计算机科学
COMPUTER SCIENCE

面向机器学习的成员推理攻击综述

陈得鹏，刘肖，崔杰，何道敬

引用本文

陈得鹏，刘肖，崔杰，何道敬.面向机器学习的成员推理攻击综述[J]．计算机科学, 2023, 50(1): 302-317.

CHEN Depeng, LIU Xiao, CUI Jie, HE Daojing. Survey of Membership Inference Attacks for Machine Learning [J]. Computer Science, 2023, 50(1): 302-317.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800227

面向机器学习的成员推理攻击综述

陈得鹏[1]　刘　肖[1]　崔　杰[1]　何道敬[2]
1 安徽大学计算机科学与技术学院　合肥 230601
2 哈尔滨工业大学(深圳)计算机科学与技术学院　广东 深圳 518055
(depengchen@ahu.edu.cn)

摘　要　随着机器学习的不断发展,特别是在深度学习领域,人工智能已经融入到人们日常生活的方方面面。机器学习模型被部署到多种场景的应用中,提升了传统应用的智能化水平。然而,近年来的研究指出,用于训练机器学习模型的个人数据时常面临隐私泄露的风险。其中,成员推理攻击就是针对机器学习模型威胁用户隐私安全的一种非常重要的攻击方式。成员推理攻击的目的是判断用户数据样本是否被用于训练目标模型(如在医疗、金融等领域的用户数据),从而直接干涉到用户隐私信息。首先介绍了成员推理攻击的相关背景知识,随后对现有的成员推理攻击按照攻击者是否拥有影子模型进行分类,并对成员推理攻击在不同领域的威胁进行了相应的总结。其次,介绍了应对成员推理攻击的防御手段,对现有的防御机制按照模型过拟合、基于模型压缩和基于扰动等策略进行分类和总结。最后,对现有的成员推理攻击和防御机制的优缺点进行了分析,并提出了成员推理攻击的一些潜在的研究方向。

关键词:机器学习;隐私保护;成员推理攻击;防御机制
中图法分类号　TP391

Survey of Membership Inference Attacks for Machine Learning

CHEN Depeng[1],LIU Xiao[1],CUI Jie[1] and HE Daojing[2]
1 School of Computer Science and Technology,Anhui University,Hefei 230601,China
2 School of Computer Science and Technology,Harbin Institute of Technology(Shenzhen),Shenzhen,Guangdong 518055,China

Abstract　Artificial intelligence has been integrated into all aspects of people's daily lives with the continuous development of machine learning,especially in the deep learning area. Machine learning models are deployed in various applications,enhancing the intelligence of traditional applications. However,in recent years,research has pointed out that personal data used to train machine learning models is vulnerable to the risk of privacy disclosure. Membership inference attacks(MIAs) are significant attacks against the machine learning model that threatens users' privacy. MIA aims to judge whether user data samples are used to train the target model. When the data is closely related to the individual,such as in medical,financial,and other fields,it directly interferes with the user's private information. This paper first introduces the background knowledge of membership inference attacks. Then,we classify the existing MIAs according to whether the attacker has a shadow model. We also summarize the threats of MIAs in different fields. Also,this paper points out the defense means against MIAs. The existing defense mechanisms are classified and summarized according to the strategies for preventing model overfitting,model-based compression,and disturbance. Finally,this paper analyzes the advantages and disadvantages of the current MIAs and defense mechanisms and proposes possible research directions for future MIAs.

Keywords　Machine learning,Privacy-preserving,Membership inference attack,Defense mechanism

1　引言

近年来,人工智能的发展如火如荼,机器学习是人工智能的核心,它在许多领域都有着广泛的应用,如图像识别、自然语言处理、推荐系统等。随着机器学习尤其是深度学习等技术的不断发展,人工智能技术达到了新的高度,这主要得益于两个方面:一方面是大量数据的获取更加方便;另一方面是计算能力也在不断增强。

深度学习技术的快速发展,人工智能技术的日新月异,使得人们可以更加充分和高效地利用海量的数据,如刷脸支付、医疗辅助诊断等新兴技术为人们的日常生活带来了极大的便利。然而,人们在享受人工智能带来的智慧生活的同时,个人

到稿日期:2022-08-24　返修日期:2022-10-06
基金项目:国家自然科学基金(U1936220,61872001,62011530046)
This work was supported by the National Natural Science Foundation of China(U1936220,61872001,62011530046).
通信作者:崔杰(cuijie@ahu.edu.cn).

隐私风险也不容忽视。理论上，训练机器学习模型的数据集相关信息不应该被泄露，因为训练机器学习模型需要大量的用户数据，这些用户数据往往涉及个人隐私，如财务信息[1]、医疗数据信息[2]、位置信息[3]等个人敏感数据。然而目前的研究表明，机器学习会泄露训练数据相关信息，一旦攻击者获得用户的这些信息，便可通过成员推理攻击，推断出该样本数据是否在训练目标模型中被使用过，从而泄露用户的隐私。

近年来，机器学习即服务（Machine Learning as a Service，MLaaS）得到了广泛的应用，一些互联网巨头如谷歌[1)]、亚马逊[2)]、微软[3)]等公司都纷纷推出了这项服务，其主要针对小型互联网公司及个人。它为服务提供商提供了一种简单便捷的方式来部署机器学习模型，同样也为用户提供了一种即时的方式，以便其在各种应用程序中使用机器学习的相关模型。具体而言，用户把数据样本作为输入，发送给服务器，服务器提供预先训练好的模型或者用户自己构建的模型，通过模型计算后，再把模型预测数据返回给用户。用户在不拥有机器学习模型的情况下，只需要付费访问模型 API 也能使用到机器学习的服务。然而越来越多的研究发现，MLaaS 平台加剧了隐私泄露的风险。这项服务使得攻击者即使不拥有机器学习模型，也能对隐私数据进行推理攻击，导致数据信息的泄露。因此，对机器学习隐私方面的保护应该引起人们的高度重视。

成员推理攻击（Membership Inference Attack，MIA）是导致上述机器学习过程中用户隐私泄露的一种新型攻击手段，近年来引起了学术界广泛的关注，相关的研究成果越来越多地出现在安全领域知名的会议期刊中。然而，目前针对其研究方法的系统性、总结性的工作很少。为此，本文将从成员推理攻击的相关背景知识、攻击及防御手段等方面来系统性地介绍该方面的研究工作。本文从成员推理攻击的相关背景知识出发，依据是否训练影子模型对现有的攻击方式进行分类，阐述了成员推理攻击在其他领域的威胁，总结了几种常见的防御手段，对现有攻击手段的弱点进行了分析，并展望了未来的一些研究趋势。最后总结全文。

2 成员推理攻击的背景知识

成员推理攻击是一种通过一定的攻击手段来判断数据样本是否被用于机器学习目标模型的训练中，从而威胁到用户隐私的攻击方式。例如，如果攻击者推理得到一个病人的医疗数据被用来训练和疾病相关的机器学习模型，而这个模型被用来确定药剂的用量或者发现该疾病的遗传基础，那么攻击者就可以直接判断该用户患这种疾病，进而侵害到用户的隐私信息。在另一个场景中，美国国家标准与技术研究所（NIST）[4]和欧盟[5]明确表明成员推理攻击能够判断出哪些数据被作为训练集来训练目标模型，这是严重侵犯个人隐私的行为。我国于 2021 年 8 月颁布的《中华人民共和国个人信息保护法》[4)]表明敏感隐私信息一旦泄露，容易导致人格尊严受到侵害或者人身财产安全受到危害。同时，随着 MLaaS 技术的发展，攻击者滥用查询服务，进一步加剧了个人隐私泄露的风险。

2.1 成员推理攻击的定义

现有的成员推理攻击主要针对有监督的机器学习模型。有监督的机器学习，主要是学习一组输入和一组相关输出对应关系的行为。其解决问题的过程主要分为两个阶段：训练阶段和预测阶段。训练阶段结束后，将会得到一个完整的机器学习模型，用于求解特定的问题。在训练模型的过程中，通过损失函数 $L(y_i, f(x_i,\theta))$ 不断地逼近最小值，并不断改变模型内部参数 θ，使模型 $f(x_i,\theta)$ 的输出预测标签越来越接近样本 x_i 的真实标签 y_i。一个机器学习模型典型的定义为：

$$f(x_i,\theta)=\boldsymbol{P} \tag{1}$$

$$x_i \in D_{\text{train}}=\{X,Y\} \tag{2}$$

其中，f 为目标模型，x_i 为数据集 $\{X,Y\}$ 中的一个数据样本，y_i 为数据样本 x_i 对应的真实标签，x_i 为目标模型 f 的输入，θ 为目标模型内部参数，$\boldsymbol{P}$ 为目标模型的输出，为向量形式。

由于机器学习是在有限的数据集上训练，因此模型可能会对同一数据样本进行多次训练，使得模型有足够的能力记住训练集中的数据样本，这将极大地影响训练数据的隐私性。

成员推理攻击通过判断确定的数据样本 x_i 是否在训练目标模型的数据集 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ 中，从而获取到数据提供者的隐私。一个典型的 MIA 的攻击过程可以分为训练攻击模型阶段和实行攻击阶段。在训练攻击模型阶段，使用目标模型的输出作为攻击模型的输入，并将在训练集 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ 中的所有样本标记为成员，将不在训练集中的样本标记为非成员。对于训练完成的攻击模型 $\mathscr{A}$，把未知状态的数据样本 x_j 作为 $\mathscr{A}$ 的输入。最后根据 $\mathscr{A}$ 的输出结果，实现 x_j 是否为成员的推理。其表达式为：

$$\mathscr{A}(f(x_j,\theta),\theta^*)=\{0 \text{ or } 1\} \tag{3}$$

其中，$\mathscr{A}$ 为攻击者模型，θ^* 为攻击者模型的内部参数。把目标模型的输出 $\mathscr{P}$ 作为攻击者模型的输入，最终攻击者模型的输出为 0 或 1。若输出为 1，表明该样本 $x_j \in \mathscr{D}_{\text{train}}$，即判断为成员；若输出为 0，表明该样本 $x_j \notin \mathscr{D}_{\text{train}}$，即为非成员。

2.2 对手知识

根据成员推理攻击中攻击者所掌握的目标模型的背景知识，将其划分为黑盒成员推理攻击和白盒成员推理攻击。具体而言，在黑盒成员推理攻击场景中，如图 1(a)所示，攻击者对目标模型的了解是受限的，攻击者无法获得目标模型的内部结构，仅能得到目标模型对数据样本 x 的预测结果 $\hat{p}(y|x)$ 通过黑盒访问目标模型。在白盒成员推理攻击场景中，如图 1(b)所示，攻击者能够获得目标模型的全部信息，即攻击者不仅知道目标模型的训练算法、结构和各层训练参数，还能够访问训练数据集的分布。相较于黑盒情况下的成员推理攻击，在白盒情况下，攻击者不仅可以根据模型的预测结果 $\hat{p}(y|x)$ 判断数据样本的成员关系，还可以通过更多模型蕴含的其他

1) https://cloud.google.com/vertex-ai

2) https://aws.amazon.com/cn/machine-learning/

3) https://azure.microsoft.com/zh-cn/services/machine-learning/

4) www.npc.gov.cn/npc/c30834/202108/a8c4e3672c74491a80b53a172bb753fe.shtml

信息，如损失值、梯度等，来判断其是否为一个成员。一般而言，由于白盒攻击者对目标模型的了解更多，因此白盒 MIAs 的攻击成功率高于黑盒 MIAs。但是黑盒 MIAs 的方式更适用于现实场景，如通过 MLaaS 平台进行访问目标模型。在黑盒 MIAs 场景下，攻击者根据自己所掌握的有限知识进行推理攻击，比白盒 MIAs 有着更广阔的实际应用场景。

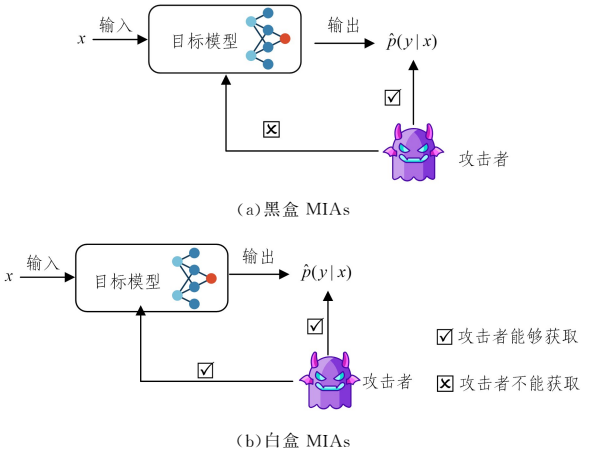


图 1 黑盒、白盒 MIAs 示意图

Fig. 1 Schematic diagram of black box and white box MIAs

3 MIA 攻击方法

近年来研究者们提出了很多关于成员推理攻击的方法，其中大量的攻击方法都着眼于如何通过更细粒度的分析或通过减少执行攻击所需的背景知识和计算能力来改进攻击方法。总的来说，现有的 MIAs 可分为两种：基于影子模型的 MIAs 和基于无影子模型的 MIAs。前者的攻击质量主要依赖于影子模型和目标模型的可转移性，影子模型和目标模型越相似，其攻击质量就越高。后者的攻击条件更为苛刻，攻击者只能获得很少的知识，例如仅能访问目标模型标签，从而实现 MIAs。

3.1 基于影子模型的 MIA

典型的基于影子模型的 MIA 框架是通过训练攻击模型来学习成员和非成员数据的预测概率分布的区别。如图 2 所示，其中第 1 步和第 2 步表示攻击前的准备过程，首先根据影子数据集训练影子模型，再根据影子模型模仿目标模型的输出结果和影子数据集，进行训练攻击模型。第 3－第 5 步为成员推理攻击的流程。现有的研究通过探索训练集中数据样本间的不同关系，或者对已知的辨别成员信息的方法进行改进，从而提高攻击方法的有效性。通过影子数据集训练得到影子模型，影子数据集和目标模型数据集相似程度越高，训练得到的影子模型和目标模型预测结果就越相似。因此，如何得到和目标模型数据集相似分布的影子数据集，也是一个值得关注的问题。除此之外，对于影子模型的结构，在白盒 MIAs 中，由于攻击者了解目标模型的内部结构，因此可以根据目标模型的结构设计出相同的影子模型结构。但在黑盒 MIAs 中，攻击者无法获得目标模型的内部结构，因此影子模型和目标模型具有不同的结构，这将限制影子模型和目标模型的可转移性。

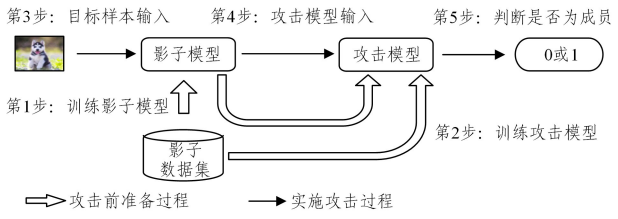


图 2 基于影子模型的 MIA

Fig. 2 MIA based on shadow model

3.1.1 多影子模型 MIA

Shokri 等[6]首次提出基于深度学习分类任务的 MIAs，这是一种通过训练攻击模型学习目标模型对于成员和非成员输出预测向量的区别，将影子模型和二分类模型作为攻击模型，来判断目标样本是否在训练集中。具体而言，如图 3 所示，攻击者是一个二分类模型，目标模型训练数据集 $\mathcal{D}_{\text{train}}^{\text{target}}$ 和影子数据集 $\mathcal{D}_1^{\text{shadow}},\cdots,\mathcal{D}_k^{\text{shadow}}$ 具有相同的分布，攻击者首先根据 K 个不相邻的影子数据集 $\mathcal{D}_1^{\text{shadow}},\cdots,\mathcal{D}_k^{\text{shadow}}$ 训练 K 个影子模型，并使用 $\mathcal{D}_1^{\text{shadow}},\cdots\mathcal{D}_k^{\text{shadow}}$ 分别作为影子模型的测试集。在训练攻击模型阶段，判断数据样本是来自 $\mathcal{D}^{\text{shadow}}$ 还是 $\mathcal{D}^{\text{test}}$，对于样本来自 $\mathcal{D}^{\text{shadow}}$ 的影子模型输出 y，将其样本全部标记为“成员”，标签$(y,1)$；对于样本来自 D^{test} 的影子模型输出 y，将其样本全部标记为“非成员”，标签$(y,0)$。

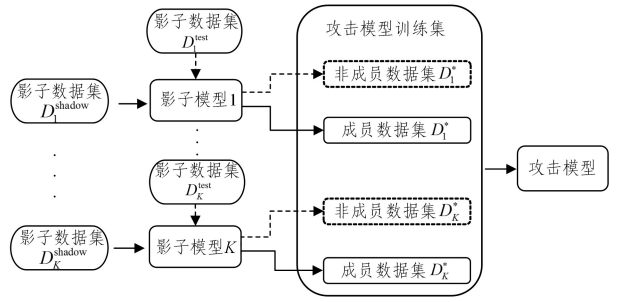


图 3 多影子模型攻击流程

Fig. 3 Multi-shadow model attack process

对于影子模型的输出 $\boldsymbol{y}$，构建攻击模型训练集 $\mathcal{D}^*$，使用 $\mathcal{D}^*$ 训练得到攻击模型 $\mathcal{A}$，训练完成的 $\mathcal{A}$ 只需将目标模型的输出 y 作为输入，便可得到 $\mathcal{A}$ 的输出结果为 0 或 1。若 A 输出为 1，则表明该样本在目标模型的训练集中，否则不在。由于影子数据集 $\mathcal{D}^{\text{shadow}}$ 和目标数据集 $\mathcal{D}_{\text{target}}$ 具有相同的分布，因此影子模型和目标模型对于同一个输入样本 x，两者的输出是相似的，故影子模型可以模仿目标模型的“行为”。对于 MIAs 攻击成功的原因，Shokri 等[6]指出，模型的过拟合越严重，就会导致越多的隐私泄露。此外，Long 等[7]认为，即使在泛化能力很好的 ML 模型中，也存在一些对目标模型产生独特影响的数据样本，这些样本较为脆弱，容易遭受 MIA 的威胁。Irolla 等[8]根据文献[6]提出的多影子模型框架，对成员推理攻击作了进一步的分析。该研究认为成员推理攻击能够攻击成功是因为攻击模型捕获到成员和非成员之间的错误分类差异，即如果目标模型正确预测了一个样本的标签，那么它就来自于训练集，否则不是训练集中的成员。

3.1.2 单影子模型 MIA

基于多影子模型的 MIAs 能够有效地攻击机器学习

模型,然而其攻击开销过大的问题也不容忽视[9,25]。首先,攻击者需要建立多个影子模型,每个模型与目标模型具有相同的结构。虽然这个过程可以利用 MLaaS 训练与目标模型相同结构的影子模型来实现,但开销很大。其次,用于训练影子模型的数据集与目标模型的训练数据需要相同的分布。这两种假设的要求比较苛刻且攻击开销大,大大缩小了成员推理攻击的实际应用范围。基于此,有研究者提出单影子模型 MIAs。

为解决上述问题,Salem 等[9]提出了 3 种不同的攻击者设定,通过逐步放宽这些假设,也能实现不错的攻击效果。对于攻击者 1 的设定:假设攻击者拥有和目标模型相同分布的数据集和模型结构,并且使用一个影子模型而不是多个,相比之前 Shokri 等[6]提出的多影子模型,单影子模型也能够达到相同的攻击效果。攻击者 2 没有和目标模型训练集相同分布的数据,且其不知道目标模型的结构,即对目标模型为黑盒访问。在对攻击者 2 的设定中,用于 MIA 的数据传输攻击在训练影子模型时,不再使用和目标模型相同分布的数据集,而是采用不同的数据集对影子模型进行训练。实验结果表明,在某些情况下,采用与目标模型相同分布的数据集训练影子模型,可以获得更高的攻击准确率。值得注意的是,在大多数情况下,结合来自不同领域的数据集所训练的影子模型,能够有效地进行相应的攻击。例如,使用 Purchase-10 训练的影子模型攻击 News 数据集训练的目标模型比同分布数据集训练的影子模型具有更高的攻击成功率。为了对其进行解释,文献[9]选择最高的 3 个成员和非成员的置信分数向量,使用 t-分布随机邻域嵌入(t-SNE)将其嵌入到二维空间中。实验发现,当攻击有效时,两个不同数据集中成员和非成员的数据点紧密地聚集在一起并且遵循一个共同的决策边界。因此,攻击模型在不同的两个数据集上具有更高的推理成功率。此外,攻击者根据目标模型返回的置信分数向量 $\mathcal{P}$,更容易推断出目标模型的训练集中的成员关系。因此,研究人员提出了隐藏目标模型的输出,攻击者仅能够得到目标模型预测标签前 K 个的概率。但文献[9]表明,即便攻击者得到目标模型的前 K 个预测标签的概率,其也能够实现较高的攻击成功率。

对于文献[9]提出的单影子模型所取得的攻击效果,有研究者使用单影子模型攻击方式对成员推理攻击做了进一步的研究和扩充。Liu 等[11]将成员推理攻击应用到通过社交媒体分析心理疾病行为中,并通过 GANs 生成和训练集相似分布的方法,通过生成、聚集合成数据,模拟目标模型得到影子模型。有研究者对单影子模型结构进行了总结,Truex 等[12]系统性地提出构建单影子成员推理攻击模型的一般框架,该框架能够适用于不同类型的目标模型的攻击。

除了对 MIAs 提出框架性的总结和新的推理方式之外,还有研究提出对成员推理攻击进行优化的理论方法。文献[13]通过贝叶斯优化策略对成员推理攻击进行优化,提出不依赖于目标模型的参数,仅依靠损失函数进行推理,使得在白盒或黑盒设定下都可以达到相同的攻击效果。随后,Nasr 等[14]通过理论分析随机梯度下降算法的隐私漏洞,对白盒成员推理攻击进行了更加细粒度的分析,发现不同层包含的信息不同,越靠近输出层,所包含的信息越多,其泄露的信息也越多,并将该攻击方法应用到集成学习和联邦学习领域中。除此之外,还有研究者研究扩展成员推理攻击的影响范围,Song 等[15]结合机器学习安全和隐私两个领域,研究成员推理攻击对对抗鲁棒性模型的攻击,并利用对抗鲁棒性的结构特征实行攻击,发现对抗鲁棒性模型比未部署防御的模型更容易受到成员推理攻击。Li 等[16]指出现有的 MIA 大多聚焦于数据样本级的攻击,这限制了其在真实世界中的应用,因而其提出了一个用户级的 MIA,其攻击过程是推理攻击者在没有精确的训练样本情况下,在训练过程中是否使用了来自目标用户的任何样本。Liu 等[17]提出在无监督学习中针对对比学习,通过黑盒访问和影子编码器对预训练的图像编码器进行成员推理攻击,从而进一步扩展成员推理攻击的范围。

除了通过区分影子模型的输出概率的分布来判断成员关系,也有研究指出通过计算影子模型的输出统计量,如交叉熵损失等,利用对比策略将其和提前设定的阈值进行比较,从而判断成员关系。图 4 给出了一个典型的基于对比策略的 MIA。攻击者提前设定一个阈值 τ,根据目标模型或者影子模型的输出结果计算最大后验概率或熵,得到 $\mathcal{A}(p)$,然后将 $\mathcal{A}(p)$与所选择的阈值进行比较,从而判断出该数据样本是否为成员,其中阈值 τ 的设定可以根据影子模型择优选择。该攻击设定也表明,即使在攻击者仅知道很少的知识这种简单的攻击设定下,也会导致训练集隐私信息的泄露。

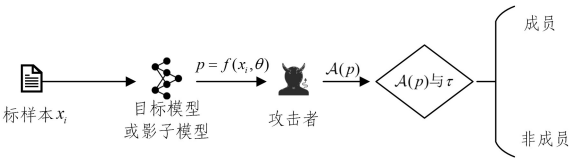


图 4 基于对比策略的 MIA

Fig. 4 MIA based on comparison strategy

Li 等[18]提出了比较交叉熵损失的迁移攻击(Transfer Attack)方案。具体而言,攻击者拥有和目标模型相同分布的数据集 $\mathcal{D}_{\text{target}}$ 和 $\mathcal{D}_{\text{shadow}}$,用 $\mathcal{D}_{\text{shadow}}$ 数据集训练影子模型 $\mathcal{S}$,从而使攻击者建立影子数据集和目标模型的联系。攻击者把待测数据样本作为影子模型 $\mathcal{S}$ 的输入,计算其交叉熵损失:

$$CE_{\text{Loss}}=-\sum_{i=0}^{K}1_y\log(p_i) \tag{4}$$

其中,p_i 为待测样本属于第 i 类的概率,K 为类别数,1_y 为真实标签 y 的独热编码。如果 CE_{Loss} 比设定的阈值 τ 更小,那么判断其为成员,否则为非成员。这是因为与非训练集相比,训练集中的样本损失值更小。Song 等[19]则认为现有的基于预测熵比较的 MIAs 没有考虑任何关于真实标签的信息,可能导致对成员和非成员的错误分类。例如正确分类为 1 和错误分类为 1 的概率,都将导致预测熵为 1,因此提出了新的度量指标来修正预测熵,如式(5)所示:

$$MH(\hat{p}(y|x),y)=-(1-p_y)\log p_y-\sum_{i\neq y}p_i\log(1-p_i) \tag{5}$$

其中,p_i 为样本 x 第 i 类的概率,p_y 为真实标签对应的概率。

上述工作所提出的 MIAs,主要根据目标模型的输出或对输出结果的统计分析来对训练集发起的攻击。如果对目标模型的输出进行扰动,在很大程度上会干扰攻击者,进而影响

MIAs 的性能。文献[20]提出抽样攻击(Sampling Attack),通过扰动函数 $pert()$对每个数据样本 x 生成 N 个扰动样本。在特定的扰动水平下,若 x 为成员样本,则标签不会发生变化;若为非成员,则标签发生变化。这种攻击的原理在于:对于训练集数据样本的一些小的扰动,训练后的目标模型可以返回一致的标签,即训练数据集中的样本比非训练集样本更加远离决策边界。但该方案攻击的准确度依赖于衡量数据样本到决策边界距离的方法。

3.2 无影子模型 MIA

上述基于影子模型 MIAs 的攻击质量主要依赖于影子模型和目标模型的可转移性,这些模型都是从代理数据集训练出离线的影子模型。为了保证 MIAs 的攻击质量,需要代理数据集和目标模型训练集拥有相似的分布,并使影子模型和目标模型具有相同的结构。然而,文献[21]在实际中难以获得和目标训练集相似分布的代理数据集和目标模型的结构。即训练出离线的影子模型也不同于目标模型,虽然输出概率分布相似,但是仍有差距,若影子模型和目标模型相差较大,那么攻击的表现也会大幅降低。因此有研究提出不依赖影子模型的成员推理攻击,这是一种更加贴合实际的攻击方式。本文将无影子模型 MIAs 划分为两类,即基于预测标签的 MIAs 和基于对比策略的 MIAs。

3.2.1 基于预测标签的 MIAs

有研究者提出基于预测标签的 MIAs,且不再依赖影子模型。Leino 等[22]提出基于标签的简单基线攻击,当模型输出正确的预测标签时,将该样本定为成员,否则为非成员。该攻击对现有的模型抵御 MIAs 提供了一个很好的评估方法,但该方案不能作为有效的 MIAs 攻击手段。基于此,Choo 等[23]提出仅仅依靠模型预测标签的成员推理攻击,该攻击不仅和基于预测向量的 MIAs 具有相同的攻击性能,而且还可以无视基于扰动目标模型输出的防御机制,如 MemGuard[24]。实验结果表明该攻击和需要获得置信概率分布的攻击一样有效。具体而言,其对于在训练集中的训练样本具有高鲁棒性,而对于非训练集中的样本更接近决策边界。当对非训练集样本 x 进行扰动时,其更容易受到干扰,从而改变目标模型的预测标签。相较于在训练集中的样本,对样本的输入进行扰动,目标模型的预测标签不易发生变化。更形式化的表述为:用一个数据点到决策边界的 L_2 范数表示数据样本到目标模型决策边界的估计距离 $dist_h(x,y)$,如果 $dist_h(x,y)>\tau$,其中 τ 为阈值,则预测 x 为成员,否则为非成员。阈值 τ 的选择可以通过影子模型进行调参。但使用 L_2 范数衡量决策边界距离 $dist_h(x,y)$的精确度也会直接影响攻击的准确率。此外,Li 等[18]也提出了仅仅依靠预测标签的边界攻击,对目标模型的输入增加一个扰动噪声,使其变为对抗样本,成员的数据样本相比非成员的样本更难发现预测标签的变化。

3.2.2 基于对比策略的 MIAs

在不需要影子模型的研究中,较多的 MIAs 是基于对比策略而发起的攻击,攻击流程如图 4 所示。文献[9]对于攻击者的设定,仅仅依赖目标模型的输出结果,使用后验统计量,根据目标模型的最大后验概率 $\mathrm{Max}(\hat{p}(y|x))$判断成员和非成员。该方法能够实现非常高的攻击准确度。同时,通过比较预测熵进行攻击,研究者认为训练集中的样本预测熵应该比测试集中的预测熵更小。样本预测熵的表达式为:

$$E(\hat{p}(y|x))=-\sum_{p_i\in y}p_i\log p_i \tag{6}$$

其中,p_i 表示第 i 类别上的预测概率。

Yeom 等[25]认为训练集中的数据样本应该比非训练集中的数据样本有着更小的预测损失,如式(6)所示。因此,他们提出攻击者可以通过比较预测损失来判断该样本是否为成员。如果目标样本预测损失小于阈值 τ,则认为样本为成员,反之为非成员。

$$\mathbb{E}[\mathscr{L}(f(x,\theta),y)-\mathscr{L}(f(X,\theta),Y)]\leqslant\tau \tag{7}$$

其中,$\mathscr{L}(f(x,\theta),y)$为样本 x 的预测损失,$\mathscr{L}(f(X,\theta),Y)$为训练集 X 的平均预测损失,$\mathbb{E}$为期望,τ 为阈值。

Rezaei 等[26]通过比较决策边界的距离将成员推理攻击应用到深度集成学习中,并通过比较构成集成模型之间的一致性,来权衡隐私和效用。文献[1]提出通过对抗样本实现成员推理攻击的方法。其基本思想是,对于数据样本 x,通过添加噪声 V 生成对抗样本,使其原有的预测标签发生改变,那么在训练集中的数据样本需要添加比非训练集数据样本更大的噪声。因此通过衡量添加噪声的大小,可以判断成员关系。

这种"一刀切"的决策方式仅适用于简单决策边界,对于决策边界复杂的目标模型则不能进行有效的攻击。因此,文献[21]提出了一种新颖的攻击方法 BLINDMI。该方法主要利用差分比较发现现有的成员和目标模型之间的关系,并通过移动目标数据集中的数据样本,观察两个数据集之间的变化差异,从而判断成员关系。具体而言,攻击者首先准备非成员数据集概率分布子空间 $S^{\mathrm{prob}}_{\mathrm{Non\text{-}mem}}$ 和目标模型数据集 $S^{\mathrm{prob}}_{\mathrm{target}}$,差分比较就是测量 $S^{\mathrm{prob}}_{\mathrm{Non\text{-}mem}}$ 和 $S^{\mathrm{prob}}_{\mathrm{target}}$ 的距离 d,然后一个样本从 $S^{\mathrm{prob}}_{\mathrm{target}}$ 移动到 $S^{\mathrm{prob}}_{\mathrm{Non\text{-}mem}}$ 中,再计算两者的距离 d'。若 d 大于 d',则移动的样本为非成员;若 d 小于 d',则移动的样本为成员,重复迭代过程,直到收敛。如果非成员样本靠近非成员数据集,那么 $S^{\mathrm{prob}}_{\mathrm{target}}$ 和 $S^{\mathrm{prob}}_{\mathrm{Non\text{-}mem}}$ 就会产生排斥,导致距离变大,反之亦然。两个数据集之间主要通过差分比较来实现。具体而言,在再生核希尔伯特空间(Reproducing Kernel Hilbert Space,RKHS),通过最大平均偏差(Maximum Mean Discrepancy,MMD)计算两个核空间在质心的距离。MMD 能够反映出两个分布的相似度,MMD 值越小,其两个概率分布也就越相似。即使攻击者不知道目标模型的结构和训练数据集的真实标签,不需要影子模型,也能实现较高的攻击成功率。此外,Rezaei 等[27]提出通过比较目标样本和使用 GANs 生成的亚样本(语义相似的样本)经过目标模型的输出概率的不同,能够有效地判断成员关系,这是一种更加轻便、开销更低的 MIA。

在图像翻译模型中 Shafran 等[28]提出在不使用影子模型和影子数据集的情况下,通过目标模型的输入和输出构架重建误差,并根据线性预测和真实标签构建预测误差,最后将重构误差减去预测误差得到成员误差,用此判断成员关系,并提出不确定性和输出维度是 MIA 攻击成功的两个因素。Yuan 等[29]首次分析了在神经网络剪枝上的隐私风险,提出了自

注意力 MIA，其通过各个输出类别间的置信度和灵敏度差异，来判断该样本是不是成员。实验结果表明被剪枝的神经网络存在大量隐私泄露的风险，在攻击者不知道剪枝方法的情况下，也能够实现高性能的 MIA。

对于成员推理攻击为什么能够成功攻击机器学习模型，之前的研究均未探究模型的内部结构，有文献提出了更加细粒度的研究。Leino 等[22]首次从模型的内部结构研究在白盒情况下的 MIA 攻击成功的原因。模型在使用特征进行分类时，对于数据集中时常出现的特征，模型可能会学习得到一个内部层，以此辨别该特征，模型的该种行为表明成员的信息会被泄露。

4 典型机器学习模型下的 MIA

4.1 生成模型

近年来，深度生成模型作为一种抽象和近似数据分布的有效方法，受到了广泛的关注。其中最常用的两种方法是变分自动编码器（Variational Auto-Encoder，VAE）和生成对抗网络（Generative Adversarial Network，GAN）。图 5 给出了 GAN 模型的基本框架，首先生成器网络 G 通过随机噪声生成一张假图片，然后将假图片和真图片输入到鉴别器 D 中，鉴别器 D 将鉴别出的结果反馈给生成器和鉴别器。若鉴别假图片为真，则需要调整鉴别器网络参数，以便更好地鉴别正确图片；若鉴别真图片为真，则将调整生成器网络参数，以便生成更加真实的假图片。生成器网络和鉴别器网络通过动态的“博弈过程”，最终得到高判断力的鉴别器网络和高伪造能力的生成器网络。

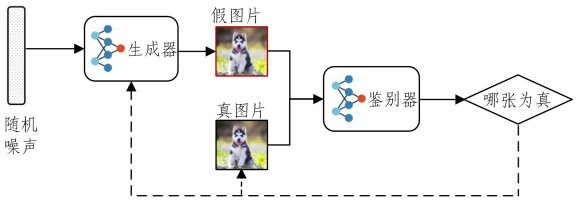


图 5 GAN 基本框架

Fig. 5 Basic framework of generative adversarial networks

有研究表明，生成模型也容易受到 MIAs 的威胁[30]，导致训练数据集中隐私的泄露。Hayes 等[30]针对 GANs 提出了比较全面的 MIAs，并且评估了该攻击和使用 MLAAS 各自的攻击成本，将其应用到 DCGAN 和 BEGAN 中，发现在白盒下能实现极高的攻击准确度。紧接着，Hilprecht 等[31]提出了适用于所有生成模型的蒙特卡洛（Monte Carlo，MC）攻击和针对 VAEs 的重建攻击，发现 VAEs 比 GANs 更容易受到成员推理攻击，但是 MC 攻击依赖训练数据特征的复制，如果样本质量不足，就无法实现有效的攻击。

前文所述的攻击生成模型都是对单个数据样本实行 MIAs，Liu 等[32]提出了针对生成模型的多个实例的 MIAs，通过最终的重构误差判断样本是否为成员。Chen 等[33]依据攻击者是否知道生成器和鉴别器的模型结构，首次提出针对 GAN 模型的 MIAs 进行分类的方法，将攻击 GANs 模型的 MIAs 分为 4 种类型：全黑盒生成器、部分黑盒生成器、白盒生成器和可访问鉴别器（全模型），并提出了一个通用的攻击模型和基于理论的攻击校准技术，在所有攻击情况下都能够不断提升攻击性能。更进一步地，文献[34]提出了一种针对 GAN 生成人脸的攻击，可以准确地识别与训练样本具有相同身份的样本，而不是先前研究中所提到的与训练样本相同的样本。为了提高针对 GANs 推理的精确度，Hu 等[35]提出了针对白盒 GANs 的成员推理攻击，通过构建 MC（Member Confidence）分数代表每个区域的过度代表程度，并对 MC 分数进行比较，得出样本和数据集的关系。

4.2 联邦学习

联邦学习（Federated Learning，FL）是在多个客户端不知道彼此数据的情况下，允许多个客户端一起训练同一个目标模型。如图 6 所示，多个客户端分别在本地训练目标模型 $f(x,w_i^{\{t\}})$，将训练完成的目标模型权值 $w_i^{\{t\}}$ 上传到中央服务器，中央服务器对各个客户端上传的权值进行整合，最后生成最终模型 $f(x,W^{\{t\}})$。联邦学习能够保证客户端数据在不离开本地的情况下，联合训练目标模型，从而保护了客户端数据的隐私。

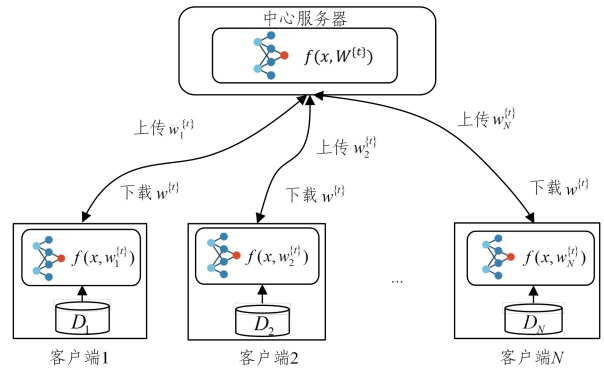


图 6 联邦学习算法框架

Fig. 6 Federated learning algorithm framework

但近来有研究指出 FL 容易受到成员推理攻击，因为 MIA 可以区分模型训练集中的成员和非成员信息。Truex 等[12]提出，在联邦学习中，客户端使用决策树模型进行预测时容易受到成员推理攻击。Chen 等[36]通过 GANs 丰富训练集数据，提出用户级的 MIA 来攻击联邦学习中各个客户端的隐私。Melis 等[37]证实客户端上传数据到中央服务器会泄露客户端训练数据的信息。Zhang 等[38]提出现有的攻击由于缺乏相应的攻击数据，不能实现很好的攻击性能。为了克服攻击数据不足的问题，他们提出攻击者可以利用 GANs 增加攻击数据的多样性，从而提升在 FL 上的 MIA 攻击性能。Hu 等[39]指出现有的针对 FL 的 MIAs 不能区别训练成员来自哪个客户端，因此提出源推理攻击。源推理攻击可以得到一个训练成员的来源的最优评估，通过贝叶斯证明一个诚实但好奇的服务器能够在不违反 FL 协议的基础上执行源推理攻击进而窃取训练成员的来源信息。Pichler 等[40]提出一个不诚实的中央服务器框架，并利用 ReLU 函数的激活特性实现高准确度的攻击。

4.3 MIA 对其他模型的攻击

成员推理攻击不关注数据样本的内容，只关注样本是否

在训练集中出现过。目前较多的成员推理攻击被应用在分类模型上，如工业物联网领域[41]。在语义分割领域，Zhang等[42]在文献[43]的研究基础上，提出了仅需要标签的MIA语义分割的框架，通过数据增强(Data Augment)和数据表示过程生成攻击模型的数据集。此外，MIAs还被应用到推荐系统[44]、知识图谱[45]、语音模型[46-47]、图神经网络[48-51]、图嵌入模型[52]、文本嵌入模型[53]、文本生成模型[54-55]、单词嵌入模型[56-57]、掩码语言模型[58]、基因序列分析[59]、彩票网络[60]、医学图像方面神经影像[61]、临床语言模型[62]等。Ye等[63]提出了一种通过成员推理攻击审计机器学习隐私风险的框架，该框架用来推导攻击策略，并分析除模型泄露之外的影响攻击性能的因素。

上述研究主要关注从头开始训练的目标模型。深度学习模型的训练需要大量的数据集和计算资源，为了数据资源的可获得性和控制训练任务的成本，引入了迁移学习相关模型。文献[64]首次提出将成员推理攻击应用到迁移学习中，该实验共设计了3种不同的攻击场景。实验结果表明学生模型不会泄露教师数据集的信息，教师模型中冻结层数的增加，会导致学生模型攻击性能下降。Chen等[65]对迁移学习中的隐私泄露风险进行了系统的分析，并对前面的研究进行了相应的分类。

针对上述常见的攻击方法，本文对主要的MIAs方法进行总结、分类，提炼了MIAs攻击方法的关键技术，并指出了其优点和不足之处。具体如表1所列。

表1 主要MIA方法总结

Table 1 Summary of main MIA methods

分类	攻击策略	对手知识	文献	关键技术	优点	不足
基于影子模型	多影子模型	黑盒	[6]	多个影子模型模仿目标模型	无须多次访问目标模型，依赖模型输出	需训练攻击模型和多个影子模型，攻击开销大
		黑盒	[7]	存在易受攻击的脆弱样本	对泛化能力好的模型，攻击仍然有效	攻击开销大
	单影子模型	黑盒	[9]	单影子模型使用不同的影子训练集	攻击开销较小，提高了攻击的灵活性	依赖影子模型和目标模型的可转移性
		黑盒	[19]	比较各个样本预测熵	无须训练攻击模型	存在错误预测的情况
		黑盒	[19]	基于比较修正预测熵	相比预测熵提高了攻击的性能	攻击依赖阈值的选择
		黑盒	[18]	提出基于损失值推理成员关系	减少了攻击者知识，不需要训练攻击模型，不受干扰预测结果防御的限制	攻击精度依赖于损失值阈值的设定，攻击效果较差
		白盒	[14]	分析目标模型的前向传播和后向传播	白盒情况下，仅需根据模型训练参数的更新信息进行攻击	对目标模型结构和参数要求高，攻击范围较小
		黑盒	[20]	通过扰动数据样本，造成预测标签变化	不需要训练攻击模型，攻击精度较高	攻击性能依赖于扰动值的设定
无影子模型	基于预测标签	黑盒	[18]	添加扰动，依靠预测标签的变化进行边界攻击	攻击开销小，攻击精度较高，不受干扰预测结果防御的限制	需要对目标模型进行多次查询
		黑盒	[22]	模型预测正确即作为成员	攻击开销小，是简单的基线攻击方法	攻击性较差
		黑盒	[23]	依据数据样本到决策边界的距离进行推理攻击	攻击开销小，攻击精度较高，不受干扰预测结果防御的限制	攻击性能依赖阈值的选择，含有不确定性
	基于对比策略	黑盒/白盒	[13,25]	仅依赖预测损失	攻击在白盒、黑盒情况下都能够实现相同的攻击效果，计算开销更小	依赖于对损失阈值的选取
		黑盒	[21]	使用差分比较方法	减少了攻击开销，适用于复杂的决策边界	需准备非成员数据集
		黑盒	[27]	通过噪声的大小和阈值比较	无视目标模型结构，不需要训练数据，攻击性能高	攻击的性能依赖于扰动和阈值的选择
		黑盒	[28]	通过成员误差攻击图像翻译模型	能够实现较高的攻击准确率	适用于图像翻译和语义分割领域
		黑盒	[29]	比较类别间的预测和灵敏度差异	具有较好的攻击效果，更细粒度分析预测的差异	受神经网络稀疏度影响

5 现有MIA的不足之处

Rezaei等[66]指出，现有的大多数成员推理攻击只关注到正类的评价指标，而忽视了对负类的评价。首先，很多研究中没有提供目标模型的训练准确度和测试准确度，即能够表示该模型的泛化能力的参考。其次，多数的成员推理攻击仅仅给出攻击的准确率、精确度和召回率，这些评价指标不足以表现MIAs的性能。该研究的实验结果表明，即使攻击者掌握了目标模型的结构和参数等，例如实验中除了使用目标模型的输出，还使用到中间层、梯度、模型权重、决策距离等参数，目前最先进的成员推理攻击也不能同时实现高精度和低误识率。研究还指出深度学习模型在正确分类的样本中经常是相似的，无论它们是否被用来训练目标模型。除此之外，在数据样本不均衡的情况下，只关注精确率和召回率会产生

攻击成功的错觉。文献[67]指出，现有的许多深度网络架构判断训练集中的数据会表现不一致，这是错误的看法。也就是说，即使有些数据样本非训练集数据，但模型对其也有可能是自信的。因此，通过使用模型的预测向量进行 MIAs，从而获取训练数据的隐私信息的能力可能被高估了。但是目前对于 MIAs 为什么能够推理成功，还没有定论。

6 MIA 防御方法

文献[7]指出成员推理攻击问题缘于目标模型会携带训练数据中个体的特征信息，这种仅能描述自己的特征信息被称为“噪声”，它能够描述其他数据的特征信息对目标模型是有益的。而这种携带特征信息的程度可以用过拟合程度来衡量。过拟合意味着机器学习模型更加能够记住自己已经训练过的数据样本，因此当目标模型训练过拟合时，相比没有遇到的数据样本，机器学习模型对于其已经训练过的数据样本会有更大的信心。基于此，有关防御成员推理攻击的研究大致有两个主要的研究思路：1)降低模型过拟合程度，增强模型的泛化能力；2)通过扰动的方法，保护数据集的隐私。

6.1 基于预防模型过拟合

现有的文献已经表明[6-7,20,68]，机器学习模型过拟合是 MIAs 高效的重要原因之一，因此可以通过降低目标模型的过拟合程度，增强模型的泛化性能，来实现防御推理攻击的目的。

在机器学习模型训练过程中，正则化技术经常被用来克服模型的过拟合，增加模型的鲁棒性，提升模型的泛化能力，这能减少模型在训练数据和测试数据的表现差异，从而实现对成员推理攻击的防御。

6.1.1 L_1，L_2 正则化

L_1 正则化的目的是使其中靠近输出层的权重参数 w 变为 0，从而减少模型的过拟合。L_2 正则化的目的是使神经网络的权值衰减，权值参数变为接近 0 的值，从而减少模型的过拟合。Shokri 等[6]通过使用 L_2 正则化来降低目标模型的过拟合程度，通过在模型的损失函数中添加 $\lambda\sum_i\theta_i^2$ 来惩罚训练模型中的大参数，鼓励较小的参数，如式(8)所示：

$$L=L(\theta)+\lambda\sum_i\theta_i^2 \tag{8}$$

其中，θ_i 是模型的参数。λ 越大，在训练过程中正则化的效果越强。但是正则化对抵御模型过拟合是有限的，因此 Li 等[68]将正则化和最大平均差异损失(MMD)相结合来弥合目标模型的训练精度和测试精度，有效降低模型的脆弱性。

6.1.2 Dropout 机制

一个全连接的神经网络包含数以万计的模型参数，这使得模型更易于过拟合。文献[69]最早提出 Dropout 机制。在训练神经网络的每次迭代中，Dropout 都以固定的概率 p 丢失神经元，从而一定程度地降低模型复杂程度以及模型的过拟合，如图 7 所示。现有文献[6,9,70-71]在模型训练的过程中，对目标模型的每层增加 dropout 机制，以降低模型的复杂度。实验证实，在所有的攻击设定下，MIA 的攻击性能都会降低，但是该防御机制仅适用于目标模型是神经网络的情况。

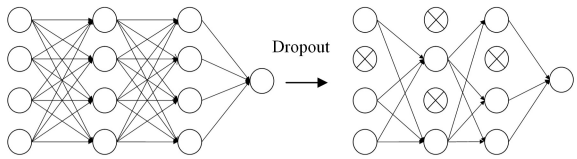


图 7 Dropout 示意图

Fig. 7 Dropout diagram

6.1.3 数据增强

数据增强意味着增加更多的数据，因此可以将数据增强应用到训练数据集中，通过增加训练数据的多样性来减少模型过拟合，增强目标模型的泛化能力。神经网络中包含众多的模型参数，如果用于训练模型的数据太少，而神经网络又很复杂，会大大降低模型的泛化能力，容易产生过拟合现象。一般而言，参加训练的数据越多，训练得到的模型泛化能力就越强。因此，为了增加训练数据量，可以对图片进行一些调整，如可以通过旋转、翻转、缩放等方法，生成更多的图片。如图 8 所示，分别对原始图像进行裁剪、缩放和反转得到新的图像。Kaya 等[72]在深度学习中应用数据增强来提高目标模型的准确度并减小模型的过拟合程度，从而降低成员推理攻击的风险。而 Yu 等[73]提出依赖增加的数据集实现对成员推理攻击的优化，并且通过理论分析得出，当输出概率分布遵循贝叶斯分布时，可推导出最优的成员推理攻击表达式。

图 8 数据增强示意图

Fig. 8 Data augmentation diagram

6.1.4 提前停止

提前停止是由文献[70]提出的。在训练过程中，机器学习模型在训练次数过多的情况下往往会产生过拟合的情况，提前停止是一种提前结束训练的策略，用来防止过拟合。如果在训练过程中连续 n 个周期都没有达到提前设置的最佳测试准确度 P 时，继续训练模型可能也达不到更好的效果，因此可以认为模型已经达到最佳训练准确度，继续训练下去模型可能会出现过拟合，此时便可以提前停止迭代。Kaya 等[74]通过比较提前停止、Dropout、标签平滑等正则化机制来抵御 MIAs，并且发现提前停止是一个更为实际的选择，其不需要任何调优就可以产生理想的预测。

6.1.5 Min-Max 机制

Nasr 等[75]提出 min-max 机制，即通过对抗训练算法最小化模型预测损失并且同时最大化推理攻击的增益，如式(9)所示：

$$\min_f(L_{\mathcal{D}}(f)+\lambda\max_h G_{f,\mathcal{D},\mathcal{D}'}(h)) \tag{9}$$

其中，$L_{\mathcal{D}}(f)$ 表示模型训练集 $\mathcal{D}$ 上的损失；$\max_h G_{f,\mathcal{D},\mathcal{D}'}(h)$ 表示在两个不相交数据集 $\mathcal{D}$ 和 $\mathcal{D}'$ 上计算推理模型的最大经验增益，即发现一个最强健的推理模型 h，进而攻击目标分类模型 f，从而得到最能抵御成员推理攻击的目标模型 f；最外层的最小化用于找到针对推理模型 h 的最好的防御分类模型 f；参数 λ 用于控制模型精度的优化和保护成员的隐私。实验

结果表明，在 Purchase100 数据集上，目标模型在得到 76.5% 的测试精度的基础上，进行 MIA 的精度为 51.8%，接近随机猜测。用 L_2 正则化作为对比实验的结果表明，当模型测试精度为 32.1%时，L_2 正则化才能达到和 min-max 机制相同水平的隐私保护。

6.1.6 模型堆叠

Salem 等[9]提出了另一种防御机制——模型堆叠(Model Stacking)。该防御机制独立于所用的 ML 分类器，其认为，如果目标模型的不同部分用不同的数据子集进行训练，那么合并后的模型将不容易发生过拟合。通过集成学习使得大量的子 ML 模型结合成最终的完整模型；并且，为了最大化防止模型过拟合，训练子模型时可以使用不相邻的数据集来训练，如图 9 所示。

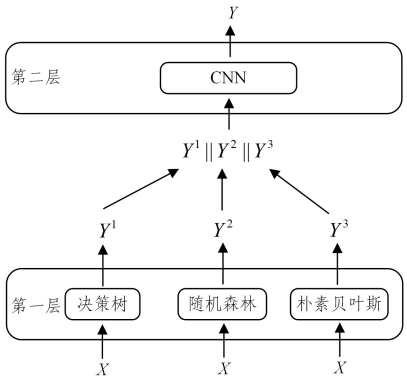


图 9 模型堆叠框架[9]

Fig 9 Model stacking framework[9]

6.1.7 RelaxLoss 机制

Chen 等[76]提出了一种基于 RelaxLoss 训练目标模型的框架，该方案能够缩小泛化误差，减少隐私泄露，同时不损失目标模型的效用。具体地，首先，为了减少成员数据和非成员数据在损失分布上的差异，该方案通过使用梯度上升策略，为目标损失设定一个更可实现的损失平均值，从而使损失值被放宽到对非成员数据更加容易实现的水平；其次，为了保证模型效用，防止错误类别预测分数超过正确类别预测分数，进而导致不正确的预测。该方案通过将错误类别预测分数展平，从而增加正确类别预测分数和错误类别预测分数的区分度，保证目标模型的准确率。

6.2 基于模型压缩技术

模型压缩大多被应用到复杂的神经网络模型中，其目的是在不影响模型性能的前提下减小模型的体积，降低模型的复杂度，减小模型规模，增强模型的泛化能力。

6.2.1 知识蒸馏

知识蒸馏(Knowledge Distillation，KD)是减小模型规模的一种常用的迁移学习方法，最早由 Hinton 等[71]提出。该方法通过教师模型的输出来训练一个规模更小的学生模型，根据教师模型训练出来的学生模型，与教师模型有着相似的预测准确度。知识蒸馏不仅可以提升模型精度，压缩网络参数，还能够将两个不同的数据集进行集成和迁移。

一般来说，KD 是为了将泛化能力(知识)从原始模型转移到蒸馏模型，而不需要进行效用退化。一旦提炼出来的模型被训练完成，它就可以在许多场景下取代原始模型，因为它的计算效率更高，对资源的依赖性更小。Liu 等[77]不仅将 KD 应用到防御 MIA 中，还将其应用到其他推理攻击的防御中。Zheng 等[78]为了更好地权衡隐私和效用，提出了两种算法：互补知识蒸馏(Complementary Knowledge Distillation，CKD)和伪互补知识蒸馏(Pseudo Complementary Knowledge Distillation，PCKD)。在 CKD 算法中，知识蒸馏的迁移数据全部来自隐私数据集，但是它们从被训练的教师模型互补数据集中产生软标签。如图 10 所示，首先隐私训练集 $\mathcal{D}$ 被划分成 K 份不相邻的子集 $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$，分别用 $\mathcal{D}-\mathcal{D}_i$ 训练教师模型 $f_{\theta_i}^T$，其教师模型训练算法和模型结构均相同。然后通过数据集 D_i 和教师模型 $f_{\theta_i}^T$ 产生预测向量，计算产生合成数据集 $\bar{D}$。最后在合成数据集 $\bar{D}$ 中选择迁移数据集 D'，从而训练得到学生模型 f_θ^S。在 PCKD 中，它减少了每个教师模型的训练集，并使用模型平均来生成传输数据的软目标。由于训练集较小会导致效用较差，因此 PCKD 利用预训练来提高教师模型的效用。

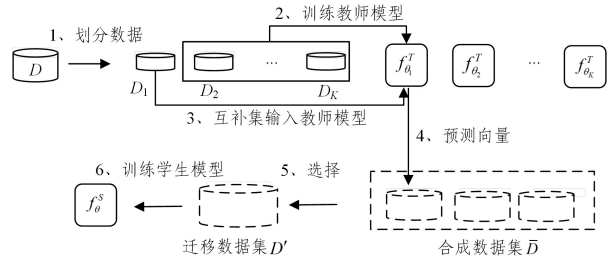


图 10 CKD 框架[78]

Fig. 10 CKD framework[78]

文献[79]提出蒸馏成员隐私(Distillation for Membership Privacy，DMP)方案。DMP 的主要目的是通过减少成员和非成员的损失差距，来保护目标模型免受成员推理攻击。该方案共分为 3 步，如图 11 所示，首先 DMP 需要一个隐私训练数据集 $\mathcal{D}_{\text{tr}}$ 和一个未标记的参考数据集 $\mathcal{D}_{\text{ref}}$，DMP 通过隐私训练数据集 $\mathcal{D}_{\text{tr}}$ 训练一个不受保护的教师模型 f_θ^T，并使用它来标记未标记的参考数据集中的数据样本；然后 DMP 从已经标记的参考数据集中选择具有低预测熵的数据样本来训练目标模型；最后 DMP 根据选择出的数据样本构成训练集训练个人模型 f_θ^S。

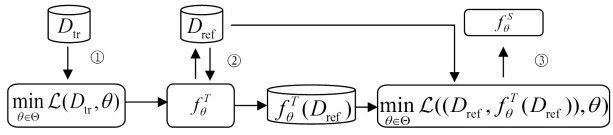


图 11 DMP 框架[79]

Fig. 11 DMP framework[79]

6.2.2 模型剪枝

剪枝(Pruning)技术是神经网络中一种压缩模型的技术，通过剪掉一部分对模型输出影响微乎其微的参数，实现对模型的压缩，同时又不会降低模型的预测精度，主要适用于计算和存储资源受限的设备。Wang 等[80]将权重剪枝应用到 MIA 的防御中，实验结果表明使用该防御机制比 min-max 防御机制的效果更好。Yuan 等[29]指出现有的大部分研究重点

集中在通过移除不显著的参数和重新训练剪枝后的模型,进而平衡剪枝网络的精度和稀疏度。但重新训练模型和重用训练样本会增加隐私泄露的风险,因此 Yuan 等提出了针对模型剪枝的成员推理攻击,此外还提出通过减少基于 KL 散度距离的预测来保护剪枝过程。

6.3 基于扰动的防御策略

目前大部分 MIAs 是根据目标模型的输出结果或者目标模型内部参数的变化,从而实现对数据样本的推理。有研究指出,可通过对模型的输出和内部参数进行扰动,从而隐藏真实的模型参数和输出,达到愚弄攻击者的目的。

6.3.1 差分隐私

差分隐私(Differential Privacy,DP)最早由 Dwork 等[81]提出。一个随机算法 $\mathcal{M}:\mathcal{D}\to\mathcal{R}$,$d,d'\in\mathcal{D}$ 是两个相邻的数据集,它们仅有一个数据记录不同,对于输出的任何子集 $\mathcal{S}\subseteq\mathcal{R}$,满足如下条件:

$$\mathcal{P}_r[\mathcal{M}(d)\in\mathcal{S}]\leqslant e^{\varepsilon}\mathcal{P}_r[\mathcal{M}(d')\in\mathcal{S}]+\delta \tag{10}$$

其中,$\mathcal{M}$ 为随机算法,目的是对原始信息产生扰动;$\mathcal{P}_r$ 为算法的输出概率;ε 为隐私预算,$\varepsilon>0$,ε 越小代表 DP 隐私保护性越好,表现为目标模型不能记住数据集中任何数据样本;δ 也为隐私预算,$0\leqslant\delta<1$,代表可容忍的隐私预算超出 ε 的概率。综上,其满足(ε,δ)差分隐私。差分隐私提供了一个对隐私概念的直观理解,即一个样本是否在数据集中并不会很大程度地改变结果的输出概率,因此差分隐私既可以保护数据集 $\mathcal{D}$ 上的个人隐私,也可以保护模型的输出隐私,但这需要损失一定的模型准确度。

Abadi 等[82]将 DP 推广到深度学习模型中,在模型训练过程中提出差分隐私随机下降算法(Differential Privacy Stochastic Gradient Descent,DP-SGD)。DP-SGD 是最具代表性的利用 DP 机制来保护 ML 模型的方案。目前部署 DP 抵御 MIAs 通常使用该机制,其通过在训练模型的过程中对梯度增加噪声,进而代替正常的梯度下降,扰动与数据相关的目标函数,并减轻了推理攻击。一般来说,在模型训练过程中,DP-SGD 添加高斯噪声(Gaussian Noise)到梯度 g。由于没有先验知识来确定单个训练样本对梯度 g 的影响,因此不能直接计算 g 的灵敏度。DP-SGD 通过裁剪梯度 g 来解决此问题,如式(11)和式(12)所示:

$$\bar{g}=\frac{g}{\max\left\{1,\frac{\|g\|_2}{C}\right\}} \tag{11}$$

$$\tilde{g}=g+N(0,\sigma^2C^2I) \tag{12}$$

式(11)表示梯度裁剪,其中 C 为设定的阈值。式(12)表示增加服从高斯分布的噪声,其中 σ 为噪声尺度,DP-SGD 对模型梯度进行扰动,从而干扰攻击者的判断。基于此,Shokri 等[6]将 DP-SGD 应用到防御成员推理攻击中,但这会损失模型的精度。Yeom 等[25]从理论上将 DP 和 MIA 联系在一起,并证实攻击的表现和隐私预算 ε 相关。DP 保证在数据集中任何单一数据样本对输出的影响是有限的。此外,Rahman 等[83]在部署最优良的 DP 的深度学习模型(Differentially Private Deep Model,DPDM)上[82]进行成员推理攻击,他们发现,DPDM 提供越高的模型效用,就越容易受到成员推理攻击。也有研究者更加细粒度地分析了 DP-SGD 隐私保护的上界和下界。文献[84]通过定量和实验的方法,分析 DP-SGD 能否提供比理论分析更好的隐私保护效果。Truex 等[85]提出在目标模型训练数据分布不均衡的情况下,会大大提升成员推理攻击的攻击成功率,并应用差分隐私权衡该情况下的隐私和效用。Naseri 等[86]提出评估联邦学习中的本地差分隐私和中心差分隐私技术的可行性和有效性。Bernau 等[87]提出在分类模型中比较本地差分隐私和中心差分隐私抵御成员推理攻击的效率。

现有的大多研究采用 DP-SGD 方法部署差分隐私机制,包括分类模型和生成模型[19,21,24,28-29,32,87-93]。文献[94]指出任何使用(ε,δ)的差分隐私训练的模型其成员推理攻击准确度的上界由(ε,δ)决定。但文献[95]认为这些通过差分隐私获得的边界证明不足以在实际场景中表示成员推理攻击的边界,因此提出使用高斯噪声机制获得最优的成员推理攻击边界。Chen 等[59]使用差分隐私抵御成员推理攻击来保护基因组数据。文献[77]表明 DP 不仅可以防御成员推理攻击,还能够防御模型反演和属性推理攻击。尽管 DP 有着广泛的应用场景,但其缺点是,要想达到较好的隐私保护,就要损失较高的模型精度,因此如何权衡隐私和效用的关系是 DP 需要解决的问题。文献[96]分别对传统机器学习和深度学习使用 DP 隐私保护进行分类,并进行了详细的探讨。

6.3.2 干预输出结果

现有文献[6-7,9]提出的攻击方法是比较成员和非成员样本输出结果的差异。基于此,可以通过对模型的输出结果进行扰动,从而干扰攻击模型的判断。文献[6]提出 Top-K 方法,即目标模型输出预测向量不再包括所有的类别预测概率,而是输出最有可能的 K 个类别预测概率。研究表明,当分类的数量较多时,在模型的预测向量中,许多类别在模型的预测向量中出现的概率可能非常小。使用过滤器仅输出最有可能的 K 个类,该模型依然有效。其中 K 值越小,所泄露的隐私信息也就越少。当 K 值为 1 时,模型的输出就是该模型只返回最有可能的类别标签,而不返回其概率信息。该方法在 K 值不小的情况下,虽然没有对目标模型的效用产生很大的影响,但是只能一定程度地抵御 MIAs。即使攻击模型仅得到 K 个输出概率,仍然可以进行有效的攻击。

为了保证模型的效用损失,并实现更好的隐私和效用的权衡,Jia 等[24]受到对抗样本的启发(对抗样本可以愚弄目标模型,也可以愚弄攻击模型),提出 MemGuard 机制。这是第一个在保证效用损失的情况下,降低黑盒 MIA 的攻击精度的防御机制。该机制通过对目标模型的输出概率加扰动噪声,从而实现愚弄攻击者的目的,使攻击者每次的推理攻击都接近于随机猜测,从而保护了目标模型中训练集的隐私信息。具体来说,MemGuard 机制总共分成两个阶段:在第一阶段,MemGuard 找到一个噪声向量并将其加入到真实置信分数向量中,两者形成对抗样本;在第二阶段,MemGuard 将噪声向量以一定的概率添加到置信分数向量中,选择该向量满足置信分数向量上给定的效用损失预算。实验通过 3 个数据集

表明该机制能够有效地防御 MIA，并在现有的防御机制中能够实现更好的隐私和效用的权衡。Yang 等[97]提出了一个防御成员推理攻击和模型反演统一的净化框架，该框架通过减少目标模型预测的置信分数向量的离散度达到净化的目的。此外可以通过对抗性学习进一步防御特定的攻击。

6.4 其他防御策略

6.4.1 在 MLaaS 领域

针对 MLaaS 所面临的预测 API 泄露的问题，Hou 等[98]提出首个通用的预防 API 威胁的框架，该防御机制是在现有的 MLaaS 框架上添加 ML Defense 机制，没有改变目标模型的任何功能，仅仅扩展了 MLaaS 的输出，因此不需要重新训练目标模型和数据集。如图 12 所示，对于授权用户，ML Defense 会通过请求，并把从 MLaaS 云服务器中返回的 $m_t(X_1)$ 经过防御机制变换为 $M_{d_t}(X_1)$；对于非授权用户，ML Defense 对其输入 X_2 进行检查，若发现不符合要求，就会给予拒绝服务。

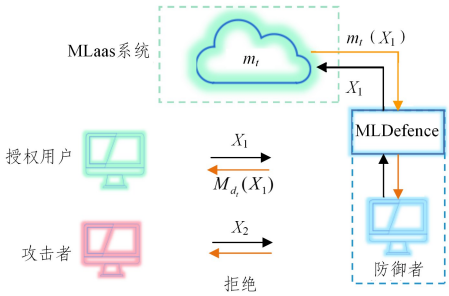


图 12 ML Defense 部署[98]

Fig. 12 Deployment of ML Defense[98]

在 ML Defense 内部，如图 13 所示，其中模拟器是用来模拟和精确地表示未知数据的分布；审计的作用是对于接收到的请求，通过比较两个模拟器给定的数据分布来判断该请求是否安全，若安全则应答，否则拒绝。

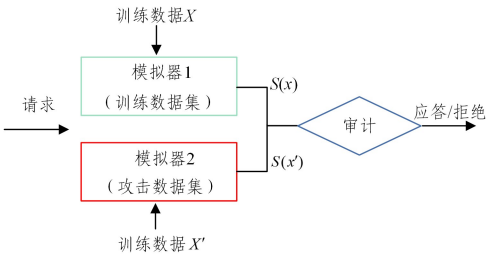


图 13 ML Defense 流程[98]

Fig. 13 Process of ML Defense[98]

对于 MLaaS，用户的输入信息是敏感的，用户不希望将自己的敏感数据上传到服务器，而服务的提供商为了保护知识产权和付费查询，也不想把模型发送给客户共享该模型。为了解决这个问题，文献[99]提出 MLCapsule 方案，这是 MLaaS 的一个受保护的离线部署。如图 14 所示，该方案将机器学习模型部署到本地客户端，以保证用户的数据不离开本地，同时通过 SGX 建立可信执行环境，来保护服务提供商的知识产权。实验结果表明该方案可以提供与传统服务器端 MLaaS 相同水平的模型安全级别和模型控制，同时保护了用户数据隐私。

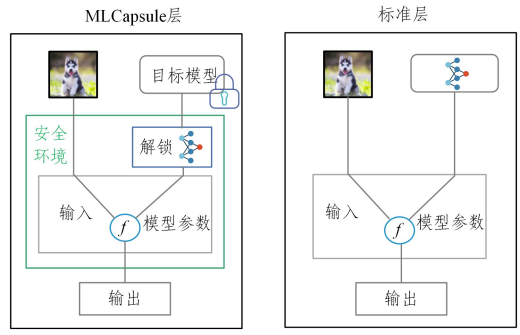


图 14 MLCapsule 层和标准层的区别[99]

Fig. 14 Difference between MLCapsule layer and standard layer[99]

6.4.2 在 GANs 领域

在 GANs 模型防御成员推理攻击中，Wu 等[100]在理论上弥合了隐私保护和 GANs 泛化能力之间的差距，并从理论上和定量上验证了训练数据集的泛化间隙和信息泄露的联系。文献[101]提出改进 GANs 模型的结构，使生成器不仅能够欺骗鉴别器，还能够有效减轻模型对训练集的记忆，从而抵御成员推理攻击。

6.4.3 生成模型抵御 MIAs

将生成模型用于防御成员推理攻击的研究中，Paul 等[102]在视网膜诊断中，基于 GANs 开发隐私防御协议，从而保护医学图像诊断系统免受隐私攻击。Webster 等[103]提出使用由 GANs 生成的图像来建立代理数据集，这些代理数据集能够有效抵御成员推理攻击。Hu 等[104]也提出使用高效的 GANs 生成的样本代替敏感数据来训练目标模型，在训练 GANs 过程中使用截断技术，以提高生成数据的效用，从而既保护了目标模型的准确度，也保护了敏感数据的隐私。此外，Yang 等[105]提出直接对原始数据进行处理，通过 VAEs 生成合成数据，使用合成数据代替原始数据训练目标模型，从而保护原始数据的隐私。Alvar 等[106]将 GANs 和知识蒸馏相结合，并将其应用到图像翻译模型中的防御策略，使用个人生成器作为教师模型，在使用未标记的代理数据集进行对抗性训练中，将知识提取到学生生成器中。Chen 等[107]更改 GANs 模型结构，使其由一个生成器和多个鉴别器组成；通过近似化训练数据的所有分区的混合分布，从而减小 GANs 模型的泛化误差。

6.4.4 迁移学习和其他策略

Papernot 等[108]在迁移学习的基础上提出了一种通用的框架 PATE，以在黑盒情况下保护训练数据的隐私。随后，Papernot 等[109]在此基础上对其进行扩展，使得 PATE 能够应用到具有大量输出类和不平衡的训练数据的学习任务中。Saeidian 等[110]通过研究单个数据的信息泄露问题来研究成员隐私，并证明在 PATE 聚合机制中存在信息泄露问题。文献[111]提出了 DAMIA，通过领域自适应地将不同领域的数据特征映射到同一个特征空间，从而增强目标领域训练。DAMIA 能够有效抵御 MIA，同时具有保护目标模型的效果。

此外，还有学者研究其他防御机制，Tople 等[112]在理论上建立隐私和因果学习模型的联系。首先证实因果学习能够有效防范隐私泄露问题，在增加相同噪声的前提下，因果学习

模型能够提供比关联模型更好的差分隐私保护。Yin 等[113]对 3 个对比模型的优化进行理论分析,并提出 3 种防御方法,实验表明将这 3 种防御方法相结合可以实现隐私和效用的平衡。Lee 等[114]提出在联邦学习场景下抵御成员推理攻击的消化神经网络模型,通过提取原始数据的有效特征,并更改原始数据,最终保证客户端的准确度。Tang 等[115]提出了一个训练隐私保护框架,使得成员数据和非成员数据有相同的表现,从而缓解成员推理攻击。Jarin 等[116]提出了 MIAShield 防御方法,在不影响模型质量的情况下,通过多个不相邻的数据集训练出多个目标模型,在预测时优先排除成员样本的模型。

针对前面所提到的防御机制,表 2 对主要的 MIAs 防御机制进行了总结和分类,分析了 MIAs 防御方法的关键技术以及优缺点。

表 2 主要 MIA 防御方法总结

Table 2 Summary of main MIA defense methods

分类	防御方法	对手知识	文献	关键技术	优点	不足
基于预防模型过拟合策略	L_1,L_2 正则化	黑盒	[6]	在模型损失函数中添加惩罚项	模型收敛较快,能更好地防止过拟合	正则化提高计算的复杂性
	Dropout	黑盒	[9]	训练过程中以一定概率丢失神经元	权重分散,降低了对某些特征的依赖	具有较大的局限性,仅适用于神经网络
	数据增强	黑盒	[72]	增加训练集中的数据量,提升模型鲁棒性	降低了样本的不均衡	可能导致数据小特征的丢失
	提前停止	黑盒	[74]	减少训练迭代的轮次,提前停止模型训练	更为实际的选择,不需要任何调优	不能提供有效的成员隐私保护
	模型堆叠	黑盒	[9]	使用大量子模型结合成最终完整的模型	独立于所使用的 ML 分类器工作	需训练多个不同结构的子模型
	Min-max game	黑盒	[75]	对抗训练算法最小化模型预测损失	获得最优的抵御 MIA 的模型	增加模型训练的时间开销,最大化隐私和效用
	Relaxloss	黑盒/白盒	[76]	使用梯度上升和后验展平	减小了泛化误差,保证模型精度的同时也保证了成员隐私	增加了模型训练的复杂性
基于模型压缩技术	知识蒸馏	黑盒	[77]	从原始模型到蒸馏模型	蒸馏模型能更快地收敛	软标签的生成增加了时间开销
		黑盒	[78]	互补知识蒸馏(CKD)和伪互补知识蒸馏(PCKD)	数据全部来自原始模型,减少了原始模型软标签的生成	蒸馏模型训练的时间开销大
		黑盒/白盒	[79]	蒸馏成员隐私(DMP)	实现更高的效用	数据可能难以满足
	模型剪枝	黑盒	[80]	利用剪枝神经网络降低模型复杂度	一定程度缓解模型受到攻击的可能性	剪枝后的模型依然存在被攻击的风险
		黑盒	[29]	减少基于 KL 散度距离的预测,保护剪枝过程	实现更好的剪枝防御策略	增大了计算开销
基于扰动的防御策略	差分隐私	黑盒	[6]	首次将差分隐私应用到防御 MIA 中	能够提供有效的隐私保证	太多的噪声影响了模型的效用
	干预输出结果	黑盒	[6]	预测向量仅输出 k 个最可能的预测	不影响模型的精度	防御效果不明显
		黑盒	[24]	目标模型输出概率中混入噪声	不影响模型精度,能有效地将攻击精度降低到随机猜测	仅能有效防御模型预测向量的攻击
		黑盒	[97]	净化预测向量的离散度	保证隐私,无效用损失	增加了训练时间开销
其他防御策略	生成模型	黑盒/白盒	[104]	用 GANs 生成的样本代替敏感数据来训练目标模型	同时保证模型的隐私和效用	增加了数据预处理的成本,模型精度依赖于生成图片的质量
	迁移学习	黑盒	[108-109]	通用 PATE 框架并对其进行扩充	模型收敛速度加快	需要准备不同数据集
		黑盒/白盒	[111]	将不同领域数据特征映射到同一个特征空间	能够高效抵御 MIA,开销更小	需要重新训练目标模型

7 未来研究方向

目前 MIAs 正处于发展起步阶段,越来越多的研究者开始关注机器学习隐私保护以及机器学习的安全问题。但当前研究 MIAs 的工作还有很多尚待解决的问题,为此本文归纳了现有的 MIAs 和防御机制的不足,并对未来进行了展望。

(1)现有的 MIAs 大多是针对攻击泛化能力差、过拟合程度高的目标模型,而对于那些泛化能力好、过拟合程度低的模型,MIAs 不能实现较好的攻击性能;而针对模型过拟合程度高的问题,往往可以通过正则化、Dropout 等方法来改善,从而一定程度地限制 MIAs。

(2)对于 MIAs 为什么能够泄露训练数据集的隐私这一问题,目前还没有定论,故未来需要进一步深入探究 MIAs 成功推理的内在机理。

(3)差分隐私防御机制以损失模型预测精度为代价,从而实现对隐私数据的保护,这是目前差分隐私保护数据隐私的最大阻碍,需要从不同角度以及不同领域来权衡隐私和效用,针对具体问题具体分析。因此在未来的工作中,如何在具体的场景中使得差分隐私在实现隐私保护强度的同时保证模型精确度需要进一步深究。

(4)目前基于密码学的技术已被用于机器学习上的隐私保护,如同态加密。这种方法很好地保护了模型的安全和输出的隐私,但需要大量的计算资源开销。还有诸如安全多方计算等技术,也需要进一步探究其适用性。现有的研究较少将密码学相关技术应用到 MIAs 中,因此未来能否将同态加密、安全多方计算以及零知识证明等密码学技术应用到 MIAs 值得进行更为深入的探究。

(5)MIAs 除了具有恶意意图外,还能够被用来检验在

机器学习中的非法数据滥用。有研究指出可以将 MIAs 应用到审计任务中，但对于应用到机器遗忘方面，目前尚未有研究，故未来可以进一步发掘 MIAs 在数据审计方面的应用。

结束语 机器学习的隐私问题在一定程度上制约着人工智能的健康发展，其中成员推理攻击是导致隐私泄露的一种重要的攻击手段。本文首先介绍了成员推理攻击的背景知识，随后对现有的成员推理攻击按照攻击者是否拥有影子模型进行分类，并对成员推理攻击的最新研究成果在不同领域的威胁进行了总结。其次，对现有的机器学习隐私防御机制从防止模型过拟合、模型压缩、基于扰动和基于加密策略等角度进行了分类和总结。最后，本文指出了现有的成员推理攻击和防御机制存在的一些问题和未来可行的研究方向。随着人工智能越来越便利人们的生活，针对机器学习模型的隐私攻击也越来越多，如模型反演攻击、模型窃取攻击等，人们的敏感数据更容易被泄露。未来如何在人们享受人工智能带来智慧生活的同时，保护人们的隐私不受侵犯，将是一个巨大的挑战。

参 考 文 献

[1] CHEN X, CHO Y H, DOU Y, et al. Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data [J]. Journal of Accounting Research, 2022, 60(2): 467-515.

[2] ZHANG Z, YAN C, MALIN B A. Membership Inference Attacks against Synthetic Health Data [J]. Journal of Biomedical Informatics, 2022, 125: 103977.

[3] PYRGELISA, TRONCOSOC, CRISTOFAROED. KnockKnock, Who's There? Membership Inference on Aggregate Location Data[J]. arXiv: 1708.06145, 2017.

[4] TABASSI E, BURNS K, HADJIMICHAEL M, et al. A Taxonomy and Terminology of Adversarial Machine Learning[OL]. https://doi.org/10.6028/NIST.IR.8269-draft.

[5] WACHTER S, MITTELSTADT B, FLORIDI L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation [J]. International Data Privacy Law, 2017, 7(2): 76-99.

[6] SHOKRI R, STRONATI M, SONG C, et al. Membership Inference Attacks against Machine Learning Models [C]// Symposium on Security and Privacy. 2017: 3-18.

[7] LONG Y, BINDSCHAEDLER V, LEI W, et al. Understanding Membership Inferences on Well-Generalized Learning Models [J]. arXiv: 1802.04889, 2018.

[8] IROLLA P, CHTEL G. Demystifying the Membership Inference Attack [C]// Conference on Cybersecurity and Privacy. 2019: 1-7.

[9] SALEM A, ZHANG Y, HUMBERT M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models [C]// Network and Distributed System Security Symposium. 2019: 1-15.

[10] REZAEI S, LIU X. An Efficient Subpopulation-based Membership Inference Attack[J]. arXiv: 2203.02080, 2022.

[11] LIU G, WANG C, PENG K, et al. SocInf: Membership Inference Attacks on Social Media Health Data With Machine Learning [J]. IEEE Transactions on Computational Social Systems, 2019, 6(5): 907-921.

[12] 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, 2019, 14(6): 2073-2089.

[13] SABLAYROLLES A, DOUZE M, OLLIVIER Y, et al. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference [C]// Proceedings of International Conference on Machine Learning. 2019: 5558-5567.

[14] 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]// Symposium on Security and Privacy. 2019: 739-753.

[15] SONG L, SHOKRI R, MITTAL P. Privacy Risks of Securing Machine Learning Models against Adversarial Examples [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

[16] LI G, REZAEI S, LIU X. User-Level Membership Inference Attack against Metric Embedding Learning [J]. arXiv: 2203.02077, 2022.

[17] LIU H, JIA J, QU W, et al. EncoderMI: Membership Inference against Pre-trained Encoders in Contrastive Learning [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2021: 2081-2095.

[18] LI Z, ZHANG Y. Membership Leakage in Label-Only Exposures [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2021: 880-895.

[19] SONG L, MITTAL P. Systematic Evaluation of Privacy Risks of Machine Learning Models [C]// Proceedings of USENIX Security Symposium. 2021: 2615-2632.

[20] RAHIMIAN S, OREKONGDY T. Differential Privacy Defenses and Sampling Attacks for Membership Inference [C]// Proceedings of ACM Workshop on Artificial Intelligence and Security. 2021: 193-202.

[21] HUI B, YANG Y, YUAN H, et al. Practical Blind Membership Inference Attack via Differential Comparisons [C]// Network and Distributed System Security Symposium. 2021: 1-17.

[22] LEINO K, FREDRIKSON M. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference [C]// Proceedings of USENIX Security Symposium. 2020: 1605-1622.

[23] CHOO C C A, TRAMER F, CARLINI N, et al. Label-Only Membership Inference Attacks [C]// Proceedings of International Conference on Machine Learning. 2021: 1964-1967.

[24] JIA J, SALEM A, BACKES M, et al. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples[C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2019: 259-274.

[25] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting [C]// IEEE Computer Security Foundations Symposium. 2018: 268-282.

[26] REZAEI S, SHAFIQ Z, LIU X. Accuracy-Privacy Trade-off in

Deep Ensemble: A Membership Inference Perspective [J]. arXiv:2105.05381,2021.

[27] GROSSO G D,JALALZAI H,PICHLER G,et al. Leveraging Adversarial Examples to Quantify Membership Information Leakage [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2022:1-13.

[28] SHAFRAN A,PELEG S,HOSHEN Y. Membership Inference Attacks are Easier on Difficult Problems [C]// International Conference on Computer Vision. 2021:14820-14829.

[29] YUAN X,ZHANG L. Membership Inference Attacks and Defenses in Neural Network Pruning [C]// Proceedings of USENIX Security Symposium. 2022:4561-4578.

[30] HAYES J,MELIS L,DANEZIS G,et al. LOGAN:Membership Inference Attacks Against Generative Models [J]. Proceedings on Privacy Enhancing Technologies,2019,(1):133-135.

[31] HILPRECHT B,HRTERICH M,BERNAU D. Reconstruction and Membership Inference Attacks against Generative Models [J]. Proceedings Privacy Enhancing Technologies,2019,(4):232-249.

[32] LIU K S,XIAO C,LI B,et al. Performing Co-Membership Attacks Against Deep Generative Models [C]// International Conference on Data Mining. 2019:459-467.

[33] CHEN D,YU N,ZHANG Y,et al. GAN-Leaks:A Taxonomy of Membership Inference Attacks against Generative Models [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2020:343-362.

[34] WEBSTER R,RABIN J,SIMON L,et al. This Person(Probably) Exists. Identity Membership Attacks Against GAN Generated Faces [J]. arXiv:2107.06018,2021.

[35] HU H,PANG J. Membership Inference Attacks against GANs by Leveraging Over-representation Regions [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2021:2387-2389.

[36] CHEN J,ZHANG J,ZHAO Y,et al. Beyond Model-Level Membership Privacy Leakage:an Adversarial Approach in Federated Learning [C]// International Conference on Computer Communications and Networks. 2020:1-9.

[37] MELIS L,SONG C,CRISTOFARO E D,et al. Exploiting Unintended Feature Leakage in Collaborative Learning [C]// IEEE Symposium on Security and Privacy. Piscataway,2019:691-706.

[38] ZHANG J,ZHANG J,CHEN J,et al. GAN Enhanced Membership Inference: A Passive Local Attack in Federated Learning [C]// IEEE International Conference on Communications. 2020:1-6.

[39] HU H,SALCIC Z,SUN L,et al. Source Inference Attacks in Federated Learning [C]// IEEE International Conference on Data Mining. 2021:1102-1107.

[40] PICHLER G,ROMANELLI M,VEGA L R,et al. Perfectly Accurate Membership Inference by a Dishonest Central Server in Federated Learning [J]. arXiv:2203.16463,2022.

[41] CHEN H,LI H,DONG G,et al. Practical Membership Inference Attack Against Collaborative Inference in Industrial IoT [J]. IEEE Transactions on Industrial Informatics,2020,18(1):477-487.

[42] ZHANG G,LIU B,ZHU T,et al. Label-Only Membership Inference Attacks and Defenses in Semantic Segmentation Models [J/OL]. IEEE Transactions on Dependable and Secure Computing. https://ieeexplore.ieee.org/abstract/document/9723588.

[43] HE Y,RAHIMIAN S,SCHIELE B,et al. Segmentations-Leak: Membership Inference Attacks and Defenses in Semantic Image Segmentation [C]// European Conference on Computer Vision. 2020:519-535.

[44] ZHANG M,REN Z,WANG Z,et al. Membership Inference Attacks Against Recommender Systems [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2021:864-879.

[45] WANG Y,HUANG L,YU P S,et al Membership Inference Attacks on Knowledge Graphs [J]. arXiv:2104.08273,2021.

[46] SHAH M A,SZURLEY J,MUELLER M,et al. Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models to Membership Inference Attacks [C]// Interspeech. 2021:891-895.

[47] MIAO Y,XUE M,CHEN C,et al. The Audio Auditor: User-Level Membership Inference in Internet of Things Voice Services [C]// Proceedings on Privacy Enhancing Technologies. 2021:209-228.

[48] OLATUNJI I E,NEJDL W,KHOSLA M. Membership Inference Attack on Graph Neural Networks [J]. arXiv: 2101.06570,2021.

[49] WU B,YANG X,PAN S,et al. Adapting Membership Inference Attacks to GNN for Graph Classification: Approaches and Implications [C]// IEEE International Conference on Data Mining. 2021:1421-1426.

[50] HE X,WEN R,WU Y,et al. Node-Level Membership Inference Attacks Against Graph Neural Networks [J]. arXiv: 2102.05429,2021.

[51] ZHANG Z,CHEN M,BACKES M,et al. Inference Attacks against Graph Neural Networks [C]// Proceedings of USENIX Security Symposium. 2022:1-18.

[52] DUDDU V,BOUTET A,SHEJWALKAR V. Quantifying Privacy Leakage in Graph Embedding [C]// International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. 2020:76-85.

[53] SONG C,RAGHUNATHAN A. Information Leakage in Embedding Models [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2020:377-390.

[54] HISAMOTO S,POST M,DUH K. Membership Inference Attacks on Sequence-to-Sequence Models:Is My Data in Your Machine Translation System? [J]. Transactions of the Association for Computational Linguistics,2020,8:49-63.

[55] YANG Y,GOHARI P,TOPCU U. On The Vulnerability of Recurrent Neural Networks to Membership Inference Attacks[J]. arXiv:2110.03054,2021.

[56] THOMAS A,ADELANI D I,DAVODY A,et al. Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks [C]// International Conference on Text, Speech,and Dialogue. 2020:273-281.

[57] MAHLOUJIFAR S,INAN H A,CHASE M,et al. Membership Inference on Word Embedding and Beyond [J]. arXiv: 2106.11384,2021.

[58] MIRESHGHALLAH F, GOYAL K, UNIYAL A, et al. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks [J]. arXiv: 2203.03929, 2022.

[59] CHEN J, WANG W H, SHI X. Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data [C]// Proceedings of the Pacific Symposium. 2020: 26-37.

[60] BAGMAR A, MAIYA S R, BIDWALKA S, et al. Membership Inference Attacks on Lottery Ticket Networks [J]. arXiv: 2108.03506, 2021

[61] GUPTA U, STRIPELIS D, LAM P K, et al. Membership Inference Attacks on Deep Regression Models for Neuroimaging [C]// Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. 2021: 228-251.

[62] JAGANNATHA A, RAWAT B, YU H. Membership Inference Attack Susceptibility of Clinical Language Models [J]. arXiv: 2104.08305, 2021.

[63] YE J, MADDI A, MURAKONDA S K, et al. Enhanced Membership Inference Attacks against Machine Learning Models [J]. arXiv: 2111.09679, 2021.

[64] ZOU Y, ZHANG Z, BACKES M, et al. Privacy Analysis of Deep Learning in the Wild: Membership Inference Attacks against Transfer Learning [J]. arXiv: 2009.04872, 2020.

[65] CHEN C, WU B, QIU M, et al. A Comprehensive Analysis of Information Leakage in Deep Transfer Learning [J]. arXiv: 2009.01989, 2020.

[66] REZAEI S, LIU X. On the Difficulty of Membership Inference Attacks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2021: 7892-7900.

[67] HINTERSDORF D, STRUPPEK L, KERSTING K. Do Not Trust or Not To Trust Prediction Scores for Membership Inference Attacks [J]. arXiv: 2111.09076, 2021.

[68] LI J, LI N, RIBEIRO B. Membership Inference Attacks and Defenses in Classification Models [C]// Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021: 5-16.

[69] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[70] CARUANA R, LAWRENCE S, GILES C. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping [C]// Neural Information Proceedings Systems. 2000: 402-408.

[71] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network [J]. Computer Science, 2015, 14(7): 38-39.

[72] KAYA Y, DUMITRAS T. When Does Data Augmentation Helpwith Membership Inference Attacks? [C]// International Conference on Machine Learning. 2021: 5345-5355.

[73] YU D, ZHANG H, CHEN W, et al. How Does Data Augmentation Affect Privacy in Machine Learning? [C]// Proceedings of AAAI Conference on Artificial Intelligence. 2021: 10746-10753.

[74] KAYA Y, HONG S, DUMITRAS T. On the Effectiveness of Regularization Against Membership Inference Attacks [J]. arXiv: 2006.05336, 2020.

[75] NASR M, SHOKRI R, HOUMANSADR A. Machine Learning with Membership Privacy using Adversarial Regularization [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2018: 634-646.

[76] CHEN D, YU N, FRITZ M. RelaxLoss: Defending Membership Inference Attacks without Losing Utility [C]// Proceedings of International Conference on Learning Representations. 2022: 1-28.

[77] LIU Y, WEN R, HE X, et al. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models [C]// Proceedings of USENIX Security Symposium. 2021: 4525-4542.

[78] ZHENG J, CAO Y, WANG H. Resisting Membership Inference Attacks through Knowledge Distillation [J]. Neurocomputing, 2021, 452(3): 114-126.

[79] SHEJWALKAR V, HOUMANSADR A. Membership Privacy for Machine Learning Models Through Knowledge Transfer [J]. arXiv: 1906.06589, 2019.

[80] WANG Y, WANG C, WANG Z, et al. Against Membership Inference Attack: Pruning is All You Need [C]// International Joint Conference on Artificial Intelligence. 2021: 3141-3147.

[81] DWORK C, ROTH A. The algorithmic foundations of differential privacy [J]. Foundations Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[82] ABADI M, CHU A, GOODFELLOW I, et al. Deep Learning with Differential Privacy [C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. 2016: 308-318.

[83] RAHMAN M A, RAHMAN T, LAGANIÈRE R, et al. Membership Inference Attack against Differentially Private Deep Learning Model [J]. Transactions on Data Privacy, 2018, 11(1): 61-79.

[84] JAGIELSKI M, ULLMAN J, OPREA A. Auditing Differentially Private Machine Learning: How Private is Private SGD? [C]// Conference on Neural Information Proceedings Systems. 2020: 22205-22216.

[85] TRUEX S, LIU L, GURSOY M E, et al. Effects of Differential Privacy and Data Skewness on Membership Inference Vulnerability[C]// IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications. 2019: 82-91.

[86] NASERI M, HAYES J, CRISTOFARO E D. Local and Central Differential Privacy for Robustness and Privacy in Federated Learning [J]. arXiv: 2009.03561, 2020.

[87] BERNAU D, ROBL J, GRASSAL P W, et al. Comparing Local and Central Differential Privacy Using Membership Inference Attacks[C]// Annual Conference on Data and Applications Security and Privacy. 2021: 22-42.

[88] JAYARAMAN B, EVANS D. Evaluating Differentially Private Machine Learning in Practice [C]// Proceedings of USENIX Security Symposium. 2019: 1895-1912.

[89] ZHANG B, YU R, SUN H, et al. Privacy for All: Demystify Vulnerability Disparity of Differential Privacy against Membership Inference Attack [J]. arXiv: 2001.08855, 2020.

[90] XIONG A, WANG T, LI N, et al. Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension [C]// IEEE Symposium on Security and Privacy. 2020: 392-410.

[91] NASR M，SONG S，THAKURTA A，et al. Adversary Instantiation：Lower Bounds for Differentially Private Machine Learning [C]// IEEE Symposium on Security and Privacy. 2021：866-882.

[92] CHEN Q，XIANG C，XUE M，et al. Differentially Private Data Generative Models [J]. arXiv：1812. 02274，2018.

[93] WUNDERLICH D，BERNAU D，ALDÀ F，et al. On the Privacy-utility Trade-off in Differentially Private Hierarchical Text Classification [J]. arXiv：2103. 02895，2021.

[94] HUMPHRIES T，RAFUSE M，TULLOCH L，et al. Differentially Private Learning Does Not Bound Membership Inference [J]. arXiv：2010. 12112，2020.

[95] MAHLOUJIFAR S，SABLAYROLLES A，CORMODE G，et al. Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms [J]. arXiv：2204. 06106，2022.

[96] LIU B，DING M，SHAHAM S，et al. When Machine Learning Meets Privacy：A Survey and Outlook[J]. ACM Computing Surveys，2021，54(2)：1-36.

[97] YANG Z，SHAO B，XUAN B，et al. Defending Model Inversion and Membership Inference Attacks via Prediction Purification [J]. arXiv：2005. 03915，2020.

[98] HOU J，QIAN J，WANG Y，et al. ML Defense：against Prediction API Threats in Cloud-based Machine Learning Service [C]// Proceedings of the International Symposium on Quality of Service. 2019：1-10.

[99] HANZLIK L，ZHANG Y，GROSSE K，et al. MLCapsule：Guarded Offline Deployment of Machine Learning as a Service [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2021：3300-3309.

[100] WU B，ZHAO S，CHEN C，et al. Generalization in Generative Adversarial Networks：A Novel Perspective from Privacy Protection[C]// Advances in Neural Information Proceedings Systems 32. 2019：1-11.

[101] MUKHERJEE S，XU Y，TRIVEDI A，et al. privGAN：Protecting GANs from Membership Inference Attacks at Low Cost to Utility[J]. Proceedings on Privacy Enhancing Technologies，2021(3)：142-163.

[102] PAUL W，CAO Y，ZHANG M，et al. Defending Medical Image Diagnostics against Privacy Attacks using Generative Methods [J]. arXiv：2103. 03078，2021.

[103] WEBSTER R，RABIN J，SIMON L，et al. Generating Private Data Surrogates for Vision Related Tasks[C]// International Conference on Pattern Recognition. 2021：263-269.

[104] HU L，LI J，LIN G，et al. Defending against Membership Inference Attacks with High Utility by GAN [J/OL]. IEEE Transactions on Dependable and Secure Computing. https：//ieeexplore. ieee. org/document/9773984/authors＃authors.

[105] YANG R，MA J，MIAO Y，et al. Privacy-preserving Generative Framework Against Membership Inference Attacks [J]. arXiv：2202. 05469，2022.

[106] ALVAR S R，WANG L，PEI J，et al. Membership Privacy Protection for Image Translation Models via Adversarial Knowledge Distillation [J]. arXiv：2203. 05212，2022.

[107] CHEN J，WANG W，GAO H，et al. PAR-GAN：Improving the Generalization of Generative Adversarial Networks Against Membership Inference Attacks [C]// Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021：127-137.

[108] PAPERNOT N，ABADI M，ERLINGSSON L，et al. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data [J]. arXiv：1610. 05755，2016.

[109] PAPERNOT N，SONG S，MIRONOV I，et al. Scalable Private Learning with PATE [J]. arXiv：1802. 08908，2018.

[110] SAEIDIAN S，CERVIA G，OECHTERING T J，et al. Quantifying Membership Privacy via Information Leakage [J]. IEEE Transactions on Information Forensics and Security，2021，16：3096-3108.

[111] HUANG H，LUO W，ZENG G，et al. DAMIA：Leveraging Domain Adaptation as a Defense against Membership Inference Attacks [J]. IEEE Transactions on Dependable and Secure Computing，2022，19(5)：3183-3199.

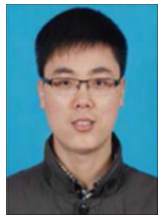
[112] TOPLE S，SHARMA A，NORI A. Alleviating Privacy Attacks via Causal Learning[C]// Proceedings of International Conference on Machine Learning. 2020：9537-9547.

[113] YIN Y，CHEN K，SHOU L，et al. Defending Privacy Against More Knowledgeable Membership Inference Attackers [C]// Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021：2026-2036.

[114] LEE H，KIM J，AHN S，et al. Digestive neural networks：A novel defense strategy against inference attacks in federated learning[J]. Computers & Security，2021，109：102378.

[115] TANG X，MAHLOUJIFAR S，SONG L，et al. Mitigating Membership Inference Attacks by Self-Distillation Through a Novel Ensemble Architecture [J]. arXiv：2110. 08324，2021.

[116] JARIN I，ESHETE B. MIAShield：Defending Membership Inference Attacks via Preemptive Exclusion of Members [J]. arXiv：2203. 00915，2022.



CHEN Depeng，born in 1988，Ph.D，lecturer，is a member of China Computer Federation. His main research interests include information security，machine leaning and IoT security.

CUI Jie，born in 1980，Ph.D，professor，Ph.D supervisor，is a member of China Computer Federation. His main research interests include applied cryptography，IoT security，vehicular ad hoc networks，cloud computing security and so on.

（责任编辑：何杨）