计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):146−150 [doi: 10.15888/j.cnki.csa.006393]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 LeaderRank 和节点相似性的多标签传播重叠社团挖掘算法①

王　林, 饶仁杰

(西安理工大学 自动化与信息工程学院, 西安 710048)
通讯作者: 饶仁杰, E-mail: 632657215@qq.com

摘　要: 针对基于多标签传播重叠社团挖掘算法 COPRA 因随机更新策略带来的不稳定性以及需要预先输入参数的局限性等问题, 提出一种基于 LeaderRank 和节点相似性的多标签传播重叠社团挖掘算法. 该算法首先利用 LeaderRank 算法对网络中的节点进行重要性排序从而确定节点的更新顺序, 减少标签不必要的更新. 在标签传播过程中, 根据节点相似性重新设计标签的更新策略, 提高算法的稳定性. 将算法应用于人工网络和真实网络中进行实验, 实验结果表明该算法在挖掘重叠社团上具有较高的准确性和稳定性.

关键词: 复杂网络; 重叠社团挖掘; 多标签传播算法; LeaderRank 算法; 节点相似性

引用格式: 王林,饶仁杰.基于 LeaderRank 和节点相似性的多标签传播重叠社团挖掘算法.计算机系统应用,2018,27(6):146–150. http://www.c-s-a.org.cn/1003-3254/6393.html

Multi-Label Propagation Algorithm for Overlapping Community Detection Based on LeaderRank and Node Similarity

WANG Lin, RAO Ren-Jie

(College of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The defects of overlapping community detection algorithm COPRA based on multi-label propagation include instability and pre-parameter limits, this study proposed a multi-label propagation algorithm for overlapping community detection based on LeaderRank and the node similarity. The algorithm uses the LeaderRank algorithm to sort the nodes in the network to determine the order of nodes updating. Then, re-design the label update strategy according to the similarity of nodes to improve the stability of the algorithm. The algorithm is applied to the artificial network and the real networks. The experimental results show that the proposed algorithm has high accuracy and stability for detecting overlapping communities.

Key words: complex network; overlapping community detection; multi-label propagation algorithm; LeaderRank algorithm; node similarity

近年来, 随着科技的快速发展, 对复杂网络的研究不断深入, 社团结构作为复杂网络中的重要属性也逐渐得到人们的重视. 所谓社团结构就是: 同一社团内节点之间连接紧密, 而不同社团内节点之间连接稀疏[1,2]. 社团挖掘是为了探索和发现复杂网络的社团结构, 有助于分析网络的结构和功能, 从而发现网络中隐含的内在规律, 因此社团挖掘具有重要的理论意义和广泛的应用前景.

随着研究的不断深入, 许多社团挖掘算法被广大研究者相继提出, 如基于节点分裂的 GN 算法[3], 基于

① 基金项目: 陕西省科技计划重点项目 (2017ZDCXL-GY-05-03)
收稿时间: 2017-10-09; 修改时间: 2017-11-01; 采用时间: 2017-11-10; csa 在线出版时间: 2018-05-28

模块度优化的 FN 算法[4]和 BGLL 算法[5], 基于标签传播的 LPA 算法[6]等. 其中, LPA 算法因为其简单、高效等优势, 得到了人们的普遍关注. 但是该算法仅仅只能用于非重叠网络中, 并且该算法存在稳定性差和随机性高的缺陷. 然而在真实网络中社团通常是相互重叠的, 如图 1 所示, 网络中的阴影节点同属于两个社团, 那么这两个社团就是重叠的, 阴影节点即为重叠节点. 为了能够挖掘重叠社团结构, Steve 在 LPA 算法的基础上进行延伸, 提出了多标签传播 COPRA 算法[7], 该算法允许每个节点携带 v 个标签, 其继承了 LPA 算法的优势, 但用于未知网络时无法预估网络中节点所属的社团个数, 假设节点所属社团个数分布不均时, 就很难找到一个合适的参数 v, 使得算法难以准确的得到社团结构, 且 COPRA 算法同时也继承了 LPA 算法的随机性缺陷, 使得划分的结果非常不稳定. Xie 等人基于标签传播的思想提出 SLPA 算法[8], 算法通过记录每一个节点在刷新迭代过程中的历史标签序列, 利用概率阈值删除出现频率小的标签最终得到社团结构, 但该算法仍需要一个合适的概率阈值参数, 并且采用的随机策略将导致结果存在随机性.

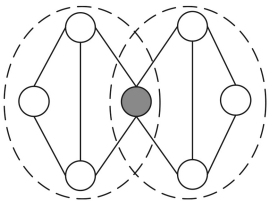


图 1 重叠社团的网络结构

为了改善现有的基于多标签传播算法的因采用随机策略导致结果不稳定, 以及需要输入额外的参数问题, 本文提出一种基于 LeaderRank 和节点相似性的多标签传播算法, 该算法引入 LeaderRank 算法[9]来衡量网络中节点重要性, 根据重要性大小确定节点的更新序列, 并重新设计了标签的更新策略, 使得到的划分结果更加稳定, 且该算法无需输入额外的参数. 实验表明本文算法较现有的多标签传播算法相比能够较准确的得到重叠社团结构.

1 基于 LeaderRank 和节点相似性的多标签传播重叠社团挖掘算法

本文利用 LeaderRank 算法对网络中的节点的更新顺序进行排序, 并结合节点的相似性重新设计了标签的更新策略提出了基于 LeaderRank 和节点相似性的多标签传播重叠社团挖掘算法.

1.1 LeaderRank 排序算法

由于先更新的节点影响传播的较远, 很多重要性较小的节点在传播时会反过来影响一些重要性较大的节点, 这样就造成了标签的逆向传播. 虽然算法在后续的迭代过程中可以修正结果, 但耗费了大量的时间和更新操作. 因此, 本文利用 LeaderRank 排序算法对节点的更新顺序进行排序减少算法不必要的更新和标签逆流现象.

算法通过在原网络中加入一个背景节点 (ground node), 将其与网络中的所有节点相连, 得到一个有 N+1 个节点的强连接网络.

首先, 给除背景节点以外的所有节点分配 1 单位的 LR 值, 然后将这 1 单位的 LR 值分配给其邻居节点, 不断重复这一过程, 直到达到稳定状态, 即公式所示:

$$s_i(t+1)=\sum_{j\in N(x)}\frac{s_j(t)}{k_j} \tag{1}$$

其中, $N(x)$表示节点 x 的邻居节点集合, k_j表示节点 j 的度, $s_j(t)$表示第 t 次迭代节点 j 的 LR 值.

达到稳定状态[9]后, 此时再将背景节点的 LR 值分配给其他所有节点, 因此, 最终的 LR 值定义为:

$$S_i=s_i(t_c)+\frac{s_g(t_c)}{N} \tag{2}$$

其中, t_c表示收敛次数, $s_g(t_c)$表示稳定状态下节点 g 的 LR 值.

在计算出所有节点 LR 值之后, 将 LR 值进行降序排序, LR 值越大说明节点的重要性越大, 从而得到了节点的更新顺序.

1.2 标签更新策略

在 COPRA 算法中, 更新节点的标签时算法简单地认为每个节点与其邻居节点之间的关系是相等的, 显然这与现实生活中的情况不符合. 比如 a 有 b, c, d, e 四个邻居节点, 但如果邻居节点 b 与节点 a 的关系更加亲密, 那么节点 a 更加容易接收来自节点 b 的标签. 因此, 本文重新设计了标签的更新策略进行标签更新. 首先给出如下定义:

定义 1. 节点间的相似性:

$$sim(x,y)=\frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{3}$$

其中$\Gamma(x)$表示节点 x 的邻居节点.

定义 2. 领导标签: 节点的标签集合中具有最大从属系数的标签.

$$Dl_x = \arg\max(b(c,x)) \tag{4}$$

结合定义 1 与定义 2 给出了新的标签更新策略, 如下式所示:

$$b(c,x) = \sum_{y\in N(x)} \theta(c,y)b(c,y)\,sim(x,y) \tag{5}$$

其中, $\theta(c,y) = \begin{cases} 1, & Dl_y = c \\ 0, & \text{其他} \end{cases}$.

根据以上定义, 标签的更新策略描述如下: 首先给每个节点分配一个唯一的标签, 并且标签的从属系数为 1. 然后为了对标签进行更新, 利用公式 (3) 计算出节点与其所有邻居节点的相似性, 当更新节点 x 时, 节点 x 只接收来自其邻居节点中从属系数最大的领导标签, 随后将节点 x 从邻居节点接收到的最大从属系数乘以对应邻居节点的相似性, 得到节点 x 更新后的标签集合L_x, 设定阈值$p = \dfrac{sum(b)}{n}$, 其中 b 表示节点 x 的标签集合L_x中所有标签的从属系数, n 表示节点 x 的标签数量, 当$b < p$时, 删除标签, 当$b \geq p$时, 则保留标签, 最后对节点 x 的标签从属系数进行标准化.

1.3 终止条件

初始时, 标签数量等于节点数量, 随着标签更新过程而减少, 最后达到最小值, 即为最终的社团数量. 定义i^t为第 t 次迭代的标签数量 (社团数量), 如果当$i^t = i^{t-1}$时, 算法停止, 可能导致输出结果不准确, 虽然此时标签数量不变, 但是每个社团内的节点数量可能在几次迭代后发生变化. 因此, 终止条件不仅要满足$i^t = i^{t-1}$, 还要注意每个社团内节点数量的变化情况. 第 t 次迭代时, 社团 c 的节点数量可以表示如下:

$$c_t = \left\{ (c,i) : c \in V \wedge i = \sum_{x\in V, b_t(c,x)>0} 1 \right\} \tag{6}$$

拥有标签 c 的节点数量会在迭代过程中发生变化, 本文观测拥有标签 c 的最小节点数的变化, 当$i^t = i^{t-1}$时,

$$m_t = \{(c,i) : \exists p \exists q((c,p) \in c_{t-1} \wedge (c,q) \in c_t \wedge i = \min(p,q))\} \tag{7}$$

否则$m_t = c_t$, 只要满足$m_t = m_{t-1}$, 算法停止.

1.4 算法描述

本文算法的主要过程可以分为 3 个阶段: 初始化、确定节点更新顺序和标签传播. 首先初始化每个节点, 给每个节点分配一个独特的标签, 并且该标签的从属系数为 1. 然后为了减少算法的迭代次数并避免标签的逆流现象, 利用 LeaderRank 算法对每个节点进行排序, 确定标签的更新顺序. 再根据本文提出的标签更新策略进行标签更新, 直到达到标签传播的终止条件结束迭代过程. 最后将具有相同标签的节点归为一个社团, 如果节点有多个标签那么该节点为重叠节点.

算法具体描述如下:

输入: 网络$G(V,E)$, 邻接矩阵 A.
输出: 重叠社团划分结果.

1) 计算网络中所有节点的 LR 值, 并按降序排序.
2) 计算网络中所有节点之间的相似性, 得到相似性矩阵.
3) 初始化, 给每个节点的标签初始化为$(c_x,1)$.
4) 节点 x 接收每个邻居节点的领导标签以及对应的从属系数.
5) 利用公式 (5) 更新节点 x 的标签从属系数.
6) 设定阈值$p=\frac{sum(b)}{n}$, 删除小于阈值的标签, 最后对剩余标签的从属系数进行标准化.
7) 重复 4)–6) 步骤, 直到达到迭代条件, 算法停止.
8) 输出社团, 具有多个标签的节点则为重叠节点.

1.5 复杂度分析

设网络中包括 n 个节点, k 为节点的平均度, m 为边的数量, L 为节点平均包含的标签个数.

利用 LeaderRank 算法对网络中的节点进行排序需要$O(ml)$, 其中 l 为 LeaderRank 算法收敛时的迭代次数. 初始化标签需要$O(n)$, 计算相似性矩阵需要$O(m)$, 更新标签时每次迭代需要$O(Lmlog(Lm/n))$, 假设共需要迭代 T 次, 所以整个标签更新过程需要$O(TLmlog(Lm/n))$, 故算法的总复杂度为$O(ml+m+n+TLmlog(Lm/n))$, 由于 L 通常非常小, 且 T 远小于 m, 所以最终的算法复杂度为$O((m+1)l+n)$.

2 实验分析

为了测试本文提出的算法的性能, 将算法应用于真实网络和人工网络数据集进行验证分析, 并与 COPRA 算法[7]、SLPA 算法[8]、BMLPA 算法[10]进行对比分析, 以验证算法的有效性.

2.1 评价指标

为了评价算法性能, 本文利用标准互信息和扩展模块度 EQ 来衡量重叠社团结构的划分结果.

1) 扩展模块度 (EQ)

模块度 Q[11]是评价社团质量的主要评价指标, 然而它只能用于评价非重叠社团质量, 并不能够准确评

价重叠社团质量. 因此, 本文引入扩展模块度 EQ[12]作为重叠社团质量的评价指标. EQ 的取值范围是 0 到 1 之间, 值越大说明重叠结构越好.

$$EQ=\frac{1}{2m}\sum_{ij}\frac{1}{O_iO_j}\left(A_{ij}-\frac{d_id_j}{2m}\right)\delta\left(C_i,C_j\right) \tag{8}$$

其中, m 表示网络中边的个数, O_i表示节点 i 属于社团的数量, d_i表示节点 i 的度, $\delta\left(C_i,C_j\right)$表示如果节点 i 和节点 j 在同一个社团则值为 1, 否则为 0.

2) 标准互信息 (NMI)

由于人工网络的社团结构是已知的, 因此采用标准互信息 NMI(Normalized Mutual Information)[13]作为社团挖掘的评价指标. NMI 的取值范围是 0 到 1 之间, 算法划分的社团结构越准确, NMI 值越大; 否则 NMI 值越小.

$$H(x|y)=1-\frac{1}{2}\left[H(x|y)_{\text{norm}}+H(y|x)_{\text{norm}}\right] \tag{9}$$

其中, x 和 y 分别表示真实的社团结构和实验产生的社团结构.

2.2 真实网络实验

为了验证本文算法在真实网络数据集上的有效性, 将其应用于 3 种真实网络数据集上, 表 1 列出了用于测试的 3 种真实网络数据集.

表 1 真实网络数据集

网络	节点	边	描述
Karate	34	78	空手道网络[14]
Dolphins	62	159	海豚网络[15]
Football	115	613	美国大学生足球网络[3]

由于 COPRA 算法和 SLPA 算法具有随机性, 每个数据集均采取运行 20 次的方式获取 COPRA 算法和 SLPA 算法的平均实验结果. 而本文算法以及 BMLPA 算法由于算法的稳定性只需本别对数据集进行一次实验. 实验中 COPRA 算法的参数 v 设为 2, SLPA 算法的概率阈值设为 0.2, BMLPA 的标签过滤阈值设为 0.75.

从表 2 中可以得出, 本文算法在选取的 3 个真实网络数据集上得出的社团结构重叠模块度较其他算法相比均有所提高, 因此本文算法可以有效提高重叠社团挖掘的重叠模块度, 得到的社团结构质量更高.

2.3 人工网络实验

为了生成 LFR 人工网络, 使用 LFR 基准程序[16]生成了如表 3 所示的 4 种类型的人工网络. 其中 N 是网络的节点数, k 是节点的平均度数, $maxk$ 是节点的最大度数, $maxc$ 是一个社团的最大节点数, $minc$ 是一个社团的最小节点数, on 是重叠节点的个数, om 是重叠节点属于的社团个数, mu 是节点与社团外部连接的边数与该节点度数的比值, 取值范围是 0 到 1 之间, 值越大说明网络的社团结构越不明显, 所以 mu 取值从 0.1 到 0.6.

表 2 真实网络数据集上的 EQ 值比较

	本文算法	COPRA	SLPA	BMLPA
Karate	0.4156	0.312	0.3728	0.3478
Dolphins	0.4926	0.3849	0.4358	0.4355
Football	0.6016	0.5876	0.5636	0.5809

表 3 人工网络的参数设置

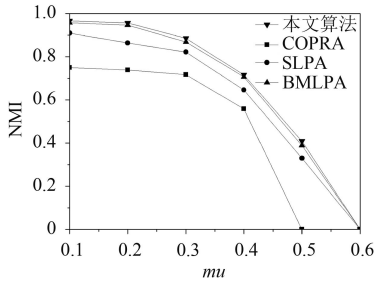
	R1	R2	R3	R4
N	1000	1000	1000	1000
k	10	10	10	10
$maxk$	30	30	30	30
$minc$	10	10	10	10
$maxc$	50	50	50	50
on	100	200	100	100
om	2	2	2~6	2~6
mu	0.1~0.6	0.1~0.6	0.1	0.3

实验中, 针对 R1 和 R2 网络, COPRA 算法的参数 v 取值 2, 针对 R3 和 R4 网络, COPRA 算法的参数 v 取值 2～6. 在 4 种网络中 SLPA 算法的概率阈值 r 均取值 0.2, BMLPA 算法的标签过滤阈值设为 0.75.

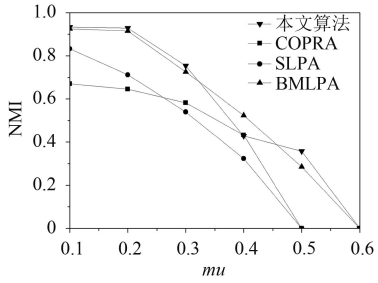
图 2 为本文算法与 COPRA 算法、SLPA 算法以及 BMLPA 算法在 4 种类型人工网络上的 NMI 值对比图. 可以看出, 本文算法仅在 R2 网络的 mu=0. 4 时未取得最优值, 这是因为本文算法运用在该网络中产生了过多的重叠节点, 因此影响了社团的质量, 其他网络所得的 NMI 值都大于对比算法. 因此, 在人工网络中, 本文算法能够挖掘出高质量的重叠社团结构.

3 结论与展望

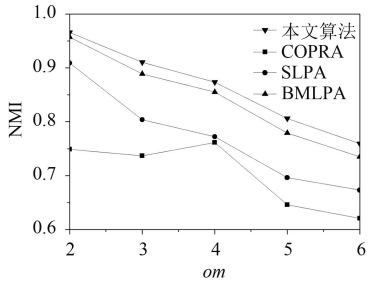
本文针对现有的多标签传播算法在挖掘重叠社团时, 稳定性较差以及需要输入额外参数的问题, 提出了一种稳定的且无需任何参数的多标签传播算法. 该算法利用 LeaderRank 算法衡量节点重要性确定更新顺序, 在进行标签更新时, 节点只接收来自邻居节点的领导标签, 并且考虑了节点与邻居节点之间相似性不同的特点, 使得更新过程更加稳定. 通过在真实网络和人工网络上进行实验, 结果表明本文算法较现有的基于多标签传播的重叠社团挖掘算法得到的社团质量更高.



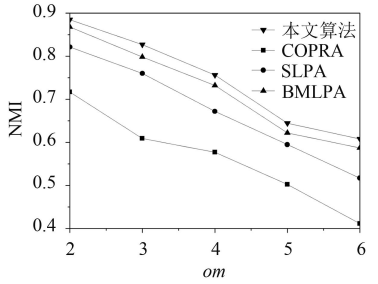
(a) R1 网络的 NMI 比较



(b) R2 网络的 NMI 比较



(c) R3 网络的 NMI 比较



(d) R4 网络的 NMI 比较

图 2 算法在 4 种网络的 NMI 比较

参考文献

1 Fortunato S. Community detection in graphs. Physics Reports, 2010, 486(3-5): 75–174. [doi: 10.1016/j.physrep.2009.11.002]

2 罗明伟, 姚宏亮, 李俊照, 等. 一种基于节点相异度的社团层次划分算法. 计算机工程, 2014, 40(1): 275–279.

3 Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821–7826. [doi: 10.1073/pnas.122653799]

4 Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 69(6): 066133. [doi: 10.1103/PhysRevE.69.066133]

5 Blondel VD, Guillaume JL, Lambiotte R, *et al*. Fast unfolding of community hierarchies in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008. [doi: 10.1088/1742-5468/2008/10/P10008]

6 Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007, 76(3): 036106. [doi: 10.1103/PhysRevE.76.036106]

7 Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010, 12(10): 103018. [doi: 10.1088/1367-2630/12/10/103018]

8 Xie J, Szymanski BK, Liu X. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. Proceedings of the IEEE 11th International Conference on Data Mining Workshops. Vancouver, Canada. 2011. 344–349.

9 Lü LY, Zhang YC, Yeung CH, *et al*. Leaders in social networks, the Delicious case. PLoS One, 2011, 6(6): e21202. [doi: 10.1371/journal.pone.0021202]

10 Wu ZH, Lin YF, Gregory S, *et al*. Balanced multi-label propagation for overlapping community detection in social networks. Journal of Computer Science and Technology, 2012, 27(3): 468–479. [doi: 10.1007/s11390-012-1236-x]

11 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2004, 69(2): 026113. [doi: 10.1103/PhysRevE.69.026113]

12 Shen HW, Cheng XQ, Guo JF. Quantifying and identifying the overlapping community structure in networks. Journal of Statistical Mechanics: Theory & Experiment, 2009, (7): 07042.

13 Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 2009, 11(3): 033015. [doi: 10.1088/1367-2630/11/3/033015]

14 Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977, 33(4): 452–473. [doi: 10.1086/jar.33.4.3629752]

15 Lusseau D, Schneider K, Boisseau OJ, *et al*. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, 54(4): 396–405. [doi: 10.1007/s00265-003-0651-y]

16 Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, Covering Statistical, Nonlinear, Biological, and Soft Matter Physics, 2008, 78(4): 046110.