

# 机器学习辅助差分拉曼光谱对蓝色中性笔色料的分类研究

黄娟娟<sup>1\*</sup>, 向万欣<sup>1</sup>, 夏婧竹<sup>2</sup>, 潘建斌<sup>3</sup>

(1. 湖南警察学院 刑事科学技术系, 湖南 长沙 410138; 2. 中科云析(南京)科技有限公司, 江苏 南京 210039; 3. 南京大学 化学化工学院, 江苏 南京 210046)

**摘要:** 利用差分拉曼光谱结合机器学习技术, 对市场上常见的43种不同品牌和型号的蓝色中性笔色料进行分类研究。首先通过差分拉曼光谱法获得样品的拉曼特征峰, 并基于差分拉曼光谱, 将样本人工分为4大类。随后采用聚类分析和随机森林算法进行分类研究, 并与人工分类结果进行对照。结果显示, 随机森林建立的分类模型的预测准确率为90.00%。差分拉曼光谱法可快速、无损分析蓝色中性笔的色料成分, 结合随机森林算法建立分类模型, 可实现快速分类。随机森林模型提高了差分拉曼光谱鉴别未知化合物的效率以及准确度, 能够更好地对样品的类别进行分类与预测。

**关键词:** 光谱; 差分拉曼; 蓝色中性笔; 机器学习; 随机森林

**中图分类号:** O657.3 **文献标识码:** A **文章编号:** 1004-4957(2025)12-2526-09

## Classification of Blue Neutral Pens Using Machine Learning Aided Differential Raman Spectroscopy

HUANG Juan-juan<sup>1\*</sup>, XIANG Wan-xin<sup>1</sup>, XIA Jing-zhu<sup>2</sup>, PAN Jian-bin<sup>3</sup>

(1. Department of Criminal Science and Technology of Hunan Police Academy, Changsha 410138, China; 2. Zhong Ke Yun Xi(Nanjing) Technology Co., Ltd, Nanjing 210039, China; 3. School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210046, China)

**Abstract:** This study employed differential Raman spectroscopy combined with machine learning to classify ink samples from 43 commercially available blue gel pens of different brands and models. Differential Raman spectra were first acquired to identify characteristic Raman peaks. Based on these spectral profiles, samples were manually categorized into four distinct groups. Subsequent classification using cluster analysis and random forest algorithms was compared against manual categorization. The random forest model achieved 90.00% classification accuracy. Differential Raman spectroscopy enables rapid, non-destructive analysis of blue gel pen inks. When integrated with random forest algorithms, it establishes an effective framework for efficient sample classification. The optimized random forest model significantly enhances both the speed and accuracy of identifying unknown ink compounds through differential Raman spectroscopy, demonstrating strong predictive capability for sample categorization.

**Key words:** spectroscopy; differential Raman; blue neutral pen; machine learning; random forest

中性笔色料鉴别技术作为文件检验领域的重要内容, 在司法鉴定实践中具有重要的应用价值。对书写色料的成分及类别鉴定, 不仅能为识别伪造、变造文件以及排查嫌疑对象、缩小侦查范围提供依据, 还可以通过确定的色料类别, 进一步开展文件制成时间鉴定排查。现代中性笔墨水体系主要由着色剂(染料/颜料)、溶剂(苯甲醇/甘油基化合物)、表面活性剂(聚氧乙烯醚类)及树脂(聚氨酯/丙烯酸酯)构成<sup>[1]</sup>。不同厂商通过专利配方(如Pilot的凝胶墨水技术<sup>[2]</sup>、Uni-ball的防伪颜料分散体系<sup>[3]</sup>)形成独特成分组合。研究表明, 市售主流品牌可检出3~5种特征染料标记物, 如三芳甲烷类染料在480 nm处呈现特征吸收峰。此外, 树脂交联度存在时变特征, 丙烯酸酯树脂在1 735 cm<sup>-1</sup>处的羰基峰强度随老化呈指数衰减, 可作为文件制成时间的鉴定依据<sup>[4]</sup>。

收稿日期: 2025-01-02; 修回日期: 2025-02-17

\* 通讯作者: 黄娟娟, 硕士, 教授, 研究方向: 文件检验, E-mail: 273207702@qq.com

网络首发时间: 2025-09-02

中性笔色料的鉴定研究经历了几个阶段,从传统的化学分析方法到现代的色谱和光谱技术,再到更精确的质谱分析技术,检验手段不断丰富。早期的中性笔色料鉴定主要依赖于传统的化学分析方法,如薄层色谱法。该方法虽然操作简单,但损坏检材,且分离效果受限于色料的性质,难以实现高通量分析<sup>[5]</sup>。色料鉴别中常见的色谱分析技术主要包括气相色谱和高效液相色谱法。但常见方法如裂解气相色谱法检测灵敏度较低<sup>[6]</sup>,高效液相色谱法的分析时间较长<sup>[7-8]</sup>,在定性和定量分析上存在一定的局限性。质谱法则因样品处理复杂和设备成本高等问题,导致应用范围受限。随着光谱技术的发展,傅里叶变换红外光谱(FTIR)、拉曼光谱等技术逐渐被引入到中性笔色料的鉴定中。FTIR 方法样本制作复杂,也有损检材<sup>[9]</sup>,相比之下,拉曼光谱对分子振动具有敏感性,能够提供丰富的分子信息,可以有效区分不同类型的色料,并且具有非破坏性和快速分析的优点<sup>[10]</sup>。然而,传统拉曼光谱在复杂样品中可能受到荧光干扰,影响结果的准确性。

差分拉曼光谱法作为一项新兴技术,使用两种波长光源对样品进行照射,获取两张光谱,理想情况下对两张谱图差分处理获得的差分谱中,受激发射谱和噪声背景会完全抵消,拉曼光谱与自身平移光谱的差分图像余留,随后通过去噪解卷积算法将拉曼光谱还原,即可得到物质的纯净谱图<sup>[11]</sup>,可有效减少荧光干扰。目前差分拉曼光谱技术已应用于纸质快递文件袋、白色购物纸袋等材料的分析<sup>[12-13]</sup>,但暂无关于该技术运用于中性笔油墨分析的研究。机器学习是一种基于数据的人工智能技术,通过机器学习已有数据中的规律,训练并建立分析模型从而实现对未来未知样本预测的功能<sup>[14]</sup>,机器学习已与质谱、色谱、光谱<sup>[15-18]</sup>等数据相结合,在解决分类问题上发挥了巨大优势。随机森林(RF)属于机器学习中的集成学习算法,可以处理特征维度高、样本量小的数据,并且对于异常值的鲁棒性较好,对特征的多重共线性也具有高容忍性。

本研究借助差分拉曼光谱技术,对市面上常见的 43 种不同品牌蓝色中性笔色料进行光谱采集和分析,并通过人工直观分类与随机森林算法分类结果的对比研究,为中性笔色料的鉴定和快速溯源提供了更优化、高效和准确的实施方案。

## 1 实验部分

### 1.1 实验器材及条件

SEDRS Portable-Base 便携式差分拉曼光谱仪(南京简智仪器设备有限公司),激光光源采用双频输出( $\lambda_1$ : 784.5 nm $\pm$ 0.5 nm,  $\lambda_2$ : 785.5 nm $\pm$ 0.5 nm),单频输出功率 $\leq$ 500 mW,线宽 $\leq$ 0.06 nm,测试光谱范围 250~2 800 cm<sup>-1</sup>,测量功率为 440 mW,积分时间设置为 3 000 ms,置信度 60%~100%(置信度是统计推断时总体参数落入置信区间内的概率,用于描述特征峰的可信程度,筛选差分的假信号,置信度低则可能为假的差分拉曼信号)。

### 1.2 样品制备

收集市场上销售的不同品牌、系列的蓝色中性笔,共计 43 种,见表 1。分别使用载玻片作为载体,用不同品牌、系列的蓝色中性笔在载体上涂写色料,阴干后测定。

表 1 蓝色中性笔样品表  
Table 1 Blue neutral pen sample table

| Number | Brand | Model       | Manufacturer   |
|--------|-------|-------------|----------------|
| 1      | 希普顿   | 009         | 安徽誉洲贸易实力供应商    |
| 2      | 晨光    | SKP61118    | 南京市经开区晨晨文体用品商行 |
| 3      | 东米    | DM-93       | 日本东米文具         |
| 4      | 無印良品  | Z026        | 济南市惠城国际贸易有限公司  |
| 5      | TEST2 | GP-380      | 义乌市立乐文具厂       |
| 6      | PILOT | G-1.100     | 日本百乐 PILOT     |
| 7      | PILOT | Juice up 05 | 日本百乐 PILOT     |
| 8      | 东米    | DM-902      | 日本东米文具         |
| 9      | 真彩    | 009         | 真彩文具股份有限公司     |
| 10     | 芮翔    | 5010        | 浙江芮翔文具有限公司     |
| 11     | 晨光    | K-35        | 上海晨光文具股份有限公司   |
| 12     | 晨光优品  | AGPA1701    | 上海晨光文具股份有限公司   |
| 13     | PILOT | LJU-10EF-SB | 日本百乐 PILOT     |

(续表 1)

| Number | Brand | Model      | Manufacturer |
|--------|-------|------------|--------------|
| 14     | PILOT | P-500      | 日本百乐 PILOT   |
| 15     | DS 点石 | DS-904 深海蓝 | 青岛点石文具用品有限公司 |
| 16     | 东米    | DM-600     | 日本东米文具       |
| 17     | 点石    | DS-911     | 青岛点石文具用品有限公司 |
| 18     | 点石    | DS-904 蓝   | 青岛点石文具用品有限公司 |
| 19     | 点石    | DS-067     | 青岛点石文具用品有限公司 |
| 20     | 点石    | DS-0163    | 青岛点石文具用品有限公司 |
| 21     | 斑马    | JJ15       | 日本斑马株式会社     |
| 22     | 斑马    | JJ77       | 日本斑马株式会社     |
| 23     | 斑马    | JJM88      | 日本斑马株式会社     |
| 24     | 三菱    | UMN-105    | 日本三菱铅笔株式会社   |
| 25     | 三菱    | UMN-207    | 日本三菱铅笔株式会社   |
| 26     | 派通    | BLN105     | 日本派通文具有限公司   |
| 27     | 派通    | BLN75      | 日本派通文具有限公司   |
| 28     | PILOT | G-1        | 日本百乐 PILOT   |
| 29     | PILOT | G-6        | 日本百乐 PILOT   |
| 30     | PILOT | 10EF JUICE | 日本百乐 PILOT   |
| 31     | 得力    | S60        | 得力集团有限公司     |
| 32     | 得力    | S08        | 得力集团有限公司     |
| 33     | 得力    | AXCEED-S01 | 得力集团有限公司     |
| 34     | 得力    | 6600ES     | 得力集团有限公司     |
| 35     | 白雪    | P1500      | 青岛昌隆文具有限公司   |
| 36     | 白雪    | P1500A     | 青岛昌隆文具有限公司   |
| 37     | 白雪    | PVR-155    | 青岛昌隆文具有限公司   |
| 38     | 白雪    | PVN-166    | 青岛昌隆文具有限公司   |
| 39     | 白雪    | G-101      | 青岛昌隆文具有限公司   |
| 40     | 晨光    | GP-1280    | 上海晨光文具股份有限公司 |
| 41     | 晨光    | AGPA9204   | 上海晨光文具股份有限公司 |
| 42     | 晨光    | GP-1008    | 上海晨光文具股份有限公司 |
| 43     | 晨光    | AGP62403   | 上海晨光文具股份有限公司 |

### 1.3 实验方法

**1.3.1 差分拉曼光谱法** 将蓝色中性笔色料均匀涂抹在载玻片上,将载玻片放置在差分拉曼光谱仪激光探头下,调整积分时间和激光功率。实验发现,功率为 440 mW,积分时间为 3 000 ms 时样品的检测效率高且信号稳定。在稳定环境下进行检验测定,每个样本测试 5 次以上。

**1.3.2 重现性实验与均匀性实验** 重现性实验是指用相同方法,同一试验材料,在不同条件下获得的单个结果之间的一致程度,主要考察仪器的稳定性。随机取样,在同样的实验条件下,对样本笔痕相同位置采用差分拉曼光谱仪进行 5 次平行测定实验。

均匀性实验是为了检验样品的均匀性,从总体样本中随机抽取一定量的样品或随机从不同部位取样检查,考察样品的均匀性。

重现性实验和均匀性实验能够有效验证仪器的稳定性和样品的均匀性,确保实验数据客观、准确。

**1.3.3 光谱预处理** 在对数据进行分析之前先对光谱数据进行预处理,所有光谱采集数据均采用归一化处理,使不同样品光谱处于同一量级。采用最大值归一<sup>[17]</sup>,将最大拉曼峰强设置为 1,以消除激光功率波动的干扰。

**1.3.4 中性笔色料拉曼光谱聚类分析** 使用 IBM SPSS Statistics 27 对数据进行系统聚类分析,初步了解不同厂家以及同一厂家不同规格样品间拉曼光谱的聚类情况,选择瓦尔德法并测量平方欧式距离,得到聚类结果树状图。

**1.3.5 机器学习模型建立** 基于预处理后的拉曼光谱数据,结合系统聚类分析结果,使用随机森林算法对样品的类别进行判断与预测。

2 结果与讨论

2.1 差分拉曼与传统拉曼结果比对

图 1 中黄色和蓝色曲线为 33 号样品两次测量的拉曼光谱图，绿色曲线为其差分拉曼光谱图。相较于传统的拉曼光谱图，差分拉曼光谱图中的特征峰更明显，可克服荧光干扰、消除干扰峰，保留较为纯净的拉曼特征峰，灵敏度更高，有效降低了环境因素的影响。

2.2 重现性实验及均匀性实验结果

随机抽取 1 号样本进行重现性实验，结果发现 1 号样本同一部位的拉曼特征峰峰形、峰数基本一致，且峰位相同。计算得到 5 次平行测定的相对标准偏差(RSD)为 4.2%，低于 5%，具有良好的重现性，说明实验环境和仪器稳定性较高。

随机抽取 15 号样本进行均匀性实验，在 5 个不同部位进行测定，结果发现 5 个不同部位的拉曼特征峰峰形、峰数基本一致，且峰位相同。计算得到其 RSD 为 5.8%，表明蓝色中性笔样本具有较好的均匀性，说明样本均匀性良好。

2.3 样品的差分拉曼特征峰

43 种样品的差分拉曼特征峰检测结果见表 2。表 2 显示，不同品牌、型号的蓝色中性笔色料的主要成分差异较大，现有蓝色中性笔色料的主要成分为主色素(此色素溶剂是蓖麻油酸、氧化蓖麻油，成分为盐基青莲和盐基品蓝)及粘性物质。不同品牌型号的样本由于其色料成分、配比不同，导致不同样本之间的差分拉曼特征峰及峰强度之间存在差异。

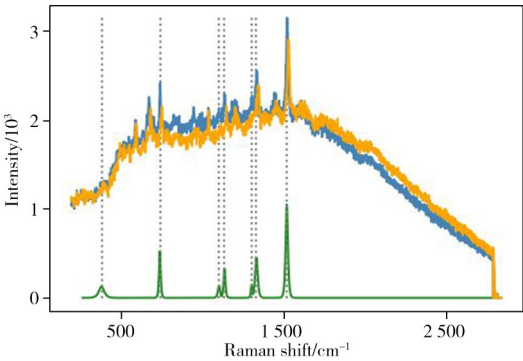


图 1 33 号样本 2 次测量的拉曼光谱图(黄色、蓝色)与差分拉曼图(绿色)比对图

Fig. 1 Comparison of Raman spectra of two measurements(yellow, blue) and differential Raman diagram (green) for sample No. 33

表 2 样本差分拉曼特征峰  
Table 2 Differential Raman shift of samples

| No. | Raman shift/cm <sup>-1</sup>                                                                            |
|-----|---------------------------------------------------------------------------------------------------------|
| 1   | 477, 588, 673, 741, 770, 947, 1 105, 1 140, 1 189, 1 212, 1 306, 1 337, 1 450, 1 525                    |
| 2   | 424, 518, 721, 794, 914, 974, 1 172, 1 363, 1 451, 1 590, 1 624                                         |
| 3   | 479, 587, 674, 741, 1 105, 1 141, 1 188, 1 307, 1 336, 1 450, 1 524                                     |
| 4   | 479, 587, 674, 742, 772, 948, 1 141, 1 200, 1 306, 1 338, 1 388, 1 451, 1 527                           |
| 5   | 477, 672, 740, 766, 946, 1 103, 1 139, 1 187, 1 331, 1 445, 1 518, 1 592                                |
| 6   | 417, 679, 1 175, 1 613                                                                                  |
| 7   | 588, 775, 949, 1 142, 1 342, 1 453, 1 532                                                               |
| 8   | 478, 589, 674, 742, 770, 825, 948, 1 141, 1 189, 1 214, 1 307, 1 337, 1 412, 1 450, 1 525               |
| 9   | 478, 589, 673, 741, 946, 1 105, 1 140, 1 187, 1 306, 1 334, 1 448, 1 521                                |
| 10  | 587, 673, 741, 1 105, 1 140, 1 187, 1 213, 1 334, 1 448, 1 521                                          |
| 11  | 478, 588, 673, 741, 826, 947, 1 104, 1 140, 1 186, 1 215, 1 307, 1 334, 1 448, 1 522                    |
| 12  | 587, 673, 740, 767, 946, 1 103, 1 139, 1 185, 1 212, 1 307, 1 333, 1 447, 1 520                         |
| 13  | 775, 951, 1 108, 1 143, 1 195, 1 217, 1 309, 1 344, 1 455, 1 534                                        |
| 14  | 479, 589, 676, 774, 949, 1 107, 1 142, 1 203, 1 309, 1 342, 1 454, 1 532, 1 613                         |
| 15  | 383, 1 180                                                                                              |
| 16  | 590, 675, 742, 832, 948, 1 004, 1 106, 1 142, 1 338, 1 451, 1 526                                       |
| 17  | 530, 1 227, 1 622                                                                                       |
| 18  | 1 173, 1 283, 1 623                                                                                     |
| 19  | 1 174, 1 622                                                                                            |
| 20  | 1 172, 1 622                                                                                            |
| 21  | 309, 476, 583, 672, 739, 769, 946, 1 001, 1 104, 1 139, 1 203, 1 306, 1 338, 1 387, 1 430, 1 451, 1 528 |
| 22  | 476, 585, 672, 739, 1 138, 1 201, 1 306, 1 338, 1 387, 1 529                                            |
| 23  | 372, 476, 584, 671, 739, 770, 859, 946, 1 103, 1 138, 1 202, 1 304, 1 337, 1 385, 1 527                 |
| 24  | 671, 738, 1 334, 1 522                                                                                  |
| 25  | 671, 739, 1 138, 1 335, 1 523                                                                           |
| 26  | 303, 526, 663, 1 171, 1 613                                                                             |
| 27  | 287, 531, 1 171, 1 373                                                                                  |
| 28  | 391, 666, 1 171                                                                                         |

(续表2)

| No. | Raman shift/cm <sup>-1</sup>                          |
|-----|-------------------------------------------------------|
| 29  | 384, 554, 640, 1 170, 1 608                           |
| 30  | 672, 739, 1 138, 1 337, 1 525                         |
| 31  | 738, 1 137, 1 333, 1 449, 1 521                       |
| 32  | 531, 669, 1 333, 1 522, 1 891                         |
| 33  | 314, 374, 737, 944, 1 137, 1 303, 1 332, 1 446, 1 520 |
| 34  | 738, 1 137, 1 334, 1 523                              |
| 35  | 303, 381, 546, 905, 1 586                             |
| 36  | 289, 516, 1 588                                       |
| 37  | 264, 533, 559, 635, 1 589                             |
| 38  | 309, 526, 630, 1 587                                  |
| 39  | 296, 387, 528, 548, 646, 1 588                        |
| 40  | 741, 1 143, 1 340, 1 530                              |
| 41  | 741, 1 340, 1 394, 1 528                              |
| 42  | 586, 742, 776, 1 530, 1 646                           |
| 43  | 672, 739, 1 521                                       |

2.4 样品分类

2.4.1 人工样品分类 通过 Origin 软件绘制样本的差分拉曼光谱数据,可以更直观地显示样本的差分拉曼特征峰。如图 2~图 5 所示,依据主要差分拉曼特征峰,可将实验样本分为 4 大类,见表 3。

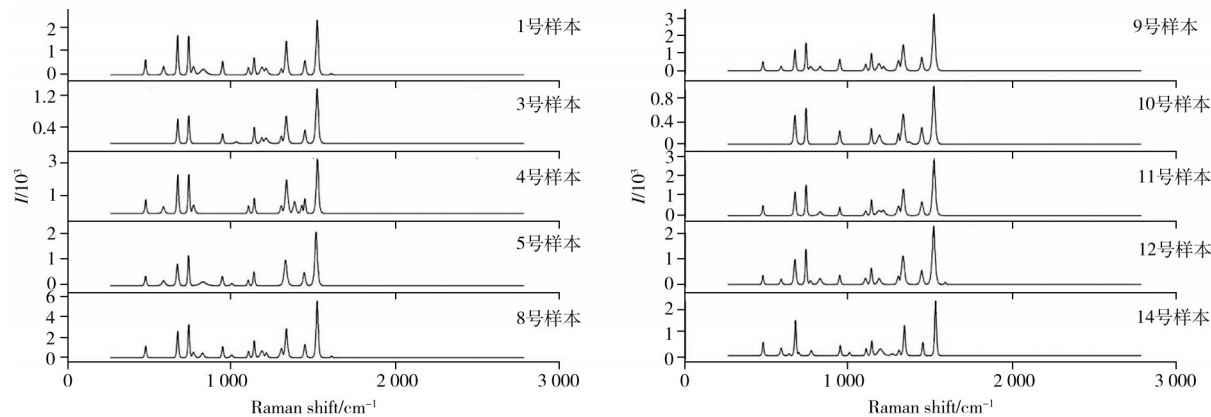


图 2 第 I 类样本的差分拉曼光谱图  
Fig. 2 Differential Raman spectra of Class I samples

研究发现,第 I、IV 类样本在 600~700 cm<sup>-1</sup> 处有较强的拉曼峰,对照表 4,说明其中含有较高含量的异硫氰酸烷基酯类物质;第 II 类样本的特征峰较少,荧光干扰大,说明成分中荧光类物质含量较高;第 I、III、IV 类样本在 1 300~1 550 cm<sup>-1</sup> 处有较强的拉曼峰,说明成分中可能含有反式二烷基乙烯类及正烷烃成分。

经过处理的 43 种蓝色中性笔样本的差分拉曼光谱图,能够更清晰、直观地反映出不同类别蓝色中性笔峰位、峰强等的差异。如从第 I 类样本的差分拉曼光谱图(图 2)中可以直观看出不同样本间大部分的拉曼特征峰峰位重合,各样本之间虽存在微小差异,但并不影响它们属于同一类样本的本质。通过直观的差分拉曼光谱处理图可以对样品进行快速分类,便于进一步鉴别不同种类的蓝色中性笔。

根据该分类方法,可以看出在第 I 类样本中有 3 号和 8 号样本同属东米品牌,11、12 号同属晨光品牌,但第 III 类样本中的 16 号样本也属于东米品牌,第 IV 类样本中的 2 号样本同属于晨光品牌。同样的情况出现在第 III 类样本中,其中 7、13 号样本同属 PILOT 品牌,但是第 I 类样本中的 14 号样本、第

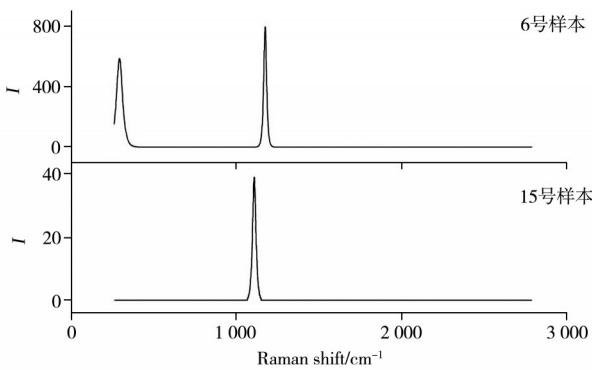


图 3 第 II 类样本的差分拉曼光谱图  
Fig. 3 Differential Raman spectra of Class II samples

Ⅱ类样本中的6号样本也属于PILOT品牌。

从以上结果可知，同品牌蓝色中性笔并不完全属于同一类样品，根据拉曼光谱处理图直观分类并不能准确无误地分类出蓝色中性笔的品牌。因为拉曼光谱处理图直观表现的是蓝色中性笔色料的成分所反映出的拉曼特征峰，即使是同一品牌的蓝色中性笔，型号不同，其拉曼特征处理图将表现出较大的差异。

**2.4.2 基于峰高比进行分类** 虽然同类样本的差分拉曼特征峰有较高的重合性，但其特征峰峰位之间的峰高存在显著差异，如最高峰与次高峰的峰高比存在差异。因此可以依据峰高比进行更加精细的分类分析<sup>[19]</sup>。

样本差分拉曼特征峰的位置主要集中在1 500~500 cm<sup>-1</sup>，峰位数较多，规律性不强，但其峰强有一定规律。这是因为，虽然不同样品的成分组成不同，但主要成分的种类及配比有一定相似性。峰强主要与物质种类及含量相关，因此，可通过比对特征峰强度、峰面积或峰高比来进一步区分同类蓝色中性笔。

为了进一步分类蓝色中性笔样本，基于43种蓝色中性笔样本的拉曼特征峰数据以及拉曼光谱处理图，通过计算第Ⅰ类样品588 cm<sup>-1</sup>和1 306 cm<sup>-1</sup>处峰高(*h*<sub>1</sub>和*h*<sub>2</sub>)的比值进行区分，结果见表5(以随机抽取的1、4、5、8号样品为例)。由表5可以看出，同大类样品之间的峰高比值差异较大，可以将样品进行更细致地区分。

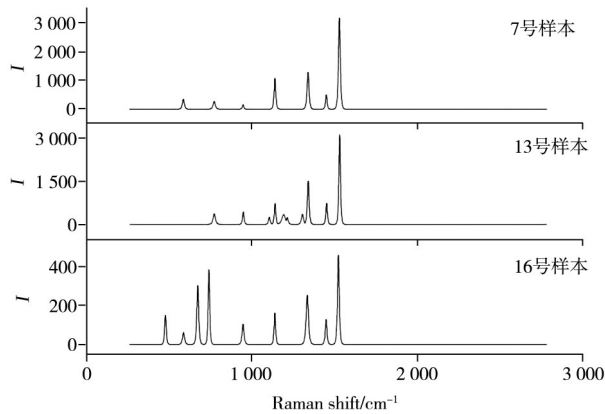


图4 第Ⅲ类样本的差分拉曼光谱图  
Fig. 4 Differential Raman spectra of Class Ⅲ samples

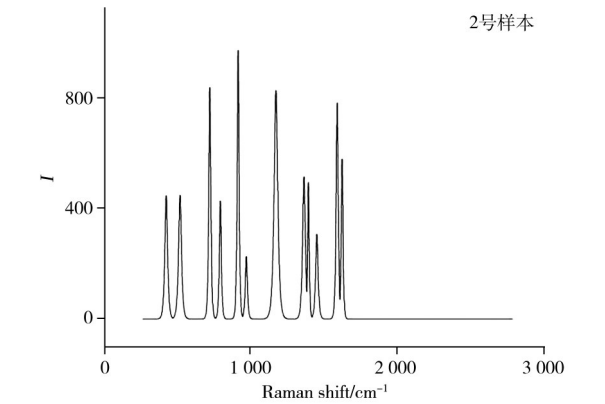


图5 第Ⅳ类样本的差分拉曼光谱图  
Fig. 5 Differential Raman spectra of a Class Ⅳ samples

表3 样品分类表

Table 3 Sample classification table

| Classification | Sample                                                                        |
|----------------|-------------------------------------------------------------------------------|
| 第Ⅰ类            | 1, 3, 4, 5, 8, 9, 10, 11, 12, 14                                              |
| 第Ⅱ类            | 6, 15, 17, 18, 19, 20, 24, 26, 27, 28, 29, 35, 36, 37, 38, 39, 40, 41, 42, 43 |
| 第Ⅲ类            | 7, 13, 16, 31, 32, 33, 34                                                     |
| 第Ⅳ类            | 2, 21, 22, 23, 25, 30                                                         |

表4 常见拉曼特征频率表

Table 4 Table of common Raman feature frequencies

| Raman characteristic peak frequency/cm <sup>-1</sup> | Chemical compound | Raman characteristic peak frequency/cm <sup>-1</sup> | Chemical compound |
|------------------------------------------------------|-------------------|------------------------------------------------------|-------------------|
| 1 634~1 622                                          | 硝酸烷基酯类            | 1 130~1 100                                          | 丙二烯类              |
| 1 616~1 571                                          | 氯代烯类              | 1 010~990                                            | 单, 间位和1, 3, 5-取代苯 |
| 1 473~1 446                                          | 正烷烃类              | 914                                                  | 四氢呋喃              |
| 1 450~1 400                                          | 异氰酸酯类             | 835~749                                              | 异丙基               |
| 1 390~1 370                                          | 萘类                | 830~720                                              | 对位二取代苯            |
| 1 375~1 360                                          | 仲硝基烷类             | 825~820                                              | 仲醇类               |
| 1 350~1 330                                          | 异丙基               | 785~700                                              | 烷基环己烷类            |
| 1 314~1 290                                          | 反式二烷基乙烯类          | 760~650                                              | 叔丁基               |
| 1 305~1 295                                          | 正烷烃               | 760~730                                              | 叔醇类               |
| 1 220~1 200                                          | 单一或1, 2二烷基环丙烷类    | 730~720                                              | 伯氯代烷类             |
| 1 196~1 188                                          | 硫酸烷基酯类            | 690~650                                              | 异硫氰酸烷基酯类          |
| 1 188                                                | 环丙烷               | 520~510                                              | 叔溴代烷类             |
| 1 145~1 125                                          | 二烷基砷类             | 484~475                                              | 二烷基乙炔类            |
| 1 140                                                | 呋喃                | 425~150                                              | 正烷烃类              |
| 1 150~950                                            | 正烷烃类              |                                                      |                   |

虽然通过峰高比可以区分出同类蓝色中性笔样本,但无法精准地区分品牌,这是因为同类品牌不同型号的蓝色中性笔,其成分和色料比例有较大差距。故峰高比只能用来判断蓝色中性笔的牌号是否相同。

由上述分类结果可知,人工分类的优点是快速、直观,能有效划分光谱差异显著的大类,但存在主观性强、对大类内细微差异不敏感、无法利用峰强度信息等问题,且不能依赖品牌进行准确分类。峰高比分类通过引入量化指标,能揭示人工分类无法区分的大类内部样本间的成分比例差异,但需人工选择特征峰,仅适用于特定大类内部,同样无法精准区分不同品牌,只能辅助判断具体牌号是否相同。上述结果表明中性笔墨水的拉曼光谱特征具有显著的多样性和复杂性,同品牌不同型号的产品可能具有完全不同的光谱特征,或在大类内具有显著不同的峰高比,品牌不能作为可靠的分类依据,成分才是关键。人工分类提供了有用的初步分组,峰高比提供了在大类内进行更细分的量化途径,但两者都存在显著局限性。

## 2.5 机器学习模型建立与预测

为克服人工分类和基于峰高比分类的主观性和局限性,并寻求更客观、高效的自动化分类方法,采用机器学习技术对样本进行分类研究。首先尝试通过无监督的聚类分析探索数据内在结构,随后采用有监督的随机森林算法建立基于人工分类结果的预测模型,并对两者的效果进行了比较评估。

**2.5.1 中性笔色料拉曼光谱聚类分析** 对样品信息进行一定归属后再进一步区分的分类结果受初期样品类别归属的准确性、科学性以及所选参考峰的影响较大,且中性笔油墨是一类复杂的混合物,不同色料既有相同成分,也存在差异,这给直观分类带来了挑战。因此选择机器学习的方法,依据墨水样品拉曼光谱特征的相似或相异程度进行分类<sup>[20]</sup>,以探索样本聚类趋势,克服人工分类主观性强、准确性低、耗时长的问题。

图6为中性笔色料的拉曼光谱系统聚类分析图,由图可知,当平方欧氏距离为1时样品分为43类,当平方欧氏距离为2、4、5、12时样品可分别分为41、17、6、1类。平方欧氏距离小于5时样品分类数目较多,大于5时则分类数过少,均无法满足分类需要。平方欧氏距离为5时分类效果较好且能满足分类需要,此时43个样品分成6类。点石、白雪等品牌内部成分高度相似,聚类结果与品牌一致;PILOT、晨光等品牌因产品线复杂(如不同型号/配方),光谱差异大,导致聚类分散。与人工分类结果相比,聚类分析仅依赖光谱相似性进行分类,不关注品牌标签。即使不同品牌,若配方相似(如均含特定树脂),也可能被归为同一类;反之,同品牌不同型号可能因配方差异被分开。此聚类结果反映了光谱数据本身潜在的相似性分组,仅靠光谱相似性不足以区分品牌,但可为后续特征选择提供参考。

**2.5.2 聚类分析结果验证** 偏最小二乘判别分析(PLS-DA)是一种有监督的判别分析方法,能够按照预先定义类别最大化组间差异,获得较好的分类效果。采用经典的偏最小二乘回归模型,对特征变量进行回归分析,可实现样本区分。本文采用PLS-DA对系统聚类分析结果进行验证,并利用PLS-DA对未知化合物样本的类别进行预测。

表6中LV表示PLS-DA提取的潜变量,1~4分别代表前4个LV(按解释方差降序排列)。 $R^2$ 和 $Q^2$ 分别为当前LV对光谱数据的方差解释率( $R^2X$ )和交叉验证预测能力( $Q^2$ ),Cum行表示前4个LV的累积值。理论上 $R^2$ 和 $Q^2$ 越接近1.0说明模型的拟合程度越好。通常情况下 $R^2$ 和 $Q^2$ 高于0.4即可接受<sup>[21]</sup>,二者之差不应过大。表6显示,构建的模型 $R^2(\text{cum})$ 为0.589、 $Q^2(\text{cum})$ 为0.056 2。43批样品的PLS-DA

表5 第I类部分样品峰高比  
Table 5 Relative peak-to-height ratio of Class I samples

| No. | $h_1$ | $h_2$ | $h_1/h_2$ |
|-----|-------|-------|-----------|
| 1   | 349   | 439   | 0.745     |
| 4   | 401   | 599   | 0.669     |
| 5   | 200   | 63    | 3.175     |
| 8   | 577   | 900   | 0.641     |

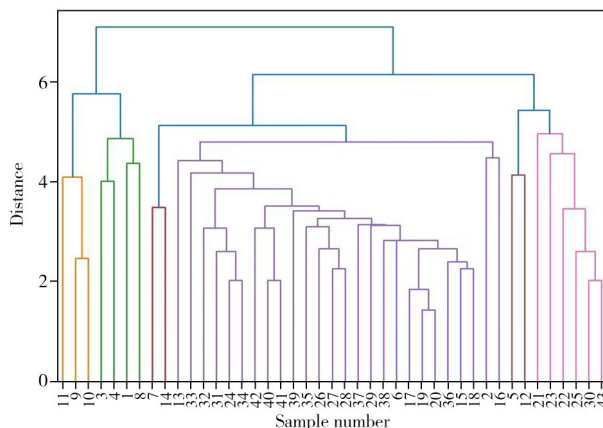


图6 中性笔色料拉曼光谱系统聚类分析图  
Fig. 6 The result of sample system cluster analysis

分析结果如图 7 所示, 由图 7 可知 1、2 和 3、4、5、6 类样本位于不同区域, 说明 1、2 和 3、4、5、6 类样本之间存在较大差异, 而 1 组样本和 2 组样本距离较近, 未完全区分; 200 次置换检验的结果(图 7B)表明模型可靠, 不存在过拟合现象。结果表明, 基于聚类分析得到的 6 类分组所建立的 PLS-DA 模型对数据的拟合度较好, 但预测能力较差, 难以有效用于未知样本的类别预测。这说明无监督聚类得到的 6 类分组在构建稳健的分类预测模型方面存在不足。

偏最小二乘分析体现的样品分类趋势同聚类分析基本一致, 但偏最小二乘判别分析算法难以捕捉特征之间的非线性关系, 当特征数量增多时分类效果变差<sup>[22]</sup>, 因此尝试采用随机森林算法做进一步区分。

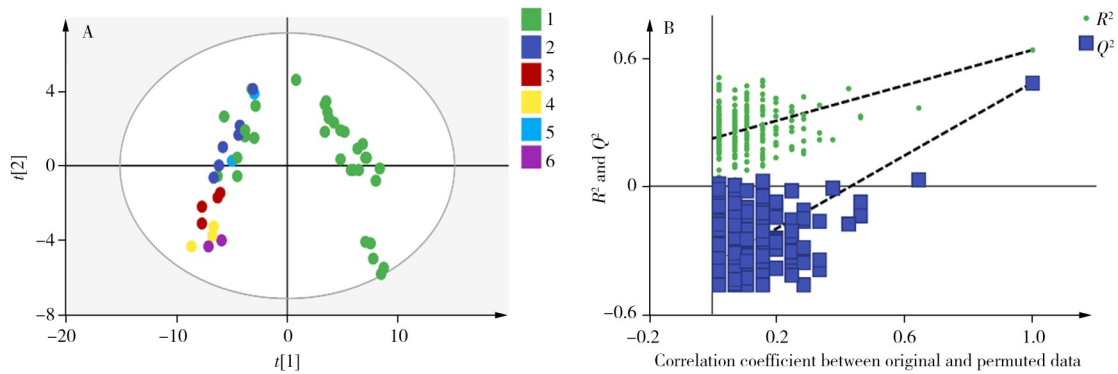


图 7 偏最小二乘判别分析

Fig. 7 Partial least squares discriminant analysis

A. sample score chart, B. results of permutation test

**2.5.3 随机森林算法分类预测结果** 鉴于聚类分析路径在建立有效预测模型上的局限性, 采用有监督的随机森林算法, 并直接基于人工分类确定的 4 种类别(见表 3)作为训练目标标签, 以建立能够预测样本所属类别的模型。随机森林模型是一种基于自主重采样技术的非线性数据处理算法, 利用 Bagging 方法构建回归树分类器<sup>[22-23]</sup>。将 43 组数据按照 8 : 2 的比例划分训练集和测试集, 建立 RF 分类模型, 其中 trees 为最大决策树数量, 设置为 10, minleaf 是模型中的最小叶子数, 设置为 2。随机森林模型对样品拉曼光谱测试集的预测准确率为 90.00%, 测试集混淆矩阵结果如图 8 所示, 可以看出第 1 类、第 2、第 3 类样品的预测准确率最高达 100%, 表明随机森林模型能够较好地对人工定义的 4 个类别进行分类与预测。与 PLS-DA 预测模型相比, 基于人工分类标签建立的随机森林模型展现出优越的分类预测性能。这表明, 本研究中利用有监督的随机森林算法直接学习人工分类的先验知识, 是比依赖无监督聚类分组更有效的构建分类预测模型的途径。

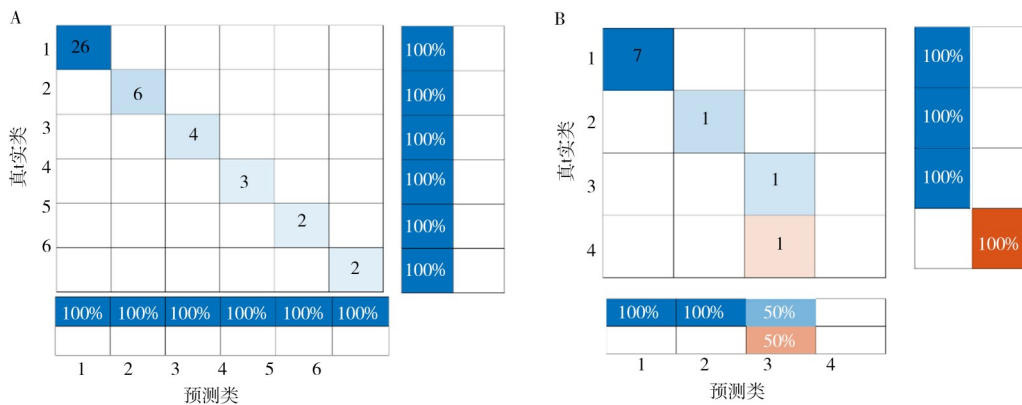


图 8 随机森林模型预测结果

Fig. 8 Prediction results of random forest model

A. confusion matrix of training set, B. confusion matrix of test set

### 3 结 论

本文使用差分拉曼光谱仪对蓝色中性笔色料进行检测,并基于特征峰的位置和峰高比的差异对43种不同品牌、型号的蓝色中性笔进行分类处理,证明了差分拉曼光谱法在分析鉴别蓝色中性笔色料种类中方便、快速、便捷、无损的优势。机器学习可以减少人工分类主观性的影响,提高差分拉曼光谱鉴定未知化合物类别的准确性,增强鉴别的可信度,也为差分拉曼光谱在高通量鉴别、预测未知化合物类别方面提供了强有力的支持。

#### 参考文献:

- [1] He H Y, Zhao P C, Li Z H, Gao J. *Shandong Chem. Ind.* (何海洋, 赵鹏程, 李泽瀚, 高静. 山东化工), **2021**, 50(7): 103-105.
- [2] Mitsubishi Pencil Co., Ltd. Japan Patent, US6132124A. 2000.
- [3] Mitsubishi Pencil Co., Ltd. Japan Patent, CN202280015331.5. 2023.
- [4] Zhang W, Liang Y, Zhao M, Tan Z, Li G. *J. Appl. Polym. Sci.*, **2023**, 40(3): 045023.
- [5] Huang J J, Mei Y, Lin D. *Phys. Test. Chem. Anal.: Chem. Anal.* (黄娟娟, 梅毅, 林岱. 理化检验-化学分册), **2008**, (8): 792-793.
- [6] Ye Z L, Wang C L, Cui L. *Chin. J. Forensic Sci.* (叶卓龙, 王长亮, 崔岚. 中国司法鉴定), **2019**, (3): 40-44.
- [7] Zhao P C, Wang Y J, Zhao Y J, Wang Y, Li X Q, Yao L J, Wang J H. *Chin. J. Chromatogr.* (赵鹏程, 王彦吉, 赵彦军, 王岩, 李心情, 姚丽娟, 王景翰. 色谱), **2007**, (5): 762-765.
- [8] Chen D W, Li B. *J. Yunnan Police College* (陈丁文, 李斌. 云南警官学院学报), **2008**, (3): 93-96.
- [9] Huang J J, Liu Z L, Mei Y. *Mod. Sci. Instrum.* (黄娟娟, 刘忠良, 梅毅. 现代科学仪器), **2009**, (3): 96-98, 101.
- [10] Hu X Y, Wang Y Z. *J. Instrum. Anal.* (胡晓燕, 王元忠. 分析测试学报), **2024**, 43(11): 1709-1724.
- [11] Chen L P, Tang Y X, Jiang H, Wang Y Y, Zhang X, Duan B, Liu F. *Packag. Eng.* (陈丽萍, 唐宇璇, 姜红, 王圆圆, 章欣, 段斌, 刘峰. 包装工程), **2023**, 44(5): 196-202.
- [12] Zhou G X, Jiang H, Hu X G, Zhang X, Huang K. *Phys. Test. Chem. Anal.: Chem. Anal.* (周贯旭, 姜红, 胡晓光, 章欣, 黄凯. 理化检验-化学分册), **2024**, 60(4): 418-422.
- [13] Xie J N, Hu X G, Jiang H, Zhang X, Huang K. *Packag. Eng.* (谢佳宁, 胡晓光, 姜红, 章欣, 黄凯. 包装工程), **2024**, 45(1): 215-222.
- [14] Bi Q, Goodman K E, Kaminsky J, Lessler J. *Am. J. Epidemiol.*, **2019**, 188(12): 2222-2239.
- [15] Xia B H, Hu Y Z, Xiong S H, Tang J, Yan Q Z, Lin L M. *China J. Chin. Mater. Med.* (夏伯候, 胡玉珍, 熊苏慧, 唐洁, 闫庆梓, 林丽美. 中国中药杂志), **2017**, 42(7): 1324-1330.
- [16] Wang S, Chen L, Zhou R R, Zhang S H, Yu Y, Wang L. *J. Anal. Sci.* (王升, 陈林, 周融融, 张水寒, 余意, 王凌. 分析科学学报), **2019**, 35(1): 85-89.
- [17] Zhang S P, Tan Z L, Liu J, Xu Z H, Du Z. *Spectrochim. Acta A*, **2020**, 228: 117781.
- [18] Chen J, Liu M, Yuan H, Huang S, Zhao J, Tao J, Xu N. *Vib. Spectrosc.*, **2018**, 99: 7-12.
- [19] Zhang H, Gao S H. *J. Instrum. Anal.* (张浩, 高树辉. 分析测试学报), **2023**, 42(7): 817-824.
- [20] Wang M. *Study of Algorithms for Clustering Categorical Data*. Zhenjiang: Jiangsu University (王敏. 分类属性数据聚类算法研究. 镇江: 江苏大学), **2008**.
- [21] Liu Z P, Zhang D, Huang Y. *Sci. Technol. Food Ind.* (刘振平, 张丹, 黄瑶. 食品工业科技), **2023**, 44(16): 306-312.
- [22] Gislason P O, Benediktsson J A, Sveinsson J R. *Pattern Recognit. Lett.*, **2006**, 27(4): 294-300.
- [23] Chen Y Y, Xia J J, Wei Y, Xu W X, Mao X R, Chen Y F, Min S G, Xiong Y M. *Chin. J. Anal. Chem.* (陈玥瑶, 夏静静, 韦芸, 徐惟馨, 毛欣然, 陈月飞, 闵顺耕, 熊艳梅. 分析化学), **2023**, 51(3): 454-462.

(责任编辑: 盛文彦)