基于太赫兹时域光谱的牛黄及其易混品分类研究

章龙，李春，李天莹，张岩，蒋玲*
南京林业大学信息科学技术学院，江苏 南京 210037

摘要 采用太赫兹时域光谱技术，结合化学计量学方法，对牛黄及其易混品进行鉴别，获取了黄连、大黄、蒲黄、人工牛黄、掺杂牛黄和天然牛黄的太赫兹时域光谱图。分别构建了随机森林(RF)模型和三种参数优化的支持向量机(SVM)模型，对六种物质的太赫兹吸收光谱进行了分类鉴别。针对样品数据集不平衡导致的随机森林模型识别率下降的问题，提出了基于合成少数类过采样技术(SMOTE)的随机森林模型。结果表明，随机森林模型和 SVM 模型均可达到 95.00%左右的分类准确率，但随机森林模型具有更快的运行速度，运行时间仅为最优 PSO-SVM 模型运行时间的 2%。基于 SMOTE 的随机森林模型可有效地解决数据不平衡情况下识别率低的问题，识别率从数据不平衡情况下的 84.17%提高到 94.17%，计算速度基本不变。研究结论为基于太赫兹光谱技术的稀有中药的鉴别提供了新方法。

关键词 光谱学；太赫兹时域光谱；天然牛黄；随机森林；不平衡数据；支持向量机
中图分类号 O433.4 **文献标志码** A **doi**：10.3788/LOP57.233001

Classification of Calculus Bovis and Its Confounding Substances Based on Terahertz Time-Domain Spectroscopy

Zhang Long, Li Chun, Li Tianying, Zhang Yan, Jiang Ling*
College of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu 210037, *China*

Abstract We employ terahertz time-domain spectroscopy (THz-TDS) combined with chemometrics to identify Calculus bovis and its confounding substances, and obtain the THz-TDS of Coptidis rhizome, Rhubarb, Cattail pollen, Calculus bovis, artificial Calculus bovis, and adulterate Calculus bovis. The random forest (RF) classification model and the support vector machine (SVM) model which adopts three kinds of parameter optimization are established, respectively. The classification and identification of the THz absorption spectra of six kinds of matter are conducted. In addition, the RF model based on the synthetic minority over-sampling technique (SMOTE) is proposed to solve the problem that the recognition rate of the RF model decreases due to the serious unbalanced sample dataset. The results show that both the RF model and the SVM model can achieve a recognition rate of about 95.00%. However, the RF model can run much faster, whose running time is only 2% of that of the optimal PSO-SVM model. The RF model based on the SMOTE technique can effectively solve the problem of low recognition rate caused by unbalanced data. The recognition rate increases from 84.17% to 94.17%, and the operation speed is basically constant. The research conclusion provides a new approach for the identification of rare Chinese medicine using terahertz spectroscopy.

Key words spectroscopy; terahertz time-domain spectroscopy; Calculus bovis; random forest; unbalanced data; support vector machine
OCIS codes 300.6495; 330.6180; 040.2235

1 引 言

牛黄为中医常用的中药材，具有很高的药用价值。黄连、大黄和蒲黄研磨成粉末后的外观与牛黄粉接近，如图 1 所示，价格却只有牛黄的 1/300～1/500。随着天然牛黄资源的减少，市场上经常会

收稿日期：2020-03-13；**修回日期**：2020-04-02；**录用日期**：2020-04-20
基金项目：国家自然科学基金(31200541)、江苏省自然科学基金(BK20161526)
* **E-mail**：jiangling@njfu.edu.cn

出现用加工后的黄连、大黄、蒲黄或人工牛黄粉冒充天然牛黄粉的不法行径。因此，如何快速准确地识别牛黄具有十分重要的意义。目前，对于牛黄的鉴别，主要为经验鉴别，辅以理化鉴别，如薄层色谱法[1]、气相色谱法和高效液相色谱法[2]。上述检测方法虽然较为成熟，但对鉴定者有较高的要求，并且操作复杂、难度大且成本较高。

图 1 样品的外观图。(a)黄连；(b)大黄；(c)蒲黄；(d)牛黄

Fig. 1 Appearances of samples. (a) Coptidis rhizome; (b) Rhubarb; (c) Cattail pollen; (d) Calculus bovis

近年来，近红外光谱技术被广泛地应用于中药牛黄的检测。徐路等[3]利用偏最小二乘回归的类模型方法，对天然牛黄、人工牛黄粉和掺杂牛黄的近红外光谱数据进行了鉴别分析。聂黎行等[4]采用近红外光谱技术，结合模式识别方法，对天然牛黄、体外培育牛黄与人工牛黄进行了鉴别。以上研究在建立判别分析模型前，需要对光谱数据进行预处理，不同的预处理方法对模型的性能指数有较大影响。同时，近红外光谱的吸收谱带主要是 C—H、N—H、O—H 等基团的倍频和合频的吸收，谱峰重叠严重[5]，且较宽的吸收峰难以识别，用作定性分析时，实验数据的重复性较差，准确度低。

太赫兹波是指频率范围为 0.1～10 THz 的电磁波，具有电子和光学的双重特性[6]。由于太赫兹波独特的指纹图谱特性，THz 光谱技术结合化学计量学的方法在药物和食品检测领域得到广泛的应用[7-9]。Chen 等[10]研究了转基因和非转基因甜菜的 THz 光谱特性，利用主成分分析 PCA(principal component analysis)、聚类分析和偏最小二乘回归对两类甜菜进行了区分。胡晓华等[11]研究了三种不同产地咖啡豆的 THz 光谱特性，并结合 PCA 和支持向量机 SVM(support vector machines)模型进行了鉴别分析。张文涛等[12]利用 THz 时域光谱(TDS)技术对八种转基因大豆油进行了检测，在此基础上构建了 PCA-SVM 模型并对其进行了鉴别。在以上利用 SVM 模型对物质进行鉴别的研究中，需要对惩罚参数和核函数参数进行优化，建模时间较长，而无需参数优化的随机森林 RF(random forest)模型在保证识别率的前提下，可大大减少建模时间。目前，随机森林算法在分析化学领域有着广泛的应用，Liu 等[13]利用 RF 模型对橘子汁和食醋的电子舌数据进行了分类研究，并将分类结果与反向神经网络(BPNN)和 SVM 模型进行了对比，结果表明，RF 模型在建模效率和预测准确率上都优于 BPNN 和 SVM 模型。Zhu 等[14]利用多层感知器、RF 模型和 SVM 模型对红茶的发酵程度进行了预测，结果表明，RF 模型的预测准确率优于另外两种模型。此外，在近红外光谱[15]和拉曼光谱[16]中，RF 模型均有广泛的应用。

本文采用太赫兹光谱技术研究牛黄及其易混中药的光谱特性，利用提取到的吸收系数谱数据训练基于网格寻优法、遗传算法 GA(Genetic Algorithms)和粒子群优化算法 PSO(Particle Swarm Optimization)的 SVM 模型以及随机森林模型，并对比了模型的分类效果。同时，对于天然牛黄稀少、价格昂贵等原因导致的天然牛黄样品数据远少于其他样品的数据不平衡问题，提出了基于合成少数类过采样技术(Synthetic Minority Oversampling Technique，SMOTE)的随机森林分类模型，以期能够解决样品数据集不平衡导致的分类识别率低的问题。

2 实验部分

2.1 仪器及参数

实验采用的设备是 Advantest 公司的 TAS7500SP 型太赫兹时域光谱仪，系统测量范围为 0.1～4 THz，分辨率为 7.6 GHz，波形幅值最大值和最小值比值的动态范围高于 60 dB，激光发射器平均功率为 20 mW，脉冲中心波长为 1550 nm，脉冲宽度为 50 fs，激光重复频率为 50 MHz±200 Hz。仪器自带样品腔，测量时对样品腔内的空气进行干燥处理以减少空气中水汽对 THz 波吸收的影响。实验采用透射模式，THz-TDS 测试系统的原理图如图 2 所示，其中 A/D 表示将模拟信号转换为数字信号，FFT 表示快速傅里叶变换。

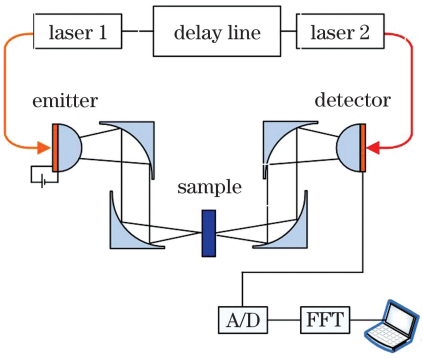

图 2 THz-TDS 测试系统的原理图

Fig. 2 Principle diagram of THz-TDS test system

2.2 样品制备及测试

实验选用的中药样品牛黄、黄连、大黄、蒲黄和人工牛黄均购于南京同仁堂中药店。使用粉碎机对样品进行粉碎，并利用筛子减小粉末颗粒，消除散射效应的干扰。使用压片模具将样品粉末压制成直径为 13 mm 和厚度为 0.90～1.20 mm 的圆形薄片，压力维持在 12 MPa 左右，两表面保持平行且光滑，每种样品分别制作 60 个样片。为减少样品混合不均匀带来的影响，对同一样品从三个不同位置分别测试一次，取三次测量的平均值作为样品的太赫兹光谱数据。

3 结果与讨论

3.1 光谱获取与分析

测试六种样品即黄连(Coptidis rhizome)、大黄(Rhubarb)、蒲黄(Cattail pollen)、人工牛黄(artificial Calculus bovis)、掺杂牛黄(adulterate Calculus bovis)和天然牛黄(Calculus bovis)的太赫兹时域光谱，其中掺杂牛黄为天然牛黄与人工牛黄按照质量比 1∶1 进行混合，六种样品的 THz 时域波形与背景信号波形如图 3 所示。可以看出，六种样品时域波形的幅值衰减以及相位偏移都存在差异，幅值衰减是样品表面的散射以及样品对太赫兹波的吸收造成的，天然牛黄的幅值衰减相比于其他样品较小。样品相位的延迟是样品厚度的差别和 THz 波在样品中的不同折射率所引起的。总体上看，牛黄类样品，包含天然牛黄、人工牛黄和掺杂牛黄，相对于黄连、大黄和蒲黄在相位上的延迟更小。为了避免样品表面颗粒大小不同以及样品厚度不均匀造成的误差，本文采用样品的吸收系数进行进一步分析。

六种样品的吸收光谱如图 4 所示。由于牛黄类样品在太赫兹较高频段(大于 1.60 THz)对信号的吸收较强，信噪比小，因此本文重点研究了 0.20～1.60 THz 频段。从图 4 可以看出，除了人工牛黄与掺杂牛黄外，其他几种样品均无明显的特征吸收峰。同时，人工牛黄、掺杂牛黄与天然牛黄吸收谱线的重叠现象较为严重，难以直接区分。针对此现象，本文结合化学计量学方法对六种样品进行鉴别区分。

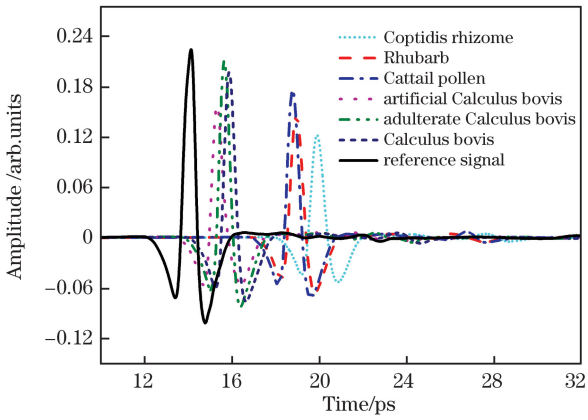

图 3 六种样品的时域光谱

Fig. 3 Time-domain spectra of six kinds of samples

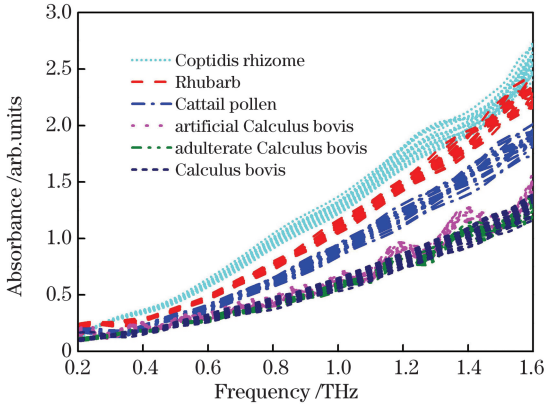

图 4 六种样品的太赫兹吸收谱

Fig. 4 THz absorption spectra of six kinds of samples

3.2 分类识别

利用吸收光谱数据分别建立了 SVM 分类模型和随机森林分类模型。将每种中药的 60 个样品数据随机分成两组，一组 40 个样品数据作为训练集，另一组 20 个样品数据作为测试集。利用训练得到的模型对测试集数据进行分类测试。同时，针对数据不平衡问题，建立了基于 SMOTE 的随机森林分类模型。

3.2.1 支持向量机识别

对于线性不可分训练数据，支持向量机用一个非线性映射函数将数据映射到高维特征空间，在高

维特征空间中构造出最优分类超平面并进行分类。　支持向量机的学习问题可表示为

$$\min_{w,b,\xi}\frac{1}{2}\left\|\boldsymbol{w}\right\|^2+C\sum_{i=1}^{N}\xi_i,\ \text{s.t.}\ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1-\xi_i,i=1,2,\cdots,N,\tag{1}$$

式中：C 为惩罚参数；$\boldsymbol{w}$ 和 b 分别为分类超平面的法向量和截距；N 为样本数；$\boldsymbol{x}_i$ 为第 i 个特征向量；y_i 为 $\boldsymbol{x}_i$ 的类标记；ξ_i 为松弛变量且 $\xi_i\geqslant 0$。通过拉格朗日乘子将约束条件融合到目标函数中，得到最优分类面：

$$f(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x},\boldsymbol{x}_i)+b,\tag{2}$$

式中：$f(\boldsymbol{x})$ 为 SVM 分类决策函数；$\boldsymbol{x}$ 为特征向量；α_i 为拉格朗日乘子；$K(\boldsymbol{x},\boldsymbol{x}_i)$ 为核函数。

本文 SVM 的核函数选用径向基核函数(radial basis function，RBF)。RBF 的定义为

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|}{2\sigma^2}\right),\sigma>0,\tag{3}$$

式中：σ 为核函数参数。

对于具有 RBF 的 SVM，需要对惩罚参数 C 和核函数参数 σ 进行参数寻优。本文对比了基于三种参数优化算法的 SVM 模型，分别是网格寻优法(K-CV-SVM)、遗传算法(GA-SVM)和粒子群优化算法(PSO-SVM)。

将黄连、大黄、蒲黄、人工牛黄、掺杂牛黄和天然牛黄分别定义为标签 1～6。计算得到三种 SVM 模型的优化参数 (C,σ)，分别为(1，2.8284)，(0.7539，3.0606)和(23.6865，2.1549)。三种 SVM 模型的识别率如表 1 所示。其中 K-CV-SVM 和 PSO-SVM 模型的识别率均达到了 94.17%，但是三种模型的训练耗时均较长。

3.2.2　随机森林识别

为减小识别模型的运行时间，提出了随机森林 RF 模型，随机森林算法是由 Breiman[17] 在决策树的基础上提出的一个机器学算法。该算法对参数设置不敏感，在决策树数量足够多的情况下，无需参数优化就能达到理想的分类效果，相比于其他算法，其计算时间显著减小。

利用随机森林算法建立牛黄及其易混品的太赫兹吸收光谱识别模型。随机森林 RF 模型仅需设置决策树的数量，实验中设置 RF 模型的初始决策树数量为 100，并反复验证，每次增加 50 棵决策树直至数量达到 1000。结果显示，随机森林模型决策树的数量对分类结果没有影响。RF 模型的识别率如表 1 所示，整体识别率为 95%，稍高于 PSO-SVM 模型，但是模型训练时间大大减小。测试得出，K-CV-SVM 模型的训练耗时为 4039.1554 s，GA-SVM 模型的训练耗时为 1258.14 s，PSO-SVM 模型的训练耗时为 412.1286 s，RF 模型的训练耗时为 8.4805 s。RF 模型的训练时间仅为 K-CV-SVM 模型的 0.2%，GA-SVM 模型的 0.6%和 PSO-SVM 模型的 2%。

表 1　模型识别率

Table 1　Identification rate of each model　　unit：%

Algorithm	K-CV-SVM	GA-SVM	PSO-SVM	RF
Coptidis rhizome	100	100	100	100
Rhubarb	95	95	100	100
Cattail pollen	90	90	90	90
Artificial Calculus bovis	95	90	90	95
Adulterate Calculus bovis	90	90	90	90
Calculus bovis	95	90	95	95
Total	94.17	92.50	94.17	95.00

3.2.3　基于 SMOTE 的随机森林识别

天然牛黄为名贵中药，数量稀少且价格昂贵。由于天然牛黄较稀少，我们将天然牛黄的训练集样本数量从 40 个减少到 5 个，其他中药的训练集样本数量保持 40 个不变，每种待测中药样本的测试集数量保持 20 个不变，其他参数设置与 3.1 小节一致。计算结果显示，三种 SVM 模型的识别率略有下降，K-CV-SVM 模型的识别率为 93.33%，GA-SVM 模型的识别率为 90.83%，PSO-SVM 模型的识别率为 93.33%。随机森林模型的训练耗时基本保持不

变，但是识别率明显变差，判断错误个数由 6 个增加到 19 个，识别率由 95.00%降为 84.17%。

为了避免过少的天然牛黄用量导致的分类模型性能的下降，需要解决不平衡数据的分类问题。所谓不平衡数据的分类问题，是指某类样本数量远小于其他类样本数量而对分类结果产生不利影响的问题。解决分类模型的数据集不平衡问题主要有两种思路，一种是对学习算法进行改进，另一种是对数据集进行平衡处理。Chawla 等[18]提出的 SMOTE 是平衡数据集的一种代表算法，对由传统过采样算法引起的分类过拟合现象有明显改善的作用，因此被广泛应用于不平衡数据集的分类中[19]。该算法的核心思想是在少数类样本集中，对邻近的样本进行插值来产生新样本，可增加稀有类样本的数目，改善数据集的不平衡状况。

为提高随机森林算法的识别率，加入 SMOTE 以改善不平衡问题。SMOTE 的主要步骤如下：

1)根据过采样倍率 N'，针对每个稀有类样本找出 k 个同类最近邻，然后在其中随机选择 N 个样本。

2)针对每个稀有类样本，利用选出的 N 个样本生成 N 个新的稀有类样本：

$$\boldsymbol{X}_{\text{new}}=\boldsymbol{X}+\text{rand}(0,1)\times(\boldsymbol{Y}_j-\boldsymbol{X}_j),\qquad(4)$$

式中，rand(0,1)表示 0 和 1 之间的一个随机数；$\boldsymbol{X}_{\text{new}}$ 为增加的新样本；$\boldsymbol{X}$ 为少数类样本；$\boldsymbol{Y}_j$ 为 $\boldsymbol{X}$ 的最邻近样本，其中 $j=1,2,\cdots,k$ 。

3)将新的样本加入原训练数据集中，形成新的训练数据集。

SMOTE 中最重要的参数是过采样倍率 N'，其代表稀有类样本增加的倍数，本文设置 N' 的初始值为 4。将 N' 值依次增加 1，反复验证随机森林模型的识别率，结果显示，当 N' 为 8 即将稀有类数据增加 8 倍时，随机森林模型的识别率提高为 94.17%，如表 2 所示。当 N' 的取值大于 8 时，随机森林的识别率呈下降趋势，原因是 SMOTE 中的新样本并不是真实样本，而是对邻近的样本进行插值产生的，如果 N' 的取值过大，会造成训练集质量下降，从而影响模型的分类准确率。结果表明，SMOTE 对三种 SVM 模型识别率的影响较小，但对随机森林模型的识别效果有较大的提升作用。基于 SMOTE 的 K-CV-SVM 模型的训练时间为 4170.1392 s，基于 SMOTE 的 GA-SVM 模型的训练时间为 1091.7193 s，基于 SMOTE 的 PSO-SVM 模型的训练时间为 407.1417 s，基于 SMOTE 的 RF 模型的训练时间为 9.5128 s。

表 2 基于 SMOTE 的 4 种模型的识别率

Table 2 Identification rates of four kinds of models basedon SMOTE unit:%

Algorithm	K-CV-SVM	GA-SVM	PSO-SVM	RF
Coptidis rhizome	100	95	100	100
Rhubarb	95	95	95	100
Cattail pollen	90	90	95	90
Artificial Calculus bovis	90	90	95	90
Adulterate Calculus bovis	90	85	90	90
Calculus bovis	95	90	90	95
Total	93.33	90.83	94.17	94.17

因此，基于 SMOTE 的随机森林模型既保持了随机森林模型运算速度快的优点，又提高了处理不平衡问题的能力，保证了模型的分类准确率。

4 结 论

利用太赫兹时域光谱技术研究了天然牛黄及其易混品的 THz 吸收谱图，并结合支持向量机模型和随机森林模型，对牛黄及其易混品进行了鉴别分析。除了人工牛黄与掺杂牛黄外，其他几种样品均无明显的特征吸收峰。同时，人工牛黄、掺杂牛黄与天然牛黄吸收谱线的重叠现象较为严重，难以直接区分。构建了随机森林模型和三种参数优化的 SVM 模型，对样品进行了分类鉴别，结果表明，随机森林模型和 SVM 模型均可达到 95.00%的分类准确率，但随机森林模型具有更快的运行速度，运行时间仅为最优 PSO-SVM 模型运行时间的 2%。对于数据不平衡问题导致的随机森林模型识别率下降的问题，提出了基于 SMOTE 的随机森林模型，改进后的随机森林模型解决了数据不平衡情况下的识别率低的问题，识别率从 84.17%提高到

94.17%,计算速度基本不变。研究结果为牛黄及其易混品的鉴别提供了高效准确的方法,也为其他类型名贵物质的鉴定提供了重要的参考。

致谢 作者感谢南京林业大学现代分析测试中心提供的设备及技术帮助。

参 考 文 献

[1] Zou Q W, Shi Y, Liu W, et al. The research of quantity test method of various components in succession medicinal substances of cow-bezoar[J]. Chinese Journal of Pharmaceutical Analysis, 2015, 35(1): 8-15.
邹秦文, 石岩, 刘薇, 等. 牛黄类药材各类成分定量检测方法研究概况[J]. 药物分析杂志, 2015, 35(1): 8-15.

[2] Li K, Qi Y X, Yu X L, et al. Comparison of HPLC fingerprint between enzymatic Calculus bovis and natural Calculus bovis[J]. Chinese Traditional Patent Medicine, 2011, 33(1): 1-5.
李珂, 齐永秀, 于秀玲, 等. 酶促牛黄与天然牛黄HPLC指纹图谱比较研究[J]. 中成药, 2011, 33(1): 1-5.

[3] Xu L, Fu H Y, Jiang N, et al. A new class model based on partial least square regression and its applications for identifying authenticity of bezoar samples[J]. Chinese Journal of Analytical Chemistry, 2010, 38(2): 175-180.
徐路, 付海燕, 姜宁, 等. 基于偏最小二乘回归的类模型方法用于中药牛黄的真伪鉴别[J]. 分析化学, 2010, 38(2): 175-180.

[4] Nie L X, Zhang Y, Hu X R, et al. Fast and non-destructive identification of Bovis Calculus, Bovis Calculus Sativus and Bovis Calculus Artifactus by near infrared spectroscopy combined with pattern recognition technology[J]. Chinese Journal of Pharmaceutical Analysis, 2017, 37(10): 1897-1903.
聂黎行, 张烨, 胡晓茹, 等. 近红外光谱法结合模式识别技术快速无损鉴别天然牛黄、体外培育牛黄和人工牛黄[J]. 药物分析杂志, 2017, 37(10): 1897-1903.

[5] Ma Q, Hao G Q, Qiao Y J, et al. Determination of the artificial bezoar powder in bezoar powder by near-infrared spectrometry and support vector machine[J]. Spectroscopy and Spectral Analysis, 2006, 26(10): 1842-1845.
马群, 郝贵奇, 乔延江, 等. 近红外光谱法结合支持向量机测定天然牛黄粉中人工牛黄的掺入量[J]. 光谱学与光谱分析, 2006, 26(10): 1842-1845.

[6] Li H T, Wang X K, Zhang Y. Study and applications of terahertz special beams[J]. Chinese Journal of Lasers, 2019, 46(6): 0614007.
李鹤婷, 王新柯, 张岩. 太赫兹特殊光束的研究与应用[J]. 中国激光, 2019, 46(6): 0614007.

[7] Peng Y, Shi C J, Zhu Y M, et al. Qualitative and quantitative analysis algorithms based on terahertz spectroscopy for biomedical detection[J]. Chinese Journal of Lasers, 2019, 46(6): 0614002.
彭滟, 施辰君, 朱亦鸣, 等. 太赫兹光谱技术在生物医学检测中的定性与定量分析算法[J]. 中国激光, 2019, 46(6): 0614002.

[8] Liu J X, Du B, Deng Y Q, et al. Terahertz-spectral identification of organic compounds based on differential PCA-SVM method[J]. Chinese Journal of Lasers, 2019, 46(6): 0614039.
刘俊秀, 杜彬, 邓玉强, 等. 基于差分-主成分分析-支持向量机的有机化合物太赫兹吸收光谱识别方法[J]. 中国激光, 2019, 46(6): 0614039.

[9] Li T, Zhang L, He J A, et al. Rapid online identification of hazardous substances in mail using terahertz technology[J]. Laser & Optoelectronics Progress, 2019, 56(23): 233001.
李涛, 张良, 何建安, 等. 基于太赫兹技术在线快速识别邮件隐匿危险品[J]. 激光与光电子学进展, 2019, 56(23): 233001.

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

[11] Hu X H, Liu W, Liu C H, et al. Rapid identification of producing area of coffee bean based on terahertz spectroscopy and support vector machine[J]. Transactions of the CSAE, 2017, 33(9): 302-307.
胡晓华, 刘伟, 刘长虹, 等. 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.

[12] Zhang W T, Li Y W, Zhan P P, et al. Recognition of transgenic soybean oil based on terahertz timedomain spectroscopy and PCA-SVM[J]. Infrared and Laser Engineering, 2017, 46(11): 1125004.
张文涛, 李跃文, 占平平, 等. 基于太赫兹时域光谱技术与PCA-SVM的转基因大豆油鉴别研究[J]. 红外与激光工程, 2017, 46(11): 1125004.

[13] Liu M, Wang M J, Wang J, et al. Comparison of

random forest, support vector machine and back propagation neural network for electronic tongue data classification: application to the recognition of orange beverage and Chinese vinegar [J]. Sensors and Actuators B: Chemical, 2013, 177: 970-980.

[14] Zhu H K, Liu F, Ye Y, et al. Application of machine learning algorithms in quality assurance of fermentation process of black tea based on electrical properties[J]. Journal of Food Engineering, 2019, 263: 165-172.

[15] Donald D, Coomans D, Everingham Y, et al. Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics [J]. Chemometrics and Intelligent Laboratory Systems, 2006, 82(1/2): 122-129.

[16] Xu H D, Lin L L, Li Z, et al. Nephrite origin identification based on Raman spectroscopy and pattern recognition algorithms [J]. Acta Optica Sinica, 2019, 39(3): 0330001.
徐荟迪, 林露璐, 李征, 等. 基于拉曼光谱和模式识别算法的软玉产地鉴别[J]. 光学学报, 2019, 39(3): 0330001.

[17] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[18] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[19] Ramentol E, Caballero Y, Bello R, et al. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory[J]. Knowledge and Information Systems, 2012, 33(2): 245-265.