doi:10.3969/j.issn.1005-3026.2016.03.022

基于多基因遗传规划的矿石品位估计

韩创益[1,2], 王恩德[1], 夏建明[1], 李光秀[2]
(1. 东北大学 资源与土木工程学院, 辽宁 沈阳 110819; 2. 金策工业综合大学 资源勘探工程学院, 平壤 999093)

摘 要: 由于矿床形成过程复杂、控制因素多,导致估计矿石品位相对困难.尽量降低矿床预测中的估计误差对矿产资源的开发和利用是至关重要的.克立格法被认为是最佳的品位估计方法,其必须满足对于品位空间分布的平稳性和内蕴假设.但实践上,大部分的品位数据具有稀疏、不规则而复杂的空间分布,这有时会导致克立格法违反平稳性和内蕴假设.本文提出基于多基因遗传规划的矿石品位估计方法,并将其与克立格法进行对比.结果显示,基于多基因遗传规划的方法不需要关于空间分布的假设.这样,简化了实施矿体品位预测的条件,并能取得较好的预测结果,可应用于复杂矿体品位的预测.

关 键 词: 矿石品位估计;多基因遗传规划;普通克立格;矿床预测;人工智能
中图分类号: P 628 **文献标志码**: A **文章编号**: 1005-3026(2016)03-0408-05

Ore Grade Estimation Based on Multi-gene Genetic Programming

HAN Chang-ik[1,2], *WANG En-de*[1], *XIA Jian-ming*[1], *LI Gwang-su*[2]
(1. School of Resources & Civil Engineering, Northeastern University, Shenyang 110819, China; 2. College of Geoexploration Engineering, Kimchaek University of Technology, Pyongyang 999093, DPRK. Corresponding author: HAN Chang-ik, E-mail: han_6130@sina.com)

Abstract: Ore grade estimation is relatively difficult due to the complexity of ore deposit formation process and numerous control factors. Evaluation of ore deposit with low estimation error is crucial in mineral resources development and usage. So far, Kriging, now known as a best estimation method of grade, is based on intrinsic assumption and stationarity about the underlying grade spatial distribution. However, most of ore grade data are spatially sparse, irregularly spaced and have complex distribution, which could result in the Kriging estimation method violating intrinsic assumption and stationarity. This article presented a new method for ore grade estimation based on multi-gene genetic programming and also compared it with ordinary Kriging. The results show that the proposed method makes no assumptions about the spatial distribution of grade data, the condition of implementing ore body grade prediction is simplified, and it can achieve better prediction effect. So, the proposed method can be used to estimate ore grade for complex ore deposit.

Key words: ore grade estimation; multi-gene genetic programming(MGGP); ordinary Kriging; ore deposit prediction; artificial intelligence

矿石品位估计属于一种空间数据的插值问题,其包括线性和非线性克立格法、距离倒数加权法、多项式回归法和仿样内插法等,其中克立格法被认为最佳的矿石品位估计方法[1-4].克立格法是在平稳性假设下,根据待估非采样点有限邻域内若干已测定的样本点数据,基于变差函数提供的空间结构随机模型,对待估未采样的位置值进行线性无偏最优估计[1-4].但是,实际品位数据经常稀疏、不规则且具有很复杂的混合分布.为了满足这些假设条件,需要大量的关于品位的空间分

收稿日期: 2015-05-22
基金项目: 国家重点基础研究发展计划项目(2012CB416800); 国家自然科学基金资助项目(41372098).
作者简介: 韩创益(1980-),男,朝鲜平壤人,东北大学博士研究生; 王恩德(1957-),男,辽宁盖州人,东北大学教授,博士生导师.

布知识，从而会导致估计方法的复杂化[1,5].

在估计矿石品位中能代替克立格的方法为多基因遗传规划. 与传统的回归分析和其他统计建模技术相比，遗传规划具有无需先假定具体的函数形式即可产出拟合的数学模型的优点[6-9]. 实践证明，各种各样的遗传规划变体，包括多基因遗传规划，在简单或复杂的各工程领域中得到了较为成功的运用[8-9].

本文探讨用多基因遗传规划估计矿石品位的新方法，并与克立格法进行对比.

1　普通克立格法

地质统计学在空间问题求解方面已经得到越来越广泛的应用. 常用于空间预测的一个重要方法就是克立格法，该方法在给出最优线性无偏估值的同时，还可计算出估计值的量化评价指标——估计方差，即克立格方差[2-3].

克立格法是一种基于最优权重的空间估值方法，待估点的变量 Z 的克立格估值是由周围一定区域内采样点的已有数据经线性加权组合得到的. 克立格估值，又称空间局部估计或空间局部插值法，它建立在变差函数理论及结构分析基础上，是在有限区域内对区域化变量的取值进行无偏最优估计的一种方法. 克立格方法包括普通克立格、泛克立格、协同克立格、对数正态克立格、指示克立格与析取克立格等，其中普通克立格为在估计矿石品位中最常用的方法[1-2]. 普通克立格为一种对空间分布数据求最优、线性、无偏内插估计量的方法.

设随机变量(也就是矿石品位)Z 为在样本点 $x_i(i=1,\cdots,n)$ 上已观测到的，而要估计在未采样的位置 x_0 上的品位值 $\hat{Z}(x_0)$，当 $E[Z(x)]=m$ 为常数(满足二阶平稳性至少准平稳或准内蕴假设)时，估计值表达式为

$$\hat{Z}(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i)\ . \tag{1}$$

式中：n 为已知的样本数目；λ_i 为对应的样本权重系数.

上面提到的平稳性假设要求样品满足在空间上的均匀性. 当区域化变量满足平稳性假设或内蕴假设后，被 $\boldsymbol{h}$ 分隔的每一对数据 $\{Z(x),Z(x+\boldsymbol{h})\}$ 都可以看成是随机变量对 $\{Z(x_0),Z(x_0+\boldsymbol{h})\}$ 的不同的实现，可以得到实验变差函数：

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

式中：$Z(x_i+\boldsymbol{h})$ 和 $Z(x_i)$ 为采样点 $x_i+\boldsymbol{h}$ 和 x_i 的实际观测值；$N(\boldsymbol{h})$ 为被向量 $\boldsymbol{h}$ 相隔的观测值数据对的对数. 这样通过平稳性假设或内蕴假设就把地质统计学的区域化理论和观测数据结合起来了. 变差函数主要用于评价在不同位置上数据的空间依赖性.

采用变差函数可以得到克立格线性方程：

$$\left.\begin{aligned}&\sum_{i=1}^{n}\lambda_i\gamma(x_i,x_j)+\psi(x_0)=\gamma(x_j,x_0),(j=1,2,\cdots,n);\\&\sum_{i=1}^{n}\lambda_i=1\ .\end{aligned}\right\}. \tag{3}$$

式中：$\gamma(x_i,x_j)$ 为已观测到的 x_i 和 x_j 两点间的变差函数值；$\gamma(x_j,x_0)$ 为已采样的位置 x_j 和未采样的位置 x_0 两点间的变差函数值；ψ 为拉格朗日乘子. 根据式(3)计算克立格权重系数 λ_i，然后把它代入式(1)中，可以得出估计值 $\hat{Z}(x_0)$.

估计值 $\hat{Z}(x_0)$ 的不确定性可用克立格方差来衡量：

$$\sigma_{\mathrm{K}}^2(x_0)=\sum_{i=1}^{n}\lambda_i\gamma(x_i,x_0)+\psi(x_0)\ . \tag{4}$$

关于普通克立格法的详细内容查看文献[2,4].

2　多基因遗传规划法

1992 年 Koza[6] 提出的遗传规划被认为是最佳的符号回归方法，它广泛应用于各种复杂过程的建模中，表现出了优异性能. 遗传规划模仿生物界的自然选择和遗传机制，在由许多可行解组成的搜索空间中，通过选择、复制、交换和突变等遗传操作，按照最优适应度逐步迭代而寻找出最优解[7]. 遗传规划的优点在于自动地生成函数形式，同时也获得其系数.

2.1　多基因遗传规划的特点

多基因遗传规划是一种遗传规划的鲁棒的变体. 该方法采用被称为“多基因”的一种新的特征，将标准遗传规划的模型结构选择能力与传统回归方法的参数估计能力有效地结合[8]. 在传统遗传规划中，模型由单一树/基因组成，而在多基因遗传规划中，模型为几个树/基因的线性组合，其中每个树/基因代表了传统遗传规划中的树/基因. 多基因遗传规划是先采用输入输出变量间的低位非线性变换，然后进行它们的线性组合，产出数据集的数学模型. 近年来，多基因遗传规划成功

应用于建模、预测、控制、符号处理、机器学习等工程领域中. 事实证明,与传统遗传规划相比,多基因遗传规划在非线性建模中具有更高的准确性及更好的有效性[8-10]. 多基因遗传规划的实施步骤如下:

1) 生成一个初始群体,其每个个体由函数符号集合与终止符号集合的组合构成;

2) 计算群体中每个个体的适应度;

3) 依据适应度以一定概率选择优良个体作为父代;

4) 通过交换、突变和复制等遗传操作,生成新的个体作为新一代的种群(也就是子代);

5) 若满足终止准则进化过程应立即停止,否则返回到 2).

终止准则一般包括两种:一是自定的最大代数,二是最小误差等适应度标准. 函数符号集合包括算术运算子(+, −, ×, ÷)、非线性函数(sin, cos, tan, exp, tanh, log)、布尔运算子(and, or, etc.)和条件运算子等;终止符集合可包括输入变量或随机常值.

多基因遗传规划与传统遗传规划之间最明显的区别在于前者参与进化的模型为几个基因/树的组合.

假设由维数为 $\mathbf{R}^{n\times m}$ 的输入变量 u 和维数为 $\mathbf{R}^{n\times 1}$ 的输出变量 y 构成的系统,其中 n 为已有的观测值的数目,m 为输入变量的数目. 采用遗传规划法的树形结构可以表达其系统的数学关系:

$$\hat{Z}=f(u_1,\cdots,u_i)\ . \tag{5}$$

在多基因遗传规划法中,每个输出变量的预测值 $\hat{Z}$ 是由多基因个体中每个树/基因的加权输出值和偏项组成的,而每个树是一个关于输入变量 u_1, …, u_i(其中 $i\geqslant 1$)的函数. 在数学上,多基因遗传规划模型可表示为

$$\hat{Z}=d_0+d_1\times \mathrm{tree}_1+\cdots+d_M\times \mathrm{tree}_M. \tag{6}$$

式中:d_0 为偏项;d_1, …, d_M 为每个树/基因的权重系数;M 为组成有效个体的树/基因的数目. 每个多基因个体的权重系数(就是回归系数)通过最小二乘法自动决定. 在多基因遗传规划中,每个符号模型是若干遗传规划树的加权线性组合,而每个树可被看为 1 个基因. 多基因遗传规划模型和其数学表达式的典型例子,如图 1 所示.

2.2 多基因遗传规划在估计矿石品位中的应用

在估计矿石品位的过程中,与克立格法相比,多基因遗传规划法不需要空间分布的假设. 克立格法估算品位前需得出空间变异函数,而多基因遗传规划法不包含此步骤. 为了估计空间变量之间的内在关系,多基因遗传规划法只需要关于输入输出变量的训练数据. 当采用多基因遗传规划对矿石品位进行估计时,输入数据将以采样位置二维空间坐标值的形式表达,且输出数据为对应位置的矿石品位值. 多基因遗传规划法将矿石品位估计问题转换为在数据坐标空间中函数拟合问题. 其数学表达式为

$$\hat{Z}=f(x,y)\ . \tag{7}$$

式中:(x,y) 为空间坐标值;$\hat{Z}$ 为在对应位置的矿石品位估计值.

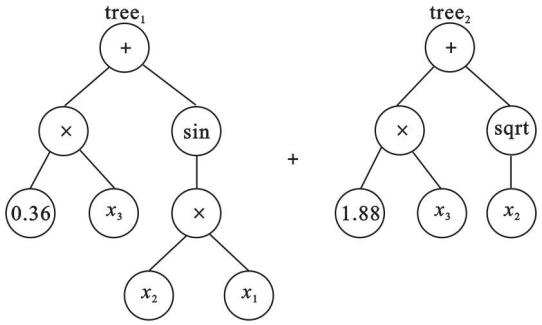

图 1 多基因遗传规划模型的例子

Fig. 1 Example of MGGP model

在估计矿石品位中,用于评估群体性能的适应度函数为矿石品位实际值与预测值间的均方根误差:

$$\mathrm{fitness}=\sqrt{\frac{\sum_{i=1}^{N}|G_i-Z_i|^2}{N}}\ . \tag{8}$$

式中:G_i 为在第 i 个空间坐标上采用多基因遗传规划得到的品位预测值;Z_i 为对应的实际观测值;N 为样本数目.

3 案例研究

为了在矿石品位估计中将多基因遗传规划法与克立格法进行对比,本文选取了用于 Clark[11] 的地质统计研究中的某种铁矿数据. 该数据能在 http://www.kriging.com/datasets/网站上得到. 本文所研究的低品位铁矿具有约 35% 的总平均铁品位,包含随机分布地垂直于矿体倾斜方向的 50 个钻孔数据.

随机选取其中 30 个样本作为估计模型的训练数据,将剩下的 20 个样本作为验证数据用于交叉验证(见图 2). 其铁矿石品位数据具有变程为 100 m、基台为 25% 和块金效应为 0 的球状变差函数模型[11]. 本文采用上述的球状变差函数模型

参数，在二阶平稳性或内蕴假设下，进行普通克立格插值.

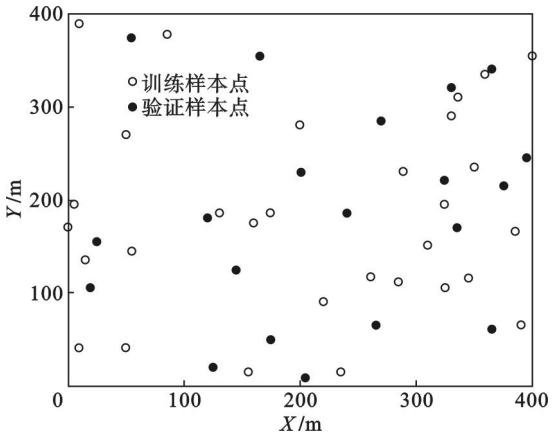

图 2　铁矿钻孔的空间分布

Fig. 2　Spatial distribution of iron ore borehole

当估计铁矿石品位时，多基因遗传规划法需要控制参数. 其参数设置值列于表 1.

表 1　多基因遗传规划法的参数设置

Table 1　Parameter setting for MGGP model

参数	设置值
群体规模	500
进化代数	200
锦标赛选择大小	20
最大树深度	6
最大基因数目	4
函数符集合(F)	+, −, ×, ÷, sin, cos, tan, atan, tanh, exp, plog, psqrt, square
终止符集合(T)	x, y, [−10 10]
交换概率	0.85
复制概率	0.10
突变概率	0.05

本文采用的多基因遗传规划输入变量为样本点的 x 和 y 空间坐标值，而输出变量为其在样本点的铁矿石品位值. 函数符号集合包括较多的要素以便能提供多种多样的非线性数学模型. 群体规模和进化代数等的参数依赖于回归问题的复杂程度. 如果训练数据的样本规模较小，那么其群体规模和进化代数应该较大，以便寻找误差最小的模型. 根据所研究问题的要求，本文将群体规模设为 500，进化代数设为 200. 最大树深度与最大基因数目能够有效地控制模型的复杂性，使多基因遗传规划法得出一个简单正确的模型. 所以，本文设最大树深度为 6，最大基因数目为 4.

基于最小适应度得出的最优模型的数学表达式为

$$\hat{Z}(x,y) = 36.09 + 6.77 \times [\operatorname{atan}(\operatorname{square}(\sin(1.802x))) - \operatorname{atan}(\operatorname{plog}(\operatorname{plog}(y)))] - 3.473 \times [\cos(\operatorname{plog}(y - x + 6.242))] + 1.911 \times [\sin(\sin(y - 5.816)) - \cos(\sin(y) - x + 6.771) + \operatorname{psqrt}(49.83 - 3.965x)] + 4.013 \times [\sin(\operatorname{plog}(y)) \times \operatorname{atan}(x - 5.395) - \tanh(\sin(\operatorname{plog}(y))) - \tanh(\operatorname{atan}(x - 5.395)) + \sin(\operatorname{plog}(y)) \times \sin(y - 5.912)].$$

多基因遗传规划法的实施结果如图 3 所示.

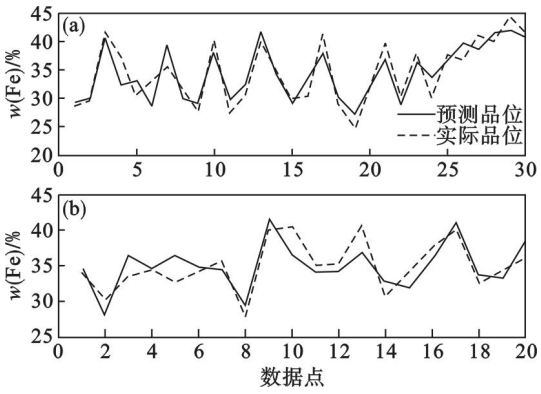

图 3　多基因遗传规划法在铁矿石品位估计中的实施结果

Fig. 3　Implementation result of MGGP model for estimating the iron ore grade

(a)—训练数据(RMSE 2.364 2)；
(b)—验证数据(RMSE 2.043 3).

最后，采用相关系数(R)、平均绝对预测误差(MAPE)和平方根预测误差(RMSPE)评估交叉验证结果，并进行了克立格法与多基因遗传规划法两种模型的对比. 其计算公式如下：

$$R = \frac{\sum_{i=1}^{N}(Z(x_i) - \bar{Z}(x_i))(\hat{Z}(x_i) - \bar{\hat{Z}}(x_i))}{\sqrt{\sum_{i=1}^{N}(Z(x_i) - \bar{Z}(x_i))^2 \sum_{i=1}^{N}(\hat{Z}(x_i) - \bar{\hat{Z}}(x_i))^2}}; \quad (9)$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N} | Z(x_i) - \hat{Z}(x_i) |; \quad (10)$$

$$\mathrm{RMSPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Z(x_i) - \hat{Z}(x_i)]^2}. \quad (11)$$

式中：$Z(x_i)$ 为在样本点 x_i 的实际观测值；$\hat{Z}(x_i)$ 为对应的预测值；N 为参与交叉验证的样本数目.

对比分析结果见表 2.

表 2　普通克立格法与多基因遗传规划法的对比结果

Table 2　Comparison of the ordinary Kriging and MGGP

方法	R	MAPE	RMSPE
普通克立格法	0.631 27	2.481 4	2.948 9
多基因遗传规划法	0.814 42	1.761 3	2.043 3

(下转第 420 页)