许思为, 周明, 邹瑞, 等. 不平衡数据集中采样比例对分类结果影响的研究[J]. 智能计算机与应用,2024,14(9):111-117.
DOI:10.20169/j.issn.2095-2163.240917

不平衡数据集中采样比例对分类结果影响的研究

许思为, 周 明, 邹 瑞, 刘吉华, 吴俊平, 秦雨露
(湖北大学 商学院, 武汉 430062)

摘 要: 各领域的发展伴随着大量不同类别数据的产生,数据集样本类别往往存在不平衡的特点,特别是医疗、金融和工业领域的数据集,以往研究专注于采样的方法和分类算法。本文针对不平衡数据集的分类问题,按原始比例抽取验证数据集,对余下数据根据不同采样比例和重采样技术构建训练数据集,运用多种分类算法,研究不同采样比例对分类结果的影响。实验结果表明,当采样比例接近原始比例时,分类器的少数类精确率表现更好;当采样比例接近平衡比例时,少数类召回率表现更佳;而最佳 F - Score 值出现在原始比例和平衡比例之间。本文为不同的应用需求提供了参考,对少数类精确率要求比较高时,使用原始数据;对少数类召回率要求比较高时,通过采样,平衡数据集的不同类别。

关键词: 重采样; 不平衡数据集; 采样比例; 召回率; 精确率
中图分类号: TP311.13 **文献标志码**: A **文章编号**: 2095-2163(2024)09-0111-07

A study on optimizing sampling ratios for improved classification results in imbalanced datasets

XU Siwei, ZHOU Ming, ZOU Rui, LIU Jihua, WU Junping, QIN Yulu
(School of Business, Hubei University, Wuhan 430062, China)

Abstract: The development in various fields is accompanied by the generation of a large amount of diverse data, often exhibiting imbalances in sample class distribution. Previous research has primarily focused on sampling methods and classification algorithms to address the challenges of imbalanced datasets. In the context of classifying imbalanced datasets, this study involved extracting a validation dataset in proportion to the original distribution. The remaining data is used to construct training datasets with different sampling ratios, applying various classification algorithms to investigate the impact of these ratios on classification outcomes. Experimental results indicated that when the sampling ratio approaches the original distribution, classifiers demonstrate better precision for the minority class. Conversely, when the sampling ratio approaches a balanced distribution, superior recall for the minority class is observed. The optimal F-score value emerged between the original and balanced ratios. This study provided a insight for diverse application requirements: original data is recommended when demanding high precision for the minority class, while sampling to balance class distribution is suggested when prioritizing high recall for the minority class.

Key words: resampling; imbalanced datasets; sampling ratios; recall; precision

0 引 言

各领域中往往会面临对不平衡数据集分类的问题,比如在医疗领域需要对肿瘤诊断、金融领域需要对风险交易识别、工业领域中零件故障诊断等[1-3]。在分类预测中,数据中不同类别样本的数量是不均衡的,比如银行的交易数据,绝大多数交易都属于正常交易,只有极少数属于风险交易,样本类别不平衡,如何识别一次交易是否存在风险,称为不平衡数据的分类问题。传统的分类算法在平衡分布的数据集中表现良好,但基于经验风险最小化,在不平衡数据分布中,更倾向预测为多数类,而对少数类预测表现不佳。

基于不平衡数据集和传统分类算法的特点,学

作者简介: 许思为(1999-),男,硕士研究生,主要研究方向:数据挖掘,机器学习; 邹 瑞(1999-),女,硕士研究生,主要研究方向:数据挖掘,数据分析; 刘吉华(1971-),男,博士,副教授,主要研究方向:神经网络,数据挖掘; 吴俊平(2001-),男,硕士研究生,主要研究方向:数据分析; 秦雨露(2001-),女,硕士研究生,主要研究方向:数据分析。
通讯作者: 周 明(1975-),男,博士,副教授,主要研究方向:服务营销。Email:dmz524@qq.com

收稿日期: 2024-03-04

者们从数据的预处理层面和分类算法层面进行了优化,数据预处理层面,主要是通过重采样技术降低数据倾斜度,这种方法不受分类器的影响,应用范围广;分类算法层面主要采用基于代价敏感、集成学习等思想,通过惩罚机制对数据不平衡进行补偿,但并没有改变数据的特性,应用范围也有限,通常将这两种方法结合起来,提高模型的分类能力[4-5]。

为了将分类器适应于不平衡数据集,许多研究者提出先对数据集进行处理,使数据集达到平衡再进行分类,数据重采样是较为常用的方法,包括过采样、欠采样和混合采样。过采样的研究中,SMOTE(Synthetic Minority Over-sampling Technique)是一种流行的方法,采用最近邻预先设定采样倍率,通过插值生成新的少数类,解决了随机过采样容易过拟合的缺点[6]。王博文[7]提出一种 SMOTENC(Synthetic Minority Over-sampling Technique for Nominal and Continuous)采样方法,基于 SMOTE 方法,在采样过程中加入了随机扰动,在驾驶人交通安全评估中获得了更精准的结果;Dong[8]为了解决 SMOTE 方法在样本稠密区域合成样本多于样本稀疏区域的问题,提出对每个少数类样本随机选择另外两个样本构成三角区,并根据采样倍率在区域内随机合成样本;谢子鹏等[9]使用聚类算法代替随机过采样的方法,提出一种 EM(Expectation-Maximization)聚类过采样算法,在少数类样本中进行聚类,把聚类中心作为过采样的样本点,避免了盲目采样的问题。在欠采样的研究中,有学者提出了 Tomek Links 和 NCL(Neighborhood Cleaning Rule)等方法。Beckmann[10]提出 K 近邻采样,对每个多数类样本进行 K 近邻计算,删除邻居计数大于阈值的多数类样本;Hido[11]基于 Bagging(Bootstrap Aggregating)的性质提出一种新的采样方法 RB-Bagging(Re-Balanced Bagging),改进不平衡分布的数据集,且充分利用了所有少数实例的欠抽样,在评价指标 AUC(Area Under the Curve)和 ISE(Integrated Squared Error)上表现良好;Lin[12]提出聚类欠采样的方法,将多数类聚类数量设置为少数类的量,使用聚类中心或中心最近邻作为代表;Fu[13]提出利用 PCA(Principal Component Analysis)算法对数据集进行降维,根据在每个主成分到少数类样本重心的距离对多数类样本进行排序和筛选;孟东霞[14]提出了一种基于特征边界信息进行欠采样的数据处理方法,先删除区域内的噪声点并筛选优质样本点,再采用欠采样的方法处理边界多数类样本,保留优质样本,以减少信息的流失。在混合采样中,Li[15]提出一种将过拟合和欠拟合结合的方法,利用 K-outline 方法将数据分为边界样本和非边界样本,对非边界进行欠采样,对边界进行过采样,实验表明分类效果有提高。

分类算法的研究主要有基于代价敏感的方法和集成学习的方法。Elkan[16]指出代价敏感学习的重要性,其基本思想是给少数类分配高额误分代价,而给多数类分配较小的误分代价,从而提高了少数类样本的重要性,减轻了分类器对多数类的偏好;周传华等[17]针对不平衡数据集中少数类样本分类识别率较低的问题,提出一种基于代价敏感卷积神经网络(CSCNN),设置特定的代价敏感指标来协同卷积神经网络的交叉熵损失函数,构建 CSCNN 神经网络;平瑞[18]提出基于聚类的弱平衡代价敏感随机森林算法,用于不平衡数据的分类,选择误分类代价下降值最大的属性划分,可以有效提高单棵决策树的性能;Tao[19]提出一种基于自适应成本权重的支持向量机代价敏感集成方法,使用成本敏感的支持向量机作为基分类器,并使用改进的成本敏感的 Boosting 方法,有利于最终分类边界略微偏离少数类。在集成方法中,王萌铎[20]提出一种基于 AdaBoost(Adaptive Boosting)集成加权宽度学习系统的不平衡数据分类方法,赋予样本和弱分类器权重,并在迭代过程中不断更新,提升模型对少数类的识别能力。

不平衡数据集问题的类别不平衡程度越高,其分类难度越大。学者们对采样方法和机器学习方法都进行了深入研究,但是在研究过程中大多都将不平衡数据集采样达到 1∶1 的水平,很少会考虑到不同比例的不平衡数据对分类结果的影响。本研究关注的问题是不同的采样比例是否会影响模型分类性能,以及如何选择适当的采样比例来提高不同应用场景下决策的有效性;通过对不平衡数据集进行适当的采样,构成不同的不平衡比例,采用不同的分类算法建模,观察不平衡比例和分类算法对预测效果的影响。

本研究提供一个了解采样比例对分类结果影响的视角,更好地解决了不平衡数据集分类问题。通过全面的实验和数据分析,探讨采样比例与不平衡数据集之间的关系,为未来研究提供参考,希望揭示采样比例选择的最佳应用实践,从而提高各个领域的决策支持和分类性能。

1 采样技术

针对不平衡数据,可以通过对样本的重采样实现数据在一定比例上的分布,数据集的重采样方法可分

为欠采样(减少多数类)和过采样(增加少数类)。

1.1　欠采样

欠采样技术的目的是减少多数类别的样本数量以平衡数据集,其主要方法是随机欠采样和基于近邻的欠采样。随机欠采样是指在多数类中进行不放回的随机抽样,使多数类数量与少数类达到一定比例平衡,再与少数类样本结合形成新数据集。由于多数类样本中包含的信息较多,采用欠采样的方法可能导致信息丢失较多,影响分类学习过程,本文考虑了不同的采样比例,尽可能保留多数类样本的信息。

1.2　过采样

过采样技术旨在增加少数类别样本以平衡不平衡数据集,其主要方法有 SMOTE,是一种合成少数类的过采样方法,其对传统随机过采样的方案进行了改进,有效避免了数据集中少数类样本大量重复出现,进而降低了模型过拟合的风险。SMOTE 采样的基本思路:计算少数类样本之间的距离,选择邻近的少数类样本进行插值,得到新样本,将人工模拟得到的新样本加入到原数据集中,使原始数据不再严重失衡。合成的新样本 x_{new} 为:

$$x_{new} = x_i + \text{rand}(0,1) \times (x_n - x_i) \quad (1)$$

其中,x_i 为一个少数类样本;x_n 为 x_i 的邻近样本;rand(0,1) 表示 0 和 1 之间的一个随机数。

2　分类算法

2.1　逻辑回归(Logistic)

逻辑回归是一种广义的线性模型,在机器学习中通常作为一种分类模型,常用于解决二分类问题。逻辑回归假设 y 服从伯努利分布,实际是使用线性回归模型的预测值映射到分类任务真实标记的对数几率,通过逻辑函数(Sigmoid)引入非线性因素,因此可以处理 0/1 分类问题。

2.1.1　K 近邻算法

K 近邻算法也是一种用于分类问题的算法,在对某一未知样本进行分类时,以全部训练样本为代表点,计算未知样本与所有训练样本之间的距离,取 K 个距离最近的样本的类别,作为划分未知样本类别的依据,通常有多数投票法和距离加权法。K 近邻算法的关键在于确定距离计算的方法以及 K 的值,K 值较小,模型容易过拟合,K 值较大,又容易学习到噪声,选择不同的 K 值可能会得到不同的分类结果。

2.1.2　决策树

决策树是一种机器学习的分类方法,一颗决策树包括一个根结点,若干内部节点和子节点,其目的是将一个具有 P 维特征的样本分到对应的类别中。构建决策树的过程:特征选择,一般选取分类能力强的特征;决策树生成,通常用信息增益、信息增益率、基尼指数来划分特征选择生成决策树;决策树剪枝,决策树对训练样本有精确的分类效果,但在测试中容易出现过拟合,所以需要剪枝处理。决策树算法把分类的过程表示成一棵树,每次通过选择一个特征进行分叉,最后得到分类结果。

2.1.3　随机森林

随机森林是通过 bagging 思想将多棵决策树集成在一起的一种分类算法,其核心思想是将一个待分类样本输入到每棵决策树中进行分类,将若干弱分类器的结果投票选择,得到样本的最终类别。在每棵树生成规则中,在训练集为 N 的样本中有放回的抽取 N 个训练样本,在 M 个特征中随机选择 m 个来构建决策树($m < M$)。在形成的随机森林中,任意两颗树相关性越大,分类错误率越高;每棵树分类能力越强,森林分类错误率越低。

2.1.4　极致梯度提升树

极致梯度提升树(XGBoost, Extreme Gradient Boosting Trees)属于 Boosting 方法中的一员,是一个基于 Boosting 增强策略的加法模型,在训练时采用前向分布算法进行贪婪学习,其基分类器支持 CART(Classification and Regression Tree)决策树,第 t 次迭代会学习一棵 CART 树,来拟合前面 $t-1$ 棵树预测结果与训练样本真实值的残差,集成为一个强分类器,与梯度提升模型相比提升了模型训练速度和预测精度。

2.2　评价指标

分类模型中,通常会根据混淆矩阵来计算评价指标,一个二分类的混淆矩阵见表 1,1 代表少数类,0 代表多数类,*TP* 表示将 1 类正确预测为 1 类、*FP* 表示将 0 类错误预测为 1 类、*FN* 表示将 1 类错误预测为 0 类、*TN* 表示将 0 类正确预测为 0 类。

表 1　混淆矩阵

Table 1　Confusion Matrix

预测值	真实值	
	1	0
1	*TP*	*FP*
0	*FN*	*TN*

评价分类模型的指标通常有准确率、精确率、召回率、F - Score 等。由于不平衡数据集的样本类别极不平衡,分类的准确率并不适用,许多应用场景所关注的指标是少数类的精确率(*precision*) 和召回

率(*recall*),分别描述了预测为少数类的样本中真实为少数类的情况,和真实为少数类样本中预测为少数类的情况:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

F - Score 是精确率和召回率的加权调和平均值,用于衡量分类模型的性能,尤其适用于不平衡数据集的评估。F - Score 的计算公式如下:

$$F - \text{Score} = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \tag{4}$$

当$\beta = 1$时,适用于对精确率和召回率同等重视的情况;当$\beta > 1$时,更重视召回率;当$\beta < 1$时,更重视精确率。

在风险识别中,不同类型的分类错误所带来的代价是极为重要的,因为这些错误可能导致实际的经济损失或其他风险。分类错误通常分为两种情况:将正常交易错误地识别为风险交易(假阳性),这种情况下,正常交易被错误地标记为风险交易,可能导致客户不便、额外的审核步骤以及潜在的客户不信任,但这通常不会带来直接的经济损失;将风险交易错误地识别为正常交易(假阴性),这种情况下,实际上的风险交易被错误地标记为正常交易,可能导致实际的经济损失,机构需要承担资金损失、风险暴露、信誉风险等后果。在交易中会更关注第二种情况,即将风险交易错误分类为正常交易的情况,因此更关注的指标通常是召回率。本文希望尽可能减少未能识别出的风险交易(假阴性)数量,即增加召回率,意味着更多的欺诈交易被正确地识别出来,降低了企业面临的实际经济损失和风险,可将β设置为大于 1 的值。

因此本文采用少数类的精确率、召回率和 F - Score($\beta = 2$)作为实验的评价指标,并观察其不同比例下的变化。

3 不平衡数据集的分类实验

3.1 数据集介绍

本文采用的是阿里天池 2018 年大数据竞赛-风险识别算法赛所提供的数据,这个数据集包含某行业对日常交易明细进行抽样审核的信息,在这个数据集中,被认为存在潜在风险的交易,都会打上一个风险标识。该数据集中包含 100 000 个样本,其中每个样本有 33 个字段,ID 代表交易 ID 标识符;V_time 表示时许;V1-V30 表示交易变量(已经过脱敏处理);Lable=(0,1),其中 0 代表此次交易无风险,1 代表此次交易有风险。该数据集中包含无风险交易样本 99 700 个,风险交易样本仅为 300 个,符合不平衡数据集的分布特征,适合本文对不平衡重采样比例问题的研究。

3.2 实验过程

首先,对数据进行预处理,删除样本无关的特征,并对特征做标准化处理;其次,对数据集进行 k 折交叉,划分为训练集和验证集;按照不同的重采样方法,并设置不同的采样比例水平,对训练集进行重采样,形成新的训练集;最后,对新的数据集应用不同的分类模型进行训练,并在不平衡的测试集中进行测试,记录其评价结果。

数据预处理。根据数据集的描述,ID 和 V_time 字段对本研究的作用不大,故应该删除;V1-V30 对应交易中的变量,可认为是特征,需要做标准化处理;Lable 为标签,0 代表多数类样本,1 代表少数类样本,不需要额外处理。

训练测试集划分。采用交叉划分是为了将样本中每个样本都可以用于训练和预测,保证实验的可靠性。本实验采用 4 折交叉将新样本集按标签类别划分为训练集:测试集为 3∶1。

数据重采样。本研究采用多种重采样技术对数据进行重采样,包括随机欠采样和基于 SMOTE 的过采样。在采样中设置为无风险样本(多数类):有风险样本(少数类)的采样比例分别为 1∶1、5∶1、10∶1、20∶1、30∶1、60∶1、90∶1、120∶1、150∶1、200∶1、250∶1、300∶1,最后设置一个原始比例的对照组,用来观察采样对分类结果的影响。

模型训练和预测。本实验采用不同的分类器对 12 组不同测试集进行拟合,包括逻辑回归、K 近邻、决策树、随机森林、极致梯度提升方法,这些分类器基于 python 开源库中的 scikit-learn 来构建,然后对拟合模型对测试集进行预测,并记录观测指标,观察不同采样水平和分类方法对预测效果的影响。

在本实验中重采样过程存在随机性,为了减少这种随机性,本实验对不同采样水平,进行重复采样 5 次,并对评价结果取平均值,增加实验结果的可信性。

3.3 结果分析

实验测试了包括随机欠采样和基于 SMOTE 的过采样,并分别使用多种分类器对不同采样比例的数据集进行测试,针对每一采样方法和分类模型重复进行实验 5 次,最终结果取平均值,实验结果见表 2~表 7,比例为多数类:少数类,逻辑回归(LR)、K-

邻近(KNN)、决策树(DTC)、随机森林(RFC)、极致梯度树(XGBC)。

表 2　随机欠采样的精确率

Table 2　Precision of random undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.070 6	0.117 8	0.027 3	0.091 0	0.066 7
5∶1	0.247 1	0.395 9	0.085 0	0.465 5	0.336 4
10∶1	0.404 5	0.585 2	0.137 5	0.701 2	0.513 5
20∶1	0.579 3	0.740 5	0.228 0	0.762 9	0.746 1
30∶1	0.665 2	0.762 4	0.293 0	0.820 6	0.782 1
60∶1	0.765 3	0.844 5	0.420 4	0.854 7	0.841 9
90∶1	0.824 6	0.869 0	0.510 7	0.861 1	0.861 7
120∶1	0.830 9	0.901 4	0.565 0	0.885 0	0.882 0
150∶1	0.844 4	0.917 0	0.630 0	0.895 7	0.895 1
200∶1	0.865 1	0.924 9	0.667 4	0.913 6	0.911 8
250∶1	0.872 6	0.936 3	0.714 1	0.925 4	0.919 4
300∶1	0.878 4	0.940 0	0.747 3	0.934 5	0.931 6
对照	0.887 3	0.943 3	0.759 4	0.934 2	0.937 5

表 2 所示，在随机欠采样中表现出相似的结果，越接近于数据的原始比例或者在原始比例附近表现出最好的精确率。通过不同的采样方法和分类模型，均能得到采样在原始比例附近时精确率表现会更好的结论。

表 3　SMOTE 采样的精确率

Table 3　Precision of SMOTE undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.101 4	0.458 3	0.399 3	0.889 9	0.788 3
5∶1	0.317 4	0.501 0	0.415 6	0.902 3	0.839 0
10∶1	0.490 4	0.546 7	0.434 5	0.900 6	0.873 1
20∶1	0.684 4	0.601 6	0.465 9	0.906 5	0.889 6
30∶1	0.749 7	0.638 8	0.501 5	0.908 5	0.907 8
60∶1	0.827 1	0.724 1	0.584 9	0.920 6	0.918 3
90∶1	0.845 4	0.787 7	0.608 4	0.923 3	0.914 8
120∶1	0.858 8	0.825 2	0.639 6	0.924 7	0.928 9
150∶1	0.863 8	0.856 9	0.655 4	0.927 9	0.927 6
200∶1	0.867 6	0.897 6	0.707 1	0.931 4	0.930 3
250∶1	0.872 8	0.925 4	0.712 4	0.935 8	0.929 4
300∶1	0.877 9	0.941 4	0.739 4	0.932 9	0.934 0
对照	0.887 3	0.943 3	0.759 4	0.934 2	0.937 5

表 3 所示，在 SMOTE 采样中各个分类器在精确率的表现，从采样比例 1∶1 到 300∶1，精确率不断上升，也就是说越接近原始数据的比例，其精确率的表现越好；反而通过采样平衡后的数据集，可能会降低精确率，例如 SMOTE 的 KNN 分类器，在采样后多数类：少数类为 1∶1 时，精确率为 0.458 3，而在 300∶1 时，却高达 0.941 4。

表 4　随机欠采样的召回率

Table 4　Recall of random undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.921 3	0.881 3	0.906 7	0.905 3	0.920 7
5∶1	0.878 7	0.854 0	0.861 3	0.853 3	0.870 7
10∶1	0.860 7	0.850 7	0.848 0	0.846 7	0.856 0
20∶1	0.844 7	0.846 0	0.842 0	0.844 7	0.842 0
30∶1	0.826 0	0.836 0	0.825 3	0.840 0	0.838 7
60∶1	0.807 3	0.813 3	0.813 3	0.834 0	0.825 3
90∶1	0.790 7	0.808 7	0.803 3	0.827 3	0.818 7
120∶1	0.783 3	0.800 0	0.794 0	0.822 0	0.813 3
150∶1	0.770 7	0.794 7	0.790 0	0.824 0	0.809 3
200∶1	0.748 7	0.792 0	0.766 7	0.814 0	0.806 0
250∶1	0.718 0	0.791 3	0.772 7	0.807 3	0.803 3
300∶1	0.700 7	0.788 0	0.760 7	0.803 3	0.800 0
对照	0.676 7	0.786 7	0.744 7	0.798 0	0.803 3

表 4 所示，在随机欠采样中各分类器的表现，从采样比例 300∶1 到 1∶1，召回率是不断上升的，也就是说通过 SMOTE 采样在数据集类别比较平衡时，召回率表现越好，在接近原始数据比例时，召回率相对会比较低。例如在逻辑回归中，采样率在 300∶1 时，召回率只有 0.676 7；而在采样率为 1∶1 时，召回率高达 0.921 3。

表 5　SMOTE 采样的召回率

Table 5　Recall of SMOTE undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.916 7	0.840 0	0.753 3	0.806 7	0.829 3
5∶1	0.862 7	0.840 0	0.790 0	0.824 7	0.826 7
10∶1	0.855 3	0.836 0	0.790 7	0.818 7	0.825 3
20∶1	0.834 7	0.831 3	0.780 7	0.820 0	0.820 7
30∶1	0.828 0	0.830 7	0.783 3	0.818 0	0.818 7
60∶1	0.806 7	0.828 7	0.774 7	0.818 0	0.817 3
90∶1	0.789 3	0.820 0	0.770 7	0.810 7	0.800 0
120∶1	0.782 0	0.813 3	0.759 3	0.811 3	0.807 3
150∶1	0.769 3	0.805 3	0.772 0	0.808 0	0.811 3
200∶1	0.749 3	0.798 7	0.760 7	0.807 3	0.798 0
250∶1	0.728 0	0.792 7	0.762 0	0.802 7	0.800 0
300∶1	0.694 7	0.788 7	0.753 3	0.799 3	0.794 7
对照	0.676 7	0.786 7	0.744 7	0.798 0	0.803 3

表 5 所示，在 SMOTE 采样中，LR、KNN、XGBC

模型表现出相似的结果,数据集在趋于平衡时,表现出最好的召回率。但这一结论在决策树和随机森林模型中要谨慎使用,在 SMOTE 采样中,决策树中召回率表现最好是在采样率为 10∶1,而随机森林出现在采样率为 5∶1,说明在数据采样后较为平衡时,召回率能得到显著的改善。

表 6 随机欠采样的 F_2 值

Table 6 F_2 of random undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.259 5	0.387 3	0.118 2	0.330 3	0.247 6
5∶1	0.578 8	0.704 0	0.297 3	0.720 4	0.649 0
10∶1	0.710 8	0.780 2	0.416 3	0.802 1	0.757 3
20∶1	0.769 0	0.819 2	0.534 1	0.830 8	0.813 0
30∶1	0.790 9	0.814 7	0.608 3	0.833 5	0.827 2
60∶1	0.797 8	0.818 2	0.673 8	0.835 1	0.825 6
90∶1	0.791 1	0.819 0	0.718 4	0.835 3	0.829 2
120∶1	0.793 1	0.818 9	0.725 3	0.835 1	0.827 6
150∶1	0.787 4	0.818 1	0.744 2	0.834 4	0.824 1
200∶1	0.761 0	0.815 7	0.750 7	0.827 3	0.826 1
250∶1	0.746 8	0.815 5	0.755 6	0.829 9	0.818 8
300∶1	0.726 3	0.812 9	0.746 2	0.825 4	0.820 5
对照	0.708 8	0.813 1	0.748 5	0.822 6	0.826 7

表 6 所示,在随机欠采样的 F_2 值表现中可以看到,综合考率精确率和召回率时,随机森林分类器在采样比例为 90∶1 时达到最高,逻辑回归、K 近邻和 XGBoost 分类器的最佳 F_2 值也都处于 5∶1 和 120∶1 之间,而决策树模型 F_2 值在 250∶1 时达到最高。

表 7 SMOTE 采样的 F_2 值

Table 7 F_2 of SMOTE undersampling

比例	LR	KNN	DTC	RFC	XGBC
1∶1	0.351 3	0.718 0	0.635 7	0.819 9	0.822 3
5∶1	0.640 3	0.736 3	0.659 0	0.836 8	0.825 3
10∶1	0.744 4	0.757 9	0.673 4	0.834 7	0.832 2
20∶1	0.797 2	0.773 5	0.693 6	0.834 4	0.831 2
30∶1	0.809 2	0.781 9	0.690 2	0.835 6	0.831 4
60∶1	0.810 4	0.803 3	0.710 6	0.834 3	0.830 0
90∶1	0.800 8	0.813 0	0.721 4	0.831 2	0.824 5
120∶1	0.796 3	0.820 3	0.742 1	0.831 9	0.826 8
150∶1	0.786 0	0.815 9	0.743 4	0.830 7	0.828 2
200∶1	0.771 7	0.816 2	0.737 3	0.828 6	0.820 9
250∶1	0.745 6	0.816 7	0.754 7	0.826 8	0.825 1
300∶1	0.726 4	0.813 4	0.751 5	0.822 8	0.817 2
对照	0.708 8	0.813 1	0.748 5	0.822 6	0.826 7

表 7 所示,在 SMOTE 采样中 F_2 值表现出相似的结果,这说明采样比例对分类结果有显著影响,在更加关注召回率的情况下,需要对不同的采样比例进行详细讨论,以确定最佳的采样比例,从而获得最优的 $F2$ 值。

4 结束语

针对不平衡数据的分类问题,本文从采样的不同比例进行探索,在极不平衡的数据集上设定不同的采样比例,测试对分类效果的影响。研究发现,不同采样比例对分类结果有很大影响。基于随机欠采样和 SMOTE 采样,各种分类器在精确率的表现中,采样比例越接近原始比例,少数类精确率表现越好;在召回率的表现中,采样比例越接近平衡比例,少数类召回率表现越好。如何确定最佳采样比例,还需要根据实际应用场景,设定 F - Score 的参数,并进行不同采样比例的实验才能得出,但在确定采样比例范围时,如果更加关注少数类精确率,可以在接近原始比例进行采样,如果更加关注少数类召回率,可以选择在接近平衡比例采样。

参考文献

[1] JAHMUNAH V, NG E Y K, SAN T R, et al. Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals [J]. Computers in Biology and Medicine, 2021, 134: 104457.

[2] STEFANOWSKI J. Dealing with data difficulty factors while learning from imbalanced data[C]//Proceedings of Challenges in Computational Statistics and Data Mining. Cham: Springer, 2015: 333-363.

[3] 李梦男, 李琨, 吴聪. 基于 IWAE 的不平衡数据集下轴承故障诊断研究[J]. 机械强度,2023,45(3):569-575.

[4] 贺指陈. 基于集成学习和代价敏感的类别不平衡数据分类算法[J]. 信息记录材料,2022,23(1):18-22.

[5] 刘嘉宇, 李贺, 谷莹,等. 不平衡数据集上在线评论有用性识别研究[J]. 情报理论与实践,2023,46(11):119-125.

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

[7] 王博文, 王景升, 吴恩重. 面向不平衡数据集的 SMOTENC-XGBoost 驾驶人交通安全评估模型[J]. 科学技术与工程,2023,23(2):831-837.

[8] DONG Y , WANG X. A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets[C]//Proceedings of International Conference on Knowledge Science, Engineering and Management. Cham:Springer, 2011:343-352.

[9] 谢子鹏, 包崇明, 周丽华, 等. 类不平衡数据的 EM 聚类过采样算法[J]. 计算机科学与探索,2023,17(1):228-233.

[10] BECKMANN M, EBECKEN N F F, LIMA B S L P. A KNN undersampling approach for data balancing [J]. Journal of Intelligent Learning Systems and Applications, 2015, 7(4):

104-116.
[11] HIDO S, KASHIMA H, TAKAHASHI Y. Roughly balanced bagging for imbalanced data[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2009, 2(5-6): 412-426.
[12] LIN W C. Clustering-based undersampling in class-imbalanced data.[J] Information Sciences, 2017,409:17-26.
[13] FU Y, ZHANG H, BAI Y, et al. An under-sampling method: Based on principal component analysis and comprehensive evaluation model[C]//Proceedings of 2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C). IEEE, 2016: 414-415.
[14] 孟东霞，李玉鑑. 基于特征边界欠采样的不平衡数据处理方法[J]. 统计与决策,2021,37(11):30-33.
[15] LI X, ZHANG L. Unbalanced data processing using deep sparse learning technique[J]. Future Generation Computer Systems, 2021, 125: 480-484.
[16] ELKAN C. The foundations of cost-sensitive learning[C]// Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI). IEEE, 2001: 973-978.
[17] 周传华，徐文倩，朱俊杰. 基于代价敏感卷积神经网络的集成分类算法[J]. 应用科学学报,2022,40(1):69-79.
[18] 平瑞,周水生,李冬. 高度不平衡数据的代价敏感随机森林分类算法[J]. 模式识别与人工智能,2020,33(3):249-257.
[19] 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.
[20] 王萌铎，续欣莹，阎高伟，等. 基于 AdaBoost 集成加权宽度学习系统的不平衡数据分类[J]. 计算机工程,2022,48(4):99-105.