

• 方药研究 •

红外光谱数据融合对栽培滇重楼产地鉴别*

丁于刚, 张庆芝[△]

(云南中医药大学中药学院, 云南 昆明 650500)

摘要: **目的** 多样的环境因素使得不同产地栽培滇重楼的化学成分也丰富多样, 不同居群栽培滇重楼的甾体皂苷类成分具有很大的差异, 多源数据融合分析能更全面的表征药材化学信息, 建立一个高效而准确的产地鉴别模型, 为其资源合理开发利用提供依据。 **方法** 以来自云南和四川的 8 个产地(保山、楚雄、大理、红河、丽江、成都、文山、玉溪)共 366 份栽培滇重楼根茎为实验材料, 采集其傅里叶变换近红外光谱(FT-NIR)和衰减全反射-傅里叶变换中红外光谱(ATR-FTMIR)数据。采用 Kennard-Stone 算法将不同产地的样品分为 2/3 的训练集和 1/3 的预测集, 基于 4 种特征变量提取方法(CARS、VIP、SPA、SO-Covsel)结合 2 种数据融合策略(低级数据融合和中级数据融合), 建立偏最小二乘产地判别分析模型。根据模型参数交叉验证均方根误差(RMSECV)和预测均方根误差(RMSEP)评估模型的稳定性, 模型训练集和预测集准确率(ACC)评估模型分类性能。 **结果** 近红外光谱和中红外光谱均能反应不同产地栽培滇重楼的化学成分差异, 在中级数据融合中, 基于 VIP 和 SPA 提取的特征变量建立的模型正确率均大于 94%。相较于中级数据融合, 低级数据融合模型得到了最为满意的结果, 其预测集分类正确率达到 100%。 **结论** 根据近红外和中红外数据建立的低级数据融合 PLS-DA 模型, 能够用于栽培滇重楼的产地鉴别分析。

关键词: 栽培滇重楼; 产地鉴别; 偏最小二乘判别分析; 特征变量提取方法; 数据融合

中图分类号: R284

文献标志码: A

文章编号: 1000-2723(2022)01-0072-07

DOI: 10.19288/j.cnki.issn.1000-2723.2022.01.014

The Origin Identification Study of *Paris Polyphylla* var. *Yunnanensis* Based on the Infrared Spectrum Data Fusion Strategy

DING Yugang, ZHANG Qingzhi

(College of Traditional Chinese Medicine, Yunnan University of Chinese Medicine, Kunming 650500, China)

ABSTRACT: Objective Various environmental factors make the chemical composition of cultivated *Paris polyphylla* var. *yunnanensis* also rich and diverse, and the steroidal saponins of *Paris polyphylla* var. *yunnanensis* in different origins are quite different. An efficient and accurate identification model of different origins is established to provide a basis for the rational development and utilization of its resources. **Methods** A total of 366 rhizomes from 8 origins in Yunnan and Sichuan provinces (Baoshan, Chuxiong, Dali, Honghe, Lijiang, Chengdu, Wenshan, Yuxi) were used as experimental materials, and their Fourier transform near-infrared spectra (FT-NIR) and attenuated total reflection-Fourier transform mid-infrared spectroscopy (ATR-FTMIR) data were collected. The Kennard-Stone algorithm was used to divide samples from different origins into 2/3 training set and 1/3 prediction set. Based on 4 feature variable extraction methods (CARS, VIP, SPA, SO-Covsel) combined with 2 data fusion strategies (Low-level data fusion and intermediate-level data fusion), the

收稿日期: 2021-10-03

* 基金项目: 云南省重大科技专项(202002AA100007); 云南省中青年学术和技术带头人后备人才项目(202005AC160032)

第一作者简介: 丁于刚(1995-), 男, 在读硕士研究生, 研究方向: 中药资源开发与利用。

△通信作者: 张庆芝, E-mail: ynkzqz@126.com

partial least squares origin discriminant analysis model was erected. According to the model parameters, cross-validation root mean square error (RMSECV) and prediction root mean square error (RMSEP) were used to evaluate the performance of the model, and the accuracy of training set and prediction set were used to evaluate the model classification performance.

Results Both near-infrared and mid-infrared spectra could reflect the differences of cultivated *Paris polyphylla* var. *yunnanensis* in different origins. In the mid-level data fusion models, the accuracy of the models established based on the feature variables extracted by VIP and SPA were both greater than 94%. Compared with the mid-level data fusion models, the low-level data fusion model obtained the most satisfactory results, and its accuracy reached 100%. **Conclusion** The low-level data fusion PLS-DA model established based on the near-infrared and mid-infrared data can be used for the identification and analysis of the cultivated *Paris polyphylla* var. *yunnanensis*.

KEY WORDS: cultivated *Paris polyphylla* var. *yunnanensis*; geographical authentication; PLS-DA analysis; feature variable extraction method; multi-blocks data fusion strategy

作为百合科多年生草本植物,滇重楼(*Paris polyphylla* var. *yunnanensis*)的根茎被广泛应用于中国、印度、不丹、尼泊尔等国家^[1]。重楼在中国有超过两千年的药用历史,最早被名为“蚤休”记录于《神农本草经》中^[2]并记其“味苦微寒,主惊痫摇头弄舌,热气在腹中,癩疾,痈疮阴蚀,下三虫,去蛇毒”,现代药理研究表明重楼具有抗肿瘤、凝血、抗氧化、抗炎镇痛、抗微生物活性等^[3]药理活性。中国是重楼最大的生产国和消费国,在过去的 40 年间,重楼药材的市场价从 2.7 元/kg 高涨到了 600~750 元/kg^[3]。随着市场需求的增大,野生重楼药材已经不能满足市场的需要。多样的环境因素使得不同产地的栽培滇重楼化学成分也各有不同,且影响滇重楼的分布的主导因子与其地理环境息息相关^[4]。因此,药材的产地溯源对其质量控制及合理开发利用具有重要意义。滇重楼化学成分的差异主要体现在次生代谢产物甾体皂苷类化合物上,有研究表明不同产地、不同年份的滇重楼其化学成分具有较大的差异^[5-6]。裴艺菲等采用多光谱技术对云南不同产地滇重楼进行了分析,结果表明单一光谱建立的模型正确率较高^[7]。目前为止,不同类型的检测分析技术被成功应用于鉴别滇重楼的来源,例如高效液相色谱法(HPLC, High performance liquid chromatography)^[8-9]、超高效液相质谱联用(UHPLC-MS/MS, Ultra-performance liquid chromatography-tandem mass spectrometry)等^[6]。然而,这些技术需要对样品进行复杂的预处理、有害试剂的使用、对实验员的高要求等,基于湿法的液相色谱法的方法不适于大量样本的快简分析。振动光谱结合化学计量学已然

成为一种绿色、快捷的方法,这使得利用红外光谱结合化学计量学对不同产地的栽培滇重楼的快速鉴别成为了可能。中级数据融合能减少大量的冗余和无关变量,在模式识别分析中更具有可靠性。通过采集云南和四川共 8 个产地的滇重楼根茎的近红外和中红外光谱数据,光谱数据表征不同产地植物样品的化学信息差异。结合多元变量提取方法,建立一个稳定可靠的产地模式识别方法,以促进云南不同产地栽培滇重楼的合理开发利用。

1 材料与方法

1.1 实验材料 本研究实验样品为栽培滇重楼的干燥根茎样品,原植物采自我国西南的云南和四川省,包括保山、楚雄、大理、红河、丽江、文山、玉溪和成都。所有采集原植物均经云南中医药大学张庆芝教授鉴定为滇重楼(*P. polyphylla* var. *yunnanensis*)。所有样品信息如表 1 所示,样品采集后经清洗干净后切片,在 50℃ 下烘干。后经打粉机粉碎后过 100 目筛,所有样品均置于干燥皿中待下一步分析。

1.2 红外光谱采集

1.2.1 衰减全反射傅里叶变换中红外光谱 衰减全反射傅里叶变换中红外光谱采集自配备氘化三甘氨酸硫酸盐(DTGS)检测器和 ZnSe 衰减全反射元件的中红外光谱仪。光谱采集范围为 4 000~650 cm^{-1} ,分辨率为 4 cm^{-1} ,每个样品累积扫描 16 次。经过 3 次重复扫描后,得到了所有样品的光谱数据。光谱采集室恒定的温度和湿度由温湿度控制仪进行调控,所有采集得到的光谱数据由 SIMCA-P+ 14.1 进行分析。

表 1 重楼样品信息表

地区	分布区域	总样品数	训练集	预测集	经度	纬度	海拔
保山	云南西部	50	35	15	98°30'52"	24°34'26"	1 916 m
楚雄	云南中部	55	36	19	101°22'57"	25°31'41"	2 168 m
大理	云南西部	47	32	15	100°28'4"	25°13'27"	2 041 m
红河	云南东南部	45	30	15	102°59'37"	23°02'20"	2 230 m
丽江	云南西北部	41	28	13	100°11'48"	26°53'04"	2 400 m
成都	四川盆地西部	44	30	14	103°17'52"	30°43'52"	2 170 m
文山	云南东南部	41	28	13	104°3'48"	23°14'32"	1 793 m
玉溪	云南中部	43	30	13	102°27'04"	24°26'19"	2 129 m

1.2.2 傅里叶变换近红外光谱 重楼样品的近红外光谱由 Antaris II 型光谱仪检测, 仪器结合漫反射附件。光谱采集波数范围为 10 000~4000 cm^{-1} , 分辨率为 4 cm^{-1} 。样品采集室的温度和湿度维持恒定的值, 温度(25℃, 30%RH)。每次采集样品光谱之前, 空气背景(CO_2 和 H_2O)会被校正以减小误差。每个样品平行检测 3 次, 3 次后的平均光谱将用于进行下一步分析。

1.2.3 红外光谱数据预处理 从光谱检测仪获得的原始数据往往会包含一些干扰信息, 它们来自环境因素、检测器和其它因素等^[10-11]。在本研究中, 采用标准正态变量变换(SNV, standard normal variate transformation)^[12]以减少由样品物理状态带来的散射干扰, 同时可以将变量平衡在 0~1 之间。

1.3 模式识别技术 在最初, 偏最小二乘算法(PLS, partial least squares algorithm)被应用于回归问题的处理。在偏最小二乘回归中, 变量矩阵 Y(Y block)与训练集 X(X block)按照常规方法配对^[13]。一般来说, 偏最小二乘判别分析(PLS-DA)应用于特定情况如 Y 是类别数据。偏最小二乘判别模型的建立主要分为 4 步: 数据预处理、降维、模型验证和决策。本研究运用了 7 折交叉外部验证, R^2 (Coefficient of determination of model fitting)、 Q^2 (Prediction)、RMSEE (Root mean squared error of estimation)、RMSEP (Root mean squared error of prediction)、RMSECV (Root mean squared error of calibration validation) 等参数被用来评价模型的优劣。一般来说, Q^2 大于 0.5 表明模型具有较好的预测能力, 而模型的稳定性则与 R^2 有关。所

有数据处理和建模均在 SIMCA-P+ 14.1 软件完成。

1.4 特征变量提取 变量投影重要性(VIP, Variable importance for the projection)是衡量 PLS-DA 模型中单个变量对整个模型的影响的参数。VIP 值为数值的均方根, 一般来说, 它代表了相关性^[14]。选取 VIP 值大于 1 的变量建立模型能减少其他不重要变量带来的干扰, 从而提高分类效果。SPA 的优势是能够消除大量的冗余信息, 适合于光谱特征波长的筛选, 并已被证实结合分类算法(如 SPA-PLS-DA)具有较好的效果^[15-16]。竞争性自适应再加权算法(CARS, Competitive adaptive reweighted sampling), 有别于其他的特征变量筛选方法, 主要通过自适应再加权采样技术选择出偏最小二乘回归系数绝对值较大的波长点, 根据交叉验证选出 RMSECV 值最小的子集^[17]。序列正交协方差特征变量选择(SO-CovSel, sequential and orthogonalized covariance selection), 通过对每个预测变量与所有变量之间的协方差进行评估, 筛选出协方差最高的变量作为特征变量^[18]。基于以上原理, 所有的预测因子和相应的响应值将会被统计, 并将重复以上过程, 直至选择出适当的变量数。

光谱数据具有复杂、多维的特点, 本研究采用多种不同的特征变量提取方法旨在从不同的角度简化和提高模型的可解释性。

1.5 多源数据融合策略 当处理来自多传感器的数据时, 多源数据融合策略(MSDF, Multi-sensor data fusion)是比较适用的方法。数据融合策略通过结合不同模块的数据, 进而分析可以得到相比单一来源数据

更准确和有效的决策^[19]。总的来说,根据融合策略的不同,数据融合方法可分为数据级融合(Low level)、特征级融合(Mid-level)和决策级融合(Decision-level)。在本文中采用了 2 种数据融合策略进行对比,以提高实验结果的准确性。

2 结果与分析

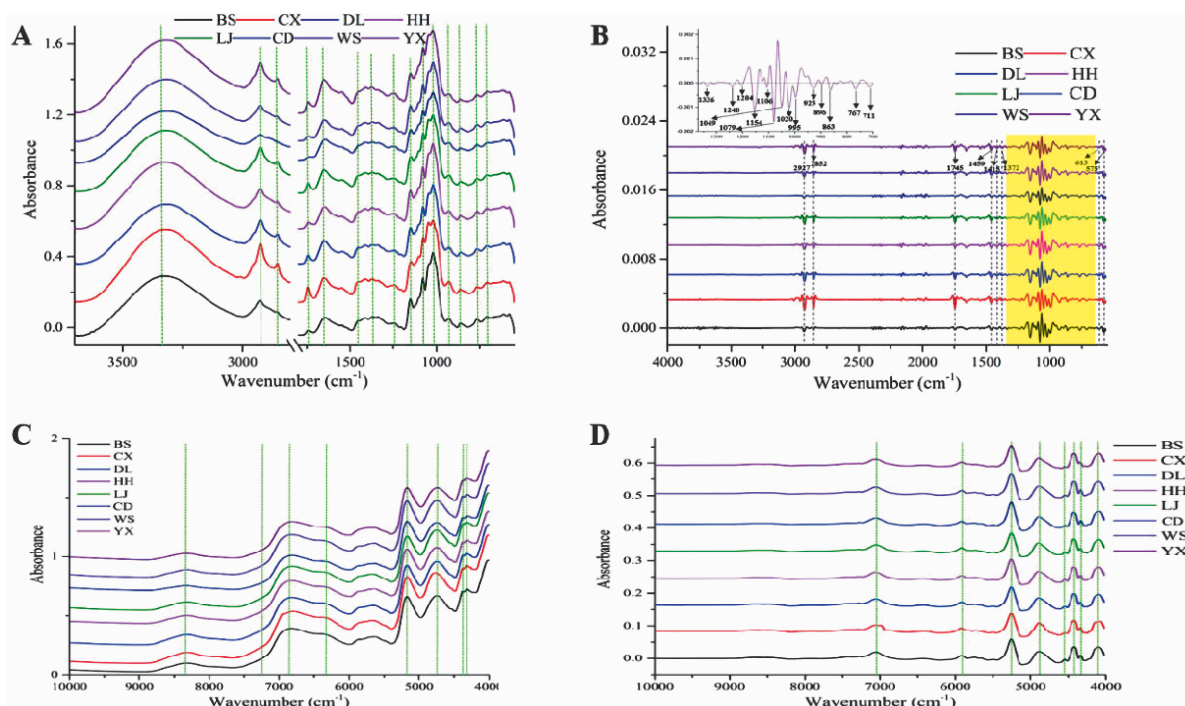
2.1 光谱的特征解释 根茎样品的傅里叶变换中红外光谱图能够反映化学信息,体现不同产地滇重楼的差异。在本研究中,来自 8 个产地的滇重楼中红外和近红外光谱数据均做了平均处理。如图 1 所示,经过二阶导数的光谱相比于原始光谱更能反映出样品的化学信息^[20]。在二阶导数的光谱中,共有 20 个显著的吸收峰(2 927, 2 852, 1 745, 1 459, 1 415, 1 372, 1 336, 1 240, 1 204, 1 154, 1 106, 1 079, 1 049, 1 020, 995, 925, 896, 863, 767, 711 cm^{-1})^[21]。其中, 3 334 cm^{-1} 的峰为羟基的伸缩振动吸收, 2 927、2 852、1 459、1 415 和 1 312 cm^{-1} 的吸收主要为来自亚甲基碳氢键的弯曲和伸缩振动。在 1 745 cm^{-1} 的较大吸收峰为碳氧双键的伸缩振动, 推测与甾体皂苷、黄酮、挥发油及多糖类物质相关^[22]。1 300~400 cm^{-1}

范围的峰较为密集且复杂,为样品的指纹图谱区。

相较于中红外光谱,来自 8 个地区的平均近红外光谱吸收峰较少。5 500 cm^{-1} ~4 200 cm^{-1} 光谱区域为其主要吸收波段,该波段的吸收可能与碳氢键的变形和氧氢、氮氢、碳氢组合模式的第二泛音有关^[23]。位于 7 200~5 500 cm^{-1} 的低频区的 2 个吸收峰归属于碳氢、氧氢和氮氢键伸缩振动的泛音,而碳氢、氧氢和氮氢键伸缩振动的第二泛音则位于 9 000~7 500 cm^{-1} 的弱吸收峰处^[24]。整体上看,不同产地的中红外和近红外光谱吸收峰位置相同,但是强度略有差异,仅凭平均光谱图的差异难以实现不同产地滇重楼的鉴别。

2.2 多源数据融合策略结合 PLS-DA 判别滇重楼产地

2.2.1 基于低级数据融合的 PLS-DA 在本次实验中,通过 SIMCA-P+ 14.1 对采集自样品的中红外和近红外光谱数据进行转换,共得到 3 346 个变量(FT-NIR: 1 557 个变量,ATR-FTMIR: 1 789 个变量)。在进行建模之前,先对所有来自 8 个大类的样品进行 KS 训练,将样品分为 2/3 的训练集和 1/3 的预测集。



注: A. ATR-FTMIR 的原始光谱; B. ATR-FTMIR 的 SD 预处理光谱; C. FT-NIR 的原始光谱; D. FT-NIR 的 SD 预处理光谱

图 1 保山、楚雄、大理、红河、丽江、成都、玉溪、文山的 ATR-FTMIR 和 FT-NIR

在对原始数据进行 SNV 预处理之后,中红外和近红外的数据被串联建立 PLS-DA 模型。PLS-DA 模型的最佳潜在变量数则是根据模型的交叉验证均方根误差和 Q^2 决定的,如图 2 所示所有模型的最佳潜在变量数均已确定。模型参数如表 2 所示,训练集和预测集的正确率均达到了 100%。

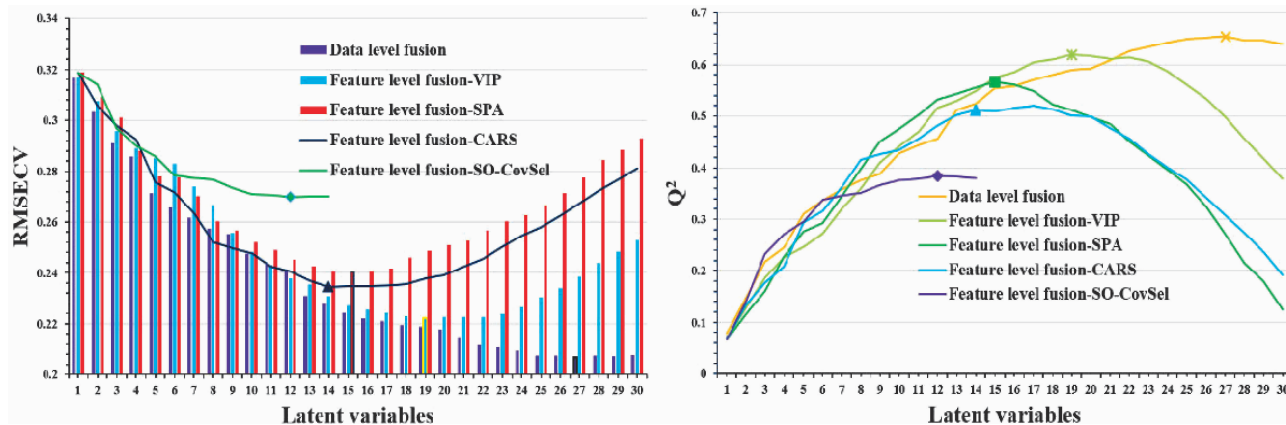


图 2 PLS-DA 模型的最佳潜在变量数

表 2 基于数据融合的多重 PLS-DA 模型的参数值

	数据集	变量数	LVs	R^2	Q^2	RMSECV	RMSEP	ACCtrain/%	ACCtest/%
数据级融合	NIR+MIR	3 346	27	0.861	0.655	0.207	0.154	100	100
特征级融合	VIP	1 005	19	0.800	0.619	0.222	0.175	99.60	98.29
	SPA	132	15	0.745	0.566	0.240	0.190	98.39	94.02
	CARS	108	14	0.649	0.512	0.234	0.210	96.79	90.60
	SO-CovSel	14	12	0.484	0.383	0.270	0.234	85.14	84.62

2.2.2 中级数据融合结合 PLS-DA 区别于低级数据融合较为庞大的计算量,特征级融合摒除了大量的冗余变量从而使模型较为简单^[25]。基于 VIP 的特征变量选择方法分别从中红外和近红外数据中筛选出 VIP 值大于 1 的变量,这 505 和 500 个变量将被视为重要变量进行下一步分析^[26-27]。根据 VIP-PLS-DA 模型可知,最佳潜在变量数为 19,训练集正确率达到 100%,训练集正确率达到 99.60%,只有预测集中有 2 个样品被错误分类。

基于 SPA 选择的特征变量来自光谱的各个波段。来自不同传感器的特征变量组成一个新的 366×132 的数据集,根据模型可知,其最佳 LVs 为 15,模型预测集正确率为 94.02%。根据竞争性自适应再加权算法的原理可知,在交叉验证均方根误差最低时筛选的变量为最佳变量数。随着模型运行次数的增加,交叉验证均方根误差呈高-低-高的变化趋势,无关

冗余变量被剔除,均方根误差减小,在最佳变量数时达到最低值。但随着采样的继续进行, RMSECV 增加是因为消除了必要的变量。最终共有 32 和 76 个变量参与模型的建立。采用 SO-CovSel 特征变量筛选方法筛选的变量数较少,分别为 (10×366) 和 (4×366) 变量并进行下一步建模分析。结果表明,基于低级数据融合的效果最佳,训练集和预测集正率均达到了 100%。而基于 SO-CovSel 方法的模型效果最差,预测集正确率仅有 84.62%。

3 讨论

从 8 个产地的滇重楼平均中红外光谱图可以得出,来自四川成都的样品在 $3\ 000 \sim 2\ 800\ \text{cm}^{-1}$ 与其它产地具有较为明显的差异,这可能是四川盆地相对独特的气候条件所致。Yang 等通过超高效液相色谱-紫外串联质谱法 (UHPLC-UV-MS, Ultra-performance liquid chromatography-ultraviolet spectroscopy-tan-

dem mass spectrometry) 检测了不同产地滇重楼的皂苷类成分含量,结果显示云南省南部地区总皂苷的平均值显著高于中部地区^[6]。表明不同产地的滇重楼,其有效成分的差异较为明显。

根据模型正确率和参数表明(表2),基于低级数据融合建立的偏最小二乘判别分析模型效果最优,低级数据融合通过串联不同传感器的信息能最大限度的保留样品的化学信息。在中级数据融合中,最终变量数较多的模型其参数较优,正确率较高。这可能是因为去除大量冗余信息的同时,许多对建立模型有效的变量也被去除。令人意外的是,基于SPA变量选择方法的模型效果达到了94.02%,因其只在较少的变量上进行分析,表明该方法具有较好的泛化能力。

4 结论

本研究通过红外光谱技术结合数据融合方法,采用PLS-DA模型对来自四川和云南8个产区的栽培滇重楼进行了鉴别分析。在特征级数据融合中采用了多种特征变量提取方法,结果表明基于VIP特征变量提取方法模型结果较优,其正确率达到99.60%,预测集中只有2个样品被错误分类。综合两种数据融合类型,初级数据融合的模型效果最优,分类正确率达到了100%。综合模型可知不同传感器来源的数据具有协同性,该方法能够成功应用于不同产地滇重楼的鉴别。

参考文献:

- [1] DING Y G, ZHAO Y L, ZHANG J, et al. The traditional uses, phytochemistry, and pharmacological properties of *Paris L.* (Liliaceae): a review[J]. *J Ethnopharmacol*, 2021, 278: 114293.
- [2] 王绍龄. 重楼的本草考证[J]. *江西中医学院学报*, 2004, 16(1): 55.
- [3] CUNNINGHAM A B, BRINCKMANN J A, BI Y F, et al. *Paris* in the spring: a review of the trade, conservation and opportunities in the shift from wild harvest to cultivation of *Paris polyphylla* (Trilliaceae) [J]. *J Ethnopharmacol*, 2018, 222: 208–216.
- [4] 石子为, 杨少华, 张丽霞, 等. 基于地形土壤因子的滇重楼生态适宜性区划研究[J]. *西南农业学报*, 2017, 30(8): 1904–1909.
- [5] 陈翠, 张浩, 谭敬菊, 等. 云南重楼生长发育特性及皂苷含量动态变化研究[J]. *西南农业学报*, 2017, 30(6): 1320–1324.
- [6] YANG Y G, JIN H, ZHANG J, et al. Quantitative evaluation and discrimination of wild *Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz from three regions of Yunnan Province using UHPLC–UV–MS and UV spectroscopy couple with partial least squares discriminant analysis[J]. *J Nat Med*, 2017, 71(1): 148–157.
- [7] 裴艺菲, 左智天, 赵艳丽, 等. FTIR, ATR–FTIR 和 UV 多光谱鉴别不同产地重楼[J]. *分析测试学报*, 2019, 38(1): 14–21.
- [8] 李海涛, 罗先文, 管燕红, 等. 云南省不同地区滇重楼皂苷含量的对比及影响因子分析 [J]. *中国中药杂志*, 2014, 39(5): 803–806.
- [9] 陈铁柱, 文飞燕, 张涛, 等. 21个产地七叶一枝花中皂苷类成分的评价[J]. *中成药*, 2017, 39(11): 2345–2350.
- [10] MISHRA P, BIANCOLILLO A, ROGER J M, et al. New data preprocessing trends based on ensemble of multiple preprocessing techniques [J]. *TrAC Trends in analytical chemistry*, 2020, 132: 116045.
- [11] ENGEL J, GERRETZEN J, SZYMAŃSKA E, et al. Breaking with trends in pre-processing? [J]. *TrAC–Trends in analytical chemistry*, 2013, 50: 96–106.
- [12] BARNES R J, DHANOA M S, LISTER S J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra [J]. *Applied spectroscopy*, 1989, 43(5): 772–777.
- [13] BARKER M, RAYENS W. Partial least squares for discrimination [J]. *Journal of chemometrics*, 2003, 17(3): 166–173.
- [14] GALINDO–PRIETO B, ERIKSSON L, TRYGG J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS): Variable influence on projection for OPLS [J]. *Journal of chemometrics*, 2014, 28(8): 623–632.
- [15] PONTES M J C, GALVÃO R K H, ARAÚJO M C U, et al. The successive projections algorithm for spectral variable selection in classification problems [J]. *Chemometrics and Intelligent Laboratory Systems*, 2005, 78(1):

- 11-18.
- [16] REN G X, WANG Y J, NING J M, et al. Highly identification of keemun black tea rank based on cognitive spectroscopy: near infrared spectroscopy combined with feature variable selection[J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2020, 230: 118079.
- [17] LI H D, LIANG Y Z, XU Q S, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. *Anal Chim Acta*, 2009, 648(1): 77-84.
- [18] BIANCOLILLO A, MARINI F, ROGER J M. SO-CovSel: a novel method for variable selection in a multiblock framework[J]. *Journal of Chemometrics*, 2020, 34(2): e3120.
- [19] XIAO F Y. Evidence combination based on prospect theory for multi-sensor data fusion[J]. *ISA Trans*, 2020, 106: 253-261.
- [20] CHEN J B, SUN S Q, ZHOU Q. Data-driven signal-resolving approaches of infrared spectra to explore the macroscopic and microscopic spatial distribution of organic and inorganic compounds in plant[J]. *Anal Bioanal Chem*, 2015, 407(19): 5695-5706.
- [21] SUN S Q, ZHOU Q, CHEN J B. Analysis of traditional Chinese medicine by infrared spectroscopy[M]. Beijing: Chemical Industry Press, 2010.
- [22] CHEN J B, GUO B L, YAN R, et al. Rapid and automatic chemical identification of the medicinal flower buds of *Lonicera* plants by the benchtop and hand-held Fourier transform infrared spectroscopy[J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2017, 182: 81-86.
- [23] FU H Y, HUANG D C, YANG T M, et al. Rapid recognition of Chinese herbal pieces of *Areca catechu* by different concocted processes using Fourier transform mid-infrared and near-infrared spectroscopy combined with partial least-squares discriminant analysis[J]. *Chinese Chemical Letters*, 2013, 24(7): 639-642.
- [24] ZHOU Y P, JIANG J H, WU H L, et al. Dry film method with ytterbium as the internal standard for near infrared spectroscopic plasma glucose assay coupled with boosting support vector regression[J]. *Journal of Chemometrics*, 2006, 20(1/2): 13-21.
- [25] 阿基业, 何骏, 孙润彬. 代谢组学数据处理: 主成分分析十个要点问题[J]. *药学报*, 2018, 53(6): 929-937.
- [26] BOLÓN-CANEDO V, ALONSO-BETANZOS A. Ensembles for feature selection: a review and future trends[J]. *Information Fusion*, 2019, 52: 1-12.
- [27] KUBINYI H. 3D QSAR in drug design: theory, methods and applications[M]. Leiden: ESCOM, 1993.