doi:10.3969/j.issn.1001-4616.2025.05.011

基于局部合力改进的 Borderline-SMOTE 过采样方法

吕　峰,宋　媚,赵　礼,祝　义,李赫男

(江苏师范大学计算机科学与技术学院,江苏省教育智能技术高校重点实验室,江苏 徐州 221116)

[**摘要**]　数据分类是保障大数据分析有效进行的关键环节,解决数据分类中的类别不平衡成为当前研究的热点.过采样技术凭借其简洁性、有效性等特点,成为处理类不平衡问题的主要途径之一. 现有的过采样技术在处理不平衡数据中类重叠时缺乏合理的采样策略,导致机器学习模型预测时出现过拟合. 因此,本文提出一种基于局部合力改进的 Borderline-SMOTE 过采样方法(IBSLG). 首先,根据少数类样本最近邻分布构建边界区域;其次,基于局部合力计算边界区域内样本的集中度,根据集中度将样本划分为低概率/高概率边界样本;然后,基于两类边界样本分布,计算缩放因子构建新边界区域;最后,基于类不平衡比,对新边界区域自适应生成新样本. 通过 IBSLG 与 6 种采样方法在 4 种分类器、8 个不平衡数据集上进行对比实验,结果表明,IBSLG 在大部分数据集上取得了最优的 F1、G-mean、AUC 和 Friedman 排名,并在大部分分类器上取得了最高的平均次优率,说明所提方法的有效性.

[**关键词**]　不平衡数据,过拟合,类重叠,过采样,Borderline-SMOTE,局部合力
[**中图分类号**]TP181　[**文献标志码**]A　[**文章编号**]1001-4616(2025)05-0093-11

An Improved Borderline-SMOTE Oversampling Method Based on Local Resultant Gravitation

Lv Feng, Song Mei, Zhao Li, Zhu Yi, Li Henan

(School of Computer Science and Technology, Jiangsu Normal University, Jiangsu Provincial Key Laboratory of Intelligent Education Technology, Xuzhou 221116, China)

Abstract: Data classification is a key process to ensure the effectiveness of big data analysis, and addressing class imbalance in data classification has become a major focus of current research. Oversampling techniques, due to their simplicity and effectiveness, have become one of the primary approaches for handling class imbalance. However, existing oversampling techniques lack rational sampling strategies when dealing with class overlap in imbalanced data, leading to overfitting in machine learning model predictions. Therefore, this study proposes an improved Borderline-SMOTE oversampling method based on local resultant gravitation (IBSLG). Firstly, the boundary region is constructed based on the nearest neighbour distribution of minority samples; secondly, samples in this region are classified into low-probability and high-probability boundary samples using a concentration measure derived from local resultant gravitation; then a new boundary region is constructed by calculating scaling factor based on these boundary samples distributions; finally, new samples are adaptively generated for this new region based on the class imbalance ratio. Comparative experiments between IBSLG and six other sampling methods on four classifiers and eight imbalanced datasets show that IBSLG achieves optimal F1, G-mean, AUC, and Friedman rankings on most datasets, as well as the highest average suboptimal ratio on most classifiers, demonstrating its effectiveness.

Key words: imbalanced data, overfitting, class overlap, oversampling, Borderline-SMOTE, local resultant gravitation

数据分类作为大数据分析的重要环节,渗透在社会的各个领域,如罕见疾病诊断[1-2]、网络病毒识别[3]、工业故障诊断[4-5]和金融信用欺诈[6]. 然而,这些领域内存在一个相同的问题:不同类的样本在数量上差距较大,定义为类别不平衡问题. 通常根据样本的数量将数量多的一类样本称作多数类样本;数量少的一类样本称作少数类样本. 由于多数类样本在整个数据集中的占比较大,分类器往往更倾向于将新样本预测为多数类样本,这种对多数类样本的“偏爱”会导致分类器在预测地位更重要的少数类样本时出

收稿日期:2025-02-25.
基金项目:国家自然科学基金项目(71503108、62077029、62401235)、江苏省教育科学规划课题项目(B-b/2024/01/47).
通讯作者:宋媚,博士,副教授,研究方向:机器学习、大数据分析. E-mail:msong@jsnu.edu.cn

错,从而影响分类器对不同样本的分类能力,这样的错误会严重影响社会的发展.

类重叠[7]与过拟合[8]是处理不平衡的数据时常见的问题. 类重叠指的是不同类别的样本在特征空间上发生重叠或交叉,通常发生在不同类样本分布的边界处,这会导致分类器无法对边界处的样本进行准确分类. 而过拟合是分类器对训练集样本过度学习的表现. 大多数的过采样会根据“平衡样本假设”[9]在训练集生成大量合成样本,使不同类的样本在数量上保持一致,然而这会导致分类器对合成样本过拟合. 特别在样本分布的边界处,基于上述假设的过采样将同时面临类重叠和过拟合问题,进而严重影响分类器的分类能力. 因此,处理不平衡的数据时,类重叠和过拟合问题,尤其是发生在边界处的类重叠和过拟合需要被充分重视.

近年来,类重叠与过拟合一直是使用过采样技术研究类不平衡问题的主要焦点. 过采样,作为最流行、有效并且简便的数据级方法[10],在数据进入训练分类器步骤之前根据数据分布的特点,复制少数类样本扩展数据集以平衡数据分布,使得分类器发挥更佳的分类效果. 而对于类重叠或过拟合,研究者们主要通过限制采样对象范围或限制合成样本数量的方式解决其中某个问题. 前一种方法针对采样对象,减少需要过采样的少数类样本的数量,后一种方法针对合成样本,减少采样对象需要生成的样本数量. 然而,这些方法缺乏合理的采样策略,导致过采样处理边界处的类重叠出现过拟合问题.

针对过采样技术处理不平衡数据中的类重叠时,由于缺乏合理的采样策略导致边界处出现过拟合,提出一种新的过采样方法. 本研究的主要贡献如下:

(1)针对过采样技术处理不平衡数据中的类重叠时出现的过拟合问题,提出了一种新的过采样方法 IBSLG.

(2)引入局部合力识别出用于重新构建边界区域的高概率边界样本与低概率边界样本,根据合理的采样策略和上述两类边界样本进行过采样.

(3)在 8 个公开数据集和 2 个合成数据集上的实验结果表明,IBSLG 方法优于经典和目前主流的过采样方法.

1　相关工作

Chawla 等[11]提出的合成少数类过采样技术(synthetic minority oversampling technique,SMOTE)是最流行且常见的过采样方法. SMOTE 通过随机线性插值的方式在所有少数类样本与其最近邻样本之间生成合成样本,导致其在边界处生成大量的合成少数类样本,造成分类器对边界处样本过拟合.

图 1 描绘了在不同类样本分布的边界处使用 SMOTE 生成的样本,其中,白色和黑色圆圈、浅色和深色三角形分别代表少数类样本、多数类样本、离类重叠处较远的少数类样本产生的合成样本和离重叠处较近或在重叠处的少数类样本产生的合成样本. 不难发现,仅靠深色三角形代表的合成样本已经清晰地划分决策边界,而浅色三角形代表的合成样本依然聚集在决策边界.

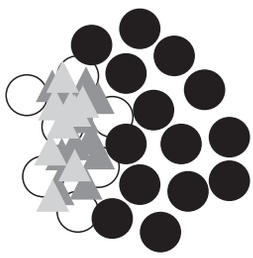


图 1　边界处样本生成图

Fig. 1　Illustration of sample-generation at the boundary

图 1 是在最近邻数 k 设置为 3 且每两个少数类样本间只生成 1 个合成样本下绘制的,而在实际情况下,浅色三角形代表的合成样本会大量聚集在已经被清晰划分的决策边界处,最终导致边界处产生大量的合成样本,造成分类器的过拟合. He 等[12]提出的自适应合成过采样(adaptive synthetic sampling,ADASYN)更关注密度高的少数类样本,并根据该类样本生成合成样本. Han 等[13]提出的 Borderline-SMOTE,也是本文需要改进的方法,该方法根据少数类样本最近邻分布构建安全区域 Safe、边界区域 Danger 和噪声区域 Noise,针对边界区域生成合成样本. ADASYN 与 Borderline-SMOTE 有效减少了少数类样本集中处的合成样本,但两种方法在边界处缺乏合理的采样策略,生成的过多的合成样本造成了无法避免的过拟合. 近些年,研究者们提出了自适应鲁棒合成少数类过采样方法(robust SMOTE,RSMOTE)[14],该方法引入相对密度对非噪声少数类样本生成合成样本,对去除噪声样本有明显作用,但是其在处理样本边界处的类重叠时缺乏合理的采样策略,引入了大量的合成样本,导致了严重的过拟合,这样糟糕的情况也发生在少数类样本集中处. 基于边界改进的自适应过采样方法

(borderline adaptive SMOTE,BA-SMOTE)[15],该方法根据少数类样本最近邻中多数类样本分布,采用基于 SMOTE 的两种不同的插值方法生成合成样本. 马贺等[16]提出改进的 Borderline-SMOTE(improved borderline-SMOTE,IBSM),其引入的合成因子为边界区域内的每个样本确定合成样本数量,有效减少了样本集中处过多的合成样本,然而也无法避免其在边界处造成的过拟合. Guo 等[17]提出的 Adaptive SV-Borderline-SMOTE-SVM 在核空间上利用 SVM 识别边界少数类支持向量(SV^+),接着,根据核距离计算 SV^+ 的最近邻居的分布以确定 SV^+的"凹凸性". 根据样本的"凹凸性"对少数类样本进行较为合理的采样. 同样的,Zhang 等[18]提出的 KDENDS-SMOTE 将样本映射到高维空间,而不是 Guo 等人映射的核空间,再使用核密度估计来导出密度比用作数据重叠程度的度量,最后与利用邻域信息计算密度比的稳定度结合构建评分机制来确定需要合成的目标样本. Lu 等[19]提出的 OMOS 与文献[17-18]均用于处理类重叠问题,OMOS 结合 mean shift 聚类算法识别出聚类前后标签一致的"安全样本",使用高斯分布模型拟合安全样本的分布特征,最后结合采样率对少数类簇进行过采样.

在上述方法中,经典的过采样方法注重于解决类不平衡问题,忽略了过采样在少数类样本集中处和边界处生成的大量合成样本导致的过拟合以及边界处的类重叠. 而近年来研究者们通过限制采样对象范围或限制合成样本数量的方式逐渐缓解了过采样在少数类样本集中处造成的过拟合以及边界处的类重叠. 然而,它们有一个共性的问题,过采样在处理边界处类重叠时会引入大量的合成样本造成过拟合,这是因为缺乏合理的采样策略,对采样区域进行全部过采样,忽视了采样区域内样本对分类的有效程度.

2　基于局部合力改进的 Borderline-SMOTE 过采样方法

2.1　局部合力

受牛顿万有引力理论启发,Wang 等提出了局部合力以研究样本与其近邻样本之间的关系,具体推导过程可见文献[20],局部合力公式如下:

$$\boldsymbol{F}_i = \sum_{j=1}^{k} \boldsymbol{D}_{ij} \sum_{j=1}^{k} \boldsymbol{u}_{ij}, \tag{1}$$

式中,$\boldsymbol{D}_{ij}$是样本 x_i 与其最近邻样本 x_j 之间的欧式距离,单位向量 $\boldsymbol{u}_{ij}$封装了 x_i 与 x_j 之间的方向信息,$\boldsymbol{F}_i$ 指的是 x_i 与其所有最近邻样本之间的局部合力.

2.2　IBSLG 算法框架图

基于局部合力改进的 Borderline-SMOTE 过采样方法的算法框架图见图 2. 在该框架中,首先,利用 Borderline-SMOTE 识别出不平衡数据集中的边界样本;其次,基于局部合力计算集中度,并根据集中度将边界

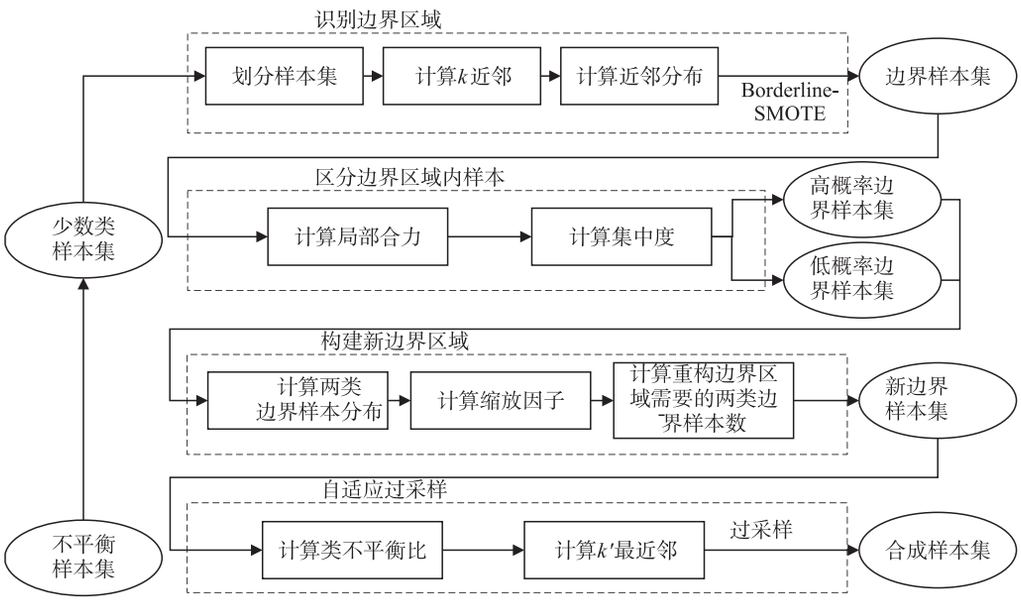


图 2　IBSLG 算法框架图

Fig. 2　Algorithm framework diagram

样本集内的样本分别划入高/低概率边界样本集. 然后,根据两类边界样本在边界区域的分布计算缩放因子,并根据缩放因子分别从两类边界样本集中随机抽取对应数量的样本构建新边界样本集. 最后,根据类不平衡比对新边界样本集内样本自适应过采样得到合成样本集. 该框架图将算法分为后面的 4 个模块.

2.3　识别边界区域模块

为了解决 SMOTE 在少数类样本集中区域造成的过拟合,本文利用 Han 等[13]提出的 Borderline-SMOTE 过采样方法初步构建边界区域 Danger,算法流程如下:

算法 1　识别边界区域算法

输入:训练集 T;最近邻个数 M;
输出:边界区域 Danger.
1. 划分训练集 T 为多数类样本集 N,少数类样本集 P,记录它们的样本数量分别为 n_{maj}、n_{min};
2. 计算 P 中的样本在 T 中的 M 个最近邻样本,得到最近邻样本中多数类样本的个数 M';
3. 将所有 $M/2 \leqslant M' < M$ 的少数类样本加入 Danger;
4. 返回 Danger.

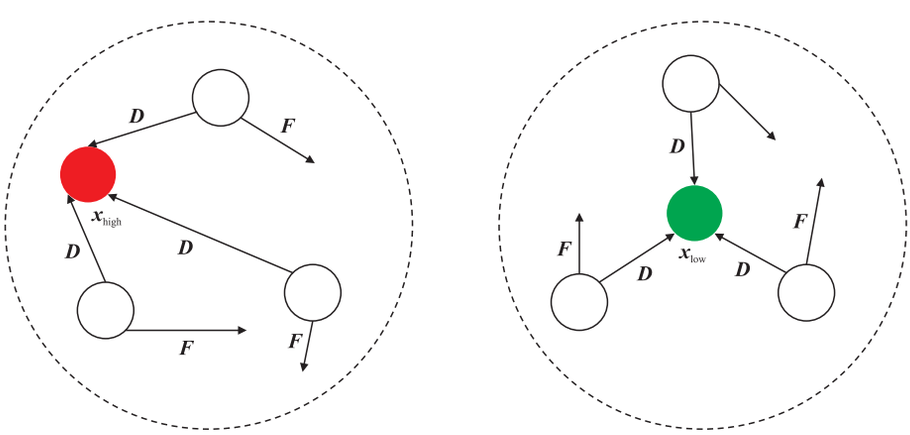
2.4　区分边界区域内样本模块

为了区分出对分类更有效果的样本,并利用该样本生成更少的、更好的合成样本解决处理 Danger 内类重叠时出现的过拟合问题,本文结合文献[20-21],根据局部合力所包含的信息计算 Danger 内样本的集中度 C_{DAS}(concentration of danger area sample),并利用 C_{DAS}对该区域内样本进行区分. 根据式(1),Danger 内某个样本 x_i 的集中度 C_{DAS_i}的计算公式如下:

$$C_{DAS_i}=\frac{1}{k}\sum_{j=1}^{k}\cos(\boldsymbol{F}_j,-\boldsymbol{D}_{ij}). \tag{2}$$

式中,$\boldsymbol{F}_j$ 是 x_i 与其所有最近邻样本间的局部合力,$-\boldsymbol{D}_{ij}$是 x_j 与 x_i 之间的距离向量,C_{DAS_i}是所有 $\boldsymbol{F}_j$ 与$-\boldsymbol{D}_{ij}$之间的余弦值均值,即 x_i 的集中度. 图 3(a)、(b)分别展示了 $C_{DAS}<0$ 与 $C_{DAS}>0$ 的样本在 Danger 内最近邻数为 3 时的分布. 两图通过 $\boldsymbol{D}$,$\boldsymbol{F}$ 分别表示距离向量与局部合力,虚线区域表示 Danger. 图 3(a)展示的样本 x_{high}的 $C_{DAS}<0$,x_{high}受到其所有最近邻样本的局部合力较小,x_{high}呈“远离”其最近邻样本的趋势,更有可能为边界样本,对提升分类作用较大,记为高概率边界样本;图 3(b)展示的样本 x_{low}的 $C_{DAS}>0$,x_{low}受到其所有最近邻样本的局部合力较大,x_{low}呈“接近”其最近邻样本的趋势,成为边界样本的概率较小,对提升分类作用较小,记为低概率边界样本. 这两类边界样本的定义见式(3). 其中,x^* 指的是 Danger 内的样本,x_{high}指的是高概率边界样本,x_{low}指的是低概率边界样本.

$$x_i^*=\begin{cases}x_{\text{high}} & C_{DAS_i}<0\\ x_{\text{low}} & C_{DAS_i}>0\end{cases}. \tag{3}$$



(a) 高概率边界样本　　(b) 低概率边界样本

样本的最近邻个数取 3

图 3　两类边界样本

Fig. 3　Two types of boundary samples

2.5　构建新边界区域模块

尽管已识别出 Danger 内的两类边界样本,但根据不合理的采样策略,比如选取 Danger 内所有样本或该区域内某一类样本构建的新边界区域,其生成的合成样本依然会导致 Danger 内出现类重叠和过拟合. 因此,本文针对 Danger 提出了一种合理的采样策略,以重新构建该区域用于采样,具体流程:首先,计算两类边界样本在 Danger 内的分布;接着,根据尽可能多地抽取高概率边界样本,尽可能少地抽取低概率边界样本的采样策略和得到的分布情况确定构建新边界区域 NewDanger 的样本数量 n_{new};最后,根据 n_{new} 分别从两类边界样本中随机抽取对应数量的样本构建 NewDanger 用于采样.

本文受到文献[22]提出合成因子研究各类样本分布的启发,提出缩放因子 scaling_factor 研究 Danger 内两类边界样本的分布情况,定义式如下:

$$\text{scaling\_factor}=\frac{n^*_{\text{positive}}-n^*_{\text{negative}}}{n^*}, \tag{4}$$

式中,n^*_{positive}、n^*_{negative}、n^* 分别指 Danger 内低概率边界样本数、高概率边界样本数和所有样本数. scaling_factor(简称为 sf)反映了高/低概率边界样本在 Danger 内的分布情况. 根据式(4),当 $sf\leqslant 0$ 时,高概率边界样本相较低概率边界样本在 Danger 内的占比较高;相反地,当 $sf>0$ 时,高概率边界样本的占比较小.

因此,根据提出的采样策略,当 $sf\leqslant 0$ 时,Danger 内的高概率边界样本相对数量足够多,可以抽取高概率边界样本集内所有(n^*_{negative} 个)样本构建 NewDanger;而 $sf>0$ 时,Danger 内的高概率边界样本相对数量较少,除了需要抽取高概率边界样本集内所有(n^*_{negative} 个)样本,还需要从低概率边界样本集内随机抽取 n_{extra} 个样本,共同构建 NewDanger. n_{extra} 与类不平衡比 IR 有关. 若 IR 比较大,则添加较少的低概率边界样本,反之则添加较多的低概率边界样本,由于 n_{extra} 与 IR 成反比,n_{extra} 的值总是较小的,即低概率边界点抽取数量总是较小的,小于 n^*_{positive}. n_{new} 和 n_{extra} 的定义式如下,其中 n_{maj}、n_{min} 和 n_{new} 分别指多数类、少数类和 NewDanger 内样本数.

$$n_{\text{new}}=\begin{cases} n^*_{\text{negative}} & sf\leqslant 0 \\ n^*_{\text{negative}}+n_{\text{extra}} & sf>0 \end{cases}, \tag{5}$$

$$n_{\text{extra}}=\frac{n^*_{\text{positive}}}{IR}, \tag{6}$$

$$IR=\frac{n_{\text{maj}}}{n_{\text{min}}}. \tag{7}$$

2.6　自适应过采样模块

为使 IBSLG 能针对不同数据集生成适量的合成样本,本文根据 Borderline-SMOTE[13,21]所用最近邻数 $k(k=5)$ 结合类不平衡比 IR 分段确定过采样需要的最近邻数 k',制定方案如下:

$$k'=\begin{cases} 1 & IR-1\leqslant 1, \\ \text{int}(IR-1) & IR-1\leqslant 5, \\ \text{int}(n_{\text{min}}/2) & IR-1>5. \end{cases} \tag{8}$$

$$x_{\text{new}}=x_i+(x_i-x_j)\times\text{rand}. \tag{9}$$

式中,x_{new}、rand、x_i 和 x_j 分别指合成样本、0~1 之间的随机数、新边界区域内的某个样本和该样本的最近邻样本. 要使两类样本数量平衡,理论上每个少数类样本都需要生成($n_{\text{maj}}-n_{\text{min}}$)/$n_{\text{min}}$ 个样本,即 $IR-1$ 个样本. 然而,基于“平衡样本假设”[9],使得不同类样本在数量上达到绝对平衡并不能达到理想的分类效果,因此,最近邻样本的数量应该被重新考虑. 本文以 Borderline-SMOTE[13]中设置的最近邻数 $k=5$ 为划分点,根据 IR 自适应分配不同的最近邻 k',最终根据式(9)利用新边界区域内的样本与 k'生成合成样本.

2.7　IBSLG 算法

本文提出的基于局部合力改进的 Borderline-SMOTE(简称为 IBSLG),首先根据算法 1 初步确定边界区域(Danger);接着,根据节(2.4)、(2.5)重新构建 Danger 区域,新边界区域记为 NewDanger;最后,对 NewDanger 区域内的样本,根据节(2.6)自适应生成合成样本. IBSLG 的算法如下:

算法 2　基于局部合力改进的 Borderline-SMOTE 过采样方法

输入:训练集 T;最近邻个数 k、M;
输出:合成样本集合 S_{new}.
1. 将 T 划分为多数类样本集合 N,少数类样本集合 P,记录它们的样本数量分别为 n_{maj}、n_{min};
2. 根据算法 1 生成边界区域 Danger,记区域内样本数量为 n^*;
3. for 每个 Danger 内的样本 x_i;
4. 根据式(2)计算 Danger 内样本的集中度 C_{DAS};
5. end for
6. 根据式(3)将 Danger 内样本分别划入高概率边界样本集 S_{high}和低概率边界样本集 S_{low},并分别记录两类样本的数量为 $n^*_{negative}$ 和 $n^*_{positive}$;
7. 根据式(4)计算缩放因子 sf;
8. 根据式(5)、(6)、(7)计算构建新边界区域 NewDanger 所需要的样本数 n_{new};
9. if $sf \leqslant 0$
10. 从 S_{high}中随机抽取 n_{new}个样本,并将抽取出的样本全部加入 NewDanger;
11. end if
12. if $sf>0$
13. 分别从 S_{high}和 S_{low}中随机抽取 $n^*_{negative}$和 $n_{new}-n^*_{negative}$个样本,并将抽取出的样本全部加入 NewDanger;
14. end if
15. for 每个 NewDanger 内的样本 x_i;
16. 根据式(8)计算 x_i 要生成的合成样本数量;
17. 根据式(9)生成合成样本 x_{new};
18. 将 x_{new}加入 S_{new};
19. end for
20. 返回 S_{new}.

3　实验结果与分析

为了证明 IBSLG 的有效性,实验采用了 6 种采样方法,分别为 SMOTE[11]、ADASYN[12]、Borderline-SMOTE[13]、RSMOTE(2021)[14]、BA-SMOTE(2022)[15]和 IBSM(2023)[16],分别简称为 SMO、BOR、ADA、RSM、BAS 和 IBS. 实验采用的分类器为决策树分类器(Decision Tree,DT)、随机森林分类器(RandomForest)、AdaBoost 分类器和 XGBoost 分类器,分别简称为 DT、Ran、Ada 和 XGB.

3.1　实验数据集

为验证 IBSLG 的有效性,本文不仅使用了来自 KEEL 和 UCI 数据库的 8 个公开数据集,其不平衡比范围为 1.5~66.67,样本数量范围为 178~17898,还选择了文献[14,23]中使用的合成数据集 circles 和 moons. 两个数据集中的多数类样本和少数类样本数量均为 750 和 250. 公开数据集信息见表 1.

表 1　公开数据集信息

Table 1　Information of public datasets

数据集名	样本数	不平衡比	特征数	简称	数据集名	样本数	不平衡比	特征数	简称
HTRU2	17 898	9.92	8	D1	vehicle2	846	2.88	18	D5
glass0	214	2.06	9	D2	wine	178	1.5	13	D6
page_blocks_1_3_vs_4	472	15.86	10	D3	yeast_2_vs_4	514	9.08	8	D7
segment0	2 308	6.02	19	D4	shuttle_2_vs_5	3 316	66.67	9	D8

3.2　评价指标

本文使用基于混淆矩阵的 F1、G-mean、AUC 作为评价指标[16,24]. F1 是精确率和召回率的调和平均值,F1 值越高表明模型越能准确地识别出少数类样本. G-mean 是召回率和特异度的几何平均值,可评估模型的平衡性. AUC 是 ROC 曲线下的面积,AUC 值越接近 1,表明模型的整体性能越好.

3.3　平均次优率

大多数实验以最优值作为评价标准,忽略了次优值对验证方法有效性和稳健性的作用. 平均次优率考虑某方法在非最优数据集中取得次优值的概率,该值一定程度上可以反映某方法的稳健性和有效性. 平均次优率越高,代表某个采样方法在非最优的情况下,越是一种有效稳健的采样方法. 平均次优率计算公式如下:

$$\text{subratio}=\frac{n_{\text{sub}}}{n_{\text{total}}-n_{\text{opt}}}, \tag{10}$$

$$\text{avesubratio}=\frac{1}{n_{\text{metric}}}\times\sum_{i=1}^{n_{\text{metric}}}\text{subratio}_i, \tag{11}$$

式中,式(10)、(11)可以分别计算方法在某个分类器、所有数据集、某个评价指标上的次优率和不同评价指标的平均次优率. subratio、n_{sub}、n_{total}、n_{opt}、avesubratio 和 n_{metric}分别指某个评价指标对应的次优率、取得次优值的数据集个数、数据集总个数、取得最优值的数据集个数、不同评价指标取得的平均次优率和评价指标的个数.

3.4　参数设置

本文的算法参数如表 2 所示,为了与 Borderline-SMOTE[13] 和文献[20]的实验环境保持一致,本文的 k、M 均取文献[13,20]中的默认值. 其中,k 是某样本的最近邻个数,M 是初步构建边界区域时某样本的最近邻个数,n_{new}在本文中取默认值 1,即新边界区域内至少存在一个样本.

表 2　算法参数

Table 2　Algorithm parameters

参数	值
k	5
M	10
n_{new}	1

3.5　结果与分析

3.5.1　二维合成数据集结果与分析

为了更直观地展示各采样方法生成样本后的样本分布情况,本文使用两个合成数据集 moons 和 circles,两者的样本数量已在节(3.1)中设置. 图 4、图 5 分别展示了使用不同采样方法后数据集 moons 和 circles 上的样本分布情况. 图 4(a)、图 5(a)是数据集的原始分布. 由图 4(b)、图 5(b)可知,SMOTE 对所有少数类样本生成合成样本,这不仅会在少数类样本集中处和样本分布的边界处生成大量的合成样本,造成分类器对这两处的样本过拟合,还生成了许多噪声影响分类结果. 由图 4(e)、图 5(e)可知,RSMOTE 适当减少了生成合成样本的数量,但少数类样本集中处依然存在大量合成样本,分类器对此处样本依然过拟合. 由图 4(c)、图 5(c)

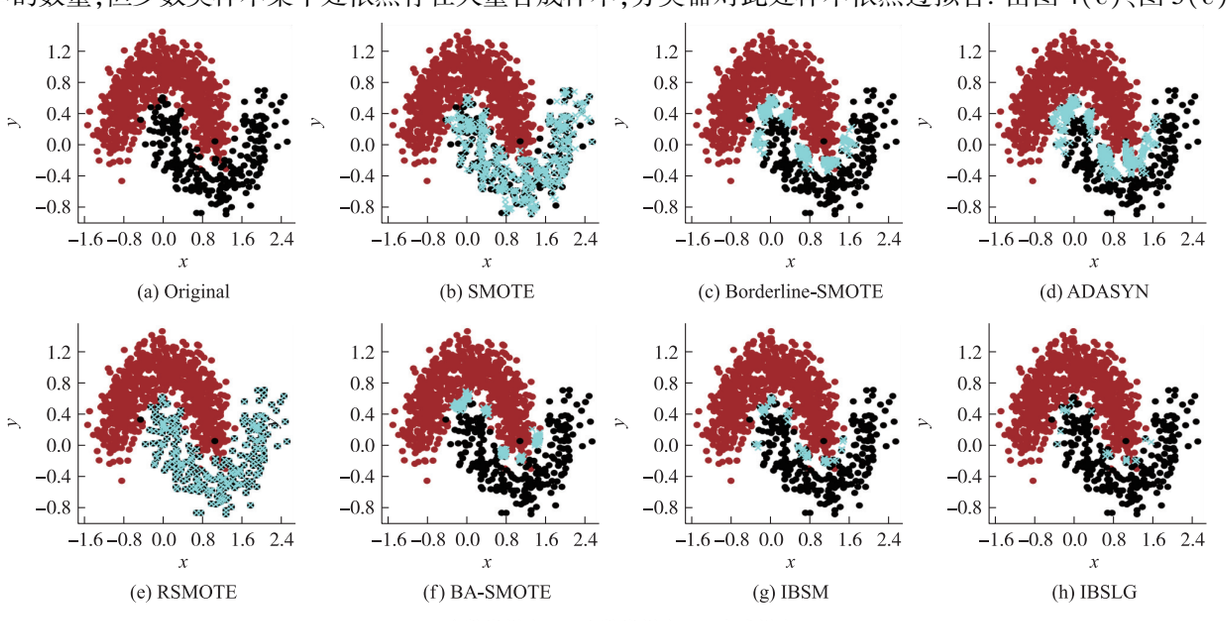


图 4　不同采样方法在 moons 数据集上采样后的数据分布图

Fig. 4　Distribution of data after sampling on moons dataset with different sampling methods

和图 4(d)、图 5(d)可知,Borderline-SMOTE 和 ADASYN 通过在识别出的边界区域生成合成样本减少了分类器对少数类样本集中处的过拟合,然而,在样本的边界处依然生成了大量的合成样本,这不仅没有缓解该区域内的类重叠,还导致分类器对该区域内样本过拟合. 由图 4(f)、图 5(f)和图 4(g)、图 5(g)可知,BA-SMOTE 和 IBSM 缓解了边界处的部分类重叠和过拟合,但在多数类与少数类样本交叉密集的区域依然存在大量合成样本,它们堆叠在一起,形成了多个稠密高亮的合成样本堆,分类器依然对边界处的这些合成样本过拟合. 由图 4(h)、图 5(h)可知,本文提出的方法不仅可以有效减少生成样本的数量,减轻分类器对边界处的过拟合,还可以缓解边界处的类重叠,加强分类器对边界处样本的识别能力.

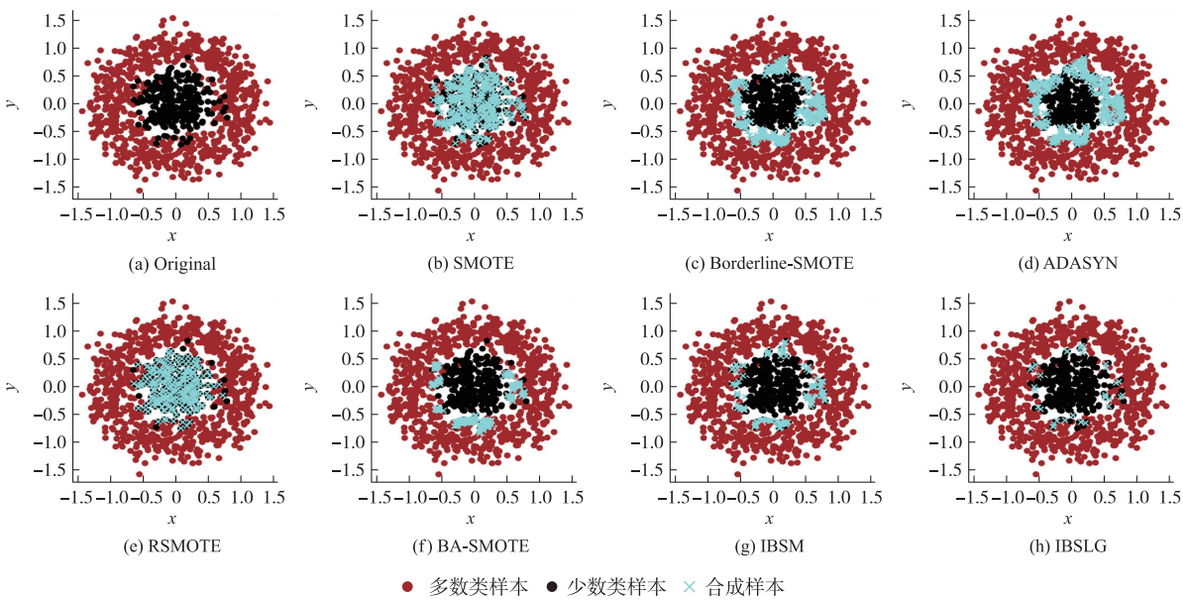


图 5　不同采样方法在 circles 数据集上采样后的数据分布图

Fig. 5　Distribution of data after sampling on circles dataset with different sampling methods

3.5.2　公开数据集结果与分析

本文实验按照 7∶3 的比例划分数据集为训练集和测试集,此外,为了避免实验的随机性,以下实验都是基于 5 次五折交叉验证的平均结果作为相应指标的最终得分. 表 3 记录了 IBSLG 和其他 6 种采样方法在 8 个数据集、4 个分类器上取得最高 G-mean、AUC 的数据集数量. 表 4 记录了各算法在 4 个分类器上 G-mean 和 AUC 值的平均次优率. 表 5 记录了各算法取得的 F1 值,表 5 中对最优值加粗标注,对 IBSLG 取得的次优值加下划线标注.

表 3　各算法取得最高评价指标的数据集数量

Table 3　Number of datasets for which all algorithms take the optimal value

评价指标	分类器	SMO	BOR	ADA	RSM	BAS	IBS	IBSLG
G-mean	DT	2	2	3	1	1	1	**4**
	Ran	3	2	1	1	1	2	**4**
	Ada	1	3	1	1	1	2	**5**
	XGB	**4**	3	2	0	2	1	**4**
AUC	DT	2	3	2	1	1	1	**4**
	Ran	3	2	1	0	1	2	**5**
	Ada	1	2	1	2	1	2	**5**
	XGB	**4**	3	2	0	2	1	**4**

表 4　各算法的 G-mean、AUC 的平均次优率

Table 4　The average suboptimal ratios of G-mean, AUC for each algorithm

分类器	SMO	BOR	ADA	RSM	BAS	IBS	IBSLG
DT	50%	8%	18%	7%	14%	0%	**50%**
Ran	0%	0%	14%	33%	29%	17%	**42%**
Ada	14%	27%	0%	23%	0%	17%	**100%**
XGB	0%	20%	17%	25%	50%	0%	25%

表 5　各方法 F1 值对比

Table 5　Comparison of F1 values for each method

数据集	分类器	SMO	BOR	ADA	RSM	BAS	IBS	IBSLG
D1	DT	0.754 5	0.800 6	0.717 9	**0.826 0**	0.778 5	0.814 8	0.749 5
	Ran	0.846 1	0.767 5	0.679 7	**0.872 3**	0.754 3	0.871 8	0.775 3
	Ada	0.813 3	0.762 8	0.623 1	**0.875 6**	0.746 4	0.867 8	0.763 6
	XGB	0.839 0	0.851 2	0.743 6	**0.879 0**	0.833 5	0.877 1	0.832 5
D2	DT	0.604 1	0.635 1	0.641 7	0.651 6	**0.662 8**	0.603 0	0.614 1
	Ran	0.667 8	0.650 4	0.650 9	0.690 0	0.654 6	0.674 4	**0.696 8**
	Ada	0.640 3	0.656 3	0.637 8	0.662 5	0.627 1	0.626 4	**0.665 9**
	XGB	0.679 2	0.669 1	0.648 5	0.684 2	0.690 8	0.683 7	**0.707 4**
D3	DT	**0.950 1**	**0.950 1**	**0.950 1**	0.921 8	0.817 6	0.897 3	0.941 8
	Ran	0.933 1	0.990 5	0.964 6	0.961 8	0.909 4	**1.000 0**	**1.000 0**
	Ada	0.939 0	0.939 0	0.939 0	**0.977 1**	0.873 8	0.939 0	0.949 4
	XGB	**0.950 1**	**0.950 1**	**0.950 1**	0.936 8	0.913 1	0.940 6	0.950 1
D4	DT	0.971 1	0.974 8	0.965 3	0.966 5	0.953 6	0.962 4	**0.975 1**
	Ran	0.971 2	0.980 0	0.975 0	0.958 1	0.960 2	0.982 0	**0.984 9**
	Ada	0.986 1	0.985 1	0.986 1	0.985 9	0.978 2	0.984 0	**0.987 1**
	XGB	0.978 1	0.978 0	0.978 0	0.976 1	0.977 3	0.979 9	0.981 0
D5	DT	0.922 8	0.916 0	0.916 5	0.923 5	0.903 2	0.913 0	**0.923 6**
	Ran	0.942 3	0.941 4	0.940 8	0.952 5	0.936 4	0.938 9	**0.965 6**
	Ada	0.951 6	0.960 9	0.956 5	0.957 8	0.948 4	0.964 4	**0.971 1**
	XGB	0.975 0	0.969 2	0.969 5	**0.975 2**	0.969 3	0.969 6	0.968 9
D6	DT	0.892 3	**0.905 8**	0.899 7	0.879 3	0.883 8	0.884 3	0.883 8
	Ran	0.969 3	0.974 7	0.974 7	0.969 3	0.975 0	0.963 6	**0.980 4**
	Ada	0.928 7	0.948 9	0.944 3	0.948 9	0.932 8	0.955 0	**0.960 0**
	XGB	**0.957 5**	0.936 2	0.931 4	0.938 2	0.928 3	0.930 9	0.936 7
D7	DT	0.720 2	0.710 2	0.719 7	0.701 8	0.660 0	0.751 6	**0.769 1**
	Ran	0.779 1	0.753 6	0.746 5	0.811 7	0.739 0	0.770 2	**0.822 0**
	Ada	0.765 6	0.770 3	0.751 9	0.775 2	0.750 0	0.782 0	**0.796 3**
	XGB	0.752 9	**0.786 1**	0.764 5	0.775 2	0.743 9	0.752 5	0.769 0
D8	DT	**1.000 0**	**1.000 0**	**1.000 0**	0.967 1	0.989 5	**1.000 0**	**1.000 0**
	Ran	**1.000 0**	**1.000 0**	**1.000 0**	0.983 8	**1.000 0**	**1.000 0**	**1.000 0**
	Ada	**1.000 0**	**1.000 0**	**1.000 0**	0.983 8	**1.000 0**	**1.000 0**	**1.000 0**
	XGB	**0.980 6**	**0.980 6**	**0.980 6**	0.945 5	**0.980 6**	**0.980 6**	**0.980 6**

由表 3、5 可知,在 8 个数据集上,本文提出的 IBSLG 与其他 6 种采样方法相比,使用随机森林、AdaBoost、XGBoost 和决策树作为分类器时,在 F1 上分别取得了 7、6、4 和 4 个最优值;在 G-mean 上分别取得了 4、5、4 和 4 个最优值;在 AUC 上分别取得了 5、5、4 和 4 个最优值. 这表明本文方法在大部分数据集上取得了最优的结果,然而,在一些数据集上,以 D1 数据集为例,IBSLG 的表现略逊于其他采样方法,这是由于数据集本身的复杂分布[14,16,25]和数据量大导致采样方法生成的较少的合成样本无法有效发挥其作用,最终使其在 F1 上未达到最优. 此外,由表 4(经式(10)(11)计算得)可知,在上述 4 个分类器中,本文方法在 3 个分类器上均取得了最高的平均次优率. 因此,实验证明本文方法在非最优的情况下,依然是一种有效稳健的采样方法. 为了进一步比较本文方法与其他方法的性能差异,本文以 IBSLG 为主控方法,与其余采样方法做 Friedman 检验[26],本文采用和文献[16,26]相同的方法,即计算各个方法在评价指标上的平均秩(秩越小算法性能越高),最终得到在不同评价指标,所有分类器上的平均秩的平均值,实验结果见图 6.

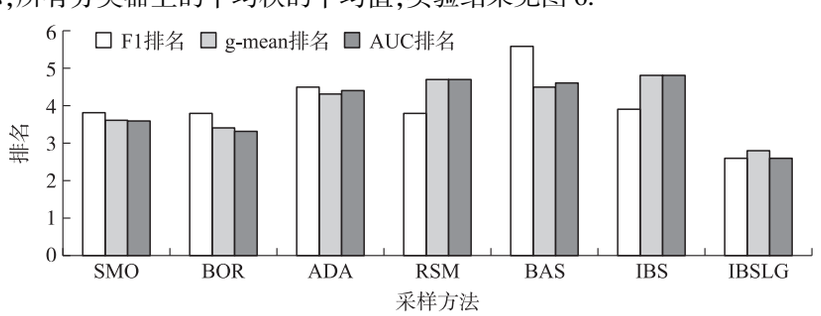


图 6　各算法的 Friedman 排名对比

Fig. 6　Comparison of Friedman rankings by all algorithms

由图 6 可知,在涵盖 4 种分类器与 8 个公共数据集的复杂实验环境下,本文提出的 IBSLG 算法在平均 F1、G-mean 与 AUC 的 Friedman 排名均稳定优于经典的(如 SMOTE)、最新的(如 IBSM)过采样算法,表明 IBSLG 算法有效且具有较强的鲁棒性.

4 结论

本文提出了一种基于局部合力改进的 Borderline-SMOTE 过采样方法,以解决在处理不平衡数据中的类重叠问题时,传统过采样技术的不合理性和容易导致的过拟合问题. 该方法基于局部合力计算缩放因子,并利用该因子和合理的采样策略构造新的边界区域,使其包含更多高概率的边界样本,相对更少的低概率的边界样本. 在该区域生成的样本不仅缓解了该区域内的类重叠问题,降低了过拟合的风险,还增强了分类器对边界处样本的识别能力. 通过在 8 个数据集和 2 个人工合成数据集上使用多种分类器与经典及最新的采样方法进行对比实验,结果表明,该方法显著提升了分类器对少数类样本的识别能力. IBSLG 方法专为二分类问题设计,如何将其有效应用于多分类仍是未来研究的方向. 此外,利用次优率检验采样方法对同类研究有一定的启发作用.

[参考文献]

[1] EL-ASSY A M,AMER H M,IBRAHIM H M,et al. A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI data[J]. Scientific reports,2024,14(1):3463-3481.

[2] HATAMI M,YAGHMAEE F,EBRAHIMPOUR R. Improving Alzheimer's disease classification using novel rewards in deep reinforcement learning[J]. Biomedical signal processing and control,2025,100(3):106920-106936.

[3] ULLAH F,ULLAH S,SRIVASTAVA G,et al. IDS-INT:intrusion detection system using transformer-based transfer learning for imbalanced network traffic[J]. Digital communications and networks,2024,10(1):190-204.

[4] LUO J,ZHANG Y,YANG F,et al. Imbalanced data fault diagnosis of rolling bearings using enhanced relative generative adversarial network[J]. Journal of mechanical science and technology,2024,38(2):541-555.

[5] FAN X,DUAN L,ZHANG N. A multi-scale graph-guided dynamic enhanced alignment network for mechanical fault diagnosis considering domain shift and data imbalance[J]. Neurocomputing,2025,625(1):129546-129559.

[6] XIE Y,LI A,HU B,et al. A credit card fraud detection model based on multi-feature fusion and generative adversarial network [J]. Computers,materials & continua,2023,76(3):2707-2726.

[7] GHOSH K,BELLINGER C,CORIZZO R,et al. The class imbalance problem in deep learning[J]. Machine learning,2024,113(7):4845-4901.

[8] CHEN L,JING X Y,CHEN R,et al. Sample-pair learning network for extremely imbalanced classification[J]. Neurocomputing,2025,634(1):129859-129871.

[9] MA H,ZHANG X,SONG M,et al. SD-CSMOTE:Over-sampling method based on SNN-DPC and improved SMOTE[J]. Neurocomputing,2025,620(1):129233-129243.

[10] AGUITAR G,KRAWCZYK B,CANO A. A survey on learning from imbalanced data streams:taxonomy,challenges,empirical study,and reproducible experimental framework[J]. Machine learning,2024,113(7):4165-4243.

[11] CHAWLA N V,BOWYER K W,HALL L O,et al. SMOTE:synthetic minority oversampling technique[J]. Journal of artificial intelligence research,2002,16(1):321-357.

[12] HE H,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). Hong Kong:IEEE,2008:1322-1328.

[13] HAN H,WANG W Y,MAO B H. Borderline-SMOTE:a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing. Berlin,Heidelberg:Springer,2005:878-887.

[14] CHEN B,XIA S,CHEN Z,et al. RSMOTE:a self-adaptive robust SMOTE for imbalanced problems with label noise[J]. Information sciences,2021,553(1):397-428.

[15] 陈海龙,杨畅,杜梅,等. 基于边界自适应 SMOTE 和 Focal Loss 函数改进 LightGBM 的信用风险预测模型[J]. 计算机应用,2022,42(7):2256-2264.

[16] 马贺,宋媚,祝义. 改进边界分类的 Borderline-SMOTE 过采样方法[J]. 南京大学学报(自然科学版),2023,59(6):

1003-1012.

[17] GUO J, WU H, CHEN X, et al. Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification[J]. Applied soft computing, 2024, 150(1): 110986-110998.

[18] ZHANG R, LU S, YAN B, et al. A density-based oversampling approach for class imbalance and data overlap[J]. Computers & industrial engineering, 2023, 186(1): 109747-109760.

[19] LU X, YE X, CHENG Y. An overlapping minimization-based over-sampling algorithm for binary imbalanced classification[J]. Engineering applications of artificial intelligence, 2024, 133(1): 108107-108120.

[20] WANG Z, YU Z, CHEN C L P, et al. Clustering by local gravitation[J]. IEEE transactions on cybernetics, 2018, 48(5): 1383-1396.

[21] 冀常鹏,尚佳奇,代巍. 不平衡数据集的 DC-SMOTE 过采样方法[J]. 智能系统学报,2024,19(3):525-533.

[22] 陶佳晴,贺作伟,冷强奎. 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.

[23] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information sciences, 2018, 465(1): 1-20.

[24] FEI Y, CHEN F, HE L, et al. Intelligent classification of antenatal cardiotocography signals via multimodal bidirectional gated recurrent units[J]. Biomedical signal processing and control, 2022, 78: 104008-104015.

[25] CHEN Q, YE A, ZHANG Y, et al. An intra-class distribution-focused generative adversarial network approach for imbalanced tabular data learning[J]. International journal of machine learning and cybernetics, 2024, 15(1): 2551-2572.

[26] YANG K, YU Z, CHENG C L P, et al. Incremental weighted ensemble broad learning system for imbalanced data[J]. IEEE transactions on knowledge and data engineering, 2022, 34(12): 5809-5822.

[责任编辑:黄　敏]