

基于图聚类结果的半监督节点分类方法

鲍舟, 刘恒

安徽工业大学计算机科学与技术学院, 安徽 马鞍山

收稿日期: 2024年7月30日; 录用日期: 2024年8月27日; 发布日期: 2024年9月4日

摘要

目前在处理现实世界中知识图谱、引文网络以及社交网络等复杂的图结构数据分类问题上, 图卷积神经网络被认为是最有效的半监督方法之一, 但存在其学习性能会被严重有限的标记数据影响的问题。本研究针对这一问题提出了一种以图聚类结果作为指导的节点分类方法。具体来说, 引入数据增强模块减少了图结构信息中的噪声, 设计了一种面向聚类的图嵌入模型作为属性图聚类网络, 并根据聚类结果预测出节点的伪标签。同时, 为了提升分类任务的性能, 筛选出高置信度伪标签来指导图节点分类任务, 并设计了一种相似度损失来提高标记节点和未标记节点之间的特征相似度。通过在基准数据集上大量的实验结果表明, 与现有的方法相比, 该方法可以克服标签数量限制, 在图节点分类任务上表现出优越的性能。

关键词

机器学习, 图卷积神经网络, 图聚类, 图节点分类

Semi-Supervised Node Classification Method Based on Graph Clustering Results

Zhou Bao, Heng Liu

School of Computer Science and Technology, Anhui University of Technology, Ma'anshan Anhui

Received: Jul. 30th, 2024; accepted: Aug. 27th, 2024; published: Sep. 4th, 2024

Abstract

Currently, graph convolutional neural networks are considered to be one of the most effective semi-supervised methods in dealing with the classification of complex graph-structured data in real-world, such as knowledge graphs, citation networks, and social networks, but there is still the problem that their learning performance can be affected by severely limited labeled data. In this study, a node classification method using graph clustering results as a guide is proposed to address this problem. Specifically, a data enhancement module is introduced to reduce the noise in the graph

structure information, a clustering-oriented graph embedding model is designed as an attribute graph clustering network, and the pseudo-labels of the nodes are predicted based on the clustering results. Meanwhile, to improve the performance of the classification task, high-confidence pseudo-labels are screened to guide the graph node classification task, and a similarity loss is designed to improve the feature similarity between labeled and unlabeled nodes. The results of extensive experiments on the benchmark dataset show that the method can overcome the label number limitation and have superior performance on the graph node classification task compared with existing methods.

Keywords

Machine Learning, Graph Convolutional Neural Network, Graph Clustering, Graph Node Classification

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

图是表示对象之间关系的非结构化数据, 可广泛用于社交媒体网络分析、化学成分分析以及知识图谱的构建等数据挖掘任务。随着深度学习的发展, 图卷积神经网络(GCNs) [1]表现出了强大的特征提取能力。GCNs 通过信息聚合过程, 将图中每个节点的信息传递给它相邻节点, 充分利用了图的节点属性和结构信息。尽管 GCNs 在处理图结构数据上取得了成功, 图的复杂结构性, 仍然使 GCNs 在属性图聚类 and 半监督节点分类等方面面临着巨大的挑战。

当前图聚类被广泛应用于社区检测、群体分割等。图聚类的目的是将图中的节点划分为多个不相交的簇, 使得簇内部尽可能相似, 每个簇之间相差越大越好。图聚类属于无监督学习, 因此, 如何捕获图的结构信息和利用节点内容属性成为图聚类的一个关键问题。传统图聚类的主要方法是通过图嵌入方法获得每个节点的特征, 如 DeepWalk [2]、Node2Vec [3]以及 SDNE [4]等, 然后通过 K-means [5]算法得出聚类结果。然而这些方法只利用了图的结构信息, 忽略了图的节点属性, 不能探索图的更深层次的潜在信息。

受经典的嵌入方法的启发, 一些基于 GCNs 模型将图卷积提取出的特征应用于 K-means, 并将重建结构信息作为监督信号指导嵌入特征的生成, 如图自编码器(GAE) [6]和图变分自动编码器(VGAE) [6]; 此外 Pan 等人在这种自编码器结构基础上, 引入了对抗的思想, 提出了对抗正则化图自编码器(ARGA) [7]。然而, 这些方法将图的嵌入和聚类的过程分开了。基于这些方法, 深度注意嵌入式图聚类(DAEGC) [8]引入了一种图注意机制, 根据嵌入特征生成目标分布作为监督信息, 指导模型端到端学习聚类; Mrabah [9]等人则通过使用高置信度节点集更新聚类友好的结构信息, 将图的嵌入和聚类过程紧密联系在一起。显然, 这些基于图自编码器的嵌入特征学习方法可以应用于纯无监督环境, 但这些方法大都依赖于重构的结构信息作为监督信号, 因此很容易受到图的结构信息中噪声的影响。Zhao [10]等人提出了一种针对图结构数据的数据增强模块, 能够删除图中冗余的边以及添加可能存在的边, 在一定程度上减少噪声的影响。

无监督学习可以在许多方法中有效地提高深度模型的性能。在计算机视觉领域, Caron [11]等人使用深度聚类生成的伪标签来指导视觉特征的学习, 显著提高了学习后的特征的识别能力; Xie [12]等人通过对样本进行数据增强, 并利用一致性损失约束预测标签在训练过程中的不变性。得利于无监督学习在计算

机视觉领域的成功, 属性图聚类等无监督学习也被用来提高图领域的半监督学习性能, 例如图节点分类。

目前 GCNs 是处理半监督节点分类任务的有效手段之一。然而, GCNs 有一个内在的局限性, 即它不能有效地将标签传播到整个图中。GCNs 的训练仍然需要大量的标记数据来进行参数调整和模型选择, 这违背了半监督学习的目的。众所周知, 训练深度神经模型通常需要大量的标记数据, 而由于标记训练数据的成本高, 在许多情况下都无法满足。因此当标记的数据被严重限制时, 现有的基于 GCNs 的技术表现不佳。许多研究表明, 如果充分利用在训练中利用未标记的数据, 可以显著提高学习的准确性。Lin [13] 等人了解决标记数据限制问题, 利用随机游走[14]发现和 GCNs 本身发现的高置信度的预测结果来扩展标签集, 然后使用扩展的标签集来继续训练模型。但这种联合训练方式对预测出的伪标签准确率要求很高, 否则训练过程中的累计误差会影响最终分类结果。此外, Sun [15]等人提出了多阶段自我训练处理, 但仅采用嵌入深度聚类, 依赖距离测度进行对齐, 没有考虑使用更合适的属性图聚类方法。在此基础上, Hui 等人提出了一种协作图卷积网络(CGCN) [16], 通过将高斯混合模型(GMM)和 VGAE 结合(GMM-VGAE)完成属性图聚类, 并根据聚类结果指导半监督学习。这些工作都证明了当可用的标记样本的数量有限时, 利用大量未标记样本的潜力来加强 GCNs 的性能是值得的。

基于上述研究, 本文设计了一种改进的图卷积神经网络。该网络包括两个部分: 属性图聚类网络和半监督节点分类网络。具体来说, 与 Hui 等人的工作不同, 本文提出的属性图聚类网络是基于 GAE 框架, 该网络首先将输入的图结构数据进行数据增强处理, 一定程度上减少图结构信息中的噪声, 然后通过加入了图注意力机制(GAT) [17]的融合编码器提取出图的嵌入特征, 最后通过重构损失和聚类分布损失指导模型学习高性能的聚类分布。同时, 根据属性图聚类网络得出的聚类结果来预测出伪标签, 并筛选出高置信度节点的伪标签来扩充标签集, 以此来指导半监督节点分类网络的训练。最后, 本文设计了一种新的相似性损失函数, 用以提高标记节点和高置信度伪标签所对应节点的嵌入特征相似性。实验证明了在基准数据集上所提出的方法的可行性, 并在对比实验中与现有方法进行指标上的对比, 进一步证明了所提出的方法在标签数据稀少的情况下, 在半监督节点分类任务上的有效性。

2. 网络模型结构

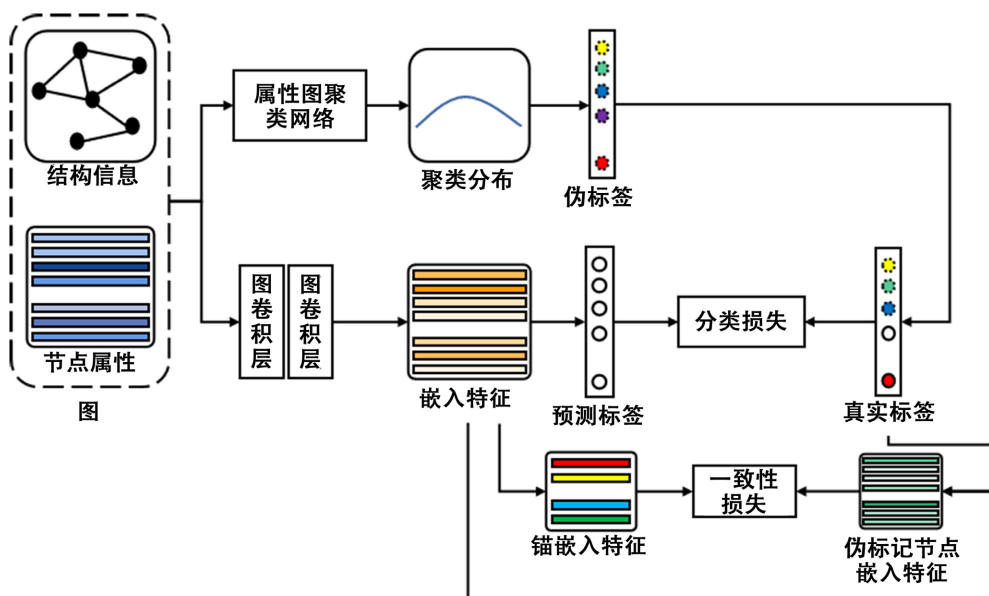


Figure 1. The overall model structure
图 1. 总体网络结构图

图 1 是改进后的图卷积神经网络的总体结构图, 网络包括两个部分: 属性图聚类网络和半监督分类网络。输入的内容为图结构数据 $G=(V, E, X)$, 其中 $V=\{v_1, v_2, \dots, v_n\}$ 表示图中的节点集, $E=\{(v_i, v_j)\}$ 表示节点 i 与节点 j 之间的边。图的结构信息可以用邻接矩阵 $A=\{a_{ij}\}$ 来表示, 对于任意 $v_i, v_j \in V$, 如过存在 $(v_i, v_j) \in E$, 则 $a_{ij}=1$; 否则 $a_{ij}=0$ 。 $X=\{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{n \times d}$ 表示图的节点属性, 其中 d 是每个节点的特征维度。

2.1. 半监督分类网络

图 G 首先经过预训练过属性图聚类网络得到聚类结果 $Q=\{q_{iu}\}$, 并预测出图中每个节点的伪标签来扩充半监督分类任务的标签集, 得到图的伪标签指示矩阵 $Y' \in \mathbb{R}^{n \times K}$:

$$\begin{cases} y'_{ik} = 1, k = \arg \max_u q_{iu} \\ y'_{ik} = 0, k \neq \arg \max_u q_{iu} \end{cases} \quad (1)$$

其中为 K 标签类别数。

然后图 G 通过两层图卷积层得到图的嵌入特征 $Z \in \mathbb{R}^{n \times h}$, 图卷积层的特征提取可以被表示为:

$$Z = \text{softmax}\left(\tilde{A} \text{ReLU}\left(\tilde{A} X W^{(0)}\right) W^{(1)}\right) \quad (2)$$

其中 $\tilde{A} = \tilde{D}^{-1/2} (A + I) \tilde{D}^{-1/2}$ 为归一化后的邻接矩阵, \tilde{D} 为矩阵 $(A + I)$ 的度矩阵, $W^{(0)}$ 、 $W^{(1)}$ 表示图卷积层的权重矩阵, h 为图的每个节点嵌入特征维度。

接下来, 根据嵌入特征 Z 以及属性图聚类网络预测的伪标签指示矩阵 Y' , 从未标记节点集 V_u 中筛选出高置信度的节点, 并作为标记数据指导半监督学习。节点如果满足公式(3)和(4)则被认为是高置信度节点, 并将其加入伪标记节点集 V_p 和标记节点集 V_l 中。

$$\max_h z_{ih} \geq \alpha_1 \quad (3)$$

$$\arg \max_h z_{ih} = \arg \max_k y'_{ik} \quad (4)$$

其中 α_1 为置信度阈值。

与此同时, 根据标记节点集的节点嵌入特征计算出每个类所对应的语义信息, 语义信息的计算可以表示为:

$$C_k = \frac{1}{|V_k|} \sum_{v_i \in V_k} Z_i \quad (5)$$

其中 $V_k \subset V_l$ 表示第 k 类标记节点集, $C \in \mathbb{R}^{K \times h}$ 为锚嵌入特征, 表示每个类所对应的语义信息。

最后, 网络预测出每个节点最终所对应的标签, 具体来说, 图中任意节点 i 的预测标签可表示为:

$$y_i'' = \arg \max_h z_{ih} \quad (6)$$

2.2. 属性图聚类网络

图 2 是属性图聚类网络的结构图。受 Zhao [10] 等人的启发, 本文对图的原始结构信息进行数据增强, 减少特征提取过程中累积噪声带来的影响, 具体是利用概率矩阵 $M=\{m_{ij}\}$ 删除图中冗余的边, 增加可能存在的边, 其中 m_{ij} 表示节点 i 和节点 j 之间存在一条边的概率。概率矩阵通过边缘预测器得出, 本文使用预训练过的 GAE 作为边缘预测器, 该过程可表示为:

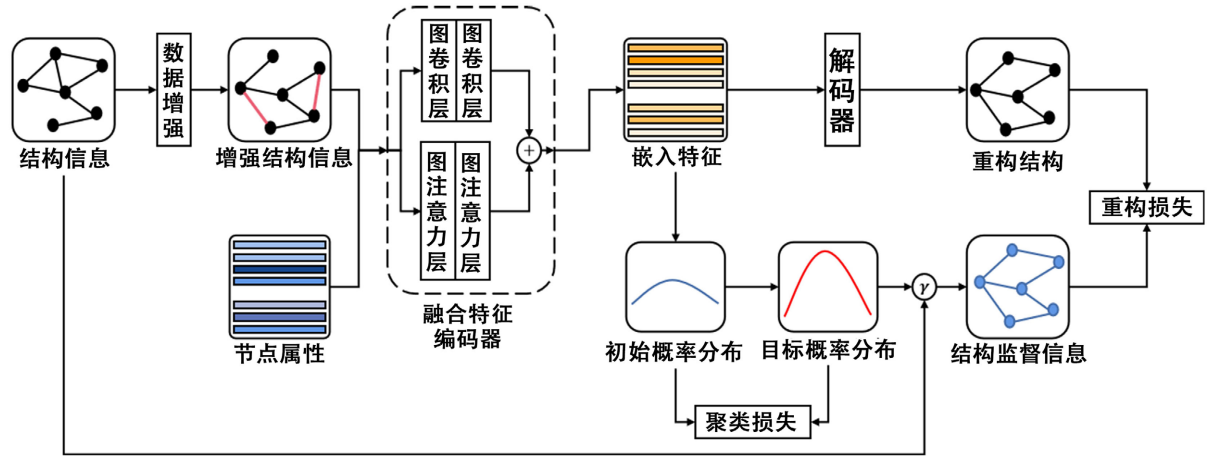


Figure 2. Attribute graph clustering network
图 2. 属性图聚类网络

$$M = \text{sigmod}\left(f_{\text{gcn}}(A, X) \cdot f_{\text{gcn}}(A, X)^T\right) \quad (7)$$

其中 $f_{\text{gcn}}(\cdot)$ 表示两层图卷积操作, 根据概率矩阵 M , 我们添加前 $\theta_1|V|$ 个可能存在的边, 并删除前 $\theta_2|V|$ 个最不可能存在的边, 得到数据增强后的结构信息 A' 。

本文引入图注意力(GAT)机制, 设计了一个融合特征编码器, GAT 的特征提取过程可以表示为:

$$f_{\text{gat}}(A', X) = Z' \quad (8)$$

$$z'_i = \frac{1}{B} \sum_{b=1}^B \sum_{j \in \mathbb{N}_i} s_{ij} W'_b x_j \quad (9)$$

其中 B 图注意力层中多头注意力的头数, \mathbb{N}_i 为节点 i 的邻居节点, w' 为图注意力层的权重。 s_{ij} 表示节点 i 和节点 j 之间的相似度系数, 可表示为:

$$s_{ij} = \text{softmax}\left(f'(W'x_i, W'x_j)\right) \quad (10)$$

其中 $f'(\cdot): \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}$ 为映射函数。

接下来, 将图的节点属性 X 和增强后的结构信息 A' 作为融合编码器的输入, 得到最终融合嵌入特征:

$$\tilde{Z} = f_{\text{gcn}}(A', X) + \beta f_{\text{gat}}(A', f_{\text{gat}}(A', X)) \quad (11)$$

其中 β 为线性融合加权系数。更进一步地, 通过解码器重构出图的结构信息, 这一过程可表示为:

$$\hat{A} = \text{sigmod}(\tilde{Z} \cdot \tilde{Z}^T) \quad (12)$$

其中 \hat{A} 为图的重构结构信息。

之后, 根据 t-分布[18]计算出计算每个节点嵌入特征和聚类质心特征的相似性, 得出属性图聚类网络输出的聚类结果, 即初始概率分布 $Q = \{q_{iu}\}$:

$$q_{iu} = \frac{\left(1 + \|\tilde{z}_i - \mu_u\|^2\right)^{-1}}{\sum_K \left(1 + \|\tilde{z}_i - \mu_K\|^2\right)^{-1}} \quad (13)$$

其中 $\{\mu_u\}_{u=1}^K$ 表示每个类的质心特征。同时, 根据初始聚类分布 Q 进一步计算目标概率分布 $P = \{p_{iu}\}$, 作为模型聚类过程中的监督信息:

$$P_{iu} = \frac{q_{iu}^2 / \sum_i q_{iu}}{\sum_k (q_{ik}^2 / \sum_i q_{ik})} \quad (14)$$

为了提高聚类性能, 本文筛选出高置信节点集 V_h , V_h 中的任意节点 i 满足:

$$\max_u q_{iu} > \alpha_2 \quad (15)$$

$$\max_u (q_{iu} | q_{iu} < \alpha_2) > \frac{\alpha_2}{2} \quad (16)$$

其中 α_2 为置信度阈值。

最后, 为了将嵌入过程和聚类过程紧密联系在一起, 本文通过 γ 算子[9], 根据高置信节点集 V_h 的初始概率分布更新结构信息 A , 并删除 A 中的冗余边, 添加可能的边来增强结构信息, 然后得到以聚类结果作为指导的结构监督信息 A'' 。

3. 损失函数

本节主要介绍网络在训练过程中所用到的损失函数, 主要包括属性图聚类网络的重构损失和聚类损失, 半监督分类网络的交叉熵损失和一致性损失。

3.1. 属性图聚类网络损失

本文通过重构损失来指导属性图聚类网络的图嵌入过程, 计算图的重构结构信息 \hat{A} 和聚类指导的结构监督信息 A'' 之间的二元交叉熵损失:

$$L_{rec} = \text{BCE}(\hat{A}, A'') \quad (17)$$

为了使初始概率分布 Q 逼近于目标概率分布 P , 本文将聚类损失定义为 KL 散度损失:

$$L_{clu} = \sum_{v_i \in V_h} \sum_u P_{iu} \log \frac{P_{iu}}{q_{iu}} \quad (18)$$

最终, 属性图聚类网络的总损失可以表示为:

$$L_{attribute} = \lambda_1 L_{rec} + \lambda_2 L_{clu} \quad (19)$$

3.2. 半监督分类网络损失

在半监督分类任务中本文计算所有标记数据的交叉熵误差, 因此分类损失可以定义为:

$$L_{ce} = -\sum_{v_i \in (V_l \cup V_p)} \sum_{k=1}^K Y_{ik} \ln Z_{ik} - \sum_{v_i \in V_p} \sum_{k=1}^K Y'_{ik} \ln Z_{ik} \quad (20)$$

其中 Y 为图的真实标签指示函数。值得注意的是, 在计算分类损失时, 对于标记数据集中新增的高置信度节点, 本文使用其伪标签而不是真实标签。

根据聚类假设, 如果图中的节点在同一聚类中, 它们很可能属于同一类。因此, 为了提高分类性能, 本文通过一致性损失函数提高标记节点和新增标记节点之间的特征相似度, 具体来说, 是计算伪标签节点集 V_p 中节点的嵌入特征和所对应类的语义信息之间的余弦相似度:

$$L_s = -\frac{1}{\|V_p\|} \sum_{v_i \in V_p} \log \text{cs} \left(z_i, C_{\arg \max_k y'_{ik}} \right) \quad (21)$$

其中 $\text{cs}(\cdot)$ 为余弦相似度函数。最终, 半监督分类网络的总损失可定义为:

$$L_{semi} = \lambda_3 L_{ce} + \lambda_4 L_s \quad (22)$$

4. 实验结果与分析

为了评估网络的性能, 本文使用了三个广泛用于属性图分析的数据集(Cora [19]、Citeseer [20]和 Pubmed [21])。

在具体实验中, 首先以 0.01 的学习率预训练属性图聚类网络 200 轮, 之后使用 0.01 学习率训练半监督分类网络 400 轮, 训练期间使用的优化器为 Adam。属性图聚类网络的图卷积层和图注意力层的输出维度都设置为[32, 16], 图注意力层的多头注意力头数 B 设置为 8, 比值系数 θ_1 和 θ_2 分别设置 0.4 和 0.02, 线性加权系数 β 设置为 0.5, 置信阈值 α_2 根据所提到的数据集顺序分别设置为[0.3, 0.4, 0.4], 超参 λ_1 和 λ_2 分别设置为 1 和 0.001。半监督分类网络的卷积层输出维度为[512, 7], 置信阈值 α_1 根据具体数据集分别设置为[0.8, 0.7, 0.5], 超参 λ_3 和 λ_4 分别设置为 1 和 2。

4.1. 定量结果及分析

在实验中本文首先使用三个标准指标来评估模型在聚类任务上的性能: 聚类精度(ACC)、归一化互信息(NMI)和平均随机指数(ARI), 这些指标的得分越高, 聚类效果越好。

Table 1. Clustering results of the attribute graph clustering network on three benchmark datasets

表 1. 属性图聚类网络在三个基准数据集上的聚类结果

	Cora			Citeseer			Pubmed		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means	0.500	0.317	0.239	0.544	0.312	0.285	0.580	0.278	0.246
DeepWalk	0.529	0.384	0.291	0.390	0.131	0.137	0.647	0.238	0.255
GAE	0.530	0.397	0.293	0.380	0.174	0.141	0.632	0.249	0.246
VGAE	0.592	0.408	0.347	0.392	0.163	0.101	0.619	0.216	0.201
DAEGC	0.704	0.528	0.496	0.672	0.397	0.410	0.671	0.266	0.278
GATE	0.658	0.527	0.496	0.616	0.401	0.381	0.673	0.322	0.299
GMM-VGAE	0.717	0.530	0.562	0.675	0.407	0.424	0.711	0.299	0.330
Ours	0.757	0.597	0.577	0.698	0.446	0.458	0.714	0.345	0.346

本文所提出的属性图聚类网络在三个基准数据集上与其他聚类方法的比较结果如表 1 所示, 其中粗体值表示聚类任务的最佳性能。从实验结果来看, 可以定量分析, 我们的方法在三个基准数据集上比其他方法获得了最好的结果。我们可以观察到, 与只使用内容信息(K-means)的方法和只使用结构信息(DeepWalk)的方法相比, 同时使用这两种信息的深度学习方法才能获得更好的结果, 如 GAE、VGAE、DAEGC 以及 GATE [22]。这表明, 图的内容信息和结构信息对图的聚类具有重要意义, 两者的结合可以进一步提高聚类性能。

更进一步地, 本文通过 ACC 指标评估半监督分类网络的性能。在控制标签率相同的情况下, 我们与标签传播(LP)、图卷积神经网络(GCN)、Li 等人提出的一些协作训练方案(Co-training、Self-training、Union 以及 Intersection)以及协作图卷积网络(CGCN)进行了大量的对比实验, 详细的对比实验结果参考表 2~4。实验证明所以出的方法在极低的标签率的情况下具有很强的竞争力, 例如从图 3 可以看出, 所提出的方法在 Cora 数据集上使用最低的标签率已经远超大部分方法, 甚至逼近大部分方法使用高标签率的性能。

Table 2. Classification accuracy on the Cora
表 2. 在 Cora 上的分类准确率

标签率	0.5%	1%	2%	3%	4%	5%
LP	0.564	0.623	0.654	0.675	0.690	0.702
GCN	0.509	0.623	0.722	0.765	0.784	0.797
Co-training	0.566	0.664	0.735	0.759	0.789	0.808
Self-training	0.537	0.661	0.738	0.772	0.794	0.800
Union	0.585	0.699	0.759	0.785	0.804	0.817
Intersection	0.497	0.650	0.729	0.771	0.794	0.802
标签率	0.5%	1%	2%	3%	4%	5%
Two-Stage	0.579	0.670	0.748	0.790	0.815	0.833
CGCN	0.643	0.724	0.768	0.801	0.827	0.842
Ours	0.732	0.763	0.806	0.823	0.832	0.843

Table 3. Classification accuracy on the Citeseer
表 3. 在 Citeseer 上的分类准确率

标签率	0.5%	1%	2%	3%	4%	5%
LP	0.348	0.402	0.436	0.453	0.464	0.473
GCN	0.436	0.553	0.649	0.675	0.687	0.696
Co-training	0.473	0.557	0.621	0.625	0.645	0.655
Self-training	0.433	0.581	0.682	0.698	0.704	0.710
Union	0.463	0.591	0.667	0.667	0.676	0.682
Intersection	0.429	0.591	0.686	0.701	0.708	0.712
Two-Stage	0.513	0.606	0.689	0.714	0.719	0.723
CGCN	0.593	0.631	0.695	0.726	0.728	0.746
Ours	0.681	0.716	0.718	0.722	0.740	0.745

Table 4. Classification accuracy on the Pubmed
表 4. 在 Pubmed 上的分类准确率

标签率	0.03%	0.05%	0.1%	0.3%
LP	0.614	0.664	0.654	0.668
GCN	0.605	0.575	0.659	0.778
Co-training	0.622	0.683	0.727	0.782
Self-training	0.519	0.587	0.668	0.770
Union	0.584	0.640	0.707	0.792
Intersection	0.520	0.593	0.694	0.776
Two-Stage	0.607	0.641	0.722	0.782
CGCN	0.647	0.692	0.778	0.803
Ours	0.708	0.718	0.733	0.810

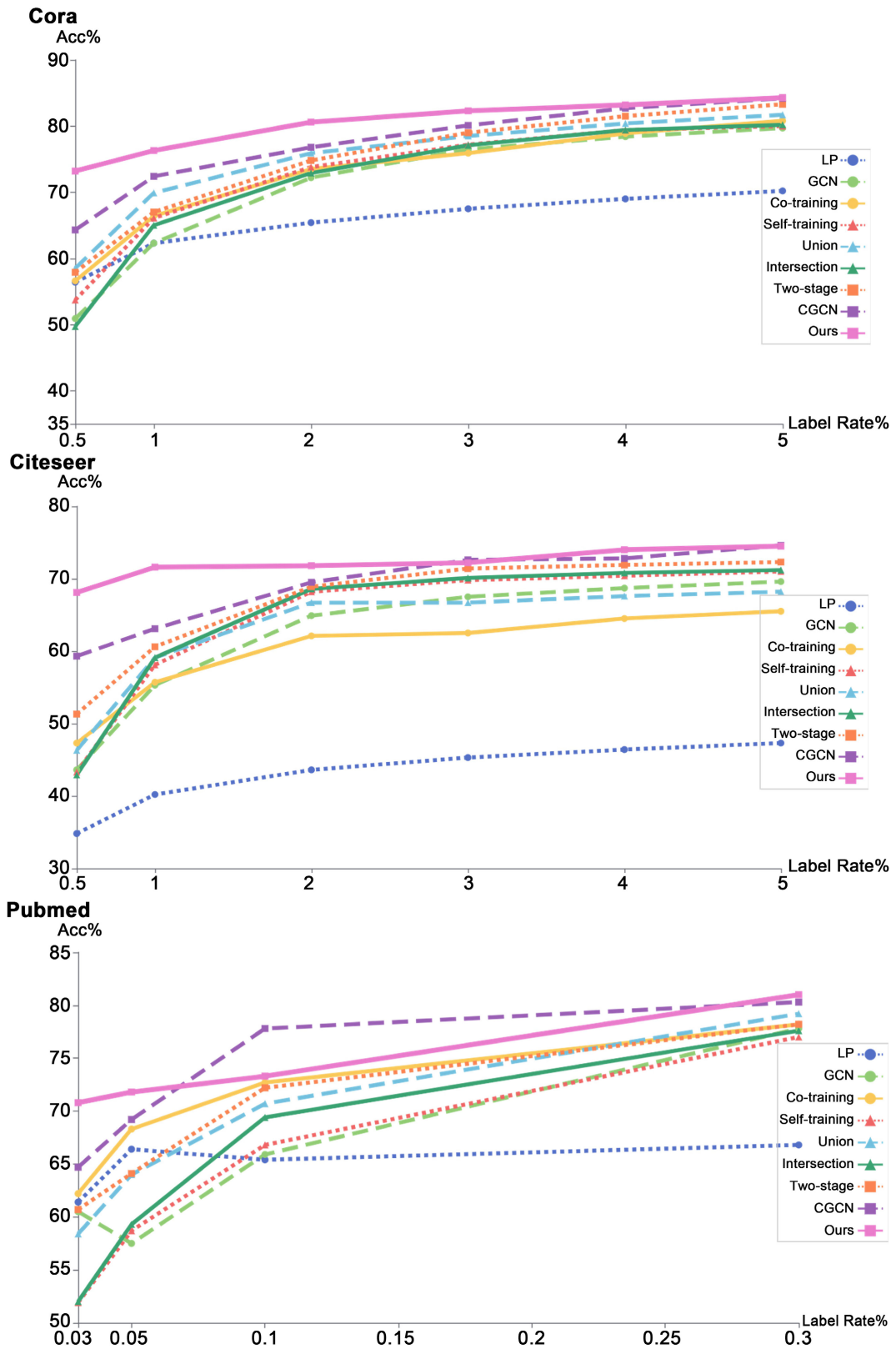


Figure 3. Classification performance of semi-supervised classification networks on benchmark datasets
 图 3. 半监督分类网络在基准数据集上的分类性能

4.2. 定性结果及分析

图 4 展示了 Cora 数据集的原始嵌入特征以及使用不同标签率训练半监督网络所获得的嵌入特征。对比图 4(a) Cora 的原始特征分布和图 4(b)中使用 0.5%的标签学习的特征,可以看出模型能挖掘到复杂图结构数据的潜在信息,即使在标签稀少的情况下依然能预测出大致分类结果。随着标签率的提升,模型性能也在不断上升,分类结果更加精确,每个类别的错误预测也在减少。

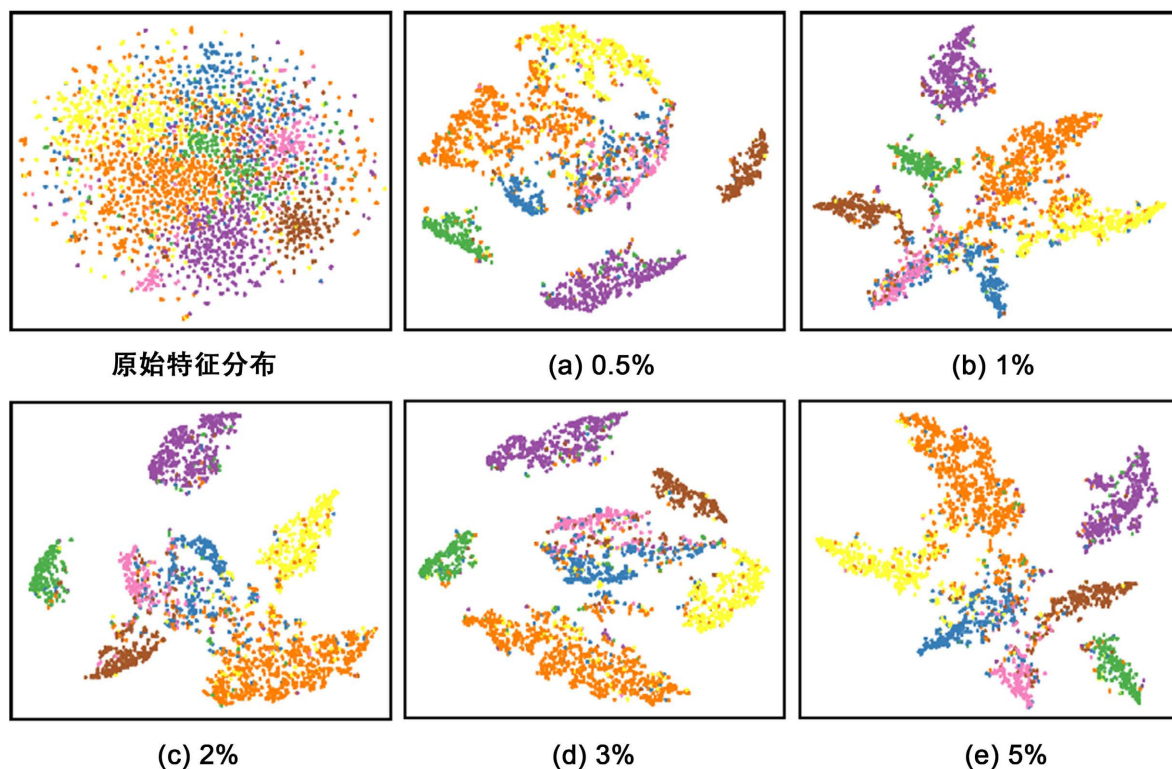


Figure 4. Visualization results of Cora features
图 4. Cora 特征的可视化结果

5. 结论与展望

本文所提出的方法有效地利用了属性图聚类,提高了半监督分类的性能。为了对图嵌入空间中的复杂数据分布进行建模,我们提出了一种面向聚类的端到端属性图聚类网络,通过特征融合机制整合图的结构信息和节点属性,高效学习图的嵌入特征。同时,为了解决半监督分类任务容易受标签数量限制的问题,利用属性图聚类网络生成的高置信度伪标签集扩充现有的标签集,为模型学习图的特征提供了良好的指导。在属性图聚类和半监督节点分类任务上与其他方法的大量对比实验验证了该模型的优越性。

基金项目

基于安徽省自然科学基金资助(2008085MF190)。

参考文献

- [1] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.
- [2] Perozzi, B., Al-Rfou, R. and Skiena, S. (2014) DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August

- 2014, 701-710. <https://doi.org/10.1145/2623330.2623732>
- [3] Grover, A. and Leskovec, J. (2016) node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 855-864. <https://doi.org/10.1145/2939672.2939754>
- [4] Wang, D., Cui, P. and Zhu, W. (2016) Structural Deep Network Embedding. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1225-1234. <https://doi.org/10.1145/2939672.2939753>
- [5] MacQueen, J. (1967) Classification and Analysis of Multivariate Observations. In: *5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Los Angeles, 281-297.
- [6] Kipf, T.N. and Welling, M. (2016) Variational Graph Auto-Encoders.
- [7] Pan, S., Hu, R., Fung, S., Long, G., Jiang, J. and Zhang, C. (2020) Learning Graph Embedding with Adversarial Training Methods. *IEEE Transactions on Cybernetics*, **50**, 2475-2487. <https://doi.org/10.1109/tyb.2019.2932096>
- [8] Wang, C., Pan, S., Hu, R., Long, G., Jiang, J. and Zhang, C. (2019) Attributed Graph Clustering: A Deep Attentional Embedding Approach. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 10-16 August 2019, 3670-3676. <https://doi.org/10.24963/ijcai.2019/509>
- [9] Mrabah, N., Bouguessa, M., Touati, M.F. and Ksantini, R. (2023) Rethinking Graph Auto-Encoder Models for Attributed Graph Clustering (Extended Abstract). 2023 *IEEE 39th International Conference on Data Engineering (ICDE)*, Anaheim, 3-7 April 2023, 3891-3892. <https://doi.org/10.1109/icde55515.2023.00372>
- [10] Zhao, T., Liu, Y., Neves, L., Woodford, O., Jiang, M. and Shah, N. (2021) Data Augmentation for Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 11015-11023. <https://doi.org/10.1609/aaai.v35i12.17315>
- [11] Caron, M., Bojanowski, P., Joulin, A. and Douze, M. (2018) Deep Clustering for Unsupervised Learning of Visual Features. In: *Computer Vision—ECCV 2018*, Springer International Publishing, Berlin, 139-156. https://doi.org/10.1007/978-3-030-01264-9_9
- [12] Xie, Q., Dai, Z., Hovy, E., et al. (2020) Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 6-12 December 2020, 6256-6268.
- [13] Li, Q., Han, Z. and Wu, X. (2018) Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 3538-3545. <https://doi.org/10.1609/aaai.v32i1.11604>
- [14] Tong, H., Faloutsos, C. and Pan, J. (2006) Fast Random Walk with Restart and Its Applications. *6th International Conference on Data Mining (ICDM'06)*, Hong Kong, 18-22 December 2006, 613-622. <https://doi.org/10.1109/icdm.2006.70>
- [15] Sun, K., Lin, Z. and Zhu, Z. (2020) Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 5892-5899. <https://doi.org/10.1609/aaai.v34i04.6048>
- [16] Hui, B., Zhu, P. and Hu, Q. (2020) Collaborative Graph Convolutional Networks: Unsupervised Learning Meets Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 4215-4222. <https://doi.org/10.1609/aaai.v34i04.5843>
- [17] Veličković, P., Cucurull, G., Casanova, A., et al. (2017) Graph Attention Networks.
- [18] Tao, Z., Liu, H., Li, J., Wang, Z. and Fu, Y. (2019) Adversarial Graph Embedding for Ensemble Clustering. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 10-16 August 2019, 3562-3568. <https://doi.org/10.24963/ijcai.2019/494>
- [19] McCallum, A.K., Nigam, K., Rennie, J. and Seymore, K. (2000) Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval*, **3**, 127-163. <https://doi.org/10.1023/a:1009953814988>
- [20] Giles, C.L., Bollacker, K.D. and Lawrence, S. (1998) Citeseer: An Automatic Citation Indexing System. *Proceedings of the 3rd ACM Conference on Digital Libraries DL '98*, Pittsburgh, 23-26 June 1998, 89-98. <https://doi.org/10.1145/276675.276685>
- [21] Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B. and Eliassi-Rad, T. (2008) Collective Classification in Network Data. *AI Magazine*, **29**, 93-106. <https://doi.org/10.1609/aimag.v29i3.2157>
- [22] Salehi, A. and Davulcu, H. (2020) Graph Attention Auto-Encoders. 2020 *IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, Baltimore, 9-11 November 2020, 989-996. <https://doi.org/10.1109/ictai50040.2020.00154>