

红外光谱结合主成分分析对纸质快递文件袋的分类研究

周贯旭¹, 姜红^{1*}, 胡晓光¹, 陈敏璠², 莫修浩²

(1.中国人民公安大学 侦查学院, 北京 100038; 2.北京鉴知技术有限公司, 北京 100084)

摘要: **目的** 建立一种快速无损的检验纸质快递文件袋的分析方法。**方法** 利用傅里叶变换红外光谱对63个纸质快递文件袋样品进行检验, 分析样品的红外光谱吸收峰的峰位, 结合主成分分析对光谱数据进行了降维处理并分类。利用费歇尔判别对快递文件袋的分类结果进行分析和验证。同时建立多层感知器神经网络和径向基函数神经网络2种分类模型, 进行分析和验证。**结果** 63个纸质快递文件袋样品可被分成四大类, 利用费歇尔分类模型进行验证, 准确率为100%; 多层感知器神经网络分类模型准确率为95.23%, 径向基函数神经网络分类模型准确率为92.06%。通过比较发现, 费歇尔判别可以实现对纸质快递文件袋更加有效地分类。**结论** 该方法简单快速, 样品用量少且无损样品, 可为快递文件袋类的物证鉴定提供科学依据。

关键词: 傅里叶变换红外光谱法; 纸质快递文件袋; 主成分分析; 费歇尔判别

中图分类号: TB486; D918.9; O657.33 **文献标识码:** A **文章编号:** 1001-3563(2023)23-0231-06

DOI: 10.19554/j.cnki.1001-3563.2023.23.028

Classification of Paper Express Document Bags by Infrared Spectroscopy Combined with Chemometrics

ZHOU Guan-xu¹, JIANG Hong^{1*}, HU Xiao-guang¹, CHEN Min-fan², MO Xiu-hao²

(1. Investigation Institute, People's Public Security University of China, Beijing 100038, China;

2. Beijing Jianzhi Technology Co., Ltd., Beijing 100084, China)

ABSTRACT: The work aims to establish a fast and nondestructive analysis method for testing paper express document bags. Sixty-three express document bag samples were tested by Fourier transform infrared spectroscopy. The peak position of the absorption peaks in the infrared spectrum was analyzed, and the spectral data were dimensionally reduced and classified in combination with principal component analysis. Fisher discriminant analysis was used to analyze and verify the classification results of express document bags. At the same time, two classification models of multilayer perceptron neural network and radial basis function neural network were established for analysis and verification. The 63 samples of paper express document bags could be divided into four categories, and the Fisher classification model could achieve 100% accuracy; the classification accuracy of the multi-layer perceptron neural network model was 95.23%, and that of the radial basis function neural network model was 92.06%. Through comparison, it was found that Fischer discriminant could achieve a more effective classification of paper express document bags. This method is simple, rapid, less sample consumption and non-destructive to samples. It can provide a scientific basis for identification of physical evidence of express document bags.

KEY WORDS: Fourier transform infrared spectroscopy (FTIR); paper express document bag; principal component analysis; Fischer discriminant

收稿日期: 2023-01-06

基金项目: 中国人民公安大学 2021 年度基科费重点项目 (2021JKF212)

*通信作者

随着快递行业的迅速发展,快递业务非常普遍。在各类案件现场,经常能提取到纸质快递文件袋物证,通过对该类物证的检验分析,并对其进行准确快速分类,有助于判断该物证来源以及主要运输途径。可为侦查破案提供线索,缩小侦查范围^[1]。

目前,用于检验纸张物证的方法主要有,红外光谱法、拉曼光谱法、扫描电镜/能谱法、X射线荧光光谱法^[2]、X射线衍射法、原子发射光谱法等^[3]。利用红外光谱结合主成分分析对纸质快递文件袋的相关研究尚未见报道。本实验采用便携式红外光谱仪对收集到63个不同品牌、不同来源的快递文件袋样品进行检验,依据快递文件袋表面的涂漆填料的不同对其进行分类,借助spss 26.0软件中的费歇尔判别式、多层感知器神经网络和径向基神经网络对上述分类结果进行分析验证,进而优选一种最适合此类样品的分类方法。

1 实验

1.1 仪器及条件

IT 2000 傅里叶变换红外光谱仪(北京鉴知技术有限公司);扫描次数为32次;光谱范围为4 000~500 cm^{-1} ;分辨率为4 cm^{-1} 。

1.2 实验样本

来自不同城市的顺丰、韵达、中通、圆通、EMS、京东、申通等7个不同快递公司的纸质快递文件袋样品63个(样品表略)。

1.3 实验方法

分别剪取约1 cm^2 大小的样品,利用红外光谱在上述实验条件下,对实验样品进行测试。

重复性实验:随机选取57#样品,对其同一部位进行5次平行实验;随机选取13#样品对其上、下、左、右、中心5个部位进行测试。

2 结果与讨论

2.1 重复性实验结果分析

为验证仪器的精密度和样品的均匀性,确保实验结果准确可靠,对样品进行了重复性实验。57#样品同一个部位的5次平行实验的红外光谱图见图1;13#样品的5个部位的红外光谱图见图2。当把一些微小的干扰峰排除后,样品在同一位置的红外光谱图基本一致,在不同位置的红外光谱图也基本一致,说明仪器具有较好的精密度,样品的均匀性较好。因此,利用该实验方法对纸质快递文件袋进行检验的结果准确可靠。

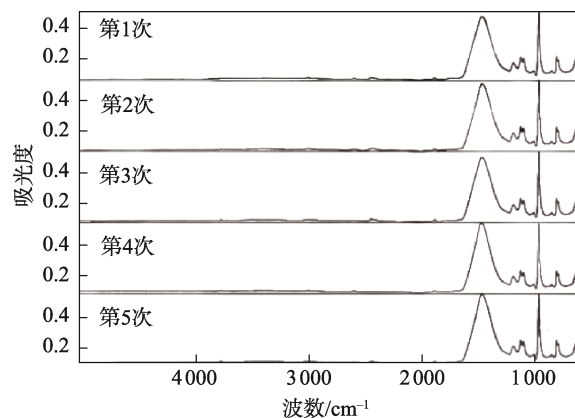


图1 57#样品的红外光谱图
Fig.1 Infrared spectrum of 57# sample

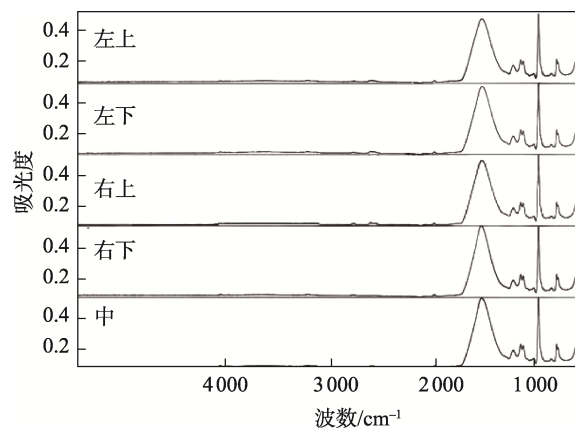


图2 13#样品的红外光谱图
Fig.2 Infrared spectrum of 13# sample

2.2 根据快递文件袋表面涂漆填料成分进行分类

纸质快递文件袋的原料主要是纤维素和半纤维素,因为快递文件袋表面一般是白色为主,因此在生产过程中,生产厂家常常会添加一定量的涂漆填料。常见的填料主要有如碳酸钙、高岭土、滑石粉、膨润土等,其中713、876、1 434、1 796 cm^{-1} 可作为识别碳酸钙的红外光谱吸收峰,542、694、916、1 037、1 117、3 622、3 655、3 696 cm^{-1} 可作为识别高岭土的红外光谱吸收峰,670、1 019 cm^{-1} 可作为识别滑石粉的红外光谱吸收峰,916、1 037、3 625 cm^{-1} 可作为识别膨润土的红外光谱吸收峰^[4]。

为满足不同纸张的性能要求,不同的快递文件袋的填料成分或配比会有所不同。本实验根据填料种类的不同,可以将样品分为四大类(分类结果见表1)。

第I类样品的主要填料是碳酸钙(见图3),其中864.39 cm^{-1} 为碳酸钙C—O面内振动弯曲^[5]吸收。碳酸钙在填料中增强了涂层耐磨损和抗腐蚀的能力,使涂层的烘烤温度降低,压缩烘烤时间,在降低成本的同时,可有效提高产品质量。

表 1 样品分类结果
Tab.1 Sample classification results

类别	填料成分	样品编号
第 I 类	碳酸钙	1 [#] ~2 [#] 、4 [#] ~19 [#] 、22 [#] 、24 [#] ~28 [#] 、31 [#] 、35 [#] ~37 [#] 、40 [#] 、49 [#] 、51 [#] 、53 [#] ~55 [#] 、58 [#] 、60 [#] 、61 [#]
第 II 类	碳酸钙+膨润土	20 [#] ~21 [#] 、23 [#] 、29 [#] 、39 [#] 、46 [#] ~48 [#] 、56 [#]
第 III 类	碳酸钙+高岭土	3 [#] 、30 [#] 、32 [#] ~34 [#] 、42 [#] ~44 [#] 、50 [#] 、52 [#] 、57 [#] 、59 [#] 、62 [#] ~63 [#]
第 IV 类	碳酸钙+滑石粉	38 [#] 、41 [#] 、45 [#]

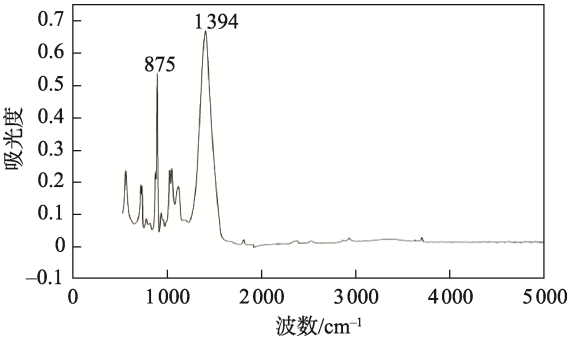


图 3 第 I 类中 1[#]样品的红外光谱图
Fig.3 Infrared spectrum of sample 1[#] in class I

第 II 类样品的主要填料是碳酸钙和膨润土(见图 4),其中 3 685.74 cm^{-1} 是膨润土内表面—OH 的吸收, 2 920.87 cm^{-1} 是—CH₃ 和—CH₂ 对称和反对称伸缩振动吸收峰^[6], 1 026.75 cm^{-1} 是 Si—O—Si 反对称伸缩振动吸收, 1 005.10 cm^{-1} 为 Si—O 伸缩振动吸收, 763.37 cm^{-1} 为 Si—O—Mg 吸收^[7]; 871.61 cm^{-1} 为碳酸钙 C—O 面内振动弯曲吸收。膨润土可以降低生产成本, 并且还能使纸张颜料分散更加均匀、提高纸张表面光洁度等。

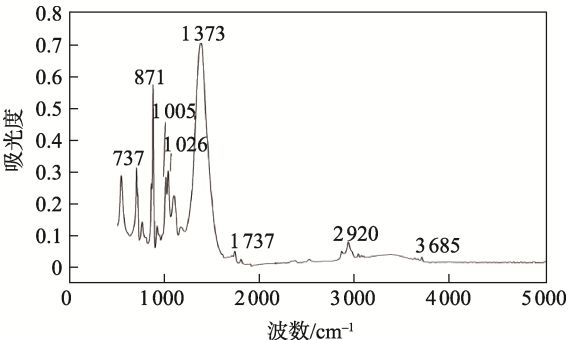


图 4 第 II 类 21[#]样品的红外光谱图
Fig.4 Infrared spectrum of sample 21[#] in class II

第 III 类中的主要填料是碳酸钙和高岭土(见图 5), 其中 870.41 cm^{-1} 为碳酸钙 C—O 面内振动弯曲吸收; 3 693.98 ~ 3 624.41 cm^{-1} 内的吸收峰均是高岭土晶体内部—OH 的吸收, 1 005.40 cm^{-1} 为 Si—O 伸缩振

动吸收, 756.79 cm^{-1} 为 Si—O—Mg 吸收^[8]。高岭土具有可塑性好、黏结性高和绝缘性能好等特点, 易分散悬浮于水中, 在提高纸张抗酸溶性、耐火性等方面具有较好的作用。

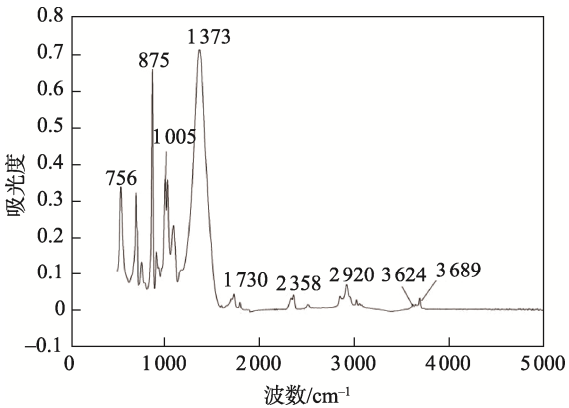


图 5 第 III 类 57[#]样品的红外光谱图
Fig.5 Infrared spectrum of sample 57[#] in class III

第 IV 类中样品的主要成分是碳酸钙加滑石粉(见图 6), 其中 876.42 cm^{-1} 为碳酸钙 C—O 面内振动弯曲吸收; 1 010.49 cm^{-1} 为滑石粉 Si—O 的伸缩振动吸收, 672.52 cm^{-1} 为—OH 弯曲振动吸收^[9]。滑石粉不但能够改善纸张的不透明度和平滑度, 而且还能提高纸张的适印刷性, 价格低廉, 具有较高的经济效益。

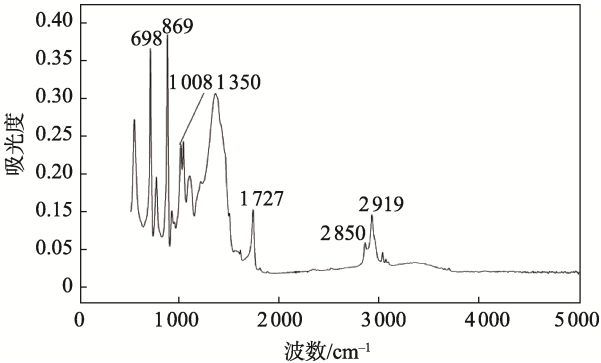


图 6 第 IV 类 41[#]样品的红外光谱图
Fig.6 Infrared spectrum of sample 41[#] in class IV

2.3 利用主成分分析法降维

主成分分析法 (Principal Component Analysis, PCA) 是一种基于统计特征的多维正交线性变换, 常用来对信号进行特征提取和对数据进行降维^[10]。可大大降低样本数据的复杂程度, 使样本数据处理更加方便。当数据累计方差贡献率大于 80%, 可以较好地解释基本原始信息^[11], 实验中将原样品数据降维成 26 个变量, 其方差累计贡献率为 99.71%, 说明 PCA 提取效果好。

2.4 费歇尔判别分析

鉴于人工进行分类可能会存在误差的情况, 利用费歇尔判别分析对分类结果进行验证, 以降维后的 26 个变量为研究对象, 建立费歇尔判别函数^[12]。对 3 个判别函数的模型摘要进行分析, 进而确定哪个判别函数在实现样本分类预测方面具有更显著的作用。判别函数的摘要统计见表 2。

表 2 判别函数的摘要统计

Tab.2 Summary statistics of discriminant function

函数	特征值	方差 百分比/%	累计 百分比/%	典型 相关性
1	47.011	74.1	74.1	0.990
2	12.418	19.6	93.6	0.962
3	4.048	6.4	100.0	0.895

费歇尔判别分别提取了 3 个特征函数^[13], 前 2 个特征函数所占方差百分比分别为 74.1%和 19.6%, 累计方差贡献率为 93.6%。其方差的百分比代表着对于该模型对于样本的分类问题的解释能力的大小, 故函数 1 和函数 2 这 2 个函数所携带的信息大于第 3 个函数, 则以函数 1 和函数 2 作为模型的判别函数绘制样品联合分布散点图 (见图 7)。

由图 7 可知, 人工分类四大类的质心在散点图上相对分离, 其中第 2 类和第 4 类的分类效果尤其显著。同时用该模型对分类结果进行了预测, 其正确判别率为 100% (见表 3), 说明人工分组的结果具有一定的准确性和科学性。

2.5 人工神经网络模型分析

人工神经网络 (ANN) 是一种监督学习算法的模型, 其信息的处理大致分为 3 个层次, 即在输入层中输入外部信息, 然后通过隐藏层进行信息的处理, 最后通过输出层将信息进行下一步的传递^[14]。多层感知器 (Multilayer Perceptron, MLP) 神经网络和径向基函数 (Radical Basis Function, RBF) 神经网络分类

模型, 都是典型的神经网络算法。

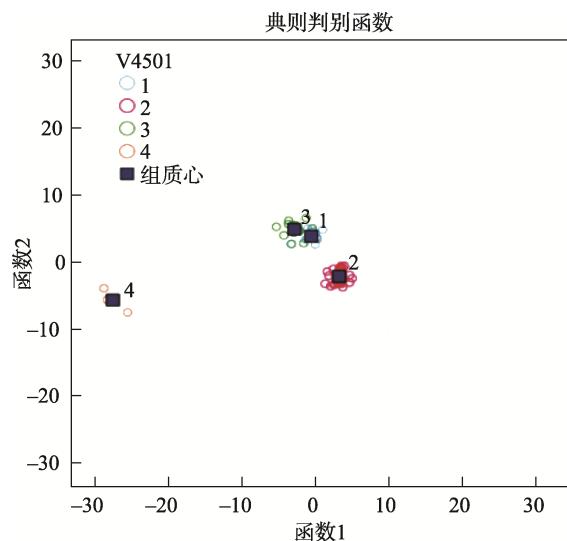


图 7 样品联合分布散点图
Fig.7 Scatter diagram of joint
distribution of samples

表 3 费歇尔模型分类预测结果

Tab.3 Classification prediction results of Fisher model

类别	观测值计数				判别率/%			
	1	2	3	4	1	2	3	4
1	9	0	0	0	100	0	0	0
2	0	37	0	0	0	100	0	0
3	0	0	14	0	0	0	100	0
4	0	0	0	3	0	0	0	100
总计	9	37	14	3	100	100	100	100

以 63 个样本的红外光谱数据为基础, 对原始数据构建 MLP 和 RBF 2 种分类模型, 将样本按照 70% 和 30% 的比例划分为训练集和测试集, 即 44 个样品作为训练集, 19 个样本作为测试集^[15]。在 MLP 模型对该类样品的测试中, 协变量是每个样品的波长作为输入层神经元, 隐藏层层数为 1 层, 隐藏层中包含 3 个神经元。选择双曲正切函数为输入函数, 令 Softmax 函数作为输出层激活函数, MLP 模型分类结果见表 4, 其中训练集的准确率为 95.45%, 测试集的准确率为 94.73%, 其总体准确率为 95.23%。

在 RBF 模型的模型参数中, 输入层神经元选择该类样品的红外波长作为 RBF 模型的协变量, 选择正态化径向基函数作为隐藏层激活函数^[16]。RBF 模型分类结果见表 5, 其中对训练集的正确率可以达到 93.18%, 反观对测试集的正确率则只有 89.47%, 其总体准确率为 92.06%。

表 4 MLP 模型分类结果
Tab.4 MLP model classification results

样本	实测	预测				正确百分比/%
		A	B	C	D	
训练	A	7	0	0	1	87.50
	B	0	25	0	0	100.00
	C	0	0	9	1	90.00
	D	0	0	0	1	100.00
	总体百分比/%	15.91	56.82	20.45	6.82	95.45
检验	A	1	0	0	0	100.00
	B	0	12	0	0	100.00
	C	0	1	3	0	75.00
	D	0	0	0	2	100.00
	总体百分比/%	5.26	68.43	15.79	10.52	94.73

表 5 RBF 模型分类结果
Tab.5 RBF model classification results

样本	实测	预测				正确百分比/%
		A	B	C	D	
训练	A	7	0	1	0	87.50
	B	0	25	0	0	100.00
	C	1	0	8	1	80.00
	D	0	0	0	1	100.00
	总体百分比/%	18.18	56.82	20.45	4.55	93.18
检验	A	1	0	0	0	100.00
	B	1	11	0	0	91.66
	C	0	0	4	0	100.00
	D	1	0	0	1	50.00
	总体百分比/%	15.79	57.89	21.06	5.26	89.47

3 结语

利用红外光谱法对快递文件袋样品进行了检验分析,依据填料的不同,可将 63 个样品分为四大类。通过 PCA 进行降维处理后,提取了 26 个特征值大于 1 的主成分,其可以解释原光谱数据中 99.71%的信息。根据费歇尔判别分析对人工分类结果进行验证,建立费歇尔判别函数和分类预测模型,该方法的识别准确率为 100%。基于样品数据建立的 MLP 和 RBF 这 2 种模型,也可验证分类方法的准确率,其准确率分别为 95.23%和 92.06%。对未知样品的判别,费歇尔判别分析模型的准确率更高。将红外光谱与主成分分析相结合,可以对纸质快递文件袋样品进行快速无损地检测,为公安机关实际办案提供帮助。

参考文献:

[1] 姜红. 刑事案件现场中纸张的检验及应用[J]. 上海造纸, 2005(3): 9-12.
JIANG Hong. On the Examination of Paper at Case Crime Scene and Its Application[J]. Shanghai Paper Making, 2005(3): 9-12.

[2] 廉哲, 梁鲁宁, 光晓俐, 等. X 射线荧光光谱在文件/纸质文物检验中的应用研究进展[J]. 刑事技术, 2022(2): 185-190.
LIAN Zhe, LIANG Lu-ning, GUANG Xiao-li, et al. Research Progress on XRF Applied into Examination of Document/Paper Cultural Relics[J]. Forensic Science

- and Technology, 2022(2): 185-190.
- [3] FISCHER T, MARCHETTI-DESCHMANN M, CRISTINA ASSIS A, et al. Profiling and Imaging of Forensic Evidence – a Pan-European Forensic round Robin Study Part 1: Document Forgery[J]. Science & Justice, 2022, 62(4): 433-447.
- [4] 付钧泽, 姜红, 陈煜太, 等. 傅里叶变换红外光谱法检验香烟水松纸的研究[J]. 中华纸业, 2017, 38(20): 43-48.
- FU Jun-ze, JIANG Hong, CHEN Yu-tai, et al. A Study on Test of Tipping Paper with Fourier Transform Infrared Spectroscopy[J]. China Pulp & Paper Industry, 2017, 38(20): 43-48.
- [5] 赵华雄, 侯小红. 制浆系统中出现的沉积物及其红外图谱简析[J]. 中华纸业, 2022, 43(12): 20-24.
- ZHAO Hua-xiong, HOU Xiao-hong. Deposit in Bleached Pulp Production System and Its Infrared Spectrum Identification[J]. China Pulp & Paper Industry, 2022, 43(12): 20-24.
- [6] 杨莹琴, 赵新富, 张永笑. 己内酰胺插层膨润土的研制及其性能[J]. 信阳师范学院学报(自然科学版), 2009, 22(4): 565-567.
- YANG Ying-qin, ZHAO Xin-fu, ZHANG Yong-xiao. Production and Performance of Caprolactam Intercalation Bentonite[J]. Journal of Xinyang Normal University (Natural Science Edition), 2009, 22(4): 565-567.
- [7] CAGGIANI M C, OCCHIPINTI R, FINOCCHIARO C, et al. Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) as a Potential on Site Tool to Test Geopolymerization Reaction[J]. Talanta, 2022, 250: 123721.
- [8] ABDEL M M I A, YOUSSEF M A, GHOBASHY M M, et al. Gamma Radiation-Induced Synthesis of Organoclay Based Polyaniline and Ilmenite Clay Minerals for Cesium Ions Removal from Aqueous Solutions[J]. Separation and Purification Technology, 2023, 305: 122434.
- [9] WU Hao-yu, HE Ming-yue, WU Shao-kun, et al. Near-Infrared Spectroscopy Study of OH Stretching Modes in Pyrophyllite and Talc[J]. Crystals, 2022, 12(12): 1759.
- [10] PFEZER D, KARST J, HENTSCHEL M, et al. Predicting Concentrations of Mixed Sugar Solutions with a Combination of Resonant Plasmon-Enhanced SEIRA and Principal Component Analysis[J]. Sensors, 2022, 22(15): 5567.
- [11] AHSAN M, MASHURI M, KHUSNA H. Comparing the Performance of Kernel PCA Mix Chart with PCA Mix Chart for Monitoring Mixed Quality Characteristics[J]. Scientific Reports, 2022, 12(1): 1-12.
- [12] ZHANG Jin, GAO Peng-ya, WU Yuan, et al. Identification of Foodborne Pathogenic Bacteria Using Confocal Raman Microspectroscopy and Chemometrics[J]. Frontiers in Microbiology, 2022, 13: 874658.
- [13] ZHANG Jin, JIANG Hong, DUAN Bin, et al. A Rapid and Nondestructive Approach for Forensic Identification of Cigarette Inner Liner Papers Using Shift-Excitation Raman Difference Spectroscopy and Chemometrics[J]. Journal of Forensic Sciences, 2021, 66(6): 2180-2189.
- [14] 赵崇文. 人工神经网络综述[J]. 山西电子技术, 2020(3): 94-96.
- ZHAO Chong-wen. A Survey on Artificial Neural Networks[J]. Shanxi Electronic Technology, 2020(3): 94-96.
- [15] 李锦, 姜红, 杨俊, 等. 差分拉曼光谱结合人工神经网络对药品塑料包装瓶的分类研究[J]. 塑料工业, 2022, 50(8): 101-107.
- LI Jin, JIANG Hong, YANG Jun, et al. Classification of Plastic Pharmaceutical Packaging Bottles Based on Differential Raman Spectroscopy and ANN[J]. China Plastics Industry, 2022, 50(8): 101-107.
- [16] 田陆川, 杨俊, 姜红. 基于 RBFNN-FDA 模型对塑料拖鞋鞋底的拉曼光谱研究[J]. 上海塑料, 2022, 50(4): 62-67.
- TIAN Lu-chuan, YANG Jun, JIANG Hong. Study on Raman Spectra of Plastic Slipper Soles Based on RBFNN-FDA Model[J]. Shanghai Plastics, 2022, 50(4): 62-67.

责任编辑: 曾钰婵