基于随机数据块与权重采样的不平衡分类集成算法

魏 勋

（江西理工大学 软件工程学院，江西 南昌 330000）

摘 要：在大量的真实问题中，数据集往往是类别不平衡的，很可能会削弱学习算法的性能。为了处理不平衡数据集，业界提出了各种类别不平衡学习算法，其中包括不少集成算法。然而，这些集成算法主要考虑在样本层面进行集成而忽视了特征层面，且常规的随机采样算法未能重点关注边界区域，此区域通常是分类困难样本。鉴于此，提出一种名为BRPE的集成采样算法进行优化。首先，对特征集进行采样；其次，以多数类样本距离少数类样本的最近距离作为权重对多数类样本进行下采样，得到一个平衡的随机数据块并将其作为训练子集；再次，在训练子集上训练一个基学习器；最后，将所有基学习器的输出组合成预测结果。在10个合成数据集和8个真实数据集上均进行了详细实验。结果表明，相比其他4种不平衡集成分类算法，BRPE能够取得更高的F1和AUC值。

关键词：不平衡数据；类别不平衡学习；集成算法；权重采样；随机数据块

DOI：10.11907/rjdk.241780 **开放科学（资源服务）标识码（OSID）：**

中图分类号：TP181；TP311.13 **文献标识码：**A **文章编号：**1672-7800（2025）003-0043-05

An Ensemble Algorithm Based on Random Patches and Weighted Sampling for Imbalanced Data Classification

WEI Xun

（*Software Engineering Academy*, *Jiangxi University of Science and Technology*, *Nanchang* 330000, *China*）

Abstract: In many real-world problems, the datasets are typically imbalanced which probably degenerate the learning algorithm. To handle these skewed datasets, there are many class imbalance learning methods are proposed, especially ensemble methods due to their efficiency. While most of these ensemble methods mainly focus on the level of samples and neglect the features aspect. And conventional random sampling method do not pay enough attention to the boundary which always contain hard classified samples. Propose an ensemble sampling method named BRPE to overcome this deficiency. BRPE firstly samples a feature subset; then down-sample majority class instances via its closest euclidean distance to minority class samples to create a balanced random patch as training subset; then trains a base learner using each of subsets, and finally obtains the output combined of these learners. Experiments on both 10 synthetic datasets and 8 real-world datasets show that BRPE can achieve higher F1 and AUC values than other four existing ensemble methods for class imbalance.

Key Words: imbalanced data; class imbalance learning; ensemble algorithm; weighted sampling; random patch

0 引言

不平衡数据通常是指数据集中的一个或一些类别样本占据大多数，这种情况在真实场景十分常见，一般被称为类别不平衡。例如：在收到大量常规邮件的同时，偶尔会收到一些垃圾邮件；在疾病检测中，检测结果为健康的通常占大多数，检测结果为患病的占少数[1]。业界通常使用不平衡比例，即多数类别样本数与少数别样本数之比，用来衡量一个数据集的不平衡程度。由于这种不平衡的样本分布，学习算法很容易发生退化。

为了解决该问题，业界通常使用对数据集进行采样的方法，通过重新调整多数类和少数类样本的数量大小以平衡训练集。采样法大致可分为3种：下采样、上采样和集成采样。相比使用单一学习器的方法，集成采样方法通常表现出更高的性能。然而，常规的集成采样算法一般只在样本层面进行采样，忽视了特征层面。

通常而言，多数类样本与少数类样本的边界区域分类

收稿日期：2024-11-11
扫描二维码阅读全文：

作者简介：魏勋（1990-），男，硕士，江西理工大学软件工程学院助教，研究方向为机器学习、数据挖掘、计算机视觉。

较为困难。而常规的采样算法一般采用随机的方式，未能重点关注分类困难区域，导致算法性能下降。

本文提出一种平衡随机数据块集成算法（Balanced Random Patches Ensemble，BRPE）以解决现实中类别不平衡的分类问题。该算法基于随机数据块算法和EasyEnsemble算法进行改进，在集成采样过程中不仅对样本采样，同时也对特征进行采样[2-3]。在采样过程中，根据样本之间的欧式距离进行权重采样，重点关注边界区域。相比于其他集成方法，该方法能够取得更高的F1值和AUC值。

1 相关工作

目前，业界存在许多方法以解决类别不平衡问题，其中采样法是一种简单且高效的方法。采样法大致可以分为3类：下采样、上采样和集成采样。

下采样和上采样一般有两种实现方式：随机方法和启发式方法。随机下采样通过对多数类样本随机地下采样以达成类别平衡，而随机上采样则通过随机重复少数类样本以平衡数据集。NearMiss是一种启发式的下采样方法，基于距离度量筛选那些离少数类样本最近的多数类样本[4]。One-sided Selection通过去除“边界”和“噪声”多数类样本，从而找到一个具有代表性的多数类样本子集以实现类别平衡[5]。SMOTE在相邻的少数类样本中随机新增合成的少数类样本[6]。文献[7]提出了一种边界过采样的图节点不平衡分类算法以提升生成样本的多样性。文献[8]提出了一种联合迁移学习和强化学习的不平衡样本分类模型，能够基于上采样方法对少数类进行样本生成。文献[9]提出了一种面向分类困难区域的DR-XGBoost模型，增强算法对边界样本和少数类样本的关注度。文献[10]提出了一种跨类别样本迁移框架下的不平衡分类方法，能够实现类别交叠区域中样本数目和分布的平衡。而针对传统采样算法固定的采样概率，文献[11]提出了一种自适应采样的不平衡分类算法。不难看出，由于丢弃了大量的多数类样本，下采样方法可能会影响学习器的性能；而上采样方法则可能会导致学习器过拟合，原因在于大量重复或者合成的少数类样本[12]。

目前，业界流行使用集成学习算法以解决类别不平衡问题，主要是因为集成算法通常具有更高的性能。RUSBoost通过对AdaBoost算法进行修改，即每次迭代时对训练集中的多数类样本进行下采样，以实现类别平衡[13]。BalanceBagging是一种保持训练集平衡的Bagging集成方法，每次使用自助法从多数类样本和少数类样本进行采样得到类别数量相同的训练子集[14]。Balanced Random Forests是一种对随机森林算法进行改造后的高效类别不平衡学习算法[15]。与随机森林算法的不同之处在于，它在每次自助采样的同时对多数类样本进行下采样以得到类别平衡的训练子集。EasyEnsemble是一种十分直接的方法，与Balanced Random Forests类似，通过对多数类样本进行下采样以得到若干个类别平衡的训练子集。而不同之处在于，前者基于训练子集去训练一个AdaBoost学习器，而后者则是训练一个决策树学习器。目前，大部分集成采样方法主要考虑在数据的样本层面上进行采样，而忽视了数据的特征空间。经实验验证，同时对数据的样本和特征进行采样，可能有益于集成模型的性能表现。

2 BRPE算法

随机下采样会丢弃大量的潜在有用信息，且随机采样一部分的多数类样本会影响模型稳定性，对边界区域关注度不够也会导致边界样本分类困难。EasyEnsemble算法能够较好地克服部分问题，通过多次对多数类样本进行随机下采样以构建不同的平衡训练子集。但是该算法主要关注数据的横向层面，即样本层面，却忽视了数据的纵向层面，即特征空间。经过充分的实验验证，同时对样本和特征进行采样的Random Patches算法比单独对样本采样的Pasting Rvotes算法更能提高集成模型的性能。而对于分类困难的边界样本，EasyEnsemble并未在采样过程中给予更高的关注度，这容易导致边界样本分类困难。因此，本文提出了BRPE算法，该算法对数据的横向与纵向层面均进行采样以构建若干个平衡随机数据块作为训练子集，并对多数类样本进行权重采样，重点关注边界区域。BRPE算法伪代码如算法1所示。为方便表示，本文将少数类视为正例，多数类视为负例。

算法1 BRPE算法

1. 输入：
2. P：正例样本集
3. N：负例样本集
4. F：样本特征集
5. α：特征集采样比例
6. n：采样子集数量
7. t：训练AdaBoost模型H_i的迭代次数
8. 步骤：
9. for i=1 to n do
10. 在特征集F上按照比例α进行采样得到P_i和N_i
11. 对N_i每个样本计算其到P_i最小欧式距离$Dist_{min}$
12. 以$Dist_{min}$的反比作为权重从N_i中采样一个子集N_i'，且$|N_i'|=|P_i|$
13. 使用N_i'和P_i组合成训练子集D_i
14. 基于D_i训练基分类器H_i，迭代t次
15. end for
16. 输出：
17. $H(x)=\mathrm{sgn}(\sum H_i-\theta)$

上述算法中，θ表示集成模型的分类阈值，即一个样本被预测为正例的投票数，该值被定为$n/2$。对Ni每个样本计算其到P_i的最小欧式距离$Dist_{min}$，其定义如式（1）所示。

$$Dist_{min}(X\in N_i)=min_{p\in P_i}(\sqrt{\sum(x_i-p_i)^2}) \quad (1)$$

以$Dist_{min}$的反比作为权重从Ni中采样一个子集，也即给予边界的多数类样本更大的采样权重，这样能够重点关

注边界区域分类困难的样本,降低算法在学习过程中的偏差,进而提高分类准确度。众所周知,一个模型的泛化误差大致可以被分解为3个部分:偏差、方差和噪声,如式(2)所示[16]。

$$E(f:D) = bias2(x) + var(x) + \varepsilon^2 \tag{2}$$

BRPE通过对正例样本P和负例样本集N进行集成采样得到若干个训练子集:$D_{en}=\{D_1,D_2,\cdots,D_n\}$;然后在这些训练子集上训练基分类器:$\{H_1,H_2,\cdots,H_n\}$;最终得到集成模型:$H(X)=\sum_{i=0}^{n}\frac{H_i(x)}{n}$。集成模型的方差如式(3)所示。

$$Var(H) = Var\left[\frac{\sum H_i(x)}{n}\right] \tag{3}$$

令Var(D_{mean})表示样本分布的方差,如式(4)所示。

$$Var(D_{mean}) = Var\left(\frac{D_{en}}{n}\right) = Var\left[\frac{\sum D_i}{n}\right] \tag{4}$$

Breiman[17]曾指出,Bagging模型的方差$Var(H)$与数据分布的方法$Var(D_{mean})$呈正相关关系。

在每次采样中,训练子集D_i包含了P_i和N_i'。很明显,P_i和N_i'相互独立,故有式(5)。

$$Var(D_i') = Var(P_i) + Var(N_i') \tag{5}$$

因此,式(4)可以表示为如式(6)所示。

$$Var(D_{mean}) = Var\left[\frac{\sum P_i}{n}\right] + Var\left[\frac{\sum N_i'}{n}\right] \tag{6}$$

如果每个训练子集都互相独立,那么集成模型的数据分布的方差会显著降低,如式(7)所示。

$$Var\left[\frac{\sum D_i}{n}\right] = \frac{Var(D_i)}{n} \tag{7}$$

如果所有训练子集都一样,则有式(8)。

$$Var\left[\frac{\sum D_i}{n}\right] = Var(D_i) \tag{8}$$

此时并不能降低集成模型的数据分布方差。而对于BRPE而言,每次采样生成的训练子集既不是完全独立,也不是完全相同。用ρ表示两个训练子集之间的相关性,并用σ^2表示方差,可以得到如式(9)所示[18]。

$$Var\left[\frac{\sum D_i}{n}\right] = \rho \times \sigma^2 + (1-\rho) \times \frac{\sigma^2}{n} \tag{9}$$

可以看出,当训练子集相关性为0时,即完全独立的时,如式(7)所示。当训练子集相关性为1时,即完全相同时,如式(8)所示。进一步,用ρ_P表示两两训练子集中正例样本的相关性,ρ_N表示两两训练子集中负例样本的相关性,由于训练子集正例数量和负例数量相等,可以得到式(10)。

$$\rho = (\rho_P + \rho_N)/2 \tag{10}$$

由于对特征集F进行了α比例的采样,可知$\rho_P = \alpha$。而由于负例样本是按距离进行权重采样,无法直接计算两两训练子集负例样本相同的期望概率,此处假设对负例样本进行随机下采样。理论上,除非采样权重方差十分巨大,此处假设是有效的。令数据集不平衡比例为r,则两两训练子集负例样本相同的概率为$1/r$,可知$\rho_N = \alpha/r$。而对于EasyEnsemble算法,$\rho_P = 1$,$\rho\_N = 1/r$。可明显看出,本文BRPE算法得到的集成模型理论上方差更低。值得注意的是,特征集F采样比例α越低,模型方差越小。但是如果数据集特征数量不足且α值过低,这会导致学习器难以学习足够的信息,进而增大模型偏差,可能会使得模型泛化性能退化。因此,对于特定的数据集,需要考虑数据集的特征数量,并设置合适大小的α值。

相比其他不平衡分类集成算法,BRPE算法的创新之处主要有两点:①BRPE不仅在样本层面采样,同时也对特征进行采样;②BRPE更加关注边界分类困难样本,对边界区域的多数类样本给予更大的采样权重。

3　实验过程与结果

3.1　评价指标

当数据存在类别不平衡问题时,对于分类器而言,准确率并非一个合适的评价标准。因此,本文使用正例F1值和AUC值作为分类器性能的评价标准。F1值和AUC值的计算均基于表1所示的混淆矩阵。

Table 1　Confusion matrix

表1　混淆矩阵

	预测正例	预测负例
实际正例	TP	FN
实际负例	FP	TN

F1值是一种综合考虑精确率和召回率的评价指标,能够很好地评估少数类样本的分类性能,AUC是一种高效且鲁棒的分类性能指标。给定一个二分类问题,改变预测阈值并产生若干个(FPR,TPR)对(FPR表示False Positive Rate,TPR表示True Positive Rate)。基于这些值对能够绘制出ROC曲线,而AUC值即为ROC曲线下的面积。对于类别不平衡数据而言,AUC已被证明是一种可靠的性能指标。相关指标计算如式(10)—式(13)所示。

$$FPR = FP/(FP + TN) \tag{10}$$

$$Recall_+ = TPR = TP/(TP + FN) \tag{11}$$

$$Precision_+ = TP/(TP + FP) \tag{12}$$

$$F1 = \frac{2 \times Precision_+ \times Recall_+}{Precision_+ + Recall_+} \tag{13}$$

3.2　实验设置

为了方便考虑,本文只对二分类问题进行实验。对于每个数据集,本文采用5折交叉验证,重复10次。最终的AUC值和G-mean值是这十次交叉验证结果的平均值。所有算法和实验都是用Python实现,并使用Scikit-Learn库作为主要框架。

本文共对5种集成模型进行了实验,这5种集成模型的简要说明与实验设置如下:

(1)BalancedBagging(BB):一种训练子集平衡的Bagging算法,这里采用下采样法保持子集平衡。基分类器使用CART,数量设置为100。

（2）BalancedRandomForests（BRF）：一种改进的平衡随机森林算法，基分类器数量设置为100。

（3）RUSBoost（RUSB）：一种基于AdaBoost改进的算法，通过在每次迭代中对样本进行下采样以达到平衡。基分类器使用CART，数量设置为100。

（4）EasyEnsemble（EE）：通过对多数类样本进行下采样以得到M个类别平衡的训练子集。这里M设置为10。基分类器为AdaBoost，并使用CART作为AdaBoost的基学习器，AdaBoost迭代次数为10。

（5）BRPE：本文提出的集成采样算法。采样子集个数设置为10，基分类器设置与EE相同。

值得一提的是，上述5种集成算法中使用的CART模型均设置最小划分样本数量为10以避免模型过拟合。此外，对于BRPE的α值，本文实验取[0.4，0.8]区间，从0.4开始，按0.1递增至0.8。

3.3 合成数据实验

本文在合成数据集上对BRPE进行实验，如上文所述，数据集的特征数量理论上对BRPE算法有很大影响，这里通过实验进行验证。此处合成了10个二分类的数据集，样本数量为1 000，类别比例设置为（0.9，0.1）。有效特征数量分别设置为10个到100个，按10递增，即合成数据集$Synth_i$包含$10 \times i$个有效特征。

由表2可知，几乎所有模型的F1值与AUC值都在随着特征数量增加而衰减，原因可能是任务的难度越来越高。对于F1值而言，当特征数量较少时，EE与BRPE表现相当，且此时α值较大的BRPE效果更好。随着特征数量增长，α值较小的BRPE表现出比其他模型明显更好的性能，能够提升超过2%。对于AUC值而言，当特征数量较少时，EE的效果比BRPE都要好，尤其是当BRPE的α值较小时。如上文所述，当数据集特征数量不足且α值过低时，会导致BRPE的基学习器难以学习足够的信息，进而增大模型偏差，使模型泛化性能发生退化。当特征数量达到40后，BRPE表现出比其他模型更高的性能，能够提升超过1.5%。并且，随着特征数量的增加，α值较小的BRPE比α值较大的BRPE表现更佳。这意味着对于此类合成数据，特征应达到一定数量，才能体现出BRPE算法的优势。同时，从实验结果可以看出，对于此类合成数据，比较适宜的α值应设置为0.4~0.6，这样能够同时兼顾模型的方差和偏差。

Table 2 The F1 value of ensemble models on synthetic datasets

表2 各集成模型在合成数据集上的F1值

F1	$Synth_1$	$Synth_2$	$Synth_3$	$Synth_4$	$Synth_5$	$Synth_6$	$Synth_7$	$Synth_8$	$Synth_9$	$Synth_{10}$	平均
BB	0.573	0.529	0.442	0.512	0.376	0.421	0.407	0.379	0.359	0.338	0.433
BRF	0.608	0.502	0.443	0.489	0.355	0.405	0.399	0.375	0.355	0.333	0.426
RUSB	0.575	0.397	0.285	0.297	0.192	0.215	0.203	0.225	0.190	0.189	0.277
EE	0.636	**0.562**	0.459	0.490	0.374	0.402	0.387	0.365	0.360	0.341	0.437
BRPE（α=0.4）	0.556	0.526	0.466	0.509	**0.395**	0.429	0.415	**0.387**	**0.382**	0.361	0.443
BRPE（α=0.5）	0.617	0.548	0.461	0.514	0.390	**0.431**	**0.416**	0.369	0.370	**0.363**	0.448
BRPE（α=0.6）	0.624	0.560	0.466	**0.525**	0.381	0.428	0.406	0.386	0.372	0.357	**0.450**
BRPE（α=0.7）	**0.653**	0.561	0.470	0.505	0.383	0.419	0.407	0.377	0.367	0.352	**0.450**
BRPE（α=0.8）	0.652	0.558	**0.471**	0.506	0.380	0.415	0.393	0.377	0.367	0.349	0.447

Table 3 The AUC value of ensemble models on synthetic datasets

表3 各集成模型在合成数据集上的AUC值

AUC	$Synth_1$	$Synth_2$	$Synth_3$	$Synth_4$	$Synth_5$	$Synth_6$	$Synth_7$	$Synth_8$	$Synth_9$	$Synth_{10}$	平均
BB	0.844	0.818	0.747	0.776	0.695	0.716	0.696	0.697	0.674	0.649	0.731
BRF	0.884	0.836	0.793	0.816	0.713	0.759	0.751	0.730	0.719	0.700	0.770
RUSB	0.740	0.654	0.596	0.603	0.556	0.566	0.561	0.572	0.552	0.553	0.595
EE	**0.901**	**0.867**	**0.812**	0.816	0.738	0.767	0.748	0.725	0.736	0.715	0.781
BRPE（α=0.4）	0.838	0.836	0.805	0.820	**0.750**	0.778	0.765	0.743	**0.751**	0.731	0.782
BRPE（α=0.5）	0.874	0.853	0.802	0.823	0.745	0.784	**0.772**	0.727	0.740	**0.733**	0.785
BRPE（α=0.6）	0.874	0.856	0.807	**0.833**	0.740	**0.787**	0.769	**0.744**	0.740	0.728	**0.788**
BRPE（α=0.7）	0.891	0.861	0.808	0.824	0.742	0.779	0.766	0.737	0.739	0.725	0.787
BRPE（α=0.8）	0.893	0.860	0.811	0.825	0.742	0.774	0.753	0.738	0.737	0.720	0.785

3.4 真实数据实验

本文进一步在真实数据集上对BRPE进行实验。真实数据集使用UCI上的Wine、Musk2、Letter和Mfeat数据集[19-20]。其中，Mfeat数据集是一个包含0—9手写数字特征的多维度数据集。Mfeat数据集包含2 000个样本，6个特征维度。①Fac：轮廓相关性，216维；②Pix：2×3窗口得到的平均像素值，240维；③Kar：Karhunen-Love系数，64维；④Zer：Zernike片刻，47维；⑤Fou：数字形状的傅立叶系数，76维；⑥Mor：形态学特征，6维。

考虑到BRPE对特征数量的要求，本文采用Mfeat数据集前5个特征维度进行实验，即Fac、Pix、Kar、Zer、Fou。数据集基本情况如表4所示。

实验结果如表5和表6所示。此处BRPE的表现取不同α值效果最好的值。对于F1值而言，BRPE相比其他模型具有明显优势，尤其是在Wine、Musk2和Kar数据集上有显著提高，相比其他模型平均性能分别提升了4.3%、3.6%、2.1%、2.8%。对于AUC值而言，相比其他模型平均性能分别提升了1.2%、0.7%、3.2%、0.4%，这说明BRPE相

Table 4　Details of UCI datasets
表 4　UCI数据集基本情况

数据集	样本量	特征量	正例类别	不平衡比例
Wine	6 497	11	7	5.0
Musk2	6 598	166	1	5.5
Letter	20 000	16	U	23.6
Fac	2 000	216	0	9.0
Pix	2 000	240	0	9.0
Kar	2 000	64	0	9.0
Zer	2 000	47	0	9.0
Fou	2 000	76	0	9.0

Table 5　The F1 value of ensemble models on UCI datasets
表 5　各集成模型在UCI数据集上的F1值

F1	Wine	Musk2	Letter	Fac	Pix	Kar	Zer	Fou	平均
BB	0.576	0.875	0.808	0.938	0.972	0.908	0.929	0.971	0.872
BRF	0.559	0.823	0.837	0.961	0.982	0.944	0.942	0.990	0.879
RUSB	0.517	0.903	0.925	0.970	0.973	0.935	0.940	0.992	0.894
EE	0.557	0.899	0.899	0.939	0.963	0.951	0.932	0.963	0.887
BRPE	**0.583**	**0.930**	**0.936**	**0.978**	**0.978**	**0.969**	**0.953**	**0.993**	**0.915**

Table 6　The AUC value of ensemble models on UCI datasets
表 6　各集成模型在UCI数据集上的AUC值

AUC	Wine	Musk2	Letter	Fac	Pix	Kar	Zer	Fou	平均
BB	0.797	0.952	0.981	0.980	0.988	0.960	0.972	0.987	0.952
BRF	**0.805**	0.945	0.987	0.983	0.989	0.978	0.981	0.992	0.957
RUSB	0.704	0.928	0.963	0.981	0.982	0.948	0.961	0.992	0.932
EE	0.803	0.968	0.992	0.983	0.987	0.980	0.980	0.990	0.960
BRPE	**0.805**	**0.973**	**0.993**	**0.988**	**0.991**	**0.986**	**0.983**	**0.994**	**0.964**

比其他模型具有一定优势。

4　结语

本文提出了一种新型的类别不平衡学习算法BRPE,该算法是一种集成采样算法,受随机数据块集成算法启发,根据多数类样本距少数类样本欧氏距离的大小进行权重采样,重点关注边界区域的分类困难样本。本文通过泛化误差分解理论简单证明了该算法在解决类别不平衡问题上的高效性。接着,在10个合成数据集和8个真实数据集上对BRPE算法的性能进行实验,并与当前流行的4种集成采样算法进行比较,从实验结果F1值和AUC值来看,BRPE表现出了一定优势。

由于对特征的下采样,BRPE算法理论上对于特征丰富的数据集更能发挥优势,且算法所需内存会更低。本文对特征采样比例α值实验的颗粒度较粗,且未系统论证其对模型偏差的影响。详细挖掘α值对集成模型的整体影响,以及在内存受限情况下BRPE的性能表现,是未来值得进一步研究的方向。

参考文献:

[1] WEI X. Research of ensemble classification methods for class-imbalance and cost-sensitive datasets[D]. Hefei: University of Science and Technology of China, 2017.
魏勋. 类别不平衡与代价敏感数据的集成分类方法研究[D]. 合肥:中国科学技术大学,2017.

[2] LOUPPE G, PIERRE G. Ensembles on random patches[C]//Bristol: European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.

[3] LIU X Y, WU J, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 39(2): 539-550.

[4] MANI I, ZHANG I. KNN approach to unbalanced data distributions: a case study involving information extraction[C]//Proceedings of ICML'2003 Workshop on Learning from Imbalanced Datasets, 2003: 1-8.

[5] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]//Proceedings of the Fourteenth International Conference on Machine Learning, 1997: 179-186.

[6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[7] WU T H, DONG M G, TAN R Q. Boundary oversampling based graph node imbalance classification algorithm[J]. Computer Engineering and Applications, 2024, 60(13): 92-101.
武天昊,董明刚,谭若琦. 基于边界过采样的图节点不平衡分类算法[J]. 计算机工程与应用,2024,60(13): 92-101.

[8] HOU C P, HUA Z H, YANG Y, et al. Unbalanced classification method combining transfer learning and reinforcement learning[J]. Computing Engineering and Design, 2022, 43(10): 2769-2776.
侯春萍,华中华,杨阳,等. 联合迁移学习和强化学习的不平衡分类方法[J]. 计算机工程与设计,2022,43(10): 2769-2776.

[9] LI K W, WANG X H, KE C H, et al. An improved method of unbalanced classification based on classification difficult regions[J]. Computing Simulation, 2023, 40(11): 452-458.
李克文,王晓晖,柯翠虹,等. 面向分类困难区域的不平衡分类改进方法[J]. 计算机仿真,2023,40(11): 452-458.

[10] YU H B, LIU J, LI Q W, et al. Imbalanced classification method based on cross-class sample migration framework[J]. Computer Engineering and Applications, 2024, 60(16): 143-158.
于海波,刘婧,李强伟,等. 跨类别样本迁移框架下的不平衡分类方法[J]. 计算机工程与应用,2024,60(16): 143-158.

[11] CHEN Q, XIE J L. Unbalanced classification method based on adaptive sampling[J]. Journal of South China University of Technology (Natural Science Edition), 2022, 50(4): 26-34, 45.
陈琼,谢家亮. 基于自适应采样的不平衡分类方法[J]. 华南理工大学学报(自然科学版),2022,50(4): 26-34, 45.

[12] DRUMMOND C, HOLTE R C. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling[C]//Washington DC: Working Notes ICML Workshop Learning from Imbalanced Data Sets, 2003.

[13] SEIFFERT C. RUSBoost: a hybrid approach to alleviating class imbalance[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2009, 40(1): 185-197.

[14] HIDO S, KASHIMA H, TAKAHASHI Y. Roughly balanced bagging for imbalanced data[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2009, 2(5-6): 412-426.

[15] CHEN C, LIAW A, BREIMAN L. Using random forest to learn imbalanced data[DB/OL]. https://statistics.berkeley.edu/tech-reports/666.

[16] ZHOU Z H. Machine learning[M]. Beijing: TsingHua University Press, 2016.
周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[17] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24: 123-140.

[18] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The elements of statistical learning: data mining, inference, and prediction[J]. Springer Series in Statistics, 2009, 27: 83-85.

[19] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[20] BREUKELEN M, DUIN R, TAX D, et al. Handwritten digit recognition by combined classifiers[J]. Kybernetika, 1998, 34(4): 381-386.

(责任编辑:孙　娟)