遗传进化算法在时间序列建模中的应用

陈晓梅[1] 杨成祥[2]
[1](辽宁志通石油化工经销有限公司,沈阳 110004)
[2](东北大学资源与土木工程学院,沈阳 110004)
E-mail:irm@mail.neu.edu.cn

摘 要 该文把时间序列建模看作是模型结构和参数的优化搜索过程,将遗传规划与遗传算法结合起来对结构和参数共存且相互影响的复杂解空间进行全局最优搜索实现模型结构和参数的共同识别。实例分析表明该方法建立的预测模型具有较高的精度和推广预测能力。

关键词 时间序列建模 预测 遗传进化算法 遗传算法 遗传规划

文章编号 1002-8331-(2005)05-0215-03 **文献标识码** A **中图分类号** TP311.13

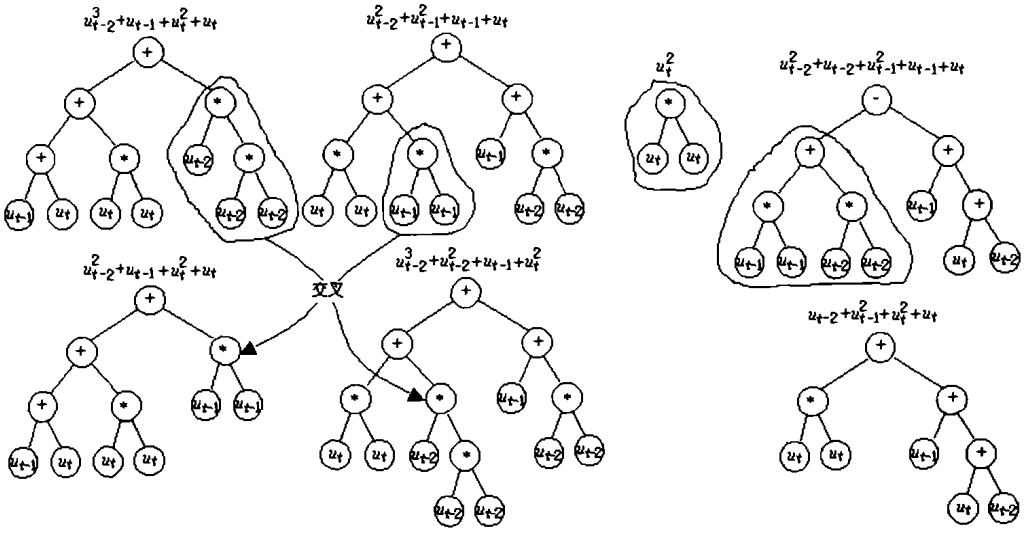
Application of Genetic Evolutionary Algorithms in Economical Time Series Modeling

Chen Xiaomei[1] Yang Chengxiang[2]
[1](Liaoning Zhitong Petrochemical Selling Ltd.,Shenyang 110004)
[2](School of Resources & Civil Engineering,Northeastern University,Shenyang 110004)

Abstract: Time series modeling can be considered as optimal search processes of model structures and model parameters.A new genetic evolutionary modeling method,combining genetic programming and genetic algorithms,was proposed for hybrid identification of model structure and model parameters by performing global optimal search in the complex solution space where the structures and parameters coexist and interact.Application results proved the high precision and generalization capacity of the predicting model obtained by the new method.

Keywords: time series modeling,prediction,genetic evolutionary algorithm,genetic algorithm,genetic programming

管理与决策活动中往往需要对大量的数据进行分析并作出正确的预测。时间序列数据是最常见也是十分重要的数据类型之一,时间序列建模及其预测技术因而多年来得到了广泛应用[1]。传统的分析方法多基于对数据的某种分步假设和对预测模型的简化[2,3]。由于现实世界中绝大多数系统都是非线性的复杂系统,这些简化假设条件难以与实际相符,导致预测结果往往与实际偏差较大。人工智能技术的发展让研究者找到了新的问题解决思路,神经网络方法逐渐成为研究热点之一并已不乏



(a)交叉(随机交换子树结构) (b)变异(随机替换子树结构)

图 1 符号回归中典型的遗传操作

基金项目:高等学校优秀青年教师教学与科研奖励计划项目资助

作者简介:陈晓梅,工程师,研究方向为计算机算法、管理与决策、优化控制。杨成祥,讲师,博士,研究方向为计算智能、计算机算法、人工智能在工程中的应用。

成功应用的例子[3-5],但由于其自身算法的一些缺陷,还不断完善和发展之中。作为计算智能的一大分支,遗传进化算法[7](Genetic Evolutionary Algorithm,GEA)在实现机器学习方面的优越性能提供了一个有着广阔应用前景的数据处理和知识挖掘手段。该文探讨基于遗传进化算法的时间序列建模与预测新技术。

1 遗传进化算法

遗传进化算法模拟自然界生物通过自身的进化与生存环境变化相适应的过程,从随机产生的一群初始解开始,按优胜劣汰的自然选择机制,通过复制、杂交、变异等遗传操作逐步改善直到找到满意解。具体实施时要对解进行编码并定义相应遗传操作方式以便于不同解个体之间的信息交换;构造评价函数评价试验解对问题的适应性作为选择依据。按其对进化过程侧重点不同,遗传进化算法发展出多种形式,其中遗传算法[6,7](Genetic Algorithm,GA)将搜索空间中的点描述为字符串形式,常见的如二进制串,通过串中位的改变调整其值。遗传规划[8](Genetic Programming,GP)将字符串推广到计算机程序,尤其是其符号回归技术(Symbolic Regression)更是直接对数学表达式进行操作,按树的遍历进行编码(如图1所示),通过对输入输出数据的分析建立描述复杂系统的数学表达式模型。

2 时间序列建模的遗传进化算法

2.1 基本思想

时间序列数据中蕴涵着丰富的系统信息,时间序列分析就是要发掘出隐含的知识(关系、规则等)。前后因果关系是其中最重要的一种,就是根据历史序列,找出反应系统演化规律的函数f建立预测模型。对等时间间隔时间序列$\{u_t\}$,就要得到如下形式的预测模型:

$$u_t=f(u_{t-1},u_{t-2},\cdots,u_{t-p}) \quad (1)$$

式中,p为输入历史信息时步数,反映了对系统当前状态影响最大的最临近的历史状态数。如果将长度为p的窗口在序列中移动,按式(1)就形成滚动多步预测。显然(1)式预测能力的好坏完全取决于f和p的正确确定。一般的做法是根据经验或简单数据拟合分析并采取一定的简化假设确定一种或几中模型(函数)作为f的待定形式,如线性函数,幂函数等,再由最小二乘等回归技术确定其中的参数,通过对比选择其中最好的一个作为最终结果,而p则按试验法来确定。然而由于现实数据的极其复杂性,往往无法获取足够的信息对来预见f的形式,片面的或局部的硬性指定将会导致大的预测偏差。因此需要从全局上对f和p值进行优化,遗传进化算法的广泛可用性和全局最优性提供了可靠的工具。

2.2 算法描述

应用遗传规划的符号回归技术,可以由函数和变量随机组合产生与(1)式相符的所有可能的表达式形式。为方便表达,引入符号GEA(函数;变量),其中函数可以为算术和数学函数。则(1)式的解集可以表示为如下形式:

$$u_{t+1}=GEA(+,-,*,/,\text{pow},\exp,\lg,\cdots;u_t,u_{t-1},u_{t-2},\cdots,u_{t-p}) \quad (2)$$

这样,事先不需要对f的形式作任何假定,按遗传进化机制对函数和变量进行选择实现全局组合优化,而p值的优化隐含在变量的选择中完成。同时,用式(2)表达的模型来描述复杂系统还需要确定其中一系列的模型参数,取值不当可能导致好的模型遭淘汰而给进化过程带来负面影响,遗传算法优良的复杂问题求解能力和快速全局寻优特点提供了解决方案。基于上述认识,可以将结构和参数混杂的组合优化过程分成相对独立而简单的结构和参数搜索问题,分别由遗传规划和遗传算法完成,实现模型结构和参数的分别进化、共同识别。主要包括模型结构进化、模型参数进化和模型结构评价三步:

(1)模型结构进化。模型结构进化过程只对函数和变量进行操作,而无需考虑模型参数的影响。按遍历树生成法则随机产生的一组表达式(框架)作为初始群体进入遗传进化循环,按遗传机制产生新群体。不同的是,由于现阶段得到的模型只是一个表达式框架,不能直接对其进行评价,而是放到模型参数进化完成后进行。

(2)模型参数进化。对模型结构进化中产生的每一个表达式框架,析取其参数信息(包括个数、范围和位置等)。这样对每个模型结构来说,问题就转化为定结构下的参数优化问题。按遗传算法搜索最佳模型参数得到当前结构下的优化模型。此时,由于当前结构下的每一组模型参数个体都对应一个完整的预测模型,可直接用事先构造的评价函数对其进行评价。

(3)模型结构评价。对模型参数进化得到的每个模型结构下的优化模型,用评价函数对其进行评价并将评价结构用于模型结构进化过程。一旦模型结构进化完成,其参数也同时确定下来。

算法从事先给定的函数和变量集开始,基本流程如图2所示。

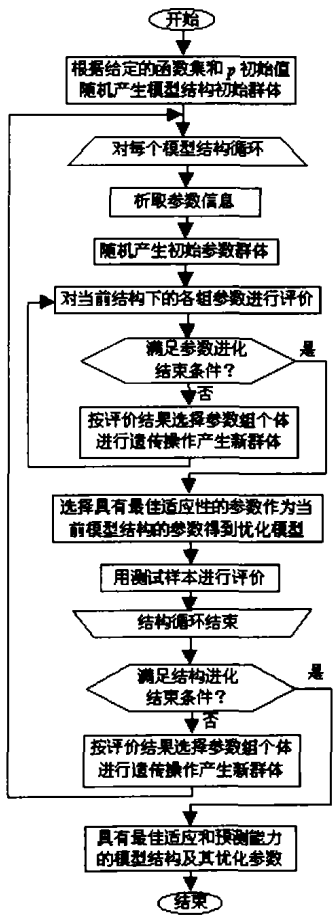


图2 算法基本流程

2.3 评价函数

试验解对问题的适应能力可以由模型预测结果和期望或

实测结果之间的比较反映出来，因此评价函数通常基于二者之间的误差分析构造，函数输出称作适应值。计算过程如下：由进化过程中产生的预测模型对应的 p 值按(1)式构造输入输出对(即学习样本)，对每个样本的输入进行预测分析，由预测输出和实际值之间的误差计算适应值。这里采用的评价函数为：

$$\text{Fitness}=\frac{1}{n}\sqrt{\sum_{i=1}^{n}(u_i-\bar{u}_i)^2} \tag{3}$$

式中，n 为样本总数，u_i 和 $\bar{u}_i$ 分别是模型预测输出和实际值。为了检验模型的推广预测能力，将学习样本分为两部分，一部分用于计算适应值，主要反映模型对样本的拟合能力，可称之为适应样本或拟合样本，另一部分用来测试模型的预测能力，称作测试样本。

3 实例分析

以我国人口自然增长率数据(1978~2001 年)为例运用上述算法进行建模和预测分析。算法参数主要包括进化种群规模和遗传操作概率等，具体设置有关文献已进行过详细的分析。该例中，模型结构进化中的群体规模为 200，交叉概率 0.8，变异概率 0.2；参数进化中群体规模为 200，交叉概率 0.98，变异概率 0.02。算法开始时给 p 一个相对较大的值 15，则模型初始变量集为 $\{u_{t-1},u_{t-2},\cdots,u_{t-15}\}$，而初始函数集为{+，-，*，/，pow，exp，lg}。(4)式是算法结束后得到的时间序列预测模型为：

$$u_t=0.961u_{t-1}-0.4095u_{t-2}+0.6993u_{t-3}-0.496u_{t-4}+0.026u_{t-5}+0.0708u_{t-6}+0.2055u_{t-7}-0.024u_{t-8}+0.0525u_{t-9}-0.1545u_{t-10} \tag{4}$$

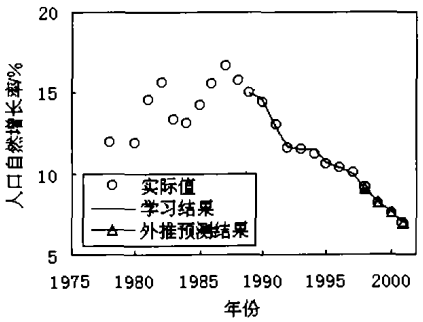


图 3 我国人口自然增长率建模和预测结果

其预测结果如图 3 所示，其中外推预测结果是将预测值看作实际值放到序列中进行下一步预测获得的结果。与实际值相比，无论是学习结果(或拟合结果)还是预测结果都具有较高的精度。图 4 给出了进化过程中最佳适应值和 p 值的变化，可以看出算法在逐步进化过程中实现了对 p 值的优化(由 15 进化为 10 个)。与初始设定的函数集比较，同样的过程也反映在函数

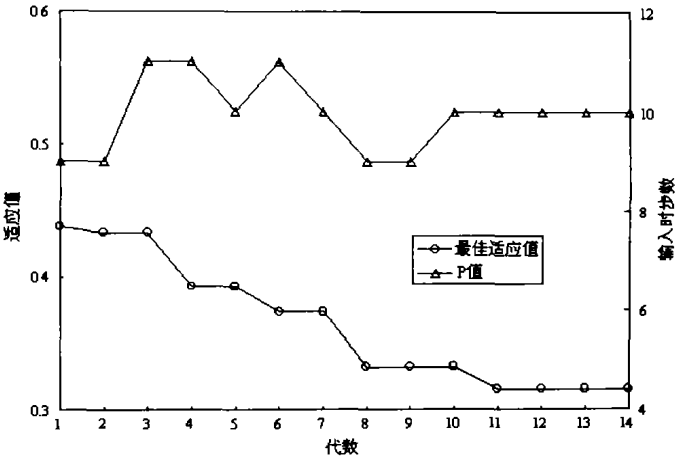


图 4 算法过程中适应值和 p 值的变化

的选择上。也就是说，算法能够自动对输入进行选择和优化组合并确定具有精简形式的预测模型。

4 结论

该文将时间序列建模看作是函数、变量和参数的组合优化过程，提出了基于遗传进化算法的全局优化搜索策略。应用遗传规划的符号回归技术对函数和变量进行优化选择，而用遗传算法优化模型参数，将二者有机结合起来从时间序列数据中发掘前后因果关系建立预测模型。实例分析表明，该算法具有较高的预测精度和推广预测能力。(收稿日期：2004 年 7 月)

参考文献

1.顾岚.时间序列分析在经济分析中的应用[M].北京：中国统计出版社，1994

2.冯春山，吴家春，蒋馥.国际石油市场的 ARCH 效应分析[J].石油大学学报(社会科学版)，2003；19(2)：18~20

3.Zhang G P.Time series forecasting using a hybrid ARIMA and neural network model[J].Neurocomputing，2003；50：159~175

4.Feng Shu-hu，Hou Yun-bing.Forecast model of energy production based on time series analysis-artificial neural network[J].Journal of Liaoning Technical University，2003；22(2)：168~171

5.钟颖，汪秉文.基于遗传算法的 BP 神经网络时间序列预测模型[J].系统工程与电子技术，2002；24(4)：9~11

6.Holland J H.Adaptation in natural and artificial systems[M].Ann Arbor：University of Michigan Press，1975

7.Goldberg DE.Genetic Algorithms in Search，Optimization，and Machine Learning [M].MA：Addison-Wesley，1989

8.Koza J.Genetic Programming：On the Programming of Computers by Natural Selection[M].MA：MIT Press，1992

(上接 32 页)

6.Paul Salama，Ness Shroff，Edward J Delp.A Bayesian approach to error concealment in encoded video streams[C].In：IEEE Image Processing，Proceedings.，International Conference on，1996；1

7.W-M Lam，A R Reilbman，B Liu.Recovery of lost or erroneously received motion vectors[C].In：Proc ICASSP，1993；5：417~420

8.H Sun，K Challapali，J Zdepski.Error concealment in digital simulcast AD-HDTV decoder[J].IEEE Trans Consumer Electron，1992；38(3)：108~116

9.J W Suh，Y S Ho.Error concealment techniques for digital TV[J].IEEE Trans.on Broadcasting，2002；48(4)：299~306

10.Marr D C，Hildreth E C.Theory of Edge Detection[C].In：Proceedings of Royal Society of London，B207，1980：187~217