Active Learning for Student Affect Detection

Tsung-Yen Yang[1], Ryan S. Baker[2], Christoph Studer[3], Neil Heffernan[4], and Andrew S. Lan[5]
[1]Princeton University, [2]University of Pennsylvania, [3]Cornell University,
[4]Worcester Polytechnic Institute, [5]University of Massachusetts Amherst

ABSTRACT

"Sensor-free" detectors of student affect that use only student activity data and no physical or physiological sensors are cost-effective and have potential to be applied at large scale in real classrooms. These detectors are trained using student affect labels collected from human observers as they observe students learn within intelligent tutoring systems (ITSs) in real classrooms. Due to the inherent diversity of student activity and affect dynamics, observing the affective states of some students at certain times is likely to be more informative to the affect detectors than observing others. Therefore, a carefully-crafted observation schedule may lead to more meaningful observations and improved affect detectors. In this paper, we investigate whether active (machine) learning methods, a family of methods that adaptively select the next most informative observation, can improve the efficiency of the affect label collection process. We study several existing active learning methods and also propose a new method that is ideally suited for the problem setting in affect detection. We conduct a series of experiments using a real-world student affect dataset collected in real classrooms deploying the ASSISTments ITS. Results show that some active learning methods can lead to high-quality affect detectors using only a small number of highly informative observations. We also discuss how to deploy active learning methods in real classrooms to improve the affect label collection process and thus sensor-free affect detectors.

Keywords

Active learning, L-MMSE estimation, student affect detection

1. INTRODUCTION

Intelligent tutoring systems (ITSs) have gradually seen more and more deployment over the years in real classrooms all over the world. Recently, large-scale randomized controlled trials have shown that they can lead to improved student learning outcomes [30] and affect [19]. However, even the state-of-the-art ITSs cannot interact with students the way human instructors can. For example, in real classrooms, instructors can detect a student's knowledge and affective states by observing their activity and behavior and then adjust their teaching strategy by changing the difficulty of practice questions or addressing negative affect [2, 23]. In particular, keeping students in positive affective states (e.g., engaged) is crucial since their affective states are found to be highly predictive of many metrics of academic performance and success, including test scores [28] and college enrollment [29]. Consequently, there exist many works on designing interventions [1, 10] to address negative affect. Examples of such interventions include selecting appropriate textual dialogues to help students engage [12], using an embodied agent to mirror and empathize with confused students [6], and providing motivational message to frustrated students [19].

Tsung-Yen Yang, Christoph Studer, Ryan S. Baker, Neil T. Heffernan and Andrew Lan "Active Learning for Student Affect Detection" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 208 - 217

1.1 Student Affect Detection

Many existing student affect detection methods employ physical and physiological sensors that make frequent observations of students when they are learning. Despite their effectiveness, these detectors are impractical for large-scale deployment in real classrooms due to cost and privacy constraints [16, 35]. On the other hand, there exists a family of "sensor-free" detectors, which uses only student activity data as they learn within ITSs to detect affect [4, 32, 37]. These detectors use machine learning-based classifiers to predict student affective states from a set of activity features [5]. These sensor-free detectors are more feasible for large-scale deployment than those sensor-dependent ones for two reasons. First, they are cost-effective since once constructed, they can operate in fully-automated fashion and can easily be integrated into ITSs and deployed at large scale. Second, they are more privacy-aware since activity data can be more effectively anonymized than data obtained from other sensors, e.g., video recordings of the students' facial expressions.

Although sensor-free affect detectors are highly automated, the student affect state label collection process remains labor-intensive. The process for collecting these labels typically consists of human observers (including trained coders and/or the teacher) making observations of students in real classrooms and encode their affect into a collection of states. For example, in the Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) for affect observation and coding, there are four affective states: boredom, confusion, engaged concentration, and frustration. The process typically proceeds in round-robin fashion, i.e., the human observer alternates among students and observe one student in each observation interval according to a pre-defined, ad hoc schedule.

However, this data collection process is insufficient since the typical round-robin schedule cannot make full use of the limited time human experts have to make observations. The reason is that, due to the inherent diversity in student activity and affect, the affect states of some students during some observation intervals are more informative to the classifiers than those in other cases; A non-adaptive, ad-hoc observation schedule leads to a lot of missed opportunities to observe these more informative cases. Therefore, it is desirable to develop methods that adaptively select the most informative students to observe in each observation interval and recommend them to human observers. These adaptive methods can potentially lead to the collection of higher-quality data for the affect detector to train on without requiring additional human effort, which will ultimately improve affect detection.

1.2 Active Learning

Active learning refers to a family of machine learning methods that adaptively select the next most "informative" observation to a classifier [34]. These methods are designed for applications where one has access to abundant unlabeled data but can only selectively label a small portion of it. In this setting, there is a need to select data instances whose labels, once obtained, result in the largest improvement in classification quality. There exist numerous active learning methods with different metrics of informativeness; these methods have been found to be effective at reducing the amount of labeled data needed when combining with many classifiers including logistic regression [38], support vector machines [15], and deep convolutional neural networks [33]. See Section 3.1 for a more formal introduction to active learning.

Existing active learning methods are not always successful in practice; in some settings, no active learning methods can outperform the simple baseline approach of randomly selecting data instances to label [38]. One such setting is the "cold-start" setting, when one does not yet have access to a sufficient amount of data to build a good classifier. In this setting, the estimate of informativeness can be highly inaccurate. In affect detection, since there are typically hundreds of features used to summarize student activity in ITSs [5], a classifier needs a significant number of labels to reach reasonable quality. Therefore, the effectiveness of existing active learning methods will be limited in the initial part of the student affect label collection process. Another such setting is when the data is highly noisy; in this case, it is hard to identify informative observations. In affective detection, the affective state labels provided by human observers are highly subjective and thus noisy; the labels provided by different human experts may differ [27] significantly. Therefore, the effectiveness of existing active learning methods in affect detection will be limited by the noisiness of the data. Therefore, it is desirable to develop new active learning methods that are robust to small and noisy data.

1.3 Contributions

In this paper, we investigate whether active learning can be used to improve the efficiency and effectiveness of student affective state label collection. We conduct a preliminary study using several classic active learning methods on an existing real-world student affect dataset collected from ASSISTments[1], a widely-used ITS. Motivated by the limitations of existing active learning methods when the data is small and noisy, we also propose a new active learning method that can excel in this setting. Our new active learning method leverages the recently proposed linear minimum mean squared error (L-MMSE) estimation framework [21, 22] to evaluate observation informativeness. This framework provides an exact, closed-form, and nonasymptotic analysis of the parameter estimation error for binary regression and is shown to be highly effective when data is small and/or noisy. Experimental results show that some active learning methods, especially our L-MMSE-based method, can reduce the number of labels needed to build high-quality, sensor-free affect detectors. We also discuss how to use active learning to improve data collection efficiency in real-world affect detection and possibly other quantitative field observation (QFO) tasks by building an interactive system that suggests human observers to make certain observations.

We emphasize that the purpose of the current work is *not* to improve affect detectors but rather to investigate whether one can collect better data to train them. Therefore, we resort to a simple logistic regression-based affect detector since it can be integrated with all existing active learning methods. More complicated, state-of-the-art deep learning-based detectors cannot be integrated with many active learning methods and thus do not offer us a complete view of active learning in affect detection.

2. RELATED WORK

ASSISTments is a free web-based platform that provides immediate feedback, on-demand hints, and scaffolding support to the many students who use it in classrooms and for daily homework [14]. The system has been used by hundreds of thousands of students and thousands of teachers, and has been found to be effective in improving learning outcomes and closing achievement gaps in a large-scale randomized controlled trial [30].

A significant amount of research has been conducted on the detection of student affect by aligning ASSISTments data to student affect labels collected in real classrooms using BROMP [27]. BROMP allows human observers to label a student in four often-studied affective states: engaged concentration [9], frustration [20], boredom [25], and confusion [8]. Initially, sensor-free affect detectors in ASSISTments leveraged a number of rule-based and statistics-based models; these models achieved performance substantially above chance, for new students from rural, suburban, and urban populations [4]. Later, the work in [36] improved upon these initial affect detectors by incorporating additional features on skills/knowledge components as well as statistics across the entire class. Most recently, the work in [5] applied deep learning methods to affect detection and produced a significant increase in detection accuracy. The key in that work is to use recurrent neural networks (RNNs), including its two popular variants in long short-term memory (LSTM) networks and gated recurrent unit (GRU) networks [13], to

[1] `https://www.assistments.org/`

capture students' changing affect over time.

3. ACTIVE LEARNING

In this section, we will first review active learning and briefly describe how it can be used to improve the efficiency in QFOs. We will then review the L-MMSE estimation framework and introduce our new, L-MMSE-based active learning method.

3.1 Background on Active Learning

Supervised learning refers to a class of machine learning approaches where the task is to learn a function (usually, a classifier) that captures the relation between input-output (feature-label) pairs. The typical setup in supervised learning is that one observes all features and labels and can use them to train the classifier. Active learning, on the other hand, deals with the setting where one has control of the data label observation process; in this case, one has access to the feature values of all feature-label pairs but can select which one gets labeled next. Naturally, the most effective strategy is to train the classifier on observed labels and select the next label that is the most "informative" to the current classifier to observe [34]. There exist numerous active learning methods with different metrics of informativeness, e.g., entropy (or observation uncertainty) [24], expected error reduction [31], expected variance reduction [40], model change [7], etc. The goal of active learning is to only observe labels that are highly informative in order to learn the function more efficiently.

Concretely, we denote the functional relation between the features and labels as

$$\mathbf{y} \sim f_{\mathbf{x}}(\mathbf{D}),$$

where $\mathbf{y} \in \mathcal{A}^N$ is the vector of labels that contains a total of N observations. $\mathcal{A}$ denotes the set of labels. $\mathbf{D} \in \mathbb{R}^{N \times P}$ denotes the matrix containing all feature values corresponding to each label. The column vectors corresponding to the rows of $\mathbf{D}$, i.e., the feature values of each observation, are denoted as $\mathbf{d}_i$, $i \in \{1, \ldots, N\}$. Correspondingly, each element in the label vector is denoted as y_i, $i \in \{1, \ldots, N\}$. $f_{\mathbf{x}}(\cdot)$ denotes the function that maps each input feature vector $\mathbf{d}_i$ to each label y_i; $\mathbf{x}$ denotes the vector containing all parameters of the function. In regression problems, $\mathbf{x}$ corresponds to the regression coefficient vector, while in neural networks, $\mathbf{x}$ corresponds to the collection of all weights and biases that characterize the connections between hidden units.

The iterative process of active learning proceeds as follows. Suppose that one now has a set of $t-1$ observations, with $t \in \{1, 2, \ldots, N\}$ and wants to select the next, t-th observation. Let $\mathcal{O}_{t-1}$ and $\mathcal{U}_{t-1}$ denote the *sets* (which contain indices) of all feature-label pairs (referred to as datasets) where the labels are observed and unobserved, respectively, and let $\hat{\mathbf{x}}_{t-1}$ denote the current estimate of the function parameters (trained on the subset of feature values and labels $\mathbf{D}_{\mathcal{O}_{t-1}}$ and $\mathbf{y}_{\mathcal{O}_{t-1}}$). Active learning methods then select the next observation i_t as

$$i_t = \underset{i \in \mathcal{U}_{t-1}}{\operatorname{argmax}}\, I(\mathbf{d}_i, \hat{\mathbf{x}}_{t-1}),$$

where $I(\cdot, \cdot)$ denotes a metric of how informative an observation i is to the current function. As an example, the simplest yet often most effective existing active learning method, uncertainty sampling, simply uses the entropy [13] as the metric of informativeness:

$$I(\mathbf{d}_i, \hat{\mathbf{x}}_{t-1}) = -\sum_{a \in \mathcal{A}} p(y_i = a) \log p(y_i = a),$$

where $p(y_i = a) = f_{\hat{\mathbf{x}}_{t-1}}(\mathbf{d}_i)$ denotes the probability of a new observation with feature values $\mathbf{d}_i$ taking on label a given the current function estimate parameterized by $\hat{\mathbf{x}}_{t-1}$. In other words, uncertainty sampling simply selects the next observation whose label the current classifier is the least certain of. After selecting the next observation, the classifier is re-trained using an updated observed dataset $\mathcal{O}_t = \mathcal{O}_{t-1} \cup \{i_t\}$. Then, in the next iteration of the active learning process, the next observation i_{t+1} is selected from an updated unobserved dataset $\mathcal{U}_t$.

In typical active learning settings, since one has access to all feature values, the unobserved dataset is simply updated by excluding the selected observation as $\mathcal{U}_t = \mathcal{U}_{t-1} \backslash i_t$. However, we emphasize that under real-world QFO settings, the set of unobserved observations can change entirely. For example, in the typical BROMP coding process, a human coder observes the affective state of one student in each observation interval (typically 20 seconds); therefore, in the next iteration, the set of unobserved dataset (which contains feature values that summarize student activity during the next observation interval) might change entirely.

3.2 Background on L-MMSE Estimation

The L-MMSE estimation framework put forward in [21,22] enables the design of new estimators for a wide range of nonlinear classification and regression problems. It also offers a closed-form, exact, and nonasymptotic analysis of the estimation error for nonlinear problems, which is typically impossible to obtain. The key insight to the L-MMSE estimation framework is that even for nonlinear problems, well-crafted linear estimators that take the nonlinearity into account can achieve comparable performance to nonlinear estimators that are computationally extensive and hard to analyze. Therefore, it is an advanced estimation technique and shall not be confused with basic linear estimation methods like least squares.

In [22], the L-MMSE estimation framework is applied to binary (especially probit) regression, which is given by

$$\mathbf{y} = \operatorname{sign}(\mathbf{D}\mathbf{x} + \mathbf{w}),$$

where $y_i \in \{-1, +1\}$ denotes the binary-valued label for feature-value pair i. The vector $\mathbf{w} \in \mathbb{R}^N$ denotes a noise vector with i.i.d. standard normal random entries. Putting a zero-mean multivariate normal prior with covariance matrix $\mathbf{C}_{\mathbf{x}}$ on $\mathbf{x}$ as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{x}})$, the L-MMSE estimator finds the best estimator of $\mathbf{x}$ that is *linear* in the observation vector $\mathbf{y}$, i.e.,

$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{y},$$

where $\mathbf{W}$ is a suitably-chosen estimation matrix that achieves the minimum mean-squared error (MSE) defined as

$$\text{MSE} = \mathbb{E}_{\mathbf{x},\mathbf{w}}\big[||\mathbf{x} - \hat{\mathbf{x}}||_2^2\big].$$

For probit regression, a variant of binary regression, the L-MMSE estimator has a closed-form expression, given by $\mathbf{W} = \mathbf{E}^T \mathbf{C}_{\mathbf{y}}^{-1}$, with its corresponding MSE given by

$$\mathrm{MSE} = \mathrm{tr}(\mathbf{C_x} - \mathbf{E}^T\mathbf{C_y}^{-1}\mathbf{E}), \tag{1}$$

where

$$\begin{aligned}\mathbf{E} &= \big(\frac{2}{\pi}\big)^{1/2}\mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big)\mathbf{D}\mathbf{C_x},\\ \mathbf{C_y} &= \frac{2}{\pi}\sin^{-1}\Big(\mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big)\mathbf{C_z}\\ &\qquad \times \mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big)\Big),\\ \mathbf{C_z} &= \mathbf{D}\mathbf{C_x}\mathbf{D}^T + \mathbf{I}.\end{aligned}$$

We note that the MSE (and also the matrix $\mathbf{W}$) depends only on the matrix $\mathbf{D}$ and not the label vector $\mathbf{y}$.

3.3 L-MMSE-based Active Learning

Results in [21, 22] have shown that the L-MMSE estimator for binary regression performs on-par with state-of-the-art, sophisticated estimators, e.g., those that require using tools in convex optimization and Markov chain Monte Carlo techniques, while having much lower computational complexity. More importantly, the L-MMSE-based estimation error analysis is shown to be more accurate than other analyses (e.g., those that rely on Fisher information) when the data is noisy and/or when the data is small, i.e., when N is not much larger than P. This advantage is highly desirable in active learning settings and especially in affect detection for two reasons. First, in active learning settings, one often work with small problem sizes: in the initial stages of the active learning process, the classifier is highly inaccurate since it is only trained on a small number of observed labels; therefore, it can lead to an unreliable metric of informativeness which is the key to active learning methods. Second, in affect detection and a lot of other educational applications, the data is inherently noisy: state-of-the-art affect detectors can only achieve area under the receiver operating characteristic curve (AUC) values of around 0.7 after many empirical tweaks [5]. This accuracy is significantly lower than that in common classification tasks [13]. Moreover, inter-coder disagreement on a student's affective state can be high in some cases [27]; this disagreement is also reported in facial expression recognition-based affect detectors [3].

Therefore, we propose a new active learning method that uses the closed-form expression of the MSE of the L-MMSE estimator given in Eq. 1 to measure informativeness since it is reliable even for small and noisy data. Note that we do not use the L-MMSE estimator to estimate $\mathbf{x}$, but only its MSE to select the next observation. Specifically, we use the negative MSE as our metric of informativeness as

$$I(\mathbf{d}_i, \hat{\mathbf{x}}_{t-1}) = -\mathrm{MSE}(\mathbf{D}_{\mathcal{O}_{t-1}\cup\{i\}}).$$

In other words, we select the t-th observation as the one corresponding to the feature vector $\mathbf{d}_i$ that *minimizes* the resulting MSE, i.e.,

$$i_t = \operatorname*{argmin}_{i\in\mathcal{U}_{t-1}} \mathrm{MSE}(\mathbf{D}_{\mathcal{O}_{t-1}\cup\{i\}}),$$

where $\mathbf{D}_{\mathcal{O}_{t-1}\cup\{i\}} = [\mathbf{D}_{\mathcal{O}_{t-1}}^T, \mathbf{d}_i]^T$.

Since the MSE is independent on the observations $\mathbf{y}$, the L-MMSE-based active learning method is likely more robust than all existing methods that rely on $\mathbf{y}$, especially during the initial stage of the active learning process when the number of observations is small. Therefore, it is likely to be highly effective in real-world QFO and especially affect detection settings. This intuition is confirmed by our experiments in Section 4.

In practice, the MSE can be computed very efficiently since the inverse of the matrix $\mathbf{C_y}^{-1}$ only needs to be computed once in every iteration; we do not need to invert it for every potential observation added to the current set of observations. For simplicity of exposition, we temporarily drop the subscripts and use $\mathbf{D}$ and $\mathbf{d}$ to denote the current feature matrix and the feature vector for a possible new observation. The new matrix $\mathbf{C'_z}$ is given by

$$\mathbf{C'_z} = \left[\begin{array}{c}\mathbf{D}\\ \mathbf{d}^T\end{array}\right]\mathbf{C_x}[\mathbf{D}^T\ \mathbf{d}] = \left[\begin{array}{cc}\mathbf{C_z} & \mathbf{D}\mathbf{C_x}\mathbf{d}\\ \mathbf{d}^T\mathbf{C_x}\mathbf{D}^T & \mathbf{d}^T\mathbf{C_x}\mathbf{d}+1\end{array}\right].$$

Now, the new matrix $\mathbf{C'_y}$ is given by

$$\begin{aligned}\mathbf{C'_y} &= \frac{2}{\pi}\sin^{-1}\Big(\mathrm{diag}\big(\mathrm{diag}(\mathbf{C'_z})^{-1/2}\big)\mathbf{C'_z}\mathrm{diag}\big(\mathrm{diag}(\mathbf{C'_z})^{-1/2}\big)\Big)\\ &= \frac{2}{\pi}\sin^{-1}\Big(\left[\begin{array}{cc}\mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big) & \mathbf{0}\\ \mathbf{0}^T & (\mathbf{d}^T\mathbf{C_x}\mathbf{d}+1)^{-1/2}\end{array}\right]\\ &\quad\cdot\left[\begin{array}{cc}\mathbf{C_z} & \mathbf{D}\mathbf{C_x}\mathbf{d}\\ \mathbf{d}^T\mathbf{C_x}\mathbf{D}^T & \mathbf{d}^T\mathbf{C_x}\mathbf{d}+1\end{array}\right]\\ &\quad\cdot\left[\begin{array}{cc}\mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big) & \mathbf{0}\\ \mathbf{0}^T & (\mathbf{d}^T\mathbf{C_x}\mathbf{d}+1)^{-1/2}\end{array}\right]\Big)\\ &= \left[\begin{array}{cc}\mathbf{C_y} & \mathbf{c}\\ \mathbf{c}^T & 1\end{array}\right],\end{aligned}$$

where $\mathbf{c} = \frac{2}{\pi}\sin^{-1}\big(\big(\mathrm{diag}(\mathbf{C'_z})^{-1/2}\big)\mathbf{D}\mathbf{C_x}\mathbf{d}(\mathbf{d}^T\mathbf{C_x}\mathbf{d}+1)^{-1/2}\big)$. Now, using the block matrix inversion rule [17], we have

$$\mathbf{C'_y}^{-1} = \left[\begin{array}{cc}\mathbf{C_y}^{-1} + h\mathbf{g}^T\mathbf{C_y}\mathbf{g} & h\mathbf{g}\\ h\mathbf{g}^T & h\end{array}\right],$$

where $\mathbf{g} = -\mathbf{C_y}^{-1}\mathbf{c}$ and $h = \frac{1}{1-\mathbf{c}^T\mathbf{C_y}^{-1}\mathbf{c}}$. Now, the new matrix $\mathbf{E}'$ is given by

$$\begin{aligned}\mathbf{E}' &= \big(\frac{2}{\pi}\big)^{1/2}\left[\begin{array}{cc}\mathrm{diag}\big(\mathrm{diag}(\mathbf{C_z})^{-1/2}\big) & \mathbf{0}\\ \mathbf{0}^T & (\mathbf{d}^T\mathbf{C_x}\mathbf{d}+1)^{-1/2}\end{array}\right]\\ &\quad\cdot\left[\begin{array}{c}\mathbf{D}\\ \mathbf{d}^T\end{array}\right]\mathbf{C_x} = \left[\begin{array}{c}\mathbf{E}\\ \mathbf{e}^T\end{array}\right],\end{aligned}$$

where $\mathbf{e} = (\frac{2}{\pi})^{1/2}(\mathbf{d}^T\mathbf{C_x}\mathbf{d}+1)^{-1/2}\mathbf{C_x}\mathbf{d}$. Therefore, plugging all of the above into Eq. 1 and some algebra, we get an expression for the new MSE after adding a new observation with feature value vector $\mathbf{d}_i$ as

$$\begin{aligned}\mathrm{MSE}' &= \mathrm{tr}(\mathbf{C_x}) - \mathrm{tr}(\mathbf{E'}^T\mathbf{C'_y}^{-1}\mathbf{E}') = \mathrm{tr}(\mathbf{C_x})\\ &\quad - \mathrm{tr}\Big([\mathbf{E}^T\ \mathbf{e}]\left[\begin{array}{cc}\mathbf{C_y} + h\mathbf{g}^T\mathbf{C_y}^{-1}\mathbf{g} & h\mathbf{g}\\ h\mathbf{g}^T & h\end{array}\right]\left[\begin{array}{c}\mathbf{E}\\ \mathbf{e}^T\end{array}\right]\Big)\\ &= \mathrm{tr}(\mathbf{C_x}) - \mathrm{tr}(\mathbf{E}^T\mathbf{C_y}^{-1}\mathbf{E}) - h\mathrm{tr}(\mathbf{E}^T\mathbf{g}^T\mathbf{C_y}^{-1}\mathbf{g}\mathbf{E})\\ &\quad - 2h\mathrm{tr}(\mathbf{E}^T\mathbf{g}\mathbf{e}^T) - h\mathrm{tr}(\mathbf{e}\mathbf{e}^T)\\ &= \mathrm{MSE} - h(\|\mathbf{E}^T\mathbf{g}+\mathbf{e}\|_2^2),\end{aligned} \tag{2}$$

where the reduction in MSE induced by making a new observation is given by the term $h(\|\mathbf{E}^T\mathbf{g}+\mathbf{e}\|_2^2)$. Therefore, we can obtain the new MSE without having to explicitly calculate $\mathbf{C'_y}^{-1}$ for every possible new observation. In our experiments, we found that this implementation speeds up the L-MMSE-based active learning method by 10 to 100 times, resulting in an empirical computational complexity that is lower than most existing active learning methods except uncertainty sampling.

4. EXPERIMENTAL RESULTS

We now perform a series of experiments on a real-world student affect dataset to explore the effectiveness of active learning methods. We start by adopting standard experimental protocols for active learning under several different settings and then present a simple example to help us understand the conditions under which active learning methods are the most effective.

4.1 Student Affect Dataset

We use an existing dataset for building sensor-free affect detectors collected in real classrooms[2] [5]. The dataset consists of $3,109$ observations, each observation contains i) a student's affective state label during a 20-second observation interval in real classrooms and ii) a set of 88 features that summarizes their activities within ASSISTments during this time interval. These features include the time each student spent on practice items, the number of hints they seek, and the correctness of their responses. We keep observations where the student is labeled as being in one of the four affect states under BROMP: bored, confused, engaged concentration, and frustrated. We leave out the few observations where the human coder indicates that either the student is not in any of the four states or that they are not sure what state the student is in. Engaged concentration is the most frequent state among the four, which occurs about 82% of the time.

Since we focus on logistic regression-based affect detectors in this paper, we need to construct a binary classification problem by detecting the presence of one of the four affective states. We start by building a detector of the engaged concentration affective state since it is the most common among the four states.

4.2 Baseline Active Learning Methods

We test four different active learning methods in our experiments: i) our L-MMSE-based active learning method, (ii) uncertainty sampling (US) [24], as introduced in Section 3.1, (iii) expected variance reduction (EVR) [40], which selects the next observation as the one that results in the largest reduction of the variance of the classifier, and (iv) model change (MC) [7], which selects the observation that changes the classifier's parameters the most. We also use random sampling (Random), which randomly selects the next observation, as the baseline method to simulate the round-robin observation schedule followed in real classrooms when the dataset was collected. We do not test another popular active learning method, expected error reduction [31], since it has very high computational complexity and does not outperform other methods in several preliminary experiments.

4.3 Engaged Concentration Detection

We start by testing active learning methods for a detector of engaged concentration vs. other affective states.

4.3.1 Experimental setup

We use cross validation to test the performance of active learning methods on the ASSISTments student affect dataset. We use two different settings for cross validation: we split the dataset at both the observation level (where each observation is regarded as a stand alone instance) and the student level (where all observation on a student is considered as an instance). We randomly select 20% and 10% of all instances as the test and validation sets, respectively, and use the rest as the training set. The test set is used to evaluate the predictive quality of the trained classifier, using the area under the receiver operating characteristic curve (AUC) metric [18]. This metric takes value in $[0, 1]$ and larger values indicate higher predictive quality.

We start by randomly selecting an initial batch of $M \in \{20, 100, 500\}$ observations (with both student activity feature vector and affective state label for each observation) from the training set; we then use them to train a base logistic regression classifier and use it as our initial affect detector. This experimental setting enables us to study the effectiveness of active learning methods when the amount of prior data available to the detector varies. Although the L-MMSE-based analysis is based on probit regression, we use the more widely-adapted logistic regression to test its robustness against model mismatch. The base classifier is trained using accelerated gradient descent [26] implemented in `TensorFlow`[3] with a $P \times 1$ zero-vector as the initializer. We do not regularize the logistic regression classifier and instead use the validation set to decide when to terminate the training process and avoid overfitting. Specifically, after each (accelerated) gradient descent step, we evaluate the current detector on the validation set, and stop once its predictive quality stops improving (as measured by AUC).

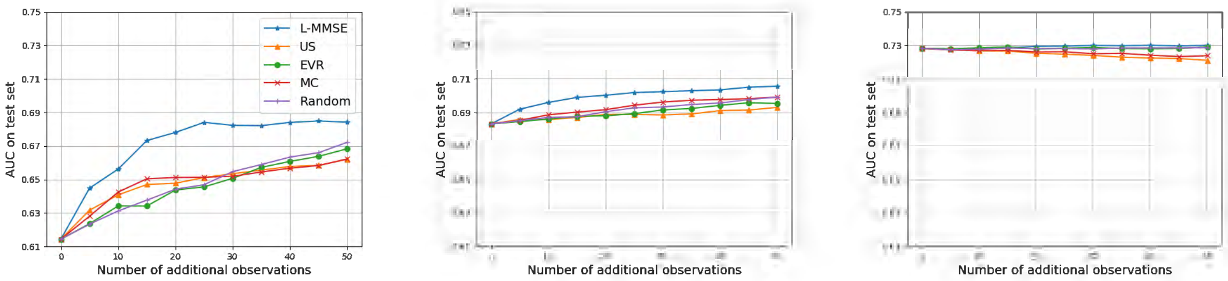
Then, in each iteration of the active learning process, we select the next observation from the remaining ones in the training set according to their feature values, for each active learning method. We then add this new observation (both its feature vector and label) to the current batch and re-train the affect detector, using the previous estimate of the regression coefficients as the initializer. We then calculate the AUC of the re-trained affect detector on the test set. We repeat these steps for a total of 50 additional observations; using more data points is unnecessary since i) we found that using 50 additional observations is enough to summarize the behavior of each active learning method and ii) the performance of the affect detector will converge to the same end point for each active learning method, after going through the entire training set. We also repeat our experiment 100 times and use a different random split of the full dataset and a different initial batch of observations each time. We then report the average results over these repetitions.

4.3.2 Results and discussion

Figure 1 plots the AUC values of the trained affect detectors on the held-out test set vs. the number of additional observations, for all active learning methods on the student affect dataset, using observation-level cross validation. We see that most active learning methods, except EVR, generally outperforms random observation selection when the quality of the affect detector is limited by the amount of data it

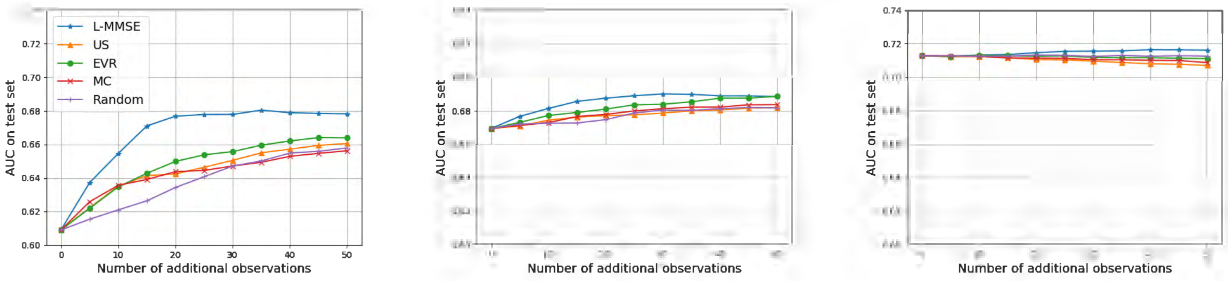
[2] This dataset is taken from `http://tiny.cc/affectdata`

[3] `https://www.tensorflow.org/`



(a) Initial batch size = 20 (b) Initial batch size = 100 (c) Initial batch size = 500

Figure 1: Comparison between different active learning methods for engaged concentration detection with observation-level cross validation. Most active learning methods, especially our L-MMSE-based active learning method, are effective at small initial batch sizes. This advantage over random observation selection diminishes as the quality of the detector saturates when a large number of observations is made.



(a) Initial batch size = 20 (b) Initial batch size = 100 (c) Initial batch size = 500

Figure 2: Comparison between different active learning methods for engaged concentration detection with student-level cross validation. The behavior of active learning methods remain largely the same as observation-level cross validation: they are most effective when the affect detector is trained on few observations.

sees (when it is trained on no more than 50 observations). Our L-MMSE-based method significantly outperforms every other method in this setting. As a concrete example, with 25 additional observations added to the 20 observations in the initial batch, the L-MMSE active learning method results in an that has an AUC of 0.685 on the test set, while no other method result in a detector that has an AUC above 0.65. This result suggests that the L-MMSE-based active learning method excels at picking out observations that are crucial to the affect detector immediately, despite the detector's limited predictive quality; its performance in this setting is impressive since the number of features is quite large ($P = 88$), which is even more than the number of observations in Figure 1(a). Moreover, to reach an AUC value of 0.685 on the test set, the L-MMSE-based active learning method only needs 45 total observations; no other active learning method can achieve this predictive quality even with 70 total observations. This result suggests that, by directing human experts at making observations that are more meaningful to the affect detector, active learning methods can potentially improve the quality of the data without requiring more human effort.

We demonstrate the statistical significance of our results using Student's t-test. Table 1 shows the p-values for rejecting the null hypothesis that the best performing active learning method (L-MMSE) over random observation selection, with an initial batch size of $M = 20$. We see that initially, when the affect detector is not highly accurate, active learning has a significant advantage over random observation selection.

No. of observations	20	30	40	70	100
p-value	3×10^{-3}	2×10^{-7}	2×10^{-9}	6×10^{-3}	4×10^{-1}

Table 1: Statistical significance of the advantage active learning (the L-MMSE-based method) exhibits over random observation selection. Active learning methods are significantly better at the initial stage of the affect observation process.

As the size of the initial batch increases ($M = 100$) and the quality of the initial affect detector improves, the advantage of the L-MMSE-based active learning method over random observation selection drops and eventually diminishes when $M = 500$. This result is not surprising since with 500 initial observations, the performance of the affect detector already saturates (the AUC on the test set after training on the entire training set is 0.74, which is consistent with the values reported in [5]). However, even in this case, the L-MMSE-based active learning method still provides some improvement compared to random observation selection (about 0.01 AUC on the test set with 100 to 150 observations in Figure 1(b)). We note that this advantage is not statistically significant (see Table 1), which is not surprising since the quality of the affect detector improves very slowly after the first 50 observations, leaving very little room for active learning to show its effectiveness.

Perhaps surprisingly, no active learning method except our L-MMSE-based method consistently outperforms random observation selection, even when the initial batch size is small. When the initial batch size is large ($M = 500$), US and MC even leads to worse affect detectors, although we suspect that the performance degradation in that case is due to randomness in cross validation not being sufficiently smoothed out rather than a poor affect detector. These results confirm our intuition that active learning methods designed for general-purpose classification tasks are not well-suited to affect detection, especially when the data size is small during the initial stage of the data collection process.

Figure 2 plots the AUC values of the trained affect detectors on the held-out test set vs. the number of additional observations for all active learning methods, using student-level cross validation. The results largely remain the same compared to observation-level cross validation. Overall, there is a small drop of about 0.01 in test set AUC, confirming the intuition that it is harder for affect detectors to generalize to unseen students than to generalize to unseen observations from current students. However, the L-MMSE-based active learning method still (perhaps even more) consistently outperforms other active learning methods and random observation selection. As a concrete example, with only 10 additional observations in addition to an initial batch of 20 observations, the L-MMSE-based active learning method achieves an AUC of 0.655 on the test set; the other active learning methods and random observation selection achieve AUC values 0.635 and 0.62, respectively. In this case, the effectiveness of using active learning methods (especially our L-MMSE-based method) to identify informative observations and use them to improve affect detection is obvious.

Our experimental results also suggest that there is a lot of redundancy in the ASSISTments student affect dataset. As we discussed above, the quality of the affect detectors saturates after training on about 500 observations. Consider that the entire training set contains more than $2,100$ observations, it seems that the majority of them do not significantly contribute to the quality of the resulting affect detector. This discovery further emphasizes the need of using smarter ways to collect higher-quality data; see Section 5 on a detailed discussion of how to use active learning methods to possibly improve data quality in practice.

4.4 Detection of Other Affective States

We now test the effectiveness of active learning methods for the detection of the other three affective states in BROMP: bored, confused, and frustrated.

4.4.1 Experimental setup

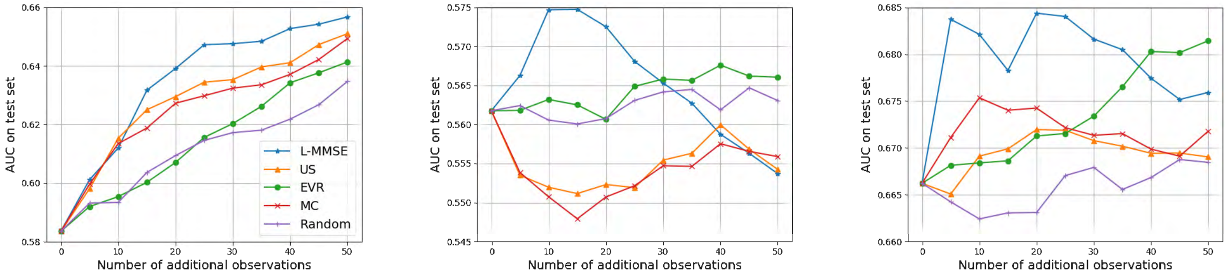
Since these affective states are rare (bored occurs about 10% of the time, while confused and frustrated each occur about 4% of the time) in the ASSISTments dataset, prior work [5, 28] uses resampling to balance among the affective states. Specifically, these works build training datasets that contain roughly equal numbers of observations corresponding to each affective state by resampling from the original training set; after affect detectors are trained on the resampled training dataset, they are then evaluated on the original, non-resampled test set.

We do not use the resampling technique since our goal is to simulate the actual affect observation setting in real-world classrooms, where the four affective states are naturally unbalanced. Therefore, we use same experimental setting as before, except that we have to resort to larger initial batch sizes to ensure that at least a few rare affective states occur in the initial batch. In our experiments, we found that using an initial batch size of $M = 100$ is sufficient.

4.4.2 Results and discussion

Figure 3 plots the AUC values of the trained detectors on the held-out test set vs. the number of additional observations, for the affective states of bored, confused, and frustrated. We used different y-axis ranges in each of the three subplots to enhance contrast since for the confused and frustrated states, the improvement in the quality of the detectors as more observations are made is small. We see that active learning methods, especially our L-MMSE-based method, can still generally outperform random observation selection in most cases (especially for the bored state). However, this advantage is much smaller for these infrequent affective states compared to engaged concentration. For the detection of confusion, two of the active learning methods (US and MC) consistently underperform random observation selection, while our L-MMSE-based method shows some improvement only initially. The only active learning method that performs on-par with random observation selection is the EVR method. One possible explanation is that for harder-to-detect affective states like the confused state, the quality of the classifier is quite low (the final AUC on the test set is only 0.67), which leaves little room for active learning to show its effectiveness.

We now present a simple example to give us some insights on the conditions under which active learning methods are most effective in affect detection. Figure 4 compares the portion of observations selected by an active learning method (US) that actually correspond to an infrequent affect (we used the bored state as an example) to that of random observation selection. We see that after using the initial batch of observations to build a (low-quality) detector, active learning methods can quickly use it to select the observations that actually correspond to the infrequent target affect. Specifically, within the first 50 additional observations, US selects about 15 observations that correspond to the bored affective state, using only student activity features, as it deems these observations more informative; this portion (about 30%) is much higher than the overall portion of the bored state in the entire training set (about 10%). This behavior of US is consistent across all affective states except the confused state, where the portion of observations it selects that actually correspond to the confused state does not exceed the overall portion. In that case, active learning methods also fail to consistently outperform random observation selection, as shown in Figure 3(b). Therefore, active learning methods seem to be effective only if they can strike the right balance between observing different affective states that occur at different frequencies.



(a) Bored (b) Confused (c) Frustrated

Figure 3: Comparison between different active learning methods for infrequent affective state detection (bored, confused, and frustrated). Active learning methods are generally effective for the bored and frustrated states but not the confused state. Their advantage over random observation selection for these states is smaller than that for engaged concentration detection.

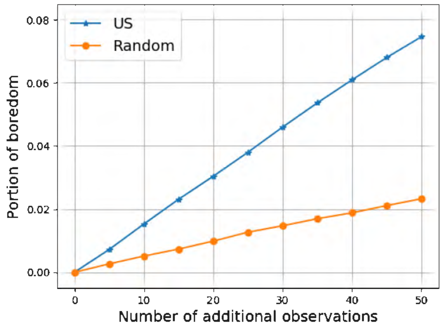


Figure 4: Portion of total infrequent affective state observations selected by an active learning method (US) versus random observation selection for boredom detection. Active learning methods can effectively select observations that actually correspond to the infrequent affect.

5. DEPLOYMENT IN CLASSROOMS

We now outline how to deploy active learning methods in real classrooms to improve the data collection efficiency for affect detection. Since active learning requires training affect detectors on-the-fly as new observations are made, there is a need to create a system that consists of three components. The first component is an interface to human observers making observations in classrooms; this interface i) suggests the human observer to observe a student at each observation interval, ii) collects their affect label on the student, and iii) send the label to the affect detector. The second component is a training paradigm for affect detectors that keeps updating the detector by re-training it after it receives each observed affect label and its corresponding feature vector. The third component is the active learning method that links the other two components together: it i) uses APIs to collect student activity data from ITSs and turn them into feature vectors, ii) selects the next observation that is the most informative to the current affect detector and sends its suggestion via the human observer interface.

There are several realistic considerations in such a system in order for it to be deployed in real classrooms. First, our experiments (see Section 4.4) have shown that active learning methods are not as effective for affective states that occur infrequently (especially the confused state). Therefore, there is a need to explore more advanced active learning methods that take class imbalance into account [11]. Second, experienced human observers may have their own understanding of the informativeness of an observation; such understanding can also be highly valuable to machine learning-based affect detectors. Therefore, the human observer interface should present an option that allows them to ignore the suggestion by active learning methods and instead propose which students to observe on their own. Third, fairness among different student subgroups [39] is critical; we want to ensure that each subgroup is well-observed in the data collection process. Therefore, there needs to be an exploration mechanism that checks whether a student subgroup is under-observed and limit active learning methods to only select among those students when that happens.

6. LIMITATIONS AND FUTURE WORK

In this paper, we have explored the problem of whether active learning methods can be used to increase the efficiency of the affective state label collection process for the development of sensor-free affect detectors. Using an existing student affect dataset collected from ASSISTments, we have shown that active learning methods are indeed effective at making observations that are the most informative to the affect detector; therefore, it can reduce the number of observation needed for the detector to reach a certain quality under most settings. We also proposed a new active learning method that is especially effective for small and noisy data; experimental results show that it outperforms existing active learning methods. At the end, we outlined how to deploy these methods in real-world systems to improve the quality of the data to be collected and discussed several necessary considerations under practical constraints.

Despite the effectiveness of active learning methods, especially our L-MMSE-based method, our work has several limitations and can be extended in many different ways. First, our experimental setting for active learning does not perfectly reflect the actual affective state observation process in real classrooms. In our experimental setting, we select the next observation from all available observations left in the training set, which was collected in many classroom sessions over a long period of time. In practice, when a human observer is making observations in real classrooms, we can only select an observation among the students in class; the most informative observation among these students is generally less informative than the most informative observation possible. Therefore, the benefit active learning methods bring to real-world affect label collection may not be as much as what we have shown in our experiments.

Second, the affect detectors we have studied in this paper are only for detecting the presence of a particular affective state, e.g., bored vs. not bored; it cannot jointly detect all possible affective states. The reason we did so is to test as many active learning methods as possible since most of them are designed only for binary classification. Unfortunately, the most effective active learning method for affect detection in our experiments (our L-MMSE-based method) only applies to binary classification tasks. Therefore, for real-world affect detection problems that are multi-class classification problems, we will extend our method so that it can be applied to multinomial logistic regression instead of binomial logistic regression.

Third, state-of-the-art affect detectors use neural networks rather than logistic regression as their base classifier [5]. While some active learning methods (e.g., uncertainty sampling) can be easily extended to neural networks, others (e.g., our L-MMSE-based method, variance reduction methods, and methods based on model change) cannot since they are either theoretically grounded in binary regression or becomes computationally intractable. Fortunately, the L-MMSE estimation framework encapsulates all the common nonlinearities used in today's state-of-the-art neural network architectures, including the hyperbolic tangent and rectified linear nonlinearities [13]. Therefore, we will extend the L-MMSE-based active learning method proposed in this paper to leverage neural networks as the base classifier.

Fourth, the workflow we outlined for the deployment of active learning in a real-world system in Section 5 presents a time mismatch challenge. In order to select an observation that the human observer should observe, we need access to the corresponding student activity feature vector; these feature values, however, are not available until the end of the observation time interval since many features summarize a student's activity during the entire period. When the teacher receives a suggestion to observe a certain student, this suggestion will be based on the student's activities during the last observation interval, which may not be the most informative observation during the current observation interval. Therefore, we will need to perform a thorough analysis of the coherence in student activity and affect over time to validate the feasibility of deploying active learning in real-world systems for affect label collection.

Finally, the essence of using active learning for affect detection is to leverage the judgement a machine learning-based detector makes on how sure it is about the affective state of a student. Simultaneously, human observers who are trained to make observations in classrooms have their own judgements on how sure they are about a student's affective state. Therefore, comparing the two sets of judgements may lead to deeper insights on how humans perceive affect. Moreover, there is an intrinsic mismatch between the two sets of judgements since one is based on a set of activity features in ITSs while the other is based on observations of activity, gesture, and facial expressions. Therefore, comparing the two sets of judgements may also lead to an analysis of the extent to which the activity features can capture student affect; these insights can potentially help us to design better student activity features or even lead to better ITS designs.

7. REFERENCES

[1] V. Aleven, F. Xhakaj, K. Holstein, and B. M. McLaren. Developing a teacher dashboard for use with intelligent tutoring systems. In *Proc. International Workshop on Teaching Analytics at the European Conference on Technology Enhanced Learning*, pages 15–23, Sep. 2016.

[2] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4):387–426, Dec. 2014.

[3] A. M. Aung and J. Whitehill. Harnessing label uncertainty to improve modeling: An application to student engagement recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 166–170, May 2018.

[4] R. S. Baker, J. Ocumpaugh, S. M. Gowda, A. M. Kamarainen, and S. J. Metcalf. Extending log-based affect detection to a multi-user virtual environment for science. In *Proc. International Conference on User Modeling, Adaptation, and Personalization*, pages 290–300, July 2014.

[5] A. F. Botelho, R. S. Baker, and N. T. Heffernan. Improving sensor-free affect detection using deep learning. In *Proc. International Conference on Artificial Intelligence in Education*, pages 40–51, July 2017.

[6] W. Burleson. Affective learning companions: Strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance. Technical report, Ph.D. Thesis, Massachusetts Institute of Technology, 2006.

[7] W. Cai, Y. Zhang, Y. Zhang, S. Zhou, W. Wang, Z. Chen, and C. Ding. Active learning for classification with maximum model change. *ACM Transactions on Information Systems*, 36(2):15, Sep. 2017.

[8] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250, Oct. 2004.

[9] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Performance*. HarperCollins Publishers, 1990.

[10] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *Proc. International Conference on Intelligent Tutoring Systems*, pages 245–254, June 2010.

[11] S. Ertekin, J. Huang, L. Bottou, and L. Giles. Learning on the border: Active learning in imbalanced data classification. In *Proc. ACM conference on Conference*

on Information and Knowledge Management, pages 127–136, Nov. 2007.

[12] K. Forbes-Riley and D. J. Litman. Adapting to student uncertainty improves tutoring dialogues. In *Proc. International Conference on Artificial Intelligence in Education*, pages 33–40, 2009.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016.

[14] N. T. Heffernan and C. L. Heffernan. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, Dec. 2014.

[15] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proc. International Conference on Machine Learning*, pages 417–424, June 2006.

[16] F. Hollands and I. Bakir. Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods. Technical report, New York, NY: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University, Aug. 2015.

[17] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[18] J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, Mar. 2005.

[19] S. Karumbaiah, R. Lizarralde, D. Allessio, B. P. Woolf, I. Arroyo, and N. Wixon. Addressing student behavior and affect with empathy and growth mindset. In *Proc. International Conference on Educational Data Mining*, pages 96–103, July 2017.

[20] B. Kort, R. Reilly, and R. W. Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proc. IEEE International Conference on Advanced Learning Technologies*, pages 43–46, Aug. 2001.

[21] A. S. Lan, M. Chiang, and C. Studer. An estimation and analysis framework for the Rasch model. In *Proc. International Conference on Machine Learning*, pages 2889–2897, July 2018.

[22] A. S. Lan, M. Chiang, and C. Studer. Linearized binary regression. In *Proc. Conference on Information Sciences and Systems*, pages 1–6, Mar. 2018.

[23] B. Lehman, M. Matthews, S. D'Mello, and N. Person. What are you feeling? Investigating student affective states during expert human tutoring sessions. In *Proc. International Conference on Intelligent Tutoring Systems*, pages 50–59, June 2008.

[24] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, July 1994.

[25] M. Miserandino. Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88(2):203–214, June 1996.

[26] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Université Catholique de Louvain, Sep. 2007.

[27] J. Ocumpaugh, R. S. Baker, and M. M. T. Rodrigo. Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. Technical report, New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences, May 2015.

[28] Z. A. Pardos, R. S. Baker, M. San Pedro, S. M. Gowda, and S. M. Gowda. 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, May 2014.

[29] M. O. Pedro, R. Baker, A. Bowers, and N. Heffernan. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proc. International Conference on Educational Data Mining*, pages 177–184, July 2013.

[30] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA Open*, 2(4):1–12, Oct. 2016.

[31] N. Roy and A. McCallum. Toward optimal active learning through Monte Carlo estimation of error reduction. In *Proc. International Conference on Machine Learning*, pages 441–448, June 2001.

[32] J. L. Sabourin and J. C. Lester. Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing*, 5(1):45–56, Jan. 2014.

[33] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proc. International Conference on Learning Representations*, pages 1–13, May 2018.

[34] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, Nov. 2012.

[35] B. Taylor, A. Dey, D. Siewiorek, and A. Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 517–528, Sep. 2015.

[36] Y. Wang, N. T. Heffernan, and C. Heffernan. Towards better affect detectors: Effect of missing skills, class features and common wrong answers. In *Proc. International Conference on Learning Analytics and Knowledge*, pages 31–35, Mar. 2015.

[37] M. Wixon, I. Arroyo, K. Muldner, W. Burleson, D. Rai, and B. Woolf. The opportunities and limitations of scaling up sensor-free affect detection. In *Proc. International Conference on Educational Data Mining*, pages 145–152, July 2014.

[38] Y. Yang and M. Loog. A benchmark and comparison of active learning for logistic regression. *arXiv preprint arXiv:1611.08618*, 2016.

[39] S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Proc. Conference on Advances in Neural Information Processing Systems*, pages 2921–2930, Dec. 2017.

[40] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proc. International Conference on Machine Learning*, pages 1081–1088, June 2006.