计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220922
61(1): 196−208, 2024

基于注意力增强元学习网络的个性化联邦学习方法

高雨佳[1,2] 王鹏飞[1] 刘 亮[2,3] 马华东[1,2]

[1](北京邮电大学计算机学院(国家示范性软件学院) 北京 100876)

[2](智能通信软件与多媒体北京市重点实验室(北京邮电大学) 北京 100876)

[3](北京邮电大学人工智能学院 北京 100876)

(gaoyujia@bupt.edu.cn)

Personalized Federated Learning Method Based on Attention-Enhanced Meta-Learning Network

Gao Yujia[1,2], Wang Pengfei[1], Liu Liang[2,3], and Ma Huadong[1,2]

[1](*School of Computer Science* (*National Pilot Software Engineering School*), *Beijing University of Posts and Telecommunications*, *Beijing* 100876)

[2](*Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia* (*Beijing University of Posts and Telecommunications*), *Beijing* 100876)

[3](*School of Artificial Intelligence*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876)

Abstract Federated learning is a distributed machine learning framework, which enables clients to conduct model training without transmitting their data to the servers. It is used to solve the dilemma of data silos and data privacy. It can work well on clients with similar data characteristics and distribution. However, in many scenarios, the differences of data distribution cause difficulties in global model training. Therefore, personalized federated learning is proposed as a new federated learning paradigm. It aims to guarantee the effectiveness of client personalized models through the collaboration between clients and the servers. Intuitively, providing a tighter collaboration for clients with similar data characteristics and distribution can facilitate the construction of personalized models. However, due to the invisibility of client data, it is a challenge to extract client features at a fine-grained level and define the collaborative relationships between clients. In this paper, we design an attention-enhanced meta-learning network (AMN) to address this issue. AMN can utilize model parameters as features and train the meta-learning network to provide an additional meta-model for each client to automatically analyze client feature similarity. According to two-layers framework of AMN, a trade-off between clients' personality and commonality can be reasonably achieved, and a hybrid model with useful information from all clients is provided. Considering the need to maintain both the meta-model and the client's local model during the training process, we design an alternative training strategy to perform the training in an end-to-end manner. To demonstrate the effectiveness of our method, we conduct extensive experiments on two benchmark datasets and eight baseline methods. Compared with the existing best-performing personalized federated learning methods, our method improves the accuracy rate by 3.39% and 2.45% on average in two datasets.

Key words federated learning; attention mechanism; deep learning; meta-learning; distributed machine learning

收稿日期: 2022-11-03; **修回日期**: 2023-03-22

基金项目: 国家自然科学基金项目(61932013, 62061146002, 62225204); 国家 111 计划项目(B18008)

This work was supported by the National Natural Science Foundation of China (61932013, 62061146002, 62225204) and the National 111 Project (B18008).

通信作者: 刘亮(liangliu@bupt.edu.cn)

摘　要　联邦学习作为一种分布式机器学习框架，客户端可以在不向服务器传输数据的情况下进行全局模型训练，解决了数据分散和数据隐私的问题. 联邦学习可以在具有相似数据特征和分布的客户端上很好地工作. 但是在很多场景中，客户端数据在分布、数量和概念上的不同，造成了全局模型训练困难. 为此，个性化联邦学习作为一种新的联邦学习范式被提出，它旨在通过客户端与服务器的协作来保证客户端个性化模型的有效性. 直观来讲，为具有相似数据特征和分布的客户端提供更紧密的协作关系可以有利于个性化模型的构建. 然而，由于客户端数据的不可见性，如何细粒度地提取客户端特征，并定义它们之间的协作关系是一个挑战. 设计了一个注意力增强元学习网络（attention-enhanced meta-learning network，AMN）来解决这个问题. AMN 可以利用客户基础模型参数作为输入特征，训练元学习网络为每个客户端提供一个额外的元模型，自动分析客户特征相似性. 基于双层网络设计，有效地实现客户端个性与共性的权衡，提供了包含有益客户信息的融合模型. 考虑到训练过程中需要同时训练元学习网络和客户本地基础网络，设计了一种交替训练策略，以端到端的方式进行训练. 为了证明该方法的有效性，在 2 个基准数据集和 8 种基准方法上进行了大量实验，相较于现有表现最优的个性化联邦学习方法，该方法在 2 个数据集中平均分别提升了 3.39% 和 2.45% 的模型性能.

关键词　联邦学习；注意力机制；深度学习；元学习；分布式机器学习

中图法分类号　TP18

联邦学习[1]作为一种分布式机器学习框架，客户端可以在不向服务器传输数据的情况下进行全局模型训练，它解决了数据分散和数据隐私的问题. 与单独工作的个体相比，联邦学习通过客户端和服务器的协作训练，获得了更好的机器学习性能. 同时，可控制的传输间隔降低了通信负载，对于时间敏感的任务，如医疗监测[2]和流量预测[3]，联邦学习可以显著减少训练延迟. 这种方法可以在具有相似数据特征和分布的客户群体中很好地工作. 但是，在许多应用程序场景中，客户端的数据集在分布、数量和概念[4]上可能不同，因此有一些限制. 这使得粗糙的全局协作在不考虑单个客户端数据属性的情况下无法获得良好的性能. 因此，在全局协同训练的基础上，为每个客户并行训练个性化模型就变得十分必要.

这种将个性化融入到联邦学习的方式被称为个性化联邦学习[5]. 常用的方法是取一个统一初始值（如全局模型），在本地训练过程中对其进行微调，以实现对客户端数据特征的适应性[6-7]. 由于通常情况下，一个模型不能适用于所有客户端，这种方式缺乏灵活性，难以解决客户的异质性问题. 同时，全局模型的粗融合也会影响全局模型的性能. 为此，我们希望探索一种更优的方式，为每个客户端提供不同的协作关系. 具体来讲，促进具有相似特征和数据分布的客户端之间的协作，并疏远具有广泛不同特征的客户端之间的协作. 因此，对于给定的客户端，如何自动提取与其类似的客户端，是一个具有挑战的问题. 此外，对于服务器-客户端框架，需要设计合理的协作训练方法.

为了解决以上问题，本文提出了一种个性化联邦学习框架 FedAMN，如图 1 所示. 和传统联邦学习架构一样，FedAMN 是由 1 个服务器和多个客户端组成的. 不同在于，FedAMN 在每个客户端均分布部署了一个注意力增强元学习网络（attention-enhanced meta-learning network，AMN），来实现客户端间协作关系的生成. AMN 以客户端的模型参数作为输入特征，学习一个元模型来分析不同客户端的相似程度，它可以自动筛选对目标客户端有益的模型参数信息，从而提升客户端个性化模型的性能. 这种方法利用客户端的本地数据进行元模型和本地个性化模型的训练，解决了元模型缺乏监督信号而难以训练的问

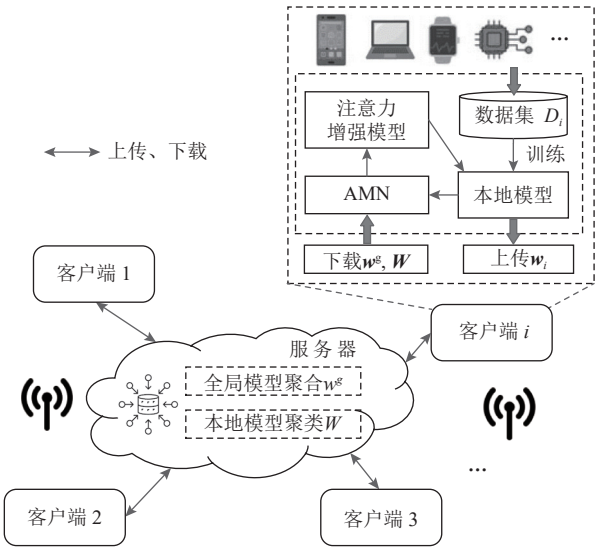


Fig. 1　Framework illustration of FedAMN

图 1　FedAMN 框架示意图

题. 并且因为数据不需要上传至服务器，客户端的数据隐私可以得到保护.

具体来讲，给定一个目标客户端，AMN 通过注意力机制，利用其他客户端的本地模型参数分析之间的相似性，训练元模型. 用元模型计算的注意力分数构建一个新的注意力增强模型，实现客户端个性与共享的平衡. 通过这样的设计，可以增强训练的灵活性，满足对客户任务精准决策的需求. 同时，为了减少 AMN 带来的额外的计算量增加，我们设计了一种渐进式的训练模式. 在开始的通信轮次中，仅对客户端的本地模型进行训练，当全局模型准确率达到稳定后，AMN 再加入训练. 在原本的客户模型基础上，进一步提升个性化模型的效果. 并且，考虑到客户端需要同时训练 AMN 和客户端的本地基础网络，提出了一种交替训练的策略，以端到端的方式对 2 个网络进行训练，最终实现客户端的个性化注意力增强模型的生成.

为了证明 FedAMN 的有效性，我们在 2 个基准数据集和 8 种基准方法上进行了大量实验. 与现有最先进的个性化联邦学习方法相比，我们的方法分别提高了至少 3.39% 和 2.45% 的模型性能. 此外，为了证明 AMN 中注意力机制的有效性，我们对 AMN 的双层网络结构进行了消融实验，并在真实世界数据集上，对 AMN 各层的注意力分数进行了可视化分析实验.最后，设计了不同数据分布和数据特征的客户端中的对比实验，探究客户端异质性对训练的影响.

总的来说，本文的贡献有 3 方面：

1) 将个性化联邦学习中的协作关系生成问题形式化为一个分布式元学习网络训练任务，并设计了一种注意力增强元学习网络(AMN)来解决客户端的本地模型的个性化问题；

2) 提出了 FedAMN 框架，采用渐进式的训练模式，缓解 AMN 引入带来的计算量增加问题，以交替训练的策略实现客户端个性化注意力增强模型的生成；

3) 对不同推理任务进行大量实验，证明我们的方法显著优于现有最先进的个性化联邦学习基线方法，并可以应用于许多具有隐私性的分布式数据个性化建模场景.

1 相关工作

本文的工作受到个性化联邦学习和元学习的思想启发，下面我们对这 2 个领域的现有工作进行简要回顾.

1.1 个性化联邦学习

一般来说，现有的个性化联邦学习方法[5] 可以分为 3 大类，分别是本地微调方法、模型正则化方法和多任务学习方法.

在本地微调方法中，每个客户端接收一个全局模型，并使用其本地数据和多个梯度下降步骤对其进行调优. 例如，PMF[8] 为用户本地模型的较高层设计了 2 个个人适配器(个人偏差、个人过滤器)，可以用客户信息对全局模型进行微调.FedPer[9] 提出了一种基础+个性化层的方法，只对基础层进行协同训练. 在模型正则化方法中，文献 [10] 在本地模型和全局模型的距离上增加了一个正则项，并使用一个混合参数来控制它们之间的加权比例. 文献 [11] 提出了一种知识精馏的方法来实现个性化，将正则化应用于本地模型和全局模型之间的预测. 这些本地微调方法的局限性在于，它们使用统一的全局模型进行个性化设置，不能为异构数据客户端上的大量潜在任务提供灵活的个性化建模. 文献 [12] 提出了一种分层异构联邦学习方法将全局模型拆分为适配不同客户端资源的子模型序列，提升建模效果. 联邦多任务学习方法[13] 将每个客户机的优化视为一个新任务，它解决了通信约束、离散和容错问题，这些问题主要集中在凸模型上. 但是由于该方法对强对偶性的刚性要求，当客户采用非凸深度学习模型时，该方法就不再保证适用. 文献 [14] 提出使用注意力函数来测量模型参数之间的差异，它在服务器上为每个客户端维护了一个个性化云模型，通过传递个性化云模型实现信息的传递. 尽管注意力函数对客户端之间的成对协作进行了建模，但它是使用模型参数作为评估标准，难以对客户端之间的动态协作关系进行识别，在当客户端模型异构不稳定时，难以对个性化云模型进行准确地融合.

1.2 元学习

元学习是近年来研究的热点. 由于它能够在有限的训练数据下很好地泛化，因此被广泛应用于少样本学习[15]、强化学习[16]、迁移学习[17] 等领域. 例如，文献 [18] 引入了一种记忆增强神经网络，用于在没有灾难性干扰的情况下利用新数据重新学习模型. 文献 [19] 使用了一个元损失函数来存储不同环境的信息. 文献 [20] 提出了一个元翻译模型，该模型可以快速适应具有有限训练样本的新领域翻译任务. 最近，元学习也被用于个性化联邦学习. 文献 [21−22] 从经验的角度研究了模型不可知论元学习(MAML)类型方法与联邦学习的不同组合. 找到一个初始的

全局模型，使当前或新的客户端能够通过执行 1 个或几个步骤的梯度下降轻松地适应本地数据集. 文献 [23] 把元学习看作是对一系列损失的在线学习，每个损失函数都会得到单个任务中的上界. 文献 [21−23] 所提的方法可以被认为属于 1.1 节中提到的本地微调方法.

与现有的工作不同，在本文中我们设计了 AMN，它以细粒度的视角，利用基于注意力机制的神经网络来学习客户端之间的协作关系，并利用客户端的本地数据作为监督信号来训练网络. 该方法可以为客户端实现自适应的协作关系生成，并且被证明在客户端数据是异构的情况下特别有效.

2 问题定义

为了方便阐述，首先给出问题定义. 在一个个性化联邦学习系统中，有 n 个客户端连接到服务器. 每个客户端具有相同类型的基础网络 $f(\cdot)$，客户端的模型参数集合 $W=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$ 表示，其中 $\boldsymbol{w}_i\in\mathbb{R}^{1\times d}$ 为客户端 i 的模型参数. 服务器收集客户端上传的模型参数，通过聚合算法维护一个全局模型 $\boldsymbol{w}^{\mathrm{g}}\in\mathbb{R}^{1\times d}$. 通过传输 $\boldsymbol{w}^{\mathrm{g}}$ 和 W 实现客户端的协同训练. 对于每个客户端，我们定义目标函数为 $F(\boldsymbol{w}_i)=L(f(\boldsymbol{w}_i);D_i)$，其中 D_i 表示客户端 i 的本地数据集，$L(\cdot)$ 是损失函数.

由于客户端的数据集是异构的，我们的目标是在不公开客户端数据集的情况下，分析客户端本地模型参数之间的相似性，并构建细粒度的协作关系，以提升 $f(\boldsymbol{w}_i)$ 的性能. 本文提出的 AMN 可以被定义为 $G_{\Theta}(\boldsymbol{w}^{\mathrm{g}},W)$，其中 Θ 为元模型参数. 此时，个性化联邦学习的优化目标定义为

$$\min_{W,\Theta}\left\{\sum_{i=1}^{n}F(\boldsymbol{w}_i)+\lambda\sum_{i=1}^{n}F(G_{\boldsymbol{\theta}^i}(\boldsymbol{w}^{\mathrm{g}},W))\right\},$$

其中 λ 为正则化系数. 优化目标可以利用元网络和客户端的基础网络共同调节 $\boldsymbol{w}_i$，考虑到大多数深度学习网络的复杂性，很难得到 W 和 Θ 的最优解. 我们采用梯度下降技术和交替训练策略来解决这个双层优化问题.

3 基于 AMN 的个性化联邦学习方法设计

3.1 FedAMN 框架概述

图 1 展示了 FedAMN 的整体框架设计，它由 1 个服务器和多个客户端组成. 客户端的数据由客户自己收集，通常为边缘传感器收集，具有客户端的个性化特征. 为了保护客户隐私，服务器与客户端仅通过上传本地模型参数 $\boldsymbol{w}_i$ 和下载全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$ 以及参数集合 W 实现协同训练，并不会对本地数据进行传输.

服务器的主要功能为：

1）收集客户端的模型参数，构成客户模型参数集合 W；

2）使用聚类算法（如 k-means[24] 或 mean-shift[25] 算法）将参数集合划分为 k 个簇；

3）根据全局聚合算法计算全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$.

在功能 2）中，聚类算法的意义在于降低引入 AMN 而带来的额外数据传输量，同时最大限度地减少模型参数传输造成的潜在隐私泄露风险. 该操作对客户进行了粗分类，为之后元学习网络的自适应的协作关系选择提供基础. 在实际操作中，k 值由网络传输量限制要求决定，理想情况为 k=1（即全部客户端在同一个簇中，不对模型参数集进行划分）. 客户端只会下载包含自己模型参数的簇 $W_k\subseteq W$. 当任务对网络传输量的限制较高时，可以通过增大 k 的取值降低单个通信轮次中的传输量.k 的取值越大，每个客户端接收到的模型参数簇就越小，单个通信轮次的传输量就越小.k 值的选择标准为选取传输限制内可以实现训练的最小值. 为了简化描述，后文我们均以 k=1 为例对 FedAMN 进行说明，并以 W 表示客户端接收到的模型参数簇.

在功能 3）中，全局聚合算法是指现有的联邦学习模型聚合算法，如 FedAvg[1] 及一些变体. 通过引入全局模型，AMN 可以引入全局信息作为个性化模型的补充，并为客户端的模型训练提供一个软启动.

客户端的主要功能为：

1）收集客户数据，并对数据进行预处理，构建本地数据集；

2）从服务器下载客户端的模型参数集合 W（或聚类后的客户模型参数集合子集 W_k）和全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$；

3）利用本地数据集训练客户端的基础网络；

4）交替训练基础网络和元网络；

5）上传本地模型参数.

由于 AMN 的训练会引入额外的计算量，为了快速启动，我们采用渐进式的训练策略. 在训练初期，仅使用全局模型实现客户端间的协同训练，当客户端的本地模型达到稳定或达到设置好的训练阈值 p 时，AMN 再加入训练. 训练阈值 p 的设定是为了防

止本地模型因数据量较少等原因导致不收敛时, 依然可以引入 AMN 的另一个限制条件.p 的取值可以根据任务实际情况设定, 设置为本地模型得到充分训练的最小通信轮次.

下面我们对 AMN 的具体结构进行描述, 再介绍 FedAMN 的训练流程.

3.2 元学习网络结构

我们发现, 现有的联邦学习方法通常利用全局模型作为信息传输的载体, 但在客户端数据异构的情况下, 单一的全局模型难以解决客户的异质性问题. 同时, 模型的粗融合也会影响全局模型的性能. 而在模型训练过程中, 具有相似特征和数据分布的客户端通常会相互提供更多的有益信息. 因此, 我们期望提出一个元模型, 通过简单的训练就可以自适应地发现相似客户端间的协作关系, 以促进模型训练. 同时减少具有不同特征的其他客户端对模型训练的负面影响. 基于以上需求, 我们设计了一种分布部署在每个客户端中的元学习网络, 实现客户端间协作关系的自动生成. 该网络被称为 AMN.

AMN 以客户端的基础网络模型参数作为输入特征, 代表客户数据特性. 并根据参数集合学习一个元模型来分析不同客户端的相似性. 为了实现模型参数的自动融合, 受到文献 [26] 的启发, 我们使用注意力机制实现客户端模型权重的自动分配. 该机制的基本思想是查询 q 和键值矩阵 $\boldsymbol{K}$ 之间的相关性, 可以被表示为

$$softmax\left(\frac{q\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right),$$

其中 d 是比例因子, 避免公式内积的值过大.

具体来讲, AMN 的网络结构如图 2 所示, 第 1 层以客户端模型参数集合 W 为键值, 客户端的本地模型 $\boldsymbol{w}_i$ 为查询, 利用注意力机制自动分析目标客户端与其他客户端的相似性, 并使用加权分数对 W 进行聚合. 第 1 层可以过滤掉对目标客户端没有帮助的其他客户端的模型参数. 第 2 层以第 1 层的输出、全局模型为键值, 以 $\boldsymbol{w}_i$ 为查询, 从而对它们进行融合, 实现客户个性和共性的平衡.

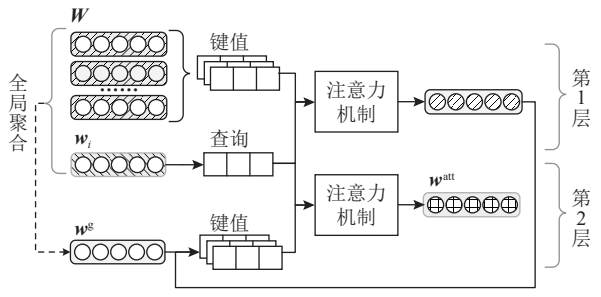


Fig. 2 Structure illustration of AMN network

图 2 AMN 网络结构示意图

因此, AMN 的优化过程可以表示为

$$\boldsymbol{w}_i^{\mathrm{att}} = \min_{\boldsymbol{\Theta}^i} G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}}, W),$$

其中 $G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}}, W) = h_2\left(\boldsymbol{w}^{\mathrm{g}}, h_1\left(W;\boldsymbol{\Theta}_1^i\right);\boldsymbol{\Theta}_2^i\right)$, $\boldsymbol{\Theta}^i = \left(\boldsymbol{\Theta}_1^i, \boldsymbol{\Theta}_2^i\right)$ 表示 AMN 中的可学习参数, 函数 $h_1(\cdot)$ 和 $h_2(\cdot)$ 分别代表 AMN 第 1 层和第 2 层的计算公式. 基于上文定义的注意力机制, $h_1(\cdot)$ 和 $h_2(\cdot)$ 被表示为

$$h_1\left(W;\boldsymbol{\Theta}_1^i\right) = softmax\left(\frac{\left(\boldsymbol{\Theta}_{1^q}^i \cdot \boldsymbol{w}_i\right)\left(\boldsymbol{\Theta}_{1^k}^i \cdot W\right)^{\mathrm{T}}}{\sqrt{d}}\right),$$

$$h_2\left(\boldsymbol{w}^{\mathrm{g}}, h_1\left(W;\boldsymbol{\Theta}_1^i\right);\boldsymbol{\Theta}_2^i\right) = softmax\left(\frac{\left(\boldsymbol{\Theta}_{2^q}^i \cdot \boldsymbol{w}_i\right)\left(\boldsymbol{\Theta}_{2^k}^i \cdot \left[\boldsymbol{w}^{\mathrm{g}}, h_1\left(W;\boldsymbol{\Theta}_1^i\right)\right]\right)^{\mathrm{T}}}{\sqrt{d}}\right),$$

其中 $\boldsymbol{\Theta}_1^i = (\boldsymbol{\Theta}_{1^q}^i, \boldsymbol{\Theta}_{1^k}^i)$, $\boldsymbol{\Theta}_2^i = (\boldsymbol{\Theta}_{2^q}^i, \boldsymbol{\Theta}_{2^k}^i)$ 分别代表第 1 层和第 2 层中的可学习参数. 通过训练得到使 AMN 效果最佳的参数 $\boldsymbol{\Theta}^i$, 该参数即为客户端 i 的 AMN 元模型参数.

3.3 FedAMN 训练流程

在 FedAMN 中, 我们引入额外的元学习网络 AMN 进行个性化联邦学习, 因此在每个通信轮次中, 客户端需要共同训练基础网络 $f(\cdot)$ 和元学习网络 $G(\cdot)$. 这 2 个网络之间具有复杂的依赖关系, 具体来讲, 元学习网络的输出会成为基础网络的模型参数, 基础网络的参数又会作为元学习网络的输入. 为了更好地训练这 2 个网络, FedAMN 采用渐进式的训练模式, 缓解 AMN 引入带来的计算量增加问题. 以交替训练的策略实现客户端个性化注意力增强模型的生成.

FedAMN 需要客户端和服务器配合完成协作训练, 算法 1 详细说明了在服务器进行的模型的收集与聚合过程. 算法 2 描述了在客户端进行的基础网络和元学习网络的交替训练策略. 客户端和服务器通过传输模型参数实现信息交换. 下面, 我们对客户端中进行的渐进式交替训练模式进行介绍.

算法 1. 模型的收集与聚合(服务器).

输入: 客户端的本地模型参数集合 $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n\}$, 聚类参数 k;

输出: 全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$.

① 初始化 $\boldsymbol{w}^{\mathrm{g}}$;

② 发送 $\boldsymbol{w}^{\mathrm{g}}$ 至各客户端;

③ for t=1, 2, … do

④ 等待接收客户端的模型参数集合

$W=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$;

⑤ 对 W 进行聚类，得到 k 个簇 $W_1,W_2,\cdots,W_k$;

⑥ 根据全局模型聚合算法计算全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$;

⑦ 将 $W_j\in\{W_1,W_2,\cdots,W_k\}$ 和 $\boldsymbol{w}^{\mathrm{g}}$ 发送给对应客户端;

⑧ if 达到停止训练条件

⑨ break;

⑩ end if

⑪ $t=t+1$;

⑫ end for

⑬ return $\boldsymbol{w}^{\mathrm{g}}$.

算法 2. 基础网络和元学习网络的交替训练策略（客户端）.

输入：全局模型参数 $\boldsymbol{w}^{\mathrm{g}}$，对应的模型参数集合簇 W_j，渐进式训练阈值 p;

输出：客户端本地模型参数 $\boldsymbol{w}_i$.

①初始化 $\boldsymbol{w}_i$ 和 $\boldsymbol{\Theta}^i$;

② for t=1, 2, ⋯ do

③ 从服务器下载 W_j 和 $\boldsymbol{w}^{\mathrm{g}}$;

④ for *epoch*=1, 2, ⋯ do

⑤ if $t<p$ then

⑥ for $D_i\leftarrow$ *minibatch* do

⑦ $\boldsymbol{w}_i\leftarrow\boldsymbol{w}_i-\eta_1\dfrac{\partial F(\boldsymbol{w}_i)}{\partial\boldsymbol{w}_i}$;

⑧ end for

⑨ else

⑩ for $D_i\leftarrow$ *minibatch* do

⑪ $\boldsymbol{\Theta}^i\leftarrow\boldsymbol{\Theta}^i-\eta_2\dfrac{\partial H(W,\Theta)}{\partial\boldsymbol{\Theta}^i}$;

⑫ end for

⑬ $\boldsymbol{w}_i\leftarrow G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}},W)$;

⑭ for $D_i\leftarrow$ *minibatch* do

⑮ $\boldsymbol{w}_i\leftarrow\boldsymbol{w}_i-\eta_1\dfrac{\partial H(W,\Theta)}{\partial\boldsymbol{w}_i}$;

⑯ end for

⑰ end if

⑱ if 达到停止训练条件

⑲ break;

⑳ end if

㉑ $t=t+1$;

㉒ 上传 $\boldsymbol{w}_i$ 至服务器;

㉓ end for

㉔ end for

㉕ return $\boldsymbol{w}_i$.

在开始的通信轮次中，客户端只负责训练基础网络. 该方式与通常的联邦学习训练方式相同，在这里不做专门描述. 当本地模型达到稳定或设定的阈值后，AMN 与基础网络交替训练.图 3 为客户端 i 中 AMN 的训练方式. 首先，在 AMN 网络训练阶段，为了调整其可学习参数 $\boldsymbol{\Theta}^i$，我们需要将其输出嵌入到基础网络中，作为基础网络的模型参数. 固定该参数，利用客户端的本地数据集计算梯度，并回传梯度对 AMN 参数进行调整. 该过程可以表示为

$$\boldsymbol{\Theta}^i=\min_{\boldsymbol{\Theta}^i}\{F(G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}},W)):=L(f(G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}},W);D_i))\}.$$

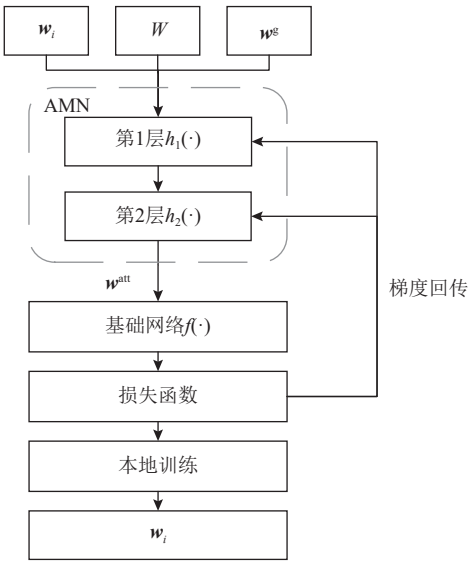


Fig. 3 Training method of meta-learning network AMN in client i

图 3 客户端 i 中元学习网络 AMN 的训练方式

算法 2 中用 W_j 表示模型参数集合簇. 如 3.1 节中服务器的功能 2)的描述，在后文我们统一用 W 表示该概念，即 k=1 时的状态. 之后，利用训练好的 $G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}},W)$ 得到注意力增强模型参数 $\boldsymbol{w}^{\mathrm{att}}$. 以 $\boldsymbol{w}^{\mathrm{att}}$ 作为基础网络的初始值，对基础网络进行训练，得到最终的客户端的个性化模型参数 $\boldsymbol{w}_i$，该过程可以表示为

$$\boldsymbol{w}_i=\min_{\boldsymbol{w}_i}\{F(\boldsymbol{w}_i):=(f(\boldsymbol{w}_i);D_i)\},$$

此时的 $\boldsymbol{w}_i$ 即为该通信轮次中的客户端的个性化模型.

上文的描述中仅对单个客户端的 AMN 训练方式进行了说明. 若考虑全部客户端，我们可以得到 FedAMN 的全局优化目标为

$$H(W,\Theta):=\sum_{i=1}^{n}F(\boldsymbol{w}_i)+\lambda\sum_{i=1}^{n}F(G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}},W)).$$

为了对 $H(W,\Theta)$ 进行优化，以一个通信轮次为例，FedAMN 中客户端的训练流程为：

1）从服务器下载 W 和 $\boldsymbol{w}_i$.

2) 当通信轮次小于训练阈值 p 时, 用 $F(\boldsymbol{w}_i)$ 训练客户端的基础网络, 以获得稳定的本地模型, 梯度下降过程表示为

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i - \eta_1 \frac{\partial F(\boldsymbol{w}_i)}{\partial \boldsymbol{w}_i}.$$

最后, 对更新后的本地模型进行微调, 得到最终的个性化模型

$$\boldsymbol{\Theta}^i \leftarrow \boldsymbol{\Theta}^i - \eta_2 \frac{\partial H(W, \Theta)}{\partial \boldsymbol{\Theta}^i}.$$

3) 当通信轮次大于训练阈值后, 使用 AMN 网络的输出更新客户端本地模型

$$\boldsymbol{w}_i \leftarrow G_{\boldsymbol{\Theta}^i}(\boldsymbol{w}^{\mathrm{g}}, W),$$

并对更新后的本地模型进行微调, 得到最终的个性化模型

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i - \eta_1 \frac{\partial H(W, \Theta)}{\partial \boldsymbol{w}_i}.$$

当通信轮次大于设定值或客户模型收敛时循环停止, 此时客户端的本地模型作为最终的客户个性化模型. 该过程与算法 2 中的描述相对应.

在 FedAMN 中, 客户端承担了更多的计算量, 服务器仅承担模型的聚合、存储、传输等功能. 这种情况下, FedAMN 允许客户在不泄露自己数据的情况下获取更多有益信息, 减少了由于本地数据量小而导致的模型过拟合问题, 并最大程度地消除了客户数据异质性带来的模型负迁移问题. 因此, 每个客户端的模型准确率可以明显高于传统联邦学习方法以及单机训练方法.

4 实验结果与分析

4.1 数据集介绍及实验设置

为了验证 FedAMN 的普适性和有效性, 本文实验数据集采用联邦学习中普遍采用的基准数据集 V-MNIST 和真实世界中的空气质量数据集 W&A China.

V-MNIST 数据集由 6 种常用的手写字体识别数据集组成, 分别是 MNIST[27]、MNIST 加旋转、MNIST 加噪声背景、MNIST 加图像背景、MNIST 加旋转和图像背景、Fashion MNIST[28]. 每个数据集都是 10 分类任务. 为了模拟联邦学习场景下客户端的个性化数据, 我们将这 6 个数据集采用独立同分布的随机抽样方法, 每个数据集按照原始数据量等比分布分为 3 个子集. 6 个数据集中, 共 72 795 个样本被划分为 18 个子数据集, 每个数据集由同一个数据的子集组成, 用于模拟 18 个客户端的本地数据集.

W&A China 数据集由中国国家环境监测中心和国家气象中心发布①, 包含 2017 年 1 月 1 日至 2017 年 12 月 31 日在全国 4 个直辖市(北京、天津、上海、重庆)中的 42 个监测站点按小时采集的共 23 万条环境数据. 每条数据包含 13 种特征元素, 分别为温度、压力、湿度、风向、风速、站号、收集时间和 6 种主要空气质量污染物的浓度. 其中根据中国环境空气质量标准(AAQS)[29], 可以按照 PM 2.5 的浓度将空气质量分为 5 个等级. 我们采用的任务为使用过去 48 h 的历史数据对未来的空气质量进行预测. 由于城市和周边环境的差异性, 每个监测站点的数据分布是异构的, 参考真实世界的设置, 在实验中, 每个站点都被认为是独立, 用于模拟 42 个客户端的本地数据集.

不同的基础网络会显著影响 FedAMN 以及基线方法的性能, 我们对 2 个数据集采用了 2 种不同的经典基础网络. 对于 V-MNIST, 我们用 AlexNet[30] 进行图像分类任务建模, AlexNet 是由 5 个卷积层和 3 个全连接层组成的卷积神经网络. 优化器为 SGD, 学习率设置为 0.02, 使用多标签交叉熵损失函数. 每个通信轮次中的 *epoch* 设置为 1, 批样本数量设置为 32.对于 W&A China 数据集, 我们构建了一个 2 层的门控循环单元网络(GRU)[31], 每层 128 个隐藏单元作为基础网络, 用于时间序列推理任务. 将结果通过全连通层输出, 并与真值进行比较, 得到预测精度. 优化器为 SGD, 初始学习率设置为 0.01, 使用多标签交叉熵损失函数. 每个通信轮次中的 *epoch* 设置为 2, 批样本数量设置为 32.

本文采用分类准确率和分类精确率作为对比各种方法的评价指标. 由于本文研究的是客户端模型的建模效果, 因此使用客户端的本地测试集进行测试, 最终结果为全部参与训练的客户端的测试结果的平均值. 对于联邦学习和个性化联邦学习方法, 我们分别对 V-MNIST 和 W&A China 数据集分别设置最大通信轮次为 150 和 175.在该轮次内, 选取在验证集中表现最佳的个性化模型或全局模型, 用于结果的测试.

全部实验均在 PyTorch1.5.0 下进行. 运行在一个 4 Tesla-P100 卡 GPU 服务器中, 搭载 Ubuntu 16.04.7 操作系统, CPU 为 Intel® Xeon® E5-2620, 128 GB 内存. 为了模拟实际情况下远程客户端的传输开销, 采用基于 socket 包的 TCP 协议进行面向连接的可靠传输.

① http://www.cnemc.cn, https://www.ncdc.noaa.gov

4.2 基线方法的对比实验

在本节中，我们将 FedAMN 与 8 种基线方法进行比较，分别是客户端单机训练、FedSGD、FedAvg[1]、FedProx[32]、FedPer[9]、FedHealth[33]、联邦迁移学习（FTL）[5] 和 FedAMP[14]. 其中 FedSGD, FedAvg, FedProx 是联邦学习方法，FedPer, FedHealth, FTL, FedAMP 是个性化联邦学习方法.

表 1 显示了不同基线方法与 FedAMN 在相同设置下的分类准确率. 可以看出，当客户端之间没有协作关系时，单机训练可以达到一个基础的分类准确率.2 个数据集代表了不同客户差异性下的任务性能. 在 V-MNIST 数据集中，基线联邦学习方法除了 FedSGD 之外，FedAvg 和 FedProx 均带来了分类准确率的提升. 这是由于图像识别任务比较复杂，有限的本地数据不足以提供必要的信息，此时客户间的协作训练非常重要. 而在 W&A China 数据集中，3 种联邦学习方法的效果均弱于单机训练. 因为 W&A China 是真实世界数据集，各城市间环境差异性大，此时使用联邦学习方法反而会造成性能下降. 并且根据实验结果我们发现，在 2 个数据集中，联邦学习方法的性能都显著低于个性化联邦学习. 与 FedAMN 相比，FedAvg 的平均分类准确率在 V-MNIST 和 W&A China 数据集中分别低了 2.9 个百分点和 11.59 个百分点. 这是由于客户端数据分布差异性导致的，说明了在联邦学习训练中构建客户个性化模型的必要性.

Table 1 Classification Accuracy of Different Methods

表 1 不同方法的分类准确率 %

方法	数据集	
	V-MNIST	W&A China
单机训练	78.95	77.36
FedSGD	62.28	68.61
FedAvg	81.76	72.57
FedProx	81.85	73.42
FedPer	82.23	79.09
FedHealth	82.81	80.21
FTL	81.07	81.51
FedAMP	81.94	82.12
FedAMN（本文）	84.66	84.16

在个性化联邦学习方法中，FedPer, FedHealth, FTL 属于本地微调方法，FedAMP 属于多任务学习方法.FedAMP 在基线方法中的性能表现最佳. 相对于 FedAMP, FedAMN 在 2 个数据集的分类准确率分别提升了 2.72 个百分点和 2.04 个百分点. 可以看出，与现有方法相比，引入 AMN 可以显著降低数据分布差异较大对模型建立带来的负面影响.

表 2 显示了不同基线方法在相同设置下的分类精确率. 分类精确率描述了在预测为正的样本中有多少是真正的正样本，体现了模型的查准率. 在 V-MNIST 数据集中，分类精确率和分类准确率表现相似. 在 W&A China 数据集中联邦学习方法的分类精确率出现了大幅度下降. 这是因为 V-MNIST 是基准数据集，不同类别样本的数量几乎分布均匀，而 W&A China 是真实世界数据集. 通常，在类别不平衡分类问题中，模型的分类精确率更低. 城市中空气质量差的样本明显少于空气质量好的样本，在此种情况下，对于样本较少的类别（如 5 级污染）查准率会更低. 尽管如此，FedAMN 相较于其他基线方法依然表现出更好的分类效果. 与最优的联邦学习方法 FedProx 相比，FedAMN 的分类精确率在 2 个数据集中分别提升了 6.49 个百分点和 17.56 个百分点. 与最优的个性化联邦学习方法 FedAMP 相比，FedAMN 的分类精确率分别提升了 2.45 个百分点和 1.55 个百分点.

Table 2 Classification Precision Rate of Different Methods

表 2 不同方法的分类精确率 %

方法	数据集	
	V-MNIST	W&A China
单机训练	79.24	51.25
FedSGD	63.30	26.86
FedAvg	78.10	46.09
FedProx	78.64	48.31
FedPer	78.42	60.16
FedHealth	79.58	62.75
FTL	81.52	63.57
FedAMP	82.68	64.32
FedAMN（本文）	85.13	65.87

此外，为了验证 FedAMN 在数据异构和本地模型异构情况下的训练表现，我们对 W&A China 数据集的数据和本地模型结构进行了调整，对比了不同方法的分类准确率.

在数据异构实验中，原本的 W&A China 数据集的每条样本数据均包含 13 种特征元素，我们对数据集中的每个站点均随机去除其中的 2 种特征. 此时，站点的数据特征元素为 11 种，各站点的数据特征元素的种类并不相同，以此模拟数据异构的情况. 实验结果如表 3 所示，可以看出，在数据异构情况下，联

Table 3 Classification Accuracy of Different Methods in the Heterogeneity Cases on W&A China Dataset

表 3 不同方法在 W&A China 数据集中不同异构情况下的分类准确率 %

方法	数据异构	模型异构
单机训练	75.27	74.54
FedSGD	54.31	42.78
FedAvg	60.49	43.51
FedProx	61.93	43.92
FedPer	73.38	60.39
FedHealth	73.26	61.14
FTL	75.85	62.27
FedAMP	76.11	73.98
FedAMN（本文）	79.63	76.48

邦学习方法 FedSGD, FedAvg, FedProx 相较于单机训练的分类准确率分别下降了 20.96 个百分点、14.78 个百分点、13.34 个百分点. 这是由于数据异构情况下，全局模型并不能很好地实现信息传递，导致模型效果变差. 个性化联邦学习方法 FedPer 和 FedHealth 的分类准确率也略低于单机训练. 此时，联邦设置并不能带来模型效果上的提升. 个性化联邦学习方法 FTL 达到了几乎等同于单机训练的效果. 在基线方法中只有 FedAMP 可以达到略高于单机训练的效果. 而 FedAMN 相较于单机训练，分类准确率提升了 4.36 个百分点，相较于表现最优的个性化联邦学习方法 FedAMP 的分类准确率提升了 3.52 个百分点. 该结果表明 FedAMN 在数据异构情况下相较于其他基线方法，可以更好地实现个性化建模.

在模型异构实验中，我们对不同站点采用不同的本地模型. 具体来讲，不同站点依然采用 2 层 GRU 作为基础网络. 但在北京和天津的站点中，每层 GRU 包含 128 个隐藏单元，上海和重庆的站点每层 GRU 包含 256 个隐藏单元，以此模拟本地模型异构的情况. 对于联邦设置中的模型聚合操作，我们仅对模型中重叠的参数进行加权平均，得出不同基线方法下的分类准确率. 实验结果如表 3 所示，可以看出模型异构对于基线方法均有较大影响，尤其是联邦学习方法，分类准确率相较于单机训练至少下降了 30.62 个百分点. 这是由于不同模型结构下的全局模型难以聚合，对于结构不同的部分无法实现加权平均. 个性化联邦学习由于有模型微调的功能，准确率略有上升，但基线方法的分类准确率依然均低于单机训练，其中表现最好的 FedAMP 的分类准确率相较于单机训练降低了 0.56 个百分点. FedAMN 是实验中效果最佳的方法，相较于单机训练，其分类准确率提升了 1.94 个百分点，相较于 FedAMP 分类准确率提升了 2.5 个百分点. 这是由于 FedAMP 可以促进具有相似特征和数据分布的客户端之间的协作，并疏远具有广泛不同特征的客户端之间的协作，从而缓解模型异构导致的模型融合效果差的情况.

4.3 训练时间对比实验

AMN 网络的引入增加了客户端的训练开销，同时增加了客户端和服务器之间的通信开销. 为了评估 FedAMN 对各项资源的消耗，我们对不同方法的总训练时间进行了测试.

实验如表 4 所示，ToAcc@70 表示不同方法的分类准确率第 1 次达到 70% 所花费的训练时间，ToAcc@80 表示不同方法的分类准确率第 1 次达到 80% 所花费的训练时间. 数值越小表示训练越快，资源消耗越小. 可以看出，单机训练方法、联邦学习方法和个性化联邦学习方法所需的训练时间依次递增. 由于 FedAMN 需要额外训练 AMN 网络，因此在 ToAcc@70 和 ToAcc@80 情况下所花费的训练时间是最长的. 相较于分类准确率最高的基线方法 FedAMP，FedAMN 的训练时间在 ToAcc@70 和 ToAcc@80 中分别增加了 6.7% 和 2.8%. 尽管 FedAMN 的训练时间相对较长，但它所带来的模型分类准确率提升是更大的. 这使得 FedAMN 更加适合对模型效果要求更高、对训练时间要求较为宽松的任务.

Table 4 Classification Accuracy of Different Methods in W&A China Dataset and Required Training Time

表 4 不同方法在 W&A China 数据集中的分类准确率以及所需的训练时间

方法	分类准确率/%	ToAcc@70/min	ToAcc@80/min
单机训练	77.36	25	-
FedSGD	68.61	-	-
FedAvg	72.57	78	-
FedProx	73.42	86	-
FedPer	79.09	82	-
FedHealth	80.21	104	159
FTL	81.51	96	142
FedAMP	82.12	133	176
FedAMN	84.16	142	181

注: -表示未能达到目标准确率.

4.4 消融实验

在 FedAMN 中我们引入 AMN 来对客户端模型参数进行融合，2 层网络结构分别承担了不同的作

用. 为了进一步说明 2 层网络设计的必要性，我们通过消融实验来说明训练过程中各层对结果的影响. 为了便于对比，我们对 AMN 进行分解，增加了 3 种变体来进行实验：

1）FTL. 优化目标与 FedAMN 相同，但客户端不加入 AMN 网络，仅利用全局模型来协助客户实现个性化模型的训练.

2）FedAMN-1 层. 客户端部署的 AMN 网络仅包含第 1 层结构，即对 W 进行融合处理.

3）FedAMN-2 层. 客户端部署的 AMN 网络仅包含第 2 层结构. 由于此时没有第 1 层，因此只使用本地模型和全局模型作为网络输入.

4）FedAMN.具备完整 2 层 AMN 结构的个性化联邦学习训练.

FedAMN 及其 3 种变体方法在 2 个数据集上的分类准确率和分类精确率结果如图 4 所示. 可以看出，FTL 方法的表现最弱，FedAMN-1 层和 FedAMN-2 层略优于 FTL，但弱于 FedAMN. AMN 的加入使得客户端分类准确率在 2 个数据集上相较于 FTL 分别提升了 3.59 个百分点和 2.65 个百分点. 同时，我们发现 AMN 中第 1 层比第 2 层为客户端个性化建模带来了更多的好处. 因为第 1 层可以调整客户间有益信息的比例，过滤相似度更弱的客户端. 但尽管第 2 层的益处略低于第 1 层，全局模型信息的加入仍然带来进一步的性能提升. 所以 AMN 的 2 层结构代表了不同层面的信息融合，两者的结合可以在模型共性和个性间达到动态平衡，为客户间协作关系的刻画提供更全面的帮助.

4.5 AMN 的可视化分析

为了对 AMN 中注意力机制的有效性进行分析，我们以 W&A China 数据集为例，对 AMN 的 2 层结构分别进行了可视化实验. 如图 5 所示，c_1~c_9 为 9 个随机选取的位于北京和上海的空气质量监测站点. 其中 c_5 和 c_9 位于上海，其他的位于北京.图 5(a)(b)为通信轮次为 2 时的 AMN 可视化结果，图 5(c)(d)为通信轮次为 50 时的 AMN 可视化结果. 网格的颜色越浅，代表相关性越强，注意力分数越高. 可以看出，在通信轮次为 2 时，AMN 的第 1 层中每个客户端的注意力权重都趋向平均，第 2 层的也相差不大(除了 c_5 因为是上海的站点，具有和其他站点更加不同的特点，$\boldsymbol{w}^{\mathrm{att}}$ 得到更高的注意力分数). 说明 AMN 在训练初期并不能得到很好的训练，实现客户模型相关性的判断. 在通信轮次为 50 时，同一个城市的客户端相互得到了更高的注意力分数. 第 2 层中的 $\boldsymbol{w}^{\mathrm{att}}$ 所占比例显著增加. c_5 和 c_9 也给予对方最高的注意力分数.

2 组对比实验表明，在训练的早期，本地模型的训练不够充分，AMN 的加入并不能带来很好的效果. 然而，在训练的后期阶段，可以很好地利用 AMN，将相似的客户端分组通过分配高注意力分数来加强协作. 这也说明了渐进式训练方式的必要性.

4.6 客户端异质性对训练的影响

为了更好地探究客户端异质性对训练的影响，在本节中，我们对 FedAMN 和其他基线方法在不同城市及站点数量情况下的分类准确率进行了测评.W&A China 数据集中包含中国 4 个直辖市(北京、天津、上海、重庆)中的 42 个监测站. 我们选取北京的 1 个站点作为观测客户，依次加入北京、天津、上海、重庆的另外 3 个站点进行联邦训练，再测量观测客户的分类准确率. 通过调整协同训练过程中参与

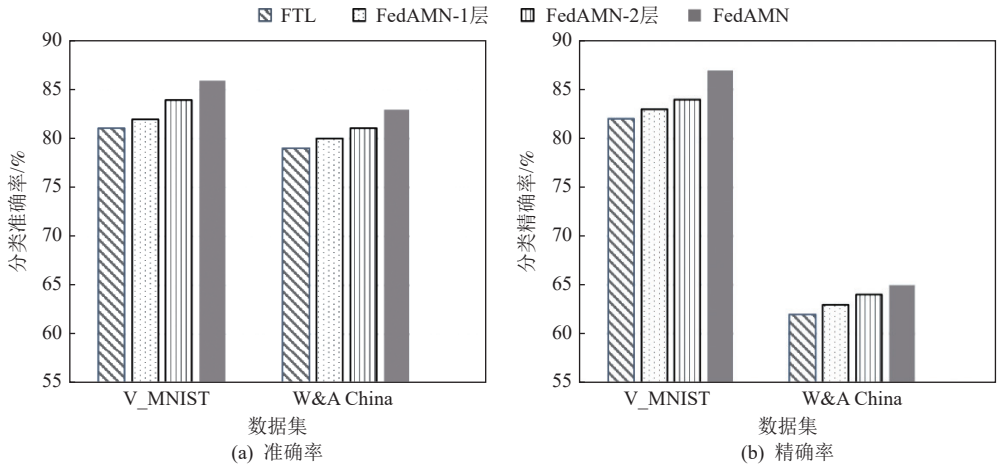
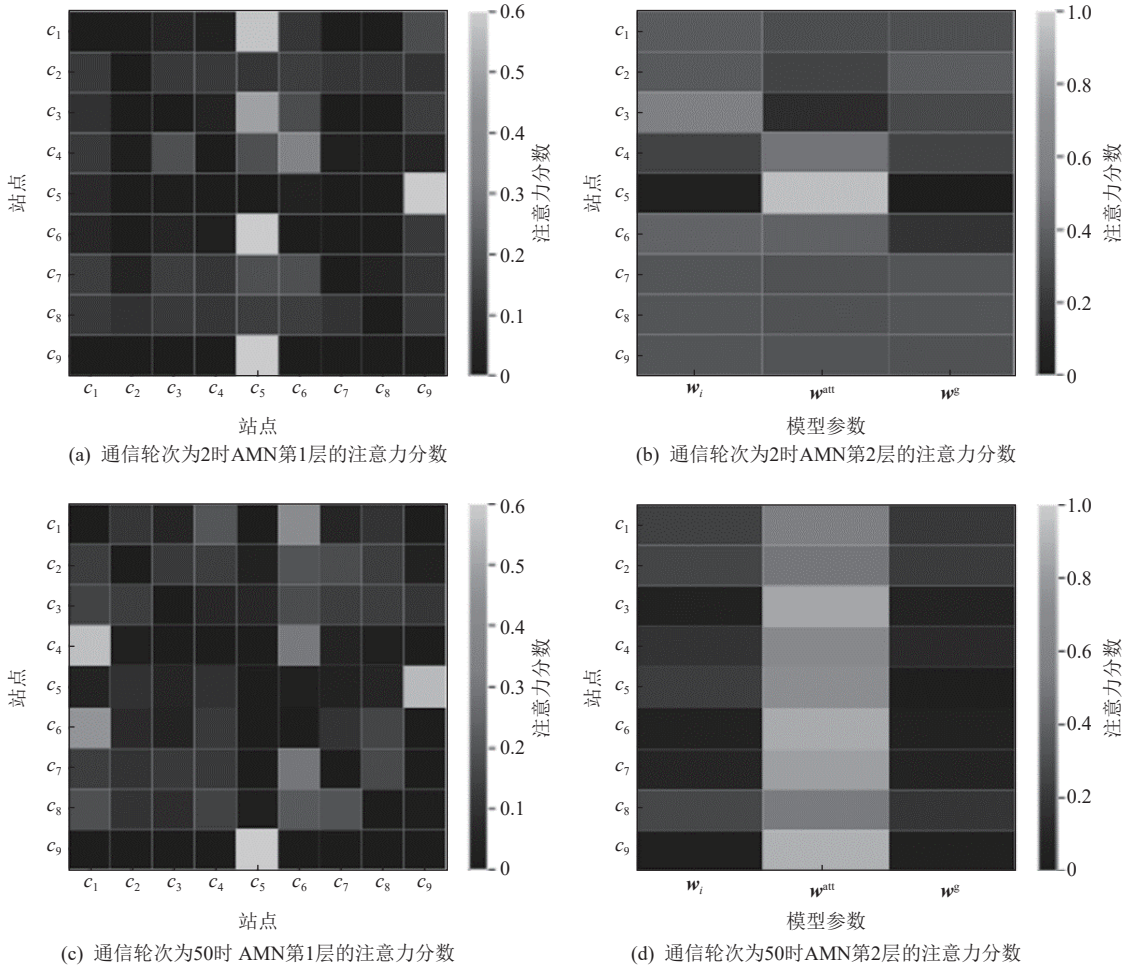


Fig. 4 Ablation experiments of FedAMN and its three variants on two data sets

图 4 FedAMN 及其 3 种变体在 2 个数据集上的消融实验



(a) 通信轮次为2时AMN第1层的注意力分数 (b) 通信轮次为2时AMN第2层的注意力分数

(c) 通信轮次为50时 AMN第1层的注意力分数 (d) 通信轮次为50时AMN第2层的注意力分数

Fig. 5 Results of attention score visualization for two-layer structure of AMN under different communication rounds

图 5 不同通信轮次下 AMN 的 2 层结构中的注意力分数可视化结果

城市和站点数量，比较不同个性化联邦学习方法对单个客户分类准确率的影响，以检验模型在面对数据异构情况下的训练表现是否良好. FedHealth 需要在服务器上构造源域数据集，在城市感知应用场景下难以实现，因此并未参与测试. 该实验中的对比方法采用联邦学习方法 FedAvg 和个性化联邦学习方法 FedPer, FTL, FedAMP.

实验结果如 6 所示，当只有观测客户独自参与训练时，可以等同于单机训练，分类准确率为 73.01 个百分点. 随着不同城市站点的依次加入，观测站点的分类准确率出现改变. 比较明显的是当加入 3 个北京站点后，所有方法的分类准确率都出现了不同程度的上升. 而随着天津站点的加入，FedAvg 的分类准确率出现了明显下降. 此时，城市间站点的异构性对全

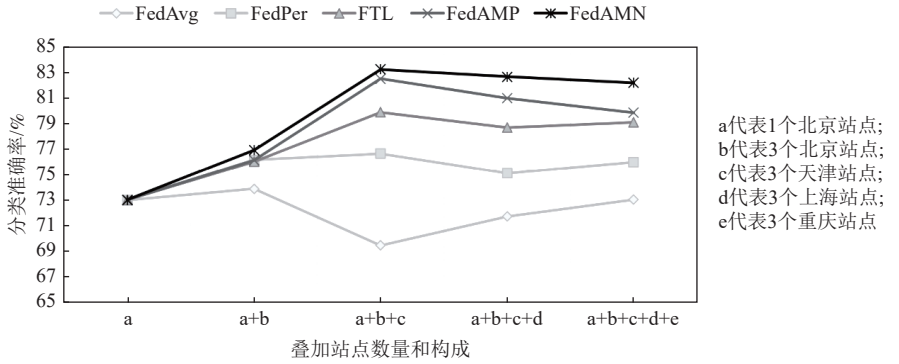


Fig. 6 Effect of changing the number of sites in different cities for collaborative training on the accuracy of observed client

图 6 不同城市及站点数量情况下的联邦训练对观测客户准确率的影响

局模型的影响初步显现，作为依赖全局模型实现模型优化的联邦学习方法，FedAvg 无法处理这种异构性，因此观测客户的分类准确率维持在较低水平. 而个性化联邦学习方法都在不同程度上有效处理此种程度的异质性. 当上海和重庆站点加入后，所有方法在观测客户上的分类准确率均有下降. 这是由于北京和天津在地理上更为接近，2 个城市具有相似的空气质量趋势，而上海和重庆作为南方城市，气候条件与北方差异较大，此时将它们放在一起训练会造成分类准确率下降. 尽管如此，FedAMN 仍然保持了最优的效果. 和次优方法 FedAMP 相比，上海和重庆的站点加入训练后的分类准确率分别提升了 1.67 个百分点和 2.34 个百分点. 和联邦学习方法 FedAvg 相比，上海和重庆站点加入训练后的分类准确率分别提升了 10.94 个百分点和 9.14 个百分点. 这表明 AMN 的加入可以自动学习客户间相似性，在面对客户间严重的数据异构问题时具有更强的健壮性和有效性.

5 总　　结

在个性化联邦学习训练过程中，由于客户端数据的隐私性，现有的方法难以细粒度地提取客户端特征，并定义它们之间的协作关系. 同时，数据异质情况下，全局模型的粗融合也会影响最终建模的性能效果. 更优的方式，是为每个客户端提供不同的协作关系. 具体来讲，促进具有相似特征和数据分布的客户端之间的协作，并疏远具有广泛不同特征的客户之间的协作. 因此，对于给定的客户端，如何自动提取与其类似的客户端，是一个具有挑战的问题. 此外，对于服务器-客户端框架，需要设计合理的协作训练方法.

为此，本文将个性化联邦学习中的协作关系刻画问题形式化为一个分布式元学习网络训练任务，并设计了一种注意力增强元学习网络（AMN）来解决客户端本地模型的个性化问题，有效地实现客户端个性与共性的权衡. 并提出了 FedAMN 框架，将 AMN 分布部署在客户端进行个性化模型训练. 考虑到客户端需要同时训练 AMN 和本地网络，我们设计了渐进式的交替训练策略，以端到端的方式对 2 个网络进行训练，缓解了元学习网络引入带来的计算量增加问题. 最终实现客户端的个性化注意力增强模型的生成. 为了证明 FedAMN 的有效性，我们在 2 个基准数据集和 8 种基准方法上进行了大量实验，相较于现有表现最优的个性化联邦学习方法，我们的方法在 2 个数据集中平均分别提升了 3.39% 和 2.45% 的模型性能.

在未来的工作中，我们会继续研究 FedAMN 中的模型压缩问题，以期进一步减少服务器和客户端之间的通信开销，实现个性化联邦学习技术在智能感知场景中的应用.

作者贡献声明：高雨佳提出了方法思路与实验方案，完成实验并撰写论文；王鹏飞提供理论指导、研究思路分析及论文修改意见；刘亮和马华东提供理论分析和技术指导，提出论文修改意见并最终审核.

参　考　文　献

[1] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proc of the 20th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273−1282

[2] Birje M N, Hanji S S. Internet of things based distributed healthcare systems: A review[J]. Journal of Data, Information and Management, 2020, 2(3): 149−165

[3] Lee Mingchang, Lin Jiachun. DALC: Distributed automatic LSTM customization for fine-grained traffic speed prediction[C]// Proc of the 34th Int Conf on Advanced Information Networking and Applications. Berlin: Springer, 2020: 164−175

[4] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1/2): 1−210

[5] Wu Qiong, He Kaiwen, Chen Xu. Personalized federated learning for intelligent IoT applications: A cloud-edge based framework[J]. IEEE Open Journal of the Computer Society, 2020, 1: 35−44

[6] Chen Hongyou, Chao Weilun. On bridging generic and personalized federated learning[J]. arXiv preprint, arXiv: 2107.00778, 2021

[7] Lu Xiaofeng, Liao Yuying, Liu Chao, et al. Heterogeneous model fusion federated learning mechanism based on model mapping[J]. IEEE Internet of Things Journal, 2021, 9(8): 6058−6068

[8] Feng Jie, Rong Can, Sun Funing, et al. PMF: A privacy-preserving human mobility prediction framework via federated learning[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(1): 1−21

[9] Arivazhagan M G, Aggarwal V, Singh A K, et al. Federated learning with personalization layers[J]. arXiv preprint, arXiv: 1912. 00818, 2019

[10] Hanzely F, Richtárik P. Federated learning of a mixture of global and local models[J]. arXiv preprint, arXiv: 2002. 05516, 2020

[11] Shen Tao, Zhang Jie, Jia Xinkang, et al. Federated mutual learning[J]. arXiv preprint, arXiv: 2006. 16765, 2020

[12] Zhong Zhengyi, Bao Weidong, Wang Ji, et al. A hierarchically heterogeneous federated learning method for cloud-edge-end system[J]. Journal of Computer Research and Development, 2022,

59(11): 2408−2422 (in Chinese)
(钟正仪, 包卫东, 王吉, 等. 一种面向云边端系统的分层异构联邦学习方法 [J]. 计算机研究与发展, 2022, 59(11): 2408−2422)
[13] Smith V, Chiang C K, Sanjabi M, et al. Federated multi-task learning[C]// Proc of the 30th Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2017: 4424−4434
[14] Huang Yutao, Chu Lingyang, Zhou Zirui, et al. Personalized cross-silo federated learning on Non-IID data[C]// Proc of the 11th Symp on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 7865−7873
[15] Jamal M A, Qi Guojun. Task agnostic meta-learning for few-shot learning[C]//Proc of the 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 11719-11727
[16] Bello I, Zoph B, Vasudevan V, et al. Neural optimizer search with reinforcement learning[C]//Proc of the 34th Int Conf on Machine Learning. New York: PMLR, 2017: 459−468
[17] Sun Qianru, Liu Yaoyao, Chua T, et al. Meta-transfer learning for few-shot learning[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 403−412
[18] Santoro A, Bartunov S, Botvinick M, et al. Meta-learning with memory-augmented neural networks[C]// Proc of the 33rd Int Conf on Machine Learning. New York: PMLR, 2016: 1842-1850
[19] Houthooft R, Chen Yuhua, Isola P, et al. Evolved policy gradients[C]// Proc of the 31st Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2018: 5405−5414
[20] Lin Jianxin, Wang Yijun, Chen Zhibo, et al. Learning to transfer: Unsupervised domain translation via meta-learning[C]// Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11507−11514
[21] Jiang Yihan, Konečný J, Rush K, et al. Improving federated learning personalization via model agnostic meta learning[J]. arXiv preprint, arXiv: 1909. 12488, 2019
[22] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning: A meta-learning approach[J]. arXiv preprint, arXiv: 2002. 07948, 2020
[23] Khodak M, Balcan M F F, Talwalkar A S. Adaptive gradient-based meta-learning methods[C]// Proc of the 32nd Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2019: 5915−5926
[24] Krishna K, Murty M N. Genetic K-means algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B (Cybernetics), 1999, 29(3): 433−439
[25] Carreira-Perpinan M A. Gaussian mean-shift is an EM algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 767−776
[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proc of the 30th Advances in Neural Information Processing Systems. San Francisco, CA: Morgan Kaufmann, 2017: 5998−6008
[27] LeCun Y. The MNIST database of handwritten digits[EB/OL]. 1998[2021-06-08]. http://yann. lecun. com/exdb/mnist/
[28] Xiao Han, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[J]. arXiv preprint, arXiv: 1708. 07747, 2017
[29] Wuxi Binhu District People's Government. GB 3095—2012 Environmental Air Quality Standard [S]. Wuxi: Binhu Ecological Environment Bureau, 2012 (in Chinese)
(无锡市滨湖区人民政府. GB 3095—2012 环境空气质量标准 [S]. 无锡: 滨湖生态环境局, 2012)
[30] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84−90
[31] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint, arXiv: 1412. 3555, 2014
[32] Li Tian, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine Learning and Systems, 2020, 2: 429−450
[33] Chen Yiqiang, Qin Xin, Wang Jindong, et al. FedHealth: A federated transfer learning framework for wearable healthcare[J]. IEEE Intelligent Systems, 2020, 35(4): 83−93



Gao Yujia, born in 1995. PhD. Her main research interests include federated learning, deep learning, and intelligent sensing.
高雨佳, 1995 年生. 博士. 主要研究方向为联邦学习、深度学习和智能感知.



Wang Pengfei, born in 1989. PhD, associate professor. His main research interests include data mining, machine learning, and recommendation system.
王鹏飞, 1989 年生. 博士, 副教授. 主要研究方向为数据挖掘、机器学习和推荐系统.



Liu Liang, born in 1982. PhD, professor. Senior member of CCF. His main research interests include Internet of things and intelligence sensing.
刘　亮, 1982 年生. 博士, 教授. CCF 高级会员. 主要研究方向为物联网、智能感知.



Ma Huadong, born in 1964. PhD, professor, fellow of CCF, fellow of IEEE. His main research interests include Internet of things, sensor networks, multimedia system and networking.
马华东, 1964 年生. 博士, 教授. CCF 会士、IEEE 会士. 主要研究方向为物联网、传感器网络、多媒体系统与网络.