[文章编号] 1007－7405(2015)06－0421－07

基于太赫兹光谱和支持向量机快速检测棉花种子

杜 勇[1]，刘建军[2]

（1. 集美大学信息工程学院，福建 厦门 361021；2. 九江学院电子工程学院，江西 南昌 330013）

[摘要] 鉴于目前对农产品品种的检测大多是基于可见光/近红外光谱的，提出了一种基于太赫兹光谱和支持向量机快速检测棉花种子的方法. 为实现棉花种子的分类识别，在频率0.2～1.2 THz范围内采集2种最新转基因及2种非转基因棉花种子，总计40个样本的太赫兹光谱，用遗传算法优化的支持向量机建立识别模型，对不同品种的棉花种子进行识别. 实验结果表明，该方法对不同品种的棉花种子综合识别率达到93.75%，由此，太赫兹光谱结合支持向量机的检测方法可为不同品种的生物辨别提供一种精确、快速、简便的检测方法.

[关键词] 太赫兹；光谱；支持向量机；棉花；种子；检测；遗传算法

[中图分类号] TN 29；O 657.3 [文献标志码] A

Rapid Detection of Cotton Seed Based on THz Spectroscopy Combined with SVM

DU Yong[1], LIU Jian-jun[2]

(1. School of Information Engineering, Jimei University, Xiamen 361021, China;
2. School of Electronic Engineering, Jiujiang University, Jiujiang 330013, China)

Abstract: At present, the detection of agricultural products is mostly based on visible/near infrared spectroscopy. In view of this, a fast and non-destructive detection method of cotton seeds based on terahertz spectroscopy combined with Support Vector Machine (SVM) was proposed. For the classification and recognition of different varieties cotton seeds, the terahertz spectra of two kinds of transgenic and two kinds of non-transgenic cotton seeds containing 40 samples in total were collected in the frequency range of 0.2～1.2 THz, using the Genetic Algorithm (GA) to optimized support vector machine. A recognition model to recognize different varieties of cotton seeds was established. The experimental results showed that the recognition rate of cotton seeds reached 93.75%. Therefore, the terahertz spectroscopy combined with support vector machine may provide an accurate, fast and simple method for the detection of different varieties of organisms.

Key words: THz; spectrum; SVM; cotton; seed; detection; GA

0 引言

太赫兹波通常是指频率在0.1～10 THz（波长在3～30 mm）之间的电磁波[1]，其波段属于远红外. 理论研究表明，大量生物分子（DNA，蛋白质等）的振动和转动能级正好处于THz的频带

[收稿日期] 2015－04－09 [修回日期] 2015－11－02
[基金项目] 福建省自然科学基金资助项目（2013J01246）
[作者简介] 杜勇（1971—），男，副教授，硕士，主要从事光电子器件研究，E-mail:duyong2001@jmu.edu.cn.

范围内，用 THz 时域光谱系统（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成为可能. 目前近红外光谱在转基因植物检测中的应用已经相当广泛[2]，文献[3]报道了近红外光谱在转基因玉米检测识别中的应用，文献[4]报道了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，谢丽娟等[5]报道了利用可见光/近红外光谱分析技术鉴别转基因番茄叶等. 但是，作为可见光/近红外光谱技术有益补充的太赫兹光谱技术在农业和食品领域的研究和探索才刚刚开始[6].

支持向量机是由 Vapanik 等人[7]提出的一种机器学习方法. 其基本思想是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性分类，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[8-10]. 目前，在支持向量机优化参数问题上采用较多的是网格搜索（Grid Search）与交叉验证相结合的优化算法，但是，该方法有个致命的缺点就是当训练样本较大时搜索过程非常费时，且计算量大，因此该方法具有一定的局限性. 而遗传算法（Genetic Algorithm，GA）具有全局搜索能力，能够在很大程度上减少计算量，使之优化支持向量机成为可能.

本文拟使用太赫兹光谱检测系统对 2 种转基因棉花种子和 2 种非转基因棉花种子的 40 个样本进行光谱扫描，并在传统支持向量机的基础上，利用遗传算法优化支持向量机，以这 4 种棉花种子的太赫兹特征吸收谱为训练集数据，对其进行识别.

1 样品的 THz 特征吸收谱线

1.1 实验装置

图 1 为本文所用的透射式太赫兹时域光谱系统（THz－TDS），其中：InAs 为 THz 发射极；ZnTe 为探测极；Chopper 为斩波器；BS 为分束器；HWP 为二分之一波片；QWP 为四分之一波片；M1～M12 为平面反射镜；PM1～PM4 为离轴抛物面镜；Sample 表示样品放置处；L1～L3 为聚焦透镜；P 为检偏器；PBS 为沃拉斯顿棱镜；Si 为硅片，可以透射太赫兹波，反射飞秒激光；Detector 为差分二极管. Detector 的输出信号接入锁相放大器，通过计算机进行数据采集. 图 2 为图 1 虚线部分的系统照片，实验测量时，该虚线部分置于氮气环境中. 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. 实验时室内相对湿度为 25%，恒温 292 K.

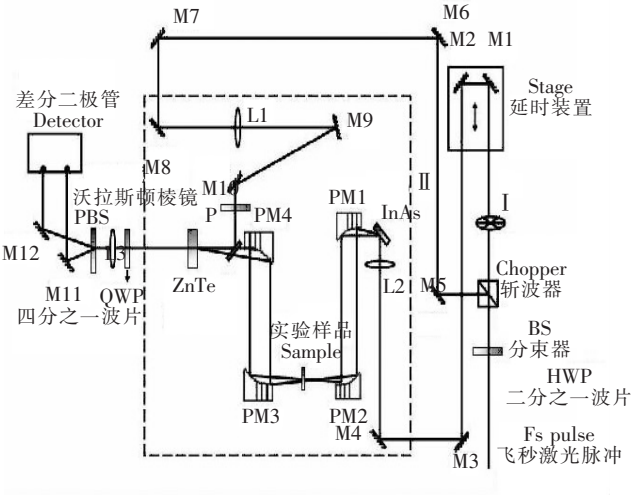


图 1 THz-TDS 系统示意图

Fig.1 The schematic diagram of THz-TDS

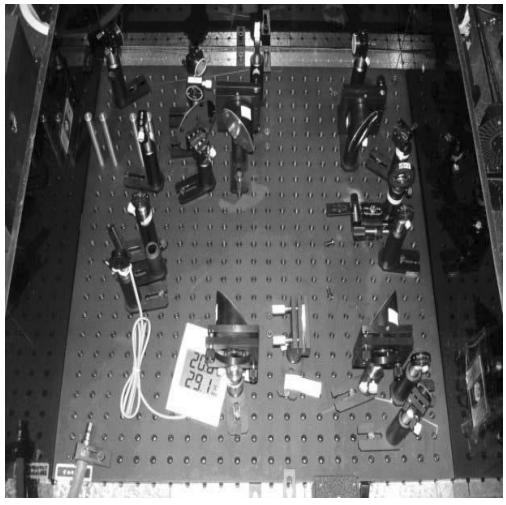


图 2 THz-TDS 系统示意图中虚线部分的照片

Fig.2 Picture of the dashed part of the THz-TDS

1.2　实验对象及样品制备

以 4 种不同品种的棉花种子为研究对象，它们是转基因种子银棉 8 号（Yinmian No. 8）、鑫秋 107 号（Xinqiu No. 107），和非转基因棉花种子新陆中 6 号（Xinluzhong No. 6）、中棉所 28 号（CCRI 28），均购于中国农业科学院生物技术研究所. 将棉花种子磨碎、烘干后，用压片机压成圆盘状，直径 13 mm. 每种棉花种子制成 10 个样片，将其中的 24 个样片（每种 6 个）作为训练集数据用于支持向量机建模校正；剩下的 16 个样片作为测试集数据用来验证模型的精确度.

1.3　样品的特征吸收谱

将制作好的样品置于 THz – TDS 中，扫描得到 4 种棉花种子的 THz 时域光谱信息，如图 3 所示. THz 时域光谱信息经过快速傅立叶变换（FFT），得到如图 4 所示的 4 种棉花种子 THz 频谱图. 可以看出，在 0.2 ~ 1.2 THz 的光谱有效区域内，4 种样品信号与自由空间的参考信号显著不同. 通过测量样品对 THz 脉冲的相位延迟和吸收可以计算出材料的吸收率. 根据测量的自由空间的 THz 参考信号以及透过物质的 THz 样品信号，得到 4 种棉花种子的特征吸收谱线，如图 5 所示.

由于 4 种棉花种子其内部分子结构不一样，可以表现出太赫兹时域及频域响应的差异. 由图 5 可知：鑫秋 107 号的吸收峰位于 0.57，0.80 THz. 需要指出的是，1.00 THz 处不是样品本身的吸收峰. 由于实验设备的分辨范围为 0.10 ~ 0.98 THz，因而 1.00 THz 以上为不确定因素引起的误差. 银棉 8 号的吸收峰位于 0.57，0.78，0.94 THz. 新陆中 6 号的吸收峰位于 0.55，0.76 THz. 中棉所 28 号在 0.20 ~ 1.10 THz 内没有明显的吸收峰. 由此，本研究可以根据不同品种的棉花种子呈现出的不同吸收峰来区分 4 种不同品种的棉花种子，但它还不能鉴别转基因与非转基因的棉花种子.

2　基于遗传算法的支持向量机

利用得到的样品的太赫兹光谱数据，建立基于遗传算法的支持向量机识别模型.

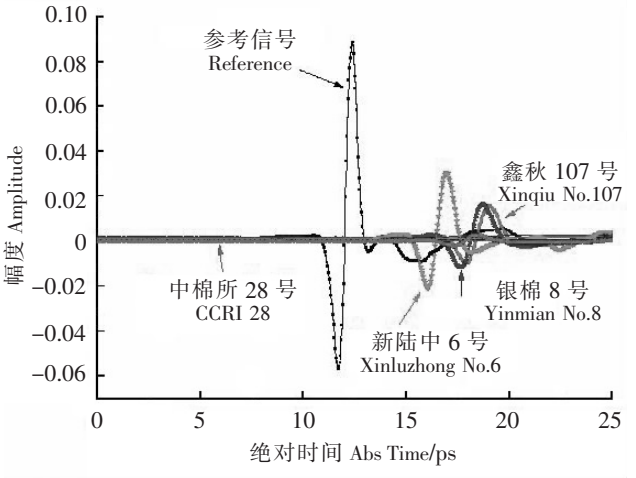


图 3　4 种棉种子的 THz 时域光谱图

Fig.3　The THz time-domain spectroscopy of four kinds of cotton seeds

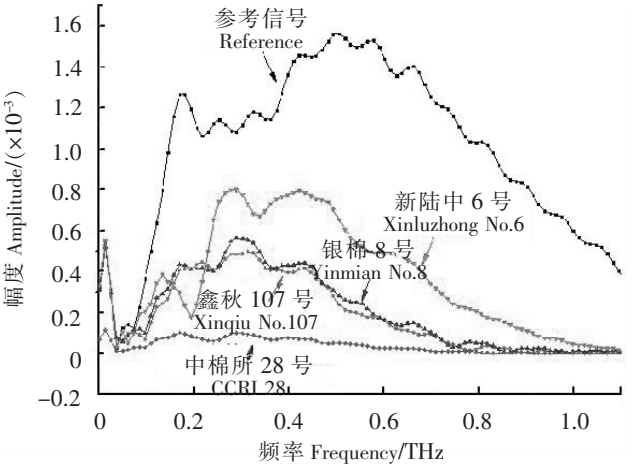


图 4　4 种棉种子的 THz 频域光谱图

Fig. 4　The THz frequency domain spectrum of four kinds of cotton seeds

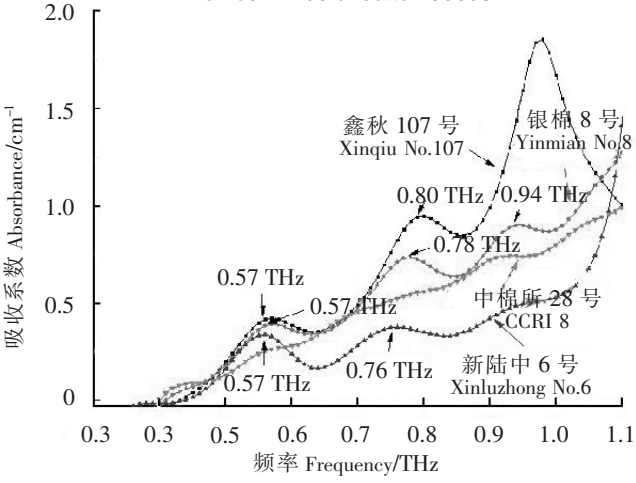


图 5　4 种棉种子的 THz 吸收峰光谱图

Fig.5　THz absorption spectra of four kinds of cotton seeds

2.1 遗传算法优化支持向量机

遗传算法（GA）从编码串群体出发，从中选择适应度值高的个体进行复制，利用变异交叉原理产生新的群体. 随着遗传算法的进行，新群体中个体适应度值不断提高，最终得到适应度高的个体，即为优化后的最优解. 用均方相对误差 D_r 评价模型的最终性能：$D_r = \sqrt{\sum_{l=1}^{l}((y_{ti} - y_{pi})/y_{ti})^2/l}$，其中 y_{ti} 表示实际值，y_{pi} 表示预测值，l 是样品数.

对于采用 BRF 核函数的支持向量机，其学习和泛化能力在很大程度上受惩罚系数 γ 和核函数值 g 的影响，因此利用遗传算法优化支持向量机的问题可以简化为寻找参数 γ 和 g 的最佳组合值，得到最优化的支持向量机参数 γ 和 g. 其优化算法为：

1）对参数 γ 和 g 进行染色体基因编码[11]，设置遗传算法交叉率、变异率，随机产生支持向量机参数值.

2）计算种群个体的适应度值，用适应度函数 $f(\gamma,g) = 1/D_r$ 衡量参数的好坏.

3）根据适应度值，利用交叉变异算子产生新个体.

4）初始设定种群数 P 和进化代数 T，利用 $P_c = \begin{cases} c_1(f_{best} - f')/(f_{best} - f_{avg}) & f' \geqslant f_{avg} \\ c_2 & f' < f_{avg} \end{cases}$，$P_m = \begin{cases} c_3(f_{max} - f)/(f_{max} - f_{avg}) & f \geqslant f_{avg} \\ c_4 & f < f_{avg} \end{cases}$，确定交叉概率 P_c 和变异概率 P_m. 其中：f_{best} 是个体最优适应度；f' 为两交叉个体中较大的适应度；f_{avg} 为平均适应度；f_{max} 是个体最大适应度；f 为变异个体适应度；$c_1, c_2, c_3, c_4 \in (0,1)$，是支持向量机控制参数.

5）重复步骤 2）—4），不断更新支持向量机参数，直到满足结束条件，得到最优化的支持向量机参数 γ_{best} 和 g_{best}.

2.2 基于遗传算法的支持向量机识别模型的建立

支持向量机识别模型建立步骤如下：

1）获取得到样品的太赫兹光谱数据，把所有样品的太赫兹光谱数据分成 n 份，将其中 k 份用来作为训练集数据.

2）读取训练集样本的太赫兹数据，随机产生一组 $\{\gamma,g\}$（即空间坐标值）作为遗传算法种群个体的初始位置，训练支持向量机.

3）根据种群个体的初始化位置，计算种群个体的适应度值，利用交叉变异算子产生新个体.

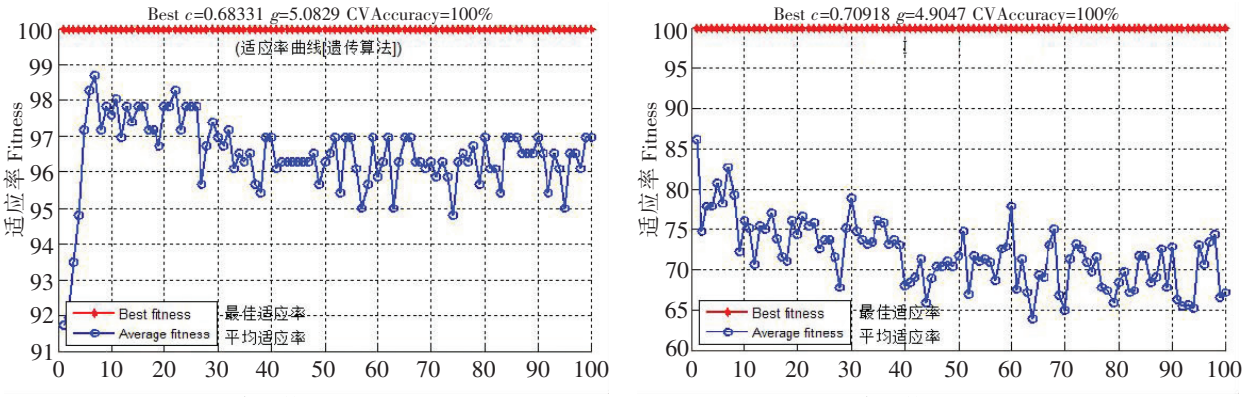
4）根据种群个体适应值不同，采取调整支持向量机惯性权重，不断更新支持向量机参数，直到满足结束条件（终止条件应根据实际情况来定，本文选择的终止条件为当训练误差率小于 1.5% 时终止算法），得到最优化的支持向量机参数 γ 和 g.

5）根据输出的最优 $\{\gamma,g\}$，建立支持向量机识别模型.

3 实验结果及分析

为了验证利用遗传算法对支持向量机的参数进行优化的有效性，本文给出了 GA 和 Grid Search 两种方法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情况. 由图 6 可知，两种方法在粒子群进化到 100 代后都可达到其最优解，但是，Grid Search 方法的适应率低于 85%，而 GA 方法适应率高于 94.8%.

用新陆中 6 号（Xinluzhong No. 6）、银棉 8 号（Yinmian No. 8）、鑫秋 107 号（Xinqiu No. 107）和中棉所 28 号（CCRI 28）等 4 种样品中的 24 个样片（每种 6 个）作为训练集数据建立遗传算法支持向量机的识别模型，将剩下的 16 个样片（每种 4 个）作为待测样品来验证本文模型的准确性.



a GA　　b Grid Search

图 6　GA 与 Grid Search 方法的迭代次数与适应度值关系曲线对比图

Fig.6　Iterations and fitness values of GA and Grid Search methods

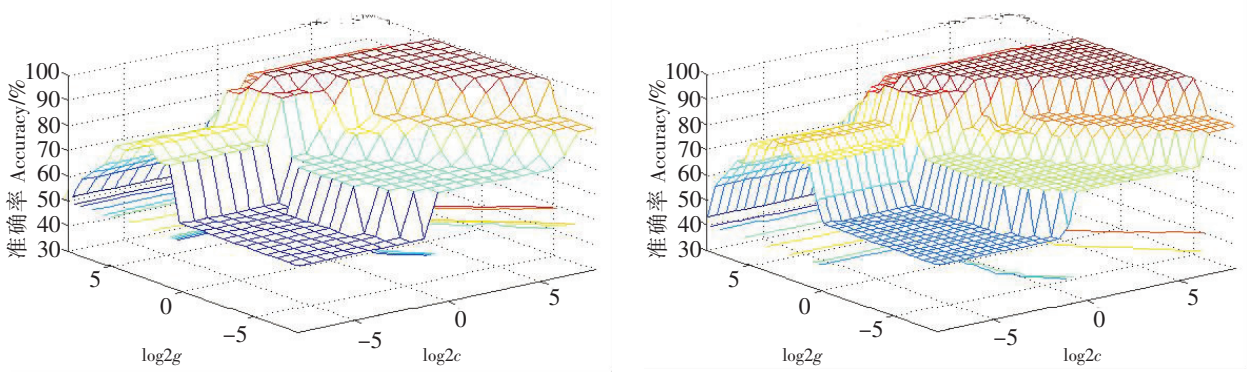
表 1 给出了经过 PCA 降维后的优化 SVM 参数值及分类正确率. 由表 1 可以看出，Grid Search 和 GA 两种方法，对训练集都有 100% 的辨别率，而对预测集，Grid Search 方法只有 81.25% 的辨别率，GA 方法辨别率则可以达到 93.75%.

表 1　Grid Search 与 GA 的各参数对比

Tab. 1　Comparison of parameters based on Grid search and GA methods

方法 Method	γ_{best}	g_{best}	种群数/种 Number of population	迭代次数/次 Iterations	交叉验证数 Cross validation	训练集准确率 Train set accuracy	测试集准确率 Test set accuracy
Grid Search	0.354	5.657	20	200	5	100% (24/24)	81.25% (13/16)
GA	1.755	9.036	20	100	5	100% (24/24)	93.75% (15/16)

图 7 从 3D 角度对 SVM 参数选择进行了比较直观的对比，从中可以看出 GA 方法能够较全面地寻找到 SVM 参数的全局最优解.



a Grid Search　　b GA

图 7　Grid Search 和 GA 优化参数选择结果 3D 对比图

Fig.7　The 3D comparison chart of optimized parameter based on the Grid Search and the GA methods

表 2 为采用 Grid Search – SVM 和 GA – SVM 两种方法对实验样品进行综合识别的结果. 由表 2 可知：GA – SVM 方法对 4 种不同品种的棉花种子的综合识别率为 93.75%，高于 Grid Search – SVM 方法（81.25%），说明该方法可有效地识别不同品种的棉花种子.

表 2 Gird Search 和 GA 方法对 4 种棉花种子识别率的对比

Tab. 2 Comparison of Gird Search and GA methods for the recognition rate of four cotton seeds

品种 Variety	项目 Item	网格算法支持向量机 Grid Search – SVM		遗传算法支持向量机 GA – SVM	
		训练集 Train set	测试集 Test set	训练集 Train set	测试集 Test set
鑫秋 107 号 Xinqiu No. 107	样品 Samples	6	4	6	4
	识别样品数 Recognized samples	6	3	6	4
	识别率 Recognition rate	100% (6/6)	75% (3/4)	100% (6/6)	100% (4/4)
银棉 8 号 Yinmian No. 8	样品 Samples	6	4	6	4
	识别样品数 Recognized samples	6	4	6	4
	识别率 Recognition rate	100% (6/6)	100% (6/6)	100% (6/6)	100% (4/4)
新陆中 6 号 Xinluzhong No. 6	样品 Samples	6	4	6	4
	识别样品数 Recognized samples	6	3	6	3
	识别率 Recognition rate	100% (6/6)	75% (3/4)	100% (6/6)	75% (3/4)
中棉所 28 号 CCRI 28	样品 Samples	6	4	6	4
	识别样品数 Recognized samples	6	3	6	4
	识别率 Recognition rate	100% (6/6)	75% (3/4)	100% (6/6)	100% (4/4)
总识别率 Total recognition rate		100% (24/24)	81.25% (13/16)	100% (24/24)	93.75% (15/16)

图 8 描述了用 Grid Search 和 GA 两种方法优化 SVM 后对 4 种棉花种子测试样本的分类情况. 从分类结果可以看出，本文方法能够更加准确地区分 4 种棉花种子. 同时从图 8 还可看出用 Grid Search 和 GA 两种方法优化 SVM 都能将 4 种棉花种子分类成两大类，其中 class1 代表的是转基因类别的棉花，class2 代表的是非转基因类别的棉花，后续研究中将进一步研究引起该现象的原因.

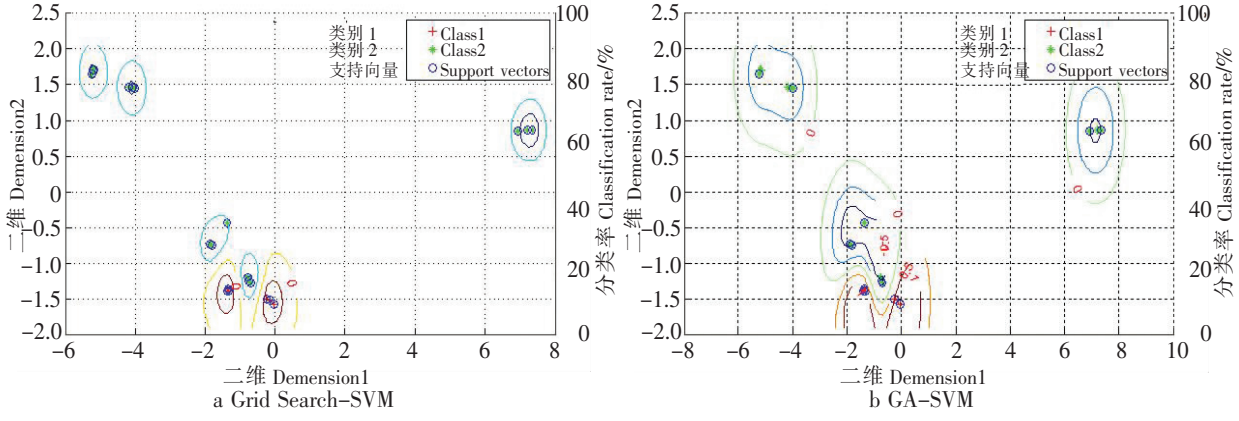


图 8 Grid Search–SVM 与 GA–SVM 对预测样本分类 2D 对比图

Fig.8 The 2D comparison chart of prediction samples classified based on Grid Search–SVM and GA–SVM methods

4　结论

利用太赫兹光谱特性，结合 GA－SVM 方法，建立了 4 种不同品种的棉花种子的识别模型. 结果表明，该模型对 4 种棉花种子的识别率达 93.75%，为定性分析模型在实际样品检测中的应用奠定了基础. 因为不同品种的棉花在基因表达和蛋白合成上存在差异，在成长过程当中就表现为生物分子的不同，因此可以利用这些不同成分在太赫兹光谱上呈现出不同的特性来对其进行鉴别. 同理，由于其他转基因作物，如水稻、大豆等代谢产物或蛋白质水平与其亲本相比会发生变化，应该会在太赫兹光谱上呈现出不同的特性，因此本文方法可能用于其他非同类物质的检测. 在实验中还发现，利用本文方法，可以将 4 种棉花种子分成转基因和非转基因两个不同的大类，但该现象具体是否由转基因引起，本文尚未做进一步研究，但为后续研究提供了方向. 鉴于此，今后应进一步研究转基因棉花与其亲本的检测方法，同时研究新的光谱数据采集和处理方法，从而提高建模的稳定性和精度，为便携式转基因产品检测的开发研制提供技术支持.

[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] MOREIRA IVANIRA, SCARMINIO IEDA SPACINO. Chemometric discrimination of genetically modified *Coffea arabica* cultivars using spectroscopic and chromatographic fingerprints [J]. Talanta, 2013, 107(30): 245-254.

[3] BORJIGIN M, ESKRIDGE C, NIAMAT R, et al. Electrospun fiber membranes enable proliferation of genetically modified cells [J]. International Journal of Nanomedicine, 2013, 8: 855-864.

[4] MILCAMPS A, RABE S, CADE R, et al. Validity assessment of the detection method of maize event Bt10 through investigation of its molecular structure [J]. Journal of Agricultural and Food Chemistry, 2009, 57(8): 3156-3163.

[5] FIEHN O, KOPKA J, TRETHEWEY R N, et al. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry [J]. Analytical Chemistry, 2000, 72(15): 3573-3580.

[6] 李斌, WANG Ning, 张伟立, 等. 基于太赫兹光谱技术的山核桃内部虫害检测初步研究 [J]. 光谱学与光谱分析, 2014, 34(5): 1196-1200.

[7] VAPNIK V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[8] BURGES C J C. A tutorial on support vector machines for pattern recognition [J]. Data Min Knowl Disc, 1998, 2(2): 121-167.

[9] SNCHEZ A V D. Advanced support vector machines and kernel methods [J]. Neurocomputing, 2003, 55(3): 5-20.

[10] VAPINK V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999. DOI: 10.1109/72.788640.

[11] 张超群, 郑建国, 钱洁. 遗传算法编码方案比较 [J]. 计算机应用研究, 2011, 28(3): 819-822.

(责任编辑　朱雪莲　英文审校　曹敏杰)