Neurocomputing 230 (2017) 345–358

Contents lists available at ScienceDirect

Neurocomputing

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

Robust regularized extreme learning machine for regression using iteratively reweighted least squares

Kai Chen[a,*], Qi Lv[a], Yao Lu[b], Yong Dou[a]

[a] National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China
[b] College of Computer, National University of Defense Technology, Changsha, China

ARTICLE INFO

Communicated by Manoj Kumar Tiwari

MSC:
00-01
99-00

Keywords:
Extreme learning machine
Iteratively reweighted least squares
Robustness
ℓ_2-norm regularization
ℓ_1-norm regularization

ABSTRACT

Extreme learning machine (ELM) for regression has been used in many fields because of its easy-implementation, fast training speed and good generalization performance. However, basic ELM with ℓ_2-norm loss function is sensitive to outliers. Recently, ℓ_1-norm loss function and Huber loss function have been used in ELM to enhance the robustness. However, the ℓ_1-norm loss function and the Huber loss function can also be effected by outliers because of their linear correlation with the errors. Moreover, existing robust ELM methods only use ℓ_2-norm regularization or have no regularization term. In this study, we propose a unified model for robust regularized ELM regression using iteratively reweighted least squares (IRLS), and call it RELM-IRLS. We perform a comprehensive study on the robust loss function and regularization term for robust ELM regression. Four loss functions (i.e., ℓ_1-norm, Huber, Bisquare and Welsch) are used to enhance the robustness, and two types of regularization (ℓ_2-norm and ℓ_1-norm) are used to avoid overfitting. Experiments show that our proposed RELM-IRLS with ℓ_2-norm and ℓ_1-norm regularization is stable and accurate for data with $0 \sim 40\%$ outlier levels, and that RELM-IRLS with ℓ_1-norm regularization can obtain a compact network because of the highly sparse output weights of the network.

1. Introduction

The extreme learning machine (ELM) [1] is proposed for training single-hidden layer feedforward networks (SLFNs). It directly approximates nonlinear mapping of input data by randomly generating the hidden node parameters without tuning. This model has been proven to exhibit the universal approximation capability [2]. ELM has the following merits: (1) *easy-implementation*, (2) *extremely fast training speed*, (3) *good generalization performance*. ELM has recently gained increasing interest in regression problems, such as stock market forecasting [3], electricity price forecasting [4], wind power forecasting [5], affective analogical reasoning [6], because of the aforementioned merits.

The performance of ELM regression crucially relies on the given labels of training data. The basic ELM with the ℓ_2-norm loss function assumes that the training labels is a normal error distribution. However, training samples for real tasks cannot be guaranteed to have a normal error distribution. Many factors can corrupt the training samples with outliers, such as instrument errors, sample errors and modeling errors. The performance of basic ELM regression is heavily deteriorated because ℓ_2-norm loss can be easily effected by the large deviations of the outliers.

To solve this problem, Deng et al. [7] proposed a regularized ELM with weighted least square to enhance the robustness. Their algorithm consists of two stages of the reweighted ELM. Zhang et al. [8] proposed the outlier-robust ELM with the ℓ_1-norm loss function and the ℓ_2-norm regularization term. They used augmented Lagrange multiplier algorithm to solve the objective loss function and effectively reduced the influence of outliers. Horata et al. [9] adopted the Huber function to enhance the robustness. They used iteratively reweighted least squares (IRLS) algorithm to solve the Huber loss function without a regularization term. The model without regularization is easy to overfit.

However, the loss functions of existing robust ELM regression, namely, ℓ_1-norm or Huber function, can also be effected by the outliers with large deviations because ℓ_1-norm or Huber loss functions are linear with the deviations. Moreover, existing robust ELM methods use only ℓ_2-norm regularization or have no regularization term. When the number of hidden nodes is large, the ℓ_2-norm regularization will train a large ELM model due to non-zero output weights of the network. In a word, there lacks a study that considers different loss functions and regularization terms simultaneously.

Thus, we conduct a comprehensive study on the loss function and regularization term of the robust ELM regression in this work. We

* Corresponding author.
E-mail address: kaenchan@163.com (K. Chen).

http://dx.doi.org/10.1016/j.neucom.2016.12.029
Received 21 February 2016; Received in revised form 15 June 2016; Accepted 8 December 2016
Available online 13 December 2016

propose a unified model for robust regularized ELM regression using IRLS (RELM-IRLS). Four loss functions (i.e., ℓ_1-norm, Huber, Bisquare and Welsch) are used to enhance the robustness, and two types of regularization (ℓ_2-norm and ℓ_1-norm) are used to avoid overfitting. These loss functions, also known as M-estimation functions, have been widely used in robust statistics [10]. IRLS is used to optimize the objective function with robust loss function and regularization term. Each IRLS iteration is equivalent to solving a weighted least-squares ELM regression. Our RELM-IRLS algorithm can also be trained efficiently because of the fast training speed of ELM. The experimental results on synthetic and real data sets show that our proposed RELM-IRLS is stable and accurate at $0 \sim 40\%$ outlier levels.

Compared to existing ELM methods for robust regression, the main contributions of this paper are highlighted as follows:

(1) A unified model is proposed for robust regularized ELM regression. Different kinds of robust loss functions and regularization terms can be used in this model.
(2) RELM-IRLS with ℓ_2-norm regularization is proposed to achieve better generalization.
(3) RELM-IRLS with ℓ_1-norm regularization is proposed to realize better generalization performance and more compact network architecture.

The rest of this paper is organized as follows. Basic ELM and its robust variants are reviewed in Section 2. In Section 3, we present the unified model and the RELM-IRLS with ℓ_2-norm regularization and ℓ_1-norm regularization. Section 4 demonstrates the experimental results of our proposed algorithms and Section 5 presents our conclusion.

2. Background and related works

2.1. ELM for regression

For a given set of training samples $S = \{(\mathbf{x}^{(i)}, y^{(i)})|i=1,..,N\}\subset\mathbb{R}^d \times \mathbb{R}$ for regression problem, ELM is a unified SLFN whose output with L hidden nodes can be represented as

$$f_L(\mathbf{x}) = \sum_{i=1}^{L} h_i(\mathbf{x})\boldsymbol{\beta}_i = \mathbf{h}(\mathbf{x})\beta, \quad \mathbf{x} \in \mathbb{R}^d \tag{1}$$

where the $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_L(\mathbf{x})]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_L]^{\top}$. $h_i(\mathbf{x})$ is the hidden layer function $g(\mathbf{a}_i, b_i, \mathbf{x})$ between the input layer and the ith hidden node. $\mathbf{a}_i$ and b_i are randomly generated independent of the training data. β_i is the output weight between the ith hidden node to the output node. The computational hidden nodes can be sigmoid function, additive and radial basis function(RBF) hidden nodes, hinging functions, wavelets, and so on. For example, $g(\mathbf{a}_i, b_i, \mathbf{x}) = \sigma(\mathbf{a}_i^{\top}\mathbf{x} + b_i)$ where σ is the sigmoid function. To minimize the least square error, we can get the objective loss function:

$$\min_{\beta}: \quad \| \mathbf{H}\boldsymbol{\beta} - \mathbf{Y} \|^2 \tag{2}$$

where

$$\mathbf{H} = [\mathbf{h}(\mathbf{x}^{(1)})^{\top}, \mathbf{h}(\mathbf{x}^{(2)})^{\top}, \ldots, \mathbf{h}(\mathbf{x}^{(N)})^{\top}]^{\top}, \mathbf{Y} = [y^{(1)}, y^{(2)}, \ldots, y^{(N)}]^{\top}.$$

The solution for Eq. (2) is provided by Huang et al. [1] as

$$\boldsymbol{\beta} = \begin{cases} (\mathbf{H}^{\top}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{Y}, & N \geq L \\ \mathbf{H}^{\top}(\mathbf{H}\mathbf{H}^{\top})^{-1}\mathbf{Y}, & N < L \end{cases} \tag{3}$$

where the two forms of $\boldsymbol{\beta}$ are equivalent based on Woodbury identity [11].

ELM is easy to overfit based on the empirical risk minimization (ERM) principle in Eq. (2). In [12] and [13], a regularized ELM is proposed to consider Structural Risk Minimization (SRM) rather than ERM only. The complexity of ELM is controlled by restricting the output weights β to small values. Thus, the objective function is

$$\min_{\beta}: \frac{1}{2} \| \boldsymbol{\beta} \|_2^2 + C\frac{1}{2}\sum_{i=1}^{N} \| \boldsymbol{\xi}^{(i)} \|_2^2 \text{ s. t. : } \mathbf{h}(\mathbf{x}^{(i)})\boldsymbol{\beta} = y^{(i)} - \boldsymbol{\xi}^{(i)}, \quad \forall\ i. \tag{4}$$

where C is the regularization parameter that trades off the norm of output weights and least squares training errors. According to [12], the solution of Eq. (4) is

$$\boldsymbol{\beta} = \begin{cases} (\mathbf{H}^{\top}\mathbf{H} + \frac{\mathbf{I}}{C})^{-1}\mathbf{H}^{\top}\mathbf{Y}, & N \geq L \\ \mathbf{H}^{\top}\left(\mathbf{H}\mathbf{H}^{\top} + \dfrac{\mathbf{I}}{C}\right)^{-1}\mathbf{Y}, & N < L \end{cases} \tag{5}$$

2.2. Robust ELM for regression

Deng et al. [7] proposed a regularized ELM with weighted least square to enhance the robustness. Their algorithm consists of two stages. A regularized ELM is trained in the first stage. Then, the weights of samples are calculated by the residual errors of the first model and the regularized ELM is retrained using the weighted samples during the second stage. This algorithm can be simply regarded as two iteration version of the iteratively reweighted method. This method may not get a convergence of only two iterations. In our method, we take the full iterations of IRLS to achieve better performance.

Horata et al. [9] proposed three robust ELM algorithms: (1) ELM based on the IRLS, where the weights of samples are iteratively recalculated by prior residual errors; (2) ELM based on the multivariate least rimmed squares (MLTS-ELM), where a subset of noisy training data is selected by minimum covariance determinant (MCD) estimator; and (3) ELM based on one-step reweighted MLTS (RMLTS-ELM), where the reweighted version of MLTS is used. The first algorithm is very similar to our proposed algorithm. However, it uses only one M-estimate function (i.e. Huber function) and lacks a regularization term. In our algorithm, more robust loss functions, namely, Bisquare and Welsch function are used. In addition, we used two kinds of regularization, namely, ℓ_2-norm and ℓ_1-norm regularization.

Zhang et al. [8] proposed an outlier-robust ELM, which adopted the ℓ_1-norm loss function to enhance the robustness and ℓ_2-norm regularization term to achieve good generalization. An augmented Lagrange multiplier algorithm based on the alternating direction technique was used to solve the objective loss function. Although the ℓ_1 loss function was robust to outlier, but it will still be effected by the outliers because it is linear with error. In our algorithm, more robust M-estimate loss functions are used, which can reduce the effects of larger errors of outliers.

3. Proposed method

In this section, we explain our proposed RELM-IRLS algorithm. First, we provide a unified model for robust regularized regression. Then, we present the RELM-IRLS with ℓ_2-norm and ℓ_1-norm regularization respectively. Finally, we discuss the advantages and disadvantages of our proposed method.

3.1. A unified model for robust regularized ELM regression

Given a regression data set $S = \{(\mathbf{x}^{(i)}, y^{(i)})|i = 1,..,N\} \subset \mathbb{R}^d \times \mathbb{R}$, the objective loss function of ELM has the general form:

$$\min_{\beta}: r(\beta) + C\sum_{i=1}^{N} \ell(\mathbf{h}(\mathbf{x}^i)\beta, y^i) \tag{6}$$

where $\beta \in \mathbb{R}^L$ is the output weight vector, $r(\cdot)$: $\mathbb{R}^L \to \mathbb{R}$ is a regularization function, and $\ell(\cdot, \cdot)$: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function. The parameter