水质模型参数优化的遗传算法实现及控制参数分析

王建平, 程声通, 贾海峰
(清华大学环境科学与工程系, 北京 100084)

摘要: 参数识别是数学模型应用的前提. 遗传算法是一种通用的全局优化算法, 结构简单. 一个实际应用问题能否利用遗传算法解决, 关键在于遗传算法的设计和控制参数的选取. 本文结合水质模型参数优化的特点, 提出采用正交试验设计的方法来考察遗传算法不同控制参数对参数优化性能的影响, 结果显示, 正交法较好地识别了关键影响因素, 并提出可能的最优方案. 表明遗传算法能较好地应用于复杂多参数水质模型的参数识别研究.

关键词: 参数识别; 水质模型; 遗传算法; 全局优化; 正交实验
中图分类号: X11, X143 **文献标识码**: A **文章编号**: 0250-3301(2005)03-0061-05

Parameter Optimization of Water Quality Model: Implementation of Genetic Algorithm and Its Control Parameters Analysis

WANG Jian-ping, CHENG Sheng-tong, JIA Hai-feng
(Department of Environmental Science and Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Parameter identification plays an important role in environmental model application. As a commonly used global optimization method, genetic algorithm (GA) has very simple structure, the key related to whether a practical issue can be solved using GA or not is algorithm design and selection of the control parameters. Based on the feature of parameter optimization of water quality model, orthogonal test method was proposed for reviewing effects of different control parameters of GA on the performance of water quality parameter optimization. The results indicate that orthogonal method could identify key factors, and also provide possible optimized experiment plan. It is concluded that GA can be applied to the research on parameter identification of complicated water quality model.

Key words: parameter identification; water quality model; genetic algorithm; global optimization; orthogonal experiments

对于选定区域的问题, 模型结构确定之后, 最重要的是如何有效地识别模型参数. 水质模型参数优化属于复杂的非线性优化问题, 参数响应曲面存在很多凹谷和平坦区域, 有大量局部极小点. 参数越多, 参数响应曲面非线性度越高[1]. 传统的优化方法, 如梯度法和单纯形法对模型结构、优化准则要求严格, 受初始条件影响较大, 且由于其非线性特征通常只能得到局部最优解. 基于随机采样的统计方法, 如 HSY 算法、GLUE 算法[2]等由于其随机抽样机制, 当参数个数增多时, 参数识别过程将非常耗时[3]. 因此, 水质模型参数识别仍然是一个值得研究的问题. J. Holland 于 1975 年受生物进化论的启发而提出的遗传算法(genetic algorithms, GA), 是一种高度并行、随机和自适应的通用优化算法, 其编码技术和遗传操作比较简单, 优化不受限制性条件的约束, 2 个最显著特点是隐含并行性和全局解空间搜索. 本文将围绕遗传算法的设计和控制参数优选来探讨其在水质模型参数识别中的应用.

遗传算法在环境模型参数识别应用中以水文模型[4]、地下水模型[5]等居多, 而且研究起步较早. 相比而言, 遗传算法用于水质模型参数率定的国内外研究文献还不多, 并以河流模型系统为主, 识别的水质模型参数较少[6]. 国内研究一般基于简单水质模型如 Dobbins BOD-DO 模型或 Streeter-Phelps 模型, 这些模型均存在解析解, 遗传算法应用效果较好[7,8].

理论上已经证明, 带有保优操作的遗传算法可以概率 1 收敛于问题最优解[9]. 作为一个通用的全局优化算法, GA 结构非常简单. 一个实际应用问题能否利用遗传算法解决, 关键在于算法的设计[9,10]. 具体包括: 确定问题的编码方案; 确定适配值函数; 设计遗传算子, 包括初始化、选择、交叉、变异和替换操作等; 选取算法参数, 包括种群数目、交叉概率、变异概率、进化代数等; 确定算法的终止条件. Grefenstette[11]将 GA 的参数选取作为一个优化问题, 提出用 GA 优化 GA 参数的二级数值方法. 尽管此方法适用范围较广, 但工作量较大, 且二级算法

收稿日期: 2004-06-28; **修订日期**: 2004-10-21
基金项目: 国家自然科学基金资助项目(50209007)
作者简介: 王建平(1977～), 男, 博士研究生, 主要从事环境系统分析的理论与应用研究. E-mail: wangjp@tsinghua.org.cn

本身的参数也有待优化,因此很少得到实际应用.同时水质模型参数识别每次参数调整均需进行水质模拟,如此操作带来的计算量将会十分巨大.因此能否通过较少的试验来考察算法控制参数的影响就显得至关重要,基于此本研究引入了正交试验设计方法解决这一问题.

正交试验设计[12]是应用数理统计观点和正交原理的科学试验设计方法.它借助合适的正交表,利用正交表的均衡分散性和整齐可比性,成功解决了理论上需要进行的试验次数与实际可行的试验次数的矛盾和实际所做的有限量试验与要求全面掌握事物内在规律之间的矛盾.

1 模型与数据

1.1 模型简介

研究采用的水质模型为 WASP 模型系统.WASP[13]是由美国国家环保局开发的用于地表水水质模拟的模型,它提供了一个灵活的动态模拟系统.如图 1 所示,WASP 可以模拟 8 个指标,分别为:氨氮(NH_3^-)、硝酸盐氮(NO_3^-)、溶解性磷酸盐(OPO_4^{2-})、叶绿素 a(Chl-a)、碳生化需氧量(CBOD)、溶解氧(DO)、有机氮(ON)和有机磷(OP).在 WASP 模型系统中水质模块 EUTRO5 的水质参数有 42 个之多,通过灵敏度分析提取了灵敏度较高的参数,结果如表 1 所示.

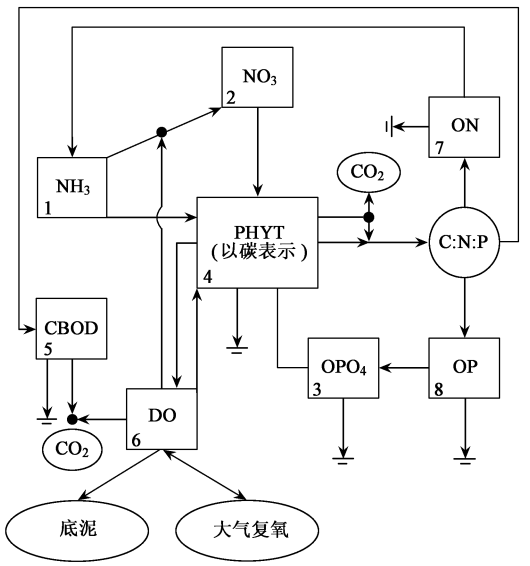

图 1 水质模拟反应动力学关系

Fig. 1 Wasp eutrophication (EUTRO) state variables relationships

表 1 待识别水质模型参数

Table 1 Parameters needed to identify in water quality model

参数名称	物理意义	参数取值	参数范围
K12C	20 ℃条件下的硝化速度系数/ d^{-1}	0.2	0.05～0.35
K20C	20 ℃条件下的反硝化速度系数/ d^{-1}	0.05	0.01～0.2
K1C	浮游植物的饱和生长率/ d^{-1}	2.5	1.5～4
K1RC	20 ℃条件下浮游植物的内源呼吸速率/ d^{-1}	0.1	0.05～0.2
NCRB	浮游植物内的氮碳比(mg/ mg),缺省值为 0.25	0.25	0.2～0.3
PCRB	浮游植物内的磷碳比(mg/ mg),缺省值为 0.025	0.025	0.02～0.03
KDC	20 ℃条件下的 CBOD 降解速率/ d^{-1}	0.03	0.01～0.1
K2	20 ℃条件下,水体的复氧速度常数/ d^{-1}	0.15	0.1～0.2
K71C	溶解有机氮的矿化速度/ d^{-1}	0.03	0.01～0.1
K83C	溶解有机磷的矿化速度/ d^{-1}	0.03	0.01～0.1

1.2 数据序列产生

本文采用合成的“观测”数据进行研究,即在已知参数值的情况下(参数值见表 1[14～17])应用 WASP 模型,产生“真实值”时间序列,得到“观测值”时间序列,然后利用这些数据进行模型参数识别.这样做的目的是为了使系统的真实情况在掌握之中,同时模型没有结构上的误差,参数估计的所有误差仅来源于参数初值和参数识别本身.这使得有可能在排除结构误差的情况下,单独研究遗传算法控制参数对于参数识别的影响.合成数据方法在许多模型分析中有广泛的使用.

2 遗传算法设计

2.1 编码

模型参数估计是多维函数优化问题.二进制编码通常会导致很大的计算量和存储量,且串长影响算法精度,在此采用双精度实数编码,将待估计参数直接用实数向量形式表示来进行优化搜索.

2.2 适配值函数

适配值函数用于对个体进行评价,也是优化过

程评价个体的依据. 本研究考察了 2 种适配值确定方法.

(1) 令系统真实输出为 $y_0(t)$,在搜索参数下的模型输出为 $y(t)$,参数估计目标就是使两者差距最小,如公式(1). 式(2)为适应值函数计算公式,同时为防止分母为 0 或者溢出,分母加上 0.01.

$$\mathrm{obj} = \sum_t [y(t) - y_0(t)]^T [y(t) - y_0(t)] \quad (1)$$

$$f = 1/\left(\sqrt{\sum_t (y(t) - y_0(t))^T (y(t) - y_0(t))} + 0.01\right) \quad (2)$$

(2) 排序思想　按目标值 obj 排序,指定三角概率分布,避免利用公式(2)计算出现某一个体适配值太高,在选择操作中重复选择,导致算法"早熟"收敛. m 为种群个体数.

$$f_i = \frac{2(m + 1 - i)}{m(m + 1)}, \quad i = 1, \cdots, m \quad (3)$$

2.3　算法参数

(1) 种群数目　种群数目是影响算法最终优化性能和效率的因素之一. 研究考察了 4 种种群个数,分别为 20、50、100 和 200 个.

(2) 交叉概率　交叉概率用于控制交叉操作的频率. 研究考察了 4 种交叉概率,分别为 0.5、0.7、0.9 和 1.0.

(3) 变异概率　变异概率是加大种群多样性的重要因素. 研究考察了 4 种变异概率,分别为 0.01、0.05、0.1 和 0.2.

(4) 代沟(替换策略)　代沟用于控制每代中父代种群被替换的比例,研究考察了 2 种代沟设置,90 %和 70 %.

2.4　遗传算子

复制操作:用于避免有效基因的损失,使高性能的个体得以更大的概率生存,从而提高全局收敛性和计算效率,一般采用轮盘赌选择方法. 交叉操作:实数编码 GA 通常采用双个体算术交叉. 本研究中复制操作和交叉操作不作变化.

变异操作:实数编码中通常采用扰动式变异,即对原先个体附加一定机制的扰动来实现变异,即 $x' = x + \eta \cdot \xi$,其中 x 和 x 分别为新旧个体,η 为扰动幅度参数,ξ 为随机扰动变量. 本研究考察了 4 种扰动分布:高斯分布、柯西分布、混沌序列和均匀分布.

2.5　算法终止条件

实际应用 GA 时不可能进行无休止地搜索,同时问题的最优解通常未知,因此必须设计一些准则来终止算法. 研究中采用给定一个最大进化步数和给定最佳搜索解的最大滞留步数的双重终止方式. 本研究中,为了使不同参数组合的试验具有对比性,设定每种组合的水质模拟次数为 10 000 次,故最大进化代数因种群个数的不同而变.

综上所述,研究将考察遗传算法中的 6 个重要因素(参数)对参数优化效果的影响,分别为适配值函数、种群大小、交叉概率、变异概率、变异分布函数和替换策略.

3　正交试验方案及结果

正交试验方案设计过程为:①明确试验目的,确定试验考核指标;②确定考察因素和水平,制定因素水平表;③选用正交表;④编写试验计划表.

本研究试验目的是考察遗传算法控制参数的影响,获得较好水质模型参数优化结果. 目标函数如式(1)所示. 考察因素包括种群大小、交叉概率、变异概率和变异分布函数 4 个四水平因素和适配值函数和替换策略 2 个二水平因素. 选用正交表 $L_{16}(4^4 \times 2^3)$,试验方案如表 2 所示,因素的水平设置是根据文献中常用的范围确定的[9,11]. 共进行 16 次试验,表 2 中水平编号后的数字为编号代表的参数实际值,只列出 1 组,其它从略.

研究中为避免随机数产生器种子(seed)的影响,每一组试验均随机模拟 5 次,参数搜索范围见表 1 最后一列. 模型运行结果见表 2,表中最后一列为 5 次运行所得最优目标函数值的平均值. 表中"1(2、3、4)水平得分之和"为对应水平的"目标函数平均值"之和;"1(2、3、4)水平平均得分"为对应"1(2、3、4)水平得分之和"除以(试验次数/水平数);"平均得分极差"为"1(2、3、4)水平平均得分"在对应因素上的极差.

4　结果分析与讨论

正交试验结果分析如下:

(1) 考察极差　极差的大小说明相应因素作用的大小. 极差大,说明该因素是灵敏的,它的变化对结果影响很大. 极差小,说明该因素是不灵敏的,它的变化对结果影响较小. 试验结果表明,变异分布函数影响最大,适配值函数的影响最小,排序为:变异分布函数 > 变异概率 > 交叉概率 > 种群大小 > 替换策略 > 适配值函数. 需要指出的是,这些因素的影响具有一定的相对性,本研究是在相关研究的基础上

进行的,适配值函数定义是合理的,它们的极差不大,说明这些适配值函数定义适用于本研究.

表 2 正交试验方案及结果

Table 2 Orthogonal experiment scheme and results

编号	因素						目标函数平均值
	种群大小	交叉概率	变异概率	变异分布函数	适配值函数	替换策略	
1	1 20	2 0.7	3 0.1	2 高斯	2 排序	1 90 %	1.72
2	3 100	4 1.0	1 0.01	2	1 计算	2 70 %	9.05
3	2 50	4	3	3 混沌	2	2	12.83
4	4 200	2	1	3	1	1	8.88
5	1	3 0.9	1	4 均匀	2	2	22.20
6	3	1 0.5	3	4	1	1	1.79
7	2	1	1	1 柯西	2	1	13.42
8	4	3	3	1	1	2	4.22
9	1	1	4 0.2	3	1	2	28.60
10	3	3	2 0.05	3	2	1	5.79
11	2	3	4	2	1	1	1.29
12	4	1	2	2	2	2	7.67
13	1	4	2	1	1	1	4.84
14	3	2	4	1	2	2	3.32
15	2	2	2	4	1	2	5.11
16	4	4	4	4	2	1	6.21
1 水平得分之和	57.36	51.48	53.55	25.80	63.79	43.95	
2 水平得分之和	32.66	19.04	23.41	19.74	73.16	93.00	
3 水平得分之和	19.95	33.50	20.56	56.10			
4 水平得分之和	26.98	32.93	39.42	35.31			
1 水平平均得分	14.34	12.87	13.39	6.45	7.97	5.49	
2 水平平均得分	8.16	4.76	5.85	4.93	9.14	11.63	
3 水平平均得分	4.99	8.38	5.14	14.02			
4 水平平均得分	6.75	8.23	9.86	8.83			
平均得分极差	7.59	8.11	8.25	9.09	1.17	6.13	

(2) 考察不同水平的平均得分 对比不同水平的得分差别,可确定相对较优的因素组合. 平均得分越低,该水平越优,逐因素比较可得较优的参数组合 ①为:种群大小(100)、交叉概率(0.7)、变异概率(0.1)、变异分布函数(高斯分布)、适配值函数(计算)、替换策略(90 %). 16 次试验中的最优组合 ②为:种群大小(50)、交叉概率(0.9)、变异概率(0.2)、变异分布函数(高斯分布)、适配值函数(计算)、替换策略(90 %). 为此,代入较优参数组合 ①再运行一次,平均目标函数值为 1.01,优于组合 ②的目标函数平均值(1.29). 可见正交试验法用于考察遗传算法不同因素对参数识别性能影响是有效的、适用的,达到了通过有限量试验来掌握算法控制参数影响的目的.

综上 2 点,可以得出如下结论:

(1) 变异概率、交叉概率和种群数目是影响算法最终优化性能和效率的重要因素. 试验结果表明,在水质模型参数识别中保持一个相对较高的变异概率,能显著提高算法的性能. 交叉概率不能太大也不能太小,以保持种群的有效进化. 种群数目太小不能提供足够的采样点,以致算法性能很差,种群增大可增加优化信息以阻止“早熟”收敛的发生,但种群太大,无疑会增加计算量,因而种群一般不能太大.

(2) 研究考察了 4 种变异分布函数对算法优化性能的差异,从不同水平平均得分来看,高斯分布和柯西分布相对较好,混沌序列和均匀分布较差. 极差结果表明,正确地选择变异分布函数非常重要,一般常用的分布函数为高斯分布或柯西分布.

(3) 适配值函数的确定. 研究表明,本文采用的 2 种适配值计算方法均可用于参数优化研究,基于公式(2)的计算方法相对较好. 但需要指出的是,对于局部极小较多的情况,进化中若较早出现某一个体适配值很高,基于公式(2)的计算方法会导致算法“早熟”收敛. 基于排序的适配值计算方法可有效避免这一情况.

(4) 替换策略. 研究表明,高替换比例要好于低替换比例,分析主要原因是低父代替换比例易导致“早熟”收敛.

(5) 16 组试验的目标函数平均值变化范围是 1.29～28.6,控制参数选择不好,会出现明显的"早熟"收敛.图 2 对比了编号为 5 的参数组合和较优参数组合①的一次搜索进程,出现"早熟"收敛的搜索结果明显很差,可见,GA 算法控制参数的优选对算法优化性能的影响显著,在应用中要慎重选择,正交试验法是获取较优参数组合的一个有效途径.研究考察了 4 水平因素 4 个,2 水平因素 2 个,若完全组合试验,需要进行 1 024 次,而正交试验只需进行 16 次就可获得较优的参数组合,计算量大大降低.表 3 为遗传算法参数组合①的水质模型参数优化结果,可见,遗传算法能较好地应用于复杂多参数水质模型的参数识别研究.

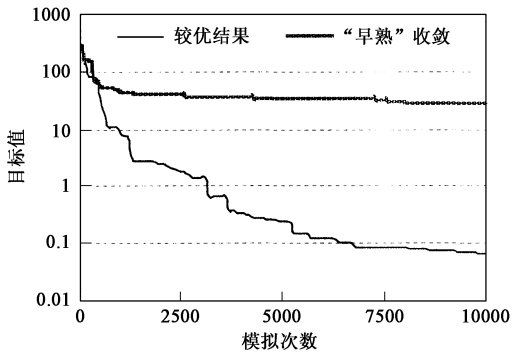

图 2　遗传算法进化过程对比

Fig. 2　Comparisons of GA search processes

表 3　参数优化结果

Table 3　Parameter optimization results

参数名称	参数"真"值	参数优化结果	相对误差/ %
K12C	0.2	0.20531	2.7
K20C	0.05	0.05055	1.1
K1C	2.5	2.5163	0.7
K1RC	0.1	0.10063	0.6
NCRB	0.25	0.24968	0.1
PCRB	0.025	0.02455	1.8
KDC	0.03	0.03044	1.5
K2	0.15	0.15132	0.9
K71C	0.03	0.03037	1.2
K83C	0.03	0.02976	0.8

5　结论

(1) 采用正交试验设计的方法来考察遗传算法不同控制参数对参数优化性能的影响,大大降低了试验计算量.试验结果表明,正交法较好地识别了遗传算法用于水质模型参数优化的关键影响因素,并提出较优的控制参数组合方案.

(2) 本研究中,通过合理的遗传算法设计和控制参数选择,成功实现了 WASP 模型系统的 10 个水质参数的识别.可见,遗传算法能较好地应用于复杂多参数水质模型的参数识别研究.

参考文献:

[1] Beven K, Binley A. The future of distributed models: model calibration and uncertainty prediction [J]. Hydrological processes, 1992, **6**: 279～298.

[2] 邓义祥. 稀疏数据条件下河流水质模型的参数识别[D]. 北京: 清华大学, 2003.

[3] Jorgensen S E. An improved parameter estimation procedure in lake modeling [J]. Lake & Reservoirs: Research and Management, 1998, **3**:139～142.

[4] Wang Q J. The genetic algorithm and its application to calibrating conceptual rainfall-runoff models [J]. Water Resources Research, 1991, **27**(9):2467～2471.

[5] McKinney D C, Min D L. Genetic algorithm solution of groundwater management models [J]. Water Resources Research, 1994, **30**(6): 1897～1906.

[6] Mulligan A E, Brown L C. Genetic algorithms for calibrating water quality models[J]. Journal of Environmental Engineering, 1998, **124**(3): 202～211.

[7] 王巍,曾光明,秦肖生. GSA 法在水质模型参数估值中的应用[J]. 上海环境科学, 2003, **22**(9): 619～623.

[8] 曾光明,洪亚雄,秦肖生. 改进的遗传算法在水环境模型参数估值中的应用研究[J]. 水电能源科学, 2002, **20**(1):38～40.

[9] 王凌. 智能优化算法及其应用[M]. 北京:清华大学出版社, 2001.

[10] Ng A W M, Pererab B J C. Selection of genetic algorithm operators for river water quality model calibration [J]. Engineering Applications of Artificial Intelligence, 2003, **16**: 529～541.

[11] Grefenstette J J. Optimization of control parameters for genetic algorithms [J]. IEEE Trans. On Systems, Man, and Cybernetics, **SMC-16**(1): 122～128.

[12] 姜同川. 正交试验设计[M]. 山东: 山东科学技术出版社, 1985. 1～71.

[13] Ambrose R B, Wool T A, Martin J L, *et al*. WASP5.x, A Hydrodynamic and Water Quality Model Model Theory, User's Manual, and Programmer's Guide[M]. Draft: Environmental Research Laboratory, US Environmental Protection Agency, 1993.

[14] DiToro D M, Matystik J W F. Mathematical Models of Water Quality in Large Lakes Part 1: Lake Huron and Saginaw Bay [M]. U. S. Environmental Protection Agency, Duluth, Minnesota. 1980, EPA/600/3-80-056.

[15] Thomann R V, Fitzpatrick J J. Calibration and Verification of a Mathematical Model of the Eutrophication of the Potomac Estuary[M]. Washington, D. C.: Prepared for Department of Environmental Services, Government of the District of Columbia, 1982.

[16] Bowie G L, Mills W B, Porcella D B, *et al*. Rates, Constants, and Kinetic Formulations in Surface Water Quality Modeling, Second Edition[M]. U. S. Environmental Protection Agency, Athens, GA. 1985, EPA/600/3-85/040.

[17] 贾海峰. GIS 强化下的水库水质模拟及其在密云水库中的应用研究[D]. 北京: 清华大学, 1999.