文章编号:1672-3961(2024)04-0059-08 DOI:10.6040/j.issn.1672-3961.0.2023.116

面向不平衡数据的提升均衡集成学习算法

白琳[1,2],俱通[1],王浩[1],雷明珠[1],潘晓英[1,2]
(1.西安邮电大学计算机学院,陕西 西安 710121;2.陕西省网络数据分析与智能处理重点实验室,陕西 西安 710121)

摘要:为有效解决欠采样技术在处理不平衡数据时的伪平衡问题,提出并设计一种基于欠采样的提升均衡集成学习算法。采用新的均衡采样机制,通过分箱操作协调数据的预测概率,生成高质量的训练子集,以此迭代训练分类器。基于基分类器在原始数据上的假阳性率和假阴性率,在迭代过程中自适应为其分配权重,避免性能较差的分类器影响整体决策,提高集成模型的泛化能力。新的算法能够在消除伪平衡的同时增加多数类样本的识别度,从而降低边界模糊对分类模型的影响。通过 18 组小型数据集和 2 组大型数据集的对比试验表明,该算法具有处理不平衡数据分类问题的优势。

关键词:欠采样;类不平衡;不平衡学习;集成学习;不平衡数据分类
中图分类号:TP391 **文献标志码**:A
引用格式:白琳,俱通,王浩,等. 面向不平衡数据的提升均衡集成学习算法[J].山东大学学报(工学版),2024,54(4):59-66.
BAI Lin, JU Tong, WAND Hao, et al. Boosted equalization ensemble learning algorithm for imbalanced data[J]. Journal of Shandong University (Engineering Science), 2024, 54(4):59-66.

Boosted equalization ensemble learning algorithm for imbalanced data

BAI Lin[1,2], JU Tong[1], WAND Hao[1], LEI Mingzhu[1], PAN Xiaoying[1,2]
(1.School of Computer Science and Technology, Xi′an University of Posts and Telecommunications, Xi′an 710121, Shaanxi, China; 2. Shaanxi Province Key Laboratory of Network Data Analysis and Intelligent Processing, Xi′an 710121, Shaanxi, China)

Abstract: In order to effectively solve the pseudo-balancing problem of the under-sampling technique in dealing with imbalanced data, a boosted equalization ensemble learning algorithm based on under-sampling was proposed. A new equalization sampling mechanism was used to train the classifier iteratively by coordinating the prediction probabilities of the data through the binning operation, so a high-quality training subset could be generated. Based on the false-positive and false-negative rates of the base classifiers on the original data, weights were assigned adaptively to them during the iterative process, so as to avoid poorly performing classifiers from influencing the overall decision and to improve the generalization ability of the ensemble model. The new algorithm was able to increase the recognition of majority class samples while eliminating pseudo-balancing, thus reducing the impact of boundary ambiguity on the classification model. Comparative experiments with 18 sets of small datasets and 2 sets of large datasets showed that the algorithm had the advantage of dealing with imbalanced data classification problems.
Keywords: under-sampling; class imbalance; imbalance learning; ensemble learning; imbalanced data classification

0 引言

在各学科实际应用中,类别不平衡问题普遍存在,如医疗诊断[1]、欺诈检测[2]、入侵检测[3]及垃圾邮件识别[4]等。不平衡问题表现为分类数据集中某些类的样本数量明显高于其他类,这种数据分布往往会导致分类算法的结果偏向多数类,从而忽略了实际中更有价值的少数类。实际应用中,少数类(即正类)往往更受关注,一般具有较高的错分代价,如网络入侵检测中,将网络攻击误判为正常会导致严重的后果。因此,数据不平衡问题给机器学习的分类任务带来严峻的挑战。本研究基于欠采样策略,提出一种提升均衡集成学习(boosted equalization ensemble learning, BoostEASE)算法有效解决类不平衡问题。

收稿日期:2023-06-01
基金项目:陕西省重点研发计划资助项目(2023-YBSF-476);西安邮电大学创新基金资助项目(CXJJYL2022043)
第一作者简介:白琳(1980—),女,陕西西安人,副教授,硕士生导师,硕士,主要研究方向为智能信息处理。E-mail:bailin@ xupt.edu

1 相关工作

当前，诸多研究工作在处理不平衡数据时采用算法级的方式[5-8]，并未直接对数据进行预处理，而是侧重改良分类器、提高少数类的决策权重，如成本敏感学习。数据级方式常发生在预处理阶段，如过采样[9-11]或欠采样[12-14]，通过增加少数类或减少多数类样本数量达到平衡的效果。混合式方法（如boosting或bagging等）将采样策略与集成学习结合[15-18]，通过集成策略组建多个基分类器，通常能获得更优的性能，但也会消耗更多的成本。实际中，由于过采样通常会进一步增加数据量，很少考虑分类器可承受的容量，而欠采样可以缩短分类器的训练时间，因此被众多研究者关注。

文献[15]提出的自步调集成（self-paced ensemble，SPE）通过对每个样本按照自身所分配的硬度（分类器的预测概率）进行分箱，采用步调函数对每个分箱采样一定数量的多数类样本，得到很好的不平衡数据分类效果；文献[16]的均衡集成（equalization ensemble，EASE）是对SPE的改进，使用少数类样本数量作为SPE分箱数量的超参数，在每个分箱中只采样一个样本，使数据达到平衡，样本分箱更加均匀，减少了欠采样的随机性，同时通过几何均值G_{mean}指标为每个分类器分配权重，减少了每个基分类器因高误报率对算法性能造成的影响。

但是，现有的欠采样算法普遍存在2个重要的问题：一是为了达到平衡数据的目的，删除了过量的负类样本（其中可能包含具有较强判别性的样本），造成伪平衡，导致分类器具有很高的假阳性率F_{P}；二是许多基于空间特征的算法运行成本过大，浪费大量时间且效果不佳。虽然EASE避免了后者，从分类器预测概率的角度出发节省了大量时间，但仍未解决第1个问题。由于G_{mean}指标更偏向正类，且欠采样会保留原始数据中全部正类样本，EASE通过原始数据的得分指标为基分类器分配权重，导致该权重因正类样本并未改动而较为稳定，因此不能在算法中体现每个分类器的性能差异。

为解决以上问题，本研究基于欠采样策略，研究能够处理不同规模的不平衡数据集成学习算法。采用基于预测概率的均衡采样机制，在不影响正类分类效果的同时解决欠采样的信息丢失问题，避免出现伪平衡；利用混淆矩阵获取原始数据的假阳性率F_{P}和假阴性率F_{N}，为每个基分类器分配更准确的权重，以此解决G_{mean}指标分数波动小的缺点。

2 提升均衡集成学习方法

2.1 相似性度量

相似性度量用于描述2个或多个样本间的相似程度。文献[13]指出，越靠近决策边界的样本，误分类的概率越大。决策边界如图1所示，数据中不可避免存在类间重叠的问题，该问题愈严重，决策边界愈模糊，模糊的边界必然导致附近的测试样本被误分类。如果模型可以有效避免边界模糊问题，则该模型的预测概率可以作为样本的相似性度量。

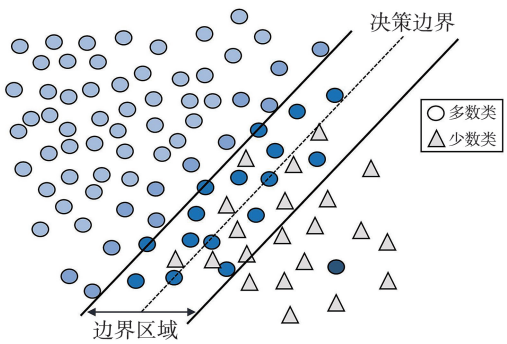


图1 决策边界示意图

Fig.1 Decision boundary diagram

2.2 提升均衡欠采样

为有效解决伪平衡问题，本研究提出BoostEASE算法。采用提升均衡采样策略获取充足的边界信息，参考boosting集成框架进行设计，使每个基分类器获得更高效的权重分配。

BoostEASE算法以预测概率作为相似性度量标准，将预测概率差异小的样本放入同一分箱，箱子的总数k与少数类样本D^+的数量保持一致，使用$p(x,y)$表示多数类样本D^-中样本(x,y)的预测概率，则第i个分箱

$$B_i=\left\{(x,y)\mid\frac{i-1}{k}\leqslant p(x,y)<\frac{i}{k}\right\},(x,y)\in D^-,\quad 1\leqslant i\leqslant k。\tag{1}$$

根据第i个分箱中多数类样本数$\hat{D}_i^-$赋予该分箱预期所需的采样量

$$S_i=\begin{cases}1, & \hat{D}_i^-=0\\ \lceil c\rceil, & \hat{D}_i^-\neq 0\end{cases},\quad 1\leqslant i\leqslant k,\tag{2}$$

式中c为少数类样本总数与非空分箱的比。

在每个分箱中随机选取一定数量的样本加入训练子集中，如果该分箱中样本量小于所分配的采样量，则按式(3)将该分箱样本全部加入训练子集，并将剩余额度分配给下一分箱。

$$S_{i+1}=S_{i+1}+S_i-|B_i|。\tag{3}$$

BoostEASE 算法改进了 EASE 的均衡思想,允许原空箱为后续非空箱提供更多的采样量,这种方式会使采样偏向安全区域的高概率样本,在不影响正类的同时,缓解了负类信息丢失的问题,因此可以解决伪平衡问题。详细步骤如算法 1 所示。

算法 1　提升均衡欠采样

输入:训练集 $D=D^+\cup D^-$,分箱个数 $k=|D^+|$,负类样本预测概率 p。

输出:训练子集。

(1)　初始化训练子集;
(2)　**for** i = 1 to k **do**;
(3)　　用式(1)为每个分箱分配样本;
(4)　　用式(2)为每个分箱分配采样量;
(5)　**end for**;
(6)　**for** i = 1 to k **do**;
(7)　　**if** $|B_i|\geqslant S_i$ **then**;
(8)　　在第 k 个分箱中随机选择S_i个样本加入训练子集中;
(9)　　**end if**;
(10)　　**if** $|B_i|<S_i$ **then**;
(11)　　将 B_i 中的样本全部加入训练子集;
(12)　　用式(3)将剩余采样量分配给下一个非空箱;
(13)　　**end if**;
(14)　**end for**;
(15)　将 D^+加入训练子集;
(16)　**return** 训练子集。

2.3　BoostEASE 算法框架

BoostEASE 算法框架如图 2 所示,其本质是一种模拟 boosting 框架的算法,在每次迭代中使用基于预测概率的均衡欠采样得到的训练子集训练分类器,每次迭代通过计算基分类器在原始训练数据上的混淆矩阵,得到假阳性样本数和假阴性样本数,为该分类器分配权重 W_t^h,计算式为:

$$W_t^h=\left(1-\frac{F_{\mathrm{P}i}}{\widehat{F_{\mathrm{P}}}}\right)\left(1-\frac{F_{\mathrm{N}i}}{\widehat{F_{\mathrm{N}}}}\right),\ 1\leqslant t\leqslant T\ ,\tag{4}$$

式中,T 为总迭代次数,$F_{\mathrm{P}i}$和 $F_{\mathrm{N}i}$分别为第 i 个基分类器 h_i 在原始数据上的假阳性率和假阴性率,$\widehat{F_{\mathrm{P}}}$为当前迭代所有基分类器在原始数据上的最大假阳性率,$\widehat{F_{\mathrm{N}}}$为最大假阴性率。

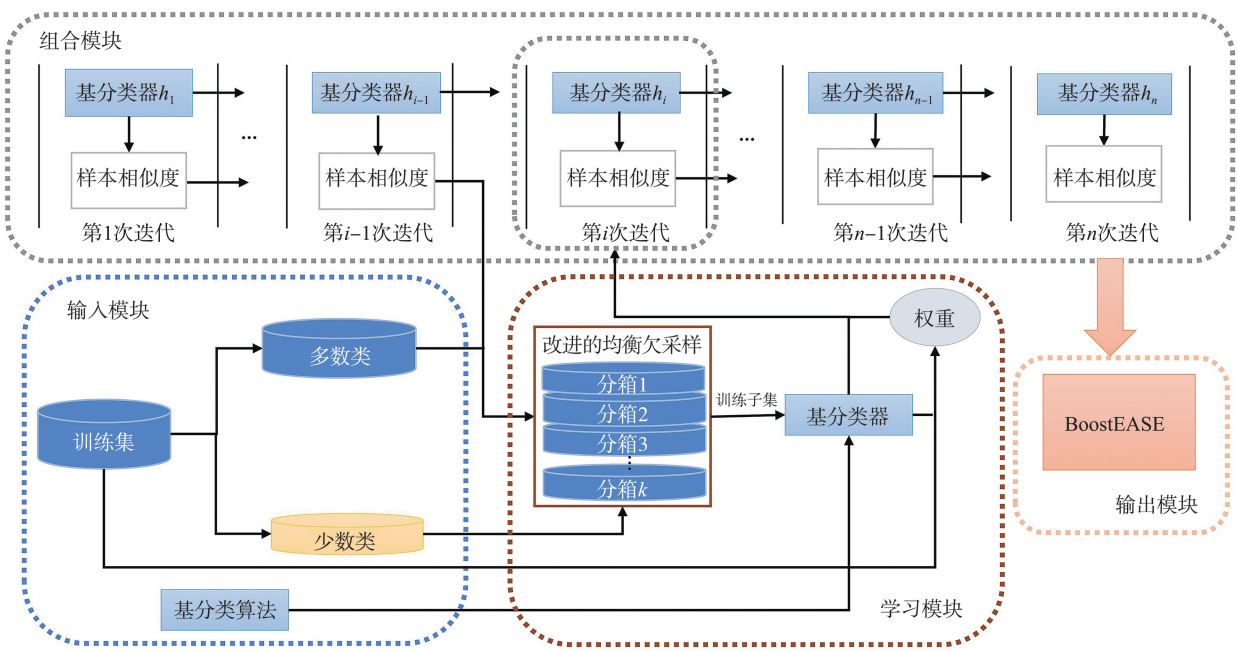


图 2　BoostEASE 算法框架

Fig.2　Diagram of BoostEASE framework

基分类器 h_i 由原始数据得到的训练子集训练而成,因此,每个基分类器的权重 W_t^h 与其在原始数据上所表现的性能成正比,与 F_{P} 和 F_{N} 成反比,$F_{\mathrm{P}i}$和 $F_{\mathrm{N}i}$越小,则$(1-(F_{\mathrm{P}i})/\widehat{F_{\mathrm{P}}})$和$(1-(F_{\mathrm{N}i})/\widehat{F_{\mathrm{N}}})$越大,它们之间的乘积也会越大,说明该基分类器 h_i 的权重更大,其性能更佳。这种权重计算方式还可以筛除对各类分类效果最差的分类器,且不会偏向任何一类。

最终,将所有基分类器平均加权,得到预测概率

$$P=\frac{1}{T}\sum_{t=1}^{T}W_t^h h_t(x,y)。\tag{5}$$

3 试验结果及分析

3.1 数据集介绍

试验数据集来自KEEL[19]及KDD Cup[16]中20个不平衡数据集,详细信息如表1所示,最后2个数据集为大型数据集。除了不平衡比率等基本信息,还基于闵可夫斯基度量统计了安全率等信息[20],具体使用5-KNN模型。将数据集进行随机划分,80%作为训练集,20%作为测试集。

表1 数据集详情

Table 1 Details of datasets

数据集	不平衡比率	样本数	负类样本数	正类样本数	特征数	安全率/%	边界率/%	异常率/%
pima	1.87	768	500	268	8	23.13	69.40	7.46
german	2.33	1 000	700	300	20	7.33	77.00	15.67
glass0123vs456	3.20	214	163	51	9	64.71	25.49	9.80
new-thyroid2	5.14	215	180	35	5	91.43	8.57	0
yeast3	8.10	1 484	1 321	163	8	51.53	38.65	9.82
yeast-2_vs_4	9.08	514	463	51	8	58.82	27.45	13.73
glass4	15.46	214	201	13	9	46.15	38.46	15.38
ecoli4	15.80	336	316	20	7	75.00	20.00	5.00
shuttle-c2-vs-c4	20.50	129	123	6	9	16.67	66.67	16.67
glass5	22.78	214	205	9	9	44.44	33.33	22.22
winequality-white-9_vs_4	32.60	168	163	5	11	0	60.00	40.00
yeast5	32.73	1 484	1 440	44	8	36.36	61.36	2.27
ecoli-0-1-3-7_vs_2-6	39.14	281	274	7	7	71.43	0	28.57
page-blocks1v5	42.72	5 028	4 913	115	10	24.35	55.65	20.00
winequality-white-3_vs_7	44.00	900	880	20	11	5.00	15.00	80.00
yeast8	73.20	1 484	1 464	20	8	0	35.00	65.00
poker-8-9_vs_5	82.00	2 075	2 050	25	10	0	36.00	64.00
poker-8_vs_6	85.88	1 477	1 460	17	10	0	29.41	70.59
creditcard	577.87	284 806	284 314	492	29	64.84	16.67	18.50
dos_vs_r2l	3 448.82	3 884 495	3 883 369	1 126	41	94.67	3.82	1.51

3.2 对比方法和参数设置

选择6个代表性对比方法:Adaptive-threshold OBU(AdaOBU)[12]、Boosted OBU(BoostOBU)[12]、SPE[15]、EASE[16]、Hashing-based Under-sampling Ensemble(HUE)[17]、SMOTEBagging(SBag)[18]。

对于集成算法,基分类器使用决策树,其数量参照文献[15]选择性能较为稳定的20,其他参数均为默认值。所有试验均进行10次重复的五折交叉验证,以减少随机误差。

3.3 评价指标

采用3种不平衡数据评价指标:F_1值、G_{mean}值、接收者操作特性曲线(receiver operating characteristic curve, ROC)下面积A_{UC}。F_1值是分类器的精确度P_{recision}和召回率R_{ecall}之和的调和平均值,可以很好地平衡正负类的预测效果:

$$F_1=\frac{2P_{\text{recision}}R_{\text{ecall}}}{P_{\text{recision}}+R_{\text{ecall}}}, \tag{6}$$

式中,P_{recision}和R_{ecall}在不受类倾斜分布影响的情况下衡量了分类器对正实例的预测准确率,$P_{\text{recision}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}}$,$R_{\text{ecall}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{N}}}$,其中$T_{\text{P}}$为正确预测的正类样本数,$F_{\text{P}}$为错误预测的正类样本数,$F_{\text{N}}$为错误预测的负类样本数。$G_{\text{mean}}$是$R_{\text{ecall}}$和特异度$S_{\text{pecificity}}$的几何平均,综合了识别正负实例的能力:

$$G_{\text{mean}}=\sqrt{R_{\text{ecall}}S_{\text{pecificity}}}, \tag{7}$$

式中,$S_{\text{pecificity}}=\frac{T_{\text{N}}}{T_{\text{N}}+F_{\text{P}}}$,其中$T_{\text{N}}$为正确预测的负类样本数。

ROC曲线展示了分类器在不同阈值下T_{P}和F_{P}之间的关系,通过计算ROC曲线下的面积得到A_{UC}指标。

对3种指标下的试验结果进行统计检验,常用方式有Friedman检验和Nemenyi检验[21]。

3.4 算法对比

3.4.1 小型数据集对比

表2~4分别展示了18个数据集上7种算法的

F_1 值、G_{mean} 和 A_{UC},由于篇幅有限且本研究算法在 G_{mean} 指标上的表现不具有优势,故表 3 中只列出 G_{mean} 指标的平均秩序和平均分数。根据表 2~4 可知:集成学习模型效果优于单一分类器,但 HUE 删除过多有用信息,导致 F_1 值指标不佳;BoostEASE 在 F_1 值和 A_{UC} 指标上平均得分最高,而在 G_{mean} 指标上略低于 SPE 和 EASE,且明显高于 AdaOBU、BoostOBU 和 SBag,这是由于 BoostEASE 提高了多数类样本采样量,减少了信息丢失。虽然这些增量专注于高概率样本区域,但无法避免会采集到边界区域的样本,从而影响到少数类的分类精度,但这种影响极微,在可接受范围内。虽然 HUE 的 G_{mean} 指标平均分数最高,但忽视了多数类的重要性,过度专注于少数类是不可取的。

表 2　小型数据集 F_1 值对比

Table 2　Comparison of F_1 values for small datasets

数据集	F_1 值						
	AdaOBU	BoostOBU	SBag	HUE	SPE	EASE	BoostEASE
ecoli-0-1-3-7_vs_2-6	32.24(4)	45.20(2)	**61.33(1)**	16.91(7)	32.09(5)	22.04(6)	42.88(3)
ecoli4	41.81(7)	58.25(6)	**80.21(1)**	63.30(5)	77.19(4)	77.96(3)	78.69(2)
german	47.97(6)	46.18(7)	51.83(5)	**61.72(1)**	58.67(2)	57.28(4)	58.55(3)
glass0123vs456	77.07(7)	80.29(6)	87.55(2)	**87.65(1)**	84.03(5)	86.60(4)	86.94(3)
glass4	53.06(7)	59.86(6)	64.28(4)	60.91(5)	77.35(2)	72.13(3)	**77.85(1)**
glass5	44.72(6)	78.93(2)	69.01(4)	42.11(7)	77.82(3)	**79.35(1)**	66.81(5)
new-thyroid2	71.10(7)	83.04(6)	92.38(4)	88.02(5)	**93.68(1)**	92.75(2)	92.74(3)
page-blocks1v5	35.56(7)	38.89(5)	66.53(4)	38.70(6)	65.94(2)	66.89(3)	**69.39(1)**
pima	55.06(7)	55.44(6)	63.18(5)	**68.20(1)**	64.99(2)	64.75(3)	63.95(4)
poker-8-9_vs_5	5.36(5)	6.88(3)	0.67(7)	4.48(6)	7.47(2)	6.64(4)	**14.20(1)**
poker-8_vs_6	3.29(7)	10.43(3)	14.60(2)	3.50(6)	7.47(4)	5.45(5)	**24.01(1)**
shuttle-c2-vs-c4	85.70(7)	89.29(5)	86.00(6)	**99.33(1)**	92.00(3)	90.00(4)	98.00(2)
winequality-white-3_vs_7	7.55(6)	16.29(4)	0(7)	14.07(5)	20.45(2)	20.25(3)	**45.03(1)**
winequality-white-9_vs_4	11.00(7)	29.81(4)	37.00(3)	16.46(6)	**41.38(1)**	17.59(5)	40.24(2)
yeast-2_vs_4	56.24(7)	67.98(6)	74.31(3)	72.16(5)	74.46(2)	72.68(4)	**76.59(1)**
yeast3	55.15(7)	55.72(6)	**77.33(1)**	76.28(2)	76.06(3)	74.04(5)	74.56(4)
yeast5	53.46(6)	66.03(5)	70.17(4)	51.87(7)	74.44(3)	74.92(2)	**76.52(1)**
yeast8	9.90(5)	10.35(4)	**58.47(1)**	7.10(7)	11.21(3)	9.00(6)	23.24(2)
平均秩序	6.39	4.78	3.56	4.61	2.72	3.72	**2.22**
平均分数	41.46	49.94	58.60	48.49	57.59	55.02	**61.68**

注:括号中数值表示该算法在该数据集下的得分秩序。

表 3　小型数据集 G_{mean} 指标的平均结果对比

Table 3　Comparison of the average results of G_{mean} for small datasets

平均对比	G_{mean}						
	AdaOBU	BoostOBU	SBag	HUE	SPE	EASE	BoostEASE
平均秩序	6.28	5.56	5.44	**1.61**	2.67	3.11	3.11
平均分数	69.16	70.98	65.73	**83.58**	82.31	81.76	80.59

表 4　小型数据集 A_{UC} 对比

Table 4　Comparison of A_{UC} for small datasets

数据集	A_{UC}						
	AdaOBU	BoostOBU	SBag	HUE	SPE	EASE	BoostEASE
ecoli-0-1-3-7_vs_2-6	79.52(7)	83.05(6)	88.96(5)	**96.46(1)**	92.82(4)	94.86(3)	95.29(2)
ecoli4	78.64(7)	84.27(6)	96.59(5)	**99.00(1)**	98.00(4)	98.57(2)	98.30(3)
german	58.38(6)	55.70(7)	76.72(4)	**77.59(1)**	77.45(3)	76.60(5)	77.47(2)
glass0123vs456	87.98(7)	89.07(6)	97.38(2)	95.95(5)	96.80(3)	96.54(4)	**97.67(1)**
glass4	80.62(6)	80.20(7)	94.24(5)	97.73(4)	98.23(3)	98.27(2)	**98.63(1)**
glass5	74.10(7)	90.87(6)	99.17(4)	**99.67(1)**	99.18(3)	99.51(2)	98.83(5)
new-thyroid2	88.50(7)	93.78(6)	99.36(4)	99.38(3)	**99.72(1)**	99.71(2)	99.21(5)
page-blocks1v5	73.31(7)	75.26(6)	97.62(5)	**98.87(1)**	**98.87(1)**	98.78(3)	98.31(4)
pima	60.80(7)	61.16(6)	**81.08(1)**	79.23(5)	80.65(2)	79.70(3)	79.34(4)

表4(续)

数据集	A_{UC}						
	AdaOBU	BoostOBU	SBag	HUE	SPE	EASE	BoostEASE
poker-8-9_vs_5	61.05(6)	59.57(7)	76.54(4)	70.63(5)	80.08(2)	79.45(3)	**84.31(1)**
poker-8_vs_6	58.73(7)	62.72(6)	**97.76(1)**	72.74(5)	81.72(4)	82.95(3)	86.50(2)
shuttle-c2-vs-c4	94.64(7)	95.92(6)	**100.00(1)**	**100.00(1)**	**100.00(1)**	**100.00(1)**	**100.00(1)**
winequality-white-3_vs_7	62.63(6)	61.89(7)	84.84(5)	86.06(4)	86.84(2)	**87.00(1)**	86.34(3)
winequality-white-9_vs_4	61.12(7)	72.21(6)	89.73(4)	92.64(2)	**95.03(1)**	90.30(3)	89.30(5)
yeast-2_vs_4	84.79(7)	87.67(6)	98.08(2)	97.76(3)	97.71(4)	97.67(5)	**98.18(1)**
yeast3	83.26(7)	83.54(6)	96.47(4)	**96.67(1)**	96.50(2)	95.97(5)	96.48(3)
yeast5	88.04(7)	89.84(6)	97.13(5)	98.89(2)	98.58(4)	98.68(3)	**99.15(1)**
yeast8	71.92(6)	68.22(7)	79.22(4)	**84.17(1)**	81.00(2)	80.17(3)	78.60(5)
平均秩序	6.72	6.28	3.61	**2.56**	**2.56**	2.94	2.78
平均分数	74.89	77.50	91.72	91.30	92.18	91.93	**92.33**

注:括号中数值表示该算法在该数据集下的得分秩序。

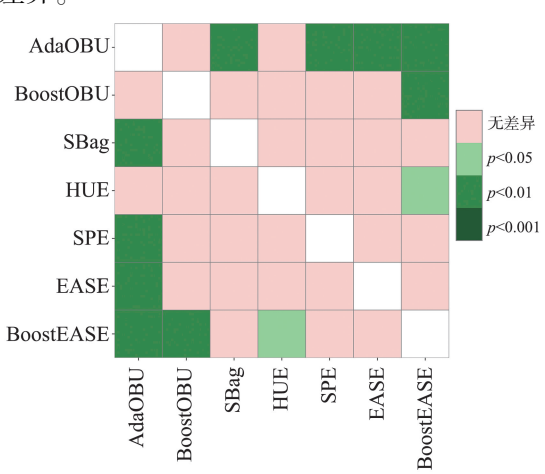
在 Friedman 统计检验中,得到 Friedman 检验统计变量 $F_{F1}=11.97$、$F_{Gmean}=32.11$、$F_{AUC}=32.26$,当 $\alpha=0.05$ 时,本研究中算法个数为 7,数据集个数为 18,由 Friedman 检验临界值计算方法[22]可知 $F_F=2.189$,由于 F_{F1}、F_{Gmeans}和 F_{AUC}均大于 F_F,在各个评价指标上拒绝了检验假设,且各算法之间性能存在差异。

后续使用 Nemenyi 检验进一步区分算法的性能,差异临界值

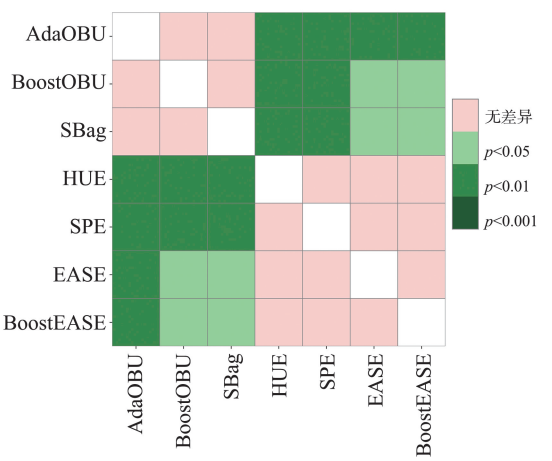
$$C_D=q_\alpha\sqrt{\frac{K(K+1)}{6D}}, \tag{8}$$

式中,q_α 为 Tukey 分布的临界值,K 和 D 分别为算法数及数据集数。

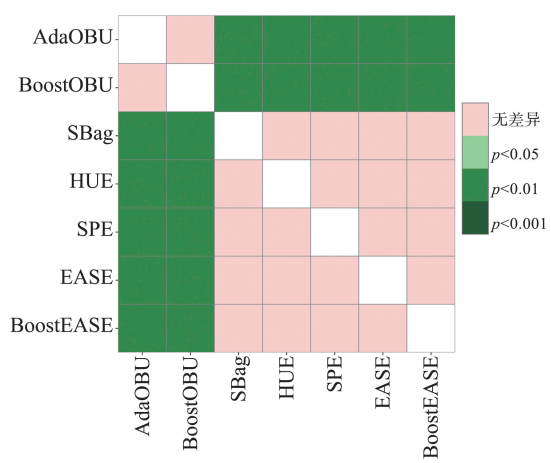
设 $\alpha=0.05$,则 Nemenyi 检验中 $q_\alpha=2.949$[22],$K=7$,$D=18$,计算得出 $C_D=2.12$,若任意 2 种算法的平均阶数之差大于 C_D,则表明 2 种算法的性能差异显著。Nemenyi 后续检验结果如图 3 所示。由图 3 可以看出:BoostEASE 与 AdaOBU、BoostOBU 有明显差异,和 SPE 及 EASE 在小型数据集上并无显著差异。



(a) F_1值指标Nemenyi后续检验



(b) G_{mean}值指标Nemenyi后续检验



(c) A_{UC}值指标Nemenyi后续检验

图 3 3 种指标下各算法差异热力图

Fig.3 Heat maps of each algorithm under three indicators

3.4.2 大型数据集对比

由于一些算法不适用于大型数据集,本节比较了 BoostEASE 和 EASE、HUE 及 SPE 算法在大型数据集上的性能。因为 HUE 算法运行时间过久,本研究直接引用文献[16]中的测试结果(试验所使用的数据预处理方式与本研究相同,试验环境、HUE 算

法参数一致),结果如表 5 所示。由表 5 可知:BoostEASE 最佳得分次数最多,在 creditcard 数据集上 F_1 值指标明显高于其他对比算法,这得益于 BoostEASE 保留了多数类重要信息,并且在偏向正类的 G_{mean} 指标上与 EASE 方法持平,HUE 性能过低,原因在于 HUE 删除了太多有用信息;dos_vs_r2l 数据集具有高维、异构且边界模糊的特点,BoostEASE 能够有效处理该数据集。

表 5　大型数据集对比

Table 5　Comparison of large datasets

数据集	HUE			SPE		
	F_1 值	G_{mean}	A_{UC}	F_1 值	G_{mean}	A_{UC}
creditcard	7.27±0.40	—	**97.99±0.69**	48.61±1.24	53.59±0.81	91.72±0.85
dos_vs_r2l	32.29±0.69	—	99.99±0.01	99.51±0.21	99.64±0.20	99.99±0.01
数据集	EASE			BoostEASE		
	F_1 值	G_{mean}	A_{UC}	F_1 值	G_{mean}	A_{UC}
creditcard	52.79±1.44	**92.00±0.98**	96.69±0.66	**79.99±1.05**	90.14±0.60	96.20±0.50
dos_vs_r2l	99.60±0.08	99.71±0.08	99.98±0.02	99.57±0.09	**99.72±0.09**	**99.99±0.01**

注:“—”表示该试验结果由于客观因素无法获取。

3.5　时间性能对比

本研究只考虑训练时间,对比结果如表 6 所示,和 3.4.2 节试验设置一样,HUE 算法运行时间过久,直接引用文献[16]中的结果作为参考。由表 6 可知:SPE 耗费时间最短,EASE 和所提算法 BoostEASE 消耗时间相近,且略慢于 SPE,在可承受范围内;HUE 算法耗费了大量时间且不可接受。

表 6　运行时间对比

Table 6　Comparison ofrunning times

数据集	运行时间/s			
	HUE	SPE	EASE	BoostEASE
creditcard	69.1	0.4	0.8	0.9
dos_vs_r2l	7 003.8	5.1	12.0	12.1

3.6　基分类器权重评估

本节评估了 BoostEASE 框架的权重分配策略,在 creditcard 数据集上分别设计以下试验:试验 1 使用 G_{mean} 指标分数作为权重,试验 2 使用基于混淆矩阵(具体使用 F_P 和 F_N)的权重策略。测试结果如表 7 所示。由表 7 可知,采用试验 2 的权重分配方式更有利于评价基分类器的好坏,在 F_1 值和 A_{UC} 指标上优于试验 1 的方式,而在 G_{mean} 指标上基本与之持平。

表 7　基分类器权重评估

Table 7　Weighted evaluation of base classifier

试验	F_1 值	G_{mean}	A_{UC}
1	73.36±1.28	**90.34±0.77**	95.65±0.57
2	**79.99±1.05**	90.14±0.60	**96.20±0.50**

4　结论

为有效处理不平衡数据,本研究提出基于预测概率的均衡集成欠采样框架 BoostEASE。采用全新的均衡欠采样策略,得到若干高质量训练子集,以此训练出若干优秀的弱分类器,并为每个基分类器自适应地分配更高效的权重。BoostEASE 解决了传统欠采样方法的伪平衡及信息丢失问题,能够有效解决不平衡数据问题,且对不同规模的数据集均有效,具有良好的可伸缩性和可拓展性;同时,集成学习算法普遍存在模型复杂、效率较低的问题,BoostEASE 虽然增加了采样量,但运行时间并未明显增加。在试验中,本研究只使用决策树作为基分类器,在多种不平衡数据集中性能提升理想(如 F_1 值、A_{UC} 等),但 BoostEASE 并不只适用于决策树,更换分类器、改进集成模型将作为进一步的研究工作。

参考文献:

[1] LIU N, LI X, QI E, et al. A novel ensemble learning paradigm for medical diagnosis with imbalanced data[J]. IEEE Access, 2020, 8: 171263-171280.

[2] LI Z, HUANG M, LIU G, et al. A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection[J]. Expert Systems with Applications, 2021, 175: 114750.

[3] DING H, CHEN L, DONG L, et al. Imbalanced data classification: a KNN and generative adversarial networks-based hybrid approach for intrusion detection [J]. Future Generation Computer Systems, 2022, 131: 240-254.

[4] LIU S, WANG Y, ZHANG J, et al. Addressing the class imbalance problem in twitter spam detection using ensemble learning[J]. Computers & Security, 2017, 69: 35-49.

[5] PASSOS L A, JODAS D S, RIBEIRO L C, et al. Handling imbalanced datasets through optimum-path forest [J]. Knowledge-Based Systems, 2022, 242: 108445.

[6] TAO X, LI Q, GUO W, et al. Self-adaptive cost

weights-based support vector machine cost-sensitive ensemble for imbalanced data classification[J]. Information Sciences, 2019, 487: 31-56.
[7] KRAWCZYK B, WOZNIAK M, SCHAEFER G. Cost sensitive decision tree ensembles for effective imbalanced classification[J]. Applied Soft Computing, 2014, 14: 554-562.
[8] 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, 2017, 29(8): 3573-3587.
[9] CHAWLA N, BOWYER K, HALL L, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.
[10] SOLTANZADEH P, HASHEMZAEH M. RCSMOTE: range-controlled synthetic minority over-sampling technique for handling the class imbalance problem[J]. Information Sciences, 2021, 542(4): 92-111.
[11] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE [J]. Information Sciences, 2019, 501: 118-135.
[12] VUTTIPITTAYAMONGKOL P, ELYAN E. Improved overlap-basedundersampling for imbalanced dataset classification with application to epilepsy and Parkinson′s disease[J]. International Journal of Neural Systems, 2020, 30(9): 2050043.
[13] VUTTIPITTAYAMONGKOL P, ELYAN E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data[J]. Information Sciences, 2020, 509: 47-70.
[14] ELHASSAN T, ALJURF M. Classification of imbalance data using Tomek link (T-link) combined with random under-sampling (RUS) as a data reduction method[J]. Global Journal of Technology & Optimization, 2016, 1: 2-11.
[15] LIU Z, CAO W, GAO Z, et al. Self-paced ensemble for highly imbalanced massive data classification[C]//2020 IEEE 36th International Conference on Data Engineering (ICDE). Dallas, USA: IEEE, 2020: 841-852.
[16] REN J, WANG Y, MAO M, et al. Equalization ensemble for large scale highly imbalanced data classification[J]. Knowledge-Based Systems, 2022, 242: 108295.
[17] NG W W Y, XU S, ZHANG J, et al. Hashing-based undersampling ensemble for imbalanced pattern classification problems[J]. IEEE Transactions on Cybernetics, 2020, 52(2): 1269-1279.
[18] WANG S, XIN Y. Diversity analysis on imbalanced data sets by using ensemble models[C]// 2009 IEEE Symposium on Computational Intelligence and Data Mining. Nashville, USA: IEEE, 2009: 324-331.
[19] DERRAC J, GARCIA S, SANCHEZ L, et al. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework[J]. Journal of Multiple-Valued Logic and Soft Computing, 2015, 17(2/3): 255-287.
[20] KOZIARSKI M. Radial-basedunder sampling for imbalanced data classification[J]. Pattern Recognition, 2020, 102: 107262.
[21] DEMSAR J. Statistical comparisons of classifiers over multiple data sets[J]. The Journal of Machine Learning Research, 2006, 7: 1-30.
[22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2021.

(编辑:孙亚彤)

(上接第 58 页)

[20] CHEN C L, 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.
[21] PRAMANIK S, BHATTACHARJEE D, NASIPURI M, et al. LINPE-BL: alocal descriptor and broad learning for identification of abnormal breast thermograms[J]. IEEE Transactions on Medical Imaging, 2021, 40(12): 3919-3931.
[22] JIN J, LI Y, YANG T, et al. Discriminative group-sparsity constrained broad learning system for visual recognition[J]. Information Sciences, 2021, 576: 800-818.
[23] WANG X, CHCENG L, ZHANG D, et al. Broad learning solution for rapid diagnosis of COVID-19[J]. Biomedical Signal Processing and Control, 2023, 83: 104724.
[24] DU J, VONG C, CHEN C L. Novel efficient RNN and LSTM-like architectures: recurrent and gated broad learning systems and their applications for text classification[J]. IEEE Transactions on Cybernetics, 2021, 51(3): 1586-1597.
[25] HOCHREITER S, JÜRGEN S. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[26] CHEN G, PENG S, ZENG R, et al. p-Norm broad learning for negative emotion classification in social networks[J]. Big Data Mining and Analytics, 2022, 5(3): 245-256.

(编辑:孙亚彤)