Signal Processing 109 (2015) 206–211

Contents lists available at ScienceDirect

Signal Processing

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

Fast Communication

An ℓ_2/ℓ_1 regularization framework for diverse learning tasks

Shengzheng Wang*, Jing Peng, Wei Liu

Merchant Marine College, Shanghai Maritime University, Shanghai 201306, PR China

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form
13 November 2014
Accepted 17 November 2014
Available online 26 November 2014

Keywords:
$\ell 2/\ell 1$ regularization
Diverse tasks
Regularized empirical risk minimization
Machine learning

ABSTRACT

Regularization plays an important role in learning tasks, to incorporate prior knowledge about a problem and thus improve learning performance. Well known regularization methods, including ℓ_2 and ℓ_1 regularization, have shown great success in a variety of conventional learning tasks, and new types of regularization have also been developed to deal with modern problems, such as multi-task learning. In this paper, we introduce the ℓ_2/ℓ_1 regularization for diverse learning tasks. The ℓ_2/ℓ_1 regularization is a mixed norm defined over the parameters of the diverse learning tasks. It adaptively encourages the diversity of features among diverse learning tasks, i.e., when a feature is responsible for some tasks it is unlikely to be responsible for the rest of the tasks. We consider two applications of the ℓ_2/ℓ_1 regularization framework, i.e., learning sparse self-representation of a dataset for clustering and learning one-vs.-rest binary classifiers for multi-class classification, both of which confirm the effectiveness of the new regularization framework over benchmark datasets.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Regularization plays an important role in learning tasks [1–6]. It helps incorporate prior knowledge into a learning problem and improve learning performance. In addition, given a finite training set sampled from an unknown distribution, the balance between bias (approximation error) and variance (estimation error) is of crucial importance for generalization, while regularization provides a valuable strategy for the bias and variance trade-off. Practically, regularization has shown great success in various learning problems, from classification, regression to learning data representations.

The ℓ_2 regularization, also known as the *Tikhonov* regularization, is one of the most used regularization methods that can be found in a wide spectrum of problems, e.g, the regularized least square regression and function learning in the reproducing Hilbert space [1]. However, the ℓ_2 regularization generally leads to nonsparse representations, while a sparse model is better interpretable and helps identify the most important factors in a problem. In contrast, the ℓ_1 regularization has received the greatest attentions in current studies for its capability to encourage sparsity. Popular sparse models with the ℓ_1 regularization include the Lasso [2], sparse coding [4], and covariance selection [5]. New types of regularization have recently been introduced to fulfil modern learning problems. For example, in multi-task learning, it is assumed that the multiple regression functions share a common sparse structure over input features, and the ℓ_1/ℓ_2 and ℓ_1/ℓ_∞ regularization are used to select these common features [7,8]. Both of these regularization functions use the ℓ_1 norm to encourage "block-sparsity" of the regression coefficients matrix, and thus a feature is jointly relevant to no regressions or mostly all ones.

In this paper, we consider a new regularization framework for learning a diverse set of tasks. In contrast to multi-task learning, where it is assumed that the tasks share a common set of features, we assume that the features are sparsely distributed among different tasks, that is when a feature is responsible for some tasks it is unlikely to be responsible for

* Corresponding author. Tel.: +8621 38282914.
E-mail address: szwang.smu@gmail.com (S. Wang).

http://dx.doi.org/10.1016/j.sigpro.2014.11.010
0165-1684/© 2014 Elsevier B.V. All rights reserved.

the rest of the tasks. Since the ℓ_1/ℓ_2 and ℓ_1/ℓ_∞ regularization used in multitask learning have opposite intuitions to learning diverse tasks, they are therefore inapplicable. To address such a new learning problem, we introduce the ℓ_2/ℓ_1 regularization framework. The ℓ_2/ℓ_1 regularization is a mixed norm defined over the parameters of the diverse learning tasks and adaptively encourages the sparse distribution of features. Further, we present a fast first-order algorithm for ℓ_2/ℓ_1 regularized empirical risk minimization. In our algorithm, the generalized gradient update in each iteration is reformed as a nonnegative least square (NLS), and further a greedy method with negligible computations is used to solve the NLS. We apply the proposed ℓ_2/ℓ_1 regularization framework to learning sparse self-representation of a dataset for clustering and to multiclass classification.

2. The framework of ℓ_2/ℓ_1 regularized multilabel learning

For convenience of presentation, we first clarify some notations to be used in the derivation. Bold lowercase letter $\mathbf{a}$ denotes a vector, while bold uppercase letter $\mathbf{A}$ denotes a matrix $\|\cdot\|$ denotes the ℓ_2 norm for a vector while the Frobenius norm for a matrix. $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{trace}(\mathbf{A}^T\mathbf{B})$. $\mathbf{sgn}(\cdot)$ and $\mathbf{abs}(\cdot)$ are sign and absolute value operations on vectors, respectively. $\odot$ denotes element-wise product between vectors. By $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$, we mean that $\mathbf{a}_j$ is the j-th column of $\mathbf{A}$, while by $\mathbf{A} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_p^T]^T$, we mean that $\mathbf{a}_j$ is the j-th row of $\mathbf{A}$. $\mathbf{e}$ denotes a vector of proper size with all elements being 1, and $\mathbf{E} = \mathbf{e}\mathbf{e}^T$. $\nabla f(x)$ denotes the gradient of function $f(x)$ at point x.

2.1. Regularized empirical risk minimization

Suppose we have m learning tasks, having training dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ and being parameterized by $\mathbf{B} \in \mathbb{R}^{m\times p}$, where p is the dimension of training examples. With a properly chosen loss function $\mathcal{L}$, the optimal parameter B can be jointly learned by the following regularized empirical risk minimization:

$$\begin{aligned}\mathbf{B}_{opt} &= \arg\min_{\mathbf{B}} \mathcal{L}(\mathbf{B};\mathcal{D}) + \varrho\mathcal{R}(\mathbf{B}) \\ &= \arg\min_{\mathbf{B}} \sum_{i=1}^{m} \mathcal{L}(\mathbf{B}(i,:);\mathcal{D}_i) + \varrho\mathcal{R}(\mathbf{B}) \end{aligned} \tag{1}$$

where $\mathcal{R}(\mathbf{B})$ is the regularization term and ϱ is the tuning parameter.

Generally, $\mathcal{R}(\mathbf{B})$ encodes the prior knowledge we assumed on the m learning tasks. For instance, if we assume sparsity on the feature distribution for each learning task, the ℓ_1 regularization $\|\mathbf{B}\|_1 = \sum_{k=1}^{m}\sum_{j=1}^{p}|\mathbf{B}(k,j)|$ can be applied. In this paper, for learning diverse tasks, we introduce the following ℓ_2/ℓ_1 regularization:

$$\mathcal{R}(\mathbf{B}) = \|\mathbf{B}\|_{2,1}^2 = \sum_{j=1}^{p}\left(\sum_{k=1}^{m}|\mathbf{B}(k,j)|\right)^2. \tag{2}$$

First, the internal ℓ_1 norm encourages the $\mathbf{B}(:,j)$ to be sparse over tasks, while the external ℓ_2 norm is used to control the complexity of entire model. But, more importantly, as we expand the regularization as below

$$\|\mathbf{B}\|_{2,1}^2 = \sum_{j=1}^{p}\sum_{k=1}^{m}|\mathbf{B}(k,j)|^2 + \sum_{j=1}^{p}\sum_{k=1}^{m} w(k,j)|\mathbf{B}(k,j)| \tag{3}$$

with

$$w(k,j) = \sum_{i\neq k}|\mathbf{B}(i,j)|, \tag{4}$$

one can see that for each feature j, it is penalized with more weight $w(k,j)$ on task k, if it contributes more to the rest of the tasks. This adaptively makes a diverse and sparse distribution of features among tasks. In contrast, though the ℓ_1 regularization also encourages sparsity, the penalties on each feature over different tasks are independent.

Combining (1) and (2), the ℓ_2/ℓ_1 regularized empirical risk minimization framework is given by

$$\mathbf{B}_{opt} = \arg\min_{\mathbf{B}} \mathcal{L}(\mathbf{B};\mathcal{D}) + \varrho\|\mathbf{B}\|_{2,1}^2, \tag{5}$$

where $\varrho > 0$ is the tuning parameter.

2.2. First-order algorithm

Now, we derive a first-order algorithm for solving the ℓ_2/ℓ_1 regularized empirical risk minimization framework (5). For simplification, we use $\mathcal{L}(\mathbf{B})$ to denote $\mathcal{L}(\mathbf{B};\mathcal{D})$, and assume that it is convex and has Lipschitz continuous gradient, which can be satisfied by choosing a proper loss function. However, as the $\|\mathbf{B}\|_{2,1}^2$ term is nonsmooth (though convex), the minimization in (5) is generally nontrivial. In addition, as the ℓ_1/ℓ_2 regularization developed for multi-task learning in the literature has a different formulation, i.e., first calculating the ℓ_2 norm of each row and then the ℓ_1 norm, which is in contrast to the ℓ_2/ℓ_1 regularization (2), the corresponding algorithm is not applicable here.

The first-order method has received considerable attentions in solving machine learning problems, because of its attractive efficiency and scalability. In this paper, we also exploit the same strategy to solve our problem (5). Specifically, in each first-order iteration, we solve the following problem:

$$\mathbf{B}^* = \arg\min_{\mathbf{Z}} \langle \nabla\mathcal{L}_n(\mathbf{B}), \mathbf{Z}-\mathbf{B}\rangle + \frac{1}{2t}\|\mathbf{Z}-\mathbf{B}\|^2 + \varrho\|\mathbf{Z}\|_{2,1}^2. \tag{6}$$

where $t > 0$ is set such that $1/t$ is larger than the Lipschitz constant of $\mathcal{L}_n$'s gradient. This can be regarded as an application of the forward–backward splitting algorithm or the majorization–minimization method and thus with guaranteed convergence [9–11].

Denoting $\mathbf{B}^* = [\mathbf{b}_1^*, \mathbf{b}_2^*, \ldots, \mathbf{b}_p^*]$ and $\nabla\mathcal{L}(\mathbf{B}) = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p]$, (6) requires p separated subproblems:

$$\begin{aligned}\mathbf{b}_j^* &= \arg\min_{\mathbf{z}} \langle \mathbf{h}_j, \mathbf{z}-\mathbf{b}_j\rangle + \frac{1}{2t}\|\mathbf{z}-\mathbf{b}_j\|^2 + \varrho\|\mathbf{z}\|_1^2 \\ &= \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}\|^2 - \langle \mathbf{b}_j - t\mathbf{h}_j, \mathbf{z}\rangle + t\varrho\|\mathbf{z}\|_1^2, \end{aligned} \tag{7}$$

with $1 \le j \le p$. Similar problems to (7), with more general settings, have actually been studied in the literature of sparse signal recovery. In particular, a coordinatewise soft-thresholding algorithm given in [12] solves (7) in finite iterations. Here, we propose an alternative strategy to solve (7). We first convert (7) to a nonnegative least squares (NLS) problem and then solve the NLS by an efficient greedy algorithm.

Download English Version:

https://daneshyari.com/en/article/562902

Download Persian Version:

https://daneshyari.com/article/562902

Daneshyari.com