◆计算科学◆

基于自监督信息增强的图表示学习*

袁立宁[1,2],文　竹[2**],冯文刚[1],刘　钊[3]
(1.中国人民公安大学国家安全学院,北京　100038;2.广西警察学院信息技术学院,广西南宁　530028;3.中国人民公安大学研究生院,北京　100038)

摘要:针对图表示学习模型依赖具体任务进行特征保留以及节点表示的泛化性有限等问题,本文提出一种基于自监督信息增强的图表示学习模型(Self-Variational Graph Auto Encoder,Self-VGAE)。Self-VGAE首先使用图卷积编码器和节点表示内积解码器构建变分图自编码器(Variational Graph Auto Encoder,VGAE),并对原始图进行特征提取和编码;然后,使用拓扑结构和节点属性生成自监督信息,在模型训练过程中约束节点表示的生成。在多个图分析任务中,Self-VGAE的实验表现均优于当前较为先进的基线模型,表明引入自监督信息能够增强对节点特征相似性和差异性的保留能力以及对拓扑结构的保持、推断能力,并且Self-VGAE具有较强的泛化能力。

关键词:自监督信息;图表示学习;图变分自编码器;图卷积网络;对比损失

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1005-9164(2024)02-0323-12
DOI:10.13656/j.cnki.gxkx.20240619.013

随着图结构数据在社交网络[1]、犯罪网络[2]以及知识图谱[3]等领域的广泛应用,构建高性能图数据分析和挖掘算法,提取数据中蕴含的有效高维非线性特征已成为当前研究的重点内容。传统基于启发式的图分析算法[4]受限于人工设计特征和对数据分布的强假设等问题,不仅具有较高的计算复杂度,而且泛化能力也较弱。基于深度学习的图表示学习算法[5]因其强大的表征能力,受到了越来越多的关注。图表示学习将节点编码为一组保留拓扑结构信息、节点属性信息等原始图信息的低维向量,进而提升节点分类、节点聚类、链接预测以及可视化等下游图分析任务的正确率与效率。

图表示学习的关键在于选择和编码原始图特征,从而生成低维的节点表示向量。由于图数据中包含局部拓扑结构、全局拓扑结构以及节点属性等多种不同类型信息,增加了特征选择的难度,因此多数图表示学习模型依赖具体的图分析任务来保留特征,例如

收稿日期:2024-01-22　　修回日期:2024-02-18
*国家重点研发计划项目(2023YFC3321604),中央高校基本科研业务费专项资金项目(2022JKF02002),广西法学会法学研究课题(GFKT2023-C3)和广西哲学社会科学研究课题(23FTQ005)资助。
【第一作者简介】
袁立宁(1995—),男,在读博士研究生,主要从事机器学习和图神经网络研究。
【**通信作者简介】
文　竹(1982—),女,硕士,副教授,主要从事机器学习和大数据技术研究,E-mail:wenzhu@gxjcxy.edu.cn。
【引用本文】
袁立宁,文竹,冯文刚,等.基于自监督信息增强的图表示学习[J].广西科学,2024,31(2):323-334.
YUAN L N,WEN Z,FENG W G,et al.Graph Representation Learning Enhanced by Self-supervised Information [J].Guangxi Sciences,2024,31(2):323-334.

节点分类关注模型对节点间相似性和差异性的保留能力,链接预测关注模型对节点间局部拓扑结构的保持和推断能力。当前,对图表示学习算法的改进方式主要分为两类,一是设计不同的神经网络结构,二是改进模型优化方式和策略。神经网络结构的设计可以通过修改图卷积编码器[6]以增强模型对原始图特征的提取能力;也可以通过在多层神经网络中引入残差连接[7]、One-Shot 聚合[8]等跨层连接方式,改善层间信息传递;还可以通过多视角[9]、多尺度[10]等形式,从不同角度关注和保留原始图信息。优化方式和策略的改进可以通过引入不同的损失函数以保留相应的拓扑或属性信息;也可以通过噪声注入[11]、对抗训练[12]等方式增强模型的泛化能力。

变分图自编码器(Variational Graph Auto Encoder,VGAE)[13]是一类重要的图表示学习算法,由变分自编码器(Variational Auto Encoder,VAE)[14]构建,编码器部分利用图卷积生成用于计算节点表示的概率分布,解码器部分利用节点表示重构邻接关系,实现链接预测任务。在 VGAE 的基础上衍生出许多变体,Salha 等[15]提出了一种线性变分自编码器模型 Linear-VGAE,其使用线性编码的简化图卷积(Simplified Graph Convolution, SGC)[16]替换 VGAE 中的图卷积编码器,减少模型的参数规模并降低计算复杂度,实现链接预测和节点聚类任务;Keser 等[17]使用残差连接构建残差变分图自编码器(Residual Variational Graph Auto Encoder, RVGAE),通过计算机视觉中常用的跨层连接方式改善层间信息传递,实现链接预测任务;Hy 等[10]提出了一种多尺度变分图自编码器(Multi-Scale Variational Graph Auto Encoder,MSVAGE),通过编码器生成多组不同尺度的低维向量表示原始图的混合概率分布,然后在每个维度上进行多次采样,并通过属性信息帮助结构表征学习,实现链路预测任务。

虽然上述改进方式提升了 VGAE 及其变体模型的性能,但是多数模型仅针对特定的图分析任务进行优化,即依赖具体任务进行特征保留,因此限制了模型生成的节点表示在不同图分析任务上的表现。针对上述问题,本文提出一种基于自监督信息[18]增强的图表示学习模型 Self-VGAE,以生成适用于多个图分析任务的泛化节点表示。Self-VGAE 主要分为两个部分:一是使用双层图卷积编码器和节点表示内积解码器构建基础 VGAE 模型,并通过重构损失和噪声约束对基础 VGAE 模型进行训练;二是使用拓扑结构和节点属性构建自监督信息,并在模型训练过程中利用构建的自监督信息约束节点表示的生成。

1 相关工作

1.1 图表示学习

开发高效的图结构分析算法已成为现在的研究热点。由于传统的路径分析、连通性分析和中心性分析等启发式算法从邻接矩阵中提取图的拓扑信息时依赖人工设计的特征,因此在应用时受限于数据的高维非线性,通常具有较高的计算复杂度和内存需求,并且人工设计的特征通常针对特定任务,在不同任务中的泛化能力十分有限[4]。图表示学习算法通过映射函数将原始图的拓扑结构和属性信息编码到潜在向量空间中,使高维、稀疏的图数据转换为低维、稠密的向量,在最大限度保留图特征信息的同时,解决图结构难以高效输入机器学习算法中的问题。图表示学习算法的一般过程如图 1 所示。

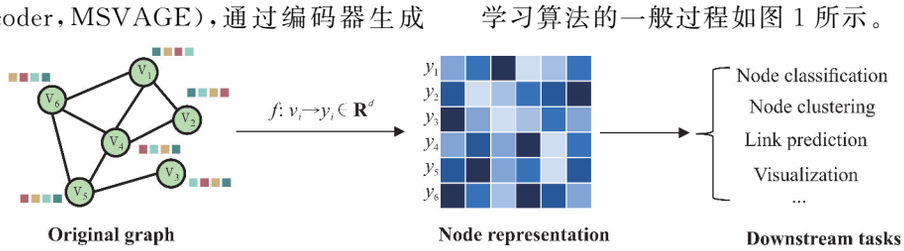


图 1 图表示学习算法的一般过程

Fig. 1 Process of graph representation learning algorithms

近年来,许多学者通过修改模型优化方式和策略来提升图表示学习算法的性能。Wang 等[11]提出基于噪声注入的训练策略,通过将噪声注入到输入图中,在防止模型参数过拟合的同时,设置合适的噪声率能够持续提高训练性能。Wang 等[19]提出能够同时提取节点属性和拓扑结构信息的图卷积自编码器模型 GASN,利用节点属性和拓扑星系重构损失,从而对编码和解码过程进行优化,增强模型对原始图信息的表征能力。Fettal 等[20]将聚类损失函数纳入表示学习过程,最小化节点表示与重建聚类表示之间的差异,提升模型在聚类任务中的表现。Li 等[21]利用对抗互信息训练模型,在 VAE 编码过程中最大化节

点特征和节点表示的互信息,使模型能够有效保留图的拓扑结构和节点属性信息。

1.2 自监督学习

自监督学习[22]是从无监督数据中提取可转移的知识或特性,将其扩展为自监督信息,并利用自监督信息对神经网络进行训练,最终生成对下游任务有价值的特征表示。当前,主流的自监督学习算法主要分为预测模型和对比模型[22]。预测模型采用监督学习方式进行训练,利用输入数据生成自监督标签,对预测结果和自监督标签之间的差异进行优化;对比模型则采用自监督学习方式生成数据表示,利用输入数据构造正负样本对,增大特征空间中同类数据表示的相似性和不同类数据表示的差异性。

由于图表示学习算法通常采用无监督学习方式进行训练,因此难以运用预测模型生成的自监督标签对特征提取过程进行约束。对比模型能够调整正负样本对的距离,因此可以将其融入到图表示学习的训练过程。对比模型的关键在于如何从输入数据中提取包含正向和负向关系的自监督信息,并构建对比约束对正负样本距离进行度量。在自监督学习中,对比约束的一般形式为

$$\mathrm{loss}(x_i)=-\log\frac{\exp(s_{i,i}/\tau)}{\sum_{k=1}^{N}1_{[k\neq i]}\exp(s_{i,k})+\exp(s_{i,i}/\tau)},\tag{1}$$

式中,τ 为对比约束中的温度系数,用于控制模型对负样本的关注度,s 为数据样本相似性度量的结果。式(1)也被称为 InfoNCE[23],对比约束可以使第 i 个数据与正样本之间的相似度 $s_{i,i}$ 尽可能大,与负样本之间的相似度 $s_{i,k}$ 尽可能小。Hu 等[24]使用图的拓扑结构生成自监督信息,然后使用 InfoNCE 构建局部结构对比约束,使仅使用节点属性作为输入的多层感知模型能够匹配图卷积神经网络(Graph Convolutional Network,GCN)[6]在节点分类任务中的性能。

本文在上述研究的基础上,使用原始图的节点属性和拓扑结构生成自监督信息,并基于 InfoNCE 设计用于图表示学习任务的自监督约束,以推动特征空间相似节点表示相互接近,不相似节点表示相互远离,在增强模型对原始图信息表征能力的同时,提升模型在节点分类、节点聚类、链接预测等多个下游图分析任务的表现。

2 图表示学习模型(Self-VGAE)

Self-VGAE 的算法原理如图 2 所示,首先给出 VGAE 模型的编解码器网络结构以及优化函数,用于提取原始图特征并生成节点向量表示,然后在 VGAE 的基础上引入增强模型表征能力的自监督损失,对节点表示进行优化,使潜在空间中相似节点彼此接近,不相似节点彼此远离。

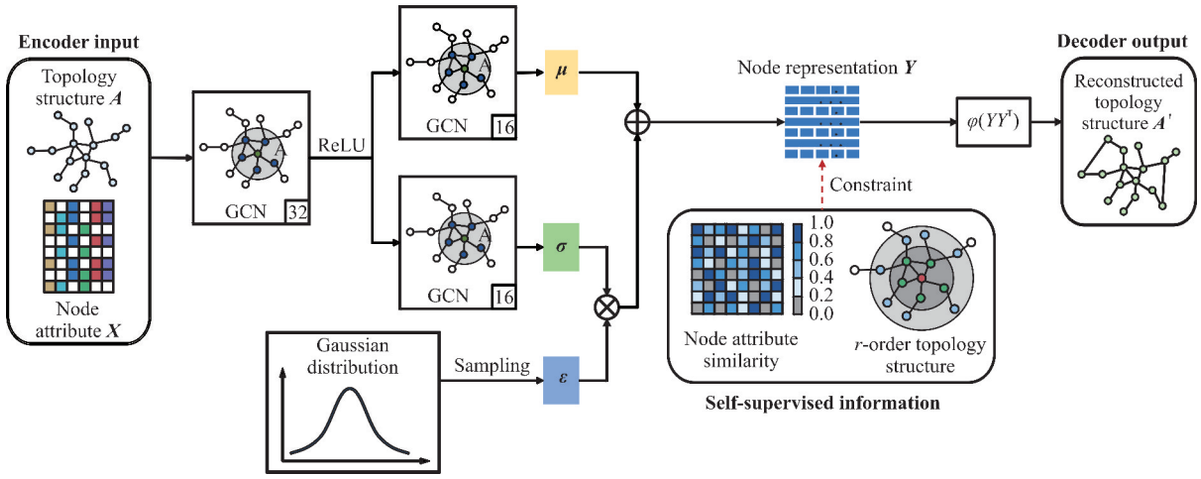


图 2 Self-VGAE 的结构

Fig. 2 Framework of Self-VGAE

2.1 图变分自编码器

VGAE 使用 GCN 编码器对原始图进行特征提取,以生成用于计算节点表示的均值向量 $\boldsymbol{\mu}$ 和方差向量 $\boldsymbol{\sigma}$。具体而言,GCN 以蕴含拓扑结构信息的邻接矩阵 $\boldsymbol{A}$ 和蕴含节点属性信息的属性矩阵 $\boldsymbol{X}$ 为输入,其层间传播公式为

$$\boldsymbol{H}^{(l+1)}=\delta(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l+1)}),\tag{2}$$

式中,$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$,$\tilde{\boldsymbol{D}}$ 为顶点度矩阵,$\boldsymbol{W}^{(l)}$ 为各层可训练的参数矩阵,$\delta(\cdot)$为激活函数(本文使用 ReLU 激

活函数),$\boldsymbol{H}^{(\cdot)}$为各层特征矩阵(当 $l=0$ 时,$\boldsymbol{H}^{(0)}$为属性矩阵 $\boldsymbol{X}$)。

Self-VGAE 模型基本结构与 VGAE 相同,编码器使用双层 GCN 进行编码,第一层表达式为

$$\boldsymbol{H}^{(1)}=\mathrm{ReLU}(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}^{(1)})。\tag{3}$$

然后基于第一层计算结果生成 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$,第二层表达式为

$$\boldsymbol{\mu}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(1)}\boldsymbol{W}^{(\boldsymbol{\mu})}$$

$$\ln\boldsymbol{\sigma}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(1)}\boldsymbol{W}^{(\boldsymbol{\sigma})},\tag{4}$$

式中,$\boldsymbol{W}^{(\boldsymbol{\mu})}$ 和 $\boldsymbol{W}^{(\boldsymbol{\sigma})}$ 分别表示用于计算 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$ 的参数矩阵。利用编码器计算的 $\boldsymbol{\mu}$ 和 $\boldsymbol{\sigma}$ 以及从高斯分布中采样的噪声编码 $\boldsymbol{\varepsilon}$,计算节点表示矩阵 $\boldsymbol{Y}$:

$$\boldsymbol{Y}=\boldsymbol{\mu}+\boldsymbol{\sigma}\odot\boldsymbol{\varepsilon},\tag{5}$$

式中,$\boldsymbol{Y}=\{\boldsymbol{y}_i\}_{i=1}^{N}$,$\boldsymbol{y}_i$ 是节点 i 的低维表示,N 为节点数量,$\odot$ 表示 Hadamard 积运算。

对于 Self - VGAE 模型,解码器利用节点表示的内积生成重构邻接矩阵 $\boldsymbol{A}'$:

$$\boldsymbol{A}'=\varphi(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}),\tag{6}$$

式中,φ 为 Sigmoid 函数。与 VAGE 的模型优化方式相似,Self - VGAE 在训练过程中也使用重构损失和噪声约束的概率形式进行优化,其表达式为

$$\mathrm{loss}_{\mathrm{VGAE}}=\mathrm{E}_{q(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{A})}[\ln p(\boldsymbol{A}\mid\boldsymbol{Y})]-\mathrm{KL}[q(\boldsymbol{Y}\mid\boldsymbol{X},\boldsymbol{A})\parallel p(\boldsymbol{Y})],\tag{7}$$

式中,$\mathrm{KL}[q(\cdot)\|p(\cdot)]$为 $q(\cdot)$ 和 $p(\cdot)$ 的 KL 散度,$p(\boldsymbol{Y})$表示高斯先验:

$$p(\boldsymbol{Y})=\prod_i p(\boldsymbol{y}_i)=\prod_i \mathcal{N}(\boldsymbol{y}_i\mid 0,\boldsymbol{I})。\tag{8}$$

式中,$\mathcal{N}$为高斯先验分布。

2.2 自监督约束

本节使用图的拓扑结构和节点属性信息生成自监督信息,并基于 InfoNCE 设计用于图表示学习模型训练的自监督约束。对于拓扑结构信息,通常使用邻接矩阵 $\boldsymbol{A}$ 表示,例如节点 i 和 j 之间存在边则邻接矩阵元素 A_{ij} 的值为 1,节点 i 和 j 之间无边则 A_{ij} 的值为 0,并且无向图中 A_{ij} 和 A_{ji} 的值相同。可见,邻接矩阵本身就包含了正负关系,即有边节点作为正样本,无边节点作为负样本。因此,在生成拓扑自监督信息时,不仅可以使用邻接矩阵构建一阶邻域关系的正负样本,还可以构建高阶邻域关系的正负样本:

$$\boldsymbol{\omega}^{\mathrm{T}}=\bar{\boldsymbol{A}}^{r},\tag{9}$$

式中,$\bar{\boldsymbol{A}}$ 表示归一化的邻接矩阵 $\boldsymbol{A}$,r 表示 $\bar{\boldsymbol{A}}$ 的阶数,$\boldsymbol{\omega}^{\mathrm{T}}$ 表示节点间的 r 阶邻域关系:

$$\boldsymbol{\omega}_{ij}^{\mathrm{T}}\begin{cases}=0,\text{节点 } j \text{ 不是节点 } i \text{ 的 } r \text{ 阶邻居}\\ \neq 0,\text{节点 } j \text{ 是节点 } i \text{ 的 } r \text{ 阶邻居}\end{cases}。\tag{10}$$

基于 InfoNCE 和 $\boldsymbol{\omega}^{\mathrm{T}}$ 构建拓扑对比约束 $\mathrm{loss}_{\mathrm{T}}$:

$$\mathrm{loss}_{\mathrm{T}}=-\log\frac{\sum_{j=1}^{N}\mathbf{1}_{[j\neq i]}\boldsymbol{\omega}_{ij}^{\mathrm{T}}\exp(\mathrm{sim}(y_i,y_j))}{\sum_{k=1}^{N}\mathbf{1}_{[k\neq i]}\exp(\mathrm{sim}(y_i,y_k))},\tag{11}$$

式中,sim 表示相似度函数(本文使用余弦相似度)。拓扑对比约束将每个节点的 r 阶邻居作为正样本,其他节点为负样本,从而激励正样本向量表示接近目标节点,负样本向量表示远离目标节点。

对于节点属性信息,通常使用属性矩阵 $\boldsymbol{X}$ 表示,例如节点 i 有属性 j 则属性矩阵元素 $\boldsymbol{X}_{ij}$ 的值为 1,节点 i 没有属性 j 则属性矩阵元素 $\boldsymbol{X}_{ij}$ 的值为 0。可见,属性矩阵表示的是节点与属性之间的关系,不能直接构建节点间的正负关系。因此,在生成属性自监督信息时,首先要对节点在属性空间的相似性进行度量:

$$S_{ij}=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{|\boldsymbol{x}_j||\boldsymbol{x}_j|},\tag{12}$$

式中,$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 表示任意两个节点 i 和 j 的属性向量,S_{ij} 表示向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的余弦相似度值。为了有效保留属性空间的自监督信息,避免过低的属性相似值带来噪声,本文根据 $\boldsymbol{S}$ 中数据元素 S_{ij} 的大小为每个节点选取前 k 个最相似的节点保留属性相似度值(其余位置为 0),构建属性相似度矩阵 $\boldsymbol{S}^{(k)}$。此时 $\boldsymbol{S}^{(k)}$ 包含了关于属性的正负关系,能够构建属性空间的正负样本:

$$\boldsymbol{\omega}^{\mathrm{F}}=\bar{\boldsymbol{S}}^{(k)},\tag{13}$$

式中,$\bar{\boldsymbol{S}}^{(k)}$ 表示归一化的属性相似度矩阵 $\boldsymbol{S}^{(k)}$,$\boldsymbol{\omega}^{\mathrm{F}}$ 表示节点在属性空间的关系:

$$\boldsymbol{\omega}_{ij}^{\mathrm{F}}\begin{cases}=0,\text{节点 } j \text{ 与节点 } i \text{ 无属性关联}\\ \neq 0,\text{节点 } j \text{ 与节点 } i \text{ 存在属性关联}\end{cases}。\tag{14}$$

基于 InfoNCE 和 $\boldsymbol{\omega}^{\mathrm{F}}$ 构建属性对比约束 $\mathrm{loss}_{\mathrm{F}}$:

$$\mathrm{loss}_{\mathrm{F}}=-\log\frac{\sum_{j=1}^{N}\mathbf{1}_{[j\neq i]}\boldsymbol{\omega}_{ij}^{\mathrm{F}}\exp(\mathrm{sim}(y_i,y_j))}{\sum_{k=1}^{N}\mathbf{1}_{[k\neq i]}\exp(\mathrm{sim}(y_i,y_k))}。\tag{15}$$

属性对比约束将存在属性关联的节点作为正样本,其他节点为负样本,鼓励正样本向量表示接近目标节点,推动负样本向量表示远离目标节点。

在 VGAE 模型优化过程中,将 $\mathrm{loss}_{\mathrm{VGAE}}$ 与基于拓扑和属性信息构建的自监督约束 $\mathrm{loss}_{\mathrm{T}}$ 和 $\mathrm{loss}_{\mathrm{F}}$ 进

行组合,构建完整的损失函数 $\text{loss}_{\text{Final}}$:

$$\text{loss}_{\text{Final}}=\text{loss}_{\text{VGAE}}+\alpha\text{loss}_{\text{T}}+\beta\text{loss}_{\text{F}}, \tag{16}$$

式中,α 和 β 是平衡自监督约束的加权系数。

3 结果与分析

3.1 数据集

本文使用基准引文网络数据集 Cora、Citeseer、Pubmed[25]评估基线模型和 Self-VGAE 在图分析任务中的实验表现。在数据集中,每个节点表示一篇论文,每条边表示不同论文间的引用关系,节点属性是论文内容的向量表示,节点标签表示论文的研究主题。数据集统计信息如表 1 所示。

表 1 数据集统计信息

Table 1 Statistics information of datasets

数据集 Datasets	#节点 #Nodes	#链接 #Edges	#属性 #Attributes	#标签 #Labels
Cora	2 708	5 429	1 433	7
Citeseer	3 327	4 732	3 703	6
Pubmed	19 717	44 338	500	3

3.2 基线模型

本文使用 VGAE 及其较为先进的变体作为基线。

VGAE:以 VAE 为基础架构,编码器使用 GCN 生成节点表示的均值和方差向量,解码器使用节点特征表示的内积恢复邻接关系,训练过程中使用重构损失和 KL 散度对模型参数进行优化。

Linear-VGAE:在 VGAE 的基础上,使用线性编码的 SGC 代替 GCN,并取消了层间激活函数,通过线性神经网络编码图信息,从而提升模型计算效率。

GC-VGAE[26]:在 VGAE 的基础上,对模型优化方式进行修改,在通过内积解码器恢复邻接矩阵的同时,引入与编码器对称的属性解码器,增强模型对拓扑结构和节点属性信息的保留。

MSVGAE:在 VGAE 的基础上,对模型编码过程进行修改,构造不同尺度的图信息,实现以多尺度和等变的方式学习特征表示,并通过图属性信息增强图结构的学习。

OSA-VGAE[8]:在 VGAE 的基础上,对编码器网络结构进行修改,通过 One-Shot 聚合改善多层 GCN 的信息传递,增强深层模型对原始图信息的表征能力。

3.3 评价指标

在节点分类实验中,使用常见的分类指标 Micro-F1 和 Macro-F1 进行比较:

$$\text{Micro-F1}=\frac{2\times P\times R}{P+R},$$

$$\text{Macro-F1}=\frac{\sum_{l\in L}\text{F1}(l)}{|L|}, \tag{17}$$

式中,P 表示精确率(Precision),R 表示召回率(Recall),F1(l)是标签 l 的 F1 分数。Micro-F1 在计算过程中考虑了每个类别中节点的数量,适用于数据分布不平衡的情况;而 Macro-F1 计算过程中没有考虑节点数量差异,而是平等地看待每一类,受高 P 值和高 R 值类的影响较大。

在节点聚类实验中,采用归一化互信息(Normalized Mutual Information,NMI)评估模型性能:

$$\text{NMI}=1-\frac{H(M_1\mid M_2)_{\text{norm}}+H(M_2\mid M_1)_{\text{norm}}}{2}, \tag{18}$$

式中,$H(\cdot)$表示信息熵。NMI 用于度量 M_1 和 M_2 聚类结果之间的相似性,值越大表明和真实结果共享的信息越多,聚类效果越好。

在链接预测中,对节点之间的边和非边进行预测,因此采用 ROC 曲线下面积(Area Under the ROC,AUC)和平均精度(Average Precision,AP)进行评估。AUC 同时考虑了分类器对于正例和负例的分类能力,把阈值设置在紧靠每个正例之下,计算负类的查全率后再取平均值,能够在样本不平衡的情况下对分类器做出合理的评价。AP 是把阈值设置在紧靠每个正例之下,计算正类的查准率后再取平均值,能够衡量模型在每个类别上的分类性能。

3.4 实验设置

为保证公平性,所有基线模型及 Self-VGAE 均参照 VGAE 原文中设置的参数进行初始化,编码器隐藏层维度设置为 32,VAE 的均值和方差向量维度为 16,训练过程中使用 Adam 优化器更新模型参数,学习率 lr 设为 0.01,迭代次数 n 设为 200。此外,Self-VGAE 的邻接矩阵阶数 r 在{1,5,10,15,20}中搜索,属性相似邻居数 k 在{1,5,10,15,20}中搜索,自监督约束系数 α 和 β 在{0.05,0.10,0.15,0.25,0.50,1.00}中搜索。综合节点分类、节点聚类和链接预测 3 个任务中的实验表现,Self-VGAE 的参数设置如表 2 所示。

表 2　Self-VGAE 参数设置

Table 2　Parameter settings of Self-VGAE

数据集 Datasets	*lr*	*n*	*r*	*k*	*α*	*β*
Cora	0.01	200	1	10	0.10	0.10
Citeseer	0.01	200	2	10	0.10	0.10
Pubmed	0.01	200	1	15	0.10	0.15

3.5　节点分类

本节通过节点分类任务评估模型性能,将各模型生成的低维节点表示作为支持向量机的输入,以 5%的间隔随机抽取 5%－25%的标签作为训练集,剩余节点中随机抽取 50%的标签作为测试集,各模型采用相同的数据集划分,记录 Micro-F1 和 Macro-F1。实验结果如图 3 至图 5 所示。

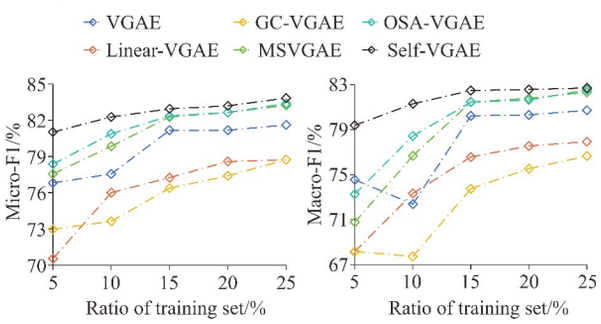


图 3　节点分类实验结果(Cora)

Fig. 3　Results of node classification experiment (Cora)

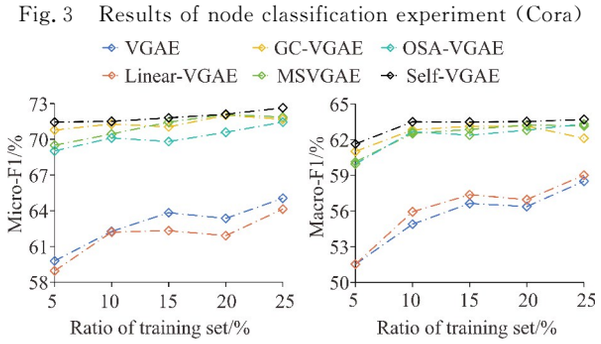


图 4　节点分类实验结果(Citeseer)

Fig. 4　Results of node classification experiment (Citeseer)

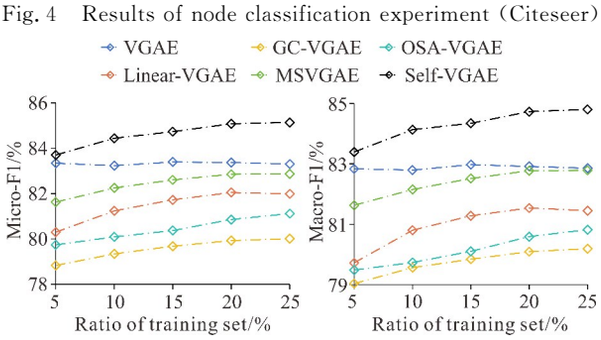


图 5　节点分类实验结果(Pubmed)

Fig. 5　Results of node classification experiment (Pubmed)

在 3 个数据集上,Self-VGAE 的 Micro-F1 和 Macro-F1 曲线始终高于基线模型,表明基于拓扑结构和节点属性构建的自监督信息能够增强模型对原始图特征信息的表征能力,使生成的低维表示中保留了丰富的节点相似性和差异性信息,进而提升节点分类任务的实验表现。基线模型中,Linear-VGAE 线性编码方式虽然降低了计算复杂度,但是在无监督条件下未能有效提取原始图的结构和属性信息,多数情况下实验表现弱于使用 GCN 编码器的 VGAE;GC-VGAE 通过恢复邻接矩阵和属性矩阵的方式保留原始图相关信息,但是在优化过程中缺少对拓扑和属性信息的平衡,使模型在不同数据集上的分类表现差异较大;保留多尺度拓扑结构信息的 MSVGAE 表现较为稳定,但仅从结构角度设计的表征方法忽略了属性信息的提取,限制了模型的分类性能。

在不同数据集上,同一基线分类结果差异较大。例如 OSA-VGAE 在 Cora 和 Citeseer 上表现出色,但是在 Pubmed 上表现较差。与基线模型不同,Self-VGAE 在 3 个数据集上均取得了最佳的实验结果,反映出模型较为强大的泛化能力。此外,随着训练集中标签数量的增加,各模型分类结果总体呈上升趋势,但部分基线模型的改善不明显,例如 VGAE 在 Pubmed 数据集上的预测结果随标签数量增加未见提升,表明基线模型生成的节点表示在保留相似性和差异性信息上能力有限,不能随监督信息的增加提升模型的分类性能。相较基线模型,Self-VGAE 能够充分利用拓扑结构和节点属性确定每个节点所属类别,仅使用 5%的训练标签即可取得良好的表现(表 3),表明模型能够保留丰富的原始图信息,增大节点特征表示的可区分性。

3.6　节点聚类

本节通过节点聚类任务评估模型性能,将各模型生成的低维节点表示作为 *K*-means＋＋算法的输入,实现无监督节点聚类,并记录 NMI(图 6)。在 *K*-means＋＋算法中,*K* 值设置为 Cora、Citeseer 和 Pubmed 数据集节点标签的类别数。

在 3 个数据集上,Self-VGAE 的 NMI 始终高于基线模型,表明基于拓扑结构和节点属性构建的自监督信息能够增强模型对社区结构拓扑信息的保留能力,使生成的低维表示中保留了丰富的有效社区信息,进而提升节点聚类任务的表现。基线模型中,Linear-VGAE 的线性编码方式在无监督条件下未能有效提取图的高阶结构特征,在多数情况下实验表现

表 3　节点分类实验结果(5%训练数据)

Table 3　Results of node classification experiment (5% training data)　　Unit:%

模型 Models	Cora		Citeseer		Pubmed	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
VGAE	76.81	74.55	59.78	51.52	83.34	82.83
Linear-VGAE	70.53	68.16	58.94	51.53	80.30	79.72
GC-VGAE	72.97	68.19	70.77	61.00	78.83	79.03
MSVGAE	77.55	70.79	69.50	60.13	81.63	81.63
OSA-VGAE	78.36	73.28	69.02	59.95	79.74	79.48
Self-VGAE	**81.02**	**79.39**	**71.44**	**61.63**	**83.70**	**83.39**

Note: bold is the best result, underlined is the second best result.

弱于使用 GCN 编码的其他模型;GC-VGAE 在优化过程中平等地计算重构拓扑结构和节点属性的损失函数,但缺少对社区结构等高阶拓扑信息的关注,因此限制了其在聚类任务中的实验表现;MSVAGE 通过编码多组不同尺度的原始图信息,实现多尺度和等变特征学习,同时采用节点属性增强拓扑结构学习的策略,使模型在一定程度上保留了高阶结构特征,取得了较好的聚类表现。

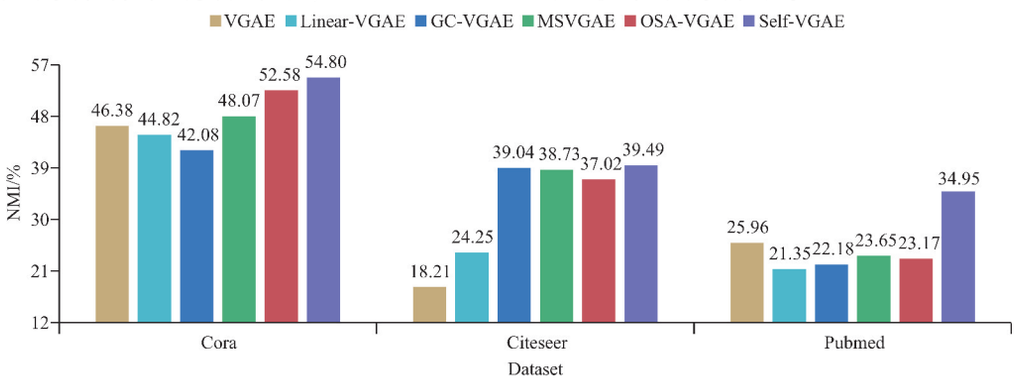


图 6　节点聚类实验结果

Fig. 6　Results of node clustering experiment

在不同数据集上,多数基线模型的聚类结果差异较大,特别是 OSA-VGAE 模型在聚类任务中的实验表现与分类任务相同,Cora 和 Citeseer 数据集上的实验表现优于原始的 VGAE 模型,但在 Pubmed 数据集上的实验表现较差,反映出该模型采用的信息传递方式对不同数据集的泛化能力有限。相反,MSVGAE 和 Self-VGAE 能够保留更丰富的拓扑结构信息,在聚类任务中的实验表现更加稳定,特别是 Self-VGAE 通过自监督信息使同一社区内的节点在特征空间更加接近,因此在 3 个数据集上均获得了最佳的聚类表现。

3.7　链接预测

通过链接预测任务评估模型实验性能。首先移除数据中 20%的链接并随机采样 20%的非链接(无链接节点对)构建测试集,然后使用剩余 80%的链接进行训练,最后利用生成的节点表示内积重构邻接矩阵,对节点间是否存在链接关系进行预测,各模型采用相同的数据集划分,记录 AUC 和 AP(表 4)。

在 3 个数据集上,Self-VGAE 的 AUC 和 AP 始终高于基线模型,表明基于拓扑结构和节点属性构建的自监督信息能够增强模型对拓扑结构信息的保持和推断能力,使生成的低维表示中保留了丰富的邻域信息,进而提升链接预测任务的表现。基线模型中,Linear-VGAE 以线性编码方式提取节点邻域关系的能力有限,在多数情况下弱于使用 GCN 编码的其他模型;GC-VGAE 增加了对属性信息的保留,但节点表示中过多的属性信息对拓扑结构预测产生了一定负面影响,使模型在 Cora 和 Citeseer 两个数据集上的实验表现较差;MSVGAE 使用的多尺度特征学习和属性增强策略在一定程度上增强了局部结构信息的保持能力,但模型仅在 Cora 和 Citeseer 数据集上表现较好,而在 Pubmed 数据集上表现较差,在链接预测任务中泛化能力有限;引入跨层信息传递策略的 OSA-VGAE 在节点分类和聚类任务中表现不佳,但

是在链接预测任务中取得了较好的实验结果。

表 4　链接预测实验结果

Table 4　Results of link prediction experiment　　**Unit:%**

模型 Models	Cora		Citeseer		Pubmed	
	AUC	AP	AUC	AP	AUC	AP
VGAE	89.60	90.87	90.46	91.90	92.60	93.21
Linear-VGAE	89.36	90.64	89.07	90.91	93.86	94.07
GC-VGAE	90.11	90.29	91.77	91.27	93.96	94.01
MSVGAE	91.93	91.97	93.66	93.79	92.05	92.21
OSA-VGAE	92.26	92.74	94.08	94.38	93.78	94.32
Self-VGAE	**93.50**	**93.65**	**95.56**	**95.92**	**94.84**	**94.95**

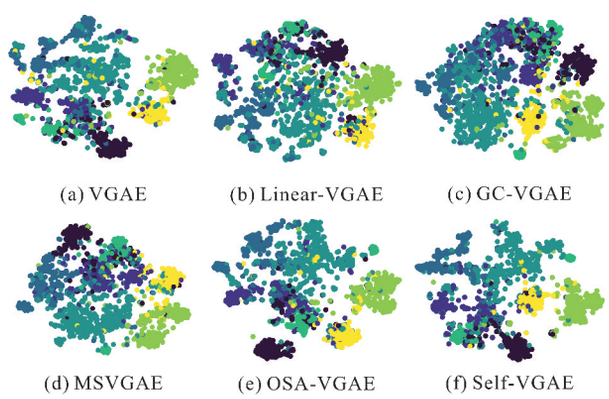
Note: bold is the best result, underlined is the second best result.

与基线模型相比,Self-VGAE 在不同数据集上的链接预测任务中同样具有较强的泛化能力。此外,Self-VGAE 和其他基线模型均由 VGAE 演化而来,区别在于采用不同的方式对模型进行改进。在 3 个数据集上,Self-VGAE 在节点分类、节点聚类和链接预测任务中的实验结果相较 VGAE 均有较为明显的提升,不仅优于当前先进的基线模型,而且在不同任务中的表现也更加稳定,证明了自监督信息的引入能够增强模型对原始图信息的表征能力,使节点表示能够保留基于节点属性和拓扑结构相似性信息以及局部和高阶结构特征,进而提升下游图分析任务的实验表现。

3.8　可视化

节点特征表示蕴含了节点属性和拓扑结构信息,可视化后能直观地反映原始图的某些特征以及模型对原始图信息的表征能力。本节通过可视化任务对模型进行评估。首先使用 t-SNE 将基线和 Self-VGAE 模型在 Cora 数据集上生成的节点特征表示降至 2 维,然后根据节点标签将二维平面上的数据点标记为 7 种不同的颜色(图 7)。

好的可视化结果通常相同颜色节点彼此接近,不同颜色节点彼此分离。由图 7 可知,所有模型均能从任意分布的原始数据中提取相关信息并形成一定的社区结构,但是不同模型可视化结果的类内相似性和类间界限区别较大。其中,Linear-VGAE、GC-VGAE 和 MSVGAE 的节点分布松散、类间差异不够明显,相反,VGAE、OSA-VGAE 和 Self-VGAE 的类内节点分布更加接近,类间距离更远,特别是 Self-VGAE 在多个类别上均有较好的可视化效果。可视化实验直观反映了模型保留同一社群节点相似特征的能力,证明了自监督信息的引入增强了模型对原始图信息的表征能力,能够使同类节点表示在特征空间更加接近。



(a) VGAE　(b) Linear-VGAE　(c) GC-VGAE

(d) MSVGAE　(e) OSA-VGAE　(f) Self-VGAE

图 7　节点表示可视化

Fig. 7　Visualization of node representations

3.9　参数分析

为了分析 Self-VGAE 性能受参数的影响,通过 Cora 和 Citeseer 数据集上的链接预测任务进行参数实验,并记录实验结果。相较学习率、迭代次数、嵌入维度等常见参数,本节仅对 Self-VGAE 特有参数邻接矩阵阶数 r、属性相似邻居数 k、自监督约束权重系数 α 和 β 进行验证。为了保证实验的公平性,除验证参数外,其余参数均按照表 2 进行设置。

为了验证邻接矩阵阶数 r 对 Self-VGAE 性能的影响,使用不同的 r 值进行实验(图 8)。邻接矩阵阶数 r 与图的高阶结构特征相关,例如三阶邻接矩阵蕴含了节点的三阶邻域关系,更高阶的邻接矩阵能够保留更高阶的邻域关系。从实验结果看,不同数据集上最优的 r 值不同,但是在 r 值较大时模型的实验表现均受到限制,这是因为较高的 r 值使得节点邻域差异性减弱,每个节点聚合了过多远距离节点的特征信息,对低阶邻域的局部特征信息保留有限,降低了实

验性能。

为了验证属性相似邻居数 k 对 Self-VGAE 性能的影响，使用不同的 k 值进行实验(图 9)。属性相似邻居数 k 与节点在属性空间的相似性有关，例如 k 值为 3 表示在属性空间为每个节点选择 3 个最接近的节点作为属性自监督信息。从实验结果看，两个数据集总体上均呈现先上升再下降的趋势，过高和过低的 k 值均未取得最佳的实验结果，这是因为 k 值过低无法保留充足的属性相似度信息，而 k 值过高导致大量属性相似度较低的噪声节点作为自监督信息，降低了实验性能。

为了验证自监督约束权重系数 α 和 β 对 Self-VGAE 性能的影响，使用不同的 α 和 β 值进行实验(图 10)。权重系数 α 和 β 用于调整自监督信息在模型训练过程中的重要程度，使节点表示关注拓扑和属性信息的保留。从实验结果看，当一个系数固定时，另一个系数的实验表现随数值的增大呈现先上升再下降的趋势，这是因为过低的系数无法有效利用拓扑或属性自监督信息，而过高的系数使模型仅侧重于保留拓扑或属性自监督信息，降低了模型对原始图信息的表征能力。实验结果表明，当 $\alpha=0$ 或 $\beta=0$(仅使用拓扑或属性作为自监督信息)、α 和 β 均取较大值时，模型的实验表现均不佳，因此在优化过程中需要设置合适的权重系数 α 和 β，平衡损失函数中拓扑与属性对比约束的比重。

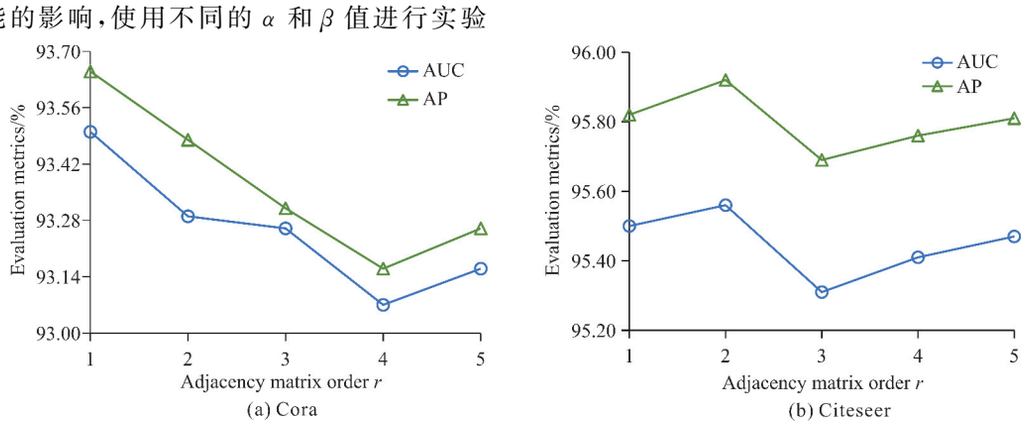


图 8　参数 r 实验结果

Fig. 8　Experimental results of parameter r

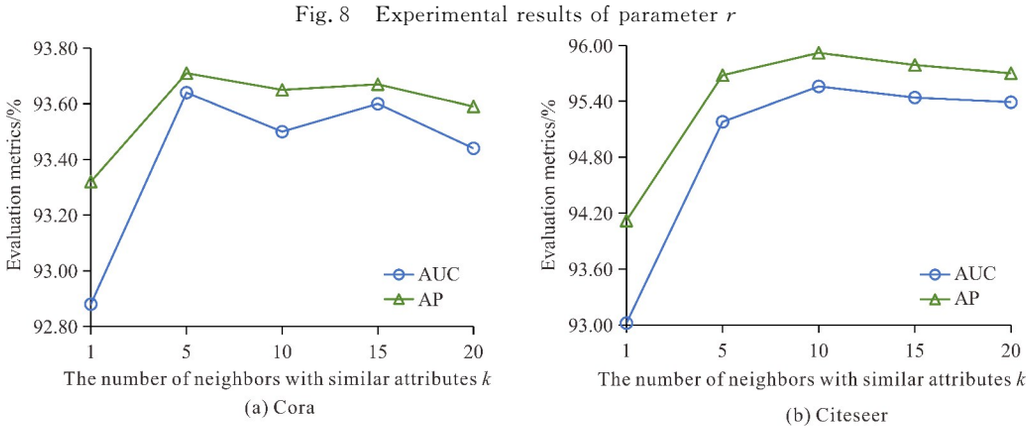


图 9　参数 k 实验结果

Fig. 9　Experimental results of parameter k

3.10　训练时间

为了比较不同模型的训练复杂度，Cora 数据集上迭代 100 次后单次迭代的平均训练时间(包括前向传播、损失函数计算、反向传播过程)如图 11 所示。以原始的 VGAE 模型训练时间为基准，Linear-VGAE 采用线性编码器进行特征提取，加快了模型训练速度；引入跨层连接的 OSA-VGAE 增加了层间特征融合步骤，使得训练速度略微增加；计算属性损失的 GC-VGAE 和计算多尺度信息的 MSVGAE 平均训练时间明显高于 VGAE。本文提出的 Self-VGAE 虽然采用与 VGAE 相同的网络结构，但是为了充分保留原始图相关信息，额外借助属性和拓扑自监督约束对节点表示生成过程进行优化，导致模型的训练时间增加。

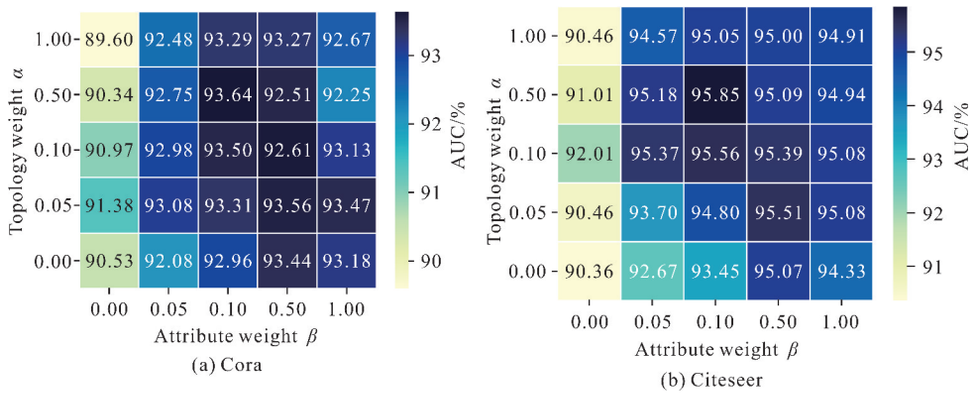


图 10　参数 α 和 β 实验结果

Fig. 10　Experimental results of parameter α and β

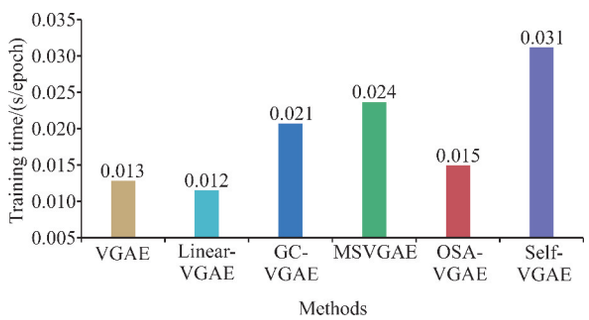


图 11　模型训练时间

Fig. 11　Model training time

4　结论

本文提出一种基于自监督信息的图表示学习模型 Self-VGAE,通过同时关注拓扑结构和节点属性的自监督约束增强了 Self-VGAE 对原始图信息的表征能力。实验结果表明,Self-VGAE 生成的节点表示具有较强的通用性和泛化性,能够有效提升模型在 3 个基准数据集上的多个图分析任务的实验表现,并优于当前较为先进的基线模型。此外,相较以往依赖具体任务进行特征保留的图表示学习模型,Self-VGAE 能够同时增强模型对节点相似性和差异性的保留能力、对拓扑结构的保持和推断能力以及高阶社区结构的表征能力。Self-VGAE 仅通过改进模型优化方式提升实验性能,未考虑使用不同的编码器增强模型对原始数据特征的提取能力。因此,在后续工作中将引入多种改进方式,同时提升模型对原始图信息的提取和保留能力,并将其应用于社交网络安全问题检测和犯罪团伙挖掘等实际任务。

参考文献

[1] GHAREHCHOPOGH F S. An improved Harris hawks optimization algorithm with multi-strategy for community detection in social network [J]. Journal of Bionic Engineering,2023,20(3):1175-1197.

[2] DIVIÁK T. Structural resilience and recovery of a criminal network after disruption: a simulation study [J/OL]. Journal of Experimental Criminology,2023 (2023-03-24)[2023-11-11]. https://doi.org/10.1007/s11292-023-09563-z.

[3] PENG C Y,XIA F,NASERIPARSA M,et al. Knowledge graphs:opportunities and challenges [J]. Artificial Intelligence Review,2023,56:13071-13102.

[4] XU M J. Understanding graph embedding methods and their applicationsm [J]. SIAM Review,2021,63(4):825-853.

[5] 邹然,柳杨,李聪,等. 图表示学习综述[J]. 北京师范大学学报(自然科学版),2023,59(5):716-724.

[6] 徐冰冰,岑科廷,黄俊杰,等. 图卷积神经网络综述[J]. 计算机学报,2020,43(5):755-780.

[7] CONG S,ZHOU Y. A review of convolutional neural network architectures and their optimizations [J]. Artificial Intelligence Review,2023,56(3):1905-1969.

[8] 袁立宁,刘钊. 基于 One-Shot 聚合自编码器的图表示学习[J]. 计算机应用,2023,43(1):8-14.

[9] HE M Y,ZHAO Q Q,ZHANG H. Multi-sample dual-decoder graph autoencoder [J]. Methods,2023,211:31-41.

[10] HY T S,KONDOR R. Multiresolution equivariant graph variational autoencoder [J]. Machine Learning:Science and Technology,2023,4(1):015031.

[11] WANG Y F,XU B Y,KWAK M,et al. A noise injection strategy for graph autoencoder training [J]. Neural Computing and Applications,2021,33:4807-4814.

[12] HUANG T J,PEI Y L,MENKOVSKI V,et al. On generalization of graph autoencoders with adversarial training [C]//Joint European Conference on Machine Learning and Principles and Practice of Knowledge Dis-

covery in Databases. Cham:Springer,2021:367-382.
[13] KIPF T N,WELLING M. Variational graph auto-encoders [EB/OL]. (2016-11-21)[2023-11-11]. http://arxiv.org/abs/1611.07308.
[14] 翟正利,梁振明,周炜,等.变分自编码器模型综述[J].计算机工程与应用,2019,55(3):1-9.
[15] SALHA G,HENNEQUIN R,VAZIRGIANNIS M. Simple and effective graph autoencoders with one-hop linear models [C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham:Springer,2021:319-334.
[16] WU F,SOUZA A,ZHANG T Y,et al. Simplifying graph convolutional networks [C]//Proceedings of the 36th International Conference on Machine Learning. Cambridge:PMLR,2019:6861-6871.
[17] KESER R K,NALLBANI I,ÇALIK N,et al. Graph embedding for link prediction using residual variational graph autoencoders [C]//Proceedings of the 28th Signal Processing and Communications Applications Conference. Piscataway:IEEE,2020:1-4.
[18] RANI V,NABI S T,KUMAR M,et al. Self-supervised learning:a succinct review [J]. Archives of Computational Methods in Engineering:State of the Art Reviews,2023,30(4):2761-2775.
[19] WANG J,LIANG J,YAO K,et al. Graph convolutional autoencoders with co-learning of graph structure and node attributes [J]. Pattern Recognition,2022,121:108215.
[20] FETTAL C,LABIOD L,NADIF M. Efficient graph convolution for joint node representation learning and clustering [C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. New York:ACM,2022:289-297.
[21] LI D J,LI D,LIAN G. Variational graph autoencoder with adversarial mutual information learning for network representation learning [J]. ACM Transactions on Knowledge Discovery from Data,17(3):45.
[22] XIE Y C,XU Z,ZHANG J T,et al. Self-supervised learning of graph neural networks:a unified review [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(2):2412-2429.
[23] KUMAR P,RAWAT P,CHAUHAN S. Contrastive self-supervised learning:review,progress,challenges and future research directions [J]. International Journal of Multimedia Information Retrieval,2022,11(4):461-488.
[24] HU Y,YOU H,WANG Z,et al. Graph-MLP:node classification without message passing in graph [EB/OL]. (2021-06-08)[2023-11-11]. http://arxiv.org/abs/2106.04051.
[25] 袁立宁,李欣,王晓冬,等.图嵌入模型综述[J].计算机科学与探索,2022,16(1):59-87.
[26] GUO L,DAI Q. Graph clustering via variational graph embedding [J]. Pattern Recognition,2022,122:108334.

Graph Representation Learning Enhanced by Self-supervised Information

YUAN Lining[1,2],WEN Zhu[2**],FENG Wengang[1],LIU Zhao[3]
(1. School of National Security,People's Public Security University of China,Beijing,100038,China;2. School of Information Technology,Guangxi Police College,Nanning,Guangxi,530028,China;3. Graduate School,People's Public Security University of China,Beijing,100038,China)

Abstract:Graph representation learning models rely on specific task to preserve features,and the generalization of node representations are limited. Aiming at the above problems,a graph representation learning model Self-Variational Graph Auto Encoder (Self-VGAE) enhanced by self-supervised information is proposed in this article. Firstly,graph convolutional encoder and node representation inner product decoder are used to

construct a VGAE. The feature extraction and coding of the original graph are performed. Then, the topology and node attributes are used to generate self-supervised information, and the generation of node representation is constrained during model training. In multiple graph analysis tasks, the experimental performance of Self-VGAE is better than the current more advanced baseline model, which shows that the introduction of self-supervised information can enhance the ability to retain the similarity and difference of node features and the ability to maintain and infer the topology. Furthermore, Self-VGAE has a stronger generalization ability.
Key words: self-supervised information; graph representation learning; graph variational auto encoders; graph convolutional networks; contrastive loss

责任编辑:陆　雁,陈少凡