基于共同进化计算模型的基因连锁问题求解*

钟求喜, 陈火旺

(国防科学技术大学 计算机学院,湖南 长沙 410073)
E-mail: qx_zhong@263.net
http://www.nudt.edu.cn

摘要: 针对传统单种群进化类算法(conventional evolutionary algorithms,简称 CEAs)求解基因连锁问题的不足,基于生物界共同进化机制提出求解 *NK* 基因连锁问题的合作式共同进化算法(Coevolutionary algorithm,简称 CoEA),探讨其子种群的合作方式与个体适应值的计算方法,并从数学上分析该算法的性能,指出共同进化算法中高于平均适应值模式的递增指数高于传统单种群进化算法.仿真结果证实了理论分析.结果表明,共同进化算法比传统单种群进化算法对求解基因连锁问题的效力和效果更好.

关 键 词: 基因连锁;合作式共同进化计算模型;进化计算
中图法分类号: TP18 文献标识码: A

进化计算是一类模拟自然进化过程来求解复杂问题的随机搜索和优化算法.在传统的单种群进化类算法中,如遗传算法(GA)[1,2],种群中一个编码串代表了问题的一个完整解.这种解完整编码方式的主要不足有:第 1,当问题解是由多部分组成时,解完整编码方式对解的好部分的利用可能会被其他较差的解部分所掩盖,从而使算法易于陷入局部最优点而过早收敛;第 2,当编码串存在基因连锁时,其适应值不易计算,且算法的求解效率随问题规模的增大而下降.

针对传统单种群进化类算法的上述不足,本文基于合作式共同进化计算模型,以 *NK* 模型为代表,提出求解基因连锁问题的共同进化方法.本文第 1 节给出 *NK* 模型的问题描述.第 2 节给出合作式共同进化计算模型,详细介绍算法求解基因连锁问题的相关细节.第 3 节是算法性能分析.第 4 节列出了算法的仿真试验结果.最后是结束语.

1 *NK* 模型的问题描述

基因连锁是指染色体中一个基因的优劣与其他基因相关[3].基因连锁问题的典型代表是 *NK* 模型(或称为 *NK* 问题)[4],其中 N 表示染色体或编码串的长度,K 表示染色体内与一个基因相关联的其他基因的个数.本文只考虑一种称为最近邻域基因关联的模型,一个基因只与其最邻近 K 个基因相关联.表 1 给出一个 *NK* 模型的例子(N=4,K=2,字母表 Σ={0,1}),每个基因都与其左右相邻的基因相关联.表 1 的下半部分列出了每个基因对编码串适应值的贡献,贡献值与相邻两个基因的贡献值相关.

取所有基因的贡献总和作为整个编码串的适应值.若编码串 $s=(l_1 l_2 \ldots l_N)$,$l_i \in \Sigma (1 \le i \le N)$,则编码串 s 的适应值可定义为

$$f(s)=\sum_{i=1}^{N} f(l_i) .$$

* 收稿日期: 2000-05-11; 修改日期: 2000-12-19
基金项目: 国家自然科学基金资助项目(69903010,69933030)
作者简介: 钟求喜(1969－),男,湖南永州人,博士,助理研究员,主要研究领域为进化计算,信息安全;陈火旺(1936－),男,福建安溪人,教授,博士生导师,中国工程院院士,主要研究领域为软件理论,人工智能,机器翻译.

其中 $f(l_i)$表示第 i 个基因的贡献值.

NK 模型中适应值空间的峰值点随着 K 的增大而增多.随着适应值空间复杂性的增加和峰值点的增多,寻找一个最优适应值编码串(Nash 平衡点[3])的难度也会增大.最优编码串是指其适应值最大.*NK* 问题的优化目标可定义为

$$\max_{s\in\{0,1\}^N}(f(s)).$$

Table 1 *NK* model for N=4 and K=2

表 **1** N=4 和 K=2 的 *NK* 模型

Sub-String① / Linkages②	Bit 1	Bit 2	Bit 3	Bit 4
000	0.818	0.067	0.477	0.121
001	0.913	0.656	0.021	0.776
010	0.940	0.204	0.379	0.324
011	0.267	0.356	0.128	0.689
100	0.803	0.670	0.327	0.621
101	0.723	0.636	0.821	0.126
110	0.640	0.214	0.509	0.344
111	0.467	0.446	0.166	0.019
Contributions③				

①子串,②关联关系,③贡献.

2 求解 *NK* 模型的共同进化方法

2.1 合作式共同进化计算模型

共同进化计算模型借鉴了生物界中的种群相互作用共同进化机制,是在现有进化计算的基础上形成的一种解空间分离编码的动态搜索和优化方法[3,5].自然界许多物种间共同进化的互利现象是普遍的,如寄生/宿主、捕食/逃逸现象等等.在共同进化计算模型中,每个子种群中的个体只代表问题解的一部分,从所有子种群中各选择一个个体共同组成问题的一个完整解.共同进化有竞争式和合作式两种主要的模型.在合作式共同进化计算模型中,各子种群中的个体通过相互合作为问题求解的总体目标作出各自的贡献.

基于同步执行方式的合作式共同进化计算模型的基本框架如下:

(1) 生成初始种群集合 $\mathrm{Pop}(t)=\{\mathrm{Pop}_i\}$,$t$=0;1≤$i$≤$p$,$p$ 为子种群个数,t 为进化代数;

(2) 根据合作关系和各子种群的状态计算所有子种群 Pop_i 中每个个体的初始适应值;采用精英策略保留最好解;

(3) 如果算法终止条件满足,转步骤(5);否则转步骤(4);

(4) 对各子种群 Pop_i 执行选择、杂交、变异等遗传操作,形成下一代种群 $\mathrm{Pop}(t+1)$;t=t+1,转步骤(2);

(5) 输出结果,算法终止.

相对于传统单种群进化算法而言,共同进化计算模型有以下几个主要优点:第 1,共同进化计算模型中由于问题的各组成部分采用不同的子种群进行进化求解,因而能有效地克服传统单种群进化算法中解完整编码方式解的好部分被差部分所掩盖的不足.第 2,共同进化是一种通用宏进化模型,各子种群的进化算法可以互不相同,更能根据特殊需要和问题特点进行问题求解.第 3,共同进化计算模型不易过早收敛和陷入局部最优点.

2.2 子种群划分

将共同进化计算模型应用于求解 *NK* 模型,子种群中可行解的编码方式与传统单种群进化算法相同,一个可行解就是一个定长的二进制串.确定共同进化算法中的子种群数,本质上是对 *NK* 模型进行适当的问题分解.

对 *NK* 模型而言,自然的问题分解方法是将长度为 *N* 的二进制串等长地划分成 *n* 段,这样就将 *NK* 问题转化成 *n* 个子问题.每个子问题用一个子种群进行进化方法求解.当 *K*=0 时,不存在基因连锁;当 *K*>0 时,则表示存在基因连锁.为了避免适应值计算的复杂性,希望子种群间的基因连锁程度越轻越好.于是,令子种群个数 *n* 为

$$n=\frac{N}{l}.$$

其中 *l* 为子种群中编码串的长度($N\geq l\geq(K+1)$).

2.3 合作方式及适应值计算

子种群中编码串的适应值是维持子种群进化的压力.合作式共同进化算法要求子种群中个体适应值的计算必须考虑其他子种群的进化状态.子种群中一个编码串的适应值与它对整个问题求解所作出的贡献有关.对子编码串贡献的评价方式确定了子编码串适应值的计算方式.而子编码串对整个问题求解的贡献与共同进化算法中的合作方式相关.

根据社会经济活动中广泛存在的"强强合作"现象,在合作式共同进化算法中也采用强强合作的简单合作方式.在评价子种群中子编码串的贡献时,将该子编码串与其他子种群中的最好子编码串组合成一个完整解,从组合的完整解的优劣来计算该子编码串的适应值.子种群 Pop_i 中子编码串 $s_i=(l_1^i,l_2^i,\ldots,l_l^i)$的适应值计算过程如下:

(1) 从其他所有的子种群 $\mathrm{Pop}_j(j\neq i,1\leq j\leq n)$中选取一个代表 $x_j=(l_1^j,l_2^j,\ldots,l_l^j)$(下标中的 *l* 为子编码串的长度且有 $nl=N$);

(2) 将 *n* 个子编码串 $x_1,x_2,\ldots,x_{i-1},s_i,x_{i+1},\ldots,x_n$,组合成一个完整解的编码串:

$$s=(l_1^1,l_2^1,\ldots,l_l^1,l_1^2,l_2^2,\ldots,l_l^2,\ldots,l_1^n,l_2^n,\ldots,l_l^n);$$

(3) 子编码串 s_i 的适应值为 $f(s_i)=f(s)$.

3 算法性能分析

Holland 的模式定理[1]奠定了遗传算法的理论基础.基于字母表 $\Sigma=\{0,1\}$上的定长二进制串编码技术,模式中的通配符*可匹配 0 或 1.模式 *S* 的阶定义为模式中固定位置的数目,记为 $o(S)$.模式 *S* 的既定长度定义为模式中第 1 个和最后一个固定位置的海明距离,记为$\delta(S)$.如模式 *S*=(1**1)可匹配 4 个串(1011,1001,1101,1111).在求解 *NK* 模型的共同进化算法中,设子种群 Pop_k 中的子模式 S_k 在第 *t* 代时的平均适应值为 $\bar{f}(S_k,t)$,

$$\bar{f}(S_k,t)=\sum_{j=1}^{q}f(s_{kj})/q,$$

其中 $s_{k1},s_{k2},\ldots,s_{kq}$ 是 *q* 个与子模式 S_k 匹配的子编码串.

子种群 $\mathrm{Pop}_k(t)$的平均适应值为 $\bar{f}_k(t)=\sum_{j=1}^{p_k}f(s_{kj})/p_k$,$p_k$ 为子种群 Pop_k 的种群规模.设$\xi(S_k,t)$为子种群 $\mathrm{Pop}_k(t)$中与子模式 S_k 匹配的子编码串个数.在 $t+1$ 代时,期望有$\xi(S_k,t+1)$个子编码串与子模式 S_k 匹配.根据按比例选择有:

$$\xi(S_k,t+1)=\frac{\bar{f}(S_k,t)}{\bar{f}_k(t)}\cdot\xi(S_k,t).$$

在求解 *NK* 问题的传统单种群遗传算法中,假设分解后的 *n* 个子问题互不关联,则按比例选择对由 *n* 个子模式组成的模式 $S=(S_1,\ldots,S_n)$的作用可表示为

$$\xi(S,t+1)=\sum_{i=1}^{n}\left(\zeta(S,t)\cdot\frac{\bar{f}(S_i,t)}{\sum_{j=1}^{n}\bar{f}_j(t)}\right)=\xi(S,t).\frac{\sum_{i=1}^{n}\bar{f}(S_i,t)}{\sum_{i=1}^{n}\bar{f}_i(t)}. \quad (1)$$

在求解 *n* 个独立子问题的合作式共同进化算法中,按比例选择对模式 $S=(S_1,\ldots,S_n)$的作用为

$$\xi(S,t+1)=\prod_{i=1}^{n}\left(\frac{\bar{f}(S_i,t)}{\bar{f}_i(t)}\cdot\xi(S_i,t)\right)=\xi(S,t).\prod_{i=1}^{n}\frac{\bar{f}(S_i,t)}{\bar{f}_i(t)}. \tag{2}$$

定理. 在基于独立子问题求解的共同进化遗传算法中,高于平均适应值的模式在算法后续代中试验次数的递增指数高于传统遗传算法的递增指数.

证明:即证明式(1)和式(2)中的递增指数满足:

$$\prod_{i=1}^{n}\frac{\bar{f}(S_i,t)}{\bar{f}_i(t)}\geq\frac{\sum_{i=1}^{n}\bar{f}(S_i,t)}{\sum_{i=1}^{n}\bar{f}_i(t)}. \tag{3}$$

为简单起见,证明过程以 n=2 为例.即证明

$$\frac{\bar{f}(S_1,t)\cdot\bar{f}(S_2,t)}{\bar{f}_1(t)\cdot\bar{f}_2(t)}\geq\frac{\bar{f}(S_1,t)+\bar{f}(S_2,t)}{\bar{f}_1(t)+\bar{f}_2(t)}. \tag{4}$$

但此证明过程可推广至一般情况.

采用分析法.由于 $\bar{f}_1(t)>0$, $\bar{f}_1(t)>0$,由不等式(4)有

$$[\bar{f}(S_1,t)\cdot\bar{f}(S_2,t)]\cdot(\bar{f}_1(t)+\bar{f}_2(t))\geq[\bar{f}(S_1,t)+\bar{f}(S_2,t)]\cdot(\bar{f}_1(t)\cdot\bar{f}_2(t)).$$

于是

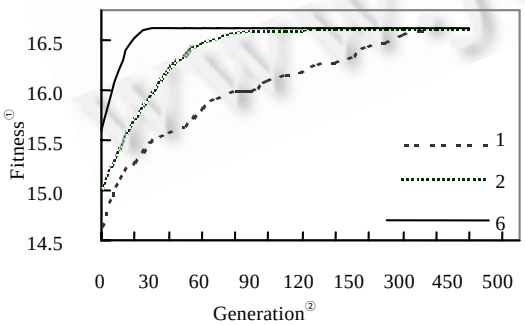
$$\bar{f}(S_1,t)\cdot\bar{f}_1(t)\cdot[\bar{f}(S_2,t)-\bar{f}_2(t)]+\bar{f}(S_2,t)\cdot\bar{f}_2(t)\cdot[\bar{f}(S_1,t)-\bar{f}_1(t)]\geq 0. \tag{5}$$

当条件 $\bar{f}(S_1,t)\geq\bar{f}_1(t)$与 $\bar{f}(S_2,t)\geq\bar{f}_2(t)$成立时,不等式(5)一定成立.由此反向推理,不等式(4)必然成立.同样,一般情况下不等式(3)也成立. □

此定理说明,在求解独立子问题的共同进化算法中,高于平均适应值的解的试验次数的递增指数也越高,因此共同进化算法的问题求解效率高于传统的进化算法.

4 仿真分析

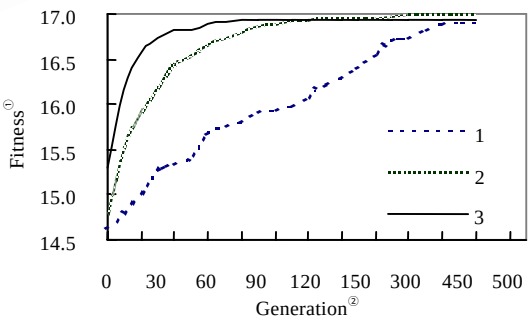
实验仿真是目前进化计算中检验算法性能的最主要的手段之一.为了检验文中算法求解 *NK* 问题的效力和效果,分别对不同规模的问题作了对比仿真实验.算法中使用单点杂交算子和位变异算子、轮盘赌规则的按比例选择和精英保留策略.实验中的重要参数有,各子种群的种群规模皆为 50/p(p 为子种群数),单点杂交概率为 0.9,位变异概率为 0.05.算法以最大的迭代代数作为终止条件(CoEA 中的 1 代是指其子种群依次各进化 1 代),最大迭代代数为 500.每个 *NK* 问题的基因适应值贡献都是 0.0~1.0 之间的随机数.采用 Potter[3]的分类法,令 N=24,分别对轻度基因连锁(K=0,1)和中度基因连锁(K=6)的 3 种情况进行仿真对比试验.图 1~图 3 分别给出了在这 3 种情况下算法在不同进化代所找到的最好解的静态性能曲线(20 次运行的平均值).图中图例标的数字代表子种群数 p(p=1 时为 CEA 算法,p>1 时为 CoEA 算法).



①适应值,②进化代数.

Fig.1 Fitness vs generation (K=0)

图 1 收敛曲线(K=0)



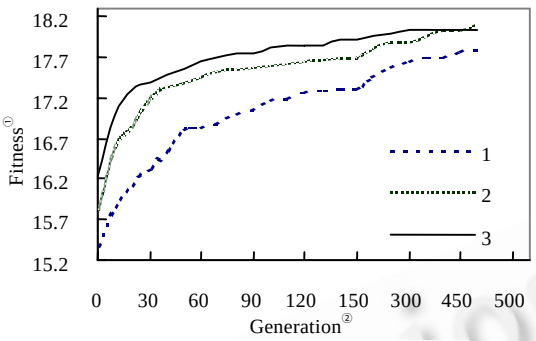
①适应值,②进化代数.

Fig.2 Fitness vs generation (K=1)

图 2 收敛曲线(K=1)

K=0 时表示各基因间不存在基因连锁,这相当于对各组成部分互不关联问题的求解.图 1 的静态性能曲线

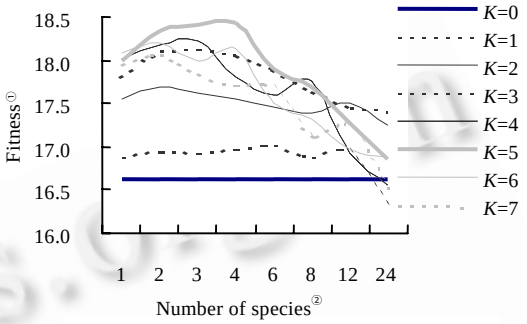
表明,随着子种群数的增多,算法的收敛性也越好.这个结果与前面的算法性能分析是一致的.图 2 和图 3 的收敛曲线也表明 CoEA 的性能优于 CEA.但同时也可以看出,算法终止时的收敛点并不总随着子种群的增多及算法收敛速度的加快而变好.这说明当子种群间相互关联时,子种群个数对算法的收敛点有重大影响.针对不同程度的基因连锁度(N=24,K=0,…,7),图 4 列出了算法在 500 代终止时的最好解与子种群个数的对比曲线(20 次运行的平均值),说明基因关联情况下子种群数宜设为 2~4 个.



①适应值,②进化代数.

Fig.3 Fitness vs generation (K=6)

图 3 收敛曲线(K=6)



①适应值,②种群数.

Fig.4 Number of species vs fitness

图 4 种群数收敛曲线

5 结束语

基于合作式共同进化计算模型,本文提出一种求解基因连锁问题的进化算法,并分析了与之相关的问题,如种群间的合作方式和子种群中个体适应值的计算等,从数学上分析了基于合作式共同进化算法的性能,指出合作式共同进化算法中好的解个体能以高于传统单种群进化算法的递增指数递增.仿真分析证实了算法的理论分析结果,表明共同进化算法比传统单种群进化算法对求解基因连锁问题的效力和效果更好.

基因连锁问题在数值优化和并行与分布式系统中任务分配与调度[6]等应用领域中是广泛存在的.由于基因连锁问题的复杂性,进一步研究共同进化算法中各子种群的相互作用方式和子种群的自适应划分,对该问题的求解和应用具有十分重要的意义.

致谢 博士后流动站的谢涛副教授和计算机学院博士生荔建琦同学对本文的完成提出了很多有益的建议,在此表示感谢.

References:

[1] Holland, J.H. Adaptation in Natural and Artificial Systems. Ann Arbor: The University of Michigan Press, 1975.

[2] Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning. Reading: Addison-Wesley Publishing Company, Inc., 1989.

[3] Potter, M.A. The design and analysis of a computational model of cooperative coevolution [Ph.D. Thesis]. George Mason University, 1997.

[4] Kauffman, S.A. Adaptation on rugged fitness landscapes. In: Stedin, D.L., ed. Lectures in the Science of Complexity. Reading: Addison-Wesley, 1989. 527~618.

[5] Li, Jian-qi. On coevolutionary computation methodology. Technical Report, TR-9901-0107, Changsha: National University of Defence Technology, 1999 (in Chinese).

[6] Zhong, Qiu-xi. Task matching and scheduling in network computing environments based on genetic algorithms [Ph.D. Thesis]. Changsha: National University of Defence Technology, 2000 (in Chinese).

附中文参考文献:

[5] 荔建琦.COGA 研究.技术报告,TR-9901-0107,长沙:国防科学技术大学,1999.

[6] 钟求喜.网络计算中任务分配与调度的遗传算法研究[博士学位论文].长沙:国防科学技术大学,2000.

Solving Epistatic Interactions Based on Computational Model of Coevolution*

ZHONG Qiu-xi, CHEN Huo-wang

(*School of Computer, National University of Defence Technology, Chansha* 410073, *China*)

E-mail: qx_zhong@263.net

http://www.nudt.edu.cn

Abstract: It is difficult for conventional single population-based evolutionary algorithms (conventional evolutionary algorithms——CEAs) to solve epistatic interaction problems. Based on computational model of cooperative coevolution inspired by the coevolutionary phenomena of natural species, a coevolutionary algorithm (CoEA) for solving *NK*-landscape problem is proposed. Some problems related to the interactions among species and individual's fitness computation are discussed. Mathematical analysis shows that the exponential increase index of CoEA is higher than that of CEA for a schema which fitness is higher than the average fitness of population. Simulation results verify the theoretical result, and show that the coevolutionary algorithm is more efficient and effective than CEA in solving epistatic interactions problems.

Key words: epistatic interaction; computational model of cooperative coevolution; coevolution computation

* Received May 11, 2000; accepted December 19, 2000

Supported by the National Natural Science Foundation of China under Grant Nos.69903010, 69933030

第 12 届中国计算机学会网络与数据通信学术会议

征 文 通 知

为推动我国在此方向的研究,探讨计算机网络与数据通信技术的发展动态与趋势,促进我国科研人员在此领域的交流与合作,中国计算机学会网络与数据通信专业委员会拟于 2002 年 12 月 2~4 日在武汉举办"第 12 届中国计算机学会网络与数据通信学术会议".会议由华中师范大学计算机科学系承办,并将邀请该领域的国际知名学者作专题特邀报告.为保证本次会议的学术质量,现向全国科技工作者公开征稿.征稿范围包括计算机通讯网络理论与工程的各个方面.本次会议的论文将结辑出版优秀论文将由计算机学会推荐给有关核心期刊发表.

一、征文要求

(1) 论文应是未公开发表过,一般不超过 6000 字;

(2) 全文电子邮件投稿,要求 Word2000 兼容的电子文档,所有内容放于一个文件中;

(3) 编排格式:

标题: 居中,2 号黑体;作者: 居中,4 号仿宋;作者地址: 5 号楷体;摘要、关键词: 5 号楷体;正文:5 号宋体,分节标题 4 号;参考文献: 小 5 号宋体.

(4) 投稿地址:华中师范大计算机科学系谭连生教授收 E-mail: L.Tan@ccnu.edu.cn

二、重要日期

论文提交截止日期: 2002 年 8 月 15 日 论文接收通知日期: 2002 年 10 月 1 日 会议注册日期: 2002 年 12 月 2 日

联系人:谭连生教授 湖北省武汉市华中师范大学计算机科学系(430079)

电话: 027–87673277 传真: 027–87876070 E-mail: L.Tan@ccnu.edu.cn