https://doi.org/10.3799/dqkx.2021.232

一种基于图神经网络的地质钻孔数据保护方案

尚 浩[1],朱恒华[1,2],李 双[1],宋晓媚[3],夏 雨[4],刘 惠[4],杨 帆[4]

1. 山东省地质调查院,山东济南 250014
2. 中国地质大学环境学院,湖北武汉 430080
3. 山东省自然资源资料档案馆,山东济南 250014
4. 中国地质大学计算机学院,湖北武汉 430080

摘 要: 随着深度学习技术的日益成熟,攻击者可以对公开的地质钻孔数据通过分类、预测等方法获取潜在的敏感信息,从而造成重要地质数据的泄露.针对上述问题,提出了一种基于图对抗攻击的地质钻孔数据保护模型*Gcntack*.一方面,基于地质数据拓扑图的度特征,产生满足同一幂律分布的攻击作为微小节点扰动,确保对抗性攻击不易被发现,同时改变了目标节点的分类结果.另一方面,引入注意力机制,使用基于可解释性的图注意力网络模型分析影响对抗攻击结果的关键节点特性,验证*Gcntack*模型中选取对抗性节点的合理性.最后,通过在基准数据集和地质钻孔数据集进行的综合实验和分析,证实了提出的地质钻孔数据保护方案能够基于较少的图结构或节点特征的对抗扰动,达到保护重要地质钻孔数据的目的.

关键词: 图卷积神经网络;图注意力网络;图对抗攻击;可解释性;地质钻孔数据保护;深度学习.

中图分类号: P628 **文章编号:** 1000-2383(2023)08-3151-11 **收稿日期:** 2021-11-23

A Geological Borehole Data Protection Based on Graph Neural Networks

Shang Hao[1], Zhu Henghua[1,2], Li Shuang[1], Song Xiaomei[3], Xia Yu[4], Liu Hui[4], Yang Fan[4]

1. *Shandong Institute of Geological Survey*, *Jinan* 250014, *China*
2. *School of Environmental Studies*, *China University of Geosciences*, *Wuhan* 430074, *China*
3. *Shandong Provicial Archives of Natural Resources*, *Jinan* 25001, *China*
4. *School of Computer Science*, *China University of Geosciences*, *Wuhan* 430074, *China*

Abstract: With the development of deep learning technology, attackers can obtain potentially sensitive information from public geological data through classification, prediction, and other methods, which could lead to the leakage of important geological data. To solve the above problems, we propose a geological drilling data protection model based on graph adversarial attack *Gcntack*. Based on the degree properties of geological data topology, we first generate attacks that satisfy the same power-law distribution as tiny node disturbance. It can ensure that the adversarial attacks are not easy to be found, and while can change the classification result of the target node. Secondly, we introduce an attention mechanism. Using a graph attention network model based on interpretability, we analyze the properties of key nodes that directly affect the results of the adversarial attacks, so as to verify the rationality of the selecting adversarial nodes in the *Gcntack* model. Finally, a comprehensive evaluation, based on the benchmark dataset and geological drilling dataset, is presented to show this proposed scheme can reduce the prediction accuracy of attackers

基金项目: 济南市科技创新发展计划(社会民生专项)项目《数字孪生城市四维可视化信息系统及在济南城区的应用》(No. 232131001).

作者简介: 尚浩(1983-),男,高级工程师,硕士,从事水工环地质,城市地质及地质环境信息化等工作.ORCID:0000-0002-2221-3774. E-mail: 181909920@qq.com

引用格式: 尚浩,朱恒华,李双,宋晓媚,夏雨,刘惠,杨帆,2023.一种基于图神经网络的地质钻孔数据保护方案.地球科学,48(8):3151-3161.

Citation: Shang Hao, Zhu Henghua, Li Shuang, Song Xiaomei, Xia Yu, Liu Hui, Yang Fan, 2023. A Geological Borehole Data Protection Based on Graph Neural Networks. *Earth Science*, 48(8): 3151-3161.

and achieve the purpose of protecting important geological drilling data.

Key words: graph neural network; graph attention network; figure adversarial attack; interpretability; data protection for geological drilling data; deep learning.

0 引言

地质大数据是地理信息系统的基础数据,是国家大数据战略的重要组成部分,也是国民经济建设和国防建设中不可缺少的战略资源(Gagula and Santillan, 2020; Marti *et al.*,2020). 伴随着地理信息产业的高速发展和地质大数据云共享环境的广泛应用,地质数据在生产、存储、传输、使用过程中极易遭受窃取、篡改、伪造、侵权、泄密等安全威胁,严重阻碍了地质数据的应用与发展(翟明国等, 2018;李虎等, 2020). 作为国家基础设施建设的核心内容,地质数据安全已成为当前地理信息领域面临的重要挑战之一.

根据地质数据的不同存储和应用需求,已有方案提出了数字水印(Peng *et al.*, 2018; Vybornova *et al.*, 2019)、数据加密(Van *et al.*, 2017; Ren *et al.*, 2020)、数字脱敏(Hui *et al.*, 2017; 江栋华和周卫, 2018)等技术和方法. 这些方法能够较好的提高地质数据在存储、传输和展示使用过程中的安全性,但大多数研究集中在栅格及矢量地图数据,针对地质钻孔数据类型的研究成果较少. 地质钻孔数据具有数字化、离散性及关联性等特点(夏丁等, 2020),采用上述传统的数据保护方案具有较多的局限性. 除此之外,深度学习技术的发展加剧了地质钻孔数据泄露的风险. 例如,攻击者可以基于地质钻孔数据所构造的图数据构建深度学习模型,通过分类、预测等方法分析地质钻孔数据之间潜在的关联关系,获取公开地质钻孔数据的敏感地质信息,从而造成敏感地质数据的泄露(Cai *et al.*,2018).

考虑到上述特征,本文提出了一种基于图对抗攻击的地质钻孔数据保护模型*Gcntack*,通过操纵目标节点与关联节点的节点数据及地质钻孔图结构,达到降低地质钻孔数据分类与预测模型精度的目的. 此外,本文通过一种基于可解释性的图注意力网络模型分析了对抗扰动节点的可靠性,以及选取对抗性节点的合理性. 本文的主要贡献如下:

(1)本文提出了一种基于可解释性图对抗攻击的钻孔数据保护框架. 引入图对抗攻击理论,生成保持图数据度幂律分布不变的节点扰动,能够有效降低地质钻孔数据分类与预测的精度.

(2)本文提出了一种基于图神经网络的地质数据保护方案. 基于统计样本检验机制生成扰动数据,能实现降低目标模型分类准确率,以及保证扰动不易察觉的目的.

(3)本文提出了一种基于图注意力模型的地质数据扰动分析方法. 基于注意力系数分析添加扰动数据的目标节点性质,验证对抗攻击模型中选取对抗性节点的合理性,有效提升模型的可解释性.

此外,本文基于基准数据集和地质钻孔数据集,从直接攻击与间接攻击、结构攻击与特征攻击等方面,通过实验分析并验证了本文所提出的地质钻孔数据保护方案的有效性和可解释性.

1 相关工作

以下结合本文研究内容,对现有的地质数据保护方法、图卷积神经网络研究与对抗攻击技术进行简要阐述.

从数据存储、传输到显示使用,地质数据保护一直是地理信息领域研究的热点. 已有的地质数据保护方法主要集中在数字水印(Qiu *et al.*, 2019; Li and Zhu, 2019)、数据加密(Van *et al.*, 2017; Pham *et al.*, 2019)、数据脱敏(江栋华和周卫, 2018; Ali-Ozkan *et al.*, 2019)等方面. Qiu *et al.*(2019)基于水印图像压缩编码和转换思想,提出一种可嵌入、可逆的矢量地图数字水印方法. Li and Zhu(2019)基于图像复杂度指数及图像自相关指数提出了一种高鲁棒性的矢量地图版权认证方案. 数字水印技术主要提供矢量地图数据的真实性和完整性保护,但一般存在数据精度影响较大、保护算法较为复杂的缺点(Qiu *et al.*, 2017). 为保证地理数据中敏感数据不被非授权用户使用,研究者基于数据加密技术,实现对矢量地图数据的安全保护. 其中,Van *et al.*(2017)提出了一种基于顶点随机化和混合变换的GIS矢量地图选择性数据加密方案. Pham *et al.*(2019)基于提取主干对象、多尺度简化等方法,提出了一种GIS矢量地图数据加密方案. 但基于数据加密技术的地理数据保护方案通常改变了原始数

据结构,数据的分发较困难(Ren *et al.*, 2020). 已有的数据脱敏方案主要通过对敏感数据执行删除、修改、替换或偏移等操作,使数据精度下降,从而达到隐藏敏感数据的目的(Ali-Ozkan and Ouda, 2019). Cuzzocrea and Shahriar(2017)分析了多种数据脱敏技术,包括混淆、替换、遮掩和删除等方法,并应用于MongoDB和Cassandra两种主流的NoSQL数据库. 江栋华和周卫(2018)基于Logistic混沌系统及Chebyshev多项式,提出了精度可控的矢量地理数据脱敏方案. 但上述地质数据保护方案通常仅考虑了栅格及矢量地图数据的安全保护,难以应用于地质钻孔数据的脱敏保护.

图卷积神经网络是目前应用较为广泛的一种图数据分析模型,典型应用于挖掘、描述数据之间的相关性信息(Cai *et al.*,2018). Li *et al.*(2019)结合词向量和语法依赖关系信息,提出了一种基于图卷积神经网络的临床医学数据分类模型. Ma *et al.*(2020)研究了如何基于图神经网络,利用表征信息及相关信息实现节点级预测任务. 其中,表征信息为模型构建节点特征提供了指导,而相关信息用于描述节点特征之间的相关性. 对抗样本的发现源自于对深度学习可解释性的探索(刘会等,2021). 现有的图对抗攻击方法主要关注对抗样本的生成方法及模型攻击的成功率(Zügner *et al*, 2018). Bojchevski and Günnemann(2019)提出了一种基于摄动理论的无监督节点表示模型的图数据投毒攻击方案. Dai *et al.*(2018)等提出一种基于强化学习的图数据攻击方案,该方案具有较高的鲁棒性. Sun *et al.*(2020)提出了一种基于马尔可夫决策以及强化学习的节点投毒攻击方法,能够降低模型整体的节点分类性能. Zügner *et al.*(2018)考虑了节点分类模型的逃避攻击和投毒攻击,基于静态代理模型执行攻击,并通过训练分类器来评估攻击效果. Chen *et al.*(2020)利用已训练的图形自动编码器模型中的梯度信息,提出了一种新的迭代梯度攻击策略.

受到已有研究工作(Zügner *et al.*,2018; Wang *et al.*,2020)的启发,本文在地质钻孔数据保护中引入图攻击理论和注意力机制,通过图卷积神经模型生成遵循度幂律分布的节点扰动,并基于重要度分布特征选取对抗性节点,以降低目标节点的分类准确度,从而保护目标节点数据并提高地质钻孔数据保护模型的可解释性.

2　模型设计

2.1　方法模型

本方案基本思想是将图对抗攻击和注意力机制引入地质钻孔数据的保护中,通过添加“微小”扰动,使得地质钻孔数据分类准确率下降,并基于注意力GAT(graph attention networks)模型对图对抗攻击选取的对抗性节点进行合理性验证. 具体方案示意图如图1所示.

方案的基本过程如下:首先,对钻孔数据进行预处理,构建拓扑图数据结构. 将上述图数据用于训练GCN模型,通过保持图 $G^{(0)}=(A^{(0)},X^{(0)})$ 幂律不变的特征来保证扰动不易被察觉. 其次,选取对抗性节点,生成针对钻孔图结构和节点特征的对抗扰动,分别使用直接攻击与间接攻击、结构攻击与特征攻击对图 $G^{(0)}$ 进行攻击,导致 $G^{(0)}$ 被干扰成为 G'. 由于对模型进行了对抗攻击,基于图数据 G' 节点分类任务将会把目标节点误分类为新类别 Y_1,使得攻击者无法正确预测目标钻孔节点的分类类别.

此外,本文基于图注意力网络模型计算节点的注意力系数,通过注意力权重值定位对抗性节点与注意力权重的对应关系,验证GCN模型中对地质钻孔节点选取的合理性.

2.2　地质数据保护模型

地质钻孔数据可以表示为图结构数据,即图结构数据 $G=(A,X)$,其中 $X\in\{0,1\}^{N\times D}$ 为钻孔节点特征,$A\in\{0,1\}^{N\times N}$ 为地质钻孔数据的邻接矩阵. 笔者用 $x_v\in\{0,1\}^D$ 表示节点 v 的 D 维特征向量,并假设节点 ID 为 $V=\{1,2,\cdots,N\}$,特征 ID 为 $F=$

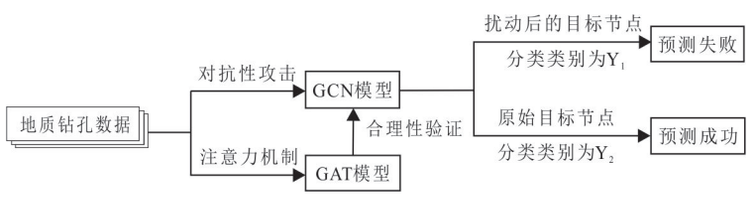

图1　基于图对抗攻击和可解释性的地质钻孔数据保护方案

Fig. 1　Geological borehole data protection scheme based on graph adversarial attack and interpretability

$\{1, 2, ..., D\}$. 此外，笔者定义节点标签和类别标签

图数据节点分类任务的目的是学习一个函数映射 $g: V \rightarrow C$，该函数将每个节点 $v \in V$ 映射到 C 中的一个类. 在这项任务中，图卷积神经网络的隐藏层 $l+1$ 可以定义为：

$$H^{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \quad (1)$$

其中：$\tilde{A} = A + I_N$ 表示增加了自连接的邻接矩阵. $W^{(l)}$ 是第 l 层的可训练权重参数，$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$，$\sigma(\cdot)$ 为激活函数（通常是 ReLU）. 基于上述图卷积神经网络模型的节点分类任务将节点特征 X 作为第一层的输入，即 $H^{(0)} = X$. 其分类任务最终输出结果可以表示为：

$$\begin{aligned} Z &= f_\theta(A, X) \\ &= \mathrm{softmax}(\hat{A}\sigma(\hat{A}XW^{(1)})W^{(2)}), \end{aligned} \quad (2)$$

其中：$\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$，输出 Z_{vc} 表示将节点 v 分配给类 c 的概率，θ 表示所有参数的集合，即 $\theta = \{W^{(1)}, W^{(2)}\}$. 本文借鉴（Kipf and Welling, 2017）提出的图卷积网络模型，将节点分类模型设置为两层分类器，即仅聚合来自两层邻域节点的节点信息. 并通过最小化标记样本 V_L 的输出交叉熵，并以半监督的方式学习参数 θ.

$$L(\theta; A, X) = -\sum_{v \in V_L} \ln Z_{v, c_v},\ Z = f_\theta(A, X), \quad (3)$$

其中：c_v 是训练集中 v 的给定标签. 训练后，Z 表示图中每个目标节点的分类概率.

2.3 图对抗攻击模型

对上述图数据 $G^{(0)} = (A^{(0)}, X^{(0)})$ 执行“微小”扰动，形成图 $G' = (A', X')$，从而降低图卷积分类模型的分类性能. 其中，对 $A^{(0)}$ 进行的扰动称为结构攻击，而对 $X^{(0)}$ 进行的扰动称为特征攻击.

具体来说，笔者将攻击特定的目标节点 $v_0 \in V$，使其分类预测值改变. 但由于图结构数据具有非独立同分布性质，v_0 的分类结果不仅取决于 v_0 节点本身，同时受到图中的周边邻接节点的影响. 因此，在本文中，笔者引入了攻击者节点集合 $A \subseteq V$，对 $G^{(0)}$ 的扰动仅限于该集合，即必须满足：

$$X'_{ui} \neq X^0_{ui} \Rightarrow u \in \mathrm{A}, A'_{uv} \neq A^0_{uv} \Rightarrow u \in \mathrm{A} \bigvee v \in \mathrm{A}, \quad (4)$$

其中：ui 表示节点 u 的特征 i，uv 为两个遭受攻击的目标节点. 如果目标 $v_0 \notin A$，即 v_0 不会直接被操纵，只能通过影响者节点间接操纵，笔者将该攻击称为间接攻击. 如果 $\{\mathrm{v}_0\} = \mathrm{A}$，笔者称之为直接攻击.

为确保对图的改动足够“微小”，笔者进一步通过预算 Δ 限制了允许更改的次数：

$$\sum_u \sum_i |X'_{ui} - X^{(0)}_{ui}| + \sum_{u<v} |A'_{uv} - A^{(0)}_{uv}| \leqslant \Delta, \quad (5)$$

其中：$\left|X'_{ui} - X^{(0)}_{ui}\right|$ 表示节点 u 允许修改的特征数量，$|A'_{uv} - A^{(0)}_{uv}|$ 表示满足条件的节点之间允许修改的边缘数量.

基于此，笔者用 $P^{G^0}_{\Delta, A}$ 表示满足式（5）的所有图 G' 的集合. 当给定一个图数据 $G^{(0)} = (A^{(0)}, X^{(0)})$，目标节点 v_0 和攻击节点集合 A，令 c_{old} 表示原始图 $G^{(0)}$ 下 v_0 所属的原始类别. 图节点扰动问题定义如下：

$$\begin{aligned} &\arg\max_{(A', X') \in P^{G^{(0)}}_{\Delta, A}} \max_{c_{\mathrm{new}} \neq c_{\mathrm{old}}} \left(\ln Z^*_{v_0, c_{\mathrm{new}}} - \ln Z^*_{v_0, c_{\mathrm{old}}}\right) \\ &s.t.\ Z^* \\ &= f_{\theta^*}(A', X') \text{ with } \theta^* \\ &= \arg\min_\theta L(\theta; A', X'). \end{aligned} \quad (6)$$

即找到一个经过扰动的图 G'，该图将目标节点 v_0 最终分类为 c_{new}，并且与正确分类 c_{old} 的概率值差距最大.

2.4 不可察觉的约束设置

通常，在图像、文本对抗性攻击中，可以通过在视觉上和使用简单约束修改输入数据. 图结构是离散的，且由于级联效应，对图数据进行对抗性扰动具有一定困难性. 图结构的最显著特征是其度数分布，它在实际网络中通常呈现出幂律形状的分布. 如果两个网络显示出非常不同的度数分布，则很容易将它们区分开. 因此，笔者的目标是产生遵循同一幂律分布的攻击作为“微小”的扰动.

笔者参考幂律分布的统计样本检验（Borgs *et al.*, 2019），即使用似然比检验来估计 $G^{(0)}$ 和 G' 两个图的度是否源自同一分布：参考 $G^{(0)}$ 的度数分布来估计幂律分布 $p(x) \propto x^{-\alpha}$ 的缩放参数 α_G，扰动后的图 G' 的幂律分布参数为 $\alpha_{G'}$，可以通过该参数判断扰动前后两个图的幂律分布是否等效. 在离散数据的情况下，虽然没有精确和封闭形式的解决方案来计算 α，但可以使用文献（Eikmeier and Gleich, 2017）中的一个近似表达式估算该值，对于图 G 而言，它转化为：

$$\alpha_G \approx 1 + |D_G| \cdot \left[\sum_{d_i \in D_G} \log \frac{d_i}{d_{\min} - \frac{1}{2}}\right]^{-1}, \quad (7)$$

其中：$d_{\min}$ 表示幂律测试中必须考虑的最小节点度，而 $D_G = \left\{d^G_v | v \in V, d^G_v \geqslant d_{\min}\right\}$ 是包含不小于最小节

点度其余节点的度集合,d_v^G 为图 G 中节点 v 的度.据此,笔者可以得到 $\alpha_{G^{(0)}}$ 和 $\alpha_{G'}$ 的估计值.样本 D_x 的对数似然可评估为:

$$l(D_x)=|D_x|\cdot\log\alpha_x+|D_x|\cdot\alpha_x\cdot\log d_{\min}+(\alpha_x+1)\sum_{d_i\in D_x}\log d_i\,, \tag{8}$$

其中:α_x 为给定的缩放参数.基于上述对数似然函数值,笔者可以估计两个样本 $D_{G^{(0)}}$ 和 $D_{G'}$ 是否来自相同的幂律分布;也就是说,笔者提供了两个相互竞争的假设,先对总体参数提出一个假设值,然后利用样本信息判断这一假设是否成立:

$$l(H_0)=l(D_{\text{comb}})\,, \tag{9}$$

$$l(H_1)=l\left(D_{G^{(0)}}\right)+l(D_{G'})\,, \tag{10}$$

其中:组合样本 $D_{\text{comb}}=D_{G^{(0)}}\cup D_{G'}$. H_0 表示原假设,也叫零假设,原假设一般是统计者想要拒绝的假设,拒绝零假设也叫Ⅰ型错误,即两个样本来自不同分布.H_1 表示备择假设,备择假设是统计者想要接受的假设,拒绝备择假设,也叫Ⅱ型错误,即两个样本来自相同分布.

在似然比检验之后,利用公式(9)、(10)中两个假设检验的对数似然函数值 $l(H_0)$ 和 $l(H_1)$,计算得到最终检验统计量 Λ 为:

$$\Lambda\left(G^{(0)},G'\right)=-2\cdot l(H_0)+2\cdot l(H_1)\,. \tag{11}$$

上述公式中的 G' 为图数据 $G^{(0)}$ 执行“微小”扰动后形成的图.通常而言,大样本数据一般服从自由度为 n 的 χ^2 分布.从统计学上讲,拒绝零假设 H_0 的典型 p 值是 0.05,如果要从相同的幂律分布中采样两个度数序列,笔者将在 95% 的置信区间内拒绝备择假设 H_1. 该临界 p 值与Ⅰ型错误具有反比关系,因此笔者将临界 p 值设置为 0.95. 在 χ^2 分布中,笔者仅接受满足度数分布如下的扰动 $G'=(A',X')$:

$$\Lambda\left(G^{(0)},G'\right)<\tau\,, \tag{12}$$

其中:笔者取 $\tau\approx10^{-3}$,由于笔者的目标是产生遵循同一幂律分布的“微小”扰动,所以计算得到的 $\Lambda\left(G^{(0)},G'\right)$ 越小越好.

2.5　可解释性的图注意力层

为了使图数据表达能力最大化,即需要将输入特征 $h=\left\{\vec{h}_1,\vec{h}_2,\cdots,\vec{h}_N\right\}$,$\vec{h}_i\in R^F$ 转换为高级特征 $h'=\{\vec{h}'_1,\vec{h}'_2,\cdots,\vec{h}'_N\}$,$\vec{h}'_i\in R^{F'}$,则需要一种可学习的线性转换表示.为此,作为初始步骤,笔者基于权重矩阵 $W\in R^{F'\times F}$ 对所有的节点进行参数化,权重矩阵对应输入特征与输出特征之间的关系.并且,基于注意力机制 a 计算注意力相关系数 e_{ij}:

$$e_{ij}=\text{attn}(W\vec{h}_i,W\vec{h}_j)\,. \tag{13}$$

上述公式中的注意力相关系数 e_{ij} 表明节点 j 的特征对节点 i 的重要性.其中,W 表示输入特征与输出特征之间关系的权重矩阵,$\vec{h}_i$、$\vec{h}_j$ 分别表示输入的第 i、j 个输入特征.一般而言,该模型允许每个节点参与其他各节点的特征变换,从而丢弃所有结构信息.笔者通过执行掩膜注意力(Masked Attention)将图结构引入模型中,模型仅计算节点 $j\in N_i$ 的 e_{ij},其中 N_i 指节点 i 的一阶邻域节点.

为了使不同节点之间的系数可以进行比较,采取 softmax 函数对所有选择的邻域节点 j 进行归一化:

$$\alpha_{ij}=\text{softmax}_j\left(e_{ij}\right)=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})}. \tag{14}$$

上述公式中 e_{ik} 表示节点 $k\in N_i$ 的注意力相关系数.注意力机制 a 是由权重向量 $\vec{a}\in R^{2F'}$ 参数化得到的单层前馈神经网络,并应用 LeakyReLU 非线性函数(输入斜率 $\alpha=-0.2$)激活.完全展开后,最终的注意力系数 α'_{ij} 可以表示为:

$$\alpha'_{ij}=\frac{\exp(\text{LeakyReLU}(\vec{a}^T[W\vec{h}_i\|W\vec{h}_j]))}{\sum_{k\in N_i}\exp(\text{LeakyReLU}(\vec{a}^T[W\vec{h}_i\|W\vec{h}_j]))}, \tag{15}$$

其中:a^{T} 代表转置操作,$\|$是串联操作,$\vec{a}^T$ 代表权重向量 $\vec{a}$ 的转置.基于邻接矩阵 A,用掩膜操作(mask)进行过滤,即邻接矩阵 A 中元素为 1 的位置,将其注意力系数设置为 α'_{ij},获得注意力系数矩阵.

注意力系数 α'_{ij} 将用于计算与其对应的特征的线性组合,得到每个节点用在输出层进行节点分类的输出特征 $\vec{h}'_i$:

$$\vec{h}'_i=\sigma\left(\sum_{j\in N_i}\alpha'_{ij}W\vec{h}_j\right), \tag{16}$$

3　实验及分析

3.1　实验环境

在本节中,笔者将对基于图卷积神经网络的地质钻孔数据保护方案进行实验评估.从模型攻击效果、可解释性等方面对钻孔数据保护方案进行分析.本文实验均基于 4 核、3.20GHz Intel i5-6500 CPU、8GB RAM、NVIDIA GeForce GTX 1050、Windows 10 Enterprise 64 位版本商用版主机,使用 Tensorflow 架构实现上述模型.

图 2 研究区钻孔节点分布图

Fig. 2 Distribution map of borehole nodes in the study area

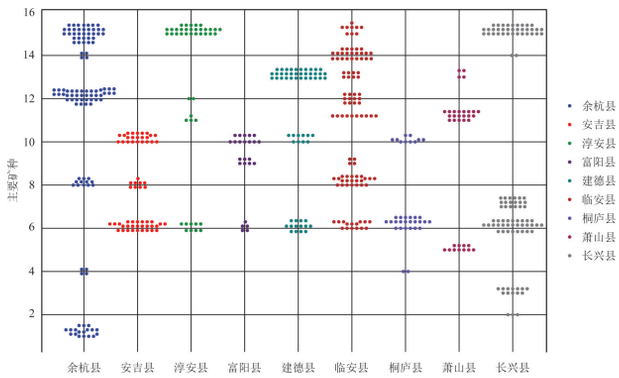

图 3 按照主要矿种－工作区构建的钻孔散点图

Fig. 3 Borehole scatter diagram constructed according to the main ore-working area

3.2 数据集

为了验证本文方案的有效性，实验采用了两类不同的数据集进行实验分析，分别使用两个公开的数据集和两个自建的钻孔图数据集对模型进行实验．下面将对数据集进行详细说明．

公开数据集采用了常用的引文网络：Cora 和 Citeseer，其中节点代表论文，边缘代表引用关系．节点数量是数据集的论文数量，特征是每篇论文在去除停用词和论文中出现频率小于 10 次的词之后剩下的独立单词，原论文和引用的论文构成了连边．Cora 数据集中共有 2 708 篇论文，其中每篇论文引用其他论文或被至少一篇其他论文引用．Cora 数据集包含 1 433 个独特单词，所以特征是 1 433 维，0 和 1 描述的是每个单词在论文中是否存在．

此外，笔者选取全国重要地质钻孔数据库服务平台（网址：http://zkinfo.cgsi.cn/）提供的钻孔数据构建地质钻孔数据集．需要说明的是，本文选取全国重要地质钻孔数据库中的公开地质钻孔数据集作为实验对象并验证方案的有效性．但本文中提出的方法不局限于地质钻孔数据保护，其他类似的地质和地理资料成果也可以采用本方法进行保护．

以浙江省的钻孔节点为例，可获得钻孔数据的钻孔类型、钻孔编号、主要矿种、终孔深度、终孔日期、工作区名称等信息．同一地区、同一终孔深度的钻孔数据通常属于相同的钻孔类型，利用地质钻孔数据的这种关联性可以构建拓扑图，该拓扑图获得的邻接矩阵和特征矩阵适用于图深度学习模型进行后续的学习任务．

地质钻孔节点是一种复杂的数据类型，其中包含化学元素、矿物类型、岩性等重要敏感地质信息．本文以矿物类型为例进行图攻击实验．

由于整个浙江省钻孔分布的工作区较多，图 2 展示局部工作区的钻孔节点分布图．其中主要矿种类型有膨润土、煤、萤石、石灰岩、重晶石、透辉石、汞、钒、钨、铁、铜、锌、金、银等．同时笔者通过可视化的手段将按照主要矿种及工作区、钻孔类型及终孔深度这两组属性关系分别将钻孔节点构建成散点图进行展示，如图 3 和图 4 所示．从图 3 中，笔者可以看出同一地区分布的钻孔数据通常主要矿种相同；地理上相邻地区的矿种是近似的，如余杭县和安吉县都存在膨润土．这说明，矿种的富集与某一地区独特的地质、天气条件有强内在关联性，相邻地区由于在地质、天气条件上有相似性，因此某些矿种可能会相同，而不同地区的矿种往往存在较大的差异．

下图 4 中，C1～C7 分别表示“非金属”、“贵金属”等 7 种钻孔类型．从图中笔者可以看出，除贵金属钻孔类型下节点较少外，其余类型下的钻孔节点均呈现出较为明显的随终孔深度不同、矿种类型聚集的的原因造成的差异化特点，这说明了地质体经过长时序的同生、影响和演化，导致了钻孔数据之间存在着较强的内在关联特征，也说明了笔者使用图神经网络来处理地质数据的合理性．

笔者针对浙江省的 670 个公开钻孔节点，分别构造可用于 GCN 分类的两个全连通图：Drilling-1 和 Drilling-2. Drilling-1 按照主要矿种及工作区两类信息将钻孔节点构建成全连通图，即：如果两个钻孔的主要矿种相同，则这两个钻孔间添加一条边；如果两个钻孔的工作区相同，则这两个钻孔间也添加一条边．Drilling-2 按照终孔深度及钻孔类型两类

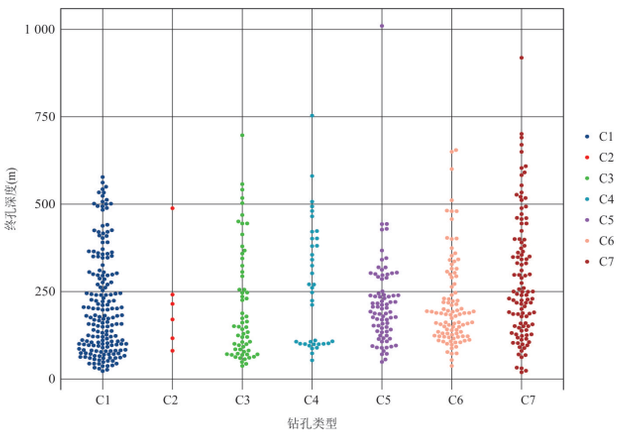

图 4　按照终孔深度－钻孔类型构建的钻孔散点图

Fig. 4　Borehole Scatter Diagram constructed according to final hole depth-type of borehole

表 1　数据集统计信息

Table 1　Data set statistics

名称	节点数量	边缘数量	特征	类别
Cora	2 708	5 429	1 433	7
Citeseer	3 327	4 732	3 703	6
Drilling-1	670	35 836	106	7
Drilling-2	670	95 472	106	7

信息将钻孔节点构建成全连通图. 构造了全连通图后,根据其连边的情况将整个图用邻接矩阵的形式进行存储和表示. 根据钻孔类型字典库将数据集分成了 7 种类别,分别是“0,非金属钻孔”,“1,贵金属钻孔”,“2,黑色金属钻孔”,“3,矿产钻孔”,“4,煤田钻孔”,“5,有色金属钻孔”,“6,综合矿产钻孔”. 钻孔特征是钻孔数据的主要矿种和工作区名称. 下表 1 显示了本文所使用的数据集信息.

笔者将数据集分为标记节点(20%)和未标记节点(80%). 接着将标记节点等分为 10% 的训练集和 10% 的验证集作为训练模型的数据. 也就是说,笔者在训练过程中从验证集中删除标签,并将其用作停止条件. 实验目标是进行节点分类,预测无标签的钻孔节点所属的钻孔类型.

3.3　实验结果分析

参考文献(Zhu *et al.*, 2019),层数设置为 2. 对于图卷积神经网络(GCN),笔者将隐藏单元数设置为 16, dropout 设置为 0.5, weight_decay 设置为 0.000 5,并对所有权重矩阵使用 Xavier 初始化,使用固定学习率 0.01,并将 epoch 数设置为 200.

基于上述实验数据与模型参数设置,分析基于图对抗攻击和可解释性的地质钻孔数据保护方案的效果.

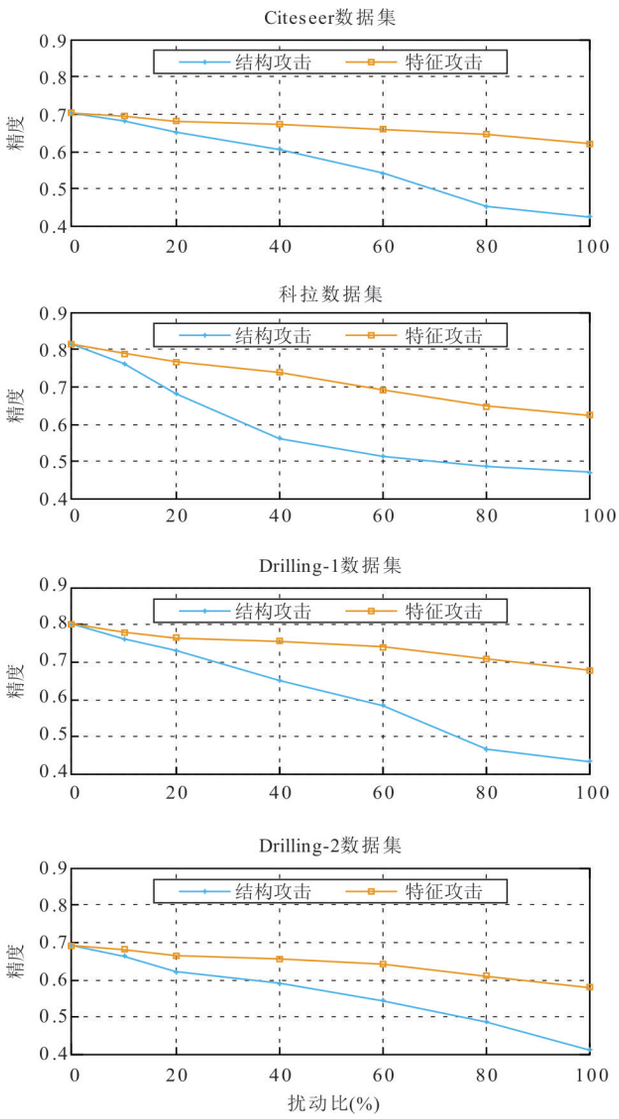

图 5　直接攻击与间接攻击的分类准确率

Fig. 5　Classification accuracy of direct and indirect attacks

3.3.1　图对抗攻击　图对抗攻击实验在 Cora 和 Citeseer 两个公开引文网络数据集以及笔者自建的两个地质钻孔数据集(Drilling-1 和 Drilling-2)上进行. 笔者分别使用直接攻击与间接攻击、结构攻击与特征攻击的方式以针对特定目标节点进行图对抗攻击,使得模型对目标节点的分类准确率下降,达到保护关键钻孔数据的目的.

首先对 4 个数据集分别进行直接攻击和间接攻击,在每个类别里面随机选取 5 个节点作为笔者攻击的目标节点,并计算运行 10 次后的平均结果.

实验结果如图 5 所示,其中,紫色的折线代表直接攻击的准确率趋势,绿色的折线代表间接攻击的准确率趋势,图 5 分别表示 Cora、Citeseer、Drilling-

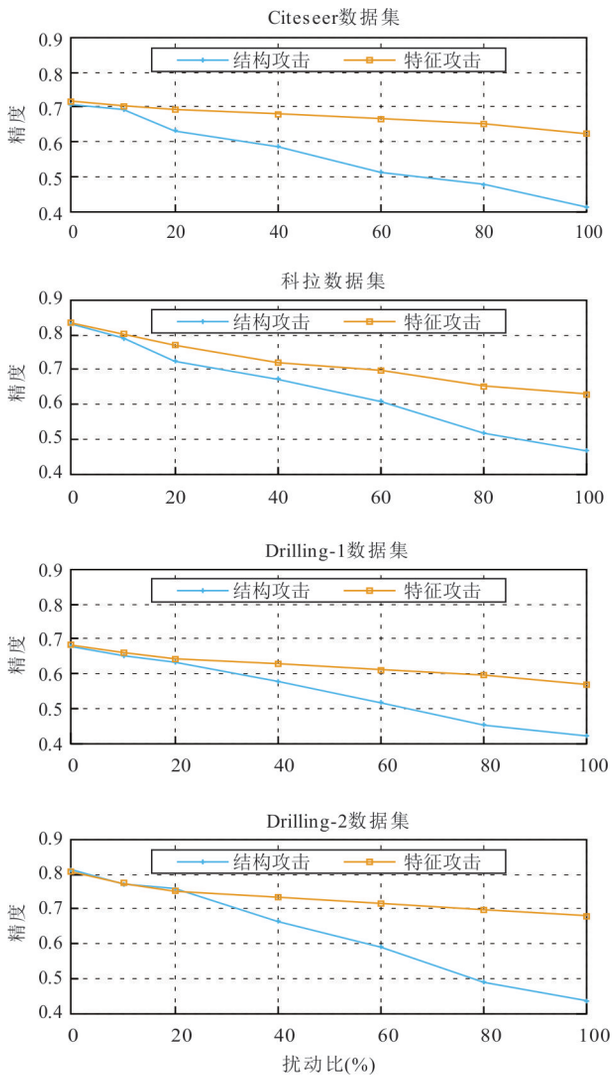

图 6 结构攻击和特征攻击的分类准确率

Fig. 6 Classification accuracy of structural and feature attacks

1、Drilling-2 的实验结果. 从图中可以看到,随着扰动次数增加,两种攻击方法的分类性能都会下降. 笔者可以看到使用间接攻击,即从目标节点的邻域中选择 5 个随机节点作为攻击对象,最终攻击的结果仍然具有相对较高的分类准确性,比直接攻击准确率高 20% 左右. 由于笔者的目标是使最终分类结果降低,根据实验结果可以看出,间接攻击通常不如直接攻击有效,因此笔者选取直接攻击进行后续实验.

在直接攻击的基础上,分别对节点进行结构攻击和特征攻击. 在每个类别中选取目标节点,对目标节点进行直接的增加/删除边缘操作,或对目标节点的特征进行增加/删除,计算运行 10 次后的平均结果. 实验结果如图 6 所示.

如上图 6 所示,*Gcntack*-Stru 代表结构攻击,

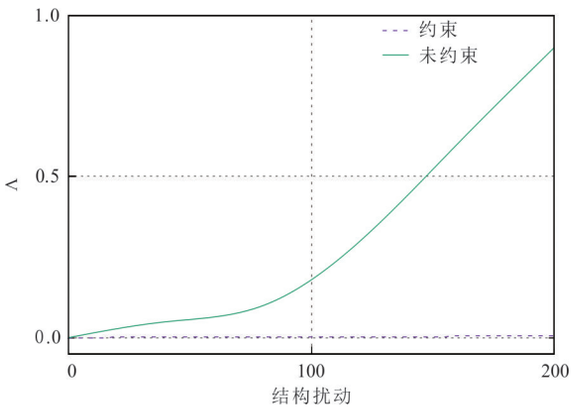

图 7 约束对图检验统计量的影响

Fig. 7 Influence of constraints on graph test statistics

Gcntack-Feat 代表特征攻击,图 6 分别表示 Cora、Citeseer、Drilling-1、Drilling-2 的实验结果. 从图中可以看出,与特征攻击相比,随着扰动次数的增加,对结构进行直接扰动导致错误分类的下降幅度更大,因此笔者选取结构攻击进行后续实验;同时笔者可以看到,相较于基于主要矿种及工作区的 Drilling-1 数据集,按照终孔深度及钻孔类型构建的 Drilling-2 数据集,在攻击前正确分类率高,攻击后分类准确率下降得更大,取得了更好的误分类效果,因此实验最终选择 Drilling-2 数据集进行后续实验.

图结构最显著的特征是其度数分布,它在实际网络结构中通常类似于幂律形状. 如果两个网络结构显示出非常不同的度分布,则很容易将它们区分开. 图 7 显示了在有或没有约束的情况下,拓扑图在似然比检验之后幂律分布的测试统计量 Λ. 如果不添加约束,随着扰动次数的增加,修改后图的幂律分布与原始图越来越不同,图中曲线逐步上升;而在添加约束的情况下,原图和扰动后的图在似然比检验之后获得幂律分布的测试统计量 $\Lambda\left(G^{(0)}, G'\right)<\tau\approx 10^{-3}$,因此在图中曲线斜率接近 0,表示幂律基本不变. 结合本节节尾的表 2 可以看出,笔者施加的幂律不变的约束仍然可以成功进行对抗攻击,同时保证了攻击不易被察觉.

总体来说,直接攻击比间接攻击有效,结构攻击比特征攻击有效. 并且在引入幂律约束后,不会影响攻击效果,同时能让产生的扰动“无法察觉”,攻击的隐蔽性更好.

3.3.2 模型可解释性 对于目标节点来说,图对抗攻击选取的对抗性节点分别有该类别下和其他类别的节点,而注意力机制只能获得与目标节点相同类别下邻域节点的注意力权重值. 为了获取不同类

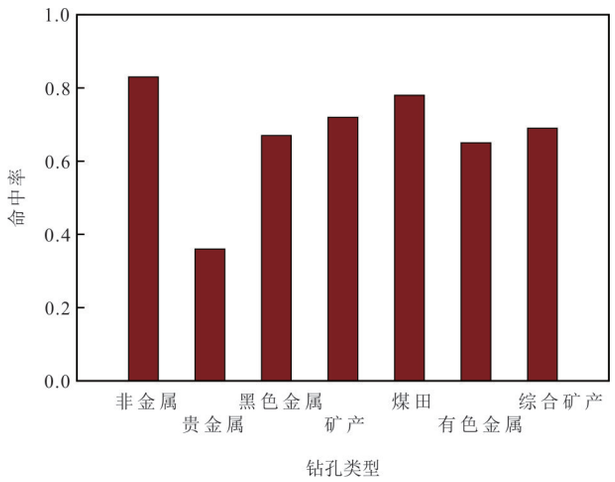

图 8　7个钻孔类型中注意力权重高的节点与对抗性节点对应的命中率

Fig. 8　The hit ratio of the nodes with high attention weight and the adversarial nodes in the seven drilling types

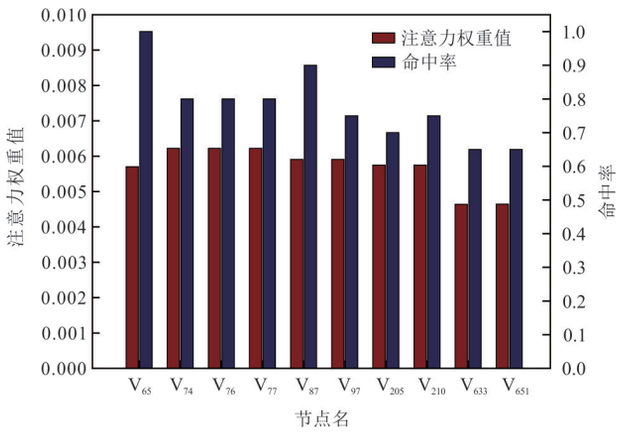

图 9　排名前10的地质钻孔注意力权重统计结果

Fig. 9　Statistical results of the top 10 geological borehole attention weights

别之间节点的注意力权重,本实验通过对主要矿种－钻孔类型、主要矿种－工作区名称、主要矿种－终孔深度3种方式分别构建全连通图;接着用3种分类方式依次计算注意力权重矩阵;再按照主要矿种和钻孔类型的对应关系进行聚合,计算钻孔类型的高注意力节点与对抗性节点的重合度,得到钻孔数据集中针对主要矿种进行分类的关键节点的注意力权重,以提高模型的可解释性.

如图8表示浙江省钻孔数据集中的7个类型下注意力权重高的节点与对抗性节点对应的命中率.从柱状图中可以看出,贵金属钻孔类型命中率较低,只有36%,和预想一致,这是因为贵金属钻孔类型下主要由金、银等金属构成,稀有金属资源分布较少导致浙江省钻孔数据集中该类别下节点较少,只有8个节点;而图对抗攻击节点的选取与节点度有关,贵金属类别下节点度均较小,导致对抗性节点选取较少.其余类别下节点数量在100左右,最终注意力权重高的节点与选取的对抗性节点对应命中率平均值达75%左右.

由于共有7个类型,每个类型下权重高的节点都不一致,接下来将详细介绍如何以注意力权重值定位找到对抗性节点与注意力权重的对应关系.以“非金属矿产地质勘查钻孔”类别为例,根据注意力权重排序,列出排名前十的地质钻孔统计结果,如图9所示:

对非金属矿产地质勘查钻孔进行了20次对抗性实验,其中节点v_{65}作为注意力权重较高的节点,其注意力权重值为0.005 7,出现在对抗性节点中共20次,命中率为100%.节点v_{633}、v_{74}、v_{76}、v_{77}、v_{87}等分别出现13、16、16、16、18次,命中率分别为65%,80%,80%,80%,90%,平均命中率为78%.通过对钻孔数据测试集的注意力权重统计分析,可以看出注意力权重和对抗性节点的命中率呈现正相关趋势,验证了本文提出的基于图对抗攻击的注意力机制在可解释性分析方面的可行性,可以得到影响对抗攻击结果的关键节点,验证了GCN模型中选取对抗性节点的合理性.

表 2　数据集上节点分类的准确性(百分比)

Table 2　Accuracy of node classification on dataset (percentage)

	Cora	Citeseer	Drilling-1	Drilling-2
GCN	0.815	0.701	0.709	0.803
GAT	0.845	0.725	0.728	0.826
Gcntack	**0.471**	**0.423**	**0.410**	**0.432**

3.3.3　实验结果对比　为进一步测试本文所提出方案的有效性,笔者首先对未受到攻击的数据集进行实验,然后再采用图对抗攻击添加扰动,得到节点分类准确率.

从表2中可以看出笔者提出的图对抗攻击方法Gcntack能够将特定目标节点的分类准确性降低到50%以下,这里*Gcntack*采用“直接攻击＋结构攻击”取得最终分类结果,同时在Drilling-2数据集上取得分类准确率下降的最大幅度.

4　结论

随着深度学习技术的的日益成熟以及地质数

据的广泛应用,攻击者可以对公开的地质数据使用深度学习技术进行分类、预测,以获取潜在的敏感信息.传统的地质数据保护方法,诸如数字水印、数据加密、数据隐写等,未考虑到地质数据之间的关联性.深度学习模型可以从地质数据的拓扑图表达中学习到潜在的知识,捕获地质数据的图结构和特征等信息,从地质数据的已知特征和结构中提炼分类依据,从而对地质资料的安全造成严重的威胁.为解决上述问题,本文提出了一种基于图神经网络的地质数据保护研究方案,核心思想是将图对抗攻击引入地质数据中,添加保持图节点幂率不变的"微小"扰动,使得特定目标节点的分类准确率下降,达到地质数据保护的效果.

本文引用了基于可解释性的深度学习模型(GAT),通过给关键邻域节点分配更小的权重,能够使模型的整体分类准确度下降.同时在GAT模型的基础上应用分层机制,通过每个层次层上的粗化操作和细化操作,可以增加节点的感受野并更有效地传递节点特征,进而可以捕获来自最相关邻居的节点信息.通过修改关键邻域节点使其权重降低,可以进一步降低全局分类准确度.本文在通用基准数据集和地质钻孔数据集上进行了广泛的实验,实验结果表明,本文的基于图对抗性攻击策略、注意力机制和分层机制的方法,能够分别有效地降低地质钻孔的局部和全局分类准确率,可以达到对钻孔数据保护的效果,从而证明了方案的合理性和有效性.本文方法针对地质数据间的关联性,为地质数据安全保护的研究和应用进行了一些有益的探索.

References

Ali-Ozkan, O., Ouda, A., 2019. Key-Based Reversible Data Masking for Business Intelligence Healthcare Analytics Platforms. *2019 International Symposium on Networks, Computers and Communications(ISNCC)*, 2019: 1－6. https://doi.org/10.1109/ISNCC.2019.8909125.

Bojchevski, A., Günnemann, S., 2019. Adversarial Attacks on Node Embeddings via Graph Poisoning. *ICML*, 97: 695－704.

Borgs, C., Chayes, J., Cohn, H., et al., 2019. An Theory of Sparse Graph Convergence I: Limits, Sparse Random Graph Models, and Power Law Distributions. *Transactions of the American Mathematical Society*, 372(5): 3019－3062. https://doi.org/10.1090/tran/7543

Cai, H. Y., Zheng, V. W., Chang, K. C. C., 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9): 1616－1637. https://doi.org/10.1109/tkde.2018.2807452

Chen, J. Y., Lin, X., Shi, Z. Q., et al., 2020. Link Prediction Adversarial Attack Via Iterative Gradient Attack. *IEEE Transactions on Computational Social Systems*, 7(4): 1081－1094. https://doi.org/10.1109/tcss. 2020. 3004059

Cuzzocrea, A., Shahriar, H., 2017. Data Masking Techniques for NoSQL Database Security: A Systematic Review. *2017 IEEE International Conference on Big Data*, 2017: 4467－4473.https://doi.org/10.1109/BigData.2017.8258486

Dai, H.J., Li, H., Tian, T., et al., 2018. Adversarial Attack on Graph Structured Data. *ICML*, 80:1123－1132.

Eikmeier, N., Gleich, D.F., 2017. Revisiting Power-Law Distributions in Spectra of Real-World Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017: 817－826.

Gagula, A.C., Santillan, J.R., 2020. Integrating Geographic Information System, Remote Sensing Data, Field Surveys, and Hydraulic Simulations in Irrigation System Evaluation. Proceedings of the 2020 IEEE REGION 10 CONFERENCE (TENCON), Osaka, 2020:626－630.

Hui, Y.U., Wei, Z., Xinnian, M.A., 2017. A Reversible Decryption Model for Vector and Raster Integration Based on Trigonometric Function. *Bulletin of Surveying and Mapping*, (10): 89－94.

Jiang, D.H., Zhou, W., 2018. Decryption Model for Vector Geographic Data Based on Chebyshev Polynomials.*Journal of Geomatics Science and Technology*. 35(3): 321－325(in Chinese with English abstract).

Kipf, T.N., Welling, M., 2017. Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations*, 1－14.

Li, A. B., Zhu, A. X., 2019. Copyright Authentication of Digital Vector Maps Based on Spatial Autocorrelation Indices. *Earth Science Informatics*, 12(4): 629－639. https://doi.org/10.1007/s12145-019-00386-z

Li, Y. F., Jin, R., Luo, Y., 2019. Classifying Relations in Clinical Narratives Using Segment Graph Convolutional and Recurrent Neural Networks (Seg-GCRNs). *Journal of the American Medical Informatics Association*, 26(3): 262－268. https://doi.org/10.1093/jamia/ocy157

Li, H., Zhu, H.H., Hua, W.H., et al., 2020.Key Technologies and Methods for Vector Geographic Data Security Protection. *Earth Science*, 45(12): 4574－4588 (in Chinese with English abstract).

Liu, H., Zhao, B., Guo, J.B., et al., 2021. Survey on Adver-

sarial Attacks Towards Deep Learning. *Journal of Cryptologic Research*, 8(2): 202－214(in Chinese with English abstract).

Ma, J.Q., Chang, B., Zhang, X., et al., 2020. CopulaGNN: Towards Integrating Representational and Correlational Roles of Graphs in Graph Neural Networks. *International Conference on Learning Representations*, 2020:1－13.

Marti, R., Li, Z. C., Catry, T., et al., 2020. A Mapping Review on Urban Landscape Factors of Dengue Retrieved from Earth Observation Data, GIS Techniques, and Survey Questionnaires. *Remote Sensing*, 12(6): 932. https://doi.org/10.3390/rs12060932

Peng, Y. W., Lan, H., Yue, M. L., et al., 2018. Multipurpose Watermarking for Vector Map Protection and Authentication. *Multimedia Tools and Applications*, 77(6): 7239－7259. https://doi.org/10.1007/s11042-017-4631-z

Pham, G.N., Ngo, S.T., Bui, A.N., et al., 2019. Vector Map Random Encryption Algorithm Based on Multi‐Scale Simplification and Gaussian Distribution. *Applied Sciences*, 9(22): 4889. https://doi.org/10.3390/app 9224889

Qiu, Y. G., Duan, H. T., Sun, J. Y., et al., 2019. Rich-Information Reversible Watermarking Scheme of Vector Maps. *Multimedia Tools and Applications*, 78(17): 24955－24977. https://doi. org/10.1007/s11042‐019‐7681-6

Qiu, Y. G., Gu, H. H., Sun, J. Y., 2017. High-Payload Reversible Watermarking Scheme of Vector Maps. *Multimedia Tools and Applications*, 77(5): 6385－6403. https://doi.org/10.1007/s11042-017-4546-8

Ren, N., Zhu, C. Q., Tong, D. Y., et al., 2020. Commutative Encryption and Watermarking Algorithm Based on Feature Invariants for Secure Vector Map. *IEEE Access*, 8: 221481－221493. https://doi. org/10.1109/access.2020.3043450

Sun, Y., Wang, S., Tang, X., et al., 2020. Adversarial Attacks on Graph Neural Networks via Node Injections: A Hierarchical Reinforcement Learning Approach. *WWW '20: The Web Conference 2020*, 2020: 673－683.

Van, B.N., Lee, S.H., Kwon, K.R., 2017. Selective Encryption Algorithm Using Hybrid Transform for GIS Vector Map. *Journal of Information Processing Systems*, 13(1):68－82. https://doi.org/10.3745/jips.03.0059

Vybornova, Y., Vladislav, S., 2019. Method for Vector Map Protection Based on Using of a Watermark Image as a Secondary Carrier. *Proceedings of the ICETE (2). Prague, Czech Republic*, 2019:284－293.

Wang, X.D., Liu, Z., Wang, N.N., et al., 2020. *Relational Metric Learning with Dual Graph Attention Networks for Social Recommendation*. PAKDD (1) 2020: 104－117.

Xia, D., Ge, Y.F., Tang, H.M., et al., 2020. Segmentation of Region of Interest and Identification of Rock Discontinuities in Digital Borehole Images. *Earth Science*, 45(11): 4207－4217(in Chinese with English abstract).

Zhai, M.G., Yang, S.F., Chen, N.H., et al., 2018. Big Data Epoch: Challenges and Opportunities for Geology. *Bulletin of Chinese Academy of Sciences*, 33(8):825－831(in Chinese with English abstract).

Zhu, D.Y., Zhang, Z.W., Cui, P., et al., 2019. Robust Graph Convolutional Networks Against Adversarial Attacks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019: 1399－1407.

Zügner, D., Akbarnejad, A., Günnemann, S., 2018. Adversarial Attacks on Neural Networks for Graph Data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 2847－2856.

附中文参考文献

江栋华,周卫, 2018.一种基于Chebyshev多项式的矢量数据几何精度脱密模型.测绘科学技术学报, 35(3):321－325.

李虎,朱恒华,花卫华,等,2020.矢量地理数据安全保护关键技术和方法.地球科学, 45(12):4574－4588.

刘会,赵波,郭嘉宝,等, 2021.针对深度学习的对抗攻击综述.密码学报, 8(2): 202－214.

夏丁,葛云峰,唐辉明,等,2020.数字钻孔图像兴趣区域分割与岩体结构面特征识别.地球科学, 45(11): 4207－4217.

翟明国,杨树锋,陈宁华,等, 2018.大数据时代:地质学的挑战与机遇.中国科学院院刊, 33(8):825－831.