近红外光谱结合不同变量筛选方法用于金银花提取过程中绿原酸量的在线监测

杜晨朝[1]，赵安邦[2]，吴志生[1*]，乔延江[1*]

1. 北京中医药大学中药学院 国家中医药管理局中药信息工程重点研究室，北京 100102
2. 新疆医科大学中医学院，新疆 乌鲁木齐 830011

摘 要：目的 采用近红外光谱技术，结合不同变量筛选方法对金银花提取过程中绿原酸量进行快速测定。**方法** 采用组合间隔偏最小二乘法（SIPLS）、竞争自适应抽样方法（CARS）、变量投影重要性（VIP）、连续投影算法（SPA）4 种不同变量筛选方法，以 HPLC 测定值作参比，建立金银花中绿原酸定量模型并进行比较，优选出最佳变量筛选方法。**结果** 经 SIPLS 方法所建绿原酸模型预测能力最好，预测集决定系数（R_{pre}^2）和预测均方根误差（RMSEP）分别为 0.990 3 和 2.316%。**结论** 近红外光谱法结合 SIPLS 变量筛选方法建立的绿原酸定量模型性能良好，满足中药提取过程实时监测分析的精度要求，可用于中药提取过程的快速分析。

关键词：近红外光谱；变量筛选；金银花；绿原酸；在线监测；组合间隔偏最小二乘法；竞争自适应抽样方法；变量投影重要性；连续投影算法

中图分类号：R286.02 **文献标志码：**A **文章编号：**0253 - 2670(2017)16 - 3317 - 05

DOI: 10.7501/j.issn.0253-2670.2017.16.09

Online control of chlorogenic acid in *Lonicerae Japonicae Flos* by near infrared spectroscopy combined with different variable selections

DU Chen-zhao[1], ZHAO An-bang[2], WU Zhi-sheng[1], QIAO Yan-jiang[1]

1. Key Laboratory of Chinese Medicine Information Engineering, State Administration of Traditional Chinese Medicine, School of Chinese Material Medica, Beijing University of Chinese Medicine, Beijing 100102, China
2. College of traditional Chinese Medicine, Xinjiang Medical University, Urumqi 830011, China

Abstract: Objective To determine the content of chlorogenic acid in *Lonicerae Japonicae Flos* by the combined near-infrared and variable selection methods. **Methods** Synergy interval partial least squares (SIPLS), competitive adaptive reweighted sampling method (CARS), variable importance in projection (VIP), and successive projections algorithm (SPA) were used to build a chlorogenic acid quantitative model in *Lonicerae Japonicae Flos* and compare. High performance liquid chromatography (HPLC) was used as a reference to select the optimum variable screening method. **Results** Study results showed that SIPLS was the most desirable method for chlorogenic acid in regression performance with R_{pre}^2 at 0.990 3 and RMSEP at 2.316%. **Conclusion** The quantitative model of chlorogenic acid established by NIR combined with SIPLS has good performance and meets the requirement of real-time analysis of traditional Chinese medicine production process.

Key words: near infrared spectroscopy; variable selection; *Lonicerae Japonicae Flos*; chlorogenic acid; online control; synergy interval partial least squares; competitive adaptive reweighted sampling method; variable importance in projection; successive projections algorithm

目前，大量研究已经证明近红外光谱（near infrared spectroscopy，NIRS）作为中药关键质量属性快速评价技术，能够有效用于中药提取、浓缩、醇沉、纯化等中药生产过程质量控制快速、无损检测的定量和定性分析[1-3]。但是，NIRS 因其吸收强度弱、谱峰重叠严重、冗余信息较多，造成模型预测性能和稳定性能较差[4]。因此，在模型的建立过程中，需采用相应的方法筛选有效变量并剔除不相关和非分析组分的干扰[5-8]。

本实验将 NIRS 技术应用于金银花提取过程[9]，

收稿日期：2017-07-24

基金项目：北京市科技新星计划项目（xx2016050）；北京中医药大学杰出青年基金项目（2015-JYB-XYQ-003）

作者简介：杜晨朝，硕士研究生，研究方向为中药过程分析与质量评价。Tel: (010)84738650 E-mail: duchenzhao12@163.com

*通信作者 吴志生，副研究员。Tel: (010)84738650 E-mail: wzs@bucm.edu.cn

乔延江，教授。Tel: (010)84738661 E-mail: yjqiao@bucm.edu.cn

实时、快速监测其指标性成分绿原酸的量[10]，同时采用组合间隔偏最小二乘法（synergy interval partial least squares，SIPLS）[11]、竞争自适应抽样方法（competitive adaptive reweighted sampling method，CARS）[12]、变量投影重要性（variable importance in projection，VIP）[13]、连续投影算法（successive projections algorithm，SPA）[14]4 种变量筛选方法建立绿原酸定量模型，比较不同变量筛选方法下模型的预测性能。以经典化学指示参数校正集决定系数（R_{cal}^2）、校正均方误差（RMSEC）、交叉验证均方根误差（RMSECV）、预测集决定系数（R_{pre}^2）和预测均方根误差（RMSEP）等作为模型评价指标[15]，进而优选最佳变量筛选方法。为 NIRS 在中药定量分析中最佳变量筛选方法的选择提供借鉴和指导，进一步提高中药生产过程的在线监测与实时控制。

1　材料与仪器

金银花购于北京本草方源药业有限公司，由北京中医药大学鉴定系刘春生教授鉴定为忍冬科植物忍冬 *Lonicera japonica* Thunb. 的干燥花蕾；绿原酸对照品（批号 110753-201314）购于中国食品药品检定研究院，质量分数大于 98%；色谱纯乙腈、色谱纯磷酸（赛默飞世尔科技有限公司）；娃哈哈纯净水（杭州娃哈哈集团有限公司）。

XL 410 近红外光谱仪及其透射光纤（AXSUN，美国）；安捷伦 1100 高效液相色谱仪（包括 G13114A 四元泵、G1379A 真空脱气机、G1313A 自动进样器、G1316A 柱温箱、G1315 二极管阵列检测器及 ChemStation 工作站，美国 Agilent 公司）。

2　方法

2.1　金银花的中试提取过程

称取金银花 150 g 置 4 L 提取容器中，加 20 倍水，提取 3 次，每次 1 h。加热阶段得到 18 个样品，沸腾后每 2 分钟采集光谱 1 次，共得到 30 个样品，上述 2 个过程共收集到 48 个样品。用 0.45 μm 滤膜滤过，同时进行 HPLC 离线分析。

2.2　在线 NIRS 采集光谱

使用近红外透射模式采集提取液光谱，采集光谱的相关参数：光谱扫描范围为 1 350～1 800 nm，扫描次数 64 次，分辨率 0.5 nm ，采用空气为背景对照，采集的金银花 NIRS 原始光谱图见图 1。

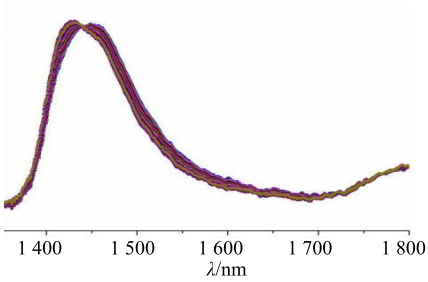

图 1　金银花 NIRS 原始光谱

Fig. 1　Raw spectra of *Lonicerae Japonicae Flos*

2.3　绿原酸的 HPLC 测定条件[16]

色谱条件为 DIKMA Diamonsil C_{18}（250 mm×4.6 mm，5 μm）色谱柱；0.2%磷酸水溶液-乙腈（15∶85）为流动相；柱温为 30 ℃；检测波长 327 nm；体积流量 1.0 mL/min；进样 10 μL。

2.4　样品的测定

按照“2.3”项液相色谱条件采集样品溶液的色谱图，采用外标一点法测定金银花中绿原酸的量。取 3 次实验的平均值为 HPLC 分析值。

2.5　数据分析

光谱预处理及模型的建立采用 Unscrambler 9.7（CAMO Software AS，Norway）软件，SIPLS、CARS、VIP、SPA 变量筛选均采用 MATLAB（MATLAB，The MathWorks，Massachussetts）软件，图形的绘制采用 ORIGIN 8 软件。

3　结果与分析

3.1　光谱预处理方法优筛

48 个金银花提取液样品采用 Kennard-Stone 方法，按照 2∶1 的比例划分成校正集（32 个样本）和验证集（16 个样本）。在 NIRS 采集过程中，光谱易受背景噪音、环境温湿度、样品自身物理变化等因素的干扰，导致基线漂移，因此在建立偏最小二乘（PLS）定量模型前，需采用合理的光谱预处理方法对样品的原始吸收光谱进行预处理。以 R_{cal}^2、RMSECV、R_{pre}^2、RMSEP 为模型性能评价指标，考察多元散射校正（multivariate scatter correction，MSC）和标准正则变换（standard normal variate，SNV）等散射校正法以及 SG 平滑（Savitzky-Golay filter smoothing，SG9 和 SG11）和矢量归一化（Normalize）等光谱预处理方法对模型性能的影响，结果见表 1。可以看出，采用 MSC 预处理后，模型的 RMSECV 和 RMSEP 较低，建模效果较好。

表 1 不同预处理方法对 PLS 模型性能的影响

Table 1 Effect of different pretreatment methods on PLS models

预处理方法	潜变量	校正集		验证集	
		R_{cal}^2	RMSECV	R_{pre}^2	RMSEP
原始光谱	2	0.987 8	4.755	0.923 6	6.483
9 点平滑（SG9）	2	0.987 1	4.885	0.914 4	6.864
11 点平滑（SG11）	2	0.986 8	4.935	0.912 1	6.956
标准正则变换（SNV）	3	0.993 3	3.816	0.966 1	4.318
多元散射校正（MSC）	3	0.994 6	3.381	0.968 5	4.160
矢量归一化（Normalize）	3	0.991 7	4.257	0.961 7	4.590

3.2 不同变量筛选方法下金银花 PLS 定量模型的建立

3.2.1 SIPLS 法 SIPLS 法是间隔偏最小二乘法的一个拓展，该方法将全光谱分解成等长的多个区间，将不同区间任意组合，以相关系数最大且 RMSECV 最小作为评价指标，选出最优区间组合。采用 SIPLS 筛选绿原酸的最佳建模波段，其筛选参数为最大潜变量因子 10，3 个区间组合，间隔数 20。结果如图 2 所示，绿原酸对应的最佳建模波段是 1 373～1 395、1 530.5～1 552.5、1 553～1 575 nm。

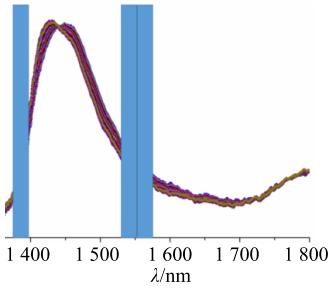

图 2 SIPLS 法绿原酸最佳建模波段考察结果

Fig. 2 Optimum subinterval combinations selected by SIPLS for quantitative determination of chlorogenic acid

3.2.2 CARS 法 CARS 是在模仿达尔文进化论中“适者生存”理论基础上提出的一种新的变量选择方法，借助自适应重加权采样技术筛选出 PLS 模型中回归系数绝对值大的波长变量，剔除权重小的波长变量，根据交叉验证均方根误差值最小原则筛选最优变量组合。

采用 CARS 法设定蒙特卡洛仿真次数为 50，选择 5 折交叉验证取交叉验证误差最小时的样本集作为最终筛选结果，筛选的变量数为 46。CARS 筛选的绿原酸的相关变量如图 3 所示。随着变量筛选过程所选择的光谱变量的变化情况见图 3-A。变量筛选过程中 RMSECV 值的变化趋势见图 3-B。变量筛选过程中光谱回归系数的变化。图中不同颜色的线代表不同的变量见图 3-C。

3.2.3 VIP 法 VIP 法是指波长变量在解释浓度变量作用的重要性，主要基于偏最小二乘回归。VIP 指标综合考虑了光谱对构造 PLS 得分的贡献和 PLS 得分对浓度变量的解释能力。某个波长变量对浓度变量的解释能力是通过得分来传递的，如果得分对浓度变量的解释能力很强，且该变量在构造这个得

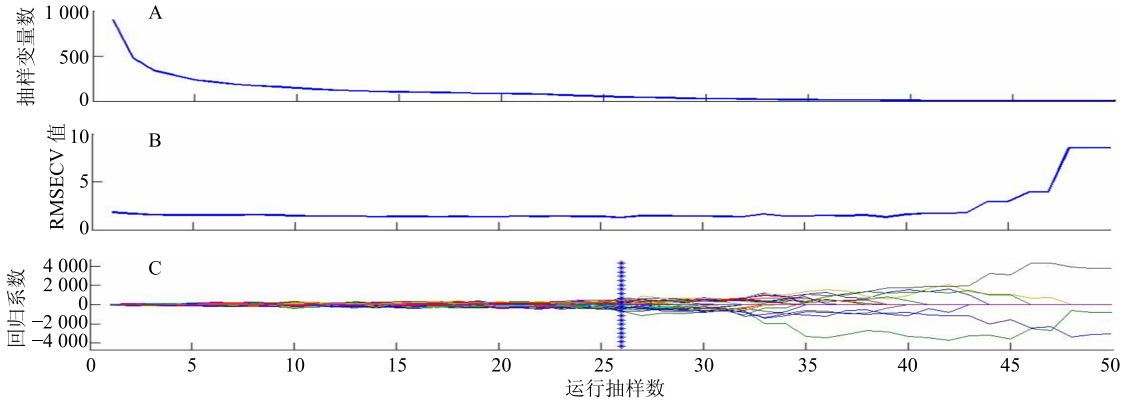

图 3 CARS 变量筛选结果

Fig. 3 CARS variable screening results

分时又起到了相当重要的作用，那么最终 VIP 指标会很大，表示该波长变量对浓度变量有很强的解释能力。本实验采用 VIP 变量筛选方法得到的变量数为 311。

3.2.4　SPA 法　SPA 法主要是应用变量投影来找到冗余信息最少和变量间共线性最小的变量组。根据最小误差均方根（RMSE）值筛选出最优变量组。SPA 变量筛选的结果如图 4 所示，当模型的变量数由 2 升至 14 时，RMSE 快速下降，表明 PLS 模型至少需要 14 个有用的光谱变量。因此，对于 SPA 变量筛选方法，共筛选出 14 个光谱变量。

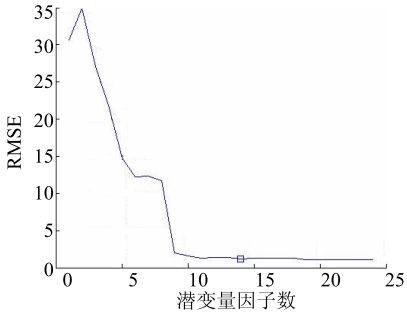

图 4　SPA 变量筛选结果

Fig. 4 SPA variable screening results

3.3　绿原酸最优定量模型的建立与预测

将金银花中绿原酸的原始光谱经过 MSC 光谱预处理后，在不同的光谱变量筛选方法下建立了不同的 PLS 模型。结果见表 2，综合 R_{cal}^2、RMSECV、R_{pre}^2、RMSEP 值分析，绿原酸采用 SIPLS 变量筛选方法建立的 PLS 模型性能明显优于其他变量筛选方法。因此，本研究选择 SIPLS 变量筛选建立绿原酸 PLS 模型并进行预测。绿原酸 NIRS 光谱预测值与 HPLC 实测值之间线性关系良好，结果如图 5 所示。绿原酸 NIRS 预测值与 HPLC 实测值的 R_{cal}^2 和 R_{pre}^2 分别为 0.998 6 和 0.990 3，RMSEC 和 RMSEP 分别为 1.503%和 2.316%，满足中药生产过程实时分析的精度要求。

4　讨论

本研究以中药金银花数据为研究载体，以 RMSEC、RMSEP、R_{cal}^2、R_{pre}^2 为评价指标，比较了 4 种变量筛选方法下所建绿原酸近红外定量模型的预测性能。结果发现，经 SIPLS 变量筛选方法筛选出的波段建立的定量预测模型性能最优，其不仅可以有效地筛选出关键建模变量剔除干扰样本，而且能有效地建立稳健、可靠、预测性能好的定量模型。研究表明，NIRS 作为一种实时、快速、无损的分析方法，为中药提取过程的实时在线监测分析提供指导和借鉴。

表 2　不同变量筛选方法建模结果比较

Table 2　Comparison on modeling results under different variables screening methods

方法	潜变量	校正集			预测集	
		R_{cal}^2	RMSEC	RMSECV	R_{pre}^2	RMSEP
SIPLS	3	0.998 6	1.503	1.739	0.990 3	2.316
CARS	2	0.988 4	4.258	4.630	0.920 1	6.629
VIP	3	0.994 6	2.910	3.361	0.688 2	13.100
SPA	2	0.986 5	4. 600	5.005	0.910 2	7.031

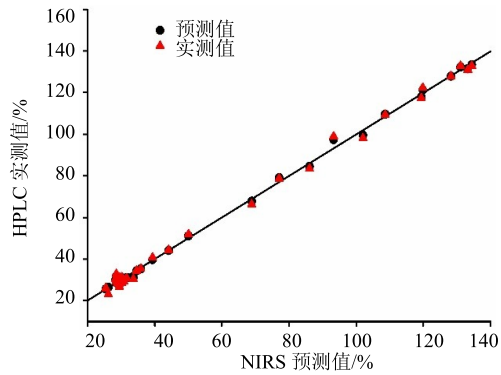

图 5　金银花中绿原酸量 NIRS 预测值与 HPLC 实测值相关关系

Fig. 5　Correction diagrams between NIRS predicted values and HPLC reference values of chlorogenic acid in *Lonicerae Japonicae Flos*

参考文献

[1]　裴艳玲, 吴志生, 史新元, 等. 中药关键质量属性快速评价 (II): NIR 光谱解析策略例证光谱学与光谱分析 [J]. 光谱学与光谱分析, 2014, 34(9): 2391-2396.

[2]　赵　娜, 吴志生, 袁瑞娟, 等. 近红外漫反射光谱法快速测定积雪草总苷中积雪草苷的含量 [J]. 世界中医药, 2013, 8(11): 1280-1286.

[3]　杜　敏, 吴志生, 巩　颖, 等. 基于近红外光谱技术的道地山药快速无损分析 [J]. 世界中医药, 2013, 8(11): 1277-1279.

[4] 彭严芳, 史新元, 李 洋, 等. 基于多变量检测限的模型变量筛选方法研究 [J]. 世界科学技术—中医药现代化, 2014, 16(5): 960-965.

[5] 孙 通, 许文丽, 林金龙, 等. 可见/近红外漫透射光谱结合 CARS 变量优选预测脐橙可溶性固形物 [J]. 光谱学与光谱分析, 2012, 32(12): 3229-3233.

[6] 李江波, 郭志明, 黄文倩, 等. 应用 CARS 和 SPA 算法对草莓 SSC 含量 NIR 光谱预测模型中变量及样本筛选 [J]. 光谱学与光谱分析, 2015, 35(2): 372-378.

[7] 闫珂巍, 王 福, 梅国荣, 等. 基于近红外光谱技术快速定性鉴别广陈皮模型的建立 [J]. 中草药, 2015, 46(20): 3096-3099.

[8] 杜晨朝, 吴志生, 赵 娜, 等. 基于两类误差检测理论金银花提取过程的 MEMS-NIR 在线分析建模方法研究 [J]. 中国中药杂志, 2016, 41(19): 3563-3568.

[9] 刘雪松, 李梦茹, 王致远, 等. 基于近红外光谱的驴胶补血颗粒浓缩过程研究 [J]. 中草药, 2016, 47(22): 3997-4002.

[10] 中国药典 [S]. 一部. 2015.

[11] Munck L, Nielsen J P, Møller B, *et al*. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics [J]. *Anal Chim Acta*, 2001, 446: 169-184.

[12] Li H, Liang Y, Xu Q, *et al*. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. *Anal Chim Acta*, 2009, 648(1): 77-84.

[13] Sills D L, Gossett J M. Using FTIR spectroscopy to model alkaline pretreatment and enzymatic saccharification of six lignocellulosic biomasses [J]. *Biot Bio*, 2012, 109(4): 894-903.

[14] Galvão R K H, Araújo M C U, Fragoso W D, *et al*. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm [J]. *Chemometr Intell Lab Syst*, 2008, 92(1): 83-91.

[15] 李 洋, 吴志生, 史新元, 等. 中试规模和不同提取时段的黄芩配方颗粒质量参数在线 NIR 监测研究 [J]. 中国中药杂志, 2014, 39(19): 3753-3756.

[16] Chen Z, Wu Z S, Shi X Y, *et al*. A study on model performance for ethanol precipitation process of *Lonicera japonica* by NIR based on Bagging-PLS and Boosting-PLS algorithm [J]. *Chin J Anal Chem*, 2014, 42(11): 1679-1686.