基于多维云概念嵌入的变分图自编码器研究

代 劲[1,2],张奇瑞[2],王国胤[1,3],彭艳辉[2],涂盛霞[4]
(1. 重庆邮电大学计算智能重庆市重点实验室,重庆 400065;2. 重庆邮电大学软件工程学院,重庆 400065;
3. 重庆邮电大学旅游多源数据感知与决策技术文化和旅游部重点实验室,重庆 400065;4. 华为技术有限公司,广东深圳 518129)

摘 要: 变分图自编码器是图嵌入研究中重要的深度学习模型,但存在着先验正态分布缺陷、训练过程中容易出现后验塌陷等问题. 本文从建立云概念空间与隐空间的映射关系入手,引入云模型数字特征对网络中的节点进行不确定性概念表示,设计了一种基于多维云模型的变分图自编码器(Variational Graph Autoencoder based on Multidimensional Cloud Model,MCM-VGAE). 该模型实现了隐空间的多维云概念嵌入及相应的漂移性损失度量,将先验分布扩展为泛正态分布,利用多维正向云发生器及云包络带修正采样算法实现了重参数化过程,有效缓解了后验塌陷现象. 在应用效果上,模型在多类型数据集上的链路预测、节点聚类、图嵌入可视化实验表现均优于基准模型,进一步说明了方法的普适有效性.

关键词: 变分图自编码器;图嵌入;多维云模型;概念嵌入;链路预测
基金项目: 国家自然科学基金(No.61936001,No.61772096);重庆市自然科学基金(No.cstc2021jcyj-msxmX0849)
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112(2023)12-3507-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220354

Research on Variational Graph Auto-Encoder Based on Multidimensional Cloud Concept Embedding

DAI Jin[1,2], ZHANG Qi-rui[2], WANG Guo-ying[1,3], PENG Yan-hui[2], TU Sheng-xia[4]
(1. *Chongqing Key Laboratory of Computation Intelligence*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
2. *School of Software Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
3. *Key Laboratory of Tourism Multisource Data Perception and Decision*, *Ministry of Culture and Tourism*, *Chongqing University of Posts and Telecommunication*, *Chongqing* 400065, *China*;
4. *Huawei Technologies Co.*, *Ltd.*, *Shenzhen*, *Guangdong* 518129, *China*)

Abstract: Variational graph autoencoder (VGAE) is a significant deep learning model in graph embedding, but there are problems such as the normal prior distribution defect and the posterior collapse during training. Focusing on establishing the mapping relationship between cloud concept space and hidden space, the uncertain concepts of nodes in VGAE network are represented by the digital features of cloud model, and an optimized VGAE model based on multidimensional cloud model (MCM-VGAE) is reconstructed. The model implements a multidimensional cloud concept embedding in the latent space and the corresponding drift loss measure, extends the prior distribution to a generic normal distribution, and uses a multidimensional forward cloud generator and a cloud envelope with modified sampling algorithm to realize the reparameterization process and effectively mitigate the posterior collapse phenomenon. In terms of application, the model outperforms the benchmark model for link prediction, node clustering, and graph embedding visualization experiments on multi-type datasets, further illustrating the universal effectiveness of the method.

Key words: variational graph autoencoder; graph embedding; multidimensional cloud model; concept embedding; link prediction
Foundation Item(s): National Natural Science Foundation of China (No.61936001, No.61772096); Natural Science Foundation of Chongqing (No.cstc2021jcyj-msxmX0849)

收稿日期:2022-04-06;修回日期:2023-01-30;责任编辑:梅志强

1 引言

客观世界中的网络往往具有小世界效应和无尺度的复杂特性,网络拓扑结构的不确定性表示是复杂网络研究的基本问题[1]. 网络表示学习(又称为图嵌入)[2]作为一项重要的图数据挖掘工作,在实际网络分析任务中扮演着十分重要的角色. 近年来,随着图神经网络(Graph Neural Networks,GNN)[3]技术的迅速发展,基于自编码器的无监督生成式图神经网络模型成为了图嵌入工作中最具价值的研究方向之一.

目前,针对图嵌入中节点的不确定性表示和处理问题,正态分布凭借其本身所具备的不确定性特性在该领域中发挥着重要作用[1]. Kipf等人[4]借助图卷积神经网络(Graph Convolutional Networks,GCN)以及正态分布理论实现了节点的低维向量表示,提出的变分图自编码器(Variational Graph Autoencoders,VGAE)在链路预测、节点分类[5]任务中相比传统方法均表现出较高的准确率和更健壮的鲁棒性;后续有许多基于先验正态分布的VGAE变体模型相继被提出[6~8],由此证明了正态分布理论用于图嵌入工作的有效性.

然而,当前基于正态分布理论的VGAE及其变体模型仍存在诸多问题,例如:(1)VGAE存在先验分布缺陷. 虽然正态分布在复杂网络的研究中占有重要地位,但是如果决定随机现象的因素单独作用不是均匀的小,相互之间并不独立,就不符合正态分布的产生条件[9]. 此时若用正态分布来对图节点的表示做简单近似,就不能真实反映网络的客观情况;(2)VGAE对于孤立节点的嵌入处理缺乏有效性,无论孤立节点的特征如何变换,VGAE始终让这类节点的嵌入结果接近于零[7];(3)变分自编码器在训练中容易出现"后验分布塌陷"问题(又称为KL散度消失)[10],导致模型在采样时"强迫"隐变量特征接近均值中心,模型退化为自编码器.

对此,云模型[11]作为描述不确定性的双向认知模型,其通过数字特征(期望Ex、熵En、超熵He)构成特定的发生器,这种结构不但放宽了形成正态分布的前提条件,而且把精确确定的隶属函数放宽到构造正态隶属度分布的期望函数,因而更具普适性[9].

为改进图自编码器的图嵌入学习能力,本文拟在VGAE框架下,借助云模型对于不确定性问题表示的优秀处理能力,提出一种基于多维云模型的变分图自编码器框架(Variational Graph Autoencoder based on Multi-dimensional Cloud Model, MCM-VGAE). 该方法将先验分布扩展为泛正态,其形成条件远比正态分布宽松,更接近于客观事实,又远比联合分布简单、可行[9]. 同时,该方法结合前沿的GNN技术分别对模型的编码器、采样过程、损失度量进行改进. 力求MCM-VGAE较其他基准模型具备更好的图嵌入性能. 本文的主要工作及创新如下:

(1) 针对VGAE中先验正态分布的缺陷,提出借助云模型实现隐空间的多维云概念嵌入并扩展先验分布的策略,在此基础上结合前沿的GNN技术进一步提出基于多维云模型的变分图自编码器框架.

(2) 针对模型在训练过程中容易出现的"后验塌陷"问题,提出一种基于多维正向云发生器及云包络带修正的隐变量采样优化算法. 经理论与实验验证,该方法实现了重参数化过程且能够有效减小采样空间的云心密度,从而缓解"后验塌陷"现象.

(3) 通过多组开源图数据集上的链路预测、节点聚类、图嵌入可视化、超参数敏感性分析实验表明:MCM-VGAE较其他基准模型具备更出色的图嵌入学习能力,且具备良好的泛化能力.

2 相关工作

在基于自编码器的图神经网络研究过程中,Tian等人[12]最早提出了稀疏图自编码器(Sparse Auto-Encoder,SAE),其利用邻接矩阵表示图中的节点特征,并通过自编码器完成了矩阵分解工作. Kipf等人[4]将GCN与变分自编码器思想推广到图上,提出的图自编码器(Graph Auto-Encoder,GAE)以及VGAE取得了重要研究成果;Salha等人[8]提出一种结构更简化的线性变分图自编码器(Linear Graph Variational Auto-Encoder,Linear-VGAE);Ahn等人[7]为提高GAE对于孤立节点的嵌入表示能力,提出一种图归一化自编码器(Graph Normalized AutoEncoders,GNAE),该模型利用L_2正则化技术有效处理了孤立节点的嵌入过程.

目前,在图自编码器隐空间表示的优化研究中,为进一步提高自编码器的图嵌入生成能力,Davidson等人[13]利用超球面分布(von Mises-Fisher,vMF)替代VGAE的先验正态分布并改进损失度量函数,在链路预测实验中获得了更高的准确率;Pan等人[5]借助对抗正则化思想,提出一种基于对抗正则化的变分图自编码器(Adversarially Regularized Variational Graph Autoencoder,ARVGA);Hasanzadeh等人[6]提出一种半隐式变分图自编码器(Semi-Implicit Graph Variational Auto-encoders,SIG-VAE),该方法采用分层变分推理框架,相比VGAE的变分推理过程更加灵活. 然而,这些针对先验分布的优化算法实现过程较为复杂,算法的泛化能力有待进一步验证.

由于自然科学、社会科学中大量的随机现象都服从正态分布,正态云模型及其普遍适用性受到了人们的广泛关注[9]. 目前,围绕云模型的研究主要有:定性概念的不确定性知识表示、多层泛概念树等[11]. 众所

周知,深度学习模型结构与参数的选择具有较多的随意性,模型的训练过程充满不确定性. 因此,正态云模型及其普遍适用性应用于深度学习领域具有一定的优越性.

3　基于多维云模型的变分图自编码器

在介绍本文方法之前,此处首先给出基本定义.

定义 1　图(Graph)[3]:图由节点与连接节点的边构成,通常记为 $G=(V,E)$. 其中 $V=\{v_1,v_2,\cdots,v_n\}$ 代表节点集合, $E=\{e_1,e_2,\cdots,e_m\}$ 表示边集合. 通用的图表示是一个五元组形式:$G(V,E,\boldsymbol{A},\boldsymbol{X},\boldsymbol{D})$,其中 $\boldsymbol{A}^{N\times N}$ 表示图的邻接矩阵,$\boldsymbol{X}^{N\times f}$ 表示节点的特征矩阵,$\boldsymbol{D}^{N\times N}$ 表示节点的度矩阵,N 和 f 分别表示节点的数量和节点的特征维度.

定义 2　云模型[11]:云模型是用自然语言表示的某个定性概念与其定量数值之间的双向认知模型. 设定性概念 T 是定量论域 U 上的概念,若 $x\in U$ 是 T 的一次随机实现,x 对 T 的确定度 $\mu(x)\in[0,1]$ 是具有稳定分布的随机数:$\mu(x):U\to[0,1],\forall x\in U$,则 x 在论域 U 上的分布称为云模型. 正态分布与正态隶属度函数的普遍性共同奠定了正态云具有普适性,正态云已成为应用最为广泛的云模型.

定义 3　多维正态云[11, 14]:多维正态云由一维正态云推广而来,能够反映多维复杂定性概念. 设 $U=\{X_1,X_2,\cdots,X_m\}$ 表示 m 维定量论域,T 是 U 上的定性概念. 若 $x(x_1,x_2,\cdots,x_m)\in U$ 是概念 T 的一次随机实现,并满足 $x=R_N\left(\mathrm{Ex}(\mathrm{Ex}_1,\mathrm{Ex}_2,\cdots,\mathrm{Ex}_m),|y|\right)$, $y=R_N\left(\mathrm{En}(\mathrm{En}_1,\mathrm{En}_2,\cdots,\mathrm{En}_m),\mathrm{He}(\mathrm{He}_1,\mathrm{He}_2,\cdots,\mathrm{He}_m)\right)$,且 x 属于 T 的确定度 $\mu(x)$ 满足隶属度分布:

$$\mu\left[x(x_1,x_2,\cdots,x_m)\right]=\exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\frac{(x_i-\mathrm{Ex}_i)^2}{y_i^2}\right\}\tag{1}$$

则 $\mathrm{drop}(x_1,x_2,\cdots,x_m,\mu)$ 称为在论域 U 上的一个云滴,所有云滴构成的分布称为 m 维正态云,记作云 $C(\mathrm{Ex},\mathrm{En},\mathrm{He})$. 其中,Ex 是云滴在论域空间分布中的数学期望,即最能够代表定性概念 T 的点;熵 En 代表定性概念的不确定性度量;超熵 He 是熵的不确定性度量,反映了云滴的离散程度和确定度的随机性. $C(\mathrm{Ex},\mathrm{En},\mathrm{He})$ 也称作正态云概念,即用于描述定性概念 T 在正态云上的整体定量属性.

针对 VGAE 模型中存在的孤立节点嵌入的有效性问题、先验正态分布在实际应用中的普适性问题、模型在训练过程中存在的“后验塌陷”问题,本文拟借助云模型与 GNN 技术理论,设计一种高性能、具备良好普适性的基于多维云模型的变分图自动编码器(MCM-VGAE),工作内容主要包括 4 个部分:(1)基于多维云概念嵌入的图归一化编码器结构设计;(2)基于多维正向云发生器以及云包络修正的隐变量采样优化算法;(3)解码器及损失函数设计;(4)算法复杂性分析. MCM-VGAE 的模型框架如图 1 所示.

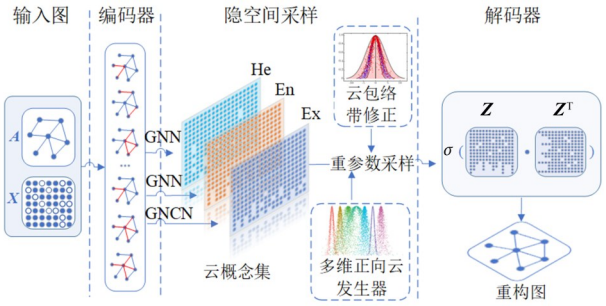


图 1　MCM-VGAE 模型整体框架

3.1　基于多维云概念嵌入的图归一化编码器

VGAE 内部的隐空间优化表征是提升模型学习能力的关键因素[13]. 针对 VGAE 中先验正态分布在实际应用中的缺陷问题,本文充分结合云模型的普适性特性,利用多维正态云对 VGAE 隐空间进行表示,在此基础上形成隐空间与云概念空间的相互映射,实现将先验分布扩展为泛正态分布、获得更高粒度的隐空间概念表征,即实现多维云概念的模型嵌入表示. MCM-VGAE 的编码器结构具体实现如下.

设 MCM-VGAE 的先验分布服从正态云的云滴分布,通过云概念的形式可描述为:先验正态云概念,记作 $C_{\mathrm{pr}}(\mathrm{Ex}_{\mathrm{pr}},\mathrm{En}_{\mathrm{pr}},\mathrm{He}_{\mathrm{pr}})$. 设输入图为 G,图卷积的嵌入维度为 d. 编码器对图 G 做嵌入处理,输出后验多维正态云的数字特征矩阵:期望 $\boldsymbol{C}_{\mathrm{Ex\_po}}^{N\times d}$、熵 $\boldsymbol{C}_{\mathrm{En\_po}}^{N\times d}$、超熵 $\boldsymbol{C}_{\mathrm{He\_po}}^{N\times d}$,且该 3 个特征距共同组成后验正态云概念集 $C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex\_po}},\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}})$,由此建立隐空间与云概念空间的相互映射关系. 编码器的整体形式如下:

$$\mathrm{Encoder}\begin{cases}\boldsymbol{C}_{\mathrm{Ex\_po}}=\mathrm{GNCN}_{Ex}(\boldsymbol{X},\boldsymbol{A},s),\ \boldsymbol{C}_{\mathrm{Ex\_po}}\in\mathbb{R}^{N\times d}\\ \ln\boldsymbol{C}_{\mathrm{En\_po}}=\mathrm{GNN}_{En}(\boldsymbol{X},\boldsymbol{A}),\ \boldsymbol{C}_{\mathrm{En\_po}}\in\mathbb{R}^{N\times d}\\ \ln\boldsymbol{C}_{\mathrm{He\_po}}=\mathrm{GNN}_{He}(\boldsymbol{X},\boldsymbol{A}),\ \boldsymbol{C}_{\mathrm{He\_po}}\in\mathbb{R}^{N\times d}\\ \text{latent space}\Leftrightarrow C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex\_po}},\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}})\end{cases}\tag{2}$$

针对 GCN 的结构设计,本文主要借助一种能够有效处理孤立节点嵌入的图归一化卷积网络(Graph Normalized Convolutional Network,GNCN)[7],本文在此基础上对网络结构进行了部分改进,实现如下:

设输入图 G 的节点特征矩阵表示为 $\boldsymbol{X}^{N\times f}$,邻接矩阵为 $\boldsymbol{A}^{N\times N}$,$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$ 表示增加了自循环的邻接矩阵,$\boldsymbol{I}_N$ 表示单位矩阵,图卷积的嵌入维度为 d. 针对后验期望特征矩阵 $\boldsymbol{C}_{\mathrm{Ex\_po}}$ 的生成,本文首先借助 Linear-

GCN[8]对图 G 做线性变换，输出特征矩阵 $\boldsymbol{H}^{N\times d}$，然后对 $\forall \boldsymbol{h}_i \in \boldsymbol{H}$ 做归一化处理，输出特征变换向量 $\boldsymbol{n}_i \in \boldsymbol{N}_\delta$，如式(3)所示：

$$\boldsymbol{n}_i = s\frac{\boldsymbol{h}_i}{\|\boldsymbol{h}_i\|}, s\in[0,1] \tag{3}$$

其中，s 表示缩放因子常数. 最后通过消息传递图神经网络(Approximate Personalized Propagation of Neural Predictions layer，APPNP)[15]处理特征矩阵 $\boldsymbol{N}_\delta^{N\times d}$，并输出后验期望 $\boldsymbol{C}_{\text{Ex\_po}}^{N\times d}$.

后验云概念集的数字特征熵 $\boldsymbol{C}_{\text{En\_po}}^{N\times d}$ 与超熵 $\boldsymbol{C}_{\text{He\_po}}^{N\times d}$ 的编码过程相同，计算形式如下：

$$\boldsymbol{C}_{\text{En\_po}} = \text{GNN}_{\text{En}}(\boldsymbol{X},\boldsymbol{A}) = \text{APPNP}\left(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_{\text{En}}\right) \tag{4}$$

$$\boldsymbol{C}_{\text{He\_po}} = \text{GNN}_{\text{He}}(\boldsymbol{X},\boldsymbol{A}) = \text{APPNP}\left(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_{\text{He}}\right) \tag{5}$$

式(4)、式(5)中，$\boldsymbol{W}$ 表示相互独立的待学习参数. 关于先验分布超参数、图卷积结构的相关超参数调优实验分析见本文第4节.

3.2 基于多维正向云发生器以及云包络带修正的隐变量采样优化算法

定义4 正态云模型包络带[16]：当 $\text{He}\neq\text{En}/3$ 时，正态云模型的外包络曲线与内包络曲线之间的区域与横轴围成的面积称为正态云模型包络带.

MCM-VGAE 的隐变量生成算法主要由两部分组成：(1)基于多维正向云发生器的初始采样特征生成；(2)基于云包络带修正的隐变量生成算法. 算法示意图如图2所示.

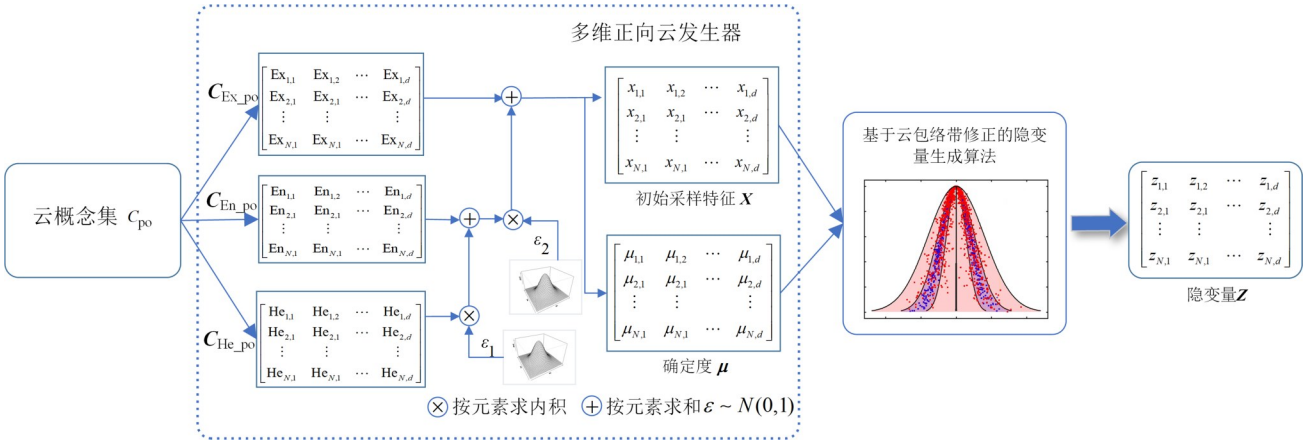


图2 基于多维正向云发生器以及云包络带修正的隐变量采样优化过程示意图

正态云通过期望 Ex，熵 En，超熵 He 构成正向云发生器，这种特定结构放宽了 VGAE 中形成正态分布的前提条件，用一个新的独立参数——超熵，来衡量偏离正态分布的程度，具有更广的普适意义[9]；此外，正向云发生器算法实现了重参数化过程，满足误差反向传播更新梯度的基本条件. 理论证明如下.

设正态云概念 $C_{\text{Pr}}=(\text{Ex},\text{En},\text{He})$，当 He=0 时，正态云模型即退化为正态分布，VGAE 的先验标准正态分布 $N(\mu,\sigma^2)=N(0,1)$ 用正态云概念的形式可表示为 $C_{\text{Pr}}(\mu,\sigma,\sigma_\theta)=C_{\text{Pr}}(0,1,0)$. 由多维正向云发生器[14]，采样特征值 $\forall x_i\in X$ 由两次连续高斯采样生成，即

$$\begin{aligned} x_i&: R_N(\text{En}_i,\text{He}_i)\to y_i, R_N(\text{Ex}_i,|y_i|)\to x_i \\ &\Leftrightarrow N(\sigma_i,\sigma_{\theta,i}^2)\to y_i, N(\mu_i,y_i^2)\to x_i \end{aligned} \tag{6}$$

由式(6)，x_i 的采样过程即运用了重参数化技巧. 依据联合概率分布，x_i 的概率密度可表示为

$$\begin{aligned} &f_X(x_i) \\ &=\int_{-\infty}^{+\infty} p(y_i)p(x_i|y_i)\text{d}y_i \\ &=\int_{-\infty}^{+\infty}\frac{1}{2\pi\text{He}|y_i|}\exp\left\{-\frac{(x_i-\text{Ex})^2}{2y_i^2}-\frac{(y_i-\text{En})^2}{2\text{He}^2}\right\}\text{d}y_i \end{aligned} \tag{7}$$

$f_X(x)$ 即为特征矩阵 $\boldsymbol{X}$ 的概率密度分布函数，计算 x_i 的期望 $\text{E}(x_i)$ 为

$$\begin{aligned} \text{E}(x_i)&=\int_{-\infty}^{+\infty}x_i f_X(x_i)\text{d}x_i \\ &=\int_{-\infty}^{+\infty}p(y_i)\text{d}y_i\int_{-\infty}^{+\infty}x_i p(x_i|y_i)\text{d}x_i \\ &=\mu_i \end{aligned} \tag{8}$$

计算 x_i^2 的期望 $\text{E}(x_i^2)$ 为

$$E\left(x_i^2\right)=\int_{-\infty}^{+\infty}x_i^2 f_X\left(x_i^2\right)\mathrm{d}x_i$$
$$=\mu_i\int_{-\infty}^{+\infty}p\left(y_i\right)\mathrm{d}y_i+\int_{-\infty}^{+\infty}y_i^2 p\left(y_i\right)\mathrm{d}y_i$$
$$=\mu_i+\sigma^2+{\sigma_\theta}^2 \tag{9}$$

x_i的方差$D\left(x_i\right)$则为

$$D\left(x_i\right)=\mathrm{E}(x_i^2)-\mathrm{E}\left(x_i\right)^2=\sigma^2+{\sigma_\theta}^2 \tag{10}$$

由式(10),当$\sigma_\theta=0$,即超熵$\mathrm{He}=0$时,特征x_i的方差$D\left(x_i\right)=\sigma^2$,云滴群退化为正态分布. 由式(2),在编码器处理节点的嵌入过程中,实际$\mathrm{He}=\mathrm{e}^{\ln \mathrm{He}}>0$,即$D\left(x_i\right)>\sigma^2$,$x_i$的采样空间始终呈现出泛正态的形式. 由此,正态云模型的数字特征超熵He可以用来反映影响因素中的不均匀或非相互独立的情况,是偏离正态分布的度量,这种泛正态更接近复杂网络中的实际情形[1].

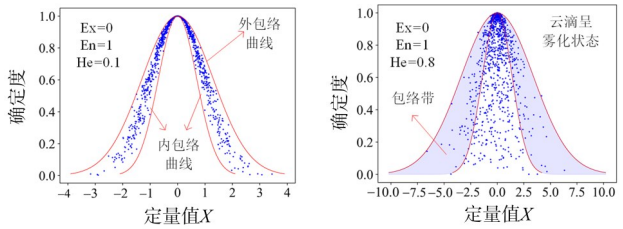
此外,针对变分自编码器在优化拟合过程中常见的后验塌陷问题[10],即:在隐变量采样过程中,模型容易忽略数据中潜在的特征信息、后验分布失效,在VGAE训练中通常表现为采样值过度向均值中心聚集,模型退化为自编码器. 对此,本文提出一种"基于云包络带修正的隐变量采样优化算法",利用云包络带约束初始采样特征$\boldsymbol{X}_o^{N\times d}$,以此缩小云心密度[17]、减轻采样过度向期望中心聚集现象,从而缓解模型训练过程中的"后验塌陷"问题. 云模型本质是一个边界模糊的泛正态分布,用云包络带来表示云滴的分布区域更能够体现整体特性[16]. 实现如下.

设正态云概念C_{Pr}的内包络曲线$\mu_{\mathrm{in}}(x)$、外包络曲线$\mu_{\mathrm{ex}}(x)$分别表示为[16]:

$$\mu_{\mathrm{in}}(x)=\exp\left(-\frac{(x-\mathrm{Ex})^2}{2(\mathrm{En}-3\mathrm{He})^2}\right) \tag{11}$$

$$\mu_{\mathrm{ex}}(x)=\exp\left(-\frac{(x-\mathrm{Ex})^2}{2(\mathrm{En}+3\mathrm{He})^2}\right) \tag{12}$$

分析云滴的分布情形,由"3He"规则[17],当He<En/3时,有99.7%的云滴落在云包络带内,内外包络曲线均存在,如图3(*a*)所示;当He>En/3时,云滴分布呈现"雾化状态",部分云滴脱离包络带内,轮廓不再清晰,如图3(*b*)所示.



(*a*) He<En/3　　(*b*) He>En/3

图3　正态云在不同情形下的包络曲线示意图

由此,本文基于云包络带修正的采样优化算法主要针对后验正态云概念集C_{po}中呈现"雾化状态"的情形,即当He>En/3时,云包络带存在且部分云滴脱离包络带区间. 具体实现如下:

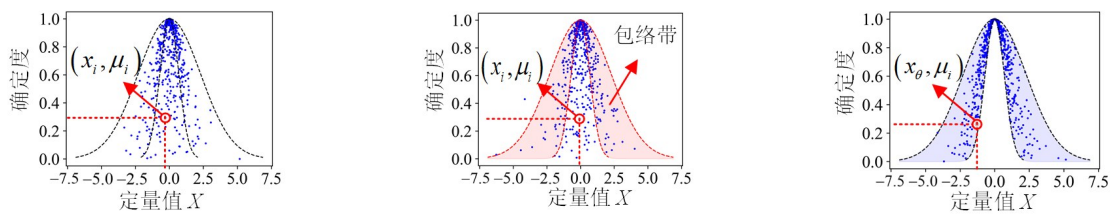
对采样特征$\forall\left(x_i,\mu_i\right)\in\left(\boldsymbol{X}_o,\boldsymbol{\mu}\right)$,$x_i$即为子云概念$c_i\left(\mathrm{Ex}_i,\mathrm{En}_i,\mathrm{He}_i\right)\in C_{\mathrm{po}}$的一次量化随机实现,由确定度$\mu_i$进行判别,将$x_i$约束至$c_i$的云包络带内,并生成修正后的隐变量特征值,形式如下:

$$\Omega_i=\mathrm{En}_{\mathrm{po},i}-3\mathrm{He}_{\mathrm{po},i} \tag{13}$$

$$x_\theta=\begin{cases}\mathrm{Ex}_{\mathrm{po},i}-\left|\Omega_i\right|\sqrt{-2\ln\mu_{\mathrm{in}}\left(x_i\right)}, & x_i<\mathrm{Ex}_{\mathrm{po},i}\\ \mathrm{Ex}_{\mathrm{po},i}+\left|\Omega_i\right|\sqrt{-2\ln\mu_{\mathrm{in}}\left(x_i\right)}, & x_i>\mathrm{Ex}_{\mathrm{po},i}\end{cases} \tag{14}$$

$$|z_i|=\begin{cases}x_i, & \mu_i>\mu_{\mathrm{in}}\left(x_i\right)\\ x_\theta, & \mu_i<\mu_{\mathrm{in}}\left(x_i\right)\end{cases} \tag{15}$$

式(14)中,x_θ表示修正后的采样值. z_i即作为隐变量$\boldsymbol{Z}$的第i个特征值. 修正示意图如图4所示.



(*a*) 初始采样特征x_i　　(*b*) 判别包络带　　(*c*) 修正后特征x_θ

图4　云包络带修正采样示意图

基于多维正向云发生器以及云包络带修正的隐变量采样优化算法完整流程如算法1所示.

3.3　解码器及损失函数

MCM-VGAE的解码器结构与VGAE相同,通过计算隐变量$\boldsymbol{Z}$的内积重构图. 形式如下:

$$p(\boldsymbol{A}|\boldsymbol{Z})=\prod_{i=1}^{N}\prod_{j=1}^{N}p(A_{ij}|\boldsymbol{z}_i,\boldsymbol{z}_j) \tag{16}$$

$$p(A_{ij}=1|\boldsymbol{z}_i,\boldsymbol{z}_j)=\sigma(\boldsymbol{z}_i^{\mathrm{T}},\boldsymbol{z}_j) \tag{17}$$

算法1 基于多维正向云发生器以及云包络带修正的隐变量采样算法流程

输入:后验云概念集 $C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex\_po}},\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}})\in\mathbb{R}^m$ 数字特征.

输出:隐变量特征矩阵 $\boldsymbol{Z}(z_1,z_2,\cdots,z_m)$.

(1) $\boldsymbol{Y}\leftarrow R_N\left(\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}}\right)$, $\boldsymbol{Y}\in\mathbb{R}^m$//生成 m 维正态随机数

(2) $\boldsymbol{X}_o\leftarrow R_N\left(\boldsymbol{C}_{\mathrm{Ex\_po}},|\boldsymbol{Y}|\right)$, $\boldsymbol{X}_o\in\mathbb{R}^m$/*生成 m 维正态随机数 $\boldsymbol{X}_o$, $\boldsymbol{X}_o$ 即为初始采样特征矩阵*/

(3) 对 $\forall x_i\in\boldsymbol{X}_o$, $\mu(x_i)\leftarrow\exp\left(-\frac{(x_i-\mathrm{Ex}_i)^2}{2y_i^2}\right)$, $y_i\in\boldsymbol{Y}$/*计算特征值 x_i 的确定度*/

(4) IF $\mathrm{He}_i>\mathrm{En}_i/3$ THEN //x_i 采样空间呈雾化状态

(5) $\mu_{\mathrm{in}}(x_i)\leftarrow\exp\left(-\frac{(x-\mathrm{Ex}_i)^2}{2(\mathrm{En}_i-3\mathrm{He}_i)^2}\right)$/*计算 x_i 在内包络曲线上相应的确定度*/

(6) IF $\mu(x_i)<\mu_{\mathrm{in}}(x_i)$ THEN

(7) IF $x_i<\mathrm{Ex}_i$ THEN // 初始采样值位于期望左侧

$$x_\theta\leftarrow\mathrm{Ex}_i-|\mathrm{En}_i-3\mathrm{He}_i|\sqrt{-2\ln\mu(x_i)};$$

(8) ELSE // 初始采样值位于期望右侧

$$x_\theta\leftarrow\mathrm{Ex}_i+|\mathrm{En}_i-3\mathrm{He}_i|\sqrt{-2\ln\mu(x_i)};$$

(9) END IF

(10) $x_i\leftarrow x_\theta$;//将修正后的采样值 x_θ 赋予 x_i

(11) END IF

(12) END IF

(13) $\boldsymbol{Z}\leftarrow\boldsymbol{X}(x_1,x_2,\cdots x_m)$/*将量化值作为隐变量 $\boldsymbol{Z}$*/

MCM-VGAE的损失函数由两部分构成:(1)重构图与输入图之间的交叉熵损失;(2)正态云概念的漂移性损失度量. 形式如下:

$$\mathcal{L}=\mathcal{L}_{\mathrm{d}D}+\beta\mathcal{L}_p,\beta\in[0,1] \tag{18}$$

式(18)中,第一项 $\mathcal{L}_{\mathrm{d}D}$ 表示重构图与原始图之间的交叉熵函数,对 $\forall y\in\boldsymbol{A}$, $\forall\hat{y}\in\hat{\boldsymbol{A}}$ 有:

$$\mathcal{L}_{\mathrm{d}D}=-\frac{1}{N}\sum y\log\hat{y}+(1-y)\log(1-\hat{y}) \tag{19}$$

式(18)的第二项 $\mathcal{L}_p$ 为正态云概念的漂移性度量函数. 超参数 β 用于调节重构项与分布项之间的重要性. 针对 $\mathcal{L}_p$ 损失函数的设计,正态云概念的漂移性度量是云模型理论研究中的重要内容之一. 目前,围绕云概念的漂移性计算方法有值系数(欧氏距离、海明距离等)和曲线相似系数(向量间的夹角余弦等)两种度量形式[11]. 同比而言,基于曲线的漂移性度量方法由云模型的3个数字特征计算而成,具备普适性,且时间复杂度较低,更适合作为深度学习模型的损失函数. 因此,本文主要借助KL散度并结合云外包络曲线实现云概念的漂移性度量(云内包络曲线可能出现消失而外包络曲线始终存在),具体如下:

设后验云概念集为 $C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex}\_po},\boldsymbol{C}_{\mathrm{En}\_po},\boldsymbol{C}_{\mathrm{He}\_po})\in\mathbb{R}^m$, 对 $\forall c_i(\mathrm{Ex}_i,\mathrm{En}_i,\mathrm{He}_i)\in C_{\mathrm{po}}$, 其云外包络曲线表示为 $\mu_i^{ex}(x)$;设先验正态云概念 $C_{\mathrm{pr}}\left(\mathrm{Ex}_{\mathrm{pr}},\mathrm{En}_{\mathrm{pr}},\mathrm{He}_{\mathrm{pr}}\right)$ 的外包络曲线为 $\mu_{\mathrm{pr}}^{\mathrm{ex}}(x)$. $\mu_i^{\mathrm{ex}}(x)$ 与 $\mu_{\mathrm{pr}}^{\mathrm{ex}}(x)$ 对应的概率密度函数分别为 $p_i(x),q(x)$,即

$$p_i(x)=\frac{1}{\sqrt{2\pi}\left(\mathrm{En}_i+3\mathrm{He}_i\right)}\mu_i^{\mathrm{ex}}(x) \tag{20}$$

$$q(x)=\frac{1}{\sqrt{2\pi}\left(\mathrm{En}_{\mathrm{pr}}+3\mathrm{He}_{\mathrm{pr}}\right)}\mu_{\mathrm{pr}}(x) \tag{21}$$

C_{po} 与 C_{pr} 之间基于非对称KL散度[10]并结合云外包络曲线的漂移度定义为

$$\begin{aligned}&D_{\mathrm{KL}}(C_{\mathrm{po}}\|C_{\mathrm{pr}})\\&=\sum_{i=1}^m\int p_i(x)\log\frac{p_i(x)}{q(x)}\mathrm{d}x\\&=\sum_{i=1}^m\left[\log\frac{\sigma_{\mathrm{pr}}}{\sigma_{\mathrm{po},i}}+\frac{{\sigma_{\mathrm{po},i}}^2+\left(\mathrm{Ex}_i-\mathrm{Ex}_{\mathrm{pr}}\right)^2}{2{\sigma_{\mathrm{pr}}}^2}-\frac{1}{2}\right]\end{aligned} \tag{22}$$

$$\sigma_{\mathrm{pr}}=\mathrm{En}_{\mathrm{pr}}+3\mathrm{He}_{\mathrm{pr}},\ \sigma_{\mathrm{po},i}=\mathrm{En}_i+3\mathrm{He}_i \tag{23}$$

此外,基于对称KL散度[11]的正态云概念漂移性度量方法定义如下:

$$\begin{aligned}&D_J\left(C_{\mathrm{po}}\|C_{\mathrm{pr}}\right)=D_{\mathrm{KL}}\left(C_{\mathrm{po}}\|C_{\mathrm{pr}}\right)+D_{\mathrm{KL}}\left(C_{\mathrm{pr}}\|C_{\mathrm{po}}\right)=\\&\frac{1}{2}\sum_{i=1}^m\left[\left(\mathrm{Ex}_{\mathrm{pr}}-\mathrm{Ex}_i\right)^2+\left({\sigma_{\mathrm{pr}}}^2+{\sigma_{\mathrm{po},i}}^2\right)\right]\left(\frac{1}{{\sigma_{\mathrm{pr}}}^2}+\frac{1}{{\sigma_{\mathrm{po},i}}^2}\right)-4\end{aligned} \tag{24}$$

本文将 D_{KL} 作为MCM-VGAE的默认 $\mathcal{L}_p$ 损失函数,D_J 则作为一种变体形式(MCM-VGAE$_{\mathrm{dj}}$),关于 D_{KL} 与 D_J 选取的实验结果分析见本文第4节.

3.4 MCM-VGAE训练算法及复杂度分析

综合3.1至3.3节,MCM-VGAE模型的整体训练流程如算法2所示.

分析算法2的时间复杂度,借鉴文献[18]的分析方

算法2 MCM-VGAE模型训练算法流程

输入:图 $G=\{V,E,\boldsymbol{A},\boldsymbol{X},\boldsymbol{D}\}$,嵌入维度 d,先验正态云概念超参数 $C_{\mathrm{pr}}\left(\mathrm{Ex}_{\mathrm{pr}},\mathrm{En}_{\mathrm{pr}},\mathrm{He}_{\mathrm{pr}}\right)$,APPNP卷积网络的迭代次数超参数 K,损失函数超参数 β,训练轮次 T.

输出:隐变量特征矩阵 $\boldsymbol{Z}$,待学习的模型参数 $\theta=\left\{\{W^{(i)}\}^3{}_{i=1},\theta_{\mathrm{APPNP}}\right\}$.

(1) $x_{\mathrm{train}},x_{\mathrm{val}},x_{\mathrm{test}}\leftarrow$从输入图 G 中划分数据集

(2) 初始化待学习模型参数 θ

(3) WHILE epochs $<T$ DO

(4) $C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex\_po}},\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}})\leftarrow\mathrm{Encoder}(x_{\mathrm{train}})$

(5) $\boldsymbol{Z}^{N\times d}\leftarrow$ Sampling from C_{po}

(6) $\hat{\boldsymbol{A}}\leftarrow\mathrm{Decoder}(\boldsymbol{Z})$

(7) $\theta\leftarrow\nabla\mathcal{L}_D(\theta);\theta\leftarrow\nabla\mathcal{L}_p(\theta)$ // 更新待学习参数

(8) END WHILE

法,编码器中Linear-GCN计算的时间复杂度为$O\left(\|\boldsymbol{A}\|_0 f+Nf^2\right)$,其中$\|\boldsymbol{A}\|_0$表示邻接矩阵$\boldsymbol{A}^{N\times N}$中非零元素的数量,$f$表示特征矩阵$\boldsymbol{X}^{N\times f}$中的特征维度;APPNP计算的时间复杂度近似为$O(Nd)$;采样过程的时间复杂度为$O(Nd)$;解码器计算隐变量内积的时间复杂度为$O\left(N^2 d\right)$. 由于嵌入维度$d$的取值通常为较小的常数,因此,可以认为MCM-VGAE在训练过程中的总时间复杂度为:

$$O\left(T\times\left(\|\boldsymbol{A}\|_0 f+Nf^2+N^2+N\right)\right) \tag{25}$$

4　实验验证及分析

为验证MCM-VGAE的有效性,本节首先通过采样云心密度对比实验[17]针对"基于云包络带修正的隐变量采样优化算法(算法1)"进行了有效性验证;然后通过三项图嵌入应用实验(链路预测、节点聚类、图嵌入可视化实验)在多类型数据集上对比各个图嵌入学习方法;最后通过模型超参数分析实验验证了MCM-VGAE相关超参数的敏感性.

4.1　实验数据集

在数据集的选取方面,由于本文的应用场景主要为静态的拓扑网络,为与VGAE等基准模型形成对照实验,实验采用了3组常用的Benchmark引文网络数据集(Cora、Citeseer、Pubmed)[19]. 其次,为进一步验证MCM-VGAE的图嵌入表征能力以及泛化能力,本文增加了3组包含较大规模数量节点或边的开源图数据集:亚马逊(Amazon)电商网络数据集(Computers、Photo)[20],NELL知识图谱数据集[19]. 所有数据集均为无向图,详细如表1所示.

表1　数据集信息统计

数据集	节点	边	特征	类别	存在孤立节点
Cora	2 708	10 556	1 433	7	否
Citesser	3 327	9 104	3 703	6	是
Pubmed	19 717	88 648	500	3	否
Computers	13 752	491 722	767	10	是
Photo	7 650	238 162	745	8	是
NELL	65 755	251 550	61 278	186	否

实验时采取数据集随机划分策略,其中验证集划分比例为5%用于超参数优化,测试集为10%用于验证模型性能,其余则用于模型训练.

4.2　对比方法

本文分别对比实验了8种基于编码器-译码器框架的GNN模型,以及本文提出的2种方法,方法描述如表2所示.

表2　对比方法及描述

方法类型	方法名	描述
基于图自编码器	GAE[4]	一种基于自编码器的图嵌入学习框架
	Linear-GAE[8]	线性图自编码器
	GNAE[7]	图归一化自编码器
基于vMF分布的变分图自编码器	S-VGAE[8]	基于超球面分布的变分图自编码器
基于先验高斯分布的变分图自编码器	VGAE[4]	变分图自编码器
	ARVGE[5]	一种基于对抗正则化思想的变分图自编码器
	Linear-VGAE[8]	线性变分图自编码器
	VGNAE[7]	变分图归一化自编码器
基于先验正态云概念的变分图自编码器(本文方法)	MCM-VGAE	基于多维云模型的变分图自编码器
	MCM-VGAE$_{dj}$	MCM-VGAE的变体,使用对称KL散度实现云概念的漂移性度量

4.3　模型训练相关超参数设置

在超参数设置上,本文设置MCM-VGAE、MCM-VGAE$_{dj}$的图嵌入维度$d=64$;先验正态云概念均设置$\mathrm{Ex}_{pr}=0$,$\mathrm{En}_{pr}=1$,$\mathrm{He}_{pr}=1$;归一化图卷积GNCN的缩放因子超参数s、APPNP图卷积的迭代次数参数K、传播概率α均与原论文中的描述保持一致(即$s=1.8$,$K=10$,$\alpha=0.1$). 模型训练过程中使用Adam优化器且设置学习率$\mathrm{lr}=0.01$;固定训练轮次$T=300$(根据模型收敛实际情况适当调整). 其余基准模型的超参数设置均与原论文保持一致. 实验环境基于PyTorch Geometric图神经网络框架[21].

4.4　MCM-VGAE采样云心密度对比实验

此处首先给出正态云的云心窗口、云滴论域、云心密度的基本定义.

定义5　云心窗口[17]:正态云滴定量值最集中的论域$[\mathrm{Ex}-\mathrm{En},\mathrm{Ex}+\mathrm{En}]$,称作大小为2En的云心窗口.

定义6　云滴论域[17]:包含所有云滴定量值的论域

范围，计算式为 $\mathrm{Max}\{x_1,x_2,\cdots,x_n\}-\mathrm{Min}\{x_1,x_2,\cdots,x_n\}$.

定义 7 云心密度[17]：云心窗口内的总云滴数/2En. 关于云心窗口、云滴论域的示意图如图 5 所示.

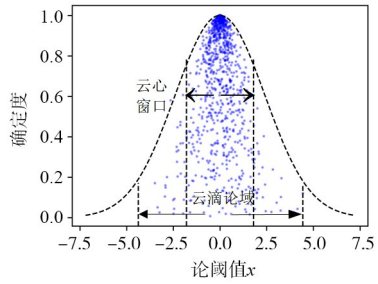
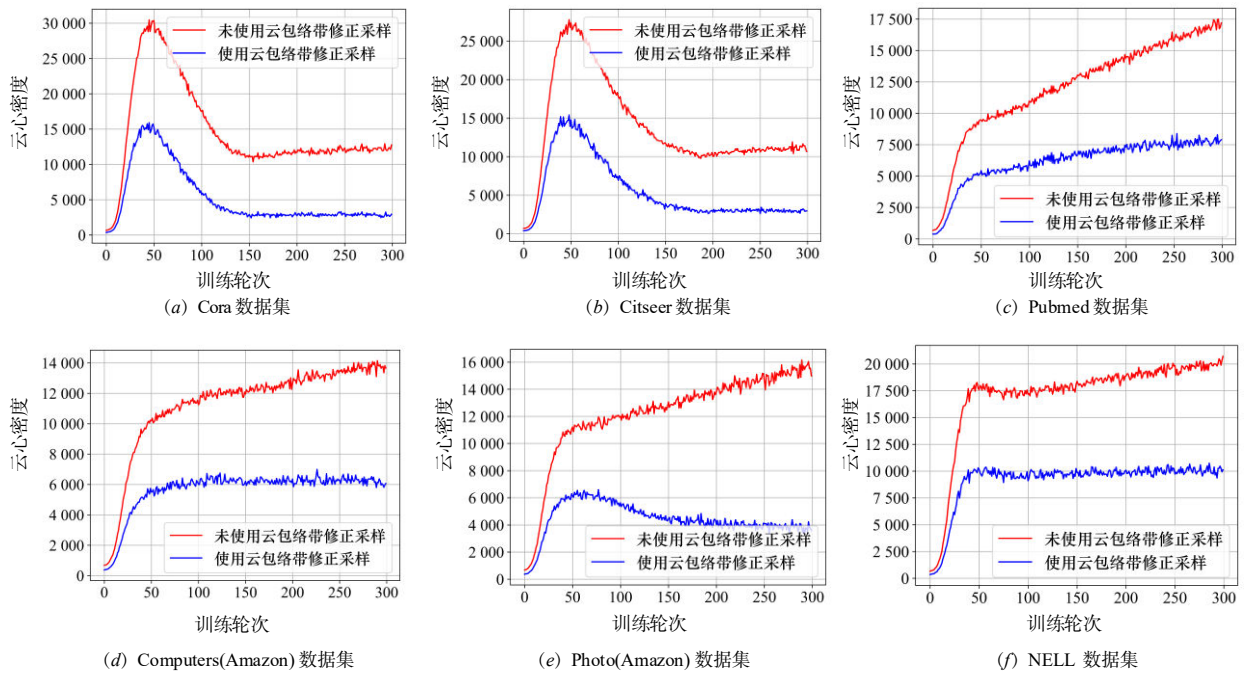


图 5 云滴分布示意图

为验证算法 1 对于变分图自编码器中存在的“后验塌陷”问题的有效性，本文借助云心密度来反映云概念空间（即模型内部隐空间）的整体采样情况. 云心密度即表示采样云滴在期望或临近期望的密度，云心密度过高则采样特征过度向期望中心聚集，KL 散度接近于零，模型退化为自编码器. 对此，实验基于 6 组数据集，针对 MCM-VGAE 在训练过程中的云心密度变化情况进行对比分析，实验步骤如下：

（1）分别训练 MCM-VGAE 以及未使用云包络带修正采样的变体模型，在训练中分别输出后验云概念集 $C_{\mathrm{po}}(\boldsymbol{C}_{\mathrm{Ex\_po}},\boldsymbol{C}_{\mathrm{En\_po}},\boldsymbol{C}_{\mathrm{He\_po}})$ 数字特征的平均值 $\widetilde{\mathrm{Ex}},\widetilde{\mathrm{En}},\widetilde{\mathrm{He}}$；（2）基于该 3 个数字特征，利用正向云发生器算法[11]生成云滴群，此处设置云滴数 $N=2\,000$；（3）分别计算 MCM-VGAE 及变体模型中采样结果的云心密度. 实验结果如图 6 所示.

实验结果表明：当模型训练结束时，使用了云包络带修正采样的 MCM-VGAE 与未使用方法相比，在 6 组数据集上的平均采样密度能够缩减 50% 至 85%，由此验证了算法 1 能够有效减轻采样过程中隐变量过度向期望中心聚集、模型退化为自编码器的现象，从而缓解“后验塌陷”问题，且具备优秀的泛化能力. 此外，在包含较大规模数量节点或边的图数据集（Computers、Photo、NELL）上，该算法表现出能够缓解模型训练过程中的“过拟合现象”，加快模型训练的收敛速度（图卷积神经网络自身具有较强的学习能力，云包络带修正可作为一种正则化手段）.



（a）Cora 数据集 （b）Citseer 数据集 （c）Pubmed 数据集

（d）Computers(Amazon) 数据集 （e）Photo(Amazon) 数据集 （f）NELL 数据集

图 6 模型训练过程中云包络带修正采样算法对云心密度的影响

4.5 链路预测实验

在图神经网络领域，链路预测是一项典型的图嵌入学习任务，其目标主要用于发现网络中缺失的链接以及未来可能出现的链接[3]. 此处实验与 VGAE 等模型的评估流程相同，将链路预测转换为二分类问题，通过计算节点嵌入向量间的相似性来预测链接是否存在，并使用衡量模型分类性能的 AUC(area under ROC)[4]、平均精准度 AP(average precision)[4]作为性能评价指标.

针对表 2 中的各图嵌入方法，本节分别在 6 组实验数据集上完成了链路预测工作. 实验过程中，由于采取随机划分数据集策略导致结果出现波动，本文在每组数组集上分别将各方法实验 10 次，并以评价指标得

分的平均值(%)、标准误差作为链路预测的实验结果，　详细结果如表3、表4所示.

表3　引文网络数据集链路预测实验结果汇总　　单位:%

Method	Cora		Citeseer		Pubmed	
	AUC	AP	AUC	AP	AUC	AP
GAE	92.2±0.76	93.0±0.71	92.5±0.89	93.4±0.82	96.4±0.13	96.5±0.17
VGAE	92.4±0.91	93.9±0.86	91.0±1.05	91.5±1.14	96.2±0.34	96.0±0.35
ARVGE	93.4±0.82	94.6±0.72	93.5±1.02	94.4±0.73	84.5±9.55	83.2±9.02
Linear-GAE	93.7±0.67	94.6±0.76	93.5±0.58	93.3±0.62	96.8±0.16	96.0±0.23
Linear-VGAE	91.3±0.88	91.5±1.06	92.0±0.93	91.5±1.10	95.7±0.19	95.3±0.23
S-VGAE	94.1±1.01	94.1±1.03	94.7±0.82	95.2±1.00	96.0±1.00	96.0±1.00
GNAE	94.4±0.66	95.1±0.51	94.5±0.62	95.5±0.49	97.0±0.16	96.8±0.19
VGNAE	95.2±0.64	95.4±0.58	95.6±0.51	95.8±0.60	97.1±0.34	97.0±0.35
MCM-VGAE	**96.8±0.63**	**97.1±0.57**	96.7±0.67	97.5±0.51	**97.9±0.15**	**97.7±0.17**
MCM-VGAE$_{dj}$	96.4±0.065	97.0±0.61	**96.8±0.67**	**97.5±0.50**	97.6±0.16	97.3±0.21

表4　Amazon电商网络、NELL数据集链路预测实验结果汇总　　单位:%

Method	Computers(Amazon)		Photo(Amazon)		NELL	
	AUC	AP	AUC	AP	AUC	AP
GAE	93.7±0.21	93.7±0.18	92.8±2.21	92.3±2.15	94.5±0.35	93.8±0.42
VGAE	90.1±0.47	91.0±0.47	92.0±1.15	91.5±1.08	95.7±0.09	95.4±0.10
ARVGE	77.6±13.96	77.7±14.1	78.7±8.00	78.8±8.58	72.1±7.80	74.2±8.37
Linear-GAE	88.4±0.10	88.6±0.14	91.2±0.20	90.9±0.27	93.6±0.06	94.8±0.08
Linear-VGAE	87.8±0.16	88.1±0.21	89.6±0.17	88.6±0.19	94.7±0.12	93.9±0.42
S-VGAE	88.8±0.61	89.1±0.53	93.7±0.82	94.2±0.78	94.0±1.00	94.6±1.10
GNAE	**96.0±0.29**	**95.9±0.31**	96.0±0.23	95.5±0.26	96.0±0.10	95.6±0.15
VGNAE	94.2±0.10	94.6±0.11	96.2±0.22	95.5±0.25	96.2±0.08	95.5±0.13
MCM-VGAE	**95.7±0.08**	**95.3±0.11**	**97.2±0.21**	**96.6±0.29**	**97.1±0.08**	**96.6±0.13**
MCM-VGAE$_{dj}$	95.5±0.07	95.0±0.09	96.6±0.13	96.0±0.16	96.8±0.09	96.4±0.10

实验结果表明:由表3,AUC、AP两个评价指标的最高得分均来自于MCM-VGAE或其变体. 在较大规模的图数据集上(表4),虽然MCM-VGAE在Computers数据集上的得分(AUC=95.7%,AP=95.3%)略低于GNAE(AUC=96.0%,AP=95.9%),但MCM-VGAE得分的标准差更低(MCM-VGAE的标准差约为0.1,GNAE约为0.3),在实验中表现出了相对更好的稳定性.

此外,ARVGE在小规模图数据集(Cora、Citeseer)上取得了较高的得分(AUC>93%,AP>94%),然而在包含较大规模数量节点或边的数据集上(Pubmed、Computers、Photo、NELL数据集),AUC、AP得分显著下降(指标得分AUC、AP均小于90%);此外,基于超球面先验分布的S-VGAE虽然在引文网络数据集上相较于使用正态分布的VGAE、ARVGE、Linear-VGAE模型表现出了更好的性能,但在Computers、NELL数据集上S-VGAE的性能却低于VGAE等,由此说明了vMF分布与先验正态分布相比,并未充分体现普适性. 相比而言,基于先验正态云概念的MCM-VGAE在6组数据集上的性能整体表现出色,具备更健壮的鲁棒性.

由此,本文通过链路预测实验验证了MCM-VGAE具备良好的泛化能力,进一步体现了云模型在复杂网络研究中的普遍适用性.

4.6　节点聚类及图嵌入可视化实验

节点聚类实验同样是图嵌入工作中的一项重要应用[2]. 图嵌入可视化实验则通常是将二维空间中的嵌入向量进行可视化展示,帮助人们更加形象地理解图嵌入的过程以及结果.

本文分别基于Cora、Computers数据集完成了节点聚类及图嵌入可视化实验. 实验步骤为:(1)训练表2中的各模型,超参数设置与4.3节保持一致;(2)利用训练完成的编码器输出节点的低维向量表示$\boldsymbol{E}_d \in \mathbb{R}^m$. 其中,MCM-VGAE取后验正态云概念集中的期望$\boldsymbol{C}_{\text{Ex_po}} \in \mathbb{R}^m$作为$\boldsymbol{E}_d$;GAE、VGAE等其余模型取均值$\boldsymbol{\mu} \in \mathbb{R}^m$作为$\boldsymbol{E}_d$;(3)利用K-means聚类算法[5]分别对各$\boldsymbol{E}_d$进行聚类,超参数$K$则与相应数据集的类别数量一致(Cora数据集包含7个类别标签,Computers含有10个类别标签);(4)借助t-SNE降维算法[22]将聚类结果映射到二维空间中进行可视化展示,节点的标签分别用不

同颜色标记.

针对节点聚类实验的评价指标选取,本文以聚类任务中常用的准确率ACC(Accuracy)[2]、标准化互信息指数NMI(Normalized Mutual Information)[5]、调整兰德系数ARI(Adjusted Rand Index)[5]、宏平均F1(Macro-F1)[5]四个评价指标得分作为节点聚类的实验结果.图嵌入可视化则主要是对嵌入结果进行展示.节点聚类结果如表5、表6所示,图嵌入可视化结果如图7、图8所示(此处基于两组数据集分别筛选了5个更具明显特征的模型).

表5 节点聚类实验结果(Cora数据集) 单位:%

Method	ACC	NMI	ARI	F1
GAE	0.547	0.402	0.269	0.561
VGAE	0.599	0.427	0.334	0.600
ARVGE	0.477	0.313	0.173	0.463
Linear-GAE	0.634	0.468	0.384	0.641
Linear-VGAE	0.670	0.489	0.441	0.648
S-VGAE	0.669	0.469	0.463	0.663
GNAE	0.697	0.502	0.455	0.682
VGNAE	0.699	0.501	0.471	0.690
MCM-VGAE	**0.723**	**0.544**	**0.514**	**0.721**
MCM-VGAE$_{dj}$	0.720	0.522	0.505	0.713

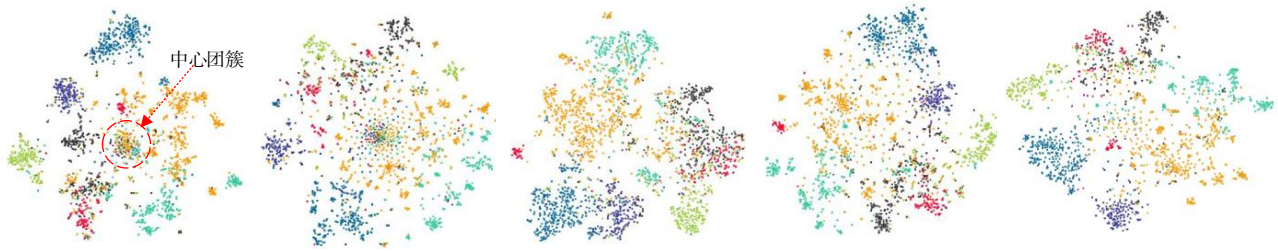
由表5、表6,MCM-VGAE在两组数据集Cora、Computers上的聚类实验中均表现出了最优的性能,Accuracy、NMI、兰德系数ARI、Macro-F1四个指标均取得了显著提升(与VGAE相比,Cora数据集上的ACC、NMI提升约0.12,ARI提升约0.18,F1提升约0.12).

表6 节点聚类实验结果(Computers数据集) 单位:%

Method	ACC	NMI	ARI	F1
GAE	0.399	0.406	0.228	0.338
VGAE	0.351	0.331	0.222	0.261
ARVGE	0.348	0.326	0.219	0.257
Linear-GAE	0.380	0.325	0.199	0.247
Linear-VGAE	0.372	0.314	0.190	0.241
S-VGAE	0.360	0.387	0.265	0.363
GNAE	0.496	0.482	0.311	0.405
VGNAE	0.501	0.485	0.321	0.405
MCM-VGAE	**0.533**	**0.514**	**0.360**	**0.433**
MCM-VGAE$_{dj}$	0.506	0.503	0.341	0.418

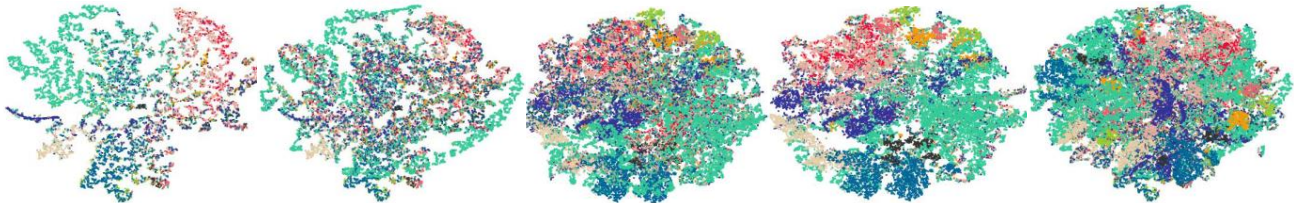
在图嵌入可视化结果中,由图7(*e*),MCM-VGAE相同类别(同种颜色)的节点分段较其他模型更为明显,且每个类的边界更加清晰,“中心团簇现象”几乎消失;观察图7(*a*)的GAE、图7(*b*)ARVGE的模型实验结果,“中心团簇现象”十分明显,且不同类别节点的重叠程度更高;图7(*e*)的Linear-VGAE、图7(*d*)VGNA模型的“中心团簇现象”同样几乎消失,但与MCM-VGAE相比,其周围节点的交错现象更显著.由图8,由于Computers数据集的节点数量远大于Cora,导致嵌入向量的可视化分布情况较为复杂,但仍可明显观测到图8(*e*)的MCM-VGAE的图嵌入结果中不同类别的轮廓更加清晰,杂糅程度明显优于基准模型GAE、ARVGE、VGAE.

综上,节点聚类、图嵌入可视化实验表明:MCM-VGAE较其他基准模型能够获得更有意义的嵌入特征分布,且性能更佳.



(*a*) GAE (*b*) ARVGE (*c*) Linear-VGAE (*d*) VGNA (*e*) MCM-VGAE

图7 图嵌入可视化结果(Cora数据集)



(*a*) GAE (*b*) ARVGE (*c*) VGAE (*d*) VGNA (*e*) MCM-VGAE

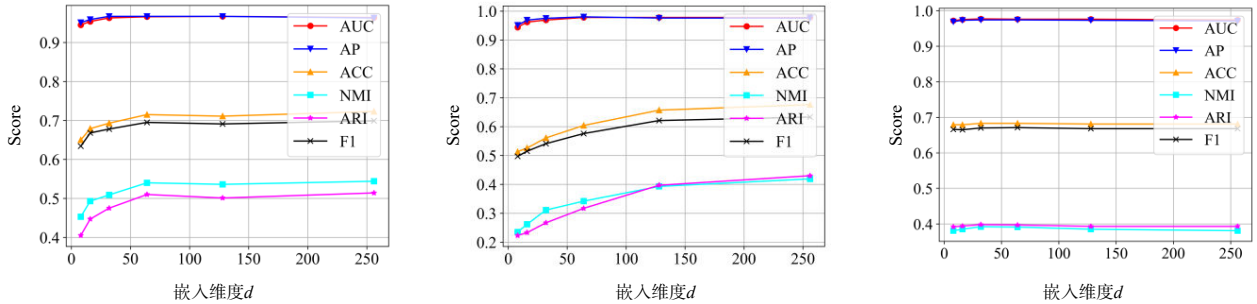
图8 图嵌入可视化结果(Computers数据集)

4.7　MCM-VGAE超参数敏感性分析

本节基于Cora、Citeseer、Pubmed引文网络数据集分别评估了MCM-VGAE的超参数节点嵌入维度d、先验正态云概念的数字特征Ex_{pr}、En_{pr}、He_{pr}对于链路预测、节点聚类实验的敏感性,实验过程如下:

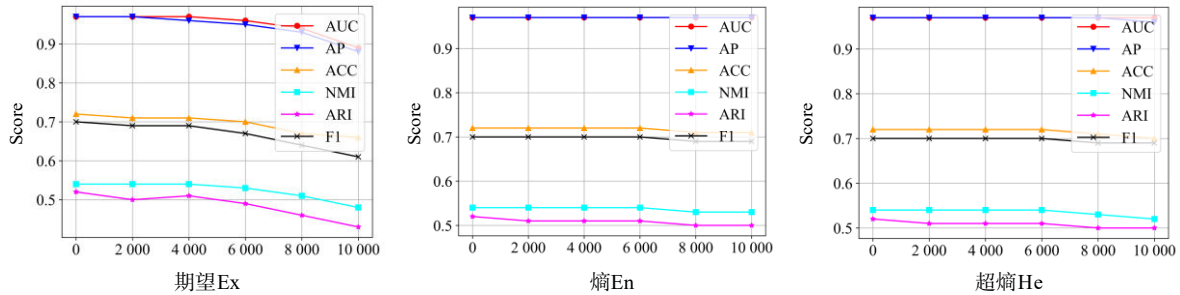
针对图嵌入维度d,本文基于3组引文网络数据集,依次扩大MCM-VGAE编码器的嵌入维度,即设置$GNCN_{Ex}$、GNN_{En}、GNN_{He}的输出维度依次为:$d=8,16,32,\cdots,256$,其它超参数设置以及训练过程与链路预测、节点聚类实验保持一致,结果如图9所示.

针对超参数Ex_{pr}、En_{pr}、He_{pr},本文基于Cora、Citeseer两组数据集,设置各数字特征初始值为$Ex_{pr}=0$,$En_{pr}=1$,$He_{pr}=1$,然后依次调整Ex_{pr}、En_{pr}、He_{pr}的取值且有$Ex_{pr},En_{pr},He_{pr}\in[0,10\,000]$,实验结果如图10所示.

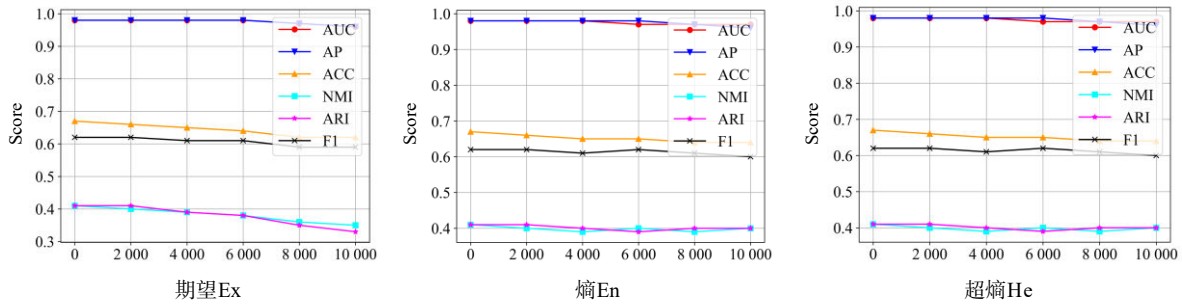


(a) Cora数据集　(b) Citeseer数据集　(c) Pubmed数据集

图9　不同嵌入维度d对于链路预测、节点聚类实验结果的影响



(a) Cora数据集



(b) Citeseer数据集

图10　不同先验云概念数字特征对于实验结果的影响

由图9,基于Cora、Citeseer数据集,当MCM-VGAE的嵌入维度由$d=8$扩大到$d=64$时,模型在链路预测(AUC、AP指标)、节点聚类(ACC、NMI、ARI、F1指标)任务上的性能表现稳步上升,且当嵌入维度$d=64$时,各评价指标增长趋于饱和,而节点聚类实验的4个评测指标仍有增长趋势;对于Pubmed数据集,MCM-VGAE的实验表现较为稳定,受嵌入维度d的影响较小,指标始终维持在较高水平(链路预测AUC、AP得分大于0.97;节点聚类实验的准确率、Macro-F1得分约为0.7,NMI、ARI指数约为0.4).

由图10,针对超参数Ex_{pr},当固定$En_{pr}=1$,$He_{pr}=1$,设置$\mu_{ex}(x)$时,MCM-VGAE在Cora、Citseer数据集上的

链路预测、节点聚类任务中均表现了良好且稳定的性能(AUC、AP得分约为0.97,其他指标均获得较高得分),当$Ex_{pr}\geqslant 4\,000$时,MCM-VGAE性能开始下降;而对于超参数En_{pr}、He_{pr},模型在两组数据集上均表现出良好的鲁棒性.

综上,MCM-VGAE在实际部署应用中,针对$C_{pr}\left(Ex_{pr},En_{pr},He_{pr}\right)$中3个数字特征超参数的设置,只需固定该取值为常规值(本文通常设置$Ex_{pr}=0$,$En_{pr}=1$,$He_{pr}=1$),而不必过多地考虑这3个数字特征的调参影响,由此验证了MCM-VGAE在实际应用部署中的可行性.

5 结论与展望

本文针对当前VGAE中存在的先验正态分布缺陷问题、孤立节点嵌入问题、模型训练过程中存在的后验塌陷问题,提出了一种基于多维云模型的变分图自编码器框架(MCM-VGAE). 经理论分析与多角度实验验证,该方法能够有效解决以上问题,有效提升模型的图嵌入学习能力.

在进一步工作,可将MCM-VGAE模型推广到大型拓扑网络或者动态网络、时序网络上进行部署应用. 此外,由于云模型可以实现对语义的不确定性表达,利用云模型理论构建带有语义的图神经网络模型,增加GNN的可解释性也是重要的研究方向.

参考文献

[1] 李德毅, 刘常昱, 杜鹢, 等. 不确定性人工智能[J]. 软件学报, 2004, 15(11): 1583-1594.
LI D Y, LIU C Y, DU Y, et al. Artificial intelligence with uncertainty[J]. Journal of Software, 2004, 15(11): 1583-1594. (in Chinese)

[2] 祁志卫, 王笳辉, 岳昆, 等. 图嵌入方法与应用: 研究综述[J]. 电子学报, 2020, 48(4): 808-818.
QI Z W, WANG J H, YUE K, et al. Methods and applications of graph embedding: A survey[J]. Acta Electronica Sinica, 2020, 48(4): 808-818. (in Chinese)

[3] 吴博, 梁循, 张树森, 等. 图神经网络前沿进展与应用[J]. 计算机学报, 2022, 45(1): 35-68.
WU B, LIANG X, ZHANG S S, et al. Advances and applications in graph neural network[J]. Chinese Journal of Computers, 2022, 45(1): 35-68. (in Chinese)

[4] KIPF T N, WELLING M. Variational graph auto-encoders [EB/OL]. (2016-11-21) [2022-03-06]. https://arxiv. org/abs/1611.07308.

[5] PAN S R, HU R Q, LONG G D, et al. Adversarially regularized graph autoencoder for graph embedding[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm: AAAI, 2018: 2609-2615.

[6] HASANZADEH A, HAJIRAMEZANALI E, DUFFIELD N, et al. Semi-implicit graph variational auto-encoders[EB/OL]. (2019-09-07)[2022-03-06]. https://arxiv.org/abs/1908.07078.

[7] AHN S J, KIM M. Variational graph normalized AutoEncoders[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 2827-2831.

[8] SALHA G, HENNEQUIN R, VAZIRGIANNIS M. Simple and effective graph autoencoders with one-hop linear models[M]//Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2021: 319-334.

[9] 李德毅, 刘常昱. 论正态云模型的普适性[J]. 中国工程科学, 2004, 6(8): 28-34.
LI D Y, LIU C Y. Study on the universality of the normal cloud model[J]. Strategic Study of CAE, 2004, 6(8): 28-34. (in Chinese)

[10] ZHU Q L, BI W, LIU X J, et al. A batch normalized inference network keeps the KL vanishing away[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 2636-2649.

[11] 杨洁, 王国胤, 刘群, 等. 正态云模型研究回顾与展望[J]. 计算机学报, 2018, 41(3): 724-744.
YANG J, WANG G Y, LIU Q, et al. Retrospect and prospect of research of normal cloud model[J]. Chinese Journal of Computers, 2018, 41(3): 724-744. (in Chinese)

[12] TIAN F, GAO B, CUI Q, et al. Learning deep representations for graph clustering[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec: AAAI, 2014: 1293-1299.

[13] Davidson T R, Falorsi L, De Cao N, et al. Hyperspherical variational auto-encoders[EB/OL]. (2018-04-03) [2022-03-06]. https://arxiv.org/abs/1804.00891.

[14] 过江, 张为星, 赵岩. 岩爆预测的多维云模型综合评判方法[J]. 岩石力学与工程学报, 2018, 37(5): 1199-1206.
GUO J, ZHANG W X, ZHAO Y. A multidimensional cloud model for rockburst prediction[J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(5): 1199-1206. (in Chinese)

[15] LIU L, ZHOU T Y, LONG G D, et al. Prototype propagation networks (PPN) for weakly-supervised few-shot learning on category graph[C]//Proceedings of the 28th

International Joint Conference on Artificial Intelligence. New York: ACM, 2019: 3015-3022.

[16] 代劲, 胡彪, 王国胤, 等. 分布轮廓与局部特征融合的云模型不确定性相似度量[J]. 电子与信息学报, 2022, 44(4): 1429-1439.
DAI J, HU B, WANG G Y, et al. The uncertainty similarity measure of cloud model based on the fusion of distribution contour and local feature[J]. Journal of Electronics & Information Technology, 2022, 44(4): 1429-1439. (in Chinese)

[17] 陈昊, 龙文佳. 高斯云模型的雾化特性[J]. 湖北大学学报(自然科学版), 2015, 37(6): 560-564.
CHEN H, LONG W J. The atomization characteristics in Gauss cloud model[J]. Journal of Hubei University (Natural Science Edition), 2015, 37(6): 560-564. (in Chinese)

[18] CHIANG W L, LIU X Q, SI S, et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 257-266.

[19] YANG Z, COHEN W W, SALAKHUTDINOV R. Revisiting semi-supervised learning with graph embeddings [EB/OL]. (2016-03-29)[2022-03-06]. https://arxiv.org/abs/1603.08861.

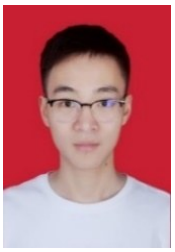
[20] SHCHUR O, MUMME M, BOJCHEVSKI A, et al. Pitfalls of graph neural network evaluation[EB/OL]. (2018-11-14)[2022-03-06]. https://arxiv.org/abs/1811.05868.

[21] FEY M, LENSSEN J E. Fast graph representation learning with PyTorch geometric[EB/OL]. (2019-03-06)[2022-03-06]. https://arxiv.org/abs/1903.02428.

[22] TU C C, ZENG X K, WANG H, et al. A unified framework for community detection and network representation learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(6): 1051-1065.

作者简介

代　劲　男,1978年出生于贵州遵义.博士,重庆邮电大学教授,硕士生导师. 主要研究方向为粒计算、认知计算、智能信息处理.
E-mail: daijin@cqupt.edu.cn



张奇瑞　男,1997年出生于四川德阳,硕士生. 主要研究方向为智能信息处理、数据挖掘.
E-mail: S201201035@stu.cqupt.edu.cn

王国胤　男,1970年出生于重庆市,博士,教授,博士生导师. 教育部“长江学者”特聘教授(2015-2019)、中组部“万人计划”科技创新领军人才(2014)、人社部“新世纪百千万人才工程”国家级人选、国务院特殊津贴专家、中科院“百人计划”专家、教育部“新世纪优秀人才”. 主要研究方向为粗糙集、粒计算、数据挖掘、认知计算、大数据、人工智能等.
E-mail: wanggy@cqupt.edu.cn

彭艳辉　女,1998年出生于重庆丰都,硕士生. 主要研究方向为智能信息处理、数据挖掘.
E-mail: S211201018@stu.cqupt.edu.cn

涂盛霞　女,1993年出生于湖北潜江,硕士,华为大数据高级工程师,主要研究方向为大数据引擎及性能优化.
E-mail: tushengxia@huawei.com