Information Sciences 294 (2015) 255–268

Contents lists available at ScienceDirect

Information Sciences

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

Optimal learning rates of l^p-type multiple kernel learning under general conditions

Shaogao Lv *, Fanyin Zhou

The School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China

ARTICLE INFO

Article history:
Received 25 March 2013
Received in revised form 7 September 2014
Accepted 13 September 2014
Available online 22 September 2014

Keywords:
Multiple kernel learning
Kernel learning
Correlation measure
Generalization ability
Local Rademacher complexity

ABSTRACT

One of the most promising learning kernel methods is the l^p-type multiple kernel learning proposed by Kloft et al. (2009). This method can adaptively select kernel function in supervised learning problems. The majority of the studies associated with generalization error have recently received wide attention in machine learning and statistics. The present study aims to establish a new generalization error bound under more general frameworks, in which the correlation among reproducing kernel Hilbert spaces (RKHSs) is considered, and the restriction of smooth condition on the target function is relaxed. In this case, the interaction between the estimation and approximation errors must be simultaneously regarded. In this investigation, optimal learning rates are derived by applying the local Rademacher complexity technique, which is given in terms of the capacity of RKHSs spanned by multi-kernels and target function regularity.

1. Introduction

Kernel-based learning methods, such as support vector machines (SVM), have been extensively applied in supervised learning tasks, including classification and regression. These methods implicitly map the input data into a high dimensional feature space, in which the implicit mapping Φ is defined by a kernel function that returns the inner product $\langle \Phi(x), \Phi(y) \rangle$ between the images of data points x and y. Hence, these kernel-learning approaches are computationally more efficient than the other methods that are required to project x and y explicitly into feature space. From a practical perspective, kernel-based algorithms are defined in an infinite functional space. However, they can efficiently work in a finite space for many applications, and can capture nonlinear structures in many real-world data sets.

Kernels and associated RKHSs are simple and can be generally applied; thus, they play an increasingly important role in machine learning and statistics. Selecting regularization parameters is an immediate concern when kernel is provided. This is typically solved by conducting cross-validation or generalized cross-validation [16]. However, most kernel-based learning algorithms critically rely on the selection of kernel function, which induces the issue of choosing the optimal kernel from a collection of candidates.

Numerous kernel selection methods have been proposed in the literature. For example, by using the outer product of the label vector as the ground-truth, The kernel target alignment actualized by Cristianini et al. [13] and Cortes et al. [9] aims to learn the entries of a kernel matrix using the outer product of label vector as ground truth. A gradient descent algorithm was developed by Chapelle et al. [12] and Bousquet and Herrmann [5] to minimize an estimate of SVMs generalization error

* Corresponding author.
E-mail address: lvsg716@swufe.edu.cn (S. Lv).

http://dx.doi.org/10.1016/j.ins.2014.09.011
0020-0255/

using the outer product of label vector as ground truth over a set of parameters defined in kernels sequence. The hyperkernels method was introduced by Ong et al. to directly analyze kernel function in an inductive setting. Alternative kernel selection approaches include the DC and semi-infinite programming realized by Argyriou et al. [1] and Gehler and Nowozin [15], respectively. Nevertheless, these approaches mentioned above often lead to non-convex optimization problems, wherein ensuring computational efficiency is difficult. Kernel learning algorithms with different settings for the width parameter of Gaussian kernels must be explored to obtain an optimal parameter from a specified interval [32]. These particular approaches can result in a standard convex optimization program, that is, the Gaussian kernels, which correspond to a sequence of capacity-limited functional spaces. However, such algorithm lacks the flexibility of modeling data with great complexity. Using classification information, Zamania et al. [37] proposed to determine a kernel function for kernel principal components and kernel linear discriminant analyses. Learning kernels with linear combinations of multiple kernel functions has recently received considerable attention in machine learning. Kloft et al. [17] proposed the so-called l^p-norm $(1 < p \leqslant 2)$ multiple kernel learning (MKL) method, which has been proven useful and effective in terms of theoretical analysis and practical applications.

l^p-norm MKL is an empirical minimization algorithm that operates on multi-kernel class, which consists of functions $\left\{f \in \mathcal{H}_{K_\theta} :\ K_\theta = \sum_{m=1}^{M} \theta_m K_m |\ \|\theta\|_p \leqslant 1, \theta \geqslant 0\right\}$, where M is the number of given candidate kernels. l^p-norm MKL has been successfully applied in solving real-world problems. For example, Kloft et al. [18] originally implemented the algorithm to clarify problems in bioinformatics. The results of his research revealed that l^p-norm MKL $(p > 1)$ achieves more accurate prediction results than the state-of-the-art methods. Yu et al. [36] developed a l^2-norm MKL algorithm and applied it to genomic data fusion. The results of their investigation showed that this algorithm achieved comparable performance with the conventional SVM–MKL algorithms. Moreover, in generic object recognition research, Nakashika and Suga [24] proposed a novel feature selection method based on MKL. Their experimental results illustrated the effectiveness of the proposed automatic feature selection method. Some researchers recently interpreted MKL from different views. For instance, Xu et al. [35] presented a novel soft margin perspective for MKL under more general frameworks, and many existing MKL can be viewed as special cases. Mao et al. [23] introduced a novel probabilistic interpretation of MKL and proposed a hierarchical Bayesian model that can simultaneously learn the proposed data-dependent prior and classification.

Over the past few years, MKL has been theoretically analyzed without a hitch. Cortes et al. [8,9] obtained the convergence rates associated with l^p-MKL of the order $\sqrt{\frac{\log(M)}{n}}$ with $p = 1$ and $\frac{M^{1-1/p}}{\sqrt{n}}$ with $1 < p \leqslant 2$. Kloft et al. [18] derived a similar convergence bound with improved constants. Bartlett et al. [6] and Kloft et al. [19] adopted localization techniques, including the local Rademacher complexities, and conventionally obtained sharp learning rates. Kloft and Blanchard [19] provided a localized convergence of l^p-MKL. However, their analysis relied on a strong condition that the underlying RKHSs are no correlated with one another. Considering the correlation between candidate kernels, Suzuki [30] derived the fast learning rates of dense-type regularization in a unifying framework, which included l^p-MKL as a special case.

However, these algorithms were conducted under a strong assumption that the target function is smooth and lies in the hypothesis space where the algorithm works. The approximation error is neglected and the relationship between regularization parameter and smoothness of the target function cannot be completely reflected. In the literature of statistical learning theory, *generalization error = estimation error + approximation error* for a given estimator (see details in [4]). Multiple kernels evidently lead to additional functional complexity. Hence, learning with these kernels can only achieve a better performance for generalization ability when it can significantly improve approximation ability. Based on this argument, the case in which the target function does not lie in the hypothesis space can be compellingly considered. In such instance, previous techniques of analyzing l^p-MKL are no longer valid because the upper bound of the estimator goes to infinity with the increase of the sample size n, which affects the upper bound of the estimation error. Therefore, more explicit analysis methods of deriving the optimal learning rates under general conditions must be developed. To the best of our knowledge, no existing study has explicitly analyzed how the correlations among RKHSs affect the learning rates under the multi-kernel learning settings.

The present study primarily aims to derive the optimal learning rates of l^p-MKL under a mild condition, that is, the target function does not lie in kernel classes. In this case, optimal rates can be obtained by uniformly bounding the second moments of functions from an adequate class by their expectations. Classical empirical process theory of the local Rademacher complexities is extended to more general cases. Thus, the optimal rates in this investigation are derived with an iterative technique. This study generally provide beneficial contributions in the following aspects:

- Optimal learning rates are obtained by considering the correlation structure of the underlying RKHSs. The final result shows that the correlation greatly affects convergence rates.
- Convergence rates are established under a mild assumption on target function that effectively relaxes function constrains.
- As a by-product, a tight bound is provided for concentration inequality by applying the local Rademacher complexity under the general conditions stipulated in Section 5. Moreover, the advantages of MKLin terms of approximation ability are discussed.

The rest of this paper is organized as follows. Section 2 formulates the classical supervised learning problem, introduces the MKL algorithm for the regression problem, and provides main assumptions for theoretical analysis. Section 3 presents