基于最大–最小策略的纵向联邦学习隐私保护方法

李荣昌[1] 刘 涛[1] 郑海斌[2,3] 陈晋音[1,3] 刘振广[4] 纪守领[5]

摘 要 纵向联邦学习 (Vertical federated learning, VFL) 是一种新兴的分布式机器学习技术, 在保障隐私性的前提下, 利用分散在各个机构的数据实现机器学习模型的联合训练. 纵向联邦学习被广泛应用于工业互联网、金融借贷和医疗诊断等诸多领域中, 因此保证其隐私安全性具有重要意义. 首先, 针对纵向联邦学习协议中由于参与方交换的嵌入表示造成的隐私泄漏风险, 研究由协作者发起的通用的属性推断攻击. 攻击者利用辅助数据和嵌入表示训练一个攻击模型, 然后利用训练完成的攻击模型窃取参与方的隐私属性. 实验结果表明, 纵向联邦学习在训练推理阶段产生的嵌入表示容易泄漏数据隐私. 为了应对上述隐私泄漏风险, 提出一种基于最大–最小策略的纵向联邦学习隐私保护方法 (Privacy preservation method for vertical federated learning based on max-min strategy, PPVFL), 其引入梯度正则组件保证训练过程主任务的预测性能, 同时引入重构组件掩藏参与方嵌入表示中包含的隐私属性信息. 最后, 在钢板缺陷诊断工业场景的实验结果表明, 相比于没有任何防御方法的 VFL, 隐私保护方法将攻击推断准确度从 95% 下降到 55% 以下, 接近于随机猜测的水平, 同时主任务预测准确率仅下降 2%.

关键词 纵向联邦学习, 属性推断攻击, 隐私保护, 最大–最小策略, 工业互联网

引用格式 李荣昌, 刘涛, 郑海斌, 陈晋音, 刘振广, 纪守领. 基于最大–最小策略的纵向联邦学习隐私保护方法. 自动化学报, 2024, **50**(7): 1373−1388

DOI 10.16383/j.aas.c211233

Privacy Preservation Method for Vertical Federated Learning Based on Max-min Strategy

LI Rong-Chang[1] LIU Tao[1] ZHENG Hai-Bin[2,3] CHEN Jin-Yin[1,3] LIU Zhen-Guang[4] JI Shou-Ling[5]

Abstract Vertical federated learning (VFL) is an emerging distributed machine learning that applies to the data distributed in various institutions to realize the joint construction of privacy preservation machine learning models. It has been widely applied to various fields such as industrial internet, financial lending, and medical diagnosis. Therefore, the privacy security research of vertical federated learning highlights its significance. Aiming at the risk of privacy leakage caused by the embedding exchanged by participants in the vertical federated learning protocol, we propose a general property inference attack initiated by the server. The adversary uses the auxiliary data and the embedding exchanged by the vertical federated learning protocol to train the attack model and steal the target privacy property of the participant. The experimental results show that the embedding representation generated by the vertical federated learning during the training and inference process can reveal the information of the personal private property. To deal with the above proposed privacy leakage risk, proposed a privacy preservation method for vertical federated learning based on max-min strategy (PPVFL), which introduces a gradient regular component to ensure the performance of the main task of the training process and adopts a construction component to hide participant′s privacy property. Finally, in steel defect diagnosis industrial scenarios, compared to VFL without any defense method, privacy-preserving method reduces attack inference accuracy from 95% to below 55%, which is close to the level of random guessing, while the main task only dropped by 2% of the prediction accuracy.

Key words Vertical federated learning (VFL), property inference attack, privacy preservation, max-min strategy, industrial internet

Citation Li Rong-Chang, Liu Tao, Zheng Hai-Bin, Chen Jin-Yin, Liu Zhen-Guang, Ji Shou-Ling. Privacy preservation method for vertical federated learning based on max-min strategy. *Acta Automatica Sinica*, 2024, **50**(7): 1373−1388

收稿日期 2021-12-26 录用日期 2022-06-12

Manuscript received December 26, 2021; accepted June 12, 2022

浙江省自然科学基金青年原创计划 (LDQ23F020001), 国家自然科学基金 (62072406), 国家重点研发计划基金 (2018AAA0100801), 浙江省自然科学基金 (LGF21F020006, LGF20F020016) 资助

Supported by Zhejiang Natural Science Foundation Youth Original Project (LDQ23F020001), National Natural Science Foundation of China (62072406), National Key Research anf Development Projects of China (2018AAA0100801), and Natural Science Foundation of Zhejiang Province (LGF21F020006, LGF20F020016)

本文责任编委 穆朝絮

Recommended by Associate Editor MU Chao-Xu

1. 浙江工业大学信息工程学院 杭州 310023 2. 浙江工业大学计算机科学与技术学院 杭州 310023 3. 浙江工业大学网络空间安全研究院 杭州 310023 4. 浙江大学网络空间安全学院 杭州 310007 5. 浙江大学计算机科学与技术学院 杭州 310007

1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023 2. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023 3. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023 4. School of Cyber Science and Technology, Zhejiang University, Hangzhou 310007 5. College of Computer Science and Technology, Zhejiang University, Hangzhou 310007

随着深度学习在诸多领域取得优异的性能, 工业互联网中不断引入深度学习技术[1−3] 赋能传统企业. 工业互联网的快速发展得益于海量的工业数据和丰富的计算资源. 然而, 随着数据隐私保护法规的颁布[4−5], 企业间难以通过直接交换私有数据的方式训练深度学习模型, 极大制约了工业互联网的快速发展. 联邦学习 (Federated learning, FL) 为上述问题提供了解决方案, 在保证隐私的前提下利用分散在各个机构的数据联合训练机器学习模型.

联邦学习按照机构间数据的分布差异[6], 通常可分为横向联邦学习 (Horizontal federated learning, HFL)、纵向联邦学习 (Vertical federated learning, VFL) 和联邦迁移学习. HFL 适用于参与方间数据特征空间相同、样本空间不同的场景, 其中特征空间指参与方用户的属性信息, 样本空间指参与方数据中的用户成员信息. 例如, 某银行在 A 地区和 B 地区设有分行, 两地业务类似, 即具有相同的特征空间; 用户差异较大, 即具有不同的样本空间. VFL 适用于参与方间的数据具有相同样本空间、不同特征空间的场景. 例如, 来自相同地区的银行和借贷机构, 银行具有该地区的经济状况, 借贷机构具有该地区的信用记录, 2 个公司具有的用户类似, 即具有相同的样本空间; 2 个公司的业务不同, 即具有不同的特征空间. 联邦迁移学习适用于参与方间数据集共享的样本空间和特征空间都有限的场景. 由于现实场景中来自同一个地区的不同机构间的合作日益紧密, VFL 逐渐受到更多关注.

随着 FL 在诸多领域得到应用[7−8], 研究者们关注到 FL 算法本身的隐私安全性, 已有研究分别从参与方和协作方 2 个角度, 讨论 HFL 中良性参与方所面临的隐私泄漏风险. 现有研究表明, 参与方或协作方可利用 HFL 训练过程中传输的中间信息发动成员推断攻击[9−10] 或数据重构攻击[11−12], 使得 HFL 中的参与方遭受隐私泄漏威胁. 现有研究针对 VFL 场景, 仅评估参与方作为攻击者时对良性参与方造成的隐私泄漏风险[13]. 协作方通常被假设为一个诚实可信的第三方, 但在现实场景中, 难以保证协作方是完全诚实可信的. 特别地, VFL 在训练过程中的良性参与方上传的嵌入表示通常包含其关键的原始信息 (包括隐私信息), 该原始信息存在泄漏的风险. VFL 中的隐私信息泄漏带来极大的社会危害. 例如现实场景中, 一个银行和一个借贷机构试图联合构建一个评估用户信誉度的 VFL 系统, 如图 1 所示. 其中, 借贷机构作为协作方和主参与方提供用户的贷款数据, 银行作为从参与方提供用户的金融数据 (如“用户负债”). 在训练过程中, 借贷机构一方面正常参与训练; 另一方面试图从银行窃取用户的“用户负债”数据, 从而恶意地推销高利贷服务. 因此, 研究 VFL 隐私保护方法显得尤为重要.

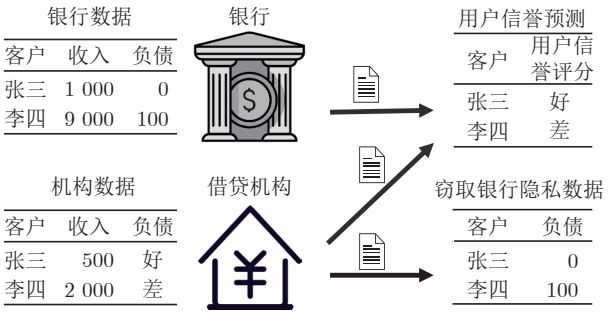


图 1 VFL 隐私泄漏示例

Fig. 1 Examples of VFL privacy leaks

为了评估 VFL 中良性参与方所面临的隐私泄漏风险, 本文提出一种由协作者发起的通用属性推断攻击. 攻击者利用良性参与方在联合训练过程中上传的嵌入表示和收集的样本隐私属性训练一个攻击模型, 并利用训练完成的攻击模型, 推断未知样本的隐私属性. 在基于全连接神经网络 (Fully connected neural network, FCNN) 构建的 VFL 框架上, 通过对实际工业场景的钢板缺陷诊断数据集上的实验结果表明, 仅当攻击者收集到参与方 1% (20 张) 样本隐私属性数据时, 可达到对良性参与方“钢板序列”隐私属性 95% 的攻击推断准确度. 此外, 这种属性推断攻击可同时窃取 VFL 中良性参与方在训练阶段的隐私属性和测试阶段的隐私属性.

VFL 研究中常见的防御方法可分为基于加密的保护方法和基于扰动的保护方法 2 种, 但无法有效防御本文提出的属性推断攻击. 其中, 现有基于加密技术 (如同态加密和多方安全计算) 构建的 VFL 框架[14] 无法防御的主要原因是协作方在基于加密技术的 VFL 协议中, 可获得解密后的真实嵌入表示, 从而发动攻击. 基于扰动的保护方法主要用于防御成员推断攻击, 如差分隐私[15] 通过注入随机噪声, 使得任意 2 个数据记录产生近似的概率, 但对主任务的预测性能损害严重, 且难以防御属性推断攻击[16].

为了同时保护参与方的隐私属性和保证主任务的预测性能, 并且降低防御的时间成本, 本文提出一种基于最大−最小策略的纵向联邦学习隐私保护 (Privacy preservation method for vertical federated learning based on max-min strategy, PPVFL) 方法. 通过对本地模型实施最大化主任务的预测性能和最小化嵌入表示的隐私信息 2 个操作, PPVFL 能够在滤除隐私属性信息的同时, 保证

VFL 主任务预测性能. 通过实验, 验证了本文提出的 PPVFL 能有效降低攻击者发动属性推断攻击的推断准确度, 并对主任务的预测准确率影响较小. 同时, 本文对 PPVFL 的通用性和参数敏感性进行讨论. 最后, 利用 t 分布式随机邻居嵌入 (t-distributed stochastic neighbor embedding, t-SNE) 可视化技术, 对 PPVFL 能有效防御属性推断攻击进行解释.

本文的主要贡献包括以下 3 个方面: 1) 针对常用的 VFL 框架, 提出一种通用的属性推断攻击, 验证了 VFL 在训练和推理阶段存在隐私数据泄漏的风险; 2) 提出一种基于最大–最小策略的纵向联邦学习隐私保护方法, 本地模型前向传播 (localforward) 时破坏嵌入表示和隐私属性间的映射关系, 同时引入梯度正则组件, 实现保护数据隐私与维持主任务预测性能的目标; 3) 通过在 3 个典型的不同模态数据集的实验, 验证了属性推断攻击和 PPVFL 方法的有效性. 此外, 在工业互联网的钢板缺陷诊断场景中, 本文提出的隐私保护框架下, 攻击者的推断准确度从 95% 下降到 55% 以下, 接近于随机猜测水平, 同时主任务的预测准确率仅下降 2%.

1　相关工作

本节介绍 VFL 方法、VFL 隐私泄漏和 VFL 隐私保护方法.

1.1　VFL 方法

近年来, 由于数据隐私保护法规限制了企业间直接交换隐私数据, 工业互联网和 FL 的结合成为新的联合训练模型技术方案, 并在诸多领域得到应用, 如故障检测、智能交通和智慧医疗等领域[16]. Lu 等[17] 提出工业互联网的分散数据采用模型共享的方式完成训练. 为了更加安全地进行数据共享, Dinh 等[18] 结合区块链和 FL 技术, 解决工业互联网中的联合训练问题[19]. 此外, Lu 等[17] 将 FL 应用到车辆互联网, 引入基站实现聚合.

随着 VFL 的不断发展, 其支持的边缘模型类型不断增加, 包括逻辑回归模型、随机森林和神经网络 (Neural network, NN) 等. Sun 等[20] 提出一种去除协作者的逻辑回归 VFL, 其不适用于其他边缘模型. 对于边缘模型为树模型的场景, Cheng 等[21] 提出一种无损的隐私保护集成系统 SecureBoost. 针对边缘模型为神经网络的 VFL, 已有研究提出了不同的 VFL 方法. 现有研究[22] 采用分裂学习的 VFL 框架, 用于广告推荐或图像识别领域. 为了将 VFL 应用于图数据挖掘领域, 出于实际应用场景的需求, 边缘模型为神经网络 VFL 方法得到广泛的应用. 因此, 本文围绕应用最为广泛的基于神经网络的 VFL 方法, 研究其隐私安全性.

1.2　VFL 隐私泄漏

随着 VFL 的应用领域不断拓展, 其算法本身的隐私安全性得到了研究者们的关注. 根据泄漏数据的类型不同, 本文将 VFL 中的隐私泄漏分为标签泄漏、成员信息泄漏和属性信息泄漏 3 类. 1) 标签泄漏. Fu 等[13] 首次针对 VFL 框架, 提出了标签推断攻击, 攻击者利用其本地模型信息, 窃取隐私标签. 2) 成员信息泄漏. 主要发生在 VFL 的数据对齐阶段. 常见 VFL 中的参与方经常在数据对齐阶段公布样本成员信息, 导致其直接泄漏. 3) 属性信息泄漏. Luo 等[22] 提出一种生成回归网络推断参与方的隐私属性数据, 该攻击过程仅发生在 VFL 推理阶段. 综上所述, 现有针对 VFL 的研究都是假设攻击者为参与方的场景, 忽略协作方本身作为攻击者的威胁场景. 此外, 已有攻击不能同时对训练阶段和推理阶段的隐私安全造成威胁.

1.3　VFL 隐私保护方法

研究者们在探索 VFL 方法面临隐私泄漏风险的同时, 提出相应的隐私保护方法. 根据防御原理, VFL 隐私保护方法可以分为基于加密的防御方法、基于扰动的防御方法和基于系统的防御方法 3 类. 表 1 总结了这些方法的优缺点.

表 1　VFL 隐私保护技术优缺点对比

Table 1　Comparison of advantages and disadvantages of VFL privacy protection technology

策略	方法	优点	缺点
基于加密的防御	同态加密[14]	可扩展性强	受限非线性函数
	MPC	准确率高	时间成本较高
基于扰动的防御	差分隐私	有理论保证	性能存在损耗
	梯度压缩[23]	通信成本低	保护效果较弱
基于系统的防御	可信执行环境[24–25]	同时抵御基于硬件攻击	经济成本较高

1) 基于加密的防御方法. Yang 等[6] 提出一种基于线性回归的 VFL 框架, 通过利用加法同态加密技术, 保证参与方交换数据的隐私; Ou 等[14] 提出一种纵向联邦期望最大算法, 引入一种贝叶斯聚合更新机制, 并结合后验采样, 来保证数据的隐私.

2) 基于扰动的防御方法. Liu 等[15] 分析了纵向联合学习的混合差分隐私框架, 将垂直分区的数据中联合学习广义线性模型. Yang 等[23] 基于深度梯度压缩方法, 提出一种通信双方参数共享和梯度压

缩的 VFL. 这种方法可以同时保护隐私和降低通信代价.

3) 基于系统的防御方法. Paramod 等[24] 引入基于可信抽象平台的验证方法, 使用硬件和软件保护将敏感计算与不受信任的软件堆栈隔离. Florian 等[25] 提出一个 Slalom 框架, 有选择地将计算外包给受信任的参与设备. 此类基于系统的方法往往不适用于特定的 VFL 工业互联网场景.

综上所述, 基于加密的防御方法时间成本较高, 通常不适用于基于神经网络的 VFL. 由于基于扰动的防御方法无法对特定的隐私属性进行保护, 所以对主任务的预测性能会造成较大损害. 因此, 如何对 VFL 中的推断攻击设计防御方法, 同时权衡主任务预测性能和隐私保护效果, 仍是挑战和难点.

2 面向 VFL 的隐私保护方法

本节首先给出 VFL 的基本定义, 然后提出一种通用的属性推断攻击方法, 最后设计一种基于最大–最小策略的隐私保护方法.

2.1 基本定义

定义 1 (纵向联邦学习). VFL 旨在建立一个基于分布式数据集的联合模型, 其中分布式数据集具有相同的样本空间以及不同的特征空间. VFL 包含训练阶段和推理阶段 2 个阶段, 2 个阶段通过交换嵌入表示实现原始数据不离开参与方本地的分布式训练, 进而保护数据隐私安全. 图 2 为 VFL 的基本框架, 包含 1 个主参与方、1 个从参与方和 1 个协作方, 其中从参与方的数量可以增加到 m.

定义 2 (主参与方). VFL 建模任务中具有任务标签数据的参与方称为主参与方. 主参与方通常具有部分特征信息, 并且拥有本地模型. VFL 中通常存在 1 个主参与方发起特定任务.

定义 3 (从参与方). VFL 中只具有部分特征信息的参与方称为从参与方. 从参与方具有提取本地嵌入表示的本地模型.

定义 4 (协作方). 协作方负责维护 VFL 功能, 同时用以传输中间通信信息、密钥分发和密文解密工作. 此外, 基于神经网络 VFL 的协作方通常存在顶部模型, 具有计算 VFL 中损失函数的功能.

定义 5 (VFL 的优化目标). 假定 VFL 中共包含 m 个从参与方, 每个从参与方具有数据样本 $\{X_1, X_2, \cdots, X_m\}$; 1 个主参与方, 其具有特征信息 X_a 和标签信息 Y; 1 个协作方. 设定参与方具有本地模型 $f_{\text{local}}(\cdot)$, 其模型参数为 $\theta_1, \theta_2, \cdots, \theta_m$; 主参与方具有本地模型参数为 θ_a, 协作方具有顶部模型 $f_{\text{top}}(\cdot)$, 其模型参数为 θ_{a+1}. VFL 中预测任务的目标函数表示为:

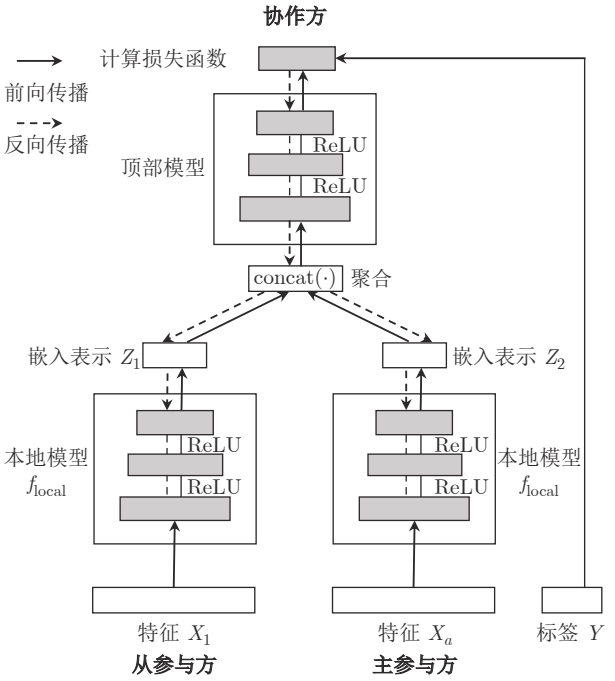


图 2 VFL 框架

Fig. 2 VFL framework

$$\min_{\theta_1, \cdots, \theta_a, \theta_{a+1}} f(\theta_1, \cdots, \theta_a, \theta_{a+1}; X_1, \cdots, X_m, Y) \tag{1}$$

定义 6 (互信息). 互信息属于信息论中的一种信息度量, 表示一个随机变量中包含的另一个随机变量的信息量. 2 个离散的随机变量 X 和 Y 的互信息 I 利用信息熵 H 表示为:

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

2.2 通用的属性推断攻击方法

2.2.1 攻击场景定义

图 3 为 VFL 场景中攻击示意图, VFL 中的通用属性推断攻击的具体定义如下:

1) VFL 场景. 假设 VFL 包含 1 个主参与方、m 个从参与方和 1 个协作方.

2) 攻击者. VFL 中的协作方.

3) 被攻击者. VFL 中的从参与方或主参与方.

4) 攻击目标. 攻击者获取 VFL 中良性参与方的隐私属性, 如“钢板材质”, 从而计划采取高额营销策略. 记推断目标样本标识号信息为 n 至 s 的隐私属性标签为 $\{P_i\}_{i=n}^{s}$.

5) 攻击者背景知识. 记攻击者收集的隐私样本标识号为 0 至 $n-1$ 的特征集合为 $\{P_i\}_{i=0}^{n-1}$, 设定攻

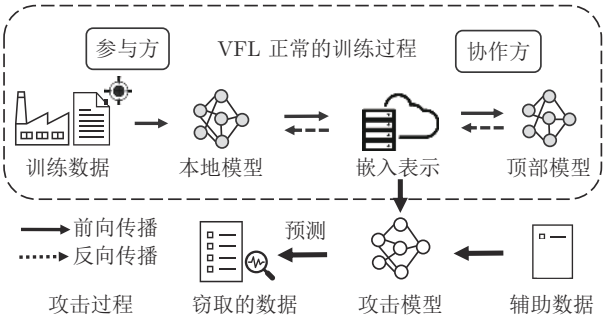


图 3　VFL 场景中攻击示意图

Fig. 3　Illustration of attack in VFL

击者按照 VFL 协议收集到样本标识号为 0 至 $n-1$ 的嵌入表示为 $\{E_i\}_{i=0}^{n-1}$.

6) 攻击原理. 本文提出的通用属性推断攻击 F_{ap} 定义为: 设隐私属性 P 的类别数为 m, 攻击者通过参与方的样本嵌入表示 E_i 推断样本的隐私属性 P, 即:

$$F_{ap}: E_i \to P \tag{3}$$

式中, $P \subseteq \{1, \cdots, m\}$.

上述属性推断攻击 F_{ap} 的原理为攻击模型正确建立隐私属性 P 和样本嵌入表示 E_i 间的映射关系 $f(\cdot)$, 即:

$$f(E_i) = P \tag{4}$$

为了正确建立上述映射关系 $f(\cdot)$, 保证隐私属性的每个类 P_i 都具有对应的样本嵌入表示 E_i, 即:

$$f(E_1, \cdots, E_m) = \{P_1, \cdots, P_m\} \tag{5}$$

式中, $E_1, \cdots, E_m$ 表示隐私属性类别 $P_1, \cdots, P_m$ 所需的辅助样本.

由式 (5) 可知, 正确建立映射关系 $f(\cdot)$ 所需的辅助样本数量与隐私属性类别数量成正比. 本文中, 攻击者采用由神经网络搭建的攻击模型, 建立样本嵌入表示和样本隐私属性的映射关系, 其中攻击模型的训练过程为减小辅助样本嵌入表示和隐私属性间的交叉熵. 当攻击者完成神经网络的训练, 攻击者可以利用完成的该攻击模型, 对样本嵌入表示预测该样本的隐私属性.

2.2.2　攻击方法描述

如图 4 所示, 攻击分为 3 个阶段. 第 1 阶段, 协作方在攻击前收集良性参与方的目标隐私属性作为辅助样本; 第 2 阶段, 攻击者利用辅助样本和 VFL 协议交换的嵌入表示训练攻击模型; 第 3 阶段, 攻击者测试攻击模型恢复参与方的目标隐私属性信息. 下面具体介绍以上 3 个阶段.

1) 局部样本属性采集阶段是为了发动通用的

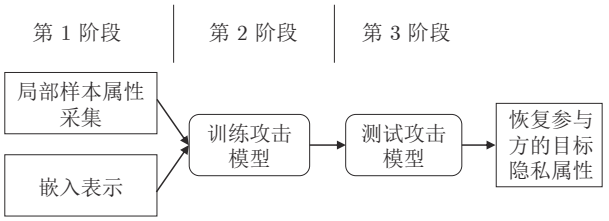


图 4　VFL 中协作方的攻击流程

Fig. 4　Attack pipeline of collaborator in VFL

属性推断攻击, 攻击者需要采集部分样本的属性特征作为背景知识. 背景知识包括参与方传输的嵌入 $\{E_i\}_{i=0}^{s}$ 和部分样本的隐私属性 $\{P_i\}_{i=0}^{n-1}$. 协作方作为攻击者在 VFL 的协议下, 可直接获取原始嵌入表示信息, 此外协作方需要在现实场景中收集部分样本的隐私属性信息. 由于在实际场景中, 属性推断攻击中的隐私属性 (如性别、种族) 具有较少的类别数量, 攻击者需要收集的样本数量也较少, 所以这种攻击在现实场景中具有实施的可能性.

2) 在样本采集阶段, 当攻击者具有背景知识后, 将训练攻击模型. 在该阶段中, 攻击者试图训练一个攻击模型 D, 该攻击模型通常由全连接神经网络构成. 训练过程中损失函数表示为:

$$L_{\text{adversary}} = d(D(E_i), P) \tag{6}$$

式中, $d(\cdot)$ 为损失函数, $E_i = f_{\theta_i}(X_i)$. 由于攻击者推断的隐私属性为类别属性, 损失函数通常设置为交叉熵损失函数.

3) 攻击模型推理阶段, 当攻击模型完成训练时, 攻击者将参与方未知属性的嵌入表示向量 $\{P_i\}_{i=n}^{s}$ 输入攻击模型, 攻击模型输出隐私属性的预测结果.

2.2.3　具体模型设置

本文采用 3 个隐藏层大小分别为 300、200、100 的 5 层全连接神经网络作为攻击模型, 其中输入层神经元个数等同于嵌入表示的维度 d, 输出层大小等同于隐私属性的类别数, 并经由 Softmax 层获得最终模型预测概率向量; 隐藏层的激活函数为 ReLU. 在训练攻击模型过程中, 利用 Adam 优化器, 以 0.005 的学习率最小化隐私属性的真实标签和攻击模型预测概率间的交叉熵损失函数.

2.3　基于最大–最小策略的隐私保护方法

为了保护 VFL 中参与方的隐私属性信息, 本文提出基于最大–最小策略的 VFL 隐私保护方法. 图 5 为 PPVFL 的流程示意图. 参与方前向传播流程和 VFL 方法隐私保护方法如算法 1 和算法 2 所示.

算法 1. 参与方 i 前向传播算法

初始化. 参与方 i 本地模型 f 的参数 θ_i, 本地数据特征

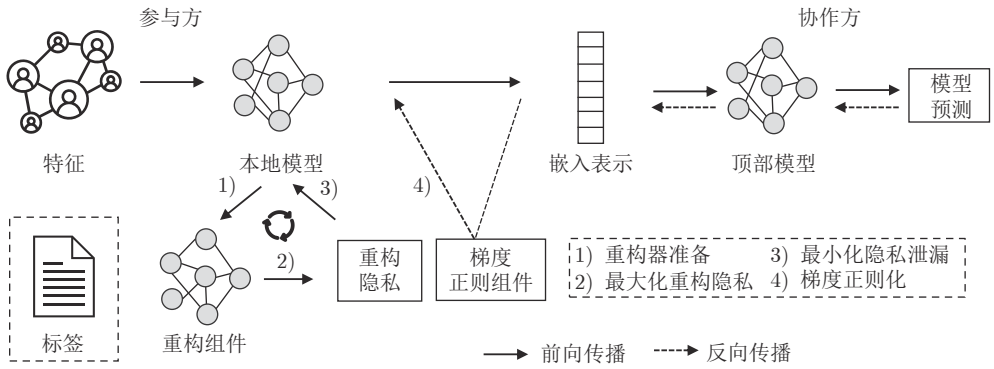


图 5 PPVFL 流程示意图

Fig. 5 Illustration of PPVFL's pipeline

X_i, 学习率 η, 隐私属性标签 P, 训练轮数 n.

1) for $t = 1,\ n$ do:

2) $E_i \leftarrow f_{\theta_i}(X_i)$;

3) $P_i' \leftarrow D_{\theta_D}(E_i)$;

4) 按照式 (15), 计算损失函数 L_{adv};

5) $g_i^D \leftarrow \frac{\partial L_{\text{adv}}}{\partial \theta_D}$;

6) $g_i^f \leftarrow \frac{\partial L_{\text{adv}}}{\partial \theta_i}$;

7) $\theta_D \leftarrow \theta_D - \eta \cdot g_i^D$;

8) $\theta_i \leftarrow \theta_i + \eta \cdot g_i^f$;

9) end for.

算法 2. VFL 隐私保护方法

初始化. 顶部模型参数 θ_{a+1}, 每个参与方本地模型的参数 $\theta_1, \cdots, \theta_M$, 学习率 η, 主任务标签 Y, 训练轮数 s;

1) for $t = 1,\ s$ do:

2) for $t = 1,\ M$ do:

3) $E_i \leftarrow \text{localforward}(\theta_i,\ X_i)$ /*本地模型前向传播*/;

4) end for;

5) $E_{\text{aggr}} = \text{concat}(E_1, \cdots, E_M, E_a)$ /*协作者聚合嵌入表示*/;

6) $Y \leftarrow f_{\theta_{a+1}}(Z_{\text{aggr}})$ /*协作者前向传播*/;

7) 按照式 (17), 计算损失函数 L_{utility};

8) $\theta_{a+1} \leftarrow \theta_{a+1} - \eta \cdot \frac{\partial L_{\text{utility}}}{\partial \theta_{a+1}}$;

9) for $i = 1,\ \cdots,\ M,\ a$ 参与方 do:

10) $g_i \leftarrow \frac{\partial L_{\text{utility}}}{\partial E_i}$;

11) if $t = 1$ do:

12) $\lambda = 1$;

13) $E_{\text{his}} \leftarrow E_{\text{aggr}}$;

14) else:

15) $\lambda = \| E_{\text{aggr}} - E_{\text{his}} \|_2$;

16) $E_{\text{his}} \leftarrow E_{\text{aggr}}$;

17) end if;

18) $g_i \leftarrow \lambda \cdot g_i \cdot \frac{\partial E_i}{\partial \theta_i}$;

19) $\theta_i \leftarrow \theta_i - \eta \cdot g_i$;

20) end for;

21) end for.

2.3.1 防御场景定义

1) VFL 场景. 假设 VFL 中存在 1 个主参与方、m 个从参与方和 1 个协作方.

2) 防御者. VFL 中的主参与方或从参与方.

3) 防御者目标. 目标 1 防止攻击者在 VFL 训练和推理阶段中推断隐私属性信息; 目标 2 保证 VFL 主任务的预测性能.

4) 防御原理. 通过掩藏参与方上传的嵌入表示和隐私属性间的映射关系, 实现对参与方的隐私属性保护. 首先, 防御者在本地维护一个重构组件, 评估当前嵌入表示的泄漏风险; 然后, 利用最小化隐私属性策略破坏嵌入与隐私属性间的映射关系; 最后, 引入梯度正则组件保证主任务预测性能.

2.3.2 防御方法描述

首先, 从信息论角度表示上述 2 个目标, 目标 1 表示为:

$$\min I(D(f_{\text{local}}(X)); P) \tag{7}$$

式中, $D(\cdot)$ 表示重构组件模型, $f_{\text{local}}(\cdot)$ 表示本地模型, P 表示隐私属性信息, $I(x; y)$ 表示变量 x 和变量 y 间的互信息.

目标 1 是最小化重构组件预测结果和隐私属性间的互信息, 即破坏本地模型输出的嵌入表示和隐私属性间的映射关系. 上述表达式等价于:

$$\min H(P) - H(P|D(E_i)) \tag{8}$$

式中, E_i 表示参与方 i 上传给协作方的嵌入表示, $H(x|y)$ 表示变量 x 和 y 的联合交叉熵. 目标 2 保证 VFL 中主任务预测性能, 表示为:

$$\max H(Y|g(E_{\text{aggr}})) \tag{9}$$

式中, 聚合完成后的嵌入表示 E_{aggr} 为:

$$E_{\text{aggr}} = \text{concat}(f_{\theta_1}(X_1), \cdots, f_{\theta_m}(X_m), f_{\theta_a}(X_a)) \tag{10}$$

式中, $\text{concat}(x, y)$ 表示 2 个变量按照相同维度拼接.

结合上述 2 个目标, 联合目标可表示为:

$$\min_{f,g}\max_{D} H(P) - H(P|D(E_i)) + H(Y|E_{\text{aggr}}) \tag{11}$$

由于 $H(P)$ 恒定不变, 式 (11) 可转化为:

$$\min_{f,g}\max_{D} C - H(P|D(E_i)) + H(Y|E_{\text{aggr}}) \tag{12}$$

式中, C 为常数.

图 6 为防御方法示意图. 图 6 左侧区域表示参与方的嵌入包含敏感属性信息, 右侧区域表示参与方的嵌入不包含敏感属性信息, 图 6 中向下的箭头表示主任务优化的方向. 为达到目标 1, 本地模型参数在第 t 次更新的方向 $\bigtriangledown\theta_f^t$ 向隐私保护方向移动.

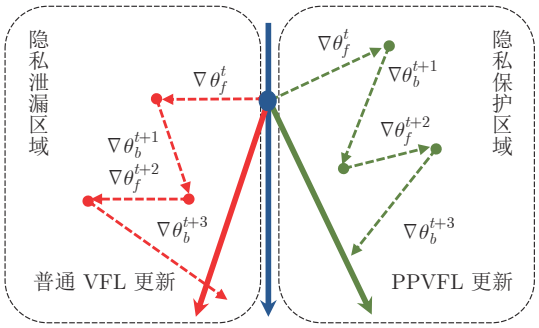


图 6　防御方法示意图

Fig. 6　Illustration of defense method

本文采用 Yaroslav 等[26] 提出的梯度反转方法解决上述优化问题, 即将计算梯度方向自动取反, 在前向传播过程中实现恒等变换:

$$\theta_i = \theta_i - \eta \cdot \lambda \cdot \frac{\partial L_{\text{adv}}}{\partial \theta_i} \tag{13}$$

式中, λ 控制梯度的更新方向和强度, 通常取为 -1.

$$\theta_i = \theta_i + \eta \cdot \frac{\partial L_{\text{adv}}}{\partial \theta_i} \tag{14}$$

由于梯度反向传播后, 本地模型更新方向朝向重构组件损失增加的方向, 重构组件损失增加将会导致攻击者推理出的隐私属性准确率下降, 达到隐私保护效果.

如算法 1 所示, 首先, 重构组件根据式 (15) 计算损失函数, 进行梯度反向传播. 在本地模型和重构组件参数更新期间, 重构组件参数按算法 1 的步骤 7) 沿梯度下降方向更新步长 η. 最后, 本地模型参数按照算法 1 的步骤 8) 沿梯度上升的反向更新步长 η, 迭代 n 次. 其中本地模型迭代次数 n 为本地模型达到收敛状态时需要的最小迭代次数. 当本地模型在训练迭代过程中, 相邻 2 次训练迭代过程中的损失差值小于 0.000 1, 则本地模型达到收敛状态.

重构组件在模拟攻击者推断隐私属性采用的损失函数表示为:

$$L_{\text{adv}} = -\sum_{i=1}^{M}\sum_{j=1}^{N}[P_{i,j}\cdot\ln(p'_{i,j}) + (1-p'_{i,j})\cdot\ln(1-p'_{i,j})] \tag{15}$$

式中, $p'_{i,j} = \arg\max D_{\theta_D}(E_i)$, 其中 θ_D 为重构组件的模型参数.

在 VFL 反向传播阶段, 为了实现目标 2, PPVFL 在随机梯度下降的基础上引入了梯度正则组件, 使得本地模型参数更新在第 $t+1$ 次时的更新方向 $\bigtriangledown\theta_b^{t+1}$ 朝向主任务优化的方向移动. 如算法 2 所示, 步骤 10) 利用反向传播算法计算参与方训练产生的梯度 g_i. 步骤 18) 通过对梯度乘正则系数 λ 更新梯度信息, 步骤 19) 完成本地模型参数的更新.

正则系数 λ 表示当前嵌入表示 E_{aggr} 和 VFL 前一轮次嵌入表示 E_{his} 的二范数值:

$$\lambda = \parallel E_{\text{aggr}} - E_{\text{his}} \parallel_2 \tag{16}$$

此外, VFL 中的主任务预测损失函数表示为:

$$L_{\text{utility}} = -\sum_{i=1}^{M}\sum_{j=1}^{N}[y_{i,j}\cdot\ln(y'_{i,j}) + (1-y_{i,j})\cdot\ln(1-y'_{i,j})] \tag{17}$$

式中, $y'_{i,j} = \arg\max f_{\theta_{\text{top}}}(E_{\text{aggr}})$.

2.3.3　具体模型设置

本文采用 4 层全连接层神经网络作为防御者的重构组件, 其中输入层大小为参与方上传给协作方嵌入表示的维度 d, 隐藏层的大小分别为 200 和 100, 输出层大小为隐私属性的类别数目, 并经由 Softmax 层获得最终模型预测概率向量; 隐藏层的激活函数为 ReLU. 在训练攻击模型过程中, 利用 Adam 优化器以 0.005 的学习率最小化隐私属性真实标签和攻击模型预测概率间的交叉熵损失函数.

2.3.4　时间复杂度分析

PPVFL 相较于没有任何防御方法的 VFL, 增加了本地模型基于最大–最小策略训练的时间代价. 这些时间代价具体产生于以下 3 个步骤:

步骤 1. 本地模型和重构组件完成前向传播, 其中引入的时间复杂度为 O(1).

步骤 2. 重构组件根据式 (15) 计算重构损失,

其中引入的时间复杂度为 O(1).

步骤 3. 重构组件和本地模型进行反向传播, 更新本地模型, 其中引入的时间复杂度为 O(1).

以上 3 个步骤依次迭代 1 次完成本地模型隐私保护训练, 时间复杂度为 O(n). 在没有任何防御方法的 VFL 中, 时间代价为本地模型完成 1 次前向传播的时间, 时间复杂度为 O(1). 由于完成正常 VFL 迭代次数为 s, 因此 PPVFL 完成模型训练的时间复杂度为 O($n \cdot s$), VFL 的训练时间复杂度为 O(s).

3 实验结果与分析

本节首先介绍实验设置, 然后从不同攻击背景知识和不同训练轮次的嵌入表示角度, 评估本文提出的属性推断攻击的有效性, 最后从 4 个方面验证、分析 PPVFL 防御的有效性, 包括隐私保护与效用的权衡、敏感性分析、通用性分析和可视化解释防御的有效性.

3.1 实验环境

实验环境基于 Python3.6 开发, 中央处理器为 I7-7700, 图形处理器为 TITAN XP, 内存为 16 GB × 4 内存 (DDR4), 操作系统为 Ubuntu16.04 (OS).

3.2 实验设置

3.2.1 数据集描述

采用在 Adults、Rochester[27] 和 Yale[27] 3 个 VFL 研究中常用的数据集评估攻击和防御性能, 这 3 个数据集包含结构化数据和图数据 2 种不同数据模态. 表 2 总结了 3 个数据集的基本统计信息.

表 2 VFL 数据集的基本统计信息

Table 2 The basic statistics of VFL datasets

数据集	样本数	连边关系	标签类别	属性特征	隐私属性
Adults	48 842	—	2	14	婚姻
Rochester	4 563	167 653	6	236	教育
Yale	8 578	405 450	6	188	种族

1) Adults 数据集是一个根据人口普查数据预测收入是否超过 5 万美元/年的二分类数据集. 该数据集包含 48 842 位统计人口的普查数据, 其中涵盖 14 个属性特征 (如年龄、受教育程度和职业等). 此数据集中攻击者推断的隐私属性为“婚姻”, 其具有 2 个标签类别.

2) Rochester 数据集是一个社交网络数据集. 该数据集包含 4 563 位用户的社交统计信息, 其中涵盖 167 653 条连边关系和 236 个属性特征 (如性别、年龄和婚姻等). 此数据集中攻击者推断的隐私属性为“教育”, 其具有 6 个标签类别.

3) Yale 数据集也是一个社交网络数据集. 该数据集包含 8 578 位用户的社交统计信息, 其中涵盖 405 450 条连边关系和 188 个属性特征 (如性别、年龄和婚姻等). 此数据集中攻击者推断的隐私属性为“种族”, 其具有 6 个标签类别.

3.2.2 数据集分割方式

1) Adults 数据集依据参与方的数量将特征进行均匀分割, 每个参与方获得相同数量的属性特征, 其中主任务的标签分配给主参与方.

2) Rochester 和 Yale 数据集依据参与方的数量将节点特征进行均匀分割, 每个参与方获得相同数量的特征, 并具有相同邻接矩阵, 其中主任务的标签分配给主参与方.

数据集按照 4:1 比例划分训练集 (已知样本的主任务标签) 和测试集 (预测样本的主任务标签). 实验结果默认为 10 次重复随机实验结果的均值.

3.2.3 评价指标

VFL 中主任务预测性能的评价指标采用主任务预测准确率 (main_acc). 由于本文隐私属性为分布平衡情况, 所以本文采用常见的文献 [27−30] 的推断准确度 (attack_acc) 来评估隐私泄漏风险. 推断准确度越低, 表示隐私保护效果越好:

$$\text{attack\_acc} = \frac{\text{ITP} + \text{ITN}}{\text{IP} + \text{IN}} \tag{18}$$

式中, ITP 表示实际为目标正样本被推断为正的样本数, ITN 表示实际为目标负样本被推断为负的样本数, IP 为目标正样本数, IN 为目标负样本数.

实验主要根据攻击模型的推断准确度来衡量 VFL 的隐私泄漏风险, 根据权衡值 T 衡量防御方法在维持主任务预测性能和隐私保护性能上的有效性:

$$T = \frac{\text{main\_acc}}{\text{attack\_acc}} \tag{19}$$

3.2.4 模型

对不同的数据集, VFL 采取不同的本地模型和顶端模型. 表 3 列举了 VFL 中本地模型和顶部模型的结构.

对 Adults 数据集, 由于该结构化数据训练简单, 本地模型采用全连接神经网络进行提取特征信息; 对 Rochester 和 Yale 图数据集, 本地模型采用图卷积神经网络 (Graph convolutional network, GCN)[31] 和简化图卷积神经网络 (Simplifying graph convolutional networks, SGC)[32] 提取图数据的

表 3　模型结构
Table 3　Model architectures

数据集	本地模型	顶部模型
Adults	FCNN-1	FCNN-2
Rochester	GCN-2	FCNN-2
Yale	SGC-2	FCNN-2

特征信息. 表 3 中 FCNN-1 表示 1 层全连接神经网络层, FCNN-2 表示 2 层全连接神经网络层, GCN-2 表示 2 层图卷积神经网络层, SGC-2 表示 2 层简化的图卷积神经网络层.

3.3　对比防御方法

实验采用随机噪声、随机丢弃、降维和差分隐私 4 种对 VFL 的先进防御方法作为对比算法.

1) 随机噪声 (Noisy). 在嵌入表示中添加高斯噪声是一种常见的隐私保护方法[22]. 实验中高斯噪声的平均值设置为 1, 标准方差 σ 设置为 3 组, 分别取值 1、5、10.

2) 随机丢弃 (Dropout). 过拟合通常使模型在训练过程中潜在地记住原始训练数据信息. Luo 等[22]将本地模型采用随机丢弃作为防御方法. 实验设置 3 组随机丢弃率 ξ 分别为 0.2、0.5、0.8.

3) 降维 (Dimensionality reduction, DR). 由于本地模型的输出维度能影响嵌入表示的稀疏性, 随着本地模型输出的嵌入表示稀疏性降低, 与主任务无关的隐私信息也会降低. Luca 等[10] 将降低输出维度作为 FL 中保护数据隐私的方法. 实验设置 4 组降维值 d 分别为 32、16、8、4.

4) 差分隐私 (Differential privacy, DP) 技术是一种常见的隐私保护技术, 应用差分隐私技术对原始数据进行扰动来保护数据隐私. 实验设置 3 组不同隐私预算的超参数 σ 分别为 0.1、0.2、0.5.

3.4　攻击实验

为了全面评估本文提出的通用属性推断攻击方法的有效性, 本节回答下面 2 个问题:

1) 攻击者拥有不同比例背景知识, 对推断准确度的影响如何?

2) 攻击者利用 VFL 中不同训练轮次产生的嵌入表示训练攻击模型, 对推断隐私准确度的影响如何?

3.4.1　不同比例背景知识

为评估攻击者拥有不同比例背景知识对攻击推断准确度影响, 本节分别对 VFL 训练数据中的隐私属性和测试数据中的隐私属性进行推测, 推断准确度结果如图 7 所示.

由图 7 可知, 在 VFL 两种不同模态的数据集上, 即使攻击者具有参与方少量的属性作为背景知识, 攻击者依然获得较高的推断准确度. 如在 Adults 数据集中, 攻击者仅具有参与方 0.03% 的训练样本隐私属性, 推断准确度可以达到 96%. 实验结果表明, 属性推断攻击对 VFL 中参与方的隐私属性具有严重威胁.

此外, 由图 7 可以发现, 属性推断攻击对训练数据和测试数据具有相近的推断准确度. 这是因为训练数据和测试数据本身具有相似的特征分布, 这使得训练数据和测试数据的嵌入表示具有相似性, 因此攻击者推断训练数据和测试数据的能力相似.

3.4.2　不同训练轮次的嵌入表示

攻击者需要对 VFL 参与方上传的嵌入表示训练攻击模型, 从而窃取隐私属性. 由于参与方在训练过程中每轮存在上传的嵌入表示, 因此本节评估

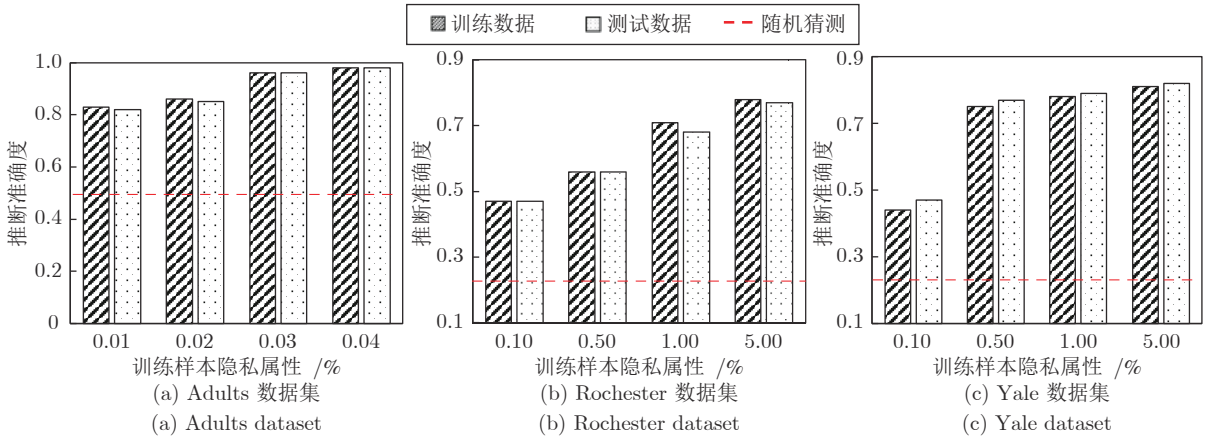


图 7　不同比例背景知识下属性推断攻击的性能
Fig. 7　Performance of property inference attack with different proportions of background knowledge

VFL 不同训练轮次后参与方上传的嵌入表示对攻击者发动属性推断攻击性能的影响, 结果如图 8 所示. 同时, 针对测试数据的隐私属性泄漏, 本节评估了 VFL 进行不同轮次训练后, 在推理阶段产生嵌入表示的隐私泄漏情况. 在图 8 中, 左侧坐标轴表示攻击者对训练数据和测试数据的隐私属性推断准确度, 右侧坐标轴表示 VFL 的主任务预测准确率.

由图 8 可知, 随着 VFL 训练轮次的增加, 主任务预测准确率逐渐上升, 攻击者利用 VFL 训练过程中任意轮次中的嵌入表示都能取得较高的推断准确度. 此外, 实验表明, 在 Adults 数据集上, 利用 VFL 第 1 轮和最后一轮传输的嵌入表示训练的攻击模型具有更高的推断准确度. 通过分析训练过程, 造成这种现象的原因是以下 2 点: 1) 在训练初期, 本地模型对主任务提取特征信息能力弱, 提取的特征信息中更多包含原始数据信息; 2) 在训练后期, 本地模型训练存在过拟合现象, 本地模型融合与主任务无关的隐私数据.

3.5 防御实验

为了全面评估本文提出的 PPVFL 隐私保护和维持主任务的性能, 本节回答下面 4 个问题:

1) PPVFL 的权衡隐私保护和主任务预测性能如何?

2) PPVFL 的参数敏感性 (即参与方数量和防御者重构组件的结构对 PPVFL 的影响) 如何?

3) PPVFL 的通用性如何? 即当攻击者使用不同的攻击模型时, PPVFL 是否同样有效? PPVFL 是否可以应用在实际工业互联网场景中?

4) 如何解释 PPVFL 的隐私保护效果?

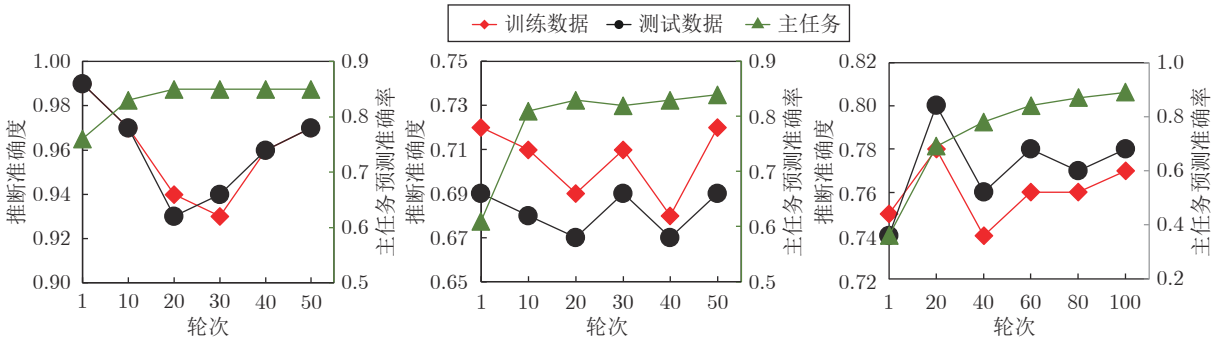
3.5.1 隐私保护与效用的权衡

为了展现 PPVFL 隐私保护和维持主任务预测性能, 本节将 PPVFL 与第 3.3 节介绍的 4 种隐私保护方法作为基准方法进行对比. 对比方法存在超参数, 第 3.3 节已给出不同组超参数进行多组对比实验; PPVFL 的超参数 λ 取值为 0.1、0.5、1.

此外, 本节展示 VFL 框架在没有任何防御方法时, 面临属性推断攻击的隐私泄漏情况. 值得注意的是, 本文假设攻击者具有的背景知识和训练轮次都在第 3.4 节中攻击者推断准确度最高设置, 以此充分评估防御方法的有效性. 分别在结构化数据集 Adults 和 2 个网络数据集 Rochester、Yale 上进行实验验证.

图 9 和图 10 展示了 PPVFL 与基准方法在权衡隐私保护任务和维持主任务预测性能上的效果对比, 其中无防御表示没有任何防御的 VFL 模型. 图中横坐标轴表示主任务预测准确率, 值越高表示主任务预测性能越好; 图中纵坐标轴表示隐私推断准确度, 值越低表示泄漏隐私的风险越低. 最佳防御方法是维持推断准确度最低且主任务预测性能最高, 即结果处于图中的右上角区域. 如图 9 所示, 对训练数据的保护中, PPVFL 数据点位于图的右上角区域. 这体现了该方法在保护数据隐私和维持主任务预测性能上取得最佳效果. 如针对 PPVFL 防御场景下的攻击, 攻击者在 Adults 数据集上的推断准确度降低到 17%, 相比没有任何防御方法的隐私属性推断准确度下降了约 63%. 攻击者在 PPVFL 防御场景下的推断准确度接近随机猜测的推断准确度.

如图 10 所示, 本文防御方法对 VFL 的测试数据也具有很好的保护效果. 这是因为在训练阶段, VFL 中的本地模型提取数据特征时, 具有滤除隐私信息的能力. 因此, 在测试阶段, 本地模型能自动滤除测试样本中的隐私属性信息.



(a) Adults 数据集 (b) Rochester 数据集 (c) Yale 数据集

(a) Adults dataset (b) Rochester dataset (c) Yale dataset

图 8 不同训练轮次后属性推断攻击的性能

Fig. 8 Performance of property inference attack with different training round

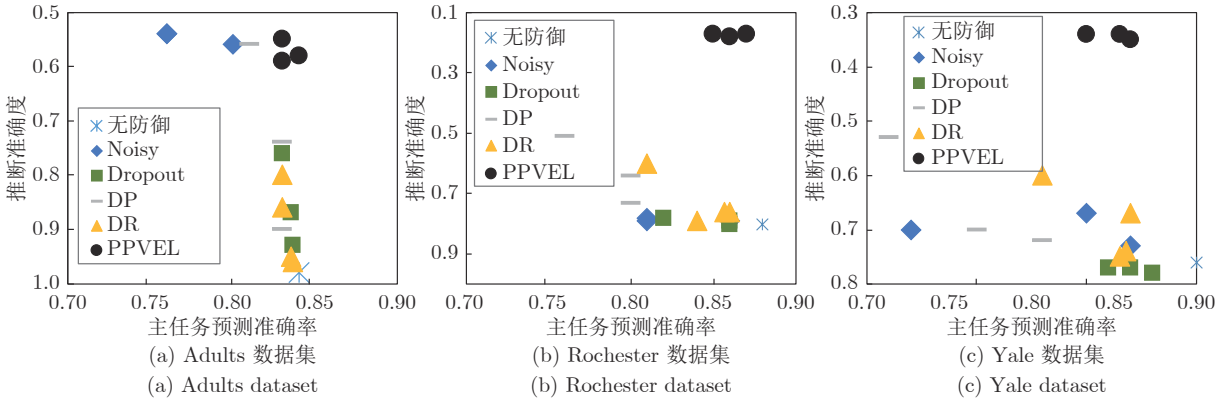


图 9　PPVFL 对训练数据的隐私保护性能

Fig. 9　Performance of PPVFL′s privacy preservation for training data

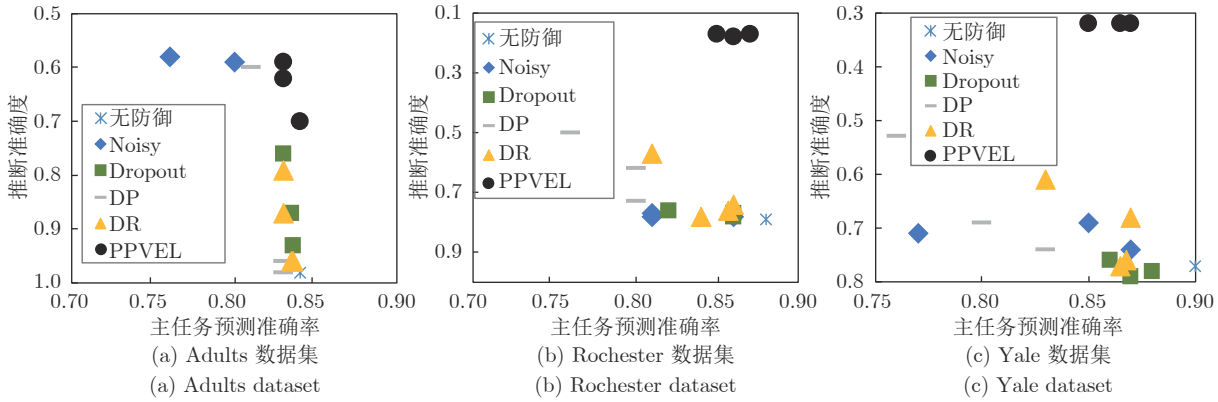


图 10　PPVFL 对测试数据隐私保护性能

Fig. 10　Performance of PPVFL′s privacy preservation for testing data

3.5.2　敏感性分析: 参与方数量

在本文实验设置中, VFL 被设定存在 2 个参与方. 然而在实际场景中, VFL 中可能存在多个参与方联合训练的场景. 为此, 本节评估 VFL 中存在多个参与方时, PPVFL 隐私保护方法的有效性. 本文选择具有特征数量最多的 Rochester 数据集进行实验验证. 在实验设置中, Rochester 数据集的特征按照参与方数量进行平均划分. 实验结果如图 11 所示.

由图 11 可知, PPVFL 在多个参与方场景下, 训练数据和测试数据的隐私属性推断准确度均降低到 30% 以下, 远低于无防御 VFL 的推断准确度. PPVFL 在多个参与方场景下防御有效的原因是本文将防御部署在参与方的本地模型, 防御方法破坏了嵌入表示和隐私属性之间的映射关系. 由于每个参与方模型各自具有提取不包含隐私的信息, 所以增加 VFL 中的参与方数量, 不会显著影响防御性能. 此外, 在实验中, 参与方数量的不同使得防御效果存在差异, 这可能是在划分数据集特征过程中引入了随机性而导致的.

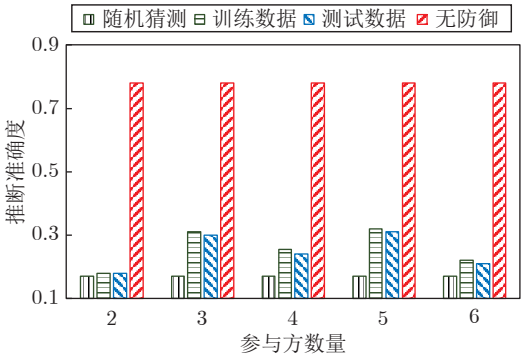


图 11　PPVFL 在多个参与方场景下隐私保护的性能

Fig. 11　PPVFL′s privacy preservation performance in multiple parties

3.5.3　敏感性分析: 重构组件结构

PPVFL 中引入了基于神经网络构建的重构组件, 其目的为模拟攻击者推断隐私属性信息. 本文从 2 个方面评估 PPVFL 的敏感性: 一是重构组件

的隐藏层层数对防御的性能影响; 二是重构组件的隐藏层神经元数量对防御的性能影响. 在实验设置中, 攻击者的攻击模型采用 3 层全连接神经网络推断隐私属性, 其中间层神经元数量为 60 000 个. 在 Rochester 数据集上的实验结果如图 12 所示.

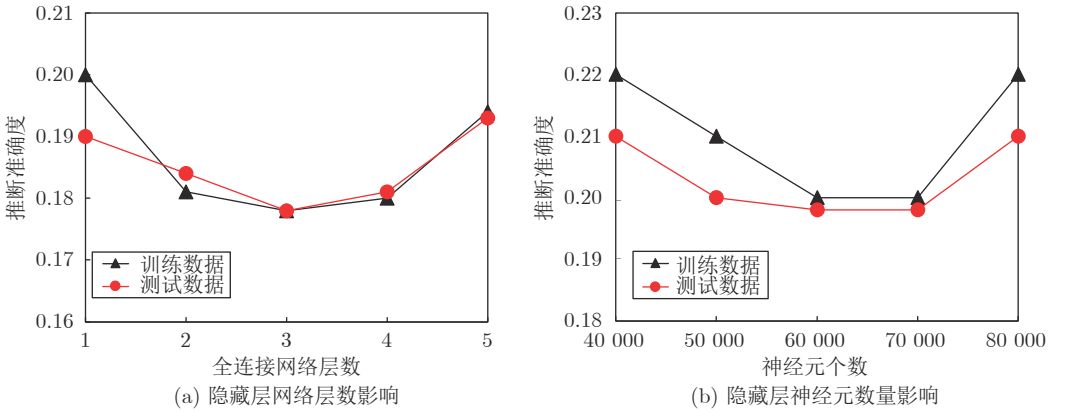
首先, 由图 12(a) 的防御效果可知, 即使防御者使用的重构组件与攻击者的攻击模型网络层数不同, 防御者可以同样达到很好的防御效果. 当防御者重构组件隐藏层仅为 1 层全连接神经网络时, 训练数据和测试数据的推断准确度都降低到 20% 左右, 这说明攻击处于随机猜测的水平.

此外, 当防御者具有的重构组件隐藏层层数和攻击者具有的攻击模型隐藏层层数相同时, 防御属性推断攻击的效果最佳. 这符合本文的猜想, 因为防御者重构组件的目的是模拟攻击者的攻击模型进行推断隐私, 当重构组件接近真实攻击模型时, 具有最佳隐私保护效果.

图 12(b) 也显示出了类似的结论. 当防御者的重构组件与攻击者的重构组件具有相同的神经元数量时, 防御属性推断攻击的效果最佳; 当两者的差距增大时, 防御效果也会随之减弱. 这对防御效果的影响有限, 推断准确度始终保持在 23% 以下, 仍具有良好的隐私保护效果.

3.5.4 通用性分析: 不同攻击模型

在第 3.5.3 节中讨论了攻击者采用全连接神经网络作为攻击模型进行属性推断攻击的场景. 在现实场景中, 攻击者可以采用不同类型的分类器发动属性推断攻击. 假设攻击者使用常见增强学习分类 (AdaBoost)、支持向量机 (Support vector machine, SVM)[33] 和梯度提升分类 (Gradient boosting, GB) 三种分类器来推断隐私属性, 图 13 展现了 PPVFL 防御方法对攻击者利用不同攻击模型进行属性推断

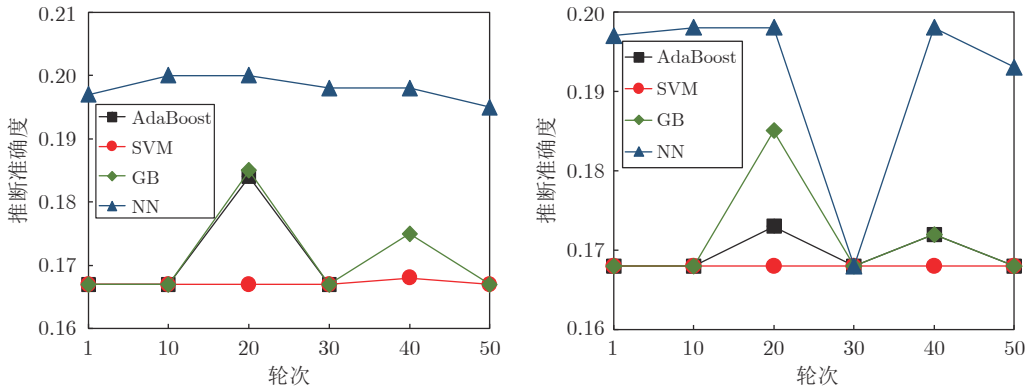


(a) 隐藏层网络层数影响 (a) The effect of the number of hidden layers

(b) 隐藏层神经元数量影响 (b) The effect of the number of hidden layer neurons

图 12 PPVFL 隐私解码器对防御性能的影响

Fig. 12 The effect of PPVFL′s privacy decoder on defense performances



(a) 训练数据隐私泄漏情况 (a) Training data privacy leakage

(b) 测试数据隐私泄漏情况 (b) Testing data privacy leakage

图 13 PPVFL 在不同攻击模型下的隐私保护性能

Fig. 13 Performance of PPVFL′s privacy preservation against different attack models

的防御效果, 可知 PPVFL 对不同的攻击模型具有很好的防御效果, 这主要得益于 PPVFL 破坏了嵌入表示和隐私属性间的映射关系.

此外, 攻击者利用神经网络作为攻击者的重构组件明显要优于 AdaBoost、SVM 和 GB. 这是因为神经网络相较于其他分类模型, 具有更好的拟合能力, 能够模拟出攻击能力强的攻击者.

3.5.5　通用性分析: 工业互联网应用

为了验证 PPVFL 在实际工业互联网任务中具有的通用性. 本文在实际工业互联网评估钢板缺陷诊断数据集上, 验证了 PPVFL 的有效性. 该数据集包含 1 941 条钢板数据记录信息, 其中有 26 个属性特征. 本节分别将“钢板序列”和“A300”作为隐私属性, 这 2 个属性属于钢板厂商的重要隐私数据. 同时, 本节评估了 PPVFL 对训练数据和测试数据的隐私属性保护性能. 评估结果如表 4 所示.

在无防御情况下, 属性推断攻击的推断准确度都保持较高水平, 表明当前的隐私属性存在严重的隐私泄漏风险. 如针对“钢板序列”属性的推断准确度达到了 95% (见表 4), 表明在实际工业互联网数据集上存在严重隐私泄漏风险, 迫切需要防御方法来保护隐私属性.

针对 PPVFL 的属性推断攻击, 攻击者的推断准确度降低到随机猜测水平, 同时主任务预测还保持在较高水平. 实验结果充分验证了 PPVFL 的通用性, 对主任务预测和隐私保护性能具有最佳的权衡效果, 适合应用在实际工业互联网中.

3.5.6　可视化解释

为了进一步理解 PPVFL 成功抵御属性推断攻击的原因, 利用 t-SNE 可视化技术进行分析. 本节将参与方本地模型输出的嵌入表示利用主成分分析方法进行聚类, 并依据不同的隐私属性类别, 将不同节点添加不同颜色. 图 14 和图 15 分别为在 Adults 和 Rochester 数据集上, 防御前和防御后的 t-SNE 示意图. 图中相同颜色的点表示样本为相同隐私的属性类别. 可以看出, 防御前, 不同隐私属性能明显地划分, 即隐私属性存在严重的泄漏风险; 防御后, 不同颜色的隐私属性混合在一起而无法划分, 达到了保护隐私属性的效果.

4　讨论

为了全面评估 PPVFL 性能, 本节回答下面 3 个问题:

1) 当攻击者已知采用 PPVFL 作为防御方法, 是否仍然可以发动自适应攻击?

2) 如果参与方直接拒绝让隐私属性参与 VFL 训练过程, 是否就不存在隐私属性泄漏问题?

3) PPVFL 如何适用于基于动态系统的 VFL?

4.1　已知采用防御方法

当攻击者已知采用 PPVFL 的防御方法, 攻击者仍然无法采取有效自适应攻击手段. 因为攻击者的背景知识是来自参与方上传的嵌入表示, 这些嵌入表示在参与方前向传播过程中被扰动. 由于嵌入表示本身不具有特定的规则, 因此即使攻击者利用现有的去噪组件[34], 也无法有效推断出隐私属性.

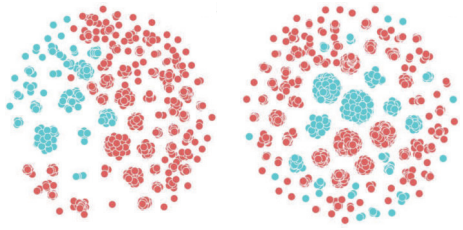
4.2　隐私属性不参与训练

当参与方的隐私属性不参与 VFL 的训练, 攻

表 4　实际工业互联网数据集上的隐私保护效果

Table 4　Privacy protection effect on actual industrial internet dataset

隐私属性	钢板序列					A300				
	训练数据		测试数据			训练数据		测试数据		
	推断准确度	权衡值	推断准确度	权衡值	主任务准确率	推断准确度	权衡值	推断准确度	权衡值	主任务准确率
无防御	0.95	0.82	0.96	0.81	**0.78**	0.74	1.00	0.72	1.03	**0.74**
Noisy ($\sigma = 1.0$)	0.66	**1.00**	0.84	0.79	0.66	0.63	0.95	0.62	0.97	0.60
Noisy ($\sigma = 5.0$)	0.60	0.93	0.55	1.02	0.56	0.60	0.83	**0.59**	0.85	0.50
Dropout ($\eta = 0.5$)	0.91	0.88	0.91	0.88	0.80	0.70	1.03	0.64	1.13	0.72
Dropout ($\eta = 0.8$)	0.86	0.86	0.86	0.86	0.74	0.70	0.96	0.64	1.05	0.67
DP ($\sigma = 0.1$)	0.56	1.21	0.56	1.21	0.68	0.67	1.06	0.65	1.09	0.71
DP ($\sigma = 0.2$)	0.90	0.79	0.89	0.80	0.71	0.68	1.06	0.67	1.07	0.72
DR ($d = 8.0$)	0.87	0.85	0.86	0.86	0.74	0.69	0.80	0.67	0.82	0.55
DR ($d = 4.0$)	0.66	0.97	0.65	0.98	0.64	0.68	0.79	0.64	0.84	0.54
PPVFL ($\lambda = 0.1$)	**0.55**	**1.38**	0.57	1.33	0.76	0.60	**1.20**	0.62	**1.16**	0.72
PPVFL ($\lambda = 0.5$)	**0.55**	1.36	**0.54**	**1.39**	0.75	**0.59**	**1.20**	0.61	**1.16**	0.71



(a) 防御前
(a) Before defense

(b) 防御后
(b) After defense

图 14 Adults 数据集上, 防御前和防御后的 t-SNE 示意图
Fig. 14 t-SNE before and after defense of Adults

(a) 防御前
(a) Before defense

(b) 防御后
(b) After defense

图 15 Rochester 数据集上, 防御前和防御后的 t-SNE 示意图
Fig. 15 t-SNE before and after defense of Rochester

击者仍能够以较高推断准确度推断出未参与 VFL 训练的隐私属性. 这是因为即使参与方的隐私属性不参与训练, 整体的嵌入表示和隐私属性仍存在映射关系. 因此, 拒绝让隐私属性参与训练无法避免隐私泄漏问题.

4.3 动态系统的 VFL

针对 VFL 动态系统的数据量小的特点, 本文提出的 PPVFL 框架在训练过程中可采用数据增强方法对原始数据进行扩增, 提高参与方本地模型的隐私保护能力; 针对 VFL 动态系统的数据实时性强的特点, 可以将 PPVFL 框架中的本地模型设计为循环神经网络, 融合时序控制处理动态系统的实时数据.

5 结束语

VFL 作为一种新兴的保护数据隐私的分布式学习技术, 受到学术界和工业界的广泛关注. 多数研究从参与方的角度分析 VFL 在实际应用场景中潜在的隐私安全问题. 在此背景下, 本文从 VFL 中协作方的角度, 构建一种通用的属性推断攻击方法, 评估了 VFL 面临的隐私泄漏风险. 为了解决 VFL 面临的上述威胁, 本文进一步提出基于最大–最小策略的 VFL 隐私保护方法. 通过对参与方的本地模型使用最大–最小化策略, PPVFL 滤除了参与方嵌入表示的隐私属性信息. 同时 PPVFL 引入梯度正则组件, 保证训练过程主任务预测性能. 本文在 3 个真实数据集上进行了大量实验, 验证了 PPVFL 的有效性. 此外, 本文还验证了 PPVFL 在实际工业互联网场景中的通用性.

然而, 该方法主要适用于边缘模型为神经网络的 VFL 框架, 还不足以有效加固基于随机森林或逻辑回归算法搭建的 VFL 框架的隐私安全. 未来工作将研究更具通用性的隐私保护方法. 此外, 在实际 VFL 场景中, 可能存在数据缺失或由于通信间断导致的数据丢失问题, 设计具有容错机制的 VFL 隐私保护方法也是未来的研究方向之一.

References

1 Luckow A, Cook M, Ashcraft N, Weill E, Djerekarov E, Vorster B. Deep learning in the automotive industry: Applications and tools. In: Proceedings of the IEEE International Conference on Big Data. Washington, USA: IEEE, 2016. 3759–3768

2 Schneider S, Taylor G W, Kremer S C. Deep learning object detection methods for ecological camera trap data. In: Proceedings of the 15th Conference on Computer and Robot Vision. Toronto, Canada: IEEE, 2018. 321–328

3 Sangineto E, Nabi M, Culibrk D, Sebe N. Self-paced deep learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **41**(3): 712–725

4 Scoon C, Ko R K. The data privacy matrix project: Towards a global alignment of data privacy laws. In: Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Tianjin, China: IEEE, 2016. 1998–2005

5 Goddard M. The EU general data protection regulation: European regulation that has a global impact. *International Journal of Market Research*, 2017, **59**(6): 703–705

6 Yang Q, Liu Y, Chen T J, Tong Y X. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 2019, **10**(2): 1–19

7 Zhang Ze-Hui, Fu Yao, Gao Tie-Gang. Research on federated deep neural network model for data privacy protection. *Acta Automatica Sinica*, 2022, **48**(5): 1273–1284
(张泽辉, 富瑶, 高铁杠. 支持数据隐私保护的联邦深度神经网络模型研究. 自动化学报, 2022, **48**(5): 1273–1284)

8 Zhang Ze-Hui, Li Qing-Dan, Fu Yao, He Ning-Xin, Gao Tie-Gang. Adaptive federated deep learning with non-IID data. *Acta Automatica Sinica*, 2023, **49**(12): 2493–2506
(张泽辉, 李庆丹, 富瑶, 何宁昕, 高铁杠. 面向非独立同分布数据的自适应联邦深度学习算法. 自动化学报, 2023, **49**(12): 2493–2506)

9 Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE, 2019. 739–753

10 Luca M, Song C, Cristofaro E D, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE, 2019. 691–706

11 Zhu L, Liu Z, Han S. Deep leakage from gradients. In: Proceed-

ings of the Advances in Neural Information Processing Systems. Vancouver, Canada: 2019. 1–11

12 Zhou Chun-Yi, Chen Da-Wei, Wang Shang, Fu An-Min, Gao Yan-Song. Research and challenge of distributed deep learning privacy and security attack. *Journal of Computer Research and Development*, 2021, **58**(5): 927–943
(周纯毅, 陈大卫, 王尚, 付安民, 高艳松. 分布式深度学习隐私与安全攻击研究进展与挑战. 计算机研究与发展, 2021, **58**(5): 927–943)

13 Fu C, Zhang X, Ji S, Chen J Y, Wu J Z, Guo S Q, et al. Label inference attacks against vertical federated learning. In: Proceedings of the USENIX Security. Boston, USA: 2022. 1–18

14 Ou W, Zeng J H, Guo Z J, Yan W Q, Liu D W, Fuentes S. A homomorphic-encryption-based vertical federated learning scheme for rick management. *Computer Science and Information Systems*, 2020, **17**(3): 819–834

15 Liu W, Cheng J H, Wang X L, Lu X J, Yin J W. Hybrid differential privacy based federated learning for internet of things. *Journal of Systems Architecture*, 2022, **124**: 1–15

16 Mehdi M, Al-Fuqaha A. Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. *IEEE Communications Magazine*, 2018, **56**(2): 94–101

17 Lu Y, Huang X H, Zhang K, Maharjan S, Zhang Y. Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles. *IEEE Transactions on Vehicular Technology*, 2020, **69**(4): 4298–4311

18 Dinh C, Pubudu N, Ming D, Aruna S. Blockchain for 5G and beyond networks: A state of the art survey. *Journal of Network and Computer Applications*, 2020, **166**: 1–45

19 Han Xuan, Yuan Yong, Wang Fei-Yue. Security problems on blockchain: The state of the art and future trends. *Acta Automatica Sinica*, 2019, **45**(1): 206–225
(韩璇, 袁勇, 王飞跃. 区块链安全问题: 研究现状与展望. 自动化学报, 2019, **45**(1): 206–225)

20 Sun H, Wang Z Y, Huang Y J, Ye J D. Privacy-preserving vertical federated logistic regression without trusted third-party coordinator. In: Proceedings of the 6th International Conference on Machine Learning and Soft Computing. Haikou, China: 2022. 132–138

21 Cheng K, Fan T, Jin Y, Liu Y, Chen T J, Papadopoulos D, et al. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 2021, **36**(6): 1–9

22 Luo X, Wu Y, Xiao X, Ooi B C. Feature inference attack on model predictions in vertical federated learning. In: Proceedings of the IEEE 37th International Conference on Data Engineering. Chania, Greece: 2021. 181–192

23 Yang K, Song Z, Zhang Y, Zhou Y F, Sun X H, Wang J X. Model optimization method based on vertical federated learning. In: Proceedings of the IEEE International Symposium on Circuits and Systems. Daegu, South Korea: IEEE, 2021. 1–5

24 Paramod S, Rohit S, Iiia L, Srinivas D, Sanjit A S. A formal foundation for secure remote execution of enclaves. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA: 2017. 2435–2450

25 Florian T, Dan H. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: 2019. 1–19

26 Yaroslav G, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: 2015. 1180–1189

27 Li K, Luo G C, Ye Y, Li W, Ji S H, Cai Z P. Adversarial privacy-preserving graph embedding against inference attack. *IEEE Internet of Things Journal*, 2020, **8**(8): 6904–6915

28 Vasisht D, Boutet A, Shejwalkar V. Quantifying privacy leakage in graph embedding. In: Proceedings of the 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. Darmstadt, Germany: 2020. 76–85

29 Zhang Z, Chen M, Backes M, Shen Y, Zhang Y. Inference attacks against graph neural networks. In: Proceedings of the USENIX Security. Boston, USA: 2022. 1–18

30 Liao P, Zhao H, Xu K, Jaakkola T, Gordon G J, Jegelka S, et al. Information obfuscation of graph neural networks. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: 2021. 6600–6610

31 Thomas N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, USA: 2017. 1–14

32 Wu F, Zhang T Y, Souza A H, Fifty C, Yu T, Weinberger K Q. Simplifying graph convolutional networks. In: Proceedings of the 36th International Conference on Machine Learning. San Francisco, USA: 2019. 6861–6871

33 Wang Jie-Ting, Qian Yu-Hua, Li Fei-Jiang, Liu Guo-Qing. Support vector machine with eliminating the random consistency. *Journal of Computer Research and Development*, 2020, **57**(8): 1581–1593
(王婕婷, 钱宇华, 李飞江, 刘郭庆. 消除随机一致性的支持向量机分类方法. 计算机研究与发展, 2020, **57**(8): 1581–1593)

34 Dou Nuo, Zhao Rui-Zhen, Cen Yi-Gang, Hu Shao-Hai, Zhang Yong-Dong. Noisy image super-resolution reconstruction based on sparse representation. *Journal of Computer Research and Development*, 2015, **52**(4): 943–951
(窦诺, 赵瑞珍, 岑翼刚, 胡绍海, 张勇东. 基于稀疏表示的含噪图像超分辨重建方法. 计算机研究与发展, 2015, **52**(4): 943–951)

李荣昌 浙江工业大学信息工程学院硕士研究生. 主要研究方向为联邦学习, 图神经网络和人工智能安全.
E-mail: lrcgnn@163.com
(**LI Rong-Chang** Master student at the College of Information Engineering, Zhejiang University of Technology. His research interest covers federated learning, graph neural network, and artificial intelligence security.)

刘 涛 浙江工业大学信息工程学院硕士研究生. 主要研究方向为联邦学习, 人工智能安全.
E-mail: leonliu022@163.com
(**LIU Tao** Master student at the College of Information Engineering, Zhejiang University of Technology. His research interest covers federated learning and artificial intelligence security.)

郑海斌 浙江工业大学网络空间安全研究院助理研究员. 分别于 2017 年和 2022 年获得浙江工业大学学士和博士学位. 主要研究方向为深度学习, 人工智能安全和公平性算法. 本文通信作者.
E-mail: haibinzheng320@gmail.com
(**ZHENG Hai-Bin** Associate professor at the Institute of Cyberspace Security, Zhejiang University of Technology. He received his bachelor and Ph.D. degrees from Zhejiang University of Technology in 2017 and 2022, respectively. His research interest covers deep learning, artificial intelligence security, and fairness algorithm. Corresponding author of this paper.)

陈晋音 浙江工业大学信息工程学院教授. 分别于 2004 年和 2009 年获得浙江工业大学学士和博士学位. 主要研究方向为人工智能安全, 图数据挖掘和进化计算.
E-mail: chenjinyin@zjut.edu.cn
(**CHEN Jin-Yin** Professor at the College of Information Engineering, Zhejiang University of Technology. She received her bachelor and Ph.D. degrees from Zhejiang University of Technology in 2004 and 2009, respectively. Her research interest covers artificial intelligence security, graph data mining, and evolutionary computing.)

刘振广 浙江大学网络空间安全学院研究员. 主要研究方向为数据挖掘, 区块链安全.
E-mail: liuzhenguang2008@gmail.com
(**LIU Zhen-Guang** Professor at the School of Cyber Science and Technology, Zhejiang University. His research interest covers data mining and blockchain security.)

纪守领 浙江大学计算机科学与技术学院研究员. 分别于 2013 年获得佐治亚州立大学博士学位, 2015 年获得佐治亚理工学院博士学位. 主要研究方向为数据驱动的安全性和隐私性, 人工智能安全性和大数据分析.
E-mail: sji@zju.edu.cn
(**JI Shou-Ling** Professor at the College of Computer Science and Technology, Zhejiang University. He received his Ph.D. degrees from Georgia Institute of Technology in 2013, and from Georgia State University in 2015, respectively. His research interest covers data-driven security and privacy, artificial intelligence security, and big data analysis.)