计算机科学
COMPUTER SCIENCE

基于特征拓扑融合的黑盒图对抗攻击

郭宇星, 姚凯旋, 王智强, 温亮亮, 梁吉业

引用本文

郭宇星, 姚凯旋, 王智强, 温亮亮, 梁吉业. 基于特征拓扑融合的黑盒图对抗攻击[J]. 计算机科学, 2024, 51(1): 355-362.

GUO Yuxing, YAO Kaixuan, WANG Zhiqiang, WEN Liangliang, LIANG Jiye. Black-box Graph Adversarial Attacks Based on Topology and Feature Fusion [J]. Computer Science, 2024, 51(1): 355-362.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600127

基于特征拓扑融合的黑盒图对抗攻击

郭宇星[1]　姚凯旋[1]　王智强[1]　温亮亮[1]　梁吉业[1,2]
1 山西大学计算机与信息技术学院　太原 030006
2 计算智能与中文信息处理教育部重点实验室(山西大学)　太原 030006
(1135408932@qq.com)

摘　要　在大数据时代,数据之间的紧密关联性是普遍存在的,图数据分析挖掘已经成为大数据技术的重要发展趋势。近几年,图神经网络作为一种新型的图表示学习工具引起了学术界和工业界的广泛关注。目前图神经网络已经在很多实际应用中取得了巨大的成功。最近人工智能的安全性和可信性成为了人们关注的重点,很多工作主要针对图像等规则数据的深度学习对抗攻击。文中主要聚焦于图数据这种典型非欧氏结构的黑盒对抗攻击问题,在图神经网络模型信息(结构、参数)未知的情况下,对图数据进行非随机微小扰动,从而实现对模型的对抗攻击,模型性能随之下降。基于节点选择的对抗攻击策略是一类重要的黑盒图对抗攻击方法,但现有方法在选择对抗攻击节点时主要依靠节点的拓扑结构信息(如度信息)而未充分考虑节点的特征信息,文中面向引文网络提出了一种基于特征拓扑融合的黑盒图对抗攻击方法。所提方法在选择重要性节点的过程中将图节点特征信息和拓扑结构信息进行融合,使得选出的节点在特征和拓扑两方面对于图数据都是重要的,攻击者对挑选出的重要节点施加不易察觉的扰动后对图数据产生了较大影响,进而实现对图神经网络模型的攻击。在3个基准数据集上进行实验,结果表明,所提出的攻击策略在模型参数未知的情况下能显著降低模型性能,且攻击效果优于现有的方法。

关键词:图神经网络;黑盒对抗攻击;信息熵;节点重要性;引文网络
中图分类号　TP391

Black-box Graph Adversarial Attacks Based on Topology and Feature Fusion

GUO Yuxing[1],YAO Kaixuan[1],WANG Zhiqiang[1],WEN Liangliang[1] and LIANG Jiye[1,2]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing(Shanxi University),Taiyuan 030006,China

Abstract　In the era of big data,the close relationship between data is widespread,graph data analysis and mining have become an important development trend of big data technology. In recent years,as a novel type of graph representation learning tool,graph neural networks(GNNs) have extensively attracted academic and industry attention. At present,GNNs have achieved great success in various real-world applications. Lately,many researchers believe that the security and confidence level of artificial intelligence is a vital point,a lot of work focuses on deep learning adversarial attacks on Euclidean structure data such as images now. This paper mainly focuses on the black-box adversarial attack problem of graph data,which is a typical non-European structure. When the graph neural network model information(structure and parameters) is unknown,the imperceptible non-random perturbation of graph data is carried out to realize the adversarial attack on the model,and the performance of the model decreases. Applying an imperceptible no-random perturbation to the graph structure or node attributes can easily fool GNNs. The method based on node-selected black-box adversarial attack is vital,but similar methods are only taking account of the topology information of nodes instead of fully considering the information of node features,so in this paper,we propose a black-box adversarial attack for graph neural network via topology and feature fusion on citation network. In the process of selecting important nodes,this method fuses the features information and topology information of graph nodes,so that the selected nodes are significant to the graph data in both features and topology. Attackers apply small perturbations on node attributes that nodes are selected by our method and this attack has a great impact on the model. Moreover,experiments on three classic datasets show that the proposed attack strategy can remarkably reduce the performance of the model without access to model parameters and is better than the baseline methods.

Keywords　Graph neural networks,Black-box adversarial attack,Information entropy,Node importance,Citation network

到稿日期:2023-06-15　返修日期:2023-09-21
基金项目:国家自然科学基金(62272285,U21A20473)
This work was supported by the National Natural Science Foundation of China(62272285,U21A20473).
通信作者:梁吉业(ljy@sxu.edu.cn)

1 引言

在现实场景中，互联网、知识图谱、化学分子、蛋白质等都属于图数据。图数据，即包含图的数据，图数据中的节点表示事物，边表示事物之间的关系。数据之间的关系无处不在，往往比数据本身更重要。因此，近期有很多与图数据相关的任务受到了人们的关注，例如节点分类[1-2]、图分类[3]、链路预测[4]等。自图神经网络[5](Graph Neural Networks，GNNs)的概念被提出以来，越来越多的人们投入到图结构和节点属性表示的研究中，进而提出了各种 GNN 框架[1,6-7]，这在很大程度上推动了图机器学习的发展。

但是，图神经网络也面临着像其他深度学习[8-16]一样的问题，对初始样本添加非随机微小扰动导致模型性能急剧下降，这种现象被称为对抗攻击(Adversarial Attack)，由此产生的样本被称为对抗样本[8](Adversarial Examples)。在图对抗攻击中，最常见的攻击手段是对图结构和节点属性进行扰动，在扰动约束范围内使模型性能下降。这些潜在严重后果的存在给模型带来了安全隐患，因此对抗攻击成为了图深度学习模型部署前鲁棒性评估的重要手段。

目前，图对抗攻击中大多数工作是在白盒设定下开展的，攻击者了解目标模型的结构、梯度、参数等信息。但是，在现实场景中，模型的这些信息是不对外公开的，而且在信息未知的情形下攻击成功后带来的危害更大，因此本文更关注模型信息无法获知情况下的图节点分类对抗攻击任务。在目标模型未知且没有与目标模型交互的情况下，攻击者仅能得到图的拓扑结构和节点特征信息，因此可以将两者作为一种有效的攻击信息源。现有基于节点重要性的黑盒图对抗攻击方法[17-18]在寻找攻击节点的过程中主要依靠节点的拓扑结构信息，而未充分考虑节点的特征信息。

受上述启发，本文以引文网络为研究对象，针对这一类型的网络(图)数据，提出了一种基于特征拓扑融合的黑盒图对抗攻击方法。该方法包括两个步骤：1)在满足节点访问限制的前提下，找出攻击节点构成攻击节点集合；2)在满足扰动特征数量限制的情况下，利用领域知识设定扰动特征集合。在寻找攻击节点的过程中综合考虑图拓扑结构和节点特征信息，挑选出重要节点后对其特征进行微小扰动，进而实现对图神经网络模型的攻击。本文在 GCN 模型以及 3 个真实基准数据集上验证了该方法的有效性。

本文的主要贡献包括 3 个方面：

1)在扰动节点和特征数量设限的情况下，提出了基于节点重要性的对抗攻击方法，攻击者无须了解模型结构和参数等细节，该设定符合现实应用场景。

2)在选择攻击节点的过程中，将图拓扑信息和节点特征信息进行融合，提出了新的节点重要性度量指标。该指标选出的节点有助于增强对目标模型的攻击。

3)实现了上述攻击策略，并在 3 个真实基准数据集上证明了所提攻击策略能大幅降低模型性能。

本文第 2 章介绍了对抗攻击的研究现状；第 3 章介绍了基础知识和攻击节点选择策略，包括符号介绍和对抗攻击设定；第 4 章通过对比实验验证了所提方法的有效性；最后总结全文并展望未来。

2 相关工作

人工智能安全问题日益凸显，在一定程度上促进了研究者对图神经网络对抗攻击的关注。Sun 等[19]对现有图对抗攻击方法进行分类，本文对其进行简单介绍。对抗任务通常分为图节点对抗攻击[20-21]和整图对抗攻击[22]。图对抗攻击通常有以下几种攻击形式可供选择，例如攻击可以是发生在模型训练期间的中毒攻击[23-24](Poisoning Attack)，也可以是发生在模型测试期间的逃逸攻击[20,25](Evasion Attack)；攻击可能是使图中特定节点分类错误的定向攻击[20-21](Targeted Attack)，也可能是使模型整体性能下降的非定向攻击[23-24](Untargeted Attack)；对抗扰动类型可以是修改节点属性[17-18,21]，也可以是增删边[20,23]。

根据攻击者对目标模型的了解程度，将其分为：白盒攻击[26-27]、灰盒攻击[18]和黑盒攻击[17-18,20,25]。在白盒攻击中攻击者可以获取与目标模型相关的所有信息，如模型参数、梯度、输入、输出等。攻击者利用模型梯度反向最大化损失函数，生成可以使模型性能下降的对抗样本；在灰盒攻击中攻击者可以获取有限信息，如训练数据和标签，攻击者利用已知信息训练替代模型近似目标模型，进而利用成功攻击替代模型的扰动去攻击目标模型；在黑盒攻击中攻击者对模型一无所知，仅可以通过查询获得相应的标签信息，攻击者利用这些信息计算伪梯度，进而实现对模型的攻击。

在白盒图节点分类对抗攻击任务中，Zügner 等[21,24]最先提出了针对图数据的对抗攻击算法 Nettack，根据目标节点在不同候选扰动攻击后损失函数的变化程度，迭代确定下一步要扰动的结构或特征。为了进一步增强对模型整体性能的攻击，他们又提出了基于元梯度的对抗攻击算法 Metattack，此方法可以显著降低模型的整体性能；随后，Chen 等[26-27]提出的快速梯度下降法 FGA 和基于动量的梯度下降法 MGA 同样被证明是有效的。

虽然白盒攻击在理论上存在可行性，但其假设太强，需要攻击者对目标模型有较全面的了解。然而，在大多数情况下攻击者无法获知目标模型的全部信息，因此在实际应用场景中黑盒设定更贴近现实。在黑盒图节点分类对抗任务中，Dai 等[20]首次将强化学习引入对抗攻击任务，并提出了 RL-S2V 方法；在对目标模型没有任何了解的情况下，GF-attack[25]将图嵌入模型建模为一种特殊的图信号处理过程，对图信号滤波器进行攻击；Ma 等[17-18]提出了 RWCS，GC-RWCS，InfMax-Unif，InfMax-Norm 等节点选择策略，其分别通过启发式算法和最大化错误分类率选出重要节点，随后对选定节点的特征进行扰动。

本文属于图节点分类、黑盒对抗攻击、特征扰动、逃逸攻击、非定向攻击。上述介绍的攻击种类在现实中都有实际的应用场景，但黑盒设定更重要。假设攻击者意图入侵银行安保系统，作为被攻击方的银行不可能公开本行所用安保系统的相关情况，攻击者攻破银行安保系统后给银行带来的损失是巨大的。由此可见黑盒对抗攻击在现实中更常见，攻击成功后带来的危害更大，因此更具有研究价值。现有基于节点重要性的黑盒图对抗攻击方法中，Ma 等[17-18]利用 n 阶随机游走概率矩阵近似目标模型损失函数，并据此选出攻击节点。

其在寻找对抗节点时主要考虑图的拓扑结构，对图节点特征信息考虑得较少，这样选出的节点限制了攻击者对目标模型的攻击。为了找到更具代表性的攻击节点，本文从图拓扑结构和节点特征信息角度出发，提出了基于特征拓扑融合的节点选择策略，该策略选出的节点经过特征扰动后，目标模型性能下降显著。

3 特征拓扑融合的黑盒图对抗攻击

本章主要对文中使用的符号、对抗攻击设定、节点选择策略进行了介绍。

3.1 预备知识

定义 $G=(V,E)$ 为一个图，其中 $V=\{v_1,v_2,\cdots,v_N\}$ 表示节点集合，$e_{i,j}=(v_i,v_j)\in E$ 表示图中连边组成的集合。在图节点分类任务中 $\boldsymbol{X}\in R^{N\times D}$ 表示节点属性，$\boldsymbol{A}\in R^{N\times N}$ 表示图的邻接矩阵，$y\in\{1,2,\cdots,K\}^N$ 表示节点标签，其中 D 表示节点特征维度数量，K 表示节点类别数量。$N_i=\{v_j\in V\mid(v_i,v_j)\in E\}\bigcup\{v_i\}$ 表示节点 i 的邻居节点(包括节点本身)。$d_i=|N_i|$ 表示节点 i 的度，$b_i=\|\boldsymbol{X}_i\|_0$ 表示节点 i 的特征多样性。

$f:R^{N\times D}\to R^{N\times K}$ 表示图神经网络 f，其输入是节点属性 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}$，输出是所有节点的概率矩阵，并用 $\boldsymbol{H}\in R^{N\times K}$ 表示所有节点的概率矩阵，即 $\boldsymbol{H}\triangleq f(\boldsymbol{A},\boldsymbol{X})$。假定所使用的模型有 L 层，其中第 l 层($0<l<L$)表示为：

$$\boldsymbol{H}^{(l)}=ReLU(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l-1)}\boldsymbol{W}^{(l)}) \tag{1}$$

其中，$\boldsymbol{W}^{(l)}$ 表示第 l 层的可学习权重矩阵，ReLU(·)是非线性激活函数，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\mathbf{I}$，其中 $\mathbf{I}$ 是单位矩阵；$\tilde{\boldsymbol{D}}$ 为度矩阵，其中对角线上的值为 $\tilde{\boldsymbol{A}}$ 对应位置的整行值之和，即 $\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$，而且 $\boldsymbol{H}^{(0)}=\boldsymbol{X}$，$\boldsymbol{H}=\boldsymbol{H}^{(L)}$。

3.2 对抗攻击设定

限制扰动的目的是使扰动尽可能与现实场景相吻合，攻击的目的是使目标模型性能下降。本文把攻击过程分为两个阶段：1)节点选择阶段，攻击者选择节点构成攻击节点集合 $S\in V$，并对攻击节点数量进行限制 $|S|\leqslant m$，其中 $0<m\ll N$；2)特征扰动阶段，允许攻击者对节点 $\boldsymbol{X}_i\in S$ 的特征施加固定微小扰动 $\boldsymbol{\epsilon}\in R^D$，进而得到扰动后的节点 $\boldsymbol{X}_i'$：

$$\boldsymbol{X}_i'\triangleq\boldsymbol{X}_i+\boldsymbol{\epsilon} \tag{2}$$

其中，扰动向量 $\boldsymbol{\epsilon}$ 是利用领域知识[17]构建的，允许在不访问图神经网络模型的情况下确定扰动特征集。在现实中，为了使扰动效果更加明显，可以根据每个节点的自身信息为其量身定制扰动向量。但本文仍沿用 Ma 等[17]所考虑到的最坏情况，黑盒场景中攻击者无法详细获知每个节点的自身信息，因此仅对攻击集合中的节点施加固定大小的扰动。

3.3 节点选择策略

本小节将按照 3.2 节中的扰动限制提出有效的节点选择策略。该算法主要分为 5 个步骤，算法流程结合公式编号进行说明，如图 1 所示。算法的流程如下：1)为图中每条边赋予权重；2)根据边权重计算出图中每个节点的权重；3)利用节点权重计算出节点信息熵，并对其进行归一化；4)将图中每个节点的加权信息熵作为节点的重要性指标；5)将节点重要性由大到小排序，取出前 m 个节点组成攻击节点集合 S。

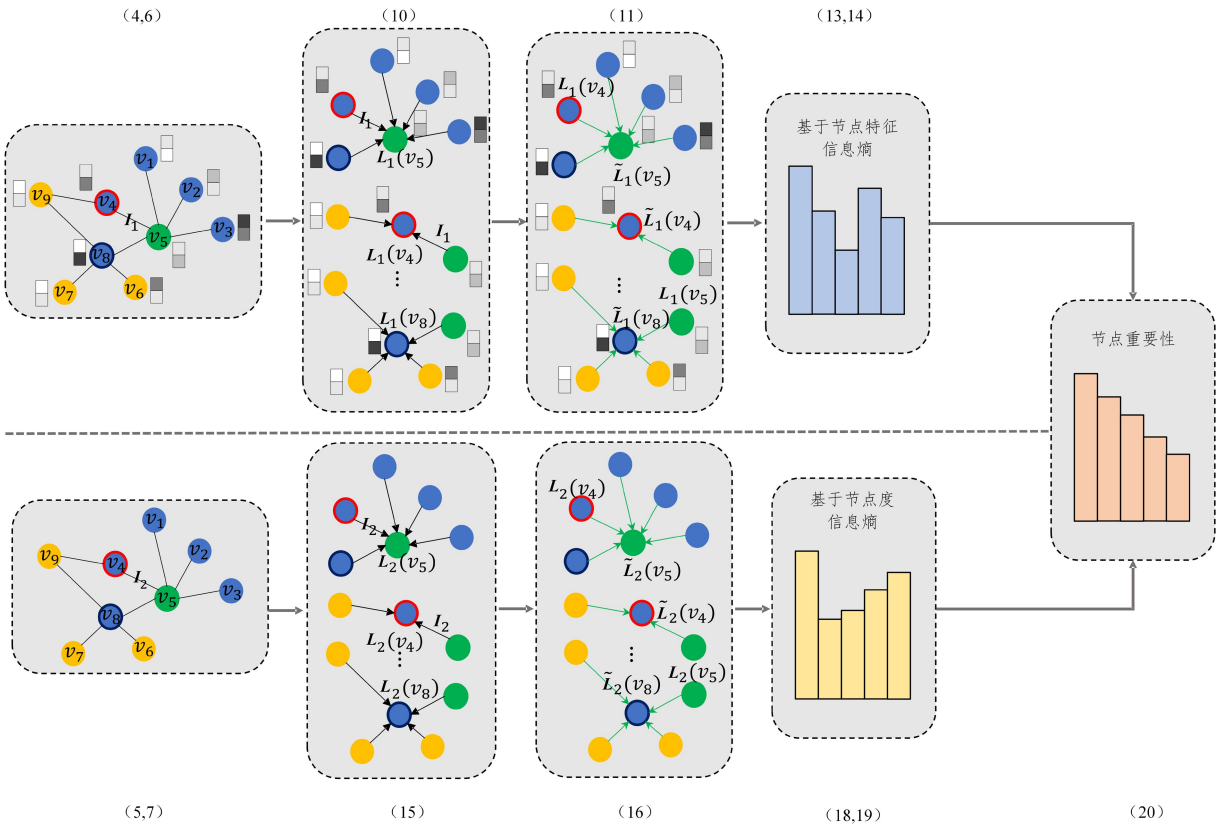


图 1 算法流程示意图

Fig. 1 Flowchart of the proposed algorithm

本文面向引文网络提出了黑盒图对抗攻击方法，在引文网络中每一个节点的特征值表示其所对应词向量的有无或词向量的出现概率。如果图中某个节点拥有的词向量越丰富，那么该节点越具有代表性。因此，在选择重要节点的过程中可以将节点特征信息纳入考虑范围。现有基于节点重要性的图对抗攻击方法在选择攻击节点时主要依靠图拓扑结构信息[28]，对节点特征信息的考虑较少。基于上述考虑，提出了特征拓扑相融合的节点选择策略。

对于给定的离散概率空间$[\boldsymbol{X},\boldsymbol{P}(x)]^{\mathrm{T}}$表示的信源，$x\in\boldsymbol{X}$为信源消息(事件)，事件$x$产生的(自)信息量定义为[29]：

$$\boldsymbol{I}(x)=\log_a\frac{1}{\boldsymbol{P}(x)}=-\log_a\boldsymbol{P}(x) \tag{3}$$

香农提出的自信息通常被用来衡量单一事件发生时所包含信息量的多寡。自信息表明，事件发生的不确定性与事件发生的概率有关。事件发生的概率越小，猜测其是否发生的难度越大，对应的不确定性越大，包含的信息量就越多；事件发生的概率越大，猜测其发生的可能性越大，对应的不确定性越小，包含的信息量就越少。以图数据中的节点v_i为例，v_i的度比图节点数量少，即直接与节点v_i相连的边很少；图中节点的特征维度比单一节点v_i真正拥有的特征数量多，即节点v_i的特征多样性差。由自信息的定义可知：节点v_i的特征与连边包含了更多有价值的信息，因此可以采用自信息构建边的权重[30]。

以图数据中节点$v_i,v_j\in V$为例，基于上述定义得到边$(v_i,v_j)\in E$的自信息，并将获得的信息作为边的权重。在基于节点特征的边权重计算中，边$(v_i,v_j)\in E$对应的概率定义为：

$$\boldsymbol{P}_1(v_i,v_j)=\frac{1}{b_ib_j} \tag{4}$$

其中，$b_i=\|\boldsymbol{X}_i\|_0$表示节点$v_i\in V$的特征丰富度，$b_j=\|\boldsymbol{X}_j\|_0$表示节点$v_j\in V$的特征丰富度。

在基于度的边权重计算中，边$(v_i,v_j)\in E$对应的概率可以定义为：

$$\boldsymbol{P}_2(v_i,v_j)=\frac{1}{d_id_j} \tag{5}$$

其中，$d_i=|N_i|$表示节点$v_i\in V$的度信息，$d_j=|N_j|$表示节点$v_j\in V$的度信息。

根据式(3)、式(4)，边$(v_i,v_j)\in E$对应的基于节点特征的自信息的计算式如下：

$$\boldsymbol{I}_1(v_i,v_j)=-\log_2\boldsymbol{P}_1(v_i,v_j)=\log_2\frac{1}{\boldsymbol{P}_1(v_i,v_j)} \tag{6}$$

其中，$\boldsymbol{P}_1(v_i,v_j)$表示基于节点特征信息计算出的边概率。

根据式(3)、式(5)，边$(v_i,v_j)\in E$对应的基于节点度的自信息的计算式如下：

$$\boldsymbol{I}_2(v_i,v_j)=-\log_2\boldsymbol{P}_2(v_i,v_j)=\log_2\frac{1}{\boldsymbol{P}_2(v_i,v_j)} \tag{7}$$

其中，$\boldsymbol{P}_2(v_i,v_j)$表示基于节点度信息计算出的边概率。

本文研究是以GCN为目标模型展开的，因此在处理图数据的方式上与原文保持一致，故边$(v_i,v_j)\in E$的权重与边$(v_j,v_i)\in E$的权重相同，有：

$$\boldsymbol{I}(v_i,v_j)=\boldsymbol{I}(v_j,v_i) \tag{8}$$

给定X是一个随机变量，它所对应的概率分布为$\vec{\boldsymbol{P}}=(p_1,p_2,\cdots,p_n)$。假设$\boldsymbol{I}(X)=(\boldsymbol{I}(x_1),\boldsymbol{I}(x_2),\cdots,\boldsymbol{I}(x_n))^{\mathrm{T}}$，那么基于式(6)、式(7)可以得到：

$$\begin{aligned}\vec{\boldsymbol{P}}\cdot\boldsymbol{I}(\boldsymbol{X})&=(p_1,p_2,\cdots,p_n)\begin{pmatrix}\boldsymbol{I}(x_1)\\\boldsymbol{I}(x_2)\\\vdots\\\boldsymbol{I}(x_n)\end{pmatrix}\\&=\sum_{i=1}^{n}p_i\boldsymbol{I}(x_i)\\&=-\sum_{i=1}^{n}p_i\log_2p_i\end{aligned} \tag{9}$$

上述定义[29]表示所有基本事件中包含信息量的期望，也叫做信息熵，因此将$\vec{\boldsymbol{P}}\cdot\boldsymbol{I}(\boldsymbol{X})$简写为$E(X)$。信息熵可以用来量化随机变量$x$中的信息，进而用信息熵来衡量节点重要性。本文使用信息熵量化节点重要性的原因是其有如下性质[29]：

性质1 假设$\boldsymbol{X}$是一个随机变量，它所对应的概率分布为$\vec{\boldsymbol{P}}$。如果$\vec{\boldsymbol{P}}$满足均匀分布，则$E(X)$达到最大值。

性质2 信息熵是一个独立于基本事件n的递增函数。

上述性质仍然适用于图数据，以节点$v_i,v_j\in V$为例，如果节点v_i的特征比图中节点v_j的特征($i\neq j$)丰富，那么节点v_i的重要性更大；如果节点v_i的度数比节点v_j的度数($i\neq j$)大，那么节点v_i的重要性更大。如果某一节点的邻居特征信息或度信息满足均匀分布，那么该节点的重要性更大[30]。基于上述说明，本文尝试从信息熵的角度衡量图节点重要性。

在计算基于节点特征信息熵之前，先计算出节点$v_i\in V$的一阶邻居边权重之和，具体计算式如下：

$$\boldsymbol{L}_1(v_i)=\sum_{v_j\in N_i}\boldsymbol{I}_1(v_i,v_j) \tag{10}$$

其中，$\boldsymbol{I}_1(v_i,v_j)$表示基于节点特征信息得到的边$(v_i,v_j)\in E$权重。

根据式(10)计算出节点$v_i\in V$的一阶邻居节点权重之和，具体计算式如下：

$$\widetilde{\boldsymbol{L}}_1(v_i)=\sum_{v_j\in N_i}\boldsymbol{L}_1(v_j) \tag{11}$$

其中，$\boldsymbol{L}_1(v_j)$表示节点$v_j\in V$的一阶邻居边权重之和。

从式(10)、式(11)可以得出，$\boldsymbol{L}_1(v_i)$反映了节点$v_i\in V$的一阶邻居边对节点的影响，而$\widetilde{\boldsymbol{L}}_1(v_i)$反映了节点$v_i\in V$二阶邻居边对节点的影响。基于上述讨论，可以将图中$v_i\in V$的邻居节点$v_j\in N_i$的特征信息贡献率定义为：

$$\boldsymbol{Q}_1(v_j)=\frac{\boldsymbol{L}_1(v_j)}{\widetilde{\boldsymbol{L}}_1(v_i)} \tag{12}$$

利用上述信息计算出基于节点特征的信息熵，具体计算式如下：

$$\boldsymbol{IE}_1(v_i)=-\sum_{v_j\in N_i}\boldsymbol{Q}_1(v_j)\log_2\boldsymbol{Q}_1(v_j) \tag{13}$$

为了便于后续信息熵的融合，在此将其进行归一化处理，具体计算式如下：

$$\overline{\boldsymbol{IE}}_1(v_i)=\frac{\boldsymbol{IE}_1(v_i)}{\sum_{j=1}^{N}\boldsymbol{IE}_1(v_j)} \tag{14}$$

上文为基于节点特征信息熵的计算流程，下面将对基于节点度的信息熵计算流程进行说明。

在计算基于节点度信息熵之前，先计算出节点 $v_i \in V$ 的一阶邻居边权重之和，具体计算式如下：

$$\boldsymbol{L}_2(v_i)=\sum_{v_j \in N_i} \boldsymbol{I}_2(v_i,v_j) \tag{15}$$

其中，$\boldsymbol{I}_2(v_i,v_j)$ 表示基于节点度信息得到的边 $(v_i,v_j) \in E$ 权重。

根据式(15)计算出节点 $v_i \in V$ 的一阶邻居节点权重之和，具体计算式如下：

$$\widetilde{\boldsymbol{L}}_2(v_i)=\sum_{v_j \in N_i} \boldsymbol{L}_2(v_j) \tag{16}$$

其中，$\boldsymbol{L}_2(v_j)$ 表示节点 $v_j \in V$ 的一阶邻居边权重之和。

根据式(15)、式(16)计算出图中 $v_i \in V$ 的邻居节点 $v_j \in N_i$ 的拓扑信息贡献率，具体计算式如下：

$$\boldsymbol{Q}_2(v_j)=\frac{\boldsymbol{L}_2(v_j)}{\widetilde{\boldsymbol{L}}_2(v_i)} \tag{17}$$

其中，$\boldsymbol{L}_2(v_j)$ 表示节点 $v_j \in V$ 的一阶邻居边权重之和，$\widetilde{\boldsymbol{L}}_2(v_i)$ 表示节点 $v_i \in V$ 的一阶邻居节点权重之和。

基于节点度的信息熵，具体计算式如下：

$$\boldsymbol{IE}_2(v_i)=-\sum_{v_j \in N_i} \boldsymbol{Q}_2(v_j)\log_2 \boldsymbol{Q}_2(v_j) \tag{18}$$

其中，$\boldsymbol{Q}_2(v_j)$ 表示基于节点度信息计算出的节点概率。

为了便于后续信息熵融合，将式(18)得到的结果进行归一化处理，具体计算式如下：

$$\overline{\boldsymbol{IE}}_2(v_i)=\frac{\boldsymbol{IE}_2(v_i)}{\sum_{j=1}^{N} \boldsymbol{IE}_2(v_j)} \tag{19}$$

其中，$\boldsymbol{IE}_2(v_i)$ 是节点 $v_i \in V$ 基于节点度得到的信息熵。

在得到基于节点度和特征的归一化信息熵后，对其进行加权求和，并将结果作为节点重要性。

$$\boldsymbol{W}(v_i)=a * \overline{\boldsymbol{IE}}_1(v_i)+(1-a) * \overline{\boldsymbol{IE}}_2(v_i) \tag{20}$$

其中，a 是一个超参数。

在得到每个节点的重要性后，将节点重要性由大到小排序，选出前 m 个节点组成攻击节点集合 S，具体计算步骤如下：

$$S=\arg\ top\text{-}m([\boldsymbol{W}(v_i)]_{i=1,2,\cdots,N}) \tag{21}$$

基于以上对各阶段的描述，本文提出的基于特征拓扑融合的黑盒图对抗攻击算法的描述如算法 1 所示。

算法 1 基于特征拓扑融合的黑盒图对抗攻击算法

输入：图数据 G(V,E)，攻击节点数 m，加权超参数 a

输出：攻击节点集合 S

1. $v_i \in V, v_j \in N_i, (v_i,v_j) \in E$：

$\mathbf{I}_1(v_i,v_j) \leftarrow b_i, b_j$

$\mathbf{L}_1(v_i) \leftarrow \mathbf{I}_1(v_i,v_j)$

$\widetilde{\mathbf{L}}_1(v_i) \leftarrow \mathbf{L}_1(v_i)$

$\mathbf{IE}_1(v_i) \leftarrow \mathbf{L}_1(v_i), \widetilde{\mathbf{L}}_1(v_i)$

$\overline{\mathbf{IE}}_1(v_i) \leftarrow \mathbf{IE}_1(v_i)$

/ * 得到基于节点特征的归一化后信息熵 * /

2. $v_i \in V, v_j \in N_i, (v_i,v_j) \in E$：

$\mathbf{I}_2(v_i,v_j) \leftarrow d_i, d_j$

$\mathbf{L}_2(v_i) \leftarrow \mathbf{I}_2(v_i,v_j)$

$\widetilde{\mathbf{L}}_2(v_i) \leftarrow \mathbf{L}_2(v_i)$

$\mathbf{IE}_2(v_i) \leftarrow L_2(v_i), \widetilde{\mathbf{L}}_2(v_i)$

$\overline{\mathbf{IE}}_2(v_i) \leftarrow \mathbf{IE}_2(v_i)$

/ * 得到基于节点度的归一化后信息熵 * /

3. $\mathbf{W}(v_i) \leftarrow \overline{\mathbf{IE}}_1(v_i), \overline{\mathbf{IE}}_2(v_i)$

$S \leftarrow \mathbf{W}$

/ * 将基于节点度和特征的归一化信息熵加权求和值作为节点重要性，并据此选出攻击节点集合 S * /

4 实验结果与分析

为了验证所提方法的有效性，以 GCN 为目标模型在 3 个真实数据集上进行实验，并与最新的基于节点重要性的攻击策略进行比较。

4.1 目标模型及数据集

本次实验中攻击的目标模型是 GCN 模型，在模型参数设定方面与文献[1]保持一致：两层模型，隐藏层大小为 32 层，学习率为 0.01，模型训练 200 次且使用 Adam 进行优化等。

本文在图神经网络常用的 3 个引文数据集 Cora，Citeseer和 Pubmed 上评估所提方法的有效性。数据集中节点表示文章，连边表示文章之间的引用关系。随着现代技术的发展，文献数量急剧增加，引文网络已经发展为一个超大规模的复杂网络，属于广义上的社会网络，因此可以用引文数据来评价攻击方法，数据集的具体情况如表 1 所列[31]。

1) Cora 数据集。Cora 数据集是一个与机器学习相关的引文网络数据集，共分为 7 类，其中有 2708 篇文章(节点)和 5429 条引用(边)。每个节点包含 1433 个特征，其中特征值表示词向量的有无。

2) Citeseer 数据集。Citeseer 数据集是学术论文的引文网络数据集，共分为 6 类，其中有 3327 篇文章(节点)和 4732 条引用(边)。每个节点包含 3703 个特征，其中特征值表示词向量的有无。

3) Pubmed 数据集。Pubmed 数据集是与生物医学相关的引文网络数据集，共分为 3 类，其中有 19717 篇文章(节点)和 44338 条引用(边)。每个节点包含 500 个特征，其中特征值表示词向量的 TF-IDF。

表 1 图数据集的基本信息统计

Table 1 Basic statistics of graph datasets

数据集	类型	节点	边	类别	特征	平均度
Cora	引文网络	2708	5429	7	1433	4.01
Citeseer	引文网络	3327	4732	6	3703	2.84
Pubmed	引文网络	19717	44338	3	500	4.50

4.2 参数设置

为了使扰动尽可能小，设定每个数据集的攻击节点数量 m 是图大小的 1%，攻击节点的特征扰动数量为单个节点特征维度的 2%。每个数据集中对特征的扰动大小是固定的，但不同数据集添加的特征扰动是不一样的，视数据集的具体情况而定。

4.3 评价指标及比较方法

本文将模型预测正确率作为实验评价指标，具体计算式如下：

$$Accuracy=\frac{\sum_{i=1}^{N} \mathbb{I}(f(x_i')=y_i)}{|V|} \tag{22}$$

其中，$\mathbb{I}(\cdot)$表示指示函数，$|V|$表示图中节点数量，$f(x_i')$表示模型对扰动图中第 i 个节点的预测。模型预测准确率越低，表明攻击方法越有效。

本文方法是基于节点重要性的攻击策略，因此在选择基准方法时主要考虑类似的工作。参照文献[25]，首先比较了从不同方面捕获节点中心性的 3 种著名网络度量，即度中心性(Degree)、中介中心性(Betweenness)和网页中心性(PageRank)，并以此命名攻击策略。其次，与当前主流的基于节点重要性的攻击方法进行比较，它们分别是 Random Walk Column Sum (RWCS)，Greedily-Corrected RWCS (GC-RWCS)，InfMax-Unif，InfMax-Norm。下面我们将简要介绍所采用的基准方法。

度中心性、中介中心性和网页中心性是典型的节点中心性算法。RWCS 和 GC-RWCS 是文献[17]中提出的黑盒攻击策略，其中 RWCS 是对最大交叉熵分类损失的近似，目的在于从图中选出重要性大的节点。GC-RWCS 是在 RWCS 方法的基础上得到的，它利用启发式方法动态更新 RWCS 的重要性分数，这样选出的节点影响力更大。基于攻击相似性的思想，该方法不再考虑选定节点的邻居。InfMax-Unif 和 InfMax-Norm 是文献[18]中提出的黑盒攻击策略。这两种方法的核心思想是将线性阈值模型上建立的原始攻击问题转换为影响最大化问题。由于该工作是在黑盒模型设定下进行的，攻击者无法直接从模型中获得阈值 θ，因此从设定的分布中对 θ 进行随机抽取。其中在均匀分布中抽取 θ 的方法命名为 InfMax-Unif，在正态分布中抽取 θ 的方法命名为 InfMax-Norm。

表 2 攻击性能总结

Table 2 Summary of attack performance

(%)

方法	Cora	Citeseer	Pubmed
None	86.32	76.39	86.33
Degree	71.72	69.47	75.78
PageRank	67.65	61.05	73.90
Betweenness	73.57	64.51	73.24
Random	79.67	72.33	80.70
GC-RWCS	65.25	66.02	71.16
RWCS	67.65	61.05	72.71
InfMax-Unif	65.80	67.97	72.43
InfMax-Norm	65.99	67.07	71.09
ours	**62.85**	**60.15**	**70.05**

4.4 实验结果及分析

本文在引文数据集上设置了 4 组实验。

实验 1 为验证所提方法的有效性，将其与上述 8 种方法进行比较，实验结果如表 2 所列。由表 2 可知，与现有基于节点重要性的方法相比，本文取得了较好的性能，表明该节点选择策略能充分提取图拓扑结构和节点特征信息，进而选出对图数据影响大的节点集合。

实验 2 为验证所提节点选择策略的有效性，分别将本方法所选不同区域的节点的特征进行攻击并输入目标模型中，对攻击后的模型性能进行比较，实验结果如图 2 所示。实验 2 是在同一个数据集中利用本文提出的节点选择策略选出前 0%～1%、前 1%～2%和前 2%～3%的节点后(它们之间相互独立并无包含关系)，将其分别组成节点数量相同的攻击节点集合，并对其特征施加攻击。实验结果表明，本方法选出的前 0%～1%的节点构成的集合优于其他节点集合，印证了本文方法的有效性。

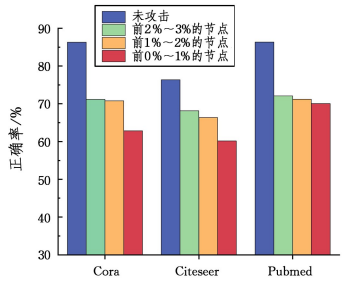
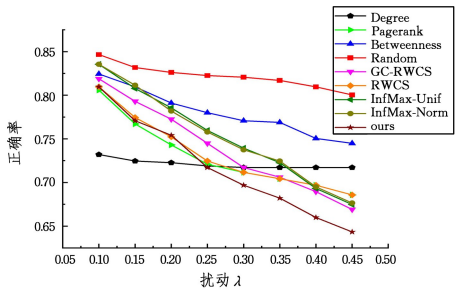


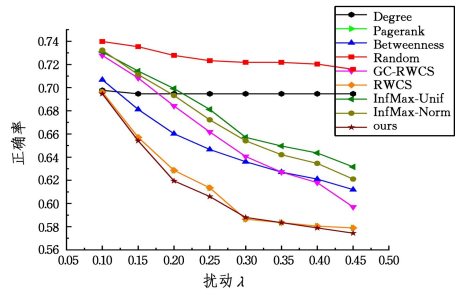
图 2 不同攻击节点集下模型正确率的变化

Fig. 2 Changes of model's accuracy on different attack sets

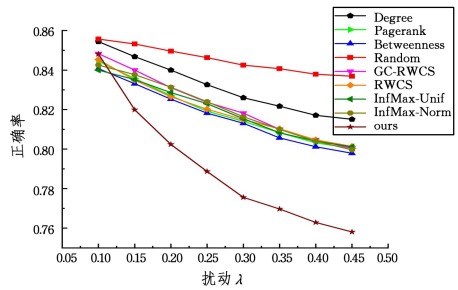
实验 3 为验证所提方法在不同扰动下对模型性能的影响，将所提方法与上述 8 种方法进行比较，实验结果如图 3 所示。实验结果表明，在一定扰动范围内所提方法在降低模型性能方面具有优势。



(a)Cora 数据集



(b)Citeseer 数据集



(c)Pubmed 数据集

图 3 不同数据集下扰动大小对模型正确率的影响

Fig. 3 Influence of perturbation on model's accuracy in different datasets

实验 4 为验证节点特征多样性对攻击的影响，将特征丰富性不同的节点组成攻击节点集合，对集合中节点特征

施加扰动后输入模型中，观察攻击后模型的性能变化，实验结果如图 4 所示。从实验结果可以看出，特征丰富度大的节点集合对模型的攻击性显著优于特征丰富度小的节点集合，这证明了在选择重要节点的过程中考虑节点特征信息的必要性。

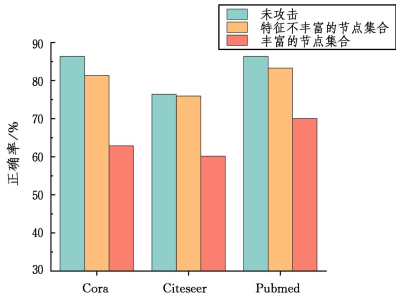


图 4 节点特征丰富性对攻击的影响

Fig. 4 Influence of node feature diversity on attack

结束语 对抗攻击在模型鲁棒性评估中具有重要意义。本文在现实扰动限制下，面向引文网络提出了一种基于特征拓扑融合的黑盒图对抗攻击方法。现有基于节点重要性的黑盒图对抗攻击方法在寻找攻击节点的过程中主要依靠节点拓扑结构信息，未充分考虑节点特征信息，而本文在选择攻击节点时将图拓扑结构信息和节点特征信息进行综合考虑，提出了新的节点重要性指标，该指标选出的节点在受到微小扰动后能使模型性能急剧下降。在 3 个基准数据集上进行实验，结果表明，所提出的攻击策略在模型参数未知的情况下能显著降低模型性能，且攻击效果优于现有的方法。图深度学习模型鲁棒性是一个重要问题，这项工作为之后更深入的研究提供了一定参考价值。目前该方法仅在引文网络方面取得进展，未来我们将进一步扩展模型的应用领域，使其能在更大范围内发挥作用。

参 考 文 献

[1] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]// Proceedings of the 5th International Conference on Learning Representations. Openreview, 2017.

[2] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]// Proceedings of the 6th International Conference on Learning Representations. Openreview, 2018.

[3] XUAN Q, WANG J H, ZHAO M H, et al. Subgraph networks with application to structural feature space expansion [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(6): 2776-2789.

[4] ZHU Z C, ZHANG Z B, XHONNEUX L P, et al. Neural bellman-ford networks: a general graph neural network framework for link prediction[C]// Proceedings of 35th Conference and Workshop on International Conference on Machine Learning. New York, NY: ACM, 2021: 29476-29490.

[5] 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.

[6] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and deep locally connected networks on graphs[C]// Proceedings of the 1st International Conference on Learning Representations. Openreview, 2014.

[7] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Proceedings of 30th Conference and Workshop on Neural Information Processing Systems. New York, NY: Curran Associates, 2016: 3837-3845.

[8] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// Proceedings of the 3rd International Conference on Learning Representations. Openreview, 2015.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]// Proceedings of the 1st International Conference on Learning Representations. Openreview, 2014.

[10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// Proceedings of the 6th International Conference on Learning Representations. Openreview, 2018.

[11] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]// Proceedings of the 6th International Conference on Learning Representations. Openreview, 2018.

[12] CHENG M H, LE T, CHEN P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[C]// Proceedings of the 6th International Conference on Learning Representations. Openreview, 2018.

[13] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]// Proceedings of 10th ACM Workshop on Artificial Intelligence and Security. New York, NY: ACM, 2017: 15-26.

[14] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box Adversarial Attacks with Limited Queries and Information[C]// Proceedings of the 6th International Conference on Learning Representations. Openreview, 2018: 2142-2151.

[15] LIU H, ZHANG Z H, XIA X F, et al. A Fast Black Box Boundary Attack Algorithm Based on Geometric Detection[J]. Journal of Computer Research and Development, 2023, 60(2): 435-447.

[16] CHEN J Y, CHEN Z Q, ZHENG H B, et al. Black-box physical attack against road sign recognition model via PSO [J]. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2785-2801.

[17] MA J Q, DING S R, MEI Q Z. Towards more practical adversarial attacks on graph neural networks[C]// Proceedings of 34th Conference and Workshop on Neural Information Processing Systems. Massachusetts, MA: MIT Press, 2020: 3837-3845.

[18] MA J Q, DENG J W, MEI Q Z. Adversarial Attack on Graph Neural Networks as An Influence Maximization Problem[C]// Proceedings of the 15th ACM International Conference on Web Search and Data Mining. New York, NY: ACM, 2022: 675-685.

[19] SUN L C, DOU Y T, YANG C, et al. Adversarial Attack and Defense on Graph Data: A Survey [J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 7693-7711.

[20] DAI H J, LI H, TIAN T, et al. Adversarial attack on graph structured data[C]// Proceedings of the 35th International Conference on Machine Learning. New York, NY: ACM, 2018: 1115-1124.
[21] ZÜGNER D, AKBARNEJAD A, GÜNNEMANN S. Adversarial attacks on neural networks for graph data[C]// Proceedings of 28th International Joint Conference on Artificial Intelligence Best Sister Conferences. San Francisco, CA: Margan Kaufmann, 2019: 6246-6250.
[22] TANG H T, MA G X, CHEN Y R, et al. Adversarial attack on hierarchical graph pooling neural networks[J/OL]. https://arxiv.org/abs/2005.115 60.
[23] SUN Y W, WANG S H, TANG X F, et al. Node Injection Attacks on Graphs via Reinforcement Learning[J/OL]. https://arxiv.org/abs/1909.06543.
[24] ZÜGNER D, GÜNNEMANN S. Adversarial Attacks on Graph Neural Networks via Meta Learning[C]// Proceedings of the 7th International Conference on Learning Representations. Openreview, 2019.
[25] CHANG H, RONG Y, XU T Y, et al. A restricted black-box adversarial framework towards attacking graph embedding models [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2020: 3389-3396.
[26] CHEN J Y, WU Y Y, XU X H, et al. Fast Gradient Attack on Network Embedding[J]. arXiv: 1809.02797, 2018.
[27] CHEN J Y, CHEN Y X, ZHENG H B, et al. MGA: Momentum Gradient Attack on Network [J]. IEEE Transactions on Computational Social Systems, 2020, 8(1): 99-109.
[28] LIU S H, CAO H Y. The Self-Information Weighting-Based Node Importance Ranking Method for Graph Data [J]. Entropy, 2022, 24(10): 1471.
[29] GRAY R M. Entropy and Information Theory [M]. Berlin: Springer, 2011.
[30] ZAREIE A, SHEIKHAHMADI A, FATEMI A. Influential nodes ranking in complex networks: An entropy-based approach [J]. Chaos, Solitons & Fractals, 2017, 104: 485-494.
[31] YANG Z L, COHEN W W, SALAKHUTDINO-V R. Revisiting Semi-Supervised Learning with Graph Embeddings[C]// Proceedings of the 33rd International Conference on Machine Learning. New York, NY: ACM, 2016: 40-48.

GUO Yuxing, born in 1998, postgraduate. His main research interests include machine learning and data mining.

LIANG Jiye, born in 1962, Ph.D, professor, Ph. D supervisor, is a member of CCF(No. 06906F). His main research interests include artificial intelligence and machine learning.

(责任编辑:喻藜)

CCF 珠海科普进校园:携手南外文华学子,共探计算机科学奥秘

2023 年 12 月 12 日 CCF 珠海会员活动中心开展了第三期“大手拉小手,科普一起走”活动,走进了深圳南外文华学校。

基于国家人才发展战略的需要,为满足中学生对计算机领域的学习需求,通过科普讲座和实践活动,激发青少年对计算机科学的兴趣,培养他们的创新能力和实践能力。12 月 12 日,CCF 珠海主席、暨南大学方俊彬教授应邀为深圳市南山外国语学校(集团)文华学校开展计算机科普讲座,文华学校校长胡丹、文华学校师生等近 300 人参加了本次讲座。

正式讲座环节,方俊彬以《计算、计算机与计算思维》作为主题,开展了一场别开生面的“计算机领域”科普演讲。方俊彬深入浅出地介绍了尖端前沿的计算机领域知识。以“数手指”“加减乘除”等贴近孩子们学习生活的概念导入,介绍了“何为计算”“计算机发展”“计算思维”三个内容丰富有趣的主题。方教授用生动有趣的语言和例子,将话题从生活中的珠算一直延伸到了深邃的量子世界,并展示了 AIGC 强大的内容创作能力,现场邀请同学上台进行 AIGC 小实验。他的分享激起了同学们的好奇心和求知欲,在文华学校学子的内心种下了一颗向往科技、求索真知的种子。

在讲座的最后,同学们展现出了对计算机和计算思维浓厚的兴趣,踊跃提问,问题涵盖了讲座内容的各个方面,既有深入的专业问题,也有充满童趣的好奇心。每一个问题都显示出他们对知识的渴望和对未知世界的好奇。方俊彬教授耐心地解答每一个问题,他的解答深入浅出,让同学们在轻松愉快的氛围中获得了新的知识和启发。

此次活动是 CCF 珠海会员活动中心积极探索科普活动走进中小学的有益实践,有利于计算机知识的普及和发展及提升中学生科学素养和创新精神。

据 CCF 微信公众号