基于基原效用差异的黄连品质辩识光谱化学表征模型构建

黄　玲[1]，齐路明[2]，王　科[3]，李　娜[1]，董继晶[1]，马云桐[1*]

1. 成都中医药大学药学院，四川 成都　611137
2. 成都中医药大学养生康复学院，四川 成都 611137
3. 成都工业学院大数据与人工智能学院，四川 成都　611730

摘　要：目的　中药多基原物种间的效用差异性受到历代医家的高度重视，联用傅里叶变换近红外光谱（fourier transform near infrared spectroscopy，FT-NIR）和傅里叶变换中红外光谱（fourier transform mid-infrared spectroscopy，FT-MIR）技术，考察光谱化学表征技术应用于多基原黄连共有物质基础测定和品质辨识的可行性。**方法**　以黄连的 4 种基原（黄连 *Coptis chinensis*、三角叶黄连 *C. deltoidea*、峨眉野连 *C. omeiensis* 和云南黄连 *C. teeta*）的共有物质基础小檗碱、黄连碱、木兰花碱、非洲防己碱和巴马汀为研究对象，基于光谱矩阵的优化和特征学习算法，串联近红外和中红外光谱特征构建偏最小二乘回归（partial least squares regression，PLSR）和支持向量回归（support vector regression，SVR）光谱化学表征模型以测定药材中共有活性成分的含量，辨识其品质差异。**结果**　SVR 模型对小檗碱的测定效果最优，剩余预测偏差（residual predictive deviation，RPD）值高达 4.842，模型对黄连碱、巴马汀和木兰花碱含量的预测 RPD 值均大于 2；PCA 结果表明所建立的模型能有效鉴别多基原黄连共有成分含量的差异，为其品质辨识提供依据。**结论**　多元光谱技术串联应用可有效表征多基原黄连药材中共有成分的含量差异，提高多基原中药品质辨识的效率。

关键词：黄连；三角叶黄连；峨眉野连；云南黄连；红外光谱技术；偏最小二乘回归；支持向量回归；光谱化学表征模型；小檗碱；黄连碱；木兰花碱；非洲防己碱；巴马汀

中图分类号：R286.2　　**文献标志码：**A　　**文章编号：**0253 - 2670(2022)20 - 6343 - 11

DOI: 10.7501/j.issn.0253-2670.2022.20.005

Construction of spectrochemical characterization model for quality identification based on utility difference of multi-source *Coptidis Rhizoma*

HUANG Ling[1], QI Lu-ming[2], WANG Ke[3], LI Na[1], DONG Ji-jing[1], MA Yun-tong[1]

1. School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China
2. School of Rehabilitation and Health Preservation, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China
3. School of Big Data and Artifical Intelligence, Chengdu Technological University, Chengdu 611730, China

Abstract: Objective　The difference in the utility of the multi-source species of traditional Chinese medicine has been paid great attention by the doctors in the past dynasties. In this study, FT-NIR and FT-MIR technologies were combined to investigate the feasibility of applying spectrochemical characterization techniques to the determination and quality identification of the common material basis of multi-source Huanglian (*Coptidis Rhizoma*). **Methods**　Taking the common material basis of four *Coptidis Rhizoma*, coptisine, magnoflorine, berberine, columbamine, and palmatine as the research objects, optimization and feature learning algorithms based on spectral matrix, combining near-infrared and mid-infrared spectroscopy characteristics to construct PLSR and SVR spectrochemical characterization models to determine the content of common active ingredients in medicinal materials and identify their quality differences. **Results**　The SVR model had the best effect on the determination of berberine, with an *RPD* value of 4.842. The predicted *RPD* values of the model for berberine, palmatine and magnoflorine content were all higher than 2; PCA results showed that the model could effectively identify the differences in the common components of multi-source *Coptidis*

收稿日期：2022-03-09

基金项目：国家重要野生植物种质资源共享平台

作者简介：黄　玲（1994—），女，在读硕士研究生，研究方向为中药品种、品质与资源开发研究。
Tel: 18482134481　E-mail: huangling0108@163.com

***通信作者：**马云桐（1963—），男，博士生导师，教授，研究方向为中药品种、品质与资源开发研究。E-mail: mayuntong06@163.com

Rhizoma, and provide a basis for its quality identification. **Conclusion** The serial application of multiple spectrum technology can effectively characterize the difference in the content of the common components of the multi-based original *Coptidis Rhizoma*, and improve the efficiency of quality identification of the multi-based original Chinese medicine.

Key words: *Coptis. chinensis* Franch.; *C. deltoidea* C. Y. Cheng et Hsial; *C. omeiensis* C. Y. Cheng; *C. teeta* Wall.; infrared spectroscopy; partial least squares regression; support vector regression; spectrochemical characterization model; berberine; coptisine; magnoflorine; columbamine; palmatine

黄连为黄连 *Coptis chinensis* Franch.、三角叶黄连 *C. deltoidea* C. Y. Cheng et Hsial、云南黄连 *C. teeta* Wall.、峨眉野连 *C. omeiensis* C. Y. Cheng 等黄连属植物的干燥根茎。依据古代典籍记载，虽然这些资源均可当作黄连药材使用，但其具体生物效用有较大差异。如《唐本草》记载：“黄连，蜀道者粗大，味极浓苦，疗渴为最；江东者节如连珠，疗痢大善”，指出川产黄连具有较好的治疗糖尿病的效果，而长江以东的黄连则更善于抑菌止痢。目前相关研究已经提供有力的证据，指出导致不同基原黄连药材效用差异的主要原因表现为其共有物质基础的含量差异[1-2]。结合课题组前期研究，次生代谢产物类型的相似性是不同基原黄连属植物根茎可作为同一中药使用的基础，也是其均可用于治疗糖尿病、阿尔茨海默症等疾病的重要原因。原小檗碱类生物碱为黄连药材的主要化学成分群，包括小檗碱、黄连碱、木兰花碱、巴马汀和非洲防己碱等[3-4]，这些活性成分的含量差异正是多基原黄连效用同中有异的主要原因。对该类化学成分的含量进行有效表征，可明确不同来源黄连药材共有物质基础的差异，为其品质差异鉴别提供有效的证据，并且对于该药材资源的合理利用具有积极的意义。

针对中药复杂体系中物质基础成分的快速测定，科研工作者已经进行了许多有益探索。其中，光谱技术表现出巨大的潜力。相对而言，该技术操作简单，无需样品试剂损耗，且能够全面展示样品中的代谢成分信息，是一种绿色环保的测定方法。但是中药的光谱信息变量复杂，易于受到外界因素干扰，如何从复杂的光谱变量矩阵中提取目标变量是阻碍其进一步应用的关键问题。目前，多种化学计量学的方法已经应用于光谱数据的优化，以求简化高维的变量矩阵并提高其可用性。例如多种优化去噪、特征学习和信息融合的算法，都是提升光谱技术应用潜力的有效方法[5-7]。通常而言，一套完整的光谱矩阵分析流程十分复杂，包括预处理、异常值诊断、特征学习和数学模型构建等多个步骤，且需要进行严格的优化。迄今，光谱化学表征技术在多基原黄连药材共有化学成分测定及其品质差异的辨识上的应用相对较少。

基于此，本研究结合课题组前期研究中的高效液相色谱数据[8]，选择多基原黄连药材中主要共有化学成分小檗碱、黄连碱、木兰花碱、非洲防己碱和巴马汀，考察光谱化学表征技术应用于多基原黄连共有活性成分测定和品质辨识的可行性。应用傅里叶变换近红外光谱（fourier transform near infrared spectroscopy，FT-NIR）和傅里叶变换中红外光谱（fourier transform mid-infrared spectroscopy，FT-MIR）技术采集不同基原黄连药材的光谱信息，结合化学计量学算法，构建一套完整的光谱矩阵分析流程。偏最小二乘回归（partial least squares regression，PLSR）和支持向量回归（support vector regression，SVR）算法被用来建立该药材的光谱化学含量的表征模型，考察其与高效液相色谱数据的相关性，探讨红外光谱应用于该类化学成分快速测定和差异辨识的可行性，为多基原黄连药材的品质辨识提供依据。

1 材料与仪器

1.1 样品

所用黄连药材来源于 4 种黄连属植物，经成都中医药大学马云桐教授鉴定，分别为黄连 *C. chinensis* Franch.、三角叶黄连 *C. deltoidea* C. Y. Cheng et Hsial、峨眉野连 *C. omeiensis* C. Y. Cheng 和云南黄连 *C. teeta* Wall.。前 3 种植物采集于四川省洪雅县黑山村人工种植基地，云南黄连采集于云南省福贡县匹河乡人工种植基地，采集样品均为 5 年生植物。在其药材采收期收集并参照产地加工方法，取其根茎部位，洗净后 60 ℃烘干，保存在阴凉干燥处。得到 4 种黄连药材分别称为味连、雅连、野连和云连。

1.2 仪器与试剂

PerkinElmer 傅里叶近红外和中红外光谱仪（美国珀金埃尔默仪器有限公司），DFT-50A 型手提式高速粉碎机（林大机械有限公司，浙江温岭），电子天平（赛多利斯科学仪器有限公司，北京），LC-20A 型高

效液相色谱仪（日本岛津公司）。对照品小檗碱（批号 110713-201814）购自于中国食品药品检定研究院，木兰花碱（批号 CHB180205）、非洲防己碱（批号 CHB180712）、黄连碱（批号 CHB180629）和巴马汀（批号 CHB180226）购于成都克洛玛试剂公司，质量分数均≥98%。

2 方法

2.1 样品检测

称取适量样品，均匀放置于样品杯中，应用傅里叶近红外和中红外光谱仪，分别采集 FT-NIR 和 FT-MIR 光谱特征。针对每一个样品，检测区间分别设定为 10 000～4000 cm^{-1} 和 4000～500 cm^{-1}，仪器分辨率均为 4 cm^{-1}，信号累积 64 次。在样品测定之前，首先测定空气中水和二氧化碳引起的背景信号，并将其从样品吸收峰中自动删除。每一个样品重复测定 3 次，平均光谱用于后续分析。色谱数据的测定参考课题组前期的研究成果[8]。

2.2 光谱数据分析流程

2.2.1 4 种光谱预处理算法应用于原始光谱优化平滑算法（11 点），去除噪音信号[9]；多元散射校正和标准正态变量，消除光散射影响[10]；导数算法，减小基线漂移并增强样品特征信息[11]。

2.2.2 异常值诊断 采用基于 PLSR 的 Hotelling T2[12]检验诊断数据分布。首先基于 *X* 和 *Y* 矩阵建立 PLSR 模型，根据 T2 值来甄别数据集中的异常值。当样本的 T2 值高于 99%置信区间，该样本被设定为异常样本并删除。

2.2.3 数据标准化 光谱和色谱数据含有不同的量纲，将两者数据标准化到同一数量级，能够提升模型的收敛速度和增加准确性，因此将数据缩减至［-1，1］之间[13]。

2.2.4 光谱特征数据筛选与评价 红外光谱数据变量复杂，代表样品全面的化学信息，也会产生大量的无关信号，不仅增加模型运行时间，还会降低模型的准确性和推广性。运用 4 种特征学习机器算法对光谱变量的重要性进行排序，分别是递归式特征消除（recursive feature elimination，RFE）[14]、Boruta 算法[15]、变量投影重要性算法（variable importance in projection，VIP）[16]和基尼指数（gini coefficient，GINI）[17]，应用十折交叉验证评价特征变量数目。

2.2.5 特征信息融合：特征级数据融合是一种中等水平的融合策略，将来源于不同传感器的特征信息加以综合，可以产生比单一信息源更精确、更完全、更可靠的估计和判断[18]。本研究将来源于 FT-NIR 和 FT-MIR 光谱的特征变量进行信息融合，进一步提高多基原黄连药材共有化学成分表征模型的准确性和有效性。

2.3 光谱化学表征模型

首先应用 PLSR[19]算法建立 5 种共有化学成分的光谱化学表征模型。该方法依据最大协方差原则，计算复杂矩阵中变量（*X*）和（*Y*）之间的关系。该方法的优点是可以很好地克服多元共线性问题，将复杂的数据矩阵降维为若干互不相关的潜在因子（latent variable，LV）。LV 的数目是 PLSR 的重要参数，基于交叉验证结果确定该参数，建立光谱化学表征模型。

SVR[20]属于基于支持向量机算法的关联模型。依据结构风险最小化理论，构建最优分类面，以允许学习模型达到全局最优。对于线性不可分的数据集，该算法将数据映射到更高维的特征平面，以求得线性可分。核函数的引入是解决这个问题的关键，径向基核函数可以有效地简化计算的复杂性，提供更加满意的准确度[21]。惩罚系数（penalty coefficient，c）和高斯核函数（gaussian kernel coefficient，g）是 SVR 模型中 2 个重要的参数。前者用于权衡算法的复杂性和偏差的关联，后者为核函数的设置参数。通常情况下，2 个参数的调节需要借助于调参算法。本研究选择遗传算法（genetic algorithm，GA）、粒子群优化算法（particle swarm optimization，PSO）和网格搜索算法（grid search，GS）调整 SVR 模型的参数，以求建立最佳光谱化学表征模型。

对于 2 种模型，主要评价参数为校正系数（correction coefficient，R^2）接近于 1 表明模型效果好；校正集均方根误差（root mean square error of estimation，RMSEE）和预测集均方根误差（root mean square error of prediction，RMSEP）分别用来评价校正集和验证集结果的偏差；交叉验证均方根误差（root mean square error of cross validation，RMSCV）基于交叉验证算法，用来估计回归模型的推广能力[22]。为保证模型的稳健性和防止模型过拟合，应用 Kennard-Stone 算法[23]将样品数据分为训练集和验证集，前者用于构建化学表征模型，后者用于测试模型的推广能力。剩余预测偏差（residual predictive deviation，RPD）是一个评价化

学表征模型效果的常用参数，通常情况下，该值越高则表明模型的效果越好。当其超过 2 时，表明模型的预测效果较好[24]。

3　结果与分析

3.1　标准数据可视化结果

课题组前期研究结果表明，小檗碱、黄连碱、木兰花碱、非洲防己碱和巴马汀是不同基原黄连药材的共有成分，也是其主要的物质基础[3, 8]。依据高效液相色谱的测定，这 5 种化学成分的标准含量结果如图 1 所示。

3.2　红外光谱数据预处理结果

主要包括光谱数据优化、异常值筛选和数据标准化 3 个步骤。FT-NIR 最好的预处理算法分别为二阶导数、二阶导数、不做处理、二阶导数结合多元散射校正和一阶导数；FT-MIR 最好的预处理算法分别为二阶导数、二阶导数、二阶导数、一阶导数结合多元散射校正和一阶导数结合标准正态变量，处理后的红外指纹图谱如图 2 所示。以 PLSR 模型输出结果为评价指标，对于黄连碱、木兰花碱、小檗碱、非洲防己碱和巴马汀的测定，FT-NIR 和 FT-MIR 结果如表 1 和表 2 所示。

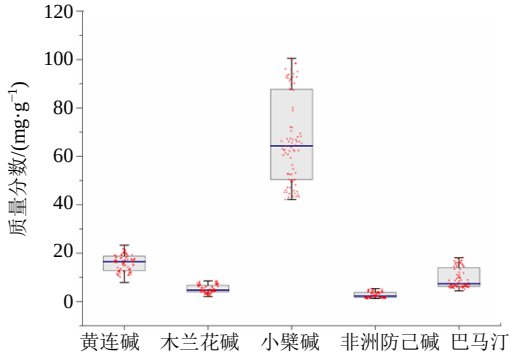


图 1　多基原黄连中 5 种共有成分的含量可视化

Fig. 1　Visualization of contents of five common components in multi-source *Coptidis Rhizoma*

采用 Hotelling T2 检验监测离群点，结果如图 3 所示。以 T2 Crit（99%）为标准，分别从 FT-NIR_黄连碱、FT-MIR_黄连碱、FT-NIR_小檗碱、FT-MIR_小檗碱、FT-NIR_巴马汀、FT-MIR_巴马汀、FT-NIR_非洲防己碱、FT-MIR_非洲防己碱、FT-NIR_木兰花碱和 FT-MIR_木兰花碱的数据矩阵中检测到 0、1、3、0、3、0、2、0、1、0 个离群点，分别有 3 个味连和野连的样品。删除异常值数据，对光谱数据进行归一化处理。

3.3　特征选择及评价

不同基原黄连药材中 5 种物质基础化学成分的数据矩阵均含有大量的信息，包括有效变量、无效变量和噪音变量。特征学习是获取目标特征，去除无效信息的关键方法。本研究应用 4 种特征学习算法（RFE、BORUTA、VIP 和 GINI）对以上光谱数据集进行特征排序。按照固定的间隔，以重复 3 次交叉验证计算的 RMSE 误差为基准来筛选最优变量，以选择出与黄连碱、小檗碱、巴马汀、非洲防己碱、木兰花碱成分关联性强的特征变量，结果如表 3 所示。如表 3 所示，无论是近红外还是中红外

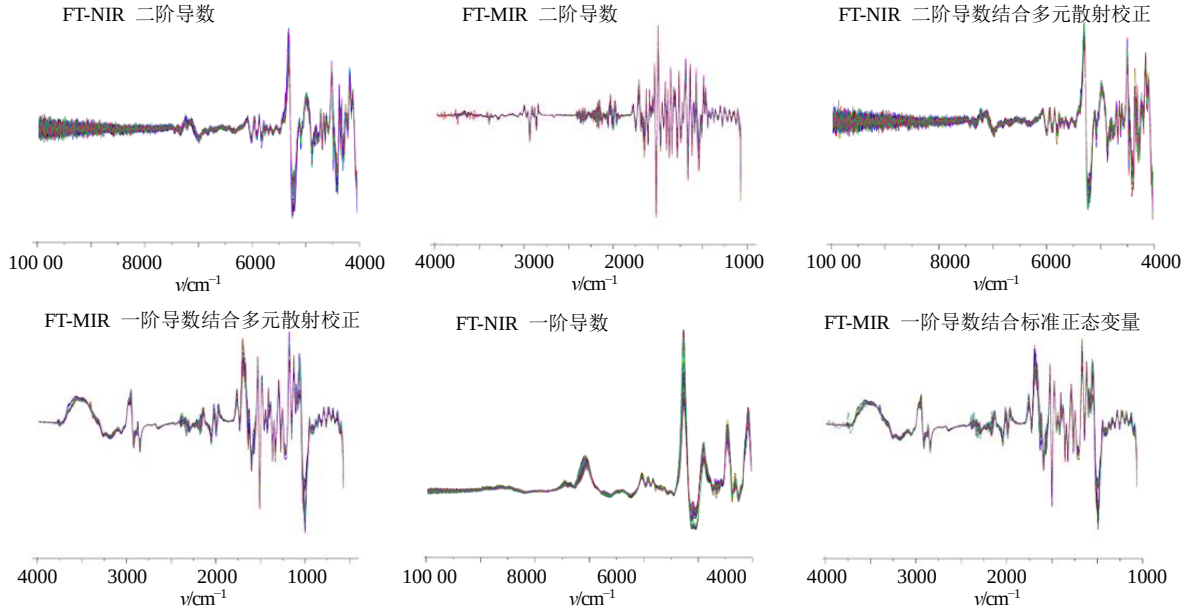


图 2　预处理后的光谱指纹图谱

Fig. 2　Spectral fingerprints after multiple pretreatments

表 1 FT-NIR 光谱优化后结果

Table 1 Results of FT-NIR after optimization

化学成分	预处理	潜在因子	R_c^2	RMSEE	RMSECV	R_p^2	RMSEP
黄连碱	Raw	7	0.85	1.46	1.94	0.61	2.16
	FD	2	0.78	1.65	1.84	0.65	1.93
	SD	4	0.98	0.58	1.46	0.66	1.86
	SNV＋FD	2	0.79	0.37	1.36	0.59	2.14
	SNV＋SD	5	0.99	0.37	1.36	0.65	1.89
	MSC＋FD	2	0.79	1.62	1.80	0.58	2.15
	MSC＋FD	2	0.79	1.65	1.84	0.65	1.89
木兰花碱	Raw	3	0.45	1.22	1.61	0.49	1.33
	FD	8	0.98	0.24	0.93	0.64	1.08
	SD	2	0.81	0.71	1.10	0.79	0.91
	SNV＋FD	1	0.51	1.13	1.21	0.55	1.25
	SNV＋SD	2	0.81	0.71	1.09	0.80	0.92
	MSC＋FD	1	0.51	1.13	1.12	0.55	1.24
	MSC＋SD	2	0.81	0.71	1.09	0.80	0.92
小檗碱	Raw	6	0.88	6.45	7.79	0.92	6.13
	FD	2	0.83	7.45	8.08	0.87	7.54
	SD	2	0.93	4.72	6.84	0.92	6.69
	SNV＋FD	2	0.87	6.62	7.16	0.90	6.96
	SNV＋SD	2	0.94	4.52	6.49	0.93	6.65
	MSC＋FD	2	0.87	6.60	7.15	0.90	6.97
	MSC＋SD	2	0.94	4.51	6.48	0.93	6.67
非洲防己碱	Raw	6	0.56	0.85	1.03	0.24	1.16
	FD	7	0.95	0.29	0.68	0.18	1.23
	SD	4	0.95	0.28	0.76	0.26	1.03
	SNV＋FD	2	0.54	0.84	0.98	0.25	1.05
	SNV＋SD	4	0.94	0.30	0.68	0.26	1.03
	MSC＋FD	2	0.54	0.84	0.98	0.25	1.04
	MSC＋SD	6	0.99	0.10	0.66	0.29	1.00
巴马汀	Raw	9	0.94	1.15	2.02	0.88	1.77
	FD	8	0.99	0.48	1.24	0.88	1.47
	SD	4	0.99	0.50	1.44	0.82	1.72
	SNV＋FD	7	0.98	0.60	1.26	0.86	1.52
	SNV＋SD	4	0.99	0.48	1.32	0.84	1.60
	MSC＋FD	7	0.98	0.60	1.26	0.86	1.52
	MSC＋SD	4	0.99	0.48	1.32	0.84	1.60

Raw、FD、SD、SNV、MSC 分别代表不做处理、一阶导数、二阶导数、多元正态变量、多元散射校正，所有数据均经过平滑处理。R_c^2 和 R_p^2 分别代表校正集和验证集的决定系数

Raw, FD, SD, SNV and MSC respectively represent the spectral pre-processing algorithms of unprocessed, first derivative, second derivative, standard normal variate and multiplicative scatter correction. All data applied with a smoothing step. Rc^2 and Rp^2 represent the coefficient of determination for calibration and the coefficient of determination for prediction respectively

表 2 FT-MIR 光谱优化后结果

Table 2 Results of FT-MIR after optimization

化学成分	预处理	潜在因子	R_c^2	RMSEE	RMSECV	R_p^2	RMSEP
黄连碱	Raw	4	0.78	1.66	1.83	0.66	1.98
	FD	2	0.75	1.78	1.89	0.67	1.94
	SD	2	0.79	1.63	1.87	0.69	1.87
	SNV＋FD	2	0.79	1.62	1.80	0.65	2.01
	SNV＋SD	2	0.80	1.57	1.84	0.69	1.88
	MSC＋FD	2	0.79	1.62	1.80	0.65	2.01
	MSC＋SD	2	0.80	1.56	1.84	0.69	1.88
木兰花碱	Raw	3	0.75	0.80	1.03	0.61	1.16
	FD	2	0.73	0.83	0.93	0.53	1.29
	SD	1	0.71	0.84	1.01	0.79	0.91
	SNV＋FD	1	0.72	0.83	0.91	0.49	1.33
	SNV＋SD	1	0.72	0.83	0.95	0.47	1.35
	MSC＋FD	1	0.72	0.83	0.91	0.49	1.33
	MSC＋SD	1	0.72	0.83	0.95	0.47	1.35

续表 2

化学成分	预处理	潜在因子	R_c^2	RMSEE	RMSECV	R_p^2	RMSEP
小檗碱	Raw	4	0.92	5.22	6.10	0.85	7.69
	FD	2	0.90	5.84	6.30	0.88	6.95
	SD	2	0.91	5.61	6.58	0.89	6.70
	SNV＋FD	4	0.96	3.67	5.09	0.86	7.23
	SNV＋SD	3	0.95	4.42	5.68	0.88	6.73
	MSC＋FD	4	0.96	3.66	5.10	0.86	7.28
	MSC＋SD	3	0.95	4.26	5.69	0.88	6.72
非洲防己碱	Raw	4	0.51	0.82	0.94	0.39	1.06
	FD	7	0.54	0.78	0.81	0.49	0.97
	SD	3	0.75	0.57	0.76	0.50	0.94
	SNV＋FD	1	0.50	0.81	0.84	0.49	0.96
	SNV＋SD	3	0.77	0.57	0.74	0.47	0.97
	MSC＋FD	3	0.69	0.65	0.73	0.54	0.91
	MSC＋SD	3	0.77	0.57	0.74	0.47	0.97
巴马汀	Raw	3	0.82	1.77	1.95	0.86	1.87
	FD	5	0.93	1.14	1.52	0.91	1.57
	SD	2	0.86	1.51	1.71	0.91	1.54
	SNV＋FD	3	0.90	1.31	1.60	0.93	1.37
	SNV＋SD	2	0.87	1.49	1.71	0.92	1.41
	MSC＋FD	3	0.90	1.32	1.61	0.86	1.37
	MSC＋SD	2	0.87	1.48	1.70	0.92	1.41

Raw、FD、SD、SNV、MSC 分别代表不做处理、一阶导数、二阶导数、多元正态变量、多元散射校正，所有数据均经过平滑处理。Rc^2 和 Rp^2 分别代表校正集和验证集的决定系数

Raw, FD, SD, SNV and MSC respectively represent the spectral pre-processing algorithms of unprocessed, first derivative, second derivative, standard normal variate and multiplicative scatter correction. All data applied with a smoothing step. Rc^2 and Rp^2 represent the coefficient of determination for calibration and the coefficient of determination for prediction respectively

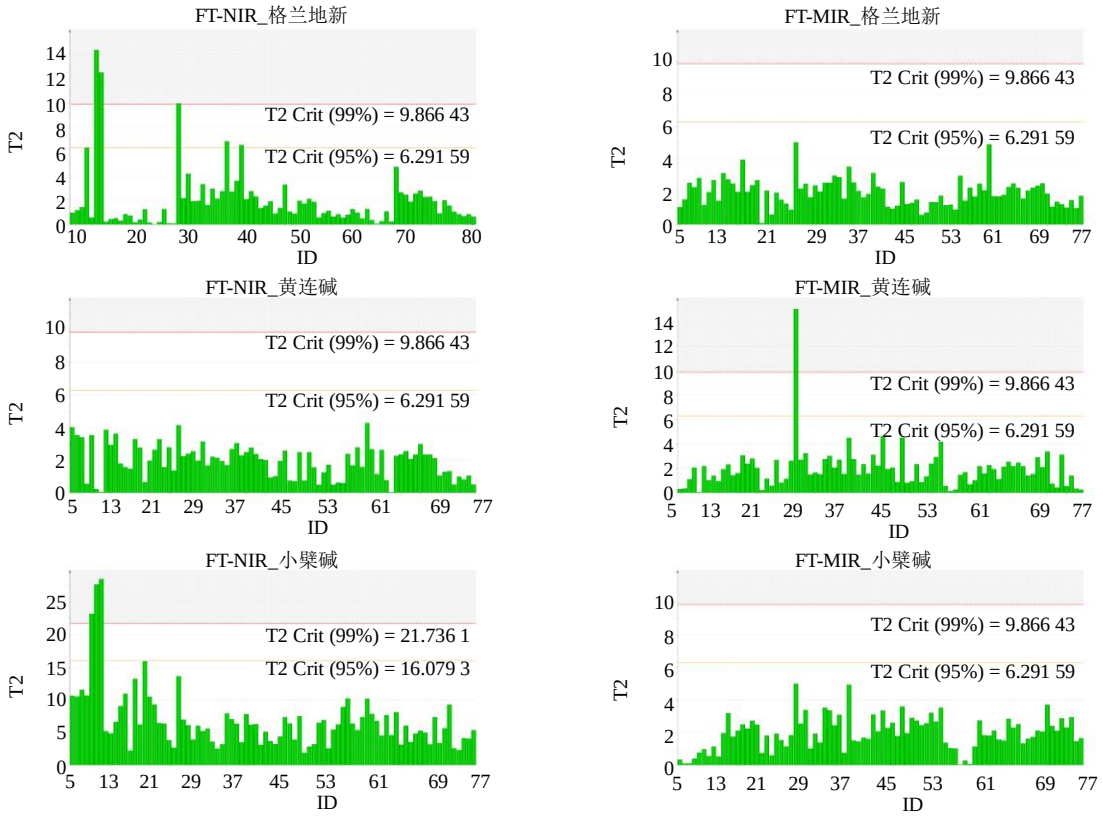


图 3　不同数据集异常值诊断结果

Fig. 3　Outlier diagnosis of different datasets

表 3 不同变量采集模型误差率

Table 3 *RMSE* based on different feature selection models

成分	红外光谱	RFE	BORUTA	VIP	GINI
黄连碱	FT-NIR	0.099±0.031	0.104±0.033	0.108±0.037	0.102±0.029
	FT-MIR	0.088±0.032	0.092±0.030	0.112±0.036	0.094±0.031
小檗碱	FT-NIR	0.123±0.050	0.130±0.048	0.133±0.050	0.123±0.049
	FT-MIR	0.074±0.027	0.077±0.026	0.083±0.026	0.077±0.022
巴马汀	FT-NIR	0.105±0.039	0.106±0.044	0.115±0.048	0.135±0.041
	FT-MIR	0.092±0.032	0.091±0.040	0.102±0.041	0.116±0.038
非洲防己碱	FT-NIR	0.131±0.036	0.126±0.037	0.194±0.038	0.185±0.036
	FT-MIR	0.119±0.033	0.122±0.034	0.203±0.034	0.190±0.042
木兰花碱	FT-NIR	0.114±0.034	0.120±0.039	0.132±0.038	0.181±0.031
	FT-MIR	0.119±0.039	0.118±0.041	0.123±0.046	0.161±0.045

光谱的数据矩阵，RFE 和 BORUTA 算法都表现出较好的特征采集能力。其中 RFE 的模型 *RMSE* 值为 0.074～0.131，针对 FT-MIR_黄连碱、FT-MIR_小檗碱、FT-NIR_巴马汀、FT-MIR_非洲防己碱和 FT-NIR_木兰花碱数据集，误差率较低；BORUTA 的模型 RMSE 值为 0.077～0.130，针对 FT-MIR_黄连碱、FT-MIR_小檗碱、FT-MIR_巴马汀、FT-MIR_非洲防己碱和 FT-MIR_木兰花碱数据集，误差率较低。对于 VIP 和 GINI 2 种特征选择算法，其误差率较高，评价效果较差。

特征变量的数目也在一定程度上反应出特征学习模型的效率，其结果见图 4。针对 FT-NIR 和 FT-MIR 数据集，RFE 和 BORUTA 在变量数目的输出中具有明显的优势，其中 BORUTA 算法的效果较好；VIP 和 GINI 2 种特征选择算法的效率较低。由图可见，BORUTA 算法基本可以将不同的数据集的变量数目缩减至 50 以内，同时保持较低的误差率。

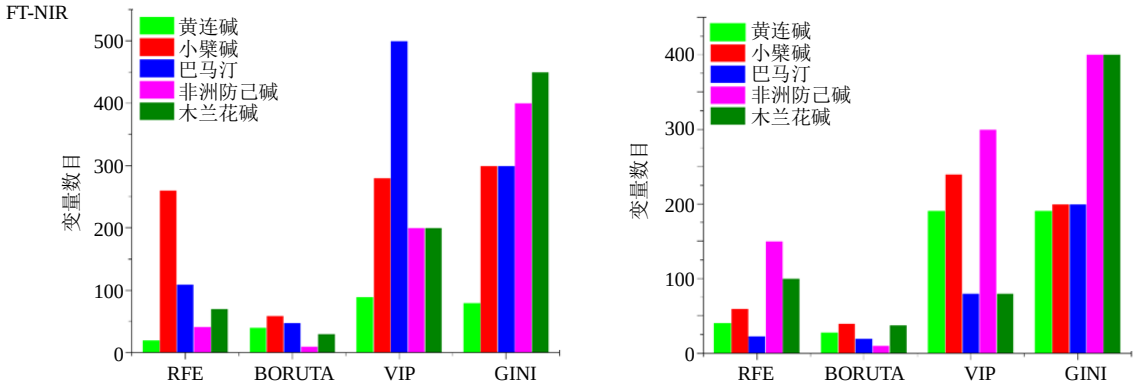


图 4 不同特征筛选后变量数目

Fig. 4 Variable number after different feature selection

结合误差率和变量数目 2 个指标综合考虑，对于 FT-NIR_黄连碱、FT-MIR_黄连碱和 FT-NIR_木兰花碱 3 个数据集，选择 RFE 特征学习算法进行变量采集，简化光谱矩阵数据。其余数据集选择 BORUTA 特征学习模型对其进行优化。经过数据预处理和变量筛选两个阶段，FT-NIR_黄连碱、FT-MIR_黄连碱、FT-NIR_小檗碱、FT-MIR_小檗碱、FT-NIR_巴马汀、FT-MIR_巴马汀、FT-NIR_非洲防己碱、FT-MIR_非洲防己碱、FT-NIR_木兰花碱和 FT-MIR_木兰花碱数据集的变量数目分别被缩减为 83×20、83×40、81×59、81×40、81×48、81×23、82×10、82×10、83×70 和 83×38。

3.4 光谱化学表征模型的建立

基于以上步骤，本研究进一步基于特征级信息融合算法，获得不同基原黄连药材中黄连碱、小檗碱、巴马汀、非洲防己碱和木兰花碱的相关特征数据集，其大小分别为 83×60、81×99、81×71、82×20 和 83×108。相较于原始数据，光谱变量的数量明显降低，可以有效增加模型的速度和准确性，同时增强其推广能力。

应用以上特征数据集，以高效液相色谱结果作为标准数据，分别建立 2 种化学计量学相关模型，以考察其预测结果和真实数据的相关性，证实模型的推广能力。PLSR 分析结果见表 4，相对于标准数

据，小檗碱的 PLSR 模型的预测效果最优。该模型将数据集缩减至 4 个 LV，其 RMSEE、RMSECV、R^2、RMSEP 和 RPD 分别是 0.075、0.097、0.928、0.096 和 3.734。对于非洲防己碱的含量预测，PLSR 模型的效果较差，其 RMSEE、RSECV、R^2、RMSEP 和 RPD 分别是 0.140、0.170、0.580、0.205 和 1.570。根据光谱化学表征模型的预测 RPD 值，表明 PLSR 模型对于黄连碱、小檗碱、巴马汀和木兰花碱的预测都取得较好效果。

SVR 模型的结果如表 5 所示，调参流程如图 5

表 4　PLSR 光谱化学表征模型的结果

Table 4　PLSR results of spectrochemical characterization model

化学成分	LV	RMSEE	RMSECV	R^2	RMSEP	RPD
黄连碱	3	0.092	0.111	0.862	0.162	2.003
小檗碱	4	0.075	0.097	0.928	0.097	3.734
巴马汀	5	0.080	0.124	0.883	0.115	2.917
非洲防己碱	3	0.140	0.170	0.580	0.205	1.570
木兰花碱	4	0.079	0.108	0.767	0.132	2.051

表 5　光谱化学表征模型的 SVR 结果

Table 5　SVR results of spectrochemical characterization model

化学成分	方法	c	g	RMSECV	R^2	RMSEP	RPD
黄连碱	GA	2.266	0.473	0.110	0.862	0.186	1.741
	PSO	1.372	0.010	0.116	0.831	0.167	1.941
	GS	1.320	0.435	0.106	0.864	0.185	1.753
小檗碱	GA	0.337	0.091	0.087	0.950	0.089	4.034
	PSO	17.238	0.010	0.094	0.957	0.075	4.842
	GS	0.758	0.027	0.088	0.951	0.079	4.558
巴马汀	GA	0.476	0.015	0.102	0.923	0.128	2.638
	PSO	14.149	0.010	0.116	0.900	0.161	2.094
	GS	2.297	0.005	0.099	0.915	0.126	2.664
非洲防己碱	GA	64.343	0.058	0.167	0.726	0.170	1.892
	PSO	3.039	0.010	0.184	0.606	0.202	1.587
	GS	1.320	0.082	0.174	0.712	0.182	1.768
木兰花碱	GA	9.971	0.032	0.114	0.763	0.134	2.030
	PSO	2.930	0.010	0.127	0.792	0.127	2.138
	GS	2.297	0.005	0.123	0.823	0.116	2.344

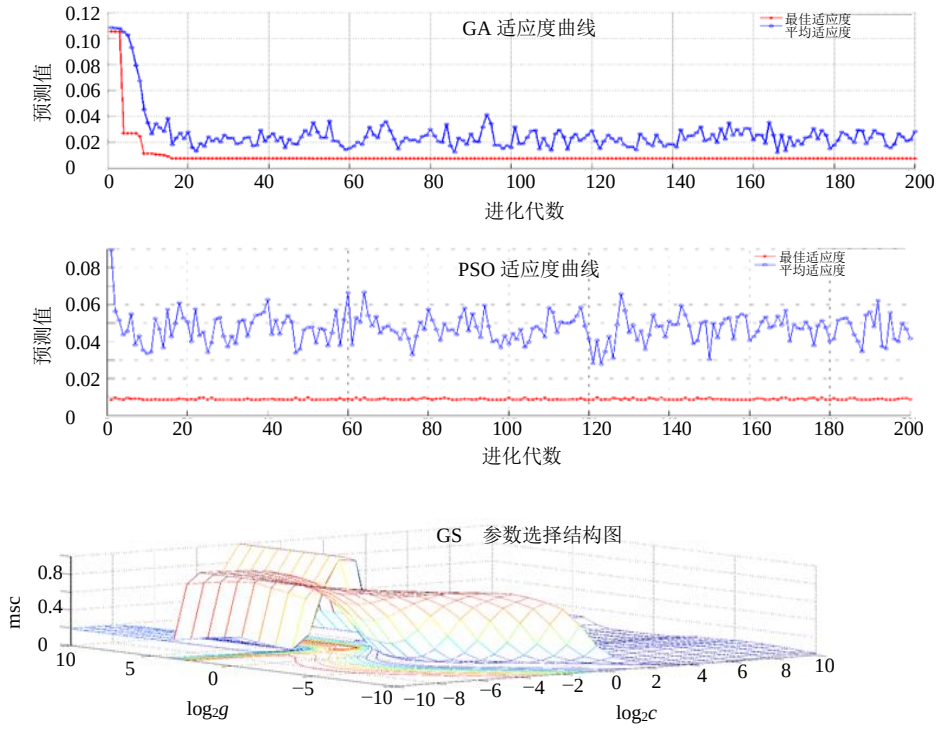
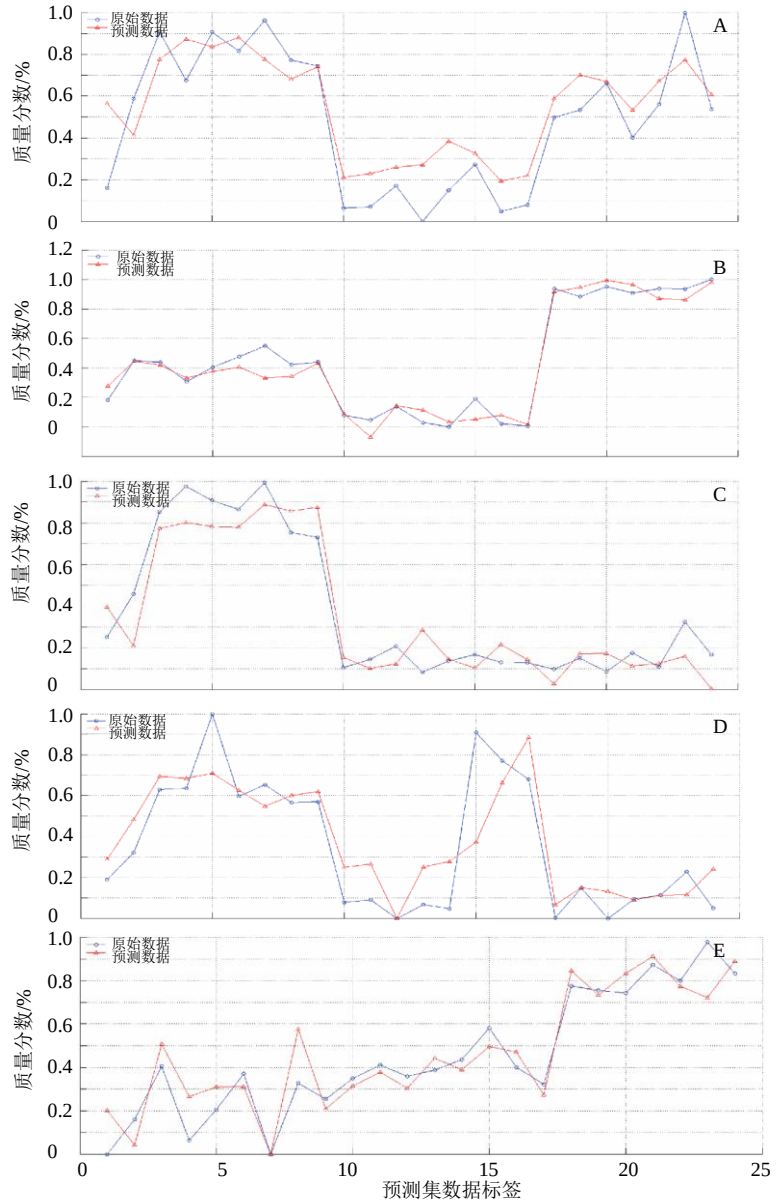


图 5　GA、PSO 和 GS 调参流程（以小檗碱数据集为例）

Fig. 5　Parameter adjustment process of GA, PSO and GS (berberine as an example)

所示。相对于 PLSR，SVR 模型可以更好地处理非线性的数据，对于某些成分取得了更好的预测效果。特别是对于小檗碱含量的预测，选择 PSO 调参算法，模型的预测效果取得了较大的提升，其 RSECV、R^2、RMSEP 和 RPD 分别为 0.094、0.957、0.075 和 4.842；其主要参数 c 和 g 分别是 17.238 和 0.010。另外针对于非洲防己碱和木兰花碱 2 个成分，SVR 模型的效果也优于 PLSR；GA 和 GS 两种调参的方法获得了更优的参数值，其 c 和 g 分别是 64.343、0.058 和 2.297、0.005。但是针对于黄连碱和巴马汀 2 个成分，SVR 模型的预测效果弱于 PLSR 相关模型。根据所建立的最优红外光谱化学表征模型，对未知样品中 5 种物质基础成分的含量进行预测。样品预测值和真实值之间的关系见图 6，两者之间趋势接近，进一步证明了模型的有效性和推广能力。应用主成分分析的方法对预测结果进行分类，结果见图 7。如图所示，味连、雅连、野连和云连样品的界限明显，可以很明显地被鉴



A-黄连碱 B-小檗碱 C-巴马汀 D-非洲防己碱 E-木兰花碱

A-coptisine B-berberine C-palmatine D-columbamine E-magnoflorine

图 6 基于最优光谱化学表征模型的真实值和预测值相关性

Fig. 6 Correlation of predicted data and actual data based on optimal spectrochemical characterization model

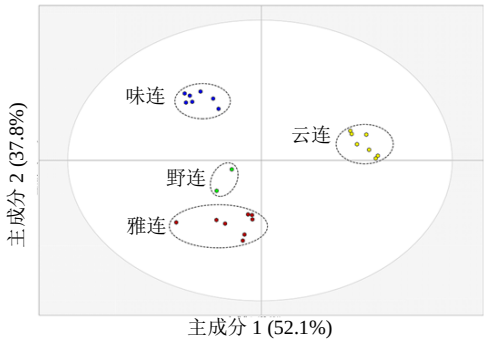


图 7　基于最优光谱化学表征模型的 PCA 结果

Fig. 7　PCA results based on optimal spectrochemical characterization model

别开。经过光谱优化和特征学习流程，光谱化学表征模型能成功预测不同基原黄连样品中主要共有物质基础的含量，可以有效对不同基原的黄连药材进行品质辨识。

4　讨论

多基原中药自古以来就是中药体系构成的重要部分，一方面保障了中医临床的有药可用，另一方面又给临床实现精准治疗提供依据。但是对于多基原品种，其基原的等效性一直是历代医家的主要研究内容，而其效用差异研究较少。因此，如何依据多基原中药的效用差异制定合理的品质评价标准，是该类药材实现临床合理用药、有效资源配置环节中亟需解决的问题。本研究以典型的多基原药材黄连为研究对象，结合课题组前期研究中得到的“共有成分的含量差异是其效用差异的主要原因”这一结论，选择黄连碱、小檗碱、巴马汀、非洲防己碱和木兰花碱为指标，应用无损绿色的光谱技术构建其含量表征模型，考察该技术应用于多基原黄连药材品质辨识的可行性。

中药的化学成分复杂，光谱变量众多且难以辨识。本研究在前人工作的基础上[25-27]，建立了一套完整的光谱矩阵分析流程：主要包括光谱信号优化、异常值诊断、数据标准化、特征学习与评价、光谱化学表征模型等步骤。结果显示，光谱预处理可以明显降低噪音信号，提高光谱数据质量；特征学习算法可有效从复杂数据矩阵中提取出和目标成分有关的光谱特征。其中 RFE 和 BORUTA 模型可以将 3000 多个黄连样品光谱变量降维到 100 个内，同时能保证较高的正确率。这两种方法在中药领域应用较少，将来可以有效地应用到中药复杂问题的解决之中。

基于主要的 FT-NIR 和 FT-MIR 变量特征，分别建立黄连中五种共有物质基础化学成分的 PLSR 和 SVR 光谱化学表征模型。其中小檗碱的预测效果最好，其 RPD 值高达 4.842；黄连碱、巴马汀和木兰花碱的数学模型的 RPD 值均高于 2，取得了满意的效果。非洲防己碱的光谱化学表征模型的 RPD 值为 1.892，其预测效果有待于进一步提高。对比 PLSR 和 SVR 光谱化学表征模型，SVR 的效果更优，可能与处理非线性问题的能力有关。

将未知样品代入最优模型之中，结果显示五种化学成分含量的真实值和预测值相关性较高，证明模型的可靠性和推广性。应用 PCA 分析最优模型的预测结果，散点图可以将不同基原黄连药材有效鉴别，表明所建立的光谱化学表征模型能够对该类药材的品质进行辨识，有进一步应用于该药材效用评价的潜力，为多基原中药的品质辨识提供一个无损、绿色和快速的方案。

利益冲突　所有作者均声明不存在利益冲突

参考文献

[1] Li J X, Yan D, Ma L N, *et al*. A quality evaluation strategy for Rhizoma coptidis from a variety of different sources using chromatographic fingerprinting combined with biological fingerprinting [J]. *Chin Sci Bull*, 2013, 58(33): 4092-4100.

[2] 刘睿颖, 任瑶瑶, 张思远, 等. 味连与雅连改善 2 型糖尿病大鼠糖及脂代谢紊乱的研究 [J]. 华西药学杂志, 2018, 33(4): 368-372.

[3] Chen Y, Qi L M, Zhong F R, *et al*. Integrated metabolomics and ligand fishing approaches to screen the hypoglycemic ingredients from four *Coptis* medicines [J]. *J Pharm Biomed Anal*, 2021, 192: 113655.

[4] Zhang H M, Guo Y N, Meng L W, *et al*. Rapid screening and characterization of acetylcholinesterase inhibitors from Yinhuang oral liquid using ultrafiltration-liquid chromatography-electrospray ionization tandem mass spectrometry [J]. *Pharmacogn Mag*, 2018, 14(54): 248-252.

[5] Zhang G W, Peng S L, Cao S Y, *et al*. A fast progressive spectrum denoising combined with partial least squares algorithm and its application in online Fourier transform infrared quantitative analysis [J]. *Anal Chim Acta*, 2019, 1074: 62-68.

[6] Shamshirband S, Petković D, Javidnia H, *et al*. Sensor data fusion by support vector regression methodology—A comparative study [J]. *IEEE Sens J*, 2015, 15(2):

850-854.

[7] Sun W J, Zhang X, Zhang Z Y, *et al*. Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb [J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2017, 171: 72-79.

[8] Zhong F R, Shen C, Qi L M, *et al*. A multi-level strategy based on metabolic and molecular genetic approaches for the characterization of different *Coptis* medicines using HPLC-UV and RAD-seq techniques [J]. *Molecules*, 2018, 23(12): E3090.

[9] Yang Y H, Pan T, Zhang J. Global optimization of Norris derivative filtering with application for near-infrared analysis of serum urea nitrogen [J]. *Am J Anal Chem*, 2019, 10(5): 143-152.

[10] Dhanoa M S, Lister S J, Sanderson R, *et al*. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra [J]. *J Infrared Spectrosc*, 1994, 2(1): 43-47.

[11] Roy I G. On computing first and second order derivative spectra [J]. *J Comput Phys*, 2015, 295: 307-321.

[12] Mahmoud S, Lotfi A, Langensiepen C. User activities outliers detection; integration of statistical and computational intelligence techniques [J]. *Comput Intell*, 2016, 32(1): 49-71.

[13] Shalabi L A, Shaaban Z. Normalization as a preprocessing engine for data mining and the approach of preference matrix. International Conference on Dependability of Computer Systems [C]. 2006: 207-214.

[14] Granitto P M, Furlanello C, Biasioli F, *et al*. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products [J]. *Chemom Intell Lab Syst*, 2006, 83(2): 83-90.

[15] Kursa M B, Rudnicki W R. Feature selection with the Boruta Package [J]. *J Stat Soft*, 2010, 36(11): 1-13.

[16] Mehmood T, Liland K H, Snipen L, *et al*. A review of variable selection methods in Partial Least Squares Regression [J]. *Chemom Intell Lab Syst*, 2012, 118: 62-69.

[17] Singh S R, Murthy H A, Gonsalves T A. Feature selection for text classification based on gini coefficient of inequality [J] *Proceed Fourth Intern* [J]. 2010, 26: 76-85.

[18] Biancolillo A, Bucci R, Magrì A L, *et al*. Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication [J]. *Anal Chim Acta*, 2014, 820: 23-31.

[19] Cheng J H, Sun D W. Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle [J]. *Food Eng Rev*, 2017, 9(1): 36-49.

[20] Drucker H, Surges C J C, Kaufman L, *et al*. Support vector regression machines [J]. *Adv Neural Inf Process Syst*, 1997: 155-161.

[21] Chen Q S, Zhao J W, Fang C H, *et al*. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM) [J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2007, 66(3): 568-574.

[22] Qi L M, Zhang J, Zuo Z T, *et al*. Determination of iridoids in *Gentiana rigescens* by infrared spectroscopy and multivariate analysis [J]. *Anal Lett*, 2017, 50(2): 389-401.

[23] Kennard R W, Stone L A. Computer aided design of experiments [J]. *Technometrics*, 2012, 11(1): 137-148.

[24] Chang C W, Laird D A, Mausbach M J, *et al*. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties [J]. *Soil Sci Soc Am J*, 2001, 65(2): 480-490.

[25] Pei Y F, Zuo Z T, Zhang Q Z, *et al*. Data fusion of Fourier transform mid-infrared (MIR) and near-infrared (NIR) spectroscopies to identify geographical origin of wild *Paris polyphylla* var. *yunnanensis* [J]. *Molecules*, 2019, 24(14): E2559.

[26] Wang Y, Zuo Z T, Shen T, *et al*. Authentication of *Dendrobium* species using near-infrared and ultraviolet-visible spectroscopy with chemometrics and data fusion [J]. *Anal Lett*, 2018, 51(17): 2792-2821.

[27] Li Y, Zhang J, Li T, *et al*. Geographical traceability of wild Boletus edulis based on data fusion of FT-MIR and ICP-AES coupled with data mining methods (SVM) [J]. *Spectrochim Acta A Mol Biomol Spectrosc*, 2017, 177: 20-27.

[责任编辑 时圣明]