基于分班图神经网络的度不平衡节点分类

刘润雨[1] 贾路楠[1]
LIU Runyu JIA Lunan

摘 要 图神经网络（GNNs）在处理非欧几里得数据上表现出了超越传统卷积神经网络（CNNs）的优势，主要的原因在于 GNNs 可以聚合节点的邻居信息来增强节点的特征表示。但是，许多图数据集都面临着度不平衡的问题，现有的工作多关注类不平衡问题，在处理此类问题时表现不佳。针对以上问题，提出了一种基于分班图神经网络的度不平衡节点分类模型（DIP-GNN）。考虑到单一模型不能同时学习不同位置节点的特征，根据节点的度信息将所有节点划分为三类，使用三个不同的模型分别训练。同时，考虑到稀疏位置节点的信息不充分，使用邻居增强的手段来增加这类节点的邻居信息，来解决现有工作中节点度不平衡的问题。在五个公共图基准数据集上的实验结果表明，通过分班模型能有效捕获图中不同位置节点的特性，节点分类准确率相较于目前常用的图卷积神经网络有更好的表现。

关键词 度不平衡；节点分类；多教师多学生；邻居增强

doi：10.3969/j.issn.1672-9528.2024.04.025

0 引言

图数据是一种重要的数据形式，广泛存在于诸多领域中，例如社交网络、知识图谱以及推荐系统等[1-2]。由于缺乏图中节点之间拓扑结构关系的有效表示，传统机器学习方法在这类数据任务上的性能往往不佳。近年来，基于图神经网络[3]（graph neural networks，GNNs）的相关方法[4-6]在图数据上取得了很大进展，能有效表示、学习节点之间的关系，相关任务上的学习性能较之以往提升显著，并开拓出一系列新的应用领域[7-8]。图中节点分类[9]是图神经网络重要的下游任务之一，其目的是基于已标注节点预测未标注节点的相关类别。例如，在引文网络中，节点表示文献，文献引用构建了节点间的连接关系。通过学习文献的文本特征（节点特征）和引用关系（节点间的拓扑关系），可以生成引文网络意义下各文献的有效特征表示，进而预测文献的研究主题（节点类别）。目前，基于 GNN 的节点分类方法普遍依赖于消息传递机制（message passing neural networks，MPNN）[10]，即在相关节点的邻域上聚合其邻居节点特征，据此丰富、平滑自身节点特征，有效支撑下游学习任务。此机制面临的问题是，图中节点位置并不平衡，有的节点处于头度位置，周围邻居节点很多，有的节点位于尾度位置，周围邻居甚为稀疏。图 1 给出了一个示例。

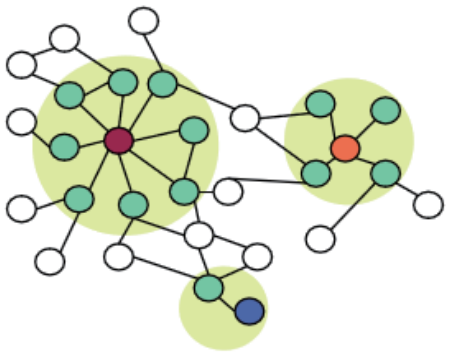

图 1 图节点度的不平衡性

实际应用中，图中中度节点通常占比较高（如表 1 所示）。如此，无差别地聚合邻域节点信息必然带来学习上的失衡，进而导致模型性能特别是尾度节点上的性能不足。针对上述节点度不平衡问题，本文提出一种“分班”图神经网络节点分类方法，分治训练头度、中度与尾度节点，取得三类节点上的性能改进。

表 1 五个节点分类数据集的基本信息

数据集	节点数	特征长度	边数	类别数	头度 / 中度 / 尾度
Cora	2485	1433	5069	7	411/1720/354
Citeseer	2110	3703	3668	6	333/1247/530
Pubmed	19 717	500	44 324	3	2896/7727/9094
A-Photo	7487	745	119 043	8	1090/5214/1183
A-Computer	13 381	767	254 778	10	3676/7236/2469

1. 福建理工大学计算机科学与数学学院 福建福州 350118

1 研究背景

1.1 图神经网络

图神经网络模型一般基于消息传递机制通过有效聚合邻域信息来学习相关节点的节点级表示，进而完成下游节点级任务，或通过池化等相关机制进一步学习图级表示、完成图级相关任务等。例如，经典图卷积神经网络（graph convolutional network， GCN）通过逐层聚合、分层传播来学习节点表示。GraphSage[11] 通过采样固定数量的邻居节点聚合生成相关节点的有效表示。GAT[12]（graph attention network）则基于注意机制实现特征聚合，将注意力机制引入到基于空间域的图神经网络。与基于谱域的图卷积神经网络不同，图注意力网络不需要使用拉普拉斯等矩阵进行复杂的计算，仅是通过一介邻居节点的表征来更新节点特征，所以算法原理从理解上较为简单。GCNII[13] 引入残差连接和恒等映射以有效聚合多阶邻居信息，一定程度上消解过平滑问题。APPNP[14] 模型将 GCN 中的邻居聚合和特征转换两个操作解耦开来，并基于个性化 PageRank 算法构建更有效聚合更高阶的邻接信息。上述模型的提出极大地促进了图神经网络的发展。

1.2 类别不平衡图神经网络

在现实场景中，类别不平衡是许多数据集的自然属性之一，包括图形数据。目前已经提出了不少先进的方法来解决特定任务图数据中的类别不平衡问题，SPARC[15] 试图通过课程自节奏策略来准确描述罕见类别，当少数类与多数类不可分离时，学习一个面向显著少数类的嵌入表示以便更好地刻画它们，并且准确地描述稀缺信息在标签信息稀缺性方面的表现。GraphENS[16] 通过基于相似性组合两个不同的自我网络来合成少数类的整个自我网络。此外，还引入了一种基于显著性的节点混合方法，以利用其他节点丰富的类通用属性，同时阻止类特定特征的注入。GraphSMOTE[17] 从多数类节点中构造了一个嵌入空间来编码节点之间的相似性，再通过这个编码空间来生成少数类的节点，同时训练一个边缘生成器来建模关系信息。

目前，图神经网络不平衡现象主要集中在研究类别不平衡的领域，很少关注到度不平衡这一图数据自带的属性。与传统用于图像识别的卷积神经网络不同，图数据在训练以及测试阶段都可以提前知道图中某个节点的具体邻居数量，合理利用这一信息可以为模型带来一定的性能提升。

2 相关概念及定义

定义 1．无权无向图。给定图 $G=(V, E)$。V 表示节点集 $\{v_1, v_2, \cdots, v_N\}$，$N=|V|$ 表示节点数，E 表示边集 $\{e_1, e_2, \cdots, e_M\}$，$M$ 表示边数。

定义 2．邻接矩阵与度矩阵。邻接矩阵 $\boldsymbol{A}$ 被定义为 $\boldsymbol{A}=\{0,1\}^{N\times N}$，如果 $(v_i,v_j)\in E$，则 $A_{i,j}=1$。$\boldsymbol{D}$ 表示度矩阵，$D_{i,i}=\sum_{j=1}^{N}A_{i,j}$。$\widetilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$ 表示添加自环的邻接矩阵，$\widetilde{\boldsymbol{D}}=\boldsymbol{D}+\boldsymbol{I}$ 表示添加自环的度矩阵，$\hat{A}=\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ 表示对应于 $\widetilde{A}$ 的归一化邻接矩阵。

定义 3．特征矩阵与标签矩阵。特征矩阵 $\boldsymbol{X}\in R^{N\times F}$，节点 v_i 被表示成一个 F 维的特征向量 $\boldsymbol{X}_i$，节点标签矩阵 $\boldsymbol{Y}\in\{0,1\}^{N\times C}$，$\boldsymbol{Y}_i\in\{0,1\}^{C}$ 是一个 One-Hot 类型向量，C 为节点对应类别数。对任意节点 $v_i\in V$，均有 $\sum_{j=1}^{C}Y_{i,j}=1$。

定义 4．多层感知机（MLP）。MLP 不使用节点的邻居信息，只靠节点自身的特征通过多层非线性变换来更新节点的特征。该过程可以定义为：

$$X^{(k+1)}=\sigma(X^{(k)}W^{k}_{(MLP)}) \tag{1}$$

定义 5．图卷积神经网络（GCN）。GCN 通过迭代聚合各节点自身特征表示及其邻居节点特征表示获得节点的平滑特征表示。该过程可表示为：

$$X^{(k+1)}=\sigma(\hat{A}X^{(k)}W^{k}_{(GCN)}) \tag{2}$$

$X^{(k)}$ 和 $X^{(k+1)}$ 分别表示第 k 层与第 k+1 层聚合后的节点特征，当 k=0 时，$X^{(0)}$ 表示初始节点特征 $\boldsymbol{X}$；$\boldsymbol{W}^{(k)}$ 表示第 k 层网络的可训练权重矩阵，$\sigma(\cdot)$ 为激活函数。

定义 6．节点分类。给定无向图 $G=(V, E)$ 与初始节点特征矩阵 $\boldsymbol{X}\in R^{N\times F}$，节点分类旨在学习映射函数。

$$f_{(X\to\tilde{Y})}: X\in R^{N\times F}\mapsto\widetilde{Y}\in R^{N\times C} \tag{3}$$

将图 G 中每一节点映射成 C 维向量，C 为节点对应类别数。一般情况下，$C<<F$，所得矩阵 $\tilde{\boldsymbol{Y}}$ 用于预测节点类别。

传统的图卷积神经网络将图数据集中的所有节点一视同仁，并没有考虑到不同节点的度信息是不同的，训练出的模型只能满足不同位置节点的平均精度，不能满足不同位置节点的各自最好精度。

3 基于分班图神经网络的度不平衡节点分类模型

给定一个节点 v，它的度为 d，$d_i=|N_{(i)}|$。节点的度 $d\geq HD$，HD 是头度（head degree），即为头度节点；节点的度 $d\leq TD$（TD，尾度（tail degree），即为尾度节点；节点的度 $TD<d<HD$，即为中度节点（mid degree）。通过比较后图中所有的节点都被分为三类：头度节点 $v_{(头)}$、中度节点 $v_{(中)}$ 和尾度节点 $v_{(尾)}$。上文根据节点的度信息将节点分为了三类，这一小节针对每一类训练一个 GCN 模型。

给定无向图 $G=(V, E)$ 和初始的节点特征矩阵 $X\in R^{M\times F}$，节点 v_i 对应一个 F 维的向量 X_i，所有节点根据度分为三类：头度节点 $v_{(头)}$、中度节点 $v_{(中)}$ 和尾度节点 $v_{(尾)}$。本文的目标是分别学习三个映射函数：

$$f_{(头)}: X \in R^{M_{(头)} \times F} \rightarrow Z \in R^{M_{(头)} \times C} \quad (4)$$

$$f_{(中)}: X \in R^{M_{(中)} \times F} \rightarrow Z \in R^{M_{(中)} \times C} \quad (5)$$

$$f_{(尾)}: X \in R^{M_{(尾)} \times F} \rightarrow Z \in R^{M_{(尾)} \times C} \quad (6)$$

将图 G 中的每一个节点映射成 C 维的向量，C 为数据集对应的类别数，一般情况下满足 $C<<F$，得到的节点表示矩阵 $Z \in R^{M \times C}$ 用于预测节点的类别。

相较于目前主流的 GCN 模型及其变体，它们都赋予所有节点相同的卷积层数，直观来看，模型更偏向于在数据集中占大部分的中度位置节点，并且尾度节点的准确率相较于其他位置更低。相比之下，本文提出的模型灵活性更好，并且可以最大化地捕获不同位置节点的特征，并通过 K- 近邻的方法来增强稀疏位置节点的邻居信息，最终的节点分类任务的精度更高。虽然本文用了三个模型来进行训练，但实际使用的时候，本文只是针对一个节点使用一个训练好的模型来分类，并且模型可以并行以加快测试，模型的流程图如图 2 所示。

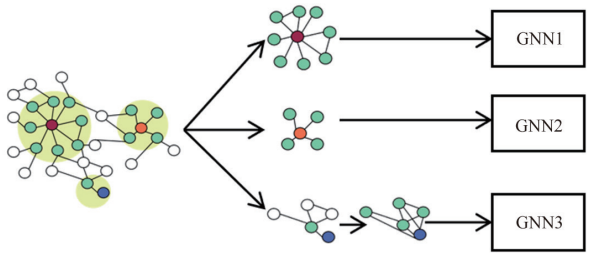

图 2 基于分班图神经网络的度不平衡节点分类模型

4 结果与分析

4.1 实验数据

本文使用文献提供的五个节点分类数据集，这五个数据集也是目前图神经网络领域的基准数据集的一部分，五个数据集的基本信息如表 1 所示。Cora、Citeseer 和 Pubmed 三个数据集为引文网络数据集，节点数代表数据集中所收录论文的数量；节点之间的边代表两篇论文之间的引用关系，一篇论文可以引用多篇论文，并且可以同时被多篇论文引用，节点的特征是由 One-Hot 编码的向量来表示。A-photo 和 A-computer 提取自 Amazon 共购图，其中节点表示产品，边表示两种产品是否经常共购，特征表示用 bag-of-words 编码的产品评论，标签是预定义的产品类别。

4.2 实验设置

实验基于 PyTorch 和 PyG，实现了 DIP-GNN 和其他方法的对比。为了进行公平比较，实验为所有中心节点添加自环以保留中心节点特征信息，并统一设置学习率（learning rate，LR）为 0.01，权重衰减率（weight decay rate，WD）为 0.000 1。由于模型参数的随机初始化和提前终止会对最终结果造成一定影响，本文采用每个方法在每个数据集上运行 100 次，取计算结果平均值。实验环境为 Inter(R) Xeon(R) E5-2680 v3 @2.50 GHz CPU 和 NVIDIA TITAN Xp 16 GB GPU，操作系统为 Windows10，内存为 32 GB，开发语言为 Python。

4.3 与传统的模型精度对比结果

本部分展示两部分实验结果，一部分是头度、中度和尾度放在一起的实验结果对比，另一部分是头度、中度和尾度三类分别实验的对比。从表 2 中可以看出本文的模型在五个数据集上均有明显提升，分别提升 0.96%、0.73%、0.66%、0.92% 和 0.53%；从表 3 中可以看出分开训练所带来的优势，尤其是对于尾度节点，提升是巨大的，在五个数据集中分别提升 2.07%、1.57%、0.15%、0.87% 和 1.22%。

表 2 五个节点分类数据集上的分类实验结果

		MLP	GCN	DIP-GNN
Cora	头度	72.67%	80.43%	**81.66%（↑ 1.49%）**
	中度	74.89%	83.06%	**83.30%（↑ 1.14%）**
	尾度	69.36%	76.64%	**78.03%（↑ 2.07%）**
Citeseer	头度	69.50%	77.04%	**77.34%（↑ 0.43%）**
	中度	63.78%	72.54%	**73.32%（↑ 0.76%）**
	尾度	60.66%	67.16%	**68.29%（↑ 1.57%）**
Pubmed	头度	71.43%	87.29%	**87.68%（↑ 0.31%）**
	中度	71.03%	86.67%	**87.35%（↑ 0.36%）**
	尾度	70.57%	84.63%	**85.55%（↑ 0.15%）**
A-Photo	头度	84.78%	93.47%	**94.85%（↑ 1.01%）**
	中度	84.98%	94.10%	**94.50%（↑ 0.20%）**
	尾度	83.78%	87.95%	**88.94%（↑ 0.87%）**
A-Computer	头度	75.91%	90.31%	**90.65%（↑ 0.65%）**
	中度	76.77%	90.00%	**90.06%（↑ 0.14%）**
	尾度	74.89%	84.96%	**86.15%（↑ 1.22%）**

表 3 五个节点分类数据集上的总实验结果

数据集	MLP	GCN	DIP-GNN
Cora	73.51%	80.04%	**81.00%**
Citeseer	64.77%	72.25%	**72.98%**
Pubmed	71.90%	86.20%	**86.86%**
A-Photo	86.42%	91.84%	**92.76%**
A-Computer	79.56%	88.42%	**88.95%**

4.4 与其他 GNN 模型的结合

本文所提模型可以与主流的所有 GNN 模型相结合并带来提升，表 4 给出了分班思想与 GAT 模型结合的实验结果，可以看出 GAT 在 GCN 的基础上有很大的提升，与本文模型结合以后性能进一步提升，证明了本文思想的普适性与先进性。

表 4 五个节点分类数据集上与 GAT 模型结合的实验结果

数据集	GCN	GAT	DIP-GNN+GAT
Cora	80.04%	81.50%	**82.34%**
Citeseer	72.25%	74.48%	**75.21%**
Pubmed	86.20%	87.86%	**88.44%**
A-Photo	91.84%	92.76%	**93.66%**
A-Computer	88.42%	89.95%	**90.30%**

5 结语

本文提出了一种基于分班图神经网络的度不平衡节点分类模型（DIP-GNN）来解决图神经网络中图数据集的节点度不平衡问题，以实现不同位置的节点都能获得更好的模型性能，同时该方法可以推广到不同的图神经网络变体模型以及相关领域（超图、异质图等）以提高模型的性能。后续将尝试在训练开销可控的情况下继续细化节点的划分，以使模型达到更优的效果。

参考文献：

[1] 吴越，孙海春．基于图神经网络的知识图谱补全研究综述 [J/OL]. 数据分析与知识发现，1-25[2024-03-28].http://kns.cnki.net/kcms/detail/10.1478.G2.20231109.1004.002.html.

[2] 马帅，刘建伟，左信．图神经网络综述 [J]. 计算机研究与发展，2022, 59(1): 47-80.

[3] THOMAS N K, MAX W. Semi-supervised classification with graph convolutional networks[EB/OL].(2016-09-09)[2024-01-06]. https://arxiv.org/abs/1609.02907.

[4] 张继杰，杨艳，刘勇．利用初始残差和解耦操作的自适应深层图卷积 [J]. 计算机应用，2022,42(1):9-15.

[5] 张嘉杰，过弋，王家辉．基于特征和图结构信息增强的多教师学习图神经网络 [J]. 计算机应用研究，2023,40(7):2013-2018.

[6] 郭梦昕．用于不平衡节点分类的集成图神经网络模型 [J]. 现代信息科技，2023,7(3):29-32.

[7] 户佐安，邓锦程，韩金丽，等．图神经网络在交通预测中的应用综述 [J]. 交通运输工程学报，2023,23(5):39-61.

[8] 杨帆，邹窈，朱明志，等．基于图注意力变换神经网络的信用卡欺诈检测模型 [J/OL]. 计算机应用，1-8[2024-03-28]. http://kns.cnki.net/kcms/detail/51.1307.TP.20231102. 1356.018.html.

[9] 张丽英，孙海航，石兵波．基于图卷积神经网络的节点分类方法研究综述 [J/OL]. 计算机科学，1-19[2024-03-28]. http://kns.cnki.net/kcms/detail/50.1075.TP.20230925.1655.162.html.

[10] HAMAGUCHI T, OIWA H, SHIMBO M, et al.Knowledge transfer for out of-knowledge-base entities:a graph neural network approach[C]//Proceedings of IJCAI. New York: Curran Associates, 2017:1802-1808.

[11] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:ACM, 2017:1025-1035.

[12] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL].(2017-10-30)[2024-01-25]. https://arxiv.org/abs/1710.10903.

[13] CHEN M, WEI Z, HUANG Z, et al. Simple and deep graph convolutional networks[J].Proceedings of the 37th International Conference on Machine Learning, 2020, 161: 1725-1735.

[14] JOHANNES K, ALEKSANDAR B, STEPHAN G. Predict then propagate: Graph neural networks meet personalized pagerank[EB/OL].(2018-10-14)[2024-02-02].https://arxiv.org/abs/1810.05997.

[15]ZHOU D, HE J, YANG H, et al. Sparc: Self-paced network representation for few-shot rare category characterization[J]. SIGKDD explorations, 2018(Udisk):2775-2784.

[16] PARK J, SONG J, YANG E. GraphENS: neighbor-aware ego network synthesis for class-imbalanced node classification[EB/OL].[2024-01-15].https://www.semanticscholar.org/paper/GraphENS%3A-Neighbor-Aware-Ego-Network-Synthesis-for-Park-Song/4d0f0212ac509445983dab3032af8cbd14a7c3e3.

[17] ZHAO T, ZHANG X, WANG S. Graphsmote: imbalanced node classification on graphs with graph neural networks[C]// WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining.New York:ACM, 2021: 833-841.

【作者简介】

刘润雨（1998—），男，河南南阳人，硕士研究生，研究方向：图神经网络。

贾路楠（1999—），女，河南南阳人，硕士研究生，研究方向：图论及其应用。

（收稿日期：2024-02-07）