doi: 10.11720/wtyht.2017.2.13
梁文轩.基于遗传算法的多母体正态分离及其地质应用[J].物探与化探,2017,41(2):278-282.http://doi.org/10.11720/wtyht.2017.2.13
Liang W X.Multi-parent normality separation based on genetic algorithm and its geological applications [J].Geophysical and Geochemical Exploration, 2017,41(2):278-282.http://doi.org/10.11720/wtyht.2017.2.13

基于遗传算法的多母体正态分离及其地质应用

梁 文 轩
(广东省核工业地质局291大队,广东 佛山 528133)

摘 要:在漫长的地质演化史中,地球表面经历了多期多阶段的地质作用,同一地层同一元素其分布状态存在差异,或其物质来源可能来自不同的母体,致使数据直方图呈现多峰形态。多峰分离是区分物质来源的基础,有利于提高地质解释质量。而地学数据大多服从正态分布(或对数正态分布),这就为多峰分离提供了数学依据。文章阐述了多母体样本数据分离的数学原理,并以遗传算法作为实现手段,从理论上证明了遗传算法对多母体样本数据的分离具有较好效果,其拟合优度可达95%。通过实例分析,遗传算法能正确区分白垩系大凤组第一段(Kd^1)地层的两种岩性,说明其在地质研究中具有实际意义。

关键词:正态分离;多母体;遗传算法
中图分类号:P631 **文献标识码**:A **文章编号**:1000-8918(2017)02-0278-05

0 引言

在处理物化探数据时常常会碰到同一地层或岩性其数据直方图呈现为明显的多峰形态,若笼统地认为物质来源于一个母体,则有可能歪曲了事物的本来面目,因此须对原始数据进行母体分离,以明确其地质意义。而自然界的事物大多服从正态分布(高斯分布),地壳表层的元素分布也不例外,服从正态分布或对数正态较为常见。因此,基于正态分布的多母体样本分离就有实际意义。但其多参量非线性求解对传统处理方法是个极大的挑战,而遗传算法在处理非线性问题及数值优化方面有较大优势。

1 正态分离的研究现状

正态分离在各行业中都有应用,其中在物理、化学[1-4]中应用最多,其次是在图像分析[5-7]方面。张建华等[8]在测定烷基葡萄糖苷(APG)的临界胶束浓度(CMC)时,应用高斯多峰拟合法对水溶液体系的可见光谱吸收叠加峰进行分离,并计算峰面积、频移和半高宽等光谱线型参数,首次发现了可见吸收光谱的半高宽对APG聚集行为敏感,并成功用于APG的CMC测量;仲岑然等[9]为了客观分析上浆纱线横截面切片图像,减少主观因素对分析结果的影响,通过对浆纱横截面切片图像灰度直方图的特征分析,使用高斯分峰方法得出浆料因素在灰度直方图上的分布区间,并计算出浆纱的几何上浆率和浆料相对分布半径,该方法比依据颗粒及其平均中心测量方法更为方便精确;朱杰等[10]通过分析鄂尔多斯盆地的石油勘探开发历程,分析了该盆地石油储量与产量的增长具有多峰形态的特点后,采用多峰高斯拟合该盆地的石油储量和产量的历史曲线,并预测了2006~2030年盆地石油储量与产量的增长趋势。

2 基本原理[11-13]

多母体正态分离是以各母体样本均服从正态分布为前提,因此需要用到正态概率密度函数,其表达式如下:

收稿日期:2015-07-20;**修回日期**:2016-07-06
基金项目:中国地质调查局"广东黄坑—百顺地区矿产地质调查"项目(12120113063800)
作者简介:梁文轩(1983-),男,物探工程师,主要从事固体矿产勘查工作。

万方数据

$$f(x)=\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\tag{1}$$

其中:μ 为随机变量的均值;σ 为随机变量的标准偏差。假设样本服从正态分布,其样本总数为 N,分组的组距为 L,分组数为 M,组中值向量为 X,频率向量为 F,频数向量为 Y,则有:

$$F=f(X)\cdot L,\tag{2}$$

$$Y=F\cdot N=f(X)\cdot L\cdot N,\tag{3}$$

令:$A=L\cdot N$,那么:

$$Y=y(X)=A\cdot f(X)。\tag{4}$$

式(4)即样本对于特定的平均值 μ 和标准偏差 σ 在 X 位置时的频数。对于多母体来源的样本,各母体的样本数(N)、均值(μ)及标准偏差(σ)均为未知,其解在可能解空间中有无数种组合,遗传算法所做的就是在无数种解中选择最优解(N 或 A、μ、σ)。

假设样本母体的个数为 k,各母体均服从正态分布。遗传算法首先在解空间中创建种群,计算各个个体的适应度值,通过选择、交叉、变异等算法选择优良基因遗传给下一代,直至算法满足终止条件为止。适应度函数的数学表达式如下,令:$C=(C_1,C_2,\cdots,C_M)$ 为样本实际分组的频数向量,$Z=(Z_1,Z_2,\cdots,Z_M)$ 为遗传算法计算各母体在组中值 X_M 中的频数和,那么有:

$$Z_i=\sum_{j=1}^{k}y_j(X_i),\quad i=1,2,\cdots,M\tag{5}$$

令

$$SSR=\sum_{i=1}^{M}(Z_i-C_i)^2。\tag{6}$$

式(6)表达的是遗传算法每次迭代后各母体对应分组频数之和与实际分组的频数的残差平方和,是编写适应度函数的核心公式,SSR 越小说明峰分离效果越好。

一般来说,样本数量都较多,为了缩小搜索空间及其与 μ 和 σ 的差距,可对式(4)变形得:

$$A=\frac{Y}{f(X)}=\frac{Y}{\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}},\tag{7}$$

令 $x=\mu$,则

$$A=\frac{y_0}{\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}}=\sqrt{2\cdot\pi}\cdot\sigma\cdot y_0,\tag{8}$$

那么式(4)可变为:

$$Y=y(X)=A\cdot f(X)=\sqrt{2\cdot\pi}\cdot\sigma\cdot y_0\cdot f(X)=y_0e^{-\frac{(x-\mu)^2}{2\sigma^2}},\tag{9}$$

其中:y_0 为 $x=\mu$ 时的频数,即正态曲线的峰值。如果 y_0 仍然很大,可于其前加一系数,使 $y_0=B\cdot y_B$,那么适应度函数最后的输入参数变为 y_B、μ 和 σ。

3　利用 MATLAB 进行处理的步骤

3.1　数据分组

数据分组及直方图绘制可使用 hist( )函数。当数据量足够大,在增加分组数时,直方图形态没有明显改变的,可采用较大的分组数,这样对降低残差平方和的值有显著效果。

3.2　变量约束

对于峰形态明显的曲线,可利用 findpeaks( )函数来寻找峰位及根据峰的形态估算其 μ 和 σ 值,但对于峰形态不明显的曲线,采用上述方法就容易得出不合理的解。适应度函数经过式(9)修改后,其参数值都较小,其中 y_0 小于直方图的最大频数,μ 在直方图的上下边界间,σ 则可设为直方图上下边界之差除以 6,在这范围内搜索,遗传算法有较高的性能。

3.3　参数设置

遗传算法的参数主要包括种群大小、选择算子、交叉算子和变异算子等,参数本身又有不同的函数及控制参数,这些参数的设置目前没有可靠的方法一次完成,较常见的设置为种群大小在 10~200 之间,交叉概率在 0.4~1 之间,变异概率在 0~0.1 之间[14-15]。通常还是靠经验来人为调整,其基本原则为:①保持种群的多样性;②防止陷入局部最小;③性能相对最佳。对于求解简单的问题,较小的种群数及较低变异率能使算法快速收敛到最优解;对于复杂问题,扩大种群、提高交叉和变异率是防止算法过早收敛的一般做法,但其牺牲了算法的性能。因此,遗传算子的选择须根据具体问题选择合适的函数及参数。

4　多母体分离的测试效果

4.1　随机数据

首先以具有 2 个母体的正态分布数据为例。令母体 1 的样本数为 800 个,平均值为 10,标准偏差为 2,母体 2 的样本数为 1 000 个,平均值为 15,标准偏差为 3。以上述参数使用 Matlab 随机生成具有正态分布的数据向量,并将母体 1 和母体 2 的数据混合成总样本,分别绘制具有相同横坐标(即相同组中值)的频数直方图(图 1)。

万方数据

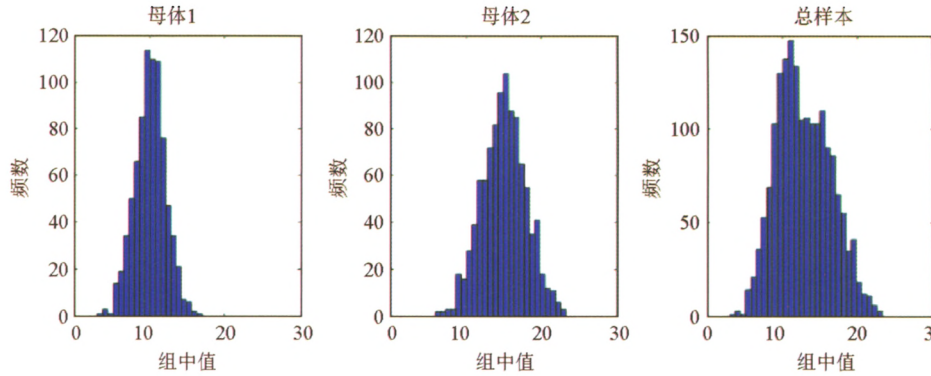

图 1　不同母体及其总样本的直方图

4.2　分离效果

对随机产生的数据作为实际测量数据，使用遗传算法进行自动搜索，其搜索过程的适应度值变化和当前最优个体如图 2 所示。

从图 2 可以看出，适应度值（即残差平方和 *SSR*）随着遗传代数的增加而迅速下降，迭代 20 多次已基本找到最优解，其最小适应度值为 609.007，可利用该值计算遗传算法的拟合优度，公式如下：

$$令\ SST = \sum_{i=1}^{M} C_i^2, \tag{10}$$

那么有：

$$R^2 = 1 - \frac{SSR}{SST}, \tag{11}$$

由上述公式计算所得的拟合优度为 94.0%，拟合效果较好，如图 3 所示。由遗传算法所求得的样本参数与实际样本参数的对比如表 1 所示。

表 1 同样说明了遗传算法对母体分离有较好的效果。对于具有 2 个以上服从正态分布的母体的数据，同样具有较好的效果。如图 4 为具有 5 个母体的分离效果图，其拟合优度为 95.2%。

但随着母体数的增加，遗传算法搜索的变量数也急剧增加，其解空间变得十分复杂，这使得算法运行的时间增加，也大大降低了算法的精度，对于均值相近的母体，其分离能力也渐渐变弱。

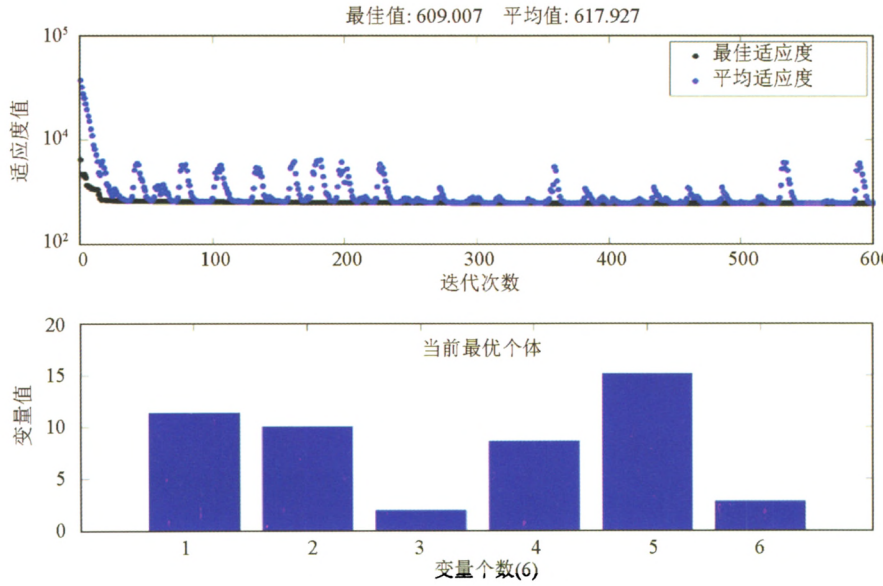

图 2　适应度变化和最优个体

表 1　样本参数对比

	母体 1	母体 2	遗传算法计算结果		相对误差/%	
样本数(N)	800	1 000	740	1 053	7.50	5.3
平均值(μ)	10	15	10.07	15.19	0.70	1.27
标准偏差(σ)	2	3	1.98	2.84	1.00	5.33

万方数据

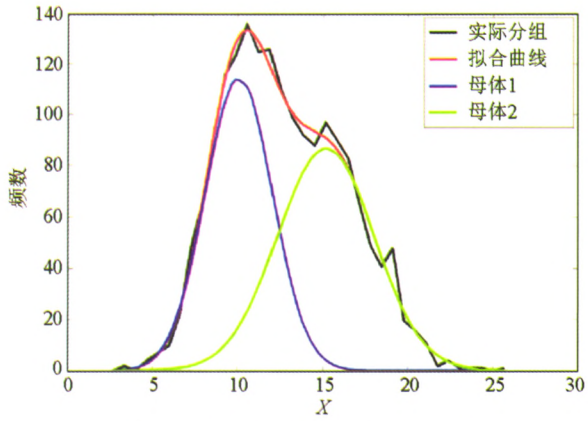

图 3　具有 2 个母体的正态分离效果

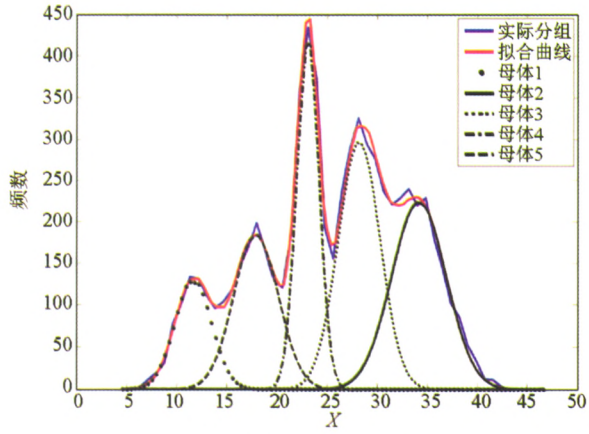

图 4　具有 5 个母体的正态分离效果

5　实例分析

以地面伽马能谱测量的铀元素为例,对 Kd^1 地层进行岩性划分析。

首先,在Matlab命令行界面执行“[nelements, centers] = hist (data, Group) ;”,其中“nelements”为分组频数向量,“centers”为各分组的上下限分界向量,“data”为待分析数据,“Group”为拟分组数。根据以上参数可以求得适应度函数的约束边界,并可从直方图中读取峰数,从而确定样本母体数。本例样本涉及两个母体,待解决问题不算复杂,遗传算法的主要参数设置如下:

种群大小(Population size):100~200;

适应度比例(Fitness scaling):Rank;

选择函数(Selection function):Stochastic uniform;

交叉概率(Crossover fraction):0.6~0.8;

交叉函数(Crossoverfunction):Constraint dependent;

变异函数(Mutationfunction):Constraint dependent。

然后执行算法,其搜索过程的适应度值变化及当前最优个体如图 5 所示。从图 5 的上半部分可以看出其收敛速度非常快,拟合优度达 88.1%,其母体分离效果见图 6。

从图 6 可知 Kd^1 地层的铀元素明显来自不同母体,其中母体 1 的 $\mu_1=2.99$,$\sigma_1=0.27$;母体 2 的 $\mu_2=4.54$,$\sigma_2=0.53$。以 $\mu_1+\sigma_1$ 和 $\mu_2-\sigma_2$ 的均值作为划分母体 1 和母体 2 的样本(或以 $\mu\pm\sigma$ 求出各母体的主体样本),并绘于地质图上,见图 7。

从图 7 可以看出,母体 1 的样本点基本分布在 Kd^1 地层的上部,母体 2 的样本点则分布在 Kd^1 地层的下部,这种具有明显分界线的现象并非偶然,事实上,母体 1 的岩性大多为紫红色粉砂岩、细砂岩,而母体 2 的岩性则为棕红色花岗质砂砾岩,岩性上的差异造成了铀含量背景值的差异。对于蚀变岩,

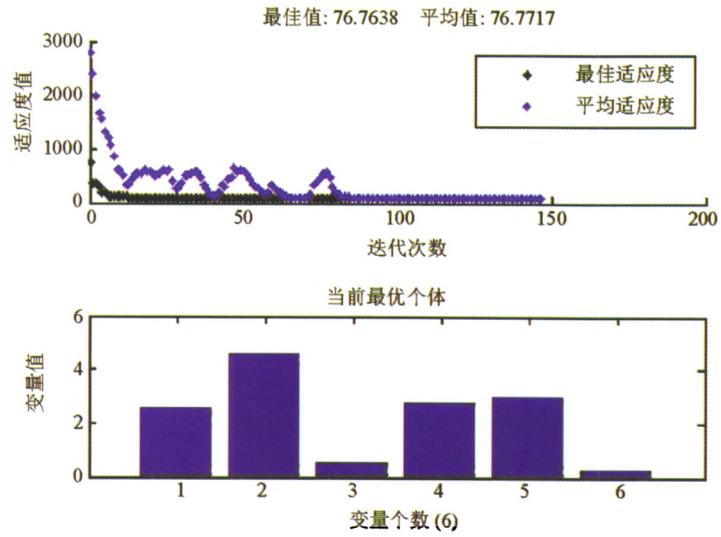

图 5　适应度变化和最优个体

万方数据

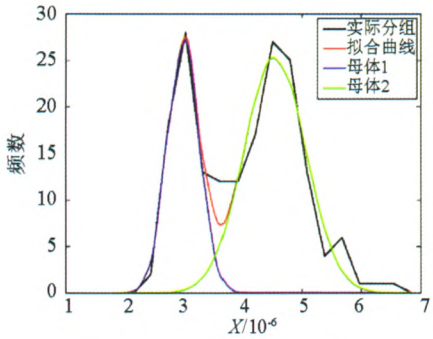

图 6　铀含量正态分离效果

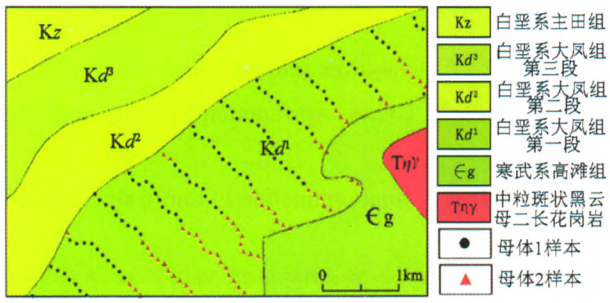

图 7　母体样本分布

果其样本足够多及其特征元素偏离正常岩，也可将其从正常岩石中分离出来。

6　结语

遗传算法的搜索速度和精度很大程度取决于输入变量的多少和算法参数的设置。频数正态分布有 3 个变量，母体多则输入变量也多，种群数量及变异系数等也随之增大。但一般情况下，日常处理的数据其直方图很少有多于 3 个峰的形态。对于峰数少于或等于 3 个的情况下，遗传算法都有较高的运行速度和精度，在实际应用中有一定的价值。

参考文献：

[1] 冯望生，房艳，徐继香，等.高斯多峰拟合在径向分布函数中的应用[J].物理化学学报，2008，24(3)：497－501.

[2] 张建华，王珺珂，毛会坚.高斯多峰拟合运用于十二烷基苯磺酸钠聚集状态的的研究[J].计算机与应用化学，2006，23(12)：1305－1308.

[3] 张建华，刘琼，陈玉苗，等.紫外-可见光吸收光谱结合高斯多峰拟合技术测定甲基红酸离解常数[N].物理化学学报，2012，28(5)：1030－1036.

[4] 李睿，金保昇，仲兆平，等.高斯多峰拟合用于生物质热解三组分模型的研究[J].太阳能学报，2010，31(7)：806－810.

[5] 尚雅层，陈静，田军委.高斯拟合亚像素边缘检测算法[J].计算机应用，2011，31(1)：179－181.

[6] 杨利红，赵变红，张星祥，等.点扩散函数高斯拟合估计与遥感图像恢复[J].中国光学，2012，5(2)：181－188.

[7] 李炜，黄心汉，王敏，等.基于二维高斯曲面拟合的图像灰度补偿算法[J].华中科技大学学报：自然科学版，2004，32(2)：43－45.

[8] 张建华，孔凯清，何争玲，等.高斯多峰拟合用于烷基葡萄糖苷临界胶束浓度测量[J].光谱学与光谱分析，2007，27(7)：1412－1415.

[9] 仲岑然，金春奎.基于 Gaussian 拟合的浆纱横截面切片图像分析[N].纺织学报，2011，32(3)：57－60.

[10] 朱杰，柳广弟，刘成林，等.基于多峰高斯模型的石油储量与产量预测[J].中国石油大学学报，2009，33(3)：45－48.

[11] 盛骤，谢式千，潘承毅.概率论与数理统计[M].北京：高等教育出版社，2010.

[12] 於崇文.数学地质的方法与应用[M].北京：冶金工业出版社，1980.

[13] 孙洪泉.地质统计学[M].徐州：中国矿业大学出版社，1990.

[14] 雷英杰，张善文，李续武，等.MATLAB 遗传算法工具箱及应用[M].西安：西安电子科技大学出版社，2006.

[15] 徐波.遗传算法及其在数据挖掘中的应用[J].电脑编程技巧与维护，2010(4)：9－16.

Multi-parent normality separation based on genetic algorithm and its geological applications

LIANG Wen-Xuan

(*No.291 Geological Party, Guangdong Geological Bureau of Nuclear Industr, Foshan　528133, China*)

Abstract: In the very long geological evolution history, earth surface has undergone multi-period and multistage geological processes, and the distribution of the same element in the same stratum differs or its material source might have come from different matrixes, which results in data histogram presenting a multimodal morphology. Multimodal separation is the basis of differentiating material sources, and helps improve geological interpretation level. The fact that geological data mostly obey the normal distribution (or logarithmic normal distribution) provides mathematical basis for multimodal separation. This paper explains the math principles of multiple matrix sample data separation and uses genetic algorithm as implementation approach to theoretically prove that genetic algorithm has a better effect on multiple matrix sample data separation, and its goodness of fit can be up to 95%. An analysis of practical examples shows that the genetic algorithm can correctly distinguish 2 kinds of lithology of the stratum, which suggests that it has practical significance in geological research.

Key words: normality separation; multi-parent; genetic algorithm　　(本文编辑：王萌)

万方数据