DOI:10.11992/tis.2016030
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0918.010.html

个体最优共享 GEP 算法及其气象降水数据预测建模

彭昱忠[1,2],元昌安[1],李洁[3],许明涛[1],陈冰廉[1]
(1.广西师范学院 计算机与信息工程学院,广西 南宁 530021; 2. 广西师范学院 北部湾环境演变与资源利用教育部重点实验室,广西 南宁 530001; 3.广西科技师范学院 数计系,广西 柳州 545004)

摘 要:针对基因表达式编程算法存在进化后期收敛慢且容易陷入局部最优而降低其数据建模的性能问题,和降水量因受诸多自然因素相互影响而难以准确地建模与预测的问题,提出了一种改进的基因表达式编程算法。该算法具有染色体最优状态记忆功能,在进化过程中可以按条件学习自身的历史经验知识,以加强局部搜索能力和促进收敛,同时尽量控制个体的趋同化而保持种群的多样性。3 组不同区域和不同类型的真实降水数据集的实验验证了其可以改善传统 GEP 算法后期收敛慢的问题,寻优能力更强,降水数据拟合和预测效果均显著优于传统 GEP 算法、BP 神经网络和 NAR 神经网络等算法。

关键词:基因表达式编程;经验共享;时间序列;气象建模;降水预测;演化计算;演化建模
中图分类号:TP391 **文献标志码**:A **文章编号**:1673-4785(2016)03-0401-09

中文引用格式:彭昱忠,元昌安,李洁,等.个体最优共享 GEP 算法及其气象降水数据预测建模[J]. 智能系统学报, 2016, 11(3): 401-409.
英文引用格式:PENG Yuzhong, YUAN Changan, LI Jie, et al. Individual optimal sharing GEP algorithm and its application in forecast modeling of meteorological precipitation[J]. CAAI transactions on intelligent systems, 2016,11(3): 401-409.

Individual optimal sharing GEP algorithm and its application in forecast modeling of meteorological precipitation

PENG Yuzhong[1,2], YUAN Changan[1], LI Jie[3], XU Mingtao[1], CHEN Binglian[1]
(1. College of Computer & Information Engineering, Guangxi Normal University, Nanning 530023, China; 2. Key Lab of Beibu Gulf Environment Change and Resource Use of ministry of Education, Guangxi Normal University, Nanning 530001, China; 3.Department of Mathematics and computer science, Guangxi Science and Technology University, Liuzhou 545004,China)

Abstract: Gene expression programming (GEP) is characterized by slow convergence and ease of falling into a local optimum in the later stages of its evolution. Many methods are difficult to model and use to accurately forecast precipitation because of the simultaneous influence of many natural factors. In this paper, we propose an improved GEP algorithm, which has an optimal state memory function, can learn from historical experience in the process of evolution to strengthen the local search ability, and can thus promote convergence and, at the same time, control the convergence of individuals and maintain the diversity of the population. The experimental results of three groups from different regions and different actual precipitation data sets show that the proposed algorithm can improve the slow convergence problem of the traditional GEP algorithm and has better search ability. Experimental results also show that the proposed algorithm´s ability to fit and forecast precipitation data is significantly better than that of traditional GEP algorithm, as well as the BP and NAR neural network algorithms.

Keywords: gene expression programming; experience sharing; time series; meteorology modeling; precipitation forecasting; evolutionary computation; evolution modeling

大气系统是个极为复杂的动态巨系统,具有高维性、多尺度性、复杂性、开放性、混沌性、非平稳性、不确定性和动态性等特点。传统上,被主要用于建立预测模型的常规统计方法难以精确描述大气系统的复杂关系,因而预测质量较低。近年来,利用先进

收稿日期:2016-03-18. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金项目(61562008、41575051);广西科学研究与技术开发计划项目(1598019-1)、广西高校科学技术研究重点项目(ZD2014083).
通信作者:李洁. E-mail:lijie980522@163.com.

的智能计算和数据挖掘方法,构建和改进气象预测的方法与模型,帮助对未知气象规律的认识和提高气象预测能力,已逐渐成为气象、数学和计算机领域专家和学者们关注的热点,多个相关国际会议上设置了相关的专题和 Workshop[1]。

近年来,被众多学者应用到气象或灾害天气的预测中的神经网络方法等智能计算方法[2-9]可有效描述气象要素间的复杂关系,但这些算法结构和参数难选定、计算量过大而不利于大容量样本学习等自身固有的缺陷,严重降低了其应用和发展的效果。基于大量历史数据进行气象数据挖掘与建模预测是个较有发展前途的研究和应用方向,已吸引了不少的学者进行研究[10-13]。但用传统数据挖掘算法难于避免由于气象数据的多层次特性造成的难以建立准确模型的缺陷,从而降低了气象预报的精度。

基因表达式编程(gene expression programming, GEP)是借鉴生物遗传的基因表达规律,融合了遗传算法(GA)和遗传编程(GP)的优点发展起来的进化计算家族中的革命性新成员。GEP 不但可以轻易地进化多种形态的复杂计算程序,构建稳健而精确、可解释性较强的计算模型,而且具有很强的问题表达能力、知识发现能力和寻优能力,可有效进行数据挖掘,发现公式、规则或规律,模型的最优化等[14]。相关研究表明,GEP 能有效克服很多智能计算方法和传统数据挖掘与知识发现的不足,求解很多复杂问题表现更出色,可望是一个具有发展前途的气象数据建模与预测研究方向。但 GEP 自身还存在复杂问题建模的进化后期寻优缓慢且易陷入局部最优的缺陷。针对此问题,本文提出了个体最优共享的改进 GEP 算法(best individual shared-based gene expression programming, BIS_GEP),能更好地解决后期寻优缓慢和局部最优问题,并通过 3 组真实降水案例的实验验证了其性能。

1 GEP 基本原理概述

GEP 的个体(染色体)由单个或者多个基因组成,基因之间可以用函数符号连接起来。GEP 的基因用长度固定的符串来表示,由头和尾两部分组成。其中头部既可以包含函数符号也可以包含终结符号,而尾部则只能包含终结符号。基因中的函数符号是问题求解过程中的所需要的数学函数和逻辑运算等所有候选的函数和操作符的表示,终结符通常是问题求解过程所需要的候选变量或常量,其中尾部长度 t 和头部长度 h 之间应该满足式(1)的关系:

$$t = h \times (n - 1) + 1 \tag{1}$$

式中 n 代表函数符集中的最大操目数(可能的最多的参数个数)。GEP 的基因有基因型和表现型两种表现形式,因此,每个基因对应一个 K 表达式(表示基因编码的有效部分)和一棵表达式树。其中,K 表达式就是基因型,表达式树就是表现型,两者之间可以相互转化。如,以 $\sqrt{x^2+xy}$ 为例来说明 GEP 的染色体编码方法,这个式子可以用基因:Q+ * * *xxxyz* 表示,该基因对应的表达式树如图 1 所示。

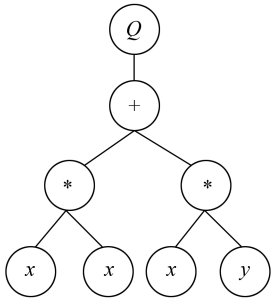


图 1 表达式树

Fig.1 The expression tree

GEP 的若干个染色体构成种群,然后通过个体在种群中不断进化而逐渐逼近问题的解。GEP 作为进化计算家族的成员,其算法的进化过程类似于 GA 和 GP。具体算法过程描述见文献[14]。

2 个体最优共享 GEP 算法

GEP 存在进化后期寻优缓慢且易陷局部最优的问题,已经引起了一些学者的注意,并试图通过控制和调节种群结构[15-17]、改进和调节遗传操作[18-20]、改变个体编码结构[21-22]等方式改进 GEP 算法,并取得了一定的成效。本文借鉴粒子群算法进化过程中粒子历史最优信息共享的机制促进粒子群算法快速收敛的思想,对 GEP 进行了改进,提出了个体最优共享 GEP(BIS_GEP)。

2.1 BIS_GEP 的基本思想

PSO 是模拟鸟群寻找食物过程的动作迁徙和群聚行为的一种启发式随机搜索的演化计算方法。GEP 和 PSO 同属仿生演化算法,本质上都是基于自然性质和行为规则随机搜索解空间寻求问题最优解。PSO 具有良好的个体最优信息共享和全局最优信息共享与更新机制,能充分利用个体自身经验和群体经验来调整自身的状态,使其位置与速度的更新具有很好的导向性。故对局部空间最优解的逼近能力很强,收敛速度快,但同时这种导向性也导致其全局搜索能力不强[23]。相对 PSO 算法,GEP 的各种遗传操作都缺乏明确的导向性,因此其对空间最优解的逼近能力不强,但这同时让 GEP 算法对空间最优解的搜索能力变得很强。经典的社会学理论认

为,人类在决策过程时,个体学习和文化传递这两类信息(即自身的经验和其他人的经验)具有极为关键的作用。对比分析 PSO 和 GEP,GEP 在对以往搜索经验的学习利用上相对较差,因为 GEP 算法的个体并不像 PSO 那样具有记忆能力,以前的知识随着种群的改变被破坏。本文认为这是导致 GEP 后期搜索慢且易陷局部最优的重要原因。BIS_GEP 算法正是借鉴了社会学理论和 PSO 的个体经验学习优势而设计的,旨在尽可能保持 GEP 自身的全局搜索优势,增强局部搜索能力和加快收敛速度。因此,在 BIS_GEP 设计上,为每个染色体设计了最优状态记忆功能,让个体在进化过程中可以充分学习自身的历史经验知识,以加强局部搜索和促进收敛。同时还需控制因过度的个体学习历史经验而引起种群个体的趋同化,尽量保持种群的多样性,让种群向全局最优移动。为了实现此目标,需要抑制个体对历史最佳状态的学习程度,避免所有个体均无节制地学习历史最佳状态而致个体快速趋同降低了种群的多样性。为此 BIS_GEP 将种群划分为两个子种群,其中一个子种群的染色体在交叉时按一定的概率与自身历史最优状态进行交叉操作,不断迭代进化,并每隔给定的 step 代通过轮盘赌选择二分之一的个体移到另一个子种群中;而另一个子种群则按常规的 GEP 算法过程进行进化,并每隔给定的 step 代排挤出适应度最差的二分之一个体移到另一子种群,同时接收选自另一子种群的个体,在迭代终止条件达到时该子种群中的最优染色体即为本次寻优过程中的最优解。BIS_GEP 算法基本思想如图 2 所示。

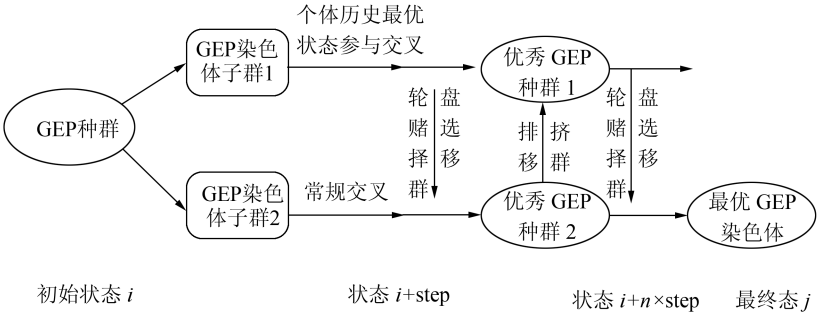


图 2　BIS_GEP 算法基本思想示意图

Fig.2　The basic algorithm idea of BIS_GEP

2.2　BIS_GEP 算法过程

根据上述主要思想,设计了如图 3 所描述的 BIS_GEP 算法流程图。

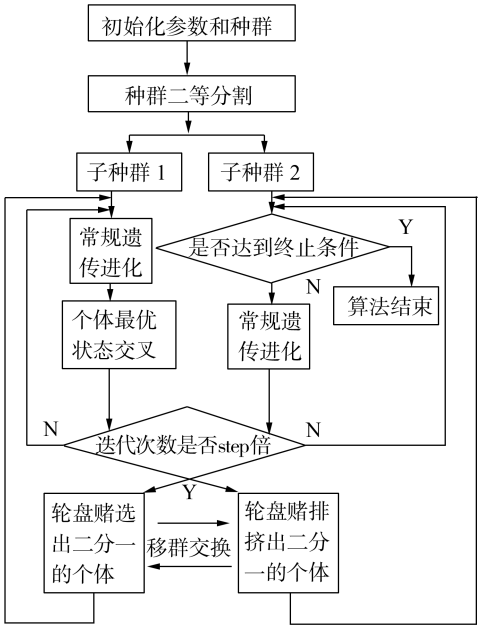


图 3　BIS_GEP 算法流程图

Fig.3　The algorithm procedure of BIS_GEP

BIS_GEP 在经典 GEP 的基础上,将种群划分为两个等规模的子种群分别按精英保留策略进行进化,然后每隔若干代即对两个相对独立的子种群进行个体选择与交换,其中的一子种群按常规的 GEP 遗传操作进化(详见文献[14]),另一子种群则在常规遗传操作的基础上增加按概率进行自身历史最佳状态(该染色体的适应度值最高时的编码表示)交叉的操作。选择个体的标准是按常规的 GEP 遗传操作进化的子种群采用轮盘赌选择取余法选择个体(选取没被轮盘赌选择法选中而排挤出的那一半),另一种群则按轮盘赌选择法进行需移群交换的个体选择。该算法通过划分子种群分别进行常规进化和外加个体历史最优交叉进化,然后隔若干代选择个体移群交换,既可通过充分学习个体自身经验加强局部搜索和促进收敛,也能保持种群的多样性,从而改善算法的寻优效果。BIS_GEP 算法描述如下:

输入　训练数据集 T,种群大小 Gs、函数集、终结符集、基因头长 HL、移群步数 step、各遗传操作率和终止条件等算法的基本参数

输出 最优函数模型 f 及其适应度

1)生成初始种群;

2)种群二等分割为 G1 和 G2;

3)种群进化过程:

```
While (终止条件 ≠ true)
    G1. CommGeneticOperateInProbability ( );//子种群 1 进行常规遗传操作
    G1. divHistoryBestCrossInProbability( );
    //子种群 1 进行个体最优状态交叉
    G1. CaculateFitness( );
    //计算子种群 1 个体适应度
    G2. CommGeneticOperateInProbability ( );//子种群 2 进行常规遗传操作
    G2. CaculateFitness( );
    //计算子种群 2 个体适应度
    If (generation Modulo step = =0)//如果当前进化代数是 step 的倍数
    G1Exchdiv=G1. RoulSelectHalf( );
    //子种群 1 进行轮盘赌选出一半的个体待交换
    G2Exchdiv=G2 - G2. RoulSelectHalf( );
  //子种群 2 进行轮盘赌排挤出一半个体待交换
    G1. Add( G2Exchdiv);
    //个体移群交换
    G2. Add( G1Exchdiv);
    End if
    G1. SelectNextPopulation( );
    //选择个体构成下一代
    G2. SelectNextPopulation( );
    generation ++;
    End while
```

4)输出结果。

3 基于 BIS_GIS 的气象降水建模与预测

3.1 数据预处理

输入数据的质量对数据挖掘与数据建模有着非常重要的影响。气象数据资料在收集过程中受到较多主观因素(如操作员认知程度等)和客观因素(如仪器设备的工作状态、环境因素等)的影响,使得气象数据不可避免地包含噪声,直接进行数据挖掘和预测建模必然会导致结果出现偏差。为了提高模型的有效性和预测结果的准确性,本文在建模前先利用菲波那契(Fibonacci)数列作为时不变线性滤波器对输入的气象数据进行滤波抑制高频噪声,然后再进行函数挖掘与建模预测。记待测时间序列为 $\{x(t),t=1,2,\cdots,N\}$,根据 Fibonacci 数列性质,取线性滤波器 H 满足式(2)[24]:

$$h_j=\begin{cases}\mathrm{Fib}(j)/\mathrm{totalWeight}, & j\leqslant K\in N\\ 0, & j>K\in N\end{cases},\quad(2)$$

$$\mathrm{totalWeight}=\mathrm{Fib}(1)+\mathrm{Fib}(2)+\cdots+\mathrm{Fib}(K)\quad(3)$$

式中:K 一般取值为滑动窗口大小减 1,则该时不变线性滤波器的输出为

$$Y_t=(\mathrm{Fib}(1)\times X_{t-k}+\mathrm{Fib}(2)\times X_{t-k+1}+\cdots+\mathrm{Fib}(K)\times X_t)/\mathrm{totalWeight}$$

3.2 建模方法

用 GEP 进行时间序列的建模和预测通常是将时间序列建模问题转换成符号回归问题,挖掘出对给定时间序列数据拟合度和对未来预测精度较高的函数模型,将用此函数模型计算未来可能的值。先求时间序列 $X(t)$ 的 M 阶延迟得到矩阵 $\boldsymbol{X}$,如式(4)所示,矩阵 $\boldsymbol{X}$ 中的元素与原序列对应关系为 $X_{ji}=x_{j+1}$,然后把矩阵 $\boldsymbol{X}$ 中的第 $N-M+1$ 列看作是所求函数模型的因变量,其余每一列看作所求函数模型的一个自变量,因而窗口大小为 $N-M+1$,而矩阵 $\boldsymbol{X}$ 的每一行即为一个样本数据,则所求的目标函数模型可记为 $x_{N-M}=f(x_0,x_1,\cdots,x_{N-M})$。接下来,GEP 根据输入样本,在给定函数符组成的所有可能函数表达空间中寻找拟合样本数据程度较佳的函数表达式。

$$\boldsymbol{X}=\begin{bmatrix}x_1 & x_2 & \cdots & x_{N-M+1}\\ x_2 & x_3 & \cdots & x_{N-M+2}\\ \cdots & \cdots & \cdots & \cdots\\ x_M & x_{M+1} & \cdots & x_N\end{bmatrix}\equiv\begin{bmatrix}x_{10} & x_{11} & \cdots & x_{1,N-M}\\ x_{20} & x_{21} & \cdots & x_{2,N-M}\\ \cdots & \cdots & \cdots & \cdots\\ x_{M0} & x_{M1} & \cdots & x_{M,N-M}\end{bmatrix}\quad(4)$$

3.3 案例实验与结果分析

3.3.1 实验数据与方案

本文分别用北京年降水量(1949-2013 年,样本长度 65,下文简称“北京降水”)、广西桂平冬季月均降水量(1951-2013 年,样本长度 63,下文简称“桂平降水”)和 UNION CITY 旱季的 6 月份降水量(1884-2006 年,样本长度 123,位于美国新泽西州东北部,下文简称“UNION 降水”)这 3 个典型的不同区域和类型的降水量作为建模预测对象,检验

BIS_GEP 预测模型实用效果。这 3 组降水案例数据的值分布如图 4 所示。其中,北京年降水数据逐年变化差异较大,突变点多而尖锐,最大值是最小值的 6 倍之多,数据的分布曲线相当复杂;桂平降水数据尽管最大值是最小值的 6 倍之多,但其逐年变化曲线比北京年降水数据逐年变化曲线平滑,突变点少;UNION CITY 降水数据波动幅度较小,数据的分布曲线相对平稳。

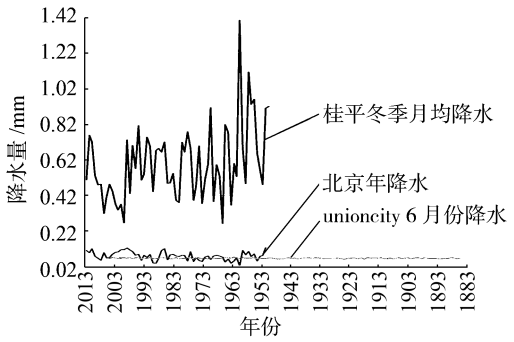


图 4　实验案例数据的值分布

Fig.4　The value distribution of experimental data

本文实验验证的主要方案是,先分别用原始 GEP 算法、GEP 改进算法 ADF_GEP 和本文所提 BIS_GEP 算法对 3 个降水案例数据集进行拟合建模,观察和比较 3 种 GEP 算法的收敛过程,验证 BIS_GEP 收敛性能改善效果。然后用这 3 种方法,以及被大气科学领域运用较多的 BP 和 NAR 等神经网络建模预测算法分别对 3 个降水案例数据集进行建模与预测,比较分析所得结果进而验证 BIS_GEP 的建模预测性能。

3 组实验均保留序列中最后 10%的样本作为测试样本,其余样本为训练样本,采用逐月/年预报形式预测测试样本的结果。3 组实验中所用的各算法的主要参数保持不变,其中,时间延迟系数都取 1,嵌入维数取 5。GEP 相关算法的主要参数如表 1 所示,其中的终结符 $a,b,c,d,e,\cdots$分别代表目标函数模型中的变量 X_0, $X_1,\cdots$, X_{N-M-1}。本文实验中的 BP 神经网络和 NAR 神经网络的均用 MATLAB 中的神经网络相关类构建,隐层数均为 20,BP 采用的其他主要参数如下:传递函数为 tansig,训练函数为 traingdm, epochs = 10 000, $lr = 0.000\ 1$, $mc = 0.5$;NAR 采用的其他主要参数如下:trainRatio = 70/100, valRatio = 15/100, testRatio = 15/100。

表 1　实验中的 GEP 及改进算法的主要参数

Table1　Main parameters of GEPs on experiments

参数名	原始 GEP	ADF_GEP	BIS_GEP
最大进化代数	2 000		
群体规模	100		
函数集	+,-,×,/, sin, cos, exp, log, sqrt		
终结符集	a,b,c,d,e,f		
头长/同源基因头长	8/无	8/3	8/无
基因数/同源基因数	5	5/3	5
交叉率(单点、两点一致)	0.2	0.2	0.2
变异率	0.25	0.25	0.25
基因迁移率(IS 和 RIS 一致)	0.1	0.1	0.1
适应度函数	MREF		

3.3.2　收敛性验证实验与结果分析

本文首先对 BIS_GEP 算法的改进性能进行验证。分别用 3 种 GEP 算法对北京降水数据集、桂平降水数据集和 UNION CITY 降水数据集进行自动建模,模型评价函数为平均相对误差。本文为避免因进化过程中的初始几代的适应度与目标值间的差异过大影响收敛过程曲线图展示效果,在画图时均忽略前 5 代的收敛过程曲线。桂平降水数据集实验的进化过程(见图 6)的前期适应度与中后期的差异较大,本文根据该收敛过程特点将其进化收敛过程图拆分成 5~125 代(见图 7(a))和 125~2 000 代(见图 7(b))两部分,以便更清晰地展示算法收敛过程的效果。

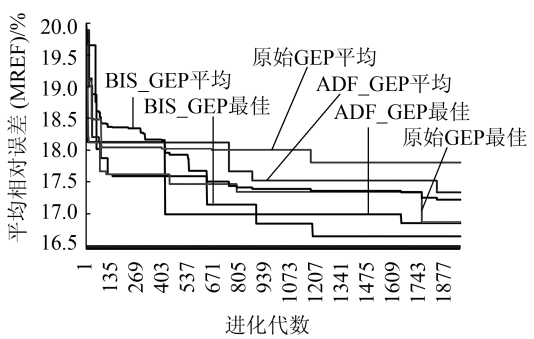


图 5　北京年降水量建模进化收敛过程

Fig5onvergence process of precipitation modeling of Beijing

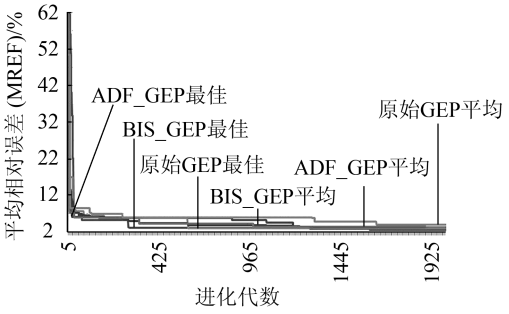
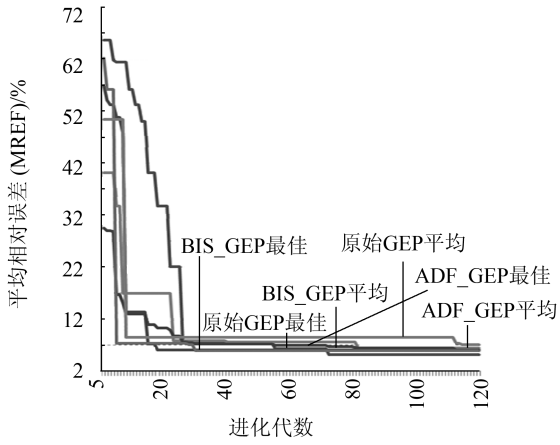


图 6 桂平冬季月均降水量建模进化收敛过程图

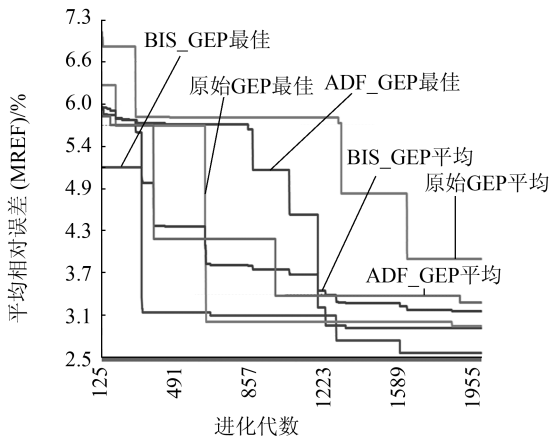
Fig.6 Convergence process of precipitation modeling of Guiping

BIS_GEP、原始 GEP 和 ADF_GEP 等对 3 组不同数据集的自动建模的进化收敛实验结果如图 5~8 所示。由图可知:1)图 5~8 均表明了 BIS_GEP 算法在这 3 组不同数据集的自动建模过程中,无论是 10 次运行结果的平均值,还是最佳运行状况,BIS_GEP 比原始 GEP 和 ADF_GEP 均有更好的收敛性能和寻优结果表现。这充分说明了本文提出的改进方法的有效性和优越性。2)图 5~8 中的左边部分显示的进化过程初期的适应度曲线均显示了在算法进化的初期,如图 5 显示的北京降水实验中的前 70 代、图 7(a)显示的桂平降水实验中的前 40 代和图 8 显示的 UNION CITY 降水实验中的前 200 代,BIS_GEP、原始 GEP 和 ADF_GEP 这 3 种 GEP 算法有近似的收敛性能表现。它们几乎都以极快速度趋于目标方向收敛,然后收敛速度逐渐减小,甚至不同程度地进入收敛缓慢状态,陷入局部最优。这说明了 GEP 算法存在着遗传算法家族常见的不足——前期收敛快,后期收敛缓慢甚至陷入局部最优。3)图 5、图 7~8 中的右边的适应度曲线均显示的进化过程中后期的 BIS_GEP 算法的适应度迭代进化比同阶段的原始 GEP 和 ADF_GEP 的更频繁,更能跳出局部最优而向全局最优方向逼近。这表明了经过本文提出的个体最优共享改进 GEP 算法可有效改善 GEP 算法后期收敛缓慢状态和易陷入局部最优的不足,寻优性能比原始 GEP 和 ADF_GEP 有显著的提高。4)在 UNION CITY 的降水实验中,BIS_GEP、原始 GEP 和 ADF_GEP 这 3 种 GEP 算法在前 600 代的适应度迭代进化较北京降水实验和桂平降水实验的表现更明显和更频繁,且更快速地逼近全局最优。5)从图 5~8 可知,UNION CITY 降水实验的平均相对误差比北京降水实验和桂平降水实验的明显小很多,桂平降水实验的平均相对误差也比北京降水实验的明显小很多。这说明了 GEP 算法在进行 UNION CITY 降水自动建模中的效果最好,在北京降水自动建模中的效果较差。

从图 4 可看出 UNION CITY 的降水量数据波动范围相对较小、数据分布相对平稳、噪声少,而北京降水数据逐年变化差异较大、突变点多而尖锐、最大值与最小值差距大、数据的分布曲线相当复杂。这些数据集的特点与 4)和 5)的情况充分表明了时间序列建模的效果与数据集的复杂程度呈强相关,建模数据分布和变化越简单,自动建模的平均相对误差越小,建模效果越好。



(a) generation 5~125



(b) generation 125~2 000

图 7 桂平冬季月均降水量建模进化收敛过程二分解图

Fig.7 Second decomposition for convergence process of precipitation modeling of Guiping

表 2　各算法的 3 组降水案例数据集建模与预测实验结果

Table2　Experimental results of mentioned algorithms of modeling and forecasting on 3 group of precipitation cases

数据与指标		BIS_GEP		ADF_GEP		原始 GEP		BP		NAR	
		拟合	预测	拟合	预测	拟合	预测	拟合	预测	拟合	预测
北京降水	MREF 最佳	16.64	23.71	16.85	24.98	16.87	26.78	27.48	47.15	29.17	36.77
	MREF 平均	17.22	24.85	17.34	27.13	17.80	29.17	28.38	51.64	30.42	37.89
桂平降水	MREF 最佳	2.57	7.16	2.92	7.83	2.95	8.04	15.69	30.86	10.97	19.65
	MREF 平均	3.16	10.25	3.28	11.51	3.90	12.18	17.08	32.17	12.29	21.83
UNION 降水	MREF 最佳	0.021 2	0.034 7	0.022 3	0.036 2	0.022 4	0.039 2	1.205 1	3.221 1	0.087 2	0.368 2
	MREF 平均	0.022 3	0.041 6	0.024 1	0.043 4	0.023 7	0.046 7	1.378 4	3.756 3	0.096 5	0.411 3

3.3.3　建模与预测效果比较验证

BIS_GEP 算法与其他 GEP 算法和气象界常用神经算法进行了自动建模与预测比较实验,取 3 组降水案例数据集的后 10%样本(北京降水和桂平降水的数据集均取 2008-2013 年的样本,UNION CITY 降水数据集取 1995-2006 年样本)作为预测的检验样本,其余样本为训练样本。采用逐月/年预报形式预测检验样本的结果,如,用 1949-2007 年真实的北京降水数据建模所得模型预测 2008 年北京降水量,接着继续用 1949-2008 年真实的北京降水数据建模所得模型预测 2009 年北京降水量,依次类推。同理,用于北京降水、桂平降水和 UNION CITY 降水实验中。比较结果如表 2 所示,MREF 最佳预测值是取 10 独立运行算法所得的 10 次各个预测检验样本预测结果平均值中的最小者,而 MREF 平均预测值是取 10 独立运行算法所得的 10 次各个预测检验样本预测结果的综合平均值。

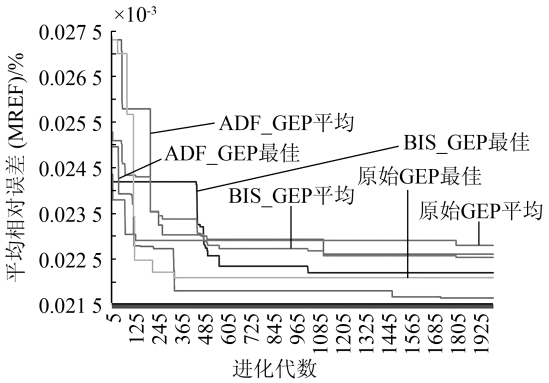


图 8　UNION CITY 每年 6 月降水量建模进化收敛过程

Fig.8　Convergence process of precipitation modeling of UNION CITY

从表 3 可看出,总体上,在 3 组不同类型和不同复杂度的真实降水数据集的实验中的 3 种 GEP 方法的拟合和预测实验结果均比另外两种神经网络方法更好。而 BIS_GEP 算法在实验上获得较其他算法更好的数据模型拟合性能和预测性能,模型具有一定的适用性。在 UNION CITY 降水数据集上的数据拟合和预测的平均相对误差 10 次运行得的最佳值分别达到 0.021%和 0.034%。据表 3 数据可知,BIS_GEP 比实验中效果第二好的 ADF_GEP 的相应 MREF 最佳值分别减少了 4.93%和 5.55%。这比实验中效果最差的 BP 的相应 MREF 最佳值分别减少了 99.45%和 99.89%。即使在逐年变化差异较大、突变点多而尖锐、最大值与最小值差距大、数据的分布曲线相当复杂的北京降水数据集上,数据拟合和预测时,BIS_GEP 算法的平均相对误差也都能分别保持在 18%和 25%以内。据表 3 数据可知 BIS_GEP 比实验中效果第二好的 ADF_GEP 的相应 MREF 最佳值分别减少了 1.25%和 5.08%。这比实验中拟合效果最差的 NAR 和预测效果最差的 BP 的相应 MREF 最佳值分别减少了 42.94%和 49.71%。这些实验对比结果充分表明了本文提出的 BIS_GEP 算法较其他算法在降水序列数据自动建模和预测上有较强优势。

4　结束语

本文提出了一种个体最优共享的 GEP 改进算法 BIS_GEP,并在 3 组真实时间序列的自动建模和预测实验中,与原始 GEP 算法、另一经典的 GEP 改进算法 ADF_GEP,以及 BP 神经网络和 NAR 神经网络进行比较。算法收敛过程实验对比结果表明 BIS_GEP 能相对改善 GEP 进化后期收敛缓慢和容易陷入局部最优的缺陷,具有更强的逼近最优能力;自动建模能力与预测能力实验对比结果表明,BIS_GEP 在 3 组不同类型的降水数据的数据拟合和数据预测实验中,10 次独立运行的最佳平均相对误差和平均相对误差均比实验中的其他 GEP 算法和神经网络算法更小,说明其自动建模能力和模型泛化

能力均有较强的优势。

对 3 组不同类型和不同复杂度的真实降水数据集的拟合和预测的对比实验结果表明,本文 BIS_GEP 算法对降雨时间序列数据的建模和预测结果比传统 GEP 及其改进算法 ADF_GEP、常用的 BP 和 NAR 神经网络自动建和预测算法的效果好,模型具有一定的适用性,同时由于该算法模型对资料要求比较单一,只需降水历史数据,因而具有广泛的应用价值。

总之,BIS_GEP 的改进是有效的,并为气象时间序列预测建模提供了一种切实可行的方法。下一步工作是进一步研究和修改 BIS_GEP 算法,并将其应用于高维多要素气象预测建模的研究和应用中。另外,该方法若在实际业务中大规模推广应用还有若干问题有待解决,如海量高维气象数据建模的适应性和稳定性问题等,都有待进一步研究。

参考文献:

[1]彭昱忠, 王谦, 元昌安, 等. 数据挖掘技术在气象预报研究中的应用[J]. 干旱气象, 2015, 33(1): 19-27.
PENG Yuzhong, WANG Qian, YUAN Chang'an, et al. Review of research on data mining in application of meteorological forecasting[J]. Journal of arid meteorology, 2015, 33(1): 19-27.

[2]金龙, 吴建生, 林开平, 等. 基于遗传算法的神经网络短期气候预测模型[J]. 高原气象, 2005, 24(6): 981-987.
JIN Long, WU Jiansheng, LIN Kaiping, et al. Short-term climate prediction model of neural network based on genetic algorithms[J]. Plateau meteorology, 2005, 24(6): 981-987

. [3]EL-SHAFIE A, JAAFER O, AKRAMI S A. Adaptive neuro-fuzzy inference system based model for rainfall forecasting in Klang River, Malaysia[J]. International journal of the physical sciences, 2011, 6(12): 2875-2885.

[4]GOSAV S, TIRON G. Artificial neural networks built for the rainfall estimation using a concatenated database[J]. Environmental engineering and management journal, 2012, 11(8): 1383-1388.

[5]VENKADESH S, HOOGENBOOM G, POTTER W, et al. A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks[J]. Applied soft computing, 2013, 13(5): 2253-2260.

[6]RAHMAN M, SAIFUL ISLAM A H M, NADVI S Y M, et al. Comparative study of ANFIS and ARIMA model for weather forecasting in Dhaka[C]//Proceedings of IEEE international conference on informatics, electronics & vision. Dhaka, Bangladesh, 2013: 1-6.

[7]ZHAO Huasheng, JIN Long, HUANG Ying, et al. An objective prediction model for typhoon rainstorm using particle swarm optimization: neural network ensemble[J]. Natural hazards, 2014, 73(2): 427-437.

[8]HE Suhong, FENG Taichen, GONG Yanchun, et al. Predicting extreme rainfall over eastern Asia by using complex networks[J]. Chinese physics B, 2014, 23(5): 059202.

[9]WU Jiansheng, LONG Jin, LIU Mingzhe. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm[J]. Neurocomputing, 2015, 148: 136-142.

[10]DHANYA C T, KUMAR D N. Data mining for evolving fuzzy association rules for predicting monsoon rainfall of India[J]. Journal of intelligent systems, 2009, 18(3): 193-210.

[11]TERZI O. Monthly rainfall estimation using data-mining process[J]. Applied computational intelligence and soft computing, 2012, 2012: 698071.

[12]BERNARD E, NAVEAU P, VRAC M, et al. Clustering of maxima: spatial dependencies among heavy rainfall in France[J]. Journal of climate, 2013, 26(20): 7929-7937.

[13]TENG Shaohua, FAN Jihui, ZHU Haibin, et al. A cooperative multi-classifier method for local area meteorological data mining[C]//Proceedings of the 18th IEEE International Conference on Computer Supported Cooperative Work in Design. Hsinchu, Taiwan, China, 2014: 435-440.

[14]FERREIRA C. Gene expression programming: mathematical modeling by artificial intelligence[M]. Portugal: Angra do Heroismo, 2002: 1-15.

[15]胡建军, 唐常杰, 段磊, 等. 基因表达式编程初始种群的多样化策略[J]. 计算机学报, 2007, 30(2): 305-310.
HU Jianjun, TANG Changjie, DUAN Lei, et al. The strategy for diversifying initial population of gene expression programming[J]. Chinese journal of computers, 2007, 30(2): 305-310.

[16]李太勇, 唐常杰, 吴江, 等. 基因表达式编程种群多样性自适应调控算法[J]. 电子科技大学学报, 2010, 39(2): 279-283.
LI Taiyong, TANG Changjie, WU Jiang, et al. Adaptive population diversity tuning algorithm for gene expression programming[J]. Journal of university of electronic science and technology of China, 2010, 39(2): 279-283.

[17]宣士斌, 刘怡光. 基于混合差异度控制的基因表达式编程[J]. 模式识别与人工智能, 2012, 25(2): 186-194.
XUAN Shibin, LIU Yiguang. GEP evolution algorithm based on control of mixed diversity degree[J]. Pattern recognition & artificial intelligence, 2012, 25(2): 186-194.

[18]TANG Changjie, DUAN Lei, PENG Jing, et al. The strategies to improve performance of function mining by gene expression programming: genetic modifying, overlapped gene, backtracking and adaptive mutation[C]//Proceedings of the 17th Data Engineering Workshop. Ginowan, Japan, 2006: 100-106.

[19]BAUTU E, BAUTU A, LUCHIAN H. AdaGEP-an adaptive gene expression programming algorithm[C]//Proceedings of IEEE International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. Timisoara, Romania, 2007: 403-406.

[20]元昌安, 唐常杰, 左劼, 等. 基于基因表达式编程的函数挖掘-收敛性分析与残差制导进化算法[J]. 四川大学学报:工程科学版, 2004, 36(6): 100-105.
YUAN Chang'an, TANG Changjie, ZUO Jie, et al. Function mining based on gene expression programming-convergency analysis and remnant-guided evolution algorithm[J]. Journal of Sichuan university :engineering science edition, 2004, 36(6): 100-105.

[21]RYAN N, HIBLER D. Robust gene expression programming[J]. Procedia computer science, 2011, 6: 165-170.

[22]ZHONG Jinghui, ONG Y S, CAI Wentong. Self-learning gene expression programming[J]. IEEE transactions on evolutionary computation, 2016, 20(1): 65-80.

[23]张鑫源, 胡晓敏, 林盈. 遗传算法和粒子群优化算法的性能对比分析[J]. 计算机科学与探索, 2014, 8(1): 90-102.
ZHANG Xinyuan, HU Xiaomin, LIN Ying. Comparisons of genetic algorithm and particle swarm optimization[J]. Journal of frontiers of computer science and technology, 2014, 8(1): 90-102.

[24]陈宇, 唐常杰, 钟义啸, 等. 基于基因表达式编程和时变强度的时间序列预测[J]. 计算机科学, 2005, 32(7 Suppl. B): 269-271.
CHEN Yu, TANG Changjie, ZHONG Yixiao, et al. Time series predication based on gene expression programming and time series vibration intensity[J]. Computer science, 2005, 32(7 Suppl. B): 269-271.

作者简介:

彭昱忠,男,1980 年生,副教授,主要研究方向为智能计算及数据挖掘。主持国家级和省级基金项目 4 项,发表学术论文 21 篇。

元昌安,男,1964 年生,教授,主要研究方向为数据库与知识工程,先后主持国家级和省级基金项目 8 项,获广西科技进步奖 5 项,发表学术论文 58 篇。

李洁,女,1980 年生,讲师,主要研究方向为智能计算及数据挖掘,发表学术论文 7 篇。