遗传聚类的社团结构发现

朱大勇[1],侯晓荣[2],张新丽[3]
(1.电子科技大学 计算机科学与工程学院,四川 成都 610054; 2.电子科技大学 自动化工程学院,四川 成都 610054; 3.成都信息工程学院 数学与信息科学系,四川 成都 610054)

摘 要:近年来在复杂网络中发现社团的结构引起了广泛的关注,目前已经提出了一些采用进化计算来分析复杂网络社团结构的方法.但大部分算法还存在处理过程复杂,空间复杂度过高等问题.通过确定网络节点的距离关系和聚类中心,提出一种新的基于遗传聚类的社团发现算法.将该算法用于真实网络的社团发现,实验结果验证了算法的可行性和有效性.

关键词:社团结构;遗传聚类;相异性指数;模块度
中图分类号: TP18 **文献标识码**: A **文章编号**: 1673-4785(2009)01-0081-04

Discovery of community structure based on genetic clustering

ZHU Da-yong[1], HOU Xiao-rong[2], ZHANG Xin-Li[3]
(1. College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China; 2. College of Automation, University of Electronic Science and Technology of China, Chengdu 610054, China; 3. Department of Math and Information, Chengdu University of Information Technology, Chengdu 610054, China)

Abstract: The discovery of community structure in complex networks has received widespread attention in recent years. Many methods based on evolutionary computation have been proposed to detect community structures in complex networks, but most of them are difficult to apply and have high degrees of space-complexity. In this paper we presented an algorithm for finding communities in complex networks using a genetic algorithm which examines distances between nodes and clustering centers. It was tested with real network datasets and the results of experiments demonstrated the feasibility of our algorithm.

Keywords: community structure; genetic clustering; dissimilarity index; modularity

社团结构的识别和发现是复杂网络的一个重要研究方向,近年来受到计算机科学、物理学和社会学等领域科研人员的广泛关注.大量的研究表明真实的复杂网络可以自然地分解为社团(communities)或模块(modules).在每个社团内部,节点之间的连接紧密,而社团与社团之间的连接相对来说较稀疏.社团结构反映了网络元素之间的拓扑关系,且对应于网络中的行为和功能单元.社团结构自动发现的研究有助于更好的理解和分析网络的结构和行为特性.

目前,已经提出了很多社团发现的算法,包括凝聚算法[1],采用介数的分裂算法[2]等等.此外,基于智能计算也提出了一些社团发现方法. Tasgin[3]应用遗传算法进行社团发现,该算法需要进行比较复杂的处理,以及需要增加额外的计算来进行纠错.文献[4]采用谱平分将网络拓扑映射为 n维空间的数据,然后采用遗传聚类的方法探测网络中的社团;但是,该算法在获得聚类中心后需要将数据进行转换才能还原为相应的社团结构.文献[5]则是采用二分法对整个网络或网络的子图进行不断的二分来发现社团,而每个染色体是包含网络所有节点的二进制串,占用空间较大.

作者提出通过网络拓扑信息确定网络节点的距离关系和聚类中心,结合模块度函数和遗传聚类分析算法进行复杂网络社团结构的分析和发现.

收稿日期: 2008-11-07.
基金项目:国家自然科学基金资助项目(NSFC-10571095);国家 973计划资助项目(NKBRPC-2004CB318003);成都信息工程学院自然科学与技术发展基金资助项目(CSRF200506).
通信作者:朱大勇. E-mail: dayongzhu75@163.com.

1 遗传聚类的社团发现方法

遗传聚类方法是通过进化迭代把数据对象划分到不同的簇中. 通常随机选取 k个对象作为初始的 k个簇的质心;然后,根据簇中心和其他对象之间的距离关系来进行簇的划分. 采用遗传聚类方法来实现复杂网络的社团发现,主要存在两个问题:一是网络的拓扑信息很难直接映射为聚类中心和其他对象之间的距离关系;二是需要根据聚类中心和网络的社团结构确定遗传算法的适应度函数.

本文提出的社团发现方法首先获取网络的拓扑信息,并计算所有节点对的相异性指数[6] (dissimilarity index);然后,通过节点的相似度度量标准进行社团划分,以网络社团结构的模块度作为算法的适应度函数控制进化过程. 该算法不需要将网络拓扑信息映射为 n维数据空间,而且染色体占用的空间较少;另外,算法最终的结果可直接对应于发现的网络社团结构.

1.1 编码及初始化种群

复杂网络社团发现的求解结果是得到已发现社团的聚类中心. 因此,在编码时不采用二进制方式,而是直接以网络节点的编号构造染色体. 一个社团的网络中心节点代表一个基因 (即问题的求解结果),多个聚类中心节点的编号连接在一起形成串. 图 1中的染色体代表 k个社团的聚类中心.

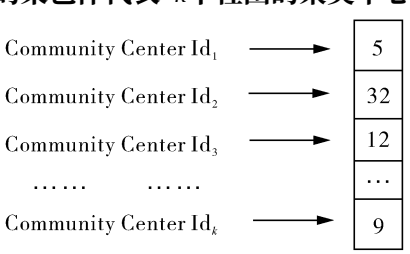

图 1 算法中染色体的编码方式

Fig 1 Chromosome representation of the algorithm

初始化种群时,随机地从网络中选取 k个节点作为 k个社团的聚类中心 (参数 k的确定将在 1.4节讨论)来建立染色体. 种群中的个体数量由参数 p设定,因此整个种群所占用的空间为 $k \times p$

1.2 适应度函数

计算种群中每个个体的适应度,首先要确定网络中节点之间的距离关系,这样才能明确各个网络节点与社团中心节点的相似程度,以便将网络节点划分到离自己最近的社团. 网络节点之间的相似度采用相异性指数[6]进行度量,也就是通过两个节点 i, j与其他网络上节点之间的最短路径来确定这两个节点之间的距离 (相似或相异程度). 相异性指数的定义如下:

$$\Lambda(i,j) = \frac{\sqrt{\sum_{k=1,k\neq i,j}^{N}[d_{ik} - d_{jk}]^2}}{(N-2)}. \quad (1)$$

其中: d_{ik}和 d_{jk}为节点 i, j到节点 k的最短路径.

Newman等人考虑了社团内部边和连接社团之间的边的数量关系,提出了模块度作为网络划分质量的度量标准[7]. 在本文中,以模块度 Q来确定种群中个体的适应度. 在 N个节点的无向网络 $G(V, E)$中, V是节点集, E是边集. 网络的模块度定义为

$$Q = \sum (e_{ii} - a_i)^2. \quad (2)$$

式 (2)中: e_{ii}是网络中第 i个社团内部连接各节点的边在网络所有边的数目中所占的比例; a_i表示与第 i个社团中的节点相连的边在网络所有边的数目中所占的比例.

在进化迭代时,根据种群中个体所代表的 k个社团中心节点,用相异性指数作为度量将每个网络节点划分到距离自己最近的社团. 得到 k个社团 (聚类)后再计算网络的模块度,以此作为种群中每个个体的适应度,用于后续的遗传操作.

1.3 遗传算子

在进行遗传操作时,根据每个个体的适应度进行选择、交叉和变异操作. 选择操作采用轮盘的方式来确定进行交叉操作的个体. 适应度越大的个体选中进行交叉的概率也就越大. 个体交叉后,从上一代种群中选择出来的一些优良的个体遗传到下一代种群中.

在交叉操作时,选取父辈种群的配对个体. 通过随机选取交叉点,交换它们的部分基因. 交叉使得遗传信息得到交换,种群获得新一代的个体. 交叉的方式如下图所示.

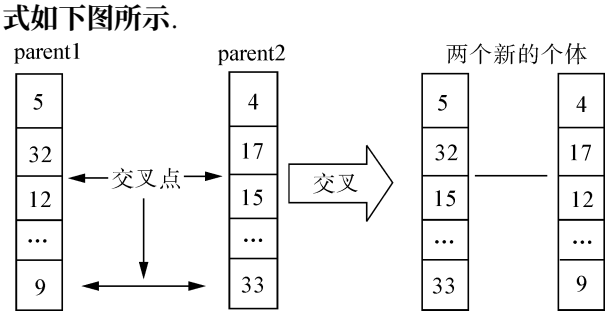

图 2 算法中染色体的交叉操作

Fig 2 Cross-over operation of chromosome in algorithm

在初始化种群和进行交叉操作时,同一个染色

体中可能出现相同聚类中心(节点)的情况,因此需要对相同的中心节点进行变异处理,即通过随机选取网络节点的方式改变原有染色体中出现的相同基因(相同的中心节点).通过这种方式,保证了在遗传过程中,进行节点聚类时社团数目不会发生变化.变异的过程如图 3所示.

图 3　算法中染色体的变异操作

Fig 3　Mutation operation of chromosome in algorithm

1.4　算法流程及参数

根据遗传聚类的方式,对网络社团中心节点进行进化迭代,算法的流程如下:

1)根据给定的网络,计算网络中所有节点对的相异性指数.

2)设置遗传迭代参数,包括迭代次数和种群个数.

3)随机初始化种群中的个体,如果发现个体中有相同的聚类中心,则对该个体进行变异操作.

4)重复进化迭代:

a)选取个体对进行交叉,如果新个体产生相同的聚类中心节点,则进行变异操作;

b)计算新种群中每个个体的适应度(计算网络划分的模块度)和整个种群的最大适应度;

c)如果上一代种群的最大适应度大于下一代种群的最大适应度,则将下一代种群中具有最小适应度的个体替换为上一代种群中具有最大适应度的个体;

d)当种群中适应度最大的个体没有改变或迭代次数到达设定的范围时则终止算法.

在复杂网络社团发现中,不用人工干涉而自动确定社团的个数 k(即聚类网络中心节点数)是一个复杂的问题.有的算法将其转换为某种阈值来进行分析,但这也仅仅是将确定社团个数变成了确定另外一个参数.在本文中首先对网络进行二分,然后根据网络模块度在优化过程中是否收敛(即平方误差准则函数是否小于指定的值)来确定参数 k

2　实验结果

为了验证所提出的方法的有效性和可行性,对 Zachary Karate Club网络和 College Football网络进行了社团发现实验.实验结果表明该算法有效并获得了较好的社团划分.同时,相对于文献[3, 5]用网络中所有节点来构造染色体,本文所提出的算法占用更少的空间.

2.1　Zachary Karate Club网络

Zachary Karate Club网络是社会学的经典研究之一.它描述了美国一所大学中空手道俱乐部会员相互间的社会关系.网络中共有节点 34个,如图 4所示.应用遗传聚类的方法,设置种群个数 50个,在迭代 50次后,网络划分为两个社团所计算的个体适应度(网络模块度)Q = 0.371 466 14,这和实际网络的社团划分情况一致.

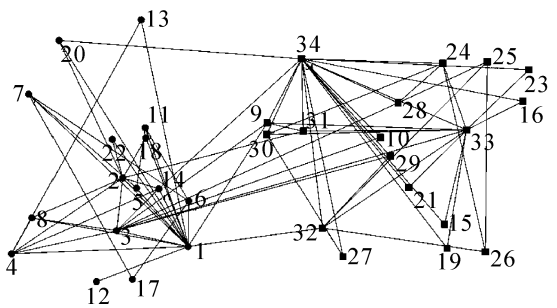

图 4　Zachary Karate Club网络的社团结构

Fig 4　Community structure of Zachary Karate Club network

2.2　College Football网络

College Football网络是 Newman等人对美国大学生橄榄球联赛的 2 000个赛季的比赛情况进行分析整理而建立的网络模型,它包含 115个节点及 616边.用本文的算法进化迭代 100次,将网络划分为 12个社团,适应度达到 0.561 835 89.该结果与实际社团的划分一致.图 5显示了 College Football网络的结构.

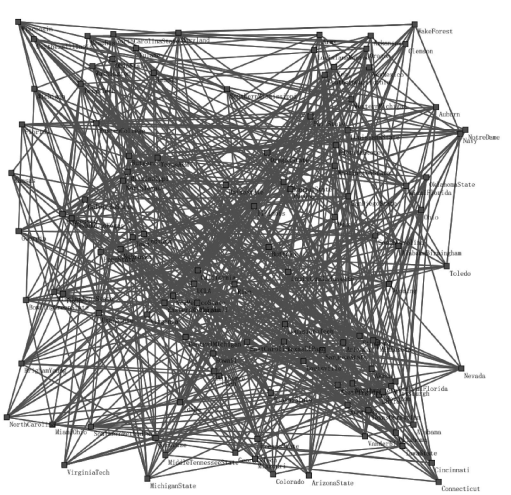

图 5　College Football网络结构

Fig 5　Network structure of College Football

3 结束语

在网络数据分析中,复杂网络的社团发现是一个非常具有挑战性的问题.本文以相异性指数作为网络节点的距离度量,结合模块度提出用遗传聚类来分析和发现网络社团结构.相对于其他算法而言,该方法以聚类中心作为染色体,减少了种群的空间占用;同时,不需要将网络拓扑信息(邻接矩阵)映射到 n维数据空间,减少了数据的失真.聚类结果也不需要进行还原处理可直接得到已划分的社团结构,降低了算法的复杂性.利用网络的拓扑信息改进算法的选择、交叉和变异操作以增加个体的多样性,提高算法的收敛速度,是今后进一步的研究方向.

参考文献:

[1] NEWMAN M E J. Fast algorithm for detecting community structure in networks [J]. Phys Rev E, 2004, 69(6): 066133.

[2] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proc Natl Acad Sci, 2001, 99: 7821-7826.

[3] TASGIN M, HERDAGDELEN A, BINGOL H. Community detection in complex networks using genetic algorithms [J/OL]. [2008-09-13]. http://arxiv.org/abs/0711.0491.

[4] 刘 婷, 胡宝清. 基于聚类分析的复杂网络中的社团探测 [J]. 复杂系统与复杂性科学, 2007, 4(1): 28-35.
LIU Ting, HU Baoqing. Detecting community in complex networks using cluster analysis [J]. Complex Systems and Complexity Science, 2007, 4(1): 28-35.

[5] LIU Xin, LI Deyi, WANG Shuliang, et al. Effective algorithm for detecting community structure in complex networks based on GA and clustering [J]. Lecture Notes in Computer Science, 2007, 4488: 657-664.

[6] ZHOU Haijun. Distance, dissimilarity index and network community structure [J]. Phys Rev E, 2003, 67(6): 061901.

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

作者简介:

朱大勇,男,1975年生,讲师.主要研究方向为复杂网络、对等计算、分布式信息检索、软件工程.参加过多项科研项目,发表学术论文 20余篇.

张新丽,女,1973年生,副教授.主要研究方向为复杂系统、神经网络.参加过多项科研项目,发表学术论文 10余篇.

侯晓荣,男,1966年生,教授、博士生导师.主要研究方向为智能推理、机器证明.参加过多项科研项目,发表学术论文 30余篇.