加权最小二乘法改进遗传克里金插值方法研究

严华雯,吴健平
(华东师范大学 地理信息科学教育部重点实验室,上海 200062)

摘 要:数据内插被广泛应用于地统计分析领域,克里金插值作为其中最为有效的方法之一,其原理是通过建立变异函数理论模型,得到可靠的权重值和拉格朗日系数,构成求解待测点的线性组合。为了有效地提高插值精度,文中利用加权最小二乘法优化遗传算法中的适应度函数,进而改进普通基于遗传算法优化的克里金插值方法。并且在MATLAB中利用外部工具箱确定模型参数,最后通过实例验证,将该方法与普通克里金插值以及遗传克里金插值结果进行对比,发现采用该方法,插值效果较好且误差也较小,证明了通过加权最小二乘法可以有效改进普通遗传克里金插值方法。

关键词:克里金;遗传算法;变异函数;加权最小二乘法
中图分类号:TP311.52 **文献标识码**:A **文章编号**:1673-629X(2012)03-0092-04

Research on Genetic Algorithm Kriging Optimized by Weight Least Square

YAN Hua-wen, WU Jian-ping
(Ministry of Education Key Lab of Geographic Information Science, East China Normal University, Shanghai 200062, China)

Abstract: Data interpolation is widely used in the field of geostatistical analysis. As one of the most effective methods of data interpolation, using Kriging could obtain reliable weight values to build up the linear combination by means of the establishment of semi-variogram model. In order to improve the accuracy of interpolation, weight least square is used in this paper to optimize the fitness function in genetic algorithm, then improve Kriging optimized by genetic algorithm only. Model parameters can be calculated by toolbox in MATLAB. Finally, verified by example data and compared with ordinary Kriging and genetic algorithm Kriging, this method is found to achieve a more interpolation result and get less error. Weight least square is proved to be an effective method to improve ordinary genetic algorithm Kriging.

Key words: Kriging; genetic algorithm; semi-variogram; weight least square

0 引 言

对于空间数据的插值,目前存在着很多内插方法,对于不同的应用,它们各自存在优缺点。克里金插值是其中比较成功的一种数值计算方法,其理论基础是变异函数和结构分析,它考虑的问题不再是单纯的点间距离,而是综合了样点的大小、相互关系和空间分布等几何特征,同时还包括观测点与待测点之间的空间关系。

对于克里金法中变异函数模型的优化,有不少专家学者提出了许多行之有效的解决方案。从优化方法上大致可以分为传统数学统计方法和人工智能算法。在使用传统数理统计法方面,一些专家提出了极大似然法、线性规划法等参与的优化模型[1,2];随着人工智能的发展,近些年来,也有不少学者结合实际情况,提出了相关的一些优化方法[3-5]。

但是,对于传统的数理统计方法而言,往往存在着正负数以及计算复杂等问题;其次,在目前被采用的人工智能算法中,粒子群算法对离散的优化问题求解不佳,并且易于陷入局部最优,而以前所采用的遗传算法优化方法都是通过最小二乘法的原理得到,前提是在值的不同测量范围内测量精度是相同的,但是实际情况却往往不是这样。文中采用加权最小二乘法的原理建立遗传算法目标函数,改进变异函数理论模型精度,并通过实例验证优化了克里金插值方法的实际应用。

收稿日期:2011-07-19;**修回日期**:2011-10-25
基金项目:国家重点基础研究发展计划(973计划)项目(2010CB951603)
作者简介:严华雯(1987-),女,上海人,硕士研究生,CCF会员,研究方向为GIS应用与开发;吴健平,博士,教授,博导,主要研究方向为GIS应用与开发。

1　克里金插值

空间数据通常是一系列采样观测值,分布往往很不规则,但用户在某些情况下需要获知相同区域内未知观测点的数据。根据空间相关性原则,这些未知点与采样点之间存在着空间上的相关性,通常情况下,距离越近的点,其特征值越相似;反之,亦然。因此空间插值算法应运而生,它是通过已知点推求未知点的计算方法[6]。克里金法是一种无偏、线性、最优统计方法[7],克里金插值的核心是变异函数,其定义为

$$\gamma(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}[Z(u_a)-Z(u_a+h)] \tag{1}$$

其中,$N(h)$ 为距离相隔为 h 的点对数,Z 是特征值,$\gamma(h)$ 表示变异函数值。

变异函数的核心思想是按照点对间的距离大小分组,对每一个组中的每一个点对进行插值计算,求算平均值,即可得到变异函数值。在普通克里金插值中,待测点的特征值是根据周围的已知点的线性组合推求得出,可以用以下数学公式表达:

$$\hat{Z}(v_0)=\sum_{i=1}^{n}\lambda_i Z(v_i) \tag{2}$$

$\hat{Z}(v_0)$ 表示待测点 v_0 的估计值;$Z(v_i)$ 表示 v_0 周围观测点 v_i 的特征值;n 为参与计算的观测点的个数;λ_i 表示权重系数。为了得到待测点的估计值,需要求出各个权重系数。

在二阶平稳条件下满足无偏性和最优性,即

$$(1)\ \sum_{i=1}^{n}\lambda_i=1; \tag{3}$$

$$(2)\ \delta^2=E[\hat{Z}(v_0)-Z(v_0)]^2。 \tag{4}$$

经过推导,建立方程组

$$\begin{cases}\sum_{i=1}^{n}\lambda_i=1\\ \sum_{i=1}^{n}\lambda_i\gamma[Z(v_i),Z(v_j)]-\mu=\gamma[Z(v_i),Z(v_0)]\end{cases} \tag{5}$$

$i,j=1,2,\cdots,n$

$\gamma[Z(v_i),Z(v_j)]$ 表示两个观测点 v_i 和 v_j 的变异函数;μ 为拉格朗日系数。

实际上,根据观测点计算得到的变异函数是一系列离散点,于是需要对这些离散点进行模型拟合,使之能够得到任意距离关系的变异函数,常用的理论模型有球面模型、高斯模型和指数模型。于是,通过变异函数拟合得到的曲线,求解方程组,求出权重系数和拉格朗日系数,从而推求出相应位置上待测点的估计值。

2　普通遗传优化克里金插值

2.1　遗传算法

遗传算法是一种宏观意义下的仿生算法,它通过对生物的进化过程中选择、交叉和变异机理的模拟,实现问题最优解的搜索[8]。具体运算过程如图 1 所示。

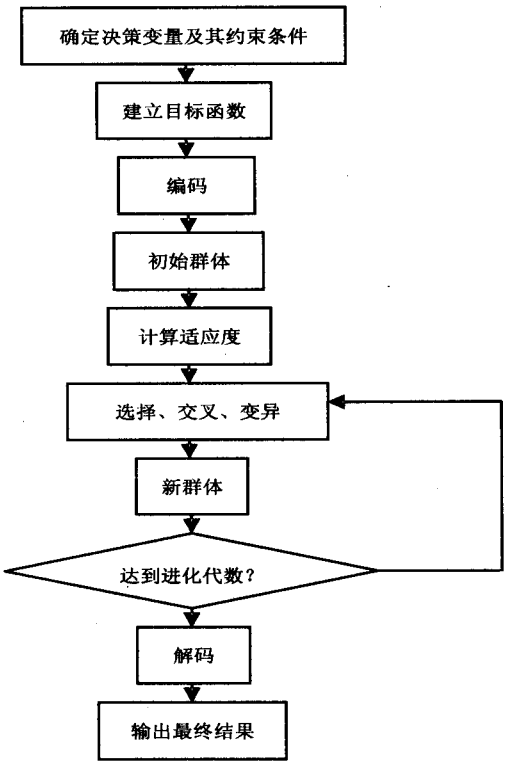

图 1　遗传算法流程图

2.2　遗传优化克里金插值

要想得到最优的插值结果,克里金插值变异函数的模型拟合起着至关重要的作用,模型的参数确定又是构造回归模型的关键。因此,在模型确定的前提下,决定克里金插值准确与否的本质即是模型参数的确定。遗传算法在函数优化方面的应用历史悠久,相比较其他优化方法,它在非线性、多模型、多目标的函数优化问题上可以方便地得出较好的结果。因此,可以采用遗传算法确定克里金插值中的三个模型参数。

根据前述遗传算法的应用步骤,只要根据克里金插值中的变异函数建立目标函数,继而通过一定规则转换为个体适应度函数,然后进行遗传操作,即可得出较好的结果。具体思路:

(1)确定决策变量及其各种约束条件。

对于克里金插值而言,决策变量就是模型参数 c_0,c,a,约束条件可以根据实际插值的特征值确定。模型参数 c_0 为块金效应值,c 为基台值,a 为自相关阈值或变程。

(2)建立优化模型。

考虑到变异函数模型拟合的实质是更准确地体现实际变异函数与理论变异函数的接近程度,因此,可以将各种距离条件下两者之间的差值之和作为目标函数[9]。

$$f=\sum_{i=1}^{m}[\hat{\gamma}(h)-\gamma(h)]^2 \tag{6}$$

(3)确定表示可行解的染色体编码方法。

也即确定出个体的基因型及遗传算法的搜索空间。通常的编码方式是,采用二进制编码串表示三个决策变量,长度根据决策变量的最大取值决定。

(4)确定解码方法。

即确定出个体基因型到个体表现型的对应关系或转换方法。

(5)确定个体适应度的量化评价方法。

即确定出由目标函数值到个体适应度的转换规则。根据实际情况,需要评价个体适应度的最小值,即

$$f_{\min} = \sum_{i=1}^{m} [\hat{\gamma}(h) - \gamma(h)]^2 \tag{7}$$

(6)遗传运算。

遗传算子都是通过对编码串进行相应操作,以达到选择、交叉和变异这些遗传规律上的应用。

(7)确定遗传算法的有关运行参数。

主要是指群体大小和终止进化代数,以及交叉概率和变异概率。

(8)判断当前运行代数与终止进化代数的大小关系,最终停止运算。

根据上述步骤,即可完成对克里金插值变异函数的优化计算。

3 加权最小二乘法改进的遗传克里金插值

遗传优化克里金插值目标函数 $f = \sum_{i=1}^{m} [\hat{\gamma}(h) - \gamma(h)]^2$ 基于最小二乘法原理,但是,最小二乘法存在一个缺点,所有数据的贡献都是均等的,它没有考虑到其中存在着影响最后结果的异常点,使得最后得到的结果不能准确地反映实际情况。

加权最小二乘法在最小二乘法基础上进行改进,有效地克服了该缺点。针对克里金插值中的变异函数,在不同距离内,插值的点对也是不同的,点对越多则说明该处的变异函数值越可信,在模型拟合的时候应该首先考虑该处,因此,可以考虑在上面目标函数的基础上增加权重系数[10]。其依据就是加权最小二乘法,实质是对原始数据实施变换,在平方和中加入一个适当的权重系数 w_i,以调整各项在平方和中的作用,根据分析,此处的权重系数 w_i 可以表示为 $\frac{n(h)}{\hat{\gamma}(h)^2}$。改变后,目标函数可以表示为:

$$f = \sum_{i=1}^{m} \frac{n(h)}{\hat{\gamma}(h)^2} [\hat{\gamma}(h) - \gamma(h)]^2 \tag{8}$$

它表示目标函数值与不同步长下的点对个数成正比,与实际变异函数值和拟合值之差成正比,同时与变异函数的拟合值成反比。相应的,遗传算法中的适应度可以表示为:

$$f_{\min} = \sum_{i=1}^{m} \frac{n(h)}{\hat{\gamma}(h)^2} [\hat{\gamma}(h) - \gamma(h)]^2 \tag{9}$$

4 实验研究

4.1 实验方法

为了验证该方法的有效性,采用 ArcGIS 自带的数据进行验证,该数据为 shapefile 矢量数据文件。考虑到 MATLAB 以矩阵和数组为基础进行计算,语言简便易操作,拥有强大的图形技术,因此使用 MATLAB 作为实际计算的辅助软件[11]。同时,选用 ArcGIS 作为最后插值以及实验对比分析的软件[12]。

实验流程图如图 2 所示。

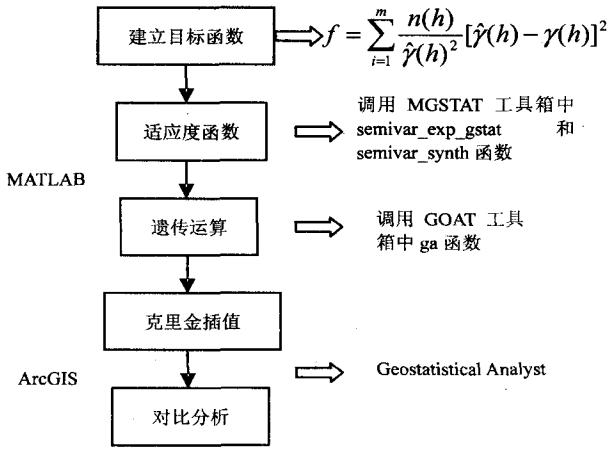

图 2 实验流程图

(1)根据目标函数建立适应度函数。

使用 map 工具箱中的 shapef 函数读取 shapefile 文件,从而获得数据的样点坐标信息 x,y 和插值数据 dval。

采用 MGSTAT 外部工具箱中的 semivar_exp_gstat 和 semivar_synth 函数,两个函数分别计算实际变异函数和拟合好后的变异函数。c_0 表示块金效应,c 表示基台值,a 表示变程。

```
% 计算实际变异函数
[garr,hc,np] = semivar_exp_gstat(pos,val);
```

选择球状模型作为最终的拟合模型,设置块金效应、基台值和变程等相关参数,并得出拟合模型下的经验变异函数,其过程如下:

```
% 变异函数拟合模型
V(1).par1 = c+c0;V(1).par2 = a;V(1).type = 'Sph';
V(2).par1 = c0;V(2).par2 = 0;V(2).type = 'Nug';
% 不同距离下的变异函数拟合值
gamma_synth = semivar_synth(V,h_arr,gstat_format);
```

根据目标函数 $f=\sum_{i=1}^{m}\frac{n(h)}{\hat{\gamma}(h)^2}[\hat{\gamma}(h)-\gamma(h)]^2$ 计算:

```
r=zeros(length(h_arr),1);
for i=1:length(h_arr)
r(i)=np(i)*((gamma_synth(i)-garr(i))/garr(i))^2;
end
objecval=sum(r);
```

(2)遗传运算。

设置遗传算法各运行参数及遗传算子:

```
[x,fval]=ga(@cobjecval,3,options);
```

该方法采用了 MATLAB 自带的 GOAT 工具箱中遗传算法的部分函数 ga, 设定三个参数可以快速求解出最终群体和适应度,第一个参数是指目标函数,这里使用 cobjecval. m 文件;第二个参数 3 表示共有三个变量参与遗传运算;第三个参数 options 则是用于设定运算过程中的运行参数和遗传算子。Options 的设定通过 Gaoptimset 函数来获得。

4.2　实验结果

根据数据的实际情况,初始群体随机产生,群体数为 100,进化代数为 1000, c_0 范围为[0,0.001],c 范围为[0,0.005],a 范围为[800000,1500000]。采用该方法得出的三个参数进行克里金插值,且都选取待测点周围 10 样点作为插值点,最终求得块金效应为 1.4199 * 10-4,偏基台值为 5.2471 * 10-4,变程为 1197300。

对比三种方法得出的插值结果,进行误差检验,发现采用加权最小二乘法的克里金插值法具有较小的误差统计值,对比结果如表 1 所示:

表 1　三种方法插值结果误差统计对比

插值方法	误差均值	误差均方根	平均标准误差
普通克里金插值	0.0003322	0.01377	0.01858
遗传优化克里金插值	0.0003052	0.01309	0.91345
加权最小二乘法遗传优化克里金插值	0.0003032	0.01312	0.01328

5　结束语

文中主要探讨了遗传算法在克里金插值优化方面的用途,并且改进了以变异函数为基础的目标函数,再通过加权最小二乘法进一步优化,并通过对实际数据进行克里金插值,验证该方法在实际应用上的可行性。通过加权最小二乘法改进的目标函数,在经过遗传运算后,能够更好地反映变异函数的实际特性,从而获取最佳的变异函数拟合模型。

参考文献:

[1] 王仁铎,胡光道. 线性地质统计学[M]. 北京:中国地质大学出版社,1986.

[2] 矫希国,刘　超. 变差函数的参数模拟[J]. 物探化探计算技术,1996,18(2):158-161.

[3] 曾怀恩,黄声享. 基于 Kriging 方法的空间数据插值研究[J]. 测绘工程,2007,16(5):5-13.

[4] 张　强,许少华,于文涛,等. 粒子群算法在克里金三维地质建模中的应用[J]. 大庆石油学院学报,2011,35(1):85-89.

[5] Li M, Li G, Azarm S. A Kriging Metamodel Assisted Multi-Objective Genetic Algorithm for Design Optimization[J]. Journal of Mechanical Design, 2008, 130(3): 031401. 1-031401. 10.

[6] 王劲峰,姜成晟,李连发,等. 空间抽样与统计推断[M]. 北京:科学出版社,2009.

[7] 彭兆璇,袁　峰,周涛发,等. 土壤中元素空间分布的体视化方法研究[J]. 计算机技术与发展,2009,19(5):195-197.

[8] 周　明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,1999.

[9] 解　皓. 基于遗传优化克里格法的城镇基准地价评估模型研究[D]. 武汉:武汉大学,2005.

[10] Kushavand B, Aghabaee H, Mohammadzadeh M J. Semivariogram Fitting with a Simple Optimizing Algorithm[J]. Journal of Applied Sciences, 2005, 5(8): 1405-1407.

[11] 张志涌,杨祖樱. MATLAB 教程 R2010a[M]. 北京:北京航空航天大学出版社,2010.

[12] Jonhnston K, Hoef J M V, Krivoruchko K, et al. Using ArcGIS Geostatistical Analyst[M]. [s. l.]: ESRI, 2001.

(上接第 91 页)

学,2009.

[7] 匡鸿博. 高动态 GPS 信号跟踪技术的研究[D]. 上海:上海交通大学,2010.

[8] 付　晓,雷建设. 3G 系统中的定位技术[J]. 电信技术,2005(8):10-11.

[9] 秦　杰,陈　希,武穆清. A-GPS 定位技术的研究与应用[J]. 数字通信世界,2007(3):5-6.

[10] 李华贵,项志华,何　伟,等. 基于 GPS 和 GPRS 车载导航定位系统的实现[J]. 计算机技术与发展,2006,16(11):10-15.

[11] Hoshen J. The GPS equations and the problem of Apollonius [J]. IEEE Trans on Aeros Ele Sys, 1996, 32(3): 1116-1124.

[12] Phatak M. Position fix from three GPS satellites and altitude: a direct method[J]. IEEE Trans on Aeros Ele Sys, 1999, 35(1): 350-354.

[13] Abbott E, Powell D. Land-vehicle navigation using GPS[J]. Proceedings of the IEEE, 1999, 87(1): 145-162.