激光与光电子学进展
先进成像

基于太赫兹检测技术的咖啡豆品种鉴别

刘燕德*，李茂鹏，胡军，徐振，崔惠桢
华东交通大学机电工程学院，江西 南昌 330013

摘要 咖啡是世界三大饮料之一，严格把控咖啡品质具有重大意义。以三个不同品种的咖啡豆为对象，利用太赫兹时域光谱技术结合化学计量学实现咖啡豆品种的快速鉴别。采用多种预处理方法减小实验误差，利用主成分分析(PCA)对光谱矩阵进行降维。建立偏最小二乘判别分析(PLS-DA)二分类模型和支持向量机(SVM)、反向传播神经网络(BPNN)、随机森林(RF)多分类判别模型。PLS-DA 二分类模型的定性判别效果较为理想，总正确率可达 98%；在多分类模型中，基于基线校正建立的 SVM 模型的效果最佳，总正确率达到 98%。本研究表明利用太赫兹光谱技术快速鉴别咖啡豆品种是可行的，建立了较优的基于基线校正后的支持向量机模型，以为太赫兹时域光谱技术在定性检测其他农产品时提供经验参考。

关键词 光谱学；太赫兹时域光谱技术；咖啡豆；定性判别；基线校正；支持向量机
中图分类号 O433.4；O439 **文献标志码** A **doi**：10.3788/LOP202158.1630002

Identification of Coffee-Bean Varieties Using Terahertz Detection Technology

Liu Yande*, Li Maopeng, Hu Jun, Xu Zhen, Cui Huizhen
School of Mechatronics & Vehicle Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, *China*

Abstract Coffee is one of the top three beverages in the world, and it is crucial to strictly control the quality of coffee. Using three different varieties of coffee beans as the object, the terahertz time-domain spectroscopy combined with chemometrics is used herein to identify three different coffee beans quickly. Various pretreatment methods are used to reduce the experimental errors, and principal component analysis is used to reduce the dimensions of the spectral matrix. The dichotomous model based on partial least squares discriminant analysis (PLS-DA) is established, and another model based on a support vector machine (SVM), a back propagation neural network, and a random forest multiclassification discriminant is then established. The PLS-DA dichotomous model exhibits the ideal qualitative discrimination effect whose accuracy is 98%. Among various classification models, the SVM model based on baseline correction is the most effective model, with the total accuracy reaching 98%. This study shows that it is feasible to use terahertz spectroscopy to quickly identify coffee bean varieties and a better support vector machine model is established based on baseline correction, which provides an empirical reference for the qualitative detection of other agricultural products using the terahertz time-domain spectroscopy.

Key words spectroscopy; terahertz time domain spectroscopy; coffee beans; qualitative discrimination; baseline correction; support vector machine
OCIS codes 300.6170; 300.6495

收稿日期：2020-10-18；**修回日期**：2020-12-16；**录用日期**：2020-12-27
基金项目："十二五"国家 863 计划(SS2012AA101306)、江西省优势科技创新团队建设计划(20153BCB24002)、南方山地果园智能化管理技术与装备协同创新中心项目(赣教高字[2014]60 号)、国家自然科学基金(2002017018)、江西省教育厅科学技术研究青年项目(GJJ190348)、江西省博士研究生创新资金项目(YC2019-B106)
通信作者：*jxliuyd@163.com

1 引 言

咖啡、茶、可可并称当今世界的三大无酒精饮料，其中咖啡是浪漫浓郁的代表，原材料为咖啡豆，不同品种的咖啡豆在口感、气味及化学组成成分上也有着很大的区别[1]。中国是世界上咖啡消费量增长最快的国家之一，咖啡产业呈现良好的发展势头与前景。咖啡品质的优劣主要取决于咖啡豆的品种，目前，针对咖啡豆的品种鉴别，国内外常用的检测方法主要有人工感官经验法、化学分析方法等。人工感官经验法依靠丰富的实践经验；化学分析方法主要包含有高效液相色谱法（HPLC）、液相色谱-质谱/质谱法（LC-MS）、气液色谱法和气相色谱-质谱联用法等检测方法。曾凡逵等[2]以云南咖啡、兴隆咖啡和越南咖啡为实验对象，使用气相色谱-质谱联用法研究不同生咖啡及不同烘焙程度下的脂肪酸含量，通过总脂肪含量和脂肪酸含量鉴别不同种咖啡豆。Moreira 等[3]利用高效液相色谱法结合近红外光谱法对传统咖啡豆和转基因咖啡豆进行了区分。这些方法为有损检测，样品预处理过程繁琐，操作程序复杂，检测成本高，耗费时间长，无法满足咖啡豆质量检测对速度和无损的要求[4]。因此，研究出一种简单、快速、无损的咖啡豆品种检测手段是十分有必要的。

近年来，太赫兹检测技术也逐渐应用在农产品/食品安全检测领域。太赫兹（THz）波段介于微波和红外区域之间，其频率范围为 0.1～10 THz（3.3～333.6 mm），该波段的太赫兹光谱具有很多独特的性质[5-6]，很多大分子（DNA、蛋白质、各种糖类、病毒等）的转动和振动能级都在此区间，所以太赫兹光谱具有丰富的信息，相比于其他光谱具有光源稳定、辐射能量小等优势。太赫兹时域光谱技术已经在农产品以及食品检测领域取得一定的进展，但对于咖啡品质的鉴别主要还是通过化学方法和经验法来判断，利用太赫兹时域光谱技术鉴别咖啡豆品种尚处于探索阶段[7]。

本文以三种不同品种的咖啡豆为研究对象，探索利用太赫兹时域光谱技术结合化学计量学鉴别咖啡豆品种的方法[8]。通过建立二分类、多分类模型定性鉴别三种咖啡豆的品种，这为咖啡豆品种的鉴别提供一种快速无损的方法，同时为进一步探究太赫兹时域光谱检测技术在农产品及食品领域的应用提供参考。

2 材料与方法

2.1 实验材料

实验室所用的咖啡豆购买于南昌市万达星巴克，选取了三大热销咖啡豆品种，主要包括肯尼亚咖啡豆、卢旺达咖啡豆、中国云南咖啡豆，三种咖啡豆采用的烘焙方式均为中度烘焙。将三种咖啡豆均放置在密封、避光、低湿度的干燥器中。由于咖啡豆颗粒较大，很难直接完成透射实验，故采用压片处理。压片之前，先用粉碎机将咖啡豆打成粉末，利用玛瑙研钵将经粉碎机后的咖啡豆粉末颗粒研磨为微细的颗粒，并过 200 目筛子滤成精细的咖啡豆粉末，以消除颗粒对太赫兹光谱的散射影响。最后称取一定质量的粉末倒入压片模具中，每片薄片的质量控制在 0.13 g 左右，设置压力为 10 MPa，压片时间为 1 min，压制成厚度为 0.85 mm 左右、直径为 13 mm，上下表面均匀、平行的薄片。每种咖啡豆制成 50 个薄片，共计 150 个样品。

2.2 太赫兹光谱成像装置

本实验装置采用的是日本 Advantest 公司的 TAS7500SU 太赫兹时域光谱系统，系统采集的频率范围为 0.1～8 THz，光谱分辨率为 7.6 GHz，飞秒激光脉冲的输出功率为 20 mW，中心波长为 1550 nm，宽度为 50 fs，重复频率为 50 MHz。图 1 为太赫兹时域光谱系统原理图，飞秒激光器发射出的发射光经分束器分成两束光，其中较强的光为泵浦光，较弱的光为探测光。泵浦光入射到砷化镓（GaAs）光电导天线上，进而激发出 THz 脉冲，该 THz 脉冲经过透镜和抛面镜入射到样品表面，THz 脉冲经过样品产生色散效应，从而引起幅值和相位的变化，THz 脉冲携带样品的信息与探测光一起聚焦到 ZnTe 晶体上。获得的光谱信号经过锁相放大器后被导入计算机中，获得物体的时域信号，再通过快速傅里叶变换（FFT）获得样品太赫兹频率信号。

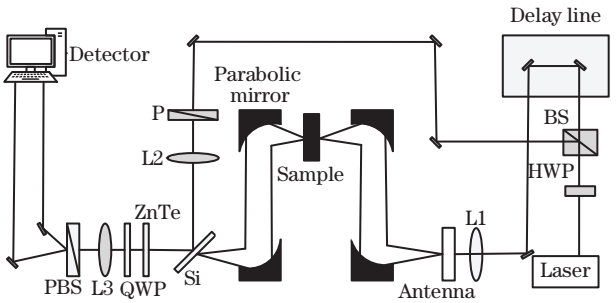

图 1 太赫兹时域光谱系统原理图

Fig. 1 Terahertz time domain spectral system schematic

2.3 样品太赫兹光谱参数提取

实验开始前，将所有的样品均放在干燥器内干燥 3 h，采用空气压缩机对太赫兹系统进行干燥，使用抽湿机降低室内的空气湿度，室温保持在

(25±0.3) ℃,空气湿度保持在 10%以下。采集光谱前,先将太赫兹设备预热 0.5 h,以便得到稳定的光谱信号,为了减小误差,分别对每个样本采集 4 个不同的点,取 4 次光谱信号的平均值作为样品的时域信号。根据 Dorney 等[9]和 Dragoman 等[10]提出的光学参数提取方法,主要研究参数包括折射率、吸收系数等,利用快速傅里叶变换(FFT)算法可以得到太赫兹脉冲光谱:

$$E(\omega)=A(\omega)\exp[-\mathrm{i}\varphi(\omega)]=\int E(t)\exp(-\mathrm{i}\omega t)\,\mathrm{d}t, \tag{1}$$

$$n(\omega)=\frac{\varphi(\omega)c}{\omega d}+1, \tag{2}$$

$$\alpha(\omega)=\frac{2k(\omega)\omega}{c}=\frac{2}{d}=\ln\frac{4n(\omega)}{A(\varphi)(n(\omega)+1)^2}, \tag{3}$$

式中:$A(\omega)$为电场幅值,ω 为角频率,t 为时间,$E(t)$是太赫兹时域波形,$\varphi(\omega)$ 是相位,$E(\omega)$是太赫兹脉冲的频谱,c 为光速,d 为样品的厚度,$n(\omega)$为样品的折射率,k 为消光系数,$\alpha(\omega)$是样品的吸收系数。

2.4　实验样品模型预处理及建模算法的基本原理

在进行数据处理前,采用了 4 种预处理方式对模型进行预处理,主要有 Savitzky-Golay 卷积平滑法、多元散射校正(MSC)、标准正态变量变换(SNV)、基线校正(Baseline)。

支持向量机(SVM)的思想来自线性判别的最优分类面,最优分类面是要求分类面不但能够将两种样本无差别地分开,而且要求分类间隙达到最大,以使非线性分类转换到线性分类,从而起到提高模型预测能力、降低分类错误率的作用[11]。常用的核函数主要有线性核和径向偏差函数(RBF),本实验用的是 RBF 测试向量的分类:

$$\mathrm{RBF}(a,b)=\exp(-|a-b|/\sigma^2), \tag{4}$$

式中:a 表示样本点;b 代表核函数中心点;σ_2 表示 RBF 核函数的方差。反向传播神经网络(BPNN)是目前应用最多的神经网络,是一种由非线性变换神经单元组成的前馈型多层神经网络,具有非常优秀的预测能力,常用于大样本的模型建立[12]。BP 神经网络通常由三层组成:输入层、隐含层和输出层。每个神经元的节点层与下一层神经元节点都相互连接,形成稳定的网络结构。BPNN 采用的是最小二乘函数作为目标函数进行训练:

$$E=\sum_{r=1}^{m}\sum_{k=1}^{n}(y_{rk}-o_{rk})^2, \tag{5}$$

式中:m 为训练样本数,n 为输出层节点个数,o_{rk} 为节点 k 处的输出值,y_{rk} 为与其对应的期望输出值。

随机森林(RF)是一种基于统计学的决策树分类集成算法,其基本思想是选择训练样本以最小化残差平方和,直至形成一个完整的决策树,从而形成多个决策树,并通过投票的方式获得最终的预测结果[13-14]。

2.5　模型评价标准

本文首先对采集的样品太赫兹光谱数据进行预处理,选择 0.5 THz 波段的光谱数据,随后通过主成分分析(PCA)将数据降维从而形成新的光谱矩阵,再运用样本选择方法 Kennard-Stone(KS)在特征空间中均匀选取样本,样本分为训练集和测试集,比例为 3∶1。针对每类咖啡豆分别选取 150 个训练集和 50 个测试集。随后对处理过的光谱矩阵进行定性建模,首先建立 PLS-DA 对咖啡豆进行二分类,随后采用支持向量机(SVM)、反向神经网络(BPNN)和随机森林(RF)对咖啡豆品种进行多分类,评价模型的参数主要是对每类咖啡豆鉴别的正确率以及总正确率,正确率越接近于 100%,模型效果越好。图 2 为三种咖啡豆品种的定性鉴别流程图。

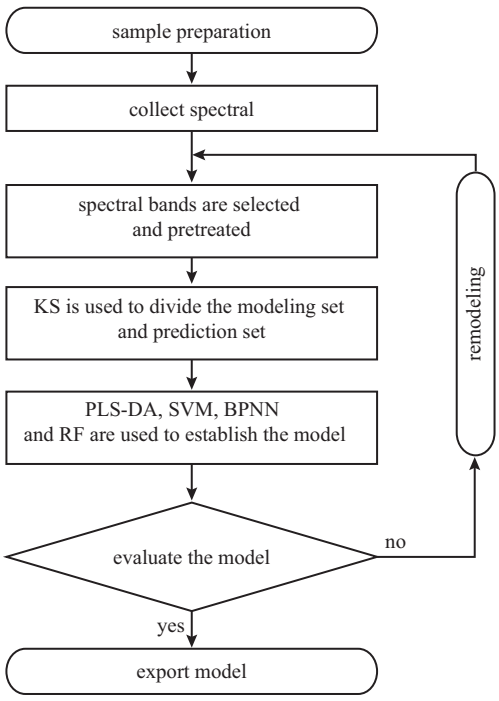

图 2　三种咖啡豆品种定性鉴别模型的流程图

Fig. 2　Flow chart of qualitative identification model of three kinds of coffee beans

3　实验结果与分析

3.1　卢旺达、云南、肯尼亚咖啡豆的太赫兹光谱特性分析

本实验共选用三种咖啡豆品种，每种咖啡豆压片 50 个，每个薄片的采集光谱数为 4 个，共采集 600 条光谱，去除首尾噪声波段，选取 0.5～1.8 THz 波段的光谱信息。利用 KS 算法将样本以 3∶1的比例分成训练集和测试集，以不同品种咖啡豆的样本代号，训练集和测试集的样本个数如表 1 所示。图 3 为三种品种咖啡豆的平均吸收光谱图，从图中可以看出，三条吸收曲线没有明显的差异，因此很难从吸收特性图判断咖啡豆的种类。

表 1　不同品种咖啡豆的训练集和测试集区分以及样本代号

Table 1　Classification of training sets and test sets and sample codes of different coffee beans

Type of coffee beans	Sample code	Total number of samples	Number of training set samples	Number of sample test sets
Kenyan coffee beans	1	200	150	50
Rwandan coffee beans	2	200	150	50
Yunnan coffee beans	3	200	150	50

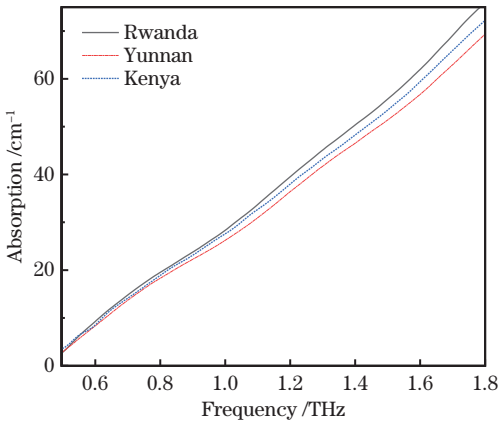

图 3　三种咖啡豆的平均吸收太赫兹光谱

Fig. 3　Mean absorption THz spectra of the three coffee beans

3.2　咖啡豆样品光谱变量 PCA 降维

主成分分析（PCA）是一种通过投影实现降维的分析方法，通过降维可以使少数几个新变量以线性组合的方式代替原始变量，这样就可以将一个大型数据空间转换为小维度的因子空间[15]。这个小型因子空间新的光谱数据调用成主成分（PC），投影的最大方差在第一个坐标（PC1）上，第二大方差在第二个坐标上（PC2），按照此规律，转换后的新变量相互正交、互不相关。选择合适的主成分个数所得到的结果既能代表原始光谱的重要信息又能使模型更加精简，如图 4 所示，当主成分个数为 7 时，结果能代表原始光谱变量 98%以上的信息，因此，最佳主成分个数为 7，所以本文选取前 7 个主成分因子构成新的变量矩阵，作为后续建模算法的输入。

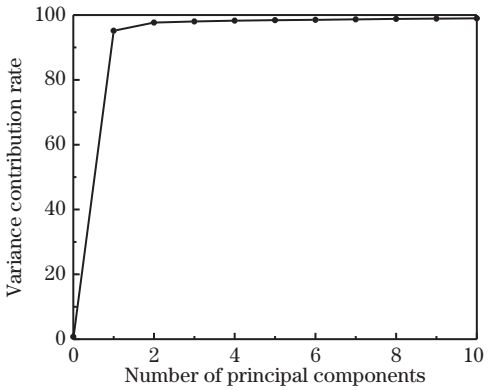

图 4　主成分个数与方差贡献率的关系图

Fig. 4　Relationship between number of principal components and variance contribution rate

3.3　咖啡豆样品二分类模型建立

通过建立 PLS-DA 模型对三种咖啡豆进行二分类定性判别，二分类指的是两两分类，评判模型好坏主要由正确率和总正确率决定，分类结果如表 2 所示。由表可看出三种咖啡豆的二分类正确率，每个品种咖啡豆的光谱数为 50 个，其中将卢旺达与云南咖啡豆作比较的时候，卢旺达咖啡豆的误判个数为 6，云南咖啡豆的误判个数为 2，总正确率为 92%；将肯尼亚和云南咖啡豆作比较时，肯尼亚咖啡豆无误判个数，云南咖啡豆的误判个数为 2，总正确率为 98%；将肯尼亚和卢旺达咖啡豆作比较时，肯尼亚咖啡豆无误判个数，卢旺达咖啡豆的误判个数为 2，总正确率为 98%。该模型的预测效果很好，可以较好地定性判别两种咖啡豆，图 5 为三种咖啡豆的真值与预测值相关图。

表 2　PLS-DA 模型分类结果

Table 2　Classification results of PLS-DA model　%

Object	Accuracy	Entirety
Rwanda	88	92
Yunnan	96	
Kenya	100	98
Yunnan	96	
Kenya	100	98
Rwanda	96	

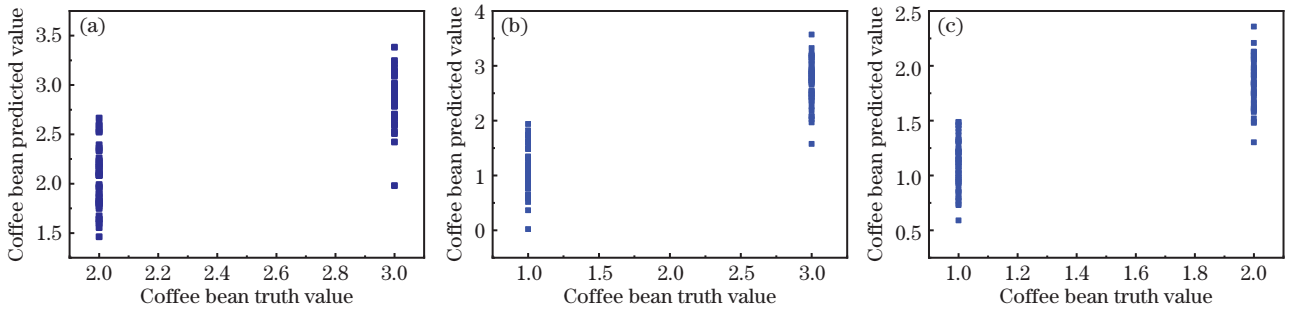

图 5 三种咖啡豆的真值与预测值相关图。(a)卢旺达-云南;(b)肯尼亚-云南;(c)肯尼亚-卢旺达

Fig. 5 Correlation between the truth value and the predicted value of the three coffee beans. (a) Rwanda-Yunnan; (b) Kenya-Yunnan; (c) Kenya-Rwanda

3.4 咖啡豆样品多分类模型建立与对比分析

为寻求最优的咖啡豆定性分类模型，采用 SVM、BPNN、RF 对预处理之后的光谱数据进行建模分析。3 种建模方法的分类比较如表 3 所示，从表中可以看出，在不经预处理的情况下，三种算法对不同品种的鉴别正确率都在 90%以上，说明利用太赫兹光谱检测技术可以很好地检测不同咖啡豆。经预处理之后，鉴别正确率均有不同程度的提升或者下降，所以对不同算法采用合适的预处理方法是至关重要的。从整体上来看，BPNN 模型分类效果最差，RF 次之，SVM 的效果最好，图 6 为 SVM 对咖啡豆样品预测集的分类结果，从图中可以看出，第一类咖啡豆有一个被误判为卢旺达咖啡豆，两个被误判为云南咖啡豆。SVM 采用网格搜索法确定最佳参数($c=2.8284$，$g=0.03125$)，每一类咖啡豆的分类正确率均高达 90%以上，总正确率都在 94%以上。综合三种多分类算法，SVM 结合 Baseline 建模效果最优，总正确率高达 98%。

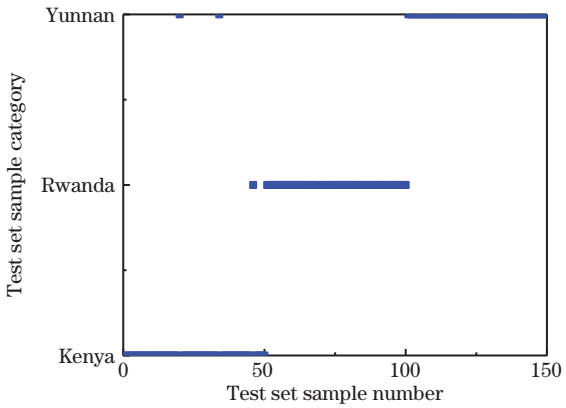

图 6 SVM 的预测集分类结果

Fig. 6 SVM prediction set classification results

表 3 三种建模算法的鉴别结果对比

Table 3 Comparison of identification results of three modeling algorithms

Modeling method	Pretreatment method	Number of principal components	Kenyan coffee beans		Rwandan coffee beans		Yunnan coffee beans		Overall accuracy / %
			Number of false positives	Accuracy / %	Number of false positives	Accuracy / %	Number of false positives	Accuracy / %	
SVM	No	9	6	88	0	100	3	94	94
	SG	9	4	92	0	100	2	96	96
	MSC	31	2	96	2	96	5	90	94
	SNV	31	0	100	1	100	2	96	98
	Baseline	**7**	**3**	**94**	**0**	**100**	**0**	**100**	**98**
BPNN	No	9	10	80	1	98	4	92	90
	SG	9	11	78	7	86	1	98	87
	MSC	31	13	74	6	88	26	48	70
	SNV	31	17	66	6	88	29	42	65
	Baseline	7	12	76	0	100	6	88	88
RF	No	9	5	90	1	98	6	88	92
	SG	9	4	92	4	92	7	86	90
	MSC	31	2	96	4	92	6	88	92
	SNV	31	5	90	6	88	2	96	91
	Baseline	7	4	92	3	94	7	86	91

4 结 论

利用太赫兹光谱技术实现对不同品种的咖啡豆鉴别研究。实验选取肯尼亚、卢旺达、云南三种常用品种的咖啡豆，采用透射式太赫兹光谱仪分别采集三种样品的时域光谱信息，并选取 0.5 THz 波段的光谱信息。结合主成分分析对原始光谱进行降维，随后经过 KS 算法将样本以 3∶1的比例分成训练集和测试集，再建立基于 PLS-DA 的二分类模型，模型显示卢旺达-云南咖啡豆鉴别的总正确率为 92%，肯尼亚-云南和肯尼亚-卢旺达咖啡豆鉴别的总正确率均为 98%，模型可以很好地对三种咖啡豆进行二分类鉴别。最后针对咖啡豆品种分别建立了 SVM、BPNN、RF 三种多分类模型，鉴别效果较为理想，通过比较，SVM 结合基线校正的建模效果最佳，其中卢旺达和云南咖啡豆能够全部被鉴别，肯尼亚咖啡豆的鉴别正确率也达到 94%，总正确率高达 98%。研究表明，基于太赫兹光谱的检测技术可以实现对不同品种咖啡豆的鉴别，可以拓展到其他农产品及食品品种的鉴别，具有较强的现实意义。

参 考 文 献

[1] Bao Y D, Chen N, He Y, et al. Rapid identification of coffee bean variety by near infrared hyperspectral imaging technology [J]. Optics and Precision Engineering, 2015, 23(2): 349-355.
鲍一丹, 陈纳, 何勇, 等. 近红外高光谱成像技术快速鉴别国产咖啡豆品种[J]. 光学 精密工程, 2015, 23(2): 349-355.

[2] Zeng F K, Liu A Q, Tan L H, et al. Fatty acid composition of coffee oil from three different origins [J]. Chinese Journal of Tropical Crops, 2011, 32(8): 1460-1463.
曾凡逵, 刘爱琴, 谭乐和, 等. 3 个不同产地咖啡脂肪酸组成比较[J]. 热带作物学报, 2011, 32(8): 1460-1463.

[3] Moreira I, Scarminio I S. Chemometric discrimination of genetically modified Coffea arabica cultivars using spectroscopic and chromatographic fingerprints[J]. Talanta, 2013, 107: 416-422.

[4] Hu J, Liu Y D, Sun X D, et al. Quantitative determination of benzoic acid in flour based on terahertz time-domain spectroscopy and BPNN model[J]. Laser & Optoelectronics Progress, 2020, 57(7): 073002.
胡军, 刘燕德, 孙旭东, 等. 基于 BP 神经网络的太赫兹时域光谱对面粉中苯甲酸定量检测研究[J]. 激光与光电子学进展, 2020, 57(7): 073002.

[5] Zhang J Y, Ren J J, Chen S H, et al. Application of wavelet denoising in terahertz nondestructive detection [J]. Chinese Journal of Lasers, 2020, 47(1): 0114001.
张霁旸, 任姣姣, 陈思宏, 等. 小波去噪在太赫兹无损检测中的应用[J]. 中国激光, 2020, 47(1): 0114001.

[6] Afsah-Hejri L, Hajeb P, Ara P, et al. A comprehensive review on food applications of terahertz spectroscopy and imaging[J]. Comprehensive Reviews in Food Science and Food Safety, 2019, 18(5): 1563-1621.

[7] Liu Y D, Du X Y, Li B, et al. Detection of purple rice adulteration by terahertz time domain spectroscopy [J]. Spectroscopy and Spectral Analysis, 2020, 40(8): 2382-2387.
刘燕德, 杜秀洋, 李斌, 等. 太赫兹时域光谱技术对紫米掺假的检测研究[J]. 光谱学与光谱分析, 2020, 40(8): 2382-2387.

[8] Ma Q X, Li C, Li T Y, et al. Research progress of terahertz spectroscopy in the field of pesticide detection[J]. Laser & Optoelectronics Progress, 2020, 57(13): 130006.
马卿效, 李春, 李天莹, 等. 太赫兹光谱技术在农药检测领域的研究进展[J]. 激光与光电子学进展, 2020, 57(13): 130006.

[9] Dorney T D, Baraniuk R G, Mittleman D M. Material parameter estimation with terahertz time-domain spectroscopy [J]. Journal of the Optical Society of America A, 2001, 18(7): 1562-1571.

[10] Dragoman D, Dragoman M. Time-frequency signal processing of terahertz pulses[J]. Applied Optics, 2004, 43(19): 3848.

[11] Li T J, Liu J J, Shao G F, et al. A novel THz spectroscopy recognition method for transgenic organisms based on APSO combined with SVM[J]. Optics and Spectroscopy, 2016, 120(4): 660-665.

[12] Wei X, Zheng W Q, Zhu S P, et al. Application of terahertz spectrum and interval partial least squares method in the identification of genetically modified soybeans[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 238: 118453.

[13] Poona N K, van Niekerk A, Nadel R L, et al. Random Forest (RF) wrappers for waveband selection and classification of hyperspectral data[J]. Applied Spectroscopy, 2016, 70(2): 322-333.

[14] Yuan H H, Yang G J, Li C C, et al. Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: analysis of RF, ANN, and SVM regression models[J]. Remote Sensing, 2017, 9(4): 309.

[15] Qin B Y, Li Z, Luo Z H, et al. Terahertz time-domain spectroscopy combined with PCA-CFSFDP applied for pesticide detection [J]. Optical and Quantum Electronics, 2017, 49(7): 1-12.