Chinese Journal of Electronics
Vol.32, No.1, Jan. 2023

A Novel Robust Online Extreme Learning Machine for the Non-Gaussian Noise

GU Jun[1], ZOU Quanyi[2], DENG Changhui[1], and WANG Xiaojun[3]

(1. *College of Information Engineering, Dalian Ocean University, Dalian 116023, China*)

(2. *School of software Engineering, South China University of Technology, Guang Zhou 510006, China*)

(3. *School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian 116025, China*)

Abstract — Samples collected from most industrial processes have two challenges: one is contaminated by the non-Gaussian noise, and the other is gradually obsolesced. This feature can obviously reduce the accuracy and generalization of models. To handle these challenges, a novel method, named the robust online extreme learning machine (RO-ELM), is proposed in this paper, in which the least mean p-power criterion is employed as the cost function which is to boost the robustness of the ELM, and the forgetting mechanism is introduced to discard the obsolescence samples. To investigate the performance of the RO-ELM, experiments on artificial and real-world datasets with the non-Gaussian noise are performed, and the datasets are from regression or classification problems. Results show that the RO-ELM is more robust than the ELM, the online sequential ELM (OS-ELM) and the OS-ELM with forgetting mechanism (FOS-ELM). The accuracy and generalization of the RO-ELM models are better than those of other models for online learning.

Key words — Extreme learning machine (ELM), Online learning, Non-Gaussian noise, Obsolescence samples, Least mean p-power (LMP).

I. Introduction

In most industrial processes, samples are inevitably contaminated by the non-Gaussian noise, and their validity gradually reduces as time passes. Models built on such samples are always with poor accuracies and generalizations. Various robust learning algorithms have been proposed by many scholars to handle the Gaussian noise. However, robust algorithms that focus on dealing with the non-Gaussian noise are relatively fewer, especially for online leaning in which the application of obsolescence samples can further decline the accuracy and generalization. Therefore, a robust online learning algorithm which can well address the non-Gaussian noise and the obsolescence samples is keenly sought.

Data-driven models have been developed and implemented in many fields for the past decades. They gain in popularity with the increasing availability of samples and feasibility of computational power. Artificial neural networks (ANNs) are a class of the most popular learning algorithms to establish data-driven models, which can approximate any nonlinear continuous functions. It is worth mentioning that in recent years, deep learning has made great success in the regression and classification problems [1]–[5]. However, ANNs (including deep learning) employ iterative technique to repeatedly adjust all the parameters of a model, slowing learning speed and reducing computational scalability, etc. Fortunately, the extreme learning machine (ELM) developed from an ANN was proposed by Huang *et al.* [6]. Compared with traditional ANNs and deep learning networks, the ELM randomly generates parameters of the hidden layer, and then the output weights of the ELM are decided by a pseudo-inverse operation. Its outstanding advantage is that the parameters of the hidden layer are randomly assigned to replace the iteratively tuning, improving the learning speed and simplifying the computing. Due to this outstanding advantage, the ELM is widely concerned in theory and applications. Various improved versions of the ELM were proposed [7], [8]. Aforementioned algorithms belong to offline learning which is based on a certain number of static samples. However, in many applications samples arrive in the order of time, such as

Manuscript Received Apr. 9, 2021; Accepted Dec. 8, 2021. This work was supported by Scientific Research Funding Project of the Education Department of Liaoning Province (LN2019Q44).

 DOI:10.23919/cje.2021.00.122

the forecasting of renewable energy generation [9]. If an offline learning algorithm is directly applied to establish a predictive model, whenever the new samples arrive, the algorithm will employ the entire samples (including the old and the new samples) to reconstruct the predictive model. Therefore, the old samples are repeatedly learned, so that the learning speed decreases, which is time-consuming.

To cope with this issue, the online sequential ELM (OS-ELM) was proposed by Liang *et al.* [10], which can learn samples chunk by chunk and the new samples instead of the entire samples are learned. However, the OS-ELM ignores the validity period of samples. Exceeding the validity period, samples will become invalid. Applying invalid samples, online models always export inaccuracy results. To handle this issue, the OS-ELM with forgetting mechanism (FOS-ELM) was proposed by Zhao *et al.* [11], where the forgetting mechanism can discard obsolescence samples and enhance the accuracy of the predictive model. Zou *et al.* [12] proposed the memory degradation based OS-ELM (MDOS-ELM) which adjusts the weights of the old and new samples by a self-adaptive memory factor, and discards invalid samples. Generally, FOS-ELM models built on samples without noise or only with the Gaussian noise can obtain satisfactory precisions. Nevertheless, built from samples contaminated by the non-Gaussian noise, the FOS-ELM models are always not accurate enough.

Aiming at handling the non-Gaussian noise and discarding obsolescence samples simultaneously, a novel method, named the robust online extreme learning machine (RO-ELM), is proposed in this paper. In the RO-ELM, on one hand, the least mean p-power (LMP) criterion [13] is employed as the cost function to boost the robustness of the ELM, which outperforms the least mean square (LMS) criterion, especially under the non-Gaussian noise environments [14]. On the other hand, the forgetting mechanism is applied to timely eliminate invalid samples. To investigate the performance of the proposed RO-ELM, both the artificial and real-world datasets from regression or classification problems are used. It is expected that the proposed RO-ELM is more robust on the samples with the non-Gaussian noise. The accuracy and generalization of the RO-ELM models are also expected to be better than those of the ELM, the OS-ELM and the FOS-ELM models for online learning.

The remainder of this paper is organized as follows. In Section II, related works are introduced, such as the OS-ELM, the LMP criterion and types of noise. In section III, the RO-ELM algorithm is proposed, and the universal approximation of the RO-ELM is given. In section IV, experiments to verify the robustness of the proposed RO-ELM and the accuracy and generalization of the RO-ELM models for online learning are performed on datasets from regression and classification problems. In Section V, the conclusion of this paper is summarized.

II. Related Works

1. The OS-ELM algorithm

The training set containing N samples is $S = \{(\boldsymbol{x}_i, t_i) | \boldsymbol{x}_i \in \mathbb{R}^{n\times 1}, t_i \in \mathbb{R}^1, i = 1, 2, \ldots, N\}$, where $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ is the input vector, t is the output variable and n is the dimension of the input variables. An ELM model with L hidden nodes can be expressed as,

$$f_L(\boldsymbol{x}) = \sum_{i=1}^{N} \beta_i G(\boldsymbol{a}_i, b_i, \boldsymbol{x}) \tag{1}$$

where β_i is the output weight of the i-th hidden node. $G(\cdot)$ is the activation function of the hidden node, and it can be "'RBF", "Sigmoid", "Sine", or "hradlim" etc. $\boldsymbol{a}$ and b are the parameters of $G(\cdot)$, which are randomly assigned and not adjusted to any optimization.

The output weight $\boldsymbol{\beta} = [\beta_1, \beta_1, \ldots, \beta_L]^{\mathrm{T}} \in \mathbb{R}^{L\times 1}$ is obtained via the least-mean method. The cost function that is based on the LMS criterion is as follows,

$$\begin{aligned} \min J_{\mathrm{MSE}} &= \frac{1}{N}\sum\nolimits_{i=1}^{N} (f_L(\boldsymbol{x}_i) - t_i)^2 \\ &= \frac{1}{N}\|\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{T}\|^2 \end{aligned} \tag{2}$$

where $\boldsymbol{T} = [t_1, t_2, \ldots, t_N]^{\mathrm{T}} \in \mathbb{R}^{N\times 1}$, and $\boldsymbol{H}$ is the output matrix of the hidden layer, and

$$\boldsymbol{H} = \begin{bmatrix} G(\boldsymbol{a}_1, b_1, \boldsymbol{x}_1) & \cdots & G(\boldsymbol{a}_L, b_L, \boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ G(\boldsymbol{a}_1, b_1, \boldsymbol{x}_N) & \cdots & G(\boldsymbol{a}_L, b_L, \boldsymbol{x}_N) \end{bmatrix} \tag{3}$$

where $\boldsymbol{H} \in \mathbb{R}^{N\times L}$. A Moore-Penrose generalized inverse operation yields the solution of (2). Namely, $\widehat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{T}$, where $\widehat{\boldsymbol{\beta}}$ denotes the estimated value of $\boldsymbol{\beta}$, and $\boldsymbol{H}^{\dagger} = \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}$, and $\boldsymbol{H}^{\dagger} \in \mathbb{R}^{L\times N}$.

As an online version of the ELM, the OS-ELM only learns newly arrived sample chunk (with fixed or varying sizes) by the recursive way.

2. The LMP criterion

The cost function based on the LMP criterion is defined,

$$\min J_{\mathrm{LMP}} = \frac{1}{N}\sum_{i=1}^{N} |e_i|^p = \frac{1}{N}\|\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{T}\|^p \tag{4}$$

where $e_i = f_L(\boldsymbol{x}_i) - t_i$ denotes the prediction error of the i-th sample. When p=2, the LMP criterion degenerates to the LMS criterion. Some researchers pointed out

that the LMP might produce a more precise solution than the LMS on samples contaminated by the non-Gaussian noise [15].

Theorem 1 $J_{\mathrm{LMP}}(\boldsymbol{\beta}) = \|\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{T}\|^p$ is a convex function on defined $\mathbb{R}^L$ and $p \geq 1$.

Proof For each $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, set $0 < \lambda < 1$,

$$\begin{aligned}&\|\boldsymbol{H}(\lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2) - \boldsymbol{T}\|^p \\ &= \|\boldsymbol{H}(\lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2) - (1-\lambda+\lambda)\boldsymbol{T}\|^p \\ &= \|\lambda(\boldsymbol{H}\boldsymbol{\beta}_1 - \boldsymbol{T}) + (1-\lambda)(\boldsymbol{H}\boldsymbol{\beta}_2 - \boldsymbol{T})\|^p \\ &\leq \lambda[\boldsymbol{H}\boldsymbol{\beta}_1 - \boldsymbol{T}]^p + (1-\lambda)\|(\boldsymbol{H}\boldsymbol{\beta}_2 - \boldsymbol{T})\|^p\end{aligned}$$

Thus,

$$\begin{aligned}&J_{\mathrm{LMP}}(\lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2) \\ &\quad\leq \lambda J_{\mathrm{LMP}}(\boldsymbol{\beta}_1) + (1-\lambda)J_{\mathrm{LMP}}(\boldsymbol{\beta}_2)\end{aligned} \tag{5}$$

The proof is completed.

The LMP criterion has the following proprieties: 1) It is applied as the cost function which is a convex function, and no local minima. 2) Algorithms based on the LMP criterion with $p > 2$ perform better convergence performance for the systems with non-Gaussian light-tailed distributions [16]. 3) Algorithms based on the LMP criterion with $p < 2$ perform better robustness for the systems with the non-Gaussian heavy-tailed distributions [13].

3. The non-Gaussian noise

Here, two kinds of the non-Gaussian noise are introduced. They are the symmetric alpha-stable (SαS) noise and the symmetry alpha-stable Gaussian (SαSG) noise which is the mixture of the independent SαS and the Gaussian noise. The SαS distribution can better simulate the noise with heavy-tailed distributions in real world [17]. The SαS noise is more universal and more realistic meaning than the Gaussian noise. Generally, the SαS distribution depicted by its characteristic function [18] is as follows,

$$\phi(t) = \exp(j\mu t - \gamma|t|^\alpha)$$

where μ is the location parameter. γ is the dispersion of the distribution and is similar to the variance of the Gaussian distribution. $\alpha\,(0 < \alpha \leq 2)$ is the characteristic exponent and determines the thickness of the tail in the distribution. As α decreases, the tail of the SαS distribution gradually becomes thick, vice versa. When $\alpha = 2$, the SαS distribution degenerates to the Gaussian distribution.

The SαSG noise appears in a variety of practical situations [19]. The SαSG distribution depicted by its characteristic function is

$$\phi(t) = \exp(-\gamma_{\mathrm{S}\alpha\mathrm{S}}|t|^\alpha - \gamma_{\mathrm{G}}|t|^\alpha)$$

where $\gamma_{\mathrm{S}\alpha\mathrm{S}}$ and γ_{G} are the dispersions of the SαS and the Gaussian distribution, respectively. $\gamma_{\mathrm{S}\alpha\mathrm{S}} > 0$ and $\gamma_{\mathrm{G}} = \sigma_{\mathrm{G}}^2/2$. σ_{G}^2 is the variance of the Gaussian distribution. When $\alpha = 2$, the SαSG becomes the sum of the two independent Gaussian distribution.

III. The RO-ELM Algorithm

Aiming at enhancing the robustness of the ELM, the LMP criterion is applied as the cost function, and simultaneously the forgetting mechanism is employed to timely discard invalid samples to improve the accuracy and generalization of online learning, and the RO-ELM is proposed here.

1. The recursive equation of the RO-ELM

According to the description in Section II.2, the cost function based on the LMP criterion in the ELM is

$$\min J_{\mathrm{LMP}} = \frac{1}{N}\sum_{i=1}^{N}|e_i|^p = \frac{1}{N}\|\boldsymbol{G}_i\boldsymbol{\beta} - t_i\|^p \tag{6}$$

where $\boldsymbol{G}_i = [G(\boldsymbol{a}_1, b_1, \boldsymbol{x}_i), G(\boldsymbol{a}_2, b_2, \boldsymbol{x}_i), \ldots, G(\boldsymbol{a}_L, b_L, \boldsymbol{x}_i)]$. The optimal solution $\boldsymbol{\beta}$ for minimizing J_{LMP} can be obtained by differentiating (6) with respect to $\boldsymbol{\beta}$ and setting the derivatives to zero. The derivatives are

$$\begin{aligned}\frac{\partial J_{\mathrm{LMP}}}{\partial\boldsymbol{\beta}} &= \frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{\partial|e_i|^p}{\partial\boldsymbol{\beta}} \\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N}p|e_i|^{p-2}e_i\cdot\frac{\partial e_i}{\partial\boldsymbol{\beta}}\end{aligned} \tag{7}$$

Inserting $|e_i| = |\boldsymbol{G}_i\boldsymbol{\beta} - t_i|$ and $\frac{\partial|e_i|}{\partial\boldsymbol{\beta}} = \boldsymbol{G}_i^{\mathrm{T}}$ into (7), we have

$$\frac{\partial J_{\mathrm{LMP}}}{\partial\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}p|e_i|^{p-2}|\boldsymbol{G}_i\boldsymbol{\beta} - t_i|\,\boldsymbol{G}_i^{\mathrm{T}} \tag{8}$$

When $\frac{\partial J_{\mathrm{LMP}}}{\partial\boldsymbol{\beta}} = 0$, (8) can be transformed into

$$\sum\nolimits_{i=1}^{N}|e_i|^{p-2}\boldsymbol{G}_i^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{\beta} = \sum\nolimits_{i=1}^{N}|e_i|^{p-2}\boldsymbol{G}_i^{\mathrm{T}}t_i \tag{9}$$

The initial sample chunk $S_0 = \left\{\left(\boldsymbol{x}_i^{(0)}, t_i^{(0)}\right)|\boldsymbol{x}_i^{(0)} \in \mathbb{R}^{n\times1}, t_i^{(0)} \in \mathbb{R}^1, i = 1, 2, \ldots, N_0\right\}$ arrives, and then

$$\sum\nolimits_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}\boldsymbol{G}_i^{(0)}\boldsymbol{\beta}^{(0)} = \sum\nolimits_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}t_i^{(0)} \tag{10}$$

where $\left|e^{(0)}\right| = \left|f_{\tilde{N}}\left(\boldsymbol{x}^{(0)}\right) - t^{(0)}\right|$.

Setting

$$\begin{aligned}\boldsymbol{P}_0 &= \sum_{i=1}^{N_0}\left|e^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}\boldsymbol{G}_i^{(0)}\\ &= (\boldsymbol{F}_0\boldsymbol{H}_0)^{\mathrm{T}}\boldsymbol{F}_0\boldsymbol{H}_0 = \boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0\end{aligned} \tag{11}$$

and

$$\boldsymbol{R}_0 = \sum_{i=1}^{N_0}\left|e^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}t_i^{(0)} = \boldsymbol{E}_0\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{T}_0 \tag{12}$$

where $\boldsymbol{H}_0 = \begin{bmatrix}\boldsymbol{G}_1^{(0)} & \boldsymbol{G}_2^{(0)} & \cdots & \boldsymbol{G}_{N_0}^{(0)}\end{bmatrix}^{\mathrm{T}}$. $\boldsymbol{F}_0$ and $\boldsymbol{E}_0$ are diagonal matrices, and the diagonal element is $\left|e_{N_i}^{(0)}\right|^{\frac{p}{2}-1}$ and $\left|e_{N_i}^{(0)}\right|^{p-2}$, respectively.

we rewrite (10) as follows,

$$\boldsymbol{P}_0\boldsymbol{\beta}^{(0)} = \boldsymbol{R}_0 \tag{13}$$

The optimal solution

$$\boldsymbol{\beta}^{(0)} = \boldsymbol{P}_0^{-1}\boldsymbol{R}_0 \tag{14}$$

When the first new sample chunk $S_1 = \left\{\left(\boldsymbol{x}_i^{(1)}, t_i^{(1)}\right) \mid \boldsymbol{x}_i^{(1)} \in \mathbb{R}^{n\times 1},\ t_i^{(1)} \in \mathbb{R}^1,\ i = 1, 2, \ldots, N_1\right\}$ arrives, according to formula (9) the following equation can be obtained

$$\begin{aligned}&\sum_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}\boldsymbol{G}_i^{(0)}\boldsymbol{\beta}^{(0)}\\ &\quad+\sum_{i=1}^{N_1}\left|e_i^{(1)}\right|^{p-2}\boldsymbol{G}_i^{(1)\mathrm{T}}\boldsymbol{G}_i^{(1)}\boldsymbol{\beta}^{(1)}\\ &= \sum_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}t_i^{(0)} + \sum_{i=1}^{N_1}\left|e_i^{(1)}\right|^{p-2}\boldsymbol{G}_i^{(1)\mathrm{T}}t_i^{(1)}\end{aligned} \tag{15}$$

Setting

$$\begin{aligned}\boldsymbol{P}_1 &= \sum_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}\boldsymbol{G}_i^{(0)} + \sum_{i=1}^{N_1}\left|e_i^{(1)}\right|^{p-2}\boldsymbol{G}_i^{(1)\mathrm{T}}\boldsymbol{G}_i^{(1)}\\ &= \boldsymbol{P}_0 + \boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{M}_1\end{aligned} \tag{16}$$

and

$$\begin{aligned}\boldsymbol{R}_1 &= \sum_{i=1}^{N_0}\left|e_i^{(0)}\right|^{p-2}\boldsymbol{G}_i^{(0)\mathrm{T}}t_i^{(0)} + \sum_{i=1}^{N_1}\left|e_i^{(1)}\right|^{p-2}\boldsymbol{G}_i^{(1)\mathrm{T}}t_i^{(1)}\\ &= \boldsymbol{R}_0 + \boldsymbol{E}_1\boldsymbol{H}_1^{\mathrm{T}}\boldsymbol{T}_1\end{aligned} \tag{17}$$

where $\boldsymbol{H}_1 = \begin{bmatrix}\boldsymbol{G}_1^{(1)}\boldsymbol{G}_2^{(1)}\ldots\boldsymbol{G}_{N_1}^{(1)}\end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{T}_1 = \begin{bmatrix}t_1^{(1)}t_2^{(1)}\ldots t_{N_1}^{(1)}\end{bmatrix}^{\mathrm{T}}$. $\boldsymbol{F}_1$ and $\boldsymbol{E}_1$ are diagonal matrices, and the diagonal element is $\left|e_{N_i}^{(1)}\right|^{\frac{p}{2}-1}$ and $\left|e_{N_i}^{(1)}\right|^{p-2}$.

We transform (15) into

$$\boldsymbol{P}_1\boldsymbol{\beta}^{(1)} = \boldsymbol{R}_1 \tag{18}$$

The optimal solution $\boldsymbol{\beta}^{(1)}$ is

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{P}_1^{-1}\boldsymbol{R}_1 \tag{19}$$

Substitute the equations (17) and (13) into (19), and then

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{P}_1^{-1}\boldsymbol{R}_1 = \boldsymbol{\beta}^{(0)} + \boldsymbol{P}_1^{-1}\boldsymbol{E}_1\boldsymbol{H}_1^{\mathrm{T}}\left(\boldsymbol{T}_1 - \boldsymbol{H}_1\boldsymbol{\beta}^{(0)}\right) \tag{20}$$

$\boldsymbol{P}_1^{-1}$ is derived from the wood-bury (20),

$$\begin{aligned}\boldsymbol{P}_1^{-1} &= \left(\boldsymbol{P}_0 + \boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{M}_1\right)^{-1}\\ &= \boldsymbol{P}_0^{-1} - \boldsymbol{P}_0^{-1}\boldsymbol{M}_1^{\mathrm{T}}\left(I + \boldsymbol{M}_1\boldsymbol{P}_0^{-1}\boldsymbol{M}_1^{\mathrm{T}}\right)^{-1}\boldsymbol{M}_1\boldsymbol{P}_0^{-1}\end{aligned} \tag{21}$$

The RO-ELM learns samples chunk by chunk with fixed or varying sizes. It assumes that any new sample chunk arrives in each unit time, and the validity period of any sample chunk is $u\,(u > 1)$ unit time. When samples exceed the validity period, they become invalid.

The current valid sample set is denoted as $\overline{S_k}$, and the k-th sample chunk $S_k = \left\{\left(\boldsymbol{x}_i^{(k)}, t_i^{(k)}\right) \mid \boldsymbol{x}_i^{(k)} \in \mathbb{R}^{n\times 1},\ t_i^{(k)} \in \mathbb{R}^1, i = 1, 2, \ldots, N_k\right\}$ arrives. If $k < u$, no samples are invalid. At the moment $\overline{S_k} = \bigcup_{i=0}^{k} S_i$. $\boldsymbol{P}_k$ can be analogized as follows,

$$\boldsymbol{P}_k = \boldsymbol{P}_{k-1} + \boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{E}_k\boldsymbol{H}_k \tag{22}$$

Based on the wood-bury formula, $\boldsymbol{P}_k^{-1}$ is expressed,

$$\boldsymbol{P}_k^{-1} = \boldsymbol{P}_{k-1}^{-1} - \boldsymbol{P}_{k-1}^{-1}\boldsymbol{H}_k^{\mathrm{T}}\left(\boldsymbol{E}_k^{-1} + \boldsymbol{H}_k\boldsymbol{P}_{k-1}^{-1}\boldsymbol{H}_k^{\mathrm{T}}\right)^{-1}\boldsymbol{H}_k\boldsymbol{P}_{k-1}^{-1} \tag{23}$$

Then, $\boldsymbol{\beta}^{(k)}$ is updated into

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + \boldsymbol{P}_k^{-1}\boldsymbol{E}_k\boldsymbol{H}_k^{\mathrm{T}}\left(\boldsymbol{T}_k - \boldsymbol{H}_k\boldsymbol{\beta}^{(k-1)}\right) \tag{24}$$

where $\boldsymbol{T}_k = \begin{bmatrix}t_1^{(k)} & t_2^{(k)} & \ldots & t_{N_k}^{(k)}\end{bmatrix}^{\mathrm{T}}$, and $\boldsymbol{E}_k$ is diagonal matrix, and the diagonal element is $\left|e_{N_i}^{(k)}\right|^{p-2}$.

If $k \geq u$, the $(k-u)$-th sample chunk becomes invalid and discarded. At the moment $\overline{S_k} = \bigcup_{i=k-u+1}^{k} S_i$, $\left(\overline{S_k} = \overline{S_{k-1}} + S_k - S_{k-u}\right)$. $\boldsymbol{P}_k$ can be analogized

$$\begin{aligned}\boldsymbol{P}_k &= \boldsymbol{P}_{k-1} + \boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{E}_k\boldsymbol{H}_k - \boldsymbol{H}_{k-u}^{\mathrm{T}}\boldsymbol{E}_{k-u}\boldsymbol{H}_{k-u}\\ &= \boldsymbol{P}_{k-1} + \begin{bmatrix}-\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}\end{aligned} \tag{25}$$

where $\boldsymbol{M}_k = \boldsymbol{F}_k\boldsymbol{H}_k$, $\boldsymbol{M}_{k-u} = \boldsymbol{F}_{k-u}\boldsymbol{H}_{k-u}$, $\boldsymbol{F}_k$ and $\boldsymbol{F}_{k-u}$ are diagonal matrices, and the diagonal element is $\left|e_{N_i}^{(k)}\right|^{\frac{p}{2}-1}$ and $\left|e_{N_i}^{(k-u)}\right|^{\frac{p}{2}-1}$.

Based on the wood-bury formula, $\boldsymbol{P}_k^{-1}$ is expressed

$$\boldsymbol{P}_k^{-1}=\left(\boldsymbol{P}_{k-1}+\begin{bmatrix}-\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}\right)^{-1}=\boldsymbol{P}_{k-1}^{-1}-\boldsymbol{P}_{k-1}^{-1}\begin{bmatrix}-\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}^{\mathrm{T}}\left(\boldsymbol{I}+\begin{bmatrix}\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}\boldsymbol{P}_{k-1}^{-1}\begin{bmatrix}-\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}^{\mathrm{T}}\right)\begin{bmatrix}\boldsymbol{M}_{k-u}\\ \boldsymbol{M}_k\end{bmatrix}\boldsymbol{P}_{k-1}^{-1} \tag{26}$$

Then, the following relation can be analogized

$$\boldsymbol{\beta}^{(k)}=\boldsymbol{P}_k^{-1}\boldsymbol{R}_k \tag{27}$$

where $\boldsymbol{R}_k=\boldsymbol{R}_{k-1}+\boldsymbol{E}_k\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{T}_k-\boldsymbol{E}_{k-u}\boldsymbol{H}_{k-u}^{\mathrm{T}}\boldsymbol{T}_{k-u}$.

And then the equation for sequentially updating $\boldsymbol{\beta}^{(k)}$ can be obtained

$$\begin{aligned}\boldsymbol{\beta}^{(k)}=&\boldsymbol{P}_k^{-1}\left(\boldsymbol{R}_{k-1}+\boldsymbol{E}_k\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{T}_k-\boldsymbol{E}_{k-u}\boldsymbol{H}_{k-u}^{\mathrm{T}}\boldsymbol{T}_{k-u}\right)\\ =&\boldsymbol{P}_k^{-1}\left(\boldsymbol{P}_{k-1}\boldsymbol{\beta}^{(k-1)}+\boldsymbol{E}_k\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{T}_k-\boldsymbol{E}_{k-u}\boldsymbol{H}_{k-u}^{\mathrm{T}}\boldsymbol{T}_{k-u}\right)\\ =&\boldsymbol{\beta}^{(k-1)}+\boldsymbol{P}_k^{-1}\begin{bmatrix}-\boldsymbol{E}_{k-u}\boldsymbol{H}_{k-u}\\ \boldsymbol{E}_k\boldsymbol{H}_k\end{bmatrix}^{\mathrm{T}}\\ &\times\left(\begin{bmatrix}\boldsymbol{T}_{k-u}\\ \boldsymbol{T}_k\end{bmatrix}-\begin{bmatrix}\boldsymbol{H}_{k-u}\\ \boldsymbol{H}_k\end{bmatrix}\boldsymbol{\beta}^{(k-1)}\right)\end{aligned} \tag{28}$$

By equations (26) and (28), the recursive relation between $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\beta}^{(k-1)}$ is obtained without repeated learning. And the obsolescence samples are also timely discarded.

2. The universal approximation of the RO-ELM

In fact, the RO-ELM is a single-hidden layer feed-forward neural-network (SLFN). So according to the Theorem 2.1 of Huang *et al*'s work [6], in a standard SFLN, $\boldsymbol{P}_k$ is invertible and $\sum_{i=1}^{N_k}\left|t_i^{(k)}-\boldsymbol{G}_i^k\boldsymbol{\beta}\right|^p=0$, where $\boldsymbol{G}_i=[G(\boldsymbol{a}_1,b_1,\boldsymbol{x}_i)\quad G(\boldsymbol{a}_2,b_2,\boldsymbol{x}_i)\quad\dots\quad G(\boldsymbol{a}_L,b_L,\boldsymbol{x}_i)]$.

From (11), $\boldsymbol{P}_0=\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0$, and $\boldsymbol{P}_0$ is an invertible matrix. That is, $\mathrm{rank}(\boldsymbol{P}_0)=L$. From (16),

$$\boldsymbol{P}_1=\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0+\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{M}_1=\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{K}_1$$

And there are N_0+N_1 distinct samples, $\mathrm{rank}\left(\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{K}_1\right)=\min(L,N_0+N_1)=L$ can be proved. According to the lemma in [20], existing $\mathrm{rank}\left(\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{K}_1\right)=\mathrm{rank}\left(\boldsymbol{K}_1\boldsymbol{K}_1^{\mathrm{T}}\right)=\mathrm{rank}(\boldsymbol{K}_1)$, thus $\mathrm{rank}(\boldsymbol{K}_1)=L$.

When $k<u$, $\boldsymbol{P}_k=\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0+\cdots+\boldsymbol{M}_k^{\mathrm{T}}\boldsymbol{M}_k=\boldsymbol{K}_k^{\mathrm{T}}\boldsymbol{K}_k$, and there are $\sum_{i=1}^{k}N_i$ distinct samples. $\mathrm{rank}(\boldsymbol{K}_k)=\min\left(L,\sum_{i=1}^{k}N_i\right)=L$ can be proved.

When $k\geq u$, $\boldsymbol{P}_k=\begin{bmatrix}\boldsymbol{M}_{k-u+1}\\ \vdots\\ \boldsymbol{M}_k\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{M}_{k-u+1}\\ \vdots\\ \boldsymbol{M}_k\end{bmatrix}=\boldsymbol{K}_k^{\mathrm{T}}\boldsymbol{K}_k$, and there are $\sum_{i=k-u+1}^{k}N_i$ distinct samples. $\mathrm{rank}\left(\boldsymbol{K}_k^{\mathrm{T}}\boldsymbol{K}_k\right)=\min\left(L,\sum_{i=k-u+1}^{k}N_i\right)$. And if $L\leq\sum_{i=k-u+1}^{k}N_i$, $\mathrm{rank}(\boldsymbol{K}_k)=L$ and $\boldsymbol{P}_k$ is an invertible matrix. If $L>\sum_{i=k-u+1}^{k}N_i$, $\mathrm{rank}(\boldsymbol{K}_k)=\sum_{i=k-u+1}^{k}N_i$ and $\boldsymbol{P}_k$ is a singular matrix.

The proposed RO-ELM consists of two phases, the initialize learning and the online sequential learning phases. However, in the initialize learning phase, the model has not been built, so $\left|e^{(0)}\right|$ can not be obtained, and this phase is replaced by the primary ELM. The proposed RO-ELM algorithm is shown in Algorithm 1.

Algorithm 1 The RO-ELM algorithm

$k\,(k=0,1,2,\dots)$ denotes the index of a sample chunk. For the initial sample chunk, k=0; and for the first sample chunk, k=1, and so on.

Step 1 Initialize learning

S_0 is the initial sample chunk. L is the number of the hidden nodes of the RO-ELM model. $G(\cdot)$ is the activation function of the hidden nodes. u is the validity period of samples.

1: Parameters $\boldsymbol{a}$ and b of the hidden layer are randomly assigned.

2: The hidden layer output matrix $\boldsymbol{H}_0$ is calculated.

3: The initial output weight $\boldsymbol{\beta}^{(0)}=\boldsymbol{P}_0\boldsymbol{H}_0^{\mathrm{T}}\boldsymbol{T}_0$ is estimated.

Step 2 Online sequential learning

The k-th $(k=1,2,\cdots)$ sample chunk S_k arrives, where N_k is the number of samples.

4: The output matrix $\boldsymbol{H}_k$ of the hidden layer is obtained.

5: The output weight $\boldsymbol{\beta}^{(k)}$ of the hidden layer is computed. When $k<u$ according to (24), $\boldsymbol{\beta}^{(k)}$ is computed. When $k\geq u$, according to (28), $\boldsymbol{\beta}^{(k)}$ is calculated.

6: When a new sample chunk arrives, k is replaced by $k+1$ and then turn to Step b). Otherwise, the learning is ended.

Remark When p=2, the proposed RO-ELM changes into the FOS-ELM. When p=2 and the timeliness of samples are not considered (i.e., $u=+\infty$), the proposed RO-ELM degenerates to the OS-ELM. Therefore, both the OS-ELM and the FOS-ELM are the special cases of our RO-ELM.

Although the proposed ROS-ELM with an appropriate period of validity can obtain better performance for online learning, the theoretic analysis to determine the exact period of validity is not very clear. This question will be discussed in future works. In specific implementation, we recommend setting a candidate set first, and then selecting it according to the prediction effect of the model.

IV. Experiments

To investigate how well the proposed RO-ELM

works on handling the non-Gaussian noise and discarding obsolescence samples simultaneously, the empirical study was performed on the artificial and the real-world datasets from regression or classification problems. The RO-ELM was compared with the ELM, the OS-ELM and the FOS-ELM on the robustness, the accuracy and the generalization for online learning. In all the runs the following procedures were used.

For any real-world dataset, samples were randomly divided into the training set (80%) and the testing set (20%). Additionally, training samples were randomly split into an initial sample chunk and some online-updating sample chunks. For regression problems, the root mean square error (RMSE) and the coefficient of determination R^2 were used as the evaluative criteria. For classification problems, the correct rate was employed as the measure criteria. The RO-ELM and its comparers adopt the "Sigmoid" function as the activation function. Any experiment was repeated executed 100 times, and the statistical results were shown. For any real-world dataset, samples were randomly divided into the training set (80%) and the testing set (20%). Additionally, training samples were randomly split into an initial sample chunk and some online-updating sample chunks. For regression problems, the root mean square error (RMSE) and the coefficient of determination R^2 were used as the evaluative criteria. For classification problems, the correct rate was employed as the measure criteria. The RO-ELM and its comparers adopt the "Sigmoid" function as the activation function. Any experiment was repeated executed 100 times, and the statistical results were shown.

1. The experiments for regression problems

For regression problems, an artificial and two real-world datasets from the UCI machine learning repository[*1] were employed.

1) On the artificial dataset

The clear samples were generated from the function of "Friedman#2" which is

$$y = x_1^2 + \sqrt{\frac{x_2 \cdot x_3 - 1}{(x_2 \cdot x_4)^2}}$$

where $x_1 \in (0, 99)$, $x_2 \in (4\pi, 564\pi - 1)$, $x_3 \in (0, 1)$ and $x_4 \in (1, 11)$. 2000 training samples and 500 testing samples were randomly generated. These training samples were divided into an initial sample chunk including 200 samples and 18 online updating-sample chunks, each of which contained 100 samples.

The random Gaussian noise with mean=0 and variance=0.4 was generated. The SαS and the SαSG noise were randomly created. For the SαS noise, exponent α= 1.2 and dispersion $\gamma_{S\alpha S}$=0.02. The SαSG noise is the sum of the Gaussian noise and the SαS noise. These four kinds of noise were added into the clear samples, respectively.

The ELM or the OS-ELM only considered the optimal number of the hidden nodes, and their learning processes were performed with different numbers of the hidden nodes from the range [2,100] with step=2. The testing RMSE of the ELM and the OS-ELM models with different numbers of the hidden nodes on the artificial datasets with three kinds of noise were shown in Fig.1. It seems that the ELM and the OS-ELM models have the similar performance. This is most likely because the data contains high noise, and OS-ELM is not robust enough to perform much better than ELM. On any noise dataset, the lowest testing RMSE was obtained when the number of the hidden nodes was 20 for both the ELM and the OS-ELM models.

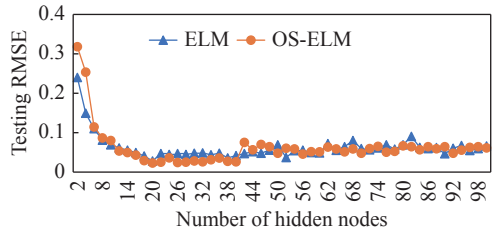


Fig. 1. The testing RMSE of the ELM and the OS-ELM models with different numbers of the hidden nodes on the artificial datasets with three kinds of noise.

For the RO-ELM model, the number of the hidden nodes was also set as 20. Additionally, it was important to select both the value of p and the validity period u of the samples. If u is too large or too small, the generalization of the RO-ELM model will be declined. Different u from the range [1,19] was selected with step=1. The value of p was chosen from the range (1,2] with step=0.2 on the Gaussian, the SαS and the SαSG datasets. Then, the best u and p were selected according to the testing RMSE of the RO-ELM model.

The testing RMSE of the RO-ELM models with different validity periods of the samples on the artificial datasets with the SαSG noise were shown in Fig.2. On any noise dataset, the lowest testing RMSE was obtained when u was 5. On the other hand, the testing RMSE of the RO-ELM models with different values of p were also shown in Fig.2. Furthermore, the detailed RMSE and the Coefficient of determination R^2 of the RO-ELM models with different values of p on the four noise artificial datasets were given in Table 1. On the Gaussian, the SαS and the SαSG datasets, the testing

[*1] https://archive.ics.uci.edu/ml/index.php

RMSE of the RO-ELM model with p=1.6 was the least, and the training and testing R^2 of the RO-ELM model with p=1.6 was the highest.

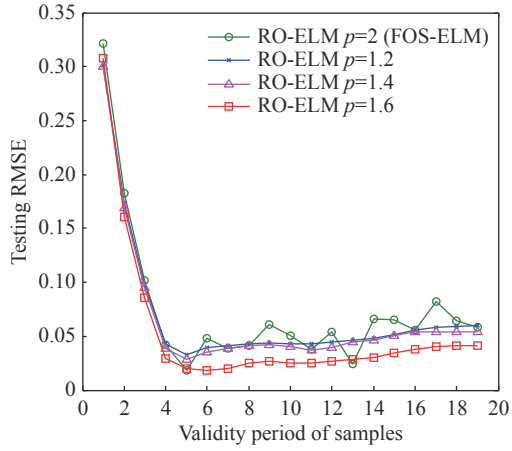


Fig. 2. The testing RMSE of the RO-ELM models with different values of p and different validity periods of samples on the artificial datasets with the SαSG noise.

Table 1. The RMSE and the coefficient of determination R^2 of the RO-ELM models with different values of p on the four noise artificial datasets

Noise type	Value of p	Training set		Testing set	
		RMSE	R^2	RMSE	R^2
Gaussian	1.2	0.2474	0.7342	0.0261	0.9423
	1.4	0.2535	0.7333	0.0251	0.9521
	1.6	**0.2364**	0.7354	**0.0164**	**0.9937**
	1.8	0.2445	**0.7422**	0.0213	0.9733
	2.0	0.2634	0.7334	0.0259	0.9745
SαS	1.2	0.7328	0.5844	0.0602	0.9502
	1.4	0.7378	0.5645	0.0503	0.9522
	1.6	**0.7242**	0.5767	**0.0205**	**0.9864**
	1.8	0.7569	0.5645	0.3451	0.9623
	2.0	0.8035	**0.7367**	0.0423	0.9604
SαSG	1.2	**0.7028**	**0.7849**	0.0567	0.9532
	1.4	0.7887	0.7543	0.0671	0.9487
	1.6	0.8032	0.7763	**0.0383**	**0.9613**
	1.8	0.7969	0.5695	0.0425	0.9552
	2.0	0.8105	0.5346	0.0424	0.9513

The RMSE and the Coefficient of determination R^2 of the ELM, the OS-ELM, the FOS-ELM and the RO-ELM models on the four noisy artificial datasets were summarized in Table 2. On the Gaussian noise dataset, the training RMSE and R^2 of the four kinds of models were quite similar. The testing RMSE of the RO-ELM model was less than that of the ELM, the OS-ELM and the RO-ELM models. The testing R^2 of the RO-ELM model was higher than that of other three models. On the SαS noise dataset, the training RMSE of the RO-ELM model was the least, and the training R^2 of the four kinds of models were quite similar. The testing RMSE of the RO-ELM model was the least, and the testing R^2 of it model was the highest. On the SαSG noise datasets, the training RMSE and R^2 of the four kinds of models were quite similar. The testing RMSE of the RO-ELM model was the least, and the testing R^2 of the four kinds of models were quite similar. The testing RMSE of the RO-ELM model was the least, and the testing R^2 of it was the highest. In a word, on the four noisy artificial datasets, the performances of the RO-ELM models were better than those of the other three models.

Table 2. The RMSE and the Coefficient of determination R^2 of the ELM, the OS-ELM, the FOS-ELM and the RO-ELM models on the four noisy artificial datasets

Noise type	Algorithms	Training set		Testing set	
		RMSE	R^2	RMSE	R^2
Gaussian	ELM	0.2474	0.7342	0.0247	0.9701
	OS-ELM	**0.2345**	0.7334	0.0239	0.9710
	FOS-ELM (p=2)	0.2634	0.7334	0.0259	0.9745
	RO-ELM (p=1.6)	0.2464	**0.7354**	**0.0164**	**0.9937**
SαS	ELM	0.7822	0.5842	0.0338	0.9689
	OS-ELM	0.7562	0.5713	0.0359	0.9691
	FOS-ELM (p=2)	0.8035	**0.7367**	0.0423	0.9604
	RO-ELM (p=1.6)	**0.7242**	0.5767	**0.0205**	**0.9864**
SαSG	ELM	0.8023	0.7365	0.0412	0.9595
	OS-ELM	**0.7922**	0.7677	0.0423	0.9581
	FOS-ELM (p=2)	0.8105	0.5346	0.0424	0.9513
	RO-ELM (p=1.6)	0.8032	**0.7763**	**0.0383**	**0.9613**

2) On the real-word regression datasets

Three real-world regression datasets with the non-Gaussian noise were employed in this experiment, and the essential information of the real-world datasets was displayed in Table 3, where R denotes Regression and C denotes Classification.

For the Pendigits dataset, there was an initial sample chunk including 2000 samples and 68 online-updating sample chunks. For the Letter dataset, there was an initial sample chunk including 2000 samples and 140 online-updating sample chunks. For the California Housing dataset, there was an initial sample chunk including 1000 samples and 64 online-updating sample chunks. Each of these online-updating sample chunks contained 100 samples.

For the RO-ELM models, the number of the hidden nodes was set as 50 and the validity period u of the samples were 5. The RMSE and the Coefficient of determination R^2 of the RO-ELM models with different values of p on the real-word regression datasets were displayed in Table 4. The results of the testing samples were more concerned, comparing with those of the training samples. On the Pendigits and California Housing datasets, with p=1.4 the RO-ELM model obtained

Table 3. The essential information of the real-world datasets

Dataset	Task type	Category	Dimensions of input features	No. of samples	
				Training	Testing
Pendigits	R	–	16	8794	2098
Letter	R	–	16	16000	4000
California Housing	R	–	8	12640	8000
Magic04	C	2	10	15216	3804
Twonorn	C	2	20	5920	1480
Mushroom	C	2	21	7311	813
Image segmentation	C	7	19	1500	810
Vehiche	C	4	18	630	216
Satellite image	C	6	26	4435	2000

the least testing RMSE and the highest testing R^2. On the Letter dataset, with p=1.6 the RO-ELM model obtained the least testing RMSE and the second highest testing R^2.

Table 4. The RMSE and the coefficient of determination R^2 of the RO-ELM models with different values of p on the real-word regression datasets

Dataset	Value of p	Training set		Testing set	
		RMSE	R^2	RMSE	R^2
Pendigits	1.2	0.0985	**0.7487**	0.0865	0.8423
	1.4	**0.0818**	0.7043	**0.0461**	**0.9021**
	1.6	0.1085	0.5976	0.0564	0.8937
	1.8	0.1108	0.6272	0.0513	0.8951
	2.0	0.0985	0.6116	0.0719	0.8753
Letter	1.2	0.1123	0.5844	0.0546	0.7150
	1.4	0.0941	0.6645	0.0582	**0.9522**
	1.6	**0.0781**	**0.7976**	**0.0486**	0.8864
	1.8	0.0947	0.6405	0.0537	0.8629
	2.0	0.0841	0.5067	0.0777	0.7604
California Housing	1.2	0.1368	0.6006	0.0697	0.7226
	1.4	0.1067	0.6678	**0.0645**	**0.9687**
	1.6	**0.0805**	**0.8060**	0.0676	0.9094
	1.8	0.1002	0.6515	0.0589	0.8621
	2.0	0.0950	0.5225	0.0864	0.7739

The RMSE and the Coefficient of determination R^2 of the ELM, the OS-ELM, the FOS-ELM and RO-ELM models on the real-word regression datasets were shown Table 5. On the Pendigits dataset, the RO-ELM model obtained the second least training and the least testing RMSE, the highest training and testing R^2. On the Letter and California Housing datasets, the RO-ELM model obtained the least training and testing RMSE and the highest training and testing R^2.

2. The experiments for classification problems

Six real-world classification datasets with the non-Gaussian noise were employed in this experiment, and the essential information of these real-world datasets was also presented in Table 3.

For any of the Magic04, Twonorn and Mushroom datasets, there was an initial sample chunk including 1000 samples and several online updating sample chunks each of which included 200 samples. On the Magic04 dataset, the number of the hidden nodes of any ELM model was 30, and the validity periods of samples was 10. On the Twonorn dataset, the number of the hidden nodes of any ELM model was 50, and the validity periods of samples was 15. On the Mushroom dataset, the number of the hidden nodes of any ELM model was 100, and the validity periods of samples was 15.

The correct classification rates (%) of the RO-ELM models with different values of p on the binary classification datasets were displayed in Table 6. For the Magic04 dataset, the RO-ELM model with p=1.6 had the highest correct classification rates for training and testing samples. For the Twonron and Mushroom datasets, the RO-ELM model with p=1.8 had the highest correct classification rates for training and testing samples.

The correct classification rates (%) of the ELM, the OS-ELM, the FOS-ELM and the RO-ELM models on the binary classification datasets were shown Table 7. On any of the three datasets, the RO-ELM model had the highest correct classification rates for the training and testing samples.

Table 5. The RMSE and the coefficient of determination R^2 of the ELM, the OS-ELM, the FOS-ELM and the RO-ELM models on the real-word regression datasets

Dataset	Algorithm	Training set		Testing set	
		RMSE	R^2	RMSE	R^2
Pendigits	ELM	0.1045	0.6091	0.1085	0.8701
	OS-ELM	**0.0834**	0.6074	0.0806	0.8710
	FOS-ELM (p=2)	0.0984	0.6169	0.0776	0.8805
	RO-ELM (p=1.4)	0.0941	**0.6645**	**0.0582**	**0.9522**
Letter	ELM	0.1191	0.5821	0.1132	0.6589
	OS-ELM	0.0993	0.5721	0.0912	0.8691
	FOS-ELM (p=2)	0.0875	0.7455	0.0794	0.8674
	RO-ELM (p=1.6)	**0.0781**	**0.7976**	**0.0486**	**0.8864**
California Housing	ELM	0.1634	0.6012	0.1143	0.6627
	OS-ELM	0.1113	0.5913	0.0933	0.8785
	FOS-ELM (p=2)	0.0894	0.7508	0.09869	0.6312
	RO-ELM (p=1.4)	**0.0805**	**0.8060**	**0.0676**	**0.9094**

Table 6. The correct classification rates (%) of the RO-ELM models with different values of p on the binary classification datasets

Dataset p	Magic04		Twonron		Mushroom	
	Training	Testing	Training	Testing	Training	Testing
1.2	82.33	85.43	81.54	82.47	82.54	82.41
1.4	92.72	93.27	82.64	84.74	83.64	84.34
1.6	**93.12**	**94.21**	88.62	89.26	88.42	88.16
1.8	82.34	85.33	**93.13**	**93.53**	**92.11**	**93.43**
2.0	88.57	89.65	88.04	88.34	88.04	88.35

Table 7. The correct classification rates (%) of the ELM, the OS-ELM, the FOS-ELM and the RO-ELM models on the binary classification datasets

Datasets	Algorithm	Training	Testing
Magic04	ELM	81.45	87.34
	OS-ELM	82.34	88.54
	FOS-ELM	88.65	89.46
	RO-ELM (p=1.6)	**93.12**	**94.21**
Twonron	ELM	84.04	85.36
	OS-ELM	86.04	87.44
	FOS-ELM	88.61	89.31
	RO-ELM (p=1.8)	**93.13**	**93.53**
Mushroom	ELM	84.44	81.36
	OS-ELM	86.34	83.44
	FOS-ELM	87.61	89.31
	RO-ELM (p=1.8)	**92.11**	**93.43**

V. Conclusions

The main innovation of this paper is that a novel method, named the robust online extreme learning machine (RO-ELM), is proposed to handle the non-Gaussian noise and discarding obsolescence samples simultaneously. The RO-ELM inherits the ELM's advantages on fast learning speed and simple architecture, and has the similar computational complexity as the OS-ELM and the FOS-ELM. In the RO-ELM, the cost function based on the LMP criterion provides a mechanism to tolerate the non-Gaussian noise, and the forgetting mechanism is applied to discard the obsolescence samples. Experiments on the artificial and real-world datasets from regression and classification problems were carried out. The results showed that the RO-ELM were more robust than the ELM, the OS-ELM and the FOS-ELM on the artificial samples with the Gaussian, the non-Gaussian, the SαS and the SαSG noise, respectively. On the real-world datasets with the non-Gaussian noise, the online models estimated by the RO-ELM outperform those estimated by the other three algorithms on the accuracy and generalization.

References

[1] B. Zou, X. Shan, C. Zhu, *et al.*, "Deep learning and its application in diabetic retinopathy screening," *Chinese Journ-*

al of Electronics, vol.29, no.6, pp.992–1000, 2020.

[2] Y. Yang, Q. M. J. Wu, X. Feng, *et al.*, "Recomputation of the dense layers for performance improvement of DCNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.11, pp.2912–2925, 2020.

[3] X. Yuan, P. He, Q. Zhu, *et al.*, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol.30, no.9, pp.2805–2824, 2019.

[4] G. Cai, G. Lyu, Y. Lin, *et al.*, "Multi-level deep correlative networks for multi-modal sentiment analysis," *Chinese Journal of Electronics*, vol.29, no.6, pp.1025–1038, 2020.

[5] Y. Zhang, K. Liu, Q. Zhang, *et al.*, "A combined-convolutional neural network for Chinese news text classification," *CTA Electonica Sinica*, vol.49, no.6, pp.1059–1067, 2021.

[6] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol.70, no.1-3, pp.489–501, 2006.

[7] G. An, Z. Jiang, X. Cao, *et al.*, "Short-term wind power prediction based on particle swarm optimization-extreme learning machine model combined with Adaboost algorithm," *IEEE Access*, vol.9, pp.94040–94052, 2021.

[8] J. Zhao, W. Wang, Q. Sun, *et al.*, "CSELM-QE: A composite semi-supervised extreme learning machine with unlabeled RSS quality estimation for radio map construction," *Chinese Journal of Electronics*, vol.29, no.6, pp.1016–1024, 2020.

[9] F. Golestaneh, P. Pierre, and H. B. Gooi, "Very short-term nonparametric probabilistic forecasting of renewable energy generation-with application to solar energy," *IEEE Transactions on Power Systems*, vol.31, no.5, pp.3850–3863, 2016.

[10] N. Liang, G. B. Huang, P. Saratchandran, *et al.*, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on Neural Networks*, vol.17, no.6, pp.1411–1423, 2006.

[11] J. Zhao, Z. Wang, and D. S. Park, "Online sequential extreme learning machine with forgetting mechanism," *Neurocomputing*, vol.87, pp.79–89, 2012.

[12] Q. Y. Zou, X. J. Wang, C. J. Zhou, *et al.*, "The memory degradation based online sequential extreme learning machine," *Neurocomputing*, vol.275, pp.2864–2879, 2018.

[13] J. Yang, F. Ye, H. J. Rong, *et al.*, "Recursive least mean p-power extreme learning machine," *Neural Networks*, vol.91, pp.22–33, 2017.

[14] S. M. Jung and P. G. Park, "Stabilization of a bias-compensated normalized least mean square algorithm for noisy inputs," *IEEE Transactions on Signal Processing* , vol.65, no.11, pp.2949–2961, 2017.

[15] B. Chen, L. Xing, Z. Wu, *et al.*, "Smoothed least mean p-power error criterion for adaptive filtering," *Digital Signal Processing*, vol.40, pp.154–163, 2015.

[16] S. C. Pei and C. C. Tseng, "Least mean p-power error criterion for adaptive FIR flter," *IEEE Journal on Selected Areas in Communications*, vol.12, no.9, pp.1540–1547, 1994.

[17] H. Sadreazami, M. O. Ahmad, and M. N. S. Swamy, "Despeckling of synthetic aperture radar images in the contourlet domain using the alpha-stable distribution," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, Lisbon, Portugal, pp.121–124, 2015.

[18] Y. Chen, "Suboptimum detectors for AF relaying with gaussian noise and interference," *IEEE Transactions on Vehicular Technology*, vol.64, no.10, pp.4833–4839, 2015.

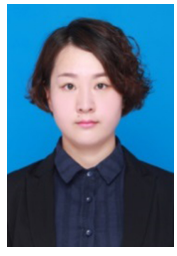
[19] J. IIow, D. Hatzinakos, and A. N. Venetsanopoulos, "Performance of FHSS radio networks with interference modeled as a mixture of Gaussian and alpha-stable noise," *IEEE Transactions on Communications*, vol.46, no.4, pp.509–520, 1998.

[20] R. D. Pierce, "Application of the positive alpha-stable distribution," in *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics (HOST)*, Banff, AB, Canada, pp.420–424, 1997.

GU Jun is a Senior Experimenter in Telecommunications Experiment Center, School of Information Engineering, Dalian Ocean University. Her research interests include electrical and electronic technology and electrical automation.

ZOU Quanyi is pursuing the Ph.D. degree in the School of Software Engineering, South China University of Technology, Guangzhou, China. His current research interests include machine learning, transfer learning, and software defect prediction.

DENG Changhui is a Professor at Dalian Ocean University, Dalian, China. His research interests include complex process modeling and control, artificial intelligence technology, and intelligent detection.



WANG Xiaojun (corresponding author) received the B.S. degree in automation from Dalian Ocean University, Dalian, China, in 2009. She received the M.S. and Ph.D. degrees in control theory and control engineering from Northeastern University, Shenyang, China, in 2011 and 2016, respectively. She is a Lecturer of School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, China. Her research interests include recommender system, deep learning, machine learning, modeling of industrial processes and soft sensors. She is the first or the corresponding author of about 20 papers.
(Email: wxjjessicaxj0903@126.com)