Accurate and Interpretable Sensor-free Affect Detectors via Monotonic Neural Networks

Andrew S. Lan[1], Anthony Botelho[2], Shamya Karumbaiah[3], Ryan S. Baker[3], Neil Heffernan[2]
[1]University of Massachusetts Amherst, [2]Worcester Polytechnic Institute, [3]University of Pennsylvania

ABSTRACT: Sensor-free affect detectors can detect student affect using their activities within intelligent tutoring systems or other online learning environments rather than using sensors. This technology has made affect detection more scalable and less invasive. However, existing detectors are either interpretable but less accurate (e.g., classical algorithms such as logistic regression) or more accurate but uninterpretable (e.g., neural networks). We investigate the use of a new type of neural networks that are *monotonic* after the first layer for affect detection that can strike a balance between accuracy and interpretability. Results on a real-world student affect dataset show that monotonic neural networks achieve comparable detection accuracy to their non-monotonic counterparts while offering some level of interpretability.

Keywords: Affect detection, interpretability, neural networks

1 INTRODUCTION

Affect detectors that can detect and monitor student affective states have become an important aspect of learning analytics research. Together with methods that can trace students' knowledge levels over time, they can support timely and personalized interventions to improve student learning outcomes. Existing student affect detection methods can be classified into two classes. One class employs physical and physiological sensors to measure students as they learn, which is accurate but invasive and not scalable, the other "sensor-free" class uses machine learning-based classifiers to detect a student's affective state from their recorded activity in the ITS, which is non-invasive, scalable, but is in some cases less accurate [Bosch et al., 2015; Henderson et al., 2019]. The trade-off a sensor-free affect detector achieves in terms of accuracy and interpretability is closely related to the type of classification algorithm it uses. Detectors based on classic algorithms such as logistic regression, i.e., [Pardos et al., 2014] can be more interpretable but less accurate, while neural network-based detectors can be more accurate but not interpretable [Botelho et al., 2017]. Therefore, there is a need to develop new classifiers that can find better trade-offs between accuracy and interpretability; we propose to use monotonic neural networks as a potential solution.

2 MONOTONIC (FULLY-CONNECTED) NEURAL NETWORKS

For sensor-free affect detection, we are given a student activity feature vector $\boldsymbol{x} \in \Re^K$, where K denotes the number of features used to summarize student activities within a learning system during an affect observation, and our goal is to detect whether or not a student is in a certain affective state y, which is (typically) binary-valued. Affect detectors are typically classifiers such as logistic regression

$$p(y=1) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = 1/(1+e^{-\boldsymbol{w}^T\boldsymbol{x}}),$$

where $\boldsymbol{w} \in \Re^K$ denotes the regression coefficient (bias is omitted for simplicity of exposition). The values of regression coefficients offer us excellent interpretability since they explicitly control the probability of the student being in this affective state via a linear relationship. Other classic algorithms such as decision trees offer reasonably high interpretability as well, e.g. [Paquette et al., 2014]. Recent research has suggested that neural networks can often achieve significantly better predictive accuracy than logistic regression for binary classification problems [Goodfellow et al., 2016]. In this paper, we use fully connected neural networks to improve the accuracy of affect detection. However, these detectors are often uninterpretable due to the presence of multiple layers and nonlinearities. In order to add interpretability to these neural networks, we propose to investigate the family of “monotonic” neural networks by i) selecting monotonic activation functions and ii) restricting weights beyond the first layer to be nonnegative. We note that common nonlinearities are monotonic, such as hyperbolic tangent ($tanh$) and rectified linear units ($ReLU$) [Goodfellow et al., 2016]. Using a two-layer neural network as an example, for hidden unit i in the first layer, we have

$$p(y = 1) = \sigma\big(W_{2,i}\, z_i + const\big) = \sigma(W_{2,i}\, \Phi(\boldsymbol{w}_i^T \boldsymbol{x}) + const),$$

where Φ denotes the nonlinearity in the first layer, z_i denotes the value of this hidden unit, and $W_{2,i}$ denotes the weight in the second layer connecting this hidden unit to the output. It is easy to show that when Φ is monotonic and $W_{2,i}$ is nonnegative, the probability of a student being in this affective state is also monotonic with respect to $\boldsymbol{w}_i^T \boldsymbol{x}$, a property shared with logistic regression. This observation can be generalized to multi-layer neural networks and enable us to interpret neural network-based affect detectors using the coefficient $\boldsymbol{w}_i$ for each hidden unit in the first layer, if weights in subsequent layers are nonnegative. Despite the presence of nonlinearities at each layer preventing us from comparing the relative importance of features using their coefficients, we can still conclude that whether a feature is positively or negatively correlated with an affective state.

3 EXPERIMENTS

We conduct a series of experiments using monotonic networks as affect detectors on the ASSISTments student affect dataset[1], which was collected in real classrooms as students work within the ASSISTments system by observers following the Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) [Ocumpaugh et al., 2015]. The dataset contains 3109 observations. Each observation contains i) a student's affective state label during a 20-second observation interval and ii) a set of 88 features that summarizes their activities within ASSISTments during this time interval. A total of 4 affective states were coded in this data set: bored, confused, engaged concentration, and frustrated. In this paper, we only analyze the detection of engaged concentration, since it is the most common.

We separate the entire dataset into a training set with 70% of the observations, a validation set with 10% of the observations, and a test set with 20% of the observations. We test four different detectors using four different classifiers: logistic regression (LR), random forest (RF), fully-connected neural network (FNN), and its monotonic version (M-FNN). For each detector, we use the validation set to select the best parameter setting and report detection performance on the test set. For the neural network-based detectors, we sweep over algorithm parameters as learning rate $\in \{1e-5, 1e-4, 1e-3\}$, number of layers $\in \{2,3\}$, number of units in each layer $\in \{5,10,20\}$, nonlinearity $\in \{tanh, ReLU\}$, and different random initializations of the network weights and biases. For the LR and RF detectors, we sweep over the learning rate and number of decision tree parameters, respectively, using a similar approach.

Table 1 shows the performance of each affect detector on the test set, with means and standard deviations calculated over 10 random partitions of the dataset. We see that neural network-based detectors significantly outperform LR- and RF-based detectors, and the monotonic version of the FNN-based detector achieves similar performance to that of its unrestricted version. Table 2 shows the top features and corresponding (regression) coefficients for most predictive features in the LR and M-FNN detectors (we selected one hidden unit in the hidden unit for the latter). We see that the top features (not coefficient values) match up reasonably closely across both cases.

[1] This dataset is taken from http://tiny.cc/affectdata.

Table 1: Engagement detection accuracy on the ASSISTments dataset for all detectors compared.

	AUC
LR	0.746 ± 0.036
RF	0.763 ± 0.029
FNN	0.782 ± 0.030
M-FNN	0.780 ± 0.032

Table 2: Most predictive features for engagement in the LR and M-FNN (1 unit) detectors.

LR		M-FNN	
Feature	Coefficient	Feature	Coefficient
max_frWorkingInSchool	-0.101	max_frWorkingInSchool	-0.471
min_correct	0.096	avg_stlHintUsed	-0.420
avg_hintTotal	-0.066	sum_hintCount	-0.379
sum_timeTaken	-0.054	sum_timeTaken	-0.369
avg_stlHintUsed	-0.032	avg_frPast8WrongCount	-0.359

4 LIMITATIONS AND FUTURE WORK

Though this approach increases interpretability, we have found that we can only interpret the directionality of each unit in the first hidden layer of the neural network separately. Moreover, our monotonic restrictions do not apply to recurrent neural networks e.g., [Botelho et al., 2017] since these restrictions would enforce monotonicity on affect over time as well as activity features. Finally, we have not yet established if similar patterns would hold for other, less frequent affective states.

REFERENCES

Bosch, N., Chen, H., Baker, R. S., Shute, V., D'Mello, S. (2015). Accuracy vs. Availability Heuristic in Multimodal Affect Detection in the Wild. In Proc. *International Conference on Multimodal Interaction*, 267-274.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In Proc. *International Conference on Artificial Intelligence in Education*, 40-51.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Henderson, N., Rowe, J., Mott, B., Brawner, K., Baker, R. S., Lester, J. (2019). 4D Affect Detection: Improving Frustration Detection in Game-based Learning with Posture-based Temporal Data Fusion. In Proc. *International Conference on Artificial Intelligence in Education*, 144-156.

Ocumpaugh, J., Baker, R. S., & Rodrigo, T. (2015). Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*, *60*.

Paquette, L., Baker, R. S., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogoff, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. In Proc. *International Conference on Intelligent Tutoring Systems*, 1-10.

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107-128.