DOI: 10.20103/j.stxb.202405151099

梁天泓,殷高方,邵新童,赵南京,张小玲,贾仁庆,徐敏,张子豪,胡翔,黄朋,董鸣,陈晓伟.浮游植物类别不均衡图像分类方法对比研究.生态学报,2025,45(7):3534-3543.

Liang T H, Yin G F, Shao X T, Zhao N J, Zhang X L, Jia R Q, Xu M, Zhang Z H, Hu X, Huang P, Dong M, Chen X W. Comparative study of class-imbalanced image classification algorithms for phytoplankton. Acta Ecologica Sinica, 2025, 45(7): 3534-3543.

浮游植物类别不均衡图像分类方法对比研究

梁天泓[1,2],殷高方[1,2,3,*],邵新童[1],赵南京[1,2,3,4],张小玲[4],贾仁庆[1],徐 敏[1,2],张子豪[4],胡 翔[1],黄 朋[1],董 鸣[3],陈晓伟[1,2]

1 中国科学院 安徽光学精密机械研究所,合肥 230031

2 中国科学技术大学,合肥 230026

3 合肥综合性科学中心环境研究院,合肥 230071

4 安徽大学,合肥 230061

摘要:自然水体中浮游植物物种丰富且类别分布不均,采集的显微图像中优势类别样本远多于劣势类别样本,导致深度学习方法在劣势类别上的分类准确率低。针对浮游植物类别不平衡引起的深度学习模型分类误差问题,分析了宏观领域类别不平衡问题的多种解决方法和策略,探究这些方法在浮游植物显微图像领域的实用性。采集了巢湖流域中常见的 29 个藻属、18044 张图像,构建了具有严重类别不平衡特性的浮游植物显微图像数据集,并提出使用微平均和宏平均综合评价模型的分类能力。实验结果表明,常规方法训练的模型预测劣势类别样本时的 F1 值较低,而使用重采样大类中平方根采样法训练的模型在微平均和宏平均两个指标上均有明显提升,分类 F1 值分别达到了 0.932 和 0.852。特别地,在样本数量最少的 10 个类别上,微平均和宏平均的 F1 值分别提高了 9.64%和 15.94%。为自然水体浮游植物群落结构自动化检测提供了更有效的深度学习模型训练方法。

关键词:浮游植物;显微;深度学习;类别不均衡;图像分类

Comparative study of class-imbalanced image classification algorithms for phytoplankton

LIANG Tianhong[1,2], YIN Gaofang[1,2,3,*], SHAO Xintong[1], ZHAO Nanjing[1,2,3,4], ZHANG Xiaoling[4], JIA Renqing[1], XU Min[1,2], ZHANG Zihao[4], HU Xiang[1], HUANG Peng[1], DONG Ming[3], CHEN Xiaowei[1,2]

1 *Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei* 230031, *China*

2 *University of Science and Technology of China, Hefei* 230026, *China*

3 *Institute of Environment, Hefei Comprehensive National Science Center, Hefei* 230071, *China*

4 *Anhui University, Hefei* 230061, *China*

Abstract: The distribution of phytoplankton classes in freshwater in imbalanced, with collected microscopic images containing significantly more samples of advantaged classes than of disadvantaged items. General deep-learning-based image classification methods trained on such datasets generally perform poorly in classifying disadvantaged classes. In addressing the classification errors caused by the class-imbalanced phytoplankton dataset in deep learning model, various solutions for handing this issue in macro-domain have been analyzed. The practicality of these methods in the domain of microscopic images of phytoplankton is explored. A dataset consisting of 29 genera and 18044 images from Lake Chaohu was collected,

基金项目:安徽省科技重大专项(202203a07020002);合肥综合性科学中心环境研究院科研团队建设项目(HYKYTD2024004);安徽省生态环境科研项目(2023hb0011);中国科学院合肥物质科学研究院院长基金(YZJJ2024QN01)

收稿日期:2024-05-15; **网络出版日期**:2025-01-07

* 通讯作者 Corresponding author.E-mail: gfyin@aiofm.ac.cn

constructing a microscopic image dataset of phytoplankton with class-imbalanced problem. An evaluation of the model's classification abilities was proposed using both micro-average and macro-average metrics. Experimental results indicate that the model trained by general method performs lower F1 values when predicting samples from disadvantaged classes. Conversely, the model trained by the square-root sampling method in the re-sampling major category exhibit significant improvement in both micro-average and macro-average metrics, with F1 values reaching 0.932 and 0.852, respectively. Particularly, on the top 10 disadvantaged genera, the F1 values for micro-average and macro-average increased by 9.64% and 15.94%, respectively. This study provides an effective method for training deep learning model for the automated detection of phytoplankton community structure in freshwater.

Key Words: phytoplankton; microscopic; deep learning; class imbalanced; image classification

浮游植物是淡水湖泊生态链中主要的初级生产者[1],其群落结构具有重要的生态学意义[2],是水生态环境检测中的重要指标[3]。目前,浮游植物鉴定的标准方法为人工镜检法,该方法对鉴定人员的浮游植物分类学专业知识和显微镜检工作经验有较高要求,且费时费力、效率低下,无法满足快速、高频次、大范围监测与调查需求[4—5]。随着计算机视觉理论和技术的发展,特别是基于深度学习的图像识别方法的突破,浮游植物显微图像自动鉴定技术有望让鉴定工作减少对专业人员的依赖,为浮游植物多样性监测和水生态研究提供高效工具,已成为领域的研究热点[6]。

目前深度学习算法已广泛地应用于浮游植物显微图像识别研究[7—11],并取得了较大的研究进展。对于类别平衡的样本,即训练集中每个浮游植物类别的图像数量相当,分类准确率已能达到较高的水平[12—13]。然而在实际应用中,淡水湖泊的浮游植物优劣势明显,优势种图像数量远多于劣势种数量[14—18],导致从中采集并拍摄的浮游植物显微图像数据呈现严重的类别不平衡特性。例如,Park 等人从韩国的监测站采集的 7 个藻种中,图像数量最多的微囊藻有 360 张,而角星鼓藻数量最少,仅有 42 张[19];Li 等人从海洋水样采集的 10 个藻种中,图像数量最多的聚球藻有 1633 张,离心列海链藻数量最少,仅有 96 张[20]。类别不均衡使模型的预测结果偏向于有更多训练数据的优势种类,导致在数据量有限的劣势种类上识别准确率不高[21]。

目前针对类别不平衡图像的分类研究大致可分为重采样和重权重两类[22—23]。其中重采样方法的核心思想是通过提高劣势类别、降低优势类别的被采样概率。Huang 等人使用了类别平衡采样法训练模型,令每个类别被采样到的概率相同,回避了类别不平衡问题[24]。Mahajan 等人提出了改进的平方根采样法,以类别样本数量的平方根作为概率进行采样,缩小了优势类别与劣势类别被采样概率的差距,同时缓解模型过分关注劣势类别、优势类别未学习充分的问题[25]。

重权重方法则是在计算损失的过程中引入类别样本数量分布,样本根据所属类别获得不同的权重,调整不同优劣势类别对损失的贡献比例。根据定义,最简易的实现为在样本损失上除以所属类别出现的频率,使模型更加关注劣势样本。Cui 等人在损失函数中引入有效样本数的概念,使损失值随类别样本数量增加单调递减[26]。Menon 等人提出的 Logit Adjusted Loss 函数[27],均将类别样本数量引入交叉熵损失函数的 Softmax 函数中,通过增大劣势类别样本的 Softmax 函数输出值,避免优势类别对劣势类别造成干扰。

重采样与重权重方法在开源的大型宏观数据集上(如 ImageNet-LT[28])均取得了良好的效果,能有效缓解类别不平衡带来的问题。针对类别不平衡引起的深度学习模型分类误差问题,以及鲜有在小型的浮游植物显微图像数据集上验证方法有效性研究的问题,本文详细分析了上述方法的核心要点与优势并探究可能存在问题。通过构建具有严重类别不平衡的巢湖浮游植物显微图像数据集,实验探究了其他领域处理方法在浮游植物显微图像上的效果,训练模型并以分类 F1 值的微平均和宏平均综合指标对模型性能进行评价和测试。

1　数据来源

本文以巢湖流域的浮游植物为分析对象,依据王徐林等人对该流域浮游藻类功能群落的研究[29],选取了

巢湖流域中具有代表性的类群进行分析。通过在龟山、东湖心、黄麓、中庙、湖滨、西湖心、新河和兆河等八个监测点位采集水样后,人工操作显微镜拍摄图像的方式,获取了 18044 张清晰的图像,涵盖了 29 个属,构建了一个具有严重类别不平衡特征的浮游植物显微图像数据集,各属图像数量分布见图 1。图中,纵轴为藻属所含图像的数量,横轴为藻属名称并按照数量从大大小排列。其中,裸藻属(*Euglena*)的图像数量最多高达到 3623 张,而拟新月藻属(*Closteriopsis*)数量最少仅有 16 张。若假设裸藻属(*Euglena*)至微囊藻属(*Microcystis*)这前 9 个包含图像数量最多的藻属为优势藻属,剩余的 20 个藻属为劣势藻属,通过分别累计优劣势藻属的图像数量可知优势藻属占据了总图像数量的 70%,而劣势藻属仅占据了总图像数量的 30%,由此可认为本文收集的巢湖流域浮游植物显微数据集具有严重的类别不均衡特性。

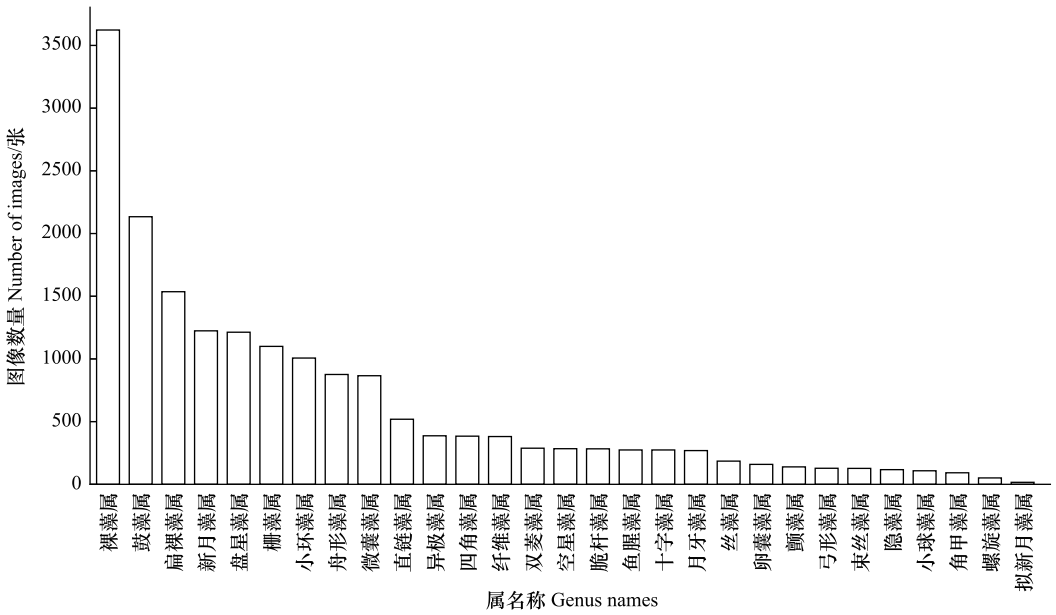

图 1 采集的巢湖浮游植物 29 个属类图像数量分布

Fig.1 The distribution of collected 29 genera of phytoplankton from Lake Chaohu

2 方法与模型

深度学习图像分类任务常采用交叉熵损失函数,其中基于 Softmax 函数能将任意特征分布转化为伪概率,并利用 KL 散度(Kullback-Leibler divergence)能够度量两个概率分布之间差异的特性指导模型训练过程。运用交叉熵损失函数可以使网络预测的类别概率分布尽可能与标签概率分布保持一致,训练流程如图 2 所示。

首先,在常规的模型训练策略中,通常使用均匀采样的数据采样方法,即每张图像的被采样概率相等。然后,被采样的藻细胞图像经由卷积神经网络获取其显微图像的形态特征。此时,网络输出的特征维度可能与模型预测的类别数量不匹配,使用一个全连接层将特征维度和类别数量对齐。接着,将对齐的分类输出输入交叉熵损失函数中,计算出损失值。最后,依据反向传播的原理更新网络参数。常规的交叉熵损失函数如式(1)所示:

$$L_{CE} = -\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T x_i + b_j}} \tag{1}$$

其中,W 和 b 分别表示全连接层的权重和偏置,x 表示卷积神经网络输出的特征向量,C 表示模型需预测的类别数量,N 表示批大小。

2.1 对比方法模型

如前文所述,本文调研的重采样与重权重方法在开源的大型宏观数据集上均取得了良好的效果,能有效

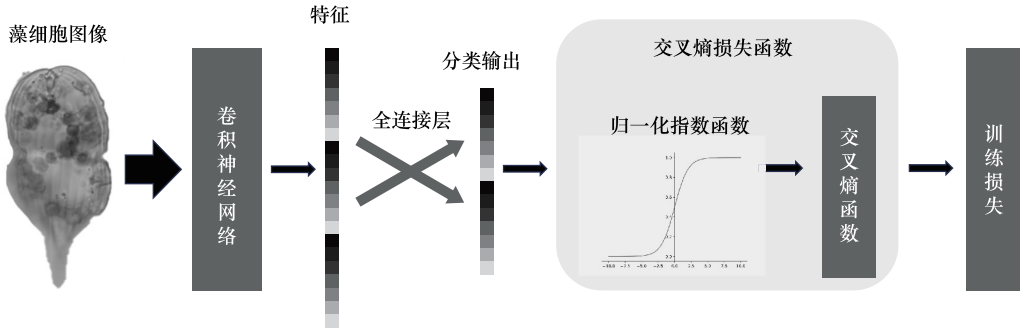

图 2　典型的的深度学习图像分类流程图

Fig.2　Flow chart of typical deep-learning-based image classification

地缓解类别不平衡造成分类准确率低的问题。然而,上述方法并不完全适用于小型的浮游植物显微图像数据集,因此本章首先将分析重采样与重权重方法的核心思想、优势与劣势,如表 1 所示。

表 1　解决类别不均衡问题方法的核心思想、优势和劣势对比

Table 1　Core ideas, advantages, and disadvantages comparison of methods to address the issue of class-imbalanced

大类 Category	方法 Method	核心思想 Core idea	优势 Advantage	劣势 Disadvantage
重采样 Re-sampling	类别平衡采样法	使优劣势类别以相同概率被采样	完全忽略了类别不平衡问题 降低了类别不平衡程度	优势类别欠拟合、对劣势类别过拟合
	平方根采样法	使优劣势类别以样本数量的平方根作为概率被采样	缓解了类别平衡采样法的拟合问题适用范围广	无明显缺陷
	解耦法	先使用常规方法训练模型学习图像特征再使用类别平衡采样法微调	对类别不平衡程度不敏感	两阶段训练,超参数(如训练轮次)设置不易
重权重 Re-weighting	重权重	样本损失值乘以所属类别样本数量占总量比例的倒数	增大劣势类别损失贡献 减小优势类别损失贡献	对超参数敏感 模型训练难度大
	有效样本数	使损失值随类别样本数量增加单调递减	增大劣势类别损失贡献 减小优势类别损失贡献	未解决如何获取有效样本数的问题 仅将其当作训练超参数,设置不易
	Logit adjusted loss	增大劣势类别样本的 Softmax 函数输出值	降低优势类别对劣势类别的干扰	具有隐形前提假设: 类别所含的样本数量等价与类别分类的难度

2.2　模型性能评价方法

在常规的图像分类任务中,各类别含有的样本数量差异小,因此可采用微平均作为模型性能的评价指标,具体公式如(2)所示。其中P_{micro}、R_{micro}和F_{micro}分别代表微查准率、微查全率和微 F1 值;TP、FP 和 FN 分别代表真正例、假正例和真负例;N 代表所有的样本。从公式可知,微平均以样本作为基本计算单位,如果某一类的样本数量多,该类别在评价中占的比重就大;对于有严重类别不平衡问题的数据集,微平均无法有效反应模型对于样本数量较少类别的预测能力。

$$\begin{cases} P_{\text{micro}} = \dfrac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i} \\ R_{\text{micro}} = \dfrac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i} \\ F_{\text{micro}} = \dfrac{2 \times P_{\text{micro}} \times R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \end{cases} \tag{2}$$

为了合理评估模型对于劣势类别的识别性能,本文选择使用宏平均作为主要的评价指标,公式如(3)所示。其中P_{micro}、R_{micro}和F_{micro}分别代表宏查准率、宏查全率和宏 F1 值;C 代表所有的类别。从公式中可以看出,宏平均以类别作为基本计算单位,每个类别的权重相等,与其所包含的样本数量无关,因此能够更好地评价模型对于劣势种类的预测能力。

$$\begin{cases} P_{\text{macro}} = \dfrac{1}{C}\sum_{i=1}^{C} P_i \\ R_{\text{macro}} = \dfrac{1}{C}\sum_{i=1}^{C} R_i \\ F_i = \dfrac{2 \times P_{\text{macro}} \times R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \\ F_{\text{macro}} = \dfrac{1}{C}\sum_{i=1}^{C} F_i \end{cases} \tag{3}$$

2.3 训练超参数设置

本文中,训练集与测试集按照 8:2 的比例随机划分,并确保各个类别对应的训练图像数量与测试图像数量比例均为 8:2,最终形成 14423 个训练样本和 3621 个测试样本。尽管本文所收集的藻类显微图像数据集数量较多,但相对比 ImageNet-LT[28] 等开源数据集仍然较少,这可能导致模型过拟合,从而影响评价结果的准确性。为降低模型过拟合的风险,本文选择使用的骨干网络和解决类别不平衡问题的算法均尽可能地简化超参数,以避免复杂的超参数设置和调节对实验结果造成干扰。

为公平比较不同策略与方法解决浮游植物类别不均衡问题的能力,本文选择使用同一骨干网络进行实验。在宏观领域解决类别不均衡问题的相关研究中,ResNet[30] 作为经典的卷积神经网络常被选作骨干网络,因此本文同样遵循这一设置。经过初步的探究,训练结果表明 ResNet 系列网络中 ResNet-18 和 ResNet-34 由于网络较小,分类准确率较 ResNet-50 低;而 ResNet-101 和 ResNet-152 则由于参数量过大,模型在训练过程中易陷于过拟合的情形,分类准确率反而不如参数量更少的 ResNet-50 模型。本文同样调研了近年来计算机视觉领域流行的模型,如 ConvNeXt[31],但实验发现这类新兴的网络结构对于数据集尺寸要求高,在本文采集的浮游植物数据集上表现不佳,具有对超参数敏感、训练过程不稳定、难以收敛等问题。综上,本文最终决定在所有实验中使用 ResNet-50 模型作为骨干网络,提取到特征后,利用交叉熵损失函数计算真实值与预测值之间的误差,后通过小批量梯度下降法更新模型参数,其中每个小批量样本数为 64、学习率为 0.1、动量因子为 0.9、权重衰减为 0.0001。使用更多更重的数据增强能够较为明显地提高模型地分类准确率,但由于准确率基数地提高,从而稀释了方法策略本身对模型准确率地影响。因此文本为了更清晰地凸显出各方法策略的效果,所有图像将长边缩放至 320 像素,短边按比例缩放并填充黑色像素;数据增强策略仅使用随机反转。本文采用余弦退火和预热轮数为 5 的学习率衰减策略;共训练 300 轮次,并取最后一轮次的模型权重作为最终训练结果,用于统计测试集的评价指标。

3 实验结果与分析

本文将上述方法应用于采集的巢湖流域 29 个属的浮游植物显微图像数据集中。对比分析了多种重采样、重权重的深度学习模型训练方法,以常规的交叉熵损失函数的实验结果为例展示类别不均衡给模型训练带来的不良影响,以及“平方根采样法”中部分劣势藻属预测错误的案例分析。实验的硬件配置如下,CPU 为 Intel Xeon Silver 4210 Process * 2,内存配置为 DDR4 2666MHz 16G * 16,GPU 配置为 Nvidia GeForce RTX 3090 * 4,模型训练所使用的框架为 PyTorch 2.0.1,Python 环境为 3.10.12。

3.1 不同方法对比实验结果

实验结果如表 2 所示,从微平均和宏平均两个指标均可以看出,重采样与重权重方法中各有表现优异的

方法。具体而言,重采样大类中的平方根采样法表现最佳,特别是在宏平均指标上从 0.810 提升至 0.852,显著提高了在模型对于劣势类别的分类准确率。Logit Adjusted Loss 方法较交叉熵损失函数均有所提升,在前文的分析中可看出两个方法的思想相近,呈现的结果也较为相近,但二者在宏平均指标上提升有限。解耦法是以常规方法训练好的网络作为预训练模型,通过类别平衡(批)采样法进行微调 10 个轮次,表现较好。然而,重权重大类中的重权重方法表现不佳,两个指标较常规方法更低。重权重方法由于简单地根据样本数量的倒数,缩小优势类别样本、放大劣势类别样本的损失值,造成了模型对优势类别欠拟合的问题,微平均指标下降严重,宏平均指标更无分析必要。

表 2　不同方法微平均与宏平均分类 F1 值比较

Table 2　Comparison of micro-average & macro-average F1 values of different methods

大类 Category	方法 Method	微平均 Micro-average	宏平均 Macro-average
	常规图像分类方法	0.910	0.810
	类别平衡采样法	0.927	0.835
重采样 Re-sampling	平方根采样法	0.932	0.852
	解耦法(类别平衡批采样)	0.918	0.826
	解耦法(类别平衡采样)	0.923	0.832
	重权重	0.878	0.756
重权重 Re-weighting	有效样本数	0.926	0.846
	Logit Adjusted Loss	0.921	0.820

3.2　常规图像分类方法与平方根采样法实验结果

如前文所述,图像分类常规方法使用交叉熵损失函数指导模型训练,难以解决数据集类别不平衡的问题;在表 2 中交叉熵损失函数的微平均值达到 0.910,而此时的宏平均值仅有 0.810,远低于微平均值,表明了类别不平衡问题对模型预测能力具有显著的负面影响。图 3 进一步对比分析了各个属类的样本数量与分类准确率的关系。F1 值最高的 5 个属类为盘星藻属(0.977)、裸藻属(0.957)、鼓藻属(0.956)、角甲藻属(0.950)、

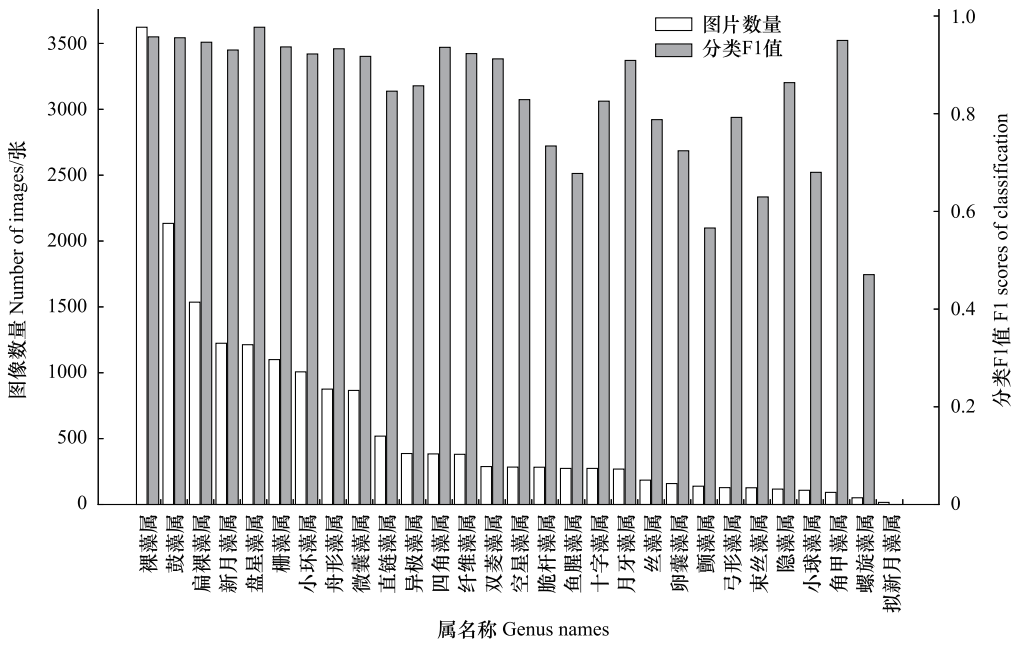

图 3　巢湖浮游植物 29 个属类图像数量常规图像分类方法 F1 值

Fig.3　The image number of 29 genera of phytoplankton from Chaohu Lake and F1 values of general image classification method

扁裸藻属(0.946);最低的 5 个属类分别为鱼腥藻属(0.678)、束丝藻属(0.630)、颤藻属(0.566)、螺旋藻属(0.471)、拟新月藻属(0.000)。从样本数量上看,F1 值前 5 名的属类对应的样本数量排名分别为第 5、第 1、第 2、第 27、第 3;最后 5 名的属类对应的样本数量排名分别为:第 17、第 24、第 22、第 28、第 29。总体趋势是,类别样本图像越多,分类的准确率就越高;反之,样本数量稀少的类别,分类准确率会受到严重的负面影响。

采用改进的平方根采样法结合交叉熵损失函数后,微平均和宏平均指标均有所提升,如表 3 所示。微平均从 0.910 增加至 0.932,提升了 2.42%;宏平均从 0.810 增加至 0.852,提升了 5.19%。特别地,利用归一化中心损失函数方法训练出的模型在对劣势类别的分类准确率上有显著提升。例如在图像数量最少的 10 个藻属中,微平均从 0.716 增加至 0.785,提升了 9.64%;宏平均从 0.646 增加至 0.749,提升了 15.94%。如前文所述,常规图像分类方法中默认使用均匀采样法对数据进行采样训练,每个类别数据的被训练次数为与其所含数据量呈正比,即样本数量越少模型学习的机会越少;正如实验结果所示,均匀采样法难以使模型关注到图像数量少的劣势藻属。而平方根采样法,其将每个类别样本数量的平方根作为被采样概率,在保留了训练次数与数据量正相关的情况下,降低了模型学习劣势类别与优势类别机会的差异性,有效增加了模型预测劣势类别的能力。具体而言,10 个藻属中有 8 个的 F1 值都有提高,卵囊藻属无变化,仅有螺旋藻属的 F1 值略有降低;特别是解决了拟新月藻属预测完全错误的问题,角甲藻属、束丝藻属和弓形藻属的 F1 值的提升也高达 17.70%、17.69% 和 15.73%。由此可见,平方根采样法能够有效提高多数劣势藻属的分类准确率,但对于一些形态特征较为明显、较容易分类的类别,如卵囊藻属、螺旋藻属等,则没有提升,甚至由于模型更加关注劣势藻属从而形成过拟合现象,使得该类别的分类准确率下降。

表 3 部分劣势类别的常规图像分类方法与平方根方法 F1 值对比

Table 3 Comparison of F1 values between general image classification method and the square-root sampling method for part of disadvantaged genera

属/评价指标 Genus/Evaluation Metrics	常规图像分类方法 General Method	平方根采样法 Square-Root Sampling	百分比变化(相对) Relative Percentage Change
微平均 *Micro-Avg*	0.910	0.932	+2.42%
宏平均 *Macro-Avg*	0.810	0.852	+5.19%
丝藻属 *Ulothrix*	0.857	0.886	+3.38%
卵囊藻属 *Oocystis*	0.774	0.774	+0%
颤藻属 *Oscillatoria*	0.622	0.627	+0.8%
弓形藻属 *Schroederia*	0.712	0.824	+15.73%
束丝藻属 *Aphanizomenon*	0.588	0.692	+17.69%
隐藻属 *Aphanocapsa*	0.826	0.844	+2.18%
小球藻属 *Chlorella*	0.791	0.851	+7.59%
角甲藻属 *Ceratocorys*	0.857	0.923	+17.70%
螺旋藻属 *Spirulina*	0.632	0.609	−3.78%
拟新月藻属 *Closteriopsis*	0.000	0.333	+∞
数量最少的 10 个属微平均 Micro-Avg of Last 10 Genera	0.716	0.785	+9.64%
数量最少的 10 个属宏平均 Macro-Avg of Last 10 Genera	0.646	0.749	+15.94%

3.3 分类错误分析

为进一步探究模型对与劣势藻属的分类能力以及缺陷,以“平方根采样法”训练的 ResNet-50 分类模型作为分析对象,详细说明数量最少的 5 个藻属的预测情况。如图 4 所示,横向为测试集中最劣势的 5 个藻属,下方为测试图像数量,纵向为 5 个劣势藻属预测错误出的类别,图中对应的数字为错误预测的个数。

数量最少的“拟新月藻属”测试的 4 个样本中,3 个样本被错误地预测为“新月藻属”,从形态上分析,“拟新月藻属”与“新月藻属”在余下的藻属中最为接近,同时由于深度学习分类通常只接受尺寸相同的输入,因

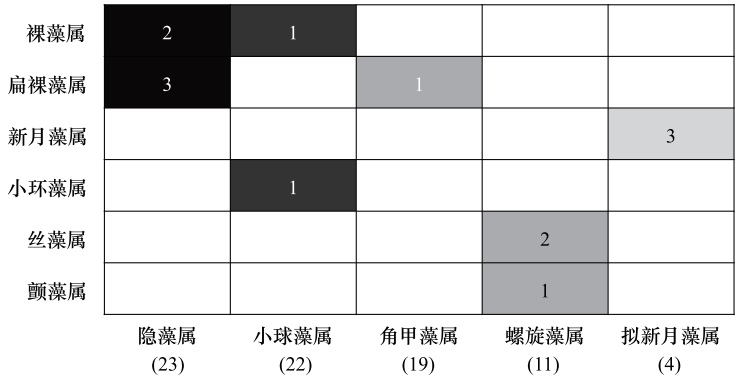

图 4　五个劣势浮游植物属类误识别结果

Fig.4　Misclassification of five disadvantaged phytoplankton genera

此等比例缩放输入图像是常见的数据前处理方法,然而这一前处理方法导致了藻细胞真实物理尺寸信息的丢失。在“螺旋藻属”测试的 11 个样本中,有 1 个样本被错误地预测为“颤藻属”、2 个被错误地预测为“丝藻属”。“角甲藻属”测试的 19 个样本中,仅有 1 个样本被错误地预测为“扁裸藻属”,对比观察其他类别的藻属可以发现,“角甲藻属”因其非常明显的一前二后的“甲鞘”特征使得其非常容易与其它藻属进行区分,导致了“角甲藻属”即使训练样本少,但模型的预测准确率非常高,如图 5 左侧所示。“小球藻属”测试的 22 个样本中,1 个被错误地预测为“裸藻属”、1 个被错误地预测为“小环藻属”,从形态上看,“裸藻属”和“小球藻属”的某些观测角度均呈现出类圆形的特征,与“小球藻属”相似性较高。类似的,“隐藻属”测试的 23 个样本中,2 个被错误地预测为“裸藻属”、3 个被错误地预测为“扁裸藻属”,如图 5 中右侧所示。观察发现,“隐藻属”与“裸藻属”和“扁裸藻”的形态特征具有较高的相似性,由于“隐藻属”属于劣势类别、训练样本较少,使得模型并不能很好地将“隐藻属”与另外两个属区分开来。综合而言,即使使用了“平方根采样法”缓解了类别不均衡问题,但模型预测仍然有对优势类别的倾向性;并且,由于深度学习中常见的等比缩放的数据前处理方法,使得模型易与错误地将劣势类别样本预测为形态相近的优势类别。

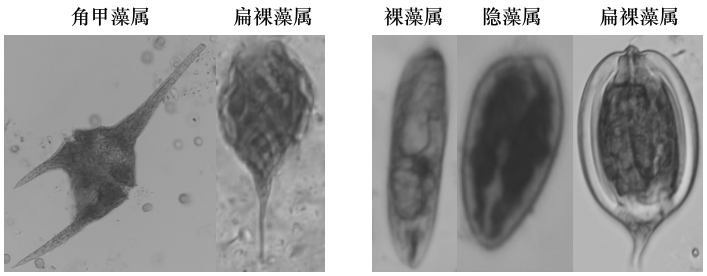

图 5　图像相似度对模型识别影响展示

Fig.5　Demonstration of the impact of image similarity on model recognition

4　结论与展望

本文首先研究发现了淡水湖泊中采集的浮游植物存在类别不平衡的特性,同时调研了国内外深度学习算法在浮游植物显微图像识别方面的应用,以及浮游植物类别不均衡问题可能造成的不良影响。在深度学习分类模型的训练中,常规的交叉熵损失函数难以处理此问题,导致模型对劣势类别的预测能力不足,从而使得现有基于深度学习的浮游植物分类方法难以准确识别实际水体中的劣势类别,增加了自动化鉴定仪器与专业人工镜检方法之间的误差。本文通过收集巢湖中常见的 29 个属类、18044 张图像,构建了一个具有严重类别不

平衡特性的浮游植物显微图像数据集;尝试了其他领域关于类别不平衡问题的多种解决方法和策略,探究这些方法在浮游植物显微图像领域的适用性。实验结果表明,重采样大类中的平方根采样法表现最佳,F1 值的微平均达到 0.932、宏平均达到 0.852,两项指标均优于其他方法。与标准方法相比,平方根采样法的 F1 值的微平均提升了 2.42%、宏平均提升了 5.19%,劣势类别的预测准确率提升明显,如拟新月藻属从 0 提升至 0.333、角甲藻属提升了 17.70%。综上所述,在其他领域中表现优异的方法中,重采样中的平方根采样法更加适用于具有严重类别不平衡特性的小型数据集,既改善了模型不易学习劣势类别的问题,也不会使得模型对优势类别过拟合;使用平方根采样法训练出的模型对不同浮游植物类别均有很好的区分能力,可以有效地应对自然水体中类别不平衡的问题,在保持或提高优势藻种预测准确率的同时,显著提高了模型对劣势藻种的分类能力。平方根采样法作为缓解数据集类别不均衡问题的有效方法,在未来可被拓展到其他领域,如浮游动物、底栖动物的自动化鉴定工作。虽然在本文的实验中重采样中的平方根采样法取得最好的训练结果,但由于采集的数据集具有地点单一、数量和类别不够多等局限性,无法说明其他方法无法缓解类别不均衡问题。同时,重采样方法与重权重方法分别作用于模型训练中的数据采样阶段和损失计算阶段,二者相对独立,因此将二类方法结合具有较高的可行性。在未来的研究中,更多采样地点、更大数据集、更多训练策略将会[32]为淡水湖泊中包含浮游植物在内的水生生物的自动化监测设备提供了模型训练的相关参考。

参考文献(References):

[1] Henley W J. The past, present and future of algal continuous cultures in basic research and commercial applications. Algal research, 2019, 43: 101636.

[2] 马建新,郑振虎,李云平. 莱州湾浮游植物分布特征. 海洋湖沼通报,2002(4): 63-67.

[3] 王洪铸,王海军,李艳. 湖泊富营养化治理: 集中控磷,或氮磷皆控. 水生生物学报,2020,44(5): 938-960.

[4] 严如玉,高桂青,杨军飞. 浮游藻类淡水生态环境评价应用现状. 人民珠江,2020,41(8): 111-116,138.

[5] 胡晓坤,赵越,孔凡洲. 流式影像仪在东海海域甲藻藻华研究中的应用. 海洋与湖沼,2022,53(2): 330-339.

[6] Garrido-Cardenas J A, Manzano-Agugliaro F, Acien-Fernandez F G. Microalgae research worldwide. Algal research, 2018, 35: 50-60.

[7] 邓杰航,何冬冬,卓家鸿. 复杂背景干扰下硅藻图像的深度网络识别与定位. 南方医科大学学报,2020,40(2): 183-189.

[8] 朱永正,张吉,程奇. 4 种深度学习图像分类算法在人工智能硅藻检验中的比较. 法医学杂志,2022,38(1): 31-39.

[9] Correa I, Drews P, Botelho S. Deep learning for microalgae classification//2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017: 20-25.

[10] Pant G, Yadav D, Gaur A. ResNeXt convolution neural network topology-based deep learning model for identification and classification of Pediastrum. Algal research, 2020, 48: 101932.

[11] 项和雨,邹斌,唐亮. 基于残差注意力网络模型的浮游植物识别. 生态学报,2021,41(17): 6883-6892.

[12] Promdaen S, Wattuya P, Sanevas N. Automated microalgae image classification. Procedia Computer Science, 2014, 29: 1981-1992.

[13] Giraldo-Zuluaga J H, Salazar A, Diez G. Automatic identification of Scene*desmus* polymorphic microalgae from microscopic images. Pattern Analysis and Applications, 2018, 21: 601-612.

[14] 王超,高越超,王沛芳. 广东长潭水库富营养化与浮游植物分布特征. 湖泊科学,2013,25(5): 749-755.

[15] 孙鑫,李兴,李建茹. 乌梁素海全季不同形态氮磷及浮游植物分布特征. 生态科学,2019,38(1): 64-70.

[16] 吴玉霖,傅月娜,张永山. 长江口海域浮游植物分布及其与径流的关系. 海洋与湖沼,2004,35(3): 246-251.

[17] 王云龙,袁骐,沈新强. 长江口及邻近海域夏季浮游植物分布现状与变化趋势. 海洋环境科学,2008,27(2): 169-172.

[18] 金海卫,徐汉祥,姚海富. 浙江沿岸夏季浮游植物分布特征. 浙江海洋学院学报自然科学版,2005,24(3): 231-235.

[19] Park J, Lee H, Park C Y. Algal morphological identification in watersheds for drinking water supply using neural architecture search for convolutional neural network. Water, 2019, 11(7): 1338.

[20] Li X, Liao R, Zhou J. Classification of morphologically similar algae and cyanobacteria using Mueller matrix imaging and convolutional neural networks. Applied optics, 2017, 56(23): 6520-6530.

[21] Zhang Y, Kang B, Hooi B. Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10795-10816.

[22] Chawla N V, Bowyer K W, Hall L O. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002, 16: 321-357.

[23] Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational intelligence, 2004, 20(1): 18-36.

[24] Huang C, Li Y, Loy C C. Learning deep representation for imbalanced classification//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5375-5384.

[25] Mahajan D, Girshick R, Ramanathan V. Exploring the limits of weakly supervised pretraining//European conference on computer vision. 2018: 181-196.

[26] Cui Y, Jia M, Lin T Y. Class-balanced loss based on effective number of samples//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277.

[27] Menon A K, Jayasumana S, Rawat A S. Long-tail learning via logit adjustment. arXiv: 2007.07314, 2020.

[28] Liu Z, Miao Z, Zhan X. Large-scale long-tailed recognition in an open world//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2537-2546.

[29] 王徐林, 张民, 殷进. 巢湖浮游藻类功能群的组成特性及其影响因素. 湖泊科学, 2018, 30(2): 431-440.

[30] He K, Zhang X, Ren S. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[31] Liu Z, Mao H, Wu C Y. A convnet for the 2020s//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[32] Barsanti L, Birindelli L, Gualtieri P. Water monitoring by means of digital microscopy identification and classification of microalgae. Environmental Science Processes & Impacts, 2021, 23(10): 1443-1457.