基于太赫兹光谱和支持向量机快速鉴别咖啡豆产地

胡晓华[1]，刘 伟[2,3※]，刘长虹[3]，钱赟惠[3]

（1. 合肥工业大学计算机与信息学院，合肥 230009； 2. 合肥学院机器视觉与智能控制实验室，合肥 230601；
3. 合肥工业大学食品科学与工程学院，合肥 230009）

摘 要：结合太赫兹时域光谱技术和支持向量机对 3 种典型产地的咖啡豆进行了鉴别。选取埃塞俄比亚（Ethiopia）、哥斯达黎加（Costa Rica）以及印度尼西亚（Indonesia）3 个产地咖啡豆样品进行压片处理，采用太赫兹透射模式获取样品的时域和频域光谱信号，并用主成分分析法对太赫兹频域光谱信号进行分析；构造了基于粒子群（partical swarm optimization，PSO）参数寻优的支持向量机（support vector machine，SVM）鉴别模型，模型对不同产地咖啡豆样品的综合识别正确率达到 95%。试验结果表明，太赫兹作为新型的检测手段结合模式识别方法可用于咖啡豆的产地鉴别。该文为一类在太赫兹波段下没有明显特征吸收峰的农产品/食品安全检测和产地追溯研究提供了一种快速、准确的方法。

关键词：光谱学；模型；支持向量机；咖啡豆；太赫兹；粒子群算法
doi：10.11975/j.issn.1002-6819.2017.09.040
中图分类号：TP274+.3; TP391.44 **文献标志码**：A **文章编号**：1002-6819(2017)-09-0302-06

胡晓华，刘 伟，刘长虹，钱赟惠. 基于太赫兹光谱和支持向量机快速鉴别咖啡豆产地[J]. 农业工程学报，2017，33(9)：302－307. doi：10.11975/j.issn.1002-6819.2017.09.040 http://www.tcsae.org
Hu Xiaohua, Liu Wei, Liu Changhong, Qian Yunhui. Rapid identification of producing area of coffee bean based on terahertz spectroscopy and support vector machine[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(9): 302－307. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2017.09.040 http://www.tcsae.org

0 引 言

咖啡是世界 3 大饮料作物之一，其产量、销售量、消费量均居世界 3 大饮料植物之首。近年来中国咖啡的进口量增长迅速，年均增长率超过 10%，已成为重要的大宗进口消费品。咖啡豆是制作咖啡的主要原材料，目前世界上咖啡豆的种植主要集中在拉丁美洲、非洲、亚洲等的热带发展中国家，如印度利西亚、埃塞俄比亚、巴西、哥伦比亚、哥斯达黎加等。不同产地的咖啡豆其外观色泽、气味以及内部化学成分存在较大差异，是影响咖啡品质的重要因素[1-3]。目前，咖啡豆的产地鉴别主要采用人工感官评定法或化学分析法[4-6]，存在方法繁琐、主观性强、效率低下等缺点。因此，如何快速准确地鉴别咖啡豆产地，保障咖啡品质，规范咖啡市场，是中国咖啡产业亟待解决的重要问题之一。

太赫兹（Terahertz，THz）是指频率在 0.1~10 THz 范围内的电磁波，研究表明，大量有机大分子（DNA、蛋白质等）的振动能级和转动能级之间的跃迁在 THz 波段，因此太赫兹光谱包含了检测对象丰富的物理、化学和构象信息[7-11]。近年来太赫兹时域光谱（THz-TDS）技术作为一种迅速发展的无损检测新技术，因其具有穿透能力强、安全性好、灵敏度高和动态范围宽等特点，在食品安全检测以及农产品质量控制等方面表现出了较强的技术优势和广泛的应用前景[12-19]。但目前太赫兹在农产品/食品领域的研究多是针对具有特征吸收峰的单一化学成分的检测，在没有光谱特征吸收峰的复杂生物体系中，太赫兹光谱特征往往分布于某些波段范围内，会造成光谱特征的高维性和不确定性等问题。因此，应用太赫兹时域光谱技术进行农产品/食品这一复杂生物体的检测尚处于探索阶段。

本文针对咖啡豆产地的快速鉴别问题，应用太赫兹时域光谱系统获取典型产地咖啡豆样品在太赫兹波段下的时域和频域光谱信息，通过主成分分析（principal component analysis，PCA）法降低光谱特征维度，通过粒子群（partical swarm optimization，PSO）算法进行模型参数优化，采用支持向量机（support vector machine，SVM）构建基于太赫兹光谱技术的鉴别模型，以期为咖啡豆产地的快速鉴别提供一种新方法，同时为太赫兹在农产品/食品中的检测应用做出探索。

1 材料与方法

1.1 试验装置及原理

设备采用 TAS7500TS HF1 THz 光谱系统（Advantest Co., Ltd, JAPAN），仪器光路示意图如图 1 所示。试验采用透射模式，激光脉冲射出后经分光分束器 CBS 分为泵

收稿日期：2017-02-22 修订日期：2017-04-16
基金项目：国家重点研发计划项目（2016YFD0401104）
作者简介：胡晓华，男，江西婺源人，主要从事太赫兹光谱无损检测研究。合肥 合肥工业大学计算机与信息学院，230009。
Email：xiaohuahu@mail.hfut.edu.cn
※通信作者：刘 伟，男，安徽寿县人，高级实验师，博士，主要从事检测技术与模式识别研究。合肥 合肥学院机器视觉与智能控制实验室，230601。
Email：lwei1524@163.com

浦光与探测光。泵浦光入射至砷化镓（GaAs）衬底的光电导天线上，激发 THz 辐射；探测光与 THz 脉冲一同聚焦在电光晶体碲化锌（ZnTe）上，其中 THz 脉冲会被吸收同时受到色散效应影响发生幅值和相位的变化，包含样品信息的 THz 波将聚焦在探测晶体上。系统通过扫描获取 THz 脉冲和探测激光脉冲的相对时间延迟，利用探测光的光电效应对 THz 脉冲电场强度进行取样测量，从而获取测量样品的 THz 时域信号波形，经快速傅里叶变换（FFT）得到频域信号。TAS7500TS HF1 的频率范围为 0.1～4 THz，光谱分辨率为 7.6 GHz，激光发射器平均功率 20 mW，脉冲中心波长为 1 550 nm，脉冲宽度为 50 fs，激光重复率为 50 MHz±200 Hz。试验在室温下进行，温度为 25 ℃，试验全过程使用空气压缩泵对测量的空间环境进行干燥，减少空气中水分对测量结果影响，提高信噪比。

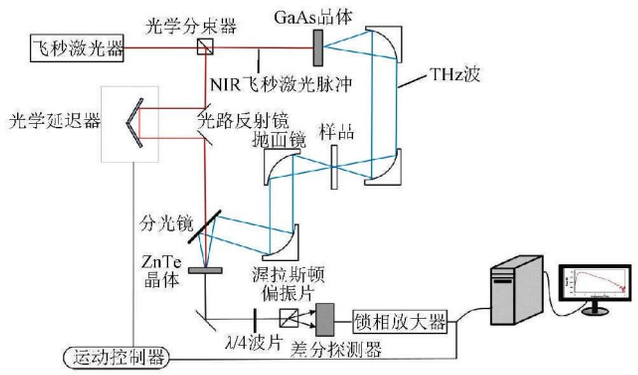

图 1　THz-TDS 系统透射模式原理图

Fig.1　Schematic of THz-TDS system in transmission-mode

1.2　试验材料

选取由星巴克合肥分公司提供的 3 个典型产区（埃塞俄比亚（Ethiopia）、哥斯达黎加（Costa Rica）以及印度尼西亚（Indonesia））的咖啡豆为试验样品。所有样品均在干燥器中存放（密封、避光的环境中）。试验编程软件采用 Matlab 2011a，试验前将埃塞俄比亚、印度尼西亚以及哥斯达黎加 3 个产地的咖啡豆样本各随机抽取 40 个共 120 个作为建模集，剩余各 20 个共 60 个作为预测集。使用粉碎机对不同产地的所有咖啡豆样品进行粉碎预处理，粉碎后的样品经孔径 0.074 mm 的筛子过滤，然后使用压片机将粉末样品进行压片处理，用 10 MPa 的压力压制成厚度约为 1 mm，直径为 13 mm、内部均匀、上下表面互相平行的薄片，每种咖啡豆各制成 60 个压片样品。

1.3　光谱获取与分析

试验前将 TAS7500TS HF1 预热半小时，以钢制背景板为系统标定板，调节螺旋测微器获取最佳焦点。将压制好的样品片放置于 TAS7500TS HF1 系统的聚乙烯样品台上，扫描得到样品的透射光谱图像。为减少测量误差，对同一样品均从不同位置测量 3 次，取平均值作为样品的光谱信号，3 种咖啡豆样品及钢制背景的时域光谱图如图 2 所示。

从图 2a 可以看出，3 种样品与背景板的太赫兹时域光谱信号在幅值与相位上均有明显差异，但样品之间的差异相对较小。在太赫兹透射的频域幅值上，哥斯达黎加咖啡豆的幅值整体上高于埃塞俄比亚和印度尼西亚的咖啡豆，而埃塞俄比亚与印度尼西亚咖啡豆在幅值上相当；在相位上，相对于参考背景哥斯达黎加咖啡豆约在 16.3 ps 产生波峰，滞后最小，印度尼西亚咖啡豆约在 16.9 ps 产生波峰，埃塞俄比亚咖啡豆约在 17.3 ps 产生波峰，滞后最大。通过对时域光谱信号进行快速傅里叶变换（FFT）得到样品的太赫兹透射频域光谱，如图 2b 所示。由图 2b 可知，太赫兹信号的有效光谱频域区域位于 0.2～1.5 THz 内，3 种咖啡豆的频谱曲线趋势一致，不同频率点下的透射能量有所差异，但在部分频率点上存在交叉，与很多复杂生物体一样，咖啡豆也没有明显特征吸收峰[20-21]。在 0.2～1.5 THz 范围内，共有 171 个频率点，高维的太赫兹光谱特征在带来丰富信息的同时，部分与样品品质相关性较弱甚至无关的信息会影响建模效果。因此，本文首先应用主成分分析方法，对不同产地咖啡豆的太赫兹光谱特征进行降维并对鉴别效果进行定性分析。

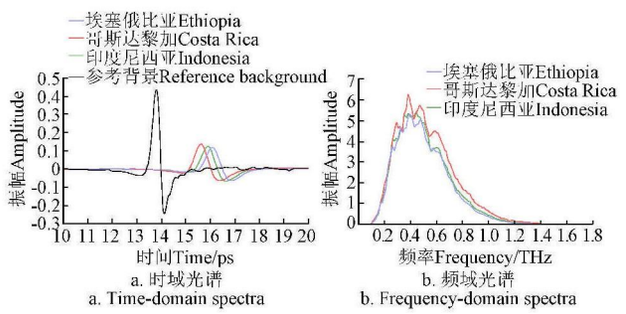

图 2　3 种不同产地咖啡豆时域光谱和频域光谱图

Fig.2　Time-domain spectra and frequency-domain spectra of coffee beans of 3 different producing areas

1.4　基于主成分的咖啡豆产地鉴别分析

主成分分析法是对多个变量间相关性进行分析的一种多元统计方法，通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量。通过主成分分析所得新变量在减少变量数目的同时，尽可能保持了原有的特征信息。本文运用主成分分析法对 3 个不同产区共 180 个咖啡豆在 0.2～1.5 THz 频域范围内的光谱数据进行处理，选取前 3 个主成分所得三维得分分布图如图 3 所示。

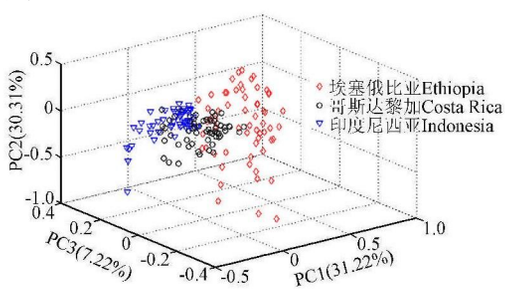

图 3　3 种不同产地咖啡豆样品光谱 3 维主成分分析图

Fig3　3D principal component analysis diagram of 3 kinds coffee bean samples from different producing areas

从图 3 可以看出，3 种咖啡豆具有较好的聚类效果，其中埃塞俄比亚与印度尼西亚咖啡豆之间没有相互交错，而哥斯达黎加咖啡豆与前两者均有交错，这与咖啡豆在太赫兹波段下没有特征吸收峰，光谱特征分布较广有关。从图 3 还可以看出前 3 个主成分的累积贡献率为 68.75%（PC1、PC2、PC3 的贡献率分别为 31.22%、30.31%、7.22%），不能完全包含太赫兹有效波段下的信息。为得到基于太赫兹光谱的咖啡豆产地识别最优主成分数，本文参考文献[22]，应用偏最小二乘判别（PLSDA）方法，选取前 3、4、…、50 个主成分(前 50 个主成分的累计贡献率可达 98.96%)分别进行咖啡豆产地鉴别。结果表明，在选取前 3 至 20 个主成分时，鉴别正确率处于上升趋势，大于 20 后鉴别效果开始下降。因此，本文选取前 20 个主成分（累积贡献率为 96.36%）作为建模的特征输入量。

2 基于粒子群参数寻优的支持向量机建模方法

2.1 支持向量机

支持向量机是一种基于有限样本统计学习理论的有监督机器学习方法，通过非线性映射将输入变量映射到一个高维的特征向量空间，并在高维空间构造最优分类超平面，较好解决了小样本、非线性、高维数、局部极小点等问题[23-25]。SVM 回归用一个非线性映射函数将数据映射到高维特征空间，在高维特征空间进行线性回归，依据结构风险最小化（structural risk minimization, SRL）原则，将其学习过程转化为凸优化问题，即

$$\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}(\beta_i + \beta_i^*) \quad \text{s.t.} \quad \begin{cases} f(x_i) - y_i \leqslant \varepsilon + \beta_i^* \\ y_i - f(x_i) \leqslant \varepsilon + \beta_i \\ \beta_i, \beta_i^* \geqslant 0; i = 1, \ldots, m \end{cases} \tag{1}$$

式中 ω 为回归函数参数；$f(x_i)$为变量 x_i 的模型回归值；y_i 为标定值；C 为边界参数；m 为样本数；β_i^*，β_i 为空间不同点的松弛变量，ε用于定义线性不敏感损失函数，其回归方程的最终表述为

$$f(x) = \sum_{i=1}^{nv}(a_i^* - a_i)K(x, x_i) + b \tag{2}$$

式中 α_i^*, α_i 为二次规划中的 Lagrange 乘子，nv 为支持向量机个数，$K(x,x_i)$为核函数，x, x_i 分别表示不同位置空间坐标点，表示采用不同的核函数可构成不同的 SVM 分类器。目前最常用的核函数分类器有线性核函数、径向基函数（radial basis function，RBF）、多项式核函数等。本文选择 RBF 作为 SVM 的函数。RBF 的定义如下

$$K(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{2\delta^2}) \tag{3}$$

式中为 δ 核宽度参数，δ>0。

对于 RBF 核函数的 SVM，有 2 个参数需要优化，即边界参数 C 和核参数 δ，这 2 个参数对 SVM 的分类性能具有相当大的影响[25]。其中边界参数 C 是 SVM 模型对结构风险和样本无误差的折中，与可容忍的误差相关；核参数 δ 反映了数据样本在高维特征空间中分布的复杂程度，决定了线性分类面的复杂度。目前在采用交叉验证（cross validation，CV）的方法下，用网格划分能够找到 CV 意义下的最高预测准确率，即全局最优解，但过程比较耗时。粒子群优化算法基于群体智能优化理论，通过群体中粒子间的合作与竞争产生的群体智能指导优化搜索。在本文中为了能够在更大范围内寻找最佳的参数 C 和 δ，提高搜索效率，采用了基于粒子群寻优的支持向量机建模方法。

2.2 粒子群算法

粒子群优化算法[26]是一种具有很强全局寻优能力的群智能优化算法，在一个 N 维的目标搜索空间，由 k 个粒子组成一个种群 $Z=\{Z_1, Z_2, \cdots, Z_k\}$，其中每个粒子所处的位置 $Z_i=\{Z_{i1}, Z_{i2}, \cdots, Z_{in}\}$都表示问题的潜在的一个解，并依据目标函数计算每个粒子的适应度。然后每个粒子都在解空间中迭代搜索，不断调整自己的位置搜索新解[27]。在每次寻优迭代过程中，粒子根据式（4）和（5）进行位置 Z_i 和速度 $V_i=\{V_{i1}, V_{i2}, \cdots, V_{in}\}$的更新。

$$v_{id}(t+1) = \lambda v_{id}(t) + \eta_1 r_1(P_{id} - z_{id}(t)) + \eta_2 r_2(P_{gd} - z_d(t)) \tag{4}$$

$$Z_{id}(t+1) = z_{id}(t) + v_{id}(t+1) \tag{5}$$

式中 λ 为惯性权重因子，η_1 和 η_2 均为取值为正数的学习因子，γ_1、γ_2 则为 0 到 1 之间的随机数，P_{id} 为粒子个体极值，P_{gd} 为粒子群的全局极值。此外，为使粒子的速度不至于过大或过小，对速度的最大值与最小值进行限制，参考文献[28-29]，设置粒子速度的最大值 v_{max} 和最小值 v_{min} 分别为 1 和-1。

2.3 基于 PSO 参数优化的支持向量机分类模型

构建基于 PSO 参数优化的支持向量机分类模型的具体步骤如下。

1）采用对 3 类咖啡豆太赫兹光谱进行主成分分析所得前 20 个主成分变量作为咖啡豆产地鉴别的特征向量，设置 SVM 模型参数的搜索范围和初始化粒子群的相关参数，如种群规模、学习因子、惯性权重、最大迭代次数等；

2）初始化粒子群。随机产生边界参数 C 和核宽度参数 δ 的值作为每个粒子的初始位置，同时随机初始化每个粒子的初始速度。

3）计算每个粒子的当前适应度。定义适应度函数如式（5），通过对训练样本的学习训练，得到各个粒子的正确分类数，用以计算各个粒子的适应度函数值。

$$F(Z_i) = \frac{T_{acc}}{T} \times 100 \tag{6}$$

式中 T_{acc} 和 T 分别表示正确分类的样本个数和样本总数。

4）计算每个粒子的当前适应值 $F(Z_i)$，并与该粒子当前自身的最优适应值 $F(P_{id})$进行比较，如果 $F(Z_i)<F(P_{id})$，则调整 $F(P_{id})=F(Z_i)$，将当前位置作为此刻该粒子的最优位置。

5）将每一个粒子自身当前最优位置的适应值 $F(P_{id})$ 与所有粒子当前最优位置的适应值 $F(P_{gd})$进行比较，若 $F(P_{id})<F(P_{gd})$，则调整 $F(P_{gd})=F(P_{id})$，将调整后的位置作为所有粒子的最优位置。

6）利用 PSO 的进化方程（4）、（5）调整粒子的速度和位置，进而得到支持向量机的参数。

7）判断是否满足给定的最大迭代次数，如果满足则停止寻优，并返回当前最优的 SVM 模型参数 C 和 δ；否则转到步骤 3）。

8）将最优参数代入 SVM 模型，对测试样本集进行有效的分类。

3　试验结果与分析

输入特征向量选用 0.2～1.5 THz 太赫兹频域光谱的前 20 个主成分，模型参数选择采用粒子群算法进行优化。试验过程中先对粒子群进行参数初始化，参考文献[28-30]中的研究结果，PSO 算法中的种群粒子设为 50，学习因子 $\alpha_1=\alpha_2=1.5$；设定变权重λ取为起始值$\lambda_{strat}=0.9$，终止值$\lambda_{end}=0.4$；C，δ 的搜索范围为 $C\in[2^{-2}, 2^2]$，$\delta\in[2^{-2}, 2^2]$，步长为 $2^{0.5}$；终止迭代次数为 100。最终通过试验，经过粒子群寻优算法得到支持向量机的最优参数结果为 C=1.393 66，δ=0.01。

为验证 PSO-SVM 分类方法的优越性，将 PSO-SVM 方法与最小二乘支持向量机（least-square-support vector machine, LS-SVM）[31]、反向神经网络算法（back propagation neural network, BPNN）[32]进行比较，结果如表 1 所示。从表中可以看出，3 种算法对不同产地咖啡豆的鉴别效果都在 80%以上，说明不同产地的咖啡豆在太赫兹波段下存在明显差异，太赫兹光谱技术可用于咖啡豆产地的鉴别；同时 3 种模型中，支持向量机的鉴别效果明显优于 BPNN，而经过 PSO 参数优化的 SVM 分类效果优于 LS-SVM。其中通过 PSO-SVM 所得的最优模型预测结果在建模集中的正确率可达 100%，在预测集中的正确率可达 95%。对 BPNN 学习算法来说，造成鉴别效果不佳的原因可能是 BPNN 学习算法对训练样本数量要求较高，高维输入特征会对神经网络的训练结果精度带来影响。粒子群算法对支持向量机参数的优化是连续的，而支持向量机本身具有小样本学习和解决高维特征的能力，所以最后能得到使分类精度更好的优化参数，获取最优的鉴别模型。

表 1　3 种建模方法的分类结果比较

Table 1　Comparison of classification results of 3 modeling methods

建模方法 Methods		埃塞俄比亚 Ethiopia		印度尼西亚 Indonesia		哥斯达黎加 Costa Rica		总正确率 Total accuracy rate/%
		误识别数 Number of misclassified samples	正确率 Accuracy rate/%	误识别数 Number of misclassified samples	正确率 Accuracy rate/%	误识别数 Number of misclassified samples	正确率 Accuracy rate/%	
BPNN	建模集 Calibration set	4	90.0	8	80.0	2	95.0	88.3
	预测集 Prediction set	3	85.0	6	70.0	2	90.0	81.7
LS-SVM	建模集 Calibration set	1	97.5	5	87.5	0	100.0	95.0
	预测集 Prediction set	1	95.0	4	80.0	0	100.0	91.7
PSO-SVM	建模集 Calibration set	0	100.0	0	100.0	0	100.0	100.0
	预测集 Prediction set	0	100.0	3	85.0	0	100.0	95.0

4　结　论

本文以太赫兹时域光谱为检测手段，研究了不同产地咖啡豆的快速鉴别问题。试验样本选取埃塞俄比亚、印度尼西亚以及哥斯达黎加 3 个典型产地的咖啡豆。采用透射式太赫兹光谱系统获取咖啡豆压片样品的太赫兹时域光谱和频域光谱信息，并结合主成分分析进行光谱特征的降维和提取，利用粒子群算法对支持向量机进行参数寻优，建立了基于太赫兹光谱特征的咖啡豆产地鉴别模型。试验结果中本论文所提方法对不同产地咖啡豆的鉴别准确率在建模集和预测集中分别高达 100%和 95%，优于 BPNN 和 LS-SVM 算法。本文的研究表明太赫兹光谱技术可用于不同产地咖啡豆的快速鉴别，采用 PSO 优化的 SVM 方法结合太赫兹光谱技术能够获得理想的鉴别模型。本文为咖啡豆产地鉴别提供了一种新方法，也为太赫兹光谱技术在其他复杂农产品/食品中的检测应用提供了思路。

[参　考　文　献]

[1] 胡双芳，卫亚西，邢精精，等. 咖啡豆的化学组分差异与感官品质的相关性分析[J]. 食品工业科技，2013，34(24)：125－129.
Hu Shuangfang, Wei Xiya, Xin Jingjing, et al. Correlation analysis between chemical components and sensory quality of coffee[J]. Science and Technology of Food industry, 2013, 34(24): 125－129. (in Chinese with English abstract)

[2] 顾文佳，李兆阶. 我国焙炒咖啡行业质量调研报告[J]. 质量与标准化，2014(12)：35－37.
Gu Wenjia, Li Zhaojie. A Survey report on the quality of roasted coffee in China[J]. Quality and Standardization, 2014(12): 35－37. (in Chinese with English abstract)

[3] Semmelroch P, Laskawy G, Blank I, et al. Determination of potent odorants in roasted coffee by stable isotope dilution assay[J]. Flavour & Fragrance Journal, 1995, 10(1): 1－7.

[4] Piccino S, Boulanger R, Descroix F, et al. Aromatic composition and potent odorants of the “specialty coffee” brew “Bourbon Pointu” correlated to its three trade

classifications[J]. Food Research International, 2014, 61(61): 264－271.

[5] 何余勤，胡荣锁，张海德，等. 基于电子鼻技术检测不同焙烤程度咖啡的特征性香气[J]. 农业工程学报，2015，31(18)：247－255.
He Yuqin, Hu Rongsuo, Zhang Haide, et al. Characteristic aroma detection of coffee at different roasting degree based on electronic nose[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(18): 247－255. (in Chinese with English abstract)

[6] Cho J S, Bae H J, Cho B K, et al. Qualitative properties of roasting defect beans and development of its classification methods by hyperspectral imaging technology[J]. Food Chemistry, 2017, 220: 505－509.

[7] Chen T, Li Z, Yin X, et al. Discrimination of genetically modified sugar beets based on terahertz spectroscopy y [J]. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 2016, 153: 586-590.

[8] Lian F, Ge H, Xia S, et al. Identification of wheat quality using THz spectrum[J]. Optics Express, 2014, 22(10): 12533－12544.

[9] Gente R, Busch S F, Stübling E M, et al. Quality control of sugar beet seeds with THz time-domain spectroscopy[J]. IEEE Transactions on Terahertz Science & Technology, 2016, 6(5): 754－756.

[10] 杨静琦，李绍限，赵红卫，等. L-天冬酰胺及其一水合物的太赫兹光谱研究[J]. 物理学报，2014，63(13)：105－111.
Yang Jingqi, Li Shaoxian, Zhao Hongwei, et al. Terahertz study of L-asparagine and its monohydrate[J]. Acta Physica Sinica, 2014, 63(13): 105－111. (in Chinese with English abstract)

[11] Liu J, Li Z. The terahertz spectrum detection of transgenic food[J]. Optik - International Journal for Light and Electron Optics, 2014, 125(23): 6867－6869.

[12] Gowen A A, O'Sullivan C, O'Donnell C P. Terahertz time domain spectroscopy and imaging: Emerging techniques for food process monitoring and quality control[J]. Trends in Food Science & Technology, 2012, 25(1): 40－46.

[13] Liu W, Liu C, Chen F, et al. Discrimination of transgenic soybean seeds by terahertz spectroscopy[J]. Scientific Reports, 2016, doi: 10.1038/srep35799.

[14] Liu W, Liu C, Hu X, et al. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics[J]. Food Chemistry, 2016, 210: 415－421.

[15] Qin J Y, Ying Y B, Xie L J. The detection of agricultural products and food using terahertz spectroscopy: A Review[J]. Applied Spectroscopy Reviews, 2013, 48(6): 439－457.

[16] Redo-Sanchez A, Laman N, Schulkin B, et al. Review of terahertz technology readiness assessment and applications[J]. Journal of Infrared, Millimeter, and Terahertz Waves, 2013, 34(9): 500－518.

[17] 谢丽娟，徐文道，应义斌，等. 太赫兹波谱无损检测技术研究进展[J]. 农业机械学报，2013，44(7)：246－255.
Xie Lijuan, Xu Wendao, Ying Yibin, et al. Advancement and trend of terahertz spectroscopy technique for non-destructive detection[J]. Transactions of The Chinese Society for Agricultural Machinery, 2013, 44(7): 246－255. (in Chinese with English abstract)

[18] Su T F, Zhao G Z, Ren T B, et al. Characterizations of physico-chemical changes of corn biomass by steam explosion[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(6): 253－256.

[19] 沈晓晨，李斌，李霞，等. 基于太赫兹时域光谱的转基因与非转基因棉花种子鉴别[J]. 农业工程学报，2017，33(增刊 1)：288－292.
Shen Xiaochen, Li Bin, Li Xia, et al. Identification of transgenic and non-transgenic cotton seed based on terahertz range spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(Supp.1): 288－292. (in Chinese with English abstract)

[20] Ge H, Jiang Y, Lian F, et al. Characterization of wheat varieties using terahertz time-domain spectroscopy[J]. Sensors, 2014, 15(6): 12560－12572.

[21] Liu J, Li Z, Hu F, et al. A THz spectroscopy nondestructive identification method for transgenic cotton seed based on GA-SVM[J]. Optical and Quantum Electronics, 2015, 47(2): 313－322.

[22] 郝勇，孙旭东，高荣杰，等. 基于可见/近红外光谱与 SIMCA 和 PLS-DA 的脐橙品种识别[J]. 农业工程学报，2010，26(12)：373－377.
Hao Yong, Sun Xudong, Gao Rongjie, et al. Application of visible and near infrared spectroscopy to identification of navel orange varieties using SIMCA and PLS-DA methods[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2010, 26(12): 373－377. (in Chinese with English abstract)

[23] Vapnik V N. An overview of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1999, 10(10): 988－999.

[24] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121－167.

[25] V David S A. Advanced support vector machines and kernel methods[J]. Neurocomputing, 2003, 55(1/2): 5－20.

[26] 焦有权，赵礼曦，邓欧，等. 基于支持向量机优化粒子群算法的活立木材积测算[J]. 农业工程学报，2013，29(20)：160－167.
Jiao Youquan, Zhao Lixi, Deng Ou, et al. Calculation of live tree timber volume based on partical swarm optimization and support vector regression[J]. Transaction of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2013, 29(20): 160－167. (in Chinese with English abstract)

[27] Venter G, Sobieszczanskisobieski J. Particle Swarm Optimization [J]. Aiaa Journal, 2013, 41(8):129-132.

[28] 刘伟，王建平，刘长虹，等. 基于粒子群寻优的支持向量机番茄红素含量预测[J]. 农业机械学报，2012，43(4)：143－147.
Liu Wei, Wang Jianping, Liu Changhong, et al. Lycopene content prediction based on support vector machine with particle swarm optimization[J]. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(4): 143－147. (in Chinese with English abstract)

[29] 刘晓峰，陈通. PSO 算法的收敛性及参数选择研究[J]. 计算机工程与应用，2007，43(9)：14－17.
Liu Xiaofeng, Chen Tong. Study on convergence analysis

and parameter choice of Particle Swarm Optimization[J]. Computer Engineering and Applications, 2007, 43(9): 14－17. (in Chinese with English abstract)

[30] Shi Y, Eberhart R C. Parameter Selection in Particle Swarm Optimization[C]. // Proceeding EP '98 Proceedings of the 7th International Conference on Evolutionary Programming VII. 1998: 591－600.

[31] Borin A, Ferrão M F, Mello C, et al. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk[J]. Analytica Chimica Acta, 2006, 579(1): 25－32.

[32] Dai H, MacBeth C. Effects of learning parameters on learning procedure and performance of a BPNN[J]. Neural Networks, 1997, 10(8): 1505－1521.

Rapid identification of producing area of coffee bean based on terahertz spectroscopy and support vector machine

Hu Xiaohua[1], Liu Wei[2,3※], Liu Changhong[3], Qian Yunhui[3]

(1. *School of Computer and Information, Hefei University of Technology, Hefei* 230009, *China*;
2. *Intelligent control and Compute vision lab, Hefei University, Hefei* 230009, *China*;
3. *School of Food Science and Engineering, Hefei University of Technology, Hefei* 230009, *China*)

Abstract: Coffee is a very popular beverage in many countries. Coffee bean from different producing area has different flavour and functional properties, and thus the identification of producing area of coffee bean is important to assure the quality of coffee bean. The feasibility of a rapid and precise determination method of producing area of coffee bean was examined by using the terahertz (THz) time-domain spectra system (TAS7500TS HF1, Advantest Co., Ltd, Japan). Coffee bean samples from 3 different typical producing areas (Ethiopia, Costa Rica, and Indonesia) were collected and pressed into pellets for THz measurements. A total of 180 pellet samples (3 classes, each had 60 pellet samples) were randomly divided into calibration set (40 pellet samples for each class) and prediction set (20 pellet samples for each class). THz time-domain spectroscopy system worked with the TAS7500TS equipment in transmission mode. Before the experiment, the dry air was injected until the relative humidity reached below 3% to reduce the absorption of the THz waves by water in air. The parameters of THz system were as follow: frequency range was from 0.1 to 4 THz, the resolution was 7.6 GHz, the short pulse width was less than 50 fs and the average power was 20 mW. For each sample, the THz time-domain spectra were measured for 3 times at different position and then the average values were obtained. The frequency-domain spectra were acquired by a fast Fourier transform (FFT). Principal component analysis (PCA) with frequency-domain spectral data was performed to examine the qualitative difference of these 3 classes of coffee beans using the first 3 score vectors. The 3 groups of different class of coffee beans were almost apart from each other in the space of the first 3 principal components (PCs), although there was some overlap among the groups, which may be due to that the first 3 PCs only accounted for the all spectral variations of 68.75%. Thus, to reduce the dimension of the model features and retain more information of the THz spectra of samples, the first 20 components were selected as the spectral characteristics for the determination of producing area of coffee bean. The support vector machine (SVM), as a learning algorithm used for classification and regression tasks, was used to get the identification model. During the iteration for the optimum parameters selection, the particle swarm optimization (PSO) was designed, which could enlarge search space and improve search efficiency. The identification results of the PSO-SVM were compared with the least squares - support vector machine (LS-SVM) and back propagation neural network (BPNN). From the comparison, it was showed that the discrimination accuracy of all 3 classes of coffee beans using the PSO-SVM was up to 95% in prediction set and 100% in calibration set, respectively, which was the best model among the 3 methods. It can be concluded that the THz frequency spectra can be used as important features to identify the producing area of the coffee bean. The model with SVM method based on PSO can get better parameters of SVM to improve the identification ability than the traditional LS-SVM. THz spectra system combined with the proposed algorithm has been proved to be a very powerful and attractive tool for identification of producing area of coffee bean.

Keywords: spectroscopy; models; support vector machine; coffee bean; terahertz; particle swarm optimization