基于免疫离散差分进化算法的复杂网络社区发现

张英杰 [1] 龚中汉 [1] 陈乾坤 [1]

摘 要 针对复杂网络社区发现问题, 在标准差分进化算法的框架下, 提出一种新型免疫离散差分进化算法 (Immune discrete differential evolution, IDDE). 该算法通过标签传播策略生成初始种群, 采用离散差分进化策略来保证种群在问题空间的全局搜索能力, 同时对种群中的优秀个体执行针对性的高频克隆变异操作, 以提高算法的局部开发能力, 改善算法的收敛性能. 在计算机生成网络与真实世界网络中的仿真实验结果表明: IDDE 算法具有较强的寻优性能与鲁棒性, 能够有效探测复杂网络中存在的社区结构.

关键词 差分进化, 克隆选择, 社区发现, 模块度

引用格式 张英杰, 龚中汉, 陈乾坤. 基于免疫离散差分进化算法的复杂网络社区发现. 自动化学报, 2015, **41**(4): 749−757

DOI 10.16383/j.aas.2015.c140018

Community Detection in Complex Networks Using Immune Discrete Differential Evolution Algorithm

ZHANG Ying-Jie[1] GONG Zhong-Han[1] CHEN Qian-Kun[1]

Abstract Aimed at the existing problem of community detection in complex networks, a novel immune discrete differential evolution (IDDE) is proposed in the framework of standard differential evolution. In the proposed method, the initial population is generated through label propagation, and the discrete differential evolution strategy is utilized to ensure the global searching ability of the IDDE; meanwhile, the high-frequency clonal selection mutation operation is applied to excellent individuals of the population to improve the local exploitation ability and the convergence performance of the IDDE. Artificial networks and several real networks are employed to test the performance of the IDDE, and the testing results show that the IDDE achieves better searching ability and stronger robustness, and that it can detect the community structure in complex networks effectively.

Key words Differential evolution, clonal selection, community detection, modularity

Citation Zhang Ying-Jie, Gong Zhong-Han, Chen Qian-Kun. Community detection in complex networks using immune discrete differential evolution algorithm. *Acta Automatica Sinica*, 2015, **41**(4): 749−757

现实世界中的许多复杂系统都可以使用复杂网络来描述, 如: 互联网、社会网络、生物网络、科学家协作网络等. 其中, 网络中的结点可表示为系统中的个体, 边可表示为个体之间的关系. 由于复杂网络在不同学科之间的广泛适用性, 目前针对复杂网络的研究已经吸引了来自物理学、数学、生物学、计算机、社会科学与复杂性科学等众多领域研究者的关注. 大量研究结果表明: 复杂网络中存在着许多不同属性的社区结构, 且网络中存在的所有社区均满足不同社区之间的结点连接相对稀疏、同一社区内部的结点连接相对紧密的特性[1−2]. 社区发现的研究目的就是对复杂网络进行划分, 以提取出各社区结构中所蕴含的信息, 这些信息将有助于人们进一步分析理解复杂网络的拓扑结构、功能以及动力学特性等. 例如对互联网社区结构进行挖掘不仅可以改善网络搜索结果与提升用户体验, 还可以实现热点话题跟踪与网络舆情分析; 对社交网络进行挖掘可以发现具有相同兴趣爱好的社交圈; 科学家协作网络中的社区可以揭示具有相同研究领域或研究兴趣的科学家团体; 而生物网络中蛋白质网络的社区发现研究可以实现对蛋白质功能的预测等. 因此, 复杂网络社区发现的研究具有十分重要的理论意义与应用价值.

近年来, 来自不同领域的研究者提出了许多新颖的社区发现算法. 这些算法可大致分为: 基于聚类的算法[3−4], 基于标签传播的算法[5], 基于模块度优化的算法[6−7], 以及基于多目标优化的算法[8−9]. 其

收稿日期 2014-01-10 录用日期 2014-12-02
Manuscript received January 10, 2014; accepted December 2, 2014
国家自然科学基金 (61440026), 教育部博士点基金 (20110161110035), 湖南省自然科学基金重点项目 (13JJA002) 资助
Supported by National Natural Science Foundation of China (61440026), Doctoral Fund of Ministry of Education of China (20110161110035), and Hunan Provincial Natural Science Foundation of China (13JJA002)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 湖南大学信息科学与工程学院 长沙 410082
1. College of Information Science and Engineering, Hunan University, Changsha 410082

中, 基于模块度优化的社区发现算法是目前研究的最多的一类算法[10]. 这类算法的核心思想是将复杂网络中的社区发现问题转化为优化问题, 以模块度为目标函数, 通过对该目标函数的优化来寻找网络社区的最优划分. 由于最大化模块度值的问题是一个 NP 完全问题[11], 一般的算法很难在规定时间内求得满意的社区划分结果. 而进化算法作为解决这一类问题的有效手段, 目前已被广泛应用于求解复杂网络社区发现问题中. 其中, Guimerá 等首次提出使用模拟退火算法来解决网络社区发现问题, 并获得了较好的社区探测效果, 但是该算法时间代价较大, 需要做进一步的改进研究[12]; 文献 [13] 中提出了一种基于离散粒子群优化的网络社区发现算法, 并在该算法中给出了一种新型粒子编码方法, 最后通过实验验证了该算法改进的有效性; 文献 [14] 提出一种基于差分进化的社区发现算法, 该算法设计了一种改进的交叉算子, 并增加了 clean-up 操作来改善试验个体的质量, 获得了相对优异的收敛性能, 能够对网络中的社区结构进行有效划分; 文献 [6] 通过对模块度函数局部单调性的分析和对遗传算法的变异策略进行改进, 提出一种结合局部搜索算子的改进遗传算法用于求解社区发现问题, 并通过实验验证了该算法的有效性和高效性; 文献 [7] 提出一种改进遗传算法用于求解网络模块度, 并在计算机生成网络与 4 个真实网络中验证了算法的有效性, 但是该方法需要预先知道网络中存在的社区个数.

虽然上述算法均能对复杂网络中存在的社区结构进行有效的探测, 但是这些算法的性能以及社区模块划分精度还具有一定的提升空间. 本文根据 "无免费午餐定理"[15], 利用基本差分进化算法简单高效的优点, 在其基础上专门针对复杂网络社区发现问题的求解设计出一种免疫离散差分进化算法 (Immune discrete differential evolution, IDDE). 该算法采用标签传播的方式来生成具有一定精度的初始种群以加快算法的收敛, 通过离散差分进化算法的相关优化策略来提高算法的全局搜索能力, 同时利用免疫克隆选择策略对精英个体进行针对性高频克隆变异, 从而增强算法的局部搜索能力, 以期进一步提升算法的优化精度, 获得更高质量的网络社区划分结果.

1 免疫离散差分进化算法

1.1 问题的定义

不同的社区发现算法在对相同的复杂网络进行划分时, 所得到的社区划分结果是不相同的. 因此, 需要采用一种统一的量化指标来评价这些不同社区划分结果的质量. Newman 与 Girvan 提出了一种用于评价复杂网络社区划分结果优劣的模块度函数[2], 该函数主要是通过比较真实网络中社区内部的边连接密度与同等规模不具有社区结构的随机网络的边连接密度之间的差异来实现对社区划分质量的评价. 由于模块度函数计算简单且容易理解, 目前它在复杂网络社区发现问题的研究领域中已经成为应用最为广泛的质量评价指标之一[16].

对于给定的复杂网络 $N(V,E)$, 其对应的模块度函数 Q 具体定义如下:

$$Q=\frac{1}{2M}\sum_{ij}\left(a_{ij}-\frac{k_ik_j}{2M}\right)\delta(i,j) \tag{1}$$

其中, $A=(a_{ij})_{n\times n}$ 为网络 N 的邻接矩阵, M 为网络 N 中边的总连接数, k_i 与 k_j 分别表示网络 N 中结点 i 与结点 j 的度数. $\delta(i,j)$ 表示网络中结点与结点之间的关系, 当结点 i 与结点 j 处于网络中同一社区时, $\delta(i,j)=1$; 否则, $\delta(i,j)=0$. 如果 Q 值大于 0, 则表示复杂网络中开始出现社区结构; Q 值大于 0.3, 则表示复杂网络中存在明显的社区结构; Q 值越接近 1, 则复杂网络中的社区结构就越明显. 在实际复杂网络中, Q 值通常处于 $[0.3,0.7]$ 之间. 尽管模块度存在 "分辨率限制" 问题[17], 即不能从大型复杂网络中探测出一些规模较小的社区, 而是趋向于将这些小规模的社区合并成较大规模的社区, 但是根据文献 [6] 所述, 模块度函数 Q 作为社区划分质量评价指标仍具有普遍适用性.

1.2 个体编码与初始化

对于具有 n 个结点的复杂网络 $N(V,E)$ 的任意划分, 都可采取下述基于字符串编码的方式进行表示:

$$X(t)=[x_1,x_2,x_3,\cdots,x_n] \tag{2}$$

其中, x_i 表示结点 i 所属的社区标签, 且用 1 到 n 之间的整数表示. 如果 $x_i=x_j$, 则表明结点 i 与结点 j 处于网络中的同一社区; 否则, 这两个结点属于网络中的不同社区.

根据上述编码方式, 本文采用标签传播的方法来对整个种群中的个体进行初始化. 在初始化过程中, IDDE 算法首先将网络中的每一个结点都初始化为唯一的标签, 每一次迭代, 如果结点 i 的邻居结点中某一社区标签出现的次数最多, 则结点 i 的标签更改为该社区的标签, 如果结点 i 的邻居结点中不同标签出现的次数一样, 则随机选取其中一个作为结点 i 的标签; 如果算法达到规定的迭代次数或下一代个体对应的模块度值小于当前个体, 则初始化过程终止. 文献 [5] 中的仿真实验结果表明, 一般只需经过 5 次迭代, 即可生成具有较好精度与多样

性的初始种群.

1.3 离散差分进化算法

差分进化算法 (Differential evolution, DE) 是由 Storn 和 Price 于 1995 年提出的一种基于浮点数编码的群体智能优化方法[18]. 该算法实现简单, 控制参数少且优化性能较好, 目前已成功应用于科学与工程领域. 由于基本 DE 算法主要是为求解连续空间优化问题而设计, 因此不能直接用于求解离散空间优化问题. 而现实世界中有许多优化问题都是基于离散空间, 如: 旅行商问题[19]、流水线调度问题[20]、复杂网络社区发现问题等[21]. 为此, 本文针对社区发现问题的求解, 在基本 DE 算法的框架下, 分别采用离散型的变异策略与交叉策略, 设计了相应的离散差分进化算法.

1) 变异策略: 对于当前种群中的每一个个体, 本文采用随机单点变异策略对其执行变异操作[20]. 具体步骤如下: 首先从个体中随机选择出一个结点 i, 然后从结点 i 的所有邻居结点中随机选择出另一个结点, 并将该结点的标签值赋给结点 i, 这一过程将重复执行 n 次. 这种变异方式不会引入新的标签值而破坏种群中个体已有的良好性状, 可有效避免算法搜索的盲目性, 提高个体的利用率与算法的收敛效率. 表 1 给出了当选取结点 v_4 且其所有邻居结点中被选中的另一个结点的标签为 1 时随机单点变异策略一次执行的操作过程.

表 1　随机单点变异的一次演示

Table 1　An illustration of random one-point mutation

V	x	x_{new}
1	2	2
2	1	1
3	1	1
4	**2**	**1**
5	2	2
6	1	1

2) 交叉策略: 由于本文采用了基于字符串的编码方式对种群中的个体进行编码, 从而导致算法种群在进化过程中可能存在多个不同个体的编码对应同一社区划分的问题. 而传统的单点交叉策略只是通过简单地交换两个父代个体编码的某一段子串来执行交叉操作, 对于社区发现问题来说, 这种交叉方式并不能使父代个体的优良性状有效地遗传给子代个体而生成优秀的新个体, 反而更有可能破坏现有个体中已形成的良好性状. 因此, 为了保证 IDDE 算法中交叉操作的有效性, 文中选用了 Tasgin 等提出的单路交叉策略[21]. 该策略以变异个体作为目标个体 x_{dest}, 从当前种群中随机选择另一个个体作为源个体 x_{src}. 在交叉过程中, 首先, 随机从网络中选取某一结点 i, 并获取该结点在源个体 x_{src} 中的所属社区 C 以及该社区对应的标签; 然后, 在目标个体中找到所有在源个体中属于社区的结点; 最后, 将这些结点的所属社区标签全部更新为 r, 从而生成交叉后的新个体 x_{new}. 单路交叉策略的具体操作过程如表 2 所示.

表 2　单路交叉的一次演示

Table 2　An illustration of one-way crossing over

V	x_{src}	x_{dest}	x_{new}
1	**2**	5	**2**
2	1	3	3
3	1	3	3
4	**2**	3	**2**
5	**2**	5	**2**
6	3	2	2

3) 选择策略: 与基本差分进化算法类似, IDDE 算法仍然通过一对一的贪婪选择策略来生成下一代群体, 即采用模块度函数 Q 对交叉后生成的新个体进行评价, 如果优于父代个体, 则替换; 否则, 仍保留原来个体.

1.4 免疫克隆选择策略

为了提高离散差分进化算法的局部搜索能力, 改善算法的收敛性能, 防止算法出现早熟收敛问题, 本文引入了人工免疫系统中的克隆选择策略. 该策略的具体操作步骤如下:

步骤 1. 将当前种群中的所有个体按照其对应的模块度值的大小进行排序, 选择其中前几个优秀个体组成一个临时的克隆种群. 同时, 对临时种群中的每一个个体, 根据其对应的模块度值大小的排名按照式 (3) 进行克隆扩增, 从而生成一个中间群体 $nPop$:

$$N_i = round\left(\frac{\beta k}{j} + b\right) \tag{3}$$

其中, $k \le 0.2NP$, j 表示临时种群中个体模块度值大小的排名, $\beta \in (0,1)$, b 等于 1, 主要用于保证临时种群中每一个个体都具有一定的克隆数量.

步骤 2. 为了保证群体 $nPop$ 中个体的多样性, 对于群体中每个个体, 首先在个体中随机选择 m 个结点, 并赋予相同的标签值, 使这些结点属于同一社区, 促进算法收敛. 然后借鉴标签传播的思想, 针对网络中所有结点, 如果某结点对应的随机概率小于免疫克隆变异概率 p_{csm}, 则对该结点的标签值进行修正, 最终得到变异群体. 其中, $m \in [2, n/2]$.

步骤 3. 将当前种群中前 k 个优秀个体分别与其在变异群体中所产生的全部变异个体中的最优变异个体进行比较, 如果当前个体次于其对应的最优变异个体, 则选择对应的最优变异个体进入当前种群; 否则保留当前个体, 以保证当前种群不出现退化.

1.5 IDDE 算法流程

根据上述算法原理, 本文在基本差分进化算法的框架下, 给出 IDDE 算法的具体执行步骤如下:

步骤 1. 初始化 IDDE 算法的相关参数, 并输入网络 N 的邻接矩阵;

步骤 2. 采用标签传播的方法对种群中的个体进行初始化;

步骤 3. 评价种群中的每一个个体, 得到对应的适应度值, 并保存当前最优个体及其对应的适应度值;

步骤 4.

If (不满足结束条件) Then /* 结束条件为达到设定的最大算法迭代次数 */

1) 根据变异概率 p_m, 采用随机单点变异策略对种群中的每个个体执行变异操作, 生成对应的变异个体;

2) 根据交叉概率 p_c, 采用单路交叉策略对种群中的每个个体执行交叉操作, 生成对应的试验个体;

3) 采用贪婪选择策略, 分别将种群中的每个个体与其对应的试验个体进行竞争比较, 选择其中的较优个体组成中间群体;

4) 对差分进化策略产生的中间群体执行免疫克隆选择操作, 生成下一代种群并返回步骤 3;

End If

步骤 5. 算法运行结束, 输出对复杂网络社区结构的最优划分结果以及对应的模块度值.

1.6 IDDE 算法时间复杂度分析

实验仿真平台为 Windows XP SP3 系统, Intel Pentium (R) CPU 2.70 GHz, 内存 4.0 GB, 仿真软件为 Matlab R2011b. 对于具有 n 个结点的复杂网络, 在本文 IDDE 算法中, 模块度函数计算时间复杂度为 $\mathrm{O}(n^2)$; 标签传播初始化策略的时间复杂度 $T_{label}(n)$ 为 $\mathrm{O}(NP \times 5 \times (n \times d + n^2))$, 其中, NP 为种群规模, d 为网络中结点的平均度数; 离散差分进化策略的时间复杂度 $T_{DDE}(n)$ 为 $\mathrm{O}(maxIter \times NP \times (n \times d + n + n^2))$, 其中, $maxIter$ 为算法最大迭代次数; 免疫克隆选择策略的时间复杂度 $T_{cl}(n)$ 为 $\mathrm{O}(maxIter \times (NP^2 + NP^2 \times d \times (n^2 + 1)))$. 因此, 本文算法的时间复杂度 $T(n) = T_{label}(n) + T_{DDE}(n) + T_{cl}(n) = \mathrm{O}(an^2 + bn + c)$, 其中, $maxIter$, NP 与 d 均为常数, $a = maxIter \times NP^2 \times d + maxIter \times NP + 5NP$, $b = maxIter \times NP \times d + maxIter \times NP + 5NP \times d$, $c = maxIter \times NP^2 \times d + maxIter \times NP^2$.

2 实验与结果分析

本文分别采用社区结构已知的计算机生成网络与 6 个真实世界网络从不同的角度测试 IDDE 算法的收敛性能, 并与其他 5 种算法 GN[2]、CNM[22]、GA[21]、MA[23] 以及 LGA[6] 进行比较. 同时, 为了减少统计误差, 使实验对比结果显得更为可信, 本文所有实验均独立运行 50 次, 取结果的平均值进行比较. 实验参数设置: 本文其他对比算法的特定参数设置分别参照各自文献的参数设置说明. 而在本文提出的 IDDE 算法中, 种群规模 NP 设为 100, 算法最大迭代次数 $maxIter$ 设为 100, 变异概率 p_m 设为 0.5, 交叉概率 p_c 设为 0.9, 免疫克隆选择变异概率 p_{csm} 设为 0.2. 其中, 变异概率 p_m 与交叉概率 p_c 的值是基本差分进化算法的经典参数配置之一, 该组参数可扩大基本差分进化算法的搜索范围, 增强算法的全局搜索能力; 免疫克隆选择变异概率 p_{csm} 是本文在文献 [9] 的基础上依经验而设定, 以充分利用克隆选择策略的局部开发能力.

2.1 计算机生成网络

为了定量分析 IDDE 算法的优化性能, 本文采用 Lancichinetti 等于 2008 年提出来的一种社区结构已知的新型扩展 GN Benchmark 网络与 LFR Benchmark 网络[24]. 其中, 扩展 GN Benchmark 网络中含有 128 个结点, 共被划分为 4 个不同的社区, 每个社区中有 32 个结点, 社区中每个结点的平均度数为 16; LFR Benchmark 网络中共含有 1 000 个结点, 被划分为多个社区, 每个社区中结点的数目处于区间 $[10, 50]$ 之间, 社区中每个结点的平均度数为 20, 最大度数为 50, 且结点度数服从幂律分布, 因此, 随着网络参数的不同, LFR Benchmark 网络可呈现出小世界或无标度的特性. 上述网络中的混合参数 μ 主要用于确定某一社区中任意一个结点与其他社区的结点之间共享边的数量, 当混合参数 μ 小于 0.5 时, 网络中的社区结构较为明显, 但随着混合参数 μ 的不断增大, 网络中的社区结构将变得越来越模糊, 将导致性能一般的社区发现算法难以探测出网络中存在的社区结构, 从而达到区分不同社区发现算法的求解性能的目的.

根据信息理论的思想, 不同社区发现算法对复杂网络的划分结果与该网络中存在的真实社区之间的区别可通过归一化互信息 (Normalized mutual

information, NMI) 来进行评价[25]. 其具体定义如下:

$$NMI(A,B)=\frac{-2\sum\limits_{c_A}^{i=1}\sum\limits_{c_B}^{j=1}C_{ij}\log\left(\frac{C_{ij}N}{C_iC_j}\right)}{\sum\limits_{c_A}^{i=1}C_i\log\left(\frac{C_i}{N}\right)+\sum\limits_{c_B}^{j=1}C_j\log\left(\frac{C_i}{N}\right)} \tag{4}$$

给定复杂网络的两种不同社区划分结果 A 与 B, 混淆矩阵中 C 的元素 C_{ij} 表示划分 A 中的社区 i 与划分 B 中的社区 j 两者之间的交集中含有的结点数目, N 表示复杂网络中结点的总数目, c_A 与 c_B 分别表示 A 划分与 B 划分中含有的社区数目, C_i 与 C_j 分别表示混淆矩阵 C 中第 i 行与第 j 列的结点总数. 归一化互信息 NMI 的值越大, 则表明算法的社区划分结果与网络中存在的真实社区越接近, 即算法的优化性能越好. 如果算法的社区划分结果与网络中存在的真实社区完全一样, 则 NMI 的值等于 1. 根据这一评价指标, 图 1 给出了 IDDE 算法与其他对比算法在扩展 GN Benchmark 上的社区划分精度比较结果.

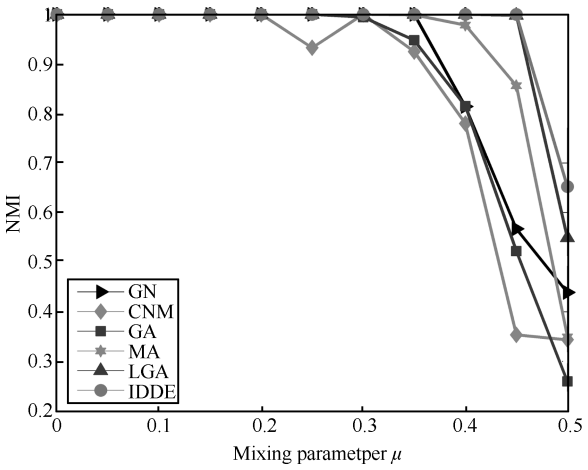


图 1　IDDE 与其他算法的平均 NMI

Fig. 1　Average NMI of IDDE and other algorithms

从图 1 可以看出, 随着混合参数 μ 的增长, IDDE 算法相对于其他 5 种对比算法的性能优势变得越来越明显. 当混合参数 μ 大于 0.35 时, 除了 IDDE 算法与 LGA 算法以外, 其他 4 种算法的社区划分质量都开始出现不同程度的下降; 即使当混合参数的 μ 等于 0.5, 网络中社区结构已经非常模糊的情况下, IDDE 算法仍能正确划分 65 % 的社区结构.

由于计算资源的限制, 在更大规模的 LFR Benchmark 网络上, 图 2 仅选取了 CNM、GA 与 LGA 三种算法与本文 IDDE 算法进行对比. 从图 2 可以看出, 当混合参数 μ 大于 0.05 时, 仅有 IDDE 算法与 LGA 算法能够正确分类网络中的结点; 当混合参数 μ 大于 0.2 时, IDDE 与 LGA 两种算法同时开始出现波动, 但与 LGA 算法相比, IDDE 算法显得更为稳定; 当混合参数 μ 大于 0.5 时, IDDE 与 LGA 两种算法性能开始出现明显下降, 但是随着混合参数 μ 的增大, IDDE 算法的性能仍优于其他算法.

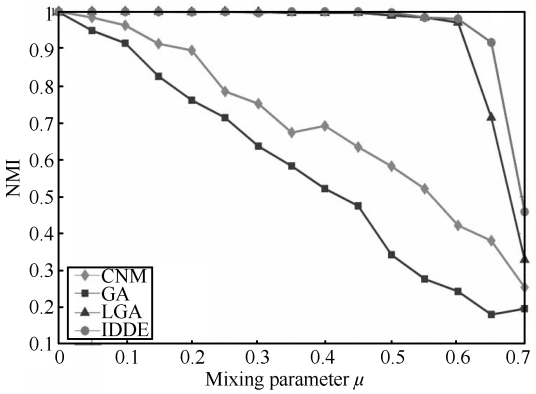


图 2　IDDE 与其他算法的平均 NMI

Fig. 2　Average NMI of IDDE and other algorithms

同时, 为了验证 IDDE 算法中初始化策略与免疫克隆选择策略对算法性能的影响, 分别选取基于随机初始化的离散差分进化算法 DDE (Random)、基于标签传播初始化的离散差分进化算法 DDE (LP)、基于随机初始化的免疫离散差分进化算法 IDDE (Random + Clonal selection) 以及基于标签传播初始化的免疫离散差分进化算法 IDDE (LP + Clonal selection) 在扩展 GN 网络上进行比较. 具体比较结果如图 3 所示.

从图 3 可以看出, 当混合参数 μ 等于 0.4 时, DDE (LP) 可以完整探测出网络中存在的真实社区结构, 而 DDE (Random) 在 μ 大于 0.3 时就已经很难发现真实的社区结构信息; 同样, 当混合参数 μ 等于 0.45 时, IDDE (LP + Clonal selection) 仍然能够完整地探测出网络中存在的社区结构, 而 DDE (LP) 此时的优化性能已经有所下降. 由此可知, 标签传播初始化策略与免疫克隆选择策略均能显著改善算法 IDDE 的性能, 但是免疫克隆选择策略明显在其中扮演了更为重要的角色, 因为该策略能够帮助 IDDE 算法从网络中发现更加模糊的社区结构.

2.2　真实世界网络

由于真实世界网络 (Real-world networks) 的拓扑结构特性与计算机生成网络不同, 因此, 为了进一步验证 IDDE 算法是否能够有效探测真实世界网络的社区结构, 本文选取 6 种不同网络对算法进行测试, 这些网络的具体描述如表 3 所示.

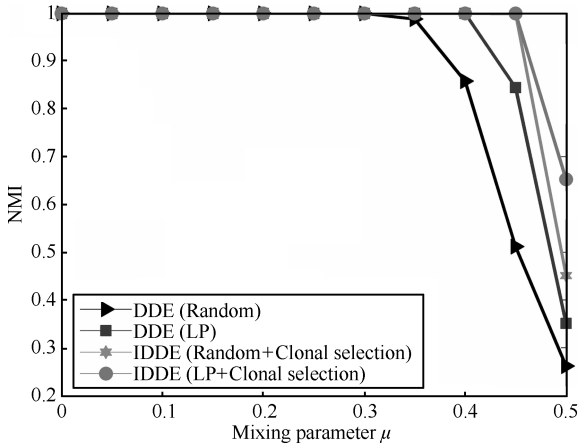


图 3 IDDE 不同优化策略对算法性能的影响

Fig. 3 Effects on algorithm performance of different optimized policies

表 3 真实世界网络

Table 3 Real-world networks

Networks	Number of nodes	Number of edges
Karate	34	78
Dolphins	62	159
Polbooks	105	441
Football	115	613
SFI	118	200
Jazz	198	2742

针对上述真实世界网络, 采用网络模块度 Q 作为算法性能的评价指标, 将 IDDE 算法与其他算法进行比较. 具体结果如表 4 所示.

Wilcoxon 秩和检验作为一种独立样本非参数统计检验方法[26], 可用于判断 6 种算法在 50 次运行过程中所求得的最优模块度值与这 6 种算法所求得的模块度值之间的差别是否具有统计学意义. 在 0.05 的显著性水平下, 如果检验结果值小于 0.05, 则拒绝零假设, 表明 IDDE 算法所求得的结果在任何情况下都是具有统计学意义的. 具体检验结果如表 5 所示, 其中 NA 表示进行检验的两个独立样本几乎完全一致.

从表 4 和表 5 可以看出, 在 6 个真实世界网络中, IDDE 算法的性能均优于 GN, CNM 以及 GA 这三种算法. 与 DDE 相比, IDDE 算法所求得的模块度值除了 Karate 网络, 其他均优于 DDE 算法; 与 MA 相比, IDDE 算法所求得的模块度均值在 6 个真实世界网络中均优于 MA, 但是从统计学角度来说, IDDE 算法仅在 Dolphins 网络, Football 网络与 Jazz 网络这 3 个网络中求得的结果与 MA 所求得的结果之间的差别具有统计学意义. 与 LGA 相比, 在 Karate 网络与 Football 网络中, IDDE 算法的划分精度与 LGA 相当; 在 Dolphins 网络与 SFI 网络中, IDDE 算法的性能要优于 LGA; 在 Jazz 网络中, IDDE 算法的性能则次于 LGA; 而在 Polbooks 网络中, IDDE 与 LGA 所求得的模块度均值相等, 但从统计学角度来说, LGA 性能稍次于 IDDE 算法. 对于 MA、LGA 与 IDDE 这三种算法, 它们的共同点在于都实现了局部搜索机制, 因此所求结果相近, 且与前三种算法相比, 取得了较大优势; 但是结合图 2 可以看出, 随着网络复杂性的增高, MA 不再适用, 而 IDDE 算法相对于 LGA 的优势也逐渐增大. 总的来说, 在求解真实世界网络社区发现问题中, IDDE 算法仍具有一定的优势.

下面通过两个社区结构已知的网络, 给出对 IDDE 算法性能更为详细的分析.

1) Karate 网络[27]. 该网络是基于美国一所大学中的空手道俱乐部成员之间的社会关系而构建的, 结点表示成员, 边表示成员之间的关系. 后来由于俱乐部中教练与管理员之间的内部分歧, 最终分裂为两个独立的俱乐部. IDDE 算法在该网络上一次随机运行所得到的划分结果如图 4 所示. 可以看出, IDDE 算法不仅能够正确划分出两个独立的俱乐部, 而且还分别从这两个俱乐部内部识别出了成员联系更紧密的两个子社区. IDDE 算法在 Karate 网络上随机运行 50 次的最优模块度值与平均模块度值分别为 0.4198 与 0.4198, 均大于该网络中真实社区结构的模块度值 0.3715, 表明 IDDE 算法在识别出真实网络社区的同时, 还能够进一步挖掘出真实社区中存在的小团体.

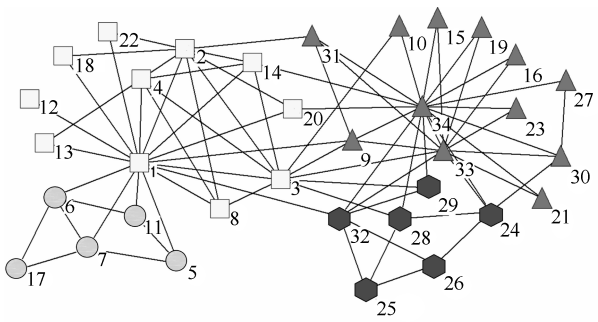


图 4 IDDE 对 Karate 网络的社区划分结果

Fig. 4 Community structure identified by IDDE algorithm on Karate network

2) Dolphins 网络[28]. 该网络是基于新西兰神奇湾中的海豚之间的联系而构建的, 结点表示海豚, 边表示海豚之间是否具有联系. 根据性别, 该网络被天然地划分为雄性海豚社区与雌性海豚社区. IDDE 算法在该网络上一次随机运行所得到的划分结果如图 5 所示. 可以看出, IDDE 算法不仅能够正确地区别雄性海豚与雌性海豚, 而且还对雌性海豚社区进行了更细致的划分. IDDE 算法在 Dolphins 网络上

表 4　IDDE 与其他算法的平均模块度

Table 4　Average modularity of IDDE and other algorithms

Networks	GN	CNM	GA	MA	LGA	DDE	IDDE
Karate	0.4013	0.3807	0.4067	0.4194	**0.4198**	**0.4198**	**0.4198**
Dolphins	0.5194	0.4955	0.5216	0.5261	0.5280	0.5275	**0.5282**
Polbooks	0.5100	0.5020	0.5222	0.5270	**0.5272**	0.5267	0.5272
Football	0.5995	0.5773	0.5911	0.6009	**0.6046**	0.6044	**0.6046**
SFI	0.7112	0.7335	0.7157	0.7505	0.7496	0.7473	**0.7506**
Jazz	0.3905	0.4389	0.4439	0.4444	**0.4449**	0.444	0.4448

表 5　IDDE 与其他算法的 Wilcoxon 秩和检验

Table 5　Wilcoxon rank sum test of IDDE and other algorithms

Networks	GN	CNM	GA	MA	LGA	DDE	IDDE
Karate	2.63E−23	2.63E−23	1.44E−08	0.3271	**NA**	**NA**	**NA**
Dolphins	6.80E−21	6.80E−21	5.98E−16	3.72E−10	3.87E−04	2.92E−09	**NA**
Polbooks	1.55E−22	1.55E−22	6.43E−16	0.0751	**0.1429**	8.55E−19	**NA**
Football	2.63E−03	2.63E−23	3.31E−20	2.15E−14	**NA**	2.63E−23	**NA**
SFI	4.27E−23	4.27E−23	4.73E−20	0.3004	2.23E−08	1.03E−22	**NA**
Jazz	1.02E−20	1.02E−20	3.60E−14	3.24E−05	**NA**	2.08E−17	7.39E−04

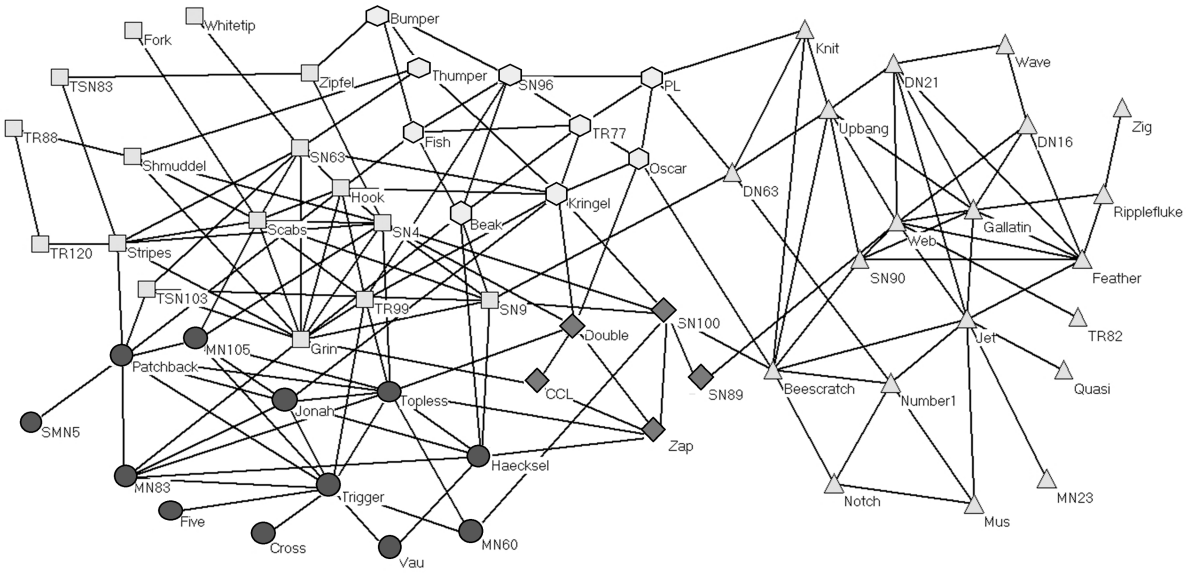


图 5　IDDE 对 Dolphins 网络的社区划分结果

Fig. 5　Community structure identified by IDDE algorithm on Dolphins network

随机运行 50 次的最优模块度值与平均模块度值分别为 0.5285 与 0.5282, 均大于该网络中真实社区结构的模块度值 0.3722, 验证了 IDDE 算法的社区挖掘性能.

3　结论

本文提出一种免疫离散差分进化算法用于求解复杂网络社区发现问题. 该算法首先通过标签传播的思想对种群进行初始化; 然后借鉴连续空间差分进化算法的基本特征, 采用随机单点变异、单路交叉及贪婪选择策略生成中间种群; 最后对中间种群执行免疫克隆选择操作, 以生成下一代种群. 上述改进模式充分利用了离散差分进化算法的全局搜索能力与免疫克隆选择策略的局部搜索能力, 有效改善了算法的收敛性能. 仿真实验结果表明: 本文提出的 IDDE 算法具有较强的寻优能力与鲁棒性, 能够有

效探测计算机生成网络及真实世界网络中存在的社区结构, 且社区划分质量与其他算法相比仍具有一定的优势.

References

1 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821−7826

2 Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**(2): 026113

3 Zhang S H, Wang R S, Zhang X S. Identification of overlapping community structure in complex networks using fuzzy means clustering. *Physica A: Statistical Mechanics and Its Applications*, 2007, **374**(1): 483−490

4 Huang Fa-Liang, Huang Ming-Xuan, Yuan Chang-An, Yao Zhi-Qiang. Spectral clustering ensemble algorithm for discovering overlapping communities in social networks. *Control and Decision*, 2014, **29**(4): 713−718
(黄发良, 黄名选, 元昌安, 姚志强. 网络重叠社区发现的谱聚类集成算法. 控制与决策, 2014, **29** (4): 713−718)

5 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale network. *Physical Review E*, 2007, **76**(3): 036106

6 Jin Di, Liu Jie, Yang Bo, He Dong-Xiao, Liu Da-You. Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, 2011, **37**(7): 873−882
(金弟, 刘杰, 杨博, 何东晓, 刘大有. 局部搜索与遗传算法结合的大规模复杂网络社区探测. 自动化学报, 2011, **37**(7): 873−882)

7 Shang R H, Bai J, Jiao L C, Jin C. Community detection based on modularity and an improved genetic algorithm. *Physica A: Statistical Mechanics and Its Applications*, 2013, **392**(5): 1215−1231

8 Huang Fa-Liang, Zhang Shi-Chao, Zhu Xiao-Feng. Discovering network community based on multi-objective optimization. *Journal of Software*, 2013, **24**(9): 2062−2077
(黄发良, 张师超, 朱晓峰. 基于多目标优化的网络社区发现方法. 软件学报, 2013, **24**(9): 2062−2077)

9 Gong M G, Cai Q, Chen X W, Ma L J. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. *IEEE Transactions on Evolutionary Computation*, 2014, **18**(1): 82−97

10 Luo Zhi-Gang, Ding Fan, Jiang Xiao-Zhou, Shi Jin-Long. New progress on community detection in complex networks. *Journal of National University of Defense Technology*, 2011, **33**(1): 47−52
(骆志刚, 丁凡, 蒋晓舟, 石金龙. 复杂网络社团发现算法研究新进展. 国防科技大学学报, 2011, **33**(1): 47−52)

11 Brandes U, Delling D, Gaertler M, Goerke R, Hoefer M, Nikoloski Z, Wagner D. Maximizing modularity is hard. arXiv: physics/0608255, 2006.

12 Guimerá R, Sales-Pardo M, Amaral L A N. Modularity from fluctuations in random graphs and complex network. *Physical Review E*, 2004, **70**(2): 025101

13 Huang Fa-Liang, Xiao Nan-Feng. Particle-swarm-optimization algorithm to discover network community. *Control Theory and Application*, 2011, **28**(9): 1135−1140
(黄发良, 肖南峰. 网络社区发现的粒子群优化算法. 控制理论与应用, 2011, **28**(9): 1135−1140)

14 Jia G B, Cai Z X, Musolesi M, Wang Y, Tennant D A, Weber R J, Heath J K, He S. Community detection in social and biological networks using differential evolution. In: Proceedings of the 6th International Conference on Learning and Intelligent Optimization Conference LION6. Heidelberg: Springer, 2012. 71−85

15 Wolpert D H, Macready W G. No free lunch theorems for optimization. *IEEE Transaction on Evolutionary Computation*, 1997, **2**(1): 62−87

16 Santo F. Community detection in graphs. *Physics Reports*, 2010, **486**(3−5): 75−174

17 Fortunato S, Barthelemy M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(1): 36−41

18 Storn R, Price K. Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 1997, **11**(4): 341−359

19 Karabulut K, Tasgetiren M F. A discrete artificial bee colony algorithm for the traveling salesman problem with time windows. In: Proceedings of the 2012 IEEE Congress Evolutionary Computation. Piscataway, NJ: IEEE, 2012. 1−7

20 Pan Q K, Mehmet F T, Liang Y C. A discrete differential evolution algorithm for the permutation flowshop scheduling problem. *Computers & Industrial Engineering*, 2008, **55**(4): 795−816

21 Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms. arXiv: 0711.0491, 2007.

22 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, **70**(6): 066111

23 Gong M G, Fu B, Jiao L C, Du H F. Memetic algorithm for community detection in networks. *Physical Review E*, 2011, **84**(5): 056101

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

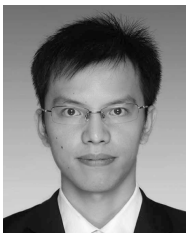
25 Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, **2005** (09): P09008

26 Derrac J, García S, Molina D, Herrera F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 2011, **1**(1): 3−18

27 Zachary W W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1997, **33**(4): 452−473

28 Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E, Dawson S M. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 2003, **54**(4): 396−405

张英杰 湖南大学信息科学与工程学院副教授. 2005 年获得湖南大学控制理论与控制工程博士学位. 主要研究方向为智能控制, 计算智能和节能优化控制. 本文通信作者.
E-mail: zhangyj@hnu.edu.cn
(**ZHANG Ying-Jie** Associate professor at the College of Information Science and Engineering, Hunan University. He received his Ph. D. degree in control theory and control engineering from Hunan University in 2005. His research interest covers intelligent control, computational intelligence, and energy-optimized control. Corresponding author of this paper.)



龚中汉 湖南大学信息科学与工程学院硕士研究生. 2011 年获得湖南大学信息科学与工程学院学士学位. 主要研究方向为进化计算和数据挖掘.
E-mail: hncl18@hnu.edu.cn
(**GONG Zhong-Han** Master student at the College of Information Science and Engineering, Hunan University. He received his bachelor degree from Hunan University in 2011. His research interest covers evolutionary computation and data mining.)

陈乾坤 湖南大学信息科学与工程学院硕士研究生. 2012 年获得河南科技大学车辆与动力工程学院学士学位. 主要研究方向为社会计算.
E-mail: s1210w103@hnu.edu.cn
(**CHEN Qian-Kun** Master student at the College of Information Science and Engineering, Hunan University. He received his bachelor degree from Henan University of Science and Technology in 2012. His research interest covers social computing.)