Smooth Trend

Peng Zhao
zhaop@lamda.nju.edu.cn
December 30, 2018

Abstract

if there is abstract...

1 Introduction

In real-life applications, data are often accumulated with time, and are inherently evolving in nature. The evolving nature, if simply ignored, will largely do harm to the performance of the learning system. Therefore, it is quite crucial to develop the approaches which are able to adapt the evolving environments.

We propose ATF (Adapting evolving data stream by Trend Filtering) to alleviate and adapt the

2 Related Work

How to effectively learn anThe evolving data strea

3 The Proposed Approach

3.1 Problem Statement

We consider the problem of learning from the data stream, where the data are coming in a batch mode. Specifically, in each time t, there comes a chunk of data $S_t = \{(\mathbf{x}_t^{(1)}, y_t^{(1)}), \ldots, (\mathbf{x}_t^{(m_t)}, y_t^{(m_t)})\}$. Each sample is denoted as $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are feature and label space, respectively.

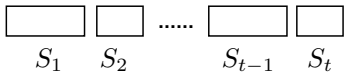

Figure 1: Illustration of data stream.

3.2 ATF Approach

We denote the overall model at time stamp t as $\mathbf{w}_t$, and decompose it into the following form,

$$\mathbf{w}_t \approx \mathbf{u}_t + \mathbf{v}_t, \tag{1}$$

where we can view $\mathbf{u}_t$ as the global trend, $\mathbf{v}_t$ as the local model, and $\boldsymbol{\epsilon}_t$ as the noise term. Essentially, (1) implies a decomposition of the final prediction, which consists of three terms, i.e., global trend prediction, local model prediction and the noise. That is,

$$\mathbf{w}_t^{\mathrm{T}} \psi(\mathbf{x}) = \mathbf{u}_t^{\mathrm{T}} \psi(\mathbf{x}) + \mathbf{v}_t^{\mathrm{T}} \psi(\mathbf{x}) + \epsilon_t$$

Therefore, it is desired to not only learn the overall model, but also consider to learn the global trend. We propose ATF approach, with the following general objective function, to learn the final model and the global trend jointly,

$$(\mathbf{w}_t, \mathbf{u}_t) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda_1 \Omega(\mathbf{w}, \mathbf{u}) + \lambda_2 \mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}) \Big\}. \quad (2)$$

In the objective function (2), $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function, and $\psi : \mathcal{X} \to \mathcal{X}'$ is any non-linear mapping on the feature space. Throughout the paper, we will adopt the linear mapping and squared hinge loss for simplicity, that is, $\psi(\mathbf{x}) = \mathbf{x}$ and $\ell(\widehat{y}, y) = \max\{0, (1 - \widehat{y}y)\}^2$. Besides, $\Omega(\cdot, \cdot)$ is the *adjustment* regularizer, which essentially depicts the bias of the local model $\mathbf{w}$ with respect to the global trend $\mathbf{u}$. Besides, $\mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2})$ is a *trend* regularizer, enforcing the global trend $\mathbf{u}$ is smooth.

In the following, we choose the trend regularizer as $\mathcal{S}(\mathbf{u}, \mathbf{u}_{t-1}, \mathbf{u}_{t-2}) = \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|_2^2$, and choose the adjustment regularizer as the squared ℓ_2 norm, that is, $\Omega(\mathbf{w}, \mathbf{u}) = \|\mathbf{w} - \mathbf{u}\|_2^2$. Then, we can specify (2) in the following form,

$$(\mathbf{w}_t, \mathbf{u}_t) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda_1 \|\mathbf{w} - \mathbf{u}\|_2^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|_2^2 \Big\}. \quad (3)$$

4 Optimization

In this section, we first develop optimization method based on first order gradient descent to address the target optimization problem (3). Then, we provide an explanation of the target function from the perspective of alternative optimization.

4.1 Nesterov's Accelerated Gradient Descent

To simplify the notations, we omitted the subscript t, specifically, using m instead of m_t, and $\mathbf{w}$ ($\mathbf{u}$) instead of $\mathbf{w}_t$ ($\mathbf{u}_t$), besides, we use $\mathbf{x}_i$ (y_i) instead of $\mathbf{x}_t^{(i)}$ ($y_t^{(i)}$), with a slight abuse of notations. The simplified version of target optimization problem becomes,

$$(\mathbf{w}^*, \mathbf{u}^*) = \arg\min_{\mathbf{w},\mathbf{u}} \Big\{ \sum_{i=1}^{m} \ell(\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_i), y_i) + \lambda_1 \|\mathbf{w} - \mathbf{u}\|_2^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|_2^2 \Big\}. \quad (4)$$

Since the objective function is jointly convex and differentiable in terms of both model term $\mathbf{w}$ and the trend term $\mathbf{u}$. Therefore, we will adopt *Nesterov's accelerated gradient descent* [Nesterov, 2004; Bubeck, 2015] to solve this problem.

To simplify the presentation, we introduce the notation $\mathbf{s} = [\mathbf{w}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2d}$ as the auxiliary variable, and immediately, we have

$$\mathbf{w} = \mathbf{s}^{\mathrm{T}} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{u} = \mathbf{s}^{\mathrm{T}} \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix},$$

where $\mathbf{0}, \mathbf{1} \in \mathbb{R}^d$ are the d-dimensional zero and one column vector, respectively. Therefore, (3) is equivalent to the following optimization with respect to $\mathbf{s}$:

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \{ \sum_{i=1}^{m} \ell(\begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}^{\mathrm{T}} \mathbf{s}, \mathbf{x}_i, y_i) + \lambda_1 \| \begin{bmatrix} -\mathbf{1} \\ \mathbf{1} \end{bmatrix}^{\mathrm{T}} \mathbf{s} \|_2^2 + \lambda_2 \| \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}^{\mathrm{T}} \mathbf{s} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1} \|_2^2 \}. \quad (5)$$

Theorem 1. *Assume the loss function $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is convex, ATF algorithm will finally convergent to the global minimum of* (3)*. Moreover, let $\mathbf{w}_t^k$ and $\mathbf{u}_t^k$ be results returned in the k-th iteration, and $\mathbf{w}_t^*$ and $\mathbf{u}_t^*$ be the global minimum, then it reaches $\mathcal{L}(\mathbf{w}_t^k, \mathbf{u}_t^k) - \mathcal{L}(\mathbf{w}_t^*, \mathbf{u}_t^*) \leq \epsilon$ after $O(1/\sqrt{\epsilon})$ iterations.*

Proof. The objective function is convex with respect to $\mathbf{s}$ (a.k.a., jointly in terms of variables $\mathbf{w}$ and $\mathbf{u}$). Therefore, we are able to utilize the celebrated Nesterov's accelerated method [Nesterov, 2004].
<<< Peng: Cheng, please add the convergence analysis of the Nesterov's accelerated method (or Momentum) here.
Of course you can write an analysis independent with notations in your own note, however, please be sure to make your notations consistent with the contexts in the draft. >>> □

4.2 An Explanation from Alternative Optimization

In fact, we can explain the formulation (3) from the alternative optimization perspective. Specifically, we adopt the alternative optimization to solve $\mathbf{w}$ and $\mathbf{u}$ instead of directly optimizing the concatenation variable $\mathbf{s}$. That is,

(i) Fix $\mathbf{w}$ to update $\mathbf{u}$: global trend filtering.

(ii) Fix $\mathbf{u}$ to update $\mathbf{w}$: local model adjustment.

Global Trend Filtering. When updating the trend term $\mathbf{u}$ with fixed model $\mathbf{w}^{(k)}$, the sub-optimization problem is,

$$\mathbf{u}^{(k)} = \arg\min_{\mathbf{u}} \lambda_1 \|\mathbf{w}^{(k)} - \mathbf{u}\|_2^2 + \lambda_2 \|\mathbf{u} + \mathbf{u}_{t-2} - 2\mathbf{u}_{t-1}\|_2^2. \tag{6}$$

Essentially, the global trend $\mathbf{u}_t$ is constrained to, on one hand, be smooth in terms of the previous trend terms (i.e., $\mathbf{u}_{t-1}$ and $\mathbf{u}_{t-2}$); and on the other hand, have a small error with the overall model (i.e., $\mathbf{w}_t$).

The problem (6) admits a close-form solution as follows,

$$\mathbf{u}^{(k)} = \frac{\lambda_1}{\lambda_1 + \lambda_2}\mathbf{w}^{(k)} + \frac{\lambda_2}{\lambda_1 + \lambda_2}(2\mathbf{u}_{t-1} - \mathbf{u}_{t-2}).$$

Local Model Adjustment. When updating the model $\mathbf{w}$ with fixed trend term $\mathbf{u}^{(k)}$, the sub-optimization problem can be written as follows,

$$\mathbf{w}^{(k+1)} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m_t} \ell(\mathbf{w}, \mathbf{x}_i, y_i) + \lambda_1 \|\mathbf{w} - \mathbf{u}^{(k)}\|_2^2. \tag{7}$$

Problem (7) aims to minimize the empirical loss over the data in the current chunk and the biased regularization term, which is biased on the global trend $\mathbf{u}^{(k)}$. In essence, the model is locally adjusted by, on one hand, the current data chunk; and on the other hand, reusing the previous model (a.k.a. the global trend).

Since the loss function and the regularizer are both convex and continuous in (7), we can use the Nesterov's accelerated gradient descent to solve this sub-problem. Moreover, if we adopt the loss function as square loss, then (7) admits a close-form solution, which essentially is the biased LS-SVM [Schölkopf *et al.*, 2001; Suykens *et al.*, 2002].

For the local model adjustment phase, we have the following generalization bound.

Theorem 2. *Assume that the loss function $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is convex, non-negative and bounded by a constant $M > 0$, and L-lipschitz continuous. Let $\widehat{\mathbf{w}}$ be the model returned by the local model adjustment of ATF algorithm, based on the trend term $\mathbf{u}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\begin{aligned} R(h_k) - \widehat{R}(h_k) = O\,(\, & \frac{1}{\sqrt{m}} \left(\sqrt{R_p} + \sqrt{\frac{R_p}{\lambda\mu}} + \sqrt[4]{\frac{R_p}{\lambda\mu m}} \right) \\ & + \frac{1}{m} \left(\sqrt{\frac{1}{\lambda}} + \sqrt[4]{\frac{1}{\lambda}} \right)\,), \end{aligned} \tag{8}$$

5 Experiments

In this section, we provide the empirical performance on both synthetic and real-world datasets to validate the effectiveness of our approach. Moreover, we also report the parameter studies.

5.1 Synthetic Datasets

We compare the proposed approach ATF_1 and ATF_2 with TIX [Forman, 2006], Learn^{++}.NSE [Elwell and Polikar, 2011], and DTEL [Sun *et al.*, 2018].

5.2 Real-World Datasets

Table 1: Basic statistics of concept drift datasets, along with the information of data chunk.

Dataset	#instance	#dim	#class	#chunk	Chunk Size
Usenet-1	1,500	100	2		
Usenet-2	1,500	100	2		
Luxembourg	1,900	32	2		
Spam	9,324	500	2		
Email	1,500	913	2		
Weather	18,159	8	2		
GasSensor	4,450	129	6		
Powersupply	29,928	2	2		
Electricity	45,312	8	2		
Covertype	581,012	54	2		

5.3 Parameter Studies

In our approach, there are two hyper-parameters, i.e., adjustment coefficient λ_1 and smoothness coefficient λ_2, which play an crucial role in trading off empirical risk minimization, global trend learning and local model adjustments. In this paragraph, we will report the empirical performance of our approach with different coefficient values.

5.4 The Effectiveness of Model Reuse

One of the baselines is to directly learn a model on the current data batch, and then use this model to predict on the future data batch.

Competitor 1 (ERM)**.**

$$\mathbf{w}_t^{\text{ERM}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda \|\mathbf{w}\|_2^2 \Big\}. \tag{9}$$

5.5 The Effectiveness of Smooth Trend Filtering

To validate the effectiveness of smooth trend filtering, we compare our proposed approach with the following two competitors.

Competitor 2 (Model Reuse)**.**

$$\mathbf{w}_t^{\text{MR}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda \|\mathbf{w} - \mathbf{w}_{t-1}\|_2^2 \Big\}. \tag{10}$$

Competitor 3 (Smooth Model Reuse)**.**

$$\mathbf{w}_t^{\text{SMR}} = \arg\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m_t} \ell(\mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}_t^{(i)}), y_t^{(i)}) + \lambda \|\mathbf{w} + \mathbf{w}_{t-2} - 2\mathbf{w}_{t-1}\|_2^2 \Big\}. \tag{11}$$

6 Discussion

References

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

George Forman. Tackling concept drift by temporal inductive transfer. In *SIGIR*, pages 252–259, 2006.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Bernhard Schölkopf, Ralf Herbrich, and Alexander J. Smola. A generalized representer theorem. In *COLT*, pages 416–426, 2001.

Yu Sun, Ke Tang, Zexuan Zhu, and Xin Yao. Concept drift adaptation by exploiting historical knowledge. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4822–4832, 2018.

Johan AK Suykens, Tony Van Gestel, and Jos De Brabanter. *Least squares support vector machines*. World Scientific, 2002.