Neural Networks 46 (2013) 172–182

Contents lists available at ScienceDirect

Neural Networks

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

Training Lp norm multiple kernel learning in the primal

Zhizheng Liang*, Shixiong Xia, Yong Zhou, Lei Zhang

School of Computer Science and Technology, China University of Mining and Technology, China

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form 4 May 2013
Accepted 5 May 2013

Keywords:
Multiple kernel learning
Manifold regularization
Primal optimization
Empirical Rademacher complexity
Data classification

ABSTRACT

Some multiple kernel learning (MKL) models are usually solved by utilizing the alternating optimization method where one alternately solves SVMs in the dual and updates kernel weights. Since the dual and primal optimization can achieve the same aim, it is valuable in exploring how to perform Lp norm MKL in the primal. In this paper, we propose an Lp norm multiple kernel learning algorithm in the primal where we resort to the alternating optimization method: one cycle for solving SVMs in the primal by using the preconditioned conjugate gradient method and other cycle for learning the kernel weights. It is interesting to note that the kernel weights in our method can obtain analytical solutions. Most importantly, the proposed method is well suited for the manifold regularization framework in the primal since solving LapSVMs in the primal is much more effective than solving LapSVMs in the dual. In addition, we also carry out theoretical analysis for multiple kernel learning in the primal in terms of the empirical Rademacher complexity. It is found that optimizing the empirical Rademacher complexity may obtain a type of kernel weights. The experiments on some datasets are carried out to demonstrate the feasibility and effectiveness of the proposed method.

1. Introduction

Kernel-based methods (Scholkopf & Smola, 2002) have been widely used to solve some machine learning problems such as classification, regression and dimensionality reduction in the past several decades. Support vector machines (SVMs), one of the most successful applications in kernel-based methods, have good generalization performance due to the trade-off between the training error and the maximization margin. However, in some classification problems one often faces a lack of sufficient labeled data in the case that manually labeling data is time-consuming or inappropriate. In order to address this problem, some semi-supervised learning algorithms have been proposed by using the unlabeled and labeled samples. The Laplacian support vector machine (LapSVM) which is one of the representative methods in semi-supervised learning imposes the manifold regularization term in the objective function to assume that the samples in the local region have similar labels. It is noted that kernel functions used in SVMs and LapSVMs can capture the similarity between a pair of samples. In some real applications one can obtain different types of features of samples where each feature representation may construct a kernel matrix, thereby resulting in multiple kernel matrices for the datasets. As a result some multiple kernel variants related to SVMs (Bach, Lanchriet, & Jordan, 2004; Lanckriet, Cristianini, Bartlett, Gaoui, & Jordan, 2004) have been proposed in recent several years.

Different from single kernel learning where one usually needs to choose proper kernel parameters, MKL usually searches for linear (nonlinear) combinations of predefined base kernels by maximizing some generalization performance measures such as the margin maximization. MKL not only provides the scheme for fusing heterogeneous data from multiple sources where each kernel may capture a type of similarities corresponding to a data source, but also gives a strategy for addressing the problem of kernel selection. Thus MKL provides more flexibility in solving some classification problems than single kernel learning. The original multiple kernel learning where linearly combinational kernels (Bach et al., 2004) are adopted can be formulated as a semi-definite programming (SDP) or second-order cone programming (SOCP) problem (Alizadeh & Goldfarb, 2003). However, due to the high computational cost of solving SDP and SOCP, this class of MKL only handles small-scale or medium-scale datasets. In order to improve the efficiency of MKL, the alternating optimization method where the kernel weights and the coefficients of the span of training samples are alternately updated is adopted. For example, in Sonnenburg, Ratsch, Schafer, and Scholkopf (2006), a semi-infinite linear program (SILP) approach is used to solve MKL where kernel weights are updated by a cutting plane method. Note that SILP may suffer from the instability of the solution of MKL. To this end, Rakotommonjy, Bach, Canu, and Grandvalet (2008)proposed SimpleMKL to overcome this problem where the kernel weights are obtained by a reduced gradient descent

* Corresponding author. Tel.: +86 0516 15996966232; fax: +86 0516 83995918.
E-mail address: cuhk_liang@yahoo.cn (Z. Liang).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.05.003

method. Obviously the existing SVM solvers are directly used to obtain the coefficients of the span of training samples in SILP or SimpleMKL.

Some MKL algorithms impose the L1 norm constraint on the kernel weights. This is usually called L1 norm MKL (Lanckriet et al., 2004). In order to improve the efficacy of L1 norm MKL, the dual augmented-Lagrangian algorithm is used to solve L1 norm multiple kernel learning in Suzuk and Tomioka (2011). L1 norm MKL generally obtains the sparse solution of kernel weights and thus it has good interpretability in kernel selection. But, as pointed out in Kloft, Brefeld, Sonnenburg, and Zien (2011), if the kernels encode orthogonal or complemental information, L1 norm MKL may yield the undesirable performance for some classification problems due to over-sparseness of kernel weights. To this end, some non-sparse MKL methods such as Lp norm MKL and MKL based on entropy regularization (Xu, Jin, Zhu, Lyu, & King, 2010) are proposed. In Vishwanathan, Sun, Ampornputn, and Varma (2010), the sequential minimization optimization (SMO) algorithm is used to solve Lp norm MKL. To make the trade-off between the orthogonal information and the sparse kernel weights, Yang, Xu, Ye, King, and Lyu (2011) also used the regularization with a linear combination of L1 norm and L2 norm to control kernel weights. In order to explore a group structure among kernels, Szafranski, Grandvalet, and Rakotomamonjy (2010) proposed composite kernel learning. In addition, the mixed norm regularization for group structures in MKL is also developed in Aflalo, Ben-Tal, Bhattacharyya, Nath, and Raman (2011), where the efficient mirror decent method is used to solve multiple kernel learning.

In some MKL algorithms, the SVM solvers are used to obtain the coefficients of the span of training samples and SVMs are usually solved in the dual. Thus these MKL algorithms are actually solved in terms of their dual representations. Similar to SVMs in the primal (Chapelle, 2007; Joachims, 2006; Shalev-Shwartz, Singer, & Srebro, 2007; Yu, Vishwanathan, Güunter, & Schraudolph, 2010), one can also implement MKL in the primal. Moreover, Chapelle (2007) also pointed out that primal optimization and the dual optimization are two equivalent ways of reaching the same aim. To this end, in this paper we will explore how to implement Lp norm MKL in the primal. Similar to previous MKL algorithms, we also resort to the alternating optimization algorithm to solve Lp norm MKL in the primal. That is, we alternately optimize SVMs and update the kernel weights in the primal. In addition, we also perform the theoretical analysis for the proposed method in terms of the empirical Rademacher complexity. It is noted that the theoretical analysis in this paper also gives some suggestions on how to obtain kernel weights in terms of the empirical Rademacher complexity. Finally, we carry out the experiments on some datasets to show the effectiveness and usefulness of the proposed method.

The rest of this paper is organized as follows. In Section 2, we briefly review the related work on supervised MKL and LapSVMs. In Section 3, we introduce MKL based on the manifold regularization in the primal and discuss how to solve it in the primal. In Section 4, we give the theoretical analysis for the proposed method in terms of the empirical Rademacher complexity. In Section 5, we demonstrate the effectiveness of Lp norm MKL in the primal on some datasets. Conclusions and further work are given in the final section.

2. Related work

2.1. Supervised multiple kernel learning based on margin maximization

Let $X = (x_1, \ldots, x_n) \in R^{m\times n}$ be a matrix whose columns consist of n training samples in an m-dimensional space. For convenience, assume that these n samples are ordered so that the first l ones are labeled, with $y_i \in \{-1, 1\}$, and the remaining u samples are unlabeled, $l + u = n$. Supervised multiple kernel learning based on the margin maximization can be formulated as the following optimization problem:

$$\min_{f\in H_\mu} r_1 \|f\|^2_{H_\mu} + \sum_{i=1}^{l} \max(0, 1 - y_i f(x_i))^s, \tag{1}$$

where s is either 1 (hinge loss) or 2 (squared hinge loss), H_μ is a reproducing kernel Hilbert space with the parameter μ, H_μ is endowed with the kernel function K taken from a linear space of base kernels $K = \sum_{i=1}^{d} \mu_i K_i$, and r_1 is a regularization parameter. When the hinge loss, i.e., $s = 1$, is used, it is proved (Xu, Jin, Ye, King, & Lyu, 2010) that Eq. (1) is equivalent to the following optimization problem.

$$\min_{\mu_j\in\Delta, f_j\in H_j} r_1 \sum_{j=1}^{d} \mu_j \|f_j\|^2_{H_j} + \sum_{i=1}^{l} \max\left(0, 1 - y_i \sum_{j=1}^{d} \mu_j f_j(x_i)\right), \tag{2}$$

where $\Delta = \{\mu : \sum_{j=1}^{d} \mu_j = 1, \mu_j \geq 0\}$, and H_j is the jth reproducing kernel Hilbert space. Eq. (2) can be solved by using the alternating optimization algorithm. In fact, one usually solves SVMs in the dual in the case of fixed kernel weights. It is noted that the kernel weights can be analytically obtained from Eq. (2). If one uses the constraint $\|\mu\|_p^p \leq 1$ to replace Δ in Eq. (2), Lp norm MKL (Kloft et al., 2011; Xu, Jin, Ye et al., 2010) can be obtained.

2.2. Laplacian SVMs

By considering the intrinsic structure of data points, Laplacian SVMs (Belkin, Niyogi, & Sindhwani, 2006; Melacci & Belkin, 2011) add the manifold regularization to the objective function of SVMs. That is, the following optimization problem is constructed:

$$\min_{f\in H} r_1 \|f\|^2_H + \sum_{i=1}^{l} \max(0, 1 - y_i f(x_i))^s + r_2 f^T L f, \tag{3}$$

where L is a Laplacian matrix obtained from n training samples and r_2 is the weight of the norm of the function in a low-dimensional manifold, which imposes the smoothness on the manifold. It is proved in Belkin et al. (2006) that f admits an expansion in terms of n training samples, denoted by

$$f(x) = \sum_{i=1}^{n} \beta_i k(x_i, x). \tag{4}$$

Substituting Eq. (4) into Eq. (3), one can obtain

$$\min_{f\in H} r_1 \beta^T K \beta + \sum_{i=1}^{l} \max(0, 1 - y_i K(:, i)^T \beta)^s + r_2 \beta^T K L K \beta, \tag{5}$$

where K is the kernel matrix obtained by using n training samples and $K(:, i)$ denotes the ith column of K. Melacci and Belkin (2011) noted that the computational complexity of solving Eq. (5) in the dual is $O(n^3)$ while the computational complexity of solving Eq. (5) in the primal is $O(n^2 k)$, where k is empirically estimated to be significantly smaller than n.

3. MKL in the primal based on the manifold regularization

Motivated by the facts that training LapSVMs in the primal is much more effective than training LapSVMs in the dual (Melacci & Belkin, 2011) and multiple kernels can capture different