

文章编号 1004-924X(2024)20-3006-11

太赫兹融合光谱结合改进 Fused Lasso 模型在转基因菜籽油鉴别中的应用

陈 涛¹, 谢光翀¹, 张绍荣^{1,2*}

(1. 桂林电子科技大学 电子工程与自动化学院, 广西 桂林 541004;

2. 桂林航天工业学院 电子信息与自动化学院, 广西 桂林 541004)

摘要: 现有基于单一光谱的转基因菜籽油分类鉴别模型, 存在包含信息少、数据维度高等问题, 导致模型运行效率较低、检测结果不够准确。针对此问题, 本研究提出一种太赫兹融合光谱结合改进 Fused Lasso 模型的转基因菜籽油分类鉴别方法。以两种转基因菜籽油和两种非转基因菜籽油为研究对象, 应用太赫兹时域光谱 (THz-TDS) 系统获取 4 种菜籽油样品在 0.2~1.6 THz 频率范围内的太赫兹吸收光谱, 采用连续投影 (SPA) 算法对样品的太赫兹吸收光谱和导数光谱进行特征提取后再融合, 引入特征选择和分类为一体的正则化稀疏模型 Fused Lasso, 通过采用一对一 (OVO) 方法将其改进为多分类模型并采用贝叶斯优化 (BO) 算法对其正则化参数寻优。结果表明, 相比传统基于单一吸收光谱的 Fused Lasso 模型, 基于融合光谱的 BO-Fused Lasso 模型对 4 种菜籽油分类效果更好, 其训练集准确率为 96.88%, 测试集准确率为 95.00%。因此, 本研究为转基因菜籽油和非转基因菜籽油的鉴别提供了一种新方法, 也为其他转基因物质鉴别提供了有价值的参考。

关键词: 太赫兹; 转基因菜籽油; 融合光谱; 正则化稀疏模型; 贝叶斯优化

中图分类号: O439 文献标识码: A doi: 10.37188/OPE.20243220.3006

Application of terahertz fusion spectrum combined with improved fused lasso model in identification of transgenic rapeseed oil

CHEN Tao¹, XIE Guangchong¹, ZHANG Shaorong^{1,2*}

(1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology,
Guilin 541004, China;

2. School of Electronic Information and Automation, Guilin University of Aerospace Technology,
Guilin 541004, China)

* Corresponding author, E-mail: zsrong@guat.edu.cn

Abstract: Existing classification and identification models for transgenic rapeseed oil based on single-spectrum analysis suffer from limited information and high data dimensionality, leading to low operational efficiency and inaccurate detection results. To address these issues, a novel classification method for transgenic rapeseed oil was proposed, leveraging terahertz fusion spectroscopy combined with an improved Fused Lasso model. Using two types of transgenic rapeseed oil and two types of non-transgenic rapeseed oil as

收稿日期: 2024-06-16; 修订日期: 2024-08-23.

基金项目: 国家自然科学基金资助项目 (No. 62261012, No. 61841502)

research subjects, the terahertz time-domain spectroscopy (THz-TDS) system was employed to obtain the terahertz absorption spectra of the four rapeseed oil samples in the frequency range of 0.2 to 1.6 THz. Features were extracted from the absorption and derivative spectra using the successive projections algorithm (SPA), and then fused. Introducing a regularization sparse model, Fused Lasso, which integrates feature selection and classification. This model was improved into a multi-class model using the one-vs-one (OVO) method, and Bayesian optimization (BO) was employed to optimize its regularization parameters. The results demonstrated that the BO-Fused Lasso model, based on fusion spectra, significantly outperformed the traditional Fused Lasso model based on a single absorption spectrum in classifying the four types of rapeseed oil. The accuracy rates for the training and testing sets were 96.88% and 95.00%, respectively. This study, therefore, presents a novel approach for accurately identifying transgenic and non-transgenic rapeseed oils and provides a valuable reference for the detection of other transgenic substances.

Key words: terahertz; transgenic rapeseed oil; fusion spectrum; regularized sparse model; Bayesian optimization

1 引言

中国是世界第一大菜籽油生产国和消费国,菜籽油在国民经济和生活中占据着重要的地位。菜籽油富含维生素E、不饱和脂肪酸和磷脂等多种营养成分,具有保护心血管健康、降低胆固醇、提供人体必要的营养成分等作用。随着转基因技术的迅速发展,转基因作物得到大量种植,根据权威数据显示^[1],2022年转基因油菜种植面积占全球油菜作物种植面积的23.8%。目前我国尚未批准转基因油菜商业化种植,但批准了部分转基因油菜籽的进口,主要用于生产菜籽油。转基因菜籽油已成为生活中常见的食用油,但截至目前还没有任何研究能够彻底否认其潜在危害^[2],公众对于转基因产品的态度也存在争议。市面上不注明转基因标示或将转基因产品标示为非转基因的行为屡见不鲜,因此对转基因菜籽油的鉴别具有重要的现实意义。目前传统的转基因检测技术主要有外源核酸检测^[3]和外源蛋白检测^[4]两类,但以上两种方法都存在检测过程繁琐、会破坏原有样品以及需要专业人员操作等缺点,因此亟须建立一种新型的更快速、无损、高效、绿色的转基因菜籽油检测方法。

太赫兹时域光谱(Terahertz Time-Domain Spectroscopy, THz-TDS)技术是一种利用太赫兹波段的电磁波进行检测和分析的技术,具有无损性、指纹性和低能性等优点。许多低能量物理现象和生物分子活动可以在太赫兹频率范围内

被观察到^[5],迄今为止许多学者已经在食品、医学、安检等领域对该技术进行了相应的研究与应用,其中也包括转基因食用油的检测研究。Zhou等^[6]利用THz-TDS技术结合极限学习机方法对转基因大豆油和非转基因大豆油进行了鉴别;Liu等^[7]利用THz-TDS技术结合加权线性判别分析方法对转基因山茶油进行了鉴别;陈涛等^[8]利用THz-TDS技术结合基于改进蜉蝣优化算法的支持向量机模型对转基因菜籽油进行了鉴别。

由于转基因和非转基因食用油的外观、气味和光谱极为相似,难以直接进行区分,需要结合机器学习方法才能准确鉴别。转基因油脂的太赫兹光谱数据多为高维度小样本数据,直接进行建模可能会导致过拟合,增加计算复杂度,且单一光谱包含的信息不足,难以取得满意的鉴别效果。因此本文首先采用连续投影算法(Successive Projections Algorithm, SPA)对菜籽油的太赫兹吸收光谱和导数光谱进行特征提取后融合,然后引入特征选择和分类为一体的正则化稀疏模型Fused Lasso并将其改进为多分类模型。由于Fused Lasso模型需要选取合适的正则化参数才能有效提升其性能,因此采用贝叶斯优化(Bayesian Optimization, BO)算法替代繁琐耗时的交叉验证过程对其正则化参数寻优,从而获得一种基于太赫兹融合光谱的改进Fused Lasso模型(BO-Fused Lasso)。实验结果表明,该模型具有较好的分类鉴别效果,对4种转基因和非转基因

因菜籽油鉴别的训练集准确率可达 96.88%, 测试集准确率可达 95.00%。

2 实验部分

2.1 实验设备

本文采用的太赫兹光谱实验测量设备为中国华讯方舟科技有限公司生产的 CCT-1800 型 THz-TDS 系统, 该系统主要由飞秒激光器、时间延时控制模块、太赫兹波产生装置以及太赫兹波探测装置 4 个部分构成, 系统原理图如图 1 所示^[9]。该系统光谱范围为 0.1~4.5 THz, 信噪比高于 75 dB, 延时扫描范围最大 1 200 ps。系统工

作原理如下: 首先飞秒激光器发射中心波长为 780 nm 的激光脉冲, 该脉冲通过分束镜分为泵浦光和探测光, 泵浦光入射到发射天线上并在偏置电压的激发下产生太赫兹脉冲, 然后太赫兹脉冲经过离轴抛物面镜聚焦后照射到实验样品上, 携带着样品信息的太赫兹脉冲再次经过离轴抛物面镜并最终与探测光一同汇聚入射到接收天线上。接收天线将处理过的数据传输到计算机, 从而获得样品的太赫兹时域光谱信号。为了确保实验数据的准确性, 在进行实验前必须消除水汽对太赫兹波吸收的干扰, 需要在样品实验舱中通入干燥氮气, 使得样品舱内相对湿度低于 2%。

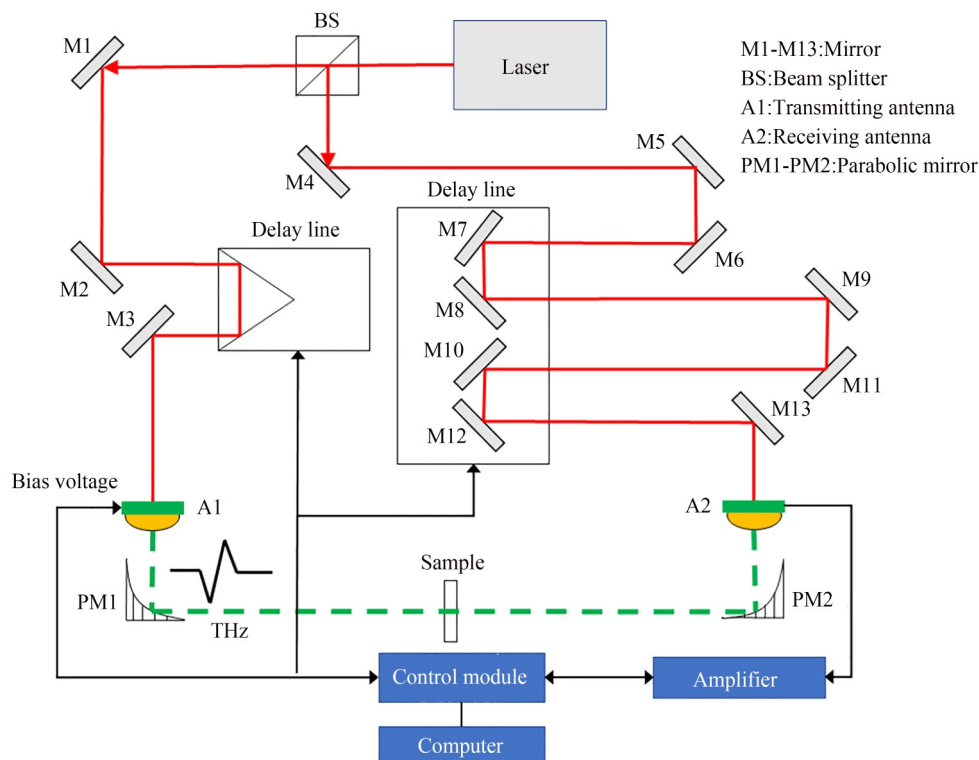


图 1 CCT-1800 型 THz-TDS 系统原理图

Fig. 1 Schematic diagram of CCT-1800 THz-TDS system

2.2 样品制备

实验样品共有 4 种, 为市面上常见的金龙鱼 (Jinlongyu) 和道道全 (Daodaoquan) 品牌的转基因菜籽油和非转基因菜籽油, 所有油样均为具有国家质量监督检验检疫认证的合格产品, 样品信息如表 1 所示。为了避免油样变质和氧化, 所有样品在制备前均在室温避光环境下储存。由于聚四

氟乙烯在太赫兹波段的低吸收特性, 实验样品架选择窗片材料为聚四氟乙烯薄膜的可拆卸液体池, 窗片的厚度为 0.5 mm, 窗片中心为面积 270 mm² 的椭圆孔。每次制样采用移液枪吸取 2 ml 的油样, 沿液体池壁将油样缓慢注满液体池, 以避免气泡的产生。每种菜籽油制作 60 个样本, 共计 240 个样本, 其中训练集 160 个, 测试集 80 个。

表 1 实验样品信息

Tab. 1 Information of experimental sample

Symbol	Brand	Type	Training set	Testing set
Non-GMO1	Daodaoquan	non transgenic	40	20
Non-GMO2	Jinlongyu	non transgenic	40	20
GMO1	Daodaoquan	transgenic	40	20
GMO2	Jinlongyu	transgenic	40	20

2.3 光学参数提取

使用 THz-TDS 技术对物质进行检测时,常用吸收系数、吸光度等光学参数表征物质的性质。其中吸光度是物质对太赫兹波吸收程度的物理量,可用于描述样品的吸收特性,并用于后续的光谱分析。为了获得样品的吸光度,需要对测得的太赫兹时域参考信号和样品信号进行快速傅里叶变换得到对应的频域信号,然后根据公式(1)获得样品的吸光度,作为 4 种菜籽油分类和识别的主要依据。

$$Absorbance(\omega) = -\log_{10} \left| \frac{E_{\text{sam}}(\omega)}{E_{\text{ref}}(\omega)} \right|^2, \quad (1)$$

其中: $E_{\text{ref}}(\omega)$ 为频域参考信号, $E_{\text{sam}}(\omega)$ 为频域样品信号。

2.4 模型评价指标

为了更好地评估分类模型的性能,采用分类精度作为模型评价指标。总体分类精度是正确分类的样本数占总样本数的比例之和,计算公式如式(2)所示:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (2)$$

其中: TP 为真正类,是模型将标签为正类的样本正确预测为正类的个数,例如将 Non-GMO1 的样本正确预测为 Non-GMO1; FP 为假正类,是模型将负类样本错误预测为正类的个数,例如将其他样本错误预测为 Non-GMO1; TN 为真负类,是模型将负类样本正确预测为负类的个数,例如将其他样本正确预测为其他样本; FN 为假负类,是模型将正类样本错误预测为负类的个数,例如将 Non-GMO1 的样本错误预测为其他样本。

3 分类模型

3.1 特征提取和特征融合

由于采用包含信息较少的单一光谱建模可

能达不到较好的分类效果,因此选择合适的特征提取方法从光谱中提取出对分类重要的信息,并将提取的特征融合可以达到提高模型分类准确率的目的。连续投影算法(SPA)是一种特征提取方法,通过循环搜索从原始变量集中选择变量来构建一组具有最小共线性以及最小冗余度的变量组合^[10],在尽可能保留原始光谱数据信息的情况下减少变量数量,从而提高建模的效率。其过程是从随机选定的波长开始,通过循环计算已选波长与未选波长之间的投影向量,选出投影向量最大的波长,逐步构建变量组合,直至达到预定的波长数量。运用连续投影算法能够从高维数据中提取重要的变量,达到降维的效果,减少建模数据的变量数量。

特征融合是对不同传感器或不同类型的数据分别进行特征提取,再将提取的信息进行融合,它可以有效综合各种数据的优势,弥补了单一光谱建模的不足^[11-12]。本文选择菜籽油的太赫兹吸收光谱和导数光谱进行特征融合,导数光谱通过对太赫兹吸收光谱求一阶导数得到,与原始光谱相比不仅可以消除或减少噪声的干扰,还能提高信噪比,准确确定吸收峰的位置^[13]。将导数光谱与吸收光谱进行融合,可以更准确地描述物质的特征。鉴于此,本文采用 SPA 分别对菜籽油的太赫兹吸收光谱和导数光谱进行特征提取,并对提取的特征进行融合得到融合光谱,以融合光谱建立分类鉴别模型,并与单一光谱进行对比。

3.2 Fused Lasso 模型

近年来正则化稀疏模型如最小绝对值收缩和选择算子(Lasso)、弹性网(Elastic Net)、平滑截断绝对偏差(Smoothly Clipped Absolute Deviation, SCAD)惩罚等方法被广泛应用于自然语言处理、金融建模、生物信息预测以及光谱的特征选择等,发挥着越来越重要的作用。在机

器学习中常遇到变量空间维度很高而样本空间维度低的数据,称为高维小样本数据,例如物质的太赫兹光谱数据。高维小样本数据会导致过拟合问题,降低模型的泛化能力^[14],而正则化稀疏模型中很多方法属于嵌入式特征选择方法^[15],即特征选择和分类融为一体,能够将数据中的冗余变量剔除,只保留最相关的解释变量,实现数据降维的同时也完成了分类任务。

Fused Lasso 是一种具有自动组效应的正则化稀疏模型^[16],自动组效应是指模型在变量选择时某些高度相关的变量作为一个组被同时选中参与模型的构造,或同时被移除不参与模型的构造,从而不会遗漏某些重要的变量。线性回归模型如式(3)^[17]:

$$y = X\beta + \epsilon, \quad (3)$$

其中: $X = [x_1, x_2, \dots, x_n]^T \in R^{n \times m}$ 为 4 种菜籽油的样本集变量, $y = [y_1, y_2, \dots, y_n]^T \in R^{n \times 1}$ 为各个样本对应的标签, $\beta = [\beta_1, \beta_2, \dots, \beta_m]^T \in R^{m \times 1}$ 是回归系数向量, $\epsilon \in R^{n \times 1}$ 为误差向量,其中 n 为样本数量, m 为样本的特征变量数。基于式(3)线性回归模型的 Fused Lasso 的目标函数如式(4)^[17]:

$$\hat{\beta} = \arg \min \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{p=2}^m |\beta_p - \beta_{p-1}| \right\}, \quad (4)$$

其中: $\frac{1}{2} \|y - X\beta\|_2^2$ 为残差平方和项,用于衡量模型预测值和实际值之间的误差。 $\lambda_1 \geq 0, \lambda_2 \geq 0$, 分别为 L1 范数罚 $\|\beta\|_1$ 和融合罚 $\sum_{p=2}^m |\beta_p - \beta_{p-1}|$ 的正则化参数, L1 范数罚能够压缩回归系数的大小,使得绝对值较小的回归系数被压缩为 0,从而产生稀疏解和实现变量选择^[18]。融合罚是对相邻变量的回归系数之差的绝对值进行惩罚,使相邻变量的回归系数几乎相等,相邻变量可以作为一个组被选择参与或不参与模型的构建。Fused Lasso 模型通过增加融合罚避免像 Lasso 对所有变量回归系数进行同等程度的压缩,使得具有相关性的相邻变量参与到模型的构建中,不会遗漏某些重要变量,达到提高

模型预测精度的目的。

3.3 Fused Lasso 模型的改进

3.3.1 应用一对一方法改进 Fused Lasso 为多分类模型

由于传统的正则化稀疏模型 Fused Lasso 只适用于二分类问题,需要将其改成多分类以区分 4 种菜籽油。一对一方法(One vs One, OVO)的主要思想是在 N 分类问题下训练 $N(N-1)/2$ 个二分类器,每个分类器区分两个类别,最终通过投票产生分类结果^[19]。在训练阶段,每个二分类器针对两个特定的类别进行训练,将一类的样本视为正类,另一类样本视为负类。在测试阶段,将测试集输入到每一个训练好的二分类器中得到 $N(N-1)/2$ 个预测结果,然后对每个样本的预测结果进行投票,得票最多的结果即为最终预测结果。OVO 方法通过将多分类问题转化成多个二分类问题,为每一对类别构建专门的二分类器,降低了分类问题的复杂性且能得到准确的结果,因此采用 OVO 方法将 Fused Lasso 改成多分类,扩展了其应用范围。本文对于 4 种菜籽油的分类问题,需要建立并训练 6 个二分类器,然后将测试集输入到训练好的模型中得到结果。

3.3.2 引入贝叶斯优化算法寻优 Fused Lasso 的正则化参数

正则化参数决定着正则化稀疏模型的性能,选择合适的正则化参数能够提高分类精度。Fused Lasso 的分类精度取决于 L1 范数罚和融合罚的正则化参数 λ_1 和 λ_2 ,正则化参数一般通过交叉验证或者网格搜索法得到,但是这两种方法繁琐耗时,针对非凸问题容易陷入局部最优,无法找到最优参数,因此本研究引入贝叶斯优化算法对两个正则化参数进行寻优。

贝叶斯优化(Bayesian Optimization, BO)是一种高效的全局优化算法,主要用于机器学习中超参数的调整。它通过高斯过程来构建超参数目标函数的概率代理模型,采集函数以概率代理模型为先验信息,在每次迭代中选择最有可能改善目标函数的模型参数用于更新模型,使得更新后的概率代理模型逐渐逼近真实的目标函数,从而获得最优的超参数值^[20]。贝叶斯优化算法的核心是概率代理模型和采集函数,基于概率代理

模型的先验信息使得采集函数能够在较少的迭代次数内获得最优的超参数值,减少了不必要的寻参次数,节省了大量的计算成本。相对于交叉验证和网格搜索法,贝叶斯优化算法有高效的寻参能力,经过OVO方法改成多分类的Fused Lasso模型计算成本较高,采用贝叶斯优化能够节省大量计算成本。因此,为提高建模效率和分类精度,本文引入贝叶斯优化替代交叉验证和网格搜索用于Fused Lasso的正则化参数寻优,将贝叶斯优化的目标函数设置为训练集的最小分类错误率,形成一种BO-Fused Lasso模型,用于转基因与非转基因菜籽油的分类鉴别。

4 结果及分析

4.1 光谱分析

应用THz-TDS系统对4种菜籽油样品进行测量,实验设置的扫描窗口长度为30 ps,光谱分辨率约为33.3 GHz,光谱测量强度的方差约为 4.3×10^{-4} 。得到4种菜籽油样品以及参考信号的太赫兹时域谱如图2所示,参考信号为系统稳定后样品架空载时获取。通过计算获得240个菜籽油样品在0.2~1.6 THz频段的太赫兹吸收光谱如图3(a)所示。由图可见,所有菜籽油样品在0.2~1.6 THz频段呈现出相似的波形和相近的幅值,无显著差异。通过对每种菜籽油60个样本

的吸光度取平均,得到4种菜籽油的太赫兹平均吸收光谱如图3(b)所示,可见4种菜籽油在分析频段内均存在明显且相似的吸收峰,在0.36, 0.55, 0.75, 1.11和1.34 THz处都有明显的吸收峰,在0.96和1.52 THz有两个不太明显的吸收峰,吸收峰所处频率位置基本一致,只是不同种类菜籽油的吸光度强度有较小差异。这说明转基因菜籽油和非转基因菜籽油的太赫兹光谱差异非常小,采用直接观察法很难对其进行鉴别,因此需要结合机器学习方法来实现对转基因和非转基因菜籽油的快速有效鉴别。

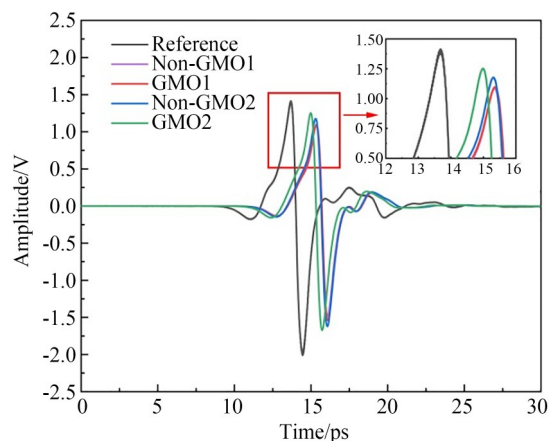
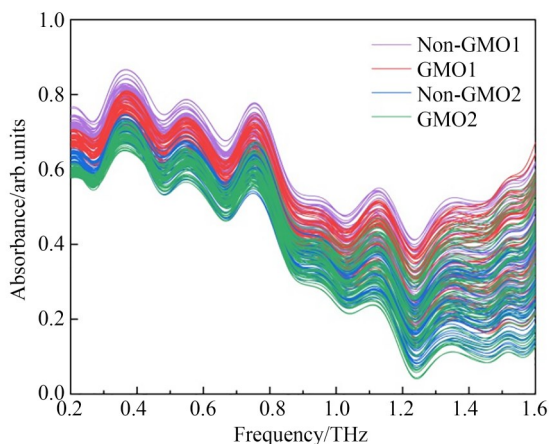
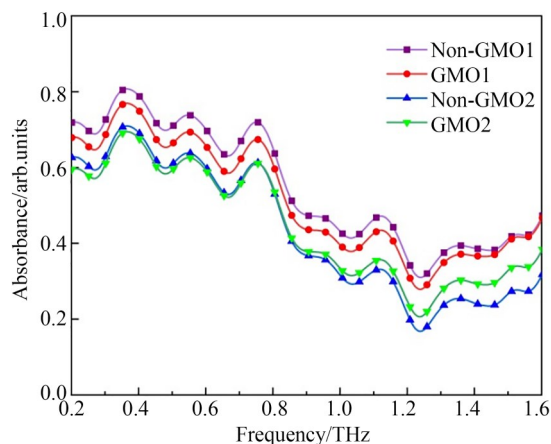


图2 4种菜籽油样品及参考信号的太赫兹时域光谱
Fig. 2 THz time-domain spectra of four rapeseed oils and reference signal



(a) 240个菜籽油样本在0.2-1.6 THz的太赫兹吸收光谱
(a) THz absorption spectra of 240 rapeseed oil samples in 0.2-1.6 THz



(b) 4种菜籽油在0.2-1.6 THz波段内的平均吸收光谱
(b) THz average absorption spectra of four rapeseed oils in 0.2-1.6 THz band

图3 不同种类菜籽油样本在0.2~1.6 THz波段内的吸收光强度

Fig. 3 THz absorption spectra of Different types of rapeseed oil samples in 0.2-1.6 THz band.

4.2 光谱特征提取及融合

图4是采用SPA对240个菜籽油样本的吸收光谱和导数光谱在0.2~1.6 THz频段进行特征提取的结果。通过设置光谱特征选择数量的区间为[10,100],SPA会计算不同数量频率点下的均方根误差(Root Mean Square Error, RMSE),根据RMSE的变化选择RMSE最小值或者不显

著大于RMSE最小值的频率点数。如图4所示,经过迭代后选取吸收光谱30个频率点为最优频率点数,RMSE最小值为0.334 9,选取导数光谱62个频率点为最优频率点数,RMSE最小值为0.494 8。将提取的吸收光谱和导数光谱特征归一化后进行融合,得到的融合光谱作为分类模型的输入,并与单一光谱建立的模型进行对比。

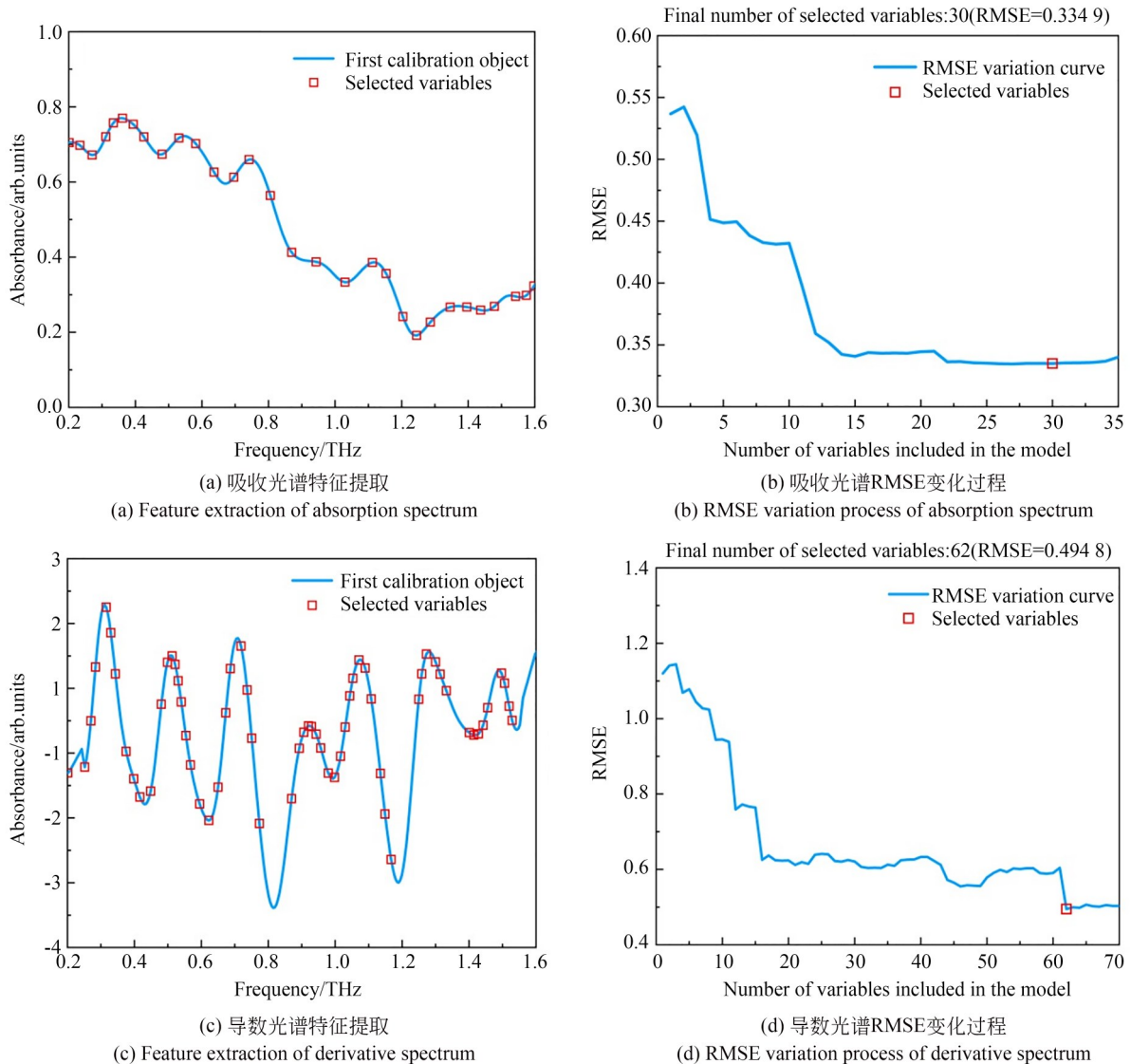


图4 SPA特征提取结果

Fig. 4 SPA feature extraction results

4.3 分类结果

以吸收光谱、融合光谱为输入建立的Fused Lasso模型以及采用贝叶斯优化的BO-Fused Lasso模型对光谱的特征选择如表2所示。由表

2可以看出,Fused Lasso多分类模型的6个二分类器对吸收光谱和融合光谱都有特征选择的作用,如分类器1(C_1)代表Non-GMO1与GMO1二分类,分类器2(C_2)代表Non-GMO1与Non-

GMO2 二分类,分类器 3(C_3)代表 Non-GMO1 与 GMO2 二分类,分类器 4(C_4)代表 Non-GMO2 与 GMO1 二分类,分类器 5(C_5)代表 Non-GMO2 与 GMO2 二分类,分类器 6(C_6)代表 GMO1 与 GMO2 二分类。每个分类器都在最优参数下选择了对分类有用的特征,剔除了部分冗余特征,体现了 Fused Lasso 作为嵌入式特征选择方法的优势。

表 2 各分类模型特征选择结果
Tab. 2 Feature selection results of each classification model

Model	Spectral data	Dimension	Number of variables for feature selection					
			C_1	C_2	C_3	C_4	C_5	C_6
Fused Lasso	Single spectrum	307	277	182	173	125	165	207
	Fusion spectrum	92	44	38	40	32	41	54
BO-Fused Lasso	Single spectrum	307	255	158	160	123	150	182
	Fusion spectrum	92	90	83	82	84	85	92

表 3 给出了各个模型的分类结果,图 5 是各个模型测试集分类结果的混淆矩阵。对于归一化处理后的吸收光谱建立的 Fused Lasso 模型,采用十折交叉验证的方法确定 L1 罚的正则化参数 λ_1 以及融合罚的正则化参数 λ_2 分别为: $\lambda_1=0.071\ 8$, $\lambda_2=0.082\ 5$,训练集准确率为 90.63%,测试集准确率为 87.50%。可见采用吸收光谱建模的测试集准确率不太理想,较多的样本被错误分类,这表明单一光谱所包含的信息较少,不能较好地将光谱相似的转基因和非转基因菜籽油区分开来。采用融合光谱建立的 Fused Lasso 模型,通过十折交叉验证确定正则化参数为: $\lambda_1=0.217\ 6$, $\lambda_2=0.041\ 2$,训练集准确率提高到了 93.75%,测试集准确率提高到了 92.50%,分类准确率较单一吸收光谱有所提高,表明融合光谱所包含的信息较多,更能反映物质的光谱特征。原因是吸收光谱和导数光谱经过 SPA 特征提取后去除了冗余变量和噪声变量,保留了含有被测

物特征信息的变量,而融合后又增加了信息量,对比高维度的吸收光谱数据,融合光谱以较少的变量反映了更多的信息。采用贝叶斯优化的 Fused Lasso 模型在吸收光谱和融合光谱上的准确率都比采用十折交叉验证的准确率要高,并且所需要的运行时间大幅减少。对于单一光谱,贝叶斯优化所确定的参数为: $\lambda_1=0.079\ 8$, $\lambda_2=0.249\ 0$,训练集准确率提高到 91.88%,测试集准确率与采用十折交叉验证相同。对于融合光谱,贝叶斯优化所确定的参数为: $\lambda_1=0.037\ 2$, $\lambda_2=0.032\ 9$,训练集准确率达到 96.88%,测试集准确率达到 95.00%,由图 5 的混淆矩阵也可看出,错误分类的样本明显减少,表明贝叶斯优化算法对参数的寻优能力和效率要优于十折交叉验证,并且适用于计算量较大的 Fused Lasso 多分类模型,证明了 BO-Fused Lasso 模型的准确性和高效性。

表 3 各分类模型结果
Tab. 3 Results of each classification model

Model	Spectral data	Parameters	Training set accuracy/%	Testing set accuracy/%	Running time/s
Fused Lasso	Single spectrum	$\lambda_1=0.071\ 8$, $\lambda_2=0.082\ 5$	90.63	87.50	294
	Fusion spectrum	$\lambda_1=0.217\ 6$, $\lambda_2=0.041\ 2$	93.75	92.50	149
BO-Fused Lasso	Single spectrum	$\lambda_1=0.079\ 8$, $\lambda_2=0.249\ 0$	91.88	87.50	40
	Fusion spectrum	$\lambda_1=0.037\ 2$, $\lambda_2=0.032\ 9$	96.88	95.00	31

True label	Non-GMO1	19	1	0	0
	GMO1	2	16	2	0
	Non-GMO2	0	0	15	5
	GMO2	0	0	0	20
		Non-GMO1	GMO1	Non-GMO2	GMO2
		Prediction label			

(a) 基于吸收光谱的Fused Lasso模型

(a) Fused Lasso model based on absorption spectrum

True label	Non-GMO1	19	1	0	0
	GMO1	2	16	2	0
	Non-GMO2	0	0	15	5
	GMO2	0	0	0	20
		Non-GMO1	GMO1	Non-GMO2	GMO2
		Prediction label			

(b) 基于吸收光谱的BO-Fused Lasso模型

(b) BO-Fused Lasso model based on absorption spectrum

True label	Non-GMO1	19	1	0	0
	GMO1	1	18	2	0
	Non-GMO2	0	0	17	3
	GMO2	0	0	0	20
		Non-GMO1	GMO1	Non-GMO2	GMO2
		Prediction label			

(c) 基于融合光谱的Fused Lasso模型

(c) Fused Lasso model based on fusion spectrum

True label	Non-GMO1	19	1	0	0
	GMO1	1	19	0	0
	Non-GMO2	0	0	18	2
	GMO2	0	0	0	20
		Non-GMO1	GMO1	Non-GMO2	GMO2
		Prediction label			

(d) 基于融合光谱的BO-Fused Lasso模型

(d) BO-Fused Lasso model based on fusion spectrum

图 5 各模型分类结果混淆矩阵

Fig. 5 Confusion matrix of classification results of each model

5 结 论

本文应用 THz-TDS 技术对转基因和非转基因菜籽油进行检测,研究发现转基因和非转基因菜籽油的太赫兹光谱极其相似,为实现转基因和非转基因菜籽油的无损准确鉴别,提出一种太赫兹融合光谱结合 BO-Fused Lasso 模型的鉴别新方法。首先,通过应用连续投影算法(SPA)对转基因和非转基因菜籽油的太赫兹吸收光谱和导数光谱特征进行提取并融合,解决了单一光谱包含物质特征信息不足的问题。然后,引入贝叶斯

优化(BO)算法替代 Fused Lasso 模型的交叉验证过程,解决了其寻参过程繁琐耗时以及可能找不到全局最优解的问题。结果表明,所提出的基于融合光谱的 BO-Fused Lasso 模型分类效果最好,对 4 种菜籽油分类的训练集准确率达到 96.88%,测试集准确率达到 95.00%,优于传统单一吸收光谱 Fused Lasso 模型的 90.63% 和 87.50%。因此,太赫兹融合光谱结合 BO-Fused Lasso 模型为转基因和非转基因菜籽油的分类鉴别提供了一种无损准确的新方法,对其他转基因物质的检测也具有较好的参考价值。

参考文献:

- [1] 李浩辉,刘彩月,张海文,等. 2022年度全球转基因作物产业化发展现状及趋势分析[J]. 中国农业科技导报, 2023, 25(12): 6-16.
LI H H, LIU C Y, ZHANG H W, *et al.* Global genetically modified crop industrialization trends in 2022[J]. *Journal of Agricultural Science and Technology*, 2023, 25(12): 6-16. (in Chinese)
- [2] USLU T. Advantages, risks and legal perspectives of GMOs in 2020s [J]. *Plant Biotechnology Reports*, 2021, 15(6): 741-751.
- [3] YIN J X, XIA L P, ZOU Z Y, *et al.* Multiplex digital polymerase chain reaction technology and its application [J]. *Chinese Journal of Analytical Chemistry*, 2022, 50(1): 25-38.
- [4] WANG Y F, BEDNARCIK M, AMENT C, *et al.* Immunoassays and mass spectrometry for determination of protein concentrations in genetically modified crops [J]. *Journal of Agricultural and Food Chemistry*, 2024, 72(16): 8879-8889.
- [5] 曾子威,金尚忠,李宏光,等. 高湿度环境下爆炸物太赫兹光谱的特征提取与精准识别[J]. 光学精密工程, 2023, 31(7): 1065-1073.
ZENG Z W, JIN S Z, LI H G, *et al.* Terahertz spectral features detection and accuracy identification of explosives in high humidity environment [J]. *Opt. Precision Eng.*, 2023, 31(7): 1065-1073. (in Chinese)
- [6] ZHOU S L, ZHU S P, LIU G H, *et al.* Rapid detection of transgenic soybean oils by terahertz (THz) spectroscopy[J]. *Journal of Nanoelectronics and Optoelectronics*, 2017, 12(9): 956-960.
- [7] LIU J J, FAN L L, LIU Y M, *et al.* Application of terahertz spectroscopy and chemometrics for discrimination of transgenic camellia oil[J]. *Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy*, 2019, 206: 165-169.
- [8] 陈涛,李欣. 太赫兹光谱在转基因菜籽油鉴别中的应用: 基于改进蜉蝣算法的支持向量机模型[J]. 物理学报, 2024, 73(5): 360-368.
CHEN T, LI X. Application of terahertz spectroscopy in identification of transgenic rapeseed oils: a support vector machine model based on modified mayfly optimization algorithm[J]. *Acta Physica Sinica*, 2024, 73(5): 360-368. (in Chinese)
- [9] ZHENG C Y, CAI S J, LI Q, *et al.* A collaborative classification algorithm with multi-view terahertz spectra [J]. *Results in Physics*, 2022, 42: 106023.
- [10] KAMRUZZAMAN M, KALITA D, AHMED M T, *et al.* Effect of variable selection algorithms on model performance for predicting moisture content in biological materials using spectral data[J]. *Analytica Chimica Acta*, 2022, 1202: 339390.
- [11] WEI C J, WANG J F, HE X L, *et al.* A fast and non-destructive approach to identify the heavy mineral oil trace evidence based on spectral fusion treatment and chemometrics [J]. *Microchemical Journal*, 2021, 163: 105924.
- [12] BIN L, YANG J L, WANG Q, *et al.* Detection of *Camellia oleifera* anthracnose based on THz combined with FT-NIR[J]. *Infrared Physics and Technology*, 2023, 133: 104833.
- [13] 殷贤华,陈慧聪,张活. 基于太赫兹光谱数据融合实现多组分橡胶添加剂的定量检测[J]. 中国激光, 2024, 51(5): 3788/CJL230807.
YIN X H, CHEN H C, ZHANG H. Quantitative detection of multi-component rubber additives based on terahertz spectral data fusion[J]. *Chinese Journal of Lasers*, 2024, 51(5): 3788/CJL230807. (in Chinese)
- [14] KANWAL A, MEHMOOD T, BUTT M M. PLS and kernel SVM based hybrid classifier for discriminating FTIR spectrum data with limited sample size [J]. *Chemometrics and Intelligent Laboratory Systems*, 2021, 215: 104365.
- [15] 张绍荣,朱志斌,冯宝,等. 基于组稀疏贝叶斯逻辑回归运动想象脑电信号分类模型的通道选择与分类新算法[J]. 仪器仪表学报, 2019, 40(10): 179-191.
ZHANG S R, ZHU Z B, FENG B, *et al.* New channel selection and classification algorithm based on group sparse Bayesian logistic regression motor imagery EEG signal classification model[J]. *Chinese Journal of Scientific Instrument*, 2019, 40(10): 179-191. (in Chinese)
- [16] 刘建伟,崔立鹏,刘泽宇,等. 正则化稀疏模型[J]. 计算机学报, 2015, 38(7): 1307-1325.
LIU J W, CUI L P, LIU Z Y, *et al.* Survey on the regularized sparse models [J]. *Chinese Journal of Computers*, 2015, 38(7): 1307-1325. (in Chinese)
- [17] HESAMIAN G, JOHANNSEN A, CHUKHROVA N. An explainable fused lasso regression mod-

- el for handling high-dimensional fuzzy data [J]. *Journal of Computational and Applied Mathematics*, 2024, 441: 115721.
- [18] 崔方晓, 李大成, 吴军, 等. 基于 Lasso 方法的污染气体自适应探测算法[J]. *光学学报*, 2019, 39(5): 0530003.
- CUI F X, LI D C, WU J, *et al.* Adaptive feature extraction algorithm based on lasso method for detecting polluted gas [J]. *Acta Optica Sinica*, 2019, 39(5): 0530003. (in Chinese)
- [19] PRABHAVATHY T, ELUMALAI V K, BALAJI E. Hand gesture classification framework leveraging the entropy features from sEMG signals and VMD augmented multi-class SVM[J]. *Expert Systems with Applications*, 2024, 238: 121972.
- [20] BAI L, ZHANG Z T, GUAN H H, *et al.* Rapid and accurate quality evaluation of *Angelicae Sinensis* Radix based on near-infrared spectroscopy and Bayesian optimized LSTM network[J]. *Talanta*, 2024, 275: 126098.

作者简介:



陈 涛(1984—),男,广西桂林人,博士,教授,博士生导师,2013年于西安电子科技大学获得博士学位,主要从事太赫兹科学技术及应用方面的研究。E-mail:tchen@guet.edu.cn

通讯作者:



张绍荣(1987—),男,广西贵港人,博士,副教授,硕士生导师,2021年于桂林电子科技大学获得博士学位,主要从事机器学习算法及应用方面的研究。E-mail:zsrong@guat.edu.cn