[文章编号] 1003－4684(2025)04-0030-07

一种子类不平衡的乳腺病理图像多分类方法

刘 敏[1,2,3]，何智子[1,2,3]，胡兰兰[1,2,3]，林 坤[1,2,3]

(1 湖北工业大学电气与电子工程学院，湖北 武汉 430068；
2 湖北工业大学太阳能高效利用及储能运行控制湖北省重点实验室，湖北 武汉 430068；
3 湖北工业大学新能源及电网装备安全监测湖北省工程研究中心，湖北 武汉 430068)

[摘 要] 针对乳腺病理图像数据集子类不平衡及子类差异细微的问题，提出了一种面向子类不平衡的乳腺病理图像多分类方法。数据层面，引入 LR-SMOTE 算法平衡样本数量；算法层面，首先为了改善子类差异细微，难以精细多分类的问题，融合位置注意力机制和通道注意力机制，建立了非局部均衡注意力模型，其次，为了进一步均衡子类不平衡，Focal loss 函数进行了优化改进。实验结果表明，所研究的方法具有较高分类精度，对不区分放大倍数的图像，准确度可达到 97.69%，召回率可达到 97.71%，对放大倍数为 40×、100×、200×和 400×的图像，比同类研究方法准确率均提高了 0.59%以上，召回率提高了 1.58%以上。

[关键词] 不平衡数据；均衡损失函数；注意力机制；乳腺癌分类
[中图分类号] TP301.6 [文献标识码] A

乳腺癌具有异质性，表现出不同的形态和生物学特征，使得乳腺癌患者具有不同的临床表现。为了给患者提供更加准确的诊断及预测肿瘤子类类型，需要进行乳腺癌病理图像精细多分类[1]。乳腺癌病理图像的精细多分类研究成为生物医学领域的研究重点[2]，多分类通常是乳腺癌的次级类别，癌细胞形态一致性很高，差异细微，因此乳腺癌病理学图像的精细多分类研究存在较大困难。

近几年来，随着深度学习技术的发展，一方面，许多学者使用深度卷积神经网络对乳腺病理图像进行了多分类研究，如文献[3-5]，但目前的探索还不足，且准确度不如二分类[6]；另一方面，乳腺病理图像的多分类中类不平衡问题也得到了广泛的关注[7]，研究者使用不同的方法克服该问题，如文献[8-9]，但还没有设计出有效的解决方案。类不平衡的数据集使深度学习模型对于少数类的样本特征学习不够充足，更容易导致错误分类。

针对子类不平衡问题，处理不平衡数据集的方法可以分为三类：数据级、算法级和混合技术。数据级方法以增加少数类样本数量或减少多数类样本数量为目的达到平衡，通常使用几何变换、随机过采样和随机欠采样，几何变换是对少数类样本进行旋转、翻转等操作，增加新的样本，如文献[3,10-11]。文献[9,12]使用随机过采样和随机欠采样方法减少不平衡的程度，使得样本较少的数据集与样本较多的数据集达到平衡。几何变换大量增加了少数类样本数量，但与此同时引入了噪声，大量的噪声数据使神经网络过拟合；而欠采样在减少样本数量时存在丢失图像重要特征的风险[13]。为了改善上述问题，出现了智能采样的方法，如 SMOTE[14] 及其变体，其中 LR-SMOTE[15] 算法能够生成的新样本分布在少数类样本中心，新样本特征更接近真实样本。

处理类不平衡的算法级方法通常使用损失函数，例如二值交叉熵损失函数、平衡交叉熵损失函数和 Focal Loss[16] 等，加强对少数群体的敏感性，减少对多数类别的偏重。然而交叉熵损失函数学习的特征零散，无法关注难分样本，Focal Loss[16] 提升了难分样本的关注度，但在多分类中难以区分不同子类，分类性能表现不佳。

混合技术就是将数据级和算法级方法用各种方式结合使用，集合了数据级和算法级方法的优点，如文献[17-18]。对于乳腺病理图像的精细多分类来说，各个类别间样本数量相差较大，使用混合技术，不仅避免了仅使用数据级的方法生成大量的噪声数据，还避免了单独使用算法级方法使得模型特征学习不足。

[收稿日期] 2023－03－03
[基金项目] 国家自然科学基金(61901165)
[第一作者] 刘 敏(1979－)，女，河南周口人，工学博士，湖北工业大学副教授，研究方向为图像处理、计算机视觉等。
[通信作者] 何智子(1995－)，女，湖北孝感人，湖北工业大学硕士研究生，研究方向为图像处理、计算机视觉。

针对乳腺病理图像子类差异细微以及子类不平衡的问题，本文使用混合技术，提出了一种精细多分类方法。数据级采用 LR-SMOTE[15] 算法增加少数类样本数量；算法级，结合通道注意力和位置注意力，建立了非局部均衡注意力模型（Resnet Non-Local Balance Selective Kernels Network，ResNls Net）网络结构，并引入优化的 Focal Loss[16] 函数结合 ArcFace loss[19] 函数，提出了一种新的均衡损失函数。为了验证本文方法的有效性，在大型公开数据集 BreakHis[20] 上对乳腺病理图像八分类开展了实验仿真研究。

1　方法

1.1　数据集均衡

训练集使用了 LR-SMOTE[15] 算法扩充样本，以 BreakHis 数据集为例，原始样本数量如图 1 所示。

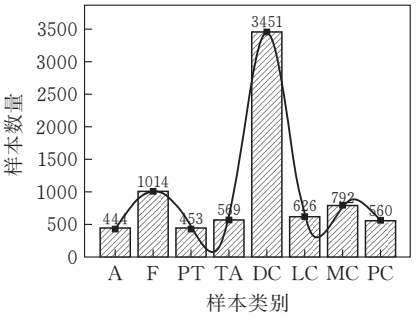


图 1　BreakHis 不平衡数据集描述

从图中看出样本数量最多的（导管癌）有 3451 个，样本数量最少的（腺病）只有 444 个，不平衡比率为 $r=7.773$，设置少数类生成的新样本数是原始数据的 3 倍，以达到辅助平衡的效果。先通过 SVM 算法将 BreakHis 数据集初步分类，将分类错误的少数类样本归为一个样本集，取每个被分错的样本的最近 k 个相邻样本，计算 k 个相邻样本的少数类样本数量 k'，删除 k' 为零的噪声样本。

对经过上述处理后的 BreakHis 数据集采用 k-means 计算样本中心 o，然后求出每个少数类样本到中心 o 的距离 f，f_{mean} 表示所有距离的平均值，n 表示计算 f_{mean} 和 f 的比值。

$$n=\frac{f_{\text{mean}}}{f} \tag{1}$$

$$x_{\text{new}}=x_i+\text{rand}(0,n)*(o_i-x_i) \tag{2}$$

式中 x_{new} 代表使用算法处理过后产生的样本，o_i 代表属于少数类别的样本中心，x_i 是少数类样本，rand$(0,n)$ 是随机生成一个 0 到 n 之间的数字。

实验中，选取 5 个邻近样本来确定未知样本类别，如图 2 所示。

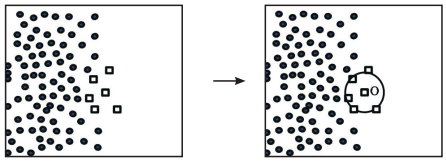


图 2　LR-SMOTE 算法选取样本

1.2　ResNls Net

1.2.1　网络结构　由于乳腺癌细胞结构复杂特征丰富且子类之间差异细微，病理图像多分类任务更需要提取全局精细准确的细粒度特征。而通道信息作为全局信息的重要部分，为了更好地获取全局精细特征，引入基于位置注意力的 Non-Local[21] 模块并结合通道注意力机制（selective-kernel-networks，SKNet），快速进行图像非局部细粒度特征提取。为了提升模型的训练速率，引入深度可分离卷积。在此基础上，使用空间金字塔池化固定特征向量，利于模型分析特征。本文研究的方法如图 3 所示。

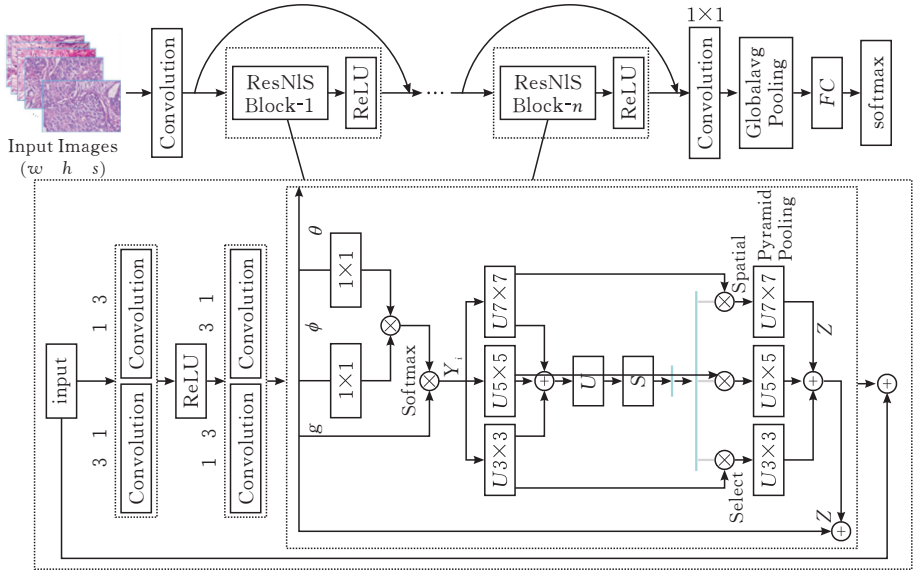
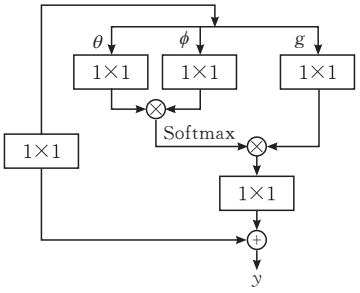
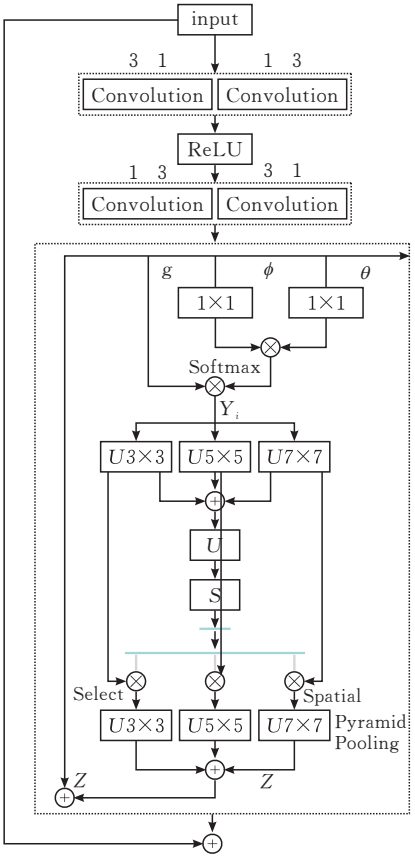


图 3　网络结构

Non-Local 模块结构如图 4a 所示，首先将输入通过矩阵乘法计算采样特征映射，然后将得到的特征映射计算为自相关矩阵，softmax 后得到注意力系数，最后将注意力系数与原始输入特征做残差运算输出结果[21]。ResNls 模块结构如图 4b 所示，原始输入经过深度可分离卷积进入位置注意力中计算长范围相关矩阵，接着进入通道注意力提取通道信息，最后与原始输入特征做残差运算输出结果。



(a) Non-Local 模块



(b) ResNls 模块

图 4 模型结构

将图 4bResNls 模块单独分析，该模块定义为：

$$Z_i = x_i + y_i \tag{3}$$

式中 $x = \{x_i\}, i = 1, 2, \cdots, n$，$x_i$ 表示 ResNls 模块的输入，而 y_i 表示 ResNls 模块的输出，ResNls 模块经过运算后并没有改变输出的维数，因此 x_i 和 y_i 具有相同的维数。网络让梯度进一步传播链接，将最终输出表示为：

$$\widetilde{x} = x_0 + z \tag{4}$$

式中 x_0 为初始输入，在每训练一部分图像后，损失函数计算模型预测值和实际值之间的误差，并得到一个损失值，用于通过反向传播更新网络的权重。ResNls 模块的梯度计算公式如下：

$$\frac{\partial \widetilde{x}}{\partial x_0} = \frac{\partial (x_0 + z)}{\partial x_0} = 1 + \frac{\partial z}{\partial x_0} \tag{5}$$

捷径连接保证了梯度在反向传播中始终大于或等于 1，从而避免了网络的梯度消失问题。

1.2.2 均衡损失函数 为了进一步优化乳腺病里图像子类不平衡及子类差异细微的问题，本文通过对均衡损失函数的 Focal loss 进行优化并结合 ArcFace loss 函数，提出了一种新的均衡损失函数，增强少数类样本权重的同时增大类间差异，用于乳腺病理图像精细多分类。

Focal loss 最早应用在目标检测中，其特点是平衡因子和聚焦参数能随着样本权重自适应调整，快速将模型集中在困难样本上，偏重于难分样本[16]。数学表达式如下：

$$Focal\ loss = (1 - P_t)^{\gamma} \log(P_t) \tag{6}$$

目标检测中，一张图片可以有多个标签，图像分为前景和背景，然而在乳腺病理图像多分类中，每张图像只有一个标签，并且不区分前景和背景。因此，为了更好地进行多分类，需要为每个类别设置权重，以提高模型从数量相对较少且难以区分的样本中学习的能力。优化的 Focal loss 函数表达式如下：

$$L_w = -W_c (1 - P_t)^{\gamma} (-X_c + \log(P_t)) \tag{7}$$

式中 P_t 是训练过程中计算的置信度，表示样本分类的难易程度，W_c 根据不同类调整权重。因乳腺病理图像中不同子类细胞结构复杂，变化多样，子类间形态极相似，癌细胞边界不清晰，使得不同子类难以区分，仅使用 Focal loss 函数无法满足乳腺病理图像精细多分类的要求。ArcFace loss 能够约束子类样本更接近中心，增强类内紧密性和增大类间差异[19]，但无法做到平衡各不同子类。本文叠加优化的 Focal loss 和 ArcFace loss 两个损失函数，二者的结合能更好地学习少数类别的特征同时增大类间距离，加强模型分类能力。新的均衡损失函数表达式如下：

$$L_{\text{total}} = -W_c (1 - P_t)^{\gamma} (-X_c + \log(P_t)) + \sum_{i}^{m} \log \frac{e^{s\cos(\theta i + \eta)}}{e^{s\cos(\theta_i + \eta)} + \sum_{j=1, j \neq yi}^{N} e^{s\cos\theta_j}} \tag{8}$$

$$\bar{\omega} = \frac{\omega}{\| \bar{\omega} \|}, x = \frac{x}{\| x \|} \tag{9}$$

$$\cos\theta_j = \bar{\omega}_j^T x_j \tag{10}$$

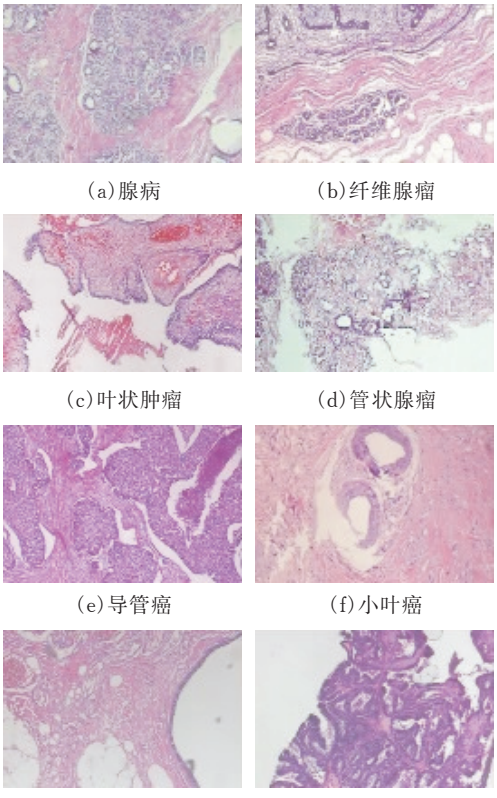
式(8)中 N 是分类数量，m 是批次内图片数量，η 表示每个类别之间的角度间距，θ 是 $\bar{\omega}$ 和 x 之间的

夹角。

2　实验结果及分析

2.1　实验设置与评价指标

本实验的实验环境基于 Python 编程语言,并使用了 PyTorch 深度学习框架,搭建了 ResNls Net 模型,实验任务是对乳腺癌病理图像数据集 BreakHis 进行多分类。BreakHis 数据集由巴西的联邦科技大学于 2014 年在临床研究中收集的匿名患者的乳腺病理图像,图像分为良性和恶性两个大类,其中良性又分为腺病、纤维腺瘤、叶状瘤和管状腺瘤四个子类,恶性分为导管癌、小叶癌、粘液癌和乳头状癌四个子类,一共分为八个子类,数据集一共有 40×、100×、200× 和 400× 四种放大倍数。考虑到实际中获取的图像倍数不是数据集中已有的倍数或获取的是未知倍数的图像,因此,模型分别在放大倍数为 40×、100×、200×、400× 的图像中进行实验,并且为了降低相同放大倍数图像对算法分类精度造成的影响,将四种放大倍数的图像混合成一个不区分放大倍数的混合数据集进行实验。图 5 列出了在 40× 放大倍数下的八种类别的样本。



(a)腺病　(b)纤维腺瘤　(c)叶状肿瘤　(d)管状腺瘤　(e)导管癌　(f)小叶癌　(g)粘液性癌　(h)乳头状癌

图 5　乳腺肿瘤子类样本图像

随意选取 BreakHis 数据集中的 72% 图像作为训练数据,随机选取 10% 的图像作为验证数据。从验证数据中随机选取 18% 的数据作为测试数据。实验中,判断分类结果所使用的衡量指标为准确度 ACC、召回率 RE、特异性 SP 和精密度 PR。计算公式如下:

$$\begin{cases} \mathrm{ACC}=\dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \\ \mathrm{RE}=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{SP}=\dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{PR}=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \end{cases} \tag{11}$$

以本实验为例,式(11)中 TP 和 TN 都表模型预测正确的样本数量,但 TP 表示模型预测为癌症的患者确实患有癌症,而 TN 表示模型预测没有患癌症的患者确实没有患癌症。FP 和 FN 表示预测错误的样本数量,分别表示实际没有患癌症的患者被预测为患癌症,患有癌症的患者却被预测为没有患癌症。

2.2　实验分析

2.2.1　消融实验　为了提高 ResNls Net 模型的训练效率,本次实验使用迁移学习,初始参数从 ImageNet 数据集上训练后提取,迁移到 ResNls Net 模型中,最终的实验结果都以测试集上的结果为准。本文对比了未使用损失函数的模型 ResNls Net (no L_{total}) 和增加本文改进损失函数的模型 ResNls Net 在不同放大倍数的实验效果(表 1)。

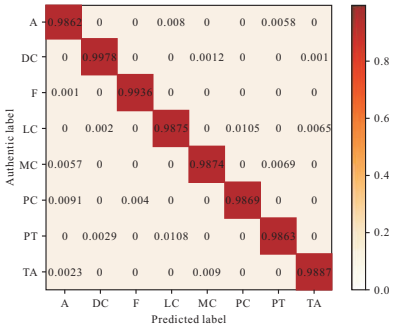
表 1　验证均衡损失函数　%

模型	40×			
	PR	RE	SP	ACC
ResNls Net (no L_{total})	96.37	96.40	95.87	96.42
ResNls Net	98.83	98.93	98.92	98.93
模型	100×			
	PR	RE	SP	ACC
ResNls Net (no L_{total})	96.40	96.39	96.36	96.43
ResNls Net	98.14	98.15	98.05	98.36
模型	200×			
	PR	RE	SP	ACC
ResNls Net (no L_{total})	96.82	96.86	96.85	96.89
ResNls Net	98.66	98.70	98.69	98.67
模型	400×			
	PR	RE	SP	ACC
ResNls Net (no L_{total})	96.51	96.39	96.52	96.50
ResNls Net	98.08	98.08	98.07	98.06
模型	不区分倍数			
	PR	RE	SP	ACC
ResNls Net (no L_{total})	92.06	93.32	91.89	94.51
ResNls Net	97.68	97.71	98.68	97.69

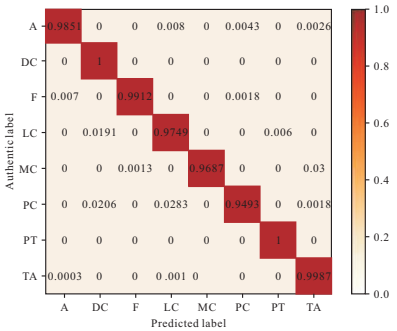
表 1 的实验数据显示,使用了本文的损失函数后,得到的测试准确度有明显提升,从实验结果看,ResNls Net 模型分类性能稳定,这是因为 ResNls Net 能够提取图像的精细特征,偏重少数类类别,增

大了类间差异，提升了模型的分类准确性。

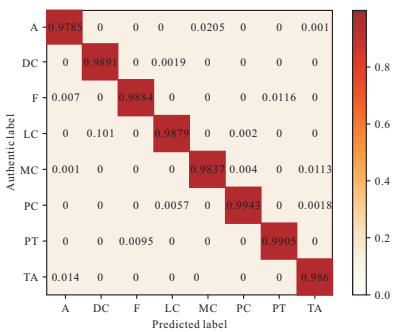
实验结果表明，ResNls Net 模型迭代 500 次时，模型已经达到了收敛，精度和损失基本停止变化，为了直观地观察模型在不同类别上的分类表现，将模型的分类情况通过混淆矩阵表现出来，如图 6 所示，观察到 ResNls Net 模型误分类样本较少，提高了对难分类样本的区分能力，并且该模型对各类别分类性能差异较小，证明该模型在解决数据集不平衡问题上效果显著。



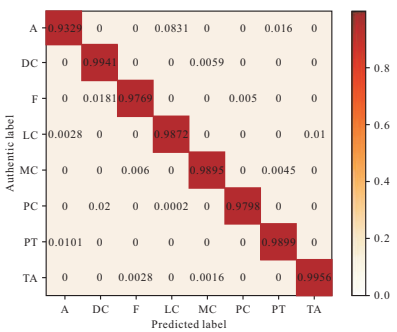
(a)40×



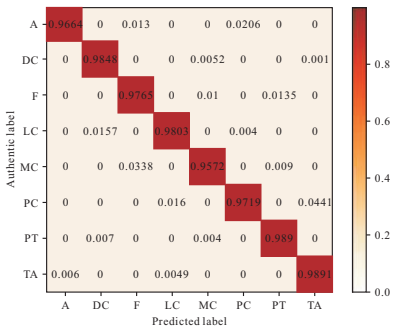
(b)100×



(c)200×



(d)400×



(e)不区分放大倍数

图 6 ResNls Net 模型在不同放大倍数上的混淆矩阵

2.2.2 对比实验 为了证明提出模型的有效性，将使用 BreaKHis 数据集的模型 GoogleNet＋SVM，AlexNet＋SVM，Inception_Resnet_v2，Csdcnn＋Aug[22]，Dblcnn ＋ MobileNetV2[23]，Efficient-Net[24]，Dnn（use of handcrafted approach as feature extractors and DNN classifiers ）[25]，Cnn＋Ensemble mode[3]和 ResNls Net 模型进行比较，为了排除训练方案不同带来的分类差异，搭建了 GoogleNet＋SVM，Inception_Resnet_v2，AlexNet＋SVM 模型并使用与该文模型相同的训练方案进行训练。实验结果如表 2 所示。根据实验数据观察到，在对比的所有模型中，ResNls Net 的 PR、RE、SP 和 ACC 均显著高于其他模型，在各种不同放大倍数和混合放大倍数的数据集上分类性能变化不大。

表 2 ResNls Net 与主流方法比较

模型	40 倍			
	PR	RE	SP	ACC
GoogleNet＋SVM	78.74	79.56	80.70	79.95
AlexNet＋SVM	81.16	82.14	83.48	82.16
Inception_Resnet_v2	94.00	93.85	93.71	93.97
Csdcnn＋Aug	—	—	—	94.12
Dblcnn＋MobileNetV2	—	—	—	96.58
Efficient－Net	97.47	96.01	—	96.99
Dnn	98	98	—	98
Cnn＋Ensemble	97.80	97.57	—	98.33
Ours	98.83	98.93	98.92	98.93

模型	100 倍			
	PR	RE	SP	ACC
GoogleNet＋SVM	81.15	80.42	80.45	80.17
AlexNet＋SVM	83.68	84.87	84.54	83.54
Inception_Resnet_v2	92.26	91.67	92.87	92.92
Csdcnn＋Aug	—	—	—	93.24
Dblcnn＋MobileNetV2	—	—	—	95.65
Efficient－Net	93.56	93.18	—	95.39
Dnn	98	98	—	98
Cnn＋Ensemble	98.58	96.98	—	97.12
Ours	98.14	98.15	98.05	98.36

续表 2　ResNls Net 与主流方法比较

模型	200 倍			
	PR	RE	SP	ACC
GoogleNet＋SVM	80.58	78.32	78.16	80.50
AlexNet＋SVM	83.93	83.55	84.83	83.40
Inception_Resnet_v2	92.21	92.18	91.09	91.23
Csdcnn＋Aug	—	—	—	94.73
Dblcnn＋MobileNetV2	—	—	—	96.43
Efficient－Net	91.42	91.63	—	94.71
Dnn	96	96	—	96
Cnn＋Ensemble	95.61	99.28	—	97.85
Ours	98.66	98.70	98.69	98.67
模型	400 倍			
	PR	RE	SP	ACC
GoogleNet＋SVM	78.04	75.53	77.37	79.98
AlexNet＋SVM	84.16	83.14	82.81	84.16
Inception_Resnet_v2	92.05	91.67	91.66	91.79
Csdcnn＋Aug	—	—	—	93.52
Dblcnn＋MobileNetV2	—	—	—	95.27
Efficient－Net	93.62	93.01	—	93.62
Dnn	97	96		96
Cnn＋Ensemble	97.54	96.49	—	96.15
Ours	98.08	98.08	98.07	98.06
模型	不分倍数			
	PR	RE	SP	ACC
GoogleNet＋SVM	73.68	72.46	74.63	74.96
AlexNet＋SVM	81.77	79.36	81.73	81.21
Inception_Resnet_v2	90.35	89.94	90.38	90.43
Csdcnn＋Aug	—	—	—	—
Dblcnn＋MobileNetV2	—	—	—	—
Efficient－Net	—	—	—	—
Dnn	—	—	—	—
Cnn＋Ensemble	—	—	—	—
Ours	97.68	97.71	98.68	97.69

与其他模型相比，ResNls Net 模型在混合放大倍数的数据集上性能没有明显的降低。表明其在任务中具有更强的分类能力和泛化能力。特别是 SP（特异性）达到 98.68%，说明 ResNls Net 模型对负类样本的识别能力非常强。

3　结论

针对数据集子类不平衡以及乳腺癌子类间差异细微的问题，提出了一种面向子类不平衡的乳腺病理图像多分类方法。本文研究的方法在数据层面使用 LR-SMOTE[15] 扩充数据；算法层面，融合了注意力机制基础上进一步提升速率，以达到提取乳腺病理图像丰富精细特征的目的，损失函数使用本文提出的新均衡损失函数，兼顾偏重少数类权重和增加类间差异两个点。通过实验验证，本文的方法有效改善了子类不平衡以及子类差异细微的问题，乳腺病理图像精细多分类的准确度高于其他同类模型。

[参 考 文 献]

[1] TSANG J, TSE G M. Molecular classification of breast cancer[J]. Advances in anatomic pathology, 2020, 27(01): 27-35.

[2] 李华，杨嘉能，刘凤，等. 基于深度学习的乳腺癌病理图像分类研究综述[J].计算机工程与应用，2020，56(13)：1-11.

[3] BARDOU D, ZHANG K, AHMAD S M. Classification of breast cancer based on histology images using convolutional neural networks[J]. Ieee Access, 2018, 6: 24680-24693.

[4] NAWAZ M, SEWISSY A A, SOLIMAN T H A. Multi-class breast cancer classification using deep learning convolutional neural network[J]. Int. J. Adv. Comput. Sci. Appl, 2018, 9(06): 316-332.

[5] YAN R, REN F, WANG Z, et al. Breast cancer histopathological image classification using a hybrid deep neural network[J]. Methods, 2020, 173: 52-60..

[6] BENHAMMOU Y, ACHCHAB B, HERRERA F, et al. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, Survey and Insights [J]. Neurocomputing, 2020, 375: 9-24.

[7] THABTAH F, HAMMOUD S, KAMALOV F, et al. Data imbalance in classification: Experimental evaluation[J]. Information Sciences, 2020, 513: 429-441.

[8] SAINI M, SUSAN S. Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-v3[C]// Iberian Conference on Pattern Recognition and Image Analysis. Springer, Cham, 2019: 409-420.

[9] ZHANG J, CHEN L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis [J]. Computer Assisted Surgery, 2019, 24(sup2): 62-72.

[10] ISMAEL S A A, MOHAMMED A, HEFNY H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks[J]. Artificial intelligence in medicine, 2020, 102: 101779.

[11] TOGACAR M, ÖZKURT K B, ERGEN B, et al. BreastNet: a novel convolutional neural network model through histopathological images for the diagnosis of breast cancer[J]. Physica A: Statistical Mechanics and its Applications, 2020, 545: 123592.

[12] SHARMA S, KUMAR S. The Xception model: A potential feature extractor in breast cancer histology images classification[J]. ICT Express, 2022, 8(01): 101-108.

[13] JOHNSAN J M, KHOSHGOFTAAR T M. Survey on deep learning with class imbalance[J]. Journal of Big Data, 2019, 6(01): 1-54.

[14] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: synthetic minority over-sampling technique. J Artif Int Res. 2002, 16(01): 321-57.
[15] LIANG X W, JIANG A P, LI T, et al. LR-SMOTE-An improved unbalanced data set oversampling based on K-means and SVM[J]. Knowledge-Based Systems, 2020, 196: 105845.
[16] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[17] LIU Z, CAO W, GAO Z, et al. Self-paced ensemble for highly imbalanced massive data classification[C]// 36th IEEE International Conference on Data Engineering. IEEE, 2020.
[18] LI J, XIN B, YANG Z, et al. Harmonization centered ensemble for small and highly imbalanced medical data classification[C]// 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021.
[19] DENG J, GUO J, YANG J, et al. ArcFace: Additive angular margin loss for deep face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 4690-4699.
[20] SPANHOLL F A, OLIVEIRA L S, PETTJEAN C, et al. A dataset for breast cancer histopathological image classification[J]. Ieee transactions on biomedical engineering, 2015, 63(07): 1455-1462.
[21] WANG X, GISHICK R, GUPTA A, et al. Non-local neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, America, June 18-22, 2018 : 7794-7803.
[22] HAN Z, WEI B, ZHENG Y, et al. Breast cancer multi-classification from histopathological images with structured deep learning model[J]. Scientific reports, 2017, 7(01): 1-10.
[23] WANG C, GONG W, CHENG J, et al. DBLCNN: Dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images[J]. Biomedical Signal Processing and Control, 2022, 73: 103451.
[24] ABDELSAMEA M M, ZIDAN U, SENOUSY Z, et al. A survey on artificial intelligence in histopathology image analysis[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(06): e1474.
[25] JOSEPH A A, ABDULLAHI M, JUNAIDU S B, et al. Improved multi-classification of breast cancer histopathological images using handcrafted features and deep neural network (dense layer)[J]. Intelligent Systems with Applications, 2022, 14: 200066.

A Multi-Classification Method for Breast Pathology Images with Subclass Imbalance

LIU Min[1,2,3], HE Zhizi[1,2,3], HU Lanlan[1,2,3], LIN Kun[1,2,3]

(1 *School of Electrical and Electronic Engineering, Hubei Univ. of Tech., Wuhan* 430068, *China*;
2 *Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei Univ. of Tech., Wuhan* 430068, *China*;
3 *Hubei Engineering Research Center for Safety Monitoring of New Energy and Power Grid Equipment, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: To address the problems of imbalanced subclasses and fine subclass differences in breast pathology image datasets, a multi-classification method for breast pathology images with imbalanced subclasses is proposed. At the data preprocessing stage, LR-SMOTE synthesizes additional samples for underrepresented classes, thereby reducing the impact of imbalanced distribution; At the algorithmic level, we first propose a non-local equilibrium attention framework that integrates position and channel attention to balance subclass representation, then enhance focal loss with adaptive probability weighting for robust feature discrimination. The experimental results show that the studied method has high classification accuracy, and the accuracy can reach 97.69% and the recall rate can reach 97.71% for images without distinguishing magnification, and for images with magnification of 40×, 100×, 200× and 400×, the accuracy of this paper's method is improved by more than 0.59% and the recall rate is improved by more than 1.58% than those of similar research methods.

Keywords: imbalanced data; equilibrium weighted loss function; attention mechanism; breast cancer Classification

[责任编校: 闫 品]