

基于网络表示学习的深度社团发现方法

潘雨^{1,2} 邹军华¹ 王帅辉³ 胡谷雨¹ 潘志松¹

1 陆军工程大学指挥控制工程学院 南京 210007

2 中国人民解放军第 31436 部队 沈阳 110000

3 海军航空大学第三飞行训练基地 河北 秦皇岛 066000

(pan_yu31@163.com)

摘要 挖掘复杂网络中的社团结构有助于理解网络内部结构和功能特性,具有重要的理论价值和实际应用意义。随着信息技术的飞速发展,爆炸式增长的网络数据为社团发现任务提出了前所未有的挑战。为此,文中利用深度神经网络将网络表示学习和社团发现领域相连接,提出一种基于网络表示学习的深度社团发现方法。算法首先根据节点潜在的社团成员相似性来量化节点之间的结构相似度,从而构造包含潜在社团结构信息的社团结构矩阵;然后建立由多个非线性函数组成的多层自编码器,将社团结构矩阵作为深度自编码器的输入,获得保存了潜在社团结构的节点低维表示;最后在网络表示上应用 K-means 聚类策略获得社团结构。在不同规模的真实网络和人工网络上进行了大量的实验,并与典型的算法进行比较,实验结果表明了算法的可行性和有效性。

关键词: 社团发现;网络表示学习;自编码器;深度神经网络;复杂网络

中图法分类号 TP393

Deep Community Detection Algorithm Based on Network Representation Learning

PAN Yu^{1,2}, ZOU Jun-hua¹, WANG Shuai-hui³, HU Gu-yu¹ and PAN Zhi-song¹

1 School of Command and Control Engineering, Army Engineering University, Nanjing 210007, China

2 The 31436 Unit of the Chinese People's Liberation Army, Shenyang 110000, China

3 The Third Flight Training Base of Naval Aeronautical University, Qinhuangdao, Hebei 066000, China

Abstract Mining the community structure in the complex network is helpful to understand the internal structure and functional characteristics of the network, which has important theoretical value and significant practical significance. With the rapid development of information technology, the explosive growth of network data poses an unprecedented challenge for community detection. In this paper, the deep neural network is utilized to connect network representation learning and community detection domains, and a deep community detection method based on network representation learning is proposed. Firstly, the structural closeness of nodes is quantified according to their potential community membership similarities, and then a novel community structure method is proposed to construct the community structure matrix. Furthermore, a deep autoencoder that has several layers with non-linear functions is developed. The community structure matrix is used as the input of the deep autoencoder to obtain the low-dimension representation of the nodes which preserve the potential community structure. Finally, the K-means clustering strategy is applied to the network representation to obtain the community structure. Extensive experiments on both synthetic and real-world datasets of different scales demonstrate that the proposed method is feasible and effective.

Keywords Community detection, Network representation learning, Autoencoder, Deep neural network, Complex network

1 引言

社团结构是广泛存在于网络中的重要结构特征,在社团内部的节点之间连接紧密,社团之间的节点连接稀疏^[1]。社团结构的挖掘为探索复杂网络的潜在特征提供了有效工具,对理解复杂网络具有重要意义。近年来,随着网络的发展和社交媒体的涌入,复杂网络逐渐呈现出大规模、稀疏、动态和高维的特征。这对社团发现提出了前所未有的挑战。

随着网络表示学习技术的发展,其为网络数据挖掘和分

析提供了有效的解决手段和途径。在对网络数据进行挖掘时,网络数据的有效分析很大程度上依赖于网络的表示方式,网络表示的好坏直接影响到后续挖掘任务的开展。网络表示学习旨在学习网络中节点潜在的低维表示,将网络中的节点表示为低维、实值、稠密的向量形式,学习到的节点低维表示同时保留了网络拓扑结构、节点属性和其他网络潜在的性质,形成更具表征性的节点表示^[2]。

随着深度学习在网络表示领域的研究愈发深入,相比于线性模型表示能力的局限性,深度模型的网络表示学习方法

可以挖掘和捕捉到网络的非线性结构;同时拥有较小的复杂度,当节点稀疏连接时,它的计算复杂度可以与网络中的节点数呈线性关系。受此启发,利用深度模型来学习网络的节点表示,然后在嵌入空间对社团进行发现,可以极大地保持高效的性能和计算速度,并且拥有可移植性和较强的学习特征能力,对网络稀疏性的问题也更有弹性。

虽然大多数网络表示学习算法得到的网络低维表示,在其上运行 K-means 等聚类策略都可以得到网络的社团结构,但学习到的网络表示没有保存面向社团的信息,因此对于社团发现任务是次优的。如何将网络表示学习和社团发现相结合,针对社团发现任务特点学习节点表示,从而获得优异的社团发现性能,主要的任务为:(1)网络中属于同一个社团的节点,在节点低维表示空间中也应该彼此靠近;(2)低维向量空间的嵌入应该具有良好的社团结构组织性,可以直接应用于后续的 K-means 等聚类策略,进而获得准确的社团结构。

针对以上挑战,本文利用深度学习技术将网络表示学习和社团发现两个领域进行结合,提出一种基于网络表示的深度社团发现方法 NECD,算法的整体流程如图 1 所示。首先,根据节点之间潜在的社团关系相似性构建保存了社团结构的

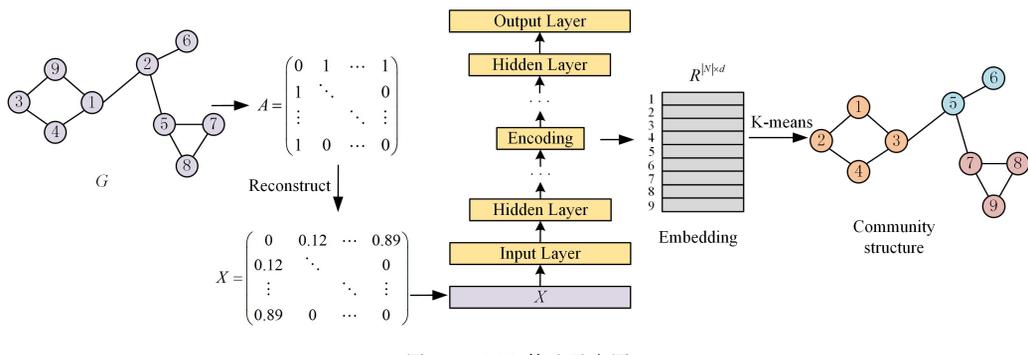


图 1 NECD 算法示意图

Fig. 1 Schematic diagram of NECD

2 相关工作

随着网络表示学习的不断发展和深度学习在网络表示学习中的广泛应用,基于网络表示学习的社团发现算法也逐渐进入研究者的视线,越来越多的研究员开拓思路,探索深度学习在社团发现中的可能性。Tian 等^[3]基于自动编码器和谱聚类之间的相似性,提出了基于稀疏自动编码器的图聚类方法 SAE。SAE 将归一化后的图相似度矩阵输入到以稀疏自动编码器为基础的深度神经网络中,并通过目标函数上引入 L_1 正则化项得到稀疏的非线性节点低维表示。然后对低维的节点表示运行 K-means 策略,得到最终的聚类结果。随后, Yang 等^[4]证明了随机生成模型和模块度最大化模型本质上的等价性,两个模型都是通过寻找网络的低秩嵌入来重构网络拓扑,这与自编码的目标不谋而合。因此,该团队提出了半监督的深度神经网络社团发现方法,将模块度矩阵作为自编码器的输入,将其映射到低维嵌入空间,然后利用 K-means 对低维空间的节点向量表示进行社团划分。Jin 等^[5]通过神经网络同时融合网络的结构信息和内容信息,将包含网络结构信息的模块度矩阵和包含网络属性信息的马尔科夫矩阵作为自编码器的输入,得到融合两种信息的网络低维表示,然后在节点表示向量上通过运行聚类算法得到最终的社团结构。DeepInNet^[6]是一个针对图聚类的深度表示学习模型,模型首

加权矩阵。具体来说,首先设计一个相似度函数来量化节点之间社团成员的相似度;然后在其基础上采用 Skip-gram 模型探索网络潜在的社团结构,得到嵌入了网络潜在社团结构的矩阵;接着将其作为深度自编码器的输入,通过深度自编码器提取社团结构矩阵的特征,将网络中的节点映射到低维嵌入空间,使在同一个社团的节点在嵌入空间中也彼此相近;最后将 K-means 聚类策略应用于节点的低维向量表示中,获得准确的社团结构。文章的主要贡献如下:

(1)根据节点潜在的社团成员相似度来量化节点之间的结构邻近度,构造出包含潜在社团结构的矩阵,将其作为网络信息的提取器。

(2)构建了一个由多个非线性函数组成的多层自编码器,通过重构社团结构矩阵,在低维表示空间中保存节点之间的社团关系。

(3)在生成的节点表示向量上运行 K-means 算法获得网络的社团结构。

(4)在多个真实和人工数据集上进行了大量的实验,实验结果表明本文提出的方法与各种基准方法相比取得了优异的社团划分结果。

先解决原始网络中的噪声和稀疏问题;然后将处理过的链接网络和节点内容网络作为栈式自编码器的输入,获得异构网络的深度表示;最后,算法引入 K-means 策略在深度集成表示中检测社团结构。Wu 等^[7]首先对网络空间信息进行重构,基于意见领袖和更近邻策略重构邻接矩阵,得到空间邻近矩阵;然后提出一种基于自动编码器和卷积神经网络的特征提取方法来提取重构矩阵的空间特征;最后在提取的空间特征上应用 K-means 算法获得社团结构。Cao 等^[8]提出了一种栈式自编码器,通过结合网络拓扑和节点属性进行社区检测。为了进一步解决网络拓扑和节点属性之间的匹配问题,又提出通过引入自适应参数用于调整两种信息之间的权重,开发了一种图正则化的自编码器方法用于社团发现^[9]。SCD^[10]是一种基于网络节点嵌入聚类的社团发现方法,将社团发现和网络表示学习相结合,通过优化轮廓度量将社团发现问题转化为网络嵌入聚类任务。

3 基于网络表示学习的深度社团发现算法

3.1 模型构建

本文定义网络表示为 $G=(V, E)$, $V=\{v_1, v_2, \dots, v_n\}$ 是节点集合, $E \subseteq (V \times V)$ 是边的集合, n 代表网络中节点的个数。设邻接矩阵 $A \in R^{n \times n}$ 代表图 G 的拓扑结构,如果 $(v_i, v_j) \in E$, 即节点 v_i 和节点 v_j 之间有边,则 $a_{ij}=1$, 否则 $a_{ij}=0$ 。

社团发现: 社团发现是将网络 $G=(V, E)$ 中的 n 个节点划分为 z 个社团, $C=\{C_1, C_2, \dots, C_z\}$, 其中社团内的节点连接紧密, 不同社团之间的节点连接稀疏。

网络表示学习: 给定一个网络 $G=(V, E)$, 网络表示学习的目标是学习一个低维、稠密的矩阵 $U \in R^{n \times m}$:

$$f(A) \rightarrow U \quad (1)$$

其中, $f(\cdot)$ 为将原始网络映射到低维表示的变换函数, m 为低维节点向量的维度, $m \ll n, d$ 。最终学习到的网络表示矩阵 $U \in R^{n \times m}$ 期望最大程度保留网络的底层结构。在嵌入空间中, 拓扑距离相近的节点应该彼此靠近。

3.2 构建社团结构矩阵

根据概率生成模型, 节点 v_i 和 v_j 之间有边相连, 这是可观察到的网络结构。有边相连的节点 v_i 和 v_j 大概率属于同一个社团, 这是未观察到的网络隐藏结构^[11], 因为节点之间边的形成很大可能是因为它们属于同一个社团所引起的。换言之, 两个节点之间边的产生一定程度上反映了它们属于一个社团的可能性。

假设网络是无向的对称图, 包含 n 个节点和 z 个社团。 π_c 表示社团 c 的概率, 网络中每个节点都有一定的概率属于一个社团, $\beta_{c,i}$ 表示社团 c 包含节点 i 的概率, $\beta_{c,i}$ 的值越大代表节点 v_i 在社团 c 中起的作用越大, 且有 $\sum_{i=1}^n \beta_{c,i} = 1$ 。节点 v_i 和 v_j 之间边的产生由以下的有限混合模型生成: (1) 以 π_c 的概率选择一个社团 c ; (2) 社团 c 以 $\beta_{c,i}$ 的概率选择节点 v_i ; (3) 同时社团 c 以 $\beta_{c,j}$ 的概率选择节点 v_j 。假设上述(2)和(3)是独立的, 那么节点 v_i 和 v_j 之间存在边的概率为:

$$\Pr(e_{ij} | \pi, \beta) = \sum_{c=1}^z \pi_c \beta_{c,i} \beta_{c,j} \quad (2)$$

从上述模型可以得出结论: 节点之间产生边的概率越大, 它们属于同一个社团的概率也越大。因为 $\beta_{c,i}$ 和 $\beta_{c,j}$ 的数值增大, $\Pr(e_{ij} | \pi, \beta)$ 的值才会增大。

综上所述, 节点之间边的形成受网络中潜在在社团结构的影响。如果两个节点之间有边相连, 则它们很可能属于同一个社团。因此, 可以通过最大化两个节点之间存在边的概率挖掘网络中潜在的社团结构。基于此, 本节根据节点潜在的社团成员相似性来量化节点的结构邻近度, 从而构造社团结构矩阵 P 。首先设计函数 R 来衡量社团成员之间的相似度; 然后基于相似度度量, 采用基于负采样的 Skip-gram 模型来进一步探究网络底层的社团结构; 最后得到能够捕捉网络潜在在社团结构的矩阵 P 。

首先, 引入社团关系指示矩阵 $H \in R^{n \times z}$, 矩阵 H 的每一行 h_i 代表对应节点隶属于每个社团的隶属度, $h_i h_j^T$ 表示节点 v_i 和 v_j 之间存在边的概率, 并且 $h_i h_j^T \geq 0$ 。因此, 本文设计如下节点相似度函数 R 来衡量两个节点之间属于同一个社团的相似度:

$$R(i, j) = 2\sigma(h_i h_j^T) - 1 = 2 \times \left(\frac{1}{1 + e^{-h_i h_j^T}} \right) - 1 \quad (3)$$

其中, $\sigma(\cdot)$ 为 sigmoid 函数, 这样 $R(i, j) \in [0, 1)$ 。因为函数 $R(i, j)$ 与 $\sigma(h_i h_j^T)$ 呈线性关系, 所以主要讨论 $\sigma(h_i h_j^T)$ 即可。

根据概率生成模型可知, 两个节点之间存在边的概率越大, 即 $R(i, j)$ 越大, 那么它们属于一个社团的概率就大。因此, 对于网络中有边相连的两个节点 v_i 和 v_j , 本文通过最大化 $\sigma(h_i h_j^T)$ 来捕捉网络潜在的社团结构。同时, 对于网络中随

机选择的两个节点, 最小化 $\sigma(h_i h_j^T)$ 。这是因为网络通常是稀疏的, 大多数的节点之间都没有连接, 因此对于网络中随机选择的两个节点, 它们之间有边的概率较低, 隶属于一个社团的概率也较低。基于此, 本文采用基于负采样的 Skip-gram 模型, 对于任意两个节点 v_i 和 v_j , 有:

$$p(i, j) = a_{ij} (\log \sigma(h_i h_j^T) + \kappa \mathbb{E}_{j_N \sim P_V} [\log \sigma(-h_i h_j^T)]) \quad (4)$$

其中, κ 为负采样的个数。本文根据节点的度来选择负样本, 网络中随机采样的节点样本服从 $P_V(i) = d_i / D$ 。其中, d_i 是节点 v_i 的度, $d_i = \sum_j a_{ij}$, $D = \sum_i d_i$ 是网络中所有节点度的和。式(4)被重写为:

$$p(i, j) = a_{ij} (\log \sigma(h_i h_j^T) + \kappa \mathbb{E}_{j_N \sim P_V} [\log \sigma(-h_i h_j^T)]) \quad (5)$$

接下来, 本文通过对 $h_i h_j^T$ 求偏导来优化式(5):

$$\frac{\partial p(i, j)}{\partial (h_i h_j^T)} = a_{ij} \sigma(-h_i h_j^T) - \kappa \frac{d_i d_j}{D} \sigma(h_i h_j^T) \quad (6)$$

由此, 我们得到 $h_i h_j^T$ 为:

$$h_i h_j^T = \log \frac{a_{ij} D}{d_i d_j} - \log \kappa \quad (7)$$

由上述公式得到的 $h_i h_j^T$ 可能为负值, 本文将负值变为 0。综上所述, 本文构建保存网络潜在在社团信息的矩阵 $P \in R^{n \times n}$, 矩阵的元素 p_{ij} 为:

$$p_{ij} = \max\{h_i h_j^T, 0\} = \max\left\{\log \frac{a_{ij} D}{d_i d_j} - \log \kappa, 0\right\} \quad (8)$$

矩阵 P 中的元素为节点之间受社团结构影响边之间的权重, 量化了节点之间结构的邻近度, 反映了网络潜在的社团结构。

3.3 深度节点表示学习

NECD 算法将上节构建的潜在在社团结构矩阵 P 作为深度自编码器的输入, 通过重构矩阵 P 在低维网络嵌入中捕获网络的社团结构, 从而确保属于同一个社团的节点在嵌入空间中彼此接近。 P 矩阵的每一行为深度自编码器的输入, 损失函数如下:

$$L_p = \sum_{i=1}^n \|\hat{p}_i - p_i\|_2^2 = \|\hat{P} - P\|_F^2 \quad (9)$$

通过训练自动编码器使重构误差最小, 可以保持嵌入空间中输入向量之间的相似性。最小化输入和输出的损失函数, 能够在隐层中最大程度保留输入数据的特征, 即网络的潜在在社团结构。虽然节点之间的结构相似性并不被显式地捕获, 但是, 由于所有节点共享自编码器的参数, 编码器期望将那些具有类似输入的节点映射到类似节点表示中, 从而隐式地保留了相似性。因此, 隐层最后一层输出的节点表示最大程度地保存了输入社团结构矩阵 P 的特征, 将其应用于后续的社团发现算法能够得到清晰准确的社团结构。

为了防止过拟合, 本文在目标函数中引入正则化项:

$$L = L_p + \gamma L_{\text{reg}} \quad (10)$$

其中, γ 为平衡过拟合的权重参数, 正则化项 L_{reg} 为:

$$L_{\text{reg}} = \frac{1}{2} \sum_{k=1}^K (\|W^{(k)}\|_F^2 + \|\hat{W}^{(k)}\|_F^2) \quad (11)$$

其中, $W^{(k)}$ 和 $\hat{W}^{(k)}$ 为第 k 层编码器和解码器的权重矩阵, $k = 1, 2, \dots, K$ 。

目标函数(10)可以通过随机梯度下降法和误差反向传播算法进行求解。基于网络表示学习的深度社团发现算法 NECD 的具体流程如算法 1 所示。

算法 1 基于网络表示学习的深度社团发现算法

输入:邻接矩阵 A ,社团个数 z ,参数 κ ,节点表示维度 m

输出:社团结构 $C = \{C_1, C_2, \dots, C_z\}$

1. 根据式(8)构建潜在社团矩阵 P
2. Repeat
3. 根据公式 $L = \sum_{i=1}^n \|p_i - p_i\|_2^2 = \|P - P\|_F^2$ 训练自编码器得到节点表示 $Y^{(K)}$
4. Until converge
5. 对网络低维表示矩阵 $Y^{(K)}$ 运行 K-means 聚类策略得到社团结构 $C = \{C_1, C_2, \dots, C_z\}$

4 实验

在实验中,本文将 NECD 与 4 个对比算法在 9 个真实数据集和 5 个人工数据集上进行大量对比实验。在 4 个基准算法中,2 个为传统社团发现算法(谱聚类算法 Spectral^[12]、非负矩阵分解算法 NMF^[13]),1 个是基于自编码器的深度社团发现算法 DNR^[4],还有 1 个是基于表示学习算法 Deepwalk+K-means^[14]。另外,实验采用标准化互信息 NMI 和纯度 Purity 作为衡量社团发现算法的性能指标。

4.1 在真实数据集上的实验

如表 1 所列,本文选取 9 个真实网络来验证 NECD 算法的有效性和准确度。

表 1 真实数据集的统计信息

Table 1 Statistics of the real-world datasets

Dataset	# Node	# Edge	# Cluster
Zachary karate club	34	78	2
Dolphins	62	159	2
School friendship 6	68	220	6
School friendship 7	68	220	7
Polbooks	105	441	4
Football	115	613	12
Polblogs	1490	16718	2
UAI2010	3363	45006	19
PubMed	19717	44338	3

Zachary karate club:Zachary karate club 数据集是 20 世纪 70 年代美国一所大学的空手道俱乐部的 34 名成员建立的一个社交网络。因为管理者和教练之间存在矛盾,34 名成员分为 2 个社团。

Dolphins:Dolphins 网络是对 62 只宽吻海豚进行观察所构建的海豚社交网络。每个节点代表每个海豚,如果两只海豚频繁在一起嬉戏,则在两个节点之间添加相应的边,由此构成了 Dolphins 社团网络。

School friendship 6 和 School friendship 7:School friendship 6 和 School friendship 7 网络是根据高中同学之间的朋友关系所构建的高校社交网络。网络中的节点为每个学生,根据每个人的自我陈述“谁是他的朋友”,在两个人之间添加边,网络将 7 到 12 年级共分为 6 个社团。其中,9 年级因为既有黑人学生又有白人学生,又分为两个子社团。

Polbooks:Polbooks 网络是 2004 年美国大选期间根据读者在亚马逊网站上购买政治类书籍所构建的网络。亚马逊网站上被购买的书籍为网络中的节点,如果有读者同时购买两本书,则在这两本书之间添加边。网络中的节点被划为 3 个社团。

Football:Football 网络是 2000 年 115 只高校足球队参加

美国高校足球联盟所构成的网络,节点是每支球队,节点之间的边代表两个高校足球队进行过对战比赛。

Polblogs:Polblogs 网络是 2004 年美国大选期间根据博客上博客之间相互转发关系所构成的网络。每篇博客为网络中的节点,如果一篇博客中有直接到达另一篇博客的链接,那么就在两个博客之间添加边。网络中的所有博客被划分为 2 个社团。

UAI2010:UAI2010 是维基百科数据集,由维基百科 2009 年 10 月出现在特色列表中的文章组成,其中包含 3067 篇文章和 45006 个链接,所有文章共分为 19 类。

PubMed:PubMed 是 19717 种科学出版物和 44338 条链接组成的引文数据集,所有科学出版物共分为 3 类。

本节将 NECD 算法与 4 个基线算法在 9 个真实数据集上进行实验。表 2 和表 3 分别展示了不同社团发现方法在真实数据集的 NMI 和 Purity 实验结果。对于每个实验,本文重复 20 次并取平均值作为最终结果。

表 2 真实数据集上社团发现的 NMI 结果

Table 2 NMI of community detection on real-world datasets

Dataset	Spectral	NMF	Deepwalk	DNR	NECD
Zachary karate club	0.7265	0.8972	0.8435	1	1
Dolphins	0.8552	0.8154	0.8487	0.8957	1
School friendship 6	0.4056	0.6489	0.6897	0.7951	0.8858
School friendship 7	0.4539	0.6741	0.6978	0.8195	0.9177
Polbooks	0.4897	0.4583	0.4724	0.4936	0.6528
Football	0.3265	0.7645	0.8514	0.8765	0.9379
Polblogs	0.4998	0.5014	0.4789	0.5457	0.7812
UAI2010	0.1578	0.1689	0.2087	0.2365	0.2954
PubMed	0.0987	0.1382	0.1987	0.2042	0.2535

如表 2 和 3 所列,NECD 算法在所有真实数据集上都取得了最佳的社团发现性能。例如,在 Polblogs 数据集上,相比于表现次优的 DNR 算法,NECD 在 NMI 和 Purity 指标上分别提高了 43.2%和 35%;与传统的 Spectral 方法相比,NECD 分别提高了 56.3%和 54.1%。相比于基于网络表示学习的方法 Deepwalk,NECD 算法在 School friendship 6, School friendship 7 和 Polblogs 数据集上的 NMI 指标分别提高了 28.4%,31.5%和 63.1%。Deepwalk 虽然也是基于网络表示的方法,但在嵌入空间中只捕捉了网络的微观结构,没有保存面向社团结构的信息,因此对于后续社团发现任务是次优的。在规模较大的 PubMed 数据集上,NECD 相比于 NMF 算法在 NMI 和 Purity 指标分别上提高了 83.4%和 35.5%,凸显出 NECD 在大规模网络中相比于传统基于拓扑算法的优越性。综上所述,本文提出的 NECD 算法在真实数据集上获得了比较准确的社团划分结果,相比于其他社团发现算法更具有竞争力。

表 3 真实数据集上社团发现的 Purity 结果

Table 3 Purity of community detection on real-world datasets

Dataset	Spectral	NMF	Deepwalk	DNR	NECD
Zachary karate club	0.8136	0.9012	0.8765	1	1
Dolphins	0.8735	0.8624	0.8868	0.9895	1
School friendship 6	0.5124	0.7324	0.7245	0.8254	0.9235
School friendship 7	0.5984	0.7565	0.7154	0.8421	0.9355
Polbooks	0.6214	0.5035	0.4895	0.5235	0.7335
Football	0.4687	0.8125	0.8754	0.9021	0.9562
Polblogs	0.5345	0.5243	0.5687	0.6098	0.8235
UAI2010	0.2151	0.2368	0.3054	0.3865	0.4525
PubMed	0.3158	0.4742	0.5168	0.5714	0.6424

4.2 在人工数据集上的实验

本节采用 Lancichinetti 等^[15]提出的 Lancichinetti-Fortunato-Radicchi(LFR)网络基准生成的幂律网络来评估算法的有效性。为了验证算法在大规模数据集上的有效性,本节通过 LFR 网络基准模型生成 5 个不同规模的人工数据集,如表 4 所列,从 LNetwork1 到 LNetwork5,网络规模呈递增趋势。在 5 个人工数据集上的社团发现结果如图 2 和图 3 所示。

表 4 LFR 人工数据集的统计信息

Dataset	# Node	# Edge	# Cluster
LNetwork1	500	864	12
LNetwork2	1 000	4734	21
LNetwork3	3 000	10 087	34
LNetwork4	5 000	30 045	42
LNetwork5	10 000	251 227	76

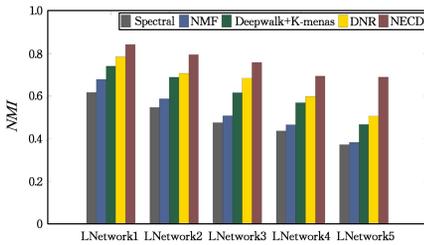


图 2 在 LFR 人工数据集上社团发现的 NMI 结果

Fig. 2 NMI of community detection on LFR synthetic datasets

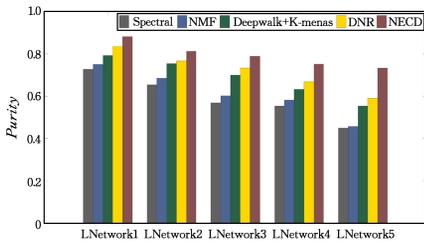


图 3 在 LFR 人工数据集上社团发现的 Purity 结果

Fig. 3 Purity of community detection on LFR synthetic datasets

从图 2 和图 3 中可以看出,NECD 算法在 5 个人工数据集上都取得了最佳的社团发现性能,特别是在规模较大的数据集 LNetwork4 和 LNetwork5 中,NECD 算法的性能明显优于传统的 Spectral 算法和 NMF 算法。在人工数据集 LNetwork4 上,NECD 算法相比于 Spectral 算法在 NMI 和 Purity 指标上分别提升了 58.9% 和 35.6%;在人工数据集 LNetwork5 上分别提升了 85.6% 和 62.7%。相比于 NMF 算法,NECD 算法在人工数据集 LNetwork4 上的 NMI 指标提升了 49.1%,Purity 指标提升了 58.9%;在人工数据集 LNetwork5 上的 NMI 和 Purity 指标分别提升了 79.6% 和 59.9%。这证明了在较大规模数据集上,本文提出的基于网络表示的深度模型相对于传统的社团发现获得了更准确的社团结构划分。这是因为 NECD 算法利用深度神经网络将网络映射到低维空间,然后在新的空间进行社团发现,能更好地捕捉网络的非线性结构,拥有较强的特征学习能力,从而得到优异的社团结构划分结果。与同样是基于网络表示的方法 Deepwalk 相比,NECD 的社团发现性能也更加优异。这是因为 Deepwalk 算法虽然将网络中节点映射到低维空间,然后运用聚类算法得到社团结构,但是低维节点表示只保存了网络的二阶和高阶邻近度,而 NECD 在节点表示中充分捕捉了网络潜在的社团信息,有助于提高后

续社团发现任务的准确度。

4.3 参数分析

本节实验对 NECD 算法的参数进行分析和讨论。首先对自编码器深度对算法性能的影响进行讨论,然后对负采样个数 κ 对模型的影响进行分析。

自编码器层数:首先,分别在 Polbooks 和 Football 数据集上实现 2,3 和 4 层的深度自动编码器结构,然后比较不同层数生成的节点表示向量用于社团发现的有效性。如图 4 所示,3 层结构的深度自编码器比 2 层结构的深度自编码器生成的低维节点表示获得了更好的社团发现性能,这是因为深层的神经网络结构可以捕捉更准确的网络社团结构,抽取更深层的网络潜在结构。在 Football 数据集中,4 层结构的自编码器社团发现结果相对于 3 层的自编码器社团发现结果有略微的下降,这可能是由于随着自编码器结构的加深,数据中的一些重要信息会丢失,从而导致性能下降。对于不同的数据集,不同深度的自编码器保留不同程度的信息,从而影响最终的社团划分结果。因此,针对不同的数据集,本文设置不同的深度自编码器结构。

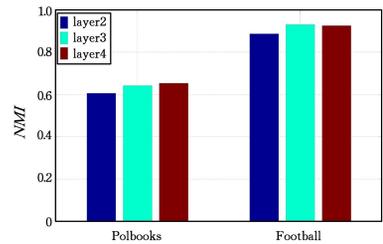


图 4 不同自编码器层数的 NMI 结果

Fig. 4 NMI of community detection for different autoencoder layers

负采样个数:本节在 Football 数据集上通过变化 $\kappa = [1, 13]$ 来观察社团发现的结果,从而对负采样的个数 κ 对社团发现结果的影响进行讨论。如图 5 所示,在 $\kappa \leq 9$ 时,算法获得了优异的社团发现性能,并且表现稳定,没有剧烈的波动。对于不同规模的数据集, κ 的不同取值将直接影响后续社团发现的结果。在上述实验中,通过在一定范围内变换 κ 的取值,采用最大的 NMI 和 Purity 值作为实验结果。

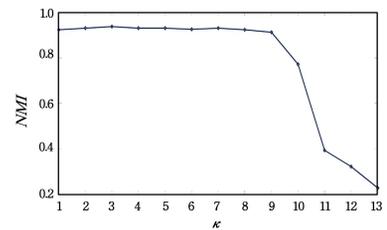


图 5 NECD 在不同负采样下的社团发现性能

Fig. 5 Community detection performance of NECD over different number of negative samples

结束语 该文将网络表示学习和社团发现两个领域相结合,提出了一种基于网络表示的深度社团发现算法 NECD。算法首先构造了保存网络潜在社团结构的矩阵作为深度自动编码器的输入,通过重构潜在社团矩阵获得低维节点表示,然后将 K-means 聚类策略作用于节点低维表示向量获得网络社团结构。相比于传统的社团发现算法,本文提出的 NECD 算法利用深度神经网络成功捕捉了网络的非线性结构,学习了更加准确和丰富的网络社团结构,为后续的社团发现打下坚实的基础。在多个不同规模的真实数据集和人工数据集上进

行了大量实验,实验表明 NECD 算法与其他社团发现算法相比取得了更好的性能,获得了更准确的社团结构。

更进一步,在接下来的工作中可以利用丰富的先验信息将 NECD 扩展为半监督的社团发现算法,利用监督信息获得更加准确的社团结构。

参 考 文 献

- [1] WANG S H. Community Detection in Signed Networks with Game Theory[J]. *Computer Science*, 2020, 47(S2): 459-463.
- [2] ZHANG D K, YIN J, ZHU X Q, et al. Network Representation Learning: A Survey[J]. *IEEE Transactions on Big Data*, 2017, 6(1): 3-28.
- [3] TIAN F, GAO B, CUI Q, et al. Learning deep representations for graph clustering[C]// *The Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014: 1293-1299.
- [4] YANG L, CAO X, HE D, et al. Modularity based community detection with deep learning[C]// *International Joint Conference on Artificial Intelligence*. 2016.
- [5] JIN D, GE M, LI Z, et al. Using Deep Learning for Community Discovery in Social Networks[C]// *2017 IEEE 29th International Conference on Tools with Artificial Intelligence*. 2017.
- [6] HU P, NIU Z, HE T, et al. Learning Deep Representations in Large Integrated Network for Graph Clustering [C] // *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering*. 2018: 101-105.
- [7] WU L, ZHANG Q, CHEN C H, et al. Deep Learning Techniques for Community Detection in Social Networks[J]. *IEEE Access*, 2020(8): 96016-96026.
- [8] CAO J, JIN D, YANG L, et al. Incorporating network structure with node contents for community detection on large networks using deep learning[J]. *Neurocomputing*, 2018, 297: 71-81.
- [9] CAO J, JIN D, DANG J. Autoencoder Based Community Detection with Adaptive Integration of Network Topology and Node Contents[C]// *International Conference on Knowledge Science, Engineering and Management*. Cham: Springer, 2018.

- [10] KRLJ B, KRALJ J, LAVRA N. Embedding-based Silhouette Community Detection[J]. *Machine Learning*, 2020, 109(1): 161-219.
- [11] REN W, YAN G, LIAO X, et al. Simple probabilistic algorithm for detecting community structure[J]. *Physical Review E*, 2009, 79(2): 036111.
- [12] NEWMAN M E J. Modularity and community structure in networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, 103(23): 8577-8582.
- [13] WANG R S, ZHANG S, WANG Y, et al. Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures[J]. *Neurocomputing*, 2008, 72(1/2/3): 134-141.
- [14] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations[C]// *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014: 701-710.
- [15] LANCICHINETTI A, FOURTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 2009, 80(2): 016118.



PAN Yu, born in 1990, doctor. Her main research interests include data processing and mining in social networks and machine learning.



PAN Zhi-song, born in 1973. Ph.D, professor, Ph.D supervisor. His main research interests include computer vision and machine learning.

(上接第 169 页)

- [5] YAN Y, WANG J X. Cylindrical fitting, method of laser scanner point cloud data[J]. *Science of Surveying and Mapping*, 2018, 43(6): 83-87.
- [6] NURUNNABI A, SADAHIRO Y, LINDENBERGH R. Robust cylinder fitting in three-dimensional point cloud data[J]. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2017, 17: 63-70.
- [7] WANG C C, WANG X W, XU X C. Study on the cylindrical surface fitting method[J]. *Engineering of Surveying and Mapping*, 2014, 23(3): 5-9.
- [8] WANG J X. A method for fitting of conicoid in industrial measurement[J]. *Geomatics and Information Science of Wuhan University*, 2007, 32(1): 47-50.
- [9] LIU Y P, ZHANG D H. Study on evaluation of cylinder geometric parameters in reverse engineering[J]. *Mechanical Science and Technology*, 2005, 24(3): 310-311.
- [10] LIU Z L G, LI M F, LU J H. Cylindrical fitting of point cloud data based on robust weighted total least squares[J]. *Modern Surveying and Mapping*, 2018, 41(4): 39-42.

- [11] ZHANG S J, LIU C J, LI J F, et al. Cylinder fitting with roundness estimate method based on projection[J]. *Journal of Geomatics Science and Technology*, 2014, 31(4): 39-42.
- [12] BAO J, WANG J X. Cylindrical fitting with improved roundness estimate method based on projection[J]. *Mine Surveying*, 2018, 46(1): 94-97.
- [13] QIN S W, PAN G R, GU C, et al. Fitting of spatial cylindrical surface based on genetic algorithm[J]. *Journal of Tongji University (Natural Science)*, 2010, 38(4): 604-607.
- [14] SHI F, WANG H, YU L, et al. 30 Examples of MATLAB Intelligence Algorithm[M]. Beijing: Beihang University Press, 2011.



GAO Shuai, born in 1988, postgraduate, engineer. His main research interests include precision measurement and mechanical design.