**Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss**

**会议：NIPS2019**

**代码：https://github.com/kaidic/LDAM-DRW**

**摘要**

当训练数据集存在严重的类不平衡问题时，深度学习算法的效果会很差，但是测试标准要求对频率较低的类进行良好的概括。 我们设计了两种新颖的方法来提高这种情况下的性能。 首先，我们提出了一种理论上原则上的标签分配感知余量（LDAM）损失，该损失是通过最小化基于余量的泛化约束来实现的。 该损失替代了训练过程中标准的交叉熵目标，并且可以与先前的训练策略结合使用，以进行类不平衡训练，例如重新加权或重新采样。 其次，我们提出了一个简单但有效的训练计划，该计划将重新加权推迟到初始阶段之后，使模型能够学习初始表示，同时避免了与重新加权或重新采样相关的一些复杂性。 我们在包括现实世界中不平衡的数据集iNaturalist 2018在内的多个基准视觉任务上测试了我们的方法。我们的实验表明，仅这些方法中的任何一种都已经可以对现有技术进行改进，并且它们的组合可以实现更好的性能提升。

**1引言**

现代现实世界中的大规模数据集通常具有长尾标签分布[Van Horn和Perona，2017； Krishna等，2017； Lin等，2014； Everingham等，2010； Guo等。 ，2016，Thomee等，2015，Liu等，2019]。 在这些数据集上，发现深度神经网络在较少表示的类上表现较差[He和Garcia，2008； Van Horn和Perona，2017； Buda等，2018]。 如果测试标准将重点放在少数群体上，则这尤其有害。 例如，均匀标签分布的准确性或所有类别之间的最低准确性就是此类标准的示例。 由于各种实际问题，例如到新域的可移植性，公平性等，这些是许多应用中的常见场景[Cao等人，2018; Merler等人，2019; Hinnefeld等人，2018]。

学习长尾数据的两种常见方法是重新加权示例损失并重新采样SGD迷你批次中的示例（请参见[Buda等人，2018； Huang等人，2016； Cui等 等人，2019年，He和Garcia，2008年； He和Ma，2013年，Chawla等人，2002年]以及其中的参考文献。他们都设计了预期接近测试分布的训练损失，因此可以在频繁类和少数类的准确性之间取得更好的权衡。但是，由于我们从根本上缺乏有关少数类别的信息和模型部署通常非常庞大，over-fitting少数类似乎是改进这些方法的挑战之一。

我们建议对少数的类别进行规范化，使其比频繁类别更为严格，以便在不牺牲模型拟合频繁类别的能力的情况下，改善少数群体的泛化误差。 要实现这一一般思想，需要使用数据依赖或标签依赖的正则化器（与标准的L2正则化不同，不仅取决于权重矩阵，还取决于标签）以区分频繁类别和少数类别。 对数据相关正则化器的理论理解很少（参见[Wei and Ma，2019; Nagarajan and Kolter，2019，Arora et al。，2018]）。

我们探索最简单，最容易理解的数据相关属性之一：训练示例的边距。 鼓励大幅度的边距可以看作是正则化，因为标准泛化误差范围（例如[Bartlett等，2017，Wei等，2018]）取决于所有示例中最小边距的倒数。 出于针对少数群体类别的泛化问题的动机，我们改为研究每个类别的最小边距，并获得每个类别和统一标签的测试错误范围。最小化所获得的范围可以在类别的边际之间实现最佳平衡 。 有关二进制分类情况的说明，请参见图1。

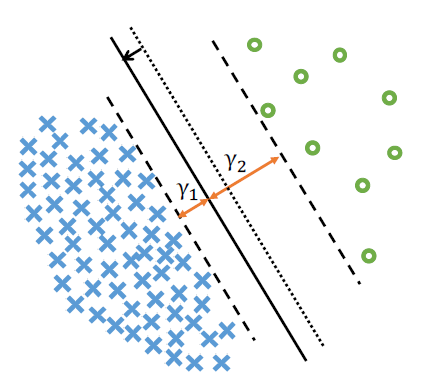


图1：对于具有线性可分离分类器的二进制分类，第i类的边距定义为第i类中的数据到决策边界的最小距离。 我们表明，具有均匀标签分布的测试错误的范围是。如此处所示，固定决策边界的方向会导致固定的，但要在之间进行权衡可以通过移动决策边界来优化。如第3.1节所述，最佳权衡是，其中ni是第i类的样本量。

受该理论的启发，我们设计了标签分配感知损失函数，该函数鼓励模型在每个类别的边距之间进行最佳权衡。 拟议的损失通过鼓励少数类获得更大的边距来扩展现有的软边距损失[Wang et al。，2018a]。 作为依赖标签的正则化技术，我们修改后的损失函数与重新加权和重新采样方法正交。 实际上，我们还设计了一个延迟的重新平衡优化程序，该程序使我们能够以更有效的方式将重新加权策略与损失（或其他损失）结合起来。

总而言之，我们的主要贡献是（i）设计标签分配感知损失函数，以鼓励少数类获得更大的边距；（ii）我们提出了一种简单的递延重新平衡优化程序，以更有效地应用重新加权；以及 （iii）我们的实际实施显示出在一些基准视觉任务上的显着改进，例如人为失衡的CIFAR和Tiny ImageNet [tin]，以及现实世界中的大规模失衡数据集iNaturalist'18 [Van Horn等，2018]。

**2相关工作**

大多数用于学习不平衡数据集的现有算法可以分为两类：重新采样和重新加权。

**重新采样**。 有两种类型的重采样技术：对少数类进行过度采样（例如，参见[Shen等人，2016； Zhong等人，2016； Buda等人，2018； Byrd和Lipton，2019]，以及对常见类别的采样不足（例如，参见[He and Garcia，2008; Japkowicz and Stephen，2002，Buda et al。，2018]以及其中的参考文献。）采样不足的缺点是它会丢弃大部分数据，因此在数据失衡非常严重时不可行。在很多情况下，过度采样是有效的，但可能导致少数群体的过度拟合[Chawla等，2002； Cui等，2019]。 针对少数群体的数据增强功能可以缓解过度拟合[Chawla等，2002； Zou等，2018]。

**重新加权**。 成本敏感型重加权为不同类别甚至不同样本分配（自适应）权重。 范本方案按频率的倒数成比例地对类别进行加权[Huang等，2016，2019，Wang等，2017]。 在极端的数据不平衡设置和大规模场景下，重新加权方法往往使深度模型的优化变得困难[Huang et al。，2016，2019]。 崔等 [2019]观察到，通过逆类频率重新加权会导致在频繁类上的表现不佳，因此建议通过有效样本数量的逆来重新加权。 这是我们经验比较的主要先验工作。

另一工作线根据每个样本的各自属性为每个样本分配权重。 焦点损失[Lin et al。，2017]降低了分类良好的示例的权重； Li等 [2018]提出了一种改进的技术，该技术可以对具有非常小的梯度或较大的梯度的示例进行加权，因为具有较小梯度的示例已得到很好的分类，而具有较大梯度的示例则具有异常值。

在最近的工作中[Byrd and Lipton，2019]，Byrd和Lipton研究了重要性加权的影响，并表明，在没有应用正则化的情况下，经验重要性加权没有显着影响，这与Soudry等人的理论预测相一致（2018），没有正则化的逻辑回归收敛到最大边际解。 在我们的工作中，我们明确鼓励稀有类具有更高的边距，因此我们不会收敛于最大边距解决方案。 此外，在我们的实验中，我们应用非平凡的L2正规化来实现最佳的泛化性能。 我们还发现，从培训开始，递延重加权（或递延重采样）比重加权和重采样更为有效。

相反，与上述这些论文正交，我们的主要技术旨在通过应用与重新加权方案正交的附加正则化来改善少数群体的泛化性。 我们还提出了一种延迟的重新平衡优化程序，以改进通用重新加权方案的优化和一般化。

边距损失。 hinge loss通常用于获得“最大余量”分类器，最明显的是在SVM中[Suykens和Vandewalle，1999]。 最近，已经提出了大边距Softmax [Liu等人，2016]，Angular Softmax [Liu等人，2017a]和加性余量Softmax [Wang等人，2018a]，以最大程度地减少类别内变化的预测和通过合并角度余量的概念来扩大类间余量。 与这些论文中与类别无关的边距相反，我们的方法鼓励为少数群体提供更大的边距。 [Li等，2002]和最近的工作[Khan等，2019，Li等，2019]也提出并研究了不平衡数据集的不均匀边界。 我们的理论通过为类的期望边际以及良好的经验进步提供具体的公式，将这一思想置于更理论的基础上。

分布鲁棒优化（DRO）是域自适应的另一种技术（请参见[Duchi等人，Hashimoto等人，2018，Carmon等人，2019]及其参考文献）。但是，该公式假设不了解目标标签分布超出了移位量的界限，这使问题变得非常具有挑战性。 在这里，我们假设您了解测试标签的分布，使用它们我们可以设计有效的方法，这些方法可以轻松地扩展到具有显着改进的大规模视觉数据集。

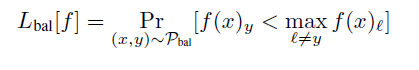
**元学习**。 元学习还用于提高不平衡数据集或小样本学习设置的性能。 我们向读者推荐[Wang等，2017，Shu等，2019，Wang等，2018b]及其中的参考文献。 到目前为止，我们通常认为，与基于元学习的方法相比，修改损失的方法在计算效率上更高。

**3主要方法**

**3.1理论动机**

**问题的设置和表示法**。 我们假设输入空间为，标签空间为。令x代表输入，y代表相应的标签。 我们假设在训练和测试时类条件分布P（x | y）是相同的。令Pj表示类别条件分布，即。我们将使用表示平衡的测试分布，该分布首先均匀地采样一个类，然后从Pj采样数据。

对于模型输出k对数，我们使用表示平衡数据分布上的标准0-1测试误差：



类似地，类别j的误差Lj定义为假设我们有一个训练数据集。令nj为类j中的示例数。 令表示对应于类j的示例索引。

将示例（x，y）的边距定义为

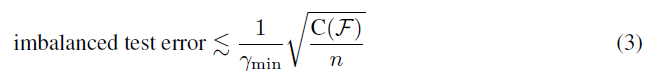


将j类的训练余量定义为：



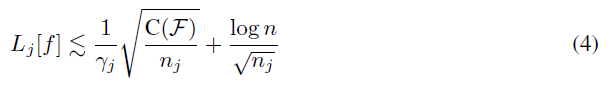
我们考虑可分离的情况（这意味着所有训练示例均已正确分类），因为神经网络经常被过度参数化并且可以很好地拟合训练数据。 我们还注意到，所有类别的最小边距是过去研究的经典训练余量概念[Koltchinskii等，2002]。

**细化的泛化误差范围**。 令F为假设类的族。 令C（F）是假设类F的某种适当的复杂性度量。最近有大量关于测量神经网络的复杂性的工作（请参阅[Bartlett等人，2017； Golowich等人，2017； Wei和 （Ma，2019]及其中的参考文献），并且我们下面的讨论与精确选择正交。当训练分布和测试分布相同时，典型的泛化误差范围在范围内。 也就是说，在我们的情况下，如果测试分布也与训练分布不平衡，则

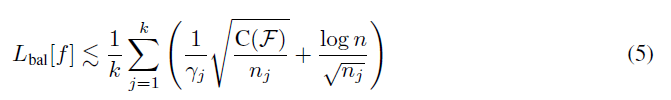


请注意，边界是标签分布所忽略的，并且仅涉及所有示例中的最小边距和数据点的总数。 通过考虑每个类的余量，我们将这种范围扩展到具有均衡测试分布的设置。 正如我们将看到的，下面的更细粒度的界限使我们能够设计针对不平衡数据集定制的新训练损失函数。

定理1（定理2的非正式和简化版本）。 在训练数据的随机性上具有很高的概率，类j的误差Lj由



我们在这里使用隐藏不变的因素。 直接的结果是

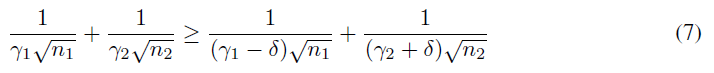


**类分配感知的边距权衡**。 每个类别的泛化误差界（4）表明，如果我们希望改善少数类别（具有较小nj的类别）的泛化，我们应该努力为他们争取更大的边距。但是，为少数群体提高边距可能会损害频繁类的边距。 类的边距之间的最佳权衡是什么？ 对于一般情况的答案可能很困难，但是幸运的是，我们可以获得针对二元分类问题的最佳折衷方案。

对于k = 2类，我们旨在优化（5）中提供的平衡泛化误差界限，该界限可以简化为（通过去除低阶项和公因子C（F））



乍一看，由于是权重矩阵的复杂函数，似乎很难理解最佳边距。但是，我们可以计算出之间的相对比例。假设最小化上面的方程，我们观察到，任何和（对于）都可以通过具有偏移偏项的相同权重矩阵实现（请参见图1）。因此，对于为最佳，他们应该满足



对于某些常数C。请参阅A节中的详细推导。

**快率与慢率，以及对边距选择的影响**。 定理1中的界限不一定要严格。 以缩放的泛化边界（或此处为不平衡类的）通常被称为“慢速”，而以1 / n缩放的泛化边界被称为“快速速率”。使用深度神经网络，并且当模型足够大时，可以将其中一些边界提高到快速速率。 有关一些最新进展，请参见[Wei and Ma，2019]。 在这些情况下，我们可以得出裕度的最佳折衷为。

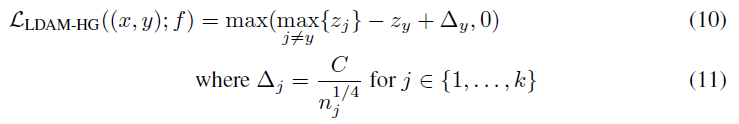
**3.2标签分配感知的边距损失**

受第3.1节中针对两个类别的类别边距之间的折衷的启发，我们建议对多种形式的下式强制实施与类别有关的边距

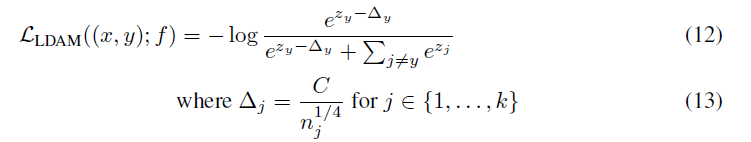


我们将设计一个软边距损失功能，以鼓励网络获得上述边距。 假设（x; y）为示例，f为模型。 为简单起见，我们使用表示第j个类的模型的第j个输出。

最自然的选择是hinge loss的多类扩展：



在此，C是要调整的超参数。 为了更轻松地调整边距，我们通过将最后一个隐藏激活正则化为L2范数1并将最后一个完全连接层的权向量正则化为L2范数1来有效地正则化logits（损失函数的输入），遵循先前的工作[Wang等，2018a]。 根据经验，hinge loss的非平滑性可能会给优化带来困难。 hinge loss的平滑松弛是具有强制余量的以下交叉熵损耗：



在之前的工作中（Liu等人，2016，2017a，Wang等人，2018a），训练集通常是平衡的，边距被选择为独立于标签的常数C，而我们的边距取决于标签 分配。

备注：细心的读者可能会发现损失有点让人想起重新加权，因为在二元分类的情况下-模型输出的是一个通过sigmoid转换成概率的单一实数，然后将这两种方法都改变了用标量因子的示例的梯度。 但是，我们注意到两个主要区别：重新加权引入的标量因子仅取决于类别，而引入的标量也取决于模型的输出； 对于多类分类问题，建议的损失以比仅引入标量因子更复杂的方式影响示例的梯度。 此外，最近的工作表明，在可分离的假设下，逻辑损失（具有弱正则化[Wei等人，2018]或不具有正则化[Soudry等人，2018]）提供了最大边际解，这反过来不受其定义的任何重新加权的影响。 正如我们在实验中所看到的，这进一步表明，损失和重新加权可能会相互补充。 （重新加权将影响不可分割数据情况下的边距，留待以后的工作。）

**3.3递延重新平衡优化计划**

成本敏感的重新加权和重新采样是应对不平衡数据集的两种众所周知且成功的策略，因为可以预期，它们有效地使不平衡训练分布更接近于统一的测试分布。 应用这些技术的已知问题是（a）如先前工作中指出的那样，当模型是一个深度神经网络时，对少数类别中的示例进行重新采样通常会导致对少数类别的过度拟合，以及（b）权衡少数群体的损失会导致优化过程中的困难和不稳定，尤其是当这些群体极度失衡时[Cui等，2019； Huang等，2016]。 实际上，崔等人 [2019]制定了新颖而复杂的学习率计划以应对优化难题。

我们从经验上观察到，在权重降低学习率之前，重新加权和重新采样均不如原始经验风险最小化（ERM）算法（所有训练示例的权重相同）。 在通过重新加权和重新采样使学习率降低之前产生的特征比ERM产生的特征更差。 （请参见图6，以对大型平衡数据集上的特征进行训练，对线性分类器进行训练，从而对特征质量进行消融研究。）

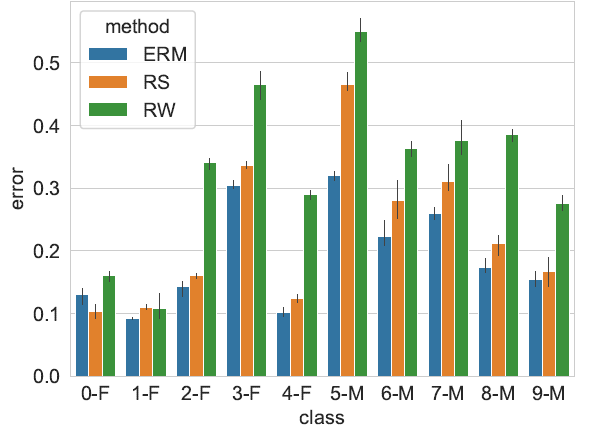
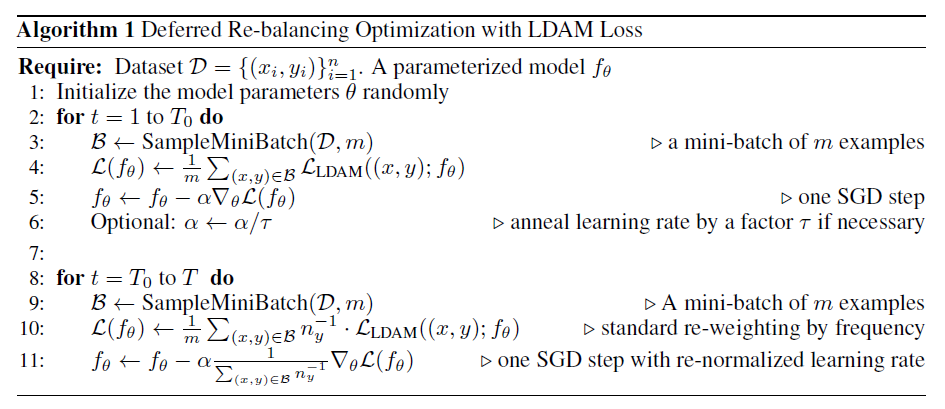


图6：在训练失调的CIFAR-10数据集中，步长不平衡= 100； = 0.5，为了在退火学习率之前测试ERM，RW和RS获得的特征的质量，我们使用平衡验证数据集的一个子集在特征之上训练线性分类器，并评估每个类其余验证数据上的验证错误（在训练线性分类器时几乎没有过度拟合的现象。）左5类很常见，并用-F表示。 从ERM设置获得的特征具有最强的性能，这证实了我们的直觉，即DRW的第二阶段始于更好的功能。 在第二阶段，DRW再次重新加权示例，调整决策边界并局部微调要素。

受此启发，我们开发了一种延迟的重新平衡训练程序（算法1），该程序首先使用具有LDAM损失的原始ERM进行训练，然后对学习率进行退火，然后部署具有较小学习率的重新加权LDAM损失。 根据经验，第一阶段的训练会导致第二阶段训练的良好初始化，并带有重新加权的损失。 由于损失是非凸的，并且第二阶段的学习率相对较小，因此第二阶段不会将权重移得太远。 有趣的是，借助我们的LDAM损失和递延的重新平衡培训，原始重新加权方案（通过按每个类中示例数量的倒数进行重新加权）与先前工作中引入的重新加权方案通用有效[Cui等，2019]。 我们还发现，通过我们的重新加权方案和LDAM，我们对早期停止的敏感性不及[Cui等人，2019]。



**4实验**

我们在人工创建的IMDB评论版本[Maas等，2011]，CIFAR-10，CIFAR-100 [Krizhevsky和Hinton，2009]和Tiny ImageNet [Russakovsky等，2015]上评估了我们提出的算法。 iNaturalist 2018 [Van Horn et al，2018]，它具有可控制程度的数据失衡以及现实世界中的大规模失衡数据集。 我们的核心算法是使用PyTorch开发的[Paszke et al，2017]。

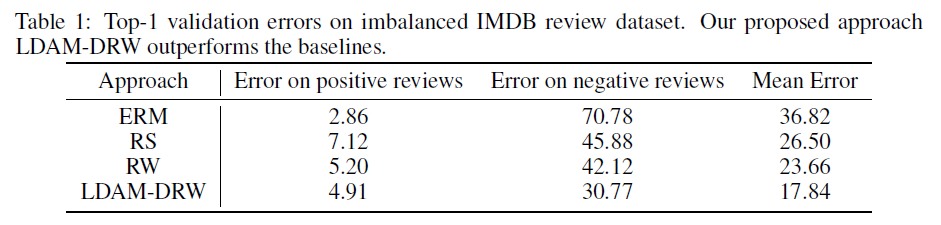
**基线**。 我们将我们的方法与标准训练和几种最先进的技术及其组合进行了比较，这些方法已被广泛采用以减轻不平衡数据集训练的问题：（1）经验风险最小化（ERM）损失：所有示例默认情况下具有相同的权重，我们使用标准的交叉熵损失。 （2）重新加权（RW）：我们按照类别样本大小的倒数对每个样本重新加权，然后重新规格化以使微型批次中的加权平均为1。 （3）重采样（RS）：每个示例都以与其类别的反样本大小成正比的概率进行采样。 （4）CB [Cui et al。，2019]：根据每个类别中有效样本数的倒数对示例进行加权或重采样，定义为， 而不是逆类频率。这个想法可以与重新加权或重新采样结合在一起。 （5）Focal：我们使用最近提出的焦点损失[Lin et al。，2017]作为另一个基准。 （6）SGD进度表：指的是标准进度表，其中学习率在某些步骤中衰减恒定因子； 我们使用标准的学习率衰减时间表。

**我们提出的算法和变体**。 我们测试了我们提出的以下技术的组合。 （1）DRW和DRS：遵循提出的训练算法1，我们使用标准的ERM优化计划，直到最后一个学习速率衰减，然后在第二阶段应用重新加权或重采样进行优化。 （2）LDAM：建议的标签分配感知边距损失，如3.2节所述。

当可以将这两种方法结合使用时，我们将首字母缩写词之间用短划线连接起来作为缩写。 我们提出的主要算法是LDAM-DRW。 请参阅B节以获取更多实施细节。

**4.1 IMDB评论数据集的实验结果**

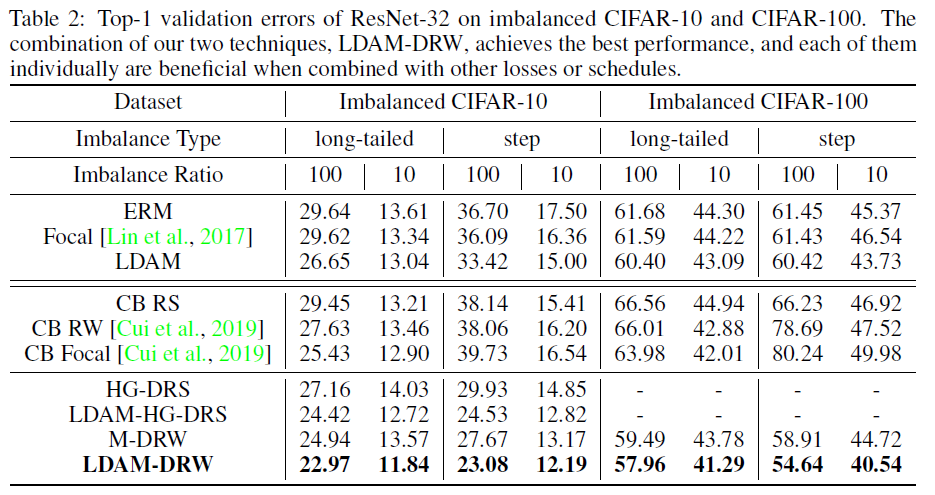
IMDB评论数据集由50,000个电影评论组成，用于二元情感分类[Maas等，2011]。 原始数据集包含平均数量的正面和负面评论。 我们通过删除90％的负面评论来手动创建不平衡的训练集。 我们使用Adam优化器训练了两层双向LSTM [Kingma and Ba，2014]。 结果记录在表1中。



**4.2 CIFAR的实验结果**

**不平衡CIFAR-10和CIFAR-100**。CIFAR-10和CIFAR-100的原始版本分别包含50,000个训练图像和10,000个验证图像，大小分别为10和100个类别，大小为32 32。 为了创建不平衡的版本，我们减少了每堂课的训练样本数量，并保持验证集不变。 为了确保我们的方法适用于各种环境，我们考虑了两种不平衡类型：长尾不平衡[Cui等人，2019]和步长不平衡[Buda等人，2018]。我们使用不平衡率来表示最频繁和最不频繁类别的样本大小之间的比率，即。长尾不平衡是由于不同类别的样本量呈指数衰减所致。 对于阶跃不平衡设置，所有少数类和所有常用类都具有相同的样本量。 这清楚地区分了少数族裔和频繁族，这对于消融研究特别有用。我们进一步将少数群体的比例定义为。 默认情况下，所有实验均设置为。

我们在表2中报告了CIFAR-10和CIFAR-100失衡版本的各种方法的前1个验证错误。我们建议的方法是LDAM-DRW，但是我们还将这两种技术的各种组合以及其他损失和培训包括在内 消融研究的时间表。

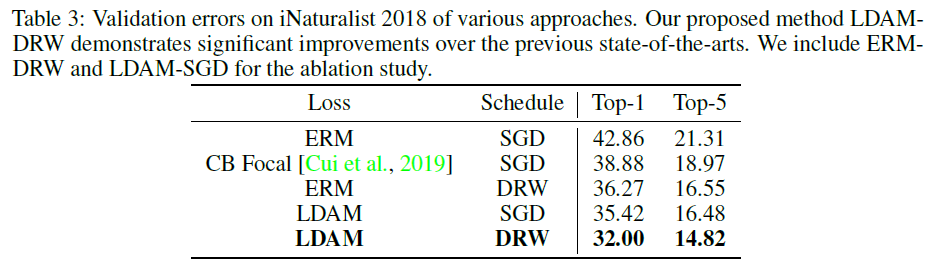


我们首先表明，提出的标签分配感知余量交叉熵损失优于纯交叉熵损失，并且它是为不平衡数据量身定制的变体之一，即焦点损失，而没有应用数据重新平衡学习计划。 我们还证明了我们的完整产品线大大优于以前的最新技术水平。 为了进一步证明拟议的LDAM损失是必不可少的，我们将其与交叉熵损失和铰链损失情况下所有类别的均匀余量进行正则化进行比较。 我们使用M-DRW表示算法，该算法使用具有均匀余量的交叉熵损失[Wang等，2018a]代替LDAM，即选择公式（13）中的作为一个不依赖于类j的调谐常数。 铰链损耗（HG）受到100个类别的优化问题的困扰，因此我们仅使用CIFAR-10限制其实验设置。

**不平衡但已知的测试标签分布**：在测试标签分布已知但不一致的情况下，我们还测试了算法扩展的性能。 有关详细信息，请参见第C.5节。

**4.3在iNaturalist 2018和不平衡的Tiny ImageNet上的视觉识别**

我们进一步验证了我们的方法在大规模不平衡数据集上的有效性。 iNatualist物种分类和检测数据集[Van Horn等，2018]是现实世界中的大规模不平衡数据集，在其2018版本中，它具有437,513个训练图像，总共8,142个类。我们为实验采用官方培训和验证拆分。 训练数据集具有长尾标签分布，并且验证集设计为具有平衡的标签分布。我们将ResNet-50用作iNaturalist 2018所有实验的骨干网络。表3总结了iNaturalist 2018的top-1验证错误。值得注意的是，我们的完整产品线能够在top-1错误率表现中比ERM基准高出10.86％和以前的最高水平高出6.88％。 有关不平衡的Tiny ImageNet的结果，请参阅附录C.2。



4.4消融研究

评估少数类的概括。为了更好地理解我们算法的改进，我们在不平衡的CIFAR-10上显示了图2中不同方法的每类错误。请参阅标题以进行讨论。

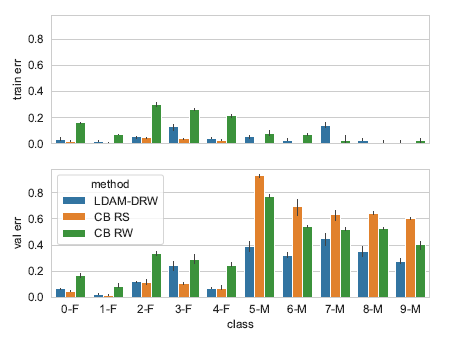


图2：CIFAR-10上每级的top-1错误，具有阶跃不平衡（= 100; = 0.5）。 0-F到4-F班是普通类，其余的是少数类。 在这种极度不平衡的情况下，RW在少数情况下会出现拟合不足的情况，而RS则会过度拟合。 相反，所提出的算法在少数类上表现出了很好的概括性，而在频繁类上的性能几乎不受影响。 这表明我们成功地更加规范了少数群体。

**评估递延的重新平衡计划。** 我们在图3中比较了递归再平衡计划的学习曲线和其他基线。在C.3节的图6中，我们进一步表明，即使第一阶段的ERM与RW和RS相比，平衡测试误差稍差或相当， 实际上，ERM学习到的功能（最后一层激活）比RW和RS更好。 这与我们的直觉相吻合，即DRW的第二阶段从更好的功能开始，调整决策边界并局部微调功能。

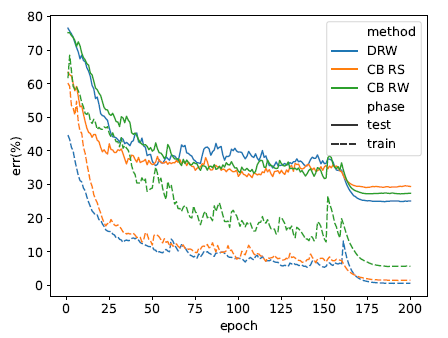


图3：带有长尾失衡（= 100）的CIFAR-10上的失衡训练误差（虚线）和平衡测试误差（实线）。对于所有算法，我们在步骤160退火衰减学习速率。 我们的DRW日程安排在降低学习率之前使用ERM，因此，如预期的那样，在此之前，ERW的表现要比RW和RS差。 但是，在降低学习速度后，它的性能明显优于其他方法。 有关更多分析，请参见第4.4节。

**5结论**

我们提出了两种在不平衡数据集上进行训练的方法：标签分配感知的边际损失（LDAM）和递延重加权（DRW）训练时间表。 我们的方法在各种基准视觉任务上实现了显着改善的性能。 此外，我们通过显示LDAM优化了统一标签泛化误差范围来提供LDAM的理论原理。 对于DRW，我们认为推迟重新加权可以使模型避免与重新加权或重新采样相关的弊端，直到模型学习到良好的初始表示之后（请参见图3和图6中的一些分析）。 但是，对于DRW成功的确切解释在理论上尚不完全清楚，我们将其作为未来工作的方向。