计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(11):73−82 [doi: 10.15888/j.cnki.csa.009297]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

Borderline-mixup 不平衡数据集分类方法①

吴振煊[1], 郭躬德[1], 王　晖[2]

[1](福建师范大学 计算机与网络空间安全学院, 福州 350117)
[2](贝尔法斯特女王大学 电子电气工程和计算机科学学院, 贝尔法斯特 BT9 5BN)
通信作者: 郭躬德, E-mail: ggd@fjnu.edu.cn; 王　晖, E-mail: h.wang@qub.ac.uk

摘　要: 不平衡数据集问题从 20 年前就已经引起人们的重视, 提出的相关解决方法层出不穷. Mixup 是这几年比较流行的数据合成方法, 其相关变体比比皆是, 但是针对不平衡数据集提出的 Mixup 变体寥寥无几. 本文针对不平衡数据集分类问题, 提出了 Mixup 的变体——Borderline-mixup, 其使用支持向量机选择边界样本, 增加边界样本在采样器中被采样的概率, 构建两个边界采样器, 替代了原有的随机采样器. 在 14 个 UCI 数据集以及 CIFAR10 长尾数据集上的实验结果表明, Borderline-mixup 相比于 Mixup 在 UCI 数据集中都有提升, 最高能达到 49.3% 的提升, 在 CIFAR10 长尾数据集中, 也能达到 3%–3.6% 左右的提升. 显然, 我们提出的 Mixup 变体在不平衡数据集分类中是有效的.

关键词: Mixup; 支持向量机; 不平衡数据集; 边界样本; 分类

引用格式: 吴振煊,郭躬德,王晖.Borderline-mixup 不平衡数据集分类方法.计算机系统应用,2023,32(11):73–82. http://www.c-s-a.org.cn/1003-3254/9297.html

Borderline-mixup Imbalanced Data Sets Classification Method

WU Zhen-Xuan[1], GUO Gong-De[1], WANG Hui[2]

[1](College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)
[2](School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5BN, UK)

Abstract: The problem of imbalanced datasets has attracted people's attention since two decades ago, and various solutions have been proposed. Mixup is a popular data synthesis method in recent years, with many variants extended. However, there are not many Mixup variants proposed for imbalanced datasets. This study proposes a Mixup variant, namely Borderline-mixup, to address the classification problem of imbalanced datasets, which uses a support vector machine (SVM) to select boundary samples and increases the probability that the boundary sample is sampled in the sampler. Two boundary samplers are constructed to replace the original random sampler. Extensive experiments have been conducted on 14 UCI datasets and CIFAR10 long-tail datasets. The results show that Borderline-mixup has outperformed Mixup consistently on UCI datasets by up to 49.3% and on CIFAR10 long-tail datasets by about 3%–3.6%. Therefore, the proposed Borderline-mixup is effective in the classification of imbalanced datasets.

Key words: Mixup; support vector machine (SVM); imbalanced data sets; boundary samples; classification

近年来, 神经网络的发展十分迅速, 在不同领域的应用都取得了优异的表现. 众所周知, 数据对于神经网络是至关重要的. 然而和许多实验中所使用的数据集不同, 真实世界的数据通常是呈不平衡分布的, 尤其在一些异常检测的应用中, 比如医疗诊断、欺诈检测、入侵检测等, 这是因为异常事件相对于正常事件而言

① 基金项目: 国家自然科学基金 (61976053, 62171131); 福建省自然科学基金 (2022J01398)
收稿时间: 2023-04-30; 修改时间: 2023-05-29; 采用时间: 2023-06-06; csa 在线出版时间: 2023-09-15
CNKI 网络首发时间: 2023-09-19

通常是罕见的. 类别不平衡问题早在 20 年前就已经得到人们的广泛关注[1,2], 在这种情况下, 数据集的不平衡分布会给大多数假设数据是相对平衡分布的机器学习算法带来严重的困难[3]. 比如在反向传播的神经网络中, 多数的类别样本往往会通过主导梯度向量来主导神经网络的训练过程, 即将类与类之间的边界由多数类推向少数类, 以减少分类误差. 这会导致神经网络在少数的类别样本上表现不佳.

针对不平衡数据集分类, 已经提出了许多相关的解决方法. 这些方法可以简单地分为 3 大类, 第 1 类是重加权, 其中包括代价敏感学习和分类器阈值调整, 目的都是为了在算法层面上给予少数类更大的权重. 代价敏感学习考虑了不同误分类情况的不同代价[4], 设置错误分类少数样本的代价大于错误分类多数样本的代价. 通过在训练期间调整不同类别的损失值来达到对类别进行重新平衡的目的. 相关的方法有 Focal loss[5]、Class-balanced loss[6] 等. 分类器阈值调整是从修正分类结果的角度出发, 通过调整阈值, 使得模型更关注少数类. 第 2 类是集成学习, 集成学习利用多个分类器, 通过各种投票机制获得最终结果, 从而提高单个分类器的准确性[7], 已经成功应用在不平衡数据集中[8], 并成为了类不平衡问题的一种流行的解决方法[9]. 第 3 类是重采样, 又可以细分为对少数类进行过采样、对多数类进行欠采样或者是两种方法结合使用, 目的是为了从数据层面上使不平衡数据集变得较为平衡. 其中, 随机采样是最简单的一种采样方法, 但是, 对少数类进行随机过采样, 容易造成少数类样本的过拟合; 对多数类进行随机欠采样, 又会损失多数类样本的相关特征信息. 于是, 有人提出基于数据生成的采样, 即对数据进行合成来增加相应类别的样本数量, 从而提升神经网络在不平衡数据集上的性能. SMOTE[10] 就是一种合成少数样本的过采样技术, 通过随机选择少数样本附近的邻近点, 在两者之间的连线上随机选择一点作为新合成的少数类样本. Mixup[11] 也是一种基于数据生成的过采样技术, 随机选择数据集中的两个样本, 将样本和样本标签分别进行混合.

有研究表明[12], Mixup 在不平衡数据集上能够有效地提升网络的性能. 虽然 Mixup 从发表至今, 提出的变体层出不穷, 但其许多变体都是对平衡的数据集进行研究实验, 在不平衡数据集上的研究比较少, 其中较为熟知的有 Remix[13]、Balanced-mixup[14]、Label-occurrence-balanced mixup[15].

本文提出一种新的不平衡数据集分类方法: 边界混合 (Borderline-mixup), 它由两个边界采样器组合而成. 在边界采样器中, 我们不再盲目地选择样本进行混合, 而是找到位于边界附近的样本, 增加它们被采样的概率. 因为边界样本最容易被错误分类, 将混合的重点放在边界区域上可能会比放在整个少数类样本区域上有更好的表现. 我们使用 4 层的多层感知机对 UCI 数据集中的 10 个二分类以及 4 个多分类的不平衡数据集进行实验, 结果表明 Borderline-mixup 在提升模型性能方面是有效的. 除此之外, 我们还在基准的不平衡数据集 CIFAR10-LT 上进行了实验, 实验结果表明, 我们提出的 Borderline-mixup 相较于 Mixup 的性能最高能提升 3.6%.

1 相关工作

1.1 重采样

重采样一般分为过采样和欠采样, 最简单的一种采样方法就是随机采样. 对少数类进行随机过采样, 虽然扩大了数据集, 但是因为对少数类样本进行了多次复制, 容易造成过拟合. 而对多数类进行随机欠采样, 会丢弃一些样本, 即有可能损失部分有用信息.

针对随机过采样的问题, 有人提出, 过采样的时候不要只是简单地复制样本, 而是通过一些方法来生成新样本, 从而降低过拟合的风险, 比如通过 SMOTE[10] 方法, 对少数类进行合成新样本, 从而达到过采样的目的. 至于随机欠采样, 有人提出了依据信息的欠采样, 主要有两种方法: EasyEnsemble 和 BalanceCascade[16], 目的是克服随机欠采样中的信息丢失.

1.2 重加权

重加权的主要思想就是根据类别样本的数量调整不同类别的权重, 以重新定义每个类别中样本的重要性, 从而达到对类别进行重新平衡的目的. 这里的权重可以是误分类的代价, 也可以是分类器的阈值.

调整误分类的代价的方法又称为代价敏感学习, 许多研究都提出了各种重新加权的方法来处理数据集不平衡的问题, 包括 Focal loss[5]、Class-balanced loss[6] 等.

调整分类器阈值也是一种重加权的方法. 有研究表明[17,18], 在数据集不平衡的情况下, 默认的分类阈值的实验结果永远不是最优的. 最优阈值通常是通过最大化某个评估指标 (比如 *g-mean*[18]、*F1-score*[19] 等) 或者是依据正类的先验概率来确定的.

1.3 集成学习

集成学习是一种利用多种机器学习算法, 根据对数据提取的特征得出预测结果, 并用投票机制获得最终结果的方法. 有效地利用了每个算法的信息, 从而使得最终得到的模型具有更好的性能. 集成方法已经被广泛运用在数据集不平衡的问题中, 许多集成模型[20–22]被提出用于解决类不平衡问题.

1.4 采样方法

常见的数据采样策略可以用式 (1) 来概括:

$$p_j = \frac{n_j^q}{\sum_{k=1}^{K} n_k^q} \tag{1}$$

在数据集$D = \{(x_j, y_j), j = 1, \cdots, N\}$中, 一共有$K$个类, n_k表示第k类里包含的样本数, 样本总数$N = \sum_{k=1}^{K} n_k$. p_j表示第j类数据被采样的概率. q常见的取值是 0, 1, 1/2. 如果$q = 0$, 则称为基于类别的采样; 如果$q = 1$则是基于实例的采样, 即随机采样; $q = \frac{1}{2}$被称为平方根采样[14].

1.5 Mixup 及其相关变体

1.5.1 Mixup

Mixup 是由 Zhang 等人[11]提出的一种正则化技术, 也是一种数据增强方法. 是为了提供神经网络的泛化能力而提出的. 其思想是随机选择数据集D中的两个样本对$(x_i, y_i)(x_j, y_j)$, 通过式 (2) 得到它们的样本及标签的凸组合$(\hat{x}, \hat{y})$, 其中y_i和y_j是对应标签的独热编码, 随后在样本的凸组合上训练网络.

$$\begin{cases} \hat{x} = \lambda x_i + (1-\lambda) x_j \\ \hat{y} = \lambda y_i + (1-\lambda) y_j \end{cases} \tag{2}$$

其中, $\lambda \sim Beta(\alpha, \alpha), \alpha \in (0, \infty)$, 得到$\lambda \in [0, 1]$.

1.5.2 Remix

Mixup 对样本和标签使用的是相同的混合因子来混合特征空间和标签空间中的样本, 而 Remix[13]给样本和标签提供不同的混合因子λ_x和λ_y, 以便于为少数类分配更高的权重.

$$\begin{cases} \hat{x} = \lambda_x x_i + (1-\lambda_x) x_j \\ \hat{y} = \lambda_y y_i + (1-\lambda_y) y_j \end{cases} \tag{3}$$

其中:

$$\lambda_y = \begin{cases} 0, & n_i/n_j \geqslant \kappa \text{ and } \lambda_x < \tau \\ 1, & n_i/n_j \leqslant 1/\kappa \text{ and } 1-\lambda_x < \tau \\ \lambda_x, & \text{otherwise} \end{cases} \tag{4}$$

这里的κ和τ是作者定义的两个超参数, 便于更加合理地控制λ_y的值. 并且, 作者通过实验表明, 设置$\kappa = 3$和$\tau = 0.5$得到的实验结果最优. 在后续的实验部分, 我们也沿用这样的设置, 用于对比实验.

1.5.3 Balanced-mixup

不同于 Mixup 使用两个基于实例的采样器来随机选择两个样本进行混合, Balanced-mixup[14]使用一个基于实例的采样器S_I和一个基于类别的采样器S_C, 采样得到的样本分别表示为x_I和x_C. 基于类别的采样器S_C能够对样本进行平衡采样, 使得采样得到的数据分布是平衡的, 这样混合得到的数据分布会更加平衡.

$$\begin{cases} \hat{x} = \lambda x_I + (1-\lambda) x_C \\ \hat{y} = \lambda y_I + (1-\lambda) y_C \end{cases} \tag{5}$$

1.5.4 Label-occurrence-balanced mixup

和 Balanced-mixup[14]类似, Label-occurrence-balanced mixup[15]使用了两个基于类别的采样器S_{C1}和S_{C2}来代替 Mixup 原有的两个基于实例的采样器, 得到的样本分别表示为X_{C1}和X_{C2}. 这样混合得到的数据是接近于完全平衡的.

$$\begin{cases} \hat{x} = \lambda x_{C1} + (1-\lambda) x_{C2} \\ \hat{y} = \lambda y_{C1} + (1-\lambda) y_{C2} \end{cases} \tag{6}$$

为了方便起见, 后面我们用 Label-mixup 指代 Label-occurrence-balanced mixup.

1.6 支持向量机

支持向量机 (support vector machine) 是一种常见的二分类模型, 通过扩展可以实现多分类的任务. 它的目标是找到特征空间上的一个超平面, 不仅要使得两类数据分开, 而且各个类别的样本点中离这个超平面最近的点, 即支持向量, 到超平面的距离要最大化. 通过确定超平面来实现分类.

以二分类为例, 数据集$D = \{(x_i, y_i)\}, i = 1, \cdots, N, y_i \in \{1, -1\}$. SVM 的目标函数可以表示为在满足$y_i(w \cdot \Phi(x_i) + b_i) \geqslant 1 - \xi_i, \xi_i \geqslant 0, \forall i$的条件下, 最小化$\frac{1}{2} \| w \|^2 + C \sum_i \xi_i$. 其中, w, b是超平面$w^{\mathrm{T}} x + b = 0$的参数, Φ是一个将样本x_i从低维到高维的映射, ξ_i是松弛变量, C是惩罚参数, 用于控制对误分类点的容忍程度.

1.7 边界混合方法

现有的边界混合方法大多都是将选取的边界样本和 SMOTE 方法结合使用, 文献 [23] 通过计算少数类中每个样本的 k 个最近邻样本中多数类样本的个数,

来确定该样本是否属于边界样本, 对取得的少数类的边界样本采取 SMOTE 方法进行过采样. 文献 [24,25] 分别定义了区分边界样本与非边界样本的标准, 对满足标准的少数类样本, 使用 SMOTE 方法进行过采样, 对非边界中的多数类样本, 则进行欠采样, 从而达到重采样的目的. 这些研究确定边界样本的方法都是通过 K-means 算法选取样本的 k 个最近邻样本, 研究这些近邻样本和被选取样本之间的关系, 从而确定被选取样本是否为边界样本. 并且只对少数类的边界样本进行过采样, 对多数类的边界样本则不进行处理.

我们提出的方法使用 SVM 确定边界样本, 即支持向量, 相比于自定义边界样本的标准, 使用支持向量作为边界样本更加合理. 且对多数类和少数类的边界样本, 我们都增加了它们的采样概率, 并且我们设置少数类的边界样本的采样概率高于多数类的边界样本, 这样不仅区分了边界样本和非边界样本、少数类和多数类的重要程度, 也对多数类和少数类的边界样本一视同仁, 相对于它们的非边界样本, 均增加了相同倍数的采样概率. 重采样之后, 我们使用 Mixup 方法进行实验, 该方法和 SMOTE 方法的根本区别在于, SMOTE 是在同一类别里进行数据合成, 即假设邻近样本共享相同的类, 而 Mixup 是随机组合, 不考虑类别, 即合成的数据可能属于同一类别, 也可能属于不同类别, 模拟了不同类别之间的邻近关系, 这给模型带来了更多的正则化好处.

2 Borderline-mixup

Mixup[11] 思想是随机选择两个数据对, 得到这两个数据对的样本和标签对应凸组合, 来达到数据增强的目的. 这里可以理解为 Mixup[11] 是利用两个随机采样器来选择数据. 类似的, Balanced-mixup[14] 是采用一个类平衡采样器和一个随机采样器来选择数据, 而 Label-mixup[15] 则是选用两个类平衡采样器来进行实验.

不管是随机采样器还是类平衡采样器, 其对于样本的选择都是一视同仁的, 即每个类的样本与样本之间, 都有着相同的被采样的概率. 而我们认为, 在分类任务中, 不应该对特征空间中的每个样本点都给予相等的重视. 那些能够帮助我们区分其他类别的样本点理应得到更多的重视.

2.1 边界采样

在数据集不平衡问题中, 少数类样本可以分为两种: 本身数量并不少, 只是相对于多数类其占的比例较少, 即相对稀缺; 以及本身数量就是很少, 即绝对稀缺. 且有研究表明[26], 相对稀缺不一定会引起分类器的性能下降. 但是对于绝对稀缺的这种情况, 则需要研究人员尽可能地挖掘出少数类样本的有效信息.

对于少数类绝对稀缺的情况, 可以从类别之间的可分性出发, 如果类别之间的边界样本重叠较少, 即可分性较强, 那么类别不平衡并不会对分类器性能造成太大的影响. 从这个角度出发, 我们认为边界样本的重要性是要高于非边界样本的, 即边界样本理应得到更多的重视.

在选择边界样本的问题上, 我们受到了支持向量机的启发, 使用其选择边界样本, 即将超平面附近的支持向量作为边界样本, 赋予它们更高的采样概率, 用于后续实验.

研究表明, 特征空间中的最优分类超平面的权重可以表示为支持向量的线性组合[27], 这就说明, 最优超平面是独立于除支持向量之外的其他样本. 文献 [26] 表明, 支持向量机对类别不平衡问题不敏感, 因为它们的分类基于少量的支持向量, 并且大量的训练数据可以被认为是冗余的, 因此, 他们认为 SVM 是处理不平衡数据集的好选择. 这也在一定程度上证明了我们在不平衡分类中选择支持向量作为边界样本的合理性.

于是, 我们设计了一个边界采样器, 具体构建过程如下: (1) 使用 SVM 对不平衡数据集进行分类, 根据每个类别的样本数量分别设置不同的惩罚参数, 其与类别的样本数量成反比. (2) 得到 SVM 中每个类的支持向量, 也就是边界样本, 将其保存下来, 用于后续操作. (3) 根据每个类的支持向量数、样本数对其进行采样概率的设计. 赋予支持向量更高的权重, 使得它们被采样的概率更大. 构建所得到的边界采样器可以用于后续的混合操作.

2.2 采样概率设计

我们构建的边界采样器, 是在类平衡采样器的基础上进行改进, 赋予我们找到的边界样本更高的采样概率. 采样概率的设计, 除了类平衡这个条件之外, 我们还需要确定边界样本和非边界样本的比例, 比例确定好了之后, 就能够得出我们的采样概率.

我们在实验中尝试了几个不同的比例, 发现边界样本和非边界样本的采样概率比为 3:1 的时候, 实验所得的结果是最好的. 在文献 [5] 中, 作者在设置 balanced

cross entropy 的正负样本的权重时, 也得出了和我们相同的结论, 只不过这篇文章讨论的是正负样本的权重比例, 而我们设置的是边界样本和非边界样本的采样概率比.

先以二分类为例, 假设D是一个二类不平衡数据集, 多数类样本数为n_1, 支持向量数为z_1, 少数类样本数为n_2, 支持向量数为z_2. 则我们设置的边界采样器中, 非支持向量、多数类的支持向量和少数类的支持向量被采样的概率p为:

$$p=\begin{cases}\dfrac{1}{2(2z_1+n_1)}, \text{多数类的非支持向量}\\ \dfrac{1}{2(2z_2+n_2)}, \text{少数类的非支持向量}\\ \dfrac{3}{2(2z_1+n_1)}, \text{多数类的支持向量}\\ \dfrac{3}{2(2z_2+n_2)}, \text{少数类的支持向量}\end{cases} \tag{7}$$

可以看到, 我们设置的某一类的支持向量的采样概率是同类中非支持向量的 3 倍, 并且重新采样后多数类和少数类能够达到近似平衡的样本比.

扩展到k分类的情况, 还是令数据集为D, n_i, $i=1,2,\cdots,k$为第i类的样本数, z_i, $i=1,2,\cdots,k$为第i类的支持向量数, 则各个类的样本的被采样概率p为:

$$p=\begin{cases}\dfrac{1}{k}\times\dfrac{1}{2z_i+n_i}, \text{第}i\text{类的非支持向量}\\ \dfrac{3}{k}\times\dfrac{1}{2z_i+n_i}, \text{第}i\text{类的支持向量}\end{cases} \tag{8}$$

2.3 边界混合采样

我们对 Mixup 方法进行了改进, 不采用两个随机采样器对数据集进行采样构成凸组合, 而是使用两个边界采样器S_{B1}和S_{B2}, 得到混合样本为:

$$\begin{cases}\hat{x}=\lambda x_{\mathrm{B1}}+(1-\lambda)x_{\mathrm{B2}}\\ \hat{y}=\lambda y_{\mathrm{B1}}+(1-\lambda)y_{\mathrm{B2}}\end{cases} \tag{9}$$

其中, $\lambda\sim Beta(\alpha,\alpha)$, $\alpha\in(0,\infty)$, 得到$\lambda\in[0,1]$.

我们把我们提出的方法称为 Borderline-mixup.

3 实验

我们在 UCI 机器学习数据库以及 CIFAR10-LT 的长尾数据集上评估了我们提出的方法. 其中我们选择的 14 个 UCI 数据集是本身就具有不平衡性质的数据集, 其不平衡的程度各不相同.

CIFAR10-LT 是根据文献 [6,28] 构建的 CIFAR10 的长尾版本. 即不同类别的样本数量呈指数衰减, 在不平衡分类中经常作为基准的数据集用于比较.

3.1 数据集

3.1.1 UCI 机器学习数据库

在二分类和多分类实验中, 我们分别使用了来自 UCI 机器学习知识库的 10 个二分类不平衡数据集和 4 个多分类不平衡数据集, 如表 1 和表 2 所示, 二分类任务中包括 Spect[29]、Blood[30]、Yeast[31]、Abalone[32]、Ecoil[33]、Ionosphere[34]、Wilt[35]、Balance Scale[36]、Bank Marketing[37]、Fertility[38] 数据集; 多分类任务中使用了 Car Evaluation[39]、Avila[40]、Balance Scale[36] 和 Chess[41] 数据集我们对原有数据集进行分层采样, 得到训练集、验证集、测试集, 分别占原有数据集的 60%、20%、20%. 其中不平衡比例是在训练集上将多数类样本数除以少数类样本数得到的.

表 1 二分类实验中使用的 UCI 数据集

数据集	特征数	样本数	不平衡比例
Spect	22	267	3.848
Blood	4	748	3.226
Yeast	8	1 484	5.1
Abalone	8	4 177	2.20
Ecoil	8	336	15.75
Ionosphere	34	351	1.79
Wilt	6	4 889	58.16
Balance Sacle	4	625	11.90
Bank Marketing	17	45 211	8.15
Fertility	9	100	7.429

表 2 多分类实验中使用的 UCI 数据集

数据集	特征数	样本数	类别数	最大不平衡比例
Car Evaluation	6	1 728	4	18.615
Avila	10	10 430	12	857.2
Balance Scale	4	625	3	5.878
Chess	6	28 056	18	58.372

在二分类实验中, 对于多类数据集, 我们采取选用其中一类为正类, 其余类为负类的方法进行实验. 表 1 展示了用于二分类实验的 10 个数据集的相关信息, 表 2 展示了用于多分类实验的 4 个数据集的相关信息.

3.1.2 CIFAR10-LT

CIFAR10-LT 是由原始 CIFAR10 数据集, 在确定不平衡比例之后, 根据指数函数$n=n_t\times u^t$, 减少每个类的训练样本数量来创建的, 其中测试集不做改变. 这里, t为类索引, n_t为原始数据集中t类训练样本的数量, $u\in(0,1)$[6,28], 在不平衡比例和类别数已知的情况下, 可以算出u的值.

不平衡比例ρ被定义为数量最多的类别样本数除以数量最小的类别样本数, 取值范围一般是在 10 到 200 之间. 在我们的实验中, 选取ρ = 100, 200, 数据集的相关信息如表 3 所示.

表 3 实验中使用的 CIFAR10-LT 数据集

数据集	ρ = 200	ρ = 100
训练样本数	11 203	12 406
最大类别样本数	5 000	5 000
最小类别样本数	25	50

3.2 实验设置

3.2.1 UCI 数据集分类实验设置

对于 UCI 数据集分类任务, 我们选择四层感知机进行实验, 隐藏层的节点设置为输入层和输出层节点之和的 2/3, 设置α = 1, *epoch*大小为 300, *batch-size* 为 128, 取 15 次实验的平均值为结果.

3.2.2 CIFAR10 长尾数据集图像分类实验设置

对 CIFAR10 长尾数据集的实验, 我们选择 ResNet32 作为主干网络, 采用随机梯度下降方法, 其中动量为 0.9, 权重衰减为2×10^{-4}, *epoch*大小为 200, *batch-size* 为 128, 学习率初始化为 0.1, 在 160 个*epoch*和 180 个*epoch*时除以 10. 我们对前 5 个*epoch*采取热身操作[42]. 设置α = 1, 取 5 次实验的平均值为结果.

3.3 评估指标

对于不平衡数据集分类, 准确率不是一个很合理的评判标准, 所以在 UCI 二分类实验中, 我们选取*recall*, *F*1-*score*, *g-mean*作为评估标准进行比较.

在二分类的混淆矩阵中, *TP*表示真阳性, *FN*表示假阴性, *FP*表示假阳性, *TN*表示真阴性. 实验中, 设置少数类为正类, 多数类为负类. 评估指标*recall*, *F*1-*score*, *g-mean*分别表示如下.

*recall*表示的是对少数类的召回率, 即:

$$recall = \frac{TP}{TP+FN} \tag{10}$$

*F*1-*score*表示的是精确率*precision*和召回率*recall*的一个调和平均值, 表示为:

$$F1\text{-}score = \frac{2\times precision\times recall}{precision+recall} \tag{11}$$

其中, *precision*表示的是对少数类预测的精确率, 即:

$$precision = \frac{TP}{TP+FP} \tag{12}$$

*g-mean*在不平衡数据集分类中是常用的一个评估指标, 它是正类准确率和负类准确率的一个综合指标.

$$g\text{-}mean = \sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{13}$$

对于 CIFAR10 长尾图像数据集分类, 我们遵循常用的设置, 对其测试集不做改变, 保持平衡, 然后采用准确率*accuracy*作为评估标准.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{14}$$

3.4 实验结果及分析

实验选择 ERM (经验风险最小化)、Mixup 以及前面提到的 Mixup 的 3 个变体 Remix、Balanced-mixup、Label-mixup 作为对比方法.

3.4.1 UCI 二分类实验结果分析

二分类的实验结果如表 4–表 6 所示, 第 1 列表示数据集的名称, 第 1 行表示使用的方法. 我们的方法在第 1 行中用加粗标明, 每个数据集的实验最优值也用加粗表示.

表 4 各方法在 UCI 数据集上的*g-mean*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	**Borderline-mixup**
Spect	0.298	0.497	0.494	0.503	0.441	**0.534**
Blood	0.335	0	0	0.529	0	**0.573**
Yeast	0.682	0.640	0.691	0.700	0.637	**0.767**
Abalone	0	0	—	0.589	0	**0.598**
Ecoil	0.5	0.678	0.622	0.634	0.583	**0.786**
Ionosphere	0.849	0.858	—	0.872	0.716	**0.904**
Wilt	0	0	0	0	0	**0.914**
Balance Scale	0	0	0	0.728	0.493	**0.764**
Bank Marketing	0.560	0.567	**0.622**	0.568	0.535	0.575
Fertility	0	0.047	0	0	0	**0.228**

可以看到, 在这 10 个数据集中, 3 个评估标准的实验结果都表明: 我们的方法在绝大多数情况下都是最优的.

在*g-mean*的比较中, 我们的方法在除了 Bank Market-

ing 数据集之外的 9 个数据集中均取得最优的结果, 尤其是在 Wilt 和 Fertility 数据集中, 比较的几个方法取得的*g-mean*值大多都为 0, 即出现了把少数类均分类为多数类的情况, 而我们的方法 Borderline-mixup 分别能取得 0.914 和 0.228 的*g-mean*值. 在*recall*的比较中, 我们的方法也是在上述的 9 个数据集上都取得了最优. *F*1-*score*值在大多的数据集上也是取得了最优的结果. 在少数的几个数据集中, 我们的方法虽然没有取得最优结果, 但都排在第 2 或者第 3, 且与第 1 的性能相差不大.

由此可见, 我们的方法在二分类不平衡数据集上是有效的, 在极度不平衡的数据集上 (例如实验中的 Wilt 数据集), 我们的方法所取得的性远远高于其他几种方法, 这足以说明边界采样策略的有效性.

表 5 各方法在 UCI 数据集上的*recall*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	**Borderline-mixup**
Spect	0.09	0.272	0.272	0.285	0.212	**0.345**
Blood	0.114	0	0	0.299	0	**0.379**
Yeast	0.479	0.419	0.493	0.521	0.422	**0.668**
Abalone	0	0	—	0.458	0	**0.481**
Ecoil	0.25	0.467	0.4	0.417	0.35	**0.633**
Ionosphere	0.72	0.741	0.376	0.773	0.515	**0.835**
Wilt	0	0	0	0	0	**0.848**
Balance Scale	0	0	0	0.622	0.4	**0.741**
Bank Marketing	0.322	0.331	**0.398**	0.402	0.297	0.345
Fertility	0	0.033	0	0	0	**0.267**

表 6 各方法在 UCI 数据集上的*F*1-*score*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	**Borderline-mixup**
Spect	0.154	0.335	0.328	0.342	0.283	**0.349**
Blood	0.195	0	0	0.397	0	**0.427**
Yeast	0.59	0.544	**0.594**	0.574	0.521	0.587
Abalone	0	0	—	0.461	0	**0.472**
Ecoil	0.4	0.606	0.522	0.553	0.507	**0.674**
Ionosphere	0.837	0.845	0.537	0.857	0.678	**0.892**
Wilt	0	0	0	0	0	**0.671**
Balance Scale	0	0	0	**0.478**	0.316	0.435
Bank Marketing	0.423	0.443	**0.492**	0.475	0.398	0.418
Fertility	0	0.044	0	0	0	**0.129**

3.4.2 UCI 多分类实验结果分析

多分类的实验结果如表 7–表 9 所示, 对于实验的 4 个不平衡数据集, 我们的方法 Borderline-mixup 在*g-mean*和*recall*这两个评价指标中都取得了第 1 的结果, 且在大多数情况下实验结果远高于第 2 名.

在*F*1-*score*指标中, 有 2 个数据集没能取得第 1 的结果. 因为*F*1-*score*是*precision*和*recall*的调和平均值, 而我们的方法在*recall*这个指标上均能取得第 1, 故需要分析*precision*值来剖析原因所在. 在打印出 Avila 和 Chess 每个类别的*precision*值之后, 我们发现, 对于类别数量较少的类, ERM 方法所得到的*precision*值大多都为 0, 对于类别数量较多的类, 取得的*precision*值较高. 而我们的方法更关注少数类, 故在多数类的*precision*中会丢失一部分的精度. 多分类的*F*1-*score*值是对所有类的*F*1-*score*值取平均得到的结果, 由表 2 可知 Avila 和 Chess 是极不平衡的数据集, 其最大不平衡比例远高于另外两个数据集, 这会导致 ERM 和 Borderline-mixup 在数据集中的多数类上*precision*差异较大. 故我们分析得知: 在 Avila 和 Chess 数据集中, ERM 的*F*1-*score*值大于 Borderline-mixup 的*F*1-*score*值, 是因为这两个数据集的最大不平衡比例较高, ERM 在多数类上的*precision*值远高于 Borderline-mixup, 通过取平均值后, ERM 的*F*1-*score*值大于 Borderline-mixup 的*F*1-*score*值.

这说明了对于有的极不平衡数据集, 我们的方法在比较*precision*的时候可能会因为更关注少数类而丢失了一部分多数类的精度, 从而在*F*1-*score*值的比较上会低于 ERM 的结果. 但是, 比较 Mixup 及其变体, 我们的方法即使在这种情况下, 也能取到最优的结果, 这也反映了我们提出的 Mixup 变体是有效的.

表 7 各方法在 UCI 数据集上的*g-mean*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	**Borderline-mixup**
Car Evaluation	0.255	0.06	0	0.474	0	**0.793**
Avila	0.258	0.129	0	0.196	0	**0.559**
Balance Scale	0.627	0.620	0	0.932	0	**0.958**
Chess	0.547	0.350	0	0.516	0	**0.587**

表 8 各方法在 UCI 数据集上的*recall*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	**Borderline-mixup**
Car Evaluation	0.292	0.253	0.25	0.508	0.25	**0.740**
Avila	0.226	0.168	0.083	0.226	0.083	**0.423**
Balance Scale	0.650	0.642	0.333	0.908	0.333	**0.945**
Chess	0.389	0.235	0.056	0.352	0.056	**0.397**

表 9 各方法在 UCI 数据集上的*F*1-*score*值

数据集	ERM	Mixup	Remix	Balanced-mixup	Label-mixup	Borderline-mixup
Car Evaluation	0.281	0.215	0.090	0.425	0.090	**0.555**
Avila	**0.223**	0.139	0.049	0.132	0.049	0.220
Balance Scale	0.624	0.618	0.211	0.845	0.211	**0.862**
Chess	**0.389**	0.212	0.010	0.257	0.010	0.260

多分类的实验结果可以表明, 我们的方法不仅在二分类中是有效的, 在多分类中也能取得优异的结果. 这进一步证明了我们设计的边界采样策略的有效性.

3.4.3 CIFAR 长尾数据集图像分类实验结果分析

CIFAR10-LT 数据集中, 测试集是保持不变, 即平衡的. 所以我们采用*accuracy*来衡量各方法的性能. 实验结果如表 10 所示.

表 10 各方法在 CIFAR10-LT 中的*accuracy* (%)

数据集	CIFAR10-LT_100	CIFAR10-LT_200
ERM	72.154	65.58
Mixup	73.13	67.102
Remix	75.76	69.92
Balanced-mixup	75.85	69.906
Label-mixup	75.51	69.806
Borderline-mixup	**76.296**	**70.722**

可以看到, 在基准的图像长尾数据集 CIFAR10-LT 当中, 我们的方法是最优的. 在不平衡比例$\rho=100$时, 相比 Mixup 我们的方法 Borderline-mixup 提升了 3.1% 左右, 在$\rho=200$时, Borderline-mixup 相比于 Mixup 提升了 3.6% 左右.

4 结语

本文提出了一种数据增强 (扩充) 方法: 边界混合 (Borderline-mixup), 旨在于提高神经网络在不平衡数据集上的分类性能, Borderline-mixup 的创新点在于, 它使用 SVM 先选择出边界样本, 依据我们给定的采样概率得到两个边界采样器的样本, 再对得到的样本进行混合. 该算法在不平衡的二分类和多分类数据集以及 CIFAR10 长尾数据集上都取得了优于 Mixup 及其相关变体的结果, 实验证明了我们提出的 Borderline-mixup 算法在处理不平衡数据集的有效性. 日后我们还需对边界样本的采集以及实验进行更深入研究和扩展.

参考文献

1 Japkowicz N. Learning from imbalanced data sets: A comparison of various strategies. Proceedings of the 2000 AAAI Workshop on Learning from Imbalanced Data Sets. Menlo Park: AAAI Press, 2000. 10–15.

2 Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1–6. [doi: 10.1145/1007730.1007733]

3 Sun YM, Wong AKC, Kamel MS. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4): 687–719. [doi: 10.1142/S0218001409007326]

4 Margineantu DD. Class probability estimation and cost-sensitive classification decisions. Proceedings of the 13th European Conference on Machine Learning. Helsinki: Springer, 2002. 270–281.

5 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense

object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.

6 Cui Y, Jia ML, Lin TY, *et al*. Class-balanced loss based on effective number of samples. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9268–9277.

7 Dong XB, Yu ZW, Cao WM, *et al*. A survey on ensemble learning. Frontiers of Computer Science, 2020, 14(2): 241–258. [doi: 10.1007/s11704-019-8208-z]

8 Feng W, Huang WJ, Ren JC. Class imbalance ensemble learning based on the margin theory. Applied Sciences, 2018, 8(5): 815. [doi: 10.3390/app8050815]

9 Guo HX, Li YJ, Shang J, *et al*. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 2017, 73: 220–239. [doi: 10.1016/j.eswa.2016.12.035]

10 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]

11 Zhang HY, Cissé M, Dauphin YN, *et al*. Mixup: Beyond empirical risk minimization. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.

12 Zhang YS, Wei XS, Zhou BY, *et al*. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 3447–3455. [doi: 10.1609/aaai.v35i4.16458]

13 Chou HP, Chang SC, Pan JY, *et al*. Remix: Rebalanced mixup. Proceedings of the 2020 European Conference on Computer Vision. Glasgow: Springer, 2020. 95–110. [doi: 10.1007/978-3-030-65414-6_9]

14 Galdran A, Carneiro G, González Ballester MA. Balanced-mixup for highly imbalanced medical image classification. Proceedings of the 24th International Conference on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 323–333. [doi: 10.1007/978-3-030-87240-3_31]

15 Zhang SY, Chen C, Zhang XJ, *et al*. Label-occurrence-balanced mixup for long-tailed recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 3224–3228. [doi: 10.1109/ICASSP43922.2022.9746299]

16 Liu XY, Wu JX, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 539–550. [doi: 10.1109/TSMCB.2008.2007853]

17 Johnson JM, Khoshgoftaar TM. Deep learning and thresholding with class-imbalanced big data. Proceedings of the 18th IEEE International Conference on Machine Learning and Applications. Boca Raton: IEEE, 2019. 755–762. [doi: 10.1109/ICMLA.2019.00134]

18 Johnson JM, Khoshgoftaar TM. Thresholding strategies for deep learning with highly imbalanced big data. Deep Learning Applications, Volume 2. Singapore: Springer, 2021. 199–227. [doi: 10.1007/978-981-15-6759-9_9]

19 Zou Q, Xie SF, Lin ZY, *et al*. Finding the best classification threshold in imbalanced classification. Big Data Research, 2016, 5: 2–8. [doi: 10.1016/j.bdr.2015.12.001]

20 Galar M, Fernández A, Barrenechea E, *et al*. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition, 2013, 46(12): 3460–3471. [doi: 10.1016/j.patcog.2013.05.006]

21 Tama BA, Lim S. Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. Computer Science Review, 2021, 39: 100357. [doi: 10.1016/j.cosrev.2020.100357]

22 Kamalov F, Moussa S, Avante Reyes J. KDE-based ensemble learning for imbalanced data. Electronics, 2022, 11(17): 2703. [doi: 10.3390/electronics11172703]

23 Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of the 2005 International Conference on Intelligent Computing. Hefei: Springer, 2005. 878–887. [doi: 10.1007/11538059_91]

24 冯宏伟, 姚博, 高原, 等. 基于边界混合采样的非均衡数据处理算法. 控制与决策, 2017, 32(10): 1831–1836. [doi: 10.13195/j.kzyjc.2016.1044]

25 侯贝贝, 刘三阳, 普事业. 基于边界混合重采样的非平衡数据分类方法. 计算机工程与应用, 2020, 56(1): 46–52. [doi: 10.3778/j.issn.1002-8331.1901-0083]

26 Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis, 2002, 6(5): 429–449. [doi: 10.3233/IDA-2002-6504]

27 He Q, Xie ZX, Hu QH, *et al*. Neighborhood based sample and feature selection for SVM classification learning. Neurocomputing, 2011, 74(10): 1585–1594. [doi: 10.1016/j.neucom.2011.01.019]

28 Cao K, Wei C, Gaidon A, *et al*. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in

Neural Information Processing Systems, 2019, 32.
29 Cios K, Kurgan L, Goodenday L. Spect heart. UCI Machine Learning Repository, 2021. [doi: 10.24432/C5P304]
30 Yeh IC. Blood transfusion service center. UCI Machine Learning Repository, 2008. [doi: 10.24432/C5GS39]
31 Nakai K. Yeast. UCI Machine Learning Repository, 1996. [doi: 10.24432/C5KG68]
32 Nash W, Sellers T, Talbot S, *et al*. Abalone. UCI Machine Learning Repository, 1995. [doi: 10.24432/C55C7W]
33 Nakai K. Ecoli. UCI Machine Learning Repository, 1996. [doi: 10.24432/C5388M]
34 Sigillito V, Wing S, Hutton L, *et al*. Ionosphere. UCI Machine Learning Repository, 1988. [doi: 10.24432/C5W01B]
35 Johnson B. Wilt. UCI Machine Learning Repository, 2014. [doi: 10.24432/C5KS4M]
36 Siegler R. Balance scale. UCI Machine Learning Repository, 1994. [doi: 10.24432/C5488X]
37 Moro S, Rita P, Cortez P. Bank marketing. UCI Machine Learning Repository, 2012. [doi: 10.24432/C5K306]
38 Gil D, Girela J. Fertility. UCI Machine Learning Repository, 2013. [doi: 10.24432/C5Z01Z]
39 Bohanec M. Car evaluation. UCI Machine Learning Repository, 1997. [doi: 10.24432/C5JP48]
40 Stefano C, Fontanella F, Maniaci M, *et al*. Avila. UCI Machine Learning Repository, 2018. [doi: 10.24432/C5K02X]
41 Bain M, Hoff A. Chess (king-rook vs. king). UCI Machine Learning Repository, 1994. [doi: 10.24432/C57W2S]
42 Goyal P, Dollár P, Girshick R, *et al*. Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv: 1706.02677, 2017.

(校对责编: 牛欣悦)