arXiv:2301.01458v1 [math.OC] 4 Jan 2023

一种改进的极限学习机混合正则化方法

周良娟 湖南大学数学学院，中国长沙

缪伟*miaow@hnu.edu.cn 湖南大学数学学院，中国长沙

摘要

极限学习机（ELM）是一种可随机初始化第一个隐藏层并能快速计算的网络模型。为提高ELM的分类性能，本文提出了一种 ℓ_2和 $\ell_{0.5}$ 正则化ELM模型（ℓ_2-$\ell_{0.5}$-ELM）。采用压缩映射不动点迭代优化算法求解 ℓ_2-$\ell_{0.5}$-ELM模型。在合理假设下，对所提方法的收敛性和稀疏性进行了讨论和分析。将所提 ℓ_2-$\ell_{0.5}$-ELM方法与反向传播、支持向量机、ELM、$\ell_{0.5}$-ELM、 ℓ_1-ELM、 ℓ_2-ELM和 ℓ_2-ℓ_1ELM进行比较，结果表明， ℓ_2-$\ell_{0.5}$-ELM的预测准确率、稀疏性和稳定性优于其他7种模型。

CCS 概念

• **计算数学** → **凸优化**； • **计算方法** → **正则化**。

关键词

高维数据，稀疏性，混合正则化，降维

ACM 参考格式：

周良娟和缪伟。2022年。改进的极限学习机混合正则化方法。载于 *2022*年第四届国际先进信息科学与系统会议（*AISS 2022 27，2022*年，*257* 10），*11*月，中国三亚。ACM，纽约，纽约州，美国，页码。https://doi.org/ 。1145/3573834.3574501

1 引言

前馈神经网络（FNNs）是最常用的神经网络之一，其数学定义如下：

$$G_N(x_i) = \sum_{i=1}^{N} \beta_i g(\langle \omega_i, x_i \rangle + b_i),$$

其中， $x_i = (x_{i1},\ x_{i2},\ \ldots,\ x_{ip}) \in \mathbb{R}^p$ 表示输入， b_i 表示偏置， g表示激活函数。$\omega_i,\ x_i = \sum_{j=1}^{p} \omega_{ij} x_{ij}$表示欧氏内积， $\omega_i = (\omega_{i1},\ \omega_{i2},\ \ldots,\ \omega_{ip}) \in \mathbb{R}^p$ 表示连接输入与第 i 个隐藏节点的权重， $\beta_i \in \mathbb{R}$ 表示连接第 i个隐藏节点与输出节点的权重。根据传统的前馈神经网络学习算法，网络中的所有参数都需要针对具体任务进行调整。一种经典的学习方法是反向传播（BP）算法，主要通过梯度下降来求解：

$$\min_{\omega_i,\beta_i,b_i} \sum_{i=1}^{n} \|t_i - G_N(x_i)\|_2^2,$$

其中，$(x_i,\ t_i)(i = 1,\ 2,\ \ldots\ldots,\ n)$表示训练样本。然而，与传统的前馈神经网络学习方式不同，Huang[10]提出了一种随机学习器模型，称为极限学习机（ELM）及其相关算法。在ELM模型中， ω_i和b_i是随机指定的，无需训练，因此只需训练 β_i。设$\mathrm{T} = [t_1, t_2, \ldots, t_n]$且

$$\mathrm{H} = \begin{bmatrix} g(\langle \omega_1, x_1 \rangle + b_1) & \ldots & g(\langle \omega_N, x_1 \rangle + b_N) \\ \vdots & \ldots & \vdots \\ g(\langle \omega_1, x_n \rangle + b_1) & \ldots & g(\langle \omega_N, x_n \rangle + b_N) \end{bmatrix}, \quad (1)$$

一旦在 $[-c, c]$中按照均匀分布随机指定输入权重和偏置，隐藏层输出矩阵在训练阶段便保持不变。因此，可以利用最小二乘法求得输出权重：

$$\min_{\beta \in \mathbb{R}^N} \left\{ \|\mathrm{H}\beta - \mathrm{T}\|_2^2 \right\}, \quad (2)$$

模型（2）的解可写为 $\beta = \mathrm{H}^{\dagger}\mathrm{T}$，其中$\mathrm{H}^{\dagger}$是隐藏层输出矩阵H[14]的Moore–Penrose广义逆。

Igelnik[11] ,提出并确立了ELM网络具备一般逼近能力的理论基础，其中随机分配的输入权重和偏置的取值范围与数据相关，并以构造模式进行分配。因此，在算法实现中，应针对不同数据集谨慎估计参数范围。另一方面，考虑到许多高维数据中输出参数 β 的稀疏性，Cao等人[4] 提出了一种基于 ℓ_1 正则化项稀疏性的 ℓ_1 正则化ELM模型，其形式如下：

$$\min_{\beta \in \mathbb{R}^N} \left\{ \frac{1}{2} \|\mathrm{H}\beta - \mathrm{T}\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (3)$$

其中， $\lambda > 0$ 是正则化参数， β 是通过迭代计算得到的输出系数。该模型称为 Lasso 模型，近年来受到许多学者的研究 [15]。

*两位作者对本研究的贡献相同。

允许免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，但不得为盈利或商业利益而制作或分发副本，且副本必须附有此声明以及首页的完整引文。对于本作品中归他人而非 ACM 所有的组件，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布到服务器或分发至列表，则需事先获得特别许可和/或支付费用。请通过 permissions@acm.org 申请许可。*AISS 2022，2022*年*11*月*25*日至*27*日，中国三亚 © 2022年 计算机协会。ACM ISBN 978-1-4503-9793-3/22/11…$15.00 https://doi.org/10.1145/3573834.3574501

针对模型（2），Fan等人 [8] 在 ELM模型中添加了 $\ell_{0.5}$ 正则化项，其生成的解 $\ell_{0.5}$ 比 ℓ_1 正则化项 [16], 更加稀疏，该模型定义如下：

$$\min_{\beta\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathrm{H}\beta-\mathrm{T}\|_2^2+\lambda\|\beta\|_{0.5}\right\}, \tag{4}$$

其中 $\lambda>0$ 为正则化参数，该模型可通过迭代半阈值算法 [16]求解。

模型（2）的另一种正则化模型涉及 ℓ_2正则化项（ℓ_2-极限学习机）[5]:

$$\min_{\beta\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathrm{H}\beta-\mathrm{T}\|_2^2+\mu\|\beta\|_2^2\right\}, \tag{5}$$

其中 μ 为正则化参数，选择参数 μ后，若表达式$\mathrm{H}^T\mathrm{H}+\mu\mathrm{I}$可逆，则模型（5）的解可写为 $\beta=(\mathrm{H}^T\mathrm{H}+\mu\mathrm{I})^{-1}\mathrm{I})^{-1}\mathrm{H}^T\mathrm{T}$。

Hai等人[9] 提出了一种 ℓ_2-ℓ_1-极限学习机混合模型，将 ℓ_1 正则化项的稀疏性与 ℓ_2正则化项的稳定性相结合，如下所示：

$$\min_{\beta\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathrm{H}\beta-\mathrm{T}\|_2^2+\lambda(\gamma\|\beta\|_1+\varepsilon\|\beta\|_2^2)\right\}, \tag{6}$$

其中，$\lambda\geq 0$、$\gamma\geq 0$ 和 $\varepsilon\geq 0$ 为正则化参数。受 ℓ_2-ℓ_1-ELM模型的启发，根据Xu等人[17], 的研究，他们发现 $\ell_p(p\in(0,\ 1))$ 正则化项的解的稀疏性具有以下特点：当 $0<p<0.5$时，ℓ_p的稀疏效果没有显著差异；当 $0.5<p<1$时，p越小，稀疏效果越好。因此，$\ell_{0.5}$ 正则化项可以作为 $\ell_p(p\in(0,\ 1))$的代表元素。由此，我们结合 ℓ_2 正则化项的稳定性和 $\ell_{0.5}$ 的稀疏性（其稀疏性高于 ℓ_1），提出了ℓ_2-$\ell_{0.5}$-ELM模型。新模型描述如下：

$$\min_{\beta\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathrm{H}\beta-\mathrm{T}\|_2^2+\lambda(\gamma\|\beta\|_{0.5}+\varepsilon\|\beta\|_2^2)\right\}, \tag{7}$$

其中，各参数的含义与表达式（6）相同。在分类 [2, 6]中可以找到在优化模型中同时加入 $\ell_{0.5}$ 和 ℓ_2 惩罚项的思路。本研究主要构建了一种迭代算法，并研究了Hai[9]所述随机学习器模型的一些性质。特别地，我们整合了ELM的特征，并提出了一种求解混合模型（7）的迭代策略。本文的主要贡献概括如下：

(i) 由于存在 $\ell_{0.5}$ 范数，整个模型是一个非凸、非光滑且非 Lipschitz 优化问题。我们提出了一种名为 ℓ_2-$\ell_{0.5}$- 极限学习机的全新算法。通过分析具有特定特征的两个凸函数之和的最小化问题，证明了该算法的有效性。

(ii) 推导了收敛性、稀疏性等关键理论性质，以保证所提方法的可行性。

(iii) 我们进行了大量实验，包括从专家系统和智能系统领域收集的一些 UCI 数据集、基因数据集以及 ORL 人脸图像数据集。实验结果表明，所提出的 ℓ_2- $\ell_{0.5}$-ELM 算法性能更优。

本文其余部分安排如下：第 2 节回顾一些基本概念和理论。第 3 节通过不动点方程阐述迭代方法，并提出一种适用于 ℓ_2 - $\ell_{0.5}$-ELM 模型的算法。第 4 节分析有关收敛性和稀疏性的一些理论结果。第 5 节展示在 UCI 数据集、基因数据集和 ORL 人脸图像数据集上的实验结果。第 6 节得出结论。

2 预备知识

本节主要介绍一些基本概念和凸优化定理。首先介绍半阈值函数[16]。$\mathscr{P}(\lambda,t)$：$\mathbb{R}\to\mathbb{R},\lambda>0$，其可写为：

$$\mathscr{P}(\lambda,t)=\begin{cases}\frac{2}{3}t\left(1+\cos\left(\frac{2(\pi-\phi(t))}{3}\right)\right) & |t|>\frac{3}{4}\lambda^{\frac{2}{3}}\\ 0 & |t|\leq\frac{3}{4}\lambda^{\frac{2}{3}}\end{cases}, \tag{8}$$

其中 $\phi(t)=\arccos\left(\frac{\lambda}{8}(\frac{|t|}{3})^{-\frac{3}{2}}\right),\pi=3.14$，相应的半阈值算子 $\mathrm{half}(\lambda,\beta):\mathbb{R}^N\to\overline{\mathbb{R}^N}$按分量作用如下：

$$[\mathrm{half}(\lambda,\beta)]_i=\mathscr{P}(\lambda,\beta_i). \tag{9}$$

接下来，我们介绍半阈值算子 [7, 16]的一个关键特性：

$$\|\mathrm{half}(\lambda,t)-\mathrm{half}(\lambda,t')\|\leq\|t-t'\|. \tag{10}$$

凸优化中的另一个重要概念是近端算子 [12]:

$$\mathrm{prox}_{\varphi}\beta=\arg\min\left\{\ \|u-\beta\|_2^2+\varphi(u)\right\},$$

其中，ϕ是定义在$\mathbb{R}^N$上的实值凸函数。近端算子的一个重要性质见命题 1[7], ，我们将利用该性质证明主要结果。

命题 1. 设 φ是定义在 $\mathbb{R}^N$上的实值凸函数。假设 $\psi(\cdot)=\varphi+\frac{\rho}{2}\|\cdot\ \|_2^2+\langle\cdot\ ,u\rangle+\sigma$，其中 $u\in\mathbb{R}^N,\rho\in[0,\infty),\sigma\in\mathbb{R}$，则

$$\mathrm{prox}_{\psi}\beta=\mathrm{prox}_{\varphi/(1+\rho)}((\beta-u)/(1+\rho)). \tag{11}$$

3 解：模型的不动点迭代算法

对于极限学习机，输出矩阵 H 是一个从$\mathbb{R}^N$到 $\mathbb{R}^m$ 的有界线性算子，这是因为激活函数 $g(\cdot)\in(0,1)$有界。为了进一步提高准确率和稀疏性，我们采用正则化模型（7）来估计网络的输出权重。简洁地定义如下：

$$p_{\gamma,\varepsilon}=\gamma\|\beta\|_{0.5}+\varepsilon\|\beta\|_2^2,$$

其中 ε、$\gamma\geq 0$、$p_{\gamma,\ \varepsilon}:\ \mathbb{R}^N\to[0,\infty)$。然后，模型（7）可以重新定义为

$$\min_{\beta\in\mathrm{R}^N}\left\{\frac{1}{2}\|\mathrm{H}\beta-\mathrm{T}\|_2^2+\lambda p_{\gamma,\varepsilon}\right\}. \tag{12}$$

此外，我们提出以下引理和定理，以用于求解我们的模型：

引理 1. 对于所有 $\lambda > 0$ 和 $\beta \in \mathbb{R}^N$，半阈值算子(8) 可表示为：

$$\mathrm{half}(\lambda, \beta) = \arg\min_u \left\{ \frac{1}{2}\|u - \beta\|_2^2 + \lambda \|u\|_{0.5} \right\}.$$

引理 2. 对于所有 $\lambda > 0$、$\gamma \geq 0$、$\varepsilon \geq 0$ 和 $\beta \in \mathbb{R}^N$，二分之一($\frac{\lambda\gamma}{1+2\varepsilon\lambda}$ $\frac{\beta}{1+2\varepsilon\lambda}$) 是 $\lambda p_{\gamma,\ \varepsilon}(\beta)$的近端算子。

定理 1. 设 $\lambda > 0$、$\gamma \geq 0$、$\varepsilon \geq 0$ 和 $\delta \in (0,\ \infty)$。那么，β 是函数 (12) 的一个最小化解，当且仅当它满足以下不动点方程：

$$\beta = \mathrm{half}\left(\frac{\delta\lambda\gamma}{1 + 2\varepsilon\lambda\delta}, \frac{(\mathrm{I} - \delta \mathrm{H}^T \mathrm{H})\beta - \delta \mathrm{H}^T \mathrm{T}}{1 + 2\varepsilon\lambda\delta} \right), \tag{13}$$

其中，单位算子 $\mathrm{I}: \mathbb{R}^N \to \mathbb{R}^N$，H 的定义见 (1)，而 H^T 表示 H的伴随算子。

此外，根据近端算子的性质，我们可以得出关于压缩映射的利普希茨常数的精确结论及相应定理，如下所示。

定理 2. 设 $\lambda > 0$、$\gamma \geq 0$、$\varepsilon \geq 0$ 和 $\delta \in (0,\ \infty)$。假设存在两个正常数 κ_0 和 κ，使得式 (1) 所示隐藏层的输出矩阵 H 的范数有界，即 $\kappa_0 \leq \|\mathrm{H}^T\mathrm{H}\|_2 \leq \kappa$，那么 β 是式 (12) 的极小值，当且仅当它是利普希茨映射 $\Gamma: \mathbb{R}^N \to \mathbb{R}^N$的不动点，即 $\beta = \Gamma\beta$，其中

$$\Gamma\beta = \mathrm{half}\left(\frac{\delta\lambda\gamma}{1 + 2\varepsilon\lambda\delta}, \frac{(\mathrm{I} - \delta \mathrm{H}^T \mathrm{H})\beta + \delta \mathrm{H}^T \mathrm{T}}{1 + 2\varepsilon\lambda\delta} \right). \tag{14}$$

选取 $\delta = \frac{2}{\kappa_0+\kappa}$ $q = 1 - \frac{2\kappa_0}{\kappa + \kappa_0} \leq$，利普希茨常数有限。

1 $\kappa_0 > 0$ Γ

特别地，如果，我们可以得到 是一个压缩映射。

定理 1 和定理 2 表明，ℓ_2-$\ell_{0.5}$-极限学习机问题可以描述为一个不动点算法。此外，下一个定理将介绍 ℓ_2-$\ell_{0.5}$-极限学习机的迭代过程。

定理 3. 假设 κ_0 和 κ 为正数，且由它们可知，式(1) 所示隐藏层的输出矩阵H 的范数有限，即 $\kappa_0 \leq \|\mathrm{H}^T\mathrm{H}\|_2 \leq \kappa$，并且序列$\{\beta\}_{l=0}^{\infty} \subseteq \mathrm{R}^N$ 按如下方式迭代生成

$$\beta_l = \mathrm{half}\left(\frac{\delta\lambda\gamma}{1 + 2\varepsilon\lambda\delta}, \frac{(\mathrm{I} - \delta \mathrm{H}^N \mathrm{H})\beta_{l-1} - \delta \mathrm{H}^T \mathrm{T}}{1 + 2\varepsilon\lambda\delta} \right), \tag{15}$$

其中 l = 1、2、3、.....、$\lambda > 0$、$\varepsilon > 0$、$\gamma \geq 0$ 和 $\delta = \frac{2}{\kappa+\kappa_0}$。因此，无论如何选择 β_0，$\{\beta_l\}_{l=0}^{\infty}$都强收敛于模型(10) 的极小值点。

注 1. 从定理 *3* 的证明中不难得出。

$$\|\beta_l - \beta^*\|_2 \leq \frac{\kappa + \kappa_0}{\kappa_0(\kappa + \kappa_0 + 4\varepsilon\lambda)} \left(\frac{\kappa - \kappa_0}{\kappa + \kappa_0} \right)^l \|\mathrm{H}^T\mathrm{T}\|_2.$$

因此，对于每个 $\xi > 0$，如果

$$\frac{\kappa + \kappa_0}{\kappa_0(\kappa + \kappa_0 + 4\varepsilon\lambda)} \left(\frac{\kappa - \kappa_0}{\kappa + \kappa_0} \right)^l \|\beta_1 - \beta_0\|_2 < \xi.$$

即，

$$l > \frac{\log\left(\frac{\|\beta_1 - \beta_0\|_2 (\kappa + \kappa_0)}{\xi \kappa_0 (\kappa + \kappa_0 + 4\varepsilon\lambda)} \right)}{\log\left(\frac{\kappa + \kappa_0}{\kappa - \kappa_0} \right)},$$

thus

$$\|\beta_l - \beta^*\|_2 < \xi.$$

总之，完整的 ℓ_2-$\ell_{0.5}$-极限学习机算法见算法 1，该算法整合了定理 3 和注 1 的结果。下一节将介绍所提出算法的一些性质。

算法 1： ℓ_2-$\ell_{0.5}$-ELM模型的算法

输入：给定一组训练样本 ℓ = $\{ x_j\ t_j \quad x_j \in \mathbb{R}^p\ t_j \in \mathbb{R}^m\ j = 1\ 2\ \ldots\ n \}$

(,) : , , , , , , 激活函数 g、隐藏节点数 N、相关正则化参数 $\lambda > 0, \gamma \geq 0, \varepsilon \geq 0$、相应参数 δ 以及可接受误差 ξ。步骤 1：为输入权重 ω_i和偏置 b_i随机指定一个适当的范围（i = 1, 2, . . . , N ）。步骤 2：计算隐藏层输出矩阵 H；步骤 3：设置 β_0 =(0, 0, . . . , 0), β_1 =half ($\frac{\delta\lambda\gamma}{1+2\varepsilon\lambda\delta}$, (I − δ $\mathrm{H}^T\mathrm{H})\beta_0 + \delta\mathrm{H}^T\mathrm{T}\ 1 + 2\varepsilon\lambda\delta)$，并令 l_{max}为大于 $\log\left(\frac{\|\beta_1 - \beta_0\|_2 (\kappa + \kappa_0)\ (\kappa + \kappa_0 + 4\varepsilon\lambda)}{\xi\kappa_0} \right) \Big/ \log\left(\frac{\kappa+\kappa_0}{\kappa-\kappa_0} \right)$ 的最小正整数。步骤 4：对于 $l = 1$： l_{max}如果 $\overline{l \geq l_{max}}$，则停止；否则 l： = $l + 1$ 并按如下方式更新 β： β_{l+1} =half ($\frac{\delta\lambda\gamma}{1+2\varepsilon\lambda\delta}$, (I − $\delta\mathrm{H}^T\mathrm{H})\beta_l + \delta\mathrm{H}^T\mathrm{T}\ 1 + 2\varepsilon\lambda\delta)$。重复步骤 4，直到得到所需的输出权重 $\hat{\beta} = \beta_{max}$。输出：返回输出权重 β;

4 ℓ_2-$\ell_{0.5}$-极限学习机的一些特性

在本节中，我们将讨论并分析有关 ℓ_2-$\ell_{0.5}$-极限学习机的估计量的一些关键特性，例如收敛性和稀疏性。

定理 4. β_l 强收敛到该最小化问题的最小值 β^*

$$\min_{\beta \in \mathrm{R}^N} \left\{ \frac{1}{2}\|\mathrm{H}\beta - \mathrm{T}\|_2^2 + \lambda p_{\gamma\varepsilon}(\beta) \right\}$$

当 $l \to \infty$时。

$\beta_{0.5}$ 在 ℓ_2-$\ell_{0.5}$-极限学习机中，是解的稀疏性的重要组成部分。因此，我们将定理 5 表述如下。

定理 5. 设 $\lambda > 0$、$\gamma > 0$，则对于任意 $\beta \in \mathbb{R}^N$，半($\frac{\lambda\gamma}{1+2\varepsilon\lambda}$, $\frac{\beta}{1+2\varepsilon\lambda}$)的支撑集都是有限的。特别地，β^* 和 β_l均具有有限支撑。

如果将正则化参数 λ and γ 固定为某些常数值，那么 β^* 和 β_l 只有少数有限的非零系数，因此（12）的解是稀疏的。

表1：6个数据集的详情

数据集	Type	样本	特征	类别
奥地利	UCI	690	14	2
电离层	UCI	151	34	2
平衡	UCL	625	4	3
结肠	gene	62	2000	2
弥漫性大B细胞淋巴瘤	gene	77	7129	2
ORL	图像	400	10304	40

5 性能评估

在新章节中，我们进行了一系列实验，使用了若干 UCI基准数据集[9] 和基因数据，以展示所提出的 ℓ_2-$\ell_{0.5}$-极限学习机方法的性能。所有实验均在 Mac Pycharm环境中进行，所用设备配备四核Intel Core i5 中央处理器（8 GB 2133 MHz LPDDR3），处理器速度为 1.40GHz。实验所用网络的激活函数为 S形函数 $g(x) = 1/(1+e^{-x})$。

ℓ_2-$\ell_{0.5}$-ELM模型与其他七种模型进行比较：反向传播、支持向量机、ELM、 ℓ_2-ℓ_1-ELM、 ℓ_2-ELM、 ℓ_1-ELM、 $\ell_{0.5}$-ELM。反向传播模型仅包含一个隐藏层和一个输出层，所有参数均通过反向传播算法训练； ℓ_1-ELM和 $\ell_{0.5}$-ELM分别是 ℓ_2-ℓ_1-ELM和 ℓ_2-$\ell_{0.5}$-ELM的简化形式。激活函数定义为： $g(x) = 1/(1+e^{-x})$。

为了检验 ℓ_2-$\ell_{0.5}$-ELM模型的算法，选取了来自UCI机器学习库的三个真实分类数据集。各数据集的基本信息如表1所示。最终结果取30次实验验证的平均值。对于这些数据集，样本量固定，但每个样本会被随机分配为训练数据或测试数据。

5.1 UCI数据集上的性能

为验证所提出的 ℓ_2-$\ell_{0.5}$-ELM模型的性能，实验采用了三类真实分类数据集，包括UCI[3], 基因表达数据集和ORL人脸数据集。UCI机器学习库（2013UCI）包含三个数据集：奥地利信用审批（奥地利）、电离层和平衡尺度（平衡）。基因表达数据集包含结肠[1] 和弥漫性大B细胞淋巴瘤[13], , 二者均为二元数据集。此外，ORL人脸数据集包含400幅图像，分为40类。每类包含10幅具有不同面部细节的图像，每幅图像的尺寸为 112×92。所有数据集的详细信息汇总于表1。此外，这些数据来自不同的应用领域，希望利用这些来自不同背景的数据，从多个角度分析 ℓ_2-$\ell_{0.5}$-ELM模型。

为减少随机误差，我们重复进行30次试验，并取平均值作为最终结果。正则化参数用于控制误差与惩罚项之间的权衡。对于奥地利数据集，采用参数（ ℓ_2-$\ell_{0.5}$-极限学习机， ℓ_2-ℓ_1-极限学习机： $\lambda = 0.8$， $\gamma = 0.1$， $\varepsilon = 0.9$）；对于电离层数据集，采用（ ℓ_2-$\ell_{0.5}$-极限学习机， ℓ_2-ℓ_1-极限学习机： $\lambda = 0.9$， $\gamma = 0.05$， $\varepsilon = 0.9$）；对于平衡尺度数据集，（ ℓ_2-$\ell_{0.5}$-极限学习机： $\lambda = 0.8$， $\gamma = 1$， $\varepsilon = 1$，对于 ℓ_2-ℓ_1-极限学习机： $\lambda = 0.005$， $\gamma = 0.5$， $\varepsilon = 0.5$），我们将可接受误差设为 $\xi =$ 0.0001，0.001，0.0001。实验中的隐藏节点数为600。表2展示了八种模型在各数据集上的运行时间、保留的节点数以及测试准确率（标准差保留4位有效数字，表格中的0.00表示标准差小于10^{-4}）。这些指标用于衡量所提方法的稀疏性、稳定性和有效性。相应的测试结果图如下所示。

表2：8种模型在3个不同数据集上的性能比较

数据集	方法	时间(s)	剩余节点	准确率(% ±%)
Austrain	BP	2.1751	600	72.58 ± 13.57
	SVM	**0.0448**	—	79.14 ± 1.98
	ELM	0.0588	600	65.37 ± 3.08
	$\ell_{0.5}$-极限学习机	5.8542	48.5	82.76 ± 0.00
	ℓ_1-极限学习机	8.1648	118.5	81.38 ± 0.00
	ℓ_2-极限学习机	8.2735	600	80.36 ± 0.00
	ℓ_2-ℓ_1-极限学习机	10.041	492.5	81.38 ± 0.00
	ℓ_2-$\ell_{0.5}$-极限学习机	7.5875	118.5	**82.76 ± 0.00**
电离层	BP	2.1751	600	72.58 ± 13.57
	SVM	0.0108	–	86.51 ± 2.09
	ELM	**0.0003**	600	91.55 ± 2.78
	$\ell_{0.5}$-极限学习机	0.0487	29.5	96.96 ± 0.00
	ℓ_1-极限学习机	5.4755	115.9	97.24 ± 1.06
	ℓ_2-极限学习机	0.0520	600	96.05 ± 1.57
	ℓ_2-ℓ_1-极限学习机	4.4093	437.5	96.84 ± 0.98
	ℓ_2-$\ell_{0.5}$-极限学习机	0.0569	193	**98.01 ± 0.00**
平衡	BP	4.3814	600	59.99 ± 25.26
	SVM	0.0215	–	88.63 ± 1.86
	EL,M	**0.0008**	600	50.72 ± 6.66
	$\ell_{0.5}$-极限学习机	0.1285	23.3	90.55 ± 0.00
	ℓ_1-极限学习机	6.5074	42.9	90.47 ± 1.66
	ℓ_2-极限学习机	0.1579	600	90.55 ± 0.00
	ℓ_2-ℓ_1-极限学习机	6.8678	246.4	90.10 ± 1.35
	ℓ_2-$\ell_{0.5}$-极限学习机	0.0974	52.7	**90.91 ± 0.00**

从结果 1-3 可以看出，ELM模型的准确率低于所有正则化 ELM 模型。ELM模型的标准差高于其他正则化 ELM 模型，这表明其稳定性较低。 ℓ_2-$\ell_{0.5}$-ELM模型在所有节点上的准确率可与其他正则化 ELM 模型进行比较，且在大多数隐藏节点上的准确率高于其他可比的正则化 ELM 模型。这表明 ℓ_2-$\ell_{0.5}$-ELM模型始终具有良好的分类预测能力。从不同节点的标准差来看， ℓ_2-$\ell_{0.5}$-ELM模型低于其他对比模型，说明该方法的分类准确率更加稳定。

我们可以详细了解 ℓ_2-$\ell_{0.5}$-ELM的性能，并得出以下结论：

(i) 在 3 个数据集中，正则化 ELM 方法（ℓ_2-$\ell_{0.5}$-ELM、 $\ell_{0.5}$-ELM、 ℓ_2-ℓ_1-ELM、 ℓ_1-ELM、 ℓ_2-ELM）的分类准确率显著高于反向传播、支持向量机和 ELM 方法，表明正则化 ELM 方法具有更好的泛化性能；并且 ℓ_2-$\ell_{0.5}$-ELM方法的分类准确率高于其他对比的正则化 ELM 方法。

(ii) 从剩余隐藏节点数来看， $\ell_{0.5}$-ELM的隐藏节点数最少。这表明

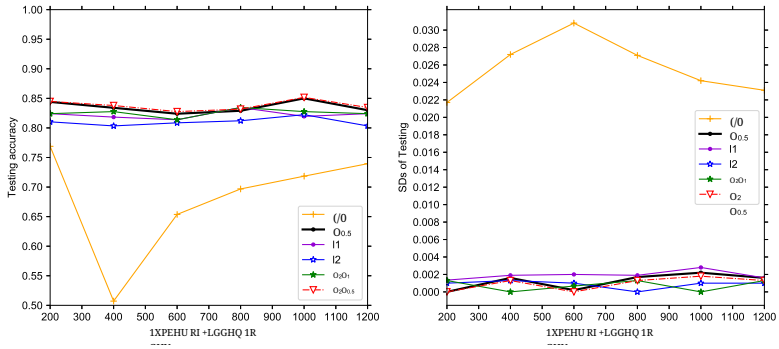
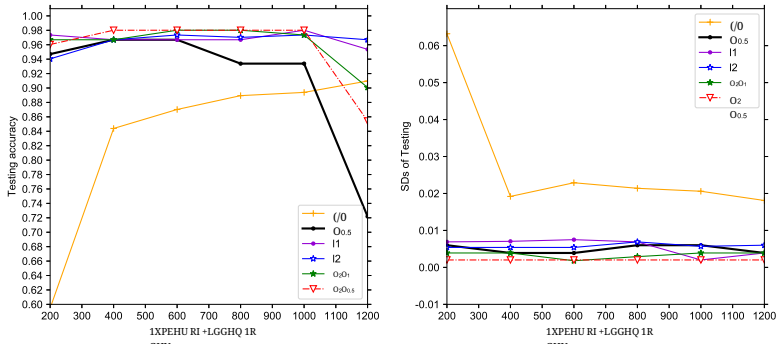


图1：奥地利数据集上6个模型的性能比较



性能比　　较6个模型在离子层　　数据集上的

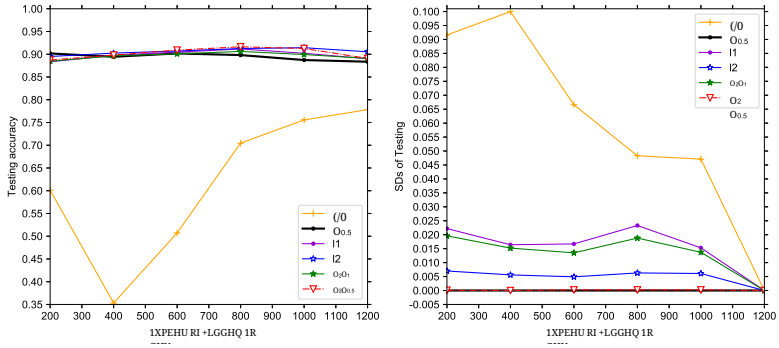


图 3：平衡数据集中6个模型的性能比较

$\ell_{0.5}$ 或 ℓ_1-正则化项有助于提高模型隐藏节点的稀疏性。与 ℓ_2-ℓ_1-ELM模型相比， ℓ_2-$\ell_{0.5}$-ELM模型在模型中加入了 $\ell_{0.5}$ 正则化项，其解更稀疏，因此具有更好的泛化能力。

(iii) 从算法运行时间来看，ELM模型运行时间最短（ELM模型可以直接获得解析解，无需迭代计算）。相比之下，SVM模型的运行速度快于所有正则化ELM方法。其次，对于5种正则化ELM模型，采用 $\ell_{0.5}$ 正则化项的模型（$\ell_{0.5}$-ELM、 ℓ_2-$\ell_{0.5}$-ELM）比采用 ℓ_1 正则化项的模型（ℓ_1-ELM、 ℓ_2- ℓ_1-ELM）运行得更快。

5.2 基因数据集的性能

本节使用结肠数据和DLBCL 数据验证 ℓ_2-$\ell_{0.5}$-ELM模型的性能。每个数据集的训练集和测试集按1：1的比例划分。正则化参数设置如下：结肠数据：(ℓ_2-$\ell_{0.5}$-ELM和ℓ_2-ℓ_1-ELM： $\lambda = 0.09$, $\gamma = 0.9, \varepsilon = 0.9$)，DLBCL 数据：($\ell_2$-$\ell_{0.5}$-ELM和 ℓ_2-ℓ_1-ELM：$\lambda = 0.005, \gamma = 0.5, \varepsilon = 0.5$)；以及 $\xi = 0.001$。每个数据集重复运行30次，并取平均值作为最终结果。如表3所示。

可以证明，单层反向传播网络的预测准确率很低，无法很好地捕捉

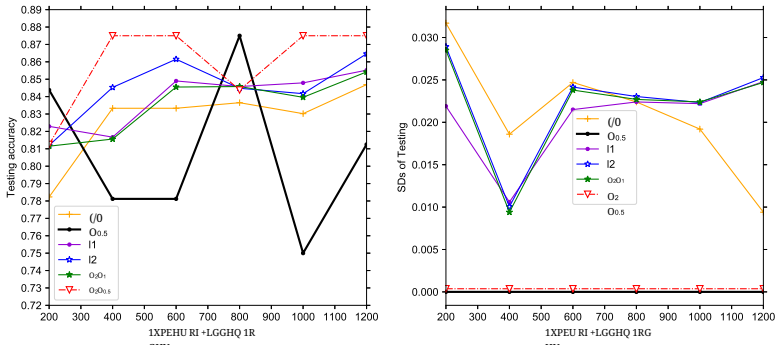


图4：结肠数据集中6 个模型的性能比较

表3：2 个基因数据集上8 个模型的性能比较

数据集	方法	时间（秒）	剩余节点	准确率(% ±%)
结肠	BP	22.2641	1000.0	55.52 ± 9.15
	SVM	0.0358	–	77.5 ± 7.28
	ELM	0.0056	1000.0	83.02 ± 1.92
	$\ell_{0.5}$-极限学习机	0.0829	370.5	75.00 ± 0.00
	ℓ_1-极限学习机	0.0488	974.5	84.79 ± 2.22
	ℓ_2-极限学习机	0.0815	1000.0	84.17 ± 2.20
	ℓ_2-ℓ_1-极限学习机	0.0401	1000.0	83.96 ± 2.24
	ℓ_2-$\ell_{0.5}$-极限学习机	0.0879	877.0	**87.50 ± 0.00**
弥漫性大B细胞淋巴瘤	BP	122.3174	1000.0	57.24 ± 12.55
	SVM	0.0968	–	87.24 ± 5.98
	ELM	0.0060	786.0	89.90 ± 5.98
	$\ell_{0.5}$-极限学习机	5.2214	242.0	**91.43 ± 0.00**
	ℓ_1-极限学习机	18.2957	188.5	89.05 ± 5.12
	ℓ_2-极限学习机	5.2324	764.0	89.51 ± 5.48
	ℓ_2-ℓ_1-ELM	15.5286	431.5	89.62 ± 6.10
	ℓ_2-$\ell_{0.5}$-ELM	5.4519	575.5	**91.43 ± 0.00**

数据的特征。还可以发现， ℓ_2-$\ell_{0.5}$-ELM模型的预测准确率略高于其他方法。采用 $\ell_{0.5}$ 正则化的ELM方法，其准确率标准差远小于反向传播、支持向量机和ELM，表明带有$\ell_{0.5}$ 正则化项的ELM模型变体能够提高解的稳定性；

$\ell_{0.5}$-ELM和 ℓ_1-ELM模型的隐藏节点数较少，也就是说，这两种正则化项的稀疏性最强，这表明在ELM模型中加入 $\ell_{0.5}$ 或 ℓ_1 正则化项可以增强模型的稀疏性，而 ℓ_2-ELM模型的隐藏节点数为1000。 ℓ_2-ELM模型的节点数为1000，这表明ℓ_2正则化项对模型没有稀疏作用。采用 ℓ_2范数，通过惩罚过大的正则化参数来提高模型的稳定性。这使 ℓ_2-$\ell_{0.5}$-ELM更加稀疏且模型更加稳定，从而获得更好的泛化能力。

从算法运行时间的角度来看，可以发现 ELM模型的运行时间最短（ELM模型可以直接获得解析解，无需迭代求解）。相比之下，SVM模型的运行速度快于所有带正则化的ELM方法。

此外，我们使用结肠数据验证不同隐藏节点数（200，400、600，800，1000、1200）对ELM相关模型稳定性的影响。对于每个隐藏节点数，我们进行了30次实验，并计算了ELM、 ℓ_2-$\ell_{0.5}$-ELM、 $\ell_{0.5}$-ELM、ℓ_2-ℓ_1-ELM、 ℓ_1-ELM、 ℓ_2-ELM的测试集准确率和标准差，如图4所示。 ℓ_2-$\ell_{0.5}$-ELM在所有节点上的测试准确率均可与所有正则化ELM模型相媲美，且在大多数隐藏节点数下，其准确率高于其他模型。 ℓ_2-$\ell_{0.5}$-ELM模型的标准差低于其他正则化ELM模型。

5.3 ORLface 数据集上的性能

采用 ORLface 数据集进行实验验证。实验中隐藏节点的数量为 1000。以 30 次实验的平均值作为最终结果。由于原始图像的维度较高，我们使用（2D）^{2}主成分分析[18] 降维技术对每幅图像进行预处理。训练集和测试集的比例为 7∶3。实验中设置的正则化参数值如下： $\ell_{0.5}$-ELM 和 ℓ_1-ELM（$\gamma = 0.05$，$\varepsilon = 0$）， ℓ_2-ELM（$\gamma = 0$， $\varepsilon = 0.5$）， ℓ_2 -ℓ_1-ELM，ℓ_2-$\ell_{0.5}$-ELM（$\gamma = 0.05$， $\varepsilon = 0.5$）；所有实验均选用 $\lambda = 0.001$ 和 $\varepsilon = 0.0001$ 。本实验从准确率和标准差两方面验证模型的性能。结果如表 4 所示。从表中可以看出，

表4：ORL人脸数据集中8 个模型的性能比较

方法	准确率(%)
BP	31.00 ± 4.90
SVM	71.53 ± 2.12
ELM	70.58 ± 2.95
$\ell_{0.5}$-极限学习机	71.00 ± 2.34
ℓ_1-极限学习机	70.85 ± 2.86
ℓ_2-极限学习机	71.17 ± 2.47
ℓ_2-ℓ_1-极限学习机	70.58 ± 2.87
ℓ_2-$\ell_{0.5}$-极限学习机	**71.67 ± 2.34**

表5：ORL人脸数据集中的6 个模型性能比较

节点	ELM	$\ell_{0.5}$-极限学习机	ℓ_1-极限学习机	ℓ_2-极限学习机	ℓ_2-ℓ_1-极限学习机	ℓ_2-$\ell_{0.5}$-极限学习机
500	52.92±3.04	66.10±2.55	60.00 ±1.77	62.63± 2.38	59.25 ±2.32	**65.83 ± 2.46**
1500	76.08±0.73	77.00±0.93	76.33 ±0.67	76.75± 0.75	76.33 ±0.76	**77.20 ±0.93**
2000	78.25±2.00	78.73±2.45	78.33 ±2.08	78.63± 2.18	78.33 ±2.08	**78.83 ±2.45**
2500	79.58±3.49	79.74±3.36	79.67 ±3.44	79.21±3.29	79.63 ±3.44	**79.76 ± 3.26**
3000	81.50±1.98	81.55±2.69	81.42 ±2.07	81.45±2.39	81.42 ±2.07	**81.58 ± 2.68**
3500	81.17±1.81	81.13±2.22	81.17 ±1.87	81.17±1.89	81.17 ±1.87	**81.25 ± 2.12**
4000	82.00±1.81	82.00±1.67	81.92 ±1.74	81.96±1.64	81.92 ±1.74	**82.08 ± 1.65**
mean	75.22±9.12	77.16±5.33	76.21 ±7.00	76.62±6.21	76.08 ±7.24	**77.26 ± 5.32**

ℓ_2-$\ell_{0.5}$-ELM模型的准确率（略高于SVM模型）略高于所有其他受测模型。

此外，我们验证了不同隐藏节点数对预测准确率的影响。实验中选取的隐藏节点数为500、1000、1500、2000、2500、3000、3500和4000。结果如表5所示， ℓ_2-$\ell_{0.5}$-ELM模型的测试准确率高于其他对比ELM模型。随着隐藏节点数变化，ELM模型的测试准确率波动最大，即不同节点的选择对其影响最大，这表明ELM模型在高维数据上的稳定性较差。相比之下，所有正则化 ELM 方法的标准差（5.33，7.00，6.21，7.24，5.32）均低于ELM方法，说明添加正则化项提高了ELM模型的稳定性。ELM方法表明，与其他5种ELM方法相比，所提方法的稳定性更好。

6 结论

为进一步提高ELM模型的稳定性和泛化能力，本文结合 $\ell_{0.5}$ 和 ℓ_2 正则化项，提出了一种 ℓ_2-$\ell_{0.5}$-ELM模型。采用迭代算法和不动点算法求解该模型，并证明了该算法的收敛性和稀疏性。此外，将所提出的 ℓ_2-$\ell_{0.5}$-ELM模型与反向传播、支持向量机、ELM、$\ell_{0.5}$-ELM、 ℓ_1-ELM、 ℓ_2-ELM和 ℓ_2-ELM模型进行了比较。在多个数据集（UCI数据集、基因数据集、ORL人脸数据集）上的实验比较表明，在这些数据集上， ℓ_2-$\ell_{0.5}$-ELM方法在预测准确率和稳定性方面优于其他7种模型。因此，该模型还可作如下改进：在计算不同隐藏节点时，不使用先前已计算节点的信息；同时，可以从增量学习的角度进行学习，从而在一定程度上减少计算时间。

参考文献

[1] U. Alon、N. Barkai、D. A. Notterman、K. Gish、S. Ybarra、D. Mack 和 A. J. Levine。1999。通过对寡核苷酸芯片检测的肿瘤和正常结肠组织进行聚类分析，揭示基因表达的广泛模式。96, 12 (1999), 6745– 6750.[2] Hai Hui Huang A B and Yong Liang C. 2018. 用于 Cox比例风险模型中基因选择的混合 $L_{1/2+2}$ 方法。Computer Methods and Programs in Biomedicine 164 (2018), 65–73. [3] K. Bache 和 M. Lichman。2013。UCI机器学习库。(2013)。[4] Feilong Cao、Yuanpeng Tan 和 Miaomiao Cai。2014。随机权重网络的稀疏算法及其应用。Expert Systems with Applications 41, 5 2014 2457 2462 ( ), – . [5] Feilong Cao、Dianhui Wang、Houying Zhu 和 Yuguang Wang。2016。一种用于随机权重前馈神经网络的迭代学习算法。Information Sciences 328 (2016), 546–557. [6] G. C. Cawley 和 Nlc Talbot。2006。采用贝叶斯正则化的稀疏逻辑回归进行癌症分类中的基因选择。Oxford University Press (2006)。[7] Patrick L. Combettes 和 Valérie R. Wajs。2005。通过近端前向-后向分裂进行信号恢复。Multiscale Model. Simul. 4 (2005), 1168–1200. [8] Qinwei Fan、Lei Niu 和 Qian Kang。2020。通过平滑组 $L_1/2$ 正则项，使用稀疏极限学习机进行回归和多类分类。IEEE Access 8 (2020), 191482–191494. [9] Hailiang, Ye, Feilong, Cao, Dianhui, and Wang. [n. d.]. 一种用于随机向量函数连接网络的混合正则化方法 - ScienceDirect。Expert Systems with Applications 140 ([n. d.]), 112912–112912. [10] Guang-Bin Huang、Qin-Yu Zhu 和 Chee-Kheong Siew。2006。极限学习机：理论与应用。Neurocomputing 70, 1 (2006), 489–501. [11] B. Igelnik 和 Y. H. Pao。1995。自适应函数逼近中基函数的随机选择与函数连接网络。IEEE Trans Neural Netw 6, 6 (1995), 1320–1329. [12] Charles Micchelli、Lixin Shen 和 Yuesheng Xu。2011。图像模型的邻近算法：去噪。Inverse Problems 27 (03 2011), 045009. [13] A. Rosenwald、G. Wright、W. C. Chan、J. M. Connors、E. Campo、R. I. Fisher、R. D. Gascoyne、H. K. Muller-Hermelink、E. B. Smeland 和 J. M. Giltnane。2011。利用分子谱分析预测弥漫性大B细胞淋巴瘤患者化疗后的生存情况。New England Journal of Medicine 346, 25 (2011), 1937. [14] W.F. Schmidt、M.A. Kraaijveld 和 R.P.W. Duin。1992。具有随机权重的前馈神经网络。(1992), 1–4. [15] Robert Tibshirani。2011。通过套索回归进行回归收缩与变量选择：回顾。Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 3 (2011), 267–288. [16] Zongben Xu、Xiangyu Chang、Fengmin Xu 和 Hai Zhang。2012。$L_1/2$ 正则化：阈值表示理论与快速求解器。IEEE Transactions on Neural Networks and Learning Systems 23, 7 (2012), 1013–1027. https: // doi.org/ 10 1109 2012 2197412 . /TNNLS. . Zong-Ben Xu、Hai-Liang Guo、Yao Wang 和 Hai Zhang。2012。 $L_q(0 < q \leq 1)$ 正则化中 $L_1/2$ 正则化的代表性：基于相图的实验研究。Acta Automatica Sinica 38, 7 (2012), 1225–1228. [18] Daoqiang Zhang 和 Zhi-Hua Zhou。2005。(2D)2PCA：用于高效人脸表示与识别的双向二维主成分分析。Neurocom- puting 69, 1 (2005), 224–231.
[17]