JMLR: Workshop and Conference Proceedings 39:266–282, 2014 ACML 2014

Online Passive Aggressive Active Learning and Its Applications

Jing Lu[†] JING.LU.2014@PHDIS.SMU.EDU.SG
Peilin Zhao[‡] ZHAOP@I2R.A-STAR.EDU.SG
Steven C.H. Hoi[†] CHHOI@SMU.EDU.SG
[†] *School of Information Systems, Singapore Management University, Singapore*
[‡] *Institute for Infocomm Research, A*STAR, Singapore*

Editor: Dinh Phung and Hang Li

Abstract

We investigate online active learning techniques for classification tasks in data stream mining applications. Unlike traditional learning approaches (either batch or online learning) that often require to request the class label of each incoming instance, online active learning queries only a subset of informative incoming instances to update the classification model, which aims to maximize classification performance using minimal human labeling effort during the entire online stream data mining task. In this paper, we present a new family of algorithms for online active learning called Passive-Aggressive Active (PAA) learning algorithms by adapting the popular Passive-Aggressive algorithms in an online active learning setting. Unlike the conventional Perceptron-based approach that employs only the misclassified instances for updating the model, the proposed PAA learning algorithms not only use the misclassified instances to update the classifier, but also exploit correctly classified examples with low prediction confidence. We theoretically analyse the mistake bounds of the proposed algorithms and conduct extensive experiments to examine their empirical performance, in which encouraging results show clear advantages of our algorithms over the baselines.

Keywords: Online Learning, Data Stream, Active Learning, Passive-Aggressive

1. Introduction

Both online learning and active learning have been extensively studied in machine learning and data mining (Freund et al., 1997; McCallum and Nigam, 1998; Balcan et al., 2006; Cesa-Bianchi and Lugosi, 2006; Crammer et al., 2006; Balcan et al., 2007; Castro and Nowak, 2007; Zhao and Hoi, 2010; Hoi et al., 2014). In a traditional online learning task (e.g., online classification), a learner is trained in a sequential manner to predict the class labels of a sequence of instances as accurately as possible. Specifically, at each round of a typical online learning task, the learner first receives an incoming instance, and then makes a prediction of its class label. After that, it is assumed to *always* receive the true class label from an oracle, which can be used to measure the loss incurred by the learner's prediction so as to update the learner if necessary. In many real-world applications especially for mining real-life data streams (e.g., spam email filtering), acquiring the true class labels from an oracle is often time-consuming and costly due to the unavoidable interaction between the learner and the environment. This has motivated the recent study of Online Active

Learning (Cesa-Bianchi et al., 2006; Dasgupta et al., 2009; Cesa-Bianchi and Lugosi, 2006; Sculley, 2007), which explores active learning strategy in an online learning setting to avoid requiring to request class labels of every incoming instance.

A pioneering and state-of-the-art technique to online active learning is known as Label Efficient Perceptron (Cesa-Bianchi and Lugosi, 2006) or Selective Sampling Perceptron (Cesa-Bianchi et al., 2006; Cavallanti et al., 2008), or called Perceptron-based Active Learning (Dasgupta et al., 2009). In particular, consider an online classification task, when a learner receives an incoming instance $\mathbf{x}_t$, the learner first makes a prediction $\hat{y}_t = sign(f(\mathbf{x}_t))$ where $f(\mathbf{x}_t) = \mathbf{w}_t \cdot \mathbf{x}_t$, and then uses a stochastic approach to decide whether it should query the class label or not, where the query probability is inversely proportional to the prediction confidence (e.g., the magnitude of the margin, i.e., $\rho/(\rho+|f(\mathbf{x}_t)|)$ where ρ is a positive smoothing constant). If no class label is queried, the learner makes no update; otherwise, it acquires the true label y_t from the environment and follows the regular Perceptron algorithm to make update (i.e., the learner will update the model if and only if the instance is misclassified according to the true class label).

In the above Perceptron-based active learning, if an incoming instance is predicted with low confidence by the current model, the learner very likely would query its class label. However, if the instance is correctly classified according to the acquired true label, this training instance will be discarded and never be used to update the learner according to the principle of the Perceptron algorithm. Clearly this is a critical limitation of wasting the effort of requesting class labels. To overcome this limitation, we present a new scheme for online active learning, i.e., the Passive-Aggressive Active (PAA) learning, which explores the principle of passive-aggressive learning (Crammer et al., 2006). It not only decides when the learner should make a query appropriately, but also attempts to fully exploit the potential of every queried instance for updating the classification model.

The rest of this paper is organized as follows. Section 2 reviews the background of passive-aggressive online learning. Section 3 presents the proposed PAA algorithms. Section 4 analyzes the mistake bounds of the proposed algorithms. Section 5 discusses our empirical study and Section 6 concludes this work.

2. Background Review

Online learning has been extensively studied in literature (Gaber et al., 2005; Hahsler and Dunham, 2011; Wang et al., 2012c). Specifically, online learning mainly aims to online optimize some performance measures, for example, accuracy (Zhao et al., 2011a; Wang et al., 2012b), balanced accuracy (Wang et al., 2014), AUC (Zhao et al., 2011b), etc. In this paper, our goal is to explore online learning techniques for optimizing the accuracy of binary classification tasks. We first introduce the problem setting of a regular online binary classification task. Let $\{(\mathbf{x}_t, y_t)|\ t = 1, \ldots, T\}$ be a sequence of input patterns for online learning, where each instance $\mathbf{x}_t \in \mathbb{R}^n$ received at the tth trial is a vector of n dimension and $y_t \in \{-1, +1\}$ is its true class label. The goal of online binary classification is to learn a linear classifier $f(\mathbf{x}_t) = sign(\mathbf{w}_t \cdot \mathbf{x}_t)$ where $\mathbf{w}_t \in \mathbb{R}^n$ is the weight vector at the tth round. For the Perceptron algorithm (Rosenblatt, 1958; Freund and Schapire, 1999), a learner first receives an incoming instance $\mathbf{x}_t$ at tth round; it then makes a prediction $\hat{y}_t = sign(f(\mathbf{x}_t))$; finally the true class label y_t is disclosed. If the prediction is correct, i.e., $\hat{y}_t = y_t$, no update is applied to the learner; otherwise, Perceptron updates the model with the misclassified instance $(\mathbf{x}_t, y_t)$, i.e., $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t\mathbf{x}_t$.

Unlike Perceptron that updates the model only when a misclassification occurs, the Passive-Aggressive (PA) algorithms (Crammer et al., 2006) make update whenever the loss function $\ell_t(\mathbf{w}_t; (\mathbf{x}_t, y_t))$ is nonzero, e.g., one can choose the hinge loss $\ell_t(\mathbf{w}_t) = \max(0, 1 - y_t\mathbf{w}_t \cdot \mathbf{x}_t)$. In particular, PA algorithms update the model $\mathbf{w}_{t+1}$ by solving three variants of the optimization task:

$$\arg\min_{\mathbf{w}} F(\mathbf{w}) = \begin{cases} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 \text{ s.t. } \ell_t(\mathbf{w}; (\mathbf{x}_t, y_t)) = 0, & \text{(PA)} \\ \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + C\ell_t(\mathbf{w}; (\mathbf{x}_t, y_t)), & \text{(PA-I)} \\ \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|^2 + C\ell_t(\mathbf{w}; (\mathbf{x}_t, y_t))^2, & \text{(PA-II)} \end{cases}$$

where $C > 0$ is a penalty cost parameter. The closed-form solutions can be derived for the above optimizations, i.e., $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$, where the stepsize τ_t is computed respectively as follows:

$$\tau_t = \begin{cases} \ell_t(\mathbf{w}_t; (\mathbf{x}_t, y_t))/\|\mathbf{x}_t\|^2, & \text{(PA)} \\ \min(C, \ell_t(\mathbf{w}_t; (\mathbf{x}_t, y_t))/\|\mathbf{x}_t\|^2), & \text{(PA-I)} \\ \ell_t(\mathbf{w}_t; (\mathbf{x}_t, y_t))/(\|\mathbf{x}_t\|^2 + 1/(2C)). & \text{(PA-II)} \end{cases} \tag{1}$$

Thus, the PA algorithms are more aggressive in updating the model than Perceptron.

3. Passive-Aggressive Active Learning

In this section, we aim to develop new algorithms for online active learning. Unlike conventional online learning (Rosenblatt, 1958) and pool-based active learning (McCallum and Nigam, 1998; Tong and Koller, 2002), the key challenges to an online active learning task are two-fold: (i) when a learner should query the class label of an incoming instance, and (ii) when the class label is queried and disclosed, how to exploit the labeled instance to update the learner in an effective way. We propose Passive-Aggressive Active (PAA) learning to tackle the above challenges. In particular, the PAA algorithms adopt a simple yet effective randomized rule to decide whether the label of an incoming instance should be queried, and employ state-of-the-art PA algorithms to exploit the labeled instance for updating the online learner.

In particular, for an incoming instance $\mathbf{x}_t$ at the tth round, the PAA algorithm first computes its prediction margin, i.e., $p_t = \mathbf{w}_t \cdot \mathbf{x}_t$, by the current classifier, and then decides if the class label should be queried according to a Bernoulli random variable $Z_t \in \{0, 1\}$ with probability equal to $\rho/(\rho + |p_t|)$, where $\rho \geq 1$ is a smoothing parameter. Such an approach is similar to the idea of margin-based active learning (Tong and Koller, 2002; Balcan et al., 2007) and has been adopted in other previous work (Cesa-Bianchi et al., 2006; Dasgupta et al., 2009). If the outcome $Z_t = 0$, the class label will not be queried and the learner is not updated; otherwise, the class label is queried and the outcome y_t is disclosed. Whenever the class label of an incoming instance is queried, the PAA algorithm will try the best effort to exploit the potential of this example for updating the learner.

Specifically, it adopts the PA algorithms to update the linear classification model w_{t+1} according to Eqn. (1). Clearly this is able to overcome the limitation of the Perceptron-based active learning algorithm that only updates the misclassified instances and wastes a large amount of correctly classified instances with low prediction confidence which can be potentially beneficial to improving the classifier. Finally, we summarize the detailed steps of the proposed PAA algorithms in Algorithm 1.

Algorithm 1 Passive-Aggressive Active Learning Algorithms (**PAA**)

```
INPUT : penalty parameter $C > 0$ and smoothing parameter $\rho \geq 1$.
INITIALIZATION : $\mathbf{w}_1 = (0, \ldots, 0)^\top$.
for $t = 1, \ldots, T$ do
    observe: $\mathbf{x}_t \in \mathbb{R}^n$, set $p_t = \mathbf{w}_t \cdot \mathbf{x}_t$, and predict $\hat{y}_t = sign(p_t)$;
    draw a Bernoulli random variable $Z_t \in \{0, 1\}$ of parameter $\rho/(\rho + |p_t|)$;
    if $Z_t = 1$ then
        query label $y_t \in \{-1, +1\}$, and suffer loss $\ell_t(\mathbf{w}_t) = \max(0, 1 - y_t\mathbf{w}_t \cdot \mathbf{x}_t)$;
        compute $\tau_t$ according to equation (1), and $\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$;
    else
        $\mathbf{w}_{t+1} = \mathbf{w}_t$;
    end if
end for
```

4. Analysis of Mistake Bounds

In this section, we aim to theoretically analyze the mistake bounds of the proposed PAA algorithms. Before presenting the mistake bounds, we begin by presenting a technical lemma which would facilitate the proofs in this section. With this lemma, we could then derive the loss and mistake bounds for the three variants of PAA algorithm. For convenience, we introduce the following notation: $\mathcal{M} = \{t | t \in [T], \hat{y}_t \neq y_t\}$, and $\mathcal{L} = \{t | t \in [T], \hat{y}_t = y_t, \ell_t(\mathbf{w}_t; (\mathbf{x}_t, y_t)) > 0\}$, where $[T]$ denotes $\{1, 2, \ldots, T\}$.

Lemma 1 *Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ be a sequence of input instances, where $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$ for all t. Let τ_t be the stepsize parameter for either of the three PAA variants as given in Eqn. (1). Then, the following bound holds for any $\mathbf{w} \in \mathbb{R}^n$*

$$\sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\big] \leq \alpha^2 \|\mathbf{w}\|^2 + \sum_{t=1}^{T} \tau_t^2 \|\mathbf{x}_t\|^2 + \sum_{t=1}^{T} 2\alpha\tau_t \ell_t(\mathbf{w}),$$

where $M_t = \mathbb{I}_{(t \in \mathcal{M})}$, $L_t = \mathbb{I}_{(t \in \mathcal{L})}$, $\mathbb{I}$ is an indicator function, and $\alpha > 0$.

Proof First of all, we need to prove the following inequality holds for every t

$$\begin{aligned}&(L_t Z_t 2\tau_t(\alpha - |p_t|) + M_t Z_t 2\tau_t(\alpha + |p_t|)\\ &\leq (\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2) + \tau_t^2\|\mathbf{x}_t\|^2 + 2\alpha\tau_t\ell_t(\mathbf{w}).\end{aligned}$$

To prove that, we enumerate all the possible cases for discussions as follows:

Case 1: "$Z_t = 0$" It is clear that the inequality holds with equality since $\mathbf{w}_t = \mathbf{w}_{t+1}$ and $\tau_t = 0$.

Case 2: "$Z_t = 1$ and $M_t = 0$" The label is requested, but no mistake occurs.

Sub-case 2.1: "$L_t = 0$" Since $\ell_t(\mathbf{w}_t) = 0$, $\tau_t = 0$ and $\mathbf{w}_{t+1} = \mathbf{w}_t$. Thus, the inequality holds.

Sub-case 2.2: "$L_t = 1$" Since $\ell_t(\mathbf{w}_t) > 0$, we have

$$\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2 = -2\tau_t y_t \mathbf{w}_t \cdot \mathbf{x}_t + 2\tau_t \alpha y_t \mathbf{w} \cdot \mathbf{x}_t - \tau_t^2 \|\mathbf{x}_t\|^2.$$

Since $\ell_t(\mathbf{w}) = \max(0, 1 - y_t \mathbf{w} \cdot \mathbf{x}_t) \geq 1 - y_t \mathbf{w} \cdot \mathbf{x}_t$, we have

$$\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2 + \tau_t^2 \|\mathbf{x}_t\|^2 + 2\alpha\tau_t \ell_t(\mathbf{w}) \geq 2\tau_t(\alpha - y_t \mathbf{w}_t \cdot \mathbf{x}_t).$$

Also $M_t = 0$ and $\ell_t(\mathbf{w}_t) > 0$ implies $0 \leq y_t \mathbf{w}_t \cdot \mathbf{x}_t < 1$. Thus, we have the inequality

$$|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2 + \tau_t^2 \|\mathbf{x}_t\|^2 + 2\alpha\tau_t \ell_t(\mathbf{w}) \geq 2\tau_t(\alpha - |p_t|).$$

Case 3: "$Z_t = 1$ and $M_t = 1$" It means the label is requested and a mistake occurs, but $L_t = 0$. Similarly, we have

$$\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2 + \tau_t^2 \|\mathbf{x}_t\|^2 + 2\alpha\tau_t \ell_t(\mathbf{w}) \geq 2\tau_t(\alpha - y_t \mathbf{w}_t \cdot \mathbf{x}_t).$$

Since $M_t = 1$ implies $y_t \mathbf{w}_t \cdot \mathbf{x}_t \leq 0$ and $-y_t \mathbf{w}_t \cdot \mathbf{x}_t = |p_t|$, we have

$$\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2 + \tau_t^2 \|\mathbf{x}_t\|^2 + 2\alpha\tau_t \ell_t(\mathbf{w}) \geq 2\tau_t(\alpha + |p_t|).$$

Combining the above cases for all $t = 1, \dots, T$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} (L_t Z_t 2\tau_t(\alpha - |p_t|) + M_t Z_t 2\tau_t(\alpha + |p_t|) \\
&\leq \sum_{t=1}^{T} (\|\mathbf{w}_t - \alpha\mathbf{w}\|^2 - \|\mathbf{w}_{t+1} - \alpha\mathbf{w}\|^2) + \tau_t^2 \|\mathbf{x}_t\|^2 + 2\alpha\tau_t \ell_t(\mathbf{w}) \\
&\leq \alpha^2 \|\mathbf{w}\|^2 + \sum_{t=1}^{T} \tau_t^2 \|\mathbf{x}_t\|^2 + \sum_{t=1}^{T} 2\alpha\tau_t \ell_t(\mathbf{w}) \ .
\end{aligned}$$

■

Based on Lemma 1, we first derive the expected mistake bound for the PAA algorithm in the separable case. We assume there exists some $\mathbf{w}$ such that $y_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 1$, $\forall t \in [T]$.

Theorem 1 *Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_T, y_T)$ be a sequence of input instances, where $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$ and $\|\mathbf{x}_t\| \leq R$ for all t. Assume that there exists a vector $\mathbf{w}$ such that $\ell_t(\mathbf{w}) = 0$ for all t. Then the expected number of mistakes made by the PAA algorithm on this sequence of examples is bounded by*

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \mathbb{E}[\sum_{t=1}^{T} M_t \ell_t(\mathbf{w}_t))] \leq \frac{R^2}{4}(\rho + \frac{1}{\rho} + 2)\|\mathbf{w}\|^2.$$

By setting $\rho = 1$, we can obtain the best upper bound as follows:

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \mathbb{E}[\sum_{t=1}^{T} M_t \ell_t(\mathbf{w}_t))] \leq R^2\|\mathbf{w}\|^2.$$

Proof Since $\ell_t(\mathbf{w}) = 0$, $\forall t \in [T]$, according to Lemma 1, we have

$$\sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\big] \leq \alpha^2\|\mathbf{w}\|^2 + \sum_{t=1}^{T} \tau_t^2 \|\mathbf{x}_t\|^2.$$

Further, the above inequality can be reformulated as:

$$\begin{aligned}
\alpha^2\|\mathbf{w}\|^2 &\geq \sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\big] - \sum_{t=1}^{T} \tau_t^2 \|\mathbf{x}_t\|^2 \\
&= \sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{\tau_t}{2}\|\mathbf{x}_t\|^2) + M_t(\alpha + |p_t| - \frac{\tau_t}{2}\|\mathbf{x}_t\|^2)\big] \\
&= \sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{\ell_t(\mathbf{w}_t)}{2}) + M_t(\alpha + |p_t| - \frac{\ell_t(\mathbf{w}_t)}{2})\big] \\
&= \sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{1 - y_t p_t}{2}) + M_t(\alpha + |p_t| - \frac{1 - y_t p_t}{2})\big] \\
&= \sum_{t=1}^{T} Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{1 - |p_t|}{2}) + M_t(\alpha + |p_t| - \frac{1 + |p_t|}{2})\big] \\
&= \sum_{t=1}^{T} L_t Z_t 2\tau_t(\alpha - \frac{1 + |p_t|}{2}) + \sum_{t=1}^{T} M_t Z_t 2\tau_t(\alpha - \frac{1 - |p_t|}{2}).
\end{aligned}$$

Plugging $\alpha = \frac{\rho+1}{2}$, $\rho \geq 1$ into the above inequality results in

$$(\frac{1+\rho}{2})^2\|\mathbf{w}\|^2 \geq \sum_{t=1}^{T} M_t Z_t \tau_t(\rho + |p_t|),$$

since when $L_t = 1$, $|p_t| \in [0, 1)$, $(\alpha - \frac{1+|p_t|}{2}) = \frac{\rho-|p_t|}{2} > 0$, and $(\alpha - \frac{1-|p_t|}{2}) = \frac{\rho+|p_t|}{2}$.

In addition, combining the fact $\tau_t = \ell_t(\mathbf{w}_t)/\|\mathbf{x}_t\|^2 \geq \ell_t(\mathbf{w}_t)/R^2$ with the above inequality concludes:

$$(\frac{1+\rho}{2})^2\|\mathbf{w}\|^2 \geq \frac{1}{R^2}\sum_{t=1}^{T} M_t Z_t \ell_t(\mathbf{w}_t)(\rho + |p_t|).$$

Taking expectation with the above inequality results in

$$\begin{aligned}
\mathbb{E}[\frac{1}{R^2}\sum_{t=1}^{T} M_t \ell_t(\mathbf{w}_t) Z_t(\rho + |p_t|)] &= \mathbb{E}[\frac{1}{R^2}\sum_{t=1}^{T} M_t \ell_t(\mathbf{w}_t)(\rho + |p_t|)\mathbb{E}Z_t] \\
&= \frac{1}{R^2}\mathbb{E}[\rho\sum_{t=1}^{T} M_t \ell_t(\mathbf{w}_t)] \leq (\frac{1+\rho}{2})^2\|\mathbf{w}\|^2.
\end{aligned}$$

■

The above mistake bound indicates that the expected number of mistakes is proportional to the upper bound of the instances norm R and inversely proportional to the margin $1/\|\mathbf{w}\|^2$, which is consistent with existing research (Crammer et al., 2006). One disadvantage of the above theorem is the linear separable assumption, since real world datasets are usually not separable. To solve this problem, we present the expected mistake bound for the PAA-I algorithm, which is suitable for the non-separable problem.

Theorem 2 *Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_T, y_T)$ be a sequence of examples where $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$ and $\|\mathbf{x}_t\| \le R$ for all t. Then, for any vector $\mathbf{w} \in \mathbb{R}^n$, the expected number of prediction mistakes made by PAA-I on this sequence of examples is bounded from above by*

$$\mathbb{E}[\sum_{t=1}^T M_t] \le \beta \left\{ (\frac{\rho+1}{2})^2 \|\mathbf{w}\|^2 + (\rho+1)C \sum_{t=1}^T \ell_t(\mathbf{w}) \right\},$$

where $\beta = \frac{1}{\rho}\max\{\frac{1}{C}, R^2\}$ and C is the aggressiveness parameter for PAA-I. Setting $\rho = 1$ leads to the following bound

$$\mathbb{E}[\sum_{t=1}^T M_t] \le \max\{\frac{1}{C}, R^2\} \left\{ \|\mathbf{w}\|^2 + 2C \sum_{t=1}^T \ell_t(\mathbf{w}) \right\}.$$

Setting $\rho = \sqrt{1 + \frac{4C\sum_{t=1}^T \ell_t(\mathbf{w})}{\|\mathbf{w}\|^2}}$ leads to the following bound

$$\mathbb{E}[\sum_{t=1}^T M_t] \le \max\{\frac{1}{C}, R^2\} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{t=1}^T \ell_t(\mathbf{w}) + \frac{1}{2}\|\mathbf{w}\| \sqrt{\|\mathbf{w}\|^2 + 4C \sum_{t=1}^T \ell_t(\mathbf{w})} \right\}.$$

Proof According to Lemma 1, we have

$$\begin{aligned}
\alpha^2\|\mathbf{w}\|^2 + \sum_{t=1}^T 2\alpha\tau_t\ell_t(\mathbf{w}) &\ge \sum_{t=1}^T Z_t 2\tau_t \big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\big] - \sum_{t=1}^T \tau_t^2 \|\mathbf{x}_t\|^2 \\
&= \sum_{t=1}^T Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{\tau_t}{2}\|\mathbf{x}_t\|^2) + M_t(\alpha + |p_t| - \frac{\tau_t}{2}\|\mathbf{x}_t\|^2)\big] \\
&\ge \sum_{t=1}^T Z_t 2\tau_t \big[L_t(\alpha - |p_t| - \frac{\ell_t(\mathbf{w}_t)}{2}) + M_t(\alpha + |p_t| - \frac{\ell_t(\mathbf{w}_t)}{2})\big] \\
&= \sum_{t=1}^T L_t Z_t 2\tau_t(\alpha - \frac{1+|p_t|}{2}) + \sum_{t=1}^T M_t Z_t 2\tau_t(\alpha - \frac{1-|p_t|}{2}).
\end{aligned}$$

Similar with Theorem 1, plugging $\alpha = \frac{\rho+1}{2}$, $\rho \ge 1$ into the above inequality will result in

$$(\frac{\rho+1}{2})^2\|\mathbf{w}\|^2 + \sum_{t=1}^T (\rho+1)\tau_t\ell_t(\mathbf{w}) \ge \sum_{t=1}^T M_t Z_t \tau_t(\rho + |p_t|).$$

Since $\tau_t \geq \min\{C, \frac{1}{R^2}\}$, the above inequality implies:

$$(\frac{\rho+1}{2})^2\|\mathbf{w}\|^2+\sum_{t=1}^{T}(\rho+1)\tau_t\ell_t(\mathbf{w})\geq\min\{C,\frac{1}{R^2}\}\sum_{t=1}^{T}M_tZ_t(\rho+|p_t|).$$

Taking expectation with the above equality and re-arranging the result conclude the theorem. ■

This theorem shows that the number of expected mistakes is bounded by a weighted sum of the model complexity $\|\mathbf{w}\|^2$ and the cumulative loss $\sum_{t=1}^T \ell_t(\mathbf{w})$ suffered by it.

Finally, we present the mistake bound for the PAA-II algorithm in the following theorem.

Theorem 3 *Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ be a sequence of examples where $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$ and $\|\mathbf{x}_t\| \leq R$ for all t. Then, for any vector $\mathbf{w} \in \mathbb{R}^n$, the expected number of prediction mistakes made by PAA-II on this sequence of examples is bounded from above by,*

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \gamma\frac{1}{\rho}\{(\frac{\rho+1}{2})^2\|\mathbf{w}\|^2 + 2C(\frac{\rho+1}{2})^2\sum_{t=1}^{T}\ell_t(\mathbf{w})^2\},$$

where $\gamma = \{R^2 + \frac{1}{2C}\}$ and C is the aggressiveness parameter for PAA-II. By setting $\rho = 1$, we can further have

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \{R^2 + \frac{1}{2C}\}\{(\|\mathbf{w}\|^2 + 2C\sum_{t=1}^{T}\ell_t(\mathbf{w})^2\}.$$

Proof Define $\mathcal{O} = \alpha^2\|\mathbf{w}\|^2 + \sum_{t=1}^T \tau_t^2\|\mathbf{x}_t\|^2 + \sum_{t=1}^T 2\alpha\tau_t\ell_t(\mathbf{w})$, $\mathcal{P} = \sum_{t=1}^T \alpha(\frac{\tau_t}{\sqrt{2C\alpha}} - \sqrt{2C\alpha}\ell_t(\mathbf{w}))^2$ and $\mathcal{Q} = \alpha^2\|\mathbf{w}\|^2+\sum_{t=1}^T \tau_t^2(\|\mathbf{x}_t\|^2+\frac{1}{2C})+\sum_{t=1}^T 2C\alpha^2\ell_t(\mathbf{w})^2$, then it is easy to verify that $\mathcal{O} \leq \mathcal{O}+\mathcal{P} = \mathcal{Q}$.

Combing $\mathcal{O} \leq \mathcal{Q}$ with Lemma 1, we get

$$\sum_{t=1}^{T}(L_tZ_t2\tau_t(\alpha - |p_t|) + M_tZ_t2\tau_t(\alpha + |p_t|) \leq \mathcal{Q}.$$

Furthermore, the above formulation can be reformulated as:

$$\begin{aligned}\alpha^2\|\mathbf{w}\|^2 + \sum_{t=1}^{T}2C\alpha^2\ell_t(\mathbf{w})^2 &\geq \sum_{t=1}^{T}Z_t2\tau_t\Big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\Big] - \tau_t^2(\|\mathbf{x}_t\|^2 + \frac{1}{2C})\\ &= \sum_{t=1}^{T}L_tZ_t2\tau_t(\alpha - \frac{1+|p_t|}{2}) + \sum_{t=1}^{T}M_tZ_t2\tau_t(\alpha - \frac{1-|p_t|}{2})\end{aligned}$$

Similar with Theorem 1, plugging $\alpha = \frac{\rho+1}{2}$, $\rho \geq 1$ into the above inequality results in

$$(\frac{\rho+1}{2})^2\|\mathbf{w}\|^2 + \sum_{t=1}^{T}2C(\frac{\rho+1}{2})^2\ell_t(\mathbf{w})^2 > \sum_{t=1}^{T}M_tZ_t\tau_t(\rho + |p_t|).$$

Taking expectation with the above inequality and using $\tau_t \geq 1/\gamma$, will conclude the theorem. ■

This bound is quite similar with the one for Theorem 2.

Remark. As proven in previous work (Cesa-Bianchi et al., 2006), the expected mistake bounds for active learning perceptron, which in our notation, could be expressed as follows:

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \frac{(2\rho + R^2)^2}{8\rho}\|\mathbf{w}\|^2 + (1 + \frac{R^2}{2\rho})\sum_{t=1}^{T}\ell_t(\mathbf{w}).$$

By setting $\rho = 1$, they further have

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \frac{(2 + R^2)^2}{8}\|\mathbf{w}\|^2 + (1 + \frac{R^2}{2})\sum_{t=1}^{T}\ell_t(\mathbf{w}).$$

We could find that generally speaking, the bounds are similar and it depends on the parameters to determine which is better. This is similar to the comparison between the PA bound (Cesa-Bianchi and Lugosi, 2006; Crammer et al., 2006) and Perceptron bound (Freund and Schapire, 1999). However, the bound for Percetron Based Active learning has a $R^4\|w\|^2$ order term, which may make it inferior to ours.

One problem in the above theorem, is that the value of ρ must be larger than 1, which may cause many requests, so we propose the following theorem, which can solve this problem

Theorem 4 *Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ be a sequence of examples where $\mathbf{x}_t \in \mathbb{R}^n$ and $y_t \in \{-1, +1\}$ and $\|\mathbf{x}_t\| \leq R$ for all t. Assume that there exists a vector $\mathbf{w}$ such that $\ell_t(\mathbf{w}) = 0$ for all t. For the PAA algorithm, if change the parameter for the Bernoulli distribution to $\rho/(\rho + 1 + |p_t|)$ and $\rho \geq 0$, then its expected number of prediction mistakes on this sequence of examples is bounded by,*

$$\mathbb{E}[\sum_{t=1}^{T} M_t)] \leq \mathbb{E}[\sum_{t=1}^{T} M_t\ell_t(\mathbf{w}_t)] \leq R^2(\frac{\rho}{4} + \frac{1}{\rho} + 1)\|\mathbf{w}\|^2.$$

When setting $\rho = 2$, we get the best upper bound

$$\mathbb{E}[\sum_{t=1}^{T} M_t] \leq \mathbb{E}[\sum_{t=1}^{T} M_t\ell_t(\mathbf{w}_t))] \leq 2R^2\|\mathbf{w}\|^2.$$

Proof According to Lemma 1,

$$\begin{aligned}\alpha^2\|\mathbf{w}\|^2 + \sum_{t=1}^{T} 2\alpha\tau_t\ell_t(\mathbf{w}) &\geq \sum_{t=1}^{T} Z_t 2\tau_t\big[L_t(\alpha - |p_t|) + M_t(\alpha + |p_t|)\big] - \sum_{t=1}^{T}\tau_t^2\|\mathbf{x}_t\|^2 \\ &= \sum_{t=1}^{T} Z_t 2\tau_t\big[L_t(\alpha - \frac{1+|p_t|}{2}) + M_t(\alpha - \frac{1-|p_t|}{2})\big].\end{aligned}$$

Plugging $\alpha = \frac{\rho}{2} + 1$, $\rho \geq 0$ into the above inequality results in

$$(\frac{\rho}{2} + 1)^2 \|\mathbf{w}\|^2 > \sum_{t=1}^{T} M_t Z_t \tau_t (\rho + 1 + |p_t|),$$

since, when $L_t = 1$, $|p_t| \in [0, 1)$, $(\alpha - \frac{1+|p_t|}{2}) = \frac{\rho+1-|p_t|}{2} > 0$, and $(\alpha - \frac{1-|p_t|}{2}) = \frac{\rho+1+|p_t|}{2}$. Taking expectation with the above inequality and using $\tau_t \geq \ell_t(\mathbf{w}_t)/R^2$ will conclude the theorem. ■

5. Experimental Results

In this section, we evaluate the empirical performance of the proposed Passive Aggressive Active Learning (PAA) algorithms for online active learning tasks.

5.1. Compared Algorithms and Experimental Testbed

We compare the proposed PAA algorithms with the Perceptron-based Active learning, and their random variants, which are listed as follows:

- “RPE”: the Random Perceptron algorithm (Cesa-Bianchi and Lugosi, 2006);
- “RPA”: the Random Passive-Aggressive algorithms, including RPA, RPA-I, RPA-II, which will uniformly randomly query labels;
- “PEA”: the Perceptron-based Active learning algorithm (Cesa-Bianchi et al., 2006);
- “PAA”: the Passive-Aggressive Active learning algorithms, including PAA, PAA-I, PAA-II.

To examine the performance, we conduct extensive experiments on a variety of benchmark datasets from web machine learning repositories. Table 1 shows the details of twelve binary-class datasets used in our experiments. All of these datasets can be downloaded from LIBSVM website [1] and UCI machine learning repository [2]. These datasets are chosen fairly randomly in order to cover various sizes of datasets.

All the compared algorithms learn a linear classifier for the binary classification tasks. The penalty parameter C is searched from $2^{[-5:5]}$ through cross validation for all the algorithms and datasets. The smoothing parameter ρ is set as $2^{[-10:10]}$ in order to examine varied sampling situations. All the experiments were conducted over 20 runs of different random permutations for each dataset. All the results were reported by averaging over these 20 runs. For performance metrics, we select F-measure, which is defined as $F\text{-}measure = 2\frac{Precision*Recall}{Precision+Recall}$.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html

Table 1: Summary of datasets in the experiments.

Dataset	#Instances	#Features
a8a	32561	123
codrna	271617	8
german	1000	24
gisette	7000	5000
ijcnn1	141691	22
magic04	19020	10
mushrooms	8124	112
spambase	4601	56
splice	3175	60
svmguide1	7089	4
svmguie3	1243	21
w8a	64700	300

5.2. Performance Evaluation

Next we evaluate the performance of online active learning tasks. Figure 1 summarizes the average performance of the eight different algorithms for online active learning. Several observations can be drawn from the results in Figure 1.

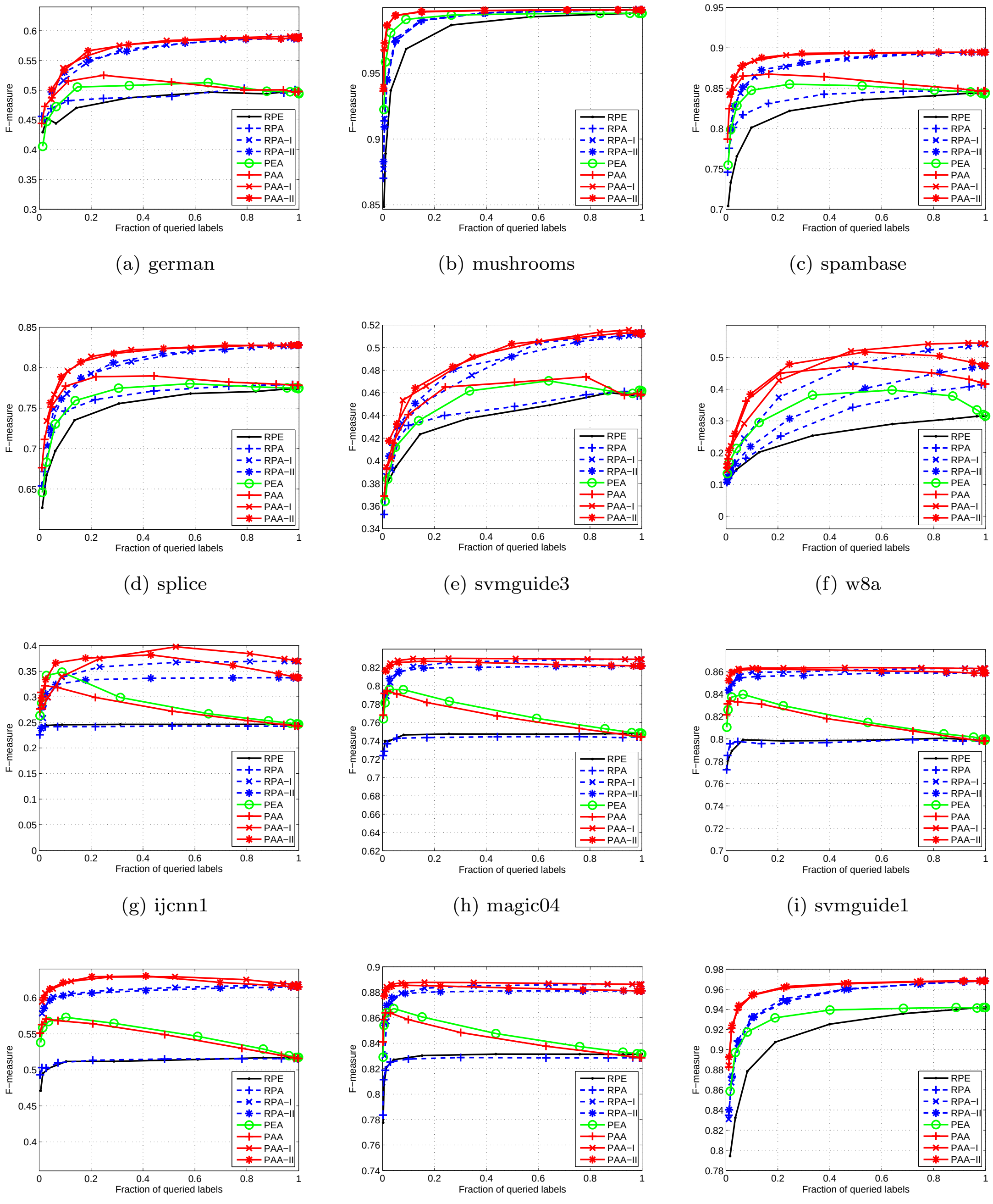
First of all, we observe that all the active learning algorithms outperform their corresponding random version in terms of F-measure results, which validates the efficacy and advantage of the active learning strategies.

Second, we found that the two soft-margin PAA algorithms (i.e., PAA-I and PAA-II) achieve similar F-measure performance on all the datasets, while the hard-margin PAA usually performs slightly worse, which may be caused by overfitting on noisy training data, since PAA conducts a more aggressive update and is thus more sensitive to noise.

Third, we found that under the same fraction of queried labels, the two soft PAA algorithms always achieve significantly higher F-measure than those of the PEA algorithm, while PAA is usually comparable with PEA. This promising result indicates that our PAA strategy can effectively exploit those requested labeled data, especially those instances that are correctly classified but with low confidence.

Fourth, we observe that the F-measure usually increases as the fraction of queried labels increases at the beginning, but saturates quickly after the fraction of queried labels exceeds some value. This result indicates the proposed online active learning strategy can effectively explore those most informative instances for updating the classifiers in a rather effective and efficient way.

Finally, it is interesting to see that on some datasets (e.g., ijcnn1, magic04, svmguide1, etc.), the F-measures achieved by PAA and PEA could decrease when increasing the fraction of queried labels. This seems a little bit surprising as we usually expect the more the labeled data queried, the better the predictive performance. We suspect this was mainly caused due to the overfitting issue on the noisy training data because the other two soft-margin algorithms (PAA-I and PAA-II) tend to be able to avoid such situations.



(a) german

(b) mushrooms

(c) spambase

(d) splice

(e) svmguide3

(f) w8a

(g) ijcnn1

(h) magic04

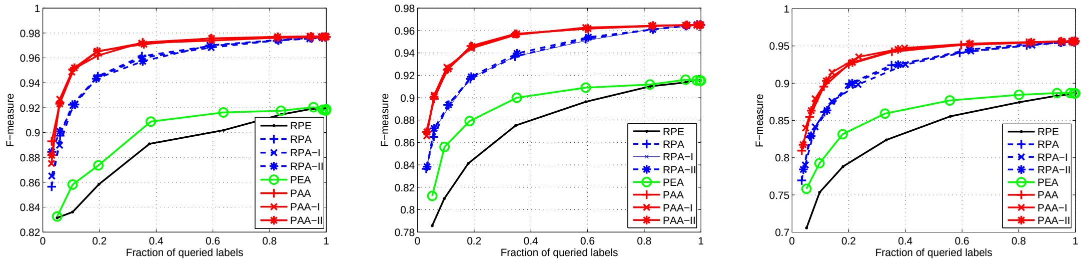
(i) svmguide1

(g) a8a

(h) codrna

(i) gisette

Figure 1: Evaluation of F-measure against the fraction of queried labels on all the datasets. The plotted curves are averaged over 20 random permutations.



(a) RCV1 (b) 20Newsgroup "comp" vs "rec" (c) 20Newsgroup "rec" vs "sci"

Figure 2: Evaluation of F-measure against the fraction of queried labels for text classification applications.

Besides, Table 2 shows detailed F-measure and running time cost of the eight different algorithms for online active learning on several randomly sampled data sets of our testbed. We adjust ρ to make the percentage of queried instances near 10% and 20% and compare the all the algorithms on a fair platform. It is easy to see that PAA algorithms always outperform previous perceptron based algorithms and randomized query approaches, furthermore, the running time cost of PAA and PEA algorithms are similar, as well as in the same order of magnitude with randomized query algorithms, which validates the efficiency of the proposed methods. Among all the algorithms, PAA-II performs best in most cases, which demonstrates the efficacy of soft-margin learning.

5.3. Application to Text Stream Classification

In this section, we apply our proposed Passive-Aggressive Active Learning algorithms to text stream classification. Our experimental testbed consists of: (i) a subset of the Reuters Corpus Volume 1 (RCV1) [3] which contains 4,086 documents with 29,992 distinct words; (ii) 20 Newsgroups datasets [4], we extract the "comp" versus "rec" and "rec" versus "sci" to form two binary classification tasks, which have a total of 8,870 and 8,928 documents, respectively. Each document is represented by a feature vector of 26,214 distinct words.
The text classification results are shown in Figure 2. We could see that Passive Aggressive based algorithms usually outperform the Perceptron based algorithms, which empirically shows the advantages of large margin approaches for active learning. Among all methods, PAA algorithms consistently perform better than random querying methods and perceptron based active learning methods, which further validates the efficacy of our proposed approaches.

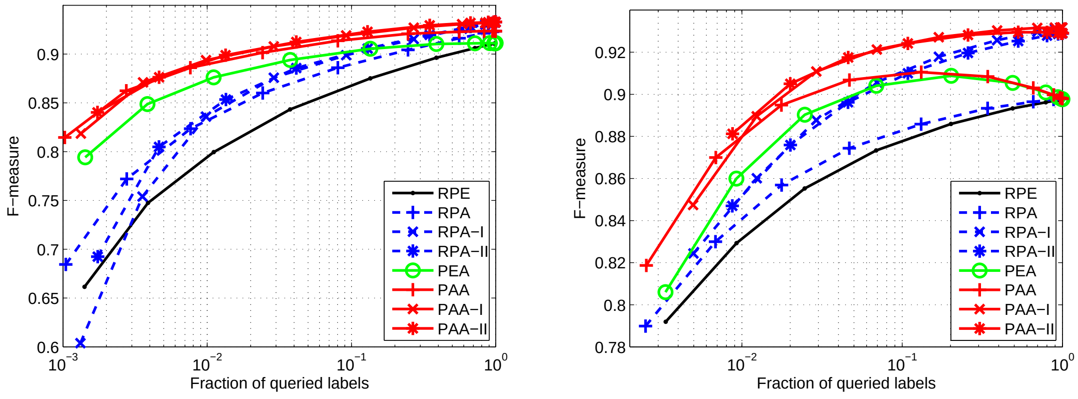
5.4. Application to Web Data Classification

To further evaluate the PAA algorithms, we apply them to web data classification tasks, which are (i) URL classification (Ma et al., 2009) which contains 1,782,206 URLs with 3,231,961 features; (ii) webspam classification (Wang et al., 2012a), which have a total of 350,000 instance with 254 features, respectively. These two datasets can be downloaded

[3] http://thedatahub.org/dataset/rcv1
[4] http://qwone.com/~jason/20Newsgroups/

Table 2: Evaluation of the PAA algorithms against other baselines on three datasets.

Dataset	Algorithm	Request 10% labels			Request 20% labels		
		F-measure	Time (s)	Query(%)	F-measure	Time (s)	Query (%)
german	RPE	0.459 ±0.019	0.010	10.450 ±1.202	0.492 ±0.021	0.011	19.600 ±0.283
	RPA	0.477 ±0.017	0.012	9.550 ±0.071	0.512 ±0.006	0.013	20.450 ±0.354
	RPA-I	0.498 ±0.025	0.012	11.100 ±0.849	0.550 ±0.000	0.013	21.150 ±1.626
	RPA-II	0.514 ±0.006	0.012	9.300 ±0.849	0.568 ±0.010	0.013	21.700 ±0.424
	PEA	0.523 ±0.013	0.062	11.450 ±0.071	0.512 ±0.020	0.062	22.050 ±0.354
	PAA	0.546 ±0.019	0.063	10.200 ±0.990	0.532 ±0.022	0.064	20.800 ±2.263
	PAA-I	0.558 ±0.001	0.063	9.750 ±0.212	0.569 ±0.001	0.065	19.400 ±1.556
	PAA-II	**0.570 ±0.000**	0.063	10.700 ±0.283	**0.575 ±0.023**	0.065	19.800 ±0.424
mushrooms	RPE	0.971 ±0.004	0.096	10.210 ±0.287	0.983 ±0.001	0.099	19.861 ±0.061
	RPA	0.984 ±0.000	0.099	9.866 ±0.339	0.993 ±0.001	0.102	20.347 ±0.035
	RPA-I	0.987 ±0.001	0.099	9.927 ±0.566	0.992 ±0.001	0.103	19.941 ±0.087
	RPA-II	0.989 ±0.000	0.099	9.780 ±0.148	0.993 ±0.000	0.103	20.218 ±0.305
	PEA	0.991 ±0.000	0.513	10.014 ±0.601	0.993 ±0.000	0.516	19.153 ±2.420
	PAA	**0.997 ±0.000**	0.515	9.564 ±0.052	0.997 ±0.000	0.520	19.879 ±0.679
	PAA-I	**0.997 ±0.001**	0.516	9.521 ±0.827	**0.998 ±0.001**	0.524	19.485 ±0.209
	PAA-II	**0.997 ±0.000**	0.515	9.847 ±0.244	0.997 ±0.001	0.520	19.116 ±0.609
spambase	RPE	0.801 ±0.011	0.049	10.411 ±0.676	0.826 ±0.008	0.049	18.952 ±0.123
	RPA	0.830 ±0.023	0.052	10.509 ±0.108	0.846 ±0.010	0.054	20.104 ±0.092
	RPA-I	0.854 ±0.001	0.053	9.324 ±0.031	0.870 ±0.007	0.056	20.170 ±0.584
	RPA-II	0.861 ±0.002	0.055	10.367 ±0.277	0.881 ±0.008	0.059	19.996 ±0.553
	PEA	0.850 ±0.011	0.284	9.965 ±0.200	0.860 ±0.006	0.287	20.680 ±0.323
	PAA	0.871 ±0.002	0.288	10.204 ±0.138	0.864 ±0.004	0.292	18.583 ±0.645
	PAA-I	0.879 ±0.006	0.290	10.802 ±1.260	0.888 ±0.000	0.295	19.170 ±0.676
	PAA-II	**0.882 ±0.010**	0.290	10.335 ±0.353	**0.891 ±0.001**	0.297	20.941 ±0.046



(a) URL

(b) Webspam

Figure 3: Evaluation of F-measure against the fraction of queried labels for web applications.

from the LIBSVM website [5]. Similar phenomenon could be observed from the results, as shown in Figure 3.

6. Conclusion

This paper investigated online active learning techniques for mining data stream. In particular, we presented a new family of Passive-Aggressive Active (PAA) learning algorithms for online active learning, which overcomes the drawback of the existing perceptron-based active learning algorithm that could waste a lot of queried labeled instances that are correctly classified but with low prediction confidence. We theoretically analyzed the mistake bounds for the proposed PAA algorithms, which share almost the same mistake bounds as those regular algorithms when requesting class labels of every instance. Our empirical study found very encouraging performance by comparing the proposed PAA algorithms with the state-of-the-art algorithms. For future work, we plan to address the open issues of online active learning for multi-class classification (Crammer and Singer, 2003) and other challenging data stream mining tasks with concept drift (Minku and Yao, 2012).

References

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *ICML*, pages 65–72, 2006.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *COLT*, pages 35–50, 2007.

Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. In *COLT*, pages 151–156, 2007.

Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. Linear classification and selective sampling under low noise conditions. In *NIPS 21*, pages 249–256, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Worst-case analysis of selective sampling for linear classification. *JMLR*, 7:1205–1230, 2006.

Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *JMLR*, 3:951–991, 2003.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *JMLR*, 7:551–585, 2006.

Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. *JMLR*, 10:281–299, 2009.

Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3):277–296, 1999.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28(2-3):133–168, 1997.

Mohamed Medhat Gaber, Arkady B. Zaslavsky, and Shonali Krishnaswamy. Mining data streams: a review. *SIGMOD Record*, 34(2):18–26, 2005.

Michael Hahsler and Margaret H. Dunham. Temporal structure learning for clustering massive data streams in real-time. In *SDM*, pages 664–675, 2011.

Steven C. H. Hoi, Jialei Wang, and Peilin Zhao. LIBOL: a library for online learning algorithms. *Journal of Machine Learning Research*, 15(1):495–499, 2014. URL http://dl.acm.org/citation.cfm?id=2627450.

Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Identifying suspicious urls: an application of large-scale online learning. In *ICML*, page 86, 2009.

Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *ICML*, pages 350–358, San Francisco, CA, 1998.

Leandro L. Minku and Xin Yao. Ddd: A new ensemble approach for dealing with concept drift. *IEEE Trans. Knowl. Data Eng.*, 24(4):619–633, 2012.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

D. Sculley. Online active learning methods for fast label-efficient spam filtering. In *CEAS*, 2007.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, 2002. ISSN 1532-4435.

De Wang, Danesh Irani, and Calton Pu. Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006. In *CollaborateCom*, pages 40–49, 2012a.

Jialei Wang, Peilin Zhao, and Steven C. H. Hoi. Exact soft confidence-weighted learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012b. URL http://icml.cc/discuss/2012/86.html.

Jialei Wang, Peilin Zhao, and Steven C. H. Hoi. Cost-sensitive online classification. *IEEE Trans. Knowl. Data Eng.*, 26(10):2425–2438, 2014. doi: 10.1109/TKDE.2013.157. URL http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.157.

Peng Wang, Peng Zhang, and Li Guo. Mining multi-label data streams using ensemble-based active learning. In *SDM*, pages 1131–1140, 2012c.

Peilin Zhao and Steven C. H. Hoi. OTL: A framework of online transfer learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 1231–1238, 2010. URL http://www.icml2010.org/papers/219.pdf.

Peilin Zhao, Steven C. H. Hoi, and Rong Jin. Double updating online learning. *Journal of Machine Learning Research*, 12:1587–1615, 2011a. URL http://dl.acm.org/citation.cfm?id=2021051.

Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 233–240, 2011b.