Neural Networks 34 (2012) 65–71

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Analysis of convergence performance of neural networks ranking algorithm☆

Yongquan Zhang, Feilong Cao*

Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou 310018, Zhejiang Province, PR China

ARTICLE INFO

Article history:
Received 15 September 2011
Received in revised form 31 March 2012
Accepted 29 June 2012

Keywords:
Ranking algorithm
Neural networks
Covering number
Convergence rate

ABSTRACT

The ranking problem is to learn a real-valued function which gives rise to a ranking over an instance space, which has gained much attention in machine learning in recent years. This article gives analysis of the convergence performance of neural networks ranking algorithm by means of the given samples and approximation property of neural networks. The upper bounds of convergence rate provided by our results can be considerably tight and independent of the dimension of input space when the target function satisfies some smooth condition. The obtained results imply that neural networks are able to adapt to ranking function in the instance space. Hence the obtained results are able to circumvent the curse of dimensionality on some smooth condition.

1. Introduction

The analysis of convergence performance of learning algorithm is an important and hot topic in machine learning research. To our knowledge, Vapnik and Chervonenkis (1971) first started to study the learning algorithm and established the analysis of convergence for classification algorithm from the statistical analysis. Since then, more different tools have been used to study the convergence performance of learning algorithms and have been applied to both classification (learning of binary-valued functions) and regression (learning of real-valued functions). In many learning algorithms, the goal is not simply classifying objects into one of a fixed number of classes; instead, a ranking of objects is desired. For example, in information retrieval problems, where one likes to retrieve documents from some databases that are 'relevant' to a given query or topic. In such problems, one needs a ranking of the documents so that relevant documents are ranked higher than irrelevant documents. Recently, the ranking problem has gained much attention in machine learning (see Agarwal & Niyogi, 2005, 2009; Clemencon, Lugosi, & Vayatis, 2008; Cohen, Schapire, & Singer, 1999; Cortes, Mohri, & Rastogi, 2007; Cossock & Zhang, 2006; Cucker & Smale, 2001, 2002). For ranking problem, we learn a real-valued function which gives scores to instances; however, these scores themselves do not matter; instead, we are only interested in the relative ranking of instances which are given by these scores.

Now, the ranking has been successfully applied to all kinds of fields, such as social choice theory (Kenneth, 1970), statistics (Lehmann, 1975) and mathematical economics (Chiang & Wainwright, 2005). However, in 1999 Cohen et al. (1999) first began to study the ranking in machine learning. From then on, many researchers started to pay attention to it and study the interesting topic from machine learning view, for example, Crammer and Singer (2002) and Herbrich, Graepel, and Obermayer (2000) considered the related ranking but distinct problem of ordinal regression. Radlinski and Joachims (2005) developed an algorithmic framework for ranking in information retrieval applications. Both Agarwal and Niyogi (2005) and Freund, Iyer, Schapire, and Singer (2003) have considered the convergence properties of ranking algorithms for the special setting of bipartite ranking respectively. Clemencon et al. (2008) have given statistical convergence properties of ranking algorithms based on empirical and convex risk minimization by using the theory of U-statistics. Agarwal and Niyogi (2009) studied the convergence properties of ranking algorithms in a more general setting of the ranking problem that arise frequently in applications and convergence error via ranking algorithmic stability. Burges et al. (2005) have developed a neural network based on the algorithm of ranking problem. Although there have been several recent advances in developing algorithms for various settings of the ranking problem, the study of generalization properties of ranking algorithms has been largely limited to the special setting of bipartite ranking (see Agarwal & Niyogi, 2005, Freund et al., 2003). Similar to Agarwal and Niyogi (2009), we study the convergence property of ranking learning algorithms in a more general setting of the ranking problem that arises frequently in applications and practice. Our convergence rates are derived by using the approximation property of neural networks and covering number instead of the notion of algorithmic stability in reproducing kernel Hilbert space in Agarwal and Niyogi (2009).

☆ This research was supported by the National Natural Science Foundation of China (No. 61101240) and the Zhejaing Provincial Nature Science Foundation of China (Nos Y6110117, Q12A010096).

* Corresponding author. Fax: +86 571 86835737.
E-mail address: feilongcao@gmail.com (F. Cao).

0893-6080/$ – see front matter
doi:10.1016/j.neunet.2012.06.012

Similar to both classification and regression, the ranking problem takes place in some hypothesis space which has good approximation property for a ranking function. It is well known that feedforward neural networks (FNNs) have universal approximation property for any continuous or integrable functions defined on a compact set; there are some algorithms to carry out the approximation. In 1989, Cybenko (1989) first proved that if the activation function in FNNs is a continuous sigmoidal function, and $I = [0, 1]^d$ is a unit cube in $\mathcal{R}^d$, then any continuous function on $I = [0, 1]^d$ is approximated by FNNs. Since then, some different methods from Cybenko (1989) have been designed. Meanwhile, a series of investigations into the condition of activation function ensuring the validity of the density theorem can be found in Chen and Chen (1995a, 1995b), Chen, Chen, and Liu (1995), Hornik (1991), and Mhaskar and Micchelli (1992). The complexity of FNN approximation mainly describes the relationship among the topology structure of hidden layer (such as the number of neurons and the value of weights), the approximation ability and the approximation rate. The study of complexity has attracted much attention in recent years (Cao, Zhang, & He, 2009; Cao, Zhang, & Xu, 2009; Chui & Li, 1992; Maiorov & Meir, 1998; Xu & Cao, 2005). In the study of machine learning, FNNs are usually used as hypothesis space to study the convergence performance of learning algorithm. For example, Barron (1993) gave the convergence rate of least square regression learning algorithm by the approximation property of FNNs. In 2006, Hamers and Kohler (2006) obtained nonasymptotic bounds on the least square regression estimates by minimizing the empirical risk over suitable sets of FNNs. Recently, Kohler and Mehnert (2011) gave an analysis of the rate of convergence of least squares learning algorithm in FNNs for smooth regression function. In this article, we study ranking learning algorithm by using neural networks, where the hypothesis space is chosen as a class of FNNs with one hidden layer.

The article is organized into six sections. Following the introduction in the present section, we describe general ranking problem in a more general setting and introduce neural networks in Section 2. In Section 3, we give approximation error of ranking algorithm by the approximation property of neural networks. Section 4 estimates the sample error. The obtained upper bound in connection with the approximation error leads to estimating the upper bound of convergence rate of neural networks ranking algorithm. In Section 5, we compare our results with the known related work. Finally, we conclude the article with the obtained results.

2. General ranking problem and neural networks

For the ranking problem, one is given some samples of ordering relationships among instances in some instance space X, and the goal is to learn a real-value function from these samples that ranks future instances. The ranking problems arise in all kinds of domains: in user-preference modeling, one wants to order movies or texts according to likes of oneself, and in information retrieval, where one is interested in retrieving documents from some databases that are 'relevant' to a given query or topic. In such problems, one wants to return a list of documents that contains relevant documents at the top and irrelevant documents at the bottom to the user; in other words, one wants a ranking of the documents so that relevant documents are ranked higher than irrelevant documents. In this article, we let $X \subset B_r = \{x \in \mathcal{R}^d, \|x\|_2 \le r\}, Y = [0, M_0] \subset \mathcal{R}$ for two positive constants r, M_0, and let ρ be a probability distribution on $Z = X \times Y$, and ρ_X, $\rho(y|x)$ are marginal probability and conditional probability on Z respectively. Denote $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ a set of labeled samples according to ρ. The goal of the ranking problem is to learn a real-valued function f ($f : X \to Y$) which orders exactly the other instances in X according to random samples. The function f is considered to rank x lower than x' if $f(x) < f(x')$, and higher than if $f(x) > f(x')$. The penalty for mistakenly ranking a pair of instances can be taken greater for mistakenly ranking a pair of instances with a larger difference between their labels. We introduce the ranking loss function in the following definition.

Definition 1 (*See Agarwal and Niyogi (2009)*)**.** A ranking loss function is a function $\ell : X^Y \times (X \times Y) \times (X \times Y) \to \mathcal{R}^+ \cup \{0\}$ that assigns to each real function $f : X \to Y$, and $(x, y), (x', y') \in (X, Y)$ a non-negative real number $\ell(f, (x, y), (x', y'))$.

The ranking loss function can be interpreted as the penalty of f in its relative ranking of two instances x and x' given their corresponding labels y and y'. We shall require that the loss function ℓ be symmetric with respect to (x, y) and (x', y'), that is $\ell(f, (x, y), (x', y')) = \ell(f, (x', y'), (x, y))$ for all f, (x, y) and (x', y'). Several ranking loss functions are useful in the study of ranking problem as follows:

(1) the 0–1 ranking loss function:

$$\ell_{0\text{-}1}(f, (x, y), (x', y')) = \mathbf{I}_{(y-y')(f(x)-f(x'))<0} + \frac{1}{2}\mathbf{I}_{f(x)=f(x')};$$

(2) the least squares ranking loss function:

$$\ell_{sq}(f, (x, y), (x', y')) = \left(|y - y'| - \operatorname{sgn}(y - y')(f(x) - f(x'))\right)^2;$$

(3) the discrete ranking loss function:

$$\ell_{\text{disc}}(f, (x, y), (x', y')) = |y - y'|\left(\mathbf{I}_{(y-y')(f(x)-f(x'))<0} + \frac{1}{2}\mathbf{I}_{f(x)=f(x')}\right);$$

(4) for $\gamma > 0$, the γ-ranking loss function is defined as follows:

$$\ell_\gamma(f, (x, y), (x', y')) = \begin{cases} |y - y'|, \\ \quad \text{if } \dfrac{(f(x) - f(x')) \times \operatorname{sgn}(y - y')}{\gamma} \le 0, \\ |y - y'| - \dfrac{(f(x) - f(x')) \times \operatorname{sgn}(y - y')}{\gamma}, \\ \quad \text{if } 0 < \dfrac{(f(x) - f(x')) \times \operatorname{sgn}(y - y')}{\gamma} < |y - y'|, \\ 0, \quad \text{if } \dfrac{(f(x) - f(x')) \times \operatorname{sgn}(y - y')}{\gamma} \ge |y - y'|, \end{cases}$$

where for $u \in \mathcal{R}$,

$$\operatorname{sgn}(u) = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u = 0, \\ -1, & \text{if } u < 0. \end{cases}$$

According to the definitions of both the discrete ranking loss function and the γ-ranking loss function, we know that for all $\gamma > 0$, there holds $\ell_{\text{disc}} \le \ell_\gamma$.

By using the ranking loss, we define the expected ℓ-*error* of the function f:

$$\mathcal{E}(f) = R_\ell(f) = \mathbf{E}_{(x,y),(x',y')\sim\rho\times\rho}\ell(f, (x, y), (x', y')). \tag{1}$$

The function f_ρ^ℓ that minimizes the error (1) is given by

$$f_\rho^\ell = \arg\min_{f:X\to Y} \mathbf{E}_{(x,y),(x',y')\sim\rho\times\rho}\ell(f, (x, y), (x', y')),$$

where the minimum is taken over all measurable functions. In the article we always assume that f_ρ^ℓ exists and satisfies $|f_\rho^\ell(x)| \le \log m$ for all $x \in X$.

The corresponding empirical ranking error of expected ranking ℓ-error is defined as follows:

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