GLCC:AGeneral Framework for Graph-Level Clustering
202200130267 张宇欣

1. 问题陈述

由于图结构的复杂性和多样性，现有的方法往往在适应不同类型的图数据时表现有限。

传统的图聚类方法通常是节点级的聚类，在多个图上的聚类（也称为图级聚类）在很大程度上还没有被探索，同时难以在图数据的无监督和半监督任务中同时获得良好的效果。

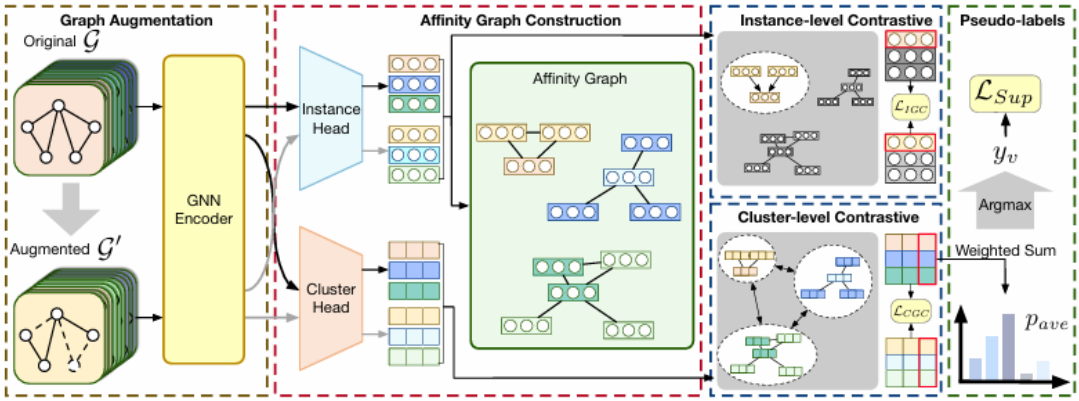
现有的图级聚类方法在处理图嵌入表示时，通常只关注部分特征，难以有效区分不同类型的图。当存在必须同类或不同类的图对时，如何在图聚类中有效应用这些约束信息也是一个问题。

2. 方法概述

作者提出的方法是什么？如果论文有模型图或算法图，可以截图放在笔记中，并用自己的话简单描述。
是否基于前人的方法？如果是的话，基于哪些方法？

作者提出了 GLCC 框架，希望通过灵活的图表示学习、自适应的聚类机制以及有效的约束条件应用，设计一个通用的、能够适应多种类型图数据的图级聚类方法。这种方法不仅在不同数据集上具有较好的泛化能力，同时也能在无监督和半监督情境中获得稳定的聚类效果。

GLCC 首先构建自适应亲和度图来链接语义相似的样本，然后在该亲和度图的基础上引入两级对比学习。一方面，GLCC 进行对比学习并结合图拉普拉斯算子，从实例级别的角度学习聚类友好的表示。另一方面，GLCC 鼓励样本与其邻居之间的一致性，以从集群级别视图捕获紧凑的集群表示。此外，使用预测的邻居感知伪标签来优化表示学习过程。这两个步骤可以交替训练以相互协作。



第一步：图增强

对输入的图数据进行数据增强操作，在生成图嵌入时创建多个视图，从而帮助模型通过对比学习的方式学到更稳健的特征表示。使用图嵌入方法将每个图实例转换为一个低维嵌入

向量。这通过 GCN 来实现，使得每个图的特征和结构信息都被编码在嵌入向量中。

第二步：亲和图构建

上一步生成的图嵌入向量共同输入到 instance head 和 cluster head 中，Instance Head 接收图嵌入向量并生成细粒度实例嵌入，用于个体图实例的相似性建模。Cluster Head 接收相同的图嵌入向量，生成更全局的簇级嵌入，用于增强聚类结构。

Instance head 的输出嵌入表示会被传递到实例级对比学习模块，优化个体图实例的表征。Cluster head 的输出嵌入表示会被传递到簇级对比学习模块，通过簇中心增强簇结构的清晰度。

instance head 的输出还用于构建 GLCC 框架中的亲和图，通过计算图实例之间的相似性来生成亲和矩阵 A，其中 A_{ij} 表示图实例 G_i 和 G_j 的相似度。使用亲和矩阵 A 来构建稀疏的亲和图 G_A，保留相似性高的邻接关系。例如，对于每个图实例，选择最相似的 k 个邻居，将这些邻居节点作为亲和图中的边。

第三步：实例级对比学习与簇级对比学习

实例级对比学习通过构造正样本对和负样本对来优化嵌入表示。GLCC 采用对比学习来提高图嵌入的区分性。通过图增强技术，生成同一个图的不同视图作为正样本对，确保模型学到的一致表示不依赖特定的增强方式。不同的图实例之间则作为负样本对，以使模型能够区分不同的图。使用对比损失L_{igc}优化嵌入向量，使正样本对之间的相似度最大化，负样本对之间的相似度最小化。

$$\mathcal{L}_{IGC} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{L_{ij}<0} -L_{ij}e^{h_i\cdot h_j'/\tau}}{\sum_{L_{ij}=0} e^{h_i\cdot h_j'/\tau}}\right).$$

簇级对比学习不仅关注单个实例的区分性，还在簇的层面上进行对比。在获得图嵌入后，先使用初步的聚类方法将图实例划分为多个簇。然后计算每个簇的中心向量（即簇中心）作为该簇的表示。簇级对比学习中，来自同一个簇的样本以及该簇中心作为正样本对，增强簇内一致性。来自不同簇的样本或簇中心作为负样本对，增强簇间的区分度。使用对比损失L_{cgc}将同簇样本之间的相似度最大化，不同簇之间的相似度最小化，从而在簇的层面上优化嵌入表示。这种方法进一步提高了聚类的准确性和稳健性。

$$\mathcal{L}_{CGC} = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{z_i\cdot z_i'/\tau}}{\sum_{j=1}^{K} e^{z_i\cdot z_j'/\tau}}\right) - H(\mathbf{Z}),$$

其中 H(·)是熵函数，以防止坍缩为同一集群的平凡输出。

第四步：伪标签生成

为未标注的数据自动生成标签的，为无监督聚类提供更多监督信息。GLCC 通过伪标签来指导模型的训练，进一步提升聚类质量。对图实例进行一次初步聚类，为每个图分配一个初始簇标签，根据初步聚类的结果，将每个图实例的簇标签作为伪标签，用伪标签生成的交叉熵损失来训练模型，指导模型聚合同类样本。伪标签作为近似的监督信号，进一步优化嵌入表示并提升聚类的精确性。

算法图：

作者使用的前人方法：

1.对比学习

GLCC 结合了实例级（instance-level）和簇级（cluster-level）对比学习，将对比学习扩展到簇层次，以捕捉聚类结构。将对比学习扩展到簇层次，以捕捉聚类结构。

2. 谱聚类

GLCC 利用亲和图来表征图实例之间的相似性，并以此作为聚类的基础。

3. 伪标签

GLCC 通过初步聚类生成伪标签，用于引导簇级对比学习。簇中心作为指导信号，使得簇内样本更加集中，簇间边界更加清晰。

4. 图增强技术

GLCC 采用图增强生成多视图样本，构建实例级和簇级对比学习的正样本对，确保模型在噪声条件下仍能生成高质量的嵌入。

5. 多投影头架构

作者设计了实例级和簇级两种投影头，分别优化个体图实例和簇的特征表示。

3. 实验

作者使用了哪些数据集？数据集是否公开？链接是什么？

作者是否提供了代码？链接是什么？

主要的 baselines 有什么？

实验的效果如何？可以截取论文的主要实验结果，并简单描述。

在两类数据集上进行了广泛的实验：生化分数据集和社交网络数据集。在生化分子数据集上，采用 TU 数据集的 DD 数据集，并在 AnchorQuery 平台上构建了 AnchorQuery-10K 和 AnchorQuery-25K 数据集。对于社交网络数据集，采用 TU 数据集中的 IMDB B、REDDIT-B 和 REDDIT-12K 数据集。

链接：AnchorQuery™ Web Page。 Welcome | TUDataset

将 GLCC 与两组 baseline 进行比较：图核方法与图对比学习方法。图核方法包括 Graphlet 核(Shervashidze et al. 2009)、最短路 径（SP）核(Borgwardt andKriegel 2005)和 Weisfeiler Lehman（WL）核(Shervashidze et al. 2011)。图对比学习方法有 InfoGraph(Sunet al.2020)、GraphCL(You et al. 2020)、CuCo(Chuet al. 2021)、JOAO(Youet al. 2021)、RGCL(Li et al. 2022)和 SimGRACE(Xia et al. 2022)。这些 baseline 首先学习图表示，然后利用 K-means (MacQueen1967)根据图表示对实例进行聚类。值得注意的是，缺少用于表示学习和聚类联合学习的图级聚类方法。

实验效果：

总的来说， GLCC 框架在所有六个数据集上与其他图核方法和图对比学习方法相比取得了最佳性能。特别是，在 ACC 方面，GLCC 在 IMDB B 的 8.2%和 REDDIT-B 的 9.9%上优于最接近的竞争对手，这证明了 GLCC 框架在图级聚类方面的卓越能力。

核方法在生物医学数据集上的表现通常比图对比学习方法差，但在社交数据集上的情况正好相反。这可能是因为传统的图核难以通过手工子结构来捕捉分子的官能团，而擅长探索社交网络数据集中关系连接的路径信息。

图对比学习方法的性能通常低于我们在 IMDB-B 和 REDDIT-B 数据集上的 GLCC，这表明只有实例级别的对比学习无法学习到聚类的有效表示。换句话说，簇级别的对比学习对于图级别的聚类至关重要。

对于聚类数量较大的数据集，如 REDDIT 12K 和 AnchorQuery-25K，GLCC 也表现出了比所有强基线的优越性，这证明了所提出框架对各种聚类数量的鲁棒性

Dataset	DD			AnchorQuery-10K			AnchorQuery-25K			IMDB-B			REDDIT-B			REDDIT-12K		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
Graphlet	0.004	0.579	-0.001	0.053	0.192	0.025	0.01	0.136	0.009	0.020	0.583	0.026	0.001	0.502	0.000	0.073	0.187	-0.004
SP	0.003	0.585	0.001	0.048	0.169	0.016	0.152	0.156	0.050	0.035	0.567	0.044	0.021	0.577	0.022	0.062	0.204	0.005
WL	0.006	0.586	0.002	0.033	0.162	0.010	0.159	0.163	0.068	0.023	0.53	0.003	0.089	0.576	0.021	0.092	0.189	0.044
InfoGraph	0.008	0.558	-0.006	0.072	0.238	0.036	0.178	0.181	0.061	0.041	0.538	0.005	0.016	0.508	0.000	0.045	0.205	0.003
GraphCL	0.019	0.573	-0.009	0.074	0.239	0.037	0.195	0.201	0.074	0.046	0.545	0.008	0.033	0.519	0.001	0.096	0.181	0.021
CuCo	0.012	0.562	-0.010	0.072	0.238	0.038	0.194	0.189	0.073	0.001	0.507	0.000	0.018	0.510	0.000	0.003	0.192	0.002
JOAO	0.012	0.578	-0.004	0.069	0.235	0.033	0.197	0.205	0.076	0.042	0.543	0.008	0.034	0.520	0.001	0.003	0.183	0.001
RGCL	0.014	0.565	-0.009	0.063	0.214	0.028	0.190	0.182	0.059	0.047	0.546	0.007	0.017	0.509	0.001	0.003	0.092	0.001
SimGRACE	0.001	0.589	0.003	0.068	0.226	0.031	0.189	0.186	0.074	0.049	0.559	0.007	0.024	0.513	0.001	0.062	0.210	0.005
Ours	**0.024**	**0.607**	**0.023**	**0.076**	**0.247**	**0.043**	**0.209**	**0.228**	**0.083**	**0.081**	**0.665**	**0.106**	**0.092**	**0.676**	**0.087**	**0.105**	**0.226**	**0.058**

Table 2: 在6个图属性预测基准上的聚类性能。最好的结果以粗体显示。

通过消融实验，可以得出无论是缺乏实例级对比还是集群级对比，都会损害集群的性能， 这表明图表示学习和聚类可以相互促进和受益。亲和度图提供的额外邻居信息有利于学习更好的聚类分配。此外，紧凑邻居能够更好地增加簇内信息，更好地服务于聚类过程。邻居感知的伪标记机制是必不可少的，并可以进一步提高 1.0%在两个数据集上的聚类精度，这表明通过聚类分配伪标记提供的有效监督信号可以反过来优化表示学习。

	Correlations				IMDB-B			AnchorQuery-25K		
	IGC	CGC	AAG	NPL	NMI	ACC	ARI	NMI	ACC	ARI
M_1	√				0.046	0.545	0.008	0.195	0.201	0.074
M_2		√			0.059	0.631	0.071	0.191	0.206	0.069
M_3	√	√			0.063	0.640	0.084	0.199	0.210	0.077
M_4	√	√	√		0.068	0.653	0.092	0.202	0.217	0.079
M_5	√	√	√	√	**0.081**	**0.665**	**0.106**	**0.209**	**0.228**	**0.083**

Table 3: IMDB-B和AnchorQuery-25K数据集的消融研究分析。IGC、CGC、AAG和NPL分别对应实例级图对比度、簇级图对比度、自适应亲和度图和邻居感知伪标签。

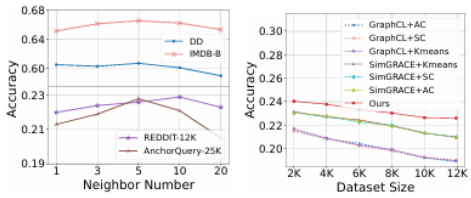


Figure 2: 在四个数据集上的性能*w.r.t.*邻居数和REDDIT-12K上的数据集大小。

经过敏感性分析，邻居数量增加时，性能逐渐提高。然而过大的近邻个数会影响对比学习的性能，因为随着距离的增大，两个样本属于同一类别的概率会降低，给对比学习带来噪声。结果显示随着数据集大小的增长，GLCC 的性能下降得比其他方法慢。表明其对各种簇样本大小的鲁棒性。

4. 思考

论文的主要贡献是什么？
论文是否存在缺陷？具体有哪些？
对未来的研究方向有什么想法？

论文介绍了一个通用框架 GLCC，用于给定多个图的图级聚类，它捕获了多粒度信息，以提供图实例的全局特征。首先构建一个自适应的亲和度图来链接语义相似的样本，然后基于亲和度图引入实例级和聚类级的对比学习。预测邻域感知伪标签来优化表示学习过程。在一系列著名的基准数据集上的广泛实验证明 了 GLCC 对于图级聚类的有效性

构建自适应亲和图和双层对比学习需要较高的计算资源，尤其在大型数据集上可能会遇到性能瓶颈。模型效果在一定程度上依赖亲和图的质量，而图结构质量不佳时可能会影响整体性能。方法虽然在 TU 数据集等上表现优异，但缺少在更大规模或真实世界复杂图数据集上的测试，泛化能力未完全验证。

对于未来的研究，可以进一步研究 GLCC 在异质图（节点和边类型多样）的表现，增强方法的通用性。探索更高效的亲和图构建方法，减轻计算负担，支持大规模数据集的应用。在

弱监督或半监督场景中，研究如何引入更少的标签数据来优化 GLCC 方法。应用方面，将 GLCC 应用于化学分子分析、生物网络聚类或社交网络分组等具体任务，验证其实用性。

5. 其它（选填）

需要特别记录的其它笔记