黑盒机器学习模型的成员推断攻击研究

刘高扬[1], 李雨桐[1], 万博睿[1], 王 琛[1,2], 彭 凯[1,2]

[1]华中科技大学电子信息与通信学院 武汉 中国 430074

[2]华中科技大学智能互联网技术湖北省重点实验室 武汉 中国 430074

摘要 近年来, 机器学习技术飞速发展, 并在自然语言处理、图像识别、搜索推荐等领域得到了广泛的应用。然而, 现有大量开放部署的机器学习模型在模型安全与数据隐私方面面临着严峻的挑战。本文重点研究黑盒机器学习模型面临的成员推断攻击问题, 即给定一条数据记录以及某个机器学习模型的黑盒预测接口, 判断此条数据记录是否属于给定模型的训练数据集。为此, 本文设计并实现了一种基于变分自编码器的数据合成算法, 用于生成与给定模型的原始训练数据分布相近的合成数据; 并在此基础上提出了基于生成对抗网络的模拟模型构建算法, 利用合成数据训练得到与给定模型具有相似预测能力的机器学习模型。相较于现有的成员推断攻击工作, 本文所提出的推断攻击无需目标模型及其训练数据的先验知识, 在仅有目标模型黑盒预测接口的条件下, 可获得更加准确的攻击结果。通过本地模型和线上机器学习即服务平台 BigML 的实验结果证明, 所提的数据合成算法可以得到高质量的合成数据, 模拟模型构建算法可以在更加严苛的条件下模拟给定模型的预测能力。在没有目标模型及其训练数据的先验知识条件下, 本文所提的成员推断攻击在针对多种目标模型进行攻击时, 推断准确率最高可达 74%, 推断精确率可达 86%; 与现有最佳攻击方法相比, 将推断准确率与精确率分别提升 10.7%及 11.2%。

关键词 机器学习; 黑盒模型; 成员推断攻击; 变分自编码器; 生成对抗网络

中图法分类号 TP181; TP309 DOI 号 10.19363/J.cnki.cn10-1380/tn.2021.05.01

Membership Inference Attacks in Black-box Machine Learning Models

LIU Gaoyang[1], LI Yutong[1], WAN Borui[1], WANG Chen[1,2], PENG Kai[1,2]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

[2] Internet Technology and Engineering R&D Center (ITEC), Huazhong University of Science and Technology, Wuhan 430074, China

Abstract In recent years, machine learning has developed rapidly and has been widely deployed in the fields of natural language processing, image recognition, and search recommendations. However, a large number of machine learning models in the wild are facing severe challenges in terms of model security and data privacy. This paper focuses on the member inference attack against black-box machine learning models: given a data record and the black-box prediction interface of a machine learning model, the aim is to determine whether the data record was used to train the target model or not. To this end, in this paper, we design a synthetic data generation algorithm based on VAE and implement it to generate a synthetic dataset which has a similar distribution with the original training data of the given model. In addition, a mimic model construction algorithm based on the generated adversary network is proposed, which can train a mimic machine learning model that can imitate the prediction behavior of the target model by using the synthetic data. Compared with the existing works of member inference attacks, the inference attacks proposed in this paper do not require the prior knowledge about the target model and its training data, and can achieve more accurate attack results only with the black-box access to the target model. Experimental results show that the data synthesis algorithm proposed in this paper can obtain high quality synthetic data. The minic model construction algorithm can simulate the predictive power of a given model under more stringent conditions. Without prior knowledge about the target model and its training data, the proposed membership inference attack against multiple target models can achieve the highest attack accuracy and precision of 74% and 86% respectively, which are 10.7% and 11.2% higher than the state-of-the-art attack method.

Key words machine learning; black-box model; membership inference attack; variational autoencoder; generative adversarial network

通讯作者: 王琛, 博士, 副研究员. Email: chenwang@hust.edu.cn。

本课题得到国家自然科学基金(No. 61872416, No. 62002104, No. 52031009, No. 62071192)、中央高校基本科研业务费(No. 2019kfyXJJS017)、湖北省自然科学基金(No. 2019CFB191)、国家大学生创新训练计划项目(No. 2020104870001, No. DX2020041)资助。

收稿日期: 2020-07-12; 修改日期: 2020-09-30; 定稿日期: 2021-03-05

1 引言

近年来, 得益于计算设备功能的持续突破, 机器学习在数据挖掘、计算机视觉、自然语言处理等领域得到了广泛的应用并取得了显著的成效。相比于传统数据挖掘技术, 机器学习技术的出现提升了计算机系统处理数据和理解数据的能力[1-4]。

广泛部署的机器学习模型在为用户提供巨大便利的同时, 也面临着泄露用户隐私的风险。例如, 为了构建一个监测某类异常行为的机器学习模型, 研究者需要向模型输入包含大量此类行为的样本作为训练数据。然而, 在该模型的训练以及发布过程中, 攻击者可利用技术手段获取模型内部信息, 并以此推断用户的部分敏感特征, 从而破坏用户的隐私安全。因此, 在开发机器学习模型时如何保证模型的安全性以及如何防止敏感数据泄露, 已经成为国内外研究者共同关注的焦点问题[5-7]。

本文重点关注针对机器学习模型的成员推断攻击: 即给定一个机器学习模型以及一个数据样本, 推断该样本是否是该模型训练数据的一部分。成员推断攻击可威胁为模型提供训练数据的用户的隐私, 也会侵犯机器学习模型所有者的利益。例如, 准确推断某个病人的临床记录是否被用来训练与某种疾病相关的机器学习模型, 可揭示该病人是否患有这种疾病并且侵害该模型所有者的权益。因而, 成员推断攻击可作为机器学习模型中隐私和数据安全防护设计中的重要参考和度量[8-10]。

尽管现有成员推断攻击已经取得了较好的攻击效果[9-14], 但其攻击所需条件在实际机器学习应用场景中难以得到满足。例如, 部分工作[11,13]假设攻击者拥有与目标模型训练集具有相同分布的数据, 而通常情况下攻击者很难直接获得具有如此特性的数据集。此外, 部分攻击[11,14]要求目标模型被部署为白盒或灰盒模型。在此条件下, 攻击者可获得目标模型所使用的云训练平台的相关信息(即灰盒模型), 或直接获得目标模型的训练算法、内部参数、模型结构、中间结果等信息(即白盒模型), 从而构建与目标模型预测能力相似的模型。然而在现实中, 机器学习模型通常被部署为黑盒模型, 即攻击者仅能查询目标模型的预测接口得到预测输出, 而无法获得目标模型及其训练数据的任何先验信息。到目前为止, 尚无相关工作针对黑盒机器学习模型的成员推断攻击提出有效的攻击方法。

本文设计了一种成员推断攻击方法, 在无需获得目标模型的训练算法、参数、结构及其训练数据的统计信息的前提下, 仅利用目标模型的黑盒预测接口实现对目标模型的成员推断攻击。我们的基本想法是通过充分利用目标模型的黑盒接口合成高质量的可用训练数据, 并基于合成数据建立一个与目标模型预测行为接近的模拟模型, 再基于模拟模型实现成员推断攻击。具体而言, 本文做出了如下贡献:

1. 提出基于变分自编码器[15] (Variational Autoencoder, VAE)的合成数据生成算法。主要分为三个阶段: 首先根据目标模型输入数据的结构, 随机采样得到采样数据; 随后通过与目标模型的交互, 从随机采样数据中动态筛选出有效部分, 并利用 VAE 模型对筛选后的数据进行数据增强; 之后对增强数据再进行二次过滤, 最终得到与黑盒目标模型训练数据具有相似分布的合成数据集。

2. 提出基于生成对抗网络[16] (Generative Adversarial Networks, GAN)的模拟模型构建算法。利用上述合成数据以及目标模型的预测接口, 将 GAN 生成器网络作为模拟模型, 若判别器网络难以区分生成器与目标模型的输出, 通过二者之间的对抗训练, 构建与目标模型具有相似预测行为的模拟模型。所得的模拟模型可被视为为白盒模型, 并且其训练集中全部训练样本均为已知。

3. 利用模拟模型对其训练数据与非训练数据的预测构建成员推断攻击模型。由于成员推断问题为二分类问题, 因此任何监督分类算法均可用于训练成员推断攻击模型。最终, 使用该攻击模型与目标模型的预测接口, 实现针对黑盒目标模型的成员推断攻击。

4. 通过本地模型和线上机器学习即服务平台 BigML 的实验结果表明, 基于本文所提算法训练得到的模拟模型对于测试数据的预测结果平均相似度为 84.1%, 相似度的最优表现为 93.1%。本文所提的攻击方法在没有目标模型及其训练数据任何先验知识的条件下, 所构建的模拟模型实施的攻击在测试数据上推断准确率最高可达 74%, 推断精确率可达 86%; 与现有最佳攻击方法相比, 将推断准确率与精确率分别提升 10.7%及 11.2%。

本文后续的内容安排如下: 第 2 节对成员推断攻击的相关工作进行回顾; 第 3 节对本文所提出的成员推断攻击方法的技术基础及预备知识进行介绍; 第 4 节介绍本文所提针对黑盒机器学习模型的成员推断攻击, 分别对生成与目标模型训练集具有相似分布的合成数据、构建与目标模型具有相似预测行为的模拟模型以及训练可执行成员推断的攻击模型

进行详细描述; 第 5 节对所提攻击算法进行性能测试, 并与现有最佳攻击方法进行对比; 最后, 第 6 节对全文工作进行总结。

2 相关工作

2.1 成员推断问题的起源

早期的成员推断问题, 是在攻击者已知整体数据统计信息的前提下, 推断单个数据是否出现在混合数据中。例如, Homer 等[23]对基因组数据进行成员推断攻击, 可以准确判断给定目标个人是否隶属于与某种疾病相关的研究群体。Wang 等[17]减少了攻击者对先验信息的需求, 利用部分统计数据即可实现成员推断攻击。Backes 等[18]针对 miRNA 表达数据集的成员推断攻击, 揭示了由 RNA 数据发布引起的数据贡献者的个人隐私泄露风险。

此外, 部分研究人员针对位置服务展开成员推断攻击, 从而揭示了通过位置信息泄露用户个人隐私的风险。例如, Pyrgelis 等[19]建立博弈模型, 利用用户的先验位置信息, 通过可识别博弈过程将其转化为是否属于特定集合成员的分类问题, 在多用户的聚合位置中推断特定用户是否存在。Xu 等[20]充分利用用户移动的独特性和规律性, 无需任何先验知识即可从聚合的移动轨迹数据中恢复目标个体的轨迹, 实现成员推断攻击。

2.2 面向机器学习模型的推断攻击

2017 年 Shokri 等[11]率先提出针对机器学习模型的成员推断攻击, 该研究揭示了机器学习模型在训练数据和测试数据上预测行为的不一致性, 表明仅通过机器学习模型的输出即可侵犯训练数据的隐私。自此, 机器学习模型的原始训练数据的隐私泄露问题引起了国内外学者较多的关注。

针对灰盒机器学习模型, 部分研究人员利用有关目标模型的先验信息, 使用与目标模型相同的训练设置构建影子模型, 从而模拟目标模型的预测行为, 进而实施成员推断攻击。Shokri 等[11]使用与目标模型一致的训练设置, 构建多个影子模型, 并基于影子模型的输出训练多个推断攻击模型。Salem 等[13]沿用了影子模型的思路, 但放宽了攻击所需条件, 只使用一个影子模型和一个攻击模型, 从而大大降低了攻击的成本, 并且可以达到比较接近的攻击效果。Liu 等[21]使用对抗训练技术构建与目标模型预测行为类似的影子模型, 并使用影子模型的预测结果构建攻击模型。Truex 等[22]使用与目标模型相同的训练设置在同一个数据集上构建影子模型, 利用目标模型的可移植性实施成员推断攻击。

针对白盒机器学习模型, 部分研究人员利用模型内部的参数或中间结果进行推断攻击。Nasr 等[23]利用深度学习中随机梯度下降算法的隐私漏洞, 基于神经节点的梯度与激活值对白盒深度学习模型实施成员推断攻击。此外, Hayes 等[24]针对生成对抗模型, 利用判别器的预测输出进行成员推断攻击。Melis 等[14]利用深度学习模型特征嵌入层的输出结果进行成员推断攻击: 若某条文本信息参与了模型的训练, 则该文本对嵌入层参数具有较小的梯度。

此外, 针对在线机器学习即服务平台[25], 也有研究人员尝试进行成员推断攻击。Shokri 等[11]表明亚马逊等公司提供的机器学习即服务平台发布的模型可能会泄露大量训练数据信息。Song 等[26]证明即使模型只提供黑盒接口发布, 使用第三方代码在敏感数据上训练机器学习模型仍然存在较大风险。

在现实中, 机器学习模型通常被部署为黑盒模型, 例如在线机器学习即服务平台, 且一般无法获得白盒模型或灰盒模型所需的信息。综合现有工作, 目前对仅基于黑盒模型预测接口能否实施成员推断攻击尚没有相关工作。

3 预备知识

3.1 变分自编码器

变分自编码器(VAE)[15,27]通常由两个部分组成, 一部分是编码器, 另一部分是解码器。VAE 模型将一个高维的输入数据通过编码器映射到一个低维的隐变量, 接着通过解码器将得到的低维隐变量解码输出得到生成数据[15]。随后将输入数据和生成数据进行比较, 通过最小化二者之间的差异来训练 VAE 模型。在整个训练过程结束之后, 通过随机采样得到隐变量, 并将该采样结果输入解码器就可以产生和原始训练数据的分布基本相似的生成数据。

如图 1 所示, VAE 输入数据 x, 编码器 $q_\varphi(z|x)$ 实现 $x \to z$ 的映射, 其中 φ 为编码器的网络参数, z 为数据隐编码。随后, 解码器 $p_\theta(z|x)$ 完成 $z \to x$ 的映射, 实现从隐编码到生成数据的重构。

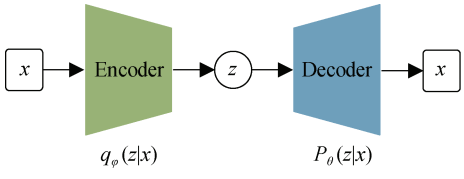

图 1 VAE 网络结构

Figure 1 Structure of VAE

为最大程度的还原输入数据, 并使隐编码服从特定分布, VAE 损失函数一般构造如下[27]:

$$l_i = -E_{z\sim q_\varphi(z|x_i)}\left[\log\left(p_\theta\left(x_i\mid z\right)\right)\right] + KL\left(q_\varphi\left(z\mid x_i\right)\|\, p\left(z\right)\right) \tag{3.1}$$

式中, 前一项是 $x\sim z\sim x$ 输入数据重建损失, 后一项 KL 散度损失[28]为模型的正则项。KL 散度是用于衡量两个概率分布之间差异的重要度量, 表示理论分布拟合真实分布时产生的信息损耗。若后验分布 $q_\varphi\left(z\mid x\right)$ 与先验分布 $p\left(z\right)$ 越接近, 则 KL 散度越小:

$$KL\left(q_\varphi\left(z\mid x_i\right)\|\, p\left(z\right)\right)=\int q_\varphi\left(z\mid x_i\right)\log\frac{q_\varphi\left(z\mid x_i\right)}{p\left(z\right)}\mathrm{d}z \tag{3.2}$$

通过 KL 正则项的约束后验分布 $q_\varphi\left(z\mid x\right)$ 向先验分布 $p\left(z\right)$ 靠近。通常假设 $p\left(z\right)$ 服从标准正态分布。由于 VAE 模型具有强大的数据生成能力, 本文将基于 VAE 模型构建合成数据生成算法, 用于合成与目标模型训练数据具有相似分布的数据。

3.2 生成对抗网络

生成对抗网络(GAN)[29-30]利用深度模型之间的对抗过程来估计生成模型。GAN 一般包含两个组成部分, 即捕获数据分布的生成模型 G(Generator)和判别数据真伪的判别模型 D(Discriminator), 其网络结构如图 2 所示。生成器 G 学习真实的数据分布, 并根据输入的随机噪音产生伪数据; 判别器 D 则对输入数据进行真伪判定。GAN 网络的训练过程即是生成器和判别器之间的博弈: G 尽可能生成接近真实的数据, 以骗过 D; 而 D 则尽可能准确地将真伪数据区分开。随着训练的进行, 最终两个网络达到动态均衡, 此时生成器生成的数据近似于真实数据。

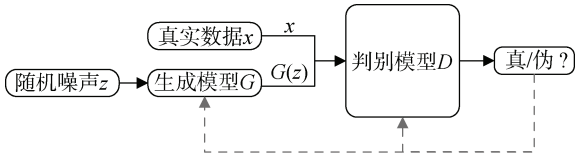

图 2 GAN 网络结构

Figure 2 Structure of GAN

G 和 D 的博弈过程描述如下: 生成模型 G 可依据随机噪声 z 生成数据 $G\left(z\right)$; 判别模型 D 的输出 x 为真实数据的概率为 $D\left(x\right)$。令 P_r 表示真实样本分布、P_g 表示生成样本分布, 则生成模型的目标函数为:

$$\min_G E_{x\sim P_g}\left[\log\left(1-D\left(x\right)\right)\right] \tag{3.3}$$

判别模型的目标函数为:

$$\min_D E_{x\sim P_r}\left[\log D\left(x\right)\right]+E_{x\sim P_g}\left[\log\left(1-D\left(x\right)\right)\right] \tag{3.4}$$

整个优化过程其实就是一个二元极大极小博弈过程, 目标函数如公式 3.5 所示:

$$\min_G\max_D V\left(D,G\right)=E_{x\sim P_r(x)}\left[\log D\left(x\right)\right]+E_{z\sim P_g(z)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \tag{3.5}$$

然而, 原始 GAN 模型具有难训练、不稳定等缺点[30], 为此, Arjovsky 等[16,31]提出 W-GAN, 利用 Wasserstein 距离度量两个概率分布之间的距离, 避免 GAN 模型中存在的问题。本文将使用 W-GAN 对抗训练生成可模拟目标模型预测行为的深度网络模型。

4 成员推断攻击

4.1 推断攻击概述

本文的攻击目标是: 给定一个目标样本 X_t 及一个机器学习模型的黑盒预测接口 F_t: $X\to Y$, 我们依据 F_t 对目标样本的预测结果, 推断目标样本 X_t 是否参与模型 F_t 的训练。我们的成员推断攻击主要包含三部分(如图 3 所示):

1) 影子数据生成: 为了构建与给定目标机器学习模型预测行为相似的模拟模型, 我们需要生成与目标模型训练数据相似的影子数据。首先, 我们随机生成部分数据并通过 F_t 得到其分类结果; 随后根据分类结果过滤筛选数据; 接着使用 VAE 模型对筛选后的数据进行增强扩充, 直到得到充足的模拟模型训练数据。最终我们对扩充数据再过滤, 得到高质量的影子数据。

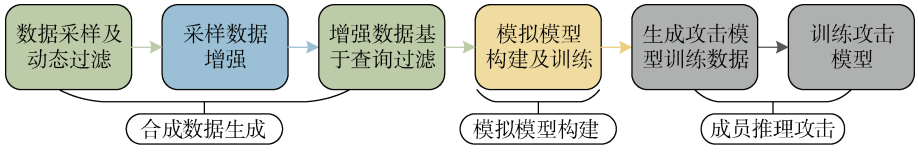

图 3 成员推断攻击基本流程

Figure 3 Procedure of Membership Inference Attack

2) 模拟模型构建: 生成影子数据后, 我们构造与目标模型具有相似预测功能的模拟模型。由于无法获得有关目标模型算法、结构、参数等信息, 我们统一使用深度网络构建模拟模型。我们对 GAN 网络进行修改, 利用其中生成器与判别器间的对抗, 使生成器的输出不断逼近目标模型的输出。当判别器难易区分目标模型和生成器对同一组数据的预测输出时, 我们将生成器提取出来, 作为模拟模型进行下一步成员推断攻击。

3) 攻击模型构建: 成员推断攻击的基本思路是利用目标模型在其训练集和测试集上的预测行为之间的差异, 因此我们将用于训练模拟模型和未用于模拟模型训练的影子数据分别输入至模拟模型中, 得到对应的分类预测结果。随后, 利用分类预测结果构建推断攻击模型。由于推断攻击本质上可被归类为二分类问题, 经典的机器学习算法均可用于训练推断攻击模型。

4.2 影子数据生成

为训练模拟模型, 首先需要获取与原始目标模型训练数据具有相似分布的影子数据。当拥有关于目标模型训练集先验信息时, 攻击者可采取不同的方式生成影子数据。若已知原始数据集合的特征分布, 可以利用这些统计信息, 通过独立地对每一个特征或是类标签进行随机抽样, 从而构造影子数据[13]; 若可以获取部分真实训练数据, 则可通过数据增强技术扩充得到影子数据; 若仅能获取目标模型的预测接口, 则仅能通过查询生成方式构造影子数据[11], 通过对预测结果对数据进行筛选。

然而, 现有成员推断攻击中的数据生成方法效率低、合成数据质量差。因此, 我们提出合成数据生成算法算法, 如图 4 所示。动态过滤噪音数据并且利用 VAE 扩充数据。在无需任何有关数据的先验信息的前提下, 可以获得高质量的影子数据。

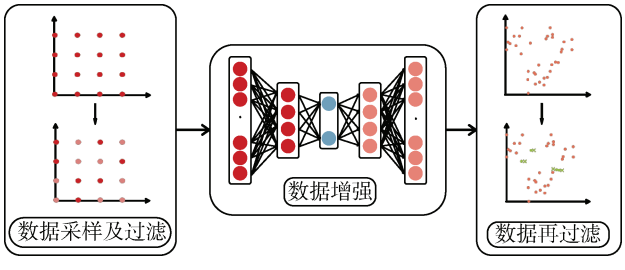

图 4 合成数据生成算法流程

Figure 4 Procedure of Synthetic Data Generation

4.2.1 数据采样及动态过滤

我们随机初始化一个数据集 X_{cand}, 其中的特征值均由均匀采样得到。如果一条数据记录位于或接近目标模型的训练数据, 则目标模型将会有较高的信息将该数据记录预测为某类。基于此, 我们设计了一种基于查询的动态数据过滤方法,具体步骤如下:

*Step*1: 利用数据集 X_{cand} 查询目标模型, 得到相应的预测结果 Y_{cand}, 根据目标模型的分类数目 C, Y_{cand} 按照预测的类别 c 分为 Y^c_{cand}, 其中 $c \in [1,2,\cdots,C]$。

*Step*2: 对于 Y^c_{cand} 进行降序排序并计算出 Y^c_{cand} 的变化速率, 选择变化速率第一个峰值所对应的概率值作为概率阈值 THR^c_P,计算出每个类别的概率阈值。

*Step*3: 如果属于类 c 的候选记录的预测概率高于 THR^c_P, 则我们筛选该候选记录到数据集 X_{smp}。由于阈值 THR_p 是根据 X_{cand} 的整体预测概率动态确定的, 所以与任何固定阈值相比, THR_p 都可以从 X_{cand} 中选择所有合适的样本。

4.2.2 数据增强

我们在上一步中采样得到了接近目标训练数据的数据集 X_{smp}。为了训练模拟模型, 需要生成与原始训练数据大小相当的样本数量。若利用上述方法生成足够数量的影子数据, 将带来巨大的时间和计算开销。为了提高数据生成模块的效率, 我们利用 VAE 模型来学习数据集 X_{smp} 的潜在分布, 然后生成类似但不同于 X_{smp} 的一个增广数据集 X_{aug}。

具体来说就是通过 X_{smp} 训练 VAE 模型, 并当训练完成时将解码器网络从 VAE 模型中分离出来。之后我们从隐空间中随机采样得到隐编码 z 并且将 z 解码得到数据集 X_{aug}。当数据集 X_{aug} 的数量与目标训练数据集的大小相当时停止数据扩充过程。

4.2.3 增强数据再过滤

上述数据增强方法将引入一些与目标模型模型训练数据的相差较大的数据样本, 因此我们需要对 X_{aug} 进行过滤, 过滤后的数据集记为 X_R。数据过滤基于以下观察: 如果 X_{aug} 中的数据记录与目标模型的训练数据接近, 那么该记录的目标模型的预测概率应该更接近于 1。具体步骤如下:

*Step*1: 利用数据集 X_{aug} 查询目标模型, 得到相

应的预测结果 Y_{aug}，根据目标模型的分类数目 C，将 Y_{aug} 按照预测的类别 c 分为 Y^c_{cand}，其中 $c \in [1,2,\cdots,C]$。

*Step*2: 对于类 c，属于 c 类的目标模型的训练数据的预测概率应接近概率向量[0, 0, …, 1, …, 0], 其中 c 类的预测概率为 1, 其他类的预测概率为 0。我们把这个概率向量表示为 Pr^c_{ref},表示 c 类的概率参考向量。计算 Y^c_{aug} 与 Pr^c_{ref} 之间的余弦距离 $D^c = dist(Y^c_{aug}, Pr^c_{ref})$，其中，距离函数选择为余弦距离函数，$D^c$ 能力表示 X^c_{aug} 和目标模型中 c 类训练数据的相似程度。

*Step*3: 设置一个固定的余弦距离阈值 THR_d (或选择一个固定数量的记录), 如果 D^c 小于阈值 THR_d，则认为此条记录和原始训练数据相似，标记为训练数据，否则，该记录被标记为测试数据。对所有的类重复上述操作，得到模拟模型的训练数据 X^R_{train} 和模拟模型的测试数据 X^R_{test}。

4.3 模拟模型构建

在得到模拟模型的训练数据 X^R_{train} 之后，接下来构建与目标模型拥有一致或相似预测能力的目标模型。由于无法获得与目标模型训练算法、模型结果、训练超参数等信息，我们使用深度网络模型构建模拟模型。为了提高模拟模型与目标模型的相似度，我们在这一步通过修改 W-GAN 网络引入对抗训练。W-GAN 网络一般由生成器和判别器两部分组成。生成器根据输入的随机噪声训练合成假数据以欺骗判别器，而判别器则尝试从合成的样本中鉴别真实的数据样本。随着生成器和判别器之间竞争的继续，两种神经网络在执行任务时表现得越来越好，最终达到二者间的均衡。

若判别器无法准确区分合成数据的预测结果是由目标模型还是生成器输出，则生成器与目标模型在相同数据上具有难以区分的预测行为，我们可将生成器作为模拟模型使用。如图 5 所示，我们用合成的 X^R_{train} 数据代替生成器的输入噪声，用目标模型对 X^R_{test} 的预测结果代替真实数据。随着生成器与判别器之间竞争的进行，生成器将逐步学习目标模型的潜在预测规则。当判别器的误差比以前减小的慢，甚至开始增大时，就停止训练过程，以生成器网络作为期望的模拟模型，记为 F_{rep}。

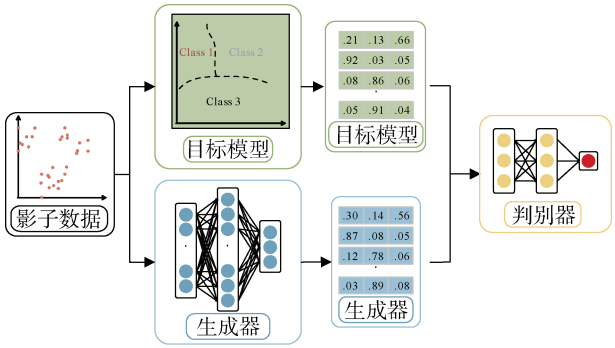

图 5 模拟模型构建算法流程

Figure 5 Procedure of Mimic Model Construction

4.4 攻击模型构建

当模拟模型训练完成后，如图 6 所示，我们将模拟模型对训练数据 X^R_{train} 的预测结果记为 Y^R_{train}，得到 (Y^R_{train}, IN)；将目标模型对测试数据 X^R_{test} 的结果记为 Y^R_{test}，得到 (Y^R_{test}, OUT)。将上述数据合并得到攻击模型训练数据 D_{attack}。由于成员推断攻击本质上为二分类问题，任何机器学习分类算法均可用于训练攻击模型 F_{attack}。当攻击模型构建完成后，对于任意给定目标数据 X_{target}，我们先将该目标数据输入至目标模型得到其对应预测结果 Y_{target}。随后将 Y_{target} 输入至攻击模型中即可得到推断结果。

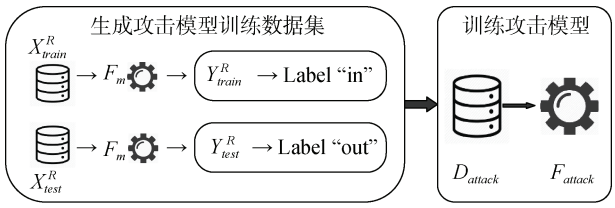

图 6 攻击模型构建流程

Figure 6 Procedure of Attack Model Construction

5 性能测试

本章在真实数据集上测试本文所提推断攻击方法的性能，并与现有攻击工作进行对比。我们将展示不同条件下对比实验的结果，并详细研究影响成员推断攻击性能的关键因素。

5.1 数据集合

本文实验所采用的数据集分别为 IMDB 电影评价数据集、Tweets 推特情绪分类数据集和 Shop 购物评论数据集。

IMDB 数据集① 该数据集是 keras 库中自带的一个数据集, 它包含 5 万条来自互联网电影数据库的严重两极分化的评论, 正面和负面评论各占 50%。

Tweets 数据集② 该数据集来源于 kaggle 网站, 是一个情绪分类数据集, 它包含了 6 个类别一共 41 万条推文, 其中分别为喜悦(13.5 万)、悲伤(11.5 万)、愤怒(5.5 万)、恐惧(5 万)、爱(4 万)、惊讶(1.5 万)。

Shop 数据集③ 该数据集共有 6 万条针对不同商品的评论, 包括十个类别分别是书籍、平板、手机、水果、洗发水、热水器、牛奶、衣服、计算机、酒店。其中热水器类别数据量较少, 我们选择其余 9 类商品的评论进行实验。

5.2 模型选择

5.2.1 目标模型

我们分别在本地和线上针对不同种类的机器学习模型进行成员推断攻击实验。

本地模型: 利用公共框架 XGBoost[32] 和 Scikit-Learn[33]提供的标准培训过程来训练不同种类的目标模型, 包括分别 SVM、XGBoost、Random Forest 和 Logistics 模型。当目标模型训练完成后, 我们将上述模型均部署为黑盒模型。

线上模型: 使用机器学习及服务平台 BigML④训练。BigML 平台仅提供少量的模型参数控制, 并且对数据所有者隐藏了训练过程。用户无法选择模型的训练算法、结构以及超参数, 所有训练过程均由平台自动完成。因此可将 BigML 训练得到的目标模型视为一个黑盒模型。

5.2.2 模拟模型

根据通用近似原理, 多层神经网络可以模拟或近似任何类型的映射。在我们的实验中利用神经网络来构建模拟模型。针对上述三种不同的数据集, 我们利用 Pytorch[34]开源框架分别搭建具有不同结构的神经网络作为模拟模型, 之后调整其网络层数以测试模拟模型规模对攻击结果的影响。

5.2.3 攻击模型

攻击模型的目标是推断给定的记录是否在目标模型的训练数据中, 此种隶属度推断攻击可以被转化为二元分类问题。我们使用不同机器学习分类算法如 XGBoost、SVM、神经网络构建攻击模型。

5.3 性能评估指标

5.3.1 攻击模型评估指标

成员推断攻击本质上为机器学习中的二分类问题, 因此我们使用机器学习中的准确率、精确率、以及召回率作为攻击模型评估指标。具体来说, 准确率表示攻击模型针对训练集与测试集进行推断的准确程度; 精确率表示被预测为训练数据集成员的数据记录确实在目标模型的训练集中的比例; 召回率表示我们可以正确推断出的训练记录在训练集中的比例。准确率或精确率可用于评估我们的推断攻击的准确程度, 而召回率可评估推断攻击对成员数据的覆盖程度。

5.3.2 模拟模型评估指标

为了比较模拟模型预测能力和目标模型预测能力的相似程度, 我们使用预测概率均方误差(MSE)与预测结果相似度(Similarity)两个指标进行评估。

预测概率均方误差(MSE) 表示在相同测试数据集上, 目标模型和模拟模型预测概率的均方误差, 其中 $y_R^{(i)}$ 和 $y_R^{(i)}$ 分别是模拟模型和目标模型的预测结果, N_{test} 是测试数据集的大小。MSE 越小说明两个模型的预测能力越接近:

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (y_R^{(i)} - y_t^{(i)})^2 \tag{5.1}$$

预测结果相似度(Similarity) 表示目标模型和模拟模型预测到同一类的数据记录与整个测试数据集的比率, 其中 N_{same} 是具有相同预测类标签的数据记录的数量。相似度越高说明两个模型的预测能力越接近:

$$Similarity = \frac{N_{same}}{N_{test}} \times 100\% \tag{5.2}$$

5.4 实验结果

5.4.1 成员推断攻击性能评估

本部分是对成员推断攻击的攻击结果进行评估, 对比实验条件如表 1 所示, 其中基线方法 A[13]为现有最佳攻击方法。我们在实验假设 A 的基础上利用不同的信息(实验假设 B、C、D)对目标模型进行模拟, 主要验证了本文提出的数据合成算法和模拟模型算法均可以提高攻击的准确率, 同时我们也讨论了不同目标模型和不同攻击模型对攻击结果的影响, 此外, 我们还探究了模拟模型和目标模型的相似程度

① https://datasets.imdbws.com/
② https://www.kaggle.com/kazanova/sentiment140
③ https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/datasets/online_shopping_10_cats
④ https://bigml.com/

对攻击结果的影响。

表 1 成员推断攻击实验假设

Table 1 Experiment Assumptions of Membership Inference Attacks

实验	实验假设
基线方法 A	假设攻击者拥有目标模型训练集的部分数据; 与Salem等人攻击假设相同, 攻击者可使用与目标模型相同的算法、训练参数构建影子模型, 随后利用影子模型的输出构建成员推断攻击模型。
攻击实验 B	假设攻击者拥有目标模型训练集的部分数据, 并利用合成数据生成算法算法进行数据增强和数据筛选; 后续步骤和Salem实验相同, 利用单个影子模型模拟目标模型并进行推断攻击。
攻击实验 C	假设攻击者拥有与目标模型训练集具有相同分布的额外数据; 利用模拟模型构建算法训练模拟模型, 随后构建攻击模型进行推断攻击。
攻击实验 D	假设攻击者拥有与目标模型训练集具有相同分布的额外数据, 并利用合成数据生成算法算法进行数据增强和数据筛选; 利用模拟模型构建算法训练模拟模型,随后构建攻击模型进行推断攻击。

a) 成员推断攻击效果整体评估

对于 IMDB 数据集来说, 基线方法 A 中平均 Acc、Pre、Rec 分别为 53.1%、54.3%、65.7%; 对比实验 B 分别为 57%、53.9%、67.9%; 对比实验 C 分别为 57.1%、55.7%、69%; 对比实验 D 分别为 58.3%、56.5%、70%。

对于 Tweet 数据集来说, 基线方法 A 中平均 Acc、Pre、Rec 分别为 57.6%、57.7%、67.7%; 对比实验 B 分别为 63.4%、58.8%、73.7%; 对比实验 C 分别为 64.2%、61%、75.5%; 对比实验 D 分别为 65.6%、62.7%、75.6%。

对于 Shop 数据集来说, 基线方法 A 中平均 Acc、Pre、Rec 分别为 61.3%、69.4%、73.1%; 对比实验 B 分别为 67.4%、71.2%、75.1%; 对比实验 C 分别为 68%、73.3%、77.7%; 对比实验 D 分别为 70%、74.5%、79%。

图 7~图 9 表示成员推断攻击在三个不同性能评估指标上的表现。攻击方法 A、B、C 和 D 的对比结果说明我们的数据合成算法和模拟模型算法都使得整体的攻击结果有所提高。从整体攻击性能上看, 当攻击由不同数据集训练得到的不同目标模型时, 方法 A 的攻击性能最差, 其平均攻击准确率仅能达到 57.3%, 平均攻击精确率仅能达到 60.4%。而方法 B 相较于基线方法, 在平均准确率上提升 5.7%, 在平均精确率上提升 3.2%。方法 C 的性能与方法 D 接近但在大部分模型上弱于方法 D, 仅在针对基于 Shop 数据集训练的 BigML 模型的攻击中, 方法 C 的攻击精确率好于方法 D。方法 C 的平均攻击准确率为 65.1%, 平均精确率为 67.3%。在此部分对比实验中, 方法 D 相较于其他方法具有最佳的攻击性能表现。与基线方法 A 相比, 方法 D 在平均准确率与精确率上分别提升 7.2%与 8.4%。在针对 Tweet-XGB 模型进行攻击时, 攻击方法 D 取得最佳性能, 准确率可达 74%, 精确率可达 86%; 同时, 与现有基线方法 A 相比, 在无需有关目标模型及其训练数据的任何先验知识的条件下, 可将攻击准确率提升 10.7%, 攻击精确率提升 11.2%。从上述实验结果可以看出, 本文所提成员推断攻击方法在无需目标模型及其训练数据的先验信息的条件下, 可准确推断目标模型的训练数据的隶属情况, 从而严重威胁模型的隐私安全。

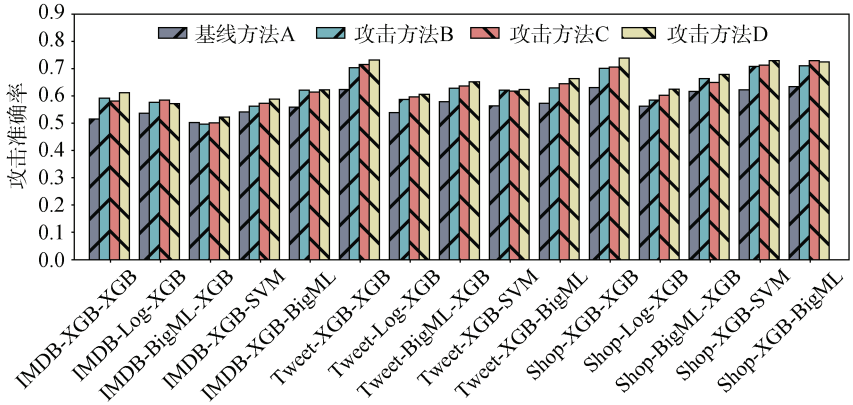

图 7 成员推断攻击准确率对比

Figure 7 Accuracy Comparison of Inference Attacks

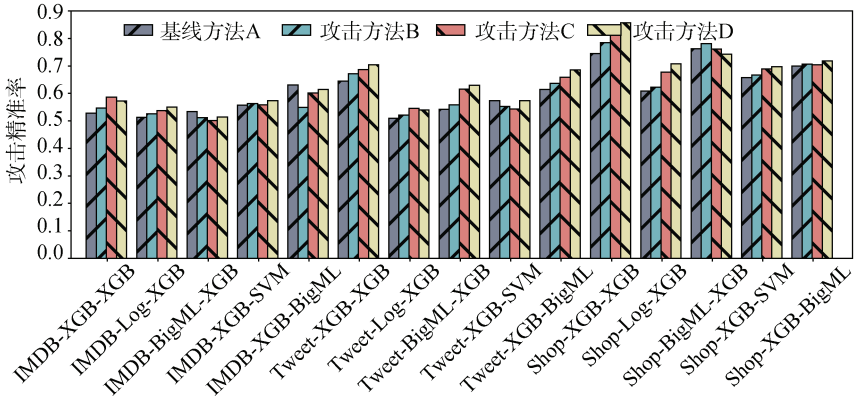

图 8　成员推断攻击精确率

Figure 8　Precision Comparison of Inference Attacks

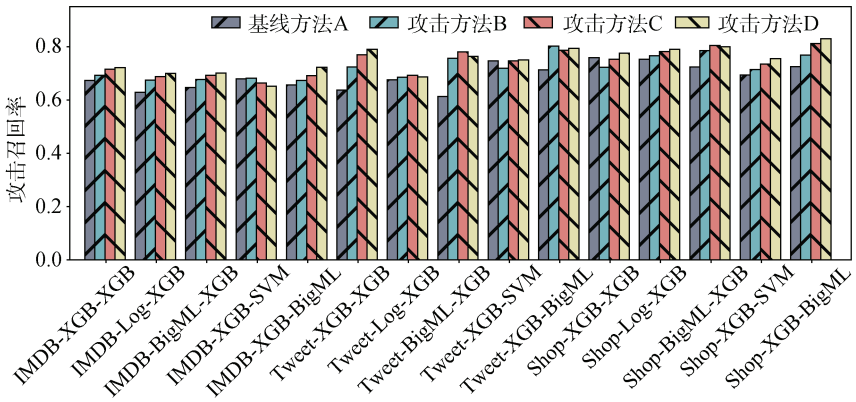

图 9　成员推断攻击召回率

Figure 9　Recall Comparison of Inference Attacks

b) 目标模型对攻击效果的影响

本部分评估成员推测攻击在不同机器学习目标模型上的表现。目标模型使用 XGBoost、Logistics 以及神经网络三种算法训练得到。如图 10 所示, 所有目标模型的攻击精度均高于 0.5, 当目标模型使用 XGB 时攻击效果最好。对于 IMDB 电影评论数据集来说, 准确率为 61.2%, 精确率为 57.2%, 召回率为 72.1%; 对于 Tweet 数据集来说, 准确率为 73.2%, 精确率为 70.4%, 召回率为 79%; 对于 Shop 购物数据集来说, 准确率为 73.9%, 精确率为 85.7%, 召回率为 77.5%。此外逻辑回归在 Tweet 和 Shop 数据集上的表现没有其他算法的表现好, 分析原因可能是该算法在多分类任务中表现较差。

c) 攻击模型对攻击效果的影响

本部分评估成员推测攻击在不同机器学习攻击模型上的表现。攻击模型使用 XGBoost、SVM 以及神经网络三种算法训练得到。如图 11 所示, 对于 IMDB 电影评论数据集来说, 当攻击模型为神经网络时, 攻击效果最好, 此时准确率为 62.2%, 精确率为 61.4%, 召回率为 72.1%; 对于 Tweet 数据集来说和 Shop 数据集来说, 当攻击模型为 XGBoost 时, 攻击效果最好, 此时准确率分别为 73.2%和 73.9%; 精确率分别为 70.4%和 85.7%。

d) 原始数据类别数量对攻击的影响

分析可知攻击模型是二分类模型, 特征维度等同于原始训练数据类别数量, 比如 Shop 数据集训练的攻击模型, 特征维度为 9。目标模型的分类越多, 攻击模型得到信息就越多, 因此效果也应该更好。通过比较 IMDB(2 分类), Tweet(6 分类), Shop(9 分类)三个数据集可以看出, 针对 Shop 数据集的攻击效果最好, 符合预期。

e) 目标模型和模拟模型的相似度对攻击效果的影响

部分主要评估目标模型和模拟模型的相似程度对攻击效果的影响, 从图 12 中可以看出, 目标模型和模拟模型的预测结果越相似, 整体的攻击效果就越好。随着相似度从 75%增加到 91%, Shop 数据集的准确率从 56.2%增加到 73.9%, 精确率从 60.8%增加到 85.7%, 召回率从 72.3%增加到 80.4%。

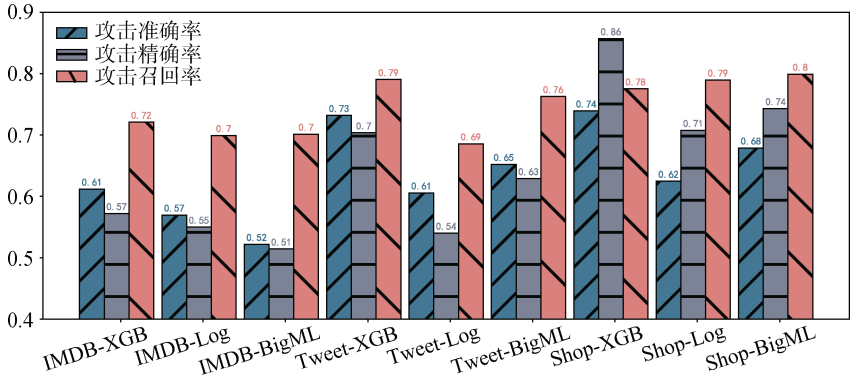

图 10　目标模型对攻击结果的影响

Figure 10　Performances Against Different Target Models

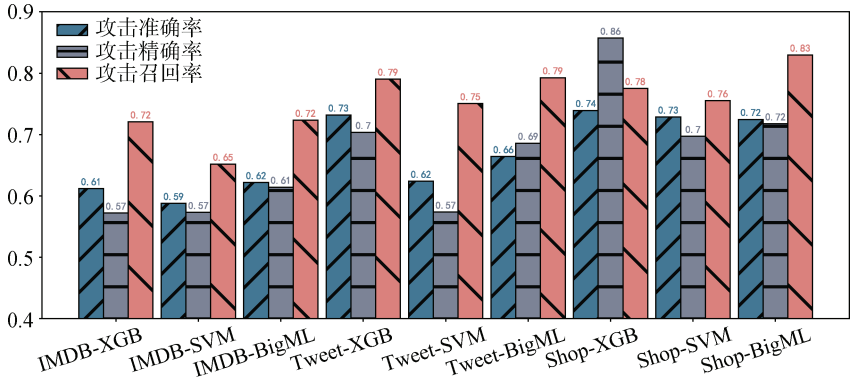

图 11　攻击模型对攻击结果的影响

Figure 11　Performances of Different Attack Models

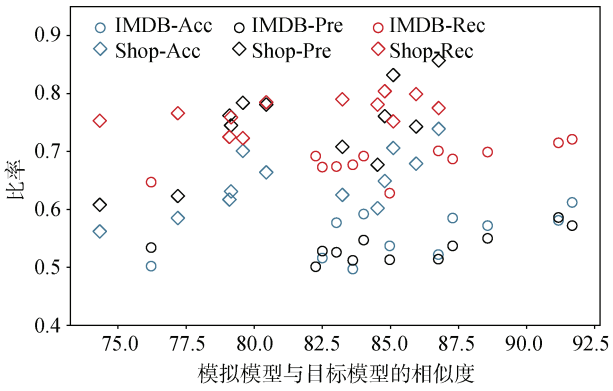

图 12　模拟模型性能对攻击结果的影响

Figure 12　The Relationship between the Mimic Model Similarity and the Attack Performance

5.4.2　评估模拟模型

模拟模型模仿的是目标模型的预测能力。模拟模型和目标模型越相似, 所训练得到的攻击模型就会越有效。所以模拟模型和目标模型的相似度是很重要的一个评估指标。本文在实验假设 A 的基础上对目标模型进行模拟, 主要验证数据合成算法的可靠性、以及模拟模型算法的有效性。

a) 模拟模型性能整体评估

图 13 表示模拟模型和目标模型两者预测概率的均方误差, 图 14 表示两者预测的相似度。基线方法 A 的平均 MSE 为 0.146, 平均相似度为 78.3%。对比实验 B 中使用合成数据生成算法算法合成数据, 平均 MSE 为 0.138, 平均相似度为 80.3%。从图 12、图 13 中可以直观地看出在大多数情况下对比实验 B 的结果要优于基线方法 A, 说明本文提出的数据合成算法可以增加样本空间的丰富性, 提高模拟模型性能。

对比实验 C 中, 平均 MSE 为 0.084, 平均相似度为 83.1%。和实验 A 进行对比, 可以看出使用模拟模型构建算法算法训练得到的模拟模型要比使用影子算法得到的模拟模型的均方误差更小, 相似度更高。

对比实验 D 中, 平均 MSE 为 0.061, 平均相似度为 84.1%。总体而言, 在不知道算法、模型参数和结构以及训练数据的情况下, 本文提出的合成数据算法是可靠的, 模拟模型算法是有效的, 由此我们可以训练出和目标模型预测能力更为接近的模拟模

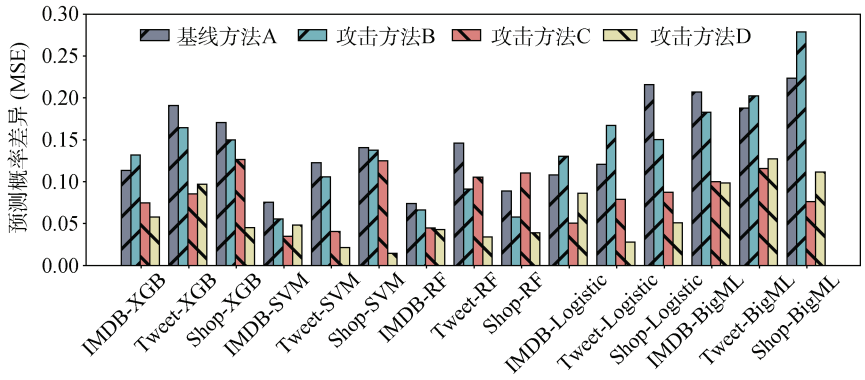

图 13　模拟模型性能评估(MSE)

Figure 13　The Prediction MSE of Mimic Models

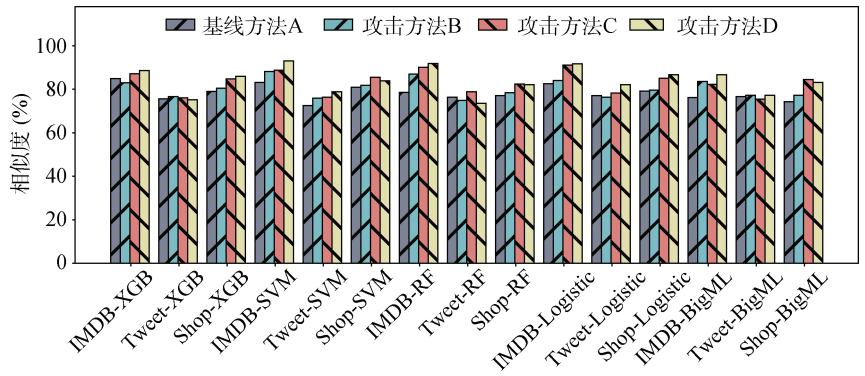

图 14　模拟模型性能评估(Similarity)

Figure 14　The Prediction Similarity of Mimic Models

型。随着模拟模型性能的提升，我们可以生成更加准确的攻击数据。

b) 合成数据数量对模拟模型性能的影响

本文利用合成数据查询目标模型，然后将获得的查询结果用于训练模拟模型，直观上合成数据集的大小将会影响模拟模型性能。我们对 XGBoost 和 RF 两种目标模型进行模拟，图 15 和图 16 显示了合成数据量和模拟模型性能之间的关系。

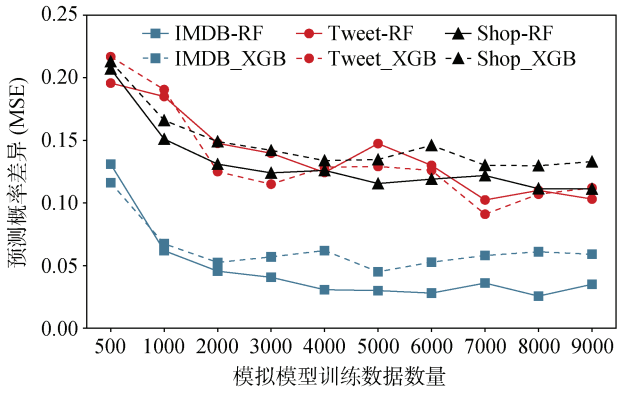

图 15　合成数据数量影响(MSE)

Figure 15　Impact of Synthetic Data Size (MSE)

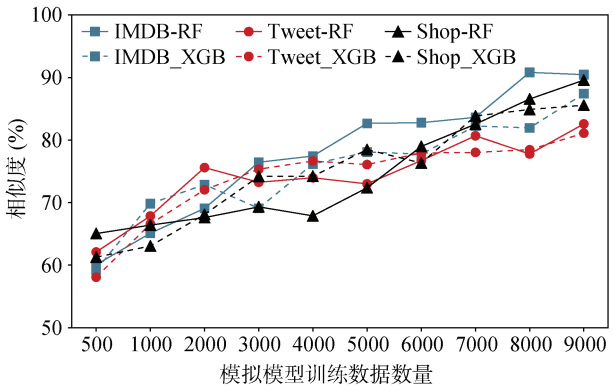

图 16　合成数据数量的影响(Similarity)

Figure 16　Impact of Synthetic Data Size (Similarity)

可以看出，随着数据数量的增加，模拟模型的预测行为和目标模型的预测行为越来越接近。以购物数据集为例，当目标模型为 RF 且模拟模型训练数据仅有 500 条时，均方误差为 0.207，相似率为 65.05%，随着数据增加，均方误差逐渐降低至 0.112，相似率增加至 89.58%。

c) 目标模型参数对模拟模型性能的影响

这部分评估不同结构的目标模型对模拟模型性

能的影响。主要关注 XGBoost 和 RF 两个目标模型及控制这两个模型的主要参数 Max-Depth 和 N-Estimator。Max-Depth 控制基学习器的最大树深度, 而 N-Estimator 控制基学习器的数量。通过调整这两个主要参数, 可以训练得到不同结构的目标模型。

图 17 与图 18 描述了目标模型的 Max-Depth 参数值对模拟模型性能的影响, 给定 N-Estimator 为 100, Max-Depth 为 3~9, 从实验结果可以看出: 对于 RF 模型来说, 当 Max-Depth 等于 6 时, MSE 为 0.032 且 Similarity 为 91.33%, 随着 Max-Depth 值的变化, MSE 和 Similarity 的波动幅度分别小于 0.03 和 8%。对于 XGB 模型来说, 当 Max-Depth 等于 6 时, MSE 为 0.043, 同时 Similarity 为 87.78%, 随着 Max-Depth 值的变化, MSE 和 Similarity 的波动幅度分别小于 0.03 和 5%。

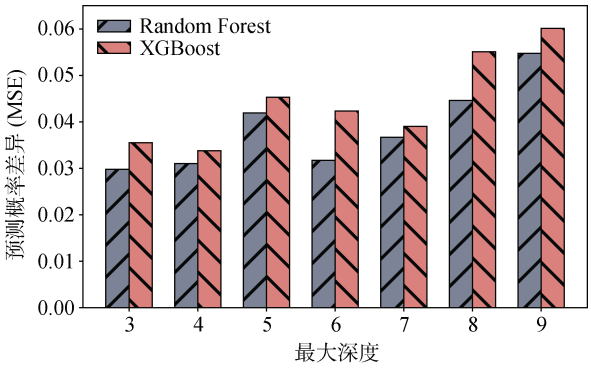

图 17 Max-Depth 参数值的影响(MSE)
Figure 17 Impact of Max-Depth (MSE)

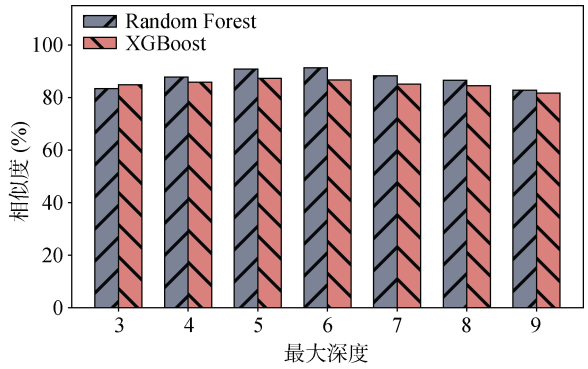

图 18 Max-Depth 参数值的影响(Similarity)
Figure 18 Impact of Max-Depth (Similarity)

图 19 与图 20 描述了目标模型的 N-Estimator 参数值对模拟模型性能的影响, 给定 Max-Depth 为 6, N-Estimator 分别为[10,40,70,100,140,170,200], 从实验结果可以看出: 对于 RF 模型来说, 当 N 等于 100 时, MSE 为 0.045 且 Similarity 为 92.07%, 随着 N 值的变化, MSE 和 Similarity 的波动幅度分别小于 0.04 和 15%。对于 XGB 模型来说, 当 N 等于 100 时, MSE 为 0.055, Similarity 为 87.33%, 随着 N 值的变化, MSE 和 Similarity 的波动幅度分别小于 0.03 和 14%。

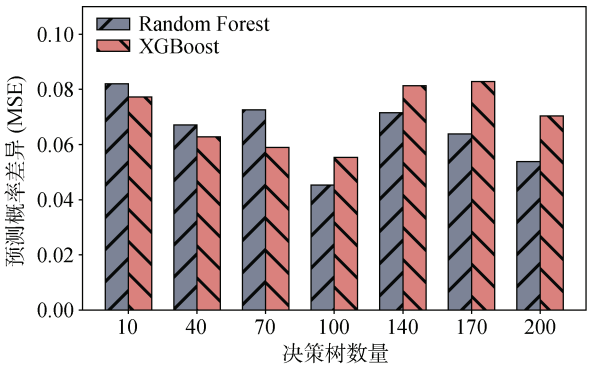

图 19 N-Estimator 参数值的影响(MSE)
Figure 19 Impact of N-Estimator (MSE)

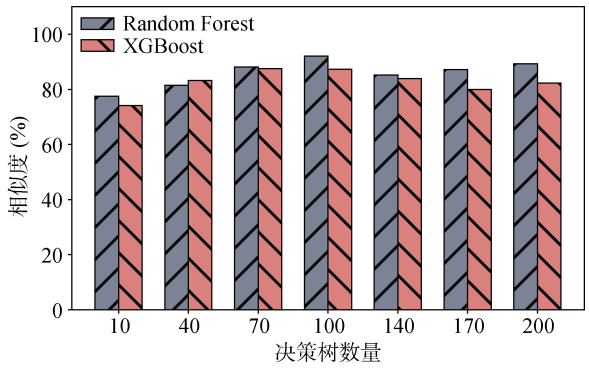

图 20 N-Estimator 参数值的影响(Similarity)
Figure 20 Impact of N-Estimator (Similarity)

d) 模拟模型参数对模拟模型性能的影响

在模拟目标模型的过程中, 我们构造了一个神经网络来模拟目标的功能。显然不同结构的神经网络模型具有不同的表示能力。因此在这一部分中, 我们评估模拟模型的规模对模拟模型性能的影响。

图 21 与图 22 显示了模拟模型的层数与性能之间的关系。容易得出, 更大规模的模拟模型具有更强

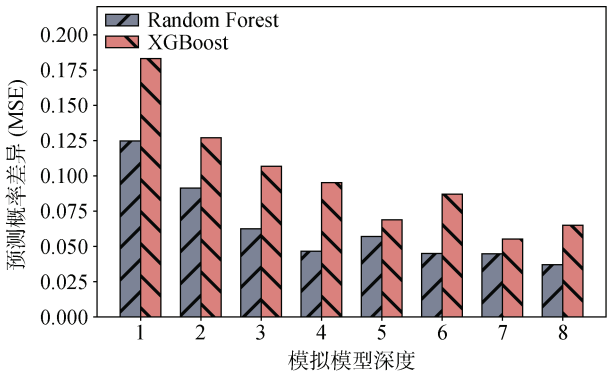

图 21 模拟模型规模对性能的影响(MSE)
Figure 21 Impact of Mimic Model Scale (MSE)

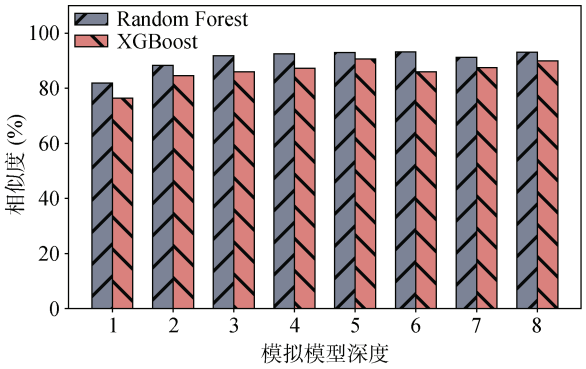

图 22　模拟模型规模对性能的影响(Similarity)
Figure 22　Impact of Mimic Model Scale (Similarity)

的模仿目标模型预测行为的能力。对于 RF 模型来说, 随着模拟模型层数的增加, MSE 从 0.125 降至 0.037, Similarity 从 81.85%增至 93.13%。对于 XGB 模型来说, 随着模拟模型层数的增加, MSE 从 0.183 降至 0.065, Similarity 从 76.38%增加到 89.9%。

6　总结与展望

本文针对黑盒机器学习模型的成员推断攻击进行了研究, 结果表明参与机器学习模型训练的用户即便在黑盒模型下依然会面临巨大的隐私泄露风险。相比于现有的成员推断攻击, 本文提出的攻击无需目标模型及其训练数据的先验知识, 在仅有目标模型黑盒预测接口的条件下, 可获得更加准确的攻击性能。本文使用了三种真实数据集, 并分别针对本地训练和 BigML 平台训练的模型进行了攻击, 实验结果证明基于本文提出的合成数据生成算法与模拟模型构建算法的成员推断攻击具有较强的攻击有效性。在未来的工作中, 我们将重点关注黑盒模型场景下对于成员推断攻击的防御以及机器学习模型安全风险的定量评估。

参考文献

[1] Toch E, Lerner B, Ben-Zion E, et al. Analyzing Large-Scale Human Mobility Data: A Survey of Machine Learning Methods and Applications[J]. *Knowledge and Information Systems*, 2019, 58(3): 501-523.

[2] Dacrema M F, Cremonesi P, Jannach D. Are we Really Making much Progress? a Worrying Analysis of Recent Neural Recommendation Approaches[C]. *The 13th ACM Conference on Recommender Systems*, 2019: 101-109.

[3] Nawaratne R, Alahakoon D, De Silva D, et al. Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(1): 393-402.

[4] Yao H X, Tang X F, Wei H, et al. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction[J]. *The AAAI Conference on Artificial Intelligence*, 2019, 33: 5668-5675.

[5] Zhao Z D, Chang X L, Wang Y X. A Survey of Privacy Preserving in Machine Learning[J]. *Journal of Cyber Security*, 2019, 4(5): 1-13. (赵镇东, 常晓林, 王逸翔. 机器学习中的隐私保护综述[J]. *信息安全学报*, 2019, 4(5): 1-13.)

[6] Al-Rubaie M, Chang J M. Privacy-Preserving Machine Learning: Threats and Solutions[J]. *IEEE Security & Privacy*, 2019, 17(2): 49-58.

[7] Yang Q, Liu Y, Chen T J, et al. Federated Machine Learning[J]. *ACM Transactions on Intelligent Systems and Technology*, 2019, 10(2): 1-19.

[8] 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.)

[9] Li J C, Li N H, Ribeiro B. Membership Inference Attacks and Defenses in Supervised Learning via Generalization Gap[EB/OL]. 2020

[10] 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]. *2019 IEEE Symposium on Security and Privacy*, 2019: 739-753.

[11] 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.

[12] Wang C, Liu G Y, Huang H J, et al. MIASec: Enabling Data Indistinguishability Against Membership Inference Attacks in MLaaS[J]. *IEEE Transactions on Sustainable Computing*, 2020, 5(3): 365-376.

[13] Salem A, Zhang Y, Humbert M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[C]. *2019 Network and Distributed System Security Symposium*, 2019: 1-15.

[14] 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.

[15] D.P. Kingma, M. Welling. Auto-encoding variational bayes[EB/OL]. *CoRR, arXiv*:1312.6114, 2013.

[16] M. Arjovsky, S. Chintala, L. Bottou. Wasserstein generative adversarial networks[C]. *International Conference on Machine Learning*, 2017: 214-223.

[17] Wang R, Li Y F, Wang X F, et al. Learning your Identity and Dis-

ease from Research Papers: Information Leaks in Genome Wide Association Study[C]. *The 16th ACM conference on Computer and communications security - CCS '09*, 2009: 534-544.

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

[19] Pyrgelis A, Troncoso C, De Cristofaro E. Knock Knock, Who's There? Membership Inference on Aggregate Location Data[C]. *2018 Network and Distributed System Security Symposium*, 2018: 1-15.

[20] F.L. Xu, Z. Tu, Y. Li, et al.. Trajectory recovery from ash: User privacy is not preserved in aggrgated mobility data[C]. *International Conference on World Wide Web*, 2017: 1241-1250.

[21] 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.

[22] S. Truex, L. Liu, M. E. Gursoy, et al. Towards demystifying membership inference attacks[EB/OL]. *CoRR, arXiv*: 1807.09173, 2018.

[23] 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.

[24] 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.

[25] Ribeiro M, Grolinger K, Capretz M A M. MLaaS: Machine Learning as a Service[C]. *2015 IEEE 14th International Conference on Machine Learning and Applications*, 2015: 896-902.

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

[27] C. Doersch. Tutorial on variational autoencoders[EB/OL]. *CoRR, arXiv*:1606.05908, 2016.

[28] Yu D, Yao K S, Su H, et al. KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition[C]. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: 7893-7897.

[29] I. Goodfellow, J. Pouget-Abadie, and M. Mirza. Generative adversarial nets[C]. *Conference on Neural Information Processing Systems*, 2014: 2672-2680.

[30] M. Arjovsky, L. Bottou. Towards principled methods for training generative adversarial networks[EB/OL]. *CoRR, arXiv*:1701.04862, 2017.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, et al. Improved training of Wasserstein GANs[C]. *Conference on Neural Information Processing Systems*, 2017: 5767-5777.

[32] Chen T Q, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016: 785-794.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python[J]. *The Journal of Machine Learning Research*, 2011(12): 2825-2830.

[34] A. Paszke, S. Gross, F. Massa, et al. Pytorch: An imperative style, high-performance deep learning library[C]. *Advances in Neural Information Processing Systems*, 2019: 8026-8037.

刘高扬 于 2016 年在华中科技大学电子信息与通信学院获得工学学士学位。目前在华中科技大学电子信息与通信学院攻读博士学位。研究领域机器学习安全。研究兴趣包括: 对抗学习、异常检测、联邦学习。Email: liugaoyang@hust.edu.cn

李雨桐 于 2018 年在华中科技大学电子信息与通信学院获得工学学士学位。目前在华中科技大学电子信息与通信学院攻读硕士学位。研究领域为数据安全。研究兴趣包括: 对抗机器学习、数据隐私保护等。Email: ytli@hust.edu.cn

万博睿 于 2018 年毕业于江西师范大学附属中学。目前在华中科技大学电子信息与通信学院电信卓越计划实验班攻读学士学位。研究领域为人工智能安全。研究兴趣包括: 对抗机器学习、数据隐私保护等。Email: raywan@hust.edu.cn

王琛 于 2013 年在武汉大学自动化系获得博士学位。现任华中科技大学电子信息与通信学院副研究员、博士生导师、ACM/IEEE/CCF 高级会员。研究领域为物联网数据安全与隐私保护。研究兴趣包括: 时空数据挖掘、社交计算、边缘智能、对抗机器学习等。Email: chenwang@hust.edu.cn

彭凯 于 2006 年在华中科技大学通信工程专业获得博士学位。华中科技大学电子信息与通信学院教授、博士生导师。研究领域为网络安全。研究兴趣包括: 移动边缘计算、大数据处理、物联网等。Email: pkhust@hust.edu.cn