第34卷，第12期
2014年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 34, No. 12, pp3267-3272
December, 2014

校正集选择方法对于积雪草总苷中积雪草苷NIR定量模型的影响

詹雪艳，赵 娜，林兆洲，吴志生，袁瑞娟，乔延江*

北京中医药大学中药学院，北京 100102

摘 要 近红外光谱定量分析中，采用合适的校正集选择方法是建立预测性能良好的近红外定量模型的关键技术之一。校正集选择方法有RS法、CS法、KS法和SPXY法等，但是对以上校正集选择方法缺乏系统地比较。本文以积雪草总苷中积雪草苷NIR定量模型为载体，对NIR定量模型的7个评价指标进行分类和筛选，比较了CS法、KS法和SPXY法三种校正集选择方法对NIR定量模型的准确性和稳健性两类评价指标的影响。结果表明，SPXY法与CS法、KS法选择校正集样本后所建近红外模型的RPD和RSEP两个准确性评价指标存在显著性差异，模型的稳健性评价指标RMSECV和|RMSEP－RMSEC|不存在显著性差异。因此，建立积雪草总苷近红外光谱的积雪草苷偏最小二乘定量模型时，SPXY校正集选择方法能显著提高该定量模型的预测准确度，但对模型稳健性的评价指标没有显著影响，以上结论为中药固体体系建立近红外定量模型确定校正集选择方法提供参考。

关键词 近红外；校正集选择；偏最小二乘回归；漫反射

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2014)12-3267-06

引 言

近红外光谱(near infrared spectroscopy，NIRS)分析技术具有快速无损、操作简单、分析成本低等优点，在中药材[1]、中成药[2]、中药生产过程[3]以及药物[4,5]的质量评价中已得到了广泛应用。

近红外光谱定量分析的前提是采用化学计量学方法建立光谱特征与待测物质化学测量值之间的定量校正模型，常用的建模方法为偏最小二乘回归法(partial least square regression，PLSR)[1-5]。选择代表性的校正集样本是获得预测性能良好的近红外定量模型的关键技术之一，合适的校正集选择方法能增强模型的预测能力[6]。校正集选择方法有随机抽样(random sampling，RS)法[7,8]、常规选择(conventional selection，CS)[5]法、Kennard-Stone(KS)法[1-3,9]、Sample set Portioning based on joint x-y distance (SPXY)法[10,11]等，但对校正集选择方法进行系统地比较以及其对近红外定量模型影响的研究少见报道。因此，本研究拟以积雪草总苷中积雪草苷NIR定量模型为载体，比较校正集样本选择方法对NIR定量模型的准确性和稳健性两类评价指标的影响，选择合适的校正集样本，建立预测性能好的积雪草苷近红外定量校正模型，实现积雪草苷含量的准确预测和积雪草总苷质量的有效监控。

1 校正集选择方法

1.1 RS法

RS法以随机选取一定数量样本组成校正集[7,8]。该校正集选择方法简单，不需要对数据进行排序、挑选或计算，但每次随机挑选校正集样本可能存在很大差异，不能保证所选样本的代表性和模型的外推能力。

1.2 CS法

CS法根据样品的某些已知因素对数据进行挑选，如产地、厂家、生产批号等，挑选建模样本时应尽可能地增大这些因素的变异，得到代表性尽可能好的校正集。当样本的化学测量值已知时，可按照组分的化学测量值进行挑选，选择

收稿日期：2013-11-27，修订日期：2014-03-15

基金项目："重大新药创制"国家科技重大专项(2010ZX09502-002)，北京市青年英才计划项目(YETP0815)，北京中医药大学青年教师专项计划项目(2012-QNJSZX009)资助

作者简介：詹雪艳，1978年生，北京中医药大学中药学院博士 e-mail：snowzhan@126.com

*通讯联系人 e-mail：yjqiao@263.net

那些分布在两端即化学测量值最高或最低的样本作为校正集样本[5]。通常将所有样本的化学测量值按大小排序后，以校正集和验证集样本数的比例按顺序将样本依次分配到校正集和验证集，而且每次分配的验证集样本的化学测量值均在校正集样本化学测量值的范围内。在样本量不多、大部分变量参数可知的情况下，该方法选择的校正集样本代表性较好，模型的预测能力达到要求。但是该校正集选择方法带有较大的主观性，当样本量较大时该方法费时费力，而且所选出的校正集代表性差，所建模型的预测性能差。

1.3 KS 法

KS 法[9]是将所有样本都看作校正集的候选样本，首先选择欧氏距离或马氏距离最远的两个样本对进入校正集，计算剩余的候选样本中每个样本到校正集中每个已选样本的距离，找出最小距离值样本和最大距离值样本，加入到校正集中，重复此步骤，直至校正集样本数目满足要求为止。在建立近红外定量模型过程中 KS 法是基于各个样本的近红外光谱数据来计算两样本间的距离，即 $d_{x(p,q)}$ 是基于近红外光谱数据 x 计算 p，q 两样本间的距离，其距离计算公式见式(1)，所选出的校正样本能够均匀地覆盖整个样本集实验区域，所建模型的预测能力较好。但是，该方法需要进行数据转换和计算两两样本空间距离，计算量大，需采用计算机识别。

$$d_x(p,q) = \sqrt{(x_p - x_q)^2} \quad p,q \in [1,n] \tag{1}$$

1.4 SPXY 法

SPXY 法是在 KS 法的基础上发展而来的，在样本间距离的计算时将近红外光谱数据变量 x 和化学测量值变量 y 同时考虑在内，p，q 两样本间距离 $d_{xy(p,q)}$ 能有效地覆盖近红外光谱数据 x 的多维向量空间和化学测量值 y 空间，基于样本间距离 $d_{xy(p,q)}$ 进行校正样本的选择能改善所建 NIR 定量模型的预测能力[11]。

SPXY 法逐步选择校正样本的过程与 KS 法相似，p，q 两样本间的距离 $d_{xy(p,q)}$ 是在 $d_{x(p,q)}$ 基础上引入了 $d_{y(p,q)}$，即基于化学测量值计算的 p，q 两样本间的距离，其计算公式见式(2)。同时为了确保样本在 x 和 y 空间具有相同的权重，将 $d_{x(p,q)}$ 和 $d_{y(p,q)}$ 分别除以它们在数据集中的最大值，在 x 和 y 空间 p 和 q 两样本的标准化距离公式见式(3)。

$$d_y(p,q) = \sqrt{(y_p - y_q)^2} \quad p,q \in [1,n] \tag{2}$$

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max_{p,q\in[1,n]} d_x(p,q)} + \frac{d_y(p,q)}{\max_{p,q\in[1,n]} d_y(p,q)} \quad p,q \in [1,n] \tag{3}$$

2 实验部分

2.1 样本

积雪草总苷样本来自市场上流通的广西、陕西、江苏三省 11 个提取物厂家的 66 个批次的样品，样品粒径均小于 80 目的干燥粉末，密封包装。

2.2 近红外光谱的采集和积雪草苷含量的测定

仪器：Antaris 傅里叶变换近红外光谱仪(美国 Thermo Nicolet 公司)配有 InGaAs 检测器、积分球漫反射采样系统、Result 操作软件和 TQ Analyst 光谱分析软件；Shimadzu LC-20AT 高效液相色谱仪(日本岛津公司)配有 LC-20AT 高压泵、DGU-20A$_5$ 在线脱气机、SIL-20A 自动进样器、CTO-10AS 柱温箱、SPD-20A 紫外检测器；色谱柱为 Agilent TC-C18 (4.6×250 mm，5 μm)；电子天平 BS 110S (北京赛多利斯仪器系统有限公司)。

试剂：积雪草苷(中国食品药品检定研究院)，β-环糊精(天津光复化工研究所)，乙腈(Fisher 赛默飞世尔科技有限公司)。

66 份积雪草总苷样品近红外光谱采集条件：积分球漫反射方式，以内置背景为参照，扫描范围 4 000～10 000 cm^{-1}，分辨率 8 cm^{-1}，扫描次数 32，增益 2，透镜 Empty，每个样品重复测定 3 次，以其平均光谱作为积雪草总苷各样本的近红外光谱，积雪草总苷的原始近红外光谱见图 1。

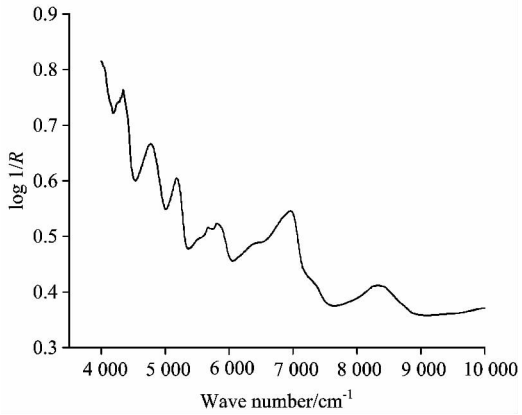

Fig. 1 NIR diffuse reflectance spectra of Centella Total Glucosides

积雪草苷含量的测定：精密称取积雪草总苷约 7.5 mg，至 10 mL 容量瓶中，乙腈-水(1∶3)定容，摇匀，称定质量，超声 20 min，用乙腈-水(1∶3)补足质量，摇匀，0.45 μm 微孔滤膜过滤，即得供试品溶液。色谱条件为：Agilent TC-C$_{18}$ (4.6 mm×250 mm，5 μm)色谱柱；乙腈-2 mmol·L^{-1} β-环糊精溶液(25∶75)为流动相，流速为 1 mL·min^{-1}；柱温 25 ℃；紫外检测波长 205 nm，进样量 10 μL，采用外标标准曲线法计算积雪草苷的含量。

2.3 校正集选择和光谱预处理

66 个积雪草总苷样本，分别采用 RS 法、CS 法、KS 法、SPXY 法选择其中 2/3 样本作为校正集，剩余 1/3 样本为验证集。除 CS 法外，其余三种校正样本的选择方法均在 MATLAB (Mathwork Inc)环境中实现，SPXY 算法原代码来自 PLS-Toolbox 2.1，其他相关计算程序均自行编写。

利用 TQ Analyst 8 软件，以 PLSR 为建模方法，采用多元散射校正(multivariate scatter correction，MSC)和标准正则变换(standard normal variate，SNV)等散射校正法、一阶导数(1st derivative，1D)和二阶导数(2 nd derivative，2D)等光谱导数法及 Savitzky-Golay 平滑(SG)和 Norris 导数(Norris derivative，ND)平滑等光谱数据的预处理方法，提高近红外分析信号的信噪比，筛选最佳光谱预处理方法及其合适的

潜变量因子数，建立积雪草苷近红外定量模型。

2.4　近红外定量模型的评价指标

近红外定量模型通常用样本的预测值和化学测量值相关系数(correction coefficient，R)[4-6](校正集中样本的预测值和化学测量值相关系数为 R_c，预测集中样本的预测值和化学测量值相关系数为 R_p)、交叉验证均方根误差(root mean squared error of cross-validation，RMSECV)[4,6]、校正集均方根误差(root mean squared error of calibration，RMSEC)[5]、验证集均方根误差(root mean squared error of prediction，RMSEP)[4,6]、验证集相对误差(relative standard error of prediction，RSEP)[12]及验证集标准偏差与标准误差的比值(ratio of standard deviation and standard error of prediction，RPD)[13]等指标来评价所建近红外定量模型的性能，并用以上模型评价指标来优化建模方法和建模过程中参数的选择[15,16]。预测集中各样本模型预测值和化学测量值相关系数 R_p 和 RPD 越大，RMSECV，RMSEC，RMSEP 和 RSEP 的值越小，说明所建模型性能越好。

$$R_p=\sqrt{1-\frac{\sum_{i=1}^{m}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{m}(\bar{y}_i-y_i)^2}}\ \text{或}\ R_c=\sqrt{1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i-y_i)^2}} \tag{4}$$

$$\text{RMSEP}=\sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i-y_i)^2}{m}}\ \text{或}$$

$$\text{RMSEC}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}} \tag{5}$$

$$\text{RSEP}=\frac{\sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i-y_i)^2}{m-1}}}{\bar{y}_p} \tag{6}$$

$$\text{RPD}=\sqrt{\frac{\sum_{i=1}^{m}(\bar{y}_i-y_i)^2}{\sum_{i=1}^{m}(\hat{y}_i-y_i)^2}} \tag{7}$$

其中，n 和 m 分别是校正集和预测集的样本数，y_i 是样本 i 的化学测量值，$\hat{y}_i$ 为样本 i 的模型预测值，$\bar{y}_i$ 是验证集化学测量值的平均值。R_p 和 R_c 分别为验证集或校正集的模型预测值和化学测量值间的相关系数。k 折交叉验证的 RMSECV 是将所有样本分成 k 个子集，每个子集做一次预测集，交叉验证重复 k 次，k 次的预测集均方根误差的平均值为其交叉验证均方根误差。

3　结果与讨论

3.1　四种校正集选择方法下校正集和验证集化学测量值的比较

采用 RS 法、CS 法、KS 法和 SPXY 法选择校正集样本，校正集和验证集化学测量值数值的分布见表 1。RS 法完全随机选取校正集样本，所选取的校正集样本中化学测量值积雪草苷含量范围为 13.73%～42.84%，不能完全覆盖验证集样本中积雪草苷含量 14.73%～45.82%，不能保证所选取校正集样本的代表性。而且计算机产生的随机数矩阵不具有重复性，每次随机挑选的校正集存在很大差异。除 RS 法外，CS 法、KS 法和 SPXY 法均属于有指导地选择校正集样本，CS 法是以化学测量值的大小为指导，KS 法是以 NIR 光谱空间样本间距离远近为指导，SPXY 法以 NIR 光谱空间和化学测量值空间中样本间距离远近为指导。表 1 中以上三种有指导的校正集样本选择方法所选择的校正集样本化学测量值的范围覆盖验证集样本的化学测量值，而且验证集化学测量值的标准偏差均小于校正集化学测量值的标准偏差，以上三种方法所选择的校正样本均具有一定的代表性。

Table 1　Content range of asiaticoside in calibration and validation sets($g\cdot g^{-1}$)

校正样本选择方法	样本集	样本数	积雪草苷含量范围/%	平均值/%	标准偏差/%
RS	校正集	44	13.73～42.84	31.29	10.26
	验证集	22	14.73～45.82	32.51	12.47
CS	校正集	44	13.73～45.82	31.88	11.46
	验证集	22	14.28～42.84	31.35	10.15
KS	校正集	44	13.73～45.82	29.86	11.98
	验证集	22	15.60～42.51	35.37	7.57
SPXY	校正集	44	13.73～45.82	29.43	11.51
	验证集	22	14.77～42.76	36.23	8.28

3.2　不同校正集样本选择方法下所建 NIR 定量模型评价指标的比较

3.2.1　模型评价指标的筛选

基于验证集的模型评价指标 R_p，RMSEP，RPD 和 RSEP 均为表征所建模型预测准确性的评价指标[12-15]。由式(4)和式(7)可推导出 $1/\text{RPD}=\sqrt{1-R_p^2}$，表明 RPD 与 R_p 正相关，R_p 越大，相应的 RPD 增大。R_p 和 RPD 两个评价指标中，RPD 可视为将 $R_p(0,1)$ 投影到 RPD$(1,\infty)$的空间，RPD 数值变化更灵敏。用 $\frac{d(\text{RPD})}{\text{RPD}}$ 和 $\frac{dR_p}{R_p}$ 表征 RPD 和 R_p 两评价指标的灵敏度，式(8)表明，当 $R_p^2\geqslant 0.5$ 时，$\frac{d(\text{RPD})}{\text{RPD}}\geqslant\frac{dR_p}{R_p}$，RPD 评价指标具有更高灵敏度。通常 NIR 定量模型的 $R_p^2\geqslant 0.5$，因此采用 RPD 来衡量 NIR 定量模型预测值和实测值间的相关性，从而表征 NIR 定量模型的预测准确度。由式(5)和式(6)知 RMSEP 是验证集预测的绝对误差，RSEP 是验证集预测的相对误差，相对误差能衡量预测误差对测量结果的影响，更能表征所建模型预测的准确度。因此，可选用 RPD 和 RSEP 两个基于验证集的模型评价参数来表征模型的预测准确度。

$$\frac{d(\text{RPD})}{\text{RPD}}=\frac{R_p^2}{1-R_p^2}\frac{dR_p}{R_p} \tag{8}$$

模型评价参数 RMSECV 是基于校正集内部样本对模型进行交叉验证，一定程度上评价所建模型的稳健性，该数值小，表明所建模型的稳健性好。同时，RMSEC 值与 RMSEP 值相差较大，会出现所建模型“欠拟合”和“过拟合”的现象。当 RMSEC 值与 RMSEP 值接近，表明所建模型不仅对校正集样本数据有好的预测准确度，而且适用于验证集样本，该模型具有较好的拓展性和稳健性。因此，|RMSEP－RMSEC

|值趋近 0，表明模型具有好的稳健性。

性能良好的 NIR 定量模型应该将准确性和稳健性两方面指标组合起来进行模型评价，如将 R_p，RSEP 结合 RMSECV 来优化模型和评价模型。根据以上对模型的准确性和稳健性两方面评价指标的分析，RPD 相对于 R_p 具有更高的灵敏度，因此，应选择模型的预测准确性评价指标 RPD，RSEP 结合模型的稳健性评价指标 RMSECV，|RMSEP－RMSEC|来评价模型的预测准确度和稳健性。

3.2.2 不同校正集样本选择方法下所建模型评价指标的比较

RS 法、CS 法、KS 法和 SPXY 法四种校正集样本选择方法中，RS 法由计算机随机挑选校正集样本，具有随机性和不重复性，所建 NIR 定量模型及其评价参数也具有不确定性，无法保证所建模型的稳健性。CS 法、KS 法和 SPXY 法所选择的校正集样本具有确定性，此三种方法在不同的光谱预处理方法下采用 PLSR 所建 NIR 定量模型的七个模型评价指标见表 2。

Table 2 Evaluation indexes of PLSR models with three algorithms for calibration set selection

校正集选择方法	光谱预处理方法	RMSEC	RMSEP	R_c	R_p	RSEP	RPD	RMSECV
CS	Raw	1.95	3.61	0.985 0	0.947 7	0.117 9	2.75	3.64
	MSC	2.16	4.42	0.981 7	0.933 9	0.144 3	2.24	3.52
	SNV	2.05	4.27	0.983 6	0.937 6	0.139 4	2.32	3.34
	MSC+2D+SG(9, 2)	1.07	3.02	0.995 2	0.953 9	0.098 6	3.28	2.99
	SNV+2D+SG(9, 2)	1.07	3.03	0.995 5	0.953 6	0.098 9	3.27	3.00
	MSC+2D+ND(3, 1)	1.19	3.04	0.994 4	0.953 1	0.099 3	3.26	3.02
	SNV+2D+ND(3, 1)	1.19	3.04	0.994 4	0.953 1	0.099 3	3.26	3.02
	MSC+1D+SG(11, 3)	2.81	3.10	0.968 7	0.951 1	0.101 2	3.20	3.28
	SNV+1D+SG(11, 3)	2.82	3.12	0.968 5	0.950 1	0.101 9	3.18	3.30
	MSC+1D+ND(3, 1)	2.81	3.10	0.968 7	0.951 0	0.101 2	3.20	3.30
	SNV+1D+ND(3, 1)	2.82	3.12	0.968 5	0.950 3	0.101 9	3.18	3.28
KS	Raw	2.36	1.96	0.980 0	0.966 4	0.056 7	3.77	4.60
	MSC	1.60	1.86	0.990 8	0.972 0	0.053 8	3.98	4.32
	SNV	2.29	1.56	0.981 2	0.977 5	0.045 1	4.74	3.94
	MSC+2D+SG(9, 2)	0.88	1.81	0.997 3	0.972 0	0.052 4	4.09	3.87
	SNV+2D+SG(9, 2)	0.88	1.81	0.997 2	0.971 8	0.052 4	4.09	3.88
	MSC+2D+ND(3, 1)	1.00	1.85	0.996 4	0.970 4	0.053 5	4.00	3.86
	SNV+2D+ND(3, 1)	1.01	1.86	0.996 4	0.970 2	0.053 8	3.98	3.88
	MSC+1D+SG(11, 3)	0.96	2.09	0.996 7	0.963 4	0.060 5	3.54	3.81
	SNV+1D+SG(11, 3)	0.94	2.08	0.996 8	0.962 4	0.060 2	3.56	3.78
	MSC+1D+ND(3, 1)	1.09	2.07	0.995 8	0.963 9	0.059 9	3.58	3.82
	SNV+1D+ND(3, 1)	1.06	2.10	0.996 0	0.961 7	0.060 8	3.52	3.80
SPXY	Raw	1.46	1.35	0.991 8	0.986 0	0.038 1	5.99	4.38
	MSC	2.63	1.29	0.973 0	0.988 2	0.036 4	6.27	4.10
	SNV	2.39	1.19	0.977 7	0.989 6	0.033 6	6.79	4.19
	MSC+2D+SG(9, 2)	1.01	1.24	0.996 1	0.988 2	0.035 0	6.52	3.47
	SNV+2D+SG(9, 2)	1.02	1.24	0.996 0	0.988 1	0.035 0	6.52	3.48
	MSC+2D+ND(3, 1)	1.17	1.25	0.994 7	0.988 0	0.035 3	6.47	3.43
	SNV+2D+ND(3, 1)	1.18	1.25	0.994 6	0.988 0	0.035 3	6.47	3.45
	MSC+1D+SG(11, 3)	1.24	1.60	0.994 0	0.983 8	0.045 2	5.05	3.79
	SNV+1D+SG(11, 3)	1.43	1.28	0.992 0	0.989 5	0.036 2	6.32	3.81
	MSC+1D+ND(3, 1)	1.40	1.58	0.992 5	0.984 7	0.044 6	5.12	3.76
	SNV+1D+ND(3, 1)	1.36	1.65	0.992 8	0.983 5	0.046 6	4.90	3.77

Note: "Raw" represent the raw spectra without data pre-processing

对三种校正集样本选择方法下所建 NIR 定量模型的评价指标 RMSEP，RSEP，R_p 和 RPD 进行配对 t 检验[17]，结果表明在相同光谱数据预处理条件下 SPXY 与其他校正集样本选择方法下所建 NIR 定量模型的 RMSEP，RSEP，R_p 和 RPD 四个评价参数均存在显著性差异，显著性水平均小于 0.001。KS 法与 CS 法所建 NIR 定量模型的 R_p，RMSEP 和 RSEP 评价参数显著性水平均小于 0.001，模型评价参数 RPD 存在显著性差异，显著性水平小于 0.05 大于 0.001。三种校正集样本选择方法下所建 NIR 定量模型的评价参数 RMSECV，不存在显著性差异，而且模型的|RMSEP－RMSEC|值相当，不存在显著性差异，由此可见，校正集样本选择方法对 NIR 定量模型的预测准确度有较大的影响，对 NIR

定量模型的稳健性没有显著影响。

SPXY 法基于 NIR 光谱变量和目标变量空间上样本间距离来选择校正集样本，具有很好的代表性，所建 NIR 定量模型的 RPD 值为 4.90～6.79，RSEP 是三组 RSEP 数值中最小的，其范围为 0.0336～0.0466，表明所建 NIR 定量模型的预测准确度好，均大于 RPD 评价参数的阈值 3.0[18]，且 RSEP 小于 10%[12]，所建的模型预测准确度符合快速定量的要求。

因此，基于该 66 个积雪草总苷样本，选取 SPXY 作为校正集样本选择方法，采用多元散射校正(MSC)结合二阶导数光谱(2D)和 Norris 导数(ND)平滑的光谱预处理方法，建立积雪草苷的近红外偏最小二乘回归模型，所得模型的

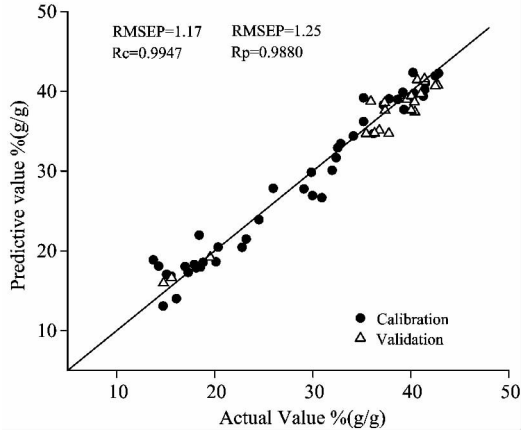

Fig. 2　Correlation plot between NIR measured values and reference values of asiaticoside

RMSECV 值为 3.43 比较小，RMSEC 和 RMSEP 接近，分别为 1.17 和 1.25，RSEP 和 RPD 分别为 0.0352 和 6.47，所建模型积雪草苷含量预测值和高效液相色谱实测值相关性见图 2。该积雪草苷 NIR 定量模型具有好的预测准确度和稳健性，能用于积雪草总苷中积雪草苷含量的快速预测。

4　结　论

以积雪草总苷中积雪草苷 NIR 定量模型为研究对象，系统地研究了 RS 法、CS 法、KS 法、SPXY 法四种校正集选择方法对 NIR 定量模型的影响。创新性工作及其主要结论如下：

(1)对 7 个近红外定量模型的评价指标进行了分类，筛选出 RSEP 和 RPD 作为近红外定量模型的准确性评价指标，RMSECV 和|RMSEP－RMSEC|作为近红外定量模型的稳健性评价指标，从模型的准确性和稳健性两方面进行评价。

(2)采用配对 t 检验对 CS 法、KS 法和 SPXY 法三种校正集选择方法下所建近红外定量模型的准确性和稳健性两类评价指标进行比较。基于积雪草总苷近红外漫反射光谱建立积雪草苷的偏最小二乘回归模型，校正集选择方法对 NIR 定量模型的预测准确度有较大的影响，SPXY 法与其他校正集样本选择方法所建模型的准确性评价指标 R_p，RMSEP，RSEP 和 RPD 存在显著性差异。在积雪草总苷固体体系建立积雪草苷的 NIR 定量模型，三种校正集选择方法中 SPXY 法选择校正集样本所建立的 NIR 定量模型预测准确度最高。三种校正集选择方法下模型稳健性的评价指标 RMSECV 和|RMSEP－RMSEC|值相当，不存在显著性差异，表明校正集选择方法对 NIR 定量模型的稳健性没有显著影响。

References

[1] Wu Zhisheng, Sui Chenglin, Xu Bing, et al. Journal of Pharmaceutical and Biomedical Analysis, 2013, 77: 16.
[2] Li W L, Xing L H, Fang L M, et al. Journal of Pharmaceutical and Biomedical Analysis, 2010, 53: 350.
[3] Xu Bing, Wu Zhisheng, Lin Zhaozhou, et al. Analytica Chimica Acta, 2012, 720: 22.
[4] Mantanus J, Ziemons E, Lebrun P, et al. Talanta, 2010, 80(5): 1750.
[5] Xiang D, Konigsberger M, Wabuyele B, et al. Analyst, 2009, 134(7): 1405.
[6] Nisha Shetty, Åsmund Rinnan, René Gislum. Chemometrics and Intelligent Laboratory Systems, 2012, 111(1): 59.
[7] Asikainen A, Ruuskanen J, Tuppurainen K. Environmental Science & Technology, 2004, 38(24): 6724.
[8] Tong W D, Hong H X, Fang H, et al. Journal of Chemical Information and Computer Sciences, 2003, 43(2): 525.
[9] Kennard R W, Stone L A. Technometrics, 1969, 11: 137.
[10] Zhu X, Shan Y, Li G Y, et al. Spectrochim Acta A Mol Biomol Spectrosc., 2009, 74(2): 344.
[11] Galvao R K H, Araujo M C U, José G E, et al. Talanta, 2005, 67(4): 736.
[12] Blanco M, Peguero A. J. Pharm. Biomed. Anal., 2010, 52(1): 59.
[13] Joubert E, Manley M, Botha M. J. Agric. Food Chem., 2006, 54(15): 5279.
[14] Chodak M. Journal of Plant Nutrition and Soil Science, 2011, 174(5): 702.
[15] Lin Hao, Chen Quansheng, Zhao Jiewen, et al. Journal of Pharmaceutical and Biomedical Analysis, 2009, 50(5): 803.
[16] Herrmann S, Mayer J, Michel K, et al. Journal of Near Infrared Spectroscopy, 2009, 17(5): 289.
[17] Kristian Linnet. Clinical Chemistry, 1999, 45(2): 314.
[18] Shen F, Niu X Y, Yang D T, et al. Journal of Agricultural and Food Chemistry, 2010, 58(17): 9809.

Effect of Algorithms for Calibration Set Selection on Quantitatively Determining Asiaticoside Content in Centella Total Glucosides by Near Infrared Spectroscopy

ZHAN Xue-yan, ZHAO Na, LIN Zhao-zhou, WU Zhi-sheng, YUAN Rui-juan, QIAO Yan-jiang*
School of Chinese Pharmacy, Beijing University of Chinese Medicine, Beijing 100102, China

Abstract The appropriate algorithm for calibration set selection was one of the key technologies for a good NIR quantitative model. There are different algorithms for calibration set selection, such as Random Sampling (RS) algorithm, Conventional Selection (CS) algorithm, Kennard-Stone(KS) algorithm and Sample set Portioning based on joint x-y distance (SPXY) algorithm, et al. However, there lack systematic comparisons between two algorithms of the above algorithms. The NIR quantitative models to determine the asiaticoside content in centella total glucosides were established in the present paper, of which 7 indexes were classified and selected, and the effects of CS algorithm, KS algorithm and SPXY algorithm for calibration set selection on the accuracy and robustness of NIR quantitative models were investigated. The accuracy indexes of NIR quantitative models with calibration set selected by SPXY algorithm were significantly different from that with calibration set selected by CS algorithm or KS algorithm, while the robustness indexes, such as RMSECV and |RMSEP－RMSEC|, were not significantly different. Therefore, SPXY algorithm for calibration set selection could improve the predicative accuracy of NIR quantitative models to determine asiaticoside content in centella total glucosides, and have no significant effect on the robustness of the models, which provides a reference to determine the appropriate algorithm for calibration set selection when NIR quantitative models are established for the solid system of traditional Chinese medcine.

Keywords Near infrared spectroscopy; Calibration set selection; Partial least square regression; Diffuse reflectance

(Received Nov. 27, 2013; accepted Mar. 15, 2014)

* Corresponding author