Reading Hidden Intentions in the Human Brain," *Current Trends in Biology*, vol. 17, no. 4, 2007, pp. 323–328.
6. R. Hutchinson et al., "Modeling fMRI Data Generated by Overlapping Cognitive Processes with Unknown Onsets Using Hidden Process Models," *NeuroImage*, vol. 46, no. 1, 2009, pp. 87–104.
7. T.M. Mitchell et al., "Predicting Human Brain Activity Associated with the Meanings of Nouns," *Science*, vol. 320, no. 5880, 2008, pp. 1191–1195.
8. K.A. Norman et al., "Beyond Mind-Reading: Multi-Voxel Pattern Analysis of fMRI Data," *Trends in Cognitive Sciences*, vol. 10, no. 9, 2006, pp. 424-430.
9. P.A. Foerster, *Algebra I*, 2nd ed., Addison-Wesley Publishing, 1990.
10. J.R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* Oxford Univ. Press, 2007.

11. A. Brunstein, S. Betts, and J.R. Anderson, "Practice Enables Successful Learning Under Minimal Guidance," *J. Educational Psychology*, vol. 101, no. 4, 2009, pp. 790–802.
12. J.R. Anderson et al., "Neural Imaging to Track Mental States While Using an Intelligent Tutoring System," *Proc. Nat'l Academy of Science*, vol. 107, 2010, pp. 7018–7023.
13. J.R. Anderson et al., "Can Neural Imaging Be Used to Investigate Learning in an Educational Task?" to be published in *Expertise and Skill Acquisition: The Impact of William G. Chase*, J. Staszewski, ed., Psychology Press, 2011.
14. R.E. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, 1989, pp. 257–286.
15. K. Murphy, *Hidden Semi-Markov Models*, tech. report, 2002, MIT AI Lab.
16. J.R. Anderson et al., "Tracking Children's Mental States while Solving Algebra Equations," to be published in *Human Brain Mapping*, 2011.
17. C.W. Hsu, C.C. Chang, and C.J. Lin, "A Practical Guide to Support Vector Classification," Dept. of Computer Science and Information Eng., Nat'l Taiwan Univ., 2009.
18. A.C. Graesser et al., "The Relationship Between Affect States and Dialogue Patterns During Interactions with AutoTutor," *J. Interactive Learning Research*, vol. 19, 2008, pp. 293–312.