一种基于概率分布分层图聚类网络的社区检测模型

徐 森[1†]，刘轩绮[1]，陈朝峰[1]，郭乃瑄[1,2]，卞学胜[1]，马芙蓉[1]，花小朋[1]，周 天[3]

(1. 盐城工学院 信息工程学院，江苏 盐城 224051；
2. 东南大学 计算机网络和信息集成教育部重点实验室，南京 210000；
3. 哈尔滨工程大学 水声工程学院，哈尔滨 150001)

摘 要: 为了捕捉网络的隐藏结构, 减少社区检测模型对初始参数选择的依赖性, 提出一种基于概率分布分层图聚类网络 (HGCPD) 的社区检测模型. 首先, 利用图卷积网络学习和缓存图中节点的特征表示; 然后, 引入一种基于节点对相似度概率的分层聚类方法, 在不同层次上递归地构建社区结构; 最后, 探究模型超参数优化问题, 设计贝叶斯优化方法自动调整参数, 从而提升模型效率. 在多个不同规模的网络数据集上的实验表明, HGCPD 模型在社区检测的准确性、有效性均优于主流方法, 并通过可视化验证了所提出模型的可解释性.

关键词: 概率分布；多尺度结构；图卷积网络；贝叶斯优化

中图分类号: TP181；TP301 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0874

引用格式: 徐森, 刘轩绮, 陈朝峰, 等. 一种基于概率分布分层图聚类网络的社区检测模型 [J]. 控制与决策.

A community detection model based on hierarchical graph clustering network with probability distribution

XU Sen[1†]，*LIU Xuan-qi*[1]，*CHEN Chao-feng*[1]，*GUO Nai-xuan*[1,2]，*BIAN Xue-sheng*[1]，*MA Fu-rong*[1]，*HUA Xiao-peng*[1]，*ZHOU Tian*[3]

(1. School of Information Engineering，Yancheng Institute of Technology，Yancheng 224051，China；2. Key Laboratory of Computer Network and Information Integration，Southeast University，Nanjing 210000，China；3. School of Underwater Sound Engineering，Harbin Engineering University，Harbin 150001，China)

Abstract: In order to capture the hidden structure of networks and reduce the dependence of community detection models on the choice of initial parameters, this paper proposes a community detection model based on hierarchical graph clustering networks with probability distribution (HGCPD). Firstly, graph convolutional networks are utilized to learn and cache feature representations of nodes in the graph. Then, a hierarchical clustering method based on node pair similarity probability is introduced to recursively construct community structures at different levels. Finally, the problem of model hyperparameter optimization is explored, and the Bayesian optimization methods are designed to automatically adjust parameters, thereby improving the efficiency of the model. Experiments on multiple network datasets of different scales show that the HGCPD model is superior to mainstream methods in terms of the accuracy and effectiveness of community detection, and the model's interpretability is verified through visualization.

Keywords: probability distribution；multi-scale structure；graph convolutional networks；Bayesian optimization

0 引 言

随着网络科学和图数据分析的迅速发展, 社区检测已成为理解和分析复杂图结构的关键任务, 广泛应用于推荐系统、社交网络分析和文献检索等领域. 最初的社区检测方法, 如 Girvan-Newman (GN) 算法[1]、Infomap 算法[2] 和 Walktrap 算法[3], 为社区检测领域奠定了基础, 然而, 它们在面对大规模和高维结构特征网络时, 往往难以同时捕捉到图的局部细节和全局结构. 随后, 专注于在大型社交网络中搜索社区的算法不断涌现, 但是也伴随着聚类效率低下、

收稿日期: 2024-07-22；**录用日期:** 2024-11-20.

基金项目: 国家自然科学基金项目 (62076215,62301473)；中央高校基本科研业务费专项资金项目 (K93-9-2022-03)；江苏省高等教育厅面上项目 (23KJB520039)；江苏省网络与信息安全重点实验室项目 (BM2003201)；江苏高校“青蓝工程”项目；盐城市基础研究计划项目 (YCBK2023008).

†通讯作者. E-mail: xusen@ycit.cn.

鲁棒性较低等现象.

深度学习的强大技术为聚类方法提供了更好的解决方案. 图神经网络 (GNN) 因其在编码图拓扑和节点特征方面的卓越能力而被广泛应用. 近期, Tan等[4] 提出了一种自监督图自编码器模型, 可通过图掩蔽提高图自编码器 (GAE) 在各种图分析任务上的通用性, 大部分基于 GNN 模型的方法均是沿着每个相邻节点间迭代地传递信息, 这使得聚合远程信息的能力尤为关键. 为此, 将层次聚类思想与 GNN 模型进行有效结合的方法诞生了, 如层次社区感知图神经网络[5] 模型通过将平面图中的节点组织为多层次的超级图来降低编码远程信息时的计算成本, 但是仍然局限于单一的社区结构.

为了在更高层次上关注社区间的关系, 提高聚类性能, 本文提出基于概率分布的分层图聚类网络 HGCPD, 结合 GNN 的表示学习和层次聚类算法的多尺度分析能力, 用概率分布的方法更精确地揭示图中社区的嵌套结构.

1 相关工作

李慧等[6] 提出了一种基于时间加权的重叠社区检测算法, 解决了局部相似度的问题; Newman 等[7] 提出了一个优化模块度度量的 Louvain 算法, 这是社区检测算法的一个里程碑; Clauset 等[8] 进一步优化了 Louvain 算法, 降低其时间复杂度. 大多数基于模块度优化的方法在处理大规模网络时仍然面临性能瓶颈, 且缺乏对网络多层次结构的考虑. 为了进一步提高算法的鲁棒性和准确性, 层次聚类算法包括 Hierarchical Louvain[9] 等模型的提出为社区检测领域带来了新的视角, 这些方法通过逐步合并节点或社区来构建网络的层次结构. 然而, 这些方法通常依赖参数设置, 尤其是在处理大规模网络时的效果很小. 本文结合全概率原理的余弦相似度方法对 Louvain 进行改进, 以减少对初始化参数的依赖.

为了学习非线性网络特性、保留复杂网络结构的低维特征嵌入, 基于深度学习的社区检测方法也得到了更多的探索. 图卷积网络 (GCN) 是一种能够对图结构进行端到端学习的模型, 可捕捉拓扑结构和节点属性信息. 对于大规模图数据, 张建朋等[10] 提出了能够有效保持原始图聚类结构的图采样算法, 此采样算法能够很好地保持原始图的内在聚类结构, 解决了计算复杂度过高、难以适用于大规模图的聚类问题. 陈洁等[11] 提出了一种基于影响力与种子扩展的重叠社区发现算法, 利用节点影响力策略找出具有紧密结构的种子社区. 然而, 大多数 GCN 模型侧重于单个层次的节点表示学习, 尽管 Cluster-GCN 模型[12] 将 GCN 与聚类进行了高效结合, 但是, 仍然很少有工作探讨如何将 GCN 与层次聚类相结合有效聚合不同层次的远程信息.

2 本文方法

定义 1 给定一个图$G=(V,E,W)$. 其中: V为节点集, $V=1,2,\ldots,n$, 节点$\forall i$、$j\in V$; E为边集, 若$e_{ij}=(v_i,v_j)\in E$, 则节点i与j间存在一条边, 且$w_{ij}=w_{ji}$为边e_{ij}的权重. 给定一社区集$C=\{C_1,C_2,\ldots,C_k\}$, 每个社区C_i均为网络G的划分.

2.1 GCN 更新节点特征

对于每个节点$v\in V$可能有一个特征向量xv, 任意两个节点i与j间存在的边表示它们的相似性, 设$A=[A_{ij}]$为一个$N\times N$维的加权邻接矩阵, 若$e_{ij}\in E$, 则$a_{ij}=1$; 反之, 则$a_{ij}=0$. 对于度矩阵$D=D_{ii}=\sum_j a_{ij}$, 将度矩阵的逆平方根对邻接矩阵进行归一化处理, 可得到一个对称归一化矩阵$\hat{A}=D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. 该矩阵只与图的结构相关, 可保证节点特征在图卷积中的稳定性和有效性. 通过 GCN 的多层卷积操作, 节点特征逐步聚合自身的特征和邻居节点的信息, 形成相似节点的低维空间表示. 具体地, 对于一个若干层的 GCN, 每层的节点表征聚合后的更新传播规则为

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

其中: $\tilde{A}=A+I$为加上自环的邻接矩阵; $\tilde{D}$为一个对角矩阵, $\tilde{D}_{ii}=\sum_j \tilde{a}_{ij}$; $\sigma(\cdot)$为一个激活函数; l 为 GCN 的层数; $H^{(l)}$为第 l 层节点的特征矩阵, 初始$H^{(0)}=x_v$; $W^{(l)}$为神经网络的权重. 经L层 GCN 后的节点特征嵌入为最终的输出, 即

$$Z_v=H^{(L)}. \tag{2}$$

最终获得节点嵌入矩阵$Z_v\in\mathbb{R}^{(N\times D)}$. 这里: N为节点数, D为节点嵌入的特征维度.

为了引导网络学习有效的节点特征表示, 本文引入了一种基于重构的自监督学习损失函数, 它通过图本身的邻接关系进行监督, 避免了对大量标注数据的需求. 损失函数如下所示:

$$\mathcal{L}_{\text{recon}}=\sum_{i=1}^{N}\sum_{j=1}^{N}[A_{ij}\log\sigma(H_i^{\mathrm{T}}H_j)+(1-A_{ij})\log(1-\sigma(H_i^{\mathrm{T}}H_j))]. \tag{3}$$

其中: H_i和H_j分别为节点i和节点j的嵌入特征, $\sigma(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$用于将节点嵌入之间的点积转换为 [0, 1] 的概率值. 该重构损失函数通过H_i与H_j的点

积来近似反映图的邻接关系, 即若节点i与节点j间有边, 则$\sigma(H_iH_j)$接近 1; 反之, 则$\sigma(H_iH_j)$接近于 0.

2.2 节点对采样

2.2.1 节点对采样概率计算

为了采样相似节点对, 基于这些嵌入特征用余弦相似度来反映节点间的相似性. 记每个节点i的嵌入特征为$h_i^{(L)}$, 节点对余弦相似度的两个向量夹角为

$$\mathrm{sim}(i,j)=\frac{h_i^{(L)}\cdot h_j^{(L)}}{\|h_i^{(L)}\|\|h_j^{(L)}\|}. \tag{4}$$

其中: $h_i^{(L)}\cdot\ h_j^{(L)}$为向量的点积, $\|\cdot\|$为向量的范数. 为了获取更精确的簇团, 利用 softmax 函数将相似度转化为节点对的采样概率, 如下所示:

$$p(i,j)=\frac{\exp(\mathrm{sim}(i,j))}{\sum\limits_{k\neq i}\exp(\mathrm{sim}(i,k))}, \tag{5}$$

这里k为除节点i外的所有邻居节点, 可由邻接矩阵$A=[a_{ij}]$获得. k用于归一化余弦相似度, 使得节点i与其所有邻居节点的相似度的和为 1, 构造出概率分布. 节点对采样概率越高, 节点对距离越小, 两个节点越相似.

2.2.2 节点对距离计算

观察到, 全概率法是通过将每个节点对的联合概率转化为边缘概率的乘积, 并通过总和得到边缘概率的特点. 为了在无监督条件下实现节点对间距离的估计, 基于节点对采样概率计算的节点对距离, 有

$$d(i,j)=\frac{p(i)p(j)}{p(i,j)}, \tag{6}$$

其中$p(i)=\sum\limits_j p(i,j)$为单个节点$i$的边缘概率, 通过采样概率的总和进行估计. 若$p(i,j)=0$, 则$d(i,j)=+\infty$, 即节点$i$与$j$不是邻居. 这种距离度量的方法结合了节点特征的嵌入表示和概率计算, 能够更准确地反映节点间的相似度, 以挖掘更具代表性的社区结构.

类似地, 考虑图聚类在社区检测任务上对应的簇即为社区C, 相应的簇对的概率分布为

$$p(a,b)=\sum_{i\in a,j\in b}p(i,j),\ \forall a,b\in C. \tag{7}$$

对于单个簇a而言, 其概率分布可表示为

$$p(a)=\sum_{i\in a}p(i)=\sum_{b\in C}p(a,b),\ \forall a\in C. \tag{8}$$

因此, 类比到不同簇对a与b间的距离采样计算可由节点对距离公式推导, 即

$$d(a,b)=\frac{p(a)p(b)}{p(a,b)}. \tag{9}$$

当$d(a,b)=+\infty$时, 簇对a与b间不存在相连的边, 即$p(a,b)=0$. 这个计算的距离结果将用于层次聚类算法中合并最接近的簇类. 为了确保同一社区内节点具有高度相似的嵌入, 社区紧凑性损失如下所示:

$$\mathcal{L}_{\mathrm{compact}}=\sum_{i\in C}\sum_{j\in C}d(i,j). \tag{10}$$

2.3 社区检测算法

2.3.1 节点对采样的 Louvain 算法

本文基于传统的 Louvain 算法进行改进, GCN 节点嵌入帮助存储边的信息; 基于概率分布的距离计算使得聚类更加精确, 这两者的结合不仅解决了局部最优的问题, 还通过简单参数的引入控制了影响采样概率相似度的阈值范围.

开始时, 每个节点作为一个单独的社区$C_0=\{\{1\},\{2\},\ldots,\{n\}\}$. 通过最近邻链方案高效精确地找到最相似的社区进行合并, 逐步构建层次聚类树. 基于社区对的采样概率, 在距离矩阵中找到距离最近的两个社区a和b, 将其合并为一个新社区$a\bigcup b$, 对于任意不同的聚类簇a、b、$c\in C$, 有

$$d(a\bigcup b,c)=\left(\frac{p(a)}{p(a\bigcup b)}\frac{1}{d(a,c)}+\frac{p(b)}{p(a\bigcup b)}\frac{1}{d(b,c)}\right)^{-1}, \tag{11}$$

其中$p(a\bigcup b)$为新合并后社区$a\bigcup b$的边缘概率. 每次动态地合并社区后, 更新新社区与其他社区的距离, 能够准确细致地捕捉图中的社区结构, 适用于不同分辨率的社区检测需求. 当然, 对于任意的不同簇a、b、$c\in C$有一定的可约性, 即

$$d(a\bigcup b,c)\geqslant\min(d(a,c),d(b,c)). \tag{12}$$

由可约性可知, 簇a与b合并的距离至少要大于等于它们到其他任何聚类簇c的最小距离, 且合并过程中聚类间的距离$d_0,d_1,\ldots,d_{n-1}$是非递减的, 这里$d_0=0$. 这意味着合并的过程中不会减少聚类间的距离, 即没有倒置现象, 这也保证了生成的树状图是规则的. 该算法每次合并均会产生一个新的聚类结果, t为合并的次数, 对于$t=0,1,\ldots,n-1$, C_t包含$n-t$个聚类.

2.3.2 模块度优化

Louvain 算法的基本思想是将节点一个接一个地移动至其他社区, 重复至无法进一步增加模块度为止, 该模块度的值即为最大化模块度, 其定义为

$$Q(C)=\frac{1}{\omega}\sum_{i,j\in V}\left(A_{ij}-\frac{\omega_i\omega_j}{\omega}\right)\delta_C(i,j). \tag{13}$$

由邻接矩阵、节点权重与节点对采样概率的转换关系, 其模块度表示可用概率分布表示为

$$Q(C)=\sum_{i,j\in V}(p(i,j)-p(i)p(j))\delta_C(i,j). \quad (14)$$

其中: $\delta_C(i,j)$为一个指示函数, 若i与j属于同一个社区, 则取值为 1, 即当$C_i=C_j$时, $\delta_C(i,j)=1$; 否则, 为$\delta_C(i,j)=0$. 由此可见, Q值其实反映了实际连接到节点的边数与期待边数的差异, 然而, 任何最大化模块的聚类均存在一定的分辨率限制, 为了提高普适性引入乘法因子γ. γ为分辨率参数, 本文采用启发式的方法动态调整分辨率参数γ来优化模块度, 以确保社区划分既能反映局部结构, 又能适应整体图结构. 参数模块度为

$$Q_\gamma(C)=\sum_{i,j\in V}(p(i,j)-\gamma p(i)p(j))\delta_C(i,j). \quad (15)$$

对应到集群层面的模块度为

$$Q_\gamma(C)=\sum_{a\in C}(p(a,a)-\gamma p(a)^2). \quad (16)$$

通过启发式方法动态调整参数主要是当模块度增加时, 增大γ来促进社区进一步细分, 揭示更详细的结构; 当模块度减少时, 减小γ以防止过度分割, 保持社区结构. 依据如下所示:

$$\gamma_{t+1}=\begin{cases}\gamma_t(1+\delta), & Q_{t+1}>Q_t;\\ \gamma_t(1-\delta), & Q_{t+1}<Q_t.\end{cases} \quad (17)$$

这里: δ为一个很小的正常数, 本文默认其值为 0.05; 并初始化$\gamma=\gamma_0$, 对于每次迭代t从 1 开始, 从而不断更新下一次迭代的γ.

2.3.3 损失函数

为了获取更好的聚类性能, 保证类内高度相似性和类间连接稀疏性, 需要综合考虑这些过程所产生的各种损耗, 因此总损失为

$$\mathcal{L}_{\text{final}}=\alpha\mathcal{L}_{\text{recon}}+\beta\mathcal{L}_{\text{modularity}}+\eta\mathcal{L}_{\text{compact}}. \quad (18)$$

其中: $\mathcal{L}_{\text{modularity}}=-Q_\gamma(C)$, α、β和η为平衡 3 个损失函数的加权超参数.

3 实 验

实验采用 4 个真实世界网络的数据集: DBLP、OpenStreet、Amazon 和 Youtube, 它们均由节点和边表示, 其中 OpenStreet 是从 OpenStreetMap5 中提取的巴黎市中心的街道图. 本文还选择了一些常用的、经典的聚类方法作为对比模型, 分别为 Label Propagation (LPA)[13]、Louvain[1]、Constrained Louvain[14]、H-Louvain[9]、Infomap[2] 和 Smart Local Moving (SLM)[15], 这 6 个基准模型适用于不同规模和结构的数据集. 本文共采用 3 种评价指标来衡量聚类的质量和准确性, 分别为 NMI[12]、ARI[12] 和 F1[12], 这 3 个评价指标的值越大, 聚类质量越高.

实验分为两部分: 1) 通过与其他经典模型在各评价指标上的对比实验, 验证所提出模型相较于其他模型性能的优越性; 2) 对模型中的超参数进行贝叶斯优化, 通过寻找合适的参数值搭配来验证所提出模型相较于其他模型受参数影响较小, 且参数调整更方便.

3.1 实验结果

3.1.1 模型对比实验

在模型评估实验中, 对损失函数中的超参数是通过贝叶斯优化自动选择的, 步骤如下.

step 1: 定义目标函数$f(\alpha,\beta,\eta)=-F1_{\text{validation}}(\alpha,\beta,\eta)$, 其中 F1 为模型在验证集上的 F1 分数, 负号即希望的最小化损失.

step 2: 定义超参数$\alpha\in[0.01,10]$, $\beta\in[0.01,10]$, $\eta\in[0.01,10]$, 并选择一个贝叶斯优化库 scikit-optimize 来实现优化过程, 使用高斯过程作为贝叶斯优化的后端模型.

step 3: 定义高斯过程上置信区间宽度 (GP-UCB) 采集函数$\text{GP}-\text{UCB}(x)=\mu(x)+k\sigma(x)$. 其中: x为参数空间中的一个点; $\mu(x)$为x处目标函数的预测均值; $\sigma(x)$为在x处目标函数预测的标准差; k用于控制探索与利用之间权衡的超参数, 调整k值来权衡探索与利用, 并运行贝叶斯优化. 表 1 为各模型在 4 个数据集上的不同评价指标结果, 通过对比可以看出各模型的性能.

由表 1 可见: 所提出方法在大多数数据集中表现出显著的性能优势, 特别是在 DBLP 数据集上, 所有评价指标的平均性能均超过了其他模型, 可见该模型在处理节点数和维度较少的这类图数据时具有更好的适应性和鲁棒性; 在较为复杂的 Amazon 和 Youtube 数据集上所提出模型表现更为突出, 与其他模型相比, 所提出模型在这两个数据集上聚类指标均有显著提高, 可见不管用户间的关系是否稳定, 所提出模型在处理大规模数据集上有一定的优势; 然而, 在 OpenStreet 数据集上的表现相对较弱, 可能是该数据集的网络结构具有较高的稀疏性和复杂的地理性, 使得社区检测任务更具有挑战性. 总而言之, 实验结果清楚地表明了 HGCPD 模型在社区检测任务中的有效性和优势, 相比于经典的基础模型可以更好地适应不同类型的数据集, 可见 HGCPD 可以更好地捕捉节点间的结构信息, 提高社区检测的准确性.

表1　基准模型和本文方法的聚类性能

数据集	指标	LPA	Louvain	Constrained Louvain	H-Louvain	Infomap	SLM	HGCPD
DBLP	ARI	0.47	0.65	0.75	0.76	0.04	0.03	**0.77**
	NMI	0.75	0.81	0.82	0.89	0.74	0.72	**0.91**
	F1	0.78	0.85	0.86	0.87	0.73	0.79	**0.88**
OpenStreet	ARI	0.39	0.64	0.72	0.73	0.57	0.47	**0.76**
	NMI	0.74	0.79	**0.92**	0.89	0.65	0.42	0.88
	F1	0.72	0.77	0.77	**0.87**	0.83	0.72	0.86
Amazon	ARI	0.07	0.63	0.71	0.70	0.31	0.42	**0.72**
	NMI	0.73	0.78	0.92	0.86	0.54	0.36	**0.94**
	F1	0.71	0.76	0.86	0.84	0.78	0.69	**0.87**
YouTube	ARI	0.05	0.62	0.66	0.67	0.38	0.39	**0.68**
	NMI	0.72	0.77	0.78	0.85	0.48	0.34	**0.86**
	F1	0.70	0.75	0.75	0.82	0.72	0.61	**0.83**

3.1.2　贝叶斯优化参数

为了提高聚类精度, 提升社区检测的准确性, 本文使用 GCN 特征更新捕捉节点的隐藏结构信息, 同时使用贝叶斯优化方法, 优化过程同上述实验. 通过对采样的数据集构建高斯模型并迭代优化采集函数, 对 GCN 中的学习率l_r、GCN 层数L、损失函数中的正则化系数λ以及分辨率参数γ进行优化. 初始设置$L=2$, $l_r=0.001$, 隐藏层为 128; $\gamma_0=1.0$, $\delta=0.05$, $N_{\text{Iter}}=200$.

记录优化前后的参数和评价指标可以发现: 收集到最佳搭配的参数值后, 各数据集上的性能均有明显提高, 且 F1 的数值平均提高了 1 %. 各参数最优值如表 2 所示.

表2　建立模型所用变量

数据集	l_r	L	λ	γ	ARI	NMI	F1
DBLP	0.001	3	0.001	0.8	0.778	0.918	0.889
Open Strteet	0.01	4	0.005	1.1	0.769	0.893	0.868
Amazon	0.002	2	0.002	1.2	0.734	0.952	0.881
Youtube	0.001	3	0.001	1.0	0.693	0.877	0.843

通过构建高斯模型近似目标函数, 并使用该模型指导参数选择, 从而逐步逼近最优全局解. 在实验中发现随着迭代次数的增加, 合适参数下各数据集中各评价指标逐渐趋于稳定, 体现了良好的收敛性. 优化后的分辨率参数在 0.8 ～ 1.2 之间, 表明不同数据集对于社区检测的分辨率需求不同, 并减少了模型对参数初始值的依赖程度, 调整分辨率参数能够显著提高聚类精度. 该实验验证了贝叶斯优化在参数调优中的有效性. 但是, 为了寻找最优的超参数组合, 需要多次迭代和评估, 整个优化过程耗时较长.

3.2　可 视 化

可视化的图表可从多个方面分析评估模型的性能和有效性, 以更直观的方式深入理解数据集的内部结构和聚类算法在不同数据集上的性能表现. 本节对 HGCPD 模型在上述 4 个数据集聚类后的社区结构进行可视化分析. 如图 1 所示, 本文通过生成具有明显的团簇分布状态来创建一个社区结构. 由图 1 可以清晰地观察到每个簇的分布状态, 同时帮助发现模型在社区划分上的不足, 从而进行调整和优化. 若簇间联系越稀疏，簇内联系越紧密, 则模型在识别社区结构上表现良好.

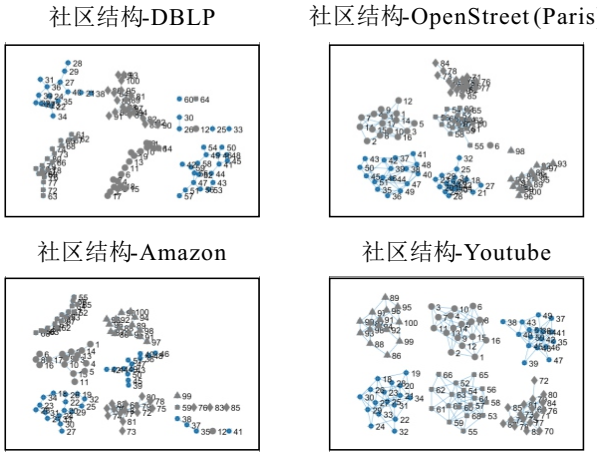

图1　社区结构可视化

可以看出, 对于大部分数据集, 所提出模型能够有效地划分各簇, 且在大规模的、复杂的数据集上相对具有清晰的聚类效果. 图 1 中: Amazon 数据集中有部分节点孤立在外, 可见其类内相似度稍低, 可能是因为该数据集中包含了用户对商品的评价和记录, 而不同用户对商品需求的不同会导致其对商品功能、质量以及评价等存在很大的差别, 个人情感色彩较浓, 因此, 数据集中包含更多的噪声数据, 易使得聚类效果不理想; DBLP 数据集的聚类存在一些界限模糊的节点, 这可能是因为 DBLP 包括“合著者”“共同任期”和“联合会议”, 而一些作者可能同时与

几个研究领域均有一定的关系, 也可能同时担任会议作者和任期作者等. 整体上这些图直观地显示了模型在社区检测上的能力, 可见所提出模型在面对噪声、不确定性等干扰时仍然具有一定的鲁棒性, 也有效地反映出模型在社交网络分析、推荐系统等领域的应用潜力.

4 结 论

本文提出了一种结合 GCN 和节点对采样概率的分层图聚类网络, 通过优化节点嵌入特征和社区结构, 提高了图数据的聚类性能. 未来希望结合其他深度学习技术将多层网络信息融合为一层, 并保留交互类型间的差异.

参考文献 (References)

[1] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[2] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118-1123.

[3] Pons P, Latapy M. Computing communities in large networks using random walks[C]. Computer and Information Sciences-ISCIS 2005. Istanbul, 2005: 284-293.

[4] Tan Q Y, Liu N H, Huang X, et al. S2GAE: Self-supervised graph autoencoders are generalizable learners with graph masking[C]. Proceedings of the 16th ACM International Conference on Web Search and Data Mining. Singapore, 2023: 787-795.

[5] Zhong Z Q, Li C T, Pang J. Hierarchical message-passing graph neural networks[J]. Data Mining and Knowledge Discovery, 2023, 37(1): 381-408.

[6] 李慧, 马小平, 张舒, 等. 基于时间加权的重叠社区检测算法研究[J]. 自动化学报, 2021, 47(4): 933-942.
(Li H, Ma X P, Zhang S, et al. Research of overlap community detection algorithm based on time-weighted[J]. Acta Automatica Sinica, 2021, 47(4): 933-942.)

[7] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[8] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 70(6 pt 2): 066111.

[9] Han Z X, Shi L L, Liu L, et al. H-Louvain: Hierarchical Louvain-based community detection in social media data streams[J]. Peer-to-Peer Networking and Applications, 2024, 17(4): 2334-2353.

[10] 张建朋, 陈鸿昶, 王凯, 等. 基于采样的大规模图聚类分析算法[J]. 电子学报, 2019, 47(8): 1731-1737.
(Zhang J P, Chen H C, Wang K, et al. A sampling-based graph clustering algorithm for large-scale networks[J]. Acta Electronica Sinica, 2019, 47(8): 1731-1737.)

[11] 陈洁, 李锐, 赵姝, 等. 面向图表示社区检测的新型聚类覆盖算法[J]. 电子学报, 2020, 48(9): 1680-1687.
(Chen J, Li R, Zhao S, et al. A new clustering cover algorithm based on graph representation for community detection[J]. Acta Electronica Sinica, 2020, 48(9): 1680-1687.)

[12] 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. Anchorage, 2019: 257-266.

[13] Hartigan J A, Wong M A. Algorithm AS 136: A K-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100-108.

[14] Yao B B, Zhu J F, Ma P J, et al. A constrained Louvain algorithm with a novel modularity[J]. Applied Sciences, 2023, 13(6): 4045.

[15] Waltman L, van Eck N J. A smart local moving algorithm for large-scale modularity-based community detection[J]. The European Physical Journal B, 2013, 86(11): 471.

作者简介

徐森 (1983−), 男, 教授, 博士, 主要研究方向为机器学习、数据挖掘、人工智能, E-mail: xusen@ycit.cn;

刘轩绮 (1998−), 女, 硕士生, 主要研究方向为深度学习、图聚类, E-mail:1138732172@qq.com;

陈朝峰 (1982−), 男, 副教授, 博士, 主要研究方向为深度学习、图像处理, E-mail: zhaofeng.chen@foxmail.com;

郭乃瑄 (1991−), 男, 讲师, 博士, 主要研究方向为网络安全、图像处理, E-mail:guonaixuan@ycit.edu.cn;

卞学胜 (1991−), 男, 讲师, 博士, 主要研究方向为深度学习与医学影像处理, E-mail:xsbian@ycit.edu.cn;

马芙蓉 (1995−), 女, 讲师, 博士, 主要研究方向为计算机视觉、图像生成, E-mail:mafurong_M@163.com;

花小朋 (1975−), 男, 副教授, 博士, 主要研究方向为机器学习、数据挖掘, E-mail:1462124471@qq.com;

周天 (1980−), 男, 教授, 博士, 博士生导师, 主要研究方向为水声目标探测与定位、水声信号处理、声纳系统设计, E-mail:zhoutian@hrbeu.edu.cn.