基于显微聚焦拉曼光谱技术的丹参产地鉴别研究

李 庆^{1,2},许 莉^{1,2},彭善贵^{1,2},罗 霄^{1,2},张蓉琴^{1,2},严铸云³,文永盛^{1,2*}

1. 成都市药品检验研究院,四川成都 610045

2. 国家药品监督管理局中药材质量监测评价重点实验室,四川成都 610045

3. 成都中医药大学,四川成都 611137

摘 要产地是影响中药材质量的重要因素,产地差异导致中药材质量参差不齐,为维护市场秩序,有必要 建立中药材产地鉴别方法,以便更加精准地判别和分析中药材品质。以多产地临床大宗药材丹参为研究对 象,收集不同产地丹参样品150份,采用显微聚焦拉曼光谱技术在无损条件下对每份丹参样品的每根药材 表面随机扫描1~n次,求每份样品扫描1~n次的平均光谱。分析原始光谱数据发现丹参表面光谱信号同时 包含了丹参酮类成分的拉曼光谱和杂质的荧光光谱,主要表现在特定波长范围内不同产地丹参存在各自的 聚集区和丹参表面光谱信号强度明显弱于或强于丹参酮类对照品的拉曼光谱信号强度。对扫描1~n次的平 均光谱数据进行预处理后运用偏最小二乘判别分析(PLS-DA)和随机森林分类算法[不筛选(RF)或筛选重要 变量(RF-VS)]建立扫描1~n次的丹参产地分类模型。结果随机扫描1次所得最优模型训练集和测试集预 测准确率分别为88%和87%,且对质量差和质量优的丹参样品区分准确率高达97%;随机扫描2次和3次 所得最优模型训练集和测试集预测准确率均分别为89%和87%,结合模型运行效率和成本,选择随机扫描 1次所得光谱,经一阶导数(1ST-D)预处理和RF-VS计算所得模型为丹参最终产地鉴别模型。综上,在无损 伤条件下显微聚焦拉曼光谱技术能建立快速、准确的丹参产地鉴别预测模型,为该技术进一步用于贵细中 药材的产地和真伪鉴别提供参考。

关键词 拉曼光谱; 丹参; 产地鉴别 中图分类号: O657.3 文献标识码: A

DOI: 10.3964/j.issn.1000-0593(2022)06-1774-07

引 言

中药材产地识别是中医药界保证临床用药安全有效的重要手段,《神农本草经》明确提出"土地所出,真伪陈新";《本草衍义》谓:"用药必择州土所宜者,则药力具,用之有据";李时珍在《本草纲目》中谓:"性从地变,质与物迁"。丹参为唇形科植物丹参 Salvia miltiorrhiza Bge.的干燥根和根茎,常用于治疗心血管疾病^[1]。丹参除河南、山东、四川传统产地外,陕西、湖北、河北、安徽、山西、江苏、云南和贵州等地也在栽培,不同产地的丹参质量差异大^[2]。常规的性状鉴别、显微鉴别和薄层色谱鉴别难以识别丹参产地。虽然液相色谱^[3]、质谱联用^[4]和分子标记^[5]也用于丹参产地研究,但这些方法耗时、消耗化学试剂,不能满足市场中药材、尤其

是贵细中药材的快速无损伤鉴定需要。

显微聚焦拉曼技术是一种微区分析技术,具快速、无 损、结果直观等优点,已在文物、宝石和生物医学等多个领 域得到应用,也用于中药质量分析^[6-7]。但样品在无损伤条 件下,增加扫描位点以获取样品整体光谱信息,再结合化学 计量学建立中药材产地鉴别模型的研究鲜有报道。本文以丹 参为研究对象,利用显微聚焦拉曼技术对不同产地丹参样品 每根药材表面随机扫描 1~n次,获取每份样品 1~n次的平 均光谱数据,经数据前处理后用偏最小二乘判别分析(partial least squares-discriminant analysis, PLS-DA)和随机森林分 类算法[不筛选(random forest, RF)或筛选重要变量(RF-VS)]建立不同扫描次数的丹参产地识别模型,为该技术应 用于中药材产地鉴别提供参考。

收稿日期: 2021-03-31,修订日期: 2021-09-19

基金项目:四川省重大科技专项(2018TZDZX0007),四川省科技厅重点研发项目(2021YFS0045),国家自然科学基金项目(81973416)资助 作者简介:李 庆,1980年生,成都市食品药品检验研究院检验员 e-mail:liqqin2001@outlook.com

1 实验部分

1.1 样品

2020年9月至12月自7个省采集150份栽培丹参样品, 除杂后于50℃烘干备用。所有样品均经成都中医药大学严 铸云教授鉴定为丹参(Salvia miltiorrhizae radix et rhizoma) 正品,样品详细信息见图1和附表1。



图 1 丹参产地及其代表性样品图

Fig. 1 Origins of danshen and their representative samples

附表 1 150 份丹参样品信息 Attached Table 1 Information of 150 samples of danshen

采样区域 (见图 1)	采样点 (见图 1)	样本数量	土壤类型	地貌
а	粉色点	17	沙质	平原
b	深蓝色点	17	沙质	平原
с	亮绿色点	67	黄棕	丘陵
d	红色点	30	黄棕	丘陵
e	黄色点	19	黄棕	丘陵

根据各采集地的地理位点远近、土壤类型及丹参外观性 状归纳丹参产地。由图 1 可知,河北、四川和山东在区位上 较为独立,分别可归为产地 a, e 和 d。图中亮绿色采样点分 布于河南、山西、陕西,区位上彼此非常接近,将这些采样 点归为产地 c。但河南两个采样点温县、禹州及安徽亳州采 样点(图蓝色位点)均位于黄河下游冲击平原,土壤为沙质 (见附表 1),这三个样点所产丹参外观表面土灰色,明显不 同于产地 c 和 d 的外观暗棕红色的丹参(见图 1),且含量测 定结果(待发表)发现该三个样点的丹参样品丹参酮类含量整 体明显低于产地 c 和 d,因此,将该三个采样点归为产地 b。

1.2 样品

隐丹参酮(110852—201807)、丹参酮 Ⅱ A(110766— 202022, TA2)和丹参酮 Ⅰ(110867—201607)购自中国食品 药品检定研究院,纯度均高于 97%。二氢丹参酮 Ⅰ(MUST-15020102)购自成都曼斯特生物科技有限公司,纯度高 于 98%。

1.3 仪器

DXR 显微聚焦拉曼光谱仪(美国,赛默飞世尔科技公司)。

1.4 数据采集

显微聚焦拉曼光谱仪检测时激光波长 780 nm;激光功 率5 mW;波数范围 50~3 350 cm⁻¹;采集曝光时间 3 s;检 测精度 1 cm⁻¹。考虑到测试成本,在无损伤条件下,本文仅 对每份样品的每根药材表面随机扫描 1~3 次,对不同扫描 次数所得光谱数据求平均,得到每份样品扫描 1 次、2 次、3 次的平均光谱数据。四个丹参酮类对照品分别扫描一次即得 拉曼光谱数据。

1.5 数据处理和模型建立

1.5.1 数据预处理和样本集的划分

为消除基线漂移,降低随机噪音,提取光谱有效信息, 本文使用常用的标准正态变换(standard normal variable transformation, SNV)、多元散射校正(multiplicative scatter correction, MSC)、1ST-D、二阶导数(second derivative, 2ND-D)、三阶导数(third derivative, 3RD-D)对原始光谱进 行前处理^[8]。

随机选取三分之二的样本作为训练集,剩余三分之一的 样本作为测试集。

1.5.2 模型建立

(1)PLS-DA 模型:在全波段条件下,利用 PLS-DA 建立 丹参产地识别模型。采用 7 折交叉验证的交叉验证均方根误 差(RMSECV)的最小值确定最适隐变量数(LVs)。使用 Simca (Version 13.0, Umetrics, Sweden)软件完成 PLS-DA 模 型的建立。

(2) RF 和 RF-VS 模型:应用 RF 建立丹参产地识别模型。对两个重要参数决策树数量(n-estimator)和最大特征数量(Max-feature)分别在 0~500 和 0~120 范围进行优化,选取袋外误差(out-of-bag score, Oob_score)最小的参数作为建模参数。

在全波段条件下建立 RF 模型。同时,随机选取训练集的五分之四样本用来计算变量的重要性,重复 500 次,使用 基尼指数评价变量的重要性;再根据变量的重要性范围,筛 选重要变量,筛选幅度为变量重要性范围的 1/40,使用五折 交叉验证寻找不同的界值,各界值对应的重要变量建立不同 的模型;根据模型平均预测准确率筛选出最优 RF-VS 模型 和相应的最优界值。使用 Python 语言完成 RF 和 RF-VS 模型。

(3)模型评价:采用训练集的预测准确率(accuracy, ACC)和测试集的预测准确率来评价模型区分能力,ACC值越大,模型性能越好。

2 结果与讨论

2.1 光谱分析

图 2(a),(b)和(c)分别为扫描1、2 和3次所得原始光谱 图,图 2(d)为丹参酮类成分丹参酮ⅡA、隐丹参酮、二氢丹 参酮和丹参酮Ⅰ的原始光谱图,波段范围均为 50~ 3 350 cm⁻¹。

由图 2(d)可知,丹参酮类成分在相同的位置具相似的吸收峰,但存在强度差异,如在 2 900,1 550 和 1 620 cm⁻¹位置分别是隐丹参酮、丹参酮 I 和丹参酮 II A 吸收峰最强。丹

参酮类成分各主要特征峰的归属^[9-11]:在吸收位置2900 cm⁻¹及其附近的峰归属于丹参酮类成分中烯烃或苯环的 C—H的伸缩振动,1620和1550 cm⁻¹归属于丹参酮类成分 酮类的 C—O 伸缩振动和烯烃或苯环的 C—C 伸缩振动, 1300和1360 cm⁻¹归属于丹参酮类成分甲基的 C—H 的弯 曲振动,1070和1190 cm⁻¹位置的吸收峰为丹参酮类成分 环氧的 C—O—C 伸缩振动,872 cm⁻¹附近的吸收峰为丹参 酮类成分环氧的 C—O 伸缩振动和烯烃或苯环的 C—H 面外 弯曲振动,705,586,506和270 cm⁻¹位置的吸收峰归属为 丹参酮类成分甲基的 CH 面外弯曲振动和其混合振动。



图 2 丹参表面不同扫描次数和丹参酮类成分的原始光谱 Fig. 2 Original spectra of surface of danshen with different scanning times and tanshinone components

由图 2(a)可知,不同产地丹参的拉曼光谱图彼此之间既 有重叠区又有各自的聚集区,如 2 100~2 800 cm⁻¹范围明显 重叠;而在 1 570~1 630 cm⁻¹范围内,由上至下依次为产地 b,产地 e 部分样品,产地 d,产地 c,产地 a 和产地 e 的部分 样品;处于高波数(大于 3 000 cm⁻¹)和低波数(小于 250 cm⁻¹)范围的不同产地样品同样存在各自聚集区。需要指出 的是,产自 a 和 b 的样品丹参酮类成分含量明显低于其他产 地样品,但在图 2(a)中,产地 a 和 b 丹参样品的拉曼光谱吸 收强度并不弱,反而产地 b 丹参样品的光谱吸收最强,这可 能是由于丹参样品表面除了丹参酮类成分,还含有其他杂质 成分,其所产生的荧光信号占主导,导致丹参酮类成分含量 低的产地 b 的丹参样品表面光谱信号反而更强。

由图 2(a),(b)和(c)可知,不同测定次数下所得平均原 始光谱图十分接近,但也存在细微差异。以1 570~1 630 cm⁻¹范围为例,图 2(a)和(b)中各产地样品的光谱曲线由上 至下的分布较一致,但图 2(c)中产地 d,c和 a 三者重叠在一 起,表明从原始数据看,增加扫描次数可能不能改善产地识 别效果。将图 2(a),(b),(c)与(d)对比,尽管丹参样品表面





Attached Fig. 1 Optimization results of random forest model parameters

(a): The relationship diagram between n_estimator and OOB_SCORE; (b): The relationship diagram between max_features and OOB_SCORE

与丹参酮类成分的拉曼光谱图较相似,但图2(d)中丹参酮类 对照品吸收峰更强更尖锐;另一个差异是在高波数和低波数 区,丹参样品的吸收强度明显大于丹参酮类成分(除丹参酮 IIA外),原因同样是丹参样品表面杂质产生的荧光效应, 减弱或增强了丹参样品表面光谱信号。综上,杂质改变了丹 参样品表面的拉曼光谱信号,不同产地的杂质成分不同,这 有利于丹参表面光谱数据用于产地溯源。

2.2 PLS-DA 模型

6个 PLS-DA 模型的详细结果见附表 2 所示,详细的结果分析见下文。

2.3 随机森林模型的参数优化

对随机森林的两个重要参数 n_estimators 和 max_features 进行优化。最终仅提供扫描一次、1ST-D 预处理后获得 的 RF-VS 模型的参数优化图(见附图 1)。如附图 1 所示,最 佳参数为 300 棵树、最大特征为 32 时的 Oob_score 最小。

2.4 随机森林模型建立

由附表 2 可知,使用 1ST-D 预处理随机扫描 1 次所得光 谱数据,经 RF-VS 计算可得最适模型(训练集和测试集的准 确率分别为 88%和 87%)。以该模型为例,计算 1ST-D 预处 理后的原始光谱数据的 3 215 个变量重要性,见附图 2,可知 变量重要性范围在0.000 009和0.009 775之间,表明不同变 量对模型的产地预测效果存在贡献差异,需提取出重要特征 建模变量。由附表 2 可知,经交叉验证筛选,最适界值为 0.009 207,其对应的变量为 167 个,即可建立最优模型。

2.5 分类模型的比较

附表 2 列出了随机扫描 1 次、2 次和 3 次所得原始数据 和五种数据前处理方法所得数据, 经 PLS-DA, RF 和 RF-VS



附图 2 经 1ST-D 处理后的光谱变量的重要性与波数的关系图 Attached Fig. 2 The relationship between the importance of spectral variables after 1st-d processing and wave numbers

计算的训练集和测试集准确率。

2.5.1 扫描一次条件下各分类模型的比较

由附表 2 可知,在 PLS-DA 建立的模型中,原始数据经 五种数据预处理后所得模型的准确率较原始数据建立的模型 准确率并没有得到较好的改善,这表明数据的前处理过程可 能丢失了更为重要的信息。仅 MSC 预处理后所得模型的准 确率有轻微改善,其训练集准确率 74%,预测集准确率 68%,表明该模型性能一般。由该模型的测试集混淆矩阵表 可知(见附表 3),该模型对质量差的产地 b 样本能 100% 区 分,但对质量差的产地 a 的 4 个样本和质量好的产地 e 的 4 个样本的预测准确率均为 0,需进一步用其他分类算法建立 区分性能更好的模型。

测定次数	模型	数据前处理方法	训练集 ACC	测试集 ACC	主成分数	变量数/个	界值
		Original data	73	68	2	3 216	
		MSC	74	65	3	3 216	
	DISDA	SNV	69	68	2	3 216	
	FLS-DA	1ST-D	74	68	2	3 215	
		2ND-D	69	68	2	3 214	
		3RD-D	69	65	2	3 213	
		Original data	83	87		3 216	0
		MSC	86	87		3 216	0
测合 岁	DE	SNV	82	68		3 216	0
测定一次	КГ	1ST-D	87	81		3 215	0
		2ND-D	87	84		3 214	0
		3RD-D	87	81		3 213	0
		Original data	86	87		87	0.013 837
		MSC	86	87		366	0.008 228
	DE VS	SNV	81	77		295	0.010 197
	KF-V5	1ST-D	88	87		167	0.009 207
		2ND-D	88	81		29	0.014 741
		3RD-D	88	87		198	0.009 325

附表 2 不同测定次数条件下所建模型结果 Attached Table 2 The results of the model under the conditions of different determination times

ム土:	74.3	÷ 6	2
ZT	INPL ->	E 2	

PLATE <br< th=""><th>· 医阳衣 Δ</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></br<>	· 医阳衣 Δ							
PLS-DAMSC746533 216SNV706523 2162ND-D706823 2143RD-D96523 2133RD-D96523 214MSC85813 2160SNV82713 2160SND85873 2160SRD-D88773 21603RD-D88773 2130MSC84773 2130SRD-D88773 2130SRD-D88773 2130.011 960SRD-D88773 2130.081 426MSC87841010.007 782SNV84773100.008 426SNV84773100.008 4262ND-D87873 2130.013 165SNV84773 2130.013 165SNV878823 213PS-DA716823 213MSC746543 216SNV84743 2160.013 165SNN85883 2130MSC81743 2160.013 165SNN84743 2160.01SNN84743 2160.01SNN84743 2160.01SNN84743 216<			Original data	71	68	2	3 216	
PLS-DASNV 1ST-D 1ST-D736623 2161ST-D 3RD-D736823 215御定二次RF696523 213御定二次NSC85843 2160MSC85813 2160MSC82713 21601ST-D88773 21502ND-D88773 21301ST-D88773 21303RD-D88773 2130MSC8784100,007 7821ST-D8784100,007 7822ND-D87873 2140.008 4261ST-D87873 2140.008 4262ND-D8790.013 1652ND-D873 2140.008 4262ND-D878790.013 1652ND-D878790.013 1652ND-D736823 214MSC716523 2162ND-D736823 2141ST-D736823 2162ND-D80813 21603RD-D89873 21603RD-D89813 21603RD-D89813 21603RD-D89813 21603RD-D89813 21603RD-D81			MSC	74	65	3	3 216	
메라이자 15T-D 73 68 2 3 215 2ND D 70 68 2 3 214 2ND D 69 65 2 3 216 0 BCD D 69 65 2 3 216 0 MSC 85 84 3 216 0 MSC 85 84 3 216 0 SNV 82 71 3 216 0 MSC 88 77 3 215 0 ARD D 88 77 3 215 0.011960 MSC 87 3 215 0.011960 MSC 87 3 213 0.00772 RF-VS SNV 81 77 3 213 0.008 426 MSC 87 84 3 215 0.013 65 SNV 81 71 68 2 3 216 MSC 73 68 2 3 216 0.013 155 SNV 71 68		DISDA	SNV	70	65	2	3 216	
행군 행군 (1)2 3 <br< td=""><td></td><td>FLS-DA</td><td>1ST-D</td><td>73</td><td>68</td><td>2</td><td>3 215</td><td></td></br<>		FLS-DA	1ST-D	73	68	2	3 215	
3RD D696523 213예약 (mind data87813 2160MSC85843 2160MSC85713 21601ST-D88773 21302ND-D88773 21303RD-D88773 2130MSC88773 2130.0119603RD-D88773 2130.011960MSC87843 2150.011960MSC87843 2130.008 4261ST-D87843 2130.008 2072ND-D87873 2130.018 426MSC87873 2130.018 4261ST-D87873 2130.018 426MSC746543 216MSC746523 216SNV716823 216SND716823 216MSC746543 216SNV716823 216SND-D696833 213MSC82743 2160SND-D89813 2160MSC813 21600SND-D89813 2160SND-D89813 2160SND-D89813 2160SND-D89813 2160			2ND-D	70	68	2	3 214	
測定 測定 調定 和子Original data MSC87813 2160MSC85843 2160SNV82713 21602ND-D88773 21403RD-D88773 2130MSC88773 2140MSC88773 2140MSC87713 2140MSC873 2130MSC873 2140.001 9601010.008 2073 2130.002 2072ND-D87843 2140.009 2072ND-D878790.014 8202ND-D8787910.014 8202ND-D8787910.014 820SNV716823 216PLS-DA736823 216SND-D686833 213PLS-DA736823 216SND-D686833 213PLS-DA89843 2160SND-D89843 21603RD-D89843 21603RD-D89843 2160SNV81713 2130SNV81713 2160SND-D89843 2160SND-D89843 2160SND-D89843 2160			3RD-D	69	65	2	3 213	
潮定 潮定 水 、 水 、 水 、 水 、 水 、 水 、 水 、 水 、 <br< td=""><td></td><td></td><td>Original data</td><td>87</td><td>81</td><td></td><td>3 216</td><td>0</td></br<>			Original data	87	81		3 216	0
謝定二次 削定二次 削定二次 飛子 肥子 第 化 第 \$ 第 第 第 			MSC	85	84		3 216	0
内圧 (A)KP1ST-D88773 21502ND-D88773 21403RD-D88773 21303RD-D88773 2130MSC87841010.007 782MSC87841010.007 782SNV84773 190.008 4261ST-D87843 2140.009 2072ND-D8787990.013 1652ND-D8787990.013 165SRD-D8987910.013 165PLS-DAMSC746543 216NSC746523 216164SNV716523 216PLS-DASNV716523 216SND-D686833 213164SND-D686833 213164SND-D68843 2160154SND-D8984743 2160MSC82873 2130154SND-D8984743 2160SND-D8984743 2160SND-D8984743 2160SND-D8984743 2160SND-D8984743 2160SND-D8984743 2160SND-D8	過点 ニック	DE	SNV	82	71		3 216	0
2ND-D88773 21403RD-D88773 21303RD-D88773 2130.011 960NSC87841010.007 782SNV84773190.008 4261ST-D87873 2130.014 3822ND-D878790.013 1653RD-D898790.013 165NSC746523 216NSC746523 216NSC746523 216SNV716823 216SNV736823 216SNV746523 216SNV736823 216SNV736823 216SNV736823 216SNV736833 213Original data85843 216SNV81743 216Original data85843 216SNV84743 216SNV84743 216Original data863 214SNV813 214ORD-D973 213SNV84743 216SND-D89813 214SND-D973 213SND-D89813 214SND-D89843 216SND-D89 <td>测走</td> <td>KF</td> <td>1ST-D</td> <td>88</td> <td>77</td> <td></td> <td>3 215</td> <td>0</td>	测走	KF	1ST-D	88	77		3 215	0
SRD-DS8773 2130Original data86843 2150.011 960MSC87841010.007 782SNV84773190.008 42611ST-D87843 2140.009 2072ND-D8787990.013 1653RD-D8987990.013 165NENE676823 216NENE716823 216NENE736823 216SNV716523 216PLS-DA736823 216SNV716523 2162ND-D696823 2162ND-D696833 213PESEXMSC82873 2163RD-D696833 213NETMSC82843 2163RD-D89843 2163RD-D89843 2163RD-D89813 2163RD-D89813 2163RD-D89813 2133RD-D89813 2143RD-D89813 2163RD-D89813 2163RD-D89813 2163RD-D81321303RD-D813 2143RD-D833 2133RD-D813 2143RD-D81			2ND-D	88	77		3 214	0
PreserveOriginal data86843 2150.011 960MSC87841010.007 782SNV84773190.008 4261ST-D87843 2140.009 2072ND-D87873 2130.014 3823RD-D8987900.013 165SNV7166523 216NSC746543 216NSC7166523 2161ST-D7366823 2162ND-D696823 2161ST-D7366823 2162ND-D696833 213Original data85843 21603RD-D89813 21601ST-D89813 21602ND-D89813 21602ND-D89813 21603RD-D89813 21603RD-D89813 21602ND-D89813 21603RD-D89813 2130RF-VSNSC85843 216ARF-VSNSC85843 2162ND-D89813 21403RD-D89813 21403RD-D89843 2160.011 5533RD-D89841560.010 3433RD-D			3RD-D	88	77		3 213	0
MSC87841010.007 782SNV84773190.008 426IST-D87843 2140.009 2073RD-D87990.013 1653RD-D87990.013 165SNV716823 216MSC746543 216NTS-D736823 216SNV716523 216IST-D736823 216SND-D696823 216Original data87833 216ORD-D696833 216ORD-D696833 216ORD-D696833 216ORD-D69843 2160MSC82873 2160IST-D89843 2160SNV813 21600SNV82873 2160SRD-D99813 2160SRD-D90773 2160SRD-D89813 2140RF-VSSNV83773 213MSC85841560.011 553MSC83772180.008 102RF-VSSNV83772180.008 102SND-D89841670.017 985SND-D89841670.017 985 <td< td=""><td></td><td></td><td>Original data</td><td>86</td><td>84</td><td></td><td>3 215</td><td>0.011 960</td></td<>			Original data	86	84		3 215	0.011 960
RF-VSSNV84773190.008 4261ST-D87843 2140.009 2072ND-D87873 2130.014 3823RD-D8987990.013 165NGC716823 216PLS-DASNV716623 2161ST-D736823 2162ND-D696823 2161ST-D736823 2162ND-D696823 2161ST-D736823 2162ND-D696833 213MSC82843 2160SNV84743 21601ST-D89843 21602ND-D89813 21603RD-D89813 21602ND-D89813 21603RD-D89813 21603RD-D89813 21603RD-D89813 2140SNV83772180.011 553MSC85841560.010 343SNV83772180.009 1682ND-D89843030.009 1682ND-D89843030.009 1682ND-D89841670.017 9882ND-D89841670.017 9883ND-D8			MSC	87	84		101	0.007 782
RF-VS 1ST-D 87 84 3 214 0.009 207 2ND-D 87 87 3 213 0.014 382 3RD-D 89 87 99 0.013 165 3RD-D 89 87 99 0.013 165 NGC 74 65 4 3 216 PLS-DA MSC 74 65 2 3 216 PLS-DA 73 68 2 3 216		DE Ve	SNV	84	77		319	0.008 426
2ND-D87873 2130.014 3823RD-D8987990.013 1653RD-D6823 216100MSC746543 216SNV716523 2161ST-D736823 2162ND-D696823 2163RD-D696823 2163RD-D686833 213MSC82873 2160MSC82873 2160MSC82873 2160SNV84743 2160MSC89813 2130NB-D90773 2130ND-D89813 2130NSC85843 2150ND-D89813 2130NSC89813 2130NSC89813 2130NSC89813 2130NSC85843 2130NSC85841560.011 553MSC85841560.003 43SNV83772180.008 1021ST-D89843030.009 1682ND-D89843030.009 1682ND-D89843030.009 1683RD-D89843030.007 9883RD-D<		KF-VS	1ST-D	87	84		3 214	0.009 207
3Rb-b8987990.013 165Original data716823 216MSC746543 216SNV716523 2161ST-D736823 2132ND-D696823 2133RD-D686833 213MSC82843 2160MSC82873 2160MSC82873 2160MSC82873 2160MSC82873 21601ST-D89843 21502ND-D89843 21503RD-D90773 21303RD-D90773 2130RF-VSMSC8584156SNV83772180.001 343RF-VSSNV83772180.009 1681ST-D89843030.009 1682ND-D89843030.009 1682ND-D89843030.009 1682ND-D89843030.009 1683RD-D89841670.017 9883RD-D89841670.017 9883RD-D89843030.009 1683RD-D89843030.009 1683RD-D89843030.007 1983RD-D89<			2ND-D	87	87		3 213	0.014 382
PLS-DA Original data 71 68 2 3 216 MSC 74 65 4 3 216 SNV 71 65 2 3 216 1ST-D 73 68 2 3 215 2ND-D 69 68 2 3 214 3RD-D 68 68 3 3 213 Original data 85 84 3 216 0 MSC 82 87 3 216 0 MSC 84 74 3 216 0 SNV 84 74 3 216 0 2ND-D 89 84 3 215 0 3RD-D 90 77 3 213 0 RF-VS MSC 85 84 156 0.011 553 <tr< td=""><td></td><td></td><td>3RD-D</td><td>89</td><td>87</td><td></td><td>99</td><td>0.013 165</td></tr<>			3RD-D	89	87		99	0.013 165
PLS-DAMSC746543 216SNV716523 2161ST-D736823 2152ND-D696823 2143RD-D686833 213MSC82873 2160MSC82873 2160MSC82873 2160MSC82873 2160MSC89843 21502ND-D89843 21502ND-D89813 21403RD-D90773 21302ND-D89813 21403RD-D90773 2130RF-VSNSC85841560.010 343SNV83772180.008 1021ST-D89843030.009 1682ND-D89841670.017 9883RD-D89841670.017 988			Original data	71	68	2	3 216	
PLS-DA SNV 71 65 2 3 216 1ST-D 73 68 2 3 215 2ND-D 69 68 2 3 214 3RD-D 68 68 3 3 213 Øriginal data 85 84 3 216 0 MSC 82 87 3 216 0 MSC 82 87 3 216 0 1ST-D 89 84 3 216 0 2ND-D 89 84 3 216 0 2ND-D 89 81 3 216 0 2ND-D 89 81 3 216 0 3RD-D 90 77 3 213 0 MSC 85 84 3 216 0 RF-VS MSC 85 84 3 213 0 1ST-D 89 84 3 60 0.011 553 1ST-D 89 84 303 0.008 102		PLS-DA	MSC	74	65	4	3 216	
PLS-DA 1ST-D 73 68 2 3 215 2ND-D 69 68 2 3 214 3RD-D 68 68 3 3 213 Øriginal data 85 84 3 216 0 MSC 82 87 3 216 0 MSC 84 74 3 216 0 1ST-D 89 84 3 215 0 2ND-D 89 84 3 215 0 2ND-D 89 81 3 213 0 2ND-D 89 81 3 216 0 3RD-D 90 77 3 213 0 RF-VS MSC 85 84 3 213 0 RF-VS MSC 85 84 3 213 0 RF-VS MSC 85 84 156 0.011 553 IST-D 89 84 303 0.008 102 IST-D 89 84			SNV	71	65	2	3 216	
			1ST-D	73	68	2	3 215	
 測定三次 RF 3 RF-VS 3 RD-D 68683 843 2160 0 0 0測定三次RF-VS 3 2ND-D89843 2150 0ND-D89813 2130 0ND-D89813 2130 0ND-D89813 2130 0ND-D89813 2130 0ND-D89813 2130 0ND-D89813 2130 0ND-D89841560.010 303ND-D89843030.009 168ND-D89843030.009 168ND-D89843070.017 988ND-D89870.008 3070.008 307			2ND-D	69	68	2	3 214	
 測定三次Original data85843 2160MSC82873 2160SNV84743 21601ST-D89843 21502ND-D89813 21403RD-D90773 2130RF-VSOriginal data8687780.011 553MSC83772180.008 1021ST-D89843030.009 1682ND-D89841670.017 9883RD-D89841670.017 9883RD-D89873070.008 811			3RD-D	68	68	3	3 213	
測定三次RFMSC82873 2160SNV84743 21601ST-D89843 21502ND-D89813 21403RD-D90773 2130RF-VSOriginal data8687780.011 553MSC85841560.010 3432ND-D89843030.009 1682ND-D89843030.009 1682ND-D89841670.017 9883RD-D89873070.008 811			Original data	85	84		3 216	0
測定三次RFSNV84743 21601ST-D89843 21502ND-D89813 21403RD-D90773 2130RF-VSOriginal data8687780.011 553MSC85841560.010 343SNV83772180.008 1021ST-D89843030.009 1682ND-D89841670.017 988 3RD-D 89873070.008 811			MSC	82	87		3 216	0
神定三次 KF 1ST-D 89 84 3 215 0 2ND-D 89 81 3 214 0 3RD-D 90 77 3 213 0 Original data 86 87 78 0.011 553 MSC 85 84 156 0.010 343 SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811	测试子子对	DE	SNV	84	74		3 216	0
2ND-D 89 81 3 214 0 3RD-D 90 77 3 213 0 Original data 86 87 78 0.011 553 MSC 85 84 156 0.010 343 RF-VS SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811	测定二伏	KF	1ST-D	89	84		3 215	0
3RD-D 90 77 3 213 0 Original data 86 87 78 0.011 553 MSC 85 84 156 0.010 343 RF-VS SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811			2ND-D	89	81		3 214	0
Original data 86 87 78 0.011 553 MSC 85 84 156 0.010 343 SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811			3RD-D	90	77		3 213	0
MSC 85 84 156 0.010 343 RF-VS SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811			Original data	86	87		78	0.011 553
RF-VS SNV 83 77 218 0.008 102 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811			MSC	85	84		156	0.010 343
KF-VS 1ST-D 89 84 303 0.009 168 2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811		DE Ve	SNV	83	77		218	0.008 102
2ND-D 89 84 167 0.017 988 3RD-D 89 87 307 0.008 811		KF-VS	1ST-D	89	84		303	0.009 168
3RD-D 89 87 307 0.008 811			2ND-D	89	84		167	0.017 988
			3RD-D	89	87		307	0.008 811

附表 3 MSC 处理后的建立 PLS-DA 模型的 测试集混淆矩阵表

Attached Table 3 Test set confusion matrix of PLS-DA model after MSC processing

				L.		0	
Memerbers ACC/%			а	b	с	d	е
4	0	а	0	0	4	0	0
2	100	b	0	2	0	0	0
12	83	с	0	0	10	2	0
9	100	d	0	0	0	9	0
4	0	е	0	2	2	0	0
0		No class	0	0	0	0	0
1	67.74	Total	0	4	16	11	0

同样地,由附表2可知,原始数据经五种光谱预处理方 法处理后,仅 MSC 所得 RF 模型性能(训练集和预测集准确 率分别为 86%和 87%)较原始数据所得模型性能(训练集和 预测集准确率分别为 83%和 87%)有所改善。表明经 MSC 处理后所建立的 RF 模型最优,该模型对产地 a 和 b 样本总 的预测准确率为 50%,优于 PLS-DA 模型的 33%预测准确 率[见附图 3(a)和附表 3]。同时 RF 模型的整体性能要明显 优于 PLS-DA 模型。但 RF 模型变量过多,运行耗时,需进 一步选择重要建模变量以改善模型性能。

经重要变量筛选所建立的 RF-VS 模型中,应用 1ST-D 对原始数据进行处理后所得模型性能最佳,其训练集和测试 集准确率分别为 88%和 87%,重要变量数为 167个,最佳界 值为0.009 207(见附表 2)。尽管最优 RF-VS 模型性能较 RF 模型性能(训练集和预测集准确率分别为 86%和 87%)仅得 到轻微改善,但最优 RF-VS 模型产地 a 和 b 样本总的预测准 确率 83%,明显优于模型 RF 的 50%和 PLS-DA 的 33%,同 时对质量差的来自产地 a 和 b 的丹参样本与产地 c、d 和 e 的 样本之间区分准确率高达 97%,高于最优 RF 模型的 90%和 最优 PLS-DA 模型的 81%区分准确率(见附图 3 和附表 3)。 另一方面,建模变量降至 167个,增加了模型的运行速率。 但总的准确率不变,经重要变量筛选后的 RF-VS 模型对其 他三产地质量较好样本的测试集准确率低于 RF 模型。

综上,选择经 1ST-D 处理后,由 RF-VS 建立的模型为 最终模型。





Attached Fig. 3 The confusion matrix of the test set of the optimal RF model (a) and the optimal RF-VS model (b)

2.5.2 不同测定次数下分类模型的比较

扫描 1 次、2 次和 3 次所得原始数据和五种数据前处理 方法所得数据,经 PLS-DA,RF,RF-VS 运算所得最优模型 见表 1。由表 1 可知,扫描 2 次和 3 次的最优模型均经 3 RD-D处理后由 RF-VS 计算所得,其训练集的准确率和测试集 的准确率均分别为 89%和 87%,但与扫描 1 次的最优模型 (训练集和预测集准确率分别为 88%和 87%)相比,性能改 善轻微,再根据每份样品随机扫描 1 次、2 次、3 次所需时间 分别大致为 150,300 和 450 min,最终选择每份样品每根药 材随机扫描 1 次所得数据,经 1 ST-D 预处理,RF-VS 计算所 得模型为最终模型。

表 1 不同扫描次数条件下所建最优模型结果 Table 1 The results of best models under the conditions of different scanning times

or unrerent scanning times									
扫描次数	模型	数据前处	训练集	测试集	变量数	田店			
		理方法	ACC	ACC	/个	介诅			
扫描一次	RF-VS	1ST-D	88	87	167	0.009 207			
扫描二次	RF-VS	3RD-D	89	87	99	0.013 165			
扫描三次	RF-VS	3RD-D	89	87	307	0.008 811			

3 结 论

显微聚焦拉曼光谱技术具快速、无损、样本需求少等优 点,本文尝试利用该技术对不同产地丹参样品每根药材的表 面在无损伤条件下进行随机取点,所得原始光谱既包含了样 本本身表面丹参酮类成分的信息,也包含了不同产地杂质的 信息,这有利于该技术用于丹参产地鉴别。比较了不同扫描 次数下,不同分类算法建立的丹参产地鉴别模型,结果样品 的每根药材表面随机扫描一次所得光谱数据经1ST-D处理, 由 RF-VS计算的丹参产地识别模型性能优良,其训练集和 测试集预测准确率分别为88%和87%;扫描次数增加为2 次和3次,所得最优模型训练集和测试集预测准确率分别为 89%和87%,模型性能轻微改善,但测试时间成倍增加,因 此选择每份样品每根药材随机扫描一次所得光谱数据建立的 最优 RF-VS 模型为最终模型。本研究为显微聚焦拉曼光谱 技术应用于中药材尤其是贵细中药材的产地溯源和真伪鉴别 提供了重要依据。

References

- [1] Fang J, Little P J, Xu S W. Medicinal Research Reviews, 2017, 38(1): 201.
- [2] DENG Ai-ping, GUO Lan-ping, ZHAN Zhi-lai, et al(邓爱平, 郭兰萍, 詹志来, 等). China Journal of Chinese Materia Medica(中国中药 杂志), 2016, 41(22): 4274.
- [3] Liang W Y, Chen W J, Wu L F, et al. Molecules, 2017, 22: 478.
- [4] Ni J L, Zhang F F, Han M Y, et al. Journal of Pharmaceutical and Biomedical Analysis, 2019, 170: 295.
- [5] Wang H, Hao N, Chen L, et al. Springerplus, 2016, 5(1): 1919.
- [6] MING Jing, CHEN Long, CHEN Ke-li, et al(明 晶,陈 龙,陈科力,等). Journal of Chinese Medicinal Materials(中药材), 2017, 40 (1): 32.

- [7] CHEN Long, MING Jing, YUAN Ming-yang, et al(陈 龙,明 晶,袁明洋,等). Chinese Journal of Experimental Traditional Medical Formulae(中国实验方剂学杂志), 2016, 22(21): 77.
- [8] Rinnan A, Berg F V, Engelsen S B. Trends in Analytical Chemistry, 2009, 28(10): 1201.
- [9] WANG Ya-xuan, TAN Feng, XIN Yuan-ming, et al(王亚轩,谭峰,辛元明,等). Spectroscopy and Spectral Analysis(光谱学与光谱 分析), 2021, 41(2): 565.
- [10] Schulz H, Baranska M. Vibrational Spectroscopy, 2007, 43(1): 13.
- [11] ZHU Zi-ying, GU Ren-ao, LU Tian-hong(朱自莹,顾仁敖,陆天虹). Application of Raman Spectroscopy in Chemistry(拉曼光谱在化学 中的应用). Changchun: Northeast University Press(长春:东北大学出版社), 1998. 295.

Research on Identification of Danshen Origin Based on Micro-Focused Raman Spectroscopy Technology

LI Qing^{1, 2}, XU Li^{1, 2}, PENG Shan-gui^{1, 2}, LUO Xiao^{1, 2}, ZHANG Rong-qin^{1, 2}, YAN Zhu-yun³, WEN Yong-sheng^{1, 2}*

1. Chengdu Institute for Drug Control, Chengdu 610045, China

- NMPA Key Laboratory for Quality Monitoring and Evaluation of Traditional (Chinese Medicine Chinese Materia Medica), Chengdu 610045, China
- 3. Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

Abstract The origin is an important factor affecting the quality of Chinese herbal medicine. The difference of origin leads to the uneven quality of Chinese herbal medicine. In order to maintain the market order, it is necessary to establish the method of identification of the origin of Chinese herbal medicine to identify and analyze the quality of Chinese herbal medicine more accurately. This article takes danshen, a major clinical medicinal material, from many origins as the research object, and 150 samples of danshen were collected from different origins. The surface of each root of danshen sample was scanned $1 \sim n$ times randomly by micro focusing Raman spectroscopy under non-destructive conditions, and the average spectrum of each sample was calculated. By analyzing the original spectral data, it is found that the surface spectral signal of danshen contains both the Raman spectra of tanshinones and the fluorescence spectra of impurities. Mainly reflected in that danshen from different origins has their aggregation regions, and the signal intensity of the surface spectral signal of danshen is significantly weaker or stronger than that of tanshinones in the specific wavelength range. After preprocessing the spectral data of $1 \sim n$ scans, the classification model of danshen origin was established by partial least squares discriminant analysis (partial least squares-discriminant analysis, PLS-DA) and random forest classification algorithm [no screening (random forest, RF) or screening of important variables (RF-VS)]. Results the training set and test set accuracy of the optimal model obtained by random 1 scanning were 88% and 87%respectively, and the samples with low quality and high quality could be distinguished with an accurate of 97%; the training set and test set accuracy of the optimal model obtained by random scanning 2 and 3 times were both 89% and 87% respectively. Combined with the operation efficiency of the model The spectrum obtained by random 1 scanning was selected, and the identification model of the origin of danshen was obtained by first derivative (1ST-D) pretreatment and RF-VS calculation. In conclusion, the micro focused Raman spectroscopy technology can establish a rapid and accurate prediction model of the origin of danshen under non-invasive conditions and provide a reference for the further application of this technology in identifying the origin and authenticity of expensive and scarce Chinese herbal medicine.

Keywords Raman spectroscopy; Danshen; Identification of the origin

(Received Mar. 31, 2021; accepted Sep. 19, 2021)

* Corresponding author