第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于单目标 PSO 的社区检测算法

杨令兴 张喜斌

(空军工程大学航空航天工程学院 西安 710038)

摘 要 在将复杂网络的社区结构检测问题建模为单目标优化问题时,采用粒子群算法进行优化。传统粒子群算法用来解决连续优化问题,而社区结构检测问题则是一种基于图的离散优化问题。应用了新的编码策略和粒子更新策略解决这一问题,在更新策略中引入了基于近邻更新的方式,保证了在一定程度上遵循邻域信息引导粒子更新,以符合真实复杂网络的特性。另外,采用拓展的模块度密度函数进行优化,以解决传统模块度密度函数的分辨率限制问题,保证能在不同分辨率发现复杂网络的社区结构。实验结果证明,本算法是有效的,能够检测出不同分辨率下的社区结构。

关键词 复杂网络,社区检测,模块度密度函数,粒子群算法,单目标

中图法分类号 O157.5 文献标识码 A

Community Detection Algorithm Based on Single Objective PSO

YANG Ling-xing ZHANG Xi-bin

(School of Aeronautics and Astronautics, Air Force Engineering University, Xi'an 710038, China)

Abstract The community detection problem is modeled as single objective optimization problem, and the PSO algorithm is adopted to solve it. The traditional PSO algorithms are used to handle with continuous optimal problems, while the community detection problem is a graph based on discrete optimal problem in our paper. In this case, a new coding scheme and particles updating scheme were used to overcome this shortcoming. Besides that, a neighbor based strategy was adopted in the updating scheme to confirm that the neighborhood information can guide their updating which is consistent with the property of the real world complex networks. Besides, the general modularity density function is taken as the objective function to overcome the resolution problem, which confirms proposed algorithm can detect the construction of community under different resolutions. Experimental results indicate that this algorithm is effective and can detect the community construction of different resolutions.

Keywords Complex networks, Community detection, Modularity density function, PSO algorithm, Single objective

1 引言

计算机领域的图分割与复杂网络中的社区检测是两个类似的问题,因此图分割中的很多方法都可以应用到社区检测中来,如 ratio-cut[1]、ratio-association[2]、normalized-cut[3] 等。但通常实际生活中的社区数目是不可预知的,因此提出了衡量网络中社区划分的模块度 Q。模块度值越小,社区结构越模糊,模块度值越大,对应的社区结构越明显。但是由于模块化函数 Q 存在着严重的分辨率极限问题[4],利用模块度检测出的社区最小尺寸取决于与网络总边数有关的某个尺度,小于这个尺度的社区即使与外界只存在一条边的全连接子图,也不能被检测出来,因此将它合并到比它更大的子图中去[5]。

考虑到模块度的分辨率限制问题,通过模块度优化所得到的结果,不能判断出检测到的社区划分是否最优。李珍萍[6]等人针对此问题提出模块度密度的概念,其改善了模块度分辨率限制的问题,在此基础上,文中还提出了一个带参数的模块度密度扩展函数。通过调节这个参数,可以在不同的分辨率下分析网络结构,从而很好地解决了分辨率的限制问题。考虑到粒子群算法(PSO)的简单和易实现性,采用此算法对其进行优化。

2 PSO 算法

PSO 算法是一种基于群体的随机优化技术,它源于对鸟群捕食的行为研究,随后被拓展开来并用于很多领域。PSO 属于进化算法的一种,与遗传算法类似,也是从随机解出发,通过设置一定的迭代次数,寻找最优解。通过适应度来评价解的品质,与遗传算法相比,PSO 算法具有更简单的规则,具体描述如下[7]:

由 m 个粒子(particle)组成的群体以一定的速度在 d 维搜索空间中飞行,每个粒子以自己搜索到的历史最佳位置和群体内(或邻域内)其他粒子搜索到的历史最佳位置为依据,来调整改变自己的位置。粒子群中的第 i 个粒子分别由以下 4 个 d 维向量组成:

(1)第 i 个粒子在 d 维空间的位置表示为矢量 $x_i=(x_{i1},$

杨令兴(1988—),男,硕士,主要研究方向为网络社区检测,E-mail:178606862@qq.com;张喜斌(1962—),男,教授,主要研究方向为计算机技术应用、网络社区检测。

$x_{i2},\cdots,x_{id})$；

(2) 飞行速度 $v_i=(v_{i1},v_{i2},\cdots,v_{id})$；

(3) 个体历史最佳位置 $pbest_i=(pbest_{i1},pbest_{i2},\cdots,pbest_{id})$；

(4) 全局历史最佳位置 $gbest=(gbest_1,gbest_2,\cdots,gbest_d)$。

PSO 算法初始化为一群随机粒子，粒子跟踪两个“极值”$(pbest,gbest)$来更新自己，并通过有限次迭代找到最优解。更新策略[8]为：

$$v_i(t+1)=v_i+c_1r_1(t)(pbest(t)-x_i(t))+c_2r_2(t)(gbest(t)-x_i(t)) \tag{1}$$

$$x_i(t+1)=x_i(t)+v_i(t+1) \tag{2}$$

其中，加速因子 c_1、c_2 是两个调节粒子自身经验和群体经验的重要参数，影响着粒子的运动轨迹，c_1 表示自身经验对粒子运动轨迹的影响，c_2 表示群体经验对粒子运动轨迹的影响，通常将 c_1、c_2 都取为 2；r_1、r_2 为两个 0～1 之间的随机数。$c_1r_1(t)(pbest(t)-x_i(t))$决定了粒子有向当前个体最优方向移动的趋势，而 $c_2r_2(t)(gbest(t)-x_i(t))$则引导粒子向全局最优方向移动。

粒子的全局极值 $gbest$ 和个体极值 $pbest$ 由每一个粒子的适应值进行更新，更新方式如下：

$$pbest_i(t+1)=\begin{cases}x_i(t+1), & x_i(t+1)\geqslant pbest_i(t)\\ pbest_i(t), & x_i(t+1)<pbest_i(t)\end{cases} \tag{3}$$

$$gbest(t+1)=\max(pbest_i(t+1)),i=1,2,\cdots,n \tag{4}$$

3 基于 PSO 的社区检测算法

传统 PSO 算法用来解决连续优化问题，而社区结构检测问题则是一种基于图的离散优化问题。这里应用了新的编码策略和粒子更新策略解决这一问题，在更新策略中引入了基于近邻更新的方式，保证了在一定程度上遵循邻域信息引导粒子更新，符合真实复杂网络的特性。另外，采用拓展的模块密度函数进行优化，解决了传统模块密度函数的分辨率限制问题，以保证能在不同分辨率发现复杂网络的社区结构。

3.1 编码与解码

在此问题中，粒子位置和速度均为离散型，因此采用如下方式定义它们。粒子位置为 $x=(x_1,x_2,\cdots,x_i,\cdots,x_n)$，其中，$x_i$ 的取值范围为 $[1,n]$，n 为社区中顶点的数目。若 $x_i=x_j$，则表示顶点 i 与 j 属于同一社区，具体如图 1 所示。

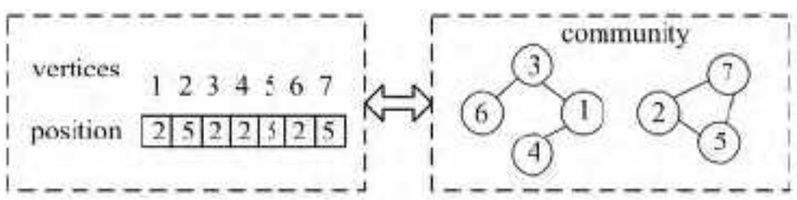


图 1 编码方式介绍

粒子飞行速度 $v=(v_1,v_2,\cdots,v_i,\cdots,v_d)$ 采用二进制编码，若 $v_i=1$，粒子位置 x_i 改变；否则，粒子位置 x_i 保持不变。

3.2 初始化

利用上述编码方式进行编码，在初始化之前，所有粒子位置设为 $(1,2,\cdots,n)$（n 为顶点数）。初始化时，对于每一个粒子顶点的随机选取，采用标签传播的方式进行，即若顶点 i 与顶点 j 互为邻居，则相应位置上的值更新为相同值，即两个顶点位于同一个社区，相关知识可参考文献[9]。

3.3 粒子更新

与传统的粒子群优化算法不同，采用如上编码方式的社区网络检测算法属于离散型 PSO 算法。为了成功实现离散编码下粒子的更新，采用文献[10]中的更新方法，其简要介绍如下：

$$v_i(t+1)=sig(\omega v_i+c_1r_1(pbest(t)\oplus x_i(t))+c_2r_2(gbest(t)\oplus x_i(t))) \tag{5}$$

$$y_i=sig(x_i)=\begin{cases}1, & rand(0,1)<sigmoid(x_i)\\ 0, & rand(0,1)\geqslant sigmoid(x_i)\end{cases} \tag{6}$$

$$sigmoid(x)=\frac{1}{1+e^{-x}} \tag{7}$$

式(5)—式(7)定义了粒子的飞行速度，符号 $\oplus$ 表示逻辑运算 XOR，ω 为 $[0,1]$ 之间的随机数，c_1、c_2 为 1.494，r_1、r_2 取值为 $[0,1]$ 间的随机数。

$$x_i(t+1)=x_i(t)\otimes v_i(t+1)=\begin{cases}x_i(t), & v_i(t+1)=0\\ Nbest_i, & v_i(t+1)=1\end{cases} \tag{8}$$

$$Nbest_i=\arg\max_r\sum_{j\in N}\varphi(x_i(t),r) \tag{9}$$

式(8)、式(9)定义了粒子的更新方法。其中，函数 $\varphi(i,j)=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$，$\arg\max_r f(r)$ 表示当 $f(r)$ 取最大值时 r 的取值，则 $Nbest_i$ 的物理含义为顶点 i 的近邻中数量最多的顶点的位置，其中 N 表示顶点 i 的近邻的集合。

图 2 详细描述了粒子的更新方法，清晰地揭示了式(5)—式(9)的物理意义。

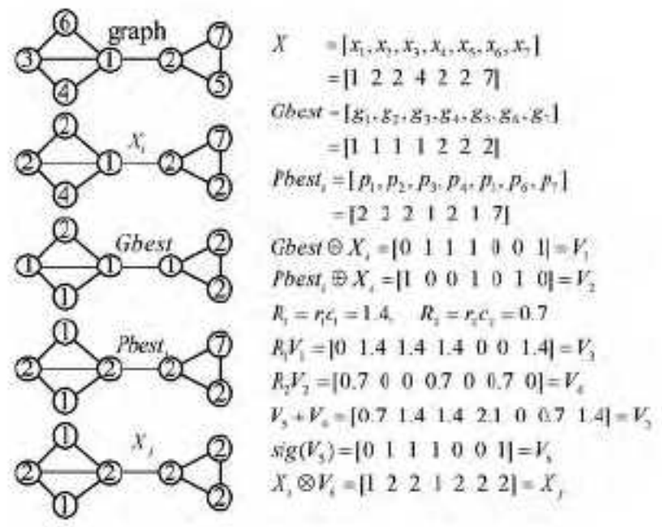


图 2 粒子更新方法介绍

3.4 算法流程

基于 PSO 的社区检测算法流程如表 1 所列。

表 1 基于 PSO 的社区检测算法流程

```
begin
  初始化粒子位置、速度、近邻等，并设置停止条件（迭代次数 gen）；
  解码计算适应度值，D_λ；
  for i=1:1:gen
    更新 gbest；
    计算每个粒子的飞行速度；
    计算粒子位置；
    解码计算适应度值；
    更新粒子近邻；
    更新 pbest；
  end
若满足终止条件，选择 gbest 输出；
end
```

4 实验结果与分析

这里将基于 PSO 的社区检测算法分别应用于人工合成网络和真实世界网络（对于这些网络，已经知道其真实划分），以验证本算法能否有效地检测出网络的社区结构，并且是否具有多分辨能力。

选取 Normalized Mutual Information(NMI)[11] 作为相似性度量，用来衡量真实的网络划分与算法检测的结果之间的相似度：

$$NMI(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}C_{ij}\log(C_{ij}N/C_{i.}C_{.j})}{\sum_{i=1}^{C_A}C_{i.}\log(C_i/N)+\sum_{j=1}^{C_B}C_{.j}\log(C_{.j}/N)} \quad (10)$$

式中，A 和 B 表示网络的两种划分方式，C 表示混淆矩阵(confusion matrix)，其中 C_{ij} 表示网络划分后，既属于 A 社区 i 中也属于 B 社区 j 中的节点个数。C_A、C_B 分别表示划分 A、B 中的社区数目，C_i 和 C_j 分别表示 C 中第 i 行或第 j 列的元素之和。N 表示节点个数，$NMI(A,B)$ 取值范围为 $[0,1]$，其值越大，表示分类越准确。当 $NMI(A,B)=1$ 时，表示 A 和 B 完全相同；$NMI(A,B)=0$，则表示 A 和 B 完全不同。

4.1 实验参数设置

在基于 PSO 的社区检测算法中，需要预先人为设定一部分参数，如粒子群规模、迭代次数。本文的主要目的在于运用 PSO 算法解决复杂网络社区检测问题，如何选取参数并不是重点讨论的问题，因此文中仅仅给出了最后的参数设置。粒子群规模设为 100，迭代次数为 100。

4.2 人工网络测试

采取 Lancichinetti 提出的基准测试网络[12]。该网络包含了 128 个节点，共分 4 个社区，每个社区包含了 32 个节点，其中每个节点的平均度为 16，由混合参数(mixing parameter) μ 控制节点外度所占的比例。μ 越小，节点和社区外节点的连接比例越小，社区结构越明显；μ 越大，节点和社区外节点的连接比例越大，社区结构越模糊。假设当 μ 取 0.5 时，每个节点平均有一半的连接都指向了社区外的节点。当 $\mu<0.5$ 时，节点外度所占比例要小于内度所占比例，此时社区结构模糊。当 μ 取 0 时，外度所占比例为 0，此时社区结构最明显。

通过在实验中调节混合参数 μ 的值，生成从 0.05 到 0.5 变化的 10 个网络，并用 NMI 衡量本检测算法的有效性。计算 20 次独立运行结果的平均值。λ 取不同的值时，对于 μ 从 0.05 到 0.5 变化的不同人工合成网络，PSO 算法所得到的平均 NMI 值变化曲线图如图 3 所示。

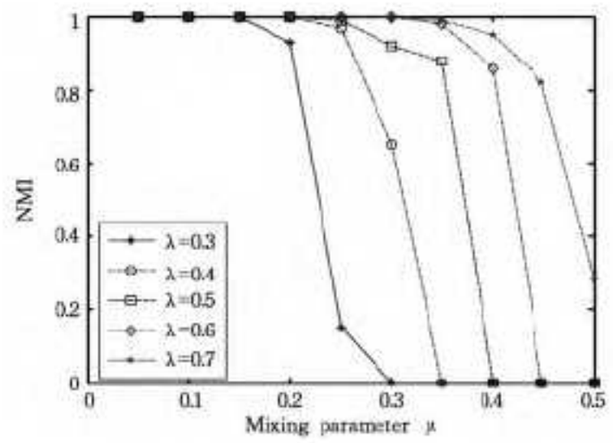


图 3 当 λ 取不同的值时，μ 从 0.05 到 0.5 变化时的不同人工合成网络算法所得到的平均 NMI 值变化曲线

由图 3 可知，当混合参数 μ 较小时，PSO 算法能够有效地发现社区结构，随着混合参数的提高，社区结构趋于模糊化时，算法的有效性也在不断下降。当 $\mu<0.25$，$NMI=1$ 时，算法能准确地发现社区结构，但当社区结构本身不明显时，算法也由于客观原因受到了限制。同时，分析 λ 值的变化曲线也可以发现，较大的 λ 值有助于发现较小的社区，而较小的 λ 值则有助于发现较大的社区。需要注意的是，当 $\mu=0.5$ 时，整个网络中的社区结构很模糊，所有算法都很难检测到社区结构。

4.3 真实世界网络

将 PSO 社区检测算法用于 Zachary 的空手道俱乐部(complex network on karate，以下简写作 karate)和美国大学足球联赛俱乐部(简写作 football)，并采用 NMI 指标进行测试。最后以 karate 为例给出社区结构的检测结果。

空手道俱乐部网络是 Zachary 在 2 年时间内观察一个具有 34 名成员的空手道俱乐部得到的。因为俱乐部教练和管理员之间发生了分歧，教练最终离开俱乐部，并带走了一半左右的俱乐部成员，此时网络被划分为两个社区，如图 4 所示。

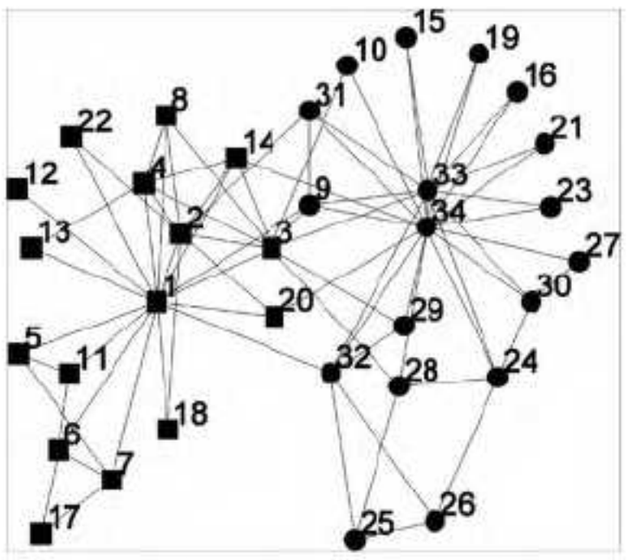


图 4 Zachary 空手道俱乐部网络的真实社区结构

而美国大学足球联赛网络是 Newman 和 Girvan 根据 2000 年秋季美国大学足球联赛常规赛季而编译的比赛网络[13]，球队是按照赛区划分的。网络中节点表示球队，网络中的边表示两个球队间之间进行的比赛。在整个赛季中，每支球队平均要打 11 场比赛，其中 7 场赛区内的比赛和 4 场赛区间的比赛。网络总共包含 115 个节点和 616 条边，被划分为 12 个组，即 12 个社区，如图 5 所示。

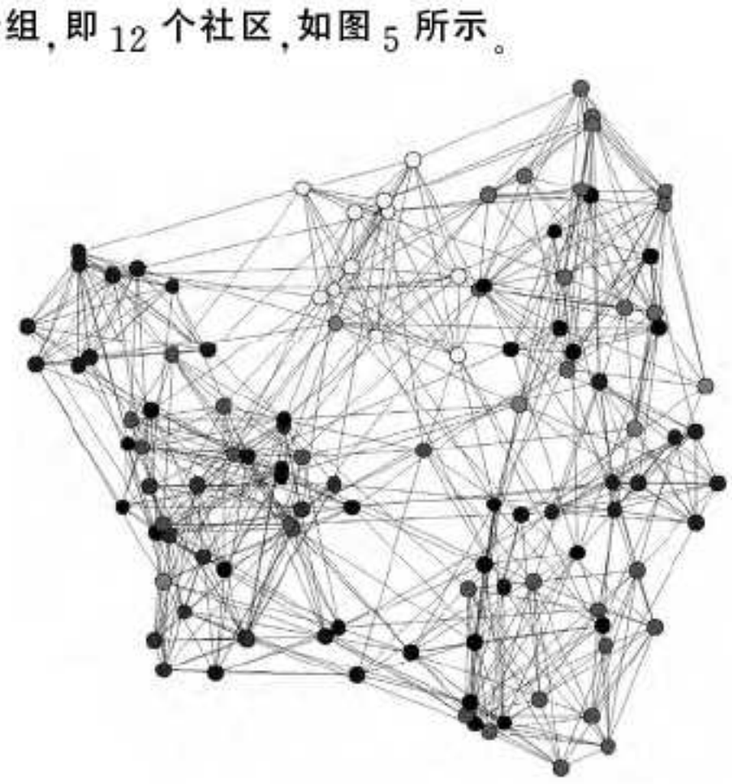


图 5 美国大学足球联赛俱乐部(football)

为了验证算法的有效性，调节目标函数中参数 λ 的值，使其在 $[0.2:0.1:0.8]$ 上取值。每组实验在真实网络 karate 和 football 上独立运行 30 次，NMI 指标的统计结果如图 6 所示。

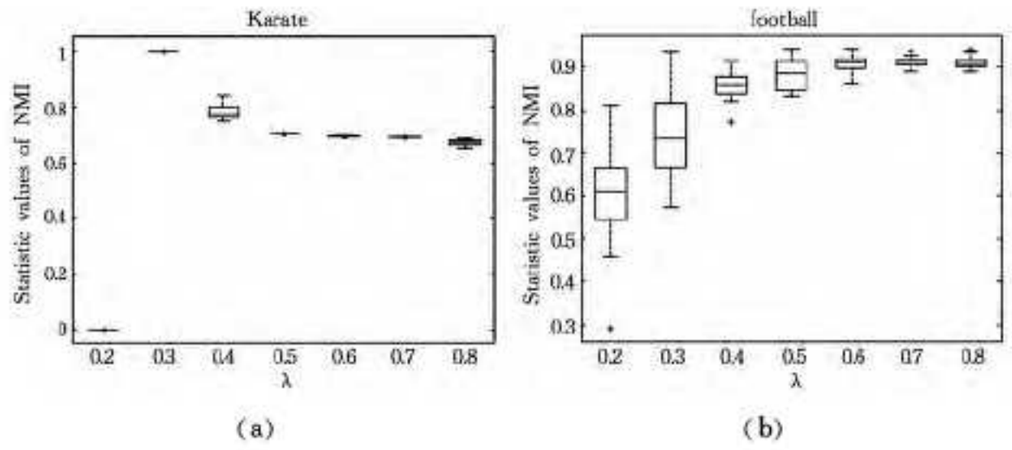
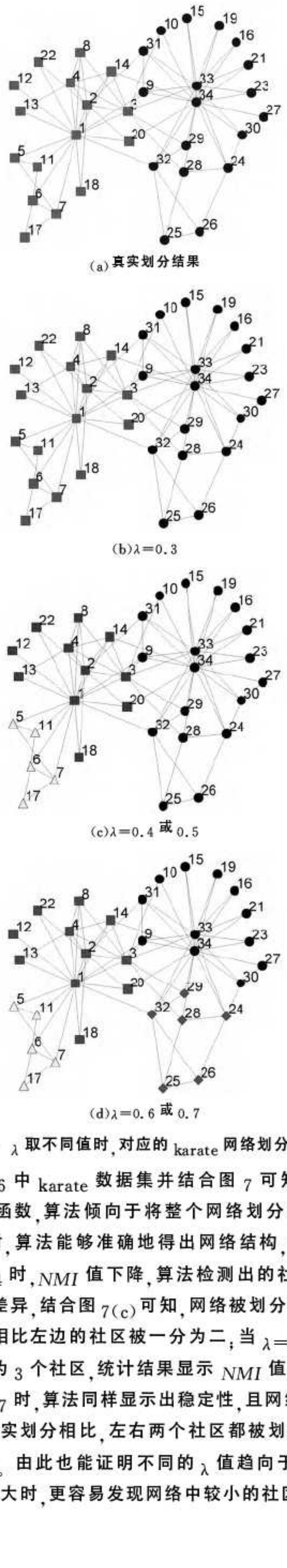


图 6 λ 取不同的值时，算法分别在(a)Zachary 空手道俱乐部，(b)美国大学足球联赛网络和真实世界网络上运行所得 NMI 的统计结果

λ 取不同值时，对应的 karate 网络划分结果如图 7 所示。



(a) 真实划分结果

(b)$\lambda=0.3$

(c)$\lambda=0.4$ 或 0.5

(d)$\lambda=0.6$ 或 0.7

图 7 λ 取不同值时，对应的 karate 网络划分结果

分析图 6 中 karate 数据集并结合图 7 可知，当 $\lambda=0.2$ 时，优化 $D\lambda$ 函数，算法倾向于将整个网络划分为一个社区；当 $\lambda=0.3$ 时，算法能够准确地得出网络结构，如图 6(b)所示；当 $\lambda=0.4$ 时，NMI 值下降，算法检测出的社区结构与真实网络存在差异，结合图 7(c)可知，网络被划分为 3 个社区，与真实划分相比左边的社区被一分为二；当 $\lambda=0.5$ 时，网络同样被划分为 3 个社区，统计结果显示 NMI 值趋于稳定；当 $\lambda=0.6$ 或 0.7 时，算法同样显示出稳定性，且网络被划分为 4 个社区，与真实划分相比，左右两个社区都被划分为两个，如图 7(d)所示。由此也能证明不同的 λ 值趋向于不同的划分结果，其值较大时，更容易发现网络中较小的社区；反之，则更容易发现大社区。由此，模块度密度的分辨率限制问题也得以解决。

结束语　根据复杂网络条件下的社区检测问题，提出了一种基于单目标 PSO 的社区检测算法。与传统的社区检测算法相比，本算法为基于局部信息的启发式算法，采用合适的编码方式和粒子群更新方法，优化不同分辨率下的模块度密度函数，以解决离散型的社区检测问题。实验结果证明，本算法是有效的，解决了模块度密度的分辨率限制问题，能够检测出不同分辨率下的社区结构。

参 考 文 献

[1] Roxborough T, Sen A. Graph clustering using multiway ratio cut (Software demonstration)[C]// Graph Drawing. Springer Berlin Heidelberg, 1997: 291-296

[2] Angelini L, Boccaletti S, Marinazzo D, et al. Identification of network modules by optimization of ratio association[J]. arXiv preprint cond-mat/0610182, 2006

[3] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[4] Fortunato S, Barthelemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41

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

[6] Li Zhen-ping, Zhang Shui-hua, Wang Rui-sheng, et al. Quantitative Function for Community Detection[J]. Phys. Rev. E, 2008, 77(3): 036109

[7] Kennedy J. The particle swarm: social adaptation of knowledge [C]// IEEE International Conference on Evolutionary Computation, 1997. IEEE, 1997: 303-308

[8] 陈琳，何嘉. 基于模糊聚类的粒子群优化算法[J]. 西南民族大学学报：自然科学版，2007，33(4)：739-742

[9] Gong M, Cai Q, Li Y, et al. An improved memetic algorithm for community detection in complex networks[C]// 2012 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2012: 1-8

[10] Gong M, Cai Q, Chen X, et al. Complex Network Clustering by Multiobjective Discrete Particle Swarm Optimization Based on Decomposition[J]. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 82-97

[11] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): P09008

[12] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110

[13] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826