# 小生境技术在遗传规划中的应用①

刘国炜，常新功

(山西财经大学 信息与管理学院，太原 030006)

**摘 要**：为了提高遗传规划算法的性能，把遗传算法中的小生境技术运用到遗传规划中，提出了改进的遗传规划算法(NGP)。该算法首先对原始训练集进行数据拟合，然后应用小生境技术跟踪拟合函数的极值点，并根据拟合函数的维数的不同，分别计算极值点在自变量维上的欧氏距离并排序，选取欧式距离较大且数量不超过原始训练集 10%的极值点，加入到原始训练集中作为新的训练集，最后用遗传规划算法处理新训练集。在符号回归实验中对 NGP 的准确率进行了测试，说明了该算法的准确性和有效性。

**关键词**：小生境；遗传规划；极值点；符号回归；效率

## Application of Niche Technology to Genetic Programming

LIU Guo-Wei, CHANG Xin-Gong

(Department of Information Management, Shanxi University of Finance & Economics, Taiyuan 030006, China)

**Abstract:** To improve the performance of genetic programming algorithm, the niche technology used in genetic algorithm is applied to genetic programming. It is improvement of genetic programming algorithm, which is called NGP in the next text. First, the algorithm fits the data with the original training set. Second, it tracks the extreme points of the fitting function, and according to the dimensions of the fitting function, calculate the extreme points' Euclidean distance in the independent variable dimension and order it. Then it selects the extreme points whose Euclidean distance is larger and does not exceed the number of the ten percent of the original training set, and added them to the original training set as new training set. Finally, it deals with new training set using genetic programming. In this paper, we use symbolic regression experiment to test the accuracy of the NGP. It illustrates the accuracy and effectiveness of the algorithm.

**Keywords:** Niche; genetic programming; extreme points; symbolic regression; efficiency

## 1 引言

科学实验涉及大量的数据，为了在错综复杂的数据中找到内在的规律就要用到数据拟合的方法。传统的线性回归和非线性回归等需事先指定拟合公式的形式,而且容易陷入局部最优解。采用遗传规划的方法的优点在于不需要给定具体的函数形式，并且在初始群体足够大而且参数设置合理的情况下，可以找到全局最优解，因而在符号回归中有很好的应用。为了提高遗传规划在数据拟合中的精度，厦门大学邵桂芳，周绮凤等人在数据拟合中提出了大规模突变和子树突变的方法[1]。同济大学王站权等对遗传规划初始种群的生成方法进行了改进，提出了最优生成法和平均生成法[2]。在遗传规划寻求自身改进的基础上，本文把遗传算法中的小生境技术[3]运用到遗传规划算法中，改进遗传规划算法的性能。

## 2 相关理论

### 2.1 小生境技术

小生境淘汰运算基本思想是:首先初始化小生境之间的距离参数 L，然后比较种群中各个个体之间的欧氏距离,若他们的欧式距离小于 L，则比较两者之间的适应度大小，并对其中适应度较低的个体施加一个

① 基金项目:山西省高校科技研究与开发项目(20081023);山西自然基金 (2010011022-1).
收稿时间:2010-05-17;收到修改稿时间:2010-06-23

较强的罚函数[4]，极大的降低其适应度。这样,使得在预先指定的距离 L 内只存在一个优良个体并且各个个体之间保持一定的距离，从而避免了算法陷入局部最优解，并使个体能够在整个约束空间中分散开来。

### 2.2 遗传规划

遗传规划的基本思想是:随机产生一个适用于所给问题环境的初始种群，种群中的每个个体表示为树，计算每个个体的适应值;依据优胜劣汰的原则，选择遗传算子对种群进行迭代进化，直到满足终止条件。

## 3 基于小生境技术的遗传规划算法

传统的遗传规划算法在解决预测和分类问题时，先使用给定的训练数据集进行训练，得出训练结果后，用于测试数据集的预测和分类。但给定的训练数据通常是随机数据，所以在利用遗传规划算法进行训练时，无法准确的得出准确的训练结果。所以，本文提出了基于小生境技术的遗传规划算法(NGP)，NGP 算法的思想是：在给定训练数据集进行训练时，首先根据最小二乘法求出拟合函数，然后运用小生境技术跟踪拟合函数的极值点，当极值点的数量小于 10%的训练数据集时，将极值点全部加入到训练数据中；当极值点的数量大于训练数据集的 10%时，根据拟合函数的维数，计算所有极值点在自变量维上的欧氏距离。当拟合函数在二维空间时，计算相邻极值点在自变量维上的欧式距离，然后按欧氏距离的大小进行排序，选取数量为 10%原始训练数据集且欧氏距离最大极值点加入到原始训练数据集中；当拟合函数在三维或三维以上空间时，计算所有极值点之间的欧式距离，选择欧式距离最大的两个点加入原始训练集，然后比较其它极值点到这两个点之间的欧式距离，选择欧式距离较大的点加入原始训练集，接着比较其它极值点到这三个点的欧式距离，选择欧式距离最大的点加入，依次类推，直到加入的极值点的数量达到原始数据集数量的 10%，最后利用遗传算法对训练数据集进行训练，得出训练结果。

NGP 的算法的伪代码：

1)对原始数据集进行拟合，求出拟合函数 F(x)；

2)确定小生境算法的输入及控制参数：主要包括:小生境距离 L、交叉概率 $P_c$、变异概率 $P_m$、进化代数和罚函数等；

3)对求出的拟合函数运用小生境技术，跟踪进化过程中求出的所有的极大值点和极小值点。对求出的极值点进行如下操作：

① 当极值点的数量小于 10%的原始数据集时，将极值点全部加入到训练数据中；

② 当极值点的数量大于训练数据集的 10%，且拟合函数在二维空间时，利用公式(1)计算相邻极值点在自变量维上的欧氏距离，

$$d=\sqrt{\sum_{i=1}^{n}(x_{i1}-x_{i2})^2} \tag{1}$$

$x_{i1}$ 表示第一个点的第 i 维坐标，$x_{i2}$ 表示第二个点的第 i 维坐标。然后按欧氏距离的大小进行排序，选取数量为 10%原始训练数据集且欧氏距离最大的极值点加入到原始数据集中。

③ 当极值点的数量大于训练数据集的 10%，且拟合函数在三维或三维以上空间时，执行以下算法：

i. 计算所有极值点在自变量维上的欧式距离，选择最大欧式距离对应的两个点加入到原始训练集，置点数 n=2；

ii. 若 n 大于 10%原始训练集结束,否则转到iii；

iii. 比较其它 n-2 个点到这两个点之间的欧式距离，选择欧式距离最大的边对应的点加入到原始训练集中；

iv. n= n+1 转 ii；

4) 确定遗传规划中的参数，包括函数集、终止符集、适应度函数、群体大小、迭代次数、交叉概率 Pc、变异概率 Pm、终止准则等；

5)随机产生一个初始种群，计算每个个体的适应度；

6)运用遗传规划算法对处理后的数据集进行训练，如果达到最大进化代数或满足终止准则，则停止；否则转 7)；

7)对种群个体进行遗传操作：

① 对群体中适应度较高的个体进行复制；

② 根据交叉概率 $P_c$，重新组合两个个体的任意选定部分；

③ 根据变异概率 $P_m$，变异选定个体的选定部分；转 6)。

## 4 NGP在符号回归中的应用

为了证明 NGP 算法的有效性，本文在符号回归实验中对其进行了测试。

### 4.1 符号回归

符号回归[5]旨在用给定的精度拟合一组相关变量的有限样本，找出一个符号形式的数学公式。传统的线性回归、非线性回归等需事先指定拟合公式的形式，通过回归确定拟合公式中的待定参数，而符号回归无需事先指定拟合公式的形式，公式的形式和其中的参数均在回归的过程中确定。因此，符号回归具有更广泛的应用范围，如数学规律的经验发现和符号方程求解等。

### 4.2 极值点的跟踪

首先给定测试目标函数：

$$y = \sin(x) \times e^{-0.1 \times x}, x \in [0,10\pi] \quad (2)$$

实验中设置交换概率 $P_c$ 为 0.9，变异概率 $P_m$ 为 0.1，种群规模为 100，最大进化代数为 100，小生境距离参数 L 设为 5，罚函数设为 10-30，通过实验，跟踪全部的极值点，计算极值点自变量之间的欧式距离，选取欧氏距离最大且数量为 10%原始数据集的数据，经过处理最后得到表 1 中的极值点。

表 1 处理后的极值点

| 极小值点 | | 极大值点 | |
|---|---|---|---|
| $x$ | $y$ | $x$ | $y$ |
| 4.54 | -0.625700000 | 1.29 | 0.844500000 |
| 10.82 | -0.333700000 | 7.79 | 0.457900000 |
| 17.26 | -0.178000000 | 14.10 | 0.244400000 |
| 23.42 | -0.095170000 | 20.29 | 0.130400000 |
| 29.72 | -0.050800000 | 26.58 | 0.069550000 |

### 4.3 极值点对数据拟合准确率的影响

遗传算法的适应度函数关系式为：

$$r(i,t) = \sum_{j=1}^{M} |S(i,j) - C(j)| \quad (3)$$

r(i,t)为第 t 代群体中个体 i 的适应度,S(i,j)为个体 i 在适应度计算案例 j 的返回值,M 为训练样本个数,C(j)为适应度计算案例 j 的实测值或正确值。显然，适应度值越小，个体越优良。此外，GP 在运行过程中需要设置的其它参数包括：进化代数、种群规模等，见表 2：

表 2 参数设置

| 项目 | 参数设置 |
|---|---|
| 函数运算符集 | $F$={+, -, *, $sin$, $cos$} |
| 种群规模 | 50 |
| 终止结点 | { $x$,$C$}，$C$ 为常数项 |
| 进化代数 | 25 |
| 终止条件 | 达到最大进化代数或满足终止条件 |
| 遗传算子概率 | 交换概率 $P_c$ 为 0.5，变异概率 $P_m$ 为 0.5 |

为了排除实验过程中随机因素造成的影响，验证极值点对回归的影响程度，我们随机产生一组训练数据，规模为 100，并进行 10 训练，选出测试结果最好的训练结果。通过比较，测试适应度最小值为 0.67434。适应度的变化情况如图 1，在模型操作的整个过程中，初始群体的适应度很高，为 8.638324；随着进化代数的增加，函数关系模型对训练样本的适应度值迅速下降，且下降的速度在开始的几代尤其快，当进化到第 5 代时，最优个体的适应度已下降到 6.438213；然后适应度下降的速度逐渐减慢，当到第 12 代时，最优个体的适应度下降为 6.102920。根据适应度值越小，个体越优良的原则，当进化 25 代后，具有最高适值的个体。

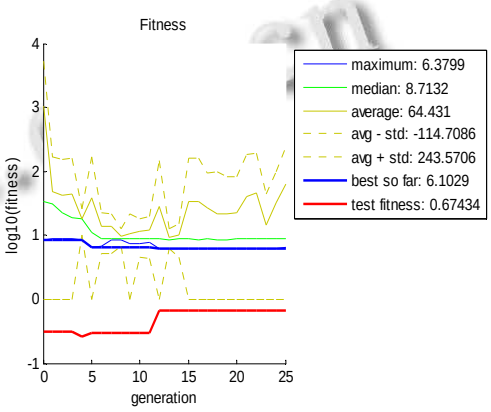


图 1 适应度变化情况

然后，用处理后得到的极值点随机代替原始数据中 10%的数据，同样进行 10 训练，选出测试适应度最小的训练结果，得到的最小测试适应度为 0.5302。图 2 为适应度的变化情况，在模型操作的整个过程中，初始群体的适应度很高，为 9.848537；随着进化代数的增加，函数关系模型对训练样本的适应度值迅速下

降，且下降的速度在开始的几代尤其快，当进化到第 3 代时，最优个体的适应度已下降到 7.042718；然后适应度下降的速度逐渐减慢，当到第 19 代时，最优个体的适应度下降为 6.774866。

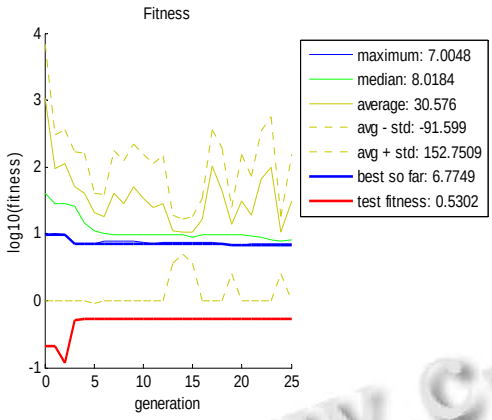


图 2　适应度变化情况

通过计算我们得出加入极值点后回归适应度的平均值为 7.00225，加入前回归适应度平均值为 6.22908；加入极值点后测试适应度平均值为 0.707849，加入前测试适应度平均值为 0.953097。通过比较，可以明显看出加入极值点后的回归适应度高于原始数据的适应度，但是测试适应度小于加入前，说明加入极值点后测试数据拟合的效果增强。为了排除数据集分布情况对实验结果的影响，试验中采用了随机替换 10%的原始数据集的方法。公式(3)中 C(j)通常是对原始数据集进行数据拟合后得到的近似函数，通过运用小生境技术跟踪 C(j)极值点，说明了极值点在数据拟合中的重要性。在而加入极值点后的训练集训练适应度的增加，是因为有随机因素的影响，如训练集分布情况的改变。

利用加入极值点前后回归得到的函数，对测试数据集进行测试，比较结果如图 3。然后利用公式：

$$e=\sqrt{\sum_{i=1}^{n}(y_i-y_0)} \qquad (4)$$

$y_i$ 是 $x_i$ 代入回归函数后计算得到的结果，$y_0$ 是测试集的原始值。进行累计误差 e 的计算。通过计算，得到加入极值点后回归函数对测试集测试的累计误差为 0.041063，加入前的累积误差为 0.063873。说明极值点的引入对回归效果显著。

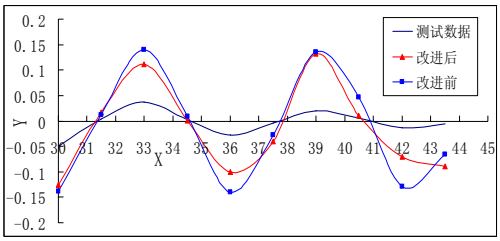


图 3　实验对比

### 4.4 NGP 算法对数据拟合准确率的提高

选取小生境算法求出的极值点代替原始训练集中的数据时，由于随机性因素，可能造成误差的增大。而且根据常识，当选择的训练数据集服从均匀分布是，图像的拟合结果为最优。为了减少误差因素造成的影响和排除训练数据的分布情况造成的影响，确保原始数据的真实性，选出的大小为原始训练集 10%的极值点(当极值点的个数小于原始训练集 10%时，则全部加入原始训练集中)，直接加入原始测试集中进行 10 次测试，选取测试适应度最小的训练结果，测试适应度最小值为 0.066829。通过计算得出 NGP 算法处理后回归适应度的平均值为 8.16405，传统遗传规划算法处理后的回归适应度平均值为 6.22908；NGP 算法处理后测试适应度平均值为 0.764873，传统遗传规划算法处理后的测试适应度平均值为 0.953097。通过比较，可以看出 NGP 算法在数据拟合中的精度有显著提高。

## 5　结论

小生境技术在寻优过程中，对目标函数没有可导的限制，在多元不可导函数中有很好的应用。为了克服遗传规划中偶然因素的影响，本文通过做 10 次试验，分别对平均适应度和最好适应度进行了比较。通过测试 NGP 算法在一元函数拟合中的应用，可以发现极值点的引入虽然带来了误差，使训练过程的适应度增加，但在测试中，却发现测试适应度明显变小，大大的提高了数据的拟合效果。NGP 算法在拟合多元函数中有更好效果，如函数

$$y=2500-(x_1^2+x_2-11)^2-(x_1+x_2^2-7)^2, x_1,x_2\in[-6,6] \qquad (5)$$

用 NGP 算法处理后，找到函数的四个极值点，如图 4 所示，可以更准的定位图像，提高数据拟合的精度。

(下转第 156 页)

降低。随着数据量不断增大，两种算法执行效率差距也越来越明显。GTSClu 算法比 CluStream 算法效率更高。

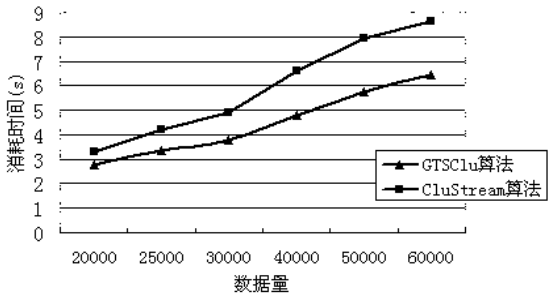


图 6　算法执行效率比较

## 5　结论与展望

GTSClu 算法分在线处理和离线聚类两部分，网格和最小生成树技术应用到聚类中。在线过程利用均匀网格对数据空间进行划分，以网格单元为单位得到数据流的概要信息及部分数据具体信息。有效提高了处理数据对象效率。离线聚类中将相关网格单元进行拆分，均匀网格转化为不均匀网格，提高聚类精度，使算法发现任意形状聚类能力更加完善，能有效排除噪声和孤立点数据对聚类影响，提高聚类质量。GTSClu 存在以下不足：一是聚类过程中网格划分参数 m 的取值会影响聚类结果；二是离线聚类过程中最小生成树构造过程待简化。下一步研究工作针对这两方面完善算法。

## 参考文献

1 Muthukrishnan S, Shah R, Vitter J. Mining Deviants in Time Series Data Streams. Proc. of the 16th International Conference on Scientific and Statistical Database Management (SSDM'04). Santorini Island, Greece, 2004. 41－50.

2 Guha S, Mishra N, Motwani R, Ocallaghan L. Clustering data Streams. Proc. of the 2000 Annual Symp. on Foundations of Computer Science. 2000. 359－366.

3 Han J, Kamber M. Data Mining: Concepts and Techniques (Second Edition). Morgan Kaufmann, Elsevier Inc, 2006. 467－589.

4 Yang YD, Sun ZH, Zhang J. Finding outliers in distributed data streams based on kernel density estimation. Computer Research and Development, 2005,42(9):1498－1504.

5 Aggarwal C, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. Proc. of 29th International Conference on Very Large Databases (VLDB'03). Berlin, Germany, 2003. 81－92.

6 邱保志,沈钧毅.基于网格技术的高精度聚类算法.计算机工程,2006,32(3):12－13.

7 何勇等.基于动态网格的数据流聚类分析.计算机应用研究,2008,25(11):2－4.

8 严蔚敏,吴伟民.数据结构.北京:清华大学出版社,1997.173－175.

9 Hsu CM, Chen MS. Subspace clustering of high dimensional spatial data with noises. Heidelberg: Springer, 2004. 31－40.

(上接第 52 页)

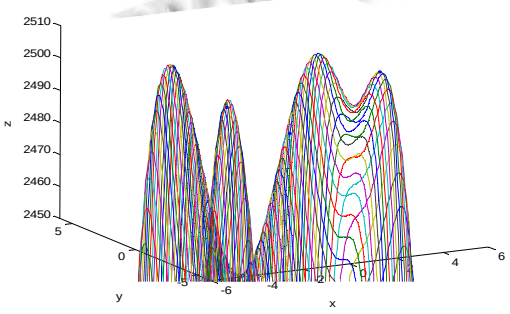


图 4　小生境技术求得极值点

## 参考文献

1 邵桂芳,周绮凤,陈桂强.基于改进遗传规划算法的数据拟合.计算机应用研究,2009,26(2).

2 王战权,云庆夏,杨东援.改进的遗传规划研究.系统工程理论与实践,2000,(5):66－69.

3 朱筱蓉,张兴华.基于小生境遗传算法的多峰函数全局优化研究.南京工业大学学报,2006,28(3):39－42.

4 袁丽华,黎明,李军华.进化优化小生境遗传算法控制参数的研究.计算机工程,2006,32(13).

5 代杰,吴军,田社平,颜德田.遗传规划在符号回归中的应用.传感器与微系统,2007,26(11):108－110.