

基于近红外光谱技术的金银花产地鉴别与酚酸类成分预测研究

刘长青^{1,2,3}, 郝楠^{2,3}, 缪培琪^{2,3}, 王彦帅^{2,3}, 王晓梅⁵, 李正^{1,4*}

1. 天津中医药大学中药制药工程学院, 天津 301617

2. 天津现代创新中药科技有限公司, 天津 300392

3. 国家现代中药创新中心, 天津 300392

4. 现代中医药海河实验室, 天津 301617

5. 珀金埃尔默企业管理(上海)有限公司, 上海 201203

摘要: **目的** 建立一种基于近红外光谱(near infrared spectroscopy, NIRS)技术的金银花 *Lonicerae Japonicae Flos* 产地鉴别和酚酸类成分含量预测的分析方法。**方法** 采集不同产地不同批次的金银花的 NIRS 数据, 采用高效液相色谱法测定其中的酚酸类成分含量, 然后采用多种数据预处理方法提高光谱数据的信号特征, 利用特征变量筛选方法提取与检测目标相关的特征波段。利用 K 近邻(K-nearest neighbor, KNN)和线性判别分析(linear discriminant analysis, LDA)建立产地鉴别模型, 采用加权软投票法整合 KNN 和 LDA 模型的预测结果。利用支持向量机(support vector machine, SVM)和随机森林(random forest, RF)算法分别建立绿原酸、异绿原酸 A 和异绿原酸 C 成分的含量预测模型, 并从中为每种成分选取合适的建模方法。**结果** 采用加权软投票法构建的产地鉴别模型的预测性能最佳, 对预测集的预测准确率为 95.7%、对数损失值为 0.214、Kappa 系数为 0.907, 对巨鹿县、封丘县和平邑县 3 个产地的识别准确率分别为 98.8%、83.3%和 94.4%。构建的 SVM 模型对预测集金银花样品中的绿原酸成分的预测决定系数(coefficient of determination, R^2)为 0.810 2, 均方根误差(root mean square error, RMSE)为 0.313 6%, 相对预测偏差(relative percent deviation, RPD)为 2.305 2, 对异绿原酸 A 成分的预测 R^2 为 0.804 0, RMSE 为 0.135 4%, RPD 为 2.268 2, 构建的 RF 模型对异绿原酸 C 成分的预测 R^2 为 0.271 0, RMSE 为 0.052 0%, RPD 为 1.176 3。**结论** NIRS 技术结合化学计量学方法可以有效区分不同产地的金银花, 并可以较为准确的预测绿原酸和异绿原酸 A 成分的含量, 为金银花质量的快速、无损评价提供方法依据。

关键词: 近红外光谱; 金银花; 产地鉴别; 酚酸类; 异绿原酸 A; 绿原酸; 无损检测

中图分类号: R286.2 文献标志码: A 文章编号: 0253-2670(2025)16-5959-11

DOI: 10.7501/j.issn.0253-2670.2025.16.024

Origin identification and phenolic acid compositions prediction of *Lonicerae Japonicae Flos* based on near-infrared spectroscopy technique

LIU Changqing^{1,2,3}, HAO Nan^{2,3}, MIAO Peiqi^{2,3}, WANG Yanshuai^{2,3}, WANG Xiaomei⁵, LI Zheng^{1,4}

1. College of Pharmaceutical Engineering of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China

2. Tianjin Modern Innovative TCM Technology Co., Ltd., Tianjin 300392, China

3. National Innovation Center for Modern Chinese Medicine, Tianjin 300392, China

4. Haihe Laboratory of Modern Chinese Medicine, Tianjin 301617, China

5. PerkinElmer Management (Shanghai) Co., Ltd., Shanghai 201203, China

Abstract: Objective To establish an analytical method for the identification of *Lonicerae Japonicae Flos* (LJF) origin and the prediction of phenolic acid contents based on near infrared spectroscopy (NIRS) technology. **Methods** NIRS data of different batches of LJF from different origins were collected, and the content of phenolic acid components was determined by high performance liquid chromatography (HPLC), and then multiple data preprocessing methods were used to improve the signal

收稿日期: 2025-02-01

基金项目: 现代中医药海河实验室科研项目(22HHZYSS00003)

作者简介: 刘长青, 高级工程师, 从事中药智能装备研发和车间智能化等应用研究。E-mail: cq_liu@cntmic.com

*通信作者: 李正, 研究员, 从事中药智能制造关键技术及装备的研发与产业化应用基础研究。E-mail: lizheng@tjutcm.edu.cn

characteristics of the spectral data, and feature bands related to the detection targets were extracted by using the feature variable screening methods. K-nearest neighbour (KNN) and linear discriminant analysis (LDA) were used to establish the origin identification models, and the prediction results of KNN and LDA models were integrated using the weighted soft voting method. Support vector machine (SVM) and random forest (RF) algorithms were used to establish the content prediction models for chlorogenic acid, isochlorogenic acid A and isochlorogenic acid C components, respectively, and select appropriate modeling methods for each component. **Results** The origin identification model constructed using the weighted soft voting method had the best prediction performance, with a prediction accuracy of 95.7% a Log loss value of 0.214, a Kappa coefficient of 0.907 on the prediction set, and the identification accuracies for samples from Julu, Fengqiu, and Pingyi counties were 98.8%, 83.3%, and 94.4%, respectively. The constructed SVM model had a coefficient of determination (R^2) of 0.810 2, a root mean square error (RMSE) of 0.313 6%, and a relative percent deviation (RPD) of 2.305 2 for predicting chlorogenic acid content on the prediction set of LJF samples. For predicting isochlorogenic acid A content, the R^2 was 0.804 0, RMSE was 0.135 4%, and RPD was 2.268 2. Additionally, the constructed RF model for predicting isochlorogenic acid C content with an R^2 of 0.271 0, RMSE of 0.052 0%, and RPD of 1.176 3. **Conclusion** The NIRS technique combined with chemometrics can effectively distinguish LJF from different origins, and can accurately predict the content of chlorogenic acid and isochlorogenic acid A components, which provides a methodological basis for the rapid and non-destructive evaluation of the quality of LJF.

Key word: near-infrared spectroscopy; *Lonicerae Japonicae Flos*; origin identification; phenolic acids; chlorogenic acid; isochlorogenic acid A; nondestructive testing

金银花为忍冬科植物忍冬 *Lonicera japonica* Thunb.的干燥花蕾或带初开的花^[1],具有清热解毒、疏散风热的功效,可用于治疗和预防严重急性呼吸综合征、甲型流感病毒和手足口病等严重病毒性疾病^[2-3]。随着金银花的市场需求量的不断增加以及其本身较强的适应能力,金银花在中国各地均有种植^[4]。其中,山东省平邑县和河南省封丘县是金银花的道地产区^[5]。不同地区的土壤、气候、海拔、降雨量等自然条件会影响金银花的生长和发育,进而导致不同产地的金银花在质量上存在较大的差异^[5-6]。金银花的主要活性成分包括酚酸类、黄酮类、三萜皂苷类和挥发油类等^[7],其中酚酸类成分含量是金银花的重要质量控制指标之一^[8]。《中国药典》2020 年版中对绿原酸、异绿原酸 A 和异绿原酸 C 3 种酚酸类成分的含量进行了明确的规定。因此,为了保障金银花及其相关中药制剂的质量稳定性,对金银花进行产地鉴别和关键指标性成分含量的准确测定显得尤为重要。

目前,元素分析技术^[9]、高效液相色谱技术^[10]和电子鼻技术^[11]被用于金银花的产地溯源研究,但这些分析技术不可避免的存在分析时间长、操作复杂、成本高等问题。近年来,近红外光谱(near infrared spectroscopy, NIRS)分析技术以其快速、无损、绿色的优点被广泛用于金银花的产地鉴别和有效成分含量测定的研究中,但多数基于 NIRS 技术的研究仍需对金银花样品进行粉碎过筛等处理,难以满足金银花大批量分析及快速、在线质量控制

的需求^[12-14]。大光斑的 NIRS 仪器可以在一定程度上降低样品表面不平整、存在空隙等因素对光谱测量结果的影响^[15],非常适用于分析中药这类质地分布不均匀、形状不规整的复杂样本体系。本研究为解决中药原材料质量在线监测的问题,以不同产地的金银花为研究对象,采用 NIRS 技术结合化学计量学方法,建立了金银花的产地鉴别模型以及金银花中酚酸类成分的含量预测模型,为中药制造过程原料质量的稳定、均一、可控,提供一种无损快速的检测方法。

1 材料与仪器

1.1 药材

在河北省、河南省和山东省 3 个主要产区收集金银花的花蕾样本,每个产区均覆盖多个县乡,共计 147 批次。经国家现代中药创新中心刘岱琳教授鉴定为忍冬科植物忍冬 *L. japonica* Thunb.的花蕾。样品信息如表 1 所示。

表 1 金银花样品产地来源信息

Table 1 Origin information of *Lonicerae Japonicae Flos* samples

样本编号	产地	样本数量
1~23	河南省新乡市封丘县	23
24~115	河北省邢台市巨鹿县	92
116~137	山东省临沂市平邑县	22
138~141	山东省临沂市费县	4
142~147	未知	6

1.2 试剂

对照品绿原酸(批号 110753-202119)、异绿原酸 A(批号 111782-202208)和异绿原酸 C(批号 111894-202205)均购自中国食品药品检定研究院,质量分数大于 98%;色谱纯甲醇(赛默飞世尔科技公司,美国);磷酸为分析级;离子水经 Cascada 超纯水系统(颇尔公司,美国)制得。

1.3 仪器

DA 7440 型近红外光谱仪(波通公司,瑞典), UltiMate 3000 型高效液相色谱仪(赛默飞世尔公司,美国), Quintix224-1CN 型电子分析天平(赛多利斯科学仪器有限公司,德国),超声清洗仪(天津奥特赛恩斯仪器有限公司,FW100 型高速粉碎机(天津市泰斯特仪器有限公司)。

2 方法与结果

2.1 NIRS 数据采集

首先,对金银花进行净制处理,剔除其中的叶子、枝条、碎石等异物。随后,将金银花置于直径为 10.5 cm、深度为 1.5 cm 的圆形样品盘中,摊平,以内置背景为参比扣除背景信息。NIRS 设备的采集波长为 950~1 650 nm,测量时间为 8 s,采用非接触漫反射方式采集光谱信息,光谱仪与样品之间的距离约为 40 cm。每批次样品均进行 3 次独立采集,每次采集完成后更换新的样品进行下一次采集。得到金银花样品的 NIRS 曲线,如图 1 所示。

2.2 酚酸类成分含量测定

采集完 NIRS 数据后,对金银花进行粉碎处理,并制备粒径小于 250 μm 的粉末样品。参照

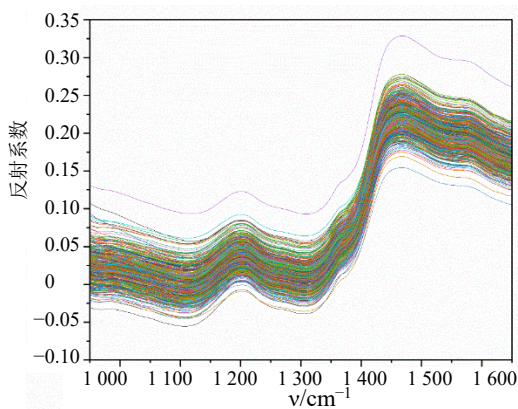


图 1 金银花样品的原始 NIRS 曲线

Fig. 1 Original NIRS curves of *Lonicerae Japonicae Flos* samples

《中国药典》2020 年版一部中金银花含量测定方法测定 147 批次的金银花样品中绿原酸、异绿原酸 A 和异绿原酸 C 的含量,并将其含量作为 NIRS 数据建模的标准参考值。所采用的色谱柱为 Hypersil GOLD™ C₁₈ 色谱柱(250 mm×4.6 mm, 5 μm)。色谱柱温度保持在 25 $^{\circ}\text{C}$,进样体积均为 2 μL 。检测波长为 327 nm。金银花样品的 HPLC 图谱如图 2 所示。

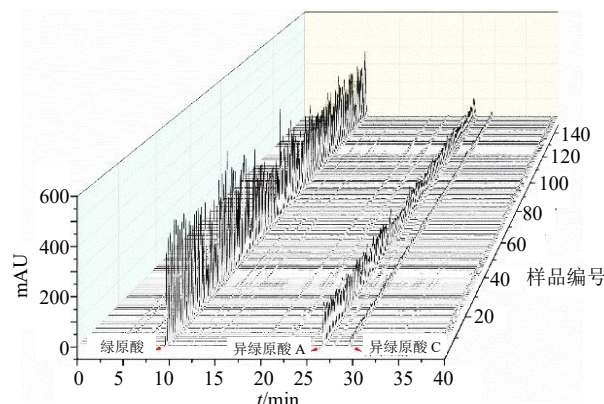


图 2 金银花样品的 HPLC 图谱

Fig. 2 HPLC chromatogram of *Lonicera Japonica Flos* samples

2.3 产地鉴别模型的建立

由于从费县收集的样本较少,不能反映该地区金银花样本的特点。此外,所构建的模型难以捕捉该地区样本的特异性特征,模型泛化能力较差。仅使用封丘县、巨鹿县和平邑县 3 个产地金银花样品的 NIRS 进行产地鉴别模型构建。

2.3.1 建模方法及评价指标的确定 K 近邻(K-nearest neighbor, KNN)算法通过计算待分类样本与校正集中每个样本之间的距离,将 K 个距离最近的样本中出现次数最多的样本标签认定为待分类样本的类别^[16]。KNN 算法的关键是距离度量方法和 K 值的选择,本研究选择欧氏距离计算样本间的距离,以对校正集的留一法交叉验证预测准确率(accuracy of leave-one-out cross validation, LOO-CV ACC)的平均值最小为标准,在 5~15 的中选择最佳的 K 值。

线性判别分析(linear discriminant analysis, LDA)算法通过线性投影,将高维空间中的数据映射到低维空间,使得投影同类样本的投影点尽可能接近,而异类样本的投影点尽可能远离^[17]。奇异值分解(singular value decomposition, SVD)适用于

处理高维小样本数据,因此本研究采用 SVD 求解 LDA 模型。

采用准确率 (accuracy, ACC)、对数损失 (Log loss) 以及 Kappa 系数对分类模型的性能进行评价。一个性能较好的分类器应该具有较高的 ACC 和 Kappa 系数值、较低的 log loss 值。

2.3.2 样本集划分及离群样本剔除 首先从 3 个产地的样品中分别随机选取 30% 数量的样品作为预测集。采用 Savitzky-Golay (SG) 算法、一阶导数 (first derivative, 1st Der) 和标准正态变量变换 (standard normal variate transformation, SNV) 算法依次对划分出预测集后剩余的样本集进行数据预

处理,以增强光谱数据的信号特征。然后采用蒙特卡罗交叉验证 (monte carlo cross-validation, MCCV) 的方法对预处理后的数据集进行离群样本检验^[18]。具体来说,使用蒙特卡罗采样的方法从样本集中选取 70% 的样本,使用 70% 的样本构建 KNN 分类模型并进行模型的训练与优化,对剩余的 30% 的样本进行预测。蒙特卡罗采样重复进行 3 000 次。根据 KNN 分类器对 30% 样本的预测性能 (以 log loss 为评价指标) 对 3 000 进行升序排序,其中出现次数明显偏移其他样本的被判断为离群样本。使用 MCCV 方法得到的累积概率如图 3 所示,共检测出 237、244 和 266 3 个离群样本。

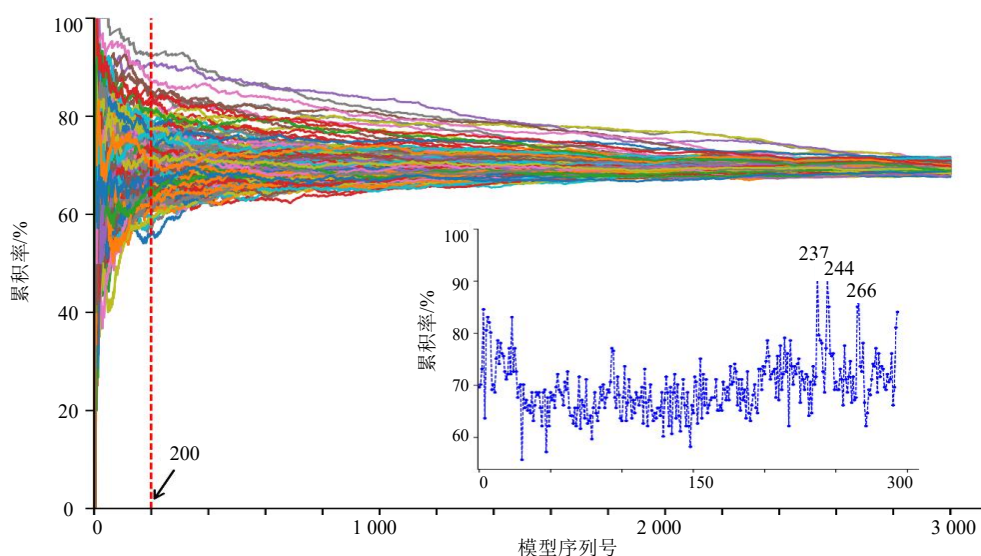


图 3 使用 MCCV 的方法获得的累积概率图

Fig. 3 Accumulated probability plot obtained using MCCV method

离群样本的存在会降低分析模型的预测性能。通过构建 KNN 分类器检测采用 MCCV 方法识别出来的离群样本是否为真正的离群样本。表 2 展示了逐一剔除离群样本前后分类模型的预测性能。从中可以看出剔除 237、244 和 266 3 个样本后 KNN 分类器的性能得到提升。剔除 3 个离群样本后,采用分层随机抽样的方法将剩余样本以 7:3 的比例划分为校正集和验证集。

2.3.3 数据预处理方法选择 NIRS 数据在采集过程中容易受到样本表面颗粒分布不均匀、仪器噪声和基线漂移等因素的影响,因此需要采用适当的预处理方法降低或消除这些因素的影响。根据预处理目的的不同,光谱预处理方法可以分为基线校正、散射校正、平滑处理和尺度缩放 4 类^[19]。其中,基线校正为了扣除仪器背景或漂移对光谱信号

表 2 检测到的离群样本对 KNN 分类器的影响

Table 2 Impact of detected outlier samples on KNN classifier

删除的样本	样本数	K 值	ACC/%	Log loss	LOO-CV AC/%
—	294	7	91.5	0.205	86.4
No.237	293	7	91.8	0.201	86.7
No.244	293	7	91.8	0.203	87.4
No.266	293	7	91.8	0.204	87.0
No.237, No.244	292	7	92.1	0.199	87.7
No.237, No.266	292	7	92.1	0.200	87.3
No.244, No.266	292	7	92.1	0.201	88.0
No.237, No.244, No.266	291	7	92.4	0.197	88.3

的影响,常用方法有 1st Der、二阶导数 (second derivative, 2nd Der) 和连续小波变换 (continuous wavelet transform, CWT) 等。散射校正用于消除由于颗粒分布不均匀及颗粒大小不同产生的散射对

光谱的影响，常用方法有多元散射校正（multiplicative scatter correction, MSC）和 SNV。SG 平滑处理能够消除或降低随机噪声的影响，提高光谱信号的信噪比。尺度缩放用于消除数据尺度差异过大带来的不良影响，常用方法有均值中心化（mean centering, MC）、最大最小归一化（min max Scaler, MMS）和标准化（autoscaling, AS）等^[20]。从表 3 中展示的预处理方法中，依次从每列选取一种预处理方法，排列组合形成 72 种不同的数据预处理方法及其组合方法。

表 3 每种预处理方法的一些常用算法

Table 3 Some commonly used algorithms for each preprocessing method

基线校正	散射校正	平滑	尺度缩放
1 st Der	MSC	SG	MC
2 nd Der	SNV		MMS
			AS

使用 72 种数据预处理方法及其组合方法分别对校正集和验证集进行处理，然后使用校正集分别构建相应的 KNN 和 LDA 分类器，并对验证集进行预测。图 4 展示了使用 72 种预处理及其组合方法处理后的数据构建的分类器对验证集的预测结果。

从中可以发现，NIRS 数据经过预处理后，构建的分类器的预测性能有了一定的提高，其中 KNN 分类器的预测性能提升较为显著，表明数据预处理是数据分析过程中不可缺少的关键环节。最后，以分类器对验证集的预测 ACC 和 Log loss 值为筛选标准，选择 2nd Der-MSC-AS 作为 KNN 分类器的预处理方法，选择 SG-MC 作为 LDA 分类器的预处理方法。使用优选的预处理方法处理后 NIRS 数据构建的分类器的预测结果如表 4 所示，其中 KNN 模型对验证集的预测 ACC 为 92.0%、Log loss 值为 0.270、Kappa 值为 0.842，LDA 模型对验证集的预测 ACC 为 94.3%、Log loss 值为 0.438、Kappa 值为 0.888。

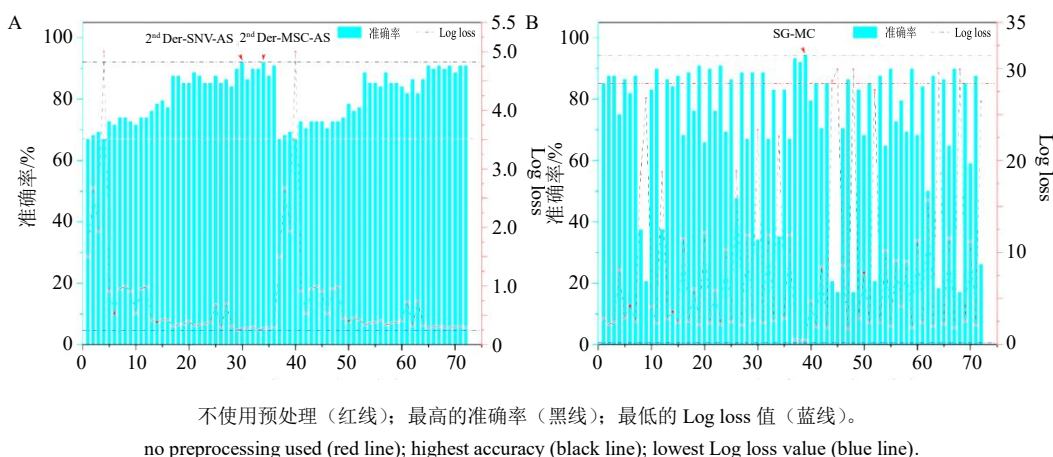


图 4 使用 72 种预处理及其组合方法处理后的数据构建的 KNN (A) 和 LDA (B) 分类器对验证集的预测结果

Fig. 4 Prediction results of validation set by KNN (A) and LDA (B) classifiers constructed from data processed by 72 preprocessing and their combination methods

表 4 使用最优预处理方法处理的数据和原始数据构建的 KNN 和 LDA 分类器的预测结果

Table 4 Prediction results of KNN and LDA classifiers constructed using optimal preprocessing method and raw data

分类器	数据预处理方法	K值	校正集		验证集		
			ACC/%	LOO-CV ACC/%	ACC/%	Log loss	Kappa系数
KNN	—	15	70.4	69.0	67.0	1.501	0.066
	2nd Der-MSC-AS	15	94.1	91.1	92.0	0.270	0.842
LDA	—	—	100.0	89.7	85.2	2.800	0.738
	SG-MC	—	99.5	95.6	94.3	0.438	0.888

2.3.4 特征变量选择 采集到的 NIRS 数据中包含大量的信息，同时也含有噪声和冗余信息。需要采用特征变量选择方法从复杂的信号中筛选出与分

析目标高度相关的变量，从而降低数据维度、降低计算成本。采用排列重要性（permutation importance, PI）方法分别与 KNN 和 LDA 算法进

行组合。该方法通过随机打乱某列特征值后分类器的预测性能变化评估该列特征的重要程度。在 1~100 的范围中寻找使 KNN 和 LDA 分类器预测性能最佳的特征变量数,使用筛选得到的特征波段构建的分类器的预测结果如图 5 所示。从中可以看出,随着选择的特征变量数的增加,2 个分类器的对验

证集的预测 ACC 呈现出先增加后趋于稳定的趋势,但 Log loss 值的波动较大。最后对于 KNN 分类器,选择了 63 个特征变量;对于 LDA 分类器,选择了 90 个特征变量。使用选择的特征变量构建的分类器的预测结果如表 5 所示,对于每个样本的预测结果如图 6 所示。与表 4 中使用全部光谱数据构建的分

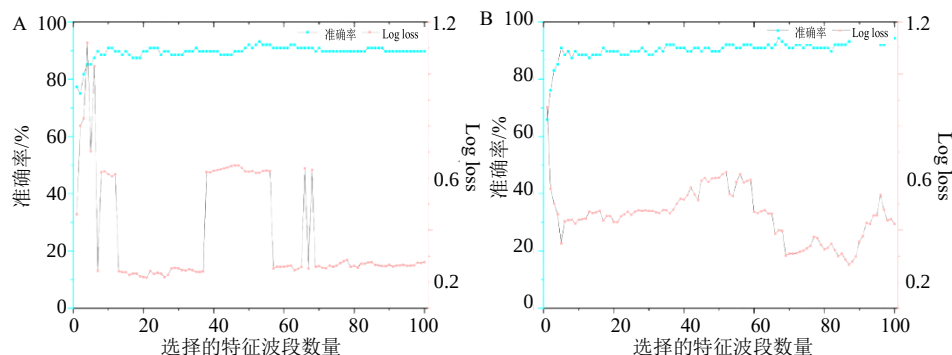


图 5 在 1~100 内选择的特征波段构建的 KNN (A) 和 LDA (B) 分类器对验证集的预测结果

Fig. 5 Prediction results of (A) KNN and (B) LDA classifiers constructed from feature bands selected in range of 1—100 on validation set

表 5 使用 PI 算法筛选得到特征变量构建的分类器的预测结果

Table 5 Prediction results of classifiers constructed using feature variables filtered using PI algorithm

分类器	数据预处理方法	保留的变量数	K值	校正集		验证集		
				ACC/%	LOO-CV ACC/%	ACC/%	Log loss	Kappa系数
KNN	2nd Der-MSC-AS	63	14	0.951	0.931	0.920	0.246	0.842
LDA	SG-MC	90	—	1.000	0.966	0.943	0.356	0.889

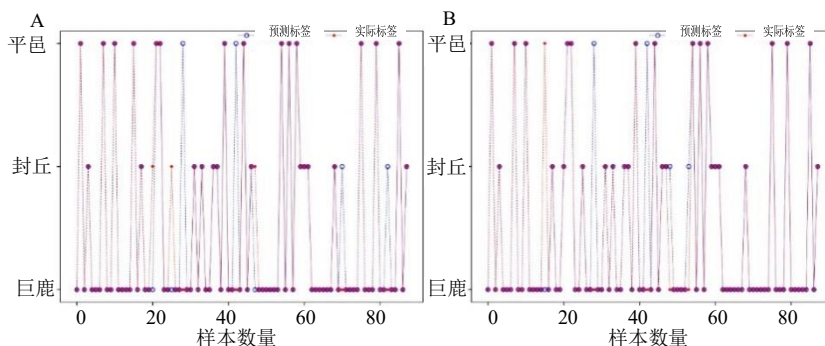


图 6 使用 PI 算法筛选得到特征变量构建的 KNN (A) 和 LDA (B) 分类器的预测结果

Fig. 6 Prediction results of (A) KNN and (B) LDA classifiers constructed using feature variables filtered using PI algorithm

类器相比,使用特征变量构建的分类器的预测 Log loss 值得到了轻微的降低,说明经过特征变量筛选后模型的性能得到提升。

2.3.5 集成学习模型的构建 从图 6 可以看出,由于不同的分类器在计算原理和计算方法上的差异,它们对同一样本进行预测时可能会得到不同预测结果。为了充分利用 2 个分类器的优势,采用加权软投票法整

合 KNN 和 LDA 分类器的预测结果。集成学习模型的构建示意图如图 7 所示。即采用优化的数据处理方法分别对校正集和验证集进行数据处理,分别构建相应的分类器,然后将 2 个分类器对数据集中样本的预测概率的平均值作为集成学习模型的预测结果。

2.3.6 模型性能评价 为了检验所构建的模型在实际应用中的预测性能,采用预测集对优化后的模型进

行性能评价。模型的预测结果如表 6 所示。从中可以看出, KNN 模型对预测集的预测 ACC 为 84.6%、Log loss 值为 0.342、Kappa 系数为 0.693; LDA 模型对预测集的预测 ACC 为 89.4%、Log loss 值为 0.420、Kappa 系数为 0.783, LDA 模型的预测性能优于 KNN 模型。集成学习模型相比于 2 个基学习器模型的性能提升显

著, 对预测集的预测 ACC 提升至 95.7%、Log loss 值提升至 0.214、Kappa 系数提升至 0.907。图 8 展示了构建的集成学习模型对每个产地金银花的识别准确率, 其中对巨鹿县产的金银花的识别 ACC 为 98.8%, 对封丘县产的金银花的识别 ACC 为 83.3%, 对平邑县所产的金银花的识别 ACC 为 94.4%。

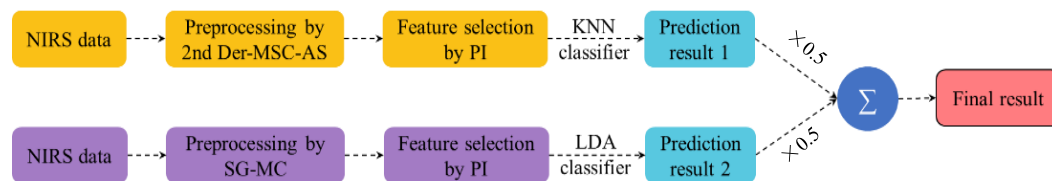


图 7 集成学习模型的构建流程图

Fig. 7 Flowchart of construction of integrated learning model

表 6 优化后的 KNN 和 LDA 分类器以及集成学习模型的预测结果

Table 6 Prediction results of optimized KNN and LDA classifiers and integrated learning model

分类模型	K值	校正集		预测集		
		ACC/%	LOO-CV ACC/%	ACC/%	Log loss	Kappa系数
KNN	14	95.1	93.1	84.6	0.342	0.693
LDA	—	100.0	96.6	89.4	0.420	0.783
集成学习	—	100.0	—	95.7	0.214	0.907

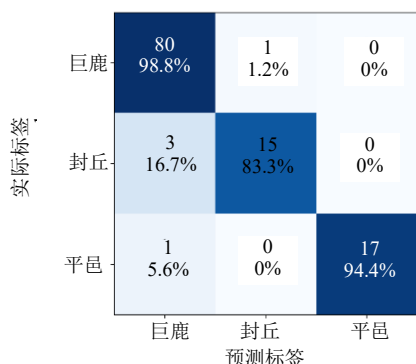


图 8 构建的集成学习模型对预测集的预测混淆矩阵

Fig. 8 Predictive confusion matrix of constructed integrated learning model on prediction set

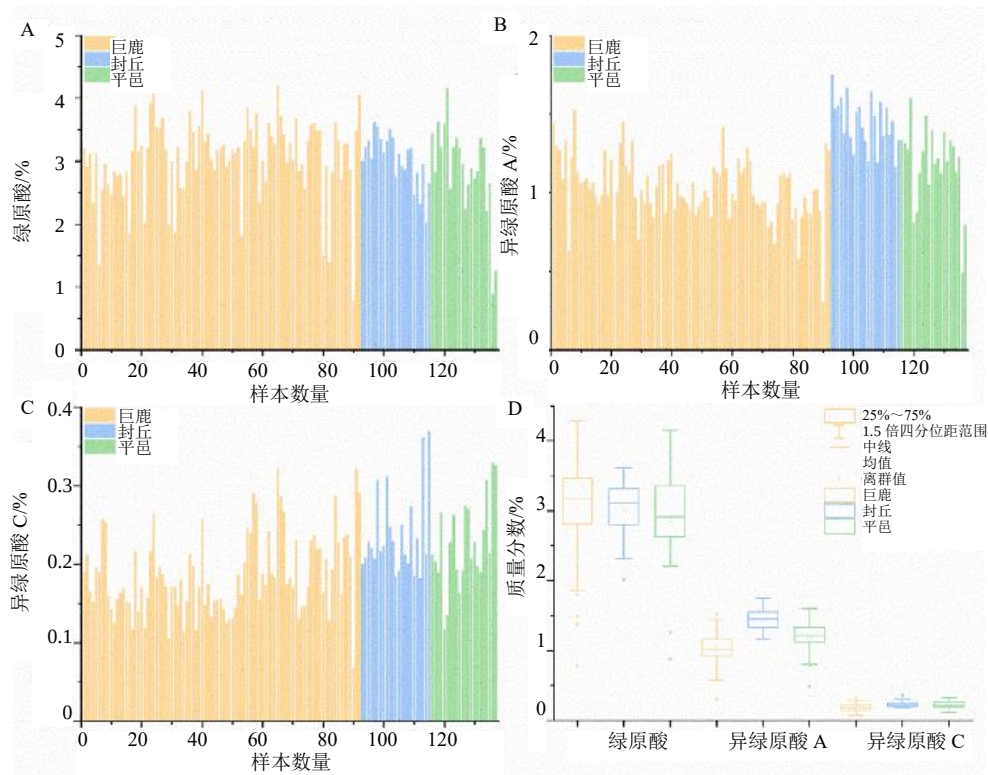
2.4 酚酸类成分预测模型的建立

2.4.1 酚酸类成分含量的统计分析 对收集到的巨鹿县、封丘县、平邑县的金银花中绿原酸、异绿原酸 A 和异绿原酸 C 成分进行统计学分析。图 9 展示了 3 个产地金银花中 3 种酚酸类成分的含量分布差异。从图 9-A~C 可以看出不同产地 3 种酚酸类成分含量存在较大的差异。图 9-D 可以看出, 封丘县和巨鹿县的金银花中绿原酸成分的平均含量较高, 封丘县和平邑县的金银花中异绿原酸 A 和异绿原酸 C 成分的平均含量较高, 说明封丘县生产的金

银花质量较佳。此外, 封丘县的金银花中 3 种酚酸类成分的箱线图的箱体长度明显短于其他 2 个产地, 说明封丘县生产的金银花质量较为稳定。然而, 同一产地不同批次的样品在质量上仍然存在较大的差异, 因此需要对每批次的金银花进行质量评价。

2.4.2 建模方法及评价指标的确定 支持向量机 (support vector machine, SVM) 回归算法的原理是通过最小化预测误差来拟合数据, 并且在拟合过程中保持一个边界 (间隔), 使得大部分数据点落在边界之内^[21]。SVM 算法的关键参数包括核函数、惩罚系数 C 和核函数宽度 gamma。本研究使用的核函数为径向基核函数 (radial basis function, RBF), 使用贝叶斯参数优化算法对 C (范围 0.01~100) 和 gamma (范围 0.01~100) 2 个参数进行优化。

随机森林 (random forest, RF) 算法采用有放回地随机抽样方式生成多个样本集并构建决策树, 将多棵决策树预测结果的平均值作为最终预测结果, 以得到更准确、更稳定的预测结果^[22]。采用贝叶斯参数优化算法对最大深度 (范围 1~20)、再划分所需最小样本数 (范围 2~10)、节点最少样本数 (范围 1~10)、决策树的个数 (范围 100~500) 以



A-绿原酸; B-异绿原酸 A; C-异绿原酸 C; D-成分含量对应的箱线图。
A-chlorogenic acid; B-isochlorogenic acid A; C-isochlorogenic acid C; D-box plots corresponding to component content.

图 9 不同产地的金银花中 3 种酚酸成分含量的统计分析结果

Fig. 9 Statistical analysis results of content of three phenolic acid components in *Lonicerae Japonicae Flos* from different regions

及最大特征数（范围 0.100~0.999）进行优化。RF 和 SVM 算法均通过寻找 10 折交叉验证下，使校正集的交叉验证均方根误差（root mean square error of cross-validation, RMSECV）最小的参数组合。

采用决定系数（coefficient of determination, R^2 ）、交叉验证决定系数（coefficient of determination of cross validation, R^2_{CV} ）、均方根误差（root mean square error, RMSE）、RMSECV 以及相对预测偏差（relative percent deviation, RPD）评价回归模型的预测精度和稳定性。一个稳定可靠的模型应该具有较高的 R^2 和 RPD 以及较低的 RMSE。

2.4.3 样本集划分 采用 Kennard-Stone (KS) 算法将划分出预测集后剩余的样本按照 7 : 3 的比例划分为校正集和验证集。4 个样本集中 3 种成分的含量分布如表 7 所示。其中校正集和验证集中 3 种成分的含量分布差异较小，有利于构建稳健的分析模型。预测集的含量分布范围略大于校正集和验证集，可用于检验模型的泛化能力。

2.4.4 数据预处理方法选择 采用 72 种数据预处理方法对校正集和验证集进行数据预处理，然后构建相应的 SVM 和 RF 模型。根据模型对验证集的预测性能，为每种算法选择合适的预处理方法。表 8

表 7 不同样本集中 3 种酚酸类成分的含量统计结果

样本集	绿原酸/%			异绿原酸A/%			异绿原酸C/%		
	范围	平均值	标准差	范围	平均值	标准差	范围	平均值	标准差
所有样本	0.78~4.28	3.02	0.62	0.30~1.75	1.13	0.25	0.07~0.37	0.20	0.06
校正集	1.25~4.28	3.02	0.62	0.63~1.66	1.13	0.22	0.11~0.37	0.20	0.05
验证集	1.25~4.12	3.01	0.46	0.67~1.66	1.12	0.22	0.12~0.32	0.20	0.06
预测集	0.78~4.20	3.01	0.72	0.30~1.75	1.13	0.31	0.07~0.36	0.21	0.06

表 8 使用最优预处理方法处理的数据和原始数据构建的 SVM 和 RF 回归器的预测结果

Table 8 Prediction results of SVM and RF regressors constructed using optimal preprocessing method and raw data

成分	回归模型	数据预处理方法	超参数	校正集				验证集		
				R^2	RMSE/%	R^2_{cv}	RMSE _{cv} /%	R^2	RMSE/%	RPD
绿原酸	SVM	—	100.000, 4.443	0.733 5	0.318 6	0.489 3	0.402 5	0.500 3	0.326 2	1.422 0
		SG-2 nd Der-SNV	7.211, 1.673	0.969 8	0.107 2	0.745 0	0.280 1	0.752 7	0.229 5	2.021 5
	RF	—	11, 1, 2, 137, 0.999	0.881 8	0.212 2	0.201 7	0.539 9	0.268 0	0.394 8	1.174 9
		SG-2 nd Der-SNV	19, 1, 2, 462, 0.221	0.965 6	0.114 5	0.703 0	0.314 4	0.610 1	0.288 1	1.609 8
异绿原酸A	SVM	—	94.419, 7.090	0.760 5	0.109 8	0.5442	0.146 6	0.578 7	0.145 0	1.548 7
		1 st Der-SNV	42.565, 0.977	0.890 7	0.074 2	0.720 8	0.112 3	0.749 6	0.111 8	2.008 8
	RF	—	3, 1, 2, 356, 0.406	0.245 5	0.194 9	-0.091 7	0.227 6	0.046 3	0.218 2	1.029 3
		SG-1 st Der-SNV	15, 1, 2, 154, 0.500	0.959 3	0.045 3	0.655 7	0.126 6	0.698 3	0.122 7	1.829 9
异绿原酸C	SVM	—	0.115, 12.941	-0.051 7	0.054 2	-0.178 3	0.055 0	-0.121 3	0.058 0	0.949 3
		SG-MSC-MMS	25.384, 0.037	0.066 6	0.051 0	-0.049 3	0.051 8	0.230 5	0.048 1	1.145 9
	RF	—	17, 1, 3, 288, 0.830	0.869 7	0.019 1	0.045 0	0.050 3	-0.078 6	0.056 9	0.967 9
		2 nd Der-SNV	12, 1, 2, 448, 0.795	0.927 2	0.014 3	0.426 0	0.039 1	0.567 9	0.036 0	1.529 2

展示了使用优选的预处理方法处理后的数据集构建的回归模型的预测结果。相较于使用原始 NIRS 数据构建的模型，其性能得到极大的提升。对于绿原酸成分，SVM 和 RF 模型均选择 SG-2nd Der-SNV 预处理方法；对于异绿原酸 A 成分，SVM 和 RF 模型分别选择 1st Der-SNV 和 SG-1st Der-SNV 预处理方法；对于异绿原酸 C 成分，SVM 和 RF 模型分别选择 SG-MSC-MMS 和 2nd Der-SNV 预处理方法。

2.4.5 特征变量筛选 采用无信息变量消除 (uninformative variable elimination, UVE) 算法提取与酚酸类成分相关的特征波段。鉴于 UVE 算法每次引入的噪声的随机性，每次运行 UVE 算法得到特征波段均有差异。本研究通过重复运行 20 次 UVE 算法，从中选取对验证集预测性能最佳的特征波段组合。表 9 展示了使用 UVE 算法筛选得到的特征变量构建的回归模型的预测结果，从中可以看出经

过变量筛选后，大多数模型的预测性能得到提升。只有使用特征变量构建 SVM 回归模型预测绿原酸成分时，相比于全部光谱数据构建的模型的性能有所降低。为了降低无关信息变量对模型性能的影响，使用 UVE 筛选得到的特征变量构建相应的成分含量预测模型。其中绿原酸和异绿原酸 A 成分含量使用 SVM 模型进行预测，异绿原酸 C 成分含量使用 RF 模型进行预测。

2.4.6 模型性能评价 采用预测集评价优化后的回归模型的预测性能，模型的预测结果如表 10 所示。回归模型的预测结果与实际测量结果的散点图如图 10 所示。其中对绿原酸成分的预测 R^2 为 0.8102，RMSE 为 0.313 6%，RPD 为 2.305 2，对异绿原酸 A 成分的预测 R^2 为 0.804 0，RMSE 为 0.135 4%，RPD 为 2.268 2，对异绿原酸 C 成分的预测 R^2 为 0.271 0，RMSE 为 0.052 0%，RPD 为 1.176 3。说明所构建的

表 9 使用 UVE 算法筛选得到的特征变量构建的 SVM 和 RF 回归模型的预测结果

Table 9 Prediction results of SVM and RF regression models constructed using feature variables obtained by UVE algorithm

成分	回归模型	变量数	超参数	校正集				验证集		
				R^2	RMSE/%	R^2_{cv}	RMSE _{cv} /%	R^2	RMSE/%	RPD
绿原酸	SVM	49	27.761, 1.274	0.932 2	0.160 7	0.773 6	0.265 3	0.729 2	0.240 1	1.931 6
	RF	61	12, 1, 2, 326, 0.322	0.965 8	0.114 2	0.718 0	0.307 3	0.617 5	0.285 4	1.625 4
异绿原酸A	SVM	41	67.292, 1.041	0.896 6	0.072 1	0.782 6	0.097 9	0.760 1	0.109 5	2.052 1
	RF	53	13, 1, 2, 478, 0.999	0.960 6	0.044 5	0.647 5	0.127 4	0.708 3	0.120 7	1.861 3
异绿原酸C	SVM	102	9.767, 0.048	0.075 9	0.050 8	-0.039 6	0.051 6	0.243 6	0.047 7	1.155 8
	RF	59	19, 3, 2, 431, 0.916	0.891 4	0.017 4	0.467 9	0.037 5	0.580 2	0.035 5	1.551 5

表 10 优化后的回归模型对预测集的预测结果

Table 10 Prediction results of optimized regression models on prediction set

成分	分类器	超参数	预测集		
			R^2	RMSE/%	RPD
绿原酸	SVM	27.761、1.274	0.810 2	0.313 6	2.305 2
异绿原酸A	SVM	67.292、1.041	0.804 0	0.135 4	2.268 2
异绿原酸C	RF	19、3、2、431、0.916	0.271 0	0.052 0	1.176 3

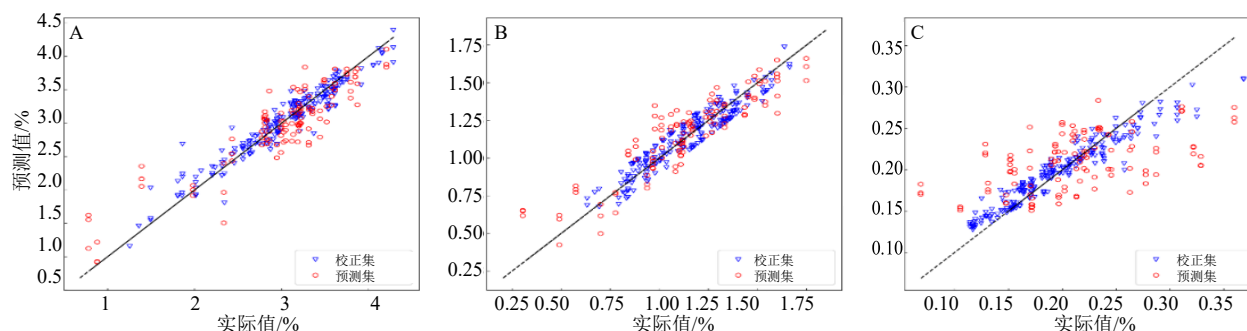


图 10 优化回归模型的预测结果和绿原酸 (A)、异绿原酸 A (B) 和异绿原酸 C (C) 的实际结果的散点图

Fig. 10 Scatter plots of predicted results of optimized regression models and actual results for chlorogenic acid (A), isochlorogenic acid A (B) and isochlorogenic acid C (C)

回归模型可以较为准确的预测绿原酸和异绿原酸 A 成分的含量,但是对异绿原酸 C 成分的预测效果较差。这可能是因为异绿原酸 C 成分在金银花中含量相对较低,所使用的 NIRS 设备的检测限较高。在不破坏金银花样品的情况下,难以充分采集异绿原酸 C 成分相关的信息。

3 讨论

金银花作为一味药食两用的中药材,具有极大的经济价值。本研究建立了基于 NIRS 技术的金银花产地鉴别方法和酚酸类成分的含量测定方法,其中基于加权软投票法构建的集成学习模型可以更为准确的预测金银花的产地,对预测集样本的预测 ACC 为 95.7%、Log loss 值为 0.214、Kappa 系数为 0.907。此外,所建立定量分析模型可以较为准确的预测绿原酸和异绿原酸 A 成分的含量,绿原酸成分 $R^2=0.810\ 2$, $RMSE=0.313\ 6\%$, $RPD=2.305\ 2$; 异绿原酸 A 成分 $R^2=0.804\ 0$, $RMSE=0.135\ 4\%$, $RPD=2.268\ 2$,但对异绿原酸 C 成分的预测效果较差, $R^2=0.271\ 0$, $RMSE=0.052\ 0\%$, $RPD=1.176\ 3$ 。结果表明,本方法适用于金银花的产地鉴别以及绿原酸和异绿原酸 A 成分的含量测定,可以用于快速、无损的评价金银花的质量。目前,高光谱成像技术已被应用于金银花中质量标志性成分含量的无损检测研究,且预测精度较高^[23]。然而,后续数据分

析过程中需要从图像中提取光谱信息,耗时较长,在一定程度上限制了该技术在实时监测方面的应用。相较之下,NIRS 技术能够直接获取光谱信息,同时兼顾了成本效益,为金银花质量检测提供了一种更为高效、经济的解决方案。本研究证明了金银花的产地鉴别和质量标志物含量测定可以在不对其进行粉碎处理的前提下进行,基于本方法可开发一种基于 NIRS 技术的金银花质量的在线评估设备,实现金银花生产过程中的实时监测与质量控制。

此外,金银花中酚酸类成分的含量测定结果显示,不同产地、同一产地不同批次间的金银花中的酚酸类成分含量存在较大的差异。这说明金银花的质量不仅与土壤、气候、水文等生态因子的紧密关联,而且受到种植管理、采收时间、加工方式等多重因素的影响。因此,为了提高金银花的质量稳定性和一致性,需要从生态环境选择、种植管理优化、采收时间控制和加工方式改进等多个方面入手,建立科学、规范的生产和质量控制体系。多项研究发现,金银花的外观性状(如颜色、花蕾直径和长度)与其内部主要有效成分之间存在较为显著的相关性^[24-26]。然而,传统“辨状论质”高度依赖个人经验与主观判断,面临着传承困难与现代化挑战,其局限性日益凸显。未来可以探索可见光图像等机器视觉技术与 NIRS 分析技术相结合的新方法,借助

集成自动化图像处理和智能识别算法,实现对采集信息的即时、高效处理。进一步,通过综合分析其外观性状与内部成分信息,构建一个全面且精准的在线品质评估体系,从而满足大规模生产和质量控制场景需求。

综上所述,NIRS 分析技术作为一种高效、无损的检测手段,在金银花原材料的质量初步评估中展现出了巨大的潜力和价值。同时本研究方法也为其他中药材的质量评估提供了重要的科学参考。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 中国药典 [S]. 一部. 2020: 230-231.
- [2] Yang B X, Zhong Z H, Wang T T, *et al.* Integrative omics of *Lonicera japonica* Thunb. flower development unravels molecular changes regulating secondary metabolites [J]. *J Proteomics*, 2019, 208: 103470.
- [3] Li M W, Wang Y X, Jin J, *et al.* Inhibitory activity of honeysuckle extracts against influenza A virus *in vitro* and *in vivo* [J]. *Virol Sin*, 2021, 36(3): 490-500.
- [4] 马涵玉, 钱琪, 牛丽颖. 基于文献计量学的金银花研究现状与热点分析 [J]. *药物评价研究*, 2022, 45(7): 1426-1434.
- [5] Zhang J, Yu X, Yang R, *et al.* Quality evaluation of *Lonicerae Japonicae Flos* from different origins based on high-performance liquid chromatography (HPLC) fingerprinting and multicomponent quantitative analysis combined with chemical pattern recognition [J]. *Phytochem Anal*, 2024, 35(4): 647-663.
- [6] Gu L F, Xie X Q, Wang B, *et al.* Discrimination of *Lonicerae Japonicae Flos* according to species, growth mode, processing method, and geographical origin with ultra-high performance liquid chromatography analysis and chemical pattern recognition [J]. *J Pharm Biomed Anal*, 2022, 219: 114924.
- [7] 刘璐, 吕凌, 李晨, 等. 基于多指标成分和模式识别分析不同产地金银花的质量差异 [J]. *中华中医药杂志*, 2023, 38(9): 4328-4335.
- [8] Cai Z C, Wang C C, Chen C H, *et al.* Quality evaluation of *Lonicerae Japonicae Flos* and *Lonicerae Flos* based on simultaneous determination of multiple bioactive constituents combined with multivariate statistical analysis [J]. *Phytochem Anal*, 2021, 32(2): 129-140.
- [9] 魏静娜, 李安, 张强, 等. 基于稳定同位素与矿物元素指纹特征的金银花产地溯源研究 [J]. *农产品质量与安全*, 2023(1): 73-80.
- [10] 刘天亮, 杨林林, 董诚明, 等. 基于化学模式识别的不同产地金银花 HPLC 指纹图谱研究 [J]. *中草药*, 2022, 53(15): 4833-4843.
- [11] 刘鹏, 白上圆, 张维瑞, 等. 电子鼻技术鉴别金银花真伪、产地、贮藏年份 [J]. *中成药*, 2021, 43(7): 1960-1963.
- [12] 耿泽宇, 左旭丽, 潘秀珍, 等. 金银花近红外定性定量快检模型的建立 [J]. *中南药学*, 2023, 21(8): 2076-2082.
- [13] 耿姝, 陈勇, 金叶, 等. 基于近红外技术的金银花药材多指标成分快速检测 [J]. *中国现代应用药学*, 2016, 33(2): 144-150.
- [14] 郝海群. 近红外光谱法测定金银花中绿原酸含量 [J]. *河南中医*, 2015, 35(3): 640-642.
- [15] 韩燕雨, 吴志生, 戴平, 等. 多源信息融合的中药制造白术饮片多维质量属性一致性的诊断策略 [J]. *中国中药杂志*, 2024, 49(9): 2434-2440.
- [16] 侯富国, 桂新景, 王艳丽, 等. 基于智能视觉技术的白及饮片真伪快速辨识方法研究 [J]. *中草药*, 2023, 54(2): 509-519.
- [17] Bellou E, Gyftokostas N, Stefan D, *et al.* Laser-induced breakdown spectroscopy assisted by machine learning for olive oils classification: The effect of the experimental parameters [J]. *Spectrochim Acta Part B At Spectrosc*, 2020, 163: 105746.
- [18] Liu Z C, Cai W S, Shao X G. Outlier detection in near-infrared spectroscopic analysis by using Monte Carlo cross-validation [J]. *Sci China Ser B Chem*, 2008, 51(8): 751-759.
- [19] 第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究 [J]. *光谱学与光谱分析*, 2019, 39(9): 2800.
- [20] Wu S J, Wang L, Zhou G M, *et al.* Strategies for the content determination of capsaicin and the identification of adulterated pepper powder using a hand-held near-infrared spectrometer [J]. *Food Res Int*, 2023, 163: 112192.
- [21] Wedding B B, Wright C, Grauf S, *et al.* The application of FT-NIRS for the detection of bruises and the prediction of rot susceptibility of 'Hass' avocado fruit [J]. *J Sci Food Agric*, 2019, 99(4): 1880-1887.
- [22] Maione C, Batista B L, Campiglia A D, *et al.* Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry [J]. *Comput Electron Agric*, 2016, 121: 101-107.
- [23] Wang Q Q, Liu Y H, Gao X W, *et al.* Potential of hyperspectral imaging for nondestructive determination of chlorogenic acid content in *Flos Lonicerae* [J]. *J Food Meas Charact*, 2019, 13(4): 2603-2612.
- [24] 王玲娜, 孙希芳, 张芳, 等. 不同发育时期金银花颜色与活性成分的相关性分析 [J]. *中草药*, 2017, 48(15): 3182-3188.
- [25] 唐志强, 程雪梅, 刘谦, 等. 金银花外观性状与主要有效成分含量的相关性研究 [J]. *山东农业科学*, 2023, 55(11): 115-121.
- [26] 熊乐文, 金莹, 黄玮慕, 等. 金银花色泽与酚酸类成分含量相关性研究 [J]. *中成药*, 2022, 44(8): 2736-2739.

[责任编辑 时圣明]