超越同质性假设的双通道属性图聚类

安俊秀，柳源，杨林旺
（成都信息工程大学软件工程学院，四川 成都 610207）

摘　要：属性图聚类的研究近些年取得了显著进步，但现有方法大多基于同质性假设，忽略了异质图的应用场景，导致在聚类过程中高频信息的丢失和聚类效果不佳。为解决此问题，提出了一种新颖的双通道属性图聚类方法（DCAGC）。该方法采用混合高斯模型预测节点连接的同质性，并基于这一预测构建同质和异质两种视图，以便从不同角度捕捉图中的低频和高频信息。同时，通过融合对比学习和聚类，实现了更精准的节点嵌入。与其他方法相比，DCAGC在处理异质图数据集时聚类效果显著，且具有较强的抗异常连接能力。

关键词：属性图聚类；自监督学习；异质图学习
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025009

Dual-channel attribute graph clustering beyond the homogeneity assumption

AN Junxiu, LIU Yuan, YANG Linwang
School of Software Engineering, Chengdu University of Information Technology, Chengdu 610207, China

Abstract: In recent years, significant progress has been made in the research of attribute graph clustering. However, existing methods are mostly based on the homogeneity assumption, thereby neglecting the application scenarios of heterogeneous graphs, leading to the loss of high-frequency information and poor clustering results during the clustering process. To address this issue, a novel dual-channel attribute graph clustering (DCAGC) method was proposed. A mixture of Gaussian models was used to predict the homogeneity of node connections and two views of homogeneous and heterogeneous were built, based on this prediction to capture low-frequency and high-frequency information in the graph from different perspectives. Simultaneously, by integrating contrastive learning and clustering, more precise node embeddings were achieved. Compared to other methods, DCAGC demonstrates significant clustering performance when handling heterogeneous graph datasets and exhibits strong resilience to anomalous connections.

Key words: attribute graph clustering, self-supervised learning, heterogeneous graph learning

收稿日期：2024−07−31；**修回日期：**2024−11−06
通信作者：柳源，liuyuan86@foxmail.com
基金项目：国家社会科学基金资助项目（No.22BXW048）
Foundation Item: The National Social Science Foundation of China (No.22BXW048)

0 引言

属性图聚类（attributed graph clustering）是图数据挖掘的一个重要方向，它不仅考虑图中节点的连接结构，还考虑节点属性。其核心目的是将图中的节点划分为若干个簇，确保同一簇内的节点在结构和属性上具有高度相似性。然而，在无标签环境下对属性图进行聚类，是一项不小的挑战[1]。近年来，图神经网络在属性图聚类任务中展现了卓越的性能[2]，这得益于其能够同时提取节点的结构和属性信息。大多数基于图神经网络的属性图聚类方法都建立在同质性假设的基础上。同质性假设认为，在图或网络中，相连节点往往具有相似的特征或属性[3]。以社交网络为例，如果两个人之间是朋友关系（即图中两节点通过一条边相连），则根据同质性假设，可推测他们拥有相似的兴趣、年龄或教育背景（即节点属性）。然而，现实生活中的图结构可能包含非同质连接，即属性不相似的节点之间也可能存在连接关系。同样以社交网络为例，兴趣或生活背景截然不同的两个人也可能因为工作需要而建立联系。因此，单纯依赖同质性假设进行属性图聚类可能会导致结果偏离真实情况。

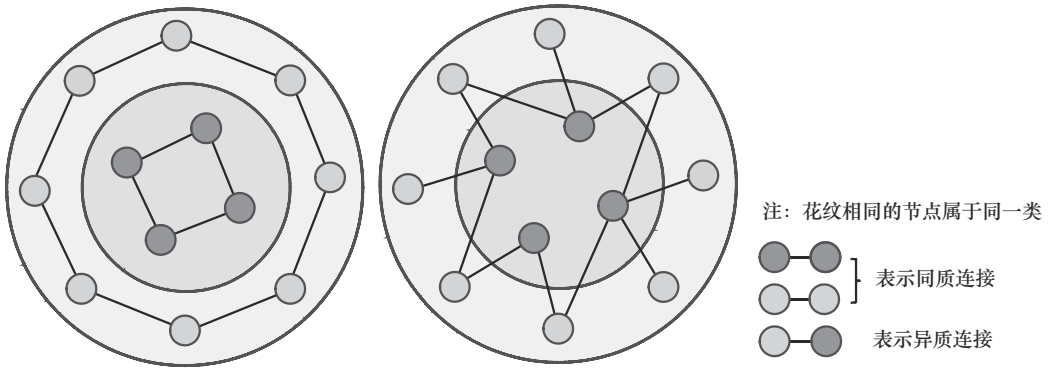
首先明确相关概念。连接属性相似节点的边通常被称为同质连接，而连接属性不相似节点的边则被称为异质连接。从整体层面上看，如果图中同质连接的数量超过异质连接，那么这样的图就被称为同质图；反之，如果异质连接数量占多数，则被称为异质图。同质图与异质图如图1所示。

同质性假设的局限性意味着基于该假设所建立的模型仅适用于同质图聚类任务，而对于异质图或混合图可能效果不佳。下面通过实验更直观地说明这一现象。

在图中，有些节点对是直接相连的，而有些是随机从图中抽取的。前者被称为直接相连节点对（connected node pairs，CNP），后者被称为随机抽样节点对（randomly sampled node pairs，RNP）。RNP的相似度分布反映了图中全部节点的同质性水平；而CNP的相似度分布则反映了图中相连节点的同质性水平，二者的相对大小能够从侧面反映图中连接关系的总体性质。

原始数据集和模型嵌入表示的相似度分布如图2所示。在具体的数据集验证中，本文发现：在同质图中，CNP的相似度明显高于RNP的相似度，同质图数据集Cora原始节点相似度分布如图2（a）所示；相反，在异质图中，CNP和RNP的相似度分布情况大致相同，异质图数据集Texas原始节点相似度分布如图2（b）所示。

进一步的，本文以经典的图神经网络方法GAE[6]为例，进行了嵌入表示的分析。在同质图数据集（如Cora）上，该方法能够很好地捕捉节点间的连接关系，使得相连节点的特征趋于一



（a）同质图　　（b）异质图

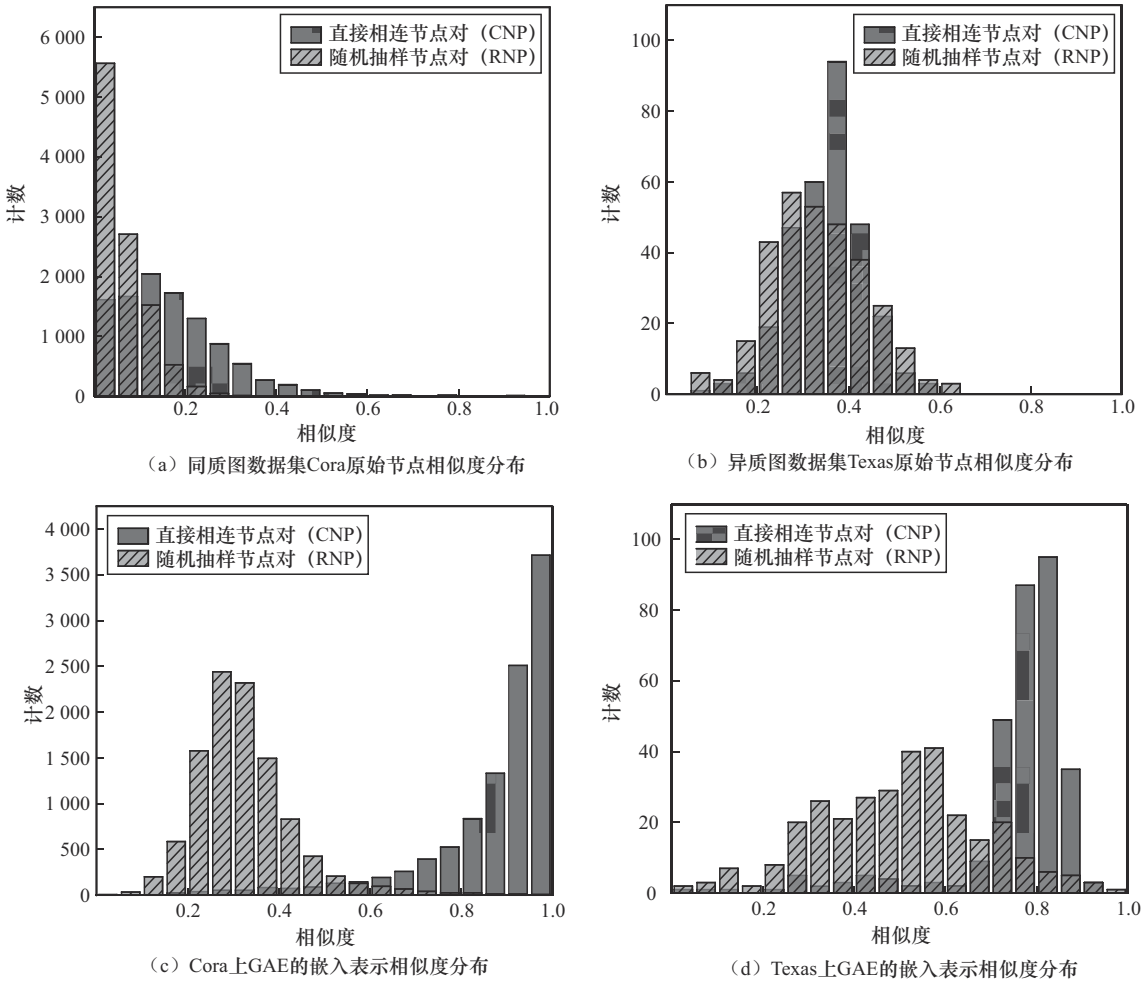
图1　同质图与异质图

致，这是符合预期的，Cora 上 GAE 的嵌入表示相似度分布如图 2（c）所示。但在异质图数据集（如 Texas）上，该方法可能导致相邻节点的属性特征过度融合，使得本该不相似的节点对变得相似，从而产生邻居节点同质化效应，Texas 上 GAE 的嵌入表示相似度分布如图 2（d）所示。

因此，在属性图表示学习的研究中，如何超越同质性假设的局限性并克服邻居节点同质化效应是一项很大的挑战。当前已有研究尝试解决这一问题，例如，通过拓展邻域结构来捕捉更高阶的节点关系信息[7-8]，或引入自适应滤波器来更好地适应图的异质性[9]。尽管这些方法提供了一些新的思路，但它们大多依赖于监督信号，这在无监督属性图聚类任务中受到了限制。

为了更有效地解决这个问题，需要聚焦于以下 3 个关键点：（1）在无监督学习环境中，如何准确区分同质连接和异质连接，这是实现无标签数据下节点关系识别的前提；（2）在（1）的基础上，如何同时捕获并融合节点的同质性和异质性关系特征，这需要构建一个高效的表示学习方法，该方法不仅能整合这两方面的信息，还要具备更强的特征表达能力，以便更全面地描述节点间的复杂关系；（3）需要探讨如何将前面学到的表示有效整合到一个深度聚类框架中，以实现更精准的节点分类和群组发现。

针对上述 3 个关键点，本文提出了一种超越同质性假设的双通道属性图聚类（dual-channel attribute graph clustering beyond the homogeneity



（a）同质图数据集Cora原始节点相似度分布

（b）异质图数据集Texas原始节点相似度分布

（c）Cora上GAE的嵌入表示相似度分布

（d）Texas上GAE的嵌入表示相似度分布

图2　原始数据集和模型嵌入表示的相似度分布

assumption，DCAGC）。首先，本文提出了一个基于高斯混合模型的视图构造方法，该方法能够在无监督的条件下预测连接的同质性水平，进而生成同质视图和异质视图，其目的是从不同的角度发现图中的同质信息和异质信息。其次，本文构建了一组分别用于同质视图和异质视图特征提取的双通道编码器，该编码器对同质视图和异质视图进行分离编码，确保特征提取的独立性，避免单通道特征提取带来的邻居节点同质化效应，导致图中的异质信息丢失。接下来，为了学习节点跨同质视图与异质视图之间的一致性表示并有效减少异常连接所引起的噪声，本文采用对比学习方法准确获取节点的嵌入表示。最后，本文联合优化特征表示学习任务和聚类任务，从高置信度的目标分布中迭代细化簇结构，以达到最终的聚类效果。

本文提出的DCAGC能够解决两个方面的实践问题：（1）在类似社交网络的异质图聚类分析中，用户之间的连接可能不仅仅基于相似兴趣或行为，还可能包含其他不同类型的关系，DCAGC有助于发现和区分这些不同类型的连接关系，从而更准确地聚类用户或社交群体，这一点在对比试验中有所体现；（2）在噪声图聚类分析中，图数据存在许多噪声连接或异常连接，这些连接可能会对聚类算法的性能产生负面影响，导致算法错误地将噪声点归类到某个簇中，从而影响最终的聚类结果，DCAGC可以有效地避免噪声连接带来的不良影响，提高聚类效果，这一点在后续的抗噪声连接分析中有所体现。

本文贡献可以总结为以下几点。

（1）提出一种新的属性图聚类方法，该方法超越了同质性假设的局限性，能够更好地适用于异质图和混合图的聚类任务。

（2）精心设计了基于高斯混合模型的视图构造器以及双通道图编码器，这一创新性的视图分离编码策略为异质图特征提取开辟了新的路径。

（3）本文所提模型巧妙地结合了对比学习方法，不仅实现了双视图的完美融合，同时显著提升了模型的鲁棒性，有效抵御了噪声连接的干扰，进而大幅提高了节点表示的精准度。

1 相关工作

深度属性图聚类近年来成为研究的热点，其核心目标是通过神经网络对节点进行编码，进而将它们划分为互不相连的群集。目前，主流的深度属性图聚类方法主要基于生成式模型架构。文献[6]提出了图自编码器（graph autoencoder，GAE），该方法通过最小化重建误差来学习图中节点的低维表征，并在此基础上实现聚类操作。继此之后，文献[10]引入了注意力机制，提出了深度注意力嵌入图聚类（deep attentional embedded graph clustering，DAEGC）模型，有效提升了聚类性能。文献[11]则提出了结构化深度聚类网络（structured deep clustering network，SDCN），该网络将图的拓扑结构融入深度聚类过程中。近年的研究趋势倾向于解决更具体的问题。文献[12]提出了细颗粒属性图聚类（fine-grained clustering，FGC）方法，旨在充分利用节点特征和结构信息。而文献[13]则设计了双重相关性减少网络（dual correlation reduction network，DCRN），通过降低对偶模型中的信息相关性，从而防止表征崩溃的发生。文献[14]采用先学习再融合的方式，分别学习多视图的共享表示与特定表示再进行融合，更细粒度地学习多视图的一致信息和互补信息，得到了更好的多视图属性图聚类效果。

对比学习[15-16]成为近期无监督学习研究的热点，在不同任务中取得了优异的表现[17-19]。在深度属性图聚类任务中，对比学习方法发挥出卓越的效果。为了保持网络的跨视图结构一致性，文献[20]提出了一种面向邻居的对比损失，而这一方法也极大限度地提升了聚类效率。与此类似，文献[21]采用多头图注意力机制自动

学习具有不同自适应拓扑邻接矩阵的增强视图。此外，文献[22]在文献[23]的基础上，将簇级对比损失引入图对比聚类，取得了不错的效果。文献[24]提出了基于聚类引导的硬样本挖掘策略，从而有效提高了多视图属性图聚类效果。尽管如此，无论是基于生成式模型的属性图聚类方法，还是基于对比学习的属性图聚类方法，它们大多基于同质性假设，仅适用于同质图，难以直接应用于异质图。本文所提方法在理论上跳出同质性假设，通过同质视图和异质视图两个角度挖掘聚类信息。

此外，与文献[20-24]类似，本文也是基于对比学习的属性图聚类方法，但最大不同在于样本的增强方式。现有方法通常通过引入噪声来增强样本，而本文则采用基于高斯混合模型的视图构造器有针对性地增强样本，并通过双通道编码器将增强效果融入嵌入特征中，从而提升网络的识别性能。

2　双通道属性图聚类

2.1　符号描述与问题定义

符号描述：定义图为 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，其中 $\mathcal{V}=\{v_1,v_2,\cdots,v_N\}$ 是可以被分成 K 类的 N 个节点的集合，$\mathcal{E}$ 表示边的集合，将连接节点 v_i 和 v_j 的边记作 $e_{i,j}$。使用 $\boldsymbol{X}\in\mathbb{R}^{N\times d}$ 表示图的属性矩阵，$\boldsymbol{A}\in\mathbb{R}^{N\times N}$ 表示图的邻接矩阵，其中 d 表示节点属性的维度，则无向图也可以用 $\mathcal{G}=(\boldsymbol{X},\boldsymbol{A})$ 表示。设度矩阵为 $\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\cdots,d_N)\in\mathbb{R}^{N\times N}$，其中 $d_i=\sum_{(v_i,v_j)\in\mathcal{E}}a_{ij}$，$a_{ij}$ 表示矩阵 $\boldsymbol{A}$ 的第 i 行第 j 列，将对称归一化的邻接矩阵记作 $\tilde{\boldsymbol{A}}=\hat{\boldsymbol{D}}^{-\left(\frac{1}{2}\right)}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\left(\frac{1}{2}\right)}$。将图拉普拉斯矩阵的定义为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$，那么对称归一化图的拉普拉斯矩阵表示为 $\tilde{\boldsymbol{L}}=\boldsymbol{I}-\hat{\boldsymbol{D}}^{-\left(\frac{1}{2}\right)}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\left(\frac{1}{2}\right)}$。

属性图聚类的形式化描述：属性图聚类的目标是将给定的 N 个未标记节点划分为 K 个互不相连的簇 $\{C_1,\cdots,C_k,\cdots,C_K\}$，使同一簇 C_k 中的节点彼此之间具有较高的相似性。

2.2　框架概述

DCAGC 模型框架如图 3 所示。首先，图数据由基于高斯混合模型的视图构造器进行处理，从而产生同质视图和异质视图，随后这两个视图分别进入低通编码器和高通编码器进行提取特征，以得到节点的嵌入表示。在嵌入表示上同时

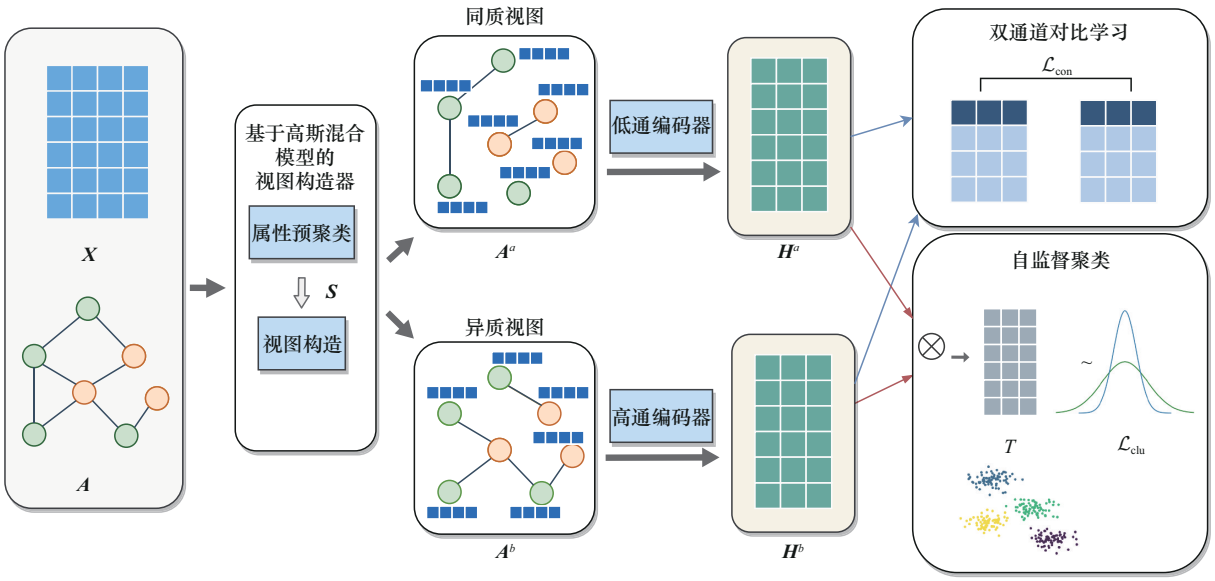


图 3　DCAGC 模型框架

优化对比学习目标和聚类目标，以获取节点聚类分配。

2.3 基于高斯混合模型的视图构造器

本文致力于通过构建同质视图与异质视图，以全面捕捉图的特征，从而防止高频信息的丢失。实现这一目标的核心在于对图中的同质连接与异质连接进行精确区分。而在无监督环境下，准确判断连接的同质性颇具挑战性，因为这需要对节点相似度进行精细化的比对。为解决这一问题，本文提出了一种简洁而高效的方法，以进行边的判别与视图构建。

本文方法先通过预聚类来确定每个节点属于各簇群的概率，即节点的软分配。这样，就能更准确地了解节点与各簇群的关系。然后，利用这些概率信息来计算边的权重，从而更容易地区分同质连接和异质连接，以便构建相应的视图。这种方法既提高了连接的识别准确度，也为构建视图提供了可靠基础。

具体来说，首先利用期望最大化算法[25]对属性特征$\boldsymbol{A}$进行高斯混合模型聚类，每个节点被分配到若干概率分布参数化的簇中，从而生成软分配矩阵$\boldsymbol{S}\in\mathbb{R}^{N\times K}$，其中$\boldsymbol{S}_{ik}$表示节点$v_i$分配到簇$k$的概率。对于每条边$e_{i,j}\in\mathcal{E}$所连接的两个节点$v_i$与$v_j$，二者同质的概率为：

$$\omega_{i,j}=\sum_{k=1}^{K}\boldsymbol{S}_{ik}\boldsymbol{S}_{jk} \tag{1}$$

该概率是基于节点v_i和节点v_j对应的簇概率分布的加权和。接下来对每条边进行权重再分配，构造同质视图的邻接矩阵$\boldsymbol{A}^a$和异质视图的邻接矩阵$\boldsymbol{A}^b$。

$$\boldsymbol{A}^a_{i,j}=\omega_{i,j},\boldsymbol{A}^b_{i,j}=1-\omega_{i,j},e_{i,j}\in\mathcal{E} \tag{2}$$

2.4 双通道编码器

现有大多数研究都采用了图神经网络对图结构数据进行编码。但如前文所述，图神经网络本质上是一个低通编码器，其平滑特性容易忽略异构视图中的高频信息（即异质连接），进而引发邻居同质化现象，严重影响模型的表达能力。为解决这一问题，本文针对同质视图和异质视图分别设计专门的编码器。

在同质视图中，节点间的连接代表着它们之间具有相似特征。因此，本文设计了一个低通编码器来捕捉并强化这些相似性信息，提取视图中的低频特征。低通编码器的表示如下：

$$\boldsymbol{H}^a_0=\mathrm{M}^a(\boldsymbol{X});\boldsymbol{H}^a_l=\tilde{\boldsymbol{A}}^a\boldsymbol{H}^a_{l-1},l=1,2,\cdots,L \tag{3}$$

其中，M^a是一个简单的多层感知机，$\tilde{\boldsymbol{A}}^a$是$\boldsymbol{A}^a$对称归一化的邻接矩阵，l是编码器的层数。

在异质视图中，边用于连接差异明显的节点，凸显节点间的不同。有研究表明，拉普拉斯矩阵在捕获高频信息方面表现出色[26]。为捕捉异质图中的高频信息并强化节点间的区别，本文基于拉普拉斯矩阵设计了高通编码器。高通编码器的表示如下：

$$\boldsymbol{H}^b_0=\mathrm{M}^b(\boldsymbol{X});\boldsymbol{H}^b_l=\left(\boldsymbol{I}-\alpha\tilde{\boldsymbol{A}}^b\right)\boldsymbol{H}^b_{l-1},l=1,2,\cdots,L \tag{4}$$

其中，M^b是一个简单的多层感知机，$\tilde{\boldsymbol{A}}^b$是$\boldsymbol{A}^b$对称归一化的邻接矩阵，l是编码器的层数。α是表示编码器捕获能力的参数。

最终对两个视图的嵌入表示进行融合，以得到节点嵌入表示：

$$\boldsymbol{H}=\frac{1}{2}\left(\boldsymbol{H}^a+\boldsymbol{H}^b\right) \tag{5}$$

2.5 基于同质性信息增强的双通道对比学习

对视图进行编码后，本文进一步结合了对比学习的思想，旨在使模型学习节点在不同视图间的一致性表示，并减少由错误边连接引入的噪声。对比学习的关键在于最大化正样本对之间的相似性，同时最小化负样本对之间的相似性。本文将同一节点在不同视图中的嵌入特征视为正样本对，其余视为负样本对。

为了获得特征的紧凑且具有区分性的表示，本文使用全连接投影网络M^{p}，将视图的潜在特征$\boldsymbol{H}^a$和$\boldsymbol{H}^b$投射到低维特征空间中，得到低维特

征$\boldsymbol{Z}^a=\mathrm{M}^{\mathrm{p}}\left(\boldsymbol{H}^a\right)$和$\boldsymbol{Z}^b=\mathrm{M}^{\mathrm{p}}\left(\boldsymbol{H}^b\right)$。然后通过计算余弦相似度来评估任意两个向量$\boldsymbol{z}_i^{k_1}$与$\boldsymbol{z}_j^{k_2}$之间的相似度程度：

$$s\left(\boldsymbol{z}_i^{k_1},\boldsymbol{z}_j^{k_2}\right)=\frac{\left(\boldsymbol{z}_i^{k_1}\right)\left(\boldsymbol{z}_j^{k_2}\right)^{\mathrm{T}}}{\left\|\boldsymbol{z}_i^{k_1}\right\|\left\|\boldsymbol{z}_j^{k_2}\right\|} \tag{6}$$

其中，$k_1,k_2\in\{a,b\}$，$i,j\in[1,N]$。在不失一般性的情况下，对于给定样本的损失为：

$$l_i^a=-\log\frac{\exp\left(s\left(\boldsymbol{z}_i^a,\boldsymbol{z}_i^b\right)/\tau_{\mathrm{C}}\right)}{\sum_{j=1}^{N}\left[\exp\left(s\left(\boldsymbol{z}_i^a,\boldsymbol{z}_j^a\right)/\tau_{\mathrm{C}}\right)+\exp\left(s\left(\boldsymbol{z}_i^a,\boldsymbol{z}_j^b\right)/\tau_{\mathrm{C}}\right)\right]} \tag{7}$$

其中，τ_{C}为温度超参数，用于控制相似度的敏感程度。最后，给出整体的对比损失函数：

$$\mathcal{L}_{\mathrm{con}}=\frac{1}{2N}\sum_{v=1}^{V}\left(l_i^a+l_i^b\right) \tag{8}$$

2.6　自监督聚类

图聚类本质上是一个无监督的任务，缺乏标签指导训练。为此，本文使用概率分布衍生的软标签作为一种自我监督机制进行聚类增强，从而有效地将聚类效果叠加在嵌入信息上。具体来说，首先需要获得嵌入点的软聚类分配概率。与现有的工作[27]类似，本文使用学生 t 分布[28]作为嵌入点$h_i\in H$的软分配：

$$q_{iu}=\frac{\left(1+\left\|h_i-\mu_u\right\|^2/\eta\right)^{-\frac{\eta+1}{2}}}{\sum_{u'}\left(1+\left\|h_i-\mu_{u'}\right\|^2/\eta\right)^{-\frac{\eta+1}{2}}} \tag{9}$$

其中，聚类中心μ_u由 K-means 对来自预训练的自编码器的嵌入进行初始化。η是自由度参数，在无监督学习中学习η是多余的[25]，因此在实验中设$\eta=1$。接下来定义具有高置信度的辅助分配p_{iu}：

$$p_{iu}=\frac{q_{iu}^2/\sum_i q_{iu}}{\sum_{u'}\left(q_{iu'}^2/\sum_i q_{iu'}\right)} \tag{10}$$

其中，$\sum_i q_{iu}$是质心u的软簇频率。为了提高集群凝聚度，最小化软分配和辅助分配之间的 KL 散度（Kullback-Leibler divergence）损失，迫使当前的软分配接近高置信度的辅助分配：

$$\mathcal{L}_{\mathrm{clu}}=\sum_i\sum_u p_{iu}\log\frac{p_{iu}}{q_{iu}} \tag{11}$$

KL 散度能够更加温和地更新模型，避免对比损失对嵌入的干扰，以更好地适应模型的整体优化目标。最后，取节点软分配最大值对应的簇作为聚类结果，节点i的聚类结果表示为：

$$Y_i=\mathrm{argmax}\left(q_{iu}\right) \tag{12}$$

2.7　优化策略

在 DCAGC 框架中，视图构造器模块独立运行，不参与优化过程。而其他模块则协同进行优化操作。DCAGC 的总体损失函数由两个部分组成，一是对比损失，二是聚类损失。该总损失函数表示如下：

$$\mathcal{L}=\mathcal{L}_{\mathrm{con}}+\mathcal{L}_{\mathrm{clu}} \tag{13}$$

DCAGC 采用标准的反向传播算法对损失函数进行优化。具体优化细节将在随后的实验部分予以详细描述。

2.8　时间复杂度

在数据预处理部分着重关注视图生成模块的时间复杂度。首先，通过预聚类确定每个节点属于各簇群的概率，这一过程的时间复杂度为$O(ink)$，其中i为迭代次数，n为节点数量，k为聚类数量。接着，进行权重计算，其时间复杂度为$O(mk)$，其中m为图中边的数量。

在双通道编码器中，编码器的时间复杂度为$O(ndd_e+md_eL)$，这里d_e表示最终节点嵌入表示的维度，L表示编码器的层数。

对于模型优化中使用的损失函数，对比损失

的时间复杂度为 $O(n^2 d_e)$，而聚类损失的时间复杂度为 $O(nd_e k)$。

3 实验设置

3.1 数据集

为评估DCAGC的有效性，在10个基准数据集上进行广泛实验，具体来说包含5个同质图数据集：Cora、CiteSeer[4]、AMAP[13]、EAT、BAT[29]，以及5个异质图数据集Texas、Cornell、Wisconsin、Washington[5]、Squirrel [30]。数据集统计信息见表1，其中同质率这项指标表示同质连接占所有连接的比例，该指标通过对每个数据集的计算得到。本文将同质率低于0.3的图视为异质图。

表1 数据集统计信息

分 类	数据集	样本数	维度	边数	类数	同质率
同质图	Cora	2 708	1 433	5 429	7	**0.81**
	CiteSeer	3 327	3 703	4 732	6	**0.74**
	AMAP	7 650	745	119 081	8	**0.83**
	EAT	399	203	5 994	4	**0.41**
	BAT	131	81	1 038	4	**0.51**
异质图	Texas	183	1 703	325	5	**0.06**
	Cornell	183	1 703	298	5	**0.12**
	Wisconsin	251	1 703	515	5	**0.17**
	Washington	230	1 703	786	5	**0.14**
	Squirrel	5 201	2 089	217 073	5	**0.22**

3.2 对比方法与评价指标

本文将DCAGC与传统方法以及近些年提出的一些先进的聚类方法进行比较，包括GAE[6]、DAEGC[10]、SUBLIME[31]、FGC[12]、DCRN[13]、NACL[21]、SCGC[20] 、MEFGC[32]，这些方法在相关研究中已经进行过简要描述。

为了评估本文提出的算法性能，选择两种常见的聚类算法评估指标，分别是聚类准确度（ACC）和归一化互信息（NMI），其值越高代表聚类效果越好。

3.3 实现细节

本文提出的DCAGC使用PyTorch平台实现，使用Adam算法对模型进行优化。所有实验均在一台配有Intel Core i7-8700 3.20 GHz CPU、GeForce RTX 3060 GPU 和32 GB RAM 的计算机上进行。学习率设置为0.000 1，超参数 τ_1 和 τ_2 分别设置为0.5和1.0。在DCAGC中使用卷积神经网络提取图像特征，使用全连接网络提取其他类型数据特征。

4 结果分析

4.1 聚类性能比较

同质图数据集上的对比实验结果见表2，详细罗列了DCAGC模型与其他基线模型在同质数据集上的聚类性能比较。本实验的主要目标是验证DCAGC对于处理同质图与异质图的通用性。结果显示，DCAGC在聚类性能上与近年来提出的其他模型相媲美，且在一半以上的数据集中取得了最佳聚类效果。总体来看，DCAGC在不同数据集上表现出较为稳定的聚类性能，反映了该模型具有较强的灵活性和适应性，适合于不同领域数据的聚类分析。此外，相较于基于生成式的方法（如GAE等），基于对比学习的方法在性能上普遍更优，这凸显了对比学习在特征表示方面的优越性。

异质图数据集上的对比实验结果见表3，展示了DCAGC与基线模型在异质图数据集上的聚类性能对比。可以很明显地看出DCAGC在处理异质图时的聚类性能显著优于其他模型，在所有数据集中均取得了最佳或次佳的效果。例如，在Texas数据集上，DCAGC的ACC指标比次优模型高出22.83%；在难度较高的Squirrel数据集上，比次优模型高出6.99%。这一性能优势的主要原因是其他模型在对异质图进行特征提取时丢失了关键的高频信息，从而导致错误的节点嵌入表示，而DCAGC通过在原始图数据基础上生成同

表2　同质图数据集上的对比实验结果

数据集	Cora		CiteSeer		AMAP		EAT		BAT	
评价指标	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE	44.27	30.55	60.4	33.55	71.42	61.39	45.59	16.10	54.69	31.39
DAEGC	71.80	51.94	66.71	38.10	54.26	41.83	30.77	12.45	56.62	36.26
SUBLIME	71.54	53.75	58.30	37.54	26.43	7.76	37.17	14.75	47.31	37.58
FGC	72.37	**55.70**	68.28	43.55	70.90	65.19	43.88	28.74	68.70	44.45
DCRN	46.71	26.62	69.74	44.84	**79.98**	**72.24**	53.30	11.34	42.96	27.85
NACL	53.86	31.66	58.17	35.91	66.90	63.17	37.72	21.14	46.32	23.17
SCGC	**72.95**	53.93	71.71	44.81	76.79	67.86	**59.46**	34.87	**78.48**	53.89
MEFGC	69.24	52.89	71.93	50.68	65.31	63.8	40.34	23.71	75.24	40.01
DCAGC	72.51	53.19	**72.47**	**65.20**	72.37	71.46	58.53	**35.35**	77.78	**62.35**

表3　异质图数据集上的对比实验结果

数据集	Texas		Cornell		Wisconsin		Washington		Squirrel	
评价指标	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE	37.65	11.37	47.73	23.83	47.43	16.72	42.65	22.13	21.91	4.56
DAEGC	45.16	12.21	40.43	10.78	41.39	17.54	40.85	21.57	26.71	1.57
SUBLIME	55.81	11.80	51.92	17.69	53.28	19.69	44.56	15.27	29.20	9.42
FGC	56.76	9.41	44.42	9.51	49.48	11.30	54.41	22.14	24.48	3.57
DCRN	55.99	16.59	59.43	19.32	51.82	9.26	61.17	17.36	28.89	4.62
NACL	58.21	16.90	55.84	16.90	64.31	14.34	64.38	25.22	25.34	3.19
SCGC	49.29	13.18	41.10	11.32	53.48	20.22	49.59	22.96	23.8	6.52
MEFGC	57.24	14.42	47.28	9.63	66.83	19.04	**65.12**	**30.91**	22.32	4.97
DCAGC	**71.50**	**33.29**	**66.72**	**27.55**	**72.65**	**39.91**	63.14	21.72	**31.24**	**12.65**

质视图和异质视图，并分别采用低通编码器和高通编码器进行编码，有效地保留了更多图结构信息。

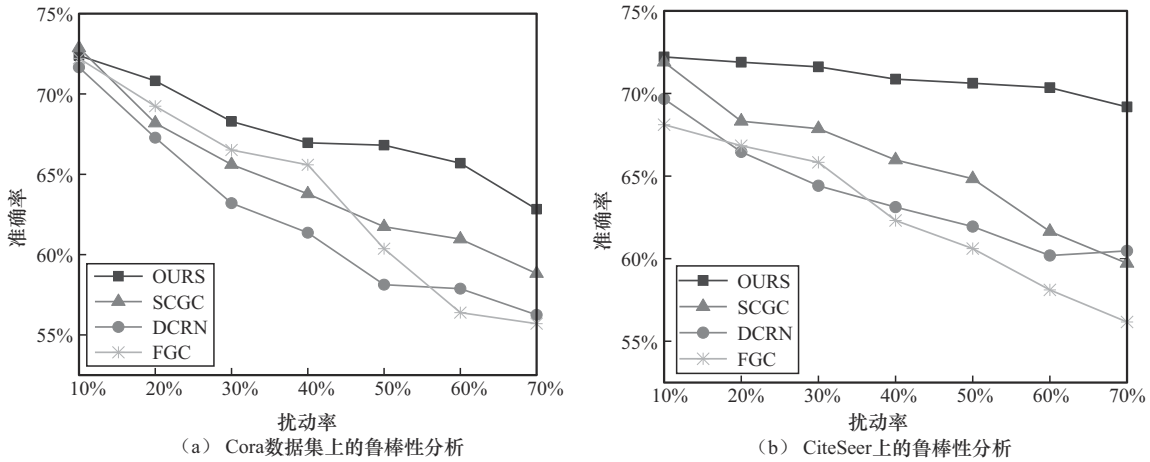
综上所述，DCAGC不仅在同质图数据集上表现出稳定的性能，在异质图数据集上亦展现出突出的聚类效果。这超越了同质性假设的局限，本文将在后续实验中进一步证明DCAGC的这一优势。

4.2　抗噪声连接分析

本节实验通过向原始图中随机添加噪声边对图的结构进行扰动，以此来考察模型的鲁棒性，研究了不同扰动率下模型的聚类性能。DCAGC在不同数据集上的鲁棒性分析如图4所示。从图4展示的结果来看，DCAGC模型的聚类性能在总体上优于其他比较模型。值得注意的是，随着扰动率的提高，DCAGC的优势愈加明显，特别是在CiteSeer数据集上进行的实验中，DCAGC展现出了较高的稳定性。本实验结果表明，DCAGC对于图结构中的噪声干扰有着良好的抑制作用，证实了模型在面对结构扰动时具有较强的鲁棒性。

4.3　聚类可视化分析

为直观展示DCAGC的优势，本节通过t-SNE技术[33]对DCAGC学习的嵌入表示进行可视化分析，并且以GAE作为参考进行对比。聚类可视化分析如图5所示。结果表明，DCAGC能够得到更加准确的分配结果与紧凑的聚类结构，这主要得益于基于对比学习的特征表示学习和在嵌入表示上进行的聚类结构优化。



（a） Cora数据集上的鲁棒性分析　　（b） CiteSeer上的鲁棒性分析

图4　DCAGC在不同数据集上的鲁棒性分析

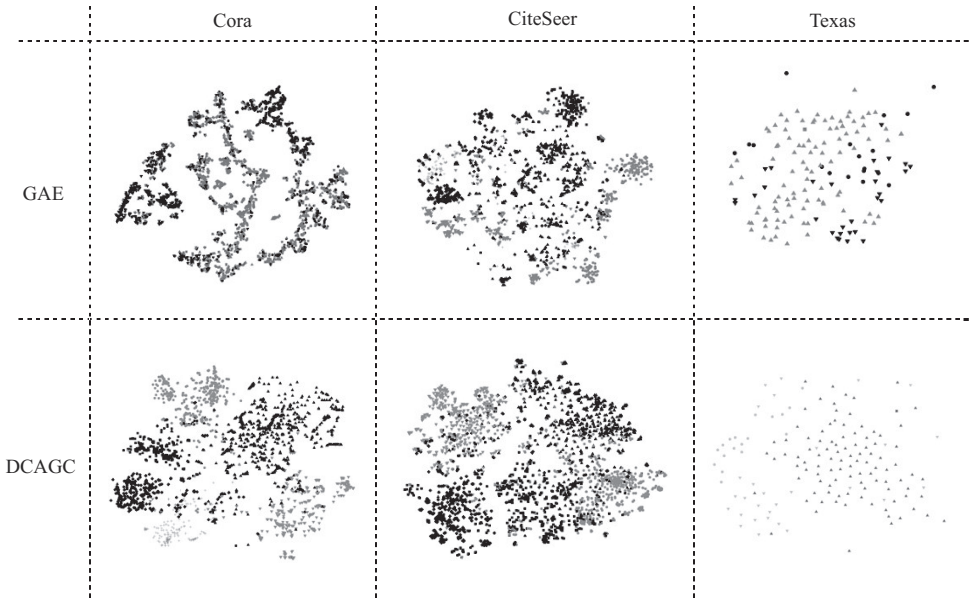
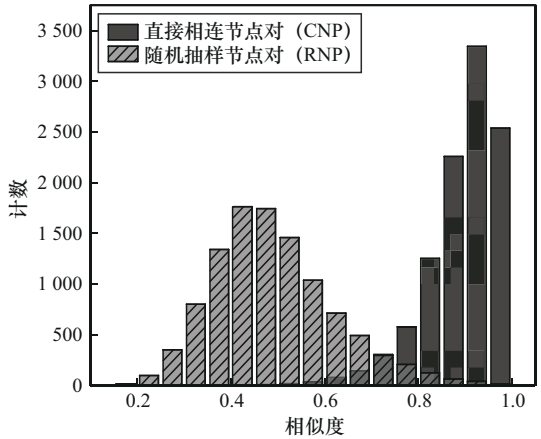


图5　聚类可视化分析

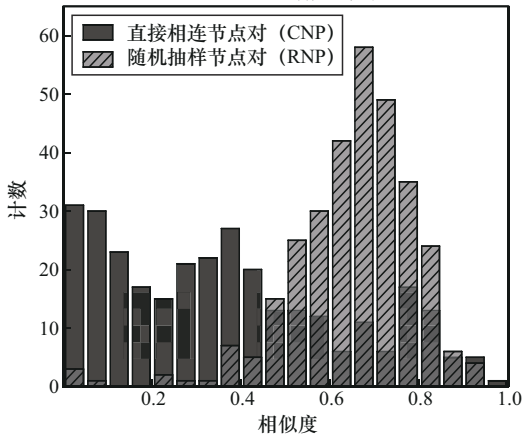
4.4　嵌入表示相似度分布分析

在前文中，针对数据集特征的相似度分布进行了初步分析，并阐明了当前大多数基于图神经网络（GNN）的图表示学习模型在处理非同质图数据时所遭遇的挑战。这些现有模型往往倾向于产生过度平滑的节点特征表示，其后果在异质图中尤为显著：模型倾向于将邻接节点的特征表示同质化，从而导致错误的特征表示被学习。

DCAGC嵌入表示的相似度分布如图6所示。为了更清楚地分析DCAGC针对这一问题的优化，图6（b）展示了基于非同质数据集Texas上对DCAGC的嵌入表示相似度分布进行的分析。实验结果表明，与其他模型相比，DCAGC能够倾向性地保留图中的高频成分，有效缓解了相邻节点特征同质化的问题，并更为准确地学习了异质图的节点特征表示。如图6（a）所示，在同质数据集Cora上，DCAGC也展现了学习平滑特征表示的能力。这一能力主要归因于视图构造器对边缘同质性的动态识别能力。



（a） Cora上的嵌入表示



（b）Texas上的嵌入表示

图6　DCAGC嵌入表示的相似度分布

4.5　消融实验

为了验证DCAGC各损失项的有效性以及不同功能模块对整体性能的贡献，本文进行了一系列消融实验，消融实验结果见表4。

表4　消融实验结果

模型	Cora	CiteSeer	AMAP	Texas	Cornell	Wisconsin
DCAGC	72.51	72.47	69.37	71.5	66.72	72.65
DCAGC–$\mathcal{L}_{clu}$	68.27	69.84	63.83	65.68	61.80	52.80
DCAGC–$\mathcal{L}_{con}$	42.14	39.66	29.22	45.80	38.22	40.48
DCAGC－VC	71.99	71.47	67.73	63.91	55.66	61.88
DCAGC－DE	71.14	70.51	66.97	62.94	56.58	64.40

在损失项的消融实验中，首先从DCAGC的总损失函数中移除聚类损失项$\mathcal{L}_{clu}$（DCAGC－$\mathcal{L}_{clu}$），然后移除对比学习损失项$\mathcal{L}_{con}$（DCAGC－$\mathcal{L}_{con}$），以评估这些损失项的功效。实验结果表明，移除$\mathcal{L}_{con}$对模型的聚类性能产生了显著的负面影响，这证实了$\mathcal{L}_{con}$在表示学习中的关键作用，它是模型能否获得有效特征表示的决定性因素。而$\mathcal{L}_{clu}$能够指导模型形成更加紧密的簇结构。

在功能模块的消融实验中，首先去除了视图构造器模块。为了保证模型仍能进行对比学习，本文采用了文献[22]中提出的方法对原始数据进行扰动，生成两个差异性视图（DCAGC－VC）。接着去除了双通编码器，改为使用两个不共享参数的图神经网络（DCAGC－DE）。实验结果显示，视图构造器模块和双通编码器模块对模型的聚类性能均有积极影响。特别是在异质图数据集Texas、Cornell和Wisconsin上，缺失这些模块会导致聚类性能显著下降，这进一步证实了DCAGC在处理异质数据集方面的结构优势。

4.6　参数敏感性研究和收敛分析

为确保模型达到性能最佳，本节对参数进行分析。对于本模型的优化目标，式（8）中的温度参数τ_C会对对比学习的效果产生直接影响，其参数敏感性分析如图7所示。可见参数τ_C的敏感性较低，当τ_C处于0.4和0.7之间时，模型聚类效果达到最优。

5　结束语

本文主要针对在属性图聚类领域同质性假设的局限性进行讨论，提出了合理的解决方法，设计出双通道对比属性图聚类模型，并通过实验验证了其出色的聚类性能。此外，本文所提方法具有良好的可扩展性，在自监督学习过程中所学习的节点嵌入表达能够应用于各类下游任务。

在未来的研究中：（1）将DCAGC运用于社交网络分析中，解决实际的社会学分析问题，弥补传统聚类方法在社交网络分析中的不足；（2）设计可训练的视图构造器，将视图构造器的优化融入模型整体的优化过程，通过无监督学习从特

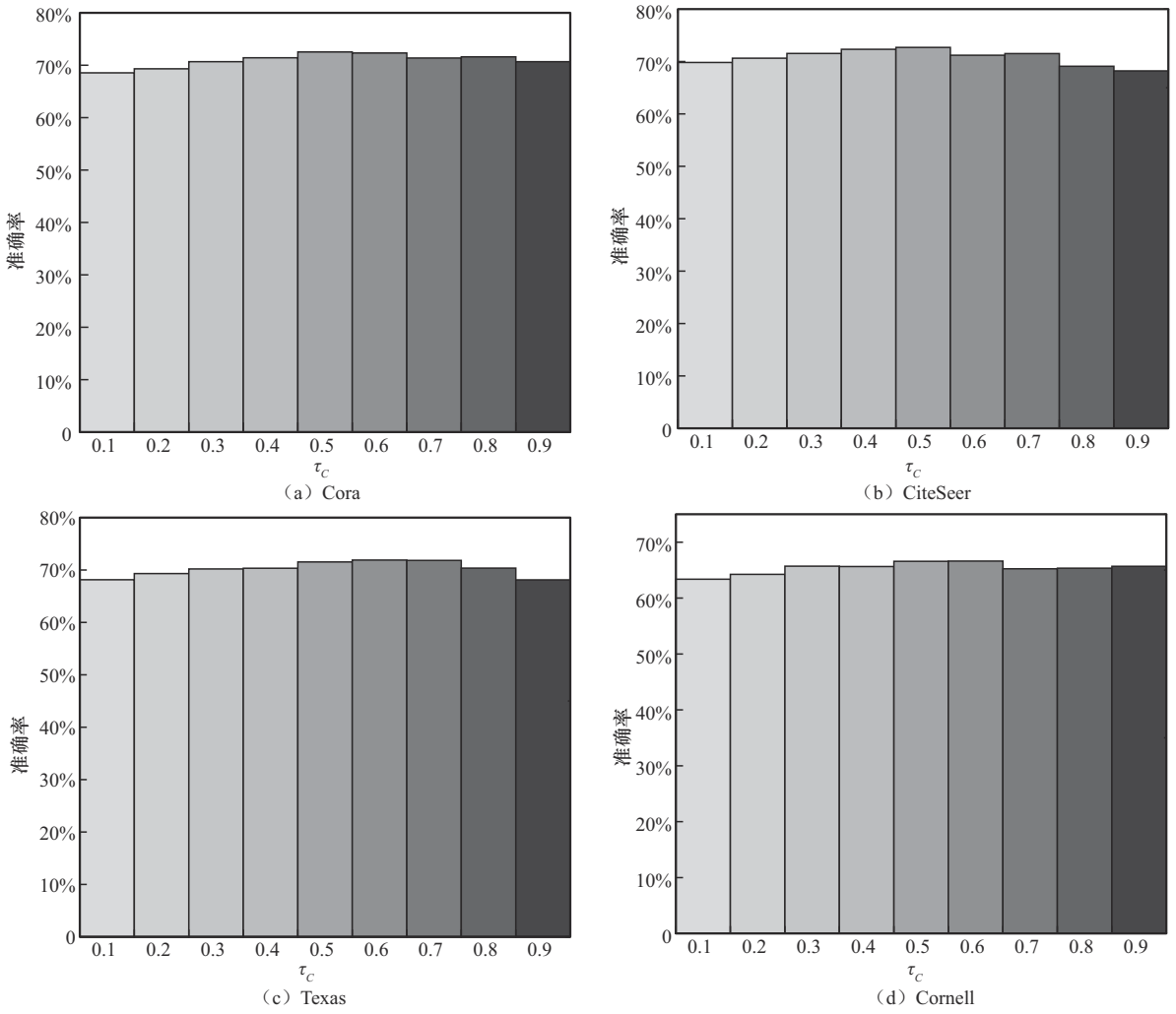


图 7　参数敏感性分析

征和结构信息中推断出边缘同质性；（3）考虑运用基于密度的聚类方法，根据数据特征自动确定最佳聚类数量，避免预设的聚类数量限制；（4）使用时间效率更高的采样方法，提升模型效率，使其能够处理大规模图数据集。

参考文献：

[1] NGUYEN V H, SUGIYAMA K, NAKOV P, et al. FANG: leveraging social context for fake news detection using graph representation[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020: 1165-1174.

[2] FAN S H, WANG X, SHI C, et al. One2Multi graph autoencoder for multi-view graph clustering[C]//Proceedings of the Web Conference 2020. New York: ACM Press, 2020: 3070-3076.

[3] FANG R Y, WEN L J, KANG Z, et al. Structure-preserving graph representation learning[C]//Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2022: 927-932.

[4] KULATILLEKE G K, PORTMANN M, CHANDRA S S. SCGC: self-supervised contrastive graph clustering[EB]. 2022: 2204.12656.

[5] PEI H B, WEI B Z, CHANG K C C, et al. Geom-GCN: geometric graph convolutional networks[EB]. 2020: 2002.05287.

[6] KIPF T N, WELLING M. Variational graph auto-encoders [EB]. 2016: 1611.07308.

[7] DEVVRIT F, SINHA A, DHILLON I, et al. S3GC: scalable self-supervised graph clustering[J]. Advances in Neural Information Processing Systems, 2022, 35: 3248-3261.

[8] WANG Y, DERR T. Tree decomposed graph neural network [C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM

Press, 2021: 2040-2049.

[9] WANG X, ZHANG M. How powerful are spectral graph neural networks[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2022: 23341-23362.

[10] WANG C, PAN S R, HU R Q, et al. Attributed graph clustering: a deep attentional embedding approach[EB]. 2019: 1906.06532.

[11] BO D, WANG X, SHI C, et al. Structural deep clustering network[C]//Proceedings of the Web Conference 2020. New York: ACM Press, 2020: 1400-1410.

[12] KANG Z, LIU Z Y, PAN S R, et al. Fine-grained attributed graph clustering[C]//Proceedings of the 2022 SIAM International Conference on Data Mining (SDM). Philadelphia: Society for Industrial and Applied Mathematics, 2022: 370-378.

[13] LIU Y, TU W X, ZHOU S H, et al. Deep graph clustering via dual correlation reduction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022, 36 (7): 7603-7611.

[14] 曹付元, 陈晓惠. 共享和特定表示的多视图属性图聚类[J]. 软件学报, 2024: 1-14.
CAO F Y, CHEN X H. Multi-view attributed graph clustering based on shared and specific representation[J]. Journal of Software, 2024: 1-14.

[15] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2020: 1597-1607.

[16] 岑科廷, 沈华伟, 曹旖, 等. 图对比学习综述[J]. 中文信息学报, 2023, 37(5): 1-21.
CEN K T, SHEN H W, CAO Q, et al. A survey on graph contrastive learning[J]. Journal of Chinese Information Processing, 2023, 37(5): 1-21.

[17] 尹梦冉, 梁美玉, 于洋, 等. 面向跨模态检索的查询感知双重对比学习网络[J]. 软件学报, 2024, 35(5): 2120-2132.
YIN M R, LIANG M Y, YU Y, et al. Query aware dual contrastive learning network for cross-modal retrieval[J]. Journal of Software, 2024, 35(5): 2120-2132.

[18] 张颖, 张冰冰, 董微, 等. 基于语言-视觉对比学习的多模态视频行为识别方法[J]. 自动化学报, 2024, 50(2): 417-430.
ZHANG Y, ZHANG B B, DONG W, et al. Multi-modal video action recognition method based on language-visual contrastive learning[J]. Acta Automatica Sinica, 2024, 50(2): 417-430.

[19] 柳源, 安俊秀, 杨林旺. 多角度语义标签引导的自监督多视图聚类[J]. 计算机应用研究, 2024: 1-8.
LIU Y, AN J X, YANG L W. Multi-view clustering with self-supervised learning guided by multi-angle semantic labels[J]. Application Research of Computers, 2024: 1-8.

[20] LIU Y, YANG X H, ZHOU S H, et al. Simple contrastive graph clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 13789-13800.

[21] SHEN X, SUN D W, PAN S R, et al. Neighbor contrastive learning on learnable graph augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023, 37(8): 9782-9791.

[22] XIA W, WANG Q Q, GAO Q X, et al. Self-consistent contrastive attributed graph clustering with pseudo-label prompt[J]. IEEE Transactions on Multimedia, 2023, 25: 6665-6677.

[23] LI Y F, HU P, LIU Z T, et al. Contrastive clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press 2021, 35(10): 8547-8555.

[24] 钱李烽, 李静, 邹徐熹, 等. 基于双重对比学习和硬样本挖掘的多视图图聚类算法[J]. 计算机工程, 2024: 1-16.
QIANG L F, LI J, ZOU X X, et al. Multi-view graph clustering algorithm based on dual contrastive learning and hard sample mining[J]. Computer Engineering, 2024: 1-16.

[25] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1977, 39(1): 1-22.

[26] DONG Y S, DING K Z, JALAIAN B, et al. AdaGNN: graph neural networks with adaptive frequency response filter[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2021: 392-401.

[27] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//International conference on machine learning. Piscataway: IEEE Press, 2016: 478-487.

[28] MAATEN L, POSTMA E, HERIK J. Dimensionality reduction: a comparative review[J]. Journal of Machine Learning Research, 2009, 10(66-71): 13.

[29] MRABAH N, BOUGUESSA M, TOUATI M F, et al. Rethinking graph auto-encoder models for attributed graph clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(9): 9037-9053.

[30] ROZEMBERCZKI B, ALLEN C, SARKAR R. Multi-scale attributed node embedding[J]. Journal of Complex Networks, 2021, 9(2): cnab014.

[31] LIU Y X, ZHENG Y, ZHANG D K, et al. Towards unsupervised deep graph structure learning[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 1392-1403.

[32] LIU H T, LU X B, CHENG K F, et al. A multi-embedding fusion network for attributed graph clustering[J]. Applied Soft Computing, 2024, 165: 112073.

[33] VAN D M L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2589-2605.

[作者简介]

安俊秀（1970− ），女，成都信息工程大学软件工程学院教授，主要研究方向为云计算与大数据技术、人工智能。

柳源（1999− ），男，成都信息工程大学软件工程学院硕士生，主要研究方向为深度聚类、对比学习。

杨林旺（2000− ），男，成都信息工程大学软件工程学院硕士生，主要研究方向为深度聚类、数据挖掘、自然语言处理。