

近红外光谱结合不同变量筛选方法用于黄芩提取过程中 黄芩苷含量预测

刘雪松¹, 张丝雨¹, 赵曼茜², 王 钧³, 李页瑞³, 代 军², 滕传震²,
柯 潇², 陈 勇¹, 吴永江^{1*}

(1. 浙江大学药学院, 浙江 杭州 310058; 2. 成都康弘药业集团股份有限公司, 四川 成都 610036;
3. 苏州泽达兴邦医药科技有限公司, 江苏 苏州 215000)

摘要: 近红外光谱技术 (NIRS) 结合化学计量学可以实现中药过程分析中的快速定量, 变量筛选算法的应用可以有效去除光谱中的冗余信息并筛选出与成分信息相关的关键变量, 与全光谱模型相比可以显著降低模型复杂度, 并提高预测精度。本文将近红外光谱技术结合不同变量筛选算法用于黄芩提取过程黄芩苷含量的快速测定, 基于 SPXY 法划分数据集, 采用竞争自适应加权采样法 (CARS)、随机青蛙算法 (RF)、连续投影算法 (SPA) 3 种不同变量筛选方法, 以偏最小二乘法 (PLS) 为基础, 建立并比较了黄芩药材提取过程黄芩苷含量的定量校正模型。经 CARS 法、RF 法和 SPA 法分别筛选出 92、10、17 个变量, CARS-PLS 法建立的黄芩苷模型具有最佳性能, CARS 法筛选的关键变量与指标成分黄芩苷的化学结构也有着较好的对应关系。模型的校正均方根误差 (RMSEC) 和预测均方根误差 (RMSEP) 分别为 0.528 2 和 0.720 2, 与全光谱模型相比, 模型的校正集相关系数 (R_c) 从 0.917 0 上升到 0.979 9, 相对预测偏差 (RSEP) 从 10.58% 降低到 5.59%。

关键词: 近红外光谱; 变量筛选; 黄芩苷; 竞争自适应加权采样; 随机青蛙算法; 连续投影算法; 偏最小二乘法
中图分类号: R917 文献标识码: A 文章编号: 0513-4870(2019)01-0138-06

The prediction of baicalin content in the extraction process of *Scutellaria baicalensis* by near-infrared spectroscopy combined with different variable selection methods

LIU Xue-song¹, ZHANG Si-yu¹, ZHAO Man-qian², WANG Jun³, LI Ye-rui³, DAI Jun²,
TENG Chuan-zhen², KE Xiao², CHEN Yong¹, WU Yong-jiang^{1*}

(1. College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China; 2. Chengdu Kanghong
Pharmaceutical Group Co., Ltd., Chengdu 610036, China; 3. Suzhou ZeDaXingBang Pharmaceutical Co., Ltd.,
Suzhou 215000, China)

Abstract: Near-infrared spectroscopy (NIRS) combined with chemometrics can achieve rapid detection in process analysis. After variable selection, the redundant information is effectively removed and the characteristic variables related to the response values are selected. Compared with global model, the complexity is significantly reduced and the prediction accuracy is also improved. In this study, near-infrared spectroscopy analysis combined with different variable selection methods was applied to achieve the rapid detection of baicalin in the extraction process of *Scutellaria baicalensis*. Data sets were divided based on sample set portioning based on joint x - y

收稿日期: 2018-08-06; 修回日期: 2018-10-31.

基金项目: 重大新药创制科技重大专项 (2018ZX09201010).

*通讯作者 Tel: 86-571-88208455, E-mail: yjwu@zju.edu.cn

DOI: 10.16438/j.0513-4870.2018-0712

distance (SPXY) method. Competitive adaptive weighted resampling method (CARS), random frog (RF) and successive projections algorithm (SPA) were applied to variable selection. Partial least squares (PLS) models were constructed based on above three methods, and the prediction results were compared. After CARS, RF and SPA method, 92, 10 and 17 variables were screened out respectively. According to the performance of the models, CARS method is found to be more effective and suitable than RF and SPA. Furthermore, the characteristic variables selected by CARS method have a better correspondence with the chemical structure of baicalin. The root mean square error (RMSEC) of the calibration set and the root mean square error (RMSEP) of the prediction set are 0.528 2 and 0.720 2 respectively. Compared with the global PLS model, the correlation coefficient of the calibration set (R_c) is increased to 0.979 9 from 0.917 0, and the relative standard errors of prediction (RSEP) is reduced to 5.59% from 10.58%.

Key words: near infrared spectroscopy; variable selection; baicalin; competitive adaptive reweighted sampling method; random frog; successive projections algorithm; partial least squares

近红外光谱 (near infrared spectroscopy, NIRS) 自 20 世纪末以来发展迅猛, 因为其具有快速测量、无消耗和非破坏性的特点^[1], 已经在中药定性和定量分析中展现了巨大优势。目前, 近红外光谱技术作为过程分析的重要手段被广泛应用于中药原药材鉴别^[2,3]、提取^[4]、浓缩^[5]、醇沉^[6]等各个生产环节。但是, 近红外光谱中所反映的信息繁杂, 除了关键信息外, 光谱中通常还会包含大量的冗余信息和不相关变量, 传统的光谱预处理方法无法完全消除其影响。因此, 分析时通常采用一些波段选择方法, 筛选与待测化合物结构关联密切并受干扰程度较小的典型波数区域^[7]。

波长选择可剔除无信息波长和干扰波长, 有效避免模型过拟合, 使模型训练器运行更加快速、高效, 以建立预测能力强、稳健性好的定量校正模型^[8]。研究证明, 无信息变量的消除和关键变量的优选可显著改善模型的预测精度和稳定性^[9,10]。因此, 针对不同指标成分, 选择合适的方法筛选近红外光谱变量成为模型建立中的重要步骤。

本文以黄芩提取过程为研究对象, 采用 NIRS 技术分析和监控黄芩苷在提取过程中的含量变化, 定量校正模型的建立基于偏最小二乘算法 (partial least squares, PLS)。采用竞争自适应加权重采样法^[11] (competitive adaptive reweighted sampling, CARS)、随机青蛙算法^[12] (random frog, RF)、连续投影算法^[13] (successive projections algorithm, SPA) 3 种变量筛选方法对模型进行优化, 全面比较了基于 3 种变量筛选的模型性能, 优选出黄芩苷的最佳定量模型, 并对筛选出的变量进行了解释和化学结构对应, 为实现黄芩提取过程中黄芩苷含量的快速定量检测以及黄芩苷关键变量的筛选提供理论依据和支持。

材料与方法

材料与仪器 Antaris II 傅立叶变换近红外光谱

仪 (Thermo Nicolet, USA), 仪器配有液体透射检测器、TQ Analyst 8.0 数据处理软件, 石英比色皿内径为 2 mm; Agilent 1200 高效液相色谱仪 (Agilent, 美国), 配有二元梯度泵、自动进样器、柱温箱、UV 检测器和 Chemstation 工作站; XS105 分析天平 (Mettler Toledo, 瑞士); JCS-600 电子天平 (永康凯丰集团有限公司)。黄芩饮片购自杭州孙泰和医药馆, 由浙江大学徐娟华副教授鉴定为唇形科植物黄芩 (*Scutellaria baicalensis* Georgi) 的干燥根。甲醇、磷酸 (分析纯, 阿拉丁公司); 纯化水 (娃哈哈集团); 黄芩苷对照品 (MUST-18010410) 购自成都曼斯特生物科技有限公司, 纯度大于 98.0%。

提取过程样本采集 取每批次黄芩 60 g, 加 10 倍量水提取。从不同批次的黄芩提取过程共采集到 86 个样品。提取过程中, 在每个取样点取提取物样品 2 mL 用于近红外光谱和高效液相色谱定量分析。

近红外光谱采集 对收集到的样品进行近红外透射光谱扫描, 以空气作为参比, 扫描次数为 32 次, 光谱点范围为 4 000~10 000 cm^{-1} , 分辨率为 8 cm^{-1} , 承载样品的样品杯为石英比色皿, 光程为 2 mm。每个提取点的提取液重复扫描 3 次, 经 TQ Analyst 8.0 数据处理软件得到平均光谱。

含量测定方法及色谱条件 本实验中黄芩苷的含量测定参考 2015 版《中国药典》一部黄芩药材中黄芩苷的定量检测方法^[14]。以高效液相色谱法测得样本中的含量作为近红外光谱定量预测中的标准对照方法的参比值。色谱柱: XBridge Shield RP18 C18 色谱柱 (250 mm×4.6 mm, 5 μm); 流动相: 甲醇-0.2% 磷酸溶液 (47:53); 流速: 1 $\text{mL}\cdot\text{min}^{-1}$; 检测波长: 280 nm; 柱温: 30 $^{\circ}\text{C}$; 进样量: 10 μL 。

模型性能评价指标及数据处理 以相关系数 (correlation coefficient, R)、校正集均方根误差 (root mean square error of calibration, RMSEC) 及预测集均

方根误差 (root mean square error of prediction, RMSEP) 作为优化建模参数的指标, 并检查模型预测性能。模型相关系数越接近 1, 模型分析的准确度越高。当 RMSEC 和 RMSEP 较小且彼此更接近时, 该模型更加稳健和具有预测性。同时, 以相对预测偏差 (relative standard errors of prediction, RSEP) 来考察模型的预测误差, RSEP 越小模型精度越高^[15]。

光谱预处理、CARS、RF、SPA 变量筛选和模型的建立均采用 Matlab R2017b 软件, 图形绘制采用 Origin 8.0 软件。

结果

1 校正集与验证集的划分

目前常用的划分方法包括随机抽样法、常规选择法、Kennard-Stone 法和 SPXY (sample set portioning based on joint x - y distance) 法, 在以上方法中, SPXY 法同时考虑光谱数据变量和化学测量值变量两方面, 并基于样本间距离进行校正样本的选择, 可以提高 NIR 模型的预测能力^[16]。SPXY 法进行样本间距离计算的公式如下:

$$d_x(p, q) = \sqrt{(x_p - x_q)^2} \quad p, q \in [1, n]$$

$$d_y(p, q) = \sqrt{(y_p - y_q)^2} \quad p, q \in [1, n]$$

$$d_{xy}(p, q) = \frac{d_x(p, q)}{\max_{p, q \in [1, n]} d_x(p, q)} + \frac{d_y(p, q)}{\max_{p, q \in [1, n]} d_y(p, q)} \quad p, q \in [1, n]$$

其中 $d_x(p, q)$ 是基于近红外光谱数据 x 来计算 p, q 两样本间的距离, $d_y(p, q)$ 是基于化学测量值计算 p, q 两样本间的距离, $d_{xy}(p, q)$ 是在 x 和 y 空间 p 和 q 两样本的标准化距离。

本研究将采集到的黄芩提取液样品采用 SPXY 算法, 使得选择的校正样本尽量覆盖全部光谱向量空间和化学测量值空间, 以提高模型稳健性。通过软件计算样本间距离, 按照样本间距离进行排序, 根据排序结果, 覆盖距离最大的前 60 个样本作为校正集, 其余 26 个样本作为验证集。

2 光谱预处理方法优筛

近红外光谱的采集过程中易受样品状态以及外界环境等影响, 导致光谱中含有冗杂的噪声信息, 影响建模效果^[17,18]。基于以上问题, 对光谱叠加一些预处理方法通常可在建模前帮助消除噪声及其他影响。本研究比较了原始光谱和经过多种不同预处理方法后的模型效果, 考察了一阶导数法 (first derivative, 1stD)、二

阶导数法 (second derivative, 2ndD)、多元散射校正 (multiplicative scatter correction, MSC)、标准正交变换 (standard normal variate correction, SNV) 和 Savitsky-Golay 滤波平滑 (S-G) 以及叠加预处理方法对模型效果的影响, 结果如表 1 所示。通过比较结果可知, 1stD+SG 预处理后模型的 RMSEC 和 RMSEP 值分别为 1.057 和 1.362, 校正集的相关系数 R 值为 0.917 0, RSEP 值为 10.58%, 1stD 法可以消除峰与峰之间的重叠, 平滑可以降低由于导数法引入的随机噪声^[19], 因此本研究选择 1stD+S-G 平滑, 平滑点数 15, 作为最佳的预处理方法。收集的样品原始光谱显示在图 1A 中, 预处理的光谱显示在图 1B 中。

Table 1 Influence of different pretreatment methods on performance of calibration models. 1stD: First derivative; 2ndD: Second derivative; SNV: Standard normal variate correction; S-G: Savitsky-Golay; MSC: Multiplicative scatter correction; RMSEC: Root mean square error of calibration; RMSEP: Root mean square error of prediction; R : Correlation coefficient; RSEP: Relative standard errors of prediction

Pretreatment method	Factor	Calibration set		Validation set		
		R_c	RMSEC	R_p	RMSEP	RSEP/%
Raw	8	0.790 3	1.623	0.798 8	1.486	11.52
1 st D	3	0.756 5	1.733	0.780 9	1.543	11.97
1 st D+SNV	7	0.953 9	0.795 4	0.755 3	1.618	12.55
1 st D+S-G	8	0.917 0	1.057	0.833 5	1.362	10.58
1 st D+MSC	7	0.954 0	0.794 6	0.752 6	1.626	12.61
2 nd D	6	0.985 8	0.444 7	0.648 2	1.880	14.59
2 nd D+SNV	6	0.985 5	0.449 7	0.642 9	1.891	14.67
2 nd D+S-G	7	0.948 8	0.837 0	0.718 9	1.716	13.31
2 nd D+MSC	6	0.985 5	0.449 0	0.642 9	1.891	14.67

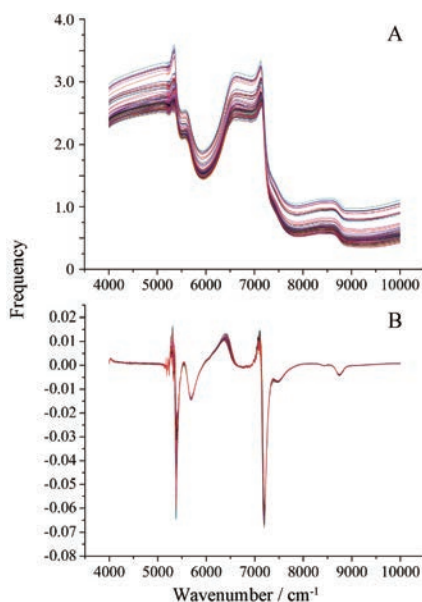


Figure 1 Raw NIR spectra (A) and spectra pretreated with first derivative and S-G (B) of extraction samples

3 不同变量筛选方法及结果

3.1 CARS法 竞争自适应加权重采样法 (competitive adaptive weighted resampling method, CARS) 首先采用蒙特卡洛采样法建立 PLS 模型, 然后基于指数衰减函数消除变量, 强制去除波长点。进一步筛选变量是基于自适应加权重采样技术, 选择具有较大权重的波长点, 以及计算和比较变量子集的交叉验证均方 (root mean square error of cross-validation, RMSECV) 值。最小的变量子集即为最优子集^[20]。

在使用 CARS 进行有效波段的筛选时, 设置蒙特卡洛的采样次数为 100, 其他参数均为默认值。选择 RMSECV 最小时的变量子集作为最终建模的关键变量结果。经过 CARS 筛选变量后, 变量数目由 1 557 减少到 92。筛选全谱中关键变量的 CARS 方法的结果如图 2 所示。图 2A 中显示了随着变量筛选过程所选择的光谱变量的减少速度由快变慢。关键变量筛选过程中的 RMSECV 值的变化趋势在图 2B 中显示, 图中的趋势表明, 随着采样次数增加和冗余信息的消除, RMSECV 值不断变小, 达到最小值后继续采样, RMSECV 值反而增大, 说明关键指标信息随冗余信息被一同消除, 因此选择 RMSECV 值最小的变量子集即为最优。图 2C 中显示了变量筛选过程中光谱回归系数的变化。

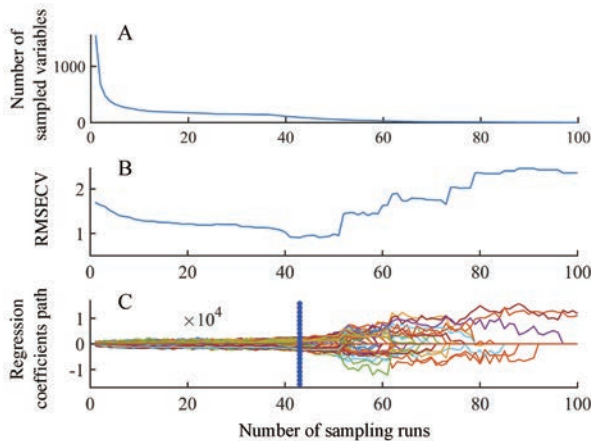


Figure 2 Competitive adaptive weighted resampling method (CARS) variable screening results. RMSECV: Root mean square error of cross-validation

3.2 RF法 随机青蛙算法 (random frog, RF) 基于逆跳跃马尔可夫链蒙特卡罗方法, 使用一组预定义数量的变量 (V_0) 和另一组不同数目的变量 (V_*) 作为候选子集来初始化子集, 初始化的子集在每次迭代时由候选集以一定概率更新。通过多次迭代后, 以每个变量的选择概率大小作为评价变量的重要性的依据^[21,22]。

采用 RF 法筛选后的各波长点的选择概率见图 3。在该实验中, 根据算法给出的选择概率结果将波数变量由高到低进行排序, 因为高选择概率代表了较高的特征性, 因此将选择概率排序前 10 的波数变量用作黄芩苷模型建立的特征变量。

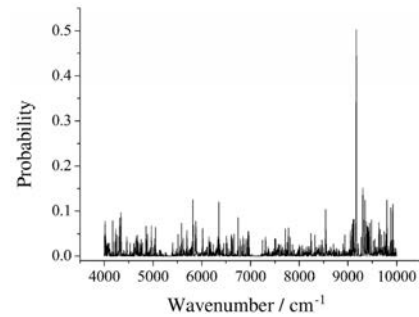


Figure 3 Random frog (RF) variable screening results

3.3 SPA法 作为经典的前向周期波长选择方法, 连续投影算法 (successive projections algorithm, SPA) 可以搜索包含最少冗余信息的变量组的频谱信息, 显著减少建模中使用的变量数量, 最小化变量之间的共线性, 从而显著提高建模的速度和效率^[23]。通过 SPA 方法筛选变量的原理是优化变量组, 其均方根误差 (root mean square error, RMSE) 最小。图 4 显示了在 SPA 方法中优化关键变量的过程中 RMSE 值与关键变量数量的变化。结果显示, 随着变量数的增加, RMSE 值变化的趋势先是不断下降最后趋于平缓, 经过 SPA 方法进行关键变量筛选后得到的变量数为 17, 作为关键变量进行 PLS 模型建立。

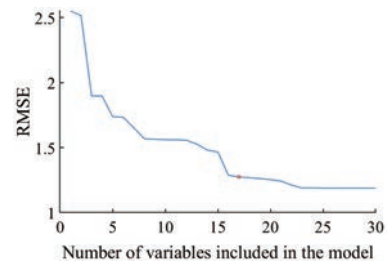


Figure 4 Successive projections algorithm (SPA) variable screening results. RMSE: Root mean square error

CARS 方法、RF 方法和 SPA 方法用于变量选择是基于不同的原理和选择依据, 故筛选出的关键变量也各不相同; 以不同变量筛选的结果进行建模, 模型的预测效果也有所不同, 更加突出了多变量筛选方法比较的重要性。比较各筛选结果和全光谱可发现, 变量由 1 557 显著减少到最多 92 个, 减少了至少 90% 以上, 起到了简化模型的作用。经过 CARS 法、RF 法和 SPA 法

筛选后的关键变量见表2。

3.4 黄芩苷定量模型的建立与预测 PLS模型的建立分别基于全光谱和各关键变量筛选后的结果进行,结果见表3。综合各指标的分析结果可知,经过CARS法变量筛选后的模型明显优于全谱和其他方法。因此,本研究选择CARS法进行变量筛选,并将结果应用于黄芩苷定量校正模型的建立和预测。黄芩苷NIRS预测值和标准对照方法参比值的相关分析见图5。黄芩苷定量校正模型的相关系数 R 值达到0.97以上, RMSEC和RMSEP值分别为0.528 2和0.720 2, RSEP值为5.59%,符合中药提取过程分析的准确度和精度要求。

Table 3 Comparison on modeling results by different variables screening methods

Method	Factor	Calibration set		Validation set		
		R_c	RMSEC	R_p	RMSEP	RSEP%
Full	8	0.917 0	1.057	0.833 5	1.362	10.58
CARS	9	0.979 9	0.528 2	0.956 5	0.720 2	5.59
RF	7	0.873 5	1.524	0.870 7	1.215	9.42
SPA	4	0.753 3	1.743	0.794 4	1.500	11.63

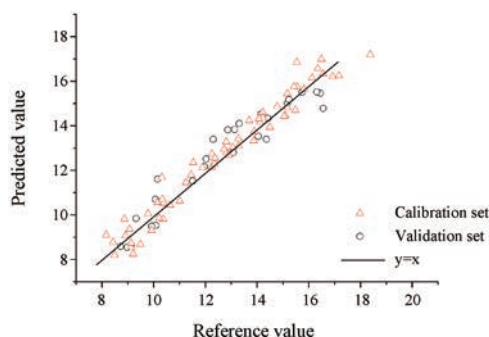


Figure 5 Correlation diagram between NIRS predicted values and reference values

讨论

本研究基于PLS建立了黄芩苷的定量校正模型,采用SPXY法划分数据集,并采用3种不同的变量筛选方法进行关键变量筛选并比较对模型的影响。结果表明,3种变量筛选方法筛选出的关键变量都可以去

掉冗余的变量信息,降低模型的复杂性。

SPA法筛选变量后仅得到17个变量,可以大大减少冗余的变量信息,但其建模效果并不理想,模型的相关系数为0.753 3, RSEP值为11.63%,模型效果甚至差于全谱模型。SPA法旨在最大程度的降低变量的共线性,这就有可能导致从变量中提取到的有效信息不足,数据被过度压缩,主要的化学信息被去除,从而影响了模型效果,说明SPA法不适合于选择黄芩苷NIRS分析数据的变量。

RF法和CARS法在压缩变量的同时还可以提高模型的预测能力和稳健性。其中RF-PLS模型的预测精度和鲁棒性与全光谱-PLS相比仅有微小提高, RSEP值降低至10%,模型预测结果较准确。但其参与建模的变量数从1 557减少到10,仅占全谱变量的0.6%。RF法筛选出的关键变量普遍集中在9 881~8 539 cm^{-1} 以及5 812 cm^{-1} 和6 352 cm^{-1} ,这部分的吸收主要由C-H、 CH_2 和 CH_3 的二级倍频和组合频以及氢键键合的O-H产生^[24],而黄芩苷中其他的化学结构信息则无法从筛选变量中得到体现。建模的结果可以说明其压缩变量的可行性,但由于部分关键化学信息被过度压缩,模型效果不够理想。

多种算法比较后,以CARS法筛选的变量建立的定量校正模型预测效果最好,与全光谱模型的预测性能相比,模型的相关系数从0.917 0上升到0.979 9, RSEP值从10.58%降低到5.59%。黄芩苷是黄酮类化合物,其结构中存在苯环、酚羟基、醇羟基、羧基以及醚键,提取液中还存在水溶液对光谱的影响,近红外光谱中最主要的吸收峰来自C-H的吸收谱带和O-H的吸收谱带。CARS法筛选出的波段也对应了原始光谱中的出峰位置,其中7 247~7 000 cm^{-1} 内的吸收主要是由游离O-H的一级倍频产生,7 140~6 940 cm^{-1} 区域出现酚O-H的一级倍频,羧酸单体游离O-H伸缩振动的一级倍频吸收出现在6 920 cm^{-1} 附近,6 850~6 240 cm^{-1} 之间的宽吸收峰由氢键键合的O-H产生,5 697~4 999 cm^{-1} 归属于C=O的二级倍频和醇的O-H合频吸

Table 2 Optimal wavenumbers selected by different variables screening methods

Method	Total of the number	Wavenumber
CARS	92	4 007, 4 011, 4 015, 4 019, 4 023, 4 030, 4 042, 4 046, 4 073, 4 077, 4 104, 4 108, 4 135, 4 169, 4 173, 4 246, 4 273, 4 300, 4 308, 4 320, 4 381, 4 501, 4 543, 4 632, 4 636, 4 663, 4 686, 4 694, 4 736, 4 752, 4 756, 4 775, 4 779, 4 790, 4 829, 4 852, 4 860, 4 883, 4 894, 4 902, 4 914, 4 999, 5 026, 5 029, 5 053, 5 118, 5 122, 5 126, 5 180, 5 188, 5 353, 5 357, 5 373, 5 431, 5 450, 5 697, 6 329, 6 333, 6 441, 6 453, 6 487, 6 507, 6 522, 6 599, 6 603, 6 615, 6 626, 6 661, 6 703, 6 742, 6 757, 6 780, 6 842, 6 850, 6 861, 6 946, 6 962, 6 966, 6 977, 6 993, 6 996, 7 000, 7 012, 7 031, 7 035, 7 081, 7 093, 7 108, 7 155, 7 174, 7 205, 7 247
RF	10	9 168, 9 303, 5 812, 9 800, 9 349, 6 352, 9 924, 9 881, 8 539, 9 299
SPA	17	7 151, 7 162, 5 319, 7 139, 5 365, 5 307, 5 431, 6 349, 5 377, 5 334, 7 128, 7 182, 5 288, 7 251, 7 108, 6 484, 7 085

收峰, 不饱和键 C-H 的合频吸收出现在 $4\ 660\ \text{cm}^{-1}$ 左右, $4\ 381\sim 4\ 007\ \text{cm}^{-1}$ 的吸收主要来自 C-H、 CH_2 、 CH_3 的组合频^[25], 这些结构在黄芩苷中均存在。而提取液中水的 O-H 伸缩振动的一级倍频和二级倍频吸收分别出现在 $6\ 944$ 和 $10\ 420\ \text{cm}^{-1}$, 合频吸收出现在 $5\ 155$ 和 $8\ 197\ \text{cm}^{-1}$, 这部分干扰变量未出现在筛选出的关键变量中, 说明 CARS 法能够筛选出与黄芩苷相关的关键变量, CARS-PLS 建立的定量校正模型可以为黄芩提取过程中黄芩苷的含量预测提供理论依据。

References

- [1] Lu WZ, Chu XL. Near infrared spectroscopy and process analytical technology [J]. Mod Sci Instrum (现代科学仪器), 2007, (4): 13-17.
- [2] Liu XS, Sun FF, Jin Y, et al. Application of near infrared spectroscopy combined with particle swarm optimization based least square support vector machine to rapid quantitative analysis of Corni Fructus [J]. Acta Pharm Sin (药学报), 2015, 50: 1645-1651.
- [3] Geng S, Chen Y, Jin Y, et al. Rapid quantitative determination of multi-indicators in Lonicerae Japonicae Flos by near-infrared spectroscopy [J]. Chin J Mod Appl Pharm (中国现代应用药学), 2016, 33: 144-150.
- [4] Jin Y, Ding HY, Wu YJ, et al. Near infrared spectroscopy on-line and real-time monitoring of extraction process of Xuebijing Injection [J]. Chin J Pharm Anal (药物分析杂志), 2012, 32: 1214-1221, 1234.
- [5] Liu XS, Chen JS, Chen GQ, et al. On-line detection of concentration process of Ganmaoling granules by near infrared spectroscopy combined with automatic control system [J]. Acta Pharm Sin (药学报), 2017, 52: 462-467.
- [6] Zhang Y, Wu YJ, Wu CY, et al. Online monitor multi-component content of Danhong injection in the alcohol precipitation process using near-infrared spectroscopy [J]. Chin J Mod Appl Pharm (中国现代应用药学), 2015, 32: 67-72.
- [7] Bai G, Ding GY, Hou YY, et al. Rapid identification and evaluation of Chinese materia medica quality by near infrared spectroscopy [J]. China J Chin Mater Med (中国中药杂志), 2016, 41: 3501-3505.
- [8] Bin J, Fan W, Zhou JH, et al. Application of intelligent optimization algorithms to wavelength selection of near-infrared spectroscopy [J]. Spectrosc Spectr Anal (光谱学与光谱分析), 2017, 37: 95-102.
- [9] He J, He Y, Zhang AC. Determination and visualization of peimine and peiminine content in *Fritillaria thunbergii* Bulbi treated by sulfur fumigation using hyperspectral imaging with chemometrics [J]. Molecules, 2017, 22: 1402.
- [10] Xia Z, Zhang C, Weng H, et al. Sensitive wavelengths selection in identification of *Ophiopogon japonicus* based on near-infrared hyperspectral imaging technology [J]. Int J Anal Chem, 2017: 6018769.
- [11] Zhan BA, Ni JH, Li J. Hyperspectral technology combined with CARS algorithm to quantitatively determine the SSC in Korla fragrant pear [J]. Spectrosc Spectr Anal (光谱学与光谱分析), 2014, 34: 2752-2757.
- [12] Zheng J, Zhou Z, Zhong SM, et al. Chestnut browning detected with near-infrared spectroscopy and a random-frog algorithm [J]. J Zhejiang A F Univ (浙江农林大学学报), 2016, 33: 322-329.
- [13] Li JB, Guo ZM, Hung WQ, et al. Near-infrared spectra combining with CARS and SPA algorithms to screen the variables and samples for quantitatively determining the soluble solids content in strawberry [J]. Spectrosc Spectr Anal (光谱学与光谱分析), 2015, 35: 372-378.
- [14] Chinese Pharmacopoeia Commission. Pharmacopoeia of People's Republic of China (中华人民共和国药典) [S]. Vol 1. 2015 ed. Beijing: China Medical Science Press, 2015: 301.
- [15] Jin Y, Wu ZZ, Liu XS, et al. Near infrared spectroscopy in combination with chemometrics as a process analytical technology (PAT) tool for on-line quantitative monitoring of alcohol precipitation [J]. J Pharm Biomed Anal, 2013, 77: 32-39.
- [16] Zhan XY, Zhao N, Lin ZZ, et al. Effect of algorithms for calibration set selection on quantitatively determining asiaticoside content in *Centella* total glucosides by near infrared spectroscopy [J]. Spectrosc Spectr Anal (光谱学与光谱分析), 2014, 34: 3267-3272.
- [17] Suo T, Wang H, Shi X, et al. Combining near infrared spectroscopy with predictive model and expertise to monitor herb extraction processes [J]. J Pharm Biomed Anal, 2018, 148: 214-223.
- [18] Li TT, Hu T, Nie L, et al. Rapid monitoring five components of ethanol precipitation process of Shenzhiling oral solution using near infrared spectroscopy [J]. China J Chin Mater Med (中国中药杂志), 2016, 41: 3543-3550.
- [19] Jin Y, Yang K, Wu YJ, et al. Application of particle swarm optimization based least square support vector machine in quantitative analysis of extraction solution of safflower using near-infrared spectroscopy [J]. Chin J Anal Chem (分析化学), 2012, 40: 925-931.
- [20] Wang Y, Jiang F, Gupta BB, et al. Variable selection and optimization in rapid detection of soybean straw biomass based on CARS [J]. IEEE Access, 2018, 6: 5290-5299.
- [21] Hu M, Zhai G, Zhao Y, et al. Uses of selection strategies in both spectral and sample spaces for classifying hard and soft blueberry using near infrared data [J]. Sci Rep, 2018, 8: 6671.
- [22] Zhang C, Kong W, Liu F. Measurement of aspartic acid in oilseed rape leaves under herbicide stress using near infrared spectroscopy and chemometrics [J]. Heliyon, 2016, 2: e00064.
- [23] Chen LY, Zhao ZG, Liu F. mRMR-based wavelength selection for quantitative detection of Chinese yellow wine using NIRS [J]. Anal Methods, 2018, 10: 667-675.
- [24] Lacy JH. Interpretation of infrared vibration vibration-rotation spectra of interstellar and circumstellar molecules [J]. Astrophys J, 2013, 765: 1035-1045.
- [25] Nishimura Y, Lee YP, Irle S, et al. Critical interpretation of CH- and OH- stretching regions for infrared spectra of methanol clusters $(\text{CH}_3\text{OH})_n$ ($n = 2-5$) using self-consistent-charge density functional tight-binding molecular dynamics simulations [J]. J Chem Phys, 2014, 141: 094303.