文章编号:1001-6112(1999)02-0170-05

基于遗传算法的模糊模式分类法在油样属性分析中的应用

陈 锋 胡上序 俞蒙槐
(浙江大学智能信息研究所,杭州 310027)

摘要:本文采用模糊模式识别和遗传算法相结合,对塔里木盆地油样属性进行了比较。用遗传算法确定最优聚类中心和权向量,并确定每种油样的隶属度。对 20 种油品的研究结果表明,此方法性能良好,对油样属性作出定量的描述,是油样属性分析的一种新手段。

关键词:遗传算法;模糊模式分类法;油样属性分析
中图分类号:TE19 **文献标识码**:B

油样属性分析是地球化学研究的一个重要课题,它直接服务于含油气盆地的油气远景评价及油气田的勘探开发。通过研究可以了解含油气盆地中的石油、天然气、源岩之间的成因关系,油气运移的方向和距离以及油气的次生变化,从而进一步圈定可靠的油源区,确定勘探目标,有效地指导油气的勘探和开发工作。尤其在含油构造为断层切断的生产区和多油层油田的开发中,还可提供不同断块和不同生产层之间联通情况的信息。这在油气勘探上有着深远的意义。

塔里木盆地自震旦纪以来经历了 6 个大的演化阶段,其地质构造复杂,造成了盆地的多油源性。而且由于油气迁移的影响,不同油源的原油也会发生"混合",使来源于不同生油层的原油的原始地球化学特征发生变化,这种现象使原油的来源在某种程度上呈现模糊现象,原油之间的可比性有所下降。因此,根据原油的物理性质判断其形成环境和来源往往有一定困难,用通常的方法只能区分原油的类别,不易定量地知道原油的"混合"程度。用本文的方法能较好地解决这个问题,并得到更多的信息。

模糊模式识别[1~3]是当前模糊数学应用的重要领域之一。由于原油的"混合"现象,基于模糊模式区别方法识别和预测样品油的种类,定量地确定原油的类别,提高正确判断原油的形成环境和来源的依据,有效地指导勘探工作。

遗传算法[4](GA)是一种模拟生物群体和进化机理的优化算法,到目前为止,已在许多领域得到应用。遗传算法对于解决优化问题,如条件选择等,有很多优势。特别是具有很高的搜索次序且搜索具有探索性和自进化能力。基本的遗传算法包括交叉、选择、变异操作。

本文将 GA 引进模糊识别系统,保留遗传算法的全局搜索能力,提出通过 GA 操作算子——交叉、突变、选择算子,确定模糊模式识别的最优聚类中心模糊矩阵及权向量,并用模糊分类方法确定最优模糊分类隶属度矩阵,这种方法具有自进化能力。

1 基于遗传算法的模糊模式分类

设有待分类的 n 个样品组成的样品集合

$$[X]=\{X_1,X_2,\cdots,X_j,\cdots,X_n\} \quad (1)$$

每个样品有据以分类的 m 个指标特征

$$X_j=\{X_{1j},X_{2j},\cdots,X_{3j},\cdots,X_{mj}\}^T \text{ 其中 } j=1,2,3,\cdots,n \quad (2)$$

则样品集可用矩阵表示:

$$[X]=(X_{ij})_{m\times n} \text{ 其中 } i=1,2,3,\cdots,m;j=1,2,3,\cdots,n \quad (3)$$

收稿日期:1998-05-10
作者简介:陈锋(1965—),男(汉族),浙江象山人,博士,主要从事工业自动化研究。

由于样品 m 个特征的物理量不尽相同，在进行分类时，要先消除各特征的物理量量纲的影响，必须使特征值规格化，经常采用的规格化公式为：

$$r_{ij}=(X_{ij}-X_{i,\min})/(X_{i,\max}-X_{i,\min}),(\text{用于越大越优指标}) \quad (4A)$$

$$r_{ij}=(X_{i,\max}-X_{ij})/(X_{i,\max}-X_{i,\min}),(\text{用于越小越优指标}) \quad (4B)$$

式中：r_{ij} 为规格化特征值；X_{ij} 为样品特征值；$X_{i,\max}$、$X_{i,\min}$ 分别为 n 个样品第 i 特征的最大值、最小值。

矩阵(5)为经规格化后变换为元素在[0,1]区间的模糊特征矩阵：

$$[R]=(r_{ij})_{m\times n}\text{其中 } i=1,2,\cdots,m;\ j=1,2,\cdots,n.\ 0\leqslant r_{ij}\leqslant 1 \quad (5)$$

设将容量为 n 的样本分为 C 类，样品的每一类都有一个中心位置，称为聚类中心，描述该聚类中心位置的聚类中心矩阵表示为：

$$[S]=(S_{ih})_{m\times c} \qquad 0\leqslant S_{ih}\leqslant 1 \quad (6)$$

S_{in} 表示第 h 类聚类中心特征 i 的聚类中心值，$i=1,2,3\cdots,m$；$h=1,2,3,\cdots c$。第 h 聚类中心的 m 个特征以向量表示为：

$$\vec{S_h}=(S_{1h},S_{2h},,\cdots,\ S_{3h},\cdots,\ S_{mh})^T \quad (7)$$

将 n 个样品依据 m 个特征划分为 C 类，设类别隶属度矩阵为：

$$[U]=(u_{hj})_{c\times n} \quad (8)$$

U_{hj} 为样品 j 从属于第 h 类的相对隶属度，矩度(8)应满足条件

$$0\leqslant u_{hj}\leqslant 1,\qquad \sum_{h=1}^{c}u_{hj}=1 \quad (9)$$

由于每个特征对分类的影响不同，故有不同的权重，其权向量为：

$$W=(W_1,W_2,\cdots,W_m)\qquad \sum_{i=1}^{m}W_i=1 \quad (10)$$

这样，设第 j 个样品与第 h 类中心的差异用广义距离

$$d_{hj}=\|W\|r_j-\vec{S_h}\|\|=\sqrt[p]{\sum_{i=1}^{m}[W_i(r_{ij}-s_{ih})]^p} \quad (11)$$

表示。式中 p 为距离参数，在本文中取 2。

为求解最优类别隶属度矩阵，本文提出的目标函数是使全体样本对于各类中心之间的加权广义距离平方和最小，可以求得样品 j 属于中心模式 h 的隶属度定义为[2]：

$$u_{hj}=\frac{1}{\sum_{k=1}^{c}\left[\frac{\sum_{i=1}^{m}[w_i(r_{ij}-s_{ih})]^2}{\sum_{i=1}^{m}[w_i(r_{ij}-s_{ik})]^2}\right]} \quad (12)$$

这就是求解类别隶属度矩阵的公式(12)，本文利用遗传算法的全局搜索性求取模糊聚类中心位置。对于遗传算法产生的每一个体的类别隶属度矩阵，用遗传算法产生和进化模糊聚类中心矩阵，产生新的聚类中心群体，然后用公式(12)求得新的类别隶属度矩阵，根据已知样本作为训练集，我们取遗传算法的适应度函数为：

$f(X)$ = 根据模糊分类矩阵判定是正确的样品数/总的样品数 × 100。

故 $0\leqslant f(X)\leqslant 100$。本文设定当 $f(X)\geqslant 99$，或者遗传算法的代数达到设定的限值时，即可结束训练。当 $f(X)\geqslant 99$ 时，可认为已求得最优类别隶属度矩阵

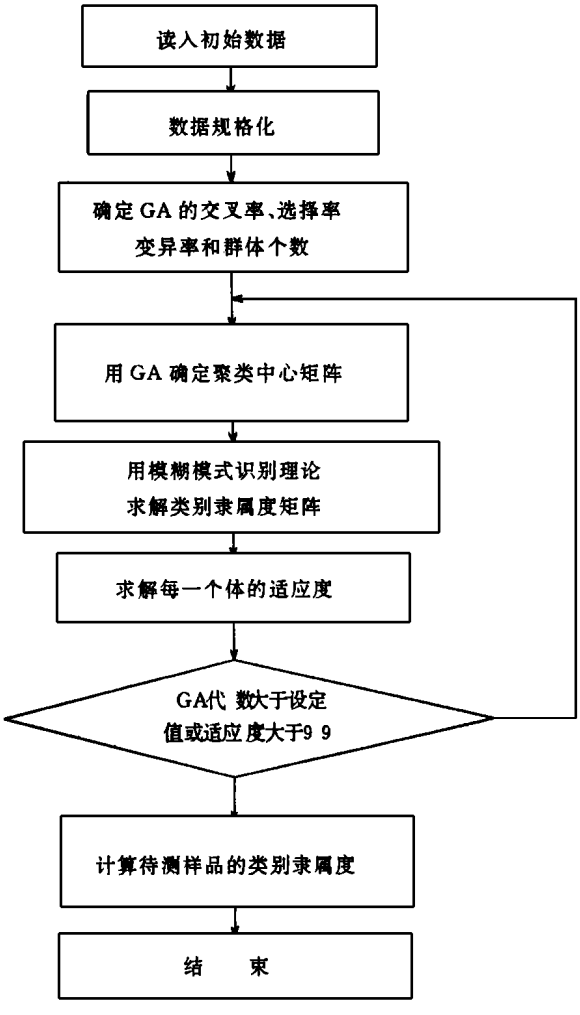

图 1　信息流程图

Fig. 1　Flow chart of the programming

和最优模糊聚类中心矩阵。这时可根据具体的样品和求得的最优模糊聚类中心，即可预测样本的类型。程序的信息流程见图 1。

由于此法把权向量也作为未知量用遗传算法进行优选，充分考虑了样本的实际情况因为，一般来说，样本的不同指标不宜采用相等的权重，各指标对样本分类的贡献是不一样的，这种方法特别适用于样品本身来源具有模糊性的样品。

本文遗传算法的选择操作[5]根据串的适应度大小来选择"父代"，其选择的概率是根据适应度较大的串被选中的概率比适应度较小的概率被选中的概率大。

交叉操作首先随机地把选择操作选择的"父代"两两配对，然后按照一定的概率部分地交换相互配对的二个串。变异操作则随机地改变串中某一位的值。具体的例子见图 2。

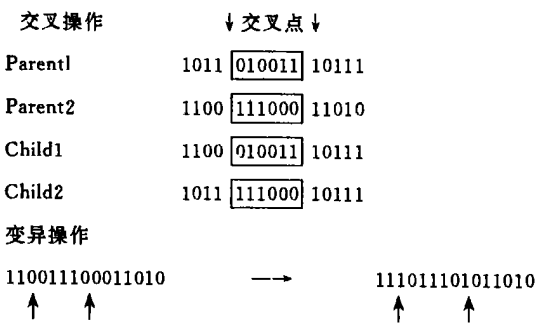

图 2 遗传算法操作算子

Fig 2 The operators of genetic algorithm

2 油品属性的比较

本文研究的样本来自塔里木盆地轮南、东河、塔中和英买力地区的 20 个原油样品(见表 1)，这些油样经处理后用三维全扫描荧光[6]采集数据，再用R

表 1 塔里木盆地原油样品及预测结果

Table 1 Prediction result of crude oils in the Tarim Basin

序 号	井 号	层 位	深 度 (m)	构造带	属于第二类原油的隶属度
1	YM2	O	5926～6050	英买力构造	0.04
2	YM2	O	5940～5953	英买力构造	0.05
3	TZ4	C	3597～3807	塔中四号构造	0.38
4	TZ4	C	3712～3720	塔中四号构造	0.21
5	TZ402	C	3259～3268	塔中四号构造	0.08
6	TZ402	C	3705～3708	塔中四号构造	0.14
7	LN10	T	4722～4754	轮南断垒带	0.61
8	LN14	C	5117～5124	桑塔木断垒带	1.00
9	LN22	C	5090～5124	桑塔木断垒带	0.92
10	LN39	O	5521～5549	桑塔木断垒带	0.92
11	LN48	O	5436～5470	桑塔木断垒带	0.90
12	LN54	O		桑塔木断垒带	0.92
13	LN46	O	6144	桑塔木南背斜带	0.93
14	DH-1	C		东河塘构造	0.12
15	DH-11	C	5698～5705	东河塘构造	0.27
16	JF122	C	5199～5216	桑塔木断垒带	0.99
17	LN55	T	4296～4429	吉拉克构造带	0.75
18	LN58	T		吉拉克构造带	0.98
19	LN46	T		桑塔木南背斜带	0.92
20	LN46	O	6114～6144	桑塔木南背斜带	0.98

型聚类方法处理后获得 4 个荧光指标[臻獐]。塔里木盆地 1977 年发现柯克亚油田以来，特别是 1984 年 9 月位于塔北隆起中部的沙参 2 井喷出高产油气流以来，地球化学工作者已投入了很大的力量开展地球化学研究工作。在判别油品属性方面，已经知道除陆相油外，塔里木盆地的原油初步分为两类。第一类原油样品包括了所有的东河塘、塔中地区、英买力地区的原油。第二类原油包括了轮南地区大部分油样，包括 LN^{14}、LN^{22}、LN^{39} 等样品。但是还存在另一类原油，可认为是一种过渡性的原油，如 LN^{10} 等。它们在某些特征上与第二类原油相似，但在很多方面也显示了第一类原油的特征。说明这类原油可能是在形成第二类原油的情况下，又受到第一类原油的影响。用本文的方法不但可以判别某一样品属于第一类原油或第二类原油，也能说明其属于第一类的隶属度或第二类原油的隶属度。因而可以看出两类原油的"混合"程度，对于预测过渡性原油的"混合"程度起到指导作用，有效地指导油气的勘探和开发工作，更有利于判别原油的形成环境以及迁移对原油性质的影响。

用 R 型聚类分析得到的 4 个指标$\frac{\text{芘}}{\text{菲}}$、R 值、菲、荧光特征值作为特征进行数据处理，由表 1 可知，1～8 一般属于第一类原油，9～20 一般属于第二类原油，为了消除各荧光指标的物理量对划分的影响，先用公式(4)进行规格化，样本处理的方法是采用留一法(leave-one-out)，每次对 19 个样品进行计算，求得最优模糊聚类中心矩阵，然后对留下的最后一个进行预测，在此算法中，遗传算法的群体取 66 个，选择率取 0.8，交叉率取 0.75，变异率取 0.1，一般经 10 代选择后，其适应度函数的值即可达到 100。以 7 号样品为例进行预测，对其余样品进行计算以求最优模糊聚类中心矩阵和最优权向量，当样本的分类归属与样品已知类型相符时，即可认为求得最优模糊聚类中心矩阵$[S^*]$和最优权向量 W^*，此时可得最优模糊聚类中心矩阵为：

$$[S^*]=\begin{Bmatrix}0.26 & 0.80\\ 0.12 & 0.61\\ 0.34 & 0.77\\ 0.18 & 0.62\\ 0.19 & 0.81\end{Bmatrix}$$

最优权向量 W^* 为：

$$W^*=[0.08\quad 0.30\quad 0.37\quad 0.25]$$

最优类别隶属度矩阵$[U^*]$为：

$$[U^*]=\begin{Bmatrix}0.10 & 0.90\\ 0.05 & 0.95\\ 0.41 & 0.59\\ 0.35 & 0.65\\ 0.14 & 0.86\\ 0.13 & 0.87\\ 0.06 & 0.94\\ 0.13 & 0.87\\ 1.00 & 0.0\\ 0.90 & 0.10\\ 0.93 & 0.03\\ 0.88 & 0.12\\ 0.90 & 0.10\\ 0.96 & 0.04\\ 0.99 & 0.01\\ 0.70 & 0.43\\ 0.96 & 0.04\\ 0.84 & 0.16\\ 0.95 & 0.05\end{Bmatrix}$$

可以看出分类情况与已知的情况完全一致，将 7 号样品的数据及有关的数据代入(12)，即可得到 7 号样品的最优类别隶属度 U_9^* 为：

$$U_9^*=[0.61\qquad 0.39]$$

可知其属于第一类原油的隶属度为 0.61，属于第二类的隶属度为 0.39。表明此号样品经预测较多地属于第一类原油。

由表 1 可知，此方法预测分类的结果完全准确，效果令人满意，而且这种方法可使我们了解第一类原油和第二类原油的"混合"程度，从最优模糊聚类矩阵中可看出，对于第 7 号样号 LN^{10}，虽然被划分为第一类原油，但划分为第一类原油的隶属度只有 0.61，而划分为第二类原油的隶属度为 0.39，这说明 LN^{10} 在形成第一类原油后，又混入了较多的第二类原油，致使最后形成的原油属性具有模糊性。

臻　候镜德等．采用轻烃(C_5-C_{20})进行油源对比研究．浙江大学分析测试中心，1993

獐　陈锋．应用三维全扫描荧光光谱在塔里木盆地上进行油源对比．浙江大学硕士论文，1995

3 结论

油气勘探是一项难度很大的工作，单靠一种方法难以下定论，应综合各种方法并辅以地质上的解释。本文把遗传算法和模糊识别理论相结合应用于油油对比中，充分考虑了原油样品属性的不确定性。其基本要点是按已知类型的样品，用遗传算法调整权向量和模糊聚类中心，用模糊识别模型求解类别隶属度矩阵，使之达到样品已知类型与最优模糊分类矩阵相对较优符合，从而确定相对较优指标权向量和最优模糊聚类中心矩阵。然后用另外的样本进行预测识别，确定样品所属类型。这种算法具有自进化能力，可以随着实测资料的增多不断进行参数的调整，如果用其它方法求得聚类中心，也可用本算法探索更优的聚类中心，使其能更确切地反映实际情况，并避免线性加权平均模型对评判结果的平均化，使得到的优选结果更易于决策和客观。经实验验证，取得的结果令人鼓舞，此方法提供了解决问题的新途径。

本文提出的算法有一定的通用性，也可应用于其它的分类问题。

参 考 文 献

1 陈守煜，赵瑛琪．模糊模式识别理论模型与水质评价．水利学报，1991，(6)

2 陈守煜，陈晓冰．化工方案选择的模糊优选方案．化工学报，1990，(2)

3 陈守煜．模糊划分在理论分析模型及其在水文中的应用．水利学报，1991，(12)

4 Goldberg D．Genetic Algorithm in search，optimazation，and machine learning．Addison Wesley：Reading．MA，1989

5 Antonia J Jones．Genetic algorithms and their application to the design of neural networks．*Neural Computing & Application*，1993，(1)：32～45

6 Brooks J M and Kennicutt．Three dimensional total scanning fluorescence．Technical Report，Department of Oceanography，Texas A & M University，1984

APPLICATION OF FUZZY CLASSIFICATION METHOD BASED ON GENETIC ALGORITHM TO THE IDENTIFICATION OF CRUDE OIL NATURE

CHEN Feng HU Shangxu YU Menhuai

(*Institute of Intelligent Information Engineering*，*Zhejiang University*，*Hangzhou* 310027，*China*)

Abstract

In this paper，Fuzzy Pattern Classification Method based on Genetic Algorithm(GA) is proposed．It's used to the identification of crude oil nature in the Tarim Basin．The clustering center and weight vector are determined with GA and the degree of compatibility of every crude oil sample is determined too．The results of classification for 20 specimen indicated that the performance of this method is good，and it might be referred as an assissant technique to the identification of crude oil nature．

Key words：genetic algorithm；fuzzy pattern classification；identification of crude oil nature