基于 GAN 的联邦学习成员推理攻击与防御方法

张佳乐[1,2]，朱诚诚[1,2]，孙小兵[1,2]，陈兵[3]

（1. 扬州大学信息工程学院，江苏 扬州 225127；2. 江苏省知识管理与智能服务工程研究中心，江苏 扬州 225127；
3. 南京航空航天大学计算机科学与技术学院，江苏 南京 211106）

摘 要：针对联邦学习系统极易遭受由恶意参与方在预测阶段发起的成员推理攻击行为，以及现有的防御方法在隐私保护和模型损失之间难以达到平衡的问题，探索了联邦学习中的成员推理攻击及其防御方法。首先提出 2 种基于生成对抗网络（GAN）的成员推理攻击方法：类级和用户级成员推理攻击，其中，类级成员推理攻击旨在泄露所有参与方的训练数据隐私，用户级成员推理攻击可以指定某一个特定的参与方；此外，进一步提出一种基于对抗样本的联邦学习成员推理防御方法（DefMIA），通过设计针对全局模型参数的对抗样本噪声添加方法，能够在保证联邦学习准确率的同时，有效防御成员推理攻击。实验结果表明，类级和用户级成员推理攻击可以在联邦学习中获得超过 90%的攻击精度，而在使用 DefMIA 方法后，其攻击精度明显降低，接近于随机猜测（50%）。

关键词：联邦学习；成员推理攻击；生成对抗网络；对抗样本；隐私泄露

中图分类号：TP391

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023094

Membership inference attack and defense method in federated learning based on GAN

ZHANG Jiale[1,2], ZHU Chengcheng[1,2], SUN Xiaobing[1,2], CHEN Bing[3]

1. School of Information Engineering, Yangzhou University, Yangzhou 225127, China
2. Jiangsu Engineering Research Center Knowledge Management and Intelligent Service, Yangzhou 225127, China
3. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract: Aiming at the problem that the federated learning system was extremely vulnerable to membership inference attacks initiated by malicious parties in the prediction stage, and the existing defense methods were difficult to achieve a balance between privacy protection and model loss. Membership inference attacks and their defense methods were explored in the context of federated learning. Firstly, two membership inference attack methods called class-level attack and user-level attack based on generative adversarial network (GAN) were proposed, where the former was aimed at leaking the training data privacy of all participants, while the latter could specify a specific participant. In addition, a membership inference defense method in federated learning based on adversarial sample (DefMIA) was further proposed, which could effectively defend against membership inference attacks by designing adversarial sample noise addition methods for global model parameters while ensuring the accuracy of federated learning. The experimental results show that class-level and user-level membership inference attack can achieve over 90% attack accuracy in federated learning, while after using the DefMIA method, their attack accuracy is significantly reduced, approaching random guessing (50%).

Keywords: federated learning, membership inference attack, generative adversarial network, adversarial example, privacy leakage

收稿日期：2022−11−25；**修回日期：**2023−03−06

通信作者：朱诚诚，1229653863@qq.com

基金项目：国家自然科学基金资助项目（No.62206238）；江苏省自然科学基金资助项目（No.BK20220562）；江苏省高等学校基础科学（自然科学）研究基金资助项目（No.22KJB520010）；扬州市科技计划项目–市校合作专项基金资助项目（No.YZ2021157, No.YZ2021158）

Foundation Items: The National Natural Science Foundation of China (No.62206238), The Natural Science Foundation of Jiangsu Province (No.BK20220562), The Natural Science Foundation of Jiangsu Higher Education Institutions of China (No.22KJB520010), The Yangzhou City-Yangzhou University Science and Technology Cooperation Fund Project (No.YZ2021157, No.YZ2021158)

0 引言

近年来，随着物联网、边缘计算、5G 等技术的不断发展及用户终端数量的爆炸式增长，传统云计算架构下的集中式机器学习模型由于其高时延、高并发、弱隐私保护等缺陷，已经逐渐演化为能够支撑边缘智能化应用的分布式联邦学习架构。联邦学习在结构上具有特殊的隐私保护性，它允许各个参与方（用户）下载全局模型到本地，利用本地数据对模型进行训练并更新参数，最终这些参数被汇总到服务器端进行聚合平均，生成新的全局模型[1-4]。在这一过程中，由于参与方的数据保留在本地终端，其隐私得到了较大程度的保护。可以说，联邦学习已呈现出极具实用性的发展潜力，有关联邦学习的研究方向也被国内外学者广泛关注。

联邦学习尽管在隐私保护方面取得了一定突破，但是仍面临着诸多安全与隐私问题。大量研究成果表明，联邦学习框架极易受到各种推理攻击的威胁，如成员推理[5]、特征推理[6]、属性推理[7]和梯度推理[8]。其中，成员推理是针对训练数据集的主动攻击之一，其目的是确定某个数据样本是否被用于模型训练过程。Shokri 等[9]首次在机器学习模型中提出了通过黑盒应用程序接口（API）构造的成员推理攻击方法，并证实了通过区分训练和非训练样本在模型输出结果上的差异，能够成功获得某一预测样本是否为模型训练数据成员的信息。在联邦学习的场景中，攻击者的角色是复杂多变的，其既可以作为参与方加入训练过程（如图 1 所示），不断获取服务器反馈的全局模型参数并进行成员推理攻击[10-11]，也可以作为不可信的中央服务器，通过收集参与方上传的本地参数来推理训练数据[12]。

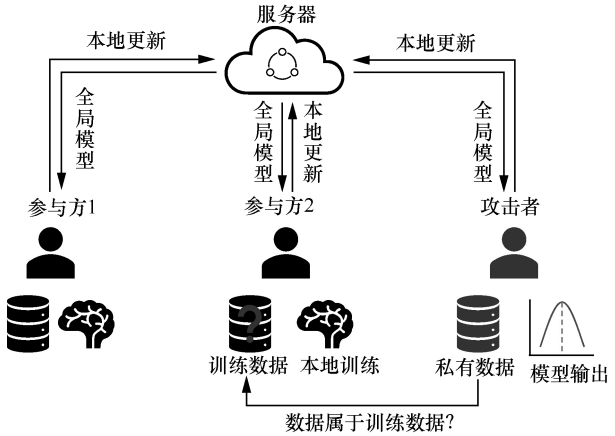

图 1 联邦学习中的成员推理攻击模型

此外，生成对抗网络（GAN, generative adversarial network）的广泛应用进一步加强了成员推理攻击对联邦学习的威胁强度[13]。通过使用 GAN 中的辨别器和生成器，攻击者可以利用训练模型中的参数来生成伪样本，或者获取与其他参与方训练数据集相同分布的数据样本[14]，这些生成的数据样本对成员推理攻击的发起提供了有力途径。

目前，传统机器学习模型的成员推理防御机制主要集中于参数保护方面，即防止攻击者获得未受保护的模型参数，常用的方法有安全聚合[15]、同态加密[16]和差分隐私[17]等。除此之外，也有部分研究表明，通过在攻击模型中加入一个特定的噪声向量，能够成功地将隐私效用权衡的效果最大化[18]。然而，上述对模型参数做直接修改的防御方法将导致模型精度的严重下降，同时，复杂的密码算法也给资源受限的参与方带来巨大的计算资源消耗。此外，联邦学习中参与方训练数据的不可见性也增加了防御难度。

针对上述问题，本文首先探索了 GAN 模型在联邦学习成员推理攻击中的增强效果，并基于 GAN 生成的增强数据，提出 2 种针对不同联邦学习场景的成员推理攻击方法：类级和用户级成员推理攻击。其中，类级成员推理攻击主要适用于横向联邦学习场景，即攻击者仅关注所推理样本是否属于训练数据，并不在意该样本具体属于哪个用户；与之相反，用户级成员推理攻击旨在推断出联邦学习中特定参与方的训练数据信息，攻击者不仅能够推理出某一指定样本是否属于训练数据，还可以判断出该样本属于哪个用户，用户级攻击可应用于参与方训练数据的类标签相互独立的场景，如纵向联邦学习。此外，针对上述 2 种攻击类型，本文进一步提出一种基于对抗样本的联邦学习成员推理攻击防御方法（DefMIA），在保证全局模型准确率的同时，使联邦学习中成员推理攻击准确率接近于随机猜测。

本文的主要贡献包括以下 3 个方面。

1) 提出了 2 种基于 GAN 的联邦学习成员推理攻击方法：类级和用户级成员推理攻击，其中，类级成员推理攻击旨在泄露所有参与方的训练数据，用户级成员推理攻击则面向某一特定的联邦学习参与方，并证明了这 2 种攻击在联邦学习场景中的有效性。

2) 提出了 DefMIA 方法，将全局模型的输出置信度向量构造为对抗样本，使其能够误导攻击模型的正确分类结果，进而有效防御成员推理攻击。相

较于传统基于差分隐私或密码系统的防御方法，DefMIA 方法不需要对模型参数进行任何修改，最大化地保证了全局模型的可用性。

3) 在 3 个基准数据集上对所提 2 种攻击方法进行了实验测试，并验证了 DefMIA 方法的有效性。实验结果表明，无论是类级成员推理攻击还是用户级成员推理攻击，都能在联邦学习上获得较高攻击准确率。然而，在使用 DefMIA 方法后，上述 2 种方法的攻击准确率大大降低，接近于随机猜测。

1　相关工作

1.1　对抗性攻击

尽管联邦学习能够在架构上对参与方的训练数据隐私提供一定的保护作用，但模型安全问题仍然存在[19]。其主要原因在于联邦学习中的本地训练数据对中心服务器是不可见的，导致服务器无法验证上传参数的准确性。本文将针对联邦学习模型发起的攻击类型统称为对抗性攻击，其主要通过干扰联邦学习训练或推理过程，来影响联邦学习模型的收敛速度或推理结果。近年来，有关联邦学习对抗性攻击的研究成果包括模型反转攻击[20]、投毒攻击[21-22]、对抗样本攻击[23]等，这些攻击可按不同目的分为保密性攻击、完整性攻击和可用性攻击[24]。其中，保密性攻击指泄露、窃取用户敏感数据的攻击类型，该类攻击不仅试图窃取本地训练数据，还试图暴露隐私数据或反向推断出训练模型[3]，但该攻击中的攻击者不会干扰训练进度或修改模型，而是通过扮演参与方的角色来间接发起攻击；与保密性攻击不同，完整性攻击的目的是通过毒化模型使模型产生错误的输出，典型的完整性攻击包括标签翻转[25]和后门攻击[26]，它们通过训练误导目标模型，使其输入攻击者指定的预测结果[27]；可用性攻击的目的是攻击分类的可用性，通过恶意后门、对抗样本等方法[28]，使联邦学习中的目标模型无法使用。

1.2　成员推理攻击

成员推理攻击是一种针对机器学习模型的隐私攻击类型，通过获取模型的预测向量，以确定预测数据是否来自模型的原始训练数据集[11]。在传统的机器学习与联邦学习中都能够进行成员推理攻击，可对用户信息安全造成严重的威胁[29]。Shokri 等[9]首次提出了可用于分类器模型的成员推理攻击方法，通过模拟与目标模型相似的影子模型来获取模型的预测输出向量，进而创建有关成员信息的二分类模型，以判断某一给定样本是否属于原始训练数据集。随后，Nasr 等[12]提出针对联邦学习系统的成员推理攻击，在该攻击场景中，攻击者可以假扮良性参与方，从服务器获取聚合模型，以发起面向所有参与方训练数据的成员推理攻击。此外，基于 GAN 模型在联邦学习中的数据增强特性，Zhang 等[30]探索了 GAN 模型在构造成员推理攻击方案中的作用，攻击者通过部署一个 GAN 模型来生成攻击模型训练样本[31]，从而有效提升成员推理攻击的准确率；Chen 等[10]提出在联邦学习场景下，成员推理攻击的发起方也可以是不可信服务器，通过在服务器端部署多任务 GAN 模型，攻击者能够获取用户级的伪样本[6]，从而利用各参与方上传的本地模型发起细粒度成员推理攻击。值得注意的是，不管是用户还是服务器，攻击者均能够部署 GAN 模型以生成伪样本，进而增加攻击成功率。

1.3　成员推理防御

在联邦学习场景下，由于各参与方从服务器端获取全局模型，且本地训练过程完全自主化，全局模型参数很容易被恶意参与方获取。因此，与传统的机器学习相比，联邦学习更易受到成员推理攻击。针对上述问题，研究者进行了广泛的研究，文献[32]表明，在训练阶段引入对抗样本可以误导攻击模型的预测结果。根据这一发现，Jia 等[18]提出了一种基于效用损失的方法来防御黑盒设置下的成员推理攻击，通过将精心设计的噪声添加到模型的置信度向量中，攻击模型可能会被误导为随机结果。该方法尽管对成员推理攻击起到了一定的抵制作用，但会对模型收敛速度与模型精度造成负面影响。

2　联邦学习中的成员推理攻击

2.1　威胁模型

不失一般性，本文假设攻击者为联邦学习中的某一恶意用户[9,10,12,30]，该用户通过在本地部署 GAN 模型，生成接近真实样本的伪样本并将其作为攻击模型的训练数据，进而实现高精度的成员推理攻击。具体来说，本文提出的联邦学习成员推理攻击威胁模型包括以下 3 个方面。

1) 攻击目标。对于类级成员推理攻击，攻击者的目标是通过获取联邦学习中的传输参数，在不影响全局模型的前提下，隐蔽推理出某一给定样本的成员信息；对于用户级成员推理攻击，攻击者的目标不局限于推理出某一指定样本是否属于模型训练

数据，其最终目的为判断出该样本属于哪个用户。因此，本文所提成员推理攻击威胁模型的分类任务可以由以下指标来衡量：①成员推理准确率，即给定样本在成员推理模型上的分类性能；②主任务准确率，即联邦学习全局模型的性能。

2) 攻击知识。在联邦学习中，攻击者可以观察到系统初始化模型结构和每个通信轮次下发的全局模型参数。因此，攻击者可以获得联邦学习模型的每个细节，包括模型结构、学习算法 L 和模型参数 θ，这些信息可用于训练 GAN 模型生成伪样本以及成员推理攻击的初始化。

3) 攻击能力。攻击者具有以下能力：①获得每个通信轮次的全局模型；②作为参与方控制本地训练和本地数据；③修改本地模型的超参数；④随机更新和选择本地参数。同时，攻击者存在以下限制：①无法获得其他参与方上传的本地模型参数（各参与方的本地模型参数仅用于在服务器端的加权平均过程）；②无法访问其他参与方的本地数据。

2.2 类级成员推理攻击

2.2.1 攻击概述

在联邦学习场景下，全局模型参数由所有参与方的本地模型参数聚合产生，因此，由恶意参与方发起的成员推理攻击主要用于判断成员信息和类标签信息，即类级成员推理攻击。具体来说，在联邦学习场景下，攻击者可以观察到全局模型结构和参数信息，其攻击模型对于攻击者来说是一个白盒设置。基于上述白盒设置，攻击者作为参与方加入联邦学习并发起类级成员推理攻击，即攻击者使用 GAN 生成的数据训练二分类攻击模型，以区分训练数据中的成员和非成员样本。图 2 所示为类级成员推理攻击模型。利用目标模型的真实标签和预测结果来训练攻击模型，攻击模型通过学习模型输出的分布来区分目标模型的成员和非成员。本文定义 $f_{\text{target}}()$ 为目标模型，$D_{\text{target}}^{\text{train}}$ 为目标模型的训练集，其中，数据样本 $(x\{i\},y\{i\})_{\text{target}} \in D_{\text{target}}^{\text{train}}$，$x\{i\}$ 表示目标模型的输入，$y\{i\}$ 表示来自标签类 c_{target} 的 $x\{i\}$ 的真实标签。由于联邦学习模型的预测结果 $Y = f_{\text{target}}(x)$ 高度依赖于真实标签，因此本文使用目标模型的预测输出结果 $f_{\text{target}}(x)$ 来训练成员推理攻击模型，进而保证所训练的攻击模型可以区分给定样本是否来自目标模型的训练数据集。然而，在联邦学习的场景下，攻击者拥有的训练数据十分有限，导致目标模型的训练集 $D_{\text{target}}^{\text{train}}$ 在数量和多样性上十分匮乏。

为了解决这个问题，本节提出基于 GAN 的类级成员推理攻击模型。在数据增强阶段，攻击者通过在本地部署一个 GAN 模型以生成与训练数据相同分布的伪样本[31]，这些生成的样本将被用于训练攻击模型，最终以较高的推理精度成功发起成员推理攻击。从 GAN 模型的工作机理可以看出，辨别器 D 需要直接在真实样本上训练才能迫使生成器 G 生成同分布的伪样本[13]。而在联邦学习中，攻击者无法直接获取真实样本来训练一个分类良好的辨别器，针对这个问题，本节利用联邦学习中全局模型的迭代机制来直接更新辨别器。换言之，标准联邦学习协议中的参与方（包括攻击者）能够获取服务器发布的多个全局模型参数，这些模型参数由所有参与方本地训练参数聚合而来，而参与方的本地参数则是在真实样本上进行训练的。因此，用联

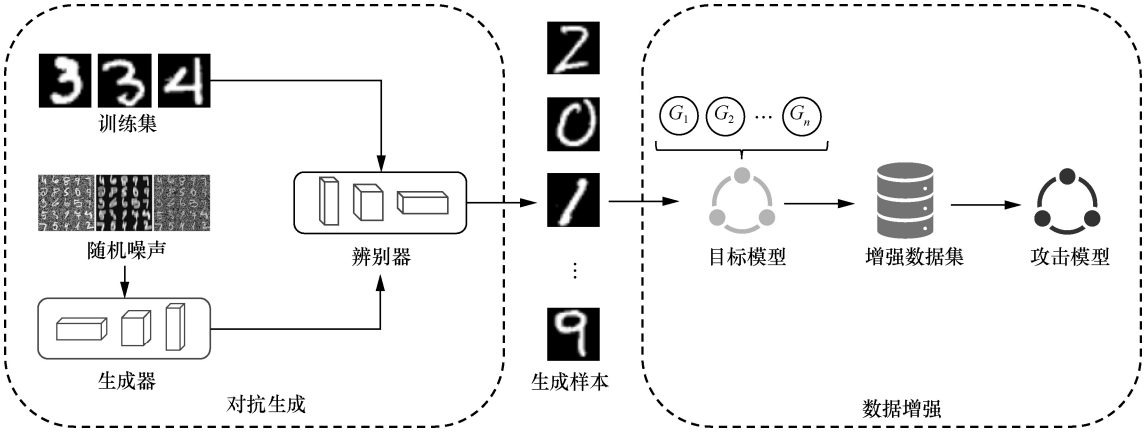

图 2 类级成员推理攻击模型

邦学习的全局模型参数来更新辨别器就相当于直接在真实样本上训练 D。这种巧妙的结合方式使生成器能够生成与真实样本相似的伪样本。

2.2.2　基于 GAN 的训练数据生成

本文利用 GAN 生成与原始数据相同分布的伪样本 x_{gen}，以克服类级攻击中训练数据量不足和多样性差的问题。如图 2 所示，生成器 G 的初始输入为随机噪声 $g(z;\theta_G)$，与此同时，辨别器 D 通过目标联邦学习模型 $f(x;\theta_D)$ 进行初始化和更新。随后，生成器 G 通过随机噪声生成与原始数据同分布的伪样本，而辨别器 D 则对生成的伪样本进行判断，确定生成的伪样本 x_{gen} 是否来自原始数据集，两者相互博弈，直到 D 无法区分 x_{gen} 与原始训练样本的差异。在此过程中，辨别器会引导生成器生成训练数据。通过一定轮次的迭代，生成数据 x_{gen} 的质量十分接近原始数据。值得注意的是，通过生成器 G 生成的伪样本在目标联邦学习模型上的分类精度与原始样本一致，主要原因在于本文采用目标联邦学习模型参数更新辨别器 D，因此生成的伪样本与原始训练样本分布相同。GAN 模型的生成过程如式(1)所示。

$$\min_{\theta_G}\max_{\theta_D}\left(\sum_{i=1}^{n_+}\log f(x_i;\theta_D)+\sum_{j=1}^{n_-}\log(1-f(g(x_{\text{gen}};\theta_G);\theta_D))\right) \tag{1}$$

其中，x_i 表示原始数据，x_{gen} 表示生成数据。然而，基于上述方式生成的 x_{gen} 是无标记的，无法直接用于成员推理攻击模型的训练，通常采用人工识别或运行目标模型 2 种方式标记。在联邦学习场景下，由于参与方可以轻易获取多轮次的目标模型（全局模型）参数，因此，本文采用运行目标全局模型的方式来对生成的伪样本 x_{gen} 进行标记。最终，原始数据和标记后的数据将被合并为一个数据集来训练成员推理攻击模型。

2.2.3　攻击模型训练

如图 2 所示，在数据增强阶段后，将原始数据和生成数据整合作为训练数据集。被整合的数据集包括可以被攻击模型学习的预测、真实标签和成员状态。最终，成员推理攻击模型的训练数据 $x\{i\}_{\text{attack}}$ 由预测、真实标签和表示是否由目标模型成员的 2 个属性“in”和“out”组成。若某一预测样本在目标模型上的输出与目标模型在原始训练样本上的输出结果十分相近，则二分类模型将该预测样本标记为“in”，否则标记为“out”。此外，根据 GAN 模型，可从 $D_{\text{target}}^{\text{train}}$ 中查询预测结果 Y，进而生成带有 (record,label) 的数据集 $D_{\text{target}}^{\text{gen}}$。最终，$D_{\text{target}}^{\text{ori}}$ 和 $D_{\text{target}}^{\text{gen}}$ 共同组成增强数据集 $D_{\text{target}}^{\text{train}}$。

由于成员推理攻击模型的目标是通过围绕真实标签的预测分布对成员状态进行分类，因此，可将训练数据集 $D_{\text{target}}^{\text{train}}$ 进一步划分为 n 个类别，其中 $f_{\text{attack}}()$ 代表攻击模型，模型的输入 x_{attack} 为 $(y,Y,\text{in/out})$。每个类别将被用于训练一个可以从给定的数据记录中对成员状态进行分类的攻击模型。类级成员推理攻击成功发起的主要原因是训练数据的多样性，其中，GAN 以其在数据增强方面的出色性能成为提高数据集多样性的有效方法。

2.3　用户级成员推理攻击

上述类级成员推理攻击方法能够在特定场景下转化为细粒度的用户级成员推理攻击，即恶意参与方能够通过推理攻击模型来判断某一给定样本的成员信息和所属用户身份信息。用户级成员推理攻击的发起依赖于“标签相互独立”的假设，下面进行详细说明。

1) 假设。在联邦学习模型训练之前，各个参与方需要向服务器声明本地训练数据的标签，由于标签信息并不会反映出数据特征，因此本地训练样本的隐私信息也不会暴露。上述标签声明现象在纵向联邦学习场景中经常发生，例如，医疗数据分析需要整合不同医院的诊断数据，服务方希望各医院提供的诊断数据属于不同分类，进而充分发挥联邦学习的协同训练作用，因此，参与方则率先声明自己的标签信息，进而用不同的标签数据来丰富训练集。在上述场景中，各个参与方的数据标签可以被认为是相互独立的，以 MNIST 数据集为例，某一参与方拥有标签“0”和“1”，而其他所有参与方都不会有“0”和“1”标签的数据。该假设的目的是便于将攻击模型的结果与之前声明的标签信息进行比较，从而发起用户级成员推理攻击。

2) 攻击构造。用户级成员推理攻击模型如图 3 所示。假设有 N 个参与方参与联邦学习，其中 V 表示受害者（良性参与方），A 表示作为参与方参与联邦学习的攻击者。在 k 轮训练迭代后，A 和 V 从中央服务器处得到了相同的全局模型参数 θ_d。通常情况下，A 和 V 会使用全局模型和本地数据训练一个新的本地模型，然后将更新的参数 θ_u 上传到服

务器。随后，联邦学习中央服务器对所有参与方更新的参数进行加权平均以更新全局模型。在加权平均机制的作用下，A 几乎不可能直接对某一特定 V 进行细粒度成员推理攻击。

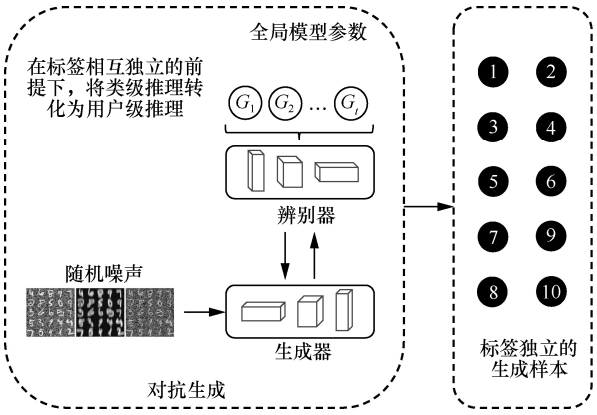

图 3 用户级成员推理攻击模型

针对这个问题，本节提出了一种基于标签声明的用户级成员推理攻击。具体来说，该攻击模型引入了与 2.2.2 节相同结构的 GAN 模型，使用 θ_d 迭代更新辨别器 D，迫使生成器生成与真实数据相似的伪样本。为了实现用户级成员推理，本文通过标签声明的方式进一步判断所生成的伪样本属于哪个参与方，进而达到与基于多任务 GAN 的用户级数据生成相同的效果[6]，同时保证攻击的发起方为恶意参与方，而非不可信服务器。由此，可用生成的数据来训练成员推理二分类器，一旦该分类器的输入为原始训练样本分布特征，则分类器输出预测结果为“in”，这表明成员推理结果与声明信息一致，否则标记为“out”。在获得成员信息以后，攻击者根据训练前声明的标签信息对推理样本进行身份信息判断，最终实现用户级成员推理。

3 防御方法

3.1 问题定义

本节设计了一种新型防御方法 DefMIA，以防御利用全局模型参数作为特征的成员推理攻击，其主要目标是在保证全局模型准确率的同时，使攻击者失去发起成员推理攻击的能力。具体来说，DefMIA 主要依赖于以下攻击事实，即联邦学习成员推理攻击的发起需要不断获取全局模型参数，以得到推理数据的置信度向量，进而创建一个二分类攻击模型判断该推理数据是否属于训练数据中的成员。因此，本文通过在下发全局模型的参数中加入特定噪声，将其转化成一个能够导致二分类攻击模型误分类的全局模型参数，进而成功防御成员推理攻击，如式(2)所示。

$$F' = F + N \tag{2}$$

其中，F 表示模型的真实参数，N 表示噪声，此时，攻击者获得带有噪声 N 的模型参数 F'。然而，噪声 N 不仅会影响攻击模型，也会降低全局模型的性能。此外，防御者还存在以下问题：①防御者无法获取攻击者的二分类攻击模型，即很难观察到攻击分类器的准确性；②防御者不能观察到每个参与方的本地训练过程，无法有针对性地控制噪声大小。

为了解决上述问题，本文方案在服务器端训练一个防御分类器，它能够模仿攻击过程来发动成员推理攻击。防御分类器的决策函数 g 表示相应数据样本是训练集成员的概率。联邦学习的训练模式表明参与方能够获得多个通信轮次的全局模型，这意味着攻击者将拥有多个版本的全局模型，进而可整合所有的模型来提高成员推理攻击的性能。为此，本文通过引入一个遵循约束条件的噪声，在中央服务器分发全局模型之前将其添加到模型参数中。此外，攻击者会选择全局模型的多个参数作为攻击特征，而防御者很难将噪声添加到所有参数中，例如，梯度会影响全局模型的性能，但若在所有梯度上添加噪声，计算成本会很高。因此，要实现为每一个攻击特征添加噪声，其优化问题可以概括为

$$\begin{aligned} &\min E(g(\boldsymbol{s}+\boldsymbol{n};W)-0.5) \\ &\min E(L(f(x;W))) \end{aligned} \tag{3}$$

其中，$\boldsymbol{s}$ 表示模型输出置信度向量，$\boldsymbol{n}$ 表示添加的噪声向量，$\min E(g(\boldsymbol{s}+\boldsymbol{n};W)-0.5)$ 表示攻击模型的性能被限制在 0.5 左右，即攻击模型无法区分哪个样本属于目标模型训练数据；$\min E(L(f(x;W)))$ 表示最小化全局模型的损失以获取最佳分类精度。

3.2 DefMIA 方法架构

本节详细介绍本文所提 DefMIA 方法，包括噪声配置和噪声添加 2 个阶段。在 DefMIA 方法中，本文使用对抗样本来误导成员推理攻击中的二分类攻击模型，在考虑模型效用损失的同时应用对抗性方法产生噪声。基于前文分析的联邦学习成员推理攻击威胁模型，攻击者能够伪装成联邦学习的良性参与方并且参与目标模型的训练过程。由于攻击者能够获取每个通信轮次中的全局模型参数，目标模型对攻击者相当于白盒场景，攻击者可以使用联

邦学习全局模型的梯度参数和置信度向量作为攻击特征来训练攻击模型。作为一个被动的参与方，攻击者在每个通信轮次中仅能获得一个全局模型，并且可以通过组合这些模型参数来提高成员推理攻击的准确性。

基于以上判断，本节提出一种基于对抗样本的成员推理防御方法。通过设计一个包含两阶段优化的噪声生成算法来误导攻击者的二分类攻击模型，同时保证全局模型的准确率。具体来说，DefMIA 方法在第一阶段优化中找到添加到攻击特征中的噪声，进而通过第二阶段优化将攻击特征变成一个对抗样本。如图 4 所示，中央服务器在将全局模型分发给参与方之前，将噪声添加到置信度向量中，使带噪声的置信度向量对于成员推理攻击模型来说是一个对抗样本，而在全局模型的分类任务上表现正常。换言之，DefMIA 方法的可行性主要依赖于选取一个合适的噪声，使其与置信度向量相结合后，能够在成员推理攻击模型错误分类的同时不改变联邦学习模型的分类边界。此外，联邦学习中的攻击者可以结合全局模型的不同版本来获取攻击特征以提高成员推理攻击的准确率，因此防御者需在训练时保存训练过程中的历史模型。

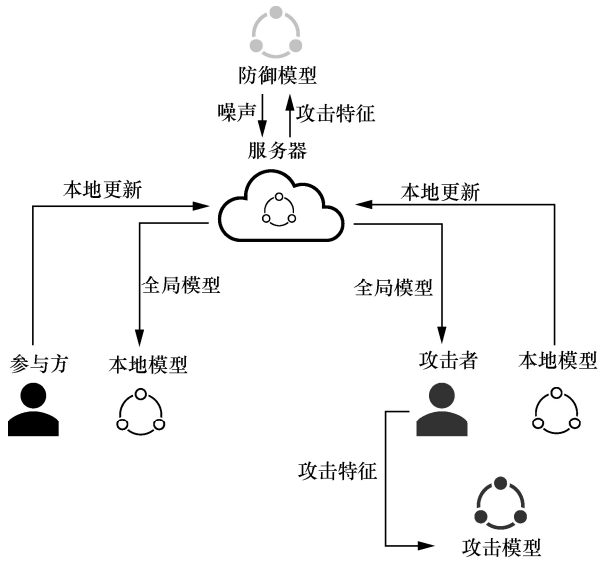

图 4　DefMIA 方法架构

3.3　噪声配置

在 DefMIA 方法中，一个适合的噪声必须满足 2 个约束条件：①联邦学习全局模型损失最小化；②成员推理攻击准确率接近 0.5。例如，在第 t 次迭代中，中央服务器将所有参与方上传的梯度聚合，以更新 $t+1$ 轮的全局模型参数。在更新全局模型下发之前，服务器首先利用更新后的全局模型参数，生成满足以上 2 个约束条件的噪声，并将这个精心选择的噪声添加到全局模型参数中，进而分发给所有参与方，以进行 $t+1$ 轮的本地训练过程。因此，DefMIA 方法首先生成一个满足防御模型输出约束的随机噪声向量，防御模型的约束条件为

$$g(\mathrm{softmax}(\boldsymbol{l}+\boldsymbol{e}))=\frac{1}{1+\exp(-\boldsymbol{h}(\mathrm{softmax}(\boldsymbol{l}+\boldsymbol{e})))} \tag{4}$$

其中，$\boldsymbol{l}$ 和 $\boldsymbol{h}$ 分别表示目标模型和防御模型的概率输出向量，$\boldsymbol{e}$ 表示未经归一化的噪声向量。当 $\boldsymbol{h}=\boldsymbol{0}$ 时，式(4)恒等于 0.5，因此，成员推理攻击模型的损失函数可定义为

$$L_1=|\boldsymbol{h}(\mathrm{softmax}(\boldsymbol{l}+\boldsymbol{e}))| \tag{5}$$

其次，对于目标联邦学习模型的损失函数来说，本文要使对抗样本与原置信向量间的距离尽可能小，以达到不改变联邦学习模型分类边界的目的。此外，为保证目标模型的效用，模型的输出预测结果（即预测标签）不能改变。因此，约束条件为

$$\min d(\boldsymbol{s}+\boldsymbol{n},\boldsymbol{s}) \tag{6}$$

$$\arg\max_j\{l_j+e_j\}=\arg\max_j\{l_j\} \tag{7}$$

其中，式(6)表示最终的输出结果与原始的输出结果的距离应该尽可能小，$\boldsymbol{s}+\boldsymbol{n}$ 为 $\boldsymbol{l}+\boldsymbol{e}$ 经过激活函数后的结果；式(7)表示模型的输出结果保持不变以保证模型的可用性。本文假设 y 为目标模型对于数据预测的标签结果，要保证模型的预测结果不变，即有 $l_y+e_y\geqslant\max_{j|j\neq y}\{l_j+e_j\}$，因此式(7)的约束条件可以进一步表示为

$$L_2=\max\{0,\max_{j|j\neq y}\{l_j+e_j\}-l_y-e_y\} \tag{8}$$

在形式化本文的目标优化问题后，优化问题可以转化为

$$\min_e L_{\mathrm{def}}=L_1+c_1L_2+c_2d \tag{9}$$

本文的目标是使目标模型置信度向量失真尽可能小的同时使攻击者攻击模型不准确，对于每个给定的 c_2，使用梯度下降法来寻找满足条件的噪声向量，为了保证噪声能够满足 2 个约束条件，在每一轮的梯度下降后都要检查所得噪声是否满足条件。

3.4　添加噪声

在噪声配置阶段后，获得了满足条件的噪声向

量 $\boldsymbol{n}$；在添加噪声阶段，考虑如何将噪声添加到攻击模型的特征中。当攻击者训练攻击模型时，他们只能使用带噪声的攻击特征来训练成员推理攻击模型。噪声添加问题可以表示为

$$p=\begin{cases}0, |g(\boldsymbol{s})-0.5| \leqslant |g(\boldsymbol{s}+r)-0.5| \\ \min\left(\dfrac{\varepsilon}{d(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{n})},1.0\right), \text{其他}\end{cases} \tag{10}$$

其中，ε 是置信度向量的失真预算，用于控制添加噪声的大小。本文使用 L1 范数来衡量目标模型置信向量的失真度。在攻击者获取置信度向量之前，防御者以一定的概率向其中添加精心选择的噪声来干扰成员推理攻击的准确率。由于联邦学习多轮迭代的特性，攻击者可以获取训练过程中不同版本的全局模型，因此在每轮模型下发之前，防御者需要更新噪声。DefMIA 方法的伪代码如算法 1 所示。

算法 1 DefMIA 方法

输入 $w_t, \boldsymbol{s}, \max$

输出 r

/*服务器执行*/

1)初始化 w_0

2) for iteration $t \in (1,2,\cdots,T)$ do

3) for client $k \in (1,2,\cdots,K)$ do

4) $W_{t+1}^k \leftarrow \text{ClientUpdate}(k, W_t)$

5) $W_{t+1} \leftarrow \dfrac{1}{N}\sum_{k=1}^{n} W_{t+1}^k$

6) end for

7) end for

更新 W_{t+1} 并发送至防御模型

随机生成噪声 $\boldsymbol{n}'$

8) for $i=0$ to max do

9) if $h(\text{softmax}(\boldsymbol{l}))h(\text{softmax}(\boldsymbol{l}+\boldsymbol{e}))>0$

10) $\min d(\boldsymbol{s}+\boldsymbol{n}, \boldsymbol{s})$

11) $\min E(g(\boldsymbol{s}+\boldsymbol{n}; W)-0.5)$

12) $\min L(f(x; W))$

13) end if

14) end for

15) 返回噪声 $\boldsymbol{n}$ 至中央服务器

更新全局模型

/*参与方执行*/

1) for each local epoch e do

2) for batch b do

3) $W_{t+1} \leftarrow W_t - \eta\nabla(W; b)$

4) end for

5) end for

更新本地参数 (k, w)

4 实验评估

本节首先介绍了数据集和实验设置，然后评估了类级成员推理攻击、用户级成员推理攻击和防御方法的性能。

4.1 数据集和实验设置

1) 数据集

本文采用图像识别领域中的 3 个基准数据集 MNIST、Fashion MNIST（F-MNIST）和 CIFAR-10 来进行实验评估。

MNIST 是一个手写体数字的图像数据集，由 60 000 个训练数据和 10 000 个测试数据组成，包括从数字“0”到“9”的 10 个类别，其中每个样本都是像素大小为 28×28 的灰度图像[33]。

F-MNIST 是一个服装分类的图像数据集，由 60 000 个训练数据和 10 000 个测试数据组成，包括 T 恤、裙子、鞋子等 10 个类别，其中每个样本也是 28 像素×28 像素的灰度图像[34]。

CIFAR-10 是一个包含 10 类、共计 60 000 张彩色的图像数据集（如飞机、汽车、鸟等），其中包括 50 000 张训练图像和 10 000 张测试图像。CIFAR-10 数据集中的图像在输入处理时被标准化为 32 像素×32 像素的三通道输入。

2) 实验设置

本节所有实验均在具有 32 GB RAM 和 NVIDIA Quadro P4000 GPU 的服务器上运行，其软件环境为 Ubuntu 16.04 Linux，Python 3.6 下的 Pytorch 和 Tensorflow+Keras 框架。

类级成员推理攻击配置。在 3 个数据集上应用一个基本的 CNN 模型作为成员推理模型。为联邦学习中设置了 100 个参与方，其中每个参与方拥有 600（CIFAIR-10 数据集上为 500）个数据样本，此外，每个参与方以 0.001 的学习率训练 10 个 epoch。

用户级成员推理攻击配置。MNIST 数据集的神经网络模型包含 2 个卷积层和 2 个深度层，卷积核的大小设置为 5×5，每个参与方以 0.01 的学习率训练 30 个 epoch；F-MNIST 数据集的神经网络模型拥有 4 个卷积层和 2 个深度层，卷积核大小为 3×3，每个参与方以 0.000 1 的学习率训练 60 个 epoch。

CIFAR-10 数据集采用 Resnet18[35]网络，所有实验均运行了 400 个联邦学习的通信轮次。

防御配置。目标联邦学习模型的参与方数量被设定为 5 个，其中一个参与方被认为是攻击者。每个参与方拥有 12 000 个训练样本，以 0.01 的学习率训练 10 个 epoch。此外，为了进一步体现 DefMIA 方法的有效性，将 DefMIA 方法部署在基于 CNN 的 MIA（CNN- based MIA）[9]、白盒 MIA（White-box MIA）[12]和 GAN 增强的 MIA（GAN-enhanced MIA）[30]这 3 种攻击场景下进行防御验证。

4.2　数据增强的有效性

为了评估在联邦学习中使用 GAN 进行数据增强的有效性，本文在参与方和样本数量保持不变的前提下对 GAN 重建数据过程进行可视化。不失一般性，实验采用文献[6]和文献[31]中相同结构的 GAN 模型进行数据增强评估分析，其中，生成器生成长度为 100 的样本，并将其重塑为 28 像素×28 像素，当全局模型的精度达到 90%时，生成器开始生成样本。在 GAN 模型的配置中，本文将由本地模型平均聚合得到的全局模型参数用以更新辨别器，随着联邦学习全局模型精度的不断提升，生成器将生成与原始样本十分相似的伪样本。图 5 展示了不同数据集在不同通信轮次下的数据重建结果。

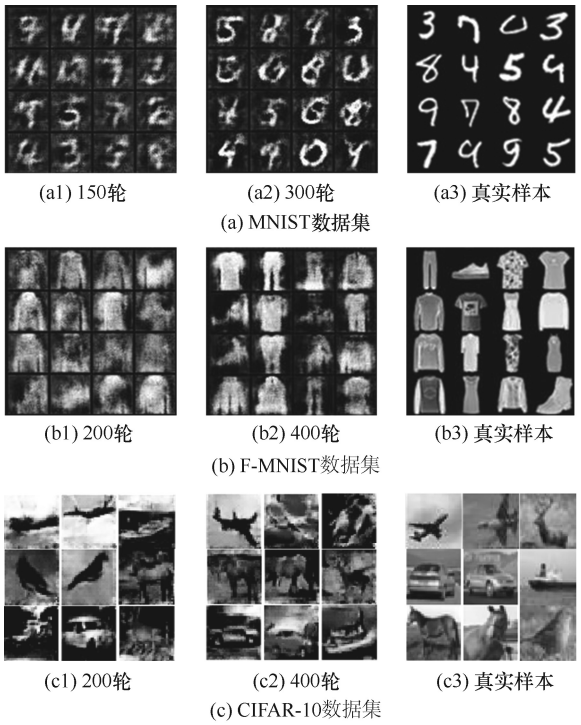

图 5　不同数据集在不同通信轮次下的数据重建结果

由图 5 可知，随着迭代的进行，生成器的效果不断优化，生成的图像也更加清晰。具体来说，在 MNIST 数据集中，当联邦学习运行到 150 轮时，已经能够基本识别手写数字以及所属类别；当联邦学习运行到 300 轮时，生成图像的轮廓更加清晰。此外，在 F-MNIST 和 CIFAR-10 数据集上，由于实验采用的神经网络结构有所差异，当联邦学习运行到 200 轮时，才能够生成具有一定轮廓的图像；当联邦学习运行到 400 轮时，生成图像及其类别越加明显，在视觉上与原始样本十分接近。

4.3　类级攻击评估

本文采用模型的攻击精度和召回率 2 个指标来衡量类级成员推理攻击的性能表现，其中，攻击精度代表了攻击的准确率，而攻击方法的覆盖率则由召回率来衡量。实验通过 GAN 模型生成 5 000 个伪样本，并将这些生成的伪样本作为攻击模型的训练集。表 1 展示了类级成员推理攻击模型在 3 个基准数据集上的性能。

表 1　类级成员推理攻击模型在 3 个基准数据集上的性能

数据集	攻击精度	召回率	F1 分数
MNIST	0.976	0.884	0.94
F-MNIST	0.955	0.872	0.92
CIFAR-10	0.901	0.854	0.87

由表 1 可知，本文提出的类级成员推理攻击方法在 MNIST、F-MNIST 和 CIFAR-10 数据集上能够取得超过 0.9 的攻击精度，召回率分别为 0.884、0.872 和 0.854。此外，实验采用 F1 分数衡量联邦学习全局模型的分类能力，结果显示，在 3 个基准数据集上，F1 分数分别达到 0.94、0.92 和 0.87，这意味着所提方法具有良好的泛化和成员推理能力，与前 2 个数据集相比，CIFAR-10 上的 F1 分数有所下降，但仍然能够达到期望的效果，即攻击精度大于 0.9。

此外，本文验证了 MNIST 数据集下不同规模的生成数据对成员推理攻击方法在每一类上的预测影响，结果如图 6(a)所示。当增强数据为 100 个样本时，成员推理攻击准确率并未有明显提升，仍接近于随机猜测；当增强数据为 1 000 个样本时，推理攻击的平均准确率得到大幅提升（约 0.817）；当增强数据达到 5 000 个样本后，准确率显著上升到 0.972。进一步地，为了探索本地训练参数对成员推理攻击的准确率影响，在不同的本地 epoch 下进行了实验，结果如图 6(b)所示。100 个 epoch 的

成员推理攻击精度远高于 10 个 epoch，分别为 0.964 和 0.813，这意味着本地模型过拟合将导致全局模型更容易受到成员推理攻击的影响。

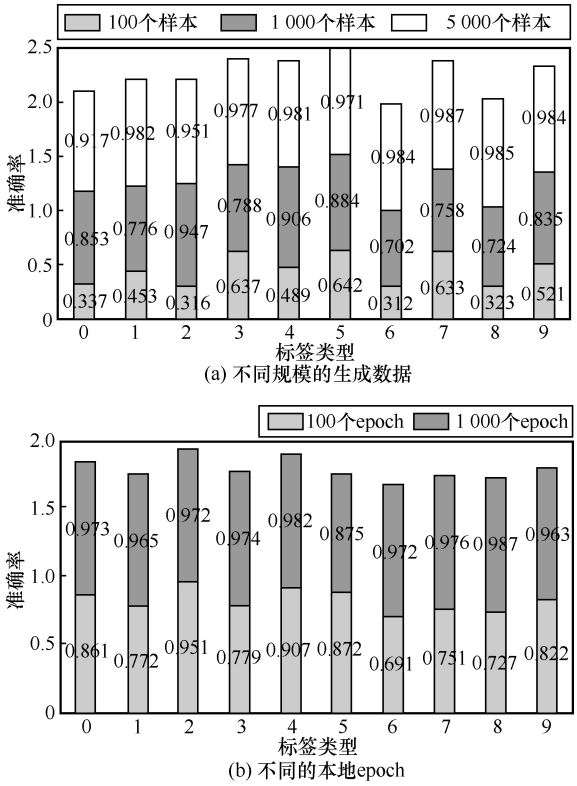

图 6　不同参数下类级推理攻击的评估结果

为了进一步评估本文所提类级成员推理攻击方法的有效性，本节将其与现有典型成员推理攻击方法（文献[9]和文献[12]）进行了实验对比，其中，文献[9]针对 MLaaS（机器学习即服务）通过黑盒 API 构造成员推理攻击方法，文献[12]在联邦学习场景下实现了白盒成员推理攻击。表 2 从主任务准确率和类级成员推理攻击精度 2 个方面展示了对比实验结果。由表 2 可看出，本文所提方法明显优于典型成员推理攻击方法。此外，在攻击成功发起的同时，联邦学习主任务准确率并未受到影响。特别地，由于类级成员推理攻击方法采用数据集重建方法以增强攻击数据集，其攻击精度相较于文献[12]整体上提升了 13%左右，进一步说明本文所提方法的有效性。

表 2　　类级推理攻击对比实验结果

数据集	文献[9]攻击精度	文献[12]攻击精度	类级 MIA 精度	主任务准确率
MNIST	0.904	0.847	0.976	0.963
FMNIST	0.873	0.826	0.955	0.925
CIFAR-10	0.866	0.822	0.901	0.897

4.4　用户级攻击评估

用户级成员推理攻击通过本地部署的 GAN 生成足够的样本来训练攻击模型，在成员推理攻击之后，进一步从标签的角度衡量攻击效果。首先，验证了联邦学习模型的分类精度，如图 7(a)所示，联邦学习模型在 MNIST、F-MNIST 和 CIFAR-10 数据集上的准确率分别达到 0.981、0.937 和 0.857，表明联邦学习能够完成测试数据集上的所有分类任务，进而间接保证了成员推理攻击的有效性。此外，图 7(b)显示了用户级成员推理攻击在上述 3 个数据集上的攻击效果，其中，TP 代表真阳性，FN 代表假阴性。实验比较了受害者所在持有不同样本数量下的成员推理攻击的有效性，假设受害者持有多个标签的数据，分别为 1、2、3、4、5。从图 7(b)中可以看出，受害者持有的数据类别越多，不同推理样本数量下的 FN 也越高，这意味着成员推理攻击的有效性越差。

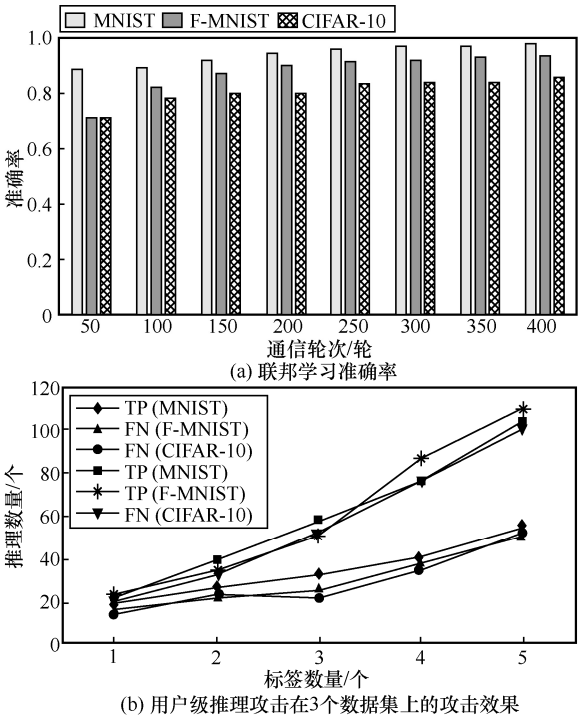

图 7　联邦学习准确率和用户级推理攻击的评估结果

4.5　防御方法评估

为了评估本文所提 DefMIA 方法在不同基准数据集上的性能表现，本节验证了 DefMIA 在 MNIST、F-MNIST 和 CIFAR-10 数据集上不同类别之间的防御效果。表 3 给出了不同数据集下 DefMIA 防御方法的实验结果。从表 3 可以看出，成员推理攻击在 3 个数据集上的攻击准确率（Before）能够达到 0.874

以上。其中，MNIST 数据集中类别“3”和类别“4”上的攻击准确率最高，为 0.977；CIFAR-10 数据集中类别“3”上的攻击准确率最低，为 0.874，这说明成员推理攻击在获得较高攻击准确率的同时，也会受到数据样本特征数量的影响。部署 DefMIA 方法后，成员推理攻击在 MNIST、F-MNIST 和 CIFAR-10 数据集上的平均准确率降低为 0.542、0.515 和 0.498，结果十分接近于随机猜测（0.5），这充分证明了 DefMIA 方法的有效性。

表 3　不同数据集下 DefMIA 防御方法的实验结果

类别	MNIST		F-MNIST		CIFAR-10	
	Before	DefMIA	Before	DefMIA	Before	DefMIA
0	0.903	0.501	0.942	0.512	0.876	0.483
1	0.961	0.506	0.922	0.502	0.889	0.491
2	0.953	0.551	0.925	0.524	0.879	0.507
3	0.977	0.575	0.921	0.504	0.874	0.498
4	0.977	0.597	0.953	0.533	0.901	0.503
5	0.896	0.546	0.932	0.517	0.877	0.488
6	0.923	0.549	0.897	0.503	0.894	0.502
7	0.924	0.535	0.933	0.513	0.886	0.514
8	0.923	0.529	0.967	0.539	0.882	0.489
9	0.921	0.535	0.929	0.507	0.906	0.504

此外，实验进一步评估了 DefMIA 在联邦学习中不同成员推理攻击方法下的性能表现，分别为 CNN-based MIA[9]、White-box MIA[12] 和 GAN-enhanced MIA[30]。其中，CNN-based MIA 为机器学习中典型的黑盒成员推理攻击方法，其利用影子模型和置信度向量来判断某一给定样本是否属于模型训练数据集；White-box MIA 和 GAN-enhanced MIA 为针对联邦学习场景下的白盒成员推理攻击模型，主要使用模型辅助信息（如参数、结构、中间结果等）以实现增强推理攻击准确率的目标。图 8 给出了不同攻击下 DefMIA 方法的有效性评估结果。从图 8 可以看出，上述 3 种成员推理攻击方法对在 3 个基准数据集上的攻击准确率均超过 0.725。其中，White-box MIA 和 GAN-enhanced MIA 的攻击准确率分别为 0.836 和 0.901，主要原因在于这 2 种攻击方法均采用不同辅助信息以实现增强攻击数据集的目标，进而获得了较高的攻击准确率。但在 DefMIA 方法下，上述 3 种攻击方法的准确率为 0.508～0.602，充分证明了本文所提 DefMIA 方法能够有效降低攻击准确率。此外，在不同数据集上的结果表明，所提 DefMIA 方法并不受数据分布影响。综上所述，DefMIA 方法可以成功防御联邦学习中的成员推理攻击，同时将全局模型的性能保持在一个较高的水平。

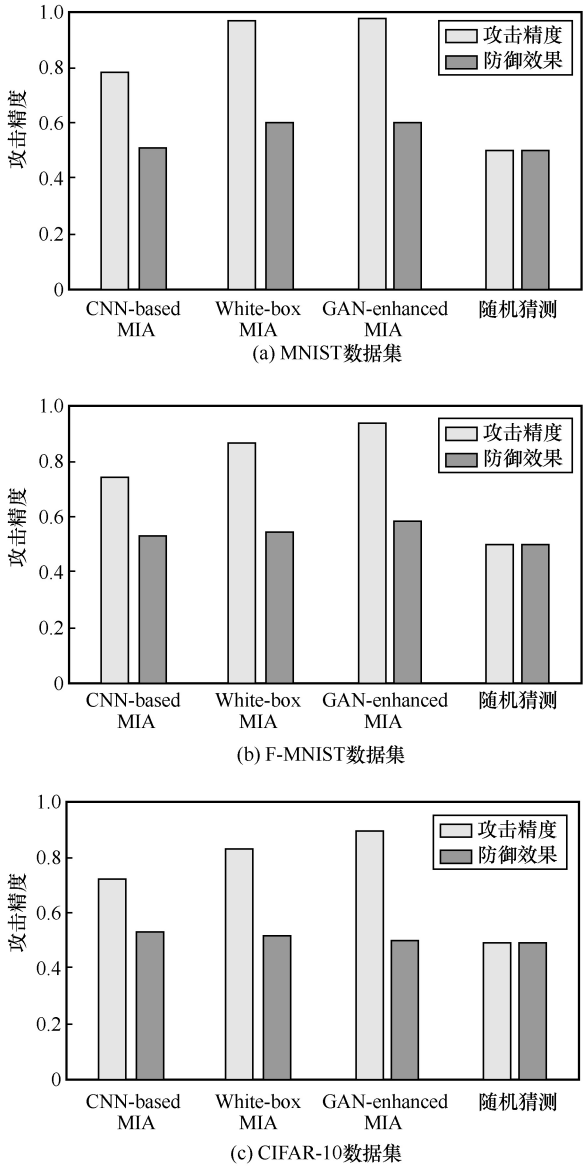

图 8　不同攻击下 DefMIA 方法的有效性评估结果

最后，为了进一步说明本文所提 DefMIA 方法的优势，本节将 DefMIA 与已发表的其他 2 种成员推理攻击防御方法进行了对比。对比方法分别选择了黑盒和白盒模型下的最新研究成果，对应参考文献[18,36]。文献[18]在黑盒攻击模型下采用对抗样本进行成员推理攻击防御，所提方法能够在黑盒设定下实现效用损失保证，不需要重新训练模型。文献[36]提出了 Nirvana 算法，可在白盒模型下最小化神经网络隐藏层之间的泛化误差，进而实现成员推理攻击的有效防御。表 4 给出了在 3 种典型成员推

理攻击下的防御方法对比实验结果。从表 4 中可以看出，本文所提 DefMIA 方法取得了更好的性能表现，平均攻击精度为 0.566。其中，在 White-box MIA 防御场景下，由于 Nirvana 算法对神经网络具有更强的泛化能力，其防御效果最佳，达到 0.556。而在 GAN-enhanced MIA 中，DefMIA 相较于其他 2 种防御方法具有更强的防御能力，主要原因在于 DefMIA 中的噪声生成算法可以针对不同轮次的联邦学习模型输出相应的噪声大小，导致成员推理攻击模型不断输出错误分类结果，进而保证 DefMIA 方法的有效性。

表 4　在 3 种典型成员推理攻击下的防御方法对比实验结果

攻击	文献[18]	文献[36]	DefMIA
CNN-based MIA	0.657	0.527	0.517
White-box MIA	0.698	0.556	0.586
GAN-enhanced MIA	0.753	0.684	0.594

5　结束语

本文对联邦学习中基于 GAN 的成员推理攻击及防御方法进行了全面探索。首先提出了基于 GAN 的两类成员推理攻击：类级成员推理攻击和用户级成员推理攻击。其中，类级成员推理攻击中的 GAN 模型被用来增加和填补攻击数据的多样性，从而提高二分类攻击模型的准确性；用户级成员推理攻击的目的是推断出特定参与方的成员信息。此外，针对上述攻击类型，进一步提出了基于对抗样本的防御方法 DefMIA。在每个通信轮次中，中央服务器以一定概率向全局模型的置信度分数向量中添加噪声，从而误导全局模型的成员推理攻击。在未来的工作中，将探索在一个不受信任的联邦学习环境中的成员推理攻击，在这种情况下，如何保证推断某些数据记录的成员信息的攻击准确率是一个新的挑战。

参考文献：

[1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint, arXiv: 1602.05629, 2016.

[2] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[3] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[4] SATTLER F, WIEDEMANN S, MÜLLER K R, et al. Robust and communication-efficient federated learning from non-i.i.d. data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(9): 3400-3413.

[5] TRUEX S, LIU L, GURSOY M E, et al. Demystifying membership inference attacks in machine learning as a service[J]. IEEE Transactions on Services Computing, 2021, 14(6): 2073-2089.

[6] WANG Z B, SONG M K, ZHANG Z F, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]//Proceedings of IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2019: 2512-2520.

[7] MELIS L, SONG C Z, CRISTOFARO E D, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 691-706.

[8] ZHU L, LIU Z, HAN S. Deep leakage from gradients[J]. arXiv Preprint, arXiv:1906.08935, 2019.

[9] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 3-18.

[10] CHEN J L, ZHANG J L, ZHAO Y C, et al. Beyond model-level membership privacy leakage: an adversarial approach in federated learning[C]//Proceedings of 2020 29th International Conference on Computer Communications and Networks (ICCCN). Piscataway: IEEE Press, 2020: 1-9.

[11] HAYES J, MELIS L, DANEZIS G, et al. LOGAN: membership inference attacks against generative models[C]//Proceedings of Privacy Enhancing Technologies Symposium. Berlin: Springer, 2019: 133-152.

[12] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2019: 739-753.

[13] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[14] QU Y Y, YU S, ZHANG J W, et al. GAN-DP: generative adversarial net driven differentially privacy-preserving big data publishing[C]//Proceedings of 2019 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2019: 1-6.

[15] JONSSON K V, KREITZ G, UDDIN M. Secure multi-party sorting and applications[J]. IACR Cryptology ePrint Archive, 2011: doi. eprint.iacr.org/2011/122.

[16] AONO Y, HAYASHI T, WANG L, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345.

[17] 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. New York: ACM Press, 2016: 308-318.

[18] JIA J Y, SALEM A, BACKES M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 259-274.

[19] ZHOU Y H, YE Q, LV J C. Communication-efficient federated learning with compensated overlap-FedAvg[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(1): 192-205.

[20] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2015: 1322-1333.

[21] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems[C]//European Symposium on Research in Computer Security. Berlin: Springer, 2020: 480-501.

[22] ZHANG J L, CHEN J J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets[C]//Proceedings of 2019 18th IEEE International Conference on Trust, Security and Privacy In Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). Piscataway: IEEE Press, 2019: 374-380.

[23] MOTHUKURI V, PARIZI R M, POURIYEH S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619-640.

[24] PROUDFOOT D. Anthropomorphism and AI: Turing's much misunderstood imitation game[J]. Artificial Intelligence, 2011, 175(5-6): 950-957.

[25] ZHANG J L, CHEN B, CHENG X, et al. PoisonGAN: generative poisoning attacks against federated learning in edge computing systems[J]. IEEE Internet of Things Journal, 2021, 8(5): 3310-3322.

[26] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]//International Conference on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938-2948.

[27] XU G W, LI H W, LIU S, et al. VerifyNet: secure and verifiable federated learning[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 911-926.

[28] LU Y L, HUANG X H, DAI Y Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2020, 16(6): 4177-4186.

[29] SALEM A, ZHANG Y, HUMBERT M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//Proceedings of 2019 Network and Distributed System Security Symposium. Reston: Internet Society, 2019: 1-15.

[30] ZHANG J W, ZHANG J L, CHEN J J, et al. GAN enhanced membership inference: a passive local attack in federated learning[C]//Proceedings of 2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-6.

[31] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 603-618.

[32] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 427-436.

[33] DENG L. The MNIST database of handwritten digit images for machine learning research[best of the Web][J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.

[34] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv Preprint, arXiv: 1708.07747, 2017.

[35] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[36] WU D, QI S Y, QI Y, et al. Understanding and defending against White-box membership inference attack in deep learning[J]. Knowledge-Based Systems, 2023, 259: 110014.

[作者简介]

张佳乐（1994− ），男，安徽蚌埠人，博士，扬州大学讲师、硕士生导师，主要研究方向为人工智能安全、联邦学习、数据隐私保护等。

朱诚诚（2000− ），男，安徽临泉人，扬州大学硕士生，主要研究方向为联邦学习安全与隐私保护。

孙小兵（1985− ），男，江苏姜堰人，博士，扬州大学教授、博士生导师，主要研究方向为软件安全、人工智能安全、区块链安全等。

陈兵（1970− ），男，江苏南通人，博士，南京航空航天大学教授、博士生导师，主要研究方向为无线网络、人工智能安全、网络空间安全、智能无人系统等。