Neurocomputing 119 (2013) 434–438

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Generalization performance of support vector classifiers for density level detection

Hong Chen [a], Yicong Zhou [b], Yi Tang [c], Yuan Yan Tang [b], Zhibin Pan [a,*]

[a] College of Science, Huazhong Agricultural University, Wuhan 430070, China
[b] Department of Computer and Information Science, University of Macau, Macau 999078, China
[c] School of Mathematics and Computer Science, Yunnan University of Nationalities, Kunming 650031, China

ARTICLE INFO

Article history:
Received 29 October 2012
Received in revised form
11 January 2013
Accepted 23 March 2013
Communicated by D. Tao
Available online 25 April 2013

Keywords:
Learning rate
Density level detection
Rademacher average
Iterative technique

ABSTRACT

This paper investigates the generalization performance of support vector classifiers for density level detection (DLD) when the input term belongs to a separable Hilbert space. The estimate of learning rate for DLD problem is established by Rademacher average and iterative techniques, which is independent of the assumption of covering number used in the previous literature.

1. Introduction

A classification framework for density level detection (DLD) problem has been proposed in [11] and its error analysis has been well established in [10,6] based on the capacity assumption of covering numbers. The theoretical result is important to better understand the mathematical foundation of classification method for DLD. It is well known that the Rademacher complexity has been used successfully for mathematical analysis of machine learning algorithms, see e.g., [2,3,20]. In this paper, we consider establishing the generalization error analysis of the DLD problem by combining the Rademacher complexity with the iterative technique in [13,19,7].

Let us recall the background of the density level detection problem in Hilbert spaces (see [11,10,6]). Let $(H, \|\cdot\|)$ be a separable Hilbert space (possibly infinite dimensional) and let $X \subset H$ with $\|x\| \le B$ for all $x \in X$. Let Q be an unknown data-generating distribution on X. One of the most common ways to define anomalies is by saying that anomalies are not concentrated. A reference distribution μ on X is introduced to describe the concentration of Q. Assume that Q has a density h with respect to μ, i.e. $dQ = h\, d\mu$. Given $\rho > 0$, the set $\{x : h(x) > \rho, x \in X\}$ is called ρ-level set of density h. To define anomalies in terms of the concentration one only has to fix a threshold $\rho > 0$ so that a sample $x \in X$ is considered to be anomalous whenever $h(x) \le \rho$. The main task of the DLD problem is to find ρ-level set $\{x : h(x) > \rho, x \in X\}$. In this paper, we assume that $\{x : h(x) = \rho, x \in X\}$ is a μ-zero set and hence it is also a Q-zero set (see e.g., [9,17]).

Let $S = \{x_i\}_{i=1}^k$ be a training set which is drawn independently from Q. Given S, a DLD algorithm learns a function $f_S : X \to \mathbb{R}$ such that the set $\{x : f_S(x) > 0\}$ is a good estimate of ρ-level set. For a measurable function $f : X \to \mathbb{R}$ the approximation performance is measured (see [11]) by

$$S_{\mu,h,\rho}(f) := \mu(\{f > 0\} \Delta \{h > \rho\}),$$

where Δ denotes the symmetric difference.

Unfortunately, there is no known method to estimate $S_{\mu,h,\rho}(f)$ from empirical data, and hence empirical comparison in terms of $S_{\mu,h,\rho}(f)$ is difficult. To overcome this difficulty, a novel performance measure has been proposed in [11] by interpreting the DLD problem as a binary classification problem. Let $Y := \{-1, 1\}$. The measure is defined as below.

Definition 1. Let Q and μ be probability measures on X and $s \in (0,1)$. Then the probability measure $Q \ominus_s \mu(A)$ on $X \times Y$ is defined by

$$Q \ominus_s \mu(A) = sE_{x\sim Q} I_A(x,1) + (1-s)E_{x\sim\mu} I_A(x,-1)$$

for all measurable subsets $A \subset X \times Y$. Here I_A denotes the indicator function of a set A.

From the definition we know that $P := Q \ominus_s \mu$ can be associated with a binary classification problem where positive samples are drawn from Q and negative samples are drawn from μ.

* Corresponding author. Tel.: +86 15387173655.
E-mail address: zhibinpan2008@gmail.com (Z. Pan).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.03.014

The misclassification risk for a measurable function $f: X \to \mathbb{R}$ and a distribution P on $Z := X \times Y$ is defined by

$$\mathcal{R}_P(f) = P(\{(x,y) : \mathrm{sign}(f)(x) \neq y\}),$$

where sign $t := 1$ if $t > 0$ and sign $t := -1$ otherwise.

It is well known that the Bayes classifier $f_c = \mathrm{sign}(2P(y=1|\cdot)-1)$ minimizes the misclassification risk $\mathcal{R}_P(f)$. Moreover, for $P := Q \ominus_s \mu$ and $s = 1/(1+\rho)$, $f_c = I_{\{h > \rho\}} - I_{\{h \le \rho\}}$.

As shown in [12], $S_{\mu,h,\rho}(f) \to 0$ if and only if $\mathcal{R}_P(f) \to \mathcal{R}_P(f_c)$. Thus, the problem of DLD can be transformed into finding a good function f such that $\mathcal{R}_P(f) \to \mathcal{R}_P(f_c)$. Based on the interpretation, a kernel-based method is introduced in [11] to realize DLD.

Recall that $K: X \times X \to \mathbb{R}$ is a Mercer kernel if it is continuous, symmetric, and positive semi-definite. The candidate reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ 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 $\langle \cdot,\cdot \rangle_K$ defined by $\langle K_x, K_y \rangle_K = K(x,y)$ (see [1]). The reproducing property is given by $\langle K_x, f \rangle_K = f(x)$ for all $x \in X$ and $f \in \mathcal{H}_K$.

For given positive labeled data $T^+ = \{x_i\}_{i=1}^n$ drawn independently from Q, the empirical quantity

$$\frac{1}{n(1+\rho)} \sum_{i=1}^{n} (1-f(x_i))_+ + \frac{\rho}{1+\rho} E_{x' \sim \mu}(1+f(x'))_+.$$

is considered in [11]. As pointed out by Steinwart et al. in [11], although the measure μ is known, the expectation $E_{x' \sim \mu}(1+f(x'))_+$ can be numerically computed through finite evaluation of f on $T^- = \{x'_j\}_{j=1}^m$. Here $T^- = \{x'_j\}_{j=1}^m$ are randomly drawn independently according to μ. The empirical risk of f is defined as

$$\mathcal{E}_T(f) = \frac{1}{n(1+\rho)} \sum_{i=1}^{n} (1-f(x_i))_+ + \frac{\rho}{m(1+\rho)} \sum_{j=1}^{m} (1+f(x_j'))_+.$$

The following regularized algorithm has been proposed in [11]

$$f_T = \arg\min_{f \in \mathcal{H}_K} \{\mathcal{E}_T(f) + \lambda \|f\|_K^2\}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter.

Under the assumption on covering numbers, the convergence of (1) is well understood in [10,6]. In this paper, inspired by theoretical analysis in [3,5,7], we adopt the Rademacher average as the capacity measure of hypothesis space. Dimension-free bound of capacity can be derived in terms of the structural properties of Rademacher average. Without the covering number assumption, satisfactory learning rate is obtained by combining the Rademacher complexity with the iteration technique.

The rest of this paper is organized as follows. In Section 2, we introduce the necessary definitions and present the main result on learning rate. A detailed proof of the main result is provided in Section 3.

2. Error analysis

To establish the relationship between $S_{\mu,h,\rho}(f_T)$ and the excess risk $\mathcal{R}_P(f_T) - \mathcal{R}_P(f_c)$, we recall the following assumption [11,10].

Definition 2. Let μ be a distribution on X and let $h: X \to [0,\infty]$ be a measurable function with $\int h \, d\mu = 1$, i.e. h is a density with respect to μ. For $\rho > 0$ and $0 \le q \le \infty$, we say h has ρ-exponent q if there exists a constant $c > 0$ such that for all $t > 0$

$$\mu(\{|h-\rho| \le t\}) \le c t^q.$$

The assumption on h is closely related to the definition of Tsybakov noise in [18] for binary classification. If h has ρ-exponent $q \in (0,\infty]$, Theorem 10 [11] shows that there exists a constant $c > 0$ such that

$$S_{\mu,h,\rho}(\mathrm{sign}(f)) \le c(\mathcal{R}_P(f) - \mathcal{R}_P(f_c))^{q/(q+1)}. \tag{2}$$

According to $\mathcal{E}_T(f)$, we introduce the expected risk with a convex loss

$$\mathcal{E}(f) = \frac{1}{1+\rho} E_{x \sim Q}(1-f(x))_+ + \frac{\rho}{1+\rho} E_{x' \sim \mu}(1+f(x'))_+.$$

We know that for every measurable function $f: X \to \mathbb{R}$

$$\mathcal{R}_P(f) - \mathcal{R}_P(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_c) \tag{3}$$

according to Theorem 2.1 in [21] or Theorem 9.21 in [8].

Define the data independent regularization function

$$f_\lambda := \arg\min_{f \in \mathcal{H}_K} \{\mathcal{E}(f) + \lambda \|f\|_K^2\}. \tag{4}$$

From the definitions of f_T in (1) and f_λ in (4), we have

$$\mathcal{E}(f_T) - \mathcal{E}(f_c) \le \mathcal{E}(f_T) - \mathcal{E}(f_c) + \lambda \|f_T\|_K^2 \le \mathcal{S}(T,\lambda) + \mathcal{D}(\lambda), \tag{5}$$

where the sample error

$$\mathcal{S}(T,\lambda) = \{\mathcal{E}(f_T) - \mathcal{E}_T(f_T)\} + \{\mathcal{E}_T(f_\lambda) - \mathcal{E}(f_\lambda)\}$$

and the approximation error

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_c) + \lambda \|f_\lambda\|_K^2.$$

The bounding technique for sample error $\mathcal{S}(T,\lambda)$ relies on complexity measure of hypothesis function space $\mathcal{H}_K$. To derive a dimension-free estimate, we introduce Rademacher complexity [2] as the measure of capacity.

Definition 3. Let ϱ be a probability distribution on a set X and suppose that $x_1, \ldots, x_m$ are independent samples selected according to this distribution. Let $\mathcal{F}$ be a class of real-valued functions defined on X. The empirical Rademacher average of $\mathcal{F}$ is defined by

$$\hat{\mathcal{R}}_m(\mathcal{F}) = \mathrm{E}_\sigma \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right| : x_1, \ldots, x_m \right\},$$

where $\sigma_1, \ldots, \sigma_m$ are independent uniform $\{\pm 1\}$-valued random variables. The Rademacher complexity of $\mathcal{F}$ is $\mathcal{R}_m(\mathcal{F}) = \mathrm{E}\hat{\mathcal{R}}_m(\mathcal{F})$.

In this paper, we adopt the following condition for approximation error, which has been extensively used in the literature. See e.g., [4,19,8,20,6].

Definition 4. We say the target function f_c can be approximated with exponent $0 < \beta \le 1$ in $\mathcal{H}_K$ if there exists a constant $c_\beta \ge 1$, such that

$$\mathcal{D}(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0. \tag{6}$$

It is now a position to present our main result on learning rate. The detailed proof will be given in the next section.

Theorem 1. *Let $\rho > 0$. Let μ and Q be distributions on X such that Q has a density h with respect to μ. For $s = 1/(\rho+1)$ we write $P = Q \ominus_s \mu$. Assume that h has ρ-exponent q and f_c can be approximated with exponent β in $\mathcal{H}_K$. Then, for any $0 < \delta < 1$, choosing $\lambda = (1/\sqrt{m} + 1/\sqrt{n})^{2/(\beta+1)}$, we have with confidence $1-\delta$*

$$S_{\mu,h,\rho}(f_T) \le C \sqrt{\ln\left(\frac{k_0+1}{\delta}\right)} \left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)^{q\beta/(\beta+1)(q+1) - q/2^{k_0}(\beta+1)(q+1)},$$

where C is a constant independent of m, n, δ, and k_0 is a constant satisfying $(\sqrt{mn}/(\sqrt{m}+\sqrt{n}))^{2\beta/(\beta+1)2^{k_0+1}} = O(k)$.

From the result in Theorem 1, we know that the balance of samples is crucial to reach the fast learning rate. In particular, learning rate of f_T can be close to $O(n^{-q/(4q+4)})$ when $m = O(n)$ and $\beta \to 1$. It is worth noting that the presented convergence analysis is independent of the assumption on covering numbers in [6].