

基于模块度优化的标签传播社区发现算法^①

李磊, 倪林

(中国科学技术大学 电子工程与信息科学系, 合肥 230027)

摘要: 标签传播算法(LPA)是一种快速高效的社区发现算法, 算法无需社区数量等先验信息, 但存在大量随机性, 稳定性较差. 为了提高标签传播算法的稳定性, 提出了一种改进的标签传播算法(LPAMP). 该算法分为两个阶段, 第一阶段以模块度贪婪为依据, 进行节点粗聚类; 第二阶段在粗聚类的基础上, 进行节点标签传播. 实验结果表明, 所提算法降低了标签传播算法的随机性, 增强了稳定性, 并且提高了准确率.

关键词: 标签传播; 社区发现; 模块度; 贪婪; 优化

Community Detection for Label Propagation with Modularity Optimization

LI Lei, NI Lin

(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China)

Abstract: Label Propagation Algorithm (LPA) is a fast and efficient community detection algorithm and this algorithm does not need to know the prior information such as the number of communities. However, this algorithm has a large number of randomness, which leads to unstable results. In order to improve the stability of label propagation algorithm, we propose an improved label propagation algorithm (LPAMP). The algorithm is divided into two phases. In the first phase, vertices are clustered roughly by optimizing the modularity greedily; in the second phase, the labels propagate through the network based on the result of the first phase. Experimental results show that the proposed algorithm not only reduces the randomness of the label propagation algorithm, but also improves the stability and increases the accuracy.

Key words: label propagation; community detection; modularity; greed; optimization

网络在现实世界中无处不在. 许多复杂系统, 如社交网络、交通网络、生物网络等都是网络的不同表现形式. 研究发现, 复杂网络具有小世界效应^[1]、幂律分布和无标度性^[2]. 2002年, Girvan 和 Newman 等指出社区结构^[3]在复杂网络中广泛存在. 随着计算机和信息技术的快速发展, 特别是社交网络的出现, 社区结构的研究显得愈发具有价值, 社区发现在科学研究、商业推广、公共安全等领域有了越来越多的应用, 得到了越来越广泛的关注.

社区发现是指将社区中的节点按照连接的紧密程度划分为不同的类, 同一个类的节点之间连接相对紧密而不同类的节点之间连接相对稀疏. 2007年, Raghavan^[4]等人提出了标签传播算法(LPA), 该算法具有接近线性的复杂度, 且无需社区数量等先验知识,

适合在大规模网络中进行社区发现, 得到了广泛应用. 然而, 该算法存在具有大量随机性, 社区划分结果不够稳定, 极易得到一些较差的划分. 在该算法基础上, Barber^[5]等提出了一种模块化标签传播算法, 有效地防止将所有节点划分到一个网络; Cordasco^[6]等人提出了一种基于染色模型的并行算法, 提高了标签的更新速度; 文献[7]提出了基于节点重要性和标签影响力的标签传播算法, 提高算法的稳定性.

本文对 LPA 算法进行研究, 分析了该算法中存在的随机性和不稳定性. 为了提高社区划分的质量, 在 LPA 基础上提出了 LPAMP 算法, 该算法首先以模块度贪婪为依据, 进行节点的粗聚类, 然后在粗聚类的基础上进行标签传播. 实验表明, LPAMP 算法降低了 LPA 算法的随机性, 提高了社区划分结果的稳定性和

^① 收稿时间:2016-01-18;收到修改稿时间:2016-03-22 [doi:10.15888/j.cnki.csa.005416]

准确性。

1 LPA算法

LPA(Label Propagation Algorithm)快速高效,可以对大规模网络进行社区发现。在复杂网络中,同一个社区内的节点之间连接相对紧密,不同社区的节点之间连接相对稀疏。文献[4]认为,如果社区 C 包含了一个节点最多的邻居节点,那么该节点也属于社区 C 。算法的具体步骤如下:

- 1) 将网络中每个节点初始化为唯一且不同的社区标签;
- 2) 将所有节点进行随机排列;
- 3) 按照排列后的顺序依次将所有节点进行标签更新,统计节点的所有邻居节点的标签,将节点更新为包含其最多邻居节点的社区标签,若两个或两个以上标签的频率同时取得最高,则随机选取一个;
- 4) 重复 2)和 3),直到每个节点的标签都与其邻居节点中出现频率最高的标签相同。

在步骤 3)中,原文采用异步更新策略,即节点的更新依次进行,数学描述为 $C_v(t) = f(l_{v_{i1}}(t), \dots, l_{v_{im}}(t), l_{v_{i(m+1)}}(t-1), \dots, l_{v_{ik}}(t-1))$, 其中, $l_v(t)$ 表示节点 v 在 t 时刻的社区标签, $v_{i1}, \dots, v_{im}$ 表示在当前迭代已经更新过的节点, $v_{i(m+1)}, \dots, v_{ik}$ 表示该轮迭代还未更新的节点。同步更新则可以描述为 $C_v(t) = f(l_{v_1}(t-1), \dots, l_{v_k}(t-1))$, 虽然更快,适合于并行框架,但同步更新可能会震荡,特别是存在二分网络结构的时候。LPA 标签更新策略的数学描述如下:

$$l'_v = \operatorname{argmax}_l \sum_{u \in \sigma(v)} \delta(l_u, l)$$

其中, l_u 表示节点当前的社区标签, l'_v 表示节点 v 更新后的社区标签, $\sigma(v)$ 表示节点 v 的邻居节点集合, δ 表示狄利克雷函数。

2 LPAMP算法

2.1 基于模块度贪婪的预处理

LPA 算法将每个节点初始化为唯一且不同的社区标签,标签传播前将所有节点随机排列,若邻居节点中有不止一个标签出现频率同时取得最高,随机选取一个,这使得 LPA 算法具有大量随机性,降低了划分

结果的稳定性和准确性。特别的,第一轮传播过程中,大量的节点会随机选择其邻居节点的标签。这导致了节点之间一些无意义的影响,产生了所谓的“逆流”现象^[8],带来大量无价值的标签更新,这一方面影响了标签的更新速度,另一方面影响了算法的稳定性和准确性。因此,在标签传播前,进行一个粗聚类,便能有效地降低前期迭代中由于不止一个标签频率最高导致的随机性,从而提高算法的稳定性和准确性。

模块度(Modularity)^[9]的提出对于社区发现有重要的意义,它不仅是社区划分的一种评价标准,基于模块度的优化也成为了一类社区发现方法^[10-14]。模块度计算公式为:

$$Q = \sum_{c=1}^{n_c} \frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \quad (1)$$

其中, m 是网络中边的总数, n_c 是社区数目, l_c 是社区 c 中边的总数, d_c 是指社区 c 中所有节点的度之和。

由公式(1)知,将社区 c_2 中的节点 i 移到社区 c_1 的前后,两个社区模块度的和分别为:

$$Q_{\text{before}} = \frac{l_{c1}}{m} - \left(\frac{d_{c1}}{2m} \right)^2 + \frac{l_{c2}}{m} - \left(\frac{d_{c2}}{2m} \right)^2 \quad (2)$$

$$Q_{\text{after}} = \frac{l_{c1} + k_{i,c1}}{m} - \left(\frac{d_{c1} + k_i}{2m} \right)^2 + \frac{l_{c2} - k_{i,c2}}{m} - \left(\frac{d_{c2} - k_i}{2m} \right)^2 \quad (3)$$

(3)-(2), 得到

$$\Delta Q = \frac{1}{m} \left[k_{i,c1} - k_{i,c2} - \frac{k_i(d_{c1} - d_{c2} + k_i)}{2m} \right] \quad (4)$$

其中, $k_{i,c1}$ 表示节点 i 与社区 c_1 中节点间连接边数, $k_{i,c2}$ 表示与社区 c_2 中节点间的连接边数。

基于模块度贪婪预处理的主要思想是依次计算节点并入其各个邻居节点所在社区后模块度的增量,将其并入到模块度增量最大的社区,如果最大模块度增量是非正,则保持原社区不变。重复进行并入,直到指定迭代次数或直到模块度不再增加为止。由于模块度贪婪是基于局部信息,所有该预处理部分的时间复杂度 $O(m)$,效率非常高。

2.2 算法流程

本文提出的算法主要分为两个阶段,首先以模块度贪婪为依据进行粗聚类,然后在此基础上进行标签传播。具体流程如下:

- 1) 将所有节点初始化为唯一且不同的社区标签;

2) 依次更新每个节点的社区标签. 方法是计算节点并入其邻居节点所在社区带来的模块度增量, 将其更新为模块度增量最大且为正的社区标签, 如果最大增量非正, 则保留在原社区;

3) 将 2) 重复进行指定的 K 次或者直到模块度不再增加为止, 将结果作为后续标签传播的初始划分;

4) 将所有节点进行随机排列, 依次更新所有节点的标签, 更新为邻居节点中出现频率最高的社区标签, 若多个标签频率同时最高, 则随机选择;

5) 重复 4), 直到所有的节点标签都与其邻居节点中出现频率最高的标签相同.

其中, 3) 中的 K 取值越大, 基于模块度优化的初始聚类对最终社区发现结果的影响越大. 实验表明, 若社区结构简单明显, K 应取较小值, 此时初始聚类得到的类个数较多, 类均节点数量较少, 标签传播的结果稳定准确; 若社区结构复杂模糊, K 更适合取较大值, 此时初始聚类得到的类个数较少, 类均节点数量较多, 模块度优化聚类的影响较大, 使得标签传播在社区结构复杂模糊的网络中也能有较好的表现.

3 实验分析

实验部分对 LPA 及本文提出的算法进行对比. 标签传播算法具有随机性, 因此所有实验结果均是运行 100 次取平均值. 实验环境为 Windows7 系统, Inter(R) Core(TM) i5-3337U CPU@1.80GHz 1.80 GHz, 4.00GB 内存.

3.1 实验数据

3.1.1 LFR 基准网络数据集

LFR 网络^[15]是社区发现中经典的人工生成测试数据集. 度数和社区大小均服从参数为 γ 和 β 的幂律分布, 本实验设置参数节点个数 N 为 1000, 平均节点度 k 为 25, 最大节点度 max_k 为 50, 单个社区最小、最大节点个数 min_c 、 max_c 分别为 20、100, 四组混合系数 μ 分别取 0.45、0.50、0.55、0.60.

3.1.2 真实网络数据集

本文选取了四个经典的真实网络, 分别为 Karate、Dolphin、Jazz、Email 网络. 四个网络的节点数和边数分别为 34、78、62、159、198、2742、1133、5451.

3.2 社区划分评价标准

LFR 基准网络、Karate 和 Dolphin 网络都有已知的社区结构, 因此我们以 NMI、模块度两个标准, 衡量社区划分的结果; 而 Jazz 和 Email 网络, 由于没有已

知社区结构, 只能以模块度为标准来衡量. 模块度的计算公式见 2.1 节, 模块度表示实际网络中的社区结构和与其具有相同期望度序列的随机网络之间的差异, 而该随机网络是被认为没有社区结构的, 因此, 模块度越大, 社区结构往往越明显. NMI 计算公式为:

$$NMI = \frac{-2 \sum_{i,j} N_{ij} \log(N_{ij} / N_i N_j)}{\sum_i N_i \log(N_i / N) + \sum_j N_j \log(N_j / N)}$$

其中, N 为混合矩阵, N_{ij} 表示 i 、 j 两个类中公共节点数, N_i 、 N_j 分别表示 N 中第 i 行和第 j 列的和. NMI 表示算法得到的社区划分与真实的社区划分之间的相似性. NMI 的取值在 0 到 1 之间, 越接近于 1, 表示算法得到的社区划分与实际社区结构越一致, 划分结果越好.

3.3 实验结果

3.3.1 LFR 基准网络数据集

实验中, LFR 基准网络节点数为 1000, 四组混合系数分别为 0.45、0.50、0.55、0.60, 混合系数 μ 较小时, 网络社区结构较为简单清晰. μ 取 0.45、0.50、0.55 时网络社区结构较明显, K 取 0, 即只进行一轮模块度优化聚类; μ 取 0.60 时, 进行多轮模块度优化聚类, 直到模块度不能再增加为止. 从表 1 和表 2 的实验结果可以看出, 当社区结构简单清晰时, LPAMP 与 LPA 在 NMI 和模块度上表现相当; 然而, 随着 μ 增大, 网络结构更加复杂模糊, 相比于 LPA, LPAMP 在 NMI 和模块度上都有较大提高. 特别的, LPA 在 $\mu=0.60$ 时, NMI 和模块度都在 0.1 以下, 已经失去社区划分的意义, 但 LPAMP 仍然能较较好地划分, NMI 和模块度分别为 0.8356、0.2958.

表 1 LPAMP 和 LPA 在 LFR 网络上 NMI 的对比

μ	0.45	0.50	0.55	0.60
算法				
LPA	0.9726	0.8827	0.5344	0.0826
LPAMP	0.9872	0.9572	0.9462	0.8356

表 2 LPAMP 和 LPA 在 LFR 网络上模块度的对比

μ	0.45	0.50	0.55	0.60
算法				
LPA	0.4796	0.3973	0.2183	0.0329
LPAMP	0.4865	0.4314	0.3878	0.2958

3.3.2 真实网络数据集

真实数据集上, 本文选取了 Karate、Dolphin、Jazz、

Email 四个网络进行测试。由于真实网络社区结构相对复杂, K 取较大值, 使得模块度不再增加为止。前两个分别作了 NMI 和模块度的对比, 后两个由于没有已知社区结构, 只进行模块度对比。从表 3 和表 4 的实验结果可以看出, 在四个网络中, LPAMP 在模块度上比 LPA 有明显提高, 分别提高了 8.36%、2.71%、19.30%、36.34%; 在前两个网络中, LPAMP 比 LPA 的 NMI 分别提高了 7.06%、12.16%。上述结果表明, 本文所提 LPAMP 算法能够有效地进行社区发现, 社区划分结果比 LPA 稳定、准确。

表 3 LPAMP 和 LPA 在真实网络上 NMI 的对比

网络 算法	Karate	Dolphin	Jazz	Email
LPA	0.7140	0.5937	----	----
LPAMP	0.7644	0.6659	----	----

表 4 LPAMP 和 LPA 在真实网络上模块度的对比

网络 算法	Karate	Dolphin	Jazz	Email
LPA	0.3530	0.4432	0.3534	0.3553
LPAMP	0.3825	0.4552	0.4216	0.4844

4 结语

本文针对 LPA 算法中的不稳定性, 提出了一种基于模块度贪婪预处理的标签传播算法。该算法利用局部信息, 以模块度贪婪为依据, 对节点进行粗聚类, 并将其结果作为初始社区划分进行标签传播。在 LFR 基准网络和真实网络数据集上的实验结果表明, 相较于 LPA, 本文所提算法发现的社区结构质量更高, 划分结果更稳定。本文的算法思路是异步串行, 下一步工作将考虑如何设计算法, 将其部署到并行计算架构, 从而快速实现超大规模网络的社区发现。

参考文献

- Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*, 1998, 393(6684): 440–442.
- Barabási AL, Albert R. Emergence of scaling in random networks. *Science*, 1999, 286(5439): 509–512.
- Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc. of the National Academy of Sciences*, 2002, 99(12): 7821–7826.
- Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, 76(3): 036106.
- Barber MJ, Clark JW. Detecting network communities by propagating labels under constraints. *Physical Review E*, 2009, 80(2): 026129.
- Cordasco G, Gargano L. Community detection via semi-synchronous label propagation algorithms. 2010 IEEE International Workshop on Business Applications of Social Network Analysis (BASNA). IEEE, 2010, 1–8.
- 黄佳鑫, 郭昆, 郭红. 融入节点重要性和标签影响力的标签传播社区发现算法. *小型微型计算机系统*, 2015, 36(6): 1171–1175.
- 赵卓翔, 王铁彤, 田家堂, 等. 社会网络中基于标签传播的社区发现新算法. *计算机研究与发展*, 2011, (S3): 8–15.
- Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, 69(2): 026113.
- Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, 2013, 86(11): 1–14.
- Hassan EA, Hafez AI, Hassanien AE, et al. Community detection algorithm based on artificial fish swarm optimization. *Intelligent Systems' 2014*. Springer International Publishing, 2015: 509–521.
- Danon L, Díaz-Guilera A, Arenas A. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006, 2006(11): P11010.
- Naeni LM, Berretta R, Moscato P. MA-Net: A reliable memetic algorithm for community detection by modularity optimization. *Proc. of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, Volume 1. Springer International Publishing, 2015. 311–323.
- Duch J, Arenas A. Community detection in complex networks using extremal optimization. *Physical Review E*, 2005, 72(2): 027104.
- Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 2008, 78(4): 046110.