基于DNA 编码的人工免疫模型在土壤质量评价中的应用

杨海东[1]，胡月明[2※]，邓飞其[1]
(1. 华南理工大学自动化科学与工程学院，广州 510640； 2. 华南农业大学地理信息工程研究所，广州 510642)

摘 要：针对目前土壤质量评价方法中存在的不足之处，该文通过分析人工免疫模型中二进制编码所存在的问题，提出采用DNA 编码对其进行改进，构造一种基于DNA 编码的人工免疫模型进行土壤质量评价。利用该模型对东莞赤红壤现代农业试验区进行土壤质量评价，将试验区土壤质量分为4等，根据实地抽样对照评价的结果，结果表明采用基于DNA 编码的人工免疫模型进行土壤质量评价时与实际相符，并具有稳定、结果可靠等特点，能较好地解决在进行土壤质量评价时，对于具有空间特性、模糊性、不确定性以及多指标的对象难以评价等问题。

关键词：DNA 编码；人工免疫模型；GIS；评价；土壤质量
中图分类号：TP183 **文献标识码**：B **文章编号**：1002-6819(2005)06-0040-05

杨海东，胡月明，邓飞其 基于DNA 编码的人工免疫模型在土壤质量评价中的应用[J]. 农业工程学报，2005，21(6)：40- 44
Yang Haidong，Hu Yueming，Deng Feiqi. Artificial immune model based on DNA-encoding and its application in evaluating the soil quality[J]. Transactions of the CSAE，2005，21(6)：40- 44 (in Chinese with English abstract)

0 引 言

目前，土壤质量评价方法处于起步阶段，而土地评价方法则比较丰富和系统，主要包括参数法、模型法、景观生态法、土地系统分析法和地理信息系统法，其中很多评价方法和原理可以借鉴到土壤质量的评价之中[1,2]，例如主成分分析、逐步回归分析、层次分析法、多元回归分析、相关系数检验、灰色关联度分析、模糊综合判别等都是在土壤质量评价过程中可利用的一些经验模型。

但将这些模型和方法应用于土壤质量评价时表现出一定的局限性。土壤质量评价所需的数据中包含大量的拓扑和/或距离信息，通常按复杂的、多维的空间索引结构组织数据，其访问是通过空间数据的访问方法，经常需要空间推理、地理计算和空间知识表示技术。而这些模型和方法在空间推理、地理计算和空间知识表示技术等方面存在一定的缺陷，并且缺乏知识自学习的能力。利用人工免疫模型的知识表达、逻辑推理、多样性、自学习和自适应能力以及地理信息系统(GIS)强大的空间数据管理、表示和地理计算等能力可以解决这些复杂问题。针对人工免疫模型中二进制编码所存在的不足，提出一个基于DNA 编码的人工免疫模型，并将其与GIS 结合起来，对广东省东莞赤土壤现代农业试验区(以下简称试验区)进行土壤质量评价。根据实地抽样对照检验的结果，基于DNA 编码的人工免疫模型所得出的评价结果均与试验区土壤资源的实际情况大致相符。同其它模型的结果相比，用基于DNA 编码的人工免疫模型进行土壤质量评价具有稳定、结果可靠等特点，能非常好地解决在进行土壤质量评价时，对于具有空间特性、不确定性以及多指标的对象难以评价等问题。

收稿日期：2004-08-19 修订日期：2004-12-28
基金项目：国家自然科学基金资助项目(69934030)；广东省自然科学基金资助项目(980150，011629)
作者简介：杨海东(1973-)，男，博士，主要研究方向：进化算法，空间数据挖掘。广州 华南理工大学自动化科学与工程学院，510640。Email：yanghd@gsta.com
通讯作者：胡月明(1964-)，男，教授，主要研究方向：地理信息系统应用。广州 华南农业大学地理信息工程研究所，510642

1 人工免疫模型的基本原理及其编码

免疫系统是一个复杂的、高度进化的功能系统，具有学习、记忆和自适应调节能力，维持机体内环境的稳定。免疫系统在显示识别、学习、记忆、适应性、多样性、分布性机制等应用于不同计算任务的方法方面可以给人们提供丰富的灵感和启示。人工免疫模型提取和反映生物机体免疫系统这些特点，是一种具有导向性的随机全局搜索方法[3]。

在大多数人工免疫模型的实现过程中，抗原与抗体的识别、抗体的进化过程是建立在编码机制的基础上，编码对于人工免疫模型的性能如搜索能力和抗体多样性等影响较大。目前在人工免疫模型中的抗原、抗体基因编码方法中最常用的是二进制字符串(binary strings)的编码形式，但是二进制编码本身存在Hamming悬崖[4]，因而采用二进制编码的人工免疫模型进行优化存在严重的缺陷。另外二进制编码还存在连续函数离散化时映射误差，当个体编码串较短时，达不到搜索精度要求；而当个体编码串较长时，对提高精度有所帮助，但使人工免疫模型的搜索空间急剧扩大，造成人工免疫模型性能降低[4,5]。

基于上述二进制编码存在的不足，本文考虑从DNA 计算及其多样性的角度[6]，设计了一种基于DNA 编码机理的人工免疫模型。考虑单串DNA，且将其表达为4字母的集合{G，C，T，A}，并用DNA 串来编码用于表达待解决问题的候选解的抗体，从而在此基础上发展基于DNA 编码的人工免疫模型的各种操作算子，如变异、抗体的产生和抑制等，并开发基于DNA 编码的人

工免疫模型来解决土壤质量评价问题。

2　基于DNA 编码的人工免疫模型的设计

基于DNA 编码的人工免疫模型的结构与常规人工免疫模型的结构类似。待解问题通过 4 字符集 $\sum = \{A, T, C, G\}$ 编码以形成染色体, 即DNA 串。模型的任务是从DNA 串群体出发, 模拟进化过程, 最后搜索出优秀的群体和个体, 满足求解问题的优化要求。下面具体说明基于DNA 编码的人工免疫模型求解问题时的步骤[6-9]:

1) 分析问题

对问题及其解的特性进行分析和了解, 设计解的合适表达形式并采用DNA 对其进行编码。采用N 个具有任意DNA 串的个体组成初始群体C_0。一个DNA 串由 4 种碱基A, T, C, G 的结合体构成。

2) 初始抗体群体的产生

如果记忆库有初始抗体, 则从中提取M 个DNA 串(其中M 为记忆库中DNA 串的数量), 并随机产生剩余$N - M$ 个DNA 串构成初始的子代DNA 串群体C_0; 否则随机产生N 个DNA 串构成初始的子代DNA 串群体C_0。

3) 编码及适应度计算

在DNA 串中, 每 3 个连续碱基对应一个密码子, 共有 64 个密码子, 对应 20 个氨基酸。译码就是结合具体问题, 建立氨基酸与待求解问题的映射关系, 一般来讲, 可以将每个氨基酸对应一个值, 例如, 用- 10～+ 10 的自然数表示, 则值域为〔- 10, + 10〕, 然后, 结合具体问题将其映射到合理的值域上即可, 就可以求得对应DNA 串的适应度。

4) 选择

对DNA 串的选择是以DNA 串的期望繁殖率e_i 为基准。设当前的子代DNA 串群体为C_{k-1}, 分别计算C_{k-1} 中N 个DNA 串的适应度ax_v 和期望繁殖率e_i。具体计算过程如下:

① 计算DNA 串V 的浓度c_v

$$c_v = [\sum_{w=1}^{N} ac_{vw}]/N \tag{1}$$

其中

$$ac_{vw} = \begin{cases} 1 & a_{y,w} \geq Tac \\ 0 & a_{y,w} < Tac \end{cases} \tag{2}$$

式中　$a_{y,w}$ —— 抗体y 和w 之间的结合度; Tac —— 一预先确定的值。

② 计算DNA 串V 的期望繁殖率

$$e_i = ax_v/c_v \tag{3}$$

将C_{k-1} 中的DNA 串按e_i 的降序排列, 同时取前M 个DNA 串存入记忆库中。判断是否满足结束条件, 如果是, 则结束; 否则继续下一步操作。

5) 变异

变异率p_m 按期望繁殖率值自动调整, 具体公式如下

$$p_m = \begin{cases} \dfrac{k_1(e_{max} - e)}{e_{max} - e_{avg}} & e > e_{avg} \\ k_2 & e < e_{avg} \end{cases} \tag{4}$$

式中　e_{max}—DNA 串群中最大期望繁殖率值; e_{avg} —— 每代DNA 串群的平均期望繁殖率值; e—— 要变异DNA 串的期望繁殖率值。

设定k_1, k_2 取(0, 1) 区间的值, 则p_m 的值可以根据(4) 式自适应调整。从(4) 式可以看出,DNA 串期望繁殖率值越接近最大值, p_m 就越小。以一定的概率p_m 从DNA 群体B_k 中随机选取若干个DNA 个体, 对于选中的DNA 串个体, 随机地选取某一位进行DNA 串中碱基序列的变化, 从而得到DNA 种群C_k。

6) 将产生的新一代DNA 群体C_k 返回第 3) 步, 再进行评价、选择、变异等操作。

3　基于DNA 编码的人工免疫模型的土壤质量评价

以广东东莞赤红壤现代农业试验区为例, 采用基于DNA 编码的人工免疫模型, 在地理信息系统工具软件Arc Info 的支持下进行土壤质量评价。

3.1　试验区土壤质量概况

试验区位于东经 114°07′, 北纬 23°03′, 地处珠江三角洲经济开发区, 面积约 1.4 km^2。海拔 12.5～48.3 m, 属低丘陵地带, 区内丘陵岗地坡度平缓。由于地处南亚热带, 高温多雨, 年均气温 21.9℃, 日均温≥10℃, 积温为 7600℃, 无霜期长达 350 d, 平均日照时数为 1929 h, 年均降雨量 1790 mm, 4～9 月为雨季, 占年降雨量的 80% 以上, 10～3 月为旱季, 干湿交替较为明显, 雨热同季。区内成土母岩主要为砂页岩。自然植被已不复存在, 仅在低丘顶部保留有少量以马尾松为主的次残林和草丛灌木, 如芒萁、桃金娘、岗松等, 其余已开垦为旱地或种植果树, 现种植的果树有荔枝、龙眼、芒果和杨桃等。作物一年两熟至三熟, 主要农作物为蔬菜和水稻等。

试验区的土壤类型包括 3 个土类、5 个土属及 9 个土种。低丘陵顶部分布着自然赤土壤, 耕型赤土壤分布于山坡及部分谷地, 只有极少低平谷地分布着水稻土及厚熟土(菜园土) 等。各种土壤类型的面积分配情况见表 1。

表 1　试验区的土壤类型及面积

Table 1　Soil types and areas in the experimental field

土类	亚类	土属	土种	面积/hm^2	占土壤总面积/%
赤土壤	赤土壤	砂页岩赤土壤	薄厚页赤土壤	4.2	3.54
			中厚页赤土壤	3.5	2.95
			厚厚页赤土壤	3.3	2.78
		砂页岩赤红地	页赤红泥地	18.8	15.86
			页赤红砂泥地	13.9	11.73
			页果园土	67.5	56.96
厚熟土	普通厚熟土	菜田	菜田	4.9	4.13
		菜地	菜地	2.1	1.77
水稻土	潴育型水稻土	泥肉田	泥肉田	0.3	0.28

表 2 土壤质量评价因素指标体系

Table 2 Evaluation factor index system of the soil quality

等 级	Ⅰ	Ⅱ	Ⅲ	Ⅳ
土地利用类型	水田、菜地	园地	牧草地、林地	未利用地、建设用地
土壤质地	壤质	黏壤质、砂壤质	黏质、砂质	粗骨
有机质含量/%	> 3	2～3	1～2	< 1
全氮含量/%	> 0.15	0.10～0.15	0.075～0.10	< 0.075
速效磷含量/ppm	> 20	10～20	5～10	< 5
速效钾含量/ppm	> 150	100～150	50～100	< 50
pH 值	6.5～7.5	5.5～6.5	4.5～5.5	< 4.5
等级分值	4	3	2	1

3.2 评价因素及指标的确定

根据科学性、完备性和数据易获取等原则，选取土壤有机质含量、全氮含量、速效磷含量、速效钾含量、土壤 pH 值、土壤质地、土地利用类型等 7 个因子作为试验区土壤质量评价的评价因素。S_{index} = (土壤有机质含量、全氮含量、速效磷含量、速效钾含量、土壤 pH 值、土壤质地、土地利用类型)。

参照广东省土壤普查和土地利用详查成果资料中的土壤性状分级标准[10]，根据东莞试验区内多年试验研究的结果，结合专家建议，确定试验区土壤质量评价的评价因素指标体系，见表 2。

3.3 评价单元的划分

土壤类型采用 1993 年完成的试验区土壤类型调查成果图；土地利用类型、土壤质地、土壤 pH 值、土壤性状采用土地利用现状调查所获得的图件、表格、文字材料及室内分析数据。对所选因素的专题地图，经过图形预处理，应用地理信息系统软件 ArcInfo 进行数字化，生成单因素图层。

采用 ArcInfo 的多边形拓扑叠加功能，通过对各评价因素的单因素图层进行叠置分析，用生成图层的图斑作为试验区土壤质量评价的评价单元。共分 175 个评价单元。

评价所需的属性数据从评价单元图的属性数据表中获取。

3.4 评价结果的计算

采用基于 DNA 编码的人工免疫模型计算评价结果。其基本步骤如下:

1) 土壤质量等级划分的标准

本文在利用基于 DNA 编码的人工免疫模型进行土壤质量评价时，结合聚类方法的基本原理进行土壤质量等级划分。其划分的基本标准是: 使属于同一土壤质量等级中评价单元的特性尽可能地相似，而不同土壤质量等级间的特性差异尽可能地大。

2) 初始种群的产生及其编码

在 ArcInfo 的支持下，对试验区土壤资源进行矢量叠置、图斑融合、属性数据处理等一系列的操作过程之后，将试验区土壤资源划分为 175 种一级土壤资源类型，即 175 个土壤评价单元。在试验区土壤评价单元集合中随机选取 10 组对象，并根据试验区土壤质量的评价要求将其划分为 4 个等级，按此要求确定每组 4 个点，代表着土壤评价的 4 个中心。将每一组对象编码为一个 10×4×3 位的 DNA 种群个体 c，随机选取的 10 组对象构成了初始种群 C_0。即对搜索空间进行 10×4×3 维的编码，则 10 为初始 DNA 群体的规模。

3) 计算适应度

当前群体中的每一 DNA 个体 c，都对应着 4 个中心。按照试验区土壤质量评价的要求将集合分为 4 簇，对于第 i 簇 C_i，定义其簇内距离为

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i} \| x - z_i \| \tag{5}$$

式中 z_i—— 簇 C_i 的均值。定义簇 C_i 和簇 C_j 之间的簇间距离为

$$d_{ij} = \| z_i - z_j \| \tag{6}$$

并定义

$$R_i = \max_{i\neq j}\left\{\frac{S_i + S_j}{d_{ij}}\right\} \tag{7}$$

则试验区土壤质量评价的目标函数定义为

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_i \tag{8}$$

个体 c 的适应度定义为

$$f(c) = 1/DB \text{ [7]} \tag{9}$$

在计算完一个个体的适应度之后，用所产生的 4 个均值 $\{z_i \mid i = 1, 2, 3, 4\}$ 构成新的个体，并以此代替当前的个体 c。然后，判断停机条件(重复迭代 100 次或两次之间的群体簇内距离之差小于 0.1)。若条件满足，则停止运行并输出结果。否则，继续。

4) 选择

根据公式(1)、(2)、(3) 计算个体的期望繁殖率，然后，按期望繁殖率进行克隆选择。

5) 变异

根据公式(4) 计算个体的变异率 p_m。以一定的概率 p_m 从 DNA 群体中随机选取若干个 DNA 个体，对于选中的 DNA 串个体，随机地选取某一位进行 DNA 串中碱基序列的变化，从而得到 DNA 种群 C_k。

6) 转向步 3)。

对试验区 175 个土壤评价单元的集合进行了评价，试验重复进行了 1000 次，每次评价的初始值独立随机产生。基于 DNA 编码的人工免疫模型通常在 10 次迭代之内收敛到最佳适应度，收敛在 20 次迭代以上的概率要小于 1‰。

3.5 评价结果分析

基于 DNA 编码的人工免疫模型的试验区土壤质量评价结果及采用传统的模糊综合评判法得到试验区土壤质量评价结果，见表 3 和图 1、2。

表 3　试验区土壤质量评价结果表

Table 3　Evaluation result of soil quality in the experimental field

质量等级	基于DNA 编码的人工免疫模型			模糊综合评判法		
	土壤类型单元数	面积/hm^2	占试验区土壤面积/%	土壤类型单元数	面积/hm^2	占试验区土壤面积/%
Ⅰ	25	15.77	13.31	35	27.31	23.05
Ⅱ	29	17.74	14.97	27	17.37	14.66
Ⅲ	9	8.35	7.05	7	5.02	4.24
Ⅳ	112	76.64	64.68	106	68.8	58.08
合计	175	118.5	100	175	118.5	100

图 1　基于DNA 编码的人工免疫模型的试验区土壤质量评价结果图

Fig. 1　Evaluation results of the artificial immune model based on DNA-encoding in the experimental field

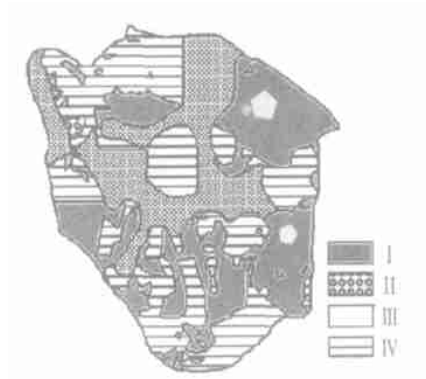

图 2　基于模糊综合评判法的试验区土壤质量评价结果图

Fig. 2　Evaluation results of the fuzzy comprehensive evaluation of soil quality in the experimental field

由表 3 和图 1、2 可以看出, 采用基于DNA 编码的人工免疫模型和模糊综合评判法进行试验区的土壤质量评价, 所得出的结果基本相近。结果表明: 一等地面积为 15.77 hm^2, 占试验区土壤总面积的 13.31%。这类土壤实际种植的作物主要是对土壤质量要求较高的龙眼、荔枝、杨桃、杨梅, 说明评价结果基本上能反映土壤质量的实际情况。二等地面积为 17.74 hm^2, 占试验区土壤总面积的 14.97%, 主要种植野生蔬菜、荔枝、杨桃。三等地面积仅为 8.35 hm^2, 占试验区土壤总面积的 7.05%, 主要为荒地, 只有很少面积种植有杨桃、芒果。四等地面积为 76.64 hm^2, 占试验区土壤总面积的 64.68%, 该等地除部分种植芒果外, 其余全分布在未规划利用区, 主要为荒地、荒草地、灌木林或废弃的荔枝/龙眼种植区, 植被稀疏, 土壤质量以退化为主, 且绝大部分土壤退化非常明显。

4　结论与讨论

本文充分利用地理信息系统(GIS)提供的强大空间数据处理和分析能力, 将之与基于DNA 编码的人工免疫模型进行有效的集成, 并首次应用于土壤质量评价, 该模型具有以下优点: (1)能够有效地量化空间数据与非空间数据; (2)能较好地解决在进行土壤质量评价时, 对于具有空间特性、模糊性、不确定性以及多指标的对象难以评价等问题。(3)本文提出基于DNA 编码的人工免疫模型属于一种新的土壤质量评价模型, 该模型的构造不但丰富了土壤质量的评价方法, 而且拓宽了人工免疫模型的应用范围。

从东莞赤红壤现代农业试验区应用实例来看, 本文所采用 2 种评价方法所得出的评价结果均与试验区的土壤资源的实际情况大致相符, 根据实地抽样对照检验的结果, 其中以基于DNA 编码的人工免疫模型得出的结果更接近实际情况。该模型能够揭示东莞赤红壤现代农业试验区的土壤质量状况, 为该区土壤质量的合理开发利用与保护、为农业的可持续发展提供可靠资料与决策依据。

但在采用基于DNA 编码的人工免疫模型进行土壤质量评价时, 由于记忆库规模、初始抗体数等参数的值都是根据实际经验选取, 模型的收敛速度和评价结果也受到人为主观的影响。因此如何进一步改进人工免疫模型, 将是下一步需要研究的问题。

[参　考　文　献]

[1]　孙　波, 赵其国. 红壤退化中的土壤质量评价指标及评价方法[J]. 地理科学进展, 1999, 18(2): 118-128.

[2]　Hu Yueming, Daijun, Wang Renchao. GIS-based red soil resources classification and evaluation [J]. Pedosphere, 1999, 9(2).

[3]　莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨: 哈尔滨工业大学出版社, 2002: 92-136.

[4]　Jerne N K. The immune system [J]. Scientific American, 1973, 229(1): 51-60.

[5]　Perelson A. Immune network theory [J]. Immunological Review, 1989, 110: 5-36.

[6]　董亚非, 王淑栋, 许　进. DNA 计算原理及系统分析[J]. 计算机工程与应用, 2003(9): 70-73.

[7] 行小帅, 潘 进, 焦李成. 基于免疫规划的 K-means 聚类算法[J]. 计算机学报, 2003, (26): 50- 54.
[8] Farmer J D, Packard N H, Perelson A S. The immune system, adaptation, and machine learning[J]. Physica D, 1986, 22: 187- 204.
[9] 任立红, 丁永生, 邵世煌. 采用 DNA 遗传算法优化设计的 TS 模糊控制系统[J]. 控制与决策, 2002, (7): 117- 121.
[10] 广东省土壤普查办公室. 广东土壤[M]. 北京: 科学出版社, 1993: 67- 85.

Artificial immune model based on DNA-encoding and its application in evaluating the soil quality

Yang Haidong[1], Hu Yueming[2※], Deng Feiqi[1]
(1. *School of Automation Science and Engineering, South China University of Technology, Guangzhou*, 510640, *China*;
2. *GIS Laboratory, South China Agricultural University, Guangzhou* 510642, *China*)

Abstract: Based on the shortage of the evaluation methods for soil quality, the artificial immune model was applied to evaluate soil quality. After analyzing the shortage of binary-encoding in the artificial immune model, the authors proposed to integrate DNA with immune algorithm and to construct an artificial immune model based on DNA-encoding. Then the new evaluation method was used to evaluate the soil quality in Dongguan Agricultural Modernization Experimental Field in Guangdong Province. The results showed the new evaluation method was stable and reliable. The artificial immune model based on DNA-encoding was able to solve the soil quality evaluation problems which were uncertain, fuzzy, and had a spatial characteristic. It was also able to solve the problem of multiple objects under multi-index condition for soil quality evaluation.

Key words: DNA-encoding; artificial immune model; GIS; evaluation; soil quality