机器学习中成员推理攻击和防御研究综述

牛 俊[1,3], 马骁骥[4,3], 陈 颖[4,3], 张 歌[3,5], 何志鹏[3,6], 侯哲贤[3,7], 朱笑岩[2], 伍高飞[3,7], 陈 恺[8,9], 张玉清[3,4,6,7]

[1]西安电子科技大学计算机科学与技术学院 西安 中国 710071
[2]西安电子科技大学通信工程学院 西安 中国 710071
[3]国家计算机网络入侵防范中心 中国科学院大学 北京 中国 101408
[4]海南大学网络空间安全学院 海口 中国 570228
[5]西安电子科技大学广州研究院 广州 中国 510555
[6]西安邮电大学网络空间安全学院 西安 中国 710121
[7]西安电子科技大学网络与信息安全学院 西安 中国 710126
[8]中国科学院信息工程研究所 信息安全国家重点实验室 北京 中国 100195
[9]中国科学院大学 网络空间安全学院 北京 中国 100195

摘要 机器学习被广泛应用于各个领域, 已成为推动各行业革命的强大动力, 极大促进了人工智能的繁荣与发展。同时, 机器学习模型的训练和预测均需要大量数据, 而这些数据可能包含隐私信息, 导致其隐私安全面临严峻挑战。成员推理攻击主要通过推测一个数据样本是否被用于训练目标模型来破坏数据隐私, 其不仅可以破坏多种机器学习模型(如, 分类模型和生成模型)的数据隐私, 而且其隐私泄露也渗透到图像分类、语音识别、自然语言处理、计算机视觉等领域, 这对机器学习的长远发展产生了极大的安全威胁。因此, 为了提高机器学习模型对成员推理攻击的安全性, 本文从机器学习隐私安全攻防角度, 全面系统性分析和总结了成员推理攻击和防御的基本原理和特点。首先, 介绍了成员推理攻击的定义、威胁模型, 并从攻击原理、攻击场景、背景知识、攻击的目标模型、攻击领域、攻击数据集大小六个方面对成员推理攻击进行分类, 比较不同攻击的优缺点; 然后, 从目标模型的训练数据、模型类型以及模型的过拟合程度三个角度分析成员推理攻击存在原因, 并从差分隐私、正则化、数据增强、模型堆叠、早停、信任分数掩蔽和知识蒸馏七个层面对比分析不同防御措施; 接着, 归纳总结了成员推理攻击和防御常用的评估指标和数据集, 以及其在其他方面的应用。最后, 通过对比分析已有成员推理攻击和防御的优缺点, 对其面临的挑战和未来研究方向进行了展望。

关键词 机器学习; 成员推理攻击; 隐私安全; 防御措施
中图法分类号 TP18 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.11.01

A survey on membership inference attacks and defenses in Machine Learning

NIU Jun[1,3], MA Xiaoji[4,3], CHEN Ying[4,3], ZHANG Ge[3,5], HE Zhipeng[3,6], HOU Zhexian[3,7], ZHU Xiaoyan[2], WU Gaofei[3,7], CHEN Kai[8,9], ZHANG Yuqing[3,4,6,7]

[1] School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[2] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[3]National Computer Network Instrusion Protection Center, University of Chinese Academy of Sciences, Beijing 101408, China
[4]College of Cyberspace Security, Hainan University, Haikou 570228, China
[5]School of Guangzhou Research Institute, Xidian University, Guangzhou 510555, China
[6]School of Cyberspace Security, Xi'an University of Posts & Telecommunications, Xi'an 710121, China
[7]School of Cyber Engineering, Xidian University, Xi'an 710126, China
[8]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China
[9]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100195, China

Abstract The newly emerged machine learning (ML) methods have been widely applied to various applications, and have become a strong driving force to revolutionize a wide range of industries, which have greatly promoted the prosperity and development of artificial intelligence. Meanwhile, the training and inference of the machine learning model are based

通讯作者: 朱笑岩, 博士, 教授, Email:xyzhu@mail.xidian.edu.cn; 张玉清, 博士, 教授, Email:zhangyq@ucas.ac.cn。
本课题得到国家自然科学基金项目(No. U1836210, No. 61772406); 海南省重点研发计划项目(No. ZDYF202012); 陕西省自然科学基础研究计划资助项目(No. 2021JQ-192); 中央高校基本科研业务费专项资金(No. JB211508)资助。
收稿日期: 2022-07-04; 修改日期: 2022-10-06; 定稿日期: 2022-10-08

on a large amount of data, which always contains some private information. And the privacy and security of the ML has faced serious challenges. Membership inference attacks (MIAs) mainly aim to infer whether a data record was used to train a target model or not. MIAs have not only been shown to be effective on various ML models (e.g., classification models and generative models), but also have been penetrated into the fields of image classification, speech recognition, natural language processing, computer vision and so on, which creates a great security threat to the long-term development of machine learning. Therefore, in order to better improve the security of ML models for membership inference attacks, in this paper, we systematically introduce and analyze the basic principles and characteristics of the MIAs and their defenses from a ML attack-defense perspective. Firstly, we introduce the definitions and threat models of the MIAs, and classify these MIAs from six different perspectives such as attacks' principles, scenarios, background knowledge, target models, fields and the size of attack datasets, and we compare their advantages and disadvantages. Secondly, we summary the reasons caused the MIAs from three aspects, namely diversity of training data, types of target models and overfitting of target models. Thirdly, we survey defensive techniques for MIAs as well as their characteristics by differential privacy, regularization, data argumentation, model stacking, early stopping, confidence score masking and knowledge distillation. Futhermore, we institute the evaluation metrics and datasets used in MIAs, and the other applications of the MIAs. Finally, by comparing and analyzing the existing MIAs and their defenses, we discuss the challenges and future research directions.

Key words machine learning; membership inference attacks; privacy & security; defensive techniques

1 引言

近年来, 海量可获得的数据、不断更新的硬件设备、强大的计算设施以及日益完善的智能算法, 极大地推动了人工智能(Artificial intelligence, AI)理论和技术的飞速发展, 促使传统行业的智能化变革。机器学习(Machine learning, ML)作为 AI 技术的一种实现方式, 在各个领域扮演者重要的角色并取得了巨大的成功, 比如图像识别[1], 自然语言处理[2], 图数据应用[3]、脑电路分析[4]、数据挖掘[5]、计算机视觉[6]、电子邮件过滤[7]、检测信用卡欺诈[8]、能源勘探等。虽然, 机器学习促使人们的生活更加方便、快捷和智能, 但其需要大量的数据进行训练, 而这些数据中包含隐私敏感数据, 比如用户文件、位置轨迹等信息。这使得机器学习的安全性、隐私性和公平性也面临着更加严峻的挑战。同时也使机器学习隐私安全问题受到了广泛的关注[9-10]。

众所周知, 机器学习会无意识记住它的训练数据信息, 从而容易遭受各种隐私攻击, 比如模型提取攻击[11], 属性推理攻击(也叫模型逆向攻击)[12], 特征推理攻击[13], 以及成员推理攻击[14]。其中, 成员推理攻击(Membership inference attacks, MIAs)主要推测一个数据样本是否被用来训练目标机器学习模型。这种攻击对个人造成了极大的隐私威胁, 比如: 通过识别某个医疗记录被用来训练一个和特定疾病相关的模型, 可推测某人的健康隐私信息。2019 年的一项报告[15]特别强调: 成员推理攻击是隐蔽的隐私破坏。而且, 成员推理攻击可导致机器学习服务(Machine learning as a service, MLaaS)提供商违反隐私法规。比如, Veale 等人[16]指出成员推理攻击增大了数据被视为隐私个人信息的风险。Homer 等人[17]首次在生物领域提出成员推理攻击的概念, 其主要推测某个特定基因是否在基因数据集中。Shokri 等人[14]首次在机器学习领域提出针对分类模型的成员推理攻击。

目前, 已有很多英文综述研究机器学习中的不同的隐私攻击[18-24], 这些隐私风险综述主要侧重研究机器学习中的隐私安全, 介绍了成员推理攻击的基本概念以及基本的讨论, 并没有进行深入的归纳和总结。另外, 针对成员推理攻击和防御的综述一共有 3 篇[25-27], 其中 2 篇中文综述、1 篇英文综述。表 1

表 1 已有成员推理攻击和防御综述与本文比较

Table 1 Comparisons of existing surveys of MIAs and Defenses with this paper

文献	发表时间	语言		调研数量	攻击定义	威胁模型	攻击分类	原因分析	防御分类	应用	评估指标	数据集
[26]	2019	中文	15	7(攻击) 8(防御)	1 种	黑/白盒	3 个方面	2 种	6 个方面	2 种	×	×
[25]	2022	中文	29	24(攻击) 5(防御)	1 种	黑/白盒	5 个方面	3 种	4 个方面	3 种	×	×
[27]	2022	英文	108	78(攻击) 30(防御)	1 种	黑/白盒	5 个方面	3 种	4 个方面	3 种	7 个	49 个
本文	2022	中文	189	118(攻击) 71(防御)	3 种	黑/灰/白	6 个方面	3 种	7 个方面	3 种	14 个	88 个

(注: ×指综述文献“不包括”这部分内容)

从发表时间、使用语言、调研数量、攻击定义、威胁模型、攻击分类、攻击原因分析、防御分类、攻击应用介绍、评估机制和数据集 11 个方面比较已有的 3 篇成员推理攻击和防御综述与本文工作, 可以看出: 已有 3 篇综述[25-27]的调研文献数量逐渐增多, 但仍缺少对 2021 年 9 月以后成员推理攻击和防御文献的总结和对比研究; 仅给出了基于二元分类器的 MIAs 定义, 缺少对基于评估机制和基于数据集差异的 MIAs 定义; 仅研究了黑盒和白盒两种威胁模型, 缺少对灰盒威胁模型以及开源情况的归纳研究; 仅从攻击原理、攻击场景、背景知识、攻击的目标模型、攻击领域等方面对 MIAs 进行分类, 缺少从攻击数据集大小进行的分类, 而该因素在现实中对攻击者至关重要; 从信任分数处理、正则化、差分隐私和知识蒸馏对 MIAs 防御进行分类和比较, 不够细致和严谨; 缺少对 MIAs 攻击和防御评估指标和新兴领域所用数据集的归纳和总结。

鉴于成员推理攻击发展迅猛, 现有的机器学习中的隐私安全综述[18-24]和成员推理攻击的中英文综述[25-27]并不能对成员推理攻击和防御进行全面的介绍和归纳, 为此我们对成员推理攻击和防御进行了系统性分析和研究, 调研了从 2017 年到 2022 年 6 月的 190 篇左右的相关文献, 跨越了图像分类、语音识别、自然语言处理、计算机视觉等领域, 其中成员推理攻击 118 篇, 成员推理攻击防御 71 篇。这些文献基本是发表在安全、隐私和机器学习领域顶级会议和期刊上的论文, 以及在公共平台上(如 arXiv)上的预发表论文。其中四大安全高水平会议包括: IEEES&P、CCS、USENIX Security 和 NDSS; 人工智能高水平会议包括: ICML、AAAI、IJCAI 和 TPAMI; 计算机视觉高水平会议 CVPR、ECCV 等。表 1 表明: 我们的工作在调研数量、攻击定义、威胁模型、攻击分类、防御分类、评估指标和数据集等方面优于已有的 3 篇成员推理攻击和防御综述。

图 1(a)展示了 2017—2022 年 6 月成员推理攻击和防御相关的研究数据, 可知从 2017—2022 年(6 月)论文发表数量呈上升趋势。图 1(b)和(c)显示了从 2017—2022 年 6 月, 成员推理攻击和防御按攻击场景、背景知识和攻击领域分类后, 分别对应的文献数量。可知, 成员推理攻击从集中式转变为分布式、黑盒演变到白盒、图像分类扩展到文本处理等领域。

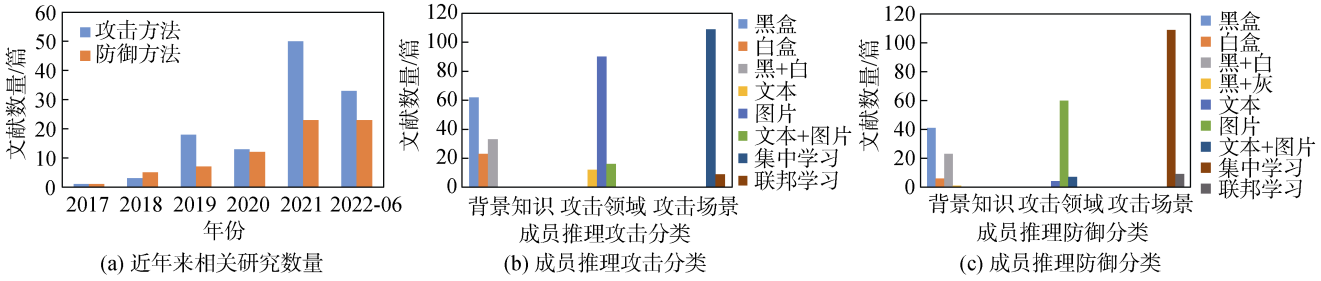

图 1 近年来相关研究数量和成员推理攻击和防御分类

Figure 1 The number of publications in recent years and classifications of the membership inference attacks and defenses

本文主要研究了成员推理攻击的定义、威胁模型、攻击方法、存在原因、防御机制、评估指标、使用数据集、实际应用以及有价值的现象与结论等, 极大地增强了成员推理攻击综述的宽度和深度。主要包括 6 个方面的贡献:

1) 调研了 189 篇成员推理攻击和防御文献, 系统性对成员推理攻击的攻击原理、攻击场景、背景知识、攻击的目标模型、攻击领域和攻击数据集大小进行了梳理、归纳和总结, 并分析了不同成员推理攻击方法的优缺点;

2) 从目标模型的训练数据、模型类型以及模型的过拟合程度 3 个角度, 对成员推理攻击存在原因进行了深入剖析和归纳;

3) 从差分隐私、正则化、数据增强、模型堆叠、早停、信任分数掩蔽和知识蒸馏 7 个维度, 对成员推理攻击防御措施进行了分类和归纳, 并比较了不同防御方法的优缺点;

4) 从图片数据、文本数据、图数据以及二元数据 4 个方面, 对成员推理攻击和防御所使用的数据集进行了归纳和总结; 并统计分析了成员推理攻击和防御的 14 类评估指标, 对比了各自的优缺点;

5) 从隐私审计、知识产权保护和疾病预测 3 个方面, 介绍成员推理攻击被作为隐私审计工具、模型版权保护手段、疾病筛查方法等的现实应用;

6) 深入分析已有成员推理攻击和防御的优缺点, 以及尚未解决的问题和原因, 对其面临的挑战和未

来研究方向进行了展望。

2 成员推理攻击

本小节主要介绍成员推理攻击的定义、攻击原理、威胁模型以及成员推理攻击的分类。

2.1 成员推理攻击的定义

机器学习中的成员推理攻击(MIAs)主要推测一个数据样本是否被用来训练一个目标机器学习模型。一个典型的 MIA 分为三个阶段: 训练、推测和攻击。现有成员推理攻击主要分为三类, 分别为基于二元分类器的 MIAs、基于评估机制的 MIAs 和基于数据集差异的 MIAs, 这三类 MIAs 的训练阶段相同, 主要区别在于推测和攻击阶段(主要由于攻击原理不同)。图 2 展示了基于二元分类器的 MIAs 的具体流程: 训练阶段主要是将目标数据集输入机器学习模型中, 并利用已有的机器学习算法(如, 决策树等)训练得到一个目标模型 Γ_t, 然后该目标模型 Γ_t 被部署在各种机器学习平台上(如谷歌、亚马逊、微软等)用于提供机器学习服务(MLaaS)。用户可通过 API 接口查询目标模型 Γ_t, 获得查询结果, 并向机器学习平台支付相应的费用。预测阶段是指攻击者通过 MLaaS 的 API 接口, 输入一些其认为和目标模型训练集分布相似的数据给目标模型 Γ_t, 获得对应的输出概率, 并利用查询数据和输出概率训练一个二元分类器, 作为最终的攻击模型 Γ_a。 ⊙

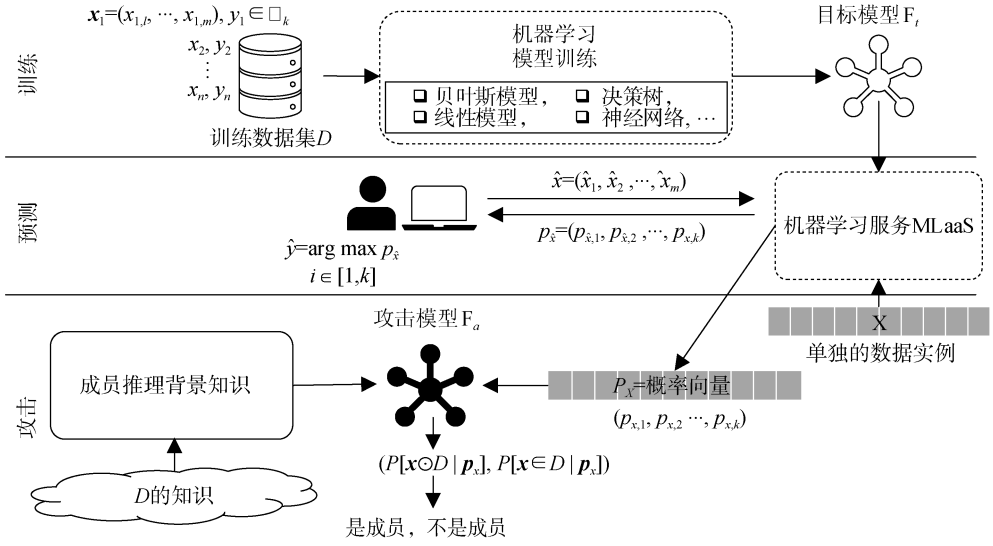

图 2 基于二元分类器的成员推理攻击流程图

Figure 2 The workflow of a Membership Inference Attack based on Binary-classifier

攻击阶段, 攻击者将某个感兴趣的数据点 x_i 输入到目标模型中得到输出概率, 再将输出概率输入攻击模型 Γ_a 中, 如果 Γ_a 输出是“1”, 则认为 x_i 是成员; 否则 x_i 不是成员。

基于评估机制的 MIAs 在推测阶段首先制定一个阈值, 并通过 API 接口将需要攻击的数据输入目标模型 Γ_t, 获得相应的模型输出(比如, 输出概率和输出标签), 攻击者不需要训练攻击模型; 在攻击阶段, 攻击者只需要比较得到的模型输出和预先定义阈值的相对大小, 如果模型输出大于预先定义的阈值, 则认为是成员, 否则认为是非成员。

基于数据集差异的 MIAs 在预测阶段构建两个数据集, 一个是目标攻击数据集, 另一个是非成员数据集, 其中目标数据集中既包含成员样本又包含非成员样本, 而非成员数据集只包含非成员样本; 接着攻击者将这两个数据集中的数据样本输入目标模型 Γ_t 中, 获得对应的模型输出概率, 攻击者不需要训练攻击模型, 需计算这两个数据集模型输出概率间的最大均值差异(Maximum mean discrepancy, MMD); 在攻击阶段, 攻击者从目标攻击数据集中任意移动一个数据样本到非成员数据集, 比较移动该数据样本前后这两个数据集间 MMD 距离的相对大小, 如果移动该数据样本后, 这两个数据集模型输出概率的 MMD 距离变小, 则认为该样本是成员, 否则认为该样本是非成员。

2.2 成员推理攻击的威胁模型

根据攻击者所拥有的背景知识, 我们将成员推理攻击中的威胁模型分为 3 类, 即黑盒、灰盒和白盒威胁模型(参见表 2)。

表 2 威胁模型(敌手知识)

Table 2 Threat models (adversarial knowledge)

威胁模型	数据分布	目标模型		
		输入	输出	架构和参数
黑盒	×	√	√	×
灰盒	√	√	√	×
白盒	√	√	√	√

(注: √和×分别指攻击者需要或不需要该知识实施攻击)

2.2.1 黑盒威胁模型

黑盒威胁模型(Black-box attacks)是指攻击者对目标模型的“训练数据知识”、学习算法、系统架构、学习参数一无所知, 其只能查询机器学习服务(MLaaS)中的目标模型, 并获得相应的预测输出, 其是 3 种攻击威胁中假设最弱的, 且在现实生活中最为常见。

2.2.2 灰盒威胁模型

灰盒威胁模型(Gray-box attacks)是指攻击者对目标模型的学习算法、系统架构、学习参数一无所知, 但其不仅可以查询目标模型, 获得相应的预测输出, 而且可获得和目标模型训练数据集分布相同的数据, 并利用相应的数据增强技术(如, GANs 或 VAEs)生成更多数据, 训练更强的攻击模型。灰盒威胁模型介于白盒和黑盒之间。

2.2.3 白盒威胁模型

白盒威胁模型(White-box attacks)是指攻击者可以获得目标模型的所有信息, 即目标模型的训练集数据分布、训练算法、系统架构、学习参数, 其是三种威胁模型中假设最强的, 也是现实中最不常见的。

2.3 成员推理攻击的分类

本小节, 主要从攻击原理、攻击场景、背景知识、攻击的目标模型、攻击领域和攻击数据集大小 6 个方面对成员推理攻击进行全面细致的分类和介绍, 图 3 表示成员推理攻击发展历程。

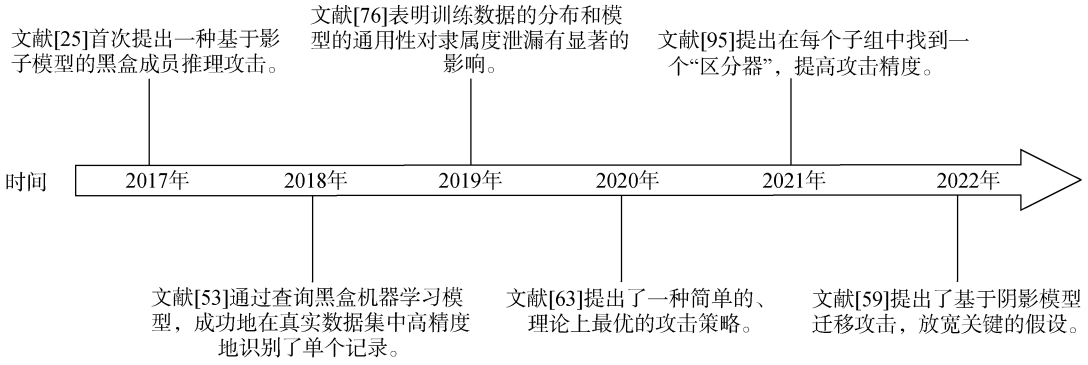

图 3 成员推理攻击发展历程

Figure 3 The development of membership inference attacks

2.3.1 按攻击原理

已有研究[47-50]发现机器学习模型(如, 深度神经网络 DNN)通常是过参数化的, 会记住其训练数据并存储在模型参数中[14,48,51]。根据成员推理攻击的不同攻击原理, 我们将已有成员推理攻击方法分为 3 类: 基于二元分类器的 MIAs、基于评估机制的 MIAs 和基于数据集差异的 MIAs(参见表 3)。

1) 基于二元分类器的攻击:

基于二元分类器的成员推理攻击是指: 攻击者利用一个或多个影子模型, 其主要模仿目标模型的性能, 根据已有数据和影子模型的预测输出训练一个二元分类器作为最终的成员推理攻击模型。

Shokri 等人[14]提出一个有效的影子模型训练技术来训练一个二元分类器作为攻击模型, 采用多个影子模型。与 R. Shokri 等人[14]的方法相似, Long 等人[52]测试一个样本“在”与“不在”训练集中时, 影子模型输出差异, 来区分成员和非成员。Long 等人[53]主要在训练数据集中识别出“易受攻击样本”和“增强样本”, 并分别实施有目标和无目标的成员推理攻击。Salem 等人[54]提出一个更一般的成员推理攻击, 主要放宽了以前攻击的主要条件, 发现只有一个影子模型也可实施有效的成员推理攻击。

同时, Truex 等人[55]提出了一个黑盒成员推理攻击的一般形式, 并联合不同机器学习模型研究模型选择对模型脆弱性的影响。Chen 等人[56]首次研究机器去学习领域中无意识信息泄露问题, 提出一个新的基于版本差异的成员推理攻击方法。Shokri 等人[57]用成员推理攻击研究了基于特征的隐私泄露模型解释。Liu 等人[58]提出一个名为“SocInf”的成员推理攻击方法, 并利用构建的模仿模型对训练和测试数据表现的差异来区分成员和非成员。Chen 等人[59]首次研究工业物联网中的成员推理攻击, 提出一个迁

移遗传影子训练技术, 并放宽了已有的假设。Song 等人[60]研究语言模型中的成员推理攻击, 提出一个新的针对深度学习模型审计技术。Wang 等人[61]首次研究边缘智能中的成员推理攻击, 提出一个针对多等级边缘智能的攻击模型。

基于二元分类器的 MIAs 可在黑盒、灰盒、白盒 3 种情况下实施攻击, 但需要根据影子模型和额外数据训练二元攻击模型, 攻击开销有时较大。

2) 基于评估机制的攻击:

基于评估机制的成员推理攻击是指: 攻击者根据预先定义的成员评估机制来进行成员和非成员判断, 包括: 模型输出阈值、损失阈值、样本标签阈值、交叉熵损失阈值、对抗扰动和假设检验。

① 模型输出阈值:

基于模型输出阈值的成员推理攻击是指: 攻击者根据目标模型的输出信任分数制定一个阈值, 当某个样本的输出信任分数大于该阈值时, 认为该样本是成员, 否则不是成员。

Salem 等人[54]提出了 Global-TopOne 和 Global-TopThree 攻击, 当某个样本的预测输出大于 top1 或 top3 特征阈值时, 认为该样本是成员。Irolla 等人[62]通过理论证明: 信任分数在大多数情况下, 对成功的成员推理攻击起到了很少的表示作用。Bentley 等人[63]研究目标模型的泛化误差如何影响黑盒成员推理攻击的有效性, 并发现泛化误差越大, 成员信息越容易泄露。Sablayrolles 等人[64]基于参数分布的假设来研究成员推理攻击, 并推导出成员推理攻击的最优策略。

② 损失阈值:

基于损失阈值的成员推理攻击是指: 攻击者先根据目标模型对成员的输出信任分数计算成员的平均损失, 并制定一个阈值, 当某个样本的输出信任分数损失小于该阈值时, 认为该样本是成员, 否则不是成员。

Ye 等人[36]提出一个基于蒸馏的损失阈值攻击, AUC 面积可达 87.6%。Sablayrolles 等人[64]基于参数分布的假设来研究成员推理攻击, 发现最理想的攻击仅依赖于损失函数。Yeom 等人[65]在白盒场景下研究当一个训练样本的模型损失小于某个阈值(训练集的平均损失)时, 认为该样本是成员。

③ 样本标签阈值:

基于样本标签阈值的成员推理攻击是指: 攻击者根据目标模型的输出标签来识别成员, 当某个样本的预测标签和真实标签一致时, 认为该样本是成员, 否则不是成员。

Yeom 等人[65]提出一种基于目标模型预测标签的成员推理攻击方法, 当样本的预测标签和其 ground-truth 标签一致时, 则认为其是成员。Choquette 等人[66]通过评估模型对扰动后输入数据预测标签的鲁棒性来推测成员关系。Li 等人[67]提出了基于决策的成员推理攻击——基于迁移和基于边界的攻击, 并研究多个防御机制。Rahimian 等人[68]提出一种基于标签的样本攻击, 攻击成功率很高(如 100%)。

④ 交叉熵损失阈值:

基于交叉熵损失阈值的成员推理攻击是指: 攻击者将已有数据输入目标模型得到预测的信任分数, 计算这些信任分数的交叉熵, 并制定一个交叉熵阈值, 当某个样本的交叉熵小于该阈值时, 认为该样本是成员, 否则不是成员。Salem 等人[54]提出了一种基于交叉熵损失阈值的成员推理攻击, 当某个样本的交叉熵小于制定的阈值时认为其是成员。Song 等人[69]也提出了一个基于预测熵变化的新的成员推理攻击方法, 以及攻击风险评估准则。

⑤ 对抗扰动:

基于对抗扰动的成员推理攻击是指: 攻击者给样本添加扰动使得目标模型对该样本的预测标签发生变化, 并利用添加扰动的大小来识别成员和非成员。Choquette 等人[66]通过评估模型对扰动后输入数据预测标签的鲁棒性来推测成员关系。Li 等人[67]提出一种基于决策的成员推理攻击, 主要通过给样本添加的对抗扰动大小来判断成员和非成员。

⑥ 假设检验:

基于假设检验的成员推理攻击是指: 攻击者首先假设“成员条件”和“非成员条件”, 当某个样本的“成员条件”的概率大于“非成员条件”的概率, 认为该样本是成员, 否则不是成员。Long 等人[52]提出了一个基于假设检验的实际成员推理攻击方法, 当样本假设检验的值小于一个切断的阈值时, 认为该样本是成员。同时, Long 等人[53]根据成员样本对模型独特的影响, 在黑盒场景下提出基于假设检验的成员推理攻击方法。

基于评估机制的 MIAs 不需要训练攻击模型, 节省攻击成本, 但阈值选择有时花费时间较多。

3) 基于数据集差异性比较的攻击:

基于数据集差异性比较的成员推理攻击是指: 攻击者利用在两个数据集中任意移动一个样本后, 两数据集移动前后的距离差异进行成员和非成员的识别。Hui 等人[70]利用差异比较提出一种实际盲成员

表 3 根据攻击原理对成员推理攻击方法进行分类

Table 3 Classifications of MIAs based on attacks' principles

攻击原理	英文缩写	威胁模型	原理描述	分类	优点	缺点	参考文献
基于二元分类器	Bi-nary-classifier based MIAs	黑/灰/白盒	利用一个或多个影子模型的预测输出	单个影子模型	攻击成本低	可能遗漏某些有效特征, 造成攻击不准确	[54]
				多个影子模型	攻击较准确	攻击成本高	[14, 52-53, 55-61]
基于评估机制	Metrics-based MIAs	黑/灰/白盒	预先定义的成员评估机制	模型输出阈值	可直接获得, 不需训练攻击模型	对抗扰动会影响攻击效果	[54, 62-64]
				损失阈值	不需要训练攻击模型	很难选择合理的损失阈值	[36,64-65]
				样本标签阈值	可直接获得, 不需训练攻击模型	攻击效果依赖标签的好坏	[65-68]
				交叉熵损失阈值	不需要训练攻击模型	很难选择合理交叉熵阈值	[54, 69]
				对抗扰动	不需要训练攻击模型	需要寻找合适的扰动大小	[66-67]
				假设检验	不需要训练攻击模型	需要设置合理的假设条件	[52-53]
基于数据集差异性比较	Differential compara-tions-based MIAs	黑/灰/白盒	移动前后两数据集距离差异	盲数据集差异 (BlindMI-DIFF)	不需要影子模型和模型预测输出	需要构造非成员数据集	[70]

推理攻击—BLINDMI, 并通过一个新的差异分布方法来提取成员语义信息。

2.3.2 按攻击场景

根据成员推理攻击实施的不同场景——集中式和分布式场景(参见图 4), 我们将已有成员推理攻击方法分为 2 类: 集中式和分布式成员推理攻击(参见表 4)。

1) 集中式成员推理攻击(centralized MIAs):

集中式机器学习是指将所有用户或设备的信息收集到一个服务器上, 并利用机器学习算法进行训练得到对应的目标模型。

Jayaraman 等人[71]提出一种现实假设中的成员推理攻击——Merlin, 利用基于正例预测值联合成员增益的机制来评估隐私信息泄露情况。Hilprecht 等人[72]针对现有生成网络提出蒙特卡洛攻击和重构攻击, 分别在黑盒和白盒场景下对 GANs 和 VAEs 网络进行攻击。同时, Hayes 等人[73]研究生成网络生成的样本如何泄露一般模型或过拟合模型的隐私信息。Chen 等人[74]研究深度生成模型中的成员推理攻击, 首次提出一个一般性的攻击模型——利用 Parzen 窗口密度估计近似计算一个样本由受害生成器生成的概率, 并根据概率的大小进行成员和非成员的判断。Leino 等人[75]发掘模型内部特征和运作规律, 捕捉更有效的信息作为“成员证据”。

集中式 MIAs 将数据收集起来训练模型, 不需要上传训练梯度和参数, 但需要上传数据, 增加隐私泄露风险, 且对于某些不能公开收集的数据, 无法完成模型训练。

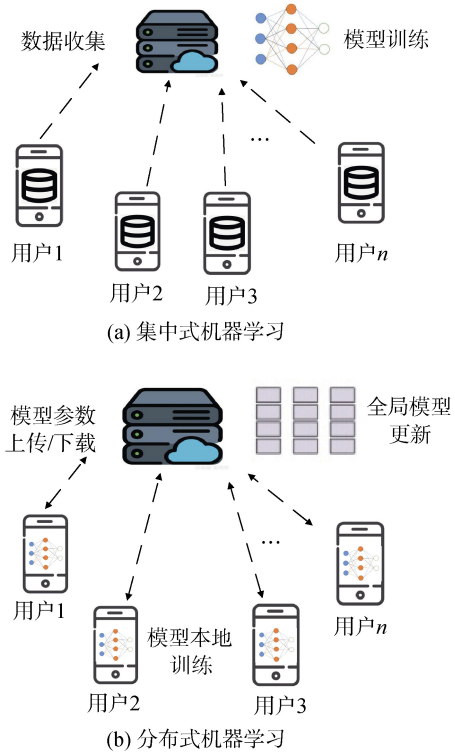

图 4 集中式机器学习和分布式机器学习

Figure 4 Examples of centralized and federated machine learning

表 4 根据攻击场景对成员推理攻击进行分类

Table 4 Classifications of MIAs based on attacks' scenarios

攻击场景	英文缩写	威胁模型	场景描述	优点	缺点	参考文献
集中式	Centralized MIAs	黑/灰/白盒	所有数据收集到一起, 集中训练一个目标模型	不需要上传训练梯度和参数	需要上传数据, 增加隐私泄露风险	[14, 52-54, 56, 64-67, 69-75]
分布式	Federated MIAs	黑/灰/白盒	多个实体以一种协作的方式共同训练一个目标模型, 其不共享数据, 只共享训练参数和梯度	不需要上传数据, 降低隐私泄露风险, 打破数据孤岛	存在恶意参与者或参数服务器, 增加隐私泄露风险	[76-82]

2) 分布式成员推理攻击(federated MIAs):

分布式机器学习是指各个用户或设备不需要上传自身数据, 仅从参数服务器上下载模型架构, 并在本地利用机器学习算法进行训练, 只将训练的梯度和参数上传给参数服务器, 参数服务器收集整合所有用户上传的梯度和参数, 更新全局模型。Nasr 等人[76]提出一个联邦学习场景中基于梯度信息的成员推理攻击, 发现最后一层梯度信息的攻击效果好于利用泛化误差实施的成员推理攻击。Zhang 等人[77]提出一种针对联邦学习的被动成员推理攻击, 主要利用生成对抗网络来增加数据多样性, 攻击成功率可达 98%。Chen 等人[78]提出一个用户层面的联邦学习黑盒成员推理攻击, 并采用生成对抗网络来增加数据量。Hu 等人[79]提出了一个新的名为源推理攻击的联邦学习成员推理攻击方法, 可获得对训练成员源头的理想化评估。

同时, Melis 等人[80]研究联邦学习中无意识隐私信息泄露问题, 并且利用这些信息可以发起被动或者主动的成员推理攻击。Wang 等人[81]研究联邦场景下由恶意服务器发起的成员推理攻击, 在服务器端将 GAN 网络和多功能鉴别器相结合实施攻击。Gupta 等人[82]研究深度回归模型用于神经影像中的成员推理攻击, 发现上传的训练参数仍可泄露数据隐私。

分布式 MIAs 不需要上传数据, 降低隐私泄露风险, 打破数据孤岛, 但存在恶意参与者或参数服务器, 增加隐私泄露风险。

2.3.3 按攻击者的背景知识

根据攻击者所拥有的背景知识——“训练数据知识”和“目标模型知识”, 我们将成员推理攻击分为黑盒、灰盒和白盒成员推理攻击(参见表 5)。

1) 黑盒成员推理攻击:

黑盒成员推理攻击是指攻击者对目标模型的“训练数据知识”、学习算法、系统架构、学习参数一无所知, 只知道目标的预测输出。已有的大部分攻击都是黑盒成员推理攻击, 如基于二元分类器的 MIAs[12,52-58]、基于评估机制的 MIAs[64-69]、基于差异性比较的 MIAs[70], 以及集中式 MIAs[71], 其具体内容已在 2.3.1 和 2.3.2 节详细介绍过, 在此不再赘述。Kulynych 等人[83]深入分析了黑盒成员推理攻击不同的脆弱性, 并提出一个满意的框架来解决现实中的这些问题。黑盒 MIAs 需要的背景知识最少, 在现实中最常见, 但仅依赖目标模型和影子模型输出易导致攻击效果有限。

2) 灰盒成员推理攻击:

灰盒成员推理攻击是指攻击者对目标模型的学习算法、系统架构、学习参数一无所知, 但是攻击者可以获得目标模型的预测输出和与目标模型训练数据分布相同的数据。Hui 等人[70]基于差异比较提出一种实际盲成员推理攻击—BLINDMI-DIFF, 其在所有场景下通过一个新的差异分布方法来提取成员语义信息。灰盒 MIAs 需要适中的背景知识且攻击效果较黑盒有所提高, 但现实中有时无法获得与训练数据分布相同的数据, 导致攻击效果提升有限。

3) 白盒成员推理攻击:

白盒成员推理攻击是指攻击者可以获得目标模型的所有信息, 即目标模型的训练集数据分布、训练算法、系统架构、学习参数。2.3.2 节介绍的集中式 MIAs[72-75]和分布式 MIAs[76,80]都属于白盒成员推理攻击, 在此不再赘述。

Rezaei 等人[84]在白盒场景下, 研究深度模型中成员推理攻击的不可行性。Carlini 等人[85]研究基于深度学习的生成序列模型中的白盒成员推理攻击。Sablayrolles 等人[86]在白盒场景下, 提出针对成员推理攻击的贝叶斯优化策略。Song 等人[87]研究深度学习中, 白盒和黑盒的成员推理攻击和数据版权保护问题。Ha 等人[109]提出成员特征分解网络, 从数据特征的角度来研究成员推理。Gu 等人[110]设计了联邦学习中局部和全局攻击推理算法。Zhang 等人[111]提出了一种基于对抗鲁棒性的成员推理攻击增强方法。Pichler 等人[112]研究了一个不诚实的中央服务器的框架。Watson 等人[113]提出成员推理攻击可以从难度校准中获得巨大好处。Hu 等人[139]提出了一种基于成员推理的过度代表性攻击。Rezaei 等人[146]提出已有成员推理攻击性能报告具有误导性。白盒 MIAs 攻击效果最好, 但需要的背景知识最多, 现实中很难实现。

表 5 根据攻击者的背景知识对成员推理攻击进行分类

Table 5 Classifications of MIAs based on attacks' scenarios

背景知识	英文缩写	拥有知识	攻击力度	优点	缺点	参考文献	开源和商业化
黑盒	Black-box MIAs	目标模型预测输出	最弱	最少的背景知识	攻击效果有限	[14,53,55,57,58,64-65,68,71,83] [52,54,56,66-67,69-70] [69]	自己训练 开源 商业化
灰盒	Gray-box MIAs	目标模型预测输出和训练数据分布	适中	适中的背景知识	攻击效果提升有限(相比黑盒)	[70]	开源
白盒	White-box MIAs	目标模型预测输出、学习算法、系统架构、参数及训练数据分布	最强	攻击效果最高	背景知识最多	[72-73,75-76,80,85-87,109-110,139] [74,84,111-113]	自己训练 开源

2.3.4 按攻击的目标模型

根据攻击者所攻击的目标模型, 我们将成员推理攻击分为针对分类模型的 MIAs、针对回归模型的 MIAs、针对嵌入模型的 MIAs、针对生成模型的 MIAs、针对临床语言模型的 MIAs 和针对彩票网络的 MIAs(参见表 6)。

1) 分类模型的 MIAs:

针对分类模型(classification models)的 MIAs 包括针对二分类模型的 MIAs 和针对多分类模型的 MIAs。

① 针对二分类模型 MIAs:

针对二分类模型的 MIAs 是指攻击者攻击的目标模型是二分类模型, 主要判断某个样本是否是该二分类模型训练集中的成员。已有的部分 MIAs 针对二分类模型, 如基于二元分类器的 MIAs[14,52-57]、基于差异性比较的 MIAs[70]、集中式 MIAs[75]和黑盒 MIAs[83], 其具体内容已在 2.3.1、2.3.2 和 2.3.3 节介绍过, 在此不再赘述。Humphries 等人[88]在理论和实验上研究了差分隐私在什么情况下会遭受成员推理攻击。

② 针对多分类模型 MIAs:

针对多分类模型的 MIAs 是指攻击者攻击的目标模型是多分类模型, 主要判断某个样本是否是该多分类模型训练集中的成员。已有的部分 MIAs 针对多分类模型, 如基于二元分类器的 MIAs[14,52-57]、基于评估机制的 MIAs[64-69]、基于差异性比较的 MIAs[70]、集中式 MIAs[71,75]和分布式 MIAs[76,80], 黑盒 MIAs[83]和白盒 MIAs[84], 其具体内容已在 2.3.1、2.3.2 和 2.3.3 节介绍过, 在此不再赘述。

Song 等人[89]首次联合研究机器学习服务中隐私性和安全性的关系。Truex 等人[90]提出了一个成员推理攻击一般的表示形式, 研究模型在什么样的条件下易遭受黑盒成员推理攻击。Rahman 等人[91]系统性研究了差分隐私模型中的成员推理攻击。Li 等人[92]提出模型泛化误差与成员推理攻击脆弱性间的数字关系。Kaya 等人[93]研究了正则化在抵御成员推理攻击中的有效性, 并给出正则化的下界。Liu 等人[94]首次提出一个机器学习中推理攻击的整体风险评估方法——ML-DOCTOR。Chang 等人[95]研究了算法的公平性如何影响训练数据隐私泄露, 其从成员推理攻击角度分析组公平性的隐私风险。He 等人[96]研究图神经网络中(graph neural networks, GNNs)基于节点的成员推理攻击。Iyiola 等人[97]也研究了图神经网络中的成员推理攻击问题。

针对分类模型的 MIAs 可在不同情况下进行攻击, 且攻击效果较好(比如, 攻击精确率 85%, 召回率接近 100%), 但需要训练多个影子模型, 使得攻击模型训练开销较大。

2) 回归模型的 MIAs:

针对回归模型的 MIAs 是指: 攻击者预测一个样本是否被用来训练一个回归模型(比如, 深度回归模型)。Tan 等人[33]提出一个过度参数化的未被充分探索的隐藏代价。Gupta 等人[98]首次提出深度回归模型中的成员推理攻击, 结合参数梯度、激活函数、模型预测和样本标签的构造攻击模型。

3) 嵌入模型的 MIAs:

嵌入模型是指将原始的目标(比如, 文字、句子和图)映射成实值向量的数学函数, 主要为了捕捉和保存原始目标的重要语义信息。针对嵌入模型的 MIAs 是指: 攻击者推测一个样本是否在嵌入模型的训练数据集中。Song 和 Raghunathan [103]首次提出单词和句子嵌入模型中的成员推理攻击, 主要推测一个文字滑动窗口或一对句子是否在嵌入模型的训练数据集中。Mahloujifar 等人[104]表明即使嵌入层和嵌入模型不暴露给攻击者, 仍会遭受成员推理攻击。

表 6　根据攻击的目标模型对成员推理攻击进行分类

Table 6　Classifications of MIAs based on target models

攻击的目标模型	英文缩写	威胁模型	攻击原理	分类	优点	缺点	参考文献
分类模型	Classification models	黑/灰/白	推测一个数据样本是否是成员	二分类	背景知识较少	仅对特定类进行攻击	[14,52-57,70,75,83,88]
				多分类	可同时攻击多个类	训练攻击模型开销大	[12,52-57,64-71,75-76,83-84,89-97]
回归模型	Regression models	白盒	推测一个数据样本是否是成员	深度回归模型	白盒时攻击成功率较高	白盒假设现实中很难实现	[28,33,98]
嵌入模型	Embedding models	黑/白	推测一个文字滑动窗口或一对句子是否是成员	单词句子嵌入模型	不管嵌入层是否暴露都可进行攻击	只进行一个滑动窗口文字或一对句子的推测, 无法进行单个单词的推测	[30,103-104,106]
				图嵌入模型	在黑盒和白盒时都可实现攻击	图嵌入模型推测的精确性仍需提高	[105]
				图片嵌入	用途较广	训练成本高	[107]
生成模型	Generative models	黑/白	鉴别器对训练成员的输出信任分数更高	针对 GANs	仅利用鉴别器的输出就可实现攻击	成员和相似样本间的距离阈值很难选择	[29,72-74,108,114-116]
				针对 VAEs	利用 VAEs 对成员的重构损失进行推测	仅仅实现对单个样本的推测, 无法实现对多个样本推测	[72,74,108]
临床语言模型	Clinical Language Models	黑/白	推测一个数据样本是否是成员	针对 BERT 和 GPT2	分别提出两种黑盒和白盒攻击, 隐私泄露达 7%	该一般的攻击方法, 无法对特定模型和数据实施有效攻击	[119]
彩票网络	Lottery Ticket Networks	黑盒	推测一个数据样本是否是成员	针对剪枝网络	利用影子训练技术提出可迁移的攻击	攻击不通用, 未涉及其他模型(比如 Grasp 和 SNIP)	[168]

Duddu 等人[105]首次研究图嵌入模型中的成员推理攻击, 提出一个黑盒影子模型攻击以及一个基于信任分数的白盒成员推理攻击。Klakow 等人[106]研究了在各种预训练词嵌入模型(如 GloVe、ELMo 和 BERT)上的隐私泄露问题。Liu 等人[107]提出了第一个在对比学习上预训练的图片嵌入器中的成员推理攻击(EncoderMI 方法)。

针对嵌入模型的 MIAs 可适用单词嵌入、图嵌入、图片嵌入等不同情况, 攻击准确率可达 83.04%, 但假设攻击者的背景知识较强, 现实中实现困难且攻击模型训练成本较高。

4) 生成模型的 MIAs:

针对生成模型的 MIAs 主要是指: 攻击者推测一个样本是否被用于训练一个生成模型, 包括: 生成对抗网络(GANs)和变分自动编码器(VAEs)。2.3.2 节介绍的集中式成员推理攻击[72-74]主要是针对生成模型。

Liu 等人[108]提出一个共同的成员推理方法, 其对于不同的输入数据需要重新训练新的神经网络, 而集合成员推理攻击方法[73]仅利用生成器合成的固定数据。文献[72,74,108]统一表明 VAEs 比 GANs 更易遭受成员推理攻击。Wu 等人[114]提出各种针对成员推理攻击方法来研究泛化好的 GANs 的隐私泄露情况。Mukherjee 等人[115]提出一个新的生成对抗网络架构 privGAN, 其生成器不仅要欺骗鉴别器, 而且还能抵御 MIAs。Webster 等人[116]通过构建一个新的成功的成员推理, 来挑战 GANs 生成的图片是新创造的假设。

针对生成模型的 MIAs 可攻击不同生成模型(如 GANs 和 VAEs), 但很难选择合适的攻击阈值(如, 距离阈值), 无法对多样本实施攻击。

5) 临床语言模型的 MIAs:

针对临床语言模型的 MIAs 主要是指: 攻击者推测一个样本是否被用于训练一个临床语言模型。Jagannatha 等人[119]提出一个针对临床语言模型的 MIA, 主要针对针对 BERT 和 GPT2 模型, 分别提出两种黑盒和白盒攻击, 隐私泄露达 7%, 但对特定模型和数据无法实施更准确攻击。

6) 彩票网络的 MIAs:

针对彩票网络的 MIAs 主要是指: 攻击者推测一

个样本是否被用于训练一个彩票网络, 彩票网络是指利用剪枝技术得到一个神经网络的子网络。Bagmar 等人[168]利用影子训练技术提出一个针对彩票网络的成员推理攻击, 且不同网络的攻击具有可迁移性, 但其只适用于 ResNet18 和 ResNet50, 对 Grasp 和 SNIP 模型无法实施攻击。

2.3.5 按攻击的不同领域

根据成员推理攻击涉及的不同领域, 我们将其划分为图像分类、自然语言处理、计算机视觉、音频、推荐系统、迁移学习、对比学习、图神经网络、在线学习、机器去学习、医疗场景、工业物联网和边缘智能等领域(参见表 7)。

1) 图像分类:

针对图像分类的 MIAs 是指攻击者判断图像分类领域的 MIAs。已有大部分成员推理攻击针对图像分类, 如基于二元分类器的 MIAs[14,52-57]、基于差异

表 7 根据攻击的不同领域对成员推理攻击进行分类

Table 7 Classifications of MIAs based on different domains

攻击的领域	英文缩写	威胁模型	攻击原理	分类	优点	缺点	参考文献
图像分类	Image Classifications	黑/灰/白	推测一个数据样本是否是成员	二分类和多分类	对多类别实施攻击, 且新提出的基于阈值的评估机制可有效评估攻击	背景知识较多, 综合的评估指标无法全面评估攻击效果, 阈值指标无法评估所有样本	[14,28,31-32,34,40-44,52-57,64-71,75-76,80,84,90-94,99-101,135]
自然语言处理	Natural language processing	黑/白	推测一个文字滑动窗口或一对句子是否是成员	文本分类	攻击场景多样, 攻击准确率和精确率较高	攻击成本较高, 模仿模型对攻击影响较大	[28,39,41,58,80,102,117]
				文本生成	攻击精确性较高, 可迁移到其他模型	攻击成本较高, 评估指标单一	[60,118]
				词嵌入	攻击模型多样, 攻击可适应分类和生成任务	背景知识较多, 需要对数据进行预处理	[103-104,106,119-120]
计算机视觉	Computer Vision	黑/白	推测一个数据样本是否是成员	图像生成	攻击方法多样, 可攻击多个模型	背景知识较多, 需要额外训练神经网络	[72-74,108,114-115]
				语义分割	背景知识较少, 攻击效果较高(如, AUC 面积为 87.1%)	攻击方法少, 需要选择合适的成员特征	[121-122]
音频	Audio	黑	推测一个语音信息是否是成员	/	攻击精确率较高(75%), 在较少语音信息下仍可实现有效攻击	仅有黑盒攻击, 需要探索攻击性更强的攻击	[123-124]
推荐系统	Recommender system	黑/白	推测一个用户是否是成员	/	攻击效果较高(AUC 面积 99.8%), 提出一个用户级的 MIA	攻击方法单一, 需要生成标签数据和训练影子模型	[125]
迁移学习	Transfer learning	黑/白	推测一个数据是否属于教师模型的成员	/	攻击场景多样, 对深度迁移模型进行分类并分别实施攻击	背景知识较多, 攻击的适用领域有限(如, 人脸识别)	[126,129]
对比学习	Contrastive learning	黑/白	推测一个数据样本是否是成员	/	背景知识较少, 攻击准确性较高(72.6%)	攻击方法单一, 综合的评估指标很难反应真实攻击	[127]

续表

攻击的领域	英文缩写	威胁模型	攻击原理	分类	优点	缺点	参考文献
				知识图谱	背景知识较少, 三类攻击可实现医疗和金融知识图谱的攻击	采用的评估指标无法有效反应攻击效果	[128]
图神经网络	Graph neural networks (GNNs)	黑/白	推测一个数据样本是否是成员	节点分类	攻击场景多样, 提出一个三节点级攻击	背景知识较多, 仅采用精确性无法有效评估攻击效果	[96-97,105]
				图分类	提出两类攻击且 f1 分数较高(70%)	综合的评估指标会造成评估结果不准确	[130]
在线学习	Online learning	黑/白	推测一个数据样本是否是成员	/	背景知识较少, 提出单个和多个样本攻击	攻击方法单一, 需要构造编码器和训练影子模型	[131]
机器去学习	Machine unlearning	黑/白	推测被去掉的数据是否是成员	/	提出两种机器去学习中的评估指标	攻击方法单一, 需生成后验概率, 并重构特征	[132]
医疗场景	Medical data	黑/白	推测一个数据样本是否是成员	/	攻击效果较好, 可重构医疗图片	攻击方法单一, 需训练攻击模型	[133]
工业物联网	Industrial IoT	黑盒	推测一个数据样本是否是成员	/	提出一个迁移遗传影子训练技术	采用综合的评估指标不能对攻击实现有效评估	[59]
边缘智能	Edge Intelligence	黑盒	推测一个数据样本是否是成员	/	首次提出一个针对多等级边缘智能的攻击	准确性指标不能很好表示攻击效果	[61]

性比较的 MIAs[70]、基于评估机制的 MIAs[64-69]、集中式 MIAs[71]、白盒 MIAs[75-76,80,84]、分类模型 MIAs[90-94], 其具体内容已在2.3.1、2.3.2、2.3.3 和2.3.4 节详细介绍过, 在此不再赘述。

Jagielski 等人[28]提出利用一个或多个模型更新的新成员推理攻击。Zhang 等人[31]提出语义分割模型中单标签隶属度推断攻击。Rezaei 等人[32]提出一种利用目标模型对语义相似样本输出差异的攻击。Yuan 等人[34]提出了一种剪枝神经网络的自注意成员推断攻击。GMR 等人[40]研究隶属度推理攻击对传统机器学习算法的影响。Li 等人[42]提出了一个用户成员推理攻击。Del 等人[43]提出了基于训练模型识别数据隐私风险的标准方法。Pedersen 等人[44]提出攻击者可通过简单攻击策略达到隐私损失的下限。Long 等人[52]研究实际场景中的成员推理攻击, 并设计了一个新的可在样本个体层面而不是聚集的训练集层面的攻击方法。Jayaraman 等人[71]发现均衡的先验概率在实际中不可行, 并提出一个新的成员推理攻击 Merlin。Carlini 等人[99]提出了一个似然比攻击。Mahloujifar 等人[100]提出了攻击者进行成员推理攻击的增益边界。Duddu 等人[101]提出了 SHAPr 指标, 用来量化模型对单个训练数据的记忆。

针对图像分类的 MIAs 可对多类别实施攻击, 且新提出的基于阈值的评估机制可有效评估攻击, 但背景知识较多, 综合的评估指标无法全面评估攻击效果, 基于阈值的评估指标无法评估所有样本。

2) 自然语言处理:

针对自然语言处理的 MIAs 是指攻击者判断自然语言处理领域的 MIAs。根据不同的任务, 其可分为三类: 文本分类中的 MIAs, 文本生成中的 MIAs 和词嵌入中的 MIAs。

① 文本分类中的 MIAs:

针对文本分类中的 MIAs 是指攻击者判断某个分类文本数据是否是训练集成员。Zhong 等人[39]提出一种新的隐私泄露差异符号, 其量化了不同子组间的 MIAs。Li 等人[41]提出了黑盒成员推理攻击 l-Leaks。Liu 等人[58]提出一个名为“SocInf”的黑盒成员推理攻击方法。Melis 等人[80]研究联邦学习中无意识隐私信息泄露问题。Yang 等人[102]研究递归网络中的成员推理攻击。Wunderlich 等人[117]研究了差分

隐私分层文本分类中的隐私-可用性平衡问题, 并识别超出平衡的网络架构。

② 文本生成中的 MIAs:

针对文本生成中的 MIAs 是指攻击者判断某个生成的文本数据是否是训练集成员。Song 等人[60]研究语言模型中的成员推理攻击, 提出一个新的针对深度学习模型审计技术。Hisamoto 等人[118]研究黑盒场景下端到端模型的成员推理攻击问题。

③ 词嵌入中的 MIA:

针对词嵌入中的 MIAs 是指攻击者判断某个词或句子是否是嵌入模型的训练集成员。2.3.4 节嵌入模型的 MIAs[103-104]主要研究词嵌入中的 MIAs。Jagannatha 等人[119]在黑白盒场景下, 设计和评估了 Bert 和 GPT2 模型的隐私安全。Carlini 等人[120]研究语言模型 GPT2 训练数据泄露情况。Klakow 等人[106]研究了各种预训练词嵌入模型(如 GloVe、ELMo 和 BERT)上的隐私泄露问题。

针对自然语言处理的 MIAs 攻击场景多样, 攻击实现 83.04%的准确率, 83.2%的 f1 分数, 并可迁移到其他模型和任务(如, 分类和生成任务), 但需要的背景知识较多, 模仿模型对攻击影响较大, 且需要对数据进行预处理。

3) 计算机视觉:

针对计算机视觉的 MIAs 是指攻击者判断计算机视觉领域的 MIAs, 包括图像生成中的 MIAs 和图像语义中的 MIAs。

① 图像生成中的 MIAs:

针对图像生成中的 MIAs 是指攻击者判断某个生成的图像是否是训练集成员。集中式 MIAs[72-74]和生成模型[108,114-115]主要针对图像生成中的 MIAs, 具体内容已在 2.3.2 和 2.3.4 节介绍过, 在此不再赘述。

② 图像语义分割中的 MIAs:

针对图像语义分割中的 MIAs 是指攻击者判断某个分割图像是否是训练集成员。He 等人[121]研究了语义图像分割领域的成员推理攻击问题, 并提出相应的隐私防御方法。Shafran 等人[122]发现具有高维输入输出的数据易遭受成员推理攻击, 并研究图像变换和语义分割模型中的成员推理攻击。

针对计算机视觉的 MIAs 攻击方法多样, 可攻击多个模型, 攻击效果较高(如, AUC 面积为 87.1%), 但背景知识较多, 且需要额外训练神经网络, 以及选择合适的成员特征实施有效攻击。

4) 音频:

针对音频的 MIAs 是指攻击者判断某个音频数据是否是训练集成员。Shah 等人[123]研究语音识别模型中的成员推理攻击, 可实现 60%的黑盒攻击准确率和召回率。Miao 等人[124]主要研究语音服务中的黑盒成员推理攻击, 设计了一个音频审计器且准确性可达 80%。针对音频的 MIAs 攻击精确率较高(如, 75%), 在较少语音信息下仍可实现有效攻击, 仅有黑盒攻击, 需要探索攻击性更强的攻击。

5) 推荐系统:

针对推荐系统的 MIAs 是指攻击者判断某个用户是否是推荐系统训练集成员。Zhang 等人[125]首次研究了推荐系统中的成员推理攻击, 提出一个新的方法来表示不同列表中的用户, 并采用一个影子推荐器来训练攻击模型。该方法攻击效果较高(AUC 面积 99.8%), 提出一个用户级的 MIA, 但攻击方法单一, 需要生成标签数据和训练影子模型。

6) 迁移学习:

迁移学习是指将各大预训练模型(称作教师模型)应用于下游任务(称作学生模型), 用于提高学生模型的准确性。针对迁移学习的 MIAs 是指攻击者根据学生模型推测某个样本是否是教师模型训练集的成员。Chen 等人[126]研究深度迁移学习中的隐私风险, 并提出了隐私防御方法。同时, Liew 等人[129]研究了迁移学习中的成员推理攻击问题, 提出一个新的策略从集成信息中进行推测。

针对迁移学习的 MIAs 攻击场景多样, 对深度迁移模型进行分类并分别实施攻击, 但背景知识较多, 攻击的适用领域有限(如, 人脸识别)。

7) 对比学习:

针对对比学习的 MIAs 是指攻击者根据对比学习的模型差异, 推测某个样本是否是源域的训练集成员。He 等人[127]首次研究对比学习中的成员推理攻击, 发现在图片数据集上训练的对比学习模型遭受成员推理攻击要小于属性推理攻击, 并提出一个隐私保护对比学习方案。该方法背景知识较少, 攻击准确性较高(如, 72.6%), 但攻击方法单一, 综合的评估指标很难反应真实攻击效果。

8) 图神经网络:

针对图神经网络的 MIAs 是指攻击者判断图神经网络领域的 MIAs。根据不同的任务可以分为知识图谱的 MIAs、节点分类中的 MIAs 和图分类中的 MIAs。

① 知识图谱的 MIAs:

针对知识图谱的 MIAs 是指攻击者判断某个样本是否是知识图谱训练集成员。Wang 等人[128]研究知识图谱中的成员推理攻击, 其主要在 4 个标准的知识图嵌入模型上实施成员推理攻击。

② 节点分类中的 MIAs:

针对节点分类中的 MIAs 是指攻击者判断图中某个节点是否是训练集成员。He 等人[96]研究图神经网络中基于 3 个节点级的成员推理攻击。yiola 等人[97]也研究了图神经网络中的成员推理攻击问题。Duddu 等人[105]首次研究图生成模型中的图嵌入隐私泄露问题。

③ 图分类中的 MIAs:

针对图分类中的 MIAs 是指攻击者判断某个样本是否是图分类网络的训练集成员。Wu 等人[130]首次研究图神经网络中的成员推理攻击, 提出基于训练和基于阈值的攻击, 发现已有的图模型可遭受成员推理攻击且 f1 分数达到 70%。

针对图神经网络的 MIAs 攻击场景多样, 可实现医疗和金融知识图谱的攻击, 攻击准确性较高(如, 72.6%), 但需要的背景知识较多, 采用综合的评估指标会造成评估结果不准确。

9) 在线学习:

针对在线学习中的 MIAs 是指攻击者判断某个样本是否是在线学习的训练集成员。Salem 等人[131]研究了在线学习场景下的成员推理攻击, 比较黑盒在线学习数据集更新前后的差异, 并提出四种基于编码器-解码器形式的攻击。该方法所需背景知识较少, 提出单个和多个样本攻击, 但攻击方法单一, 需要构造编码器和训练影子模型。

10) 机器去学习:

针对机器去学习中的 MIAs 是指攻击者判断某个被去掉的样本是否是机器去学习模型的训练集成员。Chen 等人[132]首次研究机器去学习领域中无意识信息泄露问题, 提出一个新的基本版本差异的成员推理攻击方法。该方法提出两种机器去学习中的评估指标, 但攻击方法单一, 需生成后验概率, 并重构特征。

11) 医疗场景:

针对医疗场景的 MIAs 是指攻击者判断某个样本是否是医疗模型的训练集成员。Wu 等人[133]表明其攻击会以很高的置信度重构真实医疗影像以及临床记录, 并提出一些防御机制。该方法攻击效果较好, 可重构医疗图片, 但攻击方法单一, 需训练攻击模型。

12) 工业物联网:

针对工业物联网的 MIAs 是指攻击者判断某个样本是否是工业物联网的训练集成员。Chen 等人[59]首次研究工业物联网中的协作成员推理攻击, 提出一个迁移遗传影子训练技术, 并放宽了已有假设, 但综合的评估指标无法准确反应攻击效果。

13) 边缘智能:

针对边缘智能的 MIAs 是指攻击者判断某个样本是否是边缘智能系统中的训练集成员。Wang 等人[61]首次提出一个针对多等级边缘智能的攻击, 但攻击准确性无法反应攻击真实情况。

2.3.6 按攻击的数据集大小

根据成员推理攻击可攻击的数据集大小, 我们将其划分为攻击整个数据集和攻击部分数据集两类(参见表 8)。

1) 攻击整个数据集:

针对攻击整个数据集的 MIAs 是指攻击者利用成员推理攻击方法可以对数据集所有样本进行攻击, 其主要在所有样本上采用综合的评估指标(比如, 准确率、精确率、召回率、f1 分数等)。已有大部分成员推理攻击主要攻击整个数据集, 其具体内容已在 2.3.1、2.3.2、2.3.3、2.3.4 和 2.3.5 节详细介绍过, 在此不再赘述。

针对攻击整个数据集的 MIAs 使用综合的评估指标可以对整个测试集中所有样本进行评估, 但是召回率和 f1 分数在大多数情况下很高, 精确率在某些情况下很低, 且有高的假阳率, 严重损害攻击者的利益。

2) 攻击部分数据集:

针对攻击部分数据集的 MIAs 是指攻击者利用成员推理攻击方法只对数据集中特定精心挑选的样本进行有效攻击, 而对其他样本无法实施攻击, 其主要通过设计不同的评估机制, 并选择满足条件的阈值来实施基于评估机制的成员推理攻击。已有小部分成员推理攻击主要攻击部分数据集, 其具体内容已在 2.3.1、2.3.2、2.3.3、2.3.4 和 2.3.5 节详细介绍过, 在此不再赘述。

针对攻击部分数据集的 MIAs 利用不同的评估机制, 根据设定的特定阈值, 对满足阈值条件的样本, 可降低假阳率, 提高攻击质量, 但是这些攻击只对满足阈值要求的样本可实施较高质量攻击, 对其他样本无法实施有效攻击。

3 成员推理攻击存在的原因

已有一些工作[63,65,136-137]分析了成员推理攻击存在的原因, 但仍缺乏更科学严谨的解释。本小节, 主要从目标模型的训练数据、模型类型、过拟合程度三个角度, 分析成员推理攻击存在的原因。

3.1 目标模型的训练数据

目标模型训练数据集越具有代表性, 其遭受隐

表 8 根据攻击数据集大小对成员推理攻击进行分类

Table 8 Classifications of MIAs based on the size of the attack datasets

攻击数据集大小	攻击原理	威胁模型	评估指标	攻击效果	优点	缺点	参考文献
整个数据集	基于二元分类器	黑/灰/白	准确率、精确率、召回率、f1 分数、成员优势、AUC 面积	精确率 51.7% 召回率 93%	使用综合的评估指标可以对整个测试集中所有样本进行评估	召回率和 f1 分数在大多数情况下很高, 精确率在某些情况下很低, 且有高的假阳率, 严重损害攻击者的利益	[14,28,31,35,40,42,51,54-56,58-61,67,76-77,81-83,85,87,90,93-94,96-97,102,107,109-110,115-116,118,121,123-127,129,131,135,194]
	基于评估机制	黑/灰/白		精确率 50.05% 召回率 99%			[32-34,39,41,43,47-50,54,57,61-68,72,74-75,78-80,84,86,89-90,93,95,98,100,103-106,108,111-112,114,117,119-120,122,126-127,130-131,133,136-137,139,164]
	基于数据集差异	黑/灰/白		精确率 50.01% 召回率 98.18% f1 分数 97.24%			[70]
部分数据集	基于评估机制	黑/白	沙普利值	精确率 94.06% 召回率 88.06%	利用不同的评估机制, 根据设定的特定阈值, 对满足阈值条件的样本, 可降低假阳率, 提高攻击质量	只对满足阈值要求的样本可实施较高质量攻击, 对其他样本无法实施有效攻击	[101]
			难度校准分数	AUC 面积 64.8%, 准确率 61.8%			[113]
			隐私风险分数	准确率 79.4% 精确率 88.2%			[69]
			正例预测值	成员优势 26.8%			[71]
			基于蒸馏的损失阈值	AUC 面积 87.6%			[36]
			低假阳率下的真阳率	准确率 82.6% 0.1%假阳率下的真阳率 27%			[99]
			Log 损失值和影响分数	精确率 95.05%			[52]

私泄露的风险越低。因为训练数据很好地表示了整个数据集的分布, 利用其训练出的模型能很好的捕捉和表征数据特征, 从而使得目标模型有很好的泛化性。文献[12]表明: 训练数据越多, 越不容易区分成员和非成员。然而, 现实中数据类型繁多, 且没有一个系统客观评估训练数据代表性的方法, 从而导致训练得到的目标模型极易遭受各种隐私攻击(如, 成员推理攻击), 极大地破坏了机器学习的安全性和隐私性。

3.2 目标模型的类型

除目标模型的训练数据外, 已有研究[55]表明目标模型的类型对其遭受成员推理攻击的风险起着至关重要的作用, 其通过对 DNN、逻辑回归、朴素贝叶斯、k-近邻和决策树等模型进行成员推理攻击后, 发现: 决策树是 6 个模型中攻击精确率最高的, 而朴素贝叶斯则是最低的, 原因是对于朴素贝叶斯模型来说, 单个训练数据只能在边缘影响给定类的预测; 而对于决策树模型而言, 一个样本就代表一个独一无二的特征, 可使决策树产生一个新的分支, 并改变分类边界。因此, 不同的模型遭受攻击的风险不同。

3.3 目标模型的过拟合

除目标模型的训练数据和类型外, 已有研究[14,54,65,74-75]表明目标模型的过拟合是导致其遭受成员推理攻击的最主要原因。文献[138]表明造成模型过拟合的主要原因是模型的高复杂性和训练数据集有限的数据量。深度学习模型通常是过参数化的, 而且具有很高的复杂性, 会有很强的能力记住噪声或者给定数据集的细节信息[47-50]。此外, 机器学习模型在训练时需要重复的在相同的样本训练很多个 epochs, 从而导致训练样本很容易被模型记住。同时, 有限的训练数据量很难完整表示整个数据分布, 限制了模型泛化性, 从而很难捕捉成员和非成员特征。

4 成员推理攻击的防御

本小节, 主要从差分隐私、正则化、数据增强、模型堆叠、早停、信任分数掩蔽和知识蒸馏七个方面介绍 ML 模型中成员推理防御(参见表 9)。图 5 是 MIAs 防御发展历程。

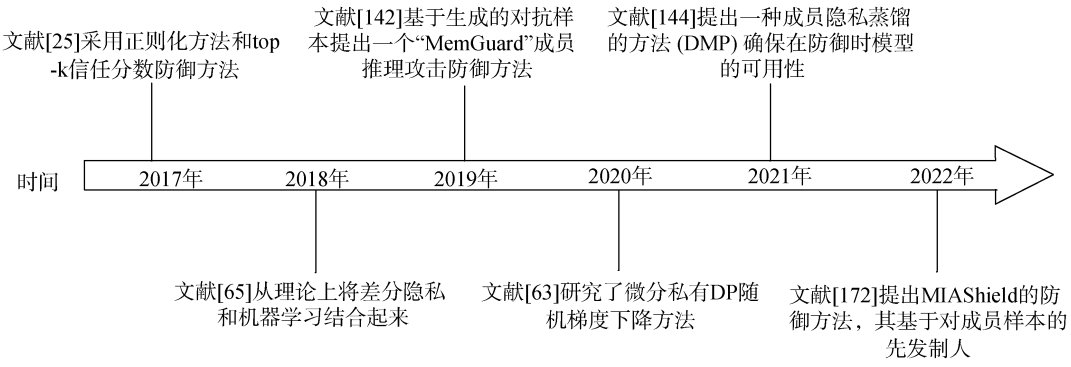

图 5　成员推理攻击防御发展历程

Figure 5　The development of membership inference attacks' defenses

4.1　基于差分隐私

利用差分隐私技术抵御成员推理攻击[32,65-66,70-71,74-75,88,90-92,114,140-145]是指攻击者给样本添加噪声抵御 MIAs。

1) 差分隐私抵御分类模型中的 MIAs

Shokri 等人[14]首次讨论了差分隐私抵御成员推理攻击。Yeom 等人[65]从理论上将差分隐私和机器学习联系起来。Rahimian 等人[68]提出一种只在预测阶段给输入样本的 logits 添加噪声方法(DP-Logits), 并限制查询的次数。Truex 等人[90]评估了在类别差异和不平衡数据上训练时, 差分隐私如何影响模型。Rahman 等人[91]首次系统性地评估了成员推理攻击在差分隐私深度神经网络上的攻击效果, 发现其会降低模型可用性。随后, Jayaraman 等人[141]在多个差分隐私机制上进行了系统性研究。

2) 差分隐私抵御生成模型中的 MIAs

有研究[114,140,150-155]用差分隐私抵御生成模型中的 MIAs。Hayes 等人[32]首次评估了 MIAs 在差分隐私生成对抗网络(DP GAN)[152]中的攻击效果, 隐私预算会影响攻击效果。Wu 等人[114]系统性的证明使用差分隐私训练 GANs 的泛化误差有界限。Chen 等人[140]发现差分隐私可降低 GANs 的 MIAs 风险, 但会导致 GANs 生成的样本质量变差, 并增加计算开销。Hu 等人[147]提出了一种新的针对 GANs 的 MIAs 防御方法。Chen 等人[148]提出了一个新的防御框架 RelaxLoss, 能抵御广泛的攻击。Bernau 等人[149]评估了差分隐私可变自动编码器的强重建 MIAs。Alvar 等人[156]提出了对抗性知识提炼抵御图像翻译模型中的 MIAs。Chen 等人[157]提出一种增强型混合训练的防御方法。Yang 等人[171]提出了一个通过 VAE 和扩展的差异隐私机制构建隐私保护系统框架。

虽然, 差分隐私给成员隐私提供了理论保障, 但其几乎不能提供一个可接受的隐私-可用性平衡, 当隐私预算较大时会导致模型不可用[141]。

4.2　基于正则化

正则化技术主要通过降低目标模型的过拟合来抵御成员推理攻击, 根据已有防御方法[14,54,69-70,92,144,159-161], 我们将从 L2 正则化[14]、Dropout[162]、标签平滑[163]、对抗正则[159,161]、Mixup + MMD[70,92], 介绍基于正则化的 MIAs 防御方案。

1) L2 正则化

基于 L2 正则化的 MIAs 防御是指攻击者给损失函数添加 L2 正则化保护数据隐私。Shokri 等人[14]主要将 L2 正则化添加到其损失函数中降低模型的过拟合并保护数据隐私。

2) Dropout

基于 dropout 的 MIAs 防御是指攻击者随机去掉一些神经元来保护数据隐私。Salem 等人[54]也采用 dropout 方法抵御 MIAs。Srivastava 等人[162]首次提出利用 dropout 来保护数据隐私, 在每次训练的过程中任意去掉一些神经元。

3) 标签平滑

基于标签平滑的 MIAs 防御是指攻击者对标签进行平滑处理来保护数据隐私。Szegedy 等人[163]提出一种标签平滑的成员推理攻击防御方法, 将样本的原始标签分布和一个给定的分布进行混合计算, 并作为最后的标签。

4) 对抗正则

基于对抗正则的 MIAs 防御是指攻击者采用生成对抗网络的对抗思想来保护数据隐私。Hu 等人[159]研究 GANs 的各种变体遭受成员推理攻击的情况, 并提出一种基于 Least Square GANs (LSGANs)的增强对抗正则方法来保护隐私。随后, Nasr 等人[161]采用对抗正则的方法抵御 MIAs, 提出一个基于生成对抗网络的 MIN-MAX 博弈的方法。

5) Mixup + MMD

基于 MMD+Mix-up 的 MIAs 防御是指攻击者结合 MMD 和 Mix-up 技术来保护数据隐私。Li 等人[92]提出基于 MMD+Mix-up 的正则化矩阵法实现隐私保护。Hui 等人[70]在 Mixup + MMD 防御方法上验证他们攻击方法 BlindMI-DIFF 的可行性。

基于正则化的 MIAs 防御方法可任何情况下保护数据隐私, 但很难提供满意的隐私和可用性平衡。

4.3 数据增强

基于数据增强的 MIAs 防御是指攻击者对数据进行增强处理来保护数据隐私。Kaya 等人[160]提出一个损失排序相关性机制来评估不同机制间的相似性, 并提出实际的隐私-可用性平衡的数据增强方法。该方法可通过数据降低过拟合, 但其需要额外数据, 进一步增大防御成本。

4.4 模型堆叠

基于模型堆叠的 MIAs 防御是指攻击者将多个弱模型组合成一个强模型保护数据隐私。Salem 等人[54]利用模型堆叠来抵御成员推理攻击, 主要将若干个弱的机器学习模型组合成一个强的机器学习模型, 从而降低泛化误差。该方法联合多个模型的优点可实现更强的防御, 但是联合相同模型效果欠佳, 联合不同模型又增大防御开销。

4.5 早停

基于早停的 MIAs 防御是指攻击者利用很少的训练 epochs 来实现高的模型攻击准确性和低的隐私风险之间的平衡。Song 等人[69]比较了他们的成员推理防御方法和早停方法。该方法可在训练阶段控制 epochs, 通过简单操作实现防御, 但是 epochs 大小不好控制, 需要花费训练时间和成本。

4.6 基于信任分数掩蔽

基于信任分数掩蔽的隐私保护方法通过隐藏目标分类器输出的真实信任分数来保护成员隐私。主要包括: 只输出前 k 个信任分数(top-k); 只输出预测标签; 给信任分数添加精心设计的噪声。

1) top-k 信任分数向量

Shokri 等人[14]首次在全连接网络中发现 top-3 信任分数仍不能抵御基于影子模型的 MIAs。Salem 等人[54]利用部分的信任分数实现和利用完整信任分数相似的攻击效果。

2) 只输出预测标签

文献[14]表明只返回预测标签可降低攻击准确

表 9 成员推理攻击防御方法

Table 9 Defenses of Membership inference attacks

防御方法	英文缩写	威胁模型	防御原理	优点	缺点	分类	参考文献
差分隐私	Diferential privacy	黑/灰/白	给隐私信息提供理论隐私保证	给隐私信息提供理论隐私保证	损害模型可用性	分类模型	[14,68,134,149]
						生成模型	[114,140,147-149,150-156]
正则化	Regularization	白盒	降低目标模型的过拟合	可在任何情况(黑/灰/白)进行保护	很难提供一个满意的隐私和可用性平衡点	L2 正则化	[25]
						dropout	[162,164]
						标签平滑	[163]
						对抗正则	[159,161]
						Mixup + MMD	[70,92]
数据增强	Data argumenta-tion	黑盒	降低目标模型的过拟合	可通过数据降低过拟合	需要额外数据, 增大防御成本	/	[160]
模型堆叠	Model stacking	黑盒	降低目标模型的过拟合	联合多个模型的优点	相同模型效果欠佳	/	[54]
早停	Early stopping	黑盒	降低目标模型的过拟合	可在训练阶段控制 epochs	epochs 大小不好控制	/	[69,165-167]
信任分数掩蔽	Confidence score masking	黑/白	隐藏目标分类器输出的真实信任分数向量	不需要重新训练目标模型, 不影响目标模型的分类准确性	不能够提供足够的隐私保证	top-k 信任分数	[12,54]
						只输出预测标签	[14,66-67]
						噪声信任分数	[69,142]
知识蒸馏	Knowledge distillation	黑/白	将教师模型知识迁移到学生模型	减少对隐私数据的依赖	蒸馏数据好坏难以衡量, 仍存在隐私泄露风险	知识蒸馏	[144,172-173]

率。Li 等人[67]提出了基于输出标签的成员推理攻击, Choquette 等人[66]也研究了只有输出标签的 MIAs, 发现只输出预测标签仍会泄露数据隐私。

3) 添加噪声的信任分数

Jia 等人[142]提出一个基于对抗样本的“Mem-Guard”防御方法, 主要给信任分数添加噪声。Song 等人[69]重新评估了 Mem-Guard[142]的有效性, 发现 Mem-Guard[142]仍易遭受成员推理攻击。

基于信任分数掩蔽的 MIAs 防御方法无需重新训练目标模型, 不影响目标模型的分类准确性, 但不能提供足够的隐私保证。

4.7 基于知识蒸馏

知识蒸馏是指利用大的教师模型的输出来训练一个小的学生模型, 将大的教师模型上的知识迁移到小的学生模型上, 并允许学生模型拥有和教师模型相似的准确率[172]。基于知识蒸馏的 MIAs 防御是指攻击者利用知识蒸馏处理数据后再进行模型训练。Shejwalkar 等人[144]提出一种成员隐私蒸馏方法确保在防御时模型的可用性以及新的标准。Zheng 等人[173]提出两个互补性知识蒸馏。基于知识蒸馏的 MIAs 防御方法减少对隐私数据的依赖, 但蒸馏数据的好坏影响防御效果且难以衡量, 仍存在隐私泄露风险。

5 成员推理攻击的评估指标和数据集

本小节, 总结了成员推理攻击和防御的评估指标以及使用的数据集。

5.1 评估指标

本小节, 主要介绍目标模型和攻击模型的评估指标(参见表 10)。

5.1.1 目标模型的评估指标

1) 准确率(model-side accuracy): 是指预测正确的样本占所有样本的比例;

2) 泛化误差(generalization errors): 目标模型训练准确率和测试准确率之间的差, 反映目标模型的过拟合程度; 泛化误差越大, 目标模型的过拟合程度越高。

5.1.2 攻击模型的评估指标

1) 攻击准确率(attacker-side accuracy)是指: 攻

表 10 成员推理攻击评估指标

Table 10 Evaluation metrics of Membership inference attacks

评估指标	英文缩写	优点	缺点	参考文献
准确率	Attacker-side accuracy	整体反映模型预测能力	不能客观全面的反映模型的性能	[14,54-55,57-59,60-64,66,69,72-73,76-78,83-84,86,89-90,93,96,105,107,118,124,142-144,161,168,173-192,195]
精确率	Attacker-side precision	反映对正例的预测能力	无法全面评估模型整体性能	[14,52-55,58-60,63,65-66,69,71-72,75,77,80,84,89,97,107,124,126,173,192]
召回率	Attacker-side recall	反映对真正例的预测能力	无法反映对假正例的表现	[14,53-54,59-60,63,65,69,75,80,84,89,97,107,124,151,173,192]
f1-分数	Attacker-side f1-score	精确率和召回率的调和平均数, 评估较全面	无法直接反映对假正例的表现	[70,77,84,124,187,191,194]
假阳率	Attacker-side FPR	可以对反例的预测能力	无法反映对正例的表现	[84,151]
成员优势	Attacker-side membership advantage	联合召回率和假阳率, 评估较全面	仅间接反映预测能力	[65,71,75,88,92-93,103,105,117,141,160,182]
AUC 面积	Area-under the-ROC-curve	考虑召回率和假阳率, 评估较全面	在基于评估机制的攻击中不适用	[60,68,74,97,108,114,117,121-122,126,187,192,196-199]
沙普利值	Shapley values	对满足特定阈值要求的样本可实现高精确率、低假阳率的攻击	仅对满足特定阈值要求的样本可实现高质量攻击, 无法对其他样本实施有效攻击, 且所有攻击方法仅适用于基于评估机制的 MIAs, 并不适用于所有类型的 MIAs	[101]
校准分数	Calibrated score			[113]
隐私风险分数	Privacy risk scores			[69]
正例预测值	Positive predictive value			[71]
基于蒸馏的损失阈值	Distillation-based loss threshold			[36]
低假阳率下的真阳率机制	A true-positive rate at low false-positive rates metric			[99]
Log 损失值	Log loss value			[52]

击模型预测正确的样本占所有样本的比例;

2) 攻击精确率(attacker-side precision)是指: 在预测为正例的样本中, 攻击模型预测正确的正例占所有预测为正例的比重;

3) 攻击召回率(attacker-side recall)是指: 攻模型预测正确的正例占所有真实正例的比例;

4) 攻击f1分数(attacker-side f1-score): 是指攻击精确率和召回率的调和平均数;

5) 攻击假阳率(attacker-side false positive rate, FPR)是指: 攻击模型预测错误的反例占所有反例的比例;

6) 成员优势(attacker-side membership advantage)是指: 攻击模型的召回率和假阳率之间的差, 其反应攻击模型预测一个样本是成员的优势;

7) AUC面积: 是ROC曲线下的面积, ROC曲线是指横轴是FPR, 纵轴是TPR(recall), 且ROC曲线越靠近左上方分类模型性能越好; AUC面大, 攻击模型性能越好;

8) 沙普利值(shapley values, SV)是指: 当一个样本“在”与“不在”训练数据子集中时, 目标模型的预测准确率的平均变化率;

9) 难度校准分数(calibrated score)是指: 目标模型对某个样本的成员分数(模型对数据样本的损失)与影子模型对该样本的成员分数的差值;

10) 隐私风险分数(privacy risk scores)是指: 攻击者在观察了目标模型对一个样本的表现后, 判断其来自于训练集的后验概率;

11) 正例预测值(positive predictive value, PPV)是指: 当给一个样本添加噪声后模型的预测损失变小的比例, 以及目标模型对单个样本损失的下界阈值和上界阈值, 当一个样本添加噪声后同时满足三个阈值要求时, 攻击者认为样本是成员, 否则是非成员;

12) 基于蒸馏的损失阈值(distillation-based loss threshold)是指: 攻击者将模型在蒸馏的数据集上进行训练, 并根据不同模型和的数据集设定一个在可忍受假阳率范围内的阈值, 当某个样本的损失阈值大于该设定的可忍受假阳率阈值时, 认为该样本是成员, 否则是非成员;

13) 低假阳率下的真阳率机制(a true-positive rate at low false-positive rates metric)是指: 攻击者在对数尺度上绘制ROC曲线, 并报告一个固定的低假阳率下的真阳率的大小;

14) Log损失值(log loss value)是指: 攻击者首先定义log损失值, 再根据训练集中的数据样本的邻居个数选择易受攻击样本, 计算一个样本对另一个样本的影响分数以及增强样本, 当这些易受攻击样本的模型损失大于该log损失值时, 认为该易受攻击样本是成员, 否则是非成员。

5.2 数据集

本小节, 我们将MIAs数据集划分为图片数据、文本数据、图数据以及二元数据(参见表11), 其主要被用于分类、生成新数据、分类和生成、以及语义分割等任务。

6 成员推理的应用

成员推理攻击不仅可以推测数据隐私, 而且在现实中也有一些应用。

6.1 隐私审计

已有一些开源工具[38]和机器学习图书馆[37]利用成员推理攻击进行ML模型隐私评估。Song等人[35]设计了一个针对文本生成模型的隐私审计模型, 评估某个文本是否被未授权使用。Ye等人[36]基于假设检验提出一种新的成员推理攻击方法, 可被用作模型隐私审计和评估的一种工具。Miao等人[158]提出一个语音审计模型, 可推测用户的语音数据是否被非法训练和使用。

表11 成员推理攻击和防御中所使用的数据集

Table 11 Datasets used in Membership inference attacks

数据类别	ML任务	数据集	数据量	类别个数	特征个数	参考文献
图片数据	分类	Colored-MNIST	70000	2	28×28×1	[183]
		CH-MNIST	5000	8	150×150×1	[68,70,142]
		SVHN	99289	10	32×32×3	[89,197]
		Yale Face	2414	38	168×192×1	[200]
		RCV1X	800000	103	/	[71]
		Birds-200	11788	200	/	[70]
		FaceScrub	100000	530	/	[188]
		Market-1501	26051	1501	128×64	[42]

续表

数据类别	ML 任务	数据集	数据量	类别个数	特征个数	参考文献
图片数据	分类	PRID-2011	71657	934	/	[42]
		QMNIST	402953	10	28×28×1	[44]
		USPS	7291	10	16×16	[101]
		FLOWER	3670	5	320×240	[101]
		CREDIT	30000	2	24	[101]
		MEPS	15830	2	/	[101]
		CENSUS	48842	2	103	[101]
		Stanford Dogs	20580	120	/	[146]
		Synthetic	/	/	/	[79]
		GTSRB	51839	/	64×64	[67]
		Face	13000	1680	/	[67]
		MotionSense	70610	/	/	[149]
		CMP Facade	3475	/	/	[156]
	生成	ImageNet	1281167	1000	/	[64,84,168,190,197]
		MIMIC-III	46520	/	1,071	[74]
		Insta-NY	34336	/	4048	[56,74]
		CelebA	202599	10177	218×178×3	[74,108,122]
		ChestX-ray8	108948	32717	1024×1024×1	[108]
		IDC	277524	2	50×50×3	[114,187]
		EyePACS	88702	5	/	[70,73]
	分类和生成	MNIST	70000	10	28×28×1	[14,52-56,59,61-62,65-68,72,75,77-78,84,90,92, 151,175,179,185,197]
		Fashion-MNIST	70000	10	28×28×1	[62,68,72,93,177,197]
		CIFAR-10	60000	10	32×32×3	[54-57,59,61-64,66,68,72-73,75,78,84,89-90,93, 107,137,144,151,168,173,175,177,185,189-191,197]
		CIFAR-100	60000	100	32×32×3	[14,54,57,59,64-70,75-76,84,90,92-93,143-144, 161,168,173,179,185,190, 192, 197]
		LFW	13233	5749	62×47×3	[54,67,73,75,80,90,114,174]
	语义分割	BDD100K	100000	/	/	[194]
		Cityscapes	20000	30	/	[122,194]
		Mapillary-Vistas	25000	37	/	[194]
		Broward	7200	2	8	[39]
		IMDb	50000	2	/	[169]
		Texas	67330	100	6169	[34]
		Location	5010	30	446	[34]
文本数据	分类	WikiText-103	100000000	/	/	[99]
		ADM	1000000000	/	/	[125]
		Lastfm-2k (lf-2k)	1000000000	/	/	[125]
		Movielens-1m	1000000	/	/	[125]
		Multi30K	30000	/	/	[102]
		BestBuy	51646	/	/	[117]
		RCV1	800000	/	/	[117]
		DBPedia	337739	/	/	[117]
		CSI	1412	2	/	[80]
		Review	364038	2	/	[189]
		Tweet EmoInt	7097	4	/	[58]
		Yelp-health	17938	10	/	[80]
		News	20000	20	/	[54]
		Weibo	23000	/	/	[58]

续表

数据类别	ML 任务	数据集	数据量	类别个数	特征个数	参考文献
文本数据	生成	Reddit comments	83293	/	/	[60]
		Dialogs	220579	/	/	[60]
		SATED	2324	/	/	[60]
		WMT18	/	/	/	[118]
	嵌入	BookCorpus	14000	/	/	[103]
		Wikipedia	150000	/	/	[103]
图数据	分类	PROTEIN full	1113	2	/	[130]
		DD	1178	2	/	[130]
		ENZYMES	600	6	/	[130]
		OGBGPPA	158000	37	/	[130]
		WN18RR	40943	/	11	[128]
		FB15K237	14541	/	237	[128]
		NELL-995	75492	/	200	[128]
		Cora	5429	7	/	[97]
		CiteSeer	4715	6	/	[97]
		PubMed	44338	3	/	[97]
		Flickr	449878	7	/	[97]
		Reddit	57307946	41	/	[97]
		FFHQ	70000	/	62×64	[139]
		Pubmed	19717	3	500	[97,105]
		Citeseer	3327	6	3703	[96-97,105]
		Cora	2708	7	1433	[96-97,105]
		Lastfm	7624	18	7842	[96]
		Hepatitis	155	2	19	[75]
		Cancer	699	2	10	[52-53,75]
		Diabetes	768	2	8	[75]
		German credit	1000	2	20	[75,187]
		Adult	48842	2	14	[14,52-54,61,66,70,75,83,88,90,183]
二元数据	分类	Hospital	101766	2	127	[57]
		US-Accident	3000000	3	30	[56]
		Foursquare	528878	30	446	[14,54,66,69-70,80,142,177,198-199]
		Texas-100	67330	100	6170	[14,57,66,68-71,76,92,142,144,161,174-175,183,197]
		Purchase-100	197324	100	600	[14,54-55,57,61,66,68-70,72,76,90,92,143-144,151,161,173-174,177,183,198-199]
		UTKFace	20705	106	/	[83]

6.2 知识产权保护

文献[165]利用基于影子模型的成员推理攻击方法[12]筛选成员样本, 并任意选择 20%成员样本嵌入目标模型中, 实现目标模型的版权保护。成员推理技术还被用于系统发布前的隐私质量评估, 判断模型是否具备发布标准, 以及监管部门对用户个人隐私的非法滥用、对用户位置[46,170]和信用进行监测。

6.3 疾病预测

成员推理在医疗领域[166-167]也得到了广泛应用, 比如推测某个基因是否在基因库中, 或某人是否患有某种疾病。

7 主要挑战和未来研究方向

本小节, 我们将讨论成员推理攻击和防御的主要挑战和潜在研究方向, 为该领域的研究者提供一些建议(参见表 12)。

7.1 成员推理攻击的主要挑战和研究方向

7.1.1 研究全面高效的成员推理攻击方案

目前, 几乎所有研究主要关注成员样本的预测和评估, 很少涉及非成员样本的检测。而现实的成员

推理任务中, 大部分都是非成员样本, 且攻击者对其需要评估的样本一无所知, 此时已有的攻击无法准确评估成员和非成员。此外, 成员评估的 f1 分数和召回率几乎在大部分情况下都很高, 而精确率有时很低, 进一步增加了攻击者区分的难度, 甚至误导攻击者得出错误结论。因此, 如何全面有效评估现实中的成员和非成员是成员推理攻击面临的一大挑战, 研究全面高效成员推理攻击方案是一个亟待解决的问题。

表 12 成员推理攻击防御主要挑战和未来研究方向

Table 12 Main challenges and future research directions of membership inference attacks

研究内容	主要挑战	未来研究方向
成员推理攻击	已有成员推理攻击方法很难在现实中全面有效评估成员和非成员	如何全面有效评估现实中的成员和非成员, 并研究全面高效成员推理攻击方案
	现实生活中攻击者由于拥有的背景知识有限, 其很难确定一个模型是否过拟合	研究已有的针对过拟合模型的成员推理攻击是否对非过拟合模型, 并提出针对非过拟合模型的高效攻击方法
	已有研究对一些新兴模型和领域的成员推理攻击方法还未涉及和探索, (如自监督学习、元学习等)	探索和研究针对新兴模型和领域的成员推理攻击方案
	针对成员推理攻击在现实中应用(如模型隐私安全审计等)的研究较少, 且涉及机器学习很小的范围	需要进一步完善成员推理攻击在现实和其他任务上的应用(如联邦学习中如何准确确定某个参与者的身份等)
成员推理攻击防御	已有的成员推理防御方案很难防御所有场景和模型中的成员推理攻击	结合密码学、对抗训练等知识对已有防御方法进行优化和改进, 并探索新的鲁棒性高的成员推理攻击方案
	差分隐私成员推理防御方案会破坏模型可用性; 而其他防御方法不具有理论保障, 也很难兼顾隐私和模型效用	研究和设计一个隐私-可用性平衡的成员推理攻击防御方案
	对于一些其他模型(如生成模型等)的成员推理防御方案还很少, 很难评估无监督模型是否过拟合	研究如何判断一个模型的过拟合程度, 并设计有效的成员推理防御方案
	已有研究很少关注不同设置下成员推理有效性, 很难适应现实中类别个数、特征分布不均衡的情况	探究不同设置对成员推理攻击的影响, 并提出针对不同设置的成员推理防御方案

7.1.2 研究针对非过拟合模型的高效攻击方法

已有研究表明成员推攻击在过拟合模型中取得了巨大的成功, 但现实生活中攻击者由于拥有的背景知识有限, 其很难确定一个模型是否过拟合。同时, 随着软硬件技术的飞速发展和革新、易获得的的海量数据, 会不断提升训练模型的质量, 进一步降低其过拟合的风险。因此, 已有的针对过拟合模型的成员推理攻击是否对非过拟合模型也有效还有待进一步探索和研究, 需要研究针对非过拟合模型的高效攻击方法。

7.1.3 研究其他模型和领域的成员推理攻击方法

机器学习中的众多模型和领域已遭受成员推理攻击, 如分类模型、生成模型等。但对一些新兴模型和领域(如自监督学习、元学习、同质联邦学习等)的成员推理攻击研究较少; 而这些模型在机器学习中扮演着越来越重要的角色, 探索和研究针对这些模型和领域的成员推理攻击方案, 将有助于对其进行更好的隐私防御, 从而推动机器学习的蓬勃发展。

7.1.4 研究成员推理攻击的其他用途

已有成员推理攻击主要研究如何提高攻击成功率, 也有少部分文献研究利用成员推理攻击方法进行模型隐私安全审计等任务。但已有应用只涉及机器学习很小的范围, 也处于刚起步阶段, 还需进一步完善。因此, 成员推理攻击在机器学习其他任务上的用途仍需探索和研究。比如, 联邦学习中如何准确确定某个参与者的身份, 如何根据成员信息研究有效的机器去学习方案等等。

7.2 成员推理防御的主要挑战和研究方向

7.2.1 研究优化高效的成员推理攻击防御方案

已有的成员推理防御方案虽已取得一定的防御效果, 但是机器学习场景繁多、模型多样, 其很难防御所有场景和模型中的成员推理攻击。因此, 需结合密码学、对抗训练等知识对已有防御方法进行优化和改进, 并探索和研究新的鲁棒性更强的成员推理攻击方案。

7.2.2 研究隐私可用性平衡的成员推理防御方案

差分隐私成员推理防御方案虽具有一定的理论保障, 但其会破坏模型可用性; 而其他防御方法不具有理论保障, 也很难兼顾隐私和模型效用。因此, 需要研究和设计一个隐私-可用性平衡的成员推理攻击防御方案。

7.2.3 研究针对其他模型的成员推理防御方案

目前, 成员推理防御方案主要针对分类模型, 但是对其他模型(如生成模型等)的成员推理防御方案还很少。同时, 对于无监督模型来说, 很难评估其是否过拟合。因此, 如何判断一个模型的过拟合程度,

并设计有效的成员推理防方案是未来的研究方向。

7.2.4 研究不同设置下成员推理防御方案

目前, 大部分成员推理攻击和防御方案主要针对整个数据集的所有样本, 但很少研究不同设置下成员推理方案的有效性, 而现实中类别个数、特征分布不均衡的现象(如长尾现象)很常见, 探究不同设置对成员推理攻击的影响, 以及如何提出针对不同设置的成员推理防御方案, 仍是一个开放和值得研究的问题。

8 总结

本文首先介绍了机器学习在实现人工智能时取得的巨大成功, 以及面临的隐私安全威胁, 尤其是成员推理攻击; 其次, 介绍了成员推理攻击的定义和威胁模型, 并从攻击场景、背景知识、目标模型、攻击原理、攻击领域、攻击数据集大小等方面对成员推理攻击进行全面细致的分类和归纳; 然后, 从目标模型的训练数据、模型类型、过拟合程度分析成员推理攻击存在的原因; 随后, 从差分隐私、正则化、数据增强、模型堆叠、早停、信任分数掩蔽和知识蒸馏这七个方面对现有成员推理攻击防御措施进行分析和比较。接着, 总结了成员推理攻击和防御的评估指标、数据集, 以及其在现实中的应用; 最后, 分析讨论了其面临的隐私威胁和挑战, 并给出未来研究方向, 进一步推动该领域繁荣发展。

参考文献

[1] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[2] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. 2018: arXiv: 1810.04805. https://arxiv.org/abs/1810.04805

[3] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[EB/OL]. 2016: arXiv: 1609.02907. https://arxiv.org/abs/1609.02907

[4] Litjens G, Kooi T, Bejnordi B E, et al. A Survey on Deep Learning in Medical Image Analysis[J]. *Medical Image Analysis*, 2017, 42: 60-88.

[5] Liao G H, Liu J Y. A Malicious Code Detection Method Based on Data Mining and Machine Learning[J]. *Journal of Information Security Research*, 2016, 2(1): 74-79.
(廖国辉, 刘嘉勇. 基于数据挖掘和机器学习的恶意代码检测方法[J]. *信息安全研究*, 2016, 2(1): 74-79.)

[6] Chen X Y, Xiang S M, Liu C L, et al. Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks[J]. *IEEE Geoscience and Remote Sensing Letters*, 2014, 11(10): 1797-1801.

[7] Wittel G L, Wu S F. On attacking statistical spam filters[C], *Proc of Confon Email & Anti-spam*. *Mountain View: CEAS*, 2004.

[8] Ling C T. Evolutionary Neural Network for Credit Card Fraud Detection[J]. *Microelectronics & Computer*, 2011, 28(10): 14-17.
(凌晨添. 进化神经网络在信用卡欺诈检测中的应用[J]. *微电子学与计算机*, 2011, 28(10): 14-17.)

[9] de Cristofaro E. An Overview of Privacy in Machine Learning[EB/OL]. 2020: arXiv: 2005.08679. https://arxiv.org/abs/2005.08679

[10] Jere M S, Farnan T, Koushanfar F. A Taxonomy of Attacks on Federated Learning[J]. *IEEE Security & Privacy*, 2021, 19(2): 20-28.

[11] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]. *The 25th USENIX Conference on Security Symposium*, 2016: 601-618.

[12] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]. *The 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015: 1322-1333.

[13] Ganju K R, Wang Q, Yang W, et al. Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018: 619-633.

[14] Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 3-18.

[15] Tabassi E, Burns K J, Hadjimichael M, et al. A taxonomy and terminology of adversarial machine learning[J]. *Journal of research of the national institute of standards and technology*, 2019: 1-29.

[16] Veale M, Binns R, Edwards L. Algorithms that Remember: Model Inversion Attacks and Data Protection Law[J]. *Philosophical Transactions Series A*, *Mathematical*, *Physical*, *and Engineering Sciences*, 2018, 376(2133): 20180083.

[17] Homer N, Szelinger S, Redman M, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays[J]. *PLoS Genetics*, 2008, 4(8): e1000167.

[18] Liu B, Ding M, Shaham S, et al. When Machine Learning Meets Privacy: A Survey and Outlook[J]. *ACM Computing Surveys*, 2022, 54(2): 31.

[19] Liu H C, Wang Y Q, Fan W Q, et al. Trustworthy AI: A Computational Perspective[EB/OL]. 2021: arXiv: 2107.06641. https://arxiv.org/abs/2107.06641

[20] Mireshghallah F, Taram M, Vepakomma P, et al. Privacy in Deep Learning: A Survey[EB/OL]. 2020: arXiv: 2004.12254. https://arxiv.org/abs/2004.12254

[21] Rigaki M, Garcia S. A Survey of Privacy Attacks in Machine Learning[EB/OL]. 2020: arXiv: 2007.07646. https://arxiv.org/abs/2007.07646

[22] Rosenberg I, Shabtai A, Elovici Y, et al. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain[J]. *ACM Computing Surveys*, 2022, 54(5): 108.

[23] Serban A, Poll E, Visser J. Adversarial Examples on Object Recognition: A Comprehensive Survey[J]. *ACM Computing Surveys*, 2021, 53(3): 66.

[24] Yin X F, Zhu Y M, Hu J K. A Comprehensive Survey of Privacy-Preserving Federated Learning: A Taxonomy, Review, and

Future Directions[J]. *ACM Computing Surveys*, 2022, 54(6): 131.

[25] Gao T. Research Progress and Challenges of Membership Inference Attacks in Machine Learning[J]. *Operations Research and Fuzziology*, 2022(1): 1-15.
(高婷. 机器学习成员推理攻击研究进展与挑战[J]. *运筹与模糊学*, 2022(1): 1-15.)

[26] Wang L L, Zhang P, Yan Z, et al. A Survey on Membership Inference on Training Datasets in Machine Learning[J]. *Cyberspace Security*, 2019, 10(10): 1-7.
(王璐璐, 张鹏, 闫峥, 等. 机器学习训练数据集的成员推理综述[J]. *网络空间安全*, 2019, 10(10): 1-7.)

[27] Hu H S, Salcic Z, Sun L C, et al. Membership Inference Attacks on Machine Learning: A Survey[J]. *ACM Computing Surveys*, 2022, 54(11s): 235.

[28] Gupta U, Stripelis D, Lam P K, et al. Membership Inference Attacks on Deep Regression Models for Neuroimaging[C]. *Medical Imaging with Deep Learning*, 2021: 228-251.

[29] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership Inference Attacks Against Generative Models[EB/OL]. 2017: arXiv: 1705.07663. https://arxiv.org/abs/1705.07663

[30] Grover A, Leskovec J. Node2vec: Scalable Feature Learning for Networks[J]. *KDD*: *Proceedings International Conference on Knowledge Discovery & Data Mining*, 2016: 855-864.

[31] Zhang G, Liu B, Zhu T, et al. Label-Only Membership Inference Attacks and Defenses In Semantic Segmentation Models[J]. *IEEE Transactions on Dependable and Secure Computing*, 2022: 1545-5971.

[32] Rezaei S, Liu X. An Efficient Subpopulation-Based Membership Inference Attack[EB/OL]. 2022: arXiv: 2203.02080. https://arxiv.org/abs/2203.02080

[33] Tan J, Mason B, Javadi H, et al. Parameters or Privacy: A Provable Tradeoff between Overparameterization and Membership Inference[EB/OL]. 2022: arXiv: 2202.01243. https://arxiv.org/abs/2202.01243

[34] Yuan X Y, Zhang L. Membership Inference Attacks and Defenses in Neural Network Pruning[EB/OL]. 2022: arXiv: 2202.03335. https://arxiv.org/abs/2202.03335

[35] Gomrokchi M, Amin, Aboutalebi H, et al. Where did You Learn that From? Surprising Effectiveness of Membership Inference Attacks Against Temporally Correlated Data in Deep Reinforcement Learning[EB/OL]. 2021: arXiv: 2109.03975. https://arxiv.org/abs/2109.03975

[36] Ye J Y, Maddi A, Murakonda S K, et al. Enhanced Membership Inference Attacks Against Machine Learning Models[EB/OL]. 2021: arXiv: 2111.09679. https://arxiv.org/abs/2111.09679

[37] https://blog.tensorflow.org/2020/06/introducing-new-privacy-testing-library. Html.

[38] https://github.com/privacytrustlab/ml_privacy_meter.

[39] Zhong D, Sun H P, Xu J, et al. Understanding Disparate Effects of Membership Inference Attacks and Their Countermeasures[C]. *The 2022 ACM on Asia Conference on Computer and Communications Security*, 2022: 959-974.

[40] Ruiz de Arcaute G M, Hernández J A, Reviriego P, et al. Assessing the impact of membership inference attacks on classical machine learning algorithms[C]. *2022 18th International Conference on the Design of Reliable Communication Networks*, 2022: 1-4.

[41] Li S H, Wang Y J, Li Y Z, et al. L-Leaks: Membership Inference Attacks with Logits[EB/OL]. 2022: arXiv: 2205.06469. https://arxiv.org/abs/2205.06469

[42] Li G Y, Rezaei S, Liu X. User-Level Membership Inference Attack Against Metric Embedding Learning[EB/OL]. 2022: arXiv: 2203.02077. https://arxiv.org/abs/2203.02077

[43] del Grosso G, Jalalzai H, Pichler G, et al. Leveraging adversarial examples to quantify membership information leakage[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 10389-10399.

[44] Pedersen J, Muñoz-Gómez R, Huang J N, et al. LTU Attacker for Membership Inference[J]. *Algorithms*, 2022, 15(7): 254.

[45] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. 2013: arXiv: 1312.6199. https://arxiv.org/abs/1312.6199

[46] Pyrgelis A, Troncoso C, de Cristofaro E. Measuring Membership Privacy on Aggregate Location Time-Series[J]. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2020, 4(2): 1-28.

[47] Carlini N, Liu C, Erlingsson Ú, et al. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks[C]. *The 28th USENIX Conference on Security Symposium*, 2019: 267-284.

[48] Murakonda S K, Shokri R. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning[EB/OL]. 2020: arXiv: 2007.09339. https://arxiv.org/abs/2007.09339

[49] Song C Z, Ristenpart T, Shmatikov V. Machine Learning Models that Remember too much[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 587-601.

[50] Zhang C Y, Bengio S, Hardt M, et al. Understanding Deep Learning (still) Requires Rethinking Generalization[J]. *Communications of the ACM*, 2021, 64(3): 107-115.

[51] Backes M, Berrang P, Humbert M, et al. Membership Privacy in microRNA-Based Studies[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 319-330.

[52] Long Y H, Wang L, Bu D Y, et al. A pragmatic approach to membership inferences on machine learning models[C]. *2020 IEEE European Symposium on Security and Privacy*, 2020: 521-534.

[53] Long Y H, Bindschaedler V, Wang L, et al. Understanding Membership Inferences on Well-Generalized Learning Models[EB/OL]. 2018: arXiv: 1802.04889. https://arxiv.org/abs/1802.04889

[54] 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 2019 Network and Distributed System Security Symposium*, 2019: 1-15.

[55] 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.

[56] Chen M, Zhang Z K, Wang T H, et al. When Machine Unlearning Jeopardizes Privacy[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 896-911.

[57] Shokri R, Strobel M, Zick Y. On the privacy risks of model explanations[C]. *The 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021: 231-241.

[58] Liu G Y, 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.

[59] Chen H X, Li H W, Dong G S, et al. Practical Membership Inference Attack Against Collaborative Inference in Industrial IoT[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(1): 477-487.

[60] Song C Z, Shmatikov V. Auditing Data Provenance in Text-Generation Models[C]. *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019: 196-206.

[61] Wang K H, Hu Z X, Ai Q S, et al. Membership Inference Attack with Multi-Grade Service Models in Edge Intelligence[J]. *IEEE Network*, 2021, 35(1): 184-189.

[62] Irolla P, Châtel G, Communication N A B T, et al. Demystifying the membership inference attack[C]. *2019 12th CMI Conference on Cybersecurity and Privacy*, 2020: 1-7.

[63] Bentley J W, Gibney D, Hoppenworth G, et al. Quantifying Membership Inference Vulnerability via Generalization Gap and other Model Metrics[EB/OL]. 2020: arXiv: 2009.05669. https://arxiv.org/abs/2009.05669

[64] Sablayrolles A, Douze M, Ollivier Y, et al. White-Box Vs Black-Box: Bayes Optimal Strategies for Membership Inference[EB/OL]. 2019: arXiv: 1908.11229. https://arxiv.org/abs/1908.11229

[65] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]. *2018 IEEE 31st Computer Security Foundations Symposium*, 2018: 268-282.

[66] Choquette-Choo C A, Tramer F, Carlini N, et al. Label-only Membership Inference Attacks[EB/OL]. 2020: arXiv: 2007.14321. https://arxiv.org/abs/2007.14321

[67] Li Z, Zhang Y. Membership Leakage in Label-only Exposures[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 880-895.

[68] Rahimian S, Orekondy T, Fritz M. Sampling Attacks: Amplification of Membership Inference Attacks by Repeated Queries[EB/OL]. 2020: arXiv: 2009.00395. https://arxiv.org/abs/2009.00395

[69] Song L W, Mittal P. Systematic Evaluation of Privacy Risks of Machine Learning Models[EB/OL]. 2020: arXiv: 2003.10595. https://arxiv.org/abs/2003.10595

[70] Hui B, Yang Y C, Yuan H L, et al. Practical blind membership inference attack via differential comparisons[C]. *Proceedings 2021 Network and Distributed System Security Symposium*, 2021.

[71] Jayaraman B, Wang L X, Knipmeyer K, et al. Revisiting Membership Inference under Realistic Assumptions[EB/OL]. 2020: arXiv: 2005.10881. https://arxiv.org/abs/2005.10881

[72] Hilprecht B, Härterich M, Bernau D. Monte Carlo and Reconstruction Membership Inference Attacks Against Generative Models[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(4): 232-249.

[73] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership Inference Attacks Against Generative Models[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(1): 133-152.

[74] Chen D F, Yu N, Zhang Y, et al. GAN-Leaks: A Taxonomy of Membership Inference Attacks Against Generative Models[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 343-362.

[75] Leino K, Fredrikson M. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference[C]. *The 29th USENIX Conference on Security Symposium*, 2020: 1605-1622.

[76] Nasr M, Shokri R, Houmansadr A, et al. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 739-753.

[77] Zhang J W, Zhang J L, Chen J J, et al. GAN enhanced membership inference: A passive local attack in federated learning[C]. *ICC 2020 - 2020 IEEE International Conference on Communications*, 2020: 1-6.

[78] Chen J L, Zhang J L, Zhao Y C, et al. Beyond model-level membership privacy leakage: An adversarial approach in federated learning[C]. *2020 29th International Conference on Computer Communications and Networks*, 2020: 1-9.

[79] Hu H S, Salcic Z, Sun L C, et al. Source inference attacks in federated learning[C]. *2021 IEEE International Conference on Data Mining*, 2022: 1102-1107.

[80] Melis L, Song C Z, de Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 691-706.

[81] Wang Z B, Song M K, Zhang Z F, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning[C]. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019: 2512-2520.

[82] Zhang Z K, Chen M, Backes M, et al. Inference Attacks Against Graph Neural Networks[EB/OL]. 2021: arXiv: 2110.02631. https://arxiv.org/abs/2110.02631

[83] Yaghini M, Kulynych B, Troncoso C. Disparate Vulnerability: On the Unfairness of Privacy Attacks Against Machine Learning[J]. 2019: arXiv preprint arXiv:1906.00389.

[84] Rezaei S, Liu X, Processing C A. On the difficulty of membership inference attacks[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 7888-7896.

[85] Song C, Shmatikov V. The natural auditor: How to tell if someone used your words to train their model[J]. 2018: arXiv preprint arXiv:1811.00513.

[86] Long Y H, Bindschaedler V, Gunter C A. Towards Measuring Membership Privacy[EB/OL]. 2017: arXiv: 1712.09136. https://arxiv.org/abs/1712.09136

[87] Song C Z, Shokri R. Robust Membership Encoding: Inference Attacks and Copyright Protection for Deep Learning[EB/OL]. 2019: arXiv: 1909.12982. https://arxiv.org/abs/1909.12982

[88] Humphries T, Rafuse M, Tulloch L, et al. Differentially Private Learning does not Bound Membership Inference[EB/OL]. 2020: arXiv: 2010.12112. https://arxiv.org/abs/2010.12112

[89] Song L W, Shokri R, Mittal P, et al. Membership inference attacks against adversarially robust deep learning models[C]. *2019 IEEE Security and Privacy Workshops*, 2019: 50-56.

[90] Truex S, Liu L, Gursoy M E, et al. Effects of differential privacy

and data skewness on membership inference vulnerability[C]. *2019 First IEEE International Conference on Trust*, *Privacy and Security in Intelligent Systems and Applications*, 2020: 82-91.

[91] 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.

[92] Li J C, Li N H, Ribeiro B. Membership Inference Attacks and Defenses in Classification Models[C]. *The Eleventh ACM Conference on Data and Application Security and Privacy*, 2021: 5-16.

[93] Kaya Y, Hong S, Dumitras T. On the Effectiveness of Regularization Against Membership Inference Attacks[EB/OL]. 2020: arXiv: 2006.05336. https://arxiv.org/abs/2006.05336

[94] Liu Y G, Wen R, He X L, et al. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models[EB/OL]. 2021: arXiv: 2102.02551. https://arxiv.org/abs/2102.02551

[95] Chang H Y, Shokri R, Communication N A B T, et al. On the privacy risks of algorithmic fairness[C]. *2021 IEEE European Symposium on Security and Privacy*, 2021: 292-303.

[96] He X L, Wen R, Wu Y X, et al. Node-Level Membership Inference Attacks Against Graph Neural Networks[EB/OL]. 2021: arXiv: 2102.05429. https://arxiv.org/abs/2102.05429

[97] Olatunji I E, Nejdl W, Khosla M, et al. Membership inference attack on graph neural networks[C]. *2021 Third IEEE International Conference on Trust*, *Privacy and Security in Intelligent Systems and Applications*, 2022: 11-20.

[98] Tramèr F, Atlidakis V, Geambasu R, et al. FairTest: discovering unwarranted associations in data-driven applications[C]. *2017 IEEE European Symposium on Security and Privacy*, 2017: 401-416.

[99] Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles[C]. *2022 IEEE Symposium on Security and Privacy*, 2022: 1897-1914.

[100] Mahloujifar S, Sablayrolles A, Cormode G, et al. Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms[EB/OL]. 2022: arXiv: 2204.06106. https://arxiv.org/abs/2204.06106

[101] Duddu V, Szyller S, Asokan N. SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning[EB/OL]. 2021: arXiv: 2112.02230. https://arxiv.org/abs/2112.02230

[102] Yang Y H, Gohari P, Topcu U. On the Privacy Risks of Deploying Recurrent Neural Networks in Machine Learning Models[EB/OL]. 2021: arXiv: 2110.03054. https://arxiv.org/abs/2110.03054

[103] Song C Z, Raghunathan A. Information Leakage in Embedding Models[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 377-390.

[104] Mahloujifar S, Inan H A, Chase M, et al. Membership Inference on Word Embedding and beyond[EB/OL]. 2021: arXiv: 2106.11384. https://arxiv.org/abs/2106.11384

[105] Duddu V, Boutet A, Shejwalkar V. Quantifying Privacy Leakage in Graph Embedding[C]. *MobiQuitous'20*: *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems*: *Computing*, *Networking and Services*, 2020: 76-85.

[106] Thomas A, Adelani D I, Davody A, et al. Investigating the Impact of Pre-Trained Word Embeddings on Memorization in Neural Networks[C]. *Text*, *Speech*, *and Dialogue*: *23rd International Conference*, *TSD 2020*, *Brno*, *Czech Republic*, *September 8-11*, *2020*, *Proceedings*, 2020: 273-281.

[107] Liu H B, Jia J Y, Qu W J, et al. EncoderMI: Membership Inference Against Pre-Trained Encoders in Contrastive Learning[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 2081-2095.

[108] Liu K S, Xiao C W, Li B, et al. Performing Co-membership attacks against deep generative models[C]. *2019 IEEE International Conference on Data Mining*, 2020: 459-467.

[109] Ha H, Jang J, Jeong Y, et al. Membership Feature Disentanglement Network[C]. *The 2022 ACM on Asia Conference on Computer and Communications Security*, 2022: 364-376.

[110] Gu Y H, Bai Y B, Xu S B. CS-MIA: Membership Inference Attack Based on Prediction Confidence Series in Federated Learning[J]. *Journal of Information Security and Applications*, 2022, 67: 103201.

[111] Zhang Z X, Zhang L Y, Zheng X F, et al. Evaluating Membership Inference through Adversarial Robustness[EB/OL]. 2022: arXiv: 2205.06986. https://arxiv.org/abs/2205.06986

[112] Pichler G, Romanelli M, Vega L R, et al. Perfectly Accurate Membership Inference by a Dishonest Central Server in Federated Learning[EB/OL]. 2022: arXiv: 2203.16463. https://arxiv.org/abs/2203.16463

[113] Watson L, Guo C, Cormode G, et al. On the Importance of Difficulty Calibration in Membership Inference Attacks[EB/OL]. 2021: arXiv: 2111.08440. https://arxiv.org/abs/2111.08440

[114] Wu B, Zhao S, Chen C, et al. Generalization in generative adversarial networks: a novel perspective from privacy protection[J]. *Proceedings of the 33rd International Conference on Neural Information Processing Systems,* 2019: 307-317.

[115] Mukherjee S, Xu Y X, Trivedi A, et al. PrivGAN: Protecting GANs from Membership Inference Attacks at Low Cost to Utility[J]. *Proceedings on Privacy Enhancing Technologies*, 2021, 2021(3): 142-163.

[116] Webster R, Rabin J, Simon L, et al. This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces[EB/OL]. 2021: arXiv: 2107.06018. https://arxiv.org/abs/2107.06018

[117] Wunderlich D, Bernau D, Aldà F, et al. On the Privacy-Utility Trade-off in Differentially Private Hierarchical Text Classification[EB/OL]. 2021: arXiv: 2103.02895. https://arxiv.org/abs/2103.02895

[118] 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.

[119] Jagannatha A, Rawat B P S, Yu H. Membership Inference Attack Susceptibility of Clinical Language Models[EB/OL]. 2021: arXiv: 2104.08305. https://arxiv.org/abs/2104.08305

[120] Carlini N, Tramer F, Wallace E, et al. Extracting Training Data from Large Language Models[EB/OL]. 2020: arXiv: 2012.07805. https://arxiv.org/abs/2012.07805

[121] He Y, Rahimian S, Schiele B, et al. Segmentations-Leak: Mem-

bership Inference Attacks and Defenses in Semantic Image Segmentation[C]. *Computer Vision- ECCV 2020*: *16th European Conference*, *Glasgow*, *UK*, *August 23-28*, *2020*, *Proceedings*, *Part XXIII*, 2020: 519-535.

[122] Shafran A, Peleg S, Hoshen Y. Membership Inference Attacks are Easier on Difficult Problems[EB/OL]. 2021: arXiv: 2102.07762. https://arxiv.org/abs/2102.07762

[123] 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*, 2021: 891-895.

[124] Miao Y T, Xue M H, Chen C, et al. The Audio Auditor: User-Level Membership Inference in Internet of Things Voice Services[J]. *Proceedings on Privacy Enhancing Technologies*, 2021, 2021(1): 209-228.

[125] Zhang M X, Ren Z C, Wang Z H, et al. Membership Inference Attacks Against Recommender Systems[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 864-879.

[126] Chen C, Wu B Z, Qiu M H, et al. A Comprehensive Analysis of Information Leakage in Deep Transfer Learning[EB/OL]. 2020: arXiv: 2009.01989. https://arxiv.org/abs/2009.01989

[127] He X L, Zhang Y. Quantifying and Mitigating Privacy Risks of Contrastive Learning[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 845-863.

[128] Wang Y, Huang L F, Yu P S, et al. Membership Inference Attacks on Knowledge Graphs[EB/OL]. 2021: arXiv: 2104.08273. https://arxiv.org/abs/2104.08273

[129] Liew S P, Takahashi T. FaceLeaks: Inference Attacks Against Transfer Learning Models via Black-Box Queries[EB/OL]. 2020: arXiv: 2010.14023. https://arxiv.org/abs/2010.14023

[130] Wu B, Yang X W, Pan S R, et al. Adapting membership inference attacks to GNN for graph classification: Approaches and implications[C]. *2021 IEEE International Conference on Data Mining*, 2022: 1421-1426.

[131] Salem A, Bhattacharya A, Backes M, et al. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning[C]. *The 29th USENIX Conference on Security Symposium*, 2020: 1291-1308.

[132] Chen M, Zhang Z K, Wang T H, et al. When Machine Unlearning Jeopardizes Privacy[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 896-911.

[133] Wu M Q, Zhang X Y, Ding J H, et al. Evaluation of Inference Attack Models for Deep Learning on Medical Data[EB/OL]. 2020: arXiv: 2011.00177. https://arxiv.org/abs/2011.00177

[134] Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 308-318.

[135] Melis L, Song C Z, de Cristofaro E, et al. Exploiting Unintended Feature Leakage in Collaborative Learning[EB/OL]. 2018: arXiv: 1805.04049. https://arxiv.org/abs/1805.04049

[136] Farokhi F, Kaafar M A. Modelling and Quantifying Membership Information Leakage in Machine Learning[EB/OL]. 2020: arXiv: 2001.10648. https://arxiv.org/abs/2001.10648

[137] Jha S K, Jha S, Ewetz R, et al. An Extension of Fano's Inequality for Characterizing Model Susceptibility to Membership Inference Attacks[EB/OL]. 2020: arXiv: 2009.08097. https://arxiv.org/abs/2009.08097

[138] Theodoridis S, Koutroumbas K. Introduction[M]. Pattern Recognition. Amsterdam: Elsevier, 2006: 1-11.

[139] Hu H L, Pang J. Membership Inference Attacks Against GANs by Leveraging Over-Representation Regions[C]. *The 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021: 2387-2389.

[140] Chen Q R, Xiang C, Xue M H, et al. Differentially Private Data Generative Models[EB/OL]. 2018: arXiv: 1812.02274. https://arxiv.org/abs/1812.02274

[141] Jayaraman B, Evans D. Evaluating Differentially Private Machine Learning in Practice[C]. *The 28th USENIX Conference on Security Symposium*, 2019: 1895-1912.

[142] Jia J Y, Salem A, Backes M, et al. MemGuard: Defending Against Black-Box Membership Inference Attacks via Adversarial Examples[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 259-274.

[143] Naseri M, Hayes J, De Cristofaro E. Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy[J]. 2020: arXiv preprint arXiv:2009.03561.

[144] Shejwalkar V, Houmansadr A. Membership Privacy for Machine Learning Models through Knowledge Transfer[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(11): 9549-9557.

[145] Ying Z B, Zhang Y, Liu X M. Privacy-Preserving in Defending Against Membership Inference Attacks[C]. *The 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 2020: 61-63.

[146] Hintersdorf D, Struppek L, Kersting K. To Trust or not to Trust Prediction Scores for Membership Inference Attacks[EB/OL]. 2021: arXiv: 2111.09076. https://arxiv.org/abs/2111.09076

[147] Hu L, Li J, Lin G B, et al. Defending Against Membership Inference Attacks with High Utility by GAN[J]. *IEEE Transactions on Dependable and Secure Computing*, PP(99): 1.

[148] Chen D F, Yu N, Fritz M. RelaxLoss: Defending Membership Inference Attacks without Losing Utility[EB/OL]. 2022: arXiv: 2207.05801. https://arxiv.org/abs/2207.05801

[149] Bernau D, Robl J, Kerschbaum F. Assessing Differentially Private Variational Autoencoders under Membership Inference[EB/OL]. 2022: arXiv: 2204.07877. https://arxiv.org/abs/2204.07877

[150] Beaulieu-Jones B K, Wu Z S, Williams C, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing[J]. *Circulation: Cardiovascular Quality and Outcomes*, 2019, 12(7): e005122.

[151] Nasr M, Songi S, Thakurta A, et al. Adversary instantiation: lower bounds for differentially private machine learning[C]. *2021 IEEE Symposium on Security and Privacy*, 2021: 866-882.

[152] Triastcyn A, Faltings B. Generating Artificial Data for Private Deep Learning[C]. *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies, AAAI Spring Symposium Series*, 2019: 33-40.

[153] Xie L Y, Lin K X, Wang S, et al. Differentially Private Generative Adversarial Network[EB/OL]. 2018: arXiv: 1802.06739. https://arxiv.org/abs/1802.06739

[154] Xu C G, Ren J, Zhang D Y, et al. GANobfuscator: Mitigating In-

formation Leakage under GAN via Differential Privacy[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(9): 2358-2371.

[155] Zhang X Y, Ji S L, Wang T. Differentially Private Releasing via Deep Generative Model (Technical Report)[EB/OL]. 2018: arXiv: 1801.01594. https://arxiv.org/abs/1801.01594

[156] Alvar S R, Wang L J, Pei J, et al. Membership Privacy Protection for Image Translation Models via Adversarial Knowledge Distillation[EB/OL]. 2022: arXiv: 2203.05212. https://arxiv.org/abs/2203.05212

[157] Chen Z Q, Li H W, Hao M, et al. Enhanced Mixup Training: A Defense Method Against Membership Inference Attack[M]. Information Security Practice and Experience. Cham: Springer International Publishing, 2021: 32-45.

[158] Miao Y T, Zhao B Z H, Xue M H, et al. The Audio Auditor: Participant-Level Membership Inference in Voice-Based IoT[EB/OL]. 2019: arXiv: 1905.07082. https://arxiv.org/abs/1905.07082

[159] Hu H S, Salcic Z, Dobbie G, et al. EAR: an enhanced adversarial regularization approach against membership inference attacks[C]. *2021 International Joint Conference on Neural Networks*, 2021: 1-8.

[160] Kaya Y, Dumitras T. When Does Data Augmentation Help With Membership Inference Attacks?[C]. *International conference on machine learning*, 2021: 5345-5355.

[161] Nasr M, Shokri R, Houmansadr A. Machine Learning with Membership Privacy Using Adversarial Regularization[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018: 634-646.

[162] 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.

[163] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2818-2826.

[164] Li Z, Zhang Y. Label-leaks: Membership inference attack with label[J]. 2020: arXiv preprint arXiv:2007.15528.

[165] Caruana R, Lawrence S, Giles L. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping[C]. *The 13th International Conference on Neural Information Processing Systems*, 2000: 381-387.

[166] Hagestedt I, Zhang Y, Humbert M, et al. MBeacon: privacy-preserving beacons for DNA methylation data[C]. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019: 72-87.

[167] Yao Y, Rosasco L, Caponnetto A. On Early Stopping in Gradient Descent Learning[J]. *Constructive Approximation*, 2007, 26(2): 289-315.

[168] Bagmar A, Maiya S R, Bidwalka S, et al. Membership Inference Attacks on Lottery Ticket Networks[EB/OL]. 2021: arXiv: 2108.03506. https://arxiv.org/abs/2108.03506

[169] Jagielski M, Wu S, Oprea A, et al. How to Combine Membership-Inference Attacks on Multiple Updated Models[EB/OL]. 2022: arXiv: 2205.06369. https://arxiv.org/abs/2205.06369

[170] Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? membership inference on aggregate location data[C]. *Proceedings 2018 Network and Distributed System Security Symposium*, 2018: 1-15.

[171] Yang R K, Ma J F, Miao Y B, et al. Privacy-Preserving Generative Framework Against Membership Inference Attacks[EB/OL]. 2022: arXiv: 2202.05469. https://arxiv.org/abs/2202.05469

[172] Jarin I, Eshete B. MIAShield: Defending Membership Inference Attacks via Preemptive Exclusion of Members[EB/OL]. 2022: arXiv: 2203.00915. https://arxiv.org/abs/2203.00915

[173] Zheng J X, Cao Y Z, Wang H P. Resisting Membership Inference Attacks through Knowledge Distillation[J]. *Neurocomputing*, 2021, 452: 114-126.

[174] Bernau D, Robl J, Grassal P W, et al. Comparing Local and Central Differential Privacy Using Membership Inference Attacks[M]. Data and Applications Security and Privacy XXXV. Cham: Springer International Publishing, 2021: 22-42.

[175] Chen J J, Wang W H, Gao H C, et al. PAR-GAN: Improving the Generalization of Generative Adversarial Networks Against Membership Inference Attacks[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 127-137.

[176] Chen J J, Wang W H, Shi X H. Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data[J]. *Pacific Symposium on Biocomputing*, 2021, 26: 26-37.

[177] Duddu V, Boutet A, Shejwalkar V. GECKO: Reconciling Privacy, Accuracy and Efficiency in Embedded Deep Learning[EB/OL]. 2020: arXiv: 2010.00912. https://arxiv.org/abs/2010.00912

[178] Grosse K, Smith M T, Backes M, et al. Killing four birds with one Gaussian process: The relation between different test-time attacks[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 4696-4703.

[179] Hou J H, Qian J W, Wang Y, et al. ML Defense: Against Prediction API Threats in Cloud-Based Machine Learning Service[C]. *The International Symposium on Quality of Service*, 2019: 1-10.

[180] Galinkin E. The Influence of Dropout on Membership Inference in Differentially Private Models[EB/OL]. 2021: arXiv: 2103.09008. https://arxiv.org/abs/2103.09008

[181] 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.

[182] Tang X Y, Mahloujifar S, Song L W, et al. Mitigating Membership Inference Attacks by Self-Distillation through a Novel Ensemble Architecture[EB/OL]. 2021: arXiv: 2110.08324. https://arxiv.org/abs/2110.08324

[183] Tonni S M, Vatsalan D, Farokhi F, et al. Data and Model Dependencies of Membership Inference Attack[EB/OL]. 2020: arXiv: 2002.06856. https://arxiv.org/abs/2002.06856

[184] Tople S, Sharma A, Nori A V. Alleviating Privacy Attacks via Causal Learning[C]. *The 37th International Conference on Machine Learning*, 2020: 9537-9547.

[185] Wang Y J, Wang C H, Wang Z G, et al. Against membership inference attack: Pruning is all You need[C]. *The Thirtieth International Joint Conference on Artificial Intelligence*, 2021: 1-7.

[186] Webster R, Rabin J, Simon L, et al. Generating private data surrogates for vision related tasks[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 263-269.

[187] Wu B Z, Chen C C, Zhao S W, et al. Characterizing Membership Privacy in Stochastic Gradient Langevin Dynamics[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(4): 6372-6379.

[188] Yang Z Q, Shao B, Xuan B H, et al. Defending Model Inversion and Membership Inference Attacks via Prediction Purification[EB/OL]. 2020: arXiv: 2005.03915. https://arxiv.org/abs/2005.03915

[189] Yin Y, Chen K, Shou L D, et al. Defending Privacy Against more Knowledgeable Membership Inference Attackers[C]. *The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021: 2026-2036.

[190] Yu D, Zhang H S, Chen W, et al. How does Data Augmentation Affect Privacy in Machine Learning? [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(12): 10746-10753.

[191] Zhang T W, He Z C, Lee R B. Privacy-Preserving Machine Learning through Data Obfuscation[EB/OL]. 2018: arXiv: 1807.01860. https://arxiv.org/abs/1807.01860

[192] Zou Y, Zhang Z K, Backes M, et al. Privacy Analysis of Deep Learning in the Wild: Membership Inference Attacks Against Transfer Learning[EB/OL]. 2020: arXiv: 2009.04872. https://arxiv.org/abs/2009.04872

[193] Hidano S, Murakami T, Kawamoto Y, et al. TransMIA: membership inference attacks using transfer shadow training[C]. *2021 International Joint Conference on Neural Networks*, 2021: 1-10.

[194] Hitaj B, Ateniese G, Perez-Cruz F. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 603-618.

[195] Shejwalkar V, Inan H A, Houmansadr A, et al. Membership inference attacks against nlp classification models[C]. *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021: 1-13.

[196] Hanzlik L, Zhang Y, Grosse K, et al. MLCapsule: guarded offline deployment of machine learning as a service[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021: 3295-3304.

[197] Rezaei S, Shafiq Z, Liu X. Accuracy-Privacy Trade-off in Deep Ensemble: A Membership Inference Perspective[EB/OL]. 2021: arXiv: 2105.05381. https://arxiv.org/abs/2105.05381

[198] Zhao B Z H, Agrawal A, Coburn C, et al. On the (In)feasibility of attribute inference attacks on machine learning models[C]. *2021 IEEE European Symposium on Security and Privacy*, 2021: 232-251.

[199] Zhao B Z H, Asghar H J, Bhaskar R, et al. On Inferring Training Data Attributes in Machine Learning Models[EB/OL]. 2019: arXiv: 1908.10558. https://arxiv.org/abs/1908.10558

[200] Lee K C, Ho J, Kriegman D J. Acquiring Linear Subspaces for Face Recognition under Variable Lighting[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(5): 684-698.

牛俊 于 2016 年在西安电子科技大学信息安全专业获得硕士学位。现在西安电子科技大学计算机科学与技术专业攻读博士学位。研究领域为人工智能安全。研究兴趣包括: 人工智能安全。Email: niujun@stu.xidian.edu.cn

马骁骥 于 2022 年在中北大学软件工程专业获得学士学位。现在海南大学电子信息专业攻读硕士学位。研究领域为网络空间安全。研究兴趣包括网络攻击与防范、安全漏洞的挖掘与利用。Email: 835563116@qq.com

陈颖 于 2021 年在皖西学院计算机科学与技术专业获得学士学位。现在海南大学电子信息专业攻读硕士学位。研究领域为网络空间安全。研究兴趣包括: 网络攻击、网络防御。Email: 1286572504@qq.com.

张歌 于 2020 年在东北电力大学信息与计算科学专业获得学士学位。现在西安电子科技大学网络与信息安全专业攻读硕士学位, 研究领域为迁移场景下的对抗样本攻击。研究兴趣包括 AI 安全、物联网安全。Email: zhangg@nipc.org.cn

何志鹏 于 2020 年在西安邮电大学电子信息工程专业获得学士学位, 现在西安邮电大学网络与信息安全专业攻读硕士学位。研究领域为机器学习下的模型架构安全。研究兴趣包括 AI 安全。Email: hezhp@nipc.org.cn

侯哲贤 于 2020 年在河南农业大学电子信息科学与技术(网络信息技术)专业获得学士学位。现在西安电子科技大学电子信息专业攻读硕士学位, 研究领域为迁移场景下数据投毒攻击。研究兴趣包括 AI 安全、物联网安全。Email: houzhx@nipc.org.cn

朱笑岩 于 2009 年在西安电子科技大学获得博士学位。现任西安电子科技大学教授, 博士生导师。主要研究方向为大数据、移动互联网、云计算、在线社交网络、机器学习、车辆自组网、推荐系统中的隐私安全。Email: xyzhu@mail.xidian.edu.cn

伍高飞 于 2015 年在西安电子科技大学获得博士学位。现任西安电子科技大学讲师, 硕士生导师。主要研究方向为网络与信息系统安全、AI 安全、密码学。Email: wugf@nipc.org.cn

陈恺 于 2010 年在中国科学院研究生院信息安全专业获得博士学位。现为中国科学院信息工程研究所研究员。研究领域为系统安全、人工智能安全。Email: chenkai@iie.ac.cn

张玉清 于 2000 年在西安电子科技大学获得博士学位。现任中国科学院大学教授, 博士生导师。主要研究方向为网路与信息系统安全。Email: zhangyq@nipc.org.cn