基于核心度和偏移量的社区检测算法

辛慧英,刘向阳
(河海大学 理学院,江苏 南京 211100)

摘 要:为减少社区检测算法中大量中间结果的计算对社区划分的影响,同时能够准确检测到网络的社区划分以及网络的核心社区,提出了一种基于核心度和偏移量的社区检测算法,其中核心度和偏移量定义了任意节点作为社区核心的程度。首先针对复杂网络的邻接矩阵,应用广度优先搜索算法计算网络中节点之间的边介数,基于边介数确定网络中每条边的权值,计算得到网络的加权邻接矩阵及全局距离矩阵;然后计算网络节点的核心度和偏移量,来确定社区的核心节点和核心社区;最后对其余节点进行划分以完成社区检测。在数据集 Karate、Dolphins、Football 上的实验结果表明,该算法具有很好的稳定性,并且可以很好地检测出社区结构,相比其他的方法,该算法复杂度更低,计算量更少,更高效。

关键词:边介数;距离矩阵;核心度;偏移量;核心社区
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2020)10-0037-05
doi:10.3969/j.issn.1673-629X.2020.10.007

Community Detection Algorithm Based on Core Degree and Distance

XIN Hui-ying,LIU Xiang-yang
(School of Science,Hohai University,Nanjing 211100,China)

Abstract:In order to reduce the impact of the calculation of a large number of intermediate results in the community detection algorithms on community partitioning,and to accurately detect the community division of the network and the core community of the network,we propose a community detection algorithm based on core degree and distance which define the degree to which any node is the core of the community. Firstly,based on the adjacency matrix of the complex network,the breadth-first search algorithm is applied to calculate the betweenness in the network. The weight of each edge is determined based on betweenness,and the weighted adjacency matrix and global distance matrix of the network are calculated. Then,the core degree and distance of the network node are calculated to determine the core nodes and core communities. Finally,the remaining nodes are dispatched to complete the community detection. The experimental results on the datasets Karate,Dolphins,and Football show that the proposed algorithm can well detect the community structure with high stability. Compared with other methods,it has lower complexity,less calculation and more efficiency.

Key words:betweenness;distance matrix;core degree;distance;core community

0 引 言

在复杂网络中,社区可以表示具有相似性、共同特点或关系的共同体,它具有“社区内部节点之间连接比较紧密,与外部节点连接相对稀疏”的特点。对复杂网络进行社区检测可以清晰地认识到社区内部的结构及组织,明确网络社区结构,有助于获得更多网络未知区域的信息,在物理、生物、计算机科学等诸多领域应用广泛。

Girvan 和 Newman 在 2002 年提出了社区检测第一算法:基于模块度优化的划分算法[1](GN),该算法奠定了社区检测研究领域的基石,但该算法需要多次计算节点之间的边介数,大大增加了计算复杂度。而后,相继提出的一些基于标签传播的社区检测算法,基于信息流动的算法[2-3],基于优化块模型的算法[4-5],基于动态随机游走的算法[6]以及基于网络拓扑结构的算法[7]都在社区检测中得到了广泛的应用。在一定意义上,社区检测就是某种的聚类,这两者的区别在于:社区检测算法中可能存在孤立点,而聚类一般假设任意对象都是互相连接的,只是相似度不同,数据集可以表示为一张完全连通图,如:层次,密度聚类等。但这两个过程非常相似,因此产生许多将聚类算法的思想应用于社区检测算法的例子。黄岚等[8]通过网络中

收稿日期:2019-10-24 **修回日期**:2020-02-25
基金项目:国家自然科学基金(61001139)
作者简介:辛慧英(1995-),女,硕士研究生,研究方向为复杂网络、社区划分;通讯作者:刘向阳(1977-),男,副教授,博士,研究方向为复杂网络分析、数据分析和机器学习。

节点的相似度来定义节点间的距离,将密度峰值聚类算法[9]应用于社区检测中;文献[10]引入箱线图来确定核心节点,将密度峰值聚类算法应用于重叠社区检测中;文献[11]应用启发式算法来划分社区的自然结构;郭玉全[12-13]等以两阶段盒子覆盖法为基础,进一步提出分形聚类检测方法[14-15],核心思想是通过两阶段盒子覆盖法来完成分形聚类过程,并且在分形聚类过程中形成分形树,最后通过对分形树进行分割进而得到复杂网络的社区结构。上述算法虽然成功将聚类算法引入到社区检测中,但是却由此引入一些新的参数,并且每个参数的选择都需要大量的实验数据来获取,大大增加了算法的复杂度,通常只对一种或某几种特定网络的结果较好,无法适应性地对多种社区进行划分,算法性能存在不足。

针对不依赖于大量实验数据获得参数的方法,文中提出一种基于偏移量和核心度的社区检测算法。首先求出网络中节点之间的距离矩阵,然后给出核心度、偏移量、核心社区的概念,并求出节点的核心度,来确定社区核心和核心社区,最后对其余节点进行分派,进而确定社区划分结果。该方法的优点在于,首先不需要过多的参数,不需要多次计算节点之间的边介数,其次不仅可以给出社区划分,而且可以很好地确定核心社区。实验结果充分验证了该方法的高效性和稳定性。

1 算法思想

社区检测算法的一个基本假设是:社区内部节点的相似度比较高,社区间节点相似度比较低;也可以理解为:社区内部节点间的距离比较小,社区间节点的距离比较大。基于节点间的距离求出节点的核心度,基于核心度确定社区核心,假设:(1)社区核心的核心度比它周围邻居节点的核心度高;(2)社区核心比其他任意核心度更高节点的相对偏移量更大。该算法不需要重复大量的计算,只需要一次计算,即可得出复杂网络的社区划分。算法流程如图 1 所示。

2 社区检测算法

社区检测算法基于两种基本的思想,一种是凝聚算法(agglomerative methods),这类算法是从一个个孤立的节点开始,计算任意两个节点的相似度,相似度高则这两个节点在同一个社区。另一种社区检测的思想是分裂算法(divisive methods),原始的输入是一个完整的网络图,要做得就是基于节点之间边的某些特性来删除图中的一些边。要删除的边是基于 Newman[16] 提出的删除具有最大边介数的边。文中具体思路的灵感来自于此,并不是要去删除节点之间的边,但是在最初完整的网路中计算边介数作为每条边的权重,基于完整网络中每条边的权重,计算提出的加权邻接矩阵。

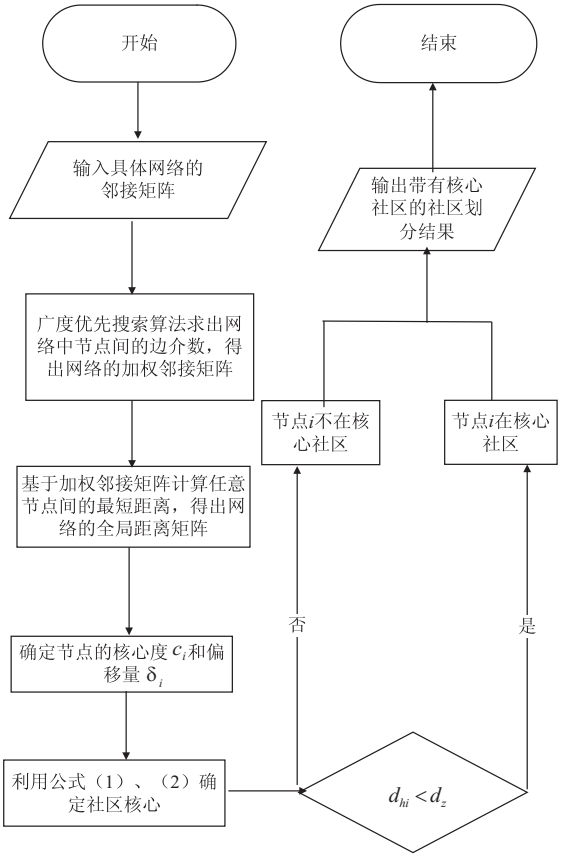

图 1 算法流程

2.1 计算加权邻接矩阵

首先计算网络中每个节点的权重 w,节点权重为从源点 s 到该节点最短路径的数目:

(1)将初始节点 s 的 w 值设为 1;

(2)每个与 s 相邻的节点 i 的权重 w 也设为 1;

(3)对于每一个与(2)中节点 i 相邻的节点 j,有如下计算规则:

(a)若 j 的权重还未设置,令 $w(j) = w(i)$;

(b)若 $w(j)$ 已经存在,则令 $w(j) = w(i) + w(j)$。

计算得到每个节点的权重之后,计算每条边的边介数,自底向上,对边界的节点:即与之直接相连的边数为 1,那么设该边介数为 1;对于一个节点有多条边与之直接相连,则边介数为 $w(i)/w(j)$,其中 i 是上方的节点。对于上层节点:其与更高一层节点连接的边介数等于所有与之直接相邻的子边介数之和加 1,(如果该节点只有一个与之直接相连的上层节点),或等于所有与之直接相邻的子边介数之和加 1 乘 $1/n$(如果该节点有 n 个直接与之相连的上层节点)。计算完所有边介数,将以此作为节点间边的权重,计算得出网络的加权邻接矩阵。

2.2 全局距离矩阵

全局距离矩阵用来描述任意节点之间的最短距

离,应用广度优先搜索算法求出复杂网络中的边介数作为边的权值,基于边的权值得到复杂网络的加权邻接矩阵,计算得到任意节点间的最短路径和最短距离。全局距离矩阵的计算步骤如下:

(1) 基于加权邻接矩阵生成网络的无向图;

(2) 基于加权无向图,用函数计算任意两点的最短路径和最短距离;

(3) 由任意两点的最短距离构成网络的全局距离矩阵。

2.3 节点的核心度和偏移量

2014 年 Alex Rodriguez[9] 在新聚类算法中提出了局部密度和"距离"的概念,基于此,文中提出将网络中节点的核心度以及偏移量应用于复杂网络的社区检测。核心度和偏移量定义了任意节点作为社区核心的程度。

采用高斯核函数计算核心度,这样计算不同的节点具有相同的核心度的概率更小。计算公式为:

$$c_i = \sum_{j \neq i} \exp(-(d_{ij}/d_c)^2) \tag{1}$$

其中, $\exp(.)$ 为指数函数, d_c 为截断距离,是算法中的可变参数,实验时取网络中所有节点间距离的 1% ~2%, d_{ij} 为网络中节点之间的距离。当计算得到核心度 c_i 后,让核心度按从大到小的顺序进行排列,即 $\{c_1, c_2, \cdots, c_m\}$。

节点的偏移量计算公式为:

$$\delta_i = \begin{cases} \max\limits_{j}(d_{ij}), i = 1 \\ \min\limits_{j:\rho_j > \rho_i}(d_{ij}), i > 1 \end{cases} \tag{2}$$

即当 x_i 的核心度最大时, δ_i 表示网络中与 x_i 距离最大的节点到 x_i 之间的距离;否则, δ_i 表示在所有核心度大于 x_i 的节点中,与 x_i 距离最小的那个节点到 x_i 之间的距离。

2.4 社区检测

2.4.1 确定社区核心和核心社区

根据假设,社区核心为核心度比它周围邻居节点的核心度高且与任意其他核心度更高节点的偏移量相对更大的节点,由公式(1)和公式(2),可以计算得出网络中节点的核心度和偏移量,从而确定社区核心。核心社区"C"为社区中距离社区核心的距离相对比较小的节点集,确定核心社区的计算公式为:

$$C = \begin{cases} i \in C, d_{hi} - d_z \leqslant 0 \\ i \notin C, d_{hi} - d_z > 0 \end{cases} \tag{3}$$

其中, d_z 是核心社区内节点距离社区核心的距离阈值,实验中取网络中所有节点间距离的 50% ~60%, d_{hi} 为网络中节点与社区核心的距离。

2.4.2 对其他节点的处理

将其余节点按照核心度大小依次归类到比它们核心度更大,节点之间相似性更大的节点所属的类别。对社区边界的节点,首先定义社区边界区域,即分配到该社区又与其他社区节点的距离小于截断距离的节点的集合,然后为每个社区找到其边界区域中平均核心度的最大值,并以这个核心度作为阈值来筛选节点,对不满足该阈值的节点,排除在该社区之外。至此,即可完成社区划分。

3 实验与分析

为了验证算法的有效性、可行性和稳定性,选取了数据集 Karate、Dolphins、Football 进行实验。

3.1 实验数据集描述

选取三个真实数据集进行实验,数据集中的节点之间有一定的社会关系,所以这三个数据集都具有已知的社区结构,分别为空手道数据集(Karate)、海豚数据集(Dolphins)、美国大学橄榄球数据集(Football)。Karate 数据集是由 34 个节点和 78 条边组成,其中每个空手道俱乐部成员代表一个节点,如果两个节点之间存在边,则表示相对应的成员之间联系交往密切,俱乐部会长和俱乐部的教练由于一些个人原因,他们之间存在一些分歧,所以最终俱乐部成员们形成了两个小团体(俱乐部)。Dolphins 数据集是由 62 个节点和 159 条边组成,每只生活在新西兰 Doubtful Sound 海峡的宽吻海豚代表一个节点,如果两个节点之间存在边,那么代表对应的两只海豚有非常多的交流,经过很长一段时间的跟踪观察及记录,发现可以将这些海豚分为两个社区,但是更仔细观察分类,它们可以被分为 5 个社区。Football 数据集是由 115 个节点和 609 条边组成,每支参加 2000 年美国大学秋季学期橄榄球赛的球队代表一个节点,如果两支球队曾经有过一场比赛,则表示两个节点之间存在边。这些球队隶属于 12 个不同的球会,且在球会内部进行的比赛比较多。

3.2 实验结果和分析

将文中算法在三个真实数据集上进行实验,图 2 ~图 5 为文中算法在 Karate、Dolphins 数据集上的划分结果以及带有核心社区的社区划分结果,此时都取 d_z 为网络中所有节点间距离的 59%。图 6 为 Dolphins 数据集更仔细地被划分为 5 个社区的情况。图 7 为文中算法在 Football 数据集上的划分结果。

图 2、图 3 表明,文中算法几乎完美地将 Karate 数据集的 34 个节点分为两个社区,只有一个节点(Karate 数据集中的 3 号节点)被错误地分类,但该节点是在群体之间的边界上,所以可能是一个模糊的情况,是可以理解的。文中算法可以找出核心社区,在图 3 中,标签为 3 的节点和标签为 4 的节点分别为两个社区的核心社区,这些节点在网络中占据重要的地位,

发现并准确定位这些节点将有很大的现实意义。

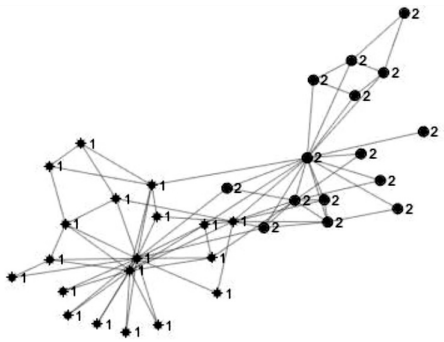

图 2 Karate 数据集社区划分结果
(34 个节点被分为两个社区)

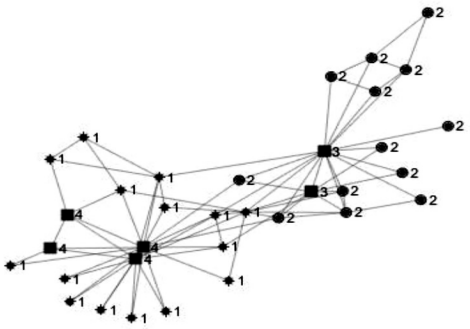

图 3 Karate 数据集带有核心社区的社区划分结果
(其中标签为 3 和 4 的节点分别代表
两个社区的核心部分)

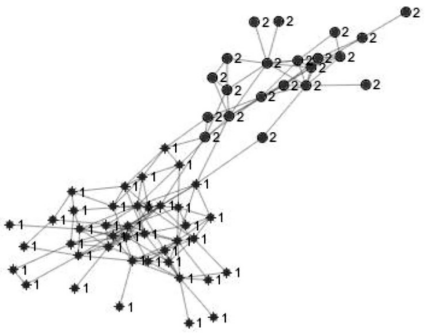

图 4 Dolphins 数据集社区划分结果
(62 个节点分为两个社区)

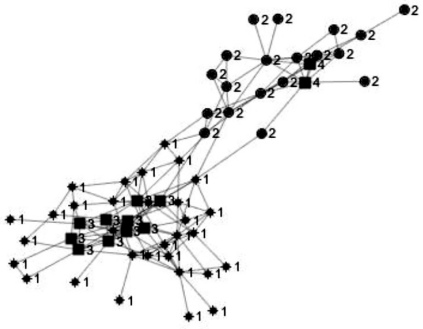

图 5 Dolphins 数据集带有核心社区的社区划分结果
(其中标签为 3 和 4 的节点分别代表
两个社区的核心部分)

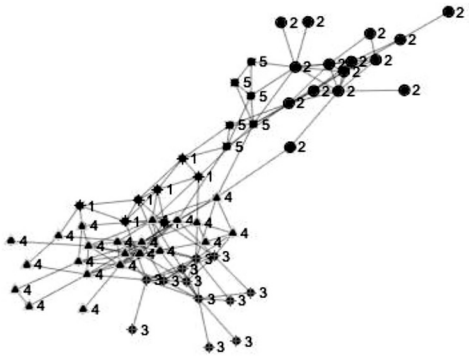

图 6 Dolphins 数据集社区划分结果
(62 个节点更仔细地分,可以分为 5 个社区)

图 4 ~ 图 6 表明,文中算法可以准确将 Dolphins 数据集划分为 2 个社区,社区间有明显的界限,并且进一步划分,Dolphins 数据集可以被划分为 5 个社区。文中算法可以找出核心社区,在图 5 中,标签为 3 的节点和标签为 4 的节点分别为两个社区的核心社区。

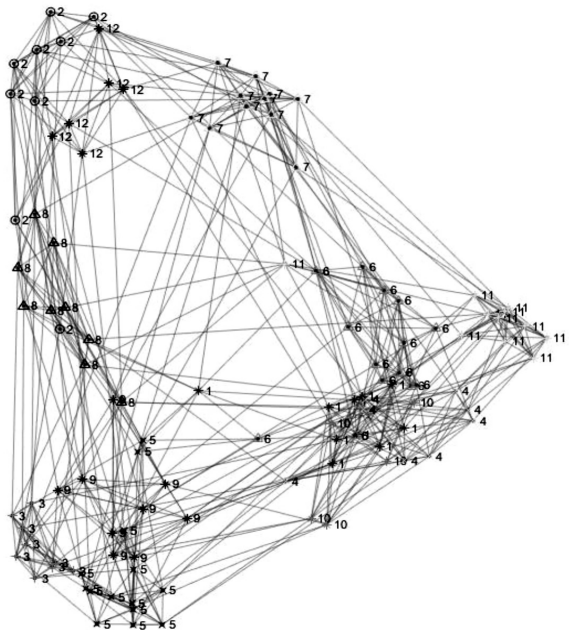

图 7 Football 数据集社区划分结果
(115 个节点被分为 12 个社区)

图 7 表明,文中算法将 Football 数据集中 115 个节点分为 12 个社区,图 7 中不同的节点标签分别代表不同的社区。可以看出,文中算法发现的社区数目与真实的社区数目相同,并且可视化效果良好,只有少数几个节点游离在社区之间。

4 结束语

提出一种基于节点核心度和偏移量进行社区检测的算法,在三个真实数据集上的实验结果表明,该算法简洁、高效、准确,算法运行良好,不仅可以准确地对数据集 Karate、Dolphins、Football 进行社区检测,而且可

以找到核心社区。相比于一般的社区发现算法只能发现社区的一般整体划分,该算法可以检测出社区核心和核心社区,基于大多数网络具有拓扑结构的特性,核心社区的发现具有很好的现实意义,未来将尝试应用于更加复杂的网络。

参考文献:

[1] 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.

[2] VROSVALL M,BERGSTROM C T. An information theoretic framework for resolving community structure in complex networks[J]. Proceedings of the National Academy of Sciences,2007,104(18):7327-7331.

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

[4] XU Baochang,ZHANG Yingying. An improved gravitational search algorithm for dynamic neural network identification[J]. International Journal of Automation and Computing,2014,11(4):434-440.

[5] NAENI L M,BERRETTA R,MOSCATO P. Ma-net:a reliable memetic algorithm for community detection by modularity optimization[C]//Proceedings of the 18th Asia Pacific symposium on intelligent and evolutionary systems,volume 1. [s. l.]:Springer International Publishing,2015:311-323.

[6] PONS P,LATAPY M. Computing communities in large networks using random walks[M]. Berlin Heidelberg:Springer,2005:284-293.

[7] LANCICHINETTI A,RADICCHI F,RAMASCO J J,et al. Finding statistically significant communities in networks[J]. PloS One,2011,6(4):e18961.

[8] 黄　岚,李　玉,王贵参,等. 基于点距离和密度峰值聚类的社区发现方法[J]. 吉林大学学报:工学版,2016,46(6):2042-2051.

[9] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.

[10] WANG Guisheng,HUANG Lan,WANG Yan,et al. Link community detection based on line graphs with a novel link similarity measure[J]. International Journal of Modern Physics B,2016,30(6):1650023.

[11] HENNIG C,HAUSDORF B. Design of dissimilarity measures:a new dissimilarity measure between species distribution areas[M]//Data science and classification,studies in classification,data analysis,and knowledge organization. Berlin,Germany:Springer,2006:29-38.

[12] 郭玉泉,李雄飞. 复杂网络社区的分形聚类检测方法[J]. 吉林大学学报:工学版,2016,46(5):1633-1638.

[13] 王冰玉,吴振宇,沈苏彬. 一种社交网络的增量社区检测算法及实现优化[J]. 计算机技术与发展,2018,28(10):64-69.

[14] 孙延维,彭智明,李健波. 基于粒子群优化与模糊聚类的社区发现算法[J]. 重庆邮电大学学报:自然科学版,2015,27(5):660-666.

[15] 陈新泉. 基于单元网格近邻势的聚类方法[J]. 重庆邮电大学学报:自然科学版,2014,26(6):771-777.

[16] NEWMAN M E J,GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E,2004,69(2):026113-1-026113-13.