面向不平衡数据集的一种基于邻域的过采样算法

孟国庆[1],高 源[2],梅 颖[3,4],卢诚波[3,4]

(1. 浙江理工大学 计算机科学与技术学院,浙江 杭州 310018;2. 国网浙江省电力有限公司丽水供电公司,浙江 丽水 323050;3. 丽水学院 数学与计算机学院,浙江 丽水 323000;4. 浙江得图网络有限公司,浙江 丽水 310011)

摘 要: 过采样是一种通过合成新的同类样本解决数据集中类分布不平衡问题的常用方法。针对数据集中样本分布不平衡的问题,提出一种基于邻域概念的PSON算法。该算法定义每个少数类样本的影响力,依据不同影响力对少数类样本进行过采样以获得平衡数据集。在50个数据集上对8种过采样算法得到的数据集进行分类测试,通过威尔科克森符号秩检验比较7种分类性能指标,结果表明采用PSON算法后分类准确率提升显著。

关键词: 不平衡数据集;过采样;分类;逆近邻

DOI: 10.11907/rjdk.232015 开放科学(资源服务)标识码(OSID):

中图分类号: TP18 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0116-06

A Neighborhood-Based Over-Sampling Algorithm for Imbalanced Datasets

MENG Guoqing[1], GAO Yuan[2], MEI Ying[3,4], LU Chengbo[3,4]

(1. *School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*; 2. *State Grid Lishui Power Supply Company, Lishui* 323050, *China*; 3. *School of Mathematics and Computer, Lishui University, Lishui* 323000, *China*; 4. *Zhejiang Detu Network Co., Ltd, Lishui* 310011, *China*)

Abstract: Oversampling is a commonly used method to solve the problem of imbalanced class distribution in a dataset by synthesizing new samples of the same class. A PSON algorithm based on neighborhood concept is proposed to address the issue of imbalanced sample distribution in the dataset. This algorithm defines the influence of each minority class sample and oversamples the minority class samples based on different influences to obtain a balanced dataset. Classification tests were conducted on datasets obtained from 8 oversampling algorithms on 50 datasets. The Wilcoxon symbol rank test was used to compare 7 classification performance indicators, and the results showed that the use of PSON algorithm significantly improved classification accuracy.

Key Words: imbalanced dataset; over-sampling; classification; reverse neighbors

0 引言

类不平衡指的是在一个数据集中,不同类别先验概率不平衡,通常表现为某些类别的样本数量远少于其他类别[1]。在多类不平衡数据中,样本较多的类称为多数类,样本较少的类称为少数类。在二分类问题中,通常将样本数量较少的类别定义为正类(C^+),样本数量较多的类别定义为负类(C^-)[2]。类不平衡现象在许多领域中普遍存在,例如欺诈检测[3]、贷款风险预测[4]、故障检测[5]、医疗诊断[6]和文本分类[7]等。由于正类样本的数量较少,常规的学习模型往往会对负类样本产生更强的偏向性,将导致模型在分类时将一些正类样本误判为噪声或者负类,从而降低学习模型分类性能[8]。

在处理类不平衡数据的分类问题时,主要挑战为如何获得一个均衡的数据分布及如何训练出具有更强分类性能的模型。对类不平衡分类的解决方案通常分为:①数据层面。通过重新采样,即删除原有样本或合成(添加)新样

收稿日期: 2023-10-09

扫描二维码阅读全文:

基金项目: 国家自然科学基金项目(12171217);浙江省自然科学基金项目(LY18F030003)

作者简介: 孟国庆(1997-),男,浙江理工大学计算机科学与技术学院硕士研究生,研究方向为数据挖掘;高源(1977-),男,国网浙江省电力有限公司丽水供电公司高级工程师,研究方向为数据挖掘;梅颖(1977-),女,硕士,丽水学院数学与计算机学院副教授,研究方向为机器学习、人工智能;卢诚波(1977-),男,博士,丽水学院数学与计算机学院教授、硕士生导师,研究方向为数据挖掘、机器学习。本文通讯作者:卢诚波。

本减小类不平衡产生的负面影响，根据采样方式不同可分为过采样、欠采样以及过采样与欠采样组合的采样方法。②算法层面。通过优化现有分类模型或设计新的分类算法适应不平衡数据的分布情况，例如代价敏感学习[9]、内核学习[10]和特征选择[11]等。

数据层面的解决方案因独立于底层分类器而得到广泛应用[12]。SMOTE 算法（Synthetic Minority Oversampling Technique）因在过采样方面的有效性被广泛应用，成为国内外研究的热点，促使研究人员针对 SMOTE 算法的不足进行改进和优化，因此许多 SMOTE 改进算法被提出[13]。Han 等[14]提出一种 Borderline-SMOTE 算法，将少数类样本根据最近邻的多数类样本比例分为边界样本、安全样本和噪声，仅对边界样本进行插值来优化分类器的决策边界。He 等[15]提出 ADASYN（Adaptive Synthetic Sampling）方法，根据学习难度对不同少数类样本使用加权分布，然而更难学习的少数类样本需要合成的新样本数目更多。ADASYN 增加了边界样本合成的新样本密度，在一定程度上增强了决策边界。Safe-Level-SMOTE 算法为每个少数类样本分配一个安全系数，使新合成的样本更接近安全系数高的样本，从而确保新样本分布在安全区域，避免与多数类重叠[16]。陶叶辉等[17]提出一种基于高斯混合模型聚类的 SMOTE 过采样算法，利用 GMM 算法对少数类样本集进行聚类，再删除与聚类中心点重叠的冗余样本，最后根据不同聚类进行 SMOTE 过采样使数据平衡。针对传统聚类过采样算法的边界样本损失问题，樊东醒等[18]提出一种融合改进的 k 中心点算法的过采样算法 KmedlodSMOTE，引入聚类准则函数和边界阈值减少边界样本损失。

然而，上述方法并未解决数据集中样本分布不平衡的问题。为此，本文提出一种基于邻域概念的 PSON 算法（Prior Synthetic Oversampling based on Neighbors，PSON），对具有影响力的少数类样本进行过采样以实现类别分布均衡化，从而提升少数类样本分类准确率。首先，计算每个少数类样本的影响力因子以评估样本点对少数类样本分布影响力，该因子能有效应对样本分布密度不同的数据集，以更好区分数据集中影响力不同的样本点；其次，去除掉影响力较小的样本点，对欠采样数据集应用过采样算法。

1　相关工作

SMOTE 相较于通过对少数类进行随机复制达到类别平衡的随机过采样算法，不同之处在于：在特征空间上通过随机选择少数类样本的同类近邻样本进行插值，生成无重复的、新的少数类样本，从而有效缓解了由随机过采样引起的过拟合问题[13]。

在 SMOTE 算法中若选取的少数类样本及其相邻样本中存在噪声，则合成的新样本将可能成为噪声从而影响分类模型的准确性。Batista 等[19]使用 SMOTE 进行过采样后通过 ENN（Edited Nearest Neighbors）[20]或 Tomek Links[21]进行清理步骤。NRAS 算法（Noise Reduction A Priori Synthetic Oversampling）引入了先验合成过采样降噪技术，在合成新样本之前去除被认为是噪声的少数类样本[22]。Tomek 等[21]根据 NRAS 算法思想提出 k-IN（k-Influential Neighborhood）邻域作为识别噪声的过采样方法，k-IN 邻域可解释为一个区域在特征空间中包含目标样本与对该样本具有直接影响的 k 个最近邻及具有间接影响的逆向最近邻。k-NN(x)：在欧式距离空间中与样本点 x 最近的 k 个近邻。点 k-RNN(x)：在欧式距离空间中，如果样本点 x 为在样本点 s 的 k-NN(s)中，则样本点 s 为样本点 x 的逆近邻点。k-IN(x)：k-IN(x) = $\{x\} \cup k\text{-NN}\{x\} \cup k\text{-RNN}\{x\}$。修正 k-IN(x)：修正 k-IN(x)为 k-IN(x)中样本点 x 的同类样本点集合。由此可知，k-IONS 算法首先去除修正 k-IN 邻域中样本点个数小于 τ 的少数类样本，然后对处理后的少数类样本进行过采样来提升过采样算法的鲁棒性。

2　PSON 算法

k-INOS 算法引入了 k-IN 邻域的概念，以少数类样本为中心包含 k 近邻和逆近邻的少数类样本区域。k-IN 邻域中少数类样本的数量反映了少数类样本对数据分布的影响力，数量越大代表少数类样本越重要，越应该被采样。因此，k-INOS 算法在一定程度上降低了噪声对分类器性能，但从实验结果发现 k-INOS 的准确率显著下降，原因为该算法只选择具有影响力的少数类样本进行过采样，忽略了密度较低且靠近分类边界的少数类样本，导致分类器在学习过程中偏向于密度较高的区域，对密度较低的区域关注度下降，从而影响了分类器对边界区域少数类样本的识别能力[23]。此外，逆近邻表示相对密度，不同数据集对 k 值的选取较敏感，将影响 k-INOS 算法的泛化性能并在一定程度上限制了算法应用。

为了克服 k-INOS 算法的缺点，改善数据集类分布不均衡的问题，本文提出一种基于邻域概念的 PSON 过采样算法。首先考虑少数类样本逆近邻中同类样本点数量的变化程度，以克服采用逆近邻数量在描述影响力过程中不同密度数据集存在的缺陷，即算法分类性能会对选取的参数较为敏感。本文定义了影响因子以重新描述影响力：

$$IC(x_i) = \frac{|\{kNN(x_i)\} \cap \{RNN(x_i)\}|}{|\{kNN(x_i)\}|} \tag{1}$$

式中：x_i 为少数类样本点；样本点 x_i 的 k 近邻在全体数据集中计算求得；$\{kNN(x_i)\}$ 表示 k 近邻集中的同类样本点集合；$\{RNN(x_i)\}$ 表示逆近邻集中的同类样本点集合；$|\cdot|$ 表示集合中元素个数。

以图 1 为例说明 k-INOS 与 PSON 对数据集的影响差异，空心圆和实心圆分别表示少数类样本和多数类样本。当近邻值 k=3 时，7 个少数类样本点的 k 近邻、逆近邻、k-IN

邻域个数和IC值如表1所示，其中k近邻和逆近邻为同类样本点的集合。

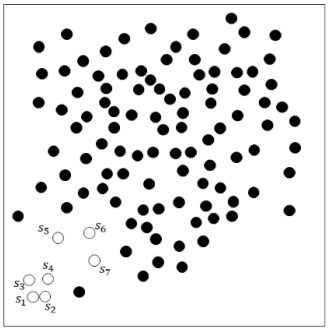

Fig. 1 Unbalanced dataset distribution

图1 不平衡数据集分布

Table 1 Number of k-nearest neighbors, inverse neighbors, k-IN neighborhoods, and IC values for each minority class sample

表1 各少数类样本的k近邻、逆近邻、k-IN邻域个数和IC值

少数类样本点	kNN	RNN	k-IN邻域个数	IC值
S_1	$\{S_2, S_3, S_4\}$	$\{S_2, S_3, S_4\}$	5	1
S_2	$\{S_1, S_3, S_4\}$	$\{S_1, S_3, S_4\}$	5	1
S_3	$\{S_1, S_2, S_4\}$	$\{S_1, S_2, S_4\}$	5	1
S_4	$\{S_1, S_2, S_3\}$	$\{S_1, S_2, S_3\}$	5	1
S_5	$\{S_6\}$	$\{S_6\}$	2	1
S_6	$\{S_5, S_7\}$	$\{S_5, S_7\}$	3	1
S_7	$\{S_6\}$	$\{S_6\}$	2	1

由表1可知，少数类样本点的k近邻和逆近邻中同类样本点个数不确定，相对密度不同的样本点的k-IN邻域个数方差较大，因此在不同数据集上选择统一的k-INOS算法参数容易去除密度相对稀疏且非噪声的样本点。PSON则采用k近邻与逆近邻的交集表示样本点对少数类样本点分布的影响力，相似的概念在文献[24]中表示影响空间。此外，PSON算法采用k近邻中同类样本个数作为分母，有助于降低相对密度不同的样本点对k值敏感性的影响，表1中密度相对密集的样本点s_1、s_2、s_3、s_4与密度相对稀疏的样本点s_5、s_6、s_7具有相同IC值。

为了进一步提升算法泛化性能，本文采用迭代搜索方式选出邻域参数，当所有少数类样本点的逆近邻中同类样本集合为空集或不在变化时停止搜索。当所有样本点都有逆近邻或逆近邻个数为0的所有样本不变时，近邻搜索达到自然稳定状态[25]。

算法1 PSON算法

输入：训练集$\boldsymbol{D}$，其中少数类样本为$\boldsymbol{D}_{min}$，参数ε。

输出：新的训练样本集$\boldsymbol{D}_{new}$。

1. 初始化搜索次数$r = 1$，$\mathrm{KNN}(x_i) = \varnothing$，$\mathrm{RNN} = \varnothing$，$x_i \in \boldsymbol{D}_{min}$。

2. $r = r + 1$。

3. 当所有的x_i的$\mathrm{RNN}(x_i) \neq \varnothing$或$\{x_i | \mathrm{RNN}(x_i) = \varnothing\}$不再变化时搜索停止，否则转至步骤2。

4. $k = r$，计算每个样本点$x_i \in \boldsymbol{D}_{min}$的$IC(x_i)$。

5. 去除掉$\mathrm{IC}(x_i) < \varepsilon$和$\mathrm{KNN}(x_i) = \varnothing$的少数类样本点，得到新的不平衡数据集$\boldsymbol{D}'$。

6. 使用过采样算法对不平衡数据集$\boldsymbol{D}'$进行过采样。

3 实验结果与分析

3.1 实验数据集

本文实验数据集为KEEL数据库、UCI数据库中50个二分类不平衡数据集，数据集属性全部为实数值，不平衡率（负类样本与正类样本个数的比例）范围为1.80~100.14。数据集详细描述如表2所示，Examples表示数据集样本数，Features表示特征数，IR为不平衡率。

3.2 评估指标

通常情况下，在不平衡数据集分类性能评价指标中使用几何平均值（G-mean）、正类检验值（F-measure）和AUC（Area Under the Curve）作为评估标准。其中，G-mean衡量分类器对正、负类样本分类的平衡性能，如式（2）所示；F-measure综合反映分类器对负类样本的分类性能，如式（3）表示；AUC通过ROC曲线评判模型分类效果的量化标准，点坐标由横坐标FPR（表示负类样本分类错误概率）和纵坐标TPR（表示正类样本分类正确概率）构成，FPR、TPR计算公式如式（4）、式（5）所示。为了更全面评价本文算法分类性能，还选用了Accuracy、Recall、Precision和Specificity作为辅助评价指标。

$$\mathrm{G-mean} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}} \tag{2}$$

$$\mathrm{F-measure} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

上述评价标准建立在混淆矩阵上，如表3所示。其中，TP表示实际为正类被预测为正类的样本数；FN表示实际为正类被预测为负类的样本数；FP表示实际为负类被预测为正类的样本数；TN表示实际为负类被预测为负类的样本数。

3.3 分类器与过采样算法

本文为了研究算法的普适性，分别以支持向量机（SVM）、决策树（DT）、k近邻（kNN）、Adaboost和随机森林（RF）作为分类模型，采用开源的基于Python语言的机器学习工具包sklearn[26]。其中，k近邻算法中k=5，决策树算法中树的最大深度为3，此外其他参数均为默认参数。

为了进一步验证本文算法的有效性，共选取了8种过采样算法进行比较，分别为ADASYN[15]、Borderline-SMOTE（BL_1，BL_2）[14]、MWMOTE[27]、RandomOverSampling（ROS）、Safe-Level-SMOTE（SL_SMOTE）[16]、SMOTE[13]和RWO[28]。同时，为确保比较算法性能的一致性，参数设置与原论文参数保持相同，即ADASYN、BL_1、BL_2、Safe-Level-SMOTE和SMOTE的k近邻参数设置为

Table 2　Unbalanced dataset

表 2　不平衡数据集

Name	Examples	Features	IR
cylinder-bands	277	40	1.8
tic-tac-toe	958	10	1.89
cmc_0_1	962	10	1.89
breast-cancer	277	10	2.42
haberman	306	4	2.78
ecoli1	336	8	3.36
hepatitis	80	20	5.15
ecoli2	336	8	5.46
ecoli	336	8	8.6
ecoli3	336	8	8.6
yeast-2_vs_4	514	9	9.08
ecoli-0-6-7_vs_3-5	222	8	9.09
ecoli-0-2-3-4_vs_5	202	8	9.1
yeast-0-3-5-9_vs_7-8	506	9	9.12
yeast-0-2-5-6_vs_3-7-8-9	1 004	9	9.14
yeast-0-2-5-7-9_vs_3-6-8	1 004	9	9.14
ecoli-0-1_vs_2-3-5	244	8	9.17
ecoli-0-2-6-7_vs_3-5	224	8	9.18
ecoli-0-3-4-6_vs_5	205	8	9.25
yeast-0-5-6-7-9_vs_4	528	9	9.35
ecoli-0-6-7_vs_5	220	7	10
ecoli-0-1-4-7_vs_2-3-5-6	336	8	10.59
spectrometer	531	94	10.8
glass-0-1-4-6_vs_2	205	10	11.06
glass2	214	10	11.59
us_crime	1 994	101	12.29
yeast_ml8	2 417	104	12.58
scene	2 407	295	12.6
cleveland-0_vs_4	177	14	12.62
libras_move	360	91	14
ecoli4	336	8	15.8
abalone9-18	731	9	16.4
solar_flare_m0	1 389	33	19.43
oil	937	50	21.85
yeast-2_vs_8	482	9	23.1
wine_quality	4 898	12	25.77
yeast4	1 484	9	28.1
yeast_me2	1 484	9	28.1
yeast-1-2-8-9_vs_7	947	9	30.57
yeast5	1 484	9	32.73
abalone-21_vs_8	581	9	40.5
yeast6	1 484	9	41.4
mammography	11 183	7	42.01
abalone-19_vs_10-11-12-13	1 622	9	49.69
kr-vs-k-zero_vs_eight	1 460	7	53.07
poker-8-9_vs_6	1 485	11	58.4
shuttle-2_vs_5	3 316	10	66.67
kddcup-land_vs_satan	1 610	42	75.67
kddcup-land_vs_satan	1 610	42	75.67
kddcup-rootkitimap_vs_back	2 225	42	100.14

Table 3　Confusion matrix for binary classification problems

表 3　二分类问题混淆矩阵

类别	预测为正类	预测为负类
正类	TP	FN
负类	FP	FN

5;MWMOTE 中 k_1=5,k_2=3,k_3=5;k-INOS 算法中 k 近邻参数设置为 11,t=3。实验中,采用五折交叉验证方法将所有数据集分为训练集和测试集,依据威尔科克森符号秩检验算法结果是否存在显著差异,显著性水平设置为 0.05[29]。

3.4　结果分析

本文通过两组实验对 PSON 算法进行评价,第一组实验比较使用 SON 算法和不使用 PSON 算法的过采样算法性能,第二组实验将 PSON 算法与 k-INOS 算法进行比较。参数设置方面,PSON 算法在两组实验中 ε 取值在 $(1/k, 2/k)$ 区间时性能最好,k 为近邻参数并在算法中计算求得。

表 4、表 5 为各算法与未使用 PSON 算法和使用 k-INOS 算法的 8 种过采样算法在 7 个性能指标上的平均数。其中,“▲”表示在均值上 PSON 算法显著提升;“△”表示在均值上 PSON 算法存在提高但并不显著;“-”表示在均值上 PSON 算法既没有提高也没有下降;“▼”表示在均值上 PSON 算法显著下降;“▽”表示在均值上 PSON 算法下降但并不显著。由表 5 可知,在 Accuracy、G-mean、Specificity 指标与 5 种分类器的组合后,PSON 算法显著提升的个数占比高达 100%。在 AUC、F-measure、Recall、Precision 性能指标与 5 种分类器的组合后,PSON 算法显著提高的个数占比达到 95% 以上。实验表明,PSON 算法相较于 k-INOS 算法的分类性能更优。

表 6 展示了过采样算法在 7 种性能指标中的情况。在分类器和过采样算法组合中 PSON 算法性能指标 Accuracy 显著提升的个数占比为 77.5%(31/40)。相较于 ADASYN、MWMOTE 算法,PSON 算法显著提高个数为 2,但相较于其他过采样算法较少。总体而言,在 Accuracy 上 PSON 算法性能提升显著;在 AUC 上 PSON 算法显著提升的个数占比与 Accuracy 相同,但在 SVM 中显著提高的个数只有 2 个;在 F-measure 上 PSON 算法显著提升的个数占比为 80%(32/40),仅与 RWO 算法在 Adaboost 分类器的比较中显著下降;在 Recall 上 PSON 算法显著提升的占比为 65%(26/40),相较于 ADASYN、MWMOTE 算法分别在 Adaboost 和 SVM 上显著下降;在 Precision 上 PSON 算法显著提升的个数占比为 62.5%(25/40),但在 SVM 分类器上显著提升的个数为 1,虽然在 Adaboost 分类器上 PSON 算法性能有所提升,但只相较于 ROS 提升显著;在 G-mean 上 PSON 算法显著提升的个数占比为 77.50%(31/40),在 Adaboost 分类器上相较于 8 个过采样算法而言,PSON 算法显著提升的个数为 3;在 Specificity 上 PSON 算法显著提升的个数占比为 87.5%(35/40)。综上,PSON 算法在 7 种分类性能指标中均有所提升,在 280 个性能指标比较个数中显著提升的个数占比为 75.36%(211/280),证明了 PSON 算法的有效性。为

Table 4 Comparison between various algorithms and oversampling algorithms without using k − INOS

表 4 各算法与未使用 k–INOS 的过采样算法比较

Name	ADA	BDL-1	BDL-2	MW	ROS	RWO	SL	SMOTE
Accuracy								
Adaboost	△	▲	▲	–	▲	▲	▲	▲
5-NN	△	▲	▲	▲	▲	△	▲	▲
RF	▲	▲	▲	△	▲	▲	▲	▲
SVM	▽	△	△	–	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
AUC								
Adaboost	▲	▲	▲	▲	▲	△	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	△	▲
RF	△	▲	▲	▲	▲	▲	▲	▲
SVM	△	△	△	–	▲	△	▲	△
DT	▲	▲	▲	▲	▲	▲	▲	▲
F-measure								
Adaboost	▲	▲	▲	▲	▲	▼	▲	▲
5-NN	△	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	△	▲	△	▽	–	▽	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
Recall								
Adaboost	▼	△	△	▼	▲	△	△	△
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	△	▲	▲	▲	△	▲	▲	▲
SVM	▼	▲	△	▼	▲	△	△	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
Precision								
Adaboost	△	△	△	△	▲	△	▲	▲
5-NN	▲	▲	▲	▲	▲	△	▲	▲
RF	▲	△	▲	▲	▲	▲	▲	▲
SVM	▼	△	△	△	▲	▼	△	△
DT	▲	▲	▲	▲	▲	▲	▲	△
G-mean								
Adaboost	–	△	△	▽	▲	△	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	△	▲	▲	▲	△	△	▲	△
DT	▲	▲	▲	▲	▲	▲	▲	▲
Specificity								
Adaboost	△	▲	▲	▲	▲	▽	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	△	▲	▲
SVM	△	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	△	▲	▲

Table 5 Comparison between various algorithms and the k − INOS algorithm

表 5 各算法与 k–INOS 算法比较

Name	ADA	BDL-1	BDL-2	MW	ROS	RWO	SL	SMOTE
Accuracy								
Adaboost	▲	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
AUC								
Adaboost	△	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
F-measure								
Adaboost	△	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
Recall								
Adaboost	△	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
Precision								
Adaboost	▼	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
G-mean								
Adaboost	▲	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲
Specificity								
Adaboost	▲	▲	▲	▲	▲	▲	▲	▲
5-NN	▲	▲	▲	▲	▲	▲	▲	▲
RF	▲	▲	▲	▲	▲	▲	▲	▲
SVM	▲	▲	▲	▲	▲	▲	▲	▲
DT	▲	▲	▲	▲	▲	▲	▲	▲

进一步对表 4 进行阐述，本文还统计出各分类器和过采样算法组合中分类性能指标显著提高的个数，如表 7 所示。由此可知，在 8 种过采样算法与 5 种分类器的 40 个组合中，有 7 种分类性能指标显著提升的个数≥4 的个数有 33 个，但在 ADASYN 与 SVM 的组合中 7 种性能指标都均未得到显著提升。

4 结语

本文为了提升过采样算法在不同数据集上的准确性，提出一种基于邻域概念的混合过采样算法 PSON 算法。通过采样具有影响力的样本促进样本分布平衡，相较于 k–

Table 6　Situation of oversampling algorithm in 7 performance indicators

表6　过采样算法在7种性能指标中的情况

index	▲	△	-	▼	▽
Accuracy	31(77.50%)	6(15.00%)	2(5.00%)	0(0.00%)	1(2.50%)
AUC	31(77.50%)	8(20.00%)	1(2.50%)	0(0.00%)	0(0.00%)
F-measure	32(80.00%)	4(10.00%)	1(2.50%)	1(2.50%)	2(5.00%)
Recall	26(65.00%)	10(25.00%)	0(0.00%)	4(10.00%)	0(0.00%)
Precision	25(62.50%)	13(32.50%)	0(0.00%)	2(5.00%)	0(0.00%)
G-Mean	31(77.50%)	6(15.00%)	1(2.50%)	0(0.00%)	2(5.00%)
Specificity	35(87.50%)	4(10.00%)	0(0.00%)	0(0.00%)	1(2.50%)
Total	211(75.36%)	51(18.21%)	5(1.79%)	7(2.500%)	6(2.14%)

Table 7　The number of performance indicators improved after combining various classifiers and oversampling algorithms

表7　各分类器和过采样算法组合后性能指标提升个数

index	ADA	BDL-1	BDL-2	MW	ROS	RWO	SL	SMOTE
Adaboost	3	4	4	3	7	1	6	6
5-NN	5	7	7	7	7	5	6	7
RF	5	6	7	6	6	6	7	7
SVM	0	4	2	2	5	2	5	4
DT	7	7	7	7	7	6	7	6

INOS算法在提升分类准确率的同时减少了分类信息损失,还解决了近邻参数选取敏感的问题。

具体为,通过近邻参数k值搜索算法获取最佳的参数k,从而进一步提升PSON算法的泛化能力。通过多个数据集的实验表明,PSON算法优于其他比较的过采样算法。未来,将基于现有研究,将重采样方法扩展应用到不平衡数据多标签分类问题中,以提升算法的适用性。

参考文献:

[1] GUZMÁN-PONCE A, SÁNCHEZ J S, VALDOVINOS R M, et al. DBIG-US: a two-stage under-sampling algorithm to face the class imbalance problem[J]. Expert Systems with Applications, 2021, 168: 114301.

[2] LIU J. A minority oversampling approach for fault detection with heterogeneous imbalanced data[J]. Expert Systems with Applications, 2021,184: 115492.

[3] KHAN A T, CAO X, LI S, et al. Fraud detection in publicly traded US firms using beetle antennae search: a machine learning approach[J]. Expert Systems with Applications, 2022, 191: 116148.

[4] PARK M S, SON H, HYUN C, et al. Explainability of machine learning models for bankruptcy prediction[J]. IEEE Access, 2021, 9: 124887-124899.

[5] JAN S U, LEE Y D, KOO I S. A distributed sensor-fault detection and diagnosis framework using machine learning [J]. Information Sciences, 2021, 547: 777-796.

[6] LYU X C, WANG W H, LIU H F. Cluster-wise weighted NMF for hyperspectral images unmixing with imbalanced data [J]. Remote Sensing , 2021,13(2):1-19.

[7] BRUNI R, BIANCHI G. Website categorization: a formal approach and robustness analysis in the case of E-commerce detection [J]. Expert Systems with Applications, 2020, 142: 113001.

[8] GARCÍA V, SÁNCHEZ J S, MARQUÉS A I, et al. Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data [J]. Expert Systems with Applications, 2020, 158: 113026.

[9] REN Z J, ZHU Y S, KANG W, et al. Adaptive cost-sensitive learning: improving the convergence of intelligent diagnosis models under imbalanced data[J]. Knowledge-Based Systems, 2022, 241: 108296.

[10] YANG M P, WANG ZHE, LI Y Q, et al. Gravitation balanced multiple kernel learning for imbalanced classification [J]. Neural Computing and Applications, 2022, 34(16): 13807-1382.

[11] KIM J, KANG J, SOHN M, et al. Ensemble learning-based filter-centric hybrid feature selection framework for high-dimensional imbalanced data[J]. Knowledge-Based Systems, 2021, 220: 106901.

[12] YE X C, LI G M, IMAKUR A, et al. An oversampling framework for imbalanced classification based on Laplacian eigenmaps[J]. Neurocomputing , 2020, 399: 107-116.

[13] 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.

[14] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing, 2005: 878-887.

[15] HE H, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning [C]// 2008 IEEE International Joint Conference on Neural Networks, 2008: 1322-1328.

[16] BUNKHUMPORNPAT C, SINAPIROMSARAN K, LURSINSAP C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009: 475-482.

[17] TAO Y H, ZHAO S W. Improved SMOTE algorithm based on gaussian mixture clustering for imbalanced data sets[J]. Software Guide, 2022, 21(5): 110-114.
陶叶辉, 赵寿为. 面向不平衡数据基于高斯混合聚类的SMOTE改进算法[J]. 软件导刊, 2022, 21(5): 110-114.

[18] FAN D X, YE C M. Credit unbalanced data classification based on clustering oversampling algorithm[J]. Software Guide, 2021, 20(11): 70-74.
樊东醒, 叶春明. 融合聚类过采样算法的信贷不平衡数据分类[J]. 软件导刊, 2021, 20(11): 70-74.

[19] BATISTA G E, 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.

[20] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1972 (3): 408-421.

[21] TOMEK I. Two modifications of CNN [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1976(11): 769-772.

[22] RIVERA W A. Noise reduction a priori synthetic over-sampling for class imbalanced data sets[J]. Information Sciences, 2017, 408: 146-161.

[23] DE MORAIS R F A B, VASCONCELOS G C. Boosting the performance of over-sampling algorithms through under-sampling the minority class [J]. Neurocomputing, 2019, 343: 3-18.

[24] VADAPALLI S, VALLURI S R, KARLAPALEM K. A simple yet effective data clustering algorithm[C]// 6th International Conference on Data Mining, 2006: 1108-1112.

[25] ZHU Q, FENG J, HUANG J. Natural neighbor: a self-adaptive neighborhood method without parameter K[J]. Pattern Recognition Letters, 2016, 80: 30-36.

[26] FERNÁNDEZ A, GARCÍA S, DEL JESUS M J, et al. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets[J]. Fuzzy Sets and Systems, 2008, 159 (18): 2378-2398.

[27] 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, 2012, 26(2): 405-425.

[28] ZHANG H, LI M. RWO-Sampling: a random walk over-sampling approach to imbalanced data classification[J]. Information Fusion, 2014, 20: 99-116.

[29] WOOLSON R F. Wilcoxon signed-rank test [M]. New Jersey: John Wiley & Sons, ltd, 2005.

(责任编辑:刘嘉文)