ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Online multiclass learning with "bandit" feedback under a Passive-Aggressive approach

Hongliang Zhong[1,3], Emmanuel Daucé [2,3] and Liva Ralaivola[1]

1- Laboratoire d'Informatique Fondamentale
CNRS UMR 7279 - Aix-Marseille Université

2- Institut de Neurosciences des Systèmes
INSERM UMR 1106 - Aix-Marseille Université

3- Ecole Centrale Marseille, France

Abstract. This paper presents a new approach to online multi-class learning with bandit feedback. This algorithm, named PAB (Passive Aggressive in Bandit) is a variant of Online Passive-Aggressive Algorithm proposed by [2], the latter being an effective framework for performing max-margin online learning. We analyze some of its operating principles, and show it to provide a good and scalable solution to the bandit classification problem, particularly in the case of a real-world dataset where it outperforms the best existing algorithms.

1 Introduction

Online learning is an effective way to deal with large scale applications, especially applications with streaming data. Online Passive-Aggressive learning (PA) provides a generic framework for online large-margin learning, with many applications [4,5]. PA uses hypotheses from the set of linear predictors. But it only works in the conventional supervised learning paradigm, in which, the learner has access to the true labels of the data after making its prediction. In contrast, the multi-class classification Bandit setting provides a framework where the label information is not fed back to the classifier. The learner only receives binary response telling whether the prediction was correct or not. This paradigm applies to lots of domains, including many web based applications.

There are several classification algorithms that address the bandit setting. The Banditron [1], based on the Perceptron algorithm, is the most "classical" one, having a number of mistakes asymptotically bounded. For the case where the data is linearly separable, the number of mistakes is bounded in $O(\sqrt{T})$ in T rounds. Another bandit algorithm, named "Confidit", was proposed by [3]. In the confidit approach, the bound of the regret (sum of mistakes with respect to the optimal classifier) is improved from of $O(T^{2/3})$ to $O(\sqrt{T}\log T)$. At last the Policy Gradient [7], stemming from the Reinforcement Learning framework, also provides an efficient methodology to deal with the problem [6].

In this paper, we discuss a new algorithm: Passive-Aggressive in Bandit(PAB), i.e. we adapt the PA approach [2] to the bandit setting. With PA's advantage, PAB should in principle perform a max-margin with partial feedback.

In next sections, we will discuss the new bandit algorithm PAB, including its update rules. And we provide some experiments to compare the cumulative

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

loss on synthetic and real-world datasets.

2 Preliminaries

Online learning is applied in a sequence of consecutive rounds. On round t, the learner is given an instance vector $x_t \in \mathbb{R}^d$ and is required to predict a label out of a set of multi-class $[k] = \{1, ..., k\}$. We denote by $\hat{y}_t$ the predicted label . In the general setting, after its prediction, it receives a correct label associated with x_t, which we denote by $y_t \in [k]$. In bandit setting, the feedback is a partial information $\mathbb{I}[\hat{y}_t = y_t]$, where $\mathbb{I}[\hat{y}_t = y_t]$ is 1 if $\hat{y}_t = y_t$, others equals to 0. It's telling whether the prediction is correct or not.

The prediction at round t is chosen by a hypothesis $h_t : \mathbb{R}^d \to [k]$, where h_t is taken from a class of hypothesis $\mathbb{H}$ parameterized by a $k \times d$ matrix of reals w, and is defined to be:

$$h_t = \underset{i \in [k]}{\operatorname{argmax}}\ x_t w_i^T \tag{1}$$

where w_i is the i^{th} row of the matrix $\mathbb{R}^{k*d}$.

Consistently with [2]'s writing, a feature function: $\Phi(x, i)$ is a $k \times d$ matrix which is composed of k features vectors of size d. All rows of $\Phi(x, i)$ are zero except the i^{th} row which is set to x_t. It can be remarked that $< \Phi(x, i), \Phi(x, j) >= \|x\|^2$ if $i = j$ and 0 otherwise.

3 The algorithm Passive-Aggressive in Bandit

In this section, we introduce a new learning algorithm, which is a variant of PA adapted to the bandit setting.

3.1 Online Passive-Aggressive learning algorithm

The goal of online learning is to minimize the cumulative loss for a certain prediction task from the sequentially arriving training samples. PA achieves this goal by updating some parameterized model w in an online manner with the instantaneous losses from arriving data $x_{t,t\geq 0}$ and corresponding responses $y_{t,t\geq 0}$. The losses $l(w; (x_t, y_t))$ can be the hinge loss. PA's update derives its solution from an optimization problem:

$$\underset{w}{min} \frac{1}{2}\|w - w_t\|^2 \quad s.t. l(w; (x_t, y_t)) = 0 \tag{2}$$

Namely, each instance x_t is associated with a single correct label $y_t \in \mathbb{Y}$ and the prediction $\hat{y}_t$ extends by Eq.(1). A prediction mistake occurs if $y_t \neq \hat{y}_t$. With the cost-sensitive setting, there is a cost $\rho(y, y')$ associated with predicting $y' \neq y$ when the correct label is y.

PA incorporates the cost function into the online update. The update in Eq.(2) has the closed form solution,

$$w_{t+1} = w_t + \frac{\langle w_t, (\Phi(x_t, \hat{y}_t) - \Phi(x_t, y_t)\rangle + \rho(\hat{y}_t, y_t)}{\|\Phi(x_t, y_t) - \Phi(x_t, \hat{y}_t)\|^2 + \frac{1}{2C}}(\Phi(x_t, y_t) - \Phi(x_t, \hat{y}_t)), \tag{3}$$

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

where the constant $C > 0$ is defined y the user.

Intuitively, if w_t suffers no loss from the new data, i.e., $l_\rho(w_t; (x_t, y_t)) = 0$, the algorithm passively assigns $w_{t+1} = w_t$; otherwise, it aggressively projects w_t to the feasible zone of parameter vectors that attain zero loss.

3.2 Online algorithm in bandit setting: PAB

We now present the PAB in Algorithm 1, which is an adaptation of PA for the bandit case.

Similar to PA algorithm, at each round the prediction $\hat{y}_t$ is chosen by Bayesian probability according to the current weight matrix w_t, to make a reference to Eq(1). Unlike the conventional learning paradigm, if $\hat{y}_t \neq y_t$, it is difficult to get a PA update because the true labels' information is not supported. So we need to perform an *exploration*, i.e sample a label randomly from [k] with parameter γ[1] and contrast this random prediction with a bandit return $\mathbb{I}(\tilde{y}_t = y_t)$, where $\tilde{y}_t$ is the result of a random draw from a certain distribution $\mathbb{P}(\tilde{Y}|\hat{y})$.

The above intuitive argument is formalized by defining the update matrix $\tilde{U}_t$ to be a function of the random prediction $\tilde{y}_t$. We show in the following that the expectation of the PAB's update is exactly the PA's update.

PAB starts with the initiation of matrix $w_1 = \mathbf{0}$. Its update contains two items:

$$\begin{aligned} w_{t+1} &= w_t + U_{PAB}(x_t, \hat{y}_t, \tilde{y}_t) = w_t + U_{t,1} + U_{t,2} \\ U_{t,1} &= \frac{\mathbb{I}(\tilde{y}_t = y_t)}{\mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)} U_{PA}(x_t, \hat{y}_t, \tilde{y}_t) \\ U_{t,2} &= \frac{\mathbb{I}(\tilde{y}_t = y_t) - \mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)}{\mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)} \cdot \rho_c \frac{\Phi(x_t, \hat{y}_t)}{2 \parallel x_t \parallel^2 + \frac{1}{2C}} \end{aligned} \tag{4}$$

where $U_{PA}(x, \hat{y}, y)$ is the classical passive-agressive update. PAB's update contains two items. The first item is controlled by the indicator $\mathbb{I}(\tilde{y}_t = y_t)$, and is nonzero only when the true label is predicted. The role of second term is to smooth the learning process when few correct labels are available. It means that whenever the process is blind to the true label, the loss is estimated to a fixed number ρ_c; this parameter is chosen empirically.

3.2.1 Simple PAB

A simple choice is $\rho_c = 0$. The item $U_{t,1}$ is very similar to the PA's update. The following lemma is easy to prove:

Lemma 1. *Let $U_{t,1}$ be defined as in eq.(4) and let $U_{PA}(x_t, \hat{y}_t, y_t)$ be defined according to eq.(3). Then, $\mathbb{E}_{\tilde{Y}}[U_{t,1}] = U_{PA}(x_t, \hat{y}_t, y_t)$.*

[1]The parameter γ refers to the definition of [1]

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

3.2.2 Full PAB

The second term $U_{t,2}$ is used to reduce the variance of the update. When $\rho_c > 0$, we need both $\mathbb{E}_{\tilde{Y}}[U_{t,2}] = 0$ (so that $\mathbb{E}_{\tilde{Y}}[U_{PAB}(x_t, \hat{y}_t, \tilde{Y})] = U_{PA}(x_t, \hat{y}_t, y_t))$) and $\mathbb{E}_{\tilde{Y}}[< U_{t,1}, U_{t,2} >] \leq 0$.

Lemma 2. *Let $U_{t,2}$ be defined as in eq.(4), $\mathbb{E}_{\tilde{Y}}[U_{t,2}] = 0$.*

Lemma 3. $\mathbb{E}_{\tilde{Y}}[< U_{t,1}, U_{t,2} >] \leq 0$

The lemma 2 is easy to prove. For the lemma 3, consider:

$$\begin{aligned}\mathbb{E}_{\tilde{Y}}[< U_{t,1}, U_{t,2} >] &= \sum_{i=1}^{k} \mathbb{P}(i|\hat{y}_t)\mathbb{I}(i = y_t)\frac{U_{PA}(\hat{y}_t)}{\mathbb{P}(i|\hat{y}_t)}\frac{\mathbb{I}(i = y_t) - \mathbb{P}(i|\hat{y}_t)}{\mathbb{P}(i|\hat{y})}\frac{\rho_c \Phi(x_t, \hat{y}_t)}{2 \parallel x \parallel^2 + \frac{1}{2C}} \\ &= \frac{1 - \mathbb{P}(y_t|\hat{y}_t)}{\mathbb{P}(y_t|\hat{y}_t)}\rho_c f(x_t, \hat{y}_t, y_t)\end{aligned}$$

with

$$f(x_t, \hat{y}_t, y_t) = \frac{< U_{PA}(x_t, \hat{y}_t, y_t), \Phi(x_t, \hat{y}_t) >}{2 \parallel x \parallel^2 + \frac{1}{2C}}$$

then:

$$\mathbb{E}_{\tilde{Y}}[< U_{t,1}, U_{t,2} >] = \mathbb{I}(\hat{y}_t = y_t)\frac{1 - \mathbb{P}(\tilde{Y} = y_t|\hat{y}_t)}{\mathbb{P}(\tilde{Y} = y_t|\hat{y}_t)}\rho_c f(x_t, \hat{y}_t, y_t) + \mathbb{I}(\hat{y}_t \neq y_t)\frac{1 - \mathbb{P}(\tilde{Y} = y_t|\hat{y}_t)}{\mathbb{P}(\tilde{Y} = y_t|\hat{y}_t)}\rho_c f(x_t, \hat{y}_t, y_t)$$

When $\hat{y}_t = y_t$, $U_{PA} = 0$ so that $f(x_t, \hat{y}_t, y_t) = 0$. When $\hat{y}_t \neq y_t$, it can be shown that $f(x_t, \hat{y}_t, y_t) \leq 0$, so that:

$$\mathbb{E}_{\tilde{Y}}[< U_t^1, U_t^2 >] \leq 0$$

For an appropriate value of ρ_c, the role of $U_{t,2}$ is thus to *reduce* the variance of the PAB update, and thus improve the speed of the learning process.

Algorithm 1: The Passive-Aggressive in Bandit

Input: $w_1 = \vec{0}$.
for $t = 1, 2, ..., T$ **do**
  Receive $x_t \in \mathbb{R}^d$
  Set $\hat{y}_t = \underset{r \in [K]}{\text{argmax}}(W_t \Phi(x_t, r))$
  $\forall i \in [k]$, $\mathbb{P}(\tilde{Y} = i|\hat{y}_t) = \mathbb{I}(\hat{y}_t = i) \cdot \gamma + \frac{(1-\gamma)}{k}$
  Randomly sample $\tilde{y}$ according to $\mathbb{P}(\tilde{Y} = i|\hat{y}_t)$
  Receive the feedback $\mathbb{I}(\tilde{y}_t = y_t)$
  $l_t =< w_t, \Phi(x_t, \hat{y}_t) - \Phi(x_t, y_t) > + \mathbb{I}(\hat{y}_t = y_t)$
  $U_{t,1} = \frac{\mathbb{I}(\tilde{y}_t = y_t)}{\mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)} U_{PA}(x_t, \hat{y}_t, \tilde{y}_t)$
  $U_{t,2} = \frac{\mathbb{I}(\tilde{y}_t = y_t) - \mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)}{\mathbb{P}(\tilde{Y} = \tilde{y}_t|\hat{y}_t)} \cdot \frac{\rho_c}{2\|x_t\|^2 + \frac{1}{2C}} \Phi(x_t, \hat{y}_t)$
  $U_{PAB,t}(x_t, \hat{y}_t, \tilde{y}_t) = U_{t,1} + U_{t,2}$
  Update:$W_{t+1} = W_t + U_{PAB,t}(x_t, \hat{y}_t, \tilde{y}_t)$
end

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

4 Experiments

In this section, we present experimental results for the PAB and other bandit algorithms on two synthetic and one real world data sets. The cumulative loss is presented for each data set.

The first data set, denoted by SynSep, is a 9-class, 400-dimensional synthetic data set of size 10^5. The method to generate the sample is found in [1]. The second data set, denoted by SynNonSep, is constructed in the same way as SynSep except that a 5% label noise is added, which makes the data set non-separable. The third data set is collected from the Reuters RCV1 collection. This set is made of 47236-dimensional vectors, contains 4 classes, and has a size of 10^5 .

Fig.1 gives the cumulative loss obtained on the 3 datasets for different online learning algorithms. On the SynSep data set, the Confidit algorithm provides the best results, but the basic PAB is second best. Three out of five algorithms attain a zero loss. The worst in that case is the Banditron, but the full PAB also fails to reach a final zero loss. On the SynNonSep data set, the results are rather poor in general. The Confidit and Policy gradient obtain the best performances, with a stable final error rate around 5%. On the Reuters data, in contrast with the synthetic datasets, the full PAB overtakes the other methods, with a final error rate around 2.5% while the other algorithms attain 5% error, and even worse in the case of Confidit (8% error). Besides, the PAB error rate is constantly reducing during the learning process.

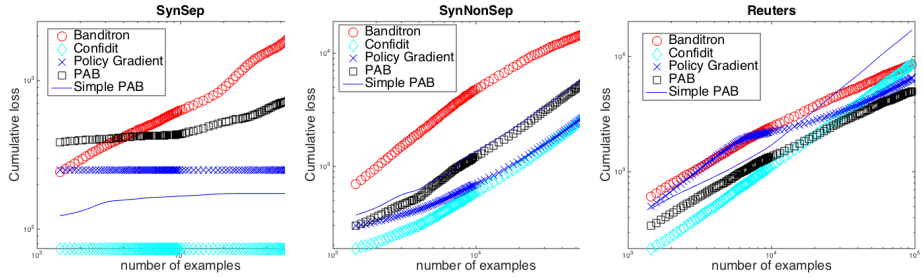

Fig. 1: Cumulative loss of Banditron, Policy Gradient, Confidit, Full PAB and Simple PAB on the SynSep, SynNonSep and Reuters data sets. The parameters are $\gamma = 0.014$ for Banditron; $\alpha = 1, \eta = 1000$ for Confidit; $\eta = 0.01, \lambda = 0.001$ for Policy Gradient; C= 0.001, $\gamma = 0.7$, $\rho = 0$ for Simple PAB and $\rho = 1$ for Full PAB; all these parameters are under the SynSep data; $\gamma = 0.006$ for Banditron; $\alpha = 1, \eta = 1000$ for Confidit; $\eta = 0.01, \lambda = 0.001$ for Policy Gradient; C=0.00001, $\gamma = 0.7$ for PAB under the SynNonSep data and the last parameters are under the Reuters data: $\gamma = 0.05$ for Banditron; $\alpha = 1, \eta = 100$ for Confidit; $\eta = 0.1, \lambda = 0.001$ for Policy Gradient; C=0.0001, $\gamma = 0.6$ is for the PAB.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

5 Conclusion

With the advantage of the Passive-Aggressive max-margin principle, the simple and full PAB appear effective to address the bandit online learning setting. Their first advantage is their linear complexity in space that allows to treat high dimensional datasets on the contrary to second-order methods. On separable data samples, the basic PAB overtakes most of the other approaches, at the exception of the Confidit algorithm, with a much lower complexity. It is however found to perform rather poorly on noisy and real world datasets. In contrast, the full PAB is expected to vary more smoothly over time, and is found to perform particularly well on the Reuters dataset. In that case, Confidit and Policy Gradient seem to fall in a local stable solution, while the full PAB constantly improves, issuing a better classifier.

However, the performance of the algorithm is found to depend on three free parameters γ, ρ_c and C. In order to avoid fastidious cross-validation, additional investigation is needed in order to find analytic estimates of their optimal values. Additional work on regret bounds is also needed in order to analytically compare our approach with the other ones.

ACKNOWLEDGEMENT

This work is partially supported by the ANR-funded projet GRETA – Greediness: theory and algorithms (ANR-12-BS02-004-01).

References

[1] Sham M.Kakade, Shai Shalev-Shwartz, Ambuj Tewari. Efficient Bandit Algorithms for Online Multiclass Prediction. In*Proceedings of the 25th international conference on Machine learning.*ACM,2008 .

[2] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer. Online Passive-Aggressive Algorithms. In *Journal of Machine Learning Research*, 2006 .

[3] Koby Crammer, Claudio Gentile. Multiclass Classification with Bandit Feedback using Adaptive Regularization. In*The 28th International Conference Machine learning.*ICML,2011.

[4] Ryan McDonald, Koby Crammer, Fernando Pereira. Online Large-Margin Training of Dependency Parsers. In*ACL*,2005.

[5] David Chiang, Yuval Marton, Philip Resnik. Online Large-Margin Training of Syntactic and Structural Translation Features. In *EMNLP*,2008.

[6] Emmanuel Dauce, Timothee Proix, Liva Ralaivola. Fast online adaptivity with policy gradient: example of the BCI "P300" speller. Proc. of the 21th European Symposium on Artificial Neural Networks, computational intelligence and machine learning(ESANN 2013), Verleysen, M.ed: 197-202, April 24-26, Bruges,Belgium, 2013.

[7] R.J. Williams. Simple Statistical Gradient Following Algorithms for Connectionnist Reinforcement Learning. *Machine Learning 8*: 229-256, 1992.