引用格式：赵华健, 杨钦程, 胡兆龙. 基于集成学习的不平衡图节点分类算法 [J]. 电子科技大学学报, 2025, 54(3): 455-463.
ZHAO H J, YANG Q C, HU Z L. Unbalanced graph node classification algorithm based on ensemble learning[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(3): 455-463.

基于集成学习的不平衡图节点分类算法

赵华健，杨钦程，胡兆龙*

（浙江师范大学 计算机科学与技术学院，金华 321004）

摘要：图神经网络（GNN）被广泛应用于节点分类。然而，现有研究集中于平衡数据集，但是不平衡数据却普遍存在。传统处理不平衡数据集的方法，如重采样和重加权，往往需要进行较多的预处理或提出新的网络结构，容易引入新的偏差并导致信息丢失。该文提出了一种改良的装袋（Bagging）集成学习方法，对不平衡图数据集进行了k折划分，并采用GNN为基础模型对子数据集进行训练得到多个不同的子模型。最后，通过融合不同模型来提升节点的分类精度而不引入过多的预处理。基于不平衡图数据集的实验结果，表明所提出的方法在准确性和鲁棒性上优于基本分类器，此外，还发现分类精度随着k的增加先提高后降低。

关键词：图神经网络；节点分类；图网络结构；不平衡图数据集；集成学习
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2024084

Unbalanced graph node classification algorithm based on ensemble learning

ZHAO Huajian, YANG Qincheng, and HU Zhaolong*

(School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China)

Abstract: Graph neural network (GNNs) has been widely employed in node classification over the past few years. However, existing research has predominantly focused on balanced datasets, whereas imbalanced data is prevalent. Traditional approaches to handling imbalanced datasets, such as resampling and reweighting, often require substantial preprocessing or proposing new network structures, which can introduce new biases and lead to information loss. An enhanced bootstrap aggregating (Bagging) ensemble learning method is proposed to address imbalanced graph datasets. It involves partitioning the data into k folds and training multiple distinct sub-models using GNNs as the base model. Finally, by fusing different models, the node classification accuracy is improved without introducing excessive preprocessing. Experimental results on imbalanced graph datasets demonstrate that the proposed method outperforms the base classifier in terms of accuracy and robustness. Additionally, it is observed that classification accuracy initially increases and then decreases with the increase of k.

Key words: graph neural network; node classification; graph network structure; imbalanced graph data set; ensemble learning

图是一种在日常生活中十分常见的数据结构，即由节点和边构成的网络结构，比如社交网络中节点表示用户，边表示用户与用户之间的好友、关注等关系[1-2]。图数据出现在许多应用中，包括社交媒体和推荐系统，图数据挖掘一直是研究热点之一[3-4]。然而，由于图难以被嵌入至欧几里得空间中，一直缺乏有效编码图结构的方法，直到图神经网络（graph neural network, GNN）的出现才改变这一现状[3]。近年来，GNN在图相关邻域取得了显著进步，其中一种典型任务是半监督节点分类，它旨在根据图的拓扑结构和部分标记节点推断剩余未标记节点的类别。节点分类问题的传统方法大多基于图

收稿日期：2024−04−10
基金项目：国家自然科学基金（62103375）；浙江省自然科学基金（LY23F030003）
作者简介：赵华健，主要从事节点分类方面的研究。
*通信作者 E-mail: huzhaolong@zjnu.edu.cn

拉普拉斯正则化[5]，然而最近几年 GNN 已经成为半监督节点分类领域的热门方法。与传统方法相比，GNN 在解决节点分类问题上具有更好的性能，但它很容易受到数据集不平衡的影响而趋向于学习多数类的特征。同时，现有研究大都只考虑了数据集平衡的场景，忽视了数据集不平衡问题[6]。

不平衡数据通常指的是所有数据中不同类别样本数量差异显著的情况。具体来说，当某些类别的样本数量远远少于其他类别，而这种差异可能达到数倍甚至更多时，就可以认为数据不平衡。不平衡数据在现实世界中普遍存在，如在社交网络中，网红和明星的占比远低于普通人，却有成千上万的粉丝，他们的影响力自然也远超普通人；当推广到不平衡分布的图上时，由于多数类的节点占主导地位，从而使得训练过程偏向多数类。另外，由于 GNN 通常通过在图结构中传递和聚合节点间的信息来学习节点表示，这种方法可能导致信息更频繁地在多数类节点之间传播和聚合，从而加剧多数类节点对于表示学习的影响，削减了对少数类节点特征的学习，进一步增强了多数类压倒少数类的趋势[7]。

传统的解决不平衡方法主要分为数据级和算法级[8]。数据级方法试图通过预处理以重新平衡先前的类分布，其中包括过采样少数类和欠采样多数类[9]。但这些方法可能导致过拟合或丢失有价值的信息。算法级方法试图修改或组合现有的方法以强调少数类，如代价敏感学习和集成学习[10]。代价敏感学习中如重加权试图提高少数类的权重以提高少数类的识别精度，但在少数类样本数量过少时，代价敏感学习则不再适用。

集成学习方法在提升非平衡数据分类性能方面通常比数据级方法更为有效[11]，且相比于代价敏感学习具有更好的鲁棒性[12]。在不平衡数据集上，单个模型难以准确稳定地对少数类样本进行分类，从而限制了整体性能。集成学习通过整合多个基分类器的预测结果来提高分类器的泛化能力。相比在传统机器学习领域，集成学习被广泛应用于提升多类不平衡数据的分类准确度[13-15]。在卷积神经网络（convolutional neural network, CNN）中，文献 [16] 基于增强随机特征子空间的集成 CNN，会在训练过程中自适应地重采样数据集以生成多个分类器，并将它们整合成级联的集成模型，提高了最终分类的准确性。

传统的重采样和重加权方法计算复杂，容易丢失信息引入新的误差进而过拟合。尽管已有较多采用集成学习研究不平衡数据分类的问题，但是采用基于图神经网络的集成学习方法研究不平衡图数据节点分类却鲜有报道。受集成学习启发，为提高整体分类性能，本文提出了一种改进的 Bagging 方法用于处理不平衡数据集上的节点分类问题。该方法将训练集平均分成k个子集，并选择不同的多个 GNN 模型。每个模型在k个子训练集上进行训练，得到k个不同的子模型，最终通过模型投票得到预测结果。

1 相关工作

现有的不平衡图节点分类方法主要包括重采样、重加权和集成学习。

1.1 重采样

重采样具体可分为两种类型：通过对少数类样本插值或基于生成对抗训练生成新的少数类样本，称为过采样。由于基于少数样本生成，过采样会导致样本集中在某些区域，使得模型的泛化能力下降，且对训练集中的噪声十分敏感。另一种则是通过对多数类样本进行选择性丢弃，称为欠采样。欠采样可能会导致关键特征的丢失，使得模型只学习了训练集的部分特征，进而导致模型的欠拟合[17]。GraphSMOTE[18] 和 mGNN[19] 使用合成少数类节点嵌入到图中，但它们也继承了 SMOTE 方法的缺点：合成样本过度集中在某些区域、对噪声十分敏感。GNN-CL[20] 进一步利用注意力机制来改进图中合成节点和原始节点之间的边缘生成过程，但是增加了算法的复杂度和过拟合风险。

1.2 重加权

重加权的基本思想是通过改变不同类别在模型训练过程中的权重，以平衡不同类别的影响。即在不平衡问题中，通常给予少数类别样本更高的权重，使得模型更加关注少数类，以达到平衡不同类别的效果[21]。RA-GCN[22] 利用加权网络来学习节点的权重。ReNode[23] 提出了改变区域中心节点的权重来解决图上的拓扑不平衡问题，但缺点在于计算量巨大，且只适用于均匀连通图。

1.3 集成学习

集成学习是一个聚合多个基本模型以提高性能的过程。在处理不平衡相关问题上，集成学习中较为成熟的实现主要采用 AdaBoost 方法[24]，它是

Boosting 方法的一种。AdaGCN[25] 集成了 AdaBoost 和 GCN，以获得更深层次的网络模型。另外，Stacking[26] 也是集成学习方法的一种，它可以在不同模型输出的结果上构建次级模型重新训练，但这种方法极易过拟合，需要采取许多正则化方法缓解过拟合。此外，还有一些其他方法来处理不平衡数据集问题。文献 [27] 基于迁移学习的方法通过将从多数类学习到的特征迁移到少数类来解决问题。文献 [28] 使用领域自适应方法处理不同类型的数据，并学习如何自适应地重新加权。

1.4　图神经网络

GNN 是一类专门用于处理图结构数据的深度学习模型，它们能够学习节点和边在图结构中的复杂关系，从而实现对节点的特征表示和图结构的分析。在 GNN 中，每个节点都有初始的特征表示，通常表示为一个向量，GNN 通过消息传递的方式在图上传播信息，它通过图中的边将邻居节点的信息传递给目标节点，目标节点通过从邻居节点汇聚信息更新当前的节点表示，这个过程可以迭代多次以提高模型的表达能力。

以最经典的 GCN 为例，图的信息传播规则为：

$$\boldsymbol{H}^{(L+1)}=\sigma\left(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(L)}\boldsymbol{W}^{(L)}+\boldsymbol{b}^{(L)}\right) \tag{1}$$

式中，$\boldsymbol{H}^{(L+1)}$表示第$(L+1)$层的节点表示矩阵；$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$是邻接矩阵$\boldsymbol{A}$加上单位矩阵；$\hat{\boldsymbol{D}}$是$\hat{\boldsymbol{A}}$的度矩阵；$\boldsymbol{W}^{(L)}$是第$L$层的权重矩阵；$\sigma$是激活函数，常用的激活函数如 ReLu。研究发现，通过叠加多层 GNN 模型可以增强模型的表达能力[29]。

2　模型设计

2.1　相关定义

本文首先对文中使用到的符号进行定义和描述。在一个无向图$G(\boldsymbol{A},\boldsymbol{X},Y)$中，$\boldsymbol{A}\in\mathbb{R}^{N\times N}$表示无自环的邻接矩阵，$N$表示节点的数量。$A_{ij}=1$表示在节点$i$和节点$j$之间存在一条连边，$A_{ij}=0$表示节点$i$和节点$j$之间没有连边。$\boldsymbol{X}\in\mathbb{R}^{N\times d}$表示特征矩阵，其中每一行代表一个节点的特征向量，$d$表示特征向量的维度。$Y$表示所有节点的类别集合，节点$i$的标签用$Y_i$表示，$Y_i\in\{1,2,\cdots,L\}$，$L$表示类别的数量，节点集$V$包含所有节点。

本研究专注于不平衡图上的节点分类问题。在不平衡问题中，如果在训练模型时忽视不平衡的分布，可能会导致无法有效地对少数样本进行分类。因为多数类节点占据主导地位，压倒了少数类节点，从而导致训练过程偏向于多数类。上述给定图G即为不平衡图，本文的目标是构建一个节点分类器f：

$$f(V,\boldsymbol{A},\boldsymbol{X})\rightarrow Y \tag{2}$$

2.2　改进的 Bagging 集成方法

由于传统的重采样和重加权方法在大规模数据集上计算复杂，往往需要设计新的网络结构，还会丢失信息，所以本文设计了一种简单的改进的 Bagging 集成方法。首先，图 1 选择了 3 个不同类型的图神经网络（GNN）模型，分别是 GCN、GAT[30]、ChebNet[31]。这些模型各自具有不同的图特征提取能力。接着，采用了k折交叉验证的方法来训练这些模型。首先，将训练集随机划分为k个大小相等的子训练集。随后，依次选择 1 个子训练集作为验证集，其余$k-1$个子训练集合并作为训练集，对每个 GNN 模型分别进行训练。这样，每个 GNN 模型都会得到 k 个基于不同训练集的版本，一共得到$3k$个模型。最后，利用这些经过训练的模型对测试集进行预测，每个模型对测试集的样本进行分类。最终的分类结果通过对这些模型的预测结果进行投票或者平均，来确定每个样本所属的类别。

这种方法结合了不同 GNN 模型和交叉验证的思想，能够更充分地利用数据进行模型训练，并且通过集成多个模型的预测结果，可以提高分类的准确性和鲁棒性。

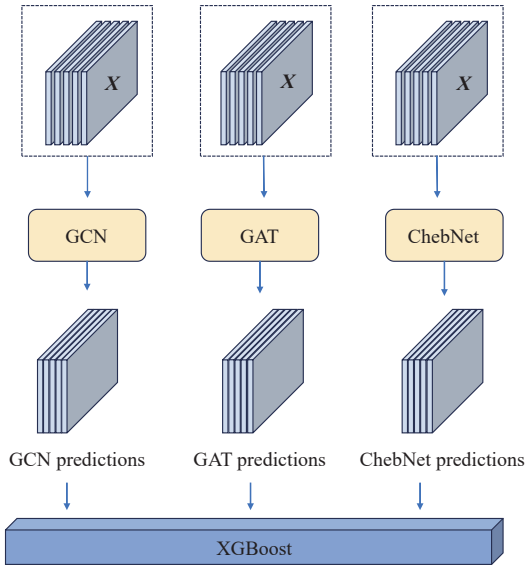

图 1　改进 Bagging 集成方法结构图

2.3 改进的 Stacking 集成方法

在 Stacking 集成方法中，与改良 Bagging 集成方法类似，同样使用 k 折交叉验证的方法划分训练集。图 2 以 GCN，GAT 和 ChebNet 为例，将训练集随机划分为k个大小相等的子训练集。然后，依次选择其中一个子训练集作为验证集，其余$k-1$个子训练集合并作为训练集。这样，针对每个训练轮次，每个 GNN 模型都会得到一个预测集，其大小等于未使用子训练集的样本数量。经过k次迭代后，基于图中输入的 3 个 GNN 模型，将会得到对应的 3 个预测集，每个预测集的大小都等于完整训练集的样本数量。接下来利用这些预测集作为训练集来构建次级模型。这个次级模型不需要具备很强的特征提取能力，因为第一层的基模型已经提取出了足够强的特征。在这里，选择使用 XGBoost[32] 作为次级模型。最终，利用构建好的次级模型对测试集进行预测，以得到最终的目标预测结果。

Stacking 方法能够充分利用基模型的预测结果来建立次级模型，从而进一步提高模型的泛化能力和预测性能。

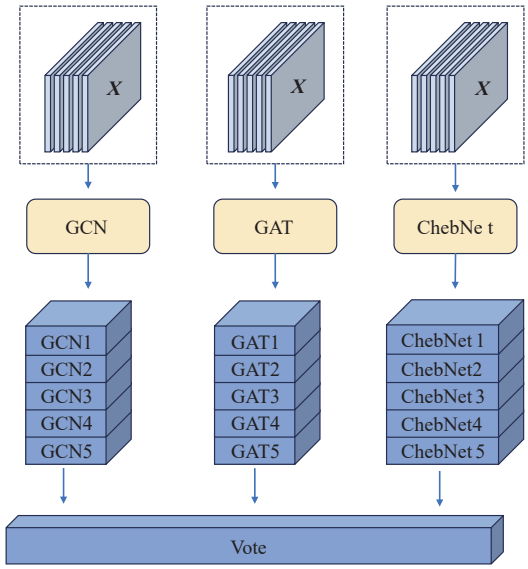

图 2 改进 Stacking 集成方法结构图

3 实验结果与分析

3.1 数据集

为了证明模型的有效性，本文在广泛使用的 33 个引文网络（包括 Cora[33]、Citeseer[33]、Pubmed[34]）和一个商品网络 Computers 上进行了实验。引文网络中节点表示文章，节点的标签代表文章的类别，两个节点之间的连边表示文章之间的引用关系。商品网络中节点表示计算机相关的商品，两个节点之间的连边表示商品被一起购买。数据集的具体情况如表 1 所示。

表 1 数据集信息

数据集	节点数	边数	类别数	特征数
Cora	2 708	5 429	7	1 433
Citeseer	3 327	4 732	6	3 703
Pubmed	19 717	44 338	3	500
Computers	13 752	491 722	10	767

3.2 基线模型

1）图卷积神经网络（GCN）利用节点邻居的信息聚合更新每个节点的表示，基于卷积操作在图上进行学习，使得模型能够有效地学习和捕获图数据中的局部和全局特征。

2）图注意力网络（graph attention network, GAT）利用注意力机制在图结构数据中进行学习，允许节点自适应地聚合邻居信息，从而为每个节点赋予不同的重要性，以实现更精准的信息传递和学习。

3）切比雪夫网络（chebyshev network, ChebNet）使用切比雪夫多项式来近似图卷积操作，该方法基于拉普拉斯特征映射，利用 k 阶切比雪夫多项式来近似卷积操作，通过局部邻域聚合节点特征，实现了在图数据上的卷积操作。

4）图采样聚合（GraphSAGE[35]）采用多轮采样节点的邻居子图，并通过聚合这些子图中节点的特征来更新节点表征。这种方法允许节点根据其邻居的局部信息自适应地聚合特征，使得模型更适用于大规模图数据。

5）图同构网络（GIN[36]）通过改进卷积层设计，在处理图同构性方面表现更为强大，提升了模型在图结构数据上的表示能力。

6）个性化 PageRank（APPNP[37]）利用节点之间的传播特性来提高节点分类的性能，通过迭代式地近似计算节点间的信息传播，融合传播和神经网络的优势，从而提高节点分类任务的准确性和效率。

7）混合跳跃网络（MixHop[38]）采用了多项式的混合方式来提取节点特征，通过不同阶数的多项式来捕获不同层次的图结构信息，将多项式卷积与图神经网络相结合，从而更好地处理图数据任务。

3.3 参数设置

对于所有算法，学习率初始化为 0.01，权重衰减为 0.000 5。在 GAT 中，注意力头设置为 8；在 ChebNet 中，多项式阶数设置为 2；在 APPNP 中，

传播步数设置为 10，阻尼系数设置为 0.1。不平衡比例指在数据集中数量较多的类别与数量较少的类别之间的比值。参考文献 [7] 和文献 [23] 的不平衡比例设置，发现在 Cora、Citeseer 等数据集中占比最多的类别是最少类别的 2～4 倍，所以为了进一步评估模型在极端情况下的性能和统一实验参数，将不平衡比例设置为 3 表示数据集轻度不平衡，设置为 5 表示数据集中度不平衡，设置为 10 表示数据集严重不平衡。如未指定，数据集不平衡比例默认为 10，折数k默认置为 5。

3.4　实验结果分析

对于合成的不平衡数据集，GCN、GAT、ChebNet、GraphSAGE、GINConv、APPNP、MixHop、Stacking 和改进 Bagging 的分类精度如表 2 所示。结果表明，本文方法在处理不平衡数据集方面优于经典的 GNN 模型和 Stacking 模型。Stacking 方法通过不同基模型卷积得到新的节点表示，再通过次级模型进行进一步融合，得到最终的节点预测，并且由于多层模型的复杂性，本文使用了k折交叉验证缓解过拟合，但是从结果来看 Stacking 方法只比大部分基模型好，稍微降低了模型之间的方差，依然存在较为严重的过拟合现象。改进 Bagging 方法使用k折划分分别训练子模型，直接使用投票融合模型，结果显示 GNN 模型的分类精度显著提高，相比于 Stacking 方法进一步降低了方差，并且过拟合现象不明显，在 Cora、Citeseer、Pubmed 和 Computers 数据集中，与原始 GNN 模型相比，在不平衡比例为 3 时，数据集的平衡程度类似于数据集的原始状况，各个基本模型的表现相近，Bagging 模型的精度在 4 个数据集上分别至少提高了 1.21%、1.00%、1.70% 和 1.20%；在不平衡比例为 10 的情况下至少提高了 3.14%、4.17%、2.00% 和 3.80%，在不平衡比例为 5 的情况也至少提高了 3.57%、3.17%、1.00% 和 4.80%。

表 2　不同模型的节点分类精度　　%

模型	Cora			Citeseer			Pubmed			Computers		
	比例3	比例5	比例10	比例3	比例5	比例10	比例3	比例5	比例10	比例3	比例5	比例10
GCN	82.50	75.57	73.86	60.33	57.00	53.67	76.00	67.67	57.33	83.10	67.30	54.30
GAT	82.29	76.14	74.00	59.50	56.00	53.17	77.67	72.00	60.67	85.30	80.00	75.30
Cheb	81.00	74.00	60.71	59.67	53.83	51.00	76.67	66.33	59.67	49.70	40.30	24.30
GraphSAGE	82.43	76.86	68.43	60.00	56.33	53.00	75.67	67.67	59.33	42.00	20.70	25.20
GIN	80.00	74.57	72.71	57.83	55.50	51.00	77.67	70.33	65.33	33.40	17.90	10.00
APPNP	81.71	75.43	70.86	62.17	58.00	54.50	57.33	56.00	55.67	79.20	78.30	63.80
MixHop	79.71	70.43	60.29	62.33	58.33	54.67	74.00	65.33	58.33	80.90	69.10	57.50
Stacking	77.10	76.28	67.85	60.50	55.16	53.00	78.30	74.66	61.33	85.90	78.40	67.10
Bagging	**83.71**	**80.43**	**77.14**	**63.33**	**61.50**	**58.83**	**80.00**	**75.67**	**67.33**	**87.10**	**84.80**	**79.10**

对单个基础模型进行了 Bagging 实验。表 3 和表 4 显示，对单个 GNN 模型进行 Bagging 训练多个子模型，并不能保证在数据集不平衡程度较轻时模型精度的提升，其中，B 表示 Bagging。具体而言，当不平衡比例为 5 时，大多数图神经网络（GNN）模型在采用 Bagging 后精度出现下降，不平衡比例增至 10 时，大多数 GNN 模型的精度才得到提升。这表明集成不同的子模型并非总能够提升模型的精度，甚至可能引起过拟合。

为此，本文进行了逐个增加基础模型的 Bagging 实验。表 5 显示了逐个累计增加基础模型的精度变化，结果显示增加基础模型的个数有益于精度的提升，降低了算法的方差，其中，B 表示 Bagging。同时，我们也注意到这种精度提升随着基础模型数量的增加呈现放缓的趋势，精度的提升越来越受限于 GNN 模型本身。

表 3　不平衡比例为 5 时单个模型使用 Bagging 后的精度　　%

模型	数据集			
	Cora	Citeseer	Pubmed	Computers
GCN	**75.57**	57.00	**69.67**	67.30
B-GCN	75.00	**57.33**	64.67	**68.50**
GAT	**76.14**	**56.00**	**72.00**	80.00
B-GAT	74.29	55.17	67.33	**82.40**
Cheb	**74.00**	53.83	**66.33**	40.30
B-Cheb	68.29	**54.33**	66.00	**45.30**
GraphSAGE	**76.86**	**56.33**	**67.67**	20.70
B-GraphSAGE	75.57	55.33	60.67	**34.80**
GIN	**74.57**	**55.50**	**70.33**	17.90
B-GIN	73.00	52.33	68.33	**21.00**
APPNP	**75.43**	**58.00**	**56.00**	**78.30**
B-APPNP	75.00	55.50	54.67	77.10
MixHop	**70.43**	**58.33**	**65.33**	69.10
B-MixHop	69.43	57.00	62.33	**74.20**

表 4 不平衡比例为 10 时单个模型使用 Bagging 后的精度 %

模型	数据集			
	Cora	Citeseer	Pubmed	Computers
GCN	73.86	53.67	57.33	54.30
B-GCN	**75.57**	**54.00**	**61.67**	**58.50**
GAT	**74.00**	53.17	60.67	75.30
B-GAT	73.43	**55.17**	**61.00**	**76.70**
Cheb	60.71	51.00	59.67	24.30
B-Cheb	**62.14**	**51.83**	**61.00**	**26.50**
GraphSAGE	68.43	**53.00**	59.33	25.20
B-GraphSAGE	**70.00**	52.67	59.33	**31.70**
GIN	**72.71**	51.00	**65.33**	10.00
B-GIN	71.86	**53.83**	59.67	**12.10**
APPNP	70.86	**54.50**	**55.67**	63.80
B-APPNP	**71.14**	54.33	55.33	**67.90**
MixHop	60.29	**54.67**	58.33	**57.50**
B-MixHop	**60.71**	54.50	58.33	56.70

表 5 使用 Bagging 逐个增加模型后的精度 %

模型	Cora		Citeseer		Pubmed		Computers	
	比例5	比例10	比例5	比例10	比例5	比例10	比例5	比例10
B+GCN	75.00	75.57	57.33	54.00	64.67	61.67	68.50	58.50
+GAT	76.71	76.14	58.67	55.67	68.00	62.33	73.49	64.58
+Cheb	77.29	76.43	59.83	56.50	70.71	64.67	77.41	68.43
+GraphSAGE	78.29	76.71	60.67	56.67	72.00	65.83	79.12	71.17
+GIN	79.71	76.86	61.00	57.33	73.33	66.91	81.87	74.15
+APPNP	79.91	77.14	61.29	58.17	74.33	67.17	83.55	77.22
+MixHop	80.43	77.14	61.50	58.83	75.67	67.33	84.80	79.10

在本文方法中，采用不同的基础模型训练各自的子模型，以充分学习数据集的多样特征、减少对异常值的敏感性，进而增强模型的鲁棒性。这种方法也进一步验证了在数据集不平衡程度较轻时，传统 GNN 模型仍然具备其优点，这一发现为理解和利用不同程度不平衡数据集的模型行为提供了重要解释。

3.5 不同k折数的影响

通过改变划分数据集的个数 k，来观察得到子模型个数的变化对分类器性能的影响。图 3 的实验结果显示，在一定程度上增加 k 值（增加训练集的划分数量），可以提高模型的性能。在 Cora、Citeseer 和 Pubmed 数据集上，当k从 3 增至 5 时，模型的精度略有提升。然而，当k从 5 提升至 7 和 9 时，模型的精度出现下降。在 Computers 数据集中，当k从 3 增至 5 时，模型大幅提升，6 之后精度大幅下降。结合 4 个数据集的结果，当k在 5 左右时 Bagging 模型的性能最好。这说明并非划分数量越多越好，过多的划分可能导致模型退化成为传统模型，进而使准确率下降。

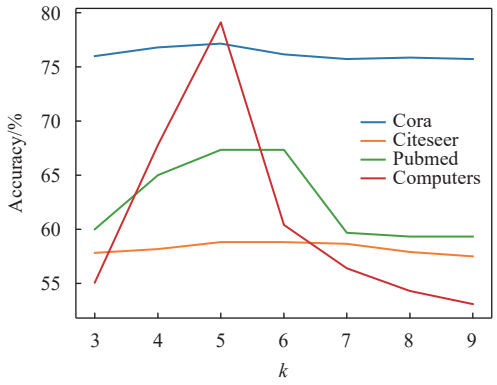

图 3 不同k值对精度的影响

这一结果提示在划分数据集时需要平衡划分数量和模型性能之间的关系。合适地划分可以提高模型的泛化能力，但划分过多可能引起过拟合问题。因此，在训练模型时，需要仔细调整划分数量，以确保模型在处理数据集时能够取得最佳性能。

3.6 鲁棒性分析

在图数据中随机添加虚假边来增加扰动，以测试不同模型在这种情况下的鲁棒性，结果见表 6～表 8。

表 6 Cora 数据集上不同假边比例对精度的影响 %

模型	假边比例								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GCN	69.71	64.86	62.57	59.86	58.43	51.14	51.71	48.86	50.29
GAT	70.57	62.57	61.14	62.29	55.14	51.43	51.43	47.29	47.43
Cheb	61.00	57.00	54.86	54.71	51.71	48.71	49.00	47.86	48.43
Graphsage	64.71	55.86	60.43	58.57	55.71	54.86	59.43	50.71	50.00
GIN	59.86	59.57	54.71	49.29	56.14	42.57	35.71	44.57	34.71
APPNP	62.71	57.43	54.00	53.00	49.71	48.57	49.86	48.86	47.71
Mixhop	57.71	56.29	53.00	53.86	51.29	50.86	48.86	51.29	50.14
Bagging	73.00	66.57	62.43	60.43	57.43	56.57	54.71	53.57	55.00

表 7　Citeseer 数据集上不同假边比例对精度的影响

%

模型	假边比例								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GCN	49.50	49.33	47.50	44.83	43.00	41.00	39.67	39.33	37.17
GAT	47.00	48.00	45.83	43.83	42.83	40.17	38.83	35.83	37.50
Cheb	52.50	52.00	48.33	47.33	45.17	46.67	45.17	41.33	40.17
Graphsage	50.33	51.00	49.67	46.67	43.83	44.83	40.33	39.67	35.67
GIN	47.33	47.33	45.83	45.50	43.17	38.33	36.50	34.50	32.33
APPNP	51.00	49.00	48.33	42.17	40.33	37.83	40.33	39.17	38.83
Mixhop	55.50	53.33	50.67	46.83	44.00	41.50	41.33	40.83	42.00
Bagging	56.33	53.33	50.00	46.17	46.83	45.33	44.83	43.00	42.00

表 8　Pubmed 数据集上不同假边比例对精度的影响

%

模型	假边比例								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GCN	60.67	55.00	54.33	53.00	50.67	49.67	52.00	48.00	47.67
GAT	61.33	57.33	53.33	53.00	48.67	49.00	48.00	49.33	47.00
Cheb	57.67	56.33	57.67	56.33	52.67	55.67	54.67	50.67	48.33
Graphsage	61.00	60.67	59.00	56.67	54.67	54.00	55.67	53.33	51.17
GIN	61.33	63.33	64.00	54.67	52.33	51.67	52.67	51.67	49.43
APPNP	52.33	53.33	51.00	52.00	50.33	51.33	52.33	52.00	46.17
Mixhop	59.00	55.67	55.67	56.00	57.67	54.67	54.00	49.67	40.00
Bagging	66.00	64.67	63.67	62.33	62.33	60.00	59.86	57.67	55.50

图 4～图 6 展示了在 Cora、Citeseer、Pubmed 数据集上，不同虚假边比例下，各模型的分类准确度。相较于基础模型，在大多数情况下，本文方法表现出更高的准确度。尤其值得注意的是，相对于其他传统的 GNN 模型，本文方法在增加假边比例时，准确度下降的速率更为平缓。在 Cora 数据集中，在假边比例达到 0.9 时，本文方法准确度仅下降了 23.57%，而 GCN 下降了 25.00%，GAT 下降了 26.57%，GIN 甚至下降了 38.00%。实验结果表明，随着扰动率的增加，本文模型表现出更强的鲁棒性。

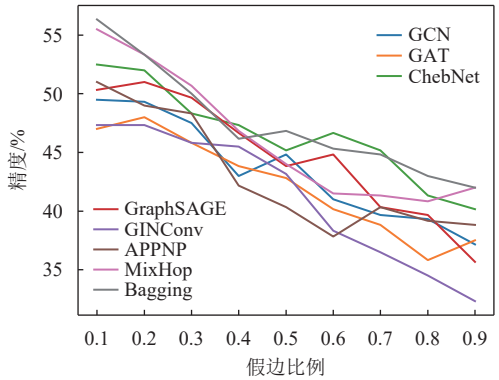

图 5　Citeseer 数据集上不同假边比例对精度的影响

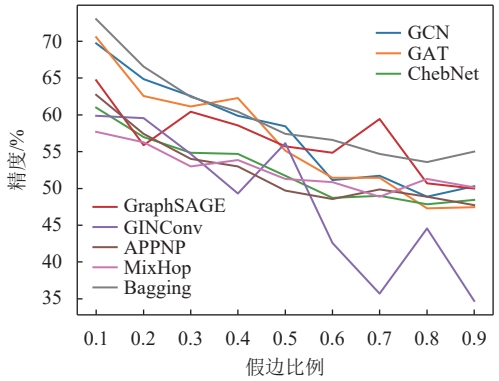

图 4　Cora 数据集上不同假边比例对精度的影响

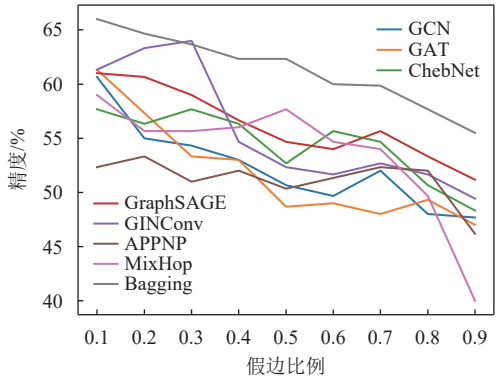

图 6　Pubmed 数据集上不同假边比例对精度的影响

4 结束语

本文探讨了在不平衡图上进行节点分类任务的问题，并提出了一种改进的 Bagging 集成学习方法。这一方法不仅提高了节点分类的准确性，还缓解了传统图神经网络（GNN）所面临的非鲁棒性问题。本文提出了集成不同模型的 Bagging 方法，并采用k折划分训练不同的子模型，以充分利用数据集中的不同信息。因此，每个模型以及每个划分下的子数据集的差异性，都降低了模型对部分节点和边的依赖性，从而减轻了过拟合和过度平滑的问题。通过在 4 个数据集上的实验结果显示，本文方法在节点分类任务上超越了传统的 GNN 模型，这突显了在框架中使用集成学习的重要性。对于在引文网络数据集中随机添加连接的实验也表明，本文算法在鲁棒性方面具有优势。

在未来的研究中，将探索不同编码方式之间在不平衡数据集中是否存在相互作用，以及框架是否适用于其他基于图的任务，如链路预测任务和图的分类任务。

参 考 文 献

[1] BRODER A, KUMAR R, MAGHOUL F, et al. Graph structure in the web[J]. Computer Networks, 2000, 33: 309-320.

[2] 邢玲, 邓凯凯, 吴红海, 等. 复杂网络视角下跨社交网络用户身份识别研究综述[J]. 电子科技大学学报, 2020, 49(6): 905-917.
XING L, DENG K K, WU H H, et al. Review of user identification across social networks: The complex network approach[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(6): 905-917.

[3] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4-24.

[4] 张千桢, 郭得科, 赵翔. 面向时序图数据的季节突发性子图挖掘算法[J]. 软件学报, 2024, 35(12): 5226-5543.
ZHANG Q Z, GUO D K, ZHAO X. Mining seasonal-bursting subgraphs in temporal graphs[J]. Journal of Software, 2024, 35(12): 5226-5543.

[5] WESTON J, RATLE F, COLLOBERT R. Deep learning via semi-supervised embedding[EB/OL]. [2023-11-10]. https://icml.cc/Conferences/2008/papers/340.pdf.

[6] BUDA M, MAKI A, MAZUROWSKI M A. A systematic study of the class imbalance problem in convolutional neural networks[J]. Neural Networks, 2018, 106: 249-259.

[7] ZHU Z, XING H, XU Y. Balanced neighbor exploration for semi-supervised node classification on imbalanced graph data[J]. Information Sciences, 2023, 631: 31-44.

[8] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[9] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673-688.
LI Y X, CHAI Y, HU Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.

[10] DONG Q, GONG S, ZHU X. Imbalanced deep learning by minority class incremental rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1367-1381.

[11] KHOSHGOFTAAR T M, FAZELPOUR A, DITTMAN D J, et al. Ensemble vs. data sampling: Which option is best suited to improve classification performance of imbalanced bioinformatics data?[C]//IEEE 27th International Conference on Tools with Artificial Intelligence. [S. l.]: IEEE, 2015: 705-712.

[12] 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.

[13] SUN Z, SONG Q, ZHU X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[14] MALEK N H A, YAACOB W F W, WAH Y B, et al. Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2023, 29: 598-608.

[15] 贾承丰, 韩华, 吕亚楠, 等. 利用集成分类器处理链路预测中的分类不平衡问题[J]. 计算机应用研究, 2018, 35(12): 3644-3647.
JIA C F, HAN H, LYU Y N, et al. Handing class imbalance in link prediction using ensemble classifier[J]. Application Research of Computer, 2018, 35(12): 3644-3647.

[16] TAHERKHANI A, COSMA G, MCGINNITY T M. AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning[J]. Neurocomputing, 2020, 404: 351-366.

[17] SHEN L, LIN Z, HUANG Q. Relay backpropagation for effective learning of deep convolutional neural networks[C]//European Conference on Computer Vision. Cham: Springer, 2016: 467-482.

[18] ZHAO T, ZHANG X, WANG S. GraphSMOTE: Imbalanced node classification on graphs with graph neural networks[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. [S.l.]: ACM, 2021: 833-841.

[19] WANG K, AN J, ZHOU M, et al. Minority-weighted graph neural network for imbalanced node classification in social networks of Internet of people[J]. IEEE Internet of Things Journal, 2022, 10(1): 330-340.

[20] LI X, FAN Z, HUANG F, et al. Graph neural network with curriculum learning for imbalanced node classification[J]. Neurocomputing, 2024, 574: 127229.

[21] SUN Y, KAMEL M S, WONG A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.
[22] GHORBANI M, KAZI A, BAGHSHAH M S, et al. RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data[J]. Medical Image Analysis, 2022, 75: 102272.
[23] WU D, LUO X, GUO X, et al. Concordant contrastive learning for semi-supervised node classification on graph[C]//International Conference on Neural Information Processing. Cham: Springer, 2021: 584-595.
[24] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[EB/OL]. [2023-11-15]. https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf.
[25] SUN K, ZHU Z, LIN Z. AdaGCN: Adaboosting graph convolutional networks into deep models[EB/OL]. [2024-01-10]. http://arxiv.org/abs/1908.05081v3.
[26] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.
[27] YIN X, YU X, SOHN K, et al. Feature transfer learning for face recognition with under-represented data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5704-5713.
[28] ROY S, SIAROHIN A, SANGINETO E, et al. Unsupervised domain adaptation using feature-whitening and consensus loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9471-9480.
[29] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2023-11-10]. 1609.02907. http://arxiv.org/abs/1609.02907v4.
[30] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2023-11-15]. https://arxiv.org/abs/1710.10903.
[31] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[EB/OL]. [2023-11-25]. http://arxiv.org/abs/1606.09375v3.
[32] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [s.l.]: ACM, 2016: 785-794.
[33] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93-93.
[34] NAMATA G, LONDON B, GETOOR L, et al. Query-driven active surveying for collective classification [EB/OL]. [2023-12-12]. https://people.cs.vt.edu/~bhuang/papers/namata-mlg12.pdf
[35] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[EB/OL]. [2023-12-10]. http://arxiv.org/abs/1706.02216v4.
[36] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[EB/OL]. [2023-12-13]. http://arxiv.org/abs/1810.00826v3.
[37] GASTEIGER J, BOJCHEVSKI A, Günnemann S. Predict then propagate: Graph neural networks meet personalized PageRank[EB/OL]. [2023-12-20]. http://arxiv.org/abs/1810.05997v6.
[38] ABU-EI-HAIJA S, PEROZZI B, KAPOOR A, et al. MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing[EB/OL]. [2023-12-24]. http://arxiv.org/abs/1905.00067v3.

编　辑　叶　芳