Toward Individualized Real-time Sensor-based Affective Modeling with Intelligent Tutoring Systems

Keith Brawner[1], Jonathan Rowe[2]

[1]United States Army Research Laboratory, [2]North Carolina State University

1 keith.w.brawner.civ@mail.mil, 2 jprowe@ncsu.edu

Abstract. Human tutors do not simply deliver content; they pay attention to the cognitive and affective states of the instructed learners and use this knowledge to adjust their instructional strategies. Thus, a key component of human tutoring is the ability to recognize affect in a learner, and intelligent tutoring systems (ITS) which recognize and classify emotion from data collected on a group of students are prevalent in the literature. However, AI-based software systems that use group-based affective modeling face challenges -- models trained and evaluated with data from *groups* of students may not be effective for *individual* learners. An alternative to this approach is individualized models – highly customized models specific to each individual learner, continuously modified over time based on individual observations. This paper examines individualized modeling techniques for affective state recognition. It reports results from an initial evaluation of individualized modeling techniques using data from WestPoint cadets interacting with a serious game for combat casualty care training.

Keywords: Intelligent Tutoring, Affective Computing, Real-time modeling

1 Introduction and Motivation

Tutoring by an expert human tutor is extraordinarily effective. There is some debate within the literature about how effective human tutors are, but it is commonly cited that tutoring yields between one and two standard deviations of improvement for learners, which corresponds to roughly one to two letter grades [1, 2]. Learning in ITS systems is typically measured in terms of “learning gains”; improved performance in equal time. This is a tradeoff, and could instead represent equivalent performance in less time, improved retention, or other measures of learning outcomes.

Theory indicates that learner data inform learner states which inform instructional strategy selection which influences learning gains [3]; adaptable and individualized tutoring requires automatically assessing the cognitive and affective states of individual learners for personalized instruction [4, 5]. As an example, extensive work has been performed to recognize the emotional state of a learner through incorporating behavioral and physiological sensors [6-10]. The remainder of the paper discusses prior work in generalized modeling, the need for individualized modeling, different AI approaches

Back to Table of Contents

for individualized modeling, the successful results of their application, and recommendations for industrial applications.

2 Background

In 2006, Mott and Lester [7] investigated the inclusion of sensors for affect detection in Crystal Island, an intelligent game-based learning environment that teaches middle school microbiology concepts. This research made use of a variety of features, including temporal interactions, location features, intentional features, physiological response from blood volume pulse and galvanic skin response. These measurements were collected and classified using various machine learning algorithms [7], including Naïve Bayes, decision trees, Support Vector Machines (SVMs), and n-grams. Each of these techniques showed significant predictive accuracy, when compared to baseline accuracy measures. However, when the generalized models were applied in situ, they were found to have worse than baseline classification accuracy [11]. Their 2011 study is one of only two published research articles with validation results across multiple studies, where cross-fold validated models are placed into practice, where Sabourin et al. reported data from 260 learners from two schools; representing a remarkably similar population, and included the injection of experimenter knowledge of student tasks into the models, which is undesirable for transference reasons.

Partially in response to this work and others [12], a new study was designed and conducted to investigate Kinect-based runtime affect modeling [13]. This study used students within a single school different from previous studies, in different semesters, in an attempt to apply the offline-created models to a new setting, without the injection of experimenter knowledge. These models failed to trigger in the operational educational settings at the appropriate times, representing another study which experienced difficulties in application transition. This dataset is used for consideration of the current results and recommendations.

2.1 Individualized Motivation

To date, offline-created, group-based models of learner affect have encountered several challenges in real-world runtime settings. Offline-created, individual-based models present an alternative. Individualized approaches to affective data analysis are rare in the ITS literature, but authors of generalized modeling publications have pointed to individualization as a possible solution to the problem for transferring models into production [9]. Certain types of signals, such as electroencephalography (EEG), naturally lend themselves to individualized approaches (e.g. human brains are very individualistic and modeled as such).

Other researchers indicate that the models are poorly fit for practice when assuming that the underlying concept is stationary, when in fact it is drifting across the sampling space [10, 14]; models should be adaptive and continuously adjusting for the reasons enumerated above. As such, they hypothesize that nonlinear algorithms could successfully deal with the dynamic nature of the signal. AlZoubi et al. empirically show this

Back to Table of Contents

success through an injection of real-time adaptive algorithmic techniques, such as windowed Bayes Networks, which diminished overall classification error by 40% [10]. Generally speaking, the individualized modeling techniques have shown superior performance in other research. Inspired by the prior work, all of the algorithmic approaches in the current work are nonlinear and adaptive.

3 Dataset

There are two datasets subject to analysis in this paper, one from each of 2013 and 2016 [13]. They were both collected from a class of United States Military Academy (USMA) at WestPoint cadets as they interacted with the Tactical Combat Casualty Care Simulation (TC3Sim), with 116 cadets from 2013 and 101 cadets from 2016. TC3Sim is a serious game used to train US Army combat medics and combat lifesavers on tasks associated with dispensing tactical field care and care under fire. Participants in both studies interacted with the system for approximately an hour of total protocol, while approximately 25 minutes were spent within the TC3Sim game. The participants were monitored via within-system interactions as well as via Microsoft Kinect sensor. While the participants interacted with the system, the BROMP protocol [15] was used in order to label the "ground truth" data of affective states of the learners, as observed. There are advantages and disadvantages to different labeling schemes [16], but in-field observations have been found to be relatively stable over time [15].

The initial 2013 collection followed the traditional offline- and group-based model creations, and saw the development of various feature extraction methods, used in both studies to compare benchmark performance. The same features and models from the 2013 study were used in 2016. Of the 91 vertices recorded by the Kinect sensor, only three are utilized for posture analysis: top_skull, head, and center_shoulder. These vertices were selected based on prior work investigating postural indicators of emotion with Kinect data [17]. Derived statistical and windowed features were calculated over top of these items, including the minimum observed, maximum observed, median, variance; each of these features is additionally calculated for 5/10/20 second windows. Further information on the dataset can be found in prior work [13, 18, 19]. 78 input features were used, including raw data, such as CENTER_SHOULDER_DISTANCE reported from the Kinect, and computer features, such as the net_dist_change_20sec. Generally, the raw input features reflect the position and orientation of the head, skull, shoulders, and center of mass, while the computed input features reflect the changes, maximums, minimums, and variances during a 3/5/10/20 second time window. This represents non-extensive feature engineering.

4 Algorithmic Implementations

In order for models to be individualized, the models must be created as new data arrives and operate on under strict time constraints. As such, only machine learning algorithms which have algorithmic complexity of O(1) are appropriate for the task, and the "1" processing requirements of the O(1) operation must be less than the frequency of data

Back to Table of Contents

per user. The algorithms used to create models within this work are the same that have been implemented previously by the lead author, in identical configuration to prior methodologies [12, 20, 21]. They are, in short, an online incremental clustering technique, Adaptive Resonance Theory (ART), and a linear regression approach called Vowpal Wabbit (VW).

5 Results

5.1 Previous Performance Benchmarks

The previous benchmarks for this work, using a variety of offline and generalized classification schemes are shown for the 2013 and 2016 datasets in the tables below, respectively [13]. It is worth noting that the 2013 affect classifiers were applied to the 2016 dataset, but no Kappa value above 0.00 was observed in situ – they were not usable in practice, as referenced in the earlier sections of this work. Additionally, the reader should note that no 'boredom' labels were observed in the 2016 study. The below table represents the best performance of a variety of offline methods given an unlimited amount of modeling time in a cross-validation approach. Naturally, different machine learning methods had different performance, with the best-performing classification approach varying between data signals, and noted in the below table.

Table 1. Performance of detectors of affect, 2013, 2016

Affect	Classifier	A', 2013	A', 2016
Boredom	Logistic Regression	0.528	-
Confusion	Jrip	0.535	0.489
Engaged Concentration	J48	0.532	0.546
Frustration	SVM	0.518	0.331
Surprise	Logistic Regression	0.493	0.51

5.2 Evaluation Methodology

Before a discussion of the results, it is useful to consider how the algorithms operate and are assessed. For each individual a model is created over time in supervised, unsupervised, and semi-supervised fashions. These samples of the model performance represent "best possible algorithmic performance", "worst possible algorithmic performance", and "realistic performance that can be expected in practice", respectively. The semi-supervised models represent effectively unsupervised models with ~6 labeled points for the largest clusters and are majority-labeled – the labeled datapoints represent a direct user query for the label on the 6 minute time scale and are allowed to influence classification boundaries afterwards. As an example, the first 6 minutes of data would be modeled as an unsupervised problem with the next 6 minutes of data being modeled as a mostly unsupervised problem (only one labeled datapoint). Given the sparseness of labeling information in this work (all, none, or 6) in the different implementations,

Back to Table of Contents

overfitting is not a particular concern; 6 labels is not enough to overfit. Further, considering that each created model is started uninitialized with standard model hyperparameters and created for a single individual, the comparing or using this model for another individual wouldn't be sensible; each model is custom to each student. In order to create an evaluation metric which might be compared with the prior work (A' metric) the models are evaluated over time in accordance with the assessment algorithm described in Pseudo-Code 1, feeding an incremental amount of data in, labeling all unknown clusters as the majority class of the true labels, making an A' metric over all data seen so far, and then destroying the evaluated model, which is now polluted with significant labeling information. Additional metrics for the ability to model the near-term past (last 10% of observed data) and near-term future (predictions on the next 10% of data) were found empirically to have within 10% of the overall error of this approach and to generally be measuring the same error rate in prior work [12, 20, 21].

```
For x from 10-100, in increments of 10
 Feed x% of the data to the algorithm
 For each class created by unlabeled class
boundaries
    Label this class the majority label of true set
     Evaluate for AUC ROC accuracy through
     input of data for classification (next, previ-
     ous, all
```

Pseudo-Code 1: Assessment Algorithm

As a byproduct of the evaluation algorithm, each of the models begins with 100% accuracy – a single datapoint generates a single cluster and the majority-class of the cluster is correctly labeled. Gradually, as more data about both the user and labels comes available, the overall accuracy of the model decreases. This decrease represents coming progressively closer to the true accuracy of the approach. This paper answers the question of whether the individual real-time modeling approach is valid. As such, it is useful to see the overall effect of the model, and how useful it would have been, on average, for a given unit of time, and to be able to compare to prior metrics. The algorithm used to assess the performance of each of the methods, per individual, is described below in Pseudo-Code 1. Using this assessment methodology generates 10 assessment points per user. These results are averaged for the group to generate a single metric to compare against prior results.

Back to Table of Contents

5.3 Tabular Results

Table 2. Clustering Performance, 2013[1], 2016[2]

Affect	Prior Best[1]	Sup[1]	Un-Sup[1]	Semi-Sup[1]	Prior Best[2]	Sup[2]	UnSup[2]	Semi-Sup[2]
Boredom	0.528	0.891	0.886	0.888	0.51	-	-	-
Confusion	0.535	0.831	0.820	0.820	0.489	0.750	0.615	0.642
E. Concentration	0.532	0.780	0.765	0.765	0.546	0.647	0.595	0.595
Frustration	0.518	0.936	0.936	0.939	0.331	0.851	0.851	0.851
Surprise	0.493	0.952	0.949	0.949	0.51	0.932	0.932	0.932

Table 3. ART Performance, 2013[1], 2016[2]

Affect	Prior Best[1]	Sup[1]	Un-Sup[1]	Semi-Sup[1]	Prior Best[2]	Sup[2]	UnSup[2]	Semi-Sup[2]
Boredom	0.528	0.886	0.878	0.878	0.51	-	-	-
Confusion	0.535	0.830	0.802	0.802	0.489	0.642	0.630	0.630
E. Concentration	0.532	0.783	0.677	0.677	0.546	0.643	0.558	0.558
Frustration	0.518	0.941	0.939	0.939	0.331	0.851	0.851	0.851
Surprise	0.493	0.955	0.954	0.954	0.51	0.932	0.932	0.932

Table 4. VW Performance, 2013[1], 2016[2]

Affect	Prior Best[1]	Sup[1]	Un-Sup[1]	Semi-Sup[1]	Prior Best[2]	Sup[2]	UnSup[2]	Semi-Sup[2]
Boredom	0.528	0.722	0.718	0.718	0.51	-	-	-
Confusion	0.535	0.699	0.703	0.703	0.489	0.577	0.588	0.588
E. Concentration	0.532	0.716	0.682	0.682	0.546	0.568	0.565	0.565
Frustration	0.518	0.719	0.733	0.733	0.331	0.664	0.655	0.655
Surprise	0.493	0.712	0.710	0.710	0.51	0.663	0.661	0.661

Table 5. Summary Best Semi-Supervised (Realistic, Industrial) Performance

Affect	2013 Method	2013 Value	2016 Method	2016 Value
Boredom	clustering	.888	-	-
Confusion	clustering	.820	clustering	.642
E. Conc.	clustering	.765	clustering	.595
Frustration	Tie	.939	tie	.851
Surprise	ART	.954	tie	.932

6 Discussion and Industrial Applications

Overall, the model performance is favorable, with the indication that the individualized and real-time modeling approach is effective. Naturally, this is an unfair comparison

Back to Table of Contents

to the previous models; these results are comparing an aggregate of many individual models to a single model which models the population. A highlight of these results was previously published in another work [20], which discussed that this performance improvement is not a “free lunch”, and that real-time models should 1) have relatively stable labeling, on the order of minutes, and 2) make use of the created features from offline models, which are shown to help online models. This paper finds similarly.

Recommendations for industrial implementation, based on the above, are for a setup for affective state detection within an intelligent tutoring system to have the following features:

- Sensors of physiological state
- Existing feature extraction shown useful in other contexts – such as the feature extraction performed in this work
- Participant able to label affect states as they come available – a system able to request these items
- Use of one of more machine learning measures, such as ART or incremental clustering, shown above to be the best-performing of the three selected.

This type of implementation can be performed relatively easily within the confines of the Generalized Intelligent Framework for Tutoring (GIFT) system. A specific implementation would be for the Sensor Module to collect, filter, and feature extract the data as above. This data is then sent to the Learner Module, which has the ability to stitch it together with survey-queried ground truth data and models which are created on the fly with algorithmic complexity of O(1). The GIFT system is set up to integrate these types of models with only configuration parameters, rather than any significant module addition or re-architecting.

References

1. B. S. Bloom, "The 2-Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring". *Educational Researcher,* vol. 13, pp. 4-16, 1984.
2. K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educational Psychologist,* vol. 46, pp. 197221, 2011.
3. R. A. Sottilare, K. W. Brawner, B. S. Goldberg, and H. A. Holden, "The Generalized Intelligent Framework for Tutoring (GIFT)," 2012.
4. Department of the Army, "The U.S. Army Learning Concept for 2015," TRADOC2011.
5. B. P. Woolf, "A Roadmap for Education Technology," vol. 0637190, 2010.
6. S. K. D’Mello, R. Taylor, and A. C. Graesser, "Monitoring Affective Trajectories during Complex Learning," in *Proceedings of the 29th Annual Cognitive Science Society*, D. S. McNamara and J. G. Trafton, Eds., ed Austin, TX: Cognitive Science Society, 2007, pp. 203-208.
7. S. McQuiggan, S. Lee, and J. Lester, "Early prediction of student frustration," *Affective Computing and Intelligent Interaction,* pp. 698-709, 2007.
8. S. K. D’Mello, S. D. Craig, B. Gholson, S. Franklin, R. W. Picard, and A. C. Graesser, "Integrating Affect Sensors in an Intelligent Tutoring System," in *Affective Interactions: The*

Back to Table of Contents

Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces, ed New York: AMC Press, 2005, pp. 7-13.
9. R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, pp. 18-37, 2010.
10. O. AlZoubi, R. Calvo, and R. Stevens, "Classification of EEG for Affect Recognition: An Adaptive Approach," *AI 2009: Advances in Artificial Intelligence*, pp. 52-61, 2009.
11. J. Sabourin, B. Mott, and J. C. Lester, "Generalizing Models of Student Affect in Game-Based Learning Environments," in *Affective Computing and Intelligent Interaction*. vol. 6975, S. D. Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., ed Berlin Heidelberg: Springer-Verlag, 2011, pp. 588-597.
12. K. W. Brawner, "Modeling Learner Mood In Realtime Through Biosensors For Intelligent Tutoring Improvements," Doctor of Philosophy in EECS, Department of Electrical Engineering and Computer Science, University of Central Florida, 2013.
13. J. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W. Mott, *et al.*, "Detecting and Addressing Frustration in a Serious Game for Military Training," *International Journal of Artificial Intelligence in Education* 2017.
14. G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," 2001, pp. 97-106.
15. J. Ocumpaugh, R. S. J. d. Baker, and M. M. T. Rodrigo, "Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0.," *New York, NY: EdLab.*, 2012.
16. K. Brawner and M. Boyce, "Establishing ground truth on pyschophysiological models for training machine learning algorithms: Options for ground truth proxies," presented at the International Conference on Augmented Cognition a part of the Human Computer and Intelligent Interaction (HCII) multi-conference, 2017.
17. J. Grafsgaard, J. Wiggins, K. E. Boyer, E. Wiebe, and J. Lester, "Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue," in *Educational Data Mining 2014*, 2014.
18. J. P. Rowe, B. W. Mott, and J. C. Lester, "It's All About the Process: Building SensorDriven Emotion Detectors with GIFT," presented at the GIFTSym2, Pittsburgh, PA, 2014.
19. J. Rowe, E. V. Lobene, and J. Sabourin, "Run-Time Affect Modeling in a Serious Game with the Generalized Intelligent Framework for Tutoring," in *AIED 2013 Workshops Proceedings Volume 7*, 2013, p. 95.
20. K. Brawner, "Lessons Learned For Affective Data And Intelligent Tutoring Systems," presented at the Defense and Homeland Security Simulation, 2017.
21. K. W. Brawner and A. J. Gonzalez, "Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness," *Theoretical Issues in Ergonomics Science*, vol. 17, pp. 183-210, 2016.

Back to Table of Contents