

文章编号: 1672-8785(2022)11-0041-08

基于 FTIR 光谱技术的霉菌 菌种判别研究

邹正宇 刘学斌 赵 鑫

(华阴兵器试验中心, 陕西 华阴 714200)

摘 要: 为了找出一种快速、简便、准确的方法来探究霉菌试验后武器装备表面生长的霉菌种类, 按照标准的试验方法进行了霉菌试验, 利用傅里叶红外光谱仪对经过不同菌种腐蚀后的样本进行了测试, 获取了光谱数据, 并通过对数据进行初步的主成分分析确定了光谱数据的识别区域。采用最小距离匹配、光谱角匹配、光谱信息散度、光谱协方差、主成分分析、偏最小二乘判别分析(Partial Least Squares Discriminant Analysis, PLS-DA)、随机森林等分类算法建立了识别模型。研究结果表明, 随机森林算法能够很好地识别霉菌种类, 准确率预期在 98% 以上。基于合适的分类算法, 傅里叶红外光谱(Fourier Transform Infrared, FTIR)技术能够实现对菌种的有效鉴别。

关键词: 傅里叶红外光谱; 偏最小二乘判别分析; 光谱信息散度; 随机森林; 霉菌试验; 菌种识别

中图分类号: O433 文献标志码: A DOI: 10.3969/j.issn.1672-8785.2022.11.007

Research on Identification of Mold Species Based on FTIR Spectroscopy

ZOU Zheng-yu, LIU Xue-bin, ZHAO Xin

(Huayin Weapons Test Center, Huayin 714200, China)

Abstract: In order to find out a fast, simple and accurate method to explore the types of mold growing on the surface of weapons and equipment after mold test, the mold test according to the standard test method is conducted in this paper. Fourier infrared spectrometer is used to test the samples corroded by different strains. Spectral data are obtained, and the identification region of spectral data is determined by principal component analysis. The minimum distance matching, spectral angle matching, spectral information divergence, spectral covariance, principal component analysis, PLS-DA and random forest are used to establish the recognition model. The results show that mold species can be identified well by the random forest algorithm, and the accuracy is expected to be more than 98%. Based on the appropriate classification algorithm, FTIR technology can realize the effective identification of bacterial species.

Key words: FTIR; PLS-DA; spectral information divergence; random forest; mold test; species identification

收稿日期: 2022-08-30

作者简介: 邹正宇(1991-), 男, 吉林省吉林市人, 助理工程师, 硕士, 主要研究方向为装备环境试验方法及评估。E-mail: 18761683681@163.com

0 引言

在武器装备的环境适应性考核中,霉菌试验是 GJB150A-2009 中规定的试验项目,也是武器装备防腐蚀考核的重要环节之一^[1]。但是依据 GJB150A-2009 实施霉菌试验时,在试后评价中通常只能给研制单位提供装备长霉面积这一信息。然而很多装备研制单位想要获悉对被试品影响较大的霉菌菌种种类,进而对装备进行针对性的改良。目前,霉菌菌种的鉴定主要依托菌种保藏中心,鉴定周期长且过程繁琐。因此,寻求一种能够快速识别装备生长霉菌种类的方法,对于加快武器装备试验鉴定进程十分重要。

FTIR 技术是当前常用的物质种类鉴别光谱技术,因具备扫描速度快、精度高、灵敏度高以及可便携现场测量等优势而被广泛应用于民用食品、农业、医药和石油化工领域的成分分析及物质种类判别^[2]。翟明阳等^[3]基于关节软骨切片的显微结构和傅里叶红外光谱,运用化学计量学算法实现了正常和病变软骨的分类判别;张东等^[4]利用傅里叶红外光谱对黄柏、生地等各品质中药材进行了检测,并依据红外光谱的特征峰形实现了不同质量的黄柏、生地等药材的无损鉴别。在霉菌等微生物鉴别领域,红外光谱技术的应用尚处于探索研究阶段。王芳等^[5]对地衣芽孢杆菌等进行了红外光谱测试,通过测试发现不同菌种的光谱吸收峰有一定的不同。傅里叶红外光谱在微生物种类识别领域的研究报道还较少,目前依然缺乏行之有效的菌种快速判别手段。

本文基于 FTIR 测试技术,依据国军标霉菌试验标准选定了部分霉菌菌种,将其培养在不同装备材料上,并进行了傅里叶变换衰减全反射红外光谱(Attenuate Total Reflection Fourier Transform Infrared, ATR-FTIR)测试,找出了菌种光谱指纹区域,对光谱数据进行了有效的预处理,建立了光谱特征识别模型,实现了对霉菌光谱的有效识别。

1 材料与方法

1.1 试验实施

(1) GJB150.10A-2009 中给出了常用的 10 种霉菌菌种。参考各型号装备霉菌试验结果,选择 4 种对装备材料影响明显的菌种作为培养菌种。表 1 列出了试验中选择的菌种种类信息。

表 1 试验中选择的菌种种类信息

霉菌名称	菌种编号	主要影响材料
黑曲霉	AS3.928	织物、乙烯树脂、橡胶
土曲霉	AS3.935	帆布、纸板、涂层
绳状青霉	AS3.875	棉、塑料
黄曲霉	AS3.950	皮革、织物

(2) 选择军事装备中常见的易受霉菌腐蚀的材料(见表 2)作为试验用基材^[6]。

表 2 试验基材种类

项目	内容
皮革	荔枝纹牛皮、超纤 PU 革、山羊皮、油珠皮、油浸皮
涂层	聚氨酯涂层、氨基树脂涂层
橡胶	硅橡胶、异戊橡胶、丁晴橡胶
纤维	环氧树脂玻璃纤维

(3) 试验流程如下:①配置孢子悬浮液;②使用喷枪将孢子悬浮液均匀喷洒在基材上;③设定试验环境温度为 30℃,环境湿度为 95%,时间为 28 天。

1.2 数据采集与处理

测试仪器采用赛默飞 Nicloet iS20 红外光谱仪,扫描区间为 4000~400 cm⁻¹,搭配金刚石漫反射模块对霉菌试验后的样品进行检测。对每个样品选择 5 个长霉点分别进行检测,得到原始光谱数据集。将基线校正(airPLS)、一阶微分(D1)、多项式平滑算法(SG)和多元散射校正(MSC)算法相互组合,对数据进行预处理;采用处理后的数据集来验证最小距离匹配、光谱角匹配、光谱信息散度、光谱协方差、主成分分析、PLS-DA 等分类算法的分类效果^[7]。

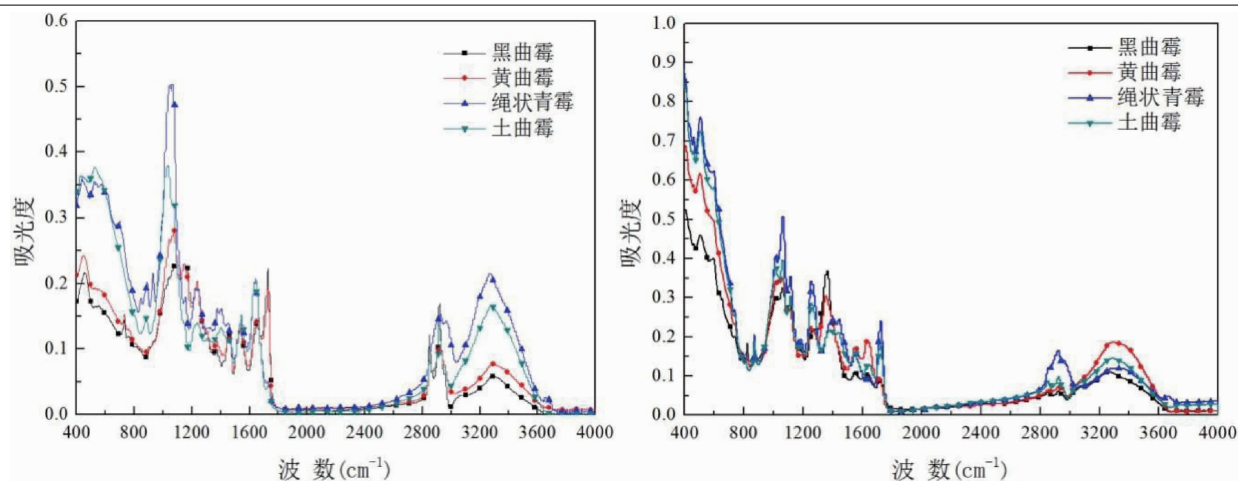


图 1 部分样品的 ATR-FTIR 光谱图

2 结果及分析

2.1 光谱识别区域选择

图 1 为部分样品的光谱图像。为了更加准确地筛选光谱的识别区域, 首先使用一种常用的无监督学习分析方法——主成分分析法对光谱数据进行降维。该方法可从数据中选出少量参数并尽可能地反映数据的完整信息, 是一种常用的降维度数据分析方法^[8]。经主成分分析, 前 4 个主成分的贡献率分别为 55.7%、20.8%、7.5%、6.8%。提取部分样品数据前两个主成分而构成的主成分得分图如图 2 所示。可以看出, 主成分分析能够很好地区分黄曲霉、黑曲霉、无霉菌三种情况, 但是土曲霉和绳状青霉在主成分分析中十分相近, 难以区分。因此, 依据这两个主成分无法较好地识别霉菌种类。

提取部分样品数据前三个主成分而构成的主成分得分图如图 3 所示。可以看出, 各类霉菌样品聚类明显, 区分度较高。前三个主成分的累计贡献率为 84%, 因此它们就反映了光谱数据的主要信息。

分析前三个主成分(PC1、PC2、PC3)的权重, 结果如图 4 所示。在 1800~500 cm⁻¹ 波段, 三个主成分有较多特征峰值。其中, PC1 的特征量较大值位于 650 cm⁻¹、1050 cm⁻¹、1730 cm⁻¹处; PC2 的特征量较大值位于 1090 cm⁻¹、1270 cm⁻¹、1550 cm⁻¹、1740 cm⁻¹、

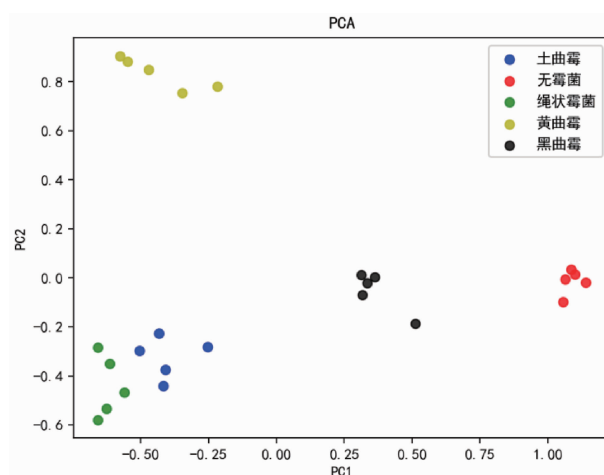


图 2 部分样品前两个主成分的 PCA 分析

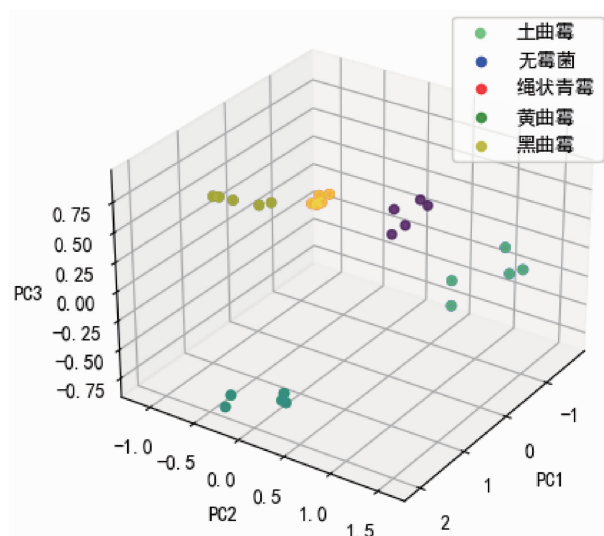


图 3 部分样品前三个主成分的 PCA 分析

1720 cm⁻¹处; PC3 的特征量较大值位于 800 cm⁻¹、1275 cm⁻¹、1650 cm⁻¹处。为了在保证

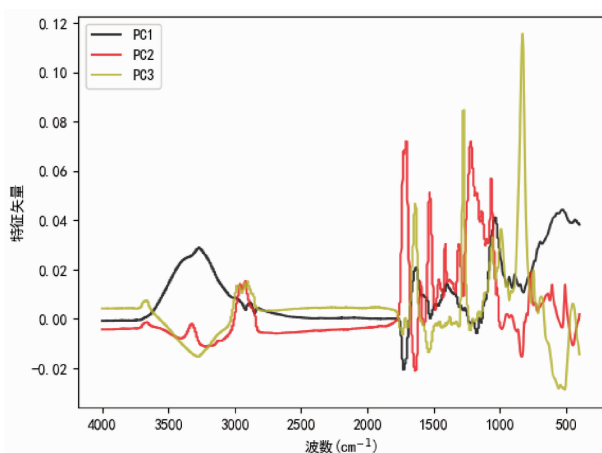


图 4 PC1、PC2、PC3 权重分析图

数据解释性的同时, 尽量减少数据信息量并提升后续算法的计算效率, 选择 $1750\text{ cm}^{-1}\sim 600\text{ cm}^{-1}$ 作为识别区域进行进一步研究。

2.2 分类算法初步判别

光谱数据统计分析方法有最小距离匹配、光谱角匹配、光谱信息散度、光谱协方差、主成分分析、PLS-DA 分类算法等。为选出判别效果最好的算法, 需逐一一对算法进行测试和调试。最小距离匹配、光谱角匹配、光谱信息散度、光谱协方差算法不需要调参, 因此先对这四种算法的识别能力进行分析。由于样本数据较多, 使用上述算法对所有数据进行检测的时间成本较高, 因此先选用荔枝白皮革腐蚀样品的光谱数据作为初步筛选数据集。表 3 列出了最小距离匹配、光谱角匹配、光谱信息散度、光谱协方差算法的光谱向量夹角计算结果。其中的“无”、“黑”、“黄”、“绳”、“土”分别表示没有霉菌污染、黑曲霉污染、黄曲霉污染、绳状霉菌污染、土曲霉污染的各五个样品的光谱平均值, 而“黑 1”则代表黑曲霉污染的荔枝白皮革的 1 号光谱样本, 以此类推。

为初步筛选算法, 选择了 20 个样本作为测试集。其中, 1~10 号为不同霉菌污染的样本, 11~20 号为同种霉菌污染的样本。最小距离匹配、光谱角匹配、光谱信息散度所得结果越相似, 说明两类光谱信息越相似。因此要求最小距离匹配、光谱角匹配、光谱信息散度

算法对不同霉菌污染样本的计算结果尽可能大, 而对同种霉菌污染样本的计算结果尽可能小。分析表 3 可知, 在最小距离匹配算法计算结果中, 当阈值设定为 4 时, 算法准确率最高。黑 1/黑 2、黑 3/黑 2、黄 2/黄 4、绳 2/绳 4、土 1/土 2 这五类样品结果有误, 准确率为 75%。同理可知, 当阈值设定为 0.05 时, 光谱角匹配算法准确率最高为 85%; 当阈值设定为 0.031 时, 光谱信息散度算法准确率最高为 90%; 光谱协方差算法与前几种算法相反, 不同霉菌污染样本的计算结果尽可能小, 而同种霉菌污染样本的计算结果则尽可能大, 其准确率为 25%。综上可知, 这四种算法中光谱信息散度准确率最高。

光谱信息散度算法考察的是两张光谱之间的相似性。同种霉菌在不同的基底上时, 由于营养物质不尽相同且存在基底物质反射光谱的干扰, 光谱数据中难免混进一些同基底相关的信息进而影响识别。因此在上述计算中, 算法输出的结果为“某种霉菌生长于某种基底”, 而并非理想的“某种霉菌”。在使用该方法进行判别时, 样品的霉菌种类和基底种类都必须为数据集已包含的类别。

2.3 PLS-DA 算法优化验证

建立 PLS-DA 模型来对光谱数据进行优化, 仅保留识别区域作为训练集并将其导入模型。通过分析不同主成分个数下的模型参数, 确定模型的最优主成分个数, 最后用模型对样本数据进行预测。

通过 Y 方向累计解释率($R^2Y(\text{cum})$)和模型累计预测率($Q^2(\text{cum})$)来判断模型的拟合效果。数值越接近 1, 模型的预测能力越强。分别对模型建立 1~34 个主成分, 计算模型累计解释率和模型累计预测率(结果见图 5)。可以看出, $R^2Y(\text{cum})$ 和主成分个数之间接近线性正相关, $Q^2(\text{cum})$ 呈现先增大后减小的趋势。因此, 取 $Q^2(\text{cum})$ 为最高点时的个数 19 为最优主成分数。

表 3 部分分类计算结果

序号	样本	最小距离匹配	光谱角匹配	信息散度	光谱协方差	序号	样本	最小距离匹配	光谱角匹配	信息散度	光谱协方差
1	黑/无	21.645	0.238	0.319	0.971	11	黑 1/黑 2	41.344	0.020	0.003	0.100
2	黑/黄	7.469	0.070	0.293	0.997	12	黑 3/黑 2	30.464	0.220	0.003	0.100
3	黑/绳	7.950	0.075	0.054	0.997	13	黄 1/黄 3	3.460	0.014	0.001	0.100
4	黑/土	73.589	0.088	0.028	0.996	14	黄 2/黄 4	29.318	0.084	0.029	0.996
5	黄/绳	4.98	0.058	0.139	0.998	15	绳 2/绳 4	9.085	0.045	0.008	0.999
6	黄/土	66.1	0.126	0.050	0.992	16	绳 3/绳 5	2.982	0.044	0.010	0.999
7	黄/无	21.695	0.234	0.208	0.972	17	土 3/土 5	3.543	0.043	0.004	0.999
8	绳/土	65.639	0.123	0.717	0.992	18	土 1/土 2	12.308	0.055	0.011	0.998
9	绳/无	18.151	0.201	0.142	0.980	19	无 1/无 5	2.033	0.023	0.006	0.100
10	土/无	62.095	0.213	0.295	0.977	20	无 2/无 4	1.162	0.016	0.003	0.100

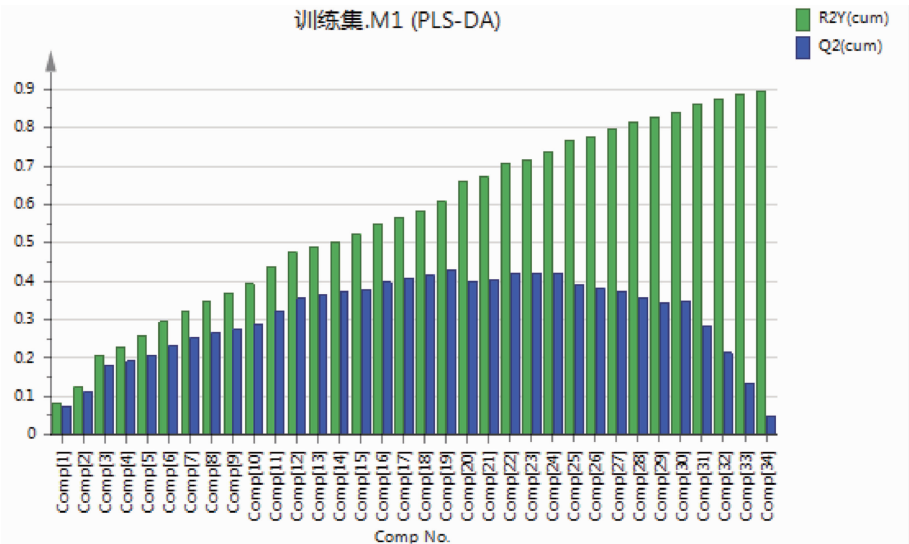


图 5 1~34 个主成分下 PLS-DA 模型的 R2Y(cum)和 Q2(cum)值

表 4 PLS-DA 模型的训练及预测结果

方法	R2Y(cum)	Q2(cum)	最优成分数	训练集	验证集
原始数据	0.61	0.42	19	85.72%	75%
airPLS	0.711	0.417	22	94.34%	83.66%
airPLS、D1	0.621	0.413	15	88.66%	78.17%
airPLS、MSC	0.651	0.406	18	92.19%	29.17%
airPLS、SG	0.767	0.583	28	96.88%	93.75%
airPLS、SG、D1	0.724	0.447	23	94.22%	91.7%
airPLS、SG、MSC	0.772	0.594	24	98.44%	58.33%

鉴于各种客观因素，数据往往会有噪声和误差，需要合适的预处理方法进行消除和纠正。为确定合适的预处理方法，通过将 airPLS、D1、SG 和 MSC 算法相互组合来对数据进行预处理。为防止模型过拟合，采用 10 折

交叉验证分割数据，建立 10 组不同训练集和验证集，并建立 PLS-DA 模型；计算不同预处理方法下模型的 R2Y(cum)和 Q2(cum)，确定最优主成分数并计算 10 组训练集和验证集的平均准确率(结果见表 4)。

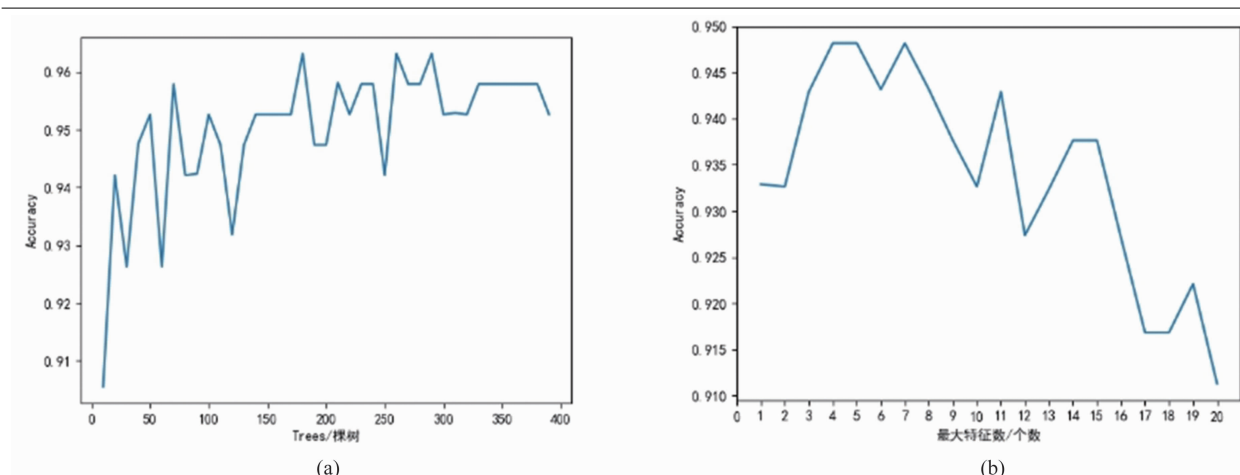


图 6 准确率变化曲线：(a)树数为 10~400；(b)最大特征数为 1~20

由表 4 可知，数据集经过 airPLS 和 SG 算法预处理后，PLS-DA 模型的预测准确率最好，训练集准确率为 96.88%，验证集准确率为 93.75%，无过拟合迹象。

2.4 随机森林算法优化验证

随机森林分类算法有两个主要的超参数：一是预测前构建树的数量，二是拆分节点的最大特征数。首先确定最优树数，设定树数范围为 10~400，步进为 10，其他参数设为默认。计算 10 折交叉验证下验证集的平均准确率，结果如图 6(a)所示。在 10~180 区间内，准确率呈增长趋势；在 180~400 区间内，准确率趋于平稳。设置树数为 180，最大特征数区间为 1~20，步进为 1，计算 10 折交叉验证下验证集的平均准确率，结果如图 6(b)所示。在 1~4 区间内，准确率呈上升趋势；在 4~6 区间内，准确率趋于平稳，然后逐渐下降。为尽可能降低模型的复杂程度，选定最大特征数为 4。

为确定合适的预处理方法，通过将 airPLS、D1、SG 和 MSC 算法相互组合来对数据进行预处理。提取识别区域数据的前 20 个主成分进行降维处理，采用 10 折交叉验证建立不同的训练集和验证集，计算 10 组训练集和验证集的平均准确率(结果见表 5)。

表 5 随机森林模型的训练及预测结果

方法	训练集/%	验证集/%
原始数据	95.22	95.83
airPLS	96.39	93.75
airPLS、D1	97.39	97.92
airPLS、MSC	95.34	97.92
airPLS、SG	94.76	95.83
airPLS、SG、D1	97.42	95.83
airPLS、SG、MSC	98.47	100
airPLS、SG、MSC、D1	97.95	97.92

由表 5 可知，airPLS、SG、MSC 等方法都可在一定程度上提高模型预测准确率。其中，三种算法联用的预处理方法效果最佳，训练集准确率为 98.47%，验证集准确率达到 100%。

2.5 支持向量机算法优化验证

支持向量机模型的主要参数为误差惩罚因子 C 和核函数参数 γ 。为确定合适的预处理方法，采用网格搜索法找出模型参数。设定的参数取值区间如下： C 的取值区间为 10~1000 (步进为 10)， γ 的取值区间为 0.1~1 (步进为 0.1)以及 10、100、1000 和 10000。基于 10 折交叉验证法，找出不同预处理方法下训练集平均准确率最大时的参数组合，并计算其对应的验证集准确率(结果见表 6)。

表 6 支持向量机模型的训练及预测结果

方法	最优 C 值	最优 γ 值	训练集/%	验证集/%
原始光谱	140	0.1	95.33	92.62
airPLS	40	0.2	92.68	95.83
airPLS、D1	20	1000	97.37	93.75
airPLS、MSC	170	0.2	94.21	91.67
airPLS、MSC、D1	210	100	98.42	93.75
airPLS、SG	110	0.2	90.61	93.75
airPLS、SG、D1	20	10000	93.71	89.582
airPLS、SG、MSC	280	0.3	92.66	93.75
airPLS、SG、MSC、D1	120	1000	93.71	93.75

由表 6 可知, 经过 airPLS、MSC 和 D1 处理后, 模型的预测效果最好。此时模型的惩罚因子 C 为 210, 核函数参数 γ 为 100, 训练集准确率为 98.42%, 验证集准确率为 93.75%。

3 结果与讨论

本文采用傅里叶红外光谱仪对 GJB150A 霉菌试验中的部分霉菌菌种进行了 ATR-FTIR 测试, 依据主成分分析结果确定了菌种光谱指纹区域为 $1750 \sim 600 \text{ cm}^{-1}$ 。另外还采用最小距离匹配、光谱角匹配、光谱信息散度等传统方法对数据进行了识别分类。其中, 光谱信息散度的识别率最高。该方法适用于已知基底材料的情况, 准确率接近 90%。

通过选择合适的预处理方法和模型参数, 对 PLS-DA、随机森林和支持向量机三种算法进行了优化, 找出了最优预处理方法, 确定了最佳模型参数并进行了训练验证。优化后的随机森林模型最为理想, 即采用 airPLS、SG 和 MSC 算法进行数据预处理, 树数为 180, 最大特征数为 4, 训练集准确率为 98.47%, 验证集准确率为 100%。

下一步可以利用傅里叶光谱技术对更多试验标准中提到的其他霉菌种类进行测试分析, 进一步完善分析模型, 帮助从业人员分析不同菌种对不同装备材料的影响, 从而完善霉菌试验鉴定评价体系。

参考文献

- [1] 刘福江, 宁薇薇. 霉菌试验标准的方法探讨 [J]. 环境技术, 2018, 36(1): 79–81.
- [2] 王明智. 傅里叶红外光谱仪 (FTIR) 的基本原理及其应用 [J]. 科技风, 2014, 8(6): 112–113.
- [3] 翟明阳, 赵远, 尚林伟, 等. 关节软骨的红外光谱成像及支持向量机判别分析 [J]. 光谱学与光谱分析, 2018, 38(S1): 25–26.
- [4] 张东, 刘志江, 牛佳. 黄柏生地白术的 FTIR 鉴别研究 [J]. 沈阳理工大学学报, 2005, 32(1): 47–49.
- [5] 王芳, 仇大剑, 夏红岩, 等. 太赫兹时域光谱技术在识别鉴定菌制剂中的初探 [J]. 红外与激光工程, 2016, 45(7): 81–87.
- [6] 齐俊臣, 李新俊, 刘春和. 武器装备的霉菌防护技术 [J]. 装备环境工程, 2004, 1(2): 69–72.
- [7] 井诗博, 杨丽明, 李军会, 等. 半监督极限学习机及其在近红外光谱数据分析中的应用 [J]. 计算机应用, 2016, 36(2): 387–391.
- [8] 杨仁杰, 刘蓉, 杨延荣, 等. 用二维相关近红外谱和多维主成分分析判别掺杂牛奶 [J]. 光学精密工程, 2014, 22(9): 2352–2358.
- [9] 常明, 郎海丽, 魏晓菲, 等. 二甲硅油 CH_3 变形振动模式 ATR 二维红外光谱研究 [J]. 红外技术, 2016, 38(1): 59–63.
- [10] 董学锋, 戴连奎, 黄承伟. 结合 PLS-DA 与 SVM 的近红外光谱软测量方法 [J]. 浙江大学学报(工学版), 2012, 46(5): 824–829.
- [11] Dai Q. Recent advances in de-noising methods and their applications in hyperspectral image processing for the food industry [J]. Comprehensive Reviews in Food Science and Food Safety, 2014, 13(6): 1207–1218.
- [12] 尼珍, 胡昌勤, 冯芳. 近红外光谱分析中光谱预

- 处理方法的作用及其发展 [J]. 药物分析杂志, 2008, **28**(5): 824–829.
- [13] Andy L, Matthew W. Classification and regression by Random Forests [J]. *R news*, 2002, **2**(3): 18–22.
- [14] Erc B, Ron K. An empirical comparison of voting classification algorithms: bagging, boosting and variants [J]. *Machine Learning*, 1999, **36**(1): 105–139.
- [15] 徐少成, 李东喜. 基于随机森林的加权特征选择算法 [J]. 统计与决策, 2018, **34**(18): 25–28.
- [16] Xia J S, Liao W Z, Chanussot J, et al. Improving Random Forest with ensemble of features and semisupervised feature extraction [J]. *IEEE Geoscience and Remote Sensing Letters*, 2015, **12**(7): 1471–1475.
- [17] 路世广, 路昌日, 周此明. 针对大规模样本集的训练策略 [J]. 广东技术师范学院学报, 2008, **39**(9): 30–33.
- [18] 张浩然. 支持向量机算法及应用研究 [D]. 上海: 上海交通大学, 2003.

(上接第 19 页)

- [3] 李言谨, 何力, 杨建荣, 等. 碲镉汞红外焦平面器件热失配应力研究 [J]. 红外与毫米波学报, 2008, **27**(6): 409–412.
- [4] 张伟婷. 碲镉汞大面阵红外焦平面探测器的可靠性技术研究 [D]. 北京: 中国科学院大学, 2021.
- [5] 浦磊鑫. 粘接接头强度实验断裂应力计算方法改进研究 [D]. 长春: 吉林大学, 2019.
- [6] 姚仲鹏, 王瑞君. 传热学 [M]. 北京: 北京理工大学出版社, 2003.