·人工智能与模式识别·　　文章编号：1000-3428(2022)04-0099-07　　文献标志码：A　　中图分类号：TP391

基于AdaBoost集成加权宽度学习系统的不平衡数据分类

王萌铎，续欣莹，阎高伟，史丽娟，郭　磊

（太原理工大学 电气与动力工程学院，太原 030024）

摘　要：宽度学习系统(BLS)是一种浅层的神经网络结构，具有快速训练、增量学习等特征，在处理类别不平衡数据时提取到的少数类别特征较少，导致识别结果不理想。提出一种基于AdaBoost集成加权宽度学习系统（AdaBoost-WBLS）的不平衡数据分类方法，通过迭代实现权重的动态更新，获得更符合数据特征的权重，提升集成模型对少数类的识别能力。基于KKT条件，对加权宽度学习系统的加权优化过程进行推导，验证了对角权重对BLS模型误差的抑制作用。在AdaBoost-WBLS模型集成初始化时，采用基于类别信息的初始化权值策略，使模型具有更高的集成训练效率。在集成权重更新时，不同数据类别采用不同的正则化更新方式，保留数据的类内特征并增加类间区分度。在实验过程中，对AdaBoost-WBLS模型的不同参数进行寻优，得到相关参数在有限范围内的最优取值。实验结果表明，AdaBoost-WBLS模型相比AdaBoost和BLS类相关模型能有效改善少数类别特征的提取能力，并且在Satimage数据集上相比加权过采样的深度自编码器模型的G-mean高出4.36个百分点，明显提升了不平衡数据的识别能力。

关键词：宽度学习系统；AdaBoost模型；不平衡数据；加权宽度学习系统；集成学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：王萌铎，续欣莹，阎高伟，等. 基于AdaBoost集成加权宽度学习系统的不平衡数据分类[J]. 计算机工程，2022，48(4)：99-105，112.

英文引用格式：WANG M D，XU X Y，YAN G W，et al. Imbalanced data classification based on ensemble weighted broad learning system with AdaBoost[J]. Computer Engineering，2022，48(4)：99-105，112.

Imbalanced Data Classification Based on Ensemble Weighted Broad Learning System with AdaBoost

WANG Mengduo，XU Xinying，YAN Gaowei，SHI Lijuan，GUO Lei

（School of Electrical and Power Engineering，Taiyuan University of Technology，Taiyuan 030024，China）

【Abstract】Broad Learning System（BLS）is a novel shallow network structure having advantages such as rapid training and incremental learning. When dealing with imbalanced data，BLS extracts fewer minority class features，which can reduce the performance of the these classes. To solve this problem，this study proposes an imbalanced data classification method based on the ensemble Weighted Broad Learning System（WBLS）with AdaBoost（AdaBoost-WBLS）to improve the recognition ability of minority classes through dynamic updating of weights，to better match the characteristics of the data. Based on the KKT condition，the weighting optimization process of WBLS is derived theoretically to verify the inhibition effect of the diagonal weights on BLS errors. The initialization of AdaBoost-WBLS is based on category information，which can increase the ensemble training efficiency of the model. In the process of weight updating，different regularized updating modes are adopted according to the different data categories，not only to retain the features within the classes but also to increase the degree of distinction between the classes. In this study，many experiments are carried out on the AdaBoost-WBLS model with the parameters of different data optimized in a limited range. The experimental results show that，compared with both AdaBoost- and BLS-related models，the AdaBoost-WBLS model improves the extraction feature ability of minority classes. On the Satimage dataset，the G-mean of the AdaBoost-WBLS model is 4.36 percentage points higher than that of the Weighted Minority Oversampling Deep Auto-encoder（WMODA）model，which shows that the recognition ability of the AdaBoost-WBLS model for imbalanced data is significantly improved.

【Key words】Broad Learning System（BLS）；AdaBoost model；imbalanced data；Weighted Broad Learning System（WBLS）；ensemble learning

DOI：10. 19678/j. issn. 1000-3428. 0061001

基金项目：国家自然科学基金面上项目(61973226)；山西省自然科学基金(201801D121144)。

作者简介：王萌铎(1996—)，男，硕士研究生，主研方向为多模态融合、宽度学习系统；续欣莹，教授、博士；阎高伟，教授、博士、博士生导师；史丽娟，硕士研究生；郭　磊，博士研究生。

收稿日期：2021-03-03　　**修回日期：**2021-05-08　　**E-mail：**wangmengduo66@163.com

0 概述

在故障诊断、金融诈骗[1-3]等分类任务中，数据分布通常是不平衡的，类别分布极端时就会形成不平衡数据集。由于少数类别的数据数量相对较少，对准确率的影响也相对较小[4]。在处理不平衡数据集时，目标识别模型受数据自身分布制约学习到的多数类类别特征更多且忽视了少数类别。数据类别分布不平衡现象制约了模型对少数类别目标的识别性能[5-6]。

针对不平衡数据，ZHANG等[7]提出一种使用新保角函数扩展最优间隔分布机（Optimal-margin Distribution Machine，ODM）核矩阵以提高特征空间可分性的不平衡分类方法（Kernel Modified ODM，KMODM）。ZHU等[8]提出一种类权重随机森林（Class Weights Random Forest，CWsRF）算法，用于处理医学数据的不平衡分类。SUN等[9]提出一种加权过采样的深度自编码器（Weighted Minority Oversampling Deep Auto-encoder，WMODA），用于检测实际旋转机械过程中的故障。KHAN等[10]提出一种代价敏感深度神经网络（Cost-Sensitive Deep Neural Network，CS-DNN），用于自动学习多数和少数类的鲁棒特征表示。

由于类别分布不平衡数据会制约模型分类性能，因此为提升模型的不平衡处理能力，采用组合模型的方式增强算法对少数类别数据的特征提取能力。AdaBoost作为一种高效集成学习方法，是提升分类模型不平衡数据分类能力的重要手段之一[11-12]，通过调整样本权重和弱分类器权值，将弱分类器组集成为一个强分类器。

宽度学习系统（Broad Learning System，BLS）结构简单且分类精度较高[13]。BLS系统模型结构为数据提取稀疏特征后输入随机向量函数链接神经网络（Random Vector Functional Link Neural Network，RVFLNN）的单层可横向扩展网络[14]。BLS模型相比深度网络模型[10]训练时间短、易于训练与再训练[15]。大量实验结果表明，标准的BLS容易受数据集自身分布的影响，改进的BLS模型相继被提出。XU等[16]提出一种用于预测多元时间序列的R-BLS（Recurrent BLS）模型。CHU等[17]采用加权方式提升了BLS模型对有噪声和异常值工业非线性数据的预测能力。BLS-CCA与CNN的级联模型[18]提升了系统对多模态数据的分类能力。徐鹏飞等[19-20]基于加权极限学习机（Weighted Extreme Learning Machine，WELM），提出一种有效的DDbCs-BLS模型处理不平衡数据，该模型的本质是在训练样本上增加一个额外的权重，以得到更好的分类边界线位置，以改善BLS性能。

为进一步提升BLS的不平衡数据处理能力，本文提出一种可实现权重动态更新的集成加权宽度学习系统（Weighted Broad Learning System，WBLS），在KKT条件下，分析比较BLS与WBLS的优化过程，在误差项上添加对角矩阵权重，降低训练误差，提升分类性能。将WBLS集成到AdaBoost模型中，通过基分类器WBLS数据权重的训练实现WBLS权重的动态更新，获得更符合数据分布特征的权重，并将所有基分类器加权集成为一个具备不平衡数据识别能力的新模型AdaBoost-WBLS。

1 宽度学习系统

本节将简要介绍标准BLS结构。与深度学习模型不同，BLS是特征横向排布模型，本质是将数据提取稀疏特征后输入随机向量函数链接神经网络。

当输入数据为$\boldsymbol{X}\in\mathbb{R}^{u\times v}$的矩阵形式时，可表示为$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_u]^{\mathrm{T}}$。BLS通过稀疏特征映射得到映射特征层$\boldsymbol{Z}_m$，可表示如下：

$$\boldsymbol{Z}_m=\varphi(\boldsymbol{X}\boldsymbol{W}_k+\boldsymbol{\beta}_k),m\in(1,N_1),k\in(1,N_2) \tag{1}$$

其中：$\boldsymbol{W}_k$、$\boldsymbol{\beta}_k$是随机生成的权重和偏差；φ是非线性激活函数；N_1是特征层节点数；N_2是特征层数。

映射提取到的特征可作为RVFLNN层的输入，再经特征选择后得到N_3维的增强特征层$\boldsymbol{Z}_{e_l}$，可表示如下：

$$\boldsymbol{Z}_{e_l}=\xi(\boldsymbol{X}\boldsymbol{W}_{e_l}+\boldsymbol{\beta}_{e_l}),l=1,2,\cdots,N_3 \tag{2}$$

映射特征层与增强特征层横向扩展为平层宽度特征$\boldsymbol{A}$，如式（3）所示。通过链接权重$\boldsymbol{W}$分配不同大小的权值进行输出，如式（4）所示。最终模型的目标输出为$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_u]^{\mathrm{T}}$。

$$\boldsymbol{A}=[\boldsymbol{Z}_m,\boldsymbol{Z}_e] \tag{3}$$

$$\boldsymbol{A}\boldsymbol{W}=\boldsymbol{Y} \tag{4}$$

BLS的链接权重$\boldsymbol{W}$是通过岭回归的优化方式快速求得。岭回归是一种快速求解伪逆的方法，本文中其对应的目标函数和计算公式分别如式（5）和式（6）所示：

$$\underset{\boldsymbol{W}}{\arg\min}\left(\|\boldsymbol{A}\boldsymbol{W}-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2\right) \tag{5}$$

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

2 AdaBoost集成的WBLS

2.1 WBLS

在处理实际数据集时，多数据集都存在不同程度的类别不平衡现象。文献[3,14]提供了为浅层网络添加敏感损失权重的方法来处理不平衡数据，以实现类间再平衡。与文献[14]的权重形式不同，权值矩阵可采用对角矩阵形式，将权重添加到数据所对应特征上，采用这种权重形式使模型可以与AdaBoost结合。

式（5）与极限学习机（Extreme Learning Machine，ELM）[14]等单层网络最小化训练误差、最大化类间距离的过程相似。与LS-SVM的优化方式相似，本节基于KKT条件[15]，对BLS与WBLS约束条件下的凸函数进行优化。通过比较推导结果，分析所添加对角权重$\boldsymbol{W}_p$在BLS模型中的作用。

BLS在输入数据$\boldsymbol{X}\in\mathbb{R}^{u\times v}$中提取到的宽度特征表示为$\boldsymbol{A}$，宽度特征对输出的链接权重矩阵表示为$\boldsymbol{W}$。与WELM[10-11]等模型的优化过程类似，BLS的优化过程可表示如下：

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}\left(\|\boldsymbol{AW}-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2\right) \tag{7}$$

$$\text{Minimize}\ \left(\|\boldsymbol{AW}-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2\right) \tag{8}$$

式(8)可简化如下：

$$\text{Minimize}\ \left(\|\boldsymbol{\xi}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2\right)$$

$$\text{Subject to}\ \boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i=\boldsymbol{y}_i^{\mathrm{T}}-\xi_i,i=1,2,\cdots,u \tag{9}$$

其中：$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_u]^{\mathrm{T}}$是模型的目标输出；$\lambda$是模型的正则化项参数，抑制模型的过拟合，也是影响模型性能的重要参数；$\boldsymbol{\xi}=[\xi_1,\xi_2,\cdots,\xi_u]^{\mathrm{T}}$是模型的预测误差。

在KKT条件下，BLS模型的优化过程可表示如下：

$$\boldsymbol{L}_{\mathrm{BLS}}=\frac{1}{2}\lambda\|\boldsymbol{W}\|^2+\frac{1}{2}\sum_{i=1}^{u}\xi_i^2-\sum_{i=1}^{u}\alpha_i[\boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i-\boldsymbol{y}_i+\xi_i] \tag{10}$$

其中：α_i是$\boldsymbol{x}_i$的特征映射对应的Lagrange乘子。

接下来分别求式(10)中$\boldsymbol{W}$、$\boldsymbol{\alpha}$、$\boldsymbol{\xi}$偏导数为0的解。由$\frac{\partial\boldsymbol{L}_{\mathrm{BLS}}}{\partial\boldsymbol{W}}=0$、$\frac{\partial\boldsymbol{L}_{\mathrm{BLS}}}{\partial\boldsymbol{\alpha}}=0$、$\frac{\partial\boldsymbol{L}_{\mathrm{BLS}}}{\partial\boldsymbol{\xi}}=0$可得：

$$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_u]=\lambda^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\alpha}=\lambda^{-1}\sum_{i=1}^{u}\boldsymbol{A}(\boldsymbol{x}_i)^{\mathrm{T}}\alpha_i \tag{11}$$

$$\sum_{i=1}^{u}\boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i-\boldsymbol{y}_i+\sum_{i=1}^{u}\xi_i=0,i=1,2,\cdots,u \tag{12}$$

$$\alpha_i=\sum_{i=1}^{u}\xi_i,i=1,2,\cdots,u \tag{13}$$

WBLS的L2范数凸优化目标可表示如下：

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}\left(\frac{1}{2}\times\boldsymbol{W}_p\times\|\boldsymbol{AW}-\boldsymbol{Y}\|^2+\frac{\lambda}{2}\|\boldsymbol{W}\|^2\right) \tag{14}$$

式(14)可简化如下：

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}\left(\frac{1}{2}\times\boldsymbol{W}_p\times\|\boldsymbol{\xi}\|^2+\frac{\lambda}{2}\|\boldsymbol{W}\|^2\right)$$

$$\text{Subject to}\ \boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i=\boldsymbol{y}_i^{\mathrm{T}}-\xi_i,i=1,2,\cdots,u \tag{15}$$

根据KKT理论，WBLS优化过程可等价表示如下：

$$\boldsymbol{L}_{\mathrm{WBLS}}=\frac{\lambda}{2}\|\boldsymbol{W}\|^2+\frac{\boldsymbol{W}_p}{2}\sum_{i=1}^{u}\xi_i^2-\sum_{i=1}^{u}\alpha_i[\boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i-\boldsymbol{y}_i+\xi_i] \tag{16}$$

分别对式(16)中的$\boldsymbol{W}$、$\boldsymbol{\alpha}$、$\boldsymbol{\xi}$求偏导可得最优解，由$\frac{\partial\boldsymbol{L}_{\mathrm{WBLS}}}{\partial\boldsymbol{W}}=0$、$\frac{\partial\boldsymbol{L}_{\mathrm{WBLS}}}{\partial\boldsymbol{\alpha}}=0$、$\frac{\partial\boldsymbol{L}_{\mathrm{WBLS}}}{\partial\boldsymbol{\xi}}=0$可得：

$$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_u]=\lambda^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\alpha}=\lambda^{-1}\sum_{i=1}^{u}\boldsymbol{A}(\boldsymbol{x}_i)^{\mathrm{T}}\alpha_i \tag{17}$$

$$\sum_{i=1}^{u}\boldsymbol{A}(\boldsymbol{x}_i)\boldsymbol{w}_i-\boldsymbol{y}_i+\sum_{i=1}^{u}\xi_i=0,\ \ i=1,2,\cdots,u \tag{18}$$

$$\alpha_i=\boldsymbol{w}_{p_i}\sum_{i=1}^{u}\xi_i,i=1,2,\cdots,u \tag{19}$$

对比BLS与WBLS在KKT条件下的优化结果的式(13)和式(19)可知，输入数据添加的权重$\boldsymbol{W}_p$是在模型的误差项上，且所加权重$\boldsymbol{w}_{p_i}$与Lagrange乘子α_i成反比。对比式(11)与式(17)可知，在WBLS中α_i又与输入数据所映射的特征层$\boldsymbol{A}$共同决定了链接权重$\boldsymbol{W}$。由此可得，权重$\boldsymbol{W}_p$改变了不同数据特征的比重。

添加的权重有多种形式，文献[5]采用将敏感损失权重添加到所对应的数据层面。本节直接采用对角矩阵权重$\boldsymbol{W}_p=\mathrm{diag}(\boldsymbol{w}_{p_1},\boldsymbol{w}_{p_2},\cdots,\boldsymbol{w}_{p_u})$，$\boldsymbol{W}_p$计算公式如下：

$$\boldsymbol{W}_p=\begin{cases}g/\#(p_i),p_i>\mathrm{AVG}(p_i)\\1/\#(p_i),p_i\leqslant\mathrm{AVG}(p_i)\end{cases} \tag{20}$$

其中，$i=1,2,\cdots,u$；$\#(p_i)$表示第i个数据所属类别的数据量；$\mathrm{AVG}(p_i)$表示平均类别的数据量。

2.2　AdaBoost-WBLS模型

为提升BLS模型对不平衡数据的识别能力，上文从理论上分析了在BLS的误差项上添加权重的作用。为进一步提升模型对于少数类的识别能力，将WBLS集成到AdaBoost.M1框架中，以获得更符合数据分布特征的权重形式。

AdaBoost是一种高效集成学习方法[21]，主要思想是在训练空间上生成一个分布D，初始分配每个训练样本的权值为$1/u$，其中u为训练样本个数。利用迭代训练基分类器，动态更新分类器的权重，并根据多数投票规则将基分类器集成为一个强分类器。本文的基分类器是WBLS，其将T个基分类器迭代训练，从而集成为一个分类能力更强的分类器AdaBoost-WBLS。

在AdaBoost原始框架中，训练样本的分布权值是通过动态迭代实现对基分类器的权重更新。在WBLS处理不平衡数据时，添加权重$\boldsymbol{W}_p$可抑制少数类样本的误差，提升分类器对少数类的识别能力。本文将WBLS集成到AdaBoost，实现了对权重$\boldsymbol{W}_p$的动态更新，可获得更合理的权重形式。与文献[5]的加权方式不同，本文权重采用对角矩阵形式$\boldsymbol{W}_p=\mathrm{diag}(w_{p_1},w_{p_2},\cdots,w_{p_u})$，仅在不同数据对应的特征上添加一维常数的权重。

与传统Boosting类模型集成过程不同，当模型输出数据的类别数为j时，本文对AdaBoost-WBLS的集成过程进行如下改进：

1)在传统的AdaBoost模型中，第1个基分类器的起始数据的权重通常选用$1/u$，而本文采用特殊起始权重$1/(j\times\boldsymbol{W}_p)$。这种将类别数据引入模型初始化过程的方式，可增加模型的类别信息，提升AdaBoost-WBLS对少数类样本的识别效率与识别能力。权重初始化公式如下：

$$\boldsymbol{D}_1=\begin{cases}g/(j\times\boldsymbol{W}_p),p_i>\mathrm{AVG}(p_i)\\1/(1\times\boldsymbol{W}_p),p_i\leqslant\mathrm{AVG}(p_i)\end{cases} \tag{21}$$

其中：$i=1,2,\cdots,u$。

2)在传统Boosting类模型中，集成学习过程是对所有训练样本之间进行归一化迭代处理，而本文模型采用在类别内部归一化的方法，以达到提升类

间平衡度的目的，即分布权值 $\boldsymbol{D}_t(\boldsymbol{x}_i), i=1,2,\cdots,u$ 对不同类别分别累加，依次更新。更新公式如下：

$$\boldsymbol{D}_{t+1}=\frac{\boldsymbol{D}_t(\boldsymbol{x}_i)\exp(-\alpha_t H_t(\Omega_t(\boldsymbol{x}_i),\boldsymbol{y}_i))}{Z_t} \tag{22}$$

其中：$\Omega_t(\boldsymbol{x}_i)$ 是第 t 个基础分类器对数据 $\boldsymbol{x}_i$ 的预测结果；$H_t(\cdot)$ 采用满足最优错误率的激活形式；Z_t 是正则化参数，满足 $\sum_{\boldsymbol{x}_i\in \text{class}\,j}\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)=1/j$ 且 $\sum_{i=1}^{v}\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)=1$。$\alpha_t$ 和 $H_t(\Omega_t(\boldsymbol{x}_i,y_i))$ 的计算公式如下：

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)=\frac{1}{2}\ln\left(\frac{\sum_{i:\,\Omega_t(\boldsymbol{x}_i)=\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}{\sum_{i:\,\Omega_t(\boldsymbol{x}_i)\neq\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}\right) \tag{23}$$

$$H_t(\Omega_t(\boldsymbol{x}_i,\boldsymbol{y}_i))=\begin{cases}1,\Omega_t(\boldsymbol{x}_t)=\boldsymbol{y}_i\\-1,\Omega_t(\boldsymbol{x}_t)\neq\boldsymbol{y}_i\end{cases} \tag{24}$$

算法1 AdaBoost-WBLS算法

输入 训练集 $P=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_u,\boldsymbol{y}_u)\}$，迭代次数（即BLS基分类器个数）$T$

输出 对于测试数据 $\boldsymbol{x}$，$\boldsymbol{\Theta}=\underset{k}{\text{argmax}}\sum_{t=1}^{T}\alpha_t[\Omega_t(\boldsymbol{x})=k]$

步骤1 初始化权重 $\boldsymbol{D}_1(\boldsymbol{x}_i)=\frac{1}{j\times\boldsymbol{W}_p}$。

步骤2 循环迭代更新 $\boldsymbol{W}_p$：

1）$t=1,2,\cdots,T$。

2）$\boldsymbol{W}_p=\text{diag}(\boldsymbol{D}_t(\boldsymbol{x}_i)), i=1,2,\cdots,u$。

3）训练BLS分类器，输出结果 Ω_t。

4）按照类别分别更新权重，对于第 j 类：

$$\alpha_t^j=\frac{1}{2}\ln\left(\frac{\sum_{\boldsymbol{x}_i\in\text{class}\,j:\,\Omega_t(\boldsymbol{x}_i)=\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}{\sum_{\boldsymbol{x}_i\in\text{class}\,j:\,\Omega_t(\boldsymbol{x}_i)\neq\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}\right)$$

$$\forall\boldsymbol{x}_i\in\text{class}\,j,\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)=\frac{\boldsymbol{D}_t(\boldsymbol{x}_i)\exp(-\alpha_t^j H_t(\Omega_t(\boldsymbol{x}_i),j))}{Z_t^j}$$

其中：Z_t^j 是AdaBoost中的正则化参数，值满足 $\sum_{\boldsymbol{x}_i\in\text{class}\,j}\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)=1/j$。

5）令 $T=t-1$，$\sum_{i:\,\Omega_t(\boldsymbol{x}_i)=\boldsymbol{y}_i}\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)\leqslant\sum_{i:\,\Omega_t(\boldsymbol{x}_i)\neq\boldsymbol{y}_i}\boldsymbol{D}_{t+1}(\boldsymbol{x}_i)$。

步骤3 计算第 t 个BLS基分类器的投票权重 $\alpha_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)=\frac{1}{2}\ln\left(\frac{\sum_{i:\,\Omega_t(\boldsymbol{x}_i)\neq\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}{\sum_{i:\,\Omega_t(\boldsymbol{x}_i)\neq\boldsymbol{y}_i}\boldsymbol{D}_t(\boldsymbol{x}_i)}\right)$。

3 实验验证

为验证AdaBoost-WBLS性能，将其分别与Boosting类、BLS类模型进行消融实验研究，之后与KMODM[7]、CWsRF[8]、WMODA[9]、CS-DNN[10]这4种不平衡分类模型进行对比研究。实验环境为Windows 10系统，8 GB内存，Intel Core i7 6500 CPU，编程环境为Matlab 2016b。采用 $\underbrace{\{-1,-1,\cdots,1,\cdots,-1\}}_{j}$ 输出形式，共输出 j 个类别，在输出类别的位置上设置为1，其余位置均设置为−1。

映射特征层节点数、特征层数、增强节点层数、正则化参数取值范围分别为 $N_1=10$、$N_2\in\{1,3,\cdots,21\}$、$N_3\in\{1,10,20,\cdots,500\}$、$\lambda\in\{2^{-40},2^{-39},\cdots,2^0,\cdots,2^{20}\}$。

引入不平衡率（Imbalance Ratio，IR），评价不同的不平衡数据集中数据的分布形式。在二分类中IR的计算公式如下：

$$I_{\text{IR}}=\frac{\#(\text{minority})}{\#(\text{majority})} \tag{25}$$

其中：#(minority)、#(majority)分别表示数据集中多数类与少数类的样本数。

在多分类中IR的计算公式如下：

$$I_{\text{IR}}=\frac{\text{Min}\#(p_i)}{\text{Max}\#(p_i)} \tag{26}$$

3.1 评价指标选取

在对数据进行分类时，准确率是分类任务常用的评价指标，但是在不平衡分类任务中，使用准确率作为评价模型性能的唯一指标，不能准确表征模型对少数类的分类能力。以二分类为例，在一些极端的分布中，少数类与多数类的比例可能达到99∶1，模型即使不具备对少数样本的分类能力，依然可以得到较高的准确率，但此时的全局准确率不能用于评价其对于少数类的识别能力。因此，本文还选用G-mean评价指标来评价不平衡数据的分类结果。

在二分类中，将少数类作为正样本（+1），多数类作为负样本（−1），则二分类混淆矩阵如表1所示。

表1 二分类混淆矩阵

Table 1 Binary confusion matrix

真实值	预测值为+1	预测值为−1
+1	T_{TP}	F_{FN}
−1	F_{FP}	T_{TN}

在表1中，T_{TP} 为正样本被分类为正确类的统计量，F_{FP} 为负样本被分类为正样本的统计量，F_{FN} 为正样本被分类为负样本的统计量，T_{TN} 为负样本被分类为正确类的统计量。

准确率（Accuracy）表示所有样本的准确识别率，计算公式如下：

$$A_{\text{Accuracy}}=\frac{T_{\text{TP}}+T_{\text{TN}}}{T_{\text{TP}}+T_{\text{TN}}+F_{\text{FP}}+F_{\text{FN}}} \tag{27}$$

召回率（Recall）表示正样本（少数类）的识别率，计算公式如下：

$$R_{\text{Recall}}=\frac{T_{\text{TP}}}{T_{\text{TP}}+F_{\text{FN}}} \tag{28}$$

特异率（Specificity）表示负样本（多数类）的识别率，计算公式如下：

$$S_{\text{Specificity}}=\frac{T_{\text{TN}}}{F_{\text{FP}}+T_{\text{TN}}} \tag{29}$$

G-mean值表示各类别识别率的几何平均值。在二分类任务中，G-mean是召回率与特异率的几何平均值，计算公式如下：

$$G_{\text{G-mean}}=\sqrt{R_{\text{Recall}}\times S_{\text{Specificity}}} \tag{30}$$

在多分类任务中，分类目标数大于2。此时，G-mean采用一对多（One-Against-All，OAA）的统计方式，分别计算各类别的识别准确率，再求整体G-mean。当有j个类别时，G-mean计算公式如下：

$$G_{\text{G-mean}}=\sqrt{R_{\text{Recall}_1}\times R_{\text{Recall}_2}\times\cdots\times R_{\text{Recall}_j}} \tag{31}$$

3.2　数据集选取

选取UCI数据库中15个不平衡数据集作为消融实验与对比实验对象。数据集具体情况如表2所示，其中，12个数据集是二分类数据集，3个数据集是多分类数据集，不平衡率分布范围为0.007 6~0.912 8。Abalone数据集与Yeast数据集为生物数据集，前者通过物理测量预测鲍鱼的年龄，后者可对核蛋白和非核蛋白的核定位信号进行判别分析。New-thyroid为甲状腺疾病数据集，Glass、Vehicle与Satimage数据集为普通分类数据集。

表2　实验数据集设置

Table 2　Setting of experimental dataset

数据集	特征数	类别数	训练集样本数	测试集样本数	IR
Yeast1	8	2	1 187	297	0.406 6
Yeast3	8	2	1 187	297	0.123 4
Yeast6	8	2	1 187	297	0.024 2
Yeast1vs7	8	2	367	92	0.069 9
Glass1	9	2	171	43	0.554 5
Glass4	9	2	171	43	0.062 1
Abalone19	8	2	3 338	836	0.007 6
Abalone9vs18	8	2	584	147	0.060 0
Vehicle0	18	2	677	169	0.307 6
Vehicle1	18	2	677	169	0.345 0
Vehicle2	18	2	677	169	0.347 1
Vehicle3	18	2	677	169	0.334 5
New-thyroid	6	3	172	43	0.200 0
Vehicle	5	4	647	169	0.912 8
Satimage	8	7	4 435	2 000	0.387 1

3.3　λ参数的作用

本文对宽度学习模型中的正则化参数λ进行实验讨论。在相关研究中，参数λ通常采用固定值$\lambda=2^{-30}$。因此，通过实验分析不平衡数据处理时，参数λ变化对实验结果的影响。

实验对象为不平衡数据集Glass4，N_1、N_2和N_3分别选取10、20和500，使参数λ成为唯一变量。实验参考了大量研究对λ的取值方式，选取取值范围为$\lambda\in\{2^{-40},2^{-39},\cdots,2^{0},\cdots,2^{20}\}$。通过实验对比了BLS、$g=1$时W1-BLS和$g=0.618$时W2-BLS的G-mean结果，如图1所示。

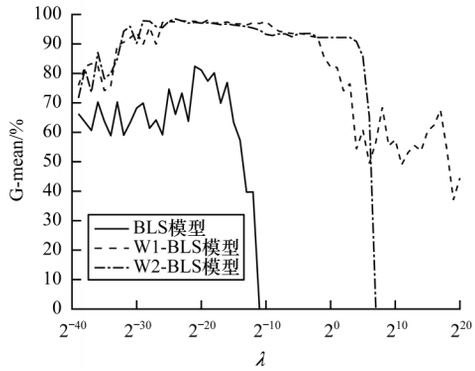

图1　Glass4数据集上随λ变化的G-mean

Fig.1　G-mean when λ varies on the Glass4 dataset

根据实验结果可知，在λ从2^{-40}变化到2^{20}的过程中，G-mean值基本呈现先上升后下降的趋势。当λ逐渐增大时，会达到最优的G-mean。当继续增大时，模型将会出现过拟合现象，导致G-mean值迅速降低。根据对比可知，在BLS内添加形如$\boldsymbol{W}_p=\text{diag}(w_{p_1},w_{p_2},\cdots,w_{p_u})$的权重，不仅可以提升模型的G-mean峰值，而且相对提高了模型的稳定性。

3.4　消融实验

3.4.1　Boosting类模型实验验证

本文设计一种将WBLS作为基分类器并在AdaBoost框架中嵌入WBLS以提升不平衡数据分类性能的优化方法。设置N_1、N_2、N_3、λ分别为10、20、500、2^{20}。AdaBoost-WBLS与DDbCs-BLS等加权宽度学习模型的最大不同点在于：基于AdaBoost模型可以实现自动更新训练样本所添加的权值。在AdaBoost中，分布权重是训练样本的重要性表征。在训练过程中，被错误分类的样本通过获得相比较被正确分类样本更大的分布权重以提升其重要性。因此，本文采用训练样本所添加的分布权值$\boldsymbol{W}_p$作为AdaBoost-WBLS中的训练样本对应的权值。

在Yeast1vs7数据集上，对AdaBoost-WBLS与传统Boosting框架的BLS迭代过程中G-mean的变化情况进行比较，结果如图2所示。由图2可知，AdaBoost-WBLS模型的G-mean曲线上升更快，获取最优基分类的迭代次数更少，稳定性更强，并且峰值更高，表明了学习到的特征更丰富。

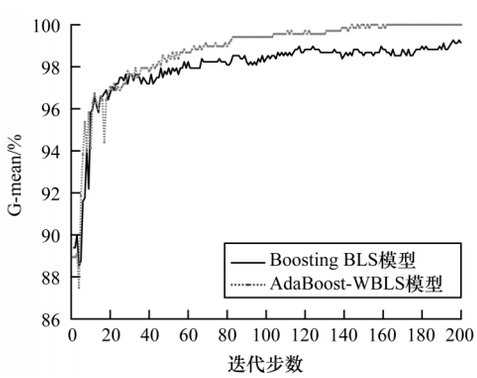

图2　Yeast1vs7数据集上AdaBoost-WBLS与Boosting-BLS模型的G-mean

Fig.2　G-mean of AdaBoost-WBLS and Boosting-BLS model on Yeast1vs7 dataset

在5个数据集上对Boosting-WELM、AdaBoost-WELM、Boosting-WBLS、AdaBoost-WBLS这4种Boosting模型进行性能对比，G-mean结果如表3所示，Accuracy结果如表4所示，其中Boosting-WELM和AdaBoost-WELM模型的结果引自文献[3]。BLS参数通过网格搜索设置为最优参数，其中，λ为正则化参数，L为网络节点数。

比较表3、表4中AdaBoost-WBLS与Boosting-WBLS模型结果可以看出：前者在多数数据集上的G-mean都相对更高，且具有相对较高的Accuracy；在Yeast3数据集上G-mean高0.90个百分点，Accuracy基本相等；在Yeast6数据集上G-mean高5.17个百分点，Accuracy下降了0.98个百分点；在Abalone19数据集上G-mean高1.75个百分点，Accuracy却下降了3.25个百分点，这说明AdaBoost-WBLS模型更关注少数类，而Boosting-WBLS模型更关注多数类的总体准确率。比较表3、表4中AdaBoost-WBLS、Boosting-WELM与AdaBoost-WELM模型结果可以得出，在经过网格搜索得出最佳参数后，BLS模型具有更高的G-mean与Accuracy。

表3 Boosting类相关模型消融实验的G-mean

Table 3 G-mean of Boosting-related model ablation experiments

数据集	Boosting-WELM			AdaBoost-WELM			Boosting-WBLS				AdaBoost-WBLS			
	λ	L	G-mean/%	λ	L	G-mean/%	λ	N_2	N_3	G-mean/%	λ	N_2	N_3	G-mean/%
Yeast3	2^{6}	160	92.92	2^{8}	400	84.47	2^{-21}	21	60	94.37	2^{-38}	21	490	95.27
Yeast6	2^{-2}	430	89.29	2^{14}	260	75.50	2^{-24}	19	140	85.99	2^{-40}	9	450	90.16
Glass1	2^{20}	250	76.81	2^{20}	510	75.72	2^{-18}	17	10	76.85	2^{-22}	7	30	78.08
Abalone19	2^{-10}	870	77.06	2^{20}	890	50.11	2^{-25}	5	140	81.21	2^{-30}	11	460	82.86
Abalone9vs18	2^{6}	230	90.64	2^{10}	420	75.22	2^{-34}	21	90	90.87	2^{-7}	7	70	92.18

表4 Boosting类相关模型消融实验的Accuracy

Table 4 Accuracy of Boosting-related model ablation experiments

数据集	Boosting-WELM			AdaBoost-WELM			Boosting-WBLS				AdaBoost-WBLS			
	λ	L	Accuracy/%	λ	L	Accuracy/%	λ	N_2	N_3	Accuracy/%	λ	N_2	N_3	Accuracy/%
Yeast3	2^{14}	230	92.99	2^{10}	350	95.22	2^{-28}	15	70	95.94	2^{-28}	21	490	95.85
Yeast6	2^{-2}	430	89.29	2^{14}	260	75.50	2^{-37}	21	490	95.08	2^{-35}	9	450	93.10
Glass1	2^{20}	250	76.81	2^{20}	510	75.72	2^{-14}	9	20	80.47	2^{-24}	7	30	79.74
Abalone19	2^{-10}	870	77.06	2^{20}	890	50.11	2^{-40}	20	490	85.11	2^{-40}	10	460	81.86
Abalone9vs18	2^{6}	230	90.64	2^{10}	420	75.22	2^{-33}	21	90	96.03	2^{-28}	7	320	94.99

3.4.2 BLS类模型实验验证

在6个二分类数据集上比较了AdaBoost-WBLS、BLS、DDbCs-BLS模型的G-mean与Accuracy，结果如表5、表6所示。由表5、表6可以看出：与BLS模型相比，AdaBoost-WBLS模型的G-mean结果均得到了改善，在Yeast3数据集上AdaBoost-WBLS模型提升了9.31个百分点，在Vehicle1数据集上提升了2.38个百分点；与DDbCs-BLS[19]模型相比，AdaBoost-WBLS模型的G-mean在Yeast1数据集上高出3.67个百分点，在Vehicle2数据集上提高了0.8个百分点。由此可见，本文提出的不平衡数据分类方法在结合Boosting模型后，提升了集成模型的局部泛化能力。

表5 BLS类相关模型消融实验的G-mean

Table 5 G-mean of Boosting-related model ablation experiments %

数据集	BLS	DDbCs-BLS	AdaBoost-WBLS
Vehicle0	95.57	97.62	99.21
Vehicle1	84.03	85.31	86.41
Vehicle2	96.43	98.65	99.45
Vehicle3	82.61	83.52	86.97
Yeast1	75.75	77.00	80.67
Yeast3	85.96	93.03	95.27

表6 BLS类相关模型消融实验的Accuracy

Table 6 Accuracy of Boosting-related model ablation experiments %

数据集	BLS	DDbCs-BLS	AdaBoost-WBLS
Vehicle0	97.90	98.82	98.17
Vehicle1	86.39	87.35	87.17
Vehicle2	98.29	98.65	97.42
Vehicle3	82.96	83.67	85.01
Yeast1	75.75	77.00	80.75
Yeast3	88.96	87.03	88.27

3.5 对比实验

在Vehicle0、Vehicle3、Yeast3等3个二分类与New-thyriod、Vehicle、Satimage等3个多分类数据集上对比AdaBoost-WBLS与KMODM[7]、CWsRF[8]、WMODA[9]、CS-DNN[10]模型的不平衡数据分类性能。G-mean结果如表7所示。由表7可以看出，AdaBoost-WBLS的G-mean明显高于其他4种模型，在Vehicle0数据集上比KMODM模型高出3.74个百分点，在New-thyriod数据集上比CWsRF模型高出3.09个百分点，在Satimage数据集上比WMODA模型高出4.36个百分点，在Vehicle数据集上比CS-DNN模型

高出1.15个百分点。实验结果验证了AdaBoost-WBLS模型通过多个加权BLS组合成的新分类器可有效处理不平衡数据。

表7　对比实验的G-mean

Table 7　G-mean of contrast experiments　%

数据集	KMODM	CWsRF	WMODA	CS-DNN	AdaBoost-WBLS
Vehicle0	95.47	96.76	95.89	97.86	99.21
Vehicle3	84.19	86.56	82.16	85.95	87.97
Yeast3	92.10	93.21	89.08	95.96	95.27
New-thyriod	98.07	96.63	94.18	98.35	100.00
Vehicle	84.61	86.15	82.54	85.73	86.97
Satimage	87.86	89.54	87.15	89.89	90.58

Accuracy结果如表8所示，可以看出相比其他4种模型，AdaBoost-WBLS模型的Accuracy相对较高。在New-thyriod数据集上比WMODA模型高出4.65个百分点，达到100%。可见，AdaBoost-WBLS模型在提升对少数类识别能力的同时，具有较高的识别精度。

表8　对比实验的Accuracy

Table 8　Accuracy of contrast experiments　%

数据集	KMODM	CWsRF	WMODA	CS-DNN	AdaBoost-WBLS
Vehicle0	95.47	97.38	94.96	99.05	99.17
Vehicle3	84.19	85.27	82.16	86.35	84.07
Yeast3	92.10	82.81	88.95	86.22	85.01
New-thyriod	98.07	96.91	95.35	97.45	100.00
Vehicle	84.61	87.13	83.22	87.21	89.86
Satimage	87.86	89.91	86.00	87.62	88.17

4　结束语

本文研究旨在通过集成AdaBoost与WBLS提升BLS的不平衡数据集处理能力。基于KKT条件推导验证了WBLS的有效性。将加权宽度学习的数据特征与AdaBoost中分类器的权重结合，在算法层面进行AdaBoost与BLS的融合。在AdaBoost-WBLS集成初始化时，WBLS采用基于类别信息的权重，使基分类器具有先验类别信息并且模型更快收敛。在迭代训练过程中，对WBLS基分类器数据权重的更新方式进行调整。对不同类别数据对应的权重采用不同的正则化准则，使权值具有更高的类间区分度，同时显著提升模型的训练效率。实验结果表明，AdaBoost-WBLS模型相比同类模型在二分类与多分类数据集上G-mean均有显著提升，准确率较高，且具有较好的不平衡数据的处理能力。下一步将使用集成BLS的AdaBoost模型，解决多模态数据分类等问题。

参考文献

[1] 韩涛，兰雨晴，肖利民，等. 一种增量并行式动态图异常检测算法[J]. 北京航空航天大学学报，2018，44(1)：117-124.
HAN T，LAN Y Q，XIAO L M，et al. Incremental and parallel algorithm for anomaly detection in dynamic graphs[J]. Journal of Beijing University of Aeronautics and Astronautics，2018，44(1)：117-124. (in Chinese)

[2] 陈龙，韩中洋，赵珺，等. 数据驱动的综合能源系统运行优化方法研究综述[J]. 控制与决策，2021，36(2)：283-294.
CHEN L，HAN Z Y，ZHAO J，et al. Review of research of data-driven methods on operational optimization of integrated energy systems[J]. Control and Decision，2021，36(2)：283-294. (in Chinese)

[3] LI K，KONG X F，LU Z，et al. Boosting weighted ELM for imbalanced learning[J]. Neurocomputing，2014，128：15-21.

[4] GUO H X，LI Y J，LI Y N，et al. BPSO-AdaBoost-KNN ensemble learning algorithm for multi-class imbalanced data classification[J]. Engineering Applications of Artificial Intelligence，2016，49：176-193.

[5] YEN S J，LEE Y S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications，2009，36(3)：5718-5727.

[6] 古平，杨扬. 面向不均衡数据集中少数类细分的过采样算法[J]. 计算机工程，2017，43(2)：241-247.
GU P，YANG Y. Oversampling algorithm oriented to subdivision of minority class in imbalanced data set[J]. Computer Engineering，2017，43(2)：241-247. (in Chinese)

[7] ZHANG X G，WANG D X，ZHOU Y C，et al. Kernel modified optimal margin distribution machine for imbalanced data classification[J]. Pattern Recognition Letters，2019，125：325-332.

[8] ZHU M，XIA J，JIN X Q，et al. Class weights random forest algorithm for processing class imbalanced medical data[J]. IEEE Access，2018，6：4641-4652.

[9] SUN J，LI H，FUJITA H，et al. Class-imbalanced dynamic financial distress prediction based on AdaBoost-SVM ensemble combined with SMOTE and time weighting[J]. Information Fusion，2020，54：128-144.

[10] KHAN S H，HAYAT M，BENNAMOUN M，et al. Cost-sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems，2015，29(8)：3573-3587.

[11] XING H J，LIU W T. Robust AdaBoost based ensemble of one-class support vector machines[J]. Information Fusion，2020，55：45-58.

[12] 张旭，周新志，赵成萍，等. 基于犹豫模糊决策树的非均衡数据分类[J]. 计算机工程，2019，45(8)：75-79，91.
ZHANG X，ZHOU X Z，ZHAO C P，et al. Unbalanced data classification based on hesitant fuzzy decision tree[J]. Computer Engineering，2019，45(8)：75-79，91. (in Chinese)

[13] CHEN C L P，LIU Z L. 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.

[14] ZHANG L，SUGANTHAN P N. A comprehensive evaluation of random vector functional link networks[J]. Information Sciences，2016，367/368：1094-1105.

[15] JIN J W，CHEN C L. Regularized robust broad learning system for uncertain data modeling[J]. Neurocomputing，2018，322：58-69.

[16] XU M L，HAN M，CHEN C L P，et al. Recurrent broad learning systems for time series prediction[J]. IEEE Transactions on Cybernetics，2020，50(4)：1405-1417.

[17] CHU F，LIANG T，CHEN C L P，et al. Weighted broad learning system and its application in nonlinear industrial process modeling[J]. IEEE Transactions on Neural Networks and Learning Systems，2020，31(8)：3017-3031.

（下转第112页）

（上接第105页）

[18] 王召新，续欣莹，刘华平，等. 基于级联宽度学习的多模态材质识别[J]. 智能系统学报，2020，15(4)：787-794.
WANG Z X，XU X Y，LIU H P，et al. Cascade broad learning for multi-modal material recognition[J]. CAAI Transactions on Intelligent Systems，2020，15(4)：787-794. (in Chinese)

[19] 徐鹏飞，王敏，刘金平，等. 基于数据分布特性的代价敏感宽度学习系统[J]. 控制与决策，2021，36(7)：1686-1692.
XU P F，WANG M，LIU J P，et al. Data distribution-based cost-sensitive broad learning system[J]. Control and Decision，2021，36(7)：1686-1692. (in Chinese)

[20] ZONG W W，HUANG G B，CHEN Y Q. Weighted extreme learning machine for imbalance learning[J]. Neurocomputing，2013，101：229-242.

[21] TOH K A. Deterministic neural classification[J]. Neural Computation，2008，20(6)：1565-1595.

编辑　陆燕菲