融合簇边界移动与自适应合成的混合采样算法

高雷阜[1],张梦瑶[2],赵世杰[1]
(1. 辽宁工程技术大学运筹与优化研究院,辽宁阜新 123000; 2. 辽宁工程技术大学优化与决策研究所,辽宁阜新 123000)

摘 要: 针对伪负采样算法(Pseudo-Negative Sampling,PNS)存在的类内子聚集和类别重叠问题,提出一种融合簇边界负样本移动策略(Cluster Boundary Negative Movement Strategy,CBNMS)与自适应正样本合成技术(Adaptive Positive Synthesis Technology,ADPST)的改进混合采样算法(Improved Cluster Boundary Negative Movement Strategy,ICBNMS),以提升非均衡数据的整体分类性能和正类识别精度. CBNMS策略采用凝聚层次聚类对正负类样本进行划分,并通过各局部样本间相似关系识别潜在负类中且与正类相关性较大的簇边界负样本,提高采样的局部精确性和时效性. 为进一步加强CBNMS策略对正样本重叠区域的识别性能,ICBNMS算法在簇边界负样本移动均衡化基础上,引入ADPST技术,利用稀疏度与距离复合因子组合加权以自适应确定最优样本生成区域,从而有效削弱样本的重叠性且丰富样本的多样性. 实验结果表明,相比其他采样算法,ICBNMS算法在10个非均衡数据集的多组实验中G-mean和F-measure等指标获得最优值,且时间效率比CDSMOTE和PNS算法分别提升了32.27%和27.88%,凸显出更优越的鲁棒性和泛化性.

关键词: 非均衡数据分类; 凝聚层次聚类; 簇边界负样本移动; 自适应正样本合成; 混合采样
中图分类号: TP181;TP39 **文献标识码:** A **文章编号:** 0372-2112(2022)10-2517-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210265

Mixed-Sampling Algorithm Combining Cluster Boundary Movement and Adaptive Synthesis

GAO Lei-fu[1],ZHANG Meng-yao[2],ZHAO Shi-jie[1]
(1. *Institute for Optimization and Decision Analytics*, *Liaoning Technical University*, *Fuxin*, *Liaoning* 123000, *China*;
2. *Institute of Optimization and Decision*, *Liaoning Technical University*, *Fuxin*, *Liaoning* 123000, *China*)

Abstract: For the problem of intra-class sub-gathering and class-overlapping in pseudo-negative sampling(PNS) algorithm, an improved mixed-sampling algorithm combining cluster boundary negative movement strategy(CBNMS) and adaptive positive synthesis technology(ADPST) is proposed to boost the overall classification performance and positive class identification accuracy of imbalanced data. The CBNMS strategy adopts AGENS(Agglomerative Hierarchical Cluster) to divide positive and negative samples, identifies the cluster boundary negative samples in the potential negative class with a large correlation with the positive class by the similar relationship between each local sample, and increases the local accuracy and timeliness of sampling. In order to further strengthen the identification performance of the CBNMS strategy for the overlap area of positive samples, the ICBNMS(Improved Cluster Boundary Negative Movement Strategy) algorithm introduces ADPST technology on the basis of moving equalization of negative samples at the cluster boundary and utilizes the combination of sparsity and distance composite factor weighting to adaptively determine the optimal sample generation area, thereby effectively weakening the overlap of samples and enriching the diversity of samples. Experiment results show that compared with other sampling algorithms, the ICBNMS algorithm can obtain the optimal values of G-mean, F-measure and other indicators in multiple experiments of 10 imbalanced data sets, and its time efficiency has improved by 32.27% and 27.88% respectively compared with the CDSMOTE and PNS algorithms, highlighting more superior robustness and generalization.

Key words: imbalanced data classification; agglomerative hierarchical cluster; cluster boundary negative sample movement; adaptive positive sample synthesis; mixed-sampling

收稿日期:2021-02-22;修回日期:2021-10-21;责任编辑:孙瑶
基金项目:辽宁省教育厅重点攻关项目(No.LJ2019ZL001)

1 引言

大数据时代,分类算法在各种应用中广受青睐,其优越性能依赖理想均衡数据,即各类样本数量大致相等或是分布无显著差异.而现实领域中所得到的往往是非均衡数据,即一个或多个类别与其他类别相比是非常稀少的,通常将其分别定义为正类(少数类)和负类(多数类).此时若仍采用标准分类算法训练模型,如支持向量机(Support Vector Machine,SVM)[1]、随机森林(Random Forest,RF)[2]及K近邻(K-Nearest Neighbor,KNN)[3]等,模型通常偏向负类,从而导致正类识别精度下降.非均衡数据分类问题普遍存在于各领域中,例如医疗疾病诊断[4]、金融欺诈检测[5]及网络入侵检测[6]等,数据本质上是严重失衡且正类样本通常携带更重要的信息.因此,提高分类模型对正类的识别能力已成为非均衡数据分类问题的研究重点.

针对非均衡数据分类问题的解决方案主要包括两个层面:算法层面匹配非均衡数据和数据层面适应算法.算法层面是在模型学习过程中修改算法机理或微调代价权重,然而由于难以确定不同类别的代价权重,算法修改难度增加.而数据层面则是在模型学习之前通过平衡非均衡数据集来提升分类器的性能,如欠采样、过采样以及混合采样.

随机欠采样(Random Under-Sampling,RUS)[7]、邻域清洗规则(Neighborhood Cleaning Rule,NCL)[8]是最经典、最简单的欠采样方法,然而因其受限于随机性,价值较高的样本易丢失.因此,学者们借鉴不同技术提出一系列改进算法,如基于聚类与近邻思想.Tsai等人[9]和Vattipittayamongkol等人[10]先后提出融合聚类和实例选择的CBIS(Cluster-Based Instance Selection)算法和近邻变体的欠采样(NB-Based Under-Sampling),较好地克服了随机性且有效增强了分类效果.鉴于典型的过采样方法,如随机过采样(Random Over-Sampling,ROS)[7]、合成正类过采样(Synthetic Minority Oversampling Technique,SMOTE)[11]等方法局限于过拟合问题且易合成噪声和冗余样本,一些改进算法相继被提出.如Douzas等人[12]针对相似样本的重叠、过度泛化等问题而提出G-SMOTE算法,提高分类器的分类性能.而混合采样方法中,主要采取的是一种将欠采样与过采样相结合的策略,如基于距离提出的经典混合采样SMOTE+Tomek links[13],有效克服单一采样方法的弊端,然而无法解决由类别重叠引起合成错误样本的问题.因而,基于样本分布特征,Elyan等人[14]提出CDSMOTE混合采样算法,以类分解欠采样与类合成过采样来平衡数据集,但该方法对正类过采样仍会导致生成的样本易重叠且缺乏多样性.为保留具有代表性的原始样本,张永清等人[15]提出样本空间型伪负采样算法(Pseudo-Negative Sampling,PNS),着重挖掘潜藏在负类中的伪负样本,以平衡样本非均衡分布,但是该算法无法准确描述存在类内子聚集及类别重叠分布特征的非均衡数据,从而削弱采样精准率并降低分类器的性能.鉴于凝聚层次聚类(Agglomerative Hierarchical Cluster,AGENS)[16]具有刻画样本局部分布特征的能力并能有效解决类内子聚集及类别重叠等问题,混合采样的难点在于如何根据样本原始特征并充分利用聚类优势以达到自适应最佳采样的效果.

为解决上述采样方法无法较好地克服类内子聚集及类别重叠问题,本文提出一种融合簇边界负样本移动策略(Cluster Boundary Negative Movement Strategy,CBNMS)与自适应正样本合成技术(Adaptive Positive Synthesis Technology,ADPST)的改进混合采样算法(Improved Cluster Boundary Negative Movement Strategy,ICBNMS).考虑类内子聚集对采样精准率的影响,本文提出CBNMS策略,根据凝聚层次聚类对正负类样本的划分与样本间相似关系的度量来识别簇边界负样本,以提高采样精确性并达到均衡数据集的目的.为减轻类别重叠对CBNMS策略的负面影响而增强正类识别精度,ICBNMS算法则在此基础上引入ADPST技术,依据复合因子组合加权自适应生成最优合成样本,从而减弱样本重叠性并增加样本多样性.数值实验验证了ICBNMS算法及其包含的两个技术的分类性能与时间效率相较于对比算法明显提高,且在解决类内子聚集与类别重叠问题时具有出色的表现与良好的泛化能力.

2 理论框架

2.1 PNS算法理论框架

PNS算法[15]首次提出"伪负样本"(PN)概念,其具体指虽被划分到负类中,但与正类样本相关性最大且冗余度最小的样本.定义所有样本构成集合为$D=\{(x_i,y_i)|i\in[1,n],y_i=0或1\}$,其正类样本集合为$P=\{(x_i^+,y_i^+)(x_i^+,y_i^+)\in D,i\in[1,m],y_i^+=1\}$,负类样本集合为$N=\{(x_i^-,y_i^-)(x_i^-,y_i^-)\in D,i\in[1,n],y_i^-=0\}$,$m$为正类样本总数,$n$为负类样本总数;伪负样本集合$S^*=\{(x_i^*,y_i^*)(x_i^*,y_i^*)\in N,i\in[1,l]\}$,$l$是伪负样本总数.该算法基于$P$和$N$来确定伪负样本集$S^*$,将其添加到正类并从负类中删除,其优势在于整个混合采样过程以挖掘潜在负样本来调整整体样本空间而非改变样本数量的方式成功克服欠采样与过采样的缺陷,且无需确定采样比例即可完成自适应采样,从而得到均衡数据集.然而,PNS算法处理类内子聚集及类别重叠数据集的能力明显有限.PNS算法主要包括3个步骤:

Step1 计算正样本的平均值作为空间中心C;

Step2 计算正样本与空间中心C的欧氏距离平均

值并判断是否满足伪负样本的相似性评价阈值;

Step3　计算每个负样本与空间中心 C 的距离,判断其是否小于阈值,若满足,则将该负样本标记为 PN 样本并添加至正样本中.

2.2　AGENS 算法理论框架

AGENS[16]是一种自底向上聚合的层次聚类算法,初始状态为每个数据点各属一类,通过计算两类数据点间的距离来衡量二者相似性大小,其中距离越小,表示相似度越高. 而后将距离最近的两个数据点或类别进行合并,反复迭代,生成聚类树,直至满足终止合并条件,得到聚类结果簇. AGENS 算法流程如图 1 所示.

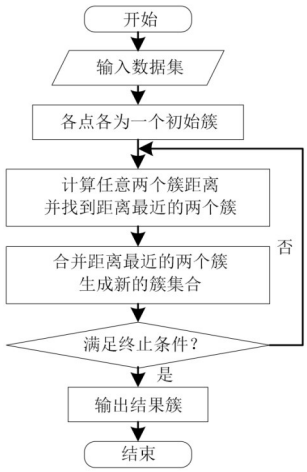

图 1　AGENS 算法流程图

鉴于凝聚层次聚类算法具有逐点划分聚簇、逐簇筛选合并的固有优势,文献[17]自适应采样类内子聚集并分配权重以避免样本重叠及过度泛化等问题,但簇内样本局部特征缺乏充分考虑将会对采样产生一定程度的不良影响.

3　改进混合采样算法

PNS 算法试图将正负类样本的增加与删除简化为两类样本的内部移动调整,一定程度上缓解由增加样本而产生的过拟合问题或是由删除样本而导致的信息丢失问题. 然而,当正样本中存在不止一个概念表达或重叠区域较大时,即存在类内子聚集和类别重叠现象. 空间中心 C 可能会位于决策边界且偏向负类的区域内,此时负样本与空间中心 C 的距离较小,相似性较大,不满足阈值判断条件,以致无法识别 PN 样本. 图 2 描述了当非均衡数据集中存在类内子聚集和类别重叠问题时,PNS 算法失效示意图. 其根本原因主要包含两个:一是非均衡数据中存在的类内子聚集及类别重叠会使正负类之间的相似关系表达不充分,且无法准确识别负类中潜在的与正类相似性较大的样本,而这些样本会反过来加剧类内子聚集和类别重叠问题,进而影响分类效果;二是正样本整体空间中心位置的确定方式,只能体现样本的总体平均分布特征,而不能有效刻画处于类内子聚集中的特殊样本,影响采样的精准性与分类的有效性.

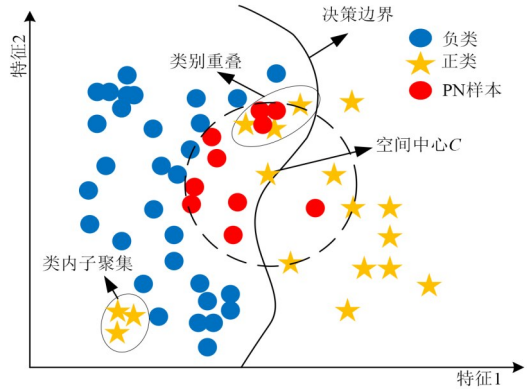

图 2　PNS 算法失效原因示意图

综上所述,类内子聚集和类别重叠是导致正类识别性能下降的主要因素. 相关研究应关注样本局部分布特征与空间分布特征,以提高分类精度.

为弱化样本原始的非均衡特征而强化局部特征,CBNMS 策略采用聚类思想划分样本,不仅可以充分表达各局部样本之间的相似性关系,挖掘潜在簇边界负样本,扩大正负类样本决策边界,而且可以有效解决类内子聚集问题,提高算法运行效率. 为更深入地加强 CBNMS 对正样本重叠区域的处理,ICBNMS 算法在上述策略均衡基础上引进 ADPST 技术,通过复合因子组合加权自适应确定最优合成样本,有效缓解类别重叠样本对分类性能的影响. ICBNMS 算法框架图如图 3 所示.

3.1　簇边界负样本移动策略

针对类内子聚集和类别重叠问题,充分考虑样本内部分布特征,由于某些负样本携带信息往往与正样本密切相关,不妨将这类样本视为隐藏在负类中的潜在边界样本,因而提出 CBNMS 策略,详细步骤如下.

Step1　采用凝聚层次聚类算法划分负子簇构成集合为 $C^-=\{C_i^-|i=1,2,\cdots,m\}$,进而划分正子簇构成集合为 $C^+=\{C_i^+|i=1,2,\cdots,n\}$.

Step2　计算各正子簇中所有正样本的平均值作为各正子簇的簇中心 M_i,构成集合为 $M=\{M_i|i=1,2,\cdots,n\}$,有

$$M_i=\frac{\sum_{k=1}^{S_i^+}x_k^+}{S_i^+} \tag{1}$$

其中,$i=1,2,\cdots,n$;x_k^+ 为正样本;S_i^+ 为第 i 个正子簇的规模.

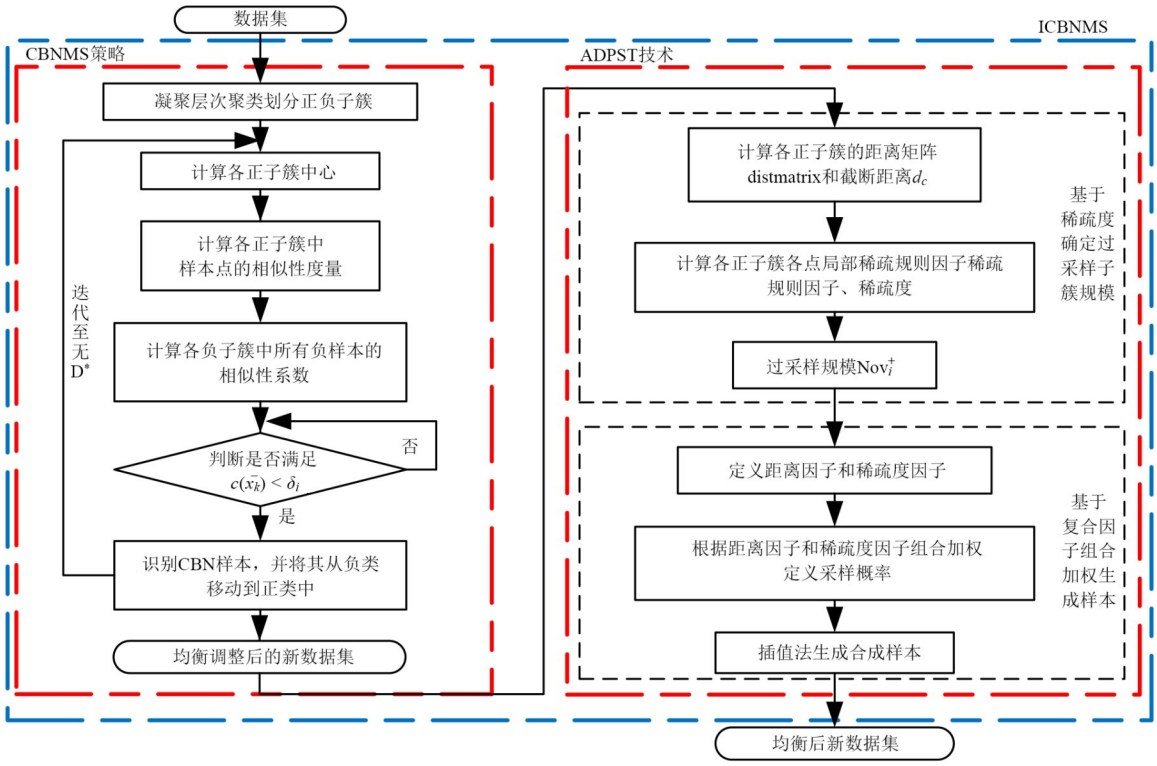

图3 ICBNMS算法框架图

Step3 计算各正子簇中的某一正样本x_k^+与其对应正子簇中心的欧氏距离平均值，作为各正子簇的相似性度量δ_i，构成集合为$\delta=\{\delta_i|\,i=1,2,\cdots,n\}$，有

$$\delta_i=\frac{\sum_{k=1}^{S_i^+}\mathrm{dist}\left(x_k^+,M_i\right)}{S_i^+} \tag{2}$$

其中，$i=1,2,\cdots,n$；dist(·)为距离函数；S_i^+为第i个正子簇的规模.

Step4 计算各负子簇中的某一负样本x_k^-与距离其最近的正子簇中心的距离，作为该负样本具有的相似系数$c(x_k^-)$，同时记录该正子簇下标p，有

$$c\left(x_k^-\right)=\min\left\{\mathrm{dist}\left(x_k^-,M_p\right)\right\} \tag{3}$$

其中，$k=1,2,\cdots,S_i^-$；$i=1,2,\cdots,m$；$p=1,2,\cdots,n$；S_i^-为第i个负子簇的规模.

Step5 判断是否满足条件$c(x_k^-)<\delta_i$，若满足，则定义该负子簇中的负样本x_k^-为“簇边界负样本”(CBN)，构成集合记为D^*. 同时记录该负样本在其所属负子簇中的下标i，有

$$D^*=\left\{x_k^-|c\left(x_k^-\right)<\delta_i\right\} \tag{4}$$

其中，$k=1,2,\cdots,S_i^-$；$i=1,2,\cdots,m$.

Step6 将Step5中的“簇边界负样本”集合D^*添加到正类中，并从负类中移除.

Step7 迭代步骤Step2~6，直至无法识别出CBN样本，算法结束.

图4(*a*)和图4(*b*)为正类样本出现类内子聚集现象时，PNS算法与CBNMS算法对比图. 若采用图4(*a*)中的PNS算法，则样本A满足阈值判断条件，而样本B不满足，但是由图直观观察到样本B也是其周围3个正样本的PN样本. 而若采用图4(*b*)中的CBNMS算法，针对以O_1和O_2为空间中心的正子簇C_1^+与C_2^+，负样本A与负样本B的相似性系数小于其对应正子簇C_1^+与C_2^+的相似性度量，于是负样本A与B被准确识别为CBN样本. 因而理论上CBNMS算法优于PNS算法.

3.2 自适应正样本合成技术

CBNMS策略仅通过调整局部簇边界负样本来解决具有类内子聚集的非均衡数据，但仍无法解决整体类间非均衡问题，尤其针对高非均衡比率数据集，两类样本在数量上仍处于非均衡分布. 为此，ICBNMS算法继续增强CBNMS策略，引入ADPST技术自适应确定最优样本生成区域，减少样本的重叠性并丰富样本的多样性. ADPST技术主要包括2个步骤：(1)根据各正子簇的稀疏度自适应确定所需过采样的子簇规模；(2)根据距离与稀疏度的复合因子组合加权生成合成样本.

3.2.1 基于稀疏度自适应确定采样子簇规模

针对CBNMS策略调整后的正类样本，可以根据稀疏度确定其所属正子簇的过采样规模，正子簇的稀疏度越大，所需合成样本越多，从而有效避免在密集区域

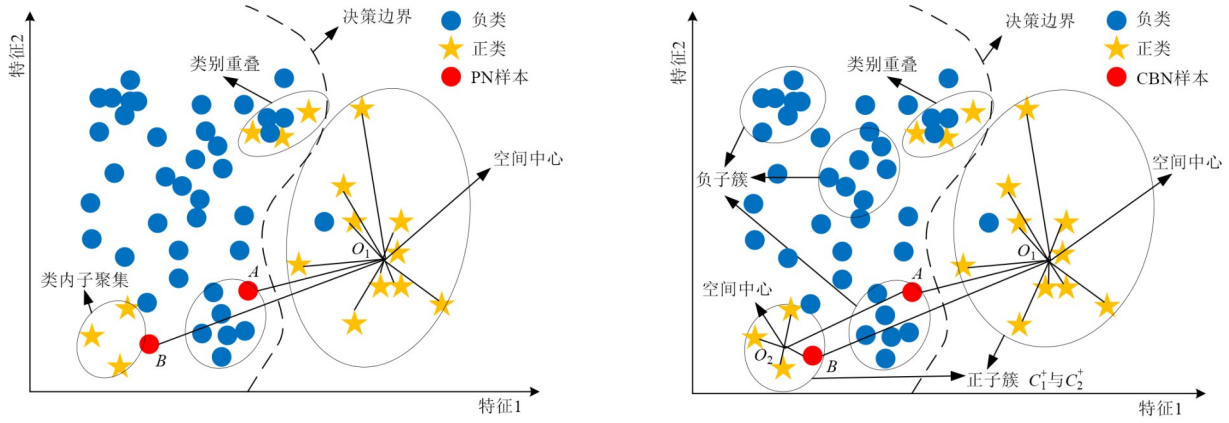

图4　PNS算法与CBNMS算法对比

集中产生样本与在合成正类样本时引起的类别重叠问题. 算法详细步骤如下.

Step1　计算各正子簇中各点的距离与截断距离 $d_c=\mathrm{sort}_{\mathrm{distmatrix}}(n\times \mathrm{percent}/100)$. 其中, distmatrix 为距离矩阵;$n$为距离总数;percent区间为[1,2],此处设为2.

Step2　计算各正子簇的局部稀疏规则因子L_i、稀疏规则因子SL_i^+及稀疏度SP_i^+,有

$$L_i=\sum_{j=i+1}^{N}\exp\left(-\left(\frac{d_{i,j}}{d_c}\right)\right),\mathrm{SL}_i^+=\sum_{j=1}^{S_i^+}L_j,\mathrm{SP}_i^+=\frac{1}{\mathrm{SL}_i^+}\quad(5)$$

其中,$i=1,2,\cdots,n$,n为正子簇个数;S_i^+为第i个正子簇的规模.

Step3　计算各正子簇所需过采样的样本数Nov_i^+,有

$$\mathrm{Nov}_i^+=\frac{\mathrm{SP}_i^+}{\sum\mathrm{SP}_i^+}\times\left(N^--N^+\right)\times\varepsilon\quad(6)$$

其中,$i=1,2,\cdots,n$;$\left(N^--N^+\right)$表示负样本与正样本的差值;$\varepsilon\in[0,1]$为采样规模调节系数.

3.2.2　基于复合因子组合加权生成合成样本

确定各正子簇所需过采样规模后,需寻找最优区域与最优方式生成合成样本. 一般地,对正类中含有不止一个类内子聚集的情况,采用近邻方法合成样本可能会从密集或小型子集群中生成重复或错误样本,致使分类模型学习难度加大. 鉴于此,根据距离和稀疏度的复合因子组合加权,以自适应确定最优样本区域并合成样本,具体步骤如下.

Step1　计算各正子簇中下标为t的样本x_t^+与其$k(k=5)$近邻中负类样本的欧氏距离d_t,归一化得$\hat{d}_t=d_t/\mathrm{dim}$($\mathrm{dim}$为正样本的维度),取其倒数的均值作为距离因子$\theta_d=\mathrm{mean}\left(1/\hat{d}_t\right)$,计算$x_t^+$的局部稀疏规则因子,取其倒数作为稀疏度因子$\theta_L=1/L_t$.

Step2　根据距离与稀疏度因子的复合因子,组合加权计算样本x_t^+的概率$P(x_t^+)$,有

$$P\left(x_t^+\right)=\alpha\times\frac{\theta_d}{\sum\theta_d}+\beta\times\frac{\theta_L}{\sum\theta_L}\quad(7)$$

其中,$i=1,2,\cdots,n$;$\alpha,\beta\in[0,1]$分别为距离、稀疏度调节系数,$\alpha+\beta=1$.

Step3　根据概率$P\left(x_t^+\right)$对该正子簇过采样,随机选择样本a与其属于同一子簇的最近邻样本b,并在a与b之间生成新样本c,$c=\delta a+(1-\delta)b$,$\delta\in[0,1]$为一随机数. 重复此步骤直至各子簇达到所需过采样规模并得到均衡样本.

4　实验与分析

为验证ICBNMS算法及其包含的CBNMS策略与ADPST技术的可行性与有效性,共设计5组实验:第1组实验对比探究改进算法在SVM,RF及KNN这3种分类器上的分类性能;第2组实验探究ADPST技术的参数敏感性;第3组实验以4种对比算法探究改进算法的优越性;第4组实验探究验证改进算法处理高非均衡比率数据集的有效性;第5组实验探究不同采样算法的时间效率. 以上实验环境均为Inter i7 CPU 2.20 GHz,RAM 8 GB,Windows 10操作系统,MATLAB R2017b数值实验平台. 数据集选自UCI,KEEL[18]数据库,详细信息见表1.

为避免数据量纲差异对实验结果造成的影响,采用离差标准化对数据预处理,将数据各属性值映射到[0,1]. 同时为消除数据随机分组对某类产生偏倚,参考文献[15]中的实验设置,即所有实验均采用5折交叉验证,每组数据独立运行5次,并取所有实验结果的平均值和标准差作为最终结果.

为验证算法的有效性,同时为确保算法不受特定分类器的限制,选择不同分类算法训练模型,从而使实验结论具有通用性. 比较各种分类算法,SVM[1]本质上

表1 非均衡数据集信息

Id	Dataset	#Att.	#Ex.	#Pos.	#Neg.	IR
1	Ecoli	7	336	35	301	8.60
2	SatImage	36	6 435	626	5 809	9.28
3	Yest_ME2	8	1 484	51	1 433	28.10
4	yeast1289vs7	8	947	30	917	30.56
5	yeast4	8	1 484	51	1 433	28.41
6	yeast5	8	1 484	44	1 440	32.78
7	abalone19	8	4 174	32	4 142	129.44
8	poker89vs5	9	3 316	49	3 267	66.67
9	shuttle2vs5	10	2075	25	2005	82.00
10	yeast6	8	1 484	35	1 449	41.40

注:#Att.表示数据集属性;#Ex.表示样本总数;#Pos.表示正类样本;#Neg.表示负类样本;IR表示非均衡比率,是负类样本个数与正类样本个数之比.

是研究二分类问题,它在处理小样本与高维数据时具有良好的泛化性能;RF[2]随机选择每个决策树作为基分类器,保证决策树间的差异性,具有良好的泛化性能和高效的训练速度;KNN[3]则对输入数据无要求且参数较少,仅依赖周围有限近邻样本实现分类. 因此,选择SVM,RF,KNN分别作为基分类器构建分类模型.

为凸显采样算法的特点,各分类器的参数均为默认值. SVM中,核函数为RBF核,γ值为1,惩罚因子C为1. RF中,单决策树个数为50. 同时,为彰显ICBNMS算法的优越性,将其中2个技术模块即CBNMS与ADPST以及4个采样算法即NCL[8],SMOTE[11],CDSMOTE[14]和PNS[15]作为对比算法,其中的参数设置均为默认值.

非均衡数据分类问题中,准确率不再是合适的评价指标,因为其只能反映分类器整体分类精度,无法强调正类的分类精度. 因此,文献[19~21]并在表2展示混淆矩阵及定义下述评价指标.

表2 混淆矩阵

	预测为正类	预测为负类
实际为正类	真正TP(True Positive)	假负FN(False Negative)
实际为负类	假正FP(False Positive)	真负TN(True Negative)

注:TP(FN)表示实际为正类,被预测为正类(负类)的样本数量;FP(TN)表示实际为负类,被预测为正类(负类)的样本数量.

正类召回率(也称为灵敏度),表示原实际正样本被预测为正样本的概率,即

$$\mathrm{TPR}=\mathrm{Sen}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{8}$$

正类精准率,表示预测为正样本中实际为正样本的概率,即正类预测准确率,即

$$\mathrm{PPV}=\mathrm{Pre}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{9}$$

负类特异度,表示原实际负样本被预测为负样本的概率,即

$$\mathrm{TNR}=\mathrm{Spe}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) \tag{10}$$

负类精准率,表示预测为负样本中实际为负样本的概率,即负类预测准确率,即

$$\mathrm{NPV}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FN}) \tag{11}$$

F-measure是正类召回率和精准率的调和均值,描述分类器对正类的识别性能,即

$$\text{F-measure}=\frac{2\times\mathrm{Recall}\times\mathrm{Pre}}{\mathrm{Recall}+\mathrm{Pre}} \tag{12}$$

G-mean是召回率和特异度的几何均值,体现分类器对整体的识别性能,即

$$\text{G-mean}=\sqrt{\mathrm{Sen}\times\mathrm{Spe}} \tag{13}$$

ROC曲线则能直观反映分类器性能的优劣,当曲线越趋于左上角时,曲线下面积AUC值越大,分类器性能越强.

在实际非均衡分析和应用中,不同角度将会产生不同的预测差异. 如若从提高正类分类精度的角度考虑,则可以将评价指标设置为F-measure指标最大化作为评价标准;如若需要同时从提高正类精度和分类器对数据集的整体分类性能角度考虑,则可以将F-measure指标与G-mean指标的特定加权作为评价标准. 总之,具体评价标准的设定需要根据实际需求进行综合考虑和选择性设定,不同的评价标准设定将会产生不同的效果.

4.1 改进算法及其技术的性能对比实验

为比较CBNMS,ADPST及ICBNMS算法在不同分类器上的分类性能,分别在SVM,KNN及RF分类器上进行实验. 表3为CBNMS,ADPST及ICBNMS算法在Ecoli,SatImage,Yest_ME2,yeast1289vs7,yeast4,yeast5共6个数据集,3个分类器上的G-mean和F-measure评价指标的平均值和标准差结果. 平均值越大且标准差越小说明分类准确度越高、稳定性越好,其中各评价指标在3个分类器中最优结果以加粗标注.

分析表3实验结果可知,相比于KNN,SVM分类器,绝大多数数据集的G-mean与F-measure指标在RF分类器上取得最优结果. 改进算法ICBNMS与其包含的2个技术模块即CBNMS策略、ADPST技术相比,总体性能明显进一步提升,而单个技术模块仅在个别数据集的某些指标上性能突出. 例如CBNMS策略在RF分类器上的F-measure指标获得最优性能,其中在SatImage,yeast1289vs7及yeast5这3个数据集上结果最佳,但在SVM与KNN分类器上表现均稍差,ADPST技术则仅在yeast5数据集上获得F-measure指标最优,表明以上2种技术对正类识别性能较强;而融合CBNMS策略与ADPST技术的ICBNMS算法几乎在多数数据集的RF分类器上的性能表现较好,如在Ecoli,yeast1289vs7,yeast4及yeast5这4个数据集上G-mean指标结果最优,表明融合算法ICBNMS对正负类整体识别性能更优.

表3　CBNMS,ADPST及ICBNMS算法在不同分类器上的评价指标结果

数据集	算法	SVM		RF		KNN	
		G-mean	F-measure	G-mean	F-measure	G-mean	F-measure
Ecoli	CBNMS	0.644±0.052	0.957±0.003	0.709±0.019	0.960±0.003	0.666±0.028	0.956±0.002
	ADPST	0.891±0.005	0.959±0.002	0.917±0.013	0.963±0.005	0.876±0.010	0.955±0.007
	ICBNMS	0.960±0.003	0.963±0.003	**0.963±0.001**	**0.963±0.001**	0.935±0.004	0.949±0.001
SatImage	CBNMS	0.724±0.006	0.968±0.000	0.723±0.007	**0.969±0.000**	0.803±0.006	0.967±0.000
	ADPST	0.902±0.042	0.965±0.006	0.941±0.000	0.947±0.000	0.912±0.038	0.960±0.003
	ICBNMS	**0.948±0.000**	0.953±0.000	0.944±0.000	0.947±0.000	0.936±0.000	0.942±0.000
Yest_ME2	CBNMS	0.543±0.077	0.982±0.001	0.435±0.029	0.982±0.001	0.382±0.047	**0.983±0.000**
	ADPST	0.887±0.003	0.981±0.003	0.901±0.003	0.917±0.003	0.789±0.012	0.979±0.004
	ICBNMS	**0.931±0.002**	0.920±0.002	0.900±0.005	0.885±0.005	0.898±0.004	0.883±0.004
yeast1289vs7	CBNMS	0.714±0.073	0.984±0.003	0.359±0.121	**0.984±0.001**	0.660±0.052	0.983±0.000
	ADPST	0.873±0.008	0.978±0.004	0.901±0.003	0.917±0.003	0.804±0.014	0.981±0.005
	ICBNMS	0.904±0.003	0.937±0.003	**0.929±0.002**	0.913±0.002	0.862±0.004	0.866±0.004
yeast4	CBNMS	0.643±0.075	0.982±0.000	0.447±0.029	0.984±0.000	0.411±0.036	0.983±0.000
	ADPST	0.786±0.002	0.983±0.002	0.763±0.028	**0.989±0.000**	0.859±0.108	0.981±0.015
	ICBNMS	0.903±0.004	0.922±0.005	**0.932±0.003**	0.888±0.004	0.897±0.005	0.884±0.005
yeast5	CBNMS	0.688±0.023	0.990±0.000	0.789±0.042	**0.991±0.001**	0.782±0.012	0.991±0.001
	ADPST	0.901±0.029	0.991±0.006	0.900±0.003	0.885±0.004	0.924±0.005	0.990±0.022
	ICBNMS	0.986±0.000	0.988±0.000	**0.988±0.001**	0.989±0.001	0.950±0.004	0.989±0.000

在RF分类器上,ICBNMS算法总体性能均优于CBNMS策略与ADPST技术.原因在于CBNMS策略只对正负样本进行移动调整,仅能缓解因类内子聚集引起的潜在边界样本不易识别问题;ADPST技术则着重强调各个子簇合成样本的质量而忽视簇边界样本的重要性;而ICBNMS算法通过融合以上2个技术来处理类内重叠样本,充分利用聚类优势并刻画样本间稀疏程度,这不仅保证了算法在聚类之前获取更多样本信息,还解决了过采样合成样本缺乏多样性的问题.

图5直观展示了改进算法ICBNMS及CBNMS,ADPST在3个分类器上的G-mean和F-measure指标的平均与总体水平的差异.

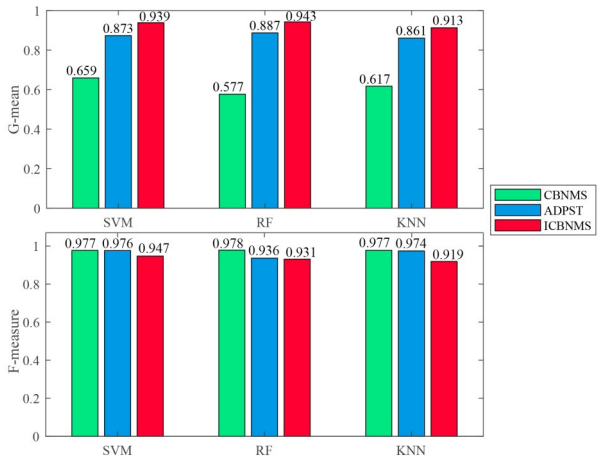

图5　不同分类器上的G-mean和F-measure均值结果

分析图5可知:

(1)针对G-mean指标,CBNMS策略在RF上仅仅获得0.577的中庸结果,分类性能稍次于SVM和KNN;而ADPST与ICBNMS算法的结果则大幅度增加,尤其在RF上性能表现突出,在SVM与KNN上结果仅次之,分别由高到低获得(0.887,0.873,0.861)与(0.943,0.939,9.913)的结果.但是仅对比3种算法在该指标值上的分类性能,ICBNMS算法明显优于CBNMS与ADPST算法,表明ICBNMS算法显著提升分类器整体分类性能.

(2)针对F-measure指标,3种改进算法在3个分类器上的总体效果无显著差异,CBNMS策略仅仅略胜一筹,说明其对正类样本有更强的识别能力且具有良好的泛化性能.

以上结果表明,3个改进算法在SVM与KNN分类器上的效果均欠佳,故RF分类器更适合非均衡数据分类问题.因此,综合以上2个指标,后续实验均选取RF作为基分类器处理均衡后数据集.

4.2　ADPST技术的参数性能对比实验

ADPST技术中,采样规模调节系数ε、距离与稀疏度调节系数α,β,均对该技术采样的效果产生一定影响.为评估ε和α,β的影响,选取4.1节中的6个数据集并以RF为基分类器,分别评估不同参数下AUC,G-mean,F-measure指标结果.

为评估采样规模调节系数对ADPST技术的影响,实验设置区间$\varepsilon\in[0.1,0.6]$,步长为0.1,经实验测试

初步筛选 $\varepsilon=0.2,0.4,0.6$ 的实验结果且最优结果见表4中加粗部分，由表4可以直观看出，当 $\varepsilon=0.6$ 时，AUC，G-mean，F-measure指标在大多数据集上均获最优.

为评估距离与稀疏度调节系数对ADPST技术的影响，实验设置当 $\varepsilon=0.6$ 时，(α,β) 不同组合为(0.2, 0.8)，(0.4, 0.6)，(0.5, 0.5)，(0.6, 0.4)，(0.8, 0.2). 实验结果如表5所示，最优结果以加粗标注，由表5可以直观看出，当 α,β 分别均为0.5时，所有数据集上分类性能均较优. 这是由于距离因子与稀疏度因子在合成样本时均具有同等代价权重，即距离越大、稀疏度越大，表示位于簇边界、稀疏域样本应赋予较大概率以合成样本.

表4 不同采样规模调节系数 ε 下的指标结果

数据集	评价指标	$\varepsilon=0.2$	$\varepsilon=0.4$	$\varepsilon=0.6$
Ecoli	AUC	0.972±0.003	0.981±0.002	**0.986±0.002**
	G-mean	0.916±0.011	0.941±0.006	**0.952±0.005**
	F-measure	**0.958±0.002**	0.957±0.004	0.957±0.004
SatImage	AUC	0.988±0.001	0.989±0.001	**0.989±0.001**
	G-mean	0.924±0.003	0.931±0.003	**0.934±0.003**
	F-measure	0.979±0.000	0.980±0.000	**0.980±0.000**
Yest_ME2	AUC	0.989±0.001	0.994±0.000	**0.996±0.000**
	G-mean	0.933±0.002	0.963±0.001	**0.973±0.001**
	F-measure	0.983±0.000	**0.983±0.000**	0.982±0.001
yeast1289vs7	AUC	0.972±0.003	0.985±0.002	**0.992±0.001**
	G-mean	0.937±0.004	0.958±0.002	**0.969±0.002**
	F-measure	**0.982±0.001**	0.980±0.001	0.979±0.001
yeast4	AUC	0.989±0.001	0.994±0.000	**0.995±0.000**
	G-mean	0.929±0.003	0.962±0.002	**0.973±0.001**
	F-measure	0.982±0.000	**0.983±0.001**	0.982±0.000
yeast5	AUC	0.984±0.006	0.981±0.006	**0.984±0.006**
	G-mean	0.779±0.031	**0.786±0.032**	0.754±0.032
	F-measure	0.991±0.000	0.991±0.000	**0.991±0.000**

4.3 其他采样算法的性能对比实验

为验证ICBNMS算法及其2个模块CBNMS与ADPST的优越性，依据4.2节参数设置 $\varepsilon=0.6$ 和 $\alpha=0.5,\beta=0.5$，将4个采样算法即NCL[8]，SMOTE[11]，CDSMOTE[14]及PNS[15]在RF分类器上的AUC，G-mean，F-measure，PPV，TPR，TNR及NPV共计7项指标的分类性能进行对比实验，实验结果如表6所示，最优结果以加粗标注. 为直观展示上述改进算法的鲁棒性，绘制ROC曲线图(图6).

由表6分析可知：

(1)相较于其他对比算法，ICBNMS算法在大多数数据集上的7项评价指标中有超过一半的指标值都显著高于其他对比算法，如AUC，G-mean，TNR，NPV指标均获得最优结果. 该算法TPR指标虽然没有达到最优值，但TNR指标明显优于其他算法，于是G-mean指标综合取得最优，表明ICBNMS算法以防造成误判能够尽可能多地识别负类，且对正负类整体识别精度较高.

表5 不同距离与稀疏度调节系数组合 (α,β) 下的指标结果

数据集	评价指标	(0.2, 0.8)	(0.4, 0.6)	(0.5, 0.5)	(0.6, 0.4)	(0.8, 0.2)
Ecoli	AUC	0.984±0.002	0.983±0.001	**0.985±0.001**	0.984±0.003	0.984±0.002
	G-mean	0.948±0.006	0.950±0.005	**0.951±0.004**	0.949±0.004	0.949±0.002
	F-measure	0.955±0.005	0.956±0.004	**0.957±0.003**	0.955±0.004	0.955±0.002
SatImage	AUC	0.993±0.000	0.993±0.001	**0.995±0.000**	0.994±0.000	0.994±0.000
	G-mean	0.957±0.000	0.956±0.001	**0.959+0.000**	0.957+0.000	0.957+0.000
	F-measure	0.970±0.000	0.971±0.000	**0.974+0.000**	**0.974+0.000**	0.970+0.000
Yest_ME2	AUC	0.995±0.000	0.995±0.000	**0.996±0.000**	0.995±0.000	**0.996±0.000**
	G-mean	0.973±0.001	0.973±0.000	**0.974±0.001**	0.972±0.001	0.973±0.001
	F-measure	0.978±0.001	0.982±0.000	**0.983±0.000**	0.981±0.001	0.982±0.000
yeast1289vs7	AUC	0.991±0.001	0.990±0.001	**0.992±0.001**	0.991±0.001	0.990±0.001
	G-mean	0.967±0.001	0.967±0.001	**0.969±0.002**	0.966±0.002	**0.969±0.002**
	F-measure	0.977±0.001	0.978±0.000	**0.980±0.001**	0.978±0.001	0.977±0.001
yeast4	AUC	0.989±0.000	0.992±0.000	**0.996±0.000**	0.995±0.000	0.994±0.000
	G-mean	0.972±0.001	0.972±0.000	**0.975±0.001**	0.972±0.001	0.973±0.001
	F-measure	0.979±0.000	0.977±0.001	**0.983±0.001**	0.981±0.001	0.982±0.000
yeast5	AUC	0.980±0.008	0.982±0.007	**0.986±0.005**	0.982±0.008	0.979±0.008
	G-mean	0.764±0.020	0.777±0.032	**0.786±0.027**	0.785±0.028	0.767±0.028
	F-measure	0.984±0.000	0.981±0.000	**0.992±0.001**	0.991±0.000	0.991±0.000

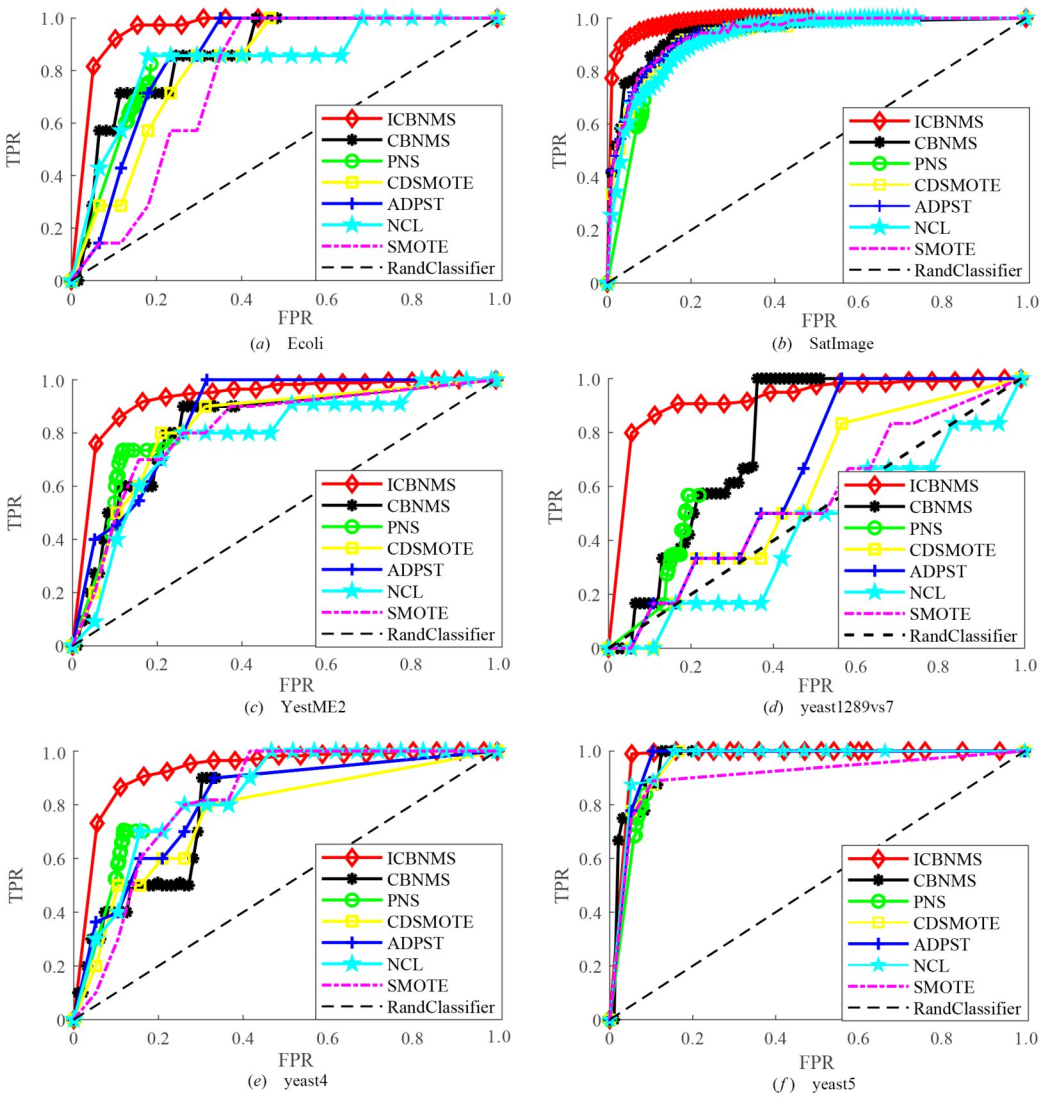

图6　其他采样算法的ROC曲线图

(2)CBNMS策略几乎在所有数据集上TPR指标取得最优,表明正类召回率越高,由该策略训练的模型能正确预测更多的正类样本,模型的效果越好. CBNMS策略与ADPST技术仅在Yest_ME2,yeast4,Ecoli数据集上的F-measure指标具有良好表现,而改进的融合算法ICBNMS则对实际为正类的样本具有准确的预测能力,说明该融合算法具有充分表达其内部技术模块优势的能力. 另外,CBNMS,ADPST及ICBNMS算法的各个指标相较于其他算法具有较小的标准差,说明其随机性远小于其他算法且普遍适用绝大多数数据集.

(3)其他对比算法在个别指标上也获得较优结果. 如CDSMOTE算法在PPV指标上表现较为突出,证实该算法针对预测为正样本的预测准确度较高;PNS算法在F-measure和TPR指标上与CBNMS算法达到同样水平,这表明该算法针对实际为正样本的预测准确度较高. 由此可知,不能仅以某项指标评估算法性能,需要从不同角度综合分析模型性能.

由图6分析可知,ICBNMS算法几乎在所有数据集上均取得最大曲线下面积,而CBNMS策略在多数数据集上也获得次大曲线下面积的结果,如Ecoli,SatImage,yeast1289vs7数据集. 针对SatImage这种具有高维、大量、高非均衡比率的数据集,ICBNMS,CBNMS,ADPST算法均取得较优的ROC曲线下面积,从而验证了以上3个改进算法具有有效性且对此类数据均具有普适性.

4.4　高非均衡比率数据集的性能对比实验

为验证CBNMS,ADPST及ICBNMS算法同样适用更高非均衡比率和特征较多的数据集,再选取4个高非均衡比率数据集,分别为abalone19,poker89vs5,shuttle2vs5及yeast6数据集,进行不同采样算法的对比实验,实验条件设置同4.3节,实验结果如表7所示,最优结果以加粗标注.

表6 其他采样算法的评价指标结果

数据集	算法	AUC	G-mean	F-measure	PPV	TPR	TNR	NPV
Ecoli	NCL	0.928±0.007	0.820±0.047	0.944±0.005	0.969±0.007	0.921±0.011	0.748±0.071	0.538±0.030
	SMOTE	0.917±0.009	0.729±0.040	0.949±0.002	0.951±0.007	0.950±0.004	0.582±0.065	0.590±0.022
	CDSMOTE	0.960±0.018	0.903±0.023	0.686±0.023	0.914±0.015	0.852±0.016	0.908±0.023	0.431±0.026
	PNS	0.929±0.007	0.703±0.033	0.960±0.003	0.946±0.004	**0.974±0.003**	0.525±0.038	0.788±0.037
	CBNMS	0.929±0.008	0.709±0.039	0.960±0.001	0.947±0.006	0.974±0.002	0.531±0.059	0.734±0.023
	ADPST	0.976±0.004	0.917±0.013	**0.963±0.005**	**0.969±0.005**	0.957±0.005	0.882±0.021	0.849±0.015
	ICBNMS	**0.985±0.002**	**0.942±0.003**	0.950±0.003	0.968±0.002	0.932±0.007	**0.952±0.004**	**0.904±0.010**
SatImage	NCL	0.954±0.001	0.799±0.002	0.961±0.000	0.963±0.000	0.960±0.001	0.666±0.004	0.645±0.005
	SMOTE	0.960±0.001	0.773±0.006	0.959±0.000	0.959±0.000	0.976±0.001	0.613±0.009	0.740±0.008
	CDSMOTE	0.961±0.001	0.808±0.005	0.965±0.000	**0.965±0.000**	0.966±0.001	0.677±0.008	0.687±0.011
	PNS	0.956±0.001	0.717±0.006	0.968±0.000	0.950±0.000	0.988±0.001	0.521±0.009	0.827±0.010
	CBNMS	0.956±0.002	0.723±0.002	0.969±0.000	0.951±0.000	**0.988±0.000**	0.531±0.004	0.833±0.005
	ADPST	0.989±0.000	0.941±0.000	0.947±0.000	0.946±0.001	0.948±0.001	0.934±0.002	**0.936±0.001**
	ICBNMS	**0.993±0.000**	**0.957±0.000**	**0.970±0.000**	0.956±0.001	0.984±0.000	**0.975±0.001**	0.930±0.001
Yest_ME2	NCL	0.905±0.016	0.558±0.017	0.980±0.000	0.976±0.000	0.985±0.001	0.329±0.020	0.478±0.032
	SMOTE	0.909±0.016	0.465±0.051	0.904±0.000	0.973±0.001	0.990±0.001	0.281±0.051	0.516±0.067
	CDSMOTE	0.916±0.009	0.647±0.019	0.973±0.001	**0.980±0.000**	0.967±0.002	0.445±0.020	0.335±0.026
	PNS	0.895±0.021	0.341±0.030	0.983±0.000	0.972±0.000	0.994±0.001	0.207±0.016	0.655±0.153
	CBNMS	0.908±0.013	0.435±0.087	**0.983±0.000**	0.972±0.001	**0.995±0.000**	0.217±0.055	0.601±0.085
	ADPST	0.968±0.000	0.897±0.004	0.882±0.004	0.887±0.005	0.879±0.006	0.916±0.004	0.911±0.004
	ICBNMS	**0.996±0.000**	**0.973±0.001**	0.982±0.000	0.972±0.001	0.993±0.001	**0.954±0.002**	**0.989±0.002**
yeast 1289vs7	NCL	0.720±0.026	0.375±0.114	0.983±0.000	**0.974±0.001**	0.993±0.000	0.193±0.054	0.460±0.168
	SMOTE	0.759±0.035	0.327±0.039	0.913±0.000	0.974±0.001	0.991±0.002	0.200±0.033	0.422±0.055
	CDSMOTE	0.736±0.015	0.602±0.045	0.552±0.059	0.549±0.070	0.668±0.018	0.666±0.074	0.579±0.030
	PNS	0.719±0.018	0.353±0.051	0.984±0.000	0.971±0.000	**0.996±0.000**	0.166±0.027	0.643±0.114
	CBNMS	0.709±0.032	0.309±0.069	0.980±0.000	0.972±0.000	0.995±0.000	0.146±0.129	0.476±0.171
	ADPST	0.968±0.001	0.901±0.003	0.917±0.003	0.899±0.003	0.937±0.006	0.866±0.005	0.918±0.007
	ICBNMS	**0.991±0.001**	**0.969±0.001**	**0.984±0.001**	0.969±0.001	0.990±0.001	**0.949±0.002**	**0.984±0.001**
yeast4	NCL	0.909±0.015	0.553±0.057	0.881±0.001	0.972±0.001	0.986±0.001	0.341±0.046	0.511±0.059
	SMOTE	0.898±0.010	0.441±0.053	0.873±0.001	0.973±0.001	0.990±0.001	0.239±0.037	0.455±0.102
	CDSMOTE	0.958±0.013	0.794±0.073	0.812±0.027	0.842±0.030	0.813±0.017	0.822±0.050	0.138±0.017
	PNS	0.908±0.015	0.436±0.039	0.983±0.000	0.972±0.000	0.995±0.000	0.207±0.018	0.650±0.087
	CBNMS	0.901±0.011	0.447±0.037	**0.984±0.000**	0.972±0.001	**0.996±0.001**	0.215±0.040	0.727±0.076
	ADPST	0.967±0.002	0.900±0.003	0.885±0.004	0.881±0.004	0.890±0.008	0.911±0.003	0.918±0.005
	ICBNMS	**0.996±0.000**	**0.973±0.002**	0.982±0.001	**0.976±0.001**	0.992±0.001	**0.954±0.002**	**0.987±0.002**
yeast5	NCL	0.984±0.006	0.882±0.017	0.889±0.000	0.893±0.000	0.990±0.000	0.792±0.027	0.725±0.020
	SMOTE	0.977±0.005	0.840±0.031	0.887±0.000	0.891±0.001	0.992±0.001	0.739±0.051	0.750±0.035
	CDSMOTE	0.984±0.008	0.969±0.014	0.940±0.008	0.987±0.000	0.930±0.017	0.958±0.009	0.331±0.015
	PNS	0.984±0.007	0.791±0.042	0.990±0.001	0.985±0.001	0.994±0.000	0.644±0.018	0.850±0.044
	CBNMS	0.983±0.010	0.789±0.036	0.989±0.001	0.986±0.001	0.994±0.001	0.635±0.054	0.779±0.053
	ADPST	0.982±0.006	0.763±0.028	0.989±0.000	**0.987±0.000**	**0.995±0.000**	0.602±0.027	0.800±0.036
	ICBNMS	**0.998±0.000**	**0.978±0.001**	**0.991±0.000**	0.986±0.001	0.992±0.000	**0.965±0.003**	**0.980±0.001**

由表7分析可知：

(1)针对以上4个高非均衡比率数据集，ICBNMS算法基本在所有指标上均获得最优结果，而CBNMS算法仅在yeast6数据集的F-measure和TPR指标上略胜出，这验证了ICBNMS算法具有良好的综合性能，适用了不同分布、不同维度和不同非均衡比率的数据.

表7　高非均衡比率数据集上对比算法的评价指标结果

数据集	算法	AUC	G-mean	F-measure	PPV	TPR	TNR	NPV
abalone19	NCL	0.697±0.013	0.737±0.048	0.806±0.032	0.902±0.000	0.854±0.000	0.825±0.015	0.879+0.042
	SMOTE	0.723±0.041	0.736±0.064	0.863±0.000	0.852±0.000	0.876±0.001	0.771±0.018	0.390±0.011
	CDSMOTE	0.972±0.010	0.723±0.020	0.769±0.023	0.816±0.015	0.742±0.046	0.819±0.026	0.773±0.039
	PNS	0.947±0.038	0.877±0.065	0.893±0.000	0.972±0.000	0.882±0.021	0.872±0.025	0.812+0.062
	CBNMS	0.994±0.012	0.874±0.020	0.983±0.001	**0.992±0.000**	0.979±0.000	0.687±0.010	0.887+0.001
	ADPST	0.995±0.000	0.964±0.002	0.967±0.001	0.965±0.002	0.969±0.001	0.958±0.002	0.963±0.001
	ICBNMS	**0.999±0.000**	**0.989±0.000**	**0.993±0.000**	0.990±0.000	**0.995±0.000**	**0.993±0.000**	**0.983±0.001**
poker89vs5	NCL	0.936±0.018	0.811±0.022	0.981±0.000	0.969±0.002	0.989±0.002	0.683±0.028	0.908±0.014
	SMOTE	0.934±0.016	0.855±0.025	0.972±0.005	0.971±0.004	0.968±0.008	0.762±0.044	0.714±0.067
	CDSMOTE	0.934±0.023	0.839±0.016	0.975±0.020	0.867±0.023	0.853±0.018	0.847±0.017	0.874±0.012
	PNS	0.939±0.013	0.913±0.056	0.978±0.005	0.971±0.006	0.984±0.005	0.690±0.077	0.807±0.082
	CBNMS	0.976±0.054	0.966±0.013	0.993±0.004	0.972±0.005	0.992±0.005	0.979±0.001	0.976±0.002
	ADPST	0.994±0.000	0.952±0.003	0.989±0.000	0.984±0.001	0.993±0.000	0.978±0.001	0.974±0.002
	ICBNMS	**0.998±0.002**	**0.983±0.001**	**0.993±0.000**	**0.987±0.000**	**0.999±0.000**	**0.989±0.001**	**0.983±0.001**
shuttle2vs5	NCL	0.964±0.011	0.782±0.013	0.870±0.013	0.882±0.011	0.896±0.018	0.870±0.014	0.890±0.019
	SMOTE	0.887±0.009	0.793±0.021	0.868±0.011	0.850±0.018	0.888±0.011	0.712±0.037	0.783±0.021
	CDSMOTE	0.885±0.009	0.879±0.003	0.968±0.006	0.837±0.002	0.810±0.013	0.673±0.005	0.805±0.023
	PNS	0.980±0.022	0.899±0.026	0.907±0.031	0.895±0.012	0.828±0.056	0.879±0.015	0.830±0.050
	CBNMS	0.989+0.004	0.955+0.002	0.932+0.007	0.942+0.006	0.934+0.011	0.996+0.000	0.991+0.004
	ADPST	0.992+0.001	0.954+0.006	0.947+0.003	0.951+0.002	0.927+0.011	0.995+0.001	0.996+0.002
	ICBNMS	**0.993+0.001**	**0.959+0.002**	**0.968+0.002**	**0.977+0.001**	**0.953+0.002**	**0.997+0.000**	**0.998+0.001**
yeast6	NCL	0.913±0.015	0.705±0.028	0.790±0.000	0.987±0.000	0.878±0.000	0.514±0.035	0.552±0.039
	SMOTE	0.890±0.011	0.731±0.036	0.766±0.001	0.889±0.001	0.884±0.002	0.565±0.046	0.484±0.033
	CDSMOTE	0.873±0.010	0.897±0.019	0.890±0.001	0.893±0.010	0.837±0.037	0.879±0.012	0.743±0.032
	PNS	0.842±0.008	0.805±0.031	0.910±0.022	0.987±0.001	0.891±0.001	0.802±0.043	0.611±0.063
	CBNMS	0.885±0.018	0.834±0.036	**0.991±0.000**	0.986±0.000	**0.996±0.000**	0.742±0.042	0.726±0.021
	ADPST	0.993±0.001	0.956±0.004	0.979±0.001	0.972±0.003	0.987±0.001	0.927±0.008	0.965±0.003
	ICBNMS	**0.996±0.000**	**0.985±0.001**	0.989±0.001	**0.988±0.001**	0.992±0.000	**0.987±0.001**	**0.978±0.002**

(2)对比算法在处理高非均衡比率数据集时,往往达不到预期效果,这是由于采样时增加和删除大量样本严重破坏样本的原始分布特征,进而影响分类效果.

为直观描绘改进算法的优越性,绘制4个高非均衡比率数据集对比算法的ROC曲线图(图7). 由图7可知,ICBNMS算法相较于其他对比算法具有更大的曲线下面积,说明该算法具有明显的性能优势. ADPST, CBNMS算法则仅在abalone19, yeast6数据集上获得次优结果,其他对比算法与ICBNMS算法相差甚远,如欠采样NCL算法因丢失价值较高样本而使分类器学习不充分,这表明这些算法处理高非均衡比率数据集时随机性较大、稳健性较差.

4.5　不同采样算法的时间对比实验

表8为不同采样方法在SVM, RF, KNN分类器上的时间对比与时间累积之和对比,时间单位为s,表中最优结果以加粗标注.

由表8分析可知,对比不同采样算法在3个分类器上的时间消耗,欠采样NCL算法用时累计最少仅为44.49s且明显比过采样SMOTE算法的用时(154.81s)的1/3还低. 混合采样算法累计用时也超过欠采样NCL算法,如CDSMOTE与PNS算法均比其高近2倍. 而CBNMS, ADPST及ICBNMS算法的累计用时与NCL算法相比无明显差距,但是平均效率与上述两种混合采样算法相比分别提升了32.27%和27.88%,尤其是CBNMS的效率提升更多,分别提升了38.65%和34.68%,这是由于该策略仅移动调整原始样本,而没有因增删样本而增加时间复杂度. 同样,ICBNMS算法虽然通过引入ADPST技术的聚类增加正样本,但是时间开销增加较少. 因此,改进混合采样算法ICBNMS及其包含的技术均具有显著的时间复杂度优势.

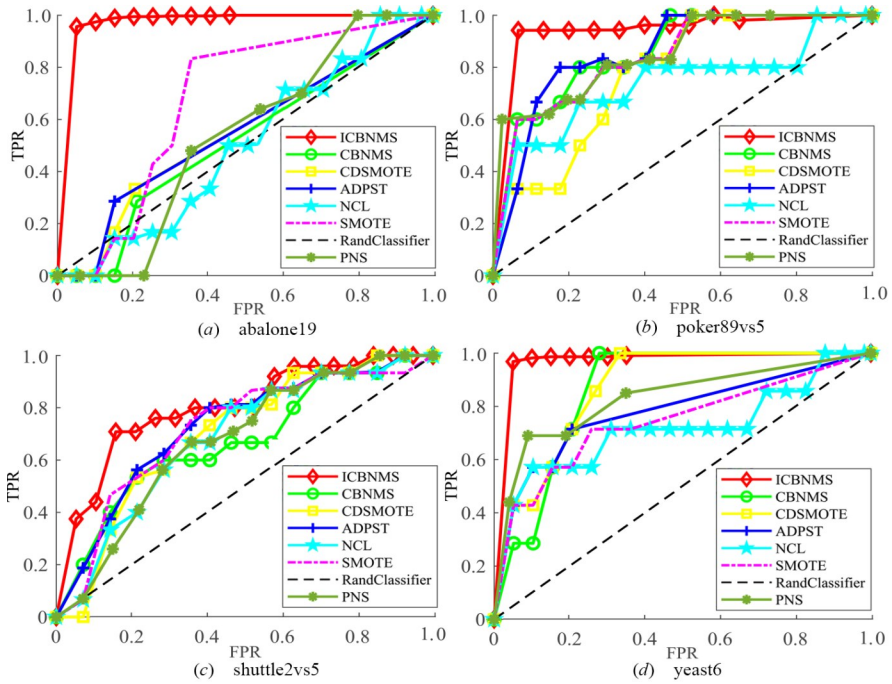

图7 高非均衡比率数据集上对比算法的ROC曲线图

表8 不同对比算法在不同分类器上的时间对比结果

数据集	分类器	NCL	SMOTE	CDSMOTE	PNS	ADPST	ICBNMS
Ecoli	SVM	**0.05**	0.09	0.37	0.32	0.09	0.09
	RF	0.53	0.65	1.64	1.44	0.57	0.54
	KNN	**0.11**	0.31	1.34	1.14	0.16	0.20
SatImage	SVM	**10.08**	28.59	12.45	12.56	10.57	12.74
	RF	**2.06**	7.30	4.64	4.43	2.13	3.12
	KNN	**24.02**	93.95	31.65	36.01	31.08	28.36
Yest_ME2	SVM	0.17	0.45	0.42	0.44	0.17	0.25
	RF	0.61	1.29	0.96	1.64	0.54	**0.52**
	KNN	1.34	5.12	2.63	1.54	**1.06**	1.21
yeast 1289vs7	SVM	**0.13**	0.30	0.55	0.42	0.17	0.15
	RF	**0.44**	0.91	1.88	1.47	0.56	0.47
	KNN	**0.61**	2.41	2.50	0.87	1.06	1.10
yeast4	SVM	**0.16**	0.46	0.48	0.36	0.26	0.24
	RF	**0.51**	1.10	1.04	1.26	0.58	0.55
	KNN	1.38	5.50	4.42	1.97	2.06	1.89
yeast5	SVM	**0.16**	0.41	0.50	0.47	0.17	0.31
	RF	0.61	0.89	2.96	2.76	0.59	**0.53**
	KNN	1.52	5.08	5.78	2.48	**1.06**	2.95
	总计	**44.49**	154.81	76.21	71.58	52.88	55.22

5 结论

为解决类内子聚集和类别重叠问题,本文提出一种融合簇边界负样本移动策略与自适应正样本合成技术的改进混合采样算法(ICBNMS). 该算法通过CBNMS策略对正负类样本有效划分以平衡数据的非均衡分布,并更深入引进ADPST技术,利用稀疏度与距离复合因子组合加权自适应确定采样规模,以成功规避类别重叠问题,增强算法局部采样的准确性和合成样本的多样性. 实验结果表明,ICBNMS算法的2个技术模块使其分类性能和运行时间效率相较于对比算法显著提高,验证了该算法具有良好的泛化性和稳健性,为非均衡数据的识别提供了新思路. ICBNMS算法虽然在

解决二分类非均衡数据问题上性能突出,但就如何使样本分布特征表达更为充分以及处理多类非均衡数据问题仍需继续进行研究.

参考文献

[1] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[2] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[3] HART P. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516.

[4] FOTOUHI S, ASADI S, KATTAN M W. A comprehensive data level analysis for cancer diagnosis on imbalanced data[J]. Journal of Biomedical Informatics, 2019, 90: 103089.

[5] MAKKI S, ASSAGHIR Z, TAHER Y, et al. An experimental study with imbalanced classification approaches for credit card fraud detection[J]. IEEE Access, 2019, 7: 93010-93022.

[6] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报, 2018, 46(1): 135-144.
HU F, WANG L, ZHOU Y. An oversampling method for imbalance data based on three-way decision model[J]. Acta Electronica Sinica, 2018, 46(1): 135-144. (in Chinese)

[7] HE H B, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[8] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine. Berlin: Springer, 2001: 63-66.

[9] TSAI C F, LIN W C, HU Y H, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019, 477: 47-54.

[10] VUTTIPITTAYAMONGKOL P, ELYAN E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data[J]. Information Sciences, 2020, 509: 47-70.

[11] 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.

[12] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences, 2019, 501: 118-135.

[13] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[14] ELYAN E, MORENO-GARCIA C F, JAYNE C. CDSMOTE: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification[J]. Neural Computing and Applications, 2021, 33 (7): 2839-2851.

[15] 张永清, 卢荣钊, 乔少杰, 等. 一种基于样本空间的类别不平衡数据采样方法[J]. 自动化学报, 2020, DOI: 10.16383/j.aas.c200034.

[16] VOORHEES E M. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval [J]. Information Processing & Management, 1986, 22(6): 465-476.

[17] NEKOOEIMEHR I, LAI-YUEN S K. Adaptive semi-unsupervised weighted oversampling(A-SUWO) for imbalanced datasets[J]. Expert Systems with Applications, 2016, 46: 405-416.

[18] ALCALÁ -FDEZ J, SÁNCHEZ L, GARCÍA S, et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems[J]. Soft Computing, 2009, 13 (3): 307-318.

[19] JAPKOWICZ N. Assessment metrics for imbalanced learning[M]//Imbalanced Learning. Hoboken: John Wiley & Sons, Inc., 2013: 187-206.

[20] CHEN B Y, XIA S Y, CHEN Z Z, et al. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise[J]. Information Sciences, 2021, 553: 397-428.

[21] GAO X, REN B, ZHANG H, et al. An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling[J]. Expert Systems with Applications, 2020, 160: 113660.

作者简介

高雷阜　男,1963 年 2 月出生,辽宁阜新人. 博士,教授,博士生导师. 主要研究方向为最优化理论与方法、机器学习与数据分析.
E-mail: gaoleifu@163.com

张梦瑶　女,1996 年 1 月出生,内蒙古呼伦贝尔人. 硕士研究生. 主要研究方向为机器学习与数据分析.
E-mail: mengyaoz119@163.com