正交空间样本选择在金银花多批醇沉过程中的应用*

□林兆洲 徐 冰 史新元 乔延江**
(北京中医药大学中药学院 北京 100102)

摘 要:本研究探讨模型传递方法在金银花醇沉过程中的适用性。在金银花醇沉过程中,以高效液相色谱法(HPLC)为对照分析方法,测定提取液中绿原酸的含量,分别运用 iPLS 算法、SPLS 算法和模型更新(MU)建立绿原酸的 NIR 定量校正模型,检测另外两个批次金银花醇沉过程中绿原酸的含量变化。为使选择的样本更具代表性,本研究在 OSC 的基础上提出了一种新的样本选择方法——正交空间样本选择(OS),并与常用的 K-S 法进行了对比。结果表明,采用模型更新可以实现定量校正模型在不同的批次之间稳健传递。且用 OS 法选择更新样本能够明显提高模型预测的准确性和稳健性。

关键词:NIR 模型传递 正交空间样本选择 SPLS 金银花醇沉过程
doi: 10.3969/j.issn.1674-3849.2012.06.014

模型传递(或称仪器标准化)是指经过数学方法处理后,使在一台仪器上建立的校正模型能够直接在其它仪器上应用,从而减少重新建模所带来的大量的人力、物力和时间的消耗[1]。模型传递算法主要用于校正不同仪器间响应误差。除此之外,还可用于校正因数据采集策略的改变、光学元件老化等所导致的响应误差。目前,常用模型传递方法大多是有标传递算法,如 SBC 算法(Simulated Bee Colony, SBC)[1,2]、分段直接校正(Piecewise direct standardization, PDS)[3]、光谱回归(Spectral Regression)[4]、CCA 算法(Canonical Correlation Analysis, CCA)[5]等。这些算法都成功实现了模型在不同的仪器间的传递,但对于组成变异较大的分析对象(如中药原药材、中药制剂中间体)效果较差。而无标传递算法对此类问题有明显的优势,这类算法包括 FIR 滤波(Finite Impulse Response)[6]、SST 算法(Spectral Space Transformation)[7]、模型更新(Model Updating, MU)[8]等。

近年来,近红外光谱技术在中药生产过程质量控制领域的应用已有大量研究,但并没有实现大规模的工业应用。究其原因,可能由于原药材质量波动较大而使常规建立的校正模型的性能难以在长时间内保持稳定。虽然借助实验设计可以将药材有效成分变化对近红外光谱的影响纳入建模过程,在一定程度上提高模型的稳健性。但实际上,受原药材品质的多变性和一些不可控因素的影响,这样处理的效果并不理想。这使近红外光谱技术在中药领域的实际应用面临极大的挑战。醇沉过程是中药制剂生产过程的关键环节,对醇沉过程进行质量控制对于确保中药制剂质量的稳定具有

收稿日期:2012-02-24
修回日期:2012-05-08
* 科学技术部国家"重大新药创制"科技重大专项(2010ZX09502-002):清开灵注射液安全性关键技术研究,负责人:乔延江。
** 通讯作者:乔延江,本刊编委,教授,主要研究方向:中药信息学,Tel:010-84738650,E-mail:yjqiao@263.net。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

重要作用。本研究以金银花醇沉过程为载体，选择模型更新（MU）、间隔偏最小二乘算法（Interval Partial Least Squares Algorithm, iPLS）[9,10]和 SPLS[11]（Stacked Partial Least Squares）算法，建立校正模型，探讨模型传递方法在中药制剂生产过程质量控制中的适用性。

一、原理和算法

1. 间隔偏最小二乘算法（iPLS）

iPLS 算法由Norgaard 等[9]在 2000 年提出后，在近红外光谱定量中得到了广泛的应用。它首先将整条光谱划分成 n 个等宽的连续片段，然后在每段光谱上分别建立 PLS（Partial Least Squares）回归模型。因为每段光谱区间所携带的信息不尽相同，所以在每段光谱上所建立的 PLS 模型的潜变量因子数不尽相同。计算每段光谱上所建立的局部模型的 RMSEP（Root Mean Square Error of Predictions），选择 RMSEP 最小的局部（子）模型作为最终的预测模型。RMSEP 的计算公式如下：

$$RMSEP=\sqrt{\frac{\sum_{i=1}^{m}(\hat{c}_i-c_i)^2}{m}} \quad (1)$$

其中，m 为测试集样本数；c_i 为样本 i 的实测值，$\hat{c}_i$为样本 i 的预测值。

某波段的 RMSEP 越小，其所含的与被分析对象相关的信息越多，对基质变化越不敏感，所建立的定量校正模型用于预测新批次样本中绿原酸的含量的效果也就越理想。

2. SPLS 算法

SPLS 是一种集合建模（模型融合）的方法。与 iPLS 类似，首先将光谱分割成 n 个等宽的间隔，然后在每段光谱区间分别建立 PLS 回归模型。二者的差异体现在对建立的子模型的利用方式上。SPLS 将所得的每个模型按一定的权重组合成一个整体去预测新样本的性质。如方程（2）所示加权的目的是为了使 SPLS 模型的交叉校验误差最小[1]。

$$\hat{w}=\mathrm{ARG}\min(\mathrm{y}-\sum_{k=1}^{n}w_k y_k)^2 \quad (2)$$

其中，y 是样本的实测值，$\hat{y}_k$是第 k 个模型的预测值，w_k 是第 k 个模型的权重；ARG min()表示取括号内式子的最小值。

w_k 可以通过第 k 个模型的交叉验证误差得到：

$$w_k=\frac{s_k^2}{\sum_{k=1}^{n}s_k^2} \quad (3)$$

其中，s_k 是第 k 个模型的交叉验证误差的倒数。为了提高模型的性能，需要对间隔的宽度（间隔数，n）进行适当的调整。因 SPLS 在模型传递时不需要标准样品，所以本研究中将 SPLS 用于醇沉过程绿原酸定量校正模型的建立。

3. 模型更新（MU）

模型更新，即按一定的规则从新的量测数据中选择适当的样本加入到训练集中，然后重新计算模型的参数。这样新建立的模型就既包含原训练集中的变异信息，又包含了新样本的变异信息，从而拓宽了模型的适用范围，准确的预测新的量测数据。为了使新选入的样本能够有效代表新的量测数据，通常选择 K-S 法选择样本。但因 K-S 在整个光谱空间内选择样本，选出的样本特异性不强，不能有效体现被分析成分的信息。因此，本文提出一种基于正交信号校正（OSC）[13]所建立的样本选择方法 OS。OS 法的实现过程与 K-S 类似，只不过样本间欧氏距离的计算不是在原始的光谱空间，而是在经过 OSC 校正后的光谱空间，这样选择的数据能够有效代表光谱中所携带的与物质待测性质相关的信息，提高了选入样本的特异性。

二、实验部分

1. 样品含量的测定和光谱的采集

金银花醇沉过程持续 30 min，在醇沉过程中每隔 30 s 取 1.5 mL 醇沉夜，离心备用。本研究中的 3 个批次醇沉操作，均在实验室中完成。醇沉液中绿原酸的含量按照《中国药典》（2010 版一部）金银花药材项下规定的方法进行测定[14]。

用 Antaris Nicolet FT-NIRsystem（Thermo Fisher Scientific Inc., USA）近红外光谱仪采集样本的透射光谱。每条光谱扫描 16 次，光谱范围 10000~4000 cm^{-1}，分辨率 8 cm^{-1}。每个样本重复测量 3 次，取 3 次测量的平均值进行后续计算[14]。

2. 评价参数和计算软件

用 RMSECV、RMSEP 作为评价上述建模方法的性能指标。所有的计算用 Matlab 7.9 完成。除 iPLS 算法*外，其它算法的函数均自行编写。

* http://www.models.life.ku.dk/algorithms.

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

三、结果与讨论

1. iPLS 定量模型的建立

将第一批所得的 61 个样品，按取样时间排序，然后剔除离群值，将剩余的 58 个样品作为训练集。采用十折交叉验证，以 RMSECV 为指标选择最佳间隔数。从图 1 中可以明显看出，当间隔数为 33 时，所对应的 RMSECV 最小，对应的光谱区间为 5654~5832 cm^{-1}。用所建立的 iPLS 模型预测另外两个批次样品的绿原酸含量，结果如图 2 所示。两批样本的预测值与液相测量值之间有较好的线性相关性，相关系数 r 分别为 0.9982 和 0.9976。但这两批样品的 RMSEP 均较高，约是训练集 RMSECV 10 倍（表 1）。这表明在另外两个批次样品中出现了新的变异。与全谱 PLS 模型的结果对比，虽然 iPLS 对训集的拟合能力和新样本的预测准确性都有一定程度的提高，但对于新样本中出现的变异仍不够稳健。经 SBC 校正后，两个批次样品的 RMSEP 都有大幅的下降，说明不同批次的金银花提取物之间可能存在一定的系统误差。

2. SPLS 定量校正模型的稳健性

用上述所得的训练集，以 RMSECV 为优化目标选择适当的间隔数建立 SPLS 定量校正模型。如图 1 所示，RMSECV 曲线在间隔数为 33 时达到最低点，即间隔数为 33 时，所建立的 SPLS 模型最优，所对应的光谱区间为 5654~5832 cm^{-1}。由表 1 可知，SPLS 模型对训练集的拟合效果优于 iPLS 模型，从图 1 中也可得出相同的结论。图 3 是 SPLS 模型对另两个批次金银花样本的预测结果，两个批次金银花的预测值都离理想的校正曲线较远。此外，从表 1 中可以看出，SPLS 在本研究中能够更好的拟合训练集的数据，但对金银花提取物批次间的变异更为敏感。类似的，对 SPLS 模型预测的结果经 SBC 校正后也使得模型预测结果的准确性得到了很大的改善。

3. 模型更新

本研究对每一批样品分别用 K-S 法选出 20 个样品进行模型的更新。虽然有研究表明，模型的预测性能会随选入样本量的不同，而呈现出差异。但本研究中每个批次的样本仅有 61 个，在剔除离群值后则更少，因此本研究不再对选入的样本量进行优化。

近红外光谱是一种由分子振动和外部环境共同作用而产生的吸收光谱。光谱除携带与被分析物质相关的信息外，还含有其它理化性质的信息。为了在选出的样本集中更好的体现与被分析物质相关的信息，本研究在经 OSC 校正的光谱空间中进行样本的选择，并与 K-S 法进行对比，结果如表 2 所示。不难看出，在经过模型的更新以后，模型对新样品的预测准确性有了大幅提高。与其它方法相比，虽然更新后的模型对训练集中数据的拟合能力略有下降，但仍能满足实际分析工作的需求。对两种样本选择方法进行对比后发现，在对两批样本进行更新时，用 OS 法所得模型的预测结果更贴近于理想的校正曲线，即所得模型的预测性能和对新变异的稳健性要优于 K-S 法（图 4）。这可能因 OS 法仅考虑在与所预测物质属性相关的光谱空间上进行样本的划分，所得模型的预测性能就在一定程度上

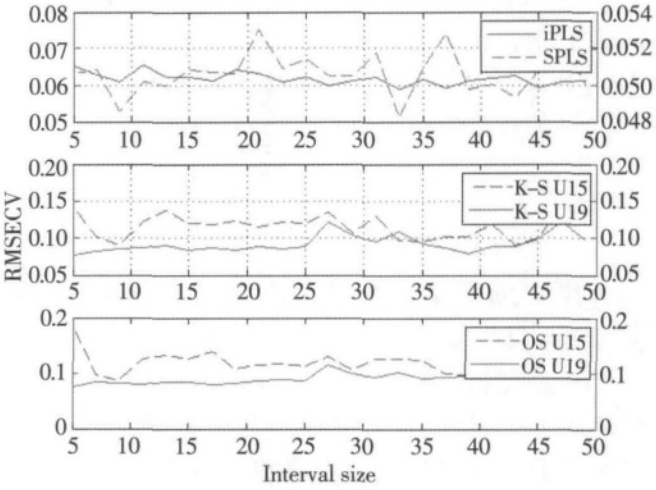

图 1 不同方法交叉验证均方根误差随窗口数目的变化

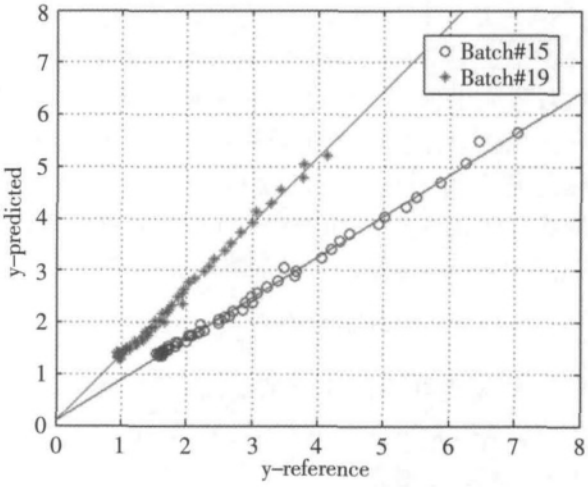

图 2 两批样品中绿原酸含量的预测结果

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1　iPLS 和 SPLS 传递性能的比较

Methods	[1]Batch #1	[2]Batch #2	[2]Batch #3
PLS	0.0672	0.6117	0.7801
iPLS	0.0587	0.5823	0.5982
iPLS[SBC]	—	0.0871	0.0584
SPLS	0.0484	0.6493	0.7084
SPLS[SBC]	—	0.0715	0.0455

注：[SBC] 结果经 SBC 校正（SBC corrected results）；交叉验证均方根误差 RMSECV（Root Mean Square Error of Coross-Validation）；[2] 预测均方根误差，RMSEP（Root Mean Square Error of Prediction）。

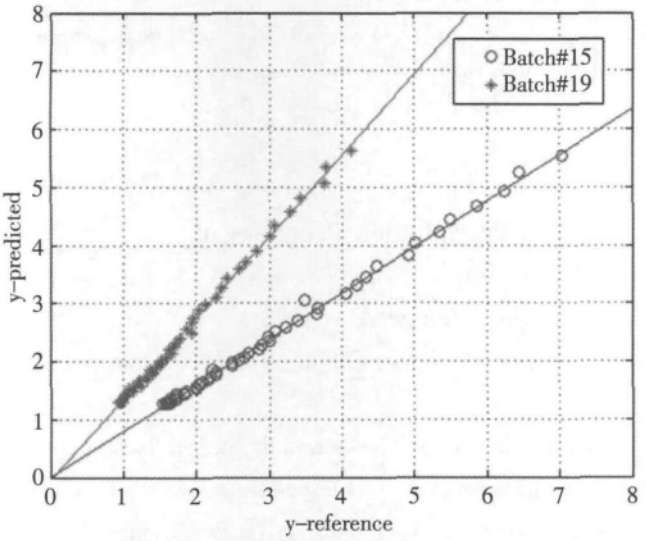

图 3　SPLS 算法的预测结果

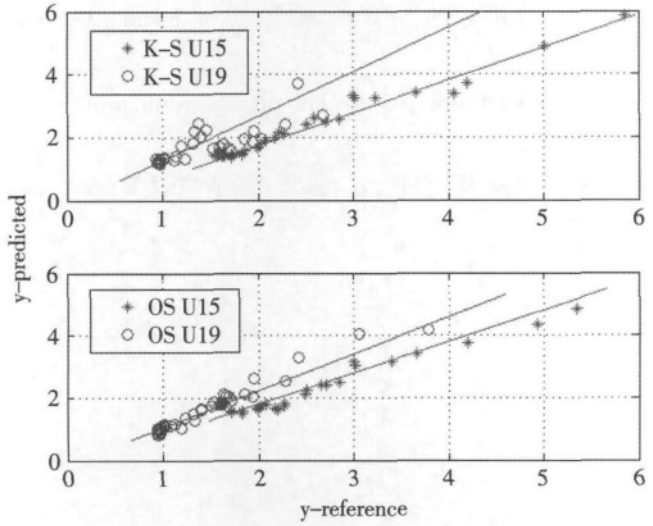

图 4　两种不同样本选择方法的结果对比

表 2　模型更新的预测结果

Methods	Training set		Test set	
	M2[1]	M3[1]	[2] Batch #2	[2] Batch #3
MU-KS	0.0896	0.0770	0.3123	0.0461
MU-OS	0.0863	0.0760	0.2333	0.0371

注：M2 用原训练集和利用样本选择方法选取的 20 个第 2 批样品拟合得到的模型 Model updated with the raw training set and 20 samples of Batch #2；M3 用从第 3 批样品中选出的 20 个样本更新的模型；Model updated using the 20 samples selected from Batch #3；交叉验证均方根误差 RMSECV（Root Mean Square Error of Cross-Validation）；[2] 预测均方根误差，RMSEP（Root Mean Square Error of Prediction）。

得到了提高。在更新后的模型中，因引入了一些新的变异信息，所以更新后模型的预测准确性有了很大的改善。

四、讨　论

本研究以金银花提取物醇沉过程为研究对象，探讨了不同模型传递方法（变量筛选、模型融合、模型更新）对校正模型传递的适用性。结果表明，变量筛选方法 iPLS 和模型融合方法 SPLS 在本研究体系中所建立的定量校正模型在预测新的样本时表现不够稳健。在与 SBC 法结合后，模型预测结果的准确性得到了大幅提高。但因为中药饮片自身的多变性，iPLS-SBC 和 SPLS-SBC 在实际应用时可能会受到一定的限制。对模型进行更新后，因所建立的模型中含有饮片有效组分变化后的信息，所得较正模型的预测准确性得到了一定的提高。在联合新的样本选择方法 OS 后，模型更新的结果趋于稳健。综上所述，模型更新可用于定量校正模型在中药制剂醇沉过程中进行传递。

作为分析化学领域的一个重要分析对象，中药及其制剂和中间体有着自身的一些特点。药材来源变化和药材生长环境、加工炮制条件的变化给校正模型的稳健性提出了很高的要求。对模型传递在中药生产过程在线的应用尚需深入研究。

参考文献

1　Feudale RN, Woody NA, Tan H, *et al*. Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems*, 2002, 64(2):181~192.

2　Pereira CF, Pimentel MF, Galvão RKH, *et al*. A comparative study of calibration transfer methods for determination of gasoline quality

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

parameters in three different near infrared spectrometers. *Analytica Chimica Acta*, 2008, 611(1):41~47.

3 Bouveresse E, Hartmann C, Massart DL, *et al*. Standardization of Near -Infrared Spectrometric Instruments. *Analytical Chemistry*, 1996, 68(6):982~990.

4 Peng J, Peng S, Jiang A, *et al*. Near-infrared calibration transfer based on spectral regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2011, 78(4):1315~1320.

5 Fan W, Liang Y, Yuan D, *et al*. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. *Analytica Chimica Acta*, 2008, 623(1):22~29.

6 Blank TB, Sum ST, Brown SD, *et al*. Transfer of Near-Infrared Multivariate Calibrations without Standards. *Analytical Chemistry*, 1996, 68(17):2987~2995.

7 Du W, Chen ZP, Zhong LJ, *et al*. Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Analytica Chimica Acta*, 2011, 690(1):64~70.

8 Woody NA, Feudale RN, Myles AJ, *et al*. Transfer of Multivariate Calibrations between Four Near-Infrared Spectrometers Using Orthogonal Signal Correction. *Analytical Chemistry*, 2004, 76(9):2595~2600.

9 Norgaard L, Saudland A, Wagner J, *et al*. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy*, 2000, 54:413~419.

10 Cramer JA, Kramer KE, Johnson KJ, *et al*. Automated wavelength selection for spectroscopic fuel models by symmetrically contracting repeated unmoving window partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 2008, 92(1):13~21.

11 Ni W, Brown SD, Man R. Stacked PLS for calibration transfer without standards. *Journal of Chemometrics*, 2011, 25(3):130~137.

12 倪网东, 满瑞林. 叠加多元校正分析. 分析化学, 2010, 38 (3): 367~371.

13 Wold S, Antti H, Lindgren F, *et al*. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 1998, 44(1-2):175~185.

14 Xu B, Wu Z, Lin Z, *et al*. NIR analysis for batch process of ethanol precipitation coupled with a new calibration model updating strategy. *Analytica Chimica Acta*, 2012, 720:22~28.

Application of Orthogonal Selection on the Alcohol Precipitation Processes of *Flos Lonicerae Japonicae* Extractions

Lin Zhaozhou, Xu Bing, Shi Xinyuan, Qiao Yanjiang

(School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing 100102, China)

Abstract: In this article, orthogonal selection (OS) was proposed in the sample selection in order to update the raw calibration model. Models described the relationship between concentrations of chlorogenic acid. And the corresponding spectra were used to monitor the alcohol precipitation processes. A comparison between models built using model updating (MU) and interval partial least squares (iPLS). Stacked partial least squares (SPLS) was executed on three batches of *Flos Lonicerae Japonicae* extractions. It was found that MU can be used to successfully correct batch-wise variations. The performance of model updating with OS outperforms that of Kennard-Stone algorithm.

Keywords: NIR, model transfer, orthogonal selection, stacked partial least square, alcohol precipitation processes

(责任编辑 李沙沙 张志华,责任译审 王 晶)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net