智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于自然邻域和数据引力的多标签不平衡数据过采样方法

刘志强, 谭浩宇, 韩奥坤, 王炜清, 严远亭, 张燕平

引用本文:
刘志强, 谭浩宇, 韩奥坤, 等. 基于自然邻域和数据引力的多标签不平衡数据过采样方法[J]. 智能系统学报, 2026, 21(3): 651-665.
LIU Zhiqiang, TAN Haoyu, HAN Aokun, et al. Multi-label imbalanced data oversampling based on natural neighborhood and data gravity[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 651-665.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505019

DOI: 10.11992/tis.202505019
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260204.0948.002

基于自然邻域和数据引力的多标签不平衡数据过采样方法

刘志强，谭浩宇，韩奥坤，王炜清，严远亭，张燕平
（安徽大学 计算机科学与技术学院，安徽 合肥 230601）

摘　要：在处理多标签不平衡数据分类问题中，过采样方法是主流技术之一。然而，如何设计有效的采样策略以捕捉样本局部分布信息，同时避免合成过程引入重叠样本而导致类间区分度降低，始终是过采样面临的关键挑战。针对该挑战，提出了一种基于自然邻域和数据引力的多标签不平衡数据过采样方法。该方法首先基于特征空间构建自然邻域结构，以自适应学习样本的局部分布信息。其次利用标签相似性来引导辅助样本选择，为相对安全的辅助样本赋予更高的权重，降低类重叠风险。最后建立数据引力模型构建动态标签分配机制，自适应生成标签信息，避免固定标签分配规则可能引发的类间冲突问题。在 14 个不平衡数据集上的实验表明，所提算法相较于 SOTA 方法在 3 个主要指标上均取得了更优的性能表现。

关键词：多标签；不平衡学习；过采样；自然邻域；数据引力；标签分配；类重叠；分类
中图分类号：TP311　　**文献标志码**：A　　**文章编号**：1673−4785(2026)03−0651−15

中文引用格式：刘志强，谭浩宇，韩奥坤，等．基于自然邻域和数据引力的多标签不平衡数据过采样方法 [J]．智能系统学报，2026, 21(3): 651–665.
英文引用格式：LIU Zhiqiang, TAN Haoyu, HAN Aokun, et al. Multi-label imbalanced data oversampling based on natural neighborhood and data gravity[J]. CAAI transactions on intelligent systems, 2026, 21(3): 651–665.

Multi-label imbalanced data oversampling based on natural neighborhood and data gravity

LIU Zhiqiang, TAN Haoyu, HAN Aokun, WANG Weiqing, YAN Yuanting, ZHANG Yanping
(College of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: In multi-label imbalanced data classification, oversampling has emerged as a mainstream technique. However, how to design effective sampling strategies that capture the local distribution information of samples while avoiding the introduction of overlapping samples during the synthesis process, and reducing the inter-class separability, remains a key challenge for oversampling methods. To this end, we propose a novel multi-label oversampling method based on natural neighborhood and data gravitation. Firstly, the method constructs adaptive natural neighborhood structures in feature space to capture local distribution information. Then, it employs label similarity to guide auxiliary sample selection, assigning higher weights to relatively safe auxiliary samples to mitigate class overlapping risk. Finally, it constructs a dynamic label assignment mechanism with the data gravitation model to generate label information, and avoiding the possible inter-class conflicts inherent in fixed label allocation rules. Experimental resultson 14 imbalanced datasets demonstrate that the proposed algorithm outperforms state-of-the-art methods in three performance metrics.

Keywords: multi-label; imbalanced learning; oversampling; natural neighborhood; data gravitation; label assignment; class overlap; classification

不平衡学习已成为机器学习、数据挖掘等领域的研究热点[1-2]。传统分类模型以最小化整体分类误差为目标，在处理不平衡数据时，模型易忽略对少数类的识别能力，限制了模型在不平衡场景下的泛化能力和分类性能[3-5]。此外，在实际应用中，少数类样本往往更值得关注。以癌症患者筛查为例，患者样本通常属于少数类，将患者误诊为健康个体的代价远比将健康个体误诊为患者更为严重[6]。因此，不平衡数据分类问题的研究不仅具有理论价值和现实意义。

收稿日期：2025−05−23.　**网络出版日期**：2026−02−04.
基金项目：国家自然科学基金项目 (62376002).
通信作者：严远亭. Email: ytyan@ahu.edu.cn.

近年来，涌现了许多针对二分类和多分类场景的不平衡数据分类方法。然而，这些方法大多属于单标签分类范畴，即每个样本仅与一个标签相关联。但在文本分类[7]、复合故障检测[8]、情绪分类[9]等应用中，样本通常与多个标签相关联，且每个标签相关的样本数量远少于不相关的样本数量[10]，从而引发了多标签不平衡问题。多标签不平衡数据的特点包括标签数量的不均衡性、标签的多样性以及标签的关联性[11]，这些特点增加了多标签不平衡数据的复杂性，相较于传统的二分类或多分类问题更具挑战。

针对多标签不平衡数据学习问题，研究者提出了多种解决方法[12]，主要分为两类：1) 重采样方法[13]，即通过调整数据空间中不同标签的样本数量来重新平衡标签分布；2) 算法自适应方法[14]，即通过修改或设计新的分类器以适应不平衡数据的特性，从而提高算法对少数类的识别能力。其中，重采样方法因其简单有效且独立于后续分类器的特点，渐渐成为了处理多标签不平衡数据集的主流策略。然而，由于多标签数据中每个样本通常与多个标签相关联，传统的二分类或多分类重采样技术难以直接应用到多标签场景[15]。因此，需要设计适用于多标签不平衡数据集的重采样方法，以有效处理多标签之间的复杂关系。

在多标签分类问题中，数据不平衡可归纳为全局和局部两种表现形式。传统方法通过对少数类样本进行随机过采样，或对多数类样本实施随机欠采样来缓解类别失衡问题[16]。这类方法虽聚焦于类别层面的样本数量均衡，但忽视了分类器决策边界优化的本质需求：即提升边界样本的表征能力[17]。针对该问题，研究者逐渐将注意力转向决策边界的处理，通过关注局部不平衡因素，提升学习性能。当前通过缓解样本的局部不平衡以提高分类性能的主流方法可分为两类：1）通过评估样本的局部不平衡程度来选择困难样本进行合成[18]。2）通过识别边界样本并设计针对性生成策略来合成样本[19]。但现有方法受限于静态规则的局部适应性不足问题，导致样本分布失真与决策边界模糊现象，其症结主要体现在以下两个层面：一是固定 k 值的 KNN(K-nearest neighbors) 算法因无法自适应局部密度分布，存在邻域尺度敏感性问题[20-21]。固定的 k 值设定会导致邻域信息捕获失真，造成合成样本的分布偏离真实数据分布情况。二是现有方法[22]在样本插值生成阶段受限于静态规则约束，缺乏对样本分布差异性的感知能力，导致不合理标签集的分配，可能引发类重叠问题。

针对上述问题，本文提出一种基于自然邻域和数据引力的多标签不平衡过采样方法。首先，基于自然邻域[20]建模自适应捕捉样本的局部分布特征；其次，设计融合全局不平衡和局部不平衡的样本权重估计方式，精准量化样本差异以优化样本选择过程；最后，基于标签相似性引导样本合成的方向和范围，结合数据引力模型动态调整标签分配，提升分类模型性能。本文的主要贡献如下：

1) 提出了一种基于自然邻域和数据引力的多标签过采样方法，突破静态规则约束，显著改善多标签不平衡问题。

2) 基于自然邻域理论构建多标签数据空间拓扑关系，提出融合全局不平衡和局部不平衡的样本权重估计方式，以选择更具代表性的合成样本。

3) 设计一种动态样本合成策略，通过标签相似性和数据引力模型指导样本合成，获得更可靠的标签信息，同时有效降低类重叠风险。

1　相关工作

多标签不平衡数据分类问题主要可分为重采样方法和算法自适应方法两大类[9]。其中，重采样方法因其独立于后续分类模型的特性，成为处理多标签类不平衡问题的主流手段。根据处理样本策略的不同，重采样方法可进一步分为欠采样和过采样方法[23]。

欠采样方法通过删除与多数类标签相关的实例来纠正类别分布倾斜。MLRUS(multi-label random undersampling)[13]通过删除携带多数类标签的样本来缓解标签间的不平衡。LPRUS(label powerset random undersampling)[13]基于 LP(label powerset) 策略[24]，将多标签问题转化为多类问题，通过随机删除最频繁标签集的实例来解决标签集间的不平衡。MLTL(multi-label Tomek link)[25]采用经典的 Tomek Link[26]欠采样算法对频繁标签进行数据清洗。MLeNN(multi-label edited nearest neighbor) [27]则利用编辑最近邻 (edited nearest neighbor, ENN) 技术识别并移除可能对分类器性能产生负面影响的样本。Liu 等[18]提出的 MLUL(multi-label undersampling based on local label imbalance) 是一种基于局部不平衡的多标签欠采样方法，其通过从多数类样本中筛选最具代表性的样本，同时保留少数类样本的完整性以实现数据再平衡。

过采样方法通过增加与少数类标签相关的实例来调整失衡的标签分布。MLROS(multi-label

random oversampling) 和 LPROS(label powerset random oversampling)[13] 通过随机复制低频少数类样本或标签集进行重采样。MLSMOTE(multi-label synthetic minority oversampling technique)[22] 采用启发式选择少数类种子样本及其邻居样本来合成新实例。MLSOL(multi-label synthetic oversampling based on local label imbalance)[18] 通过考虑所有信息标签并度量局部不平衡程度, 选择学习困难样本以合成更多样化的新实例, 解决局部区域的不平衡。LCOS(label correlation guided borderline oversampling)[19] 方法通过标签相关性识别关键边界区域, 并采用距离加权策略合成少数类样本, 改善不平衡数据分类性能。MLONC(multi-label oversampling with natural neighbor and label correlation)[28] 是一种基于自然邻域和相关性的过采样方法, 通过依赖标签自适应搜索邻居, 精准捕获标签分布特性, 优先选择决策边界附近的样本进行合成, 同时分配最相关标签以提升合成样本质量。此外, REMEDIAL(resampling multilabel datasets by decoupling highly imbalanced labels)[29] 方法通过将包含不同不平衡水平标签的复杂样本分解为两个较简单样本, 有效缓解了标签间的耦合不平衡问题。该样本重构机制可作为独立过采样策略的前置处理步骤。例如 RHwRSMT[30] 结合了 REMEDIAL 和 MLSMOTE。

2　本文方法

本文所提方法主要通过自然邻域和数据引力合成新样本, 以处理不平衡的多标签数据集。具体而言, 首先运用自然邻域算法捕获样本的分布特征。然后, 设计融合全局不平衡和局部不平衡的样本权重估计方法, 来有效评估样本的学习困难程度。最后, 通过约束合成区域的范围并基于数据引力模型自适应分配样本标签, 确保合成样本的可靠性。本文算法的框架如图 1 所示。

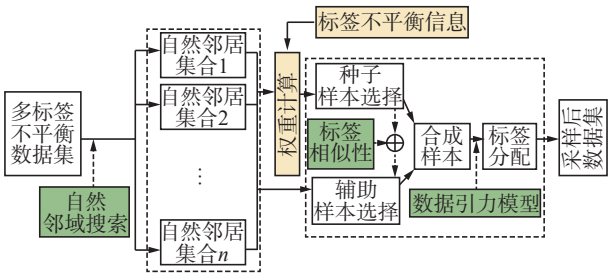


图 1　MLNNDG 算法的框架

Fig. 1　Framework of the proposed MLNNDG

为了方便介绍, 首先对本文所涉及的相关概念进行如下形式化定义: 给定一个多标签数据集 $X=\{(x_i,y_i)|1\leqslant i\leqslant n\}$, 其中 n 表示样本数量。对于任意标签 $j(j\in\{1,2,\cdots,q\})$, 其中 q 是标签总数, 定义 s_j^b 表示该标签 j 中类别 $b(b\in\{0,1\})$ 的实例数量。基于此, 引入标签内部不平衡比率的评估指标 $I_j=s_j^{G_j}/s_j^{g_j}$, 其中 $G_j=\arg\max_{b\in\{0,1\}}s_j^b$ 定义表示标签 j 内部的多数类, $g_j=\arg\min_{b\in\{0,1\}}s_j^b$ 为标签 j 内部的少数类。需要特别说明的是, 本文中所提及的多数类和少数类均特指同一标签内部的多数类和少数类。为便于理解, 本文所使用的关键符号及其含义如表 1 所示。

表 1　变量描述

Table 1　Description of variable

变量	描述	变量	描述
X	多标签数据集	$N_{\mathrm{N}}^r(x_i)$	样本x_i的r近邻集合
q	标签总数	λ	自然特征值
n	样本数量	r	迭代搜索轮次
I	标签内部不平衡比例	$N_{\mathrm{aN}}(x_i)$	样本x_i的自然邻居集合
s_j^b	标签j中类别b的实例数量	$n_{\mathrm{b}}(x_i)$	样本x_i的近邻数量
G_j	标签j内部的多数类	$\mathrm{dist}(\cdot)$	样本间欧氏距离
g_j	标签j内部的少数类	$S_k^j(x_i)$	样本x_i在标签j上同类k近邻集合
S_{noise}	噪声集合	$D_k^j(x_i)$	样本x_i在标签j上异类k近邻集合

2.1　无监督的自然邻域搜索

传统 KNN 算法采用的静态邻域搜索机制存在局限性。如图 2 所示, 当样本处于不同密度区域时, 其最优邻域尺度存在显著差异: 对于低密度区域中的样本 s_6, 为了更加精确地探测 s_6 的邻域信息以避免合成低质量样本, 需采用较小的邻域范围 ($k<5$, 绿色区域)。而处在高密度区域的样本 s_7, 则需要更大的邻域尺度 ($k>5$, 蓝色区域) 才能挖掘更完整的局部结构特征。固定不变的 k 值选择可能无法准确反映局部数据结构, 导

致生成样本的分布与实际数据分布产生偏差。

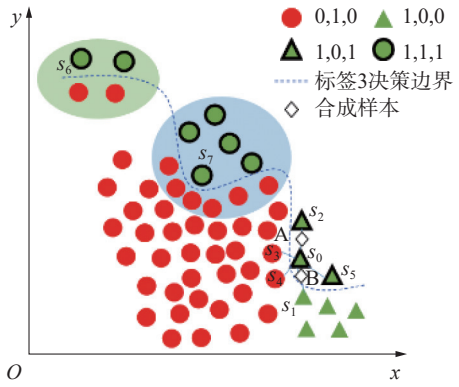


图 2　多标签局部问题示意

Fig. 2　Sketch of multi-label local problem

本文引入的自然邻域算法[20]基于整个特征空间搜索样本的局部近邻，此算法无需设置固定近邻参数且自适应捕捉数据集的局部分布特征。具体而言，对任意给定的$x_i \in X$，依次搜索其 r 近邻 (从 r=1 开始)，并且通过迭代扩展邻居搜索范围，直至满足以下条件：除异常值外，每个样本均与其他样本x_j互为邻居，此时形成稳定的自然邻域结构 (natural neighbor stable structure, NSS)。NSS 满足条件：

$$(\forall x_i)(\exists x_j)(r \in n) \wedge (x_i \neq x_j) \rightarrow (x_i \in N_{\mathrm{N}}^{r}(x_j)) \wedge (x_j \in N_{\mathrm{N}}^{r}(x_i))$$

式中$N_{\mathrm{N}}^{r}(x_i)$表示样本的x_i的 r 近邻集合。当达到稳定自然邻域结构时，迭代搜索的轮次 r 即为自然特征值 λ，其定义为

$$\lambda = r_{r \in n}\{r \mid (\forall x_i)\left(\exists x_j\right)(r \in n) \wedge \left(x_i \neq x_j\right) \rightarrow \left(x_i \in N_{\mathrm{N}}^{r}\left(x_j\right)\right) \wedge \left(x_j \in N_{\mathrm{N}}^{r}\left(x_i\right)\right)\}$$

在稳定的自然邻域结构中，样本x_i是样本x_j的自然邻居 (natural neighbor, NaN) 需满足条件：

$$x_i \in N_{\mathrm{aN}}\left(x_j\right) \Leftrightarrow x_i \in N_{\mathrm{N}}^{\lambda}\left(x_j\right) \wedge x_j \in N_{\mathrm{N}}^{\lambda}\left(x_i\right) \tag{1}$$

图 3 给出自然邻域算法的搜索流程，首先搜索每个样本的 1 近邻，如图 3 (a) 中所示，x_1和x_2互为其 1 近邻，则两个样本构成自然邻居，且近邻数增加 1，继续搜索，并依次迭代搜索次数，直到 r=4 时，除孤立样本x_6外，每个样本x_i均与至少一个样本x_j互为邻居，此时形成 NSS。如图 3(c) 中所示，此时x_1与x_2, x_3, x_4, x_5互为自然邻居，自然特征值λ=4。

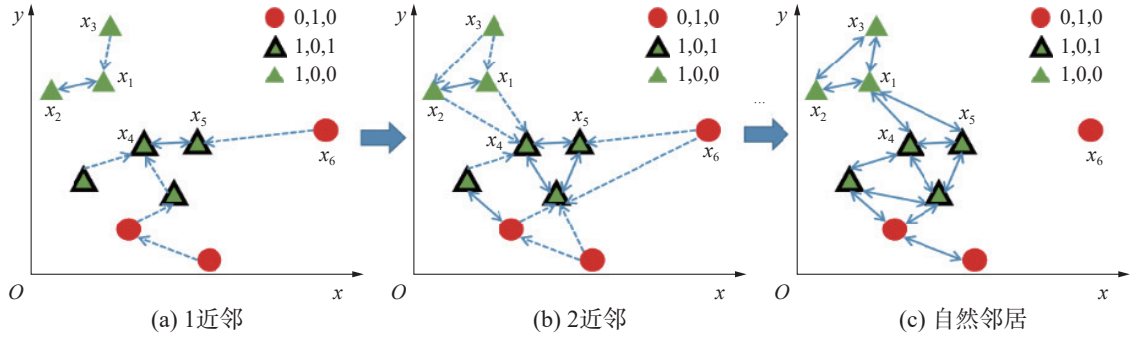


图 3　自然邻域搜索过程示意

Fig. 3　Sketch of natural neighborhood search process

自然邻域搜索的伪代码如算法 1 所示。

算法 1　自然邻域 (NaN) 搜索

输入　多标签数据集 X。

输出　自然邻域集合$N_{\mathrm{aN}}(X)$；自然特征值λ。

1) 对每个样本$x_i \in X$，初始化近邻数量 $n_{\mathrm{b}}(x_i)=0$，当前迭代次数$r=0$并创建数据集对应的 K-D 树 T；

2) 令$r=r+1$，并利用树 T 找到$x_i \in X$的r近邻x_j，更新$n_{\mathrm{b}}(x_j)=n_{\mathrm{b}}(x_j)+1$；

3) 对 $\forall x_i \in X$ 根据式 (1) 计算其自然邻域$N_{\mathrm{aN}}(x_i)$；

4) 计算统计满足$n_{\mathrm{b}}(x_i)=0$的x_i的数量 n；

5) if n 与上一轮相比未变化，$\lambda=r$；

6) else $r=r+1$，跳往 2)；

7) 输出：$\lambda, N_{\mathrm{aN}}(X)=\{N_{\mathrm{aN}}(x_i), i=1,2,\cdots,|X|\}$。

2.2　融合全局不平衡和局部不平衡的样本权重估计

针对多标签分类任务中存在的样本粒度的局部不平衡与标签粒度的全局不平衡问题，本节提出融合全局不平衡和局部不平衡的样本权重估计方法。为解决权重计算中噪声样本的干扰问题，本研究首先基于相对密度因子 (relative density factor, RDF) 构建噪声识别机制，其核心思想是通过量化样本与其同类/异类分布的密度关联来增强噪声辨识能力[31]，以克服传统自然邻域方法在噪声识别中仅通过异类近邻占比而忽略标签局部密度分布差异的局限性。具体而言，对于每个标签 j 上的少数类样本x_i，$R_{x_i}^{j}$定义为其同类局部密度和异类局部密度的比值，其值越小，越接近异类的高密度区域：

$$R_{x_i}^{j}=\frac{\sum_{p\in S_k^j(x_i)}\operatorname{dist}(x_i,p)}{\sum_{p\in D_k^j(x_i)}\operatorname{dist}(x_i,p)}$$

式中: dist(·)表示欧氏距离; $S_k^j(x_i)$定义为样本x_1在标签j上的同类k近邻集合; $D_k^j(x_i)$定义为样本x_i在标签j上的异类k近邻集合; k为近邻数, 默认取自然邻居特征值λ, 若在标签j上, 其内部少数类的样本数量小于$s_j^{g_j}$, 则取值为$s_j^{g_j}$:

$$k=\arg\min(\lambda,s_j^{g_j})$$

当$R_{x_i}^{j}<\theta$时, 判定y_{ij}为噪声标签($y_{ij}\in S_{\text{noise}}$)。阈值$\theta$与算法性能之间的关系将在实验部分3.6节进行详细讨论。

局部不平衡和全局不平衡是多标签数据集学习面临的重要挑战[18]。在计算局部不平衡前, 对任意少数类样本$x_i\in X$, 首先构建其近邻集合, 优先采用自然邻居作为局部邻域, 当自然邻居不存在时退化为λ近邻, 防止稀疏区域样本被忽略:

$$N_{\text{b}}(x_i)=\begin{cases}N_{\text{aN}}(x_i), & |N_{\text{aN}}(x_i)|\neq 0\\ N_{\text{N}}^{\lambda}(x_i), & \text{其他}\end{cases}\tag{2}$$

式中: $N_{\text{aN}}(x_i)$为样本x_i的自然邻居, $|N_{\text{aN}}(x_i)|$为自然邻居数量, $N_{\text{N}}^{\lambda}(x_i)$为样本$x_i$的$\lambda$近邻。针对每个标签$j$上的少数类样本$x_i$, 定义局部不平衡$L_{ij}$为其近邻中异类样本占比:

$$L_{ij}=\begin{cases}\dfrac{\sum_{x_m\in N_{\text{b}}(x_i)}\Delta(y_{mj}\neq y_{ij})}{|N_{\text{b}}(x_i)|}, & y_{ij}\notin S_{\text{noise}}\\ 0, & \text{其他}\end{cases}\tag{3}$$

式中$\Delta(\cdot)$为指示函数, 当样本标签不一致时取值为1, 否则为0。此外, 为了针对标签粒度的长尾分布问题, 本研究引入全局不平衡率I作为权重因子, 为每个标签j赋予全局重要性系数U_j, 以增加对罕见标签的重视程度。为了抑制极端不平衡值的影响, 做对数变换平滑数据:

$$U_j=\ln(1+I_j)\tag{4}$$

综合局部不平衡和全局不平衡特征, 提出样本级权重计算方式。具体地, 给定一个少数类样本x_i, 权重w_i由局部不平衡和全局不平衡因子共同决定:

$$w_i=\sum_{j=1}^{q}\frac{L_{ij}}{\sum_{i=1}^{n}L_{ij}}\times U_j\tag{5}$$

式中: q为标签数量, n为样本总数。对L_{ij}归一化处理则是为了消除不同标签间局部不平衡的尺度差异。样本的权重w_i越大, 表明其学习困难程度越高。因此, 在过采样过程中, 依据权重w_i采用轮盘赌法选择种子样本进行过采样。

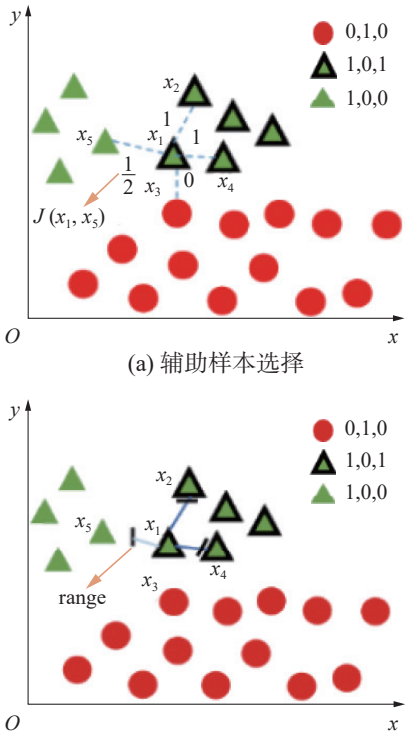
2.3　基于标签相似性的辅助样本选择机制

自然邻域算法虽然能够捕获特征空间上相互邻近的样本, 但是忽略了标签的关联性。然而选择特征空间邻近, 但标签差异显著的样本进行插值时, 易导致类别边界重叠问题。如图2所示, 以样本s_0为例, 随机选择辅助样本策略允许其与存在类间差异的样本s_3和s_4进行合成, 容易造成类别边界模糊问题。针对此不足, 本研究提出基于标签相似性的辅助样本选择机制。不同于传统的随机选择策略, 本文利用标签相似性在插值阶段实施动态空间范围约束, 有效维护分类边界的清晰性。首先, 计算样本x_i和x_j的标签Jaccard相似度$J(x_i,x_j)$, 即标签集合的交集与并集之比, 其值越大, 标签相似性越高:

$$J(x_i,x_j)=\frac{y_i\cap y_j}{y_i\cup y_j}\tag{6}$$

具体步骤为, 对选取的种子样本x_i, 计算其与邻居样本$x_j\in N_{\text{b}}(x_i)$的Jaccard相似度, 优先选择相似度高的样本作为辅助样本x_r。如图4(a)所示, 当x_1被选为种子样本时, 其邻域中x_2和x_4因具有更高的标签相似度, 优先被选为辅助样本, 而x_3则不会被选为辅助样本。然而, 当标签集不同的近邻样本进行插值合成时, 仍需施加范围约束以避免类别冲突。为此, 本节设计动态插值范围调节函数range(·):

$$\operatorname{range}(x_i,x_r)=J(x_i,x_j)\tag{7}$$



(a) 辅助样本选择

(b) 合成范围约束

图 4　辅助样本选择过程示意

Fig. 4　Sketch of auxiliary sample selection

在种子样本x_i和辅助样本x_r间进行线性插值：

$$x_{\text{new}} = x_i + \text{rand}(0,1) \times \text{range}(x_i, x_r) \times (x_r - x_i) \tag{8}$$

如图 4(b) 所示，当x_1为种子样本时，若选择x_2作为辅助样本，则range为 1，而选择x_5作为辅助样本时，则依据range缩小范围从而有效缓解合成过程中可能引发的类重叠问题。

2.4　基于数据引力的标签分配机制

传统多标签采样方法普遍采用直接复制[13]或多数投票[22]的固定分配规则，导致标签分配僵化，严重制约了合成样本标签的可靠性。如图 1 所示，以样本s_0为例，基于投票策略的传统标签分配方式，容易受多数类主导，使新样本 A、B 在合成过程中丢失少数类标签 3，从而引发少数类分布偏移。为此，本研究创新性地将万有引力定律拓展至多标签空间，提出动态标签分配机制。算法通过构建数据引力场建模标签竞争过程：将每个样本视为携带标签质量参数的“引力源”，其质量大小由对应标签的全局和局部不平衡动态决定，空间引力效应则表征样本间的竞争强度。

对于任意种子样本x_s，计算其与合成样本x_c在第 j 个标签维度产生的引力场强度$D(y_{sj}, y_{cj})$：

$$D(y_{sj}, y_{cj}) = G\frac{m_{x_s}^j m_{x_c}^j}{\text{dist}(x_s, x_c)^2}$$

式中$m_{x_s}^j$和$m_{x_c}^j$分别为样本x_s和x_c在第 j 个标签上的“质量”。为了简化公式，将引力常数 G 设置为 1。对于参与竞争的样本x_i在第 j 个标签上的“质量”$m_{x_i}^j$的构建则充分考虑标签学习的难易程度，通过融合全局不平衡系数与局部不平衡系数进行动态加权保证少数类在标签分布的优势，而待分配标签样本的“质量”$m_{x_c}^j$则假设为 1(作为竞争中标准参考点)。为建立统一量化标准，对第 j 个标签的全局不平衡系数进行归一化处理，其中$k \in \{1,2,\cdots,q\}$：

$$u_j = \frac{U_j - \min(U_k)}{\max(U_k) - \min(U_k)} \tag{9}$$

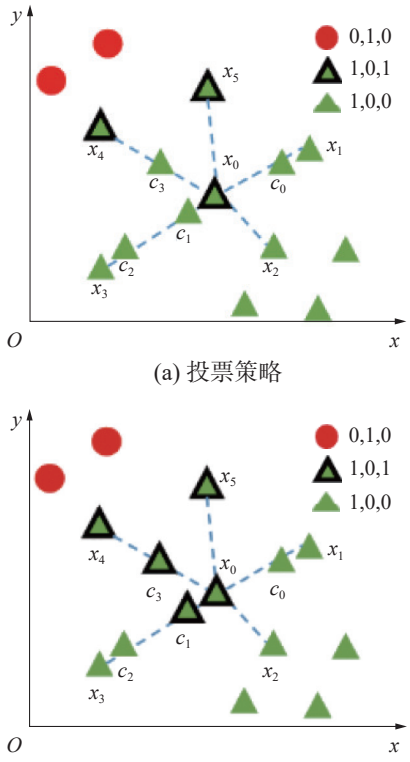
启发于 EGDRNN(entropy and gravitation based dynamic radius nearest neighbor)[32]，$m_{x_i}^j$对少数类样本采用复合权重增强，对多数类样本保持单位权重 (权重为 1，即不平衡系数均为 0)，以防止多数类主导标签分配：

$$m_{x_i}^j = \begin{cases} \left(e^{L_{ij}}\right)^2 \times (e^{u_j})^2, & y_{ij} = g_j \\ 1, & \text{其他} \end{cases} \tag{10}$$

标签分配决策机制包含 3 个核心规则：首先，若种子样本标签y_{sj}为噪声，直接继承辅助样本标签y_{cj}以防止噪声扩散；其次，当种子样本x_s与辅助样本x_r在特定标签 j 上类别一致时，合成样本直接继承该标签类别y_{sj}；最后，当两者标签的类别不同时 (以$y_{sj} = g_j$(少数类)，$y_{rj} = G_j$(多数类) 为例)，构建引力竞争方程进行判别：

$$F(y_{cj}) = G\frac{m_{x_s}^j m_{x_c}^j}{\text{dist}(x_s, x_c)^2} - G\frac{m_{x_r}^j m_{x_c}^j}{\text{dist}(x_r, x_c)^2} = \frac{\left(e^{L_{sj}}\right)^2 \times (e^{u_j})^2}{\text{dist}(x_s, x_c)^2} - \frac{1}{\text{dist}(x_r, x_c)^2} \tag{11}$$

在特定标签 j 上，当$F(y_{cj}) > 0$时，表明种子样本具有更强的引力优势，合成样本x_c继承其标签y_{sj}；反之则继承辅助样本的标签y_{rj}。特别地，当种子样本标签y_{sj}属于多数类 (此时，辅助样本标签y_{rj}为少数类)，通过样本角色交换强制少数类成为引力博弈的主动方，保证合成样本标签分配向少数类倾斜。如图 5 所示，以x_0为种子样本，c_*为合成样本，投票策略将标签 [1,0,0] 分配给所有的合成样本，而引力竞争策略则通过样本的“质量”和空间位置动态分配标签，例如c_1分配标签为 [1,0,1]，而c_2则为 [1,0,0]。



(a) 投票策略

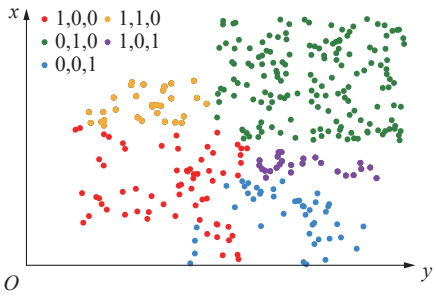
(b) 引力竞争策略

图 5　标签分配策略示意

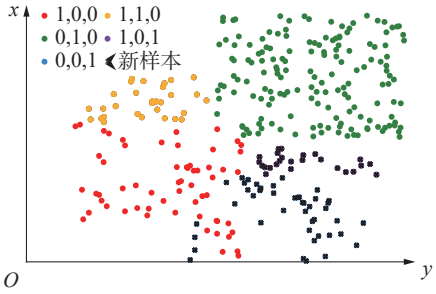
Fig. 5　Sketch of label assignment strategy

图 6(a) 是一个包含 3 个标签 (5 个类别) 的人工数据集。图 6(b)~(f) 可视化了 MLROS、MLSMOTE、MLSOL、MLONC 和 MLNNDG(multi-label oversampling with natural neighborhood and data gravity)5 种过采样方法的采样效果对比。相较于

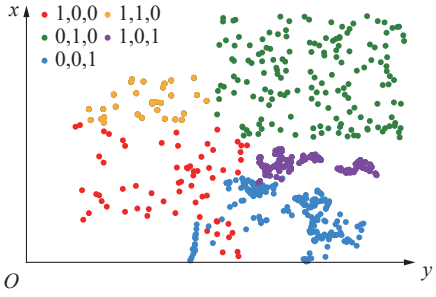
MLROS、MLSMOTE 和 MLONC 方法, MLSOL 和 MLNNDG 方法均表现出对决策边界区域的重点关注, 能够有效合成具有较高学习难度的样本。然而, MLSOL 采用的固定近邻选择策略和刚性标签分配机制, 虽然在边界区域增加了样本密度, 但容易引发类重叠现象, 甚至在部分边界区域标签分布混乱。相比之下, MLNNDG 方法通过自适应邻域搜索策略动态调整近邻范围, 在优化样本生成位置的同时降低跨类别样本干扰; 基于数据引力的标签分配机制, 既保证样本标签集分配的合理性, 又通过范围控制有效维护了类别边界的可区分性。



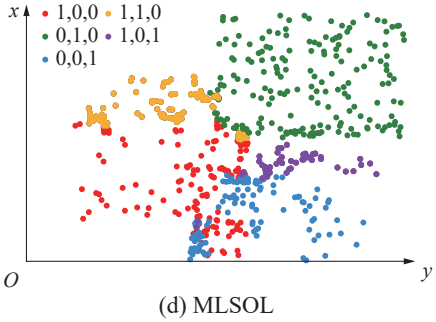
(a) DEFAULT



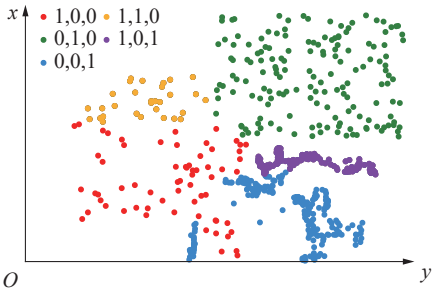
(b) MLROS



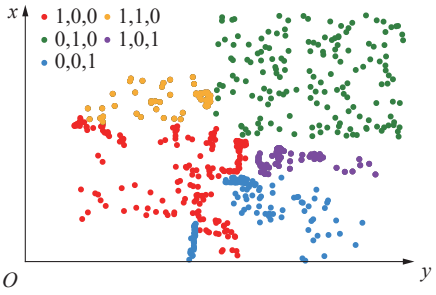
(c) MLSMOTE



(d) MLSOL



(e) MLONC



(f) MLNNDG

图 6　在二维数据集上的对比实验结果示意

Fig. 6　Sketch of comparative experimental results on 2D datasets

MLNNDG 的伪代码如算法 2 所示。

算法 2　基于自然邻居和数据引力的多标签不平衡数据过采样方法 (MLNNDG)

输入　数据集X, 过采样比例p, 邻域$N_{\mathrm{aN}}(X)$。

输出　过采样后数据集X_1。

1) $X_1 \leftarrow X$, 合成数量$n \leftarrow |X| \times p$;

2) 对标签j上$\forall x_i \in X$, 判断是否$y_{ij} \in S_{\mathrm{noise}}$;

3) 对$\forall x_i \in X$, 根据式 (2) 找到x_i的邻域$NN(x_i)$;

4) 对$\forall x_i \in X$, 根据式 (3)~(5) 计算的选择权重w_i;

5) While $n > 0$ do

6)　基于权重选择种子实例 (x_s, y_s);

7)　根据式 (6) 选择辅助样本 (x_r, y_r);

8)　根据式 (7) 和式 (8) 完成种子样本x_s和辅助样本间线性插值得到x_c;

9)　for $j = 1 \rightarrow q$ do

10)　　switch (y_{sj}的状态):

11)　　case $y_{sj} \in S_{\mathrm{noise}}$:

12)　　$y_{cj} \leftarrow y_{rj}$;

13)　　case $y_{sj} = y_{rj}$:

14)　　$y_{cj} \leftarrow y_{sj}$;

15)　　case $y_{sj} = g_j$:

16)　　交换x_s和x_r的索引;

17)　　式 (9)~(11) 计算方程$F\left(y_{cj}\right)$;

18)　　$y_{cj} \leftarrow (F(y_{cj}) > 0)\ ?\ y_{sj} :\ y_{rj}$;

19)　　End for

20)　　$X_1 \leftarrow X_1 \cup (x_c, y_c)$;

21)　　$n \leftarrow n-1$;

22)End While

2.5　时间复杂度分析

MLNNDG 算法的整体时间复杂度由算法 1 和算法 2 共同决定。根据文献 [20] 分析，算法 1 的时间复杂度为 $O(\lambda \times |X| \times \log|X|)$，其中 λ 表示自然邻居特征值。算法 2 的时间复杂度主要来源于 3 个关键步骤：噪声识别、权重计算和样本合成。具体而言，相对密度噪声识别的时间复杂度为 $O(|X|^2 \times q)$；权重计算阶段的时间复杂度为 $O(|X| \times q)$；样本合成过程包括辅助样本选择和标签分配两个子步骤，其时间复杂度分别为 $O(|X| \times p)$ 和 $O(|X| \times p \times q)$。综合上述分析，MLNNDG 算法的总体时间复杂度可近似估计为 $O(\lambda \times |X| \times \log|X| + |X|^2 \times q + |X| \times p \times q)$。

3　实验设置

3.1　数据集描述

为验证所提方法的有效性，本文从图像、音频、文本等领域选取了 14 个公开的多标签数据集 (https://cometa.ujaen.es/)，其详细信息如表 2 所列。其中，n 表示样本数量，d 表示特征维数，q 表示标签数量，Card 是每个样本的平均标签数，Dens 为基数与标签数量的比值，MeanIR 为最大标签数量与各标签数量比率的平均值，MeanImR 为每个标签下正负类样本数量比率的平均值，Scumble 则反映了每个样本中多数类和少数类标签的共现程度。从表 2 可以看出，所选数据集在各项指标上均存在显著差异，且均表现出不同程度的不平衡特性。此外，为保证实验的可比性，本文参照 MLONC[28] 的方法对数据集进行了统一的预处理。

表 2　多标签部分数据集描述

Table 2　Description of multi-label partial datasets

数据集	代称	领域	n	d	q	Card	Dens	MeanIR	MeanImR	Scumble
GnegativeGo	D1	biology	1 392	1 725	8	1.046 0	0.130 7	18.447 6	45.100 0	0.009 6
GnegativePseAAC	D2	biology	1 392	448	8	1.046 0	0.130 7	18.447 6	45.100 0	0.009 6
GpositivePseAAC	D3	biology	519	444	4	1.007 7	0.251 9	3.860 5	8.003 0	0.001 0
cal500	D4	music	502	242	174	26.043 8	0.149 7	20.577 8	22.340 0	0.337 2
emotions	D5	music	593	78	6	1.868 5	0.311 4	1.478 1	2.320 0	0.011 0
foodtruck	D6	recommend	407	33	12	2.289 9	0.190 8	7.094 5	8.975 0	0.103 5
water-quality	D7	chemistry	1 060	16	12	5.073 0	0.362 0	1.767 0	2.105 0	0.473 0
enron	D8	text	1 702	1 054	53	3.378 4	0.063 7	73.952 8	136.900 0	0.302 8
medical	D9	text	978	1 494	45	1.245 4	0.027 7	89.501 4	328.100 0	0.047 1
flags	D10	images	194	26	7	3.391 8	0.484 5	2.254 7	2.753 0	0.060 6
yeast	D11	biology	2 417	117	14	4.237 1	0.302 6	7.196 8	8.954 0	0.104 4
stackex_coffee	D12	text	225	1 886	123	1.986 7	0.016 2	27.241 5	144.900 0	0.169 1
VirusGo	D13	biology	207	749	6	1.217 4	0.202 9	4.041 2	8.615 0	0.007 9
VirusPseAAC	D14	biology	207	446	6	1.217 4	0.202 9	4.041 2	8.615 0	0.007 9

3.2　评价指标

在多标签分类性能评估指标中，Macro-average 通过计算所有标签性能指标的平均值，赋予每个标签相同的权重，而不考虑各标签样本数量差异。这一特性使其特别适用于多标签不平衡学习场景的性能评估。在类别不平衡问题中，F1(F1 score)、AUC(area under the curve)、AUCPR(area under the precision-recall curve) 是最常见的评价指标[18,28]。因此，本研究选取了 Macro-F1、Macro-AUC 和 Macro-AUCPR 这 3 个性能指标来评估方法的分类性能，指标值越高表明分类性能越优。

1)F1 是精确率和召回率的调和平均值，Macro-F1 则由各标签的 F1 值取平均得到：

$$M_{\text{acroF}_1} = \frac{1}{q}\sum_{i=1}^{q} F_1^i$$

2)AUC 是 ROC 曲线下的面积，ROC 曲线以真阳性率为纵轴，假阳性率为横轴绘制而成。Macro-AUC 则由各标签的 AUC 值取平均得到：

$$M_{\text{acroAUC}} = \frac{1}{q}\sum_{i=1}^{q} A_{\text{UC}}^i$$

3)AUCPR 是 PR 曲线下的面积, PR 曲线展示了模型在不同阈值下的精确率和召回率变化关系。Macro-AUCPR 则由各标签的 AUCPR 值取平均得到:

$$M_{\text{acroAUCPR}} = \frac{1}{q}\sum_{i=1}^{q} A^{i}_{\text{UCPR}}$$

3.3　实验设置

所有实验均通过 2 次 5 折交叉验证进行, 并报告平均结果。所选用的 8 种对比方法包括 2 种欠采样方法 MLRUS[13] 和 MLTL[25], 4 种过采样方法 MLROS[13]、MLSMOTE[22]、MLSOL[18] 和 MLONC[28], 1 种混合采样方法 RHwRSMT[30] 以及不进行任何采样的方法 DEFAULT。除此之外, 本文还考虑了 6 种不同的分类器, 分别是 BR(binary relevance)[33]、CC(classifier chains)[34]、MLKNN(multi-label K-nearest neighbors)[35]、CLR(calibrated label ranking)[36]、HOMER(hierarchy of multilabel classifiers)[37] 和 ECC(ensembles of classifier chains)[34]。对比方法和分类器的参数设置均采用原文默认设置, 此处不作赘述。

3.4　实验结果对比与分析

本节将提出的 MLNNDG 与其他 8 种对比方法在 3 种评价指标上进行比较。表 3~5 给出了 MLNNDG 与对比方法在 ECC 分类器上 Macro-F1、Macro-AUC 和 Macro-AUCPR 详细结果, 其中最优值用粗体突出显示, 次优值则用下划线显示。表格的最后两行分别统计了各方法在所有数据集上的性能均值及其平均排名。实验结果表明, 本文提出的 MLNNDG 方法在平均值和平均排名上优于所有对比方法。此外, 在 14 个数据集的性能指标评估中, 所提方法在 Macro-F1(7 项最优, 3 项次优)、Macro-AUC(6 项最优, 4 项次优) 和 Macro-AUCPR(5 项最优, 4 项次优), 共计获得 29 项最优或次优性能表现。需要指出的是, MLNNDG 在部分数据集 (如 D3、D10 等) 上的 Macro-AUC 和 Macro-AUCPR 效果性能不佳。本文认为, 这源于算法为降低重叠风险而赋予相对安全样本更高权重的选择策略, 该策略在少数类样本稀疏的场景中, 可能限制了对其增强的幅度, 进而影响了侧重正例识别的指标。表 6 则给出了各算法在 BR、CC、MLKNN、CLR、HOMER 和 ECC 分类器上取得的性能均值, 其中最优值用粗体突出显示。可观察到, MLNNDG 在 6 个分类器上的 3 种评价指标的均值都为最高。

表 3　MLNNDG 与所有对比方法在 ECC 上的 Macro-F1

Table 3　Comparison of Macro-F1 between MLNNDG and competing methods on ECC

数据集	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
D1	0.8337	0.8358	0.8286	**0.8551**	0.8292	0.8167	0.8411	0.8400	0.8511
D2	0.3338	0.3215	0.3196	0.3374	0.3116	0.2910	0.3675	0.3467	**0.4045**
D3	0.4676	0.4684	0.4903	0.4973	**0.5169**	0.4768	0.4878	0.4756	0.4953
D4	0.1099	0.1220	0.1104	0.1160	0.1279	0.0099	0.1368	0.1346	**0.1381**
D5	0.6700	0.6460	0.6478	0.6686	0.6709	0.4743	**0.6738**	0.6736	0.6708
D6	0.1668	0.1707	0.1534	0.1942	0.1265	0.0528	**0.2267**	0.2184	0.2115
D7	0.4986	0.5331	0.4899	**0.5745**	0.4996	0.1733	0.5623	0.5516	0.5596
D8	0.1253	0.1613	0.1277	0.1486	0.1131	0.0451	0.1731	0.1908	**0.2182**
D9	0.5175	0.5566	0.5319	0.5452	0.4697	0.4730	0.5618	0.5392	**0.6024**
D10	0.6507	0.6807	0.6692	0.6670	0.6652	0.4195	0.6906	0.6877	**0.7120**
D11	0.4012	0.4050	0.3909	0.4083	0.3968	0.2600	**0.4327**	0.4041	0.4281
D12	0.2416	0.2773	0.2628	0.3263	0.0401	0.0833	0.3096	0.3514	**0.3650**
D13	0.8683	0.8769	0.8949	0.8816	0.8888	0.6637	**0.9092**	0.8924	0.9071
D14	0.3508	0.3768	0.3526	0.3979	0.3317	0.2845	0.3817	0.3788	**0.3994**
平均值	0.4454	0.4594	0.4479	0.4727	0.4277	0.3231	0.4825	0.4775	**0.4974**
平均排名	6.87	5.27	6.27	3.80	6.53	8.67	2.27	3.53	**1.80**

注: 加粗代表最优结果, 横线代表次优结果。

表 4　MLNNDG 与所有对比方法在 ECC 上的 Macro-AUC

Table 4　Comparison of Macro-AUC between MLNNDG and competing methods on ECC

数据集	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
D1	0.8329	0.8415	0.8282	0.8637	0.8346	0.8369	0.8577	0.8520	**0.8798**
D2	0.3796	0.4143	0.3887	0.3887	0.3636	0.4256	0.4041	0.4336	**0.4429**
D3	0.5459	0.5331	0.5311	**0.5804**	0.5734	0.5315	0.5437	0.5398	0.5662
D4	0.1928	0.1991	0.1933	0.1952	0.1965	0.1761	0.1993	0.2012	**0.2042**
D5	0.6925	0.6793	0.6812	0.6861	0.6931	0.6304	0.6888	**0.6968**	0.6868
D6	0.2644	0.2775	0.2472	0.2857	0.2340	0.2167	0.2876	0.2889	**0.2982**
D7	0.5433	**0.5515**	0.5407	0.5335	0.5434	0.4844	0.5479	0.5496	0.5500
D8	0.1563	0.2028	0.1537	0.1837	0.1354	0.1173	0.1880	0.2163	**0.2293**
D9	0.5584	0.5832	0.5582	0.5916	0.5189	0.5491	0.5802	**0.6005**	0.5984
D10	0.7117	0.7005	0.7008	0.6901	0.6960	0.6696	0.7095	**0.7160**	0.7154
D11	0.4469	0.4514	0.4452	0.4468	0.4474	0.4061	**0.4519**	0.4462	0.4494
D12	0.4506	0.4787	0.4265	0.4869	0.0774	0.2763	0.4706	**0.5063**	0.4741
D13	0.8942	0.8940	0.8870	0.9021	0.8754	0.9118	**0.9280**	0.9140	0.9210
D14	0.4506	0.4711	0.4496	0.4765	0.4257	0.3377	**0.4779**	0.4746	0.4691
平均值	0.5086	0.5199	0.5022	0.5222	0.4725	0.4693	0.5239	0.5311	**0.5346**
平均排名	5.80	4.53	7.20	4.53	6.67	7.87	3.40	2.67	**2.27**

注: 加粗代表最优结果, 横线代表次优结果。

表 5　MLNNDG 与所有对比方法在 ECC 上的 Macro-AUCPR

Table 5　Comparison of Macro-AUCPR between MLNNDG and competing methods on ECC

数据集	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
D1	0.8329	0.8415	0.8282	0.8637	0.8346	0.8369	0.8577	0.8520	**0.8798**
D2	0.3796	0.4143	0.3887	0.3887	0.3636	0.4256	0.4041	0.4336	**0.4429**
D3	0.5459	0.5331	0.5311	**0.5804**	0.5734	0.5315	0.5437	0.5398	0.5662
D4	0.1928	0.1991	0.1933	0.1952	0.1965	0.1761	0.1993	0.2012	**0.2042**
D5	0.6925	0.6793	0.6812	0.6861	0.6931	0.6304	0.6888	**0.6968**	0.6868
D6	0.2644	0.2775	0.2472	0.2857	0.2340	0.2167	0.2876	0.2889	**0.2982**
D7	0.5433	**0.5515**	0.5407	0.5335	0.5434	0.4844	0.5479	0.5496	0.5500
D8	0.1563	0.2028	0.1537	0.1837	0.1354	0.1173	0.1880	0.2163	**0.2293**
D9	0.5584	0.5832	0.5582	0.5916	0.5189	0.5491	0.5802	**0.6005**	0.5984
D10	0.7117	0.7005	0.7008	0.6901	0.6960	0.6696	0.7095	**0.7160**	0.7154
D11	0.4469	0.4514	0.4452	0.4468	0.4474	0.4061	**0.4519**	0.4462	0.4494
D12	0.4506	0.4787	0.4265	0.4869	0.0774	0.2763	0.4706	**0.5063**	0.4741
D13	0.8942	0.8940	0.8870	0.9021	0.8754	0.9118	**0.9280**	0.9140	0.9210
D14	0.4506	0.4711	0.4496	0.4765	0.4257	0.3377	**0.4779**	0.4746	0.4691
平均值	0.5086	0.5199	0.5022	0.5222	0.4725	0.4693	0.5239	0.5311	**0.5346**
平均排名	5.80	4.53	7.20	4.53	6.67	7.87	3.40	2.67	**2.27**

注: 加粗代表最优结果, 横线代表次优结果。

表 6　MLNNDG 与对比方法在 6 种分类器上的性能指标均值

Table 6　Mean performance metrics of MLNNDG and competing methods on six classifiers

分类器	性能指标	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
BR	Macro-F1	0.4355	0.4536	0.4337	0.4610	0.4092	0.3061	0.4663	0.4695	**0.4808**
	Macro-AUC	0.6674	0.6703	0.6669	0.6723	0.6463	0.6370	0.6833	0.6804	**0.6863**
	Macro-AUCPR	0.4237	0.4292	0.4196	0.4338	0.3949	0.3852	0.4381	0.4402	**0.4432**
CC	Macro-F1	0.4425	0.4567	0.4362	0.4639	0.4167	0.3254	0.4634	0.4662	**0.4787**
	Macro-AUC	0.6650	0.6712	0.6664	0.6716	0.6487	0.6260	0.6833	0.6748	**0.6873**
	Macro-AUCPR	0.4225	0.4303	0.4207	0.4338	0.3951	0.3807	0.4370	0.4354	**0.4411**
MLKNN	Macro-F1	0.3475	0.3919	0.3421	0.4079	0.3499	0.2467	0.4299	0.4202	**0.4394**
	Macro-AUC	0.7278	0.7306	0.7276	0.7325	0.7117	0.7138	0.7363	0.7360	**0.7407**
	Macro-AUCPR	0.4715	0.4798	0.4697	0.4855	0.4520	0.4556	0.4852	0.4851	**0.4941**
CLR	Macro-F1	0.4431	0.4556	0.4352	0.4619	0.4068	0.3138	0.4696	0.4702	**0.4821**
	Macro-AUC	0.7558	0.7580	0.7544	0.7603	0.7168	0.7420	0.7646	0.7644	**0.7705**
	Macro-AUCPR	0.5097	0.5138	0.5038	0.5150	0.4577	0.5085	0.5160	0.5203	**0.5233**
HOMER	Macro-F1	0.4419	0.4502	0.4403	0.4515	0.4146	0.3944	0.4624	0.4632	**0.4665**
	Macro-AUC	0.6485	0.6572	0.6591	0.6547	0.6418	0.6416	0.6647	0.6674	**0.6734**
	Macro-AUCPR	0.4111	0.4168	0.4133	0.4165	0.3916	0.4112	0.4246	0.4260	**0.4288**
ECC	Macro-F1	0.4454	0.4594	0.4479	0.4727	0.4277	0.3231	0.4825	0.4775	**0.4974**
	Macro-AUC	0.7393	0.7443	0.7387	0.7486	0.7124	0.6965	0.7554	0.7591	**0.7622**
	Macro-AUCPR	0.5086	0.5199	0.5022	0.5222	0.4725	0.4693	0.5239	0.5311	**0.5346**

注: 加粗代表最优结果。

为了更清晰地观察不同方法的最优和次优结果的差距, 遵循于 MLCIO(a novel ensemble oversampling approach based Chebyshev inequality for imbalanced multi-label data)[38] 中的实验方法, 将最佳结果的权重设置为 1, 次优结果的权重设置为 0.7, 以验证各实验方法的性能, 例如表 7 中 MLNNDG 在 Macro-F1 上计算结果: 58.5=41×1+25×0.7。表 7 给出了各方法的排名权重以及最优/次优个数对比, 可以看出: 本方法在所有评估指标上均表现最佳, 无论是最优和次优结果数量还是权重都是最高的。

为进一步对比算法的性能, 本文将 8 种对比算法分别与 MLNNDG 在不同分类器上进行了 Wilcoxon 符号秩检验, 显著性水平选取为 0.05。表 8 给出每种方法的平均排名, 以及与其他方法对比, 在 3 个评估指标上的显著胜/负次数。实验结果表明, MLNNDG 在所有评价指标中均取得最高平均排名, 并且对比其他算法, 本算法获得了最多统计显著优势且未出现统计显著劣化现象。

表 7　各采样方法排名权重和最优/次优个数对比

Table 7　Comparison of ranking weights and best/second-best counts across sampling methods

指标	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
Macro-F1	0.7 (0/1)	5.2 (1/6)	0.0 (0/0)	15.3 (9/9)	1.0 (1/0)	2.0 (2/0)	29.6 (17/18)	30.5 (13/25)	**58.5 (41/25)**
Macro-AUC	4.1 (2/3)	4.1 (2/3)	3.1 (1/3)	14.0 (7/10)	2.4 (1/2)	7.5 (4/5)	30.6 (18/18)	25.7 (11/21)	**53.4 (38/22)**
Macro-AUCPR	2.7 (2/1)	5.1 (1/6)	3.4 (2/2)	17.0 (10/10)	4.5 (1/5)	13.0 (13/0)	25.1 (16/13)	32.1 (16/23)	**41.8 (25/24)**

注: 加粗代表最优结果。

表 8　各采样方法排名以及成对 Wilcoxon 符号秩检验显著胜/负次数

Table 8　Rankings of sampling methods and number of significant wins/losses in pairwise Wilcoxon Signed-Rank tests

指标	分类器	DEFAULT	MLROS	MLRUS	MLSMOTE	MLTL	RHwRSMT	MLSOL	MLONC	MLNNDG
Macro-F1	BR	6.47(1/5)	4.60(4/3)	7.07(1/5)	4.00(4/1)	7.47(1/6)	8.20(0/8)	2.93(5/1)	2.73(5/1)	**1.40(8/0)**
	CC	5.93(2/5)	4.20(4/1)	6.73(1/5)	3.60(4/1)	7.33(1/6)	8.93(0/8)	3.47(4/1)	3.07(4/1)	**1.60(8/0)**
	MLKNN	6.67(1/5)	4.67(4/3)	7.33(1/5)	3.93(4/2)	6.80(1/5)	8.27(0/8)	2.33(6/0)	2.93(5/0)	**2.07(6/0)**
	CLR	6.40(2/5)	4.73(4/3)	6.87(1/5)	4.20(4/1)	7.40(1/6)	7.87(0/8)	2.87(5/0)	2.80(5/1)	**1.80(7/0)**
	HOMER	6.20(1/3)	5.13(3/3)	6.27(1/4)	4.40(2/1)	7.13(0/5)	7.80(0/7)	3.07(5/0)	2.80(5/0)	**2.13(6/0)**
	ECCRU	6.87(1/5)	5.27(3/4)	6.27(1/5)	3.80(5/2)	6.53(1/4)	8.67(0/8)	2.27(6/0)	3.53(5/1)	**1.80(7/0)**
	平均(总和)	6.42(8/28)	4.77(22/17)	6.76(6/29)	3.99(23/8)	7.11(5/32)	8.29(0/47)	2.82(31/2)	2.98(29/4)	**1.80(42/0)**
Macro-AUC	BR	6.00(2/3)	5.33(2/3)	5.80(2/3)	4.60(2/2)	7.47(0/7)	7.40(0/7)	2.87(6/0)	3.73(5/1)	**1.67(7/0)**
	CC	6.00(2/5)	4.73(3/2)	5.87(1/2)	4.53(3/2)	7.53(0/6)	7.40(0/7)	2.47(6/0)	4.27(3/1)	**1.87(7/0)**
	MLKNN	5.07(1/1)	5.20(2/1)	5.93(1/2)	4.13(2/1)	7.27(0/7)	7.00(0/4)	4.07(2/0)	3.47(3/0)	**2.60(6/0)**
	CLR	5.40(1/3)	4.73(2/2)	5.80(1/3)	5.13(1/1)	8.00(0/7)	6.53(0/4)	3.60(4/0)	3.20(5/0)	**2.40(6/0)**
	HOMER	6.53(0/5)	5.13(3/2)	4.93(2/3)	5.33(0/2)	6.80(0/4)	7.20(0/5)	3.60(4/0)	3.13(6/0)	**2.20(6/0)**
	ECCRU	6.00(2/3)	5.53(2/3)	6.27(2/3)	4.87(2/3)	7.27(0/7)	8.13(0/7)	2.80(6/0)	2.27(6/0)	**1.87(6/0)**
	平均(总和)	5.83(8/20)	5.11(14/13)	5.77(9/16)	4.77(10/11)	7.39(0/38)	7.28(0/34)	3.24(28/0)	3.35(28/2)	**2.10(38/0)**
Macro-AUCPR	BR	6.00(2/4)	5.07(4/3)	6.60(0/5)	4.33(3/0)	7.27(0/6)	7.20(0/6)	3.13(5/0)	2.73(5/0)	**2.53(5/0)**
	CC	6.13(2/4)	4.47(4/1)	6.13(2/5)	4.00(4/0)	7.80(0/7)	6.87(0/7)	3.20(4/0)	3.93(3/0)	**2.33(5/0)**
	MLKNN	5.13(2/1)	5.07(2/1)	6.00(0/3)	4.67(3/1)	6.93(0/6)	7.60(0/6)	3.80(2/0)	3.13(3/0)	**2.40(6/0)**
	CLR	6.07(1/3)	5.00(1/2)	6.40(0/3)	4.93(1/0)	7.40(0/6)	5.33(0/0)	3.93(3/0)	3.07(4/0)	**2.73(4/0)**
	HOMER	6.07(0/3)	4.53(1/2)	5.60(0/3)	5.00(1/1)	7.53(0/5)	6.67(0/1)	4.00(3/0)	3.27(4/0)	**2.20(6/0)**
	ECCRU	5.80(2/4)	4.53(3/2)	7.20(1/6)	4.53(4/1)	6.67(0/5)	7.87(0/7)	3.40(4/1)	2.67(5/0)	**2.27(7/0)**
	平均(总和)	5.87(9/19)	4.78(15/11)	6.32(3/25)	4.58(16/3)	7.27(0/35)	6.92(0/27)	3.58(21/0)	3.13(14/0)	**2.40(33/0)**

注: 加粗代表最优结果, (n_1/n_2)表示显著性检验中获得的胜/负次数。

3.5　消融实验

本文所提方法包含 3 个模块: 自然邻域算法提取局部信息, 基于标签相似性选择辅助样本和基于数据引力模型分配标签。为验证每个模块的有效性, 进行了消融实验, 主要涉及以下方法。

A: 基准方法 (不进行过采样处理)。

B: 不使用自然邻域, 使用 KNN(k = 5) 捕捉局部信息。

C: 不考虑相似性, 随机选择辅助样本进行合成。

D: 不使用引力模型, 采用投票策略分配标签。

E: 完整方法 (MLNNDG)。

表 9 给出了各方法在 6 个基分类器和 14 个数据集上 Macro-F1、Macro-AUC、Macro-AUCPR 的平均排名。实验结果表明, 完整方法 E 在所有评估指标上均取得最优排名, 这充分证明了 MLNNDG 通过合理的模块整合, 有效发挥了每个模块的优势。具体而言, 方法 B 与 E 的实验结果对比表明, 自然邻域算法相比于固定近邻值的 KNN 方法, 能够自适应搜索邻域, 从而更准确地捕捉局部信息。同时, E 在所有指标上的均优于 C 和 D 的结果, 这一结果有利证实了本文所提出的基于标签相似性的辅助样本选择机制和基于引力模型的标签分配策略的有效性。

表 9　消融实验结果平均排名对比

Table 9　Average ranking comparison of ablation study results

指标	A	B	C	D	E
Macro-F1	4.75	2.93	2.08	3.18	**2.05**
Macro-AUC	4.29	2.93	2.45	3.09	**2.14**
Macro-AUCPR	4.34	2.68	2.69	2.99	**2.26**

注: 加粗代表最优结果。

3.6　参数分析

本节主要研究 MLNNDG 算法中过采样率

p和噪声识别阈值θ对模型性能的影响。实验选用性能表现最佳的 ECC 分类器进一步详细探讨参数敏感性。在 14 个数据集上, 通过 Macro-AUC 指标评估参数敏感性, 设置过采样率$p\in\{0.1,0.3,0.5,0.7,0.9\}$以控制少数类样本的生成规模; 噪声识别阈值$\theta$则依据噪声的相对密度分布特征[31] (噪声样本的相对密度远低于平均值), 设置为$\theta=\mu-k\sigma$(其中μ、σ表示少数类相对密度的均值和标准差, $k\in\{1,2,3\}$)。

表 10 列出了不同采样比 (p) 和不同噪声阈值 (θ) 下的 Macro-AUC 性能结果, 随着p的增大, MLNNDG 的性能逐渐上升, 表明适度地过采样能够有效缓解类不平衡问题。然而, 当$p=0.9$时, 大多数数据集上的模型性能开始下降, 这可能是由于过度合成样本导致类分布扭曲, 从而损害了分类性能, 而具体数据集最佳采样比存在差异。另一方面, 当$\theta=\mu-3\sigma$时, 算法在 14 个数据集上的 Macro-AUC 达到最优, 验证了其对噪声样本的有效筛选。综合来看, 当$\theta=\mu-3\sigma$, $p=0.7$时, 相应的 Macro-AUC 最高。

表 10　不同采样比 (p)/噪声阈值 (θ) 对 Macro-AUC 性能指标的影响
Table 10　Impact of sampling ratio (p) / noise threshold (θ) on Macro-AUC

数据集	p				θ			
	0.1	0.3	0.5	0.7	0.9	$\mu-3\sigma$	$\mu-2\sigma$	$\mu-\sigma$
D1	0.9507	0.9501	0.9497	0.9453	**0.9567**	**0.9567**	0.9504	0.9420
D2	0.7923	0.7986	0.8173	0.8175	**0.8273**	**0.8273**	0.8219	0.8269
D3	0.7412	0.7640	**0.7784**	0.7652	0.7427	0.7784	0.7779	**0.7869**
D4	0.5376	0.5415	**0.5430**	0.5325	0.5393	0.5430	0.5433	**0.5436**
D5	0.8365	0.8288	0.8210	0.8310	**0.8411**	**0.8411**	0.8305	0.8356
D6	0.5524	0.5772	0.5833	**0.5960**	0.5746	**0.5960**	0.5948	0.5895
D7	0.7099	**0.7103**	0.7099	0.7078	0.7091	0.7103	0.7118	**0.7165**
D8	0.6084	0.6301	0.6403	**0.6524**	0.6454	**0.6524**	0.6422	0.6484
D9	0.8365	0.8613	0.8756	**0.8795**	0.8708	**0.8795**	0.8726	0.8625
D10	0.7274	**0.7507**	0.7499	0.7436	0.7400	**0.7507**	0.7429	0.7492
D11	0.6590	0.6579	0.6576	**0.6646**	0.6605	0.6646	**0.6717**	0.6625
D12	0.7707	0.7714	0.7818	**0.7898**	0.7766	**0.7898**	0.7887	0.6533
D13	0.9669	**0.9738**	0.9718	0.9704	0.9694	0.9738	**0.9742**	0.9717
D14	0.6839	0.6704	0.6870	**0.7068**	0.7048	**0.7068**	0.7001	0.6992
平均值	0.7410	0.7490	0.7548	**0.7573**	0.7542	**0.7622**	0.7588	0.7491

注: 加粗代表最优结果。

4　结束语

针对多标签数据集中的类不平衡问题, 本文提出了一种基于自然邻域和数据引力的过采样方法 (MLNNDG)。通过借助自然邻域捕获的局部信息, 继而在决策边界附近合成低频标签来增强少数类样本的可见性, 并引入标签相似性和数据引力模型来灵活分配标签, 从而获得更可靠的标签信息。在 14 个多标签不平衡数据集上的对比实验表明, MLNNDG 显著提高了分类模型在 Macro-F1、Macro-AUC 和 Macro-AUCPR 上的性能。在未来工作中, 如何拓展思路以应对更复杂的挑战值得深入研究, 如将自然邻域算法拓展至动态流数据环境, 以处理标签分布随时间演变的不平衡问题等。

参考文献:

[1] YAN Yuanting, ZHENG Zhong, ZHANG Yiwen, et al. CPS-3WS: a critical pattern supported three-way sampling method for classifying class-overlapped imbalanced data[J]. Information sciences, 2024, 676: 120835.

[2] 严远亭, 马迎澳, 任艳平, 等. 基于构造性神经网络与全局密度信息的不平衡数据欠采样方法[J]. 计算机科学, 2023, 50(10): 48−58.

YAN Yuanting, MA Ying'ao, REN Yanping, et al. Imbalanced undersampling based on constructive neural network and global density information[J]. Computer science, 2023, 50(10): 48−58.

[3] 范洪旗, 严远亭, 张以文, 等. 学习困难与泛化能力感知的软件缺陷预测过采样方法[J]. 计算机集成制造系统, 2024, 30(8): 2663−2671.
FAN Hongqi, YAN Yuanting, ZHANG Yiwen, et al. Software defect prediction oversampling technique with generalization and difficulty-aware[J]. Computer integrated manufacturing systems, 2024, 30(8): 2663−2671.

[4] 徐贞顺, 郑顺国, 苏梦瑶, 等. 基于双自适应约束的对抗学习过采样方法[J/OL]. 计算机科学, 1−16[2026−02−10]. https://link.cnki.net/urlid/50.1075.tp.20251219.1013.002.
XU Zhenshun, ZHENG Shunguo, SU Mengyao, et al. Adversarial learning with dual adaptive constraints for oversampling method[J/OL]. Computer Science, 1−16[2026−02−10]. https://link.cnki.net/urlid/50.1075.tp.20251219.1013.002.

[5] YAN Yuanting, ZHU Yuanwei, LIU Ruiqing, et al. Spatial distribution-based imbalanced undersampling[J]. IEEE transactions on knowledge and data engineering, 2023, 35(6): 6376−6391.

[6] LILOGLOU T, MALONEY P, XINARIANOS G, et al. Sensitivity and limitations of high throughput fluorescent microsatellite analysis for the detection of allelic imbalance: application in lung tumors[J]. International journal of oncology, 2000, 16(1): 5−19.

[7] JIANG Ting, WANG Deqing, SUN Leilei, et al. LightXML: transformer with dynamic negative sampling for high-performance extreme multi-label text classification[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(9): 7987−7994.

[8] HE Zengxiang, CHU Pengpeng, LI Chenxi, et al. Compound fault diagnosis for photovoltaic arrays based on multi-label learning considering multiple faults coupling[J]. Energy conversion and management, 2023, 279: 116742.

[9] HAN Meng, WU Hongxin, CHEN Zhiqiang, et al. A survey of multi-label classification based on supervised and semi-supervised learning[J]. International journal of machine learning and cybernetics, 2023, 14(3): 697−724.

[10] AHMADI Z, KRAMER S. A label compression method for online multi-label classification[J]. Pattern recognition letters, 2018, 111: 64−71.

[11] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International journal of data warehousing and mining (IJDWM), 2007, 3(3): 1−13.

[12] KOZIARSKI M. Potential Anchoring for imbalanced data classification[J]. Pattern recognition, 2021, 120: 108114.

[13] CHARTE F, RIVERA A J, DEL JESUS M J, et al. Addressing imbalance in multilabel classification: measures and random resampling algorithms[J]. Neurocomputing, 2015, 163: 3−16.

[14] ZHANG Minling, LI Yukun, YANG Hao, et al. Towards class-imbalance aware multi-label learning[J]. IEEE transactions on cybernetics, 2022, 52(6): 4459−4471.

[15] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014, 26(8): 1819−1837.

[16] NARESHPALSINGH J M, MODI H N. Multi-label classification methods: a comparative study[J]. International research journal of engineering and technology, 2017, 4(12): 263−270.

[17] SÁEZ J A, KRAWCZYK B, WOŹNIAK M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets[J]. Pattern recognition, 2016, 57: 164−178.

[18] LIU Bin, BLEKAS K, TSOUMAKAS G. Multi-label sampling based on local label imbalance[J]. Pattern recognition, 2022, 122: 108294.

[19] ZHANG Kai, MAO Zhaoyang, CAO Peng, et al. Label correlation guided borderline oversampling for imbalanced multi-label data learning[J]. Knowledge-based systems, 2023, 279: 110938.

[20] ZHU Qingsheng, FENG Ji, HUANG Jinlong. Natural neighbor: a self-adaptive neighborhood method without parameter K[J]. Pattern recognition letters, 2016, 80: 30−36.

[21] 冯骥, 张程, 朱庆生. 一种具有动态邻域特点的自适应最近邻居算法[J]. 计算机科学, 2017, 44(12): 194−201.
FENG Ji, ZHANG Cheng, ZHU Qingsheng. Adaptive nearest neighbor algorithm with dynamic neighborhood[J]. Computer science, 2017, 44(12): 194−201.

[22] CHARTE F, RIVERA A J, DEL JESUS M J, et al. MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation[J]. Knowledge-based systems, 2015, 89: 385−397.

[23] ARYUNI M, FATICHAH C, YUNIARTI A. Resampling methods for imbalanced datasets in multi-label classification: a review[C]//2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics. Penang: IEEE, 2024: 472−477.

[24] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multilabel classification[C]//Machine Learning: ECML 2007. Berlin: Springer, 2007: 406−417.

[25] PEREIRA R M, COSTA Y M G, SILLA C N Jr. MLTL: a multi-label approach for the Tomek Link undersampling algorithm[J]. Neurocomputing, 2020, 383: 95−105.

[26] 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.

[27] CHARTE F, RIVERA A J, DEL JESUS M J, et al. MLeNN: a first approach to heuristic multilabel undersampling[C]//Intelligent Data Engineering and Automated Learning. Cham: Springer, 2014: 1−9.

[28] LIU Bin, ZHOU Ao, WEI Bingkun, et al. Oversampling multi-label data based on natural neighbor and label correlation[J]. Expert systems with applications, 2025, 259: 125257.

[29] CHARTE F, RIVERA A J, DEL JESUS M J, et al. Dealing with difficult minority labels in imbalanced mutilabel data sets[J]. Neurocomputing, 2019, 326: 39−53.

[30] CHARTE F, RIVERA A J, DEL JESUS M J, et al. REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization[J]. Neurocomputing, 2019, 326: 110−122.

[31] 许茂龙, 姜高霞, 王文剑. 基于异常检测的标签噪声过滤框架[J]. 计算机科学, 2024, 51(2): 87−99.
XU Maolong, JIANG Gaoxia, WANG Wenjian. Label noise filtering framework based on outlier detection[J]. Computer science, 2024, 51(2): 87−99.

[32] WANG Zhe, LI Yanqiong, LI Dongdong, et al. Entropy and gravitation based dynamic radius nearest neighbor classification for imbalanced problem[J]. Knowledge-based systems, 2020, 193: 105474.

[33] BRINKER K, FÜRNKRANZ J, HÜLLERMEIER E. A unified model for multilabel classification and ranking[C]//17th European Conference on Artificial Intelligence. Riva del Garda: IOS Press, 2006: 489−493.

[34] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification[J]. Machine learning, 2011, 85(3): 333−359.

[35] ZHANG Minling, ZHOU Zhihua. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038−2048.

[36] FÜRNKRANZ J, HÜLLERMEIER E, LOZA MENCÍA E, et al. Multilabel classification via calibrated label ranking[J]. Machine learning, 2008, 73(2): 133−153.

[37] SECHIDIS K, TSOUMAKAS G, VLAHAVAS I. On the stratification of multi-label data[C]//Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2011: 145−158.

[38] REN Weishuo, ZHENG Yifeng, ZHANG Wenjie, et al. A novel ensemble over-sampling approach based Chebyshev inequality for imbalanced multi-label data[J]. Neurocomputing, 2025, 612: 128717.

作者简介:

刘志强, 硕士研究生, 主要研究方向为机器学习、数据挖掘。E-mail: 1040921276@qq.com。

谭浩宇, 硕士研究生, 主要研究方向为机器学习、软件缺陷预测。E-mail: 2151476673@qq.com。

严远亭, 教授, 博士生导师, 博士, 主要研究方向为机器学习、数据挖掘。主持国家自然科学基金面上项目 1 项、国家自然科学基本青年项目 1 项, 发表学术论文 40 余篇。E-mail: ytyan@ahu.edu.cn。

[责任编辑: 丁钰]