Neural Networks 53 (2014) 119–126

Contents lists available at ScienceDirect

Neural Networks

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

Extreme learning machine for ranking: Generalization analysis and applications

Hong Chen [a,*], Jiangtao Peng [b,c], Yicong Zhou [c], Luoqing Li [b], Zhibin Pan [a]

[a] College of Science, Huazhong Agricultural University, Wuhan 430070, China
[b] Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China
[c] Department of Computer and Information Science, University of Macau, Macau 999078, China

ARTICLE INFO

Article history:
Received 8 September 2013
Received in revised form 9 January 2014
Accepted 24 January 2014

Keywords:
Learning theory
Ranking
Extreme learning machine
Coefficient regularization
Generalization bound

ABSTRACT

The extreme learning machine (ELM) has attracted increasing attention recently with its successful applications in classification and regression. In this paper, we investigate the generalization performance of ELM-based ranking. A new regularized ranking algorithm is proposed based on the combinations of activation functions in ELM. The generalization analysis is established for the ELM-based ranking (ELMRank) in terms of the covering numbers of hypothesis space. Empirical results on the benchmark datasets show the competitive performance of the ELMRank over the state-of-the-art ranking methods.

1. Introduction

The extreme learning machine (ELM) proposed by Huang, Zhu, and Siew (2006) can be considered as a learning system like feedforward neural networks (FNNs). Compared with FNN, the main feature of ELM is that the hidden node parameters are independent not only with the training data but also with each other, and can be generated before seeing the training data (Huang, Wang, & Lan, 2011). Recently, extensive studies have been paid on the ELM-like learning system through empirical evaluations (Bueno-Crespoa, García-Laencinab, & Sancho-Gómez, 2013; Cao, Liu, & Park, 2013; Huang, Zhou, Ding, & Zhang, 2012; Wang, Cao, & Yuan, 2011) and theoretical analysis (Huang, Ding, & Zhou, 2010; Liu, Lin, & Xu, 2013; Zhang, Lan, Huang, & Xu, 2012).

The previous studies of ELM usually focus on the classification and regression problems. The natural question is: Is the ELM-like learning system suitable for other learning tasks? To the best of our knowledge, the generalization analysis for ranking under the ELM framework remains untouched. In this paper, we consider the generalization performance of ELM-based least square ranking.

The ranking problem has gained increasing attention in machine learning with the fast development of ranking techniques on searching engines and information retrieval. From different perspectives, many ranking algorithms have been proposed including RankSVM (Herbrich, Graepel, & Obermayer, 2000; Joachims, 2002), RankNet (Burges, Ragno, & Le, 2007; Burges et al., 2005), RankBoost (Freund, Iyer, Schapire, & Singer, 2003), and MPRank (Cortes, Mohri, & Rastogi, 2007). The generalization analysis for the ranking problem has been established via stability analysis (Agarwal & Niyogi, 2009; Cossock & Zhang, 2008), uniform convergence estimate based on the capacity of hypothesis spaces (Clemencon, Lugosi, & Vayatis, 2008; Rejchel, 2012; Rudin, 2009; Zhang & Cao, 2012), and approximation estimate based on the operator approximation (Chen, 2012; Chen et al., 2013).

In this paper, inspired by the theoretical analysis in Liu et al. (2013), we propose an ELM-based ranking (ELMRank) algorithm to search a ranking function in a coefficient-based regularization scheme. The representer theorem and generalization bound are established for the proposed algorithm. Because the random node function in ELM has flexible forms, we use the uniform convergence analysis based on covering numbers to establish the generalization bounds.

Now, we highlight some features of this paper.

- A new ranking algorithm, called ELMRank, is proposed based on the hypothesis space of ELM. The representer theorem is provided to show that ELMRank also inherits the computation feasibility of ELM.
- Generalization analysis of ELMRank is established in terms of the capacity of the hypothesis spaces. This extends the previous analysis for regression in Liu et al. (2013) to the ranking settings.

* Corresponding author. Tel.: +86 18971089571.
E-mail addresses: chenhongml@163.com, chenh@mail.hzau.edu.cn (H. Chen), pengjt1982@126.com (J. Peng), yicongzhou@umac.mo (Y. Zhou), lilq@hubu.edu.cn (L. Li), zhibinpan2008@gmail.com (Z. Pan).

http://dx.doi.org/10.1016/j.neunet.2014.01.015
0893-6080/

- Experiments on public datasets demonstrate the competitive ranking prediction performance of ELMRank.

The remainder of this paper is organized as follows. In Section 2, we introduce the ELM-based learning system for least square ranking. The representer theorem is also proved in this section. The generalization analysis is established in Section 3 and the experimental evaluations are given in Section 4. Finally, a brief conclusion is presented in Section 5.

2. ELM-based ranking

Now we recall some basic concepts of the ranking problem (Agarwal & Niyogi, 2009). Let $\mathcal{X} \in \mathbb{R}^d$ be a compact metric space and $\mathcal{Y} = [0, M]$ for some $M > 0$. A probability distribution ρ, defined on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, describes the relation between the input $x \in \mathcal{X}$ and the output $y \in \mathcal{Y}$. x is ranked higher than x' if $y > y'$, and lower than x' if $y < y'$. In particular, there is no ranking preference between x and x' if $y = y'$.

In this paper, the least square ranking loss

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

is used to describe the difference between $y - y'$ and $f(x) - f(x')$. The expected risk (also called the generalization error) 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 empirical ranking risk is defined as

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

The least square ranking aims at finding a function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathcal{E}(f)$ is as small as possible.

Following the kernel methods for classification and regression, many ranking algorithms are proposed under a Tikhonov regularization scheme associated with a Mercer kernel (Agarwal, Dugar, & Sengupt, 2010; Agarwal & Niyogi, 2009; Chen, 2012; Chen et al., 2013). The reproducing kernel Hilbert space (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 $\langle K(x, \cdot), K(x', \cdot) \rangle_K = K(x, x')$. Then, $\|f\|_K^2 = \sum_{i,j=1}^m \beta_i \beta_j K(x_i, x_j)$ for $f = \sum_{i=1}^m \beta_i K(x_i, \cdot) \in \mathcal{H}_K$.

Agarwal and Niyogi (2009) proposed the following regularized ranking algorithm:

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

where $\gamma > 0$ is the regularization parameter.

Let $\lambda = \frac{m-1}{2m}\gamma$ and

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

The regularized scheme (1) can be transformed as below:

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

It is worth noticing that the minimizer (2) admits a representation of the form (Chen, 2012)

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

Hence, the kernel-based regularized ranking focuses on searching the coefficients in a data dependent hypothesis space.

Inspired by the computation feasibility of ELM (Huang et al., 2012, 2006; Liu et al., 2013), in this paper we consider a regularized ranking scheme in an ELM-based hypothesis space. Let $\phi(\alpha_i, \cdot) : \mathbb{R}^d \to \mathbb{R}$ be the random node function for the hidden parameter $\alpha_i \in \mathbb{R}^l$ and $n \in \mathbb{N}$ be the number of hidden nodes. The ELM-based hypothesis space is defined as

$$\mathcal{M}_n = \left\{ \sum_{i=1}^{n} \beta_i \phi(\alpha_i, \cdot) : \alpha_i \in \mathbb{R}^l, \beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n \right\},$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^{n \times l}$ are randomly drawn from a uniform distribution μ in $\mathbb{R}^{n \times l}$. Here, $\mathcal{M}_n$ can be considered as a hypothesis space of three layer FNNs with n hidden nodes and one output node whose hidden connection is α and output connection is β (Huang et al., 2010; Liu et al., 2013). That is to say $\{\phi(\alpha_i, \cdot)\}_{i=1}^n$ map the first layer to the hidden layer and $\sum_{i=1}^n \beta_i \phi(\alpha_i, \cdot)$ forms the output layer by the output weights β. In ELM, the sigmoid and Gaussian functions are two popular random node functions.

For m training samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$, the output of the ELM-based ranking (ELMRank) with n hidden nodes is

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{M}_n} \left\{ \tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda \|f\|_{\ell_2}^2 \right\}, \tag{3}$$

where

$$\|f\|_{\ell_2}^2 = \inf \left\{ \sum_{i=1}^{n} \beta_i^2 : f = \sum_{i=1}^{n} \beta_i \phi(\alpha_i, \cdot) \right\}.$$

Denote $f_{\mathbf{z},\lambda} = \sum_{i=1}^n \beta_{\mathbf{z},i} \phi(\alpha_i, \cdot)$. From (3), we know that the output weights $\beta_{\mathbf{z}} = (\beta_{\mathbf{z},1}, \ldots, \beta_{\mathbf{z},n})^T$ can be determined by

$$\beta_{\mathbf{z}} = \arg\min_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{m^2} \sum_{j,k=1}^{m} \left(y_j - y_k - \left(\sum_{i=1}^{n} \beta_i \phi(\alpha_i, x_j) - \sum_{i=1}^{n} \beta_i \phi(\alpha_i, x_k) \right) \right)^2 + \lambda \sum_{i=1}^{n} \beta_i^2 \right\}. \tag{4}$$

Compared with the kernel-based regularized ranking, there are two key differences for ELMRank: one is that the parameter α of the hidden node is independent of the samples $\mathbf{z}$; the other is that ϕ is the activation function or its composition in the FNN framework.

Recently, ELM for learning to rank has been well discussed for relevance ranking (Zong & Huang, 2013). Although our paper is closely related with Zong and Huang (2013), there are two features for our analysis and applications: In theory, we establish the generalization bound of ELMRank which fills the gap on generalization analysis of ranking under the ELM framework; In applications, we focus on learning a score function for the recommendation task and drug discovery, while Zong and Huang (2013) consider the document retrieval via linear ranking models.

Let H be the hidden layer output $m \times n$ matrix $[\phi(\alpha_i, x_j)]$ and let H^i be the $m \times n$ matrix $[a_t]_{t=1}^n$, where $a_t = (\phi(\alpha_t, x_i), \ldots, \phi(\alpha_t, x_i))^T \in \mathbb{R}^m$. Let $Y = (y_i)_{i=1}^m = (y_1, \ldots, y_m)^T$ be the target vector, $Y^i = (y_i, \ldots, y_i)^T$, and let I_m be the m-order unit matrix. Denote

$$A = \frac{2}{m} H^T H + \lambda I_m - \frac{1}{m^2} \sum_{i=1}^{m} (H^i)^T H - \frac{1}{m^2} \sum_{i=1}^{m} H^T H^i \tag{5}$$

and

$$B = \frac{2}{m} H^T Y - \frac{1}{m^2} \sum_{i=1}^{m} (H^i)^T Y - \frac{1}{m^2} \sum_{i=1}^{m} H^T Y^i. \tag{6}$$