文章编号：2095-2163(2019)03-0190-05 中图分类号：S567.9 文献标志码：A

基于 MATLAB 的长春花生物碱含量的分析与预测

陈志远，王云耿，赵万里，贺耀钦，穆丽新，刘 英
（东北林业大学，信息与计算机工程学院，哈尔滨 150040）

摘 要：长春碱是一种重要的天然抗癌药物。土壤营养成分（土壤含水量，土壤 PH，有机碳，全氮，全磷，速效磷，碱解氮），和激素（6-BA（6-苄氨基腺嘌呤），IAA（吲哚-3-乙酸），ABA（脱落酸））对长春碱含量有重要影响。本文采集了这 10 个研究条件和长春碱含量的原始数据，并且用 MATLAB 中的人工神经网络和遗传算法工具箱进行分析。结果显示土壤条件中全磷含量和土壤含水量的降低有利于提高长春碱含量，土壤 PH，有机碳，全氮，碱解氮含量的升高有利于提高长春碱含量，土壤中速效磷和长春碱含量的关系不明显。激素中，IAA 含量的降低有利于提高长春碱含量，6-BA，ABA 含量的升高有利于提高长春碱含量。

关键词：长春花；土壤条件；激素；长春碱；人工神经网络；遗传算法

Analysis and prediction of vinca alkaloid content based on MATLAB

CHEN Zhiyuan, WANG Yungeng, ZHAO Wanli, HE Yaoqin, MU Lixin, LIU Ying
(College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)

【Abstract】 Vinblastine is an important natural anticancer drug. Soil nutrients (soil water content, soil PH, organic carbon, total nitrogen, total phosphorus, available phosphorus, alkali - hydrolyzed nitrogen), and hormones (6 - BA (6 - benzylaminoaminoadenine), IAA (indole - 3 - acetic acid), ABA (abscisic acid)) have important effects on the content of vinblastine. In this paper, the ten conditions and the original data of vinblastine content are collected, and the analysis is carried out by using the artificial neural network and Genetic Algorithm toolbox in MATLAB. The results show that the decrease of total phosphorus content and soil water content is beneficial to the increase of vinblastine content. The increase of soil PH, organic carbon, total nitrogen and alkali-hydrolyzed nitrogen content is beneficial to the increase of vinblastine content. The relationship between available phosphorus and vinblastine content in soil is not obvious. In hormones, the decrease of IAA content is beneficial to the increase of vinblastine content, while the increase of 6-BA and ABA content is beneficial to the increase of vinblastine content.

【Key words】 periwinkle; soil conditions; hormone; vinblastine; artificial neural network; GA

0 引 言

长春花（Catharanthus roseus（L.）G.Don）是夹竹桃科（Apocynaceae）长春花属（Catharanthus G.Don）植物，又称雁来红、日日新、四时春、三万花等，中医临床以全株入药。原产于非洲马达加斯加岛西印度一带的热带森林地区，早在宋代以前就传入中国[1]。长春花中含有多种生物碱。其中长春碱主要用于治疗何杰金氏病和绒毛上皮癌，对何杰金氏病治疗的有效率为 68%，完全缓解率为 30%，对淋巴肉瘤、黑色素瘤、卵巢癌、白血病等也有一定疗效[2]。长春碱是一种重要的药用生物碱。

植物次生代谢的概念最早于 1891 年由 Kossel 明确提出。植物的次生代谢是指由植物体内有机化合物的初生代谢途径衍生而来，最终合成一些具有种属特异性的有机化合物的代谢过程[3]。长春花体内可以产生大量的次生代谢产物，主要是化学结构属于萜类的生物碱，这类生物碱具有非常重要的药用价值。长春碱（Vinblastine）是其中的一种。植物的次生代谢是植物在长期进化中与环境（生物的和非生物的）相互作用的结果，次生代谢产物在植物提高自身保护和生存竞争能力、协调与环境关系上充当着重要的角色，其产生和变化比初生代谢产物与环境有着更强的相关性和对应性[4]。由于环境条件在次生代谢产物合成积累中具有重要诱导作用，在植物药材种植中为了保证和提高所需成分的质量，研究产物产量和环境因素的关系即已成为亟待探索的焦点研发课题。土壤和激素是植物生长所需环境因素的重要组成部分，土壤条件和激素对植物次生代谢产物有着重大影响。对此拟展开研究阐述如下。

1 研究现状

长春花作为重要的药用植物，体内含有丰富的次生产物，因其独特的药用价值、尤其是抗肿瘤成分

基金项目：黑龙江省教育厅科学技术研究项目（12533021）。

作者简介：陈志远（1997-），男，本科生，主要研究方向：智能应用技术、软件开发。

收稿日期：2017-03-16

而使长春碱备受多方关注。经过多年研究，植物领域中已经有多种手段用在生物碱含量的提高上。例如，采用遮光培育或干旱胁迫等方法对长春花植株进行干预，目前均已取得了一定的成果。但是当下国内外对长春花中生物碱含量与土壤和激素之间关系发表的研究，则主要集中在实验手段上并且对于直接作用于相应的土壤条件和激素，改变土壤水分含量、微量元素含量等方面研究较少。因为长春花培育需要一定的周期，土壤和激素中可能影响长春花生物碱含量的研究对象比较多，单纯用实验方法取得数据进行分析将严重影响后续研究。迄今为止也还未见到有利用计算机的聚类分析方法以及仿真方法对长春花生物碱含量与土壤、激素关系进行分析的先例。

2　BP 神经网络和遗传算法

2.1　BP 神经网络简介

BP 神经网络是人工神经网络的一个基础组成部分。人工神经网络(Artificial Neural Networks)，是一种模仿动物神经网络行为、进行分布式信息存储的数字算法模型[5-6]。通过将模拟神经元逐层排列，人工神经网络可以在无需事前揭示描述映射关系的数学方程的情况下，建立输入-输出模式映射关系。人工神经网络现已成为人工智能研究的重要领域之一。BP(Back Propagation)神经网络是一种单向传播的多层前馈网络，是目前应用最广泛的神经网络模型之一[7]。BP 神经网络的核心思想在于将神经网络的预测值和原始数据的真实值加以比较，将输出误差以某种形式逐层反传，即将误差分摊给各层的所有单元，通过各层单元的误差信号来修正各单元权值。这种训练方式使得 BP 神经网络对输入变量较多的复杂问题有良好的应用效果。在收集到土壤条件和长春碱含量的基础上利用 BP 神经网络可以建立土壤中 10 个输入变量和长春碱的输入-输出模式映射。

2.2　遗传算法简介

遗传算法是一种进化算法，其基本原理是仿效生物界中的“物竞天择，适者生存”的演化法则。遗传算法是把问题参数编码为染色体，再利用迭代的方式进行选择、交叉以及变异等运算来交换种群中染色体的信息，最终生成符合优化目标的染色体[8]。当问题的输入输出函数或模式映射确定时，用遗传算法可以先利用染色体(二进制串)随机生成对应问题输入变量的几组输入值。当获得输入值后，自然而然地可以得到输入值对应的输出。如果希望得到解空间内的近似最大值或最小值，就可分别根据输出选择输出值更大/更小的染色体。让其发生交叉变异，生成新的一代继续参与运算。最后经过指定代数的选择、交叉和变异，就可以得到输入输出函数或模式映射内的近似最大或最小输出以及对应的输入值。在使用 BP 神经网络处理原始数据后，研究得到了土壤的 10 个输入变量到长春碱的输入输出映射，此时利用遗传算法就可以找到长春碱含量的近似最大值和对应的输入变量的取值，从而为下一步的分析做准备。遗传算法的研发包括以下几个步骤：

(1)初始化：设置进化代数计数器 $t = 0$，设置最大进化代数 T，随机生成 M 个个体作为初始群体 $P(0)$。

(2) 个体评价：计算群体 $P(t)$ 中各个个体的适应度。

(3)选择运算：将选择算子作用于群体。选择的目的是把优化的个体直接遗传到下一代或通过配对交叉产生新的个体再遗传到下一代。选择操作是建立在群体中个体的适应度评估基础上的。

(4)交叉运算：将交叉算子作用于群体。遗传算法中起核心作用的就是交叉算子。

(5)变异运算：将变异算子作用于群体。即是对群体中的个体串的某些基因座上的基因值作变动。

群体 $P(t)$ 经过选择、交叉、变异运算后得到下一代群体 $P(t + 1)$。

(6)终止条件判断：若 $t = T$，则以进化过程中所得到的具有最大适应度个体作为最优解输出，终止计算。

3　实验方法

3.1　获得初始数据

在实验区域划分 4 块土地，对每块土地的土壤进行不同的操作，从而建立 4 个实验组。4 个实验组分别为：对照组(CK)、对土壤使用一氧化碳供体的 SNP 的 SNP 组(SNP)、遮阴组、SNP+遮阴组，测量 4 组处理下土壤的 7 种物质，即：土壤含水量(%)、土壤 PH、有机碳(g/kg)、全氮(g/kg)、全磷(g/kg)、速效磷(mg/kg)、碱解氮(mg/kg)和植物中的激素的含量，即：6-BA(6-苄氨基腺嘌呤)、IAA(吲哚-3-乙酸)、ABA(脱落酸)(以叶片为研究对象，单位为 ug/g)与对应的长春花中长春碱(以叶片为研究对象，单位为 ug/g)的含量，每个操作组收集 3 组数据建立初始数据表格，见表 1。

表 1 初始数据表

Tab. 1 Initial data table

含水量	土壤 PH	有机碳	全氮	全磷	速效磷	碱解氮	6-BA	IAA	ABA	处理	长春碱
24.297 4	5.06	78.6	0.987	0.556 7	4.515 0	56	0.032 4	0.038 8	0.056 5		0.091 85
22.829 7	5.10	91.8	0.973	0.531 1	3.731 4	70	0.043 5	0.028 6	0.048 7	CK	0.080 97
26.010 2	5.04	85.2	1.001	0.651 9	3.623 4	63	0.027 4	0.048 9	0.067 9		0.085 32
29.056 8	5.16	73.2	1.029	0.670 3	3.967 4	70	0.164 0	0.112 0	0.170 0		0.117 26
19.925 6	5.11	86.4	1.001	0.541 3	4.519 5	70	0.258 0	0.246 0	0.256 7	SNP	0.102 90
20.256 5	5.21	62.4	1.057	0.272 5	3.335 3	63	0.195 7	0.217 0	0.201 0		0.095 22
32.283 4	5.13	83.4	0.973	0.643 2	3.515 3	70	0.032 7	0.047 7	0.199 6		0.090 52
34.294 8	5.06	87.0	0.931	0.349 9	4.075 4	56	0.046 7	0.037 8	0.184 6	遮阴	0.096 07
31.173 9	5.05	65.4	1.015	0.487 5	3.955 4	77	0.029 8	0.050 2	0.180 4		0.092 82
34.151 2	5.28	97.2	1.015	0.540 0	4.003 4	70	0.212 5	0.196 5	0.232 5	SNP	0.257 52
36.063 4	5.24	95.4	0.987	0.623 4	4.199 4	63	0.207 0	0.185 0	0.215 7	+	0.200 34
33.172 8	5.19	91.8	1.043	0.739 2	4.619 5	77	0.195 7	0.205 0	0.248 2	遮阴	0.188 36

3.2 确定 BP 神经网络结构

3.2.1 网络层数的设计

人工神经网络拓扑结构的确定对训练效果有很大影响。隐含层一般为 1~2 层,考虑到本实验有 10 个输入参数,复杂性较高,故设置 2 层隐含层,即 BP 神经网络的拓扑结构包括 4 层:输入层、2 个隐含层和输出层。

3.2.2 各层神经元数目的确定

一般来说,输入层和输出层神经元的数目由具体问题决定。输入层节点数目等于输入变量的个数,本实验有 10 个条件,因此输入层节点数取 10。输出层节点的数目等于输出变量的个数,本实验研究长春碱含量,因此取 1。隐含层节点的数目对于网络的性能起着至关重要的作用,但目前尚无统一的标准来确定,常以经验公式作为参考。隐含层节点数在采用经验取值的基础上反复试凑,测试显示隐含层采用 21-21 结构的平均相对误差最小,最终确定 BP 神经网络的拓扑结构为 10-21-21-1。

3.2.3 学习函数与传递函数以及其它参数的设定

本实验使用训练速度最快的 *traincgf* 作为学习函数,考虑到长春碱输出的范围在 0 ~ 1 之间,使用 *logsig* 作为传递函数。本模型的学习率 *lr* 的初始值取为经验值 0.01。 为了能够使神经网络达到比较小的误差,研究选定的训练目标为 1e-8,训练次数为5 000次。

3.3 选择训练集与测试集训练并测试神经网络

神经网络的训练效果不仅要看其是否能够很好地符合给定的训练数据,也要视其是否能够对新数据做出合理的预测,因此有必要在原始数据中划分训练集和测试集。

取表 1 中第六行、第九行、第十二行作为测试集,其它作为训练集。考虑到 10 个输入变量由于自身性质在土壤中含量相差很大,直接训练会导致训练误差和预测误差过大。因此需要对 2 个集合中的 10 个输入变量做适当处理,例如将某个输入变量扩大,某个输入变量缩小,使得数据间的数值差距减小。在关于如何处理数据、从而获得好的训练效果上,目前还没有固定模式。通常的做法是不断进行测试,找到对每个自变量合理的处理方式。经过大量测试发现,当土壤含水量、土壤 PH、6-BA、IAA、ABA、有机碳、全氮、全磷、速效磷、碱解氮这 10 个变量分别扩大 0.1、1、100、100、100、0.1、10、10、10、1 倍后训练效果最好,误差最小。

训练集负责训练 BP 神经网络,测试集负责测试训练好的神经网络的预测能力。具体方法为:输入测试集的 10 个输入变量,将神经网络给出的预测值和真实值进行比较,用 MATLAB 默认的 *MSE* 函数计算误差。

3.4 用遗传算法寻找有利于提高长春碱含量的土壤条件

遗传算法可以生成设定范围内的值作为训练好的神经网络的输入变量,这里则是把初始数据中每个输入变量的最大、最小值当作该变量取值范围。输入变量的最大精度为小数点后四位,考虑到每个输入变量的取值范围,个体长度取 20 即可满足需求。其它参数设定较为宽松,研究将其设为经验值。遗传代数取为 20,代沟设为 0.95,重组(交叉)概率和变异概率分别设为 0.7 和 0.01。经过 20 代的遗

传，可以得出在长春碱取值较大时输入变量的取值。记录下 10 组数据。将初始数据中每个输入变量的最大、最小值的平均值设为 A。统计 10 组数据中每个输入变量小于 A 和大于等于 A 的个数，分别设为 B 和 C，当 B 与 C 之间的差值在 4 或 4 以上时可以认为该输入变量取升高或降低有利于长春碱含量提高。否则认为该输入变量与长春碱含量联系不显著。

4　实验结果与结论

选定的 BP 神经网络经过训练后取得了比较好的效果，使用 MATLAB 软件默认的误差计算方式 *MSE* 使训练集误差达到了 7.28e−11，同时神经网络对测试集的预测值和实际值之间的误差也达到了 9.72e−05。

神经网络训练过程中误差的变化情况如图 1 所示。由图 1 可以看出，*traincgf* 函数的训练速度非常快，在第 26 次训练时就达到了 1e−8 的目标误差。

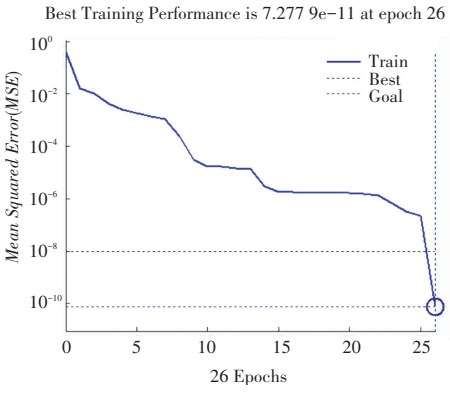

图 1　神经网络训练情况

Fig. 1　Training of neural networks

用遗传算法寻找长春碱获取近似最大值时，可得各条件的预测值，并记录 10 组数据，结果见表 2。在理想条件下，神经网络预测长春碱的含量可以进一步提高到 0.27 ug/g 以上。

表 2　长春碱取近似最大值时各条件的预测值

Tab. 2　Predictive value of various conditions for approximate maximum value of vincristine

含水量	土壤 PH	有机碳	全氮	全磷	速效磷	碱解氮	6-BA	IAA	ABA	长春碱
21.359 7	5.260 8	92.211 0	1.016 7	0.308 4	3.873 3	75.748 4	0.246 3	0.105 2	0.195 9	0.270 4
22.847 0	5.091 3	87.380 1	1.037 5	0.376 2	3.560 5	76.996 3	0.228 8	0.033 5	0.224 5	0.270 5
23.025 1	5.279 5	74.485 5	1.054 1	0.309 3	3.895 0	76.121 1	0.241 7	0.051 0	0.198 8	0.270 5
25.198 9	5.275 7	93.683 6	1.030 0	0.426 5	4.474 1	74.486 9	0.242 1	0.075 8	0.202 5	0.270 4
27.263 2	5.211 5	77.903 7	1.003 3	0.395 0	3.732 5	76.638 2	0.246 3	0.064 7	0.166 7	0.270 3
19.959 0	5.252 1	94.903 7	1.038 2	0.295 8	4.544 2	75.745 4	0.241 1	0.032 6	0.194 8	0.270 5
22.463 5	5.160 0	93.688 9	1.029 6	0.283 8	3.583 5	76.552 6	0.246 0	0.038 0	0.238 0	0.270 5
22.844 1	5.276 5	93.468 6	1.005 6	0.405 2	3.642 3	76.179 3	0.249 9	0.116 8	0.191 6	0.270 4
20.724 4	5.261 0	86.523 0	1.039 4	0.467 8	4.386 2	76.213 9	0.243 1	0.041 7	0.198 9	0.270 4
21.080 7	5.279 1	83.961 1	1.054 7	0.486 5	4.296 9	75.325 9	0.244 4	0.073 3	0.183 4	0.270 4

研究求得初始数据的长春碱平均含量为 0.124 929 ug/g，遗传算法找到的 10 组数据中长春碱的平均含量为0.270 437 ug/g，且数值波动小，相对于初始数据有较大提高。

将初始数据中每个输入变量的最大、最小值的平均值设为 A。统计 10 组数据中每个输入变量小于 A 和大于等于 A 的个数，分别设为 B 和 C，当 B 与 C 之间的差值在 4 或 4 以上时，可以认为该输入变量取升高或降低，有利于长春碱含量提高。否则认为该输入变量与长春碱含量联系不显著。经过统计，得到结果见表 3。

表 3　统计结果

Tab. 3　Statistical results

数据类别	含水量	土壤 PH	有机碳	全氮	全磷	速效磷	碱解氮	6-BA	IAA	ABA
A	27.995 4	5.16	79.8	0.994	0.505 8	3.977 4	66.5	0.142 7	0.137 3	0.152 7
B	10	1	2	0	10	6	0	0	10	0
C	0	9	8	10	0	4	10	10	0	10
结论	降低	升高	升高	升高	降低	不明显	升高	升高	降低	升高

由表 3 中可以看出在设定的取值范围内，土壤条件中全磷含量和土壤含水量的降低有利于提高长春碱含量，土壤PH、有机碳、全氮、碱解氮含量的升高有利

（下转第 198 页）