计算机科学
COMPUTER SCIENCE

路径掩码自编码器引导无监督属性图节点聚类

丁新宇, 孔兵, 陈红梅, 包崇明, 周丽华

引用本文

丁新宇, 孔兵, 陈红梅, 包崇明, 周丽华. 路径掩码自编码器引导无监督属性图节点聚类[J]. 计算机科学, 2025, 52(1): 160-169.

DING Xinyu, KONG Bing, CHEN Hongmei, BAO Chongming, ZHOU Lihua. Path-masked Autoencoder Guiding Unsupervised Attribute Graph Node Clustering [J]. Computer Science, 2025, 52(1): 160-169.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100117

路径掩码自编码器引导无监督属性图节点聚类

丁新宇[1]　孔　兵[1]　陈红梅[1]　包崇明[2]　周丽华[1]
1 云南大学信息学院　昆明 650504
2 云南大学软件学院　昆明 650504
(13623489952@163.com)

摘　要　图聚类的目的在于发现网络的社区结构。针对目前聚类方法无法很好地获取网络深层潜在社区信息,且不能对特征进行合适的信息整合导致节点社区语义不清晰的问题,提出了一种路径掩码自编码器引导无监督属性图节点聚类模型(Path-Masked Autoencoder Guiding Unsupervised Attribute Graph Node Clustering,PAUGC)。该模型通过对网络进行随机路径掩码后使用自编码器来深度挖掘网络拓扑结构,从而获得良好的全局结构语义信息,利用规范性方法来对特征进行信息整合,使节点特征能够更好地表征特征的类别信息。此外,模型结合模块最大化来抓取整个图中的底层社区群落信息,目的在于更合理地将其融合到低维度节点特征中。最后通过自训练聚类来不断迭代优化更新聚类表示以获得最终的节点特征。通过在8个基准数据集上与11种经典方法进行大量实验对比,证明了PAUGC的有效性。

关键词:深度图聚类;无监督学习;特征信息整合;模块最大化;聚类自训练
中图分类号　TP391

Path-masked Autoencoder Guiding Unsupervised Attribute Graph Node Clustering

DING Xinyu[1],KONG Bing[1],CHEN Hongmei[1],BAO Chongming[2] and ZHOU Lihua[1]
1 School of Information Science and Engineering,Yunnan University,Kunming 650504,China
2 School of Software,Yunnan University,Kunming 650504,China

Abstract　The purpose of graph clustering is to discover the community structure of the network. Aiming at the problem that the current clustering methods can not well obtain the deep potential community information of the network,and can not make suitable information integration of the features,resulting in unclear semantics of the node community,a path-masked autoencoder guiding unsupervised attribute graph node clustering(PAUGC)model is proposed. This model utilizes an autoencoder to deeply dig the network topology structure by randomly masking network paths,thereby obtaining excellent global structural semantic information. Utilizing a normative method for information integration of the features,so that the node features are able to better characterize the class information of the features. In addition,the model combines modularity maximization to capture the underlying community clusters information in the whole graph,aiming to more reasonably fuse it into the low-dimensional node features. Finally,the model iteratively optimizes and updates the clustering representation through self-training clustering to obtain the final node features. By conducting extensive experiments and comparisons with 11 classical methods on 8 benchmark datasets,PAUGC has been proven to be effective compared to current mainstream methods.

Keywords　Deep graph clustering,Unsupervised learning,Feature integration,Module maximization,Self-training for clustering

1　引言

图是一种极其重要的信息存储结构,在日常生活、商业应用和科学研究中得到了广泛应用,如推荐系统[1]、社交网络[2]、生物网络[3]等,它由一系列包含属性的节点和反映节点紧密关系的边构成。社区是由大量具有共同特征、紧密关系

到稿日期:2023-11-19　返修日期:2024-04-27
基金项目:国家自然科学基金(62062066,61762090,61966036,62276227);云南省基础科研项目(202201AS070015);云南省中青年学术和技术带头人后备人才项目(202205AC160033);云南省智能系统与计算重点实验室(202205AG070003);云南大学专业学位研究生实践创新项目(ZC-23234311)
This work was supported by the National Natural Science Foundation of China(62062066,61762090,61966036,62276227),Yunnan Fundamental Research Projects(202201AS070015),Young and Middle-aged Academic and Technical Leaders Reserve Talent Project in Yunnan Province(202205AC160033),Yunnan Key Laboratory of Intelligent Systems and Computing(202205AG070003) and Practical Innovation Project of Postgraduate Students in the Professional Degree of Yunnan University(ZC-23234311).
通信作者:孔兵(kongbing@ynu.edu.cn)

的个体组合而成的集群，一个良好的社区具备以下性质：社区内部节点联系紧密而与外部节点联系稀疏。生活中存在很多社区关系，如一个公司内部不同部门之间构成了多个不同的社区。如何识别良好的社区结构对于阐述图形内部复杂结构是十分重要的[4]。

图聚类算法通过分析不同节点特征之间的相似性和节点之间的联系程度来总结网络底层结构信息，以实现社区分割。传统聚类算法 K-means[5]，HDBSCAN[6] 在社区划分上取得了不错的效果，但它们过分依赖节点特征信息，忽略了网络中信息传递的重要性。

随着深度学习的快速发展，神经网络在深度图聚类上展现出了强大的竞争力。以多层感知机为网络层的模型将节点特征嵌入到更低的维度，以达到压缩节点语义信息的目的，通过不断训练来实现聚类的进一步优化。然而，与传统聚类方法一样，该方法只考虑了节点属性而没有融合网络的结构信息；同时，普通神经网络并不能很好地捕获图的非欧几里得信息，而良好的欧几里得信息能有效地揭示节点之间潜在的关联性。近年来，随着 Kipf 等[7]提出了一种能够融合节点特征信息与网络结构的神经网络——图卷积神经网络（Graph Convolutional Network，GCN），上述问题得到了一定程度的缓解。更进一步，一些模型使用属性重构和图结构重构来更有效地捕获网络的底层信息，希望获得解释性更好的节点嵌入。然而，如何全面有效地提取特征信息，保留细节，提升网络算法判别精度，依然困扰着许多研究者[8]。由于图网络的内在复杂性，上述方法最终实现的特征提取和聚类效果依旧不太理想。

虽然当前主流聚类方法在很多方面有了长足进步，但依旧存在以下问题：1）大部分聚类方法希望通过各种手段来尽可能获取网络深层潜在语义信息，同时减少由于节点间信息交互带来的噪音，但是效果不是很理想；2）大部分方法只注意到了网络结构，忽视了社区群落结构在整个模型中的重要性，社区群落结构越清晰详细，节点聚类效果越好；3）没有对当前获取到的特征进行进一步规范化处理，导致特征在聚簇分类任务上的区分度不够明显，降低了聚类效果。

为此，本文提出了一种路径掩码自编码器引导无监督属性图节点聚类模型（Path-Masked Autoencoder Guiding Unsupervised Attribute Graph Node Clustering，PAUGC），它具备很大的潜力，可以充分挖掘网络拓扑信息、社区群落结构信息和节点关系。本文的主要贡献如下：

1）使用路径掩码结合自编码器实现路径重构，从而尽可能地识别网络中的深层结构信息。此时节点间的信息传递只能通过仅存的部分路径，这会导致特征包含强烈特定结构倾向的语义信息。当经过多次随机路径掩蔽处理后，可得到蕴含完备底层结构信息的节点特征。

2）借用模块最大化使节点之间具备特定的潜在模块化关系，通过最大化模块度函数值来将同属一类的样本进行聚簇而不属于同一类的样本推离，以获取更加清晰的社区分布，从而更加精确地捕捉底层社区群落结构，获得完备的聚类语义信息特征。

3）利用聚类特异性分布来规范化特征信息，通过不断削减高置信度伪样本的可靠性来促进聚类样本的正确划分，进一步提升节点特征在特定空间中的区分度。随着模型的不断训练，特征可以很好地表征聚类结构，提升聚类效果。

4）使用聚类自训练来进一步提升模型聚类性能。该模块的目的在于将低维空间中的节点特征分离开来，使其形成一个个内部联系紧密的整体，最终得到具有高质量判定标准的特征。

2 相关工作

2.1 图聚类

早期的方法通过传统图聚类方法实现聚类簇的划分，但由于传统方法只能在数据表层获取极其有限的语义信息，因此聚类效果不佳。为了解决传统聚类算法的问题，引进了基于神经网络实现的图聚类算法。例如，MGAE[9] 使用图自动编码器来对破坏的节点属性进行恢复，使用生成的低维度节点嵌入进行下游任务，通过不断还原节点属性，弥补了传统方法只局限于数据表面的不足。为解决图卷积导致的特征分解计算复杂的问题，GALA[10] 构建了可以避免拉普拉斯锐化而引起网络数据不稳定的对称自编码器，用于学习图节点的表示。ARGA[11] 构建了一个对抗网络，通过对抗训练模块来强迫潜在码匹配先验分布。DAEGC[12] 采用了多阶邻居的相似矩阵来存储网络的结构信息，形成了具有明确目标导向的节点特征，经过自编码器的处理后形成的嵌入可以较好地应用到下游任务中。然而，上述方法只是简单地利用特征和结构进行节点嵌入，并没有充分利用两者的内部潜在信息。GATE[13] 通过自动编码器来重构图的节点属性和结构，以获得最终的节点嵌入表示。CGCN[14] 将变分自编码器与高斯混合模型相结合，通过编码器重构图的拓扑结构来获得解释性更好的节点嵌入。这两种方法通过重构来深度融合特征和结构信息，不断优化聚类标签并进行合理分配，以实现更好的聚类效果，但是它们无法较好地捕获清晰的社区网络结构。为了缓解该问题，CDBNE[15] 通过注意力自编码器来重构整个图的节点属性，同时结合模块最大化[16] 来捕获网络的社区结构，从而使节点嵌入包含更多的社区结构信息，进而获得语义信息更加丰富的节点特征向量。此外，对比方法在该任务上取得了显著的成果，例如 CCGC[17] 构造高置信度聚类正样本和可信负样本，通过分别最大化和最小化两者之间的横向余弦相似性拉近同一聚类的样本，同时推离其他聚类的样本，极大地提高了聚类性能。另外，非深度学习方法 BGCA[18] 通过 PCA 评估函数挖掘区分性特征子图，结合分支定界加速子图挖掘以及新颖的嵌入式特征选择算法，在中、小规模图聚类上表现优异。

2.2 模块最大化

为了获取底层的社区结构[19]，研究者们提出了很多方法，如生成网络模型的最大似然估计、谱方法和矩阵分解方法[20-21]，但最具代表性的还是模块最大化方法。模块最大化最早是由 Newman 提出的，通过最大化模块化指标可以不断增强网络社区结构强度[22]。通过在训练过程中不断增大模块化的值，可以获得良好的社区网络结构，将社区网络结构信息融入到节点嵌入中，从而获得更加出色的簇分类效果。

2.3 自动编码器

自编码器由编码器和解码器两部分组成，编码器用于将特征映射到某一特征维度，解码器将该维度的特征进行还原，从而用于提取特征中的关键信息。图自编码器最早是在GAE[23]中提出的，它构建了一个由两层GCN组成的编码器，当特征映射到低维时，使用内积解码器进行网络结构的还原。更进一步，VGAER[24]使用了变分编码器，利用多维高斯分布替换图嵌入，使模型不再局限于隐空间中的一点。之后，大多数的编码器都采用以GAE框架为基础的结构重构或者特征结构重构模式(如DGVAE[25]，HGMAE[26]等)。

3 研究方法

3.1 概念和问题定义

给定一个属性图，表示为 $G=(V,E,\boldsymbol{X})$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 是由 n 个节点组成的节点集合，E 是 n 个节点之间相互连接的边的集合，$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbb{R}^{n\times d}$ 表示图的属性矩阵，$\boldsymbol{x}_i$ 是节点 v_i 的属性向量，d 表示节点特征向量维度。图 G 的拓扑结构可以表示为 $\boldsymbol{A}=(a_{ij})_{n\times n}\in\mathbb{R}^{n\times n}$，度矩阵可以表示为 $\boldsymbol{D}=diag(d_1,d_2,\cdots,d_n)\in\mathbb{R}^{n\times n}$，其中 d_i 表示节点 v_i 度之和。

对于属性社区网络 G，社区发现的主要目的是为了发掘出 K 个互不相交的社区，每一个社区具有以下性质：1)社区内节点具有相似的特征与属性；2)节点之间有着较为紧密的直接或间接的联系。

3.2 模型整体框架

模型构造了一个具有两阶段的掩码图自编码器，用于对属性网络进行聚类分割，整体结构如图1所示。

第一阶段包括以下模块。1)信息提取模块(见图1(a))。该模块旨在不影响信息传递的前提下随机消除图网络中的部分路径，通过将图节点特征矩阵 $\boldsymbol{X}$ 与邻接矩阵 $\boldsymbol{A}$ 输入编码器得到特定节点嵌入 $\boldsymbol{Z}$；接下来，将 $\boldsymbol{Z}$ 输入到解码器中来映射原始数据空间，通过包含更多网络信息的高维节点嵌入表征来不断还原拓扑关系，从而达成高效捕获图深层拓扑结构信息的目的；最后，使用解码器输出的网络结构指导矩阵来构建损失 L_{GAES}。相较于其他通过重构节点属性信息的模型来说，特征融合了高质量的结构信息后会更加利于底层信息的发掘。2)最大化模块(见图1(b))。该部分将编码器输出的特定节点嵌入 $\boldsymbol{Z}$ 作为输入得到社区分配矩阵 $\boldsymbol{H}$，从而进一步得到模块最大化损失 L_{MM}。该模块的目的在于将潜在的深层社区群落信息结合到特征向量中，同时捕获更加底层的网络社区结构，通过最大化该模块的指标来获取更加清晰的社区群落结构，优化聚类效果。3)聚类特异性分布模块(见图1(c))。该模块的输入依旧为 $\boldsymbol{Z}$，经过正则化处理之后相乘以得到聚类特异性分布损失 L_{CSD}。它将属于同一簇节点的杂乱分布调整为近似或者相同的分布，从而让同一社区的节点具有相似的特征向量，进一步促进聚类簇的清晰分割。

第二阶段主要由自训练聚类(Self-training Cluster)模块构成(见图1(d))。该阶段希望能够进一步处理第一阶段获得的具有良好的结构信息和底层社区群落信息的节点特征向量，使具有紧密联系的同属类节点不断聚簇，以实现类内距离最小化。通过不断促进正负样本之间的分离，最终推动网络提取到高质量的低维节点特征。

3.3 信息提取

本文使用自编码器(Auto-encoder，AE)来实现信息提取。模块采用GCN作为编码器，将相邻节点属性特征与节点自身属性特征进行融合，从而更新节点属性特征。在多次的消息传递之后，节点属性特征可以感知整个网络，经过编码器的压缩处理得到具有全局信息的低维节点属性嵌入 $\boldsymbol{Z}$。在编码器第 l 层网络中，根据上一层的嵌入向量 $\boldsymbol{Z}^{l-1}$ 和遮蔽邻接矩阵 $\boldsymbol{A}'$，使用 $\mathcal{H}$ 表示编码器来获得本层输出 $\boldsymbol{Z}^l$。

$$\boldsymbol{Z}^l=\mathcal{H}(\boldsymbol{Z}^{l-1},\boldsymbol{A}')=\sigma[\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{I}+\boldsymbol{A}')\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Z}\boldsymbol{W}^l] \tag{1}$$

需要注意的是，式(1)中的 $\boldsymbol{A}'$ 并不是一个完整的图拓扑结构表示，而是只保留了部分路径的邻接矩阵；$\boldsymbol{W}^l\in\mathbb{R}^{f^{l-1}\times f^l}$ 是 l 层的可学习的权重参数矩阵，其中 f^{l-1} 代表上一层的输出维度，f^l 代表本层的输出维度；σ 代表编码器所使用的非线性激活函数，此处选择ELU函数而不是ReLU函数，原因是相比ReLU来说，ELU有两个显著的优点：1)结果具有零均值分布特性，可以有效提升训练速度；2)ELU具备单侧饱和特性，可以更好地收敛。

解码器使用多层感知机MLP。由于当前模型中反映拓扑结构的 $\boldsymbol{A}'$ 是不完整的，如果使用GCN来实现反转编码，则不会取得较好的结果。这是因为解码器无法自然地捕获完整的网络结构信息，即网络之间信息的传递是片面的，所以解码输出并不利于后续的初始拓扑结构恢复。

在经过多层网络处理后可以得到编码器输出，即 $\boldsymbol{Z}\in\mathbb{R}^{n\times d'}$，其中 d' 代表节点嵌入的维度。解码器首先将表征属性 $\boldsymbol{Z}$ 还原到较高维度，以进一步丰富属性特征的语义信息；得到较高维属性特征后，在解码器的最后一层依靠当前节点属性特征来更好地推理出完整的图结构。解码器的输出由式(2)定义：

$$\boldsymbol{F}_{ij}=\text{sigmoid}[MLP(\boldsymbol{Z}_i\circ\boldsymbol{Z}_j)] \tag{2}$$

其中，v_i，v_j 是一对有拓扑联系的节点对，可从 $\boldsymbol{Z}$ 得到该拓扑节点对的嵌入向量 $\boldsymbol{Z}_i$，$\boldsymbol{Z}_j$。式(2)中$\circ$代表矩阵点乘，解码器使用非线性激活函数sigmoid。$\boldsymbol{F}_{ij}$ 是一个0～1的实数，该结果表示节点 v_i 和 v_j 的亲疏关系，值越大表示节点间关系越亲密，反之则越疏远，它暗示着节点之间存在连接的可能性。本实验使用二元交叉熵损失来衡量结构重建效果，拓扑结构重构损失如式(3)所示：

$$L_{\text{GAES}}=-\left[\frac{1}{|e^+|}\sum_{(u,v)\in e^+}\log h_D(\boldsymbol{Z}_u,\boldsymbol{Z}_v)+\frac{1}{|e^-|}\sum_{(u',v')\in e^-}\log(1-h_D(\boldsymbol{Z}_{u'},\boldsymbol{Z}_{v'}))\right] \tag{3}$$

其中，e^+ 为正边集合(即图中自然存在的边)，e^- 为负边集合(即人为构建的非自然存在的边)，函数 h_D 表示解码过程。等式右边前半部分不断提升正节点对之间边连接的置信度，后半部分代表对负边的惩罚。随着该等式中值的不断减小，正节点对之间边的置信度在不断提高，同时负边的惩罚会逐渐减小，即模型逐步恢复正节点对的连接而剔除负节点对的连接，从而最大程度地还原网络拓扑结构；同时，该损失的减小，表示编码器输出的节点属性特征逐渐被优化，减少了全局无关信息的影响，提升了节点质量。

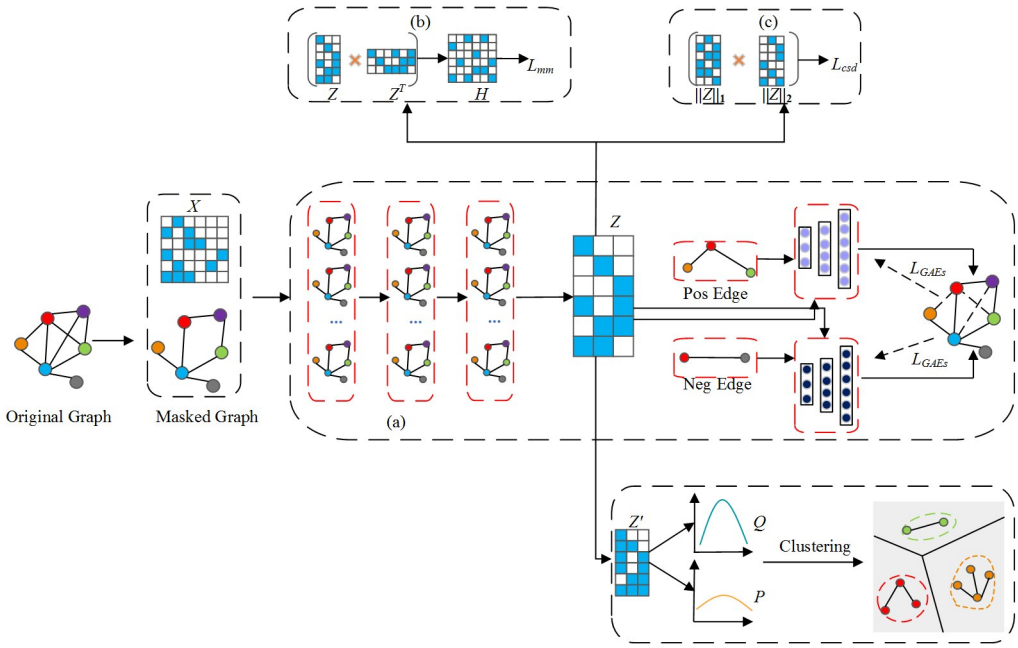


图 1 PAUGC 模型的整体框架

Fig. 1 Overall framework of PAUGC model

3.4 模块最大化

为了更好地体现社区广泛潜在的群落结构，本文引入模块最大化。使用指标 Q 来表示网络整体模块化强度，模块化越强，社区群落结构节点越清晰，聚类效果越好。具体定义如式(4)所示：

$$Q=\frac{1}{2m}\sum_{i,j}\left[\left(\boldsymbol{A}_{ij}-\frac{k_i k_j}{2m}\right)\boldsymbol{C}(i,j)\right] \tag{4}$$

其中，k_i 表示节点 i 的度，$\boldsymbol{C}(i,j)$表示节点 i 与 j 是否隶属于同一个社区，若是同一个社区则值为 1，否则值为 0。此外，m 代表图网络中的总边数。

为进一步简化上述等式，引入模块化矩阵 $\boldsymbol{B}\in\mathbb{R}^{n\times n}$和社区分配矩阵 $\boldsymbol{H}\in\mathbb{R}^{n\times n}$。通过模块化矩阵的处理后，所有节点之间都会具有模块关系，这进一步为节点间信息传递提供了强力的结构支撑。模块化矩阵 $\boldsymbol{B}$ 如式(5)所示：

$$\boldsymbol{B}_{ij}=\boldsymbol{A}_{ij}-\frac{k_i k_j}{2m} \tag{5}$$

接下来介绍社区分配矩阵 $\boldsymbol{H}$，该矩阵用于记录节点所处的社区信息。H 的维度设置为 n，其不是社区的数量，这是由于节点的数量远多于网络中的社区数量，这样做的好处在于能够获得更加丰富的语义。$\boldsymbol{H}$ 的定义如式(6)所示：

$$\boldsymbol{H}=\mathrm{sigmoid}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}) \tag{6}$$

这里使用内积编码来实现 $\boldsymbol{H}$，当引入了社区分配矩阵 $\boldsymbol{H}$ 后，可以将模块最大化表示式(4)转化为式(7)，其中 Tr 代表矩阵的迹。式(7)的定义如下：

$$Q=\frac{1}{2m}Tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{H}) \tag{7}$$

然而求解模块化指标 Q 是一个 NP 难问题，为此，利用 VGAER 提出的方法将 $Tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})=n$ 作为条件对问题进行松弛，即通过 $\max(Q)$的值来获得宽松的模块化优化问题。模块最大化损失函数如式(8)所示：

$$L_{\mathrm{MM}}=F(\max(\boldsymbol{Q})) \tag{8}$$

由于 Q 的值普遍较大，因此构建了一个同等放缩的函数 F 来尽量缩小它的值但不影响其捕获社区群落结构的能力。

3.5 聚类特异性分布

本节介绍聚类特异性分布(Cluster Specificity Distribution，CSD)在模型中的功能。在 Auto-Encoder 的模型中，编码器在编码后立即将特征向量送入解码器中进行解码，并未考虑空间节点中的聚类特异性分布。当进行簇分割时，显而易见的一点是：处于同一社区的节点不仅有相似的特征，它们在该空间的分布也应该是相似的。聚类特异性分布的损失定义如式(9)所示：

$$L_{\mathrm{CSD}}=\sum_{i=1}^{n}\sum_{j=1}^{n}\|Z\|_1\circ\|Z\|_2 \tag{9}$$

分别使用 l_1 和 l_2 正则化处理编码器输出 $\boldsymbol{Z}$，之后让二者进行点乘，最终求和得到聚类特异性分布损失。通过最小化该损失，可以得到这样的节点特征：特征向量中除了少数维度元素非 0，大部分维度的元素都为 0 或者接近 0。最终，每个特征都会保留为数不多的几个明显的判别特征维度，这为确定节点隶属于哪个社区提供了更强的置信水准，同时会得到特异性分布更好的节点特征。

3.6 自训练聚类

自训练聚类是模型的第二阶段，该阶段使用学生 t 分布[27]来衡量特征 $\boldsymbol{z}_i$ 与社区中心向量 $\boldsymbol{\mu}_j$ 之间的相似性，式(10)定义了衡量过程。

$$q_{ij}=\frac{(1+\|\boldsymbol{z}_i-\boldsymbol{\mu}_j\|^2)^{-1}}{\sum_k(1+\|\boldsymbol{z}_i-\boldsymbol{\mu}_k\|^2)^{-1}} \tag{10}$$

其中，软分布 q_{ij} 表示节点 i 分配到社区 j 的概率，$\boldsymbol{\mu}$ 则是由 K-means 算法处理得到的社区中心向量，k 代表社区中心向量的数量。同时，理想分布 p_{ij} 的定义如式(11)所示：

$$p_{ij}=\frac{q_{ij}^2/\Sigma_i q_{ij}}{\sum_k(q_{ik}^2/\Sigma_i q_{ik})} \tag{11}$$

式(12)定义了自训练聚类的损失：

$$L_{st}=\min[-\Sigma_i\Sigma_j p_{ij}\log q_{ij}] \tag{12}$$

式(12)希望真实分布 q 与理想分布 p 在训练中不断接近，通过二者的互相对抗监督来不断进行聚类优化，使同属类节点靠近，不同属类节点疏远，最终得到具有清晰边界轮廓的社区。

3.7 模型训练

合并式(3)、式(8)和式(9)得到第一阶段损失函数：

$$L_1=[\alpha L_{GAES}-\beta L_{MM}+\gamma L_{CSD}] \tag{13}$$

其中，α，β，γ 为 L_1 的权重参数，用于调整各个模块的损失比例。第二阶段的损失函数定义为：

$$L_2=L_{st} \tag{14}$$

模型训练总体流程如下：首先将节点特征与网络拓扑结构输入到模型中，此时通过信息提取、模块最大化和聚类特异性分布这 3 个模块来进行第一阶段的处理；当第一阶段结束后，由编码器输出的具有明显判别特征的低维节点嵌入蕴含了丰富的拓扑结构和社区群落信息，此时进行第二阶段的处理，将之前的节点嵌入在特定空间中进行调整和约束，以获得更优的聚类效果。

算法描述如算法 1 所示。

算法 1 路径掩码自编码器引导无监督属性图节点聚类算法

输入：邻接矩阵 **A**，节点特征矩阵 **X**，聚簇数量 K，第一阶段预训练迭代次数 T_1，第二阶段正式训练迭代次数 T_2，运行次数 R，更新间隔 t

输出：低维节点表示矩阵 **Z**

1. 第一阶段
2. 初始化模型参数；
3. for epoch_1=0 to T_1 do
4. 根据式(13)计算预训练损失并反向传播以更新参数，从而最小化预损失；
5. end for
6. 使用 K-means 计算社区中心 **μ**；
7. 第二阶段
8. for run=0 to R do
9. for epoch_2=0 to T_2 do
10. if epoch_2 % t==0 then
11. 计算 ACC，NMI，ARI；
12. 与当前最好指标进行对比，如果大于该指标就进行更新，否则不更新；
13. end if
14. 按照式(10)计算真实分布 q；
15. 按照式(11)计算理想分布 p；
16. 将计算出来的 q 和 p 带入式(14)中计算正式训练损失，更新该部分框架；
17. end for
18. end for
19. 将 R 次迭代的指标取平均得到最终结果。

4 实验

4.1 实验数据集

为了验证模型的可靠性，选择在 8 个数据集上进行实验，包括 Cora[28]，Citeseer[28]，Pubmed[28]，Photo[29]，Acm[29]，Eat[17]，Uat[17]，Amap[17]。数据集统计信息如表 1 所列。

表 1 数据集统计信息

Table 1 Datasets statistics

数据集	节点数量	边数量	特征维度	种类
Cora	2708	10556	1433	7
Citeseer	3327	9104	3703	6
Pubmed	19717	88648	500	3
Photo	7650	238162	745	8
Acm	3025	29281	1870	3
Eat	399	5994	203	4
Uat	1190	13599	239	4
Amap	7650	119081	745	8

4.2 实验环境

实验环境如下：CPU 为 Intel I9 11900k，GPU 为 NVIDIA GeForce RTX 3090，RAM 为 64GB，实验框架为 Pytorch。

4.3 对比方法

为了证明本文方法的有效性，将其与其他 11 种方法在 8 个数据集上进行了聚类性能的比较。这 11 种方法包括：

1) K-means[5]：一种主流的传统聚类算法。该方法初始化聚类中心，然后通过迭代更新优化聚簇中心的位置，不断降低类簇的 SSE 以得到最终的聚类结果。

2) MGAE[9]：采用了去噪自编码器的思想，首先通过手动破坏网络节点内容来模拟噪声，接着使用图自动编码器进行节点编码，最后使用谱聚类算法对节点进行合理聚类分割。

3) ARGA[11]：采用了图自编码器来获取网络拓扑结构和节点属性，此外，通过对抗训练来迫使潜在节点分布逼近先验以获得更加健壮的节点表示。

4) DAEGC[12]：该方法使用了基于图注意力的编码器，结合聚类对齐损失来不断优化节点特征。

5) AGE[28]：设计了一个合适的拉普拉斯平滑滤波器，可以在缓解高频噪声的同时有效解决滤波器和权重矩阵之间的纠缠，之后对正负样本进行自适应对比来不断训练编码器。

6) SDCN[30]：该方法首先使用初始节点构造 KNN 图，然后联合自动编码器和 GCN 对 KNN 图和初始特征进行表征信息结合，最后使用双重自监督来不断细化聚类过程。

7) DFCN[31]：提出了一个将网络结构和属性信息融合的模块，便于进行特征共识表示学习，同时构建了一个三重自监督策略来不断监督目标生成分布，以提升聚类性能。

8) MVGRL[32]：该方法首先对结构视图进行增强处理，并将视图中的节点表示与另一个视图的图进行对比，从而不断强化节点表示。

9) AGC-DRR[33]：构造了结构增强子网络和聚类子网络，将图的结构增强和样本聚类统一为一个整体，从而有效减少特征空间中的信息冗余。

10) AFGRL[34]：提出了一种不需要数据增强和负采样的方法，而是通过 KNN 确定可以作为正样本的节点组成正视角。通过其中一个视角中的节点信息表征来预测另一个视角中的表征，使两个视角中的表征尽量相似，从而不断提高模型效率。

11) GDCL[35]：构建了一个图去偏对比学习方法框架，用于解决负抽样中的采样偏差问题，使用样本表示与聚类信息对齐来进行优化，以促进样本特征更具识别性。

4.4 实验设置

为了让实验的对比结果更加具有信服力，所有对比实验的超参数均采取原文设置。对于 PAUGC，第一阶段的训练迭代次数依次设置为 100～800，第二阶段的训练迭代次数统一设置为 500。模型的编码器由两层 GCN 组成，两层网络维度分别为[初始嵌入维度，128]和[128，64]，解码器由两层 MLP 组成，两层网络输出维度分别为 32 和 1。α，β，γ 分别设置为 1，0.01，0.01，通过网格搜索来不断调整超参数的值，以获取到更好的指标结果。第一阶段的学习率设置为 0.01，第二阶段学习率范围为 0.01。通过早停机制来调整学习率，最大间隔次数为 50。

4.5 实验指标

实验采用的聚类指标为准确率 ACC、标准化互信息 NMI 和调整兰德系数 ARI。

ACC 表示正确预测样本数与整体样本数之比，其其定义如式(15)所示：

$$ACC(r,s)=\frac{\sum_{i=1}^{N}\delta(r_i,map(s_i))}{n} \tag{15}$$

其中，r_i 表示第 i 个样本的真实种类标签，s_i 表示模型对第 i 个样本的预测类别标签；$map()$ 函数是一个映射函数，用于协助将预测的类别标签转化为更精准的类别标签；$\delta(x,y)$ 是一个二分类的函数，$x=y$ 时其值为 1，否则值为 0。该指标范围在 0～1 之间，指标越大表示预测越精准，聚类效果越好。

在引入标准化互信息 NMI 之前，需要介绍互信息 MI。MI 是一种信息度量手段，可视为一个变量中包含关于其他变量的信息量。MI 的定义如式(16)所示：

$$I(X;Y)=H(X)+H(Y)-H(X,Y) \tag{16}$$

其中，$H(X)$ 函数代表 X 的熵，$H(X,Y)$ 代表 X,Y 的联合信息熵。有了互信息的定义之后，可以得到标准化互信息 NMI 的定义，具体如下：

$$NMI(r,s)=\frac{2I(r;s)}{H(r)+H(s)} \tag{17}$$

其中，$H(r)+H(s)$ 分别代表真实种类标签的熵和预测种类标签的熵。NMI 的范围为 0～1，值越高表示两个聚类结果越相似。

同样地，调整兰德系数 ARI 是建立在兰德系数 RI 基础上的。RI 的计算式如下：

$$RI=\frac{a+b}{C_n^2} \tag{18}$$

其中，a 表示样本对既属于真实种类标签 r 也属于预测种类标签 s 的个数，b 表示样本对既不属于真实种类标签 r 也不属于预测种类标签 s 的个数。该公式的分母表示所有可能的样本对个数。虽然 RI 取得了一定效果，但是该方法无法保证随机划分的聚类结果值接近 0。在此基础上，提出调整兰德系数 ARI，其定义如下：

$$ARI=\frac{RI-E[RI]}{\max(RI)-E[RI]} \tag{19}$$

其中，E 表示期望，max 表示最大值。该指标取值范围在 −1～1 之间，值越大代表预测聚类簇与真实聚类簇的相似程度越高。

4.6 实验结果

实验结果如表 2 与表 3 所列，最好的结果使用黑色粗体表示。可以观察到：

1)PAUGC 相较于其他基线在 3 个指标上展现出了很大的优势。在 Cora 数据集上，本文方法比第二好的方法在指标 ACC 上提高了 5.53%，在 NMI 上提高了 3.61%，在 ARI 上提高了 9.43%；在 Amap 数据集上，其相较于第二好的方法，ACC 提升了 4.25%，NMI 提升了 8.02%，ARI 提升了 7.49%。

2)以 Cora 数据集为例，PAUGC 在 ACC 指标上分别比 K-means 和 SDCN 高出 34.75%和 43.43%，体现了本文模型的优越性。其原因在于这两个对比模型都未能很好地捕获底层结构语义信息，而是将注意力集中在表层节点特征属性上，使得在获得低维嵌入特征时引入了多余的冗杂信息。PAUGC 可以联合多个模块，从而很好地从复杂网络中将抽象的拓扑与社区结构信息融合到特征向量中，同时通过路径掩码策略有效避免表征冗余，有利于得到分布合理、轮廓紧密的拓扑群落。

3)更进一步，为了体现出 PAUGC 中模块最大化的有效性，以 DAEGC 方法为对比模型。该方法同样通过内积解码器将编码器输出的特征进行邻接矩阵重建，虽然在一定程度上 DAEGC 能够较好地融合特征和结构信息，但是其忽略了网络的社区群落微观结构，使得低维特征在空间上分布杂乱。而 PAUGC 可以有效挖掘网络社区群落语义，将这些语义注入到低维特征中，使表征具有强大的社区结构信息，从而提升节点的聚类可靠性。

4)此外，以上基线模型都忽视了一个问题，即如何让特征具有更加明显的社区区分度。PAUGC 使用聚类特异性分布得到了具有显著区分效果的特征，可以增强对样本节点的判别能力，有效缓解因为噪音导致的节点社区划分错误的问题。总之，PAUGC 是高效的，这得益于各个模块能够强有力地捕获各部分的关键语义信息。

表 2 各算法在 Cora 等数据集上的指标对比

Table 2 Performance comparison of multiple algorithms on datasets like Cora

(%)

Methods	Cora			Citeseer			Pubmed			Photo		
	ACC	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
K-means	44.28	27.36	16.12	55.82	29.42	26.72	59.53	31.34	28.19	29.19	16.88	6.69
MGAE	43.38	28.78	16.43	61.35	34.63	33.55	59.30	28.20	24.80	43.88	8.16	41.98
DAEGC	70.43	52.89	49.63	64.54	36.41	37.78	67.10	26.60	27.80	76.00	65.30	58.10
ARGA	71.04	51.06	47.71	61.07	34.40	34.32	68.10	27.60	29.10	69.30	58.40	44.20
SDCN	35.60	14.28	7.78	65.96	38.71	40.17	64.20	22.87	22.30	53.40	44.90	31.20

（续表）

Methods	Cora			Citeseer			Pubmed			Photo		
	ACC	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
AGE	73.50	57.58	50.10	69.73	44.93	45.31	71.10	31.60	33.40	75.90	65.40	55.90
MVGRL	70.47	55.57	48.70	62.83	40.69	34.18	67.01	31.59	29.42	50.40	43.31	23.79
AGC-DRR	68.61	51.05	48.00	68.32	43.28	45.34	59.96	16.14	16.46	76.80	66.50	60.10
AFGRL	26.25	12.36	14.32	31.45	15.17	14.32	65.16	31.27	28.26	76.77	65.71	57.34
GDCL	70.83	56.30	48.05	66.39	39.52	41.07	70.19	34.48	33.07	43.80	37.30	21.60
PAUGC	**79.03**	**61.19**	**59.53**	**71.00**	**45.57**	**46.60**	**73.11**	**38.49**	**37.48**	**79.24**	**73.06**	**62.16**

表 3 各算法在 Acm 等数据集上的指标对比

Table 3 Performance comparison of multiple algorithms on datasets like Acm

（%）

Methods	Acm			Eat			Uat			Amap		
	ACC	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
K-MEANS	69.19	33.97	32.77	39.85	18.92	10.53	43.03	22.61	15.96	29.61	17.06	6.78
MGAE	88.10	62.10	68.70	44.61	15.60	13.40	48.97	20.69	18.33	71.57	62.13	48.82
DAEGC	86.94	56.18	59.35	36.89	5.57	5.03	52.29	21.33	20.50	75.96	65.25	58.12
ARGA	62.90	28.34	29.42	52.13	22.48	17.29	49.31	25.44	16.57	69.28	58.36	44.18
SDCN	90.45	68.31	73.91	39.07	8.83	6.31	52.25	21.61	21.63	53.44	44.85	31.21
DFCN	90.90	69.40	74.90	49.37	32.90	23.25	33.61	**26.49**	11.87	76.82	66.23	58.28
AGE	90.91	69.42	74.96	47.26	23.74	16.57	52.37	23.64	20.39	75.98	65.38	55.89
MVGRL	86.73	60.87	70.10	32.88	11.72	4.68	44.16	21.53	17.12	41.07	30.28	18.77
AGC-DRR	89.30	65.30	71.00	37.37	7.00	4.88	42.64	11.15	9.50	76.81	66.54	60.15
AFGRL	89.98	67.94	72.87	37.42	11.44	6.57	41.50	17.33	13.62	75.51	64.05	54.45
GDCL	67.60	29.08	27.52	33.46	13.22	4.31	48.70	25.10	**21.76**	43.75	37.32	21.57
PAUGC	**91.50**	**70.57**	**76.51**	**56.89**	**29.18**	**26.8**	**55.46**	25.32	20.02	**81.07**	**74.25**	**65.77**

4.7 消融实验

本节通过消融实验来验证聚类特异性分布与模块最大化的有效性。表 4 与表 5 列出了消融实验的结果，使用 B 来代表由图自动编码器和 GAE 损失为主构成的基础框架模型，B-CSD，B-MM 和 B-CSD-MM 分别代表添加了聚类特异性分布模块、模块最大化技术和两者都使用的方法。表中粗体数据表示最优的结果，仔细分析该表可以得出以下结论：

1）基线 B 经过最大化模块处理之后，在大部分数据上都有了显著提高，但是由于缺乏节点区分的强力表征信息，特征很杂乱，这无形中降低了聚类的质量。

2）在基线 B 的基础上使用聚类特异性分布模块后，虽然在 3 个指标上都取得了较为理想的效果，但由于没有考虑深层聚簇信息，聚类性能并未达到预期。

3）从结果可以看出，B，CSD 和 MM 三者同时使用的聚类性能最好。以 CORA 和 AMAP 数据集为例，在 CORA 数据集上使用 B，CSD 和 MM 的 ACC，NMI，ARI 分别超过 B 5.8%，3.39%，5.73%，在 AMAP 数据集上使用 B，CSD 和 MM 的 ACC，NMI，ARI 分别超过 B 11.82%，5.53%，12.81%。这是因为基线集合了 CSD 和 MM 之后能够很好地获取网络的拓扑信息和社区信息，进一步减少节点特征中携带的结构噪音，同时能够提高特征节点的社区辨别能力，从而构建出适合下游任务的低维特征嵌入。

表 4 在 Cora 等数据集上进行消融实验

Table 4 Ablation experiments on datasets like Cora

（%）

Methods	Cora			Citeseer			Pubmed			Photo		
	ACC	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
B	73.23	57.80	53.80	66.01	41.00	40.54	68.13	27.97	28.78	60.82	58.47	34.30
B-CSD	74.37	58.48	51.95	68.53	43.32	43.56	65.67	25.14	24.63	77.53	69.30	60.17
B-MM	73.01	57.19	49.63	68.65	43.38	44.25	72.59	37.55	36.33	68.80	67.35	51.80
B-CSD-MM	**79.03**	**61.19**	**59.53**	**71.00**	**45.57**	**46.60**	**73.11**	**38.49**	**37.48**	**79.24**	**73.06**	**62.16**

表 5 在 Acm 等数据集上进行消融实验

Table 5 Ablation experiments on datasets like Acm

（%）

Methods	Acm			Eat			Uat			Amap		
	ACC	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*	*ACC*	*NMI*	*ARI*
B	68.00	50.37	45.99	55.89	28.47	24.79	54.45	24.41	17.02	69.25	68.72	52.96
B-CSD	91.44	70.29	76.33	55.14	27.69	23.80	50.50	21.79	12.78	79.46	73.31	62.63
B-MM	67.97	49.11	44.65	56.64	29.11	25.28	48.74	23.47	15.22	79.12	73.23	62.08
B-CSD-MM	**91.50**	**70.57**	**76.51**	**56.89**	**29.18**	**26.80**	**55.46**	**25.32**	**20.02**	**81.07**	**74.25**	**65.77**

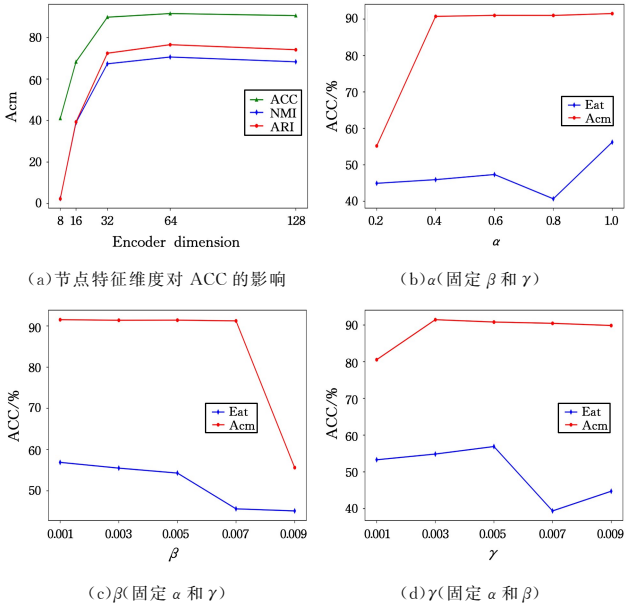
4.8 参数敏感性分析

本节研究参数对模型效果的影响，主要分为以下两部分：

1）节点特征维度分析。以 Acm 数据集为实验标准，为了直观显示节点特征向量维度对潜在信息的影响，模型在固定其他设置的基础上，将编码器输出的特征维度分别设置为 8，16，32，64，128，以测试对指标的影响，具体结果如图 2（a）

所示。由此可以得出结论：随着维度的增加，编码器对于网络关键信息把握得越准确，模型的性能也越来越好。随着维度逐渐变大，模型的性能开始下降，这可能是因为模型中混入了较多噪音，使得聚类的效果变差。

2）目标函数权重参数比例分析。我们采用 Eat 与 Acm 数据集为测试对象来测试指标 ACC，对于目标函数的 3 个权重参数 α，β，γ，通过控制其中两个参数的值不变，而修改另外一个参数的值来进行参数敏感性分析，结果如图 2(b)－图 2(d)所示。对于 β 与 γ，选用较小的权重值，这是因为最大化模块和特异性分布模块的值较大，当参数值较大时整体数值也就越大，对模型的影响就越大。通过分析图形可以得出以下结论：在 Acm 数据集上，3 个参数的变化较为稳定，在大多数情况下都可以取得较好的结果。在 Eat 数据集上，虽然模型在一定程度上会被参数变化影响，但是当 α 在[0.2，0.6]、β 和 γ 在[0.001，0.005]范围内时模型较为稳定。



(a)节点特征维度对 ACC 的影响　(b)α(固定 β 和 γ)

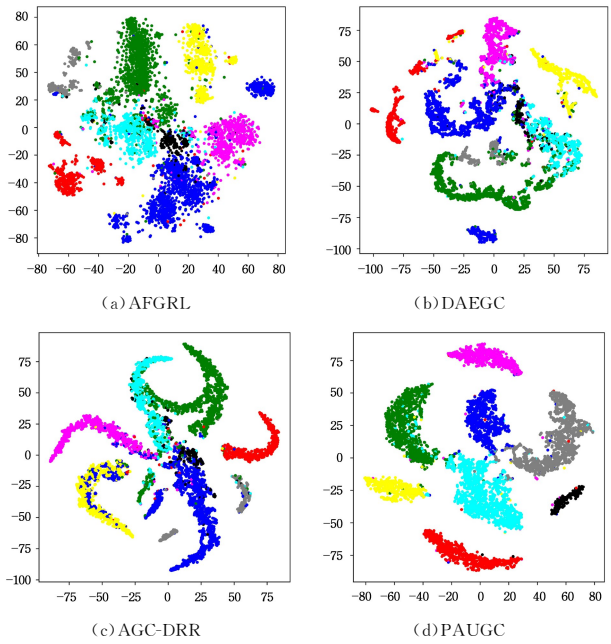
(c)β(固定 α 和 γ)　(d)γ(固定 α 和 β)

图 2　超参数对模型的影响

Fig. 2　Influence of hyperparameters on model

4.9　可视化分析

为了直观验证模型 PAUGC 的有效性，使用 t-SNE 聚类可视化来分析 PAUGC 与 AFGRL，DAEGC 以及 AGC-DRR 在 Amap 数据集上的聚类效果。图 3 直观显示了 4 种方法在二维空间中的低维表征可视化，Amap 数据集一共有 8 种类型，分别使用 8 种颜色表示。在 4 种方法中，AFGRL 的效果最差，图中大量没有关系的节点堆叠纠缠在一起，并且相同节点之间并不紧凑，降低了聚类效果。相较于 AFGRL，DAEGC 与 AGC-DRR 的效果要好，前者内部联系较为紧密，后者边界轮廓更加清楚，但是它们内部混入了大量的不同颜色节点，说明这两种方法获得的低维特征嵌入没有较好的解释性。PAUGC 改进了它们的问题，对于相同类别的节点，PAUGC 可以很好地将它们聚合到一起，形成一个个内部关系牢固的整体，同时在一定程度上减少了不同类别相似节点间的纠缠，最终形成清晰而又相互独立的空间轮廓。另一方面，PAUGC 明显减小了不同颜色节点混杂的程度，提高了嵌入质量。



(a)AFGRL　(b)DAEGC

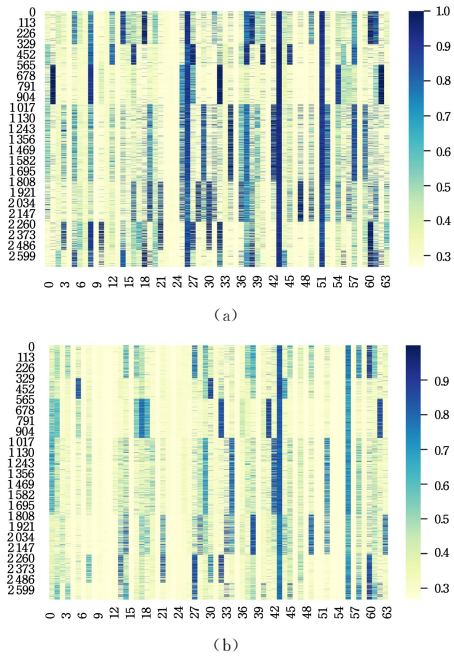
(c)AGC-DRR　(d)PAUGC

图 3　AMAP 数据集二维可视化

Fig. 3　Two-dimensional visualization of AMAP dataset

4.10　亲和图分析

图 4(a)给出了 Cora 数据集在未使用聚类特异性分布 CSD 处理后的特征嵌入，图 4(b)给出了使用 CSD 之后的特征嵌入。可以发现，后者具有更加清晰的特征表示，即特征维度中大部分值很小，只有某些维度的值较大，这为节点簇归属提供了强有力的语义支撑，提高了节点特征的聚类分割置信度。



(a)

(b)

图 4　Cora 数据集特征嵌入亲和图

Fig. 4　Feature embedding affinity graph of Cora dataset

结束语　本文提出了一种路径掩码自编码器引导无监督属性图节点聚类方法 PAUGC，通过构建掩码自编码器来发掘深层潜在结构信息。在此基础上，结合模块最大化来得到丰富的社区群落语义信息，从而获得高质量嵌入特征；同时，

模型使用聚类特异性分布模块来提升将节点划分到正确社区的能力，避免因冗余噪声带来不利影响而降低聚类性能。此外，PAUGC使用自训练聚类在低维隐空间中对节点特征进行约束和整合，从而不断提升相应节点聚类分配的置信度，进一步提升聚类效果。在大量实验中，PAUGC在ACC，NMI，ARI这3个指标上相比当前主流模型取得了较为理想的效果。

但是，目前该模型还存在一些明显的弊端。PAUGC的所有任务都是以同质图和单视图为条件实现的；其次，由于现实世界中的某些图数据是非常庞大的，如何在大型图上高效实现聚类仍是一项艰巨的任务。因此，接下来的任务是将模型拓展到多视图、异构图、大型图。

参 考 文 献

[1] LIU C, WEN L, KANG Z, et al. Self Supervised Consensus Representation Learning for Attributed Graph[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 2654-2662.

[2] WANG M, WANG C, YU J X, et al. Community Detection in Social Networks: An In Depth Benchmarking Study with A Procedure Oriented Framework[J]. Proceedings of the VLDB Endowment, 2015, 8(10): 998-1009.

[3] GARCIA J O, ASHOURVAN A, MULDOON S, et al. Applications of Community Detection Techniques to Brain Graphs: Algorithmic Considerations and Implications for Neural Function [J]. Proceedings of the IEEE, 2018, 106(5): 846-867.

[4] KRISHNAMURTHY B, WANG J. On Network-Aware Clustering of Web Clients[C]// Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. New York: ACM Press, 2000: 97-110.

[5] CHIANG M M T, MIRKIN B. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads[J]. Journal of Classification, 2010, 27: 3-40.

[6] MCINNES L, HEALY J, ASTELS S. HDBSCAN: Hierarchical Density Based Clustering[J]. Open Source Softw, 2017, 2(11): 205.

[7] KIPF T N, WELLING M. Semi-Supervised ClassificAtion with Graph Convolutional Networks[J]. arXiv: 1609.02907, 2016.

[8] TANG J M, HAN H, HUANG L. Coarse grained and Fine-grained Features Extraction Based on Unsupervised Learning in Pedestrian Reidentification [J]. Computer Engineering, 2022, 48(4): 269-275, 283.

[9] WANG C, PAN S, LONG G, et al. Mgae: Marginalized Graph Autoencoder for Graph Clustering[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM Press, 2017: 889-898.

[10] PARK J, LEE M, CHANG H J, et al. Symmetric Graph Convolutional Autoencoder for Unsupervised Graph Representation Learning[C] // Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE Press, 2019: 6519-6528.

[11] PAN S, HU R, LONG G, et al. Adversarially Regularized Graph Autoencoder for Graph Embedding [C] // Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: IJCAI Press, 2018: 2609-2615.

[12] WANG C, PAN S, HU R, et al. Attributed Graph Clustering: A Deep Attentional Embedding Approach[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCA Press, 2019: 3670-3676.

[13] SALEHI A, DAVULCU H. Graph Attention Auto-Encoders [C]// Proceedings of the 32nd IEEE International Conference on Tools with Artificial Intelligence. ELECTR NETWORK: IEEE Press, 2020: 989-996.

[14] HUI B, ZHU P, HU Q. Collaborative Graph Convolutional Networks: Unsupervised Learning Meets Semi-Supervised Learning [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 4215-4222.

[15] ZHOU X, SU L, LI X, et al. Community Detection Based on Unsupervised Attributed Network Embedding[J]. Expert Systems with Applications, 2023, 213: 118937.

[16] NEWMAN M E J. Modularity and Community Structure in Networks[J]. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577-8582.

[17] YANG X, LIU Y, ZHOU S, et al. Cluster Guided Contrastive Graph Clustering Network[C]// Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023: 10834-10842.

[18] ZHANG S, PANG J, LIAO M X. Graph Clustering Algorithm Based on Hybrid Feature Selection[J]. Journal of Chinese Computer Systems, 2024, 45(3): 606-612.

[19] SCHAUB M T, DELVENNE J C, ROSVALL M, et al. The Many Facets of Community Detection in Complex Networks [J]. Applied Network Science, 2017, 2(1): 1-13.

[20] FORTUNATO S. Community Detection in Graphs[J]. Physics Reports, 2010, 486(3/4/5): 75-174.

[21] FORTUNATO S, HRIC D. Community Detection in Networks: A User Guide[J]. Physics Reports, 2016, 100(659): 1-44.

[22] AGARWAL G, KEMPE D. Modularity Maximizing Graph Communities via Mathematical Programming [J]. The European Physical Journal B, 2008, 66(3): 409-418.

[23] KIPF T N, WELLING M. Variational Graph Auto-Encoders [J]. arXiv: 1611.07308, 2016.

[24] QIU C, HUANG Z, XU W, et al. VGAER: Graph Neural Network Reconstruction Based Community Detection [J]. arXiv: 2201.04066, 2022.

[25] LI J, YU J, LI J, et al. Dirichlet Graph Variational Autoencoder [J]. Advances in Neural Information Processing Systems, 2020, 33: 5274-5283.

[26] TIAN Y, DONG K, ZHANG C, et al. Heterogeneous Graph Masked Autoencoders[C]// Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2023: 9997-10005.

[27] XU H, XIA W, GAO Q, et al. Graph Embedding Clustering: Graph Attention Auto Encoder with Cluster-Specificity Distribution[J]. Neural Networks, 2021, 142: 221-230.

[28] CUI G, ZHOU J, YANG C, et al. Adaptive Graph Encoder for Attributed Graph Embedding [C] // Proceedings of the 26th

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Electr Network:KDD Press,2020:976985.

[29] PENG Z,LIU H,JIA Y,et al. Deep Attention Guided Graph Clustering with Dual Self Supervision[J]. IEEE Transactions on Circuits and Systems for Video Technology,2023,33(7):3296-3307.

[30] BO D,WANG X,SHI C,et al. Structural Deep Clustering Network[C]// Proceedings of the 29th World Wide Web Conference. 2020:1400-1410.

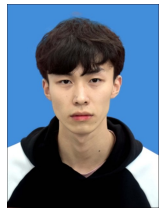
[31] TU W,ZHOU S,LIU X,et al. Deep Fusion Clustering Network [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Electr Network:AAAI Press,2021:9978-9987.

[32] HASSANI K,KHASAHMADI A H. Contrastive Multi-View Representation Learning on Graphs[C]// Proceedings of the 37th International Conference on Machine Learning. Electr Network:ICML,2020:4074-4084.

[33] GONG L,ZHOU S,TU W,et al. Attributed Graph Clustering with Dual Redundancy Reduction[C]// Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna:IJCAI Press,2022:3015-3021.

[34] LEE N,LEE J,PARK C. Augmentation-Free Self-Supervised Learning on Graphs[C]// Proceedings of the 36th AAAI Conference on Artificial Intelligence. Electr Network:AAAI Press,2022:7372-7380.

[35] ZHAO H,YANG X,WANG Z,et al. Graph Debiased Contrastive Learning with Joint Representation Clustering[C]// Proceedings of the 30th International Jo int Conference on Artificial Intelligence. Electr Network:IJCAI Press,2021:3434-3440.



DING Xinyu, born in 1997, postgraduate, is a member of CCF (No. R3213G). His main research interests include deep graph clustering and data mining.

KONG Bing, born in 1968, Ph.D, associate professor. His main research interests include social network analysis and machine learning.

(责任编辑:何杨)