

对比感知的生成式图级补全聚类网络

胡天予, 杨子硕, 谢 夏

(海南大学 计算机科学与技术学院, 海南 海口 570228)

摘 要: 现有图级聚类方法通常假设输入图的节点属性是完整的,但在实际应用中属性缺失问题普遍存在,严重影响聚类性能。为此,提出一种面向属性缺失的生成式图级补全聚类网络CGAE-GC,构建端到端的联合优化框架。该网络首先利用对抗生成网络恢复缺失属性;其次通过对比学习提升图级嵌入的判别性;最后利用软分配与目标分布的迭代优化进一步提高聚类性能。在5个公开基准数据集上,CGAE-GC的平均准确率和F1值分别达到62.0%和53.9%,相较于最优基线分别提升了3.3%和6.7%,验证了该方法在属性缺失场景下的聚类有效性。

关键词: 对比感知;属性缺失;图级聚类;对抗生成;软分配

DOI: 10.11907/rjdk.251833

中图分类号: TP183

文献标识码: A

文章编号: 1672-7800(2026)003-0165-07

扫描二维码阅读全文:



Contrast-Aware Generative Graph-Level Completion Clustering Network

HU Tianyu, YANG Zishuo, XIE Xia

(School of Computer Science and Technology, Hainan University, Haikou 570228, China)

Abstract: The existing graph level clustering methods usually assume that the node attributes of the input graph are complete, but the problem of attribute missing is common in practical applications, which seriously affects the clustering performance. For this reason, a generative graph level completion clustering network CGAE-GC for attribute missing is proposed to build an end-to-end joint optimization framework. Firstly, the network restores the missing attributes by using the confrontation generation network; Secondly, the discriminability of graph level embedding is improved through comparative learning; Finally, the iterative optimization of soft allocation and target distribution is used to further improve the clustering performance. On the five open benchmark datasets, the average accuracy and F1 value of CGAE-GC reached 62.0% and 53.9% respectively, which were 3.3% and 6.7% higher than the optimal baseline, respectively, and verified the clustering effectiveness of this method in the attribute missing scenario.

Key Words: contrast-aware; missing attributes; graph-level clustering; adversarial generation; soft assignment

0 引言

图级聚类(Graph-level Clustering)旨在量化图结构间的相似性,将具有相似拓扑与属性特征的图划分至同一簇。该操作能挖掘复杂数据中潜在的群体结构信息,在生物化学分子分类、社交网络群体发现以及金融风险等领域具有广泛应用^[1-3]。传统的图级聚类方法,如谱聚类(Spectral Clustering)和图核方法(Graph Kernels),主要依赖手工设计的特征或核函数,在大规模数据集上的泛化能力

受限^[4-6]。近年来,随着深度学习的快速发展,通过优化图级嵌入、结合对比学习来提升聚类性能的方法已逐渐成为该领域的主流范式^[7-11]。

1 相关研究

1.1 图级聚类方法

图级聚类旨在度量不同图样本之间的相似性并据此发现簇结构,早期工作主要依赖手工相似性度量与传统聚类器。早期的图级聚类方法通常采用两阶段策略:首先度

收稿日期: 2025-12-21

作者简介: 胡天予(2000-),男,海南大学计算机科学与技术学院硕士研究生,研究方向为知识图谱、图神经网络;杨子硕(2000-),男,海南大学计算机科学与技术学院硕士研究生,研究方向为知识图谱、自然语言处理;谢夏(1978-),女,博士,海南大学计算机科学与技术学院教授、博士生导师,研究方向为知识图谱、数据挖掘。本文通讯作者:谢夏。

量图样本间的相似性,其次应用K-Means等算法实现聚类^[12]。例如,Shi等^[4]提出利用谱嵌入与K-means实现图样本的划分,并通过构建图数据集的亲矩阵(Affinity Matrix)实现图级聚类。Von Luxburg^[13]进一步完善了谱聚类理论,证明了谱嵌入在揭示数据点潜在的非凸流形结构上的有效性,为基于相似性度量的图级聚类方法提供了理论基础。然而,该类方法计算复杂度很高,难以扩展至大规模图数据,且缺乏归纳能力。为克服图结构匹配的计算瓶颈,基于图核的方法通过隐式地将图映射到高维空间来度量样本间的相似性,进而完成聚类。例如,Vishwanathan等^[5]系统阐述了图核框架,通过比较图的局部子结构来计算图间核矩阵,避免了昂贵的图同构匹配。在此基础上,针对传统图核在大规模图上存在的效率问题,Shervashidze等^[6]进一步提出Weisfeiler-Lehman核,利用迭代式的节点邻域聚合策略将特征提取的复杂度降低至线性级别,极大提升了大规模图数据集聚类任务的计算效率。陈梓杨等^[14]针对图编辑距离,提出一种高效的近似策略以量化图间差异。然而,这些方法依赖手工设计相似性与子结构,且提取的特征无法针对聚类目标进行自适应优化。

随着深度学习的发展,基于图神经网络的无监督学习成为主流。例如,Sun等^[15]提出的InfoGraph利用互信息最大化原则学习图的整体表达。You等^[16]提出GraphCL框架,通过设计4种图数据增强策略,生成具有语义一致性的视图,并利用对比损失学习鲁棒的图表示。尽管这些方法增强了特征语义,但仍沿用“先表示、后聚类”的两阶段范式。为实现端到端优化,Cai等^[8]提出DGLC框架,利用伪标签作为监督信号反向指导表示学习。然而,由于DGLC依赖于伪标签的质量,且伪标签分配在迭代初期过于强硬,容易导致错误收敛。朱云华等^[17]采用自适应加权与图对比模块强化节点与邻居的一致性,该方法聚焦于局部结构对齐,缺乏对全图语义的建模。

1.2 面向结构缺失的节点级聚类

图数据的完整性往往无法保证,直接破坏了基于拓扑的消息传递机制。Xia等^[18]指出结构重构是恢复图拓扑的重要方法,可通过学习潜在的边分布来弥补观测数据的不足。Shen等^[19]提出迭代优化邻接矩阵,有效恢复了单视图下的图拓扑一致性。然而,该类方法局限于单视图内部的结构修复,无法利用多源数据中的互补信息。为此,Wen等^[20]提出基于高置信度局部结构的共识图学习框架,通过挖掘跨视图的公共结构与互补信息,实现了对缺失视图的有效填补与聚类。然而,由于依赖局部结构的可信度,在观测数据极其稀疏时,其难以从局部推断全局。为提升稀疏数据下的推断能力,Zhang等^[21]提出一种两阶段图传播方法,先利用图传播机制推断缺失的实例项,再基于推断结果进行结构学习,显著提升了不完整实例到图结构的映射质量。

现有方法普遍遵循一个关键前提,即输入的图数据具

有完整的节点属性。然而,在实际场景中,节点属性缺失是一种普遍现象。尽管文献[22-24]已关注图结构不完整的问题,但专门针对节点属性缺失的图级聚类任务仍被忽视。属性缺失会使模型在表示学习过程中累积偏差,而事后独立的补全步骤则无法与下游的聚类目标协同优化,最终导致学习到的图嵌入失真,严重损害聚类性能。为此,本文提出一种面向属性缺失的生成式图级补全聚类网络(Contrast-Aware Generative Graph-Level Completion Clustering Network, CGAE-GC),通过构建属性补全、表示学习与聚类目标的联合优化框架,实现了特征学习与高层聚类目标的动态对齐,解决了表示学习与聚类分离的问题,并利用基于概率的软分配策略有效缓解了硬性划分带来的偏差累积,使模型在属性缺失的背景下依然能保证优秀的聚类性能。本研究拓展了图级聚类在真实环境下的适用边界,具有重要的理论意义与应用价值。

2 图级聚类基本概念与理论

设图数据集为 $\mathcal{G} = \{G_1, G_2, \dots, G_N\}$,每个样本图 $G_i = (V_i, E_i, X_i)$ 由节点集合 $V_i = \{v_{i,1}, \dots, v_{i,|V_i|}\}$ 、边集合 $E_i \subseteq V_i \times V_i$ 以及节点属性矩阵 $X_i \in \mathbb{R}^{|V_i| \times F}$ 构成,其中 F 为节点属性的特征维度,邻接矩阵记为 $A_i \in \{0, 1\}^{|V_i| \times |V_i|}$ 。在节点属性缺失的背景下, X_i 仅部分可观测。为刻画缺失模式,引入掩码矩阵 $M_i \in \{0, 1\}^{|V_i|}$,当 $(M_i)_u = 1$ 时表示节点 u 的属性可观测,否则为缺失。相应地,不完整属性矩阵记为 X_i^{inc} 。在无监督设定下,目标为在未知标签的情况下将 $\mathcal{G}$ 划分为 K 个簇,并输出每个图的簇指派 $\{y_i\}_{i=1}^N$ 。

图级聚类通常遵循“表示学习—聚类优化”的基本范式,即首先学习每个图的低维表示;其次在嵌入空间中进行聚类。给定一张图 G_i ,图编码器 $f_\theta(\cdot)$ 将其映射为图级嵌入向量,表示为:

$$H_i = f_\theta(A_i, X_i) \quad (1)$$

式中: θ 为模型参数。图编码器可由图神经网络或基于注意力的Graph Transformer实现。当编码器输出为节点级表示 $H_i \in \mathbb{R}^{|V_i| \times d}$ 时,通常通过读出/池化算子 $READOUT(\cdot)$ 得到图级嵌入,表示为:

$$z_i = READOUT(H_i) \quad (2)$$

式中: $READOUT(\cdot)$ 可以是全局平均池化、最大池化或注意力池化等。

在聚类优化方面,常见方法是引入对比学习机制来构造嵌入空间的几何结构,通过构造数据增强后的多视图嵌入,将同一图样本在不同视图下得到的嵌入作为正样本对,使其在表示空间中更加接近。同时将不同图样本的嵌入视为负样本对,促使其相互远离,从而提升簇内紧凑性与簇间可分性。具体来说,设 $\mathcal{T}_1$ 与 $\mathcal{T}_2$ 为两种随机增强算子,针对图 G_i 生成两种视图并得到嵌入 z_i 与 $\tilde{z}_i$,采用基于温度参数 τ 的对比目标进行优化,表示为:

$$\mathcal{L}_{con} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(z_i, \tilde{z}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_i, \tilde{z}_j)/\tau)} \quad (3)$$

式中: $\text{sim}(\cdot, \cdot)$ 表示相似度函数(如余弦相似度)。该优化过程能够在无监督条件下强化表示的判别性, 为后续聚类划分提供结构基础。

3 CGAE-GC构建

CGAE-GC的网络架构如图1所示。该框架通过端端的联合优化策略, 集成了3个核心组件: 首先, 生成补全模块利用对抗生成机制, 根据图的拓扑结构恢复缺失的节点属性; 其次, 图嵌入学习模块通过多视图增强与对比训练, 从补全后的图中提取具有判别性的图级嵌入; 最后, 联合优化模块通过迭代优化软分配分布与目标分布, 驱动网络学习聚类友好的特征空间。以下详细阐述各模块的实现细节。

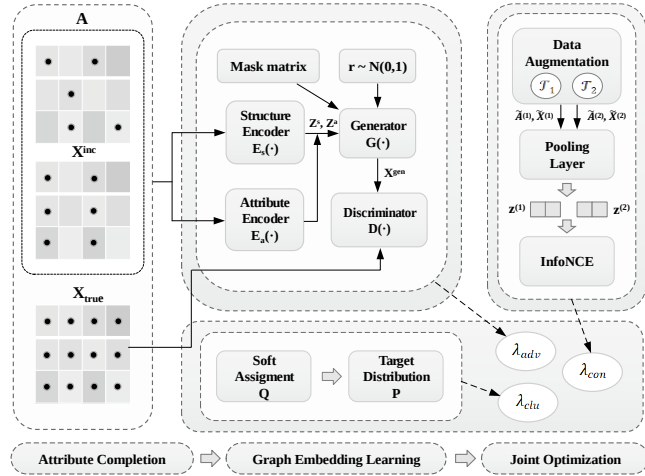


Fig. 1 CGAE-GC network architecture

图1 CGAE-GC网络架构

3.1 生成补全模块

给定图 $G_i = (V_i, E_i, X_i^{inc})$ 及其邻接矩阵 A_i , 采用图同构网络(Graph Isomorphism Network, GIN)作为图编码器, 分别抽取结构信息与属性信息^[25]。结构编码器 $E_s(\cdot)$ 仅依赖结构输入, 输出结构嵌入, 表示为:

$$Z_i^s = E_s(A_i) \quad (4)$$

属性编码器 $E_a(\cdot)$ 以结构与不完整属性为输入, 输出属性嵌入, 表示为:

$$Z_i^a = E_a(A_i, X_i^{inc}) \quad (5)$$

式中: $Z_i^s \in \mathbb{R}^{|V_i| \times d}$ 与 $Z_i^a \in \mathbb{R}^{|V_i| \times d}$ 均为节点级表示; d 为隐空间维度。

在生成阶段, 使用一个由多层感知机(Multilayer Perceptron, MLP)构成的生成器 $G(\cdot)$ 在结构嵌入与属性嵌入的条件下生成候选属性矩阵。为提升生成多样性并缓解模式坍塌, 引入随机噪声向量 $r \sim \mathcal{N}(0, I)$ 。生成公式为:

$$X_i^{gen} = G(Z_i^s, Z_i^a, r) \quad (6)$$

式中: $X_i^{gen} \in \mathbb{R}^{|V_i| \times F}$, 为生成器输出的候选属性矩阵。

为确保观测属性不被覆盖, 引入缺失掩码矩阵 $M_i \in \{0, 1\}^{|V_i|}$, 并采用掩码融合得到最终补全属性, 表示为:

$$X_i^{comp} = M_i \odot X_i^{inc} + (1 - M_i) \odot X_i^{gen} \quad (7)$$

为提升生成属性与原属性的适配度, 引入判别器 $D(\cdot)$ 对真实完整属性与生成属性进行判别。判别器的输入为 X_i^{true} 或 X_i^{gen} , 输出为真实样本的概率。通过生成器与判别器的对抗训练, 生成器被迫生成更符合真实分布的属性, 从而提升补全质量。相应的对抗损失记为 $\mathcal{L}_{adv}$, 其参与联合训练。

3.2 图嵌入学习模块

尽管属性补全能够缓解属性缺失带来的信息不足问题, 但补全属性中仍包含一些不确定性与生成噪声。因此, 本文引入对比学习机制, 通过多视图一致性约束引导模型关注图中对扰动不敏感的本质结构与属性模式。给定补全后的图 $G_i^{comp} = (A_i, X_i^{comp})$, 采用随机图增强操作构造两种不同的视图, 增强方式采用随机边或节点属性遮罩。设 $\mathcal{T}_1$ 与 $\mathcal{T}_2$ 为两种随机增强算子, 则可得到 $(\tilde{A}_i^{(k)}, \tilde{X}_i^{(k)}) = \mathcal{T}_k(A_i, X_i^{comp})$ 。为了在多视图条件下获得一致的表示, 采用共享参数的图编码器 $f_\theta(\cdot)$ 对两个增强视图进行编码, 该编码器由 Graph Transformer 构成^[26]。对于第 $v \in \{1, 2\}$ 个视图, 其节点级表示为:

$$H_i^{(v)} = f_\theta(\tilde{A}_i^{(v)}, \tilde{X}_i^{(v)}) \quad (8)$$

通过图池化算子 $Pool(\cdot)$ 将节点级表示聚合为图级嵌入, 从而得到同一图在不同增强视图下的图级嵌入 $z_i^{(1)}$ 与 $z_i^{(2)}$ 。表示为:

$$z_i^{(v)} = Pool(H_i^{(v)}) \quad (9)$$

在对比学习阶段, 同一图的不同视图构成正样本对, 不同图之间的表示构成负样本对。通过最大化正样本对的相似性并最小化负样本对的相似性, 模型被迫学习在增强扰动下保持不变的判别特征。采用基于 InfoNCE 的对比损失函数, 该函数鼓励同一图在不同视图下的嵌入保持一致, 同时拉开不同图样本之间的表示距离^[27]。表示为:

$$\mathcal{L}_{con} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(z_i^{(1)}, z_i^{(2)})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_i^{(1)}, z_j^{(2)})/\tau)} \quad (10)$$

式中: $\text{sim}(\cdot, \cdot)$ 表示相似度函数; τ 为温度系数。

3.3 联合优化模块

深度嵌入聚类(Deep Embedding Clustering, DEC)通过将聚类中心的学习与图嵌入的更新耦合, 从而实现端到端的聚类导向优化^[28]。设通过图嵌入学习模块获得的图级嵌入为 $\{z_i\}_{i=1}^N$, 为保证后续深度聚类过程的稳定性, 首先对 $\{z_i\}$ 执行 K-Mean 聚类, 并将得到的 K 个簇中心作为初始聚类原型。该初始化策略能有效避免随机初始化导致的聚类塌陷问题, 并为自训练优化提供合理起点。表示为:

$$\{\mu_k\}_{k=1}^K = KMeans(\{z_i\}_{i=1}^N) \quad (11)$$

在初始化聚类中心后, 采用 Student's t 分布来度量图

嵌入与聚类中心之间的相似性,并计算软分配概率。对于第 i 个图与第 k 个聚类中心,软分配公式为:

$$q_{ik} = \frac{(1 + \|z_i - \mu_k\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'} (1 + \|z_i - \mu_{k'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (12)$$

式中: α 为自由度参数。

软分配矩阵 $Q = [q_{ik}]$ 为每个图提供了关于其隶属各个簇的概率性估计。直接基于 Q 进行优化容易受到低置信分配的干扰。为强化高置信样本在优化过程中的作用,构造目标分布 $P = [p_{ik}]$ 。该加权方式放大了具有较大 q_{ik} 样本的权重,从而使聚类优化过程更关注可靠分配。表示为:

$$p_{ik} = \frac{q_{ik}^2 / \sum_i q_{ik}}{\sum_{k'} q_{ik'}^2 / \sum_i q_{ik'}} \quad (13)$$

在获得目标分布后,通过最小化目标分布 P 与当前软分配 Q 之间的 KL (Kullback-Leibler Divergence) 散度来更新图嵌入与聚类中心。公式为:

$$\mathcal{L}_{clu} = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \quad (14)$$

该损失直接作用于图级嵌入 z_i ,从而反向推动图编码器与前序模块的参数更新。在联合训练阶段, $\mathcal{L}_{clu}$ 与对抗损失 $\mathcal{L}_{adv}$ 和对比损失 $\mathcal{L}_{con}$ 共同构成整体优化目标。因此,总体损失函数表示为:

$$\mathcal{L}_{total} = \lambda_{adv} \mathcal{L}_{adv} + \lambda_{con} \mathcal{L}_{con} + \lambda_{clu} \mathcal{L}_{clu} \quad (15)$$

式中: λ_{adv} 、 λ_{con} 和 λ_{clu} 分别用于调整补全质量、表示学习与聚类性能的权重分布。

4 实验方法与结果分析

4.1 实验设置

实验在一台配备 Nvidia 4090 GPU 与 Intel (R) Xeon (R) Silver 4310 CPU @2.10GHz 的服务器上完成。CGAE-GC 各模块的配置为:①在生成补全模块中,结构编码器 $E_s(\cdot)$ 与属性编码器 $E_a(\cdot)$ 均采用 GIN 作为骨干网络;生成器 $G(\cdot)$ 采用 MLP 并引入噪声向量 $r \sim \mathcal{N}(0, I)$,判别器 $D(\cdot)$ 同样采用 MLP,并与生成器进行对抗训练。缺失掩码 $M \in \{0, 1\}^{|V|}$ 的缺失率设置为 30%;②图嵌入学习模块采用 Graph Transformer 作为共享图编码器 $f_\theta(\cdot)$,并通过平均池化获得图级嵌入。数据增强采用边扰动(随机添加/删除少量边)和节点特征掩蔽(随机将部分节点特征置零)两种方法,从而形成两种随机视图用于对比学习,对比目标采用 InfoNCE 损失 $\mathcal{L}_{con}$;③在联合优化模块中,聚类簇数 K 设定为各数据集的真实类别数,软分配的自由度参数 α 设定为 1.0,总损失函数中的权重系数比例设定为 $\lambda_{adv} : \lambda_{con} : \lambda_{clu} = 1 : 1 : 1$ 。所有超参数在验证集上进行优化,并在所有数据集上保持一致的训练流程以保证结果的可比性。

4.2 数据集

为全面评估 CGAE-GC 在不同领域图数据中的聚类性能,选取 5 个常用图分类基准数据集进行实验,这些数据集规模、结构复杂度与类别粒度上差异显著。具体信息如表 1 所示。其中 BZR 和 COX2 均属于分子图数据集,BZR 中为中等规模的分子结构图,COX2 比 BZR 略大;DD 为生物信息学图数据集,共包含 1 178 个图样本,平均节点数为 284.32,平均边数为 715.66,是节点规模与连接复杂度最高的数据集之一;IMDB-MULTI 为社交网络图数据集,包含 1 500 个图样本,具有小规模但相对稠密的结构特征;Letter-low 为计算机视觉相关图数据集,包含规模小且簇数较多的场景,常用于检验模型在细粒度聚类任务中的判别能力。

Table 1 Details of datasets

表 1 数据集详情

数据集	图数量	类别数	平均节点数	平均边数
BZR	405	2	35.75	38.36
COX2	467	2	41.22	43.45
DD	1 178	2	284.32	715.66
IMDB-MULTI	1 500	3	13.00	65.94
Letter-low	2 250	15	4.68	3.13

4.3 基线方法与评价指标

为验证 CGAE-GC 在节点属性缺失条件下的优越性,从传统相似性度量方法与深度学习方法中选取多种具有代表性的强基线进行比较。主要包括 3 类:①基于图核的传统方法。该类方法通过提取和比较图的子结构将图映射到高维空间以度量样本间的相似性,包括利用最短路径长度分布的最短路径核 $SP^{[5]}$ 以及基于局部子图频次统计的 Graphlet Kernel (GK)^[29];②无监督图嵌入学习方法。该类方法通常先学习图级嵌入再采用聚类器完成划分,包括 InfoGraph^[16]、GraphCL^[30];③图级聚类方法。该类方法将聚类目标与表示学习紧密结合,包括 GLCC^[11]、UDGC^[31]、DGLC^[8] 与 DCGLC^[32],DGLC 遵循其官方代码与参数配置建议。为保证比较的公平性,所有方法均在相同的硬件环境下运行,并在每个数据集上重复实验 5 次后计算平均结果。此外,DGLC 遵循其官方代码与参数配置建议。

采用 4 个被广泛使用的聚类评估指标,分别为聚类准确率 (Clustering Accuracy, ACC)、归一化互信息 (Normalized Mutual Information, NMI)、调整兰德指数 (Adjusted Rand Index, ARI) 以及 Macro-F1 分数 (Macro-F1 Score)。其中,ACC 衡量了聚类结果与真实标签的整体吻合程度;NMI 用于度量聚类结果与真实划分之间的信息一致性,能够在不同簇数与样本规模下保持可比性;ARI 基于样本对的一致性统计对聚类结果进行校正,对随机分配具有鲁棒性,并能较好地应对类不平衡带来的偏置;Macro-F1 为各类别 F1 分数的算术平均值,基于查准率 (Precision) 与查全率 (Recall) 评估模型在每个簇上的分类质量。

4.4 比较实验

表 2 为 CGAE-GC 与多个基线方法在 5 个数据集上的聚类结果。可以看出, CGAE-GC 在 4 个指标上均表现出明显优势, 具体分为: ①在跨领域数据集上的整体优势。在不同领域的数据集上, CGAE-GC 在 ACC 与 Macro-F1 上均取得最优或接近最优的结果, 与传统图核方法 (SP、GK) 相比优势显著, 表明仅依赖固定相似性度量难以应对属性缺失带来的分布偏移, 而端到端学习的表示能够更好地捕捉图的判别性结构与属性模式; ②对属性缺失噪声的鲁棒性。无监督图嵌入学习方法 InfoGraph、GraphCL 虽然能够学习无监督表征, 但其目标主要为表示一致性, 缺乏对聚类可分性的显式约束, 因此在 NMI、ARI 等更强调簇结构一致性的指标上表现不佳。相比之下, CGAE-GC 一方面利用生成式补全在结构条件下恢复缺失属性, 缓解了由 X^{inc} 引起的消息传递偏差; 另一方面通过对比学习增强嵌入的扰动不变性, 并进一步借助 DEC 机制将嵌入空间持续拉向

“簇内紧凑、簇间分离”的结构, 从而在 ACC、ARI 与 Macro-F1 指标上实现更一致的提升; ③与先进图级聚类方法的比较。与 GLCC、UDGC 等先进图级聚类框架相比, CGAE-GC 在所有数据集上均展现出较优性能, 且在结构复杂度较高与类别粒度较细的场景下优势尤为突出。具体而言, 在 DD 数据集上, CGAE-GC 的 ACC 达到 72.2%, 显著优于次优模型 DGLC (69.3%), 且在 NMI 指标上达到了 17.8%, 而同期基线普遍低于 5%, 证明了其在复杂拓扑下仍能保持较高的聚类纯度; 同时, 在多类别数据集 Letter-low 上, CGAE-GC 展现出优异的细粒度划分能力, NMI 高达 42.1%, 远超 DGLC (1.2%) 与 DCGLC (4.2%)。如图 2 所示, CGAE-GC 的平均表现处于领先地位; DGLC 在 ACC 与 Macro-F1 上具备一定竞争力, 但在 NMI 与 ARI 这类更强调簇结构一致性的指标上存在明显差距; UDGC 的整体均值偏低, 对数据分布变化更敏感; DCGLC 在部分指标上表现良好, 但未达到 CGAE-GC 的平均水平。

Table 2 Clustering results of CGAE-GC and multiple baseline methods on five data sets

表 2 CGAE-GC 与多个基线方法在 5 个数据集上的聚类结果

(%)

方法	ACC					NMI				
	BZR	COX2	DD	IMDB-MULTI	Letter-low	BZR	COX2	DD	IMDB-MULTI	Letter-low
SP	65.2±0.3	51.2±0.2	53.5±0.2	22.1±0.1	15.1±0.7	3.9±0.1	0.1±0.1	0.6±0.1	0.2±0.1	23.2±2.4
GK	56.3±2.7	44.5±0.1	53.0±1.1	23.3±0.1	13.2±0.3	0.7±0.0	0.0±0.0	0.4±0.2	0.6±0.1	16.2±1.4
InfoGraph	63.7±2.4	51.0±1.4	53.1±4.5	30.2±0.1	13.6±0.4	1.2±0.5	2.3±1.7	0.7±0.2	0.5±0.0	13.1±0.6
GraphCL	60.5±3.1	48.5±1.7	50.6±5.0	28.0±0.3	12.2±0.4	0.9±0.7	2.4±1.9	0.5±0.3	0.3±0.1	11.7±0.6
GLCC	67.2±2.3	54.6±1.3	57.0±4.1	32.7±0.1	15.3±0.3	1.7±0.4	3.2±1.4	0.9±0.1	0.6±0.1	14.5±0.3
UDGC	60.2±8.0	69.4±3.1	53.5±0.1	32.8±3.8	29.3±5.3	1.3±0.9	1.1±0.6	2.1±0.1	1.3±0.0	33.7±1.2
DGLC	70.9±4.7	78.7±4.1	69.3±2.8	38.8±2.2	36.1±0.9	3.4±0.0	1.3±0.1	3.9±0.5	1.5±0.3	1.2±0.1
DCGLC	69.6±4.1	78.2±3.5	59.9±2.0	40.5±2.7	44.6±2.0	7.4±1.4	1.7±0.7	1.8±0.1	3.2±1.1	4.2±1.6
CGAE-GC	74.0±5.2	78.5±2.8	72.2±1.4	40.6±0.2	45.7±8.2	8.8±1.1	2.9±0.9	17.8±1.1	2.2±0.1	42.1±0.2

方法	ARI					Macro-F1				
	BZR	COX2	DD	IMDB-MULTI	Letter-low	BZR	COX2	DD	IMDB-MULTI	Letter-low
SP	3.9±0.2	0.3±0.1	0.1±0.1	0.2±0.0	15.0±1.4	46.6±1.2	50.7±0.6	52.1±0.2	16.8±0.1	9.4±0.1
GK	2.8±0.7	0.1±0.0	0.0±0.0	0.1±0.0	11.1±0.2	43.7±5.6	40.8±1.0	50.1±1.2	17.7±0.1	13.1±0.7
InfGraph	3.9±0.9	3.1±1.3	0.1±0.0	0.3±0.0	13.2±0.3	55.8±0.7	50.2±3.4	49.6±0.2	21.5±0.5	12.2±0.8
GraphCL	3.2±1.1	2.5±1.5	0.1±0.1	0.2±0.1	11.5±0.2	52.5±1.4	47.7±3.7	48.3±0.3	19.8±0.7	10.8±1.0
GLCC	4.6±0.8	3.6±1.2	0.2±0.1	0.4±0.1	14.8±0.5	60.4±0.8	53.1±2.9	51.8±0.1	24.0±0.3	14.4±0.6
UDGC	1.2±0.1	0.3±0.0	2.9±0.2	0.4±0.3	15.8±0.6	56.6±0.7	36.3±3.9	51.8±3.3	29.3±0.4	31.0±2.8
DGLC	8.4±0.8	4.7±1.2	4.5±0.2	1.4±0.1	0.1±0.1	61.7±3.9	66.0±3.0	60.5±1.1	32.1±1.9	15.8±0.6
DCGLC	17.2±1.9	5.0±1.0	1.1±0.1	3.1±1.3	32.8±1.0	63.2±3.2	69.4±2.7	54.8±2.0	33.7±2.2	12.5±1.8
CGAE-GC	17.9±0.5	5.1±0.0	18.1±0.7	3.3±0.2	33.7±0.8	69.8±3.7	71.1±1.9	68.5±5.5	37.2±0.1	30.8±0.0

4.5 消融实验

分别对生成补全模块、图嵌入学习模块以及联合优化模块进行消融实验, 结果如表 3 所示。从聚类结构一致性指标 (NMI、ARI) 来看, 缺少任何模块都会削弱簇结构的可辨识性, 尤其是在结构复杂、节点规模较大的 DD 数据集上, 去除生成补全模块会使 NMI 出现显著回落, 表明在属性缺失的条件下缺乏生成式图级补全会使模型更难恢复与结构一致的属性分布, 从而影响聚类结构的可分性。图嵌入学习模块通过多视图对比约束提升表示的判别性与鲁棒性, 移除后模型整体指标出现回落。联合优化模块的移除对 ACC 和 Macro-F1 指标影响较大, 这是由于在缺乏聚类优化时, 嵌入空间出现明显退化。实验结果验证了 3

个模块是实现高质量图级聚类的关键。

4.6 缺失率敏感性分析

为全面评估 CGAE-GC 在不同数据质量下的适应能力, 对属性缺失率进行敏感性分析。将缺失率扩展为 50% 和 70%, 并在所有数据集上记录模型的性能变化, 实验结果如表 4 所示。从整体聚类性能来看, 随着缺失率的增高, CGAE-GC 在各数据集上的指标仅呈现出轻微下降趋势, 并未发生严重的性能坍塌, 尤其是在 BZR 和 COX2 数据集上, 即使面对 70% 的严重缺失, 模型的 ACC 仍维持在较高水平, 这是由于生成补全模块在信息极度匮乏的条件下仍能有效恢复潜在的数据分布。此外, 各项指标的标准差普遍出现扩大的趋势, 反映了在高缺失率场景下, 生成

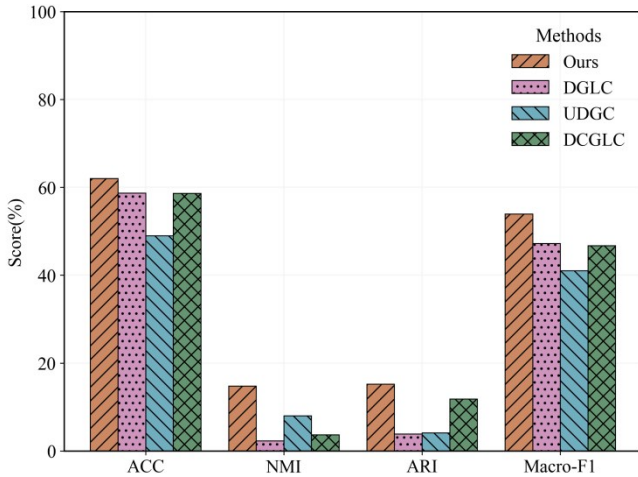


Fig. 2 Average value of indicators of the four best methods on five data sets

图2 4种最优方法在5个数据集上的指标平均值

Table 3 Module ablation test results

表3 模块消融实验结果

指标	模块			BZR	COX2	DD	IMDB-MULTI	Letter-low
	生成补全	图嵌入学习	联合优化					
ACC	✓	✓	✓	70.5±3.2	74.8±2.5	68.7±2.8	37.2±1.8	42.1±2.3
	✓	✓		68.2±3.5	73.9±2.8	67.5±3.0	36.4±2.0	41.3±2.1
	✓		✓	71.8±2.9	75.9±2.4	69.6±2.7	38.1±1.9	43.2±2.5
NMI		✓	✓	6.7±0.9	2.2±0.5	14.5±1.2	1.8±0.4	38.6±1.8
	✓	✓		6.9±1.0	2.3±0.4	15.1±1.1	1.7±0.3	39.0±1.7
	✓		✓	7.9±0.8	2.6±0.6	16.2±1.3	1.9±0.5	40.3±1.9
ARI		✓	✓	14.1±0.7	2.0±0.3	15.1±0.8	2.7±0.4	30.7±1.0
	✓	✓		14.8±0.6	2.5±0.2	15.2±0.7	2.7±0.3	30.2±0.9
	✓		✓	16.4±0.8	2.8±0.4	16.5±0.9	3.0±0.5	32.1±1.1
Macro-F1		✓	✓	66.5±1.5	68.3±2.0	65.7±1.8	34.8±1.2	20.5±0.8
	✓	✓		65.9±1.6	67.8±1.9	65.1±1.7	34.3±1.1	19.9±0.7
	✓		✓	68.2±1.7	69.7±2.1	66.8±1.9	35.6±1.3	21.2±0.9

补全的不确定性对下游聚类产生了一定影响。总体来说, CGAE-GC 在面对严重属性缺失时依然具备强大的鲁棒性与适应能力。

Table 4 Model performance under different missing rates

表4 不同缺失率下的模型性能

指标/%	缺失率/%	BZR	COX2	DD	IMDB-MULTI	Letter-low
ACC	30	74.0±5.2	78.5±2.8	72.2±1.4	40.6±0.2	45.7±8.2
	50	73.2±5.3	77.9±3.5	70.8±2.1	39.8±0.5	44.2±8.9
	70	70.1±6.8	75.6±3.5	68.5±3.2	38.5±1.0	41.8±9.8
NMI	30	8.8±1.1	2.9±0.9	17.8±1.1	2.2±0.1	42.1±0.2
	50	8.6±1.5	2.6±1.2	16.5±1.6	2.0±0.3	40.5±0.8
	70	7.9±2.0	2.1±1.6	14.9±2.3	1.7±0.5	38.2±1.5
ARI	30	17.9±0.5	5.1±0.0	18.1±0.7	3.3±0.2	33.7±0.8
	50	16.5±0.9	4.5±0.3	16.8±1.3	2.9±0.4	31.9±1.5
	70	14.8±1.5	3.8±0.6	15.0±2.1	2.4±0.7	29.4±2.2
Macro-F1	30	69.8±3.7	71.1±1.9	68.5±5.5	37.2±0.1	30.8±0.0
	50	67.9±4.3	69.5±2.6	66.8±6.3	36.5±0.4	29.9±0.5
	70	66.4±4.9	67.2±3.5	64.5±7.5	35.1±0.8	28.5±1.1

4.7 权重系数敏感性分析

为考察 CGAE-GC 对联合损失中各项权重系数的敏感度,在保证其余训练设置不变的前提下,分别采用均衡权重与3种单项占优的对照设置进行比较,并选用相对复杂的 DD 作为数据集,实验结果如图3所示。可以看出,当提高聚类权重时,模型在 ACC 与 Macro-F1 上达到最佳,表明更强的聚类牵引有助于形成更可分的嵌入空间。当偏重对抗补全或对比学习时,模型的 ACC 和 Macro-F1 均有所下降,说明仅强调补全质量或表示一致性不足以最大化聚类性能。模型的 ARI 和 NMI 在偏重补全时略高,反映出更强的补全约束对簇结构一致性指标更有利。

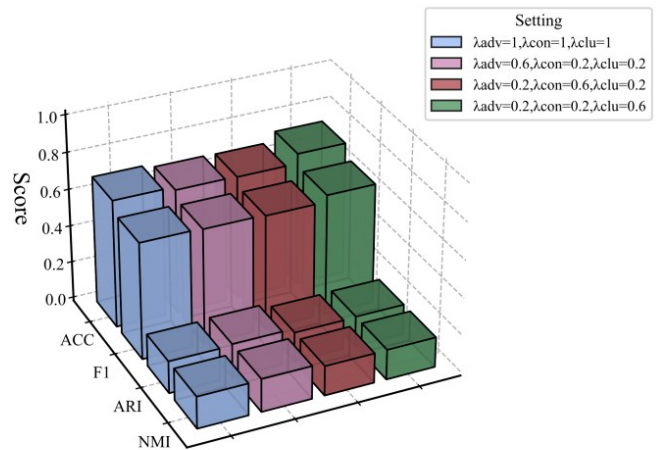


Fig. 3 Model performance under different weight configurations

图3 不同权重配置下的模型性能

4.8 时间效率分析

为评估 CGAE-GC 的计算效率,对其时间复杂度进行分析。给定一个包含 $|V|$ 个节点和 $|E|$ 条边的输入图 $\mathcal{G}$, 输入特征维度为 F , 隐层特征维度为 d , 图编码器层数为 L , 聚类簇数为 K , 训练样本规模为 N 。CGAE-GC 的单次计算开销主要由3个部分组成:①生成补全阶段。该阶段开销主要由基于 GIN 的结构/属性编码器和基于 MLP 的生成器/判别器构成。对于 GIN 编码器,其核心为消息传递与聚合机制,计算复杂度与图中的边数成线性关系,即 $\mathcal{O}(L(|V| + |E|)d)$ 。生成器与判别器主要由全连接层构成,对每个节点独立操作,复杂度为 $\mathcal{O}(|V|d^2)$ 。因此,该阶段的复杂度为 $\mathcal{O}(L(|V| + |E|)d + |V|d^2)$;②图嵌入学习阶段。该阶段采用 Graph Transformer 作为共享编码器对增强视图进行处理,自注意力机制(Self-Attention)计算节点间相关性的复杂度为 $\mathcal{O}(|V|^2d)$, 对比损失计算样本两两相似度的复杂度为 $\mathcal{O}(N^2d)$ 。因此,该阶段的复杂度由 $\mathcal{O}(|V|^2d)$ 主导;③联合优化阶段。该阶段负责计算图级嵌入与聚类中心的软分配概率及 KL 散度,复杂度为 $\mathcal{O}(N \cdot K \cdot d)$ 。与图神经网络计算开销相比,其参数更新过程几乎可以忽略不计。综上所述,CGAE-GC 的总体时间复杂度为 $\mathcal{O}(L(|V| + |E|)d + |V|^2d + |V|d^2)$ 。

5 结语

本文针对图级聚类中存在的节点属性缺失问题, 提出一种端到端的对比感知生成式补全聚类网络 CGAE-GC, 通过协同优化对抗生成、对比图嵌入与聚类目标, 有效缓解了属性缺失对聚类性能的影响, 并在多个领域数据集上展现出优秀的分类性能与鲁棒性。然而, 当前研究主要基于随机缺失假设构建补全机制, 而缺失本身可能蕴含着特定的分布偏置和语义信息。因此, 未来工作将致力于探索面向非随机缺失机制的自适应补全策略, 通过挖掘缺失模式背后的潜在关联, 进一步提升模型在复杂环境下的泛化能力与应用价值。

参考文献:

- [1] ZHANG H, XU H, LIN T E, et al. Discovering new intents with deep aligned clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 14365–14373.
- [2] WANG L, CHENG X Q. Dynamic community discovery and evolution in online social networks[J]. Chinese Journal of Computers, 2015, 38(2): 219–237.
王莉, 程学旗. 在线社会网络的动态社区发现及演化[J]. 计算机学报, 2015, 38(2): 219–237.
- [3] LI M. Financial investment risk prediction under the application of information interaction firefly algorithm combined with graph convolutional network[J]. PLoS One, 2023, 18(9): e0291510.
- [4] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905.
- [5] VISHWANATHAN S V N, SCHRAUDOLPH N N, KONDOR R, et al. Graph kernels[J]. The Journal of Machine Learning Research, 2010, 11: 1201–1242.
- [6] SHERVASHIDZE N, SCHWEITZER P, VAN LEEUWEN E J, et al. Weisfeiler-lehman graph kernels[J]. Journal of Machine Learning Research, 2011, 12(9): 2539–2561.
- [7] SMITH J, TU W, WU J, et al. Dual boost-driven graph-level clustering network[DB/OL]. <https://arxiv.org/abs/2504.05670>.
- [8] CAI J, HAN Y, GUO W, et al. Deep graph-level clustering using pseudo-label-guided mutual information maximization network[J]. Neural Computing and Applications, 2024, 36(16): 9551–9566.
- [9] HAN R, YAO G, ZHANG W, et al. Multi-relation graph-kernel strengthen network for graph-level clustering[DB/OL]. <https://arxiv.org/abs/2504.01605>.
- [10] HU M, CHEN C, LIU W, et al. Learning uniform clusters on hypersphere for deep graph-level clustering[DB/OL]. <https://arxiv.org/abs/2311.13953>.
- [11] JU W, GU Y, CHEN B, et al. Glcc: a general framework for graph-level clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 4391–4399.
- [12] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of Berkeley Symposium on Mathematical Statistics & Probability, 1965: 281–297.
- [13] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395–416.
- [14] CHEN Z Y, WANG X, ZHOU J F, et al. An efficient method for computing graph edit distance[J]. Intelligent Computer and Applications, 2020, 10(12): 94–98.
- 陈梓扬, 王臻, 周军锋, 等. 一种高效的图编辑距离计算方法[J]. 智能计算机与应用, 2020, 10(12): 94–98.
- [15] SUN F Y, HOFFMANN J, VERMA V, et al. Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization[DB/OL]. <https://arxiv.org/abs/1908.01000>.
- [16] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing Systems, 2020, 33: 5812–5823.
- [17] ZHU Y H, KONG B, ZHOU L H, et al. Graph contrastive learning guided multi-view clustering network[J]. Journal of Computer Applications, 2024, 44(10): 3267–3274.
朱云华, 孔兵, 周丽华, 等. 图对比学习引导的多视图聚类网络[J]. 计算机应用, 2024, 44(10): 3267–3274.
- [18] XIA R, LIU H, LI A, et al. Incomplete graph learning: a comprehensive survey[J]. Neural Networks, 2025, 190: 107682.
- [19] SHEN X, SHI L, GONG X, et al. Unsupervised deep graph structure and embedding learning[C]//IJCAI International Joint Conference on Artificial Intelligence, 2024: 2342–2350.
- [20] WEN J, LIU C, XU G, et al. Highly confident local structure based consensus graph learning for incomplete multi-view clustering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 15712–15721.
- [21] ZHANG X, PU X, CHE H, et al. Two-step graph propagation for incomplete multi-view clustering[J]. Neural Networks, 2025, 183: 106944.
- [22] DONG Z, JIN J, XIAO Y, et al. Subgraph propagation and contrastive calibration for incomplete multiview data clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(2): 3218–3230.
- [23] BRAUN G, TYAGI H, BIERNACKI C. Clustering multilayer graphs with missing nodes[C]//International Conference on Artificial Intelligence and Statistics, 2021: 2260–2268.
- [24] ZHANG Z, XIE J, CHEN X, et al. Dynamic deep graph learning for incomplete multi-view clustering with masked graph reconstruction loss[DB/OL]. <https://arxiv.org/abs/2511.11181>.
- [25] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[DB/OL]. <https://arxiv.org/abs/1810.00826>.
- [26] YUN S, JEONG M, KIM R, et al. Graph transformer networks[DB/OL]. <https://arxiv.org/abs/1911.06455>.
- [27] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[DB/OL]. <https://arxiv.org/abs/1807.03748>.
- [28] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//International Conference on Machine Learning, 2016: 478–487.
- [29] PASS-LANNEAU A, IGARASHI A, MEUNIER F. Perfect graphs with polynomially computable kernels[J]. Discrete Applied Mathematics, 2020, 272: 69–74.
- [30] ZHUO J, LU Y, NING H, et al. Unified graph augmentations for generalized contrastive learning on graphs[J]. Advances in Neural Information Processing Systems, 2024, 37: 37473–37503.
- [31] HU M L, CHEN C C, LIU W M, et al. Learning uniform clusters on hypersphere for deep graph-level clustering[DB/OL]. <https://arxiv.org/abs/2311.13953>.
- [32] CAI J, ZHANG Y, FAN J, et al. Dual contrastive graph-level clustering with multiple cluster perspectives alignment[C]//Proceedings of the 33rd International Joint Conference on Artificial Intelligence, 2024: 3770–3779.

(责任编辑: 尹晨茹)