Neurocomputing 198 (2016) 135–147

Contents lists available at ScienceDirect

Neurocomputing

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

Reweighted stochastic learning

Vilen Jumutc *, Johan A.K. Suykens

KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
7 July 2015
Accepted 16 August 2015
Available online 10 March 2016

Keywords:
Linear SVMs
Stochastic learning
l_0 penalty
Sparsity
Regularization

ABSTRACT

Recent advances in stochastic learning, such as dual averaging schemes for proximal subgradient-based methods and simple but theoretically well-grounded solvers for linear Support Vector Machines (SVMs), revealed an ongoing interest in making these approaches consistent, robust and tailored towards sparsity inducing norms. In this paper we study reweighted schemes for stochastic learning (specifically in the context of classification problems) based on linear SVMs and dual averaging methods with primal–dual iterate updates. All these methods favor properties of a convex and composite optimization objective. The latter consists of a convex regularization term and loss function with Lipschitz continuous sub-gradients, *e.g.* l_1-norm ball together with hinge loss. Some approaches approximate in a limit the l_0-type of a penalty. In our analysis we focus on a regret and convergence criteria of such an approximation. We derive our results in terms of a sequence of convex and strongly convex optimization objectives. These objectives are obtained via the smoothing of a generic sub-differential and possibly non-smooth composite function by the global proximal operator. We report an extended evaluation and comparison of the reweighted schemes against different state-of-the-art techniques and solvers for linear SVMs. Our experimental study indicates the usefulness of the proposed methods for obtaining sparser and better solutions. We show that reweighted schemes can outperform state-of-the-art traditional approaches in terms of generalization error as well.

1. Introduction

In many domains dealing with online and stochastic learning, the input instances are of very high dimension, yet within any particular instance several features are non-zero. Therefore specific stochastic and online approaches crafted with sparsity inducing regularization are of particular interest for many machine learning researchers and practitioners. This paper investigates an interplay between Regularized Dual Averaging (RDA) approaches [1] (along with other techniques for solving linear SVMs in the context of stochastic learning [2]) and parsimony concepts arising from the application of sparsity inducing norms, like the l_0-type of a penalty.

One can see an increasing importance of correctly identified sparsity patterns and proliferation of proximal and soft-thresholding subgradient-based methods [1,3,4]. There are many important contributions of the parsimony concept to the machine learning field. One may allude to the understanding of the obtained solution and simplified or easy to extract decision rules [5–7]. On the other hand the informativeness of the obtained features might be useful for a better generalization on unseen data [5]. Approaches based on l_1-regularized loss minimization were studied in the context of stochastic and online learning by several research groups [1,3,8,9] but we are not aware of any l_0-norm inducing methods which were applied in the context of Regularized Dual Averaging and stochastic optimization.

In this paper we are trying to provide a supplementary analysis and sufficient regret bounds for learning sparser linear Regularized Dual Averaging (RDA) [1] models from random observations. We extend and modify our previous research [10,11] and present complementary proofs with fewer assumptions and discussion for the reported theoretical findings. We use sequences of (strongly) convex reweighted optimization objectives to accomplish this goal.

This paper is structured as follows. Section 2 describes previous work on l_0-norm induced learning and some existing solutions to stochastic optimization with regularized loss. Section 3.1 presents a problem statement for the reweighted algorithms. Sections 3.2 and 3.5 introduce our reweighted l_1-RDA and l_2-RDA methods respectively while Section 3.8 presents completely novel approach based on probabilistic reweighted Pegasos-like linear SVM solver. Sections 3.4 and 3.7 provide a theoretic background for our reweighted RDA approaches. Section 4 presents our numerical results and Section 5 concludes the paper.

* Corresponding author.
E-mail addresses: vilen.jumutc@esat.kuleuven.be (V. Jumutc), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

http://dx.doi.org/10.1016/j.neucom.2015.08.123
0925-2312/

2. Related work

Learning with $\|w\|_0$ pseudonorm regularization is a NP-hard problem [12] and can be approached via the reweighting schemes [13–16] while lacking a proper theoretical analysis of convergence in the online and stochastic learning cases. Some methods, like [17], consider an embedded approach where one has to solve a sequence of QP-problems, which might be very computationally- and memory-wise expensive while still missing some proper convergence criteria.

In many existing iterative reweighting schemes [14,18] the analysis is provided in terms of the Restricted Isometry Property (RIP) or the Null Space Property (NSP) [19,14]. These approaches solely rely on the properties which are difficult to access beforehand in a data-driven fashion. This might be crucial if one decides to evaluate methods for their potential applicability. For instance in case of the Restricted Isometry Property, which is characterizing matrix Φ, one is interested to find a constant $\delta \in (0,1)$ such that for each vector w we would have:

$$(1-\delta)\|w\|_2 \leq \|\Phi w\|_2 \leq (1+\delta)\|w\|_2.$$

The RIP was introduced by Candes and Tao [20] in their study of compressed sensing and l_1-minimization. But it cannot be directly applied in the context of online and stochastic optimization because we cannot observe matrix Φ immediately. This fact directly impedes the successful implication of convergence guarantees based on the RIP or other related properties.

Other groups stemmed their research from the follow-the-regularized-leader (FTRL) family of algorithms [1,9,21] and complementary analysis for sparsity-induced learning. In primal–dual subgradient methods arising from this family of algorithms one aims at making a prediction $w_t \in \mathbb{R}^d$ on round t using the average subgradient of the loss function. The update encompasses a trade-off between a gradient-dependent linear term, the regularizer $\psi(w_t)$ and a strongly convex term h_t for well-conditioned predictions. Our research is based on FTLR algorithms with primal–dual iterate updates, such as RDA [1], and corresponding theoretical guarantees are very much along the lines of the latter.

3. Reweighted methods

3.1. Problem statement

In the stochastic Regularized Dual Averaging approach developed by Xiao [1] one approximates the loss function $f(w)$ by using a finite set of independent observations $\mathcal{S} = \{\xi_t\}_{1 \leq t \leq T}$. Under this setting one minimizes the following optimization objective:

$$\min_w \frac{1}{T}\sum_{t=1}^{T} f(w,\xi_t) + \psi(w), \tag{1}$$

where $\psi(w)$ represents a regularization term. Every observation is given as a pair of input–output variables $\xi = (x,y)$. In the above setting one deals with a simple classification model $\hat{y}_t = \text{sign}(\langle w, x_t\rangle)$ and calculates the corresponding loss $f(w,\xi_t)$ accordingly.[1] It is common to acknowledge Eq. (1) as an online learning problem if $T \to \infty$.

The problem in Eq. (1) can be approached using a sequence of strongly convex optimization objectives. The solution of every optimization problem at iteration t is treated as a hypothesis of a learner which is induced by an expectation of possibly non-smooth loss function, *i.e.* $\mathbb{E}_\xi[f(w,\xi)]$. One can regularize it by a reweighted norm at each iteration t. This approach in case of satisfying the sufficient conditions will induce a bounded regret *w.r.t.* the loss function which is generating a sequence of stochastic subgradients endowing our dual space E^* [22].

For promoting sparsity we define an iterate-dependent regularization $\psi_t(w) \triangleq \lambda\|\Theta_t w\|$ which in the limit ($t \to \infty$) applies an approximation to the l_0-norm penalty. At every iteration t we will be solving a separate convex instantaneous optimization problem conditioned on a combination of the diagonal reweighting matrices Θ_t. Specific variations of $\psi_t(w)$ for different norms (*e.g.* l_1- and l_2-norm) will be presented in the next subsections. By using a *simple dual averaging* scheme [22] we can solve our problem effectively by the following sequence of iterates w_{t+1}:

$$w_{t+1} = \arg\min_w \left\{ \frac{1}{t}\sum_{\tau=1}^{t} (\langle g_\tau, w\rangle + \psi_\tau(w)) + \frac{\beta_t}{t} h(w) \right\}, \tag{2}$$

where $h(w)$ is an auxiliary 1-strongly convex smoothing term (proximal operator defined as $h(w) = \frac{1}{2}\|w - w_0\|$, where w_0 is set to origin), $g_t \in \partial f_t(w_t)$ represents a subgradient and $\{\beta_t\}_{t\geq 1}$ is a non-negative and non-decreasing sequence, which determines the boundedness of the regret function of our algorithms.[2]

In detail Eq. (2) is derived using a different optimization objective where we have replaced static regularization term $\psi(w)$ in Eq. (1) with the iterate-dependent term $\psi_t(w)$. In the latter case our optimization objective becomes

$$\min_w \frac{1}{T}\sum_{t=1}^{T} \phi_t(w), \tag{3}$$

where composite function $\phi_t(w)$ is defined as $\phi_t(w) \triangleq f(w,\xi_t) + \psi_t(w)$. Using the aforementioned *dual averaging* scheme from [22] it is easy to show that the sequence w_t in Eq. (2) will approximate an optimal solution to Eq. (3) if we linearly approximate an accumulated loss function $f(w,\xi_t)$ from $\phi_t(w)$ and add a smoothing term $h(w)$. For exact details the interested reader can refer to Eq. (2.14) or in depth to Theorem 1 in [22].

3.2. Reweighted l_1-Regularized Dual Averaging

For promoting additional sparsity to the l_1-Regularized Dual Averaging method [1] we define $\psi_t(w) = \psi_{l_1,t}(w) \triangleq \lambda\|\Theta_t w\|_1$. Hence Eq. (2) becomes:

$$w_{t+1} = \arg\min_w \left\{ \frac{1}{t}\sum_{\tau=1}^{t} (\langle g_\tau, w\rangle + \lambda\|\Theta_\tau w\|_1) + \frac{\gamma}{\sqrt{t}}\left(\frac{1}{2}\|w\|_2^2 + \rho\|w\|_1\right) \right\}. \tag{4}$$

For our reweighted l_1-RDA approach we set $\beta_t = \gamma\sqrt{t}$ and we replace $h(w)$ in Eq. (2) with the parameterized version:

$$h_{l_1}(w) = \tfrac{1}{2}\|w\|_2^2 + \rho\|w\|_1. \tag{5}$$

Each iterate has a closed form solution. Let us define $\eta_t^{(i)} = \frac{\lambda}{t}\sum_{\tau=1}^{t} \Theta_\tau^{(ii)} + \gamma\rho/\sqrt{t}$ and give an entry-wise solution by:

$$w_{t+1}^{(i)} = \begin{cases} 0, & \text{if } |\hat{g}_t^{(i)}| \leq \eta_t^{(i)} \\ -\frac{\sqrt{t}}{\gamma}(\hat{g}_t^{(i)} - \eta_t^{(i)}\,\text{sign}(\hat{g}_t^{(i)})), & \text{otherwise} \end{cases}, \tag{6}$$

where $\hat{g}_t^{(i)} = \frac{t-1}{t}\hat{g}_{t-1}^{(i)} + \frac{1}{t}g_t^{(i)}$ is the i-th component of the averaged $g_t \in \partial f_t(w_t)$, *i.e.* $\hat{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$.

3.3. Reweighted l_1-RDA algorithm

In this subsection we will outline and explain our main algorithmic scheme for the Reweighted l_1-RDA method. It consists of a simple initialization step, drawing a sample $\mathcal{A}_t \subseteq \mathcal{S}$ from the dataset $\mathcal{S}$, computation and averaging of the subgradient g_t,

[1] Throughout this paper we will fix $f(w)$ to the hinge loss $f(w,\xi_t) = \max\{0, 1 - y_t\langle w, x_t\rangle\}$.

[2] See Sections 3.4 and 3.7.