DOI: 10.13382/j.jemi.2017.03.007

基于太赫兹光谱和 APSO-SVM 的转基因棉籽识别方法*

潘学文[1] 尹向东[1] 赵永红[2,3]

(1. 湖南科技学院 电子与信息工程学院 永州 425199;2. 西安电子科技大学 机电工程学院 西安 710071;
3. 桂林电子科技大学 广西自动检测技术与仪器重点实验室 桂林 541004)

摘 要:针对目前转基因产品检测是基于可见光/近红外光谱,在支持向量机建模中参数难以确定及光谱数据计算量过大的问题,提出了一种基于太赫兹光谱和自适应动态粒子群优化的支持向量机算法,以实现对转基因棉种子的分类判别。为实现转基因棉种子的分类识别,在波长 150 μm ~ 3 mm 采集 3 种最新转基因棉种子 165 个样本的太赫兹光谱,并用基于自适应动态粒子群优化的支持向量机对 165 个转基因棉种子进行识别。实验结果表明,综合识别率达到97.3%,太赫兹光谱结合自适应动态粒子群支持向量机可为转基因棉种子类型辨别提供一种快速、精确、简便、无损的检测方法。

关键词: 太赫兹;支持向量机;转基因棉种子;分类辨别;自适应粒子群;
中图分类号: O433.4;TN06 文献标识码: A **国家标准学科分类代码**: 510.20

Distinguish of genetically modified cotton seed by using terahertz spectroscopy and APSO-SVM

Pan Xuewen[1] Yin Xiangdong[1] Zhao Yonghong[2,3]

(1. School of Electronics and Information Engineering, Hunan University of Science and Engineering, Yongzhou 425199, China;
2. School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China;
3. Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: Aiming at the inspection of genetically modified product is mainly based on visible or near infrared spectroscopy at present, the support vector machine (SVM) modeling parameter is difficult to determine and the problem of the large amount of spectrum data calculation, a support vector machine (SVM) algorithm based on terahertz spectroscopy and adaptive particle swarm optimization (APSO) is proposed to distinguish genetically modified cotton seed. To achieve distinguish genetically modified cotton seed, the present invention is train of thought collect 165 samples of three kinds of latest genetically modified cotton seed of terahertz spectroscopy in range of 150 μm ~ 3 mm wavelength and identification of 165 genetically modified cotton seeds based on APSO-SVM. The experiment results show that the comprehensive recognition rate reached 97.3%. It can provide a precise, fast, convenient, and nondestructive detection method to distinguish genetically modified cotton seed by using terahertz spectroscopy couple to APSO-SVM.

Keywords: terahertz; SVM; genetically modified cotton seed; classification to distinguish; APSO

1 引 言

近年来,随着转基因技术的普及以及转基因产品的推广,转基因产品的安全检测和评价得到了广泛的重视。目前,检测转基因食品的分子鉴定手段是主要是 PCR (polymerase chain reaction)[1-2]。PCR 方法具有检测灵敏度高的优点,但进行转基因食品检测时需要国际标准的

收稿日期:2016-09 Received Date: 2016-09
*基金项目:湖南省教育厅科学研究优秀青年项目(14B070)、广西自动检测技术与仪器重点实验室开放基金(YQ16204)、湖南科技学院科研项目(2015XKY004)、湖南科技学院重点学科建设项目(电路与系统)资助

转基因标准样品，故 PCR 方法在实际应有中较不方便。利用可见/远红外光谱在对转基因产品进行检测，会带来优化参数难以确定、光谱数据计算量大等问题。因此，开发新的检测转基因食品手段势在必行。

太赫兹是一种频率在 0.1 ~ 10 THz、波长在 30 μm ~ 3 mm 的电磁波，波段位于微波与红外光之间，属于远红外波段。理论研究表明，大量生物分子如 DNA、蛋白质等的振动和转动能级正好处于 THz 的频带范围内，用 THz 时域光谱技术（tera Hertz-time domain spectrum，THz-TDS）探测生物样品能产生共振吸收峰，从而使利用太赫兹光谱识别生物样品将成为可能[2]。目前，利用近红外光谱进行转基因产品的检测识别已较多，如近红外光谱技术在检测转基因玉米上的应用；近红外光谱技术在检测转基因油菜籽中芥酸和硫甙上的应用；可见/近红外光谱技术鉴别转基因番茄叶的应用等。然而，在国内外利用太赫兹鉴别转基因食品的应用几乎没有，故利用太赫兹光谱鉴别转基因食品具有重要的理论和现实意义[3-4]。

本研究使用太赫兹光谱检测系统对 165 个转基因棉种子样本进行光谱扫描，并采用基于动态粒子群优化的支持向量机分类对得到的数据进行建模，实现了转基因棉种子种类的识别，实验效果良好，结果令人满意。本方法具有快速、精确、简便、无损等特点，对实际生活中的转基因棉种子检测有较高的指导意义和应用价值。

2　实验部分

2.1　实验样品

选取不同种类的转基因棉种子（陆中6号、鑫秋k638、鲁棉研36号，均购于中国农业科学生物技术研究所）。分别将每种转基因棉种子制作成55个片剂，3种转基因棉种子产生165个样品，将165个转基因棉种子样品分为两组，第1组90个样品作为训练集用于自适应粒子群支持向量机建模校正；第2组75个样品作为预测集用于待判样品预测，并保证3类转基因在训练集和预测集中都存在。

2.2　实验仪器及方法

本文所用的 THz-TDS 系统为透射式太赫兹时域光谱系统，如图1所示。激光器的中心波长为800 nm，为保证实验的准确性，系统内注入氮气直至内部相对湿度达到0.2%以下。室内相对湿度为25%，恒温292 K。

按照2.1节中的方法，将不同种类转基因棉种子分别去壳，烘干后碾压成粉末，然后用压片机对转基因棉种子粉末进行压片，制作165个带测样品。

3　数据处理方法

3.1　支持向量机

支持向量机（support vector machine，SVM）的实现是

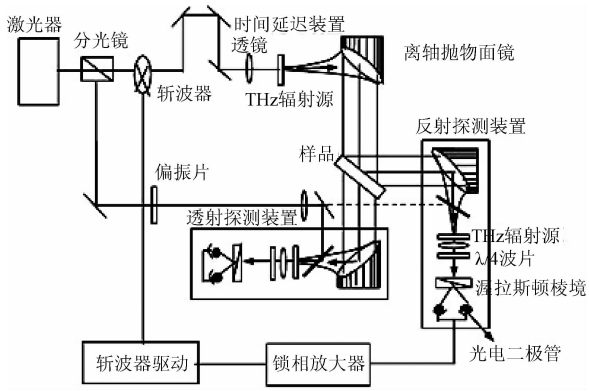


图1　THz-TDS 系统

Fig. 1　Schematic diagram of THz-TDS

通过在初始阶段选择一个非线性变换方法，将输入向量由低维非线性样本空间映射到高维或无穷维，使样本空间的非线性分类转化为线分类，并基于结构风险最小化在特征空间中寻找最优超平面，解决线性分类问题[5]。

SVM 被其描述为[6-10]：

$$\begin{cases}\min\limits_{\omega,b,\xi} J_p(\omega,\xi) = \dfrac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \gamma\sum\limits_{k=1}^{N}\xi_k \\ \text{s.t. } y_k[\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\varphi}(x_k) + b] \geqslant 1 - \xi_k, k = 1,\cdots,N \\ \xi_k \geqslant 0, k = 1,\cdots,N\end{cases} \tag{1}$$

式中：ω 为惯性权重，ξ 为松弛系数，γ 为惩罚系数。

3.2　粒子群优化算法

粒子群优化算法是由 Eberhart 博士和 Kennedy 博士提出的一种进化计算技术。粒子群优化算法源于对鸟群捕食行为的研究，其基本思想是通过群体中个体之间的信息传递及信息共享来寻找最优解[11-12]。其经典算法如下所示[13-17]：

$$v_{ij}(t+1) = \omega v_{ij}(t) + c_1 r_{1j}(t)(p_{ij}(t) - x_{ij}(t)) + c_2 r_{2j}(t)(p_{gj}(t) - x_{ij}(t)) \tag{2}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \tag{3}$$

式中：$i = 1,2,\cdots,M$，M 是该群体中粒子的总数，i 表示群体中第几个粒子；j 表示微粒的第 j 维，即算法优化的第 j 个参数；ω 为惯性权重因子，其值为负，值的大小会对整体寻优能力产生影响；t 表示当前时刻优化的代数；$v_{ij}(t)$ 表示 t 时刻粒子 i 在 j 维的空间速度；c_1 和 c_2 为加速因子，通常在0~2取值；r_{1j} 和 r_{2j} 为两个[0,1]之间变化的相对独立的随机函数；$p_{ij}(t)$ 为粒子 i 的历史最好解（个体最优）的 j 维值，即单个粒子 i 在所优化的第 j 个参数中的历史最好解；$p_{gj}(t) = \min\{p_{ij}(t)\}$ 为所有粒子在 t 时刻的历史最好解（群体最优）的 j 维值，即所有粒子在所优化的第 j 个参数中的历史最优解；$x_{ij}(t)$ 为 t 时刻粒子 i 在 j 维空间位置。

3.3　自适应粒子群优化

粒子群优化算法是通过粒子跟踪自身记忆的个体最优向种群记忆的全局最优靠近以逐渐逼近更优位置。动态环境下粒子自身记忆的个体最优位置与全局最优位置对应的适应度值是变化的，粒子将重新对先前环境进行寻优，有可能导致寻优算法失效。故在动态环境下普通粒子群算法很难逼近最优位置。

为在动态环境下粒子群算法获的最优解，本文对算法进行了如下改进：1）加入感知能力，使粒子群获得感知外部环境变化的能力；2）引入更新机制，在感知到环境变化后，采用某种响应方式对粒子群进行更新，以适应实际动态环境。其思想是利用粒子感知外部环境是否发生变化，将可行域空间分为 n_1 个均匀的子空间，在每个子空间内随机抽取 n_2 个粒子作为感知粒子进行初始化，在迭代过程中计算感知粒子的适应度 f_i，并比较相邻 2 次迭代的适应度差值 Δf，然后对所适应度差值求和 F，公式如下：

$$\Delta f_i = f_i(k+1) - f_i(k) \tag{4}$$

$$F = \sum_{i=1}^{n} |\Delta f|,\ n = n_1 \cdot n_2 \tag{5}$$

式(5)中，如果 $F \neq 0$，则表明外部环境已发生变化，此时应设定一个响应阈值 F_T，当 $F > F_T$ 时触发响应，更新机制为按一定方式重新初始化感知粒子和粒子速度，描述如下：

$$\text{if } F > F_T \text{ then} \begin{cases} V(i) = rand(M) \times V_{\max} \\ X(i) = rand(M) \times X_{\max} \end{cases} \tag{6}$$

式中：$rand(M)$ 为 M 维向量；$V(i) = [V_{i1}, V_{i2}, \cdots, V_{im}]$，$X(i) = [X_{i1}, X_{i2}, \cdots, X_{im}]$ 表示从粒子群中选出的重新进行初始化的第 i 个感知粒子，m 表示维数；$V_{\max}$ 为感知粒子最大速度，一般取 $V_{\max} = X_{\max}$。

4　结果与讨论

4.1　转基因棉种子的太赫兹光谱识别

研究表明大多数分子振动频率都在太赫兹频段内，主要表现在分子的低频集体振动模式，其位置和强度与分子机构、所处的环境及分子间相互作用等因素有关。对于不同转基因棉种子其内部分子结构不一样可以表现为太赫兹时域及频域响应的差异，如图 2 和 3 所示。图 4 所示为 3 类转基因棉种子在太赫兹下的特征吸收峰，从图 4 中可以看出 3 种转基因棉种子均呈现出独特的吸收峰，其中鲁研棉 36 号的吸收峰位于在 0.57、0.80、0.98 THz；鑫秋 638 号的吸收峰位于 0.57、0.75、0.94 THz；新陆中 6 号的吸收峰位于 0.55 THz。由此，本文可以根据不同转基因棉种子呈现出的不同吸收峰对转基因棉种子进行区分。

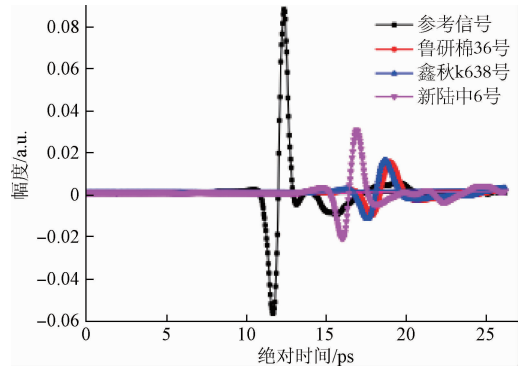


图 2　3 种转基因棉种子及参考信号 THz 时域光谱图

Fig. 2　THz time domain spectroscopy diagram of three kinds of genetically modified cotton seed and reference signal

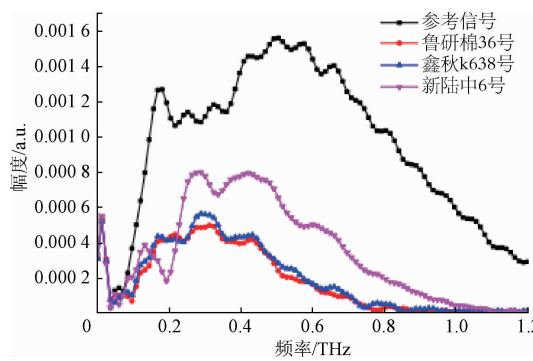


图 3　3 种转基因棉种子及参考信号 THz 频域光谱图

Fig. 3　THz frequency domain spectroscopy diagram of three kinds of genetically modified cotton seed and reference signal

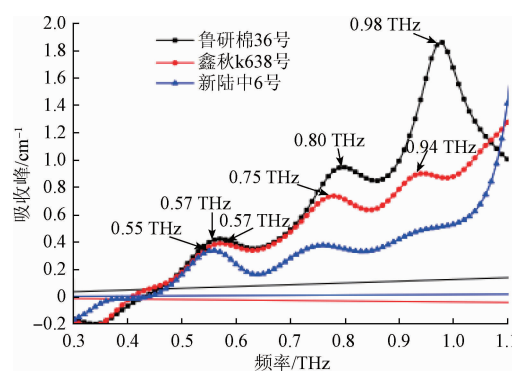


图 4　3 种转基因棉种子 THz 吸收峰光谱图

Fig. 4　THz absorption peak spectroscopy diagram of three kinds of genetically modified cotton seed

4.2　APSO-SVM 模型的优化

APSO-SVM（adaptive particle swarm optimization and support vector machine）的算法思想如下：1）在可行域空间内，把粒子群分为 n_1 均匀子空间，在每个子空间内随机选择 n_2 个感知粒子进行初始化，共产生 $n_1 \times n_2$ 个感知

粒子;2)初始化参数:惯性权重 ω,松弛系数 ξ,惩罚系数 γ,学习因子 c_1,c_2,动态响应触发阈值 F_T;3)按照适应度函数计算每个感知粒子的适应度,计算局部最优和全局最优;4)按式(4)和(5)迭代产生新感知粒子群,并计算适应度值;5)比较新粒子群的适应度值和个体局部最优值、群体局部最优值,并更新个体局部最优值和群体局部最优值;6)计算并判断 F,若 $F > F_T$,则按比例更新粒子群及粒子速度,转3算法,当结束条件满足时,则算法结束。

图5所示为通过PSO-SVM(particle swarm optimization and support vector machine)和APSO-SVM优化支持向量机得到的粒子群迭代次数与适应度值关系曲线对比图。从图5可见,APSO-SVM模型在粒子群进化到50代后可达到其最优解,PSO-SVM模型要迭代100次才达到最优解。其具体的参数值及分类正确率见表1所示。

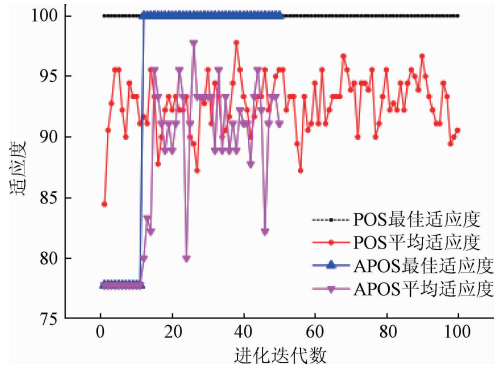


图5　PSO与APSO迭代次数与适应度值关系曲线

Fig. 5　Relation curve of the number of iterations and fitness value of PSO and APSO

表1　PSO与APSO各参数对比

Table 1　Comparison of parameters of PSO and APSO

方法	C1	C2	Best c	Best g	γ	σ	种群数	迭代次数	交叉验证数	训练集正确率	测试集正确率
PSO	0.5	1.7	0.01	39.458 1	750.45	3.99	20	100	5	98.9%(89/90)	89.3%(67/75)
APSO	0.5	1.7	0.01	1.632 9	563.48	3.86	10	50	5	98.9%(89/90)	97.3%(73/75)

4.3　判别分析

将3类转基因棉种子的165个样本分成2组,第1组90个作为训练集,第2组75个作为测试集,并保证每类转基因棉种子的30个样本在训练集中,25个样本在测试集中,如表2所示。表3、4给出了两种不同方法对转基因棉花的分类结果。从表1中可以看出PSO-SVM对转基因棉种子的综合识别率为89.3%,APSO-SVM对转基因棉种子的综合识别率为97.3%.由此可以证明该方法可有效地识别转基因类别。

表2　实验样品表

Table 2　The experimental samples table

种类	训练集	测试集
鲁研棉36号	30	25
鑫秋k638号	30	25
新陆中6号	30	25

表3　PSO-SVM模型对样品训练集和测试集的识别情况

Table 3　The results of identification of sample train set and test set based on PSO-SVM model

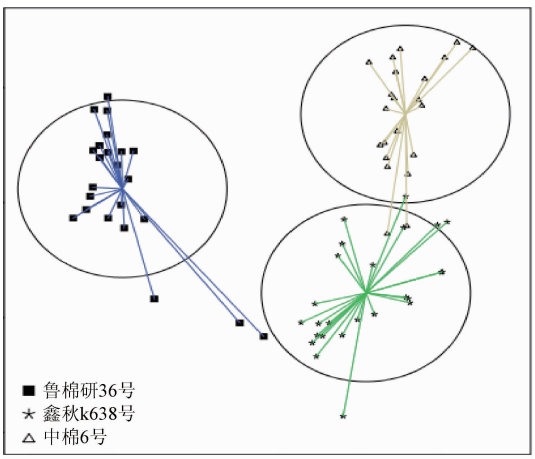
PSO-SVM模型		样品名	鲁棉研36号	鑫秋k638号	新陆中6号	识别数
训练集	识别情况	鲁棉研36号	29	1	/	29
		鑫秋k638号	/	30	/	30
		新陆中6号	/	/	30	30
	识别率/%	鲁棉研36号	96.7	3.3	/	96.7
		鑫秋k638号	/	100	/	100
		新陆中6号	/	/	100	100
测试集	识别情况	鲁棉研36号	23	2	/	23
		鑫秋k638号	1	22	2	22
		新陆中6号	/	3	22	22
	识别率/%	鲁棉研36号	92	8	/	92
		鑫秋k638号	4	88	8	88
		新陆中6号	/	12	88	88

表 4　APSO-SVM 模型对样品训练集和测试集的识别情况

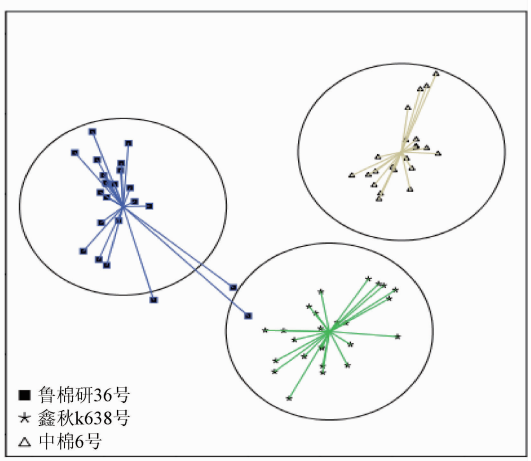
Table 4　The results of identification of sample train set and test set based on APSO-SVM model

PSO-SVM 模型		样品名	鲁棉研 36 号	鑫秋 k638 号	新陆中 6 号	识别数
训练集	识别情况	鲁棉研 36 号	29	1	/	29
		鑫秋 k638 号	/	30	/	30
		新陆中 6 号	/	/	30	30
	识别率/%	鲁棉研 36 号	96.7	3.7	/	96.7
		鑫秋 k638 号	/	100	/	100
		新陆中 6 号	/	/	100	100
测试集	识别情况	鲁棉研 36 号	23	2	/	23
		鑫秋 k638 号	/	25	/	25
		新陆中 6 号	/	/	25	25
	识别率/%	鲁棉研 36 号	92	8	/	92
		鑫秋 k638 号	/	100	/	100
		新陆中 6 号	/	/	100	100

为了更加直观的比较 PSO-SVM 和 APSO-SVM 模型的识别结果，两种方法对 3 种类别的样品的识别结果二维图如图 6 所示，从图中可以看出，PSO-SVM 方法中 3 种类别的样品都出现了不同程度的误判，而 APSO-SVM 方法只对鲁棉研 36 号出现少量误判，对其他两种类别的样品识别率达到了 100%，由此可以得出本文提出的方法能够有效对转基因样品进行鉴别。



(a) PSO-SVM方法
(a) Method of PSO-SVM



(b) APSO-SVM方法
(b)Method of APSO-SVM

图 6　PSO-SVM 与 APSO-SVM 方法对 3 种样品的识别结果对比

Fig. 6　The comparison of identification results of three kinds of samples of PSO-SVM and APSO-SVM

5　结　论

通过太赫兹光谱检测，结合 APSO-SVM，建立了转基因棉种子的识别模型。结果表明，该模型对转基因棉种子的识别率达 100%，为定性分析模型在实际样品检测中的应用奠定了基础。该方法为转基因棉种子类型辨别提供了一种精确、快速、简便、无损的检测方法，也对其它转基因产品检测有一定的指导意义。

尽管本实验取得了较好的效果，但所用实验样品类型和数量较少，可能会对实验结果的稳定性和精度带来一些影响。鉴于此，今后的研究方向可向新的光谱数据采集和处理方法进行展开及对多种样品进行测试，从而提高建模及实验结果的稳定性和精度，为便携式转基因产品检测技术提供帮助和支持。

参考文献

[1] LEE J H, CHOUNG M G. Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy[J]. Food Chemistry, 2011, 126(1): 368-373.

[2] IVANIRA M, SPACINO S I. Chemometric discrimination of genetically modified coffea arabica cultivars using spectroscopic and chromatographic fingerprints[J]. Talanta, 2013, 107(3): 416-422.

[3] LIANG M Y, SHEN J L, WANG G Q. Identification of illicit drugs by using SOM neural networks[J]. Journal of Physics D Applied Physics, 2008, 41(13): 306-310.

[4] LIU J, LI ZH, HU F, et al. Identification of transgenic organisms based on terahertz spectroscopy and hyper

sausage neuron[J]. Journal of Applied Spectroscopy, 2015,82(1):104- 110.

[5] 范玉妹,郭春静. 支持向量机算法的研究及其实现[J]. 河北工程大学学报:自然科学版,2010,27(4):106-112.
FAN Y M, GUO CH J. Study and implement of support vector machine algorithm[J]. Journal of Hebei University of Engineering: Natural Science Edition, 2010, 27(4): 106-112.

[6] WANG L, HE Y, LIU F, et al. Rapid detection of sugar content and pH in beer by using spectroscopy technique combined with support vector machines[J]. Journal of Infrared and Millimeter Waves, 2008, 27(2): 51.

[7] 王道明,鲁昌华,蒋薇薇,等. 基于粒子群算法的决策树SVM多分类方法研究[J]. 电子测量与仪器学报,2015,49(4):611-615.
WANG D M, LU CH H, JIANG W W, et al. Study on PSO-based decision-tree SVM multi-class classification method[J]. Journal of electronic measurement and instrument, 2015, 49(4): 611-615.

[8] 焦卫东,林树森. 整体改进的基于支持向量机的故障诊断方法[J]. 仪器仪表学报,2015,36(8):1861-1870.
JIAO W D, LIN SH S. Overall-improved fault diagnosis approach based on support vector machine[J]. Chinese Journal of Scientific Instrument, 2015, 36(8): 1861-1870.

[9] 邬啸,魏延,吴瑕. 基于混合核函数的支持向量机[J]. 重庆理工大学学报:自然科学,2011,25(10):66-70.
WU X, WEI Y, WU X. Support vector machine based on hybrid kernel function[J]. Journal of Chongqing University of Technology: Natural Science, 2011, 25(10): 66-70.

[10] LOU D C, LIU C L, LIN C L. Message estimation for universal steg analysis using multiclassification support vector machine[J]. Computer Standards & Interfaces, 2009, 31(2): 420-427.

[11] 张艳梅,姜淑娟,陈若玉,等. 基于粒子群优化算法的类集成测试序列确定方法[J]. 计算机学报,2016,39(55):1-18.
ZHANG Y M, JIANG SH J, CHEN R Y. Chinese class integration testing order determination method based on particle swarm optimization algorithm[J]. Journal of Computers, 2016, 39(55): 1-18.

[12] OBEIDAT F A, BELACEL N, CARRETERO J A, et al. An evolutionary framework using particle swarm optimization for classification method PROAFTN[J]. Applied Soft Computing, 2011, 11(8): 4971-4980.

[13] 梁旭,刘才慧. 基于混合粒子群算法的在线检测路径规划[J]. 国外电子测量技术,2015,34(12):30-34.
LIANG X, LIU C H. Path planning for on machine verification system based on hybrid particle swarm optimization algorithm[J]. Foreign Electronic Measurement Technology, 2015, 34(12): 30-34.

[14] JIANG Y, HU T, HUANG C C, et al. An improved particle swarm optimization algorithm[J]. Applied Mathernatics and Computation, 2007, 193(1): 231-239.

[15] 胥小波,郑康锋,李丹,等. 新的混沌粒子群优化算法[J]. 通信学报,2012,33(1):24-31.
XU X B, ZHENG K F, LI D, et al. New chaos-particle swarm optimization algorithm[J]. Journal on Communications, 2012, 33(1): 24-31.

[16] ZHANG W, LIU J, NIU Y Q. Quantitative prediction of MHC-II binding affinity using particle swarm optimization[J]. Artificial Intelligence in Medicine, 2010, 50(2): 127-132.

[17] KARAKUZU C. Parameter tuning of fuzzy sliding mode controller using particle swarm optimization[J]. International Journal of Innovative Computing, Information and Control, 2010, 6(10): 4755-4770.

作者简介

潘学文,1983年出生,湖南科技学院电子与信息工程学院讲师,主要研究方向为太赫兹检测技术。
E-mail:xwpan2005@sina.com.cn

Pan Xuewen was born in 1983, lecturer in School of Electronics and Information Engineering, Hunan University of Science and Engineering. His main research interest is terahertz detection technology.

尹向东,1976年出生,湖南科技学院电子与信息工程学院教授,研究方向为无线通信技术、信息安全等。
E-mail: 390375334@qq.com

Yin Xiangdong was born in 1976, professor in School of Electronics and Information Engineering, Hunan University of Science and Engineering. His main research interests include wireless communication technology, information security and so on.

赵永红, 1980年出生,西安电子科技大学测试计量技术及仪器专业博士研究生,主要研究方向为太赫兹检测技术。
E-mail: zhyho00824@126.com

Zhao Yonghong was born in 1980, Ph.D. candidate in School of Mechano-Electronic Engineering, Xidian University. His main research interest is terahertz detection technology.