

苹果近红外预测模型的传递研究

吉纳玉, 韩东海*

(中国农业大学食品科学与营养工程学院, 北京 100083)

摘要: **目的** 实现对相同型号和不同型号水果便携近红外仪进行模型传递。**方法** 采用标准变量变换(SNV)预处理, 无信息变量消除法(UVE)及与连续投影算法相结合(UVE-SPA)进行特征波长选择后, 再使用直接校正法(DS)进行模型传递。**结果** 对于相同型号的仪器 K1, K2, 经 SNV-UVE-SPA-DS 算法处理后, 建立在 K1 上的苹果糖度校正模型成功传递到了 K2 上, $R=0.9730$, $RMSEP=0.2375^{\circ}\text{Brix}$, 结果略优于截距法; 而经过 SNV-UVE-DS 算法, 传递到与 K1 不同型号 J 仪器上的模型预测结果为 $R=0.9375$, $RMSEP=0.3458^{\circ}\text{Brix}$ 。**结论** 合适的波长选择结合 DS 算法可以有效地用于相同型号及不同型号水果便携仪上的模型传递。

关键词: 近红外; 模型传递; 直接校正法

Study on near-infrared prediction model transfer for apples

JI Na-Yu, HAN Dong-Hai*

(College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, China)

ABSTRACT: Objective To achieve the model transfer among the same type and different types of fruit portable near infrared instruments. **Methods** Standard normalized variant (SNV) was firstly used as a pre-treatment technique. Uninformative variables elimination (UVE), which also combined with successive projections algorithm (SPA) was used to select effective wavelengths before using direct standardization (DS) for model transfer. **Results** By using SNV-UVE-SPA-DS algorithm, the prediction model was successfully transferred between the same type instruments (K1, K2), $R=0.9730$, $RMSEP=0.2375^{\circ}\text{Brix}$, which was slightly better than bias correction. The best transfer result of different type instruments was obtained by SNV-UVE-DS algorithm, $R=0.9375$, $RMSEP=0.3458^{\circ}\text{Brix}$. **Conclusion** The result highlights that an apposite wavelength selection method combined with the direct standardization (DS) algorithm can be used for model transfer among the same type and different types effectively.

KEY WORDS: near-infrared; model transfer; direct standardization

1 引言

近红外光谱和化学计量学理论相结合形成的近

红外光谱分析技术, 在检测水果完整性及质量方面有很广泛的应用, 如苹果、梨等的可溶性固形物(SSC)含量及硬度等^[1]。

基金项目: 国家自然科学基金项目(31071555)、2013 年度国家国际科技合作专项(2013DFG32760)、公益性行业(农业)科研专项经费项目(201303075)

Fund: Supported by the National Natural Science Foundation of China(31071555), Program of International Science and Technology Cooperation(2013DFG32760) and Special Fund for Agroscientific Research in the Public Interest(201303075).

*通讯作者: 韩东海, 教授, 主要研究方向为食品无损检测。E-mail: handh@cau.edu.cn

*Corresponding author: HAN Dong-Hai, Professor, College of Food Science and Nutritional Engineering, China Agricultural University, No.17, Qinghua East Road, Haidian District, Beijing 100083, China. E-mail: handh@cau.edu.cn

在近红外仪器的实际应用中, 通常会出现这样的情况, 一台仪器(称主机, Master)建立的多元校正模型, 在另一台仪器无法使用, 或产生较大的偏差, 解决这类问题的过程称为模型传递(Model Transfer), 也称仪器的标准化(standardization of spectrometric instruments)^[2]。台间测量误差主要是由于部件、使用时间长短以及周围环境如温度和湿度的差异造成的。

仪器的标准化手段, 主要是从两个方面着手。一方面是完善仪器加工标准化, 减少仪器零部件误差; 另一方面就是从化学计量学角度, 通过信号处理或光谱转换以减少台间差。模型传递是化学计量学的研究内容之一, 具有很大的实际应用价值。因为建立多元校正模型通常需要耗费大量的时间、人力和物力, 如果无法实现仪器间的模型传递, 就会使近红外仪器的推广和发展受到很大限制。

目前针对近红外光谱仪的模型传递的校正方法主要分为两种: 一种是对预测结果的校正, 如斜率/偏差算法^[3](slope/bias, S/B); 另一种是对光谱的校正, 如单变量标准化校正方法中的 Shenk's 算法^[4], 多变量标准化校正方法中的直接校正法^[5](direct standardization, DS), 分段直接校正法^[6](piecewise direct standardization, PDS), 反向标准化算法(reverse standardization, RS), 分段反向标准化算法^[7](piecewise reverse standardization), 小波变换域标准化校正方法^[8](standardization in the wavelet domain, SWD), 以及近几年提出的 Tikhonov 规则法^[9], 虚拟化标准方法^[10]等。针对不同型号仪器之间的传递也有很多研究, 如 Alamar 等^[11]提出的傅里叶变换仪器与阵列检测器之间的转换, Soldado 等^[12]对实验室的色散型近红外便携仪和在线的近红外阵列检测器进行了模型传递。

本文主要利用 DS 算法针对苹果糖度(SSC)的模型传递进行了研究。由于通常标准化样品的数量远少于波长点的数量, 可能会出现过拟合现象, 本文就此提出了一种 UVE-SPA 结合 DS 的算法, 实现了对光谱数据的降维, 使模型更加稳定, 并且解决了两台相同型号的仪器之间以及不同型号仪器之间的模型传递问题。

2 材料与方法

2.1 仪器稳定性测试

K1 与 K2 分别为型号相同生产时间不同的阵列检测器近红外光谱仪, J 为与 K1、K2 生产厂家和型号均不同的阵列检测器近红外光谱仪。K 型仪器是内置光源, 参比为空气, 光谱扫描范围是 500~1010 nm, 分辨率为 2 nm; J 型仪器为外置光源, 参比为聚四氟乙烯, 光谱扫描范围是 600~1000 nm, 分辨率为 1 nm。为确保测量结果的可靠性, 在室温(20 ℃)下对三台仪器进行了准确性和稳定性测试。

三台仪器的积分时间(10 ms)和扫描次数(50 次)相同, 稳定性测试中 K 型仪器的参比为空气, J 型仪器的参比为聚四氟乙烯, 准确性测试参比均为标准锗铟玻璃, 每隔一小时测定一组数, 每组均为三次重复实验。

2.2 样品光谱采集

实验样品为市售山东烟台红富士苹果 60 个, 外部无明显缺陷, 颜色和大小均匀, 实验前将苹果置于室温 12 h, 以使苹果温度与环境温度达到一致, 苹果样品的光谱采集和标准值测定当天完成。设置三台仪器的积分时间为 100 ms, 平均扫描次数为 10 次。对苹果赤道部位进行扫描, 相隔 90°共做采集 4 次光谱, 取平均, 作为原始光谱数据。

2.3 标准值测定

苹果的糖度测定具体参照 GB12295-1990《水果、蔬菜制品可溶性固形物含量的测定——折射仪法》, PAL-1 型手持数字式糖度仪(日本 Atago 公司), 读数结果为 °Brix, 具备自动温度校正功能。

2.4 无信息变量消除法(UVE)原理^[13]

在 PLS 模型中, 光谱矩阵 X 和苹果糖度(SSC)矩阵 Y 存在如下的关系:

$$Y = Xb + e \quad (1)$$

其中, b 是系数向量, e 是误差向量。UVE 就是把一定变量数目的随机变量矩阵(数量级为 10^{-10})加入光谱矩阵中, 然后通过交叉验证建立 PLS 模型, 得到系数矩阵 B , 分析系数向量 b 的平均值和标准偏差的商 C 的稳定性, 即:

$$C_i = \frac{\text{mean}(b_i)}{S(b_i)} \quad (2)$$

其中: $\text{mean}(b)$ 表示回归系数向量 b 的平均值, $S(b)$ 表示回归系数向量 b 的标准偏差, i 表示苹果 SSC 矩阵中第 i 列向量。根据 C_i 的绝对值大小确定是否把第 i 列变量用于最后的 PLS 模型中。

2.5 连续投影算法(SPA)原理^[14]

根据 SPA 算法, 记 $x_{k(0)}$ 为初始迭代向量, N 为需要提取的变量个数, 光谱矩阵为 j 列。

(1) 迭代开始前, 任选光谱矩阵的 1 列 j , 把建模集的第 j 列赋值给 x_j , 记为 $x_{k(0)}$;

(2) 把未选入的列向量位置的集合记为 s , $s = \{j, 1 \leq j \leq J, j \notin \{k(0), \dots, k(n-1)\}\}$;

(3) 分别计算 x_j 对剩余列向量的投影:

$$Px_j = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1}, j \in s; \quad (3)$$

(4) 记 $k(n) = \arg(\max \|Px_j\|), j \in s$; (4)

(5) 令 $x_j = Px_j, j \in s$;

(6) $n=n+1$, 如果 $n < N$, 回到 b 循环计算

对应于每一个 $k(0)$ 和 N , 循环一次后进行多元线性回归分析(MLR), 得到验证集的预测均方根误差(RMSEV), 最小的 RMSEV 值对应的 $k(0)$ 和 N 就是最优值。

2.6 DS 模型传递原理^[5]

直接标准化(DS)的原理是建立主机光谱和从机光谱的函数关系, 由确定的函数关系转换成仪器的光谱信号, 使其尽可能的与主机光谱相接近, 从而实现模型传递。

DS 模型传递算法的流程如下: 设主仪器光谱矩阵为 S_m , 从仪器光谱矩阵为 S_s , 其维数为 $m \times p$, m 为样品个数, p 为波长个数, 为了进行转换, 进行矩阵运算

$$S_m = S_s F + 1b_s^T \quad (5)$$

其中 $F(p \times p)$ 为转换矩阵, $b_s(p \times 1)$ 为背景校正矩阵。

待测样品在从机上测定的光谱 X_s , 经过转化后可用于分析:

$$X_{s, std} = X_s F + 1b_s^T \quad (6)$$

2.7 标准样品选取

DS 算法为有标算法, 其标样选择对于模型的传递效果有非常重要的作用。选择的转换样品集应该具有代表性。Kennard/Stone 算法是一种有效的、应用

广泛的转换集选取方法, 通过计算样本之间的欧式距离来选择。在利用 K/S 方法选择转换集样本前, 首先要对测试样品中存在奇异样品(Outlier)进行检验, 否则会影响转换集样本的选择。

2.8 实验方法和数据处理

根据主仪器光谱, 利用 K/S 算法选择校正集(45 个)和验证集(15 个), 并从校正集中选择一定数量的转换集。首先采用标准变量变换(SNV)方法对主机和从机光谱进行预处理, 然后利用 UVE 以及 UVE-SPA 相结合的算法对主机 K1 光谱选择有效的建模波段。对主机光谱进行 PLS 建模后, 对从机的预测集进行 DS 转换, 并直接进行预测, 以得到的相关系数 R 与均方根误差 RMSEP 作为评价指标, 并与截距法相比较。以上所有的数据和算法均采用 Matlab R2013b 软件处理。

3 结果与讨论

3.1 主仪器上模型的建立及特征波长的优化

经对三台仪器的稳定性和准确性测试, 其结果表明三台仪器具有良好的稳定性和准确性。因主机 K1 与从机 J 的光谱范围及分辨率有差异, 所以只取其共同的近红外波段进行模型传递, 即建模范围为 600~1000 nm, 分辨率为 2 nm。应用 SNV 对主机全波长光谱进行预处理, 然后使用 UVE 以及 UVE-SPA 两种算法对经 SNV 预处理和未经预处理的光谱进行波段选择, 对比结果如表 1 所示。

表 1 经波长选择后的变量个数
Table 1 Variable numbers after wavelength selection

	UVE	UVE-SPA
无预处理	27	4
SNV 预处理	36	4

表 2 DS 传递前模型的预测效果
Table 2 The prediction results before DS

	Slave1	Slave2
R	0.7948	-0.1233
RMSEP (° Brix)	0.7653	3.0627

无预处理时进行 UVE 选择变量后的结果见图 1, 中间竖线的左侧为 201 个变量的稳定性 C 分布曲线,

右侧为 UVE 中产生的 201 个随机变量的稳定性 C 分布曲线。两条水平虚线表示变量选择的阈值上下限, 在虚线外的数值对应的变量被保留, 在虚线内的数值对应的变量不用于建模。

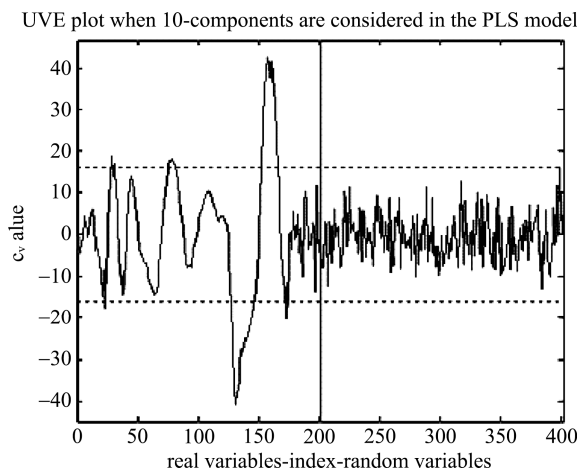


图 1 经 SNV 预处理的 UVE 选择结果
Fig. 1 Result of UVE after SNV pre-treatment

经过无信息变量消除法(UVE)选择过的变量个数仍比较多, 但由于变量间存在较大相关性, 即还有冗余信息, 因此需要利用连续投影算法(SPA)对经过 UVE 选择的变量组进行进一步降维。

无预处理时利用 UVE-SPA 选出的 4 个变量是 862、886、910、930 nm, 经 SNV 预处理利用 UVE-SPA 选出的 4 个变量是 854、884、916、930 nm。从结果可以看出, 是否进行 SNV 预处理对 UVE-SPA 选出的变量都是非常接近。

3.2 转换集样品的选择

选取转换集样品是整个模型传递中非常重要的一步, 若标样数太少, 会导致转换矩阵包含的信息不充分, 若标样数太多, 会增加实际应用的困难。本文首先采用 Kennard/Stone 算法选择样品集, 由于验证集的 RMSEP 随转换集样品个数的增多而逐渐下降, 当增至 12 个时, RMSEP 变化幅度趋于平稳, 所以转换集样品个数确定为 12 个。

3.3 DS 算法模型传递结果

首先观察一下某个样品在主、从仪器上测量得到的吸光度光谱, 从图 2 可以看到 Slave1 与 Master 波形相似但对应的吸光度值有差异, 而 Slave2 与 Master 波形差异非常明显, 图 3 为 Master 与经 DS 算

法转换后的 Slave1, Slave2 的光谱图, 三条光谱基本重合。以下分别对两台型号相同的仪器以及两台型号不同仪器之间的 DS 模型传递情况进行了对比。表 2 为未经 DS 算法转换, 直接采用主机建立的 PLS 模型, 对从机验证集的预测结果, 结果表明未经模型传递对两台光谱仪直接预测的结果均无法使用。

3.3.1 两台型号相同仪器间的模型传递结果

从图 2 中看出, 两台型号相同的仪器间的光谱呈线性关系, 首先按照 K 仪器自带的截距法进行了结果处理, 采用主机模型分别对主机和从机的预测集进行计算, 得到二者预测值平均值之差, 从机的预测值加上差值后与真值之间的相关系数 R 及均方根误差 RMSEP 分别是 0.9406 和 0.2405°Brix。

通过采用不同的预处理方式以及不同波段选择方式选择的变量进行 DS 模型传递, 得到的结果如表 3 所示:

表 3 相同仪器间经 DS 模型传递结果
Table 3 Model transfer results between the same type instruments after DS

		无预处理			SNV 预处理		
		Raw	UVE	UVE-SPA	raw	UVE	UVE-SPA
因子数	f	10	7	4	10	5	4
	R	0.5145	0.9362	0.9335	0.8453	0.9712	0.9730
Slave1	RMSEP (°Brix)	0.9062	0.3126	0.3277	0.9191	0.2437	0.2375

从表 3 中看出, 光谱经标准变量变换(SNV)预处理的结果优于不进行预处理, 校正了苹果的附加散射影响。通过无信息变量消除法(UVE)的变量选择, 从机光谱的预测模型比原始光谱的模型效果得到明显提升, 说明通过 UVE 算法去掉了与苹果糖度(SSC)无关的变量以及噪声, 使模型的传递效果得到提升。经 UVE-SPA 处理后, 去掉了共线性变量, 模型传递效果与 UVE 相当, 但变量数减少为 4 个, 模型更加稳定, 预测结果如图 4 所示。

与截距法相比, 经 SNV 预处理和 UVE-SPA 选择波段后, 得到的结果略优于截距法的结果, 经 UVE-SPA 选择波长后应用 DS 算法转换, 可以成功地用于两台相同型号仪器间的模型传递, 并且优于全波长的 DS 算法。但是对预测的结果直接进行截距法校正, 也能保证结果的准确性。

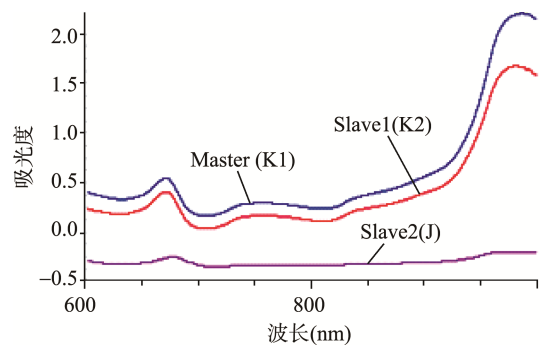


图 2 某样品在模型传递前在三个仪器上测得的光谱
Fig. 2 Spectra of one sample in three instruments before calibration transfer

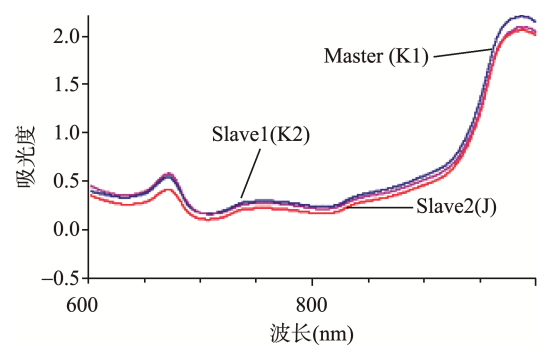


图 3 模型传递后同一样品在三台仪器上的光谱
Fig. 3 Spectra of the same apple in three instruments after calibration transfer

3.3.2 两台型号不同仪器间的模型传递结果

从图 2 中观察出, 两台仪器光谱为非线性关系, 因此不能采用简单的截距法进行结果的转换。经 SNV 预处理后的结果也优于不进行预处理的结果, 但经 UVE-SPA 波长选择后的结果却比只进行 UVE 的结果差。因为 UVE-SPA 的波长选择方式仅针对主机光谱进行处理, 但由于两台仪器光谱差异太大, 当变量个数太少的时候, 从仪器 Slave2 的有效信息没有被完全提取出来, 导致预测的结果相比 UVE 较差。

经过 SNV 预处理, UVE 选择波段后, 对 Slave2 的预测结果最好, $R=0.9375$, $RMSEP=0.3458^{\circ}\text{Brix}$, 结果明显优于不进行波长选择的 DS 算法, 说明经 UVE 进行波长选择进行 DS 转换可以成功用于两台不同型号仪器间的模型传递, 结果如图 5 所示。

3.3.3 讨论

红富士苹果作为天然产物, 化学成分虽然非常复杂, 但在短波近红外波段有比较明显特征的只有水分(85%)、可溶性糖(13.8%)、酸物质(0.3%)等含量

表 4 不同仪器间模型传递结果
Table 4 Model transfer results between the same type instruments after DS

		无预处理			SNV 预处理		
		Raw	UVE	UVE-SPA	raw	UVE	UVE-SPA
因子数	f	10	6	4	10	5	4
	R	0.7888	0.9064	0.8150	0.7946	0.9375	0.8163
Slave2	$RMSEP (^{\circ}\text{Brix})$	0.8857	0.3837	0.5357	1.1670	0.3458	0.5177

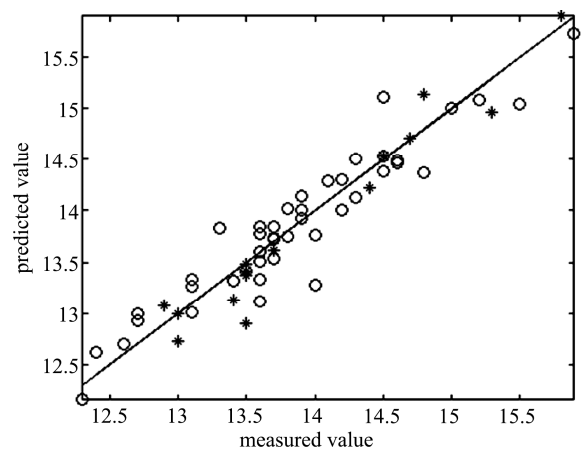


图 4 同型号仪器间 DS 模型传递最优结果
Fig. 4 The best result of the same type after DS calibration transfer

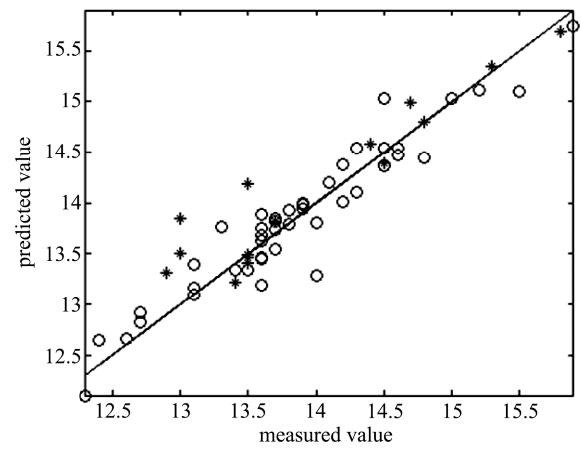


图 5 不同型号仪器间 DS 模型传递最优结果
Fig. 5 The best result of the different type after DS calibration transfer

极高的含 H 基团成分, 因此可以选择少量的特征变量, 建立稳健的模型。对于型号相同, 光谱间仅存在吸光度差异的仪器, 可以采用截距法对预测值进行

转换, 也可以通过少量的特征变量, 对模型进行 DS 转换得到很好的预测效果; 对于型号不同的仪器, 光谱间的差异非常大, 只能用 UVE 结合 DS 算法进行模型传递, 但保留变量不能太少, 否则因光谱峰值不对应而丢失部分特征信息。

4 结 论

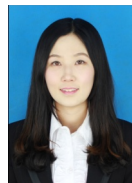
在 K1, K2 与 J 三台水果便携近红外光谱仪上进行了苹果糖度的模型传递实验, 结果表明对于相同型号的仪器, 经过经 SNV 预处理和 UVE-SPA 选择 PLS 的建模波段后, K2 的验证集在 K1 的模型上的预测结果最好, $R=0.9730$, $RMSEP=0.2375^\circ\text{Brix}$, 采用对预测结果直接处理的截距法也可以达到相似的效果; 而对于不同型号的仪器, 不能采用简单的截距法进行预测, 模型传递时保留的变量数不能太少, 经 SNV 预处理结合 UVE 选择 PLS 的建模波段得到的结果较好, $R=0.9375$, $RMSEP=0.3458^\circ\text{Brix}$ 。经波长选择后的 DS 算法不仅仅适用于具有线性关系的两台同型号仪器之间的模型传递, 对于不具有线性关系的仪器, 也能够通过合适的波长选择方式结合 DS 算法实现。

参考文献

- [1] 陈明林. 基于便携式近红外光谱仪的水果内部品质无损检测研究[D]. 杭州: 浙江大学, 2010.
Chen ML. Nondestructive detection of fruit internal quality based on portable near infrared spectrometer[D]. Hangzhou: Zhejiang University, 2010.
- [2] 张学博, 冯艳春, 胡昌勤. 近红外多元校正模型传递的进展[J]. 药物分析杂志, 2009, (8): 1390–1399.
Zhang XB, Feng YC, Hu CQ. Progress in calibration transfer of near-infrared multivariate model[J]. Chin J Pharm Anal, 2009, (8): 1390–1399.
- [3] 赵龙莲. 近红外光谱定量分析数学模型的传递[D]. 北京: 中国农业大学, 1999.
Zhao LL. Calibration transfer of near-infrared spectroscopy quantitative analysis[D]. Beijing: China Agricultural University, 1999.
- [4] 陈斌, 王豪. 专利算法在白酒酒精度近红外光谱分析模型传递中的应用[J]. 红外技术, 2006, 28(4): 245–248.
Calibration transfer between near-infrared spectrometric instrument for the determination of wine alcoholicity using shenk's algorithm[J]. Infrared Technol, 2006, 28(4): 245–248.
- [5] 李庆波, 张广军, 徐可欣, 等. DS 算法在近红外光谱多元校正模型传递中的应用[J]. 光谱学与光谱分析, 2007, 27(5): 873–876.
Li QB, Zhang GJ, Xu KX, et al. Application of DS Algorithm to the Calibration Transfer in Near Infrared Spectroscopy[J]. Spectrosc Spectr Anal, 2007, 27(5): 873–876.
- [6] Wülfert F, Kok WT, Noord OE, et al. Correction of temperature-induced spectral variation by continuous piecewise direct standardization[J]. Anal Chem, 2000, 72(7): 1639–1644.
- [7] Pereira CF, Pimentel MF, Galvão RKH, et al. A comparative study of calibration transfer methods for determination of gaso-line quality parameters in three different near infrared spectrometers [J]. Anal Chim Acta, 2008, 611(1): 41–47.
- [8] Walczak B, Bouveresse E, Massart DL. Standardization of near-infrared spectra in the wavelet domain[J]. Chemometr Intell Lab Systems, 1997, 36(1): 41–51.
- [9] Kalivas JH, Siano GG, Andries E, et al. Calibration maintenance and transfer using Tikhonov regularization approaches[J]. Appl spectrosc, 2009, 63(7): 800–809.
- [10] Cooper JB, Larkin CM, Abdelkader MF. Calibration transfer of near - IR partial least squares property models of fuels using virtual standards[J]. J Chemometr, 2011, 25(9): 496–505.
- [11] Alamar MC, Bobelyn E, Lammertyn J, et al. Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of apple[J]. Postharvest Biol Technol, 2007, 45(1): 38–45.
- [12] Soldado A, Fearn T, Martínez-Fernández A, et al. The transfer of NIR calibrations for undried grass silage from the laboratory to on-site instruments: Comparison of two approaches[J]. Talanta, 2012, 105: 8–14.
- [13] Centner V, Massart DL, de Noord OE, et al. Elimination of un-informative variables for multivariate calibration[J]. Anal Chem, 1996, 68(21): 3851–3858.
- [14] 成忠, 张立庆, 刘赫扬, 等. 连续投影算法及其在小麦近红外光谱波长选择中的应用[J]. 光谱学与光谱分析, 2010, (4): 949–952.
Cheng Z, Zhang LQ, Liu HY, et al. Successive projections algorithm and its application to selecting the wheat near-infrared spectral variables[J]. Spectrosc Spectr Anal, 2010, (4): 949–952.

(责任编辑: 赵静)

作者简介



吉纳玉, 硕士, 主要研究方向为食品无损检测。

E-mail: jinayu.linda@hotmail.com



韩东海, 博士, 教授, 主要研究方向为食品无损检测。

E-mail: handh@cau.edu.cn