控制与决策
Control and Decision

一种混合CGAN与SMOTEENN的不平衡数据处理方法

刘宁, 朱波, 阴艳超, 李岫宸

引用本文:
刘宁,朱波,阴艳超,李岫宸. 一种混合CGAN与SMOTEENN的不平衡数据处理方法[J]. 控制与决策, 2023, 38(9): 2614−2621.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.1780

一种混合CGAN与SMOTEENN的不平衡数据处理方法

刘 宁, 朱 波†, 阴艳超, 李岫宸

(昆明理工大学 机电工程学院，昆明 650500)

摘 要: CGAN能够从数据中学习其分布特性,被引入不平衡数据处理中对少数类样本进行过采样,可以生成符合原始数据分布的新样本,因此比传统的重采样方法具有更好的处理效果. 然而,CGAN对数据分布特性的学习易受限于样本规模,在少数类样本规模较小时不能充分学习其分布特性,难以保证生成样本的质量. 针对这一问题,提出一种将CGAN与SMOTEENN相结合的不平衡数据平衡化处理方法. 首先,从既有的少数类样本出发,采用SMOTEENN方法生成一定规模的少数类样本;然后,在此基础上训练CGAN模型,保证其能够生成符合原始少数类样本分布特征的新样本;最后,再利用CGAN重新生成符合原始少数类样本分布的新样本构建平衡数据集. 为验证所提出方法的有效性,基于公开的不平衡数据集开展对比实验研究. 实验结果表明,相对几种经典的不平衡数据处理方法与近期文献报道的方法,所提出方法在几项不平衡数据分类评价指标上表现出明显的优势.

关键词: 不平衡数据；数据平衡化处理；重采样方法；CGAN；SMOTEENN

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.1780 **开放科学(资源服务)标识码(OSID):**

引用格式: 刘宁,朱波,阴艳超,等. 一种混合CGAN与SMOTEENN的不平衡数据处理方法[J]. 控制与决策, 2023, 38(9): 2614-2621.

An imbalanced data processing method based on hybrid CGAN and SMOTEENN

LIU Ning, ZHU Bo†, YIN Yan-chao, LI Xiu-chen

(Faculty of Mechanical and Electrical Engineering，Kunming University of Science and Technology，Kunming 650500，China)

Abstract: Conditional generative adversarial networks (CGAN) can learn its distribution characteristics from the data, and is introduced into the imbalanced data processing to oversample the minority class samples, which can generate new samples that conform to the original data distribution, so it has a better processing effect than traditional resampling methods. However, the learning of data distribution characteristics by a CGAN is easily limited by the sample size. When the sample size of the minority class is small, its distribution characteristics cannot be fully learned, and it is difficult to ensure the quality of the generated samples. To solve this problem, this paper proposes an unbalanced data balance processing method combined with the CGAN and the synthetic minority over-sampling technique edited nearest neighbor (SMOTEENN). Firstly, starting from the existing minority class samples, the SMOTEENN method is used to generate a certain scale of minority class samples, and then the CGAN model is trained on this basis to ensure that it can generate consistent the new samples with the distribution characteristics of the original minority class samples. Finally, the CGAN is used to regenerate new samples that conform to the original minority class sample distribution to construct a balanced dataset. The experimental results show that, compared with several classical imbalanced data processing methods and methods reported in recent literature, the proposed method has obvious advantages in several imbalanced data classification evaluation indicators.

Keywords: imbalanced data classification；data balancing processing；resampling method；CGAN；SMOTEEN

0 引 言

不平衡数据分类问题广泛存在于实际应用场景的多个领域中,如医疗诊断[1]、故障检测[2]、金融欺诈[3]等分类应用领域. 在数据不平衡条件下直接构建分类模型会导致分类器将更多的关注度集中于多数类样本,难以保证少数类样本的识别精度. 然而,少数类所包含的信息通常更受关注,其误判的代价也更高,因此,在数据分布不平衡条件下提升少数类的分

收稿日期: 2021-10-15；**录用日期:** 2022-04-15.
基金项目: 国家自然科学基金项目(52065033).
†通讯作者. E-mail: zhubo20110720@163.com.

类精度十分必要,成为近年来学术研究的热点.

目前,国内外学者主要从数据平衡化处理、分类算法改进或两者结合的途径解决分类中的数据不平衡问题. 数据平衡化处理以过采样和欠采样为主要手段,独立于分类器,具有对各种具体分类算法的广泛适用性,因此应用较为广泛,也是本文关注的重点. 其中过采样和欠采样分别通过扩增少数类样本和删除部分多数类样本构建平衡数据集,但欠采样在删除样本过程中易删除对分类影响较大的多数类样本从而造成重要信息丢失,因此众多学者更侧重于过采样方法的相关研究,较具代表性的如SMOTE (synthetic minority oversampling technique)方法[4]、Borderline-SMOTE[5]、 ADASYN[6]. Barua 等[7] 在ADASYN方法的基础上提出了以难以学习的少数类样本和多数类样本的欧氏距离作为采样权重的过采样方法. Batista 等[8] 提出了一种 Tomek links 混合采样方法. Cheng 等[9] 基于噪声过滤机制提出了一种GSMOTE-NFM 方法. Li 等[10] 提出了一种 SMOTE-NaN-DE 过采样方法. Batista 等[11] 提出的SMTOEENN 方法在SMOTE 的基础上利用编辑近邻过采样方法(edited nearest neighbor, ENN)实现了对数据的深度清洗. 然而,SMOTE 以及相关改进方法本质上均是从少数类样本的局部邻域出发,没有充分挖掘数据的分布信息,使得生成的新样本不能较好地还原原始样本的分布特性,用其处理后的不平衡数据对分类性能的提升有限.

近年来,随着大数据和人工智能技术的兴起,类不平衡条件下基于深度学习的数据扩增方法开始成为新的研究热点,如Huang 等[12] 针对液压系统智能故障诊断中所存在的问题提出了一种多速率数据样本的深度学习模型. Jiang 等[13] 针对工业过程所采集的数据存在不平衡、匹配不准确和部分缺失问题,提出了一种混合粒度的数据增强策略. 易令等[14] 提出了一种光谱数据扩增方法用于解决小样本原油总氢物性回归预测问题. 此外,生成对抗网络(generative adversarial networks, GAN)[15] 作为深度学习领域中的新兴技术之一,也被引入不平衡数据处理中. 目前GAN已在金融欺诈[16] 和故障检测[17] 等领域得以应用并取得初步成效. 但由于GAN模型过于自由,在训练过程中难以达到稳定,易出现梯度消失和模型崩溃的问题. 而条件生成对抗网络(conditional generative adversarial networks, CGAN)在GAN的基础上增加外部标签信息用以指导生成对抗网络的训练,在一定程度上解决了GAN存在的问题[18]. Douzas 等[19] 通过实验验证了采用CGAN过采样处理后的样本集可以提高少数类的分类精度. 赵海霞等[20] 针对具有类别重叠的不平衡数据集,提出了一种RECGAN重抽样方法,并在不同数据集中验证了所提出方法的有效性. 然而,CGAN在训练过程中存在一个问题,即需要带标签的样本达到一定数量才能够有效学习数据的分布特性,若不平衡数据集中少数类样本规模较小,条件生成对抗网络将难以充分学习少数类样本的分布特征,从而也无法保证生成样本的质量.

鉴于CGAN在过采样上存在的优势及其学习数据分类特性的能力受限于少数类样本集规模的问题,本文提出一种将CGAN与SMOTEENN相结合的不平衡数据处理方法. 为验证所提出方法的有效性,在8个公开的不平衡分类数据集上,以 F_1 值、 Recall、AUC和G-mean为评价指标进行了对比实验. 实验结果表明,相比经典的不平衡数据处理方法与近期文献中报道的几种新方法,所提出方法具有明显优势.

1　相关理论

1.1　SMOTEENN

SMOTE方法采用线性插值方式在少数类样本与k近邻样本之间合成新的少数类样本,在一定程度上解决了随机过采样造成的信息冗余,但在合成样本中易出现样本重叠和噪声样本等问题. SMOTEENN是在SMOTE的基础上采用ENN方法对后者所生成的数据进行深度清洗,已在多个标准数据集上证明其性能通常优于其他经典采样方法[21],也因此成为一种广受关注的SMOTE改进方法.

1.2　CGAN

GAN由生成网络G和判别网络D两部分组成. 生成网络通过对原始数据分布特征信息的学习,可将输入的多维随机噪声转化为类似原始数据分布的生成数据;判别网络是一个二分类器,其目的是对输入数据进行辨别,判断输入数据是真实样本还是合成样本的概率. 在两者迭代优化并达到纳什均衡时,理论上生成网络可以生成无限接近于真实数据分布特性的样本. 设输入的随机噪声为z,生成网络G将其转化为生成样本$G(z)$. 判别网络D输出为$[0,1]$范围内的实数$D(x)$. GAN的网络结构如图1所示,GAN模型的目标函数为

$$\min_G \max_D V(G,D) = E_{x\sim P_r}\{\log[D(x)]\} + E_{z\sim P_z}\{\log[1-D(G(z|y))]\}. \tag{1}$$

其中: x为真实样本, P_r为真实样本分布, P_z为随机噪声分布.

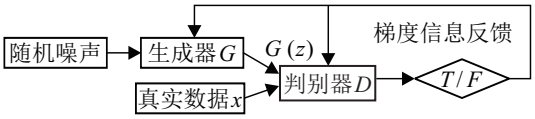

图 1 生成对抗网络原理

Mirza等[18]针对GAN存在的问题, 在GAN的基础上额外添加了条件y, 提出了CGAN用以引导GAN的训练. 随机噪声z和条件值y(如数据的标签)作为生成网络的输入, $G(z|y)$为生成网络的生成样本. 真实样本x、生成样本以及条件y一同输入判别模型进行判别. CGAN的目标函数在GAN的基础上作出了修改, 新的目标函数如下式所示:

$$\min_G \max_D V(G,D) = E_{x\sim P_r}\{\log[D(x|y)]\} + E_{z\sim P_z}\{\log[1-D(G(z|y))]\}. \tag{2}$$

2 SMOTEENN_CGAN方法

CGAN作为过采样方法, 在处理不平衡数据时易受限于样本规模的问题已隐含地反映在目标函数中[22], 若少数类样本规模较小, 则CGAN在起初训练过程中的生成器会因为没有学到少数类的分布特征导致生成的样本多数为噪声样本, 此时判别器易于根据真实数据分辨出噪声数据, 随着训练的进行很快进入判别器的最佳状态, 并可以轻松地识别出所输入的数据为生成数据, 输出$D(G(z|y))=0$并导致$\text{Loss}_G=\log(1-D(G(z|y)))\approx 0$. 对于深度学习模型而言, 其模型参数的更新受限于函数值, 若函数值为0则会造成生成网络的参数处于停滞状态, 因此其输出结果无法得到保证.

针对以上问题, 本文提出一种改进CGAN的过采样方法. 该方法首先采用SMTOEENN方法对少数类样本进行扩增, 并使其达到一定的规模; 然后通过既有规模的少数类样本训练CGAN并保证其能够充分学习少数类样本的分布特性; 最后再利用可以较好学习少数类样本分布特性的CGAN模型实际扩增少数类样本. 具体流程如图2所示.

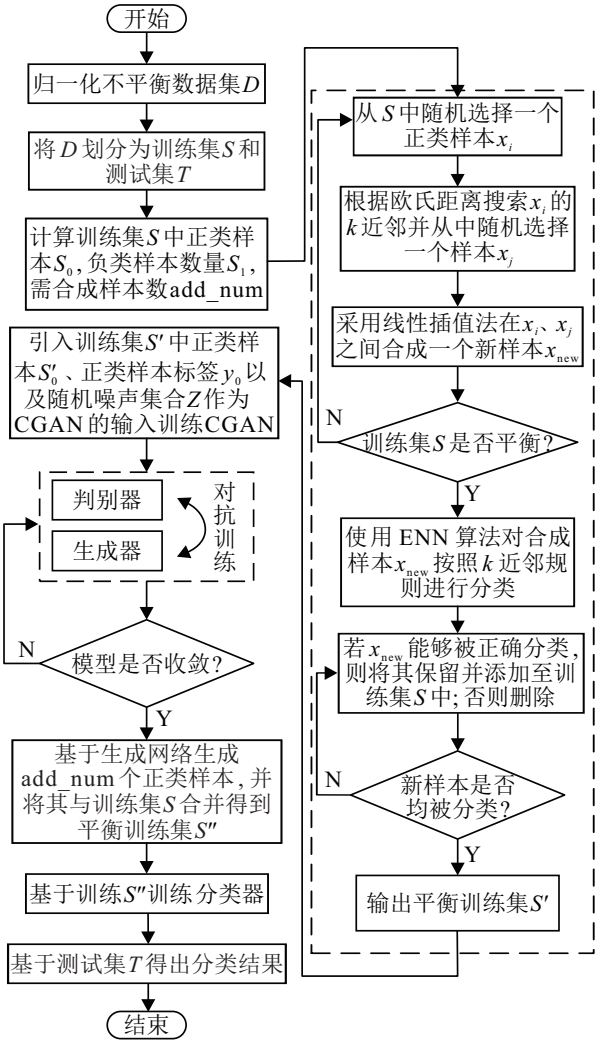

图 2 SMOTEENN_CGAN方法流程

3 实验分析

3.1 数据来源和评价指标

为了验证所提出方法的有效性和实用性, 从KEEL和UCI数据库中选取了8组不平衡率在2～10的公开不平衡数据集进行实验验证, 表1为8组数据集的具体信息. 为了提升模型的收敛速度和精度, 本文首先对原始数据集归一化处理, 然后选取70%的样本作为训练集, 其余30%用作测试集.

表 1 数据集特性描述

名称	总样本	少数	多数	特征数	IR
pima	768	268	500	8	1.87
phoneme	5 404	1 586	3 818	5	2.41
yeast_1	1 484	429	1 055	8	2.46
haberman	306	81	225	3	2.78
ecoli_1	336	77	259	7	3.36
newthyroid	215	35	180	5	5.14
wine_red	1 599	199	1 400	11	7.04
yeast_3	506	50	456	8	9.12

对于不平衡数据分类性能的评价, 整体分类正确率并不能较好地衡量其分类能力, 因此本文采用F_1值、Recall、AUC和G-mean这4个指标对不平衡数据的分类性能进行评估. 其中F_1值为精确率Precision与召回率Recall的调和平均, 是不平衡数据分类评价中最常见的评价标准; Recall值衡量模型对正类样本的敏感度; AUC等价于ROC曲线下方的面积, 与数据分布无关, 是从总体上评价分类器性能更便利的一种

方法. G-mean值用于度量分类器在两类数据上的平均性能.

3.2　CGAN超参数设置

文中判别网络和生成网络均采用全连接神经网络, 合适的神经网络隐藏层数和隐藏层单元数可提高模型的特征提取能力, 从而进一步提升其泛化性能. 本文通过多组预实验对不同数据集构建多种网络模型, 基于上述评价指标比较不同模型的泛化性能并最终确定不同数据集对应的生成网络以及判别网络的隐藏层和隐藏层节点数, 如表2所示. 本文隐藏层均采用LeakyRelu激活函数. 随机性可提高训练过程的稳定性, 对于含有多隐藏层的生成网络使用Droupout可防止CGAN在训练过程中以各种方式“卡住”. 在判别网络各层级间采用批标准化加速模型收敛并减缓过拟合. 生成网络和判别网络的输出层分别采用tanh和sigmoid函数.

表 2　生成对抗网络隐藏层及隐藏层节点数

数据集	生成网络 G		判别网络 D	
	隐藏层数	神经元数	隐藏层数	神经元数
pima	2	256/512	2	256/512
phoneme	2	256/512	2	128/256
$yeast_1$	3	256/512/1 024	3	256/512/1 024
haberman	1	16	1	8
$ecoli_1$	3	256/512/1 024	3	128/256/512
newthyroid	1	64	1	48
wine_red	2	256/512	2	128/256
$yeast_3$	2	128/256	2	64/128

3.3　方法验证与结果分析

3.3.1　SMOTEENN_CGAN验证及可视化对比分析

首先, 以不平衡数据相关研究中使用频率较高的$ecoli_1$数据集为例对SMOTEENN_CGAN方法进行方法验证, 随着训练的进行, 生成网络和判别网络的损失值不断降低并最终达到平衡. 训练过程中生成网络和判别网络的损失函数图像如图3所示, 由图3可见, 随着训练的进行, 对抗网络模型逐渐收敛, 生成网络的损失值和判别网络的损失值均不再发生变化, 表明此时所提出方法生成的数据已具备与真实数据分布一致的能力, 模型训练完毕. 然后基于SVM分类器绘制不同采样方法的ROC曲线, 综合图4中不同方法的ROC曲线和AUC值可以看出, 使用所提出方法处理后的数据在SVM分类器上进行分类实验得到的结果优于其他采样方法所得到的分类结果.

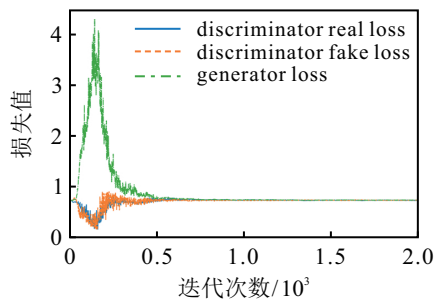

图 3　生成对抗网络原理

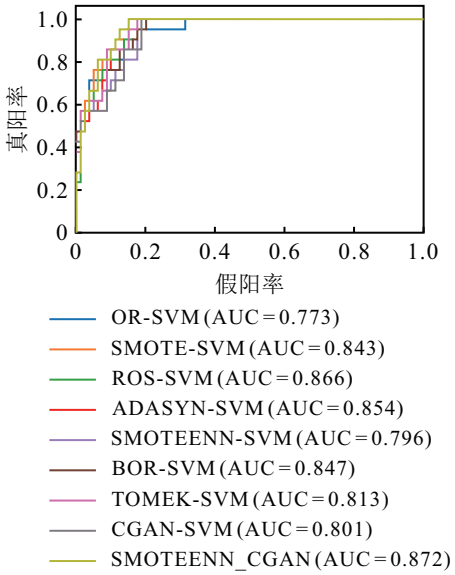

图 4　不同采样方法的ROC曲线

为了直观地体现所提出方法生成的样本能够较好地模拟原始数据分布特征, 本文将通过可视化展示对各种方法进行比较. 首先基于Python 3.7生成不平衡比为10, 属性特征为2且存在交叠区的不平衡数据集, 其中少数类样本量为50, 原始数据的分布效果如图5(a)所示; 然后分别对不同采样方法生成的数据进行可视化对比分析, 如图5所示. 由图5可见, SMOTE方法通过线性插值方式所生成的样本易产生一定数量的重复样本, 由于在交叠区的少数样本同样有机会被作为类范本以合成新样本, 会造成交叠区变得更加繁杂. SMOTEENN在SMOTE方法的基础上添加了数据清晰功能, 可在一定程度上减少交叠区的噪声样本, 但由于ENN算法在分类决策上仅依靠最临近的几个样本决定新样本的所属类别, 其分类结果会受到K值的影响, 难以保证对SMOTE方法所生成的样本完全分类正确, 也因此不能完全去除SMOTE方法在交叠区生成的新样本, 依然会在交叠区留有小部分的噪声样本. ADASYN在生成样本时更偏向于边界区域的少数类样本, 易于在边界区域生成更多的类重叠样本, 加剧交迭区的繁杂程度. CGAN作为一种需要一定规模数据支撑生成器和判别器训练收敛的神经网络, 对原始少数类样本分布特性的学习会受到少数类样本规模的影响, 在样本规模较小时难以达到一个满意的纳什平衡, 且易陷入模式崩溃, 导致在样本边

界附近产生噪声样本,对样本多样性的提升也无明显帮助. 所提出方法首先通过SMTOEENN过采样处理少数类样本并使其达到一定规模后供CGAN训练学习,由于CGAN的生成网络在GAN的基础上添加了标签信息,在生成样本时具有利于分类的导向性,在训练过程中会更多关注于SMOTEENN在非交叠区生成的规模较大的高质量样本,充分挖掘其分布特性并最大限度上减少交叠区小部分噪声样本对自身学习造成的干扰,最终保证所生成的样本尽可能远离交叠区. 下文将基于上述评价指标对不同采样方法的分类性能作出进一步比较.

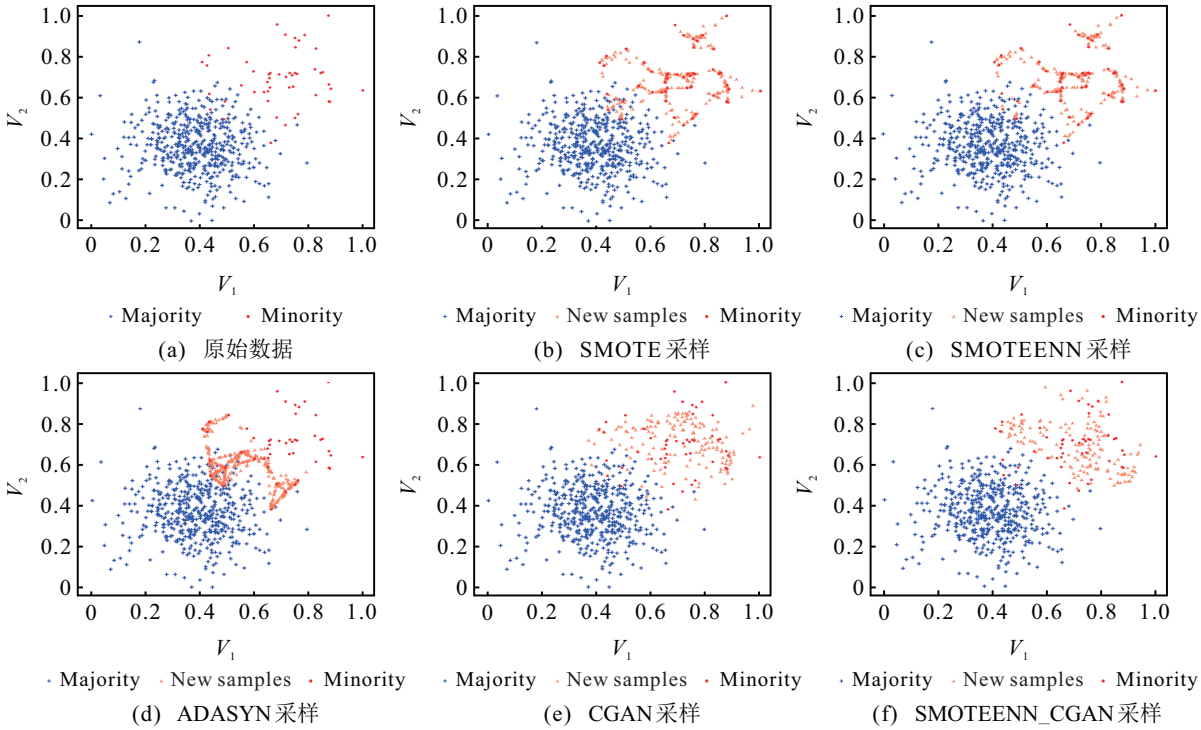

图 5 相同训练集下不同采样方法采样效果

3.3.2 不同过采样方法分类性能对比分析

实验采用SVM分类器对所提出方法开展对比实验研究,将所提出方法与SMOTE、ROS、ADASYN等7种采样方法进行比较. 此外,本文将原始数据的分类结果也作为比较对象,为保证实验结果不受随机因素的干扰和影响,首先将归一化处理后的原始样本划分为训练集和测试集,然后使用不同采样对少数类样本过采样处理得到平衡训练集,在此基础上基于SVM分类器作分类实验,最后在测试集上进行测试并得出Recall、F-measure、G-mean和AUC的值. 如此重复10次并取测试结果的平均值作为最终的实验结果. 将每个评价指标下的最大值用粗体标出,不同指标下的实验结果如表3所示. 由表3可见,所提出方法对不平衡数据的分类效果有较大地提升且优于其他几种采样方法,尤其是在F_1、AUC、G-mean这3个综合评价指标上表现突出,在不同数据集上的分类结果基本均达到了最优值,虽然在小部分数据集中的分类测试中没有达到最好的效果,但差距也在可接受的范围内. 观察另一个指标Recall,所提出方法也表现出一定的优势,仅在部分数据集上所得出的Recall值略低但也取得了不错的分类效果. 分析其原因,除了与模型的训练次数、超参数有一定的关系,也可能是由于原始数据集中存在较多的重叠样本,在生成样本时会引入部分噪声样本. 综上,所提出方法在召回率与精确率之间得到了较好的平衡,在召回率取得不错表现的同时在F_1、AUC、和G-mean这3个综合性指标上均有着明显的优势,可以说所提出方法对不平衡数据分类性能的提升优于其他方法.

为了进一步验证所提出方法的有效性,将所提出方法与文献[20]提出的类似方法RECGAN、文献[9]提出的GSMOTE-NFM方法在数据集、评价指标以及分类器相同的情况下作以比较,对比结果如表4所示. 由表4可见,所提出方法SMOTEENN_CGAN在多个数据集上表现出一定的优势,尤其在少数类样本规模较小的3个数据集(newthyroid,少数类样本个数为35; yeast_3,少数类样本个数为50; ecoli_1,少数类样本个数为77)上表现突出,以此可以表明所提出方法在处理不平衡数据尤其是少数类样本规模较小时表现更具有优势.

表 3　SVM分类器下不同数据集的分类性能

评价指标	采样方法	数据集							
		$yeast_3$	$ecoli_1$	haberman	$yeast_1$	phoneme	pima	newthyroid	wine_red
F_1 值	原始数据	0.303	0.768	0.173	0.405	0.66	0.627	0.855	0.115
	SMOTE	0.355	0.761	0.437	0.549	0.703	0.653	0.904	0.422
	ROS	0.329	0.759	0.455	0.555	0.705	**0.656**	0.901	0.417
	ADASYN	0.312	0.76	0.434	0.549	0.698	0.652	0.887	0.408
	SMOTENNN	0.316	0.761	0.451	0.553	0.705	0.653	0.896	0.408
	Borderline-SMOTE	0.354	0.763	0.435	0.543	0.697	0.648	0.882	0.407
	SMOTE-Tomek	0.355	0.761	0.438	0.549	0.703	0.653	0.904	0.422
	CGAN	0.396	0.78	0.473	0.561	0.684	0.638	0.88	0.373
	SMOTEENN_CGAN	**0.415**	**0.79**	**0.503**	**0.573**	**0.714**	0.655	**0.909**	**0.462**
AUC	原始数据	0.595	0.841	0.55	0.62	0.76	0.719	0.91	0.538
	SMOTE	0.697	0.881	0.624	0.69	0.805	0.73	0.965	0.743
	ROS	0.691	0.885	0.636	0.693	**0.808**	0.732	0.966	0.752
	ADASYN	0.673	0.893	**0.652**	0.687	0.806	0.728	0.963	0.739
	SMOTENNN	0.697	0.886	0.62	0.688	0.808	0.724	0.96	0.754
	Borderline-SMOTE	0.691	0.892	0.61	0.682	0.805	0.723	0.958	0.725
	SMOTE-Tomek	0.697	0.882	0.62	0.689	0.805	0.731	0.965	0.743
	CGAN	0.706	0.899	0.642	0.708	0.788	**0.737**	0.926	0.686
	SMOTEENN_CGAN	**0.746**	**0.906**	0.646	**0.712**	0.798	0.735	**0.962**	**0.766**
Recall	原始数据	0.196	0.724	0.119	0.298	0.624	0.621	0.787	0.09
	SMOTE	0.575	0.897	0.523	0.658	0.834	0.756	0.94	0.746
	ROS	0.609	0.918	0.482	0.683	0.847	0.798	0.933	0.793
	ADASYN	0.58	**0.943**	0.532	0.712	**0.869**	0.777	0.946	0.761
	SMOTENNN	0.677	0.915	0.481	**0.74**	0.848	**0.803**	0.94	**0.85**
	Borderline-SMOTE	0.557	0.936	0.577	0.714	0.868	0.8	0.939	0.689
	SMOTE-Tomek	0.574	0.898	0.53	0.657	0.834	0.754	0.94	0.746
	CGAN	0.721	0.89	0.377	0.724	0.805	0.737	0.932	0.623
	SMOTEENN_CGAN	**0.725**	0.904	**0.596**	0.733	0.788	0.745	**0.963**	0.746
G-mean	原始数据	0.425	0.828	0.309	0.523	0.742	0.706	0.882	0.183
	SMOTE	0.677	0.88	0.614	0.671	0.801	0.723	0.955	0.737
	ROS	0.676	0.884	0.638	0.676	**0.804**	0.724	0.951	0.748
	ADASYN	0.658	0.891	0.628	0.668	0.799	0.717	0.951	0.733
	SMOTENNN	0.689	0.884	0.611	0.669	0.803	0.707	0.956	0.746
	Borderline-SMOTE	0.667	0.891	0.6	0.662	0.799	0.711	0.95	0.717
	SMOTE-Tomek	0.676	0.88	0.611	0.67	0.801	0.722	0.955	0.737
	CGAN	0.695	0.887	0.625	0.642	0.785	0.711	0.958	**0.752**
	SMOTEENN_CGAN	**0.731**	**0.895**	**0.641**	**0.677**	0.792	**0.732**	**0.962**	0.745

表 4　与其他文献方法的性能对比

数据集	本文方法			文献[20]		文献[9]	
	F_1	AUC	G	F_1	AUC	F_1	G
$ecoli_1$	**0.79**	**0.91**	**0.9**	0.76	0.88	0.78	0.89
$yeast_1$	0.57	**0.71**	–	**0.59**	0.71	–	–
$yeast_3$	**0.42**	**0.75**	**0.73**	0.34	0.61	0.38	0.68
haber	**0.5**	0.6	0.64	0.49	**0.64**	0.49	**0.65**
newth	0.91	**0.96**	0.96	0.63	0.77	**0.94**	**0.99**
pima	0.66	0.74	**0.73**	–	–	**0.66**	0.73

3.4　SMOTEENN_CGAN方法适用性分析

所提出方法对CGAN进行改进,旨在解决将其作为过采样方法处理不平衡数据时易受限于少数类样本规模从而导致所生成样本质量欠佳的问题,因

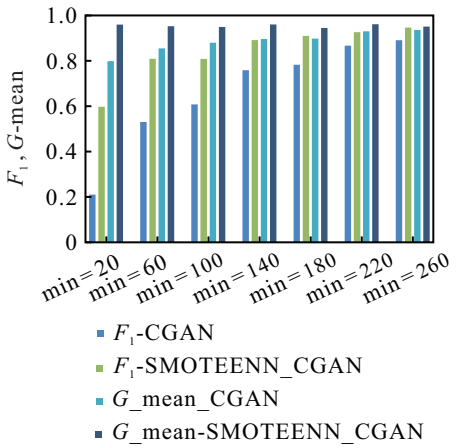

图 6　不同少数类样本规模下 F_1 和 G-mean

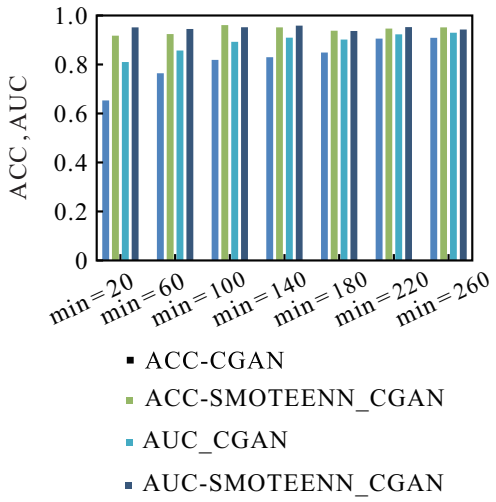

图7 不同少数类样本规模下ACC和AUC

此,不同的少数类样本规模对所提出方法的最终性能也会产生一定的影响,为了验证少数类样本规模对本文过采样方法性能的影响,在不同的少数类样本规模下测试CGAN和SMOTEENN_CGAN的性能变化,如图6和图7所示. 其中测试数据集使用第3.3节中的人工数据集,在保证多数类样本量为500的条件下,少数类样本由20每隔40取一次直至少数类样本数量为260.

由图6和图7可见,与CGAN相比,所提出方法在少数类样本为小样本下[23]优势更为突出,尤其在F_1综合评价指标上表现明显. 随着少数类样本规模的增加,CGAN能够从中学习更多的分布信息,生成的样本质量也在逐渐增加,与所提出方法获得的分类性能也在逐渐缩小.

4 结 论

针对CGAN作为过采样方法时易受限于少数类样本规模的问题,本文提出了一种混合CGAN与SMOTEENN的不平衡数据处理方法. 首先使用SMOTEENN快速合成少数类样本,使得少数类样本达到与多数类样本相当的规模,在此基础上训练CGAN以保证CGAN能够充分学习正类样本的分布特征,从而提高生成网络生成样本的质量. 在8个公开的不平衡分类数据集上,将所提出方法与其他7种采样方法进行对比,考察其F_1、Recall、AUC和G-mean四项不平衡分类性能指标值. 实验结果表明,SMOTEENN_CGAN方法处理后的不平衡数据集在各项分类性能指标上均表现出明显的优势. 通过与近期文献中报道的方法相比,所提出方法处理不平衡数据尤其是在少数类样本规模较小时更有优势. 最后对SMOTEENN_CGAN方法的适用性进行定量分析,验证不同规模的少数类样本对其分类性能的影响,实验结果表明,所提出方法在少数类样本为小样本[23]时更具优势. 然而,随着少数类样本规模和所反映分布信息的增加,运用SMOTEENN是否还有必要,虽然少数类样本规模较大时所提出方法亦能够微弱地提高分布学习效果,但是也会花销更多的学习时间,如何从中寻求一个合适的平衡点,将是下一步研究的重点.

参考文献(References)

[1] 许召召, 申德荣, 寇月, 等. 嵌入重采样技术的C4.5决策树集成分类算法的临床医学预测[J]. 控制与决策, 2021, 36(6): 1342-1350.
(Xu Z Z, Shen D R, Kou Y, et al. Clinical prediction of C4.5 decision tree classification algorithm with embedded resampling technique[J]. Control and Decision, 2021, 36(6): 1342-1350.)

[2] Cho S, Kim S, Choi J H. Transfer learning-based fault diagnosis under data deficiency[J]. Applied Sciences, 2020, 10(21): 7768.

[3] Somasundaram A, Reddy S. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance[J]. Neural Computing and Applications, 2019, 31(1): 3-14.

[4] 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(1): 321-357.

[5] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science, 2005, 3644(5): 878-887.

[6] He H B, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. IEEE International Joint Conference on Neural Networks. Hong Kong, 2008: 1322-1328.

[7] Barua S, Islam M M, Yao X, et al. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405-425.

[8] Batista G E, Bazzan A L C, Monard M C. Balancing training data for automated annotation of keywords: A case study[C]. In Proceedings of the 2nd Brazilian Workshop on Bioinformatics. Brazilian, 2003: 10-18.

[9] Cheng K, Zhang C, Yu H L, et al. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data[J]. IEEE Access, 2019, 7: 170668-170681.

[10] Li J N, Zhu Q S, Wu Q W, et al. SMOTE-NaN-DE: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and

differential evolution[J]. Knowledge-Based Systems, 2021, 223: 107056.

[11] 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.

[12] Huang K, Wu S, Li F, et al. Fault diagnosis of hydraulic systems based on deep learning model with multirate data samples[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, DOI: 10.1109/TNNLS.3083401.

[13] Jiang X Y, Ge Z Q. Augmented multidimensional convolutional neural network for industrial soft sensing[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

[14] 易令, 吕忠元, 丁进良, 等. 面向原油总氢物性预测的数据扩增预处理方法[J]. 控制与决策, 2018, 33(12): 2153-2160.
(Yi L, Lyu Z Y, Ding J L, et al. Data pretreatment approach for crude oil hydrogen properties prediction[J]. Control and Decision, 2018, 33(12): 2153-2160.)

[15] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[16] Fiore U, de Santis A, Perla F, et al. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection[J]. Information Sciences, 2019, 479: 448-455.

[17] Mao W T, Liu Y M, Ding L, et al. Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study[J]. IEEE Access, 2019, 7: 9515-9530.

[18] Mirza M, Osindero S. Conditional generative adversarial nets[J/OL]. Computer Science, 2014, arXiv: 1411.1784.

[19] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Systems with Applications, 2018, 91: 464-471.

[20] 赵海霞, 石洪波, 武建, 等. 基于条件生成对抗网络的不平衡学习研究[J]. 控制与决策, 2021, 36(3): 619-628.
(Zhao H X, Shi H B, Wu J, et al. Research on imbalanced learning based on conditional generative adversarial networks[J]. Control and Decision, 2021, 36(3): 619-628.)

[21] 严远亭, 戴涛, 张以文, 等. 邻域感知的不平衡数据集过采样方法[J]. 小型微型计算机系统, 2021, 42(7): 1360-1370.
(Yan Y T, Dai T, Zhang Y W, et al. Neighborhood-aware imbalanced oversampling[J]. Journal of Chinese Computer Systems, 2021, 42(7): 1360-1370.)

[22] 李伟. 生成对抗网络(GAN)模型优化方法研究[D]. 武汉: 武汉大学, 2019: 23-26.
(Li W. Generative adversarial network optimization strategy research[D]. Wuhan: Wuhan University, 2019: 23-26.)

[23] 程小红, 杨浩菊. 戈塞特及其小样本理论[J]. 西北大学学报: 自然科学版, 2015, 45(6): 1017-1019.
(Cheng X H, Yang H J. Gosset and his small-sample theory[J]. Journal of Northwest University: Natural Science Edition, 2015, 45(6): 1017-1019.)

作者简介

刘宁 (1996−), 男, 硕士生, 从事机器学习与数据挖掘的研究, E-mail: 365990458@qq.com;

朱波 (1978−), 男, 讲师, 博士, 从事智能制造、制造业信息化等研究, E-mail: zhubo20110720@163.com;

阴艳超 (1977−), 女, 教授, 博士生导师, 从事智能制造、知识工程等研究, E-mail: yinyc@163.com;

李岫宸 (1996−), 男, 硕士生, 从事智能制造、文本挖掘的研究, E-mail: 505441420@qq.com.