Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

One-pass online learning: A local approach

Zhaoze Zhou [a,b], Wei-Shi Zheng [a,f,*], Jian-Fang Hu [c], Yong Xu [d], Jane You [e]

[a] School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China
[b] Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha 410073, China
[c] School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, China
[d] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[e] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[f] Guangdong Provincial Key Laboratory of Computational Science, China

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form
25 May 2015
Accepted 1 September 2015

Keywords:
One-pass online learning
Local modeling
Classification

ABSTRACT

Online learning is very important for processing sequential data and helps alleviate the computation burden on large scale data as well. Especially, one-pass online learning is to predict a new coming sample's label and update the model based on the prediction, where each coming sample is used only once and never stored. So far, existing one-pass online learning methods are globally modeled and do not take the local structure of the data distribution into consideration, which is a significant factor of handling the nonlinear data separation case. In this work, we propose a local online learning (LOL) method, a multiple hyperplane Passive Aggressive algorithm integrated with online clustering, so that all local hyperplanes are learned jointly and working cooperatively. This is achieved by formulating a common component as information traffic among multiple hyperplanes in LOL. A joint optimization algorithm is proposed and theoretical analysis on the cumulative error is also provided. Extensive experiments on 11 datasets show that LOL can learn a nonlinear decision boundary, overall achieving notably better performance without using any kernel modeling and second order modeling.

1. Introduction

We are concerning the one-pass online learning without keeping any tracks of passed samples. The key idea is to update current model by retaining the new learned model close to the current one and meanwhile imposing a margin separation on the most recent sample. After prediction and updating current model, the sample is abandoned, which means it could not be used for training again. Therefore, one-pass training reduces the consuming of memory so greatly that it is very practical in some circumstances. For example, a closed-circuit television camera with very limited resources is allowed to learn from video stream in one-pass manner to enhance its performance of recognition. Hence, one-pass online learning reduces the burden of the learning system and makes machine learning models more applicable and flexible.

One of the most widely known one-pass online learning methods is the first-order Passive-Aggressive (PA) method [1], which updates a marginal classifier according to the feedback of the prediction of each sequential data point. In order to explicitly consider the uncertainty of weights of linear classifier, confidence-weighted (CW) learning [2] as well as its variants soft confidence-weighted (SCW-I, SCW-II) [3], adaptive regularization of weight vectors (AROW) [4] have been recently investigated. However, the Passive-Aggressive and its related confidence-weighted learning methods still assume that samples are almost linearly separable, which is not always true since data points are always nonlinearly separable in the original input space.

To address the nonlinear separation problem in online learning, there are indeed some nonlinear algorithms, but not all of them are one-pass based. These algorithms include direct application of kernel trick on online linear classifiers [5,6], budget-based online models [7–10], and kernel approximation mapping based models [11] by using random Fourier features or Nyström method. However, limitations exist in these methods, including (1) memory overflow after processing a large amount of data due to keeping historical wrong classified samples as support vectors (SVs) in [5,6], (2) large computational burden caused by processing on SVs in the budget in [7–10], and (3) data-independence and not adaptation to data stream in [11].

In offline learning, besides kernelizing linear classification models to solve the linearly non-separable problems, local classifiers have also been investigated recently to assign data samples to

* Corresponding author.
E-mail addresses: zhouzhaoze@gmail.com (Z. Zhou), wszheng@ieee.org (W.-S. Zheng), hujianf@mail2.sysu.edu.cn (J.-F. Hu), yongxu@ymail.com (Y. Xu), csyjia@comp.polyu.edu.hk (J. You).

http://dx.doi.org/10.1016/j.patcog.2015.09.003
0031-3203/

a set of prototypes and then infer the weights for model combination of local classifiers. Locally linear support vector machine (LLSVM) [12], and local deep kernel learning (LDKL) [13] were proposed to solve non-linear classification tasks using aweighted combination of multiple local hyperplanes, whose weights can be determined by local coding. As an extreme, the 1-nearest prototype classifier (1-NN) [14] combines prototype learning and nearest neighbor search to perform classification. Local classifiers has better adaptability to various types of data distribution than kernel classifiers, and the advantage of combining local classifiers is to avoid kernel modeling, so as to avoid computational expensive for large scale data. However, existing works mentioned above are not specifically designed for online learning tasks. Hence how to derive online local learning and how local online approach works for on-the-fly classification are still unknown.

In this paper, we propose a novel online approach by jointly learning multiple local hyperplanes to nonlinearly process sequential data in an one-pass manner. In particular, we extend the single hyperplane Passive-Aggressive method to a multiple local hyperplanes one. All local hyperplanes will be connected by a common component and optimized simultaneously. In our modeling, the local specific components of hyperplanes allow our model to make more accurate prediction locally (i.e. sensitive to a probe data point), while the common component shared by these local hyperplanes alleviates the over-fitting caused by the local information. A novel optimization algorithm is proposed and the theoretical relationship between the single and multiple hyperplane Passive Aggressive algorithms is derived from the aspect of cumulative error. We call our proposed model the *local online learning* (LOL) method. Our method achieves notable better performance on various tasks, especially on multi-class classification tasks, such as Multi-PIE Identity recognition of 249 subjects, achieving more than 95% accuracy and at least 13% higher than the compared methods. The details will be discussed in the experiment section (Section 5).

It is indeed that there are related online developments in pattern recognition such as online tracking [15], online face recognition [16,17], incremental object matching [18], and online action recognition [19]. There are also works on using sequential algorithm to learn a classifier when only limited source is available for computation such as [20]. However, these models are not generally designed to make themselves be applicable for other applications, or they only conduct the learning when the size of dataset is increasing but are not one-pass based suitable for locality sensitive classification. Compared to these work, we aim to learn an online classifier for general purpose and make it suitable for one-pass online learning.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related one-pass online learning algorithms. Sections 3 and 4 detail the proposed local online learning algorithms and the corresponding theoretical analysis. Experimental results are presented in Section 5. Finally we draw the conclusion in Section 6.

2. Related work

2.1. Passive Aggressive algorithm (PA)

Like Perceptron [21], the first order Passive-Aggressive algorithm (PA)[1] focuses on learning linear classification model for each new sample $\mathbf{x} \in \mathbb{R}^d$, formed as

$$g(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x}), \tag{1}$$

where sign function outputs the prediction label (-1 or $+1$) of the input and $\mathbf{w}$ is a weight vector. Passive-Aggressive method aims to use the pre-learned globally linear hyperplane to guide the prediction of new observed sample $\mathbf{x}_t$ labeled with $y_t \in \{-1, 1\}$ at time step t by solving the following problem:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|^2 \quad \text{s.t. } \ell^{pa}(\mathbf{w};(\mathbf{x}_t,y_t)) = 0 \tag{2}$$

where $\mathbf{w}_t$ is the model/hyperplane at time step t before update. Here, ℓ^{pa} is the hinge loss function, i.e. $\ell^{pa}(\mathbf{w};(\mathbf{x}_t,y_t)) = \max\{0, 1-y_t \cdot \mathbf{w}^T\mathbf{x}_t\}$. In practice, a non-negative slack variable ξ is introduced so that the above optimization problem becomes

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|^2 + C\xi, \quad \text{s.t. } \ell^{pa}(\mathbf{w};(\mathbf{x}_t,y_t)) \le \xi \quad \text{and} \quad \xi \ge 0 \tag{3}$$

where C is a positive parameter. This optimization problem can be understood as searching a new optimal hyperplane that does not differ from the current one too much in order to meet the loss constraint with respect to a new input sample. The solution of the problem (3) is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta_t y_t \mathbf{x}_t, \quad \eta_t = \min\left\{C, \frac{\ell^{pa}}{\|\mathbf{x}_t\|^2}\right\} \tag{4}$$

where η_t is the learning rate. Obviously, Passive-Aggressive method learns a globally linear decision function without considering the local distribution of data.

2.2. Confidence weighted online learning algorithm family

Confidence weighted algorithm (CW) [2], soft confidence weighted algorithms (SCW-I, SCW-II) [3] and adaptive regularization of weight vectors (AROW) [4] were proposed to explore the underlying structure of features. Confidence-weighted learning is actually inspired by Passive-Aggressive learning but holds a Gaussian distribution assumption over the weights. Confidence weighted algorithm assumes that a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is imposed on linear weights . The distribution is obtained by minimizing the Kullback–Leiber divergence between the new distribution (parameterized by $\mathcal{N}(\mu_{t+1}, \boldsymbol{\Sigma}_{t+1})$) and the old one (parameterized by $\mathcal{N}(\mu_t, \boldsymbol{\Sigma}_t)$) in a Passive-Aggressive way, forcing the probability of accurate prediction of current sample $\mathbf{x}_t$ greater than a threshold η below

$$(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}) = \arg\min_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} D_{KL}(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)) \quad \text{s.t. } Pr_{\mathbf{w}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[y_t \cdot (\mathbf{w}^T\mathbf{x}_t) \ge 0] \ge \eta \tag{5}$$

The problem (5) has a solution of the following form:

$$\mu_{t+1} = \mu_t + \alpha_t y_t \boldsymbol{\Sigma}_t \mathbf{x}_t, \quad \boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \beta_t \boldsymbol{\Sigma}_t \mathbf{x}_t^T \mathbf{x}_t \boldsymbol{\Sigma}_t \tag{6}$$

Soft confidence-weighted (SCW-I, SCW-II) [3] and adaptive regularization of weight vectors (AROW) [4] are extensions of confidence weighted algorithm [2], sharing the same update form but with different rules to learn the coefficients. These methods model the globally linear weight with a distribution and thus introduce variations of linear hyperplanes into the modeling. However, this modeling is still globally linear.

2.3. Online kernel (approximation) algorithms

Inspired by the successful application of kernel tricks [22] to enhance the linear classifiers, some researchers intend to employ kernel tricks for online linear classifier learning. Correspondingly, a lot of kernel approximation methods have been developed to reduce the computational complexity caused by the employment of kernel tricks [23,24]. However, these method need to keep all historical wrong classified samples as support vectors (SVs).

Please cite this article as: Z. Zhou, et al., One-pass online learning: A local approach, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.09.003