基于进化策略的 CHC 遗传算法及岩性波谱识别

张振飞，胡光道，杨明国

（中国地质大学数学地质遥感地质研究所，湖北武汉 430074）

摘要：野外实测岩性波谱数据的数据挖掘可以为高光谱遥感建模提供依据. 针对实测波谱数据的特点，设计了一种基于 Monte Carlo 抽样进化机制的 CHC（cross generation elitist selection, heterogeneous recombination, cataclysmic mutation，跨世代精英选拔、异物种重组、灾变变异）遗传算法用于多类岩性判别. 应用于云南北衙金矿蚀变岩的识别，表明该方法具有快速高效性.

关键词：CHC 遗传算法；进化策略；岩性波谱识别；北衙金矿；云南.
中图分类号：P628 **文献标识码**：A
文章编号：1000－2383(2003)03－0351－05

作者简介：张振飞(1961－)，男，副教授，2001 年毕业于中国地质大学研究生院，获地球探测与信息技术专业博士学位，主要从事数学地质及矿产勘查研究.

0 引言

CHC GA（cross generation elitist selection, heterogeneous recombination, cataclysmic mutation, genetic algorithm）是 Eshelman[1] 1991 年提出的一种对传统遗传算法的改进算法，它有 3 个特点：跨世代精英选拔（将父代和子代个体混合起来选择新一代个体）、异物种重组（当 2 个父代个体有充分差异时才进行交叉，否则该 2 个体之间不交叉）和灾变变异（随机选择部分个体进行完全初始化，而不是个体个别位值的翻转）. 进化策略（evolutionary strategies, ES）是进化计算的一种[2]，它通过搜索方向和步长的自适应调节，直接在解空间上进行交叉、变异等操作.

随着高光谱遥感的兴起，光谱数据挖掘成为近年来的一个研究热点[3]. 野外实测岩性波谱数据与高光谱遥感数据虽然随测量仪器设备不同而在波长分辨率、波段范围、光谱变质情况等方面会有一定差异，但其基本内容方面是一致的. 前者对于岩性识别具有高度可靠性，是建立高光谱遥感地面模型的重要依据. 我们利用美国 ASD 公司生产的野外光谱仪 FieldSpectr Fr，在云南北衙金矿获取了不同岩性露头的一批波谱数据. 本文以这些数据为试验对象，将 ES 和 CHC 结合起来并加以改造，设计了一种适于岩性光谱识别的独特的遗传算法，显示了很好的应用效果.

收稿日期：2002－10－08
基金项目：国家自然科学基金项目（No. 40272122）；国土资源部重点攻关项目（No. 20010305）.

1 野外光谱数据及岩性识别数学模型

1.1 数据

FieldSpectr Fr 野外光谱仪测量的数据是岩石露头（及其他地物）在 350～2 500 nm 范围内各波长点上的太阳光反射率与白色标准板反射率的比值. 在可见光－近红外波段（350～1 050 nm）取样间隔为 0.7 nm，光谱分辨率 3 nm；在短波红外线（900～2 500 nm）部分，采样间隔为 2 nm，光谱分辨率为 10～12 nm. 仪器对数据自动进行插值处理，使全部取样间隔都表现为 1 nm. 在约 1 350～1 420 nm 和 1 850～1 920 nm 2 个大气水吸收带，标准白板和地物反射率都很低，其比值波动剧烈、信噪比难以掌握，故分析时避开了这些波段. 由于野外光谱测量中影响因素很多，波谱数据一般只有相对准确性，光谱曲线的高低不能准确反映岩石露头实际反射率大小，而曲线形态反映不同波长上反射率相对大小比较可靠. 为了弱化绝对值影响而突出曲线形态的作

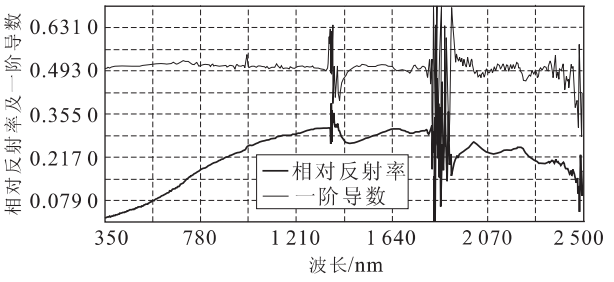

图 1 深灰色白云质灰岩露头的相对反射率及其规格化一阶导数曲线

Fig. 1 Spectrum curves of relative reflectance and its normalized one-order differential of a dark gray dolomitic limestone outcrop

用,笔者采用求一阶、二阶导数和相对吸收指数的方法进行数据预处理. 例如一个典型样品的相对反射率及其一阶导数曲线(图 1).

1.2 数学模型

为进行岩性识别,构造包含一个线性判别函数和一个判别得分阈值向量的多类判别模型. 设共有 S 个已知样品(岩石露头),分为 G 类. 线性判别函数及阈值向量为:

$$f_i = c_0 + \sum c_j x_j,\ \boldsymbol{T} = \{t_k,\ k = 1,2,\cdots,G-1\}. \tag{1}$$

其中 x_j 为每个波长点上的反射率比值一阶导数(也可为其他预处理结果),c_j 和 c_0 是系数和常数项. 阈值向量 $\boldsymbol{T}$ 将实数值域分为 G 个区间,每一类对应一个区间,使各已知样品的目标值 f_i 尽量落在它们所属的类所对应的区间内. 为了获得 $\boldsymbol{T}$、c_0 和 c_j,采用以下 2 种优化方案:(1)直接以已知样品的判对率最大为优化目标. (2)以距离比指标 I 尽量大为优化目标. I 的计算如下:

$$I=[2\overline{D}_{\mathrm{Inter}G}-(D_{\mathrm{Inter}G,\ \max}-D_{\mathrm{Inter}G,\ \min})]/\overline{D}_{\mathrm{In}G}. \tag{2}$$

式中 $\overline{D}_{\mathrm{Inter}G}$ 表示目标值 f 的平均类间距离;$D_{\mathrm{Inter}G,\ \max}$ 和 $D_{\mathrm{Inter}G,\ \min}$ 表示组间距离的最大和最小值;$\overline{D}_{\mathrm{In}G}$ 为组内离差平方和的平均值. 当 I 值达到足够大时,各类样品的目标值将分别集中于其重心附近,而不同类之间距离将会拉开,此时,按重心值排序后,取各对相邻类重心的中值为临界值,可以期待较好的判对率. 这里 I 比一般 Fisher 准则[4] 多考虑了组间距离的极差. 在多于两类判别的情况下,考虑到类间距离的极差有助于在优化过程中使类间距离趋向均匀,避免使其中有些类之间相距远到无必要的程度,而另一些类之间很近以致无法区分.

2 遗传算法设计

遗传算法的一般过程是,在一系列算法设置如编码方法、染色体定义、群体规模、适应度定义、搜索终止条件等之后,按照"生成初始种群→计算适应度→终止条件判断→选择→交叉→变异→计算适应度"的步骤循环搜索[5]. 笔者针对岩石波谱数据特点,各环节设计如下.

2.1 算法设置和初始种群

由于变量很多(在剔除大气水吸收带后仍有约 2 000 个),故为典型的高维搜索,此时采用一般的二进制编码是不合适的[6]. 笔者采用实数编码,根据前述 2 种优化方案来定义适应度、染色体(个体)和基因:当直接以判对率为优化目标时,判对率即为适应度;以距离比为优化目标时,I 值为适应度. 2 种方案均以系数向量和相应的阈值向量共同构成染色体,这些向量的实值元素为基因. 群体规模(用 p 表示)对搜索效率有影响,本文的大量试算表明,$p=120$ 左右是较好的选择. 终止条件除最大世代数、最长搜索时间外,2 种方案均当达到满意判对率时终止. 这是因为当以距离比为适应度时,常在 I 值尚未达到其极值前的某些世代,各类间距离已充分变大,类内离差也已充分缩小,达到了最好判对率,此时的多类判别函数已具备了准确判别的效能,不必继续搜索(若继续搜索,很可能使判对率回落);另外,I 的极大值不可预知,不能用 I 确定适当的满意适应度. 初始种群用一定区间内的随机数来产生:设用符号 $R(a,\ b)$ 表示区间 $(a,\ b)$ 上均匀分布的随机数,r 为任意选择的 10.0～100.0 之间的一个正数,我们用 $R(-r,\ r)$ 来初始化每个基因.

2.2 判对率的计算

无论采用上述哪种适应度定义,计算判对率之前都需进行局部搜索以获取最佳的阈值向量. 对应于每个个体的系数向量 $\boldsymbol{C}$,求出所有已知样品的判别目标值并求出各类重心,将类重心从小到大排序,设排序后 2 个相邻类重心为 a_k 和 a_{k+1},生成 p 个随机数 $t_{ki}=R(a_k,\ a_{k+1})$ 作为阈值向量中第 k 个候选阈值,从而得到阈值向量的群体,选其中判对率最大者为本世代与 $\boldsymbol{C}$ 关联的阈值向量.

2.3 选择

选择的具体方法因下述交叉算子的不同而不同. 若采用"均匀分布交叉",直接从父代群体中选择最好(适应度最大)的 $p/2$ 个体;若采用"高斯分布交叉",

则只选择 2 个父代个体,一个是最好的个体,另一个代表本世代平均水平(按适应度排序后取第 $p/2$ 个).

2.4　交叉

不同于一般遗传算法中截取交换位串的交叉方法,我们通过 Monte Carlo 抽样来实现交叉. 交叉操作仅作用于系数向量(而阀值向量则如上述,通过局部搜索得到),方法有 2 种:

(1)均匀分布交叉. 将所选出的 $p/2$ 个个体随机两两配对,每对生 4 个孩子,取代 2 个父代个体,最终得到 p 个子代个体. 用符号 $\max(a, b)$ 和 $\min(a, b)$ 分别表示 2 个数 a 和 b 中较大和较小者,生成子代个体方法如下:设 2 个父代个体的第 j 个基因分别为 a_j 和 b_j,整个父代种群中最好个体的第 j 个基因为 v_j,欲产生的子代个体的第 j 个基因为 x_j,按下式生成 x_j:

$$x_j=\begin{cases}R(\max(a_j, b_j)-h, v_j+h),\\ \quad 当\ \max(a_j, b_j)\leqslant v_j;\\ R(v_j-h, \min(a_j, b_j)+h),\\ \quad 当\ \min(a_j, b_j)\geqslant v_j;\\ R(v_j-h, v_j+h),\\ \quad 当\ \min(a_j, b_j)<v_j<\max(a_j, b_j).\end{cases} \tag{3}$$

其中 $h=|a_j-b_j|/2.0$.

(2)高斯分布交叉. 用 $N(m, d)$ 表示均值为 m,标准差为 d 的正态分布随机数. 设所选出的最好父代个体的第 j 个基因 a_j,代表父代种群中平均水平的个体的第 j 个基因为 b_j,欲产生的子代个体的第 j 个基因为 x_j,按下式生成 p 个 x_j:

$$x_j=N(a_j, \max(|a_j-b_j|, r/2)). \tag{4}$$

以上 2 种交叉算子都通过调整搜索区间而体现了算法的自适应进化,同时,每个世代每个基因都会在不同的(可能重叠的)区间内进行搜索,由此体现"异物种重组". 这里借鉴了 Georges 等[7] 1999 年提出的压缩遗传算法(compact GA)中独立处理每个基因的思路,实验证明效率很高.

2.5　变异

采用如下的灾变变异策略:从交叉后产生的子代群体中随机选择 $R(1, p/4)$ 个个体进行变异. 设欲生成的个体的第 j 个基因为 x_j,整个父代群体中第 j 个基因的最大值和最小值分别为 $c_{j,\max}$ 和 $c_{j,\min}$,则 $x_j=R(c_{j,\min}-r/2, c_{j,\max}+r/2)$. 这意味着在更大的区间内随机初始化.

2.6　跨世代精英选拔

为确保每代最好个体不差于父代,随机选一个子代个体,用父代最好个体取而代之. 这样可能恰好去掉了一个比父代最好个体还好的子代个体,但这种风险不大,因为如果群体规模 $p>100$,则该事件发生的概率将小于 0.01.

2.7　算法流程

2 种交叉方法和 2 种适应度定义可相互组合形成 4 种算法流程:①均匀分布重组—判对率;②高斯分布重组—判对率;③均匀分布重组—距离比;④高斯分布重组—距离比. 这些流程的其他环节一致.

3　应用及结论

以云南北衙金矿笔架山 18 个露头波谱岩性识别为例,共 3 类岩性,故分为 3 类判别问题. 选每类岩性的部分样品为已知样品,留一部分假设为"未知"样品,用于检验识别效果. 数据预处理方法为求一阶导数,选 350～1 320,1 450～1 820 和 1 920～2 400 nm 共 1 823 个波长点(变量)参与计算. 取群体规模 $p=120$,种群初始化时取 $r=50.0$,满意判对率 0.95. 4 种流程的进化曲线如图 2,岩性识别结果如图 3.

由图 2 可见,流程①、②在第 1 和第 2 世代时即达到搜索目标,流程③、④分别在第 7 代和第 14 代时达到搜索目标(已知样品判对率达 100%,超过满意判对率). 运行效率都很高(运行时间最长的流程④约需 1s),但总的来说,直接以判对率为适应度比以距离比为适应度要好,前者可以更充分地体现遗传算法黑箱式自适应进化的特点;均匀分布交叉比高斯分布交叉效果好,后者较多地强调了每一世代父代群体中最好个体的利用,这在进化的初期是不

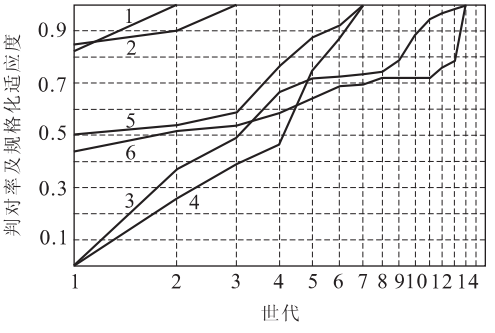

图 2　4 种流程的进化曲线

Fig. 2　Evolutionary curves of the four algorithms

1. 流程①进化曲线;2. 流程②进化曲线;3. 流程③进化曲线;4. 流程④进化曲线;5. 流程③判对率曲线;6. 流程④判对率曲线