基于麻雀搜索算法的土壤重金属X射线荧光光谱重叠峰解析

陈 颖[1]，刘峥莹[1]，肖春艳[2]，赵学亮[1,3]，李 康[3]，庞丽丽[3]，史彦新[3]，李少华[4]

1. 燕山大学电气工程学院河北省测试计量技术及仪器重点实验室，河北 秦皇岛 066004
2. 河南理工大学资源与环境学院，河南 焦作 454000
3. 中国地质调查局水文地质环境地质调查中心，自然资源部地质环境监测工程技术创新中心，河北 保定 071051
4. 河北先河环保科技股份有限公司，河北 石家庄 050000

摘 要 近年来随着土壤重金属污染的加剧，和人们环境意识的逐渐提高，科研人员对快速检测土壤重金属含量方法的研究正在不断深化。目前，X射线荧光分析法(XRF)是广泛应用于土壤重金属污染检测的方法。但由于X射线荧光光谱仪的能量分辨率有限，而一些重金属元素的荧光产额较低，一些元素的相邻谱峰出现了重叠现象。针对XRF法中元素相邻谱峰的重叠问题，提出了一种基于麻雀搜索算法(SSA)的光谱重叠峰解析方法。首先，将从河北保定地区采样得到的土壤，制备出不同含水率、不同重金属元素含量的样本并用X射线荧光光谱仪获取原始光谱数据。接着，对光谱数据进行预处理，采用谱聚类算法剔除异常光谱样本，采用Savitzky-Golay五点二次去噪法和线性本底法完成对光谱的去噪和本底扣除，并对光谱净计数用随机数法生成大量模拟光谱数据，以备后续算法使用。然后，用期望最大化法(EM)对重叠峰进行初步解析，首先设置EM算法的初始参数，并将生成的模拟光谱数据代入EM算法，当达到迭代次数时，即可初步得到高斯混合模型(GMM)中各高斯峰的期望、方差和权重参数。但由于EM算法容易受初始参数设置的影响，且易陷入局部最优而导致结果不准确，还需对EM算法进一步优化。本研究采用SSA对GMM的各参数进行全局优化，在设置SSA算法的基本参数后，将100组由EM算法得到的参数作为该算法的初始种群，并设置合适的适应度函数，通过迭代，最终得到全局最优参数，实现了重叠峰的分解。SSA受参数设置的影响较小，相比于一些传统的优化算法，如遗传算法(GA)、蚁群算法(ACO)、粒子群算法(PSO)等，具有收敛速度快、不易陷入局部最优的特点，因此，采用此算法，可以达到较好的优化效果。通过对重叠峰解析结果的分析表明，该算法可在较少的迭代次数下得到较准确的解析结果，可广泛应用于能谱重叠峰解析。

关键词 X射线荧光分析法；高斯混合模型；期望最大化法；麻雀搜索算法；重叠峰解析

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)07-2175-06

引 言

近年来土壤重金属污染愈发严重，重金属随食物链进入日常饮食，对人们的健康造成极大影响[1]，如何快速检测出土壤中各种重金属含量成研究热点。X射线荧光分析法(X-ray fluorescence analysis，XRF)是其中应用较广泛的检测法[2]。但因光谱仪能量分辨率及重金属Pb和As的荧光产额较低，Pb和As的荧光光谱严重重叠，这将给重金属含量预测带来严重误差，故需进行重叠峰解析，以便后续计算各元素含量。

目前，国内外研究学者已对重叠峰解析方法做了大量研究。其中，周世融[3]等人提出峰锐化法结合双树复小波变换的重叠峰解析方法；刘红莉[4]等提出基于粒子群算法(PSO)的重叠峰解析方法；Liu[5]等提出基于新型小波变换的重叠峰解析方法；Michael[6]等提出用演化因子法进行重叠峰解析；Xiong[7]等提出基于多阶差分法和遗传算法(GA)的重叠峰解析法。以上方法均能解析重叠峰，但有时准确度不高，或易陷入局部最优，尤其是用全局优化算法实现重叠峰分解时，这些问题时常发生。如使用GA时常因变异率设置不当

收稿日期：2020-07-06，**修订日期**：2020-11-20

基金项目：国家重点研发计划项目(2018YFC1800903，2016YFC1400601-3)，河北省重点研发计划项目(19273901D，20373301D)，河北省自然科学基金项目(F2020203066)，中国博士后基金项目(2018M630279)，河北省博士后择优资助项目(D2018003028)，河北省高等学校科学技术研究项目(ZD2018243)资助

作者简介：陈 颖，女，1980年生，燕山大学电气工程学院教授 e-mail：chenying@ysu.edu.cn

导致过早收敛；PSO 易陷入局部最优[8]。这些问题多由于参数设置不当导致，麻雀搜索算法(sparrow search algorithm，SSA)需要设置的参数少，能一定程度避免因参数设置带来的问题，且因其独特的算法思路，相比同类算法更不易陷入局部最优，且收敛速度快。将该法用于对期望最大化法(expectation-maximization algorithm，EM)得到的高斯混合模型(Gaussian mixture model，GMM)参数进行全局优化，可在较少的迭代次数下得到较准确的结果。

1 实验部分

实验采用华北平原的自然土，去除杂质，将土壤磨细并烘干。因华北平原土壤含水率大致分布在 10%～25%之间，故分 10 个梯度配制含水率为 10%～25%的土壤样本。因同时含 As 和 Pb 元素的样本荧光光谱会产生重叠，故需分别获取含 As 元素、Pb 元素，和同时含 As 和 Pb 元素样本的荧光光谱。取实验土壤 30 g，根据要配制样本的重金属含量向土壤中加入 As 和 Pb 标准溶液，充分搅拌。因 As 和 Pb 的标准溶液元素含量低，将溶液加入土壤会导致含水率远超 25%，故需将混合好的土壤烘干到含水率满足实验需求，再制作压片和样品盒样本。按此方法，分别制出含量为 100，200，400 和 600 mg·kg^{-1}的 As 土壤样本和 Pb 土壤样本，和这几种浓度相互组合的 As 和 Pb 元素混合的土壤样本。

用能量分辨率为 128 eV 的 CIT-3000SYB 能量色散 X 荧光分析仪，对各样本分别采集 6 次 X 射线荧光光谱数据供后续使用。观察光谱图 1 可发现，As 和 Pb 的谱峰严重重叠，需对重叠部分进行分解。

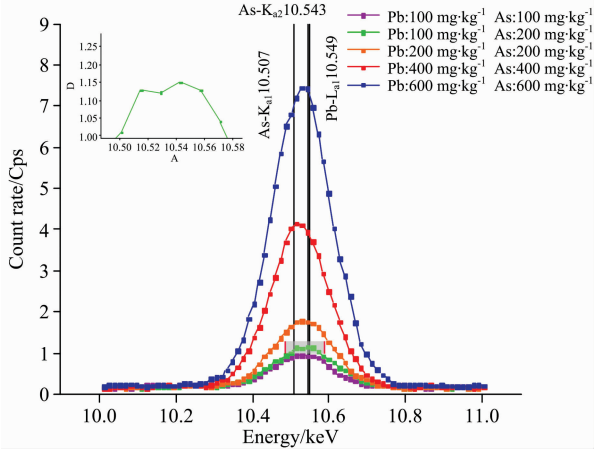

图 1 重叠峰示意图

Fig. 1 Diagram of overlapping peaks

2 数据预处理

选取谱峰重叠部分，10.3～10.8 keV 能量范围的荧光光谱为研究对象。对该部分光谱进行剔除异常样本、去噪、扣除本底和根据净计数生成随机数的预处理操作。

2.1 剔除异常数据样本

对实验中因粗大误差、环境温湿度变化等因素导致的异常样本，在重叠峰解析前需予以剔除。选取含水率相同但重金属含量不同的几组样本数据，每一组数据中包含对同一样本的 6 次采样数据。对各光谱数据提取重金属含量特征，并按该特征的欧式距离聚类。若有某一含量的样本数据聚类后被归到其他含量类别中，则将该样本数据视为异常，予以剔除。

2.2 光谱去噪及本底扣除

在光谱采集过程中，因光子辐射、环境细微变化等原因，会不可避免地将噪声引入光谱数据。为使最终结果准确，采取 5 点 2 次 Savitzky-Golay 卷积平滑去噪法，将光谱数据中相邻的五个点用二次多项式拟合，并以此代替原光谱，依次移动，直到遍历所有光谱数据，即完成对光谱的去噪。去噪后光谱更平滑，该法有效降低了光谱噪声。

另由于原级 X 射线在样品中会发生康普顿散射和瑞利散射，样品产生的射线与仪器相互作用，加上宇宙射线和电子线路的扰动，会使光谱中自带背景，即本底。为得到净荧光强度，需将本底扣除。采用线性本底法，对光谱图中谱峰底部的拐点依次用线段连接，并将连线下的部分扣除，由此获得净光谱数据。去噪和本底扣除的前后对比图如图 2 所示。

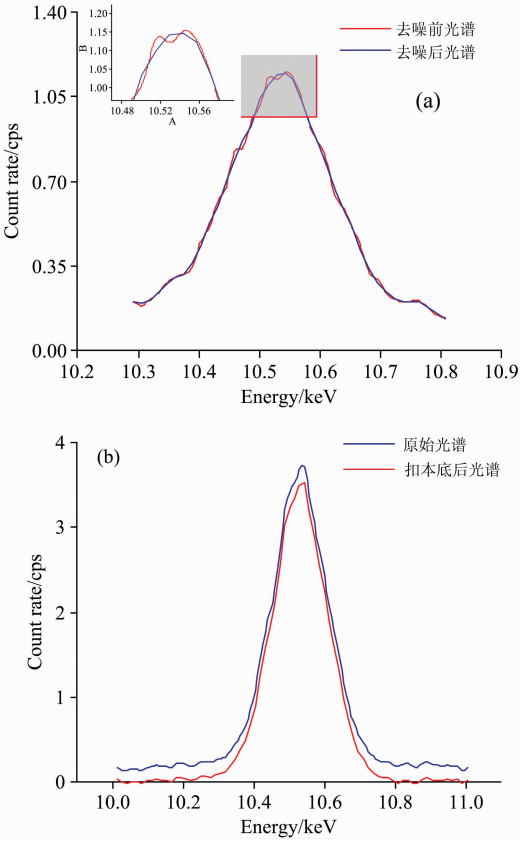

图 2 光谱去噪、扣除本底前后对比图

(a)：S-G 平滑去噪；(b)：线性本底法扣除本底

Fig. 2 Spectral denoising and background subtraction before and after comparison

(a)：S-G smooth denoising；

(b)：Linear background minus background

2.3 根据净计数生成随机数

因不能将光谱净计数直接代入以下算法中，故需根据预处理过的净计数生成大量随机数，得到由随机数分布而成的模拟能谱。先对感兴趣区中的净计数归一化，即将每一个净计数依次除以整个感兴趣区中净计数的加和，以此得到每一净计数对应的道数在随机数中出现的概率，即得到了模拟能谱分布的概率密度函数。再根据此概率，生成由各道数组成的随机数。生成的随机数越多，得到的模拟能谱越准确，这里生成 2 万个随机数。

3 重叠峰解析算法

3.1 期望最大化法(EM)初步得到解析参数

重叠峰可看做几个高斯峰叠加而成的高斯混合模型(GMM)，因此可通过得到各高斯峰的峰位、方差、面积权重参数，实现重叠峰的分解。本次要分解的重叠峰主要由三个子峰叠加而成，分别是 As-$K_{\alpha1}$，As-$K_{\alpha2}$，Pb-$L_{\alpha1}$。

EM 算法通过迭代不断逼近含有隐变量的概率模型参数[9]。因不知道重叠峰光谱中每一个数据来源于 GMM 中的哪一个峰，而只能得到最终的观测值，故重叠峰光谱数据是含有隐变量的，运用 EM 算法较合适。EM 算法的基本步骤如下：

(1) 设置初始参数：在进行算法迭代前，需设置初始参数 θ 作为迭代起点。θ 包括三部分：各子峰的峰位 μ_m，方差 σ_m^2，和子峰占总峰面积的权重 a_m，$m=1, 2, 3$。即 $\theta=[\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, a_1, a_2, a_3]$，其中各权重和 $\sum_{m=1}^{3} a_m = 1$。若对这些参数有一定经验值，可借鉴经验值进行参数设置，以提高迭代效率。

(2) E 步：求完全数据的对数似然函数的期望，因 θ 含多个参数，计算极大似然函数困难，故用求似然函数的期望极值来代替求解似然函数。假设已知观测数据 $x(m)$ 来自 GMM 的哪一个峰，并用 $z(n)$ 指示这个来源，则将不完全数据转换为完全数据，用 y 表示完全数据$[x(m), z(n)]$，求解期望 $E[L(\theta)]$[10]，将其记为 $Q(\theta, \theta^{(i)})$，如式(1)所示

$$Q(\theta, \theta^{(i)}) = E[L(\theta)] = E[\ln P(y \mid \theta)] = \sum_{m=1}^{M}\sum_{n=1}^{N} \frac{P(x(m) \mid z(n), \theta) a_m}{\sum_{m=1}^{M} P(x(m) \mid z(n), \theta) a_m} \cdot [\ln P(x(m) \mid n, \theta) + \ln a_m] \tag{1}$$

其中，$P(x(m)|z(n), \theta)$为高斯混合函数。

(3) M 步：最大化 E 步得到的期望值，具体操作时，可按式(2)，式(3)和式(4)完成最大化 E 步得到的期望。完成此步骤，可得到新一轮的峰位、方差和权重。

$$\mu_m = \frac{\sum_{j=1}^{N} x_{jm} x_j}{\sum_{j=1}^{N} x_{jm}}, \quad m = 1, 2, 3 \tag{2}$$

$$\sigma_m^2 = \frac{\sum_{j=1}^{N} x_{jm} (x_j - \mu_m)^2}{\sum_{j=1}^{N} x_{jm}}, \quad m = 1, 2, 3 \tag{3}$$

$$a_m = \frac{n_m}{N} = \sum_{j=1}^{N} \frac{x_{jm}}{N}, \quad m = 1, 2, 3 \tag{4}$$

其中，x_{jm} 表示来自第 m 个峰的第 j 个光谱数据。μ_m，σ_m^2 和 a_m 分别为新一轮迭代的各参数。当迭代到规定次数时，完成求解。因 EM 算法易收敛于局部极值，且易受初始参数影响，造成结果不准确。因此，本文结合麻雀搜索算法(SSA)，对参数进行全局寻优。

3.2 麻雀搜索算法(sparrow search algorithm, SSA)

SSA 是新提出的模仿麻雀觅食、反捕食行为的群智能优化算法，它将一个种群中的麻雀分为发现者、追随者，并随机分布有一定数量的警示者[11]。

发现者有较好的适应度值，能主动寻找食物并为追随者提供捕食方向。每次迭代中发现者按式(5)更新位置

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \exp\left(\dfrac{-i}{\alpha \mathrm{iter}_{\max}}\right) & \text{if } R_2 < ST \\ X_{i,j}^{t} + \boldsymbol{Q}\boldsymbol{L} & \text{if } R_2 \geqslant ST \end{cases} \tag{5}$$

其中，t 为当前迭代次数，$j=1, 2, \cdots, d$ 为要优化的参数的维度，这里的参数由各峰位、方差、权重组成，为 9 维；$\mathrm{iter}_{\max}$为最大迭代次数；$X_{i,j}^{t}$为迭代第 t 次时维度 j 的第 i 只麻雀；α 和 Q 为随机数，R_2 和 ST 为警报值和安全阈值；$\boldsymbol{L}$ 为 $1\times d$ 的全 1 矩阵。种群中随机分布着一定数目的警示者，若警示者发出的警报值小于安全阈值，说明种群暂不面临危险，发现者进入搜索食物模式。否则，发现者要带领种群转移到安全区。

追随者是适应度值较差的个体，他们监视到发现者找到了食物，便改变当前位置为食物竞争，没争到食物的追随者更愿到其他食物充足的地方去。追随者的位置更新公式如式(6)所示

$$X_{i,j}^{t+1} = \begin{cases} Q\exp\left(\dfrac{x_{\mathrm{worst}}^{t} - x_{i,j}^{t}}{i^2}\right) & \text{if } i > n/2 \\ X_{p}^{t+1} + \left| X_{i,j}^{t} - X_{p}^{t+1} \right| \boldsymbol{A}^{+} \boldsymbol{L} & \text{otherwise} \end{cases} \tag{6}$$

式(6)中，$\boldsymbol{A}^{+} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$；$X_P$ 是最佳的生产者位置，X_{worst}是全局最差位置；$\boldsymbol{A}$ 为元素被随机置为 1 和 -1 的 $1\times d$ 矩阵。

假设警示者占总数的 10%～20%，初始位置随机生成，按式(7)更新位置。

$$X_{i,j}^{t+1} = \begin{cases} X_{\mathrm{best}}^{t} + \beta \left| X_{i,j}^{t} - X_{\mathrm{best}}^{t} \right| & \text{if } f_i > f_g \\ X_{i,j}^{t} + K\left(\dfrac{X_{i,j}^{t} - X_{\mathrm{worst}}^{t}}{(f_i - f_w) + \varepsilon}\right) & \text{if } f_i = f_g \end{cases} \tag{7}$$

式(7)中，X_{best}是全局最佳位置；β 是迭代步长；f_i 为当前麻雀的适应度；f_g 和 f_w 为全局最好和最差的适应度值，K 为$[-1, 1]$间的随机数；ε 是防除零的最小常量。

按式(7)可完成一轮位置更新，每次更新，会使种群向着适应度值更好的方向变更。完成多次迭代后，可得到最佳的适应度值和该值对应的参数。

3.3 建立 SSA 全局优化模型

首先，需要设置初始参数。针对该问题，设种群中麻雀的数目为 100，由 EM 算法得到的 100 组参数 θ 组成，每一只麻雀对应一组参数。发现者占种群总数的 20%。本实验设置迭代次数为 50 次。

因适应度函数值是评判个体好差的标准，故选择合适的适应度函数至关重要[12]。本研究计算出感兴趣区内每一个数据属于该高斯混合模型的概率，并将这些数据的概率求和，取相反数，作为适应度函数。其值越小，说明该高斯混合模型越能准确地表述重叠峰。适应度函数如式(8)

$$o(x) = -\sum_{m=1}^{M}\sum_{n=1}^{N} P(x(m) \mid z(n), \theta)a_m \tag{8}$$

通过 SSA 对 EM 算法得到的参数全局寻优，得到适应度函数值最小时对应的最优参数。其具体步骤如图 3 所示。

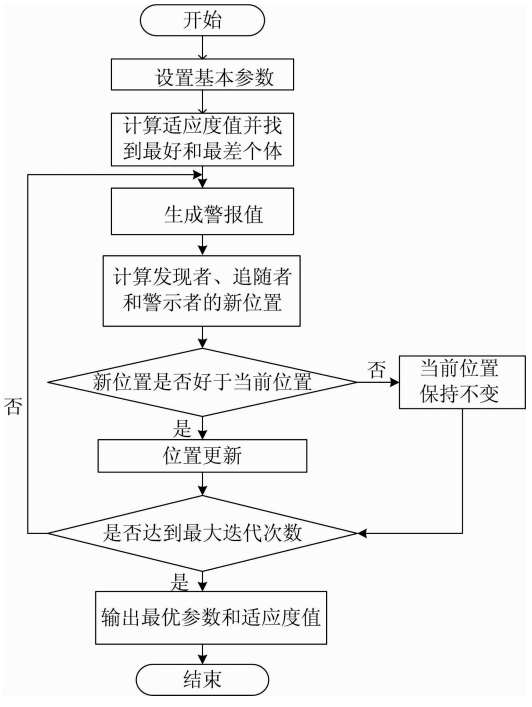

图 3 SSA 流程图

Fig. 3 SSA flow chart

4 重叠峰解析结果讨论

As 和 Pb 的重叠峰由三个子峰组成。用 SSA 对 EM 算法得到的参数全局寻优，可得到各子峰的最优参数。SSA 迭代过程中适应度值随迭代次数的变化曲线如图 4 所示。可看出，该算法收敛较快[13]。

由优化结果可得 Pb：600 mg·kg^{-1}，As：400 mg·kg^{-1} 的重叠峰分解示意图如图 5。

将最终的权重参数与重叠峰面积相乘得每一子峰面积，将其与单独元素的实际峰面积比较，得分解误差。重叠峰各子峰的实际峰位分别为：Pb-As-$K_{\alpha 2}$：10.507 keV，$L_{\alpha 1}$：10.549 keV，As-$K_{\alpha 1}$：10.543 keV，将其与分解得到的峰位比较，可得分解后峰位的误差。

仅用 EM 算法进行重峰解析时，设置不同初值，将得到不同的分解结果。以 Pb 含量 600 mg·kg^{-1}，As 含量 400 mg·kg^{-1} 的重叠峰为例，选取三组不同的初始值，分别代入 EM 算法迭代。已知该浓度下 Pb 元素的实际峰面积为 54.05，用 EM 算法分解并计算面积误差、峰位误差。再将 SSA 用于重叠峰分解，由误差结果可看出，仅用 EM 算法分解重叠峰时，设置不同初始值可能得到不同结果，且有时误差较大。进一步经过 SSA 优化后，可提高准确度，分解后特征峰位误差较小，几乎可以控制在 1%以内；峰面积最大误差为 6.4%，但随重金属浓度升高其误差有下降趋势，主要原因是当重金属元素浓度较低时，其光谱强度较弱，更易受外界因素干扰，而浓度较高时相对不易受干扰。仅用 EM 算法，和用进行全局寻优两种方法得到的峰面积误差如表 1、表 2 所示，峰位误差如表 3 和表 4 所示。

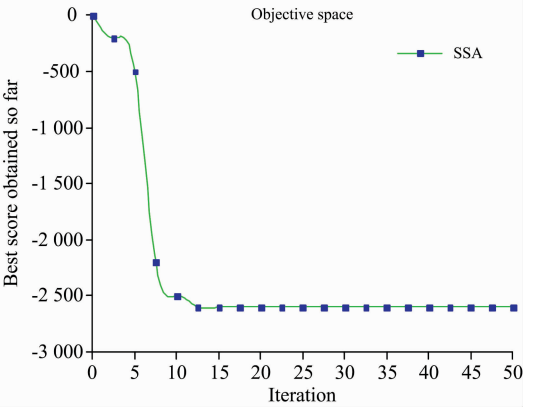

图 4 SSA 适应度值变化曲线

Fig. 4 Fitness value change curve of SSA

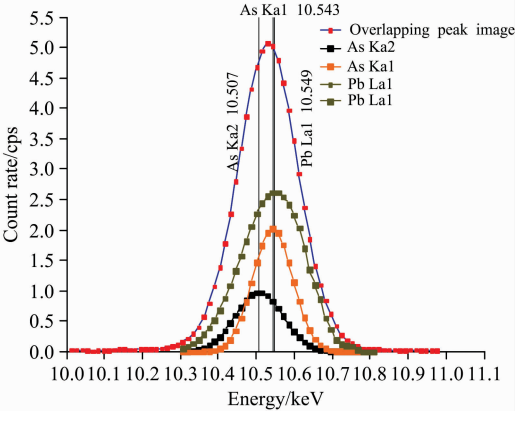

图 5 重叠峰分解结果示意图

Fig. 5 Diagram of overlapping peak decomposition results

表 1 EM 算法分解后 Pb 预测面积的误差率

Table 1 The error rate of Pb predicted area after EM algorithm decomposition

组号	预测面积	误差率/%
1	41.54	24.27
2	48.24	12.05
3	51.05	5.55

表 2　经 SSA 优化后 Pb 预测面积的误差率

Table 2　The error rate of Pb predicted area optimized by SSA

重金属含量/(mg·kg^{-1})	预测峰面积	实际峰面积	相对误差率/%
100	12.38	14.88	0.064
200	21.87	20.78	0.05
400	46.03	44.06	0.041
600	54.85	54.05	0.014

表 3　EM 算法分解后各特征峰位误差

Table 3　Position error of each peak after EM algorithm decomposition

重金属含量/(mg·kg^{-1})	Pb-$L_{\alpha1}$ 预测值	Pb-$L_{\alpha1}$ 误差率/%	Pb-$K_{\alpha1}$ 预测值	Pb-$K_{\alpha1}$ 误差率/%	Pb-$K_{\alpha2}$ 预测值	Pb-$K_{\alpha2}$ 误差率/%
第一组	10.607	0.55	10.525	0.17	10.443	0.61
第二组	10.608	0.56	10.517	0.25	10.418	0.85
第三组	10.552	0.03	10.488	0.52	10.386	1.15

表 4　SSA 优化后的特征峰位误差

Table 4　Position error of each peak after SSA optimization

重金属含量/(mg·kg^{-1})	Pb-$L_{\alpha1}$ 预测值	Pb-$L_{\alpha1}$ 误差率/%	As-$K_{\alpha1}$ 预测值	As-$K_{\alpha1}$ 误差率/%	As-$K_{\alpha2}$ 预测值	As-$K_{\alpha2}$ 误差率/%
Pb100-As200	10.552 9	0.036	10.533	0.09	10.453 9	0.5
Pb100-As400	10.571 8	0.2	10.537 6	0.05	10.448 3	0.55
Pb400-As600	10.608 8	0.56	10.526 1	0.16	10.442 6	0.61
Pb600-As400	10.542 1	0.06	10.67	1.2	10.511 8	0.045

5　结　论

通过光谱预处理、用 EM 算法初步解析重叠峰、SSA 对 EM 算法得到的参数全局寻优这几个步骤，最终完成了重叠峰解析。通过该研究，可得到如下结论：

(1)该法可较好完成重叠峰解析，解析后峰位误差基本上在 1%以内，误差较小。

(2)该法能避免 EM 算法易陷入局部最优的缺陷，且 SSA 比常见的优化算法收敛速度快、稳定性好、不易陷入局部最优。

(3)该法将新算法用于重叠峰解析领域，为重叠峰解析提供了新的思路，为其进一步研究提供了有效的参考。

References

[1] Shard A G, Wright L, Minelli C. Biointerphases, 2018, 13(6): 061002.

[2] Borba W, Silva J, Kemerich P, et al. Water, Air, & Soil Pollution, 2020, 231(4): 2727.

[3] ZHOU Shi-rong, HE Jian-feng, REN Yin-quan, et al(周世融, 何剑锋, 任印权, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(4): 1221.

[4] LIU Hong-li, HUANG Hong-quan, YANG Xi, et al(刘红莉, 黄洪全, 杨　熙, 等). Nuclear Electronics & Detection Technology(核电子学与探测技术), 2019, 39(1): 83.

[5] Liu M H, Dong Z R, Xin G F, et al. Chemometrics and Intelligent Laboratory Systems, 2018, 182: 1.

[6] Michael A, Zhou Y N, Yavuz M, et al. Thermochimica Acta, 2018, 665: 53.

[7] Xiong J Y, Liang W, Liang X B, et al. Process Safety Progress, 2020, 39: 1.

[8] WANG Xiu-yan, LIU Yan-min, ZHANG Ge-wen, et al(王修岩, 刘艳敏, 张革文, 等). Journal of System Simulation(系统仿真学报), 2018, 30(8): 3074.

[9] Nakashima S, Sughiyama Y, Kobayashi T J. Bioinformatics, 2020, 36(9): 2829.

[10] Zhao R F, Li Y Z, Sun Y K. Electronic Journal of Statistic, 2020, 14(1): 632.

[11] Xue J K, Shen B. Systems Science & Control Engineering, 2020, 8(1): 22.

[12] Li X J, Shao Z W, Cheng H P, et al. International Journal of Communication Systems, 2020, 38(8): e4370.

[13] CHEN Ying, ZHANG Can, XIAO Chun-yan, et al(陈　颖, 张　灿, 肖春艳, 等). Acta Optica Sinica(光学学报), 2020, 40(10): 180.

Overlapping Peak Analysis of Soil Heavy Metal X-Ray Fluorescence Spectra Based on Sparrow Search Algorithm

CHEN Ying[1], LIU Zheng-ying[1], XIAO Chun-yan[2], ZHAO Xue-liang[1, 3], LI Kang[3], PANG Li-li[3], SHI Yan-xin[3], LI Shao-hua[4]

1. Hebei Province Key Laboratory of Test/Measurement Technology and Instrument, School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
2. School of Resources and Environment, Henan Polytechnic University, Jiaozuo 454000, China
3. Center for Hydrogeology and Environmental Geology, China Geological Survey, Geological Environment Monitoring Engineering Technology Innovation Center of The Ministry of Natural Resources, Baoding 071051, China
4. Hebei Sailhero Environmental Protection Hi-tech Co., Ltd., Shijiazhuang 050000, China

Abstract In recent years, with the aggravation of soil heavy metal pollution and the gradual improvement of people's environmental awareness, the research on the rapid detection method of soil heavy metal content has been strengthened rapidly. At present, X-ray Fluorescence analysis (XRF) has been widely used to detect heavy metal pollution in soil. However, due to the limited energy resolution of the X-ray fluorescence spectrometer and the low fluorescence yield of some heavy metal elements, overlapping phenomena occurred in adjacent spectral peaks of some elements. In the cause of overlapping phenomenon often appears between adjacent peaks in X-ray Fluorescence analysis (XRF), a new overlapping peak analysis method based on Sparrow Search Algorithm (SSA) was proposed. Firstly, samples with different moisture content and heavy metal element content were prepared, and original spectral data were obtained by X-ray fluorescence spectrometer from the soil sampled of Baoding, Hebei. Then, the spectral data were preprocessed, the spectral clustering algorithm removed the abnormal spectral samples, the spectral denoising and background subtraction were completed by the Savitzky-Golay five-point quadratic denoising method and the linear background method. The random number method is used to generate a large number of simulated spectral data for the use of subsequent algorithms. After that, expectation-maximization (EM) was applied to analyze overlapping peaks preliminarily. Set the initial parameters of the EM algorithm, and put simulation spectra data into the EM algorithm. When it reached the maximum number of iterations, can preliminarily get parameters of the Gaussian Mixture Model (GMM), expectation, variance and weights of each Gaussian peaks. However, the EM algorithm is easily affected by the initial parameter and is prone to fall into the local optimum, leading to inaccurate results. Therefore, further optimization of the EM algorithm is needed. In this study, SSA was used for global optimization of parameters of the GMM. After setting the basic SSA algorithm parameters, 100 groups of parameters obtained by the EM algorithm were taken as the initial population of the algorithm, and then set appropriate fitness function. Finally, the optimal global parameters were obtained through iteration, and the decomposition of overlapping peaks was realized. Sparrow Search algorithm (SSA) is less affected by parameter setting. Compared with some traditional optimization algorithms, such as GA, ACO, PSO, etc. SSA has fast convergence speed and is not easy to fall into local optimal. Therefore, this algorithm can achieve better optimization results. The analysis of overlapping peaks shows that the algorithm can get more accurate results with fewer iterations and be widely used in energy spectrum overlapping peaks analysis.

Keywords X-ray fluorescence analysis; Gaussian mixture model; Expectation maximization; Sparrow search algorithm; Overlapping peaks analysis

(Received Jul. 6, 2020; accepted Nov. 20, 2020)