·研究速报·

基于太赫兹时域光谱技术的烟草组分识别

周 博[1]，朱文魁[1※]，王赵改[2]，蒋鹏飞[2]，刘洪坤[1]，李智慧[1]，张 柯[1]，付丽丽[1]
（1. 中国烟草总公司郑州烟草研究院，郑州 450000；2. 河南省农业科学院农副产品加工中心，郑州 450000）

摘 要：为了准确识别不同烟草配方组分，利用太赫兹时域光谱技术，针对烟草工业常用的叶丝、梗丝和再造烟叶丝 3 种烟草配方组分开展太赫兹光谱特性分析和分类识别方法研究。对 0.35～1.50 THz 范围内 3 种烟丝的吸收系数谱和折射率谱进行分析，通过低方差滤波结合主成分分析（Principal Component Analysis，PCA）进行光谱特征提取和降维，分别建立针对吸收谱和折射谱的支持向量机（Support Vector Machine，SVM）分类模型、最邻近分类（K-Nearest Neighbor，KNN）模型和袋装树（Bagged trees）分类模型。结果表明，基于吸收系数谱的分类模型准确率最高，低方差滤波结合 PCA 的特征提取算法能显著提高分类效果，其中 KNN 模型准确率达到 98.3%。对频域光谱使用连续投影算法（Successive Projections Algorithm，SPA）特征提取并结合 SVM 模型，分类准确率也在 90%左右。研究表明太赫兹时域光谱技术可应用于不同烟草组分的分类判别，为太赫兹技术在烟草物料无损检测的应用提供参考。

关键词：光谱；模型；烟草组分；太赫兹时域光谱；分类识别；最近邻分类；低方差滤波
doi：10.11975/j.issn.1002-6819.2022.10.037
中图分类号：O657.3 **文献标志码**：A **文章编号**：1002-6819(2022)-10-0310-07

周博，朱文魁，王赵改，等. 基于太赫兹时域光谱技术的烟草组分识别[J]. 农业工程学报，2022，38(10)：310-316.
doi：10.11975/j.issn.1002-6819.2022.10.037 http://www.tcsae.org
Zhou Bo, Zhu Wenkui, Wang Zhaogai, et al. Identification of tobacco materials based on terahertz time-domain spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(10): 310-316. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2022.10.037 http://www.tcsae.org

0 引 言

太赫兹作为一种波长介于 0.03～30.00 mm 范围内的电磁波，具有高穿透性、指纹谱性等特性[1]，在农产品[2-3]和食品无损检测[4-6]领域获得了很多关注。王远等[7]基于太赫兹时域光谱技术利用支持向量机和随机森林建立 5 种红木的分类模型，变量筛选采用连续投影算法，最终总体分类正确率分别为 94%和 96%。孙旭东等[8]根据红茶基质和夹杂昆虫异物的红茶之间的吸收系数谱和介电损耗谱，采用支持向量机和线性判别分析的方法鉴别夹杂昆虫的红茶样品，结果表明太赫兹时域光谱技术对红茶夹杂昆虫的无损检测具有可行性。但在烟草原料加工领域，尚未有太赫兹技术应用于配方烟丝组分判别的报道。

烟草原料加工涉及烟片、烟梗和再造烟叶 3 种主要物料[9]，这 3 种原料经过制丝处理后构成了成品烟丝中基本配方组分[10]，其中烟片和烟梗来自烟草植株不同的部位，再造烟叶则是利用烟叶边角料和外加纤维经过相关工艺制得的薄片状烟草制品[11]，3 种烟草原料在物理化学性质方面有着显著的区别。不同类型、不同规格的烟草制品包含不同的烟丝配方组分，便捷准确地识别烟草中叶丝、梗丝和再造烟叶丝等组分，对于烟草制品的鉴别、在制品配方加工稳定性及多点同质化生产均有重要意义[12-13]。目前烟草配方组分的检测方法主要包括图像法和光谱法[14-16]。图像法聚焦于不同组分的颗粒形态等外观特征识别，如高震宇等[17]提出一种将配方烟丝局部微观形态特征输入卷积神经网络进而识别烟丝组成成分的方法，准确率最高为 85%，但该方法在烟丝组分形态和物理性状接近时识别效果较差。光谱法目前主要集中于近红外光谱检测法，梅吉帆等[18]采用近红外高光谱成像技术开展配方烟丝组分判别研究，建立变量筛选和支持向量机判别模型，组分判别率为 86.97%。

本文在前人图像法和光谱法的基础上利用太赫兹技术对不同种配方烟丝进行了识别，针对前人鉴别准确率较低的缺陷提高分类准确率，并提出一种低方差滤波结合主成分分析的特征提取方法。本文分析了配方烟丝中的叶丝、梗丝、薄片丝在太赫兹时域光谱技术下的吸收系数谱、折射率谱和频域谱的特征，利用 3 种分类模型对 3 种烟丝进行分类判别，以期为太赫兹技术在烟草配方组分无损快速检测提供参考。

1 材料与方法

1.1 仪器与参数

试验采用 ADVANTEST 公司生产的 TAS7400TS 太赫兹时域光谱系统，该系统构成和工作原理详见文献[8]。本研究设置的扫描频谱范围为 0～5 THz，光谱分辨率为

收稿日期：2021-12-15 修订日期：2022-04-15
基金项目：国家烟草专卖局重点研发科技项目（110202102009）；科技部对发展中国家常规性科技援助项目（KY202002007）
作者简介：周博，研究方向为烟草工艺。Email：17812118762@163.com
※通信作者：朱文魁，研究员，研究方向为烟草工艺技术。
Email：wkzhu79@163.com

1.9 GHz，平均每次扫描次数为 1 024 次。测试环境温度为 25 ℃，环境相对湿度为 5%～10%。

1.2　样品准备

本文研究所用烟支叶组配方的叶丝、梗丝和造纸法薄片丝由河南中烟提供，原料都来自于烟丝掺配生产线。考虑到样品中水分对太赫兹波吸收影响分类结果，对 3 种烟丝进行低温干燥预处理，将烟丝放置干燥箱中采用 45 ℃绝干热风通风干燥 4 h 至恒量待用。在 25 ℃左右和低于 15%相对空气湿度的环境下用旋风磨将 3 种烟丝充分粉碎，过 40 目筛后再用压片机压为致密圆片。本研究共制备 60 组样品，其中每种烟丝各 20 组。同时，制备了 3 种混合烟丝样品，混合样品 1 的质量比为叶丝∶梗丝∶薄片丝=3∶7∶3，混合样品 2 的质量比为叶丝∶梗丝∶薄片丝 =2∶2∶8，混合样品 3 的质量比为叶丝∶梗丝∶薄片丝 =8∶2∶2。按比例称取 3 种烟丝，将其混合后放入旋风磨粉碎，各混合样品压制 3 组圆片，共 9 个样品，求每组样品中光谱均值。进行太赫兹检测之前先检查太赫兹仪器的密封性，向仪器中吹入 30 min 干燥空气，等待系统相对湿度低于 10%，环境温度为 25 ℃左右。打开太赫兹激光光源，先获得背景信号。将压片后的样品放置在太赫兹仪的光圈位置，调整样品在光圈上的位置再测量 3 次取光谱平均值。本文使用五折交叉检验将 60 组样品分为 5 份，依次以其中 4 部分作为训练集，剩下一部分为验证集，分类模型的准确率定义为 5 次验证集模型的识别正确率。

1.3　光谱参数获取

Dorney 等[19]提出太赫兹时域光谱数学模型，给出了获得太赫兹吸收系数和折射率的公式。用复折射率来表达物质的光学性质。

$$\tilde{n}(w)=n(w)-ik(w) \tag{1}$$

式中 $\tilde{n}(w)$ 是复折射率，$n(w)$ 是实折射率，实折射率可以描述太赫兹波穿过样本的色散程度；i 为虚数单位，$k(w)$ 为消光系数，单位为 1，用于描述样本对太赫兹波的吸收能力。经过计算可获得实折射率表达式为

$$n(w)=\frac{\varphi(w)c}{wd}+1 \tag{2}$$

吸收系数表达式为

$$a(w)=\frac{2}{d}\ln\frac{4n(w)}{\rho(w)[n(w)+1]^2} \tag{3}$$

式中 $a(w)$ 为吸收系数，cm^{-1}；d 为样品厚度，cm；c 为电磁波速，cm/s；w 为角频率，rad/s；$\rho(w)$ 为太赫兹波穿过样本与参考信号的振幅比；$\varphi(w)$ 是相位差，rad。

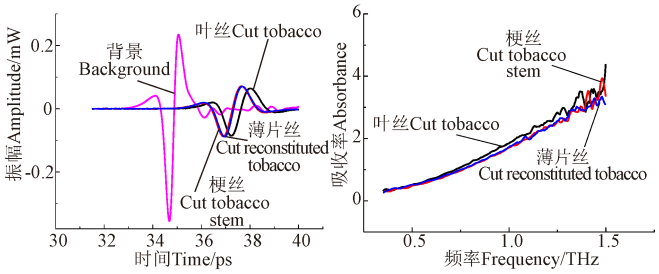
1.4　数据处理

用 S-G 曲线平滑算法对提取出的吸收系数和折射率谱进行平滑去噪处理，建立分类模型对平滑后数据进行分类学习。采用连续投影算法[20]对频域信号波长进行特征提取。本研究使用支持向量机（Support Vector Machine，SVM）[21]、最邻近分类（K-Nearest Neighbor，KNN）[22]和袋装树（Bagged trees）[23]3 种模型对烟丝分类。

2　结果与分析

2.1　烟丝太赫兹光谱响应特征分析

由于物料对高频波吸收较强[24]而导致高频区噪声较大，低频区由于信息密集但是频率分辨率不足，从而造成信号波动较大，因此本研究选择 0.35～1.50 THz 频域的数据作为分析依据。对上述频率段的吸收系数谱和折射率谱测量 3 次求平均值可减少随机误差。从图 1 的时域谱和吸收率谱看出，梗丝和薄片丝差别不明显，而叶丝明显区别于梗丝和薄片丝。



a. 太赫兹时域谱　　b. 太赫兹吸收率谱

a. Terahertz time-domain spectrum　　b. Terahertz absorption spectrum

图 1　三种烟丝太赫兹时域光谱和吸收率谱对比

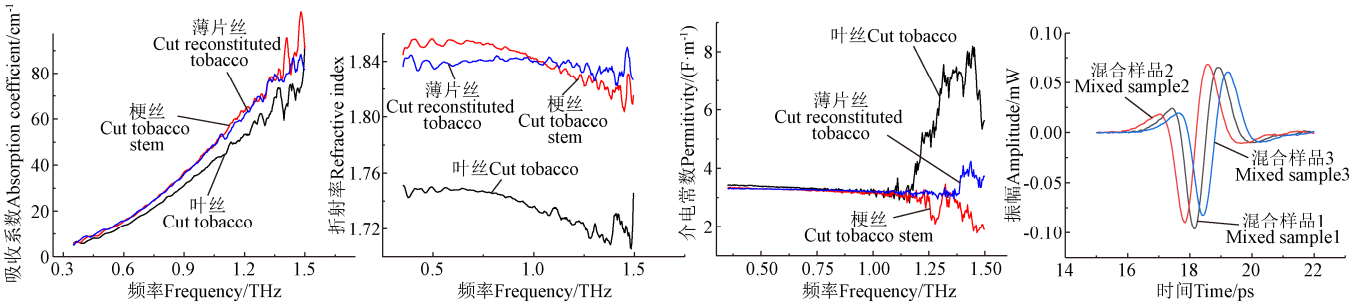
Fig.1　Comparison of terahertz time-domain spectrum and absorbance spectrum of three kinds of tobacco cuts

图 2 依次为 3 种烟丝的吸收系数谱、折射率谱、介电常数谱和 3 种混合样品的时域谱。观察发现 3 种烟丝吸收系数和折射率谱线具有一定的可分性。以折射率为例，叶丝的折射率最低，但是梗丝和薄片丝的折射曲线交叉，在低频区薄片丝折射率低于梗丝，而在 1.1 THz 附近两种烟丝折射曲线存在交点，在高频率区薄片丝折射率高于梗丝。叶丝折射率在 1.70～1.75 范围之间，梗丝折射率范围为 1.80～1.86，薄片丝折射率范围为 1.82～1.86。对于图 2a 吸收系数谱，3 种烟丝吸收系数范围差异不明显，尤其是梗丝和薄片丝吸收系数曲线在 0.35～1.00 THz 区域几乎重合，无法进行区分。结合 3 种烟丝介电常数数据，在 0.350～1.125 THz 频率范围，叶丝介电常数在 3.4 F/m 附近波动，梗丝和薄片丝在 3.3 F/m 左右波动，介电常数不能区分出梗丝和薄片丝。但在吸收系数大于 1 THz 和介电常数大于 1.125 THz 频域范围可以从对三种烟丝进行区分。不同种烟丝在太赫兹光谱曲线上的区别决定了太赫兹光谱可用来分类烟丝种类。

烟草作物是化学体系最为复杂的作物之一，烤烟烟叶中已经发现了 5 229 种化合物。这些化合物中部分大分子的骨架振动和偶极子的旋转和振动对应太赫兹波段[25]。对叶丝、烟梗[26]和薄片丝[27]化学成分的研究显示，叶丝总糖质量分数一般在30%左右，梗丝为20%左右，薄片丝为10%以下；对于纤维素含量而言，叶丝纤维素质量分数一般在10%以下，梗丝在 20%～28%之间，薄片丝在 20%以上。此外，叶丝内部主要是由栅栏组织细胞骨架构成的网状孔隙结构，梗丝内部为较为致密的管束结构，薄片丝主要由纤维交织而成的网状结构[28]。化学组成和物理结构的差异

可能导致了 3 种烟丝在太赫兹光学参数上的差异。葡萄糖等还原糖水合物的介电常数高于纤维素的介电常数[29-30]，峰的时延是由于太赫兹在通过介电常数高的物质时传播速度变慢，峰值高意味着样品对太赫兹吸收少，透过率高[31]，可以推测叶丝时域光谱时延比另两种烟丝大与糖类和纤维素的含量差异有关。

本研究还对 3 种不同混合比例的叶丝、梗丝、薄片丝样品测试了太赫兹时域光谱，见图 2d 所示。由图知叶丝含量较多的样品 3 弛豫时间较大，薄片丝含量较多的样品 2 弛豫时间较小。太赫兹谱图可得出 3 种烟丝在太赫兹时域光谱下的响应不同，叶丝与另外两种烟丝区分度较大，梗丝和薄片丝也有一定区别。



a. 吸收系数 a. Absorption coefficient　b. 折射率 b. Refractive index　c. 介电常数 c. Permittivity　d. 时域光谱 d. Time-domain spectrum

图 2　样品太赫兹时域光谱图

Fig.2　Terahertz time-domain spectrum of samples

2.2　吸收系数和折射率分类模型对比

用噪声较小的 0.35～1.50 THz 数据作为原始输入变量，先对输入变量进行 S-G 平滑，然后分别建立基于吸收系数和折射率原始变量的 SVM、KNN 和 Baggedtrees 烟丝分类模型。五折交叉检验用于验证和评价模型性能，可以提高模型泛化能力，避免过学习和欠学习现象。

基于吸收系数和折射率数据建立 SVM、KNN 和 Baggedtrees 模型，SVM 模型参数通过遗传算法筛选，结果如表 1 所示。

表 1　0.35～1.50 THz 频域吸收系数和折射率的三种模型分类准确率

Table 1　Classification accuracy of three models of absorption coefficient and refraction index in the range of 0.35-1.50 THz

分类模型 Classification model	分类准确率 Classification accuracy /%	
	吸收系数 Absorption coefficient	折射率 Refraction index
支持向量机 SVM	81.1	71.7
最近邻 K-Nearest Neighbor	83.1	66.0
袋装树 Baggedtrees	67.9	67.4

结果表明基于吸收系数的 SVM 模型和 KNN 模型分类效果更好，准确率达到 80%以上，而基于吸收系数的分类模型整体准确率高于基于折射率的模型，这说明对于烟丝来说，吸收系数的数据可分性更好，接下来的数据处理针对吸收系数进行。

2.3　低方差滤波结合主成分分析降维处理

上述分类模型中用到的 0.35～1.50 THz 频率数据较为冗余，不仅包含烟丝成分信息，还包含噪声、背景信息与低分类相关信息等干扰信息[32]，因此通过变量筛选的手段来降低输入数据的维数。

对原始输入数据进行低方差滤波，用变异系数来评估每维特征的离散程度。先对准备数据进行归一化处理，然后计算每一维变量不同样本数据的变异系数，如图 3 所示。

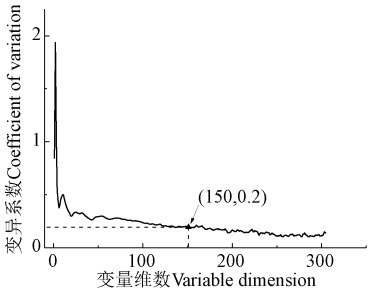


图 3　不同变量维数对应的变异系数

Fig.3　Coefficient of variation corresponding to different variable dimensions

图 3 可以看出变量的变异系数随着维数的增加而降低。选择变异系数为 0.2 作为降维的阈值，将原本 305 个数据点降至 105 个包含更多差异特征的信息，即 0.35～0.91 THz 区域。采用 0.35～0.91 THz 区域数据训练上述 3 种机器学习模型并进行参数优化，结果见表 2。

表 2　低方差滤波降维后的吸收系数谱分类准确率

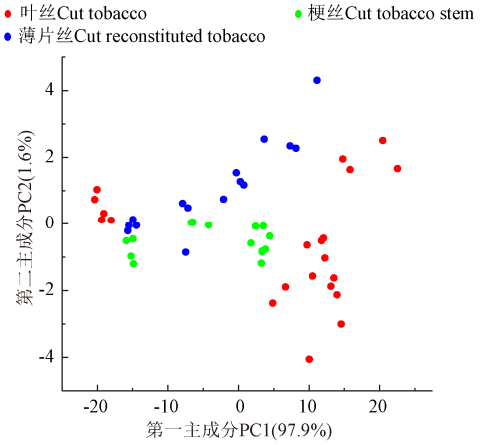
Table 2　Classification accuracy of absorption coefficient spectral after dimension reduction by low variance filter

分类模型 Classification model	分类准确率 Classification accuracy /%
支持向量机 SVM	92.5
最近邻 K-Nearest Neighbor	96.2
袋装树 Baggedtrees	88.7

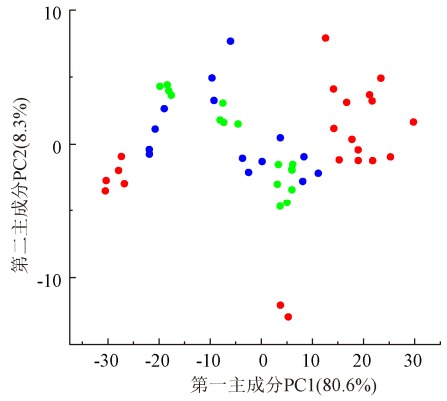
结果表明，用低方差滤波可以在降低维数的同时，大幅提高模型的正确率。这可能因为光谱数据在 150 维以后即 0.91～1.50 THz 频域噪声过大，或包含与分类相

关性不大甚至负相关的冗余信息。

使用主成分分析（Principal Component Analysis，PCA）对数据再次降维。本研究尝试将低方差滤波后数据 PCA 结果与原始输入数据 PCA 进行对比，验证低方差滤波对信息去除冗余的效果。低方差滤波后数据的 PCA 结果前 2 个主成分累计贡献率为 99.5%，高于原始数据前 2 个主成分累计贡献率 88.9%。低方差滤波结合 PCA 降维数据与 0.35～1.50 THz 数据直接 PCA 降维数据的主成分分布可视图如图 4 所示。



a. 低方差滤波结合主成分分析
a. Low variance filter combined with PCA



b. 原始数据主成分分析
b. PCA of raw data

图 4　三种烟丝太赫兹光谱数据主成分分析降维结果

Fig.4　Terahertz spectral data dimensionality reduction results of PCA of three kinds of tobacco cuts

由图 4 可知，低方差滤波结合主成分分析降维处理后的 3 种烟丝数据明显可分，而原始数据直接主成分分析降维处理的第一、第二主成分显示梗丝和薄片丝主成分交叉在一起，可分性较低。这说明低方差滤波可以去除冗余信息，提高模型准确率。

用以上 3 种机器学习模型对主成分分析结果进行分类，模型使用遗传算法和线性搜索算法进行参数优化，结果如下表 3 所示。

结果表明，基于低方差滤波结合主成分分析算法筛选得到的变量构建的 3 种不同分类模型的准确率都达到 90%以上，其中最高的为 KNN 模型，准确率为 98.3%。与用 0.35～1.50 THz 原始数据建立的分类模型对比，使用低方差滤波结合主成分分析算法筛选数据可以提高模型的准确度。

表 3　吸收系数低方差滤波结合主成分分析分类准确率

Table 3　Classification accuracy of absorption coefficients treated by low variance filtering combined with PCA

分类模型 Classification model	准确率 Classification accuracy /%
支持向量机 SVM	94.3
最近邻 K-Nearest Neighbor	98.3
袋装树 Baggedtrees	90.4

2.4　太赫兹频域分类模型

压片法作为光谱技术的常用样品制备方法，其厚度、表面粗糙度等、粉末粒径和紧密程度会影响光谱参数的提取计算，本研究进一步使用不涉及压片物理参数的太赫兹频域信号进行光谱分析处理，以消除制样阶段对提取光谱参数的影响。频域数据由时域数据快速傅里叶变换（Fast Fourier Transform，FFT）获得。烟丝、梗丝和薄片丝的频域信号如图 5 所示，选择 0～1.15 THz 区域数据作为分类模型的输入。

由频域图可以看出，3 种烟丝的频域信号都集中在 0.10～1.15 THz。3 种烟丝频域谱有着明显的区别，烟丝的频域振幅较大，薄片丝居中，梗丝最小，并且在局部肩峰也有差别。先将 3 种烟丝频域数据用 S-G 平滑方法，用低方差滤波结合主成分分析和连续投影算法进行特征提取，将特征输入分类模型中进行分类。连续投影算法（Successive Projections Algorithm，SPA）将数据分为建模集和预测集的，通过建模集选取特征波长建立多元回归模型，计算预测集的均方根误差（Root Mean Square Error，RMSE），选出均方根误差最小的 37 个特征点如图 6 所示，RMSE 为 0.612。可以看出，SPA 选择的特征点大多为拐点或局部极值点，这些点往往能反映不同烟丝之间的差异。

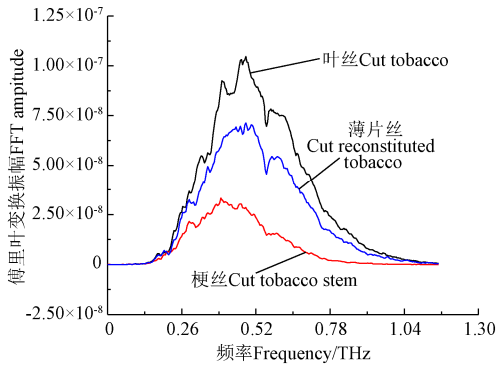


图 5　三种烟丝太赫兹频域光谱图

Fig.5　Terahertz frequency-domain spectrum of three kinds of cut tobacco

表 4 为太赫兹频域光谱的支持向量机分类结果图，连续投影算法结合 SVM 分类正确率为 92.1%，而低方差滤波结合 PCA 再分类的结果为 87.9%，这说明烟丝太赫兹频域信号也包含烟丝之间的差异信息。总体上以吸收系数谱建模的结果优于以频域信号作为输入的结果，这是由于吸收系数代表了具体的物理含义，而频域信号包含了吸收系数等数据融合的复杂信息。

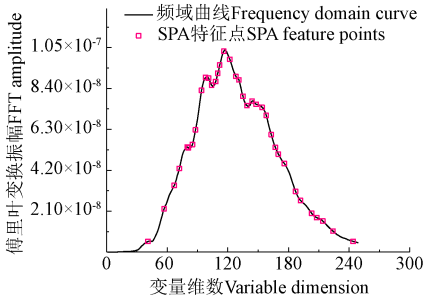


图 6 SPA 特征点选择

Fig.6 SPA feature point selection

表 4 三种烟丝太赫兹频域谱支持向量机分类结果

Table 4 Classification results of SVM based on terahertz frequency-domain spectrum of three kinds of cut tobacco

特征提取方法 Feature extraction method	准确率 Accuracy/%
连续投影算法 SPAsuccessive projections algorithm	92.1
低方差滤波结合 PCA Low variance filtering combined with PCA	87.9

3 结 论

本研究基于太赫兹时域光谱技术，结合低方差滤波和主成分分析对 0.35～1.50 THz 范围内的 3 种烟丝（叶丝、梗丝、薄片丝）的吸收系数谱和折射率谱进行分类模型构建，试验结果证明了 3 种烟丝的太赫兹光谱分类的可行性。得到的结论为：

1）基于吸收系数谱的烟丝分类模型整体分类效果更好，3 种分类模型的准确率均高于基于折射率谱分类模型，其中最邻近分类（K-Nearest Neighbor，KNN）模型正确率优于其他模型。采用 S-G 平滑和低方差滤波结合主成分分析（Principal Component Analysis，PCA）处理能提高数据的可分性，KNN 模型分类正确率为 98.3%。

2）对比低方差滤波结合 PCA 的方法和连续投影算法两种特征提取方法对 3 种烟丝太赫兹频域谱数据进行特征提取，支持向量机（Support Vector Machine，SVM）分类模型准确率分别为 87.9%和 92.1%。研究表明太赫兹频域光谱技术可应用于烟丝组分分类判别，且时域光谱分类结果优于频域。本研究为发展烟草配方组分快速检测技术提供参考。

[参 考 文 献]

[1] 刘盛纲，钟任斌. 太赫兹科学技术及其应用的新发展[J]. 电子科技大学报，2009，38(5)：481-486.
Liu Shenggang, Zhong Renbin. New developments in terahertz science and technology and its applications[J]. Journal of University of Electronic Science and Technology of China, 2009, 38(5): 481-486. (in Chinese with English abstract)

[2] 苏同福，赵国忠，任天宝，等.蒸汽爆破前后生物质秸秆的物理化学特性[J].农业工程学报，2015，31(6)：253-256.
Su Tongfu, Zhao Guozhong, Ren Tianbao, et al. Characterizations of physico-chemical changes of corn biomass by steam explosion[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(6): 253-256. (in Chinese with English abstract)

[3] 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.
Hu Xiaohua, Liu Wei, Liu Changhong, et al. Rapid identification of roducing 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)

[4] 徐振，刘燕德，胡军，等. 基于太赫兹时域光谱技术的掺假川贝母检测[J]. 农业工程学报，2021，37(15)：308-314.
Xu Zhen, Liu Yande, Hu Jun, et al. Detection of adulterated fritillariae using terahertz time domain spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(15): 308-314. (in Chinese with English abstract)

[5] Qin J, Xie L, Ying Y. Feasibility of terahertz time-domain spectroscopy to detect tetracyclines hydrochloride in infant milk powder[J]. Analytical Chemistry, 2014, 86(23): 11750-11757.

[6] 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)

[7] 王远，折帅，周南，等. 基于太赫兹时域光谱技术的红木分类识别[J]. 光谱学与光谱分析，2019，39(9)：2719-2724.
Wang Yuan, She Shuai, Zhou Nan, et al. Classification and identification of rosewood based on terahertz time-domain spectroscopy[J]. Spectroscopy and Spectral Analysis, 2019, 39(9): 2719-2724. (in Chinese with English abstract)

[8] 孙旭东，刘俊彬. 茶叶夹杂昆虫异物 THz 光谱检测研究[J]. 光谱学与光谱分析，2021，41(9)：2723-2728.
Sun Xudong, Liu Junbin. THz spectroscopy detection of insect foreign body hidden in tea porducts[J]. Spectroscopy and Spectral Analysis, 2021, 41(9): 2723-2728. (in Chinese with English abstract)

[9] 《卷烟工艺规范》制订小组. 卷烟工艺规范[M]. 北京：轻工业出版社，1985.

[10] 陈帅伟，胡苏林，崔宁，等. “三丝”掺配比例对卷烟理化指标的影响研究[J]. 西南农业学报，2015，28(6)：408-411.
Cheng Shuaiwei, Hu Sulin, Cui Ning, al et. Effects of ‘three kinds of wires’ blending ratio on cigarette physical and chemical indicators[J]. Southwest Agricultural Journal, 2015,

28(6): 408-411. (in Chinese with English abstract)

[11] 罗一鸣，张献英，何国兴. 造纸法再造烟叶加工技术分析[J]. 广东化工，2021，48(14)：104-105.
Luo Yiming, Zhang Xianying, He Guoxing. Analysis on processing technology of paper-making reconstituted tobacco[J]. Guangdong Chemical Industry, 2021, 48(14): 104-105. (in Chinese with English abstract)

[12] 李瑞丽，刘玉叶，李文伟，等. 利用近红外光谱技术快速检测配方烟丝掺配均匀性[J]. 食品与机械，2019，35(5)：83-87.
Li Ruili, Liu Yuye, Li Wenwei, et al. Study on rapid determination of tobacco blending uniformity by near infrared spectroscopy[J]. Food and Machinery, 2019, 35(5): 83-87. (in Chinese with English abstract)

[13] 刘峘. 烟草加工中固体物料混合的探讨[J]. 烟草科技，2002(7)：6-8，35.
Liu Huan. Discussion on mixing of solid materials in tobacco primary processing[J]. Tobacco Science and Technology, 2002(7): 6-8, 35. (in Chinese with English abstract)

[14] 祝元元，陈永宽，刘志华，等. 近红外光谱技术在烟草行业的应用进展[J]. 应用化工，2010，39(11)：1750-1753.
Zhu Yuanyuan, Chen Yongkuan, Liu Zhihua, et al. Advance on near infrared spectroscopy[J]. Applied Chemical Industry, 2010, 39(11): 1750-1753. (in Chinese with English abstract)

[15] 刘晓萍，李斌，于川芳，等. 基于近红外光谱的卷烟配方结构识别[J]. 烟草科技，2006(10)：16-18，27.
Liu Xiaoping, Li Bin, Yu Chuanfang, et al. Recognition of cigarette blend make-up based on NIR spectrum[J]. Tobacco Science and Technology, 2006(10): 16-18, 27. (in Chinese with English abstract)

[16] 胡立中，张胜军，余小平，等. 均匀设计 PLS-NIR 法预测卷烟配方烟丝中梗丝及薄片丝含量[J]. 中国烟草学报，2010，16(2)：26-30.
Hu Lizhong, Zhang Shengjun, Yu Xiaoping, et al. Predicting the content of shredded stems and shredded shreds in cigarette formula cut tobacco by uniformly designed PLS-NIR method[J]. Chinese Journal of Tobacco, 2010, 16(2): 26-30. (in Chinese with English abstract)

[17] 高震宇，王安，董浩，等. 基于卷积神经网络的烟丝物质组成识别方法[J]. 烟草科技，2017，50(9)：68-75.
Gao Zhenyu, Wang An, Dong Hao, et al. Identification of tobacco components in cut filler based on convolutional neural network[J]. Tobacco Science and Technology, 2017, 50(9): 68-75. (in Chinese with English abstract)

[18] 梅吉帆，李智慧，李嘉康，等. 基于高光谱成像技术的配方烟丝组分判别[J]. 分析测试学报，2021，40(8)：1151-1157.
Mei Jifan, Li Zhihui, Li Jiakang, et al. Component discrimination for formula tobacco based on hyperspectral imaging[J]. Journal of Instrumental Analysis, 2021, 40(8): 1151-1157. (in Chinese with English abstract)

[19] 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-Optics Image Science & Vision, 2001, 18(7): 1562-1571.

[20] Chen J, Bai T C, Zhang N N, et al. Hyperspectral detection of sugar content for sugar-sweetened apples based on sample grouping and SPA feature selecting methods[J]. Infrared Physics and Technology, 2022(125): 104240.

[21] Kaur P, Pannu H S, Malhi A K. Plant disease recognition using fractional-order Zernike moments and SVM classifier[J]. Neural Computing and Applications, 2019, 31(12): 8749-8768.

[22] Saputra R A, Suharyanto, Wasiyanti S, et al. Rice leaf disease image classifications using KNN based on GLCM feature extraction[J]. Journal of Physics: Conference Series, 2020, 1641(1): 012080.

[23] 刘超. 建立基于 Bagged-Trees 算法的早期乳腺癌前哨淋巴结转移预测模型[D]. 大连：大连医科大学，2019.
Liu Chao. Establishment of A Bagged-trees-based Model for Prediction of Sentinel Lymph Node Metastasis for Early Breast Cancer Patients[D]. Dalian: Dalian Medical University, 2019. (in Chinese with English abstract)

[24] 李小霞，邓琥，廖和涛，等. 室温下中药附子的太赫兹波谱分析[J]. 激光与红外，2013，43(11)：1282-1285.
Li Xiaoxia, Deng Hu, Liao Hetao, et al. Terahertz spectroscopic analysis of traditional Chinese medicine aconite at room temperature[J]. Laser and Infrared, 2013, 43(11): 1282-1285. (in Chinese with English abstract)

[25] 李允植. 太赫兹科学与技术原理[M]. 北京：国防工业出版社，2012.

[26] 程向红，王培锋，彭玉富，等. 烟梗主要化学成分特征及其与巴豆醛释放量的关系[J]. 中国烟草科学，2018，39(1)：85-90.
Cheng Xianghong, Wang Peifeng, Peng Yufu, et al. Analysis of chemical components in tobacco stem and correlation with crotonaldehyde[J]. China Tobacco Science, 2018, 39(1): 85-90. (in Chinese with English abstract)

[27] 王茹楠，李晓瑜，张利涛，等. 关键工序对造纸法再造烟叶主要化学成分的影响[J]. 安徽农业科学，2020，48(8)：175-178.
Wang Runan, Li Xiaoyu, Zhang Litao, et al. Effect of key processes on the main chemical components in paper-making reconstituted tobacco[J]. Anhui Agricultural Sciences, 2020, 48(8): 175-178. (in Chinese with English abstract)

[28] 韩李锋，陈良元，李旭，等. 不同烟草材料中水分赋存状态的低场核磁共振分析[J]. 烟草科技，2017，50(4)：65-71，102.
Han Lifeng, Chen Liangyuan, Li Xu, et al. Low-field NMR Analysis of moisture occurrence in different tobacco materials[J]. Tobacco Science and Technology, 2017, 50(4): 65-71, 102. (in Chinese with English abstract)

[29] 陈玉娟，卓克垒，康磊，等. 278. 15-313. 15K 下糖-水二元体系的介电常数[J]. 物理化学学报，2008，24(1)：91-96.

Chen Yujuan, Zhuo Kelei, Kang Lei, et al. Dielectric constants for binary saccharide-water solutions at 278. 15-313. 15 K[J]. Acta Physico-Chimica Sinica, 2008, 24(1): 91-96. (in Chinese with English abstract)

[30] 陈涛，蔡治华，胡放荣，等. 结构相似单糖和二糖分子的太赫兹时域光谱研究[J]. 光谱学与光谱分析，2019，39(3)：686-692.
Chen Tao, Cai Zhihua, Hu Fangrong, et al. A study of terahertz spectra of monosaccharides and disaccharides with structure similarities[J]. Spectroscopy and Spectral Analysis, 2019, 39(3): 686-692. (in Chinese with English abstract)

[31] 刘欢，韩东海. 基于太赫兹时域光谱技术的饼干水分定量分析[J]. 食品安全质量检测学报，2014，5(3)：725-729.
Liu Huan, Han Donghai. Quantitative detection of moisture content of biscuits by terahertz time-domain spectroscopy[J]. Journal of Food Safety and Quality Inspection, 2014, 5(3): 725-729. (in Chinese with English abstract)

[32] 陈孟秋，何明霞，李萌，等. 太赫兹光谱结合特征谱区筛选算法在发动机润滑油含水量定量分析中应用研究[J]. 光谱学与光谱分析，2021，41(5)：1393-1397.
Chen Mengqiu, He Mingxia, Li Meng, et al. Application of interval selection methods in quantitative analysis of water content in engine oil by terahertz spectroscopy[J]. Spectroscopy and Spectral Analysis, 2021, 41(5): 1393-1397. (in Chinese with English abstract)

Identification of tobacco materials based on terahertz time-domain spectroscopy

Zhou Bo[1], Zhu Wenkui[1※], Wang Zhaogai[2], Jiang Pengfei[2], Liu Hongkun[1], Li Zhihui[1], Zhang Ke[1], Fu Lili[1]

(1. *Zhengzhou Tobacco Research Institute of CNTC*, *Zhengzhou* 450000, *China*; 2. *Agricultural and Sideline Products Processing Research Center of Henan Academy of Agricultural Sciences*, *Zhengzhou* 450000, *China*)

Abstract: Tobacco leaves can often be classified in detail by variety, stalk position, and place of production. Different tobacco components include leaf shreds, stem shreds, and reconstituted tobacco leaf shreds. It is a high demand to classify and identify the tobacco components in recent years. This present work aims to analyze the absorption coefficient spectrum and refractive index spectrum of three tobacco components in the range of 0.35-1.50 THz using the terahertz time-domain spectroscopy. The low-variance filtering combined with Principal Component Analysis (PCA) was performed for the spectral feature extraction and dimension reduction on spectroscopy data. Three classification models were developed to determine the specific absorption and refraction spectra of tobacco, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Bagged trees. The results show that a higher accuracy was achieved in the classification model using the absorption coefficient spectrum. The low variance filter combined with the PCA feature extraction significantly improved the classification accuracy, and the KNN model presented an accuracy rate of 98.3%. Furthermore, the Successive Projections Algorithm (SPA) feature extraction was also utilized for the frequency domain spectrum combined with the SVM model, where the classification accuracy was also about 90%. Consequently, the terahertz time-domain spectroscopy technology can be expected to serve the classification of cut tobacco. The finding can provide a strong reference for the application of terahertz time-domain spectroscopy to the non-destructive detection of tobacco materials.

Keywords: spectroscopy; models; tobacco components; terahertz time-domain spectroscopy; classification and recognition; K-nearest neighbor classification; low-variance filtering