

基于指数 Laplace 损失函数的 回归估计鲁棒超限学习机^{*}

王快妮^{1,2}, 曹进德², 刘庆山²

(1. 西安石油大学 理学院, 西安 710065;

2. 东南大学 数学学院, 南京 211189)

(本刊编委曹进德来稿)

摘要: 实际问题的数据集通常受到各种噪声的影响, 超限学习机 (extreme learning machine, ELM) 对这类数据集进行学习时, 表现出预测精度低、预测结果波动大。为了克服该缺陷, 采用了能够削弱噪声影响的指数 Laplace 损失函数, 该损失函数是建立在 Gauss 核函数基础上, 具有可微、非凸、有界且能够趋近于 Laplace 函数的特点。将其引入到超限学习机中, 提出了鲁棒超限学习机回归估计 (exponential Laplace loss function based robust ELM for regression, ELRELM) 模型。利用迭代重赋权算法求解模型的优化问题。在每次迭代中, 噪声样本点被赋予较小的权值, 能够有效地提高预测精度。真实数据集实验验证了所提出的模型相比较于对比算法具有更优的学习性能和鲁棒性。

关键词: 神经网络; 超限学习机; 鲁棒; 指数 Laplace 损失函数; 迭代重赋权算法

中图分类号: TP181

文献标志码: A

DOI: 10.21656/1000-0887.400240

引言

超限学习机 (ELM) 是由 Huang 等^[1-3]提出的一种单隐层前馈式神经网络算法。与传统的神经网络训练算法不同, ELM 对隐节点的输入权值随机赋值, 通过求解线性方程组得到输出权值。ELM 具有训练速度极快、泛化性能强和调节参数较少等优点, 已广泛地应用于机器学习等领域^[3-5]。在实际问题中, 由于受诸多因素的影响, 数据集通常包含噪声^[6-7], ELM 对这类数据集的预测精度低且预测结果波动较大。

正则化 ELM^[3]将输出权值的范数作为目标函数的正则化项, 通过最小化结构风险, 显著提高了 ELM 的学习性能。根据训练误差, Deng 等^[8]提出加权正则化 (weighted ELM, WELM), 通过两次迭代加权方法来削弱噪声的影响。然而, WELM 在很大程度上依赖于初始迭代中学习的准确性, 并且未必仅在两次迭代后收敛。上述方法都采用 2-范数损失函数来构造 ELM 模型。该损失函数只有在训练点误差服从正态分布时是最优的。然而, 许多实际应用问题并不能保证误差服从正态分布, 这导致超限学习机对噪声样本点较敏感、鲁棒性差。

* 收稿日期: 2019-08-20

基金项目: 国家自然科学基金 (61833005; 61907033); 中国博士后科学基金 (2018M642129)

作者简介: 王快妮 (1982—), 女, 讲师, 博士 (E-mail: wangkuaini1219@sina.com);

曹进德 (1963—), 男, 教授, 博士, 博士生导师, 欧洲科学院院士 (通讯作者。E-mail: jdcao@seu.edu.cn)。

在机器学习领域,鲁棒损失函数是鲁棒算法设计的关键思路之一.研究人员提出了 Huber 损失函数^[9]、1-范数损失函数^[10]和 Pinball 损失函数^[11]及其相应的改进 ELM 模型.然而,Huber、1-范数和 Pinball 损失函数的鲁棒性仍然较差,因为它们与训练误差之间是线性关系.Chen 等^[12]同时考虑正则化项(2-范数或 1-范数)和不同的损失函数(1-范数、Huber、bisphere 和 Welsch),提出了统一的鲁棒正则化 ELM,并通过迭代加权算法求解所建模型.该方法在不同离群值比例下具有对数据集稳定且良好的泛化性能.除了文献[12]中提到的 bisphere 和 Welsch 函数外,上述方法均采用凸损失函数.

随着研究的深入,研究人员发现基于非凸损失函数的机器学习算法对受噪声干扰的数据集具有较强的鲁棒性^[13-16].非凸损失函数也称为 ramp 损失函数^[13]、truncated 损失函数^[14]或 capped 损失函数^[16].经典机器学习方法中的损失函数如 hinge 损失、 ε -敏感损失、2-范数被非凸损失函数所取代,构造了相应的鲁棒学习算法.近年来,随着信息论的不断发展,相关熵被用来解决监督和非监督学习问题^[17-18].相关熵是由 Liu 等^[17]于 2007 年提出的一种建立在 Gauss 核函数上的非线性局部相似性度量,能够弱化噪声在学习过程中的作用,已经被成功地应用于鲁棒学习和信号处理领域^[18].Xing 等^[19]提出了基于最大相关熵准则的 ELM 模型,利用半二次规划求解所建模型,数值实验表明所提出的方法优于 ELM 和正则化 ELM.由相关熵衍生的 C-损失函数^[20]和非凸光滑损失^[21]及其相应学习模型被证实对噪声数据具有较强的鲁棒性.非凸光滑损失^[21]是文献[12]中 Welsch 函数的另一种形式.作为建立在 Gauss 核函数上的度量,相关熵及其衍生损失函数^[19-21]本质上是非凸指数函数,且指数部分均为 2-范数.考虑到 2-范数损失函数过分强调训练误差的作用,文献[11,22]提出了基于相关熵 Laplace 损失函数的 ELM 分类模型^[11]和支持向量机模型^[22].该模型^[11]虽然能够弱化噪声对学习性能的影响,但需要迭代求解二次规划,这种方式不仅计算复杂度高,而且丧失了 ELM 模型运行速度极快且易于实现的显著优势.

受上述研究的启发,本文在相关熵的基础上,提出指数 Laplace 损失函数,并将其引入到 ELM 框架,建立能够弱化噪声负面作用的鲁棒 ELM 回归估计(ELRELM)模型.该模型的损失函数具有对称、非凸、有界且能够趋近于 Laplace 损失函数的性质,可通过迭代求解加权 ELM 实现.在每次迭代中,只需通过求解线性方程组就能得到模型的最优解,能够保持 ELM 解的结构优势.同时,在学习过程中,噪声样本点被赋予较小的权值,能够有效地提高预测精度.真实数据集在不同比例噪声下的数值实验验证了提出的 ELRELM 模型的有效性.

1 超限学习机

已知回归估计训练集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N, \mathbf{x}_i \in R^d, y_i \in \mathbf{R}$.具有 L 个隐节点和激活函数 $h_i(\mathbf{x})$ 的单隐层 ELM 的输出函数的数学表示如下:

$$f(\mathbf{x}) = \sum_{i=1}^L h_i(\mathbf{x})\boldsymbol{\beta} = \mathbf{h}(\mathbf{x})\boldsymbol{\beta}, \quad (1)$$

其中 $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_L]^T$ 是输出权值向量, $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_L(\mathbf{x})]$ 为变量 \mathbf{x} 的隐层输出.记 $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$, 隐层输出矩阵 $\mathbf{H} = [h(\mathbf{x}_1), h(\mathbf{x}_2), \dots, h(\mathbf{x}_N)]^T$, ELM 模型^[2-3]可表达为如下优化问题:

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{2} \|\boldsymbol{\beta}\|^2 + \frac{C}{2} \|\mathbf{H}\boldsymbol{\beta} - \mathbf{y}\|^2 \right), \quad (2)$$

其中 C 为正则化参数.由最优性条件,优化问题(2)的最优解为

$$\beta = \begin{cases} (\mathbf{H}^T \mathbf{H} + \mathbf{I}/C)^{-1} \mathbf{H}^T \mathbf{y}, & N \geq L, \\ \mathbf{H}^T (\mathbf{H} \mathbf{H}^T + \mathbf{I}/C)^{-1} \mathbf{y}, & N < L, \end{cases} \quad (3)$$

其中 \mathbf{I} 为单位矩阵。

2 指数 Laplace 损失函数及性质

2.1 指数 Laplace 损失函数

在信息理论学习中,相关熵被认为是一个广义的局部相似度量.它是建立在 Gauss 核函数 $g(z) = e^{-z^2/\lambda}, \lambda > 0$ 基础上的非凸函数. Singh 等^[20] 和 Feng 等^[21] 分别建立了相关熵的衍生函数: C-损失函数和非凸光滑损失函数.它们都是非凸指数函数,能够明确限定噪声数据所造成的损失上界,抑制噪声对学习模型的影响.注意到这两个衍生函数的指数部分都是 2-范数,这将在学习过程中过分强调噪声数据的负面作用.为了削弱 2-范数函数对噪声的敏感性,文献[11, 22]提出了基于相关熵 Laplace 函数的支持向量机模型和超限学习机分类模型.针对回归估计问题,本文提出指数 Laplace 损失函数:

$$l_\lambda(z) = \lambda(1 - e^{-|z|/\lambda}), \quad (4)$$

其中 $\lambda > 0$. 图 1 给出了不同 λ 值下的指数 Laplace 损失函数 $l_\lambda(z)$ 的图形.

2.1 性质

性质 1 指数 Laplace 损失函数 $l_\lambda(z)$ 非负、对称、有界,且满足

$$0 < l_\lambda(z) \leq \lambda, \lim_{z \rightarrow +\infty} l_\lambda(z) = \lambda.$$

性质 2 除 $z = 0$ 外, $l_\lambda(z)$ 可微且其导数有界,且 $\lim_{z \rightarrow +\infty} l'_\lambda(z) = 0$.

分析 $l'_\lambda(z) = \text{sgn}(z) e^{-|z|/\lambda}$, 易得到 $\lim_{z \rightarrow +\infty} l'_\lambda(z) = 0$. 从 M-估计的角度,当数据集存在噪声即误差 $|z|$ 较大时, $l_\lambda(z)$ 是鲁棒损失函数.

性质 3 当 $k = 1$ 时, $\lim_{\lambda \rightarrow +\infty} l_\lambda(z) = |z|$.

分析 由 Taylor 展开式 $e^z = 1 + \sum_{k=1}^{+\infty} \frac{z^k}{k!}$, 得到

$$\lim_{\lambda \rightarrow +\infty} l_\lambda(z) = \left[\sum_{k=1}^{+\infty} \frac{(-1)^k + 1 \lambda^{-k} |z|^k}{k!} \right] = \begin{cases} |z|, & k = 1, \\ 0, & k \geq 2. \end{cases}$$

因此,当 $k = 1$ 时, $\lim_{\lambda \rightarrow +\infty} l_\lambda(z) = |z|$.

3 基于指数 Laplace 损失函数的回归估计鲁棒 ELM

及其迭代求解算法

由图 1 和 $l_\lambda(z)$ 的性质可知,该有界损失函数能够有效削弱噪声对模型的影响.因此,本文提出基于指数 Laplace 损失函数 $l_\lambda(z)$ 的鲁棒 ELM 模型,其优化问题具有如下形式:

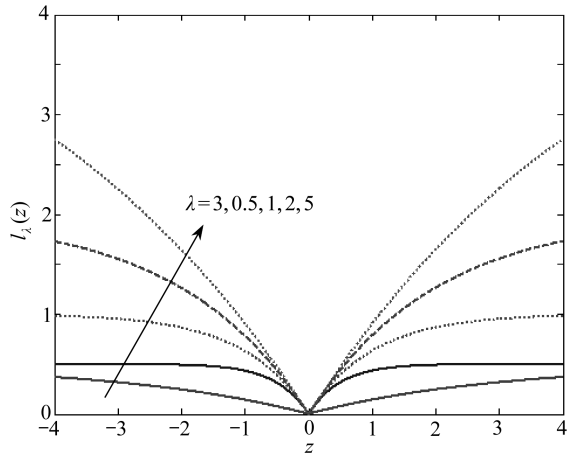


图 1 不同 λ 值下的指数 Laplace 损失函数 $l_\lambda(z)$

Fig. 1 Exponential Laplace loss function $l_\lambda(z)$ with different λ values

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N l_{\lambda}(y_i - \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta}) \right). \quad (5)$$

由于建立在相关熵 Laplace 损失函数的超限学习机^[11]需要迭代求解二次规划,该方法不仅计算复杂度高,而且丧失了 ELM 模型求解较小规模线性方程组以及易于实现的优势.为了获得更快的学习速度,同时保持 ELM 解的结构优势,本文利用迭代重赋权算法求解所建模型(5).

令

$$w_i = \frac{\partial l_{\lambda}(\xi_i)}{\partial \xi_i} / \xi_i = \frac{1}{\max(|\xi_i|, 10^{-6})} e^{-|\xi_i|/\lambda}, \quad (6)$$

其中 $\xi_i = y_i - \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta}$. 由最优性条件和文献[12],可得优化问题(5)的最优解:

$$\boldsymbol{\beta} = \begin{cases} (\mathbf{H}^T \mathbf{W} \mathbf{H} + \mathbf{I}/C)^{-1} \mathbf{H}^T \mathbf{W} \mathbf{y}, & N \geq L, \\ \mathbf{H}^T (\mathbf{W} \mathbf{H} \mathbf{H}^T + \mathbf{I}/C)^{-1} \mathbf{W} \mathbf{y}, & N < L, \end{cases} \quad (7)$$

其中对角矩阵 $\mathbf{W} = \text{diag}(w_1, w_2, \dots, w_N)$. 当 \mathbf{W} 为单位矩阵时, ELRELM 为 ELM.

算法 1 迭代重赋权算法求解 ELRELM

输入: 训练集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$.

初始化: 参数 C , 隐节点数 L , 隐层输出矩阵 \mathbf{H} , 最大迭代次数 k_{\max} 和很小的 $\rho > 0$.

令 $\mathbf{W}^{(0)} = \mathbf{I}$ 和 $k = 1$.

输出: 最优解 $\boldsymbol{\beta}^*$.

Step 1 由下式计算 $\boldsymbol{\beta}^{(k)}$:

$$\boldsymbol{\beta}^{(k)} = \begin{cases} \left(\mathbf{H}^T \mathbf{W}^{(k-1)} \mathbf{H} + \frac{\mathbf{I}}{C} \right)^{-1} \mathbf{H}^T \mathbf{W}^{(k-1)} \mathbf{y}, & N \geq L, \\ \mathbf{H}^T \left(\mathbf{W}^{(k-1)} \mathbf{H} \mathbf{H}^T + \frac{\mathbf{I}}{C} \right)^{-1} \mathbf{W}^{(k-1)} \mathbf{y}, & N < L. \end{cases} \quad (8)$$

Step 2 由式(1)得到 $f^{(k)}(\mathbf{x})$, 利用式(6)更新 $w_i^{(k)}$ 和 $\mathbf{W}^{(k)}$, 其中 $\xi_i(k) = y_i - f^{(k)}(\mathbf{x}_i)$.

Step 3 通过式(8)更新 $\boldsymbol{\beta}^{(k+1)}$.

Step 4 若 $k < k_{\max}$ 且 $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\| \geq \rho$, 则令 $k := k + 1$, 返回 step 2.

Step 5 否则终止迭代,输出 $\boldsymbol{\beta}^*$, 得到决策函数 $f(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta}^*$.

4 数值实验

为了验证本文 ELRELM 的有效性,对 9 个真实数据集进行数值实验,并与正则化 ELM^[3]、加权 ELM(WELM)^[9]和迭代重加权 ELM(iteratively re-weighted ELM, IRWELM)^[12]进行比较.正则化参数 C 来自 $\{2^{-19}, 2^{-18}, \dots, 2^{18}, 2^{19}\}$, ELRELM 中参数 λ 的搜索范围为 $\{0.05, 0.1, 0.15, \dots, 0.85, 0.9, 0.95, 1, 1.5, 2, \dots, 4, 4.5, 5\}$.所有最优参数均通过网格搜索来实现.采用 sigmoid 函数为激活函数,隐节点数目 $L = 200$, 最大迭代次数 $k_{\max} = 20$.均方根误差 $\varepsilon_{\text{RMSE}}$ 和平均绝对误差 ε_{MAE} 是评价模型学习性能的两个常用指标^[15]:

$$\varepsilon_{\text{RMSE}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2},$$

$$\varepsilon_{\text{MAE}} = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i|,$$

其中 m 为测试样本点数目, y_i 和 \hat{y}_i 分别表示测试样本点输出的真实值和预测值.通常来说,评

价指标 $\varepsilon_{\text{RMSE}}, \varepsilon_{\text{MAE}}$ 值越小,说明模型的学习性能越好.

表 1 Pyrim、servo 和 bodyfat 数据集的实验结果

Table 1 Experimental results of pyrim, servo and bodyfat datasets

experimental result	noise ratio $\delta / \%$	algorithm	(C^*, λ^*)	$\varepsilon_{\text{RMSE}}$	ε_{MAE}
pyrim	0	ELM	$2^1, -$	0.111 4±0.020 4	0.065 1±0.007 3
		WELM	$2^0, -$	0.106 0±0.029 7	0.061 2±0.007 7
		IRWELM	$2^0, -$	0.107 7±0.029 1	0.062 8±0.006 6
		ELRELM	$2^{-1}, 0.2$	0.106 1±0.026 7	0.062 5±0.006 6
	10	ELM	$2^{-8}, -$	0.135 2±0.026 4	0.103 2±0.018 7
		WELM	$2^{-2}, -$	0.110 4±0.033 1	0.067 1±0.011 4
		IRWELM	$2^{-1}, -$	0.109 4±0.033 5	0.066 6±0.011 4
		ELRELM	$2^{-4}, 0.25$	0.109 3±0.030 4	0.065 4±0.009 7
	20	ELM	$2^{-10}, -$	0.144 8±0.018 8	0.107 9±0.017 4
		WELM	$2^{-5}, -$	0.131 1±0.028 9	0.094 6±0.016 6
		IRWELM	$2^{-2}, -$	0.112 7±0.036 2	0.068 3±0.013 6
		ELRELM	$2^{-4}, 0.15$	0.111 4±0.027 3	0.068 3±0.007 4
servo	0	ELM	$2^{16}, -$	0.617 7±0.101 3	0.397 8±0.058 1
		WELM	$2^{19}, -$	0.554 7±0.168 6	0.322 5±0.069 8
		IRWELM	$2^{18}, -$	0.615 1±0.166 6	0.347 0±0.061 4
		ELRELM	$2^{15}, 0.25$	0.574 3±0.175 8	0.312 3±0.049 5
	10	ELM	$2^{10}, -$	0.868 7±0.139 8	0.633 1±0.094 9
		WELM	$2^{10}, -$	0.739 2±0.165 4	0.488 5±0.061 0
		IRWELM	$2^{12}, -$	0.738 6±0.180 1	0.437 7±0.074 2
		ELRELM	$2^7, 0.8$	0.689 5±0.201 4	0.411 0±0.080 3
	20	ELM	$2^7, -$	1.276 7±0.151 6	1.059 4±0.105 6
		WELM	$2^8, -$	0.916 6±0.158 4	0.666 2±0.100 8
		IRWELM	$2^{11}, -$	0.740 0±0.206 2	0.483 0±0.082 5
		ELRELM	$2^6, 0.2$	0.695 3±0.207 8	0.445 3±0.092 3
bodyfat	0	ELM	$2^2, -$	0.002 7±0.001 5	0.001 2±2.407 3E-4
		WELM	$2^{13}, -$	0.002 2±0.001 7	5.721 1E-4±2.311 5E-4
		IRWELM	$2^{14}, -$	0.002 1±0.001 8	4.546 0E-4±2.737 5E-4
		ELRELM	$2^4, 1$	0.002 1±0.001 8	4.607 9E-4±2.572 7E-4
	10	ELM	$2^{-15}, -$	0.022 5±0.002 8	0.017 9±0.002 4
		WELM	$2^2, -$	0.002 8±0.001 5	0.001 3±2.362 4E-4
		IRWELM	$2^{10}, -$	0.002 2±0.001 8	4.964 2E-4±2.949 4E-4
		ELRELM	$2^0, 0.15$	0.002 1±0.001 8	4.693 4E-4±2.738 6E-4
	20	ELM	$2^{-16}, -$	0.022 7±0.003 3	0.018 2±0.002 7
		WELM	$2^{-1}, -$	0.005 2±0.001 8	0.003 8±0.001 2
		IRWELM	$2^7, -$	0.002 3±0.001 8	5.536 6E-4±2.998 0E-4
		ELRELM	$2^1, 0.05$	0.002 1±0.001 8	4.751 4E-4±2.732 6E-4

在实验过程中,对每个数据集,采用文献[12]的噪声干扰方式,对训练集的输出加入不同比例的来自于 $[y_{\min}, y_{\max}]$ 的随机数噪声,测试集不加入任何噪声.对 5 个数据集 pyrim、servo、bodyfat、yacht 和 heating 在加入 0%、10% 和 20% 噪声情形进行实验.表 1 和表 2 给出了 5 个数

据集的实验结果,包括在不同比例噪声下各算法的最优参数、预测精度 $\varepsilon_{\text{RMSE}}$ 和 ε_{MAE} . 实验结果均为 10 次独立实验的平均结果,粗体表示这些算法中的最好结果.

由表 1 和表 2 可知,除了在加入 0% 噪声情形的 pyrim、servo 和 bodyfat 数据集外,ELRELM 都获得了优于 3 种对比算法的预测精度,具有最小的 $\varepsilon_{\text{RMSE}}$ 和 ε_{MAE} 值.并且,ELRELM 在这 3 个数据集没有噪声干扰情形的预测精度排名第二.当加入噪声时,这 4 种算法的预测精度均不同程度地下降,体现在各算法的 $\varepsilon_{\text{RMSE}}$ 和 ε_{MAE} 值增加.随着噪声比例的增加,ELM 的学习性能急剧下降,WELM 和 IRWELM 的预测精度也有所下降.ELRELM 始终保持稳定且较好的学习性能.以 20% 噪声情形的 Yacht 数据集为例,ELRELM 算法所获得的 $\varepsilon_{\text{RMSE}}$ 比 ELM 提高了 57.78%,比 WELM 提高了 36.14%,比 IRWELM 提高了 11.18%.

表 2 Yacht 和 heating 数据集的实验结果

Table 2 Experimental results of yacht and heating datasets

experimental result	noise ratio $\delta / \%$	algorithm	(C^*, λ^*)	$\varepsilon_{\text{RMSE}}$	ε_{MAE}
yacht	0	ELM	$2^{19}, -$	1.772 8±0.166 6	1.446 3±0.096 2
		WELM	$2^{19}, -$	1.773 9±0.168 4	1.446 6±0.096 4
		IRWELM	$2^{19}, -$	1.786 7±0.191 4	1.449 9±0.099 4
		ELRELM	$2^{19}, 1$	1.032 0±0.258 5	0.685 0±0.133 7
	10	ELM	$2^{15}, -$	6.714 0±1.081 4	4.765 2±0.752 1
		WELM	$2^{18}, -$	2.638 3±0.748 9	1.806 5±0.293 8
		IRWELM	$2^{18}, -$	2.139 7±0.460 7	1.587 4±0.152 5
		ELRELM	$2^{18}, 0.05$	1.732 1±0.616 4	1.044 5±0.256 7
	20	ELM	$2^7, -$	8.966 1±1.180 6	7.454 2±1.007 3
		WELM	$2^8, -$	5.927 4±0.922 8	4.452 1±0.577 4
		IRWELM	$2^{17}, -$	4.261 7±2.639 5	2.447 4±0.977 7
		ELRELM	$2^{11}, 0.35$	3.785 1±1.528 1	2.421 0±0.682 0
heating	0	ELM	$2^{19}, -$	0.899 9±0.030 6	0.707 6±0.027 1
		WELM	$2^{19}, -$	0.926 5±0.030 9	0.717 6±0.025 2
		IRWELM	$2^{16}, -$	1.049 5±0.051 6	0.836 3±0.030 9
		ELRELM	$2^{19}, 0.9$	0.821 7±0.137 1	0.572 8±0.042 8
	10	ELM	$2^1, -$	4.012 6±0.149 2	3.314 0±0.115 1
		WELM	$2^{15}, -$	1.976 8±0.364 7	1.386 6±0.159 5
		IRWELM	$2^{17}, -$	1.054 5±0.091 9	0.826 4±0.043 7
		ELRELM	$2^{16}, 0.1$	0.981 2±0.071 1	0.719 1±0.051 2
	20	ELM	$2^0, -$	6.080 7±0.193 4	5.420 9±0.143 1
		WELM	$2^2, -$	3.031 9±0.175 4	2.232 4±0.089 5
		IRWELM	$2^{16}, -$	1.486 2±0.706 2	1.036 7±0.230 4
		ELRELM	$2^{11}, 0.15$	1.416 1±0.474 0	0.975 0±0.179 2

为了更充分地验证 ELRELM 在噪声环境下的学习性能,进一步对 4 个数据集 triazines、pollution、MCPU 和 diabetes 在噪声比例 0%、5%、10%、15%、20%、25%、30%、35% 和 40% 情形下进行数值实验.图 2 给出了 4 个数据集在 9 种不同比例噪声情形的预测精度 $\varepsilon_{\text{RMSE}}$. 由图 2 可知,在低比例噪声环境下,ELM 在大多数情况下都取得了较好的预测精度.ELRELM 在这类情形下学习性能较差.随着噪声比例的增加,尤其是噪声比例超过 25% 时,ELRELM 的预测精度明显优于这 3 种对比算法.在 Pollution 和 MCPU 数据集上,ELRELM 所获得了最小的 $\varepsilon_{\text{RMSE}}$ 值,

且与其他3种对比算法的预测精度 $\varepsilon_{\text{RMSE}}$ 相差较大.这是因为本文的指数 Laplace 损失函数是有界函数,明确限制了噪声所造成的最大影响;同时,在训练过程中对噪声样本点赋予较小的权值,削弱这类样本点在构建回归函数过程中的作用.因此,ELRELM 对含有噪声的数据集具有较稳定的泛化能力和较强的鲁棒性.

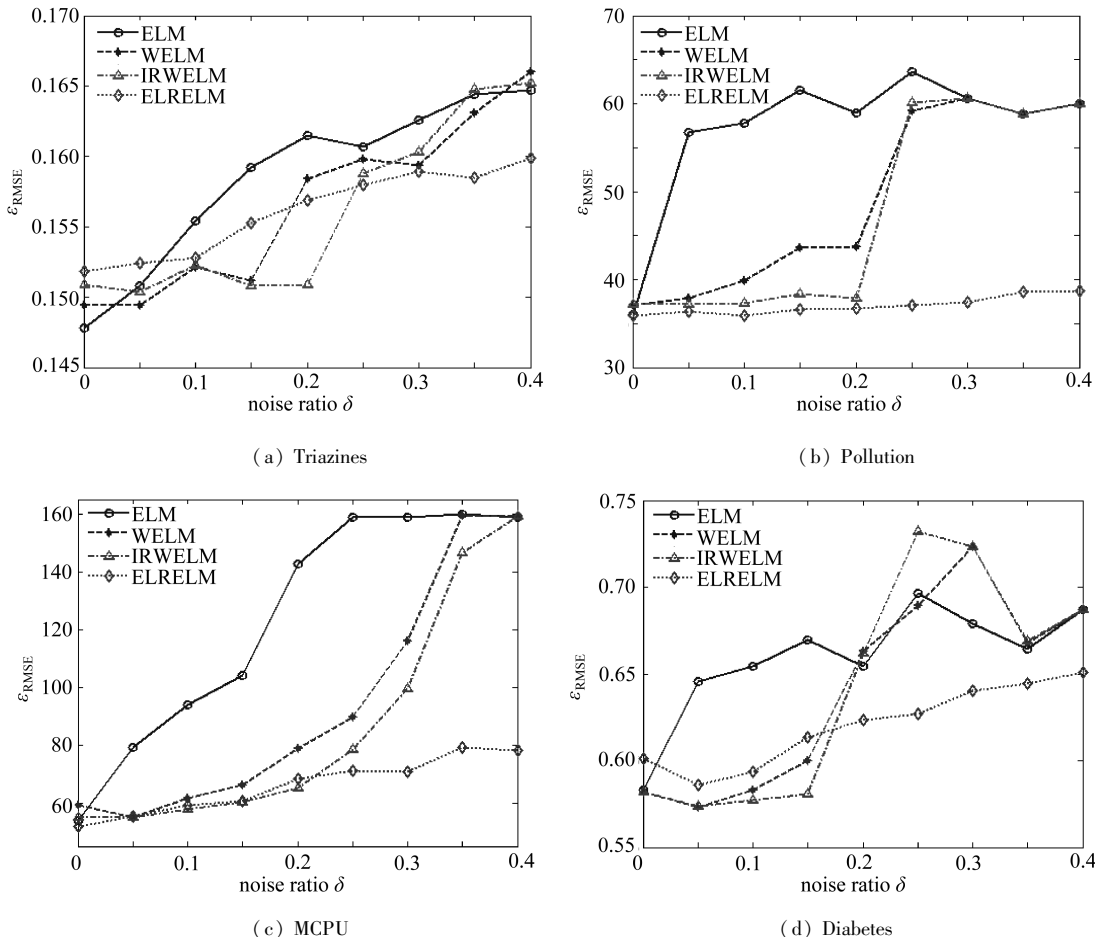
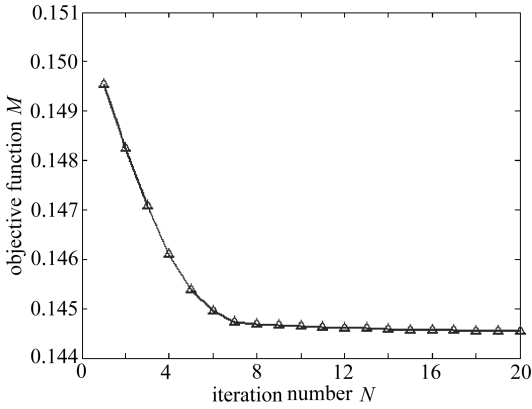


图2 Triazines、pollution、MCPU 和 diabetes 数据集在不同比例噪声下的预测精度

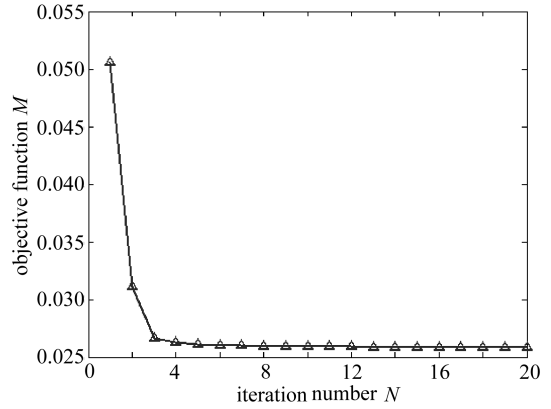
Fig. 2 Experimental results of triazines, pollution, MCPU and diabetes datasets with different noise levels

由于 ELRELM 是一个迭代求解算法,迭代次数对其收敛速度具有重要影响.通过数值实验分析迭代次数与 ELRELM 的目标函数值的关系.在实验过程中,最大迭代次数设置为 20.图 3 给出了 diabetes 和 pollution 数据集在 5% 噪声情形下迭代次数和目标函数值的影响图.根据图 3,目标函数值曲线在前几次迭代中下降很快,在迭代次数不超过 15 收敛.因此,本文提出的 ELRELM 能够在较少的迭代次数内收敛,具有较快的收敛速度.

接下来分析隐节点数目 L 对 ELRELM 算法的预测结果 $\varepsilon_{\text{RMSE}}$ 的影响.对 diabetes 和 pollution 数据集在 10% 噪声情形进行数值实验.隐节点数目 L 的取值范围为 $\{100, 200, 300, 400, 500, 600, 700, 800, 900, 1\ 000\}$.对不同的 L 值,ELRELM 在最优参数 (C^*, λ^*) 情况下得到 $\varepsilon_{\text{RMSE}}$,实验结果见图 4.由图 4 可以看出,隐节点数目对 ELRELM 的预测结果 $\varepsilon_{\text{RMSE}}$ 的影响不明显,这与文献[3]的结论一致,也是 ELM 框架算法的一个优势.因此,为了缩短训练时间,在实验中可选取较小的 L 值.本文中 $L = 200$ 就是一个较合适的选择.



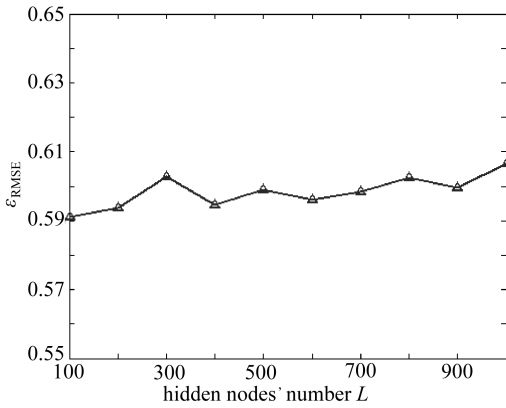
(a) Diabetes



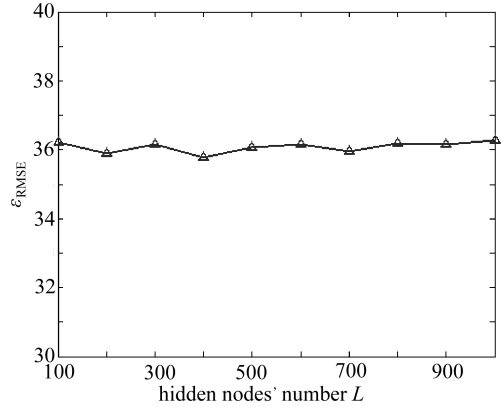
(b) Pollution

图3 Diabetes 和 pollution 数据集上 ELRELM 的迭代次数与目标函数值的关系

Fig. 3 Relationships between the iteration number and the objective function for diabetes and pollution datasets



(a) Diabetes



(b) Pollution

图4 Diabetes 和 pollution 数据集上 $\varepsilon_{\text{RMSE}}$ 值随隐节点数目 L 的变化情况Fig. 4 Relationships between hidden nodes' number L and $\varepsilon_{\text{RMSE}}$ for diabetes and pollution datasets

5 结束语

针对回归估计问题,在超限学习机框架下,提出指数 Laplace 损失函数处理受噪声干扰的数据集学习问题,并在此基础上建立了鲁棒 ELM 模型.利用迭代重赋权算法求解相应的优化问题.在每次迭代过程中,只需求解较小规模的线性方程组.求解算法能够在较少的迭代次数内收敛.通过 9 个真实数据集在不同噪声环境下的数值实验验证了所提出的 ELRELM 模型具有较稳定且更优的学习性能.对机器学习其他问题如半监督问题、无监督问题和单分类问题,均可考虑构造相应的指数损失函数,这将是笔者下一步的研究重点.

参考文献 (References):

- [1] HUANG G, ZHU Q, SIEW C. Extreme learning machine: a new learning scheme of feedforward neural networks[C]//2004 IEEE International Joint Conference on Neural Networks. Budapest, Hungary, 2004: 985-990.
- [2] HUANG G, ZHU Q, SIEW C. Extreme learning machine: theory and applications[J]. *Neuro-*

- computing*, 2006, **70**(1/3): 489-501.
- [3] HUANG G, ZHOU H, DING X, et al. Extreme learning machine for regression and multiclass classification[J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, **42**(2): 513-529.
- [4] 姜琳颖, 余东海, 石鑫. 基于加权极限学习机的肿瘤基因表达谱数据分类[J]. 东北大学学报(自然科学版), 2017, **38**(6): 798-803.(JIANG Linying, YU Donghai, SHI Xin. Tumor microarray gene expression data classification based on weighted extreme learning machine[J]. *Journal of Northeastern University (Natural Science)*, 2017, **38**(6): 798-803.(in Chinese))
- [5] 柯道, 邹嘉伟, 杜明智, 等. 基于蒙特卡罗数据集均衡与鲁棒性增量极限学习机的图像自动标注[J]. 电子学报, 2017, **45**(12): 2925-2935.(KE Xiao, ZOU Jiawei, DU Mingzhi, et al. The automatic image annotation based on monte-carlo data set balance and robustness incremental extreme learning machine[J]. *Acta Electronica Sinica*, 2017, **45**(12): 2925-2935.(in Chinese))
- [6] HAMPEL F, RONCHETTI E, ROUSSEEUW P, et al. *Robust Statistics*[M]. Wiley, 2005.
- [7] FRENAY B, VERLEYSEN M. Classification in the presence of label noise: a survey[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(5): 845-869.
- [8] DENG W, ZHENG Q, CHEN L. Regularized extreme learning machine[C]//*2009 IEEE Symposium on Computational Intelligence and Data Mining*. Nashville, TN, USA, 2009: 389-395.
- [9] HORATA P, CHIEWCHANWATTANA S, SUNAT K. Robust extreme learning machine[J]. *Neurocomputing*, 2013, **102**: 31-44.
- [10] ZHANG K, LUO M. Outlier-robust extreme learning machine for regression problems[J]. *Neurocomputing*, 2015, **151**(3): 1519-1527.
- [11] REN Z, YANG L. Robust extreme learning machines with different loss functions[J]. *Neural Processing Letters*, 2019, **49**(3): 1543-1565.
- [12] CHEN K, LÜ Q, LU Y, et al. Robust regularized extreme learning machine for regression using iteratively reweighted least squares[J]. *Neurocomputing*, 2017, **230**: 345-358.
- [13] WANG L, JIA H, LI J. Training robust support vector machine with smooth ramp loss in the primal space[J]. *Neurocomputing*, 2008, **71**(13/15): 3020-3025.
- [14] SHEN X, NIU L, QI Z, et al. Support vector machine classifier with truncated pinball loss[J]. *Pattern Recognition*, 2017, **68**: 199-210.
- [15] ZHONG P. Training robust support vector regression with smooth non-convex loss function [J]. *Optimization Methods & Software*, 2012, **27**(6): 1039-1058.
- [16] JIANG W, NIE F, HUANG H. Robust dictionary learning with capped l_1 norm[C]//*Twenty-Fourth International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina, 2015.
- [17] LIU W, POKHAREL P, PRINCIPE J. Correntropy: properties and applications in non-Gaussian signal processing[J]. *IEEE Transactions on Signal Processing*, 2007, **55**(11): 5286-5298.
- [18] HE R, HU B, YUAN X, et al. *Robust Recognition Via Information Theoretic Learning*[M]. Springer, 2014.
- [19] XING H, WANG X. Training extreme learning machine via regularized correntropy criterion[J]. *Neural Computing and Applications*, 2013, **23**(7/8): 1977-1986.
- [20] SINGH A, POKHAREL R, PRINCIPE J. The C-loss function for pattern classification[J]. *Pattern Recognition*, 2014, **47**(1): 441-453.

- [21] FENG Y, YANG Y, HUANG X, et al. Robust support vector machines for classification with nonconvex and smooth losses[J]. *Neural Computation*, 2016, **28**(6): 1217-1247.
- [22] YANG L, REN Z, WANG Y, et al. A robust regression framework with Laplace kernel-induced loss[J]. *Neural Computation*, 2017, **29**(11): 3014-3039.

An Exponential Laplace Loss Function Based Robust ELM for Regression Estimation

WANG Kuaini^{1,2}, CAO Jinde², LIU Qingshan²

(1. School of Science, Xi'an Shiyou University, Xi'an 710065, P.R.China;

2. School of Mathematics, Southeast University, Nanjing 211189, P.R.China)

(Contributed by CAO Jinde, M. AMM Editorial Board)

Abstract: Datasets are often contaminated by various noises in many practical applications. The classical extreme learning machine (ELM) shows poor prediction accuracy and large fluctuation of prediction results in dealing with such datasets. To overcome this drawback, an exponential Laplace loss function was proposed, which can weaken the influences of noises. The proposed loss function is based on the Gauss kernel function, and is differentiable, non-convex, bounded and able to approach the Laplace function. Then the proposed loss function was introduced into the ELM to build a robust ELM model for regression estimation. The iterative re-weighted algorithm was employed to solve the resultant optimization problem. In each iteration, the training samples with noises were given smaller weights, which can effectively improve the prediction accuracy. Experiments on real-world datasets show that, the proposed model has better learning performance and robustness.

Key words: neural network; extreme learning machine; robustness; exponential Laplace loss function; iterative re-weighted algorithm

Foundation item: The National Natural Science Foundation of China(61833005;61907033); China Postdoctoral Science Foundation(2018M642129)

引用本文/Cite this paper:

王快妮, 曹进德, 刘庆山. 基于指数 Laplace 损失函数的回归估计鲁棒超限学习机[J]. 应用数学和力学, 2019, **40**(11): 1169-1178.

WANG Kuaini, CAO Jinde, LIU Qingshan. An exponential Laplace loss function based robust ELM for regression estimation[J]. *Applied Mathematics and Mechanics*, 2019, **40**(11): 1169-1178.