基于人工免疫系统的沉积微相自动识别

李国和,赵决正,江 希
LI Guo-he,ZHAO Jue-zheng,JIANG Xi

中国石油大学(北京) 计算机科学与技术系,北京 102249
Department of Computer Science and Technology,China University of Petroleum,Beijing 102249,China

LI Guo-he,ZHAO Jue-zheng,JIANG Xi.Recognition of sedimentary microfacies based on Artificial Immune System. Computer Engineering and Applications,2008,44(11):220-222.

Abstract: In order to recognize sedimentary microfacies automatically by well-logging curves,by means of coding the tendency of well-logging curves and implementing the operators such as clone immunity and aberrance,the clustering well-logging curves is presented by variant feature vectors,and then recognition model of sedimentary macrofacies with the well-logging curves is constructed by the basis on Artificial Immune System(AIS).The recognition model is applied to recognizing 150 sedimentary microfacies of ShengLi oil field,and the accuracy of recognition is up to 95%,which proves the recognition model based on AIS is very efficient in recognition of sedimentary microfacies.

Key words: artificial immune algorithm;pattern recognition;time-series data;well-logging curve;sedimentary microfacies

摘 要: 为了采用测井曲线实现沉积微相的自动识别,通过测井曲线变化趋势的编码和人工免疫系统的克隆免疫、变异等算子,建立基于人工免疫系统的测井曲线识别模型,实现了不等长特征曲线匹配过程的快速收敛。对胜利油田150个沉积微相进行识别,正确率达到95%,证实了该模型应用的有效性。

关键词: 人工免疫算法;模式识别;时序数据;测井曲线;沉积微相
文章编号: 1002-8331(2008)11-0220-03 **文献标识码:** A **中图分类号:** TP391.4

近年来,生物免疫系统又成为一个新兴的生物信息研究课题,模拟生物免疫系统的信息处理能力解决工程和科学问题[1,2]。目前,人工神经网在曲线识别中也得到许多应用[3],但是,人工神经网在训练过程中容易出现局部最优化问题,而且是针对等长特征的曲线。而人工免疫方法通过免疫算子进行全局最优搜索,从而提高了识别精度。本文把人工免疫引入到石油测井曲线处理领域[5],通过定义一个特征长度变异函数,完成对人工免疫中重要参数的优选,建立基于人工免疫系统的沉积微相识别模型,提高沉积微相的识别精度。测井曲线是一种离散采集的时序数据,因此,本文对具有时序特性的数据[4]识别具有参考价值。

1 基本原理及方法

生物免疫系统是一个高度复杂的自适应系统,对侵入机体的非己成分以及发生了突变的自身细胞具有精确识别、适应度应答和有效排除的能力。生物免疫系统对抗原的识别是一个选择抗体的进化过程。从生物信息处理的角度看,生物免疫系统具有分布自律性、多样性、动态稳定性和自适应鲁棒性[1,2]。

人工免疫系统是模仿自然免疫系统功能的一种智能方法,通过学习外界物质的自然防御机理的学习技术, 提供噪声忍耐、无教师学习、自组织、记忆等进化学习机理,具有提供新颖的解决问题的潜力。通过克隆免疫,遗传和免疫细胞在增殖中的基因突变,形成了免疫细胞的多样性。抗原与抗体的反应导致细胞克隆性增值,该群体具有相同的抗体特异性,其中某些细胞克隆分化为抗体生成细胞, 另一些形成免疫记忆细胞,以参加以后的二次免疫反应,克隆选择是生物体免疫系统自适应抗原刺激的动态过程。

本文提出的沉积微相识别将利用人工免疫系统的以下几个重要特性来体现免疫机制:

(1)基于适应度的克隆选择、基于亲和力的保留父代的高频超变异模式以及加入随机产生的新细胞, 以保证抗体多样性,体现自组织网络的学习机制;

(2)基于刺激度的细胞死亡、新细胞的不断加入,诱导群体优化,体现网络的动态稳定性;

(3)产生记忆细胞池,以最少记忆细胞识别最多抗原作为追求目标,提高网络对异体的响应速度,体现免疫系统的网络表达和保存所学知识的机制;

(4)抗体之间的互相抑制,限制同类抗体的浓度和不同类抗体间相互否定选择,体现动态网络的自适应调节机制。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60473125);中国石油(CNPC)石油科技中青年创新基金资助项目(No.05E7013)。

作者简介: 李国和(1965-),男,博士,教授,主要研究领域:人工智能、知识发现等智能信息处理;赵决正(1973-),男,硕士研究生,主要研究领域:智能信息处理等;江希(1985-),男,硕士研究生,主要研究领域:智能信息处理等。

收稿日期: 2007-07-31 **修回日期:** 2007-10-18

而基于人工免疫的沉积微相识别模型建立过程为:

(1)对测井曲线的数据预处理,去掉噪声和干扰因素;

(2)测井曲线的编码表示;

(3)定义适合二级编码字符串的亲和力计算函数;

(4)训练和模型建立;

(5)通过过滤器建立多层次的识别模型。

2　测井曲线的预处理

测井曲线预处理包括平滑处理和归一化处理。设测井数据 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$,平滑处理:

$$x_i=\frac{\sum_{j=1}^{m}X_{\left(\frac{i+j-(m-1)}{2}\right)}}{m} \tag{1}$$

式中,x_i 为第 i 个采集点值;$m=2k+1$,k 为正整数。当 $i>k$ 时执行式(1),当 $i\leqslant k$ 时,从 $j=k$ 开始计算。

归一化处理:

$$y_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

式中,y_i 为 x_i 规格化后的变量;$x_{\min}$ 和 $x_{\max}$ 分别为 x_i 的最小和最大值。这样就把各个因素的数值规格化到了[0,1]之间。

3　测井曲线的编码与亲和力计算

3.1　测井曲线的编码

将曲线趋势分为五个区域(如图 1),并形成编码表(A,B,C,D,E),分别表示曲线段在([3π/10,π/2],[π/10,3π/10],[-π/10,π/10],[-3π/10,-3π/10],(-π/2,-3π/10))的大体位置,称为主编码 c;用(0,1)表示曲线段在某区域内的偏离程度即(下半区,上半区),称为辅编码 ρ。

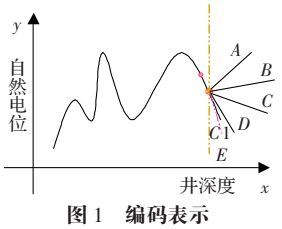

图 1　编码表示

形态编码(c_i,ρ_i)由该第 i 点与 $i-1$ 点的斜率确定,计算如式(3):

$$\begin{aligned}&\beta_i=A\tan((x_i-x_{i-1})/k)\\&c_i=seek1(\beta)\\&\rho_i=seek2(\beta)\end{aligned} \tag{3}$$

式中,β_i 表示弧度,x_i 为第 i 个采集点值,k 为采样间距,$seek1(\,)$、$seek2(\,)$分别确定该点的主编码和辅编码。

这样,一条测井曲线就可表示为:

$$Curve=\{<c_1,\rho_1>,<c_2,\rho_2>,\cdots,<c_n,\rho_n>\}=<Sn,Pn>$$

式中,$Sn=<c_1,c_2,\cdots,c_n>$,$Pn=<\rho_1,\rho_2,\cdots,\rho_n>$。

3.2　亲和力计算

亲和力计算分为主编码相似度计算和辅编码相似度计算,并且各自有着不同的权重,其中主编码相似度计算由两部分构成,设抗原 $Ag=<Sm,Pm>$,抗体 $Ab=<Sn,Pn>$。

$$\begin{aligned}aff(<Sm,Pm>,<Sn,Pn>)=u*[v*Sim_{s1}(Sm,Sn)+\\(1-v)*\ Sim_{s2}(Sm',Sn')]+(1-u)*Sim_p(Sm,Pm,Sn,Pn)\end{aligned} \tag{4}$$

式中,m、n 分别为抗原和抗体编码长度,$Sim_{s1}(\,)$为主编码相似度,$Sim_{s2}(\,)$为压缩主编码相似度,$sim_p(\,)$为辅编码相似度;$0<u<1$,$0<v<1$。

(1)$Sim_{s1}(Sm,Sn)$计算

设 $diff(Sm,Sn)$为 Sm,Sn 间的最小编辑距离;

$sim_sequence(Sm,Sn)$为相似串集合。

$$diff(Sm,Sn)=\begin{cases}m,n=0\\n,m=0\\\min\{diff(Sm-1,Sn),diff(Sm,Sn-1),diff(Sm-1,Sn-1)\},c_m=c_n\\\min\{diff(Sm-1,Sn),diff(Sm,Sn-1),diff(Sm-1,Sn-1)\}+1,c_m\neq c_n\end{cases}$$

$$Sim_{s1}(S_m,S_n)=1-\frac{diss(Sm,Sn)}{\max(m,n)} \tag{5}$$

(2)$Sim_{s2}(S_m',S_n')$计算

设 $Sm'=\{<C_1,k_1>,<C_2,k_2>,\cdots,<C_{ms},k_{ms}>\}$

$Sn'=\{<C_1,k_1>,<C_2,k_2>,\cdots,<C_{ns},k_{ns}>\}$,$C_i$ 表示压缩后的主编码,k_i 表示该编码的影响度,$q=\min(ms,ns)$。

$$Sim_{s2}(S_m',S_n')=1-\frac{\sum_{i=1}^{q}ki*diss(si,si)}{\max(\sum_{i=1}^{ms}mi,\sum_{j=1}^{ns}nj)} \tag{6}$$

(3)$Sim_p(Sm,Sn)$计算

利用主编码相似度计算过程中生成的相似串集合 $sim\_sequence(Sm,Sn)$,生成匹配辅编码串 $sim\_sequence(Pm,Pn)$:

$$sim\_sequence\ (P_{mL},P_{nL})=match\ (<Sm,Pm>,<Sn,Pn>,sim\_sequence(Sm,Sn)$$

$$Sim_p(Sm,Sn)=\frac{\sum_{i=1}^{L}(|Pmi-Pni|)}{L} \tag{7}$$

式中,$L=length(sim\_sequence(Sm,Sn))$。

3.3　相关参数的设定

人工免疫算法以随机的方式进行全局搜索,并保证以概率 1 收敛到最优解,其目标函数为:

$$C(w,p)=\frac{k}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}w_{ij}(aff(x_j,\boldsymbol{p}_i))} \tag{8}$$

式中,$w_{ij}\in\{0,1\}$ 表示抗原 x_j 属于第 i 类的隶属度,$\boldsymbol{p}_i=[p_{i1},p_{i2},\cdots p_{im}]^{\mathrm{T}}$ 表示第 i 类的聚类中心,k 为聚类中心个数,$aff(.)$表示亲和力。

人工免疫系统在训练和识别过程中进行参数设置和自适应调整,其中克隆选择阈值 λ_1 和否定选择阈值 λ_2 随着抗体亲和度的不断提高而增大。两参数定义如下:

$$\lambda_1=\frac{1}{1+c(w,p)} \tag{9}$$

$$\lambda_2=\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}aff(Abi,abj)}{\frac{N(N-1)}{2}} \tag{10}$$

其中,N 为抗体总数。

根据亲和力的大小确定变异率 p_c,既保护优良抗体,直接进入下一代,又保证新抗体的产生,实现了抗体的多样性。采用有监督的变异,对抗体(测井曲线)编码中不对相似串集合中的编码进行删除、增加和修改等变异操作。

4　沉积微相识别模型建立和应用

4.1　识别模型的建立

根据上述原理,以测井曲线为对象建立沉积微相识别模

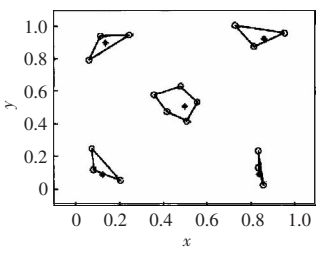

图 2　记忆细胞池的二维形态空间分布

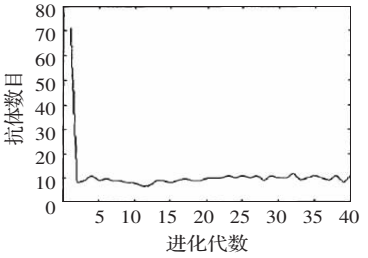

图 3　抗体数目收敛过程

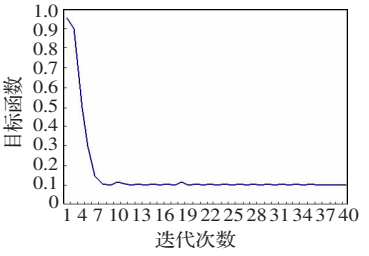

图 4　目标函数收敛效果

型。测井记录模式为 Lu（DEPTH,GR,RXO,RT,RI,SP,DEN,CNL,AC），其中 DEPTH 为该井测试点的深度,GR,RXO,RT,RI,SP,DEN,CNL,AC 依次为分别为自然伽码测井、冲洗带电阻率测井、原状电阻率测井、侵入电阻率测井、自然电位测井、密度测井、中子测井、声波测井。这些测井的探测值均为实数，每米测试八个点。本训练以胜利油田 213 井数据为主。在沉积微相识别中,SP 数据具有很好的代表性,本训练和测试仅以 SP 测井数据进行。具体流程如下：

步骤 1　L=1,针对输入抗原,产生抗体 $Ab(L)$；

步骤 2　每一个抗原 xi,进行如下运算：

(1)根据式(4)分别计算与所有抗体的亲和力并排序,按式(9)计算阈值 λ_1 进行删除；

(2)对满足条件的抗体进行克隆和变异,高亲和力的抗体(10%)直接进入下一代；

(3)重新计算抗体与抗原集的亲和力并排序,将亲力最高的 ζ%个抗体作为记忆细胞,存储于 Mp 中；

(4)在 Mp 中,按式(10)计算 λ_2,将抗体-抗体亲和力大于 λ_2 的记忆抗体删除；

(5)将 $Ab(L)$与 Wp 结合形成新的 $Ab(L)$；

步骤 3　在 $Ab(L)$中,计算每一个抗体与其它抗体的亲和力。若某抗体与 k 个以上抗体的亲和力大于 λ_2,则表明该抗体处于模糊边界上,删除该抗体(k 为簇内抗体个数)；

步骤 4　用随机产生的抗体代替部分亲和力最差的抗体；

步骤 5　L=L+1,若目标函数 $c(w,p)$误差小于 ε,训练结束；否则,返回步骤 2

通过上述流程得到一个记忆细胞池。在记忆细胞池中,明显形成具有聚类特性的若干个记忆细胞组。每组对应一类沉积微相,而组中的每个记忆细胞为沉积微相的一个实例(该实例是进化得到的)。整个记忆细胞池构成了基于实例(记忆细胞)的沉积微相分类器。如图 2 所示,图中 x 轴表示测井曲线的相对中心值,y 轴表示测井曲线的点平均值。通过 213 井数据的训练,表明该分类器的形成收敛速度快,在 5 次迭代后基本稳定,并能以概率 1 收敛到最优(如图 3、图 4 所示)。

4.2　识别结果

已知 213 井 2 199~2 202 m 河口坝粉细砂岩段数据，共有 20 个沉积微相。通过沉积微相的识别模型对该组数据进行沉积微相识别实验。识别结果如图 5 所示,其对应的统计信息如表 1 所示。可以看出,正确识别 19 个,错误识别 1 个(将远沙坝识别为河道),识别正确率高达 95%,大于 85%生产要求。错误识别原因主要是该段地层厚度太薄,其次是这两种微相曲线有一定的相似性。

表 1　实验结果统计表

沉积微相	正确数	测量数	错误数
河道	7	8	1
河道间	6	6	0
河口坝	2	2	0
远砂坝	3	2	-1
前扇三角洲	2	2	0

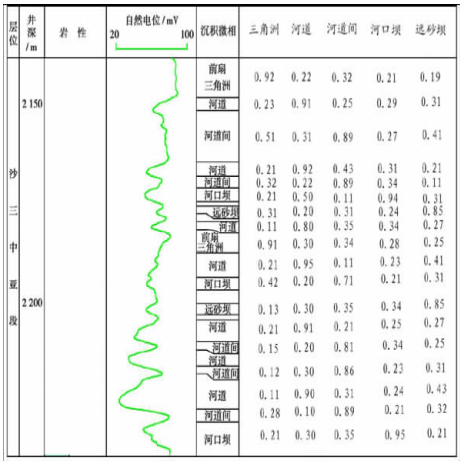

图 5　对沉积微相的识别结果

5　结束语

基于人工免疫系统的沉积微相识别,有以下的特点：

(1)用符号自然地表达测井曲线变化趋势和形态,将连续空间的时间序列转换为离散空间的二级字符串，既简洁直观、易于理解,又降低了问题的复杂度。

(2)在抗体变异过程中,实现有启发的变异和自适应调整,既保存抗体的优良特性,又提高整体抗体的亲和力,保证以概率 1 的最优收敛。

(3)通过克隆去掉了模糊边界上的抗体,以记忆抗体细胞为沉积微相实例,实现基于实例的沉积微相识别。

(4)每个记忆细胞都是进化来的,并成为一类沉积微相实例,使得记忆细胞的识别具有一定的鲁棒性,提高了沉积微相的识别正确率。

测井曲线是以可视化形式表示的离散采集的时序数据。本文的实现方法可以推广到其他领域的时序数据识别。

参考文献：

[1] 李涛.计算机免疫学[M].北京:电子工业出版社,2004.

[2] 肖人彬,王磊.人工免疫系统:原理、模型、分析与展望[J].计算机学报,2002,25(12):1281-1293.

[3] 张福明,李洪奇,邵才瑞,等,应用神经网络模式识别技术进行测井沉积学研究[J].石油勘探与开发,2003,30(3):121-123.

[4] 何书元.应用时间序列分析[M].北京:北京大学出版社,2004.

[5] 雍世和,洪有密.测井资料综合解释与数字处理[M].北京:石油工业出版社,1984.