计算机科学
COMPUTER SCIENCE

基于知识蒸馏的差分隐私联邦学习方法

谭智文，徐茹枝，王乃玉，罗丹

引用本文

谭智文，徐茹枝，王乃玉，罗丹. 基于知识蒸馏的差分隐私联邦学习方法[J]. 计算机科学，2024，51(6A): 230600002-8.

TAN Zhiwen, XU Ruzhi, WANG Naiyu, LUO Dan. Differential Privacy Federated Learning Method Based on Knowledge Distillation [J]. Computer Science, 2024, 51(6A): 230600002-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600002

基于知识蒸馏的差分隐私联邦学习方法

谭智文　徐茹枝　王乃玉　罗　丹
华北电力大学控制与计算机工程学院　北京 102206
(xuruzhi@ncepu.edu.cn)

摘　要　差分隐私技术作为一种隐私保护方法,在联邦学习领域得到了广泛应用。现有的差分隐私应用于联邦学习的研究,或是未考虑无标签公共数据,或是未考虑客户端之间的数据量差异,限制了其在现实场景的应用。文中提出一种基于知识蒸馏的差分隐私联邦学习方法,引入无标签公共数据集并考虑到客户端之间数据量的差异,为此场景设计了专用的差分隐私方案。首先,按数据量大小将客户端分组为"大数据量客户端"和"一般客户端",用大数据量客户端的数据训练教师模型,教师模型为公共数据集添加伪标签,然后,公共数据集作为"特殊客户端"与"一般客户端"共同进行联邦训练。采用差分隐私技术保证客户端的数据隐私,由于特殊客户端的数据只有标签涉及隐私,在联邦训练中为其分配比一般客户端更多的隐私预算;限制隐私预算总量,设联邦训练阶段的隐私预算为定值,根据客户端对隐私性的需求和隐私预算平行组合性质,调整伪标签添加阶段的隐私预算。在MNIST数据集和SVHN数据集上的实验表明,在同等的隐私预算消耗下,训练得到了精度比传统方法更高的模型。本方案具有可拓展性,高灵活度的隐私预算分配使其可以满足复杂的隐私需求。

关键词:联邦学习;差分隐私;知识蒸馏;隐私保护;隐私预算
中图分类号　TP309.2

Differential Privacy Federated Learning Method Based on Knowledge Distillation

TAN Zhiwen,XU Ruzhi,WANG Naiyu and LUO Dan
School of Control and Computer Engineering,North China Electric Power University,Beijing 102206,China

Abstract　Differential privacy technology,as a privacy protection method,has been widely applied in federated learning. The existing research on the application of differential privacy in federated learning either fails to consider unlabeled public data or the difference in data volume between clients,which limits its application in real-world scenarios. This paper proposes a differential privacy federated learning method based on knowledge distillation,which introduces unlabeled public datasets and considers the differences in data volume between clients. A dedicated differential privacy scheme is designed for this scenario. Firstly,the clients are grouped into "large data clients" and "general clients" based on the size of the data. The teacher model is trained using the data from the large data clients,and the teacher model adds pseudo labels to the public dataset. Then,the public dataset is used as a "special client" to jointly conduct federated training with the "general client". Adopting differential privacy technology to ensure the data privacy of clients,as the data of special clients only involves privacy with labels,more privacy budgets are allocated to them in federated training compared to general clients. Limit the total amount of privacy budget,set the privacy budget for the federal training stage as a fixed value,and adjust the privacy budget for the pseudo label addition stage based on the client's privacy needs and the parallel combination property of privacy budget. Experiments on the MNIST and SVHN datasets show that,under the same privacy budget consumption,the trained model has higher accuracy than traditional methods. This scheme has scalability,and its high flexibility of privacy budget allocation enables it to meet complex privacy needs.

Keywords　Federated learning,Differential privacy,Knowledge distillation,Privacy protection,Privacy budget

1　引言

近年来,由于计算机算力的进步和机器学习技术的广泛应用,海量数据的处理效率飞速提升。同时出现了数据隐私泄露和数据孤岛的问题,各机构一方面渴望通过合作得到更多的数据,另一方面又害怕本地数据隐私的泄漏。为解决这些问题,联邦学习[1]应运而生,其保证各客户端的数据不离开本地,客户端仅与中心服务器进行模型参数的传输,这种方式有效保障了用户的隐私安全。然而,已有研究结果表明,联邦学习中用户的隐私仍面临被动和主动攻击的威胁[1]。因此,需要隐私保护技术来进一步保证客户端的数据隐私。

差分隐私[2]作为一种被证实的具有强隐私保护能力的技术,主要通过随机噪声(如拉普拉斯噪声[2]或高斯噪声[3])的加入,将数据查询操作的实际结果隐藏起来。由于差分隐私良好的数据保护能力,越来越多的研究将其引入联邦学习中。以深度学习模型中数据处理阶段为依据,可以将差分隐私

基金项目:国家自然科学基金(61972148)
The work was supported by the National Natural Science Foundation of China(61972148).
通信作者:徐茹枝(xuruzhi@ncepu.edu.cn)

部署的位置分为输入层、隐藏层和输出层[4]。在隐藏层应用差分隐私时，常见的做法是在梯度上加入噪声，Abadi 等[5]提出一种将差分隐私机制与随机梯度下降算法相结合的隐私保护深度学习方法，该方法主要通过在小批量步骤后利用噪声干扰本地梯度实现隐私保护；Lu 等[6]提出了一种差分隐私异步联邦学习(DPAFL)方案，通过将本地差分隐私引入联邦学习中，在本地模型的 SGD 更新中加入高斯噪声以保护隐私性。随着联邦学习领域中差分隐私技术的运用日渐成熟，Wei 等[7]对差分隐私与联邦学习的结合进行了深入的分析，证明存在最优的 K 值($1\leqslant K\leqslant$总客户端数 N)，可以在固定的隐私保护级别上实现最佳的收敛性能；Mc-Mahan 等[8]使用差分隐私加密全局模型更新，进一步证明了如果参与联邦学习的客户端数量足够多，对模型更新信息的加密就会以增加计算量为代价而不会降低模型精度。

现有的联邦学习研究存在以下两种情况。1)除客户端本地存储的数据之外，还存在着同样可以利用的无标签公共数据。2)客户端的数据量不相同，甚至存在较大差异。针对以上两种场景，分别有了解决方案，但缺少同时考虑这两种情况的研究，于是本文提出方案，在有效利用公共数据集的知识的同时，处理客户端数据量存在差异的问题。

本文提出的基于知识蒸馏的差分隐私联邦学习方法首先将客户端按数据量大小分组为大数据量客户端和一般客户端，以大数据量客户端的数据训练教师模型，教师模型为公共数据集添加伪标签，然后公共数据集作为特殊客户端与一般客户端共同进行联邦训练。为算法设计了专用的隐私保护方案，由于特殊客户端的数据只有标签涉及隐私，在联邦训练中为特殊客户端分配比一般客户端更多的隐私预算；限制隐私预算总量，设联邦训练阶段的隐私预算为定值，根据客户端对隐私性的需求和隐私预算平行组合性质，调整伪标签添加阶段的隐私预算。

本文贡献如下：

1)针对存在无标签公共数据且客户端数据集数量存在一定差异的场景，提出一种基于知识蒸馏的差分隐私联邦学习方法。

2)基于提出的联邦学习方法，设计了专用的差分隐私保护方案，调整了客户端之间以及算法各阶段之间隐私预算的分配。

3)实验表明，在同等的隐私预算消耗下，我们的方案相比传统算法得到的模型精度更高。

4)本方案具有可拓展性，高灵活度的隐私预算分配使其可以满足复杂的隐私需求。

2 相关工作

第 1 章提到了联邦学习研究存在两种情况，针对第一种情况中如何利用无标签公共数据的问题，现有的工作常用知识蒸馏[9]来学习公共数据集的知识，例如在联邦训练中传输的信息不再是传统的梯度，而是本地模型的输出[10-12]，提升模型性能，降低通信开销；有的工作则是利用公共数据集来调整模型，在服务器聚合后，用公共数据集对模型进行微调以实现更高的性能[13-15]；Papernot 等[16]提出的 PATE 框架通过教师模型的投票将隐私数据的知识转移到公共数据集上；为了减少通信开销，Kerkouche 等[17]在训练中只交换部分参数，参数的选择由服务器在公共数据集上的训练决定。针对第二种情况中，客户端的数据量存在较大差异的问题，主流的处理方法是将其归类于 non-iid 联邦学习问题来处理[18]。本文提出的方案则综合考虑了两种情况，在有效利用公共数据集的知识的同时，处理了客户端数据量存在差异的问题，该方案具有广阔的应用前景。

此外，在本文提出的算法框架上，我们设计了专用的隐私预算分配方案。在差分隐私领域中，已经有许多工作对隐私预算分配方案进行了研究，例如 Shi 等[19]依据客户端在公共数据集上的测试精度将其分组，再动态调整相应模型的隐私预算；Zhang 等[20]根据每个本地模型在公共验证数据集上的准确性，为每一轮动态调整相应模型的隐私预算；Huang 等[21]将隐私预算分为两部分，在每次迭代中，一部分用来为梯度加噪，另一部分用来选择最优步长，并自适应地调整这两部分的大小；Liu 等[22]设定每个客户端隐私预算和 δ 不同，为“公共”客户端分配更大的权重，同时减少“私人”客户端的权重，希望不那么混乱的“公共”客户端更新将提高全局模型的效用。本文算法中，我们不仅提出了客户端之间的隐私预算分配方案，也提出了算法不同阶段之间的隐私预算分配方案。

3 预备知识

3.1 伪标签添加

使用大数据量客户端集群训练得到的教师模型[16]来为无标签的公共数据集添加标签。

定义 1(标签计数) 假设标签种类数为 k，每个教师模型 T_i 输出其对公共数据 $\vec{x}$ 的标签的预测结果 $f_i(\vec{x})$。对于给定类别 $j\in[k]$和数据 $\vec{x}$，认为 $\vec{x}$ 的标签为 j 的教师数量记为 $n_j(\vec{x})$，表达式如下：

$$n_j(\vec{x})=|\{i:i\in[m],f_i(\vec{x})=j\}| \tag{1}$$

定义 2(伪标签) 使用 $n_j(\vec{x})$计数最多的类别作为该数据的伪标签，考虑到数据隐私安全，引入噪声。数据的伪标签表达式如下：

$$f(x)=\arg\max_j\{n_j(\vec{x})+noise\} \tag{2}$$

3.2 联邦学习

本文的算法设计中采用的是联邦学习[23]中应用广泛的联邦平均算法[1]。

定义 3(联邦平均算法) 中心服务器初始化模型参数，训练执行若干轮，每轮选取至少 $1\sim K$ 个客户端参与训练，每个被选中的客户端 k 同时在本地根据服务器下发的本轮模型 w_t，用本地数据训练得到本地模型 w_{t+1}^k，并将 w_{t+1}^k 上传回服务器。服务器将收集到的各个客户端的模型根据各方样本数量用加权平均的方式进行聚合，得到下一轮的模型 w_{t+1}。

$$w_{t+1}\leftarrow\sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k \tag{3}$$

其中，n_k 表示客户端 k 上的样本数量，n 表示所有被选中客户端的总样本数量。

3.3 差分隐私

为了保证训练过程中的数据隐私，采用差分隐私[24]技术来保证客户端的数据安全。

定义 4（差分隐私） 设有随机算法 φ，P_φ 为 φ 所有可能的输出构成的集合。对于任意两个邻近数据集 D 和 D' 以及 P_φ 的任何子集 S_φ，若算法 φ 满足下式：

$$\Pr[\varphi(D)\in S_\varphi]\leqslant \exp(\varepsilon)\times\Pr[\varphi(D')\in S_\varphi] \tag{4}$$

则称算法 φ 提供 ε-差分隐私保护，其中参数 ε 称为隐私保护预算[25]。

在伪标签添加阶段，如式(2)所示，将噪声添加到标签计数上，防止教师模型训练数据集的隐私泄露；而在联邦学习阶段，在本地模型的每一层参数上添加噪声，防止服务器窃取客户端信息。

为了计算隐私预算消耗，引入 moments accountant[5]。

定义 5 设有随机机制 $\varphi: D\to R$，d，d' 是相邻数据集，aux 为辅助输入。moments accountant 定义为：

$$\alpha_\varphi(\lambda)\triangleq\max_{aux,d,d'}\alpha_\varphi(\lambda;aux,d,d') \tag{5}$$

对于输出 o，隐私损失变量的生成函数为 $\alpha_\varphi(\lambda;aux,d,d')\triangleq\log\mathbb{E}_{o\sim\varphi(aux,d)}[\exp(\lambda c(o;\varphi,aux,d,d'))]$，$c(o;\varphi,aux,d,d')\triangleq\log\dfrac{\Pr[\varphi(aux,d)=o]}{\Pr[\varphi(aux,d')=o]}$。

实际应用中，常需要多次用到差分隐私保护算法，隐私预算的分配遵循以下性质。

性质 1[26]（顺序组合性质） 算法 $\varphi_1,\varphi_2,\varphi_3,\cdots,\varphi_k$ 若分别满足 ε_1-DP，ε_2-DP，ε_3-DP，$\cdots$，ε_k-DP，那么对于同一数据集 D，由这些算法构成的组合算法 $\varphi(\varphi_1(D),\varphi_2(D),\varphi_3(D),\cdots,\varphi_k(D))$ 满足 $(\sum_{i=1}^{k}\varepsilon_i)-DP$。

性质 2[26]（平行组合性质） 算法 $\varphi_1,\varphi_2,\varphi_3,\cdots,\varphi_k$ 若分别满足 ε_1-DP，ε_2-DP，ε_3-DP，$\cdots$，ε_k-DP，而数据集 D_1，D_2，$\cdots$，D_k 不相交，由这些算法构成的组合算法 $\varphi(\varphi_1(D_1),\varphi_2(D_2),\cdots,\varphi_k(D_k))$ 满足 $(\max_{i\in[1,\cdots,k]}\varepsilon_i)-DP$。

4 基于知识蒸馏的差分隐私联邦学习方法

表 1 列出了本文中用到的数学符号。

表 1 常用符号说明

Table 1 Symbol description

符号	描述
m	大数据量客户端的数量
n	一般客户端的数量
c^l	大数据量客户端
c^s	一般客户端
c^{pub}	特殊客户端
D^{pub}	公共数据集
T	教师模型
k	数据标签的种类数
$f_i(\cdot)$	教师模型 i 的标签预测结果
$\vec{x}$	公共数据集的数据
$n_j(\cdot)$	预测数据标签为 j 的教师模型的个数
ε_1	第一阶段(伪标签添加)消耗的隐私预算总量
ε_2	第二阶段(联邦训练)消耗的隐私预算总量
ε_{pub}	特殊客户端的隐私预算
ε_{pri}	一般客户端的隐私预算
N	每轮训练参与训练的客户端数量

4.1 方案背景

假设所有参与训练的客户端都是可信的，不会攻击其他客户端；服务器是诚实且好奇的，可能会窃取客户端的隐私，鉴于此，引入差分隐私技术，保护各客户端的数据安全。

存在标签未知的公共数据，与之相对的，标签已知的隐私数据储存在各个客户端本地。各客户端的目的是联合训练一个全局模型，且在过程中保证本地数据的隐私。客户端之间的数据量存在差异，根据数据量大小的差异，我们将客户端分成两类：大数据量客户端 $C^l=\{c_1^l,c_2^l,c_3^l,\cdots,c_m^l\}$、一般客户端 $C^s=\{c_1^s,c_2^s,c_3^s,\cdots,c_n^s\}$。

算法框架如图 1 所示。

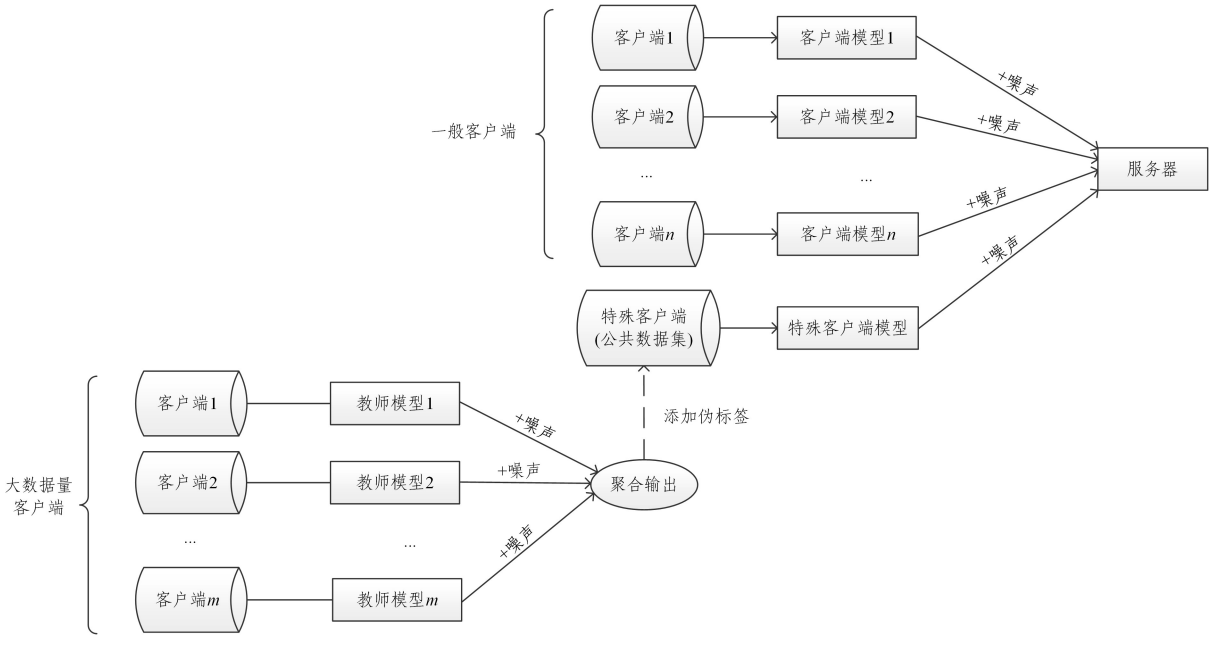


图 1 算法框架图

Fig. 1 Algorithm framework diagram

4.2 方案细节

本文提出的基于知识蒸馏的差分隐私联邦学习方法分为两个阶段：第一阶段是公共数据集的伪标签添加；第二阶段是联邦训练。其中第一阶段通过知识蒸馏技术，利用大数据量客户端的知识来为公共无标签数据进行伪标签添加，得到带标签的公共数据集，建立新的客户端，即特殊客户端；第二阶

段，一般客户端和特殊客户端共同进行联邦训练。两个阶段中都应用了差分隐私技术，以保护客户端的数据隐私。

4.2.1 第一阶段：伪标签添加

在这个阶段，我们使用大数据量客户端 $C^l=\{c_1^l, c_2^l, c_3^l, \cdots, c_m^l\}$ 训练得到教师模型 $T_1, T_2, \cdots, T_m$，然后用得到的教师模型对公共数据进行标签预测。

在进行标签预测的过程中，参考 PATE 框架[16]对标签计数添加噪声，既防止了客户端的隐私泄露，也不会严重影响伪标签的质量。由于噪声添加在标签计数上，标签计数是连续数值，因此噪声机制可以选择拉式噪声[2]、高斯噪声[3]。

算法 1 伪标签添加

输入：大数据量客户 $C^l=\{c_1^l, c_2^l, c_3^l, \cdots, c_m^l\}$，无标签公共数据集 D^{pub}，隐私预算 ε_1，数据标签种类数 k

输出：特殊客户端 c^{pub}

1. 每个大数据量客户端训练一个教师模型 T_i，$i=\{1,2,3,\cdots,m\}$
2. while $i<|D^{pub}|$ do//对每个公共数据集的数据添加标签
3. while $v<k$ do//对于每一类标签
4. while $j<m$ do//教师模型进行预测
5. 每个教师模型给出预测 $f_i(D_i^{pub})$
6. $n_v(D_i^{pub})=|\{j: j\in[m], f_j(D_i^{pub})=v\}|$
7. $label(D_i^{pub})=\arg\max_j\{n_j(\vec{x})+noise(\varepsilon_1)\}$
8. end while
9. end while
10. end while
11. D^{pub} 和 $label(D^{pub})$ 共同构成特殊客户端 c^{pub}
12. return c^{pub}

原本的公共数据集有了标签，作为特殊客户端 c^{pub} 参与后续的联邦学习。

4.2.2 第二阶段：联邦训练

联邦训练中，为避免数据重叠，我们并非简单地将公共数据集添加到每个参与联邦训练的客户端中，而是将公共数据视作一个特殊客户端 c^{pub} 参与联邦训练，即一般客户端 $C^s=\{c_1^s, c_2^s, c_3^s, \cdots, c_n^s\}$ 和特殊客户端 c^{pub} 共同进行联邦训练。在每一轮训练中，随机采样部分客户端进行训练。考虑到特殊客户端数据的共享性，设定特殊客户端参与每一轮的训练。在联邦学习中引入差分隐私技术，在每个客户端上传的模型参数处添加噪声。

算法 2 联邦训练

输入：一般客户端 $C^s=\{c_1^s, c_2^s, c_3^s, \cdots, c_n^s\}$，特殊客户端 c^{pub}，一般客户端的隐私预算 ε_{pri}，特殊客户端的隐私预算 ε_{pub}，隐私预算消耗上限 ε_2

输出：最终模型 w

1. $\varepsilon=0$//初始隐私消耗量设为 0
2. $t=0$//初始轮次设为 0
3. while $\varepsilon<\varepsilon_2$ do//只要未超过预算则训练
4. 在 C^s 中随机选择 N−1 个客户端，与 c^{pub} 共同构成本轮参与训练的客户端集群
5. for 客户端 k in 参与训练的客户端
6. 客户端 k 在 w_t 上用本地数据训练得到本地模型 w_{t+1}^k
7. if 客户端 k 为特殊客户端 c^{pub}：
8. $w_{t+1}^k=w_{t+1}^k+noise(\varepsilon_{pub})$
9. else：
10. $w_{t+1}^k=w_{t+1}^k+noise(\varepsilon_{pri})$
11. end if
12. end for
13. 本地模型上传到服务器，服务器根据联邦平均算法进行聚合得到全局模型 $w_{t+1}\leftarrow\sum_{k=1}^{N-1}\frac{n_k}{n}w_{t+1}^k$
14. 服务器下发聚合后的模型 w_{t+1}
15. $t=t+1$
16. $\varepsilon=\varepsilon+$本轮消耗隐私预算值
17. end while
18. return 最终模型 w

4.3 隐私预算分配

在联邦学习阶段，设定特殊客户端的隐私预算为 ε_{pub}，一般客户端的隐私预算为 ε_{pri}。根据性质 1，联邦学习阶段每轮消耗的隐私预算总量为 $\varepsilon_{pub}+(N-1)\times\varepsilon_{pri}$。

在整个训练过程中，标签添加阶段的隐私预算消耗为 ε_1，联邦学习阶段的隐私预算消耗为 ε_2。根据性质 2，整个训练阶段消耗的隐私预算总量为 $\max(\varepsilon_1,\varepsilon_2)$。

特殊客户端包含了公共数据集和大数据量客户端这两部分的知识，其本身也是大数据量客户端以一种更为间接的方式参与联邦训练的体现。特殊客户端包含的数据中只有标签涉及隐私，于是我们在训练中，为其分配比其他客户端更少的噪声。具体的噪声量分配情况以一般客户端和特殊客户端的隐私预算分配比例 $\varepsilon_{pri}/\varepsilon_{pub}$ 体现。在 5.4.1 节中，将通过实验探究隐私预算比例的取值对最终模型的影响。

根据性质 2，整个算法消耗的隐私预算总值取决于两阶段中消耗的最大值。由于服务器诚实而好奇的性质，在联邦训练中添加噪声是必要的；由于我们设定客户端都是可信的，第一阶段的噪声添加量则可以根据客户端对隐私性的需求决定。我们控制隐私预算消耗总量一定，控制联邦训练阶段隐私预算消耗一定，并且，由于第一阶段存在不添加噪声的可能性，设定 $\max(\varepsilon_1,\varepsilon_2)=\varepsilon_2$。在满足 $\varepsilon_1\leqslant\varepsilon_2$ 的前提下，第一阶段的隐私预算取值 ε_1 根据大数据量客户端对隐私性的需求来定，若是客户端要求较高的隐私性，则 ε_1 取较小值，反之则取较大值。在 5.4.2 节中，调整 ε_1 以匹配客户端不同的安全水平要求，探究 ε_1 对最终模型的影响。

5 理论分析

5.1 第一阶段：伪标签添加

在用 moment accountant 计算隐私预算消耗时，遵循以下性质。

性质 3 假设算法 φ 包含一系列自适应算法 $\varphi_1, \cdots, \varphi_k$，$\varphi_i: \prod_{j=1}^{i-1}R_j\times D\rightarrow R_i$，$\varphi_i(i<k)$ 的输出是 o_i。对于任意输入序列 $o_1, \cdots, o_{k-1}$ 和任意 λ：

$$\alpha_\varphi(\lambda; d, d')=\sum_{i=1}^{k}\alpha_{\varphi_i}(\lambda; o_1, \cdots, o_{i-1}, d, d')$$

证明：为简洁起见，用 $\varphi_{1:i}$ 表示 $(\varphi_1, \cdots, \varphi_i)$，$o_{1:i}$ 表示 $(o_1, \cdots, o_i)$。对于相邻数据集 $d, d'\in D^n$ 和一系列输出 $o_1, \cdots, o_k$，有：

$$c(o_{1:k}; \varphi_{1:k}, o_{1:(k-1)}, d, d')$$
$$=\log\frac{\Pr[\varphi_{1:k}(d; o_{1:(k-1)})=o_{1:k}]}{\Pr[\varphi_{1:k}(d'; o_{1:(k-1)})=o_{1:k}]}$$
$$=\log\prod_{i=1}^{k}\frac{\Pr[\varphi_i(d)=o_i\mid\varphi_{1:(i-1)}(d)=o_{1:(i-1)}]}{\Pr[\varphi_i(d')=o_i\mid\varphi_{1:(i-1)}(d')=o_{1:(i-1)}]}$$

$$=\sum_{i=1}^{k}\log\frac{\Pr[\varphi_i(d)=o_i\mid\varphi_{1:(i-1)}(d)=o_{1:(i-1)}]}{\Pr[\varphi_i(d')=o_i\mid\varphi_{1:(i-1)}(d')=o_{1:(i-1)}]}$$

$$=\sum_{i=1}^{k}c(o_i;\varphi_i,o_{1:(i-1)},d,d')$$

$$\mathbb{E}_{o'_{1:k}\sim\varphi_{1:k}(d)}[\exp(\lambda c(o'_{1:k};\varphi_{1:k},d,d'))\mid\forall\, i<k;o_i'=o_i]$$

$$=\mathbb{E}_{o'_{1:k}\sim\varphi_{1:k}(d)}[\exp(\lambda\sum_{i=1}^{k}c(o_i';\varphi_i,o_{1:(i-1)},d,d'))]$$

$$=\mathbb{E}_{o'_{1:k}\sim\varphi_{1:k}(d)}[\prod_{i=1}^{k}\exp(\lambda c(o_i';\varphi_i,o_{1:(i-1)},d,d'))]$$

$$=\prod_{i=1}^{k}\mathbb{E}_{o_i'\sim\varphi_i(d)}[\exp(\lambda c(o_i';\varphi_i,o_{1:(i-1)},d,d'))]$$

$$=\prod_{i=1}^{k}\exp(\alpha_{\varphi_i}(\lambda;o_{1:(i-1)},d,d'))$$

$$=\exp(\sum_{i=1}^{k}\alpha_i(\lambda;o_{1:(i-1)},d,d'))$$

$\alpha_\varphi(\lambda;d,d')=\sum_{i=1}^{k}\alpha_{\varphi_i}(\lambda;o_1,\cdots,o_{i-1},d,d')$得证。

性质 4 对于任意 $\varepsilon>0$，若 $\delta=\min_{\lambda}\exp(\alpha_\varphi(\lambda)-\lambda\varepsilon)$，则算法 φ 满足$(\varepsilon,\delta)-DP$。

证明

$$\Pr_{o\sim\varphi(d)}[c(o)\geqslant\varepsilon]=\Pr_{o\sim\varphi(d)}[\exp(\lambda c(o))\geqslant\exp(\lambda\varepsilon)]$$

$$\leqslant\frac{E_{o\sim\varphi(d)}[\exp(\lambda c(o))]}{\exp(\lambda\varepsilon)}$$

$$\leqslant\exp(\alpha-\lambda\varepsilon)$$

设 $B=\{o;c(o)\geqslant\varepsilon\}$，对于任意的 S 有：

$$\Pr[M(d)\in S]=\Pr[M(d)\in S\cap B^c]+\Pr[M(d)\in S\cap B]$$

$$\leqslant\exp(\varepsilon)\Pr[M(d')\in S\cap B^c]+\Pr[M(d)\in B]$$

$$\leqslant\exp(\varepsilon)\Pr[M(d')\in S]+\exp(\alpha-\lambda\varepsilon)$$

5.2 第二阶段：联邦训练

在联邦训练中，每一轮在每一个客户端上的隐私预算消耗之和为此阶段的隐私预算总消耗量。此过程遵循差分隐私的顺序组合性质。

给出 3.3 节中性质 1 顺序组合性质的具体证明如下：

性质 1 证明：

先证明：若算法 φ_1,φ_2 分别满足 $\varepsilon_1-DP,\varepsilon_2-DP$，那么对于同一数据集 D，由这些算法构成的组合算法 $\varphi(D)=\varphi_2(\varphi_1(D),D)$满足$(\varepsilon_1+\varepsilon_2)-DP$。

设 D 和 D' 是相邻数据集，设算法 φ_1 的输出集为 $Range(\varphi_1)$，记为 S，对于任意的输出结果 $t\in Range(\varphi_2)$有：

$$\Pr[\varphi_2(\varphi_1(D),D)=t)]$$

$$=\sum_{s\in S}\Pr[\varphi_1(D)=s]\Pr[\varphi_2(s,D)=t]$$

$$\leqslant\sum_{s\in S}e^{\varepsilon_1}\Pr[\varphi_1(D')=s]e^{\varepsilon_2}\Pr[(\varphi_2(s,D')=t]$$

$$=e^{\varepsilon_1+\varepsilon_2}\Pr[(\varphi_2(\varphi_1(D'),D')=t]$$

得证算法 $\varphi(D)=\varphi_2(\varphi_1(D),D)$满足$(\varepsilon_1+\varepsilon_2)-DP$。

若有算法 $\varphi_1,\varphi_2,\cdots,\varphi_k$ 分别满足 $\varepsilon_1-DP,\varepsilon_2-DP,\cdots,\varepsilon_k-DP$，对于数据集 D，记由这些算法构成的组合算法为 $\varphi(D)=\varphi(\varphi_1(D),\varphi_2(D),\varphi_3(D),\cdots,\varphi_k(D))$。$\varphi(D)$在数据集 D 上依次进行查询$\langle t_1,t_2,\cdots,t_k\rangle$ $t_1=\varphi_1(D)$，$t_2=\varphi_2(t_1,D),\cdots,t_k=\varphi_k(\langle t_1,\cdots,t_{k-1}\rangle,D)$。我们已经证得算法 $\varphi(D)=\varphi_2(\varphi_1(D),D)$ 满足$(\varepsilon_1+\varepsilon_2)-DP$，所以推导可得组合算法 $\varphi(D)=\varphi(\varphi_1(D),\varphi_2(D),\varphi_3(D),\cdots,\varphi_k(D))$满足$(\sum_{i=1}^{k}\varepsilon_i)-DP$。

所以性质 1 成立。

5.3 隐私预算总消耗

记第一阶段消耗隐私预算量为 ε_1，第二阶段消耗隐私预算量为 ε_2，根据遵循差分隐私的平行组合性质，隐私预算总消耗量为 $\max(\varepsilon_1,\varepsilon_2)$。

给出 3.3 节中性质 2 平行组合性质的具体证明如下：

性质 2 证明：设 D 和 D'是相邻数据集，不失一般性，假设 D 比 D'多包含一个元素。设 D 划分为 $D_1,D_2\cdots D_k$，D'划分为 $D_1',D_2'\cdots D_k'$。存在 j 满足：1）D_j 比 D_j'多包含一个元素；2）对于任意 $i\neq j$，$D_i\neq D_i'$。用 $\varphi(D)$表示 $\varphi_1(D_1)$，$\varphi_2(D_2),\cdots,\varphi_k(D_k)$。因为 k 种算法分别运行在不相交的数据集 D_i 上，所以对于任意输出序列 $t=(t_1,t_2,\cdots,t_k)$，$t_i\in Range(\varphi_i)$，有：

$$\Pr[\varphi(D)=t]=\Pr[(\varphi_1(D_1)=t_1)\wedge\cdots\wedge(\varphi_k(D_k)=t_k)]$$

$$=\Pr[\varphi_j(D_j)=t_j]\prod_{i\neq j}\Pr[\varphi_i(D_i)=t_i]$$

$$\leqslant e^{\varepsilon_j}\Pr[\varphi_j(D'_j)=t_j]\prod_{i\neq j}\Pr[\varphi_i(D_i')=t_i]$$

$$\leqslant e^{\max_{i\in[1,\cdots,k]}\varepsilon_i}\Pr[\varphi(D')=t]$$

6 实验

本章首先介绍了实验设置与数据集，然后是具体的实验。实验分为 3 个部分：参数分析、对比试验分析、隐私预算分配探究。

6.1 实验设置与数据集

实验使用的是美国国家标准与技术研究所发起整理的手写数字数据集 MNIST[27-28] 和来自谷歌的 SVHN[29] 数据集。

MNIST 数据集训练集由 250 个不同的人手写的数字构成，其中 50%是高中学生，50%来自人口普查局的工作人员，测试集也是同样比例的手写数字数据。图像像素值为 0～255 之间，标签是 0-9 的数值，因为是黑白图像，图像格式为 28×28。训练数据集有 60000 个，测试数据集有 10000 个。

SVHN 数据集来源于谷歌街景门牌号码，每张图片中包含一组 0－9 的阿拉伯数字。训练集中包含 73257 个数字，测试集中包含 26032 个数字，另有 531131 个附加数字。数据为彩色图像，格式为 32×32×3。

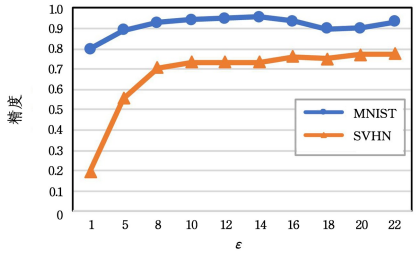
实验选用训练数据集来划分为数据量不等的客户端，最终模型在测试集上测试精度。以客户端总量为 50 为例，对于 MNIST 数据集，将训练集分为 4 个数据含量为 2500 的大数据量客户端和 49 个数据含量为 1000 的一般客户端，剩余的作为公共数据集；对于 SVHN 数据集，将训练集分为 3 个数据含量为 4419 的大数据量客户端和 49 个数据含量为 1200 的一般客户端，剩余的作为公共数据集。

6.2 参数分析

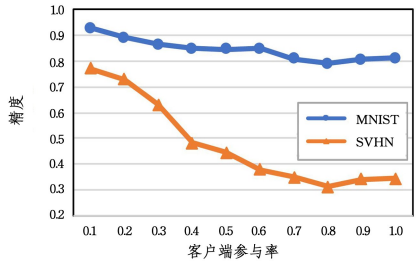
在本文算法中，影响实验结果的参数有隐私预算和客户端训练参与率。

对于隐私预算对实验结果的影响分析实验，本文将隐私预算的取值区间设置为[1,22]；而对于客户端训练参与率影响的分析实验，本文设置客户端训练参与率的取值区间设置为[0.1,1]。

如图 2 所示，在其他条件不变的情况下，随着隐私预算的增加，模型精度不断提升，对于 MNIST 数据集来说，提升趋势较为缓慢，而对于 SVHN 数据集来说，在低隐私预算时模型精度的提升较为明显，而在高隐私预算时趋势渐缓。



(a)隐私预算的影响



(b) 客户端参与率的影响

图 2 参数分析

Fig. 2 Parameter analysis

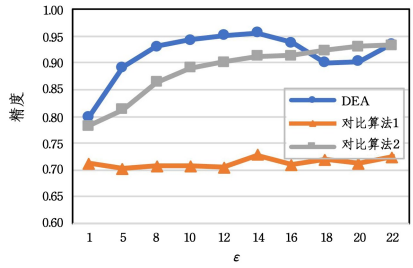
而对于客户参与率来说，在其他条件不变的情况下，随着参与率的提高，两个数据集的模型精度都出现了下降趋势，这是因为即使提高客户端参与率可以增加训练数据，但同时训练方的增加导致了噪声的增加，从而造成了精度的下降。其中，MNIST 数据集下降趋势较为平缓，而 SVHN 数据集下降趋势更为明显。推测其原因是，SVHN 数据集相比 MNIST 数据集来说数据维度更大，通常用更复杂的模型来学习。为了确保公平性，我们使用同样的模型对两个数据集进行训练，于是，在面对更多噪声引入时，SVHN 数据集精度下降趋势更为明显。

6.3 对比实验分析

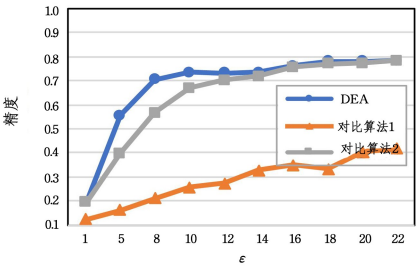
本文选取同样建立在存在无标签公共数据集的数据背景下，且适用于分布式架构的算法——PATE 来作为对比算法 1，PATE 算法同样使用差分隐私来保护客户端隐私。将同样数据背景下的传统算法作为基线，来观察本文算法的效果。对比算法 2 在本文算法的基础上，将第一阶段参与训练的大数据量客户端再次添加到联邦学习的训练阶段，与之进行对比是为了说明在第一阶段将大数据量客户端的知识转移到公共数据集后，在第二阶段的联邦学习中大数据量客户端参与训练会对模型带来的影响。

本文设定轨迹隐私预算值 ε 的取值区间为[1,22]，实验结果如图 3 所示。从实验结果可以看出，随着隐私预算增加，本文算法、对比算法 1、对比算法 2 的整体变化趋势一致，精度缓慢上升。隐私预算从 14 开始呈现下降趋势，到 18 又开始上升。这看起来不符合预期，但精度波动最大也不超过 5%，属于正常波动。在分配相同隐私预算值的情况下，本文算法的精度普遍更高。这种精度上的优势在与对比算法 1 进行比较时十分明显，这种精度上的领先在 MNIST 数据集上最高能达到 25%左右，在 SVHN 数据集上最高能达到 46%左右。究其原因，在对比算法 1 中，大量的隐私数据被用来训练教师模型为无标签公共数据集添加伪标签，而在最终模型的训练中只用到了公共数据集，公共数据集数量远少于隐私数据，最终模型的训练数据量影响了实验效果。

需要注意的是，在分配更高隐私预算值时，MNIST 数据集上对比算法 2 的效果更好一些，另外在 SVHN 数据集上，高隐私预算值时，其表现与本文算法不相上下。分析是因为对比算法 2 在联邦学习中相比本文算法增加了额外的训练方和额外的训练数据。在隐私预算较小的情况下，相比本文方法，对比算法 2 的训练数据的增加并不足以抵消噪声对模型的负面影响，而在较高隐私预算时，则体现出了训练数据增加的好处。



(a) MNIST



(b) SVHN

图 3 对比实验

Fig. 3 Contrast test

6.4 隐私预算分析

由图 4—图 6 可见，面临客户端间隐私预算分配调整时，模型效果仍保持高度稳定，精度浮动最大不超过 0.4%(MNIST)，4%(SVHN)；面临算法各阶段间隐私预算分配调整时，模型效果仍保持稳定，精度浮动最大不超过 7%(MNIST)，4%(SVHN)。

由此可见，模型在不同隐私预算分配方案下仍可以保持精度的稳定，隐私预算分配灵活度高。因此，本方案具有可拓展性，可以满足复杂的隐私安全需求。比如，某类客户端要求分配更多的隐私预算时，可以通过调整客户端间隐私预算分配或者算法各阶段间隐私预算分配，来满足新的隐私安全需求。

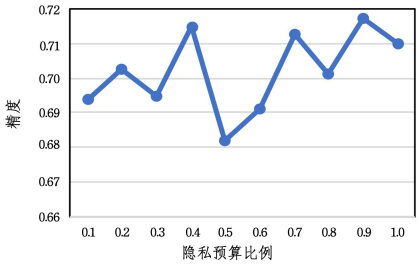
接下来探讨不同隐私预算分配方案对模型的影响。

6.4.1 客户端间隐私预算分配方案

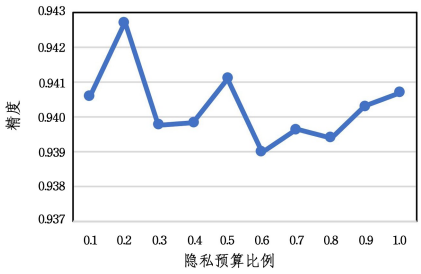
在本文算法的第二阶段联邦训练中存在两类客户端，一类是建立在隐私数据上的一般客户端，另一类是建立在公共数据上的特殊客户端。我们控制隐私预算消耗的总量一定，也就是控制模型的安全性能一定。在这样的前提下，调整一般客户端和特殊客户端的隐私预算分配比例 $\varepsilon_{pri}/\varepsilon_{pub}$，以最终模型的精度为指标，探究了隐私预算分配比例对最终模型效果的影响。如 4.3 节所述，我们的方案为特殊客户端分配比一般客户端更少量的噪声，于是 $\varepsilon_{pri}/\varepsilon_{pub}$ 取值小于 1。

如图 4 所示，在两个数据集的实验表明，在[0,1]范围内调整隐私预算分配比例对模型的精度并无明显的影响，随着 $\varepsilon_{pri}/\varepsilon_{pub}$ 比例取值变大，精度只是呈现出不规则的波动，且不超过 3%。于是可以得出结论，即使两类客户端的数据真实性并不一致(特殊客户端的标签并不是真实数据)，为两者分配不同的隐私预算对模型精度的影响可以忽略不计。这也从侧面说明了第一阶段得到的伪标签十分可靠，使得特殊客户端

在联邦训练中和其他客户端之间的差异并不明显。



(a)MNIST



(b)SVHN

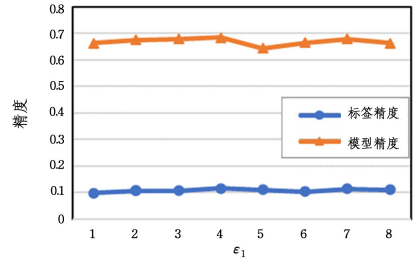
图 4　客户端之间隐私预算分配(ε=8)

Fig. 4　Privacy budget allocation between clients(ε=8)

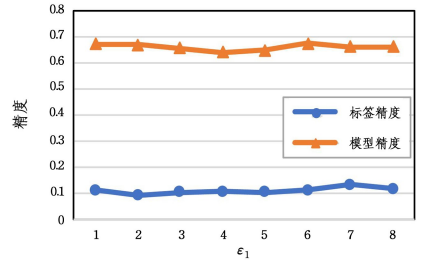
6.4.2　阶段间隐私预算分配方案

如第 2 章所述，本文算法分为两个阶段，且两个阶段都有隐私预算的消耗。在实验中，通过调整隐私预算的取值，来观察其对标签精度和模型精度的影响。如图 5、图 6 所示，当隐私预算取值很大时，标签精度才会有较大的变化幅度。我们分别观察了在 $\max(\varepsilon_1,\varepsilon_2)=\varepsilon_2=8$，$\max(\varepsilon_1,\varepsilon_2)=\varepsilon_1>8$ 两种情况下，ε_1 对最终模型的影响。

如图 5 所示，在控制总隐私预算为 8 的情况下，ε_1 的取值对标签精度和最终模型精度都无较大影响。如图 6 所示，在 $\varepsilon_1\geqslant 8$，也就是总预算消耗为 ε_1 时，若 ε_1 取到一定的值，其对标签精度确实会有较大的影响，但对于最终模型精度无明显影响。而且，在这种取值情况下，ε_1 已达到上百甚至上千，实际应用中这种超大隐私预算的情况较为罕见。



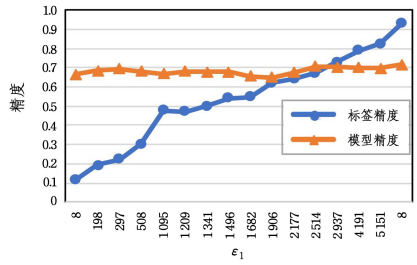
(a)MNIST



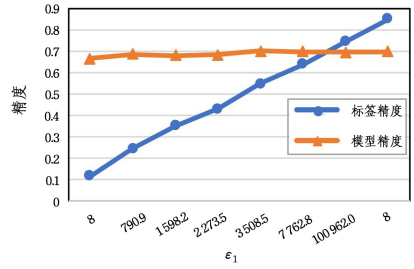
(b)SVHN

图 5　阶段之间隐私预算分配(ε=8)

Fig. 5　Allocation of privacy budget between stages(ε=8)



(a)MNIST



(b)SVHN

图 6　阶段之间隐私预算分配(ε≥ 8)

Fig. 6　Allocation of privacy budget between stages(ε≥ 8)

于是我们得出结论，在控制 ε_2 的前提下，ε_1 的取值对最终模型无明显影响，因此对于 ε_1 取值的选择只需考虑客户端对隐私性的需求。

结束语　本文综合考虑了无标签公共数据集的存在和客户端之间数据量的差异，提出了一种基于知识蒸馏的差分隐私联邦学习方法。首先，将客户端按数据量大小分组为大数据量客户端和一般客户端，以大数据量客户端的数据训练教师模型，教师模型为公共数据集添加伪标签，然后公共数据集作为特殊客户端与一般客户端共同进行联邦训练。为算法设计了专用的隐私保护方案，由于特殊客户端的数据只有标签涉及隐私，在联邦训练中为特殊客户端分配比一般客户端更多的隐私预算；限制隐私预算总量，设联邦训练阶段的隐私预算为定值，根据客户端对隐私性的需求和隐私预算平行组合性质，调整伪标签添加阶段的隐私预算。

相比传统的结合差分隐私技术的联邦学习，本文考虑到了无标签公共数据集的存在和客户端之间数据量的差异，且实验表明，本文方法能够在同等的隐私预算消耗下，经过训练得到更高精度的模型。此外，本方案具有可拓展性，高灵活度的隐私预算分配使其可以满足复杂的隐私需求。

本文方案建立在客户端数据量有较大差异且存在无标签公共数据集的数据背景下，而现有工作中，数据背景相似的研究十分少见，这导致了本文对比算法的缺乏。此外，实验表明本文算法的隐私预算分配存在一定的灵活性，这为本研究在个性化联邦领域中的拓展应用提供了可能性。

参 考 文 献

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017: 1273-1282.

[2] DWORK C, MC-SHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography Conference. 2006: 265-284.

[3] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our Da-

ta, Ourselves: Privacy Via Distributed Noise Generation[C]// International Conference on Advances in Cryptology-eurocrypt. DBLP, 2006: 486-503.

[4] PHAN N H, WU X, HU H, et al. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning[C]// IEEE International Conference on Data Mining. 2017: 385-394.

[5] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM, 2016: 308-318.

[6] LU Y, HUANG X, DAI Y, et al. Differentially private asynchronous federated learning for mobile edge computing in urban informatics[J]. IEEE Transactions on Industrial Informatics, 2020, 16(3): 2134-2143.

[7] WEI K, LI J, DING M, et al. Federated learning with differential privacy: algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[8] MC MAHAN H B, RAMAGE D, TALWAR K, et al. Learning differentially private recurrent language models[J]. arXiv: 1710.06963, 2017.

[9] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503.02531, 2015.

[10] LIU L, ZHANG J, SONG S H, et al. Communication-Efficient Federated Distillation with Active Data Sampling[J]. arXiv: 2203.06900, 2022.

[11] ITAHARA S, NISHIO T, KODA Y, et al. Distillation-Based Semi-Supervised Federated Learning for Communication-Efficient Collaborative Training with Non-IID Private Data[J]. IEEE Transactions on Mobile Computing, 2021(1): 191-205.

[12] SUN L, LYU L. Federated Model Distillation with Noise-Free Differential Privacy[C]// International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2021: 1563-1572.

[13] ZHAO Y, LI M, LAI L, et al. Federated learning with non-iid data[J]. arXiv: 1806.00582, 2018.

[14] YAO X, HUANG T, ZHANG R X, et al. Federated learning with unbiased gradient aggregation and controllable meta updating[J]. arXiv: 1910.08234, 2019.

[15] ZHU L, LIU X, LI Y, et al. A Fine-Grained Differentially Private Federated Learning Against Leakage From Gradients[J]. IEEE Internet of Things Journal, 2022, 9(13): 11500-11512.

[16] PAPERNOT N, ABADI M, ERLINGSSON L, et al. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data[J]. arXiv: 1610.05755, 2016.

[17] KERKOUCHE R, CS G, CASTELLUCCIA C, et al. Constrained Differentially Private Federated Learning for Low-bandwidth Devices[J]. arXiv: 2103.00342, 2021.

[18] PKAIROUZ H, MCMAHAN B, AVENT B, et al. Advances and Open Problems in Federated Learning[J]. arXiv: 1912.04977, 2019.

[19] SHI H, ZHANG Y, SHEN Z, et al. Towards Communication-Efficient and Privacy-Preserving Federated Representation Learning[J]. arXiv: 2109.14611, 2021.

[20] ZHANG T, SONG A, DONG X, et al. Privacy-Preserving Asynchronous Grouped Federated Learning for IoT[J]. IEEE Internet of Things Journal, 2022, 9(7): 5511-5523.

[21] HUANG X, DING Y, JIANG Z L, et al. DP-FL: a novel differentially private federated learning framework for the unbalanced data[J]. World Wide Web, 2020, 23: 2529-2545.

[22] LIU J, LOU J, XIONG L, et al. Projected Federated Averaging with Heterogeneous Differential Privacy[J]. PVLDB, 2022, 15(4): 828-840.

[23] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [J]. Artificial Intelligence and Statistics, 2017: 1273-1282.

[24] DWORK C. Differential Privacy[J]. Lecture Notes in Computer Science, 2006, 26(2): 1-12.

[25] HAEBERLEN A, PIERCE B C, NARAYAN. Differential privacy under fire[C]// Proceedings of the 20th USENIX Conference on Security. San Francisco, USA, 2011: 33-33.

[26] MCSHERRY F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis[J]. Communications of the ACM, 2010, 53(9): 89-97.

[27] LECUN Y. The MNIST Database of Handwritten Digits[OL]. http://yann.lecun.com/exdb/mnist/.

[28] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]// NIPS Workshop on Deep Learning & Unsupervised Feature Learning. 2011.

TAN Zhiwen, born in 1999, postgraduate. Her main research interests include federated learning and differentially privacy.

XU Ruzhi, born in 1966, Ph.D, professor, master supervisor. Her main research interests include information safety, application of information technology in smart grid, and computer control.