10.11895/j.issn.0253-3820.140465

Bagging 偏最小二乘和 Boosting 偏最小二乘算法的金银花醇沉过程近红外光谱定量模型预测能力研究

陈 昭 吴志生* 史新元 徐 冰 赵 娜 乔延江*
(北京中医药大学中药信息工程中心, 北京 100102)

摘 要 建立金银花醇沉过程中稳健的近红外光谱(Near infrared spectroscopy,NIR)定量模型,为金银花醇沉过程的快速评价提供方法。研究基于金银花醇沉过程绿原酸的 NIR 数据,通过建立 Bagging 偏最小二乘(Bagging-PLS)模型、Boosting 偏最小二乘(Boosting-PLS)模型与偏最小二乘(Partial Least Squares,PLS)模型,实现对模型性能比较;在此基础上,采用组合间隔偏最小二乘法(Synergy interval partial least squares,siPLS)和竞争自适应抽样(Competitive adaptive reweighted sampling,CARS)法分别对光谱进行变量筛选,建立模型, 实现了对模型预测性能的考察。实验结果表明, Bagging-PLS 和 Boosting-PLS(潜变量因子数设为 10)的预测性能均优于 PLS 模型。在此基础上,两批样品采用 siPLS 筛选变量,第一个批次金银花筛选波段 820 ~ 1029.5 nm 和 1030 ~ 1239.5 nm,第二个批次金银花醇沉筛选波段为 820 ~ 959.5 nm 和 960 ~ 1099.5 nm;采用 CARS 方法变量筛选,两批样品分别选择 5 折交叉验证和 10 折交叉验证,取交叉验证均方根误差(RMSECV)值最小的子集作为最终变量筛选的结果。经过变量筛选的两批金银花醇沉过程中的绿原酸含量 Bagging-PLS 和 Boosting-PLS 模型的预测均方根误差(RMSEP)值降低了 0.02 ~ 0.04 g/L,预测相关系数提高了 4% ~ 5%。综上,Baggning-PLS 和 Boosting-PLS 算法可作为金银花醇沉过程 NIR 定量模型的快速预测方法。

关键词 过程分析技术; 金银花; 醇沉; Bagging 偏最小二乘算法; Boosting 偏最小二乘算法

1 引 言

近红外光谱(Near infrared spectroscopy,NIR)技术作为一种快速、无损、环保的光谱分析技术,已经广泛应用于医药领域[1,2]。对于中药复杂体系中多组分低含量特征,近红外结合各种算法实现了对其软测量分析。在金银花醇沉过程研究中,加醇过程和最终量是中药制药过程控制的关键点。在加醇过程控制方面,采用多变量统计过程控制(MSPC)监控模型[3],使监控模型更加灵敏、稳健。在加醇终点检测方面[4],采用主成分分析结合移动块相对标准偏差(PCA-MBRSD)法,从正常加醇过程 NIR 数据中获得理想终点样本,由理想终点样本构成加醇过程终点的设计空间,进而实现准确判断加醇终点。在金银花(*Lonicera japonica*)醇沉过程中绿原酸含量偏最小二乘法(Partial least squares, PLS)模型中,采用准确性轮廓分析绿原酸含量, 该 PLS 模型具有稳健性和准确性[5]。

以金银花醇沉过程中绿原酸的 NIR 数据为载体,运用 Bagging-PLS 和 Boosting-PLS 算法,建立准确、稳健的 NIR 模型。Bagging 和 Boosting 作为两种代表性较强的集成算法,具有较高预测精度。将 Bagging 和 Boosting 引入到经典的 PLS 定量模型中,提高模型泛化能力,减小模型预测方差[6],这给中药 NIR 定量模型快速预测提供较好的借鉴。

2 实验部分

2.1 实验数据

研究基于两批金银花醇沉过程中绿原酸含量变化的近红外光谱数据[7],该数据来源于本课题组金银花醇沉过程研究,实验中绿原酸参考值采用 HPLC 测定,两批金银花药材来源于北京本草方源药业有限公司。每批次样本均有 216 个样本,PLS 模型采用 Kennard-Stone(K-S)算法划分样本集(144∶72),而

2014-05-29 收稿;2014-07-24 接受
本文系国家自然科学基金项目(No. 81303218)和高等学校博士学科点专项科研基金(No. 20130013120006)资助
* E-mail: yjqiao@263.net; wzs@bucm.edu.cn

Bagging-PLS 和 Boosting-PLS 模型从自身内部的样本进行抽样,不需要用 K-S 算法划分样本集。样品光谱波长范围为 400 ~ 2500 nm,两个批次的 HPLC 参考值统计结果如表 1 所示。

表 1 两批金银花醇沉过程中绿原酸 HPLC 参考值统计结果

Table 1 Two batches statistics of HPLC reference value (chlorogenic acid) in ethanol precipitation process of *Lonicera japonica*

批次 Batch	最小值 Minimum	最大值 Maxium	平均值 Average	标准偏差 SD	相对标准偏差 RSD (%)
1	1.57	3.01	2.28	0.37	0.2
2	1.48	2.92	2.24	0.373	0.2

注:HPLC 参考值浓度单位为:g/L。

Notes: Concentration unit of HPLC reference value: g/L.

2.2 Bagging 和 Boosting 算法简介

Bagging 是 Bootstrap aggregating 的缩写,在抽样方式上采取有放回地抽样,随机抽取与原训练集样本数相同的多个成员训练样本集。Bagging 方法通过重新选取训练集,增加了模型集成的差异度,提高泛化能力。本研究以成员模型指标的趋势图以及用简单平均方式评价模型预测结果。

Boosting 对各个成员模型的训练集选择是独立的,各轮训练集的选择与前面各轮的学习结果有关,且 Boosting 各个预测函数不像 Bagging 没有权重。

Boosting 算法实现步骤如下:

对于原训练集中每一个样本赋予相同的初始化样本权重,且 Boosting 最大迭代次数为 T:

$$\omega_i^{(t)} = 1/M, i = 1,2,\cdots,M \tag{1}$$

M 原始训练集的样本数, 取迭代次数 $t = 1,2,\cdots,T$, 重复以下步骤(ⅰ ~ ⅵ);

(ⅰ)按照样本权重抽取出第 t 轮的 M 个成员训练集样本(允许重复抽样);

(ⅱ)由第 t 轮的 M 个成员训练集样本,采用 PLS 算法建立成员模型 h_t;

(ⅲ)用成员模型 h_t 对原始训练集每个样本进行预测分析,并计算每个样本的预测值误差:

$$L_l^t = |\hat{y}_i^{(t)} - y_i| / \max|\hat{y}^{(t)} - y_i| \quad (i = 1,2,\cdots,M) \tag{2}$$

其中, $\hat{y}_i^{(t)}$ 为第 t 个模型每个样本的预测值。

(ⅳ)计算第 t 轮的加权误差和:

$$\bar{L}_t = \sum_{i=1}^{M} L_i^{(t)} \omega_i^{(t)} \tag{3}$$

(ⅴ)计算模型可信度:

$$\beta_t = \bar{L}_t / (1 - \bar{L}_t) \tag{4}$$

可信度表征了模型预测的可靠性,取值为 0 ~ 1,值越大,模型的可靠度越低。

(ⅵ)计算样本的新权重:

$$\omega_i^{(t+1)} = \omega_i^{(t)} \beta_i^{(1-\bar{L}_t)} \tag{5}$$

对于某个未知样本,可以得到 T 个预测值,再通过加权中值方式[8]、加权平均数方式[9],得到最终未知样本的预测结果。

2.3 Bagging-PLS 和 Boosting-PLS 应用于金银花醇沉过程 NIR 定量模型

将 Bagging-PLS 和 Boosting-PLS 引入到金银花醇沉过程中,建立稳健和预测性能较好的 NIR 模型。Bagging 是一种集成学习方法,应用于集成模型[10~12]中,具有较快的预测速度。在分类和回归树中,Bagging 能够提高泛化能力,获得较高的精度[13,14]。Boosting 是通过给定一个弱的学习算法,经过每次训练不断校正样本权重,最终得到一个预测函数。该算法通过产生数个简单的估计,再将这些规则集成构造出一个高精度的估计[15]。在金银花醇沉过程 NIR 模型中,将 Bagging 和 Boosting 作为两种不同的集成建模策略[16],建立 Bagging-PLS 和 Boosting-PLS 模型,快速预测金银花醇沉过程中绿原酸含量变化,为中药制药过程分析提供参考。

2.4 软件和装备设置

研究所涉及的算法均在 Matlab 7.10 (MatlabWorks Inc., U.S.)平台上实现,PLS 算法使用 PLS_Toolbox(Eigenvector Research Inc.,U.S.),近红外仪为 XDS Rapid Liquid Analyzer 近红外光谱仪(美国 FOSS 公司),运用 Unscrambler9.7(挪威 CAMO 软件公司)软件对光谱进行预处理和相关计算。CARS 算法来源于 http://code.google.com/p/carspls/,其它算法自行编写。模型预测性能用 r(相关系数)和

RMSEP(预测均方根误差)来指示,模型迭代次数选用 500 和 1000 作为对比研究。

3　结果与讨论

3.1　全谱模型比较

基于两个批次的金银花醇沉过程中 NIR 数据如图 1 所示。从图 1 可见，光谱在 2000 ~ 2500 nm 的组合频谱区有较大噪音。经过不同预处理方法,建立了 PLS 模型。模型的预测相关系数 r_p、校正集的平方相关系数 R^2_{cal}、验证集的平方相关系数 R^2_{val}值见表 2 和表 3。

由表 2 和表 3 可知,经过不同预处理方法的 PLS 模型 r_p 在 0.92 ~ 0.97 之间波动,且采用不同的预处理方法,结果差异比较大,这说明在建模过程中预处理选取对 PLS 模型产生较大影响。其中,PLS 模型采用 K-S 算法划分样本集,而 Bagging-PLS 和 Boosting-PLS 采用自身内部的样本进行抽样,集成多个模型,可以获得稳健的预测模型。

一般情况下,Bagging-PLS 和 Boosting-PLS 迭代次数达到 500 后，迭代指标基本上达到稳定。研究中未经变量筛选的 Bagging-PLS 和 Boosting-PLS 模型潜变量因子设置为 10,图 2 和图 3 选取 1000 次迭代,以更好地反映预测相关系数 r_p 值和预测均方根误差 RMSEP 值的变化趋势，随着迭代次数的增加，r_p 值和 RMSEP 值逐步趋于稳定,由此可对模型的稳定性与准确性做出判断。

为了考察不同迭代次数对预测结果的影响,光谱在不经过任何预处理和变量筛选的情况下,选用 Bagging-PLS 和 Boosting-PLS 算法对金银花醇沉过程 NIR 定量模型进行预测。预测结果如表 4 所示，两

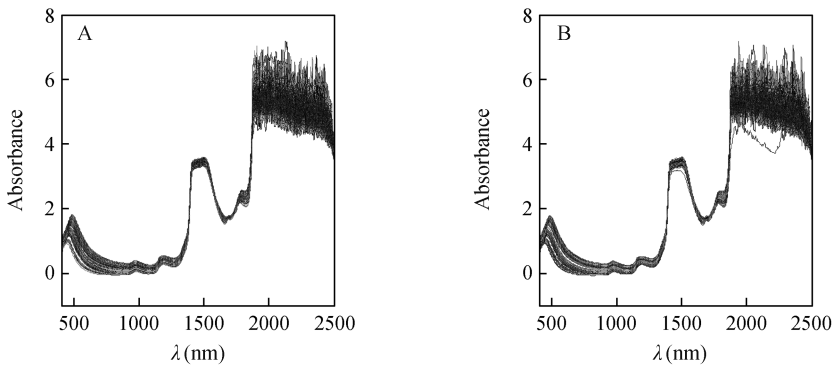

图 1　第一批(A)和第二批(B)的金银花醇沉过程 NIR 图

Fig. 1　NIR spectra of ethanol precipitation process of *Lonicera japonica* (A is the first batch and B is the second batch)

表 2　第一批次金银花醇沉过程 PLS 模型结果

Table 2　Results of PLS models of ethanol precipitation process of *Lonicera japonica* in the first batch

预处理方法 Pretreatment method	潜变量 Later factors	校正集 Calibration set		验证集 Validation set		r_p
		R^2_{cal}	RMSEC	R^2_{cal}	RMSEC	
Raw	4	0.9667	0.0682	0.9643	0.0705	0.9342
Baseline	2	0.9364	0.0942	0.9196	0.1069	0.9547
SNV	4	0.9783	0.0550	0.9767	0.0569	0.9421
SNV+Baseline	4	0.9714	0.0611	0.9733	0.0631	0.9446
Baseline+SNV+Noise	3	0.9931	0.0311	0.9926	0.0231	0.9523
S-G	3	0.9661	0.0687	0.9637	0.0711	0.9341
S-G+Noise	2	0.9899	0.0375	0.9895	0.0382	0.9441
SNV+Noise	4	0.9848	0.0466	0.9837	0.0483	0.9432
ST	4	0.9582	0.0763	0.9554	0.0788	0.9647
SNV+ST	4	0.9563	0.0780	0.9523	0.0816	0.9717
Baseline+ST	2	0.9296	0.0991	0.9259	0.1016	0.9663
MSC	2	0.9439	0.0885	0.9138	0.1099	0.9543

Raw：原始光谱，S-G：Savitaky-Golay 平滑，ST：光谱吸光度到 Kubelka-Munk 的转换，SNV：标准归一化，MSC：多元散射校正，Noise：加噪。

Raw: Original spectrum, S-G: Savitaky-Golay, ST: Spectroscopic transformation, absorbance to Kubelka-Munk transformation, SNV: Standard normal variate, MSC: Multiplicative scatter correction, Noise: Added noise; RMSEC: Root mean square error of calibration; RMSECV: Root mean square error of cross-validation.

表 3 第二批次金银花醇沉过程 PLS 模型结果

Table 3 Results of PLS models of ethanol precipitation process of *Lonicera japonica* in the first batch

预处理方法 Pretreatment method	潜变量 Later factors	校正集 Calibration set		验证集 Validation set		r_p
		R^2_{cal}	RMSEC	R^2_{cal}	RMSEC	
Raw	4	0.9693	0.0659	0.9663	0.0691	0.9234
Baseline	4	0.9749	0.0608	0.9718	0.0632	0.9415
SNV	3	0.9691	0.0662	0.9671	0.0683	0.9394
S-G	4	0.9691	0.0662	0.9660	0.0694	0.9236
SNV+Baseline	4	0.9750	0.0595	0.9732	0.0617	0.9347
Baseline+SNV+Noise	3	0.9963	0.0224	0.9962	0.0232	0.9495
SNV+Noise	3	0.9746	0.0600	0.9730	0.0619	0.9392
ST	4	0.9552	0.0797	0.9468	0.0869	0.9316
SNV+ST	4	0.9583	0.0769	0.9546	0.0803	0.9419
MSC	2	0.9439	0.0885	0.9138	0.1099	0.9543

Raw：原始光谱，S-G：Savitaky-Golay 平滑，ST：光谱吸光度到 Kubelka-Munk 的转换，SNV：标准归一化，MSC：多元散射校正，Noise：加噪。

Raw：Original spectrum，S-G：Savitaky-Golay，ST：Spectroscopic transformation，absorbance to Kubelka-Munk transformation，SNV：Standard normal variate，MSC：Multiplicative scatter correction，Noise：Added noise.

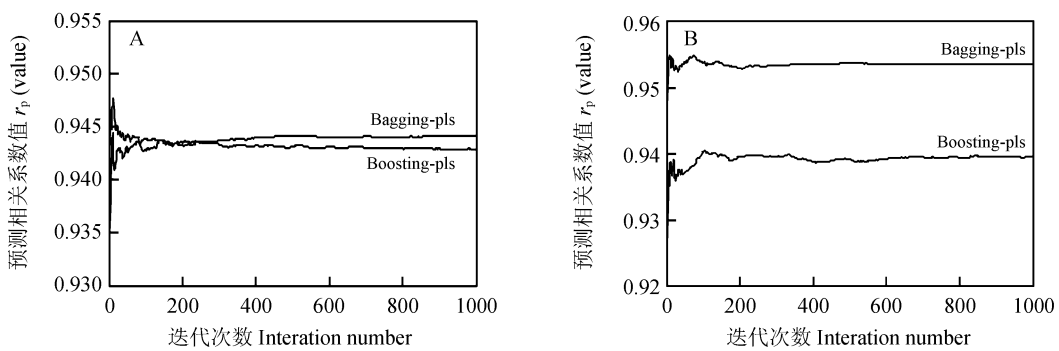

图 2 两批金银花 NIR 模型的 r_p 值趋势图(A 为第一批，B 为第二批)

Fig. 2 r_p values for NIR models of ethanol precipitation process of *Lonicera japonica* (A is the first batch, B is the second batch)

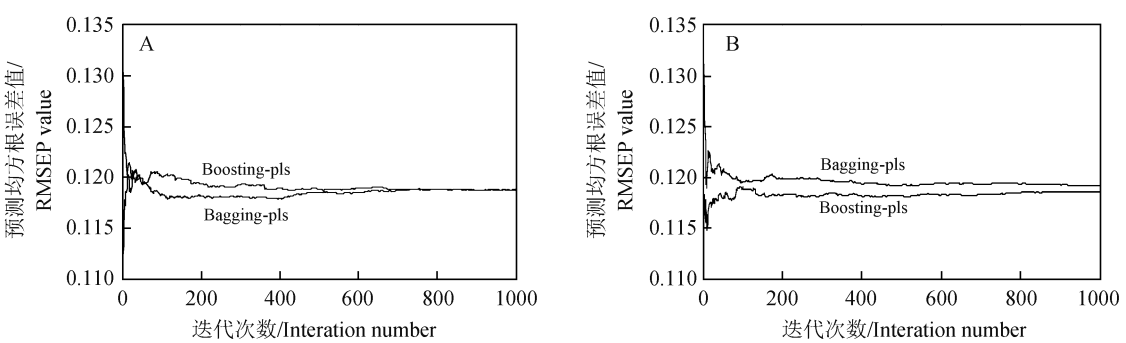

图 3 两批金银花醇沉 NIR 模型的 RMSEP 值趋势图(A 为第一批，B 为第二批)

Fig. 3 RMSEP values for NIR models of ethanol precipitation process of *Lonicera japonica* (A is the first batch, B is the second batch)

个批次金银花的 Bagging-PLS 和 Boosting-PLS 模型的 r_p 值范围在 0.93 ~ 0.95 之间，模型稳健性较好。RMSEP 值在 0.12 左右变化，且波动范围较小，进一步说明迭代次数到达 500 次后模型基本稳定。

3.2 变量筛选后模型预测性能

3.2.1 基于 siPLS 的变量筛选 为了进一步研究优化模型，采用组合间隔偏最小二乘法(Synergy interval partial least squares, siPLS)算法对两个批次的光谱进行波段筛选。siPLS 算法是在 Norgaard 的 iPLS[17] 基础上的扩展，它计算所有联合区间的PLS模型(一般间隔联合数取2，3或4)，最后选择交叉

表 4　两个批次金银花 NIR 模型预测结果

Table 4　Prediction results of NIR models for two batches of ethanol precipitation process of *Lonicera japonica*

算法 Algorithm	样品批次 Batches of samples	迭代次数 Interation number	r_p 值(平均) r_p(mean)	RMSEP(平均) RMSEP(mean)
Bagging-PLS	1	500	0.943 2	0.117 9
		1000	0.942 8	0.118 5
	2	500	0.938 9	0.118 2
		1000	0.940 8	0.116 4
Boosting-PLS	1	500	0.943 1	0.118 4
		1000	0.942 8	0.118 8
	2	500	0.938 3	0.118 0
		1000	0.938 7	0.117 8

验证均方根误差(Root mean square error of cross-validation,RMSECV)值最小的变量区间作为 NIR 定量模型变量筛选的结果。

RMSECV 计算公式：$\mathrm{RMSECV}=\sqrt{\frac{\sum_{i=1}^{n}(y_{\mathrm{NIR}_i})-y_{\mathrm{Ref}_i})^2}{n}}$，其中 y_{NIR_i}是 PLS 模型计算值，y_{Ref_i}为 HPLC 参考值。

采用 siPLS 方法筛选变量,图 4 和图 5 中 r 值表示为全样本的相关系数。两个批次的金银花醇沉过程NIR分别以RMSECV为筛选波段的指标,第一批次金银花筛选波段组合数(间隔数为10)为

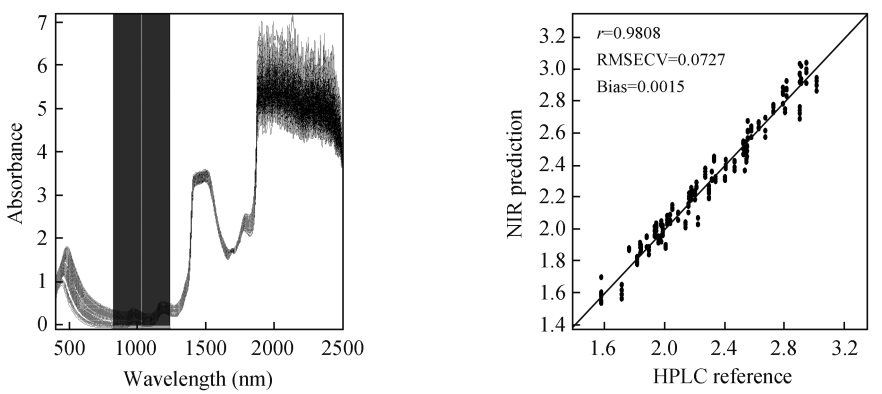

图 4　第一批次金银花醇沉 siPLS 变量筛选

Fig. 4　Ethanol precipitation of *Lonicera japonica* in the first batch using synergy interval PLS (siPLS) to select variables

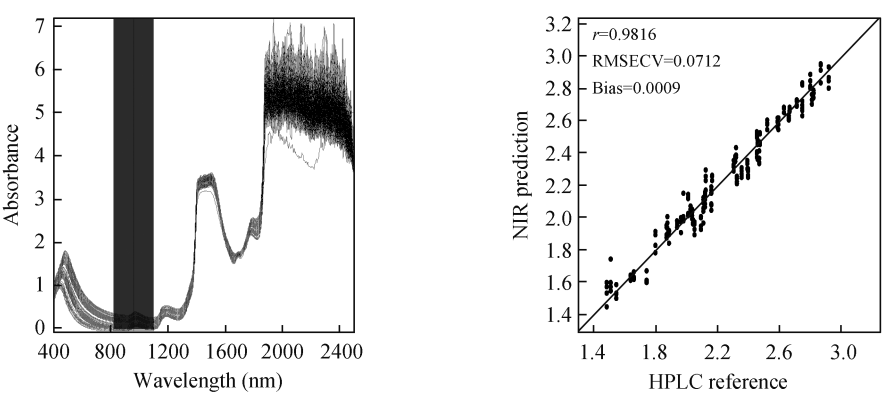

图 5　第二批次金银花醇沉 siPLS 变量筛选

Fig. 5　Ethanol precipitation of *Lonicera japonica* in the friest batch using synergy iterval PLS (siPLS) to select variables

3 和 4，其对应波段分别为 820 ~ 1029.5 nm，1030 ~ 1239.5 nm；第二个批次金银花筛选波段（间隔数为 15）组合数为 4 和 5，其对应波段分别为 820 ~ 959.5 nm，960 ~ 1099.5 nm。

采用已筛选好的波段分别对两个批次金银花醇沉过程中的 NIR 建模，潜变量因子为 8。如图 6 和图 7 所示。结果表明：两个批次的预测相关系数 r_p 值比未筛选波段的建模结果高出 2% ~4%；两个批次样本模型的 RMSEP 值比未筛选波段的模型低 0.02 ~ 0.04 g/L。这在一定程度上反映了经过 siPLS 筛选波段后的模型准确性有一定提高。

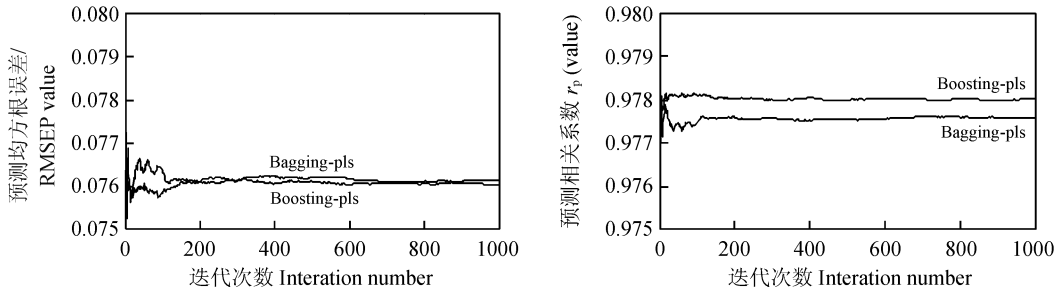

图 6 第一批次金银花醇沉 siPLS 波段筛选后模型结果

Fig. 6 NIR models results of ethanol precipitation of *Lonicera japonica* in the first batch after variables selection using siPLS

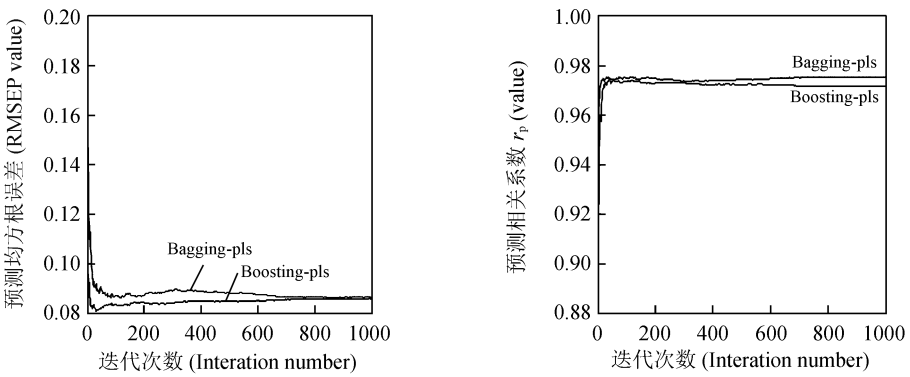

图 7 第二批次金银花醇沉 siPLS 波段筛选后模型结果

Fig. 7 NIR models results of ethanol precipitation of *Lonicera japonica* in the second batch after variables selection using siPLS

3.2.2 基于 CARS 变量筛选 CARS (competitive adaptive reweighted sampling)[18] 即竞争自适应抽样。此方法可以通过找到一个最优组合的关键波长来解释物质化学性质。其实现是一个连续的迭代过程，是被选变量子集逐渐收敛于特征变量[19]，迭代过程结束后选择 RMSEP (root mean square error of prediction) 值最小的变量子集作为特征样本集。设定蒙特卡洛仿真次数为 1000，分别选择 5 折交叉验证和 10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，两批样品得到的变量数分别为 290 和 261。将筛选的变量分别建立 CARS-Bagging-PLS 和 CARS-Boosting-PLS 模型。由表 5 可知，经过 CARS 变量筛选后的模型预测能力有了较大的提高，随着迭代次数的不断增加，r_p 值较未进行变量筛选的值高出了 4 ~ 5 个百分点，并且其值稳定在 0.98 附近。

表 5 两个批次的金银花醇沉过程模型

Table 5 The models for two batches of ethanol precipitation of *Lonicera japonica*

算法 Algorithms	批次 Batches	变量筛选 Variable selection	RMSEP 平均值 RMSEP (mean)	r_p 平均值 r_p (mean)
Bagging-PLS	1	未筛选 Without selected	0.1189	0.9425
		CARS	0.0723	0.9805
	2	未筛选 Without selected	0.1183	0.9391
		CARS	0.0713	0.9817
Boosting-PLS	1	未筛选 Without selected	0.1191	0.9423
		CARS	0.0727	0.9803
	2	未筛选 Without selected	0.1173	0.9391
		CARS	0.0708	0.9821

CARS: Competitive adaptive reweighted sampling; RMSEP: Root mean square error of prediction.

从表 5 可见，经过 CARS 筛选出的波段进行建模，r_p值有了较大提高，并且相对应的 RMSEP 也有了明显的降低。由此可以说明，在金银花醇沉过程中，运用 Bagging 和 Boosting 策略建模时，用 CARS 对光谱进行变量筛选，所建的模型提高了 r_p 值，并且减小 RMSEP 值，这为金银花醇沉过程中绿原酸含量的快速预测提供方法。

4　结　论

研究表明，Bagging-PLS 和 Boosting-PLS 两种模型表现出较稳健的预测性能。相比于 PLS 模型，两种模型能够提高模型的准确性。在 NIR 近红外在线检测实际应用中，可以采用这类方法建立模型，指导中药生产过程质量的在线监测。此外，经过变量筛选后，两种模型的预测性能和稳健性有了较大的提高，说明了特征性波段的 Bagging-PLS 和 Boosting-PLS 模型能够建立更准确、稳健的模型。Bagging-PLS 和 Boosting-PLS 两种模型提供了预测性能较好的金银花醇沉 NIR 定量模型，这将为中药的 NIR 在线质量分析和控制提供可借鉴的建模方法。

致 谢　感谢天津工业大学卞希慧博士对本研究算法的指导。

References

1　LIU Shu-Hua, ZHANG Xue-Gong, ZHOU Qun, SUN Su-Qin. *Spectroscopy and Spectral Analysis*, **2006**, 26(4): 629-632
刘沭华, 张学工, 周 群, 孙素琴. 光谱学与光谱分析, **2006**, 26(4): 629-632

2　WU Zhi-Sheng, TAO Ou, CHENG Wei, YU Lu, SHI Xin-Yuan, QIAO Yan-Jiang. *Chinese J. Anal. Chem.*, **2011**, 39(5): 628-634
吴志生, 陶 欧, 程 伟, 郁 露, 史新元, 乔延江. 分析化学, **2011**, 39(5): 628-634

3　XU Bing, SHI Xin-Yuan, QIAO Yan-Jiang, DU Min, SUI Cheng-Lin, LIU Qian. *China Journal of Traditional Chinese Medicine and Pharmacy*, **2012**, 4: 784-788
徐 冰, 史新元, 乔延江, 杜 敏, 隋丞琳, 刘 倩. 中华中医药杂志, **2012**, 4: 784-788

4　XU Bing, LUO Gan, LIN Zhao-Zhou, AI Lu, SHI Xin-Yuan, QIAO Yan-Jiang. *Chemical Journal of Chinese Universities*, **2013**, 34(10): 2284-2289
徐 冰, 罗 赣, 林兆洲, 艾 路, 史新元, 乔延江. 高等学校化学学报, **2013**, 34(10): 2284-2289

5　Wu Z S, Xu B, Du M, Sui C L, Shi X Y, Qiao Y J. *Journal of Pharmaceutical and Biomedical Analysis*, **2012**, 62: 1-6

6　CHENG Long, WANG Gui-Zeng. *Journal of Tsinghua University* (*Science and Technology*), **2008**, 48(s2): 1780-1784
程 龙, 王桂增. 清华大学学报(自然科学版), **2008**, 48(s2): 1780-1784

7　Wu Z S, Du M, Sui C L, Shi X Y, Qiao Y J. *Analytical Methods*, **2012**, 4(4): 1084-1088

8　Drucker H. *Proceedings of the Fourteenth International Conference on Machine Learning*, **1997**

9　Shao X, Bian X, Cai W. *Analy. Chim. Acta*, **2010**, 666(1-2): 32-37

10　ZHU Hong-Bin. *Computer Applications and Software*, **2010**, 27(1): 234-236
朱红斌. 计算机应用与软件, **2010**, 27(1): 234-236

11　HE Ming, LI Guo-Zheng, YUAN Jie, WU Geng-Feng. *Journal of Shanghai University*(*Natural Science Edition*), **2006**(4): 415
何 鸣, 李国正, 袁 捷, 吴耿锋. 上海大学学报(自然科学版), **2006**, (4): 415

12　WANG Li, ZHU Xue-Feng. *Control Engineering of China*, **2009**, 16(1): 59-61, 79
王 立, 朱学峰. 控制工程, **2009**, 16(1): 59-61, 79

13　Breiman L. *Machine Learning.*, **1996**, 24(2): 123-140

14　Zhang H, Ishikawa M. *International Congress Series*, **2007**, 1301: 184-187

15　YU Ling, WU Tie-Jun. *Pattern Recognition and Artificial Intelligence*, **2004**, 17(1): 52-59
于 玲, 吴铁军. 模式识别与人工智能, **2004**, 17(1): 52-59

16　CHU Xiao-Li, XU Yu-Peng, LU Wan-Zhen. *Chinese J. Anal. Chem.*, **2008**, 36(5): 702-709
褚小立, 许育鹏, 陆婉珍. 分析化学, **2008**, 36(5): 702-709

17　Norgaard L, Saudland A, Wagner J, Nielsen J P, Munck L, Engelsen S B. *Applied Spectroscopy*, **2000**, 54(3): 413-419

18 Li H, Liang Y, Xu Q, Cao D S. *Anal. Chim. Acta*, **2009**, 648(1): 77-84
19 LIN Zhao-Zhou, SHI Xin-Yuan, QIAO Yan-Jiang. *World Science and Technology/Modernization of Traditional Chinese Medicine and MateriaMedica*, **2012**, 14(4): 1760-1766
林兆洲, 史新元, 乔延江. 世界科学技术-中医药现代化, **2012**, 14(4): 1760-1766

A Study on Model Performance for Ethanol Precipitation Process of *Lonicera japonica* by NIR Based on Bagging-PLS and Boosting-PLS algorithm

CHEN Zhao[1,2], WU Zhi-Sheng[*2], SHI Xin-Yuan[2], XU Bing[2], ZHAO Na[2], QIAO Yan-Jiang[*2]
[1](*Fujian University of Traditional Chinese Medicine*, *Fuzhou* 350122, *China*)
[2](*Research Center of TCM Information Engineering*, *Beijing University of Chinese Medicine*, *Beijing* 100102, *China*)

Abstract To provide the methodology for rapid quality evaluation of *Lonicera japonica*, we have established the stable quantitative model of near infrared spectroscopy (NIR). The performance of Bagging partial least squares (Bagging-PLS) model and Boosting partial least squares (Boosting-PLS) model was compared with that partial least squares (PLS) model based on the NIR data of ethanol precipitation process of *Lonicera japonica*. On this basis, the performance of these two models after variables selection was also studied by the methods of siPLS (synergy interval partial least squares) and CARS (competitive adaptive reweighted sampling). The experimental results showed that the prediction performance of Bagging-PLS and Boosting-PLS models was superior to PLS model with the latent factor of 10. The band of 820-1029.5 nm and 1030-1239.5 nm for the first batch was selected by the method of siPLS. In addition, the band of 820-1029.5 nm and 1030-1239.5 nm was selected for the second batch sample in the same method. Furthermore, the method of CARS was taken to select variables for the two batches samples with 5-fold cross-validation and 10-fold cross-validation. And the lowest RMSECV(root mean square error of cross-validation) values were used to take subset. Compared to the model performance without the method of CARS, the RMSEP value of the Bagging-PLS model and Boosting-PLS model for the concentration of chlorogenic acid reduced by 0.02-0.04 g/L and r_p(correlation coefficient of prediction) value increased by 4%-5%. Generally, Bagging-PLS and Boosting-PLS could be regarded as rapid prediction methodsfor NIR quantitative models of ethanol precipitation process of *Lonicera japonica*.

Keywords Process analysis technology; *Lonicera japonica*; Ethanol precipitation; Bagging-partial least squares model; Boosting-partial least squares moldel

(Received 29 May 2014; accepted 24 July 2014)

This work was supported by the National Natural Science Foundation of China (No. 81303218) and the Special Research Foundation for the Doctoral Program of Higher Education (No. 20130013120006)