

引用格式:涂佩文,周金和. 基于节点合并信息熵的社团结构检测[J]. 微电子学与计算机, 2020, 37(7): 42-46. [TU P W, ZHOU J H. Community structure detection based on node merging information entropy [J]. Microeletronics & Computer, 2020, 37(7): 42-46.]

基于节点合并信息熵的社团结构检测

涂佩文,周金和

(北京信息科技大学 信息与通信工程学院,北京 100101)

摘 要:为解决目前社团划分存在的复杂度高及划分精度低等问题,从信息熵理论角度出发提出一种新的社团结构检测算法.针对节点划分概率系统事件发生的不确定性,采用相似性指标计算各节点合并所提供的信息量,结合全局重要度进一步构造节点合并信息熵模型,通过熵函数的值判断具有最小不确定度的合并方案,运用层次聚类的思想实现最终社团划分.通过对真实网络数据集的对比分析,验证了该算法的有效性,且划分社团具有较高的模块度.

关键词: 复杂网络;社团结构;节点相似性;不确定度;合并信息熵

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2020)07—0042—05

Community structure detection based on node merging information entropy

TU Pei-wen, ZHOU Jin-he

(College of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: In order to solve the problems of high complexity and low division accuracy of the current community division, a new community structure detection algorithm is proposed from the perspective of information entropy theory. In view of the uncertainty of the occurrence of node division probability system events, the similarity index is used to calculate the information provided by node combination, and the further construction of node merge information entropy model is made by combining the global importance degree. The value of the entropy function is used to judge the merging scheme with the least uncertainty. Apply the idea of hierarchical clustering to realize the final community division. Through the comparison and analysis of real network datasets, the effectiveness of the algorithm is verified, and the divided community has a high degree of modularity.

Key words: complex network; community structure; node similarity; uncertainty; information entropy

1 引言

信息时代的发展让大量数据如雨后春笋般不断涌现,如何有效地分析和挖掘数据内部的隐藏信息是当前研究的热点问题.正如科学界许多系统都可以有效地表示为网络,数据之间错综复杂的关系也

可以通过网络表现出来.而社团结构作为复杂网络异质性与模块化的反映,深入研究有助于了解其各个层面的内部特征,在很多领域都得到了有效应用,例如内容个性化推荐^[1]、网络链路重连^[2]等.

近年来,基于节点相似性的算法越来越为广大研究者们所青睐,文献[3-5]的方法有一定改进,但

大多只是在相似性指标上做研究,只考虑网络的局部特性,仍具有时间复杂度较高,划分精度偏低等不足。

为了快速且准确地发现最优的社团结构,本文提出了一种基于节点合并信息熵(Merging Information Entropy)的划分算法(简称 MIE 算法),把节点相似性转化为合并信息量的度量之一,将整个社团检测看作是平均不确定度丢失的过程,从而达到有效的社团划分。在多个真实数据集上对算法性能进行评估,结果表明本算法具有较高的准确性和划分精度。

2 节点合并信息熵模型

复杂网络系统可以被描述为一个无权无向图 $G = (V, E)$, 其中, $V = \{v_1, v_2, \dots, v_N\}$ 表示 N 个网络节点的集合, $E = \{e_1, e_2, \dots, e_M\}$ 表示 M 条网络连边的集合。当社团扩张时,社团内部节点的连接可以暂时忽略,只考虑社团之间的连接,因此可以把合并后的社团收缩为新节点。用 Γ_i 和 Γ_j 分别表示节点 v_i 和节点 v_j 的邻节点集合,则合并收缩过程可表示为将 $\delta_{\Gamma_i}(\Gamma_i \cap \Gamma_j)$ 重连接到新节点 $v_{j'}$ 上。

定义节点 v_i 的合并概率动态系统为 $\{X_{i1}, X_{i2}, \dots, X_{in_i}\}$, 事件 $X_{ij} (i \neq j)$ 发生的概率为 $p_{ij} (i \neq j)$, 产生的信息量为 $H_{ij} = -\log_2 p_{ij}$, 则此概率系统的平均信息量定义如下:

$$H_m(i) = \frac{k_i'}{\sum_{i=1}^n k_i} \beta_i \sum_{j=1(i \neq j)}^{n_i} p_{ij} \log_2 \frac{1}{\beta_i p_{ij}} \quad (1)$$

式中, n_i 表示经过 t 次合并后网络中剩余的节点数。

$\frac{k_i'}{\sum_{i=1}^n k_i}$ 为节点全局重要度,是节点度与网络总度数的

比值。由于合并时新节点吸收了原节点的邻居节点,在社团扩张过程中属于强吸力节点,更不容易被其他节点吞并,其重要度也应进行叠加。 β_i 为调节因子向量 β 的第 i 个元素,满足 $\beta_i \sum_{j=1}^n p_{ij} = 1$ 。事件 $X_{ij} (i \neq j)$ 表示节点 v_i 与节点 v_j 进行合并,一般来说,两个节点拥有越高的相似度,它们合并的可能性就越大。采用总体性能最佳的 RA 指标^[6] 进行合并概率计算,其定义如下:

节点 v_i 把 $\{\Gamma_i \cap \Gamma_j\}$ 当成枢纽向节点 v_j 传递信息,规定每个枢纽可容纳的资源是一定的,且每个枢纽发送给 v_j 的资源根据其度值进行平均分配,则节

点 v_i 和 v_j 的相似度为 v_j 能够接收到的资源数,即

$$p_{ij} = S_{ij}^{\text{RA}} = \sum_{z \in \Gamma_i \cap \Gamma_j} \frac{1}{k_z} \quad (2)$$

类似地,得到整个网络在 t 次合并后的合并信息熵模型为 $\{H_m(1), H_m(2), \dots, H_m(n_t)\}$, 由于每个节点产生合并概率的事件个数不同,为了使其在同一标准下进行判断,需要将各节点的合并信息熵进行归一化处理,如下式所示:

$$H_{\text{mnor}}(i) = \frac{H_m(i)}{\log_2 \text{Num}(p_{ij} \neq 0)} \quad (3)$$

式中, $H_{\text{mnor}}(i)$ 表示归一化后节点 v_i 的合并信息熵, $\text{Num}(p_{ij} \neq 0)$ 为其合并概率集合中非零元素的个数。

3 MIE 算法

Shannon 用信息熵来描述信源的平均不确定性,当所有可能的事件发生概率一样时,信源下一时刻产生哪个符号最难预测,表明此系统具有最大平均不确定度。将此概念引入到 MIE 算法中,假设经过 t 次合并后网络的合并信息熵模型为 $\{H_{\text{mnor}}(1), H_{\text{mnor}}(2), \dots, H_{\text{mnor}}(n_t)\}$, 其中节点 v_i 具有最小的 H_{mnor} 值,则其合并概率动态系统在全局具有最小不确定度,即合并概率值的取值范围跨度最大,此时选择 $p_{ij\text{max}}$ 进行合并能提供最小的信息量,因为它在整个系统中相当于一个几乎必然发生的事件。

MIE 算法使用模块度 Q 函数作为评定社团划分质量的标准^[7]。假设网络被划分为 c 个社团,定义一个对称矩阵 $\mathbf{B} = (e_{uv})_{c \times c}$, 其元素 e_{uv} 代表社团 u 和 w 之间的连边占整个网络边数的比例,则一端与社团 u 中节点相连的边数的比例为 $a_u = \sum_{w=1}^c e_{uw}$ 。模块度定义为社团内部连边数与对应零模型连边数的差值占网络总边数 M 的比例,即:

$$Q = \sum_{u=1}^c (e_{uu} - a_u^2) = \text{Tre} - ||\mathbf{B}^2|| \quad (4)$$

式中, $\text{Tre} = \sum_{u=1}^c e_{uu}$ 代表矩阵 $\mathbf{B}$ 对角线上各元素之和, $||\mathbf{B}^2||$ 为对矩阵 $\mathbf{B}^2$ 的所有元素进行求和。

理论上来说,模块度值越高,社团结构的质量越好,真实网络模块度值通常在 $0.3 \sim 0.7$ 范围内^[8]。

3.1 算法具体描述

输入:一个无权无向图 $G = (V, E)$;

节点合并信息熵阈值 α 。

输出:算法划分出的社团结构。

步骤1 计算节点之间的共同邻居集合 $\Gamma(i) \cap$

$\Gamma(j)(i, j = 1, 2, \dots, n_i, i \neq j, n_{i0} = N)$, 按照式(2) 计算 $P_{ij}(i \neq j)$, 构建各节点的合并概率动态系统 $\{X_{i1}, X_{i2}, \dots, X_{ij}, \dots, X_{in_i}, i \neq j\}$.

步骤 2 按照式(1) 计算各节点的合并信息熵, 并按照式(3) 进行归一化处理, 得到整个网络的合并信息熵模型 $\{H_{\text{minor}}(1), H_{\text{minor}}(2), \dots, H_{\text{minor}}(n_t)\}$.

步骤 3 选取合并信息熵最小的节点, 比较其熵值与阈值 α 的大小, 若满足 $H_{\text{minor}}(i)_{\min} \leq \alpha$, 则执行 $P_{ij\max}$ 对应合并概率动态系统中的事件 X_{ij} , 否则将其合并到与连边最多的社团中.

步骤 4 将合并后的社团收缩为新的节点, 标记节点所属社团, 同时将节点重要度进行叠加, 计算模块度 Q .

步骤 5 重复执行步骤 1 ~ 4, 直到所有的节点都处理完毕.

步骤 6 输出具有最大模块度的社团结构 $C(Q_{\max})$.

需要说明的是, 为了有效提高社团结构检测的准确率, 引入合并信息熵阈值 α 参数, 若所有节点的 H_{minor} 均不小于 α , 则将其合并到连边最多的社团中. 因为如果 H_{minor} 很大, 节点合并概率系统的不确定度就大, 无法判断此节点更应该与哪个节点进行合并, 仍旧将此属性当作划分依据易导致随机性错误, 从而影响社团划分质量.

3.2 时间复杂度分析

MIE 算法的计算主要包括两个部分: 计算节点合并信息熵和节点合并收缩. 其中, 计算节点合并信息熵重点在于计算节点合并概率, 而计算节点合并概率只需获得其二级以内的邻居节点信息. 若用 $\langle k \rangle$ 表示网络中节点的总平均度, 计算一个节点与其 $\langle k \rangle$ 个邻居节点相似度的时间复杂度为 $O(\langle k \rangle)$, 计算该节点与其二级邻居节点相似度的时间复杂度为 $O(\langle k \rangle^2)$, 则包含 N 个节点的网络所需要的时间复杂度为 $O(\langle k \rangle + N \langle k \rangle^2)$. 若每个节点都进行一次合并收缩, 则时间复杂度可近似看作 $O(N)$. 通过以上分析可知, MIE 算法的时间复杂度为 $O[N(\langle k \rangle + N \langle k \rangle^2)] \approx O(N^2)$

$(\langle k \rangle \ll N)$. 而一些经典算法, 例如谱平分法^[9-10]的时间复杂度为 $O(N^3)$, FN 算法^[11]的时间复杂度为 $O[(M+N)N]$, GN 算法^[12]的时间复杂度为 $O(M^2N)$, 不适用于一些较为密集的大型网络, 所以本算法具有一定的优势.

4 实验结果与分析

为了验证 MIE 算法用于社团结构检测的性能

及有效性, 首先用三社团网络分析其可行性, 而后选取 3 个被广泛使用的真实网络数据集进行测试.

4.1 可行性分析

三社团网络由 19 个节点和 37 条边组成, 是具有显著社团结构的经典验证网络. 社团划分时节点的动态收缩过程如表 1 所示, 可以看出在步骤 16 执行完毕后模块度 Q 达到最大值, 此时网络划分为三个社团, 划分结果与实际网络完全相同, 验证了本算法识别社团结构的可行性.

表 1 三社团网络社团划分动态表

步骤 i	待划分节点	合并入节点	模块度 Q_i
1	1	4	-0.058 4
2	6	4	-0.062 8
3	16	14	-0.067 2
4	18	14	-0.074 9
5	12	10	-0.080 4
6	2	4	-0.038 0
7	3	4	-0.022 6
8	15	14	0.040 9
9	13	10	0.104 5
10	5	4	0.170 9
11	19	14	0.228 6
12	7	4	0.289 3
13	17	14	0.360 8
14	8	10	0.463 1
15	11	10	0.515 0
16	9	10	0.611 8
17	14	10	0.420 4
18	4	10	0

4.2 真实网络数据集

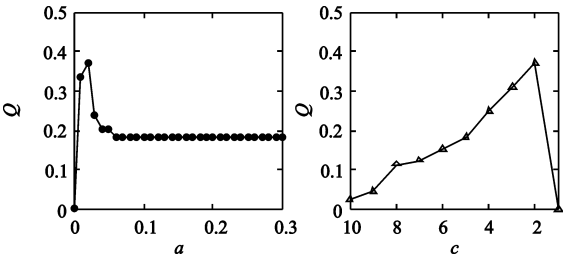
实际网络往往具有不均匀的度分布和不明显的社团结构, 因此采用 3 个真实网络做进一步验证, 数据集的基本信息如表 2 所示.

表 2 各数据集的基本信息

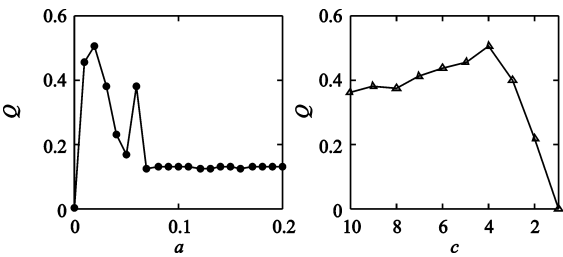
数据集	节点数	边数	平均聚类系数	平均路径长度
Karate	34	78	0.588	2.408
Dolphins	62	159	0.303	3.357
Football	115	613	0.403	2.508

用 MIE 算法分别对这三个网络进行实验, 通过调节阈值 α 得到其对应的离散模块度值, 取 Q 达到峰值时对应的 α 进行社团划分, 得到社团个数 c 与 Q 的关系曲线, 划分过程中各参量的关系如图 1 所示. 从图中可以看出不同的网络由于连接特性不同, 所

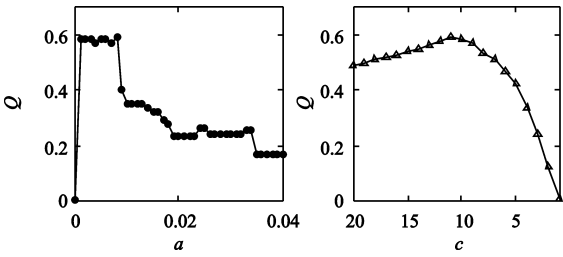
以节点合并信息熵的取值分布范围有差异,但 Q 值的变化趋势皆为先增大后减小. 与其他两个网络相比,美国大学橄榄球队网络 α 取值具有一定的区间性.



(a)空手道俱乐部网络($\alpha=0.02$)



(b)海豚社会网络($\alpha=0.02$)



(c)美国大学橄榄球队网络($\alpha=0.008$)

图 1 划分过程中的参量关系

空手道俱乐部网络的划分结果信息如表 3 所示,与俱乐部真实分裂出的两个社团完全相同.

表 3 空手道俱乐部网络社团划分结果

社团	节点
1	1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22
2	9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34

海豚社会网络的划分结果如图 2 所示,MIE 算法最终检测出了包含 4 个群落的社团结构,与真实社团结构(图中虚线分隔出的 2 个社群)对比可知,本算法将较大的群体进一步划分成 3 个小群体,检测出了更小的社团结构. 因为本算法以模块度函数作为划分过程的控制函数,所以更倾向于发现使 Q 呈增大趋势的划分方案.

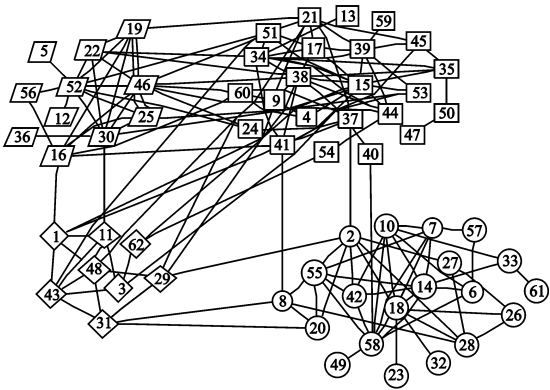


图 2 海豚社会网络划分结果

美国大学橄榄球队的划分结果如图 3 所示,与真实联盟的球队信息对比,发现一共有 9 个节点(29,37,43,59,81,83,91,98,111)划分到其他社团. 经过分析发现,真实结构中有一个联盟有 5 个球队(37,43,81,83,91),而在该联盟内只有 81 和 83 两队进行了比赛,所以这 5 个球队在划分过程中被归类到与之进行比赛次数较多的社团中,另外几个节点亦是如此,符合 MIE 算法的合并划分过程.

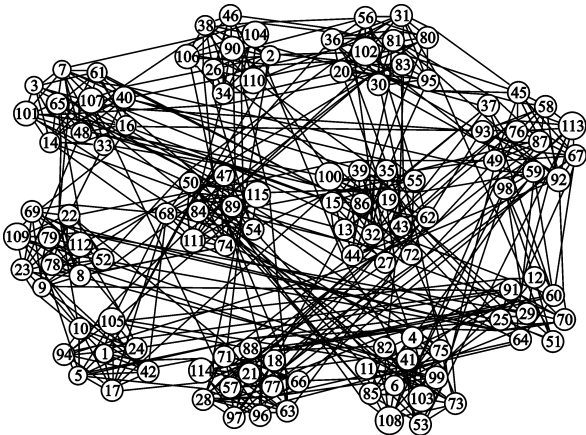


图 3 美国大学橄榄球队网络划分结果

4.3 与其他算法分析比较

表 4 是 MIE 算法与其他算法在社团个数、准确率和模块度 3 个评价指标上的结果对比,其中准确率为划分正确的节点个数与总节点数的比值. 可以看出,本算法划分社团的有效性要优于其他算法,且具有更高的划分细粒度和模块度.

经分析可知,FN 算法多次计算不同节点对合并产生的 ΔQ ,选择使 ΔQ 最大的合并方案,这种全遍历迭代使算法效率极大降低. NC 算法^[13]引入节点相异度解决了初始核心节点的位置依赖问题,但划分时仍采用全局优化模块度的策略,具有较高的

表 4 各算法社团划分结果比较

数据集		FN	NC	IPLPA	MIE
Karate	CN	2.000	4.000	3.000	2.000
	A	0.971	0.971	0.941	1.000
	Q	0.381	0.402	0.382	0.372
Dolphins	CN	2.000	4.000	4.000	4.000
	A	0.952	0.887	0.774	1.000
	Q	0.385	0.544	0.451	0.508
Football	CN	6.000	7.000	11.000	11.000
	A	0.783	0.870	0.862	0.922
	Q	0.546	0.552	0.483	0.592

注:CN 表示社团个数 (Community Number); A 表示划分准确率 (Accuracy); Q 表示模块度 (Modularity).

时间复杂度 $O[(M+N)M]$. 而 IPLPA 算法^[14] 尽管时间复杂度较低, 但仅依靠节点重要度决定标签传播迭代顺序具有局限性, 导致准确率较低. 本文提出的算法将节点在整个网络中的重要度及合并概率作为计算合并信息熵的依据, 计算过程涵盖了局部信息和全局信息. 在社团结构检测过程中, 利用节点合并收缩与邻居节点的重连接进一步细化网络结构, 具有较低的算法复杂度, 适用范围较广.

5 结束语

本文基于节点相似度的凝聚思想, 在算法中引入熵的概念构造了节点合并信息熵模型, 充分运用节点本身及其连接特性, 通过测量合并概率事件系统所具有的不确定度, 快速找到最优的社团划分方案. 通过在真实网络数据集上与 3 个经典划分算法的实验对比, 验证了本文算法对于社团结构检测的优势, 且具有较高的划分细粒度. 在以后的研究中, 考虑将概率事件动态系统与动态网络相结合, 实现更加实时有效的社团划分.

参考文献:

[1] 李超逸, 张仰森, 佟玲玲. 一种基于社区发现的微博个性化推荐算法[J]. 微电子学与计算机, 2017, 34(6):40-44.

[2] 张姣, 刘三阳, 白艺光. 基于社团结构的组合信息重连策略[J]. 复杂系统与复杂性科学, 2019, 16(2): 1-8.

[3] CHEN L J, ZHANG Z K, LIU J H, et al. A vertex similarity index for better personalized recommendation [J]. Physica A: Statistical Mechanics and its Applications, 2017(466):607-615.

[4] 张虎, 吴永科, 杨陟卓, 等. 基于多层节点相似度的社区发现方法[J]. 计算机科学, 2018, 45(1): 216-222.

[5] 丁明珠, 马英红, 李云. 基于相似度指标的社团划分算法[J]. 计算机工程, 2019, 45(2):201-207.

[6] ZHOU T, LINYUAN LÜ, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4):623-630.

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

[8] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6):066133.

[9] DONETTI L, MUNOZ M A. Improved spectral algorithm for the detection of network communities[C]// 8th International Conference on Modeling Cooperative Behavior in the Social Sciences. New York, USA, 2005:104-107.

[10] SHIGA M, TAKIGAWA I, MAMITSUKA H. A spectral clustering approach to optimally combining numerical vectors with a modular network[C]//13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2007: 647-656.

[11] GIRVAN M, NEWMAN M E 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.

[12] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6):066133.

[13] 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[J]. 计算机工程, 2014, 40(1): 275-279.

[14] 邓凯旋, 陈鸿昶, 黄瑞阳. 基于标签传播能力的改进 LPA 算法[J]. 计算机工程, 2018, 44(3):60-64.

作者简介:

涂佩文 女, (1996-), 硕士研究生. 研究方向为复杂网络、绿色网络.

周金和(通讯作者) 男, (1966-), 硕士, 教授. 研究方向为绿色网络与通信、无线通信信号处理.

E-mail:zhoujinhe@bistu.edu.cn.