计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200220

面向图垂直联邦学习的对抗攻击方法

柏 杨 陈晋音 郑海斌 郑雅羽
浙江工业大学信息工程学院 杭州 310023
(211124030094@zjut.edu.cn)

摘 要 图垂直联邦学习是一种结合图数据和垂直联邦学习的分布式机器学习方法,广泛应用于金融服务、医疗健康和社交网络等领域。该方法在保护隐私的同时,利用数据多样性显著提升模型性能。然而,研究表明图垂直联邦学习容易受到对抗攻击的威胁。现有的针对图神经网络的对抗攻击方法,如梯度最大化攻击、简化梯度攻击等方法在图垂直联邦框架中实施时仍然面临攻击成功率低、隐蔽性差、在防御情况下无法实施等问题。为应对这些挑战,提出了一种面向图垂直联邦的对抗攻击方法(Node and Feature Adversarial Attack,NFAttack),该方法分别设计了节点攻击策略与特征攻击策略,从不同维度实施高效攻击。首先,节点攻击策略基于度中心性指标评估节点的重要性,通过连接一定数量的虚假节点以形成虚假边,从而干扰高中心性节点。其次,特征攻击策略在节点特征中注入由随机噪声与梯度噪声构成的混合噪声,进而扰乱分类结果。最后,在6个数据集和3种图神经网络模型上进行实验,结果表明NFAttack的平均攻击成功率达到80%,比其他算法提高了约30%。此外,即使在多种联邦学习防御机制下,NFAttack仍展现出较强的攻击效果。

关键词:垂直联邦学习;图神经网络;图数据;节点分类;对抗攻击
中图分类号 TP387

Adversarial Attack on Vertical Graph Federated Learning

BAI Yang,CHEN Jinyin,ZHENG Haibin and ZHENG Yayu
College of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China

Abstract Graph vertical federated learning(GVFL) is a distributed machine learning approach that integrates graph data with vertical federated learning,widely applied in fields such as financial services,healthcare,and social networks. This method not only preserves privacy but also leverages data diversity to significantly enhance model performance. However,studies indicate that GVFL is vulnerable to adversarial attacks. Existing adversarial attack methods targeting graph neural networks(GNN),such as Gradient Maximization Attack and Simplified Gradient Attack,still face challenges when applied in the GVFL framework. These challenges include low attack success rates,poor stealth,and inapplicability under defense conditions. To address these issues,this paper proposes a novel adversarial attack method for GVFL,termed Node and Feature Adversarial Attack(NFAttack). NFAttack designs node and feature attack strategies to conduct efficient attacks from multiple dimensions. The node attack strategy evaluates node importance using degree centrality metrics and disrupts high-centrality nodes by connecting a certain number of fake nodes to form adversarial edges. Meanwhile,the feature attack strategy introduces hybrid noise-composed of random noise and gradient noise-into node features,thereby affecting classification results. Experiments conducted on six datasets and three GNN models demonstrate that NFAttack achieves an average attack success rate of 80%,approximately 30% higher than other methods. Furthermore,NFAttack maintains strong attack performance even under various federated learning defense mechanisms.

Keywords Vertical federal learning,Graph neural network,Graph data,Node classification,Adversarial attack

1 引言

随着数据隐私和安全问题日益受到关注,传统的集中式机器学习方法在隐私保护上面临着巨大的挑战。联邦学习[1](Federated Learning,FL)作为一种分布式机器学习方法应运而生,它允许多个参与方在不共享原始数据的情况下共同训练模型,从而有效降低数据隐私泄露的风险。然而,尽管联邦学习在隐私保护方面表现出潜力,它在实际应用中仍然面临诸多安全威胁和挑战。其中,对抗攻击是联邦学习中的一个关键安全问题[2]。对抗攻击是一种通过向模型输入恶意扰动数据,来误导模型输出错误结果的攻击手段。在联邦学习中,特别是在参与方之间交换模型参数的过程中,对抗攻击可能对模型性能产生严重影响。

与此同时,图数据在联邦学习中的应用日益增多,尤其是

基金项目:国家自然科学基金(62072406,62406286);浙江省自然科学基金(LDQ23F020001);浙江省重点研发计划(2022C01018);国家重点研发计划(2018AAA0100801)
This work was supported by the National Natural Science Foundation of China(62072406,62406286),Zhejiang Provincial Natural Science Foundation(LDQ23F020001),Key R&D Projects in Zhejiang Province(2022C01018) and National Key R&D Projects of China(2018AAA0100801).
通信作者:陈晋音(chenjinyin@zjut.edu.cn)

在社交网络、交通网络和生物信息网络等领域，对抗攻击的风险进一步加剧。图数据结构复杂且节点间高度关联，这使得对抗攻击更具有破坏性。攻击者可以通过微小的结构或特征修改，显著影响模型的性能。例如，篡改图中的节点或节点特征来破坏模型的训练过程，从而误导模型的预测结果。

针对上述安全问题，本文提出了一种面向图垂直联邦学习的对抗攻击方法（Node and Feature Adversarial Attack，NFAttack）。具体而言，该方法分为节点攻击策略和特征攻击策略。首先，节点攻击策略利用度中心性指标评估节点重要性，通过与高中心性节点连接一定数量的虚假节点形成虚假边，以实现攻击。其次，特征攻击策略向节点特征添加由随机噪声和梯度噪声构成的混合噪声，进而影响分类结果。最后，NFAttack 在 6 个数据集和 3 种图神经网络模型上验证了攻击的有效性，攻击成功率平均可以达到 80%。

综上所述，本文的主要贡献如下：

1）针对现有对抗攻击方法应用于图联邦环境中攻击成功率低的问题，本文提出了一种面向图垂直联邦学习的对抗攻击方法——NFAttack，以实现对节点分类效果的破坏。

2）NFAttack 攻击方法包括节点攻击策略和特征攻击策略，节点攻击策略修改图数据中的节点连边，特征攻击策略针对节点特征进行扰动。两个策略分别从结构和特征两个不同的角度实施攻击，以全面评估模型在不同维度上的脆弱性。

3）大量实验在 6 个数据集和 3 种图神经网络模型上进行攻击效果验证，结果表明，NFAttack 能够平均达到 80%的攻击成功率，相对于对比算法的攻击效果提高约 30%，并且验证了 NFAttack 在防御后仍然有效。

2 相关工作

在联邦学习的研究领域中，联邦学习因其在数据隐私保护和跨组织协作中的关键作用，已经成为机器学习的重要研究方向，尤其是垂直联邦学习在金融[3]和医疗[4]等数据敏感领域的应用日益广泛。同时，图神经网络以其处理图数据的卓越能力，已在引文网络分析[5]、推荐系统[6]和生物信息学[7]等领域取得显著成果。随着图神经网络的普及，图数据对抗攻击问题也日益突出，通过对图结构或节点特征的微小扰动，攻击者可以明显影响模型的预测性能。尽管已有大量研究探索了联邦学习、图神经网络和图数据对抗攻击，但针对垂直联邦学习中图数据对抗攻击的研究仍然缺乏。

2.1 垂直联邦学习

随着机器学习在金融[3]、教育[8]和医疗[4]等领域的广泛应用，“数据孤岛”[9]问题逐渐显现。为了解决这一问题，Hard 等[10]提出了联邦学习。联邦学习是一种分布式机器学习框架，旨在通过多方数据持有者协同训练，来实现数据隐私保护。联邦学习根据分布方式的不同[11]，可分为垂直联邦学习（Vertical FL，VFL）、水平联邦学习（Horizontal FL，HFL）和联邦迁移学习（Transfer FL，TFL）。其中，VFL 作为联邦学习中的关键形式，适用于各参与方拥有相同用户但不同特征的数据场景。与水平联邦学习相比，VFL 更加注重在特征空间重叠的环境下实现隐私保护的协同训练，例如银行和电商平台合作进行用户风险评估。在这种情况下，各参与方能够通过联合训练模型共享知识，而无需交换原始数据。

VFL 允许不同数据持有者在无需共享原始数据的情况下联合训练模型，主要应用于具有不用特征但共享相同样本 ID 的数据集。VFL 通过安全多方计算和同态加密保护等技术确保参与方在模型训练过程中仅交换加密或部分隐私保护的信息，从而保护数据隐私和安全。

2.2 图神经网络

图神经网络（Graph Neural Network，GNN）[12]是一种专门用于处理图数据的神经网络模型，它广泛应用于文本分类、社交网络分析和推荐系统等领域。与传统的神经网络模型不同，图神经网络能够有效地捕获图数据中的拓扑结构和节点之间的关系，从而在各种应用中取得了巨大成功。在处理节点分类任务时，常用的 GNN 有图卷积网络（Graph Convolutional Network，GCN）、图注意力网络（Graph Attention Network，GAT）和图采样聚合网络（Graph Sample and Aggregate，GraphSAGE）等。

图数据[13]是一种非常灵活和强大的数据结构，用于表示实体之间的关系。在图数据中，实体通常被称为节点，而它们之间的关系被称为边。其中，节点的度指与该节点相连边的数量。图数据可以描述各种复杂的关系网络，如引文网络、社交网络和知识图谱等。

GCN 是处理图结构数据的深度学习模型，由 Kipf 等[14]于 2016 年提出，通过对图的邻接矩阵进行归一化处理并聚合节点特征，有效捕捉图的局部和全局信息。GraphSage 是处理图结构数据的一种高效图神经网络模型，由 Hamilton 等[15]于 2017 年提出，其通过对邻居节点进行采样和特征聚合，来有效捕捉图的局部和全局信息。与传统的 GCN 不同，GraphSage 不使用全量邻居节点，而是通过采样部分邻居进行特征聚合，显著提高了模型在大规模图数据上的处理效率和扩展性。GAT 由 Velickovic 等[16]于 2018 年提出，其通过对邻居节点特征加权求和，来更精准地反映节点的重要性和关系。GAT 通过引入注意力机制，自适应地为每个节点分配不同权重，有效捕捉图的局部和全局信息。GCN，GAT 和 GraphSage 等均在节点分类、链路预测等任务中表现出色。

2.3 对抗攻击方法

图神经网络很容易被一些细微的变化扰动，这类扰动被定义为对抗性扰动。对抗攻击[17]发生在模型的测试阶段，根据模型结构精心设计细微的扰动，并添加在原始数据中，从而使图神经网络模型误判，达到欺骗图神经网络模型的目的。

对抗攻击方法按照攻击目标可以分为无目标攻击和有目标攻击。无目标攻击不针对特定节点或边，而是旨在整体上降低图模型的性能，攻击者可通过随机或系统性地添加或删除图中的边来扰乱模型的全局性能；有目标攻击针对特定节点或边，目标是误导模型对这些节点或边的预测，例如攻击者可以通过修改特定节点的特征或其邻接关系来实现特定的攻击目标。

另外，根据攻击者对目标模型内部信息掌握程度的不同，对抗攻击可分为白盒攻击、黑盒攻击和灰盒攻击。白盒攻击假设攻击者完全了解目标模型的所有信息，包括模型的架构、参数、训练数据和梯度信息。在这种情况下，攻击者可以利用这些信息设计高度针对性的对抗性扰动，以最大化攻击效果。因此，白盒攻击通常能够生成极具破坏性的对抗样本。黑盒攻击则假设攻击者对模型的内部结构和参数一无所知，仅能通过查询模型的输入输出关系来设计扰动。此时，攻击者需要通过探索和试验逐步推测模型行为，并根据输出反馈不断调整扰动策略。灰盒攻击介于白盒和黑盒攻击之间，假设攻

击者对模型有部分了解，如已知模型的架构或训练数据，但不了解具体的参数或梯度信息。在这种情况下，攻击者可以利用已有的信息设计扰动，并通过进一步探索来弥补未知部分。

针对节点分类任务[18]，常见的对抗攻击方法有 GradArgmax，SGA，NIPA 和 NETTACK 等。GradArgmax 由 Dai 等[19]于 2018 年提出，该方法首先根据损失函数计算每条边的梯度，以评估对模型性能的影响。然后，利用贪心算法选择那些对损失函数影响最大的进行修改，从而有效提高攻击成功率。通过这种方式，GradArgmax 能够在保持攻击效果的同时，尽量减少对原始数据结构的破坏；SGA 是 Li 等[20]于 2021 年提出的，其核心思想是首先提取以目标节点为中心的多阶子图，以捕捉与目标节点相关的局部结构信息。在此基础上，SGA 通过分析梯度信息，选择并删除或添加子图中梯度最大的边，来有效地生成对抗性攻击。NETTACK 是一种白盒攻击方法，由 Zügner 等[21]于 2018 年提出，针对图神经网络的节点分类任务，NETTACK 通过优化一个损失函数来选择最优的边和特征修改，以最大化目标节点分类错误的可能性。在较小的扰动预算下，NETTACK 展现了高效的攻击效果。NIPA 是一种投毒攻击方法，由 Sun 等[22]于 2019 年提出，投毒攻击通过在训练阶段注入恶意样本来降低模型整体性能，NIPA 计算每个节点对目标模型输出的影响力，选择影响力最大的节点进行投毒，通过添加或删除边以及修改节点特征来实现攻击。NIPA 能在训练数据中注入少量恶意样本的情况下显著降低模型性能。

2.4　联邦防御方法

在图垂直联邦学习中的防御研究中，对抗性训练和差分隐私是常用的防御方法。对抗性训练由 Goodfellow 等[23]于 2014 年提出，在机器学习和深度学习的安全性研究中得到了广泛关注。其核心思想是通过动态生成对抗样本，不断训练模型，最终使得模型具备抵御对抗性扰动的能力。许多研究表明，对抗性训练在各类图神经网络中均有显著的防御效果，特别是在针对白盒攻击的防御中表现出较强的鲁棒性。

差分隐私最早由 Dwork 等[24]于 2006 年提出，最初应用于数据库查询、数据挖掘和统计分析等领域，通过在数据共享和模型训练过程中添加噪声，来保护个体样本的隐私，确保攻击者无法从输出数据中推断出数据来源。Abadi 等[25]于 2016 年将差分隐私扩展到深度学习，使得用户无法通过输出结果辨别数据集的具体来源，进一步提升了隐私保护效果。随着联邦学习的发展，差分隐私的应用逐渐增加，Wang 等[26]提出了用于 VFL 的混合差分隐私框架，证明了在垂直分区数据上联合学习广义线性模型的可行性，且成本可以忽略不计。差分隐私在应对图结构上的攻击时表现出显著优势，并且随着隐私预算的优化，性能损失较小，展示了其在保护隐私和维持模型性能之间的良好平衡。

这两种方法为提升图垂直联邦学习的安全性提供了全面的防护体系，既能够增强模型抵抗攻击的鲁棒性，又能保护参与方数据的隐私安全。

3　准备工作

3.1　图神经网络

图神经网络是一类能够在图结构数据上进行有效学习的模型，传统的深度学习模型在处理像图这种复杂数据结构时表现不佳。因为图数据中的节点和边构成的复杂关系无法直接用于经典的卷积或全连接神经网络中，而图神经网络通过聚合节点邻域的特征来学习节点或图的表示。经典的图神经网络有 GCN，GAT，GraphSage 等。

3.1.1　图卷积神经网络

GCN 基于谱图理论，通过邻接矩阵对节点特征进行聚合。GCN 的设计初衷是有效地对图结构数据进行卷积操作，使得每个节点可以整合其邻居节点的特征信息，生成有意义的节点表示。GCN 的计算式如下：

$$\boldsymbol{H}^{l+1}=\sigma(\hat{\boldsymbol{D}}^{(-1/2)}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{(-1/2)}\boldsymbol{H}^{(l)}\boldsymbol{W}^{l}) \tag{1}$$

其中，$\boldsymbol{H}^{(l)}$ 表示第 l 层的节点特征矩阵，$\boldsymbol{H}^{(0)}=X$ 为输入的节点特征矩阵；$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$ 是邻接矩阵加上自环；$\hat{\boldsymbol{D}}$ 是度矩阵；σ 为激活函数（如 ReLU）。

GCN 通过邻接矩阵聚合邻居节点的特征，同时其包含节点自身的信息，从而生成更具代表性的节点嵌入。由于其计算效率高，GCN 被广泛应用于节点分类等图结构任务，但在大规模图或动态图上的表现有一定局限性。

3.1.2　图注意力网络

图注意力网络通过引入注意力机制，在聚合邻居节点特征时动态分配不同的权重。特征更新计算式为：

$$h_i'=\sum_{j\in N(i)}\alpha_{ij}\boldsymbol{W}h_j \tag{2}$$

其中，h_i 表示节点 i 的特征向量；$\boldsymbol{W}$ 是线性变换矩阵；α_{ij} 是节点 i 和节点 j 之间的注意力权重，其定义为：

$$\alpha_{ij}=\frac{e^{LeakyReLU(a^{\mathrm{T}}[\boldsymbol{W}h_i\,\|\,\boldsymbol{W}h_j])}}{\sum_{k\in N(i)}e^{LeakyReLU(a^{\mathrm{T}}[\boldsymbol{W}h_i\,\|\,\boldsymbol{W}h_k])}} \tag{3}$$

其中，$\|$ 表示向量的连接操作，a 是可训练的权重向量。

GAT 通过这种自适应的加权机制，使模型能够更灵活地处理异构图或噪声较多的图数据，能够更好地捕捉图中的重要局部结构。

3.1.3　图采样聚合网络

GraphSage 是一种通过采样固定数量的邻居节点进行特征聚合的图神经网络模型，与 GCN 和 GAT 不同，GraphSage 采用采样机制来应对大规模图上的训练挑战，并通过聚合函数从采样的邻居节点中提取信息。其核心的聚合计算式为：

$$h_i^{l+1}=\sigma(\boldsymbol{W}^{l}\cdot AGG(\{h_j^{l},\forall\, j\in N(i)\})) \tag{4}$$

其中，AGG 是灵活的聚合函数（如平均、池化等）。

3.2　垂直联邦学习

垂直联邦学习是一种分布式机器学习框架，适用于多个数据拥有者之间的数据样本重叠但特征不重叠的场景（见图 1）。在 VFL 中，各参与方在不直接共享其原始数据的前提下，通过协同训练模型来实现共同学习的目标。本文将垂直联邦学习框架应用于节点分类任务，多个客户端在本地训练节点分类模型，并通过在服务器端聚合各客户端的分类结果，来实施后续的攻击。以节点分类为例，有以下推导。

设有 K 个客户端，客户端 k 拥有特征 X_k 和标签 Y_k。那么，客户端的节点分类可以表示为：

$$f_k:X_k\to Y_k \tag{5}$$

每个客户端在本地训练模型，得到分类器 C_k，然后进行预测。得到节点预测结果为：

$$\hat{Y}_k=C_k(X_k) \tag{6}$$

在本地训练完成后，每个客户端将其输出结果（预测标

签)发送至服务器端。服务器端接收到所有客户端的预测结果后,进行聚合。假设服务器端平均聚合不同客户端的预测结果,聚合后的结果表示为:

$$\hat{Y}=\frac{1}{K}\sum_{i}^{K}\hat{Y} \tag{7}$$

通过上述操作,完整地模拟了垂直联邦学习,为后续的操作提供了基础。例如,对聚合后的预测结果实施攻击,以达到对攻击方法性能测试的目的。

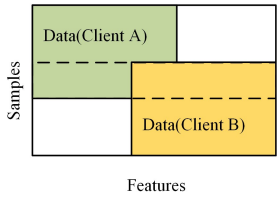

图 1 垂直联邦学习场景

Fig. 1 Vertical federated learning scenarios

3.3 威胁模型

本文提出了一种基于垂直联邦学习的节点分类任务协同框架。多个客户端在 VFL 框架中分别完成本地节点分类训练,并将预测结果传输至中央服务器,由服务器端进行结果聚合。攻击者目标为聚合结果,利用多维度攻击(如结构维度、特征维度)干扰节点分类任务,同时攻击者仅能访问预测结果及部分特征和梯度信息。训练与测试过程中,中央服务器的结构和参数均对客户端保持不可见,确保系统的隐私保护。

攻击者可以是系统内的恶意客户端或具备访问权限的外部实体,拥有获取并干预聚合结果的权限。攻击者在聚合输出层面添加对抗性噪声或者虚假节点,从而破坏模型整体的预测性能,而非针对特定标签进行误导。采用无目标攻击策略,攻击者的目标在于尽量降低模型整体预测的准确性与稳定性。

系统中的善意客户端或中央服务器端可以担负起防御职责,致力于保持分类准确率和系统稳定性以抵御外部攻击。防御策略包括善意客户端在模型输入阶段或输出阶段注入噪声,以增加对抗干扰的鲁棒性;或在本地模型训练阶段加入对抗样本,使得模型能够抵御噪声攻击。最终目的是使攻击对模型整体准确率和损失的影响降至最低,保障系统的鲁棒性和安全性。

4 面向图垂直联邦的对抗攻击方法

4.1 系统框架

本文提出了一种高效的地图联邦对抗攻击方法(Node and Feature Adversarial Attack, NFAttack),其系统框架如图 2 所示。

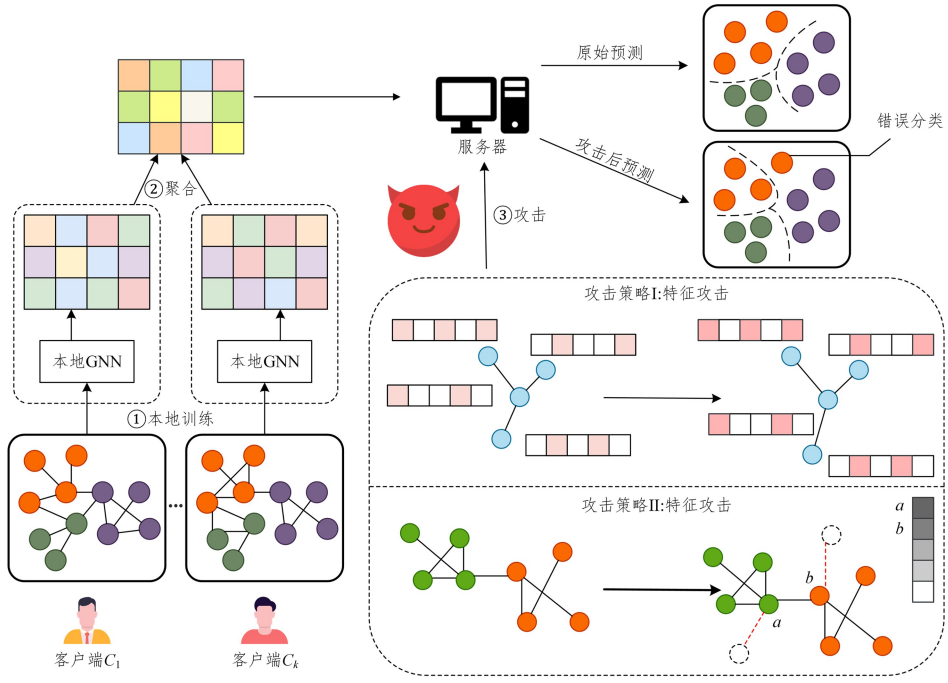

图 2 NFAttack 系统框图

Fig. 2 Framework of NFAttack

首先,各个客户端独立进行模型训练,以完成节点分类任务,并生成相应的输出结果;然后,服务器端对这些输出结果进行聚合,得到聚合结果;最后,针对聚合结果实施设计的攻击方法。具体而言:

1)客户端进行模型训练:每个客户端独立完成节点分类任务,基于分割好的本地数据对模型进行训练和推理,生成初步的节点分类结果。随后,每个客户端将训练好的分类结果传送给服务器,以便进行后续的聚合操作。值得注意的是,客户端的数量将在敏感性分析实验中具体分析。

2)聚合输出结果:服务器端负责接收来自多个客户端的节点分类结果,并采用某种聚合机制(如平均聚合、最大值聚合、最小值聚合等)生成聚合输出,为后续的攻击步骤提供基础。

3)实施攻击:基于聚合结果,实施 NFAttack 攻击,包括节点攻击策略和特征攻击策略,并且这两种策略独立实施。节点攻击策略通过筛选出节点度中心性高的节点,并将一定数量的虚假节点与筛选出来的节点相连以构成虚假边,从而对服务器端的聚合结果造成攻击。特征攻击策略通过对节点特征添加微量的混合噪声来实现攻击,其中混合噪声由随机噪声和对抗性梯度噪声组成。

4.2 针对节点特征的攻击方法实现

在本小节中实现了针对节点特征的对抗攻击方法——混合噪声攻击方法。混合噪声攻击方法旨在通过组合随机噪声

和对抗性梯度噪声两种干扰源，来最大化扰乱图神经网络的输出，从而显著降低其分类准确率。该方法充分利用了对抗性噪声的攻击性以及随机噪声的不可预测性，使得聚合模型在面对攻击时表现出明显的脆弱性。

假设有多个客户端，其中每个客户端独立地训练模型，并生成相应的输出。这些输出在经过聚合后，形成最终的聚合结果。混合噪声攻击的目标是通过向这个聚合结果添加精心设计的噪声，来诱导模型做出错误的节点分类决策。具体实现如下：

1）计算聚合输出：在多模型模拟联邦学习中，假设有 N 个图神经网络模型 $\{f_1, f_2, \cdots, f_N\}$，即每个模型的输出为 f_i。这些模型的输出聚合 y_{combined} 通过对所有模型的输出进行平均得到：

$$y_{\text{combined}} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{8}$$

2）生成对抗性噪声：对抗性梯度噪声的生成依赖于对模型损失函数的梯度计算。设损失函数为：

$$\mathcal{L}(y_{\text{combined}}, y_{\text{true}}) \tag{9}$$

其中，y_{true} 为真实标签。通过对损失函数关于聚合输出 y_{combined} 求导得到梯度 $\nabla_{y_{\text{combined}}}\mathcal{L}$：

$$n_{\text{grad}} = sign(\nabla_{y_{\text{combined}}}\mathcal{L}) \tag{10}$$

3）生成随机噪声：为了增加攻击的随机性和不可预测性，混合噪声攻击还引入了随机噪声 n_{rand}，该噪声从标准正态分布中采样得到：

$$n_{\text{rand}} \sim N(0,1) \tag{11}$$

随机噪声的引入使得每次攻击的具体效果不同，进一步增加了模型防御的难度。

4）混合噪声攻击：首先将对抗性梯度噪声和随机噪声按比例 α 进行线性组合，生成最终的混合噪声 n_{mix}：

$$n_{\text{mix}} = \alpha n_{\text{rand}} + (1-\alpha) n_{\text{grad}} \tag{12}$$

其中，α 是噪声比例参数，决定了随机噪声和对抗性噪声在混合噪声中的权重。通过调整 α，可以平衡攻击的随机性和定向攻击的强度。

最后，将生成的混合噪声乘以攻击强度 ϵ 并添加到聚合输出 y_{combined} 上，得到攻击后的输出 y_{attacked}：

$$y_{\text{attacked}} = y_{\text{combined}} + \epsilon\, n_{\text{mix}} \tag{13}$$

为了确保攻击后的输出在合法范围内，我们使用夹紧操作 $clamp$ 将攻击后的输出限制在[0,1]的区间内：

$$y_{\text{attacked}} = clamp(y_{\text{combined}} + \epsilon\, n_{\text{mix}}, 0, 1) \tag{14}$$

通过这种方式，模型的输出被成功扰乱，攻击的效果体现在模型的分类准确率显著下降。

混合噪声攻击方法通过综合利用对抗性梯度噪声的定向攻击能力和随机噪声的不可预测性，有效地扰乱了图神经网络的输出。这种攻击不仅能够精准地削弱模型的性能，还能通过引入随机性增加攻击的复杂性，使得传统的防御方法难以应对。随着 α 和 ϵ 的调整，攻击者可以在攻击的强度和随机性之间找到最佳平衡，从而在不同场景下实现最佳攻击效果。

4.3　针对节点的攻击方法实现

在前一小节中，针对节点特征进行了攻击。然而，仅仅针对特征的攻击并不足以全面评估模型的安全性。因此，接下来在本小节中，引入针对节点的对抗攻击方法，通过输入虚假节点和虚假连边，系统性地设计图联邦框架下的对抗攻击方法。这一策略不仅能有效干扰模型的学习过程，还能深入解释图神经网络在面对潜在威胁时的脆弱性。

针对节点攻击方法的实现步骤中，数据划分与聚合步骤和特征攻击方法的步骤一致，在此处省略。其余步骤如下：

1）度中心性计算：计算节点的度中心性可以识别图网络中最具影响力的节点，这些节点将成为攻击时的连接节点。具体来说，度中心性 $D(i)$ 通过邻接矩阵 $\boldsymbol{A}$ 计算，具体表达式为：

$$D(i) = \sum_{j \in N(i)} \boldsymbol{A}_{ij} \tag{15}$$

其中，$N(i)$ 是节点 i 的邻居集合。通过对所有节点的度中心性进行排序，可以选择度数最高的前 K 个节点，这些节点将作为与虚假节点连接的目标，以增强虚假节点的攻击能力。

2）虚假节点添加：在这一步骤中，生成 N_f 个虚假节点，其特征通过从正态分布中随机采样得到。

$$X_{\text{fake}} \sim N(0,1)(N_f \times F) \tag{16}$$

其中，每个虚假节点与度中心性最高的部分真实节点进行连接，形成新的边集合 E_{new}：

$$E_{\text{new}} = E \cup \{(N_{\text{fake}}, N_j) \mid N_j \in N_{\text{top}\,K}\} \tag{17}$$

经过这样的设计可以显著地增加图结构中的连接复杂性，从而对聚合后的结果产生有效的干扰。

3）去除自环：为确保图的结构合理性，对新生成的边集合执行去自环操作，得到最终的边集合 E_{final}：

$$E_{\text{final}} = remove\_self\_loops(E_{\text{new}}) \tag{18}$$

去除自环可以防止由于自连接引起的错误信息传播，从而提升了图的有效性。

通过以上步骤，成功实施了节点攻击方法，在图结构中引入虚假节点以及虚假边。这种攻击策略有效评估了模型在面对潜在安全威胁时的脆弱性，为后续针对图垂直联邦框架的防御措施提供了一定的基础。

4.4　NFAttack 伪代码

本文提出的 NFAttack 方法包括两个攻击策略：节点攻击策略和特征攻击策略。节点攻击策略利用度中心性指标评估节点重要性，通过与高中心性节点连接一定数量的虚假节点形成虚假边，以实现攻击。特征攻击策略向节点特征添加由随机噪声和梯度噪声构成的混合噪声，进而影响分类结果。具体的伪代码如算法 1 所示。

算法 1　节点攻击策略

输入：Aggregate_data，虚假节点数量

输出：attack_out_data

1. 生成虚假节点特征 fake_features
2. 将 fake_features 添加到原图的特征矩阵
3. 计算原图节点度中心性并排序
4. 为每个虚假节点添加虚假边，并连接到度中心性最高的节点
5. 去除自环
6. 构建新图 new_data(data_x，data。y，data。mask…)
7. return new_data

算法 2　特征攻击策略

输入：Aggregate_data，噪声比 α，噪声强度 ϵ

输出：attack_out_data

1. 计算 Aggregate_data 梯度
2. 生成噪声：

2.1. 随机噪声 $n_{\text{rand}} \sim N(0,1)$

2.2. 对抗性梯度噪声 $n_{grad}=\text{sign}(\nabla_{y_{combined}}\mathcal{L})$

3. 生成混合噪声：$n_{mix}=\alpha n_{rand}+(1-\alpha)n_{grad}$

4. 添加噪声：$y_{attacked}=\text{clamp}(y_{combined}+\epsilon\, n_{mix},0,1)$

5. return attack_out_data

4.5 算法复杂度分析

时间复杂度分析：NFAttack 的时间复杂度主要受到图数据规模和客户端数目的影响。因此，在计算时间复杂度时需考虑这两个因素。对于图数据，其时间复杂度为 $O(E)$，E 为图数据中边的数量，由于垂直联邦学习会涉及 K 个客户端，因此总的时间复杂度为 $O(E\times K)$。

空间复杂度分析：空间复杂度表示算法的存储空间与数据规模之间的关系，在 NFAttack 中有 K 个客户端，每个客户端处理的数据规模为 E。因此，该算法的空间复杂度为 $O(K\times E)$。

综上，经过复杂度分析可以发现随着客户端数量或数据规模的增加，算法的内存消耗也会增加。

5 实验分析

本章将全面分析 NFAttack 攻击方法的有效性和参数敏感性，深入探讨其在不同设置下的表现。同时，对实验环境进行详尽说明，包括所选数据集、模型的具体配置以及攻击参数的设置和调整等。

5.1 实验设置

5.1.1 实验环境

本文的实验环境的具体配置如下：CPU 型号为 Intel Xeon Gold 6240-2.60 GHz，GPU 型号为 Tesla V100-SXM3-32 GB，操作系统为 Ubuntu 20.04.4，编程语言为 Python 3.7，实验中使用的深度学习框架为 Pytorch-cuda-181。

5.1.2 数据集

本文共使用了 6 个数据集，即 Cora，CiteSeer，PubMed[27] 以及 Computers，Photo 和 CS[28]。其中 Cora，CiteSeer，PubMed 和 CS 均为学术引用网络，节点表示论文，边表示论文之间的引用关系，节点特征通常是论文的文本特征，标签则是论文的类别（例如所属领域）。Photo 和 Computers 来源于亚马逊共购网络，节点表示商品，边表示商品之间的共购关系，节点特征是产品的属性（如类别或文本描述的向量化表示），标签则是产品所属的分类。每个数据集的具体参数如表 1 所列。

表 1 实验数据集

Table 1 Experimental datasets

Datasets	Nodes	Edges	Features	Labels
Cora	2708	10556	1433	7
CiteSeer	3327	9104	3703	6
PubMed	19717	88648	500	3
Computers	13752	491722	767	10
Photo	7650	238162	745	8
CS	18333	163788	6805	15

5.1.3 实验模型

本文在垂直联邦学习场景下，针对 6 个数据集分别使用了 GCN[14]，GAT[16] 和 GraphSage[15] 这 3 种图神经网络模型进行节点分类实验，并在此基础上实施对抗攻击实验，具体模型结构与参数如表 2 所列。集中式学习和联邦学习框架下的节点分类结果如表 3 所列，相比于集中式学习的节点分类准确率，垂直联邦学习的分类准确率稍有下降，但这对后续的攻防研究的影响可以忽略不计，这些分类结果为后续的对抗攻击实验提供了坚实的基础。

表 2 本地 GNN 模型的结构与参数

Table 2 Structure and parameters of the local GNN

模型	层数	隐藏层维度	激活函数	训练周期
GCN	2	16	ReLU	200
GAT	2	8	ELU	200
GraphSage	2	16	ReLU	200

表 3 集中式学习与垂直联邦学习下节点分类准确率的对比

Table 3 Comparison of node classification accuracy under centralized learning and vertical federated learning

(%)

Datasets	集中式学习			垂直联邦学习		
	GCN	GAT	GraphSage	GCN	GAT	GraphSage
Cora	80.20	80.70	79.50	80.00	80.50	78.70
CiteSeer	69.50	68.00	68.50	68.50	67.80	68.00
PubMed	68.90	75.00	77.40	66.70	74.20	77.40
Computers	79.97	83.34	79.05	79.45	82.86	78.06
Photo	89.88	91.18	90.00	89.69	90.97	89.92
CS	91.60	87.80	89.60	91.50	87.70	89.60

5.1.4 对比算法

在此小节中，将 NFAttack 与 3 种不同的对抗攻击方法进行对比，包括 GradArgmax[19]，SGA[20] 和 NETTACK[21]。GradArgmax 根据损失函数计算连边上的梯度，并采用贪心算法选择需要修改的连边。SGA 计算图神经网络的梯度信息，针对图结构数据中节点特征进行微调，以实现对模型的破坏。NETTACK 通过对节点特征进行微小扰动，使模型在攻击目标节点上的预测发生错误。

上述对比算法均在 6 个数据集和 3 种模型上分别进行攻击实验，利用这两种攻击方法和 NFAttack 在相同条件下分别实施对抗攻击，通过分析这 3 种攻击方法之间的攻击成功率来比较其攻击效果的好坏，进一步证明 NFAttack 攻击方法的有效性。

5.1.5 攻击算法设置

在特征攻击实验中，攻击算法采用了混合噪声的方式对聚合后的模型输出进行干扰。首先，针对目标节点的输出，通过反向传播计算梯度噪声，同时生成与模型输出形状相同的随机噪声。然后，利用噪声比例参数 α 将随机噪声与对抗性梯度噪声进行加权混合，形成混合噪声。最后，通过参数 ϵ 来控制混合噪声的强度，将其添加至模型输出以实现攻击效果。

在节点攻击实验中，攻击引入的虚假节点数量为 200，每个虚假节点连接到度中心性排名靠前的真实节点，形成虚假连边，以最大化对模型的影响。这些参数的合理配置使得攻击能够有效地评估图神经网络模型在垂直联邦学习环境中的鲁棒性。为了确保实验结果的可重复性，设置随机种子为 100，在数据加载、模型训练及攻击过程中统一使用。

5.1.6 联邦学习算法设置

本文在攻击方法的有效性实验中设置 3 个客户端数目进行分析，以初步评估不同客户端数量对攻击效果的影响。然而，在随后的敏感性分析实验中，为了更全面地探索这一影响，客户端的数目被逐步扩展到 2，3，4，5，6。这一设计旨在

深入分析客户端数量变化对攻击效果的具体影响，并提供更为细致的实验数据。此外，主体实验中采用了平均聚合方式来聚合各客户端的输出结果，而在敏感性分析实验中，除了使用平均聚合外，还引入了最大值聚合和最小值聚合的策略。这种多样化的聚合方式可以更全面地理解不同聚合策略对攻击效果的影响。

5.1.7　评价指标

采用攻击成功率（Attack Success Rate，ASR）作为攻击效果的主要评价指标，用于衡量攻击方法对模型性能的影响。本文使用攻击前后的准确率变化来计算 ASR：

$$ASR=\frac{ACC_{\text{before}}-ACC_{\text{after}}}{ACC_{\text{before}}} \tag{19}$$

其中，ACC_{before} 表示攻击前的准确率，ACC_{after} 表示攻击后的准确率。ASR 值越大，表明攻击效果越明显。

选取余弦相似性（Cosine Similarity，CS）[29] 作为衡量攻击方法隐蔽性的指标，旨在评估攻击方法在成功影响模型性能后，其隐蔽性的程度。具体而言，通过对比攻击前后的图数据，计算余弦相似性值以评估攻击的隐蔽性：

$$CS(y_{\text{before}},y_{\text{after}})=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_{\text{before}}^{i}\cdot y_{\text{after}}^{i}}{\|y_{\text{before}}^{i}\|\cdot\|y_{\text{after}}^{i}\|}\right) \tag{20}$$

其中，y_{before}^{i} 是攻击前第 i 个节点的输出，y_{after}^{i} 为攻击后对应节点的输出，N 为测试集中的节点数。CS 值只能在[−1，1]范围内，当 CS 取值越靠近 1 时，说明该攻击方法越隐蔽。

5.2　NFAttack 攻击方法的有效性分析

本小节针对 Cora，CiteSeer，PubMed，Photo，Computers，CS 这 6 个数据集，在 GCN，GAT，GraphSage 这 3 种模型上进行了 4 种不同攻击方法的对比实验。实验通过对比 NFAttack 与其他 3 种攻击方法的效果，旨在全面地验证 NFAttack 在不同数据集和不同模型条件下攻击的有效性。为确保实验结果的公平性和可对比性，客户端参数等实验设置均保持相同。同时，为了更直观地展现 4 种攻击方法在各个数据集和模型上的效果差异，通过绘制直方图可更显著地突出 NFAttack 在攻击效果上的优势。需要注意的是，第一列直方图为 NFAttack 方法的攻击策略对比结果，第二列为特征攻击策略的对比结果，图中每一行直方图分别对应 GCN，GAT，GraphSage 模型上的对比结果。遗憾的是，由于 NFAttack 的节点攻击策略在 CS 数据集和 GraphSage 模型上对计算资源的需求较高，当前条件下无法满足，因此未能获得该部分的实验结果。

实验结果如图 3 所示，首先可以观察到，NFAttack 在 6 个数据集和 3 种模型上均可以实现对抗攻击，且各场景下的攻击成功率相对较高，进一步验证了 NFAttack 方法的有效性。具体来看，NFAttack 的攻击成功率不仅显著高于传统方法 GradArgmax 和 NETTACK，在大多数实验条件下也优于 SGA 的攻击成功率，尤其在高噪声和复杂数据集环境中，NFAttack 表现出更强的适应性和鲁棒性。因此，从实验对比中可以得出结论，所提出的 NFAttack 方法在应对多种对抗场景中具有较高的攻击成功率，进一步验证了 NFAttack 方法的有效性。

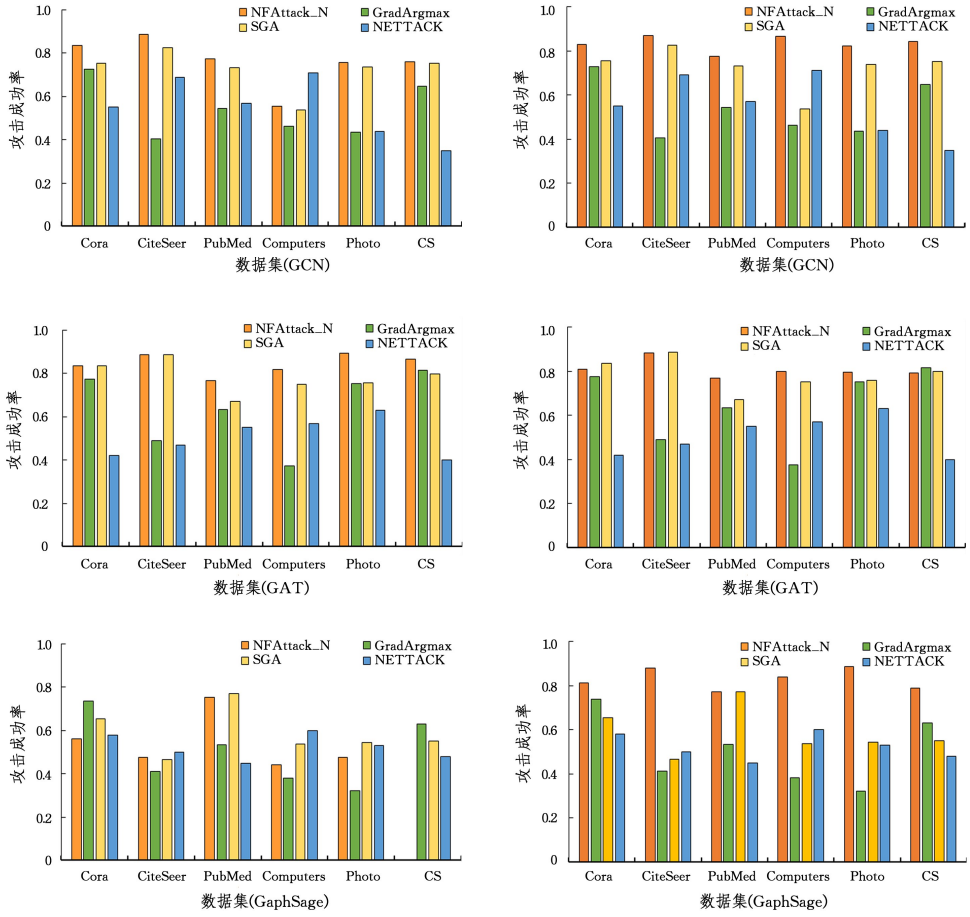

图 3　NFAttack 攻击方法与其他攻击方法的比较

Fig. 3　Comparison between NFAttack and other adversarial attack methods

5.3 NFAttack 攻击方法的隐蔽性分析

本小节对 NFAttack 的隐蔽性进行了分析，通过计算攻击前后预测结果的余弦相似度，实验结果如表 4 所列，其中 NFAttack_N 表示节点攻击策略，NFAttack_F 表示特征攻击策略。将 NFAttack 的不同攻击策略与其他攻击方法进行对比，结果显示，NFAttack 的两种策略的 CS 值均接近 1，表明其隐蔽性优越。与此相比，虽然 NETTACK 的 CS 值更接近 1，表示其隐蔽性更优，但攻击效果不如 NFAttack。这进一步验证了在保证攻击效果的前提下，NFAttack 具有最优的隐蔽性。

表 4 NFAttack 的隐蔽性分析

Table 4 Cryptic analysis of NFAttack

攻击方法	ASR	CS
GradArgmax	0.73	0.79
SGA	0.75	0.85
NETTACK	0.55	0.99
NFAttack_N	0.84	0.93
NFAttack_F	0.83	0.92

5.4 NFAttack 攻击方法的鲁棒性分析

为验证 NFAttack 攻击方法的鲁棒性，本实验分别在 NFAttack 前加入两种不同的防御方法，并将无防御情况下的 NFAttack 攻击效果与引入防御方法后的攻击效果进行对比，从而探索 NFAttack 在应对不同防御策略下的表现。本文选取了对抗性训练和差分隐私作为防御手段。

对抗性训练通过在训练阶段加入对抗性样本，来提升模型对攻击的抗干扰能力。具体而言，本实验使用 IG-FGSM[30] 方法来产生对抗样本，并将其与正常样本结合用于训练，以增强模型的鲁棒性。差分隐私[31] 则是一种基于隐私保护的防御机制，旨在通过数据噪声来减小攻击者推断单个样本信息的可能性。在本实验中，差分隐私通过在原始特征数据上添加拉普拉斯噪声来进行防御，随后将噪声处理后的数据用于模型训练，从而可以有效地抵御对抗攻击。

实验结果如表 5 所列，尽管引入防御方法后，NFAttack 的攻击成功率在少部分数据集模型条件下有所下降，但整体来看，该方法仍然具有较高的攻击效果，进一步验证了攻击方法的有效性与鲁棒性。从表 5 中还可以观察到，特征攻击策略的攻击效果普遍优于节点攻击策略的攻击效果，并且在应对不同防御方法时，特征攻击策略的攻击效果仍表现出较高的攻击成功率。遗憾的是，由于 NFAttack 的节点攻击策略在 CS 数据集和 GraphSage 模型上对算力需求较高，暂时无法满足，因此未能获得该部分实验结果。

因此，可以认为，NFAttack 不仅在无防御条件下具有较强的攻击能力，即使在防御条件下也能够有效实施攻击，进一步证明了其作为对抗攻击方法的可行性。

表 5 防御前后攻击成功率对比(ASR)

Table 5 Comparison of attack success rate before and after defense

(%)

攻击策略	模型	数据集	无防御	对抗训练防御	差分隐私防御
特征攻击策略	GCN	Cora	82.79	77.35(−5.44)	82.00(−0.79)
		CiteSeer	86.80	82.24(−4.56)	83.59(−3.21)
		PubMed	77.36	73.80(−3.56)	77.22(−0.14)
		Computers	86.68	84.75(−1.93)	85.03(−1.65)
		Photo	82.07	75.90(−6.17)	80.15(−1.92)
		CS	84.39	81.17(−3.22)	83.92(−0.47)
	GAT	Cora	80.85	68.81(−12.04)	80.43(−0.42)
		CiteSeer	88.42	82.17(−6.25)	88.36(−0.06)
		PubMed	76.86	74.07(−2.79)	76.04(−0.82)
		Computers	79.87	76.58(−3.29)	74.62(−5.25)
		Photo	79.51	66.01(−13.50)	76.34(−3.17)
		CS	79.18	64.08(−15.10)	72.50(−6.68)
	GraphSage	Cora	81.06	75.73(−5.33)	80.36(−0.70)
		CiteSeer	87.84	83.96(−3.88)	86.30(−1.54)
		PubMed	77.10	75.97(−1.13)	76.10(−1.00)
		Computers	83.74	81.17(−2.57)	78.12(−5.62)
		Photo	88.73	75.77(−12.96)	86.24(−2.49)
		CS	78.78	64.48(−14.30)	71.64(−7.14)
节点攻击策略	GCN	Cora	83.68	81.73(−1.95)	76.49(−7.19)
		CiteSeer	88.79	69.35(−19.44)	61.47(−27.32)
		PubMed	77.36	47.79(−29.57)	75.10(−2.20)
		Computers	55.58	50.01(−5.57)	40.02(−15.56)
		Photo	75.74	75.64(−0.10)	52.43(−23.11)
		CS	76.16	65.98(−10.18)	69.01(−7.15)
	GAT	Cora	83.65	81.19(−2.46)	83.09(−0.56)
		CiteSeer	88.68	71.01(−17.67)	54.42(−34.26)
		PubMed	76.86	70.86(−6.00)	75.64(−1.22)
		Computers	81.71	75.36(−6.35)	72.39(−9.32)
		Photo	89.28	88.95(−0.33)	54.16(−35.12)
		CS	86.49	70.83(−15.66)	75.32(−11.17)
	GraphSage	Cora	56.12	34.43(−21.69)	55.54(−0.67)
		CiteSeer	47.56	40.11(−7.45)	38.73(−8.83)
		PubMed	75.32	73.69(−1.63)	67.55(−7.77)
		Computers	47.66	41.22(−6.44)	41.54(−6.12)
		Photo	44.06	40.29(−3.77)	35.62(−8.84)
		CS	—	—	—

5.5　参数敏感性分析

本小节旨在分析垂直联邦学习中聚合方式和客户端数量对攻击效果的影响，以便为参数选择提供客观依据。具体而言，聚合方式的不同会导致攻击效果的显著差异，因为不同的聚合策略在整合各个客户端的模型参数时具有不同的特点和优势。图 4 和图 5 分别给出了不同聚合方式和不同客户端数量对 NFAttack 攻击效果的影响。实验结果清晰地表明，使用不同的聚合策略（最大值聚合、平均聚合和最小值聚合）时的攻击效果有明显差异。此外，客户端数量的变化也显著影响了攻击效果。因此，通过对这些因素的深入分析，可以为选择适当的参数配置提供有效依据，从而使攻击效果更优。

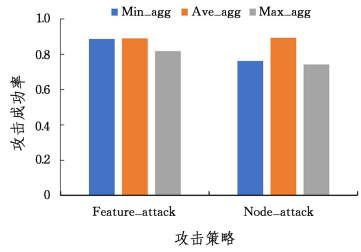

图 4　不同聚合方式

Fig. 4　Different aggregation

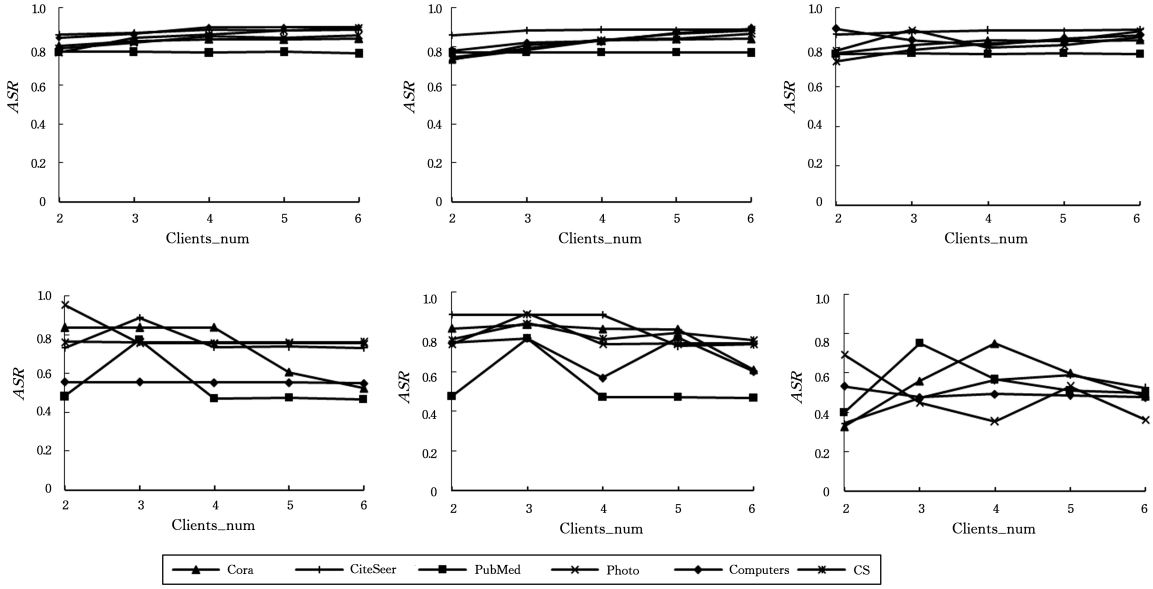

图 5　客户端数目的影响

Fig. 5　Attack on multiparticipant-based NFAttack

5.5.1　不同聚合方式对攻击效果的影响

本实验探究了不同聚合方式对攻击效果的影响，包括最大值聚合、平均聚合、最小值聚合。图 4 表明，在 3 种聚合方式中，平均聚合后的攻击效果显著优于最大值聚合和最小值聚合的攻击效果。这一现象主要是由于 NFAttack 攻击是对局部进行扰动，平均聚合能够更好地整合各个客户端的模型参数，使得聚合结果对 NFAttack 的攻击更为敏感。相比之下，最大值聚合和最小值聚合的选择更倾向于极端值，这增强了对局部扰动的抵抗力，从而削弱了在聚合结果上的攻击效果。

5.5.2　客户端数目对攻击效果的影响

图 5 中，第一行展示了特征攻击实验结果，第二行展示了节点攻击实验结果，三列分别对应使用了 GCN，GAT 和 GraphSage 模型。需要注意的是，由于数据集的限制，客户端数量设置受限。实验结果表明，随着客户端数目的增加，NFAttack 的攻击效果呈现下降趋势。这是由于客户端数量增加后，平均聚合方式将更多客户端的模型参数进行平均处理，导致模型参数趋于平滑和稳定，从而削弱了攻击的效果。此外，客户端数量的增加意味着需要攻击的数据量增多，通过对多个客户端的聚合，攻击影响被稀释，模型的鲁棒性得到增强，抗攻击能力也随之提高。整体而言，NFAttack 的攻击效果随着参与客户端数量的增加逐渐减弱。同时可以观察到，当客户端数目为 3 时，攻击效果较高。这一结果为本文其他实验提供了参考，将实验中客户端数目设置为 3 时，能够更有效地进行攻击。遗憾的是，由于 NFAttack 的节点攻击策略在 CS 数据集和 GraphSage 模型上对计算资源的需求较高，当前条件下无法满足，因此未能获得该部分实验结果。

结束语　针对现有的对抗攻击方法应用到图垂直联邦学习中的攻击效果不高的问题，本文提出了面向图垂直联邦的对抗攻击方法 NFAttack，该方法包括节点攻击策略和特征攻击策略。特征攻击通过向节点特征添加随机噪声和梯度噪声的混合噪声来影响分类结果，节点攻击策略通过利用高度中心性节点连接一定数量的虚假节点形成虚假边实现攻击。同时，在 6 个数据集和 3 种模型上展开了丰富的攻击实验，验证了 NFAttack 攻击方法的可迁移性和有效性。此外，通过添加防御机制，进一步验证了 NFAttack 的鲁棒性。

本文提出的攻击方法仍存在一些问题，在节点攻击策略中，添加虚假节点的同时添加了虚假边，仅针对边的攻击策略需进行进一步研究。此外，随着攻击效果的提升，对算力资源的需求也随之增加，因此未来的研究目标是确保高攻击效果的同时，降低算力资源消耗并提高攻击的隐蔽性。同时，在防御机制的研究中，应探索更有效的防御策略，以抵御多维度的攻击。

参 考 文 献

[1]　ZHANG C，XIE Y，BAI H，et al. A survey on federated learning [J]. Knowledge-Based Systems，2021，216：106775.

[2]　LIU P，XU X，WANG W. Threats，attacks and defenses to fed-

erated learning: issues, taxonomy and perspectives[J]. Cybersecurity, 2022, 5(1): 4.
[3] HENRIQUE B M, SOBREIRO V A, KIMURA H. Literature review: Machine learning techniques applied to financial market prediction[J]. Expert Systems with Applications, 2019, 124: 226-251.
[4] KONONENKO I. Machine learning for medical diagnosis: history, state of the art and perspective[J]. Artificial Intelligence in Medicine, 2001, 23(1): 89-109.
[5] CUMMINGS D, NASSAR M. Structured citation trend prediction using graph neural networks[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2020: 3897-3901.
[6] GAO C, WANG X, HE X, et al. Graph neural networks for recommender system[C]// Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1623-1625.
[7] ZHANG X M, LIANG L, LIU L, et al. Graph neural networks and their current applications in bioinformatics[J]. Frontiers in Genetics, 2021, 12: 690049.
[8] LUAN H, TSAI C C. A review of using machine learning approaches for precision education[J]. Educational Technology & Society, 2021, 24(1): 250-266.
[9] YU B, MBO W, LV Y, et al. A survey on federated learning in data mining[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(1): 1-20.
[10] HARD A, RAO K, MATHEWS R, et al. Federated learn-ing for mo-bile keyboard prediction[J]. arXiv: 1811. 03604, 2018.
[11] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology(TIST), 2019, 10(2): 1-19.
[12] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4-24.
[13] ZHAO T, JIN W, LIU Y, et al. Graph data augmentation for graph machine learning: A survey[J]. arXiv: 2202. 08871, 2022.
[14] KIPF T N, WELLING M. Semi-supervised classification with gra-ph convolutional networks[J]. arXiv: 1609. 02907, 2016.
[15] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-11.
[16] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv: 1710. 10903, , 2017.
[17] LI Y, CHENG M, HSIEH C J, et al. A review of adversarial attack and defense for classification methods[J]. The American Statistician, 2022, 76(4): 329-345.
[18] ZHANG T, LIAO B, YU J, et al. Benchmarking and Analysis for Graph Neural Network Node Classification Task[J]. Computer Science, 2024, 51(4): 132-150.
[19] DAI H, LI H, TIAN T, et al. Adversarial attack on graph structured data[C]// International Conference on Machine Learning. PMLR, 2018: 1115-1124.
[20] LI J, XIE T, CHEN L, et al. Adversarial attack on large scale graph[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 82-95.
[21] ZÜGNER D, AKBARNEJAD A, GÜNNEMANN S. Adversarial attacks on neural networks for graph data[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2847-2856.
[22] SUN Y, WANG S, TANG X, et al. Node injection attacks on graphs via reinforcement learning[J]. arXiv: 1909. 06543, 2019.
[23] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412. 6572, 2014.
[24] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Proceedings of Theory of Cryptography Conference. 2006: 265-284.
[25] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016: 308-318.
[26] WANG C, LIANG J, HUANG M, et al. Hybrid differentially private federated learning on vertically partitioned data[J]. arXiv: 2009. 02763, 2020.
[27] YANG Z, COHEN W, SALAKHUDINOV R. Revisiting semi-supervised learning with graph embeddings[C]// International Conference on Machine Learning. PMLR, 2016: 40-48.
[28] SHCHUR O, MUMME M, BOJCHEVSKI A, et al. Pitfalls of graph neural network evaluation[J]. arXiv: 1811. 05868, 2018.
[29] SUN M, TANG J, LI H, et al. Data poisoning attack against unsupervised node embedding methods[J]. arXiv: 1810. 12881, 2018.
[30] WU H, WANG C, TYSHETSKIY Y, et al. Adversarial examples for graph data: deep insights into attack and defense[C]// Proceedings of the Twenty Eighth International Joint Conference on Artificial Intelligence(IJCAI). 2019: 4816-4823.
[31] SUN M, DING X N, CHENG Q. Federated Learning Scheme Based on Differential Privacy[J]. Computer Science, 2024, 51(S1): 230600211-6.

BAI Yang, born in 2001, postgraduate. His main research interests include aerificial intelligence and internet security.

CHEN Jinyin, born in 1982, Ph.D, professor, is a member of CCF (No. 14348M). Her main research interests include data mining, intelligent computing and complex network analysis.