计算机科学
COMPUTER SCIENCE

基于特征插值的深度图对比聚类算法

杨希洪, 郑群, 章佳欣, 王沛, 祝恩

引用本文

杨希洪, 郑群, 章佳欣, 王沛, 祝恩. 基于特征插值的深度图对比聚类算法[J]. 计算机科学, 2024, 51(11): 157-165.

YANG Xihong, ZHENG Qun, ZHANG Jiaxin, WANG Pei, ZHU En. Feature Interpolation Based Deep Graph Contrastive Clustering Algorithm [J]. Computer Science, 2024, 51(11): 157-165.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000209

基于特征插值的深度图对比聚类算法

杨希洪[1]　郑　群[2]　章佳欣[1]　王　沛[1]　祝　恩[1]
1 国防科技大学计算机学院　长沙 410073
2 中国科学技术大学地球和空间科学学院　合肥 230001
(yangxihong@nudt.edu.cn)

摘　要　Mixup是图像领域中一种有效的数据增强方法,它通过对输入图像以及标签进行插值来合成新的样本进而扩大训练分布。然而,在图节点聚类任务中,由于图数据拓扑结构的不规则性和连通性以及无监督的场景,设计有效的插值方法成为一项具有挑战性的任务。为了解决上述问题,首先通过设计不共享参数的编码器来获取视图的嵌入特征,有效融合节点的特征和结构信息。然后将视图的嵌入特征及其对应的伪标签进行混合插值,从而将Mixup引入聚类任务中。为了确保伪标签的可靠性,设置了阈值来筛选高置信度的伪标签,并通过EMA的方式更新模型参数,使模型平稳优化的同时考虑了训练的历史信息。此外,设计了一个图对比学习模块,以保证特征在不同视图下的一致性,从而减少信息冗余,提高模型的判别能力。最终,通过在6个数据集上的大量实验证明了所提方法的有效性。

关键词:数据增强;图对比聚类;EMA;Mixup;图神经网络
中图分类号　TP391

Feature Interpolation Based Deep Graph Contrastive Clustering Algorithm

YANG Xihong[1],ZHENG Qun[2],ZHANG Jiaxin[1],WANG Pei[1] and ZHU En[1]
1 College of Computer Science and Technology,National University of Defense Technology,Changsha 410073,China
2 School of Earth and Space Sciences,University of Science and Technology of China,Hefei 230001,China

Abstract　Mixup is an effective data augmentation technique in the field of computer vision. It is widely used for expanding the training distribution by interpolating input images and labels to generate new samples. However,in the context of graph node clustering tasks,designing robust interpolation methods poses challenges due to the irregularity and connectivity of graph data,as well as the unsupervised nature of the problem. To address these challenges,we propose a novel approach that leverages a dedicated encoder with non-shared parameters to extract embedding features from different views of graph. This allows us to effectively integrate both the node features and structural information. We then introduce Mixup into the clustering task by performing mixed interpolation on the embedding features along with their corresponding pseudo-labels. To ensure the reliability of these pseudo-labels,we apply a threshold to filter out high-confidence predictions,while incorporating an exponential moving average (EMA) mechanism for updating model parameters and considering the historical information during training. Furthermore,we incorporate a graph contrastive learning module to enhance feature consistency across different views,reducing information redundancy and improving the discriminative power of the model. Extensive experiments on six datasets demonstrate the effectiveness of the proposed method.

Keywords　Data augmentation,Graph contrastive clustering,EMA,Mixup,Graph neural network

1　引言

近年来,图神经网络[1-4]因具有强大的表示学习能力,已成为聚类[5]以及推荐系统[6]等诸多领域的研究热点。图节点聚类任务旨在学习图数据的嵌入表示,并将节点分为不同的簇,是一项关键且具有挑战性的图学习任务。

最近研究人员提出了许多图聚类算法,主要可以分为3类。其中,第一类是经典的聚类算法,例如DEC[7],DCN[8]。该类方法通过自编码器来提取数据特征,尽管实现了一定的聚类性能,但是在特征提取的过程中没有考虑图的拓扑结构信息。为了弥补上述缺点,第二类方法,即深度图聚类算法,该类方法通过图自编码器的结构来对图数据进行特征编码(如MGAE[9],ARGA[10],AdaGAE[11]),从而有效利用了图的拓扑信息,实现了对图信息的挖掘。对比学习作为一类新型

到稿日期:2023-10-30　返修日期:2024-04-05
基金项目:国家科技重大专项(2022ZD0209103)
This work was supported by the National Science and Technology Major Project (2022ZD0209103).
通信作者:祝恩(enzhu@nudt.edu.cn)

的方法，是图信息挖掘中的重要技术，但是深度图聚类算法中缺乏专门设计的对比学习策略，因此聚类性能也受到影响。因此，在第三类深度图对比聚类算法中（如 GCA[12]，AF-GRL[13]，AutoSSL[14]，MCGC[15]），研究人员设计了多样化的对比学习策略来进一步提升图聚类的性能，但是潜空间中特征信息的冗余性逐渐成为影响图节点聚类性能的重要原因[16]。

此外，Mixup 是一种在图像分类领域中提出的数据增强方法[17]，它通过对成对的随机图像以及标签进行线性插值，融合生成用于训练的合成图像，其具体过程如图 1 所示。Mixup 不强制要求标签在变化的数据中保持一致，相反，Mixup 的过程反映的是特征融合会导致对应的标签融合。因此，Mixup 可以构建出虚拟的训练样本来扩大训练分布，能够作为一种有效的正则化策略以训练网络模型，平滑决策边界，并提高网络的判别能力。虽然在图像领域中 Mixup 是一种有效的数据增强方法，但是设计一种面向图节点聚类任务的 Mixup 方法仍然具有挑战性。

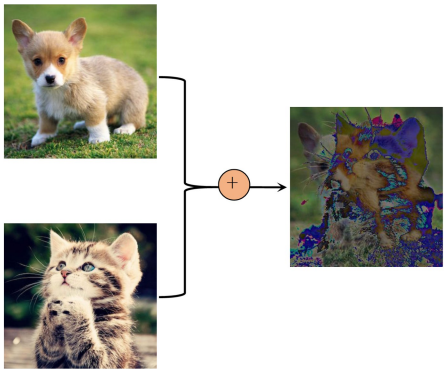


图 1 Mixup 插值融合

Fig. 1 Interpolation fusion of Mixup

首先，图不同于图像数据，其具有不规则性和连通性，并且图卷积神经网络在每层网络中通过“消息传递”的方式聚合每个节点与其邻居之间的表示，这使得图上嵌入特征的获取极大地依赖于图的拓扑结构。并且节点是无序地放置在不规则的网格上，这使得节点在不同的（子）图中进行混合变得困难。其次，由于节点之间的连通性，在不同的节点对上使用 Mixup 可能会相互干扰，从而导致融合特征可能存在冲突和干扰。最后，Mixup 需要对标签进行插值，在无监督的情况下，如何设计可靠的监督信息融合是一个难点问题。

基于上述问题，本文的目标是将 Mixup 引入图节点聚类任务中，同时设计策略来降低图嵌入潜空间中特征的冗余性。具体来说，通过设计不共享参数的编码器来获取图的嵌入表示，从而避免了使用常规数据增强，保证了图数据语义信息的可靠性，有效地融合了节点的特征和结构信息。在此基础上，对嵌入特征以及对应的伪标签进行线性插值，使得输入特征与模型预测呈线性变化；将 Mixup 引入图节点聚类任务中，进而提高了模型的判别能力。无监督场景下，伪标签的可信性决定了模型优化的质量。为了保证伪标签的可信性，本文通过设置阈值来筛选高置信度的伪标签。此外，为了确保伪标签的可靠性，本文设置了阈值来筛选高置信度的伪标签，并通过 EMA 的方式来更新模型参数，在动量的驱动下使得模型平稳优化，同时考虑训练的历史信息。

为了学习跨视图特征的一致性，减少冗余信息，本文设计了一个图对比模块，从而提高特征的判别能力。具体的模型图如图 2 所示。

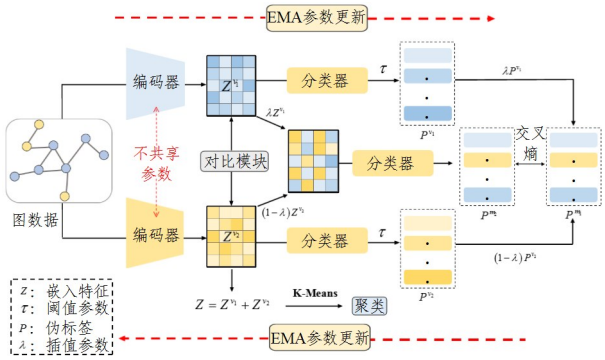


图 2 模型框架图

Fig. 2 Overall framework of the proposed model

综上，本文的主要贡献如下：

1）将 Mixup 引入无监督图节点聚类任务中，解决了图拓扑结构和伪标签可靠性的问题，并有效提高了模型的判别能力；

2）通过 EMA 方式来更新模型的参数，有效保存了训练信息，使得模型平稳优化，此外，通过阈值来进一步筛选高置信度的伪标签；

3）提出了一个图对比模块，能够保证样本特征在跨视图情景下的一致性，并且将相似度矩阵对齐单位矩阵，有效减少了冗余信息；

4）在 6 个数据集上进行了大量实验，实验结果证明了所提方法的有效性。

2 相关工作

2.1 图节点聚类

聚类是一项基本但具有挑战性的任务，旨在学习节点的语义表示并将节点划分为不同的簇。

具体而言，首先以无监督的方式训练神经网络 f 来对节点属性和结构信息进行编码。

$$\boldsymbol{Z}=f(\boldsymbol{X},\boldsymbol{A}) \tag{1}$$

其中，$\boldsymbol{X}$ 是属性矩阵，$\boldsymbol{A}$ 是邻接矩阵，$\boldsymbol{Z}$ 是经过网络编码的图嵌入表示。在此基础上，通过聚类算法将 $\boldsymbol{Z}$ 分成 k 个簇。

$$\Phi=C(\boldsymbol{Z}) \tag{2}$$

其中，C 为聚类算法，通常为 K-Means，Φ 为聚类结果。

随着深度学习技术的发展，深度图聚类逐渐成为图表示学习领域内的研究热点。受到图编码器（GAE）[18]的启发，MGAE[9]首先采用图编码器对节点进行编码，随后对潜在特征进行聚类。DAEGC[19]采用注意机制[20]来提升聚类性能；ARGA[10]通过对抗机制提高了样本的判别能力；SDCN[21]通过将 GAE 和自编码器集成到统一的框架中，缓解了过平滑

问题。此外，对比学习也已成为深度图聚类任务中的常用方法，例如 CCGC[5]，DCRN[16]，HSAN[22]。

2.2 数据增强

数据增强是一种通过对原始数据进行一系列可逆的转换来生成新的训练样本的技术，它可以通过一系列操作来扩充训练数据集。通过这些数据增强操作，可以生成具有多样性的训练样本，有助于提高深度对比图聚类模型的鲁棒性和泛化能力。这样的方式近年来受到了研究者的广泛关注。图数据增强是一种用于处理和增强图数据的技术，它在图机器学习和图深度学习任务中发挥着重要作用。由于图数据具有图结构和节点属性的特点，因此，与传统的图像数据增强不同，图数据增强关注的是在图结构中进行操作和扩充，以增加训练数据的多样性，主要方法包括子图采样、节点属性扰动、图结构扰动、子图重构以及图平移和旋转等。

图数据增强技术在深度对比图聚类研究中发挥着重要作用。具体而言，现有的深度图对比聚类方法通过对原始的图数据施加不同的数据增强方法来构造副本视图，然后在此基础上进行对比学习。例如 MVGRL[23]，GDCL[24] 中使用图扩散作为数据增强。

与上述方法不同的是，SCAGC[25] 通过随机添加或者删除边的方式来构建副本数据。尽管这些方法在聚类任务上取得了一定的效果，但是已有研究证明，常规的数据增强方法会导致图上语义信息的漂移，从而影响最终的聚类性能。不同于之前的方法，本文针对语义漂移的问题，通过不共享参数的编码器来构建新的视图，避免了语义信息的改变。

3 方法

3.1 符号描述

为了便于后续描述，本文使用 $G=(X,A)$ 来表示一个无向图。令 $V=(v_1,v_2,\cdots,v_N)$ 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，$\boldsymbol{A}\in R^{N\times N}$ 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 $\widetilde{\boldsymbol{L}}=\boldsymbol{I}-\widetilde{\boldsymbol{D}}^{-1/2}\overset{\wedge}{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-1/2}$，其中 $\overset{\wedge}{\boldsymbol{A}}=\boldsymbol{A}+\mathbf{I}$，$\mathbf{I}$ 为单位矩阵。详细符号说明如表 1 所列。

表 1 符号说明

Table 1 Notation descriptions

符号	含义
$\boldsymbol{X}$	属性矩阵
$\boldsymbol{A}$	邻接矩阵
$\mathbf{I}$	单位矩阵
$\boldsymbol{S}$	相似度矩阵
$\widetilde{\boldsymbol{L}}$	规范化的拉普拉斯矩阵
$\boldsymbol{Z}$	嵌入特征
$\boldsymbol{\lambda}$	插值参数

3.2 特征编码模块

在本节中，为了避免图卷积滤波器和权值矩阵的纠缠导致性能损失，采用广泛使用的拉普拉斯滤波器来聚合图上的邻居信息[26]。

$$\widetilde{\boldsymbol{X}}=(\mathbf{I}-\widetilde{\boldsymbol{L}})^t\boldsymbol{X} \tag{3}$$

其中，t 表示进行拉普拉斯滤波的次数，$\widetilde{\boldsymbol{L}}$ 为对称规范化的拉普拉斯矩阵，$\mathbf{I}$ 为单位矩阵。基于此，为避免常规数据增强（随机属性屏蔽、随机边删除等）导致的增强视图语义信息的漂移，本文通过不共享参数的特征编码器得到两个图 G 的图嵌入表示。

$$\begin{aligned}\boldsymbol{Z}^{v_1}&=f_1(\widetilde{\boldsymbol{X}})\\ \boldsymbol{Z}^{v_2}&=f_2(\widetilde{\boldsymbol{X}})\end{aligned} \tag{4}$$

其中，f_1 和 f_2 表示特征编码器，具体通过多层 MLP 来实现，令其结构相同，但是参数不同；$\boldsymbol{Z}^{v_1}$ 和 $\boldsymbol{Z}^{v_2}$ 是图 G 所生成的两个节点嵌入表示。随后，本文使用 l_2 范数对编码得到的图嵌入进行标准化，具体过程可以表示为：

$$\begin{aligned}\boldsymbol{Z}^{v_1}&=\frac{\boldsymbol{Z}^{v_1}}{\|\boldsymbol{Z}^{v_1}\|_2}\\ \boldsymbol{Z}^{v_2}&=\frac{\boldsymbol{Z}^{v_2}}{\|\boldsymbol{Z}^{v_2}\|_2}\end{aligned} \tag{5}$$

通过不共享参数的特征编码器可以生成不同的图数据语义，从而避免了常规数据增强对图语义信息的破坏。

3.3 特征编码增强

Mixup[27] 是图像领域中常见的数据增强方法，对于任意两对数据 (x_1,y_1) 和 (x_2,y_2)，Mixup 的操作为：

$$\begin{aligned}\widetilde{x}&=\boldsymbol{\lambda}x_1+(1-\boldsymbol{\lambda})x_2\\ \widetilde{y}&=\boldsymbol{\lambda}y_1+(1-\boldsymbol{\lambda})y_2\end{aligned} \tag{6}$$

其中，λ 是由 β 分布产生的，其值位于 0～1 之间。Mixup 通过合并先验知识来扩展训练分布，即特征的插值会导致相关标签的插值，进而增强数据模型的判别能力。

为了将 Mixup 的思想引入图节点聚类任务中，本文提出了一种适用于无监督图节点聚类的特征插值增强方法。由于图数据的拓扑结构，无法直接对原始的图数据进行融合插值，为此，基于 3.2 节中获取的图嵌入特征，本文通过 λ 对其进行特征插值融合。

$$\boldsymbol{Z}_m=\boldsymbol{\lambda}\boldsymbol{Z}^{v_1}+(1-\boldsymbol{\lambda})\boldsymbol{Z}^{v_2} \tag{7}$$

其中，H_m 是融合的图嵌入表示，在图嵌入层次上进行融合插值，避免了复杂图结构的影响。此外，在无监督的情况下，如何进行监督信息的融合是一个具有挑战性的问题。为了解决上述问题，本文通过筛选高于阈值 τ 图嵌入特征的伪标签来作为监督信息，其过程表示为：

$$\begin{aligned}P^{v_1}&=g(\boldsymbol{Z}^{v_1})\\ P^{v_2}&=g(\boldsymbol{Z}^{v_2})\end{aligned} \tag{8}$$

其中，P^{v_1} 和 P^{v_2} 分别为图嵌入 $\boldsymbol{Z}^{v_1}$ 和 $\boldsymbol{Z}^{v_2}$ 的伪标签，g 表示分类器。在此基础上，通过 λ 对伪标签进行融合，进而得到融合的伪标签 P^m：

$$P^{m_1}=\lambda P^{v_1}+(1-\lambda)P^{v_2} \tag{9}$$

类似地，通过分类器可以获取融合特征的伪标签：

$$P^{m_2}=g(\boldsymbol{Z}_m)=g(\lambda\boldsymbol{Z}^{v_1}+(1-\lambda)\boldsymbol{Z}^{v_2}) \tag{10}$$

在此基础上，通过交叉熵损失函数来约束融合的伪标签 P^{m_1} 和融合特征的伪标签 P^{m_2}，使其线性变化，进而提高模型的判别能力[28]。具体可以表示为：

$$L_C = CE(P^{m_1}, P^{m_2}) \tag{11}$$

其中，$CE(\cdot)$表示交叉熵函数。本节中，本文提出了一种基于特征插值的方法将 Mixup 引入图节点聚类任务中，使得嵌入特征和模型的预测之间呈现出线性变化，进而增强模型的判别能力。此外，插值过程可被视为一种数据增强方式，扩展了训练数据的多样性和丰富性。这有助于缓解过拟合问题，提高模型的泛化能力和鲁棒性。

3.4 图对比模块

为了充分挖掘图的潜在信息，本文通过保持交叉视图嵌入表示的一致性来提高样本的判别能力。如图 3 所示，图对比模块首先计算两个视图图嵌入表示的相似度矩阵。

$$S_{ij} = \frac{(\boldsymbol{Z}_i^{v_1})(\boldsymbol{Z}_j^{v_2})^{\mathrm{T}}}{\| \boldsymbol{Z}_i^{v_1} \|_2 \| \boldsymbol{Z}_j^{v_2} \|_2} \tag{12}$$

其中，S_{ij}表示第一个视图的图嵌入表示 $\boldsymbol{Z}^{v_1}$ 中的第 i 行和第二个视图的图嵌入表示 $\boldsymbol{Z}^{v_2}$ 第 j 行的余弦相似度。在此基础上，通过最小化相似度矩阵的冗余信息来强制相似度矩阵 $\boldsymbol{S}$ 来对齐单位矩阵 $\mathbf{I}$，具体过程可以表示为：

$$\begin{aligned} L_R &= \frac{1}{N^2}\sum(\boldsymbol{S}-\mathbf{I})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(S_{ii}-1)^2 + \frac{1}{N^2-N}\sum_{i=1}^{N}\sum_{i\neq j}(S_{ij})^2 \end{aligned} \tag{13}$$

其中，L_R 为对比损失。具体来说，相似度矩阵 $\boldsymbol{S}$ 中的对角线元素表示相同样本在不同视图下的余弦相似度，非对角线元素表示不同样本在不同视图下的余弦相似度。L_R 的第一项将其与 1 对齐，即拉近相似样本，而第二项与 0 对齐，即拉远不相似样本。通过上述操作，可以有效扩大不同样本嵌入特征之间的距离，同时保持视图间相同样本的特征不变，进而保证视图嵌入特征的一致性，提高模型的判别能力。

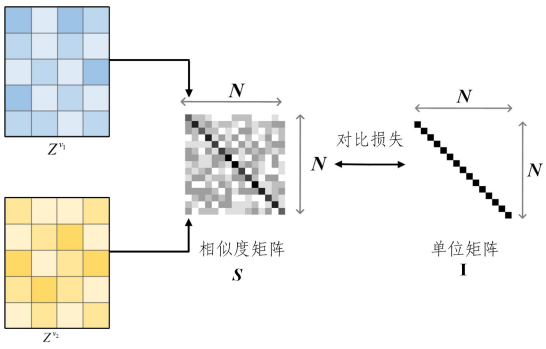


图 3 图对比模块

Fig. 3 Graph contrastive module

3.5 EMA 模型更新

EMA(Exponential Moving Average)是深度学习中用于估计变量局部均值的方法。通过 EMA 的方式，可以使得变量的更新与历史更新有关，EMA 的计算结果可以视为变量在过去一段时间内的均值。与直接更新变量值的方式相比，经过 EMA 得到的数值更加平缓光滑，波动性小，不会由于某次的异常值而引入过大的误差。

受到 MoCo[29] 和 Mean-Teache 的启发，本文引入 EMA 来更新模型的参数。具体来说，本文提出的模型包括 3 部分参数更新：特征编码器 f_1 和与其结构相同但是不共享参数的特征编码器 f_2，以及用于获得图嵌入特征伪标签的分类器 g。因此，EMA 模型更新的内容包括 3 个部分，即特征编码器 $\bar{f}_1$ 和 $\bar{f}_2$，以及分类器 $\bar{g}$。EMA 模型的参数 $\bar{\theta}$ 通过对原始模型参数 θ 进行移动平均的方式获得。

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta \tag{14}$$

EMA 模型的优点是可以在动量参数 m 的控制下平稳优化模型，同时考虑到全部的训练过程。在 EMA 模型的更新模式下，可以获得更加可信的伪标签，进而在无监督情景下提供更加可靠的监督信息。

3.6 损失函数

本文所提出模型的损失函数包含两部分，分别是无监督场景下的交叉熵损失 L_C 以及对比损失 L_R。总体损失可以表示为：

$$L = L_C + \alpha L_R \tag{15}$$

其中，α 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，设备的硬件信息为 Intel Core i7-7820x CPU，NVIDIA GeForce RTX 2080Ti GPU 和 64 GB RAM。

4.1.2 参数设置

实验的总训练次数为 400，为了避免实验的随机性，本文报告 10 次随机运行的平均值。平衡参数 α 统一设置为 1.0，阈值设置为 0.8，所有分类器的参数共享。

4.1.3 数据集描述

本文在实验中采用了 6 个被广泛使用的图数据集，包括 CORA，CITE，AMAP，BAT，EAT 和 UAT。这些数据集的具体信息如表 2 所列，基本描述如下：

CORA 数据集由 2708 份科学出版物组成，分为 7 类。引文网络由 5429 个链接组成。数据集中的每个出版物由值为 0/1 的词向量描述，该词向量指示字典中对应词不存在/存在的情形。该词典由 1433 个独特的单词组成。

CITE[5] 数据集是一个引文网络。类似于 CORA 数据集，它由 4732 个链接和 3327 份科学出版物组成，分为 6 类。数据集中的每个出版物由值为 0/1 的词向量描述，该词向量指示字典中分别对应词不存在和存在的情形。该词典由 3703 个独特的单词组成。

AMAP[16] 数据集是来自亚马逊的共同购买图。在 AMAP 中，产品由节点表示。此外，这些特征是通过单词袋编码的评论。边缘的含义为两种产品是否经常共同购买。

UAT[22] 数据集是 2016 年 1 月至 10 月从交通统计局收集的交通数据。它有 1190 个节点，13599 条边。

BAT[22] 数据集是 2016 年 1 月至 12 月从国家民用航空局(ANAC)收集的机场数据。它有 131 个节点，1038 条边。

EAT[22] 数据集是 2016 年 1 月至 11 月从欧盟统计局收集的机场数据。它有 399 个节点，5995 条边。

CORA，CITE 和 AMAP 是图机器学习常用的图数据集，具有以下特点。首先，节点表示文档或学术论文，每个节点可能具有特征，如词向量表示、关键词、摘要等。这些特征可以用于节点分类、节点聚类和链接预测等任务。其次，边表示引用关系，数据集中的边表示文档之间的引用关系或作者之间的合作关系。这些边可以用于构建图结构，并用于图分析和挖掘任务，例如链接预测、社区检测和影响力分析。最后是类别标签，这些数据集通常具有预定义的类别标签，用于节点分类任务。每个节点都被分配到一个或多个类别中，例如学科领域、主题或作者关系。

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

对上述 3 个数据集进行研究的难度主要体现在 3 个方面：首先是类别不平衡，这会导致在模型训练和评估过程中出现类别偏斜的问题，需要采取适当的策略来处理不平衡数据；其次，节点特征的稀疏性会增加模型学习和泛化的难度；最后是图结构的复杂性，具体来讲，节点之间存在的复杂关系和连接增加了模型处理的复杂性。

而 BAT，UAT 和 EAT 数据集的特点是规模较小，同时结构也相对简单。它们被主要用于研究交通流量、道路网络、机场运营等方面，适用于算法验证、模型开发以及研究探索。其难度与 CORA 等数据集相似，类别不平衡、数据稀疏性以及图结构的复杂性依旧是进行图学习的挑战。

4.1.4 评价指标

本文使用准确性(ACC)、归一化互信息(NMI)、平均随机指数(ARI)和宏观 F1 分数(F1)等指标来作为聚类结果的衡量标准。具体如下：

$$ACC=\frac{\sum_{i=1}^{n} f(l_i, map(c_i))}{n} \tag{16}$$

其中，c_i 和 l_i 分别代表预测的聚类中心以及第 i 个样本的预测标签。$\phi(\cdot)$是示性函数，其值根据式(17)进行计算：

$$\phi(l_i, map(c_i))=\begin{cases}1, & \text{if } l_i=map(c_i)\\ 0, & \text{otherwise}\end{cases} \tag{17}$$

从预测的聚类中心 c_i 到类别的最佳映射可以通过 Kuhn-Munkres 算法[30]构建，即 $map(\cdot)$。

F1 是一种常用的评估指标，用于衡量分类模型在精度(Precision)和召回率(Recall)之间的平衡。其中，精度表示模型正确预测为正例的样本数与所有预测为正例的样本数之间的比例；召回率表示模型正确预测为正例的样本数与所有实际为正例的样本数之间的比例。可以通过式(18)计算：

$$F1=2\times\frac{P\times R}{P+R} \tag{18}$$

其中，$P=TP/(TP+FP)$代表精度值，$R=TP/(TP+FN)$表示召回值，TP，FP 和 FN 分别代表真正误差、假正误差和假负误差。

归一化互信息分数(NMI)是被广泛用于聚类任务的一个指标，其计算基于信息论的概念。它使用熵和互信息来度量聚类结果和真实标签之间的相似性，NMI 的值越接近 1，表示聚类结果与真实标签之间的一致性越高，聚类效果越好。其计算式为：

$$NMI=-\frac{2\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}}{\sum_i p(x_i)\mathrm{lop}(p(x_i))+\sum_j p(y_j)\log(p(y_j))} \tag{19}$$

其中，x 和 y 分别代表预测结果和真实标签的数据分布。

ARI 是用于计算真实标签和预测值之间成对相似度的指标，计算式如下：

$$ARI=\frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}}-\overbrace{\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}^{\text{Expected index}}}{\underbrace{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]}_{\text{Max index}}-\underbrace{\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}_{\text{Expected index}}} \tag{20}$$

其中，n 表示样本对的数量，a 表示相同簇的样本对的数量，b 表示不同簇样本对的数量。

4.2 性能比较

在本节中，为了验证本文方法的优越性，将其在 6 个数据集上与 9 种方法进行了对比，包括经典的深度聚类方法(DEC[7]，DCN[8])、深度图聚类方法(MGAE[9]，ARGA[10]，AdaGAE[11])和深度图对比聚类方法(GCA[12]，AFGRL[13]，AutoSSL[14]，MCGC[14])。具体如表 3 所列。

根据表 3 中的结果，可以得出以下结论：

1)相比经典的聚类算法，本文算法取得了更好的性能。具体来说，DEC[7]和 DCN[8]通过自编码器来提取数据特征，然后通过 K-means 进行聚类。由于进行编码的网络结构简单，因此提取特征的质量相对偏低。此外，上述方法忽略了图数据中的拓扑信息，因此很难取得可靠的聚类效果。

2)本文模型的聚类性能优于深度图聚类算法，即 MGAE[9]，ARGA[10]和 AdaGAE[11]。上述方法通过图自编码器来进行数据特征的提取，尽管图的拓扑信息被考虑在内，但是由于缺乏专门设计的自监督策略，因此模型的图数据挖掘能力有所下降。

3)深度图对比聚类算法(GCA[12]，AFGRL[13]，AutoSSL[14]，MCGC[14])实现了次优性能。原因是上述方法很少考虑潜空间中的特征信息的冗余，导致特征的判别性降低。本文通过设计特征插值策略将 Mixup 引入无监督图节点分类任务中，提高了模型的判别能力。此外，通过设计对比损失，有效拉近了相似样本并推远了不相似样本，降低了特征的冗余信息，进一步提高了模型的判别能力。

表 3 对比实验

Table 3 Comparison experiment

(%)

数据集	指标	DCN	DEC	MGAE	ARGA	AdaGAE	GCA	AFGRL	AutoSSL	MCGC	Ours
CORA	ACC	49.38	46.50	43.38	71.04	50.06	53.62	26.25	63.81	42.85	**72.70**
	NMI	25.65	23.54	28.78	51.06	32.19	46.87	12.36	47.62	24.11	**55.28**
	ARI	21.63	15.13	16.43	47.71	28.25	30.32	14.32	38.92	14.33	**49.65**
	F1	43.71	39.23	33.48	69.27	53.53	45.73	30.20	56.42	35.16	**69.33**
CITE	ACC	57.08	55.89	61.35	61.07	54.01	60.45	31.45	66.76	64.76	**69.17**
	NMI	27.64	28.34	34.63	34.40	27.79	36.15	15.17	40.67	39.11	**42.75**
	ARI	29.31	28.12	33.55	34.32	24.19	35.20	14.32	38.73	37.54	**44.03**
	F1	53.80	52.62	57.36	58.23	51.11	56.42	30.20	58.22	59.64	**60.01**
AMAP	ACC	48.25	47.22	71.57	69.28	67.70	56.81	75.51	54.55	OOM	**77.46**
	NMI	38.76	37.35	62.13	58.36	55.96	48.38	64.05	48.56		**67.50**
	ARI	20.80	18.59	48.82	44.18	46.20	26.85	54.45	26.87		**58.27**
	F1	47.87	46.71	68.08	64.30	62.95	53.59	69.99	54.47		**72.14**
BAT	ACC	47.79	42.09	53.59	67.86	43.51	54.89	50.92	42.43	38.93	**75.27**
	NMI	18.03	14.10	30.59	49.09	15.84	38.88	27.55	17.84	23.11	**50.57**
	ARI	13.75	7.99	24.15	42.02	7.80	26.69	21.89	13.11	8.41	**47.76**
	F1	46.80	42.63	50.83	67.02	43.15	53.71	46.53	34.84	32.92	**75.01**
EAT	ACC	38.85	36.47	44.61	52.13	32.83	48.51	37.42	31.33	32.58	**57.64**
	NMI	6.92	4.96	15.60	22.48	4.36	28.36	11.44	7.63	7.04	**33.59**
	ARI	5.11	3.60	13.40	17.29	2.47	19.61	6.57	2.13	1.33	**27.55**
	F1	38.75	34.84	43.08	52.75	32.39	48.22	30.53	21.82	27.03	**57.37**
UAT	ACC	46.82	45.61	48.97	49.31	52.10	39.39	41.50	42.52	41.93	**55.95**
	NMI	17.18	16.63	20.69	25.44	26.02	24.05	17.33	17.86	16.64	**27.55**
	ARI	13.59	13.14	18.33	16.57	24.47	14.37	13.62	13.13	12.21	**23.16**
	F1	45.66	44.22	47.95	50.26	43.44	35.72	36.52	34.94	35.78	**55.65**

注:OOM 表示在训练过程中出现了 Out-of-Memory 的显存溢出错误。

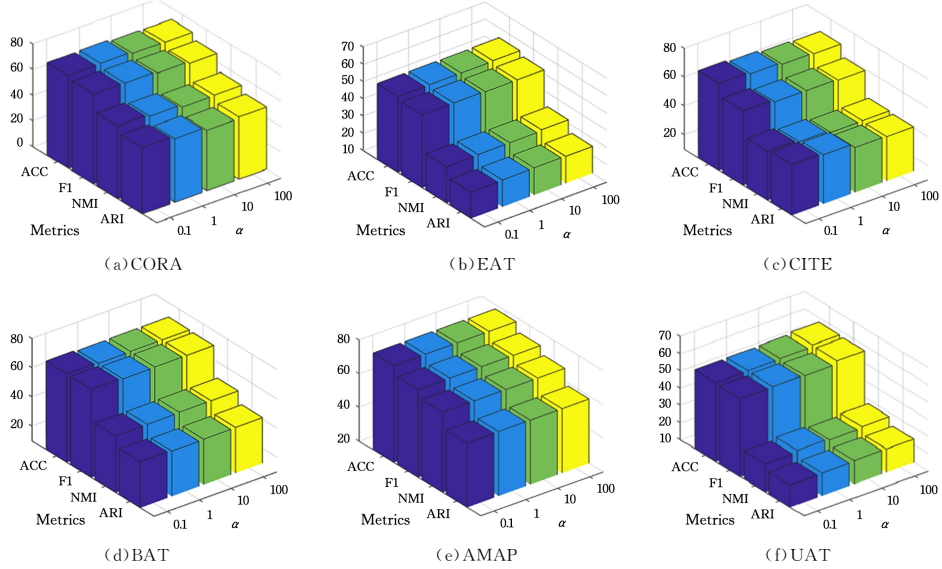
4)由表 3 中的结果可知,相比模型在 CORA,CITE 以及 AMAP 上的性能提升,本文模型在 BAT,EAT 以及 UAT 数据集上取得了更加显著的效果。这与数据集的规模密切相关。根据 4.1 节中的介绍,BAT,EAT 和 UAT 数据规模较小,结构相对简单,因此网络能够更充分地进行特征的提取,所获得表示的质量也更高,因此取得了更加优越的性能。

总之,本文方法在 4 个指标的度量下均实现了优于其他算法的实际性能。以 CORA 数据集为例,本文方法在 ACC,NMI,ARI 和 F1 4 个指标上分别较次优方法提升了 1.66%,4.22%,1.94%,0.06%。

4.3 敏感性分析实验

4.3.1 平衡参数 α 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,进一步验证所提方法对超参数的鲁棒性。具体而言,平衡参数 α 的取值范围为{0.1,1.0,10,100},如图 4 所示,α 在小范围内变化时模型性能表现平稳,这证明了本文算法的稳定性。



(a)CORA (b)EAT (c)CITE

(d)BAT (e)AMAP (f)UAT

图 4 α 的敏感性分析

Fig. 4 Sensitive analysis of α

4.3.2 阈值参数 τ 的敏感性分析

类似地,本文在 6 个数据集上对阈值参数 τ 进行了敏感性分析,结果如图 5 所示。根据结果可以得出以下结论:

1)随着阈值的逐渐增大,模型的聚类性能逐渐提升,表明筛选出的高置信度的伪标签可以优化模型的训练;

2)当阈值过大时,模型的聚类性能有所下降,这是由于

阈值过大会使得模型训练中的确认性偏差增大，随着训练的进行，误导模型的训练，因此会导致模型的性能下降。

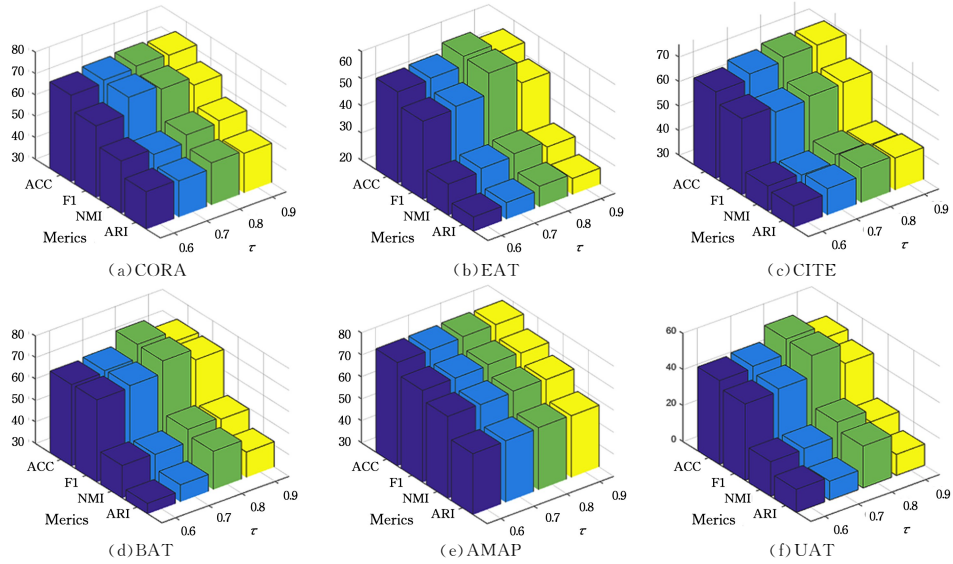


图 5 τ 的敏感性分析图

Fig. 5 Sensitive analysis of τ

4.3.3 动量参数 m 的敏感性分析

本文通过 EMA 的方式来更新模型参数，其中动量参数用于控制模型的平稳优化。本小节对其敏感性进行分析，具体实验设置为控制动量参数在 0.1～0.9 之间变化。在 CORA，CITE，AMAP 以及 BAT 数据集上进行了相关实验。

具体分析结果如图 6 所示。当动量参数值过小时（即 0.1），模型的聚类性能显著下降。根据动量参数更新模型参数的式(14)可以得出：当 m 过小时，EMA 模型参数将会缓慢地收敛到原始模型中；而设置较大的 m 值，可以使 EMA 参数更加快速地响应模型参数的变化，减少参数更新过程中的噪声，帮助模型捕捉更长期的趋势，提高模型对数据的泛化能力和鲁棒性。根据实验搜索，本文实验中将 m 的值设置为 0.9。

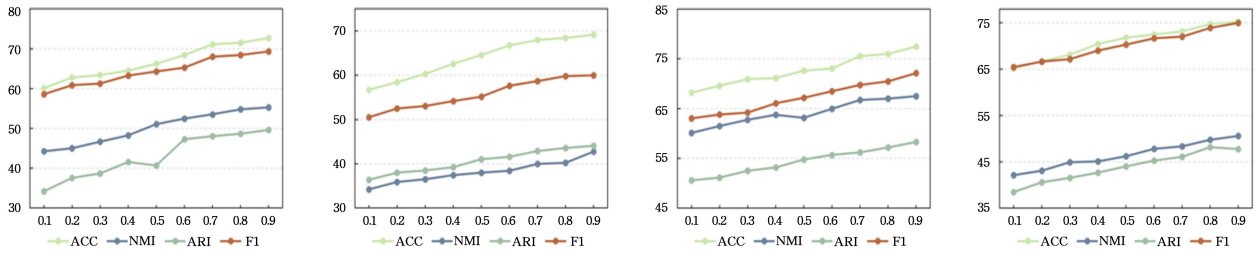


图 6 动量参数 m 的敏感性分析

Fig. 6 Sensitive analysis of m

4.4 可视化分析

为了凸显本文算法的有效性，在 AMAP 和 CORA 数据集上对潜空间中网络最后一层的输出特征使用 T-SNE 算法进行了可视化，结果如图 7 所示。与 DCN，DEC，MGAE 和 AutoSSL 的可视化结果相比，本文设计的方法更能有效地揭示数据在潜空间中的簇结构，从而实现更优的聚类性能。

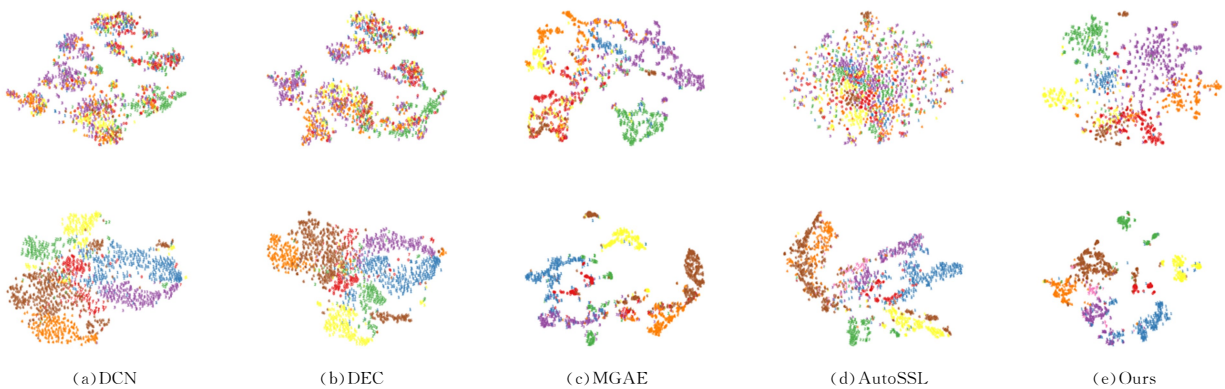


图 7 CORA 和 AMAP 数据集上的可视化实验

Fig. 7 Visualization experiments on CORA and AMAP datasets

4.5 时间以及空间消耗

本节进行了时间和空间消耗的实验。由表 4 可以观察到，与其他的深度图聚类算法相比，本文提出的算法具有更快的训练速度。其原因在于，本文使用拉普拉斯平滑的方式获取平滑特征，然后再经过 MLP 得到图数据的嵌入表示。这相比基于图神经网络的编码器，耗时更少，效率更高。

表 4 时间消耗

Table 4 Time consumption

数据集	DCN	DEC	MGAE	MCGC	Ours
CORA	47.31	91.13	7.38	118.07	**2.72**
CITE	74.69	223.95	6.69	126.06	**3.39**
AMAP	94.48	264.20	18.64	OOM	**7.42**
UAT	29.57	42.30	4.75	23.10	**1.37**
BAT	9.56	26.99	4.64	2.87	**1.17**
EAT	7.46	21.37	3.83	2.28	**1.14**
Avg.	72.16	193.09	10.90	122.07	**4.51**

尽管本文所设计的对比损失函数使得算法的运行产生了一定的空间消耗，但图 8 的结果显示，本文提出的方法在空间上的消耗是可接受的。

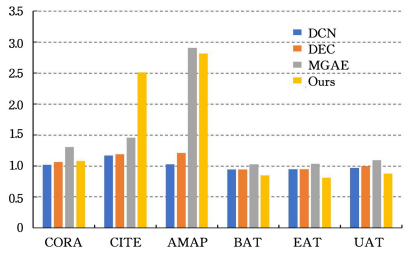


图 8 GPU 空间消耗

Fig. 8 GPU memory cost

4.6 消融实验

本节进行了消融实验来验证所提出的模块的有效性。具体实验包括验证所提出的特征插值方法、验证对比学习模块以及验证 EMA 更新模型参数，分别使用“(w/o)F”“(w/o)C”和“(w/o)EMA”来表示移除特征插值方法、移除对比学习模块以及移除 EMA 更新模型参数，具体结果如表 5 所列。

根据表 5 的结果，可以得出以下结论：

1)移除本文所提出的任意一个模块后，即移除特征插值模块、对比学习模块或者 EMA 中的任意一个，模型的性能均有所下降，这表示该模块与模型的最优性能密切相关，也反映了所提出模块的有效性。

2)本文提出的特征插值方法有效提高了特征和模型的判别能力，以 CORA 数据集为例，在 4 个指标上，结果分别提高了 8.45%，6.59%，9.23%，11.22%。

3)在无监督的场景下，图对比学习模块通过学习跨视图特征的一致性来减少相似度矩阵中的冗余信息，拉近了相似样本，同时推远了不相似样本，提高了特征的判别能力。

4)本文采用 EMA 的方式来更新模型的参数，使得模型能够在动量 m 的控制下平稳演化，同时考虑了整个优化训练过程的影响，从而提高了伪标签的可信度，发挥了其在无监督场景下的指导性能，进而提升了模型的性能。

表 5 消融实验

Table 5 Ablation study

数据集	指标	(w/o)C	(w/o)F	(w/o)EMA	Ours
CORA	ACC	71.48	64.25	70.23	**72.70**
	NMI	53.48	48.69	53.08	**55.28**
	ARI	48.65	40.42	47.40	**49.65**
	F1	65.15	58.11	65.93	**69.33**
CITE	ACC	67.63	60.49	67.72	**69.17**
	NMI	41.47	38.71	42.12	**42.75**
	ARI	41.39	36.84	42.15	**44.03**
	F1	59.01	48.73	59.31	**60.01**
AMAP	ACC	69.40	71.51	77.30	**77.46**
	NMI	60.51	59.57	66.56	**67.50**
	ARI	50.56	52.76	57.88	**58.27**
	F1	63.51	63.90	71.28	**72.14**
BAT	ACC	70.08	74.12	69.69	**75.27**
	NMI	49.03	51.06	47.68	**50.57**
	ARI	41.17	47.46	39.98	**47.76**
	F1	69.77	73.31	69.38	**75.01**
EAT	ACC	54.11	54.56	55.94	**57.64**
	NMI	31.46	30.24	33.31	**33.59**
	ARI	25.45	26.38	25.82	**27.55**
	F1	51.91	52.13	56.28	**57.37**
UAT	ACC	47.97	49.75	49.33	**55.95**
	NMI	20.53	18.63	22.36	**27.55**
	ARI	10.80	17.46	14.45	**23.16**
	F1	45.12	47.81	49.50	**55.65**

结束语 本文提出了一种基于特征插值的深度图对比聚类算法，通过设计不共享参数的编码器，将图的结构信息和特征映射到潜空间中，并且通过分类头获得特征的伪标签，通过将嵌入特征及其对应的伪标签线性插值，将 Mixup 引入到图节点聚类任务中。为了提高无监督情况下伪标签的可靠性，本文设置高置信度的阈值来筛选伪标签，并使用 EMA 的方式来更新模型参数，在动量的控制下使得模型平滑优化，同时考虑整个训练过程，减少模型偏差。此外，本文还设计了一个图对比学习模块，用于学习跨视图特征的一致性，从而减少特征的冗余信息，提高模型的判别能力。最终，在 6 个广泛使用的数据集上的实验结果证明了本文方法的有效性。

本文方法的不足之处在于伪标签的质量。算法的核心是提出了一种插值策略，即将嵌入特征及其对应的伪标签进行插值。其中，伪标签作为无监督聚类任务中的监督信息，在模型训练以及优化的过程中起着非常重要的作用。尽管本文通过设置阈值以及 EMA 的方式来更新模型参数，以获得相对高质量的伪标签，但是伪标签的质量依旧需要进一步提升。

如何在无监督情景下提高伪标签的质量是未来图节点聚类任务中的一项挑战。此外，本文采用的数据集规模较小，大规模图数据在训练过程中又会带来训练空间和时间上的消耗，因此，如何将融合特征的增强方法应用到大规模图数据上，同时保证可接受的时间和空间消耗也将会是一个有趣的研究方向。

参 考 文 献

[1] HUANG Y J, CHEN M, ZHENG Y, et al. Text Classification Based on Weakened Graph Convolutional Networks[J]. Computer Science, 2023, 50(S1): 220700039-5.

[2] LI F, JIA D L, YAO Y, et al. Graph Neural Network Few Shot

Image Classification Network Based on Residual and Self-attention Mechanism [J]. Computer Science, 2023, 50(S1): 220500104-5.

[3] YANG Y, ZHANG F, LI T R. Aspect-based Sentiment Analysis Based on Dual-channel Graph Convolutional Network with Sentiment Knowledge[J]. Computer Science, 2023, 50(5): 230-237.

[4] WANG Y L, ZHANG F, YU Z, et al. Aspect-level Sentiment Classification Based on Interactive Attention and Graph Convolutional Network[J]. Computer Science, 2023, 50(4): 196-203.

[5] 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.

[6] XU D, CHENG W, LUO D, et al. Spatio-Temporal Attentive RNN for Node Classification in Temporal Attributed Graphs [C]// IJCAI. 2019: 3947-3953.

[7] XIE J, GIRSHICK R, FARHADI A. Unsupervised Deep Embedding for Clustering Analysis[C]// International Conference on Machine Learning. PMLR, 2016: 478-487.

[8] 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.

[9] WANG C, PAN S, LONG G, et al. Mgae: Marginalized Graph Autoencoder for Graph Clustering[C]// Proceedings of the 2017 ACM Conference on Information and Knowledge Management. 2017: 889-898.

[10] 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.

[11] 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.

[12] ZHU Y, XU Y, YU F, et al. Graph Contrastive Learning with Adaptive Augmentation [C]// Proceedings of the Web Conference 2021. 2021: 2069-2080.

[13] LEE N, LEE J, PARK C. Augmentation-free Self-supervised Learning on Graphs[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 7372-7380.

[14] JIN W, LIU X, ZHAO X, et al. Automated Self-Supervised Learning for Graphs[J]. arXiv: 2106.05470, 2021.

[15] PAN E, KANG Z. Multi-view Contrastive Graph Clustering[J]. Advances in Neural Information ProcessingSystems, 2021, 34: 2148-2159.

[16] 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.

[17] YANG X, HU X, ZHOU S, et al. Interpolation-based Contrastive Learning for Few-label Semi-Supervised Learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(2): 2054-2065.

[18] GAO C, WANG X, HE X, et al. Graph Neural Networks for Recommender System[C]// Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1623-1625.

[19] 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.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-11.

[21] BO D, WANG X, SHI C, et al. Structural Deep Clustering Network[C]// Proceedings of the Web Conference 2020. 2020: 1400-1410.

[22] LIU Y, YANG X, ZHOU S, et al. Hard Sample Aware Network for Contrastive Deep Graph Clustering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 8914-8922.

[23] HASSANI K, KHASAHMADI A H. Contrastive Multi-view Representation Learning on Graphs[C]// International Conference on Machine Learning. PMLR, 2020: 4116-4126.

[24] ZHAO H, YANG X, WANG Z, et al. Graph Debiased Contrastive Learning with Joint Representation Clustering[C]// IJCAI. 2021: 3434-3440.

[25] XIA W, WANG Q, GAO Q, et al. Self-consistent Contrastive Attributed Graph Clustering with Pseudo-label Prompt [J]. IEEE Transactions on Multimedia, 2023, 25: 6665-6677.

[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] WANG Y, WANG W, LIANG Y, et al. Mixup for Node and Graph Classification[C]// Proceedings of the Web Conference 2021. 2021: 3663-3674.

[28] VERMA V, KAWAGUCHI K, LAMB A, et al. Interpolation Consistency Training for Semi-Supervised Learning[J]. Neural Networks, 2022, 145: 90-106.

[29] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.

[30] PLUMMER M D, LOV'ASZ L. Matching theory[M]. Elsevier, 1986.

YANG Xihong, born in 1999, Ph.D candidate. His main research interests include self supervised graph representation learning, recommendation system, deep multi-view learning, etc.

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：杨雪敏）