Neurocomputing 167 (2015) 528–534

Contents lists available at ScienceDirect

Neurocomputing

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

Generalization ability of extreme learning machine with uniformly ergodic Markov chains

Peipei Yuan[a], Hong Chen[b,*], Yicong Zhou[c], Xiaoyan Deng[b,**], Bin Zou[d]

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

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form
7 April 2015
Accepted 15 April 2015
Communicated by G.-B. Huang
Available online 27 April 2015

Keywords:
Generalization ability
Extreme learning machine
Uniformly ergodic Markov chain

ABSTRACT

Extreme learning machine (ELM) has gained increasing attention for its computation feasibility on various applications. However, the previous generalization analysis of ELM relies on the independent and identically distributed (i.i.d) samples. In this paper, we go far beyond this restriction by investigating the generalization bound of the ELM classification associated with the uniform ergodic Markov chains (u.e.M.c) samples. The upper bound of the misclassification error is estimated for the ELM classification showing that the satisfactory learning rate can be achieved even for the dependent samples. Empirical evaluations on real-word datasets are provided to compare the predictive performance of ELM with independent and Markov sampling.

1. Introduction

Extreme learning machine (ELM) can be considered as a single-hidden layer feedforward neural networks (FNNs), where the output weights can be adjusted while the input weights and the threshold of hidden layer are fixed randomly [6,9]. This idea of training FNNs is different from the traditional neural network theories and is related with the discussions in [13,14]. Because only the Moore–Penrose generalized inverse is necessary to be calculated, the original ELM and its variations have shown the computation feasibility in the various applications, see, e.g., [2,4,5,11,23]. With the rapid development of the ELM-based applications, there are some theoretical works for its universal consistency in [25] and generalization ability in [10,19,2]. In particular, the generalization bounds of ELM are established in [10], which demonstrate that ELM can achieve the same learning rates as FNNs under mild conditions. Moreover, analysis of the generalization ability is extended to the magnitude-preserving regularization ranking in [2]. Although these works enrich our understanding of ELM, they just consider the setting where the samples are drawn independently from an unknown distribution. In the real-world applications, the independence of samples is difficult to be verified and does not hold true usually [20,16,26,28]. Therefore, it is important to further investigate the generalization ability of ELM with dependent samples.

Recently, the Markov chain samples have attracted increasing attention in statistical learning theory. In [17], the learning rate is estimated for the online algorithm with the Markov chains. For the uniformly ergodic Markov chains (u.e.M.c), the generalization bounds are established for the regularized regression in [27] and support vector machines classification in [21,22]. Despite the rapid theoretical progresses, there is no any generalization analysis for the regularized ELM with dependent samples. To fill the theoretical gap, in this paper, we investigate the generalization ability of the ELM classification with the Markov samples. The derived results on theory and experiments demonstrate that the satisfying generalization performance can be reached by the ELM with Markov sampling.

The rest of this paper is organized as follows. ELM and some necessary definitions are introduced in Section 2. The main result on generalization analysis is presented for the ELM-based classification in Section 3. Some empirical examples are reported in Section 4. Finally, we conclude this paper in Section 5.

2. Preliminaries

Let $X \in R^d$ be the input space and $Y = \{-1, 1\}$. The training samples $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ are drawn from a probability distribution ρ on $Z = X \times Y$. Given $\mathbf{z}$, the main goal of the classification algorithm is searching a predictor $f_{\mathbf{z}} : X \to Y$ such that

[*] Corresponding author.
[**] Principal Corresponding author.
E-mail addresses: chenh@mail.hzau.edu.cn (H. Chen), dxygh@mail.hzau.edu.cn (X. Deng).

http://dx.doi.org/10.1016/j.neucom.2015.04.041
0925-2312/© 2015 Elsevier B.V. All rights reserved.

the misclassification rate is as low as possible. In learning theory, the misclassification risk is defined as

$$\mathcal{R}(f) = \int_Z I\{y \neq f(x)\}\, d\rho$$

and the Bayes risk is denoted by

$$\mathcal{R}^* = \min \int_Z I\{y \neq f(x)\}\, d\rho.$$

For the regression function $f_\rho = \int_Y y\, d\rho(y|x)$, we know that $\mathcal{R}^* = \mathcal{R}(f_c)$, where $f_c = \text{sign}(f_\rho)$, and $\text{sign}\{t\} = 1$ if $t \geq 0$ and $\text{sign}\{t\} = -1$ otherwise. The performance of a classifier is measured by the excess risk $\mathcal{R}(f) - \mathcal{R}(f_c)$. Since the indictor loss I is nonconvex and noncontinuous, we usually use the convex loss to replace it. In original ELM, the least square loss $\ell(f,z) = (y - f(x))^2$ is used.

Denote $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T \in \mathbb{R}^{n \times l}$ in which α_i is generated independently and identically according to a uniform distribution μ on $[0,1]^l$. In ELM, the hypothesis space is defined as

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

where $\phi : \mathbb{R}^l \times \mathbb{R}^d \to \mathbb{R}$ is an activation function. The activation functions include the sigmoid function, Gaussian function, hyperbolic tangent function, multiquadric function and Fourier series [7,8,5].

For $f \in \mathcal{M}_n$, define

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

Under the Tikhonov regularization scheme, the regularized ELM (see [5,6]) can be formulated as

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

where

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

is the empirical risk and $\lambda > 0$ is the regularization parameter.

The regularized ELM can be rewritten as the optimization scheme

$$\beta^* = \arg\min_{\beta} \left\{ \frac{1}{m} \|H\beta - Y\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$

where $Y = (y_1, y_2, \ldots, y_m)^T$ and

$$H = \begin{pmatrix} \phi(\alpha_1, \mathbf{x}_1) & \ldots & \phi(\alpha_n, \mathbf{x}_1) \\ & \ldots & \\ \vdots & \ldots & \vdots \\ & \ldots & \\ \phi(\alpha_1, \mathbf{x}_m) & \ldots & \phi(\alpha_n, \mathbf{x}_m) \end{pmatrix}_{m \times n}.$$

It is easy to verify that

$$\beta^* = (H^T H + \lambda m I)^{-1} H^T Y.$$

The expected convex risk, associated with the least square loss, is defined as

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

Let $L^2_{\rho_X}$ be the Hilbert space consisted all square integrable functions on X, with norm $\|\cdot\|_\rho$. For every $f \in L^2_{\rho_X}$, we have $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2$. From [24], we know

$$\mathcal{R}(f) - \mathcal{R}(f_c) \leq \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho)} = \|f - f_\rho\|_\rho.$$

This paper focuses on bounding the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ to measure the generalization ability of ELM. The current analysis is based on the u.e.M.c samples different from the previous works in [10,19].

Now we recall some preliminary definition and properties of the u.e.M.c [12,18,22]. Let $(Z, \mathcal{S})$ be a measurable space. We call $\{Z_t\}_{t \geq 1}$ is a Markov chain, if the sequence $\{Z_t\}_{t \geq 1}$ is randomly generated and its transition probability measure satisfies

$$P^k(A|Z_i) = Prob\{Z_{k+i} \in A | Z_j, j < i, Z_i = z_i\}. \tag{3}$$

Starting from the initial state z_i at time i, the probability, that the state z_{k+i} will belong to set A after k-steps, is denoted by $P^k(A|Z_i)$. Hence, if $k=1$, we have $P^1(A|Z_i) = Prob\{Z_{i+1} \in A | Z_j, j < i, Z_i = z_i\}$, which is independent of the values of $Z_j (j < i)$. For the given probabilities p_1 and p_2, the total variance distance is defined as $\|p_1 - p_2\|_{TV} = \sup_{A \in \mathcal{S}} |p_1(A) - p_2(A)|$. The definition of u.e.M.c can be described as below (see [20]).

Definition 1. A Markov chain $\{Z_t\}_{t \geq 1}$ is said to be uniformly ergodic if

$$\|P^k(\cdot|z) - \pi(\cdot)\|_{TV} \leq \gamma \tau^k, \tag{4}$$

for some $0 < \gamma < \infty$ and $0 < \tau < 1$. Here $k \geq 1$, $k \in \mathbb{N}$ and $\pi(\cdot)$ is the stationary distribution of $\{Z_t\}_{t \geq 1}$.

From [12], we know that the transition probability $P^k(A|Z_i)$ of the u.e.M.c satisfies the Doeblin condition as below.

Proposition 1. *Let $\{Z_t\}_{t \geq 1}$ be a Markov chain with the transition probability measure $P^k(\cdot|\cdot)$ and let μ be a specific nonnegative measure with nonzero mass μ_0. Assume that, for some integer t and all measurable sets A, $P^t(A|z) \leq \mu(A)$, $\forall z \in Z$. Then, we have*

$$\|P^k(\cdot|z) - P^k(\cdot|z')\|_{TV} \leq 2(1-\mu_0)^{k/t}, \quad \forall k \in \mathbb{N}, \; z, z' \in Z. \tag{5}$$

3. Generalization bound

To evaluate the generalization ability of ELM, we should estimate the approximation between $f_{\mathbf{z},\lambda}$ and f_ρ. That is to say, we should estimate the excess convex risk $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$.

Proposition 2. *For any $\mathbf{z} \in Z^m$ and $f_{\mathbf{z},\lambda}$ defined in (2), there holds*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq S_1 + S_2, \tag{6}$$

where

$$S_1 = \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_\rho))$$

and

$$S_2 = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_\rho) + \lambda \|f_{\mathbf{z},\lambda}\|_{\ell_2}^2.$$

Definition 2. For a subset $\mathcal{G}$ of a metric space and any $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{G}, \epsilon)$ is defined to be the smallest integer $l \in \mathbb{N}$ such that there exist l disks with radius ϵ and centers in $\mathcal{G}$ covering $\mathcal{G}$.

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

$$B_R = \{f \in \mathcal{M}_n : \|f\|_{\ell_2}^2 \leq R^2\}.$$

The covering number of B_R is estimated in [3].

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

$$\log \mathcal{N}(B_R, \epsilon) \leq n \cdot \log\left(\frac{4R}{\epsilon}\right). \tag{7}$$

Denote $\|\Gamma\| = \sqrt{2}/(1-(1-\mu_0)^{1/2t})$, where μ_0 and t are defined in Proposition 1. In fact, $\|\Gamma\|$ measures the "L^2-dependence" of the