doi:10.3969/j.issn.1673-4785.2010.04.003

改进基因表达式编程在股票中的研究与应用

钱晓山[1,2],阳春华[1]

(1. 中南大学 信息科学与工程学院,湖南 长沙 410083; 2. 宜春学院 物理科学与工程技术学院,江西 宜春 336000)

摘 要:简要介绍了基因表达式编程方法的基本原理,针对股票指数分析与预测问题,在经典的GEP算法基础上,提出了一种基于动态变异算子的改进的GEP算法-IGEP(improved GEP)算法,动态变异算子随着进化代数和染色体所含基因数目不同而变化,从而加快了GEP的收敛速度和精确度.还对算法进行了复杂度和收敛性分析.最后设计了一种基于IGEP的股票指数分析与预测算法,数值实验结果表明该算法优越于经典GEP算法,非常有效且具有较广泛的通用性.

关键词:基因表达式编程;复杂度分析;收敛性分析;股票指数预测

中图分类号:TP18 **文献标识码**:A **文章编号**:1673-4785(2010)04-0303-05

Improved gene expression programming algorithm tested by predicting stock indexes

QIAN Xiao-shan[1,2], YANG Chun-hua[1]

(1. School of Information Science and Engineering, Central South University, Changsha 410083, China; 2. Physical Science and Technology College, Yichun University, Yichun 336000, China)

Abstract: The authors reviewed basic principles of gene expression programming (GEP). On that basis, an improved GEP algorithm, or IGEP, was created, based on a dynamic mutation operator. The dynamic mutation operator changed with the gene number of the genome and the number of evolutionary generations. The complexity and convergence properties of the algorithm were investigated. The new IGEP was used to predict stock-market indexes. Simulation results indicated that the IGEP-based model is more accurate than the classical GEP-based model.

Keywords: gene expression programming; complexity analysis; convergence analysis; prediction in stock-price index

基因表达式编程[1](gene expression programming,GEP)是葡萄牙科学家C. Ferreira发明的一种基于基因组(genome)和表现型(phoneme)的新的遗传算法.它与遗传算法(genetic algorithms,GA)和遗传规划(genetic programming,GP)的根本区别在于他们所采用个体的本性不同:即在GA中个体是固定长度的线形串(染色体);在GP中个体是长度和形状不同的非线形性实体(分列树),而在基因表达式编程中个体首先被编码成固定长度的线形串(基因组或者染色体),然后被表达成不同长度和形状的非线形实体(简单图表示,或者表达式树).

收稿日期:2009-02-15.
基金项目:国家自然科学基金资助项目(60634020,60874069,60804037);国家"863"资助项目(2006AA04Z181).
通信作者:钱晓山. E-mail:qianxiaoshan@126.com.

1 基因表达式编程方法原理

基因表达式编程的实现技术主要包括编码方式、遗传算子、插串操作、重组算子、适应度函数选择、数值变量等几个部分[2],下面就涉及的改进部分作一介绍.

1.1 变异算子

变异(Mutation)可以发生在染色体内的任何位置,然而,染色体的结构组织必须保持完整.在基因头部,任何符号都可以变异成函数符号或者终点;在基因尾部,终点只能变异成终点.通过这种方法,染色体的结构组织得以保持,由于GEP编码方式的特点,可以预见变异产生的新个体在结构上是正确的.值得注意的是,在GEP中既没有变异种类的限制,也没有一个染色体中变异次数的限制:在所有情况中,新生的个体在句法上都是正确的.

1.2 适应度函数选择

除了个体的表示外,还需要定义个体的适应值评价函数. 这是本文的研究重点. 在 GEP 中,适应度函数设计非常重要,一般采用如下 3 种方式[3]:

$$f_i = \sum_{j=1}^{C_i} (M - | C_{(i,j)} - T_{(j)} |), \quad (1)$$

$$f_i = \sum_{j=1}^{C_i} \left(M - \left| \frac{C_{(i,j)} - T_{(j)}}{T_{(j)}} \cdot 100 \right| \right), \quad (2)$$

$$\text{if } n \geqslant \frac{1}{2} C_i, \text{then } f_i = n, \text{ else } f_i = 1. \quad (3)$$

式中:M 为一常量,f_i 是控制适应度 f_i 的取值范围;$C_{(i,j)}$ 表示第 i 个基因对应的函数表达式中利用第 j 个样本变量数据求得的函数值;$T_{(j)}$ 表示第 j 个样本中包含的实际测得的该目标函数的真实值;C_i 是测试样本数据总数,n 是正确适例的个数. 式(1)和式(2)可以解决任何一个符号回归问题. 式(3)主要解决逻辑合成问题. 在适应度的设计上,目标非常明确,让系统按照要求的方向进化.

2 改进的基因表达式编程方法算法

2.1 IGEP 算法描述

经典的 GEP 算法对变异算子的考虑不够,本文在做了大量的实验后,就变异算子的变异方法给出了一种改进的方案. 基于改进变异算子的 IGEP 算法结构如下[4]:

1)初始化种群,随机产生 60 组(过大会增大算法运行时间,过小则很难收敛)初始染色体,每个染色体由 5 个基因构成,每个基因头长度为 15(或更多);初始化时采用 KARVA 编码[4]:

Q * + - a b c d;

2)按照适应值函数求解初始种群的各染色体的适应值(本文采用的适应度函数)

$$f_{\text{fitness}}(i) = 1000 \times \frac{1}{Ei + 1}, \quad (4)$$

并将适应值排序,保存适应值最高的个体;

3)执行变异,并按照染色体所含基因的多少决定变异的基因位个数,本文选择每个基因变异一个基因位的方法;

4)执行 IS 插串、RIS 插串、Gene 插串;

5)执行单点重组、两点重组、基因重组;

6)运行代数增加 1,如果运行达到预先设定的最大代数,则停止运行,用图形输出结果并保存到记录文件中,否则转到 2)继续运行.

2.2 IGEP 算法复杂度分析:

引理 算法的的复杂度是 $O(K * L * M)$,其中 K 为种群大小,L 为总进化代数,M 为训练集的长度.

证明 在算法中计算个体针对 M 个训练样本的复杂度为 $O(M)$,算法需要计算种群中各个体的适应度值,故种群适应度计算的复杂度为 $O(K * M)$,因算法最多进化 L 代,故算法复杂度为 $O(K * L * M)$.

3.3 IGEP 算法收敛性分析

在基于 IGEP 的复杂函数参数识别反问题求解中,将适应度函数设计如(4)式所示. 其中,$E_i = \frac{1}{m} \sum_{j=1}^{m} (Y_{(ij)} - Y_j)^2$ 为均方误差和计算公式,简记为 SSE. 设遗传变异率 $P_{mn} \leqslant 0.5$,则可得如下 2 个结论:

1)IGEP 复杂函数反问题求解进化第 K 代的最小均方误差和的数学期望满足

$$\mathrm{E}(SSE^k_{\min}) \leqslant \mathrm{E}(SSE^0_{\min}) - p_r p_t p_{\text{all}} e \sum_{i=1}^{k-1} E(b_i). \quad (5)$$

2) 最小均方误差和以概率收敛,即对任意 $\varepsilon > 0$,有

$$\lim_{k \to \infty} p(SSE^k_{\min} > \varepsilon) = 0. \quad (6)$$

以上分析可看出,基于 IGEP 的复杂函数参数识别反问题求解是依概率收敛到全局最优染色体的,但是,由于不是强收敛到全局最优点,因此不能排除收敛到局部最优的可能.

3 数值实验——基于 GEP 和 IGEP 的上证指数时间序列分析

3.1 股票指数预测研究

股票市场是一个复杂的非线性动力系统,同时受多种因素的交互影响,对于股票未来价格的精确预测是非常困难的. 股市预测被认为是当前时间序列预测中最富挑战性的应用之一,受到数据挖掘界的广泛关注[5]. 股票指数涨跌数据是一种时间序列数据,它既具有一定的趋势性又具有较大随机性. 自 19 世纪股票市场建立以来,股票指数预测模型就成为各国学者研究的焦点. 在时间序列预测中,线形的概率统计模型曾得到广泛的应用,如:ARMA 模型法、AR 模型法、阈值自回归、多项式自回归、指数自回归模型等,后来还有灰色预测、混沌时间序列预测等方法. 如 White(1985)[6] 曾经尝试利用神经网络来预测 IBM 普通股每日报酬率,但是预测结果不甚理想;Bergerson and Wunsch(1991)[7] 利用 S&P 指数训练神经网络, 预测股价涨跌方向的正确率相当高;Pesaran and Timmermann(1990)[8] 对过去 25 年

的伦敦证券指数进行预测,采用神经网络技术预测指数的月变化率可以达到60%的正确性;台湾地区的研究中,张文信(1995)以总体经济变量预测股价加权指数走势,其预测正确率约为 50% ~60% 等等. 但随着人工神经网络研究的深入,人们认识到它存在的严重不足,在原理上缺乏实质性的突破,同时也缺乏理论依据[9-10],这些研究结果表明股票指数是具有可预测性的. 基于此,本文提出改进变异算子的 IGEP 应用于股票指数的预测.

3.2　上证指数时间序列模型

以上证指数 2003 年共 239 个交易日的数据为训练数据,应用 GEP 和 IGEP 方法进行时间序列分析,以 2004 年的数据作为测试数据. 时间序列分析中,一个重要的参数是历史数据长度的选择. 我们对历史数据为 1 ~13 天分别进行了模型的建立,发现对训练数据均能进行准确的模拟. 但历史数据天数太短时,由于提供的资料信息太少,预测时效果不好. GEP 的函数集可以包含运算符{ + ,- ,* ,/ }以外再加上其它初等函数. 在实验中发现基于基本运算符建立的模型就能够达到较高的精度,所以在分析中,选择这些运算符作为函数集对 13 天的历史数据进行建模.

该文在求解中,用到了数值常量集合,数值常量集合 C 由初始函数任意生成,范围在[-10,10]之间,常量的个数与基因尾部长度值相同. 实验中的参数及参数值如表 1 所示.

表 1　参数定义

Table 1　Definition of parameters

参数名	参数值
最大代数	300
种群大小	60
函数集	‘ +’,‘ -’,‘ *’,‘/’,‘s’,‘c’
基因头长度	10
基因个数	5
连接符	+
变异率	0.044
单点重组率	0.3
两点重组率	0.3
基因重组率	0.1
IS 插串率	0.1
IS 插串长度	1,2,3
RIS 插串率	0.1
RIS 插串长度	1,2,3
Gene 插串率	0.1

股票指数预测建模:取 $f_{\text{fitness}}(i) = 1000 \times 1/(E_i + 1)$ 为适应值函数,并假设股票指数与最近 13 天有关,以此最近 13 天为变量建立微分方程,求解算法简述如下:

1)将微分方程作为遗传计算对象(染色体),初始化 IGEP 种群,从训练数据中挖掘出微分方程和初始条件;

2)用 Runge-Kutta 法求解微分方程,计算出微分方程的(作为染色体)适应度,若满足终止条件,成功退出;

3)执行 IGEP 的各种操作. 产生下一代种群,转 1).

本文在实验中还引入了对于含噪时间序列的处理算法,实验中使用了显微插值法去噪,算法简述如下:

1)对含噪的时间序列用傅立叶变换,去掉高频部分;

2)把时间区间放大 m 倍 ($10 < m < 200$);

3)对上述结果作反傅里叶变换;

4)插值、光滑化、还原(缩小)时间区间.

3.3　实验数据

本文实验数据来源于 www. sohu. com 财经版.

3.3.1　GEP 算法[11]

使用 GEP 算法,参考文献[5]给出拟合函数公式为

$$D = d_{11}/(d_4/(d_2 - d_4) - (d_1 + d_4)) + d_2/(d_0/(_d2 - d_5) - d_{11}) + d_5/(d_3 + d_0/(d_0 - d_5) + d_6 + d_{12}) + (d_6 + d_4 * d_6/d_0)/((d_{11} + d_2) * (d_7 - d_8)) + d_4/(d_1/(d_2 + d_4 - d_5 - d_7) + d_7) + d_{12}.$$

式中:d_i 表示前第 13 - i 天的数据.

文献[11]指出此公式拟合曲线的相关系数为 0.953 3.

利用此式给出训练数据和测试数据的真实数据与模型数据的曲线比较图,参见图 1,图 2:

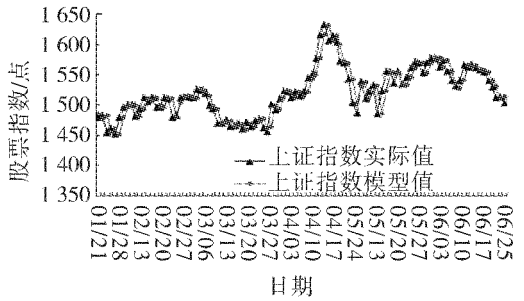

图 1　上证指数部分训练数据真实值与模型值比较曲线图

Fig. 1　The compared curve between the true value and model values of the Shanghai index part training data

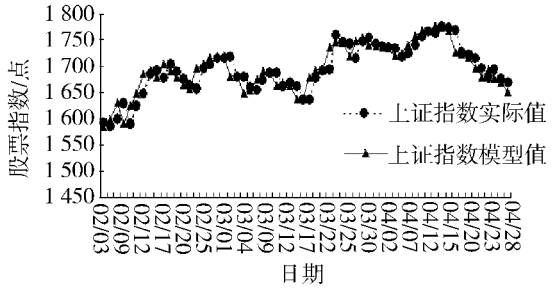

图 2 上证指数部分测试数据真实值与模型值比较曲线图

Fig. 2 The compared curve between the true value and model values of the Shanghai index part test data

4.3.2 IGEP 算法

取 $f_{\text{fitness}}(i) = 1\,000 \times 1(E_i + 1)$ 为适应值函数，本文使用 IGEP 算法得到较好的染色体如下所示（此染色体含 3 个基因，连接符采用数学运算“+”，黑体为基因头）：

/. *d*5. *. −. *d*11. *. *. −. /. /. /. *d*1. *d*0. −. *d*6. *d*8. *d*1. *d*8. *d*1. *d*1. *c*0. *d*5. *d*3. *d*2. *d*0. *c*1. *d*3. *d*3. *d*9. *d*2. *d*10 +. −. −. *d*3. +. *d*12. *d*3. /. *d*1. *. +. *d*11. +. +. *. *d*4. *d*8. *d*10. *d*3. *d*6. *d*2. *d*1. *d*8. *d*10. *d*12. *d*11. *d*10. *d*12. *d*11. *d*1. *d*9/. *d*12. *. *c*0. *. *. /. −. *. *d*1. +. *d*9. *d*0. /. *d*7. *d*2. *d*11. *d*7. *d*0. *d*2. *d*10. *d*9. *d*4. *d*11. *d*4. *d*4. *d*4. *d*6. *d*0. *d*9. *d*0/. /. /. *d*6. *. *d*0. *. −. *d*2. /. *d*4. +. *d*2. +. /. *d*12. *c*0. *d*6. *d*6. *d*0. *c*1. *c*2. *d*8. *d*1. *d*3. *d*1. *d*0. *d*11. *d*7. *d*5. *d*4*d*1. +. *d*6. /. −. *c*0. +. *d*6. /. *d*8. *d*4. *d*3. *d*1. *d*6. *. *d*8. *d*12. *d*12. *d*3. *d*9. *d*11. *d*12. *d*10. *d*9. *d*1. *d*11. *d*11. *d*4. *d*3. *d*3. *d*4/. /. /. *d*11. *. *d*10. *. −. *d*0. /. *d*12. +. *d*2. +. /. *d*12. *c*0. *d*10. *d*3. *d*0. *c*1. *d*3. *d*3. *d*7. *d*8. *d*4. *d*1. *d*10. *d*0. *d*7. *d*4/. −. *d*3. *d*1. *d*11. /. *d*6. *d*2. *d*0. +. *c*0. *d*11. *d*7. *d*11. +. *d*12. *d*2. *d*11. *d*2. *d*5. *d*1. *d*4. *d*12. *d*9. *d*11. *c*1. *d*11. *d*12. *d*4. *d*10. *c*2/. +. +. *d*5. +. *. +. *d*8. *. *d*0. *d*4. +. −. *d*2. *d*9. *d*10. *d*6. *d*0. *d*9. *d*4. *d*8. *d*10. *c*0. *d*7. *d*0. *c*1. *d*11. *d*5. *d*1. *d*11. *d*12/. *d*4. *d*0. −. *. *d*12. *. *d*3. +. *d*3. *. *d*12. *d*3. +. +. *d*3. *d*2. *d*9. *d*4. *d*11. *d*2. *d*10. *d*4. *d*5. *d*0. *d*0. *d*4. *d*8. *d*6. *d*7. *c*0

其中：

基因 1）的 $c0 = -9.979\,676$，$c1 = 1.375\,732$，

基因 3）的 $c0 = -4.226\,227$，

基因 4）的 $c0 = -7.068\,481$，$c1 = -5.105\,316$，$c_2 = -5.41\,812\,1$，

基因 5）的 $c0 = -9.595\,825$，

基因 6）的 $c0 = 6.671\,539$，$c1 = 2.517\,029$，

基因 7）的 $c0 = -6.272\,735$，$c1 = -.91\,211$，$c2 = -2.822\,785$，

基因 8）的 $c0 = 9.329\,834$，$c1 = -8.205\,658$，

基因 9）的 $c0 = -9.196\,594$.

利用 IGEP 算法提供的功能将其转化成数学模型即为

$$y = \frac{x_6}{((x_2 - x_1)\dfrac{(x_2 + 9.979\,676)}{x_7} - (\dfrac{x_9}{x_2})^2)x_{12}} - \frac{x_{12}(x_k + x_9)}{x_1 11 + x_4 + x_7 x_3} + 2x_2 + x_{13} + \frac{x_5}{x_1} + \frac{x_2 - x_{12}}{x_4} + \frac{2.517\,029 x_{12} x_{13}(x_{11} + x_4)}{(x_{13} + 6.671\,539 - x_3)x_1^2 x_{11}} + \frac{x_6 + x_9 + x_3 x_{10}}{x_1 x_5 + x_{11} + x_7 + x_1 - x_{10}} + \frac{x_1 x_2 x_3 x_{13}}{-4.226\,227 x_2 x_8^2 (x_{10} - x_1)} + \frac{2x_7^2 x_5}{(x_1 - 5.105\,316)x_1 x_3 (x_{13} - 7.068\,481 - x_3)}.$$

式中：y 表示第 14 天的值，x_i 表示前第 13 − i 天的数据.

上证指数部分训练数据真实值与模型值比较曲线如图 3 所示，其中实线为真实值，虚线为拟合值；用 IGEP 预测 04 年前 30 天的数据情况如图 4 所示.

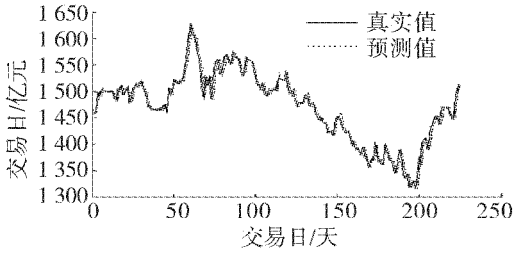

图 3 IGEP 算法上证指数部分训练数据真实值与模型值比较曲线图

Fig. 3 The IGEP algorithm for the compared curve between the true value and model values of the Shanghai index part training data

实验求出训练数据相关系数为0.977 3，测试数据相关系数为0.962 3. 由图 3 与图 4 知，其拟合效果和预测效果都优于文献[11]的 GEP 方法，证明了该文设计算法的优越性.

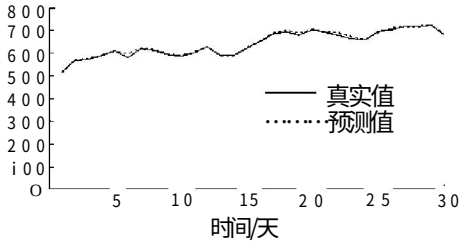

图4 IGEP算法上证指数部分训练数据真实值与模型值比较曲线图(2004年前30天)

Fig.4 The IGEP algorithm for the compared curve between the true value and model values of the Shanghai index part training data(formerly 30 days in 2004)

4 结束语

用 GEP进行预测,不需要知道各因素间的因果关系,只需要提供足够的实验或者实验数据,无须知道目标函数,就可以达到准确预测的目的.本文创新点在于提出了一种新的IGEP算法,可得到准确的函数表达式,通过实例计算和分析可知,此方法优于文献[11]的 GEP方法,并充分发挥了进化算法内含的并行性,从而使得算法在求解速度上远远快于普通的算法.

参考文献:

[1] FERREIA C. Gene expression programming: a new adaptive algorithm for solving problems [J]. Complex Systems, 2001,13(2):87-129.

[2] MITCHELL M. An introduction to genetic algorithms[M]. Cambridge: MIT Press,1996:143-164.

3] FERREIRA C. Gene expression programming: mathematical modeling by an arifcial intelgence[M].2nd Ed. Mit Press Cambridge,Massachusetts o London,England Spring-ger 2006:82-85,5,

[4] 张克俊.求解反问题的改进的基因表达式编程研究[D].南昌:江西理工大学,2006.6.

ZHANG Kejun. The study of improving gene expression pro-ogramming for solving the inverse problem[D]. Nanchang: liangxi University of science and technology,2006.6.

[5] ZUO Jie,TANG Changjie,LI Chuan,et al. Time series prediction based on gene expression programming[C]//International Conference for Web Information Age 2004. Ber-rling Heidelberg: Springer Verlag,2004: 5564.54.

6] ECONOMIC W H. Prediction using neural networks: the case of IBM daily stok returns[C] //IEE International Conference on Neural Network. 1988:3451458 58.

[7] BERGERSON K,WUNSCH D C. A commodity trading model based on a neural network-expert system Hybrid [C]//Proeedings of the IEEE International Conference on Neural Network,Seattle,1991: 1289-1293.

[8] PESARAN M H,TIMMERMANN A. A recursive modeling approach to predicting UK stock [C]//Economic Journal, Blackwell Publishing,Edward 2000:159-191.

9] 元昌安,唐常杰,左劫,等.基于基因表达式编程的函数挖掘一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 Changan,TANG Changjie,ZUO Jie,et al. Function mining based on gene expression programming-convergence analysis and remnant-guided evolution algorithm[J]. Journal of Sichuan University: Engineering Science Edition, 2004,36(6):100-105.

[10] 段磊,唐常杰,左哉,等.基于基因表达式编程的抗噪声数据的函数挖掘方法[J].计算机研究与发展,2004,41(10):1684-1689.

DUAN Lei,TANG Changjie,Zuo Jie,et al. An anti-noise method for function mining based on CEP[J]. Journal of Computer Research and Development,2004,41(10)): 1684-1689.

[11] 廖勇.基于基因表达式编程的股票指数和价格序列分析[D].成都:四川大学,2005.

LIAO Yong,Time series prediction in stock-price index and stock-price based on Gene Expression programming [D]. chengdu: Sichuan University,2005.

[12] 李曲,蔡之华,朱利等.基因表达式编程方法在采煤工作面瓦斯涌出量预测中的应用[J].应用基础与工程科学学报,2003,12(1):50-5

LI Qu,CAI Zhihua,ZHU Li,et al. Application of gene expression programming in predicting the amount of gas emited from coal face[J]. Journal of Basic Science and Engneering,2003,12(1):50-5.

作者简介:

钱晓山,男,1980年生,讲师,博士研究生,主要研究方向:复杂工业过程建模、优化控制.

阳春华,女,1965年生,教授,博士生导师,获"湖南省青年科技奖"、"湖南省十大杰出女性",主要研究方向为复杂工业过程建模、优化控制;智能信息处理.