

文章编号:1007-6735(2010)05-0441-04

基于遗传算法的二部图判定算法

李小强, 张 宁

(上海理工大学 管理学院, 上海 200093)

摘要: 应用遗传算法来判定二部图的具体过程是首先将无向图 G 的节点随机分配到两个不同社区中, 然后用遗传算法进行进化操作, 优化无向图 G 的模块化函数 Q , 当 Q 取最小值且无向图 G 的边只存在于两个社区之间, 则无向图 G 为二部图. 实例分析结果验证了算法的有效性.

关键词: 图论; 复杂网络; 二部图; 遗传算法; 模块化函数

中图分类号: TP 393 **文献标志码:** A

Bipartite graph identification using genetic algorithm

LI Xiao-qiang, ZHANG Ning

(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In order to reduce the time complexity, genetic algorithm was used to identify bipartite graphs. The nodes of graph G were allocated into two communities randomly; and then genetic algorithm was taken to optimize the modularity function Q . When obtaining the minimum value of Q and sitting all the edges between the two communities, graph G is a bipartite graph. Its accuracy was tested with an example.

Key words: graph theory; complex networks; bipartite graph; genetic algorithm; modularity function

复杂网络是一类既非完全规则又非完全随机的网络^[1,2]. 这种网络广泛存在于物理系统、化学系统、生物系统、技术系统 & 社会系统当中. 复杂网络用节点和边来描述真实世界中的复杂系统, 一个节点代表复杂系统中的一个主体, 一条边代表其所关联的两个主体间的关系. 许多真实世界的复杂网络具有模块(modularity)性, 在模块内节点之间的连接比较紧密, 而在模块间节点之间的连接比较松散. 近年来, 网络模块性研究得到了很多学者的关注^[2,3]. 而二部图 G 则是这样一种网络, 它可以分成两个非空子集 V_1 和 V_2 , 并且满足图 G 的边关联的两个节点 x, y 分别属于这两个子集, 它的两个模块内不存在边.

在现实生活中, 很多网络的拓扑结构可以用二部图来表述, 如蛋白质相互作用网络(protein-protein interaction networks)^[4]、科学家出版物合著网络^[5,6]及艺术家合作网等^[7]. 但是, 目前很多关于模块划分的一般算法对这类网络都不适用, 因为他们大多是将二部图投影到单模式网络后再划分, 这样丢失了很多信息. 为此一些学者开始探索有关二部图模块划分的算法^[8], 也有一些学者开始探索网络的二分性(bipartivity), 并定义了一些定量测量指标^[9,10]. 然而, 对于二部图判定的具体算法却很少, 大多都是根据定义来判定, 且很多算法采用了邻接矩阵作为数据结构, 并夹杂很多矩阵运算, 这就使得运算速度很慢, 不利于实施.

收稿日期: 2010-01-19

基金项目: 上海市重点学科建设资助项目(S30501); 上海市研究生创新基金资助项目(JWXCSL0902)

作者简介: 李小强(1985-), 男, 硕士研究生.

张 宁(联系人), 女, 教授. E-mail: zhangning@usst.edu.cn

本文提出的算法,以邻接表作为数据结构来存储网络节点,同时避免了很多矩阵计算,这对于节点众多的复杂网络,尤其是邻接矩阵为稀疏矩阵的复杂网络来说,可以减少很多运行时间,从而对网络拓扑结构快速做出判断.

1 二部图及其性质

定义 若图 G 的节点集 $V(G)$ 可以分成两个非空子集 V_1 和 V_2 , 并且满足图 G 的边关联的两个节点 x, y 分别属于这两个子集, 则图 G 为二分图(二部图).

定理 图 G 是二部图当且仅当图 G 不含奇圈^[11].

NEWMAN 定义了一个变量 R 来计算两个模块间边数^[12]. 具体方法是, 对于图 G , 假设可将其分成两个顶点集 g_1, g_2 , 则连接这两个集合的边的数目可表示成

$$R = \frac{1}{2} \sum_{ij} A_{ij} \delta_{ij}$$

$$A_{ij} = \begin{cases} 1 & \text{顶点 } i \text{ 与顶点 } j \text{ 相邻} \\ 0 & \text{其他} \end{cases}$$

$$\delta_{ij} = \begin{cases} 1 & \text{顶点 } i, j \text{ 属于不同顶点集} \\ 0 & \text{其他} \end{cases}$$

式中, A_{ij} 为邻接关系; δ_{ij} 为狄拉克函数.

为了便于计算, 定义了一个索引向量 S , 其分量

$$s_i = \begin{cases} 1 & \text{顶点 } i \text{ 属于集合 } g_1 \\ -1 & \text{顶点 } i \text{ 属于集合 } g_2 \end{cases}$$

则

$$R = \frac{1}{4} \sum_{ij} (1 - s_i s_j) A_{ij}$$

对于 NEWMAN 定义的变量 R , 也可以表述成以下形式

$$R = c - \frac{1}{4} S^T A S$$

其中

$$c = \frac{1}{4} \sum_{ij} A_{ij} = \frac{1}{2} m$$

式中, m 为图的边的数目.

对于一个拥有 m 条边的二部图 G , 则它的两个非空子集 V_1 和 V_2 间的边数 $R = m$. 可以以此来作为二部图识别算法的结束条件.

2 基于遗传算法的二部图识别方法

遗传算法是由达尔文进化论理论引发的一种计算模型, 是模拟生物体的适应性和进化能力的全局优

化搜索算法. 它具有全局优化、简单通用、鲁棒性强及并行处理能力强等特点, 而且它不依赖于问题的具体领域, 对问题的种类有很强的鲁棒性, 被广泛应用于许多学科, 其主要应用领域是函数优化和组合优化. 本文将遗传算法应用于二部图的判定. 将遗传算法及图的模块性结合起来, 能有效地识别二部图.

2.1 适应度函数

适应度函数一般是由所求问题的目标函数按照一定规则转换而得到的. 将图划分成两个模块后, 可以用其模块化函数 Q 来衡量其模块性, 当图的模块性最差时, 若模块间存在的边数等于图的总边数, 则图为二部图. 遗传运算优化的目标函数就是求 Q 的全局最小值, 则适应度函数

$$f = \frac{1}{Q + 2} \quad (1)$$

式中, Q 为 NEWMAN 定义的模块化函数^[13].

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (2)$$

式中, A_{ij} 代表节点 i 与节点 j 的邻接关系; k_i 代表节点 i 的度数; δ 为狄拉克函数; c_i 为节点 i 所属的社团编号.

$$A_{ij} = \begin{cases} 1 & \text{如果节点 } i \text{ 与节点 } j \text{ 相连} \\ 0 & \text{其他} \end{cases} \quad (3)$$

$$k_i = \sum_j A_{ij} \quad (4)$$

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{如果 } c_i \text{ 等于 } c_j \\ 0 & \text{其他} \end{cases} \quad (5)$$

由式(2)可知, 当模块内边数少于随机条件下的边数时, $Q < 0$, 当 Q 取值最小时若两个模块间的边数等于网络中的边的数目, 即 $R = m$, 则该网络的拓扑结构是二部图.

2.2 种群初始化

在建立初始种群的操作过程中, 给每一个节点分配一个随机的社团 ID 号(分组编号), ID 取值为 1 或 2, 如图 1 所示. 根据网络的节点数 n 及节点间邻接关系建立一条染色体. 染色体数量依据问题的具体情况确定.

在初始化的过程中, 每个节点被随机分配一个社区 ID 号, 但是, 必须有一定的机制去确保在初始化随机分配社区 ID 的同时, 节点应该尽可能按照连接关系进行合理分配. 可以随机选择几个节点, 将其 ID 号扩散到与其相连的节点上去. 该调整准则在很大程度上确保了初始化节点 ID 的过程与实际社区划分的相关性, 同时增强了本算法的收敛性, 且减

少了不必要的迭代.



图1 种群初始化

Fig.1 Population initialization

2.3 编码

编码是应用遗传算法时要解决的首要问题,也是设计遗传算法时的一个关键步骤.在遗传算法的执行过程中,对不同的具体问题进行编码,编码的好坏直接影响选择、交叉及变异等遗传操作.

本文算法采用字符集{1,2}进行编码,是因为二部图节点集可分为两个互不相交的点集.染色体的基因长度等于节点数.

2.4 选择

选择的目的是为了从当前群体中选出优良的个体,使它们有机会作为父代为下一代繁衍子孙.在初始种群建立后,每一次执行算法的迭代过程中都计算每个染色体的适应度,根据各个染色体个体的适应度,按照轮盘赌选择规则从上一代群体中选择出一些优良的个体遗传到下一代群体中,从而使优良特性得以遗传.然后再执行交叉和变异操作.

在每一代种群中,根据染色体的适应度排序,应用精英选择策略,将适应度最大的染色体保存下来.通过这种方法可以保证好的染色体产生后不会消失.

2.5 交叉

交叉操作是遗传算法的最重要的遗传操作步骤.通过交叉可以得到新一代个体,新个体组合了父辈个体的特性.将父辈群体内的个体搭配成对,交换它们之间的部分染色体.

根据选择操作所得的染色体序列,以0.8的交叉概率对它们进行交叉操作.通过交叉的方式,染色体之间可以进行信息交互,产生新的个体,使得整个染色体种群呈现新的性状.

当选定两条染色体进行交叉操作时,随机选择一个基因位,对这两条染色体进行单点一致交叉操作.单点一致交叉的具体操作描述如下:选择两个作为父辈的染色体,一个称为源染色体,另外一个称

为目标染色体.随机选择一个节点作为交叉点,将源染色体交叉点一侧的基因传递给目标染色体交叉点同侧的位置上,从而得到新的子代染色体.交叉过程如图2所示. a_1, a_2 分别为染色体a上交叉点之前、之后的基因片段, b_1, b_2 分别为染色体b上交叉点之前、之后的基因片段.

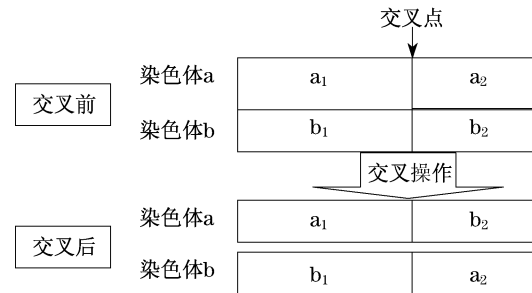


图2 单点一致交叉操作

Fig.2 One-way crossing-over

由于模块是一种相对属性,不同染色体上的不同的模块ID号有可能代表同一模块.例如,染色体解A上的ID号1的模块与染色体解B上的ID号为2的模块具有相同的节点成员,因为两条染色体上的编号并不存在对应关系.为此,需要对交叉过程做点调整,具体过程如下:对即将要进行交叉操作的每条染色体,将其第一个基因位及与第一个基因位有相同编号的ID号都设为1,然后将剩余节点的ID号设为2.例如,染色体A与染色体B具有相同适应度,然而基因编码确不同,如图3所示,如果不经调整就交叉,可能会得到退化的解,做了适当调整后,无论选择哪个点开始交叉都会保持原有性状.

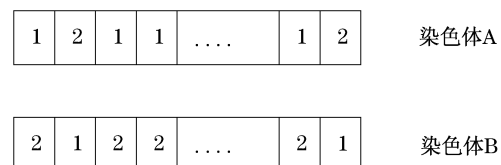


图3 具有相同适应度的两条染色体

Fig.3 Two chromosome with same fitness

2.6 变异

在执行交叉操作后,以0.1的概率对每条染色体的每一个基因位进行变异操作.被选定的基因位的值若为1,则将其改为2;否则,改为1.

3 实例分析

现用上述方法对一个实例进行分析,判断其是

否为二部图,同时验证算法的正确性.边数 $m = 11$,图 G_1 如图 4 所示,其邻接矩阵为

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

在 Java 软件中通过构造一个不平衡二维数组来作为图 G_1 的邻接表.然后设定算法的种群规模为 50,最大迭代次数为 100,变异概率和交叉概率分别为 0.1 和 0.8.独立运行程序 10 次,得到最优结果时,迭代次数 g 有 8 次取值为 1,1 次为 5,1 次为 3.说明算法基本上只要迭代 1 次就能得到适应度函数最优值.最优结果为 $f = 0.667$, $R = 11$, $g = 1$.相应染色体编码为

2	1	2	1	1	2	2
---	---	---	---	---	---	---

由 $R = m$ 知该图为二部图,依据染色体编码重绘图形如图 5 所示.

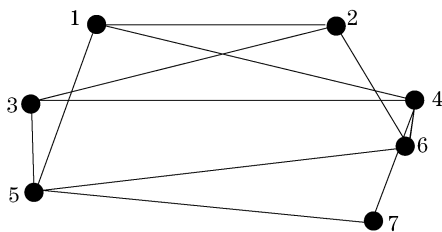


图 4 G_1 的拓扑结构

Fig.4 Topology structure of graph G_1

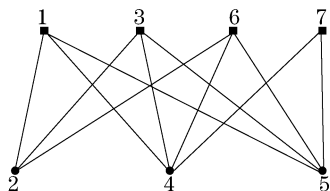


图 5 G_1 的二部图表述形式

Fig.5 Bipartite structure of graph G_1

在图 G_1 的节点 1 和 3 之间加一条边,得到图 G_2 .应用算法对图 G_2 进行判定,种群规模、交叉概率和变异概率都不变,先将迭代次数设为 1,独立运行 10 次,发现大多情况下能得到最优结果,然后将迭代次数设为 2,独立运行程序 10 次,每次都能得到最优结果, $f = 0.633$, $R = 11$,耗时 0.797 s. $R < 12$,所以,判定图 G_2 为非二部图.

4 结束语

提出了一种基于遗传算法的二部图识别方法,该方法通过优化适应度函数来获得网络模块性 Q 最小值,若此时两个模块间的边数 R 等于网络总边数 m ,则该网络具有二部图拓扑结构.由于启发式算法的特性是不一定能得到最优解,但可以得到较优解,所以,有时候需运行算法几次,取最优值.

该算法具有较快的运算速度,能够对网络结构快速做出判断.本文的实例验证了算法的有效性.

参考文献:

- [1] ALBERT R, BARABÁSI A L. Statistical-mechanics of complex networks[J]. Rev Mod Phys, 2002, 74(1): 47-97.
- [2] NEWMAN M E J. The structure and function of complex networks[J]. SIAM Rev, 2003, 45(2): 167-256.
- [3] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA, 2002, 99(12): 7 821-7 826.
- [4] MASLOV S, SNEPPEN K. Specificity and stability in topology of protein networks[J]. Science, 2002, 296(5 569): 910-913.
- [5] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proc Natl Acad Sci USA, 2001, 98(2): 404-409.
- [6] BORNER K, MARU J T, GOLDSTONE R L. The simultaneous evolution of author and paper networks[J]. Proc Natl Acad Sci USA, 2004, 101(Supple 1): 5 266-5 273.
- [7] UZZI B, SPIRO J. Collaboration and Creativity: The Small World Problem[J]. American Journal of Sociology, 2005, 111(2): 447-504.
- [8] GUIMERA R, SALES-PARDO M, AMARAL L A N. Module identification in bipartite and directed networks[J]. Phys Rev E, 2007, 76(3): 1-8.
- [9] Petter Holme, Fredrik Liljeros, Christofer R. Edling, and Beom Jun Kim. Network bipartivity [J]. Phys Rev E, 2003, 68(5): 1-10.
- [10] ESTRADA V, RODRIGUEZ-VELQUEZ V. Spectral measures of bipartivity in complex networks[J]. Phys Rev E, 2005, 72(4): 1-7.
- [11] BUCKLEY F, LEWINTER M. 图论简明教程[M]. 李慧霸, 王凤芹, 译. 北京: 清华大学出版社, 2004.
- [12] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Phys Rev E, 2006, 74(3): 1-19.
- [13] NEWMAN M E J. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2): 1-15.