基于混合聚类的多级别加权图对比学习

王笛苹 刘海洋 原继东 李方静
（北京交通大学计算机科学与技术学院 北京 100044）

摘 要 图对比学习无需依赖数据标签，能够从未标注的图数据中自动学习有价值的信息。其核心思想是通过对比图数据的不同视图来学习样本的表示。然而，现有研究大多仅基于节点特征或拓扑结构的单一角度生成对比视图，限制了对图结构数据的全面探索，单一角度的感知使模型容易忽略特征相似但非直接连接的节点或对边连接密集的节点过拟合，进而影响其在下游任务上的性能表现。此外，现有方法在进行视图对比时，无论是同级别还是跨级别的对比，通常忽视了负样本的差异性，即模型无法有效区分节点中的假负样本以及群组中不同规模负样本对全局语义的影响，从而在学习过程中出现表示偏差，影响下游任务上的性能。为了解决上述问题，我们提出了基于混合聚类的多级别加权图对比学习（multi-level weighted graph contrastive learning via hybrid clustering，HCWGC）框架。具体来说，它基于两种互补型聚类方法，分别从结构和特征两个角度生成对比视图，既能从边连接稠密性的角度挖掘局部显式结构，又能从特征相似性的角度挖掘全局隐式结构，进而全面捕获图结构数据中的复杂信息。随后，设计了多级别加权对比策略计算对比损失，在节点-节点级对比中，设计节点相似等级指标以识别假负样本，减少其噪声干扰；在节点-群组级对比中，设计群组局部强度指标以考虑不同规模的群组对全局语义的影响，使规模较大的群组负样本在对比过程中发挥更大的作用。实验结果表明，HCWGC 在多个图数据集上均展现出优越的性能。与最优基线模型相比，在节点分类任务的 Micro-$F1$ 指标上最高相对提升达到 0.45%，Macro-$F1$ 指标上最高相对提升达到 0.57%，在节点聚类任务的 NMI 指标上最高相对提升达到 4.77%，在链接预测任务部分数据集的 AP 指标上最高相对提升了 0.1%。

关键词 自监督学习；图表示学习；对比学习；混合聚类；负样本加权
中图法分类号 TP183 **DOI 号** 10.11897/SP.J.1016.2026.01075

Multi-Level Weighted Graph Contrastive Learning via Hybrid Clustering

WANG Di-Ping LIU Hai-Yang YUAN Ji-Dong LI Fang-Jing
（School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044）

Abstract Graph contrastive learning has emerged as a powerful paradigm for learning informative representations from unlabeled graph-structured data, eliminating the need for manual annotation. Its core idea is to learn meaningful representations by contrasting different views of the same graph data. However, most existing studies generate contrastive views from a single perspective, either node features or graph topology, which limits a comprehensive exploration of graph-structured data. This uni-perspective perception makes models prone to overlooking nodes that share similar features but are not directly connected, or to overfitting on nodes in densely connected regions, thereby degrading performance in downstream tasks. Furthermore, in existing methods, whether performing intra-scale or inter-scale comparisons, the differences between negative samples are often overlooked. As a result, models struggle to effectively distinguish false negatives within other node samples and to capture the impact of negative samples from

收稿日期：2025-04-25；在线发布日期：2025-11-12。本课题得到中央高校基本科研业务费专项资金（2023JBZY035）资助。**王笛苹**，硕士研究生，主要研究领域为数据挖掘、图神经网络。E-mail：dipingwang@bjtu.edu.cn。**刘海洋**，博士，讲师，中国计算机学会（CCF）会员，主要研究领域为数据挖掘、图神经网络。**原继东**（**通信作者**），博士，副教授，中国计算机学会（CCF）会员，主要研究领域为数据挖掘、时间序列分类。E-mail：yuanjd@bjtu.edu.cn。**李方静**，博士研究生，主要研究领域为数据挖掘、图神经网络。

groups of varying sizes on the global semantics. This leads to representation bias during the learning process, which also affects performance on downstream tasks. To address these limitations, we propose a novel framework called multi-level weighted graph contrastive learning via hybrid clustering (HCWGC). This framework integrates two complementary clustering strategies to construct diverse contrastive views from both structural and feature perspectives. For feature clustering, we adopt the k-means algorithm to exploit feature similarity and uncover global implicit structures. For structural clustering, we design a method called strength weighted structural clustering, which uses local strength information to optimize the degree to which each node belongs to each group, so that the resulting memberships reflect the density of the topological structure and thus make it possible to mine local explicit structures in the graph. By combining these two perspectives, HCWGC captures complex and multi-faceted information embedded in graph-structured data more effectively. Building upon these dual views, we design a multi-level weighted contrastive strategy to compute the contrastive loss. At the node-node level, we introduce a similarity-aware weighting mechanism that leverages the clustering results from both structural and feature perspectives to estimate node similarity rankings. Based on these rankings, potential false negative samples are identified, thereby reducing the noise introduced by such samples during contrastive learning. At the node-group level, a group locality strength metric is proposed to account for the influence of groups with varying sizes on global semantics, allowing larger groups to play a more significant role during contrastive learning. We conduct extensive experiments on multiple benchmark graph datasets and comprehensively evaluate HCWGC on three typical downstream tasks, including node classification, node clustering, and link prediction. The empirical results on these benchmark datasets consistently demonstrate that HCWGC achieves superior performance over a range of state-of-the-art graph representation learning methods. Compared with the best baseline models, HCWGC achieves relative improvements of up to 0.45% in Micro-$F1$ scores and up to 0.57% in Macro-$F1$ scores in node classification tasks. In node clustering tasks, the model yields relative gains of up to 4.77% in NMI scores. In link prediction tasks, it achieves relative improvements of up to 0.1% in AP scores on certain datasets. These results highlight the effectiveness of combining hybrid clustering with the multi-level weighted contrastive strategy for comprehensive graph representation learning.

Keywords self-supervised learning; graph representation learning; contrastive learning; hybrid clustering; negative sample weighting

1 引 言

图结构数据能够反映现实世界中的许多复杂关系,例如,在社交网络中,它能够反映个体间的社交关系[1];在金融领域,它能够反映客户间的交易关系[2];在生物领域,它能够反映微观原子关联关系[3]。为了生成能够准确捕捉图结构和特征的图表示向量,更好地挖掘图结构数据中的信息,近年来,图表示学习方法(graph representation learning,GRL)得到了广泛应用。随着图神经网络(graph neural networks,GNN)的发展,GNN 已经成为求解图表示的有效手段[4-9]。大多数基于 GNN 的图表示学习方法采用监督学习范式,其性能高度依赖于有标签数据。然而在部分实际应用场景中,获取足够数量且高质量的标注数据往往成本高昂[10]。

随着图对比学习(graph contrastive learning,GCL)的飞速发展,上述问题得到了缓解。GCL 是一种自监督的学习方法,通过对比图数据的不同视图来学习样本的表示,能够摆脱训练过程对数据标签的依赖,从未标注的图数据中自动学习有价值的信息[11]。然而,现有 GCL 方法大多仅从单一角度获取对比视图,且在计算对比损失时忽视了负样本之间的差异性。为了在局部或全局范围内聚合相似

的语义信息,一些研究通过聚类方法生成对比视图,例如:模型 GIC[12] 和 GRCCA[13] 侧重从节点特征的角度进行聚类,而模型 gCooL[14] 和 CAGDCL[15] 侧重于从拓扑结构的角度进行聚类,以生成群组级视图。其中,群组通常定义为在拓扑结构上具有高度密集连接或在特征空间中具有高度相似属性的节点集合。然而,不同角度的信息聚合会有不同的关注点,仅从单一的角度聚类生成对比视图存在一定局限性。如图 1 所示,结构角度的聚合更关注拓扑结构中的密集性连接,形成内部边连接稠密的群组,但容易将拓扑连接不紧密的同类节点排除在外;而特征角度的聚合更关注节点特征向量的相似性,能够捕捉未直接相连的潜在同类节点,但对图显式拓扑结构的感知不足。此外,现有 GCL 方法在利用对比视图计算对比损失时,无论是同级别还是跨级别的对比,通常忽视了负样本之间的差异性。对于节点级负样本,常见的选择策略[14,16-17] 会不可避免地引入假负样本,若不对这些样本进行区分,可能会引入大量噪声;对于群组级负样本,不同规模群组对全局语义的影响并不相同[18],若在计算对比损失时对所有群组负样本采用统一权重,则可能导致模型学习出现表示偏差。

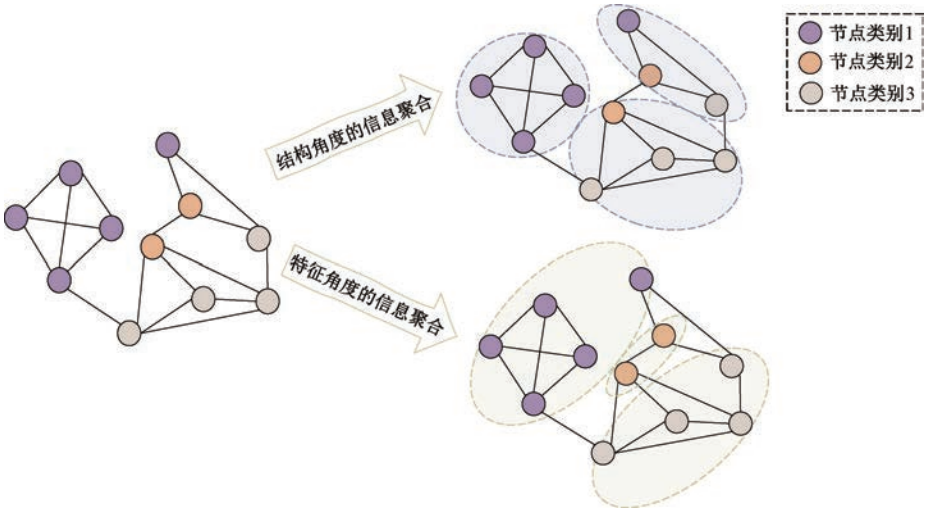


图 1　不同角度的信息聚合

针对上述问题,我们提出了基于混合聚类的多级别加权图对比学习框架(multi-level weighted graph contrastive learning via hybrid clustering, HCWGC)。通过同时使用两种互补型聚类方法生成对比视图,具体而言,使用 k-means 方法从特征角度对节点进行聚类,并设计了强度优化的结构聚类(strength-weighted structural clustering,SWSC)方法从显式拓扑结构的角度对节点进行聚类,解决了仅从单一角度探索图数据存在局限性的问题。基于构建的对比视图,提出了多级别加权对比策略以优化对比损失的计算。该策略引入相似等级和局部强度指标以分别指导对节点负样本和群组负样本的加权。最后,结构聚类损失与对比损失被加权融合,共同优化模型的训练过程。

总的来说,我们的贡献可以概括如下:

(1)提出了一种新颖的基于混合聚类的图对比学习框架 HCWGC。为了挖掘图数据中全局和局部的拓扑关系,我们基于特征和结构两种聚类方法形成对比视图。其中,对于特征聚类我们使用了 k-means 方法,能够挖掘全局隐式结构;对于结构聚类我们设计了 SWSC,利用局部强度优化节点对群组的隶属度,衡量了拓扑结构的密集性,能够挖掘局部显式结构。将得到的互补型视图应用于对比学习,能够深入挖掘节点间的关联性,全面捕获图结构数据中的复杂信息。

(2)在对比损失的计算上,我们设计了基于相似等级的节点-节点级加权对比,用以区分假负样本,从而减少其带来的噪声影响;同时设计了基于群组局部强度的节点-群组跨级别加权对比,考虑了不同的群组对全局语义的影响,从而减少学到的表示偏差。我们将 SWSC 损失与上述对比损失进行加权运算,使模型在不同训练阶段调整侧重点,提高图表示学习的精度。

(3)我们在节点分类、节点聚类和链接预测任务上验证了提出模型的有效性。实验结果表明,与现有主流模型相比,HCWGC 在多个数据集上都展现出了更优越的性能。

2 相关工作

在本节中，我们回顾了相关工作，包括图对比学习和图聚类方法。

2.1 图对比学习

图对比学习是对比学习在图结构数据上的扩展应用，希望通过不同视图的对比让相似的节点（图）学习到相似的表示，不相似的节点（图）学习到差异较大的表示。这些视图按照尺度可以划分为节点级视图、群组级视图和图级视图。基于对比视图的尺度，对比模式可以进一步分为尺度内对比和尺度间对比。

尺度内对比侧重于对比相同尺度的视图，例如：GRACE[16]采用节点-节点级对比，基于结构和属性两个方面创建节点级视图，最大化相同节点在不同视图中的一致性；GCA[17]在 GRACE 的基础上提出了自适应的图增强方法，并沿用了其对比模式；GraphCL[19]使用 4 种不同的图增强方法生成多样的图级视图，从而在图级视图间进行对比。上述尺度内对比方法能够在相同尺度下捕捉正样本对间的相似性和负样本对间的差异性，但可能忽略不同尺度视图之间的信息关联，影响总体语义信息的获取。

尺度间对比侧重于对比不同尺度的视图，例如：DGI[20]提出了节点-图级的跨级别对比方式，使用两个编码器来获得图表示和节点表示，并最大化二者之间的互信息；MVGRL[21]同样采用节点-图级对比，使用图扩散来创建图的增广结构视图，并使用鉴别器对比不同视图中的节点表示和图表示；GIC[12]和 GRCCA[13]利用 k-means 方法对节点表示进行聚类，得到群组级视图，与节点级视图构成节点-群组级对比；gCooL[14]同样采用节点-群组级对比，但是其群组视图是通过显式拓扑结构的密集性划分生成的。上述尺度间对比方法虽然有效整合了多尺度视图的信息，但它们对个体节点的差异性不再敏锐，而且大多数群组级对比视图都是从单一的角度生成的，未能实现图数据的全面探索。

无论是尺度内对比还是尺度间对比，负样本对的构建始终是对比学习中的关键环节。现有方法中，在节点级别，DGI 和 HDMI[22]通过对节点属性随机洗牌来产生负样本节点，GIC、GRCCA 和 gCooL 通过在锚点节点所属增强视图中采样其他节点来生成负样本，然而此类采样方式容易引入假负样本。针对这一问题，GCNSS[23]利用分类预测结果辅助负样本筛选，但该方法高度依赖分类器的预测准确性，容易导致新的偏差；E2Neg[24]通过谱聚类选取群组中心并进行子图拓扑重构，从结构角度出发过滤假负样本；FD4GCL[25]则从属性与结构两个维度检测假负样本，并尝试将其转化为正样本，但存在过度矫正的风险。在群组级别，负样本大多从锚点所属群组之外的其他群组中采样，然而现有方法[12-14]往往忽略了群组规模与密集程度对群组语义的潜在影响。图级别的负样本对则通常通过锚图与其他图的增强视图进行构建。本文主要关注节点级和群组级的负样本构建与加权。

2.2 图聚类

图聚类旨在以无监督的方式将图中的节点划分至不同的群组。许多传统的图聚类方式例如 k-means、谱聚类[26]、随机块模型（stochastic block model，SBM）[27]通常只能利用既得的节点表示或固有拓扑结构进行聚类。尽管这些方法有时也能取得不错的效果，但由于节点表示与聚类结果不能以端到端的方式共同优化，可能会造成性能次优[28]。随着图神经网络的不断发展，基于 GNN 的图聚类方法很好地弥补了这一不足。RGCC[29]引入拉普拉斯约束，自适应地学习图结构，并且设计了自监督的聚类模块以更好地指导节点表示学习，使其适应聚类任务；文献[30]提出了一种基于对比学习的去偏框架，能够同时进行表示学习和聚类；CDBNE[31]使用图注意力机制对拓扑结构和节点特征进行编码，并通过最大化模块度来获得聚类群组；CDC[32]通过图过滤融合结构与特征信息，并借助相似性保持正则化方法自适应学习高质量锚点；DeCA[14]利用拓扑结构的密集性进行聚类，并同时学习节点表示。尽管这些方法通过图神经网络实现了聚类和表示学习的联合优化，但鲜少考虑聚类群组规模对节点划分产生的影响，这可能会导致聚类结果出现偏差。

3 预备知识

在本节中，我们首先给出了本文中用到的符号表示和问题定义，随后介绍了模块度的基本概念。

3.1 符号表示及问题定义

设 $G=(\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{E}})$ 表示一张无向图，其中 $\boldsymbol{\mathcal{V}}=\{v_1,v_2,\cdots,v_N\}$ 代表节点的集合，$\boldsymbol{\mathcal{E}}\subseteq\boldsymbol{\mathcal{V}}\times\boldsymbol{\mathcal{V}}$ 代表边的集合，N 代表节点个数。边反映了节点间的连接关系，在实际应用中，我们通常用邻接矩阵 $\boldsymbol{A}\in\{0,$

$1\}^{N\times N}$ 来直观反映一张图中所有节点之间的连接关系。$\boldsymbol{A}_{ij}=1$ 代表节点 i 与节点 j 之间存在连边,反之则代表这两个节点间不存在连边。我们定义节点特征矩阵为 $\boldsymbol{X}\in\mathbb{R}^{N\times F}$,其中,$F$ 代表节点特征维数,$\boldsymbol{x}_i$ 代表节点 v_i 的特征向量。

在本文中,我们研究的重点是无监督的图表示学习,目标是学习一个图编码器 $f_\theta:\mathbb{R}^{N\times F}\times\mathbb{R}^{N\times N}\rightarrow\mathbb{R}^{N\times F'}$,仅依赖节点特征矩阵 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}$ 生成低维节点嵌入,从而应用于不同的下游任务。

3.2　模块度

在图对比学习中,为了增强模型对图结构的理解和表示能力,常常利用聚类方法获取图中的密集群组,从而更好地捕捉节点之间的潜在关系。利用群组级表示进行对比,能够减轻节点级对比带来的噪声影响,帮助模型获取更广泛的语义信息。

模块度[33]是用来衡量群组划分质量的经典评价指标[34],如公式(1)所示:

$$Q=\frac{1}{2m}\sum_{ij}\left(\boldsymbol{A}_{ij}-\frac{k_ik_j}{2m}\right)\delta(i,j)\qquad(1)$$

其中,k_i 和 k_j 分别代表节点 i 和节点 j 的度数,m 代表图中的总边数,定义为 $2m=\sum_i k_i$。如果边是随机分布的,那么节点 i 和节点 j 之间的期望边数为 $k_ik_j/2m$。$\delta(i,j)$ 作为条件因子,当节点 i 和节点 j 处于同一群组时取值为 1,处于不同群组时取值为 0,能够确保只在两个节点属于同一个群组的情况下进行求和。因此,模块度可以被视为群组内部实际边数的比例与随机情况下群组内部期望边数的比例之间的差异,其取值范围为$(-1/2,1]$,数值越大,代表群组内部的连接越紧密,群组划分的质量越好。

4　提出的方法

在这一部分,我们对提出的 HCWGC 进行详细介绍。HCWGC 整体架构如图 2 所示。为了增加图数据的多样性,同时使模型忽略由微小扰动引起的非语义相关变化,转而聚焦于视图间共享的语义不变性,我们对原图进行数据增强。将图增强策略 T 定义为随机丢弃边(randomly dropping edges,RDE)和随机隐藏特征(randomly masking features,RMF),通过对原始图施加这些扰动,生成两个增强图 $\widetilde{G}^U$ 和 $\widetilde{G}^V$。随后,将两个增强图输入到共享参数的编码器 f_θ 中进行编码,该编码器由 GCN 和 MLP 组成,经过编码后,得到相应的节点表示 $\boldsymbol{H}^U,\boldsymbol{H}^V\in\mathbb{R}^{N\times F'}$。将节点表示和增强图的拓扑结构分别经过两种互补型聚类方法 SWSC 和 k-means,将视图 U 和 V 中的节点划分出不同群组,并得到群组表示矩阵 $\boldsymbol{C}^U,\boldsymbol{C}^V\in\mathbb{R}^{K\times F'}$,图中相同颜色的节点隶属于相同群组。随后进行多级别加权对比,在节点-节点级对比中,我们根据上述两种聚类方法得到的群组划分结果计算锚点与其他节点的相似等级,并以此筛选假负样本,减少假负样本在对比损失中所占权重;在节点-群组级对比中,我们对各群组计算局部强度,以衡量不同群组语义对于全局语义的影响,并为局部强度更高的群组赋予更大的加权项,图中虚线的粗细程度反映了权值的大小。最后,我们将结构聚类损失、节点-节点级对比损失和节点-群组级对比损失三部分结合,计算总体损失函数,并更新编码器参数。

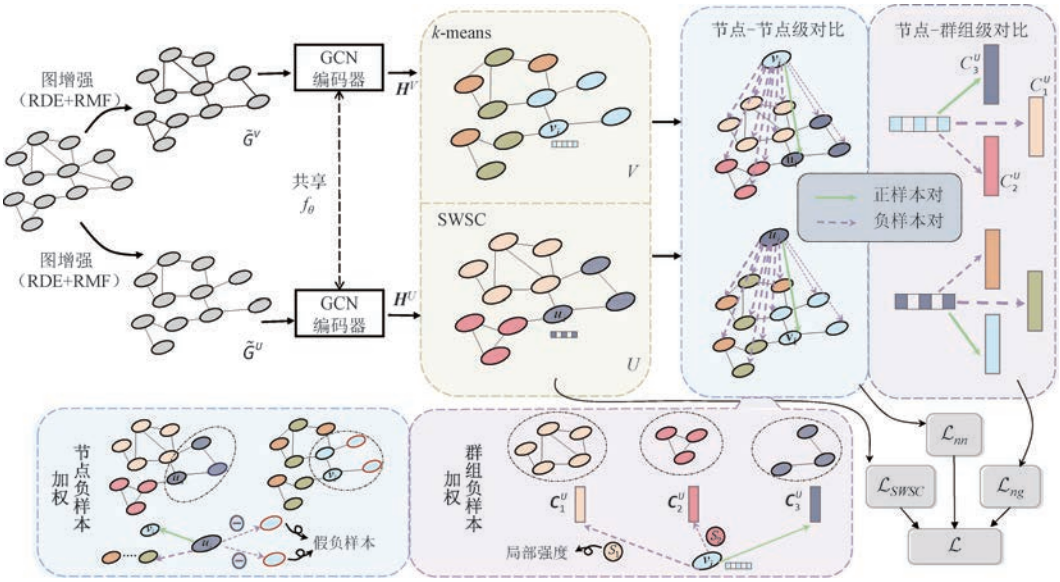


图 2　HCWGC 整体架构

4.1 强度优化的结构聚类

对比学习旨在通过对比原始数据的不同视图来学习样本的表示，我们通常希望每个视图都能有效地反映图属性，且不同视图能够保留图中的不同语义信息。因此，本文在进行图对比学习时选择使用两种侧重点不同的聚类方法生成对比视图。与其他类型的数据相比，图结构数据的显著特点在于其不仅包含特征信息，还具备显式的拓扑结构。在本文中，我们采用 k-means 方法，基于节点表示进行聚类，这样能够在全局范围内考虑节点的相似性，更多地从节点特征的角度获取语义信息。此外，根据图结构数据的特殊性，我们还需要一种聚类方法，更多地从拓扑结构的角度获取语义信息，与 k-means 方法形成互补，从而更加全面地捕获图数据中的复杂信息。

结构聚类通过聚合拓扑上密集连接的节点形成多个群组，且不同群组根据规模和密集程度对全局语义的影响不同。一般的结构聚类方法通常仅依据群组内部连通的紧密性来划分节点，忽视了群组规模对节点归属的影响，将这样的聚类结果应用于图对比学习中容易出现表示偏差。为了解决这一问题，我们提出了一种强度优化的结构聚类方法，综合考虑了节点表示相似性、群组规模和群组内部连接紧密性来确定节点归属。我们通过随机初始化生成群组中心矩阵 $\boldsymbol{C}^U$，并采用端到端的方式进行训练，并在训练过程中作为可学习参数参与反向传播，与编码器参数同步更新。

首先，我们利用节点表示的相似性对节点进行软划分，能够得到节点对于群组的隶属度矩阵 $\boldsymbol{M}\in\mathbb{R}^{N\times K}$，$K$ 代表群组的数量，具体计算如公式(2)所示：

$$\boldsymbol{M}_{ik}=normalize(sim(\boldsymbol{h}_i,\boldsymbol{C}_k^U)) \tag{2}$$

其中，$\boldsymbol{h}_i$ 为 v_i 经过编码器得到的节点表示，$\boldsymbol{C}_k^U$ 为第 k 个群组的聚类中心，$sim(\boldsymbol{h}_i,\boldsymbol{C}_k^U)=\exp(\boldsymbol{h}_i(\boldsymbol{C}_k^U)^T/\|\boldsymbol{h}_i\|\cdot\|\boldsymbol{C}_k^U\|\cdot\tau)$为指数余弦相似度函数，$\tau$ 为温度参数。将计算出的指数余弦相似度经过归一化函数 $normalize(\cdot)$处理，能够得到节点 v_i 隶属于第 k 个聚类群组的概率 $\boldsymbol{M}_{ik}$，且对于每个节点 v_i，都有 $\sum_{k=1}^{K}\boldsymbol{M}_{ik}=1$。

大多数结构聚类方法通常忽视了群组规模对节点隶属度的影响，它们假设节点归属于群组的概率仅取决于连接关系，而与群组的大小无关。然而，较大的群组通常包含更多的节点和连边，这意味着节点对于较大规模的群组的隶属度可能更高，因为它们在这个群组中可能有更多直接邻居或特征相似的节点。为此，我们定义了群组的局部强度 $\boldsymbol{S}\in\mathbb{R}^K$ 来描述群组规模，计算公式如下：

$$\boldsymbol{S}_k=\alpha\frac{|\mathcal{E}_k|}{|\mathcal{E}|}+(1-\alpha)\frac{N_k}{N} \tag{3}$$

其中，$\boldsymbol{S}_k$ 代表第 k 个群组的局部强度。局部强度由群组内边强度和群组内节点强度两部分组成。其中，$|\mathcal{E}_k|/|\mathcal{E}|$用于计算群组内边数占图中总边数的比例，$N_k/N$ 用于计算群组内节点数占图中总节点数的比例。通过采用相对比例而非绝对数量，能够提高局部强度在不同类型图结构(稀疏图和稠密图)中的普适性与鲁棒性。α 是超参数，用于权衡边和节点对局部强度的影响。

随后，我们通过广播机制沿行方向对局部强度行向量 $\boldsymbol{S}$ 进行维度扩展得到 $\boldsymbol{S}'\in\mathbb{R}^{N\times K}$，并对隶属度 $\boldsymbol{M}$ 进行如下优化：

$$\boldsymbol{M}'=\boldsymbol{M}+\beta\boldsymbol{S}' \tag{4}$$

其中，β 是超参数，用于权衡局部强度对隶属度的影响。优化后的隶属度矩阵 $\boldsymbol{M}'$充分考虑了群组规模对节点归属的影响，有助于减轻图对比学习中的表示偏差。

为了使结构聚类的结果兼具紧凑性和可分性，我们采用如下公式计算损失。

$$\mathcal{L}_{SWSC}=\underbrace{\frac{\lambda}{N(N-1)}\sum_{i,j}\sum_{k_1\neq k_2}\boldsymbol{A}_{ij}\boldsymbol{M}'_{ik_1}\boldsymbol{M}'_{jk_2}}_{inter}-\underbrace{\frac{1}{N}\sum_{i,j}\sum_{k}[\boldsymbol{A}_{ij}-ced(k)]\boldsymbol{M}'_{ik}\boldsymbol{M}'_{jk}}_{intra} \tag{5}$$

其中，λ 为系数，$ced(k)=2|\mathcal{E}_k|/(N_k(N_k-1))$为群组内边密度函数，衡量了一个群组内部节点之间的连通程度，具体表示为群组内实际存在的边数与最大可能边数之比，我们在附录 A 中针对其单调性和有界性给出证明。

相较于公式(1)，我们不再使用非 1 即 0 的方式判断某一节点是否属于某一群组，而是通过隶属度来表示节点属于某一群组的概率，从而更平滑地处理群组之间的划分界限。此外，我们不仅考虑了提高群组内部的连通性，还希望降低群组之间的连通性。$\boldsymbol{A}_{ij}-ced(k)$表示每条边的实际密度与期望密度之间的差值。为了减少计算开销，我们在实际运算中使用了 $ced(k)$的上界作为简化处理，最终的矩阵计算形式见附录 B。我们期望聚类结果具有以下特性：群组间节点应当具有较强的可分性，即群组间节点之间的期望边密度应接近于 0；同时，群组内节

点应当具有较高的紧凑性,即我们希望群组内节点的实际边密度高于期望边密度。算法 1 对 SWSC 聚类过程进行了描述:

算法 1. SWSC 聚类过程

输入:节点表示 $\boldsymbol{H}^U$

输出:隶属度矩阵 $\boldsymbol{M}'$,群组中心矩阵 $\boldsymbol{C}^U$

1. 随机初始化生成群组中心矩阵 $\boldsymbol{C}^U$;

2. 根据公式(2)对节点进行软划分,得到节点对于群组的隶属度矩阵 $\boldsymbol{M}$;

3. 根据公式(3)计算群组局部强度 $\boldsymbol{S}$;

4. 对 $\boldsymbol{S}$ 进行维度扩展得到 $\boldsymbol{S}'$;

5. 根据公式(4)对 $\boldsymbol{M}$ 进行优化,得到优化后的隶属度矩阵 $\boldsymbol{M}'$;

6. 根据公式(5)计算 SWSC 聚类损失;

7. RETURN $\boldsymbol{M}'$, $\boldsymbol{C}^U$;

4.2 多级别加权对比

经过两种互补型聚类后,我们基于得到的视图 U、V 进行多级别加权对比。我们设计了节点-节点级对比和节点-群组级对比,以更好地融合多尺度信息。在计算对比损失时,依据相似等级对节点负样本进行加权,以便更好地区分假负样本,减少噪声引入;利用局部强度对群组负样本赋予加权项,以便缓解 GCL 中的表示偏差。与现有的多级别对比方法[14,17,20-21]相比,本文提出的对比视图源自两种互补型聚类结果,能够从特征和结构两个角度更全面地捕获图数据的语义信息;此外,相较于已有的负样本加权方法[15],我们针对不同级别的负样本分别设计了精细化的加权策略,更有效地区分了不同负样本的语义表达。

4.2.1 节点-节点级对比

在进行节点级对比时,我们选取不同视图的同一节点为正样本对,不同视图的不同节点为负样本对。以图 2 为例,以节点 u_i 为锚点,节点 v_i 为其正样本,视图 V 中的其他节点为其负样本。现有的多数图对比学习方法对所有的正负样本施以同等的权重,这种方法虽然简单,但可能会引入过多假负样本,从而在学习节点表示的过程中引入过多噪声[35]。为了缓解这一问题,我们提出了基于相似等级的节点负样本加权方法。

首先,我们根据两种不同聚类方法的结果计算节点相似等级,并根据相似等级区分假负样本。我们将图中每个节点经过 SWSC 和 k-means 聚类后所属群组分别标记为 $\boldsymbol{R}^U$,$\boldsymbol{R}^V \in \mathbb{R}^n$,且 $\boldsymbol{R}_i^U = \mathrm{argmax}_{1 \leqslant j \leqslant K} \boldsymbol{M}_{ij}'$。若 $\boldsymbol{R}_i^U = \boldsymbol{R}_j^U$($\boldsymbol{R}_i^V = \boldsymbol{R}_j^V$),则代表在视图 U(V)中,节点 v_i 与节点 v_j 被划分至相同群组。我们将节点 v_i 与节点 v_j 的相似等级 $\boldsymbol{SR}_{ij}$ 定义为它们在不同视图的聚类过程中被划分至同一个群组的次数:

$$\boldsymbol{SR}_{ij} = \begin{cases} 2, \boldsymbol{R}_i^U = \boldsymbol{R}_j^U \wedge \boldsymbol{R}_i^V = \boldsymbol{R}_j^V \\ 1, \boldsymbol{R}_i^U = \boldsymbol{R}_j^U \oplus \boldsymbol{R}_i^V = \boldsymbol{R}_j^V \\ 0, \boldsymbol{R}_i^U \neq \boldsymbol{R}_j^U \wedge \boldsymbol{R}_i^V \neq \boldsymbol{R}_j^V \end{cases} \tag{6}$$

相似等级能够有效地刻画节点间的相似程度,等级越高,节点之间的相似性越强。具体来说,若节点 v_i 与节点 v_j 在两个视图的不同聚类方法中均被归为同一群组,则相似等级为 2;若仅在其中一个视图中被归为同组,则相似等级为 1;若在两个视图中均处于不同群组,则相似等级为 0。我们将与锚点相似等级为 2 的负样本定义为强相似负样本。由于假负样本通常与锚点具有高度相似的特征,并且属于同一类别,因此我们定义强相似负样本为假负样本。

随后,由于假负样本会在一定程度上对模型性能造成不利影响,我们在计算对比损失时对负样本进行重新加权:

$$\mathcal{L}_{nn} = -\log \sum_i \frac{sim(\boldsymbol{h}_i^U, \boldsymbol{h}_i^V)}{\sum\limits_{(j \neq i) \wedge SR_{ij}=2} \omega sim(\boldsymbol{h}_i^U, \boldsymbol{h}_j^V) + \sum\limits_{SR_{ij} \neq 2} sim(\boldsymbol{h}_i^U, \boldsymbol{h}_j^V)} \tag{7}$$

其中,$\omega \in [0,1]$为控制强相似负样本的权重参数。在这个损失函数中,强相似负样本对训练过程的影响将被整体弱化,能够有效缓解假负样本的噪声引入造成的不利影响。

与现有的节点负样本加权方法相比,我们对于假负样本的筛选不再仅仅依赖于简单的节点相似度计算或单次聚类的结果,而是提出一种更稳健、更可靠的策略,即基于混合聚类的综合结果计算节点相似等级。通过综合特征和结构角度聚类所获得的节点聚合情况,能够有效缓解单一聚类结果存在的随机性和偶然性问题以及节点相似度随消息传递过程递增导致的判别能力下降问题,从而更加准确地识别假负样本。

4.2.2 节点-群组级对比

为了整合多尺度的语义信息,我们在节点-节点级对比的基础上,进一步利用节点表示和群组表示进行跨级别对比。我们定义节点(锚点)与其在另一个视图中所属群组构成正样本对,与另一视图中其他所有群组构成负样本对。以图 2 中的节点 u_i 为锚点,在视图 V 中,令其对应的相同节点 v_i 所属的群组为其正样本,所有其他群组为其负样本。这种

对比方式通过最大化两种尺度之间的互信息，有利于同时捕获局部和全局信息。但现有的跨级别对比方法并未考虑到不同规模的群组会对全局语义产生不同影响，这可能会导致 GCL 中出现额外的表示偏差。针对上述问题，我们提出了基于局部强度的群组负样本加权方法。

经过 SWSC 和 k-means 两种聚类方法，我们在不同视图中划分出多个群组，并以各自的聚类中心作为群组的表示，随后以如下公式计算跨级别对比损失：

$$l_{ng}(\boldsymbol{H}^U,\boldsymbol{C}^V)=-\log\sum_i\frac{sim(\boldsymbol{h}_i^U,\boldsymbol{C}_{k_i}^V)}{sim(\boldsymbol{h}_i^U,\boldsymbol{C}_{k_i}^V)+\sum_{k_i\neq k}(sim(\boldsymbol{h}_i^U,\boldsymbol{C}_k^V)+\gamma\boldsymbol{S}_k)} \tag{8}$$

其中，第 i 个节点被划分至第 k_i 个群组。在这个损失函数中，我们利用局部强度为负样本对添加加权项，参数 $\gamma\in[0,1]$ 控制了负样本的群组局部强度对全局语义的影响，使群组规模更大的负样本在对比学习的过程中起到更大的作用，从而减轻学习产生的表示偏差。

我们综合利用两种视图计算跨级别损失：

$$\mathcal{L}_{ng}=\frac{1}{2}(l_{ng}(\boldsymbol{H}^U,\boldsymbol{C}^V)+l_{ng}(\boldsymbol{H}^V,\boldsymbol{C}^U)) \tag{9}$$

与现有的跨级别对比方法相比，我们更深入地考虑到不同规模的群组对全局语义产生的不同影响，基于局部强度对群组负样本施加权重，使局部强度更大的群组负样本在语义学习过程中发挥更显著的作用，以便减少学习的表示偏差。

4.3 模型训练

综合上述损失，我们得到了总体损失函数，具体形式如公式 (10)所示。基于此损失函数，我们对模型进行训练，以优化其性能。

$$\mathcal{L}=\mathcal{L}_{nn}+\psi(e)\mu(e)\,\mathcal{L}_{SWSC}+\psi(e)(1-\mu(e))\,\mathcal{L}_{ng} \tag{10}$$

其中，$\psi(e)=1-e/\varepsilon\eta$ 和 $\mu(e)=\exp\{-e/\eta\}$ 是衰减系数，ε 和 η 是超参数，e 为训练轮数。模型在训练初期应综合考虑跨级别粗粒度对比学习和节点-节点级细粒度对比学习，但随着训练的深入，结构聚类会逐步稳定，模型需要更专注于节点表示的细粒度对比学习。因此，我们利用 $\psi(e)$ 逐渐减少对结构聚类和跨级别损失的关注。而 $\mu(e)$ 进一步平衡了结构聚类损失和跨级别损失，使模型在早期充分训练结构聚类划分，随后平滑地将重心转向跨级别对比学习。这样的衰减策略能够帮助模型在不同训练阶段逐步调整关注重点，有效地学习多层次信息，最终提升模型的整体性能。算法 2 对整体学习算法进行了总结：

算法 2. HCWGC 学习算法

输入：图 $G=\{\boldsymbol{X},\boldsymbol{A}\}$

输出：节点表示 $\boldsymbol{H}$

1. FOR $epoch=1,2,3,\cdots$ DO
2. 向原图 G 施加增强策略 T，得到增强图 $\widetilde{G}^U$ 和 $\widetilde{G}^V$；
3. 利用共享的编码器 f_θ 对 $\widetilde{G}^U$ 和 $\widetilde{G}^V$ 分别编码，得到节点表示 $\boldsymbol{H}^U$ 和 $\boldsymbol{H}^V$；
4. 对 $\boldsymbol{H}^U$ 应用 SWSC 聚类方法得到群组表示 $\boldsymbol{C}^U$；
5. 对 $\boldsymbol{H}^V$ 应用 k-means 聚类得到群组表示 $\boldsymbol{C}^V$；
6. 利用公式(10)计算总体损失函数 $\mathcal{L}$；
7. 基于 $\mathcal{L}$，采用随机梯度下降方法优化参数 θ；
8. END FOR
9. 计算节点表示 $\boldsymbol{H}=f_\theta(G)$；
10. RETURN $\boldsymbol{H}$；

4.4 时间复杂度分析

我们分别从编码器、聚类算法和对比损失三个部分对模型时间复杂度进行分析。编码器部分由 GCN 和 MLP 共同组成，其时间复杂度为 $\mathcal{O}(Ed+Nd^2)$，其中 E 为图中边的数量，N 为节点的数量，d 为节点表示的维度。聚类算法包含 k-means 和 SWSC 两种方法，对于 k-means 方法，时间复杂度为 $\mathcal{O}(NKd)$，K 为聚类群组的个数；对于 SWSC 方法，时间复杂度为 $\mathcal{O}(NKd+EK)$。对比损失部分由节点-节点级对比损失和节点-群组级对比损失组成，时间复杂度为 $\mathcal{O}(N^2d+NKd)$。综上，模型的整体时间复杂度为 $\mathcal{O}(N^2d+Nd^2+Ed+EK+NKd)$。

5 实 验

在这一部分中，我们通过实验评估了 HCWGC 相比于其他基准模型的有效性。首先介绍了实验的基本设置；随后分别在节点分类、节点聚类和链接预测任务上评估模型的整体性能，并在节点聚类、链接预测和图分类等任务上进行了消融实验以验证所提出方法的有效性；最后，我们进行了参数分析实验。

5.1 实验设置

5.1.1 数据集

为了全面评估模型的整体性能,本研究选取了4个图数据集(Amazon-Photo、Amazon-Computers、Coauthor-CS和WikiCS)用于节点分类、节点聚类和链接预测任务,并使用MUTAG与PROTEINS两个图数据集进行图分类任务上的消融实验。数据集的详细统计信息如表1所示,其中节点数和边数为单图平均规模。

表1　数据集信息统计

数据集	图数	节点数	边数	特征维度	类别数
Amazon-Photo	1	7487	119043	745	8
Amazon-Computers	1	13381	245778	767	10
Coauthor-CS	1	18333	81894	6805	15
WikiCS	1	11701	216123	300	10
MUTAG	188	17.9	39.6	7	2
PROTEINS	1113	39.1	145.6	3	2

Amazon-Photo和Amazon-Computers[36]数据集源自亚马逊平台,都是商品的共购网络。在这些数据集中,节点代表商品,节点之间存在边表示这些商品被用户频繁一起购买。每个节点的特征由产品评论中的稀疏词袋向量表示,并且类别标签基于商品所属的类别进行划分。

Coauthor-CS[36]数据集基于微软学术图,是一个学术合作网络。在该数据集中,节点代表作者,节点之间存在边则表示两个作者曾共同撰写论文。节点特征是根据作者所发表论文的关键词提取的稀疏词袋向量,类别标签表示作者在其研究领域中最活跃的研究方向。

WikiCS[37]数据集是一个基于维基百科的引用网络。在该数据集中,节点代表与计算机科学相关的文章,边表示文章之间的超链接关系。节点特征由文章文本的预训练GloVe词嵌入的平均值表示。

MUTAG[38]数据集来自毒理学研究领域,是一套由188个化学分子构成的小型图集合。每个图对应一个硝基芳香化合物,其中节点表示原子,边表示原子之间的化学键。节点特征通常采用原子类型的独热编码。整个图的标签是二分类,表示该化合物是否对沙门氏菌具有诱变性。

PROTEINS[38]数据集包含1113个蛋白质图,其中的每个图对应一条蛋白质序列,节点是氨基酸残基;如果两个残基在蛋白质的三维结构中相距小于6埃,则在它们之间存在一条无向边。该数据集的图级标签同样是二分类,标记该蛋白质是否属于酶。

5.1.2 基准模型

我们将提出的HCWGC与三组基准模型进行对比:

(1)传统无监督模型,例如:Raw Features、Node2vec[39]和DeepWalk[40]。

(2)有监督模型,例如:GCN[8]和GAT[9]。

(3)无监督GRL模型,例如:MVGRL[21]、DGI[20]、HDMI[22]、GAE[41]、VGAE[41]、GCA[17]、gCooL[14]、GRCCA[13]、CDC[32]和E2Neg[24]。

5.1.3 细节设置

我们使用PyTorch Geometric 2.3.1[42]和PyTorch 2.0.1[43]在NVIDIA GeForce RTX 3090上进行实验(对于Coauthor-CS数据集,我们使用NVIDIA Tesla V100 GPU)。对于节点分类和节点聚类任务,我们遵循文献[14]中的实验设置。首先,以自监督的方式训练图编码器,随后使用编码后的节点表示训练逻辑回归分类器进行节点分类。同时,编码后的节点表示还用于拟合k-means模型,以执行节点聚类任务。对于WikiCS数据集,我们使用给定的20个数据划分,其余数据集按照10%、10%和80%的比例随机划分训练集、验证集和测试集。对于链接预测任务,我们遵循文献[13]中的数据集划分设置,以5%和10%的比例采样验证集和测试集,并从非直接相连的节点对中随机采样相同数量的边作为负样本。对于图分类任务,我们按照80%的比例划分训练集,并采用加和池化作为读出函数,将节点级表示聚合为图级表示。随后,在获得的图表示上训练逻辑回归分类器,以完成下游的图分类任务。

在实验中,我们使用Adam优化器更新模型参数,并使用3次独立重复实验的测试集上评价指标的平均值来衡量模型性能。我们的超参数是以网格搜索的方式确定的,在节点分类和链接预测任务上的具体数值如表2所示。其中,隐藏层维度的设置依据各数据集的特征维度与规模特性进行差异化配置,在满足高维特征空间学习需求的同时根据数据集规模进行平衡设置。在节点聚类任务中,针对Amazon-Photo和WikiCS数据集,ω的数值被调整为0.1。

5.1.4 评价指标

对于节点分类任务,我们使用Micro-$F1$和Macro-$F1$作为评价指标;对于链接预测任务,我们使用ROC曲线下面积(area under roc curve,AUC)和平均精度(average precision,AP)作为评价指标;对于图分类任务,我们使用准确率(accuracy,ACC)作为评价指标。

表 2 详细超参数

超参数	Amazon-Photo	Amazon-Computers	Coauthor-CS	WikiCS
训练轮数	2900	2900	1000	1500
隐藏层维度	256	2558	7105	1368
表示维度	128	512	300	384
学习率	0.1	0.01	0.0005	0.01
激活函数	ReLU	RReLU	RReLU	PReLU
pf	0.1	0.25	0.3	0.1
pe	0.4	0.45	0.2	0.2
ε	0.2	0.2	0.2	0.2
η	500	1000	500	500
K	8	10	15	10
τ	0.3	0.2	0.4	0.4
α	0.7	0.7	0.5	0.5
β	0.1	0.1	0.1	0.1
ω	0.3	0.1	0.5	0.7
γ	0.1	0.1	0.1	0.1

对于节点聚类任务，我们使用归一化互信息(normalized mutual information，NMI)和调整兰德指数(adjusted rand index，ARI)作为评价指标。

NMI 能够从信息论的角度评估两个聚类结果的相似性，其计算方式如公式(11)所示。

$$NMI(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)} \tag{11}$$

其中，变量 X 和 Y 分别表示聚类结果和真实标签，$H(X)$和 $H(Y)$分别代表变量 X 和 Y 的熵，$I(X,Y)$是变量 X 和 Y 之间的互信息。NMI 的取值范围为[0,1]，值越大，代表聚类的结果与真实类别更接近，聚类效果越好。ARI 基于兰德指数，通过对聚类结果与真实标签中的数据点对进行比较，来衡量聚类的一致性，其计算公式如下：

$$ARI=\frac{RI-E(RI)}{\max(RI)-E(RI)} \tag{12}$$

其中，RI 表示数据点对在聚类结果和真实标签中是否被正确分类的比例，$E(RI)$是随机情况下的期望 RI，$\max(RI)$是 RI 的最大值。ARI 的取值范围为[−1,1]，值越大，表示聚类结果与真实标签的吻合度越高。

5.2 整体性能

我们给出了模型在节点分类(表 3)、节点聚类(表 4)和链接预测(表 5)任务上的整体性能。OOM(out of memory)表示模型在该数据集上运行时超出了可用显存的上限。在节点分类任务中，我们用 X、A、$X\&A$ 和 $X\&A\&Y$ 分别表示基准模型是否只利用了图数据的特征信息、拓扑信息，或者同时考虑二者，抑或综合利用了特征信息、拓扑信息和数据标签进行训练。表中粗体标注出了最优结果，下划线标注出了次优结果。在节点分类任务中，我们的模型相较于所有基准模型都取得了更好的性能表现。在节点聚类任务中，我们基于 Amazon-Photo、Amazon-Computers 和 WikiCS 数据集评估了模型的整体性能。结果表明，我们在大多数指标上都取得了较大提升。

表 3 节点分类结果

模型	训练数据	Amazon-Photo		Amazon-Computers		Coauthor-CS		WikiCS	
		Micro-$F1$	Macro-$F1$	Micro-$F1$	Macro-$F1$	Micro-$F1$	Macro-$F1$	Micro-$F1$	Macro-$F1$
GCN	$X\&A\&Y$	92.51±1.00	90.61±1.70	81.55±1.33	75.35±3.89	92.56±0.24	90.23±0.16	78.89± 1.11	74.04±1.35
GAT	$X\&A\&Y$	92.77±0.22	90.87±0.23	85.60±1.82	78.12±6.56	89.40±1.11	86.60±1.03	79.29±0.48	74.34±1.13
Raw Features	X	78.45±0.04	76.10±0.01	73.82±0.01	70.10±0.04	90.40±0.02	89.01±0.06	72.00±0.03	70.28±0.09
Node2vec	A	89.72±0.08	87.39±0.07	84.38±0.08	82.65±0.08	85.11±0.06	82.93±0.11	71.84±0.09	70.44±0.03
DeepWalk	A	89.36±0.10	86.92±0.02	85.63±0.09	84.02±0.10	84.71±0.23	82.63±0.19	74.25±0.06	72.68±0.15
MVGRL	$X\&A$	91.74±0.09	89.93±0.09	87.42±0.07	85.92±0.11	92.11±0.10	90.50±0.12	77.50±0.08	75.62±0.00
DGI	$X\&A$	91.60±0.24	89.31±0.16	83.88±0.50	79.30±0.42	92.08±0.68	90.78±0.68	75.35±0.17	73.74±0.20
HDMI	$X\&A$	90.09±0.10	88.70±0.16	85.43±0.13	83.33±0.17	89.98±0.14	86.73±0.17	75.72±0.55	68.05±0.80
GAE	$X\&A$	91.68±0.14	89.66±0.09	85.18±0.21	83.33±0.17	90.00±0.75	88.31±0.68	70.17±0.05	68.27±0.05
VGAE	$X\&A$	92.24±0.08	90.04±0.17	86.44±0.25	83.72±0.12	92.08±0.08	90.11±0.06	75.56±0.20	74.12±0.10
GCA	$X\&A$	92.30±0.52	90.84±0.94	87.54±0.65	85.73±1.06	92.82±0.31	90.89±0.22	79.38± 0.17	76.50±0.20
gCooL	$X\&A$	93.02±0.44	91.85±0.46	88.62±0.37	87.28±0.61	92.93±0.02	91.01±0.11	79.21± 0.23	76.60±0.28
GRCCA	$X\&A$	91.15±0.23	89.50±0.17	83.66±0.41	81.87±0.23	92.67±0.30	90.50±0.60	79.40±1.37	76.18±0.87
CDC	$X\&A$	64.38±0.55	51.52±0.63	57.85±0.89	24.47±0.82	84.80±0.64	63.40±0.61	75.69±0.47	68.83±0.52
E2Neg	$X\&A$	93.14±0.08	91.48±0.01	84.81±0.83	68.39±0.47	92.91±0.49	90.48±0.61	79.62±0.55	77.07±0.37
HCWGC	$X\&A$	**93.31±0.22**	**92.06±0.32**	**89.02±0.15**	**87.78±0.37**	**93.10±0.17**	**91.35±0.23**	**79.64±0.05**	**77.12±0.07**

结果显示，与最优基线模型相比，HCWGC 在节点分类任务的 Micro-$F1$ 指标上最高相对提升达到 0.45%，Macro-$F1$ 指标上最高相对提升达到 0.57%，在节点聚类任务的 NMI 指标上最高相对提升达到 4.77%，在链接预测任务部分数据集的 AP 指标上最高相对提升了 0.1%。为定量评估 HCWGC

表 4　节点聚类结果

数据集	Amazon-Photo		Amazon-Computers		WikiCS	
	NMI	*ARI*	*NMI*	*ARI*	*NMI*	*ARI*
MVGRL	0.344±0.040	0.239±0.039	0.244±0.000	0.141±0.001	0.263±0.010	0.102±0.011
DGI	0.376±0.030	0.264±0.030	0.318±0.020	0.165±0.020	0.310±0.020	0.131±0.018
HDMI	0.429±0.014	0.307±0.011	0.347±0.011	0.216±0.006	0.238±0.002	0.105±0.000
GAE	0.616±0.010	0.494±0.008	0.441±0.000	0.258±0.000	0.243±0.020	0.095±0.018
VGAE	0.530±0.040	0.373±0.041	0.423±0.000	0.238±0.001	0.261±0.010	0.082±0.008
GCA	0.614±0.000	0.494±0.000	0.426±0.001	0.246±0.001	0.299±0.002	0.121±0.003
gCooL	0.605±0.020	0.485±0.025	0.449±0.013	**0.305±0.059**	0.377±0.023	0.189±0.022
GRCCA	0.478±0.038	0.285±0.061	0.362±0.008	0.281±0.022	0.284±0.024	0.165±0.028
CDC	0.623±0.056	0.485±0.045	0.445±0.066	0.269±0.047	0.370±0.037	0.235±0.053
E2Neg	0.464±0.019	0.229±0.018	0.356±0.017	0.235±0.122	0.341±0.006	0.152±0.001
HCWGC	**0.637±0.018**	**0.514±0.032**	**0.469±0.025**	0.300±0.063	**0.395±0.007**	**0.238±0.020**

表 5　链接预测结果

数据集	Amazon-Photo		Amazon-Computers		WikiCS	
	AUC	*AP*	*AUC*	*AP*	*AUC*	*AP*
GAE	89.58±0.18	89.20±0.22	OOM	OOM	OOM	OOM
VGAE	87.39±0.22	87.34±0.23	OOM	OOM	OOM	OOM
GCA	97.81±0.59	97.41±0.76	**98.34±0.46**	98.05±0.38	**98.76±0.37**	**98.79±0.42**
gCooL	98.47±0.11	98.17±0.09	98.33±0.05	98.06±0.06	98.73±0.07	98.74±0.07
GRCCA	OOM	OOM	OOM	OOM	OOM	OOM
CDC	97.81±0.02	97.49±0.04	97.53±0.03	97.32±0.04	96.81±0.09	96.96±0.07
E2Neg	86.06±0.17	86.01±0.16	88.62±0.89	88.75±0.84	98.45±0.14	98.51±0.14
HCWGC	**98.53±0.11**	**98.27±0.17**	**98.34±0.11**	**98.09±0.14**	98.75±0.05	98.76±0.07

与各对比模型的性能差异是否具有统计显著性,我们在节点分类任务上进行了显著性分析,结果见附录 C。相较于现有无监督模型和自监督 GRL 模型,甚至部分监督模型,我们的模型展现出了巨大的优势,在电子商务、学术研究和知识管理三个领域的数据集上都展现出卓越的性能,能够为商品推荐、新兴领域挖掘和用户行为分析等实际应用场景提供理论基础。与现有的无监督 GCL 模型相比,HCWGC 不仅均衡地综合了两个角度获得互补型视图,并在此基础上进行了多级别加权对比,有效地整合了多尺度信息,并充分考虑了不同负样本对整体语义的影响。相较于节点分类任务,我们的模型在节点聚类任务上取得了更为明显的提升,这是由于模型中 SWSC 方法的目标函数与节点聚类任务的内在需求高度契合,同时作为无监督的节点表示学习方法,避免了监督微调可能引入的语义偏移,从而在无标签依赖的场景中实现更优的群组划分。

5.3　消融实验

我们将模型的良好表现归功于所提出的 SWSC 方法、混合聚类策略和多级别加权对比训练方法。为进一步验证这些方法的贡献,我们设计了消融实验。首先,我们验证了 SWSC 方法以及混合聚类策略的优越性。随后,评估了多级别对比策略的效果。最后,确认了加权策略在对比训练中的有效性。

5.3.1　SWSC 方法性能验证

为了验证 SWSC 相较于其他结构聚类方法的优越性,我们换用不同的结构聚类方法(如 SBM 和 DeCA)与 k-means 方法相结合,并在 Amazon-Photo 数据集上的节点分类和节点聚类任务中进行了验证,结果如表 6 所示。

表 6　不同结构聚类方法下的模型性能

方法	Micro-$F1$	Macro-$F1$	*NMI*	*ARI*
SBM+k-means	93.19±0.07	92.02±0.18	0.504±0.014	0.384±0.020
DeCA+k-means	93.11±0.21	91.88±0.23	0.606±0.031	0.490±0.027
SWSC+k-means	**93.31±0.22**	**92.06±0.32**	**0.637±0.018**	**0.514±0.032**

通过实验结果可以看出,相较于其他结构聚类方法,使用 SWSC 方法与 k-means 方法结合展现了显著的性能提升。原因可以归结为以下两方面:首先,SWSC 方法不同于传统的结构聚类方法 SBM,它不仅依赖于拓扑结构信息,还适当利用了节点特征信息,从而在节点划分时提供了额外的辅助指导。其次,相比于深度结构聚类方法 DeCA,我们设计的 SWSC 方法额外考虑了群组规模对节点归属的影

响。较大规模的群组通常包含更多的节点和连边，这使得待划分的节点在这样的群组中往往拥有更多直接邻居或相似节点，从而导致对该群组的隶属度相对较高。SWSC 方法利用群组局部强度有效衡量了群组规模，并以此指导节点的划分，有效减少了模型在学习过程中的表示偏差，从而提升了模型性能。

5.3.2 混合聚类策略性能验证

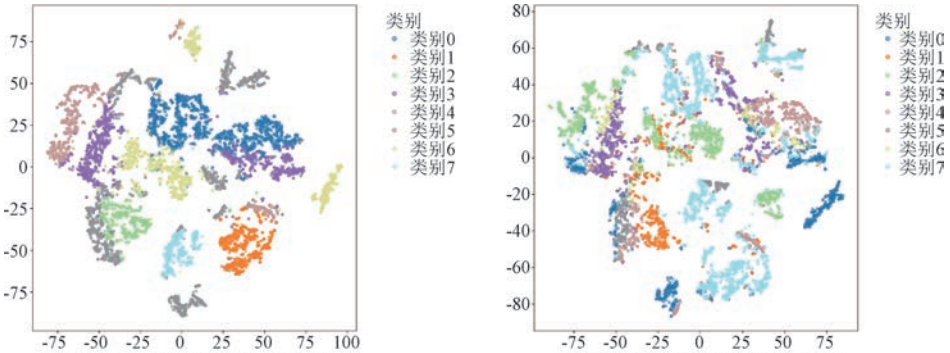
为了验证混合聚类策略的有效性，我们在 Amazon-Photo 数据集上对使用的 k-means 和 SWSC 聚类方法进行消融以探究其对模型性能的影响，结果如表 7 所示。

表 7 聚类方法的消融对模型性能的影响

方法	Micro-$F1$	Macro-$F1$	NMI	ARI
k-means	93.15±0.29	91.90±0.53	0.564±0.030	0.445±0.023
SWSC	93.27±0.33	92.04±0.16	0.591±0.073	0.471±0.101
SWSC+k-means	**93.31±0.22**	**92.06±0.32**	**0.637±0.018**	**0.514±0.032**

从结果中可以看出，采用混合聚类策略的模型在性能上明显优于仅使用单一聚类策略的模型。这一现象主要归因于单一类型的聚类结果可能存在偏差，从而造成性能的次优，而联合使用两种聚类方法，能够从不同的角度产生互补型群组视图，从更加全面的角度挖掘节点的关联性，进而提升对比学习性能。

为了进一步验证 k-means 和 SWSC 方法的聚类倾向性，我们进行了可视化分析。在使用混合聚类策略进行模型训练的过程中分别对 k-means 和 SWSC 方法的聚类结果利用 t-SNE 方法可视化，结果如图 3 所示。其中，图 3(a)和图 3(b)分别为使用 k-means 和 SWSC 方法后的可视化结果。从中可以看出，使用 k-means 方法得到的同群组节点特征分布比使用 SWSC 方法时更加紧密。这验证了 k-means 方法侧重于从全图角度将特征相似的节点划分至同一群组，能够缓解原有拓扑结构中直接邻居的限制，从而挖掘全局隐式结构。



(a) k-means聚类结果的t-SNE可视化　(b) SWSC聚类结果的t-SNE可视化

图 3 k-means 和 SWSC 聚类结果的 t-SNE 可视化比较

随后，我们使用混合聚类策略进行模型训练，分别对两种聚类方法得到的群组利用公式(13)进行平均边密度计算并进行比较，可视化结果如图 4 所示。

$$ED=\frac{1}{K}\sum_{k=1}^{K}ced(k) \tag{13}$$

从图中可以看出，随着训练轮数的增加，SWSC 方法所形成的群组平均边密度呈上升趋势，并在训练一定轮数之后稳定地优于 k-means 方法，这验证了 SWSC 方法侧重于将拓扑连接相对密集的节点划分至同一群组，能够更好地挖掘局部显式结构。

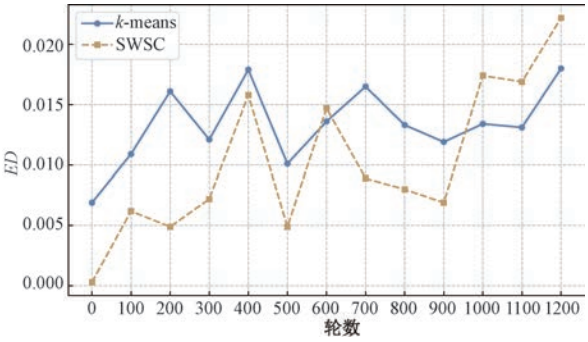


图 4 k-means 和 SWSC 聚类结果的边密度可视化比较

5.3.3 多级别对比策略效果验证

为了促进多尺度的信息融合，我们利用了节点-节点级和节点-群组级多级别对比策略。为了验证这一策略的有效性，我们在多个数据集的节点聚类、链接预测任务和图分类任务中对比了有无节点-群组级对比时模型的性能，结果分别如图 5、图 6 和图 7 所示。

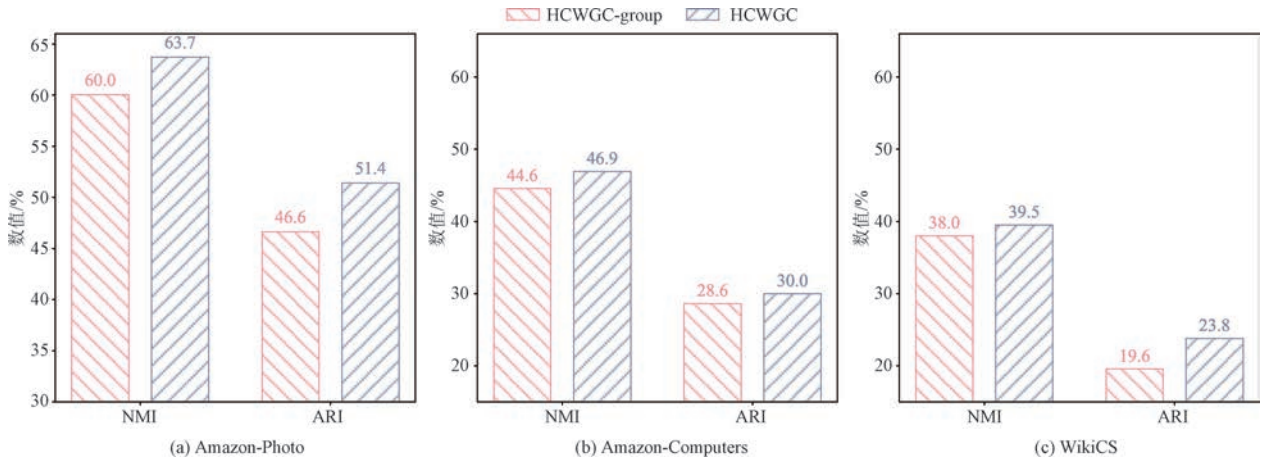


图 5　有无节点-群组级对比时节点聚类任务上模型性能对比

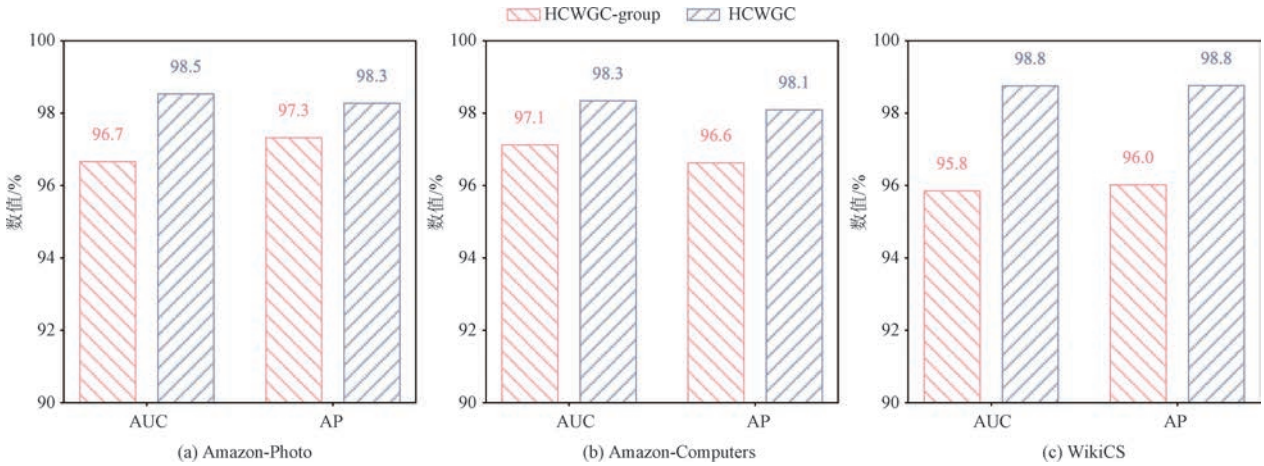


图 6　有无节点-群组级对比时链接预测任务上模型性能对比

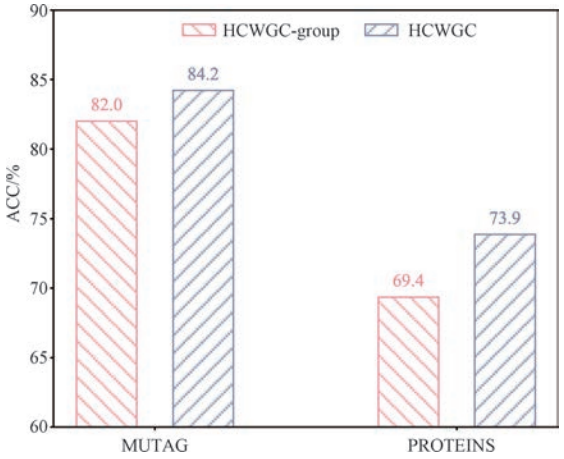


图 7　有无节点-群组级对比时图分类任务上模型性能对比

我们将删除节点-群组级对比部分后的模型命名为 HCWGC-group。从图中可以看出,移除跨级别对比策略后,模型在多个数据集上的性能显著下降,这直接验证了多级别对比策略的有效性。仅保留节点-节点级对比时,模型过于关注节点级视图的差异,忽视了局部和全局视图中的语义信息,这正是性能下降的原因之一。相比之下,基于多级别对比的策略不仅能够有效捕捉节点间的相似性和差异性,还能通过群组级表示获取局部和全局视图中的语义,更好地整合多尺度信息。

此外,多级别对比能够提升模型性能的另一个重要因素是用于对比的群组级视图是根据混合聚类策略生成的。在混合聚类策略中,SWSC 方法主要从拓扑结构的密集性角度获得群组级视图,而 k-means 方法主要从节点特征的相似性角度获得群组级视图。我们的混合聚类策略旨在从两种不同的角度生成群组级视图,使每个视图都能有效地反映图属性,且保证不同视图能够保留图中不同语义信息,这为全面探索图结构数据提供了有效保障。

5.3.4　加权策略有效性验证

为了验证我们在计算对比损失时引入的负样本加权策略的有效性,我们在多个数据集的节点聚类、链接预测任务和图分类任务上对比了有无加权策略时模型的性能,结果分别如图 8、图 9 和图 10 所示。

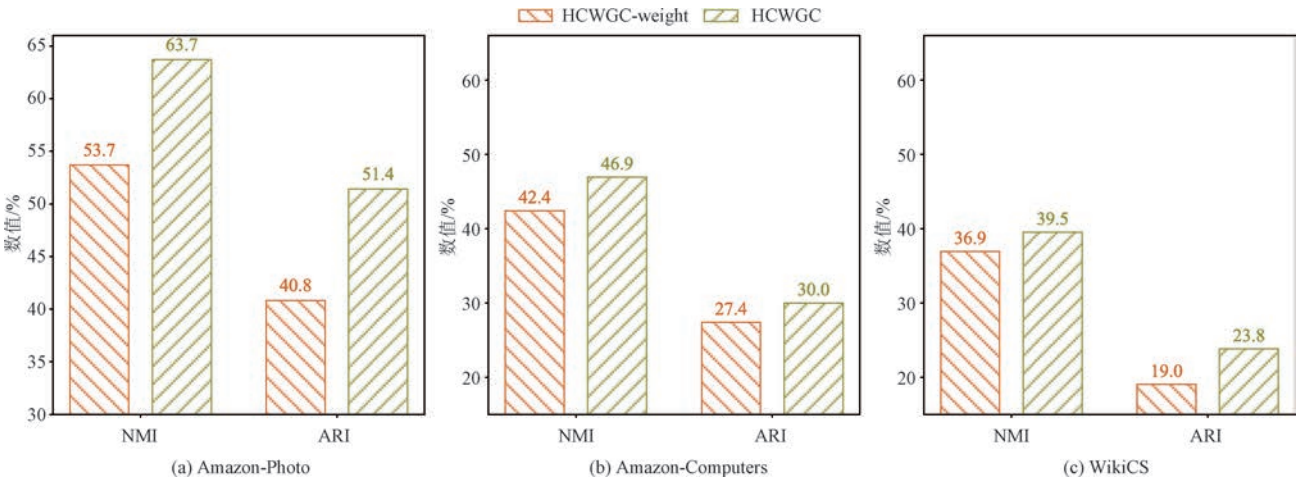


图 8 有无加权策略时节点聚类任务上模型性能对比

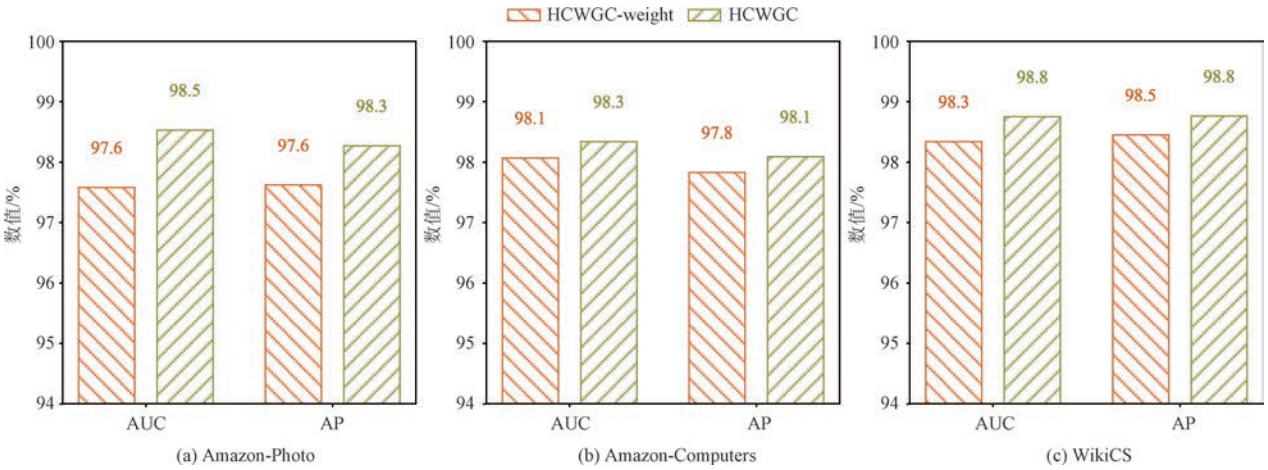


图 9 有无加权策略时链接预测任务上模型性能对比

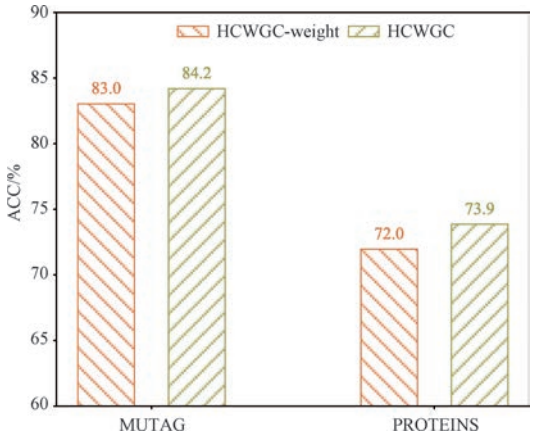


图 10 有无加权策略时图分类任务上模型性能对比

我们将删除负样本加权策略后的模型命名为 HCWGC-weight，通过对比 HCWGC-weight 与原始模型 HCWGC 的实验结果，可以直观验证负样本加权策略的有效性。在节点负样本加权中，我们基于混合聚类结果筛选假负样本，并减少这些假负样本在对比损失中的权重；在群组负样本加权中，根据群组局部强度为规模较大的群组赋予更大的加权项。因此，上述加权策略的有效性也侧面反映了我们采用的假负样本筛选方法和群组局部强度计算策略的合理性与有效性。

GCL 是自监督的方法，节点的实际类别往往是未知的，因此在负样本选取过程中，完全避免假负样本是极为困难的。正因为如此，如何有效筛选假负样本，削减其对模型性能的影响，成为提升模型表现的关键步骤。我们的假负样本筛选方法并不局限于节点表示的相似性，而是巧妙利用混合聚类的结果，全面考虑了节点特征与拓扑结构，进而更准确地评估负样本与锚点的关联性，为筛选假负样本和减少 GCL 噪声引入打开了新的思路。

此外，如何衡量群组语义对全局语义的影响也是 GCL 中需要考虑的重要问题。我们提出的群组局部强度通过评估群组内节点和边的强度来衡量群组的规模，并以此进一步评估群组语义对全局语义的影响。这一方法为解决该问题提供了一种简洁且

有效的途径，同时也为减少 GCL 模型在表示学习过程中产生的偏差拓展了新方法。

5.4　参数分析

在模型训练过程中，参数 ω 控制着在节点-节点级对比中，强相似负样本在对比损失计算中所占的权重，而参数 α 控制着群组局部强度中边强度与节点强度的分配比例。我们记录了 ω 和 α 在 0 到 1 之间变化时，模型在 Amazon-Photo 数据集上的节点聚类任务性能，结果如图 11 所示。

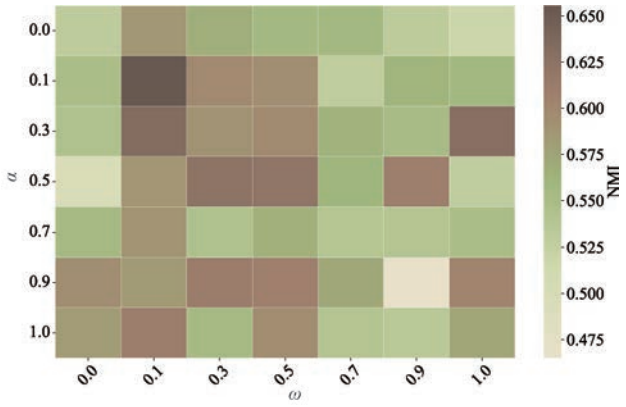


图 11　Amazon-Photo 上参数 α 和 ω 敏感性

实验结果表明，模型的性能峰值主要集中在 ω 值位于[0.1，0.5]的区间。当 ω 的取值为 1 时，代表我们对所有节点负样本“一视同仁”，为所有节点负样本分配相同的权重，这种做法忽视了假负样本的存在，会导致在学习节点表示的过程中引入大量噪声，因此无法获得最佳的性能。当 ω 的取值为 0 时，代表我们完全剔除了假负样本对模型训练的影响，虽然这样做能够让正负样本的边界更加明显，但是这种过于理想的划分方式容易导致模型过拟合，并且可能会丢失一些对于模糊样本的判别能力，因此模型也未能达到最优的性能。当 ω 的值处于[0.1，0.5]的区间范围时，代表我们适当减少了假负样本在计算对比损失时的权重占比，不仅能够减轻假负样本带来的噪声影响，还能适当保留假负样本与正样本之间的相似语义，这有助于模型更好地理解模糊样本，从而提升模型的泛化能力。在该区间内，模型在大多数数据集上的表现能够达到最优，这验证了我们假负样本选择方法的有效性。将强相似负样本定义为假负样本后，我们适当降低强相似负样本在对比损失中的权重的策略，有效减少了假负样本的影响，从而有利于模型的整体性能。

此外，当 ω 取值一定时，模型的性能峰值主要集中在 α 值位于[0.1，0.5]的区间。当 α 的取值为 0 时，代表群组局部强度完全由节点强度控制；当其取值为 1 时，代表群组局部强度完全由边强度控制。节点强度重点突出群组规模的大小，而边强度重点突出群组内部连接的紧密性。因此，综合考量节点强度与边强度的协同作用，可以更全面地描述群组的影响力。

6　总结与展望

在本文中，我们提出了一种新颖的基于混合聚类的图表示对比学习框架 HCWGC。我们通过结合基于特征的 k-means 方法和基于拓扑结构的 SWSC 两种聚类方法生成对比视图，充分挖掘图数据中全局和局部的拓扑关系，从两种不同角度获得互补的语义信息，实现了对图数据的全面探索。此外，我们的多级别加权对比策略为 GCL 中的对比损失计算提供了新的思路。在节点-节点级对比中，我们依据相似等级对节点负样本进行加权；在节点-群组级对比中，我们利用局部强度对群组负样本赋予加权项。这样的计算策略不仅能够全面整合多尺度的语义信息，还能充分考虑负样本之间的差异性。我们的方法在多个数据集上展现出有竞争力的性能，验证了实验动机的合理性和方法的有效性。

未来我们计划在现有算法的基础上进一步研究，重点探索加权参数的自适应调节机制。此外，我们的方法主要针对同质图，如何将其扩展应用到异质图上仍然是一个有待解决的问题。并且，鉴于 k-means 聚类算法在处理百万节点规模的大型图数据集时面临可扩展性瓶颈，如何将本方法高效扩展至此类大规模图数据至关重要。更值得注意的是，节点-节点级对比中对于负样本损失的计算在大规模图上会导致较高的计算复杂度和显存资源消耗。因此，未来工作也将致力于探索并实现高效的负采样策略，以提升方法在大规模图场景下的计算效率和可扩展性。

参　考　文　献

[1] Sharma K, Lee Y C, Nambi S, et al. A survey of graph neural networks for social recommender systems. ACM Computing Surveys, 2024, 56(10): 1-34

[2] Weber M, Domeniconi G, Chen J, et al. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. arXiv preprint arXiv: 1908.02591, 2019

[3] Fang Yin, Zhang Qiang, Zhang Ningyu, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nature Machine Intelligence, 2023, 5(5): 542-553

[4] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA, 2017: 1024-1034

[5] Wu F, Souza A, Zhang T, et al. Simplifying graph convolutional networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 6861-6871

[6] Zhao J, Liu X, Yan Q, et al. Multi-attributed heterogeneous graph convolutional network for bot detection. Information Sciences, 2020, 537: 380-393

[7] Hu F, Zhu Y, Wu S, et al. Hierarchical graph convolutional networks for semi-supervised node classification//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China, 2019, 4532-4539

[8] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-14

[9] Velivckovic P, Cucurull G, Casanova A, et al. Graph attention networks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[10] Wu L, Lin H, Tan C, et al. Self-supervised learning on graphs: Contrastive, generative, or predictive. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 4216-4235

[11] Ju W, Wang Y, Qin Y, et al. Towards graph contrastive learning: A survey and beyond. arXiv preprint arXiv: 2405.11868, 2024

[12] Mavromatis C, Karypis G. Graph infoclust: Maximizing coarse-grain mutual information in graphs//Proceedings of the Advances in Knowledge Discovery and Data Mining-25th Pacific-Asia Conference. Virtual, 2021: 541-553

[13] Zhang C, Yao H, Chen C L P, et al. Graph representation learning via contrasting cluster assignments. IEEE Transactions on Cognitive and Developmental Systems, 2024, 16(3): 912-922

[14] Li B, Jing B, Tong H. Graph communal contrastive learning//Proceedings of the ACM web conference 2022. Lyon, France, 2022: 1203-1213

[15] Wu P, Zhang H, Wang M, et al. Community-aware graph debiased contrastive representation learning//Proceedings of the 2024 International Joint Conference on Neural Networks. Yokohama, Japan, 2024: 1-9

[16] Zhu Y, Xu Y, Yu F, et al. Deep graph contrastive representation learning. arXiv preprint, arXiv:2006.04131, 2020

[17] Zhu Y, Xu Y, Yu F, et al. Graph contrastive learning with adaptive augmentation//Proceedings of the web conference 2021. Ljubljana, Slovenia, 2021: 2069-2080

[18] Chen H, Zhao Z, Li Y, et al. CSGCL: Community-strength-enhanced graph contrastive learning//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Macao, China, 2023: 2059-2067

[19] You Y, Chen T, Sui Y, et al. Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems, 2020, 33: 5812-5823

[20] Velickovic P, Fedus W, Hamilton W L, et al. Deep graph infomax//7th International Conference on Learning Representations. New Orleans, USA, 2019, 2(3): 4

[21] Hassani K, Khasahmadi A H. Contrastive multi-view representation learning on graphs//Proceedings of the 37th International Conference on Machine Learning. Virtual, 2020: 4116-4126

[22] Jing B, Park C, Tong H. Hdmi: High-order deep multiplex infomax//Proceedings of the Web Conference 2021. Ljubljana, Slovenia, 2021: 2414-2424

[23] Miao R, Yang Y, Ma Y, et al. Negative samples selecting strategy for graph contrastive learning. Information Sciences, 2022, 613: 667-681

[24] Huang Y, Zhao J, He D, et al. Does GCL need a large number of negative samples? Enhancing graph contrastive learning with effective and efficient negative sampling//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025, 39(16): 17511-17518

[25] Zhang B, Wang L. False negative sample detection for graph contrastive learning. Tsinghua Science and Technology, 2023, 29(2): 529-542

[26] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm//Proceedings of the Advances in Neural Information Processing Systems 14. Vancouver, Canada, 2001: 849-856

[27] Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 2011, 83(1): 016107

[28] Liu Y, Xia J, Zhou S, et al. A survey of deep graph clustering: taxonomy, challenge, application, and open resource. arXiv preprint, arXiv:2211.12875, 2022

[29] Zhao J, Sun Y, Guo J, et al. Robust graph convolutional clustering with adaptive graph learning//Proceedings of the 2022 International Joint Conference on Neural Networks. Padua, Italy, 2022: 1-8

[30] Zhao H, Yang X, Wang Z, et al. Graph debiased contrastive learning with joint representation clustering//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 3434-3440

[31] Zhou X, Su L, Li X, et al. Community detection based on unsupervised attributed network embedding. Expert Systems with Applications, 2023, 213: 118937

[32] Kang Z, Xie X, Li B, et al. CDC: A simple framework for complex data clustering. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(7): 13177-13188

[33] Newman M E J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577-8582

[34] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[35] Yu J, Ge Q, Li X, et al. Heterogeneous graph contrastive learning with meta-path contexts and adaptively weighted negative samples. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(10): 5181-5193

[36] Shchur O, Mumme M, Bojchevski A, et al. Pitfalls of graph neural network evaluation. arXiv preprint, arXiv:1811.05868, 2018

[37] Mernyei, P, Cangea C. Wiki-cs: A wikipedia-based benchmark for graph neural networks. arXiv preprint, arXiv: 2007.02901, 2020

[38] Morris C, Kriege N M, Bause F, et al. Tudataset: A collection of benchmark datasets for learning with graphs//Graph Representation Learning and Beyond (GRL+) Workshop of International Conference on Machine Learning. Vancouver, Canada, 2020: 79

[39] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 855-864

[40] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 701-710

[41] Kipf T N, Welling M. Variational graph auto-encoders//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 1-3

[42] Fey M, Lenssen J E. Fast graph representation learning with PyTorch Geometric//ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds. New Orleans, USA, 2019

[43] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library//Proceedings of the 32th Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 8024-8035

WANG Di-Ping, M. S. candidate. Her main research interests include data mining and graph neural network.

LIU Hai-Yang, Ph. D., lecturer. His main research interests include data mining and graph neural network.

YUAN Ji-Dong, Ph. D., associate professor. His main research interests include data mining and time series classification.

LI Fang-Jing, Ph. D. candidate. Her main research interests include data mining and graph neural network.

Background

Graph-structured data can effectively capture the complex relationships inherent in many real-world scenarios. With the rapid development of Graph Neural Networks (GNNs), GNNs have become powerful tools for graph representation learning. However, most GNN-based methods rely on supervised learning paradigms, which require a large amount of labeled data that is often difficult to obtain. Furthermore, the performance of these models is highly sensitive to label quality. Therefore, in this work, we focus on unsupervised graph representation learning.

With the recent advances in graph contrastive learning (GCL), the aforementioned challenges have been partially alleviated. Based on the scale of the contrasting views, existing GCL methods can be broadly categorized into intra-scale contrast and inter-scale contrast. Intra-scale contrast focuses on comparing views of the same scale, effectively capturing the similarity between positive samples and the dissimilarity between negative samples within the same scale. Inter-scale contrast, on the other hand, compares views from different scales and can better integrate multi-scale information.

However, most existing methods construct contrastive views from a single perspective, either based solely on node features or on topological structures, which limits the expressiveness of the generated views. Moreover, in the computation of contrastive loss, these methods often overlook the differences among negative samples. For node-level negative samples, the presence of false negatives can introduce noise during training. For group-level negative samples, different groups may exert varying influences on the global semantics. Treating all group-level negatives equally in the contrastive loss can therefore lead to biased representations.

To address these limitations, we propose a multi-level weighted graph contrastive learning framework via hybrid clustering. Our method leverages two complementary clustering techniques to construct diverse contrastive views, thereby overcoming the limitations of single-perspective view generation. Moreover, during the computation of contrastive loss, we assign weights to node-level and group-level negative samples based on similarity ranking and local strength, respectively, in order to fully capture the differences among negative samples. Extensive experiments on multiple graph datasets demonstrate the superior performance of HCWGC over state-of-the-art methods.

This work was supported by the Fundamental Research Funds for the Central Universities (No. 2023JBZY035).

附录 A. 群组内边密度函数

群组内边密度函数旨在衡量一个群组内部节点之间的连通程度，具体表示为群组内实际存在的边数与最大可能边数之比，其计算公式如下：

$$ced(k)=\frac{2|\mathcal{E}_k|}{N_k(N_k-1)} \tag{14}$$

我们从单调性、有界性两个角度证明其合理性。

性质 1. 单调性

对于任意群组 k，当群组内节点数 N_k 固定时，$ced(k)$ 是关于 $|\mathcal{E}_k|$ 的单调递增函数，证明如下：

$$\frac{\partial ced(k)}{\partial|\mathcal{E}_k|}=\frac{2}{N_k(N_k-1)}>0 \tag{15}$$

导函数大于 0，故单调递增，证毕。

性质 2. 有界性

对于任意群组 k，有 $0\leqslant ced(k)\leqslant 1$，证明如下：

由 $|\mathcal{E}_k|\geqslant 0$ 和 $N_k(N_k-1)>0(N_k\geqslant 2)$，易得 $ced(k)\geqslant 0$。当 $|\mathcal{E}_k|=0$ 时，分子为 0，取得下界。

当群组内边数取得最大值时，$|\mathcal{E}_k|=N_k(N_k-1)/2$，故有

$$\begin{aligned}ced(k)&=2|\mathcal{E}_k|/(N_k(N_k-1))\\&\leqslant(N_k(N_k-1))/(N_k(N_k-1))\\&\leqslant 1\end{aligned} \tag{16}$$

证毕。

附录 B. SWSC 损失

由于群组内边密度函数的计算依赖于群组的选择与每次训练的节点划分，使 SWSC 损失的计算成本高且难以向量化。因此，为了便于矩阵形式的运算，我们通过对边密度函数 $ced(k)$ 取其上界，从而得到 $\mathcal{L}_{SWSC}$ 的上界表达形式：

$$\begin{aligned}\mathcal{L}_{SWSC}=&\underbrace{\frac{\lambda}{N(N-1)}\sum_{i,j}\sum_{k_1\neq k_2}\boldsymbol{A}_{ij}\boldsymbol{M}'_{ik_1}\boldsymbol{M}'_{jk_2}}_{inter}-\\&\underbrace{\frac{1}{N}\sum_{i,j}\sum_{k}[\boldsymbol{A}_{ij}-ced(k)]\boldsymbol{M}'_{ik}\boldsymbol{M}'_{jk}}_{intra}\\\leqslant&\underbrace{\frac{\lambda}{N(N-1)}\sum_{i,j}\sum_{k_1\neq k_2}\boldsymbol{A}_{ij}\boldsymbol{M}'_{ik_1}\boldsymbol{M}'_{jk_2}}_{inter}-\\&\underbrace{\frac{1}{N}\sum_{i,j}\sum_{k}[\boldsymbol{A}_{ij}-\max(ced(k))]\boldsymbol{M}'_{ik}\boldsymbol{M}'_{jk}}_{intra}\end{aligned} \tag{17}$$

边密度函数表示群组内实际存在的边数与最大可能边数之比，其计算公式为 $ced(k)=2|\mathcal{E}_k|/(N_k(N_k-1))$，群组内实际边数的最大值对应于群组内节点全连接的情况，因此可得 $ced(k)=2|\mathcal{E}_k|/(N_k(N_k-1))\leqslant 1$。

令 $\widetilde{\boldsymbol{A}}_{ij}=\boldsymbol{A}_{ij}-\max(ced(k))$，可得

$$\begin{aligned}\widetilde{\mathcal{L}}_{SWSC}=&\frac{\lambda}{N(N-1)}\sum_{i,j}\sum_{k_1\neq k_2}\boldsymbol{A}_{ij}\boldsymbol{M}'_{ik_1}\boldsymbol{M}'_{jk_2}-\\&\frac{1}{N}\sum_{i,j}\sum_{k}\widetilde{\boldsymbol{A}}_{ij}\boldsymbol{M}'_{ik}\boldsymbol{M}'_{jk}\end{aligned} \tag{18}$$

令 $\boldsymbol{F}=\boldsymbol{M}^T\boldsymbol{A}\boldsymbol{M}'$，$\widetilde{\boldsymbol{F}}=\boldsymbol{M}^T\widetilde{\boldsymbol{A}}\boldsymbol{M}'$，其矩阵元素 $\boldsymbol{F}_{uv}=\sum_i\boldsymbol{M}'_{iu}(\boldsymbol{A}\boldsymbol{M}')_{iv}=\sum_{i,j}\boldsymbol{A}_{ij}\boldsymbol{M}'_{iu}\boldsymbol{M}'_{jv}$，$\widetilde{\boldsymbol{F}}_{uv}=\sum_{i,j}\widetilde{\boldsymbol{A}}_{ij}\boldsymbol{M}'_{iu}\boldsymbol{M}'_{jv}$。

故有

$$\mathcal{L}_{SWSC}=\frac{\lambda}{N(N-1)}\Big[\sum_{i,j}\boldsymbol{F}_{ij}-tr(\boldsymbol{F})\Big]-\frac{1}{N}tr(\widetilde{\boldsymbol{F}}) \tag{19}$$

附录 C. 显著性检验

为定量评估 HCWGC 与各对比模型在下游节点分类任务上的性能差异是否具有统计显著性，我们补充了配对 t 检验分析并报告了相应的 p 值，结果如附表 1 所示。显著性水平依据经典标准设定为 0.05，即当 p 值 <0.05 时，我们认为差异具有统计显著性。

结果显示，在节点分类任务上，HCWGC 的性能与 Raw Features、Node2vec、DeepWalk、MVGRL、HDMI 和 gCooL 相比，存在统计显著性差异。然而，与其余对比模型相比，观测到的性能差异在统计上未达到显著性水平。由于本研究仅在四个数据集上进行实验，统计检验的功效可能受到影响，使得检测微小但真实的性能差异变得更具挑战性。然而，在全部四个数据集中，我们的方法始终稳定产生正向性能增益，此一致性可以在一定程度上说明观测到的提升具有实际意义。

附表 1 显著性检验结果

模型	p 值
Raw Features	0.0443
Node2vec	0.0125
DeepWalk	0.018
MVGRL	0.0068
DGI	0.0549
HDMI	0.0003
GAE	0.0784
VGAE	0.0573
GCA	0.0841
gCooL	0.0121
GRCCA	0.1826
CDC	0.0809
E2Neg	0.3431