计算机科学
COMPUTER SCIENCE

基于图卷积神经网络的节点分类方法研究综述

张丽英, 孙海航, 孙玉发, 石兵波

引用本文

张丽英, 孙海航, 孙玉发, 石兵波. 基于图卷积神经网络的节点分类方法研究综述[J]. 计算机科学, 2024, 51(4): 95-105.

ZHANG Liying, SUN Haihang, SUN Yufa , SHI Bingbo. Review of Node Classification Methods Based on Graph Convolutional Neural Networks [J]. Computer Science, 2024, 51(4): 95-105.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600071

基于图卷积神经网络的节点分类方法研究综述

张丽英[1]　孙海航[1]　孙玉发[2]　石兵波[3]
1 中国石油大学(北京)信息科学与工程学院　北京 102249
2 石油工业出版社有限公司　北京 100011
3 中国石油勘探开发研究院　北京 100083

摘　要　节点分类任务是图领域中的重要研究工作之一。近年来随着图卷积神经网络研究工作的不断深入,基于图卷积神经网络的节点分类研究及其应用都取得了重大进展。图卷积神经网络是基于卷积发展出的一类图神经网络,能处理图数据且具有卷积神经网络的优点,已成为图节点分类方法中最活跃的一个研究分支。对基于图卷积神经网络的节点分类方法的研究进展进行综述,首先介绍图的相关概念、节点分类的任务定义和常用的图数据集;然后探讨两类经典图卷积神经网络——谱域和空间域图卷积神经网络,以及图卷积神经网络在节点分类领域面临的挑战;之后从模型和数据两个视角分析图卷积神经网络在节点分类任务中的研究成果和未解决的问题;最后对基于图卷积神经网络的节点分类研究方向进行展望,并总结全文。

关键词:图数据;节点分类;图神经网络;图卷积神经网络
中图分类号　TP391

Review of Node Classification Methods Based on Graph Convolutional Neural Networks

ZHANG Liying[1],SUN Haihang[1],SUN Yufa[2] and SHI Bingbo[3]
1 College of Information Science and Engineering,China University of Petroleum(Beijing),Beijing 102249,China
2 Petroleum Industry Press,Beijing 100011,China
3 Research Institute of Petroleum Exploration & Development,Beijing 100083,China

Abstract　Node classification is one of the important research tasks in graph field. In recent years,with the continuous deepening of research on graph convolutional neural network,significant progress has been made in the research and application of node classification based on graph convolutional neural networks. Graph convolutional neural networks are kind of graph neural network method based on convolution. It can handle graph data and have the advantages of convolutional neural networks,and have become the most active branch of graph node classification research. This paper first introduces the related concepts of graph,the definition of node classification and commonly used graph datasets. Then,it reviews two classic graph convolutional neural networks,spectral domain and spatial domain graph convolutional neural networks,and discusses the challenges of using graph convolutional neural networks to study node classification. Next,it analyzes the research progress and unresolved issues of graph convolutional neural networks in node classification tasks from the perspectives of model and data. Finally,this paper gives insights into the research direction on node classification based on graph convolutional neural networks.

Keywords　Graph structure data,Node classification,Graph neural network,Graph convolutional neural network

1　引言

图是一组点和边的集合,其中“点”表示实体,“边”表示实体间的关系[1]。图数据蕴含的信息丰富,很多问题都可以通过图模型进行求解。因此,图数据一直是研究热点之一。

围绕图数据的研究内容很多,包括图节点分类、图分类、图聚类、链路预测等任务[1]。其中,节点分类是图领域中重要的研究内容之一,常用于衡量模型学习节点表示的能力。节点分类任务是对无标签的节点进行分类识别[2]。该任务在社交网络用户分类、垃圾邮件发送者检测、文献类型确定等方面具有广泛的应用场景。除此之外,节点特征获取方法也可以应用于节点聚类、链接预测和可视化等任务中。近年来随着深度学习模型的成功,受卷积神经网络强大建模能力的启发,图卷积神经网络(Graph Convolutional Network,GCN)已经成为图数据的节点分类建模方法中最重要、最活跃的一个研究分支。

节点分类常用数据集有 Cora[2],CiteSeer[2],Pubmed[3],Reddit[4]和 PPI[5]等。表 1 列出了 5 个数据集的统计信息。其中 Cora,CiteSeer,PubMed 和 PPI 为小规模数据集[6-7],Reddit 为大规模数据集[8]。PPI 为异构图,其他 4 个数据集为同构图。Reddit 为动态图,其他 4 个数据集为静态图。

到稿日期:2023-06-08　返修日期:2023-09-13
通信作者:张丽英(lyzhang1980@cup.edu.cn)

表 1 常用数据集

Table 1 Common datasets

数据集	类型	节点数	边数	特征数	类别数	数据集规模	同/异构图	静/动态图
Cora[2]	引文网络	2708	5429	1433	7	小规模	同构图	静态图
CiteSeer[2]		3312	4552	3703	6	小规模	同构图	静态图
PubMed[3]		19717	44324	500	3	小规模	同构图	静态图
Reddit[4]	社交网络	232965	11606919	602	41	大规模	同构图	动态图
PPI[5]	生物化学结构	56944	818716	50	121	小规模	异构图	静态图

本文对基于 GCN 的节点分类研究进行综述，梳理已有研究开展的工作情况，分析总结目前节点分类中待解决的问题及存在的挑战，并对未来可能的研究方向进行了展望。

2 经典的 GCN 模型及其在节点分类中面临的挑战

2.1 经典的 GCN

GCN 是一种基于图的深度学习模型，它主要通过在图上进行卷积运算来提取节点的表示，并通过这些表示进行分类、聚类等任务。与传统卷积神经网络不同，GCN 的卷积层采用邻接矩阵、度矩阵等方式来定义图的结构，从而实现对图上节点特征的卷积操作。

图卷积神经网络是成功地将深度学习的模型移到图这类非欧氏数据结构上的一种神经网络模型。如何在图结构上构建卷积算子，是图卷积神经网络建模研究的焦点。根据卷积核的定义方法，图卷积神经网络可划分为谱域 GCN 和空域 GCN 两大类。

2.1.1 谱域 GCN

Bruna 等[9]于 2014 年提出第一个图卷积神经网络 Spectral CNN，它定义过滤器 $g_\theta=\boldsymbol{\Theta}_{i,j}^{(l)}$，其中，$\boldsymbol{\Theta}_{i,j}^{(l)}$ 是可学习的参数对角阵。Spectral CNN 的图卷积层定义如下：

$$\boldsymbol{H}_{:,j}^{(l)}=\sigma\left(\sum_{i=1}^{f_{l-1}}\boldsymbol{U}\boldsymbol{\Theta}_{i,j}^{(l)}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{H}_{:,j}^{(l-1)}\right),j=(1,2,\cdots,f_l) \tag{1}$$

其中，l 表示层数，$\boldsymbol{H}_{:,j}^{(0)}=X_{:,j}$ 表示节点状态向量第 j 维的值，f_{l-1} 表示第 $l-1$ 层通道数，f_l 表示第 l 层通道数。

Bruna 等提出的模型为谱域 GCN 的发展奠定了基石，实现了 GCN 质的飞跃，但其存在着非空间局部化和密集计算开销大的问题。为此，Henaff 等[10]引入了参数化的具有平滑系数的插值卷积核，以使其在空间上局部化，且减少了参数个数。

Duvenaud 等[11]则设计了一种新的卷积核，通过采用图拉普拉斯算子的切比雪夫展开来近似卷积核，构建了切比雪夫网络(Chebyshev network，ChebNet)。ChebNet 避免了计算拉普拉斯特征向量的步骤，极大地降低了计算复杂度，提高了计算效率，并具有一定的空间局部性。为了进一步提高 ChebNet 性能，使之具有更好的空间局部连接特性，Kipf 等[1]提出了一阶图卷积神经网络来简化 ChebNet，同时也大大简化了 GCN 的计算量。

为了进一步加速卷积过程，SGC[12]去除了连续 GCN 层之间的非线性过渡函数，提高了计算效率的同时仍能达到与传统 GCN 相当的性能。JK Nets[13]和 MixHop[14]定义了多跳图卷积，以直接访问多跳以外的邻居节点。然而，谱域 GCN 模型中的大多数模型学习滤波器依赖于整个图结构，其网络结构不能太大、太深，且计算效率较低。表 2 列出了各谱域 GCN 模型。

表 2 谱域 GCN 模型

Table 2 Spectral domain GCN model

模型	年份	数据集	开展的工作
频谱图卷积神经网络[9]	2014	MNIST	解决卷积操作如何在图上进行以及减少在图上卷积的时间复杂度
ConvNet[10]	2015	Reuters，ImageNet 等	将模型拓展到大规模数据任务上
Chebnet[11]	2016	Fingerprint	对拉普拉斯矩阵的特征值操作转为矩阵的整体进行，降低时间复杂度
简化 ChebNet[1]	2016	MNIST	提出一阶图卷积神经网络简化 ChebNet，降低 GCN 的计算量
SGC[12]	2019	Cora，Citeseer，Pubmed 等	去除连续 GCN 层之间的非线性过渡函数，提高计算效率
JK Nets[13]	2018	Cora，Citeseer，Reddit 等	利用每个节点的不同邻域范围，实现更好的结构感知表示
MixHop[14]	2019	Cora，Citeseer，Pubmed	通过重复混合不同距离上邻居的特征表示来学习一般的邻域混合关系

2.1.2 空间域 GCN

空间域 GCN 则从图数据结构的节点域出发，设计神经网络学习聚合函数，聚集每个中心节点的邻居节点特征，使用消息传播机制，来研究如何高效、准确地利用中心节点的邻居节点来学习表征中心节点的特征。空间 GCN 具有自适应具体任务和图数据结构的优点，灵活性更大。表 3 列出了各空域 GCN 模型。

最早在 2009 年 Micheli[15]就提出了空间域方法 NN4G，它由多层卷积操作组成，在每一层，NN4G 通过直接对邻居节点的特征加权求和来实施卷积操作，通过叠加神经网络层数，模型可以抓取高阶的邻居信息。通过引入残差连接和跳跃连接，NN4G 避免了网络层数较深引起的梯度消失问题。具体地，NN4G 每一层的更新过程如下：

$$\boldsymbol{h}_i^{(l)}=\sigma\left(\boldsymbol{W}_l^{\mathrm{T}}\boldsymbol{x}_v+\sum_{t=1}^{l-1}\sum_{j\in N(i)}\boldsymbol{\Theta}_t^{\mathrm{T}}\boldsymbol{h}_j^{l-1}\right) \tag{2}$$

其中，$\sigma(\cdot)$ 表示激活函数，$\boldsymbol{W}_l$ 和 $\boldsymbol{\Theta}_t$ 是每层的系数矩阵，$\boldsymbol{h}_i^{(0)}=0$。

另一项对空间方法具有重要意义的研究是 Gilmer 等提出的 MPNN[16]，它建立了一个空间方法的统一框架并被后续许多方法沿用。MPNN 认为图卷积可以视为节点之间通过相连边进行信息传递的过程。它的节点信息更新过程如下：

$$\boldsymbol{h}_i^{(l)}=\boldsymbol{F}_1\left(\boldsymbol{h}_i^{(l-1)},\sum_{j\in N(i)}\boldsymbol{F}_2\left(\boldsymbol{h}_i^{(l-1)},\boldsymbol{h}_j^{(l-1)},\boldsymbol{x}_{ij}^e\right)\right) \tag{3}$$

其中，$\boldsymbol{F}_1(\cdot)$ 和 $\boldsymbol{F}_2(\cdot)$ 表示包含可学习参数的函数，$\boldsymbol{h}_i^0=\boldsymbol{x}_i$。

除此之外，Hechtlinger 等[17]提出了 Graph CNN。该模型通过遍历图结构数据，将每一个节点作为中心节点，以基于

随机游走的概率转移矩阵为基础，为每个节点选择相连的前 k 个最近邻居组成邻域，将节点的邻居信息进行聚合，以更新节点的特征表示。

由于以上模型计算时需要将整个网络的节点特征加载进来，不便应用于大规模网络上。为此，Hamilton 等[8]提出了图采样和聚合的模型 GraphSAGE，对邻居节点做随机采样，保证每个节点的邻居节点都不多于给定的采样个数。Velickovic 等[18]则进一步考虑了聚合过程中不同邻居节点对中心节点特征的贡献的差异性，从而提出了图注意力网络(Graph Attention Network，GAT)，GAT 通过注意力机制将节点的特征表达融入到聚合函数的定义中，节点的权重计算以加权和的方式聚合到中心节点。ASGCN[19]通过自适应采样控制 GCN 训练的采样大小。Cluster-GCN[20]与 Fastgcn[21]则进一步发展了图采样和图聚类，提出了多重采样的空间域 GCN 模型。然而，上面提到的 GCN 模型都没有明确评估相邻节点的质量，也没有研究如何通过细化图结构来提高 GCN 模型的准确性。为此，Hao 等提出了邻域增强图卷积网络 NEGCN[22]。在传统的 GCN 中，每个节点的邻居节点的特征都被平均池化并与节点自身的特征进行卷积运算。然而，这种简单的聚合方法可能会丢失有关节点邻域结构的重要信息。为了解决这个问题，NEGCN 引入了邻域增强机制。具体来说，NEGCN 利用自适应邻域采样方法，对每个节点的邻居节点进行重要性采样，这样可以选择性地保留重要的邻居节点进行卷积运算，同时降低了计算复杂度。在训练过程中，NEGCN 使用节点特征和邻居节点特征来学习每个节点的表示。通过多层的卷积操作，NEGCN 可以逐渐捕捉到全局的图结构特征，从而更好地完成节点分类、图分类等图数据任务。NEGCN 相较于传统 GCN 的优势是能够更好地利用节点的邻域信息，提高模型的表示能力和性能，从理论上分析邻居质量如何影响 GCN 模型的分类性能，并专门设计邻居评估方法以提高邻居质量。

以上这些经典的图卷积神经网络模型在图节点分类任务上取得了突出的成果，但在实际应用时仍面临着巨大挑战，推动着图卷积神经网络的研究不断向前发展。

表 3 空间域 GCN 模型

Table 3 GCN model in spatial domain

模型	年份	数据集	开展的工作
NN4G[15]	2009	化合物数据集	将监督神经网络的输入域扩展到一般的图类
MPNN[16]	2017	QM9	通过在图结构上进行消息传递和图更新，来捕捉分子图的局部和全局信息，并将其映射到相应的性质预测上
Graph CNN[17]	2017	MNIST	定义了节点间的邻居关系和聚合方式，以实现有效的特征传播和表示学习
GraphSAGE[8]	2017	Citation，Reddit，PPI	通过邻居采样解决 GCN 内存爆炸问题，适用于大规模图
GAT[18]	2017	Cora，Citeseer，PPI 等	通过为同一邻域的节点分配不同权重来扩充模型尺度
ASGCN[19]	2018	Cora，Pubmed，Reddit 等	开发一种分层采样器来加速 GCNs 的训练
NEGCN[22]	2022	Cora，Pubmed，Reddit 等	引入高效的边缘分类器显式地识别有用的邻居，并修改图结构提高邻居质量

2.2 GCN 在节点分类中面临的挑战

GCN 的优点在于可以捕捉图的全局信息，从而很好地表示节点的特征。但 GCN 在节点分类任务上也面临着一些挑战，下面分别从模型视角和数据集视角来综述 GCN 在节点分类中存在的问题。

2.2.1 模型角度

(1)增加 GCN 的深度会导致梯度消失和过平滑。梯度消失是经典神经网络中会出现的问题，即网络层数堆叠过多，前面的层梯度过小而难以更新；另一个问题是过平滑问题，在堆叠了多个图卷积层之后，每个节点能覆盖到的节点会收敛到全图节点，最终会导致不同类节点的特征向量相近，产生过平滑现象，这一现象极大地限制了 GCN 的性能[23]。

(2)GCN 对动态图支持不够。谱域 GCN 的信息聚合方式是全局聚合，其经过训练得到的卷积核都依赖于 Laplacian 矩阵分解后的特征值和特征向量。Laplacian 矩阵是通过图的邻接矩阵变化而来，导致针对某一个指定图结构训练的模型并不能扩展到其他不同结构的图上，若是新增节点，整个图会发生变化，那么 GCN 的结构就会发生变化，不适用于动态图[8]。

(3)适用异构图的 GCN 研究有待深入。现实生活很多真实数据为异构图，包含多种类型的节点以及多种类型节点之间的关系。对由相同类型的节点和边组成的同构图使用 GCN 取得了很大进展，而对具有不同类型的节点和边组成的异构图使用 GCN 的研究还处于起步阶段。如社会网络多元化、复杂化数据的处理和分析都是基于异构图，需要构建适用异构图的 GCN 模型实现更深层地挖掘用户信息[24]。

(4)GCN 难以扩展到现实应用中的大型图中。GCN 需要将整个图放到内存和显存中，这非常耗内存和显存，因此，其难以扩展到现实应用中的大型图，例如社区检测通常包含数百万个节点和边[8]。

2.2.2 数据集质量方面

(1)GCN 使用原始固定图结构数据进行图卷积等操作时，数据集质量会直接影响 GCN 模型完成下游任务的效果。对于带有图结构的数据来说，其数据在采集时会不可避免地出现差错，导致收集到的数据通常会带有噪声、缺失等，从而影响模型性能[25]。

(2)高度不平衡的图数据对基于 GCN 的节点分类具有挑战性，现在大多数分类失衡的数据集都集中在 1∶4 至 1∶100的失衡比率上。在欺诈检测或化学信息学等现实应用中，可能会处理不平衡率从 1∶1 000 到 1∶5 000 的问题。当处理一个不平衡的类分布时，GCN 倾向于偏向大多数类的节点，而少数类的节点代表不足[26]。

针对 GCN 面临的挑战，近些年来陆续有深入的研究工作推动 GCN 在节点分类任务上不断取得进展。

3 GCN 针对模型问题的研究进展

3.1 梯度消失和过平滑问题

针对 GCN 增加深度会导致梯度消失和过平滑问题的研究进展如表 4 所列。这些方法可概括为设计更优深层网络的方法和设计更优提取特征的方法。

表 4 梯度消失和过平滑问题的研究

Table 4 Research on gradient disappearance and over smoothing

方法类型	代表工作	基本思想
设计更优深层网络方法	DeepGCN[27], AdaGCN[28]	设计更优将前一层的特征连接到下一层的方法，提升梯度有效传递和网络的表达能力
设计更优特征提取方法	Cluster-GCN[20], N-GCN[29]	设计更优特征提取方法，如归一化的方式与划分子图等

3.1.1 设计更优深层网络方法

DeepGCN 之前的 GCN 模型深度一般不超过 4 层，其每一层一般可以表示为式(4)[27]：

$$G_{l+1}=F(G_l,W_l)=Update((Aggregate(G_l,W_l^{agg}),W_l^{update}) \tag{4}$$

为加深图网络的深度，提升模型的表达能力，在 ResNet、DenseNet 和膨胀卷积的启发下，研究者对 GCN 进行了一系列改进升级，提出了深度更深、更加稳定、表现更好的图网络。

Kipf 等[1]认为过度平滑现象是由于图神经网络算法的原理中过分强调邻域内其他节点的特征对目标节点的影响，忽略了目标节点本身的特征信息造成的影响，于是提出了 Res-GCN。在 GCN 的基础上，Kipf 为每一层增加了残差连接，这些连接将为信息和梯度的传输提供额外的连接通道，有效解决了梯度消失的问题。即便使用了残差连接，GCN 也不可能做得太深，基本在 3 层到 5 层左右。这是因为 GCN 可以被看作低通滤波器，叠加低通滤波器具有明显的过度平滑现象。

Huang 等[30]受 DenseNet 启发，为每层图卷积衔接先前所有中间层的信息。DenseGCN[27]结合稠密连接设计了一种更为高效的特征共享方式和信息流动通道，有效融合了多级别的特征，为梯度的流动提供了良好的通道，进一步促进了特征的复用，缓解了梯度消失问题。图像领域的研究表明，膨胀卷积可以在不损失分辨率的情况下有效扩大模型感受野，研究人员通过 KNN 的方式来寻找每一层 GCN 后需要膨胀的邻域，并构建了膨胀的图结构。例如针对一个膨胀率为 d 的图，KNN 会在输入图中每隔 d 个相邻节点来构建 $k*d$ 的计算区域并返回 KNN 结果。

Chiang 等[20]认为在 GCN 上直接使用残差连接的做法是假定目标节点与所有邻接节点之间具有相同的权重，未考虑不同节点之间的重要性。因此，他们设计了一种新的正则化方法，其思想是为距离越近的节点分配越大的权重。实验结果表明，这种新的正则化策略能够使深度 GCN 实现当时最优的性能。

Sun 等[28]集成了 Adaboost 和 GCN 层，构建了 AdaGCN 模型。AdaGCN 在所有层之间共享相同的基本神经网络结构，并进行递归优化，以获得更深层次的网络模型，能够在一定程度上解决过平滑的问题。

3.1.2 设计更优特征提取方法

不同于设计更优深层网络的策略，一些研究探索设计更优特征提取方法来解决梯度消失和过平滑的问题。

Chiang 等提出的 Cluster-GCN[20]的基本思想是使用图节点聚类算法将一个图的节点划分为若干个簇，每一次选择几个簇的节点和这些节点对应的边构成一个子图，然后对子图做训练，限制节点只能在其所属的子图中进行邻域特征的提取，避免过平滑的问题。受 Inception[31]研究工作的启发，Abu-El-Haija 等[29]从网络宽度入手，设计了一种具有优良局部拓扑结构的 N-GCN 网络结构，通过对一些较小尺寸图卷积核进行组合，并将其输出结果拼接为一个高维特征图，保证模型对图数据的表征能力，同时因为模型没有提取深层的卷积特征，避免了过平滑的问题。

3.1.3 结论

设计更优深层网络或设计更优提取特征方式，可以在一定程度上解决梯度消失和过平滑问题。未来可以设计适合深层架构的卷积核、优秀的子图划分方法等，对图数据进行预处理来更好地提取特征；也可以引入外部信息，如节点之间的关系，作为监督信息来指导训练过程，减少深度训练造成的失真。

3.2 在动态图上的应用问题

GCN 的信息聚合方式是全局聚合，扩展性非常差，若新增节点，整个图会发生变化，进而导致 GCN 的结构发生变化，因此其不适用于动态图[8]。解决该问题的研究进展可概括为两个方面，如表 5 所列。

表 5 在动态图上的应用研究

Table 5 Application research on dynamic graph

方法类型	代表工作	基本思想
引入时间维度	STGCN[32], EvolveGCN[33]	在 GCN 中加入时间维度，利用 RNN 或者 LSTM 等模型来处理时间信息
引入注意力机制	DySAT[34], TGAT[35]	引入注意力机制来处理不同时间节点的信息

3.2.1 引入时间维度的方法

在动态图中，节点的特征矩阵在时间上是变化的。因此，研究者通常会在 GCN 中加入时间维度，利用 RNN 或者 LSTM 等模型来处理时间信息。2018 年提出的 STGCN[32]网络结构引入了时空卷积模块，包括时空卷积操作以及残差连接，在每个时间步上对每个节点的特征进行更新；而全局池化模块则能够捕捉整个图的全局特征，通过平均池化等方法将所有节点的特征进行整合。相比于传统的基于 RNN 或 LSTM 的方法，STGCN 不需要考虑时间序列上的顺序，网络训练速度更快，且能够处理较长的时间序列数据。同时，STGCN 还能够灵活地处理不同的空间结构，如栅格和图等。但是，STGCN 模型设计较为复杂，需要较高的计算资源和额外的数据处理步骤，例如构建时空图结构和卷积操作。这可能限制了其在某些场景下的应用。

2019 年 Pareja 提出了 EvolveGCN 模型[33]，它通过 RNN 来演化 GCN 模型的参数，实现在时间序列上的卷积操作。在每个时间步骤中，GCN 会根据当前的节点特征和邻居节点

特征,进行一次卷积操作,以获得新的节点特征表示作为下一个时间步骤的输入,与相应的邻居节点特征一起构成新的图进行下一轮的 GCN 卷积操作。相比其他传统的图神经网络模型,EvolveGCN 模型通过在 GCN 中引入动态权重矩阵,使模型能够根据节点的历史状态来更新节点的表示,并利用节点之间的时空依赖关系来进行推断和预测。但是由于 EvolveGCN 模型需要对每个时间步都进行图卷积操作,因此计算复杂度相对较高,这会导致模型在处理大规模图数据时效率低下,并可能消耗大量的计算资源;并且由于 EvolveGCN 模型依赖于时间维度的信息,因此对于没有时间标签或者时间序列长度较短的图数据,模型的性能可能受到限制;此外,如果数据中存在时间上的不一致或缺失,则可能导致模型在时空特征建模方面表现不佳。

3.2.2 引入注意力机制的方法

在动态图节点分类任务中,不同时间节点之间的联系是不同的,一些节点可能在某些时间点上很相关,而在另一些时间点上不相关。为此,研究者引入了注意力机制来处理这种情况。

DySAT[34]通过结构邻域和时间动态两个维度,联合使用多头注意力机制捕捉多方面的动态性。它以节点的贡献率为损失函数引导模型的训练,重点关注图的结构,这对于动态性更丰富的图(如结点的特征的变化)是不够的。

Xu 等受归纳表示学习的启发,提出时间感知图注意网络 TGAT[35]。该模型使用自注意力机制作为构建模块,并基于 Bochner 定理提出了一种新颖的功能时间编码技术,建模节点嵌入识别为时间的函数且能随着图的演化归纳推断新节点和被观察节点的嵌入,来有效聚合时态和拓扑邻居特征,学习到节点的时态和拓扑邻居聚合函数,再使用归纳学习表示方法快速生成节点表示,用于处理节点分类和链接预测任务。

3.2.3 结论

综上所述,对于动态图的节点分类问题,已有工作通过引入时间维度与注意力机制来研究处理动态图中节点特征的时间变化,使之更好地适应动态图数据的特点,提高节点分类任务的性能。动态图数据往往不仅涉及节点分类、边分类等单一任务,还涉及多任务联合学习,如节点分类和边预测同时进行等。未来的研究可以探索如何设计基于 GCN 的多任务学习框架,在优化多个任务时更好地结合不同任务之间的关联信息。

3.3 在异构图上的应用问题

GCN 是以同构图为基础进行研究,但现实中很多真实数据的结构为异构图,它会包含多种类型节点和多种不同类型节点之间的关系。使用 GCN 的优势来解决异构图的节点分类研究工作可概括为两方面,如表 6 所列。

表 6 在异构图上的应用研究

Table 6 Application research on heterogeneous graph

方法类型	代表工作	基本思想
多层次信息融合方法	R-GCN[24],HAN[36]	将不同类型节点间的关系进行融合
跨领域知识迁移方法	CD-GNN[37],HGCC[38]	采用跨领域知识迁移方法来增强异构图节点分类性能

3.3.1 多层次信息融合方法

传统的 GCN 模型只考虑节点和邻居节点之间的交互,忽略了不同类型节点和边的信息。针对此问题,研究者提出了使用多个模型来融合不同类型的信息方法。

Titor 等提出异构图上多关系处理的一种模型 R-GCN[24],它将不同关系分别做融合,再将结果进行叠加处理,得到节点表示。模型通过因子分解和多关系参数共享的方式,减少了多关系引起的参数剧增问题。相比于 GCN,R-GCN 没有用度矩阵及邻接矩阵作为边的权重,而是把边的权重放在模型中,通过参数自学的方式获得。

受 R-GCN 模型的启发,Wang 等将注意力机制引入异构图中的图卷积神经网络模型,提出了 HAN 模型[36]。HAN 是一种包含节点级注意力和语义级注意力的层次注意力异质图神经网络。节点级注意力学习节点与基于元路径的相邻节点之间的重要性;语义级注意力学习不同元路径的重要性,元路径是连接两个实体的一条特定的路径。HAN 通过注意力机制将不同节点类型之间的信息进行融合,其局限性表现在元路径无法处理两节点多关系的问题。

3.3.2 跨领域知识迁移方法

在某些情况下,不同领域的异构图(如不同领域中的用户关系图[37])可以共享一些相同的特征或知识,因此可以采用跨领域知识迁移方法来提升节点分类性能。

2021 年,Liu 等提出 CD-GNN 模型[37],该模型基于图神经网络的思想进行建模。它使用不同领域中的用户关系图,其中每个领域表示不同的用户行为模式或属性。CD-GNN 通过对齐不同领域图之间的节点来捕捉用户之间的关系,并使用图神经网络进行节点表示学习。与其他图神经网络中的节点分类模型相比,CD-GNN 可以处理更复杂的异构图。

2023 年,Zhang 等提出了一种新的双曲几何图卷积神经网络(Hyperbolic Graph Convolution Networks,HGCN)[38],与传统的欧氏几何空间不同,双曲几何空间具有负曲率,可用于建模更复杂的数据集,能够很好地捕捉复杂网络中的非欧几何结构,并且能够对不同域之间的节点进行信息传递和聚合。HGCN 提高了异构图节点分类的性能和效率。

3.3.3 结论

综上所述,对于异构图的节点分类问题,研究者们通过对多层次信息融合或跨领域知识迁移的方式对 GCN 模型进行改进,提出的一些新模型可以更好地适应异构图数据的特点,从而提高节点分类任务的性能。

现实世界中许多数据都是跨域的,来自不同的领域或不同的网络,因此,如何更好地将跨域的数据进行集成和分类是一个重要的研究方向。

3.4 在大规模网络上的应用问题

GCN 需要将整个图放到内存和显存内,非常耗费时间与存储空间,因此难以扩展到现实应用中包含数百万个节点和边的大型图[8],如社区检测。针对大型图的节点分类问题,目前的研究工作可归纳为 3 类方法,如表 7 所列。

表 7 在大规模网络上的应用研究

Table 7 Application research on large scale network

方法类型	代表工作	基本思想
引入邻居采样的方法	GraphSAGE[8]	对每个节点的邻居采样
引入层采样的方法	FastGCN[21], ASGCN[19]	使用分层采样,避免邻域指数扩散
引入子图采样的方法	Cluster-GCN[20], GraphSAINT[39]	通过子图采样方法加速训练过程

3.4.1 引入邻居采样的方法

邻居采样是从每个节点的邻居中选取一部分节点,并将这些节点的特征进行聚合,生成新的节点特征。GraphSAGE[8]是 2017 年提出的在层反向传播时使用固定大小的邻居样本的图神经网络模型,它解决了 GCN 网络训练时需要用到整个图的邻接矩阵的局限性。GraphSAGE 使用多层聚合函数,每一层聚合函数会将节点机器邻居的信息聚合在一起,得到下一层的特征向量。在每一层的计算过程中,GraphSAGE 对每个节点的邻居采样,对于该层中的每一个节点,都随机从邻居集合中采样出聚合节点集。式(2)中采样数量为 k,若节点邻居数少于 k,则采用有放回的抽样方法,直到采出 k 个节点。若节点邻居数大于 k,则采用无放回的抽样。

$$h_{N(v)}^{k} \leftarrow AGGREGATE_k(\{h_u^{k-1}, \forall u \in N(v)\}) \tag{5}$$

3.4.2 引入层采样的方法

层采样[21]是一种针对节点邻居数量较大的图进行卷积操作的有效采样策略。在层采样中,对于每层 GCN,仅从上一层的一部分节点邻居中采样来生成当前层节点的邻居特征。这样可以在保证采样精度的前提下,显著减少计算负载,加速模型运行。

不同于 GraphSAGE 的逐点采样,FastGCN[21]使用分层采样,即将每个节点周围的邻居节点划分为若干个块,并对每个块进行聚合,再将这些块的结果进行拼接,得到最终的节点特征。它把计算时间复杂度降低为 $O(|V|)$,极大地提高了计算效率。因此 FastGCN 不会有邻域指数扩散的问题,其效率比 GraphSAGE 高出几个数量级。但是对于一个规模大且稀疏的图来说,由于 FastGCN 采用层采样的方式来加速模型训练,因此其对图中节点的采样有一定的误差,导致模型在保证分类准确率的情况下,可能会丢失一些局部结构信息。

ASGCN[19]通过设计一种自适应的逐层采样方法,加速了图卷积网络的训练。通过自上而下地构建神经网络的每一层,根据顶层的节点采样出下层的节点,可使得采样出的邻居节点被不同的父节点所共享,并且便于限制每层的节点个数来避免过度扩张。ASGCN 使用的逐层采样方法是自适应的,能显式地减少采样方差。实验证明,该模型在准确性和有效性上明显优于 GrapheSAGE 和 FastGCN。

3.4.3 引入子图采样的方法

影响图神经网络训练的一个主要问题是“邻居爆炸”问题。GraphSAGE 需要限制邻居采样数量到很小的水平;FastGCN[21]和 ASGCN[19]进一步把邻居膨胀系数限制到了 1,但同时也遇到了规模化、精度和计算复杂度方面的挑战。为解决普通训练方法无法训练超大图的问题,Cluster-GCN[20]通过先对图进行聚类,然后在小图上进行图神经网络训练的方式,加速了训练过程,避免了“邻居爆炸”问题。

GraphSAINT[39]是一种基于子图采样的归纳学习方法,其通过对小批量上的激活输出和损失的估计偏差和方差的分析,提出了正则化和抽样方法来提高训练的效果。因为每一批数据的训练都是在子图上完成的,因此“邻居爆炸”的问题得到了有效解决。与当前 SOTA 模型对比,GraphSAINT 在精度和速度方面都有提升。

3.4.4 结论

综上所述,对于大规模数据集节点分类问题,通过在 GCN 中引入邻居采样、分层采样、子图采样,在一定程度上提升了模型的速度和精度。

未来的研究可以采用更好的图划分和并行计算策略来提高训练效率。例如,可以将图分成若干个子图,并在计算过程中灵活地调整子图的大小和数量;也可以采用更好的数据集预处理方法,例如图压缩、降维和采样等方法,以减小数据集的规模和复杂度,提高模型的训练速度和泛化性能。

4 GCN 针对数据集质量问题的研究进展

4.1 针对数据的噪声问题

在采集图结构数据时,不可避免会出现一些差错,导致收集到的数据带有噪声,从而影响模型的分类性能,增加模型的复杂性[25]。针对该问题,基于 GCN 模型开展的研究工作可概括为两大方面,如表 8 所列。

表 8 GCN 针对数据集质量问题的改进

Table 8 GCN improvement for dataset quality problems

方法类型	代表工作	基本思想
融合图滤波器的方法	SGC[12], SBGC[40], BGCN[40]	过滤掉高频噪声来平滑图上节点的特征
引入对抗学习的方法	AT-GCN[41], RGCN[42]	通过对抗学习使模型更具鲁棒性,能处理对抗性攻击和误差数据

4.1.1 融合图滤波器

图滤波器可对图中的频率分量进行增强或衰减,实现不同的滤波效果。根据滤波效果,图滤波器可分为低通、高通和带通。

针对节点分类任务,Hoang 和 Maehara[43]在常用数据集上验证输入特征由低频真实特征和噪声组成的假设,研究发现 GCNs 中的图卷积层是简单的低通滤波。受深度学习方法的复杂度变化趋势可能会继承不必要的复杂度和冗余度计算的启发,Wu 等提出了一种非常高效的模型——简单图卷积(Simple Graph Convolution,SGC)[12]。它通过反复去除 GCN 层间的非线性并将得到的函数折叠成单个线性变换,来降低 GCN 的额外复杂性。SGC 在各种基准数据集上表现出与 GCN 相当甚至更好的性能。SGC 的计算式[12]如式(6)和式(7)所示:

$$\hat{Y} = \mathrm{softmax}(S \cdots SSX^{(0)}\Theta^{(1)}\Theta^{(2)} \cdots \Theta^{(K)}) \tag{6}$$

$$\hat{Y}_{SGC} = \mathrm{softmax}(S^K X\Theta) \tag{7}$$

本质上,SGC 相当于一个固定的低通图滤波器,后面加上一个线性分类器。SGC 的性能主要取决于低通滤波器,通过过滤掉高频噪声来平滑图上节点的特征。然而,SGC 的

滤波器是一个有限脉冲响应(Finite Impulse Response，FIR)图滤波器，其谱响应为多项式，存在对图形信号中的噪声敏感、半监督节点分类的标签效率低等局限性。

无限脉冲响应(Infinite Impulse Response，IIR)图滤波器的频率响应是有理函数，相比于FIR图滤波器具有更高的精度、更高的计算效率和更大的灵活性。基于IIR图滤波器的优势，Wang等提出了双滤波图卷积网络(Bi-filtering Graph Convolutional Networks，BGCN)[40]，通过级联两个子滤波模块来实现IIR滤波器。实验结果表明，BGCN能捕获到丰富且有价值的低频特征，很好地完成节点分类任务，并取得了与GCN及其变体相当的性能。然而，BGCN的改进是以时间复杂度的增加为代价的。受SGC的启发，作者又从图信号处理的角度构造了一个简化BGCN的模型——简单双滤波图卷积模型(Simple Bi-filtering Graph Convolution framework，SBGC)。此外，针对BGCN和SBGC的实现，作者还设计了一种新的低通图滤波器，用于捕获有利于节点分类任务数据表示的低频特征。大量实验表明，SBGC不仅在性能上优于其他基准方法，而且在计算效率上也保持着较高水平。BGCN和SBGC都对特征噪声具有鲁棒性，并表现出较高的标记效率。

2023年，Huang等提出了一种中通滤波的图卷积神经网络Mid-GCN[44]。该方法对图结构数据进行中通滤波，即在保留局部详细信息的同时去掉噪声和高频信息，增强GCN的泛化能力和鲁棒性。Mid-GCN在训练和推理阶段都具有出色的性能，并在多个图数据集上取得了最先进的结果。

4.1.2 引入对抗学习

对抗学习可使模型更具鲁棒性，能够处理对抗性攻击和误差数据。2018年，Ding等基于图卷积神经网络和对抗性训练的无监督图嵌入方法，提出了AT-GCN[41]，用于学习节点的低维向量表示。该模型使用生成对抗网络来捕获数据分布的不确定性，提升了模型的鲁棒性。

2019年，Zhu等提出一种新颖的“加固”GCN对抗攻击的模型RGCN[42]。该方法不是将节点表示为向量，而是采用高斯分布作为每个卷积层中节点的隐藏表示。这样，当图受到攻击时，该模型可以自动吸收高斯分布方差中对抗性变化的影响。

4.1.3 结论

综上所述，对于有噪声以及缺失的数据集的节点分类问题，学者从融合图滤波器、引入对抗学习等方面对GCN进行了研究并提出了一些模型。这些方法可以更好地提升GCN的鲁棒性，从而应对数据集的噪声问题。

未来的研究可以通过设计更加鲁棒的模型结构来提高模型性能，如加入节点嵌入特征或结构分区等方式，同时避免过度拟合和欠拟合现象；也可以结合多种更优的噪声处理技术，例如更优的噪声过滤与数据清洗等技术，以尽可能减小噪声的影响。

4.2 在不平衡数据集上的应用问题

不平衡数据是指数据集中一个或一些类的样本数量远远大于其他类的样本数量[26]。在不平衡数据集的节点分类中，数据分布不平衡导致模型的拟合能力不足，因为多数类主导损失函数，导致少数类的表示无信息，降低了整体分类性能。现有的多数GCNs研究只考虑理想的均衡数据集，很少考虑不均衡数据集[26]。然而，图中节点的类不平衡问题广泛存在于现实应用中，如欺诈检测[45]、疾病诊断[46]和金融风险分析[47]等，这推动了图领域中不平衡数据集的研究。已有研究可总结为3方面，如表9所列。

表9 在不平衡数据集上的应用

Table 9 Applications on unbalanced data set

方法类型	代表工作	基本思想
数据级方法	GraphSMOTE[48]，GraphMixup[49]，Imgagn[50]，GraphENS[51]	使用过采样或下采样技术使数据类别分布更加平衡
算法级方法	DR-GCN[52]，GNN-INCM[53]，Boosting-GNN[26]	修改模型的底层学习或决策过程以处理类不平衡问题
混合方法	DPGNN[54]，GNNCL[55]	将数据级和算法级方法结合起来

4.2.1 数据级方法

数据级方法的基本思想是使用过采样或下采样技术使数据集类别分布更加均衡[56-57]。由于图数据节点间存在关系的特征，以前的采样算法不易直接应用于图数据。

针对图数据不平衡性的问题，GraphSMOTE[48]模型将SMOTE[56]过采样算法扩展到图数据中，把图自动编码和节点分类任务结合在一起，在该模型的输出空间执行过采样，生成更多的自然节点和关系信息来解决图数据中的不平衡问题。GraphSMOTE[48]在捕获图中生成的节点与已有节点之间的联系时，通过基于MSE的邻接矩阵重建任务来训练边生成器，然后将其用于预测生成的节点和现有节点之间是否存在边。

GraphSMOTE[48]基于MSE的矩阵重建忽略了局部和全局的结构信息，导致边生成器过分强调具有相似特征的节点间的连接而忽略了节点间的远程依赖关系。为此，Wu提出了GraphMixup[49]，通过构建分离的语义空间便于在语义级别执行语义特征混合，并针对图提出上下文边混合，设计了两个基于上下文的自监督任务来考虑图结构中的局部和全局结构信息，同时开发了一种强化混合机制来自适应地确定每个少数类的上采样比例。但是，GraphMixup算法存在一些潜在缺点。首先是关联性假设可能不准确，GraphMixup算法假设相邻节点或训练集中的不同样本之间具有一定的关联性，即它们在语义空间上相似。然而，这个假设可能不适用于所有的图结构和数据集。如果数据集中存在复杂的噪声或特殊情况，算法可能无法准确捕捉到节点之间的相关性，导致生成的样本没有意义或不准确。其次是类别边界模糊，通过混合不同类别的节点特征，GraphMixup可能会导致类别之间的界限变得模糊。这意味着生成的样本可能更难被正确分类或在原始数据中存在不明显的边界。这可能对某些应用场景造成挑战，特别是在需要高精度分类的任务中。最后是计算复杂度较高，GraphMixup算法在生成新的样本时需要对图结构进行操作和计算，可能会引入一定的计算复杂度。这可能会增加模型训练的时间与资源消耗，尤其是对于大规模的图数据集。因此，算法的实际可行性和效率需要根据具体的应用

场景和计算资源来评估。

Imgagn 模型[50]则通过使用生成的权重矩阵在整个次要节点之间插值特征来合成次要节点。若矩阵中的权重大于固定阈值,则合成节点连接到原始次要节点。但 Imgagn 仅利用相同次要类的节点来生成次要节点,且 Imgagn 主要针对二元分类,因此合成节点的样本多样性受到很大限制。为此,Park 提出了一种新的数据增强模型 GraphENS[51],其利用整个节点来合成次要节点。实验表明,尤其是图中次要类的数量较少时,该方法在节点分类任务中表现更优。

4.2.2 算法级方法

算法级方法[58]的基本思想是通过修改模型的底层学习或决策过程来改进训练过程,从而解决不平衡类数据的问题。

针对图数据中类不平衡的问题,DR-GCN[52]是第一个使用图卷积神经网络研究节点级类不平衡嵌入问题的工作。DR-GCN 首先使用一个双层图卷积网络导出在类不平衡标签上训练的节点表示,然后结合条件对抗训练过程来帮助分离不同类别的标记节点表示。此外,为了减小多数类的卷积训练对其结构附近的少数类的负面传播影响,DR-GCN 训练所有未标记的节点将相似的数据分布拟合到学习嵌入空间中来训练有素的标记节点,从而解决了多数类和少数类之间的数据不平衡问题。DR-GCN 在一定程度上缓解了图数据的类不平衡问题,但也存在缺点。一方面,它使用 GCN 直接生成少数类中的节点嵌入,通常是不准确的。此外,这些节点嵌入在执行类条件对抗训练后仍然存在不准确问题。另一方面,由于标记节点的嵌入不准确,未标记的节点被迫通过分布对齐过程导致未标记节点的嵌入不准确。因此,DR-GCN 模型的分类性能仍然有提升空间。

受 DR-GCN 模型启发,Huang 等提出了基于 GNN 的不平衡节点分类模型 GNN-INCM[53],包括基于嵌入聚类的优化和基于图重构的优化两大协作模块,以准确地为多数类和少数类学习鲁棒的节点嵌入。ECO 采用两层图卷积网络,通过谱图卷积的局部一阶近似获取并编码图中的节点嵌入,然后进行聚类分析,增强节点嵌入的代表性,提高分类的准确性。GRO 采用内积解码器重构图结构,引用图重建损失来优化 GRO 模块,最大限度地减少信息损失。特别地,设计了一种与 ECO 和 GRO 相结合的新型硬样本策略,以确保正确地表示硬节点的嵌入。此外,作者提出了一种基于硬样本的知识蒸馏方法 HSKDM,通过分布和三元组对齐损失来联合训练多个 GNNINCM 模型作为最终节点分类的集成,以提高整体分类性能。基于 3 个真实世界数据集的大量实验表明,GNN-INCM 优于现有的最先进方法,HSKDM 可以显著提高整体分类性能。

受集成学习的启发,Shi 等提出集成模型 Boosting-GNN[26],首次使用集成学习来研究 GNN 中的不平衡数据集问题。该模型将自适应提升 AdaBoost 算法与 GNN 相结合,通过序列化的方式训练 GNN 分类器,并根据计算结果对样本进行重新加权。Boosting-GNN 通过对没有正确分类的训练样本设置更高的权重,实现了更高的分类精度和更好的可靠性。

4.2.3 混合方法

混合方法是将数据级和算法级方法结合起来,从数据和算法两个层面构建模型解决类不平衡问题。

Wang 等[54]于 2021 年提出的 DPGNN 模型在算法层面通过将标记节点与每个类原型进行比较来平衡训练损失。首先,它应用类原型驱动训练来平衡不同类的训练损失;然后,利用距离度量学习来区分从每个查询节点到所有类原型的距离的每个维度;最后,通过比较节点的学习距离度量表示与类原型的相似性来执行分类。在数据层面,DPGNN 通过引入原型节点来缓解节点不平衡问题,将每个类别的节点聚类成一个原型节点,并将原型节点与其他节点之间的距离作为特征输入到神经网络中,并采用自监督学习来平滑相邻节点间的学习距离度量表示,同时分离类间原型。在 8 个真实世界数据集上的实验证明了 DPGNN 在缓解类不平衡问题方面的有效性。

2022 年,Li 等通过引入课程学习思想提出了图神经网络框架 GNNCL[55]。该框架专门设计了两个组件,第一个是自适应图过采样,其关键思想是找到与原始图结构相关的最重要的样本进行插值,动态实现图中的数据分布由不平衡变为平衡。它根据图的特性对其进行调整,对于节点的生成,通过使用特征空间中同类的 k-最近邻节点来改进原始 SMOTE 方法,从而引导模型插入新的少数类节点。为了适应原始图的生成节点生成新的边,设计了边生成器,通过平滑度和同质性衡量生成新边后图结构的质量。第二个是基于邻居的度量学习,它根据伪标签对节点与邻居的距离进行正则化,从而动态调整少数类节点嵌入在特征空间中的位置,并使用一个基于邻居的三元组损失函数来发现少数类样本的稀疏边界,达到提高图不平衡数据分类表示嵌入的质量目标。作者还在不平衡的节点分类数据集上进行了实验,实验结果证明了 GNNCL 模型相对于 GraphSmote 和 Reweighting 等经典图神经网络模型的优越性能。

4.2.4 结论

综上所述,针对不平衡数据集节点分类问题,研究者们从数据层面、算法层面及二者混合层面分别开展了工作,一定程度上提高了不平衡数据集节点分类的准确性。

数据级方法解决图上的类不平衡问题考虑的 3 个关键点是:(1)如何为少数类生成新节点及其特征?(2)如何捕获图中生成的节点与已有节点之间的联系?(3)如何确定每个少数类的上采样比例?未来数据级方法可以围绕这 3 个问题来探索其他更优的图采样方法。算法级方法可设计更有效的 GCN 架构,引入知识蒸馏方法、迁移方法、无监督方法等以在加快模型训练速度的同时提高分类性能。混合级方法需要同时从数据层面和算法层面来考虑,结合现有的方法或提出新的数据级或算法级方法,并朝着构建一个可解释的端到端的学习框架来研究。

5 展望

本章总结了基于 GCN 的节点分类未来可能的研究方向。

(1)针对深层架构的梯度消失与过度平滑问题,可以提出新的适合的卷积核;可以采用优秀的子图划分方法对图数据

进行预处理，更好地提取特征；也可以引入外部信息，减少深度训练造成的失真，如节点关系可以作为监督信息指导训练过程。

(2)在动态图的节点分类方面，往往不仅涉及节点分类、边分类等单一任务，还涉及多任务联合学习，如节点分类和边预测同时进行。未来的研究可以探索如何设计基于 GCN 的多任务学习框架，在优化多个任务时更好地结合不同任务之间的关联信息。

(3)在异构图的节点分类方面，跨域数据集成和分类是一个重要的研究方向，需要进一步探索跨域链接对节点分类效果的影响。GCN 模型可以通过更优的跨域链接实现跨域数据集成和分类。

(4)针对大规模图的节点分类问题，由于图数据的稀疏性和特殊的数据结构，训练 GCN 模型的 GPU 利用率很低，且在多节点下的图神经网络计算效率不高。可行的解决方案是设计并行算法，以在多 GPU 上运行，从而提高效率；也可以引入更有效的抽样技术，以减小大型图的规模。

(5)为提高 GCN 模型对数据集噪声的鲁棒性，未来可以探索设计更加鲁棒的模型结构，如加入节点嵌入特征或结构分区等方式，同时避免过度拟合和欠拟合现象。还可以结合多种噪声处理技术，如更优的噪声过滤和数据清洗等技术，以尽可能减少噪声的影响。

(6)在针对不平衡数据集的节点分类方面，可以采用数据级方法和算法级方法以及混合级方法 3 种方式。数据级方法可以设计更好的方法为少数类生成新节点及其特征；捕获图中生成的节点与已有节点之间的联系；确定每个少数类的上采样比例。算法级方法考虑设计能加快模型训练速度并提高分类性能的更优或全新的 GCN 架构。混合级方法考虑在多个层面对问题进行综合分析，设计更优的混合分类方法。

结束语 节点分类问题不仅是很多研究领域的基础问题，而且有着广泛的应用，具有重要的研究价值。本文对以 GCN 为核心模型的节点分类研究工作进行了综述。

首先，本文给出节点分类问题的定义和基于图卷积神经网络的节点分类领域的挑战；其次，从模型和数据集两个角度讨论基于 GCN 的节点分类方法的挑战。重点介绍了针对模型问题的 GCN 节点分类研究的进展，包括梯度消失与过平滑问题、在动态图上的应用问题、在异构图上的应用问题以及在大规模图上的应用问题。针对梯度消失与过平滑问题，一方面研究设计更优深层网络，另一方面研究设计更优提取特征方式。针对动态图应用问题，研究通过引入时间维度与注意力机制等方法来处理动态图中节点特征的时间变化，以提高模型性能。对于在异构图上的应用研究，在多层次信息融合与跨领域知识迁移的方法等方面对 GCN 模型进行了改进，使其更好地适应异构图数据的特点。对于大规模数据集节点分类问题，目前的解决方案主要分为 3 种：邻居采样、层采样、子图采样。这些方法能够较好地对大型图进行处理。

之后重点介绍了针对数据集问题的 GCN 节点分类研究进展，包括数据的噪声问题以及数据集不平衡性问题。针对数据集的噪声问题的研究通过融合图滤波器和引入对抗学习等对 GCN 进行改进，提升 GCN 的鲁棒性。针对数据集的不平衡问题，从数据层面、算法层面及二者混合层面分别开展了研究工作，一定程度上提高了不平衡数据集节点分类的准确性。

最后，展望了节点分类领域的未来研究方向并对全文进行总结。总的来说，本文对近年来基于 GCN 的节点分类领域的研究进行了综述，总结了已有方法以及未来可研究的方向，希望能为进一步的研究提供一定的参考价值。

参 考 文 献

[1] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609.02907, 2016.

[2] SEN P, NAMATA G, BILGIC M, et al. Collective Classification in Network Data[J]. AI Magazine, 2008, 29(3): 93.

[3] LU Z, KIM W, WILBUR W J. Evaluating relevance ranking strategies for MEDLINE retrieval. [J]. Journal of the American Medical Informatics Association, 2009, 16(1): 32-36.

[4] Social Network: Reddit Hyperlink Network [EB/OL]. http://snap.stanford.edu/data/soc-RedditHyperlinks.html.

[5] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.

[6] HU W, FEY M, ZITNIK M, et al. Open graph benchmark: Datasets for machine learning on graphs[J]. Advances in Neural Information Processing Systems, 2020, 33: 22118-22133.

[7] XIE Y Y, FENG X, YU W J, et al. Network Embedding Method based on Randomized Matrix Factorization [J]. Journal of Computer Research and Development, 2021, 44(3): 447-461.

[8] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]// Proc. of the Conf. on Neural Information Processing Systems (NIPS). Cambridge: MIT Press, 2017: 1024-1034.

[9] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv. 1312.6203, 2013.

[10] HENAFF M, BRUNA J, LECUN Y. Deep Convolutional Networks on Graph-Structured Data[J]. arXiv. 1506.05163, 2015.

[11] DUVENAUD D, MACLAURIN D, AGUILERA-IPARRAGUIRRE J, et al. Convolutional networks on graphs for learning molecular fingerprints[J]. arXiv: 1509.09292, 2015.

[12] WU F, SOUZA A, ZHANG T, et al. Simplifying graph convolutional networks [C] // International Conference on Machine Learning. PMLR, 2019: 6861-6871.

[13] XU K, LI C, TIAN Y, et al. Representation learning on graphs with jumping knowledge networks [C] // International Conference on Machine Learning. PMLR, 2018: 5453-5462.

[14] ABU-EL-HAIJA S, PEROZZI B, KAPOOR A, et al. MixHop: higher-Order Graph Convolutional Architectures via Sparsified Neighborhood Mixing[C]// International Conferencce on Machine Learning. PMLR, 2019: 21-29.

[15] MICHELI A. Neural network for graphs: a contextual constructive approach [J]. IEEE Transactions on Neural Networks, 2009, 20(3): 498-511.

[16] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]// International Conference on Machine Learning. PMLR, 2017: 1263-1272.

[17] HECHTLINGER Y, CHAKRAVARTI P, QIN J. A generalization of convolutional neural networks to graph-structured data [J]. arXiv: 1704.08165, 2017.

[18] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv: 1710.10903, 2017.

[19] HUANG W, ZHANG T, RONG Y, et al. Adaptive sampling towards fast graph representation learning [J]. arXiv. 1809.05343, 2018.

[20] CHIANG W L, LIU X, SI S, et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks[J]. arXiv. 1905.07953, 2019.

[21] CHEN J, MA T, XIAO C. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling [J]. arXiv. 1801.10247, 2018.

[22] CHEN H, HUANG Z, XU Y, et al. Neighbor enhanced graph convolutional networks for node classification and recommendation[J]. arXiv. 2203.16097, 2022.

[23] YING R, YOU J, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[C]// Proceedings of the Annual Conference on Neural Information Processing Systems. Cambridge, MA: MITPress, 2018: 4805-4815.

[24] TITOV I, WELLING M, SCHLICHTKRULL M, et al. Modeling relational data with graph convolutional networks [C]// Proceedings of the 15th European Semantic Web Conference. 2018: 593-607.

[25] ZHANG H Q. Research on Several Node Classification Methods based on Graph Learning [D]. Guilin: Guangxi Normal University, 2022.

[26] SHI S, QIAO K, YANG S, et al. Boosting-GNN: boosting algorithm for graph networks on imbalanced node classification[J]. Frontiers in Neurorobotics, 2021, 15: 775688.

[27] LI G, MÜLLER M, THABET A, et al. DeepGCNs: Can GCNs Go As Deep As CNNs? [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020.

[28] SUN K, ZHU Z, LIN Z. Adagcn: Adaboosting graph convolutional networks into deep models[J]. arXiv: 1908.05081, 2019.

[29] ABU-EL-HAIJA S, KAPOOR A, PEROZZI B, et al. N-gcn: Multi-scale graph convolution for semi-supervised node classification[C]// Uncertainty in Artificial Intelligence. PMLR, 2020: 841-851.

[30] HUANG G, LIU Z, LAURENS V D M, et al. Densely Connected Convolutional Networks[C]// CVPR 2017. IEEE Computer Society, 2016.

[31] SZEGEDY C, LIU W, JIA Y, et al. Going Deeper with Convolutions[C]// CVPR 2014. IEEE Computer Society, 2014.

[32] YU B, YIN H, ZHU Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting [C]// IJCAI 2017. 2017.

[33] PAREJA A, DOMENICONI G, CHEN J, et al. Evolvegcn: Evolving graph convolutional networks for dynamic graphs [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 5363-5370.

[34] SANKAR A, WU Y, GOU L, et al. DySAT: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks[C]// The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM'20). ACM, 2020.

[35] XU D, RUAN C W, KORPEOGLU E, et al. Inductive representation learning on temporal graphs [J]. arXiv: 2002.07962, 2020.

[36] MA Y, GUO Z, REN Z, et al. Streaming Graph Neural Networks[C]// The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20). ACM, 2020.

[37] LIU Z, SHEN Y, CHENG X, et al. Learning representations of inactive users: A cross domain approach with graph neural networks[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 3278-3282.

[38] ZHANG L, WU N. HGCC: Enhancing Hyperbolic Graph Convolution Networks on Heterogeneous Collaborative Graph for Recommendation[J]. arXiv: 2304.02961, 2023.

[39] ZENG H, ZHOU H, SRIVASTAVA A, et al. Graphsaint: Graph sampling based inductive learning method [J]. arXiv: 1907.04931, 2019.

[40] WANG S, PAN Y, ZHANG J, et al. Robust and label efficient bi-filtering graph convolutional networks for node classification [J]. Knowledge-Based Systems, 2021, 224: 106891.

[41] HUANG J, DU L, CHEN X, et al. Robust Mid-Pass Filtering Graph Convolutional Networks[J]. arXiv: 2302.08048, 2023.

[42] DING M, TANG J, ZHANG J. Semi-supervised learning on graphs with generative adversarial nets[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 913-922.

[43] ZHU D, ZHANG Z, CUI P, et al. Robust Graph Convolutional Networks Against Adversarial Attacks [C]// The 25th ACM SIGKDD International Conference. ACM, 2019.

[44] HOANG N T, MAEHARA T. Revisiting graph neural networks: All we have is low-pass filters[J]. arXiv: 1905.09550, 2019.

[45] MONGWE W T, MALAN K M. A survey of automated financial statement fraud detection with relevance to the South African context[J]. South African Computer Journal, 2020(1): 74-112.

[46] LARRAZABAL A J, NIETO N, PETERSON V, et al. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis [J]. Proceedings of the National Academy of Sciences, 2020, 117(23): 201919012.

[47] YU S A, YW A, XIN Y A, et al. Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending [J]. Information Sciences, 2020, 525: 182-204.

[48] ZHAO T, ZHANG X, WANG S. GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks [C]// WSDM'21. ACM, 2021.

[49] WU L, LIN H, GAO Z, et al. Graphmixup: Improving class-im-

balanced node classification on graphs by self-supervised context prediction[J]. arXiv:2106.11133,2021.

[50] QU L,ZHU H,ZHENG R,et al. ImGAGN:Imbalanced Network Embedding via Generative Adversarial Graph Networks [J]. arXiv:2106.02817,2021.

[51] JOONHYUNG P,JAEYUN S. GraphENS:Neighbor-aware ego network synthesis for class-imbalanced node classification[C]// The 9th International Conference on Learning Representations. Virtual:OpenReview.net,2021.

[52] SHI M,TANG Y,ZHU X,et al. Multi-Class Imbalanced Graph Convolutional Network Learning[C]// International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization,2020.

[53] HUANG Z,TANG Y,CHEN Y. A graph neural network-based node classification model on class-imbalanced graph data[J]. Knowledge-Based Systems,2022,244(23):108538. 1-108538. 12.

[54] WANG Y,AGGARWAL C,DERR T. Distance-wise prototypical graph neural network in node imbalance classification[J]. arXiv:2110.12035,2021.

[55] LI X,WEN L,DENG Y,et al. Graph neural network with curriculum learning for imbalanced node classification[J]. arXiv:2202.02529,2022.

[56] CHAWLA N V,BOWYER K W,HALL L O,et al. SMOTE:Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.

[57] MORE A. Survey of resampling techniques for improving classification performance in unbalanced datasets[J]. arXiv. 1608.06048,2016.

[58] THAI-NGHE N,GANTNER Z,SCHMIDT-THIEME L,et al. Cost-sensitive learning methods for imbalanced data[C]// International Joint Conference on Neural Networks. IEEE,2010.

ZHANG Liying, born in 1980, Ph.D, lecturer, master supervisor, is a member of CCF(No. 78197M). Her main research interests include spatio-temporal data mining, machine learning and deep learning on graphs.

（责任编辑：柯颖）

CCF 代表团出席日本 IPSJ 大会，CCF 副理事长胡事民应邀作特邀报告

2024 年 3 月 15－17 日，应日本信息处理学会（IPSJ）邀请，由 CCF 会士、CCF 副理事长、中国科学院院士、清华大学教授胡事民，CCF 秘书长唐卫清，CNCC 项目主任麻宇鹏组成的代表团出席了 IPSJ 第 86 届全国学术大会。CCF 会士、CCF 副理事长，中国科学院院士、清华大学教授胡事民应邀作特邀报告。

胡事民做特邀报告

胡事民的报告题目为“可视媒体计算的骨干网络和深度学习框架”。本次报告从算法和框架两个方面介绍清华大学在构建可视媒体计算基础方面所做的工作，特别介绍了用于可视媒体深度学习的基本骨干网络、计图深度学习框架和用于基础模型的深度学习技术，包括异构计算平台、基础模型的高效训练和推理等。

据 CCF 微信公众号