文章编号　1004-924X(2026)08-1219-13

硫化促进剂太赫兹光谱数据扩充策略与定量检测

殷贤华[1,2]，李　康[1,2]，孙　傲[1,2]，张富强[1,2]，张　活[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. 广西自动检测技术与仪器重点实验室，广西 桂林 541004)

摘要：为实现橡胶制品中硫化促进剂含量的快速、无损、精准检测，本研究采用太赫兹时域光谱技术，结合数据扩充与化学计量学方法，对多组分橡胶混合物中硫化促进剂进行定量分析。针对橡胶混合物光谱重叠严重、样本量偏小易导致模型过拟合、泛化能力差等问题，提出基于数据融合与最小二乘高斯拟合法(LSGF)的数据扩充策略，并构建遗传算法优化支持向量回归(GA-SVR)定量模型。为降低数据维度、提升建模效率，采用变量空间迭代收缩算法(VISSA)对原始及扩充后光谱进行特征提取。结果表明：数据扩充可显著提升模型预测性能，其中LSGF方法效果最优；经VISSA特征提取后，模型精度进一步提升，LSGF扩充数据在预测集上的相关系数Rp高达0.982 6，RMSEP低至0.002 3。该方法可为橡胶配方优化与行业绿色可持续发展提供技术参考。

关　键　词：太赫兹光谱；数据扩充；最小二乘高斯拟合；数据融合；定量分析
中图分类号：O436　　**文献标识码**：A
doi：10.37188/OPE.20263408.1219　　**CSTR**：32169.14.OPE.20263408.1219

Expansion strategy of terahertz spectral data for vulcanization accelerators and quantitative detection

YIN Xianhua[1,2], LI Kang[1,2], SUN Ao[1,2], ZHANG Fuqiang[1,2], ZHANG Huo[1,2*]

(1. *School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin* 541004, *China*;
2. *Guangxi Key Laboratory of Automatic Detection Technology and Instrument, Guilin* 541004, *China*)
* *Corresponding author, E-mail*: 13117730484@163.*com*

Abstract: To achieve rapid, non-destructive and accurate detection of vulcanization accelerator content in rubber products, this study adopted terahertz time-domain spectroscopy technology, combined with data augmentation and chemometric methods, to conduct quantitative analysis of vulcanization accelerators in multi-component rubber mixtures. Aiming at the problems of serious spectral overlap of rubber mixtures, small sample size which was prone to model overfitting and poor generalization ability, a data augmentation strategy based on data fusion and Least Squares Gaussian Fitting (LSGF) was proposed, and a quantitative model of Genetic Algorithm-optimized Support Vector Regression (GA-SVR) was constructed. To reduce data dimensionality and improve modeling efficiency, the Variable Space Iterative Shrinkage

收稿日期：2025-11-12；**修订日期**：2026-01-06.
基金项目：国家自然科学基金(No. 62161005)；广西自然科学基金项目(No. 2025GXNSFAA069795)；桂林电子科技大学研究生教育创新计划项目(No. 2025YCXS145)

Approach (VISSA) was used to extract features from the original and augmented spectra. The results show that data augmentation can significantly improve the predictive performance of the model. Among them, the LSGF method has the best effect; after VISSA feature extraction, the model accuracy is further improved, and the correlation coefficient Rp of the LSGF-augmented data in the prediction set reaches as high as 0.982 6, with the RMSEP as low as 0.002 3 . This method can provide technical reference for rubber formula optimization and green and sustainable development of the industry.

Key words: terahertz spectroscopy; data augmentation; least squares Gaussian fitting; data fusion; quantitative analysis

1 引　言

橡胶添加剂通过改善工艺加工性能、提升物理性能、增强化学稳定性及赋予特定功能特性，可实现橡胶制品的性能精准调控，满足多元应用需求。其中，硫化促进剂对于提升橡胶制品的加工效率和性能至关重要，现已成为橡胶制品中不可或缺的添加剂之一。然而，随着全球对绿色环保发展的日益重视，硫化促进剂的污染和毒害性成为人们关注的焦点。许多促进剂含有有毒化学物质，其生产和使用过程中会释放有害气体和废水，对生态环境和人体健康造成严重威胁[1]。近年来，因促进剂毒性引发的事故屡见不鲜。2018年，《橡胶制品加工业职业危害和预防》报告中提及某橡胶制品厂因使用含有二硫化四甲基秋兰姆的促进剂，导致多名工人出现呼吸系统损伤和皮肤过敏；2019年，据河北宁晋生态环境局发布，当地轮胎厂因促进剂泄漏造成周边水体污染，导致鱼类大量死亡，引发公众环保抗议。这些事故不仅暴露了硫化促进剂在安全管理方面的不足，也凸显了其在环保和健康方面的潜在风险。因此，加强橡胶制品中硫化促进剂含量的精准检测，确保其用量符合环保和安全标准，已成为推动橡胶工业实现可持续发展的关键举措。传统检测方法如化学分析[2]、色谱法[3]和红外光谱法[4]虽能准确测定硫化促进剂，但存在操作繁琐、耗时长、成本高、灵敏度不足以及实验过程中会产生废气、废液会对实验员的身体健康造成威胁等缺陷[5]，限制了我国橡胶工业的绿色可持续发展。因此，研究更加高效、环保的检测方法已成为行业的迫切需求，推动检测技术创新刻不容缓。

近年来，随着超快激光以及半导体技术的发展，太赫兹光谱技术在物质检测领域取得了显著突破。太赫兹光谱具有瞬态性、低能性、安全性、“指纹”特性[6-8]等独特的性能，此外，太赫兹光的穿透力更强，相比于其他光谱更能够表征物质的内部结构以及化学成分。太赫兹光谱技术凭借其独特的核心特性，在橡胶添加剂含量检测领域的应用正越来越广泛。例如Hirakawa等[9]人开发了一种基于太赫兹光谱技术的橡胶硫化过程及填料分散性可视化评估方法，通过光谱分析实现了对橡胶硫化反应与白炭黑宏观分散性的有效表征，为橡胶制品的质量检测提供了新的技术途径。Zhang R等[10]人采用太赫兹时域光谱技术实现了橡胶中白炭黑含量的无损测量，通过分析太赫兹信号的吸收和散射特性，建立了一种高精度的定量检测方法。Chen M等[11]人研究了太赫兹光谱技术在橡胶中氧化锌含量检测中的应用，通过光谱特征与氧化锌含量的相关性分析，提出了一种快速、可靠的检测方法。上述研究为太赫兹光谱技术在添加剂含量检测方面提供了坚实的理论支持。

目前，在采用太赫兹光谱技术进行混合物定量分析的研究中，由于光谱数据获取方式耗时费力，很难通过实验的方式来获取大量的光谱数据，数据普遍呈小样本特性[12-14]。因此，导致了后续模型分析时过拟合[15]、泛化差及预测不稳定等问题。鉴于此，本研究利用机器学习算法强大的数据处理能力，结合数据扩充策略对光谱数据的数量和维度进行优化，以提升模型分析的准确性和效率。

2　太赫兹光谱实验

2.1　实验设备

实验数据采集装置采用由深圳市华讯方舟科技有限公司生产的CCT-1800光谱仪，其实物图和原理图分别如图1～图2所示。CCT-1800光谱仪光谱范围0.05～5 THz，激光器的脉冲重复频率为80 MHz，脉冲中心波长为780 nm，脉冲宽度小于100 fs。该光谱包含太赫兹辐射发生器、太赫兹辐射探测器、延时装置和飞秒脉冲激光器。

图1　太赫兹光谱仪实物图

Fig. 1　Physical diagram of the terahertz spectrometer

2.2　样品制备

本研究以丁腈橡胶(NBR)、白炭黑(Silica)、防老剂N-苯基-N′-(1,3-二甲基丁基)-对苯二胺(44S)、含毒性的硫化促进剂亚乙基硫脲(ETU)以及无毒性的硫化促进剂二硫化四苄基秋兰姆(TBzTD)组成的混合物为研究对象。五组分混合物中各物质配比如表1所示。

表1　五组分混合物中各物质配比

Tab. 1　Ratio of each substance in the five-component mixture　(%)

Number	Mass fraction					
	NBR	Polyethylene	Silica	44S	TBzTD	ETU
1	60	5	20	5	0	10
2	60	5	20	5	2	8
3	60	5	20	5	4	6
4	60	5	20	5	6	4
5	60	5	20	5	8	2
6	60	5	20	5	10	0

样品的制备方式采用压片法，首先使用高速粉碎机分别将各种待混合物质粉碎，再使用200目筛网筛选特定粒子大小的待混合物质粉末，各

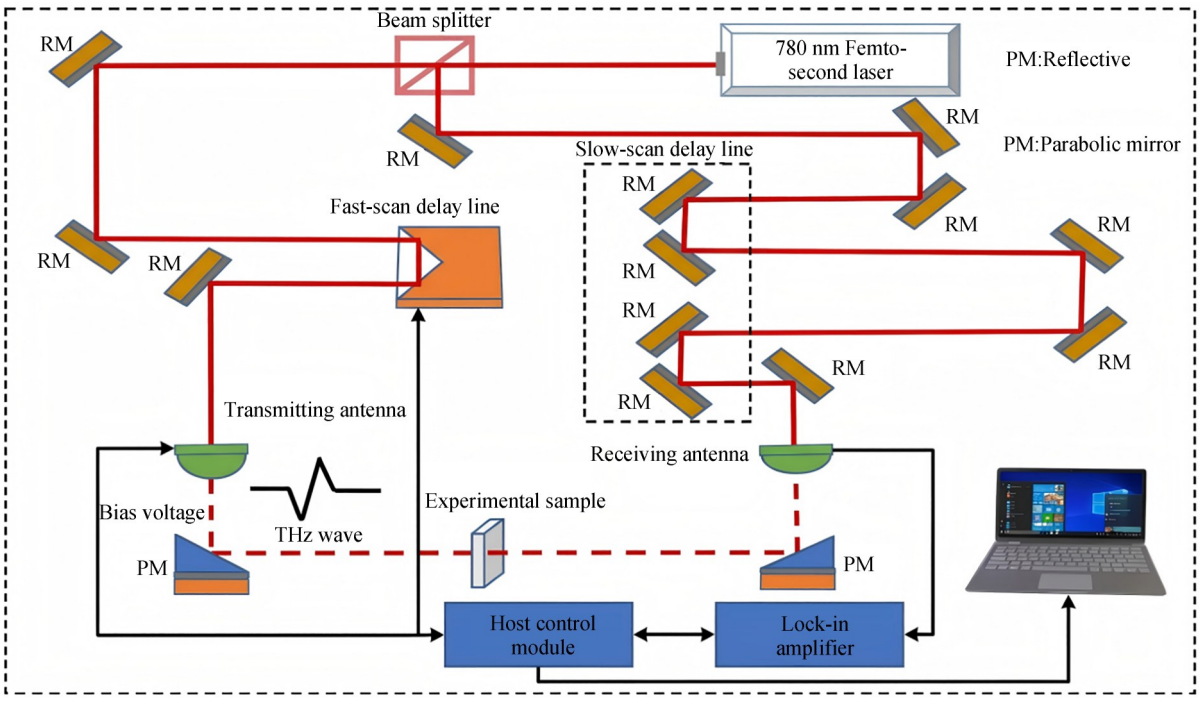


图2　太赫兹光谱仪原理图

Fig. 2　Schematic diagram of the terahertz spectrometer

物质严格按照表1的比例进行称量，称量完毕后将粉末倒入陶瓷研钵研磨直至各种物质粉末充分混合，然后倒入模具开始制样，最终制成表面光滑、直径13 mm以及厚度1 mm的圆形样片，最后将制成的样片放入恒温干燥箱干燥1.5 h后待测。由于NBR和白炭黑压片较为困难，所以在五种物质以及混合物压片时都加入5%的聚乙烯。考虑到后续所建模型的长时间鲁棒性及算法的实际适用性验证，为其在工业场景中的推广应用提供可靠的实验支撑，样品是在跨时一个月、在随机时段、多批次的条件下，严格遵循相同的实验流程与操作规范制备完成。五种纯物质及六种比例的混合物各制备36个样品，最终共制备396个样品。

2.3 光谱获取

大气中的水蒸气对于太赫兹波会强烈吸收[16]，所以在实验光谱采集时，为减少空气中水分的干扰，该系统设置了一个样品仓，将被采集样品置于密封盒中。实验过程中持续向密封盒充入干燥氮气，使相对湿度保持在3%以下。实验过程中，为避免随机误差以每个样品正反面各测3次取平均作为该样品的光谱数据。最终，通过太赫兹光谱实验采集到五种物质的时域光谱数据共180条和六种比例的五组分混合物的时域光谱数据共216条。

3 基本原理

3.1 光学参数提取

本文将表示样品对太赫兹波吸收强弱的太赫兹吸光度作为光谱分析的主要光学参数。为了计算样品的太赫兹吸光度，需要测量THz-TDS系统空载时的太赫兹时域光谱作为参考信号$E_{\mathrm{ref}}(t)$以及被测样品置于THz-TDS系统时的太赫兹时域光谱$E_{\mathrm{sam}}(t)$。通过快速傅里叶变换将时域信号$E_{\mathrm{ref}}(t)$和$E_{\mathrm{sam}}(t)$转换成频域信号的$E_{\mathrm{ref}}(w)$和$E_{\mathrm{sam}}(w)$，然后根据Dorney. T. D和Duvillaret. L等[17-18]人提出的光学参数提取模型来获取吸光度参数。吸光度计算公式见式(1)：

$$A=-\log_{10}\left|\frac{E_{\mathrm{sam}}(w)}{E_{\mathrm{ref}}(w)}\right|^{2}, \tag{1}$$

其中：$E_{\mathrm{sam}}(w)$为样品的太赫兹频域谱，$E_{\mathrm{ref}}(w)$为参考信号的太赫兹频域谱。

3.2 数据扩充

在多组分混合物的定量研究中，光谱数据普遍呈小样本特性。为解决小样本导致的过拟合与泛化能力差的问题[18-19]，本研究提出最小二乘高斯拟合法与改进的数据融合策略增加样本多样性。

3.2.1 最小二乘高斯拟合法扩充数据

最小二乘法(Least Squares Method, LSM)是一种能够拟合非线性数据以及强鲁棒性的数学优化技术[20]。本研究提出最小二乘高斯拟合法(Least Squares Gaussian Fitting, LSGF)对吸光度光谱进行1:1拟合的策略，最终将原始光谱数据样本量扩充到了1.5倍。为减小拟合偏差对数据分析的影响，拟合所得的两条曲线取平均作为新样本数据。拟合的一条光谱与原始光谱如图3所示。

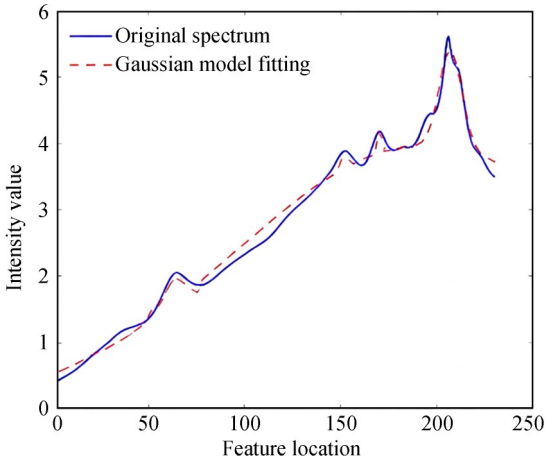


图3 原始光谱与拟合光谱

Fig. 3 Original spectrum and fitted spectrum

通过最小二乘法拟合光谱可能会引入系统误差，所以分析拟合光谱的真实性是有必要的[21]。如图3所示，高斯模型拟合光谱与原始光谱的吸收峰特征位置和强度高度吻合，证明该方法在扩充数据的同时未引入虚假特征，保证了数据的真实性与物理意义。

3.2.2 数据融合法扩充数据

数据融合(Data fusion)是指将不同来源或不同形式的数据通过特定的方式进行整合，从而生成具有更高价值的新数据的一种数据处理方法，

它不仅能够提升数据的完整性和准确性，还能够挖掘出单一数据源无法揭示的潜在信息。在本研究中，采用Savitzky-Golay算法（SG）对原始光谱进行数据预处理。SG平滑可在减少太赫兹光谱噪声的同时保留硫化促进剂核心光谱特征，处理后的数据其维度与原始吸光度光谱数据完全一致，这为数据融合提供了便利。基于这一特性，借鉴低层数据融合[22]的思想，本研究提出了一种新的数据融合方式：将原始吸光度光谱数据与SG数据预处理后的吸光度光谱数据进行融合，以达到扩充样本量的目的。这种融合方式不仅保留了原始数据的特征，还通过引入处理后的数据特征进一步丰富了数据内容，提升了信息表征可靠性与区分度、缓解小样本过拟合问题。具体融合方法如图4所示。通过采用这种融合方式，原始光谱数据样本量被扩充到了2倍，这为后续的分析提供了更加丰富的数据基础。

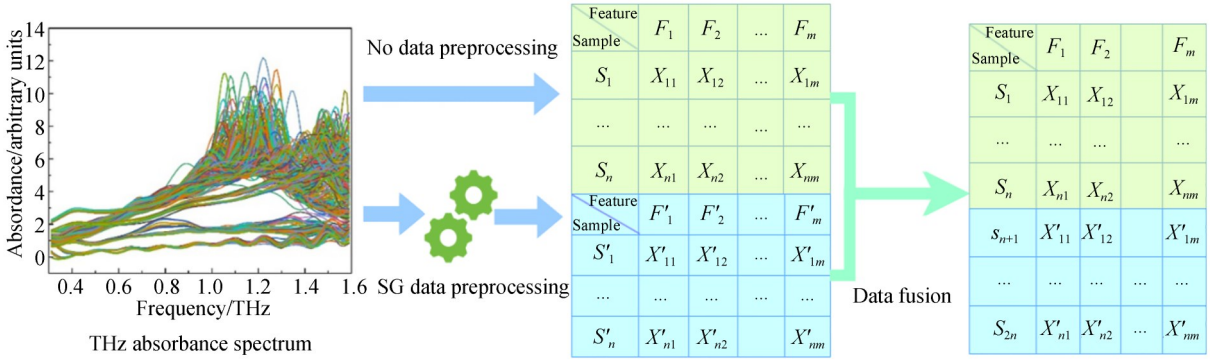


图4　数据融合法扩充数据

Fig. 4　Data fusion method for expanding data

需要说明的是，本研究在数据融合过程中已充分考虑实际工业环境中的噪声干扰问题：原始吸光度光谱数据本身和SG平滑预处理后的数据已包含实验过程中不可避免的随机噪声（如太赫兹光谱仪探测器响应波动、环境湿度微小波动、样品制备微观差异带来的噪声等），这些噪声均为实际工业场景中太赫兹光谱检测时会遇到的典型干扰，在数据融合过程中完整保留了原始光谱中的随机噪声信息，确保融合后的数据集能够真实模拟实际工业环境下的光谱特征。

3.3　数据集划分

本研究获取的太赫兹光谱经光学参数提取后，对于五种物质的光谱数据仅需分析其幅值、相位以及吸收峰等相关参数，而无需构建分类模型对其分析，所以这部分数据不进行数据划分。为了保证模型在训练时能够学习到样品数据中的大部分关键信息，对于混合物光谱数据使用Kennard-Stone（KS）算法[23]以2∶1的比例将数据划分为校准集和测试集。校正集用于模型的训练的输入，测试集不参与模型训练，只用于训练好的模型进行性能测试，这样才能真实反映模型对于未知样本的预测能力。KS算法的核心是两个不同样品欧氏距离的计算，欧氏距离的计算公式见式（2）：

$$d_x(p,q)=\sqrt{\sum_{j=1}^{N}\left[x_p(j)-x_q(j)\right]^2};\ p,q\in[1,N], \tag{2}$$

其中：x_p和x_q表示两个不同的样品，N表示样品总数。

3.4　支持向量回归

支持向量回归（Support Vector Regression，SVR）是SVM在回归问题中的扩展[24]，具有优异的非线性拟合与小样本学习能力，适用于高维光谱数据建模。然而，SVR的有效性却高度依赖于惩罚系数C和核函数宽度γ这两个超参数的适当选择[25]。为解决SVR模型中惩罚系数C和核函数宽度γ难以确定的问题，本研究利用遗传算法（Genetic Algorithm，GA）的全局搜索能力实现超参数的自适应寻优，构建了GA优化的SVR模型（GA-SVR）。该方法的科学性源于全局搜索

能力、自适应优化特性与泛化性能导向的协同作用，构建了完整的优化体系，有效避免人工调参主观性，保障参数选择的可靠性与模型泛化能力。

3.5 特征提取

为解决数据扩充带来的维度灾难及计算效率降低问题，需对光谱特征进行筛选。变量空间迭代收缩算法（Variable Iterative Space Shrinkage Approach，VISSA）能够通过迭代收缩变量空间，逐步剔除无用信息或干扰噪声，从而在降低维度的同时提升模型的预测精度。本研究利用VISSA分别提取三组数据的最优特征子集，具体流程包括标准化处理、加权采样及基于模型集群分析（Model Population Analysis，MPA）的变量重要性评估。

使用VISSA算法分别对原始光谱数据、数据融合法扩充数据以及最小二乘拟合法扩充数据特征提取后的数据特征重要性分布见图5所示。这三组数据特征维度一致，但由于数据量不同，各特征的重要性也发生了变化，因此提取到的特征数也产生了变化。在设置相同重要性筛选阈值的前提下，分别在原始光谱数据、数据融合法扩充数据以及最小二乘拟合法扩充数据提取了66个、91个和54个特征，对应数据维度分别降低了71.30%，60.43%和76.52%。

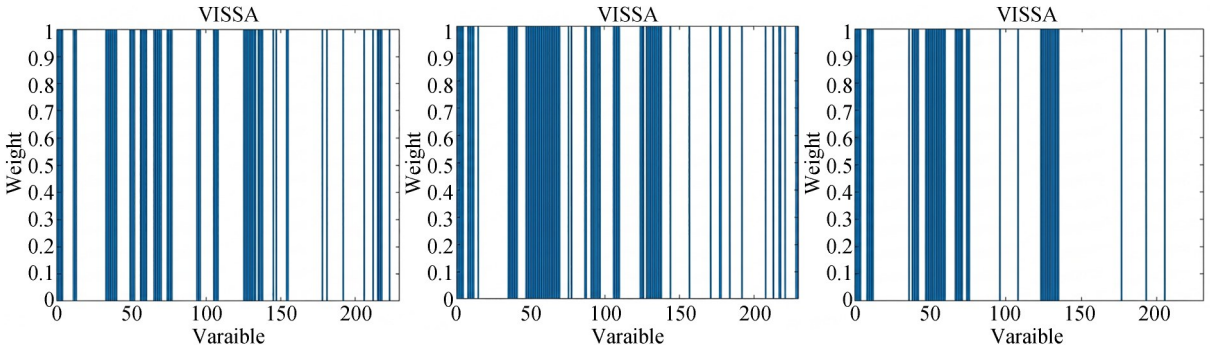


图5　数据特征重要性分布图

Fig. 5　Importance distribution of data features

4 分析与讨论

4.1 五种物质光谱分析

实验采集到五种物质的时域光谱数据共180条，为了更清晰地分析各物质样品的时域光谱中所包含的信息，分别对每一种物质的光谱采取平均操作后，绘制五种物质的平均时域光谱如图6所示。

图中样品信号的光谱与参考信号相比，在幅值上表现为减弱，这是实验样品对太赫兹波的吸收性以及菲涅尔透射系数[26]造成的；在相位上表现为滞后，这是实验样品对太赫兹波的折射反应[27]造成的。在五种物质样品制备以及数据采集环境一致的前提下，根据Beer-Lambert定律[28]

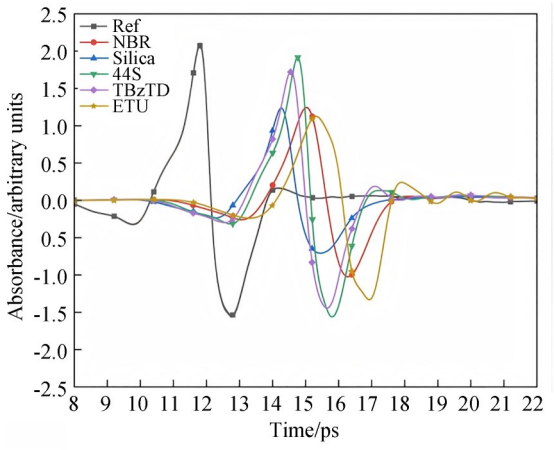


图6　五种物质的平均时域光谱图

Fig. 6　Average time-domain spectra of the five substances

可知，不同物质对太赫兹波的吸收程度的影响因子中，太赫兹时域光谱幅值的波动影响占主导地位[29]。五种物质的平均时域光谱图显示，相对于参考信号，五种物质的幅值减小且相位滞后。在同一参考信号下，太赫兹波透射防老剂 44S 样品后第一个波峰的峰值最大，透射防老剂 ETU 样品后第一个波峰的峰值最小。这表明除了样品材料不同之外，采集环境等其他条件对所有样品的影响均一致的前提下，可以推断出促进剂 ETU 对太赫兹波的吸收程度最大，防老剂 44S 样品对太赫兹波的吸收程度最小。同理，通过比较其他三种物质的平均时域光谱的第一个峰值也可推断出它们对太赫兹波的吸收程度。

为了进一步分析这五种物质对太赫兹波的吸收情况，通过光学参数提取方法获取了它们的吸光度光谱。每种物质的 36 条光谱采取平均处理，以便更直观地分析每种物质对太赫兹波的吸收情况。通过 CCT-1800 光谱仪器可以获取到 0～2.5 THz 的数据，但是由于 THz-TDS 系统的性能限制，超出 2.0 THz 范围的样品对太赫兹波的吸收趋于饱和状态，其中包含很多噪声干扰。基于对这种情况的考虑，选择 0.2～2.0 THz 范围内信噪比相对较高的数据。五种物质样品的平均吸光度光谱如图 7 所示。

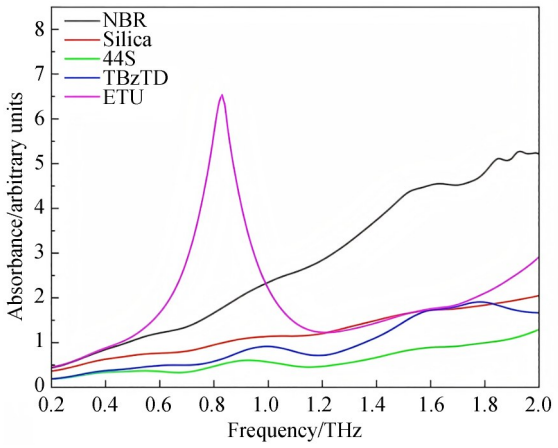


图 7　五种物质的平均吸光度光谱图

Fig. 7　Average absorbance spectra of the five substances

通过对平均吸光度光谱图的分析可以得知，丁腈橡胶样品在 0.2～2.0 THz 范围内有多个微弱的吸收峰，且在 1.5 THz 处有一个明显的吸收峰；白炭黑样品在 0.5 THz，0.9 THz 以及 1.5 THz 处分别有一个微弱的吸收峰；防老剂 44S 样品在 0.5 THz 和 0.9 THz 处分别有一个较为明显的吸收峰；硫化促进剂 TBzTD 样品在 0.6 THz，1.0 THz，1.6THz 以及 1.8 THz 处分别有一个吸收峰。硫化促进剂 ETU 样品在 0.8 THz 处有一个明显的吸收峰，在 1.5 THz 处有一个微弱的吸收峰。此外，0.2～1.0 THz 范围内硫化促进剂 ETU 样品吸收谱的值均大于其他样品的吸收谱的值，防老剂 44S 样品吸收谱的值均小于其他样品吸收谱的值，这与时域光谱图中所反映出来的信息一致。

4.2　混合物光谱分析

在橡胶及添加剂混合物实验中共采集到六种比例的混合物的时域光谱数据共 216 条，为了更准确地分析不同比例混合物样本的时域光谱中所包含的信息，同样分别对每一种比例混合物的光谱采取取平均操作，绘制五组分混合物样品的平均时域光谱如图 8 所示。相比于参考信号，混合物样品信号的幅值减小且相位滞后，这与五种物质的时域光谱分析结果相吻合。通过比较五种物质的平均时域光谱图和混合物样品的平均时域光谱可知，五种物质的第一个波峰都在 1.0～2.0 a.u. 范围内，混合物样品的第一个波峰都在 1.2～1.4 a.u. 范围内，混合物样品的波峰范围减小，这是五种物质混合形成的合理结果。进一步对混合物样品的平均时域光谱分析，随着混合物中硫化促进剂 ETU 对硫化促进剂 TBzTD 比例的减少，第一个波峰的峰值总体呈线性

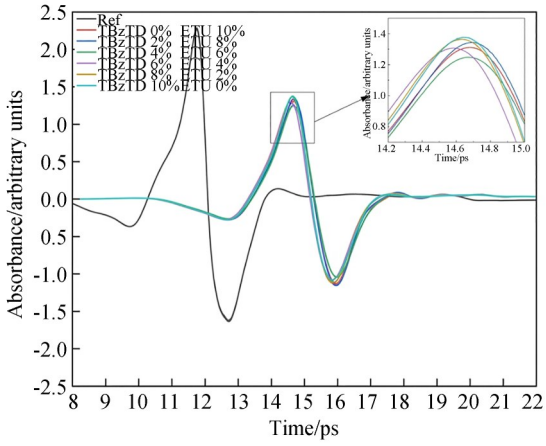


图 8　五组分混合物样品的平均时域光谱图

Fig. 8　Average time-domain spectroscopyof five component mixture sample

增长，这与五种物质的平均时域光谱分析结果相吻合。

通过分析五种物质及其混合物的时域光谱，能够得知不同比例混合物在时域光谱上信息的变化与各物质含量息息相关。为了进一步探索它们在吸光度光谱上的联系，通过光学参数提取方法获取了不同比例混合物在0.2～2.0 THz范围的吸光度光谱。获取的六种比例五组分混合物样品的平均吸光度光谱图如图9所示。

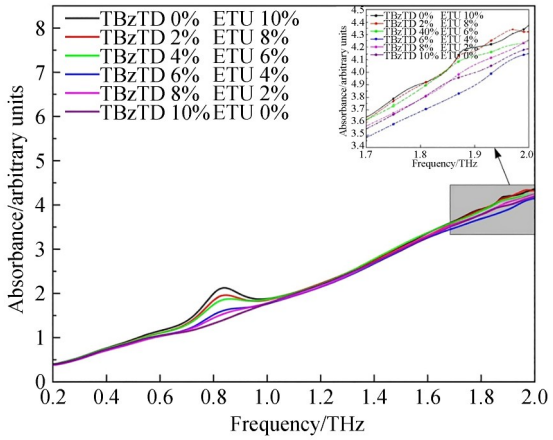


图9　五组分混合物样品的平均吸光度光谱图

Fig. 9　Average absorbance spectrum of a five component mixture sample

通过对平均吸光度光谱图的分析可以得知，混合物样品在0.8～0.9 THz出现的吸收峰的峰值随着混合物中硫化促进剂ETU含量的减少而减小，硫化促进剂ETU含量为零时混合物在此范围内的吸收峰也随之消失，可以进一步推断出混合物在该范围出现的吸收峰是由硫化促进剂ETU产生的。另外，通过图9可以看出混合物光谱重叠严重导致肉眼难以分辨，这是由于物质本身的太赫兹吸收特性、多组分相互作用、以及太赫兹波段物理机制共同导致的，对此需要借助机器学习算法构建定量模型以实现对混合物中两种硫化促进剂含量的检测。

4.3　结果分析

为了直观评估模型性能，采用结果预测图、均方根误差（Root Mean Square Error，RMSE）、相关系数（Correlation Coefficient，通常用R表示）及模型超参数来展示模型的预测结果。结果预测图中散点越接近对角线（$y=x$），表明预测值与真实值越吻合，模型精度越高。RMSE的数值越接近0，R越接近1，说明模型的拟合效果越好。需要说明的是，以下所示结果是使用校正集训练好模型并经过5次交叉验证后，以测试集作为模型输入，进行30次测试取平均得到的最终结果。

为了最合适的定量分析模型，本研究在LS-GF扩充的数据集下，对偏最小二乘回归（Partial Least Squares Regression，PLSR）、随机森林（Random Forest，RF）及反向传播神经网络（Back Propagation Neural Network，BP-NN）三种常见模型和GA-SVR模型进行了对比实验。不同模型性能对比如表2所示。结果表明，PLSR作为线性模型难以处理太赫兹光谱的复杂非线性特征，预测精度最低；RF和BPNN虽然有非线性处理能力，但在小样本高维数据下易陷入局部最优或过拟合。相比之下，GA-SVR结合了结构风险最小化原则与全局寻优能力，在Rc和Rp指标上均表现最佳。最终选取GA-SVR模型作为后续分析的基础模型。

表2　不同模型性能对比

Tab. 2　Comparison of performance of different models

Model	Data augmentation	Training set		Testing set	
		Rc	RMSEC	Rp	RMSEP
PLSR	LSGF	0.921 5	0.008 5	0.895 4	0.013 2
RF	LSGF	0.954 2	0.005 1	0.932 0	0.008 9
BP-NN	LSGF	0.961 0	0.004 8	0.941 5	0.007 5
GA-SVR	LSGF	0.973 5	0.003 7	0.975 2	0.003 1

分别以原始光谱数据、数据融合法和最小二乘高斯拟合法扩充的数据分析基础，建立了GA-SVR模型。特征提取前的GA-SVR模型结果预测图如图10所示，特征提取前GA-SVR模型定量分析相关参数如表3所示。

图10显示，使用最小二乘高斯拟合法（Least Squares Gaussian Fitting，LSGF）扩充数据的GA-SVR模型精度最高，数据融合法扩充数据GA-SVR模型精度次之，使用原始光谱数据的GA-SVR模型精度最差。另外，由于多组分橡胶混合物光谱重叠严重，三种光谱模型在同一样品不同次测量的预测值存在不同程度的波动。原始光谱数据由于小样本特性导致预测结果波动最大，数据融合法扩充的数据次之，最小二乘高斯拟合法扩充的数据最小。表明了数据融合法和最小二乘高斯拟合法两种数据扩充策略在解决多组分橡胶混合物光谱重叠严重、数据小样本特性等问题上均具备有效性。结合表3参数进一步分析可知：GA-SVR模型对原始光谱数据分析时，测试集的Rp和RMSEP分别为0.916 0和0.011 7，表明模型通过分析太赫兹吸光度光谱可捕捉混合物中两种硫化促进剂含量的微弱变化，实现含量区分；对数据融合法扩充数据建模时，测试集Rp提升至0.950 1（较原始数据提升0.034 1），RMSEP降至0.006 1（降低0.005 6），验证了数据融合法可增强光谱信息表征能力；对LSGF扩充数据建模时，测试集Rp达0.975 2（较原始数据、数据融合法分别提升0.059 2和0.025 1），RMSEP降至0.003 1（分别降低0.008 6和0.00 3），表明LSGF扩充数据的样品信息表征效果最优。进一步印证了数据融合法和最小二乘高斯拟合法两种数据扩充策略在解决多组分橡胶混合物光谱重叠严重、数据小样本特性等问题上具备优异效果。

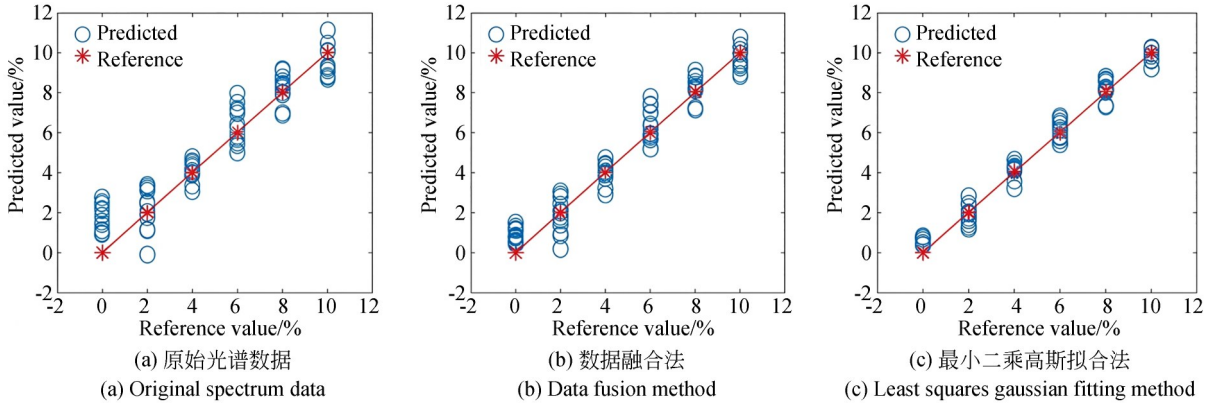


图10　特征提取前的GA-SVR模型结果预测图

Fig. 10　Prediction graph of GA-SVR model results before feature extraction

表3　特征提取前模型定量分析相关参数

Tab. 3　Parameters related to the quantitative analysis of the model before feature extraction

Model	Data augmentation	Training set		Testing set		Model parameters	
		Rc	RMSEC	Rp	RMSEP	C	γ
GA-SVR	Original spectrum	0.973 9	0.004 9	0.916 0	0.011 7	1.866 7	0.502 0
	Data fusion	0.973 4	0.004 2	0.950 1	0.006 1	1.835 3	0.337 3
	LSGF	0.973 5	0.003 7	0.975 2	0.003 1	1.827 5	0.219 6

虽然使用数据扩充策略后模型解决了多组分橡胶混合物光谱重叠严重和小样本特性的问题，但是数据量的提升会使维度和无用特征急剧增加，导致模型运行时间增长、精度下降等问题。所以要使用特征提取算法获取更高质量的数据。为了选取最优的特征提取算法，基于LSGF方法扩充的数据集和GA-SVR模型，使用VISSA算法、主成分分析（Principal Component Analysis，

PCA)及皮尔逊相关系数法(Pearson Correlation Coefficient,PCC)对全光谱进行特征提取,进行了对比实验。不同特征提取算法模型定量分析效果对比如表4所示。结果表明,VISAA各项指标均优于PCA和PCC方法,分析其原因如下:PCA作为一种无监督降维方法,虽然能够有效去除数据间的相关性并降低维度,但其主要保留方差较大的主成分,容易忽略方差较小但包含重要化学信息的特征波段,导致预测精度提升有限。PCC通过计算光谱特征与浓度之间的线性相关性进行筛选,虽然保留了相关性较高的波段,但忽略了光谱变量之间的多重共线性及组合效应。相比之下,VISSA算法作为一种基于模型集群分析的有监督变量选择方法,能够通过迭代收缩变量空间,精准剔除与待测目标变量无关的冗余信息及噪声干扰,同时保留对模型贡献度高的特征组合。这充分证明了VISSA算法在处理太赫兹光谱高维小样本数据时的有效性与优越性。

表4　不同特征提取算法模型定量分析效果对比

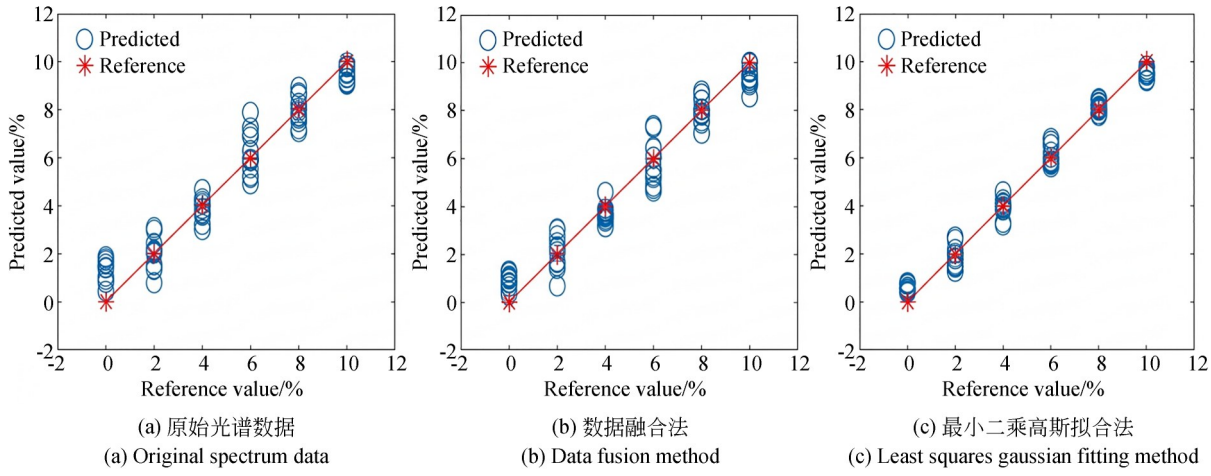
Tab. 4　Quantitative analysis effect comparison of different feature extraction algorithm models

Model	Feature extraction method	Number of features	Training set		Testing set	
			Rc	RMSEC	Rp	RMSEP
GA-SVR	Full spectrum	230	0.979 8	0.002 6	0.975 2	0.003 1
	PCA	12	0.980 5	0.002 5	0.976 8	0.002 9
	PCC	65	0.981 2	0.002 4	0.978 5	0.002 7
	VISSA	54	0.979 8	0.002 6	0.982 6	0.002 3

为了解决数据量的提升带来的问题,最终采用VISSA算法分别对特征提取之后的原始光谱数据、数据融合法扩充的数据以及最小二乘高斯拟合法扩充的数据进行特征提取,构建了GA-SVR模型进行定量分析。特征提取后的模型结果预测图如图11所示,特征提取后模型定量分析相关参数如表5所示。

图11结果表明,与特征提取前的模型结果相比,使用VISSA算法进行特征提取后,GA-SVM模型对三组数据的模型定量分析效果均得到了一定程度的提升,表明VISSA算法在降低数据维度的同时,能有效减少冗余信息,获取更高质量的数据。

进一步通过对比分析表3和表5发现,使用VISSA算法进行特征提取的GA-SVR模型在分析原始光谱数据时,测试集上的参数Rp提升幅度最大,RMSEP降低幅度最大,分析数据融合法和最小二乘高斯拟合法扩充的数据时,测试集上



(a) 原始光谱数据
(a) Original spectrum data

(b) 数据融合法
(b) Data fusion method

(c) 最小二乘高斯拟合法
(c) Least squares gaussian fitting method

图11　特征提取后的模型结果预测图

Fig. 11　Prediction graph of model results after feature extraction

表 5　特征提取后模型定量分析相关参数

Tab. 5　Parameters related to the quantitative analysis of the model after feature extraction

Model	Data augmentation	Training set		Testing set		Model parameters	
		Rc	RMSEC	Rp	RMSEP	*C*	γ
GA-SVR	Original spectrum	0.976 4	0.003 6	0.952 5	0.006 0	1.756 9	0.925 5
	Data fusion	0.979 1	0.003 7	0.960 9	0.005 4	2.001 2	0.975 6
	LSGF	0.979 8	0.002 6	0.982 6	0.002 3	77.472 8	0.250 2

的参数Rp和RMSEP变化幅度较小,表明了原始光谱数据中含有较多负特征或者无用特征,影响了模型定量分析的效果,同时也表明了VISSA算法可以在一定程度上削弱负特征或者无用特征的影响。另外,无论特征提取前后,GA-SVM模型在分析最小二乘高斯拟合法扩充的数据上取得的效果均为最优,表明了最小二乘高斯拟合法在扩充样本的同时,能有效抑制无关变量干扰,提升数据质量与模型稳定性。最终,在所有模型结果的对比中,使用最小二乘高斯拟合法数据扩充策略和VISSA特征提取算法建立的GA-SVR模型在测试集上的表现最好,Rp和RMSEP分别为0.982 6和0.002 3。

5　结　论

本研究针对多组分橡胶混合物光谱重叠严重、小样本特性易导致模型过拟合、泛化能力差等问题,提出了一种基于太赫兹时域光谱技术结合数据扩充策略与化学计量学的定量分析方法。研究表明,利用最小二乘高斯拟合法(Least Squares Gaussian Fitting,LSGF)数据扩充策略在解决扩充数据量同时,能有效抑制无关变量干扰,提升数据质量,其模型分析效果优于数据融合法数据扩充策略和原始光谱;使用变量空间迭代收缩算法(Variable Iterative Space Shrinkage Approach,VISSA)进行特征提取,可有效减少多组分橡胶混合物太赫兹光谱中的冗余信息、解决高维数据带来的问题,显著提升建模效率与精度。最终,基于LSGF扩充数据策略与VISSA特征提取算法构建的GA-SVR模型,在预测集上相关系数Rp和均方根误差RMSEP表现最优,分别为0.982 6和0.002 3,实现了多组分橡胶混合物中硫化促进剂的快速、无损、精准的检测,为橡胶配方优化及行业绿色可持续发展提供了可靠的技术支撑。

作者贡献声明:

殷贤华:开展橡胶硫化促进剂混合物检测方法调研,提出将数据融合策略用于太赫兹光谱数据处理与扩充,指导实验样品的制备、太赫兹光谱实验和定量模型的建模方法,统筹实验资源和研究经费,监督研究全过程,进行论文审阅与修改;

李康:参与实验样品制备与太赫兹光谱实验,实现数据融合法和最小二乘高斯拟合法扩充数据,进行模型建立与模型性能验证,进行光谱分析与结果分析,负责论文初稿撰写与修改;

孙傲:参与实验样品制备与太赫兹光谱实验,进行光谱预处理,负责不同特征提取算法对比实验;

张富强:参与实验样品制备与太赫兹光谱实验,参与光谱分析与结果分析,负责不同模型性能对比实验,参与论文初稿写作;

张活(通讯作者):负责实验结果准确性与可靠性验证,进行论文审阅与修改。

参考文献:

[1] 李龙飞,摆音娜,雷鸣,等. 橡胶硫化促进剂的研究进展[J]. 化学进展,2015,27(10):1500-1508.
LI L F, BAI Y N, LEI M, *et al*. Progress in rubber vulcanization accelerator[J]. *Progress in Chemistry*, 2015, 27(10): 1500-1508. (in Chinese)

[2] 延威,黄德雄,周丹青. 气相色谱-质谱联用法鉴定硫化胶中促进剂[J]. 橡胶工业,2018,65(2):223-226.
YAN W, HUANG D X, ZHOU D Q. Identification of accelerators in vulcanizates by gas chromatog-

raphy-mass spectrometry[J]. *China Rubber Industry*, 2018, 65(2): 223-226. (in Chinese)

[3] 李秋迎, 郭利娟, 薄晓文, 等. 高效液相色谱法测定卤化丁基橡胶塞中2-巯基苯并噻唑残留量[J]. 化学分析计量, 2023, 32(4): 20-23.
LI Q Y, GUO L J, BO X W, *et al*. Determination of 2-mercaptobenzothiazole residues in halogenated butyl rubber plugs by high performance liquid chromatography[J]. *Chemical Analysis and Meterage*, 2023, 32(4): 20-23. (in Chinese)

[4] ROLERE S, LIENGPRAYOON S, VAYSSE L, *et al*. Investigating natural rubber composition with Fourier Transform Infrared (FT-IR) spectroscopy: a rapid and non-destructive method to determine both protein and lipid contents simultaneously[J]. *Polymer Testing*, 2015, 43: 83-93.

[5] LIU Y, ZHANG M. Challenges and limitations of traditional analytical methods for detecting vulcanization accelerators in rubber: A comprehensive review[J]. *Talanta*, 2023, 259: 124601.

[6] TONOUCHI M. Cutting-edge terahertz technology[J]. *Nature Photonics*, 2007, 1(2): 97-105.

[7] JEPSEN P U, COOKE D G, KOCH M. Terahertz spectroscopy and imaging-Modern techniques and applications[J]. *Laser & Photonics Reviews*, 2011, 5(1): 124-166.

[8] PICKWELL E, WALLACE V P. Biomedical applications of terahertz technology[J]. *Journal of Physics D: Applied Physics*, 2006, 39(17): R301.

[9] HIRAKAWA Y, YASUMOTO Y, GONDO T, *et al*. Application of terahertz spectroscopy to rubber products: evaluation of vulcanization and silica macro dispersion[J]. *Electronics*, 2020, 9(4): 669.

[10] ZHANG R, LIU T. Non-destructive measurement of silica content in rubber using terahertz time-domain spectroscopy[J]. *Journal of Applied Physics*, 2022, 132(15): 154901.

[11] CHEN M, XU J. Terahertz spectroscopy for detection of zinc oxide content in rubber compounds[J]. *Materials Chemistry and Physics*, 2023, 298: 127321.

[12] LI X, WANG Y, ZHANG L. Terahertz spectral analysis with limited data: A machine learning approach[J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2021, 250: 119367.

[13] ZHAO F, LIU H, CHEN Y. Small-sample terahertz spectroscopy analysis using transfer learning[J]. *Optics Express*, 2022, 30(4): 5678-5692.

[14] SUN W, LI J, ZHANG Q. Improving terahertz spectral analysis with data augmentation techniques[J]. *Journal of Infrared, Millimeter, and Terahertz Waves*, 2023, 44(2): 215-230.

[15] 郑江鹏, 余平, 赵萌, 等. 利用低信噪比小样本太赫兹光谱实现心肌淀粉样变检测[J]. 中国光学, 2022, 15(3): 443-453.
ZHENG J P, YU P, ZHAO M, *et al*. Detection of myocardial amyloidosis by a small number of terahertz spectra with low signal-to-noise ratio[J]. *Chinese Optics*, 2022, 15(3): 443-453. (in Chinese)

[16] 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)

[17] 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.

[18] DUVILLARET L, GARET F, COUTAZ J L. A reliable method for extraction of material parameters in terahertz time-domain spectroscopy[J]. *IEEE Journal of Selected Topics in Quantum Electronics*, 1996, 2(3): 739-746.

[19] 王圣杰, 王铎, 梁秋金, 等. 小样本学习综述[J]. 空间控制技术与应用, 2023, 49(5): 1-10.
WANG S J, WANG D, LIANG Q J, *et al*. Few-shot learning: a survey[J]. *Aerospace Control and Application*, 2023, 49(5): 1-10. (in Chinese)

[20] DO M N, VETTERLI M. The finite ridgelet transform for image representation[J]. *IEEE Transactions on Image Processing*, 2003, 12(1): 16-28.

[21] 贾延琪, 董双丽, 肖永宝. 掺铽锗酸盐激光玻璃光谱特性定量计算与预测[J]. 发光学报, 2023, 44(5): 889-897.
JIA Y Q, DONG S L, XIAO Y B. Quantitative calculation and prediction of spectroscopic proper-

ties of thulium-doped germanate laser glass[J]. *Chinese Journal of Luminescence*, 2023, 44(5): 889-897. (in Chinese)

[22] 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): 0514001.
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): 0514001. (in Chinese)

[23] 金航峰. 基于光谱和高光谱图像技术的蚕茧品质无损检测研究[D]. 杭州: 浙江大学, 2013: 20-21.
JIN H F. *Research on Nondestructive Determination of Cocoon Quality Based on spectroscopy and Hyperspectral Imaging Techniques* [D]. Hangzhou: Zhejiang University, 2013: 20-21. (in Chinese)

[24] CORTES C, VAPNIK V. Support-vector networks [J]. *Machine Learning*, 1995, 20 (3): 273-297.

[25] AWAD M, KHANNA R. Efficient hyperparameter tuning for support vector regression using grid search and random search[J]. *Journal of Machine Learning Research*, 2015, 16(1): 1-30.

[26] SUN J, XU Y, TANG L. Derivation of Fresnel coefficients in multilayer dielectric films using recurrence relations[J]. *Journal of Suzhou University of Science and Technology: Natural Science Edition*, 2016, 33(3): 5.

[27] ZHAO S, LIU X, SHEN J. The influence of sample thickness error on the refractive index of samples in the THz band[J]. *Journal of Capital Normal University: Natural Science Edition*, 2009, 30(2): 13-15.

[28] WANG S, SHU M, GU X, *et al*. Remote sensing inversion of maize leaf area index using Beer-Lambert extinction law[J]. *Journal of Agricultural Science and Technology*, 2018, 20(12): 67-73.

[29] SU H, ZHANG Z, ZHAO X, *et al*. Analysis of the representation form of the Lambert-Beer Law in quantitative terahertz spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2013, 33(12): 7.

作者简介:

殷贤华(1974—),男,博士,教授,博士研究生导师,主要从事太赫兹检测技术、自动测试总线与系统方面的研究。E-mail:17616260660@163.com

通讯作者:

张　活(1986—),男,博士研究生,副教授,硕士研究生导师,主要从事太赫兹检测技术方面的研究。E-mail:13117730484@163.com