文章编号:1672-3961(2025)01-0077-09 DOI:10.6040/j.issn.1672-3961.0.2023.266

基于节点重要性排序的局部社区检测算法

武凯丽,陈京荣*
(兰州交通大学数理学院, 甘肃 兰州 730070)

摘要:针对目前应用广泛的社区检测算法存在时间复杂性过高、精度低、结果不稳定等缺点,提出一种基于节点重要性排序的局部社区检测算法(local community detection algorithm based on the node importance ranking, LCDIR)。根据节点重要性顺序选择核心节点,通过节点强度和网络拓扑结构特征对网络进行社区检测形成初步社区,利用内外边比例和模块化度量最大化合并弱小社区,形成最终的社区。在真实网络和人工合成网络上和 7 种社区检测算法进行对比试验,结果表明,该算法在这些网络上形成了较高质量的社区,解决现有局部社区检测算法存在核心节点选择不当的问题,具有较高模块化度量值和标准化互信息值,相较于其他社区检测算法更准确有效、性能更好、时间复杂度较低。

关键词:复杂网络;社区检测;节点重要性排序; 核心节点;节点相似性
中图分类号:O157.6;TP391 **文献标志码**:A
引用格式:武凯丽,陈京荣.基于节点重要性排序的局部社区检测算法[J]. 山东大学学报(工学版),2025,55(1):77-85.
WU Kaili, CHEN Jingrong. Local community detection algorithm based on the node importance ranking[J]. Journal of Shandong University (Engineering Science), 2025, 55(1):77-85.

Local community detection algorithm based on the node importance ranking

WU Kaili, CHEN Jingrong*
(School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou 730070, Gansu, China)

Abstract: The proposed local community detection algorithm based on node importance ranking (LCDIR) aimed to address the limitations of existing community detection algorithms, such as high time complexity, low accuracy, and unstable results. Key nodes were selected in order of their importance, and initial communities were detected based on node strength and network topology features. Weak communities were merged by the ratio of internal to external edges and maximizing modularity measure, resulting in final communities. Experiments were conducted on real and synthetic networks, compared with seven other community detection algorithms, which demonstrated that the proposed algorithm formed high-quality communities on these networks. It resolved the issue of inappropriate key node selection in existing local community detection algorithms. The algorithm achieved high modularity measure and normalized mutual information values, indicating higher accuracy, effectiveness, and better performance compared to other community detection algorithms, while maintaining low time complexity.

Keywords: complex network; community detection; node importance ranking; key node; node similarity

0 引言

现实世界中大多数复杂系统都可以抽象为复杂网络。社区检测在众多领域上具有广泛应用,包括社交媒体分析、社会网络分析、生物信息学、推荐系统等[1]。社区检测是指在网络结构中,将节点分组成具有紧密连接、内部相互关联较强、外部相互关联较弱的社区的过程[2]。通过识别社区结构,能够更好理解网络的组织原理、预测信息传播和行为传播模式[3]。

目前,社区检测领域正处于学术界广泛关注和

收稿日期: 2023-11-06
基金项目:国家自然科学基金资助项目(52362044)
第一作者简介:武凯丽(1999—),女,山西吕梁人,硕士研究生,主要研究方向为复杂网络研究.E-mail:1037697054@qq.com
*通信作者简介:陈京荣(1976—),女,甘肃兰州人,教授,硕士生导师,博士,主要研究方向为网络优化理论与算法设计研究.
E-mail:chenjr@mail.lzjtu.cn

深入研究的阶段。一般来说,社区检测方法分为全局方法和局部方法。全局社区检测方法考虑整个网络信息来识别社区,时间复杂性较高,不适用于大规模复杂网络中的社区检测[4-6]。相较于全局方法而言,局部社区检测方法基于节点结构提取社区,一般选取一组节点作为核心节点,根据相似性或模块化度量,重复添加其邻点来扩展社区,保持有较低的时间复杂度[7-9]。

文献[10]提出一种基于层次结构的社区检测算法,通过检测和去除两个社区之间的连接边来实现社区检测,是一种图聚类算法,时间复杂度为$O(n^3)$[11]。文献[12]提出的 Louvain 方法通过迭代的方式将节点合并到具有更高模块度的社区,直到不能再进行合并为止,是相对接近线性的时间复杂度。文献[13]提出的 Infomap(Information map)方法是一种基于信息论的社区发现算法,它以最小化信息流动为目标,将网络划分为多个模块,使得每个模块内节点具有相似信息,时间和空间复杂度较高。

文献[14]提出了一种 LPA(label propagation algorithm)标签传播算法作为快速启发式本地社区检测算法,具有近似线性的时间复杂度。文献[15]提出了 LPAm(modularity-specialized label propagation algorithm)方法和贪婪多步积分的组合来解决局部最优问题。文献[16]介绍了一种结合两个节点之间的共同邻居的聚类系数来获得这两个节点的相似性的方法。文献[17]提出了一种基于结构相似性度量对社区进行检测,具有近似线性的时间复杂度。文献[18]提出了一种新的 CNM(Clauset-Newma-Moore algorithm)算法,基于节点中心性发现社区。文献[19]提出了一种局部社区检测方法 ECES(expanding core nodes using extended similarity),具有使用本地知识本地提取图社区以及识别不同节点(核心或离群点)的重要性和作用的能力。文献[20]提出一种基于影响力的社区检测算法 D-LPA(degree- label propagation algorithm),提出影响度改进 LPA 标签传播算法的标签迭代顺序。

通过对相关研究工作进行学习分析,发现全局社区检测算法具有较高精度,但存在时间和空间复杂性较高等问题。局部算法不需要对网络有完全了解,仅利用节点局部信息来识别社区,陷入局部最优在局部算法中是常见的。在局部社区检测算法中考虑全局特征,可能在一定程度上可以弥补局部最优情况。

本研究提出一种基于节点重要性排序的局部社区检测算法。通过节点重要性确定核心节点,利用节点相似性确定核心节点的社区;下一步将其邻点添加入社区中,重复操作,对社区进行拓展,直至所有节点都成为其中一个社区的成员;通过内外边比例确定需要合并的弱小社区,利用模块化度量最大化合并弱小社区,形成最终的社区。该算法在对节点进行重要性排序时,考虑了全局结构,在一定程度上弥补了局部检测算法可能出现的局部最优情况,保持有较低的时间复杂度。

1 理论基础

$G=(V,E)$是一个无权无向图,节点集表示为$V(G)=\{v_1,v_2,v_3,\cdots,v_n\}$,边集为$E(G)=\{e_1,e_2,e_3,\cdots,e_n\}$,节点的数目为$|v|=n$,边的数目为$|E|=m$。

1.1 索伦森相似性指数

索伦森相似性指数(sorensen index)[21],也称为索伦森-达伊斯指数(Sorensen-Dice index),是一种度量两个集合重叠程度的相似性指标,在网络中是通过将两组邻点的交集大小的两倍除以其邻点总数来计算的,节点i和节点j的索伦森相似性指数表示为

$$S_{\text{Sorensen}}(i,j)=\frac{2*|N_i\cap N_j|}{k_i+k_j},$$

式中:N_i为节点i的邻点集合;$N_i\cap N_j$表示节点i和节点j的共同邻点集合;k_i为节点i的度,即邻点的数量。

索伦森相似性指数为0~1,值越接近1表示两个集合的重叠程度越高,相似性越大。在社区检测中,可将每个节点的所有邻点看作一个集合,计算两个节点的索伦森相似性指数,表示两个节点之间的相似性,进一步检测社区结构[22]。

1.2 枢纽抑郁指数

枢纽抑郁指数(hub depressed index, HDI)[23]是一种特定的指标,用于度量网络中节点之间的相似性,用于衡量网络中高度节点之间的相似性。节点i和节点j的枢纽抑郁指数表示为:

$$S_{\text{HPI}}(i,j)=\frac{2*|N_i\cap N_j|}{\max(k_i,k_j)},\tag{1}$$

式中,N_i为节点i的邻点集合,k_i为节点i的度,即邻点数量。

枢纽抑郁指数为0~1,值越接近1表示两个集合之间的连接性程度越高,相似性越大。在对社区进行检测时,利用枢纽抑郁指数,可以更好让位于两个社区边界节点不会被高度节点所吸引,将其放

置在一个与同一节点具有更多结构相似性社区中[24]。

1.3 模块化度量

模块化度量[25]是一种用于衡量网络社区结构优劣的指标,衡量了网络中节点的模块化程度。最流行的算法 Newman′s modularity 的度量方式由文献[26]提出,通过比较实际内部连接和预期内部连接的差异来衡量社区结构的好坏[11],即比较了网络中节点之间实际连接的数量与随机网络中相同节点度分布的期望连接数量之间差异,表达为

$$Q=\frac{1}{2m}\sum_{ij}(A_{ij}-P_{ij})\delta(C_i,C_j), \tag{2}$$

式中:A_{ij}表示节点 i 和节点 j 之间是否存在连接,存在为 1,不存在为 0;P_{ij}是一个期望连接的概率,通常可以表示为随机网络中两个节点之间连接的概率。$\delta(C_i,C_j)$是一个指示函数,当节点 i 和节点 j 属于相同的社区时为 1,否则为 0;C_i 表示节点 i 所属的社区,C_j 表示节点 j 所属的社区。

模块化度量用于衡量一个社区划分结果的质量,即衡量社区内部紧密连接和社区之间松散连接的程度。从式(2)中可以看出如果整个图被视为单个社区,则获得的模块度为零,模块度值越高,表示社区结构越好,越能反映网络内部的紧密连接。一般来说,模块化度量值在-1~1,越接近 1 表示社区结构越好。模块度值可以根据图的大小而增长,不可能针对不同大小的两个图来比较该值。

1.4 标准化互信息

标准化互信息(normalized mutual information, NMI)是一种用于衡量两个聚类结果之间的相似性的度量方法[27],NMI 值通常用于评估聚类算法的性能,在社区检测算法中,可以比较提取的社区结果和真实社区结果之间的相似性,用于评估不同的社区检测算法。假设提取的社区 $C=\{C_1,C_2,C_3,\cdots,C_q\}$ 和真实网络社区 $C'=\{C'_1,C'_2,C'_3,\cdots,C'_k\}$,NMI 值为

$$N_{\mathrm{NMI}}(C,C')=\frac{2I(C,C')}{H(C)+H(C')}, \tag{3}$$

式中,$I(C,C')$表示 C 和 C'的互信息,计算了 C 和 C'之间的共享信息,$H(C)$表示 C 的熵(即 C 的不确定性),

$$I(C,C')=H(C)+H(C')-H(C,C'),$$

$$H(C)=-\sum_{i=1}^{q}\frac{|C_i|}{n}\ln\frac{|C_i|}{n},$$

$H(C,C')$称为联合熵(即 C 和 C'的联合不确定性),表示为

$$H(C,C')=-\sum_{i=1}^{q}\sum_{j=1}^{k}\frac{|C_i\cap C'_j|}{n}\ln\frac{|C_i\cap C'_j|}{n}。 \tag{4}$$

NMI 值在 0~1,其中 0 表示两个结果是独立的,而 1 表示两个结果完全相同。在使用 NMI 评估社区检测结果的有效性,较高的 NMI 值表示预测的社区结构与真实社区结构之间的相似性更高,较低的 NMI 值则表示相似性较低。

2 算法设计及分析

针对目前应用广泛的社区检测算法存在的精度低、时间复杂度过高、结果不稳定等缺点,提出了一种基于节点重要性排序的局部社区检测算法 LCDIR。

2.1 基于节点重要性排序的局部社区检测算法设计

LCDIR 算法分为以下几个步骤:

步骤 1:根据网络中节点重要性综合得分对节点进行排序。

节点重要性综合得分公式构造为

$$K_{\mathrm{IMDD}}(i)=Km(i)\times\Big(\sum_{j\in N_{j1}}d_jKm(j)+\sum_{j\in N_{j2}}\mu_jd_jKm(j)\Big), \tag{5}$$

式中:$Km(i)$是利用混合度分解法[28]计算得到的节点 i 的混合 K-shell(Km)值,N_{j1}和 N_{j2}分别是节点 j 的邻点集合和次邻点集合,μ_i 是节点 i 的可动态调整的次邻点影响系数

$$\mu_i=\frac{D_i}{k_i+D_i}, \tag{6}$$

式中:k_i 是节点 i 的度,同时也是节点 i 的邻点数量,D_i 是节点的次邻点数量。

d_j 是节点 i 的度占邻点集合或次邻点集合中所有节点度总和的比例,即

$$d_i=\frac{k_i}{\sum k_j},\ (j\in N,\ N\subset N_{j1}\cap N_{j2})。 \tag{7}$$

节点重要性综合得分公式通过加权的方式将节点的邻点和次邻点对整个网络的影响容纳进去,充分反映了节点全局特性和局部特性。对于网络中每个节点,相邻节点与其存在直接联系,除相邻节点以外节点也与其存在间接联系,为了减少计算代价,同时考虑网络的局部特性,只考虑邻点和次邻点对网络中节点的影响。指定节点的次邻点与节点本身之间的联系是间接的,通过添加影响系数的方法将次邻点对指定节点的影响量化。

步骤 2:选取节点重要性排序前 30%节点作为

核心节点,根据索伦森相似性指数检测出核心节点初始社区。

从重要性值最高节点开始检测,如果节点只有一个重要邻居,则两个节点都属于一个社区,如果有更多的重要邻居,则根据索伦森相似性指数将最相似邻点放置在同一个社区中,使用索伦森相似性指数可以通过邻点的重叠程度来确定节点之间相似性。重复以上操作,可以得到核心节点初始社区。

步骤 3:对初始社区进行逐步拓展,形成初步的社区结构。

将前一步骤中未被选中的节点放置在围绕核心节点形成的初始社区中。从最重要节点开始,如果其邻点属于不同社区,则根据 Hub Depressed 指数的总和将节点分配给最佳社区。使用 Hub Depressed 指数时,如果一个节点位于两个社区边界,它不会被高度节点所吸引,是被放置在一个与同一节点具有更多结构相似性的社区中。继续重复上述操作对社区进行拓展,直至所有节点都成为其中一个社区成员,形成初步的社区结构。

步骤 4:对步骤 3 中形成的初步社区结构进行合并来提高社区质量。

在初步社区中,有一些社区较为稀疏,为了对社区进行最佳合并,通过对社区内边数和外边数符合下式的社区视为需要被合并社区:

$$\alpha * E_C^{\text{in}} \geqslant E_C^{\text{out}}, \tag{8}$$

式中 ,E_C^{in} 表示社区内边数,E_C^{out} 表示外边数,α 表示合并社区参数,在多个标准数据集上的试验结果表明,α 的最佳值为 1.25。

通过计算分别合并的社区模块化度量值,将需要被合并社区合并到附近的社区中,选择模块度量值最大的合并方式实现合并后社区模块度最大化。

2.2 实例分析

本节以一个包含 60 个节点和 193 条边的 Facebook 社交网络数据集为例,将 LCDIR 方法在该数据集上的社区检测步骤详细叙述。

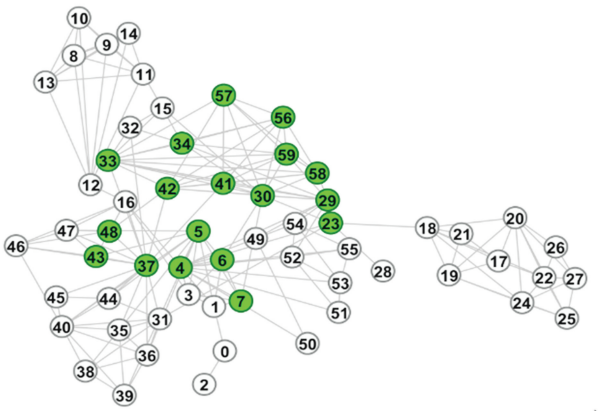
表 1 中给出网络中节点重要性排序结果。

对表 1 结果取排名前 30%的节点,得到核心节点为节点 33、29、37、42、41、4、30、58、59、23、48、5、57、6、7、34、56、43 这 18 个节点,如图 1(a)所示。图 1(b)中为核心节点,从重要性值最高的节点 33 开始检测, 考虑其邻点 29、30、34、37、41、42、56、57、58、59 这 10 个节点,41 和 59 最高相似性邻点为 33,将 33、41 和 59 放入一个社区内 ,在检测其他邻点时候,42 最高相似性邻点是 41、34 和 58 最高相似性邻点都为 59 节点,检测出来第一个核心节点社区为{33, 34, 41, 42, 58, 59, 30};继续考虑节点 29,它邻点中 30、33、34、41、42、58、59 这 7 个节点的最高相似性邻点均不为 29 且已分配社区,节点 4 虽未分配社区,最高相似性邻点也不为 29,只有节点 23 满足,检测出来的第二个社区为{29, 23};接下来检测节点 37,其邻点 41、42、33 已分配社区,48 最高相似性邻点为 37,43 的最高相似性邻点为 48,第 3 个社区为{48, 37,43};对于节点 4,它的邻点中没有最高相似性邻点是节点 4 的,节点 4 属于一个单独社区;重复上述操作,得到核心社区检测结果:{1:{33, 34, 41, 42, 56, 57, 58, 59, 30}, 2:{29, 23}, 3:{48, 43, 37}, 4:{4, 5}, 5:{6, 7}},如图 1(c)所示。图 1(d)为核心节点在整个网络中的位置。

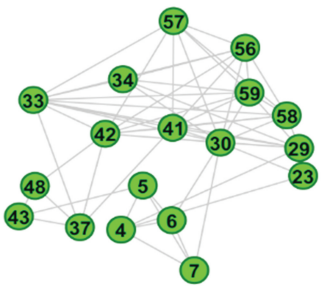
表 1 节点重要性排序结果

Table 1 Node importance ranking results

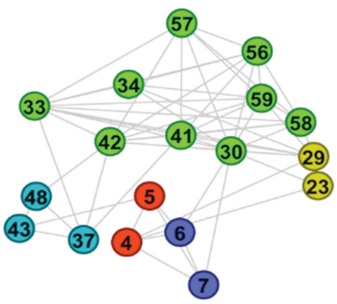
节点	K_{IMDD}	排序	节点	K_{IMDD}	排序
33	21.23	1	15	17.98	31
29	21.12	2	54	17.94	32
37	21.02	3	52	17.77	33
42	20.89	4	45	17.29	34
41	20.88	5	39	17.24	35
4	20.86	6	50	16.64	36
30	20.48	7	55	16.21	37
58	20.09	8	51	16.12	38
59	19.97	9	53	15.92	39
23	19.66	10	38	15.58	40
48	19.63	11	12	15.10	41
5	19.45	12	18	14.90	42
57	19.43	13	11	13.86	43
6	19.43	14	28	13.77	44
7	19.43	15	0	12.51	45
34	19.42	16	10	12.44	46
56	19.22	17	14	12.44	47
43	19.08	18	17	12.30	48
47	19.08	19	24	12.10	49
16	19.05	20	8	12.09	50
40	18.87	21	9	12.09	51
31	18.84	22	13	12.09	52
46	18.83	23	19	12.08	53
35	18.78	24	25	11.93	54
36	18.78	25	22	11.91	55
3	18.68	26	27	11.68	56
32	18.55	27	26	11.5	57
44	18.44	28	20	11.48	58
1	18.25	29	21	11.47	59
49	18.12	30	2	5.55	60



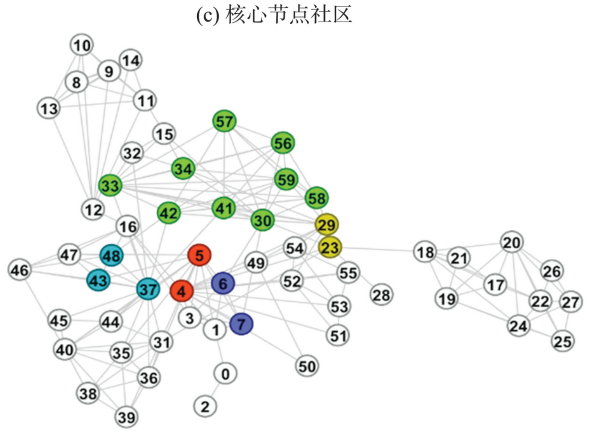
(a) 核心节点在网络中的位置



(b) 核心节点



(c) 核心节点社区

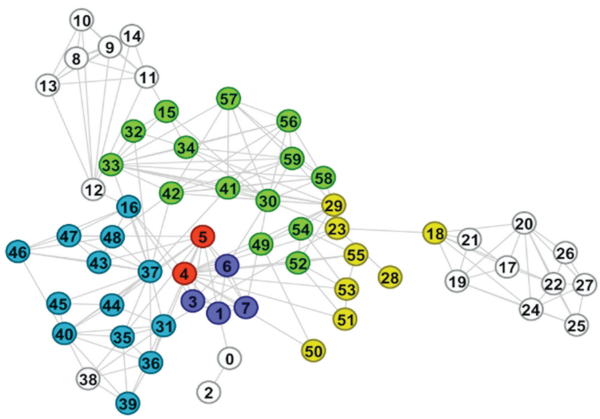


(d) 核心节点社区在网络中的位置

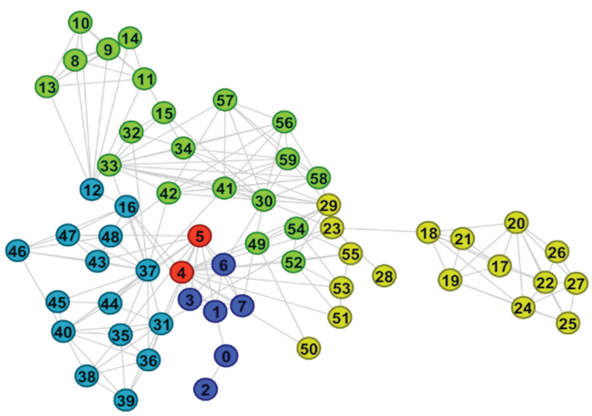
图 1　核心节点的社区检测

Fig.1　Community detection of key nodes

图 2 为在 Facebook 社交网络上由核心节点社区逐步拓展直到完成社区检测结果,由图 1(d)到图 2(a)的过程为,找出核心节点所有邻点,依旧从最高重要性节点 47 开始检测,节点 47 的邻点为 46、16、48、5、43、在这些邻点中,节点 43 和 16 还未分配社区,其余邻点分属两个社区 3:{48, 43, 37}和社区 4:{4, 5},计算社区内节点枢纽抑郁指数的总和,分别是 1.77 和 0.22,所以节点 47 归属社区 3,重复上述操作,可得到图 2(a);继续重复检测邻点对社区进行拓展,直至所有节点都成为其中一个社区的成员;得到的社区检测结果为{1:{8, 9, 10, 11, 13, 14, 15, 30, 32, 33, 34, 41, 42, 49, 52, 54, 56, 57, 58, 59}, 2:{26, 25, 27, 17, 18, 51, 50, 53, 55, 23, 19, 20, 21, 22, 28, 29, 24}, 3:{35, 36, 37, 38, 39, 40, 43, 44, 45, 46, 47, 48, 16, 12, 31}, 4:{4, 5}, 5:{0, 1, 2, 3, 6, 7}},可见图 2(b)。



(a) 第一次拓展邻点后的社区检测结果



(b) 多次拓展邻点后的最终社区检测结果

图 2　逐步执行社区检测的结果

Fig.2　Results of step-by-step implementation of community testing

在步骤 4 中,先计算每个社区的内外边比例,得出社区 4:{4, 5}为符合被合并条件的社区,当节点 4 和节点 5 合并入社区 3 时,整个网络的模块度最大。最终社区检测结果就为{1:{8, 9, 10, 11, 13, 14, 15, 30, 32, 33, 34, 41, 42, 49, 52, 54, 56, 57, 58, 59}; 2:{24, 29, 28, 17, 18, 51, 50, 53,

23, 55, 19, 25, 26, 27, 20, 21, 22}; 3: {35, 36, 37, 38, 39, 40, 4, 5, 43, 44, 45, 46, 47, 48, 16, 12, 31}; 4: {0, 1, 2, 3, 6, 7}},图 3 为最终社区检测可视化结果。

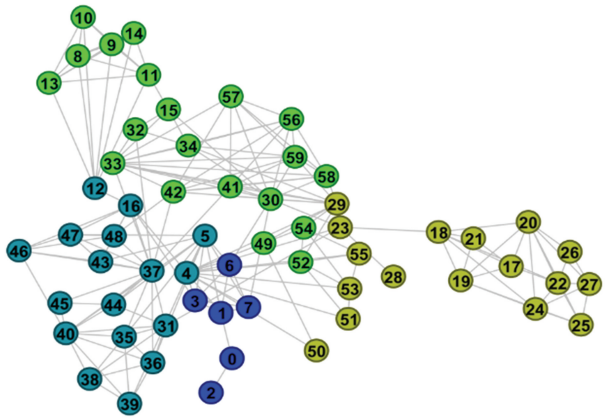


图 3 合并社区后的结果

Fig.3 Results after merging communities

2.3 时间复杂度分析

用 m 表示网络的边数,n 表示网络的节点数,$\langle k\rangle$表示网络中节点的平均度,c 表示步骤 3 中得到的社区个数,LCDIR 算法在对节点进行重要性排序时间复杂度为 $O(m+n\langle k\rangle^2)$,对核心节点进行社区检测时,时间复杂度为 $O((m+n)n)$,对核心社区进行拓展过程时间复杂度为 $O(n\log n+m+n)$,合并社区时间复杂度为 $O(c*n^2+c)$,这个社区的复杂度可以近似为 $O(n\langle k\rangle^2)$。

3 试验及结果分析

为了检验 LCDIR 方法在真实网络中的性能,在 6 个真实网络和人工合成网络上将该算法和 7 种社区检测算法,包括 CNM、InfoMap、LPA、SCAN、louvain,ECES 算法、D-LPA 算法就模块度和标准化互信息两个指标进行对比试验。

3.1 真实网络试验

在试验中使用了这 6 个真实网络,包括:Zachary 的 Karate Club 网络[29]、dolphins 网络、Polbooks 网络、Twitter 网络、Email 网络、SciMet 网络。

上述网络均处理为无权无向网络,网络规模及拓扑性质已列在表 2 中,n 表示网络中节点个数,m 表示网络中边的数目,$\langle k\rangle$表示网络中节点的平均度,$k_{\min}$表示网络中节点的最小度。$k_{\max}$表示网络中节点的最大度。

表 2 真实网络的网络拓扑性质

Table 2 Network topology properties of real network

网络	n	m	$\langle k\rangle$	$k_{\min}$	$k_{\max}$
Karate Club	34	78	4	1	17
Dolphins	62	159	5	1	12
Polbooks	105	441	8	2	25
Twitter	761	1 029	3	1	37
Email	1 133	5 451	9	1	71
SciMet	2 678	10 368	8	1	164

模块化度量是一种用来评估社区检测算法准确性的指标。在表 3 中,对比了 LCDIR 方法和其他 7 种社区检测方法在 6 个真实网络上社区检测结果和模块化度量值。试验结果清晰显示了 LCDIR 算法在大多数网络中高模块化度量值,这意味着它在准确性上优于其他算法。LCDIR 算法在根据节点重要性选择核心节点时候,充分考虑了全局特性和局部特性,通过节点之间相似性对节点进行社区检测,社区合并时实现模块最大化,节省了计算时间,也在一定程度上提高了模块化度量值,为社区检测准确性和效率带来了显著提高。LCDIR 算法在处理真实网络数据时展现出了较好性能,在模块化度量值方面具有良好表现。

表 3 在真实网络上的模块化度量值

Table 3 Modularity metric values on a real network

网络	CNM		InfoMap		LPA		SCAN		Louvain		ECES		D-LPA		LCDIR	
	Q	数量	Q	数量	Q	数量	Q	数量	Q	数量	Q	数量	Q	数量	Q	数量
KarateClub	0.380	3	0.401	4	0.390	4	0.375	3	0.421	4	0.36	3	0.393	2	0.402	2
Dolphins	0.495	4	0.483	5	0.517	4	0.513	8	0.425	4	0.462	5	0.520	3	0.528	2
Polbooks	0.501	10	0.513	5	0.469	2	0.436	12	0.506	5	0.517	2	0.485	3	0.524	4
Twitter	0.526	45	0.445	22	0.438	69	0.501	132	0.494	20	0.469	13	0.452	52	0.496	28
Email	0.492	123	0.482	11	0.459	13	0.453	127	0.493	12	0.434	28	0.508	15	0.512	38
SciMet	0.365	156	0.335	17	0.352	30	0.326	232	0.361	16	0.343	46	0.355	95	0.363	122

另一方面,标准化互信息用于验证社区检测算法的有效性。图 4 中展示了这几种算法在真实网络上进行社区检测标准化互信息值,也就是和实际社区之间的相似性。试验结果表明:LCDIR 算法在 Email 网络,dolphins 网络,Polbooks 网络这 3 个真实网络上性能明显优于其他算法;在 Twitter 网络和 SciMet 网络上模块化度量值稍小于 CNM 算法;在 Karate Club 网络上模块化稍小于 Louvain 算法。

在某些网络中模块化度量值略低于某些检测算法的原因是一些算法就是基于模块化度量值最大化,例如 CNM 算法是基于全局结构的,在这些数据集上具有较低精度。值得注意的是,LCDIR 算法基于节点重要性选择核心节点,用体现重叠程度的索伦森相似性指数对核心节点进行社区提取,用体现连接性程度的枢纽抑郁指数对核心节点的初始社区进行拓展,形成初步社区,对稀疏社区进行合并,这种综合考虑不仅提高了模块化度量值,也有助于提高社区检测算法的准确性和有效性。LCDIR 算法在真实网络上表现良好,准确性和有效性都较高。

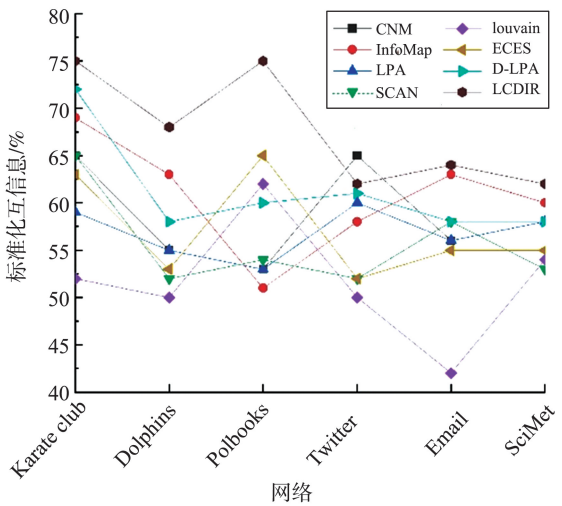


图 4　真实网络上的社区检测 NMI 值

Fig.4　NMI values of Community detection on a real networ

3.2　合成网络试验

为了更准确比较该算法,在本次试验中还采用了 4 个合成 LFR (Lancichinetti-Fortunato-Radicchi) 网络,LFR 网络的节点度和社区大小均具有幂律分布,体现了现实世界网络特征。网络参数在表 4 中给出,γ 表示社区内节点度分布的幂律指数,β 表示社区大小分布的幂律指数。

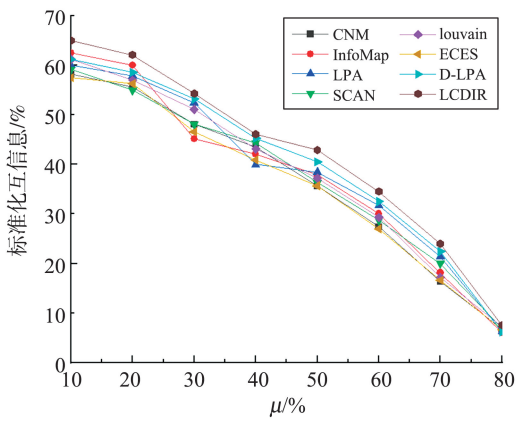
表 4　合成网络的参数设置

Table 4　Parameter settings of the synthesis network

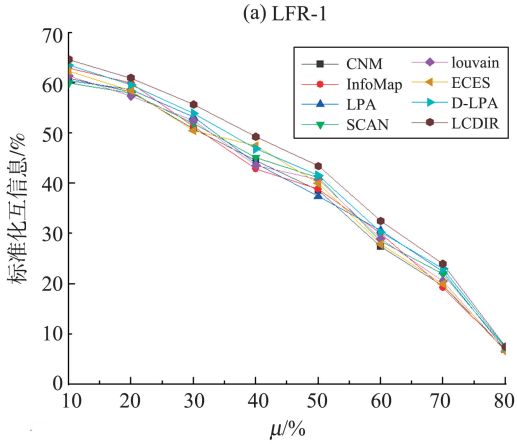
网络	n	$\langle k\rangle$	k_{max}	γ	β	C_{min}	C_{max}
LFP-1	500	20	30	5.0	1.2	10	200
LFP-2	1 000	30	50	2.5	1.5	10	200
LFP-3	5 000	50	100	2.5	1.5	20	1 000
LFP-4	10 000	100	150	2.5	1.5	50	2 000

网络的混合参数 μ 介于 0~1,用于确定内部社区链接和与外部社区链接之间比例,取值为 0~1。当 μ 接近 0 时,节点之间的连接更多地发生在其所属的内部社区之间,边的生成更倾向于社区内部的连接[31]。当 μ 接近 1 时,节点之间的连接更多地发生在不同社区之间,边的生成更倾向于社区之间连接,在试验中,取 $\mu\in[0.1,0.8]$。

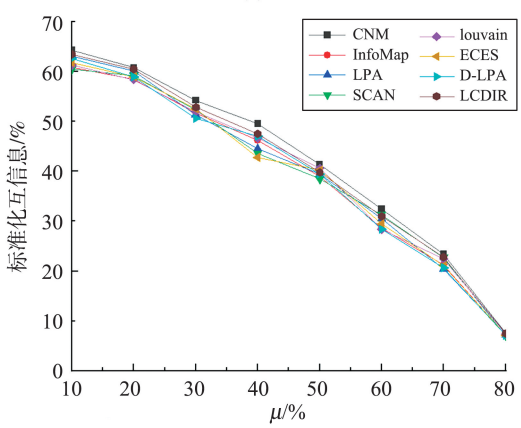
图 5 是这 7 种社区检测算法在混合参数 $\mu\in[0.1,0.8]$ 的 LFR 合成网络上试验结果,对于 μ 的所有值,LCDIR 算法都具有最高的标准化互信息值,相较于其他社区检测算法更准确有效,性能更好。算法利用不同相似性指数来对社区进行检测,从连接性和重叠度两个方面对社区结构进行检测,通过最大化模块度来对社区进行合并达到更好社区检测结果。在合成网络和现实网络中进行试验,表明该算法具有较好准确性和有效性。



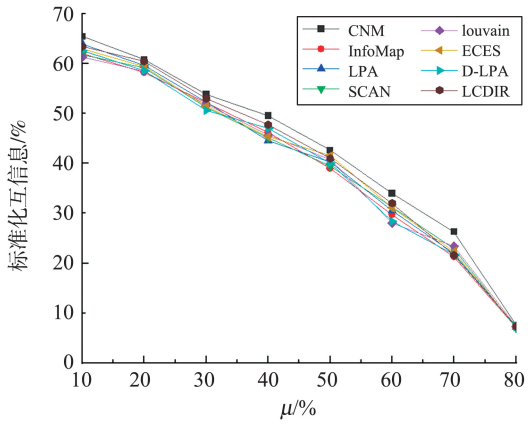
(a) LFR-1



(b) LFR-2



(c) LFR-3



(d) LFR-4

图 5 合成网络上的社区检测 NMI 值

Fig.5 NMI values of Community detection on synthetic networks

许多社区检测算法受到不确定因素的影响,例如随机参数等,在算法实现的过程中社区数量,社区里节点和网络模块化可能产生不同结果,在多次试验下才能产生最优解,在大型网络中会非常耗时,而 LCDIR 算法不存在随机参数,是一个较为稳定的算法,这是在中小型数据集中多次试验得出的结果。

4 结论

本研究提出一种基于节点重要性排序的局部社区检测算法,通过节点重要性排序得出能影响其他节点的核心节点,根据节点之间的相似性确定核心节点的社区;将其邻点添加入适宜社区中,重复添加邻点,对社区进行拓展,直至所有节点都成为其中一个社区成员;通过社区内外边比值确定需要合并的弱小社区,合并时遵循模块化度量最大化将弱小社区合并入较大社区,形成最终社区。该算法在对节点进行重要性排序时,考虑了全局结构,在一定程度上弥补了局部检测算法可能出现的局部最优情况,保持有较低时间复杂度。在 6 个真实网络和人工合成网络上将该算法和 7 种社区检测算法,包括 CNM、InfoMap、LPA、SCAN、louvain、ECES 算法和 D-LPA 算法就模块度和标准化化互信息两个指标进行对比试验。试验结果表明:该算法在这些网络上形成了较高质量社区,解决了现有局部社区检测算法存在核心节点选择不当的问题,通过节点强度和网络拓扑结构特征对网络进行社区检测,还具有较高的模块化度量值和标准化互信息值,具有较高的准确性,有效性和稳定性。

目前该算法存在的问题是在大型网络中初始社区数量很高,所需时间较长,另外只在无权无向网络中检测非折叠社区,后续的研究继续改进上述问题,提出更有效全面的社区检测算法。

参考文献:

[1] MAZZA M, COLA G, TESCONI M. Modularity-based approach for tracking communities in dynamic social networks [J]. Knowledge-Based Systems, 2023, 281: 111067.

[2] LI W, WANG J, CAI J. New label propagation algorithms based on the law of universal gravitation for community detection [J]. Physica A: Statistical Mechanics and Its Applications, 2023, 627: 129140.

[3] CHENG S, YANG S, CHENG X, et al. An effective overlapping community merging method oriented to multidimensional attribute social networks [J]. Expert Systems, 2023, 40(10): 13433.

[4] RANI S, KUMAR M. Ranking community detection algorithms for complex social networks using multilayer network design approach[J]. International Journal of Web Information Systems, 2022, 18(5): 310-341.

[5] FANG C, LIN Z Z. Overlapping communities detection based on cluster-ability optimization [J]. Neurocomputing, 2022, 494:336-345.

[6] YOU X, MA Y, LIU Z. A three-stage algorithm on community detection in social networks[J]. Knowledge-Based Systems, 2020, 187(1):104822.

[7] BERAHMAND K, BOUYER A. A link-based similarity for improving community detection based on label propagation algorithm[J]. Journal of Systems Science and Complexity, 2019, 32(3): 737-758.

[8] XU G Q, MENG L, TU D Q, et al. LCH: a local clustering H-index centrality measure for identifying and ranking influential nodes in complex networks [J]. Chinese Physics B, 2021, 30(8): 566-574.

[9] ZHANG J, ZHANG G, YANG J, et al. Local community detection algorithm based on hierarchical clustering[J]. Journal of Information & Computational Science, 2015, 12(7): 2805-2813.

[10] AGHAALIZADEH S, AFSHORD S T, BOUYER A, et al. A three-stage algorithm for local community detection based on the high node importance ranking in social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2020, 563:125420.

[11] 杨旭华,沈敏. 基于特征向量局部相似性的社区检测算法[J]. 计算机科学, 2020, 47(2):56-64.
YANG Xuhua, SHEN Min. Community detection algorithm based on local similarity of feature vectors[J]. Computer Science, 2020, 47(2): 56-64.

[12] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 2008(3): 10008.

[13] BERGSTROM C T, ROSVALL M. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123.

[14] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(9): 036106.

[15] LIU X, MURATA T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks[J]. Physica A: Statal Mechanics and its Applications, 2010, 389(7): 1493-1500.

[16] XING Y, MENG F, ZHOU Y, et al. A node influence based label propagation algorithm for community detection in networks[J]. The scientific world journal, 2014, 2014: 627581.

[17] XU X, YURUK N, FENG Z, et al. Scan: a structural clustering algorithm for networks[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, Association for Computing Machinery, 2007: 824-833.

[18] HU F, LIU Y. A new algorithm CNM-Centrality of detecting communities based on node centrality[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 446:138-151.

[19] BERAHMAND K, BOUYER A, VASIGHI M. Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes [J]. IEEE Transactions on Computational Social Systems, 2018, 5(4): 1021-1033.

[20] 蔡威林,葛斌.基于影响度的标签传播算法[J].佳木斯大学学报:自然科学版, 2022, 40(1):38-40.
CAI Weilin, GE Bin. Label propagation algorithm based on influence degree[J]. Journal of Jiamusi University: Natural Science Edition, 2022, 40(1): 38-40.

[21] SORENSEN T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons[J]. Biologiske Skrifter, 1948, 5: 1-5.

[22] WU H, DONG S, RAO B. Latitudinal trends in the structure, similarity and beta diversity of plant communities invaded by Alternantheraphiloxeroides in heterogeneous habitats[J]. Frontiers in Plant Science, 2022, 13: 1021337.

[23] BOUYER A, SABAVAND MONFARED M, NOURANI E, et al. Discovering overlapping communities using a new diffusion approach based on core expanding and local depth traveling in social networks [J]. International Journal of General Systems, 2023, 52(8): 991-1019.

[24] WANG T, YIN L, WANG X. A community detection method based on local similarity and degree clustering information[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 490: 1344-1354.

[25] FORTUNATO S, HRIC D. Community detection in networks: a user guide[J]. Physics Reports, 2016, 659: 1-44.

[26] GIRVAN M, NEWMAN M J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826.

[27] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks [J]. PhysicalReview E, 2004, 70(6): 066111.

[28] ZENG A, ZHANG C J. Ranking spreaders by decomposing complex networks[J]. Physics Letters A, 2013, 377(14): 1031-1035.

[29] ZHANG X Z, ZHANG Y B, CHEN Z L, et al. Community extraction algorithm for large-scale online social networks[J]. Journal of Northeastern University, 2015, 36(3): 342-345.

[30] YANG J, LESKOVEC J. Defining and evaluating network communities based on ground-truth [J]. Knowledge & Information Systems, 2012, 42(1): 181-213.

[31] LI C, TANG Y, LIN H, et al. Parallel overlapping community detection algorithm in complex networks based on label propagation [J]. Scientia Sinica Informationis, 2016, 46(2): 212-227.

(编辑:陈燕)