DOI:10.11918/202311030

不平衡数据集的自然邻域超球面过采样方法

周 玉[1],岳学震[1],刘 星[1],王培崇[2]

(1.华北水利水电大学 电气工程学院,郑州 450011; 2.河北地质大学 信息工程学院,石家庄 050031)

摘 要:为解决数据集类别不平衡问题,针对不平衡数据集分类提出了一种实现不平衡数据集高性能分类的自然邻域超球面过采样方法(natural neighborhood hypersphere oversampling method, NNHOS)。首先,对不平衡数据集中的每个样本点搜索其自然邻居直至形成稳定的自然邻域;接着,根据每个样本点自然邻居的标签特点,将所有样本点划分为异常点、噪声点、多数类安全点、少数类安全点和少数类边界点5个区域;然后,对每个少数类边界点构建超球面,合并完全处于大超球面中的小超球面,形成一个超球面集合;最后,根据超球面半径大小自适应地为每个超球面分配采样比例,在超球面内生成指定个数的新样本点得到平衡数据集。结果表明,利用该方法在人工数据集和真实数据集上进行过采样形成新的样本集,以CART,SVM和KNN 3个分类器进行实验,并与其他8种常用方法进行对比分析。同时,以AUC值、F_1和G_m作为评价指标,进一步证明了该方法可以更好的对不平衡数据集进行分类。

关键词:不平衡数据集;过采样;自然邻居;超球面;分类

中图分类号:TP181 **文献标志码**:A **文章编号**:0367-6234(2024)12-0081-15

A natural neighborhood hypersphere oversampling method for imbalanced data sets

ZHOU Yu[1], YUE Xuezhen[1], LIU Xing[1], WANG Peichong[2]

(1. School of Electrical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450011, China; 2. School of Information Engineering, Hebei GEO University, Shijiazhuang 050031, China)

Abstract: To address the issue of class imbalance in datasets, a natural neighborhood hypersphere oversampling method (NNHOS) for high performance classification of imbalanced data sets is proposed in this paper. First, for each sample point in the imbalanced data sets, its natural neighbors are searched until a stable natural neighborhood is formed. Then, based on the label characteristics of the natural neighbors of each sample point, all sample points are classified into five regions: outliers, noise points, safe points of the majority class, safe points of the minority class, and boundary points of the minority class. Subsequently, a hypersphere is constructed for each boundary point of the minority class. At the same time, the small hyperspheres that are completely within the large hypersphere are merged to form a set of hyperspheres. Finally, to achieve a balanced data set, each hypersphere is adaptively assigned a sampling ratio based on the hypersphere radius and a specified number of new sample points are generated within each hypersphere. The results indicate that this method utilizes oversampling on synthetic and real datasets to generate a new sample set. Experiments are conducted using the CART, SVM, and KNN classifiers, and compared with eight other commonly used methods. Additionally, AUC, F_1, and G_m are used as evaluation metrics to further demonstrate that this method can more effectively classify imbalanced datasets.

Keywords: imbalanced data sets; oversampling; natural neighborhood; hypersphere; classification

不平衡数据集是指类别分布不平衡的数据集,主要特征表现为某些类别的样本数量很少,而其他类别的样本数量很多。不平衡数据集的分类问题广泛存在于实际应用中,例如故障诊断[1-2]、人脸识别[3]、疾病检测[4-5]和企业信用评估[6-7]等,因此,对不平衡数据集进行处理进而提高分类器性能是一个十分重要的问题。一般而言,分类器训练模型为使目标函数最小,往往分类时会偏向多数类,从而导

收稿日期:2023-11-10;**录用日期**:2023-12-05;**网络首发日期**:2024-05-06
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20240430.1529.002
基金项目:国家自然科学基金(U1504622,31671580);河南省高等学校青年骨干教师培养计划项目(2018GGJS079);河北省高等学校科学技术研究项目(ZD2020344);华北水利水电大学第十五届研究生创新课题项目(NCWUYC-202315048)
作者简介:周 玉(1979—),男,副教授,硕士生导师
通信作者:周 玉,zhouyu_beijing@126.com

致少数类识别错误率很高[8]。然而,在现实生活中少数类总是包含着更多的关键信息,被错误分类的代价很高,例如疾病检测,因此,如何准确的对不平衡数据集分类,是现有分类算法所面临的难点之一。

近年来,处理数据不平衡问题的方法分为:数据级方面[9]、算法级方面[10]以及数据级和算法级两方面的结合[11]。数据级方面是通过对多数类样本进行欠采样或者对少数类样本进行过采样来实现数据集的类间平衡,例如随机过采样(random over-sampling, ROS)通过随机复制少数类样本点来实现数据集平衡,提高分类器精度。算法级方面则通过修改训练算法或对目标函数进行改进,例如代价敏感[12]和集成学习[13]作为典型的方法,可以有效提高少数类的识别精度。数据级和算法级两方面的结合是将两者进行合理结合,例如SMOTEBoost算法[14],通过将集成算法(AdaBoost)[15]和合成少数类过采样方法(synthetic minority over-sampling technique, SMOTE)[16]相结合来解决对类别不平衡数据集进行分类的问题。目前,在处理数据不平衡问题时,基于数据级方面的方法被广泛应用,因为这类方法独立于分类器并可以与多种分类器相结合。同时,由于欠采样可能导致信息缺失,从而降低分类性能,所以过采样方法被更多学者关注[17-18]。

在所有过采样方法中,随机过采样(ROS)是最简单的方法,通过随机复制部分少数类样本来达到数据平衡,但由于其不对数据做任何处理的情况下进行过采样,容易导致过拟合问题。为解决该问题,Chawla等[16]提出了合成少数类过采样方法(SMOTE),它通过与少数类样本 K 近邻[19]中的少数类进行线性插值,生成新的样本点,但SMOTE算法容易受异常点和噪声点的影响,导致在多数类区域生成样本点,加重类间重叠。对此,许多学者从不同角度出发,针对SMOTE算法的不足进行扩展或改进。一些学者考虑到决策边界对分类器的影响,尝试通过对边界数据点进行过采样,以增强边界信息。Han等[20]通过识别少数类边界点,只对边界点采用SMOTE过采样,通过增强边界信息以达到更好的分类效果,虽有效避免了异常值的影响,但因其依赖K近邻算法,当 K 选取不恰当时,容易加重边界数据的类间重叠。He等[21]通过计算每个少数类样本点的密度来自适应的确定每个少数类合成新样本点的个数,密度越低的样本点生成的新样本越多。Barua等[22]从少数类中搜索每个多数类样本的最近邻居,以识别少数类边界样本,使用聚类的方法在少数类区域生成新的样本,以保证不会引入新的噪声点,但选择合适的聚类方法成为该方法的难点。与此相反,Bunkhumpornpat等[23]提出了安全级别过采样方法,通过计算每个样本的安全系数,产生的新样本更加靠近安全系数高的样本点,有效避免了在多数类区域产生新样本,但该方法不能有效增强类边界信息。Chen等[24]基于密度将数据集分为安全区域和边界区域,根据少数类 K 近邻中少数类样本数量的比例分配不同的过采样大小,在安全区域生成更多的新样本,而边界区域生成更少的新样本。为避免SMOTE算法在处理有异常点和小分离特点的数据集时容易导致过拟合的问题,Koziarski等[25]通过引入径向基函数(radial basis function, RBF)的不平衡分布估计来找到需要生成合成样本的过采样区域。Bej等[26]采用局部随机仿射阴影采样(localized Random affine shadow-sampling, LoRAS)从少数类样本点的近似数据流形中进行过采样,有效解决了过拟合的问题,但参数适当是少数类样本流形建模的关键。由于SMOTE算法生成新的样本点时受 K 近邻算法中 K 的影响。Zhu等[27]提出了一种不带参数的最近邻方法——自然邻居。Li等[28]将SMOTE和自然邻域相结合,解决了 K 的选择问题,但并没有考虑到样本点的分布。对此,Leng等[29]通过自然邻域确定边界样本点,并根据边界样本点自然邻居中的少数类占比分配采样权重,保持数据分布并增强边界信息,但该方法也有不足之处,通过随机选取自然邻居样本进行插值,并没有扩大少数类边界样本的区域。

针对上述问题,本文提出一种基于自然邻域的超球面过采样方法。首先,对每个样本点搜索自然邻居,形成稳定的自然邻域,并根据每个样本点的自然邻居特征,将样本点划分为异常点、噪声点、多数类安全点、少数类安全点和少数类边界点5个区域。其次,通过对少数类边界点构建超球面,并把完全包含在大超球面中的小超球面进行合并,形成一个超球面集合。接着,根据每个超球面半径大小,自适应的为每个超球面分配相应的采样比例,只在超球面内进行随机过采样。最后,将剔除了异常点和噪声的原始数据集和过采样数据合并,形成平衡数据集。该方法增强边界信息的同时不会产生类间重叠问题,可以有效对不平衡样本进行处理,进而提高分类器的分类性能。

1　自然邻域理论

传统的 K 近邻本质上是利用样本点之间的距离来对测试样本点进行分类,但该方法没有考虑是否样本之间是彼此的 K 近邻。自然邻居是 Zhu 等[27]提出的一种新的邻居之间的关系,它的主要灵感来自于现实社会关系中的“友谊”。如果双方都把彼此当作朋友时,那么双方被视为彼此真正的朋友,如果每个人都有至少一个真正的朋友,便会形成一个和谐的社会环境。把这个理论推广到数据集中,如果两个样本点 x_1 与 x_2 是真正的“朋友”,即样本点 x_1 是样本点 x_2 的 λ 近邻之一,样本点 x_2 同样是样本点 x_1 的 λ 近邻之一,那么它们便被认为是彼此的自然邻居,如果每个样本点(除了异常点)都有至少一个自然邻居时,数据集中便形成了一个自然邻域 NSS,表示如下:

$$(\forall x_i)(\exists x_j)(\lambda \in n) \wedge (x_i \neq x_j) \rightarrow (x_i \in \lambda NN(x_j)) \wedge (x_j \in \lambda NN(x_i)) \quad (1)$$

式中,通过不断搜寻每个样本点的 λ 个近邻样本点,λ 从 1 开始,直到每个样本点(除异常点外)都至少有一个自然邻居(或相较于上一次搜寻得到的拥有自然邻居的样本点集合没有发生改变)时,形成稳定的自然邻域 NSS。其中 λ 近邻和自然邻居的定义如下:

定义 1(λ 近邻) 数据集 X 中的一个样本点 x_i 与其他所有样本点 $x_j(j \neq i)$ 的距离最近的 λ 个样本点,即样本点 x_i 的 λ 近邻,记为 $\lambda NN(x_i)$。

定义 2(自然邻居) 如果样本点 x_i 的 λ 近邻 $\lambda NN(x_i)$ 中包含样本点 x_j,且样本点 x_j 的 λ 近邻 $\lambda NN(x_j)$ 中同样包含样本点 x_i,则称样本点 x_i 与样本点 x_j 互为自然邻居 NaN,表示如下:

$$x_i \in NaN(x_j) \Leftrightarrow x_i \in \lambda NN(x_j) \wedge x_j \lambda NN(x_i) \quad (2)$$

为更直观的表示 NSS 的形成过程,在人工数据集上的可视化过程见图 1。图 1 中 λ 从 1 开始,近邻个数每增加 1,便把每个样本点与它所拥有的自然邻居用线连接起来,由于图 1(c)、(d)拥有自然邻居的样本集合没有发生变化,即自然邻域没有发生变化,此时形成自然邻域 NSS,其中 $\lambda = 3$,而不属于 NSS 的样本点 13 被定义为异常点。图 2 为 4 个二维数据集形成的 NSS,可以发现靠近中心的样本点拥有更多的自然邻居,λ 的大小与数据集大小有一定的关系,数据集越大,λ 越大。

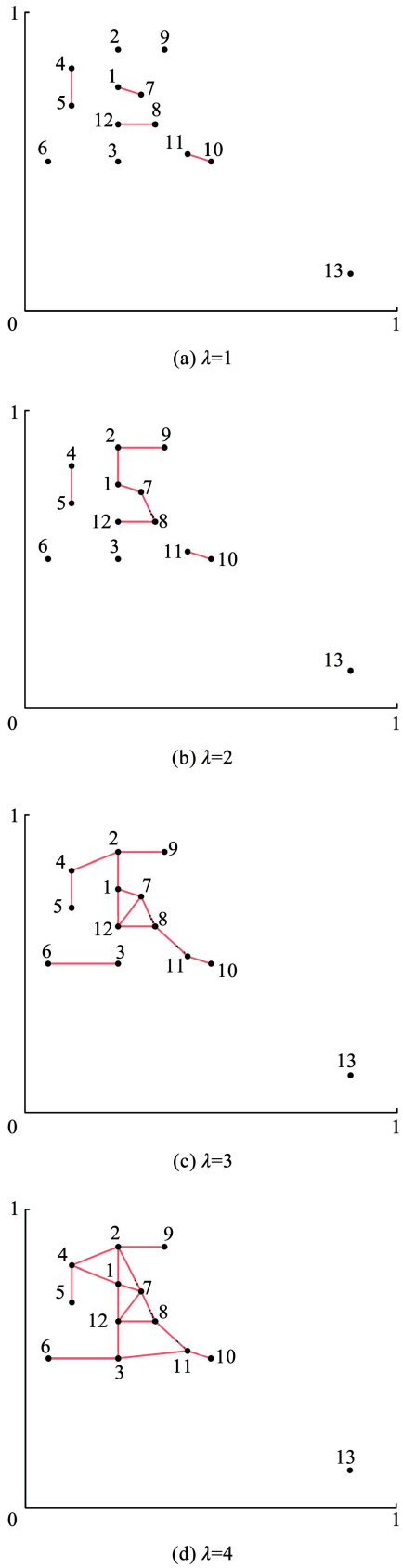
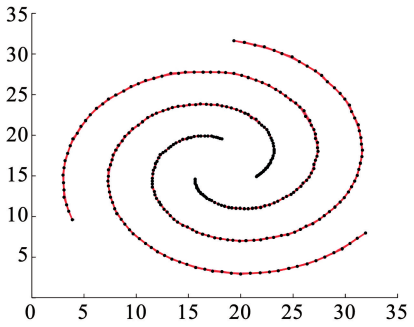
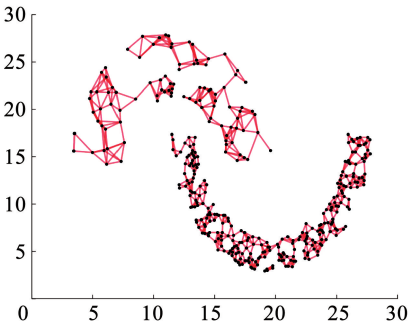
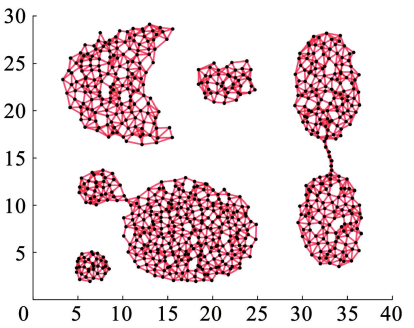


图 1　NSS 在人工数据集上的形成过程

Fig. 1　Formation process of NSS on artificial data sets


(a) λ=2


(b) λ=5


(c) λ=5

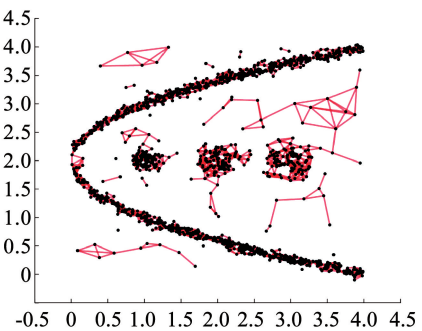

(d) λ=12

图 2 二维数据集形成的 NSS

Fig. 2 NSS formed from two-dimensional data sets

2 基于自然邻域的超球面过采样方法

2.1 基于自然邻域的区域划分

为提高采样生成的样本点的质量,根据自然邻域将原始数据集划分为 5 个区域,即异常点、噪声点、多数类安全点、少数类安全点和少数类边界点。

定义 3(异常点) 当数据集形成稳定的自然邻域后,如果样本点 x_i 不属于 NSS,则该样本点被视为异常点 $D(\text{outlier})$。

$$D(\text{outlier}) = \{x_i \mid x_i \notin \text{NSS}\} \tag{3}$$

定义 4(噪声点) 去除异常点后的数据集,对于样本点 x_i 的 λ 近邻 $\lambda NN(x_i)$,如果 $\forall x_j \in \lambda NN(x_i)$,$\text{labels}(x_j) \neq \text{labels}(x_i)$,则此时样本点 x_i 声被视为噪声点,噪声点的集合记为 $D(\text{noise})$。

$$D(\text{noise}) = \{x_j \mid \forall x_j \in \lambda NN(x_i), \text{labels}(x_j) \neq \text{labels}(x_i)\} \tag{4}$$

定义 5(多数类安全点) 去除异常点和噪声点后的多数类样本点的集合记为多数类安全点 $D(\text{safemaj})$。

定义 6(少数类安全点) 当数据集形成稳定的自然邻域后,此时对于样本点 $x_i \in \text{NSS}$ 的 λ 近邻 $\lambda NN(x_i)$,如果 $\forall x_j \in \lambda NN(x_i)$,$\text{labels}(x_j) = \text{labels}(x_i)$,则此时样本点 x_i 被视为安全点,安全点的集合记为 $D(\text{safemin})$。

定义 7(少数类边界点) 当数据集形成稳定的自然邻域后,在自然邻域 NSS 包含的的少数类样本点中,去除噪声点 $D(\text{noise})$ 和少数类安全点 $D(\text{safemin})$,剩余的少数类样本点被视为少数类边界点 $D(\text{bordermin})$

图 3 是一个人工数据集在形成自然邻域后,对各区域样本点划分的可视化结果,其中 $\lambda = 2$。

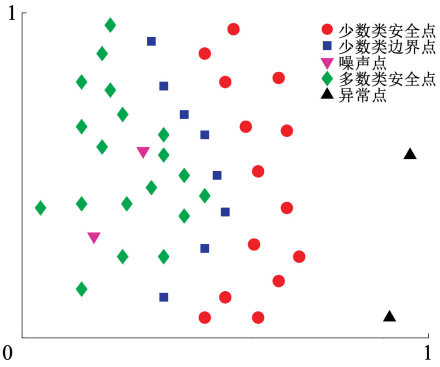


图 3 各区域样本点划分可视化

Fig. 3 Visualization of sample point division in each region

2.2 构建过采样超球面

边界样本点往往包含更多的信息,对决策边界有着重要影响。受文献[30 – 31]启发,构建以每个少数类边界点为中心的超球面,过采样过程只在超球面内进行,以增强边界信息。超球面的构建过程如图 4 所示,每个超球面的半径根据该少数类边界点与其距离最近的多数类安全点之间的欧氏距离而定,每个少数类边界点构成的超球面的半径计算如下:

$$r_i = 0.5 \times d_{x_{\text{minbor}_i\text{ne}(x_{\text{maj}})}} \tag{5}$$

式中:r_i 为超球面半径,$d_{x_{\text{minbor}_i\text{ne}(x_{\text{maj}})}}$ 为少数类边界点

到距离最近的多数类安全点之间的欧氏距离,定义超球面半径为该距离的 1/2,该策略可以有效避免类间重叠。

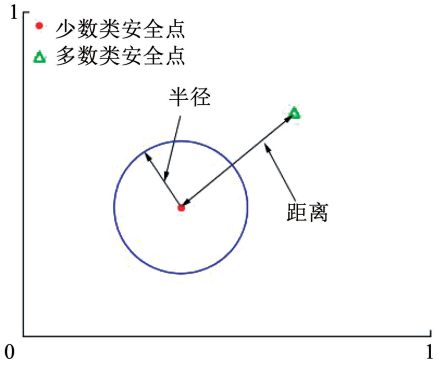


图 4　超球面半径计算

Fig. 4　Calculation of hyperspherical radius

当所有少数类边界点都构成超球面后,形成一个超球面的集合,通过计算每个超球面所包含的样本点,消除完全包含在大超球面中的小超球面,进而得到一个小的超球面集合。最终经过合并后的超球面集合包含了大多数少数类边界点,且不包含任何多数类安全点。

将多数类与少数类的数量之差作为总采样数量,以便达到类间平衡,计算方法如下:

$$N_{\mathrm{C}} = N_{\mathrm{maj}} - N_{\mathrm{min}} = (R_{\mathrm{I}} - 1) \times N_{\mathrm{min}} \tag{6}$$

式中:N_{C} 为总采样数量,N_{maj} 为去除异常点和噪声点的多数类数量,N_{min} 为去除异常点和噪声点的少数类数量,R_{I} 为不平衡比,$R_{\mathrm{I}} = N_{\mathrm{maj}}/N_{\mathrm{min}}$。

本文方法不仅通过过采样使类间达到平衡,而且根据每个超球面的半径大小自适应的为每个超球面分配相应的采样比例,越靠近分类边界的少数类边界点所构成的超球面半径越小,这些超球面对之后的分类更为重要,因此半径越小的超球面被分配的采样比例越大。每个超球面中的采样数量计算方法如下:

$$\theta_i = \mathrm{e}^{-r_i} \tag{7}$$

$$n_{c_i} = \frac{\theta_i}{\sum_{1}^{n_{\mathrm{minbor}}} \theta_i} \times N_{\mathrm{C}} \tag{8}$$

式中:θ_i 为设定的参数,用来计算每个少数类边界点构成的超球面的采样比例;r_i 为超球面的半径,n_{c_i} 为第 i 个超球面内的采样数量。

过采样过程只在超球面内进行,形成的超球面不仅覆盖了大多数少数类边界点,并且产生的新样本点不会产生类间重叠。只对少数类边界点进行过采样的策略可以增强边界信息,更有利于后续分类。当面对一些数据集按照上述方法处理后出现没有少数类边界点的情况,此时将去除异常点和噪声点后的所有少数类当作少数类边界点,超球面的半径计算如下式所示。此时根据拥有自然邻居的个数来确定每个超球面内的采样数量,处于类内边界的样本点拥有的自然邻居少,而处于类内中心的样本点拥有的自然邻居多,类内边界的样本点被分配更大的过采样数量,每个超球面内的采样数量计算方法如下:

$$r_i = d_{\max(NaN_{x_i})} \tag{9}$$

$$\theta_i = \mathrm{e}^{-n_i} \tag{10}$$

$$n_{c_i} = \frac{\theta_i}{\sum_{1}^{n_{\mathrm{minbor}}} \theta_i} \times N_{\mathrm{C}} \tag{11}$$

式中:$d_{\max(NaN_{x_i})}$ 为与样本点 x_i 的距离最远的自然邻居之间的距离,n_i 为第 i 个少数类边界点的自然邻居个数,n_{c_i} 为第 i 个超球面内的采样数量。

为避免新生成的样本点造成过拟合,在确定了每个超球面的采样数量后,采用随机采样的方法在每个超球面内进行定量的过采样,新生的样本点的生成方式如下:

$$x_{\mathrm{new}_i} = x_{\mathrm{minbor}_i} + R \times \frac{\boldsymbol{\omega}}{\| \boldsymbol{\omega} \|} \tag{12}$$

式中:x_{minbor_i} 为生成超球面的少数类边界点,R 为随机生成的半径,取值范围为 $(0, r_i)$;$\boldsymbol{\omega}$ 为随机生成的长度为 $\| \boldsymbol{\omega} \|$ 的 n 维向量。

2.3　理论分析

许多过采样方法是基于 SMOTE 算法加以改进得来,这里选用 SMOTE[16]、Borderline-SMOTE[20]、和 ADASYN[21] 这 3 种过采样方法与本文的方法进行理论分析,就目前方法的局限性,说明本文方法过采样所生成样本点的质量更高。

SMOTE 算法在面对不平衡数据集分类时,首先求得每个少数类样本点的 K 个近邻样本点,确定过采样倍率 N 后,对于每一个少数类样本点,对其 K 个近邻样本点随机选取 N 个进行过采样,采用公式如下:

$$x_{\mathrm{new}} = x_{\mathrm{cen}} + \mathrm{rand}(0,1) \times (x_{\mathrm{cen}} - x_n) \tag{13}$$

式中:x_{new} 为新生样本点,x_{cen} 为过采样样本点,x_n 为 x_{cen} 的 K 个近邻样本点之一。

按照 SMOTE 算法进行过采样时,容易受噪声点的影响,例如,当对图 5 中样本点 b 和样本点 c 进行过采样时,新生样本点 P 和 Q 位于多数类区域,造成类间重叠,不利于分类。

作为 SMOTE 算法的改进算法,ADASYN 算法和 Borderline-SMOTE 算法同 SMOTE 算法一样,都采用式(13)生成新样本点。Borderline-SMOTE 算法只对边界点使用 SMOTE 算法进行过采样,虽避免了噪声点的影响,但受 K 近邻中 K 的影响,例如当对图 5 中 a 点进行过采样,不同的 K 所决定的采样范围不同,新生样本点的质量和 K 的选取有着绝对密切的关系,例如,当 $K=9$ 时,图 5 中新生样本点 S 位于多数类区域。ADASYN 算法剔除噪声点后,根据式(14)~(16)确定每个少数类样本点的过采样个数。虽然在确定每个少数类样本点的过采样个数时,ADASYN 算法不受 K 近邻中 K 的影响,但在生成样本点阶段,同样会面临同 Borderline-SMOTE 算法一样的问题。

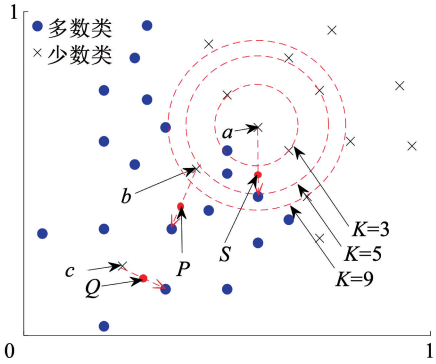


图 5 采样过程潜在问题说明

Fig. 5 Illustration of potential problems in sampling process

$$r_i = \Delta_i / K \tag{14}$$

$$R_i = \frac{r_i}{\sum_{1}^{m_s} r_i} \tag{15}$$

$$g_i = R_i \times G \tag{16}$$

式中:r_i 为设定参数,Δ_i 为少数类样本点 K 近邻中属于多数类样本点的数量,R_i 为每个少数类样本点的采样比例,G 为总采样数量,g_i 为每个少数类样本点对应的采样数量。

不同于上述 3 种过采样方法,本文方法通过对数据集构建自然邻域,根据定义 3~定义 7 将数据集划分为 5 个区域,当数据集包含少数类边界点时,由少数类边界点构建超球面,根据式(7)、(8)来计算每个超球面内的采样数量,超球面的半径越小则采样数量越多。

令

$$r_1 < r_2 < \cdots < r_{n_{\text{minbor}}} \tag{17}$$

因为指数函数 e^{-x} 是关于 x 的单调递减函数,且 $r_{n_{\text{minbor}}} > r_{n_{\text{minbor}}-1} > \cdots > r_1 > 0$,那么

$$e^{-r_1} > e^{-r_2} > \cdots > e^{-r_{n_{\text{minbor}}}} \Leftrightarrow$$

$$\theta_1 > \theta_2 > \cdots > \theta_{n_{\text{minbor}}} \tag{18}$$

所以

$$\frac{\theta_1}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C > \frac{\theta_2}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C > \cdots > \frac{\theta_{n_{\text{minbor}}}}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C$$

$$\Updownarrow$$

$$n_{c_1} > n_{c_2} > \cdots > n_{c_{n_{\text{minbor}}}} \tag{19}$$

当数据集经过区域划分后,数据集不包含少数类边界点,由自然邻域的形成过程可知,数据集的多数类与少数类有着相对独立的分布空间,例如图 2(b)所示,此时将所有的少数类归为少数类边界点。从图 1、2 可以发现,处于数据集边界的数据点所拥有的自然邻居个数少于处于数据集内部的数据点所拥有的自然邻居个数,例如在图 1(c)中,当数据集形成自然邻域后,位于数据集类内边界的样本点 5 和样本点 10 都只有 1 个自然邻居,而位于数据集类内中心的样本点 7 和样本点 12 都有 3 个自然邻居,此时由式(10)、(11)计算每个超球面内的采样数量,样本点拥有的自然邻居个数越少则采样数量越多。

令

$$n_1 < n_2 < \cdots < n_{n_{\text{minbor}}} \tag{20}$$

同上可得

$$\frac{\theta_1}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C > \frac{\theta_2}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C > \cdots > \frac{\theta_{n_{\text{minbor}}}}{\sum_{1}^{n_{\text{minbor}}} \theta_i} \times N_C$$

$$\Updownarrow$$

$$n_{c_1} > n_{c_2} > \cdots > n_{c_{n_{\text{minbor}}}} \tag{21}$$

通过上述论述可以得知,本文提出的方法不会受噪声点和 K 的影响,此外,采用式(12)在超球面内进行过采样,自适应确定超球面内采样数量,相较于 SMOTE、Borderline-SMOTE、和 ADASYN 的线性插值的采样方法,本文提出的方法可以扩展少数类边界点的未知区域,增强少数类区域边界信息,更有利于后续分类时确定决策边界。

2.4　算法步骤

Step1　计算每个样本点与其他样本点之间的欧氏距离构建距离矩阵。

Step2　根据距离矩阵,搜寻每个样本点的自然邻居,形成稳定的自然邻域。

Step3　由定义 3 ~ 定义 7 将数据集划分为 5 个区域。

Step4　对每个少数类边界点构建超球面,根据式(5)计算超球面半径大小,将完全包含于大超球面内的小超球面进行合并,形成一个超球面集合。

Step5　根据式(6)计算采样总数,并根据式(7) ~ (11)计算每个超球面内对应的采样比例,并确定每个超球面内的采样数量。

Step6　通过随机采样在每个超球面内生成指定数量的新样本点,形成增强边界信息的平衡数据集。

为更加清楚的解释本文方法的过采样过程,用二维人工数据集进行过采样,生成一个平衡数据集,其算法步骤的可视化结果见图 6。

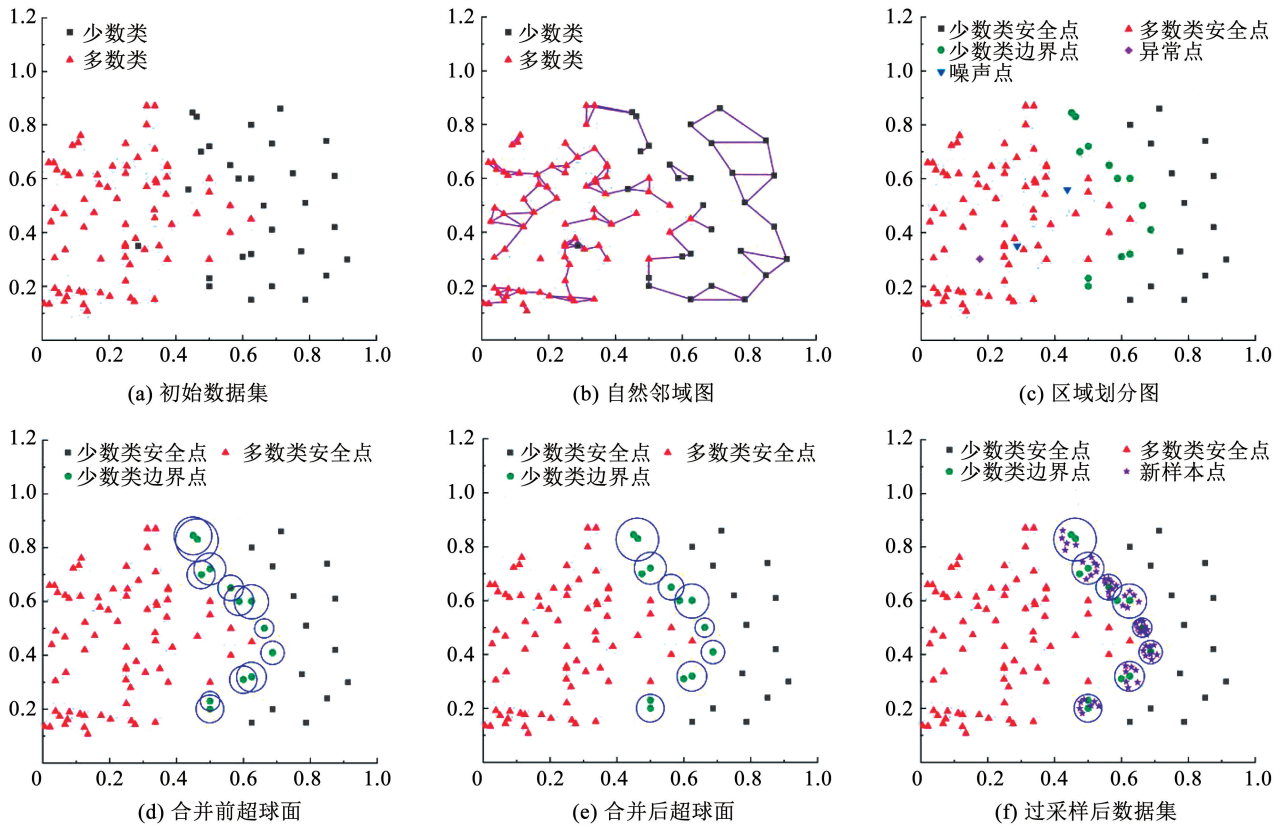


图 6　算法可视化过程

Fig. 6　Algorithm visualization flowchart

图 6(a)为初始数据集,分为多数类和少数类;图 6(b)为通过搜索每个样本点的自然邻居形成的自然邻域图,其中不属于自然邻域的样本点为异常点;图 6(c)通过每个样本点的自然邻居的标签特征,将样本点划分为异常点、噪声点、多数类安全点、少数类安全点和少数类边界点 5 个区域;图 6(d)在剔除了异常点和噪声点后,对每个少数类边界点构建超球面;图 6(e)通过计算每个超球面内的样本点信息以及半径大小,合并完全处于大超球面中的小超球面,形成过采样区域;图 6(f)通过计算每个超球面内的采样数量,在每个超球面内随机生成指定个数的新样本点,形成平衡数据集。

3　试验与分析

3.1　性能指标

对于不平衡数据集,传统分类器的分类结果会偏向多数类,此时分类器对该数据集的分类精度并不适用于评估分类器的性能。因此本文选择 AUC 值、F_1 和 G_m 作为评估分类器分类效果的性能指标[32-33],F_1 和 G_m 可由混淆矩阵(见表 1)求得。

表 1　混淆矩阵

Tab. 1　Confusion matrix

所属类别	预测类别	
	少数类	多数类
少数类	T_P	F_N
多数类	F_P	T_N

$$S = \frac{T_P}{F_P + F_N} \tag{22}$$

$$P = \frac{T_P}{T_P + F_P} \tag{23}$$

$$R = \frac{T_P}{T_P + F_N} \tag{24}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{25}$$

$$G_m = \sqrt{R \times S} \tag{26}$$

式中：S 为特异度(Specificity)，P 为精确度(Precision)，R 为召回率(Recall)，T_P 为被预测为少数类的少数类数量，F_N 为被预测为多数类的少数类数量，F_P 为被预测为少数类的多数类数量，T_N 为预测为多数类的多数类数量。

F_1 综合考虑了精确度(Precision)和召回率(Recall)的指标，F_1 越大表示对少数类的分类精度更高。G_m 表示召回率(Recall)和特异性(Specificity)的几何平均值，所以 G_m 越大表明在多数类和少数类上的表现都越好。此外还用了 AUC 值来评估分类器性能，AUC 是 ROC 曲线下的面积，它不依赖于类别分布的平衡性。AUC 是基于模型的排序能力而不是样本数量来评估性能的，即使在不平衡数据集中，正类别和负类别样本的数量差异很大，AUC 值依旧可以用来评估性能。

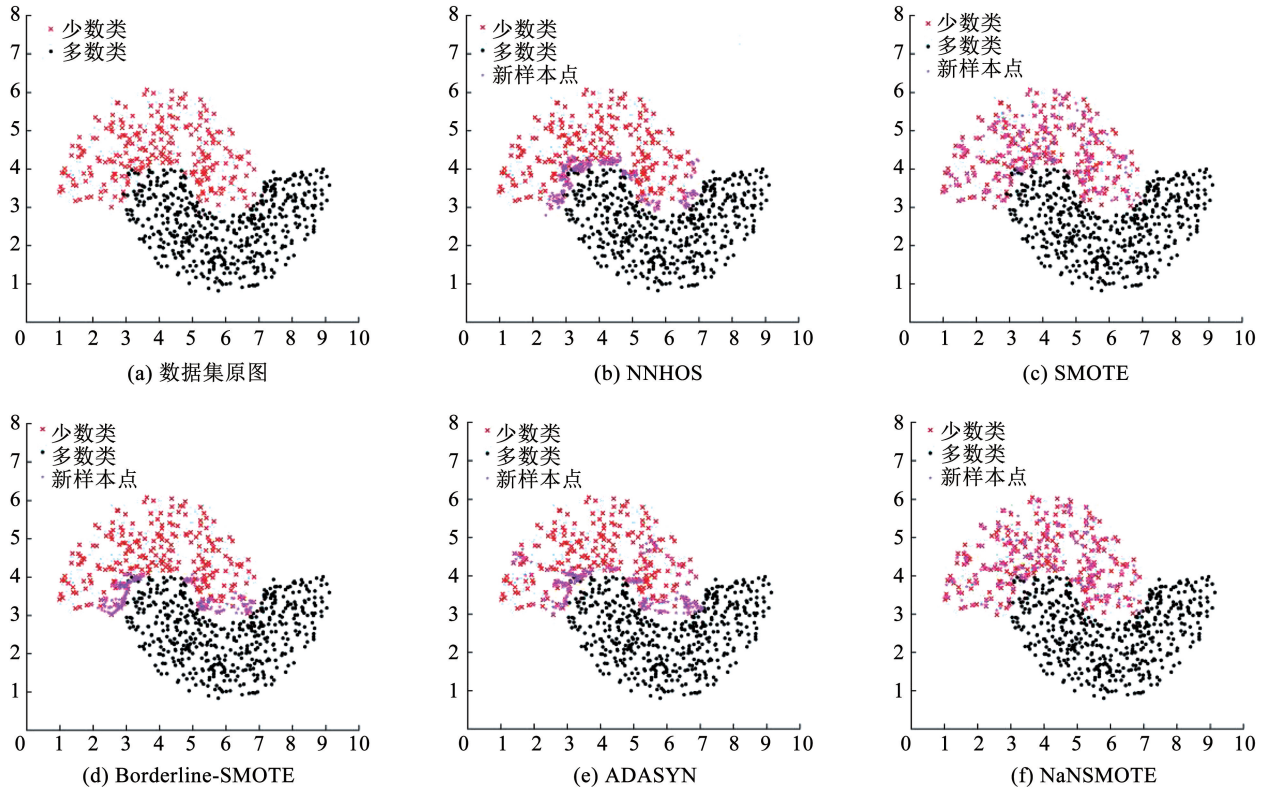
3.2 人工数据集

为了直观展现不同采样方法的特点，选用 SMOTE[16]、Borderline-SMOTE[20]、ADASYN[21] 和 NaNSMOTE[28] 这 4 种过采样方法与本文的方法在人工数据集上进行比较，其中人工数据集见表 2。不同采样方法在两个人工数据集上的采样效果见图 7、8。通过计算 AUC 值、F_1 以及 G_m 来对比不同方法的性能，表 3 ~ 5 分别是不同方法以决策树(CART)、支持向量机(SVM)和 K 近邻(KNN)作为分类器的实验结果，其中效果最好的用黑体表示。

表 2 人工数据集

Tab. 2 Artificial data sets

数据集	不平衡比	数据集数量	少数类数量	多数类数量	维度
A	2.00	750	250	500	2
B	7.96	1 505	168	1 337	2



(a) 数据集原图 (b) NNHOS (c) SMOTE

(d) Borderline-SMOTE (e) ADASYN (f) NaNSMOTE

图 7 不同方法在 A 数据集上的采样效果

Fig. 7 Sampling effects of different methods on data set A

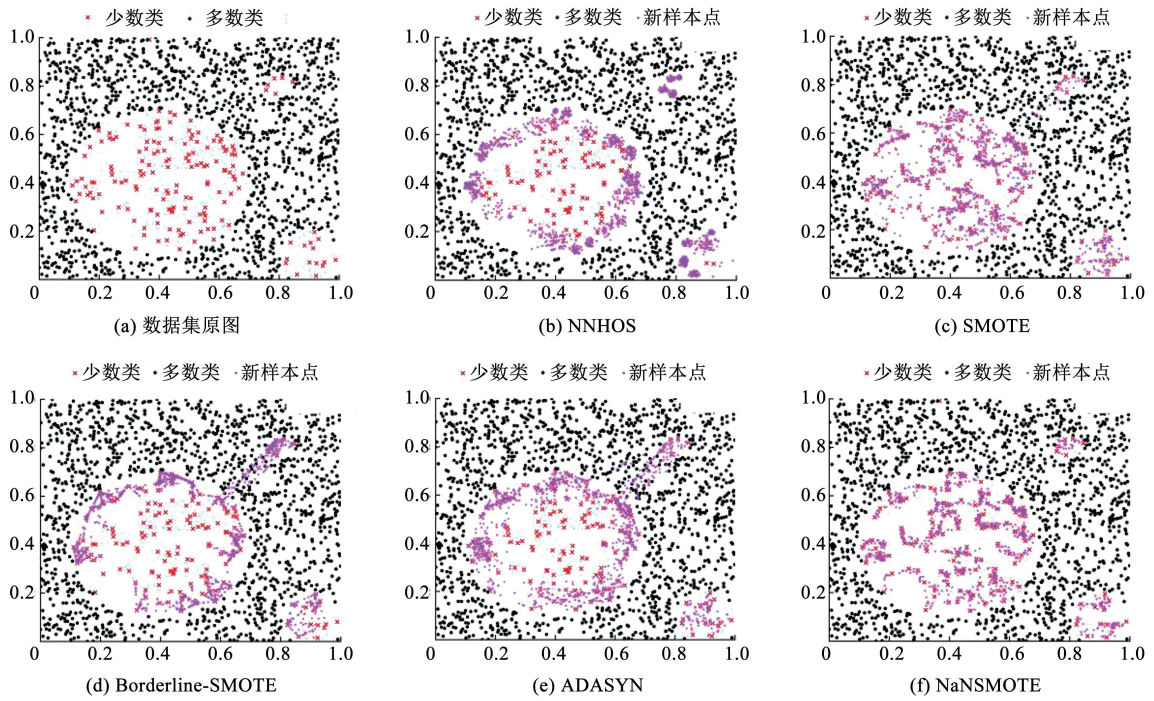


图 8　不同方法在 B 数据集上的采样效果

Fig. 8　Sampling effects of different methods on data set B

表 3　人工数据集以 CART 为分类器的实验结果

Tab. 3　Experimental results of artificial data sets based on CART classifier

数据集	指标	NNHOS	SMOTE	Borderline-SMOTE	ADASYN	NaNSMOTE
A	AUC	98.94	95.25	**99.13**	96.41	97.53
	F_1	**99.86**	96.13	99.56	96.86	97.49
	G_m	**99.27**	96.34	99.24	96.77	97.52
B	AUC	**98.55**	98.20	95.49	95.12	97.19
	F_1	**98.46**	98.18	96.28	95.40	97.07
	G_m	**98.71**	98.14	96.16	95.33	97.11

表 4　人工数据集以 SVM 为分类器的实验结果

Tab. 4　Experimental results of artificial data sets based on SVM classifier

数据集	指标	NNHOS	SMOTE	Borderline-SMOTE	ADASYN	NaNSMOTE
A	AUC	**95.85**	91.54	82.62	79.64	91.42
	F_1	**95.62**	92.31	84.31	81.31	93.33
	G_m	**94.99**	92.86	84.59	80.49	92.21
B	AUC	**80.33**	68.43	55.73	59.11	65.54
	F_1	**81.29**	69.94	54.72	59.67	65.19
	G_m	**80.46**	71.26	55.06	58.23	66.10

表 5　人工数据集以 KNN 为分类器的实验结果

Tab. 5　Experimental results of artificial data sets based on KNN classifier

数据集	指标	NNHOS	SMOTE	Borderline-SMOTE	ADASYN	NaNSMOTE
A	AUC	**1.00**	95.69	**1.00**	95.49	89.36
	F_1	**1.00**	95.73	**1.00**	96.13	88.54
	G_m	**1.00**	95.69	**1.00**	96.08	98.60
B	AUC	98.52	97.69	93.57	96.57	**98.61**
	F_1	**98.63**	98.73	93.69	96.69	98.55
	G_m	**99.36**	98.41	93.62	96.62	98.47

从图 7、8 中不难发现,由于 SMOTE、Borderline-SMOTE 以及 ADASYN 这 3 种算法在生成新的样本点时,都和 KNN 密不可分,导致的采样结果随着 K 的选择而出现显著的差异,采样效果不佳,例如图 7(c) ~7(e),SMOTE 算法因为同等的对待每一个样本点,导致产生的新的样本点与多数类产生类间重叠。Borderline-SMOTE 为了增强边界信息,只对部分少数类边界点进行过采样,但由于数据集分类边界比较模糊,新生成的样本点太过于靠近分类边界,导致产生了更加严重的类间重叠。ADASYN 通过自适应的分配少数类边界点的采样权重进行过采样,但依旧不能有效避免类间重叠。NaNSMOTE 是基于自然邻居进行过采样而非 KNN,但它对分类边界不够重视,不能在后续分类过程中加强分类边界对分类器的影响。本文提出的方法可以有效增强分类边界信息,并且不会产生类间重叠。对于试验结果来说,从表 3 ~5 可以看出,本文提出的方法在大多数情况下能取得最好的性能指标。

3.3 真实数据集

这里采用 12 个 UCI 数据集分别在决策树(CART)、支持向量机(SVM)和 K 近邻(KNN) 3 个分类器上进行试验,并与其他 8 种算法进行对比。表 6 为本次试验所用到的对比方法及本文方法。

表 7 为本文试验所用到的 UCI 数据集,其中:n 为数据量大小,n^+ 为少数类数量,n^- 为多数类数量,dim 为数据的维度。试验过程中,训练集和测试集的比例为 4∶1,训练集与测试集的不平衡比与原始数据集保持一致,试验结果为运行 30 次的平均结果。表 8 ~10 是本文方法与 8 种对比方法以不同分类器进行分类的试验结果。图 9 ~11 为本文方法与其他 8 种对比方法得到的试验结果的雷达图的形式,可以更加直观的展现试验效果。

表 6 对比方法

Tab. 6 Comparison of methods

方法	原理	方法名称简称
SMOTE[16]	KNN	M1
Borderline-SMOTE[20]	KNN	M2
ADASYN[21]	KNN	M3
SLSMOTE[23]	KNN	M4
LORAS[26]	KNN	M5
RBO[25]	KNN	M6
NaNSMOTE[28]	NaN	M7
NaNBDOS[29]	NaN	M8
NNHOS	NaN	M9

表 7 UCI 数据集

Tab. 7 UCI data sets

数据集	不平衡比	n	n^+	n^-	dim	简称
spambase	1.54	4 597	1 812	2 785	57	D1
pima	1.87	768	268	500	8	D2
seeds	2.00	210	70	140	7	D3
iris	2.00	150	50	100	4	D4
german	2.33	1 000	300	700	24	D5
wine	2.71	178	48	130	13	D6
ecoli-2	5.46	336	52	284	7	D7
yeast-2vs4	9.08	514	51	463	8	D8
ecoli-4	15.80	336	20	316	7	D9
yeast-2vs8	23.15	483	20	463	8	D10
ecoli-0137vs26	39.14	281	7	274	7	D11
yeast-6	41.40	1 484	35	1 449	8	D12

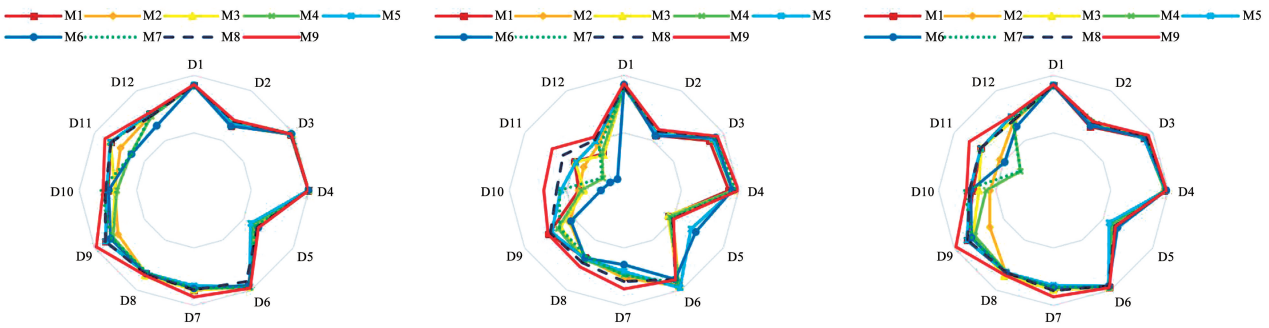
表 8 以 CART 为分类器的实验结果

Tab. 8 Experimental results of CART based classifiers

数据集	指标	M1	M2	M3	M4	M5	M6	M7	M8	M9
D1	AUC	91.22	91.43	91.38	91.36	91.55	91.23	90.89	91.52	**91.92**
	F_1	90.16	88.95	89.56	89.53	89.33	92.03	89.31	89.71	**93.01**
	G_m	91.22	91.35	90.62	91.17	91.09	90.52	91.43	91.63	**91.91**
D2	AUC	64.94	68.53	67.96	64.86	66.77	64.48	70.13	69.44	**70.33**
	F_1	55.24	57.69	58.63	56.02	57.42	54.79	60.56	59.12	**60.83**
	G_m	63.95	68.46	67.58	65.14	65.99	64.83	**69.94**	68.32	69.00
D3	AUC	97.22	97.43	98.38	97.36	97.55	98.23	**98.89**	97.52	97.92
	F_1	86.16	88.95	89.56	89.53	89.33	92.03	92.31	93.71	**94.01**
	G_m	91.22	93.35	92.62	92.17	91.09	92.52	93.43	93.63	**95.57**
D4	AUC	99.64	99.80	99.84	99.86	99.60	99.88	99.64	**99.92**	99.89
	F_1	91.37	94.95	95.86	94.55	95.36	94.32	97.65	99.63	**99.80**
	G_m	97.21	96.85	97.26	96.79	98.14	98.43	98.50	98.72	**98.55**

表 8(续)

数据集	指标	M1	M2	M3	M4	M5	M6	M7	M8	M9
	AUC	62.31	62.89	59.62	59.99	57.28	64.83	62.46	63.82	**64.88**
D5	F_1	45.01	49.00	44.37	45.83	67.31	**72.22**	47.94	49.11	49.92
	G_m	61.11	61.43	58.35	59.16	56.17	**64.53**	62.01	61.88	62.42
	AUC	95.43	95.74	97.13	97.21	96.88	95.46	97.27	91.52	**98.12**
D6	F_1	95.86	96.44	96.21	97.77	89.33	92.03	89.31	89.71	**97.01**
	G_m	97.13	96.89	97.22	**97.89**	96.11	96.52	96.95	95.81	97.61
	AUC	86.88	86.97	86.32	85.28	82.94	85.22	85.13	86.45	**92.71**
D7	F_1	75.26	76.14	72.41	73.25	70.49	64.61	73.87	79.19	**85.67**
	G_m	85.12	86.44	86.43	84.19	82.51	84.47	84.54	87.11	**92.55**
	AUC	83.37	83.76	**85.48**	83.52	83.15	83.55	82.22	83.93	85.21
D8	F_1	67.71	68.19	69.16	69.50	68.43	67.87	65.84	72.61	**76.75**
	G_m	82.96	82.94	**86.14**	81.95	82.58	82.62	82.57	82.34	84.93
	AUC	88.76	76.35	82.98	81.57	88.89	84.05	82.96	89.11	**98.26**
D9	F_1	75.82	54.85	64.11	66.46	72.88	53.41	64.87	73.99	**76.34**
	G_m	86.84	63.99	81.23	79.03	86.98	82.43	82.05	86.38	**98.13**
	AUC	74.86	69.13	68.46	67.12	75.87	74.34	**80.93**	77.02	79.17
D10	F_1	39.44	39.45	34.72	37.28	54.99	20.13	55.84	58.77	**69.92**
	G_m	70.25	55.26	64.87	58.63	73.05	72.06	**77.35**	73.15	75.64
	AUC	83.56	73.91	84.25	63.69	84.03	62.89	62.53	83.51	**89.67**
D11	F_1	49.64	40.58	49.16	21.11	48.46	13.88	22.01	61.23	**72.26**
	G_m	72.90	53.65	73.09	33.27	73.09	49.15	33.27	73.65	**84.70**
	AUC	77.13	75.35	75.37	74.12	77.16	64.89	75.87	74.67	**77.76**
D12	F_1	36.84	45.12	35.43	38.34	49.01	11.25	43.26	50.66	**53.21**
	G_m	73.93	69.24	70.34	70.02	73.96	64.18	71.52	70.04	**74.62**



(a) AUC　　(b) G_m　　(c) F_1

图 9　各种算法以 CART 为分类器的实验结果

Fig. 9　Radar chart of experimental results of various algorithms based on CART classifiers

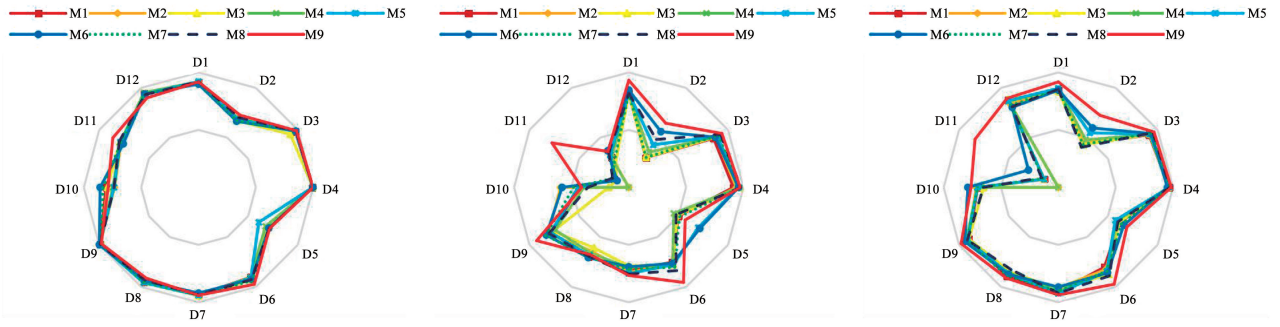
表 9　以 SVM 为分类器的实验结果

Tab. 9　Experimental results of SVM based classifiers

数据集	指标	M1	M2	M3	M4	M5	M6	M7	M8	M9
	AUC	91.12	91.56	91.25	91.98	**92.18**	89.86	91.46	92.03	91.85
D1	F_1	81.95	82.01	81.96	82.55	81.94	84.77	81.91	81.85	**93.25**
	G_m	84.16	84.93	83.52	86.06	86.15	84.61	85.12	85.43	**91.80**
	AUC	69.11	68.44	67.03	69.51	68.42	66.49	67.05	70.10	**72.54**
D2	F_1	29.84	30.21	30.11	35.43	42.66	56.03	30.14	48.16	**64.47**
	G_m	43.52	44.79	42.94	47.80	54.81	59.72	43.31	40.31	**72.47**

表 9(续)

数据集	指标	M1	M2	M3	M4	M5	M6	M7	M8	M9
	AUC	97.43	98.02	92.47	97.58	97.88	97.91	98.43	98.21	**98.62**
D3	F_1	85.38	87.63	88.14	88.06	89.23	87.19	90.27	92.35	**93.80**
	G_m	91.36	93.18	92.47	91.86	92.65	93.58	94.24	95.35	**96.34**
	AUC	99.42	99.81	99.81	99.85	99.62	99.73	**99.91**	99.86	99.88
D4	F_1	91.24	93.86	96.62	96.57	95.34	95.85	97.31	**97.64**	97.09
	G_m	96.54	95.38	96.25	96.17	96.99	95.18	97.03	97.25	**98.41**
	AUC	69.85	69.79	70.21	67.51	60.76	70.81	71.42	71.20	**71.52**
D5	F_1	50.06	49.35	48.55	45.32	70.02	**71.53**	51.16	47.61	56.77
	G_m	61.54	61.27	60.84	58.73	57.35	65.86	63.42	59.88	**69.00**
	AUC	90.35	91.26	93.11	90.53	91.27	92.42	93.74	92.56	**97.39**
D6	F_1	75.26	76.81	77.64	76.49	77.58	76.43	80.25	83.62	**95.50**
	G_m	80.20	82.54	87.38	85.49	84.31	85.82	84.26	88.94	**97.33**
	AUC	93.78	93.49	**94.31**	94.15	93.84	91.98	93.83	93.55	93.92
D7	F_1	72.56	73.99	68.42	69.86	75.06	69.14	71.26	75.38	**76.64**
	G_m	90.84	87.53	88.41	90.01	92.32	86.75	90.47	92.06	**93.61**
	AUC	94.87	94.53	94.85	**95.72**	95.64	93.78	95.47	93.29	91.23
D8	F_1	69.52	64.16	60.74	66.13	68.51	70.56	70.12	**69.41**	67.53
	G_m	89.38	87.25	84.87	86.09	86.95	87.34	89.53	82.60	**91.07**
	AUC	99.05	99.12	97.89	98.73	99.46	**99.88**	98.64	99.20	97.97
D9	F_1	78.53	80.94	76.32	74.11	78.67	82.95	83.55	80.21	**92.77**
	G_m	90.13	91.22	90.05	93.18	91.34	94.28	91.74	91.61	**97.93**
	AUC	80.66	73.25	79.51	74.83	73.64	**85.81**	83.27	73.68	78.87
D10	F_1	43.27	**59.61**	17.62	39.25	44.15	58.16	48.22	35.73	41.08
	G_m	70.31	71.64	65.28	70.14	70.62	**78.95**	70.26	65.33	76.17
	AUC	78.45	77.94	79.12	78.37	78.69	75.82	78.31	79.86	**86.28**
D11	F_1	14.62	0	12.34	0	14.33	11.81	15.96	16.17	**77.34**
	G_m	14.25	0	14.58	0	14.37	30.15	15.42	16.88	**83.91**
	AUC	94.61	93.54	94.28	**95.27**	93.62	93.86	93.91	93.20	89.92
D12	F_1	30.14	35.61	23.48	29.87	32.59	36.74	25.33	34.72	**36.12**
	G_m	87.69	83.54	88.42	86.17	87.25	80.46	82.71	80.98	**89.51**



(a) AUC (b) G_m (c) F_1

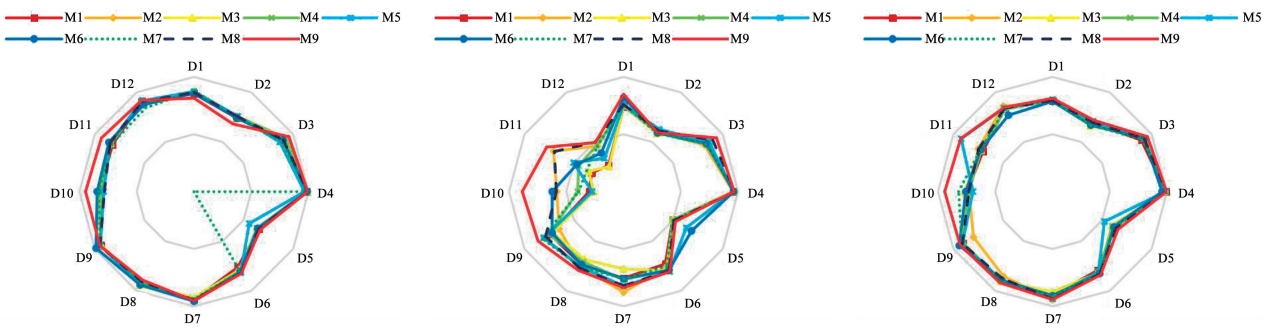
图 10 各种算法以 SVM 为分类器的实验结果

Fig. 10 Radar chart of experimental results of various algorithms based on SVM classifiers

表 10　以 KNN 为分类器的实验结果

Tab. 10　Experimental results of KNN based classifier

数据集	指标	M1	M2	M3	M4	M5	M6	M7	M8	M9
	AUC	85.42	86.83	87.35	87.14	87.26	85.71	**87.86**	86.48	81.55
D1	F_1	74.34	75.16	75.28	74.64	75.85	82.01	76.13	75.92	**84.87**
	G_m	79.13	78.22	79.57	79.85	80.20	78.71	79.36	79.52	**81.41**
	AUC	74.22	75.15	74.38	75.64	**75.99**	74.37	75.60	75.71	68.09
D2	F_1	60.14	60.03	58.39	59.67	**62.58**	59.11	58.74	61.25	59.34
	G_m	69.56	69.43	66.41	68.71	70.83	66.94	66.81	69.78	**71.89**
	AUC	90.32	89.47	90.54	93.21	86.59	89.76	87.55	90.18	**95.79**
D3	F_1	83.91	81.42	86.33	85.76	86.37	83.54	82.68	89.94	**93.68**
	G_m	89.62	92.17	93.55	92.61	91.88	92.59	90.89	92.46	**95.75**
	AUC	98.51	99.26	99.13	98.74	96.52	98.75	**99.41**	99.23	99.31
D4	F_1	95.62	95.71	96.22	96.49	95.83	96.16	**97.83**	97.18	96.63
	G_m	98.99	98.16	98.25	99.10	98.73	95.64	99.12	99.48	**99.56**
	AUC	65.32	65.17	63.48	63.52	55.76	64.25	66.23	65.19	**66.62**
D5	F_1	49.43	49.81	49.62	49.40	63.16	**68.53**	49.88	50.67	52.23
	G_m	61.84	62.53	59.76	60.31	52.33	61.91	62.87	63.69	**65.98**
	AUC	76.25	78.79	79.21	80.35	81.62	80.59	80.36	82.17	**82.52**
D6	F_1	72.85	75.34	79.62	78.91	80.36	79.14	75.12	79.68	**80.61**
	G_m	79.16	80.20	81.45	81.06	80.41	82.51	81.56	82.87	**83.94**
	AUC	94.23	94.52	92.31	94.47	94.85	**95.66**	93.71	94.58	95.19
D7	F_1	75.62	**87.85**	67.51	76.89	82.45	76.02	76.10	82.13	83.78
	G_m	90.56	**94.32**	87.45	90.89	93.46	90.87	91.52	93.73	94.13
	AUC	92.87	**94.62**	94.11	91.93	93.26	93.95	92.47	90.80	89.27
D8	F_1	73.55	75.16	68.33	70.58	75.62	73.51	75.84	76.95	**79.28**
	G_m	89.67	85.49	90.32	90.08	89.46	89.13	90.25	87.59	**91.87**
	AUC	93.71	94.02	93.54	94.65	94.33	**98.73**	94.26	94.79	94.78
D9	F_1	74.26	65.10	71.53	74.88	80.64	72.05	81.92	78.61	**86.27**
	G_m	90.20	79.65	89.74	90.28	90.96	**94.03**	91.44	90.85	93.18
	AUC	82.72	82.85	80.99	81.53	78.86	84.61	83.54	78.25	**94.73**
D10	F_1	30.11	58.35	26.32	40.19	27.51	62.34	39.26	58.78	**88.52**
	G_m	74.64	73.01	76.89	73.27	69.53	75.82	82.06	73.14	**94.19**
	AUC	82.16	84.88	83.92	84.51	84.63	85.84	82.96	84.75	**93.41**
D11	F_1	32.42	70.67	35.18	45.26	50.41	47.23	37.19	69.84	**77.57**
	G_m	70.25	74.81	71.92	71.95	92.13	72.69	72.72	74.38	**92.58**
	AUC	91.02	89.61	90.54	91.12	91.31	87.65	84.37	90.68	**91.81**
D12	F_1	26.34	45.79	25.28	46.71	33.62	38.94	45.65	**49.82**	49.40
	G_m	83.66	**85.97**	83.62	82.94	84.25	76.97	83.48	84.13	84.67



(a) AUC　　(b) G_m　　(c) F_1

图 11　各种算法以 KNN 为分类器的实验结果

Fig. 11　Radar chart of experimental results of various algorithms based on KNN classifiers

从表 8 ~ 10 可知,在以 CART、SVM 和 KNN 为分类器的 3 个性能指标中,本文提出的方法在大多数数据集上可以取得最优结果。从图 9 ~ 11 可以看出,本文提出的方法总体上要优于其他对比方法。特别在一些不平衡比例较大的数据集上,本文提出的方法效果尤为突出,这是因为在进行过采样时,只在超球面内生成新的样本点,与多数类没有交集区域,可以完全避免类间重叠的发生。

4 结 论

1)通过搜索每个样本点的自然邻居,形成稳定的自然邻域,并将数据集划分为 5 个区域,可以有效剔除异常点和噪声点对过采样的影响。

2)对每个少数类边界点构建超球面,然后合并完全包含在大超球面中的小超球面,形成一个超球面集合。

3)过采样过程只在超球面内进行,并根据每个超球面半径的大小自适应的分配每个超球面内的采样比例,可以有效避免类间重叠的同时增强边界信息,更有利于进行分类。

4)相较于其他方法,在以 AUC 值、F_1 以及 G_m 作为估计分类器的性能指标时,本文提出的方法在大多数数据集上可以取得最优结果。在未来的研究中,可以结合样本点的分布特点,有效扩大少数类的未知区域,进一步提高分类效果。

参考文献

[1] YUAN Jianhui, ZHAO Rongzhen, HE Tianjing, et al. Fault diagnosis of rotor based on Semi-supervised Multi-Graph Joint Embedding[J]. ISA Transactions, 2022, 131: 516. DOI: 10.1016/j. isatra. 2022. 05. 006

[2] PAN Haiyang, XU Haifeng, ZHENG Jinde, et al. Non-parallel bounded support matrix machine and its application in roller bearing fault diagnosis[J]. Information Sciences, 2023, 624: 395. DOI: 10. 1016/j. ins. 2022. 12. 090

[3] YANG Xiaohui, WANG Zheng, WU Huan, et al. Stable and compact face recognition via unlabeled data driven sparse representation-based classification [J]. Signal Processing: Image Communication, 2023, 111: 116889. DOI: 10. 1016/j. image. 2022. 116889

[4] REZAEIPANAH A, AHMADI G. Breast cancer diagnosis using multi-stage weight adjustment in the MLP neural network[J]. The Computer Journal, 2022, 65(4): 788. DOI: 10. 1093/comjnl/bxaa109

[5] NASROLLAHPOUR H, ISILDAK I, RASHIDI M R, et al. Ultrasensitive bioassaying of HER-2 protein for diagnosis of breast cancer using reduced graphene oxide/chitosan as nanobiocompatible platform[J]. Cancer Nanotechnology, 2021, 12(1): 10. DOI: 10. 1186/s12645 - 021 - 00082 - y

[6] CUI Lixin, BAI Lu, WANG Yanchao, et al. Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection [J]. Pattern Recognition, 2021, 114: 107835. DOI: 10. 1016/j. patcog. 2021. 107835

[7] WANG Lu, WU Chong. Dynamic imbalanced business credit evaluation based on Learn + + with sliding time window and weight sampling and FCM with multiple kernels[J]. Information Sciences, 2020, 520: 305. DOI: 10. 1016/j. ins. 2020. 02. 011

[8] 周玉, 孙红玉, 房倩, 等. 不平衡数据集分类方法研究综述[J]. 计算机应用研究, 2022, 39(6): 1615
ZHOU Yu, SUN Hongyu, FANG Qian, et al. Review of imbalanced data classification methods[J]. Application Research of Computers, 2022, 39(6): 1615. DOI: 10. 19734/j. issn. 1001 - 3695. 2021. 10. 0590

[9] MAYABADI S, SAADATFAR H. Two density-based sampling approaches for imbalanced and overlapping data [J]. Knowledge-Based Systems, 2022, 241: 108217. DOI: 10. 1016/j. knosys. 2022. 108217

[10] SUN Lin, ZHANG Jiuxiao, DING Weiping, et al. Feature reduction for imbalanced data classification using similarity-based feature clustering with adaptive weighted K-nearest neighbors[J]. Information Sciences, 2022, 593: 591. DOI: 10. 1016/j. ins. 2022. 02. 004

[11] GONG J, KIM H. RHSBoost: improving classification performance in imbalance data[J]. Computational Statistics & Data Analysis, 2017, 111: 1. DOI: 10. 1016/ j. csda. 2017. 01. 005

[12] WANG Zhe, CHEN Lilong, FAN Qi, et al. Multiple random empirical kernel learning with margin reinforcement for imbalance problems[J]. Engineering Applications of Artificial Intelligence, 2020, 90: 103535. DOI: 10. 1016/j. engappai. 2020. 103535

[13] YUAN Bowen, ZHANG Zhongliang, LUO Xinggang, et al. OIS-RF: a novel overlap and imbalance sensitive random forest[J]. Engineering Applications of Artificial Intelligence, 2021, 104: 104355. DOI: 10. 1016/j. engappai. 2021. 104355

[14] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: improving prediction of the minority class in boosting [C]// European Conference on Principles of Data Mining and Knowledge Discovery. Heidelberg: Springer, 2003: 107. DOI: 10. 1007/ 978 - 3 - 540 - 39804 - 2_12

[15] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[C]//Proceedings of the Thirteenth International Conference on Machine Learning. Bari: ACM, 1996: 148. DOI: 10. 5555/3091696. 3091715

[16] 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. DOI: 10. 1613/jair. 953

[17] AGUSTIANTO K, DESTARIANTO P. Imbalance data handling using neighborhood cleaning rule (NCL) sampling method for precision student modeling[C]//2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE). Jember: IEEE, 2019: 86. DOI: 10. 1109/icomitee. 2019. 8921159

[18] WANG Xinyue, XU Jian, ZENG Tieyong, et al. Local distribution-based adaptive minority oversampling for imbalanced data classification[J]. Neurocomputing, 2021, 422: 200. DOI: 10. 1016/j. neucom. 2020. 05. 030

[19] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21. DOI: 10.1109/TIT.1967.1053964

[20] HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]// Lecture Notes in Computer Science. Heidelberg: Springer, 2005: 878. DOI: 10.1007/11538059_91

[21] 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. DOI: 10.1109/IJCNN.2008.4633969

[22] BARUA S, ISLAM M M, YAO Xin, et al. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405. DOI: 10.1109/TKDE.2012.232

[23] BUNKHUMPORNPAT C, SINAPIROMSARAN K, LURSINSAP C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem [C]//Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer, 2009: 475. DOI: 10.1007/978-3-642-01307-2_43

[24] CHEN Baiyun, XIA Shuyin, CHEN Zizhong, et al. RSMOTE: a self-adaptive robust SMOTE for imbalanced problems with label noise[J]. Information Sciences, 2021, 553: 397. DOI:10.1016/j.ins.2020.10.013

[25] KOZIARSKI M, KRAWCZYK B, WOZNIAK M. Radial-based approach to imbalanced data oversampling [C]//International Conference on Hybrid Artificial Intelligence Systems. Cham: Springer, 2017: 318. DOI: 10.1007/978-3-319-59650-1_27

[26] BEJ S, DAVTYAN N, WOLFIEN M, et al. LoRAS: an oversampling approach for imbalanced datasets[J]. Machine Learning, 2021, 110(2): 279. DOI: 10.1007/s10994-020-05913-4

[27] ZHU Qingsheng, FENG Ji, HUANG Jinlong. Natural neighbor: a self-adaptive neighborhood method without parameter K[J]. Pattern Recognition Letters, 2016, 80: 30. DOI: 10.1016/j.patrec.2016.05.007

[28] LI Junnan, ZHU Qingsheng, WU Quanwang, et al. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors[J]. Information Sciences, 2021, 565: 438. DOI: 10.1016/j.ins.2021.03.041

[29] 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. DOI: 10.1016/j.knosys.2023.110665

[30] HO T K, BASU M. Complexity measures of supervised classification problems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 289. DOI:10.1109/34.990132

[31] TAO Xinmin, GUO Xinyue, ZHENG Yujia, et al. Self-adaptive oversampling method based on the complexity of minority data in imbalanced datasets classification[J]. Knowledge-Based Systems, 2023, 277: 110795. DOI: 10.1016/j.knosys.2023.110795

[32] 周玉, 岳学震, 孙红玉. 考虑不平衡指数的不平衡数据集分类设计方法[J]. 计算机应用研究, 2023, 40(12): 3566
ZHOU Yu, YUE Xuezhen, SUN Hongyu. Classification design method of unbalanced data sets considering unbalanced index[J]. Application Research of Computers, 2023, 40(12): 3566. DOI: 10.19734/j.issn.1001-3695.2023.04.0163

[33] ZHANG Aimin, YU Hualong, HUAN Zhangjun, et al. SMOTE-RkNN: a hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors[J]. Information Sciences, 2022, 595: 70. DOI: 10.1016/j.ins.2022.02.038

(编辑　张　红)