文章编号：1005-8451（2009）06-0007-03

基于智能计算的数据分析方法的研究与设计

井海明[1]，赵　宁[1]，兰海波[2]
（1.石家庄铁道学院　计算机与信息工程分院，石家庄　050043；
2.石家庄铁道学院　成人教育学院，石家庄　050043）

摘　要：研究智能演化技术问题和数据分析问题，结合程序设计自动化和离散性数据，运用GEP算法的知识编程，处理太阳黑子，降水量等数据。通过遗传程序设计描绘数据规律并预测数据发展趋势，实现数据的准确拟合。

关键词：智能演化；程序设计自动化；离散性数据；GEP算法
中图分类号：TP3　　**文献标识码：**A

Research and design of data analysis method based on intelligence computation

JING Hai-ming[1], ZHAO Ning[1], LAN Hai-bo[2]
(1.School of Computing and Informatics, Shijiazhuang Railway Institute, Shijiazhuang 050043, China
2.School of Adult Education, Shijiazhuang Railway Institute, Shijiazhuang 050043, China)

Abstract: In this paper, it was mainly studied on the problem of the intelligence evolution technology and data analysis. Combined with the programming design of automaticly and discrete data, it was use the knowledge programe of GEP arithmetic, to process the data of macula and rainfall,etc.It was described the rule of data and forecast the data' s developing trend to implement the data fitting correctly.

Key words: intelligence evolution; procedure design; automation discrete data; GEP algorithm

自计算机出现以来，计算机科学的一个重要目标是让计算机自动进行程序设计，即只要明确地告诉计算机要解决的问题，而不需要告诉它如何去做。遗传程序设计（Genetic Programming）便是在该领域的一种尝试。

收稿日期：2008-12-03
基金项目：河北省教育厅基金资助项目（739001）
作者简介：井海明，讲师；赵　宁，讲师。

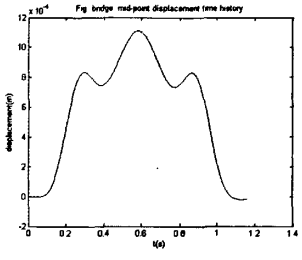

图6　桥梁跨中位移时程曲线

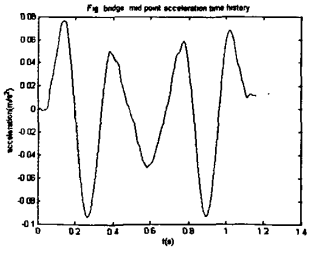

图7　桥梁跨中加速度时程曲线

向振动方程，采用Newmarkβ法解此运动方程，通过MATLAB编程实现，计算结果可靠，利用此模型能够计算移动荷载作用下桥上无砟轨道的动力响应，为桥上铺设无砟轨道提供理论依据。

参考文献：

[1] 曾庆元. 弹性系统动力学总势能不变值原理[J].华中理工大学学报，2000，28（1）：1-3.
[2] 刘晶波，杜修力. 结构动力学[M]. 北京：机械工业出版社，2005.
[3] 娄　平，曾庆元. 移动荷载作用下板式轨道的有限元分析[J]. 交通运输工程学报，2004（1）：29-33.
[4] 娄　平，曾庆元. 车辆－轨道－桥梁系统竖向运动方程的建立[J]. 铁道学报，2004（5）：71-80.
[5] 蔡成标.高速铁路列车－线路－桥梁耦合振动理论及应用研究[D]. 西南交通大学，2004.

基因表达式编程GEP（Genetic Expression Programming）是遗传计算家族的新成员，具有极强的函数发现能力和很高的效率。GEP作为一种基于基因组和表现型组的新的遗传算法(线性或非线性)，是Candida Ferreira在遗传算法（Genetic Algorithm,GA）和遗传编程（Genetic Programming,GP）的基础上发展的新概念，首批研究成果于2000年12月在互联网上发表，2001年12月正式发表[2]。

1 GEP算法概述

GEP的编码方式是先将个体编码为固定长度的线性串，待进行优化求解时再对操作对象进行编码形成基因组。个体在GEP中又称为染色体[2]，染色体是由基因通过连接运算符连接组成。Gene由头和尾组成，head包含了变量集中的变量和函数集FunctionSet中的函数，而tail只包含了变量。头和尾的长度关系为：

$|tail|=|head| \cdot (n-1)+1$

其中，n是此Gene含有的函数集中所有函数的最大目数，例如：如果Gene中有FunctionSet{+，-，*，/}，那么n=2，因为在FunctionSet中函数最大的目数就是2。Gene的长度和染色体包含的Gene个数都可以指定，一旦指定将保持不变。

GEP的编码规则可描述为，将表达式根据其语义表示为表达式树，然后从上到下，由左至右按层次遍历ET中的节点，得到的符号序列即为基因编码的有效部分。若干染色体构成整个种群。GEP模拟自然界的生物进化，按照“物竞天择，适者生存”的原则对种群实施若干次的选择、交叉、变异等基因操作，使种群一代代地进化，从中寻求出最优的个体，从而得到问题的最终解。

算法中产生初始种群，包含若干个代表不同解答方案的“个体”。选择算子作用于种群，根据达尔文的“适者生存”原则，让高适应度的个体有更高的机会生存。遗传算子作用于种群，在种群的个体之间进行遗传动作，产生出新的子代个体。在GA中遗传算子包含重组，变异两大类，重组指多个个体之间进行的遗传操作，子代个体中包含多个父代个体的遗传信息。变异则指单个个体内部进行的遗传操作，子代个体仅包含单个父代个体的遗传信息。在CEP中，除了和GA中类似的单点重组，双点重组，单点变异等以外，CEP中还包括插串（Insert Sequence, IS），根插串（Root Insert Sequence,RIS）等具有独特动作和含义的遗传算子。这些遗传算子作用在GEP特殊的遗传编码上，形成了具有特色的CEP。此类进化算法是由初始种群，各类进化的遗传操作（包括IS、RIS变换、交叉、变异、附加域变异和随机常量变异在内的共6种）、适应度判断、选择操作、收敛性判断这几个步骤构成的。加上用户设置函数集合和各类操作的参数集两个功能，整个程序就是由7个大的子块构成。

2 算法选用的依据

2.1 GEP算法与GP算法的比较

从大量看似无规则的数据中挖掘出函数关系并用于进行预测是数据挖掘的一个重要研究方向，传统的回归预测模型一般假定函数类型已知，然后借助数学方法（如最小二乘法）进行参数估计，进而确定函数表达式。决定函数类型是关键步骤，往往依赖于经验或领域知识，含有主观和盲目因素，目前不能解决复杂函数关系式和多分段函数关系式的建立。为此，李敏强等使用遗传编程（GeneticProgramming，简称GP）方法对太阳黑子时间序列进行建模预测（国际统计界一个著名的例子），并把结果与传统的几种方法进行比较，通过还原性检验，误差明显小于其它方法，并且避免了传统算法建模时事先选定函数类型的盲目性[5]。由于GP在搜索函数模型时的效率原因，该方法并没有被广泛应用于智能模型库的建立，进而融合进决策支持系统中。而CandidaFerreira提出的基因表达式编程（简称GEP）可以和GP一样实现函数表达式的挖掘。

（1）GEP采用的是符号串作为遗传编码，而GP采用树形结构作为遗传编码。从性能上来说，显然操纵符号串比操纵树形数据结构要快，例如，对于通常的单点变异操作，GEP只需要随机产生一个变异点，然后将变异点位置的符号随机变成另外一个符号；而GP则需要随机找出编码树中的一个节点，然后将以该节点为根的子树删除，再从该节点开始，重新生长出一棵子树。这两个操作的

性能差异是巨大的。

（2）对于绝大多数实际问题，GEP和GP面临的问题解空间都是无限大的。在GP中如果不采取相应措施，在进化过程中，树形结构将很容易变得非常巨大。非常巨大的树形结构不但降低了遗传算子的操作速度，更重要的是，使得搜索空间变得很大，效率急剧下降。这种现象称为代码膨胀问题。

为了能在有效的时间内搜索出有效的解，必须对搜索范围进行一定的限制。在GP中，通常都是限定编码树的生长高度不得超过指定的值，或者编码树的节点数不得超过一定的数量。在GEP中，则是对K-表达式指定了头部长度，相应的，尾部最大长度也随之固定，进而，整个遗传编码符号串就是一个定长的编码。显然，在进化过程中，GEP是不需要随时对编码进行调整，就能够保持解树的复杂度不超过一定的值。

2.2 GP对相关限制的解决机制

在GP中，各种遗传算子直接对解树进行操纵，极有可能使得新产生的解树超过指定的复杂度（高度超过范围或者节点数超过范围）；

解决的方法有两种。

（1）淘汰法，直接将超过指定复杂度的解树进行淘汰，重新进行遗传操作产生新的子代个体。

（2）罚函数法，对于超过指定复杂度的解树，在进行评价的时候增加一条罚函数，使得其适应度降低。但是，这两种方法都有很大的局限性。

实践证明，当进化进行到一定代数之后，种群中的解树复杂度一般都趋于比较接近指定的上限，因为通常来说，比较复杂一点的解树更容易且更精确地逼近问题的真实解，适应度也比较高。在这种情况下，淘汰法的进化效率将急剧下降，因为，新产生的子代个体，其复杂度很容易就超过指定的上限。而对于罚函数法，则会在大量的子代个体中引入罚函数，这使得个体的适应度在很大程度上偏离问题求解的本意，导致搜索效率也急剧下降。

3 算法仿真

我们选用两组太阳黑子数据作为测试用例，算法设置如下：运行代数3 000，种群大小3，基因个数3，头部长度6，单点交叉概率0.6，两点交叉概率0.6，变异概率0.03，IS变换概率0.1，RIS变换概率0.1,基因转化概率0.01。运算符选取加法，减法，乘法，除法，自然对数，自然指数，开方，幂指数，正弦，余弦，正切；第1组曲线拟和方程为Y=temp01+tan((exp(temp11).*temp12))+tan(temp21)，第2组曲线拟和方程为Y=tan((tan(cos(a)).*tan(temp01)))+e+tan(tan(sin(sqrt(abs((e./d))))))，其中，abcde代表x1~x5，实线为实际值，虚线为预测值。

图1　拟合曲线1

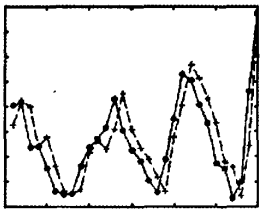

图2　拟合曲线2

4 结束语

测试过程中随机性比较大，同一个样本数据，在选取相同的测试参数的情况下所得到的拟和方程也可能不同。两次测试所得到的曲线的预测值与实际值之间的误差完全在可接受的范围内，并且所选样本种群越大预测值与实际值更接近，但出现非法个体的概率也会相应增大。虽然程序预测的数据的适应率基本在0.7～0.8之间，并不是很高，但已经大致实现了GEP算法功能。通过自动化程序设计，可使用的有图形界面的程序，实现算法功能，为进一步研究智能计算打下坚实的基础。

参考文献：

[1] 康立山，曹宏庆，陈毓屏.常微分方程组的演化建模[J].计算机学报，1999 (8).

[2] 黄　炎，蒋　培，王嘉松，杨敬安.基于可调变异算子求解遗传算法的欺骗问题[J]. 软件学报，1999，10 (2) 216-219.

[3] 曹宏庆，康立山. 动态系统的演化建模[J]. 计算机研究与发展，1999，36 (8).