计算机科学
COMPUTER SCIENCE

基于动态阈值伪标签筛选的深度图对比聚类算法

王沛, 杨希洪, 管仁祥, 祝恩

引用本文

王沛, 杨希洪, 管仁祥, 祝恩. 基于动态阈值伪标签筛选的深度图对比聚类算法[J]. 计算机科学, 2025, 52(8): 100-108.

WANG Pei, YANG Xihong, GUAN Renxiang, ZHU En. Deep Graph Contrastive Clustering Algorithm Based on Dynamic Threshold Pseudo-label Selection [J]. Computer Science, 2025, 52(8): 100-108.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700112

基于动态阈值伪标签筛选的深度图对比聚类算法

王 沛 杨希洪 管仁祥 祝 恩
国防科技大学计算机学院 长沙 410073
(wangpei@nudt.edu.cn)

摘 要 近年来,图神经网络在处理复杂结构数据方面表现出色,被广泛应用于节点分类、图分类、链接预测等领域。深度图聚类结合了GNNs强大的表示能力与聚类算法的目标,从复杂的图结构数据中发现隐藏的簇结构。然而,现有的基于伪标签的图聚类算法在进行模型优化时常使用固定阈值,根据类别对样本进行筛选,以获得高置信度的样本数据来引导模型优化。但固定阈值的方法会导致类别不平衡问题,进而影响模型聚类的性能。为了解决上述问题,提出了一种基于动态阈值伪标签的深度图对比聚类算法。具体来说,采用两个不共享参数的多层感知机(MLP)结构捕捉图数据的潜在结构特征,并使用K-Means算法得到聚类结果。在此基础上,引入信赖强度来动态调整获得伪标签的阈值,在训练过程中动态调整每个类别中高置信度的样本数量,缓解类别不平衡的问题。此外,优化了对比学习策略,改进了样本对的构造方法,提高了模型的判别能力。实验结果表明,所提方法在6个基准数据集上均表现出色,在多个评估指标上超越了现有方法,展现了其有效性。

关键词:深度图聚类;伪标签;图对比聚类;图神经网络;动态阈值
中图分类号 TP391

Deep Graph Contrastive Clustering Algorithm Based on Dynamic Threshold Pseudo-label Selection

WANG Pei,YANG Xihong,GUAN Renxiang and ZHU En
College of Computer Science and Technology,National University of Defense Technology,Changsha 410073,China

Abstract In recent years,graph neural networks have performed well in processing complex structural data,and are widely used in node classification,graph classification,link prediction and other fields. Deep graph clustering combines the powerful representation ability of GNNs with the goal of clustering algorithms to discover hidden population structures from complex graph structure data. However,the existing pseudo-label-based graph clustering algorithms often use fixed thresholds to filter samples according to categories to obtain high-confidence sample data to guide model optimization. However,the method of fixed thresholds can lead to category imbalance,which in turn affects the performance of model clustering. In order to solve the above problems,this paper proposes a contrastive clustering algorithm based on dynamic threshold pseudo-label depth map. Specifically,two multilayer perceptron(MLP) structures that do not share parameters are used to capture the latent structural features of the graph data,and the K-Means algorithm is used to obtain the clustering results. On this basis,the trust strength is introduced to dynamically adjust the threshold for obtaining pseudo-labels,and the number of high-confidence samples in each category is dynamically adjusted during the training process to alleviate the problem of category imbalance. In addition,this paper optimizes the contrastive learning strategy,improves the construction method of sample pairs,and improves the discriminant ability of the model. Experimental results show that the proposed method performs well on the six benchmark datasets,surpassing the existing methods in multiple evaluation indicators,and strongly demonstrates the effectiveness of the proposed algorithm.

Keywords Deep graph clustering,Pseudo-label,Graph contrastive clustering,Graph meural network,Dynamic threshold

1 引言

图神经网络(Graph Neural Networks,GNNs)[1-2]因其强大的表达能力和对复杂结构数据的适应性,近年来在多个领域展现出巨大的潜力。GNNs通过图的结构信息进行节点和边的嵌入表示,其多层传播机制可以同时捕获图的局部和全局特征,尤其在聚类[3-4]这样的无监督学习[5]任务中,GNNs能够利用未标记数据改进聚类结果,展现了其在复杂数据表示和聚类方面的巨大优势。

深度图聚类[6-7]作为一种先进的数据挖掘技术,结合了图神经网络的强大表示能力与聚类算法的目标,可以从复杂的图结构数据中发现隐藏的群体结构,已经成为一个热门的研究领域。深度图聚类的核心思想是通过深度学习技术自动提取图数据的特征表示,然后基于这些表示进行聚类分析。与传统的聚类方法相比,深度图聚类能够通过图神经网络更有效地捕捉图数据中的复杂结构和节点特征信息,实现更精

到稿日期:2024-07-17 返修日期:2024-10-25
基金项目:科技创新2030(2022ZD0209103)
This work was supported by the National Science and Technology Innovation 2030 Major Project(2022ZD0209103).
通信作者:祝恩(enzhu@nudt.edu.cn)

细和准确的聚类结果[8-9]。

近年来，伪标签技术[10-13]在深度图聚类中扮演着十分重要的角色。利用伪标签技术不仅可以增强模型对无标签数据的理解能力，还可以促使模型学习更加鲁棒和更具区分性的特征表示。现有的基于伪标签的图聚类算法通常采用迭代过程，模型首先根据当前特征进行聚类，然后基于聚类结果生成伪标签，再用这些伪标签反向优化模型参数。这一正向反馈机制每次迭代都在优化特征表示的同时改进了聚类质量，使得伪标签更加准确，形成良性循环。

尽管基于伪标签的深度图聚类技术取得了显著进展，但仍面临以下挑战。1）目前基于伪标签优化的深度图对比聚类算法根据聚类得到的类别使用伪标签指导模型学习时，通常使用统一的固定阈值来选取高置信度的样本[14]。这种方法忽略了不同类别间可能存在的显著差异性，导致某些稀有类别的样本被过度筛选，产生类别不平衡的问题，从而难以充分挖掘和利用无标签数据中的潜在类别信息。2）在对比学习策略的设计中，现有的方法常使用图数据增强（如随机节点/边删除、随机属性屏蔽）来构造正负样本对，忽略了对下游聚类结果产生的类别信息的利用，从而导致构造的正负样本对的质量不高。

基于上述问题，本文提出了一种基于动态阈值伪标签的深度图对比聚类方法。具体来说，本文首先通过设计不共享参数的编码器获取图的嵌入表示。在此基础上，通过结合两种视图下的节点嵌入信息，生成综合嵌入矩阵，并使用 K-Means 算法进行聚类。为了在网络训练过程中动态调整阈值以获得高置信度的伪标签指导模型的优化，本文引入信赖强度作为置信度衡量标准，通过判断不同类别的学习难度来动态调整阈值，选择高置信度样本指导模型学习，形成正向反馈机制。此外，在对比学习正负样本对的构造过程中，本文引入边缘负样本选取策略，将不同视图下不同类别的样本视为负样本对，同时通过选择距离最近的部分边缘负样本对，有效增大不同类别样本之间的间距，缓解了样本多样性不足的问题，在扩充了负样本数量的同时减少了信息的压缩和丢失。此外，选择不同视图下的同一类别内的样本作为正样本，提高了构造样本对的质量，进而提升了模型的性能。

综上所述，本文的主要贡献如下：

1）提出了一种根据聚类结果动态调整伪标签阈值的方法。引入强度信赖分值作为置信度衡量标准，动态调整阈值，从而选择高置信度样本优化网络训练。

2）本文通过聚类结果来优化图对比学习中正负样本对的构造，提高了样本对质量，减少了信息的压缩和丢失，使模型接触学习到更多样本，从而提高了模型的性能和判别能力。

3）在 6 个基准数据集上进行了大量实验，结果表明，本文方法在所有数据集上表现出色，在多个评估指标上超过了其他方法，展现出了强大的鲁棒性和广泛的适用性。

2　相关工作

2.1　深度图聚类

深度图聚类是一种利用深度学习技术对图数据进行聚类的方法。它通过学习图数据的表示来实现对图结构和节点特征的聚类，从而将具有相似特征或结构的节点归为一类。近年来，随着深度学习在图数据处理中的应用不断深入，深度图聚类受到了研究者的广泛关注，并在社交网络分析、生物信息学、推荐系统[15]等领域展现出巨大的潜力。深度图聚类的基本思想是利用深度神经网络自动提取图数据的高级特征表示，并在此基础上进行聚类分析。与传统的图聚类方法相比，深度图聚类方法不仅能够捕捉图数据的复杂结构信息，还能充分利用节点的属性信息，实现更精细和准确的聚类结果。该方法主要包括图表示学习、聚类目标函数设计等几个关键步骤。

在深度图聚类方法的发展过程中，研究者提出了多种具有代表性的方法。例如，GAE 和 VGAE[16]利用图自编码器对图数据进行无监督表示学习，实现了高效的图聚类；DAEGC[17]采用注意力机制提高了聚类性能；DGI[18]通过最大化全局图嵌入与局部节点特征之间的互信息，提高了图嵌入的质量，从而提升了聚类性能；GRACE[19]则通过结合对比学习和数据增强，构建和优化对比目标来生成鲁棒且有效的图嵌入；GCC[20]通过对比学习方法，最大化不同视图之间的互信息，生成高质量的图嵌入，有效地解决了图聚类中的语义漂移问题。

2.2　对比学习

对比学习[21-22]是一种通过构建正负样本对来优化模型嵌入表示的技术。它通过最大化正样本对之间的相似度和最小化负样本对之间的相似度，来提升模型的判别能力和表示质量。近年来，对比学习在深度学习和图数据处理中受到了广泛关注。与传统的监督学习不同，对比学习不依赖于大量标注数据，而是通过自监督的方式来学习有效的特征表示。

图对比学习[23]是一种专门用于处理图数据的对比学习技术。由于图数据具有复杂的结构和节点属性，因此图对比学习不仅关注节点和边的特征，还关注图的全局结构信息。该方法主要包括全局和局部特征对比、视图间对比，以及多尺度对比等。

图对比学习技术在深度图聚类研究中发挥着重要作用[1]。具体而言，现有的深度图对比聚类方法通过构建不同的视图来生成正负样本对，并进行对比学习。例如，DGI[18]将图的局部表示与对应的全局表示作为正样本对，通过扰动图结构或属性生成负样本，将这些负样本与全局表示作为负样本对；GRACE[19]通过自适应增强策略，从原始图中生成两个不同的视图，经过编码处理后，将来自同一个节点在两个视图中的表示作为正样本对，将来自不同节点在两个视图中的表示作为负样本对。

与上述方法不同的是，某些研究，如 GCC[20]，通过对比学习方法最大化不同视图之间的互信息，从而生成高质量的图嵌入。这些方法在聚类任务上取得了一定效果，但其性能在一定程度上都取决于所设计的数据增强策略，在进行节点丢弃或边缘扰动时，可能会丢失一些重要的信息，特别是对关键节点或关键边的扰动，可能会对最终的图表示产生负面影响。本文通过设计不共享参数的编码器构建新的视图来避免语义信息的改变，同时将动态阈值伪标签方法和对比学习相结合，利用高置信度的伪标签作为监督信息指导模型的学习。考虑聚类结果中存在类别不平衡的问题，不同于传统固定阈值的方法，本文利用聚类结果动态调整阈值，充分利用无标签数据。本文模型框架如图 1 所示。

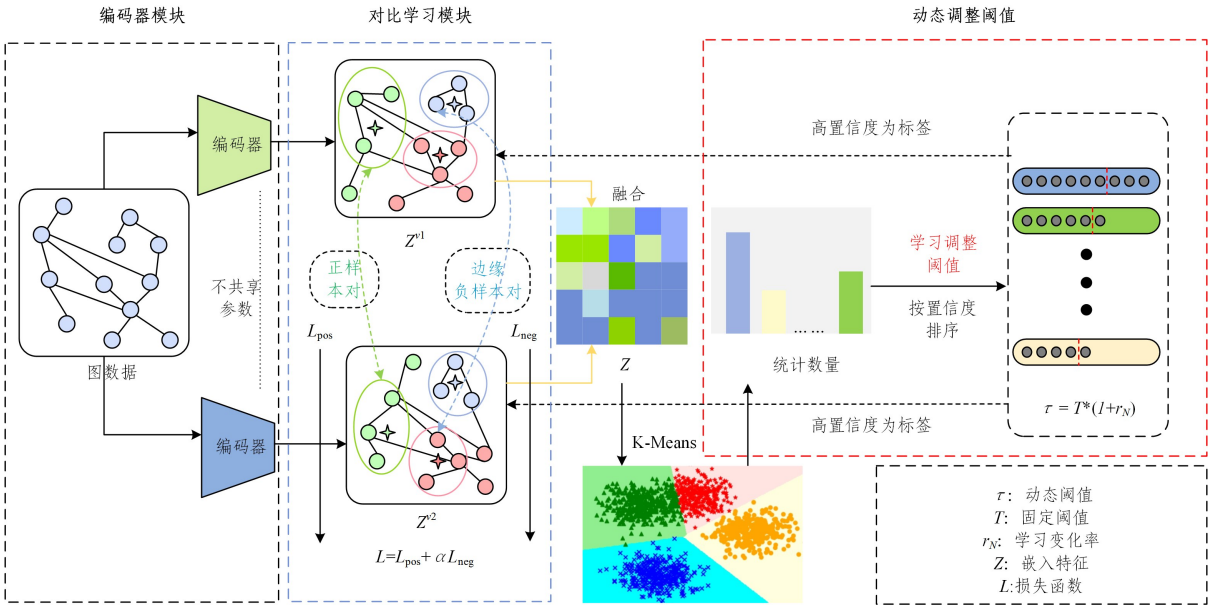


图 1 本文模型框架图

Fig. 1 Overall framework of the proposed model

3 本文方法

3.1 符号描述

为了方便后续描述，本文使用 $G=(\boldsymbol{X},\boldsymbol{A})$ 来表示一个无向图，用 $V=\{v_1,v_2,\cdots,v_N\}$ 表示具有 N 个节点的集合，E 为边的集合，G 中包含 K 类数据。令 $\boldsymbol{X}\in R^{N\times D}$ 表示属性矩阵，$\boldsymbol{X}\in R^{N\times N}$ 表示原始的邻接矩阵，度矩阵表示为 $\boldsymbol{D}=diag(d_1,d_2,\cdots d_N)\in R^{N\times N}$，此外图 G 上的对称规范化拉普拉斯矩阵表示为 $\widetilde{\boldsymbol{L}}=\mathbf{I}-\widetilde{\boldsymbol{D}}^{-1/2}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-1/2}$，其中 $\widetilde{\boldsymbol{A}}=\boldsymbol{A}+\mathbf{I}$，$\mathbf{I}$ 为单位矩阵。

3.2 编码器模块

基于 GAE 得到的启发，本文使用单位矩阵 $\mathbf{I}$、对称规范化拉普拉斯矩阵 $\widetilde{\boldsymbol{L}}$，以及表示拉普拉斯滤波器层数的 K 来构建滤波器模型。本文中所涉及部分符号如表 1 所列。

表 1 符号说明

Table 1 Symbol definition

符号	含义
$\boldsymbol{X}$	属性矩阵
$\boldsymbol{A}$	邻接矩阵
$\mathbf{I}$	单位矩阵
$\widetilde{\boldsymbol{L}}$	规范化拉普拉斯矩阵
$\boldsymbol{Z}$	嵌入特征
τ	动态阈值

$$\widetilde{\boldsymbol{X}}=(\boldsymbol{I}-\widetilde{\boldsymbol{L}})^K\boldsymbol{X} \tag{1}$$

为增强图表示的鲁棒性和语义深度，本文引入两个结构相同但参数独立的多层感知机（Multilayer Perceptron，MLP），记为 MLP_1 和 MLP_2。它们以拉普拉斯滤波结果为输入，输出两个差异性的图嵌入表示，记为 $\boldsymbol{Z}^{v_1}$ 和 $\boldsymbol{Z}^{v_2}$。

$$\begin{aligned}\boldsymbol{Z}^{v_1}&=MLP_1(\widetilde{\boldsymbol{X}})\\ \boldsymbol{Z}^{v_2}&=MLP_2(\widetilde{\boldsymbol{X}})\end{aligned} \tag{2}$$

为了确保不同图嵌入之间的可比性，消除尺度效应并进行规范化输出，本文对产生的嵌入向量和进行了归一化处理。

$$\boldsymbol{Z}^{v_i}=\frac{\boldsymbol{Z}^{v_i}}{\|\boldsymbol{Z}^{v_i}\|_2},i\in\{1,2\} \tag{3}$$

借助于此策略，各特征维度的贡献得以平衡，进而促进了后续模型层对这些高质量图嵌入的有效应用。

3.3 动态阈值聚类伪标签

本节提出了一种根据聚类结果动态调整伪标签置信度阈值的方法。首先，基于现有的视图融合策略[25-26]，本文将两种视图下的节点嵌入信息进行融合，生成综合嵌入矩阵 $\boldsymbol{Z}$。

$$\boldsymbol{Z}=\frac{\boldsymbol{Z}^{v_1}+\boldsymbol{Z}^{v_2}}{2} \tag{4}$$

通过加权融合的方式生成的综合嵌入矩阵 $\boldsymbol{Z}$ 可以有效利用来自两种视图的信息，均衡不同视图之间的差异，减少信息损失，提高融合特征的多样性，降低过拟合风险。同时，差异性的图嵌入表示可以帮助提高模型的泛化能力。

使用 K-Means 算法对 $\boldsymbol{Z}$ 进行聚类并得到聚类结果。这里，引入信赖强度值 TS 作为置信度衡量标准。

$$TS_i=\exp(-\|\boldsymbol{Z}_i-\boldsymbol{C}_{p(i)}\|^2) \tag{5}$$

其中，$\boldsymbol{Z}_i$ 表示第 i 个节点的嵌入向量，$\boldsymbol{C}_{p(i)}$ 是与第 i 个样本关联的聚类中心。对于每一个数据点 i，根据其嵌入向量 $\boldsymbol{Z}_i$ 与对应聚类中心 $\boldsymbol{C}_{p(i)}$ 之间的接近程度，量化其归属的确信程度，从而确保聚类的可靠性和稳定性。接下来，使用动态阈值筛选出合适的高置信度样本和聚类伪标签。

本文根据每个类的模型学习状态动态调整阈值。最理想的方法是计算每个类的评估准确度来缩放阈值，这在无监督学习中是不可用的。因此，使用另一种方法来估计学习状态。假设一个类的学习效果可以由预测属于该类且信赖强度值在阈值范围内的样本数量来反映，即阈值范围内预测样本较少的类被认为具有较大的学习难度。基于此，各个类的动态阈值 τ 的计算式如下：

$$\alpha_t(c)=\frac{r_t(c)}{\sum_{i=1}^{k} r_t(i)} \tag{6}$$

$$\tau_t(c)=T*(1+\alpha_{t-1}(c))$$

其中，T 表示固定阈值；$r_t(c)$表示第 t 轮聚类时得到的所有高置信度样本中，数量最多的类别的高置信度样本数与第 c 类样本数的比值：

$$r_t(c)=\frac{\max(N_t(i),i=1,2,\cdots,k)}{N_t(c)},c=1,2,\cdots,k \tag{7}$$

其中，N 表示某类别的高置信度样本数。$r_t(c)$可以反映类别 c 在第 t 轮聚类时的学习困难度，其值越大表示学习该类别越困难。将其归一化后得到 $\alpha_t(c)$，避免原始权重差异过大导致的数值溢出或者下溢问题。将各类别的样本根据 TS 值排序后，选择位于各类别前 τ 位的样本作为高置信度样本：

$$S=\{s_1,s_2,\cdots,s_n,\cdots\} \tag{8}$$

其中，动态阈值 τ 使用 $1+\alpha$ 对固定阈值 T 进行放大，可以充分利用数据进行学习。对于较难学习的类，其动态阈值 τ 较大，对于数据的接受程度较高，因此可以学习到更多的样本，提高样本的利用率。而对于较易学习的类，则选择性地学习更高质量的样本。

本文提出了一种根据聚类结果动态调整置信度阈值的方法。相较于为所有类别设定统一的固定阈值，该方法可以根据模型的学习状态来利用无标签数据。其核心在于建立了一个反馈循环，通过监测模型的聚类性能和学习进展，动态调整用于生成伪标签的置信度阈值。具体而言，随着训练的深入，模型能够自动识别哪些类别需要更多的样本支持，哪些类别已经达到了较高的学习质量。对于前者，动态阈值 τ 会相应增加，鼓励模型探索并吸收更多边界或不确定的样本，以期改善其在这些类别上的表现。对于后者，阈值 τ 则会保持在一个较高的水平，确保只将最可靠的样本纳入学习，防止过拟合或引入噪声。

3.4　对比学习模块

基于 3.3 节中得到的高置信度样本索引和聚类伪标签，本文分别构建正负样本进行对比学习。

首先，识别并提取来自两个不同视图的高置信度样本。然后，基于预设的伪标签信息，将筛选出来的高置信度样本分别归类到 K 个不相交的簇中（K 为样本的类别数），形成 $Q_p^{v_1}$（$p=1,2,\cdots,K$）和 $Q_p^{v_2}$（$q=1,2,\cdots,K$）两个簇，每个簇代表一个特定的类别。接下来，在确保簇内部一致性的同时，从对应高置信度分类中挑选样本，通过计算不同视图下正样本嵌入之间的余弦距离之和表示正对比损失，利用高置信度的聚类标签指导模型学习。其过程如式(9)所示：

$$L_{\text{pos}}=\frac{1}{N_{\text{clusters}}}\sum_{i=1}^{N_{\text{clusters}}}\sum_{j=1}^{n_i}(2-2\langle \boldsymbol{G}_{i[j,:]}^{v_1},\boldsymbol{G}_{i[j,:]}^{v_2}\rangle) \tag{9}$$

其中，N_{clusters}表示类别数，n_i 表示第 i 个类别的高置信度样本数，$\boldsymbol{G}_{i[j,:]}^{v_1}$和 $\boldsymbol{G}_{i[j,:]}^{v_2}$分别表示在两个视图下的第 i 类别的第 j 个高置信度嵌入节点，$\langle \boldsymbol{G}_{i[j,:]}^{v_1},\boldsymbol{G}_{i[j,:]}^{v_2}\rangle$表示 $\boldsymbol{G}_{i[j,:]}^{v_1}$和 $G_{i[j,:]}^{v_2}$的内积。该设计将高置信度样本作为监督信息，提高了正样本对的质量。

为了缓解样本对构建过程中类别内部多样性和边界样本处理不足的问题，本文设计了边缘负样本策略来帮助模型更加精准地区分相近类别，增强其在复杂场景下的判别能力。

$$L_{\text{neg}}=-\frac{1}{\sum_{i,j}^{K} m_{ij}}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\sum_{(u,v)\in P_{ij}}(2-2\langle \boldsymbol{G}_{i[u,:]}^{v_1},\boldsymbol{G}_{j[v,:]}^{v_2}\rangle) \tag{10}$$

其中，m_{ij}表示类别 i 和类别 j 之间的最小样本数。L_{neg2}利用得到的高置信度样本集 $\boldsymbol{G}_i^{v_1}$（$i=1,2,\cdots,K$）和 $\boldsymbol{G}_j^{v_2}$（$j=1,2,\cdots,K$），首先计算类别 $\boldsymbol{G}_i^{v_1}$ 与 $\boldsymbol{G}_j^{v_2}$ 之间所有样本对的距离（$i\neq j$），形成距离矩阵 $\boldsymbol{D}_{ij}$，接着从距离矩阵 $\boldsymbol{D}_{ij}$中选择距离最小的 m_{ij}对样本，记作样本对集合 P_{ij}。(u,v)表示在 P_{ij}中的一个样本对，u 属于类别 $\boldsymbol{G}_i^{v_1}$，v 属于类别 $\boldsymbol{G}_j^{v_2}$。选择距离最近的样本对是由于其往往位于类别边界处，这些样本对可以更好地描述不同类别间的分界边界，通过优化这些边缘样本对，可以显著增强模型对类别边界的判别能力。将数量设置为 m_{ij}是为了确保选择的样本对数量可以根据不同类别之间的样本数量动态调整，从而避免样本数量不平衡导致的偏差。分母$\sum_{i,j} m_{ij}$用于标准化损失值，使其具有可比性。因此，本文通过选择距离最近的边缘负样本对并最大化它们的距离，可以增大不同类别之间的间距。这有助于提升模型的分类性能，在面对复杂数据时做出准确的判断。

算法 1　基于动态阈值伪标签筛选的深度图对比聚类算法

输入：输入图 G=$\{\mathbf{X},\mathbf{A}\}$；迭代次数 N

输出：聚类结果 R

1. 通过式(1)获得平滑的属性 $\widetilde{\mathbf{X}}$.
2. for i=1 to N do
3.　通过式(2)将 $\widetilde{\mathbf{X}}$ 编码为两个视图；
4.　通过式(3)归一化嵌入表示 $\mathbf{Z}^{v_1}$，$\mathbf{Z}^{v_2}$；
5.　通过式(4)融合 $\mathbf{Z}^{v_1}$ 和 $\mathbf{Z}^{v_2}$ 得到 $\mathbf{Z}$；
6.　在 Z 上执行 K-Means 以获得聚类结果；
7.　通过式(7)得到 $r_t(c)$，归一化后得到学习变化率 α；
8.　通过式(6)得到各类的动态阈值 τ；
9.　各类选择前 τ 位样本作为高置信度样本 S；
10.　分别构建正样本对和边缘负样本对；
11.　通过式(9)和式(10)计算对比损失；
12.　通过最小化式(11)中的 L 更新整个网络；
13. end for
14. 在 $\mathbf{Z}$ 上执行 K-Means 来获得最后的聚类结果；
15. 返回 R.

算法 1 给出了从图编码、伪标签生成、正负样本对构建到最终聚类结果输出的全过程。

3.5　损失函数

本文的损失函数主要包括正对比损失 L_{pos}和负对比损失 L_{neg}，总的损失函数如下：

$$L=L_{\text{pos}}+\theta L_{\text{neg}} \tag{11}$$

其中，θ 为平衡参数。

4　实验

4.1　实验设置

4.1.1　实验环境

本文使用 PyTorch 进行实验，训练总次数为 400 次，计算

10 次运行的平均值来避免实验运行的随机性。硬件设备条件为:NVIDIA GeForce RTX 2080Ti GPU 和 64 GB RAM。

4.1.2 数据集描述

本文在多个标准数据集上进行实验,以验证所提出方法的有效性和鲁棒性。这些数据集包括 CORA,AMAP,CITE,EAT,BAT 和 UAT。数据集信息如表 2 所列。

表 2 数据集信息

Table 2 Dataset information

数据集	类型	样本数	维度	边数	种类
CORA	Graph	2708	1433	5429	7
CITE	Graph	3327	3703	4732	6
AMAP	Graph	7650	745	119081	8
UAT	Graph	1190	239	13599	4
BAT	Graph	131	81	1038	4
EAT	Graph	399	203	5994	4

CORA 数据集是一个广泛用于文献分类的图数据集,包含 2708 篇科学出版物,每篇文献被标注为 7 个预定义类别中的 1 个。文献之间通过引用关系形成一个图结构,边数为 5429。每篇文献用一个包含 1433 个特征的稀疏向量表示,这些特征是词频向量。

AMAP[27] 数据集是一个用于生物信息学的图数据集,包含 3483 个样本和 5207 条边。这些样本代表不同的生物分子,每个样本被标注为 4 个预定义类别中的 1 个。每个样本用一个由 2089 个特征组成的稀疏向量表示,这些特征包括各种生物属性和特征。

CITE[28] 数据集是另一个用于文献分类的引用网络数据集,包含 3327 篇科学出版物和 4732 条引用关系。与 CORA 类似,每篇文献属于 6 个预定义类别中的 1 个。每篇文献的特征向量由 3703 维的词频向量组成。

EAT[29] 数据集是一个电商推荐系统的数据集,包含4000 个产品和 5800 条边。产品之间的边表示它们经常被一起购买,每个产品被标注为 6 个预定义类别中的 1 个,并用一个由 1000 个特征组成的稀疏向量表示,这些特征包括产品的描述、类别和销售数据。

BAT[29] 数据集是一个用于社交网络分析的图数据集,包含 5300 个节点和 7130 条边。节点代表不同的用户,每个用户被分类为 5 个预定义类别中的一个。每个节点用一个包含 1500 个特征的稀疏向量表示,这些特征包括用户的行为和兴趣数据。

UAT[29] 数据集是一个用于城市交通分析的数据集,包含 2700 个交通节点和 4800 条边。节点表示不同的交通站点,每个站点被标注为 3 个预定义类别中的 1 个。每个节点用一个包含 1200 个特征的稀疏向量表示,这些特征包括交通流量、地理位置和时间模式。

这 6 个数据集是常用的图机器学习数据集,其节点代表不同类型的信息,如文档、论文、交通节点等。每个节点都附带有特征,这些特征可以是词向量表示、关键词、摘要或其他描述性信息。边表示节点之间的连接关系。对于 CORA,CITE 和 AMAP 数据集,边表示文档之间的引用关系或作者之间的合作关系。而对于 BAT,EAT 和 UAT 数据集,边表示交通流量、道路连接或机场运营中的连接关系。这些数据集通常包含预定义的类别标签,用于节点分类任务。类别标签为监督学习任务提供了基础。

4.2 性能比较

为了验证本文方法的优越性,在上述 6 个数据集上与 9 种方法进行了对比,包括深度聚类方法 DEC[30] 和 DCN[31]、深度图聚类方法 MGAE[32],ARGA[33],AdaGAE[34],以及深度图对比聚类方法 GCA[35],AutoSSL[36],CONVERT[25] 和 MGCN[37]。结果如表 3 和表 4 所列。

根据对比结果,可以得出以下结论。

1)本文方法在 6 个数据集上均表现出色,在多个评估指标上超过了其他方法。在不同规模的数据集上,和目前的先进模型相比,本文方法都能取得了优异的表现,展示了强大的鲁棒性和广泛的适用性。

2)与深度图聚类算法 MGAE[32],ARGA[33],AdaGAE[34] 相比,本文方法获得了较大的性能提升。上述方法是常见的图自编码器模型,通过学习节点的压缩表示来实现特征抽取和降维,但由于缺乏设计良好的对比学习策略,其对图数据的挖掘能力有限。

3)深度图对比聚类算法 GCA,AutoSSL,CONVERT 和 MGCN 取得了次优性能,这是由于深度图对比聚类算法结合了深度学习和图结构分析的优势,通过高维表示学习、对比学习机制、自适应聚类策略和非线性变换等手段,显著提升了聚类性能。然而,为所有类别设定统一的固定阈值,无法进行充分的学习。本文在设置伪标签时,根据聚类结果为不同的类设置了动态阈值,在保证伪标签质量的同时增加伪标签的数量,提高了模型的判别能力。此外,先前的方法在处理负样本时往往采用简单的策略,如随机选择或使用固定的聚类中心,本文引入边缘负样本作为核心负样本的补充,使模型在学习过程中可以更好地理解和区分不同类别,从而提高了模型的鲁棒性和判别能力。

本文方法在多个指标表现出优于其他算法的性能。以 CORA 数据集为例,本文方法在该数据集的 ACC,NMI,ARI 指标上均优于次优方法,分别提升了 0.85 个百分点、2.01 个百分点、2.27 个百分点。

表 3 对比实验结果

Table 3 Comparison experiment results

methods	CORA				CITE				AMAP			
	ACC	*NMI*	*ARI*	*F1*	*ACC*	*NMI*	*ARI*	*F1*	*ACC*	*NMI*	*ARI*	*F1*
DCN	49.38	25.65	21.63	43.71	57.08	27.64	29.31	53.80	48.25	38.76	20.8	47.87
DEC	46.50	23.54	15.13	39.23	55.89	28.34	28.12	52.62	47.22	37.35	18.59	46.71
MGAE	43.38	28.78	16.43	33.48	61.35	34.63	33.55	57.36	71.57	62.13	48.82	68.08
ARGA	71.04	51.06	47.71	69.27	61.07	34.40	34.32	58.23	69.28	58.36	44.18	64.30
AdaGAE	50.06	32.19	28.25	53.53	54.01	27.79	24.19	51.11	67.70	55.96	46.20	62.95

(续表)

methods	CORA				CITE				AMAP			
	ACC	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1
GCA	53.62	46.87	30.32	45.73	60.45	36.15	35.20	56.42	56.81	48.38	26.85	53.59
CONVERT	73.57	55.07	50.28	70.92	68.33	42.15	42.12	62.21	77.19	66.33	**58.77**	**73.05**
AutoSSL	63.81	47.62	38.92	56.42	66.76	40.67	38.73	58.22	54.55	48.56	26.87	54.47
MGCN	71.09	53.89	50.19	69.97	65.67	39.96	37.55	59.67	76.67	65.55	57.56	71.19
Ours	**74.42**	**57.08**	**52.55**	**71.06**	**69.72**	**43.81**	**44.42**	**62.39**	**77.70**	**67.17**	58.53	72.32

表 4　对比实验结果

Table 4　Comparison experiment results

methods	BAT				CITE				UAT			
	ACC	*NMI*	*ARI*	*F*1	*EAT*	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1
DCN	47.79	18.03	13.75	46.80	38.85	6.92	5.11	38.75	46.82	17.18	13.59	45.66
DEC	42.09	14.10	7.99	42.63	36.47	4.96	3.60	34.84	45.61	16.63	13.14	44.22
MGAE	53.59	30.59	24.15	50.83	44.61	15.60	13.40	43.08	48.97	20.69	18.33	47.95
ARGA	67.86	49.09	42.02	67.02	52.13	22.48	17.29	52.75	49.31	25.44	16.57	50.26
AdaGAE	43.51	15.84	7.80	43.15	32.83	4.36	2.47	32.39	52.10	26.02	24.47	43.44
GCA	54.89	38.88	26.69	53.71	48.51	28.36	19.61	48.22	39.39	24.05	14.37	35.72
CONVERT	**76.79**	51.53	**49.56**	76.23	**58.31**	33.25	26.97	57.32	57.52	28.77	27.17	55.12
AutoSSL	42.43	17.84	13.11	34.84	31.33	7.63	2.13	21.82	42.52	17.86	13.13	34.94
MGCN	75.23	50.03	46.36	76.12	56.57	33.52	24.72	57.19	56.77	28.51	23.18	57.12
Ours	76.56	**51.59**	49.37	**76.40**	57.57	**34.00**	**27.58**	**57.57**	**58.72**	**29.77**	**27.29**	**57.54**

注：最优结果加粗表示。

4.3　消融实验

为了验证本文方法的有效性，本节进行了消融实验。具体实验包括验证所提出的边缘负样本补充方法和动态类别阈值伪标签方法。用“(w/o)B”表示移除边缘负样本补充方法，用“(w/o)D”表示移除动态类别阈值伪标签方法。具体的实验结果如表 5 和表 6 所列。

表 5　消融实验

Table 5　Ablation study

methods	CORA				CITE				AMAP			
	ACC	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1
(w/o)B	67.80	51.07	44.34	63.92	62.71	38.79	37.69	54.84	74.65	62.68	54.85	68.18
(w/o)D	65.25	48.78	41.66	61.70	60.94	38.13	36.72	51.49	69.15	58.77	51.09	62.03
Ours	**74.42**	**57.08**	**52.55**	**71.06**	**69.72**	**43.81**	**44.42**	**62.39**	**77.70**	**67.17**	**58.53**	**72.32**

注：最优结果加粗表示。

表 6　消融实验

Table 6　Ablation study

methods	BAT				EAT				UAT			
	ACC	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1	*ACC*	*NMI*	*ARI*	*F*1
(w/o)B	74.12	49.81	45.64	74.01	56.02	32.60	27.37	54.79	54.85	26.54	22.76	54.16
(w/o)D	73.97	49.77	45.90	73.73	55.94	31.88	26.97	54.77	54.29	26.01	21.89	53.33
Ours	**76.56**	**51.59**	**49.37**	**76.40**	**57.57**	**34.00**	**27.58**	**57.57**	**58.72**	**29.77**	**27.29**	**57.54**

注：最优结果加粗表示。

根据表 5 和表 6 的结果，可以得出以下结论：

1）本文提出的边缘负样本补充方法有效提高了模型的判别能力，边缘负样本的补充可以有效减少信息的压缩和丢失，使模型在训练过程中接触到更多样化的负样本情形，进而提升模型的鲁棒性和判别能力。

2）本文提出的类别动态阈值伪标签方法有效提高了模型的性能。以 AMAP 数据集为例，类别动态阈值伪标签方法在 ACC，NMI，ARI，F1 这 4 个指标上，分别提高了 8.55 个百分点、8.4 个百分点、7.44 个百分点、10.29 个百分点。这是因为相较于固定阈值，类别动态阈值能够根据学习难度，针对每一类别帮助模型更精确地选择对模型训练有帮助的样本，从而提高模型的学习效率。这也进一步验证了上文中假设的正确性。

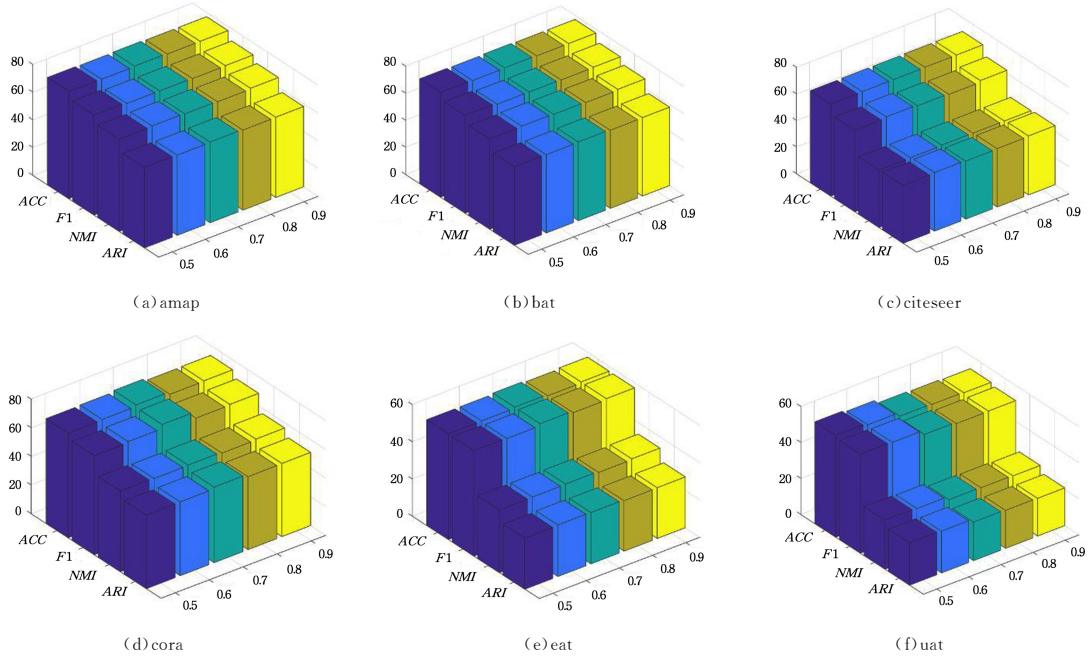
3）移除上述两种方法的任何一种方法后，模型的性能均有所下降，这表明边缘负样本补充与类别动态阈值伪标签都为模型的最优性能做出了贡献，同时也验证了所提出方法的有效性。

4.4　敏感性分析实验

4.4.1　阈值参数 T 的敏感性分析

本文在 6 个数据集上进行了敏感性分析实验，进一步验证了本文方法的鲁棒性，结果如图 2 所示。可以看出，当固定阈值参数 T 在 0.5 以上时，模型的性能十分稳定，这证实了

本文提出的类别动态阈值的方法可以在一定的阈值变化范围内自适应地调整类别阈值。当学习到的数据较少时，模型会适当降低学习难度，因而可以学习到更多样本。这证明了本文方法的稳定性。



(a) amap (b) bat (c) citeseer

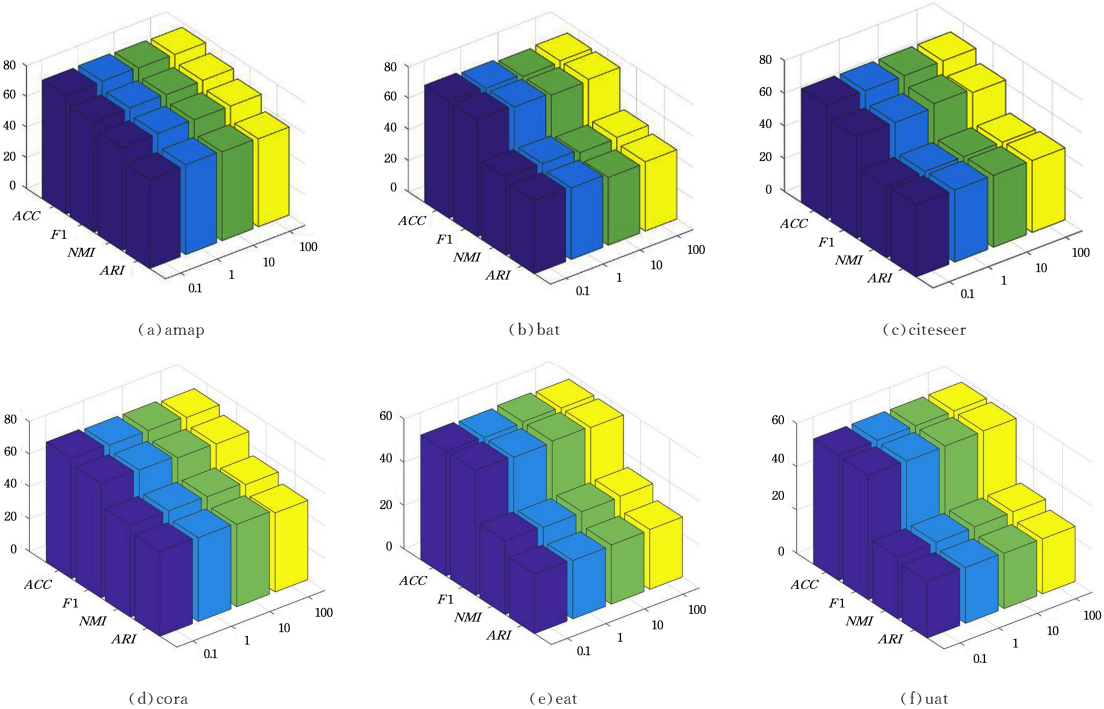
(d) cora (e) eat (f) uat

图 2 T 的敏感性分析

Fig. 2 Sensitive analysis of T

4.4.2 平衡参数 θ 的敏感性分析

此外，本文还在 6 个数据集上对平衡参数 θ 的敏感性进行了分析，以进一步验证所提出方法的鲁棒性。实验中，参数的取值范围为{0.1,1,10,100}。实验结果表明，当平衡参数在这一范围内小幅变化时，模型性能的波动较小。实验结果如图 3 所示。



(a) amap (b) bat (c) citeseer

(d) cora (e) eat (f) uat

图 3 θ 的敏感性分析

Fig. 3 Sensitive analysis of θ

4.5 MLP 权重分析实验

本文在 CORA 和 CITE 两个数据集上，对由两个 MLP 分别生成的节点嵌入表示在融合为综合嵌入矩阵 $\boldsymbol{Z}$ 时的权重分配方案进行了实验分析，结果如表 7 所列。实验结果表明，均匀赋予权重是一种有效方法，极端的权重分配可能会使模型过多地依赖某一视图，导致模型对特定类型的输入过于敏感，无法综合

两种视图的优势。而均匀分配有助于保持信息的平衡，提高模型的性能，在没有额外信息的情况下，这是一个合理的选择。

表 7　MLP 权重选择实验结果

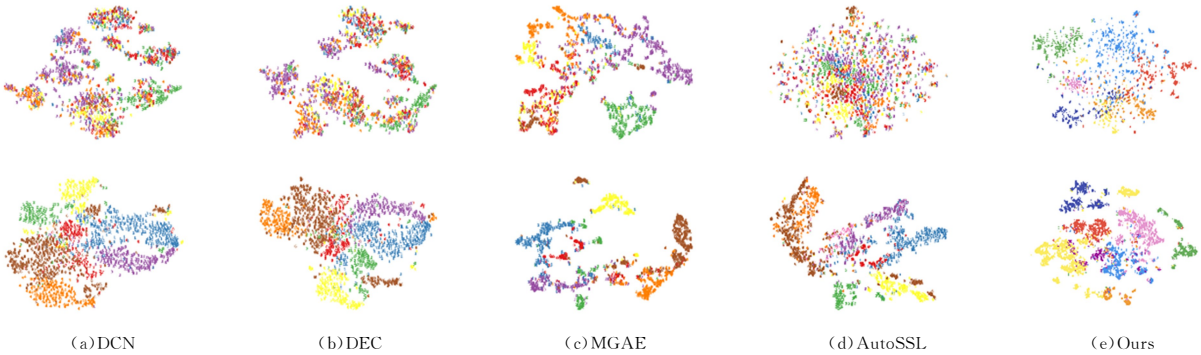
Table 7　MLP weight seletction experiment results

数据集	指标	MLP 权重分配									
		$\boldsymbol{Z}^{v_1}$ 0.1	$\boldsymbol{Z}^{v_2}$ 0.9	$\boldsymbol{Z}^{v_1}$ 0.3	$\boldsymbol{Z}^{v_2}$ 0.7	$\boldsymbol{Z}^{v_1}$ 0.5	$\boldsymbol{Z}^{v_2}$ 0.5	$\boldsymbol{Z}^{v_1}$ 0.7	$\boldsymbol{Z}^{v_2}$ 0.3	$\boldsymbol{Z}^{v_1}$ 0.9	$\boldsymbol{Z}^{v_2}$ 0.1
CORA	*ACC*	72.87		73.47		**74.42**		73.99		73.07	
	NMI	55.84		56.26		**57.08**		57.39		56.23	
	ARI	51.32		51.44		**52.55**		51.86		51.29	
	*F*1	68.83		69.85		**71.06**		68.62		67.61	
CITE	*ACC*	68.67		69.08		**69.72**		69.16		68.81	
	NMI	42.34		43.51		**43.81**		43.64		42.77	
	ARI	43.09		43.28		**44.42**		44.49		43.66	
	*F*1	60.21		61.52		**62.39**		62.12		61.65	

4.6　可视化分析

本文使用 T-SNE 算法对不同聚类方法（如 DCN，DEC，MGAE 和 AutoSSL）以及本文方法进行了可视化对比分析。

如图 4 所示，本文方法展现出更为紧密和分离度更高的聚类结构，节点分布更为集中且类别间界限更加明确。这表明本文方法在捕捉数据内在结构和提升聚类准确性方面具有优势，可以实现更优的聚类性能。



(a)DCN　(b)DEC　(c)MGAE　(d)AutoSSL　(e)Ours

图 4　CORA 和 AMAP 数据集上的可视化实验

Fig. 4　Visualization experiments on CORA and AMAP datasets

结束语　本文提出了一种基于动态阈值伪标签和对比学习的深度图聚类方法，通过引入信赖强度分值动态调整阈值，选择高置信度样本，形成正向反馈机制，提高了模型的判别能力和鲁棒性。此外，采用不共享参数的多层感知机结构和边缘负样本补充方法，进一步增强了模型对复杂图结构的理解能力和聚类性能。实验结果表明，在 CORA，CITE，AMAP，BAT，EAT 和 UAT 这 6 个基准数据集上，本文方法在多个评估指标上超过了现有的其他方法。总之，本文方法不仅在无监督学习任务中展现出卓越的性能，还为复杂图结构数据的聚类提供了新的解决思路和方法。未来可以在更大规模的数据集和更多样化的场景中进一步验证和优化该方法。

参 考 文 献

[1] WU Z，PAN S，CHEN F，et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems，2020，32(1)：4-24.

[2] GUPTA A，MATTA P，PANT B. Graph neural network：Current state of Art，challenges and applications[J]. Materials Today：Proceedings，2021，46：10927-10932.

[3] TSITSULIN A，PALOWITCH J，PEROZZI B，et al. Graph clustering with graph neural networks[J]. Journal of Machine Learning Research，2023，24(127)：1-21.

[4] TU W，GUAN R，ZHOU S，et al. Attribute-missing graph clustering network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：15392-15401.

[5] DIKE H U，ZHOU Y，DEVEERASETTY K K，et al. Unsupervised learning based on artificial neural network：A review[C]// 2018 IEEE International Conference on Cyborg and Bionic Systems(CBS). IEEE，2018：322-327.

[6] LIU Y，TU W，ZHOU S，et al. Deep graph clustering via dual correlation reduction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022：7603-7611.

[7] LI J，GUAN R，HAN Y，et al. Superpixel-Based Dual-Neighborhood Contrastive Graph Autoencoder for Deep Subspace Clustering of Hyperspectral Image[C]// International Conference on Intelligent Computing. Springer，2024：181-192.

[8] TSITSULIN A，PALOWITCH J，PEROZZI B，et al. Graph clustering with graph neural networks[J]. Journal of Machine Learning Research，2023，24(127)：1-21.

[9] WANG C，PAN S，HU R，et al. Attributed graph clustering：a deep attentional embedding approach[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019：3670-3676.

[10] XIA W，WANG Q，GAO Q，et al. Self-consistent contrastive attributed graph clustering with pseudo-label prompt[J]. IEEE

Transactions on Multimedia, 2022, 25: 6665-6677.
[11] WANG X, WU Z, LIAN L, et al. Debiased learning from naturally imbalanced pseudo-labels[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14647-14657.
[12] ARAZO E, ORTEGO D, ALBERT P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning[C]// 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.
[13] GUAN R, LI Z, LI X, et al. Pixel-superpixel contrastive learning and pseudo-label correction for hyperspectral image clustering [C]// ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 6795-6799.
[14] YANG X, LIU Y, ZHOU S, et al. Cluster-guided Contrastive Graph Clustering Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 10834-10842.
[15] XU D, CHENG W, LUO D, et al. Spatio-Temporal Attentive RNN for Node Classification in Temporal Attributed Graphs [C]// IJCAI. 2019: 3947-3953.
[16] KIPF T N, WELLING M. Variational Graph Auto-Encoders [J]. Stat, 2016, 1050: 21.
[17] WANG C, PAN S, HU R, et al. Attributed Graph Clustering: A Deep Attentional Embedding Approach[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 3670-3676.
[18] VELICKOVIC P, FEDUS W, HAMILTON W L, et al. Deep Graph Infomax[J]. Stat, 2018, 1050: 21.
[19] ZHU Y, XU Y, YU F, et al. Deep graph contrastive representation learning[J]. arXiv: 2006. 04131, 2020.
[20] QIU J, CHEN Q, DONG Y, et al. Gcc: Graph contrastive coding for graph neural network pre-training[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1150-1160.
[21] WU Z, XIONG Y, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3733-3742.
[22] CHEN X, FAN H, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[J]. arXiv: 2003. 04297, 2020.
[23] YOU Y, CHEN T, SUI Y, et al. Graph contrastive learning with augmentations[J]. Advances in Neural Information Processing Systems, 2020, 33: 5812-5823.
[24] GUAN R, LI Z, TU W, et al. Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14.
[25] YANG X, TAN C, LIU Y, et al. Convert: Contrastive graph clustering with reliable augmentation[C]// Proceedings of the 31st ACM International Conference on Multimedia. 2023: 319-327.
[26] CUI G, ZHOU J, YANG C, et al. Adaptive graph encoder for attributed graph embedding[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 976-985.
[27] LIU Y, TU W, ZHOU S, et al. Deep Graph Clustering via Dual Correlation Reduction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 7603-7611.
[28] YANG X, LIU Y, ZHOU S, et al. Cluster-guided Contrastive Graph Clustering Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 10834-10842.
[29] HASSANI K, KHASAHMADI A H. Contrastive Multi-view Representation Learning on Graphs[C]// International Conference on Machine Learning. PMLR, 2020: 4116-4126.
[30] XIE J, GIRSHICK R, FARHADI A. Unsupervised Deep Embedding for Clustering Analysis[C]// International Conference on Machine Learning. PMLR, 2016: 478-487.
[31] YANG B, FU X, SIDIROPOULOS N D, et al. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering[C]// International Conference on Machine Learning. PMLR, 2017: 3861-3870.
[32] WANG C, PAN S, LONG G, et al. Mgae: Marginalized Graph Autoencoder for Graph Clustering[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 889-898.
[33] PAN S, HU R, FUNG S, et al. Learning Graph Embedding with Adversarial Training Methods[J]. IEEE Transactions on Cybernetics, 2019, 50(6): 2475-2487.
[34] LI X, ZHANG H, ZHANG R. Adaptive Graph Auto-encoder for General Data Clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9725-9732.
[35] ZHU Y, XU Y, YU F, et al. Graph Contrastive Learning with Adaptive Augmentation[C]// Proceedings of the Web Conference 2021. 2021: 2069-2080.
[36] JIN W, LIU X, ZHAO X, et al. Automated Self-Supervised Learning for Graphs[J]. arXiv: 2106. 05470, 2021.
[37] LI X, WU W, ZHANG B, et al. Multi-scale Graph Clustering Network[J]. Information Sciences, 2024, 678: 121023.

WANG Pei, born in 2001, postgraduate. His main research interest is self supervised graph representation learning.

ZHU En, born in 1976, professor, Ph.D supervisor, is a senior member of CCF (No. 16689D). His main research interests include clustering, anomaly detection, computer vision, medical image analysis, etc.

(责任编辑:李亚辉)