中红外光谱的进口木材树种识别方法

冯国红，朱玉杰*，李耀翔

东北林业大学工程技术学院，黑龙江 哈尔滨 150040

摘 要 基于支持向量机和马氏距离探索了中红外光谱分析识别进口的卢氏黑黄檀、风车木、微凹黄檀、燃料紫檀和东非黑黄檀的能力。应用中红外光谱仪采集了500组试验样本进行分析，对试验数据进行了预处理：首先，为了保证样本的有效性，对异常光谱进行了诊断。基于莱特检验法诊断出卢氏黑黄檀和微凹黄檀各有2组异常，风车木、燃料紫檀和东非黑黄檀各有1组异常。为使样本数量统一，五种树种分别剔除了包含异常光谱在内的5组数据；其次，分析了近红外光谱的树种识别研究，结果表明：对光谱数据进行一阶导数处理，可提高识别的精度。因此，对中红外光谱数据进行了平滑处理和一阶导数处理。采用主成分分析提取了光谱数据的特征值，测试集的第一和第二主成分得分的散点图显示，平滑加一阶导数处理的测试集的各自聚类性较平滑处理好。以主成分的得分为特征，基于支持向量机和马氏距离进行了识别研究。考虑到识别方法中主成分个数的选取会直接影响识别的精度，而通常主成分的选取仅参考累计贡献率，此处为使主成分的选取更科学，在支持向量机识别方法中利用粒子群算法进行参数寻优时，对主成分的个数(范围为[5，30])与5折检验下的最佳判别准确率的关系进行了试验，结果表明：平滑处理和平滑加一阶导数处理的主成分个数在[7，11]范围内的5折检验下的最佳判别准确率较高，结合对应的判别准确率，确定了最佳的主成分个数为8个。以前8个主成分作为输入变量，基于支持向量机和马氏距离对测试集进行了测试，结果得出：两种识别方法的正确识别率均较高，支持向量机的识别率略高于马氏距离，平滑加一阶导数处理的识别率均优于平滑处理，平滑加一阶导数处理的支持向量机正确识别率达到了98%，识别效果最好。因此，中红外光谱分析可以作为木材树种识别的一种有效手段。

关键词 中红外光谱；树种识别；一阶导数；主成分分析；支持向量机；马氏距离

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)07-2128-05

引 言

随着木材需求的增加，我国从欧洲、东南亚、非洲、大洋洲等地区进口木材的数量和种类正逐年大幅增长，树种不符是进口贸易中最常见的问题之一，也是主要的欺诈手法。正确鉴定树种是进口木材执法的前提和市场流通的需要[1]。进口木材树种种类繁多，木材树种的快速、准确识别研究已成为木材科学发展中的一个备受关注的问题。近年来，有关木材树种识别的方法主要有DNA法、图像法、高光谱法和红外光谱法等。DNA法是通过提取木材的DNA进行识别的，由于DNA的提取不容易实现，目前研究的较少。图像法主要从木材的纹理特征出发，侧重于图像特征提取方法的研究[2-4]。高光谱法主要利用其窄波段特性在较小的空间尺度上能区分地表的细微变化的优势进行树种识别，该方法主要应用于树种的遥感识别[5-6]。红外光谱法主要基于木材的物质结构信息与光谱的吸收特征的关系进行识别。

光谱分析法具有绿色、高效、可实时在线分析等特点，目前，红外光谱分析已经成为发展最快、最引人瞩目的一门独立的分析技术。近年来人们开始探索基于红外光谱分析识别木材的树种。谭念等基于近红外光谱利用主成分分析和支持向量机进行了树种识别研究，识别率达到94.29%[7]。汪紫阳等基于可见/近红外光谱对树叶树种的识别进行了研究[8]。纵观目前的研究，光谱范围主要集中在近红外区域，对于中红外的研究鲜有报道。中红外的波数范围在400～4 000 cm^{-1}之间，是绝大多数有机物和无机离子的基频吸收带，是红外光谱中吸收能力最强的振动谱区，所以中红外区也被认为是最适合于进行红外光谱定性和定量分析的区

收稿日期：2019-06-10，**修订日期**：2019-10-29

基金项目：中央高校基本科研业务费专项资金项目(2572015CB04)，林业公益性行业科研专项经费项目(201504508)资助

作者简介：冯国红，女，1980年生，东北林业大学工程技术学院副教授 e-mail：fgh_1980@126.com

*通讯联系人 e-mail：782377994@qq.com

域[9-10]。

以进口的卢氏黑黄檀、风车木、微凹黄檀、燃料紫檀和东非黑黄檀为研究对象(该五种树种在日常交易中常被称为大叶紫檀、皮灰黑檀、微凹黄檀、赞比亚血檀和紫光檀)，采用中红外光谱仪获取其光谱数据，对数据进行平滑处理及一阶导数处理，运用主成分分析提取光谱数据的特征，基于常用的模式识别方法-支持向量机和马氏距离建立判别模型[11-12]，验证两种判别方法的识别效果。应用中红外光谱在木材识别领域进行探索与实践。

1　实验部分

1.1　仪器与数据采集

美国 Frontier FT-IR 的傅里叶中红外光谱仪，采用 PerkinElmer spectrum 软件采集漫反射光谱，波数范围 400～4 000 cm^{-1}。

树种试样为 6 cm×4 cm×1 cm 的木块，如图 1 所示。每块木块采集 10 组光谱数据，共采集 500 组，卢氏黑黄檀、风车木、微凹黄檀、燃料紫檀和东非黑黄檀各采集 100 组。

图 1　木块试样

Fig. 1　Wood sample

1.2　数据预处理

平滑处理：此处采用 7 点移动平滑处理。

波数的筛选：观察平滑处理的光谱图，两端的谱图噪声较大，选取 600～3 800 cm^{-1} 波数的数据为分析范围。

异常样本的诊断：诊断方法采用莱特检验法($|v_i|>3s$，v_i 为残差，s 为标准差)。对同一树种的每一组光谱数据分别计算 600～3 800 cm^{-1} 范围内的平均值，记为 $\overline{X}_1$，$\overline{X}_2$，…，$\overline{X}_{100}$，进而计算 $\overline{X}_1$，$\overline{X}_2$，…，$\overline{X}_{100}$ 的平均数 $\overline{X}$ 和标准差 s，当 $|v_i|=|\overline{X}_i-\overline{X}|>3S$ 时，$\overline{X}_i$ 对应的光谱数据为异常样本，剔除。依据该诊断方法得到卢氏黑黄檀和微凹黄檀各有 2 组异常，风车木、燃料紫檀和东非黑黄檀各有 1 组异常。为使样本数量统一，五种树种各剔除包括异常光谱在内的 5 组数据，每种树种剩余 95 组为待分析数据。

导数处理：采用一阶导数处理。

归一化处理：将数据集映射到[0，1]上。

2　结果与讨论

2.1　光谱分析

经平滑处理和一阶导数处理的五种树种的光谱图如图 2 所示。由图 2 可以看出，五种树种的光谱图在 600～1 900 及 2 900～3 800 cm^{-1} 范围内存在差异，尤其是燃料紫檀、风车木与其他 3 种檀差异性明显。经过平滑加一阶导数处理的光谱图差异性较明显。

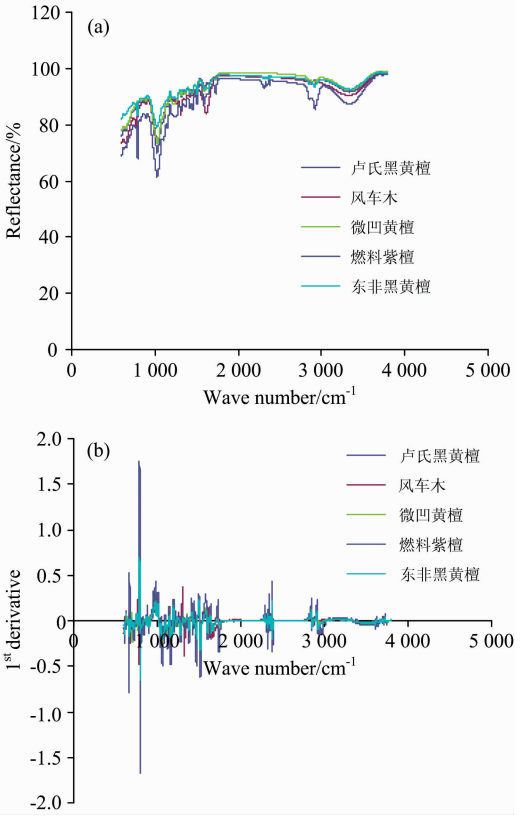

图 2　五种树种的光谱图

(a)：平滑处理；(b)：平滑＋一阶导数处理

Fig. 2　Spectra of five tree species

(a)：Smoothing processing；

(b)：Smoothing+first derivative processing

2.2　主成分分析

主成分分析法是较常用的一种数据压缩特征提取方法，简化原始高维变量的同时最大限度的保留了原始数据的信息。

对五种树种的平滑处理数据和平滑加一阶导数处理数据进行主成分分析，分别绘制测试集的第一、第二主成分得分的散点图，如图 3 所示(为避免数据点密集，此处仅给出前 10 个得分)。由图 3 可以看出，经过平滑加一阶导数处理的测试集的各自聚类性较平滑处理好。

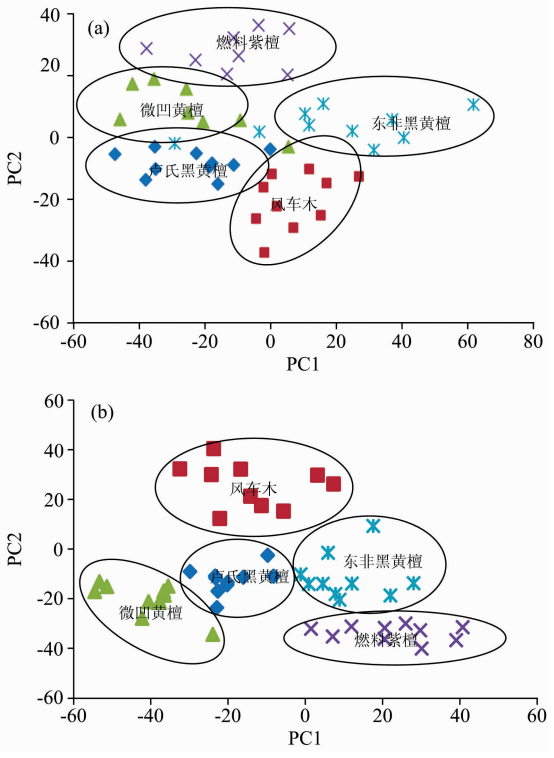

图 3 测试集前 2 个主成分的得分图

(a)：平滑处理；(b)：平滑＋一阶导数处理

Fig. 3 Score plot of the first 2 principal components of the test set

(a)：Smoothing processing;

(b)：Smoothing＋first derivative processing

2.3 支持向量机识别方法

支持向量机(support vector machine，SVM)是一种较常用的模式识别方法，SVM 能够很好的预防欠学习与过学习的发生，在解决实际问题中总是属于最好的方法之一。台湾大学林智仁教授等开发设计了 SVM 的训练与预测工具箱-LIBSVM，可快速有效的进行 SVM 模式识别。此处基于该工具箱进行识别研究。使用时需要确定核函数、惩罚因子 c 及核参数 g，此处确定的核函数为径向基核函数，c 和 g 的寻优方法采用粒子群算法(particle swarm optimization，PSO)。

2.4 SVM 识别结果与讨论

使用 PSO 进行参数寻优时，首先需要确定光谱图的特征个数，即主成分的个数，主成分个数的选取直接影响识别结果。由主成分分析的结果可知，平滑处理和平滑加一阶导数处理的前 5 个主成分的累积贡献率达到了 90%以上，此处对主成分个数为[5，30]范围进行试验验证，以获得最佳值。每种树种的 60 组数据为训练集，15 组数据用于 c 和 g 的寻优，剩下 20 组数据用于测试。利用 Matlab 软件对主成分个数为[5，30]范围进行 c 和 g 的寻优，得到的 5 折检验下的最佳判别准确率如表 1 所示。

由表 1 可以看出，平滑处理和平滑加一阶导数处理的主成分个数在[7，11]范围内的 5 折检验下的最佳判别准确率较高，达到 95%以上，主成分个数在 15 个以上时，5 折检验下的最佳判别准确率降低明显。此处，结合 15 组的判别准确率，最终确定的主成分个数为 8 个。此时得到的 PSO 参数寻优的适应度曲线如图 4 所示。

表 1 不同主成分个数的 5 折检验下的最佳判别准确率

Table 1 The best discriminant accuracy under the 5-fold test of different principal components

主成分个数	最佳判别准确率/%		主成分个数	最佳判别准确率/%	
	平滑	平滑＋一阶导数		平滑	平滑＋一阶导数
5	95	98.33	14	90	90
6	91.67	96.67	15	95	93.33
7	95	96.67	16	91.67	91.67
8	96.67	98.33	17	93.33	91.67
9	95	96.67	18	88.33	91.67
10	95	95	19	86.67	91.67
11	95	95	20	85	90
12	93.33	91.67	25	84.33	86.67
13	93.33	95	30	83.33	81.67

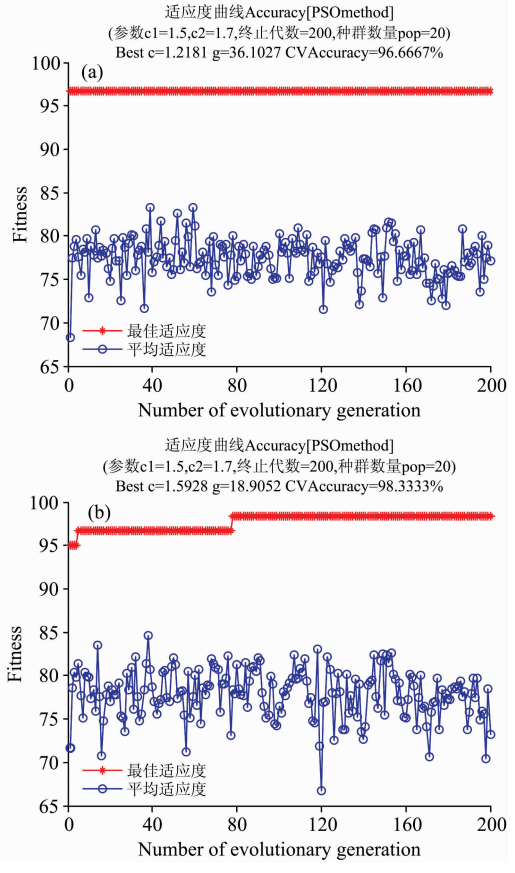

图 4 PSO 参数寻优的适应度曲线

(a)：平滑处理；(b)：平滑＋一阶导数处理

Fig. 4 Adaptability curve of PSO parameter optimization

(a)：Smoothing processing;

(b)：Smoothing＋first derivative processing

由图 4 可知，平滑处理的光谱数据 $c=1.2181$，$g=$

36.102 7 时，5 折检验下的最佳判别准确率为 96.67%；平滑加一阶导数处理的光谱数据 $c=1.592\ 8$，$g=18.905\ 2$ 时，5 折检验下的最佳判别准确率为 98.33%。以径向基函数为核函数，分别以 $c=1.218\ 1$，$g=36.102\ 7$ 和 $c=1.592\ 8$，$g=18.905\ 2$ 建立支持向量机模型，对五种树种的 100 组测试集(每种 20 组)进行分类，得到的各树种的正确识别率如表 2 所示。表 2 的结果表明，基于支持向量机法以前 8 个主成分的得分作为特征，对 5 个树种的识别效果较好。经平滑加一阶导数处理的数据，其识别效果优于平滑处理。平滑加一阶导数处理的数据仅有卢氏黑黄檀和东非黑黄檀各出现了 1 例错判，正确识别率达 98%。

表 2　支持向量机的树种识别结果

Table 2　Tree species recognition result of support vector machine

树种	测试集个数	识别的正确个数	
		平滑	平滑＋导数
卢氏黑黄檀	20	18	19
风车木	20	20	20
微凹黄檀	20	19	20
燃料紫檀	20	20	20
东非黑黄檀	20	18	19
全部	100	95	98

2.5　马氏距离识别方法

马氏距离判别法的基本思想是：首先根据已知分类的数据，分别计算各类的中心，即分类均值，在此基础上，距离判别准则是对于任意给定的一组新样品的观测值，若它与第 i 类中心距离最近，就认为它来自第 i 类。

将每个树种的 75 个校正集的前 8 个主成分得分求平均，记为该树种的中心，利用式(1)求每个验证集到各树种中心的马氏距离，验证集距离哪个中心近，则判定验证集属于该树种。

$$D(X, G_i) = \sqrt{(X-\mu_i)^{\mathrm{T}} V_i^{-1} (X-\mu_i)} \qquad (1)$$

式(1)中，X 为样品均值；G_i 为第 i 类总体；μ_i，V_i^{-1} 分别为第 i 类总体的均值与协方差。

2.6　马氏距离识别结果与讨论

利用 SPSS 软件计算得到的五种树种的 100 组测试集的树种识别结果如表 3 所示。表 3 的结果表明，在马氏距离法中，以前 8 个主成分的得分作为特征，可以获得较好的识别效果。经平滑处理的数据正确识别率达 94%，平滑加一阶导数处理的数据正确识别率达 97%。马氏距离的正确识别率整体略低于支持向量机。

表 3　马氏距离的树种识别结果

Table 3　Tree species recognition result of Mahalanobis distance

树种	测试集个数	识别的正确个数	
		平滑	平滑＋导数
卢氏黑黄檀	20	18	18
风车木	20	20	20
微凹黄檀	20	18	20
燃料紫檀	20	20	20
东非黑黄檀	20	18	19
全部	100	94	97

3　结　论

利用中红外光谱仪采集了卢氏黑黄檀、风车木、微凹黄檀、燃料紫檀及东非黑黄檀五种树种的光谱，进行了平滑处理和一阶导数处理，运用主成分分析法提取了光谱图的特征信息，由测试集的第一和第二主成分的得分，得出五种树种的光谱数据具有较好的各自聚类性，平滑加一阶导数处理的聚类性优于平滑处理。应用支持向量机进行判别研究，对主成分个数为[5，30]范围进行惩罚因子 c 和核参数 g 的寻优，结果表明：主成分个数在[7，11]范围内的 5 折检验下的最佳判别准确率较高，结合验证集的识别准确率确定的主成分个数为 8 个。取前 8 个主成分作为输入变量，基于最优的 c 和 g 进行判别，结果显示：平滑处理的正确识别率达到 95%，平滑加一阶导数处理的正确识别率达到 98%。取前 8 个主成分作为输入变量，进行了马氏距离判别，结果显示：平滑处理的正确识别率达到 94%，平滑加一阶导数处理的正确识别率达到 97%，平均识别率稍低于支持向量机。支持向量机和马氏距离识别中平滑加导数处理的识别效果优于平滑处理，燃料紫檀和风车木的识别效果最好，卢氏黑黄檀的识别率稍低。由支持向量机和马氏距离的识别率可以认为，中红外光谱可用于识别树种，具有良好的应用前景。

References

[1] MENG Qian, LUO Xin-jian, LIU Ying, et al(孟　倩，罗信坚，刘　颖，等). World Forestry Research(世界林业研究), 2017, 30(2): 73.

[2] FU Feng, WANG Xin-jie, WANG Jin, et al(傅　锋，王新杰，汪　锦，等). Remote Sensing for Land & Resources(国土资源遥感), 2019, 31(2): 118.

[3] CHEN Ming-jian, CHEN Zhi-bo, YANG Meng, et al(陈明健，陈志泊，杨　猛，等). Journal of Beijing Forestry University(北京林业大学学报), 2017, 39(2): 108.

[4] WANG Li-jun, HUAI Yong-jian, PENG Yue-cheng(王丽君，淮永建，彭月橙). Journal of Beijing Forestry University(北京林业大学学报), 2015, 37(1): 55.

[5] TAO Jiang-yue, LIU Li-juan, PANG Yong, et al(陶江玥，刘丽娟，庞　勇，等). Journal of Zhejiang A&F University(浙江农林大学学报), 2018, 35(2): 314.

[6] WANG Lu, FAN Wen-yi(王 璐, 范文义). Journal of Northeast Forestry University(东北林业大学学报), 2015, 43(5): 134.
[7] TAN Nian, SUN Yi-dan, WANG Xue-shun, et al(谭 念, 孙一丹, 王学顺, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(11): 3370.
[8] WANG Zi-yang, YIN Shi-kui, LI Chun-xu, et al(汪紫阳, 尹世逵, 李春旭, 等). Journal of Northwest Forestry University(西北林学院学报), 2019, 34(1): 229.
[9] LI Yan-kun, JIA Ming-jing, WANG Han(李艳坤, 贾明静, 王 涵). Journal of Hebei University(河北大学学报), 2018, 38(3): 262.
[10] ZHU Xiang-rong, LI Gao-yang, SHAN Yang(朱向荣, 李高阳, 单 杨). Acta Agriculturae Zhejiangensis(浙江农业学报), 2015, 27(9): 1677.
[11] LIANG Long, FANG Gui-gan, WU Ting, et al(梁 龙, 房桂干, 吴 珽, 等). Journal of Instrumental Analysis(分析测试学报), 2016, 35(1): 101.
[12] GAI Bing-liang, TENG Ke-nan, TANG Jin-guo, et al(盖炳良, 滕克难, 唐金国, 等). Systems Engineering and Electronics(系统工程与电子技术), 2019, 41(3): 686.

Identification Method of Imported Timber Species by Mid-Infrared Spectrum

FENG Guo-hong, ZHU Yu-jie*, LI Yao-xiang
Northeast Forestry Univesity, Harbin 150040, China

Abstract Based on support vector machine and Mahalanobis distance, the ability of mid-infrared spectrum analysis to identify imported rosewood, windmill wood, micro ebony, fuel rosewood and east African rosewood was explored. Five hundred group of test samples were collected and analyzed by the mid-infrared spectrometer, and the test data were preprocessed. Firstly, in order to ensure the validity of the samples, the abnormal spectra were diagnosed. Based on Wright's test, two groups of abnormalities were found in rosewood and micro ebony, one group of abnormalities was found in windmill wood, fuel rosewood and east African rosewood respectively. In order to unify the sample size, five species of trees were excluded from the five sets of data, including the abnormal spectrum. Secondly, the research of tree species recognition in near-infrared spectroscopy was analyzed. The results showed that the first derivative processing of spectral data could improve the recognition accuracy. Therefore, the mid-infrared spectroscopy data were smoothed and first derivative processing. The eigenvalues of the spectral data were extracted by principal component analysis. The scatter plots of the first and second principal component scores of the test set showed that the clustering of the smoothed plus first derivative processed test set was smooth. Based on the scores of principal components, the recognition research was based on support vector machine and Mahalanobis distance. Considering the selection of the number of principal components in the recognition method would directly affect the accuracy of recognition, and usually, the selection of principal components only referred to the cumulative contribution rate. In order to make the selection of principal components more scientific, in the support vector machine identification method, the particle swarm optimization algorithm was used for parameter optimization, the relationship between the number of principal components (range [5, 30]) and the best discrimination accuracy under the 50-fold test was tested. The results showed that the optimal discriminating accuracy of the number of principal components in the range of [7, 11] of smoothing processing and smoothing plus first-order derivative processing was relatively high, and the optimal number of principal components was determined as 8 based on the corresponding discriminating accuracy. The first eight principal components were used as input variables, and the test set was tested based on support vector machine and Mahalanobis distance. The results showed that the correct recognition rates of the two recognition methods were higher, and the recognition rate of support vector machines was slightly higher than that of Mahalanobis distance. The recognition rate of smooth distance plus first-order derivative processing was better than that of smoothing processing. The correct recognition rate of support vector machine with smooth plus first-order derivative processing reached 98%, and the recognition effect was the best. Therefore, the mid-infrared spectrum can be used as an effective means to identify timber species.

Keywords Mid-infrared spectrum; Tree species identification; First derivative; Principal component analysis; Support vector machine; Mahalanobis distance

* Corresponding author
(Received Jun. 10, 2019; accepted Oct. 29, 2019)