计算机科学
COMPUTER SCIENCE

基于无噪梯度分布的合成过采样方法

胡立彬, 张云峰, 刘培德

引用本文

胡立彬, 张云峰, 刘培德. 基于无噪梯度分布的合成过采样方法[J]. 计算机科学, 2025, 52(9): 220-231.

HU Libin, ZHANG Yunfeng, LIU Peide. Synthetic Oversampling Method Based Noiseless Gradient Distribution [J]. Computer Science, 2025, 52(9): 220-231.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000010

基于无噪梯度分布的合成过采样方法

胡立彬[1]　张云峰[2]　刘培德[3]
1 山东财经大学管理科学与工程学院　济南 250014
2 山东财经大学计算机科学与技术学院　济南 250014
3 山东财经大学山东省区块链金融重点实验室　济南 250014
(hlblydx@163.com)

摘　要　合成过采样方法(Synthetic Oversampling Method)是解决不平衡分类问题的重要手段,但当前的合成过采样方法在处理高维不平衡分类问题时仍面临诸多挑战。针对当前合成过采样方法未考虑噪声样本造成的误差累积、对样本空间距离过度依赖、合成样本的分布牺牲负类样本识别精度这3个问题,提出一种基于无噪梯度分布的合成过采样方法。首先,利用样本的梯度贡献属性作为度量样本标签置信度的指标并过滤数据集中的噪声标签样本,避免了噪声样本作为根样本造成的误差累积。其次,根据梯度贡献指标和安全梯度阈值将正类样本分配到不同的梯度区间,并选择安全梯度区间内的样本作为根样本,根样本的梯度右近邻作为辅助样本,不仅摆脱了对空间距离度量的依赖,而且保证了决策边界不断往负类样本移动。最后,设计了基于余弦相似度的安全梯度分布近似策略,用于计算每个安全梯度区间内需要生成的样本数量,该策略合成后的样本分布可以使决策边界以安全的方式向负类样本移动,因此不会明显牺牲负类样本的识别精度。在来自KEEL,UCI和Kaggle平台的数据集上的实验表明,所提出的算法在提升分类器Recall值的同时,也可以获得很好的F1-Score,G-Mean和MCC值。

关键词:梯度贡献;无噪梯度;梯度右近邻;安全梯度分布近似;合成过采样
中图分类号　TP181

Synthetic Oversampling Method Based Noiseless Gradient Distribution

HU Libin[1],ZHANG Yunfeng[2] and LIU Peide[3]
1 School of Management Science and Engineering,Shandong University of Finance and Economics,Jinan 250014,China
2 School of Computer Science and Technology,Shandong University of Finance and Economics,Jinan 250014,China
3 Shandong Key Laboratory of Blockchain Finance,Shandong University of Finance and Economics,Jinan 250014,China

Abstract　Synthetic Oversampling Method is an important means to solve imbalanced classification problem,but the current oversampling methods still have many problems when dealing with high-dimensional imbalanced classification problem. A synthetic oversampling method based on noiseless gradient distribution is proposed to address the three issues of error accumulation caused by noise samples,excessive dependence on sample space distance,and reduced recognition accuracy of negative class samples in current synthetic oversampling methods. Firstly,the gradient contribution attribute of the sample is used as the metric to measure the label confidence of the sample and the noise label samples in the data set are filtered to avoid the error accumulation caused by the noise samples as the root samples. Secondly,the positive samples are assigned to different gradient intervals according to the gradient contribution metric and the safe gradient threshold,the samples in the safe gradient interval are selected as the root samples,and the gradient right nearest neighbor of the root sample are regarded as the auxiliary samples,which not only gets rid of the dependence on spatial distance measurement,but also ensures that the decision boundary moved to the negative class samples continuously. Finally,a safe gradient distribution approximation strategy based on cosine similarity is designed to calculate the number of samples to be generated in each safe gradient interval,and the synthesized sample distribution by which can make the decision boundary moved toward the negative class samples in a safe way,so the recognition accuracy of the negative class samples will not be significantly sacrificed. Experiments on datasets from KEEL,UCI and Kaggle platforms show that the proposed algorithm can not only improve the Recall value of the classifier,but also obtain satisfactory F1-Score,G-Mean and MCC values.

到稿日期:2024-10-08　返修日期:2025-02-15
基金项目:山东省自然科学基金(ZR2022MF245);山东省重点研发计划(2023CXPT033)
This work was supported by the Natural Science Foundation of Shandong Province(ZR2022MF245) and Key R&D Program of Shandong Province (2023CXPT033).
通信作者:张云峰(yfzhang@sdufe.edu.cn)

Keywords Gradient contribution，Noiseless gradient，Gradient right neighbor，Safe gradient distribution approximation，Synthetic oversampling

1 引言

在二分类任务中，类别不平衡现象指用于训练机器学习模型的数据集中某个类别（多数类或负类）的样本数量远远多于另一个类别（少数类或正类）的样本数量。在信贷违约预测[1]、欺诈消费检测[2]、入侵检测[3]、故障诊断[4]等场景下，类别不平衡现象较为普遍。该现象导致机器学习模型对正类样本的识别精度较低，但场景任务的特殊性决定了正类样本的识别精度具有重要意义。因此，解决数据集的类别不平衡问题是人工智能领域的热点话题。

合成过采样方法因不受数据场景和机器学习模型种类的限制而成为解决类别不平衡问题的重要研究方向[5-7]。它通过为正类样本合成一定数量的伪样本来使正类样本和负类样本的数量大致相当，通过数据质量优化手段提升了机器学习模型对正类样本的识别能力。合成过采样算法涉及 3 个关键信息的确定，分别是根样本（或种子样本）、辅助样本和采样倍率（或采样权重），这 3 方面的因素决定了合成过采样方法的性能。其中根样本与辅助样本都是原始数据集中被选择出来的正类样本，二者通过线性插值的方式合成新的伪样本，采样倍率则决定了为每个根样本合成的样本数量。合成少数类过采样技术（SMOTE）[8]是早期比较经典的合成过采样方法，其在现实不平衡分类问题中也有广泛应用。它将全部正类样本作为根样本，平均分配每个根样本的平均采样权重，然后从每个根样本的 K 个空间近邻中随机选择一个样本作为辅助样本，最后通过线性插值的方式合成新的伪样本。随后，诸多针对 SMOTE 的优化方法被提出，包括以空间近邻标签确定采样权重的 ADASYN[9]、基于边界样本度量合成的 Borderline-SMOTE[10]、基于空间聚类合成的 KMeans-SMOTE[11]、基于支持向量合成的 Safe-Level-SMOTE[12]等，这些方法分别在根样本、辅助样本和采样权重方面优化了 SMOTE 方法。

首先，本文将影响不平衡分类性能的因素分为两个方面：噪声标签和类别梯度贡献不均衡。其中噪声标签是广义的概念，它包括被标记错误的标签，也包括被标记正确但使机器学习模型性能降低的样本标签，所以本文将噪声标签定义为对机器学习模型拟合起负作用的样本标签。类别梯度贡献不均衡指不同类别的样本在数量上的差异导致每个类别的样本对机器学习模型拟合的梯度贡献不均衡。负类样本在数量上的绝对优势导致其对模型的梯度贡献较大，而正类样本则相反。因此，在合成过采样方法中避免噪声标签导致的误差累积并且平衡类别间的梯度贡献是一个值得研究的方向。

当前关于合成过采样方法的研究仍然存在以下问题：

1）噪声样本会使分类器朝着错误的方向优化，分类器为了拟合这些噪声样本，会强制决策边界产生错误偏移。图 1(a)为不含噪声标签时的决策边界，但当数据集中存在噪声标签时，决策边界如图 1(b)所示。可以看出，噪声标签样本使决策边界发生错误的偏移，而这些噪声样本在合成过采样方法中无论被选择作为根样本还是辅助样本，都会将其误差逐渐累积。同时，噪声样本的存在还会对正常样本在分类器中的表现产生影响，例如噪声样本的存在使正常样本对分类器的贡献产生偏移。

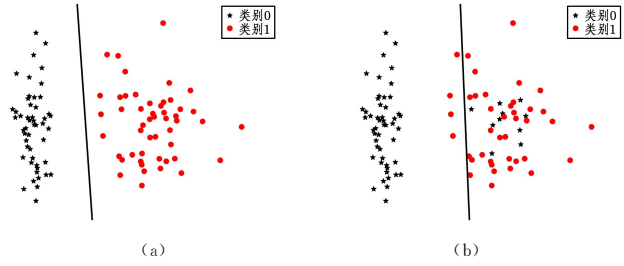


图 1 噪声样本对分类器的影响

Fig. 1 Effect of noisy samples on the classifier

2）当前的合成过采样方法对样本间的空间距离过度依赖，而高维条件下，空间的稀疏性导致样本的空间距离度量逐渐失效[13]。真实数据集上的空间聚类实验结果如图 2 所示。图 2(a)展示了将 28 维数据降低到 3 维时样本的空间分布情况。DBSCAN 和 OPTICS 这两个基于空间距离的聚类方法无法得到有效子簇，认为所有样本都是异常点。如图 2(b) 所示，用 K-means 方法强制将该数据集分为 4 个子簇时，各子簇的边界处样本分布十分模糊，因此边界附近样本的空间近邻是不确定的，这使得基于空间距离的合成过采样方法性能十分受限。

3）合成后的正类样本分布容易造成正类样本过拟合，导致负类样本的识别精度降低。

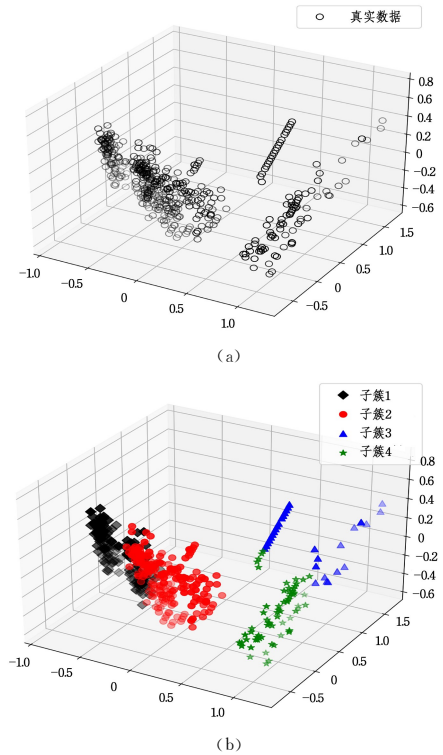


图 2 基于空间距离的聚类结果

Fig. 2 Clustering results based on spatial distance

假设样本的空间距离度量有效，当前基于空间距离的合成过采样方法在选择根样本和辅助样本时仍然存在一定问题。如图 3 所示，虚线圆内的样本只能在虚线圆内搜索空间近邻，且选择的辅助样本可能会比根样本距离决策边界更远。同时，合成的伪样本也会在虚线圆内且比根样本距离决策边界更远，无法有效推动决策边界往负类样本移动。

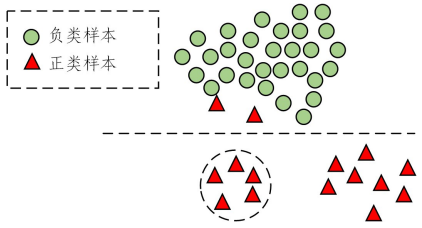


图 3 空间近邻搜索

Fig. 3 Spatial nearest neighbor search

同样假设样本的空间距离度量有效，且正类样本可以被很好地聚类从而获得子簇。当前基于聚类的合成过采样方法通常用于为更稀疏的子簇合成更多的伪样本。如图 4 所示，当虚线圆内的子簇距离决策边界较远时，为该子簇合成较多数量的伪样本将强制决策边界往负类样本移动，造成边界处的负类样本被识别错误。这也是其他合成过采样算法牺牲负类样本识别精度的原因，即合成过采样后的样本分布使决策边界以一种危险的方式往负类样本移动。

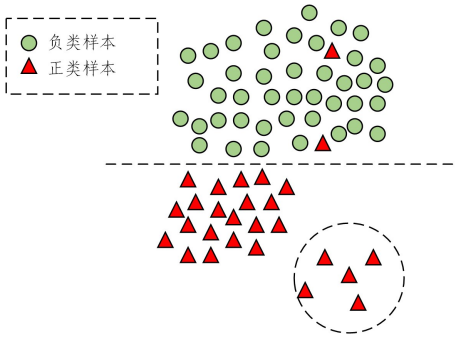


图 4 基于聚类的合成方法

Fig. 4 Cluster-based synthesis method

为解决上述问题，本文提出了基于无噪梯度分布的合成过采样方法(Noiseless Gradient Distribution Synthetic Over-sampling，NGDSO)。首先 NGDSO 初步过滤了数据集中的噪声标签样本，然后根据无噪梯度的分布信息来合成正类伪样本。噪声样本的完全识别和过滤是非常理想的情况，本文方法在前端初步过滤噪声样本的目的是剔除明显对分类器的拟合起负作用的样本，以及计算得到更精确的正常样本的梯度分布。此外，本文方法在根样本和辅助样本的选择策略设计时考虑了方法对噪声样本的鲁棒性。

本文的主要工作总结如下：

1)引入了不依赖空间距离的梯度贡献概念作为样本的置信度度量指标，并设计了基于平均梯度贡献的噪声标签过滤方法，避免了噪声样本造成的误差累积。

2)在初步过滤噪声样本后的数据集上，重新计算正类样本的梯度贡献(此时称为无噪梯度贡献)，并设计了根样本的选择策略，即安全梯度区间内的样本被选择作为根样本，进一步避免了梯度贡献较大的噪声样本造成的误差累积。

3)设计了辅助样本的选择策略，即将根样本的梯度右近邻作为辅助样本。该策略决定了辅助样本始终比根样本距离决策边界更近，使决策边界不断往负类样本的方向移动，从而提高了正类样本的识别精度。

4)设计了基于余弦相似度的安全梯度分布近似策略。该策略确定了每个安全梯度区间内需要合成的样本数量，合成后的样本分布使决策边界以安全的方式向负类样本移动。因此，NGDSO 相比其他方法不会明显降低负类样本的识别精度。

2 相关工作

近年来，许多新的合成过采样方法被提出[14]，这些研究工作可以分为 4 个方向，分别是空间特征优化、自适应倍率、基于聚类合成和基于过滤合成。

2.1 空间特征优化

Chen 等[15]首先通过信息熵定义每个正类样本被识别为边界样本的置信度，同时根据西格玛规则确定正类样本的边界点；然后沿着梯度方向迭代地合成辅助边界点；最后对正边界样本和辅助边界点样本进行过采样。该方法摆脱了对空间距离特征的依赖。Li 等[16]提出了基于自然近邻的合成过采样方法，它不受近邻个数这一参数的限制，利用中心样本与边界样本自然近邻数量不同这一特点，提高了合成样本的多样性。Wang 等[17]利用自然邻域计算样本的局部密度，并对数据集进行过滤处理使样本的类别边界更加清晰，根据局部密度将合成的样本分布到正类样本中。Li 等[18]通过样本的自然近邻数量来识别噪声样本和离群样本，根据正类样本的相对邻域密度分布分配合成伪样本的数量和位置。

2.2 自适应倍率

Leng 等[19]改进了 Li 等[16]提出的方法，在搜索自然邻域时为样本自适应地分配动态权重，充分利用了样本的原始空间分布特征。Yan 等[20]也为原始样本分配了自适应权重，并结合原始样本的局部密度和离分类边界的距离进行自适应加权过采样。Tao 等[21]设计了一种由均值移动方法引导的自适应过采样方法，它结合正类样本半径邻域内的样本分布、正类样本的密度和正类样本到负类样本的距离自适应地确定合成正类样本的数量。Zhang 等[22]提出了一种多重自适应过采样方法，它首先将正类样本进行多次自适应过采样来平衡类别分布，然后在不同不平衡率的数据集上训练不同权重的子分类器，最后利用近邻信息对难以分类的样本标签进行修正。Sun 等[23]提出了一种基于模糊邻域的自适应特征选择方法，根据样本间的方差距离和自适应模糊邻域半径构造自适应模糊邻域联合熵，用于构造原始样本和合成样本的平衡决策系统。

2.3 聚类合成

Moutaouakil 等[24]提出了一种基于最优熵遗传 C 均值的合成过采样方法，利用模糊 C 均值识别出原始正类数据中的安全样本并根据空间特征进行过滤，最后基于不平衡率阈值和数据分布指数来合成样本。Meng 等[25]利用中心偏移因子 COF(Center Offset Factor)改进了 SMOTE 方法，首先利用空

间距离把正类样本分为不同子簇,然后基于中心偏移因子识别出分布稀疏的子簇,最后在稀疏子簇内合成伪样本。Wang 等[26]首先利用 K-Means 方法将数据集分为不同不平衡率的子簇,然后将所有子簇内样本分为正域、边界域、负域,最后根据不平衡率和每个域的属性计算每个子簇的过采样权重。

2.4 过滤合成

Xu 等[27]提出一种基于高斯混合模型滤波的合成过采样方法。首先根据高斯混合模型的期望最大算法对原始数据集进行分组,并过滤掉每个分组中的噪声样本和边界样本,最后在每个分组内设置动态采样倍率来合成伪样本。Li 等[28]对 SMOTE 方法合成后的样本进行优化。首先根据样本的自然邻域分布来检测噪声样本,然后利用差分进化方法对噪声样本的位置进行优化。Park 等[29]提出了一种重标记采样方法,它不合成伪样本,而是从负类样本里过滤出疑似正类的样本,通过重新标记这些样本来达到类别平衡的目的。Liu 等[30]利用基于相对密度和绝对密度的噪声滤波器来去除噪声样本并计算稀疏性和边界权重,从而在正类样本的边界和稀疏区域合成更多的伪样本。

上述工作针对不平衡数据集的合成过采样方法提出了有效的改进措施,但大多数仍然在根样本选择、正类样本聚类、采样倍率计算方面依赖样本的空间距离,难以适应高维样本的合成过采样任务。此外,Zheng 等[31]提出了一种基于类间方差迁移的合成过采样方法。它将负类样本的偏移向量叠加至正类样本中心,丰富了少数类的特征空间。但在高维样本条件下,负类样本的协方差矩阵难以表示样本的分布特征。

3 基于无噪梯度分布的合成过采样方法

本章主要介绍提出的基于无噪梯度分布的合成过采样方法的具体细节和实现过程,包括相关定义、方法实现、方法框架和算法流程。

3.1 相关定义

定义 1(梯度贡献) 二分类交叉熵损失函数 L_{bce} 对样本预测概率 p 的导数称为样本的梯度贡献(Gradient Contribution,GC)。二分类交叉熵的损失函数如式(1)所示:

$$L_{bce}=-y\log p-(1-y)\log(1-p) \tag{1}$$

其中,y 是样本的标签,二分类时值为 0 或 1;p 为分类器将该样本预测为正类(标签为 1)的概率,$1-p$ 则表示分类器将该样本预测为负类的概率。

由于式(1)中预测概率 p 的取值为[0,1],首先需要证明 L_{bce} 函数在[0,1]内的连续性和可导性,本文利用极限的性质证明式(1)的连续性。将式(1)代入式(2)后,得到式(3)的形式,当 $h\to 0$ 时,式(3)的计算结果为 0,因此 L_{bce} 函数在[0,1]内是连续的。同时,由于 $\log(x)$ 在定义域内是可导的,而 L_{bce} 是 $\log(x)$ 函数的线性组合,因此 L_{bce} 在定义域内也是可导的。

$$L=\lim_{h\to 0}L_{bce}(p+h)-L_{bce}(p) \tag{2}$$

$$L=\lim_{h\to 0}-y\log\frac{p+h}{p}-(1-y)\log\frac{1-p+h}{1-p} \tag{3}$$

gc 的计算方式如式(4)所示。gc 的取值为[0,1],该指标反映的是样本对分类器的梯度贡献大小,即样本被拟合的难易程度。同时,其在低维样本的条件下也可以代表样本点到决策边界的距离。

$$gc=|p-y|=\begin{cases}p, & y=0\\ p-y, & y=1\end{cases} \tag{4}$$

定义 2(梯度区间) 将 0 到 1 内的梯度贡献 gc 分为等间隔的 k 个区间,同时设置安全梯度阈值 τ,$\tau\in[0,1]$。定义梯度贡献小于安全梯度阈值的区间为安全梯度区间(Safe Gradient Interval,SGI),梯度贡献大于安全梯度阈值的区间为危险梯度区间(Danger Gradient Interval,DGI)。定义离安全梯度阈值最近的安全梯度区间为临界安全梯度区间(Critical Safe Gradient Interval,CSGI),离安全梯度阈值最近的危险梯度区间为临界危险梯度区间(Critical Danger Gradient Interval,CDGI)。各定义区间的形式如图 5 所示。

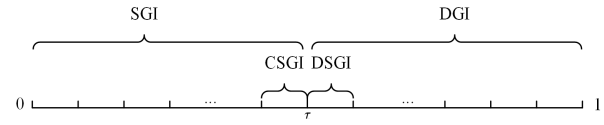


图 5 梯度区间定义

Fig. 5 Definition of gradient interval

定义 3(无噪梯度贡献) 因为数据集中存在噪声标签,所以直接按式(4)计算得到的样本梯度贡献也是带噪声的。假设直接计算得到数据集全部样本的是带噪梯度贡献,那么在过滤掉数据集中的噪声标签样本后,再利用式(4)计算得到的梯度贡献称为无噪梯度贡献。

3.2 NGDSO 方法实现

3.2.1 无噪梯度计算

将样本的梯度贡献作为样本的标签置信度度量指标,假设原始数据集为 D_{ori},标签为噪声的样本集为 D_{noise},过滤掉噪声样本后的数据集为 D_{cl}。

首先计算每个样本的梯度贡献指标,本文采用逻辑回归模型以交叉验证的方式计算得到每个样本的梯度贡献指标。由于逻辑回归模型的损失函数是交叉熵损失,而梯度贡献这一指标正是基于交叉熵损失提出的,通过交叉验证的方式可以获得更优的模型表现。也就是说,当所有样本交叉熵损失和为最小时,每个样本的梯度贡献指标更精确。

其次计算正类样本和负类样本的平均梯度贡献 $\overline{gc_{min}}$ 和 $\overline{gc_{maj}}$。将梯度贡献值大于另一类别平均梯度贡献且识别错误的样本定义为噪声标签样本。由式(4)可知,当梯度贡献值大于 0.5 时,代表该样本被错误识别,取阈值 0.5 和另一个类别的平均梯度贡献二者的最大值作为每个类别噪声标签的识别阈值,如式(5)所示:

$$D_{noise}=\begin{cases}gc>\max\{\overline{gc_{min}},0.5\}, & y=0\\ gc>\max\{\overline{gc_{max}},0.5\}, & y=1\end{cases} \tag{5}$$

将噪声标签样本过滤后的数据集表示为:

$$D_{cl}=D_{ori}-D_{noise} \tag{6}$$

在 D_{cl} 上再次利用逻辑回归模型训练所有样本,按照式(4)计算得到每个样本的无噪梯度贡献。

3.2.2 根样本与辅助样本选择策略

定义 2 中划分的每个梯度贡献区间用 I_m 表示,$m=0,1,2,\cdots,k$ 为区间索引。那么每个梯度贡献区间的梯度贡献范

围 gc_m 如式(7)所示：

$$m\times(1/k)\leqslant gc_m\leqslant(m+1)\times(1/k) \tag{7}$$

其中，$1/k$ 表示每个区间的梯度贡献间隔。每个正类样本 x_i 被分配到的区间如公式(8)所示：

$$x_i\in I_m, m=\left\lfloor\frac{gc_i}{1/k}\right\rfloor \tag{8}$$

其中，gc_i 表示第 i 个正类样本的梯度贡献。

由于梯度贡献较大的样本仍然可能为噪声样本，因此本文选择安全梯度区间内的样本作为根样本。该根样本选择策略既不依赖空间距离，又可以进一步避免可能的噪声样本作为根样本带来的误差累积。根样本的集合 x_{rs} 如式(9)所示：

$$x_{rs}=\{x_i, gc_i\leqslant\tau\} \tag{9}$$

定义 4(梯度右近邻) 某个样本的梯度右近邻定义为梯度贡献值大于该样本，且与该样本的梯度贡献差最小的样本，用 x_{grnn} 表示。用 gc_{rs} 表示根样本的梯度贡献，gc_{grnn} 表示梯度右近邻的梯度贡献，则梯度右近邻可表示为：

$$x_{grnn}=\arg\min(gc_{grnn}-gc_{rs}), gc_{grnn}\geqslant gc_{rs} \tag{10}$$

梯度右近邻不同于 K 近邻、自然近邻等邻居选择策略，它根据样本对模型的梯度贡献来确定邻居样本。该策略不仅不依赖样本的空间距离，而且梯度右近邻始终比根样本距离决策边界更近，合成的新样本也始终比根样本距离决策边界更近。梯度右近邻也可以很大程度地保证合成的新样本和根样本落在同一个梯度区间内。

对比图 3 中选择空间近邻为辅助样本与图 4 的聚类后合成，图 6 的梯度右近邻作为辅助样本合成的伪样本始终落在比根样本更靠近决策边界且不远于梯度右近邻样本的位置。该近邻的选择可以保证每一个合成的伪样本均使决策边界往负类样本移动，从而提升正类样本的识别性能。

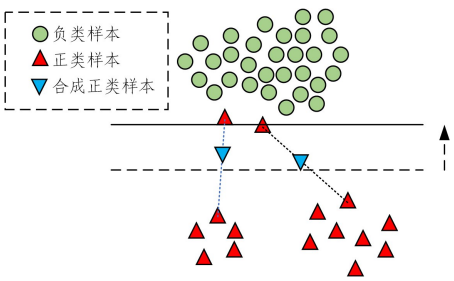


图 6 梯度右近邻合成方式

Fig. 6 Gradient right neighbor synthesis method

3.2.3 安全梯度分布近似策略

在 3.2.2 小节中确定了根样本和辅助样本的选择策略，每个安全梯度区间合成的样本数量则由本小节的安全梯度分布近似策略确定。

首先由式(11)计算需要合成的样本总数量 N_{tg}，其中 N_{maj} 代表原始数据集负类样本的总数量，N_{min} 代表原始数据集正类样本的总数量。

$$N_{tg}=N_{maj}-N_{min} \tag{11}$$

其次计算原始数据集中正类样本的安全梯度分布 $\boldsymbol{M}$，也就是每个安全梯度区间内的样本数量。假设安全梯度区间的个数是 $q(q\leqslant k)$，那么 $\boldsymbol{M}$ 的形式是一个 q 维的向量，表达式如式(12)所示：

$$\boldsymbol{M}=\{n_1, n_2, \cdots, n_q\}, q\leqslant k \tag{12}$$

其中，n_q 表示第 q 个安全梯度区间内的样本数量。假设合成后的正类样本安全梯度分布为 $\tilde{\boldsymbol{M}}$，合成后每个安全梯度区间内的样本数量为 $\tilde{n_q}$，则 $\tilde{\boldsymbol{M}}$ 的表达式如式(13)所示：

$$\tilde{\boldsymbol{M}}=\{\tilde{n_1}, \tilde{n_2}, \cdots, \tilde{n_q}\}, q\leqslant k \tag{13}$$

合成后的正类样本和原始正类样本在安全梯度区间内的分布近似是本文确定安全梯度分布近似策略的前提。默认数据集中正类样本在安全梯度区间内的梯度分布可以代表真实世界里的所有正类样本在安全梯度区间内的分布，也就是向量 $\boldsymbol{M}$ 和 $\tilde{\boldsymbol{M}}$ 是相似度极大的。

向量 $\tilde{\boldsymbol{M}}$ 内的元素，也就是合成后每个安全梯度区间内的样本数量为待求解元素。本文利用余弦相似度来度量这两个向量的相似度，当两个向量的余弦相似度为 1 时，两个向量的相似度最大，如式(14)所示：

$$\cos(\boldsymbol{M}, \tilde{\boldsymbol{M}})=\frac{\boldsymbol{M}\cdot\tilde{\boldsymbol{M}}}{\|\boldsymbol{M}\|\ \|\tilde{\boldsymbol{M}}\|}=1 \tag{14}$$

由式(14)可以求解合成后每个安全梯度区间内的样本总数量 $\tilde{n_q}$，如式(15)所示：

$$\tilde{n_q}=\frac{n_q}{\sum_1^q n_q}\times N_{maj}, q\leqslant k \tag{15}$$

每个安全梯度区间需要合成的样本数量 n_{tg}^q 由式(16)或式(17)计算得到：

$$n_{tg}^q=\tilde{n_q}-n_q, q\leqslant k \tag{16}$$

$$n_{tg}^q=N_{tg}\times\frac{n_q}{\sum_1^q n_q}, q\leqslant k \tag{17}$$

最后，通过随机线性插值的方式为每个安全梯度区间合成伪样本，如式(18)所示：

$$x_{ps}=x_{rs}+\gamma\times(x_{rs}-x_{grnn}) \tag{18}$$

其中，γ 表示 0 到 1 的随机数。

3.2.4 方法框架和算法流程

NGDSO 方法的整体框架如图 7 所示。原始不平衡数据集首先经过噪声标签过滤，然后基于正类样本的无噪梯度分布确定根样本、辅助样本和每个梯度区间需要合成的样本数量，最后经过线性插值方法得到类别平衡的数据集。

NGDSO 算法流程如算法 1 所示。

算法 1 基于无噪梯度分布的合成过采样算法

输入：不平衡数据集 D_{ori}，梯度分布计算分类器逻辑回归 f，安全梯度阈值 τ

输出：类别平衡数据集 D_{ft}

1. 初始化：正类样本集合←D_{min}，负类样本集合←D_{maj}，通过式(11)计算需要合成的样本数量←N_{tg}，合成样本集合 x_{ps}←NULL
2. 在原始数据集上以交叉验证的方式训练 f，通过式(1)和式(4)计算每个样本的梯度贡献和每个类别的平均梯度贡献$\overline{gc_{min}}$和$\overline{gc_{maj}}$
3. 按照式(5)和式(6)过滤数据集中的噪声样本
4. 在过滤后的数据集 D_{cl} 上再次训练 f，得到每个正类样本的无噪梯度贡献 gc_i
5. 通过式(8)，按照 gc_i 值将每个正类样本分配到 k 个区间内
6. 取梯度贡献值小于安全梯度阈值 τ 的 q 个安全梯度区间，通过

式(12)计算正类样本的安全梯度分布 **M**

7. 通过式(16)计算每个安全梯度区间需要合成的样本数量 n_{tg}^{q}

8. for i←1 to q do

9. 将第 i 个安全梯度区间内 D_p^i 内的样本按照梯度贡献值升序排序

10. 随机从 D_p^i 内选择 n_{tg}^i 个样本作为根样本→x_{rs}^j，j=1，2，…，n_{tg}^i

11. if x_{rs}^j 在 D_p^i 内存在梯度右近邻

12. 辅助样本 x_a^j←x_{grnn}^j（梯度右近邻）

13. else 辅助样本 x_a^j←x^j=argmin gc in D_p^{i+1}

14. 生成 n_{tg}^i 个 0～1 之间随机数的集合←γ^i

15. 通过式(18)线性插值合成 n_{tg}^i 个伪样本

16. 组合当前伪样本集合：x_{ps}←x_{ps}+x_{ps}^j

17. 组合合成的伪样本和原始数据集：D_{ft}←D_{maj}+D_{min}+x_{ps}

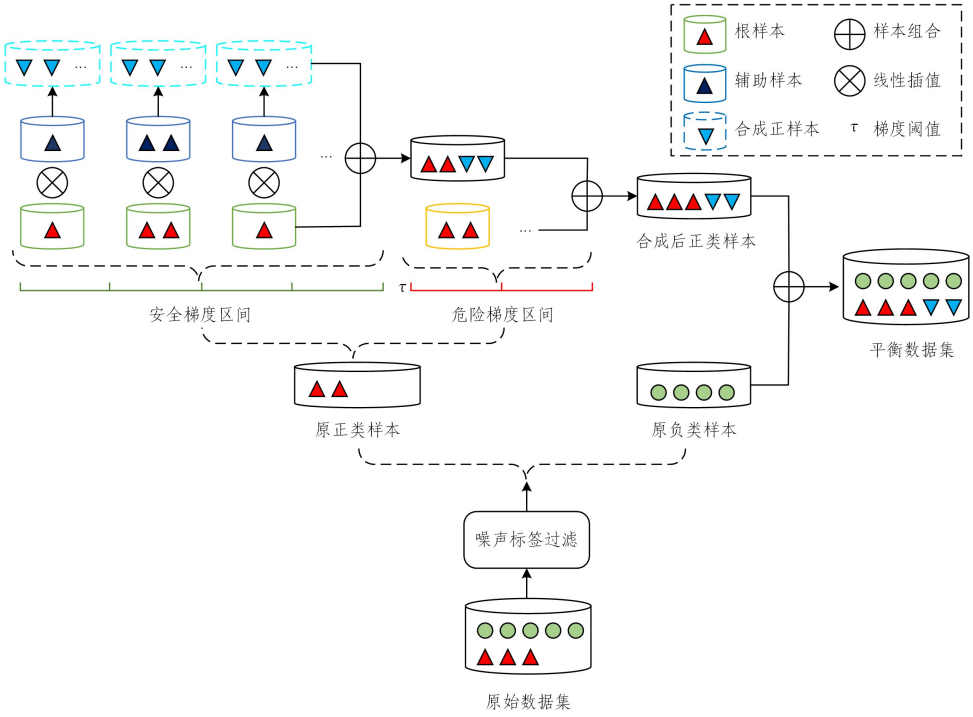


图 7 NGDSO 方法整体框架

Fig. 7 Overall framework of NGDSO

4 实验结果与分析

本章从实验环境和数据集、对比算法、测试模型及评价指标等方面介绍了实验相关细节，对实验结果进行了分析并讨论了 NGDSO 方法的超参数影响，最后分析了消融实验结果。

4.1 实验环境和数据集

实验在 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz CPU 与 32 GB 内存配置的电脑上运行，NGDSO 算法由 Python 语言编写实现，不依赖其他特殊的软硬件环境。

本文选取了两个来源于 UCI 平台的数据集，即 Australian(1 号)和 German(2 号)，以及一个来源于 Kaggle 平台的数据集 Credit Card Fraud Detection(3 号)。这 3 个数据集是真实金融场景中的数据集，1 号数据集是澳大利亚信贷数据集，2 号数据集是德国信贷数据集，3 号数据集是德国信用卡欺诈消费检测数据集。由于欺诈消费样本数量较少而正常消费样本数量较多，3 号数据集具有较高的不平衡率。此外，还选取了 4 个来自 KEEL 平台的数据集 Page-block0(4 号)，winequality-red-4(5 号)，Yeast6(6 号)，Abalone19(7 号)。这些经典数据集被频繁用于证明不平衡二分类方法的性能[17-31]，因此这些数据集上的实验结果可以证明合成过采样方法对分类器的性能提升情况。这 7 个数据集的样本数量、特征维度以及不平衡率信息如表 1 所列。

表 1 实验数据集信息

Table 1 Details of experimental datasets

数据集	样本数量	特征维度	不平衡率
Australian	690	14	1.25
German	1 000	24	2.33
Page-block0	5 472	10	8.79
winequality-red-4	1 599	10	29.17
Yeast6	1 484	8	41.40
Abalone19	4 174	8	129.44
Credit Card Fraud Detection	284 807	28	577.88

4.2 对比算法

实验选择了 8 种用于不平衡分类的主流合成过采样算法作为对比算法。其中包括 5 个基于样本空间特征的方法，分别是 SMOTE(Synthetic Minority Over-sampling Technique)，ADASYN(Adaptive Synthetic Sampling)，BL-SMOTE(Borderline-SMOTE)，SVM-SMOTE(Support Vector Machine SMOTE)，SMOTEN(SMOTE for Nominal)，1 个聚类合成的方法 KM-SMOTE(Kmeans-SMOTE)，以及 2 个混合采样的方法 SMOTE-Tomek 和 SMOTEENN(SMOTE-Edited Nearest Neighbors)。

这 8 种合成过采样算法均被证明对不平衡数据集有一定的优化效果，可以不同程度地提升分类器的性能，这些算法的主要实现思想如表 2 所列。可以看见，这 8 种对比算法虽然被广泛应用，但在选择根样本和辅助样本或者确定合成样本

数量时严重依赖样本的空间特征。因此这些算法适用于解决低维数据集的类别不平衡问题，但在高维数据集下，这些算法的性能受到高维空间距离度量的限制。

表 2 对比算法实现思想

Table 2 Idea of comparison methods

算法名称	核心思路
SMOTE	为所有的正类样本合成相同数量的伪样本，并随机选择 K 近邻中的一个样本作为辅助样本
ADASYN	利用空间近邻方法计算每个正类样本的局部密度，根据局部密度确定每个正类样本合成的伪样本数量，辅助样本是 K 近邻中的随机样本
BL-SMOTE	利用空间近邻识别出边界样本，为边界样本使用 SMOTE 方法合成伪样本
SVM-SMOTE	为支持向量样本利用 SMOTE 方法合成伪样本
SMOTEN	以 SMOTE 为基础，部分合成样本的特征值从原样本中选择，而不是线性计算得到
KM-SMOTE	利用 K-Means 方法对正类样本聚类，根据子类的稀疏程度确定其合成样本的数量，然后利用 SMOTE 方法合成
SMOTE-Tomek	首先利用 SMOTE 方法合成伪样本，再剔除那些在空间上互为近邻但属于不同类别的 Tomek-Links 对
SMOTEENN	首先用 SMOTE 方法合成伪样本，再根据空间 ENN 方法剔除噪声和重复样本

4.3 测试模型及评价指标

本文实验选择了 Logistic Regression(LR)和集成学习分类器 Light-GBM(LGBM)来测试 NGDSO 算法对类别不平衡数据集优化的效果。LR 因为对预测结果的可解释性，仍然在金融风险控制领域被广泛应用。同时，它作为大部分神经网络的基本组成单元，可以体现合成过采样方法对分类器性能提升的作用。LGBM 是一种 boosting 集成思想的模型，它基于梯度提升框架来迭代训练决策树模型从而优化损失函数值，由于其在效率和精度上具有优势，因此在现实场景中取得了优异表现。LR 分类器的正则化类型为 L1，优化器为 liblinear，最大迭代次数为 200，其他参数为通用默认参数。LGBM 的叶子数量为 30，学习率为 0.05，基分类器的个数为 30。

本文除了选取经常被用于证明分类器性能的 F1-Score，Recall，AUC 指标外，也着重考虑了能反映类别不平衡条件下分类器性能的 G-Mean，MCC，KS 指标。

4.4 实验结果

实验部分利用 8 种对比算法和 NGDSO 算法对 7 个数据集分别进行合成过采样操作，然后测试合成过采样后的数据集对 LR 和 LGBM 两个分类器在 6 个性能指标上的提升情况，实验结果如表 3－表 9 所列。

从表 3 可以看出，对于 LR 分类器而言，NGDSO 算法，在安全梯度阈值取 0.8 时相较于其他 8 种合成过采样算法在 F1-Score，Recall，AUC，G-Mean，MCC 和 KS 这 6 个指标上全部取得了最优值，分别比其他最优指标提高了 1.56，2.47，0.67，1.63，2.36，2.38 个百分点；对于 LGBM 分类器而言，NGGSO 算法在安全梯度阈值取 0.9 时获得了最优的 F1-Score，Recall，MCC 值，分别比其他最优指标提高了 0.06，1.23，0.57 个百分点，同时取得了次优的 AUC 和 G-Mean 值。

表 3 数据集 1 上的实验结果

Table 3 Experimental results on dataset 1

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.8263	0.8519	0.9208	0.8584	0.7100	0.7416
	BL-SMOTE	0.8148	0.8148	0.9130	0.8472	0.6958	0.7372
	SVM-SMOTE	0.8098	0.8148	0.9169	0.8434	0.6864	0.7337
	ADASYN	0.8199	0.8148	0.9199	0.8510	0.7053	0.7496
	SMOTEN	0.8415	0.8519	0.9195	0.8702	0.7376	0.7619
	KM-SMOTE	0.8383	0.8642	0.9220	0.8686	0.7301	0.7496
	SMOTE-Tomek	0.8293	0.8395	0.9192	0.8600	0.7174	0.7390
	SMOTE-ENN	0.7123	0.6420	0.8278	0.7588	0.5666	0.5679
	NGDSO(0.8)	**0.8571**	**0.8889**	**0.9278**	**0.8849**	**0.7612**	**0.7857**
LGBM	SMOTE	0.8024	0.8272	0.9298	0.8381	0.6698	0.7593
	BL-SMOTE	0.8171	0.8272	0.9253	0.8498	0.6972	0.7381
	SVM-SMOTE	0.8214	0.8519	0.9239	0.8545	0.7010	0.7557
	ADASYN	0.8144	0.8395	0.9272	0.8483	0.6899	0.7434
	SMOTEN	0.8313	0.8519	0.9250	0.8624	0.7191	0.7496
	KM-SMOTE	0.8193	0.8395	0.9279	0.8522	0.6990	0.7672
	SMOTE-Tomek	0.8220	0.8271	0.9271	0.8536	0.7066	0.7381
	SMOTE-ENN	0.7778	0.6914	0.9102	0.8081	0.6744	0.7743
	NGDSO(0.9)	**0.8391**	**0.8642**	0.9284	0.8607	**0.7248**	0.7293

在 2 号数据集上，在 NGDSO 算法安全梯度阈值取 0.7 时，LR 同样取得了全部 6 个指标的最优值，分别比其他最优指标提高了 5.64，2.17，0.79，4.88，8.93，4.56 个百分点。在 NGDSO 算法安全梯度阈值取 0.9 时，LGBM 取得了最优的 F1-Score 和 MCC 指标，分别比其他最优指标提高了 0.04 和 0.11 个百分点，同时取得了次优的 Recall 和 G-Mean 值。但当安全梯度阈值取 1 时，LGBM 取得了最优的 Recall 值，比其他最优指标提升了 1.09 个百分点。

在 3 号数据集上，对于 LR 而言，在 NGDSO 算法安全梯度阈值取 1 时，取得了最优的 G-Mean 和 KS 值，分别比其他最优指标提高了 0.99 和 0.45 个百分点，Recall 指标仅次于 ADASYN 算法，但 F1-Score 指标比 ADASYN 算法高出 14.42 个百分点；当安全梯度阈值取 0.9 时，NGDSO 算法取得了最优的 F1-Score 和 MCC 值，分别比其他最优指标提高了 8.75 和 8.02 个百分点。对于 LGBM 而言，在 NGDSO 算法安全梯度阈值取 0.5 时，取得了最优的 F1-Score，AUC 和 MCC 值，分别比其他最优算法提高 3.04，0.19 和 3.07 个百分点，同时取得了仅次于 SMOTE-ENN 的 KS 值；但当安全

梯度阈值取 1 时，LGBM 取得了仅次于 ADASYN 算法的 Recall 指标值，但 NGDSO 算法的 F1-Score 指标比 ADASYN 提高了 22.53 个百分点。

在 4 号数据集上，对于 LR 而言，在 NGDSO 算法安全梯度阈值取 0.8 时，取得了最优的 MCC 指标、仅次于 KM-SMOTE 的 F1-Score 值和仅次于 SMOTE-ENN 的 AUC 指标，但 NGDSO 的 Recall 值比 KM-SMOTE 高出 19.76 个百分点；当安全梯度阈值取 1 时，NGDSO 算法取得了最优的 Recall 指标和仅次于 SMOTE-ENN 的 KS 指标，Recall 指标比最优的 ADASYN 方法提升了 0.27 个百分点，同时 F1-Score 也提升了 6.55 个百分点。对于 LGBM 而言，在 NGDSO 算法安全梯度阈值取 1 时，NGDSO 算法取得了最优的 Recall 和 AUC 指标、仅次于 KM-SMOTE 的 MCC 指标和 SMOTE-Tomek 的 KS 指标。ADASYN 方法取得了与 NGDSO 算法一致的 Recall 指标，但 NGDSO 算法的 F1-Score 和 MCC 指标分别比 ADASYN 方法高出了 5.16 和 5.10 个百分点。SMOTEN 和 KM-SMOTE 方法的 F1-Score 比 NGDSO 算法高，但它们的 Recall 指标分别比 NGDSO 方法低 8.73 和 8.72 个百分点。

表 4　数据集 2 上的实验结果

Table 4　Experimental results on dataset 2

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.6263	0.6739	0.7936	0.7289	0.4460	0.4987
	BL-SMOTE	0.6051	0.6413	0.7962	0.7111	0.4174	0.5073
	SVM-SMOTE	0.6146	0.6413	0.8015	0.7176	0.4345	0.5194
	ADASYN	0.6207	0.6848	0.7970	0.7258	0.4336	0.4929
	SMOTEN	0.6138	0.6304	0.7509	0.7157	0.4366	0.4707
	KM-SMOTE	0.5970	0.6522	0.7880	0.7061	0.3994	0.4967
	SMOTE-Tomek	0.6176	0.6848	0.7843	0.7235	0.4282	0.4904
	SMOTE-ENN	0.6188	0.7500	0.7696	0.7256	0.4202	0.4760
	NGDSO(0.7)	**0.6827**	**0.7717**	**0.8094**	**0.7777**	**0.5259**	**0.5650**
LGBM	SMOTE	0.6129	0.6195	0.7925	0.7137	0.4391	0.4728
	BL-SMOTE	0.6489	0.6630	0.8057	0.7426	0.4891	0.5314
	SVM-SMOTE	0.6452	0.6522	0.8140	0.7386	0.4858	0.4971
	ADASYN	0.6321	0.6630	0.8050	0.7318	0.4593	0.5169
	SMOTEN	0.5632	0.5326	0.7588	0.6694	0.3869	0.3995
	KM-SMOTE	0.5829	0.5543	0.7904	0.6849	0.4128	0.4714
	SMOTE-Tomek	0.6161	0.6630	0.7921	0.7208	0.4309	0.5086
	SMOTE-ENN	0.5625	0.7826	0.7522	0.6606	0.3152	0.4158
	NGDSO(0.9)	**0.6493**	0.6704	0.7936	0.7393	**0.4902**	0.5019
	NGDSO(1)	0.5703	**0.7935**	0.7489	0.6681	0.3293	0.4049

表 5　数据集 3 上的实验结果

Table 5　Experimental results on dataset 3

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.1001	0.9265	0.9812	0.9497	0.2179	0.9190
	BL-SMOTE	0.1604	0.8529	0.9328	0.9171	0.2725	0.8458
	SVM-SMOTE	0.2022	0.8897	0.9596	0.9380	0.3163	0.8871
	ADASYN	0.0309	0.9559	0.9806	0.9299	0.1160	0.9181
	SMOTEN	0.7234	0.6250	0.9724	0.7905	0.7322	0.9108
	KM-SMOTE	0.7235	0.7794	0.9706	0.8826	0.7249	0.8935
	SMOTE-Tomek	0.1001	0.9265	0.9810	0.9497	0.2179	0.9189
	SMOTE-ENN	0.0957	0.9265	0.9809	0.9491	0.2127	0.9193
	NGDSO(1)	0.1751	0.9338	0.9806	**0.9596**	0.2979	**0.9238**
	NGDSO(0.9)	**0.8110**	0.8676	0.9780	0.9313	**0.8124**	0.9066
LGBM	SMOTE	0.2222	0.9044	0.9772	0.9463	0.3361	0.9094
	BL-SMOTE	0.7870	0.8014	0.9763	0.8951	0.7868	0.8829
	SVM-SMOTE	0.4552	0.8971	0.9579	0.9456	0.5220	0.8984
	ADASYN	0.0601	0.9412	0.9778	0.9471	0.1663	0.9067
	SMOTEN	0.7955	0.7721	0.9564	0.8786	0.7955	0.9011
	KM-SMOTE	0.8192	0.8162	0.9597	0.9033	0.8189	0.9093
	SMOTE-Tomek	0.1814	0.8824	0.9816	0.9334	0.2963	0.9020
	SMOTE-ENN	0.1820	0.8824	0.9811	0.9335	0.2968	0.9146
	NGDSO(1)	**0.8496**	0.8309	**0.9835**	0.9114	**0.8496**	0.9098
	NGDSO(1)	0.2854	0.9044	0.9795	**0.9476**	0.3897	0.9140

在 5 号数据集上，对 LR 分类器而言，当安全梯度阈值取 1 时，NGDSO 方法取得了最优的 F1-Score，AUC 和 MCC 指标，分别比其他最优指标提升了 1.92，0.24 和 0.72 个百分点，同时取得了仅次于 SMOTE-ENN 的 Recall 指标，但 NGDSO 的 F1-Score 指标比 SMOTE-ENN 高出 12.9 个百分点。NGDSO 也取得了仅次于 KM-SMOTE 的 KS 指标，但 NGDSO 的 F1-Score，Recall，AUC 和 MCC 指标均高于 KM-SMOTE 方法。对于 LGBM 分类器而言，NGDSO 取得了最优的 F1-Score，Recall，AUC 和 MCC 指标，分别比其他最优指标提升了 2.57，1.26，0.89 和 1.86 个百分点，同时取得了仅次于

SMOTE 方法的 G-Mean 指标和仅次于 SVM-SMOTE 方法的 KS 指标，但 NGDSO 方法的其他指标均高于这两种方法。

表 6 数据集 4 上的实验结果

Table 6 Experimental results on dataset 4

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.6726	0.8837	0.9583	0.8982	0.6487	0.8065
	BL-SMOTE	0.6585	0.9360	0.9575	0.9147	0.6439	0.8313
	SVM-SMOTE	0.6696	0.8953	0.9605	0.9021	0.6476	0.8248
	ADASYN	0.6279	0.9419	0.9584	0.9084	0.6156	0.8272
	SMOTEN	0.5761	0.9128	0.9289	0.8824	0.5597	0.8256
	KM-SMOTE	0.7301	0.6919	0.9466	0.8218	0.7017	0.7953
	SMOTE-Tomek	0.6756	0.8837	0.9577	0.8989	0.6516	0.8041
	SMOTE-ENN	0.6825	0.9186	0.9637	0.9141	0.6641	0.8511
	NGDSO(0.8)	0.7002	0.8895	0.9606	0.9065	**0.7076**	0.8221
	NGDSO(1)	0.6934	**0.9446**	0.9549	0.8972	0.6928	0.8359
LGBM	SMOTE	0.8165	0.9186	0.9859	0.9397	0.7987	0.9100
	BL-SMOTE	0.7845	0.9419	0.9863	0.9441	0.7685	0.9016
	SVM-SMOTE	0.8152	0.9360	0.9895	0.9469	0.7990	0.9044
	ADASYN	0.7762	0.9477	0.9890	0.9449	0.7610	0.9090
	SMOTEN	0.8385	0.8604	0.9832	0.9171	0.8104	0.8967
	KM-SMOTE	0.8810	0.8605	0.9871	0.9226	0.8677	0.9105
	SMOTE-Tomek	0.8189	0.9070	0.9862	0.9350	0.8002	0.9200
	SMOTE-ENN	0.7845	0.9419	0.9792	0.9441	0.7685	0.8986
	NGDSO(1)	0.8278	**0.9477**	**0.9896**	**0.9527**	0.8120	0.9182

表 7 数据集 5 上的实验结果

Table 7 Experimental results on dataset 5

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.1657	0.8824	0.8785	0.7736	0.2185	0.6064
	BL-SMOTE	0.1940	0.7647	0.8693	0.7700	0.2325	0.6200
	SVM-SMOTE	0.2553	0.7059	0.8572	0.7790	0.2848	0.6022
	ADASYN	0.1404	0.7059	0.8666	0.6996	0.1581	0.5940
	SMOTEN	0.0678	0.1176	0.6655	0.3278	0.0204	0.3128
	KM-SMOTE	0.2027	0.8824	0.8663	0.8132	0.2622	0.7009
	SMOTE-Tomek	0.1582	0.8235	0.8674	0.7509	0.1993	0.6150
	SMOTE-ENN	0.1455	0.9412	0.8840	0.7490	0.2010	0.6690
	NGDSO(1)	**0.2745**	0.9276	**0.8864**	0.7637	**0.2902**	0.6913
LGBM	SMOTE	0.2192	0.4706	0.7359	0.6494	0.2113	0.4477
	BL-SMOTE	0.1013	0.2353	0.5363	0.4537	0.0606	0.1100
	SVM-SMOTE	0.0541	0.0588	0.7704	0.2375	0.0165	0.4808
	ADASYN	0.1579	0.3529	0.7204	0.5591	0.1342	0.4088
	SMOTEN	0.1081	0.1176	0.5456	0.3363	0.0729	0.1225
	KM-SMOTE	0.1509	0.2353	0.7361	0.4680	0.1166	0.4358
	SMOTE-Tomek	0.1333	0.2941	0.7630	0.5103	0.1019	0.4456
	SMOTE-ENN	0.1429	0.4118	0.7397	0.5882	0.1243	0.4537
	NGDSO(1)	**0.2449**	**0.4832**	**0.7793**	0.6371	**0.2299**	0.4592

表 8 数据集 6 上的实验结果

Table 8 Experimental results on dataset 6

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.2593	0.7000	0.9127	0.8004	0.3054	0.6881
	BL-SMOTE	0.2979	0.7000	0.9094	0.8073	0.3388	0.7083
	SVM-SMOTE	0.2727	0.6000	0.9115	0.7493	0.2989	0.6945
	ADASYN	0.2154	0.7000	0.9216	0.7893	0.2656	0.7087
	SMOTEN	0.2133	0.8000	0.9103	0.8339	0.2808	0.7202
	KM-SMOTE	0.2609	0.6000	0.8748	0.7475	0.2887	0.6463
	SMOTE-Tomek	0.2545	0.7000	0.9202	0.7994	0.3012	0.6995
	SMOTE-ENN	0.2414	0.7000	0.9140	0.7964	0.2895	0.6991
	NGDSO(1)	**0.3095**	**0.8000**	**0.9222**	0.8333	**0.3487**	**0.7202**
LGBM	SMOTE	0.4667	0.7000	0.7810	0.8241	0.4794	0.6771
	BL-SMOTE	0.5217	0.6000	0.8429	0.7684	0.5139	0.5991
	SVM-SMOTE	0.5000	0.6000	0.8615	0.7675	0.4939	0.6280
	ADASYN	0.4667	0.7000	0.8135	0.8241	0.4794	0.6771
	SMOTEN	0.3636	0.4000	0.8581	0.6266	0.3492	0.6349
	KM-SMOTE	0.5556	0.5000	0.8558	0.7047	0.5501	0.6028
	SMOTE-Tomek	0.4375	0.7000	0.7883	0.8221	0.4551	0.6794
	SMOTE-ENN	0.3043	0.7000	0.8463	0.8084	0.3443	0.6518
	NGDSO(1)	0.5217	**0.7000**	0.8469	**0.8253**	**0.5596**	0.6667

在 6 号数据集上，对于 LR 分类器而言，当安全梯度阈值取 1 时，NGDSO 取得了除 G-Mean 指标外的全部最优指标，NDOSO 方法的 G-Mean 指标仅次于 SMOTEN 方法，但 NGDSO 方法的其他指标均高于 SMOTEN 方法。对于 LGBM 分类器而言，NGDSO 取得了最优的 Recall，G-Mean 和 MCC 指标和仅次于 KM-SMOTE 方法的 F1-Score 指标，但 NGDSO 的 Recall 指标比 KM-SMOTE 方法高出了 20 个百分点。

表 9 数据集 7 的实验结果

Table 9 Experimental results on dataset 7

分类器	算法	*F1-Score*	*Recall*	*AUC*	*G-Mean*	*MCC*	*KS*
LR	SMOTE	0.0296	0.6667	0.8107	0.7264	0.0776	0.5703
	BL-SMOTE	0.0254	0.5000	0.7843	0.6395	0.0567	0.6160
	SVM-SMOTE	0.0240	0.3333	0.7819	0.5393	0.0425	0.6152
	ADASYN	0.0284	0.6667	0.8090	0.7220	0.0747	0.5559
	SMOTEN	0.0215	0.1667	0.7164	0.3939	0.0265	0.5512
	KM-SMOTE	0.0432	0.5000	0.7963	0.6692	0.0887	0.5469
	SMOTE-Tomek	0.0294	0.6667	0.8098	0.7257	0.0771	0.5487
	SMOTE-ENN	0.0256	0.6667	0.8249	0.7108	0.0682	0.6047
	NGDSO(1)	**0.0308**	**0.6667**	0.8001	**0.7301**	0.0801	0.5791
LGBM	SMOTE	0.0373	0.5000	0.6806	0.6626	0.0793	0.3781
	BL-SMOTE	0.0449	0.3333	0.6242	0.5583	0.0745	0.2684
	SVM-SMOTE	0.0506	0.3333	0.6267	0.5607	0.0815	0.2764
	ADASYN	0.0373	0.5000	0.6792	0.6626	0.0793	0.3781
	SMOTEN	0	0	0.7718	0	−0.0052	0.6251
	KM-SMOTE	0.0317	0.1667	0.6238	0.3990	0.0403	0.2708
	SMOTE-Tomek	0.0256	0.3333	0.6727	0.5420	0.0456	0.3420
	SMOTE-ENN	0.0302	0.5000	0.6634	0.6510	0.0665	0.3476
	NGDSO(1)	**0.0580**	**0.6667**	**0.7966**	**0.7735**	**0.1268**	**0.6274**

在 7 号数据集上，NGDSO 方法使 LGBM 分类器取得了全部指标的最优值，而且每个指标相对于次优指标都有明显的性能提升，F-Score，Recall，AUC，G-Mean，MCC 和 KS 指标分别提升了 1.31，16.67，2.48，11.09，4.53，0.23 个百分点。对于 LR 分类器而言，NGDSO 方法取得了最优的 F1-Score，Recall 和 G-Mean 指标，同时取得了仅次于 KM-SMOTE 的 MCC 指标。

4.5 实验结果分析

1）NGDSO 算法的优势。在小规模数据集上，NGDSO 算法既可以取得比其他算法更高的 Recall 值，也可以取得比其他算法更高的 F1-Score，MCC 和 KS 值。也就是说，NGDSO 算法在取得最优的正类样本识别精度的同时，也提高了负类样本的识别精度，克服了其他合成过采样算法能提升正类样本识别精度但会损失负类样本识别精度的问题。在大规模数据集上，NGDSO 算法可以取得非常均衡的指标表现，但对 F1-Score 指标的提升非常明显，说明 NGDSO 算法对不同类别的样本识别能力也非常均衡。同时，NGDSO 算法可以根据场景任务的特殊要求来调节安全梯度阈值，从而达到不同类别的样本识别效果。

2）NGDSO 算法对不同复杂度分类器的提升性能。从表 3 和表 4 可以看出，在小规模数据集上，NGDSO 算法对结构相对简单的 LR 分类器的性能提升在全部指标上非常明显。对结构更复杂的 LGBM 分类器在大多数性能指标上也有所提升，但不如 LR 分类器提升效果明显。但在大规模数据集上，NGDSO 算法对不同结构复杂度的分类器的性能提升效果比较相似。

4.6 超参数分析

NGDSO 算法只有一个需要调节的超参数，即安全梯度阈值 τ，本文测试了从 0.4 到 1 内间隔为 0.1 的超参数取值对数据集优化效果的影响，1 号－3 号数据集上不同超参数的测试结果如图 8 所示。

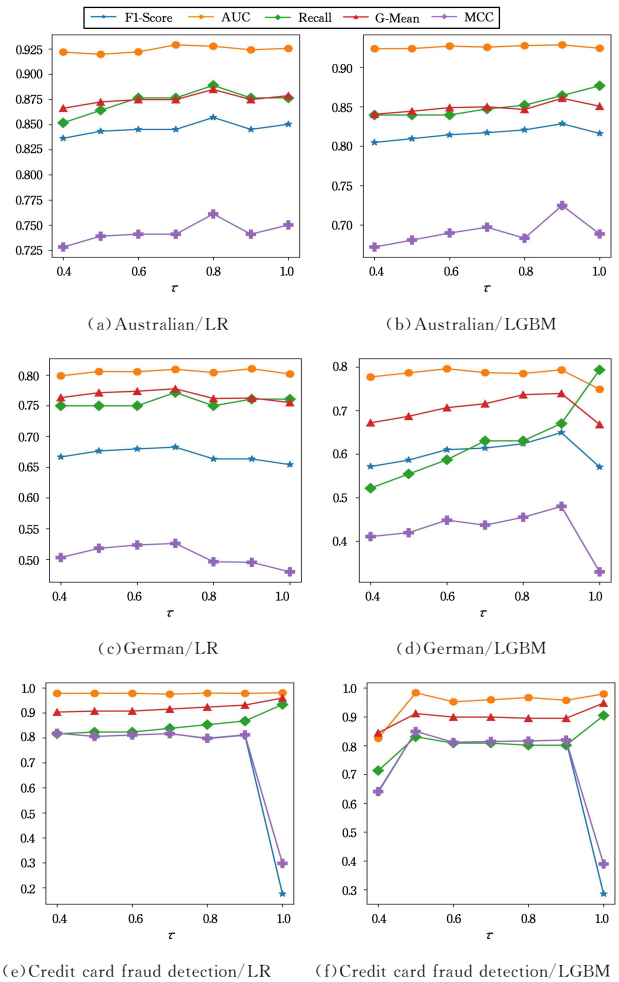


图 8 超参数测试结果

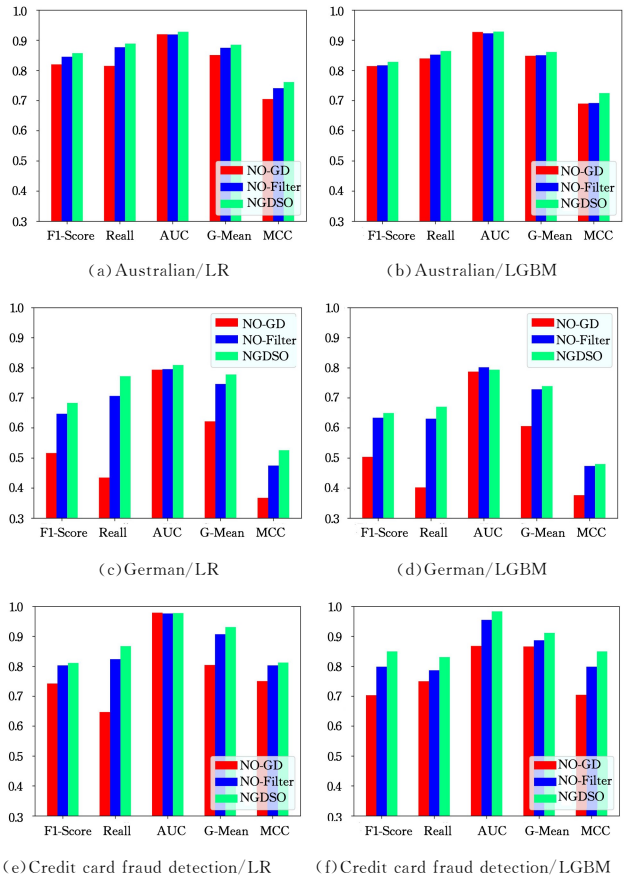
Fig. 8 Hyperparameter test results

由图 8 超参数测试结果可知，在 1 号数据集上，LR 和

LGBM 分类器分别在安全梯度阈值取 0.8 和 0.9 时取得最优分类器性能。在 2 号数据集上对应的最优安全梯度阈值分别是 0.7 和 0.9，在 3 号数据集上对应的最优安全梯度阈值均为 0.9。超参数的测试结果说明在不同数据集上，最优的安全梯度阈值是不同的。在小规模数据集上，梯度贡献较大($gc \geqslant 0.9$)的正类样本被当作根样本时，分类器的性能变化不大。而在大规模数据集上，当梯度贡献较大($gc \geqslant 0.9$)的正类样本被当作根样本时，分类器的 Recall 指标仍会上升，但 F1-Score 指标会迅速下降。该现象表明，样本被当作根样本合成过采样后，严重降低了负类样本的识别性能，可以将这些样本视为噪声标签样本。此外，该现象也说明了梯度贡献作为度量样本标签置信度指标的合理性。

4.7 消融实验

为了验证 NGDSO 算法的噪声标签过滤和梯度分布合成过采样这两个关键策略对算法的作用，本文对这两个策略进行了消融实验。只进行噪声标签过滤的实验结果称为 NO-GD，只进行梯度分布合成过采样的实验结果称为 NO-Filter，NGDSO 则表示整体算法的实验结果。消融实验的结果如图 9 所示。



(a) Australian/LR (b) Australian/LGBM

(c) German/LR (d) German/LGBM

(e) Credit card fraud detection/LR (f) Credit card fraud detection/LGBM

图 9 消融实验结果

Fig. 9 Ablation experiment results

由图 9 的消融实验结果可知，NDGSO 算法的两个策略对分类器的指标均有提升作用。大部分情况下，独立的梯度分布合成策略结果更接近 NDGSO 算法的性能，噪声标签过滤策略可以对梯度分布合成的结果进行进一步优化，但噪声标签样本过滤和梯度分布合成两个策略的作用不是简单的叠加。本文的噪声标签是广义的概念，包括错误标注、正确标注但发生概率很小的样本、部分类别重叠的样本等，而梯度分布合成时本身也会考虑部分噪声标签样本带来的误差累积。因此，噪声标签过滤和梯度分布合成对整体算法的性能提升是一个复杂的相互作用。

结束语 为解决当前合成过采样方法依赖空间距离而不适用于高维样本、噪声标签导致的误差累积，以及过采样后的样本分布降低负类样本识别精度的问题，提出了一种基于无噪梯度分布的合成过采样方法。首先利用样本的梯度贡献属性过滤数据集中的噪声标签样本，再计算正类样本的无噪梯度分布，初步避免了噪声样本造成的误差累积。其次，基于无噪梯度分布设计了新的不依赖于空间距离的根样本以及辅助样本选择策略，进一步避免了噪声样本造成的误差累积，并且使合成的样本能不断促进决策边界向负类样本移动，提高了正类样本的识别精度。最后设计了基于余弦相似度的安全梯度分布近似策略，该策略合成后的样本分布使决策边界以安全的方式向负类样本移动。因此，NGDSO 算法不仅能提升正类样本的识别精度，而且不会明显牺牲负类样本的识别精度，该特性可以使 NGDSO 算法很好地适应于当前国家的宽信用政策，既能进行精准风控，又能识别优质信贷客户的需求。

本文提出的 NGDSO 方法需要在不同的数据集上调节超参数 τ，增加了建模工作的工作量。由于多分类问题通常可以被拆解成多个二分类问题，因此本文未详细验证所提方法对于多分类问题的性能。同时，NGDSO 方法不能很好地适用于梯度分布比较极端的数据集，如何应对极端梯度分布条件下的合成过采样任务，是本文下一步考虑的研究方向。

参 考 文 献

[1] TIAN Y, BIAN B, TANG X F, et al. A new non-kernel quadratic surface approach for imbalanced data classification in online credit scoring[J]. Information Science, 2021, 563: 150-165.

[2] CHARIZANOS G, DEMIRHAN H, ICEN D. An online fuzzy fraud detection framework for credit card transactions[J]. Expert Systems With Applications, 2024, 252(PA): 124127.

[3] REB H J, TANG Y H, DONG W Y, et al. Dynamic ensemble handling class imbalance in network intrusion detection[J]. Expert Systems With Applications, 2023, 229(PA): 120420.

[4] WANG C J, XIN C, XU Z L. A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification [J]. Knowledge-Based Systems, 2021, 220: 106925.

[5] BARUA S, ISIAM M M, YAO X, et al. MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 26(2): 405-425.

[6] NEKOOEIMEHR I, SUSANA K, LAI Y. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets[J]. Expert Systems With Applications, 2016, 46: 405-416.

[7] WANG X X, LI L X, LIN H. A Review of SMOTE Algorithm Research[J]. Journal of Frontiers of Computer Science & Tech-

nology,2024,18(5):1135-1159.
[8] CHAWLA N,BOWYER W K,HAALL O L,et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. The Journal of Artificial Intelligence Research,2002,16:321-357.
[9] HE H B,BAI Y,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. 2008: 1322-1328.
[10] HAN H,WANG W Y,MAO B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [C]// Lecture Notes in Computer Science. 2005:878-887.
[11] NGUYEN H M,COOPER E W,KAMEI K. Borderline oversampling for imbalanced data classification[J]. International Journal of Knowledge Engineering and Soft Data Paradigms, 2011,3(1):4-21.
[12] BUNKHUMPORNPAT C,SINAPIROMSARA K,LURSINAP C. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem[C]// 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 2009:475-482.
[13] ZHENG J,QU H C,LI Z N,et al. A novel autoencoder approach to feature extraction with linear separability for high-dimensional data[J]. PeerJ Computer Science,2022,8:e1061.
[14] LI B Y,LIU Y,WANG X G. Gradient Harmonized Single-Stage Detector[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:8577-8584.
[15] CHEN Y Q,PEDRYCZ W,YANG J. A new boundary-degree based oversampling method for imbalanced data[J]. Applied Intelligence,2023,53(22):26518-26541.
[16] LI J N,ZHU Q S,WU Q W,et al. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors[J]. Information Science,2021,565:438-455.
[17] WANG W T,YANG L J,ZHANG J H,et al. Natural local density-based adaptive oversampling algorithm for imbalanced classification[J]. Knowledge-Based Systems,2024,295:111845.
[18] LI M,ZHOU H,LIU Q,et al. WRND: A weighted oversampling framework with relative neighborhood density for imbalanced noisy classification[J]. Expert Systems With Applications, 2024,241:122593.
[19] LENG Q K,GUO J M,JIAO E J,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] YAN Y T,JIANG Y F,ZHENG Z,et al. LDAS: Local density based adaptive sampling for imbalanced data classification[J]. Expert Systems with Applications,2022,191:116213.
[21] TAO X,ZHANG X,ZHENG Y,et al. A Mean Shift-guided oversampling with self-adaptive sizes for imbalanced data classification[J]. Information Science,2024,672:120699.
[22] ZHANG Z,TIAN H P,JIN J S. Multiple adaptive over-sampling for imbalanced data evidential classification[J]. Engineering Applications of Artificial Intelligence,2024,133(F):108532.
[23] SUN L,LI M M,DING W P,et al. AFNFS: Adaptive fuzzy neighborhood-based feature selection with adaptive synthetic over-sampling for imbalanced data[J]. Information Science, 2022,612:724-744.
[24] MOUTAOUAKIL K,ROUDANI M,QUISSARI A. Optimal Entropy Genetic Fuzzy-C-Means SMOTE(OEGFCM-SMOTE) [J]. Knowledge-Based Systems,2023,262:110235.
[25] MENG D X,LI Y J. An imbalanced learning method by combining SMOTE with Center Offset Factor[J]. Applied Soft Computing,2022,120:108618.
[26] WANG X L,GONG J,SONG Y,et al. Adaptively weighted three-way decision oversampling: A cluster imbalanced-ratio based approach[J]. Applied Intelligence,2022,53(1):312-335.
[27] XU Z Z,SHEN D R,KOU Y,et al. A Synthetic Minority Oversampling Technique Based on Gaussian Mixture Model Filtering for Imbalanced Data Classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3740-3753.
[28] LI J N,ZHU Q S,WU Q W,et al. SMOTE-NaN-DE: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution[J]. Knowledge-Based Systems,2021,223:107056.
[29] PARK S,LEE H,IM J. Relabeling & raking algorithm for imbalanced classification[J]. Expert Systems With Applications, 2024,247:123274.
[30] LIU R J. A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification[J]. Applied Intelligence,2023,53(1):786-803.
[31] ZHENG Y F,WANG M N. Oversampling Method for imbalanced Data based on Variance Transfer[J]. Computer Science, 2024,51(S1):657-662.

HU Libin, born in 1990, Ph.D, is a member of CCF(No. V6549G). His main research interests include data mining, artificial intelligence and financial intelligence risk controll.

ZHANG Yunfeng, born in 1977, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 19888M). His main research interests include graphics, artificial intelligence, data mining and visualization.

(责任编辑:何杨)