基于基因表达式编程的私人汽车拥有量建模和预测*

朱明放[1a], 王宏涛[2], 任艳玲[1b]

(1. 江苏技术师范学院 a. 计算机工程学院; b. 电气信息工程学院, 江苏 常州 213001; 2. 陕西理工学院 计算机科学与技术系, 陕西 汉中 723001)

摘 要: 准确预测私人汽车拥有量,对制定经济政策和进行经济宏观调控、保证社会经济和谐发展有重要的作用。基因表达式编程(GEP)是新的进化模型,在数据挖掘领域得到了广泛的关注和研究,对符号回归任务表现了很强的优势。阐述了 GEP 基本原理,GEP 进行序列分析的基本方法;根据 1990—2007 年全国和人汽车拥有量,基于 GEP 技术挖掘到了其模型。实验表明,基于 GEP 技术得到的私人汽车拥有量模型预测精度高、泛化能力强。

关键词: 基因表达式编程; 私人汽车量; 预测; 泛化能力

中图分类号: TP202.7 **文献标志码**: A **文章编号**: 1001-3695(2010)03-0958-03

doi:10.3969/j.issn.1001-3695.2010.03.041

Modeling and predicting total number of private cars based on GEP

ZHU Ming-fang[1a], WANG Hong-tao[2], REN Yan-ling[1b]

(1. *a. School of Computer Technology, b. School of Electronic Information, Jiangsu Teachers University of Technology, Changzhou Jiangsu* 213001, *China*; 2. *Dept. of Computer Science & Technolgy, Shaanxi University of Technology, Hanzhong Shaanxi* 723001, *China*)

Abstract: For making economic policies and controlling macro economy, it needs predict the total numbers of private cars. It is a vital role to ensure harmoniously develops the economics. Gene expression programming (GEP) is a novel evolution system and attracts many studies and attention. This paper introduced the principle of GEP, and the basic methods applied GEP to time series analysis. According to the total number of private cars from 1990 to 2007, based on GEP techniques, mined and analyzed prediction model. Illustrating from experiments, the model has a high prediction precision, and good generalization ability.

Key words: gene expression programming(GEP); total number of private cars; prediction; generalization ability

随着社会经济的发展、人民生活水平的提高,私家车已经走入了人们的生活,其拥有量反映了人民生活的水平和质量。近几年,我国私家车拥有量迅速增长,在为人们提供交通便利的同时,也给道路交通部门、政府部门以及人们生活的小区等提出新的课题。在进行道路设计、城市规划、小区建设时,车辆的流量和停位已经成为一个重要的考虑因素,它也是政府进行石油能源的定价、经济宏观调控的主要因素。

准确的预测是科学规划、有效进行工作部署和实效管理的前提和依据,而准确的预测是建立在事物发展规律基础上的科学推断。社会经济系统是一个开放的大系统,要研究事物的发展变化规律,需要分析与考察问题相关的其他因素,而这些因素的提取和抽象是一项十分困难的任务。如果积累了事物发展的一些历史数据,则时间序列分析是一个很好的技术。为了预测私人拥有汽车量,本文使用我国近几年来私家车拥有量的历史数据,通过历史数据建立预测模型,从而对私家车拥有量进行预测。

近几年来,已经出现了将许多预测技术运用到私家车拥有量预测的问题上,诸如灰色系统理论[1]、多元线性回归分析法[2]、最小二乘法及多项式拟合技术以及神经网络算法等。这些方法存在以下问题:人为选定预测模型的结构,然后通过统计方法对模型中的参数进行确定,主观因素浓厚;需要考察与预测变量相关的其他因素,增加了研究难度等。本文应用 GEP 技术对我国私家车拥有量进行预测研究,GEP 在不必预先知道预测模型的结构和参数、不必具备领域背景知识前提下,通过进化能得到结构简单、预测精度高的模型,避免了对系统进行机理分析建立其预测模型的困难,避免了以回归方法为基础的预先设定模型结构,然后通过统计方法确定参数的主观性。

1 GEP 简介

基因表达式编程[3](GEP) 由 Ferreira 于 2001 年提出,是数据挖掘的新方法,是遗传算法家族新成员。其特点是将遗传操作和个体评价的实体相分离,即它是基因型和表现型分离进化算法,是发展完备的遗传算法,其进化效率比其先辈 GP 快 2~4 个数量级。近些年来,它在数据挖掘领域得到了广泛的关注和研究,已成功应用到函数发现[3,4]、组合优化[5]、分类[6]、聚类[7]、关联规则[8]、时间序列预测[9]等领域。

GEP 的基因型是由头部和尾部组成的有结构的线性实体,头部可以包含函数符号(取自函数集 *FS*)和终结符号(取

收稿日期: 2009-07-11; **修回日期**: 2009-08-25 **基金项目**: 江苏技术师范学院博士启动基金资助项目(KYY09001)

作者简介:朱明放(1970-),男,陕西咸阳人,副教授,博士,主要研究方向为数据挖掘、进化计算(mfzhu2009@jstu.edu.cn); 王宏涛(1976-),男,陕西汉中人,讲师,主要研究方向为算法分析与设计;任艳玲(1969-),女,陕西咸阳人,高级实验师,主要研究方向为计算机通信.

自终结符集 *TS*),尾部只能包含终结符号。对每个待解决的问题,头部长度 h 预先指定,尾部长度 t 和头部长度有如式(1)的关系。

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

其中:n 表示函数集中函数最大的参数数目。

GEP 线性固定长度的基因和柔性分支结构的表达树间的转换依赖 Karva 语言,该语言简单直观,实现了 GEP 的基因编/解码。

例 1　设 $FS=\{Q,*,/,-,+\}$,$TS=\{a,b\}$,则 $n=2$。若 $h=15$,则 $t=16$,整个基因的长度 $g=15+16=31$。下面是一个可能的基因 G_1,尾部用粗体标记。

G_1: */a Q/b*a b/Qa*b* – **ababaababbabbbba**

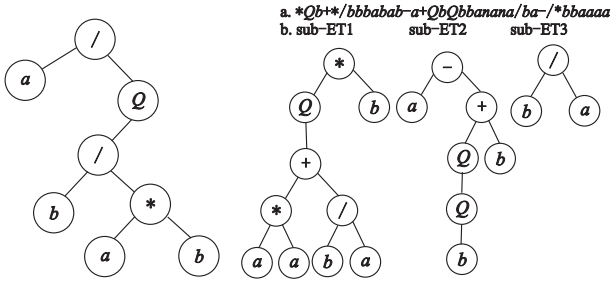
该基因经 Karva 解码,其表达树 *ET* 有八个节点,如图 1 所示。

基因 G_1 的长度是 31,开放阅读区(open reading frame,ORF)长度只有 7,位置 8 ~ 30 形成基因的中性区,GEP 的中性区以及发生在其上的遗传操作是进化成功的关键因素。该基因的数学表达式为 $a/Q(b/(ab))$。

GEP 有效进化的又一重要因素染色体(基因组)是由多基因组成。一般地,染色体由一个或几个等长的基因组成,每个基因解码为一个子表达树(sub-ET),各子表达树通过连接函数形成更复杂的多子单元的表达树(multi-subunit ET)。

例 2　设染色体有三个基因,其中 $h=6$,$FS=\{+,-,*,/,Q\}$,$TS=\{a,b\}$,则染色体的长度是 39,取连接函数是“+”。

图 2 是一个可能染色体和对应的各个子表达树,图 3 是染色体的表达树。



a)染色体　　b)子表达树

图1　基因G_1解码的表达树　　图2　多基因染色体的各子表达式

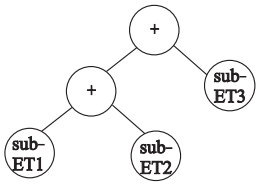


图3　多基因染色体的表达树

遗传操作是遗传算法的核心和精髓。因为 GEP 遗传操作发生在染色体的具有线性结构的基因型上,所以 GEP 有相当丰富的遗传算子,除了选择、复制外,还有变异、逆串、转位和重组等基因突变型算子。需要强调的是,GEP 的基因在任何遗传操作之后,只要保证基因的结构不变,即保证基因的头部和尾部的内容满足要求,则基因总是表达一个合法程序。

变异操作可以在染色体的任何位置操作,只要保持染色体的组织结构不变,经变异后产生的新个体均是合法的程序,即基因头元素可变异成任何 *FS* 或 *TS* 中的元素,而基因尾元素只能变异成 *TS* 中的元素,它是 GEP 技术中最重要的算子。逆算子限定在 GEP 基因的头部,随机选取头部的两个位置,将这两个位置之间的符号串取逆来引起基因突变的算子,是 GEP 中重要的操作算子作用。转位算子是选中基因的一个片断,它迁移到基因的另外一个位置。有三种转位算子,即插串算子(IS)、根插串算子(RIS)和基因插串。重组是交换双亲染色体之间的等位基因的一种遗传操作,有三种重组,即单点重组、两点重组和基因重组。这些算子的技术细节参见文献[10]。

2　基于 GEP 的时间序列预测

2.1　时间序列分析

时间序列分析的任务就是通过分析序列中历史数据所包含的信息,发现其中所蕴涵的变化规律,为其建立数学模型,运用数据模型进行分析、预测和控制。在经济、工程和其他各种领域中,到处都存在着需要进一步研究的时间序列,如国家或地方的经济分析、股市行情分析等都属于时间序列分析的范畴。对许多问题的研究和解决也就是要研究清楚有关时间序列的发展规律,实现准确的估计、预测和控制。可见,时间序列的分析研究具有非常重要的意义。

GEP 技术是一个通过人工智能进行数学建模的工具,所以 GEP 在符号回归(函数发现)领域研究得最深入,得到成果也最为丰富,同时也表现了强大的挖掘能力。时间序列分析问题是一类特殊的符号回归问题,它是利用时间序列的历史数据和当前数据对未来数据进行预测和控制。若将历史数据和当前数据看做自变量,未来的预测数据看做因变量,则可将时间序列分析转换成典型的多变量符号回归问题,因此,GEP 技术也能很好地解决时间序列分析问题。

2.2　GEP 时间序列分析的方法

在时间序列分析中,观测的序列数据有一定的间隔(相当于时间间隔),称做采样时间,如一年、一天等。时间序列预测的思想就是通过历史的观测数据确定将来的观测数据,即寻找的一个预测模型,它是一定数量的历史观测值的函数。模型中使用的历史观测值的个数在时间序列分析中称做嵌入维数 d。嵌入维数越高,则利用的历史数据越多,模型的复杂度越高。

时间序列分析中另一个参数称做延迟时间 τ,表明观测数据跳过的个数。延迟时间决定了数据是怎样处理的,如延迟时间为 1,则数据处理是连续的,τ 值越大,表明跳过的观测值越多。

形式上,假设嵌入维数为 d,延迟时间为 τ,则时间序列分析可表示为式(2),序列分析的任务就是寻找映射规则 f。

$$y_t = f(y_{t-\tau}, y_{t-\tau-1}, y_{t-\tau-2}, \cdots, y_{t-\tau-d+1}) \tag{2}$$

将一维的时间序列按照式(2)的嵌入维数和延迟函数值转换成平面表示的形式,这样,时间序列分析转换成一个具有 d 个自变量的传统符号回归问题。

时间序列预测任务就是建立模型,然后利用模型对未来进行预测,模型在运用前需要通过测试来评价模型的质量。一般地,选择时间序列前 90% 的数据进行模型建立,用余下 10% 的数据对模型进行检验和评价,如可采用预测精度对其评价等。

3　实验和性能分析

3.1　实验数据和环境

数据来自国家统计局网站的私人汽车拥有量(http://

www. stats. gov. cn/tjsj/ndsj/),选 1990—2007 年各年全国私人汽车拥有量的数据。实验中,设嵌入维数为 3,延迟时间设为 1,则 18 个序列的数据转换成了 15 行 4 列的数据矩阵。使用前 13 个数据建立预测模型,后面 2 个数据作检验,检验使用相对误差和如式(3)的 χ-方检验标准。

$$R_i = \frac{n\sum_{j=1}^{n}(T_jP_{ij}) - (\sum_{j=1}^{n}T_j)(\sum_{j=1}^{n}P_{ij})}{\sqrt{[\sum_{j=1}^{n}T_j^2 - (\sum_{j=1}^{n}T_j)^2][\sum_{j=1}^{n}P_{ij}^2 - (\sum_{j=1}^{n}P_{ij})^2]}} \quad (3)$$

其中:T_j 是第 j 个目标值,P_{ij} 是第 i 个程序对第 j 个样例的预测值。

程序用 Eclipse3. 1 + Java 编写。为了挖掘满足一定精度尽可能简单的模型,GEP 各参数设定如下:种群规模 30,头部长度为 4,基因个数 2,连接函数为“ + ”,$FS = \{+, -, *, /\}$,$TS = \{a, b, c\}$ 分别表示预测点前 1 ~3 个观测年份数据,其他参数值如表 1 所示。

适应度函数选带有选择带宽的相对误差构建的适应度函数,如式(4)所示。

$$f_i = \sum_{j=1}^{n}\left(R - \left|\frac{P_{ij} - T_j}{T_j} \times 100\right|\right) \quad (4)$$

其中:R 是选择范围,取 $R = 100$;f_i 是第 i 个个体的适应度函数值,可见,最大适应度函数值 $f_{max} = 1\ 300$。其他记号含义同式(3),精度为 10%,进化代数为 5 000 代。

3.2　实验分析

根据以上参数设定,经过多次运行,均能得到较好的预测模型。其中一次的进化模型见式(5)。

$$y_t = y_{t-1} + \frac{2y_{t-1}}{y_{t-3}} + \frac{y_{t-3}^2}{y_{t-1} + y_{t-2}} \quad (5)$$

用该模型对 2006 年和 2007 年全国私车拥有量进行预测检验。其模型值、实际值以及相对误差如表 2 所示。

表2　模型检验

年份	实际值	模型值	相对误差/%
2006	2 333. 32	2 297. 54	2. 607
2007	2 876. 22	2 864. 49	0. 195

在测试数据上,模型的适应度值达到了 197. 20,χ-方值为 1,相对平方误差根(RRSE)为 0. 159。在训练数据上,模型的适应度函数值为 1 264. 82,χ-方值为 0. 998 7,相对平方误差根为 0. 038。可见,模型有很高的精度和良好的泛化能力。应该注意到,由于是通过数据挖掘技术获得的模型,在挖掘中没有考虑模型的可理解性因素,所以式(5)看起来简单、精确,但可理解性不强。

图 4 是在一个坐标轴上展示该模型预测值与实际值各年的情况。

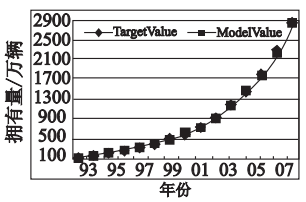


图4　模型预测值与实际值比较图

3.3　模型利用

实验验证,本文得到的模型是可靠的,有很高的准确性,故本节试用该方法得到的模型作短期预测,分别对 2008 年和 2009 年的私车拥有量进行预测。

将训练数据和测试数据合并,形成新的训练数据,依据以上参数进化模型。需要说明的是,对 2009 年总量预测时,使用的是 2008 年的模型值,而非实际值。

进化得到的模型对 2008 年预测值为 3 544. 26 万辆,2009 年为 4 350. 19 万辆,预测建立在可信的模型上。对预测结果应合理分析后再加以利用,以便为决策和管理服务。毕竟社会系统是一个开放的大系统,系统间各个因素相互制约相互影响,纷繁复杂。

4　结束语

GEP 是遗传算法家族的新成员,在符号回归领域有很强的能力,已经取得很多研究成果。时间序列分析是一类特殊的符号回归问题,本文基于 GEP 技术对我国私人汽车拥有量进行了时间序列分析,建立了可靠的预测模型。该方法克服了时间序列分析中使用机理分析得到系统动力学模型的困难,或者主观地设定系统的模型形式,然后进行参数确定方法的人为因素色彩。通过控制嵌入维数和函数符号集的方式控制模型复杂度和精度之间的矛盾。模型在测试数据集上的相对平方误差根为 0. 159,在训练集上为 0. 038,因此模型有很高的精度和良好的泛化能力。对 2008 年和 2009 年的我国私人汽车拥有量进行了预测,可帮助汽车量管理的有关部门进行科学管理和工作部署。下一步将结合领域知识,从理论的高度研究如何将领域知识融入 GEP 的挖掘模型过程中,挖到有趣的模型,增强模型的可理解性。

参考文献:

[1] 朱开永,周圣武,娄可元,等. 基于私家车保有量的预测与调控的灰色模型研究[J]. 中国矿业大学学报,2008,37(6):868-872.

[2] 鄢敏. 城市私人小汽车增长影响因素研究[J]. 交通标准化,2009(5): 142-145.

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

[4] 朱军,唐常杰, 魏大刚,等. 基于动态适应度的基因表达式编程挖掘反函数[J]. 计算机应用研究,2007,25(9):40-42.

[5] 朱明放. 基于基因表达式编程的 TSP 问题求解[J]. 计算机工程与应用,2008,44(23):53-55.

[6] DUAN Lei, TANG Chang-jie, ZHANG Tian-qing, *et al*. Distance guided classification with gene expression programming[C]//Proc of Advanced Data Mining Applications. Heidelberg: Springer-Verlag, 2006: 239-246.

[7] 陈瑜,唐常杰,叶尚玉,等. 基于基因表达式编程的自动聚类方法[J]. 四川大学学报:工程科学版,2007,39(6): 107-112.

[8] 曾涛,唐常杰,朱明放,等. 基于人工免疫和基因表达式编程的多维复杂关联规则挖掘方法[J]. 四川大学学报:工程科学版,2006,38(5):136-142.

[9] ZUO Jie, TANG Chang-jie, LI Chuan, *et al*. Time series prediction based on gene expression programming[C]//Proc of International Conference for Web Information Age. Heidelberg: Springer-Verlag, 2004:239-246.

[10] FERREIRA C. Gene expression programming: mathematical modeling by an artificial intelligence [M]. 2nd ed. Heidelberg: Springer-Verlag, 2006.