基于差分进化算法的 X 荧光重叠峰的分解

廖先莉^{1,2},黄进初^{1*},赖万昌¹,辜润秋¹,王广西¹,唐 琳²,翟 娟¹

. 成都理工大学核技术与自动化工程学院,四川成都 610059
 2. 成都大学信息科学与工程学院,四川成都 610106

摘 要 X射线荧光分析中相邻峰重叠的分解问题是十分常见的, 谱峰重叠为谱的进一步定性分析和定量 分析都带来了困难,而通过硬件手段来减少谱峰重叠的发生往往受资金和工作条件的制约,通常会选择通 过数学手段得到重叠谱中各个子峰的相关信息来完成重叠谱的分解。结合光谱形成过程的随机物理特性, 提出了一种基于高斯混合模型(GMM)的参数独立模型和参数关联模型,以及基于这两种模型和差分进化算 法的重叠峰分解方法。GMM 模型参数构成了差分进化算法个体基因,给出了目标函数的快速算法,通过随 机生成初始种群,以种群中每个个体的适应度值和各个个体参数的约束条件为选择标准,避免了初值不当 带来的局部收敛问题,并目将所有测量的随机数据参与到个体适应度值的运算当中,避免了原谱数据的损 失。对模型参数相互独立和模型参数相关联两种情况进行了解谱分析,首先,对三峰重叠和四峰重叠进行仿 真模拟分析,分解结果表明,基于 GMM 参数关联模型的解谱精度较 GMM 参数独立模型的解谱精度更高, 三峰重叠时,参数独立模型和参数关联模型分别得到的权重最大误差为8.15%和2%,峰位最大误差为 0.30%和0.06%,标准差的最大误差为7.5%和1.35%。四峰重叠时,参数独立模型和参数关联模型分别得 到的权重最大误差为 8.3%和 4.3%,峰位最大误差为 0.12%和 0.13%,标准差的最大误差为 5.04%和 0.45%。然后通过实测三峰重叠谱的解谱分析表明,用这两种模型进行重叠谱的分解,分解结果相对误差和 待测量元素的含量有关,随着待测元素含量的降低,分解结果精度会降低。仿真和实测都表明,基于高斯混 合模型和运用差分进化算法的重叠谱进行解谱时,如果能够提前得到各个相互重叠小峰权重、均值、标准差 之间的关系,建立 GMM 参数关联模型,减少寻优个体参数个数,对提高复杂峰的分解精度是非常重要的。

关键词 差分进化算法; GMM 参数独立模型; GMM 参数关联模型; 重叠峰的分解 中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2020)07-2301-06

引 言

对于谱峰重叠问题,一般用数学解析法进行重叠谱分 解,谱峰重叠数学分解方法的研究,对荧光谱进一步的定 量、定性分析都有十分重要的意义,现阶段已有不少相关的 研究报告^[1-3]。其中杨熙等提出了 GMM-SDR 模型和粒子群 算法相结合的重叠谱的解谱方法^[1];胡耀垓等运用曲线拟合 完成了光谱重叠峰解析方法^[2];徐喜荣等提出了一种基于小 波变换和连续 Hopfield 神经网络的谱图重叠峰解析策略^[3]; 目前还没有一种算法被公认为没有局限性的数学解谱方法, 比如曲线拟合度不够高,易陷入局部收敛,使用限制条件不 易满足。 本文在高斯混合统计模型的基础上,提出两种情况下的 参数模型,利用差分进化算法全局搜索优势,得到了重叠谱 的最优分解模型。两种模型下的解谱结果误差范围内都是有 效的,但是解谱精度却不相同,为类似数学解谱方法提供参 考,同时该方法的搜索速度快,种群规模对寻优结果的影响 比较小,结合了光谱的随机物理特性,保证的原谱数据的 "零损失"。

1 算法与原理

1.1 统计模型 GMM

一种融合了参数估计法和非参数估计法优点的修正模型,模型为

收稿日期: 2019-06-30, 修订日期: 2019-10-25

作者简介:廖先莉,1988年生,成都理工大学核技术与自动化工程学院博士研究生

基金项目:国家重点研发项目(2017YFC0602105)资助

(5)

$$P(x/\theta) = \sum_{i=1}^{M} a_i \frac{1}{2\pi^{1/2} \sigma_i} \exp\left[-\frac{1}{2} (x-u_i)^2 / \sigma_i^2\right]$$
(1)

其中 ai 为各分支的权重,且满足

$$\sum_{i=1}^{M} a_i = 1, \ a_i \ge 0 \tag{2}$$

式(1)中, M为分支数, u_i和 σ_i表示第 i 分支的均值和标准 差,由各个分支的权重、均值、标准差构成了差分进化算法 寻优体参数,所得的最优解,即为重叠峰分解后各个小峰的 参数。

如果不能提前确定 GMM 模型中参数间的关系,认为其 是相互独立的,建立模型为 GMM 参数独立模型,模型参数 为 $\theta = [a_1, a_2, \dots, a_M; u_1, u_2, \dots, u_M; \sigma_1, \sigma_2, \dots, \sigma_M]$ 。

如果能够知道重叠谱各个小峰参数间的关系,建立模型为GMM参数关联模型,可以通过这种关系减少差分进化算法寻优个体的参数个数,比如均值和标准差间存在线性关系 $\sigma_i = \sigma_1 u_i / u_1$,模型参数为 $\theta = [a_1, a_2, \dots, a_M; u_1, u_2, \dots, u_M; \sigma_1]$,下文中均以均值和标准差间存在的线性关系建立GMM参数关联模型。

1.2 GMM 模型参数的差分进化算法估计

差分进化算法是模拟自然界物种进化原理的一种寻优算法,算法通过对父代个体进行变异、交叉操作,生成新一代 个体,选择子代和父代个体中满足目标条件且适应度值最优 的个体作为本次寻优的结果,经过多级迭代,使搜索结果逐 渐趋近全局最优解^[4]。具体方法如下:

(1) 初始化种群

GMM 参数独立模型个体的维数为 3M, 个体中元素 a_i, u_i 和σ_i 在有限区间上随机生成,种群中粒子规模为 NP。

GMM 参数关联模型个体维数 2M+1,个体中元素 a_i , u_i 和 σ_i 在有限区间上随机生成,种群中粒子规模为 NP。整 个寻优过程需满足条件 $\sigma_i = \sigma_1 u_i / u_1$,其中 $i=2, 3, \dots, M_o$.

为了满足 GMM 模型中式(2)的条件,首先要对模型参数 *a_i* 做归一化处理,使得参数 *a_i* 的取值范围在 0 到 1 之间, 方便后续的运算。在对寻优得到的参数进行还原,方便得到的模型曲线和实际谱线对比。

(2) 适应度值的评估

种群中的每一个个体对应一个适应度值,适应度值由式 (3)算得,每一代种群中适应度值最小的个体为这一代种群 中的最优个体,多代寻优后适应度值不变时,搜索结束,得 到的参数 θ 为全局最优个体,即为我们所求的分解模型参 数。

$$Y(\theta) = -\sum_{j=1}^{n} g(j) \ln P(x(j)/\theta) = -\sum_{j=1}^{n} g(i) \sum_{i=1}^{M} a_{i} \cdot \frac{1}{2\pi^{1/2} \sigma_{i}} \exp\left[-\frac{1}{2} (x(j) - u_{i})^{2}/\sigma_{i}^{2}\right]$$
(3)

由式(3)可知, g(i)为随机序列(道址)对应的计数值, 由 随机序列 x(i)带入 GMM 模型得到的值取对数后值为负数, $P(x(j)/\theta)$ 越大 $f(\theta)$ 越大, $Y(\theta)$ 的值越小, 所得个体越优, 该方法充分使用了测量所得的所有随机数据, 保证了原谱数 据的"零损失"。

(3) 变异策略

采用式(4)的变异策略来更新种群个体,其中,V(*i*)表示变异后的第*i*个个体,X(*i*)为变异前的第*i*个个体,X(*r*(1))为变异前的第*r*(1)个个体,X(*r*(2))为变异前的第*r*(2)个个体,*r*(1)和*r*(2)是[0,NP]范围内产生了不等于*i*的整数,F为变异缩放因子。

$$V(i) = X(i) + F(\text{best}X - X(i)) + F(X(r(1)) - X(r(2)))$$
(4)

(4) 交叉策略

如式(5)所示,个体中的 X(i, j)经过变异策略后变为个 体 V(i, j),若[0,1]之间生成的随机数正好等于交叉概率 CR,或者 1 到 Dim 之间随机生成的整数 j_{rand} 等于 j 时,新生 成个体中的 U(i, j)等于 V(i, j),否则 U(i, j)为变异前的 X(i, j),该交叉策略确保了最新产生的个体中至少有一个U(i, j)=V(i, j)。

$$U(i,j) = \begin{cases} V(i,j), & \text{if}(\operatorname{rand}[0,1]) \leqslant CR \text{ or } (j = j_{\operatorname{rand}}) \\ X(i,j), & \text{otherwise} \end{cases}$$

(5) 选择策略

如式(6)所示,当经过交叉、变异后新产生的种群个体 所对应的适应度函数值小于原来个体所对应的适应度函数值 时,下一代种群这个位置上的个体变为最新产生的种群个 体,否则,下一代种群中这个位置上的个体保持不变。

$$X(i+1) = \begin{cases} U(i), & \text{if } Y(U(i)) \leqslant Y(X(i)) \\ X(i), & \text{otherwise} \end{cases}$$
(6)

(6) 终止条件

迭代次数满,或者最优适应度值 Y(θ)连续多次不变时, 算法终止,否则返回(3)继续搜索。

一直到满足终止条件得到最优个体,得到各个分支高斯 函数的权值、均值和均方差,完成重叠峰分解。

2 重叠峰的分解

采用离散直接抽样方法产生随机数值 x(1),x(2),…, x(N),由这些随机数的统计分布构成 X 荧光重叠峰,重叠 峰由几个峰位十分接近的高斯峰重叠而成^[5-7],本设计以重 叠谱解谱为重点,本底计数已扣除,下面将依次介绍两类重 叠峰的分解过程。

2.1 三峰分解

图 1 为三峰重叠谱,峰位为 210,200 和 190,面积分别 为 25 000,15 000 和 10 000,即归一化后权重分别为 0.5, 0.3 和 0.2,横坐标为道址,纵坐标为计数。

采用 GMM 参数独立的模型,设种群的个体为 $X = [a_1, a_2, a_3; u_1, u_2, u_3; \sigma_1, \sigma_2, \sigma_3]$,各参数的取值下限[0.01, 0.01, 0.01; 160, 160, 160; 4, 4, 4],取值上限[1, 1, 1; 250, 250, 250; 8, 8, 8],选择取值范围内的随机数初始化种群,种群规模 50,最大迭代次数 1000,缩放因子 CR 为 0.1 和交叉概率 F 为 0.1。采用 GMM 参数关联的模型,则提前知晓均值和标准差间的线性关系 $\sigma_i = \sigma_1 u_i / u_1$,种群个体为 $X = [a_1, a_2, a_3; u_1, u_2, u_3; \sigma_1]$,取值范围从[0.01, 0.01, 0.01; 160, 160, 160; 4]到[1, 1, 1; 250, 250, 250; 8],





Fig. 1 Three overlapping peak

如表 1 可知,参数独立模型和参数关联模型分别得到的 权重最大误差为 8.15% 和 2%,峰位最大误差为 0.30% 和 0.06%,标准差的最大误差为 7.5% 和 1.35%。

	表 1	权重、峰位和标准差的比较				
Table1	Comparision of weight, peak position an					
	standard deviation (\mathbf{I})					

GMM 模 型参数	原始值	独立 寻优值	相对误 差/%	关联 寻优值	相对误 差/%
a_1	50%	46.66%	-6.68	49.83%	-0.36
a_2	30%	31.71%	5.70	29.77 $\%$	-0.77
a_3	$20\frac{0}{0}$	21.63%	8.15	20.40%	2.00
u_1	210	210.56	0.27	209.95	-0.02
u_2	200	200.59	0.30	200	0.00
u_3	190	190.01	0.01	190.12	0.06
σ_1	7	6.71	-4.14	6.91	-1.32
σ_2	6.67	6.17	-7.50	6.58	-1.35
σ_3	6.33	6.23	-1.58	6.25	-1.19





合情况。已经能够明显观察到运用 GMM 参数关联模型对重 叠谱进行分解的精度比运用 GMM 参数独立模型对重叠谱进 行分解的精度高。



2.2 四峰分解

如图 4 所示,为四峰重叠谱,各子峰的峰位分别为 200, 210,225 和 240,峰面积分别为 20 000,30 000,30 000 和 20 000,即归一化后权重分别为 0.2,0.3,0.3,0.2,重叠严 重。



设置种群个体 50, 如果选择 GMM 独立参数的模型, 则 每个种群个体有 12 个参数, 它们的取值范围从[0.01, 0.01, 0.01, 0.01; 175, 175, 175; 4, 4, 4, 4]到[1, 1, 1, 1; 265, 265, 265, 265; 8, 8, 8, 8], 如果能够提前知道重叠谱 参数间的关系,得到 GMM 的相关联参数模型,比如均值和 标准差间的线性关系 $\delta_i = \delta_i u_i / u_1$,则每个个体参数 10 个,它 们的取值范围从[0.01, 0.01, 0.01, 0.01; 175, 175, 175, 175; 4]到[1, 1, 1, 1; 265, 265, 265, 265; 8],两种模型寻 优的迭代次数为 1 000, 交叉概率 F = 0.8, 缩放因子 CR= 0.8。

如表 2 可知,参数独立模型和参数关联模型分别得到的 权重最大误差为 8.3%和 4.3%,峰位最大误差为 0.12%和 0.13%,标准差的最大误差为 5.04%和 0.45%。

表 2 权重、峰位、均方差的比较 Table 2 Comparision of weight, peak position and standard deviation (Ⅱ)

GMM 模	百払店	独立	相对误	关联	相对误				
型参数	尿炉阻	寻优值	差/%	寻优值	差/%				
a_1	20%	18.88%	5.60	19.14%	-4.30				
a_2	30%	32.49%	-8.30	30.59%	1.97				
a_3	30%	27.80%	7.33	30.23%	0.77				
a_4	20%	20.83%	-4.15	20.03%	0.15				
u_1	200	199.76	0.12	199.74	-0.13				
u_2	210	210.10	-0.05	209.84	-0.08				
u_3	225	225.06	-0.03	224.84	-0.07				
u_4	240	239.89	0.05	240.14	0.06				
σ_1	6	6.05	-0.83	6.02	0.26				
σ_2	6.3	6.54	-3.81	6.32	0.32				
σ_3	6.75	6.41	5.04	6.77	0.32				
σ_4	7.2	7.31	-1.53	7.23	0.45				

如图 5 和图 6 所示,分别为通过独立 GMM 模型和相关 联 GMM 模型,对原始重叠峰进行分解后的各峰拟合情况。 已经能够明显观察到运用 GMM 参数关联模型对重叠谱进行 分解的精度比运用 GMM 参数独立模型对重叠谱进行分解的 精度高。



图 5 原始谱、GMM 参数独立曲线、分解峰

Fig. 5 Original spectrum, Parameters of independent GMM curve, Decomposition peak





每次寻优分解的效率和搜索结果的精度不尽相同,略有 差异,相对而言,GMM 参数关联模型运用差分进化算法搜 索结果更加稳定,精度也相对较高。

2.3 重叠谱峰的分解实例

取深井中的岩屑作为基样,加入镝粉,磨匀配置镝元素 含量分别为20,5和2μg·g⁻¹的三种样品,由于岩屑中本身 含有 Mn 和 Fe 元素,当加入 Dy 元素后,三种元素的部分特 征 X 射线会发生谱重叠现象,如图 7 所示,为三种样品用 X 荧光仪测得的全谱图。



Mn 元素的 Kβ 系的特征 X 射线能量为 6.49 keV, Fe 元 素 Kα 系的特征 X 射线能量为 6.403 keV, Dy 元素 Lα 系的 特征 X 射线能量为 6.495 keV, 由于 Dy 元素含量比较低, 为 了能够更好的分析谱重叠情况, 对图进行局部放大如图 8 所 示, 为三种元素的重叠峰。



Fig. 8 The measured three peaks overlapping spectra

将 X 荧光仪实测谱线重叠峰相关数据取出扣除本底后导入 matlab 中,根据建立的独立 GMM 模型,对重叠谱进行分解, Dy 元素含量为 20 μ g · g⁻¹的样品,分解后各峰的情况 如图 9 所示。

独立 GMM 模型对三种样品的重叠谱进行分解,得到 Dy 元素的含量分别为 19.10,5.27 和 2.27 μ g • g⁻¹,相对误 差分别为-4.5%,5.4%和 13.4%。可见用该种方法进行重 叠峰的分解时,随着元素含量降低,分解得到的结果相对误 差增大。

在已知所求元素种类的情况下,可以提前已知峰位间的 关系,建立相关联 GMM 模型, Dy 元素和 Mn 元素的特征 X 射线能量差不多, Dy 元素 L α 系的特征 X 射线能量为 6.495 keV,可得式(1)中的 $u_1 = u_2$ 和 $u_3 = 0.986$ $6u_1$,模型中参数

降低了。

个数由 9 个变成了 7 个,根据建立的相关联 GMM 模型,对 重叠谱进行分解, Dy 元素含量为 20 μ g • g⁻¹的样品,分解后



图9 原始谱、GMM 参数独立曲线、分解峰



相关联 GMM 模型对三种样品的重叠谱进行分解,计算 得到 Dy 元素的含量分别为 20.18,4.85 和 2.15 μ g · g⁻¹, 相对误差分别为 0.9%, -3%, 7.5%,可见该方法解谱得到 的结果相对误差比独立 GMM 模型解谱得到结果的相对误差 小一些,元素含量越低,脉冲计数越少,分解得到的结果相 对误差越大。

由实验可知,两种方法进行重叠谱分解时,能够知道各 峰间的关系,建立相关联 GMM 模型,减少参数个数,分解 后的精度更高,但是随着元素含量的降低,分解测量的精度

$\times 10^4$ 5 原始峰 GMM关联重叠峰 4 Dy峰 Mn峰 Fe峰 3 Counts 2 1 310 330 335 305 315 320 325 Address 图 10 原始谱、参数关联 GMM 谱、分解谱 Fig. 10 Original spectrum, Parameters of relevance GMM curve, Decomposition peak

3 结 论

运用文中方法对重叠峰的分解结果表明,两种模型下,均能实现重叠谱的分解,从模拟仿真可知,针对相对复杂的 重叠峰两种模型得到的分解结果精度都较高,但是 GMM 参 数关联模型的分解精度比 GMM 参数独立模型的分解精度明 显高一些。从三峰重叠的分解实例来看,由于实际测量过程 中对测量结果的影响因素相对复杂一些,分解计算得到的结 果较仿真计算结果相对误差大一些,两种方式的结果表明, 如果能够提前得到各个相互重叠小峰之间的关系,建立相关 联 GMM 模型,减少寻优参数个数,对提高复杂峰的分解精 度是非常重要的。

References

- [1] YANG Xi, HUANG Hong-quan, JIANG Kai-ming, et al(杨 熙,黄洪全,蒋开明,等). Spectroscopy and Spectral Analysis(光谱学与 光谱分析), 2017, 37(8); 2376.
- [2] HU Yao-gai, ZHANG Xiao-xing, ZHAO Zheng-yu, et al(胡耀垓,张晓星,赵正予,等). Journal of Chongqing University(重庆大学学报), 2012, 35(5): 76.
- [3] XU Xi-rong, PAN Zi-qi, LI Xing-hua, et al(徐喜荣,潘子琦,李兴华,等). Journal of Dalian University of Technology(大连理工大学学报), 2019, 59(1): 97.
- [4] Qin A K, Huang V L, Suganthan P N. IEEE Transactions on Evolutionary Computation, 2009, 13(2): 398.
- [5] Qiu Jingke, Deng Biao, Yang Qun, et al. Nuclear Instruments and Methods in Physics Research B, 2011, 269(22): 2662.
- [6] Angeyo K H, Gari S, Mustapha A O, et al. Applied Radiation and Isotopes, 2012, 70(11): 2596.
- [7] Sabriye Seven. Radiation Phyics and Chemistry, 2012, 81(5): 489.

Based on Differential Evolution Algorithm X Fluorescence Peak Overlapping Decomposition

LIAO Xian-li^{1, 2}, HUANG Jin-chu^{1*}, LAI Wan-chang¹, GU Rui-qiu¹, WANG Guang-xi¹, TANG Lin², ZHAI Juan¹

College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu 610059, China
 School of Information Science and Engineering, Chengdu University, Chengdu 610106, China

Abstract X-ray fluorescence analysis of adjacent peaks overlapping decomposition problem is very common, spectrum peaks overlapping spectrum for further qualitative analysis and quantitative analysis are brought difficulties, and by means of hardware to reduce the spectral peaks overlapping often occurs the restriction of the capital and the working conditions, will often go on the overlapping spectrum is obtained by mathematical means of relevant information to complete the overlap of each peak spectral decomposition. This paper proposes a model GMM parameters of the independent model and GMM parameters correlation model based on the gaussian mixture (GMM), based on these two models and differential evolution algorithm of the overlapped peaks decomposition method, GMM model parameters constitute the individual genes differential evolution algorithm, presents a fast algorithm for target function, through the randomly generated initial population, In the fitness value of each individual in a population and the constraint conditions of each individual parameter as selection criteria, avoids the local convergence of the problems of the improper initial value, and all the measurement of random data involved in the operation of individual fitness value, avoid the loss of the original spectral data. Respectively for independent parameter model and the parameters associated with the model to understand the spectrum analysis, two cases through the decomposition of three kinds of overlapping spectra show that the model based on two kinds of differential evolution algorithm for the overlapping peaks decomposition is effective. First of all, the three peak simulation analysis and four peaks overlap decomposition results show that the spectral accuracy based on GMM parameters associated model spectrum GMM parameters are independent of the model solution precision is high. Three peaks overlap, parameters independent model and correlation model respectively to get the weight of maximum error is 8.15% and 2%, a maximum error of 0.30% and 0.06%, the standard deviation of the maximum error is 7.5% and 1.35%. Four overlapping peaks, parameters independent model and correlation model respectively to get the weight of maximum error is 8.3% and 4.3%, a maximum error of 0.12% and 0.13%, the standard deviation of the maximum error is 5.04% and 0.45%. Then through measured three peaks overlapping spectra of the solutions of spectrum analysis shows that with this two kinds of model of overlapping spectrum decomposition, decomposition results relative error and the measuring element content is about, with the loss of the element under test content, the decomposition results in accuracy is reduced. Simulation and measurement show that. Using differential evolution algorithm based on gaussian mixture model and overlapping spectra for solution spectrum, if you can get ahead of the small overlapping peak weight, mean, standard deviation, the relationship between the GMM parameters correlation model is set up, and decrease the number of optimization of individual parameters to improve the accuracy of the breakdown of the complex peak is very important.

Keywords Differential evolution algorithm; GMM parameters of the independent model; GMM parameters correlation model; Decomposition of the overlapping peaks

(Received Jun. 30, 2019; accepted Oct. 25, 2019)

* Corresponding author