A Public Toolkit and ITS Dataset for EEG

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

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA. USA
{yuerany, kkchang, yanbox, mostow}@cs.cmu.edu

Abstract. We present a data set collected since 2012 containing children's EEG signals logged during their usage of Project LISTEN's Reading Tutor. We also present EEG-ML, an integrated machine learning toolkit to preprocess EEG data, extract and select features, train and cross-validate classifiers to predict behavioral labels, and analyze their statistical reliability. To illustrate, we describe and evaluate a classifier to estimate a student's amount of prior exposure to a given word. We make this dataset and toolkit publically available[1] to help researchers explore how EEG might improve intelligent tutoring systems.

Keywords: EEG; toolkit; reading comprehension; machine-learning; Project LISTEN's Reading Tutor

1 Introduction

With the rising importance of educational data mining and analytics, measuring and predicting student actions and mental states has become a key part of building better educational technologies. Electroencephalography (EEG) records a student's brain activity using electrodes on the scalp. Studies show EEG can be informative of many educationally relevant metrics including workload [1] and learning [2]. However, most of those studies were done in laboratories using laboratory-grade devices and fail to simulate the cost or environmental constraints of real classroom deployment of EEG devices. To explore the feasibility of practical classroom usage of EEG devices, we need to 1) use EEG devices simple enough for students to wear without assistance and cheap enough for schools to afford *en masse* and 2) record students' data in a realistic school setting. Collecting data in this way introduces two notable challenges: 1) the reduced dimensionality resulting from fewer sensors on a cheaper device and 2) the environmental noise inherent to an uncontrolled setting.

In this paper, we present a dataset from 3 years of school usage of Project LISTEN's Reading Tutor during which we recorded students' EEG signals. The signals were recorded with a consumer-grade single-channel device that now costs less than $100 each. We also describe EEG-ML, a machine learning toolkit to create and evaluate EEG-based classifiers of student actions and mental states. Many general-purpose EEG processing software and machine learning packages have been implemented and distributed; however, combining EEG processing with machine learning

[1] https://sites.google.com/site/its2014wseeg/eeg_ml

often involves complicated coding effort. EEG-ML simplifies the research process by providing a single pipeline for signal processing, classifier building, and cross-validated evaluation. We do not claim that the toolkit is an algorithmic innovation, rather a framework and baseline implementation to allow researchers to explore different algorithms and prediction tasks without needing to write a lot of code. We demonstrate a use case on the Reading Tutor dataset by applying this toolkit to estimate students' level of prior exposure to the word they're reading.

2 Project LISTEN's Reading Tutor EEG Dataset

Project LISTEN's Reading Tutor [3] (**Fig. 1)** is an intelligent tutoring system that displays text, listens to a student read it aloud, uses automated speech recognition to track the student's position in the text and detect miscues [4], and responds with spoken and graphical feedback. For 3 years, students 7-13 years old have worn EEG devices while they used the Reading Tutor. Our dataset consists of EEG signals and Reading Tutor logs collected during this period.

Fig. 1. On the left, NeuroSky's BrainBand. On the right, two students wear BrainBands that log their EEG data while they use the Reading Tutor.

2.1 Behavioral Data

We define a *trial* as a behavioral event with some outcome label, along with the corresponding EEG signal recorded during that time. Two types of such events are:

Sentence Encounter. During a session with the Reading Tutor, the tutor presents one sentence (or fragment) at a time, and asks the student to read it aloud. As the student reads the sentence, the words recognized by the tutor turn green.

Word Encounter. The student's speech is recognized as a time-aligned sequence of words by an Automated Speech Recognizer (ASR). The ASR estimates when each text word was read and whether the reading was correct. The tutor computes latency as the duration between successive text words. The ASR is imperfect – it detects only about 25% of misread words and falsely rejects about 2% of correctly read words.

2.2 EEG Data

We used NeuroSky's BrainBands to collect EEG data (see **Fig. 1**). The BrainBand is a wireless device with one electrode on the forehead (frontal lobe) roughly between Fp1 and Fp2 in the 10-20 system. BrainBands output raw EEG signals at a rate of 512 Hz and NeuroSky's proprietary eSense measures at 1 Hz. The BrainBand is a product for consumers so it is designed with ease of use in mind. Unlike the multi-channel electrode nets worn in labs, the BrainBand requires no gel or saline for recording, making it easier to wear and maintain. Students are able to put on the headset with minimal supervision. Students included in the dataset encountered ~160,000 sentences containing ~800,000 words and we recorded about 108 hours of EEG data, though not all sentence and word encounters have corresponding EEG data, and vice versa. Our example word-exposure classifier used data from the 17 students with the most data, who had an average of 3,600 words with aligned EEG.

3 The EEG-ML Toolkit

Machine Learning for EEG (EEG-ML) is a toolkit for studying EEG in the context of intelligent tutoring systems. The pipeline attempts to cover the complete process of signal processing, machine learning, and evaluation/analysis. Much of this pipeline has been described previously [5]. See **Fig. 2** for the pipeline's overall structure. We will describe important components below (see project website[1] for a full description).

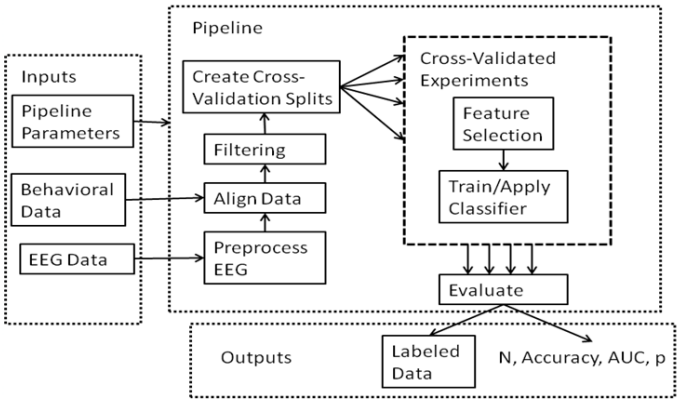

Fig. 2. Overview of pipeline of EEG machine-learning toolkit

As a demonstration of the toolkit and dataset, we use them to create a predictor for students' number of prior encounters of particular words. Notably, this measure (word exposure) is fairly well insulated from ASR error and we urge caution when studying measures that could be heavily impacted by ASR error, such as latency or correctness. We will use this classifier as a running example as we describe our pipeline.

Inputs. The pipeline's inputs are 1) a spreadsheet of behavioral data containing the label that we want to predict, 2) a spreadsheet of EEG data, and 3) a set of parameters specifying the algorithms and arguments to be used in the pipeline. For our word-exposure classifier, we labeled the first 11 encounters of a word (by each subject) as 'early encounter' and the remaining encounters as 'late encounter'. Example: if the subject saw 'cat' 30 times, the first 11 times are early encounters and the final 19 times are late encounters. We chose 11 as a threshold so that we could have roughly the same number of early and late encounters. To avoid skewing our models with subjects who have little data, we removed subject with less than 8,000 seconds of EEG recordings and analyzed 17 students who read about ~62,000 words in total.

Pipeline. Given the behavioral and EEG data, the pipeline 1) aligns corresponding EEG signals to each trial in the behavioral data, 2) filters and derives features for each trial from the EEG signals aligned to that trial, 3) splits the data into training and testing sets following a cross-validation scheme. Within each cross-validation fold, we use feature selection to reduce the dimensionality, and train a classifier on the training set and apply it to the testing set. Finally, the pipeline aggregates classification results and evaluates the classifier's performance. This entire process happens offline.

EEG Preprocessing and Filtering. Many noise sources (including eye blinks, facial expressions) can introduce artifacts into the recorded signals. To remove potential artifacts the pipeline uses soft thresholding with wavelets to denoise the signals [6]. The pipeline also allows experimenters to remove trials whose EEG signal had a certain proportion of low-quality signals. We use Neurosky's PoorSignal score as a measure of signal quality. In building our word-exposure classifier, we are aggressive in filtering; we filter out all trials where more than 50% of corresponding signals have poor reported signal quality (score of 100 or higher on NeuroSky's 0 to 200 poor signal scale).

Feature Generation. The unit of analysis is an individual trial. The pipeline breaks the trial into several *epochs* – an EEG segment of a fixed length. For example a 3-second-long trial could be broken into 3 epochs of 1 second each with no overlap, or it could be broken into 5 epochs of 1 second each with 0.5 seconds of overlap between epochs.

The pipeline uses Fast Fourier Transform to extract oscillation features from each epoch – delta (1-3Hz), theta (4-7 Hz), alpha (8-11 Hz), beta (12-29 Hz), and gamma (30-100 Hz) frequency bands. Using these per-epoch features, the pipeline derives a set of higher-level features (e.g. mean, variance) for each trial. Our word-exposure classifier used 5 features - the means of each of the alpha, beta, gamma, theta, delta features of the epochs.

Cross-Validation. The pipeline supports leave-one-out cross-validation with a *within-subject* or *between-subject* scheme. In the within-subject scheme, the training set and test set are taken from the same subject, creating a subject-specific classifier using all but 1 trial from the subject as the training set, and testing on the left-out trial. In the between-subject scheme, we train on all trials from all but one subject, and test on the trials of the left-out subject. The between-subject scheme allows us to simulate how the algorithm will perform on unseen subjects. We used within-subject cross-validation to evaluate our word-exposure classifier.

Feature Selection. In cases where we have little data but many features, we often want to use feature selection to reduce the dimensionality of our data before feeding it to a classifier, in order to learn a classifier less sensitive to noise. The pipeline supports two feature selection methods – Principal Component Analysis and T-Test based Rank Feature Selection. Because of the high level of noise in our data, our word-exposure classifier used 3 dimensional PCA to avoid over-fitting.

Train/Apply Classifier. Our pipeline supports two types of classifiers – Linear SVM and Gaussian Naïve Bayes. The Linear SVM classifier is more commonly used in brain signal processing. A Gaussian Naïve Bayes classifier allows us to train non-linear classifiers and train them more quickly than SVM. Our word-exposure classifier used a Gaussian Naïve Bayes classifier.

Evaluation. The pipeline computes 1) classification accuracy (ACC), 2) a chi-squared test comparing accuracy to chance (one over the number of categories), and 3) the receiver operating characteristic (ROC) curve and area under the curve (AUC). ACC is intuitive and widely used, but it can have issues with class size imbalance – a majority class classifier could obtain above-chance results. AUC calculates the area under the ROC curve, which is insensitive to data set imbalance. A majority class model would show a diagonal line from the bottom left to the top right corner in ROC space, and get an AUC score of 0.5.

Outputs. The outputs of the pipeline are 1) a table showing accuracy, AUC, N (number of data points), and p-value and 2) a spreadsheet where each row of the original behavioral data is annotated with the prediction made by the classifiers in the experiment so that further analysis may be done in other programs if desired. Our word-exposure classifier had an average accuracy of 57%, which was significantly above chance ($p < 0.05$), with AUC of 0.60. A measure whose accuracy is only in the high 50's is not practical on its own, but can potentially be used as a feature in combination with other features to improve student modeling, as shown by Xu et al. [7]

Though significantly above chance, our accuracy is relatively low compared to that claimed in other EEG studies of learning [8]. A subtle difference in independence assumptions might be one reason [5]. Also, we expect the lower-end device and noisy *in vivo* setting to reduce accuracy. However, further analysis of features (e.g. which bands are most useful) and algorithms (e.g. different classifiers and kernels) could produce incremental improvements to results. Indeed, a key motive for releasing this toolkit and dataset is to provide the research community with a baseline to build upon and a common dataset to evaluate different algorithms.

4 Conclusion

We present a multi-year dataset of EEG data from *in vivo* usage of an intelligent tutoring system. We also present EEG-ML, a machine learning toolkit to produce and evaluate EEG-based classifiers. We hope the dataset and toolkit will allow researchers to focus on experimentation and analysis rather than data collection and technical implementation, facilitating their research into new applications of brain signal processing in building better intelligent tutoring systems.

5 Acknowledgement

This work was supported by the National Science Foundation under Cyber-learning Grant IIS1124240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Berka, C., D.J. Levendowski, M.N. Lumicao, A. Yau, G. Davis, V.T. Zivkovic, T. Vladimir, R.E. Olmstead, P.D. Tremoulet, D. Patrice, and P.L. Craven. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 2007. *78 (Supp 1)*: p. B231-244.

[2] Miltner, W.H.R., C. Braun, M. Arnold, H. Witte, and E. Taub. Coherence of gamma-band EEG activity as a basis for associative learning. *Nature*, 1999. *397*: p. 434-436.

[3] Mostow, J. and J.E. Beck. When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In B. Schneider and S.-K. McDonald, Editors, *Scale-Up in Education*, 183-200. Rowman & Littlefield Publishers: Lanham, MD, 2007.

[4] Mostow, J. Why and How Our Automated Reading Tutor Listens. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, 43-52. 2012. Stockholm, Sweden. KTH, Computer Science and Communication, Department of Speech, Music and Hearing, SE-100 44 Stockholm, Sweden.

[5] Chang, K.M., J. Nelson, U. Pant, and J. Mostow. Toward Exploiting EEG Input in a Reading Tutor. *International Journal of Artificial Intelligence in Education*, 2013. *22*(1-2): p. 19-38.

[6] Donoho, D.L. De-noising by soft-thresholding. *IEEE Transactions on Infomation Theory*, 1995. *41*(3): p. 613-627.

[7] Xu, Y., K.M. Chang, Y. Yuan, and J. Mostow. EEG Helps Knowledge Tracing! In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems Workshop on Utilizing EEG Input in Intelligent Tutoring Systems*. 2014: Honolulu.

[8] Heraz, A. and C. Frasson. Predicting the three major dimensions of the learner's emotions from brainwaves. *World Academy of Science, Engineering and Technology*, 2007. *31*: p. 323-329.