doi：10.3969/j.issn.1004 - 4957.2020.10.011

基于局部加权偏最小二乘的近红外光谱分析方法研究

马力文[1]，郭 拓[1*]，马晋芳[1,2]，史庆龙[2]，肖环贤[3]

（1. 陕西科技大学 电子信息与人工智能学院，陕西 西安 710021；2. 中山大学 南沙研究院，广东 广州 511458；3. 江西保利制药有限公司，江西 赣州 341900）

摘 要：针对近红外光谱分析技术中分析对象非线性现象突出的情况，提出了一种新的模型计算方法——局部加权偏最小二乘法(LWPLS)。以安胎丸为研究对象，采用 LWPLS 算法进行其近红外定量模型的建立，并比较偏最小二乘法(PLS)与 LWPLS 两种算法建立定量模型的精度。结果测得两种算法建立的校正模型中，阿魏酸的模型相关系数(R^2)分别为 0.785 5、0.971 9，预测误差均方根(RMSEP)分别为 0.126 6、0.043 8，相对预测误差(RE)分别为 12.66%、9.18%；洋川芎内酯 A 的 R^2 分别为 0.886 4、0.964 9，RMSEP 分别为 0.114 8、0.077 1，RE 分别为 14.01%、7.81%，显示 LWPLS 算法建立的模型精度更高。研究表明，采用 LWPLS 算法可提高安胎丸定量模型的准确性，具有可推广性和广泛的应用性。

关键词：局部加权偏最小二乘法(LWPLS)；近红外光谱；偏最小二乘法(PLS)；阿魏酸；洋川芎内酯 A

中图分类号：O657.33；TP460.72 **文献标识码**：A **文章编号**：1004 - 4957(2020)10 - 1254 - 06

Study on a Near Infrared Spectroscopy Modeling Method Based on Local Weighted Partial Least Squares

MA Li-wen[1], GUO Tuo[1*], MA Jin-fang[1,2], SHI Qing-long[2], XIAO Huan-xian[3]

(1. School of Electrical Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; 2. Nansha Research Institute, Sun Yat - Sen University, Guangzhou 511458, China; 3. Jiangxi Poly Pharmaceutical Co. Ltd., Ganzhou 341900, China)

Abstract: A new model calculation method, local weighted partial least squares(LWPLS), was proposed to solve the problem of nonlinear phenomena in near infrared spectroscopy. Taking Antai pill as the research object, the quantitative model for Antai pill was established by LWPLS algorithm, and the accuracies of quantitative models established by PLS and LWPLS were compared. Results indicated that, the correlation coefficient(R^2) of PLS and LWPLS algorithms for ferulic acid were 0.785 5 and 0.971 9, the root mean square errors of prediction(RMSEP) were 0.126 6 and 0.043 8, and the relative prediction errors(RE) were 12.66% and 9.18%, respectively, while the R^2 values of PLS and LWPLS algorithms for senkyunolide A were 0.886 4 and 0.964 9, the RMSEP values were 0.114 8 and 0.077 1, and the RE values were 14.01% and 7.81%, respectively, which showed that LWPLS algorithm was more accurate. Therefore, the LWPLS algorithm could improve the accuracy of quantitative model for Antai pill, exhibiting a wide generalization and application potential.

Key words: local weighted partial least squares(LWPLS); near infrared spectroscopy; partial least squares(PLS); ferulic acid; senkyunolide A

现代中药工业生产过程十分复杂，基于指标成分的含量检测方法是中药质量控制的有效手段。因此，如何建立科学、高效的中药质量评价方法是中药现代化长期以来需要解决的难题。近红外(Near infrared，NIR)光谱分析技术作为一种快速、高效的检测手段，在中药生产过程中得到了广泛应用[1]。通过结合数学算法，对其近红外光谱数据与指标成分含量建立相关定量模型，以实现药物的定量分析。偏最小二乘法(PLS)作为目前近红外光谱分析中应用最广泛的线性建模方法[2]，通常用于拟合药物光

收稿日期：2020 - 07 - 15；**修回日期**：2020 - 08 - 20

基金项目：陕西省科技计划项目(2020NY - 172)

* **通讯作者**：郭 拓，博士，研究方向：信号与信息处理、机器学习及应用、光谱分析技术，E - mail：guotuonwpu@126.com

谱与指标成分含量之间的线性关系。而在实际应用过程中，由于受仪器的非线性响应和固体样品颗粒大小的不均匀性等非线性行为的影响，PLS 法难以发挥其作用[3-4]。局部加权偏最小二乘法(LWPLS)是对 PLS 法的有效改进，对每个测试集样本计算其在训练集样本上的权重，并通过加权的样本，对每一测试集样本建立局部的偏最小二乘模型，用多个局部线性模型来逼近非线性过程，该方法能在一定程度上放大光谱与性质之间的相关信息，从而使得预测结果更为准确[5]，故可以很好地解决药物光谱与指标成分含量之间为非线性关系的问题。

因此，本研究提出了基于局部加权偏最小二乘法，并结合相关系数法[6]进行波长优选，建立了对安胎丸进行指标成分定量的模型，同时与 PLS 算法建立的校正模型精度进行比较，旨在为定量模型的建立提供新算法。

1　原理与方法

1.1　局部加权偏最小二乘(LWPLS)算法

LWPLS 的基本思想是当预测某个样本的理化性质指标时，首先计算该样本与训练集之间的相似性，并将此相似度值作为该样本的权重，此时判别样本是否相近的依据通常为马氏距离或欧氏距离[7]。本文选用欧氏距离作为度量工具，即以样本间的欧氏距离作为权重，记为 δ。其建模预测流程如图 1 所示。

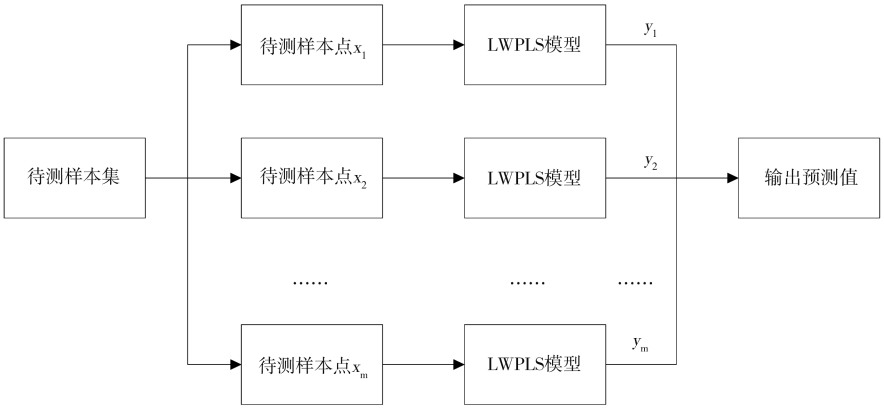

图 1　LWPLS 建模预测

Fig. 1　LWPLS modeling and prediction

假定光谱矩阵和性质矩阵分别为 $\boldsymbol{X}(n\times p)$和 $\boldsymbol{Y}(n\times q)$，其中 n 为样本数目，p 为波长点数，q 为性质数目。对于待测样本 x_m，首先根据下式计算它与其他校正集样本之间的距离：

$$\begin{cases}\delta_i = \sqrt{\sum_{j=0}^{p}(x_{ij}-x_{mj})^2}\\ \delta_i^* = \delta_i/\max(\delta_1,\delta_2,\delta_3,\cdots,\delta_n)\\ f(\delta_i^*) = \exp(-\delta_i^*/(\mathrm{std}(\delta_i^*)))\\ w_i^* = f(\delta_i^*)/\max(f(\delta_1^*),f(\delta_2^*),f(\delta_3^*),\cdots,f(\delta_n^*))\end{cases} \tag{1}$$

当待测样本的距离与校正集样本越大时，权重越小，反之则越大，故权重函数必须是减函数，本文以 $f(\delta_i^*)=\exp(-\delta_i^*/(\mathrm{std}(\delta_i^*)))$作为权重函数[8-9]。根据实际情况该权重函数可有多种表达形式，如：$(1-X^2)^2$ 或者$(1-X^3)^3$[5]。其中，δ_i^* 是标准化后的权重，h 定义为权重函数的形状参数，标准化的权重记为 w_i^*。权重系数为一个方阵，可表示为：

$$\boldsymbol{W}=\mathrm{diag}(w_1^*,\ w_2^*,\ w_3^*,\ \cdots,\ w_n^*) \tag{2}$$

LWPLS 算法思想如图 2 所示，主要包括以下几个步骤：

第一步：计算潜在成分数(ncomp)，设置其初值为 $a=1$，采用留一交叉法，将样本分为 n 组训练集和验证集。在每一组中，用训练集样本建立的模型去预测验证集样本，当预测误差平方和最小时所对应的组号即为 ncomp。

第二步：根据式(1)和(2)计算权重矩阵，并采用 K-近邻(KNN)算法，在训练集中选取与待测样本 x_m 之间欧氏距离最小的 10 个样本点，将这 10 个样本点所表示的集合记为 $\boldsymbol{X}'$，其对应的性质矩阵记为 $\boldsymbol{Y}'$。

第三步：对训练集矩阵以及待测样本进行预处理，计算 $\boldsymbol{X}_a$、$\boldsymbol{Y}_a$、$\boldsymbol{X}_{ma}$

$$\begin{cases}\boldsymbol{X}_a=\boldsymbol{X}'-1_n[\overline{x}_1,\ \overline{x}_2,\ \cdots,\ \overline{x}_P]\\ \boldsymbol{Y}_a=\boldsymbol{Y}'-1_n[\overline{y}_1,\ \overline{y}_2,\ \cdots,\ \overline{y}_Q]\\ \boldsymbol{X}_{ma}=x_m-[\overline{x}_1,\ \overline{x}_2,\ \cdots,\ \overline{x}_P]\end{cases} \tag{3}$$

$$\overline{x}_p=\frac{\sum_{i=1}^{N}w_ix_{ip}}{\sum_{i=1}^{N}w_i},\overline{y}_q=\frac{\sum_{i=1}^{N}w_iy_{iq}}{\sum_{i=1}^{N}w_i} \tag{4}$$

式中，1_n 为 1 的列向量，$p=1,2,\cdots\cdots,P$；$q=1,2,\cdots\cdots,Q$。

第四步：建立局部加权模型

$$t_a=\boldsymbol{X}_a(\boldsymbol{X}_a^T\boldsymbol{W}\boldsymbol{Y}_a\boldsymbol{Y}_a^T\boldsymbol{W}\boldsymbol{X}_a) \tag{5}$$

$$p_a=\frac{\boldsymbol{X}_a^T\boldsymbol{W}t_a}{t_a^T\boldsymbol{W}t_a} \tag{6}$$

模型回归系数：$c_a=\dfrac{\boldsymbol{Y}_a^T\boldsymbol{W}t_a}{t_a^T\boldsymbol{W}t_a}$

待测样本点的得分：

$$\begin{aligned}t_{ma}&=x_{ma}(\boldsymbol{X}_a^T\boldsymbol{W}\boldsymbol{Y}_a\boldsymbol{Y}_a^T\boldsymbol{W}\boldsymbol{X}_a)\\ \boldsymbol{X}_{a+1}&=\boldsymbol{X}_a-t_ap_a^T\\ \boldsymbol{Y}_{a+1}&=\boldsymbol{Y}_a-t_ac_a^T\\ x_{m(a+1)}&=x_{ma}-t_{ma}p_a\end{aligned} \tag{7}$$

若 $a=\mathrm{ncomp}$，则跳至下一步；否则 $a=a+1$，返回第四步。

第五步：计算待测样本值 $\hat{y}_m$

$$\begin{aligned}\boldsymbol{T}&=[t_1,\ t_2,\ \cdots,\ t_{ncomp}]\\ t&=[t_{m1},\ t_{m2},\ \cdots,\ t_{m(ncomp)}]\\ \hat{y}_m&=tc^T\end{aligned} \tag{8}$$

其中 $\boldsymbol{T}$ 为 $\boldsymbol{X}$ 的得分矩阵。

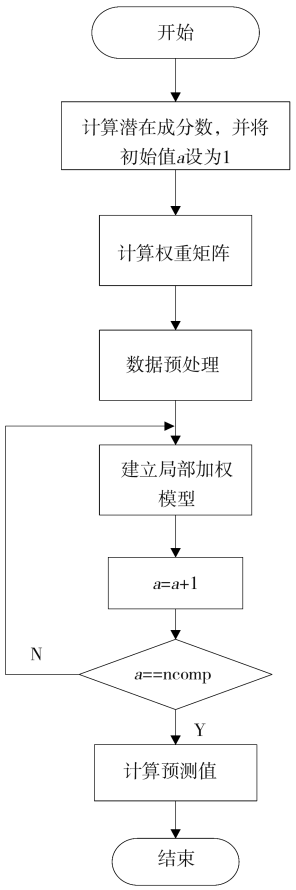

图 2 LWPLS 算法思想

Fig. 2 Algorithm idea of LWPLS

1.2 相关系数法

相关系数法是将校正集光谱矩阵中每个波长对应的吸光度向量 x_j 与性质矩阵中的待测组分性质向量 y_i 进行相关性计算，相关系数越大的波长，其信息量也越多。因此，可结合经验知识给定一个初始阈值，选取相关系数大于该阈值的波长参与建模。然后根据模型的精度调整阈值，从而确定最优的波段。相关系数 r 用下式计算[10-11]。

$$r_j=\frac{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{9}$$

其中，$\overline{x}_j=(\sum_{i=1}^{n}x_{ij})/n$，$\overline{y}=(\sum_{i=1}^{n}y_i)/n$，$j=1,2,\cdots,p$，$i=1,2,\cdots,n$。

2 实验部分

2.1 数据集

本研究参考文献的光谱采集方法[12]，采用 SupNIR1500 近红外光谱仪，应用漫反射模式，设置波

长扫描范围为 1 000～1 800 nm，分辨率为 2 nm，对 3 年生产的共 21 批安胎丸进行 NIR 光谱数据的采集；采用高效液相色谱法(HPLC)梯度洗脱，对 21 批安胎丸中的指标含量进行测定。将测得的安胎丸样本数据按 3∶1∶1 的比例分成训练集、验证集和测试集。首先随机挑选 17 个样本作为验证集，剩下的数据集采用 X－Y 共生矩阵法(Sample set partitioning based on Joint X－Y Distance，SPXY)算法分成训练集和验证集。具体结果见下表，原始数据见原文献。

表 1　安胎丸样本集的分类结果

Table 1　Classification results of Antai pill sample set

Component	Sample set	Number of samples	Maximum	Minimum	Average	Total
Ferulic acid(阿魏酸)	Calibration set	55	0.763 2	0.133 2	0.384 6	97
	Validation set	17	0.727 8	0.130 2	0.416 5	
	Test set	17	0.763 2	0.136 8	0.436 7	
Senkyunolide A(洋川芎内酯 A)	Calibration set	55	1.336 8	0.053 4	0.783 0	97
	Validation set	17	1.113 0	0.244 2	0.753 2	
	Test set	17	1.338 6	0.244 2	0.803 2	

2.2　数据预处理

由于建模过程中，近红外光谱的校正集样本中可能混杂异常光谱，会直接影响到定量模型的精确度，进而影响指标成分的预测结果。因此，本研究首先采用马氏距离法[13]对异常样本进行剔除，此处两种指标成分的马氏距离的阈值分别设为 1.112 6、1.266 0，然后建立模型。

近红外光谱的采集过程中，由于样品本身的状态、表面颗粒的不均匀程度以及仪器操作等因素的影响，导致出现光谱信息重叠及背景干扰的现象[14]。因此，建立模型之前需要对光谱数据进行预处理。在诸多的光谱预处理方法中，标准正态变量变换(SNV)可有效地消除因固体样品表面颗粒大小不均匀、样品表面光散射以及光程变化等引起的光谱噪声[15]。因此，本研究在 LWPLS 建模之前首先采用 SNV 对近红外光谱进行预处理，预处理前后的结果如图 3 所示。

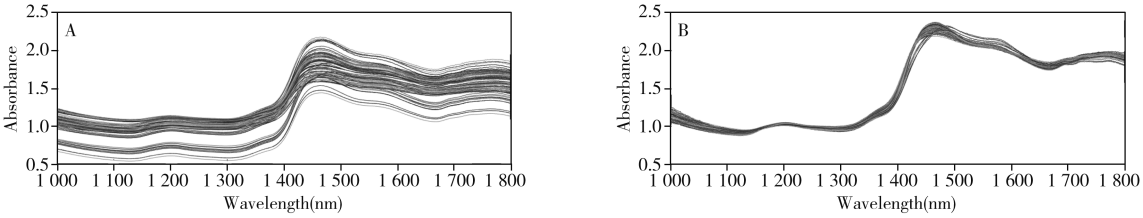

图 3　SNV 预处理前(A)后(B)的光谱图

Fig. 3　Spectra before(A) and after(B) SNV pretreatment

2.3　特征波长的选择

在近红外光谱技术的应用中，通常出现以下现象：由于波长之间有一定的相关信息，导致光谱信息中存在冗余信息，使得计算量较大[16]；由于人工误操作或者仪器自身的噪声，使得光谱某些波段会夹杂噪声，直接导致模型不稳定；或某些波段有可能受外界因素的影响导致吸光度异常，存在局部异常点。因此，NIR 校正模型建立之前进行波长选择不仅可以使计算量减少，更能使参加建模的变量中有效信息增多，进而提高校正模型的预测精度[17]，增强稳健性。本文采用相关系数法进行波长选择，并对比了 PLS 与 LWPLS 算法的建模效果。

2.4　模型的建立与评价

采用 KNN 算法选取 10 个近邻样本，利用 LWPLS 结合相关系数法对安胎丸的训练集进行阿魏酸和洋川芎内酯 A 定量模型的建立。

模型的优劣主要以模型参数，如潜在成分数(ncomp)和权重函数的形状参数(h)、预测误差均方根(RMSEP)、相对预测误差(RE)及模型相关系数(R^2)等作为评价指标，对定量模型的精度进行评估。

3　结果与分析

3.1　LWPLS 建模参数优选

本研究利用 SPXY 算法对安胎丸样本进行训练集、验证集、测试集的划分，SNV 对近红外光谱进

行预处理，相关系数法进行波长选择，并分别结合 LWPLS 与 PLS 对安胎丸进行定量模型的建立。其中 h 是LWPLS 中一个重要的参数，Lesnoff 等[5]认为主成分数的 h 一般在 0～1 之间。因此，本研究将权重函数的 h 范围设定为 0.1～0.9，在不同形状参数下比较模型的 RMSEP。

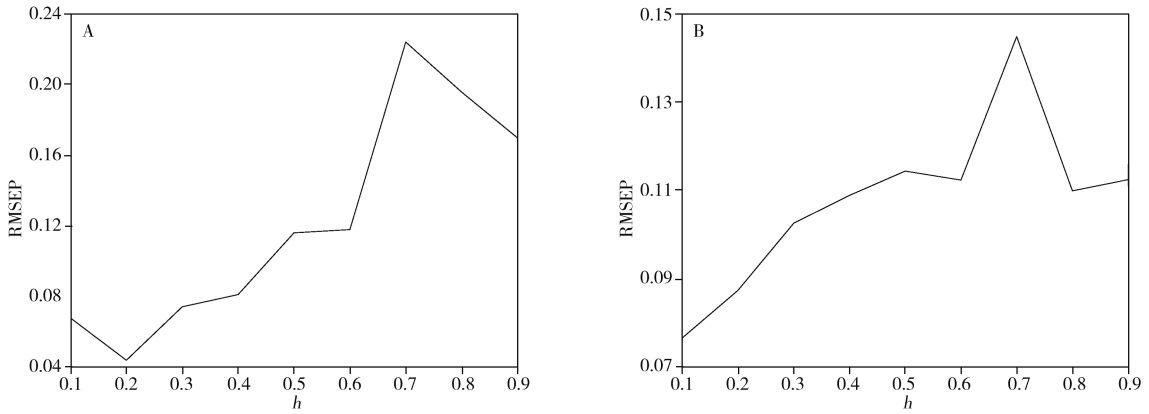

图 4 两种指标成分不同形状参数(h)下 LWPLS 建模的 RMSEP

Fig. 4 RMSEP of LWPLS modeling with two index components and different shape parameter(h)

A：ferulic acid(阿魏酸)；B：senkyunolide A(洋川芎内酯 A)

由图可知，对于指标成分阿魏酸，当 h = 0.2 时，模型效果较好(RMSEP < 0.05)；对于指标成分洋川芎内酯 A，当 h = 0.1 时，模型效果较好(RMSEP < 0.08)。因此，本研究中阿魏酸的 LWPLS 模型的 h 设为 0.2，洋川芎内酯 A 的 LWPLS 模型的 h 设为 0.1。

3.2 两种定量建模方法对模型预测性能影响的对比分析

将阿魏酸和洋川芎内酯 A 两种指标成分的 LWPLS 模型验证结果与线性模型 PLS 的验证结果进行对比。结果显示，两种指标的 LWPLS 模型的预测值与真值更接近1∶1，聚集性也优于 PLS 的结果，且 LWPLS 的结果未出现远离对角线的异常点。阿魏酸采用 PLS 和 LWPLS 建立定量模型的预测结果与真值的线性相关系数分别为 0.886 2、0.985 8(见图 5)；洋川芎内酯 A 采用 PLS 和采用 LWPLS 建立定量模型的线性相关系数分别为 0.941 4、0.982 3。

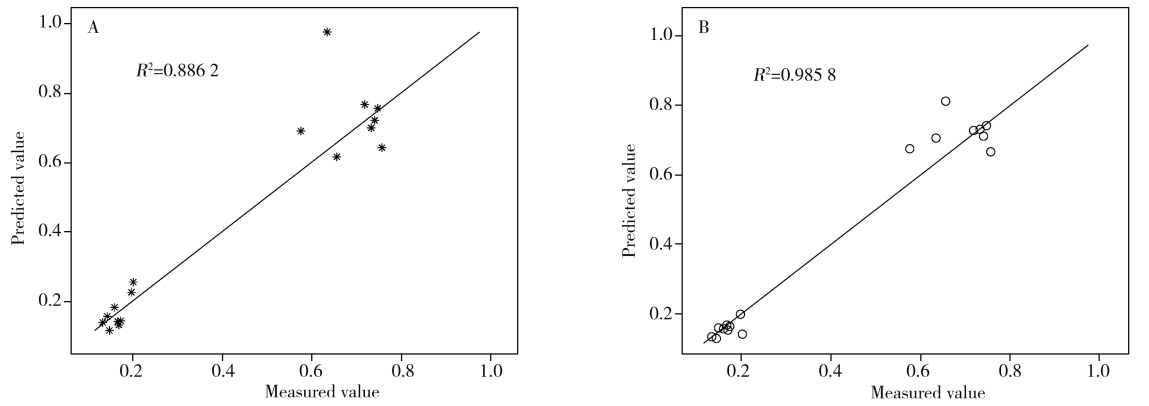

图 5 阿魏酸采用 PLS(A)和 LWPLS(B)建立定量模型的预测结果

Fig. 5 Prediction results of ferulic acid by PLS(A) and LWPLS(B)

两种指标成分的预测结果与原结果的线性相关系数均大于 0.88，其中 LWPLS 方法建立定量模型的预测结果的线性相关系数高于 PLS 方法，且 LWPLS 方法的线性相关系数更接近 1，说明其预测结果更接近真值。

以下将模型参数中选择的波长数目、ncomp、h、RMSEP、RE 与 R^2 进行比较，结果见表 2。由表 2 可以得出，采用 LWPLS 方法建立的模型，其 R^2 分别由 0.785 5、0.886 4 上升至 0.971 9、0.964 9，RMSEP 分别由 0.126 6、0.114 8 降至 0.043 8、0.077 1，RE 也分别从 12.66%、14.01% 降低至 9.18%、7.81%。数据表明：LWPLS 方法使得模型的准确性和稳定性优于 PLS 方法，且模型的指标参数得到显著提高。

表 2　安胎丸中指标成分定量模型参数的比较

Table 2　Comparison of quantitative model parameter values of index components in Antai pills

Component	Method	Number of selected wavelengths	h	ncomp	RE(%)	R^2	RMSEP
Ferulic acid	PLS	763	–	8	12.66	0.785 5	0.126 6
	LWPLS	763	0.2	8	9.18	0.971 9	0.043 8
Senkyunolide A	PLS	771	–	9	14.01	0.886 4	0.114 8
	LWPLS	771	0.1	9	7.81	0.964 9	0.077 1

4　结　论

本文研究的 LWPLS 算法，是针对每一测试集样本建立局部的 PLS 模型，将多个局部线性模型组合，其整体上为一个非线性模型。该算法成功应用于安胎丸指标成分的建模，并解决了线性建模方法 PLS 对非线性关系无法准确拟合的问题，提高了模型的性能与预测精度。该方法有望以较小的计算代价完成非线性模型的建立，并应用于实际生产过程的在线质量监测。

参考文献：

[1] Chu X L, Shi Y Y, Chen P, Li J Y, Xu Y P. *J. Instrum. Anal.*(褚小立，史云颖，陈瀑，李敬岩，许育鹏. 分析测试学报), **2019**, 38(5): 603 – 611.

[2] Zhou Z L, Li J, Huang S Q, Tian S H, Liu Y J, Lu L, Zhang Y, Huang Y S, Wang X C. *Chem. Ind. Eng. Prog.*(周昭露，李杰，黄生权，田淑华，刘玉娇，鲁亮，张扬，黄延盛，王学重. 化工进展), **2016**, 35(6): 1627 – 1645.

[3] Fan L H, Fan W X, Wei Z Q, Tan C Q, Wang J L, Wei D N, Wu B, Wu C J, Huang Y L. *Chin. J. Exp. Tradit. Med. Formulae*(范林宏，范文翔，韦志强，谭超群，王蛟龙，魏大能，吴博，吴纯洁，黄永亮. 中国实验方剂学杂志), **2019**, 25(24): 205 – 210.

[4] Zhang Y, Wang D, Li X, Zhang L X, Zhang W, Ding X X, Zhang Q, Li P W. *Food Saf. Qual. Detect. Technol.*(张勇，王督，李雪，张良晓，张文，丁小霞，张奇，李培武. 食品安全质量检测学报), **2018**, 9(23): 6161 – 6166.

[5] Lesnoff M, Metz M, Roger J M. *J. Chemom.*, **2020**, 34(5): 1 – 13.

[6] Liu J S. *Study on Quality Evaluation Method of Baphicacanthiscusiae Using Near Infrared Spectroscopy*. Guangzhou: Guangzhou University of Chinese Medicine(刘家水. 近红外光谱技术应用于南板蓝质量评价方法的研究. 广州：广州中医药大学), **2013**.

[7] Yan Y, Zhang H G, Lu J G, Shi Y Z, Chen J S. *Comput. Appl. Chem.*(鄢悦，张红光，卢建刚，施英姿，陈金水. 计算机与应用化学), **2017**, 34(5): 351 – 355.

[8] Pan B, Jin H P, Yang B, Feng L H, Chen X G. *Inf. Control*(潘贝，金怀平，杨彪，冯丽辉，陈祥光. 信息与控制), **2019**, 48(2): 217 – 223, 231.

[9] Cleveland W S, Devlin S J. *J. Am. Stat. Assoc.*, **1988**, 83(403): 596 – 610.

[10] Chen B, Wang H, Lin S, Zhao J W. *Trans. Chin. Soc. Agric. Eng.*(陈斌，王豪，林松，赵杰文. 农业工程学报), **2005**, (7): 99 – 102.

[11] Shen Y X, Zhao Q J. *Technol. Innov.*(申永祥，赵秋菊. 科技与创新), **2014**, (14): 113 – 115.

[12] Ma J F, Wang X L, Xiao X, Peng Y, Ge F H. *World Sci. Technol. – Mod. Tradit. Chin. Med.*(马晋芳，王雪利，肖雪，彭银，葛发欢. 世界科学技术 – 中医药现代化), **2018**, 20(5): 651 – 659.

[13] Liu C L, Hu Y J, Wu S N, Sun X R, Dou S L, Miao Y Q, Dou Y. *J. Food Sci. Technol.*(刘翠玲，胡玉君，吴胜男，孙晓荣，窦森磊，苗雨晴，窦颖. 食品科学技术学报), **2014**, 32(5): 74 – 79.

[14] Li J J, Wu J H, Zhang H B. *Food Saf. Qual. Detect. Technol.*(李佳洁，吴建虎，张海波. 食品安全质量检测学报), **2017**, 8(8): 3037 – 3043.

[15] Luo X, Wu F X, Xie H G, Zhu Y S, Zhang J F, Xie H A. *Spectrosc. Spectral Anal.*(罗曦，吴方喜，谢鸿光，朱永生，张建福，谢华安. 光谱学与光谱分析), **2016**, 36(3): 697 – 701.

[16] Xie Y, Zhou C, Tu C, Zhang Z L, Wang J F. *Chin. J. Anal. Chem.*(谢越，周成，涂从，张祖亮，汪建飞. 分析化学), **2017**, 45(3): 363 – 368.

[17] Lu W Z, Yuan H F, Xu G T, Qiang D M. *Beijing*: *China Petrochemical Press*(陆婉珍，袁洪福，徐广通，强冬梅. 北京：中国石化出版社), **2000**.

(责任编辑：龙秀芬)