计算机科学
COMPUTER SCIENCE

针对高维数据的动态集成堆叠宽度学习系统

云帆, 余志文, 杨楷翔

引用本文

云帆, 余志文, 杨楷翔. 针对高维数据的动态集成堆叠宽度学习系统[J]. 计算机科学, 2026, 53(4): 48-56.

YUN Fan, YU Zhiwen, YANG Kaixiang. Dynamic Ensemble Stacking Broad Learning System for High-dimensional Data [J]. Computer Science, 2026, 53(4): 48-56.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.251000068

针对高维数据的动态集成堆叠宽度学习系统

云　帆　　余志文　　杨楷翔
华南理工大学计算机科学与工程学院　广州 510006
(820832107@qq.com)

摘　要　在高维小样本分类任务中,宽度学习系统(Broad Learning System,BLS)因其高效的特性而备受关注。然而,原始的单层BLS的特征提取能力有限,难以处理复杂的高维数据。随机节点生成机制导致直接堆叠BLS隐层时出现节点冗余,模型性能难以提升。为解决上述问题,提出了一种集成堆叠BLS算法。所提算法利用前一层BLS的输出作为增强特征,将其与按分类置信度加权的原始特征进行拼接后输入下一层BLS,不断提高深层特征表达能力。通过元学习器池集成多个BLS层的输出,增强了原始单层BLS的高维特征提取能力,从而提升了模型的泛化性能。此外,考虑到高维数据复杂多变的特性,设计了动态集成框架,根据数据难度动态调整模型的复杂度。所提方法在保持模型性能的同时,进一步提升了集成效率。消融实验证明了所提算法的各个模块的有效性,对比实验证明了所提算法在高维疾病数据上的优越分类性能。

关键词:宽度学习系统;集成学习;动态结构;高维数据;堆叠
中图分类号　TP302

Dynamic Ensemble Stacking Broad Learning System for High-dimensional Data

YUN Fan,YU Zhiwen and YANG Kaixiang
College of Computer Science and Engineering,South China University of Technology,Guangzhou 510006,China

Abstract　In high-dimensional small sample classification tasks,BLS (Broad Learning System) has garnered much attention due to its efficiency. However,the feature extraction capability of the single-layer BLS is limited,making it difficult to handle complex high-dimensional data. The random node generation mechanism induces node redundancy when directly stacking BLS hidden layers,thereby hindering improvements in model performance. To address these issues,an ensemble stack BLS(E-SBLS) algorithm is proposed. E-SBLS utilizes the output of the previous BLS layer as enhanced features,concatenates them with the original feature weighted by classification confidence,and sends them into the subsequent BLS to continuously enhance the feature representation capability in deeper layers. By integrating the outputs of multiple BLS layers through a meta-learner pool,the high-dimensional feature extraction ability of the original single-layer BLS is augmented,thereby improving the generalization perfor-mance of the proposed model. Furthermore,considering the complex and variable characteristics of high-dimensional data,a dynamic ensemble framework is designed to adjust the complexity of the model dynamically based on data difficulties. The proposed method further enhances ensemble efficiency while maintaining model performance. Ablation experiments validate the effectiveness of each module in the proposed algorithm,and comparative experiments demonstrate the superior classification performance of the proposed model on high-dimensional disease data.

Keywords　Broad learning system,Ensemble learning,Dynamic structure,High-dimensional data,Stacking

1　引言

现实世界中存在着大量高维数据,特别是在疾病检测领域,随着基因组学的迅猛发展,小样本高维数据在生物信息领域随处可见。例如,在基因表达数据中,每个样本包含着数千个基因表达值,这些数据可用于疾病诊断和基因功能研究。如今,高维数据的处理受到了广泛的关注与研究。

深度学习通常用于处理高维数据。然而,在面对小样本高维数据时,深度学习会陷入过拟合的问题,导致泛化能力欠佳。宽度学习系统(Broad Learning System,BLS)是Chen等[1-2]提出的一种使用伪逆计算输出权重的单隐层神经网络。与深度学习不同,宽度学习系统不使用梯度更新权重,因此避免了梯度消失和梯度爆炸的风险。此外,BLS具有鲁棒的泛化能力和对于小数据集的高效训练方式。

到稿日期:2025-10-16　返修日期:2026-01-16
基金项目:国家自然科学基金(62572199,92467109,U21A20478);国家重点研发计划(2023YFA1011601)
This work was supported by the National Natural Science Foundation of China(62572199,92467109,U21A20478)and National Key R&D Program of China(2023YFA1011601).
通信作者:余志文(zhwyu@scut.edu.cn)

然而，当面对高维复杂数据时，结构简单的单隐层 BLS 性能不佳。因此，需要加深 BLS 的网络层数来提高其特征提取能力。由于 BLS 的节点权重和偏置都是随机生成的，机械式地堆叠 BLS 的隐藏层并不能提高其特征表达能力，过多的节点反而会产生冗余特征，降低训练效率。因此，本文使用 Stacking 的方式，设计了堆叠 BLS 模型，使用前一层 BLS 生成的输出作为增强特征，结合原始特征输入下一层 BLS，从而提升 BLS 对高维数据的特征提取能力。

高维数据通常具有复杂特征，而集成学习能够进一步提升 BLS 的泛化能力。Yun 等[3]分析了 BLS 和集成学习方法结合的可行性和有效性。然而，大多数现有的集成宽度学习方法为静态模型，难以灵活应对复杂多变的高维数据。在训练前，现有方法通常需要耗费时间搜索参数，并在训练时固定集成参数，如基学习器的数量。然而，不同数据集的分类难易程度不同，简单数据集不需要过多的基分类器即可达到令人满意的效果；而面对困难数据集，集成学习在基分类器不足时表现不佳。参数搜索费时费力，固定参数又难以使模型达到最优性能。因此，本文提出了动态模型框架，根据高维数据集的难度动态调整集成宽度模型的复杂度，从而保证所提模型能够在最小的开销下达到最优的分类性能。

本文主要贡献如下：

1）针对单层 BLS 面对高维数据时特征提取能力不足的问题，提出了 Stack BLS(SBLS)算法。现有的多层 BLS 算法简单地对隐层进行堆叠，然而，随机映射的隐层堆叠会生成大量冗余节点，难以提升模型性能；而所提出的 SBLS 结合原始特征和每层的输出特征进行训练，从而提升 BLS 的特征提取能力。

2）现有的多层 BLS 算法是单一模型，本文将 SBLS 算法的每一层作为一个基学习器进行有机集成，进一步提升模型性能。此外，设计了加权集成 Stack BLS 算法(E-SBLS)，根据先前层的 BLS 分类置信度为样本进行动态加权，从而提升模型处理复杂高维样本的能力。通过集成不同层 SBLS 的多样化输出，模型的泛化能力得到提升。

3）现有的 BLS 集成算法大多是单一静态的模型，然而，不同数据集所需的模型规模不同。针对高维数据复杂多样的特性，本文设计了动态集成框架，根据样本难度动态调整所提模型的复杂度，在保持模型优秀性能的同时降低训练开销。实验验证了所提算法动态集成 SBLS 算法(DE-SBLS)在高维小样本疾病数据集上的有效性。

2　相关工作

2.1　宽度学习系统

宽度学习系统是基于随机向量函数链接神经网络(RVFLNN)[4]提出的。如图 1 所示，BLS 的隐藏层包括两个部分：特征节点和增强节点。首先，输入数据经过线性映射变换为特征节点，然后经过非线性激活函数生成增强节点，这使得 BLS 同时具有线性和非线性特征表达能力。最后，使用伪逆直接计算输出权重，形成一种高效的训练模式。

BLS 通过增加特征节点来改进 RVFLNN 的隐藏层。原始数据不再直接映射为增强节点，而是先进行一次线性特征变换。假设 BLS 有 n 组特征节点，特征映射可以计算为：

$$Z_k=\varnothing_k(X W_k^z+\beta_k^z),k=1,2,\cdots,n \tag{1}$$

其中，X 为输入数据，$\varnothing_k$ 为线性激活函数，W_k^z 和 β_k^z 分别为随机生成的权重和偏置。为了解决随机性造成的节点冗余问题，BLS 首先使用稀疏自编码器进行特征提取，微调随机生成的权重和偏置，从而减少冗余特征，使特征表达更加紧凑。BLS 结合所有特征节点组，得到特征节点矩阵：

$$\boldsymbol{Z}_N=[Z_1,Z_2,\cdots,Z_n] \tag{2}$$

然后，生成增强节点。增强节点的计算过程与特征节点类似，不同之处在于增强节点是基于所有特征节点生成的。假设有 m 组增强节点，增强节点 E_l 可以计算为：

$$E_l=\varnothing_l(Z_N W_l^e+\beta_l^e),l=1,2,\cdots,m \tag{3}$$

其中，$\varnothing_l$ 为非线性激活函数，W_l^e 和 β_l^e 分别为随机生成的权重和偏置。针对随机矩阵中可能存在线性相关项的问题，对 W_l^e 进行正交归一化，使得矩阵中的每项均线性无关，进一步提高增强节点的非线性表达能力。结合所有增强节点，能够得到增强节点矩阵：

$$\boldsymbol{E}_M=[E_1,E_2,\cdots,E_m] \tag{4}$$

结合所有特征节点和增强节点，获得 BLS 的隐藏层矩阵 $\boldsymbol{H}=[\boldsymbol{Z}_N,\boldsymbol{E}_M]$。

BLS 的目标函数为最小化分类误差：

$$f(x)=\underset{W_o}{\operatorname{argmin}}\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_2^2 \tag{5}$$

因此，可以使用伪逆求解输出权重 $\boldsymbol{W}_o$：

$$\boldsymbol{W}_o=(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}+\lambda\mathbf{I})^{-1}\boldsymbol{H}^{+}\boldsymbol{Y} \tag{6}$$

其中，λ 为正则化系数。

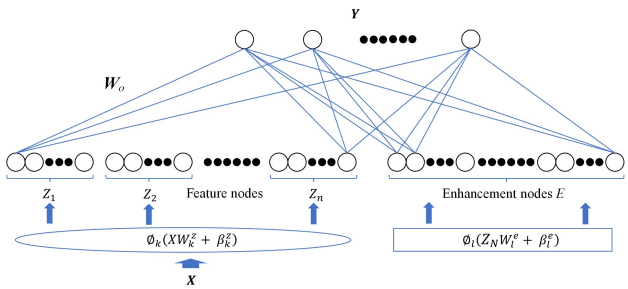


图 1　宽度学习系统结构

Fig. 1　Structure of broad learning system

BLS 的伪代码如算法 1 所示。

算法 1　宽度学习系统

输入：训练样本集 D={X,Y}，特征节点数 n，增强节点数 m，正则项系数 λ

输出：权重 $\mathbf{W}_o$

1. For k← 1 to n do
2.　随机生成特征节点权重 W_k^z 和偏置 β_k^z
3.　计算特征节点 Z_k
4. End For
5. 将所有特征节点拼接 $\mathrm{Z_N}=[\mathrm{Z}_1,\mathrm{Z}_2,\cdots,\mathrm{Z}_n]$
6. For l ←1 to m do
7.　随机生成增强节点权重 W_h^l 和偏置 β_h^l
8.　计算增强节点 E_l
9. End For
10. 拼接所有增强节点 $\mathrm{E_M}$

11. 拼接特征节点与增强节点，得到隐层权重矩阵 $\mathbf{H}=[\mathbf{Z}_N,\mathbf{E}_M]$，并计算伪逆$H^+$

12. 通过式(6)计算隐层到输出层的连接权重$\mathbf{W}_o$并保存

宽度学习系统在计算机领域备受关注，许多研究者将BLS与机器学习方法相结合[5]，包括迁移学习[6]、度量学习[7]、自编码器[8]和核函数[9-10]，并将BLS广泛应用于数据挖掘[1]、模式识别[12]、图像分类[13]、疾病检测[14]和工业控制[15]等多个领域。

2.2 多层宽度学习系统

为了提高BLS的特征提取能力，研究者通过堆叠和级联的方式加深BLS的网络结构。

堆叠通过纵向叠加多个隐藏层构建多层神经网络，各层之间全连接或局部连接，形成端到端的学习框架。Yang等[16]提出了用于工业网络入侵检测的堆叠OCBLS。OCBLS将BLS的隐藏节点进行水平式堆叠，从而提升了模型对复杂网络数据及多种类型入侵任务的处理性能与效率。Liu等[17]采用残差思想堆叠BLS，并提出了相应的增量学习机制。Xie等[18]通过堆叠多个轻量级BLS子系统来提高BLS的特征表示能力和分类性能。Xie等[19]以双层堆叠的方式整合BLS子系统，设计残差方法，使BLS子系统更加多样化。

一些研究人员则在BLS模型中级联隐层。Chen等[20]提出了5种级联BLS，包括宽度和深度上的级联，还加上了递归和卷积结构。对于聚类任务，Yu等[8]设计了3种BLS自编码器，分别级联特征节点、增强节点和所有隐藏节点，以解决单个BLS-AE中随机子空间引起的不稳定问题。Yi等[21]为BLS中的增强节点设计了一种组间级联结构ICBLS，将原始知识与当前信息相结合来确定预测结果，在保持BLS训练效率的同时，提高了混沌时间序列的动态特征提取能力。

现有的堆叠BLS方法虽提升了BLS的特征提取能力，但仍局限于单一的模型。假如将多层BLS的每一层作为基分类器，并整合多层输出，将获得更加准确和稳健的预测结果。综上所述，多层BLS适用于集成学习方法。因此，本文提出了一种新的集成堆叠BLS方法。

2.3 集成宽度学习系统

集成学习[22]是一种机器学习技术，通过组合多个基学习器来构建一个更加强大、准确的预测模型。作为基学习器，BLS已被应用于Bagging[23]，Boosting[24]和Stacking[25]等集成策略中。

基学习器越精确，越多样化，集成学习的效果越好。为了增加基学习器的多样性，集成BLS中经常使用Boostrap[26]等重采样技术。Yan等[27]在Bagging和Stacking框架中使用BLS作为基学习器，以提高模型的准确性和稳定性。Wu等[28]使用Bootstrap创建了基础BLS的子集，并通过Stacking整合初级学习器的结果。Fan等[29]使用BLS获得LncRNA蛋白质对中的不同特征，并通过Stacking输出蛋白质特征之间的关系。

与其他专注于并行组合基模型的集成方法不同，Boosting强调通过更多地关注错误分类的样本，来提高基模型的性能。Boosting算法迭代地训练基模型，后续的基模型侧重训练被先前模型错误分类的样本。Yu等[30]设计了基于核BLS的渐进式Boosting框架。PEKB使用梯度和次梯度解决了基学习器的损失问题，并将残差作为后续基学习器的训练标签。为了应对不平衡样本中异常值和噪声的挑战，Yang等[11]提出了一种加权BLS来分配和增加少数样本的权重。为了获得样本分布，AWBLS设计了基于密度的权重生成机制，同时考虑类间和类内距离。最后，提出的IWEB算法通过渐进式Boosting方法增强了AWBLS的稳定性和鲁棒性。

除了通用算法外，集成BLS被广泛应用于不同领域，特别是异常检测任务。Zhao等[31]提出了一种用于变压器故障诊断的AdaBoost加权BLS，该方法使用成本敏感矩阵来处理不平衡数据。Lin等[15]提出了SMC-OCBLS，并设计了用于不平衡网络流量数据的Boosting框架。SMC-OCBLS通过随机化特征来构建不同的特征空间，以训练多个OCBLS，从而提高了模型性能和泛化能力。

除了使用BLS作为基学习器外，一些研究还在BLS的节点生成方式中使用了集成学习方法，如Bootstrap和Stacking。Gao等[32]在生成特征节点时使用了Bootstrap方法和决策树方法。Xie等[18]使用Boostrap策略生成增强节点，在提高性能的同时轻量化BLS的网络结构。通过整合残差，D&BLS堆叠了几个轻量级的BLS子系统，以确保更强的特征表示能力和更好的分类性能。此外，Xie等[19]使用了简单的线性模型来在相邻的BLS子系统之间传输共享特征节点。然后，RST&BLS以双堆叠的方式集成这些子系统，通过残差使BLS子系统更加多样化。

3 动态集成堆叠宽度学习系统

所提动态集成堆叠方法分为3个部分。首先，提出了一种SBLS算法，以提高单层BLS的特征提取能力。其次，为了处理不同难度的数据，提出了一种加权机制，使得多层BLS能够逐层提取到更加困难的特征，使用集成学习方法来整合多层SBLS的预测。最后，设计了动态集成框架，根据数据集的难易程度动态调整模型复杂度，最小化时空开销。

3.1 堆叠宽度学习系统

作为单层神经网络，宽度学习系统通过增量方法扩展网络架构，从而提高其特征提取能力。然而，隐藏层的扩展存在着性能上限。当达到上限时，隐层节点的增加会导致节点冗余，使得BLS的性能难以提高。特别是在面对高维数据时，BLS及其增量形式并不令人满意。因此，本文提出了堆叠BLS，通过加深网络来提高BLS的特征映射能力。

BLS的节点随机生成机制导致其在直接堆叠隐藏层时会造成节点冗余。因此，使用前一层的BLS输出作为增强特征，结合原始数据特征输入下一层来完成BLS的堆叠，从而提高BLS对高维数据的特征提取能力。

图2展示了所提堆叠宽度学习系统(Stack BLS，SBLS)的网络结构。其中，第一层将原始数据$X_1=X$作为特征输入网络进行训练，如算法2所示。将第一层的输出Y_1与原始输入X合并作为增强特征，输入至第二层堆叠BLS进行训练。

$$X_2=[X_1,Y_1] \tag{7}$$

假设堆叠 BLS 共有 h 层，将前$(k-1)$层的输出和原始数据作为第 k 层的输入。

$$X_k=[X_{k-1},Y_{k-1}]=[X,Y_2,\cdots,Y_{k-1}],k>1 \tag{8}$$

通过堆叠的方式，来自先前层的信息可以逐层传输到堆叠 BLS 的后层，提高了所提算法对高维数据的特征提取能力。

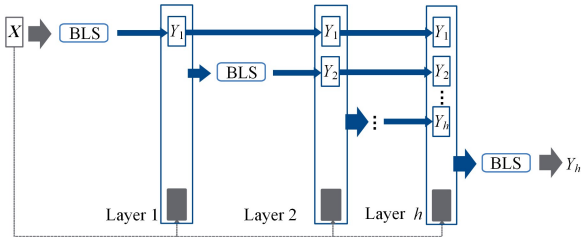


图 2　堆叠宽度学习系统的网络结构

Fig. 2　Structure of stack broad learning system

算法 2　堆叠宽度学习系统

输入：训练样本集 D={X,Y}，层数 h

输出：预测结果Y_h

1. 将 X 输入第一层 BLS 进行训练，得到预测结果Y_1

2. For k← 2 to h do

3.　根据式(8)生成X_k

4.　训练第 k 层 BLS，得到预测结果Y_k

5. End For

6. 输出最后一层 BLS 预测结果Y_h并保存

3.2　集成堆叠宽度学习系统

在所提的堆叠宽度学习系统中，每一层都将原始数据作为输入的一部分。然而，数据集中不同数据的分类难易程度不同，简单数据能够在浅层就被正确分类，而困难数据需要经过多层训练才能被正确分类。为了更有效地处理困难数据，本文提出了一种数据加权方法，通过对困难数据赋予更高的权重，来提升模型对困难数据的关注度，从而提高模型整体的分类性能。

对于堆叠 BLS 第一层，依旧将原始数据 X 作为输入，输出第一层的预测结果Y_1。然后，将输出分为两部分，作为困难数据子集：将分类正确的样本置 0，其他置 1。

$$d_k^i=\begin{cases}1, & \text{if } Y_k^i\neq Y^i\\ 0, & \text{if } Y_k^i=Y^i\end{cases} \tag{9}$$

假设数据个数为 s，第 k 层堆叠 BLS 的误差可计算为：

$$e_k=\frac{1}{s}\sum_{i=1}^{s}d_k^i \tag{10}$$

根据该层的预测结果Y_k，赋予错分的数据更高的权重。下一层的数据权重可计算为：

$$w_{k+1}^i=w_k^i\mathrm{e}^{c_k d_k^i} \tag{11}$$

其中，c_k是输出Y_k的置信度。

$$c_k=\frac{1}{2}\log\frac{1-e_k}{e_k}+\log(R-1) \tag{12}$$

其中，R 是类别数量。权重赋值提高了堆叠 BLS 对复杂高维数据的特征提取能力。

然后，归一化权重：

$$w_{k+1}^i=\frac{w_{k+1}^i}{z_k} \tag{13}$$

其中，z_k是归一化因子。

最后，得到下一层堆叠 BLS 的输入：

$$X_{k+1}=[w_kX_k,c_kY_k] \tag{14}$$

堆叠 BLS 逐层累积预测结果。通过整合每层的输出，可以获得性能更优、更稳定的集成结果。

在集成学习中，基分类器的性能越高且多样性越好，集成效果就越好。将堆叠 BLS 的每层视作一个基学习器，所提的加权机制能够提升堆叠 BLS 的多样性。此外，持续增加的特征也能够提升堆叠 BLS 在更深层网络中的性能。因此，基于多样且精确的堆叠 BLS 层，整合多层堆叠 BLS 以形成更强大而高效的集成学习方法是可行的。

与深度学习类似，堆叠的 BLS 层数越深，所包含的信息量越大，提取的特征也越全面。因此，在最终的集成模型中，为更深层的堆叠 BLS 赋予更大的权重。第 k 层的集成决策权重可以计算为：

$$a_k=kc_k \tag{15}$$

集成权重与精度和深度成正相关。归一化集成权重a_k：

$$a_k=\frac{a_k}{\sum_{k=1}^{h}a_k} \tag{16}$$

最后，堆叠 BLS 的加权集成输出计算为：

$$\hat{Y}=\sum_{k=1}^{h}a_kY_k \tag{17}$$

图 3 展示了集成堆叠 BLS（Ensemble-Stack BLS，E-SBLS）的结构，其中×表示加权操作。

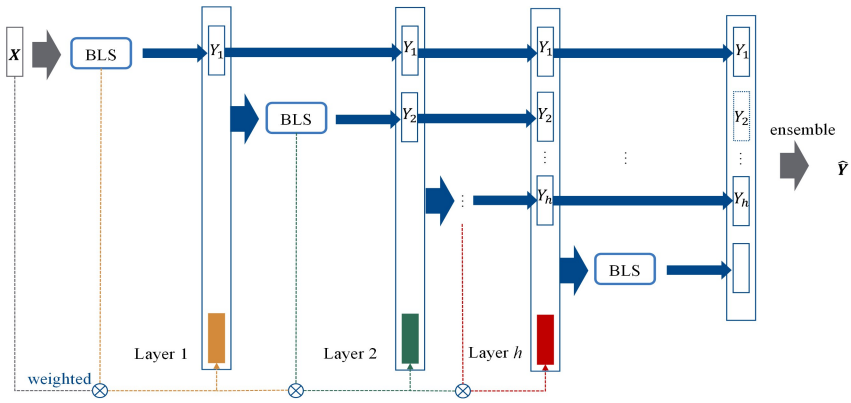


图 3　集成堆叠宽度学习系统结构

Fig. 3　Structure of ensemble-stack broad learning system

3.3 动态集成堆叠宽度学习系统

根据不同数据的难易程度，所提的集成堆叠 BLS 通过逐层为训练样本分配不同的权重，来调整每层对不同难度样本的侧重点，从而提升堆叠 BLS 处理困难高维数据的能力。事实上，不同数据集的难易程度及其所需的计算资源也存在着差异。在训练每个数据集前，传统方法通常需要进行网格搜索才能选到最佳参数，费时费力。节约简单数据的计算成本，并为复杂数据分配更多资源，是大数据时代模型的发展方向。

因此，本文进一步提出了一种动态集成框架，根据不同难度的数据动态调整集成堆叠 BLS 的复杂度，从而在保持模型性能的同时降低时空开销。

具体来说，在堆叠 BLS 的每层均设置出口，从训练集 X 中划分出一个独立的验证数据集 X^v。集成模型的效果取决于基分类器的准确性和多样性。因此，计算混淆矩阵，使用性能衡量指标：

$$f_k=(1-e_k)g_k \tag{18}$$

其中，e_k 是第 k 层的输出误差，反映模型的准确性；g_k 是前 k 层的熵，反映模型的多样性。

$$g_k=\frac{1}{s}\sum_{i=1}^{s}\frac{\min(\rho(x_i),k-\rho(x_i))}{\left\lfloor\frac{T}{2}\right\rfloor} \tag{19}$$

其中，$\rho(x)$ 表示 k 个基分类器对数据 x 分类正确的数量，$\rho(x)\in[0,k]$。

$$\rho(x)=\frac{1}{k}\sum_{i=1}^{k}(1-g_k^i) \tag{20}$$

在模型训练过程中，当连续多层迭代中验证集的准确性和多样性没有提高时，在最优层输出堆叠 BLS 的预测结果，提前停止训练，以避免过拟合问题和不必要的计算资源浪费。

最后，设计分层优化集成框架，以进一步提高所提方法的泛化能力。将堆叠 BLS 的多层输出与原始特征输入元学习器，将这些专注于不同难度数据的分类结果进行高效准确的选择和组合，以提升集成模型在新数据上的泛化能力。元学习器的输入为：

$$X_m=[X,Y_1,Y_2,\cdots,Y_h] \tag{21}$$

常用的元学习器包括逻辑回归(Logistic Regression，LR)[33]、决策树(Decision Tree，DT)[34]、随机森林(Random Forest，RF)[35]和 K 近邻(K-Nearest Neighborhood，KNN)[36]。不同的元学习器侧重点不同，在不同数据集中表现各异。LR 假设数据线性可分，适合低维线性数据；DT 无分布假设，适合低维非线性数据；RF 通过集成 DT，降低过拟合风险，适合高维复杂数据；KNN 则假设样本具有局部相似性，适合低维局部稠密特征。将 4 种元学习器组合成元学习器池，使用元学习池整合堆叠 BLS 的预测结果，根据验证集的性能指标动态选择合适的元学习器来适应多层预测结果。

图 4 展示了动态集成堆叠 BLS(Dynamic Ensemble-Stack BLS，DE-SBLS)的框架。所提算法的训练流程如算法 3 所示，多层堆叠 BLS 和元学习器池共同构建了一个更强大、动态自适应的集成堆叠 BLS。

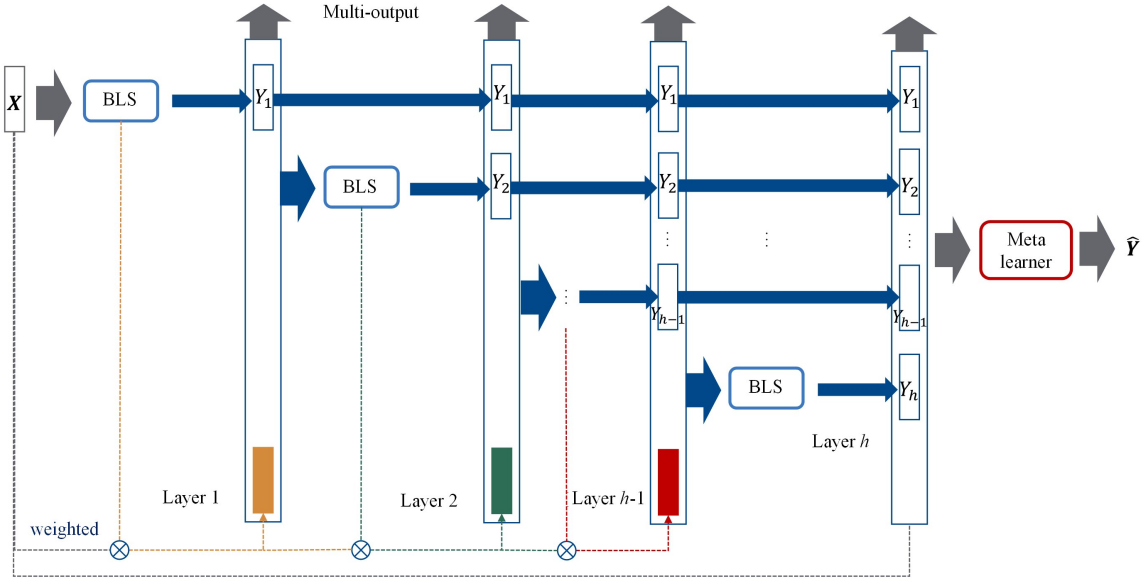


图 4 动态集成堆叠宽度学习系统结构

Fig. 4 Structure of dynamic ensemble-stack BLS

算法 3 动态集成堆叠宽度学习系统

输入：训练样本集 D={X,Y}

输出：集成预测结果 $\hat{Y}$

1. 将 X 输入第一层 BLS 进行训练，得到预测结果 Y_1，性能置信度 e_1，权重 a_1
2. While $e_{k-1}<e_k$ do
3. 根据式(13)计算 w_k
4. 根据式(14)生成 x_k
5. 训练第 k 层 BLS，得到预测结果 Y_k，性能置信度 e_k，权重 a_k
6. End While
7. 结合所有层预测结果与原始数据，将 X_m 输入元学习器池进行训练
8. 选择最优元学习器，得到集成预测 $\hat{Y}$

4 实验与分析

实验一共使用了 12 个高维疾病数据集[26]，其中包括不同规模的二分类和多分类数据集。数据集信息如表 1 所列。

表 1　数据集

Table 1　Datasets

数据集	样本数	特征数	类别数
ALLAML	72	7129	2
CLLSUB	111	11340	3
GLI-85	85	22283	2
GLIOMA	50	4434	4
lung	203	3312	5
lung_small	73	325	7
orlraws10P	80	10334	10
PCMAC	1943	3289	2
pixraw10P	80	10000	10
ProstateGE	102	5966	2
TOX-171	171	5748	4
warpPIE10P	210	2420	10

所有实验在 CPU 为 12th Gen Intel$^{(R)}$ Core$^{(TM)}$ i5-12400F，2496 MHz，内存为 16.0 GB 的计算机上运行，程序版本为 Python 3.7。实验使用精度作为性能评价指标：

$$acc=\frac{1}{s}\sum_{i=1}^{s}I(\hat{Y}=Y) \tag{22}$$

其中，s 为样本数，$I(\cdot)$ 为指示函数。所有算法经过 10 次重复实验求平均值得到的百分比作为最终结果。

4.1　参数实验

BLS 的基本参数经网格搜索法确定，如表 2 所列。特征节点数 n_1 和 n_2 的搜索空间 $k_1=[2,15]$，增强节点数 m 的搜索空间 $k_2=[20,800]$，步长均为 1。实验采用一次性划分的验证策略。网格搜索的时间复杂度为 $O(k_1*k_1*k_2)$。

表 2　参数设置

Table 2　Parameters setting

数据集	特征节点组数 n_1	每组特征节点数 n_2	增强节点数 m
ALLAML	2	5	39
CLLSUB	4	7	125
GLI-85	3	14	168
GLIOMA	5	2	181
lung	8	1	311
lung_small	11	2	502
orlraws10P	4	11	578
PCMAC	14	11	726
pixraw10P	5	2	153
ProstateGE	2	6	106
TOX-171	10	14	557
warpPIE10P	9	3	118

此外，还需要确定堆叠 BLS 的层数。固定表 2 的 BLS 基本参数，在其中 6 个数据集上进行了层数实验，结果如表 3 所列，其中最优值用粗体表示。实验表明，不同数据集根据其难易程度的不同，对应 BLS 网络所需的节点数和网络层数不同。因此，根据数据的难度动态调整模型复杂度是提高训练效率的必要条件。

表 3　堆叠层数参数实验（测试精度百分比）

Table 3　Stack layer parameter experiment(test acc)

（%）

数据集	1	2	3	4	5	6	7	8	9	10
CLLSUB	57.9710	60.8693	59.4203	60.8696	60.8696	55.0723	73.5553	**83.0435**	78.2608	69.5652
lung	67.3174	75.1221	74.6343	74.1465	78.0489	73.1710	**96.3415**	94.6343	90.2439	92.6829
lungsmall	40.0002	48.0000	41.3332	46.6666	46.6666	55.7333	46.6664	43.3332	**93.3333**	86.6667
PCMAC	86.1826	87.2428	88.0462	87.8535	87.8536	86.3110	**91.3239**	89.4602	88.6889	88.9460
TOX-171	75.5102	75.9180	73.8775	80.0000	75.5101	83.6734	76.3264	77.1428	**97.1428**	97.1428
warpPIE	85.7140	64.2860	64.2860	85.7140	73.8100	78.5710	71.4290	97.61905	**100.0000**	**100.0000**

4.2　消融实验

本文在 BLS 的基础上提出了堆叠 BLS、集成堆叠 BLS 和动态集成堆叠 BLS。在 6 个数据集上进行的消融实验如表 4 所列，其中最优值用粗体表示。

表 4　消融实验（测试精度百分比）

Table 4　Ablation experiment(test acc)

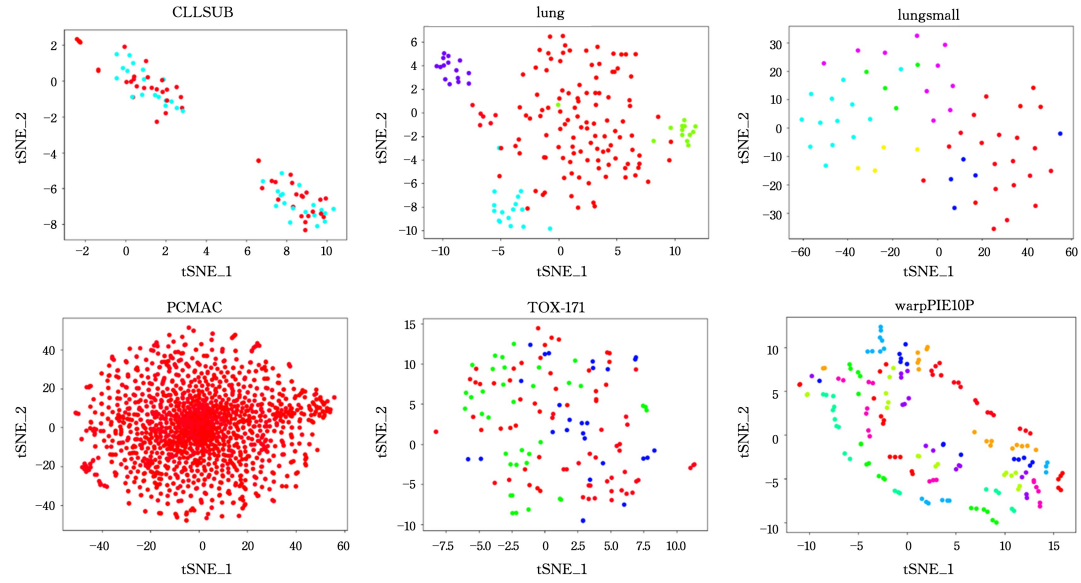
（%）

数据集	BLS	SBLS	E-SBLS	DE-SBLS
CLLSUB	60.9783	67.7019	73.2919	**83.0434**
lung	78.4555	80.0002	93.0313	**96.3414**
lungsmall	44.5333	56.1904	80.9523	**92.2222**
PCMAC	85.7875	87.4357	89.1663	**91.3239**
TOX	65.3946	77.2449	81.6326	**97.1429**
warpPIE	73.8096	84.0473	97.9591	**100.0000**

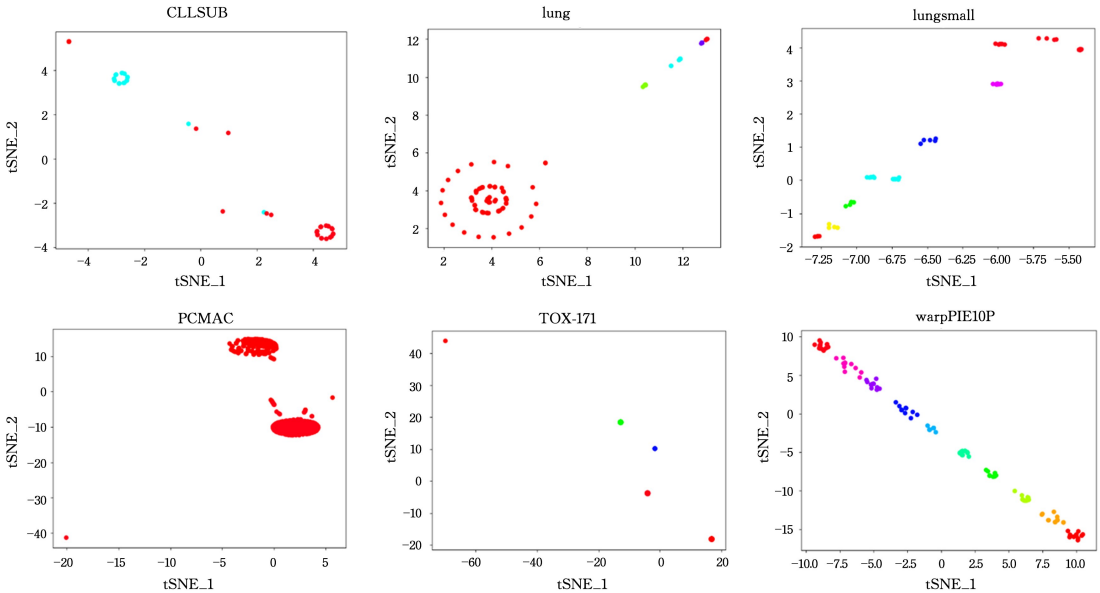
实验表明，所提集成堆叠框架提高了 BLS 的特征提取能力，所提算法性能均优于基础算法 BLS。从性能排名来看，集成堆叠 BLS 的性能优于堆叠 BLS，说明集成框架能够整合各层堆叠 BLS 的优势，提高算法性能；样本加权机制能够提高算法对困难数据集的特征提取能力。动态集成堆叠 BLS 的性能最优，说明元学习方法能够让算法性能更胜一筹。综上所述，所提算法的各个模块循序渐进，相辅相成，逐步提升了 BLS 的性能。

本文使用 t-SNE 嵌入可视化模型训练前后的特征。如图 5 所示，不同颜色的点代表不同类的样本。图 5(a)中的原始数据特征散乱排布；而图 5(b)经过 DE-SBLS 最后一层提取的特征呈现出明显的分块现象，同色同类样本相聚，异色异类样本分离，证明了所提算法具有优秀的特征提取能力。

如图 6 所示，模型运行时间与堆叠 BLS 的层数呈正相关，堆叠 BLS 的层数越多，模型耗时越长。因此，在合适的层数停止堆叠是提高模型效率的关键。为了验证所提动态框架的高效性，对堆叠集成 BLS 和动态堆叠集成 BLS 的实验时间进行了对比，结果如表 5 所列。其中，E-SBLS 的层数统一设置为 10，DE-SBLS 根据数据集难度自适应调整层数和模型复杂度，具体最优层数如表 3 所列。实验表明，动态堆叠集成 BLS 的耗时短，性能高，效率较堆叠集成 BLS 提升了 7%～28%，证明了所提动态框架的有效性。



(a)原始特征



(b)DE-SBLS 提取后特征

图 5 t-SNE 嵌入图(电子版为彩图)

Fig. 5 t-SNE embedding graph

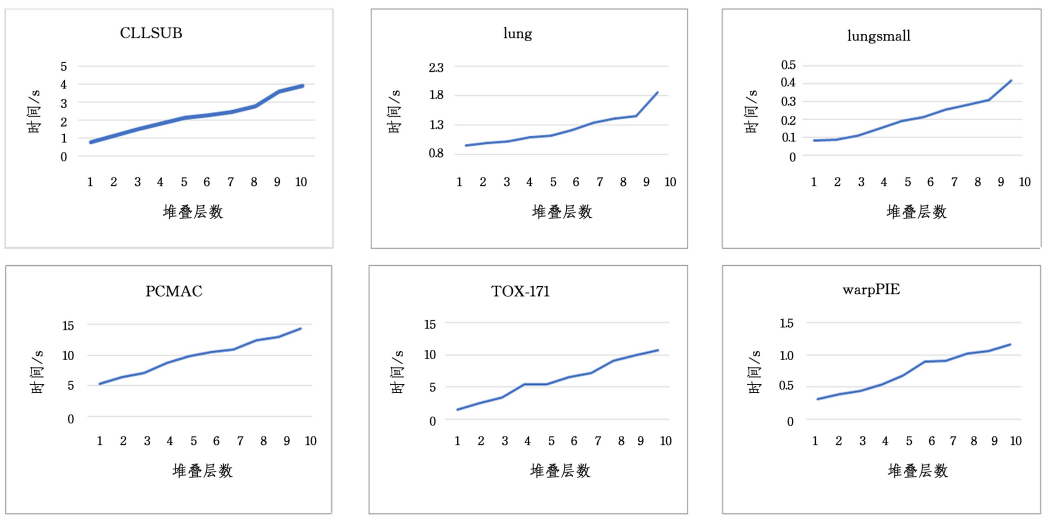


图 6 运行时间随层数变化曲线

Fig. 6 Curves of running time changing with the number of stack layers

表 5　实验运行时间

Table 5　Experiment runtime

(s)

数据集	集成堆叠 BLS	动态集成堆叠 BLS
CLLSUB	3.9189	**2.7875**
lung	1.8568	**1.3375**
lungsmall	0.4190	**0.3077**
PCMAC	14.2097	**10.8846**
TOX-171	10.6389	**9.9373**
warpPIE	1.1618	**1.0638**

4.3　对比实验

将所提算法与 10 种分类算法和集成算法进行对比，包括 BLS、逻辑回归、支持向量机（Support Vector Machine，SVM）[37]、K 近邻、随机森林、GCForest[38] 和 xgboost[39]，以及针对高维数据的方法，包括有监督降维方法线性判别分析（Linear Discriminant Analysis，LDA）、无监督降维方法主成分分析（Principal Components Analysis，PCA）和集成方法 AdaSPEL[40]。对比结果如表 6 所列，所提算法在 10/12 个数据集上获得了最优结果，平均精度达到最优。实验证明了所提动态集成堆叠 BLS 算法在高维数据分类任务上的有效性。

表 6　对比实验（测试精度百分比）

Table 6　Comparative experiment(test acc)

(%)

数据集	ours	BLS	LR	SVM	KNN	RF	GCForest	xgboost	LDA	PCA	AdaSPEL
ALLAML	**100.0000**	71.3333	97.7778	98.6667	83.3333	93.3333	78.2617	85.3333	93.3333	96.6667	97.7778
CLLSUB	**83.0435**	60.0000	70.4348	71.1957	54.3478	72.2826	73.9130	67.6288	69.5652	82.6087	78.2608
GLI-85	**94.1176**	88.5294	91.1765	83.0882	87.2549	87.3949	77.3920	82.3529	88.2352	85.2941	91.1764
GLIOMA	78.0000	38.0000	42.0000	76.2500	70.0000	72.5000	77.0000	63.7500	70.0000	75.0000	**80.0000**
lung	**96.3415**	78.4555	94.6341	96.0976	95.6098	91.7073	94.6320	88.0488	95.1219	95.9349	95.6600
lung_small	**93.3333**	44.5333	91.1111	90.0000	88.8889	82.2222	79.9996	68.0000	78.6667	93.3333	89.9999
orlraws10P	98.7500	34.5000	97.7778	**99.0000**	95.5000	98.0000	96.5000	50.0000	96.8750	96.2500	97.5000
PCMAC	**91.3239**	87.4357	90.8740	90.4627	70.8483	91.1311	84.3480	89.6915	73.7789	87.4036	87.9177
pixraw10P	**100.0000**	68.0000	99.5000	97.7778	99.0000	99.5000	98.0000	66.0000	98.3333	**100.0000**	**100.0000**
ProstateGE	**94.2857**	93.8094	93.3333	90.0000	80.0000	90.4761	77.8260	83.3333	90.4761	92.8571	93.6507
TOX-171	**97.1429**	65.3946	85.7142	96.6667	79.5238	76.6667	83.1525	64.0000	68.5714	94.2857	91.4285
warpPIE10P	**100.0000**	84.0473	99.7619	99.5238	97.1428	98.0952	73.1883	69.8412	**100.0000**	96.1538	98.0769
AVG	**93.8615**	67.8365	87.8413	90.7274	83.4541	87.7758	82.8511	73.1650	85.2464	91.3156	91.7874

结束语　本文针对高维数据的复杂特性，改进了单层 BLS 算法，提出了堆叠宽度学习系统，通过堆叠结构提取高维数据更复杂的特征，提高对高维数据分类的准确性；为了进一步提高所提堆叠 BLS 的泛化性，将每层 BLS 视为一个基学习器，设计了集成堆叠宽度学习系统；最后，针对不同高维数据集所需的计算资源不同的问题，设计了动态集成堆叠宽度学习系统，根据数据集的难易程度动态调整模型复杂度。消融实验证明了所提算法的 3 个方法环环相扣，层层递进，性能和效率逐步提升，最终形成了本文所提出的动态集成堆叠宽度学习系统，提高了 BLS 对高维数据的特征提取能力、分类泛化能力和模型效率。与其他经典分类算法对比，所提算法在高维疾病数据集上表现优秀，其可行性和有效性得到证明。

未来拟将算法扩展到不同领域的数据，如设计不同的加权方法处理不平衡数据。

参 考 文 献

[1] CHEN C L P, LIU Z. Broad learning system: A new learning paradigm and system without going deep[C]// Proceedings of 32nd Youth Academicannual Conference of Chinese Association of Automation. IEEE, 2017: 1271-1276.

[2] CHEN C L P, LIU Z. Broad learning system: An effective and efficient incremental learning system without the need for deep architecture[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(1): 10-24.

[3] YUN F, YU Z, YANG K, et al. Adaboost-stacking based on incremental broad learning system[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36: 7585-7599.

[4] PAO Y H, TAKEFUJI Y. Functional-link Net Computing: Theory, System Architecture, and Functionalities[J]. Computer, 1992, 25: 76-79.

[5] YUN F, YU Z, YANG K, Broad Learning System in Data Mining and Machine Lear- ning[J]. Data Mining and Machine Learning, 2025, 1(1): 100002.

[6] ZHANG L, YU Z, YANG K, et al. Transferable and discriminative broad network for unsupervised domain adaptation[J]. Knowledge-Based Systems, 2025, 315: 113297.

[7] HU X, CHEN C L P, ZHANG T. Broad metric learning: A fast and efficient discriminative metric learning model[J]. IEEE Transactions on Cybernetics, 2010, 55(10): 14.

[8] YU Z, ZHONG Z, YANG K, et al. Broad learning autoencoder with graph structure for data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(1): 49-61.

[9] CHEN W, YANG K, ZHANG W, et al. Double-kernelized weighted broad learning system for imbalanced data[J]. Neural Computing and Applications, 2022, 34(22): 19923-19936.

[10] CHEN W, YANG K, YU Z, et al. Double-kernel based class-specific broad learning system for multiclass imbalance learning[J]. Knowledge-Based Systems, 2022, 253: 109535.

[11] YANG K, YU Z, CHEN C L P, et al. Incremental weighted ensemble broad learning system for imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 34: 5809-5824.

[12] ZHANG T, LIU Z, WANG X, et al. Facial expression recognition via broad learning system[C]// Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. IEEE,

2018:1898-1902.
[13] LEI C,GUO J,CHEN C L P. Convbls:An effective and efficient incremental convolutional broad learning system combining deep and broad representations[J]. IEEE Transactions on Artificial Intelligence,2024,5(10):5075-5089.
[14] YUN F,YU Z,YANG K,et al. Ensemble multi-kernel broad learning system for disease diagnosis[C]// Proceedings of IEEE International Conference on Medical Artificial Intelligence. IEEE,2024:97-104.
[15] LIN M,YANG K,YU Z,et al. Hybrid ensemble broad learning system for network intrusion detection[J]. IEEE Transactions on Industrial Informatics,2024,20(4):5622-5633.
[16] YANG K,SHI Y,YU Z,et al. Stacked one-class broad learning system for intrusion detection in industry 4.0[J]. IEEE Transactions on Industrial Informatics,2023,19:251-260.
[17] LIU Z,CHEN C L P,FENG S,et al. Stacked broad learning system:From incremental flatted structure to deep model[J]. IEEE Transactions on Systems,Man,and Cybernetics:Systems,2020,51:209-222.
[18] XIE R,WANG S. Downsizing and enhancing broad learning systems by feature augmentation and residuals boosting[J]. Complex & Intelligent Systems,2020,6:411-429.
[19] XIE R,VONG C,CHEN C L P,et al. Dynamic network structure:Doubly stacking broad learning systems with residuals and simpler linear model transmission[J]. IEEE Transactions on Emerging Topics in Computational Intelligence,2022,6:1378-1395.
[20] CHEN C L P,LIU Z,FENG S. Universal Approximation Capability of Broad Learning System and Its Structural Variations [J]. IEEE Transactions on Neural Networks and Learning Systems,2019,30:1191-1204.
[21] YI J,HUANG J,ZHOU W,et al. Intergroup cascade broad learning system with optimized parameters for chaotic time series prediction[J]. IEEE Transactions on Artificial Intelligence,2022,3:709-721.
[22] DONG X,YU Z,CAO W,et al. A survey on ensemble learning [J]. Frontiers of Computer Science,2020,14(2):241-258.
[23] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996,24(2):123-140.
[24] HASTIE T,ROSSET S,ZHU J,et al. Multi-class adaboost [J]. Statistics and its Interface,2009,2(3):349-360.
[25] WOLPERT D. Stacked Generalization [J]. Neural Networks,1992,5:241-259.
[26] LI J,CHENG K,WANG S,et al. Feature Selection:A Data Perspective[J]. ACM Computing Surveys,2016,50(6).
[27] YAN Y,GUO W,WANG L. Broad learning system based on ensemble learning[C]// Proceedings of International Conference on Artificial Intelligence,Big Data and Algorithms. Chengdu:IEEE Press,2021:62-67.
[28] WU C,QIU T,ZHANG C,et al. Ensemble strategy utilizing a broad learning system for indoor fingerprint localization[J]. IEEE Internet of Things Journal,2021,9(4):3011-3022.
[29] FAN X,ZHANG S. Lpi-bls:Predicting lncrna-protein interactions with a broad learning system-based stacked ensemble classifier[J]. Neurocomputing,2019,370:88-93.
[30] YU Z,LAN K,LIU Z,et al. Progressive ensemble kernel based broad learning system for noisy data classification[J]. IEEE Transactions on Cybernetics,2022,52(9):9656-9669.
[31] ZHAO B,YANG D,ZBSG H R. Filter-wrapper combined feature selection and adaboost-weighted broad learning system for transformer fault diagnosis under imbalanced samples[J]. Neurocomputing,2023,560(1):1.1-1.12.
[32] GAO Y,MIAO H,CHEN C L P,et al. Explosive cyber security threats during covid-19 pandemic and a novel tree-based broad learning system to overcome[J]. IEEE Transactions on Intelligent Transportation Systems,2022,25(1):1-10.
[33] COX D. The Regression Analysis of Binary Sequences[J]. Journal of the Royal Statistical Society. Series B:Methodological,1958,20(2):215-242.
[34] QUINLAN J. Induction of decision trees[J]. Machine Learning,1986,1(1):81-106.
[35] BREIMAN L. Random forests [J]. Machine Learning,2001,45(1):5-32.
[36] ZHANG M,ZHOU Z. Ml-knn:A lazy learning approach to multi-label learning[J]. Pattern Recognition,2007,40(7):2038-2048.
[37] HEARST M,DUMAIS S,OSUNA E,et al. Support vector machines[J]. IEEE Intelligent Systems and their applications,1998,13(4):18-28.
[38] ZHOU Z,FENG J. Deep forest:Towards an alternative to deep neural networks[C]// Proceedings of International Joint Conference on Artificial Intelligence. Melbourne:Morgan Kaufmann,2017:3553-3559.
[39] CHEN T,GUESTRIN C. Xgboost:A scalable tree boosting system[C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco:ACM,2016:785-794.
[40] XU Y,YU Z,CAO W,et al. A Novel Classifier Ensemble Method Based on Subspace Enhancement for High-Dimensional Data Classification[J]. IEEE Transactions on Knowledge and Data Engineering,2023,35(1):16-30.

YUN Fan, born in 2000, postgraduate, is a member of CCF(No. K4411M). Her main research interests include computer science, artificial intelligence, machine learning, data mining, broad learning system and ensemble learning.

YU Zhiwen, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of CCF(No. 16933D). His main research interests include artificial intelligence, machine learning and data mining.

(责任编辑:柯颖)