软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):1587−1600 [doi: 10.13328/j.cnki.jos.007014]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多样真实任务生成的鲁棒小样本分类方法*

刘 鑫[1,2], 景丽萍[1,2], 于 剑[1,2]

[1](交通数据分析与挖掘北京市重点实验室(北京交通大学), 北京 100044)

[2](北京交通大学 计算机与信息技术学院, 北京 100044)

通信作者: 景丽萍, E-mail: lpjing@bjtu.edu.cn

摘 要: 随着大数据、计算机与互联网等技术的不断进步, 以机器学习和深度学习为代表的人工智能技术取得了巨大成功, 尤其是最近不断涌现的各种大模型, 极大地加速了人工智能技术在各个领域的应用. 但这些技术的成功离不开海量训练数据和充足的计算资源, 大大限制了这些方法在一些数据或计算资源匮乏领域的应用. 因此, 如何利用少量样本进行学习, 也就是小样本学习成为以人工智能技术引领新一轮产业变革中一个十分重要的研究问题. 小样本学习中最常用的方法是基于元学习的方法, 这类方法通过在一系列相似的训练任务上学习解决这类任务的元知识, 在新的测试任务上利用元知识可以进行快速学习. 虽然这类方法在小样本分类任务上取得了不错的效果, 但是这类方法的一个潜在假设是训练任务和测试任务来自同一分布. 这意味着训练任务需要足够多才能使模型学到的元知识泛化到不断变化的测试任务中. 但是在一些真正数据匮乏的应用场景, 训练任务的数量也是难以保证的. 为此, 提出一种基于多样真实任务生成的鲁棒小样本分类方法(DATG). 该方法通过对已有少量任务进行 Mixup, 可以生成更多的训练任务帮助模型进行学习. 通过约束生成任务的多样性和真实性, 该方法可以有效提高小样本分类方法的泛化性. 具体来说, 先对训练集中的基类进行聚类得到不同的簇, 然后从不同的簇中选取任务进行 Mixup 以增加生成任务的多样性. 此外, 簇间任务 Mixup 策略可以减轻学习到与类别高度相关的伪判别特征. 同时, 为了避免生成的任务与真实分布太偏离, 误导模型学习, 通过最小化生成任务与真实任务之间的最大均值差异(MMD)来保证生成任务的真实性. 最后, 从理论上分析了为什么基于簇间任务 Mixup 的策略可以提高模型的泛化性能. 多个数据集上的实验结果进一步证明了所提出的基于多样性和真实性任务扩充方法的有效性.

关键词: 小样本学习; 元学习; 任务 Mixup; 多样性; 真实性

中图法分类号: TP18

中文引用格式: 刘鑫, 景丽萍, 于剑. 基于多样真实任务生成的鲁棒小样本分类方法. 软件学报, 2024, 35(4): 1587−1600. http://www.jos.org.cn/1000-9825/7014.htm

英文引用格式: Liu X, Jing LP, Yu J. Diverse and Authentic Task Generation Method for Robust Few-shot Classification. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1587−1600 (in Chinese). http://www.jos.org.cn/1000-9825/7014.htm

Diverse and Authentic Task Generation Method for Robust Few-shot Classification

LIU Xin[1,2], JING Li-Ping[1,2], YU Jian[1,2]

[1](Beijing Key Lab of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

[2](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

* 基金项目: 中央高校基本科研业务费(2019JBZ110); 北京市自然基金(L211016); 国家自然科学基金(62176020); 国家重点研发计划(2020AAA0106800)

本文由“绿色低碳机器学习研究与应用”专题特约编辑封举富教授、俞扬教授、刘淇教授推荐.

收稿时间: 2023-05-15; 修改时间: 2023-07-07; 采用时间: 2023-08-24; jos 在线出版时间: 2023-09-11

CNKI 网络首发时间: 2023-11-24

Abstract: With the development of technologies such as big data, computing, and the Internet, artificial intelligence techniques represented by machine learning and deep learning have achieved tremendous success. Particularly, the emergence of various large-scale models has greatly accelerated the application of artificial intelligence in various fields. However, the success of these techniques heavily relies on massive training data and abundant computing resources, which significantly limits their application in data or resource-scarce domains. Therefore, how to learn from limited samples, known as few-shot learning, has become a crucial research problem in the new wave of industrial transformation led by artificial intelligence. The most commonly used approach in few-shot learning is based on meta-learning. Such methods learn meta-knowledge for solving similar tasks by training on a series of related training tasks, which enables fast learning on new testing tasks using the acquired meta-knowledge. Although these methods have achieved sound results in few-shot classification tasks, they assume that the training and testing tasks come from the same distribution. This implies that a sufficient number of training tasks are required for the model to generalize the learned meta-knowledge to continuously changing testing tasks. However, in some real-world scenarios with truly limited data, ensuring an adequate number of training tasks is challenging. To address this issue, this study proposes a robust few-shot classification method based on diverse and authentic task generation (DATG). The method generates additional training tasks by applying Mixup to a small number of existing tasks, aiding the model in learning. By constraining the diversity and authenticity of the generated tasks, this method effectively improves the generalization of few-shot classification methods. Specifically, the base classes in the training set are firstly clustered to obtain different clusters and then tasks are selected from different clusters for Mixup to increase task diversity. Furthermore, performing inter-cluster tasks Mixup helps alleviate the learning of pseudo-discriminative features highly correlated with the categories. To ensure that the generated tasks do not deviate too much from the real distribution and mislead the model's learning, the maximum mean discrepancy (MMD) between the generated tasks and real tasks is minimized, thus ensuring the authenticity of the generated tasks. Finally, it is theoretically analyzed why the inter-cluster task Mixup strategy can improve the model's generalization performance. Experimental results on multiple datasets further demonstrate the effectiveness of the proposed method.

Key words: few-shot learning; meta-learning; task Mixup; diversity; authenticity

随着大数据、计算机、互联网等信息技术的不断进步, 以机器学习和深度学习为代表的人工智能技术得到了飞速发展, 在诸如图像分类、人机对弈、语音识别、知识问答、无人驾驶等应用场景取得了重大进展. 比如, 残差网络 ResNet[1]在 ImageNet[2]数据集上的分类精度已经超过人类, 阿尔法狗(AlphaGo)[3]在围棋比赛中战胜人类冠军, 自然语言领域的大模型 ChatGPT[4]可以像人类一样聊天交流. 然而, 目前深度学习算法的成功应用离不开海量训练数据和强大算力的支撑. 比如, ImageNet 的训练数据包含了 1 400 万张图片, AlphaGo 学习了 6 000 万盘棋局, GPT-3 的训练数据高达 45TB, 需要数千个高端 GPU 同时进行训练, 这大大限制了深度学习模型在一些领域的应用. 比如: 由于专业性或者安全性问题, 医疗或军事领域的数据通常很难获得或者标注成本很高, 创建该领域的大规模训练数据集是十分困难的. 因此, 如何在只有少量有标签训练样本的情况下进行学习, 也就是小样本学习是以人工智能技术引领的新一轮产业变革中一个十分重要的研究问题.

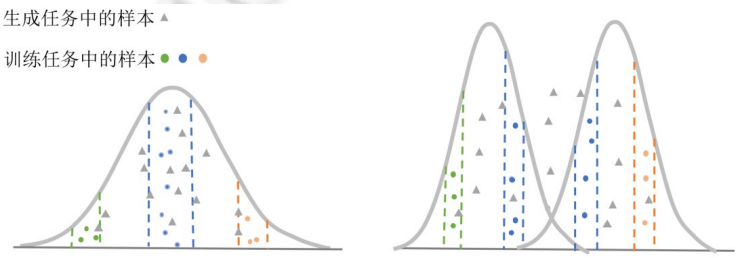
基于元学习的方法是小样本学习中主流的方法[5], 这类方法通过在大量类似的小样本任务上学习解决这类任务的元知识, 利用这些元知识帮助目标小样本任务进行快速学习. 比如: 代表性工作 MAML[6]希望学习的元知识是一个可以适用于不同小样本任务的初始点, 在新任务上只需要几步就可以达到最优; 原型网络[7]希望学习一个可以适用于不同任务的度量空间, 在该空间中, 通过比较查询样本与每个类原型之间的距离进行分类. 沿袭这两种思路, 一系列基于元学习优化[8−16]和基于度量[17−21]的小样本学习方法被提了出来. 虽然这些方法在小样本分类任务上取得了不错的效果, 但这些方法隐含着一个假设: 训练任务和测试任务的分布是一致的. 这就要求我们的训练任务数量足够多, 可以有效地代表整体任务的分布. 但事实上, 由于某些领域训练样本数量的缺乏, 训练任务的数量也是难以保证的. 这就导致在训练任务上学习的元知识不一定能很好地适用于测试任务. 此外, 在小样本分类任务中, 训练任务和测试任务的类别通常是不相交的, 因此训练任务和测试任务之间很容易存在分布偏差, 进一步加剧了元知识迁移的难度. 为此, 一些文献通过加权的方式对训练任务的分布进行修正, 使其与测试任务的分布更加一致[22,23]. 但是由于事先我们很难得知测试任务的分布, 而且实际场景中测试任务可能会随着时间发生变化, 通过加权修正训练任务分布的方式并不一定有效. 因此, 有文献提出通过任务扩充的方式来增加训练分布的多样性, 从而提高其泛化能力. 比如 Meta-MaxUp[24]组合不同的数据扩充方式来对支持集、查询集和任务进行扩充, MLTI[25]通过对随机采取的任务的 Mixup 来扩

充训练任务的分布. 虽然这些方法可以生成新的任务来扩充训练任务的分布, 但是这些方法并没有考虑生成任务与真实任务分布的关系, 难以保证生成任务的质量.

为了保证在训练任务上学习到的元知识可以有效地迁移到测试任务中, 我们认为, 训练任务应该具备以下特性.

1) 多样性. 生成的训练任务应该足够多样才能保证训练任务的分布包含未知的测试任务;
2) 真实性. 生成的任务应该服从真实任务的分布, 偏离真实任务分布的训练任务容易误导模型学习, 从而导致负迁移.

为此, 本文提出了一种基于多样真实任务扩充的鲁棒小样本分类方法(DATG). 该方法通过对已有少量任务进行 Mixup, 可以生成更多的训练任务帮助模型进行学习. 通过约束生成任务的多样性和真实性, 该方法可以有效提高小样本分类方法的泛化性. 具体来说, 我们先对训练集中的基类进行聚类得到不同的簇, 然后从不同的簇中选取任务进行 Mixup 以增加生成任务的多样性, 如图 1 所示. 此外, 基于簇间任务 Mixup 可以减轻学习到与类别高度相关的伪判别特征. 同时, 为了避免生成的任务与真实分布太偏离, 误导模型学习, 我们通过最小化生成任务与真实任务之间的最大均值差异(MMD)[26]来保证生成任务的真实性. 最后, 我们从理论上分析了为什么基于簇间任务 Mixup 的策略可以提高模型的泛化性能. 多个数据集上的实验结果进一步证明了本文提出的基于多样性和真实性的任务扩充方法的有效性.



(a) 随机任务 Mixup (b) 簇间任务 Mixup

图 1 随机任务 Mixup 和簇间任务 Mixup 的图示

本文第 1 节介绍基于元学习小样本分类任务和 Mixup 技术的相关工作和研究现状. 第 2 节介绍本文所需的基础知识, 包括小样本分类任务的定义、基于元学习的小样本分类方法中代表性方法 MAML 和原型网络. 第 3 节介绍本文提出的基于多样性和真实性的任务扩充模型. 第 4 节从理论上分析为什么基于簇间的任务 Mixup 可以提高模型的泛化性能. 第 5 节通过对比实验验证所提模型的有效性. 最后一节总结全文.

1 相关工作

1.1 基于元学习的小样本分类方法

小样本分类任务的难点是目标任务上有标签的训练样本太少, 直接在少量样本上学习容易导致模型出现过拟合现象, 进而导致模型的泛化能力下降. 为了解决这一问题, 最常用的方法是基于元学习的方法.

基于元学习的小样本分类方法希望通过学着去学习的方式来进行学习, 通过在大量不同的小样本学习任务上学习解决这类任务的元知识, 在遇到新的任务时可以利用这些元知识进行快速学习. MAML[6]是利用元学习思想进行小样本分类的算法, 它希望在不同任务上学习一个好的初始化权重, 在新任务上通过几步更新就可以达到最优. MAML 可以形式化为一个双层优化问题, 求解时需要求二阶 Hessian 矩阵, 因此存在着计算缓慢、内存消耗大等问题. 为了解决这一问题, 后续有很多工作被提出. 比如 FOMAML[7]利用一阶导数近似、IMAML[8]使用隐式双层优化求解方法、Reptile[9]用基学习器和元学习器的向量差作为梯度、MT-net[10]则通过将 MAML 的元学习器参数空间约简为由每一层的激活空间组成的子空间, 并在该子空间上进行快速学习, 进而加速整个学习过程. 为了找到更好的优化路径, Meta-SGD[11]不仅学习好的初始化权重, 还同时学习最优的

学习率和更新方向. 此外, 还有一些方法将 MAML 与贝叶斯学习联系起来, 从贝叶斯学习的角度对 MAML 进行分析和改进[12−16]. 原型网络[7]是基于元学习的另一代表性工作, 其将度量学习和元学习结合起来,希望学习一个适用于不同任务的度量空间, 在该空间中, 通过比较查询样本和类原型之间的距离进行分类. 为了学习更好的度量空间, 后续学者们提出了不同的类表示学习方法和相似性学习方法. 比如 CAN[17]利用查询样本结合注意力机制、AM3[18]利用多模态信息来对类原型进行表示. Simple CNAPS[19]使用马氏距离来度量不同特征之间的相关性, 从而更好地计算样本之间的相似性. DeepEMD[20]将每个图片表示成一个张量来更好地挖掘样本的局部信息, 采用推土机距离来考虑两个张量之间的相关性. Relation Network[21]使用神经网络代替距离函数, 将类原型和查询样本拼接后放入一个神经网络中, 自适应地学习它们之间的相似性.

虽然以上方法在小样本分类任务上取得了不错的进展, 但这些方法都需要一个假设: 训练任务和测试任务的分布是一致的. 这就要求我们的训练任务数量足够多, 可以有效地代表整体任务的分布. 但事实上, 由于某些领域样本数量的缺乏, 训练任务的数量也是难以保证的. 此外, 在小样本分类任务中, 训练任务和测试任务的类别一般都是不一致的, 因此训练任务和测试任务之间很容易存在分布偏差, 进一步加剧了元知识迁移的难度. 为了解决这一问题, 有一些研究聚焦在有分布外(out of distribution, OOD)任务的小样本分类学习上, 采用重加权技术来调整有偏差的训练任务分布. 比如 NestedMAML[22]通过双层优化对每个任务中的每个查询样本进行权重学习、Weighted Meta-Learning[23]通过计算目标任务与训练任务之间的距离来对训练任务进行加权. 但是由于我们很难事先得知测试任务的分布, 而且实际场景中测试任务可能会随着时间发生变化, 通过加权对训练任务分布修正的方法不能满足测试任务会不断变化的场景. 因此, 有文献提出通过对任务中样本或任务扩充的方式来增加训练分布的多样性, 从而提高其泛化能力. 比如: MAML_MetaMix[27]利用 Mixup 和通道变换的方式对任务中的查询样本进行扩充; Meta-MaxUp[24]比较了不同数据扩充方式在支持集、查询集和任务不同情况下的效果, 最后结合 MaxUp 策略提出一种通过组合不同的数据扩充方式来自适应对支持集、查询集和任务进行扩充的方法; MLTI[25]通过对随机采样的任务进行 Mixup 来扩充训练任务的分布. 虽然这些方法可以生成新的任务来扩充训练任务的分布, 但这些方法并没有考虑生成什么样的任务有利于模型的学习.

1.2 Mixup

Mixup[28]于 2018 年作为一种数据增强被首次提出, 通过对一对样本的输入和标签进行线性组合来合成新的样本, 被广泛使用在许多监督和半监督任务中来提高模型的泛化和对抗鲁棒性. 该方法简单有效, 有理论保证, 激发了一系列关于 Mixup 变体工作的产生. 比如: Manifold Mixup[29]把输入层的线性组合扩展到隐层, 对中间隐层的特征表示和标签进行线性组合来平滑决策边界; CutMix[30]从图像的空间角度出发, 从一张图片中随机选取一个区域裁剪后放到另一张图片对应的区域生成新的图片, 标签按照两张图片比例的线性组合来生成. Puzzle Mix[31]在 CutMix 的基础上引入显著性分析, 只选取图片的显著性区域进行裁剪, 避免裁剪的区域是与类别无关区域. PatchUp[32]同时结合 Mainfold Mixup 和 CutMix 的优点, 对中间隐层也进行裁剪. MetaMixUp[33]使用元学习方法来学习线性组合的系数, 不过, 由于元学习方法会增加计算复杂度, 目前大部分基于 Mixup 的方法还是通过从贝塔分布中采样得到的. 此外, 还有一些结合特定任务的 Mixup 方法, 比如: DAML[34]将 Mixup 的思想引入到开放域泛化任务中, 通过 Dir-mixup 来合成新域数据, 提高模型对开放环境下未知域的泛化能力; IMLT[25]将 Mixup 引入到元学习小样本任务中, 通过对任务进行 Mixup 扩充, 解决实际场景中训练训练任务缺乏的问题.

2 基础知识

本节主要介绍小样本学习中典型的 *N*-way *K*-shot 任务、元学习训练机制和两个代表性的小样本分类方法.

2.1 *N*-way *K*-shot任务和元学习训练机制

小样本分类任务旨在让机器像人类一样, 在新的类别只有少量标注数据的情况下, 可以快速学习和识别该类. 但是多少样本是少量样本并没有明确定义. 为了方便比较和评估不同的方法, 学术界通常在 *N*-way

K-shot 分类任务上进行测试, 认为: 如果一个方法可在 *N*-way *K*-shot 分类任务上取得好的结果, 该方法就具备小样本学习能力. 一个典型的 *N*-way *K*-shot 分类任务中包括两部分数据: 支持集(support set)和查询集(query set). 支持集对应于传统机器学习中的训练样本, 包含少量的有标签训练数据, 包括 $N\times K$ 个样本, 其中, N 为需要分类的类别数目, K 为每个类中的样本数目, 通常设为 1 或 5. 查询集对应于传统机器学习中的测试样本, 为待分类的无标签数据, 用于验证方法通过支持集样本对 N 个类的认知能力, 包括 $N\times Q$ 个样本.查询集的类别与支持集的类别一致, 只是每个类的样本数目为 Q 个. *N*-way *K*-shot 分类任务就是通过在 $N\times K$ 个支持集样本上学习一个模型, 用学到的模型对 $N\times Q$ 个查询样本进行分类.

但是直接利用 $N\times K$ 个样本进行学习, 十分容易出现过拟合现象. 为了解决这一问题, 目前的小样本分类方法通常借助于一个有大量有标签数据的辅助数据集. 先在该数据集上学习一些知识, 利用这些知识帮助目标小样本任务进行学习. 在小样本学习中, 把辅助数据集中的类别叫做基类(base class), 测试小样本任务中的类别叫做新类(novel class). 通常, 基类和新类的类别是不相交的. 基于元学习的小样本分类方法希望通过学着去学习的方式来进行学习, 通过在辅助数据集上构造一系列不同的 *N*-way *K*-shot 分类任务, 学习解决这类任务的元知识, 利用元知识帮助目标任务进行学习, 如图 2 所示. 一般把这样的训练方式叫做插曲式训练机制或者元学习训练机制. 测试时, 为了验证模型可以适用于不同的小样本分类任务, 通常在测试数据集上构造一系列 *N*-way *K*-shot 任务进行测试, 计算在这些任务上的平均性能作为模型的评价指标.

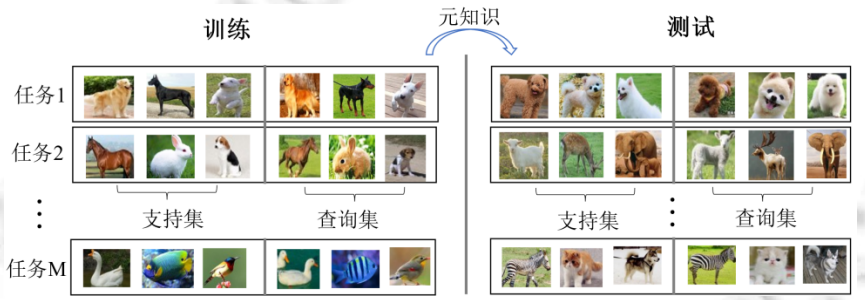


图 2 元学习训练和测试机制示意图

2.2 MAML和原型网络

在基于元学习的小样本分类方法中, 通常我们会构造一系列元训练任务 $\{\mathcal{T}_i\}_{i=1}^{M}$, 每个任务 $\mathcal{T}_i$ 中包含着两部分数据: 支持集 $\mathcal{D}_i^S=\{x_i^j,y_i^j\}_{j=1}^{N\times K}$ 和查询集 $\mathcal{D}_i^Q=\{x_i^j,y_i^j\}_{j=1}^{N\times Q}$. 模型 f_ϕ 在数据集 $\mathcal{D}$ 上的损失记为 $\mathcal{L}(f_\phi;\mathcal{D})$.

MAML[6]的目标是: 学习一个适用于不同任务的好的初始化参数 θ^*, 在新的任务上, 少量支持集样本 $\mathcal{D}^S$ 只需要更新一步或几步就可以达到当前任务的最优点. MAML 可以形式化为如下双层优化问题:

$$\left.\begin{aligned}\theta^*&=\arg\min_{\theta}\mathcal{F}(\theta),\\ \mathcal{F}(\theta)&=\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\mathcal{A}\mathrm{lg}(f_\phi^{MAML};\mathcal{D}_i^S);\mathcal{D}_i^Q)\end{aligned}\right\}\tag{1}$$

其中, $\mathcal{A}\mathrm{lg}(f_\phi^{MAML};\mathcal{D}_i^S)$ 表示利用 $\mathcal{D}^S$ 对任务 $\mathcal{T}_i$ 进行优化更新. 以一步梯度下降更新为例, $\mathcal{A}\mathrm{lg}(f_\phi^{MAML};\mathcal{D}_i^S)$ 可写成 $\mathcal{A}\mathrm{lg}(f_\phi^{MAML};\mathcal{D}_i^S)=\theta-\eta\nabla_\theta[\mathcal{L}(f_\theta^{MAML};\mathcal{D}^S)]$, η 是学习率, M 是元训练任务的数量.

原型网络[7]要学习的元知识是一个低维度量空间, 在该空间中查询样本可以根据计算与每个类原型的距离进行分类. 给定每个类的少量支持集样本, 第 r 的类原型 c_r 为该类所有支持集样本在度量空间特征表示 $f_\theta^{PN}(x^s)$ 的均值. 原型网络的损失函数如下:

$$\mathcal{L}=\frac{1}{M}\sum_{i=1}^{M}\left[-\sum_{i,k,r}\log p(y_{i,k}^q=r\,|x_{i,k}^q)\right]=\frac{1}{M}\sum_{i=1}^{M}\left[-\sum_{i,k,r}\log\frac{\exp(-d(f_\theta^{PN}(x_{i,k}^q),c_r))}{\sum_{r'}\exp(-d(f_\theta^{PN}(x_{i,k}^q),c_{r'}))}\right]\tag{2}$$

其中, $f_\theta^{PN}(x_{i,k}^q)$ 是第 i 个任务中第 k 类的查询样本经过原型网络得到的隐表示, M 是元训练任务的数量.

3 基于多样性和真实性的任务生成方法

本节详细介绍本文提出的基于多样性和真实性的任务生成方法 DATG. DATG 的主要目的是通过生成多样且符合真实分布训练任务, 缓解基于元学习的小样本学习方法在训练任务有限情况下泛化能力差的问题.

3.1 多样性任务生成

- 随机任务 Mixup

为了生成新的任务, 一种最简单的方法是从已有任务中随机选取两个任务进行 Mixup, 比如 MLTI[25]就采用了随机任务 Mixup 的策略. 具体来说, 先随机构造一对任务 $\mathcal{T}_i=\{\mathcal{D}_i^s,\mathcal{D}_i^q\}$ 和 $\mathcal{T}_j=\{\mathcal{D}_j^s,\mathcal{D}_j^q\}$, 假设模型 f 有 L 层, 样本 X 第 l 层的隐表示为 $H^l=f_{\theta^l}(X)(0\leqslant l\leqslant L^S)$, 其中, H^0=X, L^S 表示所有任务共享的层数. 在基于梯度的方法中, 只有部分层是共享的(即 L^S<L). 在基于度量的方法中, 所有层都是共享的(即 L^S=L). 然后, 再随机选择要进行 Mixup 的层 l, 并分别对支持集样本和查询集样本第 l 层的隐表示 $(H_i^{s(q),l},H_j^{s(q),l})$ 进行线性组合来生成新任务中的支持集样本和查询集样本, 具体生成公式如下:

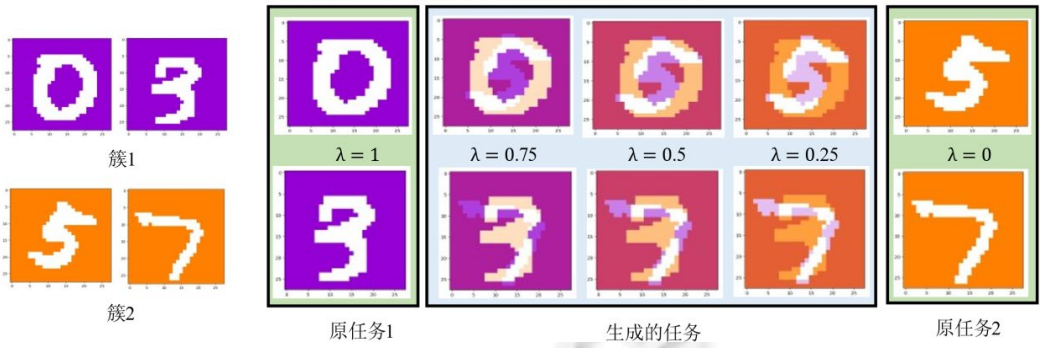
$$\tilde{H}_{cr}^{s,l}=\lambda H_i^{s,l}+(1-\lambda)H_j^{s,l},\tilde{H}_{cr}^{q,l}=\lambda H_i^{q,l}+(1-\lambda)H_j^{q,l} \tag{3}$$

其中, $\lambda\in[0,1]$是从贝塔分布 $Beta(\alpha,\beta)$采样得到的. 由于在基于元学习的小样本分类方法中, 样本的标签在不同任务中是不共享的, 也就是说同一个类的样本在第 i 个任务中的类标为 1, 在第 j 个任务中类标可能是 2. 因此, 在进行任务 Mixup 时只考虑特征之间的线性组合, 不用对标签进行线性组合. 也就是说, 不论线性组合系数λ是多少, 只需要保证由相同两个类别线性组合得到的样本属于同一类即可. 比如任务 1 中有 c_1,c_2,c_3 这 3 个类, 任务 2 中有 c_4,c_5,c_6 这 3 个类, 在生成新任务时, 我们分别将 c_1,c_4,c_2,c_6 和 c_3,c_7 中的隐层表示进行线性组合, 生成新的 3 个类的样本, 类别并没有实际含义.

- 簇间任务 Mixup

虽然随机任务 Mixup 可以生成新的任务, 但是其多样性可能不够. 如图 1(a)所示: 假设训练任务服从正态分布, 通过随机采样的方式我们得到的任务大都来自正态分布的中心位置, 通过对这些任务进行线性组合, 生成的任务也集中在中心附近, 缺乏多样性, 不能很好地泛化到处于边缘分布的任务上. 为此, 我们提出了一种簇间任务 Mixup 策略, 该策略可以有效地提高生成任务的多样性. 具体来说, 簇间任务 Mixup 策略在进行任务 Mixup 时先对所有的基类进行聚类, 将不同的类别分成不同的簇, 从不同簇中选取任务间, 按照公式(3)进行 Mixup. 本文简单地将所有类别分为 2 簇. 如图 1(b)所示: 在原始正态分布中, 处于中心位置的任务通过聚类可能被分到不同的簇中, 我们认为, 每个簇的任务都服从一个更瘦高的正态分布. 由于分簇后每个分布中的任务数量减少, 在原始分布中处于边缘位置的任务在瘦高的分布中被采样到的概率更大, 因此在进行簇间任务 Mixup 时, 生成的任务会更加多样.

此外, 基于簇间任务 Mixup 的策略可以在一定程度上减弱虚假相关性的学习, 进一步提高模型的泛化性能. 我们认为, 通过聚类可以发现一些和类别高度相关的簇信息. 为了方便说明, 我们用手写数字进行直观展示. 如图 3(a)所示: 0 和 3 这两个类的所有图片的背景都是紫色的, 5 和 7 这两个类的所有图片的背景都是橙色的, 通过聚类, 我们可以把 0 和 3 聚为一簇, 5 和 7 聚为一类. 然后, 我们只在簇间进行任务 Mixup. 如图 3(b)所示: 先分别在各自的簇内构造任务(如原任务 1 和原任务 2), 然后对来自不同簇的任务进行 Mixup, 生成新的任务. 通过改变线性组合的系数λ, 可以生成不同的任务. 比如图中生成 3 组不同任务, 不同任务中相同数字的背景颜色是不同的, 在这样的任务上学习, 可以避免学习与类别伪相关的颜色属性, 进而提高模型的泛化性能.



(a) 聚类将类别分成两簇 (b) 簇间任务 Mixup:来自不同簇的样本进行线性组合

图 3 簇间任务 Mixup 例子

3.2 任务真实性约束

除了生成任务的多样性, 我们还需要考虑生成任务的真实性. 虽然簇间任务 Mixup 可以增加任务的多样性, 但也可能生成没有意义的图片或者引入噪声, 误导模型的学习. 因此, 如何约束生成的任务符合真实的任务分布, 提高训练任务的质量也是十分重要的. MMD 是通过计算两个分布的均值距离来度量两个分布的差异, 被广泛应用于域泛化和域自适应领域来缩小源域与目标域之间的分布差距[35−37]. 受此启发, 我们通过最小化生成任务与已有任务之间的 MMD 来保证生成任务的真实性, 损失函数如下:

$$L_{MMD}=\frac{1}{M}\sum_{m=1}^{M}\left(\left\|\frac{1}{N\times(K+Q)}\sum_{i=1}^{N\times(K+Q)}\phi(\tilde{H}_{cr})-\frac{1}{N\times(K+Q)}\sum_{i=1}^{N\times(K+Q)}\phi(H)\right\|^2\right) \tag{4}$$

其中, $\phi(\cdot)$是核函数, $\tilde{H}_{cr}$ 和 H 分别是生成的任务和真实任务中的样本表示, M 是元训练任务的数量. 在计算中, 如果每次都计算合成任务与所有真实任务之间的 MMD 距离, 需要足够大的内存来存储所有任务. 因此, 在实验中我们用每个批次中的所有任务来代替全部任务.

3.3 算 法

生成任务后, 我们利用生成的任务代替原来的任务进行学习. 我们的方法主要是提出一种同时考虑多样性和真实性的任务生成策略, 可以作用于所有基于元学习的小样本分类方法上. 以 MAML 模型为基础模型, 其算法见算法 1. 可以看到, DATG 与 MAML 方法的主要不同之处在于元训练任务的构建. MAML 方法通过在随机选取的任务上学习一个好的初始点 θ, DATG 通过先对基类进行聚类(算法第 2 行), 在不同簇中选取任进行务 Mixup 以提高生成任务的多样性(算法第 4–9 行), 并通过引入 $\mathcal{L}_{MMD}$ 保证生成任务的真实性(算法第 10–14 行), 进而学习一个更具泛化性的初始点 θ.

算法 1. DATG 算法.

输入: 基类; Beta 分布; η; γ;

1: 随机初始化模型参数 θ
2: 对基类进行聚类, 得到两个簇
3: **for** $iter<n_{iter}$ **do**
4: 从一个簇中随机采样一个批次的任务 $\{\mathcal{T}_i\}_{i=1}^{|I|}$
5: **for** 每个任务 $\mathcal{T}_i$ **do**
6: 从另一个簇中随机采样一个任务 $\mathcal{T}_j$
7: 随机选择要进行 Mixup 的层 l
8: 得到任务 $\mathcal{T}_i$ 和任务 $\mathcal{T}_j$ 中样本在第 l 层的隐表示

9: 根据公式(3)生成新的任务

10: 根据公式(4)计算新生成任务与 $\{\mathcal{T}_i\}_{i=1}^{|I|}$ 和 $\{\mathcal{T}_j\}_{j=1}^{|I|}$ 的 MMD 距离和$\mathcal{L}_{MMD}$

11: 计算 MAML 网络在新生成任务中查询样本上的损失函数$\mathcal{L}_{cl}$

12: 根据$\mathcal{L}_{MMD}$+$\mathcal{L}_{cl}$, 利用梯度下降算法更新 MAML 的内循环参数

13: **end for**

14: 根据 MAML 在查询集上的损失$\mathcal{L}_{MMD}$, 利用梯度下降算法更新其外循环参数

15: **end for**

4 理论分析

本节先给出基于随机任务 Mixup 的方法对于模型泛化界的影响, 然后在此基础上分析本文提出的基于簇间任务 Mixup 策略对模型泛化界的影响. 根据文献[25,38], 在以 MAML 为代表的基于梯度的元学习小样本分类方法中, 以两层神经网络和二分类为例, 利用线性组合生成的任务来替代原任务进行学习的泛化性能分析结果如下. 定理 1 和引理 1 的详细证明过程见文献[25].

假设训练集中每个类的样本数目均为 N, 训练任务的数目为$|I|$. 对于每个任务$\mathcal{T}_i$, 其损失定义为

$$l(f^{MAML}(x),y)=\log(1+\exp(f^{MAML}(x)))-yf^{MAML}(x),$$

其中, $f_{\phi_i}^{MAML}(x_{i,k})=\phi_i^{\top}\sigma(Wx_{i,k}):=\phi_i^{\top}h_{i,k}^1$, $h_{i,k}^1$ 是样本 $x_{i,k}$ 第 1 层的隐表示.

根据公式(3)来生成新的任务, 新任务中的查询集表示为 $\mathcal{D}_{i,cr}^q=\{\tilde{H}_{i,cr}^{q,l},\tilde{Y}_{i,cr}^q\}$. 为了方便, 我们省略查询集样本上标 q. 在查询集上的经验损失为 $\mathcal{L}_t(\{\mathcal{D}_{i,cr}\}_{i=1}^{|I|})=|I|^{-1}\sum_{i=1}^{|I|}\mathcal{L}(\mathcal{D}_{i,cr})=(N|I|)^{-1}\sum_{i=1}^{|I|}\sum_{k=1}^{N}\mathcal{L}(f_{\phi_i}(x_{i,k,cr}),y_{i,k,cr})$, 那么可以得到如下引理.

引理 1. 依据公式(3)对任务进行线性组合生成新的任务, 假设λ~$Beta(\alpha,\beta)$. 令$\psi(u)=e^u/(1+e^u)^2$, $N_{i,r}$ 表示任务$\mathcal{T}_i$中第r类样本的数目. 那么存在一个常数c>0, 使得模型在新生成任务上的损失函数 $\mathcal{L}_t(\{\mathcal{D}_{i,cr}\}_{i=1}^{|I|})$ 的二阶近似如下:

$$\mathcal{L}_t(\bar{\lambda}\cdot\{\mathcal{D}_i\}_{i=1}^{|I|})+c\frac{1}{N|I|}\sum_{i=1}^{|I|}\sum_{k=1}^{N}\psi(h_{i,k}^{1\top}\phi_i)\cdot\phi_i^{\top}\left(\frac{1}{|I|}\sum_{i=1}^{|I|}\frac{1}{2}\sum_{r=1}^{2}\frac{1}{N_{i,r}}\sum_{i=1}^{|I|}\sum_{k=1}^{N_{i,r}}h_{i,k;r}^1h_{i,k;r}^{1\top}\right)\phi_i \tag{5}$$

其中, $\bar{\lambda}=\mathbb{E}_{\mathcal{D}_\lambda}[\lambda]$, $\mathcal{D}_\lambda\sim\frac{\alpha}{\alpha+\beta}Beta(\alpha+1,\beta)+\frac{\alpha}{\alpha+\beta}Beta(\beta+1,\alpha)$.

引理 1 表明: 通过公式进行线性组合生成新任务, 并利用新任务替换原始任务训练模型, 相当于对ϕ_i们增加了一项正则项约束. 为了进一步分析该正则项是如何提高模型的泛化性能的, 我们考虑公式中正则项的分布 $\mathcal{F}_\gamma=\{H^{l\top}\phi:\mathbb{E}[\psi(H^{l\top}\phi)]\phi^{\top}\Sigma\phi\leqslant\gamma\}$, 其中, $\Sigma=\mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})}\mathbb{E}_{\mathcal{T}}[H^lH^{l\top}]$, $\mu_{\mathcal{T}}=\mathbb{E}_{\mathcal{T}}[H^l]$. 假设对来自任务分布中的每个任务$\mathcal{T}$~$p(\mathcal{T})$满足下面的条件:

$$rank(\Sigma)\leqslant R,\ \|\Sigma^{\dagger/2}\mu_{\mathcal{T}}\|\leqslant U \tag{6}$$

其中, $\Sigma^{\dagger}$ 是Σ的广义逆. 在此基础上, 我们假设 H^l 的分布对于一些$\rho\in(0,1/2]$是ρ-retentive. 也就是说, 对任意非零向量 $v\in\mathbb{R}^d$, $[\mathbb{E}[\psi(v^{\top}H^l)]]^2\geqslant\rho\cdot\min\{1,\mathbb{E}[(v^{\top}H^l)^2]\}$. 这一假设在权重的 l_2 范数是有界的情况下是可以满足的. 模型在任务上的经验风险和期望风险分别定义为 $\mathcal{R}(\{\mathcal{D}_i\}_{i=1}^{|I|})$, $\mathcal{R}=\mathbb{E}_{\mathcal{T}_i\sim p(\mathcal{T})}\mathbb{E}_{(X_i,Y_i)\sim\mathcal{T}_i}[\mathcal{L}(f_{\phi_i}(X_i),Y_i)]$, 在通过线性组合对任务扩充得到的训练和期望风险之间的泛化误差界如下.

定理 1. 假设 X_i, Y_i 和ϕ的谱范数是有界的且假设公式(6)成立, 那么存在一些常数 A_1,A_2,A_3>0, 对所有的 $f_{\mathcal{T}}\in\mathcal{F}_\gamma$, $\delta\in(0,1)$, 至少以 1−δ的概率, 以下泛化界成立:

$$|\mathcal{R}(\{\mathcal{D}_i\}_{i=1}^{|I|})-\mathcal{R}|\leqslant A_1\max\left\{\left(\frac{\gamma}{\rho}\right)^{1/4},\left(\frac{\gamma}{\rho}\right)^{1/2}\right\}\left(\sqrt{\frac{R+U}{N}}+\sqrt{\frac{R+U}{|I|}}\right)+A_2\sqrt{\frac{\log(|I|/\delta)}{N}}+A_3\sqrt{\frac{\log(1/\delta)}{|I|}} \tag{7}$$

根据上面的引理和定理可知: 基于随机任务 Mixup 的方法主要是通过在原来损失函数的基础上加了一项

正则项 $c\frac{1}{N|I|}\sum_{i=1}^{|I|}\sum_{k=1}^{N}\psi(h_{i,k}^{1\top}\phi_i)\cdot\phi_i^{\top}\left(\frac{1}{|I|}\sum_{i=1}^{|I|}\frac{1}{2}\sum_{r=1}^{2}\frac{1}{N_{i,r}}\sum_{i=1}^{|I|}\sum_{k=1}^{N_{i,r}}h_{i,k;r}^{1}h_{i,k;r}^{1\top}\right)\phi_i$ 来提高模型的泛化性能, 该正则项中的 $\phi^{\top}\Sigma\phi$ 满足 $\mathcal{F}_{\gamma}=\{H^{l\top}\phi:\mathbb{E}[\psi(H^{l\top}\phi)]\phi^{\top}\Sigma\phi\leqslant\gamma\}$, 基于随机任务 Mixup 的方法通过对正则项进行优化, 得到一个更小的γ, 进而使公式(7)中的泛化误差界更紧致.

与没有任务 Mixup 策略、直接用原始任务进行训练的方法相比, 本文可以通过 Mixup 的策略增加数据的方差, 减小γ, 提高模型的泛化性. 与随机任务 Mixup 策略相比, 本文主要通过先聚类再从不同簇随机选取任务进行 Mixup, 类似于从所有任务中通过分层采样的方式进行任务采样, 通过分层采样和随机采样对整体分布估计方差比较的证明如下.

假设随机变量 X 的数学期望μ, 存在离散型随机变量 Y, $p_j=P(Y=y_j)$, $j=1,2,3,\ldots,m$, 在 $Y=y_j$ 条件下, 可以从 X 的条件分布抽样,则:

$$\mathbb{E}(X)=\mathbb{E}\{\mathbb{E}(X|Y)\}=\sum_{j=1}^{m}\mathbb{E}[X|Y=y_j]p_j \tag{8}$$

如果在 $Y=y_j$ 条件下生成的 $N_j=Np_j$ 个抽样值设为 $X_i^{(j)},i=1,2,...,N_j$, 则可以用 $\frac{1}{N_j}\sum_{i=1}^{N_j}X_i^{(j)}$ 估计$\mathbb{E}[X|Y=y_j]$, 估计μ为 $\hat{\mu}=\sum_{j=1}^{m}\frac{1}{N_j}\sum_{i=1}^{N_j}X_i^{(j)}p_j=\frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{Np_j}X_i^{(j)}$, 估计方差为

$$Var(\hat{\mu})=\frac{1}{N^2}\sum_{j=1}^{m}Np_jVar[X|Y=y_j]=\frac{1}{N}\sum_{j=1}^{m}Var[X|Y=y_j]p_j=\frac{1}{N}\mathbb{E}\{Var[X|Y]\}\leqslant\frac{1}{N}Var[X] \tag{9}$$

可以看出, 分层采样比直接用平均值法估计真实分布的方差小, 也就是说比用随机采样估计的方差少. 由于分层采样对于整体分布的估计优于随机采样, 因此可以减少对采样数目的需求. 在相同采样任务下提高模型的泛化性能.

5 实验分析

5.1 数据集

我们在小样本分类任务中使用最广泛的自然图像数据集和数据匮乏的医学领域的数据集来验证方法, 数据集具体介绍如下.

1) miniImageNet-S[39]: 该数据集来自于小样本分类中最常用的标准数据集 miniImageNet[40]. miniImageNet 包含 60 000 张的三通道彩色图像, 共有 100 个类别, 每个类别 600 样本. 该数据集最广泛采用的划分方式为训练集、验证集以及测试集, 各包含类 64, 16 以及 20 个类. miniImageNet-S 是通过从 miniImageNet 中选取 12 个训练类别构造元训练类别得到的, 其目的是通过降低训练类别数目来降低元训练任务的数量;
2) ISCI[41]: 是一个医学影响数据集. 根据文献[25], 我们选择“ISIC 2018: 皮肤病变分析到黑色素瘤检测”调整中的任务 3. 为 10 015 张医学图像分类, 类别分别为: 痣、皮肤纤维瘤、黑色素瘤、色素鲍恩氏病、良性角化病、基底细胞癌、血管性. 我们使用了样本数量最多的 4 个类别作为元训练类别, 其余 3 个类别作为元测试类别. 由于该数据集类别较少, 我们在该数据集上进行的是 2 分类任务;
3) DermNet-S[25]: 是在公共数据集 DermNet Skin Disease Atlas 基础上构造的一个数据集. 原始数据集中包括来自于 625 个细粒度类中的 22 000 多张图片. 文献[25]关注样本数不少于 30 个类的类别, 筛选出 203 个类. 该数据集的类服从长尾分布, 在使用时, 仅用前 30 个类进行元训练, 后 53 个类进行测试;
4) Tabular Murris[42]: 是基于已有的 Tabula Muris 数据集构造的一个新的单细胞转录组数据集. 原始的 Tabula Muris 数据集包含了从小鼠模型生物的 23 个器官中收集的 124 种细胞类型的 105 960 个细胞.

每个细胞由 23 341 个基因表示. 文献[42]从中选取了 2 866 个具有高标准化对数离散度的基因来表示每个细胞. 这个数据集的目标是对细胞进行分类, 用于训练、验证和测试的类别数分别为 15/4/4.

5.2 实验设置

本文的实验主要在基于元学习训练机制的代表性方法 MAML 和原型网络上进行. 为了公平比较, 与文献[25]一致, 除了 Tabular Murris 数据集, 在其他数据集上我们采用与文献[6]相同的主干特征提取网络, 它包含 4 个卷积块和一个分类器层. 每个卷积块各包括一个卷积层、批处理规范化层和 ReLU 激活层. 对于 MAML, 我们在最后一个卷积块和分类器层上应用了任务特定的自适应. 在 Tabular Murris 数据集上, 主干特征提取网络包含两个全连接模块和一个线性回归层, 其中, 每个全连接模块各包含一个线性层、批处理归一化层、ReLU 激活层和 dropout 层. Dropout 的比例为 0.2, 线性层的输出通道为 64. 训练和测试任务都是基于 N-way K-shot 分类任务, 其中, ISIC 的 N=2, 其余数据集的 N=5, 每个任务中查询样本的数目为 15.训练中, 使用动量为 0.9 的 SGD 优化器对模型进行优化, 初始学习率为 0.05, 学习率的衰减因子为 0.1. 当模型迭代 60 轮后生效, 每 10 轮衰减一次, 模型一共训练 15 000 轮. 测试时, 从测试数据集中随机采取 20 000 个任务进行测试, 计算在所有任务上的平均精度作为最终评价指标.

5.3 实验及结果分析

为了全面评估本文提出方法 DATG, 我们从以下几个方面进行实验验证: (1) DATG 在小样本分类任务的表现如何? (2) DATG 是怎么发挥作用的? (3) DATG 在什么情况下有效?

- Q1: DATG 在小样本分类任务的表现如何?

首先, 我们在标准的小样本分类任务上进行实验验证. 我们选取基于元学习中的代表性方法 MAML 和原型网络作为基线方法,在此基础上, 比较我们的方法 DATG 与不同增强方法(MetaMix, Meta-MaxUp 和 MLTI)和基于正则化的方法(Meta-Reg, TAML 和 Meta-Dropout)在 5-way 1-shot 和 5-way 5-shot 任务上的效果. 如表 1 所示, 在所有 4 个数据集上, 本文提出的 DATG 始终优于其他所有方法, 证明了我们方法在小样本分类任务上的有效性. 与 MLTI 方法只采用随机任务 Mixup 策略进行任务扩充相比, 我们的任务扩充策略考虑了任务的多样性和真实性, 可以生成质量更好的任务, 进一步提高模型对于小样本分类的性能.

表 1 5-way 1-shot 和 5-way 5-shot 分类精度比较

基准方法	策略	miniImageNet-S		ISIC		DermNet-S		Tabular Murris	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML	Vanilla	0.382 7	0.521 4	0.575 9	0.652 4	0.434 7	0.605 6	0.790 8	0.885 5
	Meta-Reg	0.383 5	0.517 4	0.585 7	0.684 5	0.450 1	0.609 2	0.791 8	0.890 8
	TAML	0.387 0	0.527 5	0.583 9	0.660 9	0.457 3	0.611 4	0.798 2	0.891 1
	Meta-Dropout	0.383 2	0.525 3	0.584 0	0.673 2	0.443 0	0.608 6	0.781 8	0.892 5
	MetaMix	0.394 3	0.541 4	0.603 4	0.694 7	0.468 1	0.635 2	0.810 6	0.897 5
	Meta-MaxUp	0.392 8	0.530 2	0.586 8	0.691 6	0.461 0	0.626 4	0.795 6	0.888 8
	MLTI	0.415 8	0.552 2	0.617 9	0.706 9	0.480 3	0.645 5	0.817 3	0.910 8
	Ours	**0.423 2**	**0.565 9**	**0.623 9**	**0.710 8**	**0.489 7**	**0.650 2**	**0.823 3**	**0.918 9**
ProtoNet	Vanilla	0.362 6	0.507 2	0.585 6	0.662 5	0.442 1	0.603 3	0.800 3	0.892 0
	MetaMix	0.397 6	0.531 0	0.605 8	0.701 2	0.477 1	0.626 8	0.807 2	0.893 0
	Meta-MaxUp	0.398 0	0.533 5	0.596 6	0.689 7	0.460 6	0.629 7	0.808 0	0.894 2
	MLTI	0.413 6	0.553 4	0.628 2	0.715 2	0.493 8	0.651 9	0.818 9	0.901 2
	Ours	**0.422 1**	**0.564 7**	**0.632 1**	**0.717 8**	**0.520 9**	**0.660 2**	**0.824 5**	**0.911 3**

为进一步验证我们的方法通过对任务进行扩充, 可以有效应对训练任务和测试任务分布之间存在差异的问题, 我们在更难的跨域的小样本分类任务上进行实验验证. 在跨域小样本分类任务中, 训练和测试的任务来自不同的数据集. 表 2 展示了在两个跨域小样本分类任务下, 我们的方法都可以有效地提升 MAML 和原型网络的分类性能, 进一步证明了我们所提出扩充策略可以生成有效的任务, 提升小样本分类模型的泛化性能.

表 2 跨域小样本分类精度比较

模型	策略	Mini→DermNet		DermNet→Mini	
		1-shot	5-shot	1-shot	5-shot
MAML	Vanilla	0.336 7	0.504 0	0.284 0	0.409 3
	+MLTI	0.367 4	0.525 6	0.300 3	0.422 5
	+Ours	**0.374 5**	**0.534 0**	**0.312 4**	**0.437 8**
ProtoNet	Vanilla	0.331 2	0.501 3	0.281 1	0.403 5
	+MLTI	0.354 6	0.517 9	0.300 6	0.422 3
	+Ours	**0.356 9**	**0.521 8**	**0.311 2**	**0.432 5**

- Q2: DATG 是怎么发挥作用的?

为了分析 DATG 的作用机制, 我们对 DATG 进行消融实验分析. 以 MAML 和原型网络为基础模型, 构造了 4 种不同的策略.

1) Vanilla: 不进行任务扩充;

2) 随机 Mixup: 随机采取任务通过 Mixup 策略进行任务扩充;

3) 簇间 Mixup: 先聚类, 从不同簇采取任务通过 Mixup 的策略进行任务扩充;

4) 我们的方法: 基于簇间 Mixup 策略生成新任务, 并通过 MMD 损失约束其与分布的一致性.

表 3 展示了 miniImageNet-S 上 4 种不同情况下的 5-way 1-shot 和 5-way 5-shot 任务下的分类精度. 可以看出, 簇间 Mixup 策略和 MMD 约束都是有效的, 也证明了为了提高模型的泛化性能, 需要同时考虑训练任务的多样性和真实性.

表 3 消融实验

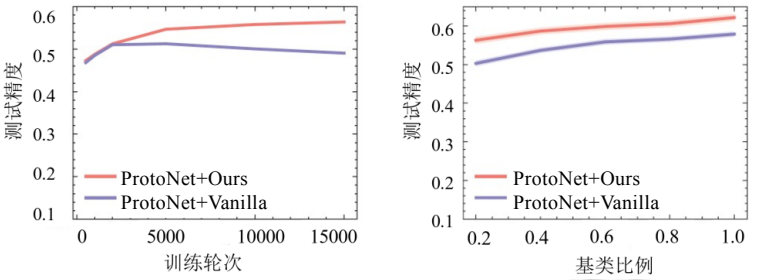
基准方法	策略	miniImageNet-S	
		1-shot	5-shot
MAML	Vanilla	0.382 7	0.521 4
	随机 Mixup	0.415 8	0.552 2
	簇间 Mixup	0.419 7	0.260 4
	Ours	**0.423 2**	**0.565 9**
ProtoNet	Vanilla	0.362 6	0.507 2
	随机 Mixup	0.413 6	0.553 4
	簇间 Mixup	0.418 9	0.559 4
	Ours	**0.422 1**	**0.564 7**

- Q3: DATG 在什么情况下有效?

由于本文方法主要是针对缺乏训练任务的场景设计的, 因此, 我们主要分析训练任务的数目对于 DATG 的影响.

首先, 我们分析在训练过程中, 随着训练任务数目的增加, DATG 的性能是如何变化的. 图 4(a)展示了在 miniImageNet 中随机选取 12 个基类进行训练时, 本文方法和原型网络的测试精度随着训练轮次的变化情况. 由于在小样本分类任务中, 任务是随机选取的, 不同轮次中训练任务是不相同的, 因此, 训练轮次的增加也代表了训练任务数目的增加. 从图 4(a)可以发现: 随着训练的进行, 本文方法的测试精度会不断上升,而原型网络的精度会出现先上升后下降的情况. 这可能是因为随着训练轮次的增加, 原型网络会出现过拟合的现象. 而本文方法因为可以增加任务的多样性, 因此不容易出现过拟合现象.

然后, 我们通过改变训练集中参与训练的基类数目来进一步分析训练任务对我们方法的影响. 通过从所有基类中随机选取一定比例的基类, 在选中的基类中构造训练任务进行模型训练. 图 4(b)展示我们的方法和原型网络在 5-way 1-shot 任务上的测试精度随着 miniImageNet 中参与训练的基类数目变化的情况.



(a) 测试精度随训练轮次的变化 (b) 测试精度随训练类别数目的变化

图 4 训练任务对测试精度的影响

可以看出:

1) 随着参与训练基类数目的增加, 我们的方法和原型网络的性能都会提升. 这是因为参与训练基类数目的增加, 意味着训练任务可采样的空间更大, 训练任务的会更加多样, 小样本模型的泛化性能更好. 这也证明了通过增加训练任务多样性来提高模型泛化性能的合理性;

2) 随着参与训练基类数目的增加, 与原型网络相比, 我们的方法提升的性能会逐渐下降. 这也是合理的, 因为随着训练类别的增加, 原始任务的多样性已经足够了, 因此, 通过增加样本多样性的方式对于小样本泛化能力的提升会下降. 由此也可以推断, 我们的方法在训练任务多样性不足的情况下更有效.

6 总 结

基于元学习的方法是通过学着去学习的方式, 可以有效避免小样本分类任务中训练样本太少容易过拟合的问题, 是小样本分类任务中最重要的一类方法. 但是这类方法通常需要大量的训练任务来保证在训练任务和测试任务来自同一分布, 大大限制了这些模型在一些训练任务也十分缺乏的场景下的应用. 为此, 本文基于多样真实任务生成的鲁棒小样本分类方法. 该方法通过簇间任务 Mixup 策略可以生成具有多样性的任务, 通过约束生成任务与真实任务分布之间的 MMD 距离来保证生成任务的真实性, 可以有效提高基于元学习的小样本方法的泛化性能. 最后, 我们从理论上分析了为什么簇间任务 Mixup 策略可以提高模型的泛化性能, 并且在多个数据集上验证了我们所提方法在小样本分类任务上的有效性. 本文只是提出了一种约束生成任务多样和真实性的方法, 未来如何更好地约束生成任务的质量, 也是一个值得研究的方向. 比如: 在多样性任务生成方面, 本文通过聚类进行簇间任务 Mixup, 不同簇之间的特征并不明确, 未来可以考虑利用因果学习进行分簇, 进一步挖掘伪相关特征, 提高模型的泛化性; 在任务真实性约束方面, 本文采用传统的 MMD 对特征分布进行对齐, 可以进一步考虑标签分布之间的对齐, 提升生成任务的真实性.

References:

[1] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770–778.

[2] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. 1106–1114.

[3] Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484–489.

[4] Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, Vol.35. 2022. 27730–27744.

[5] Wang Y, Yao Q, Kwok JT, *et al*. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR), 2020, 53(3): 1–34.

[6] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc. of the Int'l Conf. on Machine Learning. 2017. 1126−1135.

[7] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT, 2017. 4080−4090.

[8] Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. arXiv:1803.02999, 2018.

[9] Rajeswaran A, Finn C, Kakade SM, *et al*. Meta-learning with implicit gradients. In: Advances in Neural Information Processing Systems. 2019. 32.

[10] Lee Y, Choi S. Gradient-based meta-learning with learned layerwise metric and subspace. In: Proc. of the Int'l Conf. on Machine Learning. 2018. 2927−2936.

[11] Li Z, Zhou F, Chen F, *et al*. Meta-SGD: Learning to learn quickly for few-shot learning. arXiv:1707.09835, 2017.

[12] Yoon J, Kim T, Dia O, *et al*. Bayesian model-agnostic meta-learning. In: Advances in Neural Information Processing Systems. 2018. 31.

[13] Ravi S, Beatson A. Amortized Bayesian meta-learning. In: Proc. of the Int'l Conf. on Learning Representations. 2019.

[14] Patacchiola M, Turner J, Crowley EJ, *et al*. Bayesian meta-learning for the few-shot setting via deep kernels. In: Advances in Neural Information Processing Systems, Vol.33. 2020. 16108−16118.

[15] Zhang Q, Fang J, Meng Z, *et al*. Variational continual Bayesian meta-learning. In: Advances in Neural Information Processing Systems, Vol.34. 2021. 24556−24568.

[16] Chen L, Chen T. Is Bayesian model-agnostic meta learning better than model-agnostic meta learning, provably? In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2022. 1733−1774.

[17] Hou R, Chang H, Ma B, *et al*. Cross attention network for few-shot classification. In: Advances in Neural Information Processing Systems. 2019. 4003−4014.

[18] Xing C, Rostamzadeh N, Oreshkin B, *et al*. Adaptive cross-modal few-shot learning. In: Advances in Neural Information Processing Systems. 2019. 4847−4857

[19] Bateni P, Goyal R, Masrani V, *et al*. Improved few-shot visual classification. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 14481−14490.

[20] Zhang C, Cai Y, Lin G, *et al*. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 12203−12213.

[21] Sung F, Yang Y, Zhang L, *et al*. Learning to compare: Relation network for few-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1199−1208.

[22] Killamsetty K, Li C, Zhao C, *et al*. A nested bi-level optimization framework for robust few shot learning. Proc. of the AAAI Conf. on Artificial Intelligence, 2022, 36(7): 7176−7184.

[23] Cai D, Sheth R, Mackey L, *et al*. Weighted meta-learning. arXiv:2003.09465, 2020.

[24] Yao H, Huang LK, Zhang L, *et al*. Improving generalization in meta-learning via task augmentation. In: Proc. of the Int'l Conf. on Machine Learning. 2021. 11887−11897.

[25] Yao H, Zhang L, Finn C. Meta-learning with fewer tasks through task interpolation. arXiv:2106.02695, 2021.

[26] Ni R, Goldblum M, Sharaf A, *et al*. Data augmentation for meta-learning. In: Proc. of the Int'l Conf. on Machine Learning. 2021. 8152−8161.

[27] Smola AJ, Gretton A, Borgwardt K. Maximum mean discrepancy. In: Proc. of the 13th Int'l Conf. 2006. 3−6.

[28] Zhang H, Cisse M, Dauphin YN, *et al*. Mixup: Beyond empirical risk minimization. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[29] Verma V, Lamb A, Beckham C, *et al*. Manifold mixup: Better representations by interpolating hidden states. In: Proc. of the Int'l Conf. on Machine Learning. 2019. 6438−6447.

[30] Yun S, Han D, Oh SJ, *et al*. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2019. 6023−6032.

[31] Kim JH, Choo W, Song HO. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In: Proc. of the Int'l Conf. on Machine Learning. 2020. 5275−5285.

[32] Faramarzi M, Amini M, Badrinaaraayanan A, *et al*. Patchup: A regularization technique for convolutional neural networks. Proc. of the AAAI Conf. on Artificial Intelligence, 2022, 36(1): 589−597.

[33] Mai Z, Hu G, Chen D, *et al*. MetaMixUp: Learning adaptive interpolation policy of mixup with metalearning. IEEE Trans. on Neural Networks and Learning Systems, 2021, 33(7): 3050−3064.

[34] Guo H, Mao Y, Zhang R. Mixup as locally linear out-of-manifold regularization. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1): 3714−3722.

[35] Shu Y, Cao Z, Wang C, *et al*. Open domain generalization with domain-augmented meta-learning. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2021. 9624−9633.

[36] Long M, Wang J, Ding G, *et al*. Transfer feature learning with joint distribution adaptation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2200−2207.

[37] Li H, Pan SJ, Wang S, *et al*. Domain generalization with adversarial feature learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 5400−5409.

[38] Zhang L, Deng Z, Kawaguchi K, *et al*. How does mixup help with robustness and generalization? arXiv:2010.04819, 2020.

[39] Liu J, Chao F, Lin CM. Task augmentation by rotating for meta-learning. arXiv:2003.00804, 2020.

[40] Vinyals O, Blundell C, Lillicrap T, *et al*. Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. 2016. 29.

[41] Milton MAA. Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv:1901.10802, 2019.

[42] Cao K, Brbic M, Leskovec J. Concept learners for few-shot learning. arXiv:2007.07375, 2020.

刘鑫(1994－), 女, 博士生, CCF 学生会员, 主要研究领域为机器学习, 小样本学习.

景丽萍(1978－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 高维数据表示及其在人工智能领域中的应用.

于剑(1969－), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为机器学习理论, 自然语言处理.