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

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

Now we present the following representer theorem.

Theorem 1. *The minimizer $f_{\mathbf{z},\lambda}$ in (3) can be represented as*

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{n} \beta_{\mathbf{z},i}\phi(\alpha_i, x),$$

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

$$A\beta = B. \tag{7}$$

Proof. Note that

$$\begin{aligned}\tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda\|f\|_{\ell_2}^2 &= \frac{2}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 \\ &\quad - \frac{2}{m^2}\sum_{i,j=1}^{m}(y_i - f(x_i))(y_j - f(x_j)) + \lambda\|f\|_{\ell_2}^2 \\ &= \frac{2}{m}\|H\beta - Y\|_2^2 - \frac{2}{m^2}\sum_{i=1}^{m}(Y^i - H^i\beta)^T(Y - H\beta) + \lambda\beta^T\beta.\end{aligned}$$

Then

$$\begin{aligned}&\frac{\partial(\tilde{\mathcal{E}}_{\mathbf{z}}(f) + \lambda\|f\|_{\ell_2}^2)}{\partial\beta} \\ &= \frac{4}{m}\left[H^TH\beta - H^TY\right] - \frac{2}{m^2}\sum_{i=1}^{m}(H^i)^TH\beta - \frac{2}{m^2}\sum_{i=1}^{m}H^TH^i\beta \\ &\quad + \frac{2}{m^2}\sum_{i=1}^{m}(H^i)^TY + \frac{2}{m^2}\sum_{i=1}^{m}H^TY^i + 2\lambda\beta.\end{aligned}$$

Setting $\frac{\partial(\tilde{\mathcal{E}}_{\mathbf{z}}(f)+\lambda\|f\|_{\ell_2}^2)}{\partial\beta} = 0$, we get the desired result.

Theorem 1 tells us that the ELMRank can be implemented by solving the linear system (7). To improve the stability of ELMRank, we have

$$\beta = \left(A^TA + \frac{I_n}{C}\right)^{-1}A^TB, \tag{8}$$

where C is a ridge regularization parameter.

3. Generalization analysis

In this section, we will investigate the generalization performance of ELMRank (3). Note that $E_{\mathbf{z}}\mathcal{E}_{\mathbf{z}}(f) = \frac{m}{m-1}E_{\mathbf{z}}\tilde{\mathcal{E}}_{\mathbf{z}}(f) = \mathcal{E}(f)$. In the following, we will establish the upper bounds of the excess risk $E_{\mathbf{z}}\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)$, where f^* is the minimizer of $\mathcal{E}(f)$ over the measurable function space. Without loss of generality, we assume that $\|f^*\|_\infty \le M$. Some discussions for the optimal ranking function f^* can be found in the literature (Chen, 2012; Chen et al., 2013; Hu, Fan, Wu, & Zhou, 2013).

In learning theory, the excess generalization error is usually decomposed into the sample error and approximation error. From the definition of $f_{\mathbf{z},\lambda}$, we can get the following error decomposition.

Proposition 1. *For any $\mathbf{z} \in \mathcal{Z}^m$, there holds*

$$\begin{aligned}&E_{\mathbf{z}}(\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*)) \\ &\le E_{\mathbf{z}}\left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f^*))\right\} \\ &\quad + E_{\mathbf{z}}\left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2\right\} \\ &:= E_{\mathbf{z}}S_1 + E_{\mathbf{z}}S_2.\end{aligned}$$

Here, we call $E_{\mathbf{z}}S_1$ and $E_{\mathbf{z}}S_2$ as the sample error and the approximation error respectively. Because $S_1 = \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f^*) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f^*))$ is dependent on the random samples $\mathbf{z}$, we need to measure the capacity of the hypothesis space to establish the upper bound of the sample error. In this paper, the covering number is introduced to measure the capacity of the hypothesis space. In fact, the covering number has been well studied in the literature (Chen, Wu, Ying, & Zhou, 2004; Cucker & Smale, 2002; Cucker & Zhou, 2007; Zhou, 2002, 2003).

Definition 1. For $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{H}, \eta)$ is defined to be the smallest integer $l \in \mathbb{N}$ such that there exist l disks in $C(\mathcal{X})$ with radius η and centers in $\mathcal{H}$ covering the set $\mathcal{H}$.

For given $R > 0$, we define a class of functions as

$$B_R = \left\{f \in \mathcal{M}_n : \|f\|_{\ell_2}^2 \le R^2\right\}. \tag{9}$$

Cucker and Smale (2002) present the following bound for the covering number of B_R (also see Liu et al., 2013).

Lemma 1. *For any $R > 0$ and $\eta > 0$, there holds*

$$\log \mathcal{N}(B_R, \eta) \le n\log\left(\frac{4R}{\eta}\right).$$

The McDiarmid inequality is introduced to establish the relationship between the expected risk and the empirical risk.

Lemma 2. *Let $\{x_i\}_{i=1}^m$ be independent random variables taking values in a set $\mathcal{A}$ and let $\{b_i\}_{i=1}^m$ be positive constants. Assume that $\varphi : \mathcal{A}^m \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_m,\tilde{x}_i\in\mathcal{A}} |\varphi(x_1, \ldots, x_i, \ldots, x_m) - \varphi(x_1, \ldots, \tilde{x}_i, \ldots, x_m)| \le b_i$$

for every $1 \le i \le m$. Then, for every $\varepsilon > 0$,

$$\text{Prob}\{\varphi(x_1, \ldots, x_m) - E\varphi \ge \varepsilon\} \le \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{m} b_i^2}\right\}.$$

The following inequality follows the characteristic of the least square ranking loss.

Lemma 3. *Assume that the node function $\phi(\alpha_i, x) \le \kappa < \infty$ for the randomly preselected α_i and all $x \in \mathcal{X}$, $i \in \{1, \ldots, n\}$. The output $y \in [0, M]$ for $M > 0$. Then, for all $f_1, f_2 \in B_R$, $z, z' \in \mathcal{Z}$, we have*

$$|\ell(f_1, z, z') - \ell(f_2, z, z')| \le 4(M + 2\kappa R)\|f_1 - f_2\|_\infty.$$

Proof. Based on the Cauchy–Schwarz inequality, we have

$$|f(x)| \le \kappa\|f\|_{\ell_2} \le \kappa R, \quad \forall f \in B_R,\ x \in \mathcal{X}.$$

For any $f_1, f_2 \in B_R$, there is

$$\begin{aligned}&|\ell(f_1, z, z') - \ell(f_2, z, z')| \\ &\le (2|y - y'| + |f_1(x)| + |f_1(x')| + |f_2(x)| + |f_2(x')|) \\ &\quad \times (|f_1(x) - f_2(x)| + |f_1(x') - f_2(x')|) \\ &\le 4(M + 2\kappa R)\|f_1 - f_2\|_\infty.\end{aligned}$$

This completes the proof.

Now we present the uniform convergence analysis for $f \in B_R$.

Lemma 4. *For any $\varepsilon > 0$, there holds*

$$\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\sup_{f\in B_R}(\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*)))\geq\varepsilon\Big\}$$
$$\leq \mathcal{N}\Big(B_R,\frac{\varepsilon}{16(M+2\kappa R)}\Big)\exp\Big\{-\frac{\varepsilon^2 m}{1024(M+\kappa R)^4}\Big\}.$$

Proof. Let $\mathbf{z}=\{z_i\}_{i=1}^m\in\mathcal{Z}^m$ and $\mathbf{z}^k=(z_1,\ldots,z_{k-1},z_k',z_{k+1},\ldots,z_m)$. Denote $\varphi(\mathbf{z})=\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*)$, then $E\varphi(\mathbf{z})=\mathcal{E}(f)-\mathcal{E}(f^*)$. Then, for any $1\leq k\leq m$ and $f\in B_R$, we have

$$|\varphi(\mathbf{z})-\varphi(\mathbf{z}^k)|\leq\frac{2}{m(m-1)}\sum_{j\neq k}\Big(|\ell(f,z_k,z_j)-\ell(f,z_k',z_j)|+|\ell(f^*,z_k,z_j)-\ell(f^*,z_k',z_j)|\Big)\leq\frac{4(M+2\kappa R)^2+18M^2}{m}.$$

According to the McDiarmid inequality, for any $\varepsilon>0$, we get

$$\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{E\varphi(\mathbf{z})-\varphi(\mathbf{z})\geq\varepsilon\Big\}\leq\exp\Big\{-\frac{\varepsilon^2 m}{256(M+\kappa R)^4}\Big\}.$$

Now we use the technique in Cucker and Smale (2002) to obtain the uniform convergence estimate. Let $J=\mathcal{N}(B_R,\frac{\varepsilon}{16(M+2\kappa R)})$ and f_j, $1\leq j\leq J$, be the centers of disks D_j with radius $\frac{\varepsilon}{16(M+2\kappa R)}$ such that $B_R\subset\cup_{j=1}^J D_j$. Note that, for all $f\in D_j$ and $\mathbf{z}\in\mathcal{Z}^m$,

$$|\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*))-\{\mathcal{E}(f_j)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f^*))\}|\leq 8(M+2\kappa R)\|f-f_j\|_\infty\leq\frac{\varepsilon}{2}.$$

Then,

$$\sup_{f\in D_j}(\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*)))\geq\varepsilon$$
$$\Rightarrow\mathcal{E}(f_j)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f^*))\geq\frac{\varepsilon}{2}.$$

That is to say

$$\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\sup_{f\in D_j}\Big(\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*))\Big)\geq\varepsilon\Big\}\tag{10}$$
$$\leq\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\mathcal{E}(f_j)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f^*))\geq\frac{\varepsilon}{2}\Big\}$$
$$\leq\exp\Big\{-\frac{\varepsilon^2 m}{1024(M+\kappa R)^4}\Big\}.\tag{11}$$

Note that

$$\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\sup_{f\in B_R}\Big(\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*))\Big)\geq\varepsilon\Big\}$$
$$\leq\sum_{j=1}^J\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\Big\{\sup_{f\in D_j}\Big(\mathcal{E}(f)-\mathcal{E}(f^*)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*))\Big)\geq\varepsilon\Big\}.\tag{12}$$

The desired result is obtained by combining (11) and (12).

It is a position to present the main result on the generalization bound.

Theorem 2. *Assume $\phi(\alpha_i,x)\leq\kappa<\infty$ for the randomly preselected α_i and all $x\in\mathcal{X}$, $i\in\{1,\ldots,n\}$, there exists a constant $\tilde{C}$ independent of m, n such that*

$$E_{\mathbf{z}}(\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f^*))\leq 32\tilde{C}M\Big(1+\kappa\lambda^{-\frac{1}{2}}\Big)^2\sqrt{\frac{n(\log m-\log(n-1))}{m}}+\inf_{f\in\mathcal{M}_n}\Big\{\mathcal{E}(f)-\mathcal{E}(f^*)+\lambda\|f\|_{\ell_2}^2\Big\}.$$

Proof. From the definition of $f_{\mathbf{z},\lambda}$ in (4), we get that

$$\|f_{\mathbf{z},\lambda}\|_{\ell^2}^2\leq\frac{\mathcal{E}_{\mathbf{z}}(0)}{\lambda}\leq\frac{M^2}{\lambda}.$$

Hence, $f_{\mathbf{z},\lambda}\in B_R$ with $R=\frac{M}{\sqrt{\lambda}}$.

Based on Lemma 1 and Lemma 4, we get

$$E_{\mathbf{z}}(S_1)=\int_0^t\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\{S_1\geq\varepsilon\}d\varepsilon+\int_t^\infty\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\{S_1\geq\varepsilon\}d\varepsilon$$
$$\leq t+\int_t^\infty\exp\Big\{n\log\frac{64R(M+2\kappa R)}{\varepsilon}-\frac{\varepsilon^2 m}{256(M+\kappa R)^4}\Big\}d\varepsilon$$
$$\leq t+\exp\Big\{-\frac{t^2 m}{256(M+\kappa R)^4}\Big\}\int_t^\infty\Big(\frac{64R(M+2\kappa R)}{\varepsilon}\Big)^n d\varepsilon$$
$$\leq t+\exp\Big\{-\frac{t^2 m}{256(M+\kappa R)^4}\Big\}\Big(\frac{64R(M+2\kappa R)}{mt}\Big)^n\frac{tm^n}{n-1}.$$

Choose

$$t=16c(M+\kappa R)^2\sqrt{\frac{n(\log m-\log(n-1))}{m}}$$

such that $t\geq\frac{64R(M+2\kappa R)}{m}$, where c is a constant independently of m, n. Then

$$E_{\mathbf{z}}(S_1)\leq 2t\leq 32c(M+\kappa R)^2\sqrt{\frac{n(\log m-\log(n-1))}{m}}.$$

Now we estimate $E_{\mathbf{z}}S_2$. There holds

$$E_{\mathbf{z}}(S_2)=E_{\mathbf{z}}\Big\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f^*)+\lambda\|f_{\mathbf{z},\lambda}\|_{\ell_2}^2\Big\}$$
$$=E_{\mathbf{z}}\inf_{f\in\mathcal{M}_n}\Big\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*)+\lambda\|f\|_{\ell_2}^2\Big\}$$
$$\leq\inf_{f\in\mathcal{M}_n}\Big\{\mathcal{E}(f)-\mathcal{E}(f^*)+\lambda\|f\|_{\ell_2}^2\Big\}.$$

We complete the proof by combining the estimates of $E_{\mathbf{z}}(S_1)$ and $E_{\mathbf{z}}(S_2)$ with Proposition 1.

From Theorem 2, we know that the generalization ability of (3) depends on the hypothesis space $\mathcal{M}_n$ and the intrinsic ranking rule f^*. In essential, we just give the explicit convergence rate on the sample error and leave the approximation error analysis for future study. When the optimal ranking function can be approximated by the functions in the ELM-based hypothesis space, we have $\mathcal{E}(f_{\mathbf{z},\lambda})\to\mathcal{E}(f^*)$ with order $(\frac{\log m}{m})^{\frac{1}{2}}$. This polynomial decay is satisfactory to ranking and similar with the convergence results for kernel-based ranking, see, e.g., $O(m^{-\frac{1}{2}})$ (Agarwal & Niyogi, 2009), $O(m^{-\frac{1}{5}})$ (Chen, 2012), $O(m^{-\frac{1}{4}})$ (Chen et al., 2013).

The ELMRank is associated with the hypothesis space $\mathcal{M}_n$ which is independent of training samples $\mathbf{z}$. Hence, the ELMRank has much lower computation complexity than the kernel-based ranking algorithms (Agarwal et al., 2010; Agarwal & Niyogi, 2009; Chen, 2012; Chen et al., 2013). That is to say the ELMRank inherits the computation advantages of ELM and its implementation is faster than the kernel-based regularized methods.

4. Experiments

In this section, the empirical performance of ELMRank is verified on several benchmark datasets for the recommendation task and drug discovery. The experimental results show ELMRank can achieve competitive performance compared with several state-of-the-art ranking algorithms.

4.1. Algorithm and parameter selection

From Theorem 1, we know that ELMRank can be implemented easily through the linear system (7) and (8). The explicit computation steps of the ELMRank are summarized in Algorithm 1.

Algorithm 1 ELMRank

Require:

Training set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, activation function ϕ, hidden node number n, ridge parameter C, and regularization parameter $\lambda > 0$.

1: Assigning parameter $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ randomly and generating matrix H.
2: Computing the matrices A and B defined in (5) and (6).
3: Solving the linear system $A\beta = B$ by (8) to derive $\beta_{\mathbf{z}} = (\beta_{\mathbf{z},1}, \ldots, \beta_{\mathbf{z},n}) \in \mathbb{R}^n$.
4: **return** A ranking function

$$f_{\mathbf{z},\lambda}(x) = \sum_{i=1}^{n} \beta_{\mathbf{z},i}\phi(\alpha_i, x).$$

In the experiments, we adopt the sigmoid function as the nonlinear activation function. The hidden node number is set as $n = 100$. All the hidden-node parameters $\{\alpha_i\}_{i=1}^n$ are randomly generated with the uniform distribution.

The user-specified parameters are regularization parameter λ and ridge parameter C, where the regularization parameter λ is selected from $\{10^{-5}, \ldots, 10^{1}\}$ and C is chosen from the range $\{2^1, \ldots, 2^{25}\}$. The optimal parameter values are chosen based on 10-fold cross validation. Average results of 50 trials for each fixed ELM are reported in this paper.

Some remarks for the parameter selections are given as below. In (3), $\lambda\|f\|_{\ell_2}^2$ is used to restrict the function f. In general, λ is a small value such that the regularization part plays an auxiliary role in choosing a feasible f compared with the empirical term $\tilde{\mathcal{E}}_{\mathbf{z}}(f)$. The same selection for λ is also given in Agarwal et al. (2010). The regularization parameter C is used in (8) to solve the linear system of ELMRank. In solving the linear system, the ridge parameter $1/C$ makes the solution more stable but introducing bias. So, the ridge parameter $1/C$ should also be a small value while still maintaining model stability. As we know, in the original ELM, $1/C$ is chosen in a width range $\{2^{-25}, \ldots, 2^{25}\}$. For our ELMRank, in the experiments, we have found that better results are usually achieved at small ridge parameter $1/C$ (usually smaller than 1). For simplicity, we choose $1/C$ from the left half of the whole set, that is, C varies in the range $\{2^1, \ldots, 2^{25}\}$.

Table 1
Comparison of MSD (mean and standard deviation).

Dataset	MPRank	SVRank	ELMRank
MovieLens 20–40	2.01 ± 0.02	2.43 ± 0.13	**1.98** ± 0.04
MovieLens 40–60	2.02 ± 0.06	2.36 ± 0.16	**2.00** ± 0.04
MovieLens 60–80	2.07 ± 0.05	2.66 ± 0.09	**1.96** ± 0.04
Jester 20–40	51.34 ± 2.90	55.00 ± 5.14	**37.64** ± 1.37
Jester 40–60	46.77 ± 2.03	57.75 ± 5.14	**38.98** ± 1.61
Jester 60–80	49.33 ± 3.11	56.06 ± 4.26	**34.30** ± 1.21
Books	4.00 ± 3.12	3.64 ± 3.04	**2.81** ± 3.44

4.2. Experiments on the recommendation task

The recommendation task aims to produce for a given user a list of unseen movies/jokes/books ordered by the predicted preference. The ELMRank is compared with MPRank (Cortes et al., 2007), SVRank (Cortes et al., 2007) and RankBoost (Freund et al., 2003).

4.2.1. Datasets and experimental settings

The MovieLens dataset contains 1,000,209 anonymous ratings of 3883 movies made by 6040 MovieLens users, where rating belongs to $\{1, \ldots, 5\}$ and not all movies are rated. The Jester Joke Recommender System dataset contains 4.1 Million continuous ratings ranging from -10.00 to $+10.00$ to 100 jokes from 73,421 users. The book-crossing dataset contains 278,858 users and 1,149,780 ratings for 271,379 books.

For the MovieLens dataset, the reviewers are divided into three groups (20–40 movies, 40–60 movies, and 60–80 movies) according to their numbers of reviewed movies. The reviewers, reviewed between 50 and 300 movies, are used for testing. For a given test reviewer, we randomly select 300 reference reviewers from one of the three groups and their rating scores are used to form the input vectors, and use half of his rated movies for training and the other for testing. The average performance is obtained from 300 different test reviewers. The mean values and standard deviations are derived after ten repeated experiments for each of the three groups. For the Jester Joke Recommender System dataset, its experiment procedure is similar to the MovieLens dataset.

For the book-crossing dataset, only those users who have reviewed at least 200 books, and books with at least 10 reviews are considered in our experiment. This gives us a dataset including 87 books and 130 reviewers. Then, each of the 130 reviewers is selected by turn as a test reviewer, and the rest 129 reviewers are considered as the input features. The mean values and standard deviations are reported over these 130 leave-one-out experiments.

4.2.2. Experimental results

The prediction performance of ELMRank is evaluated by three measures including the mean squared difference (MSD), Mean 1-norm Difference (M1D), Pairwise Misranking Error (Cortes et al., 2007). Let $T = \{(x_i, y_i)\}_{i=1}^{m'}$ be the test set and f be the prediction function. It is worth noticing that MSD defined by

$$\tilde{\mathcal{E}}_T(f) = \frac{1}{m'^2}\sum_{i,j=1}^{m'} \big(y_i - y_j - (f(x_i) - f(x_j))\big)^2$$

is the empirical risk on the test set and reflects the generalization performance of f. In essential, our theoretical analysis in Theorem 2 reflects the generalization performance according to MSD. The other measures of ranking performance are closely related with MSD although they have different characteristics.

According to the above measures, we report the experimental results in Tables 1–3 respectively. The results of MPRank, SVRank and RankBoost come from Cortes et al. (2007).

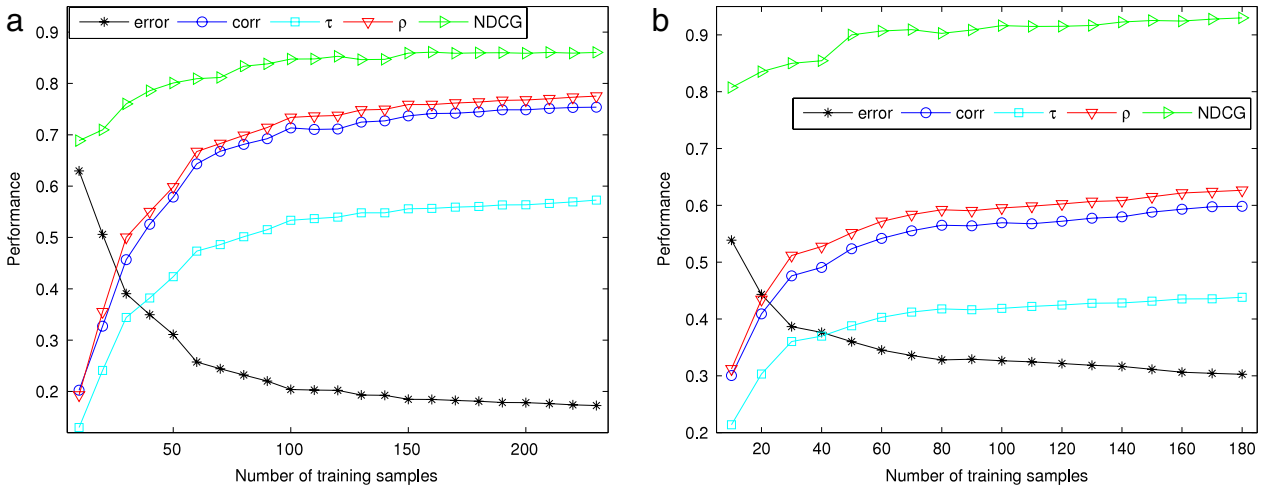

Fig. 1. The ELMRank performance vs. the number of training samples on DHFR (a) and COX2 (b).

Table 2
Comparison of M1D (mean and standard deviation).

Dataset	MPRank	SVRank	ELMRank
MovieLens 20–40	1.04 ± 0.05	1.17 ± 0.03	**1.03** ± 0.02
MovieLens 40–60	1.04 ± 0.02	1.15 ± 0.07	**1.03** ± 0.01
MovieLens 60–80	**1.06** ± 0.01	1.24 ± 0.02	1.07 ± 0.01
Jester 20–40	5.08 ± 0.15	5.40 ± 0.20	**4.57** ± 0.09
Jester 40–60	4.98 ± 0.13	5.27 ± 0.20	**4.63** ± 0.09
Jester 60–80	4.88 ± 0.14	5.25 ± 0.19	**4.37** ± 0.08
Books	1.38 ± 0.60	1.32 ± 0.56	**0.97** ± 0.77

Table 3
Comparison of pairwise misrankings (mean and standard deviation).

Dataset	MPRank	RankBoost	ELMRank
MovieLens 40–60	47.1% ± 0.5%	47.6% ± 0.7%	**45.5**% ± 0.4%
MovieLens 60–80	44.2% ± 0.5%	46.3% ± 1.1%	**43.6**% ± 0.5%
Jester 20–40	**41.0**% ± 0.6%	47.9% ± 0.8%	41.3% ± 0.6%
Jester 40–60	40.8% ± 0.6%	43.2% ± 0.5%	**39.3**% ± 0.5%
Jester 60–80	**37.1**% ± 0.6%	41.7% ± 0.8%	37.5% ± 0.5%

We can observe that ELMRank outperforms other algorithms on almost all datasets according to the measures of MSD and M1D, which verifies the effectiveness of ELMRank. For the percentage of pairwise misrankings, ELMRank has the best performance on MovieLens and loses to the MPRank on Jester Jokes. Note that ELMRank in (3) is designed to search a function in the ELM-based hypothesis space to minimize the least square ranking loss. In theory, it is a reason to induce the weak performance on the pairwise misranking loss. In fact, the selection of convex loss depends on the characteristics of ranking tasks and the hypothesis spaces. Because our main concern in this work is to investigate the generalization performance of ELMRank for the least square ranking, we leave the selection of different ranking loss functions for future study.

4.3. Experiments on QSAR analysis

We evaluate the prediction performance of ELMRank on two Quantitative Structure–Activity Relationship (QSAR) datasets, including inhibitors of dihydrofolate reductase (DHFR) and cyclooxygenase-2 (COX2). The prediction performance of ELMRank is compared with the RankSVM (Agarwal et al., 2010; Joachims, 2002) and Support Vector Regression (SVR)-based ranking (Agarwal et al., 2010; Vapnik, 1998).

4.3.1. Datasets and experimental settings

The DHFR inhibitor dataset contains 361 compounds, with pIC50 values ranging from 3.3 to 9.8; the COX2 inhibitor dataset contains 282 compounds, with pIC50 values ranging from 4.0 to 9.0. In the original DHFR dataset, 237 out of 361 compounds are selected as the training set and the rest compounds are considered as the test set. For the COX2 dataset, 188 of 292 compounds form the training set and the remaining compounds are used as the test set. In these datasets, each compound is represented by the 2.5D chemical descriptors (Sutherland, O'Brien, & Weaver, 2004). See Sutherland et al. (2004) and the references therein for details.

70 real-valued descriptors are contained in the DHFR inhibitor dataset and 74 real-valued descriptors are contained in the COX2 inhibitor dataset, where each of these descriptors is scaled to lie between 0 and 1. The experimental set-ups here follow those in Agarwal et al. (2010).

4.3.2. Experimental results

To better describe the performance of ELMRank, we investigate different measures including the ranking error, correlation, Kendall's τ ranking correlation coefficient, Spearman's ρ rank correlation coefficient, and normalized discounted cumulative gain (NDCG). The detail definitions of these measures can be found in Section 3.5 of Agarwal et al. (2010).

According to these measures, we report the results of ELMRank in Table 4. The results of RankSVM and SVR come from Agarwal et al. (2010).

From these experimental results, we can see that ELMRank has the best performance. These results further verify the effectiveness of ELMRank. We also investigate the relation between the ELMRank's performance and the number of training samples in Fig. 1. Ten samples are selected as the training set randomly from the original training dataset. Then we test it on the original test dataset. All performance evaluations are recorded. We train ELMRank by adding 10 training samples each time. The reported results are the average of ten-time evaluations with randomly chosen training samples. It can be seen from Fig. 1 that ELMRank shows the consistent performance under different training sample sizes.

Note that MSD is the empirical version of the generalization error $\mathcal{E}(f)$. To better verify the theoretical analysis in Theorem 2, we compare the prediction performance between ELMRank and the least square support vector regression (LSSVR) (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002) in terms of MSD under different numbers of training samples. The experiment results are presented in Fig. 2. It can be clearly seen

Table 4
QSAR ranking results on the DHFR and COX2 dataset.

	DHFR			COX2		
	SVR	RankSVM	ELMRank	SVR	RankSVM	ELMRank
Ranking error	0.1837	0.1726	**0.1699**	0.3138	0.3173	**0.3010**
Correlation	0.7519	**0.7618**	0.7567	0.5836	0.5703	**0.6062**
Kendall's τ	0.5571	0.5747	**0.5762**	0.4351	0.4346	**0.4390**
Spearman's ρ	0.7752	0.7758	**0.7784**	0.6100	0.6174	**0.6287**
NDCG	0.8540	0.8632	**0.8642**	**0.9399**	0.9231	0.9225

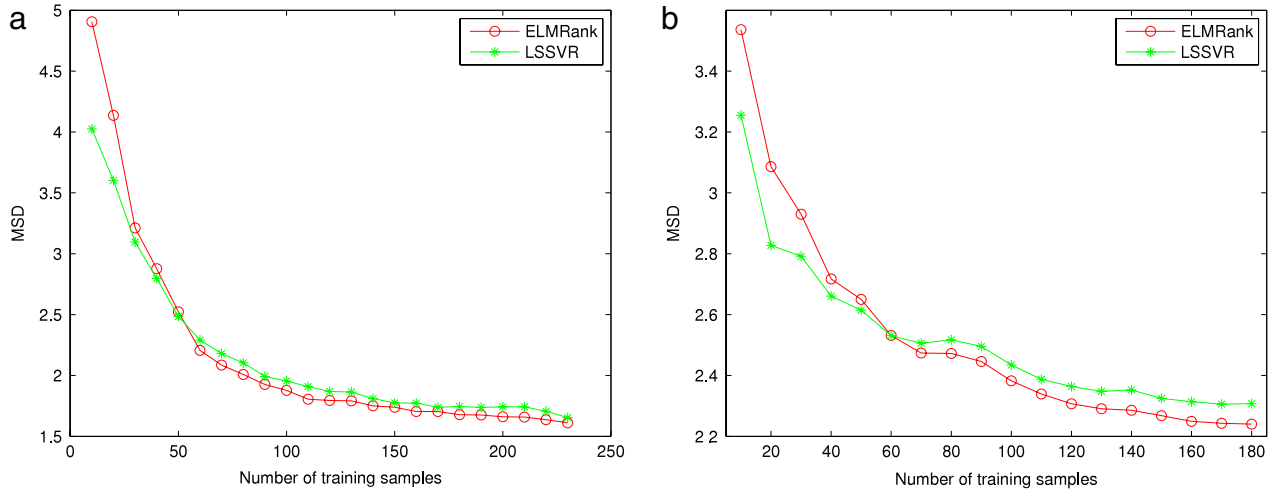

Fig. 2. ELMRank vs. LSSVR w.r.t. MSD on (a) DHFR and (b) COX2.

that the test errors decrease rapidly with the increasing number of training samples. This is consistent with the theoretical results in Theorem 2. Moreover, the superiority of ELMRank over LSSVR is consistent when enough training samples are available.

5. Conclusion

This paper investigated the convergence performance of ELMRank. In theory, we established the generalization bound of ELMRank and showed that satisfactory learning rates can be obtained under mild conditions. In applications, we evaluated the prediction performance of ELMRank on the public datasets and demonstrated its competitive performance compared with state-of-the-art algorithms. Along the line of the present work, further studies may consider to establish the generalization analysis of ELMRank with dependent samples (Zou, Li, & Xu, 2009; Zou, Li, Xu, Luo, & Tang, 2013) and with different regularization terms (Chen, Pan, Li, & Tang, 2013; Li, Chen, & Li, 2012; Tong, Chen, & Yang, 2012; Xu, Chang, Xu, & Zhang, 2012).

Acknowledgments

This work was supported partially by the National Natural Science Foundation of China under Grant Nos. 11001092, 11371007, 61300143, and by the Fundamental Research Funds for the Central Universities (Program No. 2011PY130), and by the Macau Science and Technology Development Fund (FDCT) under Grant 017/2012/A1, the Research Committee at University of Macau under Grants MYRG113(Y1-L3)-FST12-ZYC, and MRG001/ZYC/2013/FST.

References

Agarwal, S., Dugar, D., & Sengupt, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of Chemical Information and Modeling*, *50*(5), 716–731.

Agarwal, S., & Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, *10*, 441–474.

Bueno-Crespoa, A., García-Laencinab, P. J., & Sancho-Gómez, J. (2013). Neural architecture design based on extreme learning machine. *Neural Networks*, *48*, 19–24.

Burges, C., Ragno, R., & Le, Q. (2007). Learning to rank with nonsmooth cost functions. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19*. MIT Press.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., & Hamilton, N. et al. (2005). Learning to rank using gradient descent. In *Proc. 22th international conference on machine learning*.

Cao, F. L., Liu, B., & Park, D. S. (2013). Image classification based on effective extreme learning machine. *Neurocomputing*, *102*, 90–97.

Chen, H. (2012). The convergence rate of a regularized ranking algorithm. *Journal of Approximation Theory*, *164*(12), 1513–1519.

Chen, H., Pan, Z. B., Li, L. Q., & Tang, Y. Y. (2013). Learning rates of coefficient-based regularized classifier for density level detection. *Neural Computation*, *25*(4), 1107–1121.

Chen, H., Tang, Y., Li, L. Q., Yuan, Y., Li, X., & Tang, Y. Y. (2013). Error analysis of stachastic gradient descent ranking. *IEEE Transactions on Cybernetics*, *43*(3), 898–909.

Chen, D. R., Wu, Q., Ying, Y., & Zhou, D. X. (2004). Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, *5*, 1143–1175.

Clemencon, S., Luogosi, G., & Vayatis, N. (2008). Ranking and empirical minimization of *U*-satistics. *The Annals of Statistics*, *36*, 844–874.

Cortes, C., Mohri, M., & Rastogi, A. (2007). Magnitude-preserving ranking algorithms. In *Proc. 24th international conference on machine learning*.

Cossock, D., & Zhang, T. (2008). Statistical analysis of Bayes optimal subset ranking. *IEEE Transaction on Information Theory*, *54*, 5140–5154.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *American Mathematical Society. Bulletin*, *39*, 1–49.

Cucker, F., & Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*. Cambridge, UK: Cambridge Univ. Press.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933–969.

Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In *Advances in large margin classifiers* (pp. 115–132).

Hu, T., Fan, J., Wu, Q., & Zhou, D. X. (2013). Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, *14*, 377–397.

Huang, G. B., Ding, X., & Zhou, H. (2010). Optimization method based extreme learning machine for classification. *Neurocomputing*, *74*, 155–163.

Huang, G. B., Wang, D. H., & Lan, Y. (2011). Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, *2*(2), 107–122.

Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(2), 513–529.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, *70*(1), 489–501.
Joachims, T. (2002). Optimizing search engines using click through data. In *Proc. in eighth ACM SIGKDD int'l conf. knowledge discovery and data mining* (pp. 133–142).
Li, H., Chen, N., & Li, L. Q. (2012). Error analysis for matrix elastic-net regularization algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(5), 737–748.
Liu, X., Lin, S., & Xu, Z.B. (2013). Is extreme learning machine feasible? A theoretical assessment. Preprint.
Rejchel, W. (2012). On ranking and generalization bounds. *Journal of Machine Learning Research*, *13*, 1373–1392.
Rudin, C. (2009). The *P*-norm push: a simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, *10*, 2233–2271.
Sutherland, J. J., O'Brien, L. A., & Weaver, D. F. (2004). A comparison of methods for modeling quantitative structure—activity relationships. *Journal of Medicinal Chemistry*, *22*, 5541–5554.
Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.
Tong, H. Z., Chen, D. R., & Yang, F. (2012). Least square regression with ℓ^p-coefficient regularization. *Neural Computation*, *22*, 3221–3235.
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Wang, Y., Cao, F. L., & Yuan, Y. (2011). A study on effectiveness of extreme learning machine. *Neurocomputing*, *74*(16), 2483–2490.
Xu, Z. B., Chang, X. Y., Xu, F. M., & Zhang, H. (2012). $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(7), 1013–1027.
Zhang, Y. Q., & Cao, F. L. (2012). Analysis of convergence performance of neural networks ranking algorithm. *Neural Networks*, *34*, 65–71.
Zhang, R., Lan, Y., Huang, G. B., & Xu, Z. B. (2012). Universal approximation of extreme learning machine with adaptive growth of hidden nodes. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(2), 365–371.
Zhou, D. X. (2002). The covering number in learning theory. *Journal of Complexity*, *18*, 739–767.
Zhou, D. X. (2003). Capacity of reproducing kernel spaces in learning theory. *IEEE Transaction on Information Theory*, *49*(7), 1743–1752.
Zong, W., & Huang, G. B. (2013). Learning to rank with extreme learning machine. *Neural Processing Letters*. http://dx.doi.org/10.1007/s11063-013-9295-8.
Zou, B., Li, L. Q., & Xu, Z. B. (2009). The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, *75*, 275–295.
Zou, B., Li, L. Q., Xu, Z. B., Luo, T., & Tang, Y. Y. (2013). Generalization performance of Fisher linear discriminant based on Markov sampling. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(2), 288–300.