热毒宁注射液金银花和青蒿（金青）萃取过程中固形物含量近红外光谱在线监测模型的建立及萃取终点判断研究

童　枫[1,2]，徐芳芳[1,2*]，闫逸伦[1,2]，李执栋[1,2]，张永超[1,2]，刘恒旭[1,2]，章晨峰[1,2]，王振中[1,2]，张　欣[1,2*]

1. 中药制药过程控制与智能制造技术全国重点实验室，江苏 连云港　222001
2. 江苏康缘药业股份有限公司，江苏 连云港　222001

摘　要：目的　采用近红外光谱（near-infrared spectroscopy，NIRS）技术，结合机器学习算法，实现热毒宁注射液（Reduning Injection，RI）金银花和青蒿（金青）萃取过程中固形物含量（solid content，SC）的在线监测，并基于 NIRS 技术建立萃取终点判别模型，以提高金青萃取过程的质量控制水平。**方法**　采用 NIRS 技术，结合偏最小二乘法（partial least squares，PLS）和多元自适应回归样条（multivariate adaptive regression splines，MARS）算法，模型经过光谱预处理方法的优选及特征变量筛选，建立最佳 SC 的在线监测模型；采用支持向量机（support vector machine，SVM）算法建立异常光谱判别模型，通过移动块标准偏差法（moving block standard deviation，MBSD）算法建立萃取终点判别模型。**结果**　PLS 和 MARS 模型性能优异，相较于 PLS 模型，MARS 模型性能有所提升，预测相对误差（relative standard error of prediction，RSEP）由 2.87%降低至 2.64%，性能偏差比（ratio of performance to deviation，RPD）由 15.953 0 升至 17.376 1，2 种算法模型均具有模型性能好、预测精度高的优点；MBSD 算法用于萃取终点的判断，可有效提升萃取效率。**结论**　NIRS 技术结合 PLS 算法和 MARS 算法，均可用于 RI 金青萃取过程 SC 的在线监测，MARS 模型性能更佳；采用 MBSD 方法进行萃取终点判断，方法简便易行，可以满足生产实际需求。

关键词：近红外光谱；热毒宁注射液；金银花；青蒿；机器学习；在线监测；终点判断；萃取过程；固形物含量；偏最小二乘法；多元自适应回归样条；支持向量机；移动块标准偏差法

中图分类号：R283.6　　**文献标志码：**A　　**文章编号：**0253 - 2670(2024)19 - 6555 - 11

DOI: 10.7501/j.issn.0253-2670.2024.19.010

Establishment of near-infrared spectroscopy online monitoring solid content and determination of extraction endpoint during extraction of *Lonicerae Japonicae Flos* and *Artemisiae Annuae Herba* (Jinqing) in Reduning Injection

TONG Feng[1, 2], XU Fangfang[1, 2], YAN Yilun[1, 2], LI Zhidong[1, 2], ZHANG Yongchao[1, 2], LIU Hengxu[1, 2], ZHANG Chenfeng[1, 2], WANG Zhenzhong[1, 2], ZHANG Xin[1, 2]

1. State Key Laboratory on Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture, Lianyungang 222001, China
2. Jangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

Abstract: Objective　To adopt near-infrared spectroscopy (NIRS) technology and machine learning algorithms, the on-line monitoring of solid content (SC) in the extraction of Jinyinhua (*Lonicerae Japonicae Flos*) and Qinghao (*Artemisiae Annuae Herba*) (Jinqing) in Reduning Injection (RI) was realized, and the extraction endpoint discrimination model based on NIRS was established to improve the quality control level of the extraction process. **Methods**　Using NIRS technology, combined with partial least squares (PLS) and multiple adaptive regression spline algorithm (MARS), the optimal model for online monitoring of solid content (SC) was developed through the selection of spectral pretreatment methods and feature variable screening. Support vector machine (SVM)

收稿日期：2024-03-28

基金项目：国家科技部长三角科技创新共同体联合攻关项目（2023CSJGG1700）；连云港市市重点研发计划（产业前瞻与关键核心技术）：基于 PAT 的中药浓缩和萃取过程反馈调控技术研究（CG2320）

作者简介：童　枫，男，硕士，研究方向为中药制药过程新技术。E-mail: tongfeng5324@126.com

***通信作者：**张　欣，博士，研究方向为中药制药过程新技术。E-mail: zxtcm@126.com

徐芳芳，副主任药师，硕士生导师，从事中药智能制造研究。E-mail: 879164331@qq.com

algorithm was used to establish the abnormal spectrum discrimination model for data cleaning, and subsequently, the extraction endpoint discrimination model was established by moving block standard deviation (MBSD) method. **Results** Compared with the PLS model, the performance of the MARS model was improved, and the relative standard error of prediction (RSEP) was reduced from 2.87% to 2.64%, ratio of performance to deviation (RPD) increased from 15.953 0 to 17.376 1. Both of the two algorithm models have the advantages of good model performance and high prediction accuracy. The MBSD algorithm can be used to determine the extraction end point, which can effectively improve the extraction efficiency by about 18%. **Conclusion** NIRS technology, when integrated with both PLS and MARS algorithms, is effective for online monitoring of SC during the Jinqing extraction process of RI, with the MARS model showing superior performance. The MBSD method for endpoint determination is straightforward and meets the practical needs of production.

Key words: near-infrared spectroscopy; Reduning Injection; *Lonicerae Japonicae Flos*; *Artemisiae Annuae Herba*; machine learning; online monitoring; endpoint determination; extraction process; solid content; partial least squares; multivariate adaptive regression splines; support vector machine; moving block standard deviation

热毒宁注射液（Reduning Injection，RI）是江苏康缘药业股份有限公司的独家品种，由金银花、青蒿和栀子 3 味药材制得，具有清热、疏风、解毒的功效，主要用于治疗上呼吸道感染引起的咳嗽、高热、头身疼痛、痰色发黄等症状[1-3]。金银花和青蒿（金青）萃取是 RI 提取精制过程中的关键工艺单元之一，在萃取过程中，将一定倍量的萃取溶剂以一定的体积流量自下而上打入萃取塔，从而将有效成分从金青醇沉浸膏中充分萃取出来，经过萃取后，有效成分通过流通管道被输送到贮存罐中，以供后续的加工和制备流程使用。有效的金青萃取过程可以大幅缩短生产周期，提高生产效率，如果萃取不完全，则需要反复进行萃取和调整，这将导致生产成本的增加和生产周期的延长。因此，实现萃取终点的准确判断至关重要，对于提高生产效率、降低生产消耗具有积极意义，能够为 RI 的高效、稳定生产提供有力保障。目前，终点的判断仍然依赖于人工经验和传统分析方法[4]，这种方法既耗时又效率低下，因此，借助先进的过程分析技术进行萃取终点的准确监测显得尤为重要。

在 RI 金青萃取过程中，固形物含量（solid content，SC）是一个关键的质量指标，通过检测 SC，可以了解有效成分的萃取程度，从而调整工艺参数，提高生产效率。然而，常规的 SC 检测方法需要离线进行，并存在检测时间长、分析效率低、样品被破坏等缺点。近年来兴起的近红外光谱（near-infrared spectroscopy，NIRS）技术，作为过程分析技术的重要代表，相较于传统的化学分析方法，具有操作简便、对样品破坏性小、分析速度快、对环境友好等优点，目前已经广泛应用于中药制剂生产过程在线监测与终点判断等方面[5-7]。

本研究以 RI 金青萃取过程为研究对象，通过采用 NIRS 技术结合偏最小二乘算法（partial least squares，PLS）和多元自适应回归样条算法（multivariate adaptive regression splines，MARS），建立金青萃取过程在线 NIRS SC 监测模型，并结合移动块标准偏差法（moving block standard deviation，MBSD），通过建立定量和定性模型，实现对 RI 金青萃取过程终点的准确判断。

1 仪器与材料

1.1 仪器

MATRIX-F 型在线傅里叶变换近红外光谱仪，德国 Bruker 公司；DHG-9135A 型电热鼓风干燥箱，上海一恒科学仪器公司；ME104E 型电子天平，梅特勒-托利多仪器（上海）有限公司。

1.2 材料

7 批次 RI 金青萃取过程样品，批号 Z230307、Z230311、Z230317、Z230327、Z230330、Z230334、Z230338，均来自江苏康缘药业股份有限公司数字化提取工厂，过程萃取开始 30 min 后采集第 1 个样本，每隔 15～20 min 取样 1 次，直至萃取结束，7 批金青萃取过程分别收集 36、28、29、29、32、30、31 个样本，共计 215 个样本。

2 方法与结果

2.1 NIRS 的采集

在流通管道末端安装检测流通池，金青萃取液经流通管道流入贮存罐，在线近红外光谱仪通过光纤与检测流通池相连，从光源发射近红外光，经光纤到达流通池，流通池内金青萃取液自上而下流入，近红外光透射过池内萃取液后，经光纤被光谱仪重新接收，经仪器处理后完成光谱数据的采集，检测后的萃取液被重新输送回流通管道。金青萃取 NIRS 在线应用示意图如图 1 所示。

在线近红外光谱仪每隔 20 s 收集 1 张光谱，以

空气为扫描背景；光谱扫描范围 12 500～4 500 cm^{-1}；样品扫描时间 8 scans；分辨率 16 cm^{-1}；根据样品取样时间推算样本光谱采集时间，取 1 min 内连续的 3 张光谱的平均值，作为金青萃取样本光谱数据。共采集 7 个批次共 215 个 NIRS 样本，NIRS 结果如图 2 所示。

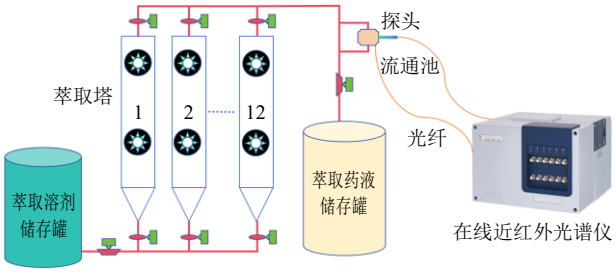

图 1　金青萃取 NIRS 在线应用示意图

Fig. 1　Schematic diagram of NIRS online application of Jinqing extraction

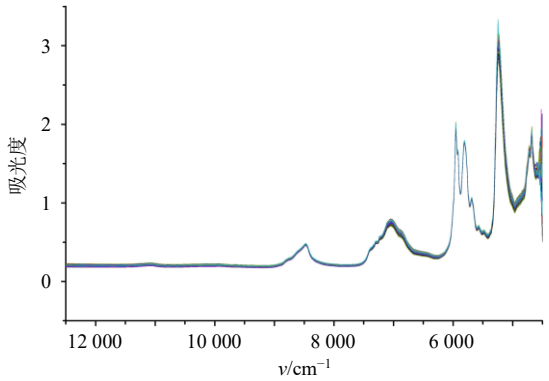

图 2　215 批样本 NIRS 平均光谱图

Fig. 2　Average NIRS of 215 samples

2.2　SC 的测定

参考文献方法[8-9]，称取约 5 g 样品至已烘干至恒定质量的称量瓶（X_0）中，称定质量（X_1），置烘箱 105 ℃条件下烘干 5 h 至恒定质量，计为 X_2。SC 计算公式如下。

$$SC=(X_2-X_0)/(X_1-X_0) \quad (1)$$

2.3　样本划分

对采集到的 7 个批次 215 个样本，进行样本划分。将 Z230338 批次的 31 个样本作为外部验证集，剩余的 6 个批次 184 个样本，采用光谱-理化值共生（sample set partitioning based on joint *x*-*y* distance，SPXY）算法[10]以 4∶1 的比例划分为校正集和验证集，SPXY 算法兼顾参考值与光谱距离，可以保证划分后的样本集的光谱和参考值都覆盖较大范围且均匀分布。最终得到校正集样本 147 个，验证集样本 37 个。样本划分结果如表 1 所示。

表 1　样本集划分结果描述统计

Table 1　Description table of sample set partitioning results

样本集	样本数	SC/%				
		最小值	下四分位数	上四分位数	最大值	平均值
校正集	147	0.344 5	0.730 7	1.675 0	2.413 4	1.230 8
验证集	37	0.349 8	0.490 1	0.770 2	2.123 4	0.670 5
外部验证集	31	0.405 0	0.585 0	1.569 0	2.404 0	1.118 0

2.4　数据处理软件

采用 Unscrambler 11.0（挪威 Camo Analytics 公司）软件进行光谱预处理、SVM 异常数据清洗模型和 MBSD 模型的建立；采用 Python 3.5.3 软件进行样本集划分、特征变量筛选以及 PLS 模型建立；采用 SPM 8.3 软件（美国 Salford Systems 公司）进行 MARS 模型建立；采用 GraphPad Prism 9.1 软件（美国 GraphPad Software 公司）绘图。

2.5　MARS 算法[11]

MARS 算法是一种非线性和多维关系建模的算法，旨在通过使用基函数构建分段线性回归模型，并将其拟合到独立变量的不同区间中，从而建立更为灵活的回归模型。在 MARS 算法中，数据训练集被划分为独立且具有不同梯度的分段线段。算法通过逐步搜索来生成基函数，并利用自适应回归算法来确定结点的位置。每个分段线段被称为基函数，而各段的端点则被称为结点。MARS 算法分为前向选择、后向剪枝 2 个步骤，在前向选择过程中，算法生成一个过拟合模型，而在后向剪枝过程中，算法会删除导致模型过拟合的基函数，从而筛选出最优模型。为了选择和剔除多余的基函数，MARS 算法采用广义交叉验证法（generalized cross-validation，GCV）来确定最佳模型，最终的 MARS 模型是 GCV 值最小的模型[12-13]。

2.6　模型性能评价

本研究以样本 SC 为因变量，以在线 NIRS 光谱为自变量，分别采用 PLS 算法和 MARS 算法建立 SC 预测模型。以校正集决定系数（R_c^2）、验证集决定系数（R_p^2）、校正均方根误差（root mean square error of calibration，RMSEC）、验证均方根误差（root mean square errors of prediction，RMSEP）、交叉验证均方根误差（root mean square error of cross validation，RMSECV）、性能偏差比（ratio of performance to deviation，RPD）、预测相对误差（relative standard error of prediction，RSEP）为指标评价模型[8]。一般

而言，性能优异的模型应具有较高的 R_c^2、R_p^2 和 RPD、较小且接近的 RMSEC 和 RMSEP 以及较小的 RSEP[8]。此外，使用 RMSEP 与 RMSEC 的比值（RMSEP/RMSEC）作为评价模型过拟合和欠拟合程度的标准，该值过大易过拟合，过小易欠拟合，通常认为该值在 0.8～1.2 时，所建立的模型是可以接受的[14]。本研究 PLS 模型使用留一交叉验证法，根据 RMSECV 结果确定 PLS 模型的最佳的潜变量数（latent variables，LVs），MARS 模型采用广义交叉验证法，以 GCV 结果确定基函数个数（number of basis functions，NBF）。上述评价指标的相关公式为式（2）～（6）。

$$R_c^2=1-\sum(y-y_i)^2/\sum(y-y_m)^2 \quad (2)$$

$$R_p^2=1-\sum(y-y_i)^2/\sum(y-y_n)^2 \quad (3)$$

$$\mathrm{RMSEC}=[\sum(y-y_i)^2/m]^{1/2} \quad (4)$$

$$\mathrm{RMSEP}=[\sum(y-y_i)^2/n]^{1/2} \quad (5)$$

$$\mathrm{RSEP}=\sum(y-y_i)^2/\sum y^2 \quad (6)$$

m、n 为校正集、验证集样本数，y 为实测值，y_i 为预测值，y_m、y_n 分别为校正集和验证集实测值的平均值

2.7 预测模型建立及结果

2.7.1 关键质量属性（critical quality attributes，CQA）确定 金青萃取过程是 RI 提取精制过程的关键工艺单元之一，根据企业内部标准，其 CQA 包括绿原酸、新绿原酸、隐绿原酸、异绿原酸 A、异绿原酸 B、异绿原酸 C、断氧化马钱子苷和 SC 共 8 个含量指标。为使模型更具代表性，需探索模型的关键质量属性，因而，本研究依据候化蕊等[15]的实验结果，采用 Person 相关系数法，计算另外 7 个关键质量属性与 SC 间的相关系数，结果如图 3 所示。结果表明，SC 与其余 7 个 CQAs 呈高度相关，且均为正相关，新绿原酸、绿原酸和隐绿原酸含量与 SC 的相关系数均大于 0.8，其余 4 个 CQAs 与 SC 的相关系数也均大于 0.6。结果表明，SC 确为 RI 提取精制过程的 CQA，通过建立在线 NIRSSC 预测模型，可以快速准确地反映产品质量变化。

2.7.2 光谱预处理方法的选择 NIRS 不仅包含了样品本身的物理结构和化学成分信息，还可能受到仪器暗电流、样品背景与状态、杂散光和环境变化等多种因素的影响，从而引入光谱噪声。这些噪声可能会对后续的建模分析造成干扰，因此，在建模前，需对采集的 NIRS 数据进行预处理，通过预处理，可以有效地减少噪声、净化无用信息，从而提高模型的精度和预测效果[16]。

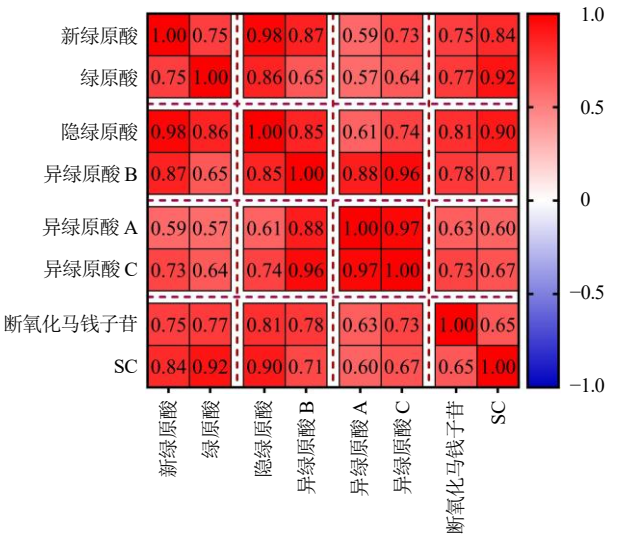

图 3 CQAs 的 Person 相关系数结果

Fig. 3 Results of Person correlation coefficients for CQAs

光谱预处理方法主要分为以下 4 类[17]：（1）基线校正：消除基线漂移的影响；常见的方法包括一阶导数、二阶导数和小波变换等[18]。（2）散射校正：消除因样本颗粒大小和分布不均匀产生的散射影响；常见的方法包括多元散射校正（multivariate scattering correction，MSC）和标准正态变量变换（standard normal transformation，SNV）等。（3）平滑校正：降低光谱的随机噪声，提高信噪比；常见的方法包括移动平均法（moving average，MA）、Savitzky-Golay（S-G）卷积平滑法和高斯滤波等。（4）尺度缩放：消除尺度差异带来的影响；常见的方法包括最大最小归一化法、标准化变换、中心化变换等。本研究采用 MA、S-G 卷积平滑、S-G 卷积平滑＋一阶导数（S-G＋1st）、去趋势、基线校正、归一化法、SNV、MA-SNV 等多种预处理方法，基于 PLS 和 MARS 算法，分别建立金青萃取过程 SC 预测模型，不同预处理方法对模型性能的影响如表 2、3 所示。以 RPD 和 RSEP 为主要评价指标，结合剩余指标综合评价模型。结果表明，针对 PLS 模型，当以 MA 为预处理方法时，相较于无预处理的模型，该预处理模型的 R_c^2 和 R_p^2 均有所提升，RMSEC 和 RMSEP 较小且接近，均小于 0.025%，此时，RPD 为 14.295 3，RSEP 为 3.21%，模型性能显著提升；针对 MARS 模型，当以 MA 为预处理方法时，相较于无预处理的模型，该预处理模型的 R_c^2 和 R_p^2 均有所提升，RMSEC 和 RMSEP 较小，均小于 0.035%，此时 RPD 为 17.415 5，RSEP 为 2.63%，模型性能显著提升。

表 2 不同预处理方法对 PLS 模型性能的影响

Table 2 Effects of different pretreatment methods on the performance of PLS model

预处理方法	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RMSEP/RMSEC	RPD	RSEP/%	LVs
无预处理	0.992 4	0.988 2	0.049 5	0.038 8	0.78	7.924 8	5.78	18
MA	0.998 0	0.997 2	0.024 9	0.021 5	0.86	14.295 3	3.21	20
S-G 平滑	0.992 2	0.988 4	0.050 1	0.040 1	0.80	7.830 8	5.85	16
S-G＋1st	0.996 6	0.988 0	0.032 9	0.036 4	1.11	8.437 2	5.43	20
归一化法	0.993 0	0.986 4	0.047 6	0.038 5	0.81	7.991 9	5.74	18
基线校正	0.993 8	0.986 0	0.044 6	0.040 5	0.91	7.589 9	6.04	20
去趋势	0.993 8	0.985 5	0.044 6	0.040 1	0.90	7.663 1	5.98	19
SNV	0.991 4	0.983 7	0.052 8	0.044 5	0.84	6.908 4	6.64	17
MA-SNV	0.997 6	0.995 6	0.028 0	0.022 5	0.80	13.667 4	3.35	19

表 3 不同预处理方法对 MARS 模型性能的影响

Table 3 Effects of different pretreatment methods on performance of MARS model

预处理方法	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RMSEP/RMSEC	RPD	RSEP/%	NBF
无预处理	0.996 5	0.981 7	0.033 7	0.041 0	1.22	7.491 7	6.12	18
MA	0.996 9	0.996 6	0.031 9	0.017 7	0.55	17.415 5	2.63	22
S-G 平滑	0.996 9	0.993 3	0.031 6	0.024 8	0.78	12.409 5	3.69	10
S-G＋1st	0.997 8	0.994 1	0.026 7	0.023 2	0.87	13.249 3	3.46	20
归一化法	0.995 5	0.988 6	0.038 3	0.032 4	0.85	9.493 0	4.83	10
基线校正	0.998 1	0.993 9	0.025 1	0.023 7	0.94	12.991 7	3.53	13
去趋势	0.998 7	0.995 8	0.020 3	0.019 6	0.97	15.666 8	2.93	23
SNV	0.978 2	0.970 7	0.084 0	0.051 9	0.62	5.920 3	7.74	26
MA-SNV	0.999 7	0.995 1	0.009 6	0.021 9	2.28	14.048 6	3.26	29

2.7.3 特征变量的筛选 NIRS 数据往往含有数千个波数点，然而并不是所有波数变量都与目标成分相关，因此，需要从完整光谱中，剔除冗余波数，筛选出具有代表性的重要波数，以此增强模型解释性，简化模型并提高模型的预测精度[19]。本研究在上述筛选出的最佳预处理方法的基础上，对比全光谱、组合间隔 PLS（synergy interval PLS，siPLS）、连续投影算法（successive projections algorithm，SPA）和竞争性自适应重加权采样法（competitive adapative reweighted sampling，CARS）建模的效果。此外，MARS 模型也可按照变量重要性（variable importance in the projection，VIP）排序，通过软件自动剔除最不重要的变量，并重新建模。

siPLS[20]通过对不同子区间进行任意组合，建立所有可能的 2、3 或 4 个区间的 PLS 回归模型。本研究将光谱区间均分为 20 个子区间，以子区间组合数为 3 建立模型，以 RMSECV 作为评价指标，筛选最优建模波数。SPA 是一种前向迭代搜索方法，从某一波长开始，然后在每次迭代中加入一个新变量，直至所选变量数达到设定值。SPA 能够在筛选出最低冗余信息变量组合的同时，有效消除变量间的共线性问题，最大限度地获取解释信息，并降低模型的复杂度[21]。CARS 是一种结合蒙特卡洛采样与 PLS 模型回归系数的特征变量选择方法，通过获得 RMSECV 最小的子集中的波数作为特征波数[22]，本研究采样迭代次数选择 50 次，并采用 10 折交叉验证。特征变量筛选结果如图 4 所示。

根据特征变量筛选结果，基于 PLS 和 MARS 算法，以 RPD 和 RSEP 为主要评价指标，结合剩余指标综合评价模型，分别建立金青萃取过程 SC 预测模型，不同特征筛选结果对模型性能的影响如表 4、5 所示。可见，针对 PLS 模型，相较于无处理的模型，经 siPLS 算法筛选的特征变量效果最佳，特征变量数目由 2 074 个降为 310 个，模型 R_c^2 和 R_p^2 变化不大，RMSEC 和 RMSEP 较小且接近，均小于 0.020%，此时，RPD 为 15.953 0，RSEP 为 2.87%，

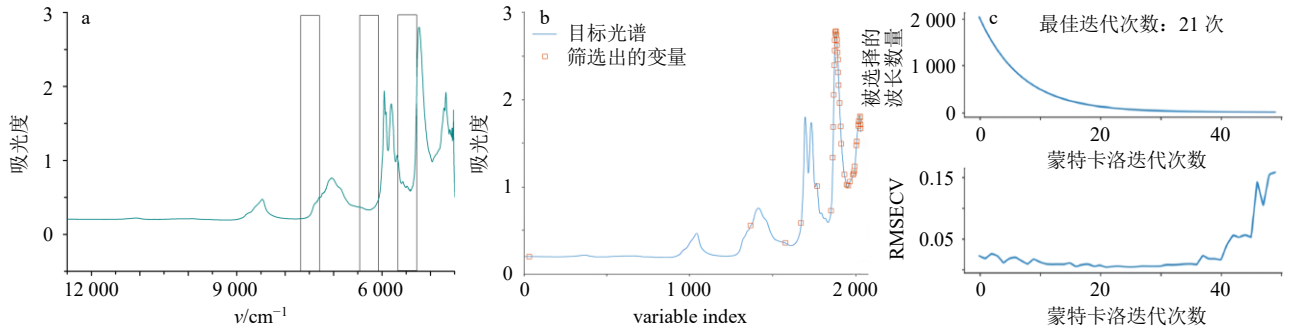

图 4　siPLS (a)、SPA (b) 和 CARS (c) 算法特征变量筛选结果

Fig. 4　siPLS (a), SPA (b), and CARS (c) algorithm feature variable screening results

表 4　不同特征变量筛选方法对 PLS 模型的影响

Table 4　Influences of different feature variable screening methods on PLS model

特征变量筛选方法	特征变量数目	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RMSEP/RMSEC	RPD	RSEP/%	LVs
无处理	2 074	0.998 0	0.997 2	0.024 9	0.021 5	0.86	14.295 3	3.21	20
siPLS	310	0.999 2	0.997 2	0.017 4	0.019 3	1.11	15.953 0	2.87	16
SPA	42	0.997 2	0.997 0	0.030 6	0.019 7	0.64	15.623 6	2.93	17
CARS	104	0.997 4	0.996 6	0.029 4	0.021 1	0.72	14.569 7	3.15	14

表 5　不同特征变量筛选方法对 MARS 模型的影响

Table 5　Influences of different feature variable screening methods on MARS model

特征变量筛选方法	特征变量数目	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RMSEP/RMSEC	RPD	RSEP/%	NBF
无处理	2 074	0.996 9	0.996 6	0.031 9	0.017 7	0.55	17.415 5	2.63	22
VIP	8	0.995 1	0.984 6	0.039 7	0.037 6	0.95	7.394 4	5.61	15
siPLS	310	0.995 1	0.984 6	0.039 7	0.037 6	0.95	7.394 4	5.61	19
SPA	42	0.996 2	0.991 5	0.034 9	0.028 0	0.80	10.989 8	4.17	25
CARS	104	0.999 1	0.996 6	0.016 9	0.017 7	1.05	17.376 1	2.64	14

模型稳定性和精度有所提升；针对 MARS 模型，相较于无处理的模型，经 CARS 算法筛选的特征变量效果最佳，特征变量数目由 2 074 个降为 104 个，模型 R_c^2 有所提升，RMSEC 和 RMSEP 较小且接近，均小于 0.020%，此时，RPD 为 17.376 1，RSEP 为 2.64%，模型性能有所提升。

2.7.4　最佳模型的建立　基于 PLS 和 MARS 算法，通过光谱预处理方法的选择和特征变量的筛选后，优选出最佳的 SC 预测模型，最佳模型结果如表 6 所示。对应 SC 最佳预测模型的预测值与实测值的相关性关系如图 5 所示。结果表明，PLS 和 MARS 的最佳预测模型性能优异，其 R^2 均大于 0.99，且 RMSEC 和 RMSECV 均小于 0.020%，RPD 分别为 15.953 0 和 17.376 1，RSEP 分别为 2.87%和 2.64%，均小于 3%。综合对比，MARS 模型性能强于 PLS 模型，且均能满足实际生产需求，可以用于金青萃取过程 SC 在线监测。为验证 PLS 和 MARS 模型预测结果的可靠性，对 SC 定量模型的验证集中参考

表 6　PLS/MARS 最佳模型性能对比

Table 6　Performance comparison of the best PLS/MARS models

算法	预处理方法	特征变量筛选方法	特征变量数目	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RMSEP/RMSEC	RPD	RSEP/%
PLS	MA	siPLS	310	0.999 2	0.997 2	0.017 4	0.019 3	1.11	15.953 0	2.87
MARS	MA	CARS	104	0.999 1	0.996 6	0.016 9	0.017 7	1.05	17.376 1	2.64

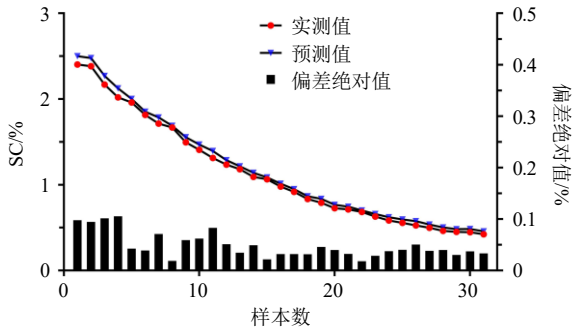

图 5 最佳模型中预测值与实测值的相关性

Fig. 5 Correlation between predicted value and measured value in the best models

值与预测值进行配对 t 检验，结果其 P 值分别为 0.207 和 0.106，均大于 0.05，说明参考值与预测值之间无明显差异。

2.8 模型外部验证

将外部验证集导入已建立的最佳模型中，即 MARS SC 预测模型中。通过比较样品的实测值和模型预测值，计算其偏差绝对值和平均相对偏差，以平均相对偏差为评价指标，验证在线 NIRS MARS 模型的泛化能力。结果如图 6 所示。结果显示，MARS 模型的预测值紧密围绕着 SC 实测值曲线，模型对金青萃取过程 SC 变化实现了较好的预测，其平均相对偏差为 4.888%，小于 5.0%，模型预测精度满足实际生产需求。

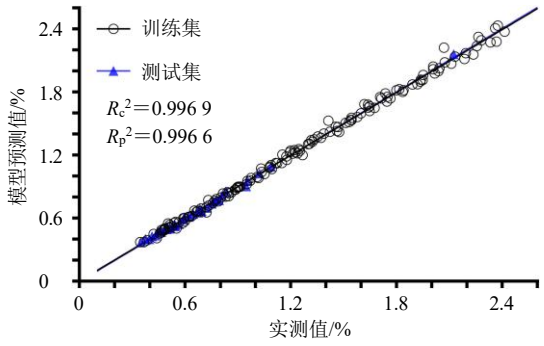

图 6 模型外部验证结果

Fig. 6 Results of external validation of model

2.9 金青萃取终点判断

2.9.1 在线 NIRS 采集 依据“2.1”项下光谱采集条件收集 7 个批次的完整金青萃取过程在线 NIRS（A1～A7，批号 20230222、20230225、20230307、20230407、20230920、20240108、20240109）样本。

2.9.2 数据清洗

（1）异常数据清洗模型建立：由于金青萃取现场工况环境复杂，采集到的原始 NIRS 中存在大量异常光谱，若直接采用全部光谱计算会干扰计算结果，为保障数据的准确性和完备性，需预先对异常数据进行清洗。考虑到光谱数据量较大，难以手动剔除异常数据，本研究采用支持向量机（support vector machine，SVM）算法对异常光谱数据进行清洗。SVM 算法是基于统计学理论提出的一种针对二分类问题的监督学习模型，其目的是通过寻求 1 个最优超平面将 2 种不同类别的样本分开，具有学习速度快、全局最优和泛化能力强的优点，黄国东等[23]基于 SVM 算法对供水管网监测数据进行清洗，获得了良好的异常检测性能。

本研究从批次 A1～A5 中挑选出正常光谱和异常光谱各 100 条，合计 200 条光谱，用于建模的异常光谱，应尽可能地包含所有可能出现的异常情况。将光谱定义为 2 类，正常光谱为第 I 类，异常光谱为第 II 类，并选择 SNV＋1st 作为光谱预处理方法，5 446.1～9 411.2 cm^{-1} 作为建模区间，通过网格搜索筛选惩罚因子 C 和核函数 γ 建立 SVM 模型，筛选结果如图 7 所示，当 C 值为 10，γ 值为 0.01 时，模型的预测精度达到 100%。SVM 模型结果如图 8 所

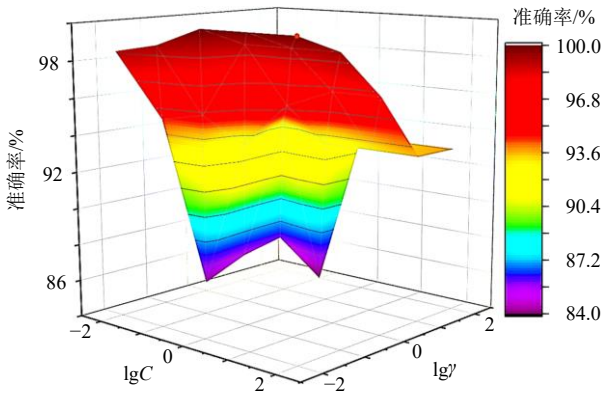

图 7 网格搜索筛选惩罚因子 C 和核函数 γ

Fig. 7 Grid search filters penalty factor C and kernel function γ

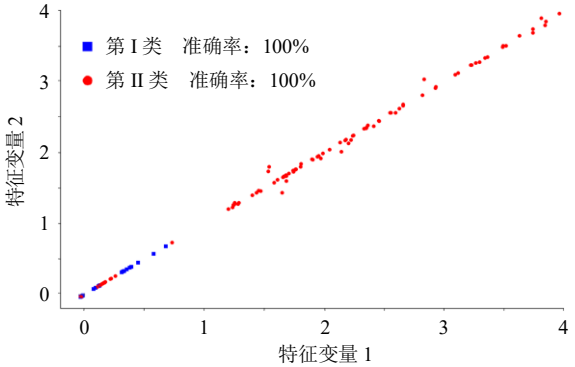

图 8 SVM 分类模型

Fig. 8 SVM classification model

示，此时，模型判别准确率达到 100%，满足数据清洗需求。

（2）数据清洗结果：使用建立的 SVM 模型分别预测上述 2 个批次（A6、A7）NIRS 样本，并根据预测结果批量对异常光谱数据进行数据清洗，清洗结果如图 9 所示，A6 和 A7 经数据清洗后，光谱矩阵分别由原始的 2 984×2 074、3 007×2 074 降为 2 216×2 074、2 248×2 074，占比约 74%。

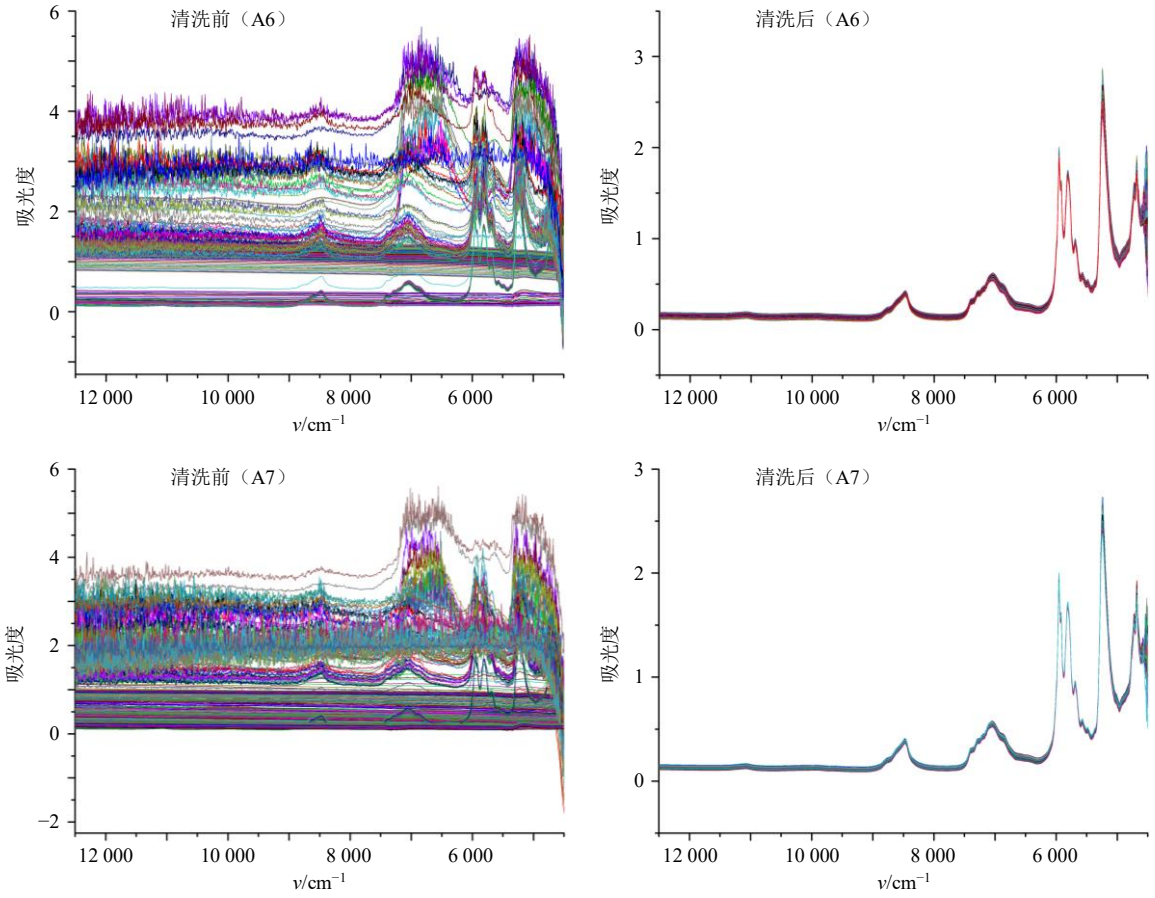

图 9　NIRS 数据清洗结果

Fig. 9　Data cleaning results of NIRS

2.9.3　终点判别模型建立

（1）移动块标准偏差法（moving block standard deviation，MBSD）模型建立：MBSD 算法是常用于工艺过程终点判断的定性算法之一，其实现过程是以 *n* 张连续光谱为 1 个窗口，计算窗口内光谱每个波数下的平均标准偏差，然后求得 *n* 张光谱的平均标准偏差，即是该窗口下的 MBSD 值，通过将窗口沿样品光谱方向选取合适的步长移动，继而表征样品光谱间的差异[6]。

本研究在上一步数据清洗的基础上，以窗口宽度 15，步长 1 建立移动块均值（moving block mean，MBM）模型和 MBSD 模型。结果如图 10 所示，当批次 A6、A7 金青萃取过程进行到 *X*（光谱序号）＝1 724、1 779 时，分别对应实际时间 9.58 h 和 9.88 h，此时光谱的 MBSD 值和 MBM 值波动较小并趋于稳定，可以认为萃取过程已到达终点。相较于实际的萃取结束时间 11.75 h 和 11.95 h，萃取效率分别提升了 18.47%和 17.32%，极大地提升萃取效率及减少溶剂浪费。

从 MBSD 趋势图可以看出，萃取过程光谱差异呈现先急剧上升后指数下降最后趋于平缓的变化趋势。推测原因是随着萃取溶剂的加入，前期大量有效物质成分被萃取出来，造成光谱间差异巨大，随着萃取的进行萃取出的有效物质成分减少，使得光谱间差异减小，直至萃取终点，有效物质成分被萃取完全，此时光谱反映的是萃取溶剂状态，因此光谱间基本无差异，趋势趋于平缓。

（2）主成分分析法（principal component analysis，PCA）算法模型建立：由于 NIRS 矩阵中含有巨量的数据点，使得计算的复杂程度增大，PCA

算法可实现对多变量数据的降维，并以图形的方式表现，可以更直观地表现样本的分类聚集情况。本研究在“2.9.2”的基础上，对清洗后的数据进行 PCA，PCA 降维后的主成分得分图如图 11 所示。从图中可以明显看出，2 批数据的主成分得分图有自右向左移动的趋势，最终聚成一团，该结果可以与 MBSD 模型结果对应，此时得分数据分布集中，光谱差异极小，可以认定此时即为萃取终点。

2.9.4 模型验证 在“2.7”项的基础上，基于已建立的最佳 SC 预测模型，预测上述 2 个批次（A6、A7）在金青萃取过程中的 SC 变化情况，SC 预测变化趋势如图 12 所示。结果表明，MARS 模型 SC 预测结果与 MBSD 模型结果一致，均呈现先增大后减小最后趋于平缓的趋势，且当萃取过程到达 X=1 724 和 X=1 779 附近后，SC 变化极小，可以认为此时已达到萃取终点。SC 预测结果表明，MBSD 法预测的萃取终点与实际终点基本一致，借助 MBSD 方法进行终点判断简便易行，可以满足实际生产需求。

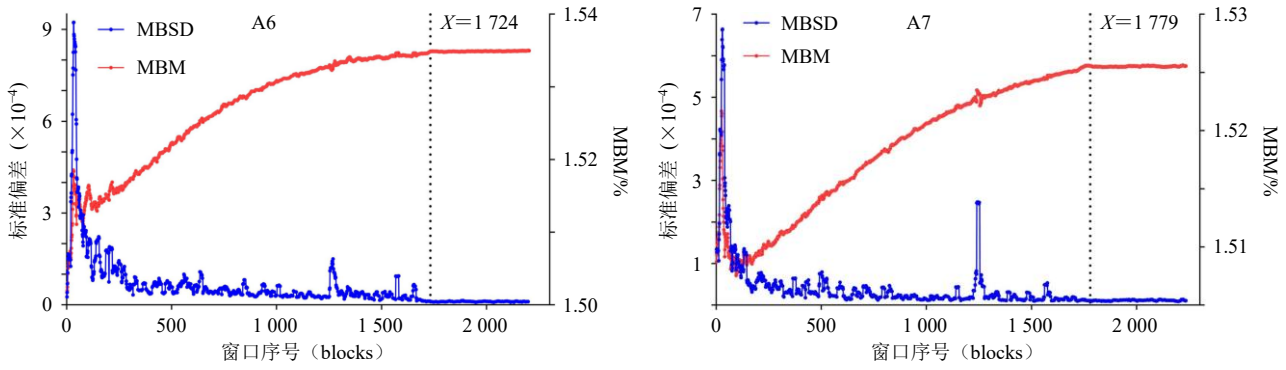

图 10 A6、A7 金青萃取过程 MBSD 及 MBM 趋势图

Fig. 10 MBSD and BBM trends during Jinqing extraction process of A6 and A7

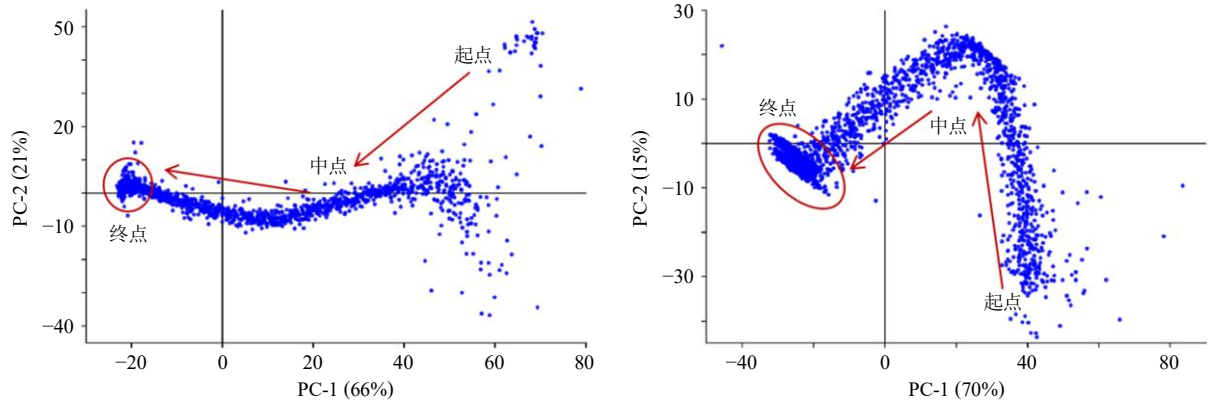

图 11 A6、A7 金青萃取过程 PCA 得分图

Fig. 11 PCA scores of A6 and A7 Jinqing extraction process

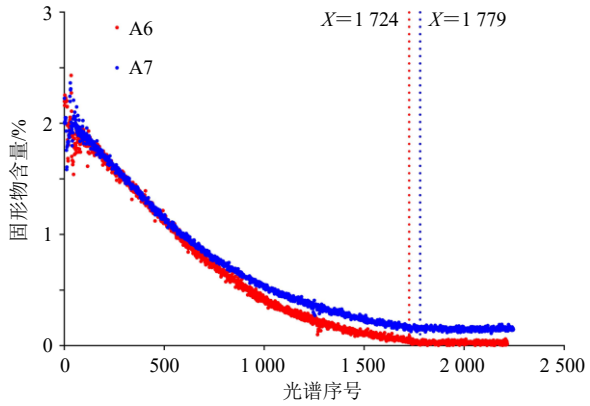

图 12 A6、A7 批次金青萃取过程 SC 预测变化趋势

Fig. 12 Prediction trend of solid content during Jinqing extraction process of A6 and A7

3 讨论

本研究以 RI 金青萃取过程为研究对象，采用 NIRS 技术，结合 PLS 和 MARS 算法，经过光谱预处理方法及建模波段筛选，分别建立了金青萃取过程在线 SC 预测模型，模型性能优异，PLS 和 MARS 模型的 RSEP 分别为 2.87%和 2.64%，均可用于金青萃取过程 SC 的在线监测。为实现 RI 金青萃取过程终点的准确判断，本研究在上一步的基础上，首先采用 SVM 算法建立异常数据清洗模型，对萃取

过程在线 NIRS 采集过程产生的异常光谱进行数据清洗，后续采用 MBSD 和 PCA 算法，建立了萃取终点判别模型，最终结合 MARS 模型对结果进行验证。研究通过 MARS 模型对萃取过程 SC 进行预测，SC 预测结果表明，通过 MBSD 模型预测的萃取终点与实际终点基本一致，MBSD 模型效果真实可靠。

在 RI 金青萃取过程 SC 的在线监测模型建立过程中，采用 MARS 算法构建的定量校正模型展现出了卓越的预测能力。相较于传统的 PLS 算法，MARS 模型的 RSEP 降低了 8.19%，模型预测性能得到提升。MARS 模型的优势在于其优化了 PLS 算法中需要手动筛选关键变量的过程，从而显著提高了模型应用的效率。此外，由于 PLS 模型依赖于潜在变量，这些变量往往难以直接解释，导致模型的解释性不如 MARS 模型直观。MARS 或 PLS 算法的选择，应基于数据特性、分析目标以及对模型解释性的具体要求。MARS 算法特别适用于数据集具有显著非线性特征且需要自动进行变量选择的情境。相反，当数据集中的预测变量数量超过样本数量，或者预测变量之间存在多重共线性时，PLS 算法可能更加适用。因此，在建立 SC 监测模型时，MARS 算法因其较高的预测性能和自动变量筛选能力而成为优选。然而，在实际生产过程中，工艺终点判断的关键在于过程拐点的精确识别，虽然运用先进算法建立的模型相比传统 PLS 回归校正模型能更准确地预测 SC，但是想得到质的提升仍极具挑战性，换言之，考虑采用便捷的定性方法或开发简易定量校正模型可能是一个更为实用的选项。因此，为了做出最佳决策，需要仔细权衡预测精度和成本（包括时间成本和计算成本等），综合考虑选择最适合的算法模型，以满足实际需求。

此外，本研究在构建 SC 在线监测模型过程中，所采纳的数据量相较于生产过程中实时采集的巨量数据而言显得相对有限，因此，存在建模样本代表性不足等潜在问题。为了确保模型的准确性和可靠性，后续研究有必要进一步扩充数据集，对模型进行验证、更新、维护以及优化，从而提升模型的预测精度，为 RI 金青萃取过程的质量控制提供更为坚实的技术支撑。而在萃取过程终点判断的研究中，本研究建立的模型相较于传统的 MBSD 模型，不依赖于固定的阈值和控制限来确定萃取终点，而是通过分析趋势图中平缓变化的区域来灵活判定萃取终点，这种方法不仅提高了判断的准确性，也增强了模型的可靠性和灵活性，为 RI 金青萃取过程的反馈调控奠定了坚实基础。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 余俭. 抗菌抗病毒新药: 热毒宁注射液 [J]. 中南药学, 2010, 8(7): 548-550.

[2] 黄小民, 柳于介, 何煜舟, 等. 热毒宁注射液治疗急性上呼吸道感染的临床研究 [J]. 中国临床药理学与治疗学, 2006, 11(4): 470-473.

[3] 王高举, 焦红军. 基于网络药理学和分子对接技术的热毒宁注射液抗 SARS、MERS 和 COVID-19 的潜在共性作用机制与活性成分研究 [J]. 药物评价研究, 2022, 45(5): 842-852.

[4] Sasić S. Chemical imaging of pharmaceutical granules by Raman global illumination and near-infrared mapping platforms [J]. *Anal Chim Acta*, 2008, 611(1): 73-79.

[5] 唐辉, 陈强, 张学荣. 近红外分析技术在流化床一步制粒工艺中水分控制的应用 [J]. 食品安全质量检测学报, 2020, 11(3): 955-960.

[6] 吴思俊. 基于近红外光谱技术的中药制药工艺终点判断方法研究 [D]. 天津: 天津中医药大学, 2021.

[7] 龙若兰, 冯丹, 罗西, 等. 藏药五脉绿绒蒿提取过程的在线近红外光谱质量控制研究 [J]. 分析测试学报, 2023, 42(08):920-929.

[8] 童枫, 徐芳芳, 张欣, 等. 基于近、中红外光谱的热毒宁注射液制剂过程投料和二次热处理工序快速检测方法研究 [J]. 中草药, 2022, 53(21): 6706-6715.

[9] 杜文俊, 刘雪松, 陶玲艳, 等. 热毒宁注射液金银花和青蒿(金青)醇沉过程中多指标的近红外快速检测 [J]. 中草药, 2015, 46(1): 61-66.

[10] Yang Z F, Xiao H, Zhang L, *et al*. Fast determination of oxide content in cement raw meal using NIR spectroscopy with the SPXY algorithm [J]. *Anal Methods*, 2019, 11(31): 3936-3942.

[11] Friedman J H. Rejoinder: Multivariate adaptive regression splines [J]. *Ann Stat*, 2007, 19(1): 1-67.

[12] 仉文岗, 洪利, 黎泳钦. 基于多元自适应回归样条的高维岩土工程问题分析 [J]. 河海大学学报: 自然科学版, 2019, 47(4): 359-365.

[13] 陈琪, 徐芳芳, 张欣, 等. 基于不同算法对桂枝茯苓胶囊内容物吸湿性预测建模研究 [J]. 中草药, 2021, 52(11): 3216-3223.

[14] 严衍禄, 陈斌, 朱大洲, 等. 近红外光谱分析的原理、技术与应用 [M]. 北京: 中国轻工业出版社, 2013: 162-175.

[15] 候化蕊, 徐芳芳, 王振中, 等. 基于近、中红外光谱技术的热毒宁注射液制备过程中金银花浓缩过程含量预

测研究 [J]. 中草药, 2023, 54(02): 520-533.

[16] 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

[17] Li Y, Wang G Z, Guo G S, *et al*. Spectral pre-processing and multivariate calibration methods for the prediction of wood density in Chinese white poplar by visible and near infrared spectroscopy [J]. *Forests*, 2022, 13(1): 62.

[18] Shao X G, Leung A K M, Chau F T. Wavelet: A new trend in chemistry [J]. *Acc Chem Res*, 2003, 36(4): 276-283.

[19] 王冬, 吴静珠, 韩平, 等. 光谱关键变量筛选在农产品及食品品质无损检测中的应用进展 [J]. 光谱学与光谱分析, 2021, 41(5): 1593-1601.

[20] 王彩虹, 黄林, 刘木华, 等. 基于 SiPLS 模型的稻壳中重金属铬 LIBS 检测 [J]. 激光与光电子学进展, 2016, 53(11): 113001.

[21] 张婷婷, 赵宾, 杨丽明, 等. 基于高光谱成像技术结合 SPA 和 GA 算法测定甜玉米种子电导率 [J]. 光谱学与光谱分析, 2019, 39(8): 2608

[22] 程惠珠, 杨婉琪, 李福生, 等. 面向 XRF 的竞争性自适应重加权算法和粒子群优化的支持向量机定量分析研究 [J]. 光谱学与光谱分析, 2023, 43(12): 3742-3746.

[23] 黄国东, 龙志宏, 朱子朋, 等. 基于支持向量机的供水管网监测数据清洗 [J]. 给水排水, 2022, 48(9): 124-129.

[责任编辑 郑礼胜]