基于代理训练集的属性推理攻击防御方法

董 恺 蒋驰昊 李 想 凌 振 杨 明
(东南大学计算机科学与工程学院 南京 211189)

摘 要 本文首次提出针对属性推理攻击的有效防御方法.属性推理攻击可以揭示出用于训练公开模型的原始私有数据集中的隐私属性信息.现有研究已经针对不同的机器学习算法提出了多种属性推理攻击.这些攻击很难防御,一方面原因是训练有素的模型总是会记住训练数据集中的显性和隐性全局属性,另一方面原因在于模型提供者无法事先知道哪些属性将受到攻击从而难以有针对性地进行防御.为了解决这个问题,本文提出了一种通用的隐私保护模型训练方法,名为PPMT(Privacy Preserving Model Training).它以迭代的方式工作.在每次迭代中,PPMT构建一个代理数据集,并在该数据集而不是私有数据集上训练模型.虽然每次迭代会同时导致隐私性的提升和功能性的降低,但隐私性的提升呈快速指数级,而功能性的降低则是缓慢线性的.经过多次迭代,PPMT在模型功能性的约束下最大化全局属性的隐私性,并生成最终的模型.本文选择了两种代表性的机器学习算法和三个典型的数据集来进行实验评估PPMT所训练出模型的功能性、隐私性和鲁棒性.结果显示,使用PPMT训练出的模型,在全局属性上会以不同速度朝不同方向改变,在功能性上的平均损失为1.28%,在超参数α保密的情况下被可能攻击倒推的成功率仅有22%~33%.这说明,PPMT不仅能保护私有数据集的全局属性隐私性,而且能保证模型有足够的功能性,以及面对可能攻击的鲁棒性.

关键词 人工智能安全;属性推理攻击;全局属性隐私;隐私增强;代理数据集
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.00907

Defending against Property Inference Attacks Based on Agent Training Datasets

DONG Kai JIANG Chi-Hao LI Xiang LING Zhen YANG Ming
(*School of Computer Science and Engineering, Southeast University, Nanjing* 211189)

Abstract We are the first to propose an effective defense against property inference attacks. A property inference attack reveals properties of the private training dataset from public classifiers trained on this dataset. Existing researches have proposed various property inference attacks for different machine learning algorithms. These attacks are difficult to defend against, since a well-trained model always remembers all the explicit and implicit global properties of the training dataset, and the model provider cannot know what properties will be attacked in advance. To address this problem, this paper proposes a generic privacy preserving model training method, named PPMT, which works in an iterative fashion. In each iteration, PPMT constructs a substitution dataset and trains a model on this dataset instead of the private one. Although each iteration leads to privacy increasing and utility decreasing, the privacy exhibits a fast and exponential increase, while the utility exhibits a slow and linear decrease. After several iterations, PPMT

收稿日期:2023-06-29;在线发布日期:2024-01-03. 本课题得到国家重点研发项目(No. 2023YFC3605804)、国家自然科学基金(No. 62072098,62072103,62232004)、江苏省重点研发项目(No. BE2022065-5,BE2022680)资助.**董 恺**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为隐私增强技术、人工智能安全、物联网和系统安全. E-mail: dk@seu.edn.cn. **蒋驰昊**,硕士研究生,主要研究方向为人工智能安全. **李 想**,硕士研究生,主要研究方向为人工智能安全、可解释机器学习. **凌 振**,博士,教授,中国计算机学会(CCF)会员,主要研究方向为网络安全和隐私保护、人工智能安全、物联网、移动系统安全、可信计算. **杨 明**,博士,教授,中国计算机学会(CCF)会员,主要研究方向为网络安全和隐私保护、人工智能安全.

generates the final model which maximizes privacy of global properties under the constraint of model utility. This paper considers two representative machine learning algorithms and three typical datasets, and conducts experiments to evaluate the utility, privacy and robustness performance achieved by models trained by PPMT. The results show that the models trained with PPMT change at different speeds in different directions in terms of global properties, with an average loss of 1.28% in terms of utility, and the success rate of inverse by a possible attack with hyper-parameter α secrecy is only 22% to 33%. Thissuggests that PPMT not only preserves privacy of the private dataset but also ensures adequate model utility and even robust to possible attack.

Keywords artificial intelligence security; property inference attack; global property privacy; privacy-enhancing techniques; agent datasets

1 引 言

近年来,研究者针对机器学习已提出多种攻击,大致可分为四类:(1)后门攻击[1-6],向模型中注入隐蔽的后门;(2)对抗攻击[7-12],通过生成对抗样本使模型预测错误;(3)模型窃取攻击[13-16],窃取模型的机密信息;(4)属性推理攻击[17-21]、成员推理攻击[22-25]和模型反演攻击[26-27],窃取训练数据集的机密信息.针对前述大多数攻击,研究人员已经提出了相应的防御措施.然而针对属性推理攻击的防御措施尚未提出.

属性推理攻击通过影子训练技术进行.具体来说,首先在影子训练集上训练一批影子模型,其中一半有特定属性,另一半则没有.然后,攻击者将每个影子模型中提取到的特征作为样本,将模型相应的数据集是否有特定属性作为标签,形成一个元训练数据集.接着,攻击者在元训练数据集上训练出一个二分类的元分类器.最后,使用元分类器攻击目标模型.

训练数据集的全局属性泄露带来了严重的隐私问题.全局属性是指训练数据集中所有用户在某些特征上的分布.属性推理攻击所揭示的属性可以是任何特征,这意味着不仅训练数据集中的任何显性特征会被暴露,而且数据集之外的本身符合的隐藏属性的隐性特征也可能被暴露.当这些用户来自同一个用户群(例如,属于同一个组织,居住在同一个地区,或使用同一个服务等),全局属性的泄漏甚至会引起单个用户的隐私问题.例如,如果训练数据集中高比例用户具备某些隐私特征(例如,"医疗史"上的"心脏病"),那么攻击者可以猜测用户大概率拥有该特征.

在私有数据集上训练一个能够抵御属性推理攻击的模型是非常困难的.因为训练有素的模型(不是欠拟合的)总是不可避免地以不同程度过拟合.此外,训练有素的模型总是倾向于描述训练数据集中的大部分而不是少部分数据.上述两点因素会使目标模型的参数产生隐藏的模式,从而可以被属性推理攻击所运用,揭示出训练数据集中显式特征或隐式特征的全局属性.

本文提出了一种保护隐私的模型训练方法,称为 PPMT(Privacy Preserving Model Training),作为对属性推理攻击的防御.其基本思想是构建一个代理数据集替代私有数据集,之后就可以在代理数据集上进行训练而非私有数据集.这个数据集应当可以保证模型拥有足够的功能性,并确保私有数据集的全局属性不被泄露.

然而,构建出这样的一个代理数据集并非易事.在这个数据集上训练出的模型应该满足两个条件.一方面,它必须达到足够的功能性(通常表现为准确率),与在私有数据集上训练的模型达到的功能性相媲美.另一方面,它不应该暴露私有数据集中的任何全局属性.一个天然的想法是向私有数据集中添加虚拟数据或者噪声,然而这个想法在推理攻击场景下是无效的.因为随着虚拟数据或者噪声的增加,训练出的模型其功能性会迅速下降.仅仅添加一个小比例的虚拟数据或者噪声(例如10%),训练出的模型其准确率就会下降很大比例(例如从 95%降到 85%).而同时,属性推理攻击的结果仅会稍有变化.

为了应对上述挑战,PPMT 使用迭代的方法来构建代理数据集并训练模型.PPMT 首先在私有数据集上训练一个模型,本文称之为旗帜模型.在每一轮迭代中,PPMT 从原始数据集的样本域(而不仅仅是原始数据集)中随机选择样本,使用旗帜模型来标记样本以构建代理数据集.接着在该代理数据集

上训练一个新的模型以代替旗帜模型.虽然在这个过程中,功能性随着隐私性的提高而降低.但是前者表现为线性缓慢降低,而后者为指数级快速增加.经过几次迭代之后,训练出的模型可以保留隐私性,同时保证足够的功能性.

本文在三个数据集上进行了实验,以评估通过PPMT方法所训练模型的隐私性和功能性,并验证了PPMT在可能攻击下的鲁棒性.实验结果表明,随着迭代轮数的增加,训练出的模型可以在保证有足够功能性的同时得到隐私保护(即在属性推理攻击下得到的属性值会发生很大变化).即使遵循开放设计原则,假设攻击者知道PPMT是否被使用并且能够获取PPMT实现的所有细节,PPMT仍然可以保障原始数据集的全局属性隐私性.

本文的主要贡献如下:

(1)本文提出了第一个针对属性推理攻击的有效防御PPMT.PPMT不是通过向训练数据集添加噪声进行隐私性和功能性的简单权衡,而是使训练出的模型忘记私有数据集的所有全局属性.

(2)本文提出了一种迭代模型训练方法,以确保当新训练的模型忘记全局属性时,模型潜在的分类准则能够很好地保留下来.

(3)本文通过实验验证了PPMT的有效性及其所训练模型的隐私性和功能性.

(4)本文根据开放设计原则,通过考虑可能的攻击来评估PPMT的鲁棒性.结果表明,在模型训练者所使用的超参保密的情况下,PPMT是安全和鲁棒的.

2　研究背景

现有的属性推理攻击大多聚焦于机器学习中的有监督分类模型,因此本文也针对该场景研究防御方法.本节简要介绍机器学习中的监督分类和属性推理攻击的概念,并解释为什么一个训练有素的分类器总是容易受到这种攻击.

2.1　监督分类

在监督学习中,分类器的训练方法如下.设X是一组数据样本,Y是一组标签或类别.给定一组数据(也称为训练数据集),每一项都是一个样本—标签对,表示为(x,y),其中$x\in X$,$y\in Y$,y是x的真实标签.给定一个分类算法M,本文用Θ来表示M中所有参数的域.设f_θ为一个由参数$\theta\in\Theta$指定的分类器.而一个最佳分类器f^*可以通过在训练数据集上进行以下计算得到:

$$\theta^*=\underset{\theta\in\Theta}{\operatorname{argmin}}\sum(L(f_\theta(x),y))+\lambda\|\omega\|_2 \quad (1)$$

其中θ^*是用于指定f^*的参数集,$L(\cdot,\cdot)$是衡量预测结果与真实标签之间距离的损失函数,λ是正则化参数,ω是分类器内神经元的权重.一般来说,分类器需要处理未曾见过的数据样本.具体的,给定一个数据样本$x'\notin X$,通过计算$f^*(x')$得到每个标签预测值,最后将具有最大预测值的标签视作分类结果.

根据是否使用神经网络,监督分类算法可以分为两类.对于基于深度学习的相对复杂分类算法,本文选择全连接神经网络(Fully Connected Neural Network, FCNN)算法作为其代表.对于基于简单机器学习的相对简单分类算法,本文选择支持向量机(Support Vector Machine, SVM)算法作为其代表.以下简要描述这两种算法:

FCNN,也被称为多层感知器,在过去的十年中得到了普及.一个FCNN分类器f由三种类型的层组成:输入层、输出层和多个隐藏层.设x为f的输入,输入层输出x,输出层会输出一个k维向量$f(x)$.位于输入层和输出层之间的隐藏层按照以下方式从x计算出$f(x)$:

$$f(x)=f_{|f|}(f_{|f|-1}(\cdots f_2(f_1(x)))) \quad (2)$$

其中,$|f|$表示层数(不包括输入层),$f_i(1\leqslant i\leqslant|f|)$表示对第$i-1$层(输入层为第0层)的输出进行的第$i$个变换.每个变换$f_i$由多个称为神经元的计算单元组成.每个神经元使用其权重和偏置,对前一层的输出进行加权线性求和.由于FCNN一般用于拟合非线性函数,因此求和结果还会通过非线性激活函数进行转换.FCNN分类器是通过求解方程(1)中定义的优化问题来训练的,其中的参数包括所有神经元的权重和偏置.最优参数可以通过使用梯度下降法找到,例如随机梯度下降(SGD)[28]和Adam(Adaptive Moment Estimation,Adam)[29].

SVM已被广泛应用于分类任务中.分离超平面的概念[30]可以解释SVM的工作原理.分离超平面是样本空间中的一个超平面,用于将不同类别分开,因此它作为决策边界来对不同样本进行分类.对于给定的训练数据集,存在无穷多个分离超平面.SVM分类器经过训练得到最优超平面,最大化不同类别样本到最优超平面的最小距离.具有最小距离的样本对约束了分类器,被称为支持向量.预测时,SVM分类器将数据样本映射到样本空间,并使用最优分离超平面得到预测类别.

2.2 属性推理攻击

属性推理攻击是一种获取模型提供者不愿公开、但在训练阶段被窃取模型记忆的私有数据集全局属性(比如男女比例)的方法.这些全局属性在分类过程中未被明确运用,但最终对模型参数产生隐性影响.属性推理攻击给社会信息安全带来了极大的威胁.假设有一个基于海量用户数据训练完好的个性化推荐模型(例如短视频推荐模型)被泄露,攻击者将可以通过属性推理攻击得到参与训练用户的全局属性(例如政治取向、收入分布等).

通常来说,属性推理攻击是一种白盒攻击,即攻击者对模型具有完全的知识和访问权限,包括模型结构、输入、输出和内部参数.攻击者拿到一个目标模型后,对训练模型所使用的私有数据集的某个全局属性感兴趣.就可以通过训练一个分类器来识别目标模型是否拥有这个属性,从而进行属性推理攻击.攻击的整个工作流程如图 1 所示.

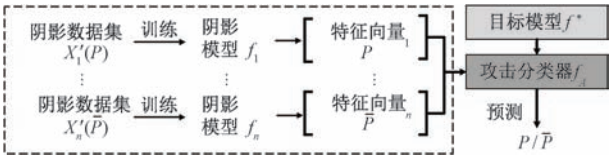

图 1 属性推理攻击的工作流程

具体来说,设 f^* 表示从目标数据集 X 训练得到的目标模型,P 表示攻击者感兴趣的属性.攻击者首先收集一个数据集 X',该数据集由 f^* 的一组合理输入组成.需要注意的是攻击者收集的数据集 X' 的分布可能与 X 的分布非常不同.攻击者接着生成 n 个所谓的阴影数据集,表示为 $\{X'_i \subseteq X' \mid 1 \leqslant i \leqslant n\}$,其中一半符合属性 P,另一半符合 $\bar{P}$.对于每一个阴影数据集,攻击者训练一个阴影模型,表示为 f_i,$1 \leqslant i \leqslant n$,并根据相应的阴影数据集将其标记为 P 或 $\bar{P}$.例如,攻击者想要知道在 X 中性别比例是否为 1∶1.那么对于每个阴影模型 f_i,如果其训练数据集 X'_i 中的性别比例为 1,f_i 就被标记为 P;否则,f_i 被标记为 $\bar{P}$.之后,攻击者在由阴影模型及其标签所组成的数据集上完成一个二分类任务(P 或 $\bar{P}$),得到分类器 $f_{\mathcal{A}}$.最后,将 f^* 作为 $f_{\mathcal{A}}$ 的输入,得到的输出就是对全局属性的预测值(例如,X 中的性别比例是否为 1).攻击者可以利用目标模型的独特特征,例如模型架构、输入、输出或内部参数中的任何部分(或全部),来训练二元分类器 $f_{\mathcal{A}}$.

研究者已经验证了训练有素的模型容易受到属性推理攻击的影响并提出了多种攻击方法.2015 年,Ateniese 等人[17]第一次提出了属性推理攻击.该攻击针对的是 SVM 和隐马尔可夫模型(HMM),在更加复杂的神经网络上效果较差.为了解决这个问题,2018 年,Ganju 等人[18]针对 FCNN 的置换不变性提出了两种改进攻击:基于排序的方法和基于集合的方法.2019 年,Melis 等人[19]针对联邦学习[31-32]提出了一种被动属性推理攻击方法以及一种主动的变种攻击方法.截至目前,学术界还没有提出一种针对属性推理攻击的有效防御方法.在文献[18]中,Ganju 等人提出了三种可能的防御方法,但是当攻击者了解这些方法后就会失效.

3 问题描述与分析

本文对以下问题感兴趣:是否可以在私有数据集上训练一个实用的模型,同时确保任何全局属性不被这个模型记住?

在本节中,提出了一个基本思路来解决这个问题.如果模型是在一个与私有数据集有很大差距的代理数据集上训练的,那么对手就只能获得代理数据集的全局属性.实现该思路的主要挑战涉及到如何构建这个代理数据集,以同时确保模型的功能性与原数据集的全局属性隐私性.本文对构建数据集的可能方法进行了详细的分析.

3.1 问题定义

假设一个模型提供者可以访问一个私有数据集 S.他在 S 上训练了一个模型 f^*,并愿意公开这个模型.用户对该模型有充分的了解,并可进行任意的访问,但不能获取私有数据集 S.这种情况下任何用户都可能是攻击者.假设攻击者对某个全局属性 P 很感兴趣,其中 P 是模型提供者所不愿分享甚至不为模型提供者所知的.攻击者通过执行包括属性推理攻击在内的各种攻击来确认 S 是否遵循 P.在这个场景下,我们希望模型提供者能达到下面两个要求:(1)发布的模型 f^* 拥有足够的功能性(即准确率足够高),(2)模型不会泄露私有数据集的任何全局属性.

学术界普遍认为,同时实现这两个目标非常困难.其根本原因在于,经过良好训练的模型(未欠拟合)总是不可避免地出现过拟合现象.而通过识别模型过拟合引起的特征,可以进行各种被动攻击,包括属性推理攻击、成员推理攻击和模型反演攻击.尽管通过修改损失函数(例如,使用方程(1)中的 $L2$ 范

数)可以控制模型的复杂度,但过拟合的问题并没有被完全解决.

此外,训练有素的模型总是倾向于描述训练数据集中的大部分数据.具体而言,私有数据集 S 中的属性 P 会将 S 划分成多个分区,而经过充分训练的模型 f^* 总是倾向于对具有最多样本数的分区做出最佳预测.这是求解方程(1)中定义的优化问题所带来的不可避免现象.对于模型而言,如果能正确预测来自两个不同分区样本的概率相同,那么具有更多样本的分区会对损失函数 $L(\cdot,\cdot)$ 的优化做出更多贡献.为了减小后续拟合过程中的分类误差,模型必然会偏向于找到大分区中样本标签分布的规则,因为这种行为在减小损失方面是最有利的.这就导致该分区内的样本具有更高的准确率.上述的模型倾向性可以被属性推理攻击中的二元分类器 $f_{\mathcal{A}}$ 所利用.

3.2　基本思路及其挑战

为了解决上述问题,本文的基本思路是构建一个数据集作为原始私有数据集的替代,之后让公开的模型在构建的数据集上进行训练.训练得到的模型应该满足两个条件:一方面,它必须达到足够的功能性,与在私有数据集上训练的模型所达到的准确率相媲美;另一方面,它必须有足够的隐私性,不暴露私有数据集的任何全局属性.

在存在公共数据集的情况下,很容易同时满足这两个约束条件.因为如果公共样本足够多,可以直接在公共数据集上训练模型.在这种情况下,不需要保护任何隐私.公共数据集中的样本数量如果不足,则可以使用私有数据集,通过混合公共数据集和私有数据集来满足隐私约束,以混淆全局属性.

本文重点关注没有公共数据集的情况.此时构建一个代理数据集以同时满足功能性约束和隐私性约束并非易事.人们可能会想到向私有数据集中添加噪声,以在隐私性和功能性之间达到权衡.例如翻转私有数据集中一些训练样本的标签.然而,已有研究验证了该方法不能抵御属性推理攻击.而且,随着插入训练数据集的噪声数量增加,训练出的模型准确率会迅速下降.为了满足功能性的约束,构建的数据集中的样本—标签对必须是合理的,比如可以来自现实世界或根据私有数据集生成.但在这种设置下,隐私性约束又很难满足.在下文中,我们列出了几种构建合理数据集的方法,并阐述了它们不能满足隐私性约束的原因.

方法 1:对私有数据集进行随机抽样.可以通过对私有数据集进行随机抽样来构建一个替代数据集.然而,从概率来说构建的数据集和私有数据集的全局属性总是相同的.因此,这种方法没有任何意义.

方法 2:选择数据以保护某些全局属性.可以通过从私有数据集中人为选择数据来构建数据集,以确保某些全局属性得到保护.然而,这种方法在普遍情况下无法使用,因为模型提供者事先不知道攻击者感兴趣的属性是什么.此外,属性推理攻击可以用于揭示训练数据集之外的一些额外信息.一个例子是攻击 MNIST 手写数字识别模型来得到训练数据集中是否存在噪声[18].由于以上原因,该方法对于保护原始数据集的全局属性也没有贡献.

方法 3:收集新数据.通过从现实世界中收集新的数据,可以构建一个全新的数据集.模型可以在新收集的数据集上训练,也可以在由新数据和私有数据集组成的混合集上训练.无论使用哪种方法,训练数据集的全局属性都会被新样本所改变.同时,训练后的模型的功能性也是足够的.然而,此时需要保护的属性转变为了这个新的私有训练数据集的全局属性.除非如上文中所讨论的那样,新数据来自公共数据集.

方法 4:引入反向数据集.在私有数据集的反向形式上训练模型不能保护全局属性.这是因为,反向数据集中的样本只改变了输入向量的方向,但没有改变样本空间中的基本分类准则.因此在反向数据集上训练的模型仍然有一些隐藏模式.这些隐藏模式是以反向形式学习的,与在私有数据集上训练的模型中的模式相同.

方法 5:使用深度学习技术生成虚拟数据.可以使用一些深度学习技术生成虚拟数据集.然而,通过引入深度学习模型生成的虚拟数据无法抵御属性推理攻击.原因在于,深度学习模型的目标是生成尽可能与真实样本的统计分布相似的数据,而这意味着虚拟数据的属性与私有数据集的属性相似.

4　总体策略

本文介绍了一种保护隐私的模型训练方法,称为 PPMT.该方法构建了用于保护原始数据集全局属性隐私的训练数据集,称为代理数据集.代理数据集由在原始数据集样本域中随机采样的噪声数据和合理的标签组成.通过 PPMT,模型提供者可以训练一个基于代理数据集而不是原始数据集的模型.

本文将这个模型称为代理模型.任何对代理模型发生属性推理攻击的攻击者,只能获得代理数据集的全局属性信息.

在前一节中,已经阐述了一些方法不能满足隐私约束的原因.PPMT 与这些方法之间的主要区别在于,PPMT 以迭代方式工作,从而在隐私性和功能性之间取得更好的平衡.在每次迭代中,PPMT 在一个新构建的数据集上训练一个新的模型.而前一轮迭代中由 PPMT 训练的模型被用来确保新训练的模型能够达到充足的准确率.同时,新构建的数据集中的样本是从样本域中采样得到的结果,而不是来自原始数据集,因此新模型拥有更高的隐私性.虽然在每次迭代中,PPMT 以模型部分功能性的牺牲为代价换取了隐私性,但 PPMT 保证了模型的隐私性提升速度远远高于功能性下降的速度.

图 2 展示了 PPMT 的完整工作流程,由以下五个步骤组成.

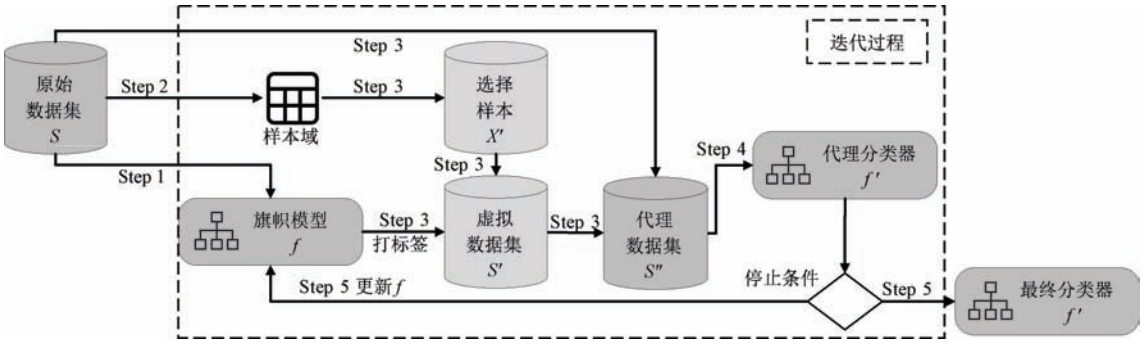

图 2 PPMT 的工作流程

第 1 步:训练旗帜模型.在这一步中,模型提供者在原始训练数据集上训练一个模型.我们将这个原始模型称为旗帜模型.模型提供者选择分类算法后,对原始数据集进行数据预处理操作,如标准化和归一化,并训练旗帜模型.与后续训练的代理模型相比,旗帜模型分类输入样本的能力更强,学习到的原始数据集中的特征更多.模型提供者需要确保旗帜模型达到最佳分类准确率,因为该准确率是后续代理模型所能达到的准确率上限.训练旗帜模型的目的有两个.首先,它指导后续代理模型的训练,因为代理模型的训练过程就是在尽可能模仿旗帜模型的行为,只不过其在代理数据集而不是原始数据集上进行训练.其次,它被视为基准模型,用于评估后续代理模型的隐私性和功能性.

第 2 步:确定样本域.在构建代理数据集以训练代理模型之前,需要确定样本域,以确保构建的数据是“合理的”.原始数据集可以根据样本的结构被分为两类:结构化数据集和非结构化数据集.结构化数据集由高度组织和格式化的数据组成,可以表示为二维表格.非结构化数据集则包括其他类型的数据,如图像和文本数据.

对于模型提供者来说,可以根据原始数据集的结构化或非结构化特性,使用以下两种操作之一来获取样本域.对于结构化数据集来说,样本域是所有可能值的集合.对于非结构化数据集来说,情况更为复杂.例如图像数据集的样本域是像素点的所有可能值的集合.文本数据集的样本域是词库中所有词素组合的集合.在一般情况下,非结构化数据集的样本域比结构化数据集的更大.因此非结构化数据集上 PPMT 的使用效果可能不如在结构化数据集上的有效.

第 3 步:构建代理数据集.这是 PPMT 中最重要的步骤.在这一步中,会构建一个由噪声样本组成的非真实数据集.我们将这个数据集称为虚拟数据集,它在经过处理后将参与代理模型训练.在构建虚拟数据集时需要解决一个基本冲突.一方面,虚拟数据集必须与原始数据集足够相似,以确保在该数据集上训练的代理模型有足够的准确率.但在另一方面,它又必须与原始数据集足够不同,以使其自身的全局属性与原始数据集不同.

PPMT 构建虚拟数据集的方法如下.首先,在样本域中根据均匀概率分布随机选择一定数量的样本.由于这些样本通常在真实世界中不存在,因此被称为噪声样本.然后,将这些噪声样本输入到第一步中训练的旗帜模型,以获得预测的标签.最后,将每个样本与其预测的标签结合起来,构成虚拟数据集.在构建过程中,原始数据集的任何统计信息(即全局属性)都被破坏.同时,预测的标签是由旗帜模型计算得出的,因此旗帜模型中分类的基本准则得以保留,并且一直可以被后续的代理模型所学习.

如果原始数据集的隐私被完全保护,那么对代理模型进行属性推理攻击时,任意有关比例的全局属性预测结果会是 1∶1.这是因为虚拟数据集是按照均匀概率分布进行采样的.然而,实际实验中得到的比例始终是介于 1∶1 和原始数据集中比例之间的值,这意味着隐私尚未得到完美保护.这个现象的原因是旗帜模型记住的属性仍然对代理模型产生潜在影响.因此,想要保留原始模型的分类潜在准则就势必会牺牲部分隐私性保护.

在 PPMT 中,代理模型是在代理数据集上进行训练的,该数据集由虚拟数据集和原始数据集混合组成.这种设置实际上是功能性和隐私性之间的权衡.虚拟数据集中的样本确保了模型的隐私性增加,同时原始数据集中的样本确保了代理模型有足够的准确率.代理模型从旗帜模型中学到的越多(即原始数据集的比例越大),其准确率就越高,但隐私级别就会相应降低.在构建代理数据集时,可以通过添加噪声(例如使用差分隐私技术,如拉普拉斯机制[33]和指数机制[34])到标签中来增加隐私.然而,这些噪声也会增加分类错误.为了达到更好的权衡,PPMT 使用迭代方法在确保充分准确率的前提下提升隐私性.该迭代过程在第 4 步中描述.

第 4 步:训练代理模型.在这一步中,使用第 3 步构建的代理数据集训练一个代理模型.然后,重复第三步和第四步进行迭代.在每次迭代中,上一次迭代训练出的代理模型被视为新的旗帜模型,并用于构建一个新的代理数据集,然后在该数据集上训练一个新的代理模型.由于预测的标签是由旗帜模型计算得出的,因此训练会保留旗帜模型中分类的基本准则,这些准则也会持续被后续代理模型所学习.

现在我们对这一步骤提供一些见解.在每次迭代结束后,需要更新旗帜模型.原因是,如果在 PPMT 中始终使用一个在原始数据集上训练的模型作为旗帜模型,那么在后续迭代中训练的代理模型都会继承这个模型的基本分类准则.最终的结果是不同迭代中得到的代理模型非常相似,使得迭代完全没有意义.相比之下,使用前一次迭代中的代理模型作为新的旗帜模型,可以确保新模型在每次迭代中得到隐私性提高,但也会导致了模型效用的轻微下降.

PPMT 中的每个迭代本质上是隐私性和功能性之间的权衡,因为隐私性始终在提升,而功能性始终在降低.最有趣的部分是,PPMT 确保一个模型的隐私增长速度远快于其功能性下降.在每轮迭代中,新的代理模型总是根据目标函数进行训练,最小化该函数的过程也是新的代理模型与新的旗帜模型在新的代理数据集上预测结果差异不断变小的过程.模型将样本空间划分为不同类别的能力在每次迭代中得以保留.出于这个原因,尽管模型功能性下降不可避免,但它呈现出缓慢且线性的趋势.相比之下,在训练过程中没有保留全局属性的目标或约束条件.旗帜模型记住的任何属性都会逐渐被代理模型所遗忘.由于这个原因,代理模型的隐私性呈现出指数级增长的趋势.

第 5 步:设置停止条件.在这一步中,模型提供者确定一个关于模型准确率损失的容忍阈值,以规定模型的最差可接受功能性.当模型准确率损失高于这个阈值时,PPMT 就会停止迭代.

PPMT 不依赖于特定的模型架构或机器学习算法.尽管在对不同的模型进行属性推理攻击防御时效果存在差异,但在训练模型时使用相同设置的 PPMT 都可以达成相应的防御目的.这是因为 PPMT 更改的是训练数据,而不会改变机器学习算法或是模型的架构.

5　算法细节

表 1 中介绍了本节所使用的符号.基于这些符号,算法 1 中给出了 PPMT 算法.

表 1　符号说明

符号	含义
x,y,X,Y	样本,标签,样本集合,标签集合
$M,L(\cdot,\cdot),\Theta$	分类算法,损失函数,参数域
f^*,f_θ	目标模型,内部参数为 θ 的模型
$P,\bar{P},f_{\mathcal{A}}$	全局属性,与 P 相反的属性,攻击分类器
f,f'	旗帜模型,代理模型
S,S',S''	原始(私有)数据集,虚拟数据集,代理数据集
$\mathcal{D},\mathcal{D}_S,\mathcal{D}_u$	样本域,结构化数据集样本域,非结构化数据集样本域
x_i,A_i	样本 x 的第 i 个属性,第 i 个属性的域
x',X',y'	构建的样本,构建的样本集合,旗帜模型预测的标签
α,ϵ	代理数据集中虚拟数据集的占比,准确率损失容忍阈值
η,η'	原始模型的准确率,代理模型的准确率

算法 1.　PPMT 算法.

输入:原始数据集 S,阈值 ϵ

输出:最终的代理模型 f'

BEGIN

1. 在 S 上训练出旗帜模型 f

2. 计算 S 的样本域 $\mathcal{D}$

3. WHILE TRUE DO
4. 从 $\mathcal{D}$ 中采样选取样本集合 X'
5. 使用 f 为 X' 打标签来构建虚拟数据集 S'
6. 按一定比例混合 S 和 S' 得到代理数据集 S''
7. 在 S'' 上训练代理模型 f'
8. IF$\eta-\eta'\geqslant\epsilon$ THEN
9. RETURN f'
10. END IF
11. $f\leftarrow f'$
12. END WHILE

END

PPMT 将原始数据集 S 和阈值 ϵ作为输入，并输出最终的代理模型 f'. 算法在第 1 步(算法的第 1 行)训练旗帜模型 f，在第 2 步(第 2 行)计算样本域 $\mathcal{D}$，然后以迭代的方式训练代理模型 f'. 每次迭代中，在第 3 步(第 4—6 行)构建代理数据集 S''，在第 4 步(第 7 行)训练代理模型 f'，在第 5 步(第 8 行)检查停止条件. 如果条件得到满足则返回代理模型 f'，否则更新旗帜模型 f 并进入第 2 步.

5.1 训练旗帜模型

正如第 4 节中所讨论的，旗帜模型所达到的准确率是后续代理模型所能达到的准确率上限. 因此，确保旗帜模型达到最佳准确率水平非常重要. 模型提供者通过解决一个优化问题，在私有数据集 S 也就是原始数据集上训练旗帜模型 f. 该优化问题如下所示：

$$\theta^{*}=\underset{\theta\in\Theta}{\operatorname{argmin}}\sum(L(f_{\theta}(x),y)) \tag{3}$$

其中 θ^{*} 表示模型 f^{*} 的内部参数，$L(\cdot,\cdot)$ 表示损失函数. 这个优化问题与方程(1)中定义的有区别，即没有使用 L2 范数来防止过拟合.

尽管过拟合会增强目标模型中可用于属性推理攻击的隐藏模式，但我们仍不使用像 L2 范数这样的正则化方法来限制旗帜模型的复杂性. 这样可以使旗帜模型更好地拟合甚至过度拟合原始数据集. 过拟合问题可以通过后续迭代中使用来自虚拟数据集而非原始数据集中的数据训练解决. 同时，由于缺乏正则化，旗帜模型更容易在推理攻击中暴露信息. 这样可以更好地展示本文方法的有效性.

5.2 确定样本域

PPMT 必须确保其构建的虚拟数据集达到以下两个方面的要求. 首先，应能够在该数据集上训练出一个有足够准确率的模型. 如果虚拟数据集中的某些样本标签不正确，将会使模型在训练阶段中产生困惑，进而导致其准确率不足. 因此，PPMT 使用训练有素的旗帜模型为每个样本确定合理的标签，即预测值最高的标签. 其次，虚拟数据集中的每个样本应是旗帜模型的合法和合理输入. 这通过确定原始数据集中的样本域 $\mathcal{D}$ 来保证. 具体地说，我们将结构化数据集的样本域表示为 $\mathcal{D}_S$，非结构化数据集的样本域表示为 $\mathcal{D}_U$.

设 S 为原始数据集，由样本-标签对组成：

$$S=\{(x,y)\mid x\in X,y\in Y\} \tag{4}$$

其中 X 是具有 d 个维度的样本集合，Y 是标签集合. 如果 S 是结构化数据集，那么每个样本 $\boldsymbol{x}$ 是一个 d 维的向量：

$$\boldsymbol{x}=(x_1,x_2,\cdots,x_d) \tag{5}$$

其中 $x_i(1\leqslant i\leqslant d)$ 表示 x 的第 i 个属性值. 该属性值的域表示为 A_i，由原始数据集中所有样本的该属性值组成：

$$A_i=\{x_i\mid\boldsymbol{x}\in X\} \tag{6}$$

样本域由所有的属性域一起组成：

$$\mathcal{D}_S=\{(x_1,x_2,\cdots,x_d)\mid x_i\in A_i,1\leqslant i\leqslant d\} \tag{7}$$

设 S 是一个非结构化数据集(例如，由图像或文本组成的数据集). S 中样本的维度数可能会很大，每个维度(属性)的取值范围也可能会非常广. 因此，最终的样本域是一个非常庞大的空间. 以图像数据集为例，一张表示为 x 的图像包含 $a\times b$ 个像素点，而每个像素点的取值域为 $\mathbb{P}$：

$$x=(x_{i,j})\in\mathbb{P}^{a\times b} \tag{8}$$

样本域表示为

$$\mathcal{D}_u=\{(x_{i,j})\in\mathbb{P}^{a\times b}\} \tag{9}$$

举个例子，对于一个低分辨率($a=b=16$)的 256 色位图，有 $|\mathbb{P}|=256$，其样本域大小为：

$$|\mathcal{D}_u|=|\mathbb{P}|^{ab}=(256^{16})^{16}\approx 3.23\times 10^{616}.$$

当样本域规模太大的时候，除非模型提供者能够构建足够数量样本的虚拟数据集，否则很难训练出具有足够准确率的代理模型. 这也是本论文所提出的 PPMT 方法的最主要局限.

5.3 构建虚拟数据集

虚拟数据集是由噪声样本组成的非现实数据集. PPMT 按照两个阶段来生成一个虚拟数据集 S'.

在第一阶段，从步骤 2 获得的样本域 $\mathcal{D}$ 中根据均匀概率分布随机选择 n 个样本，组成集合 X'. 集合中第 i 个样本由 x'_i 表示，

$$x'_i=(Rand(A_1),Rand(A_2),\cdots,Rand(A_d)) \tag{10}$$

$Rand(A)$ 表示从属性域 A 中采样的函数. 采

样得到的样本会被放入 X' 中.最后,我们有

$$X'=\{x'_i \in \mathcal{D} \mid 1 \leqslant i \leqslant n\} \tag{11}$$

需要注意的是,样本集 X' 由噪声样本组成,并不反映原始数据集的任何全局属性.

在第二阶段中,PPMT 会给所有选取的样本打上标签来构建虚拟数据集.每个样本 $x' \in X'$ 被传递给旗帜模型 f 以产生一个标签 $y' \in Y$.接着每个样本—标签对会被放入虚拟数据集中:

$$S'=\{x', f(x') \mid x' \in X'\} \tag{12}$$

虚拟数据集 S' 能够描述样本域中不同类别的边界.采样的 n 越大,样本域中所选的样本密度就越高,描述的边界也越清晰.此外,旗帜模型中的分类准则被进一步加强.虽然虚拟数据集由噪声样本组成,但由于含有旗帜模型的预测值,所以能让模型进一步学习不同类别的边界.

5.4　训练代理模型

为了更好地在隐私性和功能性之间做出权衡,PPMT 以迭代的方式训练代理模型.每轮迭代中,f' 为要训练的代理模型,本轮迭代结束后,PPMT 会用 f' 替换旗帜模型 f.

如第 4 节所述,PPMT 将随机采样的虚拟数据集和原始数据集中的数据混合在一起构建成一个新的训练数据集,称为代理数据集(表示为 S''),用于训练本次迭代的代理模型.在代理数据集中,给定比例(表示为 α)的数据来自虚拟数据集.模型提供者可以通过调整 α 在隐私性和功能性之间进行权衡.当 α 较大时,S'' 中来自原始数据集的真实样本—标签对将减少,那么在 S'' 上训练的代理模型就会获得更好的隐私性和略微的功能性损失.

人们可能会想到设置一个很小的 α 并增加迭代次数来提高最终模型的准确率,因为模型的隐私性表现为快速的指数级增长,而功能性则表现为缓慢的线性下降.然而这种方法并不可行.因为较小的 α 意味着用于训练的 S'' 中包含更多来自原始数据集的真实样本.这会导致两个问题.第一个也是最重要的一个是涉及隐私性的上限,因为具有较小 α 的代理数据集具有与私有数据集更相似的全局属性,可能会导致无法防御住推理攻击.第二个问题涉及到计算开销,因为更多的迭代轮次意味着需要训练更多的模型.一个较好的策略是在迭代中使用缓慢增大的 α,以保证隐私性在不断增加的同时功能性也能得以保留.在前几轮迭代中,使用较小的 α 来保证功能性.在若干轮迭代之后,使用更大的 α 以确保能达到足够的隐私上界.

值得注意的是,α 的值是模型提供者唯一可以保留的秘密,因为根据开放设计原则[35],攻击者不仅对目标模型有充分的了解和访问权限,而且还知道训练模型时是否使用了 PPMT 以及 PPMT 的每一个设计细节.第 6.6 节中的实验表明,PPMT 的隐私性依赖于 α 的保密.

5.5　设置停止条件

训练模型时经历的 PPMT 迭代次数越多,模型所取得的隐私性就越好.设ϵ为模型提供者可以忍受的最大准确率损失,η 为在原始数据集上训练的模型准确率,η' 为代理模型的准确率.则停止条件如下:

$$\eta-\eta' \geqslant \epsilon \tag{13}$$

当满足停止条件时,返回当前训练出的代理模型.

6　实验验证

在这一节中,我们评估了由 PPMT 训练出的最终模型的准确率表现和在属性推理攻击下的隐私性表现.进一步,我们探究了 PPMT 中超参数的影响,PPMT 在可能攻击下的鲁棒性以及 PPMT 的运行时长.

6.1　数据集

本文的实验在以下三个经典数据集上进行.

(1)US Census Income

美国人口普查收入数据集[36]是从美国人口普查局在 1994 年和 1995 年进行的人口调查结果中提取而来.它包含有加权的人口普查数据.该数据集是一个结构化的数据集,包含 41 个与人口和就业相关的属性,例如性别、种族和受教育情况等.我们使用该数据集进行二元分类任务.任务的目标是预测一个个体的年收入是否超过 5 万美元.

(2)Heart Disease Data

心脏病数据集[37]包括 4 个心脏疾病诊断数据库:Cleveland, Hungary, Switzerland 和 Long-beach-va.该数据集中的每个样本包含 76 个特征.在本文实验中,我们使用了 14 个特征,包括年龄、性别、胸痛类型、运动引起的心绞痛、空腹血糖等.分类任务是根据这 14 个特征来预测血管造影疾病的状态.

(3)CelebA

一个大规模的人脸识别数据集[38],它拥有超过 200K 张名人的图像.这些图像有丰富的人体姿态和背景,每张图像包含 40 个描述图像中人物的属性注

释.我们使用这个数据集进行人脸微笑检测.

6.2 实验设置

本文的实验在一台配有 24 GB 内存,运行 Windows11 操作系统的电脑上进行.我们使用 Scikit-learn 框架[39]在 Intel i7－8700 CPU 上训练支持向量机模型,并使用 PyTorch 框架[40]在 NVIDIA RTX 3090(24 GB) GPU 上训练所有神经网络.

(1)目标模型架构

本文在两个数据集上使用 FCNN 进行训练.对于美国人口普查收入数据集,我们使用了一个 5 层的 FCNN,其中三个隐藏层的大小分别为 32、16 和 8.在模型训练时,使用 ReLU 作为激活函数,采用学习率为 0.0001 的 Adam 优化器.这个设置与文献[18]中的设置相同.对于 CelebA 数据集,我们使用预训练网络 Facenet[41]来获取大小为 512 的图像特征表示,并在特征表示后面接上包含 2 个隐藏层的 FCNN,隐藏层大小分别为 64 和 16.在模型训练中,同样使用 ReLU 作为激活函数,采用学习率为 0.0001 的 Adam 优化器.在上述两个数据集中,我们选择了准确率高于 84%的模型作为目标模型.对于心脏病数据集,我们使用 SVM 作为分类算法.核函数为径向基函数(RBF),正则化参数大小为 1.0.对于这个数据集,我们选择准确率高于 95%的分类器作为目标模型.

(2)攻击分类器架构

对于属性推理攻击,训练数据集由两种阴影模型组成,即 512 个具有属性 P 的阴影模型和 512 个具有属性 $\bar{P}$ 的阴影模型.值得注意的是,这些模型都是彼此独立的.在每个实验中,我们考虑不同的攻击分类器架构.根据文献[18]的研究,FCNN 内部参数具有置换不变性,而基于集合的攻击架构比基于排序的攻击架构在 FCNN 上更有效.因此,我们使用基于集合的方法来实施对 FCNN 的属性推理攻击.攻击分类器采用 DeepSets 架构[42],它由四个 ϕ 网络和一个 ρ 网络组成,每个 ϕ 网络层的大小为 $42\times32\times8$、$289\times128\times8$、$145\times64\times8$ 和 73×8,ρ 网络层大小为 32×1.本文对 SVM 的属性推理攻击使用了文献[17]中提出的方法实现.攻击通过训练具有 512、256、128 和 16 大小的隐藏层的 FCNN 实现.

为了验证 PPMT 的有效性,我们在以上数据集中训练代理模型.本文的实验旨在:(1)评估 PPMT 在模型隐私性和功能性方面的性能;(2)评估 PPMT 中超参数对其效果的影响,包括停止条件中的准确率阈值ϵ和代理数据集中虚假数据比例 α;(3)评估 PPMT 在可能攻击下的鲁棒性;(4)评估 PPMT 的运行时长.

6.3 功能性表现

我们使用模型准确率来量化模型功能性.实验中一共设计了八种场景来评估 PPMT 的准确率.每个场景由一个原始数据集、一个分类任务和一个目标属性组成.八个场景被分为三组.第一组关于在 US Census Income 数据集上训练的收入预测模型.在这个组里,我们考虑了三个目标属性:男性比例为 30%(该场景记为 E_1),白人比例为 25%(记为 E_2),高中学历比例为 75%(记为 E_3).第二组关于在 Heart Disease Data 数据集上训练的心脏病预测模型.在这个组里,我们考虑了两个目标属性:运动引起的心绞痛比例为 80%(记为 E_4),空腹血糖比例为 50%(记为 E_5).第三组关于在 CelebA 数据集上训练的微笑检测模型.在这个组里,我们考虑了三个目标属性:男性比例为 10%(记为 E_6),佩戴眼镜的比例为 60%(记为 E_7),年轻人的比例为 25%(记为 E_8).所有实验场景的概述在表 2 中展示.

表 2 实验场景概述

场景	数据集	分类任务	目标属性(P)	
E_1	US Census Income	收入预测	男性比例	30%
E_2	US Census Income	收入预测	白人比例	25%
E_3	US Census Income	收入预测	高中学历比例	75%
E_4	Heart Disease Data	心脏病预测	运动引起的心绞痛比例	80%
E_5	Heart Disease Data	心脏病预测	空腹血糖比例	50%
E_6	CelebA	微笑检测	男性比例	10%
E_7	CelebA	微笑检测	佩戴眼镜比例	60%
E_8	CelebA	微笑检测	年轻人比例	25%

对于每个实验场景,我们构建具有属性 P 的原始数据集,并使用 PPMT 来训练代理模型.在实验中设置越大的 α,隐私性的提升和准确率的下降都会越明显.为了评估准确率表现,我们将 α 分别设置为 0、0.5 和 1.我们记录了每个实验场景下代理模型在前 10 次迭代中的准确率.

图 3 显示了使用 PPMT 训练的代理模型所能达到的准确率.随着迭代次数的增加,代理模型的准确率呈线性缓慢下降.每次迭代中准确率的下降通常在 1%以内,有时甚至会出现准确率提高的现象.在实验场景 E_1、E_2、E_3(结构化数据集上使用 FCNN 进行训练的模型)中,经过 10 次迭代后(即将第 0 次迭代和第 10 次迭代的准确率进行比较),当 $\alpha=1$ 时,代理模型的平均准确率下降了 2.7%;当 $\alpha=0.5$ 时,准确率下降了 1.3%;当 $\alpha=0$ 时,准确

率下降了 0.03%. 在实验场景 E_4、E_5(结构化数据集上使用 SVM 进行训练的模型)中,经过 10 次迭代后,当 $\alpha=1$ 时,准确率平均下降了 1.45%;当 $\alpha=0.5$ 时,准确率下降了 0.6%;当 $\alpha=0$ 时,准确率下降了 0%. 在实验场景 E_6、E_7、E_8(非结构化数据集上使用 FCNN 进行训练的模型)中,经过 10 次迭代后,当 $\alpha=1$ 时,准确率平均下降了 3.63%;当 $\alpha=0.5$ 时,准确率下降了 1.73%;当 $\alpha=0$ 时,准确率下降了 0.06%.

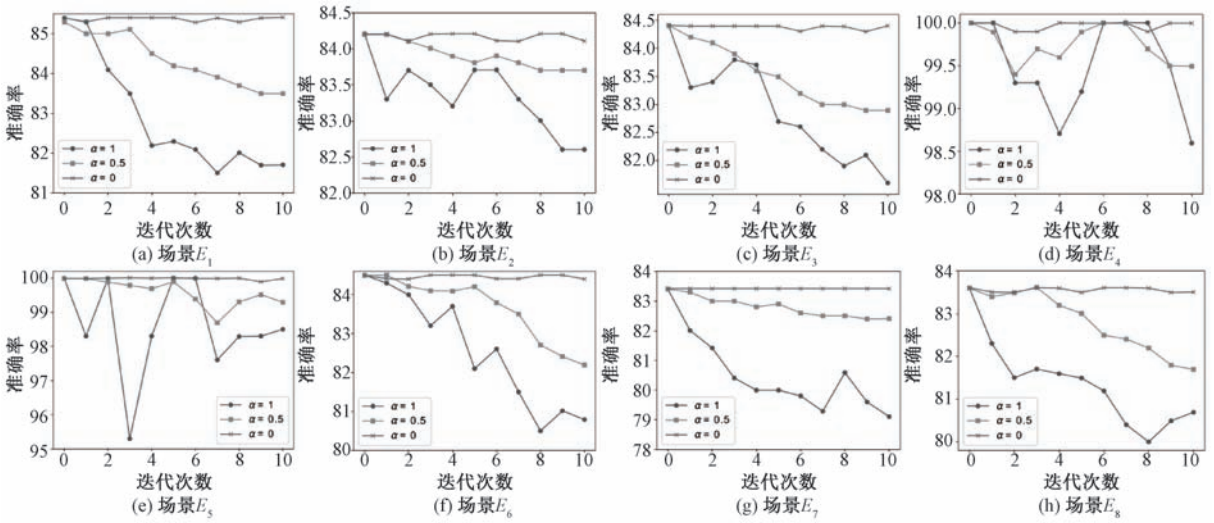

图 3　不同场景下前 10 次迭代后的准确率

我们发现,当每次迭代中只使用虚假数据集(即 $\alpha=1$)时,模型的准确率迅速下降. 当只使用原始数据集时(即 $\alpha=0$),模型的功能性几乎没有减少. 在实验场景 E_1、E_2 和 E_3 中,准确率的下降速度比实验场景 E_6、E_7 和 E_8 中的要慢. 这是因为非结构化数据集具有更大的样本域. PPMT 在非结构化数据集上表现不佳,详细原因已经在 5.2 节中阐述. 在实验场景 E_4 和 E_5 中,准确率的下降速度比其他场景要慢得多. 这意味着 PPMT 在简单机器学习算法上的表现要优于深度神经网络.

6.4　隐私性表现

我们使用模型在属性推理攻击下预测结果的改变来量化 PPMT 的隐私性能. 由于属性推理攻击主要通过二分类任务进行,所以我们在同一目标模型上实施多次属性推理攻击,并选择具有最高置信度的属性作为目标模型(私有数据集)的预测属性. 当考虑每一对属性 $<\gamma_1,\gamma_2>$ 时,将 γ_1 和 γ_2 作为要训练的攻击分类器的二分类标签. 攻击分类器的输出层使用 $Sigmoid$ 作为激活函数,因此攻击分类器的预测置信度在 $[0,1]$ 之间. 当置信度小于 0.5 时,预测的标签为 γ_1. 否则,预测标签为 γ_2.

为了评估隐私性能,我们将 α 设为 1. 在实验场景 E_1 中,我们选择对第 0、2、4、6、8 和 10 次迭代后训练出的代理模型进行属性推理攻击,并将每个标签所具有的置信度画在了图 4. 在 E_1 中,原始数据集的目标属性是“男性比例为 30%”. 0 次迭代时,属性推理攻击得到的结果在 $\rho=30\%$ 处具有最大的置信度,因此认为该模型具有男性比例为 30% 的属性. 这与原始数据集的属性相同,是合理的,因为 0 次迭代时模型的隐私信息尚未得到保护. 在第 2、4、6、8 和 10 次迭代后,属性推理攻击对代理模型男性比例的预测结果分别是 60%、100%、100%、100% 和 100%.

在其他七个实验场景中,我们以类似的方式预测原始数据集的属性,结果如表 3 所示. 我们发现,PPMT 逐渐改变了从代理模型中推理出的属性. 随着迭代次数的增加,任何推理出的属性总是朝着同一个方向变化. 不同的数据集、不同的模型可能会导致属性变化的方向不同,变化速度也不同. 在实验场景 E_4 和 E_5 中,预测属性的变化速度比 E_1、E_2 和 E_3 快,而后者又比 E_6、E_7 和 E_8 快. 使用 PPMT 训练的 SVM 模型,功能性下降比 FCNN 模型慢,而隐私性增加比 FCNN 模型快. 这再一次说明了 PPMT 在简单机器学习算法上的表现优于深度神经网络.

6.5　超参数影响

我们现在评估 PPMT 中超参数不同选值带来的影响,包括第 5.4 节中描述的比例 α 和第 5.5 节中描述的阈值 ϵ. 由模型提供者设定的 ϵ 值决定了 PPMT 在训练最终代理模型时所经历的迭代次数. 因此,我们通过比较模型在不同迭代次数后达到的

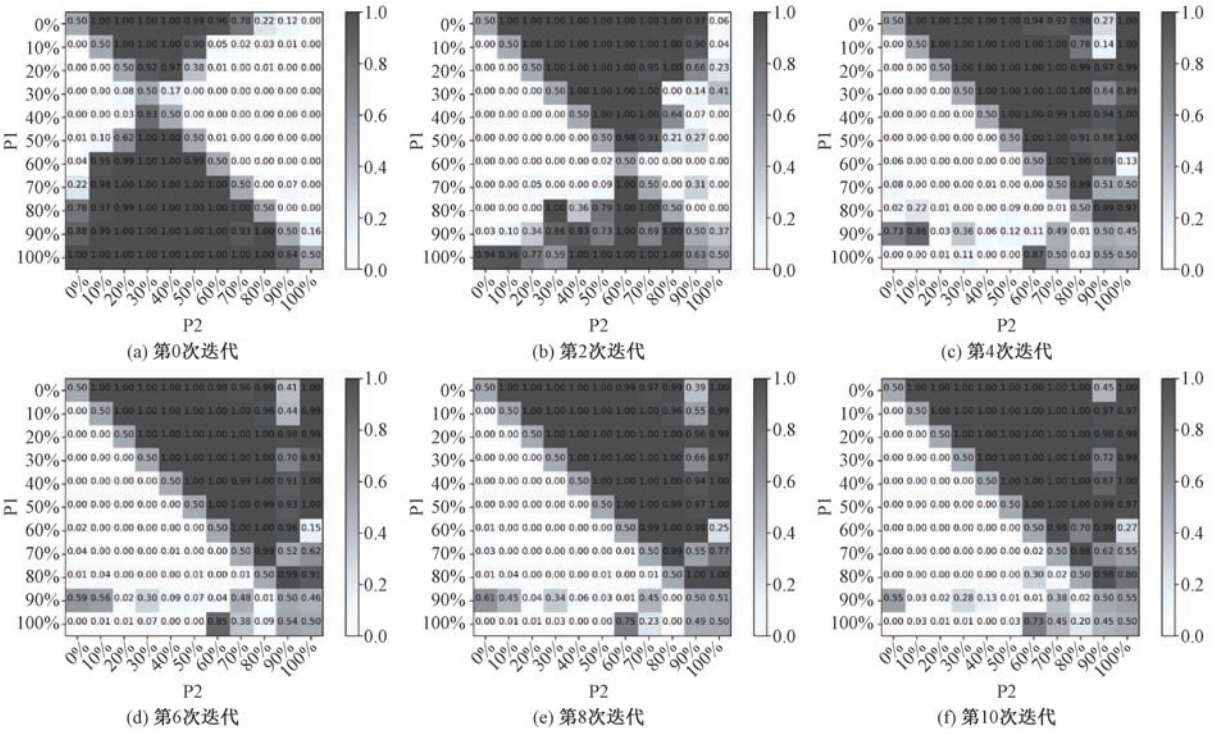

图 4 第 0、2、4、6、8、10 次迭代后的属性推理攻击结果

表 3 隐私性表现

	每轮迭代后的属性预测(%)										
场景	0-th	1-st	2-nd	3-rd	4-th	5-th	6-th	7-th	8-th	9-th	10-th
E_1	30	40	60	90	100	100	100	100	100	100	100
E_2	25	50	90	100	100	100	100	100	100	100	100
E_3	75	30	10	0	0	0	0	0	0	0	0
E_4	80	20	0	0	0	0	0	0	0	0	0
E_5	50	0	0	0	0	0	0	0	0	0	0
E_6	10	20	30	50	70	90	100	100	100	100	100
E_7	60	40	20	10	0	0	0	0	0	0	0
E_8	25	50	90	100	100	100	100	100	100	100	100

准确率来评估ϵ值的影响. α 值决定了代理数据集中虚拟数据的比例,我们通过在实验场景 E_1 的迭代中使用不同的 α 进行训练来评估该值的可能影响. 不同 α 值下训练出的代理模型的准确率如图 5(a)所示(折线自上而下为 $\alpha=0$ 到 $\alpha=1$),隐私性能如图 5(b)所示(折线自上而下为 $\alpha=1$ 到 $\alpha=0$).

随着 α 的增加,用于训练代理模型的数据集中来自原始数据集的真实样本—标签对将减少. 因此,代理模型在隐私性方面表现更好,但准确率较差. 特别是当 $\alpha=0$ 时,旗帜模型和所有代理模型直接在原始数据集上进行训练,这意味着随着迭代次数的增加,准确率和隐私性能几乎没有变化. 相反,当 $\alpha=1$ 时,准确率和隐私性能都会快速变化. 在这种设置下,需要的迭代次数相对较小,不然难以实现隐私性和功能性之间的权衡. 但是我们观察到,无论在哪种 α 设置(除了 $\alpha=0$)下,最终训练的模型总能达到足够的功能性和隐私性,只是需要迭代的次数不同. 因此参数 α 的选择只与迭代次数相关,与最终模型的隐私性和功能性几乎无关.

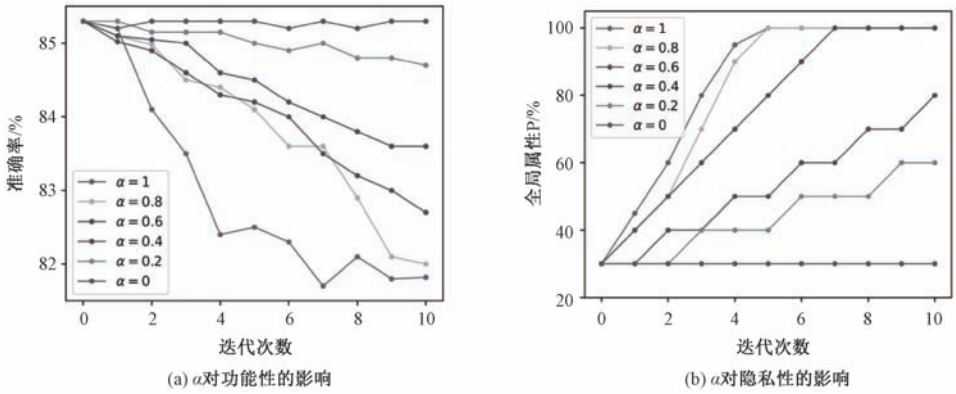

图 5 不同迭代次数下,超参数 α 选值对 PPMT 效果的影响

6.6　不同采样方法的影响

在上文的介绍中,PPMT 的虚拟数据集是通过在样本域中根据均匀概率采样得到的.为了探究不同采样方法对于 PPMT 性能的影响,我们在 E_1 场景下进行实验.对于人口普查收入数据集中的连续特征,我们使用了正态分布进行采样.而对于离散的性别特征,我们分别选择了 0%、50%、70%以及 100%的男性比例进行采样,实验结果如表 4 所示.我们发现,随着迭代次数的增加,模型被推理出来的属性同样也是朝着一个方向变化.因此不同的采样方法对于 PPMT 的隐私保护效果没有太大的影响.

表 4　不同采样方法下的隐私性表现

	每轮迭代后的属性预测(%)										
场景	0-th	1-st	2-nd	3-rd	4-th	5-th	6-th	7-th	8-th	9-th	10-th
正态采样/男性 0%	30	20	10	0	0	0	0	0	0	0	0
正态采样/男性 50%	30	40	60	70	80	100	100	100	100	100	100
正态采样/男性 70%	30	40	40	50	60	70	70	100	100	100	100
正态采样/男性 100%	30	50	60	60	70	90	100	100	100	100	100

6.7　可能攻击下的鲁棒性

基于开放设计原则,我们应该始终假设攻击者能够获取目标模型的全面知识和访问权限,包括目标模型在训练过程中是否使用了 PPMT,以及 PPMT 设计和实现的每个细节.只有模型提供者在训练特定模型时设定的超参数 α 可以被保留为秘密,不被攻击者所掌握.而超参数ϵ不能假设为秘密,因为其可以通过测量阴影模型准确率和目标模型在阴影数据集上的准确率之差来估计.假设攻击者同时得到了混合比例 α 和 PPMT 迭代次数,就可以发动改进的属性推理攻击,对私有数据集的隐私安全产生威胁.基于这个原因,本文还进行了实验来评估 PPMT 在可能攻击下的鲁棒性.

在属性推理攻击中,可以利用的目标模型信息除了传统的内部参数以外,还可以是模型在不同数据上的功能性表现(即识别准确率).因为模型的功能性表现与训练所使用的数据具有直接关联,蕴含攻击者感兴趣的全局属性信息.本文使用 η_G 来表示训练后模型的全局准确率(即在整个阴影数据上的准确率),使用 η_{A_i} 表示模型在给定属性 A_i 不同取值的数据分区上实现的准确率分布.我们感兴趣的是,是否可以利用这些额外信息和混合比例 α ,倒推 PPMT 训练过程的迭代次数,进而实现一种改进的属性推理攻击.值得注意的是,我们并没有使用模型准确率损失容忍阈值ϵ,因为它所携带的信息已经蕴含在了全局准确率 η_G 之中.

本文考虑了 7 种可能的攻击方式,其不同点在于使用的特征表示中是否包含以下值/超参数:全局准确率 η_G ,任意给定属性 A_i 的准确率分布 η_{A_i} ,以及模型提供者的秘密 α .这 7 种可能的攻击方式被表示为:

(1) $\mathcal{A}_{\eta_G}$,使用全局准确率 η_G 作为特征.

(2) $\mathcal{A}_{\eta_{A_i}}$,使用准确率分布 η_{A_i} 作为特征.

(3) $\mathcal{A}_{\eta_G+\eta_{A_i}}$,使用 η_G 和 η_{A_i} 作为特征.

(4) $\mathcal{A}_{\alpha}$,使用模型提供者的秘密 α 作为特征.

(5) $\mathcal{A}_{\eta_{A_i}+\alpha}$,使用 η_{A_i} 和 α 作为特征.

(6) $\mathcal{A}_{\eta_G+\alpha}$,使用 η_G 和 α 作为特征.

(7) $\mathcal{A}_{\eta_G+\eta_{A_i}+\alpha}$,使用 η_G、η_{A_i} 和 α 作为特征.

在上述的可能攻击下,本文评估了 PPMT 训练方法的鲁棒性.本文选择了三种分类算法来训练攻击分类器,包括 K 近邻算法(KNN)、支持向量机(SVM)和随机森林(RF).本文在八个实验场景中针对每种分类算法训练了 3000 个目标模型,并对每个目标模型进行了所有 7 种可能攻击.在每个模型的训练过程中,我们记录了前五次 PPMT 迭代后模型的全局准确率 η_G 、准确率分布 η_{A_i} 和虚拟数据集占比 α .接着,我们统计了每种分类算法和每种可能攻击组合下的攻击成功率,并使用成功率来量化相应可能攻击下 PPMT 的鲁棒性.在八个实验场景中,我们得出了相同的实验结论,其中 $E1$ 场景中的实验结果在表 5 中给出.从表 5 中可以看出除了最后两种可能攻击以外,其他攻击的平均成功率在 22%至 33%之间,构不成隐私性威胁.

在最后两种攻击中,全局准确率 η_G 和模型提供者的秘密 α 都被当作特征使用,攻击成功率接近完美.其中 η_G 应该被视为是公开的,因为攻击者可以多次在阴影数据集上迭代进行预测来得到近似的该值.这意味着,PPMT 的隐私性依赖于超参数 α 的保密.

表 5 7 种可能攻击的成功率

分类算法	$\mathcal{A}_{\eta_G}$	$\mathcal{A}_{\eta_{A_i}}$	$\mathcal{A}_{\eta_G+\eta_{A_i}}$	$\mathcal{A}_{\alpha}$	$\mathcal{A}_{\eta_{A_i}+\alpha}$	$\mathcal{A}_{\eta_G+\alpha}$	$\mathcal{A}_{\eta_G+\eta_{A_i}+\alpha}$
KNN	28.41%	28.83%	28.44%	30.61%	33.75%	95.31%	94.73%
SVM	29.42%	29.90%	30.23%	31.81%	27.22%	94.53%	94.50%
RF	26.49%	24.60%	23.72%	27.61%	24.96%	94.72%	93.00%

6.8 运行时长

我们在 E_1 和 E_4 的情景下进行实验评估 PPMT 的运行时长.我们记录了 10 次迭代完成后消耗的时间.其中 E_1 中数据集大小为 12000,而 E_4 中的数据集大小为 3000.实验结果如图 6 所示(柱形从左往右分别代表 $\alpha=0$、$\alpha=0.5$、$\alpha=1$),运行时间随着迭代次数的增加呈线性增长趋势.

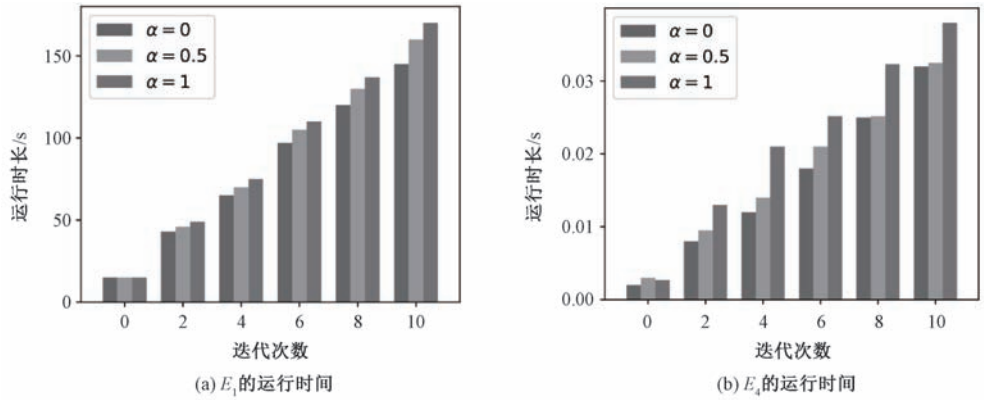

图 6 不同迭代次数下,超参数 α 选值对 PPMT 运行时长的影响

7 结论与展望

本文提出了一种可以保护隐私的模型训练方法,名为 PPMT.该方法通过构建代理数据集替代私有数据集,并在代理数据集上训练代理模型的方法,在保护私有数据集全局属性的同时,确保模型具有足够的功能性.PPMT 没有使用简单的数据集构建技术,而是采用迭代方法在隐私性和功能性之间取得更好的平衡.PPMT 进行多次迭代的依据是,代理模型的功能性只会缓慢地线性下降,而隐私性则是快速指数增加.实验证明,PPMT 在功能性、隐私性和鲁棒性方面均具有良好的表现.据我们所知,本文是首次提出属性推理攻击防御方法的研究工作.

本文所提出的防御方法有待进一步优化.首先,代理数据集中的虚拟数据是通过均匀采样获得的,在未来可以探寻效果更优的采样方法,来降低模型功能性的损失,特别是非结构化数据上的模型.其次,目前没有工作能够评估模型全局属性隐私性,因此需要研究一个具体的量化指标以评估不同防御策略.最后,可以在更多应用场景下评估我们的防御方法,比如目标检测、语义分割、自然语言处理以及图神经网络等等.

参 考 文 献

[1] Scott Alfeld, Xiaojin Zhu, Paul Barford. Data poisoning attacks against autoregressive models//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA, 2016:1452-1458

[2] Raef Bassily, Adam Smith, Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds//Proceedings of the 55th IEEE Annual Symposium on Foundations of Computer Science. Philadelphia, USA, 2014: 464-473

[3] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA, 2017: 27-38

[4] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018:6106-6116

[5] Wang Yizhen, Kamalika Chaudhuri. Data poisoning attacks against online learning. arXiv preprint arXiv:1808.08994, 2018

[6] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, Fabio Roli. Is feature selection secure against

training data poisoning? //Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 1689-1698

[7] Nicholas Carlini, David Wagner. Towards evaluating the robustness of neural networks//Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[8] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 10-17

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[10] Ian J Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[11] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, Ananthram Swami. The limitations of deep learning in adversarial settings//Proceedings of the 2016 IEEE European Symposium on Security and Privacy. Saarbrucken, Germany, 2016: 372-387

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv: 1312.6199, 2013

[13] Jacson Rodrigues Correia-Silva, Rodrigo F Berriel, Claudine Badue, Alberto F de Souza, Thiago Oliveira-Santos. Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data//Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil, 2018: 1-8

[14] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, Ananthram Swami. Practical black-box attacks against machine learning//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 506-519

[15] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, Thomas Ristenpart. Stealing machine learning models via prediction APIs//Proceedings of the 25th USENIX Security Symposium (USENIX Security 16). Austin, USA, 2016: 601-618

[16] Wang Binghui, Neil Zhenqiang Gong. Stealing hyperparameters in machine learning//Proceedings of the 2018 IEEE Symposium on Security and Privacy. San Francisco, USA, 2018: 36-52

[17] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks, 2015, 10(3):137-150

[18] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada, 2018: 619-633

[19] Luca Melis, Congzheng Song, Emiliano De Cristofaro, Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning//Proceedings of the 2019 IEEE Symposium on Security and Privacy . San Francisco, USA, 2019: 691-706

[20] Junhao Zhou, Yufei Chen, Chao Shen, Yang Zhang. Property inference attacks against gans. arXiv preprint arXiv: 2111.07608, 2021

[21] Wang Xiuling, Wendy Hui Wang. Group property inference attacks against graph neural networks//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. Los Angeles, USA, 2022: 2871-2884

[22] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov. Membership inference attacks against machine learning models//Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 3-18

[23] Jamie Hayes, Luca Melis, George Danezis, Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. Proceedings on Privacy Enhancing Technologies (PoPETs), 2019, 2019(1): 133-152

[24] Luca Melis, Congzheng Song, Emiliano De Cristofaro, Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning//Proceedings of the 2019 IEEE Symposium on Security and Privacy. San Francisco, USA, 2019: 691-706

[25] Yuan Xiaoyong, Zhang Lan. Membership inference attacks and defenses in neural network pruning//Proceedings of the 31st USENIX Security Symposium, USENIX Security. Boston, USA, 2022: 4561-4578

[26] Matt Fredrikson, Somesh Jha, Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. Denver, USA, 2015:1322-1333

[27] Zhang Yuheng, Jia Ruoxi, Pei Hengzhi, Wang Wenxiao, Li Bo, Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020:250-258

[28] Zhang Tong. Solving large scale linear prediction problems using stochastic gradient descent algorithms//Proceedings of the International Conference on Machine Learning. Banff, Canada, 2004: 69

[29] Diederik P Kingma, Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[30] Bernhard E Boser, Isabelle M Guyon, Vladimir N Vapnik. A

training algorithm for optimal margin classifiers//Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, USA, 1992: 144-152

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data//Proceedings of the Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2017: 1273-1282

[32] Reza Shokri, Vitaly Shmatikov. Privacy-preserving deep learning//Proceedings of the 22nd ACM SIGSAC conference on Computer and Communications Security. Denver, USA, 2015: 1310-1321

[33] Cynthia Dwork. Differential privacy//Proceedings of the International Colloquium on Automata, Languages, and Programming. Venice, Italy, 2006: 1-12

[34] Frank McSherry, Kunal Talwar. Mechanism design via differential privacy//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). Providence, USA, 2007: 94-103

[35] Jerome H Saltzer, Michael D Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 1975, 63(9):1278-1308

[36] Dheeru Dua, Casey Graff. UCI machine learning repository, 1996

[37] Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano. The heart disease dataset. http://archive.ics.uci.edu/ml/datasets/Heart+Disease, 1988

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang. Deep learning face attributes in the wild//Proceedings of International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 3730-3738

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 2011, 12:2825-2830

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 8024-8035

[41] Florian Schroff, Dmitry Kalenichenko, James Philbin. Facenet: A unified embedding for face recognition and clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815-823

[42] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, Alexander J Smola. Deep sets//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 3391-3401

DONG Kai, Ph. D., associate professor. His main research interests include privacy enhancing techniques, artificial intelligence security, Internet of Things and system security.

JIANG Chi-Hao, M. S. candidate. His main research interests focus on artificial intelligence security.

LI Xiang, M. S. candidate. His main research interests include artificial intelligence security and explainable machine learning.

LING Zhen, Ph. D., professor. His main research interests include network security and privacy, artificial intelligence security, Internet of Things, mobile system security and trusted computing.

YANG Ming, Ph. D., professor. His main research interests include network security and privacy, and artificial intelligence security.

Background

In recent years, machine learning has been playing an increasingly important role in various fields. However, various attacks have been proposed to induce classifier misclassification or violate model's privacy, property inference attack is one of them.

A property inference attack reveals global properties of the training dataset. It can be performed by using a shadow training technique. It is generally considered that a well-trained public classifier trained on a private dataset is always vulnerable to property inference attacks. One possible reason is that well-trained classifiers are always inevitably overfitted to some extent. Moreover, well-trained classifiers have their own 'affinities' to better describe the majority rather than the minority of data in the training dataset. Overfitting and the said 'affinities' give rise to hidden patterns in the parameters of a target classifier. These hidden patterns are utilized in a property inference attack to reveal the global properties of the training dataset.

In this paper, we propose a privacy preserving classifier training method, named PPMT, as the first defense against property inference attacks. The basic idea is to construct a dataset as the substitution of the private one. However, it is nontrivial to construct such a substitution dataset. Two constraints should be met: On the one hand, the substitution

dataset should guarantee sufficient data utility so that a classifier trained on this dataset is able to achieve an adequate model utility. On the other hand, the substitution dataset should ensure the privacy of any global property of the private dataset is not leaked under a property inference attack. To address the above challenge, PPMT uses a novel mechanism to construct the substitution dataset, and trains the classifier on that dataset in an iterative fashion. PPMT at first trains a model on the private dataset. We name it the pilot model. In each iteration, PPMT randomly samples the input domain, uses the pilot model to label the random samples to construct the substitution dataset, trains a new classifier on that dataset, and replaces the pilot model with this new classifier. After several iterations, the privacy is preserved in the trained classifier while an adequate level of accuracy is also guaranteed.

We conduct experiments on three datasets, to evaluate the performance of PPMT in terms of model utility, privacy, and robustness. The experimental results show that the utility of a trained classifier decreases slowly as the number of iterations increases, while the privacy increases rapidly.

This research is supported by National Key R&D Program of China Grant No. 2023YFC3605804, National Natural Science Foundation of China (62072098, 62072103, 62232004), Jiangsu Key R&D Program (BE2022065-5, BE2022680), Jiangsu Provincial Key Laboratory of Network and Information Security (BM2003201), Key Laboratory of Computer Network and Information Integration of Ministry of Education of China (93K-9), and Collaborative Innovation Center of Novel Software Technology and Industrialization. Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.