Neurocomputing 133 (2014) 54–62

Contents lists available at ScienceDirect

Neurocomputing

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

Learning performance of coefficient-based regularized ranking

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

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

ARTICLE INFO

Article history:
Received 7 August 2013
Received in revised form
24 November 2013
Accepted 24 November 2013
Communicated by D. Tao
Available online 13 January 2014

Keywords:
Ranking
Coefficient regularization
Generalization bound
Recommendation system
Drug discovery

ABSTRACT

The regularized kernel methods for ranking problem have attracted increasing attention recently, which are usually based on the regularization scheme in a reproducing kernel Hilbert space. In this paper, we go beyond this framework by investigating the generalization ability of ranking with coefficient-based regularization. A regularized ranking algorithm with a data-dependent hypothesis space is proposed and its representer theorem is proved. The generalization error bound is established in terms of the covering numbers of the hypothesis space. Different from the previous analysis relying on Mercer kernels, our theoretical analysis is based on much general kernel function, which is not necessarily symmetric or positive semi-definite. Empirical results on the benchmark datasets demonstrate the effectiveness of the coefficient-based algorithm.

1. Introduction

The ranking problem has gained increasing attention in machine learning with the fast development of ranking techniques on recommender systems, drug discovery, and searching engines, see, e.g., [1–6]. Usually, a ranking method is aiming to find a score function such that the predicted ranking relation is consistent as possible as the true order. From different perspectives, various ranking algorithms have been proposed including RankSVM [7,1], RankNet [8,9], RankBoost [10], the gradient descent ranking [11,12], the Bayes subset ranking [13], the *P*-norm push ranking [14], and the kernel-based regularized ranking [15–20].

The kernel-based ranking usually can be unified in a Tikhonov regularization scheme in a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel. Based on this regularization scheme, the generalization performance of these methods has been studied in terms of different techniques including stability analysis [15–17,20], uniform convergence analysis based on *U*-statistics [21,19], and approximation analysis based on operator approximation [22,12].

Despite these theoretical progresses have been made, two issues should be further addressed:

- Regularization selection. The previous regularized algorithms are dependent on the norm square regularizer in a RKHS. The natural questions are as follows: Is the regularizer suitable for various ranking tasks? Is there other selection of regularization term? In fact, the optimal regularization for different tasks may be different, and various regularization terms have been used successfully for classification and regression, see, e.g., ℓ_1-regularizer [23–25], ℓ_2-regularizer [26–28], $\ell_{1/2}-$regularizer [29], Elastic-net regularizer [30], and manifold or Hessian regularizer [31–34].
- Kernel selection. The previous theoretical results rely heavily on a Mercer kernel, which is symmetric and positive semi-definite. However, as studied in [26,24,27], the kernel is not necessarily symmetric or positive semi-definite. In theory, the wider selection of kernel provides us more flexibility.

To address the above issues, we consider to search the ranking function in a data dependent hypothesis space, which is defined to be the linear combinations of basis functions. Inspired by the coefficient-based regression algorithms in [26,27], we propose a novel ranking algorithm with coefficient-based regularization. This regularization is dependent on the empirical data and the general kernel, where the kernel is not necessarily symmetric or positive semi-definite. This characteristic tells us that the theoretical analysis in [17,22] is not valid directly in the current setting. In particular, without the reproducing property of kernel, the representer theorem should be rebuilt. In this paper, we establish its representer theorem and generalization error bound by analyzing the data dependent characteristics of the proposed algorithm.

In summary, the main contributions of this paper can be highlighted as follows:

- A novel ranking algorithm is proposed, which finds the ranking function in a data dependent hypothesis space. The representer

* Corresponding author.
** Principal corresponding author.
E-mail addresses: chenh@mail.hzau.edu.cn (H. Chen), zhibinpan2008@gmail.com (Z. Pan).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.032

theorem is provided for the proposed algorithm, which gives a simple and fast procedure to implement this algorithm.

- Generalization analysis of the proposed algorithm is established in terms of the capacity of the data dependent hypothesis spaces. To the best of our knowledge, this is the first time to touch the generalization ability for ranking with coefficient regularization.
- Experimental evaluation on public datasets demonstrates the effectiveness of the coefficient-based algorithm.

The remainder of this paper is organized as follows. In Section 2, we introduce the background of ranking and the coefficient-based algorithm. The generalization error analysis is established in Section 3 and the experimental evaluation is given in Section 4. We conclude this paper in Section 5.

2. Coefficient-based regularized least square ranking

Now we recall some basic concepts of ranking problem (see [17] and references therein for details). Let $\mathcal{X} \subset \mathbb{R}^n$ be a compact metric space and $\mathcal{Y} = [0, M]$ for some $M > 0$. The relation between the input $x \in \mathcal{X}$ and the output $y \in \mathcal{Y}$ is described by a probability distribution $\rho(x, y)$ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. x is to be ranked preferred over x' if $y - y' > 0$, and lower than x' if $y - y' < 0$. In particular, $y - y' = 0$ indicates no preference between the two inputs.

The least square ranking loss

$$\ell(f, z, z') = (y - y' - (f(x) - f(x')))^2,$$

is considered in this paper. The expected risk of a ranking function f is defined as

$$\mathcal{E}(f) = \int_{\mathcal{Z}}\int_{\mathcal{Z}} (y - y' - (f(x) - f(x')))^2 \, d\rho(x, y) \, d\rho(x', y').$$

Given samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$ independently drawn according to ρ, the least square ranking problem aims at finding a ranking function $f_{\mathbf{z}} : \mathcal{X} \to \mathbb{R}$ such that $\mathcal{E}(f)$ is as small as possible.

The ranking algorithm usually can be implemented by a Tikhonov regularization scheme associated with a Mercer kernel. We call a symmetric and positive semidefinite continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K(x, \cdot), K(x', \cdot) \rangle_K = K(x, x')$ (see [35,36]).

Given $\mathbf{z}$ and regularization parameter $\gamma > 0$, the following regularized ranking is introduced in [17]:

$$\tilde{f}_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \{\mathcal{E}_{\mathbf{z}}(f) + \gamma \|f\|_K^2\}, \tag{1}$$

where $\mathcal{E}_{\mathbf{z}}(f)$ is the empirical ranking risk defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (y_i - y_j - (f(x_i) - f(x_j)))^2.$$

Setting $\lambda = ((m-1)/2m)\gamma$, we can rewrite (1) as the following least square regularized ranking (LSRRank) (see [22,12])

$$\tilde{f}_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \{\tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda \|f\|_K^2\}, \tag{2}$$

where

$$\tilde{\mathcal{E}}_{\mathbf{z}}(f) = \frac{1}{m^2} \sum_{i,j=1}^{m} (y_i - y_j - (f(x_i) - f(x_j)))^2.$$

The generalization ability of (1) has been studied via algorithmic stability in [17] and operator approximation in [22]. In particular, the minimizer (2) admits a representation with the form (see [22])

$$\tilde{f}_{\mathbf{z},\lambda} = \sum_{i=1}^{m} \tilde{\alpha}_{\mathbf{z},i} K_{x_i}, \tilde{\alpha}_{\mathbf{z},i} \in \mathbb{R}.$$

In this paper, we consider a coefficient-based ranking scheme in a data dependent hypothesis space. We just require that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous and bounded function. The hypothesis space for the given sample $\mathbf{z}$ is defined as

$$\mathcal{H}_{K,\mathbf{z}} = \left\{ \sum_{i=1}^{m} \alpha_i K_{x_i} : \alpha = (\alpha_1, \ldots, \alpha_m)^T \in \mathbb{R}^m \right\}, \tag{3}$$

where $K_t(\cdot) = K(\cdot, t)$. Since each candidate function in $\mathcal{H}_{K,\mathbf{z}}$ is determined by the corresponding coefficients, we consider the following coefficient-based regularizer:

$$\Omega_{\mathbf{z}}(f) = m \sum_{i=1}^{m} \alpha_i^2, \quad f = \sum_{i=1}^{m} \alpha_i K_{x_i}.$$

The coefficient-based least square regularized ranking (CLSRRank) is defined as the following scheme:

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_{K,\mathbf{z}}} \{\tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f)\}. \tag{4}$$

Denote the output function $f_{\mathbf{z},\lambda} = \sum_{i=1}^m \alpha_{\mathbf{z},i} K_{x_i}$. Then, the coefficient vector $\alpha_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})$ is given by

$$\alpha_{\mathbf{z}} = \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m^2} \sum_{j,k=1}^{m} \left(y_j - y_k - \sum_{i=1}^{m} \alpha_i (K(x_j, x_i) - K(x_k, x_i)) \right)^2 + \lambda m \sum_{i=1}^{m} \alpha_i^2 \right\}. \tag{5}$$

There are two features for CLSRRank: one is the flexibility imposed by removing the symmetry and positive semi-definiteness for the kernel; the other is the efficiency on computation, where $\alpha_{\mathbf{z}}$ in (5) can be solved by a linear system of equations (see Theorem 1 as below).

Denote the matrix $[K(x_i, x_j)]_{i,j=1}^m$ by $K_{\mathbf{x}}$ and let $K_{\mathbf{x}}^i$ be the m-order matrix $[a_t]_{t=1}^m$, where $a_t = (K(x_i, x_t)), \ldots (K(x_i, x_t))^T$. Let $Y = (y_i)_{i=1}^m = (y_1, \ldots, y_m)^T$, $Y^i = (y_i, \ldots, y_i)^T$, and let I be an m-order unit matrix. Denote

$$A = \frac{2}{m} (K_{\mathbf{x}})^T K_{\mathbf{x}} + \lambda m I - \frac{1}{m^2} \sum_{i=1}^{m} (K_{\mathbf{x}}^i)^T K_{\mathbf{x}} - \frac{1}{m^2} \sum_{i=1}^{m} (K_{\mathbf{x}})^T K_{\mathbf{x}}^i \tag{6}$$

and

$$B = \frac{2}{m} (K_{\mathbf{x}})^T Y - \frac{1}{m^2} \sum_{i=1}^{m} (K_{\mathbf{x}}^i)^T Y - \frac{1}{m^2} \sum_{i=1}^{m} (K_{\mathbf{x}})^T Y^i. \tag{7}$$

Now we present the following representer theorem.

Theorem 1. *The minimizer $f_{\mathbf{z},\lambda}$ in (4) can be represented as*

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{m} \alpha_{\mathbf{z},i} K(x, x_i),$$

where $\alpha_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})^T \in \mathbb{R}^m$ is the unique solution of linear system

$$A\alpha = B. \tag{8}$$

Proof. Note that

$$\begin{aligned}
&\tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f) \\
&= \frac{2}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 - \frac{2}{m^2} \sum_{i,j=1}^{m} (y_i - f(x_i))(y_j - f(x_j)) + \lambda \Omega_{\mathbf{z}}(f) \\
&= \frac{2}{m} \|K_{\mathbf{x}}\alpha - Y\|_2^2 + \lambda m \alpha^T \alpha - \frac{2}{m^2} \sum_{i=1}^{m} (Y^i - K_{\mathbf{x}}^i \alpha)^T (Y - K_{\mathbf{x}}\alpha).
\end{aligned}$$