Journal of Statistical Planning and Inference 143 (2013) 276–282

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Learning rates of multi-kernel regression by orthogonal greedy algorithm

Hong Chen [a], Luoqing Li [b], Zhibin Pan [a,*]

[a] College of Science, Huazhong Agricultural University, Wuhan 430070, China
[b] Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China

ARTICLE INFO

Article history:
Received 22 July 2012
Accepted 3 August 2012
Available online 16 August 2012

Keywords:
Sparse
Multi-kernel learning
Orthogonal greedy algorithm
Data dependent hypothesis space
Rademacher chaos complexity
Learning rate

ABSTRACT

We investigate the problem of regression from multiple reproducing kernel Hilbert spaces by means of orthogonal greedy algorithm. The greedy algorithm is appealing as it uses a small portion of candidate kernels to represent the approximation of regression function, and can greatly reduce the computational burden of traditional multi-kernel learning. Satisfied learning rates are obtained based on the Rademacher chaos complexity and data dependent hypothesis spaces.

1. Introduction

Kernel methods have been extensively used in various learning tasks, and its performance largely depends on the data representation via the choice of kernel function. Due to the practical importance of multi-kernel learning, many studies in machine learning have been devoted to the data dependent choice of kernel recently, see, e.g., Lanckriet et al. (2004), Micchelli and Pontil (2005), Wu et al. (2007), Ying and Zhou (2007), Ying and Campbell (2010), and Chen and Li (2010).

In the regression setting, the above mentioned multi-kernel models usually can be formulated as a regularized framework in reproducing kernel Hilbert spaces. Let us recall some basic concepts of the multi-kernel regularized regression. Let X be a compact subset of $\mathbb{R}^d$ and let Y be contained in $[-M,M]$. The product space $Z := X \times Y$ is assumed to be measurable and it is endowed with an unknown probability measure denoted by ρ. Input–output pairs (x,y) are sampled according to ρ. For every $x \in X$, let $\rho(y|x)$ be the conditional (w.r.t. x) probability measure on Y and let $\rho_X(x)$ be the marginal probability measure on X. The error for a measurable function $f : X \to Y$ is the so-called expected risk

$$\mathcal{E}(f) := \|y-f\|_{L^2_\rho}^2 = \int_Z (y-f(x))^2 \, d\rho.$$

It is known that the function which minimizes $\mathcal{E}(f)$ is the regression function defined by

$$f_\rho(x) = \int_Y y \, d\rho(y|x), \quad x \in X. \tag{1}$$

From the assumption $y \in [-M,M]$, we know that $|f_\rho(x)| \le M$.

* Corresponding author.
E-mail address: zhibinpan2008@gmail.com (Z. Pan).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.08.002

Set $\mathbb{N}_m := \{1,2,\ldots,m\}$ for any $m \in \mathbb{N}$. A training set of size m is drawn by sampling m independent and identically distributed pairs according to ρ,

$$\mathbf{z} := \{z_i, i \in \mathbb{N}_m\} = \{(x_i,y_i), i \in \mathbb{N}_m\} \in Z^m.$$

Throughout the paper, we restrict our attention to a prescribed set $\mathcal{K}$ of candidate Mercer kernels. We say that $K : X \times X \to \mathbb{R}$ is a Mercer kernel if it is a continuous, symmetric, and positive semi-definite, i.e., for any finite set of distinct points $\{x_1,x_2,\ldots,x_\ell\} \subset X$, the matrix $(K(x_i,x_j))_{i,j=1}^{\ell}$ is positive semi-definite. The candidate reproducing kernel Hilbert space $\mathcal{H}_K$ associated with a Mercer kernel K is defined as the closure of the linear span of the set of functions $\{K_x := K(x,\cdot) : x \in X\}$, equipped with the inner product $\langle \cdot,\cdot \rangle_{\mathcal{H}_K}$ defined by $\langle K_x,K_y \rangle_{\mathcal{H}_K} = K(x,y)$. The reproducing property is given by

$$\langle K_x,f \rangle_{\mathcal{H}_K} = f(x), \quad \forall x \in X,\ f \in \mathcal{H}_K. \tag{2}$$

Denote $C(X)$ as the space of continuous functions on X with the supremum norm $\|\cdot\|_\infty$. Because of the continuity of $K \in \mathcal{K}$ and the compactness of X, we have

$$\kappa := \sup_{K\in\mathcal{K}} \sup_{x\in X} \sqrt{K(x,x)} < \infty.$$

So, the reproducing property above tells us

$$\|f\|_\infty \le \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K.$$

The empirical error with respect to the random samples $\mathbf{z}$ is defined as

$$\mathcal{E}_{\mathbf{z}}(f) := \|y-f\|_m = \frac{1}{m}\sum_{i=1}^{m}(y_i-f(x_i))^2,$$

where $\|\cdot\|_m$ is the L^2_ρ norm with respect to the discrete measure $(1/m)\sum_{i=1}^m \delta_{x_i}$ with δ_u is the Dirac measure of u.

In general, the regularization scheme of multi-kernel regression is defined as a two-layer minimization problem

$$f_{\mathbf{z},\lambda} := \operatorname*{argmin}_{K\in\mathcal{K}} \min_{f\in\mathcal{H}_K} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2\}, \quad \lambda > 0. \tag{3}$$

Its error analysis has been well developed with various techniques in learning theory (see, e.g., Lanckriet et al., 2004; Micchelli and Pontil, 2005; Ying and Zhou, 2007; Ying and Campbell, 2010; Chen and Li, 2010).

Here we are interested in the case when the total number n of candidate kernels $\mathcal{K} = \{K^j : j \in \mathbb{N}_n\}$ is large, but only a relatively small number of them is necessary to represent the approximation of regression function f_ρ. Note that the solution of (3) belongs to the hypothesis space

$$\mathcal{H}_{\mathbf{z},\mathcal{K}} = \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i^j K^j_{x_i} : \alpha_i^j \in \mathbb{R}, K^j \in \mathcal{K} \right\},$$

which involves expansions of all the candidate kernels and all the training data. This may result in computation burden when the number of candidate kernels is large. Thus, the sparse representation of solution is crucial to improve the efficiency of multi-kernel learning.

Only recently there are studies for concerning the sparsity of multi-kernel learning in Koltchinskii and Yuan (2008, 2010). For the multiple kernel regularized method with sparsity penalty, the oracle inequality of excess risk is established in Koltchinskii and Yuan (2008). In this paper, we consider to realize sparse representation by greedy selection of important kernels and training samples. We denote the set of candidate kernels is $\mathcal{K}_{k_{\mathbf{z}}}$, where $k_{\mathbf{z}}$ is the number of candidate kernels after k times of feature selection. The hypothesis space based on kernel selection is defined by

$$\mathcal{H}^{k_{\mathbf{z}}}_{\mathbf{z},\mathcal{K}} = \left\{ \sum_{i=1}^{m}\sum_{j=t_1}^{t_{k_{\mathbf{z}}}} \alpha_i^j K^j_{x_i} : \alpha_i^j \in \mathbb{R}, K^j \in \mathcal{K}, t_i \in \mathbb{N}_m \right\}$$

with ℓ_1 norm

$$\|f\|_{\ell_1} = \inf\left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} |\alpha_i^j| : f = \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i^j K^j_{x_i} \right\}.$$

The data dependent hypothesis space $\mathcal{H}_{\mathbf{z},\mathcal{K}}$ can be considered as a natural extension from single kernel setting in Xiao and Zhou (2010) and Shi et al. (2011) to multi-kernel setting.

Based on the hypothesis space, a new multi-kernel orthogonal greedy algorithm (MOGA) is introduced in Table 1.

The algorithm in Table 1 can be divided into two parts: selecting features $\{\phi_k\}$ and solving the empirical risk minimization to derive $\hat{f}_k$. In fact, the goal of the normalization of kernels is to provide the feasibility of error analysis, which does not affect the predictive performance of the algorithm.

There are some statistical analysis of orthogonal greedy algorithms in learning problem (Barron et al., 2008; Zhang, 2009). However, to the best of our knowledge, there is no any studies of kernel choice by greedy algorithm. Our method tries to bring together three distinct concepts that have received independent attention in learning theory: multi-kernel