基于偏最小二乘留一交叉验证法的近红外光谱建模样品选择方法的研究

白鹏利*, 王 钧, 尹焕才, 殷 建, 田晶晶, 陈名利, 高 静

(中国科学院苏州生物医学工程技术研究所, 苏州 215163)

摘 要: 目的 提出一种新的挑选定标集的方法-偏最小二乘留一交叉验证法。**方法** 以玉米为例, 通过对玉米中水分含量的实际建模与外部验证, 根据主成分数、相关系数、预测均方根差以及相对分析误差(ratio of performance to standard deviate, RPD)等因素, 综合比较 4 种定标集挑选方法的优缺点。**结果** 偏最小二乘留一交叉验证法结合样品和光谱性质, 在保持原始样品覆盖范围的基础上, 挑选出的定标集所建立的模型具有较低的模型复杂程度、较高的验证相关系数以及较高的 RPD 值。**结论** 该方法既克服了随机挑选法存在的样品代表性不足的风险, 同时也避免了含量梯度法和计算机识别法只考虑样品或者光谱的单一性质的不足, 同时该方法具有操作简单、易于推广等优点, 为食品安全检测提供了一种新的筛选样品的方法。

关键词: 近红外光谱; 偏最小二乘留一交叉验证法; 样品挑选; 定标集

Study on the sample selection methods for building calibration model by near infrared spectroscopy based on partial least squares-leave one out-cross validation

BAI Peng-Li*, WANG Jun, YIN Huan-Cai, YIN Jian, TIAN Jing-Jing, CHEN Ming-Li, GAO Jing

(*Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou* 215163, *China*)

ABSTRACT: Objective To establish a new method of selecting calibration set-partial least squares-leave one out-cross validation (PLS-LOO-CV). **Methods** According to the values of principal component number, correlation coefficient (*R*), root mean square error of prediction (RMSEP), and ratio of performance to standard deviate (RPD), the 4 selection methods of calibration sets were compared by modeling and validation of water content in rice. **Results** By PLS-LOO-CV combined with samples and spectral properties, the selected sets of the model had lower model complexity, higher correlation coefficient and higher RPD value on the basis of maintaining the original sample coverage. **Conclusion** The established method can avoid the risk of lacking of representation caused by random method as well as the efficiency of considering single property of sample or spectrum using content grads method or computer recognition method. At the same time, the PLS-LOO-CV method has the advantages of simple operation and easy popularization, which provides a novel method of screening food samples for food safety inspection.

KEY WORDS: near infrared spectroscopy; partial least squares-leave one out-cross validation; sample selection; calibration set

基金项目: 国家高技术研究发展计划(863 计划)项目(2015AA021106)

Fund: Supported by National High Technology Research and Development Program of China (863 Program) (2015AA021106)

*通讯作者: 白鹏利, 副研究员, 主要研究方向为高分子材料合成生物诊断以及红外拉曼光谱。E-mail: baipl@sibet.ac.cn

***Corresponding author:** BAI Peng-Li, Associate Researcher, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou 215163, China. E-mail: baipl@sibet.ac.cn

1　引　言

近红外光谱分析是一种快速、无损的检测方法，同时也是一种间接的分析方法，建模需要大量样品集才可以保证模型的稳健性。近红外分析方法是食品安全快速检测的重要方法之一，但是对于定标集样品的选择是至关重要的，选择比较好且具有代表性的定标集样品不但在一定程度上可以减少建模的工作量，而且可以提高模型的稳定性以及模型适用性，也进一步提高了检测的准确度。因此挑选样品参与定标是近红外技术分析的核心，同时也是优化近红外模型的关键技术之一[1]。

在近红外分析的模型中，常见的 3 种定标集样品的选择的方法为：(1)常规选择，主要是根据样品性质或组成数据的分布来选择建立定标集的样品，并通过部分样品进行验证。常规选择仅仅考虑了样品的性质，同时在定性模型中无法运用常规选择，如芦永军等[2]提出相似样本剔除算法；(2)计算机识别，则是纯粹根据光谱的性质来分布建立定标集的样品，通常是用计算机来识别所采集样品的马氏距离，通过马氏距离的差异进一步确定适合定标集的样品。吴静珠等[3]提出全局-邻域距离(global H and neighborhood H, GN) 距离法来挑选定标集，相比较于双向算法(Duplex)和 Kennard-Stone 法，模型的稳固性进一步提高，这 2 种方法都属于通过计算机来识别定标集；但这种通过计算机自动识别定标集的方法也存在一定的缺陷，如有些光谱的差异并非完全由所测样品的组成或性质差异引起，可能是由某些随机因素如样品的温度、粒径大小等因素的差异造成的。(3)随机筛选，在定性识别中的应用比较广泛，存在着很强的主观性和随机性，建立的模型的稳固性很差。

本研究基于上述方法的优缺点，将水玉米样本的含水量性质以及近红外光谱结合起来考虑，提出新的定标集样品筛选的方法-偏最小二乘留一交叉验证法。

2　材料与方法

2.1　实验样品及其制备

本研究采用的 80 个玉米数据样品，以玉米的水分含量作为研究对象，水分含量通过卡尔费休法测定。数据由 http://www.eigenvector.com/data/Corn/index.html 提供，近红外光谱的波长范围为 1100~2498 nm，分辨率为 2 cm^{-1}，采样 700 个点。玉米样品的近红外光谱图见图 1。

2.2　定标集样本挑选方法的原理

2.2.1　随机法[4]

由于在定性识别分析中，样品种类无法用一个或多个化学指标量化，通常使用随机法来作为定标集。采用随机分类方法确定建模集样品建模的最大缺点是必须积累大量的样品以供选择，同时也存在人为因素的影响。

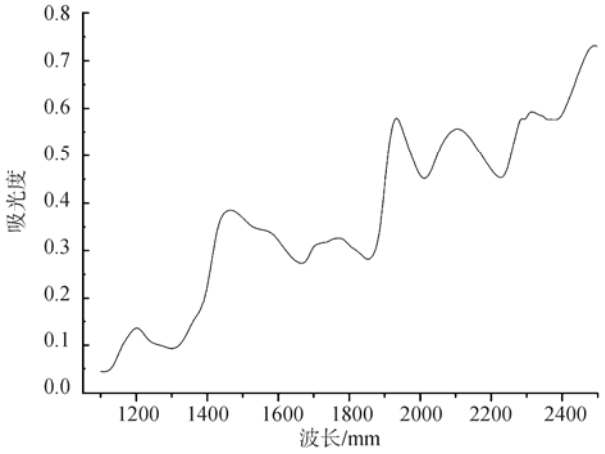

图 1　玉米样品的近红外光谱图

Fig. 1　Near infrared spectrum of corn samples

2.2.2　常规选择-含量梯度法[5]

含量梯度法是一种运用比较多也相对简单的常规选择方法，是将所有的样品集中按所测样品的化学组分的含量值排序(由小到大或反之)，然后从其中按某个序列抽取样品组成定标集或者预测集，使用该方法，需要先指定定标集的样品数，这种方法简单直观，但是定标集样品的代表性相对比较差。

2.2.3　计算机识别法-Kennard-Stone 法[6,7]

Kennard-Stone 法的设计原理是通过对光谱本身进行分析，将光谱的差异较大的样选入定标集，一些较接近的光谱差异的样品进入预测集，这样可以满足代表性的样品全部进入定标集，也解决了定标集样品分布不均匀的问题。使用这种方法，通常事先需要定标集的样品数目。该方法属于比较经典的计算机识别方法，缺陷是没有考虑到光谱的差异不仅仅是由物质的性质导致的，也有可能是仪器以及外界的噪声导致的差异。

2.2.4　偏最小二乘留一交叉验证法

上述3种挑选方法都是在假设所有原始样品集没有异常样品的情况下进行定标集样品挑选的，如果在原始的样品中存在异常样品，将异常样品挑选到定标集或者是预测集中，这样会导致模型的稳定性和准确度大打折扣。同时考虑到含量梯度法以及 Kennard-Stone 法只考虑化学组分或者光谱的单一因素的缺点。针对上述需要解决的问题，本研究结合上述方法提出了一种新的方法-偏最小二乘留一交叉验证法(partial least square-leave one out-cross validation, PLS-LOO-CV)。

PLS-LOO-CV 是一种基于 PLS 的定标集样品的挑选方法。其原理如图 2 所示：将所有的原始样品经过交叉验证得到预测值，一方面可以通过预测值和实际值比较，去除异常点；另一方面将预测值通过排序，从中选择样品作

为定标集，可以扩大定标和预测模型的范围。使用这种方法，同时也需要事先指定定标集的样品数。其优势在于: (1)异常点的存在会严重影响模型的准确度，通过该方法可以有效剔除所有样品的异常点; (2)PLS 不仅仅是对光谱 X 进行分解，也对性质 Y 进行分解，运用 PLS-LOO-CV 同时考虑了光谱矩阵 X 和性质矩阵 Y，这种方法相比较原来的方法，在模型的稳定性上有很大的提高，同时也为筛选定标集提供了一种方法。

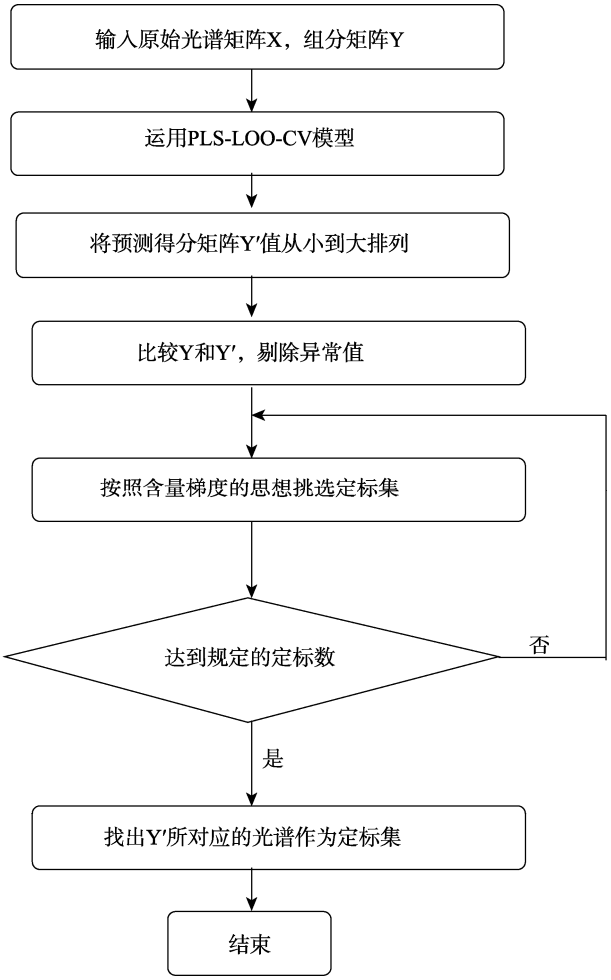

图 2 PLS-LOO-CV 分析流程图

Fig. 2 Analysis procedure of PLS-LOO-CV

2.3 模型的指标

对于建立的校正模型，通常通过相关系数(R)、交叉校验定标标准差(root mean square error of cross, RMSEC)和校验标准差(root mean square error of prediction, RMSEP)来判断模型的优劣，同时结合相对分析误差(ratio of performance to standard deviate, RPD)的大小进一步判断模型的稳定性。

3 结果与讨论

本研究共采用 80 个玉米样本，按照 4:1 的比例，选择 64 个样品作为建模集，16 个样品作为验证集，同时建模集的 64 个样品也作为交叉验证集。Kennard-Stone 法是在 Windows 2000 操作系统下, Matlab 6.5 中编程实现的。PLS 建模是在 Unscrambler X 软件中实现的。采用上述 4 种方法分别挑选定标集样品来建模，预处理采用 mean centering，使用 PLS 进行建模。

首先使用 PLS-LOO-CV 法挑选定标集，先将所有的样品经过留一验证法(leave one out cross validation)建立 PLS 模型，由图 3 可知，样品的预测值与真实值的相关性比较高，达到 0.9 以上，通过观察图 3 散点的分布，将远离线的点作为异常点，初步认定本研究中没有异常点，所有的样品进入筛选定标集和预测集。

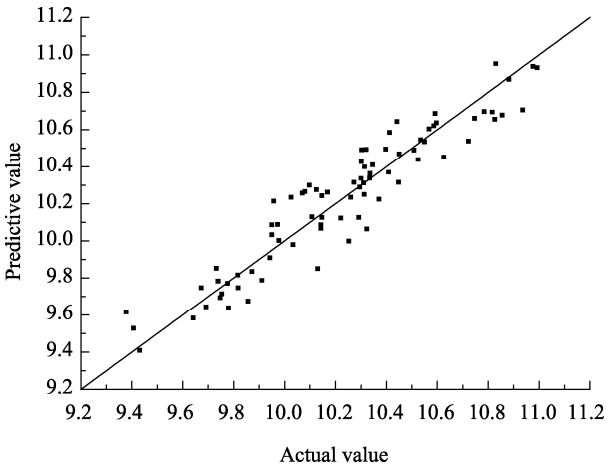

图 3 80 个玉米样品偏最小二乘法的留一验证法结果

Fig. 3 Results of 80 rice samples by left one out cross verification method of partial least squares

将上述样品的预测值从小到大或者反之进行排序，编号。将所有的编号除以 5 余 3 的对应的那一组水分含量以及对应的光谱作为预测集，其余的作为定标集。同时由于在含量梯度中，存在 2 个相同的水分含量的值，故含量梯度法有 2 种不同的选择。

4 种不同的挑选定标集的建模结果如表 1 所示:

主成分因子数: 主成分因子数决定了模型的复杂程度，从表 1 可以明显看出, PLS-LOO-CV 法挑选定标集的主成分的因子数最小，相比较而言，模型的复杂程度越低。

相关性系数[8-12]: 本研究中相关系数 R 有 3 种，一个是建模相关系数，决定所建立模型的相关性; 一个为交互验证的相关系数，决定预测样品的误差大小; 最后一个是预测相关系数，决定模型的稳定性。由表 1 可以得出，随机筛选的建模相关系数和交互验证的相关系数比较高，但是预测的相关系数最低，由于随机本身就存在着一定的不确

表 1　4 种定标集挑选方法的建模结果综合比较

Table 1　Comprehensive comparison of modeling results of 4 methods to select samples for calibration

	建模集		交叉验证集		预测集		主因子数	RPD
	R_c	RMSEC	R_{cv}	RMSECV	R_p	RMSEP		
随机筛选	0.9201849	0.1069997	0.8849688	0.1342314	0.8690789	0.1092382	7	2.854
含量梯度-1	0.9073402	0.1145559	0.8607344	0.141986	0.8700348	0.1386097	6	2.865
含量梯度-2	0.90715	0.1146734	0.8724559	0.1353249	0.8731009	0.1369648	6	2.811
Kennard-Stone	0.9128136	0.1123765	0.8775795	0.1377822	0.9229885	0.101944	7	3.954
PLS-LOO-CV	0.9012151	0.117423	0.8643751	0.1420662	0.9339452	0.1012012	6	4.285

注: *R*c: 定标集的模型的相关系数; *R*cv: 交互验证的相关系数; *R*p: 验证集的相关系数; RMSEC: 建模集的均方根误差; RMSECV: 交互验证的均方根误差; RMSEP: 验证均方根误差; RPD: 相对分析误差

定性，PLS-LOO-CV 法的预测的系数最高，说明运用 PLS-LOO-CV 法挑选定标集所建立的模型的稳定性最高。

均方根误差: 类似于相关系数, 也分为 3 种。由表 1 可以看出, PLS-LOO-CV 的预测均方根误差最低, 其次是 Kennard-Stone 法。

相对分析误差[13]: RPD 用来验证模型的稳定性和预测能力。当 RPD>3 时，模型具有较高的稳定性和良好的预测能力。由表 1 可知, PLS-LOO-CV 法的 RPD 最大(4.285), 其次是 Kennard-Stone 法。

综合以上几点考虑, PLS-LOO-CV 法可以在保持原有的浓度范围前提下, 挑选出定标集所建立的模型复杂度相对比较低, 模型具有较高的 *R*p、较低的 RMSEP 和较高的 RPD。同时相比较其他的方法, 运用 PLS-LOO-CV 法也可以剔除一些异常样品, 进一步提高模型的预测能力。

4　结　论

本研究提出一种新的定标集样品筛选的方法-PLS-LOO-CV。从对玉米样品的分析实例来看, 与其他的定标集筛选的方法相比较, 该方法所建立的模型有较低的复杂度以及较高的 RPD 值, 进一步提高了模型的预测能力, 同时该方法不仅可以运用在近红外定量定性分析模型上, 也可以运用在荧光、拉曼等光谱结合化学计量学分析上的定标集样品的选择, 结果可以使模型的预测范围变大, 同时也使模型的预测能力增强。

参考文献

[1] 陆婉珍, 袁洪福, 徐广通. 现代近红外光谱分析技术[M]. 北京: 中国石化出版社, 2000.
Lu WZ, Yuan HF, Xu GT. Modern near infrared spectroscopy analytical technology [M]. Beijing: Sinopec Press, 2000.

[2] 芦永军, 曲艳玲, 朴仁官, 等. 近红外光谱分析技术定标和预测中的相似样品剔除算法[J]. 光谱学与光谱分析, 2004, 24(2): 158–161.
Lu YJ, Qu YL, Piao RG, *et al*. The algorithm of eliminating the similar sample in the process of calibration and prediction [J]. Spectrosc Spect Anal, 2004, 24(2): 158–161.

[3] 吴静珠, 王一鸣, 张小超, 等. 近红外光谱分析中定标集样品挑选方法研究[J]. 农业机械学报, 2006, 37(4): 80–82.
Wu JZ, Wang YM, Zhang XC, *et al*. Study on algorithms of selection of representative samples for calibration in near infrared spectroscopy analysis [J]. J Agric Mach, 2006, 37(4): 80–82.

[4] 严衍禄, 赵龙莲, 韩东海, 等. 近红外光谱分析基础与应用[M]. 北京: 中国轻工业出版社, 2005.
Yan YL, Zhao LL, Han DH, *et al*. Near infrared spectroscopy analysis and application [M]. Beijing: China Light Industry Press, 2005.

[5] Ferreira DS, Galao OF, Pallone JAL, *et al*. Comparison and application of near-infrared and mid-infrared spectroscopy for determination of quality parameters in soybean samples [J]. Food Control, 2014, 35(1): 227–232.

[6] 侯振雨, 蔡文生, 邵学广. 主成分分析-支持向量回归建模方法及应用研究[J]. 分析化学简报, 2006, 34(5): 617–620.
Hou ZY, Cai WS, Shao XG. Principal component analysis-suppont vector regression and its application in infrared spectral analysis [J]. Chin J Anal Chem, 2006, 34(5): 617–620.

[7] 李彦周, 闵顺耕, 刘霞. 主成分分析在近红外定量分析校正集样品优选中的应用[J]. 分析化学研究简报, 2007, 35(9): 1331–1334.
Li YZ, Min SG, Liu X. Application of principal component analysis in calibration sample selection of near-infrared quantitative model [J]. Chin J Anal Chem, 2007, 35(9): 1331–1334.

[8] Kennard RW, Stone LA. Computer aided design of experiments [J]. Technormetrics, 1969, (11): 137–148.

[9] Zhu XR, Li SF, Shan Y, *et al*. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics [J]. J Food Eng, 2010, 101: 92–97.

[10] Meng XH, Pan QY, Ding Y, *et al*. Rapid determination of phosphplipid content of vegetable oils by FTIR spectroscopy combined with partial least-square regression [J]. Food Chem, 2014, 147(15): 272–278.

[11] 张晓伟, 王加华, 王昌禄, 等. 基于近红外光谱技术检测红曲米中的红曲色素[J]. 现代食品科技, 2014, 30(5): 273–279.
Zhang XW, Wang JH, Wang CL, *et al*. Determination of monascus pigments in red yeast rice using near infrared spectroscopy [J]. Mod Food Sci Technol, 2014, 30(5): 273–279.

[12] Zou XB, Zhao JW, Li YX. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of ‘Fuji’ apple based on BiPLS and FiPLS models [J]. Vibrat Spectrosc, 2007, 44(2): 220–227.

[13] 张小超, 吴静珠, 徐云. 近红外光谱分析技术及其在现代农业中的应用[M]. 北京: 电子工业出版社, 2012.
Zhang XC, Wu JZ, Xu Y. Near infrared spectroscopy and its application in modern agriculture [M]. Beijing: Publishing House of Electronics Industry, 2012.

(责任编辑: 刘 丹)

作者简介

白鹏利, 副研究员, 主要研究方法为高分子材料合成生物诊断以及红外拉曼光谱研究。
E-mail: baipl@sibet.ac.cn

《食品中农兽药残留检测与监控技术专题》征稿函

农药残留、兽药残留是目前食品安全最大的风险, 也一直是食药监管部门监管的重点。近几年快速、高通量、多组分残留同时检测及未知化合物的农/兽药残留筛查技术取得了一定突破。

鉴于此, 本刊特别策划了**“农兽药残留检测与监控技术”**专题, 由**华南农业大学食品学院孙远明教授**担任专题主编。专题将围绕**(1)国内国际农药售药残留管理动态与风险评估新进展; (2)痕量农兽药残留多组分高通量的样品制备新技术, 包括固相萃取、离子交换、凝胶渗透、加速溶剂萃取、衍生化、酶化学等; (3)痕量农兽药残留多组分高通量的检测新技术, 包括液相色谱、气相色谱、色谱质谱联用、免疫亲和色谱、新型快速检测技术等; (4)重要农兽药残留的毒性、控制、分析技术及其各种仪器设备技术应用; (5)农兽药残留的监测抽样和风险管理控制; (6)农兽药残留能力验证的操作技巧等**多方面展开讨论, 计划在 **2017 年 4 月**出版。

鉴于您在该领域的成就, **孙远明教授**和主编**吴永宁研究员**特邀请您为本专题撰写稿件, 综述、研究论文、研究简报均可, 以期进一步提升该专题的学术质量和影响力。请在 **2017 年 3 月 1 日**前通过网站或 Email 投稿。我们将快速处理并经审稿合格后优先发表。

投稿方式:

网站: www.chinafoodj.com

E-mail: jfoodsq@126.com

《食品安全质量检测学报》编辑部