0254-6124/2024/44(2)-0241-10 *Chin. J. Space Sci.* 空间科学学报

ZHOU Jun, TONG Jizhou, LI Yunlong, FANG Shaofeng. Study of Sample Imbalance in Deep Learning Modeling of Solar Flare Forecasting (in Chinese). *Chinese Journal of Space Science*, 2024, **44**(2): 241–250. DOI:10.11728/cjss2024.02.2023–0028

太阳耀斑预报深度学习建模中样本不均衡研究*

周 俊[1,2] 佟继周[1,3] 李云龙[1,3] 方少峰[1,3]

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

3(国家空间科学数据中心 北京 101407)

摘 要 不同等级耀斑发生的频次存在数量级上的差别, 使基于常规卷积神经网络的耀斑预报模型通常难以捕捉 M 和 X 类耀斑先兆特征, 导致高等级耀斑预报精度低的问题. 本文对于这种耀斑预报中的长尾分布问题, 通过控制变量法讨论不同深度长尾学习方法对于耀斑预报精度提升. 尝试从训练集优化、损失函数优化、网络权重优化等角度改进模型对于 M 和 X 类耀斑的预报性能. 在 SDO/HMI 太阳磁图预报未来 24 h 耀斑的实验中, 相比于常规方法训练的基准模型, 改进模型在 M 和 X 类耀斑预报的精确率分别有了 53.10% 和 38.50% 的提升, 同时在召回率上有 64% 和 52% 的提升. 表明在耀斑预报问题中, 数据长尾分布的处理至关重要, 验证了深度长尾学习方法的有效性. 这种提升尾部类预报精确率的方法不仅可以应用于耀斑预报领域, 还可以迁移到其他具有长尾分布现象的空间天气典型事件的预报分析中.

关键词 耀斑预报, 长尾分布, 残差神经网络

中图分类号 P354

Study of Sample Imbalance in Deep Learning Modeling of Solar Flare Forecasting

ZHOU Jun[1,2] TONG Jizhou[1,3] LI Yunlong[1,3] FANG Shaofeng[1,3]

1(*National Space Science Center*, *Chinese Academy of Sciences*, *Beijing* 100190)

2(*University of Chinese Academy of Sciences*, *Beijing* 100049)

3(*National Space Science Data Center*, *Beijing* 101407)

Abstract Solar flares, as violent eruptions occurring in the lower atmosphere of the Sun, exert significant impacts on human activities. Researchers globally have developed multiple prediction models for solar flares, employing empirical, physical, statistical, and other methodologies. There is an order of magnitude difference in the occurrence of different classes of flares. This makes it difficult for traditional convolutional neural network-based flare prediction models to capture M, X class flare features, which leads to the problem of low precision of high level flare prediction. With the breakthrough of deep learning

* 国家重点研发计划项目 (2022YFF0711400) 和中国科学院网信专项 (CAS-WX2022SF-0103) 共同资助

2023-02-17 收到原稿, 2023-03-26 收到修定稿

©The Author(s) 2024. This is an open access article under the CC-BY 4.0 License (https://creativecommons.org/licenses/by/4.0/)

technology in recent years, it has shown strong potential in modelling and prediction of complex problems and a number of works have begun to try to use deep learning methods to construct flare prediction models. In this paper, different deep long-tail learning methods are discussed by us to improve the precision of flare forecasting by controlling the variables for the long-tail distribution phenomenon in flare forecasting. The forecast performance of the model for M and X flares is tried to be improved from the perspectives of training set optimization, loss function optimization and network weight optimization. The experiments on SDO/HMI solar magnetogram data show that the precision of M, X class flare prediction is significantly improved by 53.10% and 38.50%, respectively, and the recall is increased by 64% and 52% compared with the baseline model trained by conventional methods. It shows that the treatment of the long-tailed distribution of data is crucial in the flare forecasting problem, and verifies the effectiveness of the deep long-tailed learning method. This method of improving the precision of tail class forecasts can be applied not only to the field of flare forecasting, but also can be transferred to the analysis of forecasting other typical events of space weather with long-tailed distribution phenomenon.

Key words Flare prediction, Long-tailed distribution, Residual neural network

0 引言

太阳耀斑是发生在太阳低层大气的剧烈太阳爆发现象, 对人类活动具有重要影响. 据此, 基于经验、物理、统计等方法提出了多种太阳耀斑预报模型. 基于统计和经验的太阳耀斑预报主要是探寻预报因子与预报对象之间的统计关系, Bloomfield 等[1] 利用泊松统计模型估计太阳耀斑的发生概率, 基于物理方法的预报侧重耀斑爆发机制及其传播过程的研究, 从理论上获取耀斑爆发的内在因素. 随着早期机器学习的发展, 统计学习也大量应用于太阳耀斑预报建模, 例如 Yuan 等[2] 将有序逻辑回归模型与支持向量机分类器相结合, 利用总无符号磁通量、强梯度磁极性反转线的长度和总磁能耗散, 进行 24 h 内耀斑预报; Qahwaji 与 Colak[3] 利用 SVM 等方法建立了太阳黑子与耀斑的关联模型, 利用太阳黑子群的 McIntosh 分类和太阳周期数据预报未来 6 h 内是否有耀斑爆发. 除此之外, 还有贝叶斯网络方法、随机森林算法、集成学习等统计学习方法实现的耀斑预报模型[4–6].

近几年随着深度学习技术的突破, 其在复杂问题建模和预测等方面都展现出较强的潜力, 有不少工作开始尝试使用深度学习方法构建耀斑预报模型. 例如 Park 等[7] 将 GoogleNet 与 DenseNet 相融合, 构建了未来 24 h 内的有无耀斑发生预报模型; He 等[8] 通过 XGBoost 方法选取特征参量, 提出了一种基于 LSTM 的耀斑预报模型, 预报未来 48 h 内是否发生 M 级别以上耀斑事件; Guo 等[9] 基于元学习方法, 设计了生成式度量模型以提升 M 类和 X 类耀斑预报精确率; Huang 等[10] 基于卷积神经网络提出一种新的耀斑预报模型, 该模型可以对不同预测周期 (6, 12, 24, 48 h) 进行预报, 且性能相差不大; Zheng 等[11] 在以往研究的基础上, 提出了一个混合卷积神经网络, 将四分类模型细分为两个二分类模型, 针对 N, C, M, X 四类耀斑进行预报; Li 等[12] 将 VGGNet 和 AlexNet 结合, 进一步发展了是否发生 M 级以上耀斑预报模型; Wan 等[13] 针对类不均衡问题提出了三种方法, 分别对应数据集、损失函数和训练过程中的策略, 结果表明, 过采样和样本加权的策略在大多数参数上优于其他策略; Deng 等[14] 先使用 GAN 方法增强样本数据以平衡不同类, 提出一种混合卷积神经网络模型预测太阳耀斑爆发, 进一步提出了基于太阳上升期和下降期的耀斑预报模型, 分别建立两个模型对不同时期的耀斑进行预报; Deshmukh 等[15] 提出了一种解决太阳耀斑预报系统中过度误报问题的两阶段架构, 第一阶段基于 VGG-16 方法从磁图提取特征并预报耀斑, 将耀斑发生概率与所得特征混合后, 输入第二阶段超随机树模型中, 并引入新的评价指标以解决数据高度不平衡问题; Kaneda 等[16] 将注意力机制 (Transformer) 引入耀斑预报研究领域中, 通过磁图模块和太阳黑子特征模块处理图像和物理特征, 引入注意力机制来建模输入特征之间的时间关系, 最后取得了比人类专家预报更准的结果.

深度学习虽然能反映出比传统方法更好的效果, 例如 Deng 等[14] 提出的深度学习模型, 在各类耀斑预

报精确率方面均优于 Bloomfield 等[1] 提出的传统统计模型, 但深度学习在耀斑预报这类长尾分布数据上仍存在缺陷, 例如 Zheng[11] 的模型对 X 类耀斑预报精确率较高, 但召回率远低于其他类别; Li[12] 的模型由于训练样本中 M 和 X 类耀斑样本数量过少, 仅能对是否发生 M 级以上耀斑进行预测; Deng[14] 的模型在没有进行数据增强之前, X 类样本太少直接导致模型性能较低, 846 个 X 类样本中有 711 个样本被识别错误.

长尾分布 (或称类不均衡问题) 是指数据样本分布极不均衡, 少数类型的样本数量占据总样本数极高的比例 (称为头部类), 而其他类别的样本数仅占极小的比例 (称为尾部类). 在第 22, 23 和 24 太阳活动周的 M 与 X 级耀斑的总发生率分别为 10.33%, 6.77% 和 5.45%[17], C 级及以下耀斑发生率占 90% 左右, 是典型的长尾分布样本. 实际上已有部分研究在对耀斑数据进行建模的过程中尝试解决样本存在的长尾分布问题, 例如 Zheng 等[11] 试图通过欠采样和数据增强方法来缓解长尾分布问题; Deng 等[14] 将生成对抗网络 (GAN) 引入耀斑预报模型的训练, 通过 GAN 扩充少数类的样本以缓解长尾分布问题, 但以上方法对于缓解太阳耀斑预报中的长尾分布问题仍不明显.

本文利用 SDO/HMI 的 2010—2021 年所有太阳活动区磁图观测数据以及 GOES 卫星的太阳耀斑观测数据构建了具有 N, C, M, X 四种类型耀斑的标准数据集, 以 ResNet50 卷积神经网络作为耀斑预报的基线模型, 对基于耀斑长尾分布的预报模型进行探究, 通过引入训练集优化、损失函数优化、网络权重优化等深度长尾学习策略, 研究不同优化方法对于耀斑预报模型的提升效果, 最后通过各类方法的组合实验提出改进后的耀斑预报模型.

研究重点在于将长尾分布概念引入太阳耀斑预报领域, 并从数据端、网络端角度将深度长尾研究方法分为训练集优化、损失函数优化、网络权重优化. 在训练集优化部分引入渐进式采样方法以及多种数据增强方法; 损失函数优化部分引入 LDAM 损失、IB 损失; 网络权重优化部分引入可学习权重缩放算法等. 同时将几种方法进行组合训练以得到更优结果, 系统全面阐述了耀斑预报领域在面对长尾分布问题时的解决方案, 为后续研究者提供研究思路. 最后通过标准数据集与去除投影效应影响后的数据集所得结果对比, 分析投影效应对于耀斑预报模型整体准确度的影响.

1　方法

基于 ResNet50 网络训练耀斑预报基准模型, 分别通过训练集优化方法对训练集进行预处理、通过权重优化方法修改模型分类头、通过损失函数优化方法改变损失函数对长尾分布数据的敏感程度, 训练获得新的模型, 比较优化前后模型性能的提升. 进一步通过组合性能较优的改进方法探究组合方案对于耀斑预报的影响. 图 1 给出了主要实验流程.

1.1　训练集优化

训练集优化的核心思想是通过改变不同类别样本的分布情况, 以缓解类不均衡问题, 具体可分为类重平衡方法[18] 与数据增强方法两类.

类重平衡方法[18] 旨在从长尾分布的样本中通过采样方法获取各类别样本数趋于均衡的分布, 这种方法不改变样本图片特征分布, 具体包括类平衡采样、

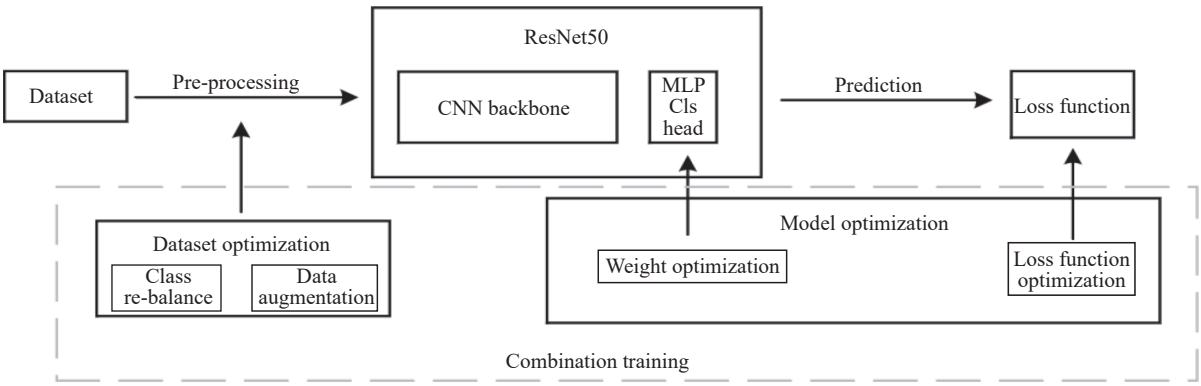

图 1　实验流程 (虚线框标出的是对建模过程的修改部分)

Fig. 1　Experimental flow chart (Dashed box marks the part of the modeling process that has been modified)

平方根采样和渐进式采样. 表 1 给出了本实验使用的类重平衡算法, 其核心公式中, $\tilde{n}_j$为第j类耀斑在经过类平衡后的样本数量; N为训练集样本总数; C为类别总数; 指数$q \in [0,1]$. 不同q对应不同的采样策略, 表 1 中$q=0,0.5,1$分别对应类平衡采样、平方根采样、实例采样. 渐进式采样公式中的T代表总的训练轮次, t代表网络当前训练轮次, p_j^{IB}代表实例采样时j类别的概率, p_j^{CB}代表类平衡采样时j类别的概率.

数据增强方法是通过算法合成新样本补充稀有类别的样本数量, 以缓解类不平衡问题, 这种方法会改变样本的特征分布. 表 2 给出了本文使用的数据增强算法. 其中 RSG 算法[19] 是通过计算多数类样本中与类别信息无关的特征位移, 将该特征位移与少数类样本结合, 形成新的少数类样本. Mixup 算法[20] 与 Manifold Mixup 算法[21] 均是通过对样本线性加权以缓解类不平衡问题. 区别在于前者在输入数据端进行修改, 后者在网络训练过程中进行修改. 其中, (X_i, Y_i)和(X_j, Y_j)代表从训练数据中随机抽取两个样本; $\boldsymbol{X}$代表图片数据矩阵值, 例如太阳磁图; Y代表该图片对应的标签, 比如耀斑类型; $\lambda \in [0,1]$是比例系数.

1.2 模型优化

在模型优化部分, 可分为损失函数优化方法和权重优化方法.

损失函数优化方法通过修改模型训练时使用的损失函数, 使其对于长尾分布中的尾部类更加敏感. 表 3 给出了本实验使用的两类损失函数的算法. 其中代价敏感学习[22] 方法中, $L(y, f(x,w))$为传统交叉熵损失函数; λ_k用于调节网络训练过程中对于不同类别样本的敏感程度; h为网络训练中隐藏层特征向量. LDAM 算法[23] 中, z_y为标签为y的样本经过模型后得到的第 y 类的输出值, z_j就是该样本得到的第 j 类的输出值, Δ_y为由样本标签分布控制的超参数.

权重优化方法[18] 是通过调整权重对于不同类别样本的敏感度以获得更优解. 表 4 给出了本实验采用的权重优化算法. 其中类重训练方法旨在固定表征学习阶段参数, 通过类平衡采样策略重新训练分类头; 最近邻分类器算法以 KNN 算法为基础, 在训练阶段, 计算出每类样本的特征中心, 在预测阶段, 通过余弦

表 1 类重平衡算法

Table 1 Class rebalancing algorithm

算法名称	简称	核心公式
实例采样(标准数据集)	Uniform	
类平衡采样	Balance	$\tilde{n}_j = N\dfrac{n_j^q}{\sum_{i=1}^{C} n_i^q},\ q=0,0.5,1$
平方根采样	Squreroot	
渐进式采样	Shift	$\tilde{n}_j = N\left[\left(1-\dfrac{t}{T}\right)p_j^{\mathrm{IB}} + \dfrac{t}{T}p_j^{\mathrm{CB}}\right]$

表 2 数据增强算法

Table 2 Data enhancement algorithms

算法名称	简称	核心公式
RSG算法	RSG	—
Mixup算法	Mixup	$\tilde{X} = \lambda X_i + (1-\lambda)X_j,\ \tilde{Y} = \lambda Y_i + (1-\lambda)Y_j$
Manifold Mixup算法	Manifold	$\tilde{X} = \lambda X_i + (1-\lambda)X_j,\ \tilde{Y} = \lambda Y_i + (1-\lambda)Y_j$

表 3 损失函数优化算法

Table 3 Loss function optimization algorithms

算法名称	简称	核心公式
代价敏感学习方法	IB Loss	$L_{\mathrm{IB}}(w) = \dfrac{1}{m}\sum_{(x,y)\in D_m} \lambda_k \dfrac{L(y, f(x,w))}{\lVert f(x,w)-y\rVert_1 \cdot \lVert h\rVert_1}$
LDAM算法	LDAM Loss	$L_{\mathrm{LDAM}}(y,z) = -\ln\dfrac{e^{z_y-\Delta_y}}{e^{z_y-\Delta_y}+\sum_{j\neq y} e^{z_j}}$

相似度或者欧几里得距离计算需要预测的样本到各类别中心点相似度或距离; τ正则化算法中的w_i代表网络权重, τ为超参数, 通过对权重进行正则化处理以调节网络对于不同类别样本的敏感程度; 可学习权重缩放算法则是在τ正则化算法的基础上, 将原本的超参数τ替换为函数f, 网络转而通过训练f得到最优解.

1.3 组合训练

在上述工作的基础上, 为进一步测试各类方法在组合训练后对各类耀斑预报精确率的提升效果, 将数据优化方法与模型优化方法进行组合. 由于计算资源限制, 在统计每种单独的优化方法后发现, 类重平衡方法整体优于数据增强方法, 因此将类重平衡方法中的实例采样、类重平衡采样、平方根采样、渐进式采样方法与权重优化方法中的类重训练算法、最近邻分类器算法、τ 正则化算法、可学习权重缩放算法进行两两组合训练, 寻找预报精确率优于单类优化方法的方案.

2 数据与评价指标

2.1 数据筛选

利用 SDO/HMI 发布的名为 hmi.SHARP_cea_720 s(以下简称 SHARP) 数据集, 选取 2010 年 5 月 1 日到 2021 年 12 月 31 日所有太阳活动区磁图观测数据以及 GOES 卫星的太阳耀斑观测数据构建 N, C, M, X 四类耀斑的标准数据集. 由于 SHARP 数据的采样间隔为 720 s, 短时间内太阳活动变化小使相邻图片相似度高, 为提升模型训练速度, 选择对样本按 1 h 等间隔采样.

原始数据中还存在较多观测区域不完整的图像, 通过统计所有图片中的缺失像素占比, 将缺失占比超过 10% 的数据判断为低质量图片, 予以舍去.

2.2 数据标注

为研究太阳耀斑对人类活动的影响, GOES 采用波长在 1Å到 8Å之间的软 X 射线辐射强度对太阳耀斑等级进行划分. 通常依据探测到的软 X 射线峰值流量将太阳耀斑分为 5 个等级, 分别为 A, B, C, M, X 级, 其释放的能量也随之递增, 每个等级对应的软 X 射线峰值流量范围详见表 5.

参考文献 [11] 中对于耀斑的分类方法, 将 A, B 级耀斑标注为 N 类, 等级为 C, M, X 的耀斑标签不变. 每一个磁图对照 SolarMonitor 网站耀斑事件列表判断在该磁图样本产生后的 24 h 内是否爆发了 C, M 或 X 耀斑, 如果存在, 则将该样本标记为对应类别, 否则将该样本标记为 N 类. 对于爆发多次不同级别耀斑的情况, 将最大级别作为该样本标签. 对于 SHARP 数据中存在的耀斑事件位置信息缺失情况, 将对应数据舍去不做标记. 一次高等级耀斑爆发后往往伴随着多次低等级耀斑爆发, 该标注方法可以有效避免低等级耀斑影响数据质量, 有效缓解了耀斑数据中固有的类不均衡问题, 增加了模型的准确度.

经过数据标注后, 共产生有标签样本 72734 张. 在划分训练集、验证集、测试集时, 随机划分方式会导致数据泄露、模型准确率虚高的问题, 本文以活动区为单位划分训练集、验证集、测试集, 不同活动区的磁图样本相似度低, 可以有效避免数据泄露问题. 按照 6:2:2 的形式划分数据集, 最终得到标准数据集列于表 6.

2.3 评价指标

为了全面评价模型结果, 采用精确率和召回率评价指标. 精确率用于形容各类耀斑预报的报准率, 即所有被预测为正类的结果中, 真正为正类样本所占比例, 其公式为

$$R_{\text{precision}} = N_{\text{TP}}/(N_{\text{TP}} + N_{\text{FP}}).$$

召回率又被称为查全率, 即正确预测为正类的样本占实际为正类样本的比例, 其公式为

表 4 权重优化算法统计

Table 4 Weight optimization algorithm statistics table

算法名称	简称	核心公式
类重训练方法	CRT	—
最近邻分类器方法	NCM	—
τ正则化算法	τ	$\tilde{w}_i = \frac{w_i}{\lVert w_i \rVert^{\tau}}$
可学习权重缩放算法	LWS	$\tilde{w}_i = f_i \cdot w_i$

表 5 太阳耀斑软 X 射线等级划分

Table 5 Solar flare soft X-ray class classification

太阳耀斑等级	软X射线峰值流量范围/(W·m^{-2})
A	$< 10^{-7}$
B	$10^{-7} \sim 10^{-6}$
C	$10^{-6} \sim 10^{-5}$
M	$10^{-5} \sim 10^{-4}$
X	$> 10^{-4}$

$$R_{\text{recall}} = N_{\text{TP}}/(N_{\text{FP}} + N_{\text{FN}}).$$

其中: N_{TP}为预测爆发某类耀斑, 实际也爆发了该类耀斑的数量; N_{FP}为预测爆发某类耀斑, 实际却没爆发了该类耀斑的数量; N_{FN}为预测没有爆发某类耀斑, 实际却爆发了该类耀斑的数量.

3 实验结果

基于 ResNet50 骨干网络, 通过训练集优化、模型优化、组合训练方法对标准数据集训练, 在测试集得到重要结果.

3.1 训练集优化实验

分析表 7 可知, 在选择实例采样, 也即使用原始样本进行训练时, M 类和 X 类的预报精确率为 0.00%, 而 N 类和 C 类预报精确率在 80% 以上, 表示当样本分布极度不均匀时, 网络学习存在过拟合现象, 也印证了长尾分布问题分析时提到的, 当数据存在长尾分布时, 模型学习的过程中倾向于将决策边界偏向头部类, 使模型在预报阶段倾向于将样本判定为多数类, 导致模型过拟合现象. 而在不同的采样方法下, N 类和 C 类 (头部类) 预报精确率虽然有所下降, 但是 M 类和 X 类 (尾部类) 预报精确率也有提升, 证明网络的过拟合现象有所缓解. 从表 7 分析, 渐进式采样在不同采样方法对比中, 其对于 X 类和 M 类耀斑报准率具有显著提升, 同时也能维持 N 类和 C 类耀斑报准率. 综合来看, 在类重平衡算法中, 渐进式采样方法的整体精确率更优.

数据增强算法相较于重采样算法而言, 改变了样本的原始特征分布. 从结果来看 (表 7 和表 8), 改变样本特征分布的方法不适用于耀斑预报研究. 其中, Mixup 方法旨在通过线性插值方式来构建新的训练样本和标签, 增加扰动, 以提升模型泛化能力, 抑制过拟合. 但在该实验中, 采用 Mixup 方法未能抑制由长尾分布带来的过拟合问题, 模型还是倾向于将样本分类为头部类. 而 Manifold 方法通过训练时对隐藏层的混合, 对于尾部类精确率有所提升, 但是同时降低了头部类的精确率, 特别是 C 类耀斑的预报精确率, 因此也不适用于耀斑预报领域.

3.2 损失函数优化方法实验

代价敏感学习算法与 LDAM 算法都是旨在通过

表 6 太阳耀斑预报训练、测试、验证数据集

Table 6 Solar flare forecast training, testing and validation dataset

数据集类型	No-flare	C-Class	M-Class	X-Class
训练集	22766	18478	2422	290
验证集	7532	6003	795	92
测试集	7493	5988	786	89

表 7 不同重采样方法下耀斑预报模型的精确率

Table 7 Precision of flare forecasting models under different resampling methods

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	87.10	83.50	0.00	0.00
类平衡采样	42.20	66.70	7.70	0.00
平方根采样	54.40	70.70	8.50	5.70
渐进式采样	58.20	64.40	18.20	15.30

表 8 不同数据增强方法下耀斑预报模型的精确率

Table 8 Precision of flare forecasting models with different data enhancement methods

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	87.10	83.50	0.00	0.00
RSG算法	54.80	48.20	0.00	0.00
Mixup	76.20	58.10	0.00	0.00
Manifold	42.30	22.90	26.70	17.60

调整损失函数来间接影响网络对于不同类别样本的灵敏度. 从结果上分析 (表 9), 两种方法对于 M 类和 X 类耀斑预报精确率均有提升, 但 LDAM 算法相较于代价敏感学习而言更加优秀, 其中 M 类耀斑预报精确率有了 18.30% 的提升, 但 N 类和 C 类耀斑预报精确率仍有下降.

3.3 网络权重优化方法实验

分析表 10 可知, 4 种权重优化方法对于 M 类耀斑的预报精确率均有较大提升, 但是对于 X 类耀斑的预报并无帮助, 同时对 No-flare 类以及 C 类耀斑的预报精确率均有所下降, 因此权重优化方法不能缓解长尾分布问题.

3.4 组合训练方法实验

将 4 种采样方法与 4 种权重优化方法两两组合, 送入 ResNet50 网络模型进行组合训练, 使用标准测试集所得结果列于表 11.

表 9 不同损失函数优化方法下耀斑预报模型的精确率

Table 9 Precision of flare forecasting models with different loss function optimization methods

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	87.10	83.50	0.00	0.00
代价敏感学习	57.90	42.10	15.50	4.30
LDAM算法	63.60	43.10	18.30	15.10

表 10 不同网络权重优化方法下耀斑预报模型的精确率

Table 10 Precision of flare forecasting models with different network weight optimization methods

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	87.10	83.50	0.00	0.00
分类器重训练算法	73.50	78.40	28.40	0.00
最近邻分类器算法	65.70	71.30	30.10	0.00
τ正则化算法	43.50	51.80	28.10	0.00
可学习权重缩放算法	63.10	73.10	26.30	0.00

表 11 不同组合训练方法下耀斑预报模型的精确率

Table 11 Precision of flare forecasting models with different combinations of training methods

Method		No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)		87.10	83.50	0.00	0.00
组合训练	Uniform+CRT	73.50	78.40	28.40	0.00
	Uniform+NCM	65.70	71.30	30.10	0.00
	Uniform+τ	43.50	51.80	28.10	0.00
	Uniform+LWS	63.10	73.10	26.30	0.00
	Balance+CRT	78.10	77.50	38.80	0.00
	Balance+NCM	73.20	75.80	56.50	0.00
	Balance+τ	75.30	78.20	33.10	0.00
	Balance+LWS	81.20	79.60	35.10	0.00
	Squareroot+CRT	42.80	71.50	50.20	2.50
	Squareroot+NCM	72.50	75.90	65.20	28.50
	Squareroot+τ	39.20	44.50	36.80	13.50
	Squareroot+LWS	73.80	75.20	30.80	15.60
	Shift+CRT	58.40	78.50	23.10	17.30
	Shift+NCM	78.20	56.80	35.90	18.30
	Shift+τ	58.10	38.50	33.70	16.50
	Shift+LWS	73.50	78.10	53.10	38.50

通过分析表 11 可知, 实例采样、类平衡采样与其他分类器学习方法结合的 8 个结果中, 可以看出均存在 X 类耀斑识别精度过低的问题. 其中, 实例采样不对原样本修改, 仅依赖分类器学习方法, 可以对 M 类耀斑的预报准确度有所提升, 但对 X 类耀斑预报的准确度却无法产生作用. 类平衡采样方法在与分类器学习方法结合后, 准确度相较于实例采样方法有所提升, 但是在 X 类耀斑预报问题上, 依然存在精度太低的问题. 在平方根采样、渐进式采样与其他分类器结合的 8 个结果中, 可以看到渐进式采样与可学习权重缩放算法的整体结果是最好的, N 类、C 类耀斑精确率分别为 73.50% 与 78.10%, M 类耀斑精确率为 53.10%, X 类耀斑精确率为 38.50%.

对于几类表现较优的优化方法, 统计了其召回率(表 12), 在渐进式采样和可学习权重缩放算法结合训练的算法中, M 类和 X 类耀斑的召回率高达 64% 和 52%, 相较于基准模型而言存在较大提升, 这也从多个角度说明了本工作所提方法的有效性, 同时由于高等级耀斑可能造成更大影响, 这种提升稀有类别预报的方法也具有一定的实用价值.

3.5 去除投影效应影响后实验

SDO/HMI 公开的 SHARP 数据集中采用 CEA 投影方法处理[15], 该处理方式可以缓解太阳两极附近的挤压效应, 但是对赤道边缘附近发生的垂直挤压效果不明显, 也就不免受到投影效应的影响. 一方面, 在实际预报中, 无法确保活动区发生在不受投影效应影响的方位, 这也是本文不对数据进行过滤处理以避免投影效应的原因. 另一方面, 从算法提升角度来看, 在模型训练之前也确有必要对数据进行筛选, 以避免低质量数据影响模型效果. 因此, 为尽可能减少投影效应对于模型训练的影响, 同时定量分析投影效果的影响大小. 基于标准数据集筛选了位于太阳中心±45°以内的磁图作为去除投影效应后数据集[11](表 13, 以下称去投影数据集). 将该数据集输入基准模型与前文中表现较优的几种模型中所得结果见表 14 和表 15.

分析表 14 和表 15 可知, 在基准模型中去投影数

表 12 几种表现较优方法召回率

Table 12 Recall rates for several better performing methods

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	83	75	0	0
Balance+LWS	77	71	46	0
Squareroot+NCM	72	73	75	38
Shift+LWS	76	69	64	52

表 13 去除阴影效应后太阳耀斑预报训练、测试、验证数据集

Table 13 Training, testing and validation datasets for solar flare forecasting after removal of shadow effects

数据集类型	No-flare	C-Class	M-Class	X-Class
训练集	16218	14735	2059	199
验证集	6518	4984	719	92
测试集	5676	4861	683	72

表 14 几种表现较优方法在去投影数据集上的精确率统计

Table 14 Precision statistics of several better-performing methods on the deprojected dataset

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	58.6	48.1	16.7	0
Balance+LWS	68.6	64.4	82.4	11.7
Squareroot+NCM	77.8	77.3	75.9	28
Shift+LWS	82.8	90.2	96.7	34.3

表 15　几种表现较优方法在去投影数据集上的召回率统计

Table 15　Recall statistics of several better-performing methods on the deprojected dataset

Method	No-flare/(%)	C-Class/(%)	M-Class/(%)	X-Class/(%)
实例采样(基准)	70.7	43.7	0.7	0
Balance+LWS	79.4	58.6	17.1	26.4
Squareroot+NCM	83.8	70	65.3	73.6
Shift+LWS	95.4	77.3	64.3	63.9

据集在 N 类和 C 类精确率与召回率方面有所下降, 而 M 类耀斑预报精确率有 16.7% 的提升. 在 Shift 与 LWS 组合训练方法中, N 类和 C 类预报精确率有 10% 左右的提升, 召回率方面更是有 10%～20% 的提升. M 类耀斑预报精确率高达 96.7%, 召回率也能达到 64.3%. 同时需要注意的是, X 类耀斑预报精确率下降了 4.2%, 但召回率上升了 11.9%. 其他方法在各类耀斑预报精确率及召回率上均有所提升. 综合来看, 去除投影效应的处理方法对耀斑预报整体精度有非常显著的提升, 对 N, C 和 M 类预报精确率和召回率有大幅提升, 而对 X 类耀斑预报精确率提升效果一般, 分析原因可能是在去除投影效应过程中, 其他类别存在大量数据位于太阳边缘地带, 去除大量干扰数据后模型对其识别精确率上升, 而 X 类耀斑受限于数据大小, 在其他类别有所提升的情况下反而有所下降. 综上, 去除投影效应的处理方法对于模型整体精度提升效果明显, 特别是在 N, C, M 类耀斑预报精确率及召回率上提升幅度较大, 因此在模型训练中应当考虑去除投影效应影响.

4　结论

通过比较本文研究中所有预报结果, 可以得到以下结论.

在 N 类与 C 类耀斑预报中, 如果不对样本或网络模型进行任何处理, 实例采样 (基准) 所得结果是最优的, 其中 N 类耀斑预报精确率为 87.10%, C 类耀斑预报精确率为 83.50%. 对样本或者网络模型进行处理的方法所得结果中, 类平衡采样与可学习权重缩放算法结合训练所得结果最优, 其中 N 类耀斑预报精确率为 81.20%, C 类耀斑预报精确率为 79.60%, M 类耀斑预报精确率为 35.10%. 对比结果显示, 在引入深度长尾学习算法后, 对 M 类耀斑预报精确率提升 35.10%, 但 N 类、C 类耀斑预报精确率存在 5% 左右下降.

在 M 类耀斑预报中, 平方根采样与最近邻分类器算法结合训练所得结果最优, 在 M 类耀斑预报精确率上的表现能达到 65.20%. 在 X 类耀斑预报中, 渐进式采样与可学习权重缩放算法结合训练所得结果表现最好, 对 X 类耀斑预报的精确率能达到 38.50%. 该方法在以低等级耀斑精确率下降 10% 左右的前提下, 对于 M 类及 X 类耀斑报准率有较大提升.

通过对实例采样所得结果与渐进式采样和可学习权重缩放算法结合训练所得结果进行对比后可以得出, 对于侧重 N, C 类耀斑预报的建模, 采用常规的深度学习建模即可. 对于侧重 M, X 类耀斑的建模, 则需要考虑长尾分布影响, 采用深度长尾学习方法以提高尾部类别的预报精确率.

通过去投影数据集与标准数据集在同一模型下的结果对比, 去除投影的处理方法对于模型整体精度提升效果明显, 特别是在 N, C, M 类耀斑预报精确率及召回率上有着大幅提升, 因此在耀斑预报研究中应当考虑去除投影效应影响, 以提升预报精度.

致谢　SDO 耀斑数据由联合科学行动中心 (JSOC) 提供, 郑艳芳与邓正两位老师为本研究提供建议. 国家科技资源共享服务平台–国家空间科学数据中心 (https://www.nssdc.ac.cn) 提供计算服务支持.

参考文献

[1] BLOOMFIELD D S, HIGGINS P A, MCATEER R T J, *et al.* Toward reliable benchmarking of solar flare forecasting methods[J]. *The Astrophysical Journal Letters*, 2012, **747**(2): L41

[2] YUAN Y, SHIH F Y, JING J, *et al.* Automated flare forecasting using a statistical learning technique[J]. *Research in Astronomy and Astrophysics*, 2010, **10**(8): 785-796

[3] QAHWAJI R, COLAK T. Automatic short-term solar flare prediction using machine learning and sunspot associ-

ations[J]. *Solar Physics*, 2007, **241**(1): 195-211

[4] YU D R, HUANG X, WANG H N, *et al.* Short-term solar flare level prediction using a Bayesian network approach [J]. *The Astrophysical Journal*, 2010, **710**(1): 869-877

[5] HAZRA S, SARDAR G, CHOWDHURY P. Distinguishing between flaring and nonflaring active regions[J]. *Astronomy & Astrophysics*, 2020, **639**: A44

[6] COLAK T, QAHWAJI R. Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares[J]. *Space Weather*, 2009, **7**(6): S06001

[7] PARK E, MOON Y J, SHIN S, *et al.* Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms[J]. *The Astrophysical Journal*, 2018, **869**(2): 91

[8] HE Xinran, ZHONG Qiuzhen, CUI Yanmei, *et al.* Solar flare short-term forecast model based on long and short-term memory neural network[J]. *Chinese Journal of Space Science*, 2022, **42**(5): 862-872 (何欣燃, 钟秋珍, 崔延美, 等. 基于长短期记忆神经网络的太阳耀斑短期预报 [J]. 空间科学学报, 2022, **42**(5): 862-872)

[9] GUO Dalei, ZHANG Zhen, ZHU Lingfeng, *et al.* Generative model-based of flare hierarchic recognition and forecast of extreme ultraviolet images in solar active region[J]. *Chinese Journal of Space Science*, 2023, **43**(1): 60-67. DOI: 10.11728/cjss2023.01.220214015 (郭大蕾, 张振, 朱凌锋, 等. 太阳活动区 EUV 图像的生成式模型耀斑分级与预报 [J]. 空间科学学报, 2023, **43**(1): 60-67. DOI: 10.11728/cjss2023.01.220214015)

[10] HUANG X, WANG H, XU L, *et al.* Deep learning based solar flare forecasting model. I. results for line-of-sight magnetograms[J]. *The Astrophysical Journal*, 2018, **856**(1): 7

[11] ZHENG Y F, LI X B, WANG X S. Solar flare prediction with the hybrid deep convolutional neural network[J]. *The Astrophysical Journal*, 2019, **885**(1): 73

[12] LI X B, ZHENG Y F, WANG X S, *et al.* Predicting solar flares using a novel deep convolutional neural network[J]. *The Astrophysical Journal*, 2020, **891**(1): 10

[13] WAN J, FU J F, LIU J F, *et al.* Class imbalance problem in short-term solar flare prediction[J]. *Research in Astronomy and Astrophysics*, 2021, **21**(9): 237

[14] DENG Z, WANG F, DENG H, *et al.* Fine-grained solar flare forecasting based on the hybrid convolutional neural networks[J]. *The Astrophysical Journal*, 2021, **922**(2): 232

[15] DESHMUKH V, FLYER N, VAN DER SANDE K, *et al.* Decreasing false-alarm rates in CNN-based solar flare prediction using SDO/HMI data[J]. *The Astrophysical Journal Supplement Series*, 2022, **260**(1): 9

[16] KANEDA K, WADA Y, IIDA T, *et al.* Flare transformer: solar flare prediction using magnetograms and sunspot physical features[C]//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer, 2022: 1488-1503

[17] WANG Ting. Statistical Analysis of Solar Flares During 22, 23 and 24 Solar Cycles[D]. Beijing: North China Electric Power University, 2021

[18] KANG B Y, XIE S N, ROHRBACH M, *et al.* Decoupling representation and classifier for long-tailed recognition [C]//8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview. net, 2020

[19] WANG J F, LUKASIEWICZ T, HU X L, *et al.* RSG: a simple but effective module for learning imbalanced datasets[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 3783-3792

[20] ZHANG H Y, CISSE M, DAUPHIN Y N, *et al.* mixup: Beyond empirical risk minimization[C]//6th International Conference on Learning Representations. Vancouver, BC, Canada: OpenReview. net, 2018: 1-13

[21] VERMA V, LAMB A, BECKHAM C, *et al.* Manifold mixup: Better representations by interpolating hidden states[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 6438-6447

[22] PARK S, LIM J, JEON Y, *et al.* Influence-balanced loss for imbalanced visual classification[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 715-724

[23] CAO K D, WEI C, GAIDON A, *et al.* Learning imbalanced datasets with label-distribution-aware margin loss [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2019: 140

作者简介

周　俊　男, 1999 年 6 月出生于湖南省岳阳市. 现为中国科学院国家空间科学中心硕士研究生, 主要研究方向空间大数据处理、基于深度学习的太阳耀斑预报.
E-mail: 1034051968@qq.com

佟继周（通信作者）　女, 1976 年 5 月出生于北京市. 现为中国科学院国家空间科学中心研究员, 主要研究方向为数据标准, 空间科学数据信息系统研发, 空间科学数据挖掘分析与应用等.
E-mail: tongjz@nssc.ac.cn