Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Selective Prediction of Student Emotions based on Unusually Strong EEG Signals

Judith AZCARRAGA[a*], Nelson MARCOS[a], Arnulfo AZCARRAGA[a], and Yoichi HAYASHI[b]
[a]*College of Computer Studies, De La Salle University, Manila, Philippines*
[b] *Department of Computer Science, Meiji University, Kawasaki, Tokyo, Japan*
*judith.azcarraga@dlsu.edu.ph

Abstract: With an electroencephalogram (EEG) sensor mounted on their head while learning mathematics using two computer-based learning software, EEG signals were collected from fifty six (56) academically-gifted students of ages 11 to 14. The EEG signals are used to predict four academic emotions, namely frustrated, confused, bored, and interested. It is shown that emotion classification accuracy is improved by selective prediction - performed only when a pre-determined proportion of EEG feature values deviate significantly from the baseline mean. The experiments on instances, where 0%, 2%, 4%, and up to 20% of the features are significantly stronger EEG signals, show that the accuracy rate of decision trees increases from 0.50, 0.59, and 0.45 (for instances with 0% special event features) to 0.74, 0.75, and 0.66 (for instances with 20% special event features) for predicting frustrated, confused and bored, respectively. Accuracy for predicting interested does not increase like for the other three emotions.

Keywords: EEG, academic emotions, prediction, decision trees

1. Introduction

Students experience various emotions while engaged in learning activities. Learning takes place through a complex interplay of cognitive and emotional states. Emotions experienced during learning, also referred to as academic emotions (Pekrun, et. al., 2002), may affect the flow of learning as well as the motivation to continue with a learning task. As such, maintaining positive academic emotions among learners, and quickly addressing negative emotions, have been the challenge for human tutors, as well as for designers and developers of learning systems.

Indeed, recent developments in the design of tutoring systems, or even learning systems in general, have attempted to render these systems more adaptive not only to the learners' cognitive state but also to their affective state. In the case of these affective tutoring systems, the affective states such as 'confident', 'frustrated', 'excited', 'interested', 'confused', 'engaged' and 'bored' may be recognized using the tutor-context and the user-profile (Baker, et. al, 2008), and sometimes in combination with signals from hardware sensors such as a camera, microphone and various other physiological sensors that capture EEG signals, electromyogram (EMG) signals, skin conductance levels, heart rate, and respiration rate (Arroyo, et. al.,2009).

Physiological signals such as brainwaves and skin conductance may provide rich information about the current cognitive (Jausovec, 2000; Chanel, 2009; Stevens, Galloway and Berka, 2007) and emotional state (Arroyo, et. al., 2009; Chanel, 2009). Unlike a facial expression, such signals are difficult to mask and can capture natural emotional expression through physiological manifestations.

Signals from brain activities have indeed been reported to provide some useful information as to emotion valence (positive or negative) and arousal (Chanel, 2009; Heraz, Razaki and Frasson, 2007) and level of frustration and distraction (Stevens, Galloway and Berka, 2007).

Even when confining experiments to those systems that only record brainwaves, and not the other physiological signals, the aptitude as well as the personality of the learner would add to the wide variety of possible factors that may influence the success or failure of learning systems. Indeed, the brainwave pattern of academic achievers, who may be loosely referred to as gifted learners, have been found to be different from average learners (Jausovec, 2000). This may not be surprising since a number of studies, particularly in psychology and education, have reported that gifted learners learn differently

as compared to average ones. Gifted learners are characterized by high or advanced intellectual ability and/or outstanding talents. They have the potential to learn fast, to deal well with complex and abstract ideas and to have a large knowledge base (Diezmann, 2005; Greene, Moos, Azevedo and Winters, 2008). As such, they are pre-disposed to be able to learn by themselves.

This study focuses on academic achievers, in order that various other academic-related variables such as general ability in mathematics and motivation, as well as the capacity for self-learning may be controlled. The brainwave patterns and their association with the affective learning behaviors of fifty six (56) academically-gifted students of ages 11 to 14 are the subject of this study. Using an EEG sensor mounted on their head, EEG signals were collected while the students were engaged in some computer-based learning systems for mathematics.

The EEG signals, once pre-processed, are analyzed and used to predict four academic emotions, namely frustrated, confused, bored, and interested. It is shown that the accuracy of EEG-based classification of academic emotions can be improved by selective prediction, i.e. prediction is only performed when a certain pre-determined percentage of EEG feature values deviate significantly from the baseline mean. The hypothesis is that when recordings of an EEG sensor are picking up something 'unusual', this signifies that the learner is experiencing some heightened emotion that may then be easier to predict. As such, when prediction is limited to only those instances where a given proportion of feature values deviate significantly from their baseline mean, the emotion prediction accuracy would tend to increase.

2. Improving Classification using Selective Prediction

The approach to improving prediction accuracy based on selective predictions was previously explored using a different dataset, and was first reported in (Azcarraga, et. al., 2011). This was subsequently studied further and described in (Azcarraga and Suarez, 2013). In this study, we further refined the previous work following the same the same line of thought, taking into more stringent account of what constitutes 'number of feature values deviate significantly from their baseline mean'. Also, pre-processing and data normalization has by now been more thoroughly and methodically executed. In addition, training and testing where performed under a more methodical, statistically-controlled environment.

The previous works described in (Azcarraga, et. al., 2011) and (Azcarraga and Suarez, 2013) were also based on an entirely different set of experimental conditions. The past works employed raw EEG values from 14 sensors which were not filtered and were not transformed to a frequency format, as they were done more systematically here. The human subjects were also a different set, and of a different age range. Finally, the academic emotions that were tested were different and the learning modules used for the experimental sessions were also quite different.

In the current work, the feature values are now standardized, ranging in values from -3 to +3. The data have been fully normalized and so therefore, the feature values are now comparable from one learner (subject) to another. Also, the validation tests employed in the previous works used a 10-fold cross validation, i.e. the original dataset was randomized and divided into 10 folds prior to validation, and each segment composed of 10% of the dataset was used for testing. Although this is a standard approach to machine learning and computational intelligence experiments, this is a weaker approach to testing and so in the experiments described here, we followed a stricter train-and-test methodology. All the instances of each student are isolated into a rolling test set, one train-test episode for each subject, and the results on all the test sets are combined to determine the prediction accuracy.

3. Data Collection and Preprocessing

The participants for this study are first year (Grade 7) and second year (Grade 8) high school students with ages 11 to 14 who are all academic achievers. During the data collection sessions, these students were asked to use two computer-based learning systems for high school mathematics, Aplusix and Scooter. Aplusix is a learning system for algebra developed in Grenoble, France (Nicaud et. al., 2002), while Scooter is a scatterplot learning software (Baker, et. al, 2008), developed as part of the Cognitive

Tutor tutoring system. A software module for collecting simultaneous data signals from the EEG sensor and the self-reported emotions has been pretested and is the one deployed in the actual experiments.

The EEG sensor used in this study is the Emotiv EPOC sensor (http://www.emotiv.com). A commercial product typically used for gaming purposes, the Emotiv sensor is equipped with 14 channels based on the International standard 10-20 locations.

All 56 students were subjected to a calibrated, video-taped learning session using the two learning systems, equipped with emotion-sensing devices that are all connected to a computer to record all the simultaneous data signals. Sensing devices which were used in order to capture various student affect include an EEG sensor and a video-camera. Each learning system was used for about 10 mins.

While using a learning system, each participant was asked to report the level of their confusion, boredom, frustration and interest by clicking on a sliding bar that range in value from 0 to 100. The participants were instructed to self-report every 2 minutes, after solving each problem or whenever they sense that there is a change of emotion or task difficulty.

Standard pre-processing and data preparation techniques are used, including computing deviations from baseline signals instead of using raw signals, data normalization, and balancing of the datasets. The EEG signals are also carefully calibrated and synchronized with the other collected information, for purposes of tagging the signals with self-reported emotions.

According to psycho-physiological literature (Ekman, 1984), emotions persist for about 0.5 to 4 seconds. Guided by this, the sampling rate (i.e. window size for the computation of average EEG signals per segment) is set at 2.0 seconds. All the pre-processed EEG data and self-reported emotion tags were carefully synchronized, merged and uniformly segmented into 2-second windows with 1-second overlap. The EEG signals are sampled by the Emotive device to 128 samples per second.

Since EEG data usually contain non-desirable artifacts/noise that seriously degrade the quality of the data signals, high-pass, low-pass filtering and moving average techniques were employed to clean up the data. Otherwise, if used directly for classification of the academic emotion, these signals would be incorrectly classified.

4. Data Preparation for Classification

To extract the specific EEG features, raw data from the 14 EEG channels were segmented (into 2-second window samples, with 1-second overlap). Each sample or segment was treated as a single instance in each subject's dataset. Each sample was filtered and transformed into alpha (8-12 Hz), low beta (12-21 Hz), high beta (21-30 Hz) and gamma (31-50 Hz) frequency bands. Peak magnitude and the mean spectral power for each band coming from each the 14 channels positioned at specific locations on a student's head are taken as EEG features. The peak magnitude is the highest magnitude or amplitude of the sample. The mean spectral power is the average power spectrum of the signals in the sample. This feature was used in the brainwaves research of Jausovec (2000). Aside from these, other features such as the average energy over brain areas (frontal, parietal, temporal, occipital), brain asymmetry scores or lateralization of the 7 right-left electrode pairs from the alpha band and the energy of beta for all the electrodes, were extracted (Chanel, 2009). All told, a total of 126 features were extracted from each segmented raw sample.

Each EEG sample, with a total of 126 feature values, is considered an instance in each subject's dataset and each instance is labeled according to the self-reported emotions. Emotion rates for frustrated, confused, bored and interested were discretized to either 'low' or 'high' which are then used to label each instance. The label is 'low' if emotion rate was below 50, otherwise, the label is set to 'high'.

Two distinct sets of EEG data for each student participant were processed. The first set is the EEG data taken during the 'resting-state' while the other is gathered during learning session. During the 'resting-state' period, the participant is asked to stare at a black screen with minimal movement that lasts for about 3 minutes. The brainwave signals captured during this period, i.e. the values of each FFT-transformed feature for each student, were averaged. The average value serves as the baseline EEG of that particular student (Davidson, et. al., 1990).

Guided by the methodology in psycho-physiological research on emotion (Davidson, et. al., 1990; Azcarraga & Suarez, 2013), the data taken during the learning sessions were converted into deviations from the baseline EEG of the 'resting-state' session. Raw signals are converted into positive or negative deviations from the baseline EEG signal (Azcarraga et. al., 2011).

For the experiments on selective prediction, Table 1 shows the number of students/learners (subjects) according to increasing percentages of feature values that deviate from the baseline mean by at least one-standard deviation. Such features are referred to as 'special event' features. The number of students are shown for each of the two sessions, namely Aplusix and Scooter. For example, 20 students for the Scooter session, in at least one of their EEG recordings, have at least 10% of the feature values (at least 13 of the 126 features) that have values of less than -1.0, or higher than +1.0. Note that since the feature values have been normalized and transformed into standardized scores (z-scores), the feature value already represents the deviation from the mean (normalized to 0.0) in terms of the number of standard deviations from the mean.

Obviously, at 0%, all the 49 students with all their instances are included in the Aplusix sessions, while all 47 are included in the Scooter session. Note that the (maximum) number is different per session because although all 56 human subjects participated in all the sessions, some of their recordings for an entire session have been dropped due to technical failures (e.g. when the EEG headset set got disengaged).

Table 1. Number of students per session according to percentage of special event features whose values deviate from the mean by at least one standard deviation.

	Percentage (%) of Special Event Detector Features										
Session	**0**	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**
Aplusix	49	40	28	25	22	15	11	11	9	9	7
Scooter (Scatterplot)	47	40	33	32	25	20	18	15	10	10	9

As the percentage (%) of 'special event' features increases, the number of students involved for testing decreases, since there would be fewer and fewer instances that deviate significantly (one standard deviation) from the mean. Of course, not all students have EEG recordings that show at least one instance with a minimum of 30% of the features having significantly high or significantly low values, i.e. values less than -1.0 or greater than +1.0. Note that by 20%, there would only be from 7 and 9 students that can be used for testing for Aplusix and Scooter, respectively.

Such features whose values deviate from the mean by at least one standard deviation are referred to here as 'special event' features. In the earlier work (Azcarraga, et. al., 2011), these were referred to as 'outliers', since mathematically, they truly are the outliers of a normal curve. In this study, however, the term 'special event feature' is instead used to refer to those 'outlier' features since these are the features that may signal that something special may have triggered an unusually low or unusually high EEG signal. In some statistical experiments, 'outliers' are sometimes removed (or treated with caution) as they tend to spoil the statistical process. Clearly, this is not the intention in this study, and hence, a new term is adopted.

For the test sets, different datasets were formed for each percentage level of special event features, each of which is classified using the Decision Tree algorithm. Separate datasets were formed for 2%, 4%, 6%, 8%, 10%, 12%, 14%, 16%, 18%, and 20% for the Aplusix and Scooter sessions. Note that these two sessions involve some interactive tasks as students learn algebra (Aplusix) and as they learn how to plot using a specially-designed software module (Scooter).

Student-fold cross validation was employed to maintain mutually exclusive train and test sets. Moreover, for each validation process, the train set is balanced according to emotion tags (high or low). Each student becomes a test set only once and only those instances that have the required percentage of 'special event features' are included in the test set. This applies to 2% to 20% detectors, whereas for the 0%, all instances of that particular student are included in the test. As for the train set, all instances of all students in the train set are included (other than the instances of the specific student in the test set).

Rapidminer was used to perform standard classification (http://www.rapidminer.com), while Octave was used to program specific software modules for data filtering, data normalization, pre-processing, and testing.

5. Results and Discussion

Table 2 presents the classification performance of a Decision Tree classifier based on different 'special event' percentages during the Aplusix and Scooter sessions. The performance measure used in the evaluation is the 'accuracy' of predicting the academic emotion.

For the Aplusix session, as shown in Table 2, we can confirm at a more detailed level the general pattern that as the percentage of the number of special event detectors increases, the prediction accuracy also increases. The accuracy for frustrated increased from 0.5 to 0.55, confused from 0.59 to 0.75, bored from 0.45 to 0.66. However, the accuracy rate for predicting interested did not display the same trend. For the Scooter, predicting interested also did not show the same trend and the most notable result is the increase in Accuracy for the frustrated emotion that increased from 0.50 to 0.74.

It must be noted that in all these experiments, the training set have all been the same. In other words, even for the test set where at least 20% of the features are 'special event features', the train set had been the same as for the 0% test set. As such, although there are some marked increases in the prediction accuracies for the 20% test sets, we imagine that the prediction accuracies will even be higher if different classification models were used for each test set by training the classifier using only those samples with the corresponding proportion of special event features. This is the subject of an ongoing study, and the results and discussion are beyond the scope of this paper.

Table 2. Accuracy for increasing percentages of special event features, for each of the four emotions during the Aplusix and Scooter sessions.

Session	**Emotion**	**0%**	**2%**	**4%**	**6%**	**8%**	**10%**	**12%**	**14%**	**16%**	**18%**	**20%**
Aplusix	Frustrated	0.50	0.50	0.49	0.56	0.6	0.53	0.58	0.55	0.52	0.60	0.55
	Confused	0.59	0.63	0.67	0.63	0.55	0.60	0.62	0.71	0.71	0.70	0.75
	Bored	0.45	0.47	0.47	0.41	0.42	0.41	0.42	0.48	0.50	0.56	0.66
	Interested	0.36	0.41	0.35	0.27	0.27	0.22	0.15	0.18	0.20	0.25	0.18
Scooter	Frustrated	0.50	0.51	0.57	0.51	0.55	0.57	0.57	0.72	0.72	0.73	0.74
	Confused	0.28	0.26	0.28	0.32	0.3	0.32	0.28	0.17	0.10	0.11	0.10
	Bored	0.29	0.3	0.33	0.26	0.27	0.29	0.29	0.33	0.37	0.40	0.41
	Interested	0.16	0.11	0.13	0.12	0.11	0.13	0.11	0.14	0.19	0.17	0.16

6. Conclusion and Future Work

Using an Emotiv EPOC sensor mounted on their head while they learn mathematics using two computer-based learning software, the EEG signals were collected from fifty six (56) academically-gifted students were pre-processed, synchronized, analyzed and used to predict four academic emotions, namely frustrated, confused, bored, and interested. It is shown that the accuracy of EEG-based classification of academic emotions can be improved by selective prediction, where prediction is only performed when a certain pre-determined proportion of EEG feature values deviate significantly from the baseline mean. The hypothesis is that when recordings of an EEG sensor are picking up something 'unusual', then the learner is experiencing some heightened emotion that may then be easier to predict. As such, when prediction is limited to only those instances where a given percentage of feature values deviate significantly from their baseline mean, the emotion prediction accuracy would tend to increase.

A study of the performance rates using decision trees - for instances where 0%, 2%, 4% up to 20% of the features are 'special event features' - shows that the accuracy rate increases from 0.50, 0.59, and 0.45 (for instances with 0% special event features) to 0.74, 0.75, and 0.66 (for instances with 20% special event features) for predicting frustrated (Scooter), confused (Aplusix) and bored (Aplusix), respectively. The accuracy rate for predicting interested did not display the same trend.

All the experiments that were performed only concern academically gifted students. Indeed, this is the sector of students who would benefit the most from independent learning – such as when using computer-based learning modules to learn about specific topics of interest on their own. The results of the study would certainly pave the way for designers of future learning systems to incorporate affect-related features in the system that respond positively to the special needs of learners.

Affect-aware learning systems in the future should have effective predictors of the academic emotion of students – that are able to recognize, with high prediction accuracy, those critical points during learning sessions when students are getting confused about the topic, or perhaps, when they are just about to get frustrated or bored with the learning session. Such affect-aware systems could then prevent frustration or boredom to set in by altering the tutoring tasks, by reinforcing the lessons on the possible sources of confusion using examples and illustrations, by adjusting the level of difficulty of the problems and questions posed, and various other pedagogical and instructional interventions.

Given the results of this study, the prediction should NOT continuously keep on trying to determine the academic emotion of the learner as this would lead to very low prediction accuracies – and hence, interventions will likely to be ineffective. Instead, the system should monitor whether a good percentage (e.g. 20%) of the features have significantly larger or significantly smaller (deviating from the mean by at least 1 standard deviation) values than the baseline mean. And only at these points, should an emotion prediction be attempted.

References

Azcarraga, J., Ibanez JR, J. F., Lim, I. R., Lumanas JR, N., Trogo, R., & Suarez, M. T. (2011). Predicting Academic Emotion based on Brainwaves Signals and Mouse Click Behavior. In T. Hirashima et al. (Ed.), *19th International Conference on Computers in Education (ICCE)* (pp. 42–49). Chiang Mai, Thailand: NECTEC, Thailand.

Azcarraga, J. & Suarez, M. (2013). Recognizing Student Emotions using Brainwaves and Mouse Behavior Data. *International Journal of Distance Education Technologies*, *11*(2), 1–15.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. M. (2009). Emotion Sensors Go To School. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. Graesser (Eds.), *AIED* (Vol. 200, pp. 17–24). IOS Press.

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.

Chanel, G. (2009). *Emotion assessment for affective computing based on brain and peripheral signals*. *University of Geneva*. Citeseer.

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology. I. *Journal of Personality and Social Psychology*, *58*(2), 330–41.

Diezmann, C. (2005). Challenging Mathematically Gifted Primary Students. *Australian Journal of Gifted Education*, *14*(1), 50–57.

Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 319–344). Hillsdale, NJ.: Erlbaum.

Greene, J., Moos, D., Azevedo, R., & Winters, F. (2008). Exploring differences between gifted and grade-level students use of self-regulatory learning processes with hypermedia. *Computers and Education*, *50*, 1069–1083.

Heraz, A, Razaki, R., & Frasson, C. (2007). Using machine learning to predict learner emotional state from brainwaves. *Advanced Learning Technologies 2007 ICALT 2007 Seventh IEEE International Conference on* (Vol. 0, pp. 853–857). IEEE Computer Society.

Jausovec, N. (2000). Differences in Cognitive Processes Between Gifted, Intelligent, Creative, and Average Individuals While Solving Complex Problems: An EEG Study. *Intelligence*, *28*(3), 213–237.

Nicaud, J.-F., Bouhineau, D., & Huguet, T. (2002). The Aplusix Editor: A new kind of software for the learning of algebra. In S. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Lecture notes in computer science* (Vol. 2363, pp. 178–187). Berlin/Heidelberg: Springer.

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.

Stevens, R.H., Galloway, T. and Berka, C. (2007). EEG-Related Changes in Cognitive Workload, Engagement and Distraction as Students Acquire Problem Solving Skills, In C. Conati, K. McCoy, and G. Paliouras (Eds.): *UM 2007*. LNAI 4511, pp. 197–206. Springer-Verlag.