基于 PSO 微粒群算法的复杂网络社区结构发现

戴飞飞,唐普英
DAI Fei-fei,TANG Pu-ying

电子科技大学 光电信息学院,成都 610054
School of Opto-Electronic Information,University of Electronic Science and Technology of China,Chengdu 610054,China
E-mail:yuebianyun@163.com

DAI Fei-fei,TANG Pu-ying.Community structure detection in complex networks using Particle Swarm Optimization algorithm.Computer Engineering and Applications,2008,44(22):56-58.

Abstract: Community structure identification has been one of the most popular research areas in recent years and there has been many algorithm proposed so far to detect community structures in complex networks in varied topics,where most of the algorithm have some drawbacks,and some of them are not suitable for very large networks because of their time-complexity.In this paper,an algorithm for detecting community structures in complex network is presented,which is based on the Particle Swarm Optimization algorithm.It doesn't need any priori knowledge about the numbers of communities and any threshold values.The algorithm is tested on the two network data named Zachary Karate Club and College Football.
Key words: complex networks;community structure;Particle Swarm Optimization(PSO) algorithm

摘 要:复杂网络社区结构划分日益成为近年来复杂网络的研究热点,到目前为止,已经提出了很多分析复杂网络社区结构的算法。但是大部分算法还存在一定的缺陷,而且有些算法由于其时间复杂度的过高导致其不适合应用于对大型网络的分析。提出了一种基于 PSO 微粒群算法的复杂网络社区结构分析方法。此方法无需预先知道组成该复杂网络的社区数量、社区内的节点数以及任何门限值。该算法的可行性用 Zachary Karate Club 和 College Football Network 模型进行验证。
关键词:复杂网络;社区结构;PSO 微粒群算法
DOI:10.3778/j.issn.1002-8331.2008.22.016 **文章编号**:1002-8331(2008)22-0056-03 **文献标识码**:A **中图分类号**:TP393

1 前言

复杂网络是现实世界中复杂系统的一种抽象表现形式。现实世界中存在很多类型的复杂网络,比如互联网,社会关系网络,以及食物链网络等等。复杂系统中的独立个体是复杂网络中的节点,节点之间的边则是系统中个体之间按照某种规则而自然形成或人为构造的一种关系。举例而言,互联网中的节点就是每台独立的计算机,边则是两台计算机之间的连接关系。

大量研究表明,许多网络是异构的,即复杂网络不是一大批性质完全相同的节点随机地连接在一起的,而是许多类型的节点的组合。相同类型的节点之间存在较多连接,不同类型的节点之间连接则相对较少。把满足同一类型的节点以及这些节点之间的边所构成的子图称为网络中的"社区[1]"。社区结构是复杂网络的一个重要特性。

复杂网络中社区发现的研究起源于社会学的研究工作 Girvan 和 Newman 以及其它学者的研究成果,Girvan 和 Newman 把社区发现问题定义为:将网络节点划分成若干组,使得组内节点之间的连接比较稠密,而不同组节点之间的连接则比较稀疏。

2 社区结构发现方法

2.1 Girvan-Newman(GN)算法[2]

目前最众所周知的算法当属 Girvan-Newman(GN)算法,这是一种基于边介移除的方法。GN 算法是一种分裂方法,它的基本思想是:通过不断地从网络中移除边介数最大的边,将整个网络分解为各个社区。边介数定义为:网络中经过每条边的最短路径的数目,它为区分一个社区的内部边和连接社区之间的边提供了一种有效的度量标准。

2.2 GN 算法的衡量标准[3]

GN 算法弥补了一些传统算法的不足,但是,在不知道社区数目的情况下,也不知道这种分解要进行到哪一步终止。

为解决此问题,Newman 等人引进了一个衡量网络划分质量的标准——模块性(Modularity)。考虑某种划分形式,它将网络划分为 k 个社区。定义一个 k×k 维的矩阵 $e=(e_{ij})$,其中元素 e_{ij} 表示网络中连接两个不同社区 i 和 j 的节点的边在所有边中所占的比例。注意,在这里所说的所有的边是在原始网络中的,而不必考虑是否被社区结构算法移除。因此,该模块性的衡量标准是利用完整的网络来计算的。

作者简介:戴飞飞(1982-),男,在读硕士研究生,主要研究领域为进化计算、智能信号处理等。
收稿日期:2007-10-10 修回日期:2008-01-21

设矩阵中对角线上各元素之和为 $Tre=\sum_i e_{ii}$ (e_{ii} 表示网络中连接社区 i 内部各节点的边在所有边的数目中所占比例),定义每行(或者列)中各元素之和为 $a_i=\sum_j e_{ij}$,它表示与第 i 个社区中的节点相连的边在所有边中所占的比例。在此基础上,用下式来定义模块性的衡量标准:

$$Q=\sum_i(e_{ii}-a_i^2)=Tre-\|e^2\| \quad (1)$$

其中 $\|x\|$ 表示矩阵 x 中所有的元素之和。式(1)的意义是:网络中连接两个同种类型的节点的边(即社区内部边)的比例减去在同样的社区结构下任意连接这两个节点的边的比例的期望值。如果社区内部边的比例不大于任意连接时的期望值,则有 Q=0。Q 的上限为 Q=1,而 Q 越接近这个值,,就说明社区结构越明显。实际网络中,该值通常位于 0.3~0.7 之间。在 GN 算法上改进的一些分裂算法。

3 基于 PSO 算法的复杂网络社区结构发现

3.1 PSO 微粒群算法[4]

3.1.1 PSO 微粒群算法

微粒群优化算法兼有进化计算和群智能的特点。起初 Kennedy 和 Eberhart[5]只是设想模拟鸟群觅食的过程,但后来发现 PSO 是一种很好的优化工具。与其他进化算法相类似,PSO 算法也是通过个体间的协作与竞争,实现复杂空间中最优解的搜索。PSO 先生成初始种群,即在可行解空间中随机初始化一群粒子,每个粒子都为优化问题的一个可行解,并由目标函数为之确定一个适应度值(Fitness Value)。每个粒子将在解空间中运动,并由一个速度决定其方向和位置。通常粒子将追随当前的最优粒子而动,并经逐代搜索最后得到最优解。在每一代中,粒子将跟踪两个极值,一个是粒子本身迄今找到的最优解 pbest,另一个是整个种群迄今找到的最优解 gbest。

通常数学描述为:设在一个 n 维的空间中,由 m 个粒子组成的种群 $X=\{x_1,\cdots x_i,\cdots x_m\}$,其中第 i 个粒子位置为 $x_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$,其速度为 $v=\{v_{i1},v_{i2},\cdots,v_{in}\}$。它的个体极值为 $p_i=\{p_{i1},p_{i2},\cdots,p_{in}\}$,种群全局极值为 $p_g=\{p_{g1},p_{g2},\cdots,p_{gn}\}$,按追随当前最优粒子的原理,粒子 x_i 将按照式(2)和式(3)改变速度和位置。

$$v_{id}^{(t+1)}=v_{id}^{(t)}+c_1r_1(p_{id}^{(t)}-x_{id}^{(t)})+c_2r_2(p_{gd}^{(t)}-x_{id}^{(t)}) \quad (2)$$

$$x_{id}^{(t+1)}=x_{id}^{(t)}+v_{id}^{(t+1)} \quad (3)$$

式中 d=1,2,…,n,i=1,2,…,m,m 为种群规模,t 为当前进化代数,r_1 和 r_2 为分布于[0,1]之间的随机数,c_1、c_2 为学习因子或加速常数(Learning Factor or Acceleration Constant),此外,为使粒子速度不致过大,可设置速度上限 V_{max}。式(2)第一部分为粒子先前的速度;第二部分为“认知(Cognition)”部分,表示粒子自身的思考;第三部分为“社会(Social)”部分,表示粒子间信息共享与相互合作。

3.1.2 算法流程

算法主要计算步骤如下:

(1)初始化,设定学习因子 c_1、c_2,最大进化代数 T_{max},当前进化代数 t=1,在定义空间中 R^n 中随机产生 m 个粒子 $x_1,x_2,\cdots,x_m$,形成初始种群 X(t);随机产生各粒子位移变化 $v_1,v_2,\cdots,v_m$,形成位移变化矩阵 V(t)。

(2)评价种群 X(t),计算每个粒子的适应度值 F(X_i)。

(3)比较粒子的当前适应值 F(X_i)和自身最优值 pbest,如果 F(X_i)优于 pbest,则置 pbest 为当前值,并置 pbest 的位置为 n 维空间中的当前位置。

(4)比较粒子当前适应度值 F(X_i)与种群最优值 gbest,如果 F(X_i)优于 gbest,则置 gbest 为当前值,gbest 对应的序号为当前粒子序号。

(5)按式(2)和式(3)更新粒子的位移方向和步长,产生新种群 X(t+1)。

(6)检查是否评价值达到给定精度,若已达到或进化代数达到 T_{max},则结束循环;否则 t=t+1,转步骤(2)。

3.2 基于 PSO 算法的复杂网络社区发现

3.2.1 微粒群初始化

在建立初始种群的操作中,给每个节点,分配一个随机的初始化社区 ID 号。根据社区内节点数量 n,建立一个初始化数组。粒子数量 m,根据求解的问题的具体情况而建立。

如表 1 所示,对于粒子 X_i($i\in[1,m]$)来说,节点 j($j\in[1,n]$)的 ID 号代表了该粒子 X_i 在第 n 维上的当前位置。

表 1 种群初始化

1st node ID	2nd node ID	3rd node ID	…	(n-1)th node ID	nth node ID	X
10	22	21	…	3	31	X_1
23	15	19	…	11	11	X_2
…	…	…		…	…	…
31	20	6	…	25	1	X_{m1}
4	28	23	…	33	14	X_m

3.2.2 粒子位置调整

必须有一定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。比如存在边的两个节点,在很大概率上是存在于同一个社区的。基于这个思想,必须对初始化分配的节点进行调整:从节点 1 开始到节点 n 结束,把每个节点的社区 ID 号,扩散到周围的所有于该节点存在边的所有节点上。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。

3.2.3 参数选择

从微粒进化方程式(2)可以看出,c_1 调节微粒飞向自身最好位置方向的步长,c_2 调节微粒向全局最好位置飞行的步长,两者通常在 0~2 间取值。$r_1\sim U(0,1)$,$r_2\sim U(0,1)$ 为两个相互独立的随机函数。为了减少在进化过程中,微粒离开搜索空间的可能性,v_{id} 通常限定于一定范围内,即 $v_{id}\in[-V_{max},V_{max}]$。

3.2.4 进化计算

根据微粒进化方程式(3),对每个微粒的位置进行进化。对于每个微粒,为避免微粒离开搜索空间的可能性,根据复杂网络的具体情况,对于新的位置取模;对于每个微粒,将其适应度与所经历过的最好位置 P_i 的适应度进行比较,若较好,则将其作为当前的最好位置。对于每个微粒,将其适应度与全局所经历的最好位置 P_g 的适应度进行比较,若较好,则将其作为当前的全局最好位置。

3.2.5 纠错

为了增加复杂网络社区结构划分的准确性,必须要对那些可能明显划分错误的节点进行纠错。基于此目的,引入一个“错误门限”(Mistake Threshold)的定义。

根据复杂网络社区结构划分的理论,一个社区内部的连接

各节点的边数应该远远大于社区外部连接各社区之间的边数。因此,与一个节点相连的节点(邻居节点)与该节点在同一个社区内的可能性是很大的。通过分析每个节点的社区ID值及其邻居节点的社区ID值来增强划分结果的准确性。定义MI(i)表示节点i的错误门限值,假如MI(i)大于某个门限值,则该节点的划分呈现明显的错误状态。

$$MI(i)=\frac{\sum_{(i,j)}f(i,j)}{degree(i)} \quad (4)$$

其中

$$f(i,j)=\begin{cases}0 & CommID(i)=CommID(j)\\ 1 & CommID(i)\neq CommID(j)\end{cases}$$

按照节点号的顺序分析所有节点的MI(i)值,假如MI(i)大于某个门限值,那么查找该节点的邻居节点的社区ID,用出现次数最多的社区ID替代该节点原有的社区ID。经过对模型的分析所得结果显示,该纠错步骤相当行之有效,极大地解决了节点错误分配的情况,弥补了PSO算法的一些固有缺陷所带来的错误。

4 实验结果

为了测试本文中所提出算法的可行性,针对两个经典模型Zachary Karate Club和College Football Network进行了验证。

4.1 Zachary Karate Club[6]

20世纪70年代初期,Zachary用了两年的时间来观察美国一所大学中的空手道俱乐部成员间的相互社会关系。基于这些成员在俱乐部内部及外部的社会关系,他构造了他们之间的关系网。在调查过程中,该俱乐部的主管与校长之间因是否抬高俱乐部收费的问题产生了争执。结果,该俱乐部分裂成了两个分别以主管和校长为核心的小俱乐部。在复杂网络的社团结构分析中Zachary Karate Club网络已经成为一个经典的问题。

用算法去分析Zachary Karate Club网络,在很多次的计算中,除了部分进化失败的情况(如模块度Q=1)以外,对该复杂网络的社区结构划分准确性超过90%。个别不准确的情况中,节点10被分配到了错误的社区中去,见图1。

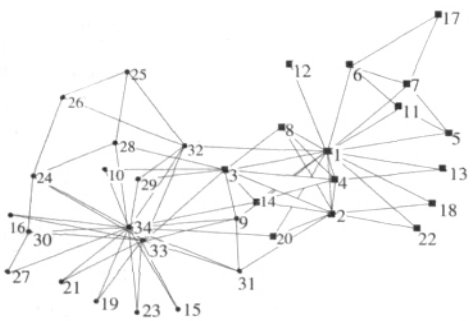

图1 PSO某次计算时的社区划分结果

4.2 College Football Network[7]

College Football Network模型是美国大学生足球联赛抽象出来的一个复杂网络模型。足球联赛中有若干支球队,网络中的节点代表一只足球队,两个节点之间的边表示两只球队之间进行过一场比赛。该网络模型收集于2000赛季的比赛数据情况,由Girvan M与Newman M收集整理而成。存在115只球队(节点)及616场比赛(边),包含了12个联盟。通过PSO算法解决球队联盟划分的问题,再次表现出为一种切实可行的方法。根据实验结果,比照联赛的真实情况,划分正确率超过80%。

5 结束语

复杂网络的社区结构发现已经成为当今一个非常具有挑战性的研究领域。本文中尝试用PSO微粒群算法来分析网络社区结构的划分情况。研究结果表明,用该算法对复杂网络的分析是行之有效的。比较其他算法而言,本方法的优点在于无需预先知道复杂网络的社区数量或者社区内的节点数量。

参考文献:

[1] Girvan M,Newman M E J.Community structure in social and biological networks[C]//Proceedings of National Academy of Science,2002,99:7821-7826.

[2] Clauset A,Newman M E J,Moore C.Finding community structure in very large networks[J].Physical Review E,2004,70.

[3] Newman M E J.Fast algorithm for detecting community structure in networks[J].Physical Review E,2004,69.

[4] Kennedy J,Eberhart R C.Particle swarm optimization[C]//Proceedings of IEEE International Conference on Neural Networks,Piscataway,NJ,1995:1942-1948.

[5] Eberhart R C,Kennedy J.A new optimizer using particle swarm theory[C]//Proceedings of the Sixth International Symposium on Micromachine and Human Science,Nagoya,Japan,1995:39-43.

[6] Zachary W W.An information flow model for conflict and fission in small groups[J].Journal of Anthropological Research,1977,33:452-473.

[7] Girvan M,Newman M E J.Community structure in social and biological networks[C]//Proceedings of National Academy of Science,2002,99:7821-7826.

[8] Newman M E J.Fast algorithm for detecting community structure in networks[J].Physical Review E,2004,69.

[9] 曾建潮.微粒群算法[M].北京:科学出版社,2004.

[10] Tasgin M.Community detection model using genetic algorithm in complex networks and its application in real-life networks[D].Graduate Program in Computer Engineering,Bogazici University,2005.

(上接55页)

参考文献:

[1] 王炳锡,屈单.实用语音识别基础[M].北京:国防工业出版社,2004.

[2] 蔡莲红,黄得志,蔡锐.现代语音技术基础与应用[M].北京:清华大学出版社,2003.

[3] 张雄伟,陈亮,杨吉斌.现代语音技术及应用[M].北京:机械工业出版社,2003.

[4] 杨行峻.语音信号数字处理[M].北京:电子工业出版社,1995.

[5] 邵央,刘丙哲,李宗葛.基于MFCC和加权矢量化的说话人识别[J].计算机工程与应用,2002,38(5):127-128.

[6] Fakhr W,Salam A A,Hamdy N.Enhancement of mismatched conditions in speaker recognition for multimedia applications[J].IEEE International Conference on Acoustics,Speech,and Signal Processing,2004.