

极限学习机前沿进展与趋势

徐 睿 梁 循 齐金山 李志宇 张树森

(中国人民大学信息学院 北京 100872)

摘 要 极限学习机(Extreme Learning Machine, ELM)作为前馈神经网络学习中一种全新的训练框架,在行为识别、情感识别和故障诊断等方面被广泛应用,引起了各个领域的高度关注和深入研究. ELM 最初是针对单隐层前馈神经网络的学习速度而提出的,之后又被众多学者扩展到多隐层前馈神经网络中. 该算法的核心思想是随机选取网络的输入权值和隐层偏置,在训练过程中保持不变,仅需要优化隐层神经元个数. 网络的输出权值则是通过最小化平方损失函数,来求解 Moore-Penrose 广义逆运算得到最小范数最小二乘解. 相比于其它传统的基于梯度的前馈神经网络学习算法,ELM 具有实现简单,学习速度极快和人为干预较少等显著优势,已成为当前人工智能领域最热门的研究方向之一. ELM 的学习理论表明,当隐层神经元的学习参数独立于训练样本随机生成,只要前馈神经网络的激活函数是非线性分段连续的,就可以逼近任意连续目标函数或分类任务中的任何复杂决策边界. 近年来,随机神经元也逐步在越来越多的深度学习中使用,而 ELM 可以为其提供使用的理论基础. 本文首先概述了 ELM 的发展历程,接着详细阐述了 ELM 的工作原理. 然后对 ELM 理论和应用的最新研究进展进行了归纳总结,着重讨论并分析了自 ELM 提出以来的主要学习算法和模型,包括提出的原因、核心思想、求解方法、各自的优缺点以及相关问题. 最后,针对当前的研究现状,指出了 ELM 存在的争议、问题和挑战,并对未来的研究方向和发展趋势进行了展望.

关键词 极限学习机;网络结构;正则化;核学习;深度学习;在线学习;并行计算

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01640

Advances and Trends in Extreme Learning Machine

XU Rui LIANG Xun QI Jin-Shan LI Zhi-Yu ZHANG Shu-Sen

(School of Information, Renmin University of China, Beijing 100872)

Abstract Extreme Learning Machine (ELM) as a new single hidden layer feedforward neural network (SLFN) learning framework has obtained extensive attention and in-depth research in various domains. It has been widely used in many applications, such as action recognition, emotion recognition, fault diagnosis, and so on. ELM was originally proposed for “generalized” single hidden layer feedforward neural networks to overcome the challenging issues faced by back-propagation (BP) learning algorithm and its variants. Recent studies show that ELM can be extended to “generalized” multilayer feedforward neural networks in which a hidden node could be a subnetwork of nodes or a combination of other hidden nodes. ELM provides an efficient and unified learning framework for regression, classification, feature learning, and clustering. The learning theories of ELM show that when learning parameters of hidden layer nodes are generated independently of training samples, as long as the activation function of feedforward neural network is non-linear and continuous, it can approach any continuous objective function or any

收稿日期:2018-06-19;在线出版日期:2019-03-01. 本课题得到国家自然科学基金项目(71531012,61762073,71601013)、国家社会科学基金(18ZDA309)、北京市自然科学基金(4172032,4174087)、北大方正集团有限公司数字出版技术国家重点实验室开放课题资助. 徐 睿, 博士研究生,主要研究方向为图像处理、机器学习. E-mail: xurui1064@ruc.edu.cn. 梁 循(通信作者),教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据挖掘、神经网络、社会计算. E-mail: xliang@ruc.edu.cn. 齐金山,博士研究生,主要研究方向为社会计算、数据挖掘. 李志宇,博士研究生,中国计算机学会(CCF)会员,主要研究方向为社会计算、自然语言处理. 张树森,博士研究生,主要研究方向为数据挖掘、社会计算.

complex decision boundary in the classification task. In ELM, the input weights and hidden biases connecting the input layer and the hidden layer can be independent of the training sample and randomly generated from any continuous probability distribution. The output weight matrix between the hidden layer and the output layer is obtained by minimizing the square loss function and solving the Moore-Penrose generalized inverse operation to obtain the minimum norm least squares solution. The only parameter that needs to be optimized is the number of hidden layer nodes. It has been shown by theoretical studies that ELM is capable of maintaining the universal approximation and classification capability of SLFNs even if it works with randomly generated hidden nodes. Different from traditional gradient-based neural network learning algorithms, which are sensitive to the combination of parameters and easy to trap in local optimum, ELM has faster learning speed, least human intervention and easy to implementation. In a word, ELM has become one of the most popular research directions in the field of artificial intelligence in recent years and received widespread attention from more and more research members domestic and abroad. To make it more suitable and efficient for specific applications, ELM theories and algorithms have been investigated extensively in the past few decades. Recently, random neurons have gradually been used in deep learning and ELM provides the theoretical basis for use. This paper aims to provide a comprehensive review of existing research results in ELM. We first give an introduction to the historical background and developments of ELM. Then we describe the principle and algorithm of ELM in detail followed by the introduction of its feature map and feature space. After an overview of ELM theory, we discuss and analyze state-of-the-art algorithms or the typical variants of ELM, including models, solution approaches and relevant problems. On this basis, the core ideas, advantages and disadvantages of each algorithm are summarized. Furthermore, the latest applications of ELM are reviewed. In the end, several controversies, open issues and challenges in ELM are pointed out together with its future research directions and trends.

Keywords extreme learning machine; network structure; regularization; kernel learning; deep learning; online sequential learning; parallel computing

1 引言

过去的几十年里基于梯度的学习方法被广泛用于训练神经网络,如 BP(Back Propagation)算法利用误差的反向传播来调整网络的权值。然而,由于不适当的学习步长,导致算法的收敛速度非常慢,容易产生局部最小值,因此往往需要进行大量迭代才能得到较为满意的精度。这些问题,已经成为制约其在应用领域发展的主要瓶颈。最近,Huang 等人^[1]提出了一种简单高效的单隐层前馈神经网络(Single Hidden Layer Feedforward Neural Network, SLFN)学习算法,称为极限学习机(Extreme Learning Machine, ELM)。ELM 随机选取网络的输入权值和隐层偏置,通过解析计算得到输出权值,有效地克服了传统 SLFN 学习算法的不足,已被广泛用于疾病

诊断、交通标志识别、图像质量评估等多个领域^[2-4]。ELM 最初只能用于处理单隐层前馈神经网络,后来又被推广到 RBF 神经网络、反馈神经网络、广义单隐层前馈神经网络和多隐层前馈神经网络^[5-8]。从设计目标来看,ELM 力求将回归、分类、聚类、压缩、特征提取等机器学习领域的研究问题统一到同一个框架下解决。从学习效率的角度来看,ELM 实现简单,具有极快的学习速度和较少的人为干预。从理论研究来看,即使在随机生成隐层神经元参数的情况下,ELM 仍然能够保持 SLFN 的插值能力^[9]、通用逼近能力^[10]和分类能力^[11]。从结构风险最小化的角度看,ELM 的 VC 维依赖于隐层神经元个数^[12],可以通过调节 ELM 隐层神经元个数来控制其 VC 维的大小,从而在训练误差和模型复杂度之间取得折中,以达到最优的泛化性能。从实现角度来看,硬件实现和并行计算技术大大加快了 ELM 的训练速度,使

得大数据处理和实时推理成为可能^[13-14].最近 ELM 还被扩展到深度学习中^[15-16],并取得了丰富的研究成果.这些新兴算法的提出,极大地推动了 ELM 的发展,作为人工智能领域的另一个研究热点,引起了工业界和学术界广泛深入的研究.

本文对近年来 ELM 的最新重要研究成果进行全面系统地分析和讨论,对 ELM 自提出以来的主要算法和模型进行梳理、归纳和分类,如图 1 所示.在此基础上,对 ELM 未来的发展方向提出了几点建议.本文第 2 节对当前主要的 ELM 算法和模型

进行全面介绍、总结和归纳;第 3 节对 ELM 网络结构优化算法进行详细分析、比较和分类;第 4 节系统地阐述 ELM 融合与加速模型;第 5 节深入讨论并分析深度 ELM 模型的研究进展;第 6 节回顾 ELM 及其变体在生物医学、计算机视觉和故障诊断等应用中的最新进展;第 7 节对 ELM 中存在的一些争议进行分析和讨论.针对当前的研究现状,指出 ELM 面临的问题与挑战,并对未来的研究方向和发展趋势做出展望;第 8 节对全文工作进行概括总结.

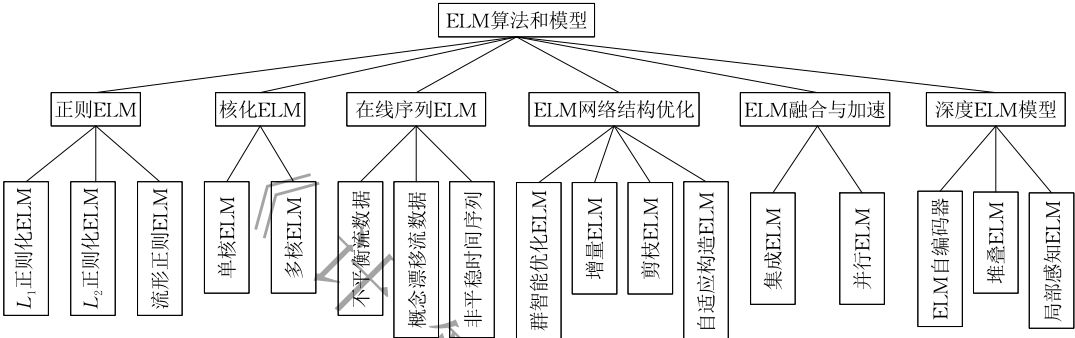


图 1 ELM 算法和模型分类

2 ELM 算法最新进展

近年来,ELM 的研究已有了较为迅速的进展,表现出广阔的发展和应用潜力,吸引了大量学术界和工业界研究人员的高度关注,并取得了丰硕的研究成果.本节将从不同角度对这些重要成果进行全面系统地分析、总结和归纳.

2.1 标准极限学习机

给定 N 个任意不同的训练样本 $\{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^N$, 其中 $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbf{R}^n$ 为输入向量, $\mathbf{t}_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbf{R}^m$ 为对应的期望输出向量. 标准的带有 n 个输入神经元, L 个隐层神经元和 m 个输出神经元且激活函数为 $g(\mathbf{x})$ 的 ELM 网络,其数学模型表示如下:

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \tag{1}$$

其中,

$$\mathbf{H} = [\mathbf{h}(\mathbf{x}_1)^T, \dots, \mathbf{h}(\mathbf{x}_N)^T]^T$$
$$= \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_N + b_L) \end{bmatrix}_{N \times L} \tag{2}$$

在 ELM 中, $\mathbf{H}$ 又被称为随机特征映射矩阵^[11], $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$ 表示连接第 i 个隐层神经元

和输入层神经元的输入权值, b_i 表示第 i 个隐层神经元的偏置, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_L]^T$ 表示输出层和隐层之间的权值矩阵, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_N]^T$ 表示训练样本期望输出矩阵. 在隐层神经元参数 $(\mathbf{w}_i, b_i)$ 根据任意连续采样分布概率随机生成并给出训练样本之后, 隐层输出矩阵 $\mathbf{H}$ 实际上是已知的, 并且保持不变. 这样, 式(1)就转化为求解线性系统 $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$ 的最小范数最小二乘解 $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^+ \mathbf{T} \tag{3}$$

其中, $\mathbf{H}^+$ 表示隐层输出矩阵 $\mathbf{H}$ 的 Moore-Penrose 广义逆.

2.2 正则极限学习机

标准的 ELM 算法^①是一个基于经验风险最小化原理的学习过程, 训练出的模型容易产生过拟合现象. 此外, 当训练样本中出现较多的离群点时, 隐层输出矩阵 $\mathbf{H}$ 具有不适定性, 导致模型的泛化能力和鲁棒性都会受到影响. 正则化理论可以用来有效地解决上述问题^[17], 正则化本质上是结构风险最小化策略的实现, 是在经验风险的基础上加入了表示模型复杂度的正则化项或惩罚项. 根据统计学习理论, 在经验风险最小化的同时, 模型越简单, 置

① ELM Source Codes. <http://www.ntu.edu.sg/home/egb-huang/>

信风险就越小,能带来更好的泛化能力. 标准正则 ELM(Regularized ELM, RELM)的数学模型可表示如下:

$$\min_{\beta \in \mathbf{R}^{L \times m}} \frac{1}{2} \|\beta\|_p^{\sigma_1} + \frac{C}{2} \|\xi\|_q^{\sigma_2} \quad (4)$$

$$\text{s. t. } \mathbf{h}(\mathbf{x}_i)\beta = \mathbf{t}_i^T - \xi_i^T, i=1,2,\dots,N$$

其中, $\sigma_1 > 0, \sigma_2 > 0, p, q = 0, 1/2, 1, 2, \dots, +\infty, \|\cdot\|_p$ 表示向量或矩阵的 L_p 范数. $\xi_i = [\xi_{i1}, \xi_{i2}, \dots, \xi_{im}]^T$ 表示样本 $\mathbf{x}_i$ 的训练误差, $\|\beta\|_p^{\sigma_1}$ 为正则化项, 表示模型的复杂度; $\|\xi\|_q^{\sigma_2}$ 是 N 个不同训练样本的总误差, 代表经验风险; $C \geq 0$ 称为正则化参数或惩罚参数, 用以平衡经验风险和模型复杂度.

特别地, 当 $\sigma_1 = \sigma_2 = p = q = 2$ 时, 式(4)是一个等式约束下的二次规划问题. 为此, 对每一个等式约束条件引入拉格朗日乘子, 通过构造拉格朗日函数转化为无约束最优化问题:

$$L(\beta, \xi, \alpha) = \frac{1}{2} \|\beta\|^2 + \frac{C}{2} \sum_{i=1}^N \|\xi_i\|^2 - \sum_{i=1}^N \sum_{j=1}^m \alpha_{i,j} (\mathbf{h}(\mathbf{x}_i)\beta_j - t_{i,j} + \xi_{i,j}) \quad (5)$$

其中, $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$, $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{im}]^T$ 为拉格朗日乘子向量.

求解式(5)可得 RELM 的输出权值矩阵:

$$\hat{\beta} = \begin{cases} \left(\mathbf{H}^T \mathbf{H} + \frac{\mathbf{I}_L}{C} \right)^{-1} \mathbf{H}^T \mathbf{T}, & L \leq N \\ \mathbf{H}^T \left(\mathbf{H} \mathbf{H}^T + \frac{\mathbf{I}_N}{C} \right)^{-1} \mathbf{T}, & L > N \end{cases} \quad (6)$$

其中, 当 $L \leq N$ 时为一般情况, 即隐层神经元个数小于训练样本个数, 此时 $\mathbf{I}$ 为 $L \times L$ 单位矩阵. 而当训练样本的个数小于隐层神经元个数 $N < L$, 利用 Woodbury 公式可以等价地求出 $\hat{\beta}$, 此时 $\mathbf{I}$ 为 $N \times N$ 单位矩阵. 显然, 在这种情况下, 计算一个 $N \times N$ 维逆矩阵要比计算 $L \times L$ 维逆矩阵高效得多.

当正则化项为参数向量的 L_2 范数时, 称为岭回归(L_2 正则化); 当正则化项为参数向量的 L_1 范数时, 称为 LASSO(L_1 正则化). 岭回归和 LASSO 是两种最典型的正则化方法, 此外还有一种称为弹性网络的正则化方法, 在损失函数中同时加入 L_1 惩罚项和 L_2 惩罚项, 对这两者进行折中. Deng 等人^[18]研究了隐层神经元为 Sigmoid 函数的 L_2 正则化 ELM, 针对数据集集中是否存在明显的噪声, 提出了无权 RELM(Unweighted RELM)和加权 RELM(Weighted RELM)算法. 加权 RELM 采用加权最小二乘法计算输出权值, 对噪声有一定的抗干扰能

力, 但由于训练过程中增加了计算误差的权值过程, 在数据量很大的时候, 时间消耗有所增加. 无权 RELM 的计算量和传统的 ELM 算法基本一样, 能够处理对实时性要求很高的应用. Zhang 等人^[19]研究了回归问题中的异常值问题, 提出了一种异常值鲁棒 ORELM(Outlier-Robust ELM)算法. 针对异常值的稀疏特性, 首先采用 L_1 正则化提高模型的鲁棒性, 然后利用 L_2 正则化提高模型的泛化性能. 此外, 一种基于交替方向迭代的增广拉格朗日法被用于求解目标损失函数. 由于在目标函数中引入了额外的正则参数需要优化, ORELM 通常需要较长的训练时间. Huang 等人^[20]利用流形正则化理论来处理未标注样例之间的关系, 提出了基于流形正则化的半监督 SS-ELM(Semi-Supervised ELM)和无监督 US-ELM(Unsupervised ELM)算法, 分别用于处理半监督学习和无监督学习任务, 大大扩展了 ELM 的适用范围. SS-ELM 和 US-ELM 继承了传统 ELM 的计算效率和学习能力, 可以快速处理多分类和多聚类问题. 文献[21]将 ELM 扩展到判别聚类, 结合迭代加权(Iterative Weighting)、线性判别分析(Linear Discriminant Analysis, LDA)和核 K -均值(Kernel K -means), 提出了 $\text{ELMC}^{\text{Iter}}$ 、 ELMC^{LDA} 和 ELMC^{KM} 三种判别聚类算法. 基于 ELM 的判别聚类算法在 ELM 输出权值的求解过程中采用正则化技术来提高聚类的准确性和鲁棒性, 不仅比 US-ELM 有更少的可调参数, 而且还能够处理非平衡聚类问题. $\text{ELMC}^{\text{Iter}}$ 、 ELMC^{LDA} 和 ELMC^{KM} 属于离线学习算法, 需要事先给定所有的训练样本, 不具备在线学习能力. Yi 等人^[22]将多图学习方法引入到 ELM 中, 提出了一种自适应多图正则化半监督 AMGR-SSELM(Adaptive Multiple Graph Regularized Semi-Supervised ELM)算法. AMGR-SSELM 结合了六种不同的图构造技术, 并且可以自适应地将不同权重分配给相应的图, 有助于刻画出数据内在的流形结构. 由于使用了多图结构, 该算法的训练时间要高于使用单一图的 SS-ELM 算法, 特别是在大型数据集上其计算复杂度是无法容忍的. Iosifidis 等人^[23]将子空间学习准则引入到 ELM 的优化过程中, 提出了图嵌入 GEELM(Graph Embedded ELM)算法. 该算法充分利用了图嵌入框架下的惩罚图和本征图, 并且可以挖掘 ELM 特征空间中数据之间的局部近邻和全局结构信息, 在人脸识别和行为识别上取得了一定的效果, 但和深度学习相比, 其表征能力仍然有限. 针对 ELM 的随机特征映射不能适应

隐层神经元之间的横向抑制问题,Yu 等人^[24]提出一种稀疏编码 ScELM(Sparse coding ELM)算法.该算法使用稀疏编码技术代替随机映射将输入特征向量映射到隐层,以提高分类准确率.在编码阶段使用基于梯度投影和 L_1 范数的优化方法,而输出权值则通过拉格朗日乘子法求得.加权 RELM 对零均值高斯噪声有很好的稳健性,但是其性能严重依赖于误差权值估计的准确性,特别是在复杂应用环境下,会产生高昂的计算成本,导致模型的效率降低.为了解决这一问题,Lu 等人^[25]提出了一种概率正则 ELM 算法.该算法考虑了建模过程中误差的分布,确保模型误差和噪声分布的一致性,通过构造一个新的目标函数来最小化模型误差的均值和方差,从而提高了模型对离群点或非高斯噪声的鲁棒性.然而,该算法无法处理具有不同噪声分布的大规模数据集.此外,如何将其扩展到处理非平衡数据、半监督和无监督的情况下还需进一步研究. Zhao 等人^[26]将模型的偏差和方差同时引入到目标函数中进行优化,并保持 L_2 惩罚项不变,提出了鲁棒 RRELM(Robust RELM)算法. RRELM 同时考虑了模型的偏差和方差,力求在两者之间取得最佳的折中,以达到增强网络泛化性能和鲁棒性的目的. 可以从理论上证明,RELM 是 RRELM 的一种特例,并且 RRELM 拥有和传统 ELM 及 RELM 相同的计算效率. Chen 等人^[27]进一步研究了不同的 M 估

计函数和正则化项在 ELM 中的作用,基于稳健回归理论,提出了一种使用迭代重加权最小二乘法 (Iteratively Reweighted Least Squares, IRLS) 的鲁棒正则 RELM-IRLS(Robust Regularized ELM Regression Using IRLS)回归模型. 在 RELM-IRLS 中, L_1 范数、Huber 损失函数、Bisquare 损失函数和 Welsch 损失函数同时被用来增强模型的鲁棒性,并使用 L_1 正则化和 L_2 正则化来防止过拟合. 由于 RELM-IRLS 需要优化的参数较多,导致算法的时间复杂度较大,因此无法有效处理大规模数据集. 针对不平衡数据的分类问题,Xiao 等人^[28]提出了 CCR-ELM(Class-specific Cost Regulation ELM)算法,通过对不同类的错分样本施加不同的惩罚因子,实现错分样本数目和模型泛化能力的折中. 和标准 ELM 相比,CCR-ELM 可以显著提高非平衡数据的分类准确率. CCR-ELM 中的隐节点个数、正负样本权重以及核参数对模型的性能影响较大,如何开发更有效的方法来确定这些参数还有待进一步研究.

上述研究表明,将正则化引入 ELM 中能够在一定程度上解决过拟合问题,提高模型的鲁棒性和泛化能力.但是由于在目标函数中增加了正则化参数需要进行优化,降低了算法的学习效率.表 1 对当前部分正则极限学习机模型进行了汇总,并对它们的算法思想、采用的正则化方法、特征映射、鲁棒性、测试数据集和应用等方面进行了比较.

表 1 正则极限学习机模型

模型	核心思想	L_1 正则化	L_2 正则化	流形正则化	特征映射	加权二乘	鲁棒性	评测数据集	应用
WRELM ^[18]	§ 结构风险 § 加权残差		✓		Sigmoid	✓	✓	※人工数据“SinC” ※13 个基准数据集	回归
ORELM ^[19]	§ 权重稀疏 § ALM	✓	✓		Sigmoid		✓	※人工数据“SinC” ※3 个 UCI 数据集	回归
SS-ELM ^[20] US-ELM ^[20]	§ 图嵌入		✓	✓	Sigmoid Gaussian	✓	✓	※5 个半监督数据集 ※3 个 UCI 数据集 ※2 个人脸数据集	分类 回归 聚类
AMGR-SSELM ^[22]	§ 多图构造	✓	✓	✓	Sigmoid		✓	※4 个人脸数据库 ※PloyU FKP 数据库	分类
GEELM ^[23]	§ 图嵌入 § 子空间学习		✓	✓	Gaussian			※8 个分类数据集 ※9 个行为数据集	分类
ScELM ^[24]	§ 稀疏编码 § 梯度投影	✓			Sigmoid			※8 个二分类数据集 ※8 个多分类数据集	分类
RRELM ^[26]	§ 偏差方差最小化		✓		Sigmoid Fourier	✓	✓	※5 个 UCI 数据集 ※发动机推力数据	分类 回归
RELM-IRLS ^[27]	§ 稀疏权重 § IRLS	✓	✓		Sigmoid	✓	✓	※人工数据“SinC” ※5 个基准数据集	回归
CCR-ELM ^[28]	§ 错分样本损失惩罚		✓		Sigmoid Gaussian	✓	✓	※19 个 UCI 数据集 ※高炉状态诊断	分类

2.3 单核极限学习机

ELM 隐层的特征映射有很多种选择,既可以是显式映射,也可以是隐式映射. 隐式映射巧妙利用核方法得到特征向量之间的内积,因此不需要显式地定义特征空间和映射函数. 核方法实质是通过核函数隐式地将输入空间中低维的线性不可分样本映射到高维甚至无限维的特征空间,使得原空间的非线性可分问题转化为特征空间中的线性可分问题,是机器学习领域里一类非常重要的方法. 传统的 ELM 算法采用显式的非线性特征映射,对于比较复杂的分类、回归等非线性模式识别任务往往需要更多的隐层神经元,导致网络的结构非常复杂. Huang 等人^[11]将核函数引入到 ELM 中,提出了基于核方法的 KELM(Kernel ELM)算法,为回归、二分类和多分类问题提供了一个统一的学习框架.

在隐层的特征映射 $\mathbf{h}(\mathbf{x})$ 具体形式未知的情况下,需要引入核函数来度量样本之间的相似度,可以根据 Mercer 条件定义 ELM 的核矩阵,表示如下:

$$\mathbf{\Omega}_{\text{ELM}} = \mathbf{H}\mathbf{H}^T: \Omega_{\text{ELM},i,j} = \mathbf{h}(\mathbf{x}_i) \cdot \mathbf{h}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) \quad (7)$$

这样, KELM 模型的输出可以表示为

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \mathbf{h}(\mathbf{x})\mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T \right)^{-1} \mathbf{T} \\ &= \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}^T \left(\frac{\mathbf{I}}{C} + \mathbf{\Omega}_{\text{ELM}} \right)^{-1} \mathbf{T} \quad (8) \end{aligned}$$

从上式可以看出,核矩阵 $\mathbf{\Omega}_{\text{ELM}} = \mathbf{H}\mathbf{H}^T \in \mathbf{R}^{N \times N}$ 仅和输入数据 $\mathbf{x}_i$ 以及训练样本的个数有关. 在 KELM 中,通过核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$ 将低维输入空间中的数据 $(\mathbf{x}_i, \mathbf{x}_j)$ 转化为高维特征空间中的内积 $\mathbf{h}(\mathbf{x}_i) \cdot \mathbf{h}(\mathbf{x}_j)$,与特征空间的维数无关,可以有效避免维数灾难问题. KELM 只需要预先选定核函数,不需要显式地定义映射函数,也不需要设置隐层神经元个数,从而节省了隐层神经元个数优化的时间. 相比于传统的 ELM 算法, KELM 用核映射代替随机映射,能够有效改善隐层神经元随机赋值带来的泛化性和稳定性下降的问题.

针对二分类问题, Liu 等人^[29]将 ELM 的学习思想引入到 SVM(Support Vector Machine)中,提出了 ESVM(Extreme Support Vector Machine)算法. 该算法首先将输入数据映射到 ELM 特征空间,然后在特征空间中采用正则最小二乘法求解分类超平面. ESVM 仅需要求解一个简单的线性方程组,和标准的 SVM 相比,具有实现简单,训练速度更快的优点. 然而, ESVM 没有考虑如何求得最大

间隔分离超平面,也没有充分利用对偶算法的优点. Frénay 等人^[30]通过定义 ELM 核(ELM kernel)将 SVM 和 ELM 算法融合在一个框架中. ELM kernel 由传统 ELM 隐层的特征映射 $\mathbf{h}(\mathbf{x})$ 扩展为核方法中的核映射得到,然后被用来代替 SVM 的核函数进行训练. 和传统的核函数相比, ELM kernel 不需要优化核参数,降低了计算复杂度,并且能够得到最大间隔分类超平面. Huang 等人^[31]进一步研究了基于标准优化方法的二分类 ELM 算法,通过对每个样本点 $(\mathbf{x}_i, \mathbf{x}_j)$ 引入一个松弛变量 $\epsilon_i \geq 0$,使函数间隔加上松弛变量大于等于 1. 此时,约束条件变成下式:

$$t_i \boldsymbol{\beta} \cdot \mathbf{h}(\mathbf{x}_i) \geq 1 - \epsilon_i, \quad i = 1, 2, \dots, N \quad (9)$$

其中, $t_i \in \{-1, +1\}$, $\mathbf{x}_i \in \mathbf{R}^n$, 当 $t_i = -1$ 时,对应的样本 $\mathbf{x}_i$ 称为负类;当 $t_i = +1$ 时,对应的样本 $\mathbf{x}_i$ 称为正类.

这样,从标准优化理论的角度来看, ELM 的学习问题可以转化为如下二次规划问题:

$$\begin{aligned} \min_{\boldsymbol{\beta}, \epsilon} \quad & \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N \epsilon_i \\ \text{s. t.} \quad & t_i \boldsymbol{\beta} \cdot \mathbf{h}(\mathbf{x}_i) \geq 1 - \epsilon_i, \quad i = 1, 2, \dots, N \\ & \epsilon_i \geq 0 \end{aligned} \quad (10)$$

其中, C 为惩罚参数, C 取值越小,对误分类的惩罚越小, C 取值越大,对误分类的惩罚越大.

从式(10)可以看出, ELM 和标准 SVM 具有相似的优化目标函数,主要区别在于:

(1) ELM 的特征映射 $\mathbf{h}(\mathbf{x})$ 具有随机性,即 $\mathbf{h}(\mathbf{x})$ 中的所有参数均可以随机产生.

(2) SVM 的判别函数中含有偏置项,而 ELM 特征空间中的分类超平面会经过原点,因此在对偶问题中少了对偏置项的等式约束条件.

基于标准优化方法构建的 ELM 模型要比 Liu^[29]和 Frénay^[30]等人提出的算法具有更少的约束条件. 此外,最小化输出权值 $\|\boldsymbol{\beta}\|$ 等同于最大化分类间隔 $2/\|\boldsymbol{\beta}\|$, 这和 SVM 的学习策略相一致,如图 2 所示. 文献[32]对 ELM 的核函数以及 ELM 和 SVM 之间的关系进行了深入的研究和比较,详细分析了在使用相同核函数的情况下 SVM 和 LS-SVM 比 ELM 更容易产生次优解的原因. 为了能够快速处理在线大规模数据学习问题, He 等人^[33]提出了一个并行增量 PIESVM(Parallel Incremental Extreme SVM)分类算法. 该算法首先基于并行编程框架 MapReduce 实现 ESVM 的并行化,提高 ESVM 处理大规模数据的训练速度,然后又提出了

一种 ESVM 增量学习算法来满足在线学习对现有模型进行更新的需求. PIESVM 不仅可以处理大规模数据,而且在加速比、规模增长性等评估指标方面具有很好的扩展性. 针对 KELM 在处理大样本数据时计算复杂度和空间复杂度过高的问题, Deng 等人^[34]采用 Nyström 方法来求解核矩阵的近似低秩分解,实现了一种加速 NKELM(Nyström Kernel ELM)算法. NKELM 的训练时间远低于传统的 KELM,适合用来处理大规模数据. NKELM 的惩罚参数和核参数通常使用交叉验证法确定,模型的泛化性能对这些参数的取值非常敏感,因此如何高效自适应确定 NKELM 的参数还需进一步研究. 此后, Deng 等人^[35]又提出了一种快速的非迭代简化核 RKELM(Reduced Kernel ELM)算法. 该算法从训练数据集中随机选择一个子集作为支撑向量,无需迭代计算节省了大量的训练时间. 理论研究表明,当核函数严格正定并且所有的训练样本都被选为支撑向量时, RKELM 就能零误差地逼近任意非线性函数. 如何将 RKELM 扩展到半监督和无监督学习任务中值得进一步研究. 针对海量数据学习问题, Bi 等人^[36]从理论角度研究了 KELM,并提出了一个分布式解决方案 DK-ELM(Distributed Kernelized ELM). 通过在 MapReduce 上实现并行化的 KELM,可以有效避免单机环境下大规模核矩阵运算的大量内存消耗问题. 然而,当数据集中存在较多离群点时, DK-ELM 的泛化能力和鲁棒性都会受到较大的影响,无法有效学习. 传统的 ELM 和 KELM 都不具有稀疏性,其逆矩阵的计算复杂度至少与训练样本的规模呈二次方关系,因此在处理大规模二分类问题时,往往需要消耗大量的存储空间和训练时间. 受序列最小优化算法的启发, Bai 等人^[37]提出了一种高效的极限学习机训练方法. 该方法将优化过程中的原始大型二次规划问题分解为一系列最小的

子二次规划问题迭代求解,避免了矩阵的逆运算,从而降低了时空复杂度. Chen 等人^[38]将 MPE(Mean p -Power Error)引入到核空间中,定义了一个新的核空间统计量 KMPE(Kernel Mean p -Power Error),并将其与 ELM 相结合,提出了一种鲁棒 ELM-KMPE 学习算法. 虽然 ELM-KMPE 能够处理非高斯信号且对噪声具有一定的鲁棒性,但是如何高效地优化 KMPE 中的核参数和 p 参数仍是一个难题.

2.4 多核极限学习机

KELM 的泛化性能在很大程度上依赖于核函数及其参数的选择,针对不同的问题如何设计相应的核函数仍然是核学习中的一个开放性问题. 常用的核函数主要有:线性核函数、多项式核函数、高斯核函数和 Sigmoid 核函数. 除了上述核函数外,还有字符串核、小波核、卷积核等,这些核函数还可以作为基函数,通过线性组合或非线性组合等多核学习方法构造出混合核函数. Liu 等人^[39]提出了一个基于多核 ELM 的多源异构数据集成框架 MK-ELM(Multiple Kernel ELM). 在 MK-ELM 中,假设最优核为一组基核函数的线性组合,在训练过程中对基核函数的组合权重和 ELM 的网络结构参数进行联合优化. 对于不同的应用,如何自适应选择基核函数并快速确定不同核的权系数,提高 MK-ELM 的训练效率仍是一个难点. 针对人体动作识别中存在不平衡数据的问题, Wu 等人^[40]提出了一种基于混合核函数的加权 MK-WELM(Mixed-kernel Based Weighted ELM)算法. 该算法根据样本的分布信息,将代价敏感函数引入到 ELM 中,以减少不平衡数据对分类器的影响. 同时,为了进一步提高识别准确率,减少核函数对分类器的影响,用高斯核和多项式核的线性组合优化核来训练模型. 由于 MK-WELM 中需要优化的参数较多,导致算法的计算复杂度很高,因此无法有效处理大规模数据集. Li 等人^[41]研究了大规模数据集中 ELM 的多核组合优化,并将其转化为一个半无限线性规划问题,提出了基于多核学习的 MKL-ELM(Multiple-Kernel Learning ELM)算法. 在多核模型中,核函数既可以是给定的不同核的凸组合,也可以是同一个核的不同参数的凸组合,而最优核由黑盒学习方法得到. 虽然多核 ELM 算法在理论上和实验结果中表现出比单核 ELM 更强的非线性表达能力和稳定性. 但是随着核函数的增加,参数的优化问题也凸显出来,通常需要更多的时间进行模型选择,增加了计算复杂度. 针对金融领域多源异构数据中的信息融合问

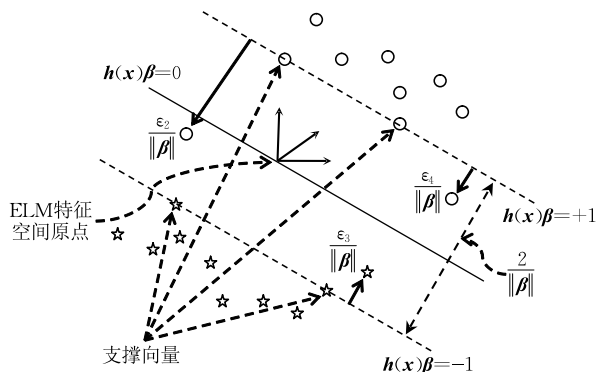


图2 ELM 特征空间中的最优分类超平面

题,Xue 等人^[42]结合增量学习和多核学习方法,提出了一种增量多核 IMK-ELM(Incremental Multiple Kernel ELM)算法,并将其用于智能理财顾问推荐系统中.该模型最初只设置少量隐层神经元,然后在训练过程中逐步添加新的神经元直到网络的训练结果稳定下来.IMK-ELM 在优化过程中同时更新训

练数据集和多源信息组合权重,并假设最优核为不同特性核函数的线性组合,融合多类核函数的优点,从而得到更高水平的综合特征来表示投资者的偏好.

表 2 从算法核心思想、特征映射、测试数据集和应用这几个方面对当前最新的核极限学习机模型进行了总结.

表 2 核极限学习机模型

模型	核心思想	特征映射	评测数据集	应用
NKELM ^[34]	§ 低秩分解	Gaussian kernel	※5 个基准数据集	分类
RKELM ^[35]	§ 随机选择支撑向量	Gaussian kernel	※6 个大规模数据集 ※15 个小规模数据集	分类 回归
DK-ELM ^[36]	§ MapReduce	Gaussian kernel	※4 个合成数据集 ※4 个 UCI 数据集	分类
Sparse-ELM ^[37]	§ SMO	Gaussian kernel Polynomial kernel	※15 个基准数据集	分类
ELM-KMPE ^[38]	§ KMPE	Gaussian kernel	※人工数据“SinC” ※16 个基准数据集	分类 回归
MK-WELM ^[40]	§ 多核学习 § 代价敏感函数	Gaussian-Gaussian Polynomial-Gaussian Polynomial-Polynomial	※LDPA 行为数据集 ※7 个二分类不平衡数据集 ※5 个多分类不平衡数据集	分类
MKL-ELM ^[41]	§ 多核学习 § 半无限线性规划 § 二次约束二次规划	Gaussian-Polynomial- Linear-Laplacian Gaussian-Gaussian	※13 个分类数据集	分类
IMK-ELM ^[42]	§ 增量学习 § 多核学习	Gaussian kernel Polynomial kernel	※BCSs 数据集	分类 回归

2.5 在线序列极限学习机

ELM 的学习过程通常都是在给定的训练集上完成的,即需要一次性获取所有的训练样本来进行训练.然而,对于实际应用环境而言,训练数据不是静态不变的,而是逐个或者逐批加入的,是一个随时间不断更新变化的过程.当有新的训练样本出现时,ELM 会对原有数据和新数据重新进行训练,从而导致学习效率低下,难以满足在线实时应用的需求.为了解决这一问题,Liang 等人^[43]提出了一种基于递推最小二乘法的在线序列 OS-ELM(Online Sequential ELM)算法.该算法首先通过少量的训练样本初始化网络的输出权值,然后在增量学习过程中,利用新增的样本或样本块递推更新上一步求出的网络输出权值.OS-ELM 只对当前产生的新样本进行更新,不需要重复训练旧的样本,能够处理训练样本逐块或逐个到达的情况,并且允许样本块的大小可变,因此具有良好的在线学习能力.虽然 OS-ELM 通过分块矩阵的方法避免了旧数据的重新训练,减少了模型的训练时间,但由于需要保存旧训练样本的更新矩阵,增加了空间复杂度.Lan 等人^[44]将集成的思想引入到 OS-ELM 中,提出了 EOS-ELM

(Ensemble of Online Sequential ELM)算法.该算法首先训练多个具有相同隐层神经元个数和激活函数的 OS-ELM 子网络,然后通过简单平均的策略对这些 OS-ELM 子学习器进行集成.EOS-ELM 在一定程度上减少了隐层神经元参数随机赋值对模型稳定性产生的影响,提高了 OS-ELM 的泛化性能.但由于学习过程中需要训练多个 OS-ELM 并对其结果进行存储,通常需要较高的时间开销和空间成本.

在诸如股票价格预测,天气预报等大量实际应用中,训练数据除了具有实时动态变化的特外性往往还具有时效性.因此,在增量学习过程中加入新样本的同时,也应该及时淘汰那些过于陈旧的样本.OS-ELM 和 EOS-ELM 显然不能很好地反映在线训练数据的时效性.针对这一问题,Zhao 等人^[45]将遗忘机制和 EOS-ELM 相结合,提出了 FOS-ELM(Online Sequential ELM with Forgetting Mechanism)算法.该算法通过设置滑动时间窗口来完成模型的在线更新,能够在学习过程中丢弃过时的数据,减少其对后续训练的不良影响,具有良好的在线数据处理能力.Gu 等人^[46]提出了一种基于 ELM 的时效在线序列 TOSELM(Timeliness Online Sequential ELM)算

法. 该算法通过计算时序数据的均值和方差并采用指数函数对新增数据分配不同的惩罚权重, 同时使用一种自适应迭代方案来更新输出权值以提高模型的稳定性和收敛性. 该方法在训练过程中引入了额外的迭代和停止条件, 增加了模型的计算复杂度. Huynh 等人^[47]将结构风险最小化理论引入到 OS-ELM 中, 提出了一种基于正则 ELM 的在线序列 ReOS-ELM(Regularized OS-ELM)算法. 该算法在经验风险的基础上加了一个控制目标函数光滑程度的 Tikhonov 正则化项, 通过调节正则化参数, 选择复杂度适中的模型, 以达到测试误差最小的学习目的. 针对在线学习中类别不平衡问题, Mirza 等人^[48]提出了加权 WOS-ELM(Weighted Online Sequential ELM)算法. 该算法采用衡量正负样本分类精度的平衡因子 G-mean 作为分类器整体性能的评价指标, 对不同的数据块施加不同的权值以使 G-mean 尽可能最大, 这样就不再需要对旧的训练样本进行权值调整, 可以缓解不平衡样本分布带来的问题.

在实际应用中, 由于计算复杂性、存储和传输等问题, 限制了大规模数据的集中处理. 为了解决这一问题, Vanli 等人^[49]对分布式多智能体系统中的非线性序列学习问题进行了研究, 并提出了一种分布式序列分裂 DSS-ELM(Distributed Sequential Splitting ELM)算法. 研究表明, DSS-ELM 的计算复杂度仅和隐层神经元的个数呈线性关系, 不仅适用于大规模数据流, 而且还能以较快的速度处理非平稳环境中的数据. 针对非平稳时间序列的预测问题, Wang 等人^[50]提出了一种基于在线序列训练的核 OS-ELMK(Online Sequential ELM with Kernels)预测算法. 在 OS-ELMK 中, 使用隐式的核映射代替显式的特征映射函数, 不需要确定隐层输出矩阵, 通过计算核函数来得到网络的最终输出, 避免了隐层神经元个数优化的难题. 然而, OS-ELMK 的抗干扰能力比较差, 无法处理带有噪声的数据流. Scardapane 等人^[51]研究了 ELM 理论和核自适应滤波之间的内在联系, 并将核递归最小二乘法扩展到 OS-ELM 框架中, 提出了一种基于核函数的 KOS-ELM(Kernel Online Sequential ELM)在线学习算法. 该算法采用近似线性依赖和固定预算这两种稀疏化准则来控制网络结构的生长. 由于在训练过程中增加了多重判断机制, KOS-ELM 通常需要较高的时间开销. Deng 等人^[52]提出了一种

在线序列简化核 OS-RKELM(Online Sequential Reduced Kernel ELM)算法, 为在线学习中的二分类、多分类和回归任务提供了一个统一的学习框架. OS-RKELM 不仅支持增量学习, 同时还支持学习模型中的递减更新. 但是, 当数据流中存在较多离群点时, OS-RKELM 的鲁棒性会变得非常差, 从而影响模型的泛化性能. 针对非静态环境中数据流的概念漂移问题, Liu 等人^[53]在 OS-RELM 的基础上提出了一种带有遗忘系数的在线序列 FP-ELM(Forgetting Parameters ELM)算法. 该算法的核心思想是对训练过程中不同时期加入的数据块施加不同的权重. 但是 FP-ELM 在计算遗忘参数的过程中引入了额外的可调节参数, 增加了计算复杂度. Zhang 等人^[54]将元认知理论扩展到 OS-ELM 中, 并采用主动学习方法选择最具代表性的样例加入样本池来扩充训练集的规模, 提出了一种基于 ELM 的元在线序列主动学习算法 SEAL-ELM(Sequential Active Learning Using Meta-cognitive ELM). 该算法由认知成分和元认知成分两部分构成, 将 OS-ELM 作为认知成分, 而元认知成分使用自我调节机制来控制认知成分的学习过程. SEAL-ELM 通过主动学习选择有价值的未标注样例进行标注, 减少了训练所需的标注代价, 同时提高了分类器的分类精度. 在实际的工业生产过程中采集的样本往往带有不同统计特征的噪声, 针对这一问题, Yang 等人^[55]提出了 RLMP-ELM(Recursive Least Mean p -Power ELM)在线序列学习算法. 该算法采用 LMP(Least Mean p -Power)误差准则来更新 ELM 的输出权值, 能够对具有不同统计噪声的变量进行在线预测. 尽管 RLMP-ELM 可以通过设置不同的 p 值范围处理重尾分布和轻尾分布数据, 但是如何确定准确的 p 值才能获得最好的泛化性能目前尚不清楚. 为了解决标注样本容量的限制问题, Yang 等人^[56]将增量拉普拉斯正则化引入到 ELM 中, 提出了一种半监督的 ILR-ELM(Incremental Laplacian Regularization ELM)在线学习算法. 该算法利用图拉普拉斯将标记样本与未标记样本合并, 因此只需要少量标记样本即可实现稳健的分类和回归模型.

表 3 对当前最新的在线序列极限学习机算法进行了归纳, 并对它们的核心思想、特征映射、评测数据集和应用等方面进行了总结.

表 3 在线序列极限学习机模型

模型	核心思想	特征映射	评测数据集	应用
OS-ELM ^[43]	§ 递推最小二乘法	Sigmoid active function Gaussian function	※3 个分类数据集 ※3 个回归数据集 ※1 个时间序列数据集	分类 回归
FOS-ELM ^[45]	§ 遗忘机制 § 集成学习	Sigmoid active function Gaussian function	※3 个分类数据集 ※11 个回归数据集	分类 回归
ReOS-ELM ^[47]	§ Tikhonov 正则化	Sigmoid active function Gaussian function	※8 个分类数据集 ※3 个回归数据集	回归
OS-RKELM ^[52]	§ 核学习方法	Gaussian kernel Sigmoid kernel ANOVA kernel Multiquadric kernel	※3 个回归数据集 ※6 个分类数据集 ※1 个时间序列数据集	分类 回归
FP-ELM ^[53]	§ 遗忘系数 § 正则化优化方法	Sigmoid active function	※4 种概念漂移数据集	分类 回归
SEAL-ELM ^[54]	§ 元认知 § 主动学习	Sigmoid active function	※9 个基准数据集	分类
RLMP-ELM ^[55]	§ 递归最小平均 p 范数	Gaussian function	※人工数据“SinC” ※2 个回归数据集 ※1 个时间序列数据集	回归
ILR-ELM ^[56]	§ 图拉普拉斯	Sigmoid active function	※3 个基准数据集	分类 回归

3 ELM 网络结构优化

在 ELM 中,网络的输入权值和隐层偏置随机生成,在训练过程中避免了大量迭代调整,只需要确定激活函数和隐层神经元个数,而输出权值可以通过最小二乘法直接计算得到,因此隐层神经元个数的确定与否直接影响到模型泛化性能的好坏.如何为不同的应用选则合适的隐层神经元数目一直以来都是神经网络研究领域中的一个难点和热点.隐层神经元数量过少,网络从训练样本中提取的信息不足以概括和反映数据的内在规律,映射能力不强,会导致欠拟合问题;隐层神经元数量太多,会使得网络结构变得非常复杂,不仅增加了训练时间,而且容易引起过拟合问题,从而影响网络的整体性能.目前,ELM 网络的隐层结构调整方法主要分为群智能优化、增量法、剪枝法和自适应这四大类,本文对这几种类型的神经元优化算法和最新研究进展进行了全面的分析和归纳.

3.1 群智能优化

虽然 ELM 算法通过随机选择隐层神经元参数提高了整个网络的学习速度,但大量研究表明^[57-61],由于输入权值和隐层偏置的随机赋值,使得 ELM 在某些应用中往往比传统方法需要更多的随机隐层神经元,而过多的随机神经元容易导致隐层不具有稀疏性和调节能力,会产生一些对模型训练作用很小甚至没有作用的冗余神经元,从而导致输出矩

阵 H 不是列满秩矩阵或者是一个病态矩阵而无法得到精确的解析解,最终影响网络的泛化能力和稳定性. Zhu 等人^[57]将差分进化引入到 ELM 中,提出了一种进化 E-ELM(Evolutionary ELM)算法.该算法利用结构简单的差分变异和交叉算子,并根据整个种群的动态调整来搜索最优的输入权值和隐层偏置,从而获得更紧凑的网络结构.针对 E-ELM 算法中个体生成策略及相应的控制参数需要手动设置的问题, Cao 等人^[58]提出了一种自适应进化 SaE-ELM(Self-adaptive Evolutionary ELM)算法.该算法不仅能够自适应调整交叉概率和缩放因子,还可以从四种子代个体生成策略中自动选择最佳变异策略,提高了算法的收敛性. SaE-ELM 无法处理不平衡数据分类问题, Tang 等人^[59]提出了一种基于 WELM 的自适应进化加权 SDE-WELM(Self-adaptive Differential Evolutionary Weighted ELM)算法,利用自适应差分进化算法优化隐层神经元参数以及训练样本权重,从而提高二类不平衡数据分类准确率. Xu 等人^[60]利用粒子群优化(Particle Swarm Optimization, PSO)算法良好的全局搜索能力优化 ELM 的隐层神经元参数,提出了一种 PSO-ELM 算法.该算法将 ELM 的输入权值和隐层偏置编码为 PSO 搜索空间中的粒子.在每一次迭代过程中,所有粒子通过自身的历史最优解和整个种群当前的全局最优解来更新自己的位置,从而在解空间中搜索最优值. PSO 优化的效果取决于其拓扑结构,即整个群体中所有粒子之间相互连接的方式以及单

个粒子和其它粒子间的相互连方式. Figueiredo 等人^[61]研究了八种不同的拓扑结构对 PSO-ELM 性能的影响, 结果表明并没有适合于所有问题的最佳拓扑结构, 但从实验中验证集上的误差来看, 使用全局拓扑结构在大多数评测数据集上要比其它拓扑结构表现更好. Lahoz 等人^[62]提出了一种基于双目标微种群遗传 μ G-ELM(Multi-objective Micro Genetic ELM)算法, 在双目标函数优化过程中同时最小化均方根误差和隐层神经元数来获取合适的输入权值和隐层偏置. μ G-ELM 的优化后期, 种群的多样性越来越少, 搜索的空间有限, 在测试数据生成中经常会出现早熟收敛现象, 从而容易陷入局部最优解, 此外 μ G-ELM 的泛化性能严重依赖于种群中个体的初始化值. Zhang 等人^[63]提出了一种基于文化基因的 M-ELM(Memetic Algorithm Based ELM)算法, M-ELM 将个体的局部启发式搜索融入到种群的全局优化框架中来自动学习最优的网络参数, 能够有效克服早熟收敛的问题.

虽然利用群智能优化隐层神经元参数提高了 ELM 的稳定性, 减少了冗余的隐层神经元, 但同时也引入了大量的超参数并且通常需要对其进行优化, 增加了算法的计算复杂度, 因此无法有效处理大规模数据集等复杂学习问题.

3.2 增量构造法

ELM 的网络结构还可以通过增量构造的方式来调整, 这类算法的主要思想是, 在网络初始化阶段仅设置少量的隐层神经元, 在训练过程中按照一定的规则逐个或逐块增加隐层神经元. Huang 等人^[64]通过增量构造法证明了 ELM 的通用逼近能力, 并提出了一种增量 I-ELM(Incremental ELM)算法. 该算法在学习过程中每次向隐层添加一个新的随机神经元, 直到隐层神经元数达到预先设定的最大值或者是误差满足预定的精度为止. 当一个新神经元加入网络后, 原先已存在于网络中的隐层神经元的输出权值固定不变, 只需要计算新增神经元和输出层之间的连接权值. 为了进一步提高 I-ELM 的收敛速度, Huang 等人^[10]提出了一种基于凸优化理论的增量 CI-ELM(Convex I-ELM)算法. 该算法和 I-ELM 的本质区别在于, 每增加一个新的隐层神经元, 不仅需要计算新增神经元的输出权值, 还会利用凸函数更新已有的隐层神经元的输出权值. 文献^[65]将 I-ELM 从实数域扩展到复数域, 并证明了只要隐层激活函数满足有界非线性分段连续, I-ELM 仍然能够逼近复数域中的任何目标函数. 在 I-ELM

中, 每次新加入的隐层神经元都是随机产生, 没有经过任何筛选, 无法保证每次加入的神经元都能够让网络误差下降最快, 因此容易产生一些对网络输出作用很小的隐层神经元. 为了解决这一问题, 文献^[66]对 I-ELM 进行了改进, 提出了 EI-ELM(Enhanced I-ELM)算法. EI-ELM 在每次迭代过程中产生 k 个随机候选隐层神经元, 通过计算并比较这 k 个神经元的误差选出拟合效果最好的一个作为新神经元添加到网络的隐层中, 这样不仅加快了算法的收敛速度, 而且还能得到结构更紧凑的网络. 当 k 取值为 1 时, EI-ELM 算法就退化为 I-ELM 算法. Feng 等人^[67]提出了一种基于误差最小化的 EM-ELM(Error Minimized ELM)算法, 该算法中隐层神经元的添加数目不再固定不变, 而是允许隐层神经元逐个增加或逐块增加, 并通过分块矩阵递归更新所有隐神经元的输出权值, 从而提高网络整体的学习速度. 为了解决回归应用中网络结构的设计问题, Lan 等人^[68]基于逐步前向选择法提出了一种构造性隐层神经元选择算法 CS-ELM(Constructive Hidden Nodes Selection for ELM). 该算法将隐层神经元的选择视为线性回归中子集模型选择或变量选择问题, 利用无偏风险估计准则 C_p 来评估模型的性能, 并选择具有最小 C_p 值的隐神经元组合作为最优子集. CS-ELM 算法仍存在问题, 当新神经元加入后可能会影响原先神经元的有效性. Yang 等人^[69]在 I-ELM 的基础上提出了一种基于误差双向反馈的双端增量 B-ELM(Bidirectional ELM)算法. 该算法利用误差的反向传播来直接计算部分隐层神经元参数, 从而降低了随机搜索最优隐层神经元的时间. 通过进一步研究发现, ELM 网络的输出权值和输出误差之间存在椭圆方程关系, 并根据此方程从理论上证明了 B-ELM 只需迭代两次就能够使网络输入误差趋向于零, 从而极大地减少了模型的训练时间. EM-ELM 的计算复杂度比原始的 ELM 高, 为了减少网络的训练时间, Ye 等人^[70]提出了一种基于 QR 分解的增量 QRI-ELM(QR Factorization Based Incremental ELM)算法. Li 等人^[71]提出了一种策略能够随着隐层神经元的增加迭代更新输出权值, 避免了传统 ELM 中的矩阵求逆运算. EM-ELM 是在经验风险最小化原则的指导下选择最优的随机隐层神经元加入到网络中, 因此在训练过程中往往会产生过拟合现象, 影响模型的泛化性能. 为了解决这一问题, Xu 等人^[72]结合结构风险最小化理论, 提出了一种基于正则化的 EIR-ELM

(Enhancement of Incremental Regularized ELM)算法. 该算法通过 RELM 更新网络的输出权值, 同时为了控制模型的复杂度, 在平方损失函数的基础上加入了一个 L_2 正则化项, 然后计算所有候选隐层神经元的误差并选择值最小的神经元添加到网络

中. 虽然 EIR-ELM 减少了过拟合问题的发生, 但由于引入了额外的惩罚参数需要优化, 增加了计算复杂度.

表 4 总结了几种典型的极限学习机群智能优化和增量构造算法.

表 4 极限学习机群智能优化和增量构造模型

模型	核心思想	特征映射	评测数据集	应用
E-ELM ^[57]	§ 差分进化	Sigmoid active function	※ 4 个基准分类数据集	分类
PSO-ELM ^[61]	§ 粒子群优化	Sigmoid active function	※ 9 个基准分类数据集 ※ 1 个基准回归数据集	分类 回归
μ G-ELM ^[62]	§ 遗传算法 § 多目标优化	Sigmoid active function	※ 8 个基准回归数据集 ※ 11 个基准分类数据集	分类 回归
M-ELM ^[63]	§ 文化基因算法	Sigmoid active function	※ 46 个 UCI 基准数据集	分类
I-ELM ^[64]	§ 增量构造	Sigmoid active function Gaussian function Fourier function	※ 人工数据“SinC” ※ 8 个 UCI 基准数据集	回归
CI-ELM ^[10]	§ 增量构造 § 凸优化	Sigmoid active function Gaussian function	※ 8 个基准回归数据集	回归
EM-ELM ^[67]	§ 增量构造	Sigmoid active function	※ 人工数据“SinC” ※ 4 个回归基准数据集 ※ 3 个分类基准数据集	分类 回归
CS-ELM ^[68]	§ 增量构造 § 无偏风险估计准则	Sigmoid active function	※ 10 个回归基准数据集	回归
B-ELM ^[69]	§ 增量构造 § 误差反向传播	Sigmoid active function Fourier function	※ 9 个 UCI 基准数据集	分类 回归
QRI-ELM ^[70]	§ 增量构造 § QR 分解	Sigmoid active function	※ 人工数据“SinC” ※ 3 个分类基准数据集	分类 回归
Inverse-Free ELM ^[71]	§ 增量构造 § 矩阵免逆	Sigmoid active function Fourier function Hardlimit function Gaussian function	※ 5 个分类基准数据集 ※ 3 个回归基准数据集	分类 回归

3.3 剪枝构造法

剪枝法和增量构造法正好相反, 它首先建立一个含有较多隐层神经元的网络, 在此基础上, 通过某种准则来衡量每个隐层神经元对模型性能的贡献度, 并逐步删除不相关的或冗余的隐层神经元, 最终实现网络结构的简化. Rong 等人^[73]对 ELM 分类器的结构设计问题进行了研究, 提出了一种剪枝 P-ELM(Pruned ELM)算法. 该算法从一个拥有大量随机隐层神经元的网络开始训练, 利用 Chi-squared (χ^2) 或者信息增益(Information Gain, IG)来计算期望输出和各隐层神经元之间的相关度并按降序排列; 然后将每个神经元的相关度和用户事先定义好的一组相关度阈值进行比较产生多个候选神经元子集, 确定每个阈值对应的候选子集的规模并计算这些子集在验证集上的分类准确率; 最后采用 AIC (Akaike Information Criterion) 准则对每个候选子集进行评估, 根据最小的 AIC 值来确定最终的隐层神经元个数. 需要指出的是, 该算法对相关度阈值的取值十分敏感, 因为阈值选取正确与否直接关系到

网络的紧凑程度, 进而影响模型的泛化性能. 当训练集中存在噪声或高度相关变量时, ELM 的性能会受到影响, 为了解决这一问题, Miche 等人^[74]提出了一种能够同时处理分类和回归问题的最优剪枝 OP-ELM (Optimally Pruned ELM) 算法^①. 其具体实现过程分为三步: 首先使用标准 ELM 算法构建一个大规模单隐层前馈神经网络; 然后利用多响应稀疏回归 (Multi-response Sparse Regression, MRSR) 方法对各隐层神经元的重要性进行排序以提高网络的收敛速度; 最后采用留一 (Leave-One-Out, LOO) 交叉验证法来确定需要保留的隐层神经元数. 在 OP-ELM 中, 隐层和输出层之间是简单的线性关系, 因此采用预测平方和 (PRediction Sum of Squares, PRESS) 作为 LOO 的误差计算公式以减少算法的计算复杂度. 在求解伪逆运算的过程中, 当隐层输出矩阵奇异或近似奇异时会导致数值计算不稳定, 从而影响模型的准确性. 针对这一问题, 文献^[75]对

① OP-ELM Source Code. <https://research.cs.aalto.fi/aml/>

OP-ELM 进行了改进,提出了一种基于最小角回归 LARS(Least Angle Regression)和 Tikhonov 的双正则 TROP-ELM(Tikhonov-regularized OP-ELM)算法.该算法和 OP-ELM 的主要区别是,在第二步中通过 LARS 算法对输出权值施加 L_1 惩罚并对各隐层神经元进行排序;在第三步中对原始 PRESS 公式及其计算方法进行改进,在计算矩阵伪逆的过程中加入了 Tikhonov 正则化参数并利用 Nelder-Mead 方法进行参数优化来加速收敛.虽然 TROP-ELM 利用正则化提高了模型的稳定性,但是由于在训练过程中增加了很多超参数需要优化,降低了计算效率. Grigorievskiy 等人^[76]将 OP-ELM 模型应用于长时间序列预测问题,实验结果表明,OP-ELM 的计算时间比线性最小二乘法慢很多,因此仅适合在变量较少的情况下使用.针对回归问题,文献[77]考虑了训练数据的先验知识,采用 Lasso、岭回归和弹性网络来剪裁 ELM 隐层中的冗余神经元.研究表明,通过 Lasso 和弹性网络对输出权值向量的大小进行约束可以得到更紧凑的网络结构,同时也能保持良好的泛化性能.文献[78]提出了一种基于 OP-ELM 的多层感知机结构选择算法 ASELM(Architecture Selection ELM),该算法利用 OP-ELM 来训练多层感知器神经网络,从而删除不相关的输入特征和冗余的隐层神经元.文献[79]提出了一种用于处理回归和多元分类问题的 1-范数 ELM,该算法在标准 ELM 损失函数的基础上增加了一个控制目标函数光滑程度的 L_1 惩罚项,从而可以产生一些严格等于零的输出权值系数,进而实现网络结构的稀疏化.传统的 ELM 算法需要大量的标记实例才能实现准确的分类,然而在实际应用中标记样本往往难以获得,针对少量标记样本的分类问题, Luo 等人^[80]将联合稀疏正则化方法引入到 ELM 中,提出了一种稀疏半监督剪枝 S3ELM(Sparse Semi-supervised ELM)算法.前面所提到的 ELM 网络结构优化方法大多都是贪婪式的.虽然可以通过扩大搜寻范围等手段减轻贪心算法产生的不利影响,但它们仍然容易陷入局部最优解. S3ELM 利用 $L_{2,1}$ 范数作为正则项来约束隐含层和输出层之间的权值矩阵 β ,这样目标函数的求解就成为一个凸规划问题,进而能够得到全局最优解.此外, S3ELM 通过引入了图拉普拉斯正则项扩展到半监督学习,充分利用标记和未标记样本的信息,在少量标记样本可用的情况下提高分类器的准确率.但是由于该算法中可调参数太多,在训练阶段计算复杂度非常高.为了进一步提高 ELM 隐层的映射能力, Zhou 等人^[81]提出

了一种基于随机傅里叶映射的 $\ell_{2,1}$ RF-ELM($\ell_{2,1}$ -norm-based Random Fourier ELM)算法.该算法首先利用随机傅里叶映射作为激活函数来逼近高斯核,然后将行稀疏诱导 $L_{2,1}$ 正则项加入到 ELM 的目标函数中,从而剔除无关的冗余神经元.

3.4 自适应构造法

在隐层神经元选取的过程中,已存在的神经元可能会对后续神经元的选择产生影响,而新神经元的加入又可能导致之前选择的神经元失效,从而变成冗余神经元,即所谓的“嵌套效应”^[82],因此增量和剪枝这两种构造方法通常都不是最理想的.自适应构造法吸收了增量构造法和剪枝法的优点,既可以向隐含层添加新的神经元,又可以删除其中的冗余神经元. Lan 等人^[83]提出了一种用于回归问题的两阶段 TS-ELM(Two-Stage ELM)算法,该算法通过前向选择和后向移除这两个阶段来完成网络结构的优化.在前向选择阶段,每次迭代中随机生成多组候选隐层神经元,然后递归计算它们对损失函数下降的贡献度以衡量各组候选神经元的重要性,选择具有最高贡献度的隐层神经元组加入到网络中并递归更新网络的输出权值,直到最终预测误差(Final Prediction Error, FPE)停止准则达到最小值时结束.后向移除阶段包括隐层输出矩阵重构和隐层神经元删除两步,这个阶段主要是对前一阶段选择的隐层神经元进行检查,从而淘汰冗余神经元. Zhang 等人^[84]提出了一种隐层神经元能够自适应增长的 AG-ELM(ELM with Adaptive Growth of Hidden Nodes)算法.该算法首先初始化一个含有少量隐层神经元的网络,然后在每次迭代中随机产生一组隐层神经元数不同的网络放入候选池,并从中选出泛化性能最优的网络.这样做的优点是可以由新生成的网络来代替现有的网络,在每一步中,可以添加一个或多个隐层神经元,并且可以同时删除一个或多个现有的隐层神经元,从而实现隐含层神经元的自适应调整.由于每次迭代都要训练多个候选网络, AG-ELM 的时间复杂度要比 I-ELM 高.此后, Zhang 等人^[85]对 AG-ELM 进行了改进,提出了一种隐层神经元动态调整的 D-ELM(Dynamic ELM)算法,从理论上进一步分析了 ELM 的通用逼近能力.该算法从仅含一个隐层神经元的网络开始训练,采用和 AG-ELM 相同的候选网络构建策略,每次迭代都会产生一组网络. D-ELM 每次从候选池中寻找一个隐层神经元最少并且泛化性能最好的网络,并使用 EM-ELM 中提出的误差最小化的方法来更新网络的输出权值,以此来提高模型的性能.研究结

果表明,大多数情况下,在 D-ELM 中使用 Sigmoid 激活函数比 RBF 激活函数表现要好. Pratama 等人^[86]围绕网络结构复杂性、数据流的不确定性、概念漂移和高维度这四个问题对 ELM 展开研究,提出了一种基于元认知理论的在线 eT2ELM(Evolving Type-2 ELM)算法.该算法融合了元认知学习的三个重要组成部分:学什么、怎样学、何时学,所有这些都遵循严格的序列学习原则.它利用在线主动学习的方法管理学什么模块,能有效处理概念漂移问题,并将提取的数据流提供给如何学模块,该模块包括隐层神经元增长机制,隐层神经元剪枝策略,基于马尔可夫覆盖的在线特征选择算法和参数学习机制这四部分.在隐层神经元增长的过程中监测数据流的焦点以应对快速变化的学习环境,随机生成由高斯函数的协方差矩阵构成的影响区域,使用相对互信息(Type-2 Relative Mutual Information, T2RMI)方法量化隐层神经元和目标变量之间的相关性,从而捕获冗余的神经元并删除. Sun 等人^[87]经过研究发现,在 CI-ELM 中,可以用网络输出权值的范数值来衡量隐层神经元的重要性,提出了一种基于 CI-ELM 的 DCI-ELM(Dynamic Convex Incremental ELM)隐层神经元选择算法. DCI-ELM 将神经元的剪枝和增长融合到一个过程中,交替完成隐层结构的调整. Zhao 等人^[88]基于 QRI-ELM 提出了两种改进的算法: GSI-ELM(Gram-Schmidt Process Based Incremental ELM)和 IGSI-ELM(Improved GSI-ELM). GSI-ELM 引入了一个概率评估准则,只考虑一个固定大小的候选神经元子集,并从该集合中

选择最佳的隐层神经元加入到网络中. IGSI-ELM 在 GSI-ELM 的基础上增加了一个剪枝策略,该算法在每次添加新隐层神经元后,都要根据事先定义的评估准则对所有隐层神经元按重要性进行排序,并挑选出重要性最差的隐层神经元.自适应构造方法虽然能够获得更加紧凑的网络结构,但是由于增加了多层判断机制,使得算法整体趋于复杂.

除了上述几类 ELM 网络结构优化方法外, Feng 等人^[89]还提出了一种 ELM 隐藏层节点参数动态调整算法 DA-ELM(Dynamic Adjustment ELM).在 DA-ELM 中,冗余隐节点的输入参数在每个步骤中沿能量误差递减的方向更新.虽然 DA-ELM 能够获得更紧凑的网络结构,提高了 SLFN 的泛化能力,但是由于在训练过程中需要对每个隐节点进行排序,增加了算法的时间复杂度.在处理高维大数据任务时,随机隐节点往往无法获得输入数据的有效表征和抽象,导致 ELM 分类器的泛化能力和鲁棒性急剧下降,学习效果不佳.针对这一问题, Deng 等人^[14]提出了一种快速的基于奇异值分解隐节点的 FSVD-H-ELM 算法(Fast Singular Value Decomposition-Hidden-nodes Based ELM).在 FSVD-H-ELM 中,隐节点参数不再随机生成,而是在原始数据集采样的多个随机数据子集上利用 SVD 产生 ELM 的输入权值和隐层偏置.在 12 个高维大数据集上的实验结果表明,和传统的 SVM 等方法相比, FSVD-H-ELM 具有较快的学习速度、良好的泛化性能和鲁棒性.表 5 对几种典型的极限学习机剪枝构造法、自适应构造法进行了总结和归纳.

表 5 极限学习剪枝和自适应构造模型

模型	核心思想	特征映射	评测数据集	应用
P-ELM ^[73]	§ 剪枝构造 § 隐神经元相关性度量	Sigmoid active function Gaussian function	※ 8 个 UCI 基准数据集	分类
OP-ELM ^[74]	§ 剪枝构造 § 多响应稀疏回归 § 留一交叉验证	Sigmoid active function Gaussian function Linearkernel	※ 4 个分类基准数据集 ※ 11 个回归基准数据集	分类 回归
TROP-ELM ^[75]	§ Tikhonov 正则化 § 最小角回归	Sigmoid active function Gaussian function Linearkernel	※ 11 个 UCI 基准数据集	回归
ASELM ^[78]	§ 多层感知器剪枝	Sigmoid active function	※ 14 个分类基准数据集	分类
NLPELM ^[79]	§ 1 范数正则化 § Newton-Armijo	Sigmoid active function Gaussian function Multiquadrics function	※ 16 个分类基准数据集 ※ 26 个回归基准数据集	分类 回归
S3ELM ^[80]	§ $L_{2,1}$ 正则化 § 图拉普拉斯 § 联合稀疏正则化	Sigmoid active function Gaussian function	※ 3 个人工数据集 ※ 4 个 UCI 基准数据集 ※ 2 个人脸数据集	分类
$\ell_{2,1}$ RF-ELM ^[81]	§ $L_{2,1}$ 正则化 § 随机傅里叶映射	Random Fourier	※ 22 个分类基准数据集 ※ 8 个回归基准数据集	分类 回归
TS-ELM ^[83]	§ 两阶段构造	Sigmoid active function	※ 6 个 UCI 基准数据集	回归
AG-ELM ^[84]	§ 自适应构造	Sigmoid active function	※ 7 个 UCI 基准数据集	回归

(续 表)

模型	核心思想	特征映射	评测数据集	应用
D-ELM ^[85]	§ 动态学习	Sigmoid active function Gaussian function	※10 个 UCI 基准数据集	回归
eT2ELM ^[86]	§ 元认知学习	Gaussian function	※12 个概念漂移数据集	分类
DCI-ELM ^[87]	§ 动态调整 § 凸优化	Sigmoid active function Gaussian function	※9 个回归基准数据集	回归
IGSI-ELM ^[88]	§ Gram-Schmid 正交化	Sigmoid active function	※13 个分类基准数据集 ※13 个回归基准数据集	分类 回归
DA-ELM ^[89]	§ 隐藏层节点排序 § 误差最小化近似	Sigmoid active function	※人工数据“SinC” ※4 个回归基准数据集 ※3 个分类基准数据集	分类 回归
FSVD-H-ELM ^[14]	§ 奇异值分解	Sigmoid active function	※12 个大规模基准数据集	分类 回归

4 ELM 融合与加速

随着日常生产和生活中数据量的迅猛增长,存储和处理大规模数据已成为企业的迫切需求,ELM 的集成与并行也因此成为另外一个研究热点.受益于单隐层前馈神经网络的简单结构和隐层神经元参数的随机取值,标准的 ELM 具有算法复杂度低和容易实现等优点,因此被广泛地应用于集成学习中,以进一步提高模型的泛化性能.此外,还可以借助矢量编程方式实现网络输出矩阵的并行计算,使得模型的训练加速.这样在使用海量训练数据的场景下,所耗费的时间将大幅缩短.本节将对基于集成和并行的 ELM 算法进行详细地阐述、分析和归纳.

4.1 集成极限学习机

集成学习的核心是构建多个不同的模型,并将这些模型聚合起来从而有效地提高算法的泛化性能.根据基学习器生成策略的不同,集成学习的方法可以分为两类:以 Bagging 为主要代表的并行方法和以 Boosting 为主要代表的顺序方法.围绕这两种方法,大量的集成 ELM 算法被提出来.针对非平稳时间序列的预测问题, Van Heeswijk 等人^[90]提出了一种自适应集成 ELM 算法.该集成模型由多个随机初始化的 ELM 组成,每个 ELM 都有一个相关的权重,作为误差的函数决定了该模型对整体预测的贡献度.在对下个时间步长进行预测之前,使用滑动窗口或增长窗口内的样本重新训练每个 ELM 子学习机,最后采用加权线性组合的方法将各子学习机的预测结果进行整合.针对模式分类问题, Liu 等人^[91]提出了一种基于集成的 EN-ELM (Ensemble Based ELM) 算法.该算法将集成学习方法和交叉验证策略引入到训练过程中,以减轻过拟合问题并提高模型预测的稳定性.在 EN-ELM 中,首先将整个训练集划分为多个不同的子集,然后在每个子集上

训练不同的 ELM 弱分类器,最终的分类结果通过投票法产生.针对回归问题, Tian 等人^[92]提出了一种基于改进 AdaBoost. RT 的 ELM 集成算法.传统 AdaBoost. RT 算法的性能对初始阈值的取值十分敏感,通常需要经过反复试验才能确定,增加了计算复杂度.改进后的 AdaBoost. RT 在每次迭代学习过程中可以根据训练误差自适应调整阈值大小,从而保证始终朝着性能最优的方向搜索最佳阈值,有效克服了阈值选择困难的问题. Cao 等人^[93]提出了一种基于投票法的集成分类 V-ELM (Voting Based ELM) 算法, V-ELM 在同一个训练集上训练多个具有相同隐层神经元数和激活函数的 ELM, 并利用集成学习的方法求得每个训练样本的后验概率, 最终的类别由多数投票法决定. 该算法能够在一定程度上解决单个 ELM 分类器的不稳定性, 并提高 ELM 的泛化性能. Xue 等人^[94]提出了一种基于遗传算法的选择性集成 GE-ELM (Genetic Ensemble of ELM) 算法. GE-ELM 采用遗传算法生成一组候选 ELM 网络, 再根据适应度值和权值范数 $\|\beta\|$ 的大小对每个候选网络进行排序, 并选出最优的几个网络进行集成, 从而提高单个 ELM 网络的鲁棒性和泛化能力. Wang 等人^[95]提出了一种基于最小最大模块化网络的并行 M³-ELM (ELM Ensemble Based on The Min-Max Modular Network) 集成算法, 以更好地解决大数据问题. 该算法首先将原始分类问题分解成较小的子问题, 这些子问题的规模都是可管理的, 因此它们可以很好的适应并行计算环境中的存储器容量和计算神经元的数量. 此外, 通过任务分解可以将不平衡问题转化为平衡问题. 然后针对每个子问题构建相应的 ELM 个体学习器, 最后利用最小最大模块化网络集成训练好的 ELM. Samat 等人^[96]将 ELM 引入到高光谱图像分类问题中, 结合集成学习提出了两种分别基于 Bagging 和 AdaBoost 的 E²LMs (Ensemble ELMs)

集成算法. Bagging 算法非常适用于 ELM 这类不稳定的学习算法,通过 Bootstrap 重采样来获得训练集的不同子集,利用这些子集构造多个相互独立的 ELM 基学习器,可以有效减少模型的方差. AdaBoost 算法首先构造一个 ELM 弱分类器,根据该分类器在训练集上的错分类样本,对训练集中每个样本赋予不同的权重.在构建新的 ELM 分类器时,优先考虑之前被错分的样本,通过重复该过程,构建多个 ELM 弱分类器,最后利用多数投票法得到最终的分类结果.基于数据分区学习的思想,Yang 等人^[97]将 ELM 学习方法从单个神经网络扩展到多神经网络系统,并从理论上证明了多网络 ELM 学习方法不仅可以逼近任意连续目标函数,还能够划分具有任意形状的不相交区域.针对区间估计问题,Ning 等人^[98]通过整合两种不同的非对称 ELM,提出了孪生 TELM(Twin ELM)算法.首先提出了一种正则化的非对称最小二乘 RALS-ELM(Regularized Asymmetric Least Squares ELM)算法,根据模型输出误差的正负来设置带有不同权重的平方损失函数,以使上述误差可以在参数学习过程中被区分开,并引入 Tikhonov 正则化来减少过拟合;然后提出了一种基于贝叶斯框架和非对称高斯分布的 AB-ELM(Asymmetric Bayesian ELM)算法,并利用最大似然算法来学习 AB-ELM 的参数;最后通过一对权重对 RALS-ELM 和 AB-ELM 进行整合来预测区间的上下界.虽然利用集成学习的方法构建合并多个 ELM 基学习器,通常可以获得比单个 ELM 更显著的泛化性能,但由于需要训练一组不同的基学习器,增加了算法的时间复杂度.此外,在集成学习中如何确定合适数量的基学习器也是一个需要解决的难题.

4.2 并行极限学习机

目前处理大数据问题的一个主要途径就是分而治之,即将大数据问题分解成规模较小的子问题来解决.如果子问题之间是相互独立的,就可以并行处理这些子问题,然后将各个子问题的结果合并,从而得到最终的解.为了高效地处理大规模数据的回归任务, Van Heeswijk 等人^[99]提出了一种基于 GPU(Graphics Processing Unit)加速的并行 ELM 集成算法.该算法使用集成学习的方法训练多个 ELM 子学习机,并采用留一交叉验证法调整每个 ELM 子学习机的隐层神经元个数.为了提高网络的训练速度以及减少模型选择的时间,首先利用 GPU 对单个 ELM 进行加速训练,然后通过多个 GUP 和多核 CPU 共同协作完成 ELM 子学习机的并行化训

练,从而大幅缩短大规模任务的计算时间.针对大数据挖掘中的回归问题,He 等人^[100]通过分析 ELM 算法的机制,基于 MapReduce 框架设计并实现了一种高效的 PELM(Parallel ELM)并行回归模型.在 PELM 中,利用 MapReduce 分布式并行编程模型定义 Map 函数和 Reduce 函数来实现特征映射矩阵 $\mathbf{H}$ 和广义逆矩阵的并行化处理. PELM 仅考虑了样本数大于隐层神经元个数的情况,并且在两个 MapReduce 作业的执行期间,会产生大量临时的结果需要转换,增加了算法的计算时间.针对大规模数据流学习问题,Wang 等人^[101]通过分析在线学习阶段 OS-ELM 中各矩阵之间更新的依赖关系,提出了一种基于 MapReduce 的并行在线序列 POS-ELM(Parallel Online Sequential ELM)算法.根据 OS-ELM 算法易知,隐层输出矩阵 $\mathbf{H}$ 和输出权值矩阵 $\boldsymbol{\beta}$ 的在线更新过程是线性的.事实上,对于每一批新加入的样本或样本块,当前样本的输出权值 $\boldsymbol{\beta}_{k+1}$ 的计算依赖于前一批样本的输出权值 $\boldsymbol{\beta}_k$,而每一批样本的隐层输出矩阵之间的更新则相互独立,因此可以先对矩阵 $\mathbf{H}$ 进行并行化处理,然后再对 $\boldsymbol{\beta}$ 进行串行更新,以缩短模型在线训练的时间.

为了一步提高 ELM 在处理大规模学习任务中的计算效率减少内存消耗,Wang 等人^[102]提出了两种并行的正则化 ELM 算法,即数据并行正则化 DPR-ELM(Data Parallel Regularized ELM)和模型并行正则化 MPR-ELM(Model Parallel Regularized ELM),统称为 PR-ELM(Parallel Regularized ELM). DPR-ELM 和 MPR-ELM 均在具有消息传递接口(MPI)环境的网络集群上实现 ELM 的并行化,不同的是 DPR-ELM 完成了数据并行的 ELM 网络训练,而 MPR-ELM 则是充分利用模型并行.针对大数据分类问题,Duan 等人^[103]提出了一种基于 Spark 框架的 SELM(ELM Based on Spark)算法. SELM 由三个高效的并行子算法组成,这些算法充分利用了 Spark 的一系列优良特性,包括高容错性、持久/缓存策略、用户控制分区策略等,来实现矩阵的加速计算和分解,从而极大地加快了 ELM 训练大数据的计算过程.使用医疗大数据集和手写字体数据集对 SELM 的性能进行评测,实验结果表明, SELM 在 10 个节点、15 个节点、20 个节点、25 个节点、30 个节点和 35 个节点的集群上分别实现了 8.71 倍、13.79 倍、18.74 倍、23.79 倍、28.89 倍和 33.81 倍的加速.在相同参数条件下, SELM 具有和传统 ELM 相当分类精度.图 3 给出了 SELM 处理大数据分类过程的详细步骤.

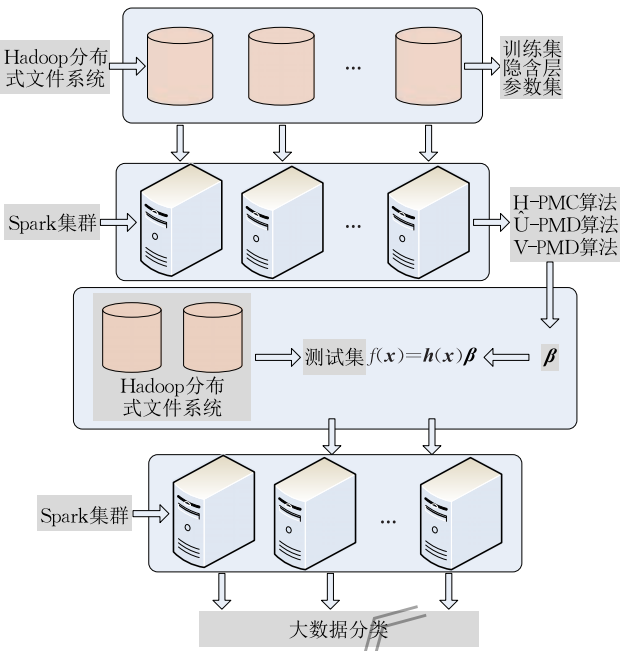


图 3 基于 Spark 框架的 SELM 大数据分类流程^[103]

Lam 等人^[104]利用 GPU 高效并行计算算法的能力,提出了一种基于 CUDA 的 ELM 算法来缩短 ELM 中 RBF 核函数的计算时间.原始的 SS-ELM 算法^[20]在训练时需要将数据全部读入内存,因此无法处理大规模数据集.针对这一问题,Chen 等人^[105]对 SS-ELM 算法进行了深入的分析,提出了一种基于 MapReduce 编程模型的并行半监督 PSS-ELM (Parallel SS-ELM)算法,该算法可以充分利用云计算平台提供的灵活性.文献[106]提出了一高度可扩展的 ELM 工具箱来处理大数据应用,该工具箱实现了基于 GPU 加速训练的 ELM 算法.

本节综述了几种主要的集成和并行 ELM 算法.一方面,利用集成学习的方法,能够有效提高单个 ELM 的泛化性能;另一方面,借助分布式并行计算框架和 GPU,使得 ELM 具备高效的大规模数据处理能力.表 6 对部分集成和并行极限学习机算法进行了归纳,并对它们的核心思想、特征映射、评测数据集和应用等方面进行了总结.

表 6 集成和并行极限学习机模型

模型	核心思想	特征映射	评测数据集	应用
EN-ELM ^[91]	§ 训练集划分 § 交叉验证	Sigmoid active function	※3 个人脸数据库 ※3 个 UCI 基准数据集	分类
Adaboost ^[92]	§ AdaBoost 集成	Sigmoid active function	※LF 钢水温数据	回归
V-ELM ^[93]	§ 多数投票法	Sigmoid active function Gaussian function	※19 个基准数据集	分类
GE-ELM ^[94]	§ 选择集成 § 遗传算法	Sigmoid active function	※5 个基准分类数据集 ※4 个基准回归数据集	分类 回归
M ³ -ELM ^[95]	§ 最小最大模块网络	Sigmoid active function	※12 个大规模数据集	分类
TELM ^[96]	§ 非对称损失函数 § 贝叶斯回归 § Tikhonov 正则化	Linear kernel Gaussian kernel	※人工数据“SinC” ※3 个 UCI 基准数据集 ※2 个工业数据集	回归
PPLM ^[97]	§ 数据分区学习	Sigmoid active function	※19 个大规模数据集	分类 回归
文献 ^[99]	§ GUP 加速计算 § 并行集成模型	Sigmoid active function	※2 个大规模数据集	回归
PELM ^[100]	§ MapReduce	Sigmoid active function	※5 个 UCI 数据集	回归
POS-ELM ^[101]	§ MapReduce § 在线学习	Sigmoid active function	※6 个基准数据集	分类
PR-ELM ^[102]	§ 并行 § 集群	Sigmoid active function	※11 个基准数据集	分类
SELM ^[103]	§ Spark 并行框架 § 矩阵分解	Sigmoid active function	※9 个合成医疗大数据集 ※MNIST 数据集	分类
UFL-RBF-ELM ^[104]	§ GPU § CUDA kernel	Gaussian kernel	※CIFAR-10 图像数据集 ※MNIST 数据集	分类
PSS-ELM ^[105]	§ MapReduce § 局部敏感哈希	Gaussian function	※5 个半监督基准数据集	分类

5 ELM 深度学习模型

ELM 最初是作为单隐层前馈神经网络学习算

法提出的,和深度学习(Deep Learning)算法相比,其特征表示能力有限.最近的一些研究表明,ELM 也可以拓展到深度神经网络的学习中,以减少深层网络的训练时间和计算复杂度.如图 4 所示,深度极

限学习机将特征学习和模型选择统一到一个框架下,其训练过程在结构上通常可以分成两个阶段:特征提取和模型学习. 第一阶段,利用 ELM 进行逐层无监督学习或监督学习,通过非线性变换实现对上一层的特征抽象,经过逐层的特征组合最终从原始输入数据中学习到更加抽象的高层特征,减少了人工干预的成本;第二阶段,ELM 或者其变体被用于对提取后的特征进行建模来完成分类、回归或聚类等学习任务.

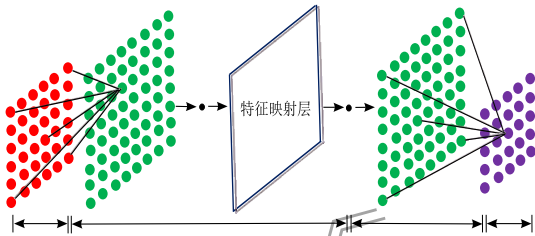


图 4 多隐层 ELM 学习框架

随着数据的规模和复杂程度不断增加,传统的人工特征设计已远远无法满足实际应用的需求. 针对这一问题,Kasun 等人^[107]将自动编码器和 ELM 相结合,提出了一种基于 ELM 自动编码器的多隐层 ML-ELM(Multi Layer ELM)模型. 该模型由多个 ELM 自编码器(ELM Auto Encoder, ELM-AE)堆叠而成,利用每个 ELM-AE 的输出权值矩阵对应的奇异值来表示输入数据的特征,前一层自编码器的隐层输出作为后一层自编码器的输入,不需要专门的领域知识,避免了繁杂、耗时的人工特征选取过程. ELM-AE 是一种无监督的单隐层前馈神经网络学习算法,自编码的本质是在训练过程中尽可能使网络的输入数据和目标输出保持一致. 在 ELM-AE 中,网络的输入权值和隐层偏置随机选取,并对它们进行正交化处理,在学习过程中保持不变. 使用非线性激活函数将输入数据映射到 ELM 特征空间中,而输出权值则是通过最小化平方损失函数,由最小二乘法求得. 根据输入数据的维度和隐层神经元个数的不同,ELM-AE 可以实现三种不同类型的特征表示:压缩表示、稀疏表示和等维表示. 与传统的基于逐层贪婪训练方法的深度学习算法相比,ML-ELM 采用 ELM 的思想训练每个隐含层,无需对各隐层的参数进行迭代微调,降低了模型的计算复杂度. 表 7 给出了 ML-ELM 模型和不同深度学习模型在 MNIST 数据集上的测试结果.

表 7 MNIST 数据集上的测试结果^[107]

模型	测试准确率	训练时间
ML-ELM	99.03%	444.655 s
Deep Belief Network (DBN)	98.87%	20 580 s
Deep Boltzmann Machine (DBM)	99.05%	68 246 s
Stacked Auto-Encoder (SAE)	98.60%	>17 h
Stacked Denoising Auto-Encoder (SDAE)	98.72%	>17 h

传统 ELM 的网络结构几乎都是全连接型的,即隐含层的每个神经元都要和输入层的所有神经元相连接,这种全连接方式虽然比较容易实现,并在很多应用中取得了不错的表现,但是由于层与层之间皆为全连接,导致需要学习的参数过多. 视觉系统的工作原理表明,局部感知野 ELM-LRF(Local Receptive Fields Based ELM)有助于生物学习系统识别输入图像的局部相关信息,受此启发 Huang 等人^[108]提出了一种基于局部感知野的算法. 在 ELM-LRF 中,每个隐含层神经元和输出层神经元之间的连接根据某种连续概率分布随机生成,这些分布会使局部相邻的输入神经元联系较为紧密,而距离较远的神经元相关性较弱,这样就实现了局部连接或稀疏连接,不但可以提取局部特征,还可以通过组合隐含层神经元进一步获取对旋转、平移和缩放不变的观测数据的显著特征. 另外,为了获得更完备的特征集合,还需要对随机输入权值进行正交化处理. 和卷积神经网络(Convolutional Neural Network, CNN)相比,ELM-LRF 提供更灵活、更广泛的局部感知野,它们可以根据不同类型的概率分布随机采样得到,而 CNN 只使用卷积隐层神经元. 此外, CNN 通常采用 BP 算法逐层实现对网络参数的调优,存在收敛速度慢和局部极小值问题. 而 ELM-LRF 随机产生输入权值,并解析求解输出权值,具有更高的计算效率和稳定性. 从网络结构的角度来看,尽管使用了一个组合隐含层,ELM-LRF 本质上仍属于浅层网络,如何构建基于 ELM-LRF 的深度堆叠网络,在提取更抽象的高层特征同时,而不丢失重要的数据信息,还有待进一步研究. 当面对极其庞大而复杂的训练数据集时,为了保证一定的分类或预测精度,ELM 需要大量的隐层神经元来逼近目标函数,从而导致内存消耗严重且计算成本非常高. 针对这一问题,Zhou 等人^[109]提出了一种专为解决复杂大规模数据问题而设计的堆栈 S-ELM(Stacked ELM)算法. 该算法将一个大型 ELM 网络分解成多个串联连接的 ELM 子网络,而每个 ELM 只包含少量的隐层神经元,这对于解决大规模复杂数据问题特别有效,因为单个 ELM 在隐层神经元数量非常大时通常难以计算.

S-ELM 利用主成份分析法来减少每层中的隐层神经元个数,具体来说就是根据输出权重的特征值选择最重要的几个神经元来代表每个隐层中的所有神经元输出到下一层并与该层中的新随机隐神经元合并,通过这种方式,可以保持整个网络的大小不变,但能够在更高维度的 ELM 特征空间中学习数据. S-ELM 的层数由用户自己决定,使其满足给定的机器学习任务. 对于大规模非结构化数据集,可以在 S-ELM 的每次迭代中加入一个 ELM 自编码器作为无监督特征提取步骤,来提高模型的泛化性能.

目前大多数深度学习方法都会遇到难以解决的非凸优化问题,Yu 等人^[110]将堆叠泛化原理引入到深度表示学习过程中,提出了一种 DrELM(Deep Representations Learning via ELM)堆叠模型,用于逐层提取输入数据的高层抽象特征,以期能够获得更有意义的表示. 传统的基于浅层学习理论的 ELM 算法特征学习能力不足,即使设置大量的隐层神经元,也无法有效获得图像和视频等复杂信号的高层抽象特征. 针对这一问题,Tang 等人^[8]提出了一种 H-ELM(Hierarchical ELM)多层感知器分层学习框架. 该框架主要由两部分组成:无监督多层前馈编码和监督特征分类. 在特征提取阶段,首先将原始数据映射到 ELM 随机特征空间,然后在 ELM-AE 的基础上引入 L_1 正则化构建 ELM 稀疏自编码器,进行逐层无监督特征学习以获得输入数据的多层稀疏表示. H-ELM 中的每个隐含层作为特征提取器都是一个独立的模块,并以前馈方式逐层进行训练,一旦上一层特征提取完成后,当前层的权值或参数也随之确定下来,不再进行微调. 这和传统的深度学习模型完全不同,其隐含层参数逐层初始化,并通过 BP 算法进行调优. 因此,H-ELM 比基于贪婪训练方法的堆栈式自动编码器、深度置信网络以及深度玻尔兹曼机具有更高的学习效率. 在特征分类阶段,对学习到的高层稀疏特征施加随机扰动以保证 ELM 的通用逼近能力,最后利用原始 ELM 进行分类决策. ML-ELM 的学习性能严重依赖于每层中隐层神经元个数的选取,此外在自编码阶段随机生成隐层神经元参数会导致模型的泛化性能不稳定. 针对上述问题,Wong 等人^[111]将核方法引入到 ML-ELM 中,提出了一种多层核 ML-KELM(MultiLayer Kernel ELM)模型^①. ML-KELM 由多个核 ELM 自编码器(KELM Auto Encoder,KELM-AE)堆叠而成,每个 KELM-AE 利用隐式的核映射代替显式的随机映射来学习输入数据的特征,继承了核 ELM 的优点,在训练过程中只要确定核参数而不需要调整每一层的

隐层神经元个数和隐层神经元参数. 因此,ML-KELM 比 ML-ELM 和 H-ELM 具有更少的人工干预,更易于特征学习. Yang 等人^[112]对具有广义隐层神经元的多层感知器网络结构进行了研究,证明了其不仅具有通用逼近能力,而且还具有表示学习的能力,并提出了一种 MLRL(Multilayer Representation Learning)表示学习算法. 在 MLRL 中,隐层神经元本身可以是一个由几个隐层神经元组成的子网络,称之为广义隐层神经元. 从这个意义上来说,一个特征映射层可以包含多个子网络. 隐层中的广义隐层神经元的个数和输出层神经元的维数是相互独立的,但要求每个广义隐层神经元中的隐层神经元个数要等于输出层的神经元个数. 与 DBN、SAE 和 ML-ELM 等基于 NN 的表示学习方法不同,其中参数的组合对最终结果的稳定性非常敏感,MLRL 在初始阶段可以随机选择参数,而不会影响到模型的泛化性能. MLRL 仅使用了一种形式的输入,因此如何将感受野的概念应用到 MLRL 中以进行局部感知并构建更复杂的深层堆叠网络还有待进一步研究. 此外,MLRL 要求激活函数存在反函数,使其应用范围受到了一定的限制.

Wang 等人^[113]将自动编码器和卷积神经网络结合到 ELM 理论框架中,提出了一种卷积自编码 CAE-ELM(Convolutional Auto Encoder ELM)算法来自动学习三维模型的形状特征. CAE-ELM 的训练过程可以分成三个阶段进行描述:首先随机生成卷积核,对输入数据进行卷积特征映射,计算每一层的特征图,并在卷积核运算后的特征图上执行池化操作以保持旋转不变性;然后将第一阶段生成的卷积特征映射作为 ELM-AE 的输入和输出进行特征学习,获得输入特征的更高阶表示;最后融合所有 ELM-AE 提取的特征作为 ELM 分类器的输入. CAE-ELM 的卷积层采用局部连接和权值共享的网络结构,使得自由参数的数量显著减少,从而降低了计算复杂度. 但 CAE-ELM 也还存在一些不足,例如在卷积阶段生成的每个特征映射都需训练相应的 ELM-AE 进行二次特征提取,在卷积核的数量很大的时候,后续 ELM 自编码器的个数随之变得非常庞大,导致整个网络结构过于复杂,不利于模型的训练. 此外,层数的选择也是一个十分重要的问题,CAE-ELM 可以使用多层卷积网络和多层 ELM 自编码器,如何以最少的隐层数提取足够好的有效特

① <https://www.fst.um.edu.mo/en/staff/fstcmv.html#software>.

征来完成学习任务还亟需进一步研究. ELM-AE 没有考虑数据的内在流形结构,即原空间中彼此邻近的数据在新的投影空间中也应该尽可能接近. 基于这个事实,Sun 等人^[114]借鉴流形学习的思想,对 ELM-AE 进行了改进,提出了一种判别图正则化 GELM-AE(Generalized ELM Auto Encoder)自动编码器. GELM-AE 具有和 ELM-AE 相似的网络结构,不同之处在于,GELM-AE 在目标函数中加入了一个图拉普拉斯正则项来约束输出权值的大小,从而确保彼此靠近的原始数据在输出空间中也能保持局部几何结构. 进一步,仿照 ML-ELM 通过堆叠多个 GELM-AE 自编码器构建 ML-GELM 深度学习模型,在挖掘数据之间的全局结构信息和局部紧邻结构信息的同时,从输入数据中提取更加抽象的高层特征,为模型获取更高的泛化性能奠定了良好的基础. 为了充分挖掘 H-ELM 框架处理大数据的能力,Chen 等人^[115]提出了一种基于 Flink 内存集群计算平台的并行 H-ELM(Parallel H-ELM, PH-ELM)算法. H-ELM 框架由许多隐含层组成,它们与迭代操作相似. Flink 提供的内存并行执行模块可以将数据缓存在集群中的分布式内存中,从而在两层之间节省大量的磁盘 I/O 操作时间,因此非常适合用来实现并行化 H-ELM 框架. 此外,作者通过分析 H-ELM 框架的执行过程,发现一些耗时的

子过程适合在 GPU 中执行,提出了 GPH-ELM(GPU-accelerated Parallel H-ELM)算法. 它利用 GPU 的高速计算能力来加速 PH-ELM,从而进一步提高大数据训练任务的速度. GPH-ELM 同时继承了 Flink 和 GPU 的优良特性,包括容错性、可扩展性和加速性,能够高效的处理大规模数据集. 这种将内存集群计算和 GPU 计算相结合的并行加速框架同样适用于其它机器学习算法. 为了提高人脸图像年龄估计的鲁棒性,Duan 等人^[116]将 CNN 和 ELM 结合在一起,提出了一种集成学习框架 CNN2ELM 来根据人脸图像估计年龄. CNN2ELM 由三个卷积神经网络和两个 ELM 网络堆叠而成,以分层的方式进行年龄估计,其中 CNN 用于提取与年龄、性别和种族相对应的特征,然后进行特征融合并在此基础上训练一个 ELM 分类器和回归器分别用于年龄分组和估计.

基于各种模型和算法,深度神经网络能够从大量的复杂数据中学习到合适且有效的特征,这些特征在解决实际问题时常常能够取得令人满意的结果,从而使得深度学习受到了工业界和学术界的广泛重视,掀起了人工智能的新浪潮,同时也推动 ELM 由浅层神经网络向深层神经网络发展. 本节综述了深度 ELM 网络框架,并对其进行了分析和总结,如表 8 所示.

表 8 深层极限学习机模型

模型	核心思想	特征映射	评测数据集	应用
ML-ELM ^[107]	§ 正交随机特征映射 § 堆叠自编码器	Sigmoid active function	※ MNIST 数据集	分类 表示学习
ELM-LRF ^[108]	§ 局部感知	Sigmoid active function	※ NORB 图像数据集	分类 表示学习
S-ELM ^[109]	§ 主成分分析 § 堆叠自编码器	Sigmoid active function	※ MNIST 和 OCR Letters ※ NORB 和 USPS	分类 表示学习
DrELM ^[110]	§ 堆叠泛化原理	Sigmoid kernel Gaussian kernel	※ 2 个合成数据集 ※ 10 个基准分类数据集 ※ 5 个基准回归数据集	分类 回归 表示学习
H-ELM ^[8]	§ 多层前馈稀疏编码	Sigmoid active function	※ 14 个基准分类数据集 ※ MNIST 和 NORB ※ 3 个计算机视觉任务	分类 表示学习
ML-KELM ^[111]	§ 堆叠核自编码器 § 核学习	Gaussian kernel	※ 20 个基准数据集 ※ Madelon 数据集	分类 表示学习
MLRL ^[112]	§ 子网络隐层神经元	Sigmoid active function	※ 10 个图像数据集 ※ 16 个分类数据集	分类 表示学习
CAE-ELM ^[113]	§ 卷积神经网络 § ELM 自动编码器	Sigmoid active function	※ ModelNet 三维模型库 ※ MNIST 和 NORB	分类 表示学习
ML-GELM ^[114]	§ 流形正则化 § 堆叠自编码器	Sigmoid active function	※ 5 个 UCI 基准数据集 ※ YALEB 和 COIL20 ※ USPST 数据集	聚类 分类 表示学习
PH-ELM ^[115] GPH-ELM ^[115]	§ Apache Flink 集群 § GPU 加速计算	Sigmoid active function	※ MINST 和 GISETTE ※ ADULT 和 WINE	分类 表示学习
CNN2ELM ^[116]	§ 卷积神经网络 § 集成	Gaussian kernel	※ MORPH-II ※ LAP-2016 ※ Adience benchmark	回归 分类 表示学习

6 极限学习机的应用

与传统的神经网络和支持向量机相比,ELM 具有学习速度快,实现简单,人为干预少等显著优点,已被应用于生物医学、计算机视觉、故障诊断等多个领域.在本节中,我们将回顾 ELM 应用的一些最新研究进展.

6.1 ELM 在医疗和生物医学中的应用

目前智能医疗信号处理已逐渐成为临床诊断的重要辅助工具.医学或生物医学数据通常具有高维特征或大量样本.因此,机器学习技术通常用于医学数据分析.准确的预测和快速的处理速度是评价智能医疗处理系统的两个重要指标.近年来,采用 ELM 进行医疗信号处理,由于其算法简单易于实现实时诊断,且性能较好,引起了学术界越来越多的关注.

Zhang 等人^[117]提出一种基于多核 ELM 的运动想象脑电信号(Electroencephalogram, EEG)分类算法 MKELM(Multi-Kernel ELM),利用高斯核和多项式核的线性组合将原始的共空间模式特征映射到不同的非线性空间中,得到更丰富的判别信息来提高 EEG 信号的分类精度.两个公共 EEG 数据集上的实验结果表明,基于 MKELM 的分类精度高于传统的 SVM 方法,为研究复杂非线性信息在专家系统和智能系统中的应用提供了一种有效的解决方案.针对癌症分类中微阵列数据的基因选择和稀疏数据分类问题, Saraswathi 等人^[118]提出了一种基于整数编码遗传算法(Integer-Coded Genetic Algorithm, ICGA)和粒子群优化相结合的极限学习机算法 ICGA-PSO-ELM.该算法首先通过 ICGA 选择最优的一组基因来降低输入特征的维度,然后根据选定的基因构建 ICGA-PSO-ELM 分类器,处理稀疏数据和样本不平衡数据,提高多种类癌症的分类准确率. GCM 数据集上的实验结果表明,该方法识别出的编码分泌蛋白的基因能对每个肿瘤类型的微环境中关键的生物学特性的本质提供重要的参考. Chyzhyk 等人^[2]利用 ELM 设计了一个基于静息态功能磁共振成像的计算机辅助诊断系统,成功地解决了精神分裂症患者的分类问题.文献^[119-121]分别使用标准 ELM、KELM 和集成 ELM 方法实现了三维计算机断层扫描(Computed Tomography, CT)

图像中的肝脏分割和肝脏肿瘤检测.

6.2 ELM 在计算机视觉中的应用

随着计算机视觉和机器学习的飞速发展,图像识别和目标检测等任务已成为 ELM 在高维大数据应用中的又一热点问题.

Huang 等人^[3]提出了一种基于 KELM 的快速有效的交通标志识别方法.首先利用改进的梯度方向直方图(Histogram of Oriented Gradient variant, HOGv)特征,在冗余度和局部细节之间保持良好的平衡;然后使用 KELM 作为分类器进行交通标志识别(Traffic Sign Recognition, TSR).针对触觉目标识别问题, Liu 等人^[122]在稀疏表示和字典学习的基础上提出了一种结合 ELM 和核字典学习的联合学习框架(Extreme Kernel Sparse Learning, EKSL).通过引入行稀疏约束,提出了一种简化的核字典学习方法,解决了现有核字典学习方法固有的困难,减少了时间成本和存储负担.采用 PHAC-2 数据集进行验证,实验结果表明,和传统的基于 SVM 和标准 ELM 的方法相比,在绝大多数触觉识别任务中, EKSL 取得了最佳的分类结果. Iosifidis 等人^[123]提出了一种针对 KELM 的核矩阵近似求解算法 AKELM(Approximate Kernel ELM),用于处理人脸表情识别问题.为了进一步提高分类器的性能,将正则化方法以及训练数据的总体和类内方差整合到目标函数的优化过程中.实验在 JAFFE、COHN-KANADE、BU 和 YouTube 四个人脸表情数据库上进行,相比传统的基于 SVM 和 KELM 的人脸表情识别方法, AKELM 在计算复杂度和内存使用上具有更好的扩展性和较快的学习速度,在大多数情况下都取得了最佳的分类性能. Yang 等人^[124]提出了一种基于 ELM 的广义单隐层前馈神经网络模型.通过将网络残差反馈回隐层,从而实现广义隐层神经元的动态增长,这里的广义隐层神经元可以表示为一个子网络,而每个子网络又可以是若干隐层神经元组成.该模型在 Caltech-101、Caltech-256、Scenes-15、Yale faces 和 Olivetti faces 这五个图像识别数据集进行验证,取得了令人满意的结果.表 9 给出了该模型和 SVM、ELM 在不同图像识别数据集上的平均分类准确率. Tang 等人^[125]将深度神经网络(Deep Neural Network, DNN)和 ELM 相结合,解决了云层遮挡等天气条件下高分辨率星载图像船舶检测问题.

表 9 不同图像识别数据集上的平均分类准确率^[124]

数据集	SVM	ELM	ELM with Subnetwork Hidden Nodes
Scenes-15	72.80%	76.14%	82.21%
Caltech-101	63.70%	74.00%	76.60%
Yale faces	90.87%	91.45%	94.45%
Olivetti faces	98.07%	98.38%	98.50%

6.3 ELM 在故障诊断中的应用

随着工业设备和过程的复杂性不断提高,故障诊断的任务变得越来越困难.为了使系统高度可靠、安全地运行,需要及时检测异常情况,快速识别和隔离最可能的故障原因.由于算法简单,训练速度快,近年来,利用 ELM 实现实时准确的自动故障诊断已成为一个热门的研究方向.

Wong 等人^[126]开发了一个基于 ELM 的燃气轮发电机实时故障诊断系统,用于监测旋转机械的组件状态.在该系统中,采用小波包变换和时域统计特征提取振动信号.然后,使用核主成份分析进一步减少冗余特征.根据故障识别时间和故障检测精度对其进行评估,实验结果表明,ELM 可以检测到燃气轮发电机系统变速箱中多种组件故障,准确率高达 98.22%.当故障模式类别增加时,传统的识别算法无法有效地更新诊断系统,针对这一问题,Yin 等人^[127]提出了一种基于增量支持向量数据描述(Incremental Support Vector Data Description, ISVDD)和具有增量输出结构的极限学习机(Incremental Output Structure ELM, IOELM)在线故障诊断方法.在设备连续状态监测过程中,采用 ISVDD 快速寻找新的故障模式.将 ELM 的固定结构转化为弹性结构,使输出节点增量增加,有效识别新的故障模式.在 11 种不同条件下对柴油机的识别实验表明,该方法独立于检测对象,能够在线检测新的故障模式,且能够在线更新诊断系统的学习能力和知识.Chen 等人^[128]利用 KELM 建立光伏阵列故障诊断模型,并采用 Nelder-Mead 单纯形算法对 KELM 的参数进行优化,提出了一种改进的基于 Simulink 的精确光伏建模方法,便于对实验室光伏阵列进行故障模拟和数据采集.研究表明,基于 KELM 的光伏阵列故障诊断模型具有较高的精度、可靠性和良好的泛化性能.

除了上述应用以外,ELM 最近还在许多其它复杂学习任务中得到了广泛的研究和实现.如工业控制和机器人^[129-130]、文本分类^[131]、语音识别^[132]、图像质量评估^[133]、高光谱图像分类^[134]、3D 打印建模^[135]、室内定位系统^[136]、短期风速预测和电价预

测^[137-138]等等.这些丰硕的成果进一步扩展了 ELM 的应用领域,极大地推动了 ELM 的发展.

7 争议及未来研究方向

7.1 存在的争议

ELM 提出至今已经在很多领域得到了广泛应用和深入研究,并且逐渐被接受和认可.然而,在其发展过程中也伴随着不少争议.比如有文章指出 ELM 只是随机向量函数连接网络的一个变种^[139-140],是伪逆学习算法的“马甲”^①,不是一个新算法,没必要重新起一个名字.本节我们将围绕这些争议展开讨论与分析.

迄今为止,人们对神经网络的训练方法进行了大量的研究工作,其中大多数都是基于梯度下降法迭代调整所有参数.随着隐含层层数和节点数的增加,为了获得更好的效果,这类学习方法往往需要进行多次迭代,存在收敛速度慢、训练时间长、梯度消失、容易陷入局部最小值等问题,这些已经成为制约它们应用领域发展的主要瓶颈.近年来,神经网络随机学习方法因其在处理大规模数据分析,快速动态建模和实时数据处理等方面具有良好的应用潜力而受到越来越多的关注^[141].从 SLFN 的结构上看,此类学习方法的训练过程主要分为两个阶段.第一阶段,随机选择网络的输入权值和隐层偏置,利用隐层神经元的激活函数将输入空间转换到隐层空间.当隐层空间到输出空间的映射为线性时,在第二阶段只需使用最小二乘法确定网络的输出权值,无需迭代就能够获得全局最优解,从而显著提高网络的训练效率和性能.

关于神经网络随机学习方法的研究最早的尝试可以追溯到 20 世纪 50 年代,受生物神经网络的启发,Rosenblatt 设计了一种感知器模型^[142].一般来说,感知器由传感器单元、关联单元和响应单元组成.传感器单元以多对多的方式随机连接到关联单元,关联单元可以连接到其它关联单元或响应单元.传感器单元与响应单元之间的权值从{0,1}中随机选取,而关联单元和响应单元之间的权值是通过强化学习实现的.研究发现,这种具有随机连接单元的系统可以在特定约束下对特定输入给出特定响应.1992 年,Schmidt 等人^[143]对具有随机权值的 SLFN

① A VEST of the Pseudoinverse Learning Algorithm. preprint arXiv: 1805. 07828, 2018. <https://arxiv.org/pdf/1805.07828.pdf>.

的性能进行了比较分析. 他们从 $[-1, +1]$ 区间内随机产生输入权值和偏置, 在训练过程中保持不变, 通过求解线性最小二乘问题计算输出权值. 该研究指出, 隐含层和输出层之间的权值更重要, 而输入权值一旦被正确初始化就不再需要调整. 然而, 这种对输入权值和偏置取值范围的设置缺乏理论依据, 无法保证最终模型的通用逼近能力. 同年, Pao 等人^[144]提出了随机向量函数连接网络(Random Vector Functional Link Neural Network, RVFL). RVFL 是一种特殊的单隐层前馈神经网络, 其中输入层直接连接到隐含层和输出层. 在 RVFL 中, 输入权值和隐层偏置随机生成, 而输出权值可以通过计算隐含层输出矩阵的 Moore-Penrose 广义逆得到. 作为一种可供选择的网络, RVFL 在一定程度上克服了传统 BP 网络训练速度慢、容易陷入局部最小值的问题^[145]. 理论研究表明 RVFL 虽然对有限维输入集上的连续函数具有良好的通用逼近能力^[146], 但其随机参数的取值范围依赖于训练样本. 1995 年, Guo 等人^[147]针对 SLFN 的训练, 提出了一种高效的非迭代伪逆学习(PseudoInverse Learning Algorithm, PIL)算法. 在 PIL 中, 输入权值矩阵可以随机产生, 也可以设为输入矩阵的伪逆矩阵. 输出权值矩阵同样采用伪逆计算得到, 因此整个训练过程不需要反复迭代, 具有极快的学习速度, 并且能够获得全局最优解. 然而, 为了保证网络的泛化能力, PIL 要求隐层节点的个数要大于等于训练样本数, 因此无法应用在数据压缩问题中, 其中要求隐层节点个数小于训练样本数. 此后, Guo 等人^[148]又将 PIL 算法从单层神经网络扩展到多层, 并对算法的性质和加速、网络结构调整以及如何利用堆叠泛化技术提高模型的性能等问题进行了深入地讨论和分析.

2004 年, Huang 等人^[1]重新提出了随机生成隐层节点参数, 通过解析计算得到输出权值的 SLFN 训练方法, 并将其命名为 ELM, 作为梯度类方法的一种替代学习算法进行大力推广. 从本质上讲, 标准 ELM 的核心思想及训练过程和上述随机类学习算法基本相同. 然而, 在 ELM 最初的相关文献^[1, 5]中并没有提及这一领域的发展, 也没有对类似的方法进行讨论和比较, 因此引发了不少争议. 之后, Huang 等人在文献^[32, 65, 149]中指出了 ELM 与部分早期随机神经网络研究工作之间的关系和主要差异. 具体而言, ELM 和 Schmidt 等人^[143]提出的方法的区别在于输出层节点是否具有偏置项, 从标准

优化理论的角度来看, 这也是 ELM 和 SVM 之间最大的区别之一. 关于输出层是否需要偏置项以及偏置项对网络泛化性能的影响, 文献^[32]进行了详细的讨论和分析. 从网络结构上看, RVFL 最明显的特征是其输入层和输出层之间存在直接连接, 而标准 ELM 的输入层仅和隐含层相连. 最近 ELM 还被扩展到广义多层前馈神经网络^[112, 124, 150], 其隐层神经元可以由多个节点组成的子网络. 传统的 RVFL 主要用来训练全连接网络, 而 ELM 算法的变种既可以训练全连接网络, 也可以训练局部连接网络, 其中网络各层之间的连接是稀疏连接^[108]. 从通用逼近能力看, Schmidt 等人没有从理论上证明随机权值神经网络的通用逼近能力. Igel'nik 等人^[146]仅证明了部分随机化参数下 RVFL 的通用逼近能力, 但对完全随机化参数情况下 SLFN 的通用逼近能力没有涉及. 文献^[11, 64]分别从理论上严格证明了即使随机生成所有隐层节点参数, ELM 仍然能够保持 SLFN 的分类能力和通用逼近能力. 文献^[12]证明了 ELM 的 VC 维以概率 1 等于其隐层节点数, 从而为计算 ELM 的泛化误差界提供了一种便捷的方法. 从随机初始化机制看, 与 RVFL 中的基于训练样本分布或由目标函数确定的概率空间随机生成部分隐层节点参数不同, ELM 中的所有隐层节点参数都可以按照任意连续采样分布独立于目标函数或训练样本随机产生^[11, 65]. 总之, 虽然存在诸多争议, 但这也反过来推动了随机权值网络学习方法的发展.

此外, 标准 ELM 作为一种单层神经网络学习方法, 对解决许多具有良好约束的简单问题非常有效, 但在处理大数据或复杂学习任务时, 往往需要指数增长的隐层神经元, 才能获得更好的泛化能力. 对大多数实际应用来说, 这显然是无法接受的. 事实上, 这也是所有浅层神经网络面临的共性问题, 它们都不具备深层神经网络强大的数据表征能力. 针对这一问题, 近年来有很多研究人员提出了不同的解决方案^[14, 97, 103, 115]. 可以利用 MapReduce、Spark 等并行框架结合 ELM 来实现大数据的快速处理和分析. 还可以利用 ELM 深度学习模型获得原始数据的分层表示, 提高复杂任务的学习效率. 总之, 在解决具体问题时, 需要我们针对现有的数据选择合适的学习方法或者设计并构建适合自己问题的网络模型或学习算法, 只有这样才能更好地推动 ELM 等机器学习技术的发展. 同时, 需要指出的是, 深度学习虽然作为人工智能领域中非常火热的一个研究方向, 但这并不代表它就能完全取代浅层学习. 在一些

实际应用中,ELM 等浅层学习方法仍将继续发挥重要作用。

7.2 未来研究方向

近年来,经过大量研究人员长期不懈的努力,ELM 的研究不仅在理论上取得重要进展,而且已经在肿瘤检测、蛋白结构预测和交通标志识别等多个实际应用领域形成了丰硕的成果。但是作为一类新兴随机连接神经网络,它仍然面临许多亟待解决的问题和巨大挑战。根据当前 ELM 的国内外研究现状,本节提出一些值得深入探讨的研究难点,同时也是今后需要进一步解决与完善的研究方向。

(1) 理论研究。目前,ELM 算法的可行性和有效性大多数都是通过实验来证明的,理论分析还不够深入。随机特征映射是 ELM 的核心思想,但是对随机特征映射的有效评价方法仍处于空白状态没有突破。ELM 中隐神经元的选取没有形成一个统一合理的模式,对 ELM 网络模型的泛化能力分析缺乏较为系统的研究。到目前为止,尚不清楚 ELM 在哪些特定的问题上有很好的表现。由于采用非迭代学习策略,ELM 可以实现极快的学习速度,但是却失去了对输入权值的调节能力,可能无法获得学习权值的语义含义,因此这种非反馈机制对网络泛化能力的影响还有待研究。另外,众所周知,生物神经网络由无数的子网络构成,它们具有不同的功能并协同工作以使整个神经系统达到最优。因此,子网络与整个神经网络之间的关系在神经科学中具有重要意义,已成为神经科学领域最流行的研究课题之一。ELM 网络和其子网络之间的类似关系对于理论与算法研究尤为重要,因为这有助于深入理解网络的属性并设计各种高效的训练算法。然而,这种关系至今还没有得到充分的研究,仍处于起步阶段。ELM 的理论基础是 ELM 发展的基石,其理论体系还有待进一步完善与扩展。

(2) 核函数的构造及优化。在多核 ELM 中,如何针对具体应用选择或设计出最合适的核函数以及确定最优权值系数组合时的效率问题都是未解决的难点。现有的基于多核的 ELM 算法多采用简单的线性组合的方法对不同特性的核函数进行融合。在新核的构造过程中,核函数参数的选取缺乏实际的理论指导,只能通过实验对比或交叉验证等方法寻找最优的参数,无法有效解决非平坦分布的数据。通过多尺度分析方法对多个核函数进行融合,在处理复杂问题时要比合成核方法更具灵活性。因此,后续研究可以考虑引入尺度空间,对多尺度核进行整体

优化并和 ELM 相结合来提高模型的泛化性和普适性。除此之外,在处理一些复杂大规模问题时,需要对高维多核矩阵进行各种分解等处理,使得训练算法的时空开销急剧增长。如何解决海量样本下多核 ELM 的高效训练问题也值得进一步研究。

(3) 隐层结构设计。对于单隐层前馈神经网络而言,隐层神经元个数的选取对网络的泛化性能起着至关重要的影响。由前文的分析可知,现有的一些增量和剪枝构造法还存在诸多不足,例如:针对不同的问题,如何选择或设计合适的隐神经元评价准则、相关度阈值和终止条件没有固定的规则可循;当新神经元加入隐层后可能会影响原先神经元的有效性,即已存在的隐层神经元可能会对后续神经元的选择产生影响,而新神经元的加入又可能使之前选择的神经元失效而变成冗余神经元,导致最终的网络结构不够稀疏。另外,在利用群智能方法优化隐层神经元参数时,经常会出现早熟收敛现象,从而容易陷入局部最优解。对于一些大规模问题,由于群智能算法中涉及到太多的参数需要调优,会消耗大量的计算时间和内存资源使得学习效率非常低下。这些问题的存在,阻碍了 ELM 的进一步推广,使得网络结构优化方法的研究任重道远。自适应构造法虽然常常能够取得不错的表现,但是在每次迭代过程中都要根据事先定义好的评估准则对所有已存在的隐层神经元和候选神经元按重要性进行排序。当隐层有神经元加入或删除时,需要重新训练调整后的网络,计算并比较新旧网络的训练误差,该过程会不断重复,直到所有隐层神经元都被扫描一遍为止,因此具有较高的计算复杂度和时间复杂度,这显然无法适应某些对实时性要求很强的应用。如何提高自适应构造法的学习效率,在减少计算量的同时还不会影响模型的整体泛化性能是进一步亟待解决的课题。此外,隐层激活函数的选择对网络的泛化性能也有很大的影响,对于不同的问题,如何确定合适的激活函数还没有非常明确的方法和理论指导。

(4) 算法的鲁棒性。在取得输入数据有效的特征表示且样本中不含噪声的情况下,随机选择网络的输入权值和隐层偏置,训练过程中可以不做任何调整也能获得良好的泛化性能和极快的学习速度。但是,当数据中存在较多的离群点或者扰动时,随机生成隐层神经元参数会使得特征映射层调节能力变弱而产生冗余神经元,造成输出矩阵不是列满秩矩阵或者是一个病态矩阵,系统的鲁棒性会受到严重的影响,从而无法得到准确的结果。目前的改进算法

主要采用正则化方法和稳健估计理论优化目标函数,同时利用加权最小二乘法计算输出权值,来提高模型的抗干扰能力。然而,这些研究大都只集中在高斯噪声的处理上,无法有效应对具有混合噪声的数据集,降低了算法应用的灵活性。在未知复杂系统中存在各种类型的噪声,而一个好的模型应该能够精确地逼近实际系统的均值和方差。针对大规模数据集、不平衡数据集、半监督、无监督等复杂数据集中的噪声问题,如何使建模误差分布和噪声分布相一致,提高模型对不同噪声的鲁棒性,还需要进一步深入研究。此外,不同的随机参数会产生不同的模型,从而导致模型的性能不稳定,但是至今仍缺乏有效的方法来指导隐节点参数的选择。

(5) 大数据在线学习。随着人们对大数据处理实时性的要求越来越高,如何对海量流数据进行实时在线训练成为机器学习领域的一大挑战。传统的 MapReduce 框架采用离线处理计算的方式,主要用于对静态数据的批量学习,并不适合用于处理流数据。目前的一些基于 ELM 的在线学习方法主要是针对小规模结构化数据集而提出的,对于海量数据流中的类别不平衡和概念漂移等问题的研究很少涉及,因此如何构建高性能分布式并行在线 ELM 模型来实现复杂大数据实时流处理值得进一步研究。

(6) 深度极限学习机。理论上只要隐层神经元数足够多,ELM 就能充分逼近地表达任意多元非线性函数,但在具体实现时往往由于需要太多的隐神经元而无法实际应用。在解决真实世界的复杂模式识别问题时,通常需要使用深层神经网络从大量数据中逐层提取特征,获得输入数据更有意义的表示。当前的深度 ELM 模型大都是由多个 ELM-AE 堆叠而成,还存在如下问题:① 模型学习性能的好坏依赖于每层中隐层神经元个数取值是否合理,一般是以经验的方式确定或者通过实验的方法不断调整各隐层的神经元个数以达到最优。显然,这是一个非常耗时的过程,因为根据应用的复杂程度,每个隐层神经元数的取值范围可以从数百到数万个;② ELM-AE 随机选择输入权值和隐层偏置进行特征映射会导致模型泛化性能不稳定;③ 在训练过程中,各隐层的重构误差累计在一起传播到输出层,容易产生次优解;④ 各隐层之间的连接权值由 ELM-AE 训练得到,也就是说前一层 ELM 自编码器的输出权值将作为后一层 ELM 自编码器的输入,因此所有 ELM 自编码器的输出权值占用的存储容量与堆叠模型的隐层层数成正比。对于那些需要大量隐

层的应用问题来说,空间复杂度会非常高,甚至还会影响模型的训练速度。此外,针对不同问题,如何确定最合适的隐层层数来提取输入数据的有效特征;如何将深度 ELM 并行化,使之能够处理大规模复杂应用;如何改善深度 ELM 模型的可解释性;如何在训练样本集规模很小的情况下,有效地避免过拟合问题的发生;如何更好地将权值共享和局部感知的概念融入到深度 ELM 中构建具有强稀疏性的网络结构来降低学习的复杂度,这些都是极具挑战性并且亟需进一步深入研究的课题。

(7) 拓展极限学习机的应用领域。目前,大多数 ELM 算法还停留在理论研究阶段,使用比较简单的公开数据集或者人工数据集进行测试,很少考虑真实世界中的复杂场景。因此,需要结合具体的任务需求进一步拓展这些算法的应用领域来解决问题,如自然语言处理、计算机视觉、智能物联网等。只有取得实际应用价值,才能让 ELM 保持长久发展。

8 结束语

ELM 作为一种新兴的前馈神经网络学习框架,是在克服反向传播算法对 SLFN 网络的训练困境中逐步建立和发展起来的。ELM 随机选择网络的输入权值和隐层偏置,通过解析计算得到隐层输出权值,和传统的基于梯度下降的训练算法相比,它具有学习速度快、易于实现、人为干预少等显著优势,为大规模数据挖掘问题提供了一种有效的解决方案,成为人工智能领域一个新的研究热点,引起大量学术界和工业界研究人员的高度关注和深入研究,并取得了一系列重要成果。

在浅层神经网络领域里,ELM 仍将继续发挥重要作用,特别是推动了神经网络中随机连接/随机神经元这类方法的发展研究。对于简单的模式识别问题,现有浅层学习算法已经足以应对,但在解决复杂大规模高维数据问题时,浅层模型往往容易出现函数表征能力不足的情况。这是因为在处理某些复杂问题时,可能需要指数级增长的隐层神经元数,而无法满足实际应用需求。相对浅层模型而言,深度学习通过训练大量数据,学习调整具有多个隐层的神经网络模型,以期获得原始输入数据的分层表示,通过逐层特征组合的方式来提取数据的有效特征,进而提高分类或预测的准确性。在深度神经网络发展过程中,遇到的主要问题是其训练算法得不到很好解决,例如采用误差反向传播算法训练时,梯度的幅度

会随着隐层层数的增加急剧减小直至消失,从而导致权值的更新非常缓慢,无法有效完成学习任务。目前深度 ELM 的理论研究还处于起步阶段,还在不断发展完善,但从当前的发展趋势来看,其在应用领域已显现出巨大能量。其中,借鉴 ELM 算法思想代替传统的逐层贪心训练算法进行高效学习的方法,为训练深度神经网络带来了希望。

本文围绕理论和算法这两大部分对 ELM 的国内外研究现状与前沿问题进行了系统分析和归纳总结。在此基础上,对 ELM 中面临的挑战以及今后的研究方向做出了展望。ELM 作为一类随机连接神经网络,也可以被视为大脑实际工作方式的某一种抽象版本。从输入层直接使用随机权值的角度看,它有可能是更靠近大脑生物工作方式的学习算法。尽管使用众多的随机神经元,ELM 仍然能高效地完成那些需要算法反复迭代执行才能达到最优的同类任务。总之,极限学习机虽然仍存在诸多难题有待解决,但这并不影响它在人工智能领域中广阔的发展和潜力。可以预见,随着这些难题的不断攻克,势必将推动神经网络的研究和应用迈向一个新的高潮。

参 考 文 献

- [1] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: A new learning scheme of feedforward neural networks// Proceedings of the 2004 IEEE International Joint Conference on Neural Networks. Budapest, Hungary, 2004: 985-990
- [2] Chyzykh D, Savio A, Graña M. Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Networks*, 2015, 68: 23-33
- [3] Huang Z, Yu Y, Gu J. An efficient method for traffic sign recognition based on extreme learning machine. *IEEE Transactions on Cybernetics*, 2017, 47(4): 920-933
- [4] Wang S, Deng C, Lin W. NMF-based image quality assessment using extreme learning machine. *IEEE Transactions on Cybernetics*, 2017, 47(1): 232-243
- [5] Huang G B, Siew C K. Extreme learning machine: RBF network case//Proceedings of the 8th International Conference on Control, Automation, Robotics and Vision (ICARCV). Kunming, China 2004: 1029-1036
- [6] Zhang Y, Wu J, Zhou C. Instance cloned extreme learning machine. *Pattern Recognition*, 2017, 68: 52-65
- [7] Ertugrul Ö F. Forecasting electricity load by a novel recurrent extreme learning machines approach. *International Journal of Electrical Power & Energy Systems*, 2016, 78: 429-435
- [8] Tang J, Deng C, Huang G B. Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, 27(4): 809-821
- [9] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications. *Neurocomputing*, 2006, 70(1-3): 489-501
- [10] Huang G B, Chen L. Convex incremental extreme learning machine. *Neurocomputing*, 2007, 70(16-18): 3056-3062
- [11] Huang G B, Zhou H, Ding X. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, 42(2): 513-529
- [12] Liu X, Gao C, Li P. A comparative analysis of support vector machines and extreme learning machines. *Neural Networks*, 2012, 33: 58-66
- [13] Yao E, Basu A. VLSI extreme learning machine: A design space exploration. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2017, 25(1): 60-74
- [14] Deng W Y, Bai Z, Huang G B, et al. A fast SVD-hidden-nodes based extreme learning machine for large-scale data analytics. *Neural Networks*, 2016, 77: 14-28
- [15] Tang J, Deng C, Huang G B. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, 53(3): 1174-1185
- [16] Kim J, Kim J, Jang G J. Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection. *Neural Networks*, 2017, 87: 109-121
- [17] Zhang L, Zhang D. Evolutionary cost-sensitive extreme learning machine. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(12): 3045-3060
- [18] Deng Wan-Yu, Zheng Qing-Hua, Chen Lin, et al. Research on extreme learning of neural networks. *Chinese Journal of Computers*, 2010, 33(2): 279-287(in Chinese)
(邓万宇, 郑庆华, 陈琳等. 神经网络极速学习方法研究. 计算机学报, 2010, 33(2): 279-287)
- [19] Zhang K, Luo M. Outlier-robust extreme learning machine for regression problems. *Neurocomputing*, 2015, 151(P3): 1519-1527
- [20] Huang G, Song S, Gupta J N D. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 2014, 44(12): 2405-2417
- [21] Huang G, Liu T, Yang Y. Discriminative clustering via extreme learning machine. *Neural Networks*, 2015, 70: 1-8
- [22] Yi Y, Qiao S, Zhou W. Adaptive multiple graph regularized semi-supervised extreme learning machine. *Soft Computing*, 2018, 22(11): 3545-3562
- [23] Iosifidis A, Tefas A, Pitas I. Graph embedded extreme learning machine. *IEEE Transactions on Cybernetics*, 2016, 46(1): 311-324
- [24] Yu Y, Sun Z. Sparse coding extreme learning machine for classification. *Neurocomputing*, 2017, 261: 50-56
- [25] Lu X J, Ming L, Liu W B. Probabilistic regularized extreme learning machine for robust modeling of noise data. *IEEE Transactions on Cybernetics*, 2018, 48(8): 2368-2377

- [26] Zhao Y P, Hu Q K, Xu J G. A robust extreme learning machine for modeling a small-scale turbojet engine. *Applied Energy*, 2018, 218: 22-35
- [27] Chen K, Lv Q, Lu Y. Robust regularized extreme learning machine for regression using iteratively reweighted least squares. *Neurocomputing*, 2017, 230: 345-358
- [28] Xiao W, Zhang J, Li Y. Class-specific cost regulation extreme learning machine for imbalanced classification. *Neurocomputing*, 2017, 261: 70-82
- [29] Liu Q, He Q, Shi Z. Extreme support vector machine classifier//Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Osaka, Japan, 2008: 222-233
- [30] Frénay B, Verleysen M. Using SVMs with randomised feature spaces: An extreme learning approach//Proceedings of the 18th European Symposium on Artificial Neural Networks-Computational Intelligence and Machine Learning (ESANN). Bruges, Belgium, 2010: 315-320
- [31] Huang G B, Ding X, Zhou H. Optimization method based extreme learning machine for classification. *Neurocomputing*, 2010, 74(1-3): 155-163
- [32] Huang G B. An insight into extreme learning machines: Random neurons, random features and kernels. *Cognitive Computation*, 2014, 6(3): 376-390
- [33] He Q, Du C, Wang Q. A parallel incremental extreme SVM classifier. *Neurocomputing*, 2011, 74(16): 2532-2540
- [34] Deng Wan-Yu, Zheng Qing-Hua, Chen Lin. Large-Scale kernel extreme learning machine. *Chinese Journal of Computers*, 2014, 37(11): 2235-2246(in Chinese)
(邓万宇, 郑庆华, 陈琳. 面向大样本数据的核化极速神经网络. *计算机学报*, 2014, 37(11): 2235-2246)
- [35] Deng W Y, Ong Y S, Zheng Q H. A fast reduced kernel extreme learning machine. *Neural Networks*, 2016, 76: 29-38
- [36] Bi X, Zhao X, Wang G. Distributed extreme learning machine with kernels based on MapReduce. *Neurocomputing*, 2015, 149(Part A): 456-463
- [37] Bai Z, Huang G B, Wang D. Sparse extreme learning machine for classification. *IEEE Transactions on Cybernetics*, 2014, 44(10): 1858-1870
- [38] Chen B, Xing L, Wang X. Robust learning with kernel mean p -power error loss. *IEEE Transactions on Cybernetics*, 2018, 48(7): 2101-2113
- [39] Liu X, Wang L, Huang G B. Multiple kernel extreme learning machine. *Neurocomputing*, 2015, 149(Part A): 253-264
- [40] Wu D, Wang Z, Chen Y. Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing*, 2016, 190: 35-49
- [41] Li X, Mao W, Jiang W. Multiple-kernel-learning-based extreme learning machine for classification design. *Neural Computing and Applications*, 2016, 27(1): 175-184
- [42] Xue J, Liu Q, Li M. Incremental multiple kernel extreme learning machine and its application in Robo-advisors. *Soft Computing*, 2018, 22(11): 3507-3517
- [43] Liang N Y, Huang G B, Saratchandran P. A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 2006, 17(6): 1411-1423
- [44] Lan Y, Soh Y C, Huang G B. Ensemble of online sequential extreme learning machine. *Neurocomputing*, 2009, 72(13-15): 3391-3395
- [45] Zhao J, Wang Z, Park D S. Online sequential extreme learning machine with forgetting mechanism. *Neurocomputing*, 2012, 87: 79-89
- [46] Gu Y, Liu J, Chen Y. TOSELM: Timeliness online sequential extreme learning machine. *Neurocomputing*, 2014, 128: 119-127
- [47] Huynh H T, Won Y. Regularized online sequential learning algorithm for single-hidden layerfeedforward neural networks. *Pattern Recognition Letters*, 2011, 32(14): 1930-1935
- [48] Mirza B, Lin Z, Toh K A. Weighted online sequential extreme learning machine for class imbalance learning. *Neural Processing Letters*, 2013, 38(3): 465-486
- [49] Vanli N D, Sayin M O, Delibalta I. Sequential nonlinear learning for distributed multiagent systems via extreme learning machines. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(3): 546-558
- [50] Wang X, Han M. Online sequential extreme learning machine with kernels for nonstationary time series prediction. *Neurocomputing*, 2014, 145: 90-97
- [51] Scardapane S, Comminiello D, Scarpiniti M. Online sequential extreme learning machine with kernels. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(9): 2214-2220
- [52] Deng W Y, Ong Y S, Tan P S. Online sequential reduced kernel extreme learning machine. *Neurocomputing*, 2016, 174: 72-84
- [53] Liu D, Wu Y X, Jiang H. FP-ELM: An online sequential learning algorithm for dealing with concept drift. *Neurocomputing*, 2016, 207: 322-334
- [54] Zhang Y, Er M J. Sequential active learning using meta-cognitive extreme learning machine. *Neurocomputing*, 2016, 173: 835-844
- [55] Yang J, Ye F, Rong H J. Recursive least mean p -power extreme learning machine. *Neural Networks*, 2017, 91: 22-33
- [56] Yang L, Yang S, Li S. Incremental Laplacian regularization extreme learning machine for online learning. *Applied Soft Computing*, 2017, 59: 546-555
- [57] Zhu Q Y, Qin A K, Suganthan P N. Evolutionary extreme learning machine. *Pattern Recognition*, 2005, 38(10): 1759-1763
- [58] Cao J, Lin Z, Huang G B. Self-adaptive evolutionary extreme learning machine. *Neural Processing Letters*, 2012, 36(3): 285-305

- [59] Tang X F, Chen L. A self-adaptive evolutionary weighted extreme learning machine for binary imbalance learning. *Progress in Artificial Intelligence*, 2018, 7(2): 95-118
- [60] Xu Y, Shu Y. Evolutionary extreme learning machine—based on particle swarm optimization//*Proceedings of the 3rd International Symposium on Neural Networks*. Chengdu, China, 2006: 644-652
- [61] Figueiredo E M N, Ludermir T B. Investigating the use of alternative topologies on performance of the PSO-ELM. *Neurocomputing*, 2014, 127(1): 4-12
- [62] Lahoz D, Lacruz B, Mateo P M. A multi-objective micro genetic ELM algorithm. *Neurocomputing*, 2013, 111: 90-103
- [63] Zhang Y, Wu J, Cai Z. Memetic extreme learning machine. *Pattern Recognition*, 2016, 58: 135-148
- [64] Huang G B, Chen L, Siew C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 2006, 17(4): 879-892
- [65] Huang G B, Li M B, Chen L. Incremental extreme learning machine with fully complex hidden nodes. *Neurocomputing*, 2008, 71(4-6): 576-583
- [66] Huang G B, Chen L. Enhanced random search based incremental extreme learning machine. *Neurocomputing*, 2008, 71(16-18): 3460-3468
- [67] Feng G, Huang G B, Lin Q. Error minimized extreme learning machine with growth of hidden nodes and incremental learning. *IEEE Transactions on Neural Networks*, 2009, 20(8): 1352-1357
- [68] Lan Y, Soh Y C, Huang G B. Constructive hidden nodes selection of extreme learning machine for regression. *Neurocomputing*, 2010, 73(16-18): 3191-3199
- [69] Yang Y, Wang Y, Yuan X. Bidirectional extreme learning machine for regression problem and its learning effectiveness. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, 23(9): 1498-1505
- [70] Ye Y, Qin Y. QR factorization based incremental extreme learning machine with growth of hidden nodes. *Pattern Recognition Letters*, 2015, 65: 177-183
- [71] Li S, You Z H, Guo H. Inverse-Free extreme learning machine with optimal information updating. *IEEE Transactions on Cybernetics*, 2016, 46(5): 1229-1241
- [72] Xu Z, Yao M, Wu Z. Incremental regularized extreme learning machine and its enhancement. *Neurocomputing*, 2016, 174: 134-142
- [73] Rong H J, Ong Y S, Tan A H. A fast pruned-extreme learning machine for classification problem. *Neurocomputing*, 2008, 72(1-3): 359-366
- [74] Miche Y, Sorjamaa A, Bas P. OP-ELM: Optimally pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 2010, 21(1): 158-162
- [75] Miche Y, Van Heeswijk M, Bas P. TROP-ELM: A double-regularized ELM using LARS and Tikhonov regularization. *Neurocomputing*, 2011, 74(16): 2413-2421
- [76] Grigorievskiy A, Miche Y, Ventelä A M. Long-term time series prediction using OP-ELM. *Neural Networks*, 2014, 51: 50-56
- [77] MartiNez-MartiNez J M, Escandell-Montero P, Soria-Olivas E. Regularized extreme learning machine for regression problems. *Neurocomputing*, 2011, 74(17): 3716-3721
- [78] Bueno-Crespo A, García-Laencina P J, Sancho-Gómez J L. Neural architecture design based on extreme learning machine. *Neural Networks*, 2013, 48: 19-24
- [79] Balasundaram S, Gupta D. 1-Norm extreme learning machine for regression and multiclass classification using Newton method. *Neurocomputing*, 2014, 128: 4-14
- [80] Luo X, Liu F, Yang S. Joint sparse regularization based sparse semi-supervised extreme learning machine (S3ELM) for classification. *Knowledge-Based Systems*, 2015, 73: 149-160
- [81] Zhou S, Liu X, Liu Q. Random Fourier extreme learning machine with $\ell_{2,1}$ -norm regularization. *Neurocomputing*, 2016, 174: 143-153
- [82] Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. *Pattern Recognition Letters*, 1994, 15(11): 1119-1125
- [83] Lan Y, Soh Y C, Huang G B. Two-stage extreme learning machine for regression. *Neurocomputing*, 2010, 73(16-18): 3028-3038
- [84] Zhang R, Lan Y, Huang G. Universal approximation of extreme learning machine with adaptive growth of hidden nodes. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, 23(2): 365-371
- [85] Zhang R, Lan Y, Huang G B. Dynamic extreme learning machine and its approximation capability. *IEEE Transactions on Cybernetics*, 2013, 43(6): 2054-2065
- [86] Pratama M, Zhang G, Er M J. An incremental type-2 meta-cognitive extreme learning machine. *IEEE Transactions on Cybernetics*, 2017, 47(2): 339-353
- [87] Sun Y, Chen Y, Yuan Y. Dynamic adjustment of hidden layer structure for convex incremental extreme learning machine. *Neurocomputing*, 2017, 261: 83-93
- [88] Zhao Y P, Li Z Q, Xi P P. Gram-Schmidt process based incremental extreme learning machine. *Neurocomputing*, 2017, 241: 1-17
- [89] Feng G, Lan Y, Zhang X, et al. Dynamic adjustment of hidden node parameters for extreme learning machine. *IEEE Transactions on Cybernetics*, 2015, 45(2): 279-288
- [90] Van Heeswijk M, Miche Y, Lindh-Knuutila T. Adaptive ensemble models of extreme learning machines for time series prediction//*Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN)*. Limassol, Cyprus, 2009: 305-314
- [91] Liu N, Wang H. Ensemble based extreme learning machine. *IEEE Signal Processing Letters*, 2010, 17(8): 754-757

- [92] Tian H X, Mao Z Z. An ensemble ELM based on modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace. *IEEE Transactions on Automation Science and Engineering*, 2010, 7(1): 73-80
- [93] Cao J, Lin Z, Huang G B. Voting based extreme learning machine. *Information Sciences*, 2012, 185(1): 66-77
- [94] Xue X, Yao M, Wu Z. Genetic ensemble of extreme learning machine. *Neurocomputing*, 2014, 129: 175-184
- [95] Wang X L, Chen Y Y, Zhao H. Parallelized extreme learning machine ensemble based on min-max modular network. *Neurocomputing*, 2014, 128: 31-41
- [96] Samat A, Du P, Liu S, et al. E² LMs: Ensemble extreme learning machines for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014, 7(4): 1060-1069
- [97] Yang Y, Wu Q M J, Wang Y, et al. Data partition learning with multiple extreme learning machines. *IEEE Transactions on Cybernetics*, 2015, 45(8): 1463-1475
- [98] Ning K, Liu M, Dong M. Two efficient twin ELM methods with prediction interval. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(9): 2058-2071
- [99] Van Heeswijk M, Miche Y, Oja E. GPU-accelerated and parallelized ELM ensembles for large-scale regression. *Neurocomputing*, 2011, 74(16): 2430-2437
- [100] He Q, Shang T, Zhuang F. Parallel extreme learning machine for regression based on MapReduce. *Neurocomputing*, 2013, 102: 52-58
- [101] Wang B, Huang S, Qiu J. Parallel online sequential extreme learning machine based on MapReduce. *Neurocomputing*, 2015, 149(PA): 224-232
- [102] Wang Y, Dou Y, Liu X. PR-ELM: Parallel regularized extreme learning machine based on cluster. *Neurocomputing*, 2016, 173: 1073-1081
- [103] Duan M, Li K, Liao X. A parallel multiclassification algorithm for big data using an extreme learning machine. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 29(6): 2337-2351
- [104] Lam D, Wunsch D. Unsupervised feature learning classification with radial basis function extreme learning machine using graphic processors. *IEEE Transactions on Cybernetics*, 2017, 47(1): 224-231
- [105] Chen C, Li K, Ouyang A. A parallel approximate SS-ELM algorithm based on MapReduce for large-scale datasets. *Journal of Parallel and Distributed Computing*, 2017, 108: 85-94
- [106] Akusok A, Björk K M, Miche Y. High-performance extreme learning machines: A complete toolbox for big data applications. *IEEE Access*, 2015, 3: 1011-1025
- [107] Kasun L L C, Zhou H, Huang G B. Representational learning with extreme learning machine for big data. *IEEE Intelligent Systems*, 2013, 28(6): 31-34
- [108] Huang G B, Bai Z, Kasun L L C. Local receptive fields based extreme learning machine. *IEEE Computational Intelligence Magazine*, 2015, 10(2): 18-29
- [109] Zhou H, Huang G B, Lin Z. Stacked extreme learning machines. *IEEE Transactions on Cybernetics*, 2015, 45(9): 2013-2025
- [110] Yu W, Zhuang F, He Q. Learning deep representations via extreme learning machines. *Neurocomputing*, 2015, 149 (Part A): 308-315
- [111] Wong C M, Vong C M, Wong P K. Kernel-based multilayer extreme learning machines for representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(3): 757-762
- [112] Yang Y, Wu Q M J. Multilayer extreme learning machine with subnetwork nodes for representation learning. *IEEE Transactions on Cybernetics*, 2016, 46(11): 2570-2583
- [113] Wang Y, Xie Z, Xu K. An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning. *Neurocomputing*, 2016, 174: 988-998
- [114] Sun K, Zhang J, Zhang C. Generalized extreme learning machine autoencoder and a new deep neural network. *Neurocomputing*, 2017, 230: 374-381
- [115] Chen C, Li K, Ouyang A, et al. GPU-accelerated parallel hierarchical extreme learning machine on Flink for big data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017, 47(10): 2740-2753
- [116] Duan M, Li K, Li K. An ensemble CNN2ELM for age estimation. *IEEE Transactions on Information Forensics and Security*, 2018, 13(3): 758-772
- [117] Zhang Y, Wang Y, Zhou G. Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces. *Expert Systems with Applications*, 2018, 96: 302-310
- [118] Saraswathi S, Sundaram S, Sundararajan N. ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011, 8(2): 452-463
- [119] Huang W, Tan Z M, Lin Z, et al. A semi-automatic approach to the segmentation of liver parenchyma from 3D CT images with extreme learning machine//*Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. San Diego, USA, 2012: 3752-3755
- [120] Huang W, Li N, Lin Z, et al. Liver tumor detection and segmentation using kernel-based extreme learning machine //*Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Osaka, Japan, 2013: 3662-3665
- [121] Huang W, Yang Y, Lin Z, et al. Random feature subspace ensemble based extreme learning machine for liver tumor detection and segmentation//*Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine*

- and Biology Society (EMBC). Chicago, USA, 2014: 4675-4678
- [122] Liu H, Qin J, Sun F. Extreme kernel sparse learning for tactile object recognition. *IEEE Transactions on Cybernetics*, 2017, 47(12): 4509-4520
- [123] Iosifidis A, Tefas A, Pitas I. Approximate kernel extreme learning machine for large scale data classification. *Neurocomputing*, 2017, 219: 210-220
- [124] Yang Y, Wu Q M J. Extreme learning machine with subnetwork hidden nodes for regression and classification. *IEEE Transactions on Cybernetics*, 2016, 46(12): 2885-2898
- [125] Tang J, Deng C, Huang G B, Zhao B. Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, 53(3): 1174-1185
- [126] Wong P K, Yang Z, Vong C M, Zhong J. Real-time fault diagnosis for gas turbine generator systems using extreme learning machine. *Neurocomputing*, 2014, 128: 249-257
- [127] Yin G, Zhang Y T, Li Z N, et al. Online fault diagnosis method based on incremental support vector data description and extreme learning machine with incremental output structure. *Neurocomputing*, 2014, 128: 224-231
- [128] Chen Z, Wu L, Cheng S, et al. Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and IV characteristics. *Applied Energy*, 2017, 204: 912-931
- [129] Yu Y, Choi T M, Hui C L. An intelligent quick prediction algorithm with applications in industrial control and loading problems. *IEEE Transactions on Automation Science and Engineering*, 2012, 9(2): 276-287
- [130] Zhang Y, Slaughter D C. Hyperspectral species mapping for automatic weed control in tomato under thermal environmental stress. *Computers and Electronics in Agriculture*, 2011, 77(1): 95-104
- [131] Zheng W, Qian Y, Lu H. Text categorization based on regularization extreme learning machine. *Neural Computing and Applications*, 2013, 22(3-4): 447-456
- [132] Xu J, Zhang W Q, Liu J, Xia S. Regularized minimum class variance extreme learning machine for language recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, 2015(1): 22
- [133] Zhou W, Yu L, Zhou Y, et al. Blind quality estimator for 3D images based on binocular combination and extreme learning machine. *Pattern Recognition*, 2017, 71: 207-217
- [134] Li W, Chen C, Su H, Du Q. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, 53(7): 3681-3693
- [135] Zhang X, Le X, Panotopoulou A, et al. Perceptual models of preference in 3D printing direction. *ACM Transactions on Graphics*, 2015, 34(6): 215
- [136] Zou H, Huang B, Lu X, et al. A robust indoor positioning system based on the procrustes analysis and weighted extreme learning machine. *IEEE Transactions on Wireless Communications*, 2016, 15(2): 1252-1266
- [137] Chen X, Dong Z Y, Meng K, et al. Electricity price forecasting with extreme learning machine and bootstrapping. *IEEE Transactions on Power Systems*, 2012, 27(4): 2055-2062
- [138] Ak R, Fink O, Zio E. Two machine learning approaches for short-term wind speed time-series prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, 27(8): 1734-1747
- [139] Wang L P, Wan C R. Comments on "The Extreme Learning Machine". *IEEE Transactions on Neural Networks*, 2008, 19(8): 1494-1495
- [140] Huang G B. Reply to "Comments on "The Extreme Learning Machine"". *IEEE Transactions on Neural Networks*, 2008, 19(8): 1495-1496
- [141] Zhang L, Suganthan P N. A survey of randomized algorithms for training neural networks. *Information Sciences*, 2016, 364: 146-155
- [142] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958, 65(6): 386
- [143] Schmidt W F, Kraaijveld M A, Duin R P W. Feedforward neural networks with random weights//*Proceedings of the 11th IAPR International Conference on Pattern Recognition*. Hague, Netherlands, 1992: 1-4
- [144] Pao Y H, Takefuji Y. Functional-link net computing: Theory, system architecture, and functionalities. *Computer*, 1992, 25(5): 66-79
- [145] Pao Y H, Park G H, Sobajic D J. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 1994, 6(2): 163-180
- [146] Igel'nik B, Pao Y H. Stochastic choice of basis functions in adaptive function approximation and the functional-link net. *IEEE Transactions on Neural Networks*, 1995, 6(6): 1320-1329
- [147] Guo P, Chen C L P, Sun Y. An exact supervised learning for a three-layer supervised neural network//*Proceedings of the International Conference on Neural Information Processing (ICONIP'95)*. Beijing, China, 1995: 1041-1044
- [148] Guo P, Lyu M R. A pseudoinverse learning algorithm for feedforward neural networks with stacked generalization applications to software reliability growth data. *Neurocomputing*, 2004, 56(1-4): 101-121
- [149] Huang G B. What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle. *Cognitive Computation*, 2015, 7(3): 263-278
- [150] Wang N, Er M J, Han M. Generalized single-hidden layer feedforward networks for regression problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(6): 1161-1176



XU Rui, Ph.D. candidate. His research interests include image processing and machine learning.

LIANG Xun, Ph. D. , professor, Ph. D. supervisor. His research interests include data mining, neural networks, and social computing.

QI Jin-Shan, Ph.D. candidate. His research interests include social computing and data mining.

LI Zhi-Yu, Ph.D. candidate. His research interests include social computing and natural language processing.

ZHANG Shu-Sen, Ph.D. candidate. His research interests include data mining and social computing.

Background

In most learning algorithms, the parameters of each layer of feedforward neural networks need to be iteratively tuned using gradient descent training methods. However, traditional gradient descent based learning methods are generally very slow due to improper learning steps or may easily converge to local minimums. In addition, many iterative learning steps are required by such learning algorithms in order to obtain better generalization performance. The slow learning speed of feedforward neural networks has become the main bottleneck in their applications for past decades.

ELM as a new learning framework has recently been proposed for generalized single hidden layer feedforward neural networks to overcome the challenging issues faced by traditional gradient descent-based methods and was extended to generalized multilayer feedforward neural networks later. ELM randomly selects the input weights and hidden biases. Thus it has faster learning speed, least human intervention and easy to implementation. ELM has become one of the most popular research directions in the field of artificial intelligence in recent years and received widespread attention from more and more research members domestic and abroad.

This paper provides a comprehensive review of existing research results in ELM. The authors first give an introduction to the historical background and developments of feedforward neural networks. Then they introduce the principle and algorithm of ELM in detail followed by its feature map and feature space. After an overview of ELM theory, the authors discuss and analyze state-of-the-art algorithms or the typical variants of ELM, including models, solution approaches and relevant problems. On this basis, the main idea of each algorithm and their strengths and weaknesses are summarized. In the end, several open issues and challenges in ELM are pointed out together with its future research directions and trends.

This paper is supported by the National Natural Science Foundation of China under Grant Nos. 71531012, 61762073, 71601013, the National Social Science Foundation of China under Grant Nos. 18ZDA309; the Natural Science Foundation of Beijing under Grant Nos. 4172032, 4174087, and the Opening Project of State Key Laboratory of Digital Publishing Technology of the Founder Group.