基于模型分解的小样本学习

叶翰嘉, 詹德川*

计算机软件新技术国家重点实验室 (南京大学), 南京 210023

* 通信作者. E-mail: zhandc@lamda.nju.edu.cn

收稿日期: 2018-12-21; 修回日期: 2019-04-12; 接受日期: 2019-07-04; 网络出版日期: 2020-04-26

国家重点研发计划“大数据分析的基础理论和技术方法”(批准号: 2018YFB1004300)、国家自然科学基金(批准号: 61773198, 61632004)和计算机软件新技术协同创新中心, 南京大学优秀博士研究生创新能力提升计划项目资助

摘要 机器学习依赖大量样本的统计信息进行模型的训练, 从而能对未知样本进行精准的预测. 搜集样本及标记需要耗费大量的资源, 因而如何基于少量样本 (few-shot learning) 进行模型的训练至关重要. 有效的模型先验 (prior) 能够降低模型训练对样本的需求. 本文基于元学习 (meta learning) 框架, 从相关的、类别不同的数据中学习模型先验, 并将这种先验应用于新类别的少样本任务. 与此同时, 本文提出“模型组合先验” (MCP, model composition prior) 方法, 通过目标函数的最优条件对模型结构进行分解, 并分别估计模型的各个组成部分, 得到有效的分类器. 这种分解方式具有较高的可解释性, 能够指导在不同小样本任务中“共享”与“独立”的成分, 从而指导元学习的具体实现. 在人造数据中, 本文方法能够恢复出小样本任务之间的关联性; 在图像数据上, MCP 方法能取得比当前主流方法更优异的效果.

关键词 小样本学习, 元学习, 模型先验, 模型分解

1 引言

尽管当前的机器学习方法在文本^[1]、图片^[2]等多个领域取得了显著的成果, 在训练过程中, 机器学习算法需要“大数据”来获得准确的分布统计信息. 而在某些实际任务中, 机器学习算法面临“小数据”输入, 即需要基于少量样本进行模型训练. 例如, 某些稀有物种的图片数据难以搜集, 只有极少的有标记图片, 此时模型需要利用仅有的图片进行训练并对未知的图片进行分类^[3]; 在利用生物特征的识别系统中, 无法要求用户做过多次的数据采集, 因此需要从用户提供的有限样本中学习^[4]; 对于能够模仿用户行为的机器人, 往往只能收到用户的一到两次演示^[5]; 推荐系统中, 对于新商品只有有限的图片、文本描述^[6]. 此外, 搜集有标记的样本经常需要耗费大量的人力、物力资源. 因此, 对于一个仅有少量有标记训练样本的任务如何获得一个可靠的分类器非常重要^[7], 这类问题在后文中统称为“小样本学习问题” (few-shot learning).

引用格式: 叶翰嘉, 詹德川. 基于模型分解的小样本学习. 中国科学: 信息科学, 2020, 50: 662–674, doi: 10.1360/N112018-00332
Ye H-J, Zhan D-C. Few-shot learning via model composition (in Chinese). Sci Sin Inform, 2020, 50: 662–674, doi: 10.1360/N112018-00332

人类具有通过少样本进行物体识别的能力: 即使对某个新事物只有单一的图片、文本描述, 人也能够对该事物进行很好的理解, 并加以识别^[8]. 人类的这种小样本识别能力归因于其对已有事物观察的丰富经历, 这种对事物的“先验”使得在观察新事物时只需要很少的样本就能对其判别信息进行抽取. 为使机器学习算法具有如人类一般通过小样本学习的能力, 需要为算法寻求一个“模型先验”, 当遇到少量的新类样本时, 该模型先验能够辅助模型快速适配一个新的任务. “元学习”(meta-learning)提供了一种学习跨任务先验的方式. 不同于以往的人工选择, 元学习能够利用任务之间的相关性学习出适合每一个任务的模型配制方式^[9,10], 从具有不同类别的相关任务中提取先验知识, 并将其应用于未知的类似的任务中. 已有的算法对模型的先验有不同的实现. 如在训练一个图像分类器时, 往往会使用 ImageNet^[11] 上预训练的模型权重来进行初始化, 或者用预训练的模型直接提取特征; 对于类似任务的优化策略与调参经验也会作为新任务优化方式的首选. 初始化、特征表示、优化策略都可看作一类模型先验, 这种和具体类别、任务无关的信息可以辅助新任务的学习, 并且也被用于小样本学习问题中^[12~18].

本文通过模型的目标函数的性质, 对模型最优解的结构进行分解, 并将这种分解方式作为一种可行的先验. 假设不同的任务使用共同的特征抽取函数, 而具有不同的线性分类器. 通过分类器优化的最优解条件, 能够得到最优分类器具有的结构. 例如在最小二乘支持向量机 (least square support vector machine, LSSVM) 中, 分类器可以表示为特征精度矩阵和类别中心矩阵两部分的组合. 基于这种分类器的结构分解, 可以分别估计对应的模型成分. 基于 LSSVM 的形式, 本文提出“模型组合先验”(model composition prior, MCP) 方法. 这种对模型结构的考虑避免盲目地构造分类器, 并且通过检查结构, 很容易确定对于所有类通用的部分 (全局信息) 与特有的部分 (本地信息). 基于 MCP, 本文还提出了一种根据当前任务自动调节先验的方法 MCP⁺. 该方法利用当前任务中少量样本反映出的特征指导特征精度矩阵的构建, 从而找到适配当前任务的分类器. 在实验中, 本文首先验证了在人造数据集上 MCP 方法能够良好估计不同小样本任务之间的共享信息. 然后, 在数据集 MiniImageNet 和 CUB 上, 本文的方法也取得了最好的效果, 证实了方法的有效性.

后文将首先介绍小样本学习的背景与符号表示, 以及元学习的框架. 在深入探讨本文的 MCP 方法之前, 将给出使用一阶条件作为模型组合的直观解释. MCP 的具体训练策略、方法扩展将在第 4 节中说明. 最后是相关工作、实验和结论.

2 问题背景与符号表示

首先定义 K 样本 N 类分类问题 (K -Shot N -Way), 即在当前任务中, 共有 N 个不同的类别, 而每一个类别中有 K 个有标记样本. 训练集 $\mathcal{D}_{\text{train}}$ 中共有 NK 个样本, 对每一个样本 $(\mathbf{x}_i, \mathbf{y}_i)$, $\mathbf{x}_i \in \mathbb{R}^D$ 和 $\mathbf{y}_i \in \mathbb{R}^N$ 是一个向量表示, 对应类别的位置为 1, 其余为 0. 对于小样本学习任务, 一般考虑 K 很小的情况, 如 $K=1$ 或 $K=5$. 在训练阶段, 通过已知类别 (seen class, $\mathcal{S}$) 上的样本学习模型参数, 而在测试阶段, 将应用于未知类别 (unseen class, $\mathcal{U}$) 组成的小样本任务上.

假设所有的小样本学习任务使用同一个映射函数 $f(\mathbf{x}; \mathcal{D}_{\text{train}}) \rightarrow \mathbb{R}^N$ 对样本 $\mathbf{x}$ 进行预测. 该映射 f 刻画了不同任务之间的模型先验, 并且有多种实现方法. 例如 f 可以实现为通过某个共同初始化基于多次梯度下降得到的分类器^[16] 或是实现为基于非参数方法的最近邻规则^[13,14]. 为模拟小样本学习任务, 在训练过程中, 会从已知类别 $\mathcal{S}$ 中抽取任务, 利用“元学习”在抽取的任务上对映射 f 进行优化. 值得注意的是, 在 $\mathcal{S}$ 中抽取的任务会具有不同的类别组合, 因此, 要求 f 学习和具体类别无关的信息. 使用 $\mathcal{D}_{\text{train}}^{\mathcal{S}} \sim \mathcal{S}$ 表示从 $\mathcal{S}$ 中抽取的一个小样本任务训练集, 而 f 的性能可以使用同样方式抽

取的测试集 $\mathcal{D}_{\text{test}}^S \sim \mathcal{S}$ 来进行评估: 即对于 $\mathcal{D}_{\text{test}}^S$ 里的每一个样本, 使用 f 对其进行类别预测, 计算损失. 上标 $\mathcal{S}$ 表示强调该数据从 $\mathcal{S}$ 中抽取, 如果没有上标表示分析对 $\mathcal{S}$ 和 $\mathcal{U}$ 中抽取的数据子集都适用. 因此, 元学习对于小样本问题的目标函数如下:

$$\min_f \mathbb{E}_{(\mathcal{D}_{\text{train}}^S, \mathcal{D}_{\text{test}}^S) \sim \mathcal{S}} \left[\sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{test}}^S} \ell(f(\mathbf{x}; \mathcal{D}_{\text{train}}^S), \mathbf{y}) \right], \quad (1)$$

其中, $\ell(\cdot)$ 是损失函数, 如交叉熵, 用于度量预测结果与真实标记之间的区别. 通过优化上述函数, f 能够在抽取的不同的的小样本任务上进行良好的预测. 因此, 对于新类别构成的小样本任务, 也可以利用 f 获得分类器, 并对未来的样本进行预测.

3 基于一阶最优条件导出的模型分解形式

本节考虑参数化形式对映射 f 进行建模. 首先将上文的映射 f 分解为两个部分: 引入特征表示映射 $\phi: \mathbb{R}^D \rightarrow \mathbb{R}^d$ 与线性分类器 ($\mathbf{W} \in \mathbb{R}^{d \times N}, \mathbf{b} \in \mathbb{R}^N$). 其中, $\phi(\cdot)$ 将所有样本从原始 D 维输入空间映射到 d 维表示空间中; 线性分类器 $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_N]$ 的每一列对应当前任务 N 类中每一类的分类器, $\mathbf{b}$ 为偏移. 基于当前任务的训练集 $\mathcal{D}_{\text{train}}$, 分类器的优化目标可写为

$$(\mathbf{W}, \mathbf{b}) = \arg \min \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{\text{train}}} \ell(\mathbf{W}^T \phi(\mathbf{x}_i) + \mathbf{b}, \mathbf{y}_i) + \lambda \|\mathbf{W}\|_F^2, \quad (2)$$

其中 $\lambda > 0$ 是正则项的权重, $\|\cdot\|_F^2$ 表示矩阵的 Frobenius 范数. 基于上述目标函数, 可以获得最优分类器的一阶条件为

$$\sum_{\mathbf{y}_i = n} \nabla_{\mathbf{w}_n} \ell(\mathbf{W}^T \phi(\mathbf{x}_i) + \mathbf{b}, \mathbf{y}_i) \phi(\mathbf{x}_i) + \lambda \mathbf{w}_n = 0, \quad \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{\text{train}}} \nabla_{\mathbf{b}} \ell(\mathbf{W}^T \phi(\mathbf{x}_i) + \mathbf{b}, \mathbf{y}_i) = 0. \quad (3)$$

此处的 $\mathbf{y}_i = n$ 为索引当前任务第 n 类中样本的简化表示. 其中, $\nabla_{\mathbf{w}_n}$ 和 $\nabla_{\mathbf{b}}$ 分别为损失函数对分类器及偏移的导数. 式 (3) 中的条件表示“模型组成先验” (MCP), 即最优分类器可表示为特征映射后训练集的线性组合. 因此, 可以通过找到一组合适的样本权重来构建分类器, 并进行预测. 这种方式在邻域搜索^[13]和核方法^[19]中有类似的应用. 基于这种模型先验组合的概念, 在获取分类器之前, 需要寻找方法确定最优分类器的样本组合权重. 下文中, 主要考虑平方损失, 而由此获得的模型分解具有较强的可解释性. 定义 $\phi(\mathbf{X}) \in \mathbb{R}^{NK \times d}$ 为使用 ϕ 映射 $\mathcal{D}_{\text{train}}$ 之后的训练集矩阵, 而 $\mathbf{Y} \in \mathbb{R}^{NK \times N}$ 为对应的标记. 使用平方损失时,

$$\min_{\mathbf{W}, \mathbf{b}} \|\phi(\mathbf{X})\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\|_F^2 = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{\text{train}}} \|\mathbf{W}^T \phi(\mathbf{x}_i) + \mathbf{b} - \mathbf{y}_i\|_2^2, \quad (4)$$

其中 $\mathbf{1}$ 为元素全为 1 的向量, 其长度可通过上下文确定. 对分类器和偏移进行求导之后, 可以得到其闭式解:

$$\mathbf{W}^* = (\phi(\mathbf{X})^T \mathbf{H} \phi(\mathbf{X}))^{-1} \phi(\mathbf{X})^T \mathbf{H} \mathbf{Y}, \quad \mathbf{b}^* = \frac{1}{NK} (\mathbf{Y}^T \mathbf{1} - \mathbf{W}^T \phi(\mathbf{X})^T \mathbf{1}), \quad (5)$$

其中 $\mathbf{H} = \mathbf{I} - \frac{1}{NK} \mathbf{1}\mathbf{1}^T$ 是中心化矩阵. 观察式 (5) 中闭式解的结构, 最优分类器 $\mathbf{W}^*$ 包含两项, 分别为特征协方差矩阵的逆 $(\phi(\mathbf{X})^T \mathbf{H} \phi(\mathbf{X}))^{-1}$ (也称为精度矩阵, precision matrix) 和中心化之后每一个类的均值 $\phi(\mathbf{X})^T \mathbf{H} \mathbf{Y}$ (类别中心). 对于此种情况, 可以将小样本学习的难度归因于无法通过少量训练

图 1 (网络版彩图) 不同方法对比

Figure 1 (Color online) Comparison between different approaches. (a) ProtoNet; (b) MAML; (c) MCP.

样本估计出数据分布的统计信息 (特征关系以及类别中心). 而通过例如式 (5) 中的模型先验分解, 本文考虑使用不同任务之间的信息“协同估计”模型的各个组成部分, 再组合成当前任务的最优分类器. 此外, 特征的精度矩阵刻画出不同特征之间的条件独立关系^[20], 由于所有样本共享同样的特征表示空间, 因此可以通过已知类别 $\mathcal{S}$ 中抽取的任务来对该矩阵进行估计, 并直接用于未知类别 $\mathcal{U}$ 构建的任务中. 对于类中心的估计, 本文在元学习过程中考虑增加“一致性”正则项, 使得在每次任务的优化中, 通过当前少量样本估计的类中心和使用 $\mathcal{S}$ 中大量样本估计的类中心保持一致, 使得特征表示 ϕ 具有即使通过少量样本也能够准确反映类别信息的能力. 后文中将详细介绍 MCP 方法的思路具体的实现方法.

4 模型分解先验方法

本文在 MCP 思想的指导下学习当前任务最优分类器的每个组成部分. 首先介绍 MCP 方法, 包括在不同任务中学习共享的精度矩阵以及从已知类别中选择数据的统计信息. 接下来是相关方法的讨论. MCP 方法和相关方法的流程对比参见图 1.

4.1 MCP 方法

根据式 (5), 最优分类器有 3 个组成部分, 分别为样本的偏移、类别均值和特征的精度矩阵. 由于所有样本都使用相同的特征映射 ϕ , 本文在 $d \times d$ 半正定矩阵空间 $\mathcal{S}_d^+$ 中建模一个所有任务、类别共享的属性精度矩阵 $\mathbf{P}$. 后文中将详细介绍剩余两个组成部分如何估计.

偏移估计. 尽管最优偏移和输入样本以及最终分类器有关, 考虑到最优偏移的作用是对当前训练集样本进行中心化, 因此可以将对偏移 $\mathbf{b}$ 的估计转化为对当前任务中所有样本均值的估计. 本文使用长短期记忆模型^[21] (long-short term memory, LSTM) 来刻画基于当前训练集 $\mathcal{D}_{\text{train}}^{\mathcal{S}}$ 对样本中心估计的能力. 当前任务训练集的所有样本将依次输入 LSTM, 具体表示为 $\hat{\mathbf{b}} = \text{LSTM}(\{\phi(\mathbf{x}_i), \mathbf{x}_i \in \mathcal{D}_{\text{train}}^{\mathcal{S}}\}) \Rightarrow (\mathbf{o}_i, \mathbf{h}_i) = \text{LSTM}(\phi(\mathbf{x}_i), \mathbf{h}_{i-1}), \mathbf{x}_i \in \mathcal{D}_{\text{train}}^{\mathcal{S}}$. 元组 $(\mathbf{o}_i, \mathbf{h}_i)$ 为 LSTM 输出和隐藏层单元. 本文使用最后一层的隐藏层单元作为当前任务偏移的估计. 同时, 根据偏移描述了所有样本中心这一性质, 考虑增加正则项 $\Omega_1 = \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{train}}^{\mathcal{S}} \cup \mathcal{D}_{\text{test}}^{\mathcal{S}}} \|\hat{\mathbf{b}} - \phi(\mathbf{x}_i)\|_2^2$ 使得通过 LSTM 估计的偏移和使用当前任务训练集、测试集所有样本估计的中心一致.

类中心修正. 考虑到类中心对最优分类器的重要性, 直接使用当前任务的训练集估计各类中心可

不够准确. MCP 方法从不同任务中学习对当前少样本估计的类中心进行修正的能力, 在修正过程中同时考虑单个样本的属性和样本之间的关系. 本文使用双层双向长短期记忆模型^[22] (bidirectional LSTM, BiLSTM) 来刻画这种修正能力, 该模型同时考虑到了训练集所有样本的全局信息. 与此同时, 底层的 BiLSTM 层抽取类内信息. 对于第 n 类样本

$$(\vec{o}_k, \vec{h}_k) = \text{LSTM}(\phi(\mathbf{x}_k), \mathbf{h}_{k-1}), k = 1, \dots, K, \quad (\vec{o}_k, \vec{h}_k) = \text{LSTM}(\phi(\mathbf{x}_k), \mathbf{h}_{k+1}), k = K, \dots, 1,$$

其中, 索引 k 用来遍历当前类别中的 K 个样本. 最后第 n 类的表示引入了同类样本的所有信息, 定义为 $\mathbf{e}_n = \frac{1}{K} \sum_{\mathbf{y}_k=n} (\phi(\mathbf{x}_k) + \vec{o}_k + \vec{o}_k)$. 针对各类别调整后的中心表示, 一个高层 BiLSTM 用于考虑类别之间的信息并对类中心再次做修正.

$$(\vec{o}_n, \vec{h}_n) = \text{LSTM}(\mathbf{e}_n, \mathbf{h}_{n-1}), n = 1, \dots, N, \quad (\vec{o}_n, \vec{h}_n) = \text{LSTM}(\mathbf{e}_n, \mathbf{h}_{n+1}), n = N, \dots, 1.$$

因此, 第 n 类中心的最终估计为 $\hat{\mathbf{e}}_n = \mathbf{e}_n + \vec{o}_n + \vec{o}_n$. 类似于偏移估计的过程, 对于类别中心的估计, 可以考虑利用已知类别中的丰富信息进行正则, 即令使用 LSTM 估计的类中心和使用已知类别中该类别所有样本计算的均值保持一致. 由于在优化过程中, 样本的特征表示 ϕ 也随优化过程而变化, 且考虑到直接计算 $\mathcal{S}$ 中某个类别所有样本的均值计算量过大, 因此引入中间变量 $\mathbf{c}_n$ 记录 $\mathcal{S}$ 中每一个类别的类中心:

$$\Omega_2 = \sum_{n=1}^N \|\hat{\mathbf{e}}_n - \mathbf{c}_n\|_2^2, \quad \Omega_3 = \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{train}}^{\mathcal{S}} \cup \mathcal{D}_{\text{test}}^{\mathcal{S}}} \|\phi(\mathbf{x}_j) - \mathbf{c}_n\|_2^2.$$

通过 Ω_3 , $\mathbf{c}_n$ 记忆了不同任务中第 n 个类中每一个样本, 累积计算了 $\mathcal{S}$ 中每一个类的样本均值, 而 Ω_2 要求使用 LSTM 估计的类中心和统计得到的样本均值保持一致. 最终 MCP 的优化目标可概括为

$$\min_{\phi, \mathbf{P}} \mathbb{E}_{(\mathcal{D}_{\text{train}}^{\mathcal{S}}, \mathcal{D}_{\text{test}}^{\mathcal{S}}) \sim \mathcal{S}} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{test}}^{\mathcal{S}}} \ell(\mathbf{P}([\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N] - \mathbf{1}\hat{\mathbf{b}}^T)(\phi(\mathbf{x}) - \hat{\mathbf{b}}), \mathbf{y}) + \lambda(\Omega_1 + \Omega_2 + \Omega_3). \quad (6)$$

在上述目标中, 通过从已知类别 $\mathcal{S}$ 中抽样小样本任务的训练和测试集, 对特征表示函数 ϕ , 参数化模型的 3 个组成部分进行优化. 具体来说, 对于参数化模型需要优化的是不同任务之间共享的特征精度矩阵 $\mathbf{P}$, 对任务偏移进行估计的 LSTM, 以及对任务类中心进行修正的 BiLSTM 参数. 对中心进行修正就是为了考虑当前任务训练集中可能存在的异常或噪声样本, 增加整个预测机制的鲁棒性; 而使用 LSTM 是考虑到其对整个训练集样本之间关系的挖掘能力, 通过 LSTM 刻画样本之间的相对关系, 使得整个训练集的样本可以基于任务进行相应的特征变换, 以便后续的分类操作. MCP 方法的主要流程参见图 2.

4.2 MCP 方法扩展

在所有任务中共享同一个特征精度矩阵 $\mathbf{P}$ 潜在假设了不同的类别之间具有同样的特征相关性. 而针对某一个任务, 只需要当前任务具有类别包含的特征相关性, 直接利用 $\mathcal{S}$ 中所有类别估计的特征相关性信息并不是最优的选择. 因此, 本小节考虑如何基于当前任务的少量样本选择出对应的特征相关性矩阵.

基于特征精度矩阵和协方差矩阵的联系: 不同条件概率下的协方差矩阵可以进行直接加权求和, 而精度矩阵是对应协方差矩阵的逆. MCP⁺ 方法考虑从已知类别 $\mathcal{S}$ 中学习并组合出最优的协方差矩阵, 从而得到对应任务的精度矩阵. 从任务层面考虑, 假设存在某一空间, 其中的每一个元素表示一个 N 类 K 样本任务, 且相似的元素具有相似的特征结构关系. 将整个空间划分为 $|\mathcal{S}|$ 个区域, 其中

图 2 (网络版彩图) MCP 方法流程图

Figure 2 (Color online) Illustration of the MCP approach

$|\mathcal{S}|$ 为已知类别数目. 使用元组 $(\mathbf{a}_s \in \mathbb{R}^d, \mathbf{P}_s \in \mathbb{R}^{d \times d})$ 代表一个区域, 其中, $\mathbf{a}_s$ 是该区域的特征表示, 而 $\mathbf{P}_s$ 是这块区域的任务对应的特征协方差矩阵. 在训练过程中, 基于 MCP 的结果可使用 $\mathcal{S}$ 中每一个类别的均值设置 $\mathbf{a}_s$, 对应类别的协方差设置 $\mathbf{P}_s$. 对于某一个小样本任务, 其特征协方差矩阵可以通过以下方式获得: 首先使用新的映射 ψ (基于当前 ϕ) 产生的特征表示得到当前任务的向量表示 $\boldsymbol{\tau} = \sum_{\mathbf{x}_i \in \mathcal{D}_{\text{train}}^s} \psi(\phi(\mathbf{x}_i))$, 即经过 ψ 再次映射之后的任务内所有样本的均值. 文献 [23] 中证明这种均值表示能够反映出原始任务样本集合的性质. 使用该任务的向量表示将其嵌入到搜索空间中; 其次利用向量表示找到相关的基, 并利用相关基元素对应的协方差矩阵辅助当前任务特征精度矩阵的估计:

$$\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_{|\mathcal{S}|}], \quad \alpha_s = \text{sigmoid}(\mathbf{a}_s^T \mathbf{M} \boldsymbol{\tau}), \quad \mathbf{P}^{-1} = \left(\mathbf{C}_0 + \sum_s \alpha_s \mathbf{P}_s \right)^{-1},$$

其中 $\mathbf{C}_0 = \mathbf{P}^{-1}$ 是所有任务具有的初始协方差矩阵, 而最终的精度矩阵是 $\mathbf{C}_0$ 和不同局部协方差矩阵加权求和的逆. $\mathbf{M} \in \mathbb{R}^{d \times d}$ 是该空间中任务相似性的度量矩阵. 如果当前任务和某个基任务足够相似, 那就会有较大的权重偏向这个基, 同时该基任务对应的协方差矩阵也会较大程度地用于构建当前任务的特征协方差矩阵.

直接计算矩阵加权求和的逆会有较高的复杂度. 因此, 本文考虑对该公式进行 Taylor 展开. 由于对矩阵 $\mathbf{A}, \mathbf{B}$ 有 $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1} - \dots$, 因此

$$\left(\mathbf{C}_0 + \sum_s \alpha_s \mathbf{P}_s \right)^{-1} \approx \mathbf{P} - \sum_s \alpha_s \mathbf{P} \mathbf{P}_s \mathbf{P} = \mathbf{P} - \sum_s \text{sigmoid}(\mathbf{a}_s^T \mathbf{M} \boldsymbol{\tau}) \mathbf{P} \mathbf{P}_s \mathbf{P}. \quad (7)$$

上述推导避免了对大量的对逆矩阵的求导过程.

4.3 方法讨论

在 MCP 方法中, 主要使用 LSTM 对训练集中的统计量 (偏移、均值等) 进行估计. 对于不同顺序的输入, LSTM 本身会输出不同的结果, 无法直接用于处理原始无序的任务. 而本方法中, LSTM 主要用于刻画训练集样本之间的相对关系, 并不需要样本本身的序关系. 同时实验结果验证了使用 LSTM 的有效性. 文献 [24] 中也提出利用最小二乘的闭式解简化模型的构造. 相比于文献 [24], MCP 方法首

先强调了特征精度矩阵的重要性, 从后续实验中可发现从不同任务中学习共享的特征精度矩阵有助于对未知类别的小样本学习. 此外, MCP 方法使用正则项保持模型估计变量和样本统计变量的一致性, 增强从不同小样本任务中模型的估计能力. MAML^[16] 方法考虑针对每一个小样本分类任务进行固定步数的梯度下降获得对应的分类器. 和 MAML 相比, MCP 方法利用平方损失的闭式解避免了由梯度下降算法带来的不稳定性, 并在测试过程中大量降低了计算的负担.

5 相关工作

通过过往经验的累积, 人类能够通过极少的样本学习新的概念^[8,25,26]. 本文将这种跨任务、累积获得的经验概括为“先验知识”. 机器学习算法通过从不同任务中抽取模型先验, 从而快速适配仅有少量样本的新任务. 以往工作对模型的先验有不同的实现方式.

从非参数化方法的角度, 模型的先验可以通过样本的特征表示实现, 这种特征表示反映了样本间相似、不相似的关系, 因此仅需要使用最近邻分类器即可完成新任务. 文献 [12] 在小样本字符识别的任务中验证特征表示的重要性; MatchingNet^[13] 通过元学习的方法能够从已知类别的数据中获得高质量的特征表示. 后续方法采用其他策略增强样本特征表示的能力: 如 ProtoNet^[14] 中利用同类样本的中心对位置样本进行判定; 文献 [15] 引入全局的距离度量指导小样本学习. 从参数化的角度, 小样本学习问题的关键在于如何基于已有的样本得到对应的分类器, 并且要防止过拟合问题. MAML^[16] 提出了一种模型无关的元学习框架, 即从不同的任务中学习一个共享的模型初始化, 在一个新的任务中, 基于该初始化做固定步数的梯度下降即可得到较好的分类器; 文献 [18,27] 等一系列方法使用元学习学习小样本任务的优化策略, 如步长的判定、梯度方向的修正等. 文献 [28,29] 从理论上证明了元学习的有效性. 其他类型的先验在已有工作中也被考虑: 如生成数据的能力^[3,30]、逻辑推理能力^[31]、利用跨模态信息的能力^[32,33]. 本文强调了模型组合方式对模型先验的影响, 通过将分类器分解为若干部分, 可以容易地基于其相应的物理意义对每个部分进行元学习, 从而达到优异的小样本分类性能.

迁移学习^[34~36] 考虑从源领域 (source domain) 抽取知识并应用于目标领域 (target domain), 一般源领域具有大量的数据, 但和目标领域的分布不同, 目标领域一般只有少量的有标记数据, 但在某些情况下会有大量的无标记数据. 将本文已知类别和未知类别的数据分别对应于迁移学习中的源领域和目标领域. 本文算法利用元学习的框架进行对小样本问题的求解, 相当于在源领域中模拟出大量的小样本学习任务, 学习对应的分类器生成策略, 并假设这种生成策略和领域无关, 遇到目标领域的小样本分类任务时就可以直接利用这种学到的策略; 而迁移学习要同时考虑源领域和目标领域的数据, 一般需要大量的目标领域的无标记样本, 以挖掘两个领域之间分布的联系, 并从源领域中选择出合适的样本辅助目标领域任务的学习. 小样本学习类似于一种基于假设的模型迁移 (hypothesis transfer). 通过特定的元学习策略学习可被迁移的模型配制 (即本文中生成任务分类器的策略).

小样本学习方法在其他任务中也有相关应用. 例如, 可用于处理具有长尾属性的类别不平衡分类问题^[37]、域适应^[38]、模仿学习^[39]、概率密度估计^[40] 和零样本学习^[14].

6 实验

在实验中, 本文首先展示了 MCP 方法在人造数据上学习任务之间相关性的能力; 此后, MCP 方法在两个典型数据集上和已有方法进行对比, 并取得了优异的结果, 验证了 MCP 算法处理小样本问题的有效性.

6.1 数据集与实现细节

本文主要实现基于已有论文的设定^[13~15],使用4层卷积网络来实现特征抽取函数 ϕ :在每一个层中,包含通道大小为64、卷积核大小为 3×3 的卷积层,批量归一化层(batch normalization, BN)^[41],ReLU激活层和 2×2 最大池化层.与以前的方法不同,本文在网络最后的输出层后增减了一个全局平均池化层,因此 ϕ 的输出保持在64维.全局池化层不仅可以很好地处理图片不同局部的空间变换,而且还可以大大减轻优化负担.模型中的所有长短期记忆模型LSTM单元都使用64维隐藏状态.该架构的复杂性几乎与MatchingNet^[13]相同.对于MCP⁺,高阶映射函数 ψ 由具有ReLU激活的双层全连接网络实现,隐藏层维度为128.所有算法均使用随机梯度下降ADAM^[42]优化器进行优化.优化过程中最大迭代轮数设置为30000.

MiniImageNet. MiniImageNet数据集^[13]是ILSVRC-12数据集^[11]的一个子集.该数据集共有100个类别,每一类有600个样本.评估过程中,本文遵循文献[18]的数据分割方式,使用64个类作为已知类别进行模型参数的学习,16个类用于参数选择,20个类用于最终模型的测试.所有图片首先调整为 $3 \times 84 \times 84$ 大小.在优化过程中,本文使用验证集来选择参数,并记录最佳模型进行测试.优化之前,本方法在训练集已知类别上预训练一个64类的分类器,用其网络给特征网络 ϕ 的权重做初始化.

CUB. Caltech-UCSD Birds 200-2011数据集^[43]一般用于测试模型的细粒度分类能力,其中含有200种鸟类共11788幅图像.由于类之间的相似性,当给出有限的训练样本时,产生的分类问题更加困难.在图片预处理过程中,本文首先根据每幅图片给定的鸟的边框裁剪图像,然后将其大小调整为 $3 \times 64 \times 64$.在算法评估过程中,本文使用前100类作为已知类别,以下50种鸟类作为验证类别,剩余50类用于测试.该设置遵循文献[15],由于该论文没有公开类别划分,本文在上述类别划分上重现了所有对比方法.

本文对所有数据集使用相同的评估协议.对于模型参数训练和评估过程中的每一个 N 类 K 样本问题,都使用相同的方式构建.首先从相关类别集合(已知类别或未知类别)中抽样出 N 个类别,并从每个类中抽取 K 个样本作为当前任务的训练集,每个类别中继续抽取15个不重叠的样本作为验证集,用于评估当前任务上分类器的能力.在MiniImageNet和CUB对模型评估时,分别抽取600和1000次任务,记录其均值和置信度.

6.2 人造数据结果验证与分析

本小节展示了MCP方法在人造数据集上学习小样本任务之间关联性的能力.

假设总共有200个类,每一个类使用64维的正态分布表示,所有类别和数据按以下方式生成:首先,从空间中随机抽取200个向量作为类中心.所有类别特征的协方差矩阵都由一个基矩阵生成,假设不同的类具有不同的均值但具有相同的协方差矩阵.协方差基矩阵由8个 8×8 的小矩阵按照分块对角方式组成.其中每一个子矩阵首先随机生成,再通过自乘的方式得到半正定的矩阵并嵌入协方差基矩阵中.本文使用前100个类进行训练(用于参数学习),后50个类用于模型验证(参数选择),最后50个类用于测试(模型评估).此时的特征抽取函数 ϕ 假定为单位映射,MCP中的其他组件的设置与前文描述的相同.本小节主要考虑30类单样本和5样本分类任务.在模型训练或评估时,每一个小样本任务均从对应的类别集合中先确定30个类别对应的分布参数,再基于确定的分布抽样出样本.模型训练过程中,最大迭代次数设置为5000,如果验证集上模型的最优准确率在最近20个迭代内没有增加,则学习率将减少一半.

表 1 人造数据集不同设定下测试结果

Table 1 Comparison results over two settings on synthetic dataset

Method	Noise	30-Way 1-Shot (%)	30-Way 5-Shot (%)	Method	Noise	30-Way 1-Shot (%)	30-Way 5-Shot (%)
NN	N	9.82 ± 0.10	13.91 ± 0.11	NN	Y	9.61 ± 0.10	13.46 ± 0.11
Proto	N	9.82 ± 0.10	25.35 ± 0.15	Proto	Y	9.61 ± 0.10	24.97 ± 0.14
FCE	N	12.80 ± 0.12	63.08 ± 0.25	FCE	Y	10.09 ± 0.10	48.53 ± 0.22
MCP	N	91.77 ± 0.19	95.46 ± 0.12	MCP	Y	89.53 ± 0.20	94.88 ± 0.12

图 3 (网络版彩图) 真实精度矩阵 (a) 与 MCP 算法在较弱 (b) 和较强 (c) 噪声环境下估计的精度矩阵

Figure 3 (Color online) The ground-truth precision matrix (a) and the ones estimated by MCP with weak (b) and strong (c) noise

由于不需要为该人造数据提取特征, 本文将 MCP 方法与以下几个基准算法进行比较, 如表 1 所示. 前两种方法模拟了 MatchingNet^[13] 和 ProtoNet^[14] 中使用的分类策略. 具体来说, NN 表示使用最近邻搜索方法对每一个任务的测试样本进行分类, Proto 表示使用训练集中每一个类的类中心进行分类 (也被称为最近类均值分类器^[44]). 对于全上下文嵌入 (fully conditional embedding, FCE)^[13], 本文遵循原始论文的方法, 学习额外的 BiLSTM 参数.

本文研究了两种情况. 首先在正常情况下, 所有任务共享相同的协方差矩阵, 这与 MCP 中的假设一致. 本文还考虑了一个有噪声的情况, 即在每个样本的采样期间, 在真实的协方差矩阵上添加值在 $[0, 0.1]$ 之间的随机噪声, 试图干扰原始协方差矩阵的分块对角结构. 精度矩阵具有与协方差矩阵相同的分块对角结构. 图 3(a) 展示了所有任务下真实的精度矩阵, 图 3(b) 和 (c) 分别是 MCP 模型在正常和噪声设定下估计出的精度矩阵. 在图 3(b) 和 (c) 中可以清楚地找到这种类型的模型结构, 这意味着即使在模式不明确的嘈杂环境中, 本文的 MCP 方法也能够学习这种共享的结构. 从表 1 中的数值结果来看, 本文的 MCP 方法在所有情况下都可以达到最高精度. 直接使用现有特征对于小样本问题很难. FCE 的提升可能来自其对样本上下文信息的考虑. MCP 的成功源于其对上下文和结构的明确建模, 以及对元训练中数据统计信息的利用. 具体来说, 该人造数据的构造直接考虑到 MCP 方法的特点, 满足 MCP 方法的假设, 即所有的小样本学习任务均由同一个特征变换矩阵生成. MCP 方法能够在学习过程中明确学习到不同任务上特征的关联性, 并利用这种任务之间特征的关联性辅助后续的小样本分类. 其余方法并没有假设这种任务特征的结构信息, 因此并不能取得好的分类效果.

6.3 实际数据结果对比与分析

MiniImageNet 数据集上的比较结果在表 2 中. 本文比较了当前主流的几种方法. 由于本文遵循

表 2 MiniImageNet 数据集上不同方法的测试均值及 95% 置信度

Table 2 Few-shot classification mean accuracy on MiniImageNet dataset, together with 95% confidence intervals

	5-Way 1-Shot (%)	5-Way 5-Shot (%)		5-Way 1-Shot (%)	5-Way 5-Shot (%)
Baseline NN	41.08 $\pm$ 0.70	51.04 $\pm$ 0.65	MAML ^[16]	48.70 $\pm$ 1.84	63.11 $\pm$ 0.92
MatchingNet ^[13]	43.40 $\pm$ 0.78	51.09 $\pm$ 0.71	Siamese	48.42 $\pm$ 0.79	—
MatchingNet (FCE) ^[13]	43.56 $\pm$ 0.84	55.31 $\pm$ 0.73	mAP-SSVM ^[15]	50.32 $\pm$ 0.80	63.94 $\pm$ 0.72
Meta-LSTM ^[18]	43.44 $\pm$ 0.77	60.60 $\pm$ 0.71	mAP-DLM ^[15]	50.28 $\pm$ 0.80	63.70 $\pm$ 0.70
RelationNet ^[45]	50.40 $\pm$ 0.80	65.30 $\pm$ 0.70	Meta-CF ^[24]	48.70 $\pm$ 0.60	65.50 $\pm$ 0.60
ProtoNet ^[14]	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	ProtoNet Pool	49.21 $\pm$ 0.79	66.80 $\pm$ 0.68
MCP	51.24 $\pm$ 0.82	67.37 $\pm$ 0.67	MCP ⁺	51.27 $\pm$ 0.81	66.93 $\pm$ 0.63

表 3 CUB 数据集上不同方法的测试均值及 95% 置信度

Table 3 Few-shot classification mean accuracy on CUB, together with 95% confidence intervals

	1-Shot 5-Way (%)	1-Shot 20-Way (%)		1-Shot 5-Way (%)	1-Shot 20-Way (%)
Siamese	47.17 $\pm$ 0.62	21.35 $\pm$ 0.21	mAP-DLM	50.07 $\pm$ 0.64	22.79 $\pm$ 0.22
MatchingNet	46.58 $\pm$ 0.66	16.61 $\pm$ 0.18	ProtoNet	48.89 $\pm$ 6.30	22.92 $\pm$ 2.17
MAML	31.04 $\pm$ 0.43	10.57 $\pm$ 0.13	ProtoNet Pool	48.96 $\pm$ 0.65	21.18 $\pm$ 0.21
MCP	50.95 $\pm$ 0.65	25.30 $\pm$ 0.25	MCP ⁺	55.85 $\pm$ 0.72	26.39 $\pm$ 0.26

文献 [13~15] 相同的评估、数据处理方法, 因此引用文献 [15] 中的实验结果作为对比. 其中, ProtoNet Pool 是本方法的基准对比方法, 实现基于 ProtoNet^[14], 但有两点不同. 首先, 在特征表示网络的末尾增加全局平均池化层. 同时, 在元学习之前, 在已知类别上优化多分类问题, 用特征层的权重初始化网络. ProtoNet Pool 获得与原始 ProtoNet 类似的结果, 特别是在单样本任务 (1-shot) 设定中. 这意味着池化层在降低计算开销的同时能保持原始的性能. 本文的 MCP 方法在所有对比方法中能够取得较好的分类结果, 并比 ProtoNet Pool 具有更好的分类精度. 和 ProtoNet Pool 的比较结果证明了 MCP 的有效性不是来自最终的池化层或预训练的权重. MCP⁺ 算法能取得更好的小样本分类效果, 说明考虑精度矩阵的特定生成模式对小样本分类更有帮助. 在测试过程中, 本文也考虑 20 类分类问题. 在 1-Shot 20-Way 的小样本分类任务上, MCP 和 MCP⁺ 的分类准确率分别为 22.68 ± 0.72 和 22.68 ± 0.63 ; 在 5-Shot 20-Way 的小样本分类任务上, 两种方法的分类性能分别为 37.62 ± 0.65 和 37.85 ± 0.68 . CUB 数据集上的比较结果在表 3 中, 在该数据上, 本文提出的 MCP 方法仍然能取得最好的效果.

7 结语

本文针对小样本学习问题进行探索, 将小样本学习的主要难点归因于“模型先验”的学习. 基于损失函数的一阶最优解条件, 本文提出“模型分解先验”这一概念, 基于平方损失函数首先将最优模型分解为不同分块, 然后利用元学习方法对不同的分块进行估计, 最后综合成最优的分类器. 这种对模型分解的思路具有较强的可解释性, 利用其对应的物理意义, 一方面可以简化估计方法, 同时也便于确定在不同任务间共享或每个任务独特的成分. 此外, 本文的 MCP⁺ 方法能够利用已知类别的信息. 本文方法在人造数据集上展示出了可解释性, 并在基准数据上获得最好的效果. 未来的工作包括将这种模型分解先验的概念扩展到一般损失函数, 并引入其他模态的信息辅助小样本学习.

致谢 作者感谢在南加州大学沙飞老师研究组访问期间沙飞老师以及组里同学提供的帮助。

参考文献

- 1 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. ArXiv:1409.0473
- 2 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012, 1097–1105
- 3 Wang Y-X, Girshick R, Hebert M, et al. Low-shot learning from imaginary data. 2018. ArXiv:1801.05401
- 4 Tan X Y, Chen S C, Zhou Z H, et al. Face recognition from a single image per person: a survey. Pattern Recogn, 2006, 39: 1725–1745
- 5 Finn C, Yu T H, Zhang T H, et al. One-shot visual imitation learning via meta-learning. In: Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, 2017. 357–368
- 6 Karlinsky L, Shtok J, Tzur Y, et al. Fine-grained recognition of thousands of object categories with single-example training. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 965–974
- 7 Zhou Z H. Learnware: on the future of machine learning. Front Comput Sci, 2016, 10: 589–590
- 8 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. Science, 2015, 350: 1332–1338
- 9 Patricia N, Caputo B. Learning to learn, from transfer learning to domain adaptation: a unifying perspective. In: Proceedings of Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 1442–1449
- 10 Li Z G, Zhou F W, Chen F, et al. Meta-SGD: learning to learn quickly for few shot learning. 2017. ArXiv:1707.09835
- 11 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252
- 12 Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, 2015
- 13 Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3630–3638
- 14 Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 4080–4090
- 15 Triantafillou E, Zemel R S, Urtasun R. Few-shot learning through an information retrieval lens. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 2252–2262
- 16 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 1126–1135
- 17 Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. 2018. ArXiv:1803.02999
- 18 Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of International Conference on Learning Representations, 2017
- 19 Schölkopf B, Smola A J. Learning with Kernels. Cambridge: The MIT Press, 2001
- 20 Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. Berlin: Springer, 2001
- 21 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
- 22 Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE Trans Signal Process, 1997, 45: 2673–2681
- 23 Zaheer M, Kottur S, Ravanbakhsh S, et al. Deep sets. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 3394–3404
- 24 Bertinetto L, Henriques J F, Torr P H S, et al. Meta-learning with differentiable closed-form solvers. 2018. ArXiv:1805.08136
- 25 Li F-F, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell, 2006, 28: 594–611
- 26 Lake B M, Salakhutdinov R, Gross J, et al. One shot learning of simple visual concepts. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, 2011
- 27 Andrychowicz M, Denil M, Colmenarejo S G, et al. Learning to learn by gradient descent by gradient descent. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3981–3989

- 28 Maurer A. Transfer bounds for linear feature learning. *Mach Learn*, 2009, 75: 327–350
- 29 Maurer A, Pontil M, Romera-Paredes B. The benefit of multitask representation learning. *J Mach Learn Res*, 2016, 17: 1–32
- 30 Hariharan B, Girshick R B. Low-shot visual recognition by shrinking and hallucinating features. In: *Proceedings of IEEE International Conference on Computer Vision*, Venice, 2017. 3037–3046
- 31 Dai W Z, Muggleton S, Wen J, et al. Logical vision: one-shot meta-interpretive learning from real images. In: *Proceedings of the 27th International Conference on Inductive Logic Programming*, Orléans, 2017. 46–62
- 32 Wang P, Liu L Q, Shen C H, et al. Multi-attention network for one shot learning. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. 6212–6220
- 33 Shyam P, Gupta S, Dukkipati A. Attentive recurrent comparators. In: *Proceedings of the 34th International Conference on Machine Learning*, Sydney, 2017. 3173–3181
- 34 Pan S J, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*, 2010, 22: 1345–1359
- 35 Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning. *J Big Data*, 2016, 3: 9
- 36 Day O, Khoshgoftaar T M. A survey on heterogeneous transfer learning. *J Big Data*, 2017, 4: 29
- 37 Wang Y X, Ramanan D, Hebert M. Learning to model the tail. In: *Proceedings of Advances in Neural Information Processing Systems*, 2017. 7032–7042
- 38 Motiian S, Jones Q, Iranmanesh S M, et al. Few-shot adversarial domain adaptation. In: *Proceedings of Advances in Neural Information Processing Systems*, 2017. 6673–6683
- 39 Yu T H, Finn X, Xie A N, et al. One-shot imitation from observing humans via domain-adaptive meta-learning. 2018. ArXiv:1802.01557
- 40 Reed S E, Chen Y T, Paine T, et al. Few-shot autoregressive density estimation: towards learning to learn distributions. 2017. ArXiv:1710.10304
- 41 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 2015. 448–456
- 42 Kingma D P, Ba J. ADAM: a method for stochastic optimization. 2014. ArXiv:1412.6980
- 43 Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report, 2011
- 44 Mensink T, Verbeek J, Perronnin F, et al. Distance-based image classification: generalizing to new classes at near-zero cost. *IEEE Trans Pattern Anal Mach Intell*, 2013, 35: 2624–2637
- 45 Sung F, Yang Y X, Zhang L, et al. Learning to compare: relation network for few-shot learning. 2017. ArXiv:1711.06025

Few-shot learning via model composition

Han-Jia YE & De-Chuan ZHAN*

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

* Corresponding author. E-mail: zhandc@lamda.nju.edu.cn

Abstract Although achieve inspiring performance in many real-world applications, machine learning methods require a huge amount of training examples to obtain an effective model. Considering the effort collecting labeled training data, the few-shot learning, i.e., learning with budgeted training set, is necessary and useful. Model prior, e.g., the feature embedding, initialization, and configuration, is the key to the few-shot learning. This study metalearns such prior from seen classes and apply the learned prior over few-shot task on unseen classes. Meanwhile, based on the first order optimal condition of the objective, the model composition prior (MCP) is stressed to decompose the model prior and estimate each component. The composition strategy improves the explainability, while guiding the shared and specific parts among those few-shot tasks. We verify the ability of our approach to recover task relationship over the synthetic dataset, and our MCP method achieves better results on two benchmark datasets (MiniImageNet and CUB).

Keywords few-shot learning, meta-learning, model prior, model composition

Han-Jia YE was born in 1991. He received his B.Sc. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2013. Currently, he is working towards a Ph.D. degree with the National Key Lab for Novel Software Technology, Department of Computer Science & Technology at Nanjing University, China. His research interests lie primarily in machine learning, including metric learning, multi-modal/multi-task learning,

meta-learning, and semantic mining.

De-Chuan ZHAN was born in 1982. He received his Ph.D. degree in computer science from Nanjing University, China, in 2010, and became a faculty member in the Department of Computer Science and Technology at Nanjing University. He is currently an associate professor in the Department of Computer Science and Technology at Nanjing University. His research interests majorly include machine learning, data mining, and mobile intelligence.