

# 基于自编码器及对比损失的图聚类方法

王静红<sup>1,2,3</sup>, 王 慧<sup>1</sup>, 袁 焯<sup>4</sup>

(1. 河北师范大学计算机与网络空间安全学院, 河北 石家庄 050024)

(2. 河北省网络与信息安全重点实验室, 河北 石家庄 050024)

(3. 供应链大数据分析与安全河北省工程研究中心, 河北 石家庄 050024)

(4. 河北工程技术学院, 河北 石家庄 050020)

[摘要] 图聚类根据图数据的内在关系找到组或社区, 是数据分析中一项重要的研究问题. 近年来, 基于自编码器的方法能够获得有效节点属性表示, 但未融合结构信息. 由于图神经网络的广泛应用, 基于半监督图卷积网络和图自编码器的模型能够融合结构信息, 与传统聚类方法相比取得了较好的效果, 但标记数据和卷积操作代价昂贵. 因此, 本文提出了一种基于自编码器及对比损失的图聚类方法. 首先该方法使用简单的多层感知器设计自编码器, 预训练自编码器学习节点属性表示. 其次结合影响对比损失学习图嵌入表示, 融合丰富的图结构信息, 然后同时迭代优化嵌入表示和自监督聚类任务. 最后, 使用多个引文网络数据集与基准模型进行对比实验. 实验表明, 聚类性能得到有效提升, 并且参数敏感性分析和变体实验验证了影响对比损失和自监督聚类的有效性.

[关键词] 图聚类, 自编码器, 影响对比损失, 图嵌入, 自监督聚类

[中图分类号] TP391 [文献标志码] A [文章编号] 1001-4616(2025)01-0075-10

## Graph Clustering Based on Auto-Encoder and Contrastive Loss

Wang Jinghong<sup>1,2,3</sup>, Wang Hui<sup>1</sup>, Yuan Chuo<sup>4</sup>

(1. College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang 050024, China)

(2. Hebei Provincial Key Laboratory of Network and Information Security, Shijiazhuang 050024, China)

(3. Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics and Data Security, Shijiazhuang 050024, China)

(4. Hebei Institute of Engineering Technology, Shijiazhuang 050020, China)

**Abstract:** Graph clustering, which can find communities or groups based on the intrinsic relationships of graph data, is an important research problem in data analysis. In recent years, Auto-Encoder based methods have effectively extracted node attribute representations, but do not include structural information. Due to the widespread application of graph neural networks, the fusion of structural information based on semi-supervised graph convolutional networks and graph Auto-Encoder models has achieved better results compared to traditional clustering methods. However, labeling data and using convolution operations are expensive. This paper proposed a graph clustering method based on Auto-Encoder and Contrastive Loss (GC-AECL). Firstly, the model used a simple multilayer perceptron to design the Autoencoders, and pretrained the Auto-Encoder learning node attribute representation. Then the model combined the influence contrastive loss to enrich structural information to learn the graph embedded representation. And then the model iteratively optimized the embedded representation and self-supervised clustering tasks at the same time. Finally, the experiments were compared with the benchmark models on multiple citation network datasets. The experimental results showed that the clustering performance had been improved, and parameter sensitivity analysis and variants study had been conducted to verify the effectiveness of impact contrastive loss and self-supervised clustering.

**Key words:** graph clustering, Auto-Encoder, influence contrastive loss, graph embedding, self-supervised clustering

图结构化数据的广泛性和强大的表示学习能力, 使得图数据的研究越来越受到重视. 图聚类是数据分析中一个重要且具有挑战性的任务, 它的目的是将图数据中相似的样本划分为同一组(社区), 将不同

收稿日期: 2023-06-12.

基金项目: 河北省自然科学基金项目(F2021205014, F2019205303)、河北省高等学校科学技术研究项目(ZD2022139)、中央引导地方科技发展资金项目(226Z1808G)、河北师范大学科技类研究基金项目(L2023J05)、河北师范大学研究生创新资助项目(XCXZZSS202315).

通讯作者: 王静红, 博士, 教授, 研究方向: 人工智能、数据挖掘. E-mail: wangjinghong@126.com

的样本划分为不同的组(社区). 图聚类任务可以应用于社交网络<sup>[1]</sup>、引文网络<sup>[2]</sup>、生物学作用网络<sup>[3-4]</sup>、推荐系统<sup>[5]</sup>等. 聚类结果表示着复杂网络节点之间的关系,因此图聚类方法对研究和理解图结构数据具有重要的现实意义.

自编码器(auto-encoder, AE)<sup>[6]</sup>作为神经网络无监督学习的经典变体,常被用于聚类方法. 自编码器使用多层神经网络,即编码器学习非线性特征,解码器基于学习特征重构原始特征. 基于深度自编码器的方法<sup>[7-10]</sup>能够有效提取节点属性表示,这些方法已经取得显著的改进和先进性能,但是图聚类问题仍具有重大挑战. 首先,传统方法只关注节点属性信息或图结构信息,只使用了一种类型的信息,会导致聚类结果较差. 其次它们大多未包含结构信息,且将聚类任务与自编码器训练分离. 图 1 是一个具有不同聚类结果的图聚类示例:在(a)中,4 个节点用无向边连接,并与二维属性向量相关联;在(b)中,只使用节点属性进行聚类;在(c)中只使用图结构;在(d)中同时使用节点属性和图结构. 节点被划分为两个集群(红圈和蓝圈),不同的节点颜色表示节点的真实类别.

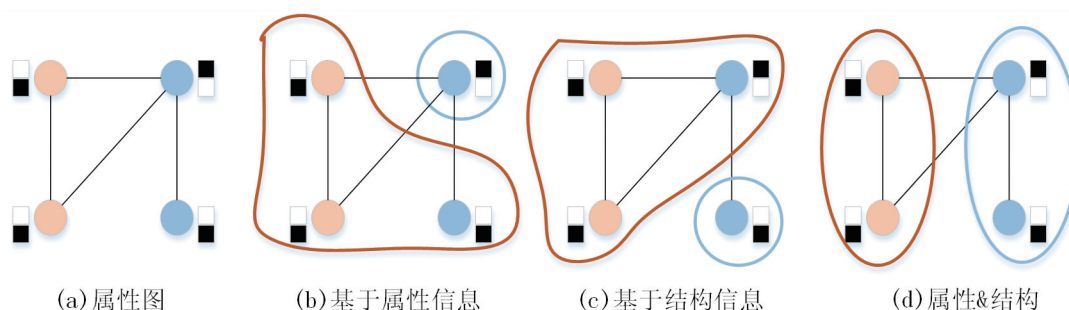


图 1 图聚类图示

Fig. 1 An illustration of graph clustering

近年来,由于图神经网络强大的图挖掘能力,引起了研究者的极大关注. 研究者们提出了许多基于图神经网络的图聚类方法,并且图神经网络已经成功应用于节点分类、图分类、时间序列分析、知识图谱、图聚类等. 由于图结构信息可以有效指导嵌入表示学习,许多方法利用图卷积网络融合图结构信息,如图卷积神经网络(graph convolutional network, GCN)<sup>[11]</sup>在许多图聚类应用中都取得了很好的性能,如社交网络和推荐系统. 已有研究<sup>[7,8,12,13]</sup>通过结合图神经网络 GNN<sup>[14]</sup>和自编码器融合更多有效图结构信息,如图自编码器(graph auto-encoder, GAE)和变分图自编码器(variational graph auto-encoder, VGAE)<sup>[12]</sup>使用自编码器重构邻接矩阵,边缘化图自编码器(marginalized graph auto-encoder, MGAE)<sup>[15]</sup>使用自编码器重构节点特征. 但基于图神经网络的图聚类方法对图结构信息进行有效融合仍存在标记数据和计算代价昂贵等问题.

图结构信息表示节点之间的依赖关系,对发现未知的潜在结构表示是非常重要的. 自监督学习方法在计算机视觉等领域的出色表现吸引了大量关注,对比学习是一种无需人工标注就能隐式获取监督信息的深度图聚类技术. 已有的一些工作<sup>[16-18]</sup>使用对比损失指导图嵌入表示,通过对比正样本和负样本来学习判别表征,如 Graph-MLP<sup>[17]</sup>使用对比损失融入简单的多层感知器(multi-layer perception, MLP)模型中进行引文网络数据的节点分类. 但这些工作大多使用图像增强、卷积、注意力等操作或以监督学习的方式表示网络数据.

本文提出了一种基于自编码器及对比损失的图聚类方法. 首先通过设计简单的线性多层感知器组成自编码器,预训练自编码器学习节点属性表示,不再使用卷积或 Softmax 操作. 其次结合影响对比损失学习图嵌入表示,融合更加丰富的结构信息,利用对比损失强化集群结构. 然后由于图嵌入表示和节点聚类是兼容的,同时优化图嵌入表示任务和节点聚类任务,使用基于概率分布的自监督训练机制迭代优化联合损失函数,得到有效的聚类结果. 最后,使用 4 种引文网络数据集与 19 种基准模型进行对比实验,实验表明,聚类性能得到有效提升,并且参数敏感性分析验证了影响对比损失和自监督聚类的有效性.

本文第 1 节介绍了图聚类的相关方法和研究现状,第 2 节介绍构建的基于自编码器及对比损失的图聚类模型,第 3 节通过对比实验验证了本文模型的有效性,第 4 节总结全文.

## 1 相关工作

### 1.1 图聚类方法

图聚类方法通常包括嵌入学习和聚类两个步骤,嵌入学习大多是生成式方法,聚类是判别式方法.生成式方法通过在输出空间中设计损失函数来学习图数据嵌入表示.其中大多数方法采用自编码器捕获潜在表示,但早期的图表示学习工作完全依赖于节点属性信息. Hinton 等<sup>[6]</sup>通过设计的自编码器网络(AE)驱动表示学习. Xie 等<sup>[9]</sup>提出了深度嵌入聚类方法(deep embedding clustering, DEC),通过聚类联合优化的特征空间中的一组数据点来学习特征表示. Guo 等<sup>[10]</sup>引入重构损失来改进 DEC,学习更好的表示法.虽然这些工作已经取得了显著的改进,但它们仅仅关注节点属性信息,而忽略了图结构信息.

近年来,很多研究提出构建高效的基于图神经网络的方法对图结构信息和属性信息融合方法. GAE<sup>[8]</sup>和 VGAE<sup>[8]</sup>使用自编码器重构邻接矩阵, MGAE<sup>[15]</sup>使用自编码器重构节点属性. Pan 等<sup>[19]</sup>在 GAE 框架的基础上,提出了对抗正则化图自编码器网络(adversarially regularized graph auto-encoder, ARGAE). Bo 等<sup>[7]</sup>通过堆叠多层 GNN,基于 DEC 框架设计结构深度聚类网络(structural deep clustering network, SDCN),传输 AE 学习表示并捕获高阶结构信息.但这些深度模型中的训练参数数量较大会导致过拟合.

### 1.2 对比学习方法

对比学习是一种自监督学习方法,在计算机视觉<sup>[20]</sup>和图形学习<sup>[21-22]</sup>中较早应用.通过最大化正样本对的相似度和负样本对的距离来学习判别特征,受计算机视觉中对比学习应用成功的启发,对比深度图聚类方法越来越多地被提出.当前一些工作将对比损失用来指导节点嵌入表示. You 等提出 GraphCL 使用图神经网络对图数据进行增强预处理对比学习<sup>[16]</sup>. Kipf 等提出 C-SWMs<sup>[18]</sup>使用连续图像的对比损失融入 MLP 模型用于对象增量学习. Chen 等提出 SimCLR<sup>[23]</sup>通过数据增强构建自监督样本,并将小批量中来自不同视图的图像视为负样本对. Hassani 等提出 MVGRL<sup>[24]</sup>比较了从邻接矩阵转换来实现表示学习的两个扩散矩阵. Zhu 等提出了 GCA<sup>[25]</sup>自适应于图结构和属性的联合数据增强方案. Velikovovic 等提出深度图信息(deep graph infomax, DGI)<sup>[26]</sup>允许节点表示保存更多的全局信息. 蒋等提出 DGCP 动态图表示学习方法,将对比学习引入动态图中,利用对比损失引导嵌入空间捕获对预测未来图结构最有用的信息<sup>[27]</sup>.但是这些模型使用数据增强、注意力、卷积操作或以监督学习的方式进行学习,因此研究融合节点对比损失的无监督表示学习方法,减少计算代价,提高算法效率.

## 2 基于自编码器及对比损失的图聚类模型 GC-AECL

本节详细介绍了基于自编码器及对比损失的图聚类模型,该模型包括三部分:(1)自编码器预训练;(2)节点对比损失;(3)自监督聚类.本文提出了基于节点影响的对比损失,融合图结构信息用于嵌入表示,并且将图对比学习和聚类任务共同学习优化,整个模型框架如图 2 所示.

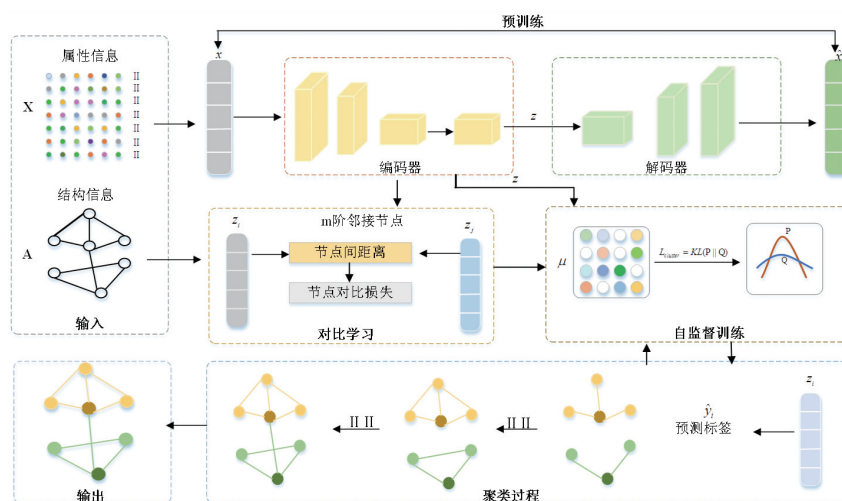


图 2 GC-AECL 模型框架图

Fig. 2 The framework of GC-AECL

文中使用  $G = \{V, E, X\}$  表示无向图,  $V$  是包含  $n$  个节点的有限集,  $E$  是无向图  $G$  中节点之间的邻接关系,  $X$  包含节点的属性特征, 即每个节点与一个  $d$  维特征向量相关联. 根据  $E$  中的邻接关系, 可将图  $G$  对应的邻接矩阵记为  $A$ , 并且可计算得到对应的度矩阵  $D$ .

## 2.1 自编码器预训练

为了提取节点属性特征并获得初始图的嵌入  $z$  和聚类中心的初始质心  $u$ , 使用自编码器 AE 进行预训练初始化. 首先使用编码器-解码器提取图节点属性信息从而获得潜在表示  $z$ , 并最小化原始数据和重构数据的重构损失.

编码器和解码器的生成嵌入过程为

$$\begin{aligned} \text{encoder: } H_{\text{encode}}^L &= \phi(W_{\text{encode}}^L H_{\text{encode}}^{L-1} + b_{\text{encode}}^L), \\ \text{decoder: } \hat{H}_{\text{decode}}^L &= \phi(W_{\text{decode}}^L \hat{H}_{\text{decode}}^{L-1} + b_{\text{decode}}^L). \end{aligned} \quad (1)$$

基于原始数据, 获得最终重构数据:

$$H_{\text{encode}}^0 = X, \hat{H}_{\text{decode}}^0 = H_{\text{encode}}^L, \hat{X} = \hat{H}_{\text{decode}}^L. \quad (2)$$

整个重构过程的损失为

$$\text{loss}_{\text{reconstruct}} = \|X - \hat{X}\|_F^2, \quad (3)$$

完成预训练获得图的初始嵌入  $Z$ , 然后使用  $K$ -means 方法获得聚类中心  $\mu$ .

$$Z = \hat{H}_{\text{decode}}^0 = H_{\text{encode}}^L. \quad (4)$$

## 2.2 节点对比损失

对比损失可以将特征空间中正样本(邻接节点对)的距离更近, 负样本(非邻接节点对)的距离更远. 使用对比损失度量图结构, 首先计算两个节点嵌入表示之间的相似度, 然后增强节点的边缘信息来确定正样本.

对于给定某一节点的不同深度的非邻接节点, 都会产生不同的影响程度. 给定深度  $m$ , 定义节点不同深度的非邻接节点产生的总影响为

$$M_{ij} = \text{Effect}_{M_{ij}} = \sum_{m=1}^m a_{ijm} \text{relationship}_m(i, j), \quad (5)$$

节点  $j$  为节点  $i$  在深度为  $m$  的非邻接节点,  $m$  通过超参数设置获得. 其中,  $a_{ijm}$  为它们的关系系数,  $M$  为节点之间的影响关系矩阵.

$$\text{relationship}_m(i, j) = \tilde{A}^m \quad (6)$$

$$M_{ij} \begin{cases} = 0, & \text{节点 } j \text{ 是节点 } i \text{ 的邻居节点} \\ \neq 0, & \text{节点 } j \text{ 不是节点 } i \text{ 的邻居节点} \end{cases} \quad (7)$$

理论上, 节点之间的潜在影响结构应与节点之间的影响关系矩阵保持一致. 传统的基于图神经网络的模型普遍设定固定深度, 并假设所有节点从不同深度获得相同的影响, 但图神经网络模型为防止过度平滑, 网络深度不超过 2 层.

定义节点  $i$  的影响增强对比损失为

$$\text{loss}_i = -\log \frac{\sum_{j=1}^B 1_{[j \neq i]} M_{ij} \exp(\text{dis}(z_i, z_j)/\tau)}{10^{-8} + \sum_{k=1}^B 1_{[k \neq i]} \exp(\text{dis}(z_i, z_k)/\tau)}, \quad (8)$$

其中  $B$  为非邻接节点的个数,  $\tau$  是一个温度参数,  $M_{ij}$  是节点  $i$  和节点  $j$  的影响关系矩阵,  $\text{dis}(z_i, z_j)$  是节点  $i$  和节点  $j$  的嵌入表示的相似度.

$$\text{dis}(z_i, z_j) = \frac{\sum_{i=1}^n (z_i, z_j)}{\sqrt{\sum_i^n (z_i)^2} \cdot \sqrt{\sum_j^n (z_j)^2}}, \quad (9)$$

真实世界中存在的复杂网络图大多是稀疏的, 因此邻接矩阵中大多数元素为 0 元素. 两个节点没有连边不意味着节点之间没有影响关系, 可能存在高阶近似的强关联性. 根据高阶近似关系受到启发, 促使



研究节点的高阶关系,将邻接矩阵扩展到  $m$  阶的节点关系矩阵,探索节点  $i$  到节点  $j$  的深度为  $m$  的路径.

根据邻接矩阵计算出度矩阵  $D$ ,

$$D_{ij} = \sum_j \tilde{A}_{ij}, \quad (10)$$

然后计算得到归一化的邻接矩阵  $\tilde{A}$ ,

$$\tilde{A} = D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}}, \quad (11)$$

其中  $I$  为单位矩阵.

计算深度  $m$  的非邻接节点的归一化邻接矩阵的累积影响,若节点  $i$  和节点  $j$  为邻接节点,则  $M_{ij}=0$ ,其余在  $m$  跳内的非邻接节点均会对节点  $i$  产生影响,则  $M_{ij} \neq 0$ .

则对比损失定义为

$$\text{loss}_{\text{contrastive}} = \frac{1}{B} \sum_{i=1}^B \text{loss}_i. \quad (12)$$

其中图的对比损失为节点  $i$  的  $m$  跳的非邻接节点的影响总和,  $B$  为非邻接节点的个数.

### 2.3 自监督聚类

在本质上,属性图聚类是无监督的学习任务,在优化过程中未使用标签信息,因此具有非常大的挑战性,本文使用自监督训练机制,生成的嵌入表示基于概率分布生成预测标签,作为聚类的指导依据.

首先计算节点  $i$  属于类别质心为  $u$  的概率  $q_{iu}$ ,基于生成的嵌入表示  $z_i$  和社区集群中心  $\mu_u$ ,使用  $t$  分布度量生成嵌入与质心之间的相似性,计算概率分布的过程为

$$q_{iu} = \frac{(1 + \|z_i - \mu_u\|^2/\eta)^{-\frac{\eta+1}{2}}}{\sum_{u'} (1 + \|z_i - \mu_{u'}\|^2/\eta)^{-\frac{\eta+1}{2}}}, \quad (13)$$

$\eta$  是  $t$  分布的自由度,保持  $\eta$  自由度为 1.

$Q=[q_{iu}]$  作为所有节点的社区集群类别的分布,节点相似度越高,距离社区集群中心越近,在同一类别的概率较高. 定义了一个目标分布  $P$  表示节点概率分布的相关性,定义为

$$p_{iu} = \frac{q_{iu} / \sum_i q_{iu}}{\sum_k (q_{ik} / \sum_i q_{ik})}, \quad (14)$$

其中  $\sum_i q_{ik}$  是节点  $i$  在质心  $u$  为类别  $k$  的频率,假设有  $k$  个类别.

为了使数据表示更接近社区集群中心,并提高社区的模块度,最小化  $Q$  分布和  $P$  分布的  $KL$  散度损失,使得真实分布  $Q$  更接近目标分布  $P$ . 通过使用  $Q$  真实分布接近于目标分布  $P$  自监督训练发现社区,然后由其监督  $Q$  分布,通过将  $KL$  散度损失最小化,则聚类损失为

$$\text{loss}_{\text{cluster}} = KL(P \parallel Q) = \sum_i \sum_u p_{iu} \log \frac{p_{iu}}{q_{iu}}, \quad (15)$$

使用最终得到优化后的  $Q$  分布获得节点所属社区标签.

$$\hat{Y}_i = \text{argmax}_u q_{iu}. \quad (16)$$

### 2.4 损失联合优化

算法流程如算法 1 所示. 模型使用简单线性多层感知器作为自编码器,对节点属性信息进行重构. 使用对比损失度量图结构相似度,基于自监督训练机制最小化生成嵌入分布与目标分布的  $KL$  散度损失,整个模型的训练损失为

$$\text{loss}_{\text{all}} = \text{loss}_{\text{reconstruct}} + \alpha \text{loss}_{\text{contrastive}}(m, \tau) + \beta \text{loss}_{\text{cluster}}, \quad (17)$$

其中  $\alpha$  是平衡对比学习结构信息,  $\beta$  是控制聚类优化.

算法 1 GC-AECL 图聚类训练过程  
Alg. 1 Training process of GC-AECL graph clustering

输入: 属性图  $G=(V, E, X)$ , 数据维度  $d$ , 自编码层数  $L$ , 社区类别个数  $k$ , 最大迭代次数  $\text{maxIter}$ ;

输出: 聚类中心  $\mu$ , 节点所属社区类别  $\hat{Y}$ ;

1. 初始化预训练自编码器的参数  $W_{\text{encode}}^L, b_{\text{encode}}^L, W_{\text{decode}}^L, b_{\text{decode}}^L$ ;

2. for epoch = 1 < maxIter do

3. 预训练自编码器模型依据公式(1)–(4);

4. 预训练得到节点嵌入表示  $Z$ , 使用 Kmeans 得到初始聚类中心  $\mu$ ;

5. end for

6. 使用节点对比损失学习图的结构信息, 依据公式(5)–(12);

7. 根据节点对比损失和节点的嵌入表示, 进行自监督聚类;

8. for epoch = 1 < maxIter do

9. 生成的概率分布  $Q$  依据式(13);

10. 计算目标分布  $P$  依据式(14);

11. 最小化聚类损失目标分布  $P$  和生成分布  $Q$  的  $KL$  散度, 依据式(15);

12. 分别计算重构损失、对比损失、聚类损失;

13. 更新整个模型的参数, 使得联合损失最小化, 依据式(17);

14. epoch = epoch + 1;

15. end for

16. 计算节点标签, 根据公式(16);

17. 返回最终聚类中心  $\mu$  和节点所属社区类别  $\hat{Y}$ .

### 3 实验及分析

#### 3.1 实验设置

##### (1) 数据集

本文在 4 种引文网络数据集上进行实验, 包括 Cora、Citeseer、Pubmed、Wiki. 表 1 给出了数据集的详细信息.

##### (2) 评价指标

本文使用准确度(ACC)和标准互信息(NMI)两种评价指标对本文模型进行评估. ACC 是正确预测的样本数与全部样本数的比值, NMI 是算法得到的聚类结果与真实类别的比值. NMI 的取值范围为(0, 1). 对于两个指标, 较大的值聚类结果更好.

##### (3) 基准模型

实验比较了七类方法, 即仅使用原始特征<sup>[28–29]</sup>, 仅使用特征的深度聚类<sup>[6]</sup>, 仅使用结构信息<sup>[29–30]</sup>, 仅使用结构信息的深度聚类<sup>[31–32]</sup>, 使用特征和 GCN 学习结构的深度聚类<sup>[7, 12, 15, 33–35]</sup>, 使用对抗训练机制和图对比学习<sup>[19, 25]</sup>, 使用特征和注意力学习结构的深度聚类<sup>[13, 36–37]</sup>. 本文方法利用对比损失学习结构信息和属性信息, 形成一个独立类型.

(4) 实验使用 Pytorch 1.10.0 框架运行在为 CPU-Intel core i5 – 12500H, GPU NVIDIA GeForce RTX 2080Ti, RAM 64GB RAM 硬件环境中, 使用 Adam 优化器最小化式(17)中的联合损失, 对学习的嵌入执行 K-means. 所有方法都进行了 10 次实验. 采用原始设置在基线方法源代码上重现结果. 表 2 为超参数的详细信息.

#### 3.2 实验结果与分析

本节通过聚类实验来评估 GC-AECL 在数据集 Cora、Citeseer、Pubmed、Wiki 上的性能, 实验结果详见表 3, 其中最佳值用粗体表示, 次优值用下划线表示.

表 1 数据集信息表

Table 1 Datasets information

数据集	节点数	边数	特征维度	社区个数	社区结构
Cora	2 708	5 429	1 433	7	是
Citeseer	3 327	4 732	3 703	6	是
Pubmed	19 717	44 338	500	3	是
Wiki	2 405	17 981	4 973	17	是

表 2 参数设置表

Table 2 Parameter settings

超参数	参数含义	设置
$L$	自编码器每层维度	500–500–2000–10
epoch	预训练迭代轮数	30
$lr$	学习率	$10^{-4}$
batch	批处理大小	256
$\alpha$	平衡对比损失参数	1
$\beta$	平衡自监督聚类参数	0.1
maxIter	最大迭代轮数	200
$m$	$m$ 阶非邻接	1, 2, 3, 4
$\tau$	温度参数	0.25

表 3 实验结果数据

Table 3 Experimental results

Methods	Cora		Citeseer		Pubmed		Wiki	
	ACC%	NMI%	ACC%	NMI%	ACC%	NMI%	ACC%	NMI%
K-means	34.65	16.73	38.49	17.02	57.32	29.12	33.37	30.20
Spectral-f	36.26	15.09	46.23	21.19	59.91	32.55	41.28	43.99
AE	55.01	31.16	57.08	27.64	56.12	30.42	—	—
Spectral-g	34.19	19.49	25.91	11.84	39.74	3.46	23.58	19.28
DeepWalk	46.74	31.75	36.15	9.66	61.86	16.71	38.46	32.28
Graph-Encoder	30.10	5.94	29.32	5.72	53.10	21.03	—	—
DNGR	49.24	37.29	32.59	18.02	45.35	15.38	37.58	35.85
GAE	53.25	40.69	41.26	18.34	64.08	22.97	17.33	11.93
VGAE	55.95	38.45	44.38	22.71	65.48	25.09	28.67	30.28
MGAE	63.43	45.57	63.56	39.75	43.88	8.16	50.14	47.97
AGC	68.92	53.68	67.00	41.13	69.78	31.59	47.65	45.28
SDCN	60.24	50.04	65.96	38.71	65.78	29.47	—	—
GUCD	50.50	32.30	58.47	27.43	63.13	26.98	—	—
SENet	71.92	55.03	67.52	41.72	67.59	33.60	—	—
ARGA	64.00	44.90	57.30	35.00	68.10	27.60	—	—
ARVGA	63.80	45.00	54.40	26.10	51.30	11.70	—	—
DGI	64.49	52.56	65.77	41.47	64.72	28.20	—	—
DAEGC	70.40	52.80	67.20	39.70	67.10	26.60	38.25	37.63
GATE	65.80	52.70	61.60	40.12	67.30	32.25	—	—
DNENC-Att	70.40	52.80	67.20	39.70	67.10	26.60	—	—
DNENC-Con	68.30	51.20	69.20	42.60	67.70	27.50	—	—
GC-AECL(ours)	<b>73.32</b>	<b>55.75</b>	<b>73.25</b>	<b>46.86</b>	<b>72.64</b>	<b>34.21</b>	<b>50.45</b>	<b>49.12</b>

GC-AECL 在 4 个数据集上优于基线方法. 具体而言,在 Citeseer 数据集上的 ACC 和 NMI 评价指标分别提高了 5.73%和 5.14%,验证了 GC-AECL 算法的有效性. 从表 3 结果可以得出以下结论:

(1)VGAE 和 GAE 优于 K-means、spectral-f、spectral-g、Graph Encoder 和 DeepWalk 方法,说明拓扑结构信息和属性信息都包含丰富的信息;

(2)GC-AECL 优于 SENet,说明 GC-AECL 重构网络属性信息的有效性. 与 SENet 相比,该方法充分利用了拓扑结构和属性信息,学习到高质量的节点特征表示;

(3)GC-AECL 优于 DAEGC,说明保持网络拓扑结构和重构属性信息的必要性. DAEGC 优于 GC-AECL-con(消融研究显示),GC-AECL-con 不包括网络拓扑结构. 这表明通过使用节点对比损失度优化了获得的节点表示,这有助于获得拓扑紧密的聚类结果;

(4)与先进的无监督学习模型 GUCD 相比,GC-AECL 获得了性能的提升,在 Cora 数据集上 ACC 提升了 22.82%,NMI 提升了 23.45%;

(5)与先进的图对比模型 DGI 相比,GC-AECL 获得了性能的提升,在 Pubmed 数据集上 ACC 提升了 7.92%,NMI 提升了 6.01%.

### 3.3 参数影响分析

图 3 为超参数  $\alpha$  和  $\beta$  对 ACC%和 NMI%的影响. 为了研究超参数  $\alpha$  和  $\beta$  的敏感性,选择在 Cora 数据

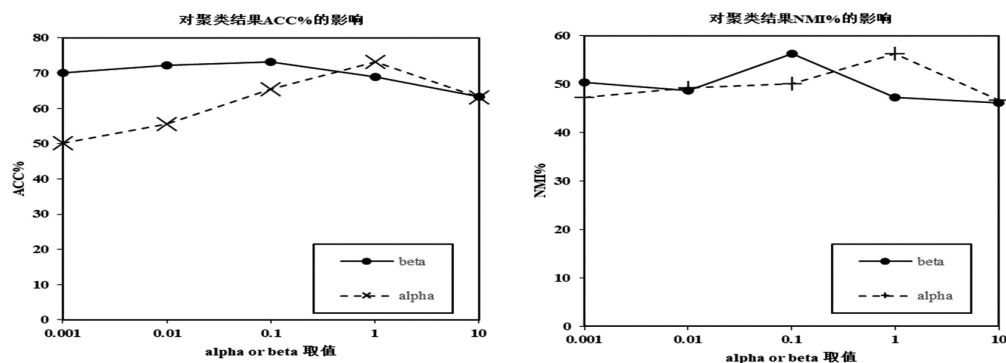


图 3 超参数取值变化对实验结果的影响(Cora 数据集)

Fig. 3 The influence of alpha and beta hyper-parameters on Cora

上进行实验,分别控制结构信息对比损失和自监督聚类的权重,参数的变化范围是 $\{0.001, 0.01, 0.1, 1, 10\}$ ,在 ACC%和 NMI%两个指标上评估聚类效果. 实验结果表明, $\alpha=1, \beta=0.1$  时,聚类效果最佳.

将本文模型与两种变体进行比较,以验证各模块在模型中的有效性. 定义了以下两种变体:

(1)GC-AECL-con,这个变体是没有对比学习模块,图嵌入学习通过自编码预训练和自监督聚类进行优化,用来验证提出的对比损失的有效性;

(2)GC-AECL-self,该变体是没有自监督聚类模块,图嵌入学习通过自编码预训练节点属性信息和节点对比损失重构结构信息进行迭代优化,用来验证提出的自监督聚类模块的有效性.

变体实验结果如图 4 所示,可以观察到两个模块对最终模型性能提升的贡献. 结果证明了对比损失对重构结构信息的有效性. 在 Cora 数据集上的实验结果,与 GC-AECL-self 相比,GC-AECL 准确度 ACC 提高了 4.3%,归一化互信息 NMI 提高了 1.65%;在 Citeseer 数据集上的实验结果,与 GC-AECL-self 相比,GC-AECL 准确度 ACC 提高了 3.05%,归一化互信息 NMI 提高了 3.8%;在 Wiki 数据集上的实验结果,与 GC-AECL-self 相比,GC-AECL 准确度 ACC 提高了 4.54%,归一化互信息 NMI 提高了 4.06%.

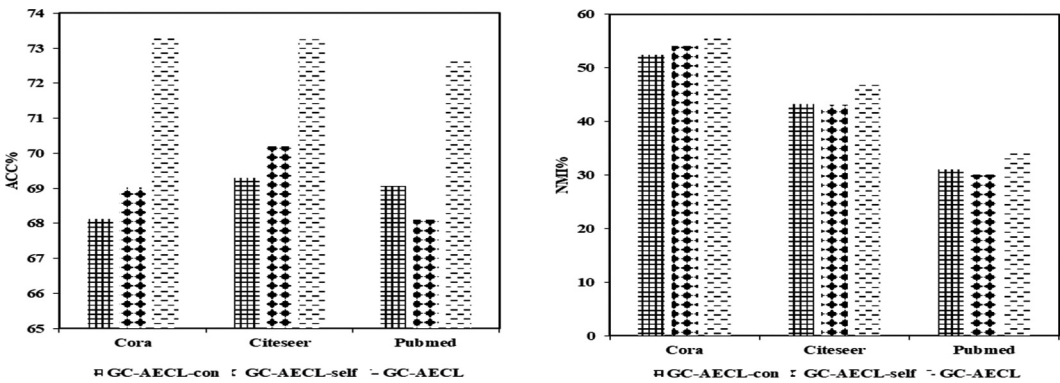


图 4 变体实验结果

Fig. 4 The results of variant study

变体结果表明,采用图嵌入学习和聚类优化相结合的策略可以获得高质量的表示. 在 Cora 数据集上的实验结果,与 GC-AECL-con 相比,GC-AECL 准确度 ACC 提高了 5.2%,归一化互信息 NMI 提高了 3.3%;在 Citeseer 数据集上的实验结果,与 GC-AECL-con 相比,GC-AECL 准确度 ACC 提高了 3.95%,归一化互信息 NMI 提高了 3.62%;在 Wiki 数据集上的实验结果,与 GC-AECL-con 相比,GC-AECL 准确度 ACC 提高了 3.58%,归一化互信息 NMI 提高了 3.16%.

4 结论

本文提出了一种基于自编码器及对比损失的图聚类方法 GC-AECL. 通过设计简单线性的多层感知器组成自编码器,结合影响对比损失融合图结构信息,对聚类损失和重构损失联合优化、迭代优化同时获得图嵌入和节点聚类结果. 与传统的基于图神经网络的方法相比具有显著优势,操作简单且效果有效. 通过多个引文网络数据集进行聚类实验,不仅学习到更加丰富的图结构信息和属性信息,还进一步提升图聚类的性能和计算效率,变体实验验证了影响对比损失和自监督聚类的有效性. 在未来的工作中,将进一步探索设计有效的负样本采样策略用于对比学习,更好解决现实场景中不同数据集样本偏差、节点噪声、网络异质性问题.

[参考文献]

[1] 李邵莹,孟丹,孔超,等. 面向社交推荐的自适应高阶隐式关系建模[J]. 软件学报,2023,34(10):4851-4869.  
[2] HUANG L, CHEN X, ZHANG Y, et al. Identification of topic evolution: network analytics with piecewise linear representation and word embedding[J]. Scientometrics, 2022, 127(9): 5353-5383.  
[3] HROVATIN K, FISCHER D S, THEIS F J. Toward modeling metabolic state from single-cell transcriptomics[J]. Molecular



- metabolism,2022,57:101396.
- [4] YUAN Q M, CHEN J W, ZHAO H Y, et al. Structure-aware protein-protein interaction site prediction using deep graph convolutional network[J]. Bioinformatics,2022,38(1):125–132.
  - [5] 刘会,张璇,杨兵,云炜,等. 用于社交推荐的增强影响扩散模型[J]. 计算机学报,2023,46(3):626–642.
  - [6] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504–507.
  - [7] BO D Y, WANG X, SHI C, et al. Structural deep clustering network[C]//Proceedings of the Web Conference 2020. New York: Association for Computing Machinery,2020:1400–1410.
  - [8] PENG Z, LIU H, JIA Y, et al. Attention-driven graph clustering network[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery,2021:935–943.
  - [9] XIE J Y, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//Proceedings of International Conference on Machine Learning. New York: Association for Computing Machinery,2016:478–487.
  - [10] GUO X F, GAO L, LIU X W, et al. Improved deep embedded clustering with local structure preservation[C]//Proceedings of IJCAI. New York: Association for Computing Machinery,2017:1753–1759.
  - [11] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[J/OL]. arXiv preprint arXiv:1609.02907,2016.
  - [12] KIPF T N, WELING M. Variational graph auto-encoders[J/OL]. arXiv preprint arXiv:1611.07308,2016.
  - [13] WANG C, PAN S R, HU R Q, et al. Attributed graph clustering: A deep attentional embedding approach[J/OL]. arXiv preprint arXiv:1906.06532,2019.
  - [14] 林晶晶,冶忠林,赵海兴,等. 超图神经网络综述[J]. 计算机研究与发展,2024,61(2):362–384.
  - [15] WANG C, PAN S R, LONG G D, et al. Mgae: Marginalized graph autoencoder for graph clustering[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: Association for Computing Machinery,2017:889–898.
  - [16] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations[J]. Advances in neural information processing systems,2020,33:5812–5823.
  - [17] HU Y, YOU H X, WANG Z C, et al. Graph-mlp: Node classification without message passing in graph[J/OL]. arXiv preprint arXiv:2106.04051,2021.
  - [18] KIPF T, VAN DER POL E, WELING M. Contrastive learning of structured world models[J/OL]. arXiv preprint arXiv:1911.12247,2019.
  - [19] PAN S R, HU R Q, LONG G D, et al. Adversarially regularized graph autoencoder for graph embedding[J/OL]. arXiv preprint arXiv:1802.04407,2018.
  - [20] YANG X H, HU X C, ZHOU S H, et al. Interpolation-based contrastive learning for few-label semi-supervised learning[J]. IEEE transactions on neural networks and learning systems,2022,35(2):2054–2065.
  - [21] XIA J, WU L, CHEN J, et al. SimGRACE: A simple framework for graph contrastive learning without data augmentation[J/OL]. arXiv preprint arXiv:2202.03104,2022.
  - [22] WANG Y, CAI Y, LIANG Y, et al. Adaptive data augmentation on temporal graphs[J]. Advances in neural information processing systems,2021(34):1440–1452.
  - [23] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of International Conference on Machine Learning. New York: Association for Computing Machinery,2020:1597–1607.
  - [24] HASSANI K, KHASAHMADI A H. Contrastive multi-view representation learning on graphs[C]//Proceedings of International Conference on Machine Learning. New York: Association for Computing Machinery,2020:4116–4126.
  - [25] ZHU J, ROSSI R A, RAO A, et al. Graph neural networks with heterophily[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: Association for Computing Machinery,2021:11168–11176.
  - [26] VELICKOVIC P, FEDUS W, HAMILTON W L, et al. Deep graph infomax[J]. ICLR,2019,2(3):4.
  - [27] 蒋林浦,陈可佳. 基于对比预测的自监督动态图表示学习方法[J]. 计算机科学,2023,50(7):207–221.
  - [28] LIKAS A, VLASSIS N, VERBEEK J J. The global k-means clustering algorithm[J]. Pattern recognition,2003,36(2):451–461.
  - [29] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and computing,2007,17:395–416.
  - [30] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th

- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2014: 701–710.
- [31] TIAN F, GAO B, CUI Q, et al. Learning deep representations for graph clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: Association for Computing Machinery, 2014.
- [32] CAO S, LU W, XU Q. Deep neural networks for learning graph representations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: Association for Computing Machinery, 2016.
- [33] ZHANG X T, LIU H, LI Q M, et al. Attributed graph clustering via adaptive graph convolution[J/OL]. arXiv preprint arXiv: 1906.01210, 2019.
- [34] HE D X, SONG Y, JIN D, et al. Community-centric graph convolutional network for unsupervised community detection[C]//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. San Francisco: International Joint Conferences on Artificial Intelligence, 2021: 3515–3521.
- [35] ZHANG X T, LIU H, WU X M, et al. Spectral embedding network for attributed graph clustering[J]. Neural networks, 2021, 142: 388–396.
- [36] SALEHI A, DAVULCU H. Graph attention auto-encoders[J/OL]. arXiv preprint arXiv: 1905.10715, 2019.
- [37] WANG C, PAN S R, CELINA P Y, et al. Deep neighbor-aware embedding for node clustering in attributed graphs[J]. Pattern recognition, 2022, 122: 108230.

[责任编辑:黄 敏]