第 28 卷第 9 期
2011 年 9 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 28 No. 9
Sep. 2011

文章编号: 1000−8152(2011)09−1135−06

网络社区发现的粒子群优化算法

黄发良[1,2], 肖南峰[1]

(1. 华南理工大学 计算机科学与工程学院, 广东 广州 510640; 2. 福建师范大学 软件学院, 福建 福州 350007)

摘要: 从优化模块度的角度出发, 提出了一种基于粒子群优化的网络社区发现的粒子群优化算法(CDPSO); 该算法根据网络连接数据的特点给出一种新的粒子编码方法, 有效地避免非法粒子的产生, 一定程度上缓解了基于二值编码的迭代二划分策略所遭遇的局部最优划分问题, 并改进了传统离散粒子群优化(PSO)的粒子位置调整策略, 使算法收敛速度更快. 实验结果表明, CDPSO能够在无先验信息的条件下快速有效地揭示网络内在的社区结构.

关键词: 粒子群优化; 社区结构; 模块度

中图分类号: TP273　　**文献标识码**: A

Particle-swarm-optimization algorithm to discover network community

HUANG Fa-liang[1,2], XIAO Nan-feng[2]

(1. School of Computer Science and Engneering, South China University of Technology, Guangzhou Guangdong 510640, China;
2. Faculty of Software, Fujian Normal University, Fuzhou Fujian 350007, China)

Abstract: For optimizing the modularity, a community discovery algorithm(CDPSO) is proposed based on particle-swarm-optimization(PSO). By the characteristics of network link data, a novel particle-encoding scheme is presented to avoid the production of illegal particles, alleviate the local optimal-partition encountered in the iterative partition approach based on Boolean encoding scheme, and improve the particle-position adjustment strategy in traditional discrete PSO to achieve better convergence. Experimental results show that CDPSO can rapidly and effectively discover the intrinsic community structure in networks without any domain information.

Key words: particle swarm optimization; community structure; modularity

1 引言(Introduction)

互联网的迅猛发展极大地推动了社会信息的网络化进程, 以即时通讯系统、邮件网络等为代表的信息网络已经深入到人们的工作、学习与生活等各种活动中去, 成为人们生产生活的重要组成部分, 是构成信息社区的基础环境. 这些各式各样的信息网络承载着人们在生产生活中形成的复杂关系, 从这些纷繁芜杂的关系结构发现隐藏的潜在有价值的关系模式是一个非常困难而又很有意义的工作.

以信息网络为代表的复杂网络可以看成是用图结构表示的异质多关系数据集, 图中节点表示网络中的个体, 边表示个体之间的某种关系(诸如WEB页面之间的链接关系). 狭义上讲, 一个网络社区就是一个具有这样特点的个体集合: 集合内的个体之间的链接稠密而集合内个体与集合外个体的链接稀疏.

从本质上说, 网络社区发现就是从连通图中识别出在某种性质达到局部最优的稠密子图. 网络社区发现正吸引着越来越多研究者的注意, 有关社区发现的各种方法不断涌现. 本文对此类方法进行了系统分析并给出了初步的分类[1]: a) 基于数据挖掘的层次聚类与划分聚类算法; b) 基于分割的方法; c) 基于模块度优化的方法; d) 基于谱分析的方法. 其中基于模块度优化的方法备受关注, 究其本质, 这是一类对聚类目标函数设计与优化的方法. 近年来, 以PSO (particle swarm optimization)算法[2]为代表的各种进化算法备受关注并广泛应用于各种领域中的优化计算. PSO算法主要是通过对鸟群、鱼群等群体行为机制的模型模仿来实现简单生物群体涌现智能, 其具有思想简单、参数不多、易于实现与收敛速度较快等优点. 但是将PSO算法用于社区发现的研究报告却很少见.

基于此, 结合网络社区发现的问题, 本文提出一种新的基于粒子群优化思想的网络社区发现算法网络社区发现的粒子群优化算法(CDPSO), 该方法首先给出了一种新的粒子编码方式, 使之有效地避免非法粒子的产生与一定程度上缓解了基于二值编码的迭代二划分策略所遭遇的局部最优划分问题,

收稿日期: 2010−04−13; 收修改稿日期: 2010−10−15.

基金项目: 国家自然科学基金与中国民用航空总局联合资助项目(60776816); 广东省自然科学基金重点资助项目(8251064101000005); 广东省科技攻关资助项目(2007B06040107); 福建省教育厅科研基金资助项目(JA10076); 国家自然科学基金资助项目(61171141).

并改进了传统离散PSO(DPSO)的粒子位置调整策略, 使算法收敛速度更快, 最后将该算法运用于3个真实数据集(Karate, Football, Dolphins), 实验结果表明, 该方法的收敛速度非常快, 并且社区划分质量比较理想.

2 离散粒子群优化算法(Discrete PSO algorithm)

PSO的搜索空间是实数空间, 而现实中的许多优化问题都是离散解空间, 为此, Kennedy与Eberhart等提出离散型PSO(DPSO)[3], 下面对DPSO简单说明. 设问题解空间为d维欧氏空间, 对于由m个粒子组成的种群, 其中一个粒子对应于该空间中的一个点, 每个粒子都有位置属性与速度属性, 第i个粒子的位置可以表示为$X_i = (x_{i1}, x_{i2}, \cdots, x_{id})$, $x_{id} \in \{0, 1\}$, 类似地, 速度表示为$V_i = (v_{i1}, v_{i2}, \cdots, v_{id})$, 将$X_i$代入目标函数可以计算该粒子的适应度值, 并根据适应度值判断该粒子的优劣, 粒子的自学习能力主要体现在它能记忆自身迄今为止搜索到的最优位置, 即$P_i = (p_{i1}, p_{i2}, \cdots, p_{id})$, 粒子还可通过协作实现群体信息的共享, 这主要表现在粒子能够感知到整个粒子群的历史最优位置, 即为$P_{\rm g} = (p_{\rm g1}, p_{\rm g2}, \cdots, p_{{\rm g}d})$. 粒子根据式(1)(2)更新其速度和位置.

$$V_i(t+1) = wv_i(t) + c_1 \cdot \mathrm{rand}(\cdot) \cdot (P_i - X_i(t)) + c_2 \cdot \mathrm{rand}(\cdot) \cdot (P_{\rm g} - X_i(t)), \tag{1}$$

$$f(x) = \begin{cases} 1, \ \rho < \mathrm{sig}(v_{ij}(t+1)), \\ 0, \ \text{其他}, \end{cases} \tag{2}$$

其中: t为进化代数, w为惯性系数, c_1与c_2称为学习因子或加速常数, 可取常数也可根据算法需要进行动态修正, $\mathrm{rand}(\cdot)$为均匀分布在[0, 1]之间的随机数, 为了防止S型函数sig饱和, 常将v_{ij}的取值限制在区间$[-4, 4]$上, ρ为预定阈值.

3 基于粒子群算法的网络社区发现(Discovering network communities based on PSO)

3.1 粒子编码(Particle encoding)

网络社区的划分问题本质上是一个网络实体的硬聚类问题, 在现有的基于进化计算的数据聚类研究中, 个体编码的方式主要有二值编码、整数编码与实数编码[4], 其中与本文工作非常类似的有基于节点标号的编码方式[5]:

令网络$G = \langle V, E\rangle$, 一个粒子对应于一种划分, 粒子的维数为$|V|$, 对于粒子$P_i = (p_{i1}, p_{i2}, \cdots, p_{id})$, 若$p_{ik} = m$ 则表示结果社区中存在边$e = \langle v_k, v_m\rangle$, 这意味着节点$v_k$与$v_m$处于同一个社区中.

基于节点标号的编码方式简单直观, 但隐含着一个缺陷, 即粒子随机产生与飞行的过程中难以避免生成这样的非法粒子: 粒子位置分量所表示的边在原网络图中不存在, 如图2(b)所示, 粒子所表示的边$\langle v_3, v_6\rangle$, $\langle v_5, v_1\rangle$与$\langle v_7, v_2\rangle$在图1(a)的网络中不存在.

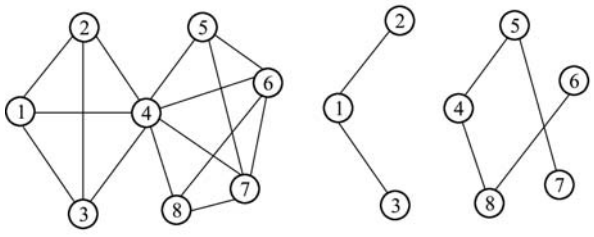


图 1 网络及其社区

Fig. 1 Network and its communities

Dim	1	2	3	4	5	6	7	8
Pos	2	1	1	5	7	8	5	4

(a) 合法粒子

Dim	1	2	3	4	5	6	7	8
Pos	1	1	1	4	3	4	2	1

(b) 非法粒子

Dim	1	2	3	4	5	6	7	8
Pos	2	1	6	5	1	8	2	4

(c) 本文方法编码的粒子

图 2 粒子编码方案的比较

Fig. 2 Comparison of particle encoding scheme

本文提出基于节点邻居有序表的编码方式, 其基本思想是: 首先对图中所有节点进行编号, 然后对每个节点根据其编号进行排序形成邻居有序表, 显然该列表的长度是该节点的度, 在初始化或粒子位置更新阶段生成新粒子时, 确保该粒子的合法性. 以图1(a)中的网络为例, 首先建立各节点的邻居有序表(表1), 根据此表可以对社区(图1(b))进行粒子编码, 结果见图2(a). 该编码方式有以下两个优势: 第一, 基于此编码方式的算法自动确定簇数, 这有效地回避了进化计算驱动的聚类算法需要用户事先指定的难题; 第二, 缓解基于二值编码的迭代二划分策略[5]所遭遇的容易陷入局部最优划分的问题.

表 1 邻居有序表

Table 1 Ordered neighbor list

中心节点	邻居节点有序列表						
1	2	3	4	—	—	—	—
2	1	3	4	—	—	—	—
3	1	2	4	—	—	—	—
4	1	2	3	5	6	7	8
5	4	6	7	8	—	—	—
6	4	5	7	8	—	—	—
7	4	5	6	8	—	—	—
8	4	6	7	—	—	—	—

3.2 粒子更新策略及其收敛性(Particle updating strategy and its convergence)

在DPSO中, 粒子的更新主要包括速度与位置的更新, 对于粒子速度, 本文继承标准的更新方法(即公式(1)). 但是标准的位置更新方法与本文的粒子编码方式不相吻合: 标准DPSO限制粒子分量i的取值范围为0, 1, 而在基于节点邻居有序表的编码方式中, 粒子分量i取值范围为$x_{id} \in \{1, 2, \cdots, \deg(v_i)\}$, 为此, 本文提出如下的位置更新策略:

$$f(x) = \begin{cases} k, & \rho < \mathrm{sig}(v_{ij}(t+1)) \text{且} \deg(v_i) > 1, \\ x_i(t), & \text{其他}, \end{cases} \tag{3}$$

其中k为除当前连接邻居外的任一随机的邻居节点, 即$k = \mathrm{ceil}(\mathrm{rand} \cdot \deg(v_i))$且$k \neq x_i(t)$, 其中: ceil为上取整函数, $\deg(v_i)$表示节点v_i的度(在网络G中与节点v_i关联的边数).

位置更新策略的收敛性极大地决定了粒子群算法的收敛性, 在此对其机理进行简单讨论: 与传统DPSO的位置更新策略相比较, 本文的位置更新策略具有更强的全局探测能力, 其原因是在传统DPSO中, 若$\rho < \mathrm{sig}(v_{ij}(t+1))$, 则$x_i(t+1)$被赋予一个固定的值1, 而在本文的位置更新策略中, $x_i(t+1)$则被赋值为除当前连接邻居外的任一随机邻居节点的编号, 显然在本文的策略中, 粒子的发散性更强. 这种更新的发散性一方面有利于粒子全局探测能力的提升, 但另一方面却使得粒子的局部搜索能力变弱, 尤其到达进化后期粒子种群发散度达到最大值. 为了粒子使既具有本文的位置更新策略所赋予的更强全局探测能力, 同时使粒子群最终很容易靠近全局最优粒子, 即具有更好的收敛性. 本文在具体实现中采用如下策略: 当进化代数到达一定值时, 将sig函数进行简单变更, 即$\mathrm{sig}\, v_{ij} = |\dfrac{1-\exp(-v_{ij})}{1+\exp(-v_{ij})}|$. 这种简单变更可以使得当让速度趋为0时, Sigmoid函数值为0, 粒子的位改变率靠近0的可能性增大, 从而使得算法逐步稳定在全局最优粒子的附近或其上.

3.3 粒子适应度(Particle fitness)

由于每个粒子对应原网络的一种社区划分方案, 故粒子适应度就是社区划分的质量. 社区划分质量的评价函数有很多种, 不同的社区定义对应着不同的社区划分质量评价函数. 尽管基于全局定义的模块度$\mathcal{Q}$值函数[6]由于其假定零模型中不存在社区结构而导致其致命的“粒度受限”(resolution limit)问题, 但由于其简单有效而成为普遍采用的社区划分质量评价函数. 在此, 本文采用粒子对应的社区划分的模块度$\mathcal{Q}$值作为粒子的适应度.

定义 1　节点内部度. 对于第i个社区中的节点v, 其内部度为与节点v邻接且属于第i个社区的节点数.

假设粒子形成网络G的k划分, 则可定义$k \times k$的矩阵E, 其对角元素E_{ii}表示第i个社区中的节点内部度在图中所有节点度和的比例, 而非对角元素E_{ij}表示社区i的节点到其他社区j的边数在图中所有边数的比例. 基于此E给出$\mathcal{Q}$值函数:

$$\mathcal{Q}(y) = \sum_i (e_{ij} - a_i^2) = \mathrm{Tr}(E) - \|E^2\|, \tag{4}$$

其中$a_i = \sum\limits_j E_{ij}$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.

3.4 CDPSO算法(CDPSO algorithm)

在PSO算法基本框架的基础上, 结合网络社区发现的具体场景, 本文提出了CDPSO算法, 算法步骤描述如下:

Step 1　建立网络各节点的邻居节点编号表;

Step 2　设置粒子位置和速度的范围以及粒子群惯性因子w, 根据网络节点数设置粒子的位置向量和速度向量的维度;

Step 3　初始化粒子群, 运用粒子位置修正策略确保粒子合法;

Step 4　复制粒子的当前的位置向量到经验位置, 并将每个粒子当前位置的适应度复制到经验适应度;

Step 5　选出适应度最高的粒子, 并将其经验位置向量和经验适应度作分别为群最优位置和群最优经验;

Step 6　根据式(1)(3)对每个粒子更新自身的位置向量和速度向量, 运用粒子位置修正策略确保粒子合法;

Step 7　根据式(4)计算每个粒子的适应度, 并与其经验适应度相比较, 如果优于其经验, 则更新该粒子的经验位置及其适应度;

Step 8　计算出群体中最优粒子, 与当前群最优适应度相比较, 如果优于当前群最优粒子, 则更新群最优位置和群最优适应度;

Step 9　如果满足停止条件, 则输出群最优划分方案和适应度, 否则, 转到Step 4.

CDPSO算法步骤可划分为两大块: 第1块是Step 1～Step 3, 主要负责初始化算法执行所需要的相关参数与数据结构; 第2块是Step 4～Step 9, 主要通过粒子群在解空间中协作飞行来实现$\mathcal{Q}$值函数的寻优, 若迭代寻优的结束条件满足, 则将最优粒子解码成原始网络的一个社区划分, 算法结束.

4 实验(Experiments)

本文采用3个真实的网络数据来检验CDPSO算法的有效性与时间性能. 实验的环境是: CPU为Intel

Core 2 Duo 2.66 G, 内存2 G, OS为Windows XP 2002.

4.1 数据集(Datasets)

3个真实网络分别是: 第1个是Karate网络(图3), 该网络是美国一所大学中的空手道俱乐部成员间的相互社会关系, 共有34个节点, 78条连接边; 第2个是Football网络(图4), 该网络是Newman等人对美国大学生橄榄球联赛的2000个赛季的比赛情况进行分析整理而建立的网络, 包含115个节点与616条边; 第3个是Dolphins(图5), Lusseau等人对栖息在新西兰Doubtful Sound峡湾的一个宽吻海豚群体(该群体由两个家族共62只宽吻海豚组成)构造的海豚关系网. 共有62个节点, 159条边.

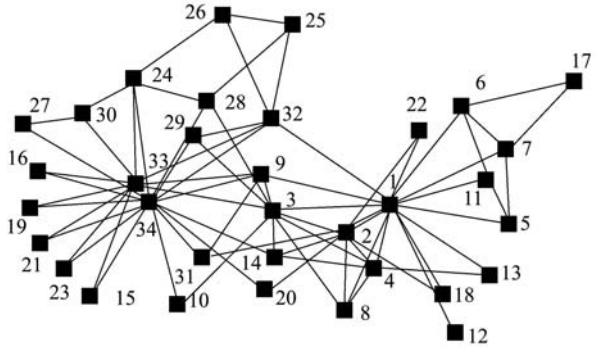


图 3 Karate网络

Fig. 3 Karate network

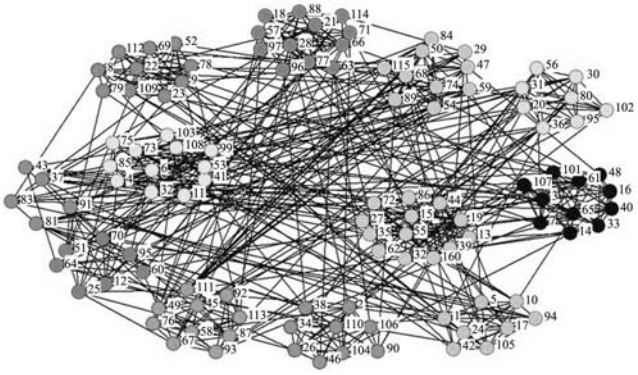


图 4 Football网络

Fig. 4 Football network

4.2 实验结果与分析(Experimental results and analysis)

4.2.1 算法收敛性分析(Algorithm convergence analysis)

首先从惯性因子来分析算法的收敛性, 从图6可以看出, 惯性因子在不同的网络中对算法收敛性的影响作用不同: 对于Karate数据集, 惯性因子大小为0.4∼0.7时, 算法收敛较快且都能收敛到最优Q值; 对于Dolphin网络, 惯性因子大小为0.6时, 算法收敛性最好; 对于Football网络, 惯性因子为0.4时算法收敛性最好. 比较分析可知, 惯性因子的影响作用与数据规模有很大关系.

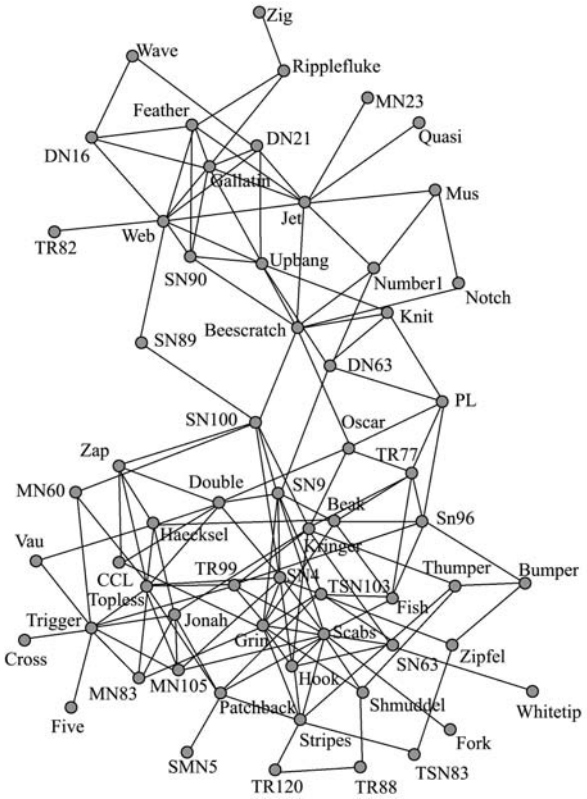
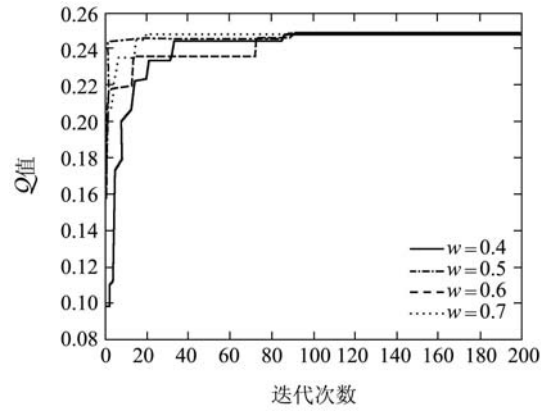
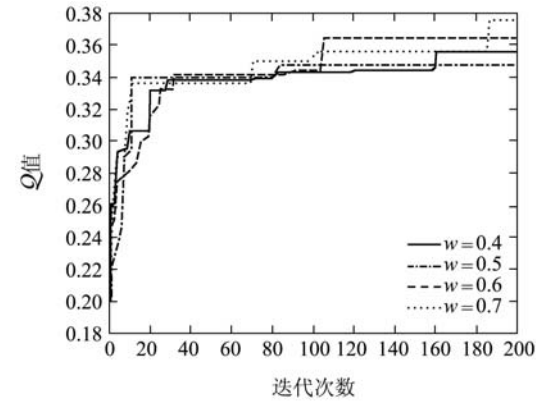


图 5 Dolphins网络

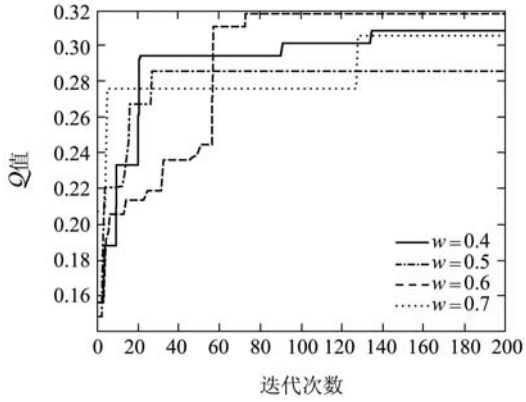
Fig. 5 Dolphins network



(a) Karate



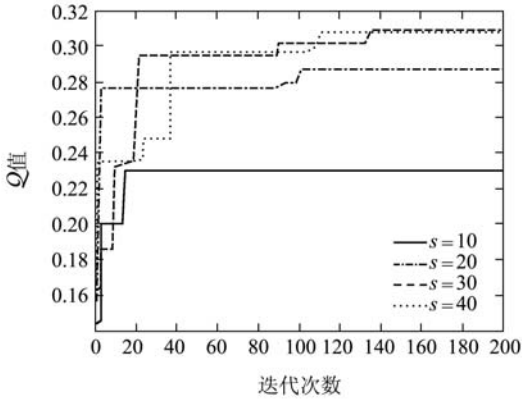
(b) Dolphins



(c) Football

图 6 惯性因子对收敛性的影响

Fig. 6 Inertia factor's impact on the convergence

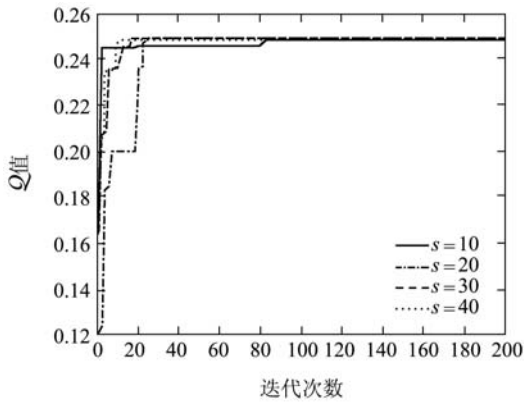


(c) Football

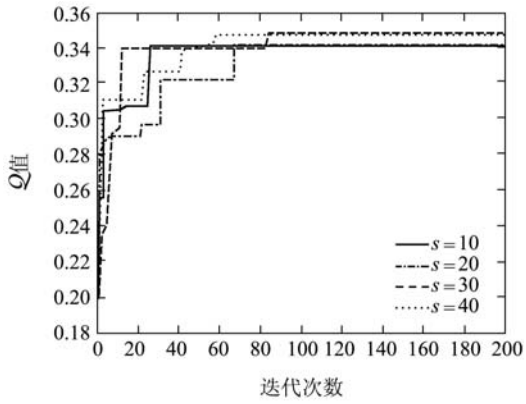
图 7 粒子数目对收敛性的影响

Fig. 7 Particle number's impact on the convergence

然后分析粒子个数对收敛性的影响, 从图7(a)可知, 对于Karate数据集, 种群大小为10∼40时, 算法收敛较快且都能收敛到最优$\mathcal{Q}$值, 而过大的种群规模并不能提高算法的效率. 在图7(b)中, 对于Dolphins网络, 种群大小为40时, 算法能较快的收敛到最大$\mathcal{Q}$值, 当种群大小为30时, 算法能收敛到最大$\mathcal{Q}$值, 但需要更多的迭代次数, 而对于种群大小为20与10时, 算法都陷入到局部最优值. 在图7(c)中, 对于Football网络, 其情形与Dolphins网络类似, 但当种群大小为10时, 其$\mathcal{Q}$值的最优性很差. 从上面分析可知, 粒子数目对算法性能的影响程度是与网络规模相关的.



(a) Karate



(b) Dolphins

4.2.2 网络社区质量分析(Quality analysis of network communities)

本文采用CE(clustering error)指标(式(5))来评价网络社区质量, 从表2中可以看出, 与GN算法[7]、GA-Net算法[8]相比较, CDPSO得到的社区质量高很多, 尤其在相对较小的网络中, 如Karate网络, 质量要高出一个数量级. 图8显示了Karate网络的社区划分情况, 该网络的实际社区两个, 分别是instructor(用圆表示)与administrator(正方形表示), 而CDPSO的结果也为两个社区, 分别instructor(用黑色标记)与administrator(灰色标记), 对比真实社区结构与计算所得社区结构可以发现, 只有节点10出现差异, 其他完全一致.

$$\mathrm{CE}(C, C^{\mathrm{true}}) = \frac{\sum\limits_{k_{\mathrm{true}}} \sum\limits_{k \neq \mathrm{true}} \left| C^{\mathrm{ture}}_{k_{\mathrm{ture}}} \cap C_k \right|}{n}. \quad (5)$$

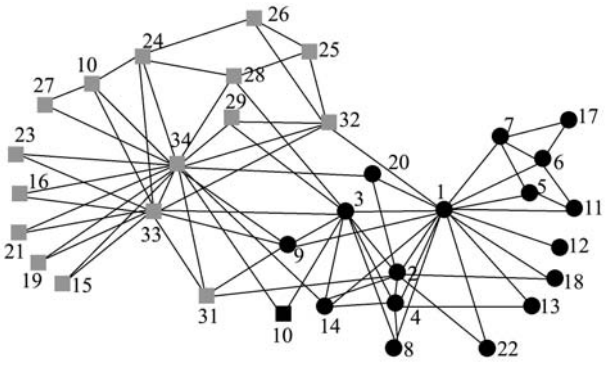


图 8 Karate网络的结果社区

Fig. 8 Resultant communities of Karate network

表 2 网络社区质量的比较

Table 2 Quality comparison of network communities

	Karate	Dolphins	Football
GN	0.2647	0. 3065	0.4261
GA-Net	0.1471	0.2581	0.5391
CDPSO	0.0294	0.0323	0.135

4.3 结论(Conclusions)

网络社区发现是网络数据挖掘领域中的一个重要研究方向. 现有的社区发现方法或者需要用户预先提供有关社区结构的先验知识, 或者对社区结构存在不合理的假设, 这使得不能有效地揭示真实复杂网络的内在社区结构. 本文从优化模块度的角度出发, 提出了一种基于粒子群优化技术的网络社区发现方法, 将社区发现问题转变为一个最大Q值寻优问题, 该方法根据网络连接数据的特点给出一种新的粒子编码方法与粒子位置更新策略. 真实网络上的实验结果表明, 该方法无须用户指定社区个数等算法参数, 能够有效地揭示网络内在的社区结构, 具有相对较好的收敛性与社区划分质量.

参考文献(References):

[1] 黄发良. 信息网络的社区发现及其应用研究[J]. 复杂系统与复杂性科学, 2010, 7(1): 64 – 74.
(HUANG Faliang. Studies on community detection and its application in information network[J], *Complex Systems and Complexity Science*, 2010, 7(1): 64 – 74.)

[2] KENNEDY J, EBERTHART R. Particle swarm optimization[C] //*Proceeding of IEEE International Conference on Neural Networks*. Perth, Australia: IEEE, 1995: 1942 – 1948.

[3] KENNEDY J, EBERHART RC, SHI Y. *Swarm Intelligence*[M]. San Francisco, CA: Morgan Kaufmann, 2001.

[4] HRUSCHKA E, CAMPELLO R J G B, FREITAS A A, et al. A survey of evolutionary algorithms for clustering[J]. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, 2009, 39(2): 133 – 155.

[5] 段晓东, 王存睿, 刘向东, 等. 基于粒子群算法的Web社区发现[J]. 计算机科学, 2008, 35(3): 18 – 21.
(DUAN Xiaodong, WANG Cunrui, LIU Xiangdong, et al. Web community detection model using particle swarm optimization[J]. *Computer Science*, 2008, 35(3): 18 – 21.)

[6] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69(2): 6113 – 6127.

[7] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences*, 2001, 99(12): 7821 – 7826.

[8] PIZZUTI C. GA-NET: a genetic algorithm for community detection in social networks[C] //*Proceedings of the 10th international Conference on Parallel Problem Solving from Nature*. Berlin, Heidelberg: Springer-Verlag, 2008: 1081 – 1090.

作者简介:

黄发良 (1975—), 男, 博士研究生, 研究方向为数据挖掘与模式识别, E-mail:faliang.huang@gmail.com;

肖南峰 (1962—), 男, 教授, 博士生导师, 研究方向为人工智能, E-mail:xiaonf@scut.edu.cn.