

最小二乘支持向量机和偏最小二乘法在栀子中间体纯化工艺近红外定量分析中的应用

吴莎¹, 刘启安^{2,3}, 李云^{2,3}, 刘茜^{2,3}, 宋亚玲^{2,3}, 吴建雄^{2,3}, 毕宇安^{2,3}, 王振中^{2,3}, 萧伟^{2,3*}

1. 北京中医药大学, 北京 100102

2. 江苏康缘药业股份有限公司, 江苏 连云港 222001

3. 中药制药过程新技术国家重点实验室, 江苏 连云港 222001

摘要: 目的 应用近红外(NIR)光谱技术建立热毒宁注射液栀子中间体纯化工艺关键质控成分的定量分析模型。方法 测定样品中山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷、绿原酸和总酸的量, 扫描 NIR 光谱, 经过异常点的剔除、最佳光谱预处理方法的选择、最佳建模波段的选择, 分别用偏最小二乘法(PLS)和最小二乘支持向量机法(LS-SVM)建立定量校正模型, 并用此模型对 18 个未知样品进行预测。结果 山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷、绿原酸和总酸的 PLS 模型和 LS-SVM 模型对未知样品的预测相对偏差(RSEP)均小于 3%, 说明 2 种方法均产生较好的拟合效果和预测能力。但是山栀苷和总酸的 LS-SVM 模型预测误差明显小于 PLS 模型, 京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸的 LS-SVM 模型和 PLS 模型预测误差接近。结论 LS-SVM 法较 PLS 法预测性能更好, 建立的模型可以用于热毒宁注射液栀子中间体纯化工艺关键质控成分的快速检测。

关键词: 近红外光谱; 偏最小二乘法; 最小二乘支持向量机法; 粒子群算法; 热毒宁注射液

中图分类号: R284.2 **文献标志码:** A **文章编号:** 0253-2670(2015)07-0990-08

DOI: 10.7501/j.issn.0253-2670.2015.07.010

Application of least squares support vector machines and partial least squares in quantitation on purification of *Gardenia jasminoides* intermediate using NIR spectroscopy

WU Sha¹, LIU Qi-an^{2,3}, LI Yun^{2,3}, LIU Qian^{2,3}, SONG Ya-ling^{2,3}, WU Jian-xiong^{2,3}, BI Yu-an^{2,3}, WANG Zhen-zhong^{2,3}, XIAO Wei^{2,3}

1. Beijing University of Chinese Medicine, Beijing 100102, China

2. Jiangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

3. State Key Laboratory of New-tech for Chinese Medicine Pharmaceutical Process, Lianyungang 222001, China

Abstract: Objective To establish the quantitative models for analyzing the content of critical quality indicators in the purification process of *Gardenia jasminoides* intermediate in Reduning Injection using near-infrared (NIR) spectroscopy. **Methods** The contents of shanzhiside, geniposidic acid, deacetyl asperulosidic acid methyl ester, genipin-1- β -D-gentiobioside, geniposide, chlorogenic acid, and total acid were determined by the reference method and NIR spectra were acquired. After removing the outliers, selecting the optimal spectral preprocessing method and selecting the best spectral wavelength, partial least squares (PLS) and the least squares support vector machines (LS-SVM) were used to build the models for predicting the contents of the above quality indicators in 18 unknown samples. **Results** For shanzhiside, geniposidic acid, deacetyl asperulosidic acid methyl ester, genipin-1- β -D-gentiobioside, geniposide, chlorogenic acid, and total acid, the relative standard errors of prediction (RSEP) was lower than 3% for PLS models and LS-SVM models, indicating both methods could exhibit the satisfactory fitting results and predictive abilities. However, the LS-SVM models of shanzhiside and total acid showed lower predictive errors than PLS models. For geniposidic acid, deacetyl asperulosidic acid methyl ester, genipin-1- β -D-gentiobioside, geniposide, and chlorogenic acid, both models have the closer predictive errors.

收稿日期: 2014-11-17

基金项目: 科技部重大新药创制: 现代中药创新集群与数字制药技术平台(2013ZX09402203)

作者简介: 吴莎(1988—), 女, 博士研究生, 研究方向为中药质量控制及过程分析。Tel: 13601215863 E-mail: wusha729@163.com

*通信作者 萧伟, 研究员级高级工程师, 博士, 研究方向为中药新药的研究与开发。Tel: (0518)81152367 E-mail: kanionlunwen@163.net

Conclusion S-SVM shows better predictive performance than PLS. The established NIR quantitative models can be used for rapidly measuring the content of critical quality indicators in the purification process of *G. jasminoides* intermediate in Reduning Injection.

Key words: near-infrared spectroscopy; partial least squares; least squares support vector machines; particle swarm optimization; Reduning Injection

热毒宁注射液处方由青蒿、金银花、栀子3味中药组成,临床上用于治疗上呼吸道感染^[1]。方中栀子为茜草科植物栀子 *Gardenia jasminoides* Ellis 的干燥成熟果实,主要含有环烯醚萜类、有机酸类和色素类成分^[2],具有抗炎、镇痛、抗病毒等药理作用^[3-4]。从2005年上市以来,热毒宁注射液因抗病毒、抗菌退热疗效显著,作用迅速,得到广大医生和患者的欢迎。继2010年热毒宁注射液被批准为甲型H1N1流感^[5]和手足口病临床用药,2014年10月热毒宁注射液又被批准为登革热诊疗用药。

栀子中间体纯化是热毒宁注射液中间体生产过程中重要工艺单元之一,它在富集药效成分的同时除去脂溶性杂质,但是纯化过程容易受到加热、搅拌、pH等工艺参数的干扰而影响药效成分量。近红外(near infrared, NIR)光谱技术凭借其样品处理简单、分析快速、无需消耗试剂的特点^[6],是目前发展最快的过程分析技术之一,被越来越多地应用于中药研究领域,包括药材产地鉴别^[7-8]、有效成分定量测定^[9-10]和制药过程的在线监控^[11-12]。因此考虑将NIR光谱分析技术引入栀子中间体纯化工艺中,实现关键质控成分的实时监测。

本研究以热毒宁注射液栀子中间体纯化工艺为研究对象,选择山栀苷(shanzhiside, SZS)、京尼平苷酸(geniposidic acid, GA)、京尼平龙胆双糖苷(genipin-1- β -D-gentiobioside, GG)、栀子苷(geniposide, GS)、去乙酰车叶草酸甲酯(deacetyl asperulosidic acid methyl ester, DAAME)、绿原酸(chlorogenic acid, CA)和总酸(total acid, TA)为关键质控成分,分别运用偏最小二乘法(partial least squares, PLS)和最小二乘支持向量机法(least squares support vector machines, LS-SVM)建立NIR定量分析模型,快速检测关键质控成分的量,为在线质量控制提供可行性实验依据与技术支持。

1 仪器与材料

Luminar 5030型AOTF近红外光谱分析仪,美国Brimrose公司;配有InGaAs检测器,SNAP光谱信号采集软件、Unscrambler 7.8数据处理软件,挪威CAMO软件公司;Matlab R2012a数据处理软

件,美国MathWorks软件公司;Agilent 1290超高效液相色谱仪,美国安捷伦科技有限公司,配有DAD检测器、四元梯度泵、在线脱气装置、自动进样器;Mettler Toledo XP6电子天平,梅特勒-托利多仪器上海有限公司;H1650-W台式高速离心机,湖南湘仪实验室仪器开发有限公司;Milli-Q Academic纯水机,美国Millipore公司。

对照品去乙酰车叶草酸甲酯(批号111786-200801)和栀子苷(批号110749-201115)均购自中国食品药品检定研究院;对照品山栀苷(批号BBP01688)购自云南西力生物技术有限公司;对照品绿原酸(批号MUST-13031401)和京尼平苷酸(批号MUST-13121502)均购自成都曼思特生物科技有限公司;对照品京尼平龙胆双糖苷(批号131120)购自成都普菲德生物技术有限公司。上述对照品质量分数均为98%以上。乙腈,色谱纯,Tedia公司;水为超纯水;栀子中间体纯化溶液来自江苏康缘药业股份有限公司。

2 方法

2.1 样品收集及划分

从栀子中间体纯化工艺开始至结束,每隔5分钟取50 mL溶液作为1份样品,每批取样16个,共取样11批,批号为Z140305、Z140306、Z140307、Z140308、Z140309、Z140310、Z140401、Z140402、Z140405、Z140406、Z140408。在Matlab软件上采用Kennard-stone样本划分方法,按照9:1的比例,将获得的176个样本分为训练集(158个样本)和预测集(18个样本)。

2.2 NIR光谱采集

在室温条件下,采集样品的NIR投射光谱。光谱扫描范围1100~2300 nm,波长增量2 nm,扫描次数300次,分辨率1 nm。每个样品采集3张光谱,取平均光谱进行分析。样品的原始光谱、一阶导数光谱和二阶导数光谱见图1。

2.3 UPLC测定

2.3.1 色谱条件 色谱柱为Agilent Zorbax SB-C₁₈(100 mm×3.0 mm, 1.8 μ m);预柱Agilent UPLC Guard Zorbax SB-C₁₈(5 mm×3.0 mm, 1.8 μ m);流

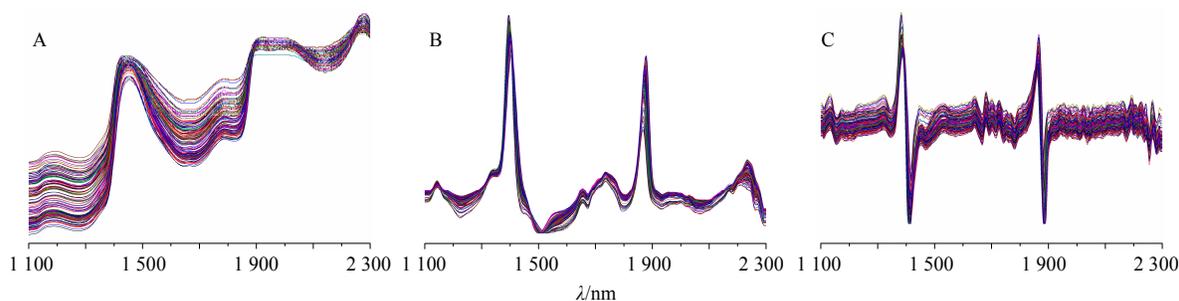


图1 NIR原始光谱(A)、一阶导数光谱(B)和二阶导数光谱(C)

Fig. 1 NIR of original spectra (A), spectra obtained with first derivative (B), and spectra obtained with second derivative (C)

动相为乙腈-0.1%磷酸水溶液, 梯度洗脱: 0~10 min, 5%~10%乙腈; 10~15 min, 10%~20%乙腈; 15~30 min, 20%~30%乙腈; 体积流量 0.4 mL/min; 进样量 2 μ L; 柱温 30 $^{\circ}$ C; 238 nm 下测定山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷和栀子苷, 327 nm 下测定绿原酸。

2.3.2 对照品储备液的制备 分别精密称取山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸对照品适量, 加 50% 甲醇制成质量浓度分别为 28.08、15.14、61.54、113.68、770.20、36.40 mg/L 的混合对照品储备液。

2.3.3 供试品溶液的制备 精密称定 50 mL 量瓶, 加入 1 mL 栀子中间体纯化溶液, 精密称定, 用 50% 甲醇稀释至刻度, 摇匀, 离心, 精密吸取 1 mL 上清液, 置 10 mL 量瓶中, 用 50% 甲醇稀释至刻度, 摇匀, 离心, 上清液过 0.22 μ m 微孔滤膜, 即得。

2.3.4 线性关系考察 分别精密量取不同体积的混合对照品储备液, 将其配成不同质量浓度的对照品溶液, 按上述色谱条件进样分析, 以峰面积积分值 (Y) 对质量浓度 (X) 进行线性回归分析, 得到各对照品的回归方程和线性范围: 山栀苷 $Y=5.512 X-1.07$, $R^2=0.999 7$, 线性范围 4.21~42.12 μ g/mL; 京尼平苷酸 $Y=6.144 X-0.90$, $R^2=0.999 7$, 线性范围 2.27~22.71 μ g/mL; 去乙酰车叶草酸甲酯 $Y=6.738 X-2.03$, $R^2=0.999 9$, 线性范围 9.23~92.31 μ g/mL; 京尼平龙胆双糖苷 $Y=3.631 X-0.73$, $R^2=0.999 9$, 线性范围 17.05~170.52 μ g/mL; 栀子苷 $Y=6.833 X-6.05$, $R^2=0.999 9$, 线性范围 115.53~1 155.30 μ g/mL; 绿原酸 $Y=13.736 X-6.95$, $R^2=0.999 8$, 线性范围 5.46~54.60 μ g/mL。

2.3.5 精密度考察 按“2.3.3”项下方法制备供试品溶液, 连续进样 6 次, 按“2.3.1”项下色谱条件测定, 记录峰面积, 结果山栀苷、京尼平苷酸、去

乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸峰面积的 RSD 分别为 0.37%、1.09%、0.32%、0.55%、0.18%、0.27%, 说明仪器的精密度良好。

2.3.6 重复性考察 按“2.3.3”项下方法重复制备 6 份供试品溶液, 按“2.3.1”项下色谱条件进样测定, 结果山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸平均质量分数分别为 8.12、2.55、12.08、22.30、145.91、6.07 mg/g, RSD 分别为 1.17%、1.36%、0.63%、1.08%、0.72%、0.92%, 说明该方法的重复性良好。

2.3.7 稳定性考察 按“2.3.3”项下方法制备供试品溶液, 室温下放置, 分别在不同时间点 (0、2、4、8、16、24 h) 进样, 按“2.3.1”项下色谱条件测定, 记录峰面积, 结果山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸峰面积的 RSD 分别为 1.08%、1.27%、0.68%、0.79%、0.42%、0.70%, 说明样品溶液中各成分在室温下 24 h 内稳定。

2.3.8 加样回收率考察 精密称定 50 mL 量瓶, 加入 0.5 mL 栀子中间体纯化溶液, 精密称定, 分别精密加入适量山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷、绿原酸对照品, 用 50% 甲醇定容至刻度, 摇匀, 离心, 精密吸取 1 mL 上清液, 置 10 mL 量瓶中, 用 50% 甲醇稀释至刻度, 摇匀, 离心, 上清液过 0.22 μ m 微孔滤膜滤过。同法制备 6 份, 按“2.3.1”项下色谱条件进样测定, 计算加样回收率, 并计算 RSD。结果平均回收率分别为 100.32%、102.86%、98.76%、100.92%、100.72%、99.04%, RSD 分别为 1.31%、1.87%、1.66%、1.51%、1.12%、1.01%。

2.4 总酸测定

2.4.1 供试品溶液的制备 精密称定 50 mL 量瓶, 加入 1 mL 栀子中间体纯化溶液, 精密称定, 用 50%

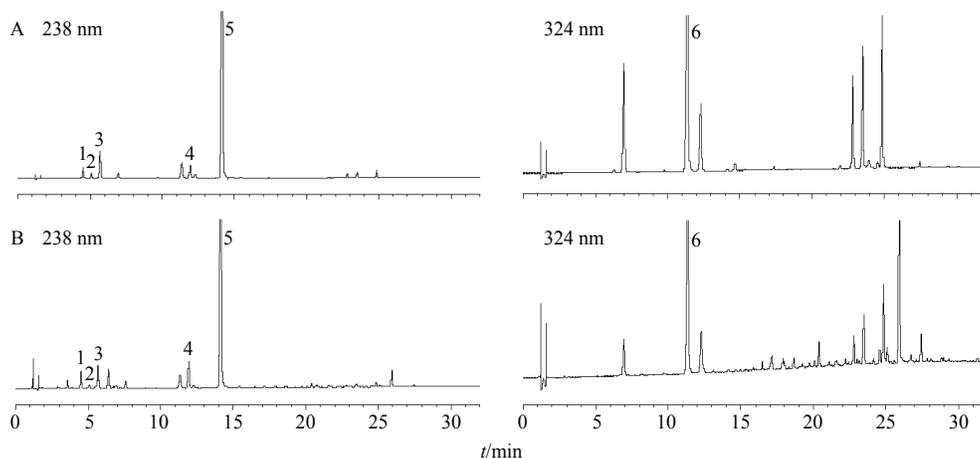
甲醇稀释至刻度，摇匀，离心，精密吸取 1 mL 上清液，置 10 mL 量瓶中，用 50% 甲醇稀释至刻度，摇匀，离心，再精密吸取 2 mL 上清液，置 10 mL 量瓶中，用 50% 甲醇稀释至刻度，摇匀，离心，即得。

2.4.2 标准曲线的绘制 精密称取绿原酸对照品 5.030 mg 置 50 mL 量瓶中，加 50% 甲醇定容至刻度，摇匀，得到含绿原酸 100.6 mg/L 的对照品储备液。精密吸取储备液各 1、2、3、4、5 mL，分别置于 20 mL 量瓶中，加 50% 甲醇稀释至刻度，摇匀，以

50% 甲醇为空白，在 324 nm 波长处测定吸光度 (A) 为 A_1 ，在 400 nm 波长下测定 A 为 A_2 ，以 $\Delta A = (A_1 - A_2)$ 为纵坐标 (Y)，以质量浓度 (C) 为横坐标 (X)，绘制标准曲线 $Y = 51.25058X - 0.00012$ ， $r = 0.9999$ 。

2.5 UPLC 色谱图

根据上述色谱条件和供试品的制备方法，检测样品溶液中山栀苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸的量。混合对照品溶液和样品溶液的 UPLC 图见图 2。



1-山栀苷 2-京尼平苷酸 3-去乙酰车叶草酸甲酯 4-京尼平龙胆双糖苷 5-栀子苷 6-绿原酸
1-shanzhiside 2-geniposidic acid 3-deacetyl asperulosidic acid methyl ester 4-genipin-1-β-D-gentiobioside 5-geniposide 6-chlorogenic acid

图 2 混合对照品溶液 (A) 和样品溶液 (B) 的 UPLC 图

Fig. 2 UPLC of mixed reference solutions (A) and sample solutions (B)

2.6 模型建立及模型性能评价

分别应用 PLS 和 LS-SVM 方法建立光谱数据与质控指标间的定量模型。以相关系数 (correlation coefficient, R)、校正均方根误差 (root mean square errors of calibration, RMSEC)、交叉验证均方根误差 (root mean square errors of cross-validation, RMSECV) 为指标优化建模参数，以 R 、预测均方根误差 (root mean square error of prediction, RMSEP) 和预测相对偏差 (relative standard error of prediction, RSEP) 为指标考察模型预测准确性。

3 结果与讨论

3.1 异常点的剔除

NIR 在采集的过程中，可能由于环境温湿度的变化或样品粒度、黏度等物理性状的变化，从而引起样品吸光度或特征峰强度异常^[13]。因此在建立模型之前，需要先剔除异常点。异常点根据杠杆值和残差的大小进行判断，杠杆值越大，残差越小，说明数据对模型贡献越大，反之，数据对模型贡献较小^[14]。以栀子苷为例，见图 3，样本 36、60、83

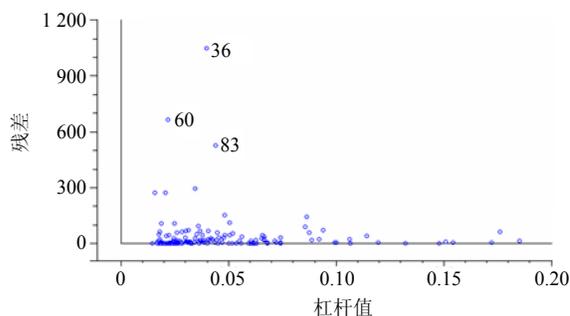


图 3 栀子苷模型中异常点

Fig. 3 Outliers in geniposide calibration model

残差很大，但杠杆值较小，说明它们为异常点，需要从校正模型中剔除。同理，剔除山栀苷模型的异常点 36、60、83，京尼平苷酸模型的异常点 14、104，去乙酰车叶草酸甲酯模型的异常点 4、14，京尼平龙胆双糖苷模型的异常点 36、83，绿原酸模型的异常点 63，总酸模型的异常点 109。

3.2 光谱预处理方法的选择

NIR 容易受到溶液的均匀性、固体颗粒杂质等因素的干扰，导致基线偏移或漂移^[15]。对光谱进行

适当的预处理, 可以消除干扰, 更好地提取特征光谱信息。从图 1 中可以看出, 光谱经过二阶导数处理后, 噪音信号明显放大, 所以分别使用原始光谱和一阶导数处理后的光谱在 1 100~2 300 nm 建模比较, 结果见表 1, 发现山梔苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、梔子苷、绿原酸和总酸光谱经过一阶导数结合 Savitzky-Golay 平滑 (S-G) 和标准正则变换 (standard normal variate, SNV) 处理后, 所建模型 RMSEC 和 RMSECV 更低, 建模效果更好。

表 1 不同光谱预处理方法对模型性能的影响

Table 1 Effect of different spectral pretreatments on calibration model performance

化合物	预处理方法	R	RMSEC	RMSECV
山梔苷	原始光谱	0.925 8	0.264 2	0.286 3
	一阶导数+S-G+SNV	0.956 1	0.204 9	0.237 8
京尼平苷酸	原始光谱	0.963 7	0.145 4	0.179 6
	一阶导数+S-G+SNV	0.969 3	0.128 2	0.156 4
去乙酰车叶	原始光谱	0.913 2	0.857 6	1.028 0
	草酸甲酯	0.935 0	0.840 4	0.996 3
京尼平龙胆	原始光谱	0.911 7	0.875 7	0.936 7
	双糖苷	0.965 7	0.616 8	0.662 0
梔子苷	原始光谱	0.933 7	4.854 5	5.263 9
	一阶导数+S-G+SNV	0.967 9	3.144 6	3.604 6
绿原酸	原始光谱	0.924 0	0.281 8	0.326 7
	一阶导数+S-G+SNV	0.941 0	0.252 6	0.283 1
总酸	原始光谱	0.937 9	1.307 0	1.385 9
	一阶导数+S-G+SNV	0.963 9	0.904 9	0.962 6

3.3 建模波段的选择

研究比较了不同的光谱波段对建模效果的影响, R 越高, RMSEC 与 RMSECV 越低说明所建模型越好, 结果如表 2 所示, 山梔苷、京尼平苷酸、京尼平龙胆双糖苷、梔子苷和总酸的最佳建模波段是 1 100~2 300 nm, 去乙酰车叶草酸甲酯的最佳建模波段是 1 500~1 850 nm, 绿原酸的最佳建模波段是 1 100~1 850 nm。

3.4 定量模型的建立

根据选定的最佳光谱预处理方法和最佳建模波段, 应用留一交叉验证法建立各成分的 PLS 和 LS-SVM 定量校正模型。对于 PLS 模型, 应用主成分分析 (principle component analysis, PCA) 选择山梔苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平

表 2 不同波段对模型性能的影响

Table 2 Effect of different wavelengths on calibration model performance

化合物	波段/nm	R	RMSEC	RMSECV	
山梔苷	1 100~2 300	0.956 1	0.204 9	0.237 8	
	1 100~1 400	0.889 2	0.314 9	0.351 2	
	1 500~1 850	0.925 7	0.265 5	0.312 6	
	1 100~1 850	0.935 6	0.246 4	0.284 5	
	1 900~2 300	0.881 7	0.314 3	0.373 7	
京尼平苷酸	1 100~2 300	0.969 3	0.128 2	0.156 4	
	1 100~1 400	0.948 2	0.172 2	0.185 6	
	1 500~1 850	0.954 6	0.160 2	0.188 6	
	1 100~1 850	0.968 0	0.135 1	0.164 1	
	1 900~2 300	0.935 0	0.187 8	0.231 0	
去乙酰车叶	1 100~2 300	0.935 0	0.840 4	0.996 3	
	草酸甲酯	1 100~1 400	0.915 4	0.943 2	1.097 4
		1 500~1 850	0.944 1	0.802 0	0.924 0
京尼平龙胆	1 100~1 850	0.916 6	0.923 3	1.092 2	
	1 900~2 300	0.892 3	0.984 5	1.162 7	
	双糖苷	1 100~2 300	0.965 7	0.616 8	0.662 0
1 100~1 400		0.915 6	0.931 9	1.023 0	
总酸	1 500~1 850	0.932 4	0.827 5	0.885 6	
	1 100~1 850	0.949 7	0.738 3	0.798 9	
	1 900~2 300	0.881 5	0.944 6	1.035 6	
梔子苷	1 100~2 300	0.967 9	3.144 6	3.604 6	
	1 100~1 400	0.925 5	4.989 1	5.287 2	
	1 500~1 850	0.916 5	5.620 8	6.121 6	
绿原酸	1 100~1 850	0.951 0	4.140 9	4.713 1	
	1 900~2 300	0.893 7	6.534 7	6.971 1	
	1 100~2 300	0.941 0	0.252 6	0.283 1	
	1 100~1 400	0.902 2	0.306 4	0.334 6	
	1 500~1 850	0.914 5	0.284 0	0.310 4	
总酸	1 100~1 850	0.947 1	0.239 3	0.271 4	
	1 900~2 300	0.883 6	0.335 0	0.404 7	
	1 100~2 300	0.963 9	0.904 9	0.962 6	
	1 100~1 400	0.924 9	1.422 4	1.473 0	
	1 500~1 850	0.941 6	1.270 6	1.379 4	
京尼平龙胆	1 100~1 850	0.954 9	1.058 1	1.118 1	
	1 900~2 300	0.931 0	1.297 5	1.435 7	

龙胆双糖苷、梔子苷、绿原酸和总酸的最佳主成分数分别是 6、8、8、7、6、6、5。对于 LS-SVM 模型, 应用径向基核函数 (radial basis function, RBF) 作为训练函数。RBF 中最重要的参数为回归误差权

重 γ 和核参数 σ^2 ，它们在很大程度上决定算法的学习和泛化能力^[16]，因此，采用粒子群算法（particle swarm optimization, PSO）对 γ 和 σ^2 进行全局寻优^[12]。LS-SVM 和 PSO 算法均在 Matlab 软件平台上自主编制。两参数 γ 和 σ^2 的搜索范围分别设为 $10^{-1} \sim 10^3$ 和 $10^{-2} \sim 10^2$ ，初始化粒子数设为 30，PSO 最大循环迭代数为 100。结果各成分用 PSO 算法搜索到的最优参数分别为山栀苷： $\gamma=52.8$ ， $\sigma^2=3.6$ ；京尼平苷酸： $\gamma=538.4$ ， $\sigma^2=89.6$ ；去乙酰车叶草酸甲酯： $\gamma=366.9$ ， $\sigma^2=73.9$ ；京尼平龙胆双糖苷： $\gamma=244.2$ ， $\sigma^2=29.6$ ；栀子苷： $\gamma=853.1$ ， $\sigma^2=44.2$ ；绿原酸： $\gamma=211.6$ ， $\sigma^2=78.6$ ；总酸： $\gamma=457.4$ ， $\sigma^2=61.8$ 。各成分的 PLS 和 LS-SVM 模型参数见表 3，从中可以看出，LS-SVM 模型 R 更高，RMSEC 和 RMSECV 更低，拟合效果更好。

表 3 校正集和交叉验证集模型参数

Table 3 Model parameters of calibration and cross-validation sets

化合物	建模方法	校正集		交叉验证集	
		R	RMSEC	R	RMSECV
山栀苷	PLS	0.956 1	0.204 9	0.950 3	0.227 8
	LS-SVM	0.987 3	0.124 2	0.975 5	0.163 9
京尼平苷酸	PLS	0.969 3	0.128 2	0.958 2	0.146 4
	LS-SVM	0.977 9	0.107 1	0.962 1	0.123 8
去乙酰车叶草酸甲酯	PLS	0.944 1	0.802 0	0.929 6	0.972 2
	LS-SVM	0.964 3	0.732 0	0.950 7	0.924 8
京尼平龙胆双糖苷	PLS	0.965 7	0.616 8	0.960 9	0.662 0
	LS-SVM	0.981 0	0.530 3	0.971 4	0.601 8
栀子苷	PLS	0.967 9	3.144 6	0.958 8	3.404 6
	LS-SVM	0.982 6	2.918 5	0.973 5	3.229 6
绿原酸	PLS	0.977 1	0.239 3	0.959 0	0.277 6
	LS-SVM	0.984 3	0.213 0	0.969 1	0.262 8
总酸	PLS	0.963 9	0.904 9	0.956 9	0.962 6
	LS-SVM	0.979 7	0.805 3	0.967 1	0.885 2

3.5 模型预测

用所建立的模型预测预测集 18 个样本中各成分的量。以 R 、RMSEP 和 RSEP 为指标评价模型预测性能，当 RSEP 小于 8%，可以认为所建模型预测准确性能够满足质量控制要求^[17]。从表 4 中可以看出，所有成分 RSEP 值皆小于 3%，说明 PLS 和 LS-SVM 2 种模型预测准确性均较好，但是 LS-SVM 模型较 PLS 模型 RMSEP 和 RSEP 值更低，预测误

表 4 预测集模型参数

Table 4 Model parameters of prediction set

化合物	建模方法	R	RMSEP	RSEP/%
山栀苷	PLS	0.948 1	0.579 9	1.86
	LS-SVM	0.961 7	0.419 2	1.35
京尼平苷酸	PLS	0.946 0	0.159 2	2.22
	LS-SVM	0.957 3	0.135 6	1.95
去乙酰车叶草酸甲酯	PLS	0.923 0	1.458 1	2.62
	LS-SVM	0.947 9	1.213 8	2.43
京尼平龙胆双糖苷	PLS	0.956 3	1.490 1	1.62
	LS-SVM	0.965 9	1.348 5	1.47
栀子苷	PLS	0.953 1	7.079 5	1.25
	LS-SVM	0.964 0	6.359 1	1.12
绿原酸	PLS	0.950 7	0.423 7	1.72
	LS-SVM	0.962 6	0.383 1	1.56
总酸	PLS	0.954 3	2.729 1	2.07
	LS-SVM	0.958 9	2.176 3	1.55

差更小。对 PLS 和 LS-SVM 模型各成分的预测值进行配对样本检验。具体方法为计算成对数据的差值 (d)，对 d 进行正态性检验，如果 d 服从正态分布，则使用 2 组配对样本 t 检验，如果 d 不服从正态分布，则使用非参数检验中的 Wilcoxon 秩和检验，结果山栀苷、京尼平苷酸和栀子苷 Wilcoxon 秩和检验的 P 值分别为 0.004、0.277 和 0.199；去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、绿原酸和总酸 t 检验的 P 值分别为 0.446、0.372、0.074 和 0.036，说明山栀苷和总酸 2 种模型的预测值间差异有统计学意义。从图 4 中可以看出，PLS 和 LS-SVM 2 种模型对山栀苷和总酸的量预测差异比较显著，LS-SVM 模型预测值与 UPLC 测定值更加接近，这与表 4 中山栀苷和总酸 LS-SVM 模型的 RSEP 值明显低于 PLS 的 RSEP 值相吻合。

4 结论

NIR 分析常用的建模方法有偏最小二乘法、人工神经网络法 (artificial neural networks, ANN) 和支持向量机法 (support vector machines, SVM) 等。LS-SVM 是支持向量机方法的扩展，能够进行线性和非线性的多元建模，它相比 SVM 提高了运算速度和收敛精度^[18]，它也可以从全局最优角度出发，有效地避免 ANN 算法容易陷入局部最优的问题^[19]。鉴于中药成分的复杂性，栀子中间体纯化溶液的 NIR 信息和质控指标的量信息间可能包含复杂的非线性关系，本实验考虑分别用 PLS 和 LS-SVM 2 种

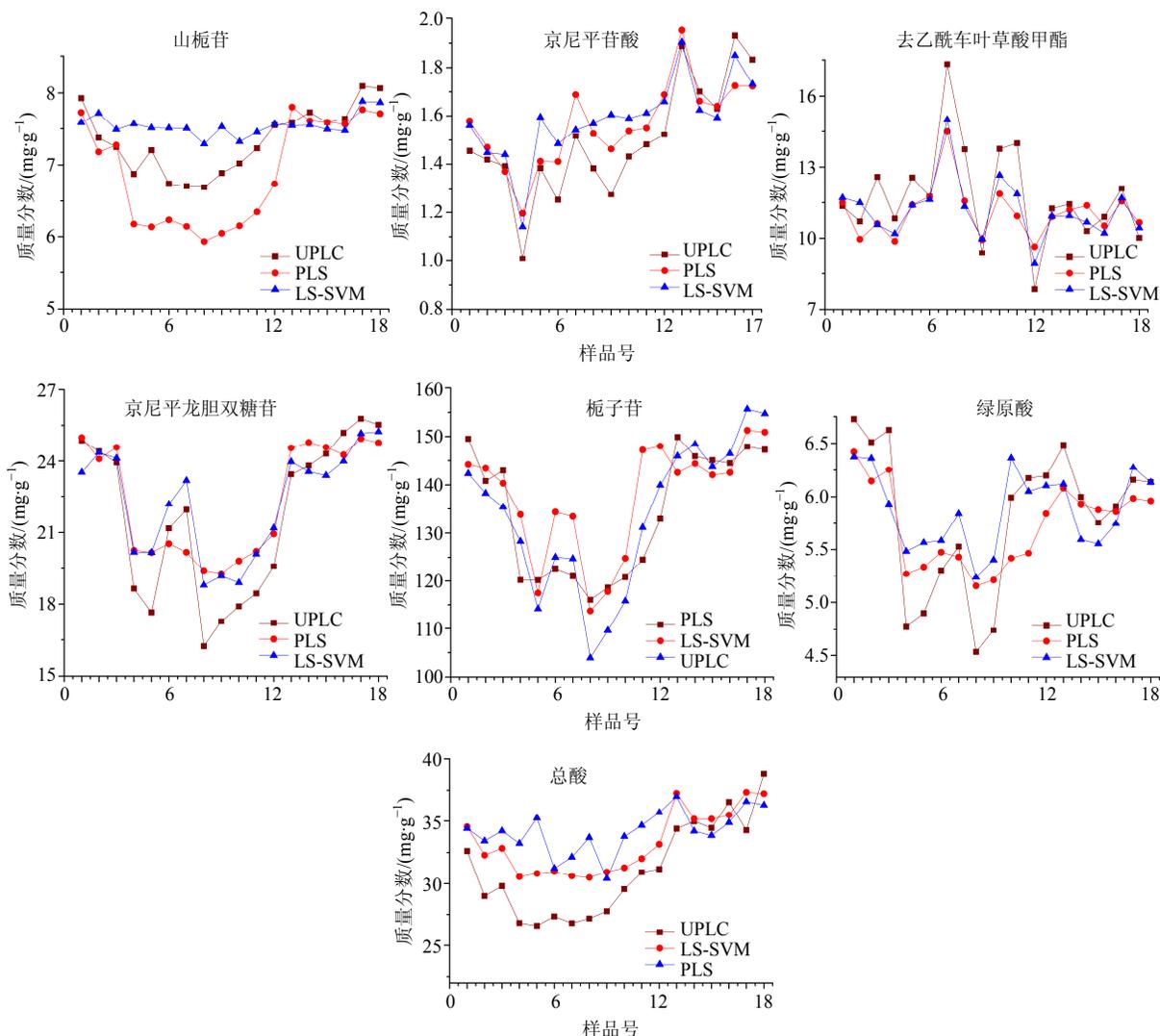


图 4 预测集样品各成分 UPLC 实测值、PLS 预测值和 LS-SVM 预测值相关性

Fig. 4 Correlation among UPLC-measured value, PLS-predicted value, and LS-SVM-predicted value of each sample in prediction set

方法构建关键质控成分的定量分析模型。

从结果来看, 山梔苷、京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷、绿原酸和总酸的 PLS 模型和 LS-SVM 模型预测集 RSEP 均小于 3%, 预测性能较好, 但是山梔苷和总酸的 LS-SVM 模型 RMSEP 和 RSEP 值明显低于 PLS 模型。对于京尼平苷酸、去乙酰车叶草酸甲酯、京尼平龙胆双糖苷、栀子苷和绿原酸, 这 2 种模型 RMSEP 和 RSEP 值差异不大, LS-SVM 模型稍优。整体来看, LS-SVM 模型对上述 7 个成分的预测性能更好。

本研究建立的栀子中间体纯化工艺关键质控成分的 NIR 定量分析模型, 操作简单, 预测误差较小, 有望在此基础上扩大校正集和预测集样本容量, 建立此过程 NIR 在线监控模型, 实现关键质控成分的

量的实时监测, 并结合自动化控制技术, 从根本上保障热毒宁注射液产品质量的安全、有效、均一、稳定。

参考文献

- [1] 罗佩施, 吴礼武. 热毒宁治疗急性上呼吸道感染伴发热的疗效观察 [J]. 临床合理用药, 2010, 16(2): 69-70.
- [2] 付小梅, 俞桂新, 王峰涛. 栀子的化学成分 [J]. 中国天然药物, 2008, 6(6): 418-420.
- [3] 方尚玲, 刘源才, 张庆华, 等. 栀子苷镇痛和抗炎作用的研究 [J]. 时珍国医国药, 2008, 19(6): 1374-1376.
- [4] 曹泽或, 常秀娟, 赵忠鹏, 等. 热毒宁注射液抗 A16 型柯萨奇病毒的研究 [J]. 中草药, 2014, 45(10): 1450-1455.
- [5] 王振中, 鲍琳琳, 孙 兰, 等. 热毒宁注射液抗甲型 H1N1 流感病毒作用机制研究 [J]. 中草药, 2014, 45(1):

- 90-93.
- [6] 杨海龙, 臧恒昌, 胡甜, 等. 近红外漫反射光谱法对不同产地山楂的定性鉴别和定量分析 [J]. 药物分析杂志, 2014, 34(3): 396-401.
- [7] 杜敏, 巩颖, 林兆洲, 等. 样品表面近红外光谱结合多类支持向量机快速鉴别枸杞子产地 [J]. 光谱学与光谱分析, 2013, 33(5): 1211-1214.
- [8] 孙丽英, 杨天鸣, 王云英, 等. 不同产地黄柏的近红外指纹图谱鉴别分析 [J]. 计算机与应用化学, 2008, 25(3): 329-332.
- [9] 雷敬卫, 樊明月, 郭艳利, 等. 近红外光谱法快速测定木香中木香烯内酯与去氢木香内酯的含量 [J]. 天然药物研究与开发, 2014, 26(7): 1062-1066.
- [10] 李伟, 孙素琴, 覃洁萍, 等. 近红外漫反射法测定杜仲中松脂醇二葡萄糖苷的含量 [J]. 中国中药杂志, 2010, 35(24): 3318-3321.
- [11] 魏惠珍, 张五萍, 毛红梅, 等. 近红外光谱法在白芍提取物纯化过程中快速质量控制研究 [J]. 中草药, 2013, 44(9): 1128-1133.
- [12] 金叶, 杨凯, 吴永江, 等. 基于粒子群算法的最小二乘支持向量机在红花提取液近红外定量分析中的应用 [J]. 分析化学, 2012, 40(6): 925-931.
- [13] 闵顺耕, 李宁, 张明祥. 近红外光谱分析中异常值的判别与定量模型优化 [J]. 光谱学与光谱分析, 2004, 24(10): 1205-1209.
- [14] Wu Z, Xu B, Du M, *et al.* Validation of a NIR quantification method for the determination of chlorogenic acid in *Lonicera japonica* solution in ethanol precipitation process [J]. *J Pharm Biomed Anal*, 2012, 62(25): 1-6.
- [15] 袁洪福. 在线近红外光谱分析技术及其应用 [A] // 第三届中国在线分析仪器应用及发展国际论坛暨展览会论文 [C]. 北京: 中国仪器仪表学会, 2010.
- [16] 徐冰, 王星, Dhaene Tom, 等. 基于遗传算法的多目标最小二乘支持向量机在近红外多组分定量分析中的应用 [J]. 光谱学与光谱分析, 2014, 34(3): 638-642.
- [17] Wu Y J, Jin Y, Li Y R, *et al.* NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of an extraction process [J]. *Vib Spectrosc*, 2012, 58(1): 109-118.
- [18] Pochet N L M M, Suykens J A K. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making [J]. *Ultrasound Obstet Gynecol*, 2006, 27(6): 607-608.
- [19] 林美成, 李亚安. 基于 ANN 与 SVM 的分类和回归比较研究 [J]. 声学技术, 2008, 27(4): 226-230.