智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种基于KNN和随机仿射的边界样本合成过采样方法

冷强奎, 孙薛梓, 孟祥福

引用本文:
冷强奎, 孙薛梓, 孟祥福. 一种基于KNN和随机仿射的边界样本合成过采样方法[J]. 智能系统学报, 2025, 20(2): 329-343.
LENG Qiangkui, SUN Xuezi, MENG Xiangfu. A borderline sample synthesis oversampling method based on KNN and random affine transformation[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(2): 329-343.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202311038

DOI: 10.11992/tis.202311038
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250121.1627.005

一种基于 KNN 和随机仿射的边界样本合成过采样方法

冷强奎, 孙薛梓, 孟祥福
(辽宁工程技术大学 电子与信息工程学院, 辽宁 葫芦岛 125105)

摘　要: 过采样是处理不平衡数据分类问题的有效策略。本文提出了一种基于 K 近邻 (K-nearest neighbor, KNN) 和随机仿射的边界样本合成过采样方法, 用于改进现有过采样方法的种子样本选择阶段和合成样本生成阶段。首先, 引入三近邻理论, 建立样本间有效的内在近邻关系, 并去除数据集中的噪声, 以降低后续分类器的过拟合风险。其次, 准确识别那些难以学习且包含丰富信息的少数类边界样本, 并将其用作采样种子。最后, 利用局部随机仿射代替线性插值机制, 在原始数据的近似流形中均匀地生成合成样本。相比于传统过采样方法, 本文方法能更充分挖掘数据集中的重要边界信息, 从而为分类器提供更多辅助以改善其分类性能。在 18 个基准数据集上, 与 8 种经典采样方法 (结合 4 种不同分类器) 进行了大量对比实验。结果表明, 本文所提方法获得了更高的 F_1 分数和几何均值 (G-mean), 可以更为有效地解决不平衡数据分类问题。此外, 统计分析也证实该方法具有更高的弗里德曼排名 (Friedman ranking)。

关键词: K 近邻; 线性插值; 边界样本; 自然分布; 过采样; 三近邻理论; 随机仿射变换; 不平衡分类
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)02−0329−15

中文引用格式: 冷强奎, 孙薛梓, 孟祥福. 一种基于 KNN 和随机仿射的边界样本合成过采样方法 [J]. 智能系统学报, 2025, 20(2): 329–343.
英文引用格式: LENG Qiangkui, SUN Xuezi, MENG Xiangfu. A borderline sample synthesis oversampling method based on KNN and random affine transformation[J]. CAAI transactions on intelligent systems, 2025, 20(2): 329–343.

A borderline sample synthesis oversampling method based on KNN and random affine transformation

LENG Qiangkui, SUN Xuezi, MENG Xiangfu
(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: Oversampling is a proven strategy for addressing imbalanced data classification challenges. This paper introduces a borderline sample synthesis oversampling method based on K-nearest neighbor (KNN) and random affine transformation to improve both the seed sample selection stage and synthetic sample generation stages of existing oversampling methods. Initially, the three nearest neighbor theory is applied to establish an effective intrinsic neighborhood relationship between samples and remove noise from the dataset. This step helps reduce the risk of overfitting by subsequent classifiers. Next, the minority-class borderline samples that are difficult to learn but contain rich information are accurately identified and treated as sampling seeds. Finally, the method replaces traditional linear interpolation with local random affine transformation, uniformly generating synthetic samples within the approximate manifold of the original data. Compared with traditional oversampling methods, the proposed method more effectively leverages important borderline information within datasets, thereby enhancing classifier performance. Extensive comparative experiments were conducted on 18 benchmark datasets, comparing the proposed method against 8 classic sampling methods, each combined with 4 different classifiers. The results show that this method achieves higher F_1 scores and geometric means (G-mean), addressing the imbalanced data classification problem more effectively. Furthermore, statistical analysis confirms that the method has a higher Friedman ranking.

Keywords: K-nearest neighbor; linear interpolation; borderline sample; natural distribution; oversampling; three nearest neighbor theory; random affine transformation; imbalanced classification

类间数据不平衡是指少数类中的样本数量远小于多数类中的样本数量[1]。这种偏斜分布在许多现实应用中普遍存在, 如医疗诊断[2]、信用风险评估[3] 和软件故障预测[4] 等。传统分类器会产生

收稿日期: 2023−11−24.　**网络出版日期**: 2025−01−22.
基金项目: 国家自然科学基金青年项目 (61602056); 国家自然科学基金面上项目 (61772249); 辽宁省教育厅项目 (JYTMS20230819); 辽宁工程技术大学博士科研启动基金项目 (21-1043).
通信作者: 冷强奎. E-mail: qkleng@126.com.

对多数类的归纳偏差，这是因为少数类对优化目标函数的贡献较小。然而，从学习的角度来看，少数类通常代表一种更关键的模式，值得更多关注[5]。因此，提高对少数类的预测能力已成为不平衡数据分类中的核心问题。

不平衡数据分类问题的解决方案可以分为4类[6-7]，即数据级方法、算法级方法、混合级方法和深度学习方法。在这些解决方案中，数据级方法特别是过采样技术被广泛采纳，因为它可以被视为预处理步骤，并且是独立于分类器的[8]。

随机过采样 (random oversampling, ROS)[9] 是最早的过采样技术，旨在通过对少数类样本的随机复制来平衡类分布。但是 ROS 容易出现过拟合的问题，因为它会放大噪声对分类器的影响。在 2002 年，Chawla 等[10] 提出了经典的合成少数类过采样技术 (synthetic minority oversampling technique, SMOTE)。它通过在原始少数类样本之间进行线性插值来生成新的合成样本。SMOTE 能够使决策区域更泛化，以此来缓解随机复制引起的过拟合问题。但是，它没有考虑数据分布，并且对每个少数类样本均等对待，容易生成无用样本。

由于许多分类器易于从边界样本中学习预测模型，因此生成具有代表性的边界样本至关重要[11]。按照这种思路，许多针对 SMOTE 的改进方法被提出，如 Borderline SMOTE[12]、ADASYN(adaptive synthetic sampling)[13] 等。这些方法为边界处的少数类样本分配更高的权重，以便它们有更大的机会被过采样。

从本质上讲，每种基于 SMOTE 的方法都可以分解为 2 个阶段[14]，即种子样本选择阶段和合成样本生成阶段。在种子样本选择阶段，通常有 K 近邻 (K-nearest neighbor, KNN) 技术的参与，但不同 K 值的选择会使过采样方法表现出显著的不适定性。以 Borderline-SMOTE 为例，当 K 取不同值时，它会将同一个少数类样本判定为不同组 (“噪声” “危险” “安全”) 的成员。在合成样本生成阶段，通常有线性插值机制的参与，但线性插值机制会将合成样本限制在原始样本之间的连线上，这就导致合成样本是噪声或位于被噪声破坏的样本之间，从而增加了数据集中两类样本间的重叠[15-16]。

针对上述问题，本文提出了一种基于 KNN 和随机仿射的边界样本合成过采样方法，用于改进过采样方法的 2 个阶段。首先，引入三近邻理论，建立了样本间的有效近邻关系。此外，去除了数据集中的少数类噪声，以降低后续分类器的过拟合风险。然后，构建边界少数类集 S_{imin}，并认为 S_{imin} 中的样本是最难以学习但最具信息性的，以用作采样种子。最后，通过局部随机仿射的方式来代替线性插值机制，对数据空间底层局部数据分布的平均值进行更精确的估计，在边界少数类集 S_{imin} 的近似数据流形中均匀地生成合成样本。

1　相关工作

1.1　合成少数类过采样技术 (SMOTE)

SMOTE[10] 在相邻的少数类样本之间进行线性插值来生成新的合成样本以平衡类分布。具体地说，假设随机选择少数类样本 x 作为采样种子，通过计算 x 到每个少数类样本间的欧氏距离，得到 x 的 K 个近邻。根据过采样率 N%，在其 K 近邻中随机选择 N 个样本并标记为：$y_1, y_2, \cdots, y_N$，然后在 x 和 y_i (i=1, 2, …, N) 之间进行线性插值，并生成一个新样本 g：

$$g = x + (y_i - x) \times \alpha \tag{1}$$

式中 α 为 [0,1] 内的随机数。由式 (1) 可知，g 将位于 x 和 y_i 之间的连线上。SMOTE 虽然克服了随机过采样容易出现的过拟合问题，但它假设了一个同质的少数类簇，并且在生成新样本时不考虑邻域中的多数类样本。当少数类样本为多聚簇分布时，SMOTE 会增加类间的重叠，从而使分类问题更加复杂。

1.2　种子样本选择方案的最新研究进展

从 SMOTE 出发，许多学者将研究重心放在种子样本的选择方案上。Han 等[12] 提出的 Borderline-SMOTE 利用 K 近邻技术确定位于分类边界处的少数类样本，并仅对这些少数类样本进行过采样，旨在强调分类边界的重要性。He 等[13] 提出的 ADASYN 对少数类样本赋予不同的权重，并根据权重分布自适应地生成数量不同的合成样本，旨在将分类边界移向那些难于学习的少数类样本。Koziarski 等[17] 提出的 RBU(radial-based undersampling) 利用互类势的概念来确定每个多数类样本的效用，并按照互类势递减的顺序进行欠采样，旨在解决传统欠采样方法容易丢失重要信息的问题。陶佳晴等[18] 利用 Tomek 链标识位于分类边界处的少数类样本，并在边界样本及其少数类近邻间进行过采样，旨在提高现有方法识别边界样本的准确率。Leng 等[19] 提出的 NanBDOS (natural neighbor based borderline oversampling) 利用自然近邻为每个少数类样本分配动态的采样权

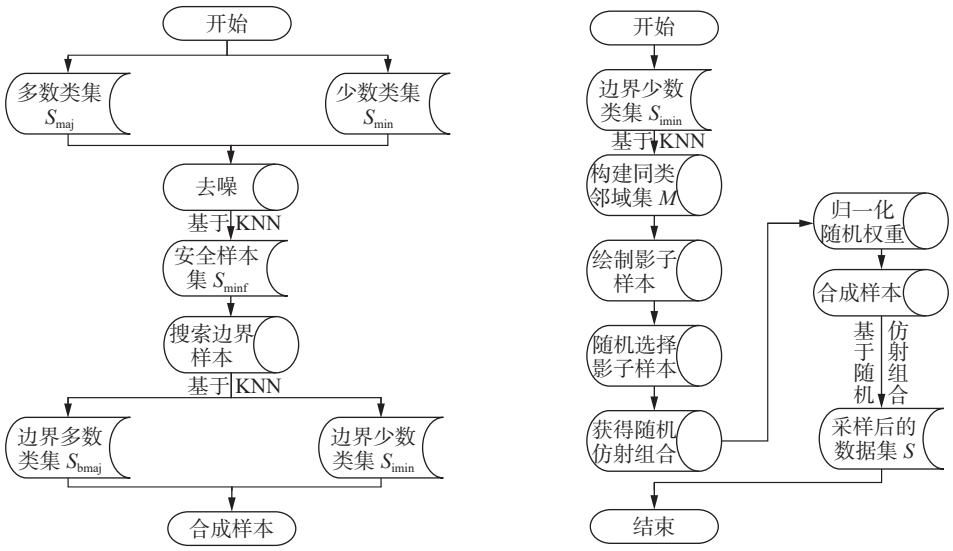
重, 并在少数类样本与其自然近邻之间进行线性插值来生成合成样本, 旨在维持过采样后数据的原始分布。Wei 等[5] 提出的 IR-SMOTE(improved random SMOTE) 利用核密度估计技术计算 K 均值聚类后每个簇的少数类密度, 并根据密度自适应地为每个簇分配不同的采样权重, 旨在使生成的合成样本具有多样性。

1.3　合成样本生成方案的最新研究进展

最近, 一些学者将过采样方法研究转向合成样本生成方案上。He 等[20] 提出的 HS-Gen(hyper-sphere-constrained generation mechanism) 在少数类样本周围的超球体区域内生成合成样本, 并且还能够动态调整超球体的大小防止生成噪声, 旨在增加合成样本的随机性和多样性。Bellinger 等[21] 提出的 ManSO(manifold-based synthetic over-sampling) 利用自编码器框架将少数类样本映射到流形上, 从而逼近少数类空间的潜在数据流形。Koziarski 等[22] 提出的 RBO(radial-based over-sampling) 利用互类势的概念将少数类区域划分为边界区域和安全区域, 并利用小步长在安全区域内生成更多的合成样本, 旨在增强分类边界的可分离性。Douzas 等[23] 提出的 G-SMOTE(geometric SMOTE) 在每个少数类样本的周围选择一个安全半径, 并在截断的超球体内生成合成样本, 旨在拓展 SMOTE 的数据生成机制。Ye 等[24] 提出的 LeO(Laplacian eigenmaps oversampling) 利用拉普拉斯特征映射来找到一个最优维空间, 并在最优维空间中构造拉普拉斯算子来生成合成样本, 旨在避免过采样后生成大量噪声。Bej 等[25] 提出的 LoRAS(localized random affine shadowsamplings) 利用正态分布绘制影子样本来构造少数类样本的随机仿射线性组合, 并在其近似数据流形中实现过采样。

2　本文方法

本文所提出的过采样方法包含 2 个阶段: 种子样本选择阶段和合成样本生成阶段, 其总体流程如图 1 所示。在种子样本选择阶段, 三近邻理论被引入用以标识重要的边界少数类样本, 并且数据中的噪声也被删除。在合成样本生成阶段, 使用局部随机仿射来代替线性插值机制, 并在边界少数类样本的近似数据流形中生成合成样本。需要说明的是, 本文方法仅针对二分类任务。接下来, 将对其进行详细描述。



(a) 种子样本选择阶段　　(b) 合成样本生成阶段

图 1　本文方法流程

Fig. 1　Flow chart of the proposed method

2.1　构建边界少数类集 S_{imin}

本文提出了一种全新的方法来识别重要的少数类样本, 并构建边界少数类集 S_{imin}。首先解释涉及样本 x 的几个术语。

最近邻集 $N_N(x)$　由 x 的 k_1 个最近邻组成的集合。

多数类最近邻集 $N_{maj}(x)$　由 x 的 k_2 个多数类最近邻组成的集合。

少数类最近邻集 $N_{min}(x)$　由 x 的 k_3 个少数类最近邻组成的集合。

图 2 给出了针对上述 3 种集合的解释。当 $k_1=k_2=k_3=3$ 时, $N_N(A)=\{C, D, R\}$, $N_{maj}(A)=\{P, Q, R\}$, $N_{min}(P)=\{B, C, D\}$。与 $N_{maj}(A)$ 和 $N_{min}(P)$ 均为同类样本不同, $N_N(A)$ 中会存在 2 类样本的混合。

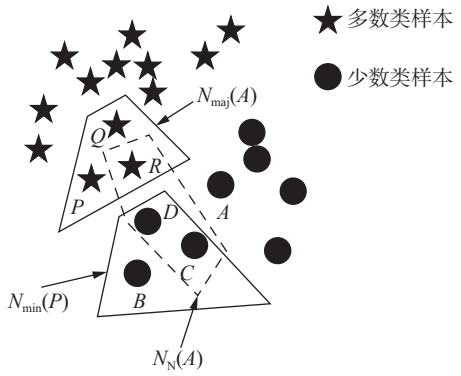


图 2　$N_N(A)$、$N_{maj}(A)$ 和 $N_{min}(P)$ 的相关解释

Fig. 2　Interpretation of $N_N(A)$, $N_{maj}(A)$ and $N_{min}(P)$

如果少数类样本中包括噪声，那么将对过采样进程产生不利影响，甚至导致新噪声样本的产生。为了解决这一问题，首先对少数类样本集进行去噪处理，并确定少数类安全样本集 S_{minf}。具体过程如下：如果某一少数类样本 x_i 的最近邻集 $N_N(x_i)$ 中只有多数类样本，则将 x_i 视为噪声并从数据集中移除。如图 3 所示，当 k_1=3 时，少数类样本 A 和 B 的 3 个最近邻均为多数类样本，则 A 和 B 被视为噪声并从少数类样本集中移除，其余样本构成少数类安全样本集 S_{minf}。

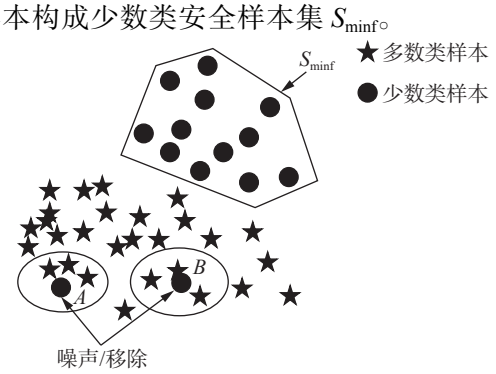


图 3　去噪并构建安全样本集 S_{minf}

Fig. 3　Noise removal and construction of the safe sample set S_{minf}

接下来，利用三近邻理论来确定边界样本。对于每个少数类样本 $x_i \in S_{minf}$，构建其多数类最近邻集 $N_{maj}(x_i)$。为了使 $N_{maj}(x_i)$ 中的样本位于分类边界附近，此时近邻参数 k_2 的值不宜设置过大，通常 k_2=3 是一个合理的选择。通过计算所有 $N_{maj}(x_i)$ 的并集，得到边界多数类集 S_{bmaj}，即 $S_{bmaj}=\cup N_{maj}(x_i)$, $x_i \in S_{minf}$。图 4 给出了 S_{bmaj} 的示例。

然后，对于每个多数类样本 $y_i \in S_{bmaj}$，构建其少数类最近邻集 $N_{min}(y_i)$，并将 $N_{min}(y_i)$ 中的所有样本组合起来构建边界少数类集 S_{imin}。考虑到 S_{imin} 中需要包含那些难以学习但具有丰富信息的少数类样本，因此近邻参数 k_3 的值不宜设置过小，应该与 S_{minf} 中样本个数存在一定的比例关系。图 5 给出了 S_{imin} 的示例。

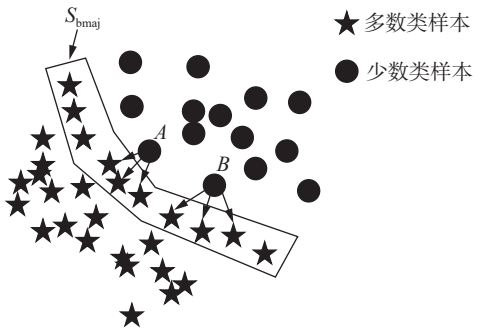


图 4　构建边界多数类集 S_{bmaj}

Fig. 4　Construction of the borderline majority class set S_{bmaj}

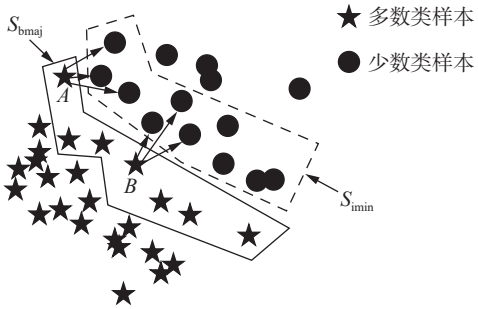


图 5　构建边界少数类集 S_{imin}

Fig. 5　Construction of the borderline minority class set S_{imin}

需要特别说明的是，当少数类样本彼此非常接近且远离多数类样本时，就会出现少数类近邻集中没有多数类样本的情况 (如图 6 所示)。但这并不会导致边界多数类集 S_{bmaj} 和边界少数类集 S_{imin} 为空。以边界多数类集 S_{bmaj} 为例，它是 S_{minf} 中每个少数类样本的 k_2 个多数类最近邻的并集，而不是任意少数类样本最近邻中的 k_2 个多数类样本的并集。由图 6 可以看出，只要数据集中有 k_2 个多数类样本，边界多数类集 S_{bmaj} 就不会因少数类样本的位置或分布而变为空。同样的解释也适用于边界少数类集 S_{imin}。

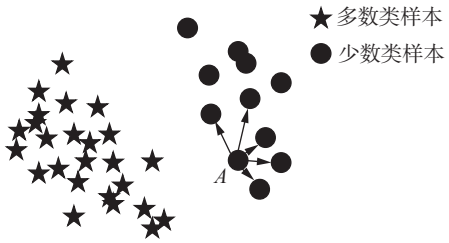


图 6　样本 A 的 K 近邻

Fig. 6　K-nearest neighbors of A

2.2　利用局部随机仿射合成新样本

本文利用局部随机仿射的方式来代替线性插值机制，在边界少数类样本周围的小区域内生成影子样本，并使用多个影子样本的凸组合来生成合成样本，旨在克服线性插值机制的缺陷。在说明本文合成方式之前，首先给出以下 4 个定义。

定义 1　将任意向量的随机仿射组合定义为

这些向量的随机仿射线性组合, 并且这些线性组合的系数是随机选择的。具体来说, 给定向量 $\boldsymbol{v}$, 且 $\boldsymbol{v}=a_1\boldsymbol{u}_1+a_2\boldsymbol{u}_2+\cdots+a_n\boldsymbol{u}_n$。如果 $a_1+a_2+\cdots+a_n=1$, $a_j\in\mathbf{R}^+$, 且 $a_1,a_2,\cdots,a_n$ 是从狄利克雷 (Dirichlet) 分布中随机选取的仿射组合系数。那么, 向量 $\boldsymbol{v}$ 就是向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n$, $\boldsymbol{u}_j\in\mathbf{R}^{|F|}$的一个随机仿射线性组合。

定义 2　给定一个类或数据集有 n 个样本和 $|F|$个特征, 若 $\lg(n/|F|)<1$, 则称该数据集为小数据集。

定义 3　对于数据集 S 中的任意一个少数类样本 x_i, 若其 K 近邻中有 n 个少数类样本, 那么该样本的同类邻域集 $M(x_i)$ 应由样本 x_i 和其 K 近邻中的 n 个少数类样本构成, 数据集 S 的同类邻域集 M 应由数据集 S 中每个少数类样本的同类邻域集 $M(x_i)$ 构成。

定义 4　若存在一组特征 $F=\{f_1,f_2,\cdots\}$, 利用正态分布 $N(0,h(\sigma_f))(f\in F$, σ_f 是特征 f 的标准差) 绘制的高斯噪声被称为影子样本。

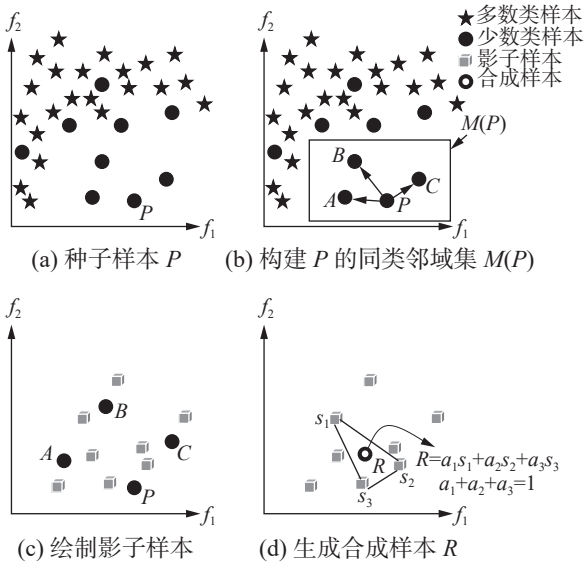
基于以上定义, 本文通过逼近边界少数类集 S_{imin} 的潜在数据流形来合成新样本, 从而避免线性插值机制存在的诸多缺陷。假设 F 表示数据的最佳特征集, 并且所有特征都同等重要, 那么过采样模型可以表示为一个函数 g:

$$\prod_{i=1}^{l}\mathbf{R}^{|F|}\to\mathbf{R}^{|F|} \tag{2}$$

式 (2) 表明 l 个选定样本可被用于合成一个新样本, 并且这些样本的特征数均为$|F|$。根据定义 1 可知, 合成样本能够利用这些选定样本的随机仿射线性组合来进行表达。也就是说, 给定特征维度为$|F|$的数据集, 假设其样本空间的流形也是$|F|$维的, 那么这个流形能够被一个$|F|-1$ 维的超平面集所逼近[25]。

为了确定这些$|F|-1$ 维的超平面, 一个近邻关系将被预先确定。根据定义, 给定样本个数为 n、特征维度为$|F|$的一个数据集, 如果满足 $\lg(n/|F|)<1$, 那么它被称为一个小数据集。显然, 一个给定样本及其 K 个近邻也构成一个小数据集。为了扩展这个小数据集, 本文根据定义 3 来构建该小数据集的同类邻域集 M。然后, 根据定义 4 为同类邻域集 M 中的每个样本绘制 S_p 个影子样本, 并要求满足 $K\times S_p>>|F|$。这样, 就可以从 $K\times S_p$ 个影子样本中选择一定数量的影子样本来创建一个具有正系数的随机仿射线性组合, 即使用影子样本的凸组合来表达合成样本。这里需要规定影子样本的选择方式: 如果选定少数类样本 P, 那么为样本 P 选择的影子样本应从其同类邻域集 $M(P)$ 绘制的影子样本中随机选择。

在 2.1 节中, 本文构造了边界少数类集 S_{imin}。S_{imin} 中的每一个样本及其同类邻域集中的样本可以构成一个小数据集。基于这个小数据集, 能够生成若干独立同分布 (正态分布) 的影子样本。然后选择一定数量的影子样本按随机仿射线性组合的方式来生成合成样本。图 7 给出了合成新样本的过程。首先, 种子样本 P 被选定 (图 7(a))。接下来, 构建 P 的同类邻域集 $M(P)=\{A, B, C\}$ (图 7(b))。然后, 为小数据集$\{P, A, B, C\}$绘制影子样本, 这些影子样本要具有与小数据集相同的正态分布 (图 7(c))。最后, 选择一定数量的影子样本$\{s_1,s_2,s_3\}$进行随机仿射线性组合, 并生成合成样本 $R=a_1s_1+a_2s_2+a_3s_3$(图 7(d))。由上述过程可知, 影子样本及合成的新样本均位于或逼近原始数据的流形。



(a) 种子样本 P　(b) 构建 P 的同类邻域集 $M(P)$

(c) 绘制影子样本　(d) 生成合成样本 R

图 7　生成合成样本 *R* 的过程

Fig. 7　Process of generating a synthetic sample *R*

2.3　算法描述

算法 1　基于 KNN 和随机仿射的边界样本合成过采样方法

输入　多数类集 S_{maj}; 少数类集 S_{min}; 近邻参数 K、k_1、k_2、k; 同类邻域集 M 中每个样本绘制的影子样本数 S_p; 正态分布的标准差 σ_f; 选择的影子样本数 n。

输出　过采样后的平衡数据集 S。

1) 初始化: 过采样后的平衡数据集 $S=\varnothing$, 影子样本集 $X=\varnothing$;

2) 对于每个少数类样本 $x_i\in S_{\text{min}}$, 计算其最近邻集 $N_{\text{N}}(x_i)$;

3) 构建安全样本集 $S_{\text{minf}}=S_{\text{min}}-\{x_i\in S_{\text{min}}:$ $N_{\text{N}}(x_i)$ 中只有多数类样本$\}$;

4) 对于每个少数类样本 $x_i \in S_{\text{minf}}$, 计算其多数类最近邻集 $N_{\text{maj}}(x_i)$;

5) 构建边界多数类集 $S_{\text{bmaj}} = \cup_{x_i \in S_{\text{minf}}} N_{\text{maj}}(x_i)$;

6) 对于每个多数类样本 $y_i \in S_{\text{bmaj}}$, 计算其少数类最近邻集 $N_{\text{min}}(y_i)$;

7) 构建边界少数类集 $S_{\text{imin}} = \cup_{y_i \in S_{\text{bmaj}}} N_{\text{min}}(y_i)$;

8) 对于每个少数类样本 $x_j \in S_{\text{imin}}$, 计算其 K 近邻;

9) 利用定义 3 构建边界少数类集 S_{imin} 的同类邻域集 M;

10) 利用定义 4 为同类邻域集 M 中的每个样本绘制 S_p 个影子样本;

11) 将绘制的影子样本 S_p 添加至影子样本集 X;

12) 从影子样本集 X 中选择 n 个影子样本;

13) 创建 n 个影子样本的随机仿射线性组合并归一化随机权重: $a_1+a_2+\cdots+a_i+\cdots+a_n=1$;

14) 计算边界少数类集 S_{imin} 中每个样本生成的合成样本数量 $N=(|S_{\text{maj}}|-|S_{\text{minf}}|)/|S_{\text{imin}}|$;

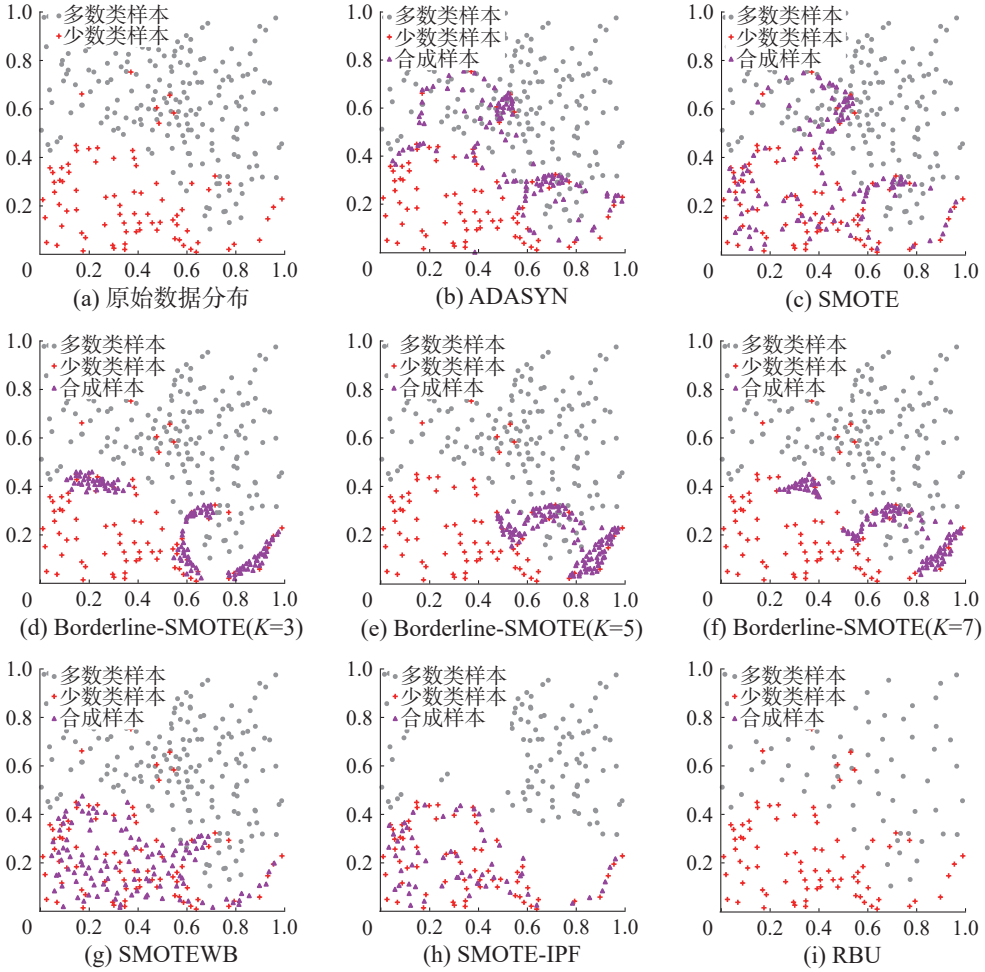
15) 生成合成样本 a_{new}: $a_{\text{new}}=a_1s_1+a_2s_2+\cdots+a_is_i+\cdots+a_ns_n$, 其中 $s_i \in X$;

16) 将全部新生成的合成样本存入 S_{minf} 中, S_{minf} 和 S_{maj} 构成过采样后的平衡数据集 S。

3　实验结果与分析

3.1　在人工数据集上的实验

为了说明本文方法的有效性, 本节在人工合成数据集上与 8 种经典的采样方法 (ADASYN[13]、SMOTE[10]、Borderline-SMOTE[12]、SMOTEWB (SMOTE with boosting)[26]、SMOTE-IPF(SMOTE with iterative-partitioning filter)[27]、RBU[17]、LoRAS[25]、RBO[22]) 进行对比。对比结果如图 8 所示, 其中少数类样本 80 个, 用“┿”表示; 多数类样本 200 个, 用“●”表示; 新生成的合成样本用“▲”表示。此外, 为了证实本文方法的抗噪能力, 在人工合成的少数类样本中包含了部分均匀分布的噪声。



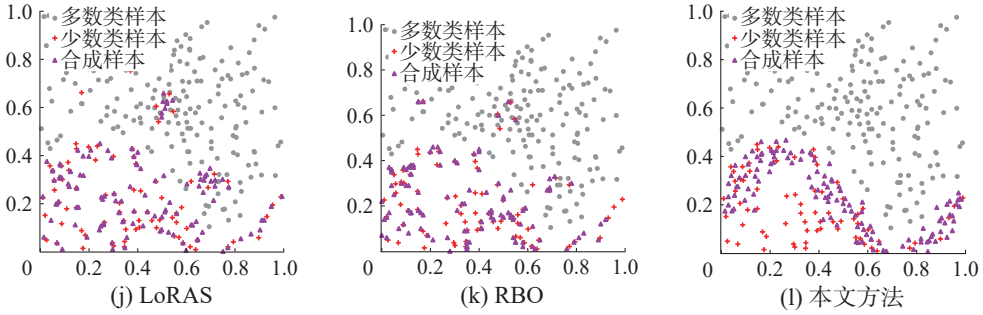


图 8　在二维数据集上的对比实验

Fig. 8　Comparative experiment on the 2D dataset

通过直观对比本文方法 (图 8(l)) 和其他 8 种方法 (图 8(b) ~ (k)) 在过采样后数据分布上的差异, 得到如下的分析结果:

1) 从图 8(b)、(c)、(i) 和 (k) 中可以看出, ADASYN、SMOTE、RBU 和 RBO 在采样后均存在大量噪声, 类间数据重叠问题并未得到缓解;

2) 从图 8(d) ~ (g) 和图 8(j) 中可以看出, 虽然类间重叠问题在 Borderline-SMOTE、SMOTEWB 和 LoRAS 中得到了一定程度的控制, 但这些方法仍然会在噪声周围生成一些合成样本。此外, Borderline-SMOTE 只考虑了样本空间的局部状态, 近邻参数 K 值的改变也会使其对种子样本的判定表现出显著的不适定性;

3) 从图 8(h) 中可以看出, 使用 SMOTE-IPF 过采样并过滤噪声后, 大量的合成样本和原始样本被删除, 导致分类边界出现偏差, 并再次造成类失衡。由于后置 IPF 过滤器直接将检测到的噪声样本删除, 因此导致数据集中重要的信息丢失;

4) 从图 8(l) 中可以看出, 本文方法能够精确识别那些具有重要信息的少数类边界样本 (采样种子), 并且在采样后不会加剧类间的数据重叠, 分类边界也接近于原始边界, 从而能够极大地保持数据的原始分布。这样, 既保留了数据原始信息和维持了数据平衡, 也减轻了噪声对分类任务造成的潜在困难。因此, 即便在数据集中存在噪声的情况下, 本文方法也取得了更好的采样结果。

3.2　在标准数据集上的实验

3.2.1　数据集及实验设置

为进一步验证本文所提方法, 使用 18 个标准数据集来进行实验。这些数据集来自 KEEL (knowledge extraction based on evolutionary learning)[28] 和 UCI(university of California Irvine machine learning repository)[29], 相关描述见表 1。在实验中, 每一个数据集均按 5 折划分为训练集和测试集, 实验结果将给出 5 次的平均值。分类器选用 AdaBoost[6]、SVM(support vector machine)[30]、BalanceCascade[31] 和 C4.5[15]。分类器和 8 种用来对比的采样方法的参数值按原始论文进行设置。为了公平对比, 本文方法参照上述设置, 具体参数值如下: 近邻参数 K=5($|S_{\text{imin}}|<100$) 或 30($|S_{\text{imin}}|\geqslant 100$)、$k_1$=5、$k_2$=3、$k_3$=$|S_{\text{minf}}|$/2; 同类邻域集 M 中每个样本绘制的影子样本数 S_p=$|F|$; 正态分布的标准差 σ_f=0.005; 选择的影子样本数 n=$|F|$。

表 1　数据集描述

Table 1　Description of the datasets

数据集名称	不平衡率	属性个数	样本个数	少数类样本数	多数类样本数	缩写
spambase	1.54	57	4 597	1 812	2 785	D1
ecoli-0_vs_1	1.86	7	220	77	143	D2
iris0	2.00	4	150	50	100	D3
german	2.33	24	1 000	300	700	D4
yeast1	2.46	8	1 484	429	1 055	D5
texture_6	2.67	40	5 500	1 500	4 000	D6
glass-0-1-2-3_vs_4-5-6	3.20	9	214	51	163	D7
texture_13	4.50	40	5 500	1 000	4 500	D8
new-thyroid1	5.14	5	215	35	180	D9
ecoli2	5.64	7	336	52	284	D10

续表 1

数据集名称	不平衡率	属性个数	样本个数	少数类样本数	多数类样本数	缩写
sgment0	6.02	19	2 308	329	1 979	D11
page-blocks0	8.79	10	5 472	559	4 913	D12
yeast-2_vs_4	9.08	8	514	51	463	D13
vowel0	9.98	13	988	90	898	D14
shuttle-c0-vs-c4	13.87	9	1 829	123	1 706	D15
glass-0-1-6_vs_5	19.44	9	184	9	175	D16
shuttle-c2-vs-c4	20.50	9	129	6	123	D17
yeast-2_vs_8	23.10	8	482	20	462	D18

3.2.2　评价指标

实验采用 F_1 分数 (F_1 score)[17] 和几何均值 (geometric mean, G-mean)[32] 作为评价指标，其计算依据为表 2 所示的混淆矩阵和查准率 ($P_{\text{recision}}=N_{\text{TP}}/(N_{\text{FP}}+N_{\text{TP}})$)、召回率 ($R_{\text{ecall}}=N_{\text{TP}}/(N_{\text{TP}}+N_{\text{FN}})$)、特异度 ($S_{\text{pecificity}}=N_{\text{TN}}/(N_{\text{TN}}+N_{\text{FP}})$)。

表 2　混淆矩阵

Table 2　Confusion matrix

样本类别	预测正类	预测负类
实际正类	TP=真正类	FN = 假负类
实际负类	FP = 假正类	TN = 真负类

$$F_1=\frac{2\times P_{\text{recision}}\times R_{\text{ecall}}}{P_{\text{recision}}+R_{\text{ecall}}}$$

$$G_{\text{mean}}=\sqrt{R_{\text{ecall}}\times S_{\text{pecificity}}}$$

式中：F_1 是查准率和召回率的调和平均，F_1 越高意味着算法对少数类样本的识别能力越强；G_{mean} 兼顾召回率和特异度，为二者的几何平均，它将 2 类样本视为同等重要。

3.2.3　实验结果分析

表 3 ~ 10 汇总了本文方法与其他 8 种采样方法 (结合 4 种分类器) 在 18 个标准数据集上取得的 F_1 和 G-mean，最高的值使用粗体表示。

表 3　本文方法与其他 8 种方法在 F_1 上的对比 (分类器使用 Adaboost)

Table 3　Comparison of the proposed method with the other 8 methods on F_1 (with Adaboost classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	92.98	93.08	95.63	**96.75**	94.42	93.10	95.84	94.55	94.71
D2	91.55	96.20	96.73	96.17	96.20	94.91	94.95	94.38	**97.98**
D3	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**
D4	56.90	56.78	57.97	57.76	60.90	56.96	60.42	58.56	**61.23**
D5	55.78	53.56	54.67	57.39	56.97	5.97	56.33	56.75	**60.87**
D6	85.94	85.67	81.34	84.40	8.02	77.86	86.67	85.29	**89.82**
D7	83.05	85.14	83.84	85.98	85.08	83.73	**87.30**	84.87	86.08
D8	71.35	75.65	73.96	75.82	73.66	73.02	76.18	74.53	**76.83**
D9	87.84	87.07	89.25	91.72	88.60	93.33	**94.00**	92.48	90.94
D10	71.23	76.27	79.73	76.91	81.82	68.74	79.84	73.64	**82.35**
D11	97.91	97.92	**98.35**	96.90	97.88	95.24	97.87	98.17	98.20
D12	78.62	80.89	81.50	81.43	78.91	78.01	86.11	**86.84**	82.94
D13	67.81	69.56	74.53	75.34	70.95	68.85	78.01	77.48	**79.38**
D14	91.25	90.68	91.42	92.38	88.45	70.07	**93.08**	91.40	92.51
D15	98.42	99.22	**100.00**	98.04	99.22	**100.00**	**100.00**	99.61	**100.00**
D16	68.86	63.33	69.33	83.05	66.10	59.69	83.90	62.00	**88.00**
D17	96.00	90.00	**100.00**	91.43	90.00	86.67	**100.00**	**100.00**	**100.00**
D18	52.86	50.94	48.48	50.58	47.71	23.29	62.62	52.55	**67.14**

表 4　本文方法与其他 8 种方法在 G-mean 上的对比 (分类器使用 Adaboost)

Table 4　Comparison of the proposed method with the other 8 methods on G-mean (with Adaboost classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	95.88	95.92	96.12	95.71	95.61	94.32	95.83	95.64	**96.31**
D2	94.36	97.22	97.62	97.61	97.22	96.17	96.54	96.18	**98.31**
D3	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**
D4	67.78	67.46	66.55	68.43	67.42	66.23	69.17	66.85	**70.68**
D5	68.41	**71.38**	66.96	69.92	66.05	65.69	65.35	65.81	69.52
D6	**93.16**	89.49	87.55	88.50	83.82	82.16	90.56	89.37	91.28
D7	88.28	90.20	88.62	91.70	90.15	89.70	90.74	88.59	**93.33**
D8	83.71	87.40	88.48	87.07	86.49	84.23	87.90	87.05	**89.15**
D9	90.98	89.64	92.58	95.01	90.11	**96.44**	94.34	94.88	93.40
D10	85.20	87.92	87.99	88.22	**91.95**	84.51	88.48	85.97	91.74
D11	99.26	**99.52**	98.87	98.96	97.47	98.66	98.75	98.93	99.19
D12	88.86	89.23	90.13	91.26	89.44	90.02	**92.68**	90.97	90.70
D13	85.79	85.80	86.65	87.06	84.55	76.51	87.67	85.23	**88.22**
D14	95.05	92.47	95.21	**98.66**	97.19	95.12	96.62	95.43	96.06
D15	99.88	99.94	**100.00**	99.85	99.94	**100.00**	**100.00**	99.97	**100.00**
D16	97.38	91.78	79.13	96.59	97.09	88.07	95.62	83.36	**99.14**
D17	99.58	98.26	**100.00**	97.50	98.26	98.33	**100.00**	**100.00**	**100.00**
D18	68.14	71.79	68.74	73.84	67.98	74.56	72.65	71.58	**77.33**

表 5　本文方法与其他 8 种方法在 F_1 上的对比 (分类器使用 SVM)

Table 5　Comparison of the proposed method with the other 8 methods on F_1 (with SVM classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	87.83	87.17	86.96	90.03	89.15	83.80	90.35	88.29	**90.52**
D2	95.67	97.37	96.20	96.77	97.37	95.10	96.21	95.74	**97.38**
D3	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
D4	77.88	78.84	76.34	**79.34**	71.92	75.30	77.17	77.59	79.05
D5	58.54	58.15	58.07	59.26	58.57	57.11	**60.15**	58.66	59.50
D6	78.25	76.35	76.14	77.60	74.83	66.94	78.53	77.02	**79.80**
D7	87.44	**90.85**	88.83	90.68	88.65	85.19	89.39	86.74	88.85
D8	57.26	59.10	60.18	60.04	58.94	54.80	64.22	59.37	**66.18**
D9	88.45	84.32	**95.80**	92.58	91.10	90.67	94.46	93.72	93.30
D10	66.78	71.84	70.29	72.92	71.49	72.32	71.47	70.58	**73.99**
D11	93.36	94.96	95.56	95.25	94.78	**97.67**	95.06	94.12	95.57
D12	48.73	54.59	55.14	53.91	55.49	52.45	50.80	51.05	**56.84**
D13	53.93	68.73	67.75	66.82	68.73	64.99	**69.24**	67.76	68.41
D14	92.36	95.38	**100.00**	97.13	94.87	77.24	**100.00**	**100.00**	**100.00**
D15	88.68	97.07	95.28	94.14	97.07	90.15	96.84	95.30	**97.92**
D16	55.71	52.76	72.00	70.11	52.76	27.72	73.26	66.00	**73.97**
D17	43.33	54.67	53.33	57.45	54.67	23.64	59.10	52.18	**60.52**
D18	19.64	54.37	52.64	60.09	54.37	56.30	65.66	**66.81**	63.67

表 6　本文方法与其他 8 种方法在 G-mean 上的对比 (分类器使用 SVM)

Table 6　Comparison of the proposed method with the other 8 methods on G-mean (with SVM classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	91.27	91.73	92.74	**94.54**	92.91	91.05	94.10	93.48	94.19
D2	97.18	97.95	97.26	97.61	97.95	94.64	97.76	96.28	**97.97**
D3	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
D4	79.92	82.58	79.76	82.17	81.25	80.39	82.64	82.87	**84.06**
D5	70.48	70.53	70.45	**71.56**	70.78	69.45	71.41	67.98	70.96
D6	83.46	82.52	82.15	83.31	78.40	72.96	**84.21**	81.98	83.57
D7	93.43	**95.47**	94.08	92.38	93.39	91.39	93.62	92.13	93.46
D8	58.09	57.45	59.83	62.27	59.74	53.26	64.82	62.10	**65.37**
D9	96.28	94.01	**97.96**	96.31	96.82	94.88	96.05	97.68	97.39
D10	88.70	90.76	89.46	91.14	89.96	90.49	91.58	90.21	**92.77**
D11	95.42	95.68	95.78	95.96	95.39	94.45	95.78	96.20	**96.91**
D12	66.51	67.45	64.65	63.89	68.01	67.09	67.75	60.92	**68.70**
D13	87.63	89.41	85.87	88.16	89.41	**89.93**	88.04	87.38	88.56
D14	99.16	99.50	**100.00**	99.83	99.44	96.94	99.44	**100.00**	**100.00**
D15	98.68	97.48	95.42	97.34	97.48	**99.18**	98.30	95.43	97.75
D16	**90.99**	90.52	79.42	86.20	90.52	82.37	89.64	73.65	87.28
D17	53.44	71.99	54.14	72.18	71.99	74.93	71.60	70.28	**75.85**
D18	68.75	74.94	71.58	77.31	74.94	72.83	79.08	75.56	**80.14**

表 7　本文方法与其他 8 种方法在 F_1 上的对比 (分类器使用 BalanceCascade)

Table 7　Comparison of the proposed method with the other 8 methods on F_1 (with BalanceCascade classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	90.96	91.82	91.30	90.11	86.70	82.98	92.00	91.53	**92.91**
D2	91.07	93.76	94.91	94.95	96.20	93.33	**97.98**	95.55	96.28
D3	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	97.90	**98.95**	**98.95**	**98.95**
D4	71.63	70.71	68.62	**72.59**	68.93	66.24	71.46	70.20	71.97
D5	42.82	49.92	52.23	57.96	52.78	43.64	56.14	51.46	**63.09**
D6	83.43	84.89	80.58	83.73	85.33	82.85	86.14	84.59	**86.62**
D7	84.60	86.25	84.88	85.83	87.72	79.44	88.80	87.92	**91.23**
D8	67.54	66.91	63.78	67.31	65.59	62.08	68.33	66.48	**69.21**
D9	89.48	88.83	85.71	91.27	84.94	91.31	87.50	**93.81**	90.95
D10	79.82	80.38	80.32	81.14	**85.71**	67.70	78.95	78.86	82.08
D11	97.02	98.80	97.74	**98.97**	96.88	96.31	97.71	97.89	98.33
D12	83.31	82.73	83.71	82.12	81.70	78.07	84.64	83.15	**86.47**
D13	71.64	63.91	72.16	72.57	63.03	69.15	**74.28**	71.33	73.23
D14	91.50	89.92	90.73	91.63	90.84	75.95	90.25	91.37	**93.06**
D15	98.42	99.22	**100.00**	99.61	98.48	**100.00**	**100.00**	99.61	**100.00**
D16	61.43	47.33	59.33	63.35	**72.67**	45.10	66.00	58.73	72.00
D17	**100.00**	96.00	**100.00**	93.33	91.43	80.00	**100.00**	**100.00**	**100.00**
D18	34.19	57.03	40.24	60.00	55.43	23.14	61.43	59.52	**65.71**

表 8　本文方法与其他 8 种方法在 G-mean 上的对比 (分类器使用 BalanceCascade)

Table 8　Comparison of the proposed method with the other 8 methods on G-mean (with BalanceCascade classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	97.14	97.00	96.90	97.10	96.54	93.83	97.11	96.32	**97.36**
D2	94.06	95.81	96.54	96.53	97.22	94.92	**98.31**	96.87	97.26
D3	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	97.95	**98.97**	**98.97**	**98.97**
D4	77.88	78.84	76.34	77.17	73.92	71.58	79.15	77.69	**79.43**
D5	55.61	61.85	63.86	67.79	65.83	56.92	61.50	63.83	**70.41**
D6	85.40	86.42	83.13	87.27	86.04	84.75	86.83	86.05	**88.14**
D7	90.62	91.23	90.25	90.08	92.50	86.09	92.09	91.20	**94.40**
D8	64.20	59.91	61.87	65.22	61.54	58.70	67.13	65.90	**68.25**
D9	91.29	89.97	91.29	93.68	95.11	89.97	91.63	**95.16**	93.42
D10	89.55	88.98	85.58	89.21	88.24	81.01	87.03	87.80	**90.96**
D11	98.73	**99.67**	98.86	99.03	97.32	97.28	98.83	98.21	99.09
D12	92.84	89.74	93.06	90.14	91.13	84.61	92.05	92.80	**94.90**
D13	84.62	79.51	84.93	85.54	83.84	78.46	85.45	81.49	**86.96**
D14	91.82	93.54	93.83	92.56	94.28	**96.62**	94.19	94.92	95.51
D15	99.88	99.94	**100.00**	99.97	99.88	**100.00**	**100.00**	99.97	**100.00**
D16	80.62	72.87	73.07	82.14	83.27	73.65	83.91	77.52	**86.85**
D17	**100.00**	99.58	**100.00**	98.94	98.71	80.00	**100.00**	**100.00**	**100.00**
D18	47.74	72.29	63.14	70.46	73.61	67.68	**74.76**	73.26	73.92

表 9　本文方法与其他 8 种方法在 F_1 上的对比 (分类器使用 C4.5)

Table 9　Comparison of the proposed method with the other 8 methods on F_1 (with C4.5 classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	94.19	92.07	92.73	93.07	92.67	90.14	94.52	92.11	**94.63**
D2	93.84	96.55	96.73	97.09	96.55	93.33	96.97	97.33	**97.98**
D3	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**	**98.95**
D4	54.02	54.19	59.25	60.07	61.14	50.91	61.21	59.20	**61.50**
D5	57.25	58.72	58.36	58.13	56.20	57.68	**61.91**	58.22	60.82
D6	**89.24**	85.27	87.42	85.97	83.15	75.20	88.48	85.98	87.86
D7	80.28	81.40	79.42	84.67	81.40	84.59	84.45	83.76	**87.72**
D8	78.70	79.76	82.75	81.48	82.75	76.53	**82.81**	80.56	81.74
D9	85.34	87.60	**92.83**	89.72	87.14	91.81	88.71	91.29	92.31
D10	76.67	75.20	78.90	77.93	74.57	75.30	73.82	75.54	**82.70**
D11	95.81	97.30	97.09	95.63	97.60	92.36	**98.04**	97.89	97.74
D12	72.40	77.20	81.45	80.26	77.30	70.85	82.65	**85.19**	84.03
D13	65.56	68.64	75.47	**76.92**	66.77	66.01	69.90	65.32	74.76
D14	89.75	89.88	92.85	90.30	90.34	67.42	92.31	91.40	**93.06**
D15	98.42	99.22	**100.00**	94.63	99.22	**100.00**	**100.00**	99.61	**100.00**
D16	57.14	54.98	**72.00**	63.81	50.22	50.67	65.72	59.88	68.00
D17	93.33	96.00	**100.00**	98.75	96.00	86.67	**100.00**	**100.00**	**100.00**
D18	23.83	46.72	47.67	52.36	50.09	36.47	56.20	53.32	**60.33**

表 10　本文方法与其他 8 种方法在 G-mean 上的对比 (分类器使用 C4.5)
Table 10　Comparison of the proposed method with the other 8 methods on G-mean (with C4.5 classifier)　%

数据集	ADASYN	SMOTE	Borderline-SMOTE	SMOTEWB	SMOTE-IPF	RBU	LoRAS	RBO	本文方法
D1	96.00	94.91	92.64	93.07	92.62	90.26	94.59	93.83	**96.02**
D2	95.79	96.61	97.60	98.15	96.61	94.93	97.20	97.96	**98.31**
D3	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**	**98.97**
D4	64.54	64.49	68.12	69.16	67.70	63.51	69.79	67.42	**70.76**
D5	69.20	70.71	70.36	70.28	68.81	69.70	**72.02**	70.25	71.43
D6	90.68	90.36	90.10	**91.00**	85.13	88.21	90.55	88.05	90.60
D7	88.18	87.99	86.20	90.75	87.99	90.00	88.68	89.72	**92.39**
D8	89.71	89.95	91.62	**91.79**	90.98	86.17	90.53	89.20	90.85
D9	95.42	93.68	96.06	94.89	92.34	95.97	91.29	94.75	**97.41**
D10	90.16	89.20	88.66	90.38	89.05	84.45	87.09	86.22	**92.59**
D11	98.14	98.78	97.98	98.73	**99.08**	98.12	99.04	98.70	98.98
D12	91.77	91.61	92.37	90.13	91.51	89.30	92.65	93.13	**93.23**
D13	86.17	**89.26**	83.99	87.85	82.51	81.79	85.07	82.52	86.63
D14	94.87	93.22	95.54	94.59	93.22	84.67	95.36	95.00	**96.11**
D15	99.88	99.94	**100.00**	99.59	99.94	**100.00**	**100.00**	99.97	**100.00**
D16	81.79	69.69	82.42	87.58	79.14	74.95	86.53	81.42	**89.82**
D17	99.60	99.58	**100.00**	99.12	99.58	98.33	**100.00**	**100.00**	**100.00**
D18	53.21	69.17	71.43	70.95	65.04	68.70	72.45	69.29	**77.65**

从表 3 ~ 4 可以看出，在使用 Adaboost 作为后置分类器时，本文方法在 12 个数据集上取得了最高的 F_1，在 11 个数据集上取得了最高的 G-mean。在与 8 种参照方法的综合对比中，F_1 平均提升了 4.66 百分点，G-mean 平均提升了 2.30 百分点。在成对比较中，本文方法与其他 8 种参照方法在 F_1 上的胜负比分别为 17∶0、17∶0、13∶2、15∶2、17∶0、15∶1、10∶5、14∶2，在 G-mean 上的胜负比分别为 15∶2、15∶2、15∶0、13∶4、15∶2、15∶1、12∶3、14∶2。

从表 5 ~ 6 可以看出，在使用 SVM 作为后置分类器时，本文方法在 11 个数据集上取得了最高的 F_1，在 10 个数据集上取得了最高的 G-mean。在与 8 种参照方法的综合对比中，F_1 平均提升了 4.16 百分点，G-mean 平均提升了 2.17 百分点。在成对比较中，本文方法与其他 8 种参照方法在 F_1 上的胜负比分别为 17∶0、15∶2、15∶1、15∶2、16∶1、16∶1、11∶5、14∶2，在 G-mean 上的胜负比分别为 15∶2、14∶3、14∶2、15∶2、15∶2、15∶2、12∶5、15∶1。

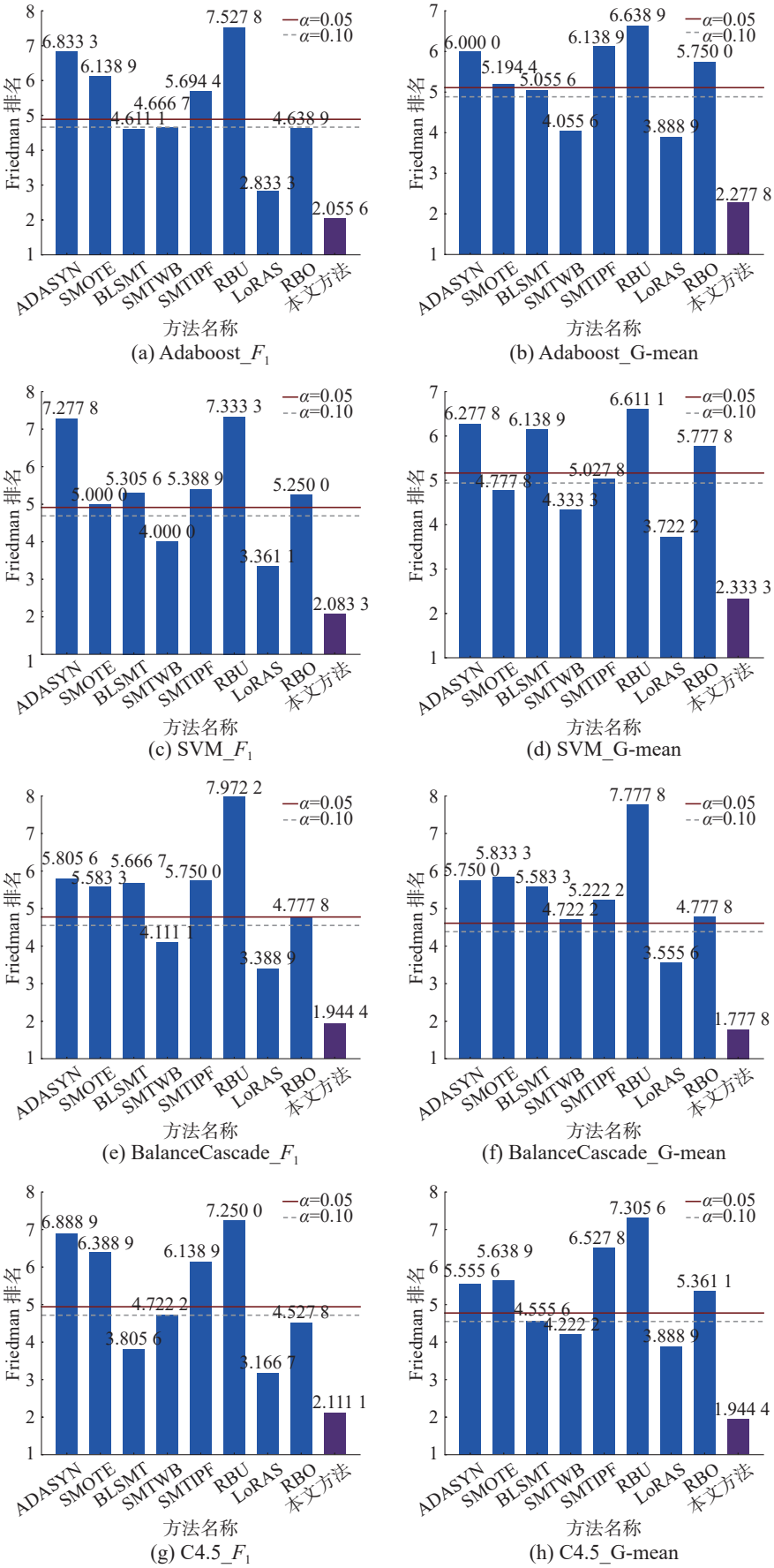
从表 7 ~ 8 可以看出，在使用 BalanceCascade 作为后置分类器时，本文方法在 11 个数据集上取得了最高的 F_1，在 13 个数据集上取得了最高的 G-mean。在与 8 种参照方法的综合对比中，F_1 平均提升了 4.59 百分点，G-mean 平均提升了 3.18 百分点。在成对比较中，本文方法与其他 8 种参照方法在 F_1 上的胜负比分别为 16∶0、16∶1、15∶0、14∶3、15∶2、16∶1、13∶2、15∶1，在 G-mean 上的胜负比分别为 16∶0、16∶1、15∶0、16∶1、16∶1、16∶1、13∶2、15∶1。

从表 9 ~ 10 可以看出，在使用 C4.5 作为分类器时，本文方法在 10 个数据集上取得了最高的 F_1，在 13 个数据集上取得了最高的 G-mean。在与 8 种参照方法的综合对比中，F_1 平均提升了 4.16 百分点，G-mean 平均提升了 2.74 百分点。在成对比较中，本文方法与其他 8 种参照方法在 F_1 上的胜负比分别为 16∶1、17∶0、11∶4、16∶1、16∶1、16∶0、11∶4、14∶2，在 G-mean 上的胜负比分别为 16∶1、16∶1、14∶1、14∶3、15∶2、16∶0、13∶2、16∶0。

实验结果表明，经本文方法过采样后，所选用的 4 个分类器均提升了对少数类样本的识别能力，即获得了更高的 F_1。在 G-mean 评价指标上，本文方法与其他 8 种方法相比也表现出明显的数值提升。这说明本文方法在关注少数类样本的同时，也能够兼顾到多数类样本。

3.2.4　非参统计检验

为了更深入地对比本文方法与参照方法之间的差异，应用 Friedman 排名和 Nemeyi 后续检验[33]来进一步执行非参统计检验。检验结果如图 9 所示，其中 BLSMT 表示 Borderline-SMOTE，SMTWB 表示 SMOTEWB，SMTIPF 表示 SMOTE-IPF。



(a) Adaboost_F_1　(b) Adaboost_G-mean

(c) SVM_F_1　(d) SVM_G-mean

(e) BalanceCascade_F_1　(f) BalanceCascade_G-mean

(g) C4.5_F_1　(h) C4.5_G-mean

图 9　本文方法与 8 种参照方法的 Friedman 排名对比

Fig. 9　Friedman ranking comparison of the proposed method with eight methods

在图 9 中，条柱的高度对应于每种方法的平均 Friedman 排名。高度越小，排名越靠前，具有最小高度的条柱对应最好的方法。由图 9 可以看出，本文方法在 8 种情况 (4 个分类器及 2 个评价指标) 下，均取得了最高的 Friedman 排名。

在 Nemeyi 后续检验方面，首先计算临界差 (critical difference, CD)：

$$I_{\mathrm{CD}} = q_{\alpha}\sqrt{\frac{c(c+1)}{6N}}$$

式中：c 是用于比较的方法的个数，N 是数据集的个数。在本文中，c=9, N=18。在显著性水平 α=0.05 下，q_{α}=0.10，由此计算得到 I_{CD1}=2.831 7；在显著性水平 α=0.10 下，q_{α}=2.855，由此计算得到 I_{CD2}=2.606 2。然后，将最小的 Friedman 排名 (即本文方法的排名) 分别与 I_{CD1} 和 I_{CD2} 相加，得出了 2 条切线，其中红色实线对应 α=0.05 显著性水平，灰色虚线对应 α=0.10 显著性水平。高于这 2 条切线的条柱所对应的方法与本文方法具有显著性差异。Nemeyi 后续检验结果表明，本文方法与 6 种参照方法 (ADASYN、SMOTE、Borderline-SMOTE、SMOTE-IPF、RBU、RBO) 在不同测度下均具有显著性差异；与 SMOTEWB 和 LoRAS 这 2 个参照方法不具有显著性差异，但取得了比它们更高的 Friedman 排名。

4　结束语

本文提出了一种基于 KNN 和随机仿射的边界样本合成过采样方法，用于在数据层面上求解针对不平衡数据的学习问题。首先，利用三近邻理论有效判别样本间的本质近邻关系，进而准确地识别出那些难以学习且包含丰富信息的边界少数类样本；其次，通过局部随机仿射的方式来代替线性插值机制，更好地估计了少数类样本 (将其视为随机变量) 底层局部数据分布的平均值，并在其近似数据流形中均匀地生成合成样本。更重要的是，本文方法解决了传统采样方法由于近邻参数 K 值改变所导致的不适定问题和线性插值机制在生成合成样本时的相关缺陷。

实验结果表明，相比于其他 8 种经典采样方法，本文方法在大部分数据集上均获得了更高的 F_1 和 G-mean。这说明本文方法不仅能够提高分类器对少数类样本的识别能力，而且也兼顾到了多数类样本。此外，Friedman 排名和 Nemeyi 后续检验也证实了本文方法的有效性和竞争力。

尽管本文方法在处理不平衡数据分类问题上取得了不错的效果，但算法的参数较多，如何有效设定这些参数的值仍然是值得研究的问题。在未来的工作中，将引入一些优化算法，尝试从自适应角度来解决参数值的设置问题。

参考文献：

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

[2] WANG Qingyong, ZHOU Yun, ZHANG Weiming, et al. Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis[J]. Expert systems with applications, 2020, 152: 113334.

[3] SHEN Feng, ZHAO Xingchao, KOU Gang, et al. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique[J]. Applied soft computing, 2021, 98: 106852.

[4] RATHORE S S, CHOUHAN S S, JAIN D K, et al. Generative oversampling methods for handling imbalanced data in software fault prediction[J]. IEEE transactions on reliability, 2022, 71(2): 747−76.

[5] WEI Guoliang, MU Weimeng, SONG Yan, et al. An improved and random synthetic minority oversampling technique for imbalanced data[J]. Knowledge-based systems, 2022, 248: 108839.

[6] GUO Haixiang, LI Yijing, SHANG J, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert systems with applications, 2017, 73: 220−239.

[7] BAO Feng, DENG Yue, KONG Youyong, et al. Learning deep landmarks for imbalanced classification[J]. IEEE transactions on neural networks and learning systems, 2019, 31(8): 2691−2704.

[8] TAO Xinmin, LI Qing, GUO Wenjie, et al. Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering[J]. Information sciences, 2020, 519: 43−73.

[9] EPENDI U, ROCHIM A F, WIBOWO A. A hybrid sampling approach for improving the classification of imbalanced data using ROS and NCL methods[J]. International journal of intelligent engineering and systems, 2023, 16(3): 345−361.

[10] 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(1): 321−357.

[11] TAO Xinmin, ZHENG Yujia, CHEN Wei, et al. SVDD-based weighted oversampling technique for imbalanced and overlapped dataset learning[J]. Information sciences, 2022, 588: 13−51.

[12] HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International Conference on Intelligent Computing. Berlin: Springer, 2005: 878−887.

[13] HE Haibo, BAI Yang, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced

learning[C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong: IEEE, 2008: 1322−1328.

[14] GAO Xin, JIA Xin, LIU Jing, et al. An ensemble contrastive classification framework for imbalanced learning with sample-neighbors pair construction[J]. Knowledge-based systems, 2022, 249: 109007.

[15] THEJAS G S, HARIPRASAD Y, IYENGAR S S, et al. An extension of synthetic minority oversampling technique based on Kalman filter for imbalanced datasets[J]. Machine learning with applications, 2022, 8: 100267.

[16] 周晶雨, 王士同. 对不平衡目标域的多源在线迁移学习[J]. 智能系统学报, 2022, 17(2): 248−256.
ZHOU Jingyu, WANG Shitong. Multi-source online transfer learning for imbalanced target domains[J]. CAAI transactions on intelligent systems, 2022, 17(2): 248−256.

[17] KOZIARSKI M. Radial-based undersampling for imbalanced data classification[J]. Pattern recognition, 2020, 102: 107262.

[18] 陶佳晴, 贺作伟, 冷强奎等. 基于 Tomek 链的边界少数类样本合成过采样方法[J]. 计算机应用研究, 2023, 40(2): 463−469.
TAO Jiaqing, HE Zuowei, LENG Qiangkui, et al. Synthetic oversampling method for boundary minority samples based on Tomek links[J]. Application research of computers, 2023, 40(2): 463−469.

[19] LENG Qiangkui, GUO Jiamei, JIAO Erjie, et al. NanBDOS: adaptive and parameter-free borderline oversampling via natural neighbor search for class-imbalance learning[J]. Knowledge-based systems, 2023, 274: 110665.

[20] HE Zuowei, TAO Jiaqing, LENG Qiangkui, et al. HSGen: a hypersphere-constrained generation mechanism to improve synthetic minority oversampling for imbalanced classification[J]. Complex & intelligent systems, 2023, 9(4): 3971−3988.

[21] BELLINGER C, DRUMMOND C, JAPKOWICZ N. Manifold-based synthetic oversampling with manifold conformance estimation[J]. Machine learning, 2018, 107(3): 605−637.

[22] KOZIARSKI M, KRAWCZYK B, WOŹNIAK M. Radial-based oversampling for noisy imbalanced data classification[J]. Neurocomputing, 2019, 343: 19−33.

[23] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information sciences, 2019, 501: 118−135.

[24] YE Xiucai, LI Hongmin, IMAKURA A, et al. An oversampling framework for imbalanced classification based on Laplacian eigenmaps[J]. Neurocomputing, 2020, 399: 107−116.

[25] BEJ S, DAVTYAN N, WOLFIEN M, et al. LoRAS: an oversampling approach for imbalanced datasets[J]. Machine learning, 2021, 110(2): 279−301.

[26] SAĞLAM F, ALI CENGIZ M. A novel SMOTE-based resampling technique trough noise detection and the boosting procedure[J]. Expert systems with applications, 2022, 200: 117023.

[27] SÁEZ J A, LUENGO J, STEFANOWSKI J, et al. SMOTE–IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering[J]. Information sciences, 2015, 291: 184−203.

[28] WANG Xinyue, XU Jian, ZENG Tieyong, et al. Local distribution-based adaptive minority oversampling for imbalanced data classification[J]. Neurocomputing, 2021, 422: 200−213.

[29] KELLY M, LONGJOHN R, NOTTINGHAM K. Machine learning repository[EB/OL]. (1988−07−01)[2023−11−24]. https://archive.ics.uci.edu.

[30] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273−297.

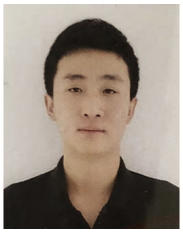
[31] YANG Kaixiang, YU Zhiwen, WEN Xin, et al. Hybrid classifier ensemble for imbalanced data[J]. IEEE transactions on neural networks and learning systems, 2020, 31(4): 1387−1400.

[32] XIE Yuxi, PENG Lizhi, CHEN Zhenxiang, et al. Generative learning for imbalanced data using the Gaussian mixed model[J]. Applied soft computing, 2019, 79: 439−451.

[33] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of machine learning research, 2006, 7: 1−30.

作者简介：

冷强奎，教授，博士生导师，博士，中国计算机学会高级会员。主要研究方向为人工智能与机器学习。主持国家自然科学基金青年项目 1 项、辽宁省博士科研启动基金项目 1 项、辽宁省自然科学基金项目 1 项、辽宁省教育厅科研项目 2 项。发表学术论文 30 余篇。E-mail：qkleng@126.com。



孙薛梓，硕士研究生，主要研究方向为人工智能与机器学习。E-mail：980048119@qq.com。

孟祥福，教授，博士生导师，博士，中国计算机学会高级会员。主要研究方向为时空大数据分析、医学影像分析、人工智能。主持国家自然科学基金项目 2 项、辽宁省高校优秀学校杰出青年学者成长计划项目 1 项、辽宁省教育厅一般项目 2 项。获发明专利授权 5 项、软件著作权 10 项，发表学术论文 80 余篇，出版专著 2 部。E-mail：marxi@126.com。