RareBoost: 一种基于稀有区域识别的稀有值自助不平衡回归方法

刘 丹，付英姿，付光辉†

(昆明理工大学 理学院，昆明 650500)

摘 要: 不平衡回归任务中, 连续目标变量的分布不均衡使得稀有值预测面临挑战, 其核心难点在于如何有效界定并识别稀有值. 针对这一挑战, 提出一种新的稀有区域识别策略 (KK-means), 该方法结合核密度估计和K-means 聚类, 能够将目标变量空间中的稀疏样本点系统性地识别并合并为连续的稀有区间. 进而, 提出一种基于稀有区域识别的稀有值自助不平衡回归方法 (RareBoost). RareBoost 先通过标签密度比加权提取稀有区间信息, 并在自助采样中动态调整样本权重, 以增强模型对稀有区域的关注; 然后通过 Stacking 元学习器集成这些具有“稀有值感知”能力的基学习器, 形成兼顾全局效率与局部精度的稀有值预测模型. 实验表明, RareBoost 在 ANLL、RMSE 与 R^2 上的平均表现分别比最佳基线方法提升 8.7%、5.8% 和 13.8%, 验证了其有效性.

关键词: 不平衡回归；稀有区域识别；Bagging；自助采样；稀有值提升；稀有值预测

中图分类号: 68T05，68T10 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1299

引用格式: 刘丹, 付英姿, 付光辉. RareBoost: 一种基于稀有区域识别的稀有值自助不平衡回归方法 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

RareBoost: A rare value self-boosting imbalanced regression method via rare region identification

LIU Dan，FU Ying-zi，FU Guang-hui†

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

Abstract: In imbalanced regression tasks, the uneven distribution of continuous target variables poses a significant challenge for predicting rare values, with the core difficulty lying in effectively defining and identifying these rare values. To address this challenge, this paper first proposes a novel rare region identification strategy, named KK-means, which combines kernel density estimation and K-means clustering to systematically identify and merge sparse sample points in the target variable space into continuous rare intervals. Furthermore, a rare value bootstrap-based imbalanced regression method (RareBoost) is proposed, based on the identified rare regions. The RareBoost first extracts information from the identified rare intervals through label density ratio weighting and dynamically adjusts sample weights during the bootstrap sampling process to enhance the model's focus on rare regions. Subsequently, a Stacking meta-learner is employed to integrate these base learners with “rare value awareness”, forming a predictive model that balances global efficiency and local accuracy for rare values. Experimental results demonstrate that the RareBoost achieves average improvements of 8.7%, 5.8%, and 13.8% over the best baseline method in terms of ANLL, RMSE, and R^2, respectively, confirming its effectiveness.

Keywords: imbalanced regression； rare region identification； Bagging； bootstrap sampling； rare value boosting； rare value prediction

0 引 言

回归问题广泛应用于金融风险预警、气象预测、医疗健康监测和工业设备诊断等领域. 然而, 现实数据通常存在显著的不平衡特征: 目标变量在数值区间上分布偏斜, 极端值或低频区间样本稀缺. 受此影响, 模型更容易拟合多数区间, 导致在稀有区域的预

收稿日期: 2025-12-17；**录用日期:** 2026-03-03.

基金项目: 国家自然科学基金项目 (12261052, 11761041)；云南省基础研究计划项目 (202501AS070103)；云南省“兴滇人才支持计划”.

责任编委: 李少远.

†通信作者. E-mail: ghuifu@126.com.

测精度不足. 实际上, 在许多应用中, 这些极端值往往具有更高的决策价值. 因此, 不平衡回归逐渐成为研究的重点[1-2].

传统的处理不平衡数据的方法, 如过采样、欠采样和类权重调整, 广泛应用于分类任务中, 通过调整不同类别样本的数量或权重来提高模型性能[3]. 然而, 当这些方法应用于连续目标变量的任务时, 往往无法达到理想效果[4]. 具体来说, 过采样方法在连续目标变量情况下难以定义需要过采样的区域[5], 并且生成的新样本可能导致过拟合[4]. 欠采样方法可能会丢失有价值的信息, 削弱稀疏区域的学习能力[4,6]. 类权重调整在回归任务中缺乏明确类别指导, 确保有效性具有挑战性[7]. 因此, 传统的不平衡数据处理技术在连续目标任务中存在局限, 降低了平衡目标值分布的效率和效果[3].

不平衡回归问题近年来逐渐引起关注, 特别是在稀有值的识别和处理方面. Branco 等人[8] 提出的 SMOGN 方法优化了回归模型的表现; Steininger[10] 提出基于密度加权的机制, 增强了对稀有值的预测能力; In 和 Kim[10] 提出基于距离的相关性函数, 为稀有值划分提供了新思路; Avelino 等人[11] 综述了不平衡回归中的重采样策略, 并提出了针对稀有值区域偏差的改进方法; Alahyari 和 Domaratzki[12] 介绍了基于生成对抗网络的 SMOGAN 方法, 用于生成稀有值区域的样本, 推动了不平衡回归研究向更先进的方向发展; 胡峰等人[13] 提出了 GSOGB-SMOTER 方法, 通过将数据划分为粒球, 并基于样本稀有度进行过采样, 生成新样本以提高稀有样本的预测准确性. 尽管如此, 稀有值的定义不统一, 且在训练数据中的稀有值仍是挑战, 影响了模型对这些区域特征的学习. 有效处理稀有值对于金融风控和医疗诊断等领域至关重要, 能够显著提升模型性能并减少潜在风险, 因此研究如何划分和处理稀有值是提高不平衡回归模型实际应用价值的关键.

集成学习作为处理连续不平衡数据的有效方法, 通过结合多个模型的预测结果, 实现"整体大于部分之和"的效果[14]. 与单一模型不同, 集成学习强调异构模型的组合, 利用每个模型独特的理解应对数据复杂性和不平衡性, 特别适用于不平衡回归问题, 能提高稀有值的识别和预测能力.

在集成学习中, 常见的两种方法是 Bagging 和 Stacking. Bagging 通过在不同子数据集上训练多个相同的基学习器, 减少模型的方差, 提高预测的稳定性和准确性[15]. 特别是在数据量大或噪声较多时, Bagging 能够显著提升模型的稳定性, 广泛应用于分类和回归问题, 特别是高方差模型如决策树[16]. Stacking 采用两层结构, 第一层使用多个基学习器, 第二层通过训练元学习器增强模型的泛化能力[17]. Stacking 在许多领域取得了显著应用, 包括能耗预测[18]、气体浓度预测[19]、森林地上生物量估计[20] 以及温室湿度预测[21]. 总体而言, 集成学习, 尤其是异构模型集成方法, 提供了有效的工具来解决不平衡问题, 显著提升预测性能.

在密度估计的应用中, 传统的密度估计方法, 如核密度估计[22]、Gamma 分布采样[23] 以及线性判别分析[24], 已广泛应用于不平衡分类问题. 尽管连续不平衡回归问题的研究较少, 但该领域正逐渐得到重视. Michael Steininger[9] 提出的基于密度加权的回归方法突出了密度信息在处理此类问题时的关键作用. 为了更好地应对这一问题, Zhao 等人[25] 提出了基于 Stacking 的密度估计方法 (Stacking Density Estimation, SDE), 它结合了集成学习和密度估计的优势, 能更精准地识别稀有值, 在精度和稳定性方面表现优越, 提供了一种高效的工具来解决连续不平衡回归问题.

本文提出了 RareBoost 方法, 该方法结合了核密度估计 (Kernel Density Estimation, KDE) 与 K-means 聚类的优势, 通过精准识别稀有区域, 并在自助采样过程中动态调整样本权重, 从而显著提高了不平衡回归任务中稀有值的预测精度. 本文的主要贡献如下:

1) 提出了 KK-means 方法, 将核密度估计与 K-means 聚类相结合, 有效识别不平衡回归任务中的稀有区域, 精确定位低密度区域, 增强模型对稀疏样本的敏感性;

2) 基于 KK-means 设计了 RareBoost 方法, 通过自助采样和基于标签密度比估计 (Label Density Ratio Estimation, LDRE) 权重的加权集成方法, 有效提高了稀有值样本的学习能力, 特别是在极端数据点的预测中表现突出. 通过集成多个基学习器和堆叠元学习器, RareBoost 能够在处理稀疏区域时兼顾全局效率与局部精度.

本研究的其余部分组织如下: 第 1 节回顾了相关领域的研究工作, 第 2 节详细阐述了基于稀有区域识别的稀有值自助不平衡回归方法. 第 3 节展示了实验设置及结果分析, 通过与其他常见方法的对比, 证明了 RareBoost 方法的优越性. 最后, 第 4 节总结了研究成果, 并讨论了未来的研究方向.

1 相关工作

1.1 不平衡回归

在过去的 15 年里, 研究人员提出了各种方法来解决不平衡回归问题, 跨越了不同的领域和技术[27]. 不平衡回归是一种特殊类型的回归任务, 具有三个方面的特征[26]:

1) 目标变量具有偏态分布;

2) 目标变量域中的值不同等重要;

3) 重点是罕见 (代表性不足) 的案例.

这类问题广泛存在于金融风险预测、气象极端天气建模、医疗异常指标诊断以及工业异常检测等场景中.

在这一背景下, 传统的回归性能指标 (如 MSE、MAE、R^2) 往往失效. 原因在于这些指标假设目标值等价, 优化目标被多数区间主导, 使得模型在常见值上表现良好, 但在稀有值上预测偏差极大. 这不仅造成应用风险, 也限制了模型的实际价值. 为此, Torgo 提出了相关性函数的形式化方法[2], 用于刻画数值域中不同目标值的重要性: $\phi: Y \to [0,1]$, 其中$\phi(y)=0$表示该值几乎无关紧要, $\phi(y)=1$表示最为关键. 给定阈值t_R, 可以将训练集划分为稀有区与常规区 $\mathcal{D}_R=\{(x_i,y_i) \mid \phi(y_i) \geqslant t_R\}$, $\mathcal{D}_N=\{(x_i,y_i) \mid \phi(y_i) < t_R\}$.

在评价指标方面, 已有方法经历了从传统误差到稀有值敏感指标的演化. Torgo 等人为不平衡回归任务引入了以下新的性能指标: 精确度 ($prec^{\phi}$) 和召回率 (rec^{ϕ})[28]. 基于这项工作, Branco[29] 提出了回归的精确度和召回率的定义如下:

$$prec^{\phi}=\frac{\sum\limits_{\phi(\hat{y}_i)\geqslant t_R}(1+U_{\phi}^{p}(\hat{y}_i,y_i))}{\sum\limits_{\phi(\hat{y}_i)\geqslant t_R}(1+\phi(\hat{y}_i))}, \tag{1}$$

$$rec^{\phi}=\frac{\sum\limits_{\phi(y_i)\geqslant t_R}(1+U_{\phi}^{p}(\hat{y}_i,y_i))}{\sum\limits_{\phi(y_i)\geqslant t_R}(1+\phi(y_i))}, \tag{2}$$

其中, $\phi(y_i)$是与观察值y_i相关的相关性, $\phi(\hat{y}_i)$是预测值$\hat{y}_i$的相关性, t_R是用户定义的阈值, 用于标记对用户来说相关的实例, $U_{\phi}^{P}(\hat{y}_i,y_i)$是对真实值$y_i$进行预测$\hat{y}_i$的效用, 归一化到$[-1,1]$的范围内. 为了直接衡量稀有区域的误差, 提出了$SER_t$指标, 该指标仅在$\phi(y_i)\geqslant t_R$的子集上计算平方误差, 从而关注稀有值的预测性能. SER_t 的局限性在于其依赖于单一阈值, 导致结果较为敏感. 为了解决这一问题, Moniz 和 Ribeiro 提出了平方误差相关面积 (Squared Error Relevance Area, SERA)[30], SERA 是对 SER_t 的积分, 表示的是整个区间内$[0,1]$下的累积误差, 目的是通过考虑所有可能的阈值t来得到模型的综合误差. SERA 计算的是 SER_t 在整个阈值区间上的总和, 通过积分形式得出, 通常用于描述模型在不同的阈值下的整体表现. SERA 成为评估极端值预测性能的里程碑指标. SER_t 和 SERA 的定义如下:

$$SER_t=\sum_{y_i\in\mathcal{D}^t}(\hat{y}_i-y_i)^2, \tag{3}$$

其中, $\hat{y}_i$是模型预测值, y_i是实际值, $\mathcal{D}^t$是满足$\phi(y_i)\geqslant t$的数据子集.

$$SERA=\int_0^1 SER_t dt=\int_0^1\sum_{y_i\in\mathcal{D}^t}(\hat{y}_i-y_i)^2 dt. \tag{4}$$

最新的 Kou 和 Fu[31] 进一步提出自适应平方误差相关 (Adapted Squared Error Relevance, ASER), 在 SERA 框架下引入权重函数, 使不同样本根据其稀有度获得动态加权, 考虑了不平衡领域中不同值的重要性. 与一些先前的指标相比, ASER 能够更好地评估模型对稀有情况的预测能力, 并避免过拟合. ASER_t 和 ASER 的定义如下:

$$ASER_t=\sum_{i\in D^t}w_i(\hat{y}_i-y_i)^2, \tag{5}$$

$$ASERA=\int_0^1 ASER_t\,dt=\int_0^1\sum_{i\in D^t}w_i(\hat{y}_i-y_i)^2 dt, \tag{6}$$

其中, w_i表示每个数据点i的真实值权重, 随目标值的稀有程度变化, 越稀有的值受到越大惩罚. $ASER_t$表示当阈值t取不同值时的误差平方和, 而 ASER 表示 ASER_t 曲线下的区域. 这一改进使得模型在稀有区的预测性能显著提升, 同时避免了 SERA 中不同阈值划分的依赖性. 综合来看, 不平衡回归的研究脉络经历了方法到指标的双重演化: 早期依靠采样与加权缓解数据不平衡, 中期引入 SERA 作为评价工具, 而近期则发展到 ASER 等稀有度敏感指标, 实现了从整体误差最小化向稀有值优先范式的转变. 这些工作不仅为模型比较提供了新的基准, 也为未来结合稀有度建模、损失函数优化与评价体系统一提供了方向.

1.2 自助集成学习模型设计

集成学习方法通过结合多个学习器来提高预测精度和稳定性, 常见的集成方法包括 Bagging 和 Stacking. Bagging 通过自助采样生成多个训练子集, 在每个子集上训练基础学习器, 并将结果结合起来,

以减少方差并提高预测稳定性. Stacking 则通过多个基础学习器的预测结果作为元学习器的新特征, 进而生成最终预测. 尽管这些方法在集成策略上有所不同, 但它们都通过多样化的学习器来增强整体性能, 特别在处理复杂问题时具有优势. 在本研究中, 我们选择合适的基础学习器和元学习器, 以便构建一个稳定且灵活的集成模型. 通过合理配置这些学习器, 我们能够更好地应对不同数据集的挑战, 提升预测性能.

基础学习器的选择在集成学习中起着关键作用. 我们选择了四个高度互补的基础学习器, 它们分别擅长不同类型的数据模式, 并且能够在集成模型中互相补充, 从而增强整体性能, 四个基学习器选择如下:

1) 广义线性模型 (Generalized Linear Model, GLM): 适用于线性关系数据, 能够有效地建模连续变量之间的线性关系[32];

2) 决策树 (rpart): 能够处理特征之间非线性关系的数据, 尤其擅长捕捉数据中的复杂分裂模式, 通过递归地基于不同特征的阈值将数据划分为不同区域[33];

3) 支持向量机 (Support Vector Machines, SVM): 特别适用于小样本和高维稀疏数据, 能够有效地进行模式分类和回归[34];

4) K-近邻 (K-Nearest Neighbors, KNN): 适合捕捉数据中的局部结构, 尽管计算复杂, 但对噪声数据具有较强的鲁棒性[35].

这些基础学习器的多样性和互补性使得集成模型能够充分利用数据中的不同特征, 从而提高对复杂数据的适应性和预测稳定性.

在 Stacking 方法中, 多个基础学习器的预测结果作为元学习器的新特征, 结合起来生成最终的预测. 为提升模型的泛化能力, 我们选择了三种具有代表性的元学习器:

1) 岭回归 (Ridge): 适用于线性关系的建模, 能够缓解多重共线性问题, 并且对数据中的噪声具有一定的鲁棒性[36];

2) 梯度提升回归树 (Gradient Boosting Regression Trees, GBRT): 擅长捕捉非线性关系, 能够处理复杂的特征交互, 因此在高维和非线性问题中表现优异[37];

3) 极端梯度提升 (Extreme Gradient Boosting, XGBoost): 在 GBRT 的基础上引入了优化, 具有强大的泛化能力和高效的计算能力, 特别适用于大规模数据集[38].

这三种元学习器在建模假设、复杂性和适用场景上各具特点. Ridge 回归强调鲁棒性, GBRT 擅长处理非线性关系, 而 XGBoost 则结合了高效计算与强大的泛化能力. 通过集成这些互补的学习器, 我们旨在构建一个在不同数据分布下都能保持高预测精度和稳定性的集成模型.

2 基于稀有区域识别的稀有值自助不平衡回归方法 (RareBoost)

本节首先提出了一种结合核密度估计与 K-means 聚类的 KK-means 算法, 用于识别不平衡回归任务中的稀有样本. 通过核密度估计目标变量的分布, 从而识别低密度区域; 然后, 采用 K-means 对数据进行局部划分, 精确定位稀有区间. 为了进一步提高稀有值预测的准确性, 我们基于 KK-means 识别的数据设计了 LDRE 权重函数, 在 Bagging 的自助采样过程中提高稀有值的采样概率, 增强模型对稀疏区域的关注. 通过多模型融合, 进一步提升整体性能. 最后, 通过 Stacking 元学习器集成这些具有”稀有值感知”能力的基学习器, 形成一个兼顾全局效率与局部精度的稀有值预测模型, 即 RareBoost. 整体流程如图 1 所示.

2.1 稀有区域识别 (KK-means)

2.1.1 核密度函数

在不平衡回归问题中, 稀疏区域的预测精度对整体模型性能至关重要. 为了估计目标变量的密度分布, 本文使用核密度估计方法. 核密度估计通过对数据的概率密度进行平滑估计, 能够捕捉到目标变量的局部变化, 特别是稀疏区域的分布. 核密度估计的表达式为:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{y - y_i}{h}), \tag{7}$$

其中, n为样本数, h为带宽参数, $K(\cdot)$为核函数 (通常采用高斯核), y_i为第i个观测值. 带宽h的选取对于核密度估计效果至关重要, 带宽过小会导致密度曲线过于波动 (欠平滑), 带宽过大则可能导致特征区间的结构信息被过度平滑而丢失. 极大光滑原则带宽定义如下.

2.1.2 极大光滑原则带宽

带宽参数h选用基于极大光滑原则的设定[27], 具体公式如下:

$$h = (\frac{R}{m \cdot n})^{0.2} \cdot \sigma, \tag{8}$$

其中, σ为样本标准差, n为数据集的样本数, m为倍数值, R为核函数区间$[-1, 1]$为时的积分常数, 本文

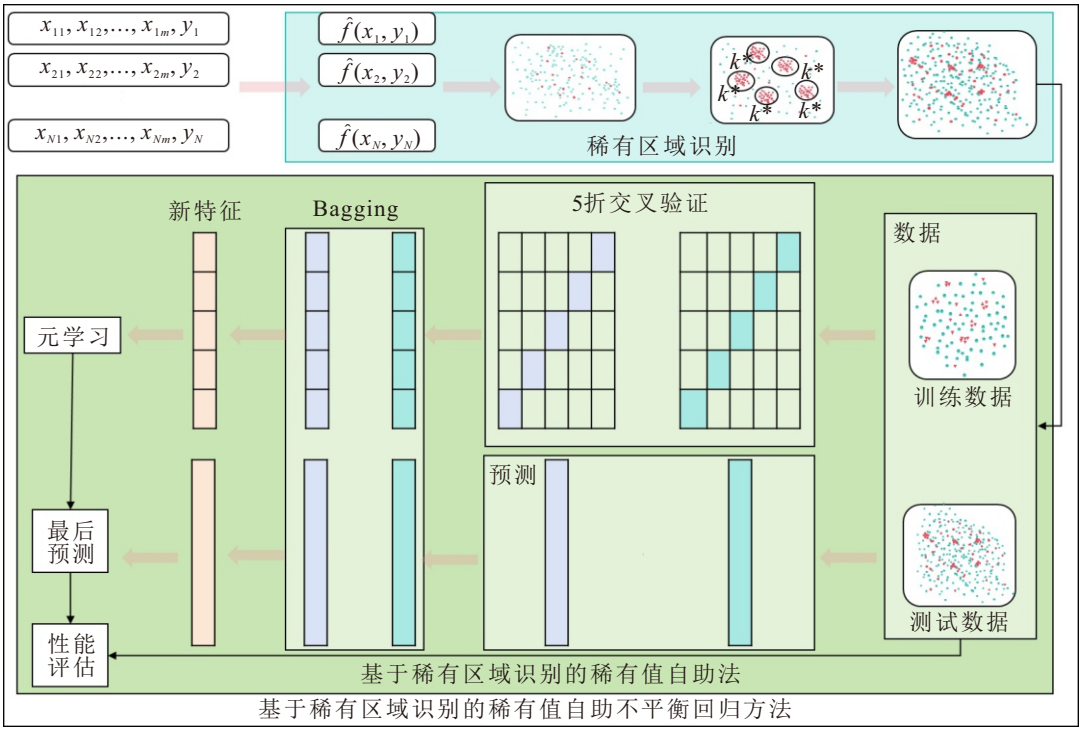


图1　RareBoost 整体流程

采用高斯核函数, 设定$R \approx 0.282$. 通过系统评估不同m取值下的模型表现, 选择m范围为$1, 2, \ldots, 10$. 通过交叉验证计算每个m对应的平均负对数似然 (ANLL) 分数, 并求出相邻m间 ANLL 变化斜率. 最终, 计算所有斜率差值的平均绝对值, 当该值降至阈值以下时, 认为该m为最优选择.

如图 2 所示, 红色虚线标出了阈值对应的m取值. 实验结果表明, 在 85% 测试数据集上, $m = 5$能够兼顾模型的拟合能力与泛化性能, 在多数回归任务中表现最为稳定. 因此, 本文后续实验均采用$m = 5$作为核密度估计带宽的默认配置.

2.1.3　稀有区域识别策略

在不平衡回归问题中, 稀有区域指的是目标变量分布中的低密度区域, 通常对应于目标变量的极端值或样本较少的区域[39],. 稀有区域的准确识别对于提高模型的预测精度至关重要, 尤其是在处理稀疏样本时, 传统回归方法往往忽视这些区域的预测. 图 3 是 KK-means 示意图.

2.1.4　算法描述

稀有区域识别方法的伪代码如算法 1 所示.

Algorithm 1　KK-means算法

1: **输入:** 目标变量样本$Y = \{y_i\}_{i=1}^{n}$; 常数$m > 0$、n、R; 候选聚类数集合$\mathcal{K} = \{k_{\min}, \ldots, k_{\max}\}$

2: 计算样本标准差$\sigma = \mathrm{sd}(Y)$、带宽h公式(8);

3: 选择核函数$K(\cdot)$, 公式(7);

4: 在$\{z_j\}_{j=1}^{M}$上进行核密度估计$\longrightarrow \hat{f}(z_j)$;

5: 构造密度点集$D_{\text{dens}} = \{(z_j, \hat{f}(z_j))\}_{j=1}^{M}$;

6: 对于每个$k \in \mathcal{K}$, 执行k-means聚类计$\longrightarrow$簇集合$\{c_j\}_{j=1}^{k}$;

7: 计算轮廓系数$S(k) \leftarrow \dfrac{1}{M}\sum_{j=1}^{M} \dfrac{b(j) - a(j)}{\max\{a(j), b(j)\}}$;

8: 选择$k^* \leftarrow \arg\max_{k \in \mathcal{K}} S(k)$;

9: 在D_{dens}上以k^*进行K-means, 得到簇中心$\{\boldsymbol{\mu}_\ell = (\bar{z}_\ell, \overline{f}_\ell)\}_{\ell=1}^{k^*}$;

10: 选择密度最低簇$\ell^\dagger \leftarrow \arg\min_\ell \overline{f}_\ell$;

11: 确定稀疏区间$[z_{\min}, z_{\max}] \leftarrow [\min_{j \in c_{\ell^\dagger}} z_j, \max_{j \in c_{\ell^\dagger}} z_j]$;

12: 计算稀疏样本索引集合$\mathcal{I}_{\text{sparse}} \leftarrow \{\, i \in \{1, \ldots, n\} \mid y_i \leqslant z_{\min} \text{ 或 } y_i \geqslant z_{\max} \,\}$;

13: **输出:** 新数据集$\mathcal{D}_S$.

2.2　基于稀有区域识别的稀有值自助法

2.2.1　标签密度比估计权重函数

在回归任务中, 标签分布通常不均衡, 极端区间样本稀缺但具有较高决策价值,. 为了应对这一问题, 本文设计了 LDRE 权重函数. 其核心目的是通过动态调整样本权重, 补偿数据分布的不均衡, 确保模型能够对稀疏区间样本给予更高的关注, 从而提高对这些关键样本的预测准确性.

LDRE 基于密度估计, 通过计算每个样本的标签密度, 并为其分配权重, 来改善训练模型的性能. 具体来说, 稀疏区间样本由于样本较少, 往往会被忽

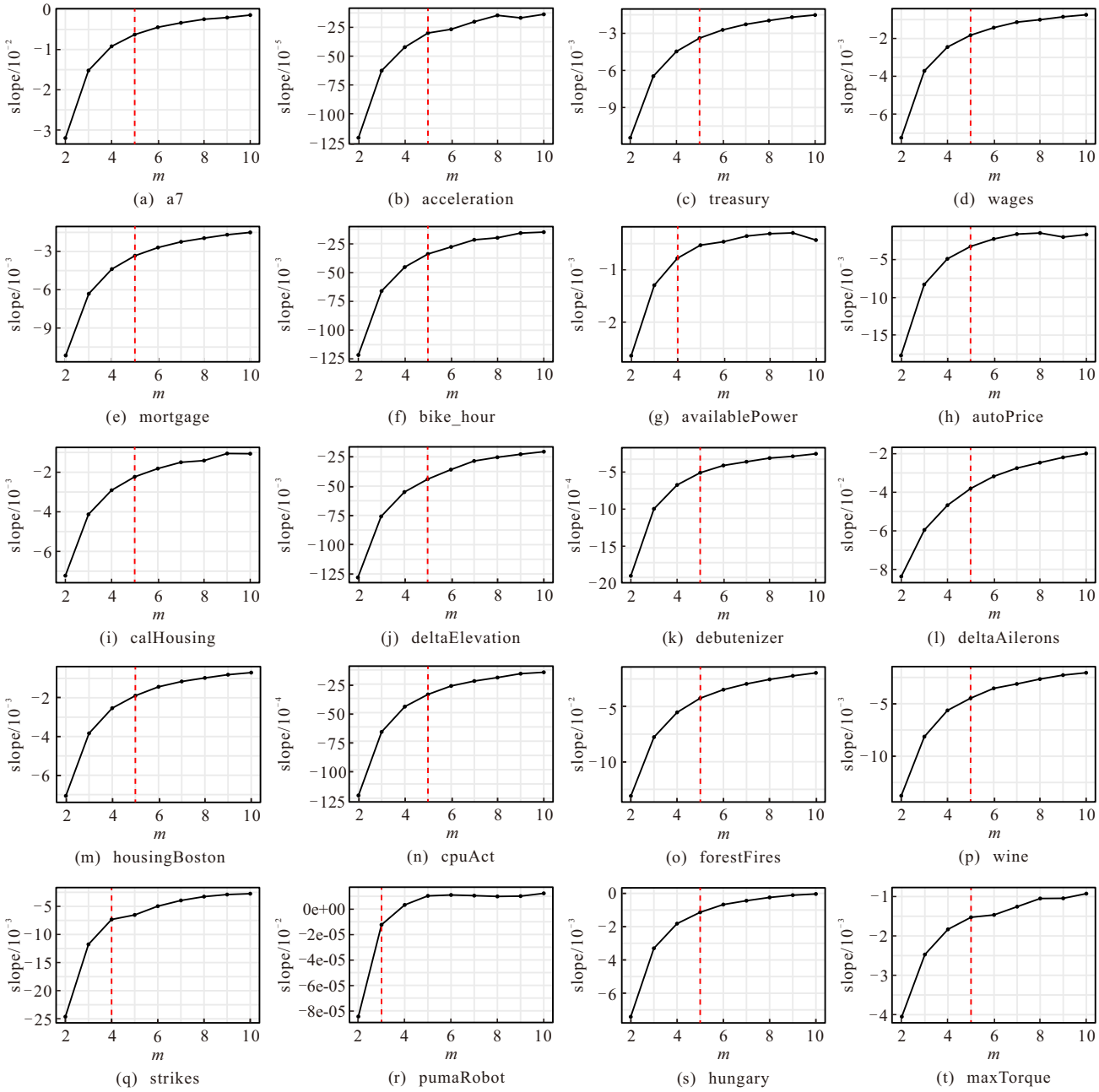


图2 ANLL 得分的斜率变化曲线

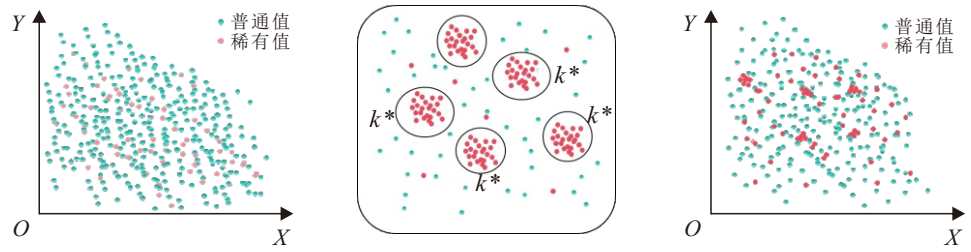


图3 稀有区域识别策略

略或低估, LDRE 通过提高这些样本的权重, 确保它们在训练中的影响力, 进而增强模型对不平衡数据的适应能力.

在标签不均衡的回归任务中, 训练数据集为 $S=\cup_{i=1}^{n}\{(x_i,y_i)\}$, 其中 n 为样本总数, $x_i\in\mathcal{X}\subset\mathbb{R}^p$ 为输入特征, $y_i\in\mathcal{Y}\subset\mathbb{R}$ 为目标变量. 标签分布通常偏斜, 部分区间样本密集, 而其他区间稀疏. 定义 $p(y)$ 为标签的经验密度分布, $p_{bal}(y)$ 为理想的平衡标签分布 (可用均匀分布 $1/(y_{max}-y_{min})$ 近似). LDRE 的表达式如下:

$$w_i = \frac{r_i}{\pi_{y_i}},\ r_i = \frac{p_{bal}(y_i)}{p(y_i)}, \tag{9}$$

其中, $p(y_i)$通过核密度估计获得, $p_{bal}(y_i)$为标签区间的均匀密度, π_{y_i}为稀有类别在数据集中的比例. 该权重设计使得稀疏区间样本权重增大, 常见区间样本权重减小, 从而缓解标签不均衡问题.

图 4 中显示了不同权重函数在LNO_2上的表现, 选择带宽计算公式 (8) 的 LDRE 权重函数最适合处理稀有极值, 灰色背景代表LNO_2的概率密度分布, 而彩色曲线则演示了加权函数$s(LNO_2)$如何根据LNO_2排放的概率密度分布来分配更高的权重给稀有极值.. 它在LNO_2的尾部区域显著增加权重, 强调了对稀有极值的关注, 这对于分析稀有事件或极端数据点非常有效. 相比之下, 逆密度权重函数 (Inverse density) 和相关函数权重 (Relevance) 在尾部区域的权重变化较小, 无法有效强化稀有极值的影响. 因此, LDRE 能够确保稀有极值在模型中占据更重要的位置, 符合处理稀有极值的目标.

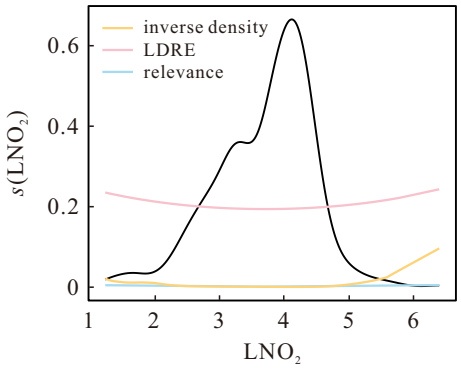
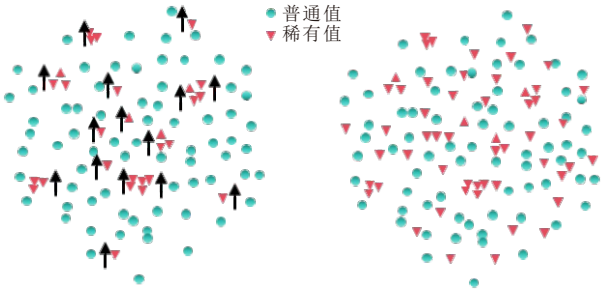


图4　不同权重函数下的LNO_2排放的概率密度分布

2.2.2　基于标签密度比估计的自助采样

自助采样是一种通过从原始数据集中随机抽取样本来生成新的训练数据集的技术. 不同于传统的随机抽样, 自助采样允许从数据集中同一样本被多次选择, 或者完全不被选择. 这种方法特别适用于集成学习算法, 用于 Bagging, 它能够通过多次采样和聚合多个模型的预测结果来提高模型的准确性和稳定性.

KK-means 稀有区域识别后的数据, 对训练数据进行自助采样. 图 5 展示了 LDRE 加权自助采样的示意图. 该图展示了如何通过调整样本权重, 增加稀有值的采样概率, 从而在训练数据中增强稀有值的表现, 进而提升模型对这些稀有样本的学习能力.



(a) 数据D_s的稀有值权重增加　　(b) 自助采样结果

图5　自助采样示意图

2.2.3　稀有值自助法

在自助采样的数据基础上训练第一级基学习器L_t, 从而得到一组预测函数$\hat{y}_i^{(t)} = L_t(x_i)$. 接着, 将基学习器输出作为新特征, 构造扩展数据集, 并在其上训练第二级学习器$\mathcal{L}$, Stacking 基于元学习器对基学习器结果汇聚输出. 最终得到的预测结果$\hat{y}$即为集成回归器的输出. 图 6 是稀有值自助法示意图.

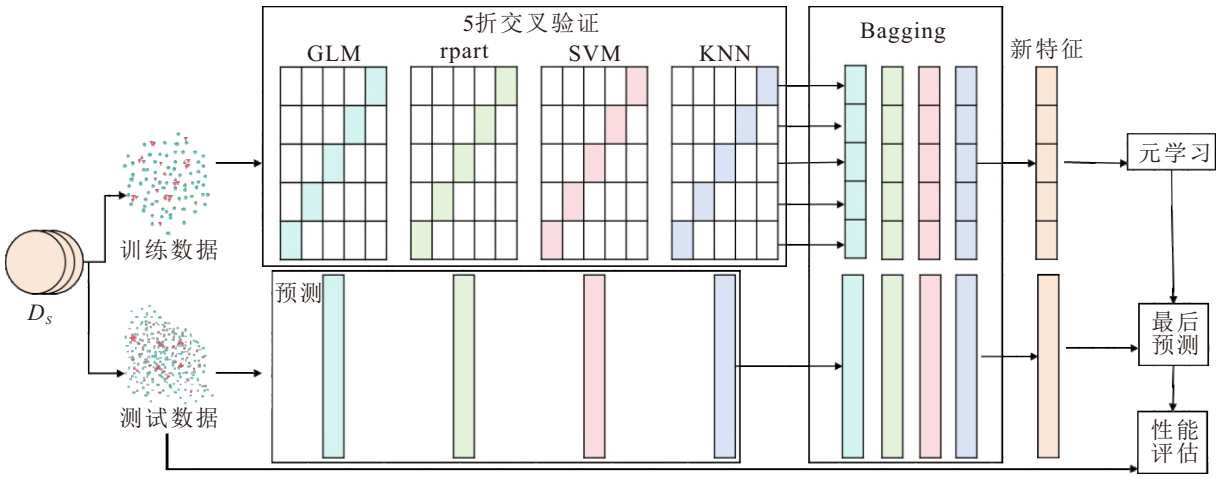


图6　稀有值自助法示意图

2.2.4　算法描述

基于稀有区域识别的稀有值自助不平衡回归方法的伪代码如算法 2 所示.

Algorithm 2　RareBoost算法

1: **输入:** 数据集$\mathcal{D}_S$; 初始化样本权重$w_i = 1$, 其中$i = 1, \ldots, n$; 自助采样子集数T, 一级基学习器族$\{L_t\}_{t=1}^{T}$, 二级学习器$\mathcal{L}$, LDRE, 轮数$B = 50$;

2: for $t = 1 \to T$ do

3: 对于$t = 1$到T

4: $\mathcal{D}_S$加权自助采样得到子集D_t, 样本的采样概率由样本权重w_i决定, 使得稀有值的样本更可能被选中

5: D_t上基于核密度估计标签密度$p_t(y)$

6: LDRE权重更新样本权重$w_i^{(t)}$

7: 更新后的、$w_i^{(t)}$将影响下一轮的自助采样

8: 自助采样得到的子集D_t训练基学习器L_t

9: 一级预测$\hat{y}_i^{(t)} = L_t(x_i)$, 其中$i = 1, \ldots, n$

10: **end for**

11: 通过Bagging汇聚一级输出, $\hat{y}_i^{(1)} \leftarrow \dfrac{1}{T}\sum_{t=1}^{T}\hat{y}_i^{(t)}$ $(i = 1, \ldots, n)$

12: 将一级预测作为新特征, 构造新的数据集$D_{\text{ext}} = \{(x_i, \hat{y}_i^{(1)})\}_{i=1}^{n}$

13: 在D_{ext}上训练二级学习器$\mathcal{L}$

14: 通过Stacking使用元学习器对基学习器结果进行汇聚

15: **输出:** 最终预测$\hat{y}$.

3 实 验

表1 实验使用的数据集

ID	Dataset	N	F	nRare	IR	Type
1	a7	160	9	7	4.58	H
2	diabetes	35	3	4	12.90	H
3	autoPrice	165	16	3	1.85	L
4	housingBoston	407	14	40	10.90	H
5	wages	429	4	1	0.23	B
6	strikes	501	7	1	0.20	H
7	mortgage	841	16	60	7.68	L
8	treasury	841	16	79	10.37	L
9	airfoild	1 203	6	11	0.92	H
10	fuelConsumption	1 413	26	27	1.95	B
11	acceleration	1 387	12	30	2.21	B
12	debutenizer	1918	8	90	4.92	H
13	space_ga	2 487	7	21	0.85	B
14	maxTorque	1 442	20	43	3.07	B
15	abalone	3 343	8	374	12.60	B
16	deltaElevation	7 615	7	1802	31.00	H
17	hungary	521	20	21	4.03	L
18	sulfur	8 065	6	606	8.12	B
19	ailerons	11 003	41	186	1.72	B
20	forestFires	416	13	7	1.71	H
21	Slump	103	9	5	19.60	L
22	Concrete	1 030	8	10	102.00	B
23	Yacht	308	6	42	6.33	H
24	kinematics8fh	6 556	9	50	0.77	B

针对所提出的基于稀有区域识别的稀有值自助不平衡回归方法, 本文使用 KK-means 稀有值区域划分后的数据, 在此相同条件下, 使用另外四个不同的权重函数来提高稀有值在自助采样中的概率, 并在集成框架中进行性能评估. 实验采用的编程语言为 R 4.4.3 版本.

3.1 实验设计及结果分析

3.1.1 数 据

在实验中, 我们从不同领域中选取了 24 个连续不平衡的数据集. 这些数据集来自 UCI 机器学习知识库和 KEEL 知识库, 用于评估机器学习算法的性能. 表 1 描述了数据集的主要特征. 对于每个数据集, 用 Tukey[40] 的箱线图规则计算稀有区间, 同时记录稀有度区间上的稀有实例数 (nRare). 该表还提供了以下信息: 实例数 (N)、特征数 (F) 和不平衡比 (IR). 最后, 我们还包括每个目标变量的不平衡类型, 如下所示: 如果调整后的箱形图仅显示低于或高于相应围栏的离群值, 则极端类型分别为低 (L) 或高 (H), 而如果它显示低于和高于围栏的离群值, 则类型均为 (B).

3.1.2 评估指标

在实验中, 选择了三个评价指标, 即 ANLL、RMSE 和 R^2, 来评估模型的性能. 计算公式如下:

$$ANLL(\hat{f}) = -\frac{1}{N}\sum_{j=1}^{N}\log \hat{f}(x_j), \tag{10}$$

其中, $\hat{f}(x_j)$表示测试样本x_j的估计概率密度, N是测试样本的数量. 注意, ANLL 越小, 估计越精确.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{12}$$

其中, y_i表示实际值, $\hat{y}_i$表示模型预测值, n为样本数量, $\bar{y}_i$表示实际值的均值. 较小的 ANLL 和 RMSE 值以及较高的R^2值表明模型的表现较好.

3.1.3 实验建立

为了保证算法的多样性, 实验中, 在基学习阶段使用使用了四种不同的机器学习算法: GLM、rpart、SVM 和 KNN. 元学习阶段使用三种不同的机器学习算法: Ridge、GBRT 和 XGBoost. 本实验的主要目的是评价稀有值集成在回归问题中的有效性. 因此, 在 24 个连续不平衡数据集使用了四种不同的权重方法: 逆密度函数 (Inverse Density, ID)、分位数尾部函数 (Quantile Tail, QT)、阈值中心函数 (Threshold

Center, TC) 和相关性函数 (PHI), 这些权重方法通过 R 实现. ID 权重函数使用 KDE 对目标变量的密度进行估算, 并为稀疏区域的样本分配较高的权重, 带宽默认大拇指法则计算. QT 权重函数通过分位数方法识别尾部样本, 并为尾部数据分配更高的权重. TC 权重函数根据样本与中位数的偏离程度分配权重, 权重越大越远离中位数. PHI 是 Torgo 提出的相关性函数$\phi : \mathcal{Y} \to [0, 1]$. ID、QT、TC 和 PHI 方法公式定义分别如下:

$$w_{ID_i} = \frac{1}{p(y_i) + \epsilon}, \tag{13}$$

$$w_{QT_i} = \begin{cases} \eta \text{ if } y_i \leqslant q_e \text{ or } y_i \geqslant q_{1-e}, \\ 1 \text{ otherwise }, \end{cases} \tag{14}$$

$$w_{TC_i} = 1 + \frac{|y_i - L|}{max(|y_i - L|) + \epsilon}, \tag{15}$$

$$w_{PHI_i} = \phi(y_i), \tag{16}$$

在上述公式中, y_i表示第i个样本的目标变量. $p(y_i)$为通过核密度估计得到的该样本输出值的概率密度估计. 为防止计算中出现除零错误, 引入极小常数ϵ进行平滑处理. 在公式 (13) 中, 权重w_{ID_i}是通过样本y_i的概率密度$p(y_i)$来计算的, 并加入ϵ作为分母平滑, 在公式 (14) 中, 权重w_{QT_i}由y_i与预设的阈值q_e和q_{1-e}比较决定. 若$y_i \leqslant q_e$或$\geqslant q_{1-e}$, 则w_{QT_i}取超参数值η(默认取 2), 否则取 1. 公式 (15) 中, 权重w_{TC_i}计算为y_i与中位数L的距离, 并通过平滑处理避免过小的值对结果的影响. 最后, 在公式 (16) 中, 权重w_{PHI_i}通过 Torgo 提出的相关性函数$\phi(y_i)$计算, 该函数基于输出值y_i来确定样本权重.

3.1.4 分析结果

表 2 至表 4 显示了不同方法在各评估指标的结果 (最佳值已加粗, 结果经四舍五入处理). 首先表 2 中, RareBoost 在 ANLL 指标上显著优于 ID、QT、TC 和 PHI 方法. 具体来看, RareBoost 的平均 ANLL 值为 2.025, 明显低于 ID(2.371)、QT(2.249)、TC (2.219) 和 PHI(2.332), 且在 24 个数据集中有 21 个取得最佳表现. 例如在数据集 6 上, RareBoost 的 ANLL 达到 0.386, 远低于其他方法 (ID: 0.734, QT: 0.554, TC: 0.555, PHI: 0.792). 即便在表现相对较弱的数据集 15 上, 其 ANLL 值仍保持较低水平, 体现了该方法在极端情况下的稳定性.

表 3 中 RMSE 指标的结果同样显示, RareBoost 显著优于 ID、QT、TC 和 PHI 方法, 其平均 RMSE 为 58.566, 低于 ID(67.541)、QT(66.709)、TC(62.164) 和 PHI(63.233), 且在 20 个数据集中表现最佳. 例如

表2　不同方法的 ANLL 值

Dataset	ID	QT	TC	PHI	RareBoost
1	6.032	6.791	5.958	6.781	5.682
2	4.305	3.568	4.225	3.583	3.246
3	4.554	4.154	3.987	4.290	3.941
4	3.628	3.535	3.850	3.969	3.385
5	3.489	3.509	3.560	3.657	3.495
6	0.734	0.554	0.555	0.792	0.386
7	0.446	0.265	0.370	0.268	−0.079
8	2.375	2.411	2.349	2.381	2.295
9	1.001	1.010	0.943	0.883	0.850
10	9.650	8.750	8.830	8.774	8.632
11	−1.149	−1.201	−1.292	−1.214	−1.334
12	−0.274	−0.488	−0.509	−0.449	−0.529
13	7.852	8.153	7.818	7.843	7.749
14	−2.618	−2.622	−2.648	−2.625	−2.644
15	2.664	2.650	2.650	2.650	2.644
16	4.408	4.638	4.624	4.659	4.343
17	−1.765	−1.713	−1.754	−1.752	−1.785
18	−7.204	−7.285	−7.285	−7.284	−7.290
19	0.738	0.711	0.711	0.714	0.708
20	2.653	2.601	2.690	2.727	2.589
21	13.150	10.824	11.371	11.211	10.416
22	5.723	6.583	5.571	6.428	5.578
23	−3.091	−3.015	−2.930	−1.917	−3.288
24	−0.400	−0.399	−0.394	−0.401	−0.404
Average	2.371	2.249	2.219	2.332	2.025

表3　不同方法的 RMSE 值

Dataset	ID	QT	TC	PHI	RareBoost
1	7.327	7.628	7.126	6.969	6.587
2	9.127	7.111	8.554	7.007	5.629
3	3.242	3.014	2.915	2.990	2.873
4	4.677	4.931	5.211	5.242	4.901
5	7.145	7.319	7.434	7.708	7.202
6	0.339	0.330	0.327	0.351	0.322
7	0.251	0.227	0.238	0.228	0.211
8	2.598	2.671	2.523	2.598	2.396
9	0.641	0.627	0.601	0.575	0.564
10	914.512	796.602	791.365	805.105	739.949
11	0.075	0.069	0.064	0.069	0.062
12	0.160	0.137	0.135	0.140	0.133
13	581.532	671.922	579.570	582.448	548.207
14	0.176	0.175	0.171	0.175	0.172
15	3.462	3.416	3.416	3.416	3.396
16	18.065	20.916	20.954	20.697	17.654
17	0.414	0.432	0.417	0.418	0.406
18	0.180	0.166	0.166	0.166	0.165
19	0.506	0.493	0.493	0.494	0.491
20	2.961	2.900	3.011	3.052	2.932
21	14.659	14.333	13.337	14.376	12.768
22	47.879	54.488	42.738	53.055	48.348
23	0.904	0.939	0.999	0.147	0.060
24	0.162	0.162	0.163	0.162	0.161
Average	67.541	66.709	62.164	63.233	58.566

表4 不同方法的R^2值

Dataset	ID	QT	TC	PHI	RareBoost
1	−0.445	−0.566	−0.367	−0.307	−0.168
2	−2.844	−1.333	−2.377	−1.266	−0.462
3	0.827	0.850	0.860	0.853	0.864
4	0.736	0.707	0.672	0.669	0.710
5	0.686	0.670	0.660	0.634	0.681
6	0.990	0.990	0.990	0.989	0.991
7	0.994	0.995	0.995	0.995	0.996
8	0.856	0.848	0.865	0.856	0.878
9	0.872	0.878	0.888	0.897	0.901
10	−0.464	−0.111	−0.096	−0.135	0.042
11	0.716	0.761	0.792	0.762	0.803
12	0.988	0.991	0.991	0.991	0.992
13	−0.228	−0.639	−0.220	−0.232	−0.091
14	0.978	0.978	0.979	0.978	0.979
15	0.496	0.509	0.510	0.510	0.515
16	0.457	0.273	0.270	0.288	0.482
17	0.510	0.466	0.503	0.500	0.530
18	0.791	0.822	0.823	0.822	0.824
19	0.835	0.844	0.844	0.843	0.845
20	0.734	0.745	0.725	0.717	0.739
21	0.967	0.968	0.972	0.968	0.975
22	0.618	0.505	0.695	0.531	0.610
23	0.988	0.987	0.985	0.968	0.995
24	0.216	0.215	0.207	0.218	0.222
Average	0.470	0.515	0.507	0.544	0.619

在数据集 2 上, RareBoost 的 RMSE 为 5.629, 较 ID、QT、TC 和 PHI 分别降低了 38.14%、20.79%、34.28% 和 19.64%. 即使在表现相对较低的数据集 12 上, 其 RMSE 也仅为 0.133, 显示出较好的稳定性.

表 4 中R^2指标的结果进一步表明, RareBoost 显著优于 ID、QT、TC 和 PHI 方法. 其平均R^2为 0.619, 高于 ID(0.470)、QT(0.515)、TC(0.507) 和 PHI(0.544), 且在 20 个数据集中表现最佳. 例如在数据集 16 上, RareBoost 的 R^2达到了 0.482, 较 ID、QT、TC 和 PHI 分别提升了 5.47%、76.69%、78.52% 和 67.36%. 这说明 RareBoost 不仅在整体上具有优势, 在特定数据集上也表现突出, 进一步验证了其处理不平衡回归任务的有效性.

3.1.5 测试与分析

本节采用 Friedman 检验[41] 与 Bonferroni-Dunn 检验[42] 对所提方法进行统计验证, 评估五种稀有值预测算法在不平衡数据上的性能差异. Friedman 检验是一种基于秩次的非参数方法, 用于判断多组算法是否存在显著性能差异. 若 P 值小于显著性水平, 则认为算法间存在显著差异. 若 Friedman 检验结果显著, 则进一步采用 Bonferroni-Dunn 检验, 判断某算法与其他算法的差异, 通过计算临界差异值 (CD) 进行判定. 其计算公式如下所示:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{17}$$

其中, k为参加对比的算法个数, N为数据集个数.

本节采用 Friedman 检验与 Bonferroni-Dunn 事后检验对五种方法的稀有值预测性能进行统计比较. 表 5 给出了 Friedman 检验的结果, 所有 P 值均小于 0.05, 说明至少有两种方法之间存在显著差异.

表5 不同方法的平均排名和 Friedman 检验结果.

Metric	ID	QT	TC	PHI	RareBoost	P	Significance
ANLL	3.833	3.375	2.958	3.708	1.125	1.847E-09	TRUE
RMSE	3.833	3.625	2.833	3.458	1.250	1.590E-08	TRUE
R^2	3.792	3.542	2.833	3.583	1.250	1.865E-08	TRUE

由于 Friedman 检验拒绝了”所有算法性能相同”的原假设 ($P < 0.05$), 进一步使用 Bonferroni-Dunn 事后检验, 识别 RareBoost 与其他四种算法 (ID、QT、TC、PHI) 之间的差异.

在本节中, 算法数$k = 5$且数据集数$N = 24$, 根据临界值表查得$\alpha = 0.05$对应的临界值$q_{\alpha=0.05} = 2.498$, 计算得到临界差异值 CD=1.245. 若两种算法的平均秩差大于 CD 值, 则认为其性能存在显著差异.

图 7 展示了多种预测方法在 ANLL、RMSE 和 R^2指标上的 Bonferroni-Dunn 检验 CD 图. 算法按平均排名从左到右排列, 平均排名越低性能越好. 红线连接性能无显著差异的算法. 在 ANLL 指标上, RareBoost 的平均秩为 1.125, 显著优于 ID、QT、TC、PHI 等方法. 同样地, 在 RMSE 和 R^2指标上, RareBoost 同样表现突出, 平均秩均为 1.250, 与其他方法存在显著差异. 虽然 ID、QT、TC、PHI 方法之间性能相近, 但 RareBoost 在三个指标上均显著优于它

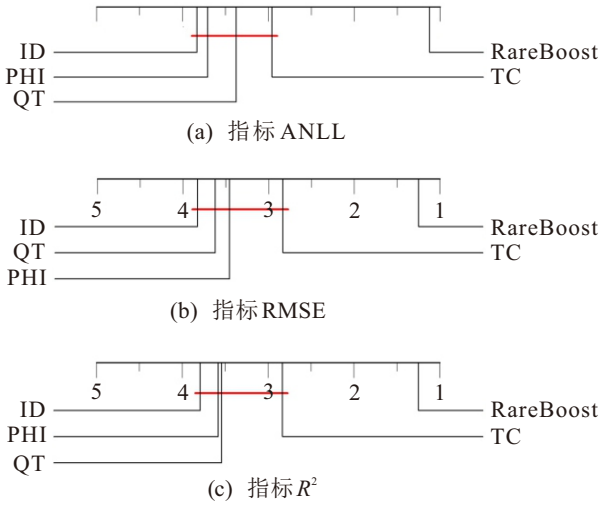


图7 在多个指标上 Bonferroni-Dunn 检验结果 CD 图

们, 表明其在预测稀有值时具有较强的竞争力. 综上, RareBoost 在 ANLL、RMSE 和R^2三个评价指标上均显著优于对比方法, 展现出明显的预测优势.

4 结 论

在许多实际应用中, 连续不平衡数据的目标变量通常分布偏斜, 导致稀疏区域和极端值的预测困难. 为此, 本文提出了基于稀有区域识别的稀有值自助不平衡回归方法 (RareBoost). 该方法结合核密度估计与 K-means 聚类, 精确识别稀有区域, 显著提高了稀有值的预测精度. 通过 LDRE 权重的自助采样加权集成方法, 进一步增强了模型对稀有样本的学习能力, 特别是在极端数据点的预测中表现突出. 实验结果表明, RareBoost 在多个公开数据集上优于传统方法, 能够平衡全局效率与局部精度, 提供了不平衡回归任务中的有效解决方案. 未来的研究可以进一步优化 RareBoost 在大规模数据集上的计算效率, 并探索其在不同场景中的泛化能力.

参考文献 (References)

[1] Branco P, Torgo L, Ribeiro R P. Relevance-based evaluation metrics for multi-class imbalanced domains[C]. Advances in Knowledge Discovery and Data Mining. Cham: Springer, 2017: 698-710.

[2] Torgo L, Ribeiro R. Utility-based regression[C]. Knowledge Discovery in Databases: PKDD 2007. Berlin, Heidelberg: Springer, 2007: 597-604.

[3] He H B, Ma Y Q. Imbalanced Learning: Foundations, Algorithms, and Applications[M]. New York: Wiley, 2013.

[4] Fernández A, García S, Galar M, et al. Learning from imbalanced data streams[M]. Learning from Imbalanced Data Sets. Cham: Springer International Publishing, 2018: 279-303.

[5] 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.

[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] López V, Fernández A, García S, et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics[J]. Information Sciences, 2013, 250: 113-141.

[8] Branco P, Torgo L, Ribeiro R P. SMOGN: A pre-processing approach for imbalanced regression[C]. First international workshop on learning with imbalanced domains: Theory and applications. PMLR, 2017: 36-50.

[9] Steininger M, Kobs K, Davidson P, et al. Density-based weighting for imbalanced regression[J]. Machine Learning, 2021, 110(8): 2187-2211.

[10] In D D, Kim H. Distance-based relevance function for imbalanced regression[J]. Stats, 2025, 8(3): 53.

[11] Avelino J G, Cavalcanti G D C, Cruz R M O. Resampling strategies for imbalanced regression: A survey and empirical analysis[J]. Artificial Intelligence Review, 2024, 57(4): 82.

[12] Alahyari S, Domaratzki M. SMOGAN: Synthetic minority oversampling with GAN refinement for imbalanced regression[J/OL]. 2025, arXiv: 2504.21152.

[13] 胡峰, 周雨龙, 苏祖强, 等. 基于网格自组织粒球模型的不平衡回归方法[J]. 控制与决策, 2025, 40(8): 2513-2524.
(Hu F, Zhou Y L, Su Z Q, et al. An imbalanced regression method based on grid self-organized granular ball model[J]. Control and Decision, 2025, 40(8): 2513-2524.)

[14] Arbib M A. The handbook of brain theory and neural networks[M]. Cambridge: The MIT Press, 2002.

[15] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[16] Kern C, Klausch T, Kreuter F. Tree-based machine learning methods for survey research[J]. Survey Research Methods, 2019, 13(1): 73-93.

[17] Ting K M, Witten I H. Issues in stacked generalization[J]. Journal of Artificial Intelligence Research, 1999, 10(1): 271-289.

[18] Khairalla M A, Ning X, Al-Jallad N T, et al. Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model[J]. Energies, 2018, 11(6): 1605.

[19] Xu Y H, Meng R T, Zhao X. Research on a gas concentration prediction algorithm based on stacking[J]. Sensors, 2021, 21(5): 1597.

[20] Zhang Y Z, Ma J, Liang S L, et al. A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets[J]. GIScience & Remote Sensing, 2022, 59(1): 234-249.

[21] Melal S R, Aminian M, Shekarian S M. A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse[J]. Journal of Agriculture and Food Research, 2024, 16: 101107.

[22] Kamalov F. Kernel density estimation based sampling for imbalanced class distribution[J]. Information Sciences: An International Journal, 2020, 512(C): 1192-1201.

[23] Kamalov F, Denisov D. Gamma distribution-based sampling for imbalanced data[J]. Knowledge-Based Systems, 2020, 207: 106368.

[24] Yan Y T, Jiang Y F, Zheng Z, et al. LDAS: Local density-based adaptive sampling for imbalanced data classification[J]. Expert Systems with Applications, 2022, 191: 116213.

[25] Zhao X R, Yi L Z, Fu G H. Stacking density estimation

and its oversampling method for continuously imbalanced data in chemometrics[J]. Chemometrics and Intelligent Laboratory Systems, 2025, 261: 105366.

[26] Branco P, Torgo L. A study on the impact of data characteristics in imbalanced regression tasks[C]. 2019 IEEE International Conference on Data Science and Advanced Analytics. Washington, 2019: 193-202.

[27] Yang G L, Zhou G H, Wang C Y, et al. A scalable thin-film defect quantify model under imbalanced regression and classification task based on computer vision[J]. Heliyon, 2023, 9(2): e13701.

[28] Torgo L, Ribeiro R. Precision and recall for regression[C]. Discovery Science. Heidelberg: Springer, 2009: 332-346.

[29] Branco P. Re-sampling approaches for regression tasks under imbalanced domains[D]. Porto: University of Porto, 2014.

[30] Ribeiro R P, Moniz N. Imbalanced regression and extreme value prediction[J]. Machine Learning, 2020, 109(9): 1803-1835.

[31] Kou Y, Fu G H. ASER: Adapted squared error relevance for rare cases prediction in imbalanced regression[J]. Journal of Chemometrics, 2023, 37(11): e3515.

[32] Hardin J W, Hilbe J M. Generalized Linear Models and Extensions[M]. The 2nd edition. College Station: Stata Press, 2007.

[33] Criminisi A, Shotton J, Konukoglu E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning[J]. Foundations and Trends in Computer Graphics and Vision, 2012, 7(2/3): 81-227.

[34] Zhang T. An introduction to support vector machines and other kernel-based learning methods[J]. Ai Magazine, 2001, 22(2): 103-104.

[35] Guo G D, Wang H, Bell D, et al. KNN model-based approach in classification[C]. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Heidelberg: Springer, 2003: 986-996.

[36] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 2000, 42(1): 80-86.

[37] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[38] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[J/OL]. 2016, arXiv: 1603.02754.

[39] 黄牛, 付光辉, 李珍珍, 等. 不平衡回归中的自适应加权采样的稀有值预测[J]. 计算机仿真, 2023, 40(8): 524-529.
(Huang N, Fu G H, Li Z Z, et al. Rare value prediction for adaptive weighted sampling in imbalanced regression[J]. Computer Simulation, 2023, 40(8): 524-529.)

[40] Chatfield C. Exploratory data analysis[J]. European Journal of Operational Research, 1986, 23(1): 5-13.

[41] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance[J]. Journal of the American Statistical Association, 1937, 32(200): 675-701.

[42] Dunn O J. Multiple comparisons among means[J]. Journal of the American Statistical Association, 1961, 56(293): 52-64.

作者简介

刘丹 (2000−), 女, 硕士生, 主要研究方向为应用统计, E-mail: 1378942532@qq.com;

付英姿 (1980−), 女, 教授, 博士, 主要研究方向为应用统计, E-mail: 1185546957@qq.com;

付光辉 (1981−), 男, 教授, 博士, 主要研究方向为生物医学统计、不平衡数据统计建模、应用统计、特征选择、统计学习、化学计量学, E-mail: ghuifu@126.com.