Two-View Online Learning

Tam T. Nguyen, Kuiyu Chang, and Siu Cheung Hui

School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798

Abstract. We propose a two-view online learning algorithm that utilizes two different views of the same data to achieve something that is greater than the sum of its parts. Our algorithm is an extension of the single-view Passive Aggressive (PA) algorithm, where we minimize the changes in the two view weights and disagreements between the two classifiers. The final classifier is an equally weighted sum of the individual classifiers. As a result, disagreements between the two views are tolerated as long as the final combined classifier output is not compromised. Our approach thus allows the stronger voice (view) to dominate whenever the two views disagree. This additional allowance of diversity between the two views is what gives our approach the edge, as espoused by classical ensemble learning theory. Our algorithm is evaluated and compared to the original PA algorithm on three datasets. The experimental results show that it consistently outperforms the PA algorithm on individual views and concatenated view by up to 3%.

1 Introduction

In applications where large amount of data arrives in sequence, e.g., stock market prediction and email filtering, simple online learning such as Perceptron [1], second-order Perceptron [2], and Passive Aggressive (PA) [4] algorithms can be easily deployed with reasonable performance and low computational cost.

For some domains, data may originate from several different sources, also known as views. For example, a web page may have a content view comprising text contained within it, a link view expressing its relationships to other web pages, and a revision view that tracks the different changes that it has undergone.

When the various data sources are independent, running several instances of the same algorithm on it and combining the output via an ensemble learning framework works well. A simple concatenation of the two sources in a vector space model could unnecessarily favor sources with larger number of dimensions. On the other hand, training a separate model on each source fails to make good use of the relationship among the sources, even for a baseline ensemble classifier.

To take advantage of data with multiple views, various methods such as SVM-2K [7] and alternatives [9] have been proposed. However, the two-view methods proposed so far utilizes support vector machine (SVM) [3], which is fundamentally a batch learning algorithm that cannot be easily tailored to work well on large scale online streaming data.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 74–85, 2012.

One simple approach to extend the online learning model to handle two view data is to train one model for each view independently, and combine the classifier outputs just like in classical ensemble learning. However, this approach ignores the relationship between the two views. Instead of using the same idea as SVM-2K where data in one view is used to improve the SVM performance [3] on another view (single view), we take advantage of the relationship between the two views to improve the combined performance. Specifically, we propose a novel online learning algorithm based on the PA algorithm, called Two-view Passive Aggressive (Two-view PA) learning. Our approach minimizes the difference between the two classifier outputs, but allows the outputs to differ as long as the weighted sum of each output leads to the correct result. In classical ensemble learning, the more diverse the classifier, the better the combined performance. In a way, the Two-view PA can be viewed as an ensemble of two online classifiers, except that the two views are jointly optimized.

2 Related Work

Online learning has been researched for more than 50 years. Back in 1962, Block proposed the seminal Perceptron [1] algorithm, while Novikoff [11] later provided theoretical findings, which started the first wave of Artificial Intelligence research in the mid twentieth century. The Perceptron is known to be one of the fastest online learning algorithms. However, its performance is still far from satisfactory in practice. Recently in 2005, Cesa-Bianchi et al. [2] proposed the Second-order Perceptron (SOP) algorithm, which takes advantage of second-order data to improve the accuracy of the original Perceptron. Compared with Perceptron, SOP works better in terms of accuracy but requires more time to train.

In 2006, Crammer et al. [4] proposed another Perceptron-based algorithm, namely the Passive Aggressive (PA) algorithm, which incorporates the margin maximizing criterion of modern machine learning algorithms. They not only have better performance than that of the SOP algorithm but also run significantly faster. Moreover, algorithms that improved upon the PA algorithm include the Passive-Aggressive Mahalanobis [10], the Confidence-Weight (CW) Linear Classifier [6], and its latest version, multi-class CW [5]. The CW algorithm updates its weight by minimizing the Kullback-Leibler divergence between the new and old weights. However, similar to the SOP algorithm, these algorithms are time consuming compared to the first-order PA.

The PA algorithm works better than the SOP in terms of both speed and accuracy. However, it can only process one data stream at one time. On the other hand, in batch learning, Farquhar et al. [7] proposed a large margin two-view Support Vector Machine (SVM) [3] algorithm called the SVM-2K, which is an extension of the well-known SVM algorithm. The two-view learning algorithm was shown to give better performance compared to the original SVM on different image datasets [7]. Thus, SVM-2K provides the inspiration for our current work.

3 Two-View Online Passive Aggressive Learning

3.1 Problem Setting

Online learning aims to learn the weight $\mathbf{w}$ of a linear prediction function $f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$. The online learning algorithm operates in rounds, as input data arrives sequentially. Let $x_t \in \mathbb{R}^n$ be an example arriving at round t. The algorithm predicts its label $\hat{y}_t \in \{-1, +1\}$, after which it receives the true label. If its prediction is correct, the learning process proceeds to the next round. Otherwise, it suffers a loss $\ell(y_t, \hat{y}_t)$, and updates its weight $\mathbf{w}$ accordingly. The loss can be modeled using the hinge-loss function, which equals to zero when the margin exceeds 1, as follows.

$$\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) = \begin{cases} 0 & \text{if} \quad y_t(\mathbf{w}_t \cdot \mathbf{x}_t) \geq 1 \\ 1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t) & \text{otherwise} \end{cases} \tag{1}$$

The overall objective is to minimize the cumulative loss over the entire sequence of examples. From this, Crammer et al. [4] formulated three optimization problems; one based on hard margin and two using soft margins, respectively named PA, PA-I, and PA-II with weight update equations as follows.

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \tau_t y_t \mathbf{x}_t$$

where the coefficient τ_t has one of three forms.

$$\begin{aligned} \tau_t &= \frac{1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)}{\| \mathbf{x}_t \|^2} && \text{(PA)} \\ \tau_t &= \min\left\{C, \frac{1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)}{\| \mathbf{x}_t \|^2}\right\} && \text{(PA-I)} \\ \tau_t &= \frac{1 - y_t(\mathbf{w}_t \cdot \mathbf{x}_t)}{\| \mathbf{x}_t \|^2 + \frac{1}{2C}} && \text{(PA-II)} \end{aligned}$$

The performance of the soft margin based PA-I and PA-II algorithms are almost identical, and both performed better than the hard margin based PA algorithm [4]. Therefore, in this work, our proposed algorithm will be developed based on the PA-I algorithm.

For the two-view online learning setting, training data are triplets $(\mathbf{x}_t^A, \mathbf{x}_t^B, y_t) \in \mathbb{R}^n \times \mathbb{R}^m \times [-1, +1]$, which arrives in sequence where $\mathbf{x}_t^A \in \mathbb{R}^n$ is the first view vector, $\mathbf{x}_t^B \in \mathbb{R}^m$ is the second view vector, and y_t is their common label. The goal is to learn the coupled weights $(\mathbf{w}_t^A, \mathbf{w}_t^B)$ of a *hybrid model* defined as follows.

$$f(\mathbf{x}_t^A, \mathbf{x}_t^B) = \text{sign}\frac{1}{2}(\mathbf{w}_t^A \cdot \mathbf{x}_t^A + \mathbf{w}_t^B \cdot \mathbf{x}_t^B)$$

To incorporate the hybrid classifier, we modify the loss function as follows.

$$\ell((\mathbf{w}_t^A,\mathbf{w}_t^B);(\mathbf{x}_t^A,\mathbf{x}_t^B,y_t)) = \begin{cases} 0 & \text{if} \quad \frac{1}{2}y_t(\mathbf{w}_t^A\cdot\mathbf{x}_t^A+\mathbf{w}_t^B\cdot\mathbf{x}_t^B)\geq 1 \\ 1-\frac{1}{2}y_t(\mathbf{w}_t^A\cdot\mathbf{x}_t^A+\mathbf{w}_t^B\cdot\mathbf{x}_t^B) & \text{otherwise} \end{cases} \tag{2}$$

3.2 Relationship between Views

The primary challenge of multi-view learning is to properly define the relatedness among the different views. In other words, the relatedness quantifies the agreement among the views. Moreover, one could simply disregard the agreement between the two prediction functions, but instead learn the hybrid prediction function. Specifically, we want the hybrid prediction function $f(\mathbf{x}_t^A,\mathbf{x}_t^B) = \text{sign}\frac{1}{2}(\mathbf{w}_t^A\cdot\mathbf{x}_t^A+\mathbf{w}_t^B\cdot\mathbf{x}_t^B)$ to optimally predict the correct labels of examples. In this case, we do not really care whether $f(\mathbf{x}_t^A)$ or $f(\mathbf{x}_t^B)$ can individually classify the example correctly; what we want is for their equally weighted sum $f(\mathbf{x}_t^A,\mathbf{x}_t^B)$ to correctly predict the class label.

Generally, we want the two views to agree with one another. This can be enforced by minimizing their L1-norm or L2-norm disagreements as follows.

$$\sum_{t=1}^{T}|\mathbf{w}_t^A\cdot\mathbf{x}_t^A-\mathbf{w}_t^B\cdot\mathbf{x}_t^B| \quad \text{or} \quad \sum_{t=1}^{T}(\mathbf{w}_t^A\cdot\mathbf{x}_t^A-\mathbf{w}_t^B\cdot\mathbf{x}_t^B)^2 \tag{3}$$

where $|\cdot|$ denotes the absolute function. Here we use L1-norm instead of L2-norm because it is harder to find a close-form solution for the latter. In the next section, we will define an optimization problem based on the L1-norm relatedness measure.

3.3 Two-View Passive Aggressive Algorithm

The ideal objective function should include both the new loss function in (2) and the view relatedness function in (3). Similar to the PA algorithm, the new weights of the two-view learning algorithm are updated based on the optimization problem as follows.

$$\begin{aligned} (\mathbf{w}_{t+1}^A,\mathbf{w}_{t+1}^B) = &\underset{(\mathbf{w}^A,\mathbf{w}^B)\in\mathbb{R}^n\times\mathbb{R}^m}{\text{argmin}} \ \frac{1}{2}\parallel\mathbf{w}^A-\mathbf{w}_t^A\parallel^2+\frac{1}{2}\parallel\mathbf{w}^B-\mathbf{w}_t^B\parallel^2 \\ &+\gamma|y_t\mathbf{w}^A\cdot\mathbf{x}_t^A-y_t\mathbf{w}^B\cdot\mathbf{x}_t^B|+C\xi \\ \text{s.t.} \qquad & 1-\frac{1}{2}(y_t\mathbf{w}^A\cdot\mathbf{x}_t^A+y_t\mathbf{w}^B\cdot\mathbf{x}_t^B)\leq\xi;\xi\geq 0 \end{aligned}$$

where γ and C are positive agreement and aggressiveness parameters respectively. While γ is used to adjust the importance of the agreement between the two views, C is used to control the aggressiveness property of the PA algorithm. Note that the y_t multiplier in the agreement is there just for subsequent derivation convenience.

For the absolute function, we have

$$|y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - y_t\mathbf{w}^B \cdot \mathbf{x}_t^B| = \max(y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - y_t\mathbf{w}^B \cdot \mathbf{x}_t^B, y_t\mathbf{w}^B \cdot \mathbf{x}_t^B - y_t\mathbf{w}^A \cdot \mathbf{x}_t^A)$$

Suppose $z = |y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - y_t\mathbf{w}^B \cdot \mathbf{x}_t^B|$, the above optimization problem can be expressed as follows.

$$\begin{aligned}(\mathbf{w}_{t+1}^A, \mathbf{w}_{t+1}^B) = \underset{(\mathbf{w}^A,\mathbf{w}^B)\in\mathbb{R}^n\times\mathbb{R}^m}{\text{argmin}} & \frac{1}{2} \parallel \mathbf{w}^A - \mathbf{w}_t^A \parallel^2 + \frac{1}{2} \parallel \mathbf{w}^B - \mathbf{w}_t^B \parallel^2 + \gamma z + C\xi \\ \text{s.t.} \qquad & 1 - \frac{1}{2}(y_t\mathbf{w}^A \cdot \mathbf{x}_t^A + y_t\mathbf{w}^B \cdot \mathbf{x}_t^B) \le \xi; \\ & \xi \ge 0; \\ & z \ge y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - y_t\mathbf{w}^B \cdot \mathbf{x}_t^B; \\ & z \ge y_t\mathbf{w}^B \cdot \mathbf{x}_t^B - y_t\mathbf{w}^A \cdot \mathbf{x}_t^A.\end{aligned}$$

Next, we define the Lagrangian of the optimization problem as follows.

$$\begin{aligned}\mathcal{L} = & \frac{1}{2} \parallel \mathbf{w}^A - \mathbf{w}_t^A \parallel^2 + \frac{1}{2} \parallel \mathbf{w}^B - \mathbf{w}_t^B \parallel^2 + \gamma z + C\xi \\ & + \tau\Big(1 - \xi - \frac{1}{2}(y_t\mathbf{w}^A \cdot \mathbf{x}_t^A + y_t\mathbf{w}^B \cdot \mathbf{x}_t^B)\Big) - \lambda\xi \\ & + \alpha(y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - y_t\mathbf{w}^B \cdot \mathbf{x}_t^B - z) + \beta(y_t\mathbf{w}^B \cdot \mathbf{x}_t^B - y_t\mathbf{w}^A \cdot \mathbf{x}_t^A - z) \qquad (4) \\ = & \frac{1}{2} \parallel \mathbf{w}^A - \mathbf{w}_t^A \parallel^2 + \frac{1}{2} \parallel \mathbf{w}^B - \mathbf{w}_t^B \parallel^2 + (\gamma - \alpha - \beta)z + (C - \lambda - \tau)\xi \\ & + (\alpha - \beta - \frac{1}{2}\tau)y_t\mathbf{w}^A \cdot \mathbf{x}_t^A + (\beta - \alpha - \frac{1}{2}\tau)y_t\mathbf{w}^B \cdot \mathbf{x}_t^B + \tau\end{aligned}$$

where α, β, τ, and λ are positive Lagrangian multipliers.

Setting the partial derivatives of $\mathcal{L}$ with respect to the weight $\mathbf{w}^A$ to zero, we have,

$$0 = \frac{\partial\mathcal{L}}{\partial\mathbf{w}^A} = \mathbf{w}^A - \mathbf{w}_t^A + (\alpha - \beta - \frac{1}{2}\tau)y_t\mathbf{x}_t^A \Rightarrow \mathbf{w}^A = \mathbf{w}_t^A - (\alpha - \beta - \frac{1}{2}\tau)y_t\mathbf{x}_t^A \quad (5)$$

Similarly, for the other view we have

$$\mathbf{w}^B = \mathbf{w}_t^B - (\beta - \alpha - \frac{1}{2}\tau)y_t\mathbf{x}_t^B \quad (6)$$

Setting the partial derivatives of $\mathcal{L}$ with respect to weight z to zero, we have

$$0 = \frac{\partial\mathcal{L}}{\partial z} = (\gamma - \alpha - \beta) \Rightarrow \alpha + \beta = \gamma \quad (7)$$

Setting the partial derivatives of $\mathcal{L}$ with respect to weight ξ to zero, we have,

$$0 = \frac{\partial\mathcal{L}}{\partial\xi} = (C - \lambda - \tau) \Rightarrow \lambda + \tau = C \quad (8)$$

Note that $\lambda \ge 0$, thus we can conclude that $0 \le \tau \le C$.

Substituting (5), (6), (7), and (8) into (4), we have,

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}(\alpha-\beta-\frac{1}{2}\tau)^2 \parallel \mathbf{x}_t^A \parallel^2 + \frac{1}{2}(\beta-\alpha-\frac{1}{2}\tau)^2 \parallel \mathbf{x}_t^B \parallel^2 \\ &+(\alpha-\beta-\frac{1}{2}\tau)y_t\Big(\mathbf{w}_t^A-(\alpha-\beta-\frac{1}{2}\tau)y_t\mathbf{x}_t^A\Big)\mathbf{x}_t^A \\ &+(\beta-\alpha-\frac{1}{2}\tau)y_t\Big(\mathbf{w}_t^B-(\alpha-\beta-\frac{1}{2}\tau)y_t\mathbf{x}_t^B\Big)\mathbf{x}_t^B+\tau \\ =& -\frac{1}{2}(\alpha-\beta-\frac{1}{2}\tau)^2 \parallel \mathbf{x}_t^A \parallel^2 - \frac{1}{2}(\beta-\alpha-\frac{1}{2}\tau)^2 \parallel \mathbf{x}_t^B \parallel^2 \\ &+(\alpha-\beta-\frac{1}{2}\tau)y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A+(\beta-\alpha-\frac{1}{2}\tau)y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B+\tau\end{aligned} \tag{9}$$

Setting the partial derivatives of $\mathcal{L}$ with respect to weight τ to zero, we have,

$$\begin{aligned}0=\frac{\partial\mathcal{L}}{\partial\tau}=&\frac{1}{2}(\alpha-\beta-\frac{1}{2}\tau)\parallel \mathbf{x}_t^A \parallel^2+\frac{1}{2}(\beta-\alpha-\frac{1}{2}\tau)\parallel \mathbf{x}_t^B \parallel^2\\ &+1-\frac{1}{2}(y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A+y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B)\end{aligned}$$

$$\Rightarrow \tau=\frac{2}{\parallel \mathbf{x}_t^A \parallel^2+\parallel \mathbf{x}_t^B \parallel^2}\Big((\alpha-\beta)(\parallel \mathbf{x}_t^A \parallel^2-\parallel \mathbf{x}_t^B \parallel^2)+2\ell_t\Big)$$

where the loss $\ell_t=1-\frac{1}{2}(y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A+y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B)$. For the sake of simplicity, we denote,

$$a=\frac{2}{\parallel \mathbf{x}_t^A \parallel^2+\parallel \mathbf{x}_t^B \parallel^2} \qquad \text{and} \qquad b=\parallel \mathbf{x}_t^A \parallel^2\parallel \mathbf{x}_t^B \parallel^2 \tag{10}$$

As mentioned in Equation (8), we have $\tau+\lambda=C$ and $\lambda\geq 0$ so we can conclude that $\tau\leq C$. Now τ can be determined as follows:

$$\tau=\min\Big\{C,a\Big((\alpha-\beta)(\parallel \mathbf{x}_t^A \parallel^2-\parallel \mathbf{x}_t^B \parallel^2)+2\ell_t\Big)\Big\} \tag{11}$$

Substituting (11) into (9), we have,

$$\begin{aligned}\mathcal{L}=&-\frac{1}{2}a\Big((\alpha-\beta)\parallel \mathbf{x}_t^B \parallel^2-\ell_t\Big)^2\parallel \mathbf{x}_t^A \parallel^2-\frac{1}{2}a\Big((\beta-\alpha)\parallel \mathbf{x}_t^A \parallel^2-\ell_t\Big)^2\parallel \mathbf{x}_t^B \parallel^2\\ &+a((\alpha-\beta)\parallel \mathbf{x}_t^B \parallel^2-\ell_t)y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A+a((\beta-\alpha)\parallel \mathbf{x}_t^A \parallel^2-\ell_t)y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B\\ &+a\Big((\alpha-\beta)(\parallel \mathbf{x}_t^A \parallel^2-\parallel \mathbf{x}_t^B \parallel^2)+2\ell_t\Big)\end{aligned} \tag{12}$$

Setting the partial derivatives of $\mathcal{L}$ with respect to weight α to zero, we have,

$$\begin{aligned}0=\frac{\partial\mathcal{L}}{\partial\alpha}&=a\Big((\alpha-\beta)\parallel \mathbf{x}_t^B \parallel^2+\ell_t\Big)b+a\Big((\beta-\alpha)\parallel \mathbf{x}_t^A \parallel^2+\ell_t\Big)b\\ &\quad+a(\parallel \mathbf{x}_t^B \parallel^2 y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A-\parallel \mathbf{x}_t^A \parallel^2 y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B+\parallel \mathbf{x}_t^A \parallel^2-\parallel \mathbf{x}_t^B \parallel^2)\\ &=a\Big((\alpha-\beta)\parallel \mathbf{x}_t^B \parallel^2+\ell_t)\Big)b+a\Big((\beta-\alpha)\parallel \mathbf{x}_t^A \parallel^2+\ell_t)\Big)b\\ &\quad+a(\parallel \mathbf{x}_t^A \parallel^2\ \ell_t^B-\parallel \mathbf{x}_t^B \parallel^2\ \ell_t^A)\end{aligned}$$

where $\ell_t^A=1-y_t\mathbf{w}_t^A\cdot\mathbf{x}_t^A$ and $\ell_t^B=1-y_t\mathbf{w}_t^B\cdot\mathbf{x}_t^B$. We also have $\alpha+\beta=\gamma$. Therefore, we can conclude that

$$\alpha = \frac{\gamma}{2} + \frac{1}{2}\frac{1}{\| \mathbf{x}_t^A \|^2 + \| \mathbf{x}_t^B \|^2}\Big(\frac{\ell_t^B}{\| \mathbf{x}_t^B \|^2} - \frac{\ell_t^A}{\| \mathbf{x}_t^A \|^2}\Big) \tag{13}$$

Similarly, we have

$$\beta = \frac{\gamma}{2} - \frac{1}{2}\frac{1}{\| \mathbf{x}_t^A \|^2 + \| \mathbf{x}_t^B \|^2}\Big(\frac{\ell_t^B}{\| \mathbf{x}_t^B \|^2} - \frac{\ell_t^A}{\| \mathbf{x}_t^A \|^2}\Big) \tag{14}$$

Recall that we have $\alpha \geq 0$, $\beta \geq 0$, and $\alpha+\beta = \gamma$. Hence, we can conclude that $\alpha \leq \gamma$ and $\beta \leq \gamma$. Finally, we obtain our Two-view Passive Aggressive formulation as shown in Algorithm 1. The optimal value of the two tuning parameters C and γ can be estimated via cross validation in practice.

Algorithm 1. Two-view Passive Aggressive Algorithm

```
Input:
  $C$ = positive aggressiveness parameter
  $\gamma$ = positive agreement parameter
Output:
  None
Process:
Initialize $\mathbf{w}_1^A \leftarrow \mathbf{0}$; $\mathbf{w}_1^B \leftarrow \mathbf{0}$;
for $t = 1, 2, \ldots$ do
    Receive instances $\mathbf{x}_t^A \in \mathbb{R}^n$ and $\mathbf{x}_t^B \in \mathbb{R}^m$
    Predict $\hat{y}_t = \text{sign}\frac{1}{2}(\mathbf{w}_t^A \cdot \mathbf{x}_t^A + \mathbf{w}_t^B \cdot \mathbf{x}_t^B)$
    Receive correct label $y_t \in \{-1, +1\}$
    Suffer loss $\ell_t \leftarrow \max\Big\{0, 1 - y_t\frac{1}{2}(\mathbf{w}_t^A \cdot \mathbf{x}_t^A + \mathbf{w}_t^B \cdot \mathbf{x}_t^B)\Big\}$
    if $\ell_t > 0$ then
        Set $\ell_t^A \leftarrow 1 - y_t\mathbf{w}_t^A \cdot \mathbf{x}_t^A$; $\ell_t^B \leftarrow 1 - y_t\mathbf{w}_t^B \cdot \mathbf{x}_t^B$
            $\alpha \leftarrow \max\Big\{0, \min\{\gamma, \frac{1}{2}\Big(\gamma + \frac{\frac{\ell_t^B}{\|\mathbf{x}_t^B\|^2} - \frac{\ell_t^A}{\|\mathbf{x}_t^A\|^2}}{\| \mathbf{x}_t^A \|^2 + \| \mathbf{x}_t^B \|^2}\Big)\}\Big\}$
            $\beta \leftarrow \max\Big\{0, \min\{\gamma, \frac{1}{2}\Big(\gamma - \frac{\frac{\ell_t^B}{\|\mathbf{x}_t^B\|^2} - \frac{\ell_t^A}{\|\mathbf{x}_t^A\|^2}}{\| \mathbf{x}_t^A \|^2 + \| \mathbf{x}_t^B \|^2}\Big)\}\Big\}$
            $\tau_t \leftarrow \min\Big\{C, \frac{(\alpha - \beta)(\| \mathbf{x}_t^A \|^2 - \| \mathbf{x}_t^B \|^2) + 2\ell_t}{\| \mathbf{x}_t^A \|^2 + \| \mathbf{x}_t^B \|^2}\Big\}$
        Update $\mathbf{w}_{t+1}^A \leftarrow \mathbf{w}_t^A - (\alpha - \beta - \frac{1}{2}\tau_t)y_t\mathbf{x}_t^A$
               $\mathbf{w}_{t+1}^B \leftarrow \mathbf{w}_t^B - (\beta - \alpha - \frac{1}{2}\tau_t)y_t\mathbf{x}_t^B$
    end
end
```

4 Performance Evaluation

In this section, we evaluate the online classification performance of our proposed Two-view PA on 3 benchmark datasets, Ads [8], Product Review [9], and WebKB [12]). The *single-view* PA algorithm serves as the baseline. We use a different PA model for each view, naming them *PA View 1* and *PA View 2*. We also concatenate the input feature vectors from each view to form a larger feature set, and report the results. We denote this alternative approach as *PA Cat*. The dataset summary statistics are shown in Table 1. We note that the Ads and WebKB datasets are very imbalanced, which led us to use F-measure instead of accuracy to evaluate the classification performance. To be fair, we choose $C = 0.1$ and $\gamma = 0.5$ for all PA algorithms. All experiments were conducted using 5-fold cross validation.

Table 1. Summary statistics of 3 datasets

	View		Sample Count		
	Name	#Dimension	#Positive	#Negative	#Total
Ads	img & dest url	929	459	2820	3279
	alt & base url	602			
WebKB	page	3000	230	821	1051
	link	1840			
Product Review	lexical	2759	1000	1000	2000
	formal	5			

4.1 View Difference Comparison

At round t, the view difference is defined as $|\mathbf{w}_t^A \cdot \mathbf{x}_t^A - \mathbf{w}_t^B \cdot \mathbf{x}_t^B|$, which shows the difference in prediction output between the two views. Figures 1(a), 2(a), and 3(a) show the view differences for the three datasets, respectively.

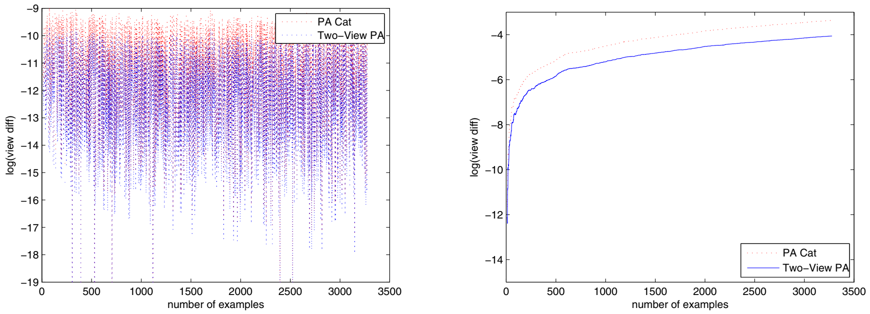
Figures 1(b), 2(b), and 3(b) plot the cumulative view difference at round t, $\frac{1}{T}\sum_{t=1}^{T} |\mathbf{w}_t^A \cdot \mathbf{x}_t^A - \mathbf{w}_t^B \cdot \mathbf{x}_t^B|$. This measures the relationship between the two views as the algorithm adapts to the dataset. The smaller it is, the more related the two views.

Compared to the Product Review and WebKB datasets, the view difference for the Ads dataset varies very much. This means that the agreement between the two views is not stable. As expected, its cumulative view difference turns out to be the largest among the three datasets. Hence, we would expect a classifier based on simple concatenation of the two views to yield poor classification performance. This is in fact confirmed subsequently in the poor PA Cat result for the Ads dataset in Table 2.

On the other end of the spectrum, both the average and cumulative view difference for the WebKB dataset is the smallest. Therefore, one should be able to combine the two views into a single view and just run a simple PA algorithm to obtain a decent classification performance. This hypothesis is confirmed in Table 2, where the PA Cat result outperforms either view by more than 2%.

4.2 Ads Dataset

The Ads dataset was first used by Kushmerick [8] to automatically filter advertisement images from web pages. The Ads dataset comprises more than two views. In this experiment, we only use four views including *image URL view*, *destination URL view*, *base URL view*, and *alt view*. The first and second original views were concatenated as View 1 and the remaining two original views were concatenated as View 2. This dataset has 3279 examples, including 459 positive examples (ads), with the remaining as negative examples (non-ads).



(a) View Difference

(b) Cumulative View Difference

Fig. 1. View Difference of the Ads Dataset

Table 2. F1 measure on 3 datasets

Dataset	PA View 1	PA View 2	PA Cat	Two-view PA
Ads	83.69 ± 3.04	76.01 ± 2.88	81.08 ± 1.99	**85.74** ± 1.97
Product Review	86.46 ± 4.59	69.20 ± 5.20	86.87 ± 3.99	**88.54** ± 1.85
WebKB	92.83 ± 1.72	92.71 ± 3.66	94.97 ± 1.80	**97.50** ± 1.80

The experimental results on the Ads dataset are shown in Table 2, where the F-measure of the proposed algorithm is the best. The Two-view PA performed up to 2% better than the runner-up, PA View 1. As previously discussed, PA View 1 is better than PA Cat since the two views have quite different classification outputs.

4.3 Product Review Dataset

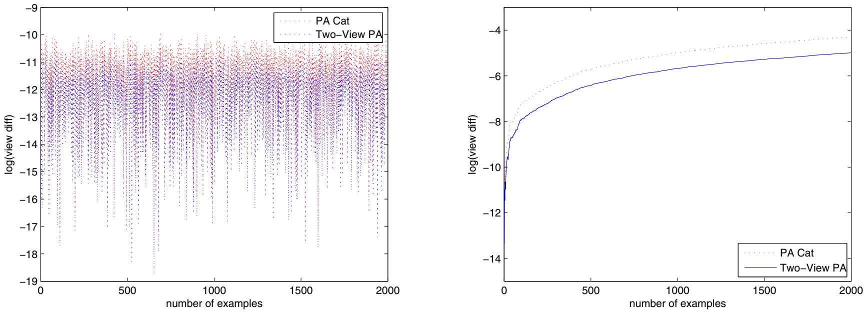
The Product Review dataset is crawled from popular online Chinese cell-phone forums [9]. The dataset has 1000 true reviews and 1000 spam reviews. It consists of two sets of features: one based on review content (*lexical view*) and the other based on extracted characteristics of the review sentences (*formal view*).

The experimental results on this dataset are shown in Table 2. Again, Two-view PA performs better than the other algorithms. The improvement is more than 2% compared with the runner-up. In this dataset, PA Cat performed better than either view alone. This is expected since the view difference between the two views are quite small, as shown in Figure 2.

Moreover, PA Cat is only 0.41% better than the best individual PA View 1. This is because PA Cat does not take into account the view relatedness information. The best performer here is the Two-view PA, which beats the runner-up by almost 2%.

4.4 WebKB Course Dataset

The WebKB course dataset has been frequently used in the empirical study of multi-view learning. It comprises 1051 web pages collected from the computer science departments of four universities. Each page has a class label, course or non-course. The two views of each page are the textual content of a web page (*page view*) and the words that occur in the hyperlinks of other web pages pointing to it (*link view*), respectively. We used a processed version of the WebKB course dataset [12] in our experiment.

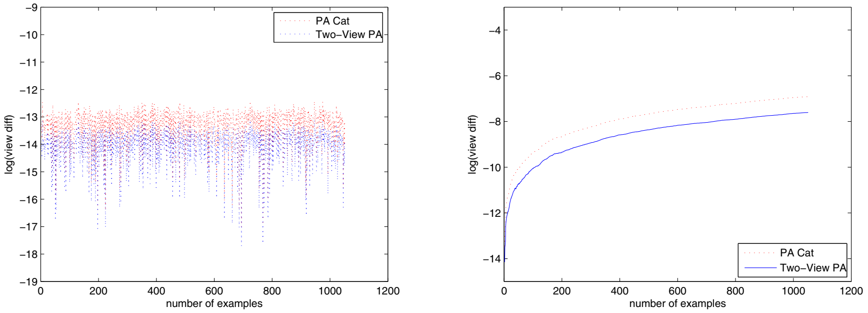


(a) View Difference (b) Cumulative View Difference

Fig. 2. View Difference of the Product Review Dataset

The performance of PA Cat here is also better than the best single view PA. However, the view difference of Two-view PA is much smaller than that of the PA algorithm as shown in Figure 3. Hence, Two-view PA performed more than 3% better than PA Cat, and 5% better than the best individual view PA.

Compared to the Ads and Product Review datasets, the view difference on the WebKB dataset is the smallest. It means that we are able to combine the two views into a single view. Therefore, the PA Cat performance on the WebKB dataset is improved more than 2% compared with the individual view PA.



(a) View Difference

(b) Cumulative View Difference

Fig. 3. View Difference of the WebKB Dataset

5 Conclusion and Open Problems

In this paper, we proposed a hybrid model for two-view passive aggressive algorithm, which is able to take advantage of multiple views of data to achieve an improvement in overall classification performance. We formulate our learning framework into an optimization problem and derive a closed form solution.

There remain some interesting open problems that warrant further investigation. For one, at each round we could adjust the weight of each view so that the better view dominates. In the worst case where the two views are completely related or co-linear, e.g., view 1 is equal to view 2, our Two-view PA degenerates nicely into a single view PA. We would also like to extend Two-view PA to handle multiple views and multiple classes. Formulating a multi-view PA is nontrivial, as it involves defining multi-view relatedness and minimizing (V choose 2) view agreements, for a V-view problem. Formulating a multi-class Two-view PA should be more feasible.

References

1. Block, H.: The perceptron: A model for brain functioning. Rev. Modern Phys. 34, 123–135 (1962)
2. Cesa-Bianchi, N., Conconi, A., Gentile, C.: A second-order perceptron algorithm. Siam J. of Comm. 34 (2005)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
4. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research, 551–585 (2006)
5. Crammer, K., Dredze, M., Kulesza, A.: Multi-class confidence weighted algorithms. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 496–504. Association for Computational Linguistics, Singapore (2009)

6. Dredze, M., Crammer, K., Pereira, F.: Confidence-weighted linear classification. In: ICML 2008: Proceedings of the 25th International Conference on Machine Learning, pp. 264–271. ACM, New York (2008)
7. Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmák, S.: Two view learning: Svm-2k, theory and practice. In: Proceedings of NIPS 2005 (2005)
8. Kushmerick, N.: Learning to remove internet advertisements. In: Proceedings of the Third Annual Conference on Autonomous Agents, AGENTS 1999, pp. 175–181. ACM, New York (1999)
9. Li, G., Hoi, S.C.H., Chang, K.: Two-view transductive support vector machines. In: Proceedings of SDM 2010, pp. 235–244 (2010)
10. Nguyen, T.T., Chang, K., Hui, S.C.: Distribution-aware online classifiers. In: Walsh, T. (ed.) IJCAI, pp. 1427–1432. IJCAI/AAAI (2011)
11. Novikoff, A.: On convergence proofs of perceptrons. In: Proceedings of the Symposium on the Mathematical Theory of Automata, vol. 7, pp. 615–622 (1962)
12. Sindhwani, V., Niyogi, P., Belkin, M.: Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 824–831. ACM, New York (2005)