基于 *K* 近邻算法的塑钢窗拉曼光谱分析

何欣龙，陈利波，王继芬，桑国通

中国人民公安大学刑事科学技术学院，北京 100038

摘要 拉曼光谱技术在法庭科学中有着广泛的应用。采用显微激光拉曼光谱分析技术和 *K* 近邻算法对 25 个塑钢窗样本进行研究。通过主成分分析提取到 5 个主成分，并运用训练样本为测试样本的方法进行交互验证。当 $K=1$ 时，测试样本的出错率最低，以区分贡献值最高的三个特征变量为参数建立分类模型，实现了对未知变量的准确归类，模型总分类准确率可达 71%，区分效果良好，比直接通过谱图比较得到的结论更加准确。

关键词 光谱学；拉曼光谱；塑钢窗；*K* 近邻算法；鉴别

中图分类号 O657 **文献标识码** A **doi**：10.3788/LOP55.053001

Raman Spectroscopy Analysis of Plastic Steel Window Based on *K* Nearest Neighbors Algorithm

He Xinlong, Chen Libo, Wang Jifen, Sang Guotong

Institute of Forensic Science and Technology, People's Public Security University of China, Beijing 100038, *China*

Abstract Raman spectroscopy has been used in forensic science widely. In this paper, laser Raman spectroscopy analysis technology and *K* nearest neighbors algorithm are used to study 25 plastic steel window samples. The five principal components are extracted by principal component analysis, and the experiment built interactive verification test with the method regarding the training sample as the test sample. When the *K* value equals to 1, the lowest training sample error rate appears. Taking the three highest contribution value characteristic variables as parameters to build the classification model to realize the accurate classification of the unknown variables, and the total correct rate is 71%. The above method is more accurate than the direct observation of the spectra.

Key words spectroscopy; Raman spectra; plastic steel window; *K* nearest neighbors algorithm; identification

OCIS codes 330.6230; 330.6170; 300.6190

1 引 言

塑钢窗的检验鉴定是理化物证中的一项重要工作[1]，在相关案件中，其能快速划定侦查方向，为侦查破案提供帮助。塑钢窗由聚氯乙烯树脂[2]、氯化聚乙烯[3]、钛白粉[4]、丙烯酸酯类共聚物[5]和其他添加剂[6]组成，是目前新型建筑的主要材料之一。拉曼光谱技术[7-10]是法庭科学微量物证检验中最常用的技术之一，具有信息丰富、制样简单、水的干扰小等优点，能实现无损分析，可用于食品安全[11]、农业工程[12]和考古研究[13]等方面。

K 近邻算法是一种基于距离度量的有效分类方法，其主要原理是从训练集中找到和新数据最接近的 *k* 条记录，并根据其主要分类决定新数据的类别。在分类过程中，新数据只与近邻的几个样本相关，不使用额外数据，也不需要事先确定类别数量，就能达到理想的分类效果。该算法克服了样本类别数量不均的弊端[14]，无需估计参数和训练，易于实现，在模式识别等领域具有广泛应用。本文使用显微激光拉曼光谱分析技术和 *K* 近邻算法对塑钢窗进行建模区分，以期为塑钢窗种类鉴别提供一定的参考依据。

收稿日期：2017-10-10；**收到修改稿日期**：2017-11-25

作者简介：何欣龙(1994—)，男，硕士研究生，主要从事刑事技术方面的研究。E-mail：1078683050@qq.com

导师简介：王继芬(1964—)，女，硕士，教授，主要从事微量物证与毒物毒品分析方面的研究。E-mail：wangjifen58@126.com(通信联系人)

2 实 验

2.1 实验样本

从收集到的 25 个不同品牌和型号的塑钢窗样本中选取 23 个作为后期建模时的训练样本。其中“海螺 1”标记为“New 1”,“实德 8”标记为“New 8”,作为未知样本。

2.2 实验仪器及参数设置

实验仪器包括:显微激光拉曼光谱仪(Nicolet Almega XR,780 nm 激光光源,1800 line/mm 光栅分光,物镜倍数为 50×,波长能量为 2.0 mW,扫描时间为 8 s,曝光次数为 4,针孔直径为 300 μm,光谱分辨率约为 1 cm^{-1},测量范围为 3100~200 cm^{-1})、手术刀、镊子、显微镜。

3 结果与讨论

3.1 拉曼光谱分析

23 个样本的拉曼光谱如图 1 所示。“实德 4”和“实德 6”的拉曼光谱如图 2 所示。由图 1 和图 2 可见:两图拉曼光谱的峰型和峰数几乎一致,相对峰高的差异较小;在 2900 cm^{-1} 处均出现了一个高强度的尖峰,且峰型与出峰位置相同;在 2800~1700 cm^{-1} 范围内未形成明显的峰;在 1500~1200 cm^{-1} 范围内,样本均出现了左高右低的三峰;在 1000~300 cm^{-1} 范围内,各样本的出峰位置、峰型与相对峰高差异较明显,谱图存在交叉重叠的现象。因为塑钢窗是混合物,其组分和添加剂会相互干扰,无法直接通过谱图实现对样本的有效区分,所以需要综合考察样本间的差异,构建对上述样品具有高鉴别能力的分类模型。

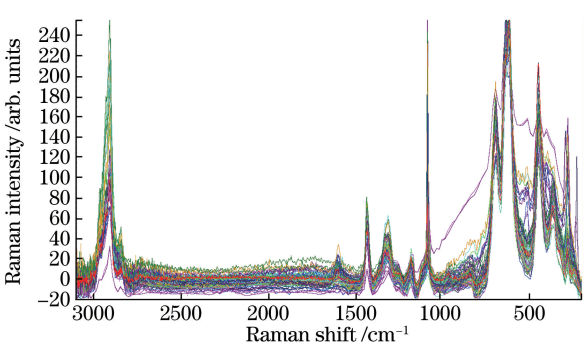

图 1 23 个样本的拉曼光谱

Fig. 1 Raman spectra of 23 samples

3.2 K 近邻算法

K 近邻算法是一种有监督模式识别的方法,具体描述如下[15]:在 N 个已知样本中找出 χ(χ 为 N 个样本中的一个样本的 K 个近邻,在这 K 个样本中,设来自 ω_1 类的样本有 N_1 个,来自 ω_2 类的样本有 N_2 个,……,来自 ω_i 的样本有 N_i 个。若 K_1,K_2,…,K_c 分别是 K 个近邻中属于 ω_1,ω_2,…,ω_c 类的样本数,则可以定义判断函数 $\varphi_i(x)$ 为

$$\varphi_i(x)=k_i,i=1,2,\cdots,c, \tag{1}$$

式中:k_i 为 K_i,即 K 个近邻中属于 ω_i 的样本数。

决策规则为:若 $\varphi_i(x)=\max\{\varphi_i(x)\}$,则决策 $x\in w_i$(w_i 为第 i 类样本)。

运用 K 近邻算法建立数学模型时,主要考虑 K 值的选取。K 即 K 个距离最近的“邻居”,在建模过程中,K 值过小会降低分类精度,过大可能会增加噪声,降低分类效果。通常,K 值的设定采用交叉检验的方式(以 $K=1$ 为基准),低于训练样本数的平方根即可。借助 K 近邻算法建模之前,需对光谱数据进行预处理,使不同数量级的数据之间可以比较,同时删去重复变量,保证新变量尽可能保持原有的信息。

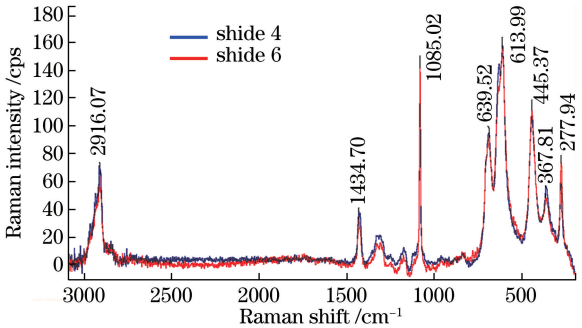

图 2 实德 4 和实德 6 样本的拉曼光谱

Fig. 2 Raman spectra of Shide 4 sample and Shide 6 sample

以样本在 3100~200 cm^{-1} 范围内的光谱数据为基础,采用 Z 标准化[16],结合主成分分析[17]法,求得协方差矩阵的特征变量和特征根,计算得到样本的 5 个主成分,部分结果如表 1 所示。

表 1 7 个样本的主成分分析结果

Table 1 Principal component analysis results of 7 samples

Sample	PCA1	PCA2	PCA3	PCA4	PCA5
Hailuo 1	5.69052	4.12079	9.62737	3.85800	0.01534
Hailuo 2	1.23356	1.38609	0.77216	0.68149	2.60546
Hailuo 3	1.84294	9.49006	3.69626	2.08477	0.39530
Hailuo 4	0.26705	8.46751	6.73757	3.13073	1.70425
Hailuo 5	2.73194	3.34770	8.22784	7.54459	0.81463
Hailuo 6	4.02364	4.39492	5.88556	0.44348	3.47548
Jinpeng 1	8.36999	6.38767	0.16108	0.36157	0.38632

以各样本的 5 个主成分为特征变量,运用训练样本即为测试样本的方法交互验证[18],建立 K 近邻算

法模型对各样本进行分类，得到了如图 3 所示的 K 选择错误统计图和如图 4 所示的特征变量重要性图。

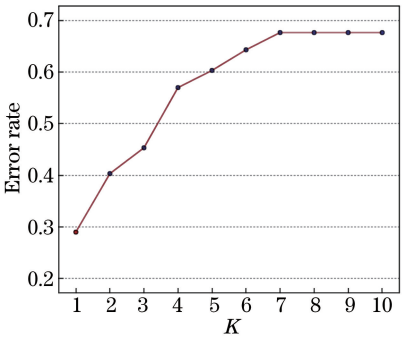

图 3 K 选择错误统计图

Fig. 3 K parameter error statistical chart

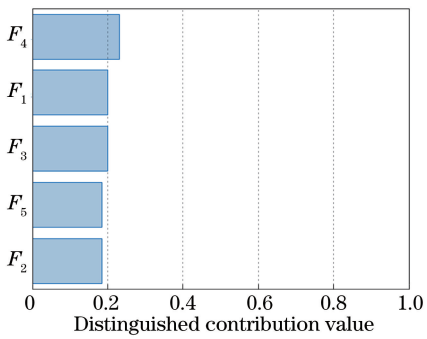

图 4 特征变量重要性图

Fig. 4 Significance chart of characteristic variables

由图 3 可知，模型分类的错误率随 K 增大呈递增的趋势，K 为 1 时的错误率最低，模型能较大程度地实现样本间的区分。在图 4 中可以看出 5 个特征变量在做出预测时的重要程度，F_4（即特征变量 4）为区分贡献最大的特征，F_2 为区分贡献最小的特征，5 个特征变量的重要程度的和为 1。在训练集中，有些样本包含的信息更为全面，是更值得“依赖”的。为保证模型建立的合理性，在计算距离时需要按重要性加权特征进行计算，即给不同的样本施加不同的权重，增大可信赖样本的权重，降低不可信赖样本的影响。

选取 K 值为 1，特征变量贡献值最大的 F_4、F_1 和 F_3 变量作为参数，23 个样本作为训练集，“New 1”和“New 8”作为未知样本，建立分类模型，得到了如图 5 所示的各样本特征空间模型的交互式图和如表 2 所示的未知样本的预测结果。

由图 5 可知：样本有 6 个聚集区域，共分成了 8 类，B、C、D 三类样本的分类情况较为明显，分类效果较好；A 和 E 两类样本分布集中，聚敛程度较大，其在特征变量 F_4、F_1 和 F_3 空间中的区分能力仍存在一定不足；G和H两类样本的离散程度较大，表明样本个

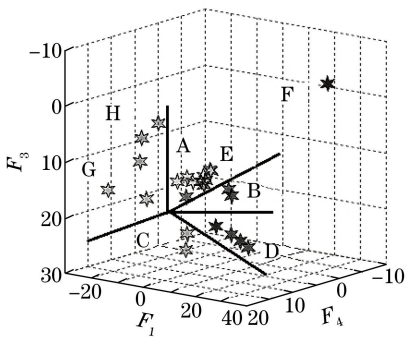

图 5 特征空间模型的交互式图

Fig. 5 Eigenvectors space model interactive graphic

表 2 未知样本的 $K(K=1)$ 个最近相邻和最近距离

Table 2 $K(K=1)$ nearest neighbor and nearest distance of unknown samples

Prediction sample	Nearest neighbor	Nearest distance
New 1	Hailuo 5(D)	0.145
New 8	Shide 3(A)	0.120

体之间的差异较为明显。由表 2 可见：当 $K=1$ 时，“New 1”和“海螺 5”最近，距离为 0.145，被划分在 D 类；“New 8”和“实德 3”最近，距离为 0.120，被划分在 A 类。两种预测样本均实现了正确的归类识别，分类正确率达 71%，分类结果较为理想。

4 结 论

基于 K 近邻算法对塑钢窗的拉曼光谱数据展开挖掘，实现了对未知样本的识别和分类。通过主成分分析实现了特征变量的提取和量化，避免了 K 近邻算法计算量较大带来的不足。通过讨论 K 近邻算法中 K 值的选择和特征变量的贡献值，合理地建立了分类模型，实现了未知样本的分类。这种分类方法借助数学运算，综合考察了变量内部的差异，比直接通过谱图得到的结论更加准确，具有一定的借鉴意义。在本实验中，分类正确率为 71%，正确率较低的原因主要是样本数量少，在建立分类模型过程中，训练样本集的可解释性比较薄弱。虽然本实验的样本数量有限，但实验方法具有普遍意义，可为今后塑钢窗等其相关物证的鉴别提供参考。

参 考 文 献

[1] He X L, Wang J F, Liu W H, *et al*. Discrimination and classification the plastic steel window based on Roman spectroscopy and cluster analysis [J]. Chemical Research and Application, 2017, 29(9): 1387-1392.
何欣龙, 王继芬, 刘文浩, 等. 拉曼光谱结合聚类分

析法区分检验塑钢窗[J]. 化学研究与应用, 2017, 29(9): 1387-1392.

[2] Zhang G F, Xiao N. Industry present situation, development trend and suggestions of PVC resin[J]. Henan Chemical Industry, 2013, 30(5): 21-23.
张国锋, 肖娜. 聚氯乙烯树脂行业现状、发展趋势及建议[J]. 河南化工, 2013, 30(5): 21-23.

[3] Ren H, Wang W Y, Zhang R, *et al*. Production status of chlorinated polyethylene and development progress of HDPE special resin[J]. Modern Plastics Processing and Applications, 2017, 29(1): 60-63.
任鹤, 王文燕, 张瑞, 等. 氯化聚乙烯生产现状及HDPE专用树脂开发进展[J]. 现代塑料加工应用, 2017, 29(1): 60-63.

[4] Zhang Z T, Liu Q, Liu Q J. Research on dispersion of titanium dioxide before surface treatment [J]. Materials Review, 2013, 27(s1): 23-25.
张智涛, 刘强, 柳清菊. 钛白粉表面处理前的分散性研究[J]. 材料导报, 2013, 27(s1): 23-25.

[5] Chen Y Y, Yang J H, Chen X M, *et al*. Effect of core structure of ACR on the toughening of PVC[J]. China Plastics Industry, 2014, 42(6): 29-32.
陈严严, 杨景辉, 陈雪梅, 等. ACR核层结构对PVC增韧的影响[J]. 塑料工业, 2014, 42(6): 29-32.

[6] Wang Y M, Ning P S, Ding Z M. The application of plastics stabilizaters in unsatured polyster glass fiber reinforced plastic (GFRP) composites[J]. Plastics Additives, 2011(5): 20-24.
王玉民, 宁培森, 丁著明. 塑料稳定剂在玻璃纤维增强不饱和聚酯树脂中的应用[J]. 塑料助剂, 2011(5): 20-24.

[7] Xu B, Lin M M, Yao H L, *et al*. Measurement of hemoglobin concentration of single red blood cell using Raman spectroscopy [J]. Chinese Journal of Lasers, 2016, 43(1): 0115003.
徐斌, 林漫漫, 姚辉璐, 等. 拉曼光谱技术测量单个红细胞的血红蛋白浓度[J]. 中国激光, 2016, 43(1): 0115003.

[8] Fang X Q, Peng Y K, Li Y Y, *et al*. Rapid and quantitative detection method of sodium benzoate in carbonated beverage based on surface-enhanced Raman spectroscopy[J]. Acta Optica Sinica, 2017, 37(9): 0930001.
房晓倩, 彭彦昆, 李永玉, 等. 基于表面增强拉曼光谱快速定量检测碳酸饮料中苯甲酸钠的方法[J]. 光学学报, 2017, 37(9): 0930001.

[9] Qu Y T, Li Y, Guan R Y, *et al*. Research progress of algae based on laser spectroscopy technology[J]. Laser & Optoelectronics Progress, 2017, 54(6): 060004.
曲颖桐, 李颖, 关冉昀. 激光光谱技术应用于藻类的研究进展[J]. 激光与光电子学进展, 2017, 54(6): 060004.

[10] Wang Q, Zhao C, Yang H N, *et al*. Simultaneous measurement of film thickness and mass fraction by Raman spectroscopy [J]. Journal of Instrumental Analysis, 2016, 53(9): 093001.
王琴, 赵畅, 杨荟楠, 等. 激光拉曼光谱法同步测量液膜厚度与浓度[J]. 激光与光电子学进展, 2016, 53(9): 093001.

[11] Chen Z Y, Lu Y L, Liang Y, *et al*. Rapid analysis of saccharin sodium salt in dried fruits by surface-enhanced Raman scattering spectroscopy[J]. Journal of Instrumental Analysis, 2017, 36(5): 650-654.
陈正毅, 卢雅琳, 梁豫, 等. 表面增强拉曼光谱法快速测定干果类食品中的糖精钠[J]. 分析测试学报, 2017, 36(5): 650-654.

[12] Li S F, Zhang X, Li J J, *et al*. Non-destructive detecting fructose and glucose content of honey with Raman spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(6): 249-255.
李水芳, 张欣, 李姣娟, 等. 拉曼光谱法无损检测蜂蜜中的果糖和葡萄糖含量[J]. 农业工程学报, 2014, 30(6): 249-255.

[13] Liu Z J, Han Y X, Yang R, *et al*. Micro-Raman analysis of the pigments in the mural paintings from a Ming dynasty tomb[J]. Chinese Journal of Lasers, 2013, 40(6): 0615003.
刘照军, 韩运侠, 杨蕊, 等. 明代古墓葬壁画颜料的显微拉曼光谱分析[J]. 中国激光, 2013, 40(6): 0615003.

[14] Peng X H, Liu F, Chen J, *et al*. Identification of gutter oil based on the model of PCA-LS-SVM[J]. Computers and Applied Chemistry, 2013, 30(10): 1207-1210.
彭秀辉, 刘飞, 陈珺, 等. PCA-LS-SVM预测模型在地沟油鉴别中的应用[J]. 计算机与应用化学, 2013, 30(10): 1207-1210.

[15] Song L M, Luo J. Pattern identification [M]. Beijing: China Machine Press, 2015: 82-85.
宋丽梅, 罗菁. 模式识别[M]. 北京: 机械工业出版社, 2015: 82-85.

[16] Zhang H, Wang Q J, Zhu J J, *et al*. Influence of sample data preprocessing on BP neural network-based GPS elevation fitting[J]. Journal of Geodesy

and Geodynamics, 2011, 31(2): 125-128.
张昊, 王琪洁, 朱建军, 等. 样本数据预处理对基于BP神经网络的GPS高程拟合的影响[J]. 大地测量与地球动力学, 2011, 31(2): 125-128.

[17] Huang A, Yang L A, Du T, *et al*. Comprehensive assessment of soil nutrients based on PCA[J]. Arid Zone Research, 2014, 31(5): 819-825.
黄安, 杨联安, 杜挺, 等. 基于主成分分析的土壤养分综合评价[J]. 干旱区研究, 2014, 31(5): 819-825.

[18] Liu L L, Wu Y W, Zhang X, *et al*. Application of Fourier transform infrared spectroscopy combined with pattern recognition method for rapid authentication of edible Oil[J]. Acta Chimica Sinica, 2012, 70(8): 995-1000.
刘玲玲, 武彦文, 张旭, 等. 傅里叶变换红外光谱结合模式识别法快速鉴别食用油的真伪[J]. 化学学报, 2012, 70(8): 995-1000.