董嘉轩, 俞卫琴. 基于去噪混合采样的医疗不平衡数据分类[J]. 智能计算机与应用,2025,15(11):68-74. DOI:10.20169/j.issn.2095-2163.251111

基于去噪混合采样的医疗不平衡数据分类

董嘉轩, 俞卫琴
(上海工程技术大学 数理与统计学院, 上海 201620)

摘 要: 医疗不平衡数据是一类常见的不平衡数据,本文提出了一种基于去噪混合采样的医疗不平衡数据分类方法。首先,使用最小协方差行列式(MCD)算法去除全局噪声,通过计算马氏距离来识别异常值;其次,结合局部离群因子(LOF)算法对样本点进行局部异常性评估,揭示其在邻域中的边界程度;通过综合全局异常分数和局部异常分数,计算得到反映样本点边界程度的权重,并将其作为混合采样的依据;最后,基于该权重,采用随机欠采样和 SMOTE 方法进行混合采样,有效平衡了数据集,同时保留了边界信息,提高了分类器的性能。实验结果表明,该方法在多个不平衡数据集上优化了分类器的性能,特别是在 G-mean 和 $F1$ 值方面取得了提升。

关键词: 医疗不平衡数据; 混合采样; 最小协方差行列式; 局部离群因子
中图分类号: TP 391 **文献标志码**: A **文章编号**: 2095-2163(2025)11-0068-07

Classification of medical imbalanced data based on denoising hybrid sampling

DONG Jiaxuan, YU Weiqin
(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Medical imbalanced data is a common type of imbalanced data, for which an improved denoising hybrid sampling method that integrates the Minimum Covariance Determinant (MCD) algorithm and the Local Outlier Factor (LOF) algorithm is proposed. Firstly, the MCD algorithm is utilized to remove global noise by identifying outliers using the Mahalanobis distance. Subsequently, the LOF algorithm is combined to assess the local abnormality of sample points, revealing their degree of marginality in the neighborhood. By integrating global anomaly scores with local anomaly scores, weights that reflect the degree of marginality of the sample points are calculated, providing a basis for hybrid sampling. Weight-based random undersampling and weight-based SMOTE methods are used for hybrid sampling, which effectively balance the dataset while preserving boundary information, thereby enhancing the performance of the classifier. Experimental results indicate that this method optimizes the performance of classifiers on multiple imbalanced datasets, particularly demonstrating excellent effects in terms of G-mean and $F1$ scores.

Key words: medical imbalanced data; hybrid sampling; minimum covariance determinant; local outlier factor

0 引 言

机器学习技术能从复杂医疗数据中挖掘有价值的信息,将机器学习方法运用到医疗数据集,在一定程度上可以帮助相关医护人员提高疾病诊断的效率[1]。诸如神经网络、朴素贝叶斯、K 近邻(K-Nearest Neighbors, KNN)和支持向量机(Support Vector Machine, SVM)等分类算法,在医学诊断问题上已被广泛应用[2]。然而,传统的机器学习分类算法通常是为平衡的数据分布而设计的,在现实世界中,医疗领域面临的数据集往往是不平衡的,健康状况的样本数量显著多于其他状况。

在医疗诊断领域中,准确地鉴别罕见病例对于初步筛查和及时治疗至关重要[3]。但这常受制于数据不平衡的问题,这种不平衡数据集为机器学习模型的准确性和可靠性带来了巨大挑战,尤其当任务涉及到识别较为罕见但极其重要的病例时,不平衡数据常常引发模型偏差,使其倾向于预测健康状态以提高总体准确率,从而导致漏诊率上升[4]。传统评估指标,如准确率,不再适用,需采用召回率、查

作者简介: 董嘉轩(1998—),女,硕士研究生,主要研究方向:不平衡数据分类。
通信作者: 俞卫琴(1982—),女,博士,副教授,主要研究方向:非线性动力学理论与应用。Email:yuweiqin1982@163.com。

收稿日期: 2024-02-02

准率和 $F1$ 值等指标去评估模型对少数类的识别能力[5]。为应对这些挑战，提出了多种方法，包括数据重采样、成本敏感学习、和集成方法等，以提高模型对罕见病例的检测准确性[6]。

合成少数类过采样（Synthetic Minority Over-sampling Technique，SMOTE）是一种经典的过采样技术，通过线性插值方法生成合成少数类样本来平衡类分布[7]。但 SMOTE 在少数类样本之间随机生成新的样本，没有考虑多数类样本的分布，可能造成类别边界模糊。为克服该问题，Nguyen 等[8]提出了 SVM-SMOTE（Support Vector Machine-Synthetic Minority Over-sampling Technique，SVM-SMOTE）方法，利用 SVM 来识别多数类和少数类样本之间的边界，并在少数类样本的支持向量之间进行过采样，以改善决策边界的划分；He 等[9]提出自适应综合过采样（Adaptive Synthetic Sampling，ADASYN）方法，侧重于那些周围被多数类样本包围的少数类样本，通过计算这些样本的近邻分布来自适应调整合成样本的生成数量；Sun 等[10]通过将代价敏感学习与 AdaBoost 算法相结合，提出一种代价敏感的 Boosting 算法；Guo 等[11]提出排他性正则化机器（Exclusivity Regularized Machine，ERM），这是一种集成 SVM 方法，通过垂直排列各 SVM 的分类界限来优化分类器之间的多样性，提升了分类器的性能。

针对处理不平衡医疗数据集时分类器性能下降的问题，本文提出了一种改进的去噪混合采样方法。通过最小协方差行列式估计（Minimum Covariance Determinant，MCD）算法清除噪声，并通过 MCD 和局部离群因子（Local Outlier Factor，LOF）算法相结合来确定每个样本的权重，并根据权重进行混合采样来平衡数据集。本文提出的去噪混合采样方法不仅可以识别并剔除噪声数据，还可以通过考虑数据整体和局部的分布特性来进行采样，减轻过拟合的风险，保留有价值的数据特征。对 3 个公开的医疗数据集进行实验，并与其他的重采样方法进行了比较，本文提出的方法有效提升了召回率、查准率、G-mean 和 $F1$ 值，验证了其对医疗不平衡数据分类的有效性。

1　相关技术

1.1　MCD 算法

MCD 算法是一种旨在估计多元正态分布的数据集中心和稳健协方差矩阵的方法，主要应用于检测数据集中的整体异常值[12]。MCD 算法的核心思想是通过最小化协方差行列式来确定数据点的一个子集，这个子集的协方差矩阵能够抵抗异常值的影响，从而确保所找到的数据点受异常值影响较小[13]。该算法通过计算数据点相对于该子集中心的马氏距离来实现异常值的识别。MCD 算法的具体步骤：

（1）从原始数据集中随机选择一个具有 h 个样本点初始子集 H，并计算该初始子集的均值 μ^H 和协方差矩阵 $\boldsymbol{\Sigma}^H$；

（2）利用均值和协方差矩阵，根据式（1）计算整个数据集中每个点相对于该子集中心的马氏距离：

$$\mathrm{MD}(x_i)=\sqrt{(x_i-\mu^H)^{\mathrm{T}}(\boldsymbol{\Sigma}^H)^{-1}(x_i-\mu^H)} \quad (1)$$

（3）根据计算出的马氏距离，选取距离中心最近的点，以形成新的子集，用新子集计算中心和协方差矩阵；

（4）多次迭代步骤（3），直到满足一定的收敛条件，例如中心位置和协方差矩阵的变化很小，或者达到预设的最大迭代次数。

1.2　LOF 算法

LOF 算法是一种基于密度的异常检测方法，其核心思想是通过比较数据点与其近邻的局部密度来鉴别异常值[14]。LOF 算法不仅仅考虑了数据点本身的密集程度，还考虑了其邻居之间的密集程度。通过计算每个数据点的局部离群因子，可以识别出相对于其邻居而言异常的数据点。LOF 算法的具体步骤如下。

1）针对每个数据点，确定其与 K 近邻之间的距离，根据距离计算局部可达密度，局部可达密度表示了该数据点相对于其邻居的密集程度；

2）计算数据点近邻的局部可达密度与自身局部可达密度的比值，得到局部离群因子，局部离群因子越大，表示该数据点相对于其邻居更可能是一个离群点；

3）依据局部可达密度之间的比率计算局部离群因子，得到的局部离群因子可以用于排名，进而通过设定阈值来鉴别哪些点是离群点。

1.3　SMOTE 算法

SMOTE 算法是一种用于处理不平衡数据集的过采样方法，其主要目标是通过合成新的少数类样本来平衡数据集，从而提升模型对少数类的识别能力。SMOTE 算法的具体步骤如下。

1）对于少数类中每一个样本 x_i，用欧氏距离计算其到少数类中所有样本的距离，得到在 n 维空间的 K 个近邻；

2)据样本不平衡比例设置采样倍率 N, 对于每一个少数类样本 x_i, 从其 K 近邻中随机选择若干个样本,假设选择的近邻为 x_n;

3)对于每一个随机选出的近邻 x_n, 分别与 x_i 按照下式生成新的样本。

$$x_{\text{new}} = x_i + \lambda \times (x_n - x_i) \quad (2)$$

其中, λ 为[0,1]间的随机数。

SMOTE 算法通过生成新的合成样本而不是简单地复制已有样本,能够有效缓解过拟合问题,增强模型的泛化能力。

2 基于 MCD 和 LOF 改进的去噪混合采样

2.1 数据去噪和边界权重计算

本文提出了一种结合 MCD 算法和 LOF 算法的新型数据去噪和边界权重计算方法,通过 MCD 算法识别和去除样本的全局异常值,再结合 LOF 算法对局部异常值的检测,能够从整体和局部两个层面评估和确定样本点的边界程度,并计算出每个样本能够代表其边界重要性的权重,为后续的混合采样提供一个更为清晰和准确的数据基础,每一步骤的原理和具体操作如下。

1)应用 MCD 算法进行数据去噪。MCD 算法是一种鲁棒的统计方法,其作用是提供一个稳健的估计来描述数据的分布特征。MCD 算法可以找出全局层面上与整体数据分布显著不同的异常值,并且对数据的分布形状没有过多的假设要求。MCD 算法选择代表数据正常模式的子集,计算该子集的稳健中心和协方差矩阵,利用得到的稳健中心和协方差矩阵,可以计算出每个样本点的马氏距离,这是一种综合考虑特征间相关性的距离度量。本文分别对少数类和多数类使用 MCD 算法,通过计算每个样本点的马氏距离,可以量化每个点相对于整体数据分布的偏离程度;将每类中马氏距离高于阈值的样本点认定为噪声,从数据集中移除,从而实现数据的整体去噪。

2)计算全局与局部异常分数。去除噪声后,对每个样本点计算两个异常分数:一个是基于 MCD 的全局异常分数;另一个是基于 LOF 的局部异常分数。尽管 MCD 算法不能直接度量类别边界,但 MCD 算法计算出的具有较高马氏距离的点通常位于数据分布的边缘,这是由于在多变量数据中,处于边界上的点通常具有与核心数据集不同的属性。因此,MCD 算法提供的全局异常分数反映了样本点与整体数据分布的偏离程度。LOF 算法衡量样本点相对于其邻近点的局部密度差异[15]。处于类别边界的点往往与其邻居相距较远,导致这些点具有较低的局部密度。这种密度差异会导致这些点的 LOF 分数相对较高,从而揭示了样本点在其邻域结构中的边界程度。因此,LOF 分数反映了样本点在其局部邻域内的密度偏离。

3)融合异常分数计算边界权重。综合全局异常分数和局部异常分数,将其相加计算出每个样本点的边界权重,这个权重是结合两个异常分数得到的,反映了样本点在全局分布和局部邻域中的异常性。这一综合指标不仅考虑了样本点是否远离全局数据中心,也考虑了其是否在局部邻域内显著的不同于其他点。通过这种方法,本文为每个样本点赋予了一个边界程度的量化指标,为后续的混合采样工作提供了重要的基础。

本文提出的结合 MCD 和 LOF 的数据去噪和边界权重计算方法,不仅去除了数据的噪声,还实现了一种多尺度的权重评估,对样本点是否属于边界的判别更加精确。

2.2 混合采样

混合采样策略中,识别数据的边界点尤为重要,在模型训练时能够提供更多的关键信息,因此采用结合 MCD 和 LOF 计算出的边界权重来指导混合采样的样本生成。该混合采样策略融合了加权随机欠采样和加权 SMOTE,旨在通过减少多数类样本数量,同时增加少数类样本数量,来达到一个更加平衡的数据集状态,同时保留样本的重要性信息。基于权重的混合采样的具体步骤如下。

1)基于权重的随机欠采样:使用计算出的样本权重,对多数类实施基于权重的随机欠采样,样本被选择的概率与其权重成反比,权重越低的样本越有可能被丢弃;

2) 基于权重的 SMOTE 方法:(1) 选择少数类的样本作为种子,并基于其权重来决定每个少数类样本生成多少合成样本;(2) 对于每个种子样本,找到其在特征空间中的 k 个最近邻,选择最近邻进行合成样本生成时考虑不同样本的权重,选择权重较高的样本作为插值的候选对象;(3) 基于种子样本与选择的最近邻进行插值,生成新的合成样本;

3) 将欠采样后的多数类样本和过采样后的少数类样本合并,形成一个新的、平衡的数据集,使用生成的平衡数据集来训练机器学习模型。

本文使用基于 MCD 和 LOF 得到的权重信息,优化了混合采样过程,为混合采样提供了精确的指

导。

2.3　基于 MCD 和 LOF 改进的去噪混合采样算法的详细步骤

输入　原始数据集 D，MCD 算法参数（子集大小 h，马氏距离阈值 T），LOF 算法参数（最近邻个数 K_{LOF}），SMOTE 过采样最近邻个数 K_{SMOTE}。

输出　平衡化处理后的数据集 D_{balanced}。

（1）将数据集划分为训练集 D_{train} 和测试集 D_{test}；

（2）计算混合采样后各类的样本数量，公式如下：

$$l = (D_{\mathrm{train}}\text{中多数类样本数量} + D_{\mathrm{train}}\text{中少数类样本数量})/2 \tag{3}$$

（3）应用 MCD 算法，子集大小为 h，对 D_{train} 中的少数类和多数类分别计算马氏距离 MD；

（4）将 MD 超过阈值 T 的数据点视为噪声点，从 D_{train} 中移除噪声点，得到去噪后数据集 D_{denoised}；

（5）对 D_{denoised} 中每个样本点 i，计算基于 MCD 的全局异常分数 M_i 和基于 LOF 的局部异常分数 L_i；

（6）对于每个样本点 i，将全局异常分数 M_i 和局部异常分数 L_i 相加，得到边界权重 W_i；

（7）对多数类样本根据 W_i 的值进行随机欠采样，保留 l 个样本，样本被保留的概率与其权重成反比，得到欠采样后的数据集 $D_{\mathrm{undersampled}}$；

（8）对少数类样本执行基于边界权重的 SMOTE 过采样，选择 W_i 较高的样本作为种子点，并使用 K_{SMOTE} 个最近邻生成新样本，得到过采样后的数据集 $D_{\mathrm{oversampled}}$；

（9）合并 $D_{\mathrm{undersampled}}$ 和 $D_{\mathrm{oversampled}}$，形成新的平衡训练数据集 D_{balanced}。

该算法通过 MCD 和 LOF 算法对样本进行多维度的异常性分析，通过结合全局和局部的异常分数，为样本赋予权重，指导后续的采样过程。这样既考虑了数据的全局分布特性，又考虑了样本的局部邻域信息，使用该算法输出的平衡训练集训练机器学习模型，获得最终的分类器，能够在处理不平衡数据集时得到更优的分类结果。

3　实验

3.1　数据集和实验环境

本文选择 3 个来自 UCI 数据库的公开医疗数据集进行实验。Heart 数据集为心脏病数据集，包括心脏病患者和非患者的各种属性，如年龄、性别、胸痛类型、静息血压等，目标是预测患者是否有心脏病；Pima 数据集为糖尿病数据集，包含来自皮马印第安人的女性的医疗记录，包括年龄、怀孕次数、胰岛素水平等，目标是预测患者是否有Ⅱ型糖尿病；WBCD 数据集为乳腺癌威斯康星（诊断）数据集，包含了乳腺癌肿瘤的特征，目标变量是肿瘤是恶性的还是良性的。实验使用数据集的具体参数见表 1，其中 IR 为不平衡率，公式如下：

$$\mathrm{IR} = \frac{\text{少数类样本数}}{\text{多数类样本数}} \tag{4}$$

表 1　实验数据集参数

Table 1　Experimental dataset parameters

数据集	特征数	样本数	少数类样本数	多数类样本数	IR
Pima	8	768	268	500	0.536
Heart	13	294	106	188	0.564
WBCD	30	569	212	357	0.594

本文实验均在 Intel（R）Core（TM）i5-1035G1 CPU @ 1.00 GHz 1.19 GHz 处理器，16 G 内存的电脑上进行，软件环境为 Python 3.9。实验采用分层随机抽样的方式来划分训练集与测试集，其中训练集与测试集的比例为 4∶1，混合采样合成的数据样本只参与分类器的训练阶段，测试阶段的数据全部为原始数据。

3.2　评价指标

评估不平衡数据分类效果时，本文采用召回率、查准率、G-mean 和 $F1$ 值作为评价指标，公式如下：

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

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} \tag{6}$$

$$\mathrm{G-mean} = \sqrt{\frac{\mathrm{TP} \times \mathrm{TN}}{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{7}$$

$$F1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

其中，TP 为真正例，指实际类别为正类且预测也为正类；FP 为假正例，指实际类别为负类，但模型预测错误为正类；FN 为假反例，指实际类别为正类，但模型预测错误为负类的情况；TN 为真反例，指实际类别为负类，预测也为负类。

在这些指标中，召回率衡量分类器识别正类样本的能力；查准率评估被识别为正类的样本中实际正类的比例；G-mean 是查准率和召回率的几何平

均,用于衡量不平衡数据上的性能; *F*1 值是查准率和召回率的调和平均数,能够综合考察查准率和召回率[16]。

3.3 实验结果及分析

为评估本文提出的去噪混合采样方法在不平衡医疗数据集上的性能,通过在 3 个医疗数据集上的实验来验证该方法的有效性及其与主流机器学习分类器的兼容性。选择 3 种重采样方法即 SMOTE、ADASYN 和基于 K-means 的欠采样方法,作为对比方法;同时,为了证明本文方法的普适性,选择 3 种经典的机器学习分类算法即 SVM、随机森林(RF)和 KNN,将这些分类算法与不同的重采样方法结合,进行了对比实验,全面评估本文方法在不同情境下的表现与适用性。使用不同分类算法时,本文方法和其他 3 种重采样方法在 3 个医疗数据集上的实验结果见表 2~表 4,其中加粗的结果为最优结果。

表 2 使用 SVM 分类时不同算法的评价指标

Table 2 Evaluation metrics for different algorithms when using SVM classification

数据集	评价指标	本文方法	SMOTE	ADASYN	基于 K-means 聚类的欠采样
Pima	召回率	**0.777 78**	0.695 74	0.762 39	0.657 09
	查准率	**0.636 36**	0.582 33	0.585 00	0.586 36
	G-mean	**0.591 11**	0.520 37	0.543 78	0.497 70
	*F*1 值	**0.700 00**	0.630 80	0.658 94	0.618 06
Heart	召回率	**0.714 29**	0.630 17	0.634 06	0.638 71
	查准率	0.526 32	0.571 08	**0.579 60**	0.469 35
	G-mean	**0.526 32**	0.464 00	0.443 42	0.411 74
	*F*1 值	**0.652 17**	0.594 57	0.600 94	0.537 12
WBCD	召回率	0.947 37	0.900 00	**0.979 17**	0.853 66
	查准率	**0.972 97**	0.918 37	0.810 34	0.972 22
	G-mean	**0.934 90**	0.843 75	0.815 97	0.841 96
	*F*1 值	**0.960 00**	0.909 09	0.886 79	0.909 09

表 3 使用随机森林分类时不同算法的评价指标

Table 3 Evaluation metrics for different algorithms when using Random Forest

数据集	评价指标	本文方法	SMOTE	ADASYN	基于 K-means 聚类的欠采样
Pima	召回率	0.722 22	0.636 40	0.709 26	**0.758 38**
	查准率	**0.661 02**	0.635 23	0.622 37	0.587 25
	G-mean	**0.577 78**	0.504 20	0.547 63	0.541 46
	*F*1 值	**0.690 27**	0.633 02	0.660 06	0.661 36
Heart	召回率	**0.809 52**	0.656 45	0.686 92	0.799 07
	查准率	0.680 00	**0.717 18**	0.709 75	0.667 47
	G-mean	**0.639 10**	0.570 48	0.575 28	0.627 03
	*F*1 值	**0.739 13**	0.680 88	0.690 10	0.719 26
WBCD	召回率	**0.973 68**	0.951 22	0.955 56	0.942 31
	查准率	**1.000 00**	0.951 22	0.934 78	0.960 78
	G-mean	**0.973 68**	0.925 16	0.914 01	0.911 91
	*F*1 值	**0.986 67**	0.951 22	0.945 05	0.951 46

表 4　使用 KNN 分类时不同算法的评价指标

Table 4　Evaluation metrics for different algorithms when using KNN classification

数据集	评价指标	本文方法	SMOTE	ADASYN	基于 K-means 聚类的欠采样
Pima	召回率	**0.685 19**	0.641 51	0.660 38	0.672 73
	查准率	**0.616 67**	0.576 27	0.479 45	0.606 56
	G-mean	**0.527 59**	0.482 72	0.411 92	0.509 64
	*F*1 值	**0.649 12**	0.607 14	0.555 56	0.637 93
Heart	召回率	**0.761 90**	0.655 17	0.692 31	0.714 29
	查准率	0.615 38	**0.678 57**	0.642 86	0.576 92
	G-mean	**0.561 40**	0.458 62	0.482 52	0.507 52
	*F*1 值	**0.680 85**	0.666 67	0.666 67	0.638 30
WBCD	召回率	**0.947 37**	0.906 98	0.940 00	0.886 79
	查准率	**1.000 00**	0.847 83	0.870 37	0.979 17
	G-mean	**0.947 37**	0.817 56	0.837 19	0.872 25
	*F*1 值	**0.972 97**	0.876 40	0.903 85	0.930 69

实验评估了本文方法和其他重采样方法在 SVM、RF 和 KNN 分类器上的性能，关注了召回率、查准率、G-mean 和 *F*1 值这 4 个指标。通过表 2～表 4 可以看出，在使用 SVM 分类器时，本文方法的多数指标超越了 SMOTE、ADASYN 以及基于 K-means 的欠采样方法，在 Heart 数据集上，本文方法的查准率相对较低；在 WBCD 数据集上，本文方法的召回率低于 ADASYN 方法。采用 RF 做为分类器时，在 Pima 数据集上，本文方法的召回率略低于基于 K-means 的欠采样方法，但在查准率、G-mean 和 *F*1 值上具有更优的结果，对于 Heart 数据集，本文方法的查准率数值相对较低，而其他指标均取得了良好结果；在 WBCD 数据集上，本文方法的各项指标均达到最优。在 KNN 分类器的评估中，本文方法仅在 Heart 数据集的查准率上数字较低，在 Pima 数据集和 WBCD 数据集上，本文方法在所有指标上均优于其他对比方法。

综合来看，本文方法在多个分类器和数据集中显示出了均衡和可靠的性能，尤其是在 G-mean 和 *F*1 值上，在使用 3 种分类器时，在所有数据集上本文方法均优于其他对比方法。G-mean 的提高说明本文方法对多数类和少数类的识别能力是均衡的。*F*1 值的提高表明本文方法在查准率和召回率的提高是平衡的，具有更好的整体分类精度。在实际应用中，*F*1 值的提高能帮助医生寻找查准率和召回率之间的最佳平衡点，尤其是在那些对查准率和召回率同样重视的场景中。在不同数据集上都能维持更高的 G-mean 和 *F*1 值，也证明了本文方法具有较好的泛化能力和鲁棒性。

4　结束语

本文提出了一种融合 MCD 算法和 LOF 算法的改进去噪混合采样方法。该方法结合 MCD 算法和 LOF 算法进行数据去噪和边界权重计算，有效去除了全局噪声，确保了数据质量，揭示了样本点的边界程度。在混合采样阶段，对于少数类过采样，优先选择权重较高的边界点进行新样本点的生成，既加强了模型对复杂决策边界的学习，又减轻了对多数类核心区域过度采样的风险，避免了过拟合；对于多数类欠采样，权重信息有助于识别并保留多数类中重要的边界样本点。在多个医疗不平衡数据集和分类器上的实验结果验证了该方法的有效性，尤其是在 G-mean 和 *F*1 值两个关键指标上，本文方法具有明显的优势，不仅提升了少数类识别的精度，也保持了整体分类的准确性。

参考文献

[1] 陈雪琴. 基于机器学习的医疗数据集填补和分类研究[D]. 武汉：华中科技大学，2020.

[2] 兰欣，卫荣，蔡宏伟，等. 机器学习算法在医疗领域中的应用[J]. 医疗卫生装备，2019，40（3）：93-97.

[3] AL-SHAMMARI A，ZHOU R，LIU C，et al. A framework for processing cumulative frequency queries over medical data streams[C]//Proceedings of 2018 19th International Conference. Cham：Springer，2018：121-131.

[4] 李艳霞，柴毅，胡友强，等. 不平衡数据分类方法综述[J]. 控制与决策，2019，34（4）：673-688.

[5] MULLICK S S，DATTA S，DHEKANE S G，et al. Appropriateness of performance indices for imbalanced data classification：An analysis

[J]. Pattern Recognition, 2020, 102: 107197.
[6] WANG L, HAN M, LI X, et al. Review of classification methods on unbalanced data sets[J]. IEEE Access, 2021, 9: 64606-64628.
[7] 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.
[8] NGUYEN H M, COOPER E W, KAMEI K. Borderline over-sampling for imbalanced data classification[J]. International Journal of Knowledge Engineering and Soft Data Paradigms, 2011, 3(1): 4-21.
[9] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Proceedings of 2008 IEEE International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2008: 1322-1328.
[10] SUN Y, KAMEL M S, WONG A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.
[11] GUO X, WANG X, LING H, et al. Exclusivity regularized machine: A new ensemble SVM classifier[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI, 2017: 1739-1745.
[12] HUBERT M, DEBRUYNE M, ROUSSEEUW P J. Minimum covariance determinant and extensions[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2018, 10(3): e1421.
[13]张若璇, 田茂再. 多元数据离群点探测的倾斜重加权方法[J]. 数理统计与管理, 2019, 38(4): 619-627.
[14]庄池杰, 张斌, 胡军, 等. 基于无监督学习的电力用户异常用电模式检测[J]. 中国电机工程学报, 2016, 36(2): 379-387.
[15]刘财辉, 刘地金. 离群点检测的邻近性方法综述[J]. 计算机工程与应用, 2022, 58(21): 1-12.
[16]刘赛可, 何晓群, 夏利宇. 不平衡数据下模型评价指标的有效性探讨[J]. 统计与决策, 2022, 38(19): 5-9.