EEG Helps Knowledge Tracing!

Yanbo Xu, Kai-min Chang, Yueran Yuan, and Jack Mostow

Carnegie Mellon University
{yanbox, kkchang, yuerany, mostow}@cs.cmu.edu

Abstract. Knowledge tracing (KT) is widely used in Intelligent Tutoring Systems (ITS) to measure student learning. Inexpensive portable electroencephalography (EEG) devices are viable as a way to help detect a number of student mental states relevant to learning, e.g. engagement or attention. In this paper, we combine such EEG measures with KT to improve estimates of the students' hidden knowledge state. We propose two approaches to insert the EEG measured mental states into KT as a way of fitting parameters *learn*, *forget*, *guess* and *slip* specifically for the different mental states. Both approaches improve the original KT prediction, and one of them outperforms KT significantly.

Keywords: EEG, knowledge tracing, logistic regression

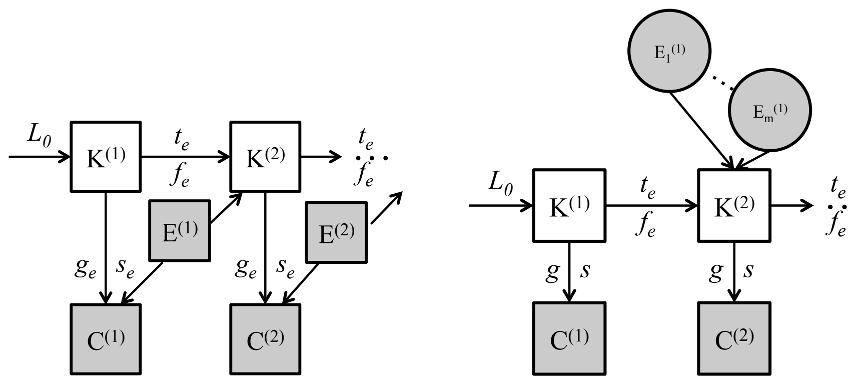
1 Introduction

Knowledge tracing (KT) is widely used in Intelligent Tutoring Systems (ITS) to measure student learning. In this paper, we improve KT's estimates of students' hidden knowledge states by incorporating input from inexpensive EEG devices. EEG sensors record brainwaves, which result from coordinated neural activity. Patterns in these recorded brainwaves have been shown to correlate with a number of mental states relevant to learning, e.g. workload [1], associate learning [2], reading difficulty [3], and emotion [4]. Importantly, cost-effective, portable EEG devices (like those used in this work) allow us to collect longitudinal data, tracking student performance over months of learning.

Prior work on adding extra information in KT includes using student help requests as an additional source of input [5] and individualizing student knowledge [6]. Thus for the first time, students' longitudinal EEG signals can be directly used as input to dynamic Bayes nets to help trace their knowledge of different skills. An EEG-enhanced student model allows direct assessment to be performed unobtrusively in real time. The ability to detect learning while it occurs instead of waiting to observe future performance could accelerate teaching dramatically. Current EEG is much too noisy to detect learning reliably on its own. However, as we show in this paper, combining EEG with KT allows us to detect learning significantly better than using KT alone.

2 Approach

KT is a type of Hidden Markov Model, which uses a binary latent variable ($K^{(i)}$) to model whether a student knows a skill at step i. It estimates the hidden variable from a sequence of observations ($C^{(i)}$'s) of whether the student has applied the skill correctly up to step i. In this paper, KT is used to capture the changes in knowledge state of a word over time (e.g., the school year), based on observations of whether or not the student read the word fluently (defined in more detail in Section 3). Standard KT usually has 4 (sometimes 5) parameters: initial knowledge (L_0), learning rate (t), forgetting rate (f) (usually set to zero, but not in this paper), guessing rate (g), and slipping rate (s). We add another observed variable ($E^{(i)}$), representing the EEG measured mental state that is extracted from EEG signals and is time-aligned to the student's performance at step i. We present two approaches to insert this variable into KT so that the student's hidden knowledge is inferred not only from the observed student's performance but also from the student's mental state measured by EEG.



(a) EEG-KT

(b) EEG-LRKT

Fig. 1: Add EEG measures into KT

Approach I: Insert 1-dimensional binary EEG measure into KT (EEG-KT). EEG-derived signals are often described as a type of measure for human mental states. For example, NeuroSky uses EEG signal to derive proprietary attention and meditation measures that indicate focus and calmness in students [7]. By adding a binary EEG input into KT, we hypothesize that a student can have a higher learning rate t given that the student is focusing at that step. Thus EEG-KT, shown in Figure 1a, extends KT by adding a binary variable $E^{(i)}$ computed from EEG input. We started with a binary (vs. continuous) EEG input for ease of implementation. This approach is reported in [8].

Approach II: Combine multi-dimensional continuous EEG measures in KT (EEG-LRKT). We also try an m-dimensional continuous variable $E^{(i)}$, denoting m EEG measures extracted from the raw EEG signal at step i. Xu and Mostow [9] proposed a method that uses logistic regression to trace multiple subskills in a Dynamic Bayes Net (LR-DBN). Without exploding the conditional probability tables in a DBN, LR-DBN combines the multi-dimensional inputs via a sigmoid function, which increases the number of parameters linearly (in number of inputs) instead of exponentially. This combination function was used in tracing multiple subskills [10]. Similarly, EEG-LRKT uses logistic regression to combine continuous EEG measures in KT. Figure 1b shows the graphical representation of EEG-LRKT, where circle nodes denote continuous variables. Hidden knowledge states are now determined by various EEG inputs. KT parameters t_e and f_e are computed by logistic regression over all m EEG measures.

3 Evaluation and Results

3.1 Data sets

Our EEG data comes from children 6-8 years old who used Project LISTEN's Reading Tutor at their primary school during the 2013-2014 school year [11]. We model the growth of students' oral reading fluency, by labeling a word as fluent if it was 1) accepted by the automatic speech recognizer (ASR) [12], as read 2) with no hesitation (the latency determined by ASR is less than 0.05s), and 3) without the student clicking on a word for help from the tutor.

EEG raw signals are captured by NeuroSky's BrainBand device at 512 Hz, and are denoised as in [11]. We use NeuroSky's proprietary algorithms to generate 4 channels: signal quality, attention, meditation, and rawwave. We then use Fast Fourier Transform to generate 5 additional channels from rawwave: delta, theta, alpha, beta, and gamma. We break EEG data into 1-second long segments, and filter out any segment with a poor EEG signal quality score (cutting off at 100 on the 0 to 200 signal quality scale provided by Neurosky). We then remove any observation for which more than 50% of its corresponding EEG signal is filtered out. We remove every word encounter whose next encounter (by the same student) has poor EEG signal quality, e.g. the first encounter of "cat" by a student is removed because the second encounter of "cat" by the same student has bad EEG quality. We keep only encounters whose next encounter had good signal quality, which reduces our data size by 1/3.

The original data set includes 16 students who read 600 distinct words. We discard 4 students who have fewer than 100 observations, resulting in 6,313 observations from 12 students. To maintain enough data for EM estimations of the parameters, however, we keep 4 students who have many more than 500 observations in the training data and cross validate the other 8 students.

3.2 Train classifiers as an extra EEG measure

We train Gaussian Naive Bayes classifiers to predict fluency. We compute the average and variance of the values of each of the 8 channels (excluding signal quality) over the duration of each word according to ASR as the classifier features (16 features in total). The validation is between-subject (i.e. training on all but one subject and testing on that remaining subject). Because the large majority class in this dataset will create overpowering priors, we pre-balance our data using under-sampling. This classifier uses a similar training pipeline as [11] with a few notable differences: 1) no feature selection due to the large training set; 2) to account for individual differences, we normalize every feature by converting features to z-scores over the distribution of that feature for that subject. Normalization is done before we train our classifier.

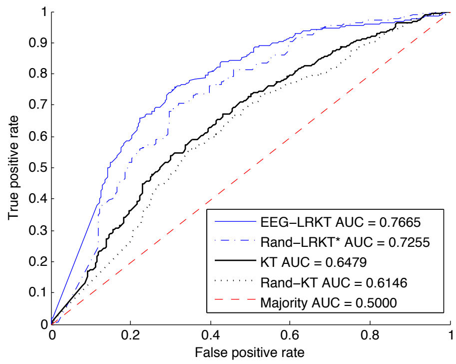
The classifier has a prediction accuracy of 61.8%. We evaluate it against a 50:50 chance classifier since we train the classifier on pre-balanced data. Our classifier performs significantly above chance on a Chi-squared test ($p < 0.05$). Finally, in Eq. 1, we compute a confidence-of-fluency (Fconf) metric as our 9th EEG measure and use it in the same way as the above 8 EEG scalar features:

$$\text{Fconf} = \Pr(\text{fluent}|2 \times 8 \text{ features}) - \Pr(\text{disfluent}|2 \times 8 \text{ features}) \tag{1}$$

3.3 Model fit with cross validation

We compare EEG-KT and EEG-LRKT to KT on a real data set. We normalize each EEG measure within student by subtracting the measure's mean and dividing by the measure's standard deviation across each student's observations. As EEG-KT requires, we discretize each measure as a binary variable: TRUE if the value is above zero; FALSE otherwise. We individually insert each of the binary EEG measures into KT and obtain in total 9 EEG-KT models: ATT(ention)-KT, MED(itation)-KT, RAW-KT, Delta-KT, Theta-KT, Alpha-KT, Beta-KT, Gamma-KT, and Fconf-KT. EEG-LRKT directly combines the 8 normalized EEG measures (excluding Fconf). Besides, we fit Rand-KT and Rand-LRKT, which replace EEG with randomly generated values from Bernoulli and standard Normal distributions respectively. We use EM algorithms to estimate the parameters, and implement the models in Matlab Bayesian Net Toolkit for Student Modeling (BNT-SM) [13, 10].

We conduct a leave-one-student-out cross validation (CV), which trains word specific models on 11 out of 12 students and tests on the remaining single student. We use receiver operating characteristic (ROC) curve and area under the curve (AUC) to assess the performance of model prediction (i.e., binary classification) since we have an unbalanced data with 83% labeled as fluent. Since we do not change the parameter of initial knowledge (L_0) in EEG-KT or EEG-LRKT, we clamp L_0 to 0.4 in our experiments in order to assess only the effect of those modified KT parameters. To test the statistical significance of differences between the proposed models and KT, we do two-tailed paired t-tests on AUC scores across the 8 students. EEG-LRKT significantly outperforms KT; the other 8 EEG measures and Rand-KT do not differ significantly from KT. Rand-LRKT seems to have a high AUC, but lacks results for half of the tested skills because of rank deficiency when fitting random values with logistic regression in DBN. Figure 2a shows a ROC graph with only the models that have significantly better AUC scores than KT with 8-fold CV; Table 2b shows a full list of AUC scores.



(a) ROC curves

Models	AUC	Models	AUC
EEG-LRKT	0.7665	Beta-KT	0.6355
Rand-LRKT*	0.7255	Gamma-KT	0.6317
Fconf-KT	0.6613	RAW-KT	0.6275
Theta-KT	0.6568	MED-KT	0.6230
KT	**0.6479**	Delta-KT	0.6224
ATT-KT	0.6435	Rand-KT	0.6146
Alpha-KT	0.6429		

(b) AUC scores

Fig. 2: Model fit comparison by 8-fold CV (Underlined if p-value <0.05 in paired T-test with KT; Rand-LRKT (starred*) is based on incomplete test data)

4 Conclusion and Future Directions

In this paper, we combine EEG measures with KT to improve estimates of the student's hidden knowledge state. Estimating Pr(K) enables us to predict performance (e.g. fluency) more accurately than estimating performance directly since the estimate of Pr(K) is conditioned on all observations so far. We present two approaches: 1) EEG-KT adds one binary EEG measure into KT, and 2) EEG-LRKT uses logistic regression on various continuous EEG measures in KT. Both approaches outperform the original KT, significantly for EEG-LRKT in terms of ROC and AUC, when predicting an unseen student's reading fluency on words in the Reading Tutor. For the first time, EEG measures are directly used to help model students' knowledge. Though not all the single-channel measures (like Theta) from EEG can help knowledge tracing, the combined EEG measure significantly improves KT predictions.

EEG studies in the neuroscience literature have better instrumentation but not longitudinal data like we have. EEG-based information (especially using a single sensor like NeuroSky's BrainBand) is noisy and is by no means a reliable, precise measure of a meaningful brain state. However, as demonstrated in this paper, longitudinal EEG does provide measurable improvement in predictive accuracy anyway.

In this paper, we focus on reading (specifically, fluency development), which is good for studying EEG-enriched KT thanks to density of sensing (many words per minute). The framework that we proposed is also applicable to other types of learning. Another future direction is to analyze the practical significance of the result in terms of impact on learning. As Beck and Gong [14] pointed out, tiny improvements in predictive accuracy don't matter - actionable intelligence does. We want to estimate the possible speedup in learning as a result of being able to use EEG to detect learning while it occurs (instead of waiting to observe future performance).

5 Acknowledgements

This work was supported by the National Science Foundation under Cyberlearning Grant IIS1124240. The opinions expressed are those of the authors and do not necessarily represent the views of the National Science Foundation.

References

1. Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Vladimir, T., Olmstead, R. E., Tremoulet, P. D., Patrice, D., Craven, P. L.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, Space, and Environmental Medicine. 78 (Supp 1), B231-244 (2007).
2. Miltner, W. H. R., Braun, C., Arnold, M., Witte, H., Taub, E: Coherence of gamma-band EEG activity as a basis for associative learning. Nature, 397, 434-436 (1999).
3. Chang, K. M., Nelson, J., Pant, U., Mostow, J.: Toward Exploiting EEG Input in a Reading Tutor. International Journal of Artificial Intelligence in Education, 22, 1-2, 19-38 (2013).
4. Heraz, A., Frasson, C.: Predicting the three major dimensions of the learner's emotions from brainwaves. World Academy of Science, Engineering and Technology, 31, 323-329 (2007).
5. Beck, J. E., Chang, K. M., Mostow, J., Corbett, A.: Does help help? Introducing the Bayesian evaluation and assessment methodology. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 383-394 (2008).
6. Pardos, Z. A., Heffernan, N. T.: Modeling individualization in a bayesian networks implementation of knowledge tracing. In User Modeling, Adaptation, and Personalization, pp. 255-266. Springer Berlin Heidelberg, (2010).
7. NeuroSky.: NeuroSky's eSense^{TM} meters and detection of mental sate. Neurosky, Inc. (2009).
8. Xu, Y., K.-m. Chang, Y. Yuan, and J. Mostow. Using EEG in Knowledge Tracing. In Proceedings of the 7th International Conference on Educational Data Mining. 2014: London, UK.
9. Xu, Y., Mostow, J.: Using logistic regression to trace multiple sub-skills in a dynamic bayes net. In Proceedings of the 4th International Conference on Educational Data Mining, 241-246 (2011).
10. Xu, Y., Mostow, J.: Comparison of methods to trace multiple subskills: Is LR-DBN best? In Proceedings of the 5th International Conference on Educational Data Mining, 41-48 (2012).
11. Yuan, Y., Chang, K. M., Xu, Y., Mostow, J.: A Toolkit and Dataset for EEG in Reading. In Proceedings of the 12th International Conference on Intelligent Tutoring Systems Workshop on Utilizing EEG Input in Intelligent Tutoring Systems (to appear).
12. Mostow, J., Beck, J. E.: When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In B. Schneider & S.-K. McDonald (Eds.), Scale-Up in Education (Vol. 2, pp. 183-200) (2007).
13. Chang, K. M., Beck, J. E., Mostow, J., Corbett, A.: A Bayes net toolkit for student modeling in intelligent tutoring systems. In Proceedings of the 8th International Conference on Intelligent Tutoring Systems, 104-113 (2006).
14. Beck, J. E., Gong, Y.: Wheel-spinning: Students who fail to master a skill. In Proceedings of the 16th Artificial Intelligence in Education, 431-440 (2013).