Shih, J. L. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Can EEG signal predict learners' perceived difficulty?

Aman KUMAR[a], Pankaj CHAVAN[b] & Ritayan MITRA[b*]
[a]*Department of Computer science and Engineering, Thapar Institute of Engineering and Technology, India*
[b]*IDP in Educational Technology, Indian Institute of Technology Bombay, India*
*rmitra@iitb.ac.in

Abstract: This study presents an approach to predict learner's perceived difficulty using features extracted from electroencephalography (EEG) data. We demonstrate how EEG signals can be used effectively to estimate learner's perceived difficulty of learning content. Student self-reports of perceived difficulty and EEG data were gathered from 9 participants who watched a video lecture. A machine learning model with random forest classifier achieved a maximum accuracy of 75.24% in estimating perceived difficulty. Furthermore, the model predicted the difficulty level of the entire video lecture for individuals fairly well. Our results have implications for intelligent tutoring systems which aim at providing the learner with an adaptive and personalized learning environment.

Keywords: Electroencephalography (EEG), DEBE feedback, perceived difficulty, intelligent tutoring systems (ITS), machine learning

1. Introduction

Assessing and monitoring the mental states of learners is important in the teaching-learning process. Detecting the mental states of the learner by collecting appropriate data and providing a personalized experience for every learner has always been the holy grail of adaptive learning (Shawky & Badawi, 2018). Intelligent tutoring systems (ITS) are known for their focus on providing the learner with an adaptive and personalized learning environment. The learner model contains knowledge about the learner's cognitive-affective states and guides individualized learning trajectories (Nkambou, Mizoguchi, & Bourdeau, 2010).

There are several approaches to create learner models. A learner's behavioural data such as keystrokes (Bixler & D'Mello, 2013), clickstream data (Baker et al., 2008; D'Mello, 2017) etc. can be used to infer learning processes. For example, keystroke data was used to estimate engagement of learners engaged with writing task (Bixler & D'Mello, 2013), and clickstream data to model whether the learner is engaged in gaming the system (Baker et al., 2008) etc. Another approach is to use a variety of psychophysiological measures to measure different mental states. Eye tracking has been demonstrated to be a good indicator of cognitive workload (Schultheis & Jameson, 2004), attention (Conati, Merten, Muldner, & Ternes, 2005), engagement (D'Mello, Olney, Williams, & Hays, 2012) and also boredom (Jaques, Conati, Harley, & Azevedo, 2014). Facial emotion recognition has been used to measure frustration (Sidney et al., 2005) and perceived difficulty of learners (Whitehill, Bartlett, & Movellan, 2008). A multimodal approach has also been found to be effective in such applications for monitoring cognitive and affective states within an ITS (Lane & D'Mello, 2019). Pressure-sensitive mouse and chair, and galvanic skin response are some other less commonly used psychophysiological sensors that have been used (Arroyo et al., 2009; Kapoor, Burleson, & Picard, 2007). Multimodal approaches have usually yielded good accuracy in predicting mental states (Arroyo et al., 2009; Kapoor et al., 2007). A relatively less common data in this context is that from electroencephalogram or EEG, which measures brain waves. It has also shown promise in capturing changes in the attention and cognitive workload of learners in ITS (Chaouachi, Jraidi, & Frasson, 2015).

As evident in the aforementioned discussion, some commonly investigated parameters in ITS research are cognitive workload, attention, engagement, frustration, boredom etc. The knowledge of

learners' perceived difficulty is also an important (Lane & D'Mello, 2019), albeit relatively less studied variable. A learner may *perceive* the content difficult to understand (regardless of the true difficulty of the topic), and his perception may differ from other students. Even then, such perception have been shown to decrease interest and increase boredom directing the focus more towards the feeling of the negative affect which can further lower the attention and motivation, eventually resulting in poor performance (Pekrun, Goetz, Titz, & Perry, 2002). There have been several attempts to measure perceived difficulty in learners (Pham & Wang, 2018; Whitehill et al., 2008) and adjust instructional material accordingly.

The DEBE framework has been recently proposed as a systemized way to collect continuous, fine-grained feedback from students on their levels of perceived difficulty (and affective states) during a lecture (Mitra & Chavan, 2019). We use student-self reports in the form of DEBE feedback to train an EEG machine learning algorithm that can be used to predict levels of perceived difficulty when interacting with a video lecture. In this study, we investigate the potential of a low-cost consumer-grade EEG device, MUSE™, in predicting perceived difficulty in learners.

2. Method

2.1 Experimental set-up

The setup of this experiment was designed keeping in mind that it should be relevant to, and closely resemble real-world conditions as opposed to providing elementary cognitive tasks that are often used in psychology studies (e.g., simple arithmetic to measure cognitive load). For this purpose, we obtained a video recording of 47 minutes duration for a lecture from the Advanced Heat Transfer course offered at the university. Ten students, aged 22-29, taking this course participated in the study. However, we have used the EEG data of 9 participants for analysis due to the presence of high noise and missing values (>25%) in one of the samples. The students were asked to come to the laboratory individually and watch the lecture video. The Institute Ethics Committee approved the study (IEC Approval Letter_Proposal No. IITB-IEC/2018/004). The students watched the video lecture and provided DEBE feedback on four parameters, namely, whether the lecture was easy, difficult, engaging or boring? The students could press the buttons as often they felt any of the four states in real-time while watching the video lecture. Details of the experimental setup and related work have been provided in (Chavan, Gupta, & Mitra, 2018). This feedback was unprompted and the students could click whenever they felt any of the four cognitive/affective states. We used a MUSE™ EEG device having 4 electrodes at AF7, AF8, TP9 and TP10 according to the 10-20 international system, sampling at 256 Hz frequency to record the EEG signals. For the purpose of this study, we chose to focus only on the instances when the students clicked either difficult or easy (engaging and boring clicks were not analyzed) with the understanding that this represents their perceived difficulty of the topic.

2.2 Data pre-processing

EEG data is noisy, complex and suffers from the curse of high dimensionality. Hence, this step aims at cleaning the EEG signal, managing its complexity by filtering out less relevant information from the signal and reducing dimensionality by selecting a subset of relevant channels. Therefore, we decided to look at only the frontal lobes that have been shown to be active during higher order processing (Zarjam, Epps, Chen, & Lovell, 2013). The data from the two temporal lobes were discarded. Besides the lower relevance of those areas for the task, they are also susceptible to more artefacts such as jaw clenches (Kappel, Looney, Mandic, & Kidmose, 2017). A 60 Hz notch filter was applied uniformly to remove the power line artefact. The cleaned data were then chunked into one-minute windows preceding each difficulty or easy click. Each such segment was divided into 5-sec epochs and all 12 epochs thus formed were categorized as either easy or difficult.

EEG data analysis is usually confined to either the frequency or the time domain. However, this dualistic approach of data analysis completely discards the information contained in one or the other domain. An alternative approach would be to retain information from both domains for the analysis. Discrete wavelet transform (DWT) is one such analysis that uses information from both domains. The

DWT method is widely used for the time-frequency analysis in EEG signals, also due to non-stationary characteristics of EEG signals and DWT does not assume signals to be stationary. Wavelet Packet Decomposition (WPD) is a wavelet transform method where the signal is passed through more filters than the discrete wavelet transform (DWT) resulting in much more extensive decomposition and offers richer signal analysis. Hence, we have used WPD for our analysis. The WPD method decomposes the signal using successive high pass and low pass filters, and this time-scale representation is generated by dilation and translation of a mother wavelet. As the choice of mother wavelet depends on the application, researchers have mentioned that Daubechies order-4 (db4) is the most suitable mother wavelet for EEG signal analysis (Adeli, Zhou, & Dadmehr, 2003). Hence, we applied level-5 Wavelet Packet Decomposition using Debauchies-4 mother wavelet to decompose the signal and then extracted approximate delta (0-4Hz), theta (4-8Hz), alpha (8-16Hz), beta (16-32Hz) and gamma (32-64Hz) bands.

2.3 Machine learning

Feature extraction aims at describing the EEG signal by a few relevant values called "features". Features were extracted from the EEG signals to construct features vectors (samples) for input into the machine learning models. Such features should capture the information embedded in the EEG signals that is relevant. In passive brain-computer interface (BCI) applications, entropy and energy are one of the most effective and commonly used features. Furthermore, researchers have found these to be statistically significant in determining cognitive states (Zarjam et al., 2013). Therefore, energy and Shannon entropy were extracted from each of the five frequency bands (alpha, beta, etc.) from both the frontal channels giving rise to a feature vector with 20 values (2*5*2=20). After extracting the feature vectors from the epochs, we created datasets corresponding to difficult and easy feedback epochs for each individual. The individual datasets were normalized (z-normalization) to account for individual differences and then merged to a single dataset of feature vectors. Initially, we had 778 instances of the difficult class and 384 instances of the easy class but, to avoid overrepresentation of the difficult class, we selected random 384 samples out of 778 samples for the difficult class. Therefore, our dataset comprised 384 instances each for difficult and easy class each.

Finally, we compared performances of various ML classifiers, namely, support vector machine (SVM), decision tree, artificial neural network (ANN) and random forest, in order to optimize and achieve the best results. We used hyperparameter tuning to achieve optimal accuracy. To test and validate the models thoroughly, we used k-fold cross-validation (k = 10). In k-fold cross-validation, the dataset is partitioned into k equal-sized non-repeating complementary subsets. Then, out of k subsets, k-1 subsets are used as training set, and 1 subset is retained as the validation set. This process is repeated k times, with each of the k subsets used exactly once as the validation set. All reported values are averages of the 10 iterations of this cross-validation.

3. Results and discussion

3.1 Performance of classifiers

We implemented SVM using three different kernels: linear, polynomial (degree = 3) and gaussian radial basis function (rbf). Further, for each kernel, we tuned the penalty parameter (C) of the error term. The best accuracy was 62.23% with the rbf kernel and C = 10. Several ANN architectures were implemented with Rectified Linear Unit (ReLU) and "logistic" as the activation function for the hidden layers and output layer, respectively. A maximum accuracy of 64.37% was achieved with two hidden layers, each having 64 neurons. The decision tree classifier with two splitting criterions, gini impurity (Gini) and information gain (entropy), achieved a maximum accuracy of 66.11% with splitting criteria = information gain (entropy). Finally, with a random forest classifier, we achieved a maximum accuracy of 75.24% with splitting criteria = Information gain (entropy) and the number of decision tree classifiers = 50.

3.2 Sensitivity analysis for random forest classification

We performed a sensitivity analysis by varying the length of the time window for data extraction preceding a feedback click and the length of epoch used to divide it further (Table 1). In this paper, we have achieved results with time window length = 60 seconds and epoch length = 5 seconds. Decreasing the epoch size increases the accuracy till 1 sec, after which the accuracy reduces, indicating a tradeoff between the size of dataset and the noise introduced by reducing the epoch size. For epoch sizes greater than or equal to 2 sec window size of 1 mimute seem to be an ideal choice as accuracy falls on both sides. However, for smaller epoch sizes, increasing window size only decreases accuracy, possibly indicating the fact that the algorithm is sensitive to the noise introduced as a result of reducing epoch size with increasing window size.

Table 1

Best classification accuracies (in percent) with Random Forest classifier with varying length of the time window for data extraction and epoch length.

	Time window (sec)			
Epoch (sec)		**30**	**60**	**120**
	0.5	75.26	73.86	71.95
	1	76.41	75.68	72.19
	2	74.79	76.10	73.96
	5	72.72	**75.24**	72.10
	10	67.78	72.38	69.40

3.3 Predicting perceived difficulty from EEG data

To further test the potential of the ML model, we used the trained model to estimate the difficulty level for the entire lecture for each individual participant. The EEG data for the entire duration of the lecture was divided into contiguous epochs of 5 seconds. Each epoch was processed and features were extracted to prepare the feature vector. Subsequently, each epoch was classified using the trained random forest classifier, which had performed the best in our analysis. At this point, it is important to note that although accuracy gives an estimate of the performance of the ML algorithm it essentially uses 5 sec epochs of one minute windows prior to all clicks to make a prediction of the category of an epoch. It does not give us an understanding of what the classifier would predict for those epochs that are further away from the clicks - either after the click or more than one minute away from one. This section attempts to address this issue so that we can form another representation of the utility of such predictions.

The model predictions tallied fairly well with the observations. We see most clicks happen at predictable times (Fig 1 a, b and c). However, the number of false positives appear to be the problem when the predicted value appear to be either difficult or easy, but no clicks were observed. This could be due to a variety of reasons. The null hypothesis would be that the algorithm is indeed not suitable for more accurate predictions. One alternative hypothesis is that the student might be experiencing difficulty without recognizing it and hence does not click. The student could also be clicking only after feeling a particular state for a reasonable yet variable duration. The difference with these two possibilities being only in whether the student is aware of the cognitive-affective state s/he is in as the action (of clicking) would be identical.

Despite the possible shortcomings of the temporal accuracy of the predictions, it is encouraging to find that for participants with many difficult (easy) clicks the algorithm is indeed showing a prolonged state of difficulty or ease (Fig. 1b and c). It is to be noted that the training set included only 384 epochs each of easy and difficult segments (with possible overlaps between epochs). Therefore, the total training interval was equal to or less than 64 minutes (384*2*5 secs). The total predicted interval was 47*9 - 64 = 359 minutes (47 minute lecture duration for 9 students). Also, the ML algorithm was trained on all positives (interval preceding either difficult or easy clicks) but the additional data for prediction (359-64 = 295 minutes) were all negatives (participants did not click either difficult or easy). This makes Figs. 1b and c even more interesting because it seems a participant clicking exclusively

difficult/easy tend to have long intervals of predicted difficulty/ease which could imply that, at the very least, the algorithm successfully predicts high average difficulty (ease) values for candidates who tend to click difficult (easy) several times during the lecture.

(a) (b)

(c)

Figure 1. Model predicted difficulty levels for three participants. Y-axis values (0=easy and 1=difficult) are averages of 40 sec or 8 epochs. For example, if 4 out of 8 epochs were predicted to be difficult, then the corresponding predicted difficulty level would be 0.5.

4. Conclusion and limitations

We have demonstrated the viability of a low-cost EEG device in predicting perceived difficulty of learners. A random forest ML algorithm was able to predict perceived difficulty in participants with an accuracy of 75.24%. Furthermore, when presented with new data from the rest of the lecture the model predictions tallied closely with the clicking behavior – students who clicked difficult (easy) more often were predicted to have high (low) perceived difficulty.

In spite of the promising results, our research has some limitations. The machine learning model was trained with data from only 9 participants. We need a considerably larger dataset to train a robust and reliable classifier that can accurately classify EEG signals into the corresponding cognitive state. It is also likely that the accuracy would increase further with increasing size of the dataset. Another limitation of this study is that there was no way to validate the mental state of the participants when there was no click. Although we did find a consistently high or low value of perceived difficulty (Figs. 1b and c, respectively) for appropriate participants, we still have no validation for the interim gaps when there were no clicks. Another limitation of this study is the possibility that the EEG data were somehow affected by the clicking itself and could have introduced artefacts that were unrelated to the construct under investigation, namely, perceived difficulty. We are so far unable to rule out such a possibility.

Acknowledgements

This work was supported by an IRCC, IIT Bombay grant (17IRCCSG013) to Ritayan Mitra.

References

Adeli, H., Zhou, Z., & Dadmehr, N. (2003). Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, *123*(1), 69–87.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. *AIED*, *200*, 17–24.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Bixler, R., & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 225–234. ACM.

Chaouachi, M., Jraidi, I., & Frasson, C. (2015). Adapting to learners' mental states using a physiological computing approach. *The Twenty-Eighth International Flairs Conference*.

Chavan, P., Gupta, S., & Mitra, R. (2018). A novel feedback system for pedagogy refinement in large lecture classrooms. *International Conference on Computers in Education.*, 464–469. Philippines.

Conati, C., Merten, C., Muldner, K., & Ternes, D. (2005). Exploring eye tracking to increase bandwidth in user modeling. *International Conference on User Modeling*, 357–366. Springer.

D'Mello, S. (2017). Emotional learning analytics. *Handbook of Learning Analytics*, 115.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, *70*(5), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. *International Conference on Intelligent Tutoring Systems*, 29–38. Springer.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, *65*(8), 724–736.

Kappel, S. L., Looney, D., Mandic, D. P., & Kidmose, P. (2017). Physiological artifacts in scalp EEG and ear-EEG. *Biomedical Engineering Online*, *16*(1), 103.

Lane, H. C., & D'Mello, S. K. (2019). Uses of Physiological Monitoring in Intelligent Learning Environments: A Review of Research, Evidence, and Technologies. In T. D. Parsons, L. Lin, & D. Cockerham (Eds.), *Mind, Brain and Technology: Learning in the Age of Emerging Technologies* (pp. 67–86). https://doi.org/10.1007/978-3-030-02631-8_5

Mitra, R., & Chavan, P. (2019). DEBE feedback for large lecture classroom analytics. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 426–430. ACM.

Nkambou, R., Mizoguchi, R., & Bourdeau, J. (2010). *Advances in intelligent tutoring systems* (Vol. 308). Springer Science & Business Media.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105.

Pham, P., & Wang, J. (2018). Adaptive Review for Mobile MOOC Learning via Multimodal Physiological Signal Sensing-A Longitudinal Study. *Proceedings of the 2018 on International Conference on Multimodal Interaction*, 63–72. ACM.

Schultheis, H., & Jameson, A. (2004). Assessing Cognitive Load in Adaptive Hypermedia Systems: Physiological and Behavioral Methods. In P. M. E. De Bra & W. Nejdl (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems* (Vol. 3137, pp. 225–234). https://doi.org/10.1007/978-3-540-27780-4_26

Shawky, D., & Badawi, A. (2018). A reinforcement learning-based adaptive learning system. *International Conference on Advanced Machine Learning Technologies and Applications*, 221–231. Springer.

Sidney, K. D., Craig, S. D., Gholson, B., Franklin, S., Picard, R., & Graesser, A. C. (2005). Integrating affect sensors in an intelligent tutoring system. *Affective Interactions: The Computer in the Affective Loop Workshop At*, 7–13.

Whitehill, J., Bartlett, M., & Movellan, J. (2008). Automatic facial expression recognition for intelligent tutoring systems. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–6. IEEE.

Zarjam, P., Epps, J., Chen, F., & Lovell, N. H. (2013). Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in Biology and Medicine*, *43*(12), 2186–2195.