

基于多种群遗传算法的复杂网络社区结构发现*

刘发升, 罗延榕

(江西理工大学 信息工程学院, 江西 赣州 341000)

摘 要: 提出了一种基于多种群遗传算法的复杂网络社区结构发现新算法, 该算法无须预先知道社区内节点的数量以及任何门限值, 同时引入并行遗传算法的思想, 进一步提高了算法的运行效率。实验结果表明, 与传统算法相比, 在无先验信息的条件下, 使用该算法对不同规模的网络图 Zachary 和 Dophins 网络结构进行验证时, 能够以较低的时间复杂度、高效并准确地完成对网络社区的有效划分。

关键词: 复杂网络; 网络社区; 社区结构; 多种群; 遗传算法

中图分类号: TP393

文献标志码: A

文章编号: 1001-3695(2012)04-1237-04

doi:10.3969/j.issn.1001-3695.2012.04.010

Community structure discovery in complex networks based on multi-population genetic algorithm

LIU Fa-sheng, LUO Yan-rong

(School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou Jiangxi 341000, China)

Abstract: This paper proposed a new model to detect community structure in complex network based on multiple-population genetic algorithm. It didn't need any prior knowledge about the numbers of community and any threshold values, introduced simultaneously parallel genetic algorithm to enhance the efficiency. The numerical experiments show that this algorithm can greatly reduce the time complexity and get more accurate optimum partiton of network structure without any prior information compared with traditional algorithm, by using this new algorithm to test the two networks with different scale named Zachary and Dophins.

Key words: complex networks; Web community; community structure; multiple-population; genetic algorithm

复杂网络社区发现的研究近年来日益受到来自物理学和计算机科学领域的科研人员的广泛关注。现实中的很多系统都可以用复杂网络来描述。复杂网络中的节点可表示为复杂系统中的个体, 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的一种关系。现实世界中包含着各种类型的复杂网络, 如互联网、社会网络、生物网络等, 这些网络都具有一种普遍的特性——社区结构。大量实证研究表明, 这些网络的构成并不是杂乱无章的, 往往是由许多小的网络社区构成的, 这些社区内部的节点之间连接非常紧密, 但是各个社区之间的连接却比较稀疏^[1]。如何对复杂网络社区进行探测和划分对于了解整个网络的结构和性质具有十分重要的意义。

1 相关的研究

网络社区结构的研究已经有很长的历史, 其起源于社会学的研究工作, 与社会学中的层次聚类(hierarchical clustering)和计算机科学中的图形分割(graph partition)有着密切的关系, 其代表性的算法有 Kernighan-Lin 算法^[2]、基于 Laplace 图特征的谱平分法^[3,4]、W-H 算法^[5]、GN 算法^[6]、Newman 快速算法^[7]等。

Kernighan-Lin 算法是一种基于贪婪算法原理, 将网络分割为两个大小已知的社区的二分法, 但该算法必须已知网络社区

的确切规模才能得到正确结果, 使其在实际网络分析中难以得到较好的应用。基于 Laplace 图特征值的谱平分法, 利用网络结构的 Laplace 矩阵中不为零的特征值所对应的特征向量和同一个社区内的节点对应的元素近似相等的原理对网络社区进行划分, 因此, 当网络明显分成两个社区时, 该算法是快速的, 否则就未必有效。Wu 和 Huberman 提出了基于电阻网络电压谱的快速分割算法, 它将网络看成一个电阻网络, 然后利用 Kirchhoff 定理^[5]求解各个节点的电压值, 绘制电压谱, 进而根据不同的阈值进行社区划分, 但该方法需已知分属于不同社区的两个节点。GN 算法就是一种分裂方法, 它的基本思想是, 通过不断地从网络中移除边数(网络中经过每条边的最短路径的数)最大的边将整个网络分解为各个社区。GN 算法弥补了一些传统算法的不足, 但是它的时间复杂度却比较大, 为 $O(n^3)$, 并且在不确定社区数目的情况下, GN 算法也不知道这种分解要进行到哪一步终止。因此, Newman 等人经过研究提出了一种度量网络社区划分质量的标准——模块度 Q 。

通过对现有网络社区发现算法的分析可以看出, 设计一种不需要先验信息且问题不相关、时间复杂度较低的算法, 进而对各种规模的网络进行划分, 具有现实的应用价值。

2 网络社区划分衡量标准——模块度 Q

一个无方向无权值的网络可以用 $G(V, E)$ 表示, 其中, V 为

收稿日期: 2011-08-19; 修回日期: 2011-09-29 基金项目: 江西省教育厅科技资助项目(GJJ08283, GJJ11463)

作者简介: 刘发升(1963-), 男, 江西大余人, 教授, 博士, 主要研究方向为数据挖掘与数据库技术; 罗延榕(1986-), 女, 硕士研究生, 主要研究方向为数据挖掘(120574553@qq.com)。

网络节点集合, E 为网络边集合。邻接矩阵 A 是社区网络的另一种表示, 对于有 n 个节点的网络, 令 $A = (a_{ij})$, $i, j = 1, 2, 3, \dots, n$, 如果节点 i 和节点 j 相连, 则 $a_{ij} = 1$, 否则 $a_{ij} = 0$ 。一般认为节点自身不存在连接, 则 $a_{ii} = 0, i = 1, 2, 3, \dots, n$ 。

设网络节点集合 V_p, V_q 是 V 的真子集即网络社区, 即 $V_p \neq \emptyset, V_q \neq \emptyset$, 且 $V_p \subset V, V_q \subset V$ 。若 $V_p \cap V_q \neq \emptyset$, 则有 $A_{pq} = \{a_{ij}\}, i \in V_p, j \in V_q$, 且 $A_{pq} \subset A$ 。记 $\|A\| = \sum_{i=1}^n \sum_{j=1}^n a_{ij}$, 因此当 $p = q$ 时, $\|A_{pp}\| = \sum_{i \in V_p} \sum_{j \in V_p} a_{ij}, a_{ij} \in A$ 为社区 V_p 内部的关系数量总和; 当 $p \neq q$ 时, $\|A_{pq}\| = \sum_{i \in V_p} \sum_{j \in V_q} a_{ij}, a_{ij} \in A$ 为社区 V_p, V_q 之间的关系数量总和。网络社区结构划分就是将节点集合划分为 m 个子集合 $V_1, V_2, \dots, V_m$, 使其满足: a) $V_p \neq \emptyset, p = 1, 2, 3, \dots, m$; b) $\bigcup_{p=1}^m V_p = V$; c) $V_p \cap V_q = \emptyset, p \neq q, p, q = 1, 2, 3, \dots, m$ 。在应用中, 定义 $e_{pq} = \|A_{pq}\| / \|A\|$, 则网络社区结构划分只要同时保证 $\max(\sum_p e_{pp}), p = 1, 2, 3, \dots, m$ 和 $\min(\sum_{p,q} e_{pq}), p \neq q$ 且 $p, q = 1, 2, 3, \dots, m$, 就可以满足社区内部的关系密集而社区之间的关系稀疏, 也即网络社区结构高内聚、低耦合的特点。因此, Newman 将度量社区结构划分的模块性指标定义为^[6]

$$Q = \sum_{p=1}^m [e_{pp} - (\sum_{q=1}^m e_{pq})^2] \quad (1)$$

若 Q 的值大于 0.3, 则网络存在明显的社区结构^[8]。 Q 值的上限为 1, 而 Q 越接近这个值, 就说明社区结构越明显。而在实际网络中, Q 的取值通常位于 0.3 ~ 0.7 之间。

3 基于多种群遗传算法的复杂网络社区结构发现

3.1 多种群遗传算法

遗传算法 (genetic algorithm, GA) 是模拟生物进化的自然选择和遗传机制的一种寻优算法^[9]。遗传算法模拟了生物的繁殖、交配和变异现象, 从任意一初始种群出发, 产生一群新的更适应环境的后代。这样一代一代不断繁殖、进化, 最后收敛到一个最适应环境的个体上。遗传算法对于复杂的优化问题无须建模和进行复杂运算, 只需要利用遗传算法的算子就能找到问题的最优解或满意解。

传统的遗传算法存在两个重大缺陷: a) 容易早熟, 即收敛提前结束, 陷入局部最优解; b) 进化后期搜索效率低, 这使得最终得到的结果往往不是全局最优解。为避免以上情况, 本文提出了一种多种群遗传算法。通过多种群并行进化的思想, 将遗传算法在多个子种群间并行进行, 并通过精英迁移策略在子种群间交换信息 (一般为最优个体), 避免过早收敛, 同时并行运算可以提高算法的效率。

3.2 基于多种群遗传算法的复杂网络社区结构发现

本文提出了一种基于多种群遗传算法的复杂网络社区结构发现的新算法, 通过加入一些改进性的操作, 解决复杂网络社区发现问题。算法采用 Newman 等人^[6,7] 提出的模块度 Q 作为遗传算法中染色体的适应度。

1) 种群初始化 分别初始化 popsize 个子种群, 设每个子种群中的节点数量为 n , 染色体个数为 m , 一条染色体则代表复杂网络的一种社区划分方案。在每个子种群的初始化时, 给网络中的每一个节点随机分配一个社区 ID 号, 如表 1 所示。社区个

数可任意设定, 一般按网络规模而定, 也可以自定义给出。

表 1 种群初始化

1 th node	2 th node...	(n-1) th node	n th node	chromosome
9	2	15	3	chromosome 1 th
23	12	4	17	chromosome 2 th
...	...	...	...	...
31	25	14	8	chromosome (m-1) th
16	5	20	33	chromosome m th

在表 1 中, 第一条染色体代表的社区划分方案为: 第 1 个节点分配到第 9 个社区, 第 2 个节点分配到第 2 个社区, 直到第 $n-1$ 个节点分配到第 15 个社区, 以及第 n 个节点分配到第 3 个社区。这些过程都是随机分配完成的。

2) 编码 在遗传算法中, 进化过程是建立在编码机制基础上的, 编码方案对算法性能的影响很大, 如搜索能力、种群多样性等。因而在遗传算法中编码方案的设计是要解决的首要问题, 目前的编码方式有二进制编码、实数编码、格雷码编码。常用的为二进制编码, 但在处理网络社区发现的问题上, 由于样本点的数量很大, 若用二进制编码会导致染色体过长而影响算法速度。为了克服二进制编码的缺点, 对于社区划分问题, 可以直接采用节点编码也即实数编码。正如表 1 所示, 每条染色体的基因值由若干个浮点数 (社区 ID 号) 来表示, 这样的编码方式既直接明了, 又增加了算法的搜索能力, 提高了运算效率。

3) 选择 在遗传算法中, 选择操作的本质目的就是实现达尔文进化论中优胜劣汰的思想。在种群中具有较高适应值的个体通过选择操作使得其作为父代为下一代繁衍子孙的概率更大; 相反, 适应值低的个体进入下一代种群的概率就更小。在本文算法中适应值的高低是由网络模块度 Q 的大小来决定的。

常用的选择算法有比例选择法、随机遍历抽样法、锦标赛法和局部选择法等。本文采用的选择方法是比例选择法也叫做轮盘赌选择法, 是最简单却十分有效的一种方法。

4) 交叉 又称做基因重组, 是按一定的概率从父代群体中选择多个个体, 随机将两个个体的基因片段相互交换, 生成两个新的子代个体。这种基因重组是在生物进化过程中, 子代个体继承父代个体优良基因的主要方式, 使得子代个体能够朝着适应环境变化的方向进化。本文采用的是离散重组交叉方式, 该方法将两个父代个体的基因片段以等概率 P_c 相互交换而产生新的子个体。例如, 图 1 中的两个父代个体均分别含有三个基因片段, 通过交换基因片段后生成两个新的子个体。

父个体 1	10	3	18
父个体 2	25	19	2
↓			
父个体 1	25	3	2
父个体 2	10	19	18

图 1 离散重组

5) 变异 在生物学中, 变异是指自然界生物在繁殖过程中, 子代个体染色体中的基因发生突变, 从而可能会生成对环境具有很强适应能力的个体。与此类似, 在遗传算法中用变异来模拟这种基因突变的现象, 期望产生具有更高适应值的新个体, 增强算法的局部随机搜索能力, 使得算法收敛到全局最优解的概率增加。本文采用的变异是基本位变异, 它是以一定的变异概率 P_m 对个体染色体中的某个或几个基因座上的基因位用等位基因来替换的操作。

6)精英迁移策略 在本文的算法中,多种群同时进行独立的遗传演化,只有在一定的时间间隔(本文算法为每隔10代)才进行子种群之间的精英迁移策略。精英迁移策略是各个子种群间进行信息交换的主要方式,即将多个种群在限定的代数后得到的最优个体进行比较,取全局最优个体替换其他种群中的最差个体,使得最优的个体在各子群中传播,从而提高了求解精度和求解速度。

4 实验结果

为了测试本文所提算法的可行性,针对两个经典的真实复杂网络模型 Zachary 和 Dolphins 网络进行了验证。本文算法均采用 MATLAB 编译,实验结果证明,基于多种群遗传算法的复杂网络社区结构发现算法是可行的,并且具有很高的精度,能很好地完成对网络社区的划分。

4.1 Zachary 社会关系网

Zachary 社会关系网^[10]是复杂网络与社会网分析领域中常用的一个小型测试网络。20 世纪 70 年代, Zachary 用两年时间观察美国一所大学空手道俱乐部成员间的社会关系,并构造了如图 2 所示的俱乐部成员社会关系网络(Zachary’s karate club network)。网络包含 34 个节点、78 条边。每个节点表示一个俱乐部成员,节点间的连接表示两个成员经常一起出现在俱乐部活动(如空手道训练、俱乐部聚会等)之外的其他场合,即在俱乐部之外他们可以被称为朋友。调查过程中,该俱乐部因为主管与教练之间的争执而分裂成两个各自以他们为核心的小俱乐部,图中不同形状的节点代表分裂后的小俱乐部成员。该网络作为一个真实的小型社会关系网,常常被用于测试社区发现方法的有效性^[6-8,11,12]。

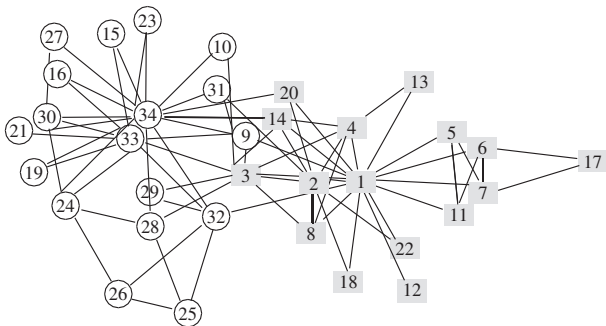


图 2 Zachary 空手道俱乐部成员关系网络
本文算法对 Zachary 网络的实验结果如表 2 所示。

表 2 本文算法对 Zachary 网络的实验结果	
模块度 Q	$Q=0.3718$
社区 1 节点	1,2,3,4,5,6,7,8,10,11,12,13,14,17,18,20,22
社区 2 节点	9,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34
结论	网络中只有第 10 个节点划分错误,实验得出的结果与实际网络基本相符

如图 3 及表 2 所示,用本文算法得到的适应值在 30 代就得到了收敛值 $Q=0.3718$,并且完成了对 Zachary 网络的基本划分,节点划分的准确性高达 97%。在文献[13]中,采用基于 DNA 的遗传算法对 Zachary 网络进行划分,则将节点 3 与节点 10 都划分错误。与之相比,本文算法充分体现出了多种群遗传算法在复杂网络社区发现中的高效性和准确性。

4.2 Dolphins 生物关系网络

Dolphins 关系网^[14,15]也是社会网分析中常用的一个真实

网络。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)进行了长达 7 年的观察,并构造了如图 4 所示的 Dolphins 关系网。图中一个节点代表一只海豚,边表示两只海豚之间接触频繁,共有 62 个节点、159 条边。较大的海豚家族包含 42 个成员节点,而较小的家族仅包含 20 个节点。

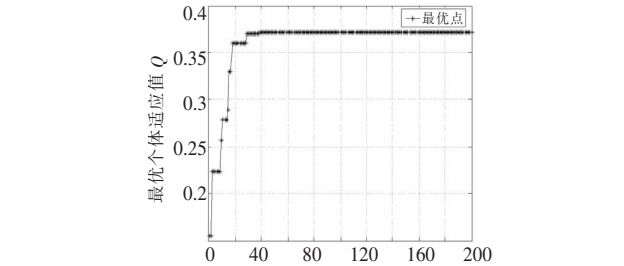


图 3 适应值的变化趋势

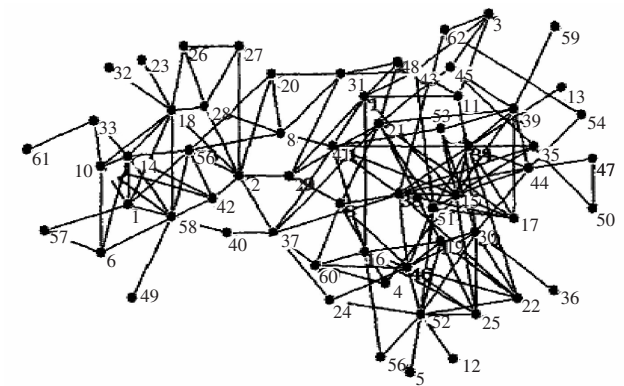


图 4 Dolphins 关系网络

本文算法对 Dolphins 网络的实验结果如表 3 所示。

表 3 本文算法对 Dolphins 网络的实验结果	
模块度 Q	$Q=0.3899$
社区 1 节点	2,6,7,8,10,14,18,20,23,26,27,28,32,33,41,42,49,55,57,58,61
社区 2 节点	1,3,4,5,9,11,12,13,15,16,17,19,21,22,24,25,29,30,31,34,35,36,37,38,39,40,43,44,45,46,47,48,50,51,52,53,54,56,59,60,62
结论	网络中第 40 个节点划分错误,实验得出的结果与实际网络基本相符

如表 3 和图 5 所示,用本文算法得到的适应值的收敛值 $Q=0.3899$,并且完成了对 Dolphins 网络的基本划分,除了第 40 个节点划分错误,其他节点都划分正确,节点划分的准确性高达 98%。这与 Newman 基于模块度矩阵的谱算法^[16]得到的划分结果相同,该算法虽然可以得到较好的社区划分,但该算法的时间复杂度却比较大,为 $O(n(m+n))$,而本文算法的时间复杂度不超过 $O(n^2)$ 。因此本文算法略优于 Newman 算法。

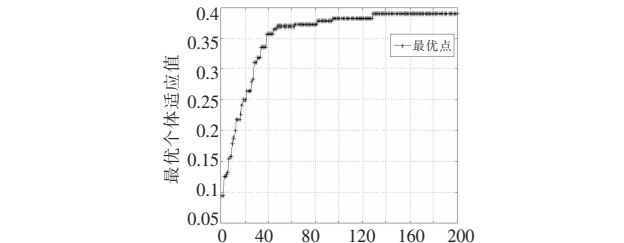


图 5 适应值的变化趋势

5 结束语

本文提出了一种基于多种群遗传算法的复杂网络社区结构发现的新算法,通过实验验证,该算法可以较好地完成对网

络社区的划分。与其他算法相比较,该算法无须预先知道复杂网络的社区数量或社区内的节点数量等先验信息,并且在种群初始化敏感度、准确率以及运算耗时等方面都具有较好的性能,因此本文提出的算法具有很强的现实应用价值。

与此同时,多种群遗传算法还将并行的思想引入到算法之中,将各子种群分别分配给多个处理器并行求解。当采用并行计算时,可以明显提高遗传算法的运行速度。在现有的普遍的串行计算机上无法体现出并行性所带来的高速计算能力,但在未来的并行计算机上,本文算法具有较强的优越性。

参考文献:

- [1] 汪小帆,李翔. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006.
- [2] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. *Bell System Technical Journal*, 1970, 49(2): 291-307.
- [3] FIEDLER M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory[J]. *Czechoslovak Mathematical Journal*, 1973, 23(298): 619-633.
- [4] POTHEN A, SIMON H D, LIOU Kan-pu. Partitioning sparse matrices with eigenvectors of graphs[J]. *SIAM Journal on Matrix Analysis and Applications*, 1990, 11(3): 430-452.
- [5] WU Fang, HUBERMAN B A. Finding communities in linear time: a physics approach[J]. *The European Physical Journal B, Condensed Matter Physics*, 2003, 38(2): 331-338.
- [6] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69(2): 26-113.
- [7] NEWMAN M E J. Fast algorithm for detecting community structure in

networks[J]. *Physical Review E*, 2004, 69(6): 66-133.

- [8] NEWMAN M E J. Detecting community structure in networks[J]. *The European Physical Journal B, Condensed Matter and Complex Systems*, 2004, 38(2): 321-330.
- [9] HOLLAND J H. Adaptation in natural and artificial systems[M]. Ann Arbor: University of Michigan Press, 1975.
- [10] ZACHARY W W. An information flow model for conflict and fission in small groups[J]. *Journal of Anthropological Research*, 1977, 33(4): 452-473.
- [11] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, 99(12): 7821-7826.
- [12] GFELLER D, CHAPPELIER J C, De Los RIOS P. Finding instabilities in the community structure of complex networks[J]. *Physical Review E*, 2005, 72(5): 056135.
- [13] 戴飞飞,唐普英. 基于 DNA 遗传算法的复杂网络社区结构发现[J]. *计算机工程与应用*, 2008, 44(3): 53-56.
- [14] LUSSEAU D, SCHNEIDER K, BOISSEAU O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait? [J]. *Behavioral Ecology and Sociobiology*, 2003, 54(4): 396-405.
- [15] LUSSEAU D, NEWMAN M E J. Identifying the role that animals play in their social networks[J]. *Processings of the Royal Society B: Biological Sciences*, 2004, 271(6): 477-481.
- [16] NEWMAN M E J. Finding community structure in networks using the eigenvectors of matrices[J]. *Physical Review E*, 2006, 74(3): 036104.

(上接第 1236 页)

表 2 CPSO 算法性能指标

问题	n	m	最好解	平均值	最差解	平均时间/s
LA01	10	5	666	666	666	17.8
LA06	15	5	930	930	930	61.1
LA11	20	5	1 222	1 222	1 222	145.6
LA16	10	10	945	945.5	946	148.3
LA21	15	10	1 024	1 024.4	1 078	733.7
LA26	20	10	1 218	1 218	1 218	2 449.5
LA31	30	10	1 789	1 789	1 789	3 883.6
LA36	15	15	1 270	1 270	1 780	3 550.2

表 3 CPSO 算法与其他 PSO 算法的最优性能比较

问题	n	m	CPSO	文献[2]	文献[3]	文献[7]
LA01	10	5	666	666	666	666
LA06	15	5	930	930	930	930
LA11	20	5	1 222	1 222	1 222	1 222
LA16	10	10	945	945	946	946
LA21	15	10	1 024	1 056	1 078	1 078
LA26	20	10	1 218	1 218	1 218	1 218
LA31	30	10	1 789	1 789	1 789	1 789
LA36	15	15	1 270	1 280	1 287	1 271

由表 2 和 3 可见, CPSO 算法对各算例均能够获得最优解, 这表明 CPSO 算法具有很好的搜索质量。同时算法多次独立运行所得平均值和最差解非常接近, 甚至相同, 这表明 CPSO 算法对初始化解具有较好的鲁棒性。

3 结束语

本文对流水车间调度问题中的最小化最大完成时间问题

进行了研究, 并采用文化粒子群算法来求解 JSP, 取得了较理想的结果。通过对典型 JSP 算例(LA 类)的实例测试, 仿真实验表明, 与基本 PSO 算法相比, CPSO 不容易陷入早熟收敛, 具有更高的求解效率; 与混合粒子群算法相比, CPSO 具有更快的收敛速度和更高的求解质量。基于 CPSO 调度算法的时间性能还有待改善, 笔者还将对此作进一步的研究。

参考文献:

- [1] 王凌. 微粒群优化与调度算法[M]. 北京: 清华大学出版社, 2008.
- [2] LIAN Zhi-gang, GU Xing-sheng, JIAO Bin. A novel particle swarm optimization algorithm for permutation Flow-Shop scheduling to minimize makespan[J]. *Chaos, Solitons and Fractals*, 2008, 35(5): 851-861.
- [3] SHA D Y, HSU C Y. A hybrid particle swarm optimization for Job-Shop scheduling problem[J]. *Computers & Industrial Engineering*, 2006, 51(4): 791-808.
- [4] REYNOLDS R G. An introduction to cultural algorithms[C]//Proc of the 3rd Annual Conference on Evolutionary Programming. River Edge, NJ: World Scientific, 1994: 131-139.
- [5] 仲纪, 刘漫丹. 文化算法研究[J]. *计算机技术与发展*, 2008, 18(5): 126-130.
- [6] 吴亚丽, 袁瑛. 一种基于文化粒子群算法的 BP 网络优化方法[J]. *系统仿真学报*, 2011, 23(5): 930-934.
- [7] TASGETIREN M F, LIANG Y C, SEVKLI M, et al. A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem[J]. *European Journal of Operational Research*, 2007, 177(3): 1930-1947.