

一种基于谱平分法的社团划分算法

谢福鼎 张 磊 嵇 敏 黄 丹

(辽宁师范大学计算机与信息技术学院 大连 116081)

摘 要 基于改进的 SNN 相似度矩阵与谱平分法,提出了一种寻找复杂网络社团结构的算法。首先计算出网络中各节点之间改进的 SNN 矩阵并将其标准化,求得该矩阵的特征值及特征向量。然后分别选取不同数目的第一非平凡特征向量作为聚类样本,利用 FCM 聚类算法对节点进行分类,并计算出每次分类结果所对应的模块度 Q 值。 Q 的最大值对应的社团结构即为最佳的网络社团结构。一些实验测试了该方法的可行性,通过与其它方法的结果进行比较,可知该算法划分社团的准确率较高。

关键词 复杂网络,社团结构,SNN 相似度矩阵,谱平分法,FCM 算法

中图分类号 TP181 文献标识码 A

Community Partitioning Algorithm Based on Spectral Bisection Method

XIE Fu-ding ZHANG Lei JI Min HUANG Dan

(College of Computer Science and Information Technology, Liaoning Normal University, Dalian 116081, China)

Abstract Based on the improved SNN similarity matrix and spectral bisection method, this paper proposed a new algorithm for detecting the community structure in complex networks. The improved SNN similarity matrix was firstly computed and normalized and its eigenvalues and eigenvectors were obtained subsequently. Then different numbers of the first non-trivial eigenvectors were chosen as clustering samples, FCM algorithm began to work and the corresponding modularity was computed. The best structure of the network was detected by mapping the largest value of modularity. The experiment shows the validity of the presented method. The result obtained here is compared with other popular ones and the conclusion is that the accuracy of the results calculated by this approach is much better than the known ones.

Keywords Complex networks, Community structure, SNN similarity matrix, Spectral bisection method, FCM algorithm

1 引言

由于复杂网络可以用来描述人与人之间的社会关系、物种之间的捕食关系、计算机之间的网络连接、词与词之间的语义联系、科学家之间的合作关系以及网页的链接结构等,因此,近年来,其研究受到越来越多的关注。随着对网络性质的物理意义和数学特性的深入研究,人们发现许多实际网络都具有一个共同性质——社团结构。也就是说,整个网络是由若干个“群”构成的,每个群内部的节点之间的连接非常紧密,但是各个群之间的链接却相对来说比较稀疏^[1]。揭示网络中的社团结构,对于了解网络结构与分析网络特性具有极为重要的意义。社团结构分析在生物学、物理学、计算机图形学和社会学中都有广泛的应用。

为了寻找复杂网络的社团结构,许多学者给出了很多好的方法,其中比较具有代表性的有:Kernighan-Lin 算法^[2]、基于谱平分的方法^[3,4]、分裂算法^[5]中的 GN 算法^[6]、凝聚算法^[5]中的 Newman 快速算法^[7]和派系过滤算法^[8]等。不管是分裂算法还是凝聚算法,最终目的都是将网络划分为若干个相互独立的社团。但是,现实中很多网络并不存在绝对的

彼此独立的社团结构,相反,它们都是由许多彼此相互关联的社团构成的。在社团结构不是十分明显的情况下,已有的很多算法对这类网络无法得到比较好的划分效果。所以,如何划分社团结构不明显的复杂网络仍是一个富有挑战性的研究问题。

本文提出了一种新的社团划分算法,该算法将 SNN 相似度矩阵与谱平分法思想相结合,进而发现复杂网络的社团结构。经实验证明,该算法对社团结构不明显的网络具有较好的分类效果,并且与目前比较流行的社团发现算法进行比较发现,利用本文算法划分得到的结果准确率较高。

2 预备知识

2.1 谱平分法

无向连通图 G 的 Laplace 矩阵是一个实对称矩阵 L , $L_{ij} = -1$, 如果节点 i 和节点 j 有边相连, 否则为 0, 对角线上的元素 L_{ii} 表示节点 i 的度。由于 Laplace 矩阵所有的行或列的和都为 0, 因此, 该矩阵总有一个特征值为 0, 且其对应的特征向量为 1。可以从理论上证明, 不为零的特征值对应的特征向量的各元素中, 同一社团内的节点对应的元素是近似相等的, 这

到稿日期:2008-12-09 返修日期:2009-03-06 本文受国家自然科学基金(10771092)和‘973’项目(2004CB318000)资助。

谢福鼎(1965—),男,教授,主要研究方向为人工智能、复杂网络、计算机代数, E-mail: xiefd@sohu.com; 张 磊(1984—),女,硕士生,主要研究方向为人工智能; 嵇 敏(1971—),女,讲师,主要研究方向为人工智能; 黄 丹(1970—),女,讲师,主要研究方向为复杂网络、计算机通信。

是谱平分的理论基础^[9,10]。因此,可以根据复杂网络的Laplace矩阵次小特征值 λ_2 对应的特征向量,所有正元素对应的那些节点都属于同一社团,而所有的负元素对应的那些节点则属于另一社团,将其分为两个社团。

谱平分法对于社团结构非常明显的网络十分有效,但当网络的社团结构不是很明显时,第一非平凡特征向量就不具有非常明显的阶梯状,此时,通过研究第一非平凡特征向量中的元素来对网络进行划分,往往不能得到理想的结构。由于社团内部节点联系十分紧密,因此,在每个特征向量内,同一个社团内部的节点相应的元素仍然比较接近。根据这一性质,通过比较多个非平凡特征向量中各节点相应元素的分布,并借助于相似度矩阵,给出了一个解决该问题的算法。与已有的结果相比,该算法的结果有比较高的准确度。

2.2 SNN 相似度矩阵

共享最近邻(shared nearest neighbor, SNN)相似度多用于多密度数据集的聚类。用 $N_i = \{n_j | a_{ij} = 1\}$ 表示节点 i 的邻居集合,其中若 i, j 两节点有边相连, $a_{ij} = 1$,否则 $a_{ij} = 0$ 。对于任意一个网络,都可以计算它的SNN相似度矩阵 S ,其中

$$s_{ij} = \begin{cases} |N_i \cap N_j| & i \neq j \\ 0 & i = j \end{cases}, \text{即节点 } i \text{ 和节点 } j \text{ 共享近邻的个数。}$$

SNN相似度反映了数据空间中点的局部结构。由于许多节点之间的SNN相似度为0,因此SNN相似度矩阵是很稀疏的。

若网络中两个节点的度相差较大,其共享近邻个数是一定的,显然这两个节点的SNN相似度对它们的影响程度是不同的。例如,网络中两个节点的度分别为6和16,它们之间的SNN相似度为4,该相似度对前者的影响程度要远远大于后者。因此,构造了一个改进的SNN相似度矩阵来反映这种情况。

设网络 G 有 n 个节点,首先计算网络的SNN相似度矩阵 S ,显然 S 为对称矩阵。定义改进的SNN相似度矩阵 S' , $s'_{ij} = \frac{s_{ij} a_{ij}}{k_i}$,其中 k_i 为节点 i 的度。 S' 反映了网络中任一节点与其他节点联系的紧密程度, S' 为非对称矩阵。

对 S' 进行行标准化的转换后得到矩阵 S'' ,即 $S'' = K^{-1} \cdot S'$, K' 是一个对角矩阵,其对角线上的元素 $k'_{ii} = \sum_j s'_{ij}$ 。由此可知 S'' 矩阵的最大特征值总是1,相应的特征向量为平凡特征向量。

3 一种新的社团发现算法

对于一个社团结构比较明显的网络,假设社团数目为 g ,则该矩阵有 $g-1$ 个非常接近1的第一非平凡特征值,而其他的特征值都与1有明显的差距。而且,这 $g-1$ 个特征值的特征向量也有一个非常明显的结构特征:在这 $g-1$ 个特征向量中,同一社团内的节点相应的元素非常接近。因此,只要研究这 $g-1$ 个特征向量,就可以利用它的元素将网络中节点划分为相应的 g 个社团^[11]。

已知一个包含 n 个节点的无向网络 G ,要将其划分为 g 个社团。首先计算出该网络的SNN相似度矩阵 S' 。将矩阵 S' 行标准化后,得到它的特征值和特征向量。由于其最大的特征值为1,按照与平凡特征值1的接近程度,由大到小依次选取 $g-1$ 个第一非平凡特征向量。将这些特征向量作为

聚类样本,用FCM算法进行分类,得到 $g-1$ 个网络的划分结果。最后,利用模块度 Q 来衡量社团结构的划分质量。计算这 $g-1$ 个划分结果的 Q 值,其中最大的 Q 值对应的划分结果即为该网络的最佳社团结构。

算法的基本步骤描述如下:

输入:具有 n 个节点的复杂网络,需要划分的社团个数为 g ;

输出:该网络的社团结构;

Step 1 计算网络改进的SNN相似度矩阵 S' 。

Step 2 将矩阵行标准化后,得到其特征值和特征向量。

Step 3 选取第一非平凡特征向量 λ_2 ,将其作为聚类样本利用FCM算法将节点进行分类。根据得到的划分结果,计算出对应的 Q 值。

Step 4 按照特征值由大到小的顺序 $\lambda_1 = 1 > \lambda_2 \geq \dots \geq \lambda_n$,依次增加第一非平凡特征向量的个数 D 直至 $g-1$ 个,重复步骤3。

Step 5 从计算出的 $g-1$ 个 Q 中选取最大的 Q 值,该 Q 值对应的划分结果即为原网络的最佳社团结构划分。

4 实验

4.1 Zachary 空手道俱乐部网络

利用本文提出的算法分析Zachary网络。Zachary通过观察美国一所大学中的空手道俱乐部成员之间的相互社会关系,构造了该网络。该俱乐部因是否抬高俱乐部收费的问题产生了争执,结果分裂成了两个小俱乐部。网络由34个俱乐部的成员组成,将其作为节点,78条边分别代表了成员之间的关系。

按照本文提出的算法,首先计算出 S' ,并且对其标准化,求出矩阵的特征值和特征向量。由于该网络已知分成2个社团,因此只需选取一个第一非平凡特征向量进行运算即可。利用FCM算法将其分类,所得结果对应的 Q 值为0.235。所分结果与原网络社团结构一致。分类结果如图1所示。

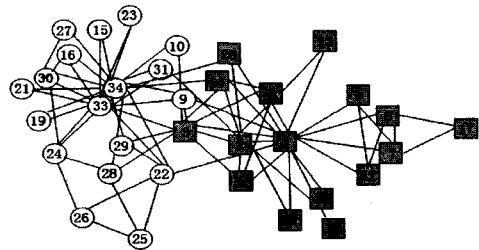


图1 Zachary研究的空手道俱乐部内部成员关系网络

4.2 23个节点构成3社团网络

利用本文算法计算图2所示的网络,其中网络中包含23个节点,该网络可以分成3个社团。

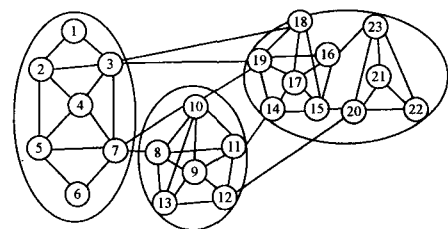


图2 由23个节点组成的3个社团网络

依照本文算法分析该网络的社团结构,结果与原网络社团结构相同。计算出相似度矩阵 S' 后,对其标准化求出矩阵的特征值和特征向量。由于该网络已知分成3个社团,因此

只需按照特征值接近 1 的程度分别选取一个和两个最接近 1 的第一非平凡特征向量,用 FCM 分别对这两次聚类样本进行分类即可。当 $D=2$ 时,对应的模块度 Q 值最大,此时 $Q=0.4116$,则其对应的分类结果与原网络社团结构相同。

4.3 足球队网络

应用本文方法对大学足球队网络进行社团结构的划分。网络中节点表示 115 支球队,613 条边表示 613 场赛事。这些球队被分成 12 个小组,每个小组有 8~12 支球队,在同一小组内的球队之间的比赛较多。

根据本文算法,首先计算出相似度矩阵 S' ,对其标准化求出矩阵的特征值和特征向量。由于网络被分为 12 个社团,因此选取 11 个第一非平凡特征向量运用 FCM 进行分类即可。选取不同个数的第一非平凡特征向量 D 对应的模块度变化如图 3 所示,横坐标表示依次选取的第一非平凡特征向量的数目,纵坐标表示利用 FCM 对选取的特征向量进行分类所得到的结果对应的 Q 值。图中实线为本文算法的结果,对应的 Q 的最大值为 0.515,虚线为利用 Laplace 矩阵代替改进的 SNN 相似度矩阵,按照本文所提出的算法进行计算,其结果对应的 Q 的最大值为 0.4。显然,利用本文算法得到的社团结构划分结果较好。

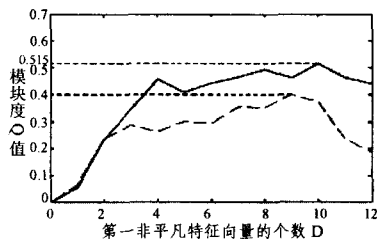


图 3 选取不同数目特征向量对应的模块度变化图

该例中,当 Q 为最大值时对应的网络社团结构划分如图 4 所示,大多数节点都与实际网络中的划分相同,只有少数节点被误分。由于节点 29,43,59,60,64,91,98,111 与同社团中节点联系较少,与同社团中节点之间的 SNN 相似度较小,因此导致被误分。

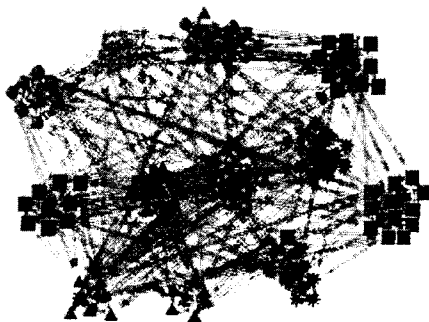


图 4 $D=10$ 时对应的最佳网络结构划分

分别利用 GN 算法^[12]、MID 方法^[13]、FCM 算法^[14]、本文算法和利用 Laplace 矩阵按照本文算法计算该例,所得的结果如表 1 所列。从表中的数据可以看出,本文算法划分社团的正确率较高。

表 1 利用不同方法分析足球队网络的结果对比

算法	社团数	模块度	正确率
GN	11	0.601	78%
MID	11	0.5985	91%
FCM	10	0.4673	90%
本文算法	12	0.515	93%
Laplace	12	0.4	70%

结束语 本文将 SNN 相似度与谱平分法的思想相结合,提出了一种新的社团划分方法。传统的谱平分法对社团结构十分明显的网络有非常好的分类效果,但是对社团结构不清晰的网络分类效果较差。谱平分法每次只能将网络平分,如果要将一个网络分成两个以上的社团,就必须对子社团多次重复这个算法。而本文算法对社团结构不明显的网络具有很好的分类效果,与目前比较流行的社团发现算法相比,准确率较高。

参考文献

- [1] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006
- [2] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49: 291-307
- [3] Fiedler M. Algebraic connectivity of graphs [J]. Czech Math J, 1973, 23: 298-305
- [4] Pothen A, Simon H, Liou K P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM J. Matrix Anal. Appl, 1990, 11: 430
- [5] Scott J. Social Network Analysis: A Handbook. London: Sage Publications [M]. 2nd ed., 2002
- [6] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proc Natl Acad Sci USA, 2002, 99: 7821-7826
- [7] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Proc. Natl. Acad. Sci., 2001, 99: 7821-7826
- [8] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435: 814-818
- [9] 解岱,汪小帆. 复杂网络中的社团结构分析算法研究综述[J]. 复杂系统与复杂性科学, 2005, 2(3): 1-12
- [10] 王林,戴冠中. 复杂网络中的社区发现——理论与应用[J]. 科技导报, 2005, 23(8): 62-66
- [11] Capocci A, Sercedio V D P, Caldarelli G, et al. Detecting communities in large networks[J]. Computer Science, 2004, 3243: 181-187
- [12] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys. Rev. E, 2004, 69: 026113
- [13] Zhang Junhua, Zhang Shihua, Zhang Xiang-sun. Detecting community structure in complex networks based on a measure of information discrepancy[J]. Physica A, 2008, 387: 1675-1682
- [14] Zhang Shi-hua, Wang Rui-sheng, Zhang Xiang-sun. Identification of overlapping community structure in complex networks using fuzzy c-means clustering[J]. Physica A, 2007, 374: 483-490