基于参数差异假设的图卷积网络对抗性攻击

吴翼腾[1],刘 伟[1],于淑乔[2]
(1. 信息工程大学信息技术研究所,河南郑州 450002; 2. 墨尔本大学,澳大利亚墨尔本 3010)

摘 要: 图神经网络容易受到对抗性攻击安全威胁. 现有图神经网络对抗性攻击思想可以概括为构造矛盾的训练数据. 矛盾数据假设不能很好地解释图神经网络过拟合训练数据的攻击场景. 本文以有效攻击前后图神经网络模型的训练参数应该具有较大差异为基本出发点,以图卷积网络为具体研究对象,建立基于参数差异假设的对抗性攻击模型. 将统计诊断的重要结果Cook距离引入对抗性攻击,提出基于Cook距离的参数差异度量方法. 采用基于Cook距离梯度的攻击方法,首次得出了攻击梯度的闭式解,并结合梯度下降算法思想和贪心算法思想提出完整的攻击算法. 最后设计实验验证了参数差异假设的合理性和基于该假设导出方法的有效性;验证了梯度信息对图场景离散数据的可用性;仿真示例说明了攻击梯度闭式解的正确性;与其他攻击方法对比分析了攻击方法的有效性.

关键词: 图卷积网络;对抗性攻击;矛盾数据假设;参数差异假设;Cook距离
基金项目: 自然科学基金创新研究群体项目(No.61521003);国家重点研发计划(No.2016QY03D0502);郑州市协同创新重大专项基金(No.162/32410218)
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112(2023)02-0330-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210222

Adversarial Attacks on Graph Convolution Networks Based on Parameter Discrepancy Hypothesis

WU Yi-teng[1], LIU Wei[1], YU Xu-qiao[2]
(1. *Institute of Information Technology*, *Information Engineering University*, *Zhengzhou*, *Henan* 450002, *China*;
2. *University of Melbourne*, *Melbourne* 3010, *Australia*)

Abstract: Graph neural networks (GNNs) are vulnerable to adversarial attacks. Existing GNN adversarial attacks can be generalized as constructing contradictory training data. However, the existing methods based on contradictory data hypothesis cannot explain well why the false outputs could be generated when GNNs fit the training data well. Firstly, based on the discrepancy of model parameters of GNNs before and after attack, poisoning attack model is proposed taking graph convolution network as a target. Secondly, a parameter difference metric, Cook distance, is proposed. The closed form solution of attack gradients is obtained for the first time, and an attack algorithm is given based on the idea of gradient descent and greedy algorithm. Finally, the rationality of the hypothesis of parameter discrepancy and the effectiveness of the proposed method are verified by experiments; the availability of gradients to discrete data of graph is verified; the correctness of closed form solution of attack gradients is illustrated by a numerical example; the effectiveness of attack method is analyzed compared with other attacks.

Key words: graph convolutional network; adversarial attack; contradictory data hypothesis; parameter discrepancy hypothesis; Cook distance
Fondation Item(s): National Natural Science Foundation of China (No.61521003); National Key R&D Plan (No.2016QY03D0502); Major Special Fund for Zhengzhou Collaborative Innovation (No.162/32410218)

1 引言

近年来随着深度学习方法的广泛应用,其安全问题也受到普遍关注. 2014年Szegedy等人[1]在研究图像数据的卷积神经网络(Convolutional Neural Networks, CNN)模型安全问题时提出"对抗样本"的概念. 图像中的对抗样本至今已从实验室水平发展至攻防实战阶段[2]. 研究对抗性攻击方法和对抗样本的产生机理[3]是深度学习方法应用于无人驾驶、智能安防等安全关键

收稿日期:2021-02-04;修回日期:2021-06-08;责任编辑:孙瑶

性领域的必要前提[4].

图数据与语义信息相结合具有表达现实数据的天然优势. 传统的复杂网络研究主要聚焦于节点和边的连接模式,研究图中关于节点、连边和局部结构的性质[5],难以将图数据和对应的属性信息有机结合. 深度学习方法在图像、文本等领域取得了目前已知的最好结果,却难以直接处理图数据. 为有效处理图数据和语义属性数据,诞生了以图卷积网络[6~9]为代表的图神经网络[10]. 图神经网络是专门针对图数据设计的端到端的深度学习模型[11~13],为含有图数据的应用场景提供了极具竞争力的学习方案.

与其他深度学习技术类似,图神经网络同样存在安全威胁. 有目的地对网络结构或属性特征施加微小扰动,会使图神经网络在具体任务中高置信度地给出错误输出. 2018年,Zügner等人[14]首次提出图神经网络对抗性攻击. 该课题近两年来逐渐受到更多关注[15~20]. 按污染数据的阶段划分,对抗性攻击分为污染训练数据的投毒攻击和污染测试数据的逃逸攻击. 与以往对图神经网络投毒攻击的理解不同,本文将图投毒攻击归结为攻击前后模型参数的显著差异. 现有文献[14,18,19,21,22]将投毒攻击形式化表达为,寻求攻击方法,使重训练的图神经网络在训练集上的损失函数达到最大,以降低测试数据上相关任务的准确率. 该表述所描述的投毒攻击可以概括为:寻求攻击方法,构造一组存在矛盾的训练数据,使图神经网络难以对其拟合,最终学习出一个“坏模型”,使其在测试集的效果显著下降. 矛盾数据假设是对投毒攻击本质的一种重要概括,但并不全面. 例如,存在这样的攻击场景——构造投毒的训练数据,使图神经网络能够对其很好地拟合,却难以对测试数据正确分类,即最终学习出一个“假模型”. 分析针对测试数据的实际攻击场景,模型结构和测试数据不变,发生改变的只有模型训练参数,可见模型训练参数对预测结果有较大影响. 本文将以此为动机展开研究.

基于以上分析,本文以攻击前后图神经网络的参数差异为基本出发点,建立基于参数差异假设的对抗性攻击模型. 主要面临如下挑战:

(1)如何衡量参数差异. 这需要建立参数差异的有效度量.

(2)如何求解攻击模型. 由于图神经网络处理的数据为离散数据,而传统观点[21]认为基于梯度的方法不适用于离散数据,但是在图神经网络攻击这一特殊场景中,基于梯度的方法是否有效不能一概而论,需要具体分析.

(3)若基于梯度的方法有效,能否得出攻击梯度的闭式解. 传统使用反向传播算法求解攻击梯度,无法得出攻击梯度的显式表达式,使得攻击方法具有黑盒性,不利于解析攻击原理和进一步寻找攻击问题的本质.

本文的主要贡献如下:

(1)基于“有效攻击前后图神经网络训练参数应该具有较大差异”的假设[17](简称参数差异假设)实施攻击,以图卷积网络为研究对象,建立新的对抗性攻击模型.

(2)从统计诊断经典文献[23~29]中引入Cook距离用以衡量攻击前后的参数差异. 设计实验验证参数差异度量方法的有效性、参数差异假设的合理性和其导出方法的有效性.

(3)推导出Cook距离对邻接矩阵和特征矩阵攻击梯度的闭式解,并对理论结果进行仿真. 设计实验验证图卷积网络对抗性攻击情景中离散数据基于梯度算法的有效性和可用性. 基于攻击梯度,根据梯度下降法和贪心算法的思想设计攻击算法,并进行对比实验.

2　基本概念

2.1　图和图卷积网络

图表示为$G(V,E)$,其中V表示节点集合,E表示连边集合. 设节点数$|V|=N$,则无权无向图可用对称的邻接矩阵$\boldsymbol{A}=\{0,1\}^{N\times N}$表示,$\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$. 图中每个节点有$n$维的特征向量,节点特征用矩阵$\boldsymbol{X}=\{0,1\}^{N\times n}$或$\boldsymbol{X}\in\mathbb{R}^{N\times n}$表示. 本文使用的特征矩阵为离散情况,$\boldsymbol{X}=\{0,1\}^{N\times n}$. 文献[7~9]将图卷积网络简化为SGC(Simple Graph Convolution),它具有低通滤波器的作用. 本文以图卷积网络SGC为攻击和研究对象. SGC模型的公式表达式为

$$\boldsymbol{Y}_{\mathrm{out}}=\mathrm{softmax}(\boldsymbol{\mathcal{A}}\boldsymbol{X}\boldsymbol{W}) \tag{1}$$

其中,$\boldsymbol{Y}_{\mathrm{out}}$为模型输出;$\boldsymbol{\mathcal{A}}$为滤波矩阵;$\boldsymbol{X}$为输入特征向量;$\boldsymbol{W}$为参数矩阵. 在文献[7]中,$\boldsymbol{\mathcal{A}}$的形式通常为

$$\boldsymbol{\mathcal{A}}=\tilde{\boldsymbol{L}}^{k},\tilde{\boldsymbol{L}}=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}},\tilde{\boldsymbol{A}}=\boldsymbol{I}+\boldsymbol{A},\tilde{\boldsymbol{D}}=\mathrm{diag}(\tilde{\boldsymbol{A}}\boldsymbol{1}) \tag{2}$$

记softmax(·)为$\sigma(\cdot)$,$\boldsymbol{Y}$表示one-hot编码的标签矩阵. $\boldsymbol{1}$表示N维全1列向量. 使用交叉熵损失函数,可得到矩阵形式的损失函数计算公式为

$$\mathcal{L}=-\mathrm{tr}\left[\boldsymbol{Y}^{\mathrm{T}}\ln[\sigma(\boldsymbol{Z}\boldsymbol{W})]\right] \tag{3}$$

2.2　图卷积网络的对抗性攻击

本文研究非指定目标、数据投毒攻击. 非指定目标攻击不指定具体的一个或几个攻击目标,需要使测试集的准确率整体下降;投毒攻击指允许图卷积网络对投毒的训练数据重新训练,重训练的图卷积网络在测试集的准确率仍然下降. 文献[14,18,19,21,22]建立了图卷积网络的对抗样本生成或投毒攻击模型. 按上述文献定义,攻击方法可以统一概括为以下约束优化问题:

$$(\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})=\underset{(\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})}{\arg\max}\ \mathcal{L}(\boldsymbol{W}^{*};\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}) \tag{4}$$

$$\begin{aligned}&\text{s.t. }\boldsymbol{W}^{*}=\underset{\boldsymbol{W}}{\arg\min}\ \mathcal{L}(\boldsymbol{W};\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})\\&\left\|\hat{\boldsymbol{A}}-\boldsymbol{A}\right\|_{0}+\left\|\hat{\boldsymbol{X}}-\boldsymbol{X}\right\|_{0}\leqslant\delta\end{aligned} \tag{5}$$

其中，$\boldsymbol{A}$ 和 $\boldsymbol{X}$ 是原邻接矩阵和特征矩阵；$\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$ 是扰动后的邻接矩阵和特征矩阵；$\|\cdot\|_0$ 为矩阵中非0元素的个数；δ 是扰动开销；$\boldsymbol{W}^*$ 是扰动后得到的训练参数. 投毒攻击允许对参数重新训练，属于双层优化问题. 若损失函数表示指定节点的损失函数，为指定目标攻击；若为全局损失函数，为非指定目标攻击. 文献[22]研究了非指定目标数据投毒攻击.

从上述投毒攻击的形式化表述中可以看出，现有攻击方法都是构造扰动后的样本数据，使得图卷积网络在扰动后数据集上损失函数最大，从而训练出不良参数. 现有方法可概括为基于“有效攻击的训练集中存在矛盾的训练数据”假设（简称矛盾数据假设）的攻击方法，攻击的实质是构造矛盾的训练数据.

3 基于参数差异假设的攻击方法

3.1 基于参数差异假设的投毒攻击模型

基于“有效攻击前后图卷积网络模型参数应该具有较大差异”的参数差异假设，建立图卷积网络的投毒攻击模型. 基于参数差异假设的投毒攻击模型可表示为如下约束优化问题：

$$(\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})=\mathop{\arg\max}_{(\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})}\left\|\boldsymbol{W}^*-\boldsymbol{W}_0\right\|_M^2 \tag{6}$$

$$\text{s.t. } \boldsymbol{W}^*=\mathop{\arg\min}_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W};\hat{\boldsymbol{A}},\hat{\boldsymbol{X}})$$
$$\left\|\hat{\boldsymbol{A}}-\boldsymbol{A}\right\|_0+\left\|\hat{\boldsymbol{X}}-\boldsymbol{X}\right\|_0\leqslant\delta \tag{7}$$

其中，$\boldsymbol{W}_0=\mathop{\arg\min}_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W};\boldsymbol{A},\boldsymbol{X})$ 为图卷积网络的最佳训练参数；$\|\cdot\|_M$ 是矩阵的某种加权范数，用以衡量参数差异，即

$$\left\|\boldsymbol{W}^*-\boldsymbol{W}_0\right\|_M=\sqrt{\mathrm{vec}^{\mathrm{T}}\left(\boldsymbol{W}^*-\boldsymbol{W}_0\right)\boldsymbol{M}\mathrm{vec}\left(\boldsymbol{W}^*-\boldsymbol{W}_0\right)} \tag{8}$$

其中，$\boldsymbol{M}$ 是 n 阶实对称权矩阵；$\boldsymbol{W}^*$ 是邻接矩阵或特征矩阵扰动后图卷积网络的训练参数；vec(·)是矩阵按列优先的拉直运算. 设模型参数采用梯度下降法训练，即

$$\boldsymbol{W}^t=\boldsymbol{W}^{t-1}-\alpha\cdot\nabla_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^{t-1}\right) \tag{9}$$

则经过 $t=t_0$ 轮训练，可得模型的训练参数 $\boldsymbol{W}^*=\boldsymbol{W}^{t_0}$. 具体地，对于式(1)的图卷积网络和式(3)的损失函数，可求得

$$\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})=\left(\hat{\mathcal{A}}\hat{\boldsymbol{X}}\right)^{\mathrm{T}}\left[\sigma\left(\hat{\mathcal{A}}\hat{\boldsymbol{X}}\boldsymbol{W}\right)-\boldsymbol{Y}\right] \tag{10}$$

$\hat{\mathcal{A}}$ 和 $\hat{\boldsymbol{X}}$ 表示扰动后的 $\mathcal{A}$ 和扰动后的 $\boldsymbol{X}$，设 $\hat{\boldsymbol{Z}}=\hat{\mathcal{A}}\hat{\boldsymbol{X}}$，则有

$$\boldsymbol{W}^*=\boldsymbol{W}^0-\alpha\cdot\sum_{t=0}^{t_0-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}\left[\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right] \tag{11}$$

注意，投毒攻击场景下，图卷积网络正向训练时所用的数据为扰动后的邻接矩阵 $\hat{\boldsymbol{A}}$ 和特征矩阵 $\hat{\boldsymbol{X}}$.

3.2 衡量参数差异的Cook距离及其计算方法

设攻击后模型求得的参数 $\boldsymbol{W}$ 的估计值为 $\boldsymbol{W}^*$，原始训练参数为 $\boldsymbol{W}_0$. 首先建立某种“广义距离”对二者差异进行度量. 本文引入统计诊断中的重要成果Cook距离及其近似方法[23~25,27~29]衡量参数差异.

定义1 （Cook距离）参数矩阵 $\boldsymbol{W}^*$ 与 $\boldsymbol{W}_0$ 的Cook距离CD定义为

$$\mathrm{CD}=\mathrm{vec}^{\mathrm{T}}\left[\nabla_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right]\left[\nabla^2_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right]^{K-2}\mathrm{vec}\left[\nabla_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right] \tag{12}$$

其中，$\left[\nabla^2_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right]^{K-2}$ 是权矩阵，K 为整数，通常 K=1,2,3.

注意，若CD严格表示距离，本该满足距离的公理化定义，但这里因循统计诊断之约定，Cook距离仍按式(12)定义. 式(12)中 $\nabla_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)=\boldsymbol{Z}^{\mathrm{T}}\left[\sigma\left(\boldsymbol{Z}\boldsymbol{W}^*\right)-\boldsymbol{Y}\right]$，$\boldsymbol{Z}=\mathcal{A}\boldsymbol{X}$，即区别于式(10)和式(11)，此处代入未扰动的邻接矩阵和特征矩阵. 这也是本文基于参数差异假设建立攻击模型与基于矛盾数据假设建立攻击模型的重要区别. 第4.3节和第4.4节将设计实验验证参数差异假设的合理性和有效性，并与矛盾数据假设做对比分析.

Cook距离CD的具体计算式由定理1给出.

定理1 （**Cook距离的计算公式**）式(12)的Cook距离CD可写成如下形式：

$$\mathrm{CD}=\mathrm{vec}^{\mathrm{T}}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{E}\right)\left[\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}^{\mathrm{T}}\right)\boldsymbol{K}_{mN}\boldsymbol{\Lambda}^*\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}\right)\right]^{K-2}\times\mathrm{vec}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{E}\right) \tag{13}$$

其中，Hessian矩阵

$$\boldsymbol{M}\left(\boldsymbol{W}^*\right)=\nabla^2_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)=\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}^{\mathrm{T}}\right)\boldsymbol{K}_{mN}\boldsymbol{\Lambda}^*\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}\right)$$

$$\boldsymbol{E}=\boldsymbol{Y}-\sigma\left(\boldsymbol{Z}\boldsymbol{W}^*\right)$$

为残差矩阵. “$\otimes$”表示矩阵的Kronecker积，有

$$\underset{m\times n}{\boldsymbol{A}}\otimes\underset{p\times q}{\boldsymbol{B}}=[\boldsymbol{A}_{ij}\boldsymbol{B}]_{mp\times nq} \tag{14}$$

$\boldsymbol{\Lambda}^*$ 为块对角阵，有

$$\underset{Nm\times Nm}{\boldsymbol{\Lambda}^*}=\begin{bmatrix}\underset{m\times m}{\boldsymbol{\lambda}_1}&&\\&\ddots&\\&&\underset{m\times m}{\boldsymbol{\lambda}_N}\end{bmatrix}$$

$$\boldsymbol{\lambda}_i=\frac{\partial\sigma(\xi_i)}{\partial\xi_i} \tag{15}$$

$$\boldsymbol{Z}\boldsymbol{W}^*=\begin{bmatrix}\boldsymbol{\xi}_1^{\mathrm{T}}\\\vdots\\\boldsymbol{\xi}_N^{\mathrm{T}}\end{bmatrix}$$

$$\begin{aligned}\boldsymbol{\lambda}_i &= \frac{\partial \sigma(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i} \\ &= \frac{\partial \mathrm{softmax}(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i} \\ &= \frac{\mathbf{1}^{\mathrm{T}} \exp(\boldsymbol{\xi}_i)\mathrm{diag}\left[\exp(\boldsymbol{\xi}_i)\right] - \exp(\boldsymbol{\xi}_i)\exp^{\mathrm{T}}(\boldsymbol{\xi}_i)}{\left[\mathbf{1}^{\mathrm{T}} \exp(\boldsymbol{\xi}_i)\right]^2}\end{aligned} \tag{16}$$

定理1的证明将在附录中给出. 当K=1时,涉及对Hessian矩阵$\nabla^2_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^*)$求逆,实用中$\nabla^2_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}^*)$通常不可逆,因此$K$=1时仅用于理论分析. 实际计算CD时为避免求逆而取K=2,3,或在具体场景中寻找K=1的等价方法或近似方法. 本文取K=2,3.

3.3　攻击梯度的闭式解

对于邻接矩阵$\boldsymbol{A}=\{0,1\}^{N\times N}$和特征矩阵$\boldsymbol{X}=\{0,1\}^{N\times n}$均为离散值的情况,显然优化问题(式(6)、式(7))属于NP难问题. 可以使用动态规划、贪心算法等求解. 以贪心算法为例,在扰动开销δ的范围内,每增加第i个扰动元素,对特征矩阵$\boldsymbol{X}$的攻击需要计算$Nn-i+1$次扰动对CD的影响,即大约需要计算δNn次CD(对邻接矩阵$\boldsymbol{A}$扰动大约需要计算δN^2次CD). 为减少计算CD的次数,基于参数差异的攻击模型设计了利用CD梯度信息的求解算法. 对于具体的图卷积网络,式(13)建立了CD关于扰动后邻接矩阵$\hat{\boldsymbol{A}}$和扰动后特征矩阵$\hat{\boldsymbol{X}}$(包含在参数$\boldsymbol{W}^*$中)的显式表达式,根据该表达式可以推导出CD对矩阵$\hat{\boldsymbol{A}}$和$\hat{\boldsymbol{X}}$梯度的闭式解. 闭式解利于理论研究、解析图卷积网络的攻击原理,容易对产生攻击的主要影响因素追踪定位和溯源,以避免反向传播等梯度求解方法的黑盒性.

采用基于梯度的攻击方法,首先求解Cook距离CD对扰动后的邻接矩阵和特征矩阵的梯度,可得关于攻击梯度的如下两个定理.

定理2　(**特征矩阵的攻击梯度**)设图卷积网络(式(1))和损失函数(式(3))通过梯度下降法(式(9))训练,将权矩阵$\boldsymbol{M}^K(\boldsymbol{W}^*)$与式(9)中的参数矩阵$\boldsymbol{W}^t(t=0,1,2,\cdots,t_0-1)$视为常数矩阵. 则Cook距离CD对扰动后的特征矩阵$\hat{\boldsymbol{X}}$的梯度可以表示为

$$\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD} = 2\alpha\cdot\hat{\boldsymbol{\mathcal{A}}}\sum_{t=0}^{t_0-1}\left(\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}} - \hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right) \tag{17}$$

其中,$\hat{\boldsymbol{E}}^t=\boldsymbol{Y}-\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)$为梯度下降法训练至第$t$轮(即$\boldsymbol{W}=\boldsymbol{W}^t$)时的残差矩阵,$\boldsymbol{H}_1=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{E}$,即

$$\mathrm{vec}\left(\boldsymbol{H}_K\right) = \boldsymbol{M}\left(\boldsymbol{W}^*\right)\mathrm{vec}\left(\boldsymbol{H}_{K-1}\right) \tag{18}$$

$$\mathrm{vec}\left(\boldsymbol{Q}^{t\mathrm{T}}\right) = \hat{\boldsymbol{\Lambda}}^t\mathrm{vec}\left(\boldsymbol{H}_K^{\mathrm{T}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\right) \tag{19}$$

$\hat{\boldsymbol{\Lambda}}^t$的形式同式(15),只是将其中的$\boldsymbol{Z}\boldsymbol{W}^*$换成$\hat{\boldsymbol{Z}}\boldsymbol{W}^*$.

定理2的证明将在附录中给出.

定理3　(**邻接矩阵的攻击梯度**)设图卷积网络(式(1))和损失函数(式(3))通过梯度下降法(式(9))训练,将权矩阵$\boldsymbol{M}^K(\boldsymbol{W}^*)$、含自连边的度矩阵$\tilde{\boldsymbol{D}}$与式(11)中的参数矩阵$\boldsymbol{W}^t(t=0,1,2,\cdots,t_0-1)$视为常数矩阵. 则Cook距离CD对扰动后的含自连边的邻接矩阵$\hat{\boldsymbol{A}}$的梯度可以表示为

$$\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD} = 2\alpha\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\left[\sum_{i=0}^{k-1}\hat{\boldsymbol{L}}^i\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}} - \hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)\left(\hat{\boldsymbol{L}}^{k-i-1}\hat{\boldsymbol{X}}\right)^{\mathrm{T}}\right]\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tag{20}$$

其中,$\hat{\boldsymbol{L}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$. 特别地,当$k$=1时,有

$$\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD} = 2\alpha\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}} - \hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)\hat{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tag{21}$$

定理3的证明将在附录中给出. 注意,CD的计算式(13)中,$\boldsymbol{Z}=\boldsymbol{\mathcal{A}}\boldsymbol{X}$,其中$\boldsymbol{\mathcal{A}}$和$\boldsymbol{X}$为未扰动的归一化邻接矩阵和特征矩阵而非扰动后的矩阵,因此求导时看作常数矩阵. 扰动信息体现在参数$\boldsymbol{W}^*$的训练过程(式(11))之中. 这也体现了Cook距离CD的物理意义,即扰动后的数据诱导图卷积网络训练出一组异于原参数的参数$\boldsymbol{W}^*$,使新参数$\boldsymbol{W}^*$偏离未扰动的损失函数$\mathcal{L}$的驻点,达到攻击目的.

3.4　基于参数差异假设的投毒攻击算法

根据上述理论,基于梯度下降算法与贪心算法的思想设计基于参数差异假设的图卷积网络投毒攻击算法(Parattack). 攻击邻接矩阵或特征矩阵算法的主要步骤包括:

(1)根据上一轮的扰动列表disturb_list求得扰动后的邻接矩阵$\hat{\boldsymbol{A}}$和特征矩阵$\hat{\boldsymbol{X}}$为$\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$=get_modified($\boldsymbol{A}$,$\boldsymbol{X}$,disturb_list).

(2)图卷积网络进行正向训练,得训练参数$\boldsymbol{W}^*$=train_GNN($\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},\boldsymbol{Y}$).

(3)根据式(17)和式(20)或其等价方法求攻击梯度矩阵$\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD}$,$\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}$为$\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD}$,$\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}$=get_CD_gradient($\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},\boldsymbol{W}^*,\boldsymbol{Y}$).

(4)对攻击梯度矩阵$\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD}$,$\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}$的符号进行处理,保留对扰动筛选有效的攻击梯度,得许用梯度矩阵$\boldsymbol{G}_A,\boldsymbol{G}_X$. $\boldsymbol{G}_A,\boldsymbol{G}_X$=sign_process($\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD},\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD},\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$). 对于扰动后的邻接矩阵$\hat{\boldsymbol{A}}$或特征矩阵$\hat{\boldsymbol{X}}$的元素值为1的位置,当梯度值为负数时有效. 对于元素值为0的位置,当梯度值为正数时有效;为保证无权无向网络邻接矩阵的对称性,再将邻接矩阵的有效梯度转置相加,得$\boldsymbol{G}_A,\boldsymbol{G}_X$. $\boldsymbol{G}_A,\boldsymbol{G}_X$称为许用梯度矩阵. 许用梯度矩阵的具体计算方式如下:首先计算$\boldsymbol{S}_A=(-2\hat{\boldsymbol{A}}+1)\odot\nabla_{\hat{\boldsymbol{A}}}\mathrm{CD}$,$\boldsymbol{S}_X=(-2\hat{\boldsymbol{X}}+1)\odot\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}$;然后将$\boldsymbol{S}_A,\boldsymbol{S}_X$中的正值置为1,负值置

为0,得许用位置矩阵$\boldsymbol{F}_A,\boldsymbol{F}_X$;最后求得许用梯度矩阵$\boldsymbol{G}_A=\boldsymbol{F}_A\odot\boldsymbol{S}_A+(\boldsymbol{F}_A\odot\boldsymbol{S}_A)^{\mathrm{T}},\boldsymbol{G}_X=\boldsymbol{F}_X\odot\boldsymbol{S}_X$.

(5)根据许用梯度矩阵和攻击类型(攻击邻接矩阵、特征矩阵或邻接矩阵和特征矩阵的联合攻击),每轮筛选许用梯度的绝对值最大的$\gamma n_0(0<\gamma<1)$个位置,对$\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$增量式地进行扰动,达到扰动总数$n_0$;达到扰动总数后,继续对扰动列表进行更新,直至满足迭代总次数iters停止. disturb_list=disturb_AX($\boldsymbol{G}_A,\boldsymbol{G}_X,\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},n_0$).

算法伪代码如表算法1所示.

算法1 基于参数差异假设的图卷积网络投毒攻击算法(Parattack)

输入: 邻接矩阵$\boldsymbol{A}$,特征矩阵$\boldsymbol{X}$,标签$\boldsymbol{Y}$,攻击点数n_0,迭代次数iters.

输出: 扰动列表disturb_list.

disturb_list=[];

FOR i in range (iters):

$\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$=get_modified($\boldsymbol{A},\boldsymbol{X}$,disturb_list);

$\boldsymbol{W}^*$=train_GNN($\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},\boldsymbol{Y}$);

$\nabla_{\hat{A}}$CD,$\nabla_{\hat{X}}$CD=get_CD_gradient($\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},\boldsymbol{W}^*,\boldsymbol{Y}$);

$\boldsymbol{G}_A,\boldsymbol{G}_X$=sign_process($\nabla_{\hat{A}}$CD,$\nabla_{\hat{X}}$CD,$\hat{\boldsymbol{A}},\hat{\boldsymbol{X}}$);

disturb_list=disturb_AX($\boldsymbol{G}_A,\boldsymbol{G}_X,\hat{\boldsymbol{A}},\hat{\boldsymbol{X}},n_0$);

RETURN disturb_list

该算法是基于梯度下降算法和贪心算法的思想设计的,当$\gamma=1/n_0$,每次允许不重复地扰动1个元素时,即为贪心算法. 然而,图卷积网络对抗攻击中的优化对象为邻接矩阵和特征矩阵,二者中的元素均为离散的0,1整数变量,与连续取值的情况不同. 从理论上,我们暂未得到算法收敛性的有关结论. 尽管暂未从理论上证明该算法的收敛性,算法仍然具有有效性和可用性. 本文通过实验分析的方法,设计实验说明算法的可用性,即攻击梯度对本文场景的离散数据扰动筛选仍然具有重要的指导意义.

4 实验验证

实验验证部分首先对攻击梯度的闭式解给出算例;其次设计实验说明基于梯度的算法对于图卷积网络对抗攻击场景的可用性;然后针对节点分类任务,设计实验验证参数度量的合理性、参数差异假设的合理性和其导出方法的有效性;最后将本文提出的基于参数差异假设的攻击方法与其他方法进行对比实验. 实验中采用4个数据集,即polblogs[30]数据集、cora数据集、cora_ml[31]数据集和citeseer[32]数据集. 各个数据集的统计特性如表1所示.

表1 实验数据集的统计特性

数据集	节点数	连边数	特征维数	分类数
polblogs	1 222	16 714	1 490	2
cora	2 485	5 069	1 433	7
cora_ml	2 810	7 981	2 879	7
citeseer	2 110	3 668	3 703	6

4.1 攻击梯度闭式解的算例

设邻接矩阵$\boldsymbol{A}$、特征矩阵$\boldsymbol{X}$、初始参数$\boldsymbol{W}^0$和标签矩阵$\boldsymbol{Y}$分别为

$$\boldsymbol{A}=\begin{bmatrix}0&1&1&0\\1&0&1&1\\1&1&0&0\\0&1&0&0\end{bmatrix},\boldsymbol{X}=\begin{bmatrix}1&0&1\\0&1&1\\1&0&1\\1&1&1\end{bmatrix}$$

$$\boldsymbol{W}^0=\begin{bmatrix}1.2&1.9&2.1\\1.8&1.1&1.3\\2.0&2.2&2.1\end{bmatrix},\boldsymbol{Y}=\begin{bmatrix}0&0&1\\1&0&0\\0&0&1\\1&0&0\end{bmatrix}$$

设扰动后的邻接矩阵$\hat{\boldsymbol{A}}$和特征矩阵$\hat{\boldsymbol{X}}$分别为

$$\hat{\boldsymbol{A}}=\begin{bmatrix}0&1&1&1\\1&0&0&1\\1&0&0&0\\1&1&0&0\end{bmatrix},\hat{\boldsymbol{X}}=\begin{bmatrix}1&0&1\\0&1&1\\1&1&1\\1&1&0\end{bmatrix}$$

设图卷积网络(式(1))使用梯度下降法(式(9))训练的迭代次数为$t_0=3$,学习率$\alpha=0.1$. 经过3轮训练得到的参数和矩阵$\boldsymbol{H}_3$为

$$\boldsymbol{W}^1=\begin{bmatrix}1.30&1.79&2.12\\1.90&1.00&1.31\\2.04&2.15&2.11\end{bmatrix}$$

$$\boldsymbol{W}^2=\begin{bmatrix}1.37&1.70&2.13\\1.97&0.92&1.31\\2.08&2.10&2.12\end{bmatrix}$$

$$\boldsymbol{W}^3=\begin{bmatrix}1.43&1.62&2.15\\2.04&0.85&1.31\\2.10&2.07&2.13\end{bmatrix}$$

$$\boldsymbol{H}_3=\begin{bmatrix}1.75&-1.86&0.10\\1.26&-1.31&0.05\\2.48&-2.62&0.14\end{bmatrix}$$

$\hat{\boldsymbol{E}}^t(t=0,1,2)$和$\boldsymbol{E}$矩阵为

$$\hat{\boldsymbol{E}}^0=\begin{bmatrix}-0.25&-0.31&0.56\\0.72&-0.31&-0.41\\-0.28&-0.33&0.61\\0.74&-0.30&-0.44\end{bmatrix}$$

$$\hat{\boldsymbol{E}}^1=\begin{bmatrix}-0.31&-0.24&0.55\\0.67&-0.26&-0.41\\-0.32&-0.29&0.60\\0.68&-0.23&-0.45\end{bmatrix}$$

$$\hat{\boldsymbol{E}}^2=\begin{bmatrix}-0.36&-0.19&0.55\\0.63&-0.22&-0.41\\-0.34&-0.26&0.60\\0.62&-0.30&-0.44\end{bmatrix}$$

$$\boldsymbol{E}=\begin{bmatrix}-0.33 & -0.24 & 0.57\\ 0.63 & -0.20 & -0.43\\ -0.33 & -0.24 & 0.57\\ 0.60 & -0.21 & -0.39\end{bmatrix}$$

$\boldsymbol{Q}^t(t=0,1,2)$矩阵为

$$\boldsymbol{Q}^0=\begin{bmatrix}1.19 & -1.40 & 0.21\\ 0.93 & -1.04 & 0.10\\ 0.70 & -0.79 & 0.09\\ 1.23 & -1.40 & 0.17\end{bmatrix}$$

$$\boldsymbol{Q}^1=\begin{bmatrix}1.29 & -1.22 & -0.06\\ 0.98 & -0.94 & -0.04\\ 0.73 & -0.74 & 0.02\\ 1.33 & -1.21 & -0.11\end{bmatrix}$$

$$\boldsymbol{Q}^2=\begin{bmatrix}1.33 & -1.06 & -0.27\\ 1.00 & -0.86 & -0.15\\ 0.74 & -0.70 & -0.04\\ 1.36 & -1.04 & -0.32\end{bmatrix}$$

根据式(17)和式(20),CD对$\hat{\boldsymbol{X}}$和$\hat{\boldsymbol{A}}$的梯度分别为

$$\nabla_{\hat{X}}\mathrm{CD}=\begin{bmatrix}-1.24 & 0.34 & -1.14\\ -1.02 & 0.19 & -0.97\\ -0.45 & 0.22 & -0.37\\ -1.44 & 0.27 & -1.37\end{bmatrix}$$

$$\nabla_{\hat{A}}\mathrm{CD}=\begin{bmatrix}-0.30 & -0.18 & -0.65 & -0.26\\ -0.98 & -0.51 & -0.88 & -0.92\\ -0.05 & 0.10 & 0.01 & 0.08\\ -1.50 & -0.81 & -1.53 & -1.43\end{bmatrix}$$

采用PyTorch的自动求导功能,可得出与上述基于式(17)和式(20)相同的梯度结果.

4.2　梯度攻击算法的可用性实验

为验证本文场景中,邻接矩阵与特征矩阵为离散取值情况下,基于梯度下降法和贪心算法思想设计的扰动算法的可用性,设计一组离散取值的梯度方法实验. 实验采用式(13)的Cook距离CD(K=2). 为验证"在一定条件下,基于梯度的算法对离散数据同样有效"这一结论,本组实验将参数矩阵$\boldsymbol{W}^*$视为常数矩阵,以排除图卷积网络重训练环节对梯度有效性的影响. 实验中直接求解CD对邻接矩阵$\boldsymbol{A}$的梯度,根据梯度进行扰动,并与遍历方法比较,观察CD值的变化趋势与收敛情况.

实验1　对数据集中图的连边逐一扰动,并计算Cook距离CD与初始CD之差ΔCD,按扰动量从大到小排序;再计算各个扰动连边对应梯度的平均值,将二者绘制于同一图中,如图1所示. 再根据算法中的步骤4,由梯度得出使CD值增大的许用梯度,与对应的值的增量,按照ΔCD的降序绘制于同一图中,如图2所示. 图中所示ΔCD与梯度均除以对应的最大值做了归一化处理. 本实验基于polblogs数据集完成.

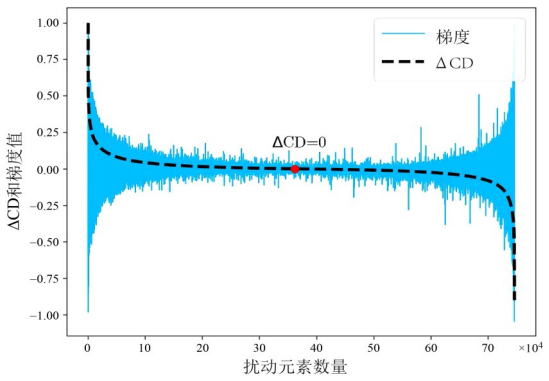


图1　梯度与ΔCD的关系图

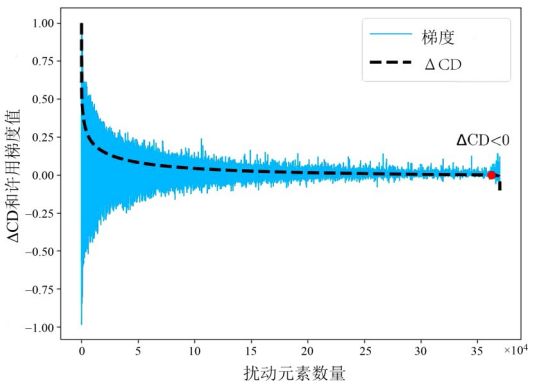


图2　许用梯度与ΔCD的关系图

从图1与图2可以看出,扰动一条连边后,CD的实际增幅与梯度之间有强相关关系:对梯度的绝对值大的位置扰动,CD的实际增幅大;对梯度的绝对值小的位置扰动,CD的实际增幅也小. 计算图2中扰动量ΔCD排序前n的ΔCD值与对应许用梯度绝对值的Cosine相关系数,如表2所示. 可见,CD的实际增量与梯度高度相关,且增幅的排序越靠前(扰动增量越大或梯度绝对值越大),相关性越强. 由此得出结论:梯度信息可为对抗攻击场景的离散数据扰动提供重要指导.

表2　ΔCD与对应许用梯度绝对值的Cosine相关系数

n	10	50	100	200	500
相关系数	0.992	0.992	0.989	0.984	0.971
n	1 000	2000	5 000	10 000	100 000
相关系数	0.971	0.972	0.972	0.969	0.965

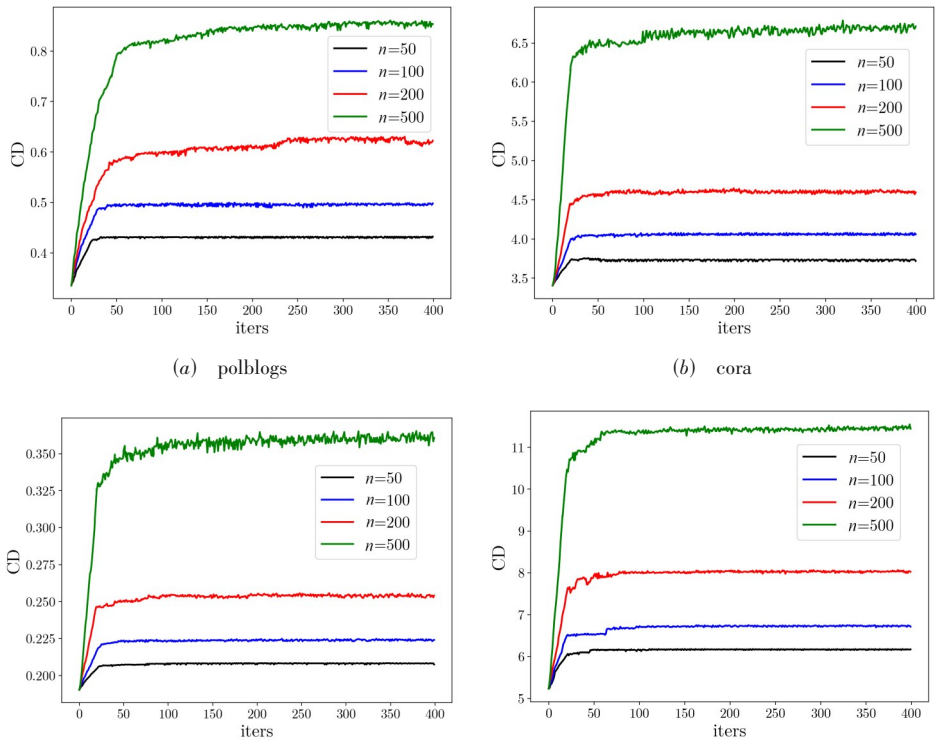
由于图2所示的是理论上使CD增长的许用梯度,而红色点后面的数据表示ΔCD<0,且图2尾部存在略微的凸起,这部分的数据表明许用梯度标致的扰动方向与CD的实际增减方向不一致. 从绝对数量上来看,这些数据占总数的2.016%. 事实上,ΔCD<−0.05的只有35个,仅占0.009 5%,而绝大多数ΔCD<0即许用梯度产生的错误梯度,为梯度的绝对值极小时,梯度信息已经失去意义而带来的误差. 考虑到对抗攻击的实际

场景，实际扰动元素个数不足元素总数的0.1%（连边总数的10%以内），即仅需使用许用梯度绝对值最大的前0.1%的梯度，许用梯度的上述错误对实际攻击场景无任何影响.

实验2 基于梯度下降法与贪心算法的思想，每轮选择许用梯度中梯度的绝对值最大的γn_0条连边对邻接矩阵$\boldsymbol{A}$增量式地进行扰动，达到扰动总数n_0实验中设置$n_0=50,100,200,500$，γ设置为0.04~0.07的随机数). 达到扰动总数后，继续对扰动列表进行更新，直至满足迭代总次数iters停止（实验中设置iters=400). 记录迭代轮数对应的CD值，并绘制曲线如图3所示.

从实验结果图中可以看出，在4个数据集中，对于不同的扰动连边数，算法均有收敛的趋势. 在一定轮数后，CD值在某一范围内小幅度波动，具有可用性.

从以上2组实验结果来看，实验1表明梯度信息对离散数据扰动有重要理论指导，实验2表明算法有收敛的趋势. 实验结果支持了“对于图卷积网络中使用的离散型0,1数据，当确定损失函数后，基于梯度的攻击方法对损失函数的增长具有有效性和可用性”这一结论. 由此进一步增加了基于梯度的攻击方法在实用中的说服力. 然而究竟在何种条件下能够充分保证这种有效性，以及基于梯度的离散算法在使用时有什么制约条件，需要进一步的理论证明. 上述实验结果也对理论研究的可行性提供了证据支持.



(a) polblogs (b) cora

(c) cora_ml (d) citeseer

图3 算法1在4个数据集上的收敛性曲线

4.3 参数差异度量的合理性实验

采用式(13)的Cook距离CD(K=2,K=3)衡量参数差异. 实验中随机生成参数初始值并代入图卷积网络(式(1))训练，迭代次数iters=100. 每隔20轮记录1次训练参数，计算CD值，实验结果如表3所示. 每组实验CD的最小值用黑体标出.

根据理论分析，随着训练轮数的增加，训练参数应该与最佳参数$\boldsymbol{W}_0$逐渐靠近，即与最佳参数$\boldsymbol{W}_0$的距离应该逐渐减小，从而表现为CD值递减. 表中4个数据集在不同K值的实验结果显示，随着训练轮数增加，CD值有减小的趋势，与理论分析一致.

4.4 参数差异假设的合理性实验

本文基于参数差异假设建立攻击模型(式(6)和式(7))，并将现有的其他攻击方法概括为基于矛盾数据假设的攻击方法. 为验证参数差异假设的合理性和由其导出方法的有效性，设计以下实验对参数差异假设和矛盾数据假设做对比分析.

实验1 实验中采用本文建立的Cook距离CD(K=2,3)作为基于参数差异假设的攻击方法Parattack. 为了与基于矛盾数据假设的攻击方法对比，控制Cook距离表达式的形式不变，基于矛盾数据假设，将表达式

表3　参数差异的度量结果

		0	20	40	60	80	100
polblogs	K=2	348.69	301.92	262.28	228.88	200.81	**177.18**
	K=3	110.59	95.12	82.02	71.03	61.82	**54.10**
cora	K=2	635 230.06	3 629.67	1 272.43	695.66	453.81	**321.55**
	K=3	24 486 698.00	7 124.50	1 825.30	796.38	424.10	**251.75**
cora_ml	K=2	20 759.96	6 339.43	2 835.57	1 599.68	1 035.22	**732.67**
	K=3	33 737.08	6 339.66	2 616.48	1 393.28	850.18	**567.34**
citeseer	K=2	1 063 964.12	4 593.94	1 904.96	1 034.97	607.26	**389.56**
	K=3	78 890 432.00	2 192 633.00	5 563.29	2 226.66	835.56	**385.31**

(13)中$\boldsymbol{Z}=\boldsymbol{\mathcal{A}X}$更换为$\hat{\boldsymbol{Z}}=\hat{\boldsymbol{\mathcal{A}}}\hat{\boldsymbol{X}}$,记为$\widehat{\mathrm{CD}}(K=2,3)$. 实验中划分40%训练集和60%的测试集,$k$=1和$k$=2时,允许对现有连边总数的10%进行攻击. 参数γ取0.03~0.05的随机数,算法1的迭代次数iters=30. 重复10次实验,记录10次攻击结果准确率下降的平均值,得到表4的实验数据.

从表4可以看出,k=2相较于k=1时攻击效果提升明显,即图卷积网络的攻击难度下降. 控制Cook距离CD的输入数据为唯一变量,从整体上该场景中基于参数差异假设的Cook距离CD的攻击效果好于矛盾数据假设修改的$\widehat{\mathrm{CD}}$,部分结果准确率下降幅度成倍数增长. 在部分取值如k=1时,cora和citeseer数据集,基于参数差异假设的攻击效果相比基于矛盾数据假设的攻击效果提升并不显著. 这提示我们:矛盾数据假设仍是对抗攻击本质的一种重要概括;矛盾数据假设是参数差异假设的一种重要实现形式. 在该实验场景下,基于矛盾数据假设导出的方法,实质上也是诱导图卷积网络训练出错误参数,这一观点在文献[22]中同样得到了印证.

表4　参数差异假设与矛盾数据假设导出方法的攻击结果(k=1,2)

		polblogs	cora_ml	cora	citeseer
k=1	CD (K=2)	6.11%	17.33%	21.81%	15.55%
	$\widehat{\mathrm{CD}}$(K=2)	3.21%	9.52%	20.96%	14.31%
	CD (K=3)	5.98%	23.14%	21.34%	15.50%
	$\widehat{\mathrm{CD}}$(K=3)	3.46%	6.59%	20.29%	14.29%
k=2	CD (K=2)	14.21%	67.76%	65.30%	51.24%
	$\widehat{\mathrm{CD}}$(K=2)	5.94%	64.51%	64.04%	48.25%
	CD (K=3)	17.26%	70.16%	69.29%	51.00%
	$\widehat{\mathrm{CD}}$(K=3)	5.63%	62.96%	68.24%	46.38%

实验2　设计实验对比分析基于矛盾数据假设的攻击方法Mettack和基于参数差异假设的攻击方法Parattack(K=2,3),以验证参数差异假设的合理性. 在不同扰动量(1%~5%)下,每种攻击方法均计算自身的损失函数和其他2种方法的损失函数,共3种损失函数值,即计算式(3)的$\mathcal{L}$和式(13)的CD(K=2)和CD(K=3). 实验设置与实验1相同,所有数值均为10次实验的平均值. 选取Cora数据集k=1和k=2的实验结果如表5和表6所示. 其他数据集和不同参数下的实验结果与所列结果相似,实验结论相同,限于篇幅不再列出.

观察表5和表6的实验结果. 从横向看,无论是基于矛盾数据假设的攻击方法Mettack,抑或基于参数差异假设的Parattack(K=2,3),随着扰动比例的增加,衡量参数差异的CD值(K=2,3)均呈稳定增大趋势. 这表明Mettack方法与Parattack攻击方法一样,事实上也是诱导图神经网络训练出异于原始参数的训练参数. 从纵向看,相同扰动量,Parattack对应的3种损失函数均大于Mettack的3种损失函数,说明了基于参数差异假设度量方法的有效性. 值得注意的是,如表5所示,Mettack方法在k=1时损失值$\mathcal{L}$并未稳定增加,而存在波动. 原因可作以下分析. 损失函数$\mathcal{L}$中有2个变量,即扰动后的邻接矩阵$\hat{\boldsymbol{A}}$和重训练参数$\boldsymbol{W}^*$. 由3.1节可知,对抗攻击分为两个阶段:一是根据式(6)实施投毒攻击;二是根据式(7)进行对抗训练. 投毒攻击阶段基于固定的$\boldsymbol{W}^*$扰动$\hat{\boldsymbol{A}}$使$\mathcal{L}$增大;而对抗训练阶段固定$\hat{\boldsymbol{A}}$训练$\boldsymbol{W}^*$使$\mathcal{L}$减小. 最终可能产生两种结果:若对抗攻击阶段占优,则最终损失函数$\mathcal{L}$增加;若对抗训练阶段占优,则最终损失函数$\mathcal{L}$减小. 两种结果均可出现,这与表5的Mettack损失函数$\mathcal{L}$存在波动的实验结果一致. 与损失函数$\mathcal{L}$相比,Cook距离CD中只有一个变量$\boldsymbol{W}^*$,大幅减弱了该种不一致性而提高攻击损失的灵敏度. 对于k=2时,表6的结果显示的损失函数均稳定增加,对于损失函数$\mathcal{L}$表明其对抗攻击阶段占优,这与实验1得出的k=2时图卷积网络的攻击难度下降的结论一致.

以上实验表明,参数差异假设具有合理性和较为一般的适用性.

4.5　与其他攻击方法的对比实验

本节将本文提出的基于参数差异假设的攻击方法Parattack(K=2,3)与基准方法进行对比实验. 基准方法包括随机攻击方法Random和基于矛盾数据假设的攻

表5 参数差异假设的合理性验证(Cora数据集, SGC模型取k=1)

方法		未扰动	1%	2%	3%	4%	5%
Mettack	$\mathcal{L}$	44.19	74.09	67.75	59.41	53.27	46.71
	CD (K=2)	9.81	303.06	794.30	1 571.02	3 135.19	5 625.58
	CD (K=3)	2.28	3 865.35	15 574.70	37 658.80	89 857.40	178 598.00
Parattack (K=2)	$\mathcal{L}$	44.01	90.22	98.59	103.03	102.39	99.31
	CD (K=2)	9.75	402.82	1 754.35	4 355.42	6 491.25	8 567.13
	CD (K=3)	2.26	11 776.00	93 541.30	346 686.00	463 502.00	583 897.00
Parattack (K=3)	$\mathcal{L}$	44.19	88.99	95.03	100.08	109.78	100.64
	CD (K=2)	9.80	312.38	1 314.19	3 230.10	8 447.78	7 479.64
	CD (K=3)	2.27	9 956.53	69 866.20	218 892.00	745 062.00	572 026.00

表6 参数差异假设的合理性验证(Cora数据集, SGC模型取k=2)

方法		未扰动	1%	2%	3%	4%	5%
Mettack	$\mathcal{L}$	88.50	132.89	1 090.86	9 216.20	54 523.90	144 609.00
	CD (K=2)	13.97	243.53	3 854.18	13 369.40	27 487.30	43 982.20
	CD (K=3)	2.59	5 265.44	103 406.00	256 005.00	306 480.00	441 769.00
Parattack (K=2)	$\mathcal{L}$	88.08	212.14	4 489.40	36 598.90	138 945.00	381 598.00
	CD (K=2)	13.86	675.67	9 962.60	26 537.40	73 422.50	112 915.00
	CD (K=3)	2.57	26 488.00	208 945.00	439 317.00	940 865.00	1 004 410.00
Parattack (K=3)	$\mathcal{L}$	88.29	200.95	5 794.97	44 233.90	147 960.00	425 041.00
	CD (K=2)	13.92	1 175.71	11 471.30	36 288.10	50 882.00	141 631.00
	CD (K=3)	2.58	55 506.60	244 881.00	645 585.00	536 679.00	1 239 370.00

击方法DICE,Mettack. 下面分别对基准方法简要介绍.

(1)随机攻击方法Random. 该方法从训练集中随机选择连边增加或删除.

(2)“删除同类连接异类”攻击方法DICE(Delete Internally Connect Externally). 该方法根据“从同类节点中删除连边,在不同类节点间增加连边”这一规则,随机增删连边. 根据本文观点,该方法属于基于矛盾数据假设的攻击方法.

(3)基于元学习的攻击方法Mettack. 该攻击方法的损失函数与训练图卷积网的损失函数相同,损失函数代入扰动后的投毒训练数据,可由矛盾数据假设导出;然后根据损失函数得攻击梯度,并采用贪心算法进行扰动筛选.

实验将数据集随机地划分为40%的训练集和60%的测试集,允许使用3%的连边数进行攻击. 根据攻击梯度,Mettack方法与Parattack方法均使用离散梯度下降算法,参数γ取0.03~0.05的随机数,算法的迭代次数iters=50. 对邻接矩阵中的连边进行扰动,重复10次实验,记录10次攻击结果准确率下降的平均值,得到表7的实验数据. 同时表8列出了不同k值时,图卷积网络SGC在各个未受到对抗攻击污染的数据集上的准确率. 实验结果表明,随机攻击方法Random和基于规则的DICE几乎无法实施有效的数据投毒攻击,即对重训练的SGC几乎没有攻击效果;部分实验结果为负值,攻击后准确率不减反增,但多为随机波动. 原因是该两种方法或随机选取连边,或基于规则随机增删连边,未进

表7 Parattack(CD)与其他方法的攻击结果

		polblogs	cora_ml	cora	citeseer
k=1	Random	0.07%	−0.16%	0.49%	−0.18%
	DICE	0.06%	0.07%	0.09%	−0.13%
	Mettack	3.99%	2.46%	12.85%	4.83%
	CD(K=2)	**4.36%**	6.26%	**15.62%**	**6.57%**
	CD(K=3)	3.80%	**6.79%**	11.13%	5.35%
k=2	Random	−0.11%	0.03%	0.13%	4.07%
	DICE	0.25%	0.05%	−0.45%	0.06%
	Mettack	3.67%	40.75%	46.00%	23.31%
	CD(K=2)	3.12%	**52.50%**	**46.48%**	**29.26%**
	CD(K=3)	**4.40%**	48.69%	43.04%	29.19%
k=3	Random	−0.15%	−0.02%	0.15%	0.33%
	DICE	0.31%	0.03%	0.01%	2.77%
	Mettack	30.83%	53.96%	58.66%	27.22%
	CD(K=2)	**42.16%**	**56.99%**	63.59%	**52.82%**
	CD(K=3)	30.07%	56.74%	**64.29%**	49.93%

表8 图卷积网络SGC在未受攻击的数据集上的准确率

	polblogs	cora_ml	cora	citeseer
k=1	91.84%	86.69%	85.20%	76.44%
k=2	94.88%	83.36%	86.13%	77.82%
k=3	94.79%	82.77%	86.19%	77.56%

行重训练. 缺少针对性的训练环节是导致攻击无效的主要原因. 对于具备重训练环节的 Mettack 和 Parattack,实验结果表明,基于参数差异假设的攻击方法 Parattack 相比基于矛盾数据假设的攻击方法 Mettack 更具有效性. 大部分攻击效果都有较大幅度提升,从而验证了基于参数差异假设导出方法的有效性. 仅攻击特征以及结构和特征联合的攻击方法,与上述方法完全一致,实验结果大致相仿,不再赘述.

5　结论

本文基于参数差异假设,将统计诊断的重要研究成果 Cook 距离引入图卷积网络对抗性攻击,用以衡量模型的参数差异. 本文主要得出了以下结论:

一是对于邻接矩阵为 0,1 的离散数据场景,基于梯度的攻击方法具有有效性和可用性. 基于攻击梯度的离散数据实验结果支持了这一结论. 实验结果也为算法的收敛性等理论研究的可行性提供了证据支持.

二是推导得出 Cook 距离的计算式和基于 Cook 距离的攻击梯度闭式解. 攻击梯度的闭式解为进一步寻找攻击有效性的本质和简化攻击方法提供了必要前提.

三是设计了矛盾数据假设和参数差异假设的对比实验,说明参数差异假设的合理性. 基于参数差异假设的攻击方法与其他方法的对比实验结果说明了参数差异假设导出方法的有效性.

图卷积网络乃至图神经网络对抗性攻击尤其是对抗性攻击存在性本质,仍具有丰富而深刻的问题有待进一步研究. 例如,参数差异假设是否包含矛盾数据假设,矛盾数据假设是否必然蕴含参数差异假设,是否存在仅满足矛盾数据假设的对抗性攻击而不满足参数差异假设,或是否存在满足参数差异假设的数据样本而不属于对抗性攻击样本,矛盾数据假设与参数差异假设何时等价,对抗性攻击问题本质能否更紧凑地定量表达,等等,需要进一步的理论和实验研究.

附录

定理 1 的证明.

$$\mathrm{d}\left(\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})\right)=\boldsymbol{Z}^{\mathrm{T}}\mathrm{d}[\sigma(\boldsymbol{ZW})]\boldsymbol{I}_m$$

$$\begin{aligned}\mathrm{vec}\left[\mathrm{d}(\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}))\right]&=\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}^{\mathrm{T}}\right)\mathrm{vec}\left[\mathrm{d}\sigma(\boldsymbol{ZW})\right]\\&=\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}^{\mathrm{T}}\right)\frac{\partial\sigma^{\mathrm{T}}(\boldsymbol{ZW})}{\partial(\boldsymbol{ZW})}\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}\right)\mathrm{vec}(\mathrm{d}\boldsymbol{W})\end{aligned}$$

代入 Cook 距离的定义式(12)可得定理 1.

证毕

定理 2 的证明.

设 $v\left(\boldsymbol{W}^*\right)=\mathrm{vec}\left[\nabla_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right]$,则

$$\mathrm{vec}\left(\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}\right)=\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{X}}}\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$$

其中,$\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{X}}}$,$\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}$,$\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}$,$\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$,$\nabla_{\hat{\boldsymbol{X}}}\mathrm{CD}$ 按如下定义,设矩阵 $\underset{M\times m}{\boldsymbol{F}}$,$\underset{N\times n}{\boldsymbol{X}}$,有

$$\underset{Nn\times Mm}{\frac{\partial\boldsymbol{F}}{\partial\boldsymbol{X}}}=\frac{\partial\mathrm{vec}(\boldsymbol{F})}{\partial\mathrm{vec}(\boldsymbol{X})},\ \underset{N\times n}{\nabla_{\boldsymbol{X}}\mathrm{CD}}=\left[\frac{\partial\mathrm{CD}}{\partial\boldsymbol{X}_{ij}}\right]$$

(1)求 $\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$

$$\begin{aligned}\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}&=2\left(\nabla^2_{\boldsymbol{W}}\mathcal{L}\left(\boldsymbol{W}^*\right)\right)^{K-2}v\left(\boldsymbol{W}^*\right)\\&=-2\boldsymbol{M}^{K-2}\left(\boldsymbol{W}^*\right)\mathrm{vec}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{E}\right)\end{aligned}$$

(2)求 $\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}$

$$\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}=\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}^{\mathrm{T}}\right)\boldsymbol{K}_{mN}\boldsymbol{\Lambda}^*\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\boldsymbol{Z}\right)=\boldsymbol{M}\left(\boldsymbol{W}^*\right)$$

$$\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}=-2\boldsymbol{M}^{K-1}\left(\boldsymbol{W}^*\right)\mathrm{vec}\left(\boldsymbol{H}_1\right)=-2\mathrm{vec}\left(\boldsymbol{H}_K\right)$$

(3)求 $\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}$

$$\begin{aligned}&\mathrm{dvec}\left\{\hat{\boldsymbol{Z}}^{\mathrm{T}}\left[\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right]\right\}\\&=\mathrm{d}\left\{\left(\boldsymbol{I}_m\otimes\hat{\boldsymbol{Z}}^{\mathrm{T}}\right)\mathrm{vec}\left[\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right]\right\}\\&=\mathrm{d}v_1\left(\hat{\boldsymbol{Z}}\right)+\mathrm{d}v_2\left(\hat{\boldsymbol{Z}}\right)\end{aligned}$$

$$\mathrm{d}v_1\left(\hat{\boldsymbol{Z}}\right)=\boldsymbol{K}_{mn}\left\{\boldsymbol{I}_n\otimes\left[\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right]^{\mathrm{T}}\right\}\mathrm{vec}\left(\mathrm{d}\hat{\boldsymbol{Z}}\right)$$

$$\mathrm{d}v_2\left(\hat{\boldsymbol{Z}}\right)=\left(\boldsymbol{I}_m\otimes\hat{\boldsymbol{Z}}^{\mathrm{T}}\right)\boldsymbol{K}_{mN}\hat{\boldsymbol{\Lambda}}\boldsymbol{K}_{Nm}\left[\left(\boldsymbol{W}^t\right)^{\mathrm{T}}\otimes\boldsymbol{I}_N\right]\mathrm{vec}\left(\mathrm{d}\hat{\boldsymbol{Z}}\right)$$

$$\begin{aligned}&\frac{\partial\mathrm{vec}\left\{\hat{\boldsymbol{Z}}^{\mathrm{T}}\left[\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right]\right\}}{\partial\hat{\boldsymbol{Z}}}\\&=\left[\boldsymbol{I}_n\otimes\left(\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right)\right]\boldsymbol{K}_{nm}\\&\quad+\left(\boldsymbol{W}^t\otimes\boldsymbol{I}_N\right)\boldsymbol{K}_{mN}\hat{\boldsymbol{\Lambda}}\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\hat{\boldsymbol{Z}}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}&=\frac{-\alpha\sum_{t=0}^{t_0-1}\partial\hat{\boldsymbol{Z}}^{\mathrm{T}}\left[\sigma(\hat{\boldsymbol{Z}}\boldsymbol{W}^t)-\boldsymbol{Y}\right]}{\partial\hat{\boldsymbol{Z}}}\\&=-\alpha\sum_{t=0}^{t_0-1}\left[\boldsymbol{I}_n\otimes\left(\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right)\right]\boldsymbol{K}_{nm}\\&\quad+\left(\boldsymbol{W}^t\otimes\boldsymbol{I}_N\right)\boldsymbol{K}_{mN}\hat{\boldsymbol{\Lambda}}^t\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\hat{\boldsymbol{Z}}\right)\end{aligned}$$

$$\frac{\partial \boldsymbol{W}^*}{\partial \hat{\boldsymbol{Z}}}\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial \boldsymbol{W}^*}\frac{\partial \mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$$

$$=2\alpha\left\{\begin{array}{l}\sum_{t=0}^{t_0-1}\left[\boldsymbol{I}_n\otimes\left(\sigma\left(\hat{\boldsymbol{Z}}\boldsymbol{W}^t\right)-\boldsymbol{Y}\right)\right]\boldsymbol{K}_{nm}\\+\left(\boldsymbol{W}^t\otimes\boldsymbol{I}_N\right)\boldsymbol{K}_{mN}\hat{\boldsymbol{\Lambda}}^t\boldsymbol{K}_{Nm}\left(\boldsymbol{I}_m\otimes\hat{\boldsymbol{Z}}\right)\end{array}\right\}\mathrm{vec}\left(\boldsymbol{H}_K\right)$$

$$=2\alpha\mathrm{vec}\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)$$

其中，$\mathrm{vec}\left(\boldsymbol{Q}^{t\mathrm{T}}\right)=\hat{\boldsymbol{\Lambda}}^t\mathrm{vec}\left(\boldsymbol{H}_K^{\mathrm{T}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\right)$.

(4) 求 $\dfrac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{X}}}$

$$\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{X}}}=\boldsymbol{I}_n\otimes\boldsymbol{\mathcal{A}}$$

$$\mathrm{vec}\left(\nabla_{\hat{X}}\mathrm{CD}\right)=\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{X}}}\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$$

$$=2\alpha\left(\boldsymbol{I}_n\otimes\hat{\boldsymbol{\mathcal{A}}}\right)\mathrm{vec}\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)$$

$$\nabla_{\hat{X}}\mathrm{CD}=2\alpha\cdot\hat{\boldsymbol{\mathcal{A}}}\sum_{t=0}^{t_0-1}\left(\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)$$

证毕

定理3的证明.

用归纳法容易证明，有

$$\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{A}}}=\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\otimes\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\right)\left[\sum_{i=0}^{k-1}\left(\hat{\boldsymbol{L}}^{k-i-1}\hat{\boldsymbol{X}}\right)\otimes\hat{\boldsymbol{L}}^i\right]$$

$$\mathrm{vec}\left(\nabla_{\hat{A}}\mathrm{CD}\right)=\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{A}}}\frac{\partial\boldsymbol{W}^*}{\partial\hat{\boldsymbol{Z}}}\frac{\partial v\left(\boldsymbol{W}^*\right)}{\partial\boldsymbol{W}^*}\frac{\partial\mathrm{CD}}{\partial v\left(\boldsymbol{W}^*\right)}$$

$$=2\alpha\frac{\partial\hat{\boldsymbol{Z}}}{\partial\hat{\boldsymbol{A}}}\mathrm{vec}\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)$$

$$=2\alpha\mathrm{vec}\left\{\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\left[\sum_{i=0}^{k-1}\hat{\boldsymbol{L}}^i\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)\left(\hat{\boldsymbol{L}}^{k-i-1}\hat{\boldsymbol{X}}\right)^{\mathrm{T}}\right]\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\right\}$$

$$\nabla_{\hat{A}}\mathrm{CD}=2\alpha\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\left[\sum_{i=0}^{k-1}\hat{\boldsymbol{L}}^i\left(\sum_{t=0}^{t_0-1}\boldsymbol{Q}^t\boldsymbol{W}^{t\mathrm{T}}-\hat{\boldsymbol{E}}^t\boldsymbol{H}_K^{\mathrm{T}}\right)\left(\hat{\boldsymbol{L}}^{k-i-1}\hat{\boldsymbol{X}}\right)^{\mathrm{T}}\right]\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$$

证毕

参考文献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2013) [2021]. https://arxiv.org/abs/1312.

[2] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2016) [2021]. https://arxiv.org/abs/1607.02533v1.

[3] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]//Advances in Neural Information Processing Systems. Vancouver: NIPS, 2019: 125-136.

[4] YUAN X Y, HE P, ZHU Q L, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30 (9): 2805-2824.

[5] COSTA L DA F, RODRIGUES F A, TRAVIESO G, et al. Characterization of complex networks: A survey of measurements[J]. Advances in Physics, 2007, 56(1): 167-242.

[6] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017)[2021]. https://arxiv.org/abs/1609.02907v4.

[7] WU F, ZHANG T, SOUZA JR A H d, et al. Simplifying graph convolutional networks[EB/OL]. (2019) [2021]. https://arxiv.org/abs/1902.07153v1.

[8] LI Q M, WU X M, LIU H, et al. Label efficient semi-supervised learning *via* graph filtering[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 9574-9583.

[9] NT H, MAEHARA T. Revisiting graph neural networks: All we have is low-pass filters[EB/OL]. (2019) [2021]. https://arxiv.org/abs/1905.09550v1.

[10] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[11] WU Z H, PAN S R, CHEN F W, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[12] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755-780.
XU B B, CEN K T, HUANG J J, et al. A survey on graph convolutional neural network[J]. Chinese Journal of Computers, 2020, 43(5): 755-780. (in Chinese)

[13] 白铂, 刘玉婷, 马驰骋, 等. 图神经网络[J]. 中国科学: 数学, 2020, 50(3): 367-384.
BAI B, LIU Y T, MA C C, et al. Graph neural network [J]. Scientia Sinica(Mathematica), 2020, 50(3): 367-384. (in Chinese)

[14] ZÜGNER D, AKBARNEJAD A, GÜNNEMANN S. Ad-

versarial attacks on neural networks for graph data[C]// KDD' 18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 2847-2856.

[15] ENTEZARI N, AL-SAYOURI S A, DARVISHZADEH A, et al. All You need is low (rank): Defending against adversarial attacks on graphs[C]//WSDM´20: Proceedings of the 13th International Conference on Web Search and Data Mining. Houston: ACM, 2020: 169-177.

[16] LI J, ZHANG H L, HAN Z C, et al. Adversarial attack on community detection by hiding individuals[C]//WWW' 20: Proceedings of The Web Conference 2020. Taipei: ACM, 2020: 917-927.

[17] WU Y T, LIU W, HU X B, et al. Parameter discrepancy hypothesis: Adversarial attack for graph data[J]. Information Sciences, 2021, 577: 234-244.

[18] CHEN L, LI J T, PENG JY, et al. A survey of adversarial learning on graphs[EB/OL]. (2020) [2021]. https://arxiv.org/abs/2003.05730.

[19] JIN W, LI Y, XU H, et al. Adversarial attacks and defenses on graphs: A review and empirical study[EB/OL]. (2020)[2021]. https://arxiv.org/abs/2003.00653v2.

[20] LI Y, JIN W, XU H, et al. DeepRobust: A PyTorch library for adversarial attacks and defenses[EB/OL]. (2020) [2021]. https://arxiv.org/abs/2005.06149.

[21] BOJCHEVSKI A, GÜNNEMANN S. Adversarial attacks on node embeddings via graph poisoning[C]//International Conference on Machine Learning. Long Beach: PMLR, 2019: 695-704.

[22] ZÜGNER D, GÜNNEMANN S. Adversarial attacks on graph neural networks via meta learning[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019: 1112.

[23] COOK R D. Detection of influential observation in linear regression[J]. Technometrics, 1977, 19(1): 15-18.

[24] COOK R D. Influential observations in linear regression [J]. Journal of the American Statistical Association, 1979, 74(365): 169-174.

[25] COOK R D, WEISBERG S. Residuals and Influence in Regression[M]. New York: Chapman and Hall, 1982.

[26] WEI B C, SHIH J Q. On statistical models for regression diagnostics[J].Annals of the Institute of Statistical Mathematics, 1994, 46(2): 267-278.

[27] 费宇, 陈飞, 喻达磊. 线性和广义线性混合模型及其统计诊断[M]. 北京: 科学出版社, 2013.

[28] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京: 高等教育出版社, 2009.
WEI B C, LIN J G, XIE F C. Statistical Diagnostics[M]. Beijing: Higher Education Press, 2009. (in Chinese)

[29] 韦博成, 鲁国斌, 史建清. 统计诊断引论[M]. 南京: 东南大学出版社, 1991.

[30] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93.

[31] MCCALLUM A K, NIGAM K, RENNIE J, et al. Automating the construction of Internet portals with machine learning[J]. Information Retrieval, 2000, 3(2): 127-163.

[32] ADAMIC L A, GLANCE N. The political blogosphere and the 2004 US election: divided they blog[C]//Proceedings of the 3rd International Workshop on Link Discovery. New York: ACM, 2005: 36-43.

作者简介

吴翼腾　男,1992年出生,吉林省吉林人. 博士,工程师. 主要研究方向为网络空间安全、人工智能安全、对抗机器学习.
E-mail: wuyiteng1992@163.com

刘　伟　男,1992年出生,河北人保定人. 硕士,工程师. 主要研究方向为人工智能安全、自然语言处理.

于溆乔　女,1998年生,河南郑州人. 硕士. 主要研究方向为数据科学.