文章编号：2095-2163(2022)01-0058-07 中图分类号：TP391 文献标志码：A

基于边界增强和去噪的自适应双权重过采样方法研究

高子寒，宋 燕
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：类别不平衡是现实世界普遍存在的问题，其严重影响着各种预测模型的预测效果，使得这些模型仅能准确识别出多类样本，却不易识别出少类样本。本文提出一种基于边界增强和去噪的自适应双权重过采样（Adaptive Double-Weight Enhanced Boundary and Denoising Oversampling，ADWEBDO）方法，以处理不平衡问题。ADWEBDO 的主要思想是：引入 K 近邻（K Nearest Neighbor，KNN）去噪技术，降低噪声样本合成的可能性；提出一种基于类间距离和少类簇大小的双重权重样本分配方法，有效避免了类重叠现象的产生；采用模糊 C 均值（Fuzzy C-Means，FCM）聚类算法，对样本进行聚类分析，提高了合成少类样本的可靠性；提出一种基于特征空间的合成样本策略，增加了合成少类样本的多样性和合理性。最终，本文提出的方法在 7 个 UCI 数据集上进行实验，并取得了令人满意的结果。

关键词：类别不平衡；边界增强；去噪；双权重；过采样

An adaptive double-weight enhanced boundary and denoising oversampling approach

GAO Zihan，SONG Yan
（School of Optical-Electrical and Computer Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China）

【Abstract】Class imbalance is a common problem in the real world，which seriously affects the prediction effect of various prediction models. These models can only identify majority class samples accurately，but it is difficult to identify minority class samples. In this paper，an Adaptive double-weight Enhanced Boundary and Denoising Oversampling（ADWEBDO）method is proposed to deal with imbalance problems. The main idea of ADWEBDO is to introduce K Nearest Neighbor（KNN）denoising technology to reduce the possibility of noise sample synthesis；a dual-weight sample allocation method based on inter-class distance and cluster capacity of minority class is proposed to avoid the overlapping of classes effectively；Fuzzy C-means（FCM）clustering algorithm is used to analyze the sample clustering，which improves the reliability of the sample synthesis；a feature space-based strategy for synthesizing samples is proposed，which increases the diversity and rationality of synthesized minority class samples. Finally，the proposed method is tested on seven UCI datasets with satisfactory results.

【Key words】class imbalance；boundary enhancement；denoising；double-weight；oversampling

0 引 言

在机器学习领域，对于不平衡数据的学习是一项极具挑战性的任务[1-2]。不同类别、不同数量的数据构成的数据集，称为不平衡数据集[3]。如果上述数据集仅包含两个类别，则具有样本数量多的类别称为多数类，其余样本所在类别称为少数类。与此同时，不平衡数据的学习对于研究界也是至关重要的，因为其普遍存在于各种常见的分类任务中。例如，电信欺诈检测[4]、癌症基因检测[5]和推荐系统[6]等等。

传统的机器学习分类器在处理不平衡数据时得到的分类精度通常是不理想的，特别是对少数类的分类效果不佳。这是由于少数类样本的数量太少[7]、所含信息不足，导致分类结果趋向于多数类。但少数类样本中含有的特征往往更重要，因此提高对少数类样本的分类精度是处理不平衡问题的关键所在。通常情况下，不平衡指的是类间不平衡，即两类样本数量的差异程度，而其只是影响不平衡学习的因素之一。影响不平衡的其它因素还包括：类内不平衡、重叠区域的大小和离群点等[8-10]。

为了更好地解决类别不平衡问题，研究人员提

基金项目：上海市自然科学基金（18ZR1427100）。

作者简介：高子寒（1997-），男，硕士研究生，主要研究方向：大数据模型应用；宋 燕（1979-），女，博士，教授，博士生导师，主要研究方向：大数据算法、图像处理、预测控制。
通讯作者：宋 燕 Email：sonya@usst.edu.cn

收稿日期：2021-10-01

出了两大解决方案，即基于算法和基于数据的解决方案。基于算法的解决方案是通过改变算法的学习方式，增强模型对少数类样本的识别能力，最终降低数据不平衡对分类器带来的消极影响，主要分为代价敏感学习、单类学习和集成学习[11-13]。而基于数据的解决方案，是通过采样技术改变不同类别的样本数，从而达到多类和少类的相对平衡，其主要包括欠采样和过采样两种方法。与欠采样相比较而言，过采样则充分保留了数据样本所含的重要信息。因此，本文着重于过采样方法的研究。

过采样通常是经过采样[14]或生成合成数据样本[15-18]来实现的。为了保证合成样本的质量，合成样本应尽可能满足以下要求：不含有干扰信息的噪声且包含有用信息的样本；合成的新样本不能落在多数类区域，以避免类重叠现象的产生。

基于以上基础，本文提出一种新颖的基于边界增强和去噪的自适应双权重过采样方法(ADWEBDO)。该方法通过充分考虑多数类与少数类间的数据分布信息，以及少数类内部的数据分布信息，增加了对边界少数类样本的采样权重，一定程度上避免了类重叠现象的产生；由于利用去噪技术对原始数据进行了去冗余处理，降低了合成噪声样本的可能性；同时基于不同少数类簇的样本特征空间，提出一种基于特征边界组合新样本的策略，不仅保证了合成样本与原样本之间的相似性，而且还增加了合成样本的多样性。

1　相关工作

1.1　模糊 C 均值聚类算法

模糊 C 均值(FCM)聚类算法是一种基于划分的聚类算法，其基本思想是利用模糊聚类分析方法，将所有对象划分到 C 个簇中，使得划分到同一个簇的对象之间的相似度最大，划分到不同簇的对象之间的相似度最小，以达到聚类的目的。FCM 的聚类模型如公式(1)所示：

$$\begin{cases} J(U,\bar{c}_1,\bar{c}_2,\bar{c}_C)=\sum_{i=1}^{C}\sum_{j=1}^{n}u_{ij}^{m}\parallel x_j-\bar{c}_i\parallel^2 \\ \text{s.t}\sum_{j=1}^{n}u_{ij}=1 \end{cases} \tag{1}$$

其中，$\boldsymbol{x}_j(j=1,2,\cdots,n)$ 是数据集 $\{x_1,x_2,\cdots,x_n\}$ b 维空间的样本向量；$\boldsymbol{U}$ 是包含集合 $\{u_{ij}\}(i=1,2,\cdots,C;\ j=1,2,\cdots,n)$ 的划分矩阵；u_{ij} 是第 i 个簇的第 j 个样本；$\bar{c}_i$ 是数据集的第 i 个簇的簇中心；m 是模糊因子($m\geqslant 1$)。为了最小化式(1)中的目标函数 $\boldsymbol{J}$，$\bar{c}_i$ 和 u_{ij} 的更新规则分别如式(2)、(3)所示：

$$\bar{c}_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{2}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{\parallel x_j-\bar{c}_i\parallel}{\parallel x_j-\bar{c}_k\parallel}\right)^{\frac{2}{m-1}}} \tag{3}$$

其中，式(2)、(3)中的变量和参数与式(1)中定义的变量和参数相同。

1.2　多层感知机分类算法

多层感知机(MLP)分类算法是一种非线性分类算法，是神经网络的一种。多层感知机的基本结构包括输入层、隐藏层和输出层。不同层之间是全连接的，即上一层的任何一个神经元与下一层的所有神经元都有连接。多层感知机中最基本的单元叫做神经元，图 1 表示的是著名的“M-P 神经元模型”。

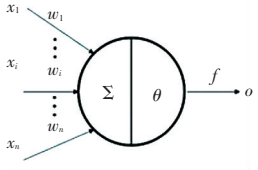

图 1　神经元结构

Fig. 1　Neuron structure

神经元接收来自不同 n 个神经元的输入信号 $(x_1,x_2,\cdots,x_i,\cdots,x_n)$，这些信号通过 n 个权重 $(w_1,w_2,\cdots,w_i,\cdots,w_n)$ 加权，传递并将总输入值与阈值 θ 比较，最后经过激活函数 f 得到输出 o。神经元的输入和输出关系如公式(4)所示：

$$o=f\left(\sum_{i=1}^{n}w_ix_i-\theta\right) \tag{4}$$

其中，x_i 表示输入的第 i 个神经元；w_i 表示第 i 个神经元的对应的连接权重；θ 表示阈值；$f(\cdot)$ 表示激活函数；o 表示神经元的最终输出。

在多层感知机分类算法中，数据首先经过输入层，接着在隐藏层中进行转换，最后在输出层中作出预测。除输入输出层外，中间可以有多个隐层。最简单的多层感知机模型只含一个隐藏层，即 3 层结构，如图 2 所示。本文将利用 MLP 分类器作为后续过采样算法的验证模型。

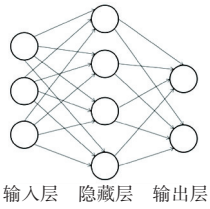

图 2 含一层隐藏层的多层感知机模型

Fig. 2 Multi-layer perceptron model with one hidden layer

2 基于边界增强和去噪的过采样方法

在本文中提出的 ADWEBDO 方法,对少数类样本簇使用基于类间距离和簇规模的双指标,自适应为其分配样本合成数,从而增加了对边界少数类样本的采样权重,避免了类重叠现象的产生;同时经过去噪处理,降低了合成噪声样本的可能性;最后基于本文提出的基于特征空间合成样本策略进行过采样,增加了合成样本的多样性。此方法主要包含:去除噪声样本、聚类分析、自适应分配合成样本数、基于特征空间的过采样 4 部分。

2.1 去除噪声样本

首先,根据公式(5)计算不同数据集需要新合成的少数类样本的数目 G。

$$G = n_{maj} - n_{\min} \tag{5}$$

其中, n_{maj} 是原始数据集中多数类样本数,$n_{\min}$ 是少数类样本数。

如果选择的目标样本是噪声样本,则新合成的样本也极有可能是噪声样本,最终导致模型性能下降。因此通过去噪可以降低噪声样本生成的可能性。对于原始数据集 I 中的每个少数类样本,通过欧式距离计算其 K 近邻。如果此目标样本的所有 K 近邻都是多数类样本,则将其视为噪声样本并直接将其剔除。同样,对于每个多数类样本执行相同的操作。最后,将剩余的样本添加到集合 X 中,记作过滤后的数据集 X。

2.2 聚类分析

聚类分析步骤需要为少数类中的不同簇自适应地分配合成样本数。首先,使用 FCM 聚类算法,对数据集 X 中的少数类和多数类进行聚类分析,同时获得对应的簇划分结果。少数类划分为 $|A|$ 个簇,每个簇的聚类中心用 a_i 表示,每个簇中包含的少数类样本的数量用 n_i 表示;多数类划分为 $|O|$ 个簇,每个簇的聚类中心用 o_q 表示,其中 i 从 1 到 $|A|$,q 从 1 到 $|O|$。

2.3 自适应分配合成样本数

为了更好地确定各个少数类簇的合成样本数,充分考虑了簇内数据分布和类间数据分布,提出一种自适应双指标样本分配策略。

第一个指标称为类间距离,即类与类之间的距离越小,多数类和少数类则越接近。为了增加边界少类样本的采样权重,需要增加该少数类簇的权重。类间距离 L_i 如式(6)所示。

$$L_i = \frac{1}{|O|}\sum_{q=1}^{|O|} \| a_i - o_q \|_2 \tag{6}$$

其中, i 从 1 到 $|A|$,q 从 1 到 $|O|$。

当多数类经 FCM 聚类之后簇数较多时,不同少数类簇与所有多数类簇的距离之和差距较小。因此,为了增强少数类与多数类的类间距离,通过式(7)将 L_i 变换得到 R_i,并对 R_i 通过式(8)进行标准化处理。

$$R_i = \frac{1}{e^{L_i}} \tag{7}$$

$$\hat{R}_i = \frac{R_i - \min R_i}{\max R_i - \min R_i} \tag{8}$$

其中,min 表示最小值函数,max 表示最大值函数。

第二个指标称为簇的大小。描述了少数类各簇中所含样本的多少。若该簇所含样本数较多,则过采样时该簇将越应该着重考虑。

为了降低数量级带来的负面影响,式(9)用于获得标准化的簇大小 $\hat{n}_i$。

$$\hat{n}_i = \frac{n_i - \min n_i}{\max n_i - \min n_i} \tag{9}$$

其中,min、max 含义同上。

此外,使用上述两个指标的参数 λ 和 θ 的加权组合,来构建新的指标 F_i,如式(10)所示。

$$F_i = \lambda \hat{R}_i + \theta \hat{n}_i \tag{10}$$

其中, $\lambda \in [0,1]$;$\theta \in [0,1]$;λ、θ 表示相应指标的重要性。

为了对每个少数类自适应分配相应数量的合成样本,利用式(11)对 F_i 通过归一化处理。$\hat{F}_i$ 代表着每个簇合成的新样本数量所占比例。通过式(12),为少数类的簇 i 分配 h_i 个新的少数类样本。

$$\hat{F}_i = \frac{F_i}{\sum_{i=1}^{|A|} F_i} \tag{11}$$

$$h_i = \hat{F}_i \cdot G \tag{12}$$

2.4 基于特征空间的过采样

对少数类的每个簇进行过采样,最终在每个簇

中分别进行新样本的合成。传统 SMOTE 的过采样，是通过目标少类样本与其近邻样本的线性插值进行新样本的合成，其合成质量的优劣主要取决于随机因子的大小。该随机因子通常取 0～1 之间某个值，若此随机因子取值不当，便直接造成合成样本质量下降，甚至导致合成冗余样本。因此，随机因子的取值至关重要。

为了避免单一随机因子取值带来的偶然性，本文通过考虑少数类样本自身的数据分布，提出一种新的合成样本策略。首先确定少类样本簇 i 的特征边界，簇 i 的样本 j 表示为 $x_j^i(i=1,2,\cdots,a_i;j=1,2,\cdots,n_i)$，簇 i 中样本 j 的特征 f 表示为 $x_{jf}^i(f=1,2,\cdots,b)$，找到对于簇 i 中所有样本的特征 f，计算其均值 μ_f^i 和方差 σ_f^i。然后，对于簇 i，使用公式(13)、(14)确定所有样本特征 f 的上边界 u_f^i 和下边界 l_f^i。

$$u_f^i=\mu_f^i+\eta\sigma_f^i \tag{13}$$

$$l_f^i=\mu_f^i-\eta\sigma_f^i \tag{14}$$

其中，$\eta\in R$ 控制标准差的影响范围，即边界范围。在 u_f^i 和 l_f^i 之间生成一个随机数 δ，利用公式(15)得到新合成的簇 i 中样本 p 的特征 s_{pf}^i，

$$s_{pf}^i=\delta l_f^i+(1-\delta)u_f^i \tag{15}$$

其中，$\delta\in[0,1]$，$l_f^i\leqslant s_{pf}^i\leqslant u_f^i$。

至此即可得到第 i 个簇的第 p 个新合成的样本 $s_p^i=\{s_{p1}^i,s_{p2}^i,\cdots,s_{pi}^i,\cdots,s_{pf}^i\}$ $(p=1,2,\cdots,h_i;f=1,2,\cdots,b)$。重复上述步骤，直至合成 G 个样本，最后将所有合成的样本添加至集合 S 中。

2.5　算法描述

I 为训练集，$I=\{(x_1,y_1),(x_2,y_2),\cdots,(x_L,y_L)\}$；$L$ 是训练集样本的总个数，其中少数类样本数为 T。样本 $\boldsymbol{x}_i\in\boldsymbol{R}^b$，是 b 维特征向量，类标签为 $y_i\in\{P,N\}$，P 对应少数类(正样本类)，N 对应多数类(负样本类)。

ADWEBDO 方法的完整步骤如下：

输入：训练集 I、参数 K

输出：合成数据集 S

Step 1　根据式(5)计算需要合成的新少类样本个数 G；

Step 2　对数据集 I 中的所有少数类样本计算其 K 近邻，对 K 个近邻均是多数类样本的少数类样本进行剔除，同理对于多数类样本也进行此操作，剔除噪声样本后的数据集记为 X；

Step 3　利用 FCM 聚类算法，对数据集 X 中的所有少数类样本和多数类样本分别进行聚类分析，根据 Xie-Beni 指标，产生 $|A|$ 个少类簇和 $|O|$ 个多类簇；

Step 4　根据双指标，即类间距离和簇的大小，给每个少类簇自适应分配相应数量的合成样本数(利用式(6)～(12)计算得到)；

Step 5　对每个簇而言，利用式(13)～(15)合成新的少类样本，并重复合成样本的操作，直至满足相应簇需要合成的样本数，并将所有新合成的样本添加至集合 S 中。

3　实验与结果分析

3.1　模型评价指标

本文采用 FCM 聚类算法对少数类样本和多数类样本分别进行聚类分析，Xie-Beni(XB)的度量标准是确定 FCM 算法需要预先设置的最佳聚类数。此度量标准包含有关每个样本和数据结构的信息。其表达形式如式(16)所示。

$$XB=\frac{\sum_{i=1}^{C}\sum_{j=1}^{n}u_{ij}^m\|x_j-\bar{c}_i\|_2^2}{n\cdot\min\limits_{i,i\neq j}\|\bar{c}_i-\bar{c}_j\|_2^2} \tag{16}$$

其中，x_j 表示数据集中第 j 个样本；元素 u_{ij} 代表样本 x_j 属于第 i 个簇的隶属度；参数 $m>1$ 表示模糊因子；$\bar{c}_i$ 表示数据集的第 i 个簇的聚类中心；$\bar{c}_j$ 表示数据集第 j 个簇的聚类中心。式中分子表示用于测量聚类内部的紧密度，而分母表示用于测量类与类之间的距离。XB 度量值越接近 0，则聚类效果越好。

在二元分类问题中，对于不平衡数据的评价方法，大多都建立在混淆矩阵基础之上，见表 1。

表 1　混淆矩阵

Tab. 1　Confusion matrix

	预测正类	预测负类
真实正类	*TP*	*FN*
真实负类	*FP*	*TN*

对于类别不平衡问题，主要关注样本数量少的类是否可以被正确分类。因此，对于不平衡数据的分类，选择准确率(*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、*F*1 值(*F*1 − *score*) 和 *ROC* 曲线下面积(*AUC*) 作为评价指标，其计算方式如式(17)～(20)所示。*Accuracy*、*Precision*、*Recall*、*F*1 − *score* 和 *AUC* 的值越大，意味着预测模型的性能越好。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

$$Precison=\frac{TP}{TP+FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{20}$$

3.2 实验设置

为了研究 ADWEBDO 方法的有效性,本文选择来自 UCI 数据库的 7 个不平衡数据集进行实验,数据集信息见表 2。

表 2 UCI 数据集

Tab. 2 Datasets from UCI

数据集	特征维数	样本总数	少类样本数	不平衡率
BT[1]	4	748	178	3.2
HCV	28	1 385	332	3.2
Yeast	8	693	30	22.1
Abalone	7	731	42	16.4
Libras	90	360	72	4.0
PC1	37	759	61	11.4
Haberman	3	306	81	2.8

(注释:BT 的全称为 BloodTransfusion)

本文选择多层感知机(Multilayer Perceptron, MLP)作为分类器,MLP 的参数均使用默认参数。实验中对比了 5 种传统的过采样方法,分别为 SMOTE、ADASYN、Borderline - SMOTE1 (BS1)、Borderline-SMOTE2(BS2)和 CBSO。为了客观比较各个方法,实验将数据集的 2/3 作为训练集,1/3 作为测试集,使用十折交叉验证,重复 5 次取均值,作为最终实验结果。

所有实验均是在一台 Ubuntu 操作系统的电脑上实现,其主要参数为: 2.2 GHz CPU、16 GB 内存,同时借助 Python 语言编程实现。其它参数选择如下: K 值的选取参照文献[15]设置为 5,FCM 算法中参考文献[19] m 取2,根据文献[20], η 选取2时,实验结果最佳。根据 FCM 聚类算法的评价指标,各数据集的最佳少数类簇数和最佳多数类簇数如图 3、图 4 所示。

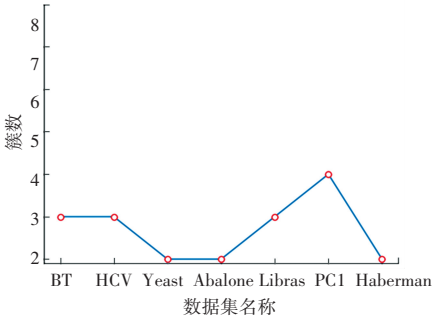

图 3 7 个数据集的多数类的最佳聚类簇数示意图

Fig. 3 Schematic diagram of the optimal number of clusters for the majority class of seven data sets

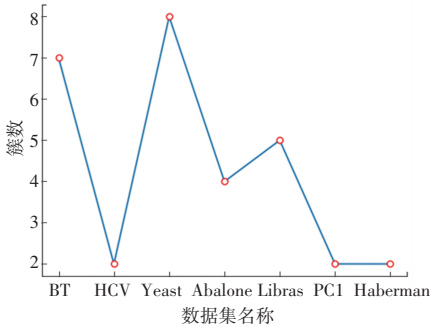

图 4 7 个数据集的少数类的最佳聚类簇数示意图

Fig. 4 Schematic diagram of the optimal number of clusters for the minority class of seven data sets

最佳 λ 依次为:0.9、0.9、0.7、0.5、0.8、0.5、0.8,最佳 θ 依次为:0.1、0.1、0.3、0.5、0.2、0.5、0.2。

3.3 结果分析

实验结果见表 3~表 7,其中对最优结果均已进行加粗标注。

表 3 6 种过采样方法在 7 组数据集上的 Accuracy 结果

Tab. 3 Results of 6 oversampling methods on Accuracy for 7 datasets

数据集	SMOTE	ADASYN	BS1	BS2	CBSO	ADWEBDO
BT	0.708	0.665	0.733	0.715	0.730	**0.751**
HCV	0.704	0.650	0.756	0.647	0.658	**0.761**
Yeast	0.887	0.856	**0.898**	0.855	0.890	0.891
Abalone	0.902	0.874	0.901	0.890	0.903	**0.913**
Libras	**0.979**	0.958	0.962	0.956	0.976	0.960
PC1	0.706	0.700	0.716	0.697	0.699	**0.788**
Haberman	0.589	0.566	0.591	0.501	**0.621**	0.599

表 4　6 种过采样方法在 7 组数据集上的 Precision 结果

Tab. 4　Results of 6 oversampling methods on Precision for 7 datasets

数据集	SMOTE	ADASYN	BS1	BS2	CBSO	ADWEBDO
BT	0.733	0.669	0.783	0.806	**0.825**	0.805
HCV	0.658	0.550	0.716	0.678	0.606	**0.731**
Yeast	0.788	0.767	0.759	0.755	0.790	**0.892**
Abalone	0.803	0.730	0.747	0.806	0.818	**0.853**
Libras	0.973	0.938	0.958	0.924	**0.981**	0.960
PC1	0.764	0.740	0.742	0.713	0.746	**0.788**
Haberman	0.594	0.587	0.692	0.565	0.593	**0.699**

表 5　6 种过采样方法在 7 组数据集上的 Recall 结果

Tab. 5　Results of 6 oversampling methods on Recall for 7 datasets

数据集	SMOTE	ADASYN	BS1	BS2	CBSO	ADWEBDO
BT	0.826	0.752	0.830	0.709	0.885	**0.891**
HCV	0.517	0.589	0.620	0.582	0.683	**0.689**
Yeast	0.810	0.749	0.836	0.879	0.768	**0.897**
Abalone	0.641	0.654	0.665	0.658	**0.703**	0.696
Libras	**0.993**	0.966	0.938	0.914	0.948	0.969
PC1	0.745	0.678	0.768	0.758	0.747	**0.775**
Haberman	0.525	0.517	**0.586**	0.544	0.584	0.560

表 6　6 种过采样方法在 7 组数据集上的 F1-score 结果

Tab. 6　Results of 6 oversampling methods on F1-score for 7 datasets

数据集	SMOTE	ADASYN	BS1	BS2	CBSO	ADWEBDO
BT	0.722	0.658	0.710	0.714	0.760	**0.801**
HCV	0.744	0.727	0.768	0.635	**0.797**	0.743
Yeast	0.890	0.860	0.901	0.862	**0.911**	**0.911**
Abalone	0.906	0.883	0.910	0.906	0.908	**0.923**
Libras	0.980	0.968	0.961	0.957	**0.988**	0.952
PC1	0.704	0.700	0.712	0.715	0.746	**0.790**
Haberman	0.668	0.685	0.690	0.580	0.683	**0.698**

表 7　6 种过采样方法在 7 组数据集上的 AUC 结果

Tab. 7　Results of 6 oversampling methods on AUC for 7 datasets

数据集	SMOTE	ADASYN	BS1	BS2	CBSO	ADWEBDO
BT	0.853	0.754	0.825	0.840	0.803	**0.876**
HCV	0.548	0.549	0.608	0.572	0.592	**0.644**
Yeast	0.781	0.750	0.842	0.868	0.877	**0.890**
Abalone	0.798	0.738	0.767	0.794	**0.817**	0.797
Libras	0.796	0.750	0.817	0.794	0.767	**0.834**
PC1	0.750	0.837	0.805	0.812	0.802	**0.872**
Haberman	0.703	0.689	0.675	0.648	0.699	**0.752**

由表 3 可见，对于 Accuracy 评价指标，本文提出的算法，在 BT、HCV、Abalone 和 PC1 4 个数据集上均明显优于其它过采样算法。

表 4 中的 Precision 指标，ADWEBDO 在 5 个数据集的表现均优于其它 5 种过采样方法。其中，在 HCV 数据集上 ADWEBDO 比表现较差的 ADASYN 提高了 1.81 个百分点。虽然 ADWEBDO 在 BT 和 Libras 两个数据集上表现不是最好，但其综合排名为第三，表现良好。

Recall 的结果见表 5。在 7 个数据集中，本文算法仅在 4 个数据集上的表现优于其它过采样方法。

表 6 中的 $F1-score$ 值，在 5 个数据集上均优于其它过采样方法。CBSO 在 3 个数据集上表现良好，其结果仅次于 ADWEBDO。

AUC 作为不平衡数据分类的重要指标之一，由表 7 可知，ADWEBDO 在 7 个数据集中有 6 个均是最优结果，这表示 ADWEBDO 具有较好的泛化能力。

通过对比 6 种过采样方法在 7 个 UCI 数据集上的表现，ADWEBDO 过采样在 Accuracy、Precision、

Recall、F1-score 和 AUC 上表现相较于其它 5 种过采样方法,均取得了不错的结果。

4 结束语

在不平衡数据的分类问题中,多类样本和少类样本在数量上差距较大,导致分类器的分类性能急剧下降。因此,在实际的分类任务中,必须有效地处理数据不平衡问题。本文提出一种基于边界增强和去噪的自适应双权重过采样方法(ADWEBDO),考虑类间距离的同时,也考虑了少类各样本簇的规模,增加了对边界少类样本的采样权重,一定程度上避免了类重叠现象的产生。同时,基于少类簇特征空间合成新样本策略,使得合成的样本更加合理。实验结果表明,ADWEBDO 在 7 个不同规模、不同不平衡率的数据集上性能表现稳定,对不平衡数据分类问题的学习具有一定的指导作用。

参考文献

[1] 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报, 2009, 4(2): 148-156.

[2] C G H A B, B L Y A, D J S, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73:220-239.

[3] 刘定祥, 乔少杰, 张永清, 等. 不平衡分类的数据采样方法综述[J]. 重庆理工大学学报(自然科学), 2019, 33(7): 102-112.

[4] WANG S. A Comprehensive Survey of Data Mining - Based Accounting - Fraud Detection Research [M]. IEEE Computer Society, 2010,1:50-53.

[5] YU H, NI J, DAN Y, et al. Mining and Integrating Reliable Decision Rules for Imbalanced Cancer Gene Expression Data Sets [J]. Tsinghuaence and Technology, 2012, 17(6): 666-673.

[6] 胡至洵, 杜宇, 刘潇月. 基于用户兴趣分类的书籍自动推荐系统设计[J]. 现代电子技术, 2021, 44(6): 58-62.

[7] F. Provost, Machine Learning from Imbalanced Data Sets 101, Proc[C]//Learning from Imbalanced Data Sets: Papers from the Am. Assoc. Artificial Intelligence Workshop, 2000, 68(2000): 1-3.

[8] 石凤兴. 针对类内不平衡样本分类方法的研究[D]. 哈尔滨: 哈尔滨工业大学, 2016.

[9] 刘杜钢. 基于聚类和类重叠分析的近邻分类[J]. 计算机系统应用, 2015, 24(9): 1-8.

[10] JO T, JAPKOWICZ N. Class Imbalances versus Small Disjuncts [J]. ACM SIGKDD Exploration Newsletter, 2004, 6(1): 40-49.

[11] 曹婷婷, 张忠林. 代价敏感的 KPCA-Stacking 不均衡数据分类算法[J]. 计算机工程与科学, 2021, 43(3): 525-533.

[12] Maldonado, Sebastin, Montecinos C. Robust classification of imbalanced data using one - class and two - class SVM - based multiclassifiers[J]. Intelligent Data Analysis, 2014, 18(1): 95-112.

[13] 陈丽芳, 代琪, 赵佳亮. 不平衡数据多粒度集成分类算法研究[J]. 计算机工程与科学, 2021, 43(5): 917-925.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority over - sampling technique [C]// Artifificial Intell, 2002: 321-357.

[15] HE H. Adasyn: Adaptive synthetic sampling approach for imbalanced learning[C]// in Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell., 2008: 1322-1328.

[16] HAN H, WANG W Y, MAO B H. Borderline-smote: A new oversampling method in imbalanced data sets learning [C]// in Proc. Int. Conf. Intell. Comput., 2005: 878-887.

[17] BARUA S, ISLAM M M, YAO X, et al. MWMOTE majority weighted minority oversampling technique for imbalanced data set learning[J]. IEEE Trans. Knowledge and Data Eng., 2014, 26(2):405-425.

[18] BARUA S, ISLAM M M, MURASE K. A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning[C]// Springer-Verlag. Springer-Verlag, 2011: 735-744.

[19] PAL N R, BEZDEK J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy systems, 1995, 3(3):370-379.

[20] SHARMA S, BELLINGER C, KRAWCZYK B, et al. Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance[C]// IEEE International Conference on Data Mining. IEEE, 2018:447-456.

(上接第 57 页)

[11] MARINO R, SCALZI S, NETTO M. Nested PID steering control for lane keeping in autonomous vehicles[J]. Control Engineering Practice, 2011, 19(12) : 1459-1467.

[12] ANG R J, YIN G D, JIN X J. Robust adaptive sliding mode control for nonlinear four - wheel steering autonomous vehicles path tracking systems[C]// Proceedings of the 8th International Power Electronics and Motion Control Conference. Hefei, China: IEEE, 2016: 1 -8.

[13] ANTONELLI G, CHIAVERINI S, FUSCO G. A fuzzy - logic - based approach for mobile robot path tracking [J]. IEEE Transactions on Fuzzy Systems, 2007, 15(2) : 211-221.

[14] HAJJAJI A E, BENTALBA S. Fuzzy path tracking control for automatic steering of vehicles [J]. Robotics & Autonomous Systems, 2003, 43(4) : 203-213.

[15] BEN-MESSAOUD W, BASSET M, LAUFFENBURGER J P, et al. Smooth Obstacle Avoidance Path Planning for Autonomous Vehicles[C]// 2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES). IEEE, 2018:1-6.