国产新型高密度光栅光谱仪数据处理方法研究

张甜甜[1]，李 兵[2*]，蔡贵民[2]，李军会[1*]，马雁军[3]，马 莉[3]，赵龙莲[1]，吴树恩[2]

1. 中国农业大学信息与电气工程学院，北京 100083
2. 上海棱光技术有限公司，上海 200023
3. 上海烟草集团北京卷烟厂，北京 101121

摘 要 由上海棱光技术有限公司与中国农业大学联合研发的S450型近红外高密度光栅光谱仪，使用高速采集技术可得到高密度光谱(波长范围900～2 500 nm，采集间隔0.1 nm，光谱包含16 001个数据点)，本文采用该仪器并以小麦、烟草样品为实验对象，针对高密度光谱的数据特点，采用S. G.(savitzky-golay)平滑、固定窗口组合滑动窗口平滑(FCMWS和一阶导数(FD)等数据处理方法，并应用偏最小二乘法(PLS)对小麦粗蛋白、烟草烟碱及总糖含量进行建模和预测，对仪器整体性能以及数据处理方法的参数优化等，进行了评价和比较研究。结果表明：(1)小麦、烟草样品的原光谱经S. G. 平滑结合一阶导数预处理后，模型性能大幅提高，通过对参数拟合阶次M和平滑点数N进行优化得出，当M一定时，N可选取范围较宽，且当$M=2$和N处于201～801区间时模型效果理想且稳定；(2)FCMWS方法对小麦、烟草样品的原光谱进行两层平均平滑，经调整优化平滑参数K_1和K_2(K_1为第一层平滑的固定窗口大小，K_2为第二层滑动窗口大小)得出：两层平滑参数相乘约为150～310时，模型性能稳定且较优，同时FCMWS方法极大地提高了建模速度；(3)以小麦样品为对象，同时在两台S450型光谱仪上采集样品光谱，对比分析了仪器间的性能差异，结果表明光谱经S. G. 平滑或FCMWS方法处理后，不同仪器模型间相互预测数据的相对偏差小于2.00%，远低于预测值与参考值间的相对偏差，说明上述两种方法均可降低仪器的台间差异，实现台间模型的稳定传递。研究结果表明，国产S450型高密度光栅光谱仪结合数据平滑去噪技术，已满足小麦、烟草等农产品品质检测和模型传递的性能要求，且该光栅型仪器成本相对较低，对农业领域推广近红外快速检测技术的应用具有实际意义。

关键词 近红外；光栅光谱仪；平滑去噪；模型传递

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)08-2651-06

引 言

近红外光谱仪器已经历半个世纪的发展，在这期间，仪器从设计到性能以及测量方法都经历了巨大的变化，随着近红外光谱仪器的数字化程度不断提高，加之功能强大的计算机和化学计量学分析软件的辅助，使其应用领域更为广泛，尤其在欧美等发达国家，近红外光谱仪器已被视为品质管理实验环节中必备的仪器[1]。我国在20世纪80年代初就进行了近红外光谱技术的应用研究，大约在20世纪90年代中后期，经过一些厂家和科研单位的积极合作与努力，在近红外光谱仪器的研制、软件开发方面取得了可观的成绩。如瑞利分析仪器公司研制了傅里叶变换近红外分析仪，石油化工科学研究院研制了采用电荷耦合检测器(CCD)的多通道近红外光谱仪器，中国农业大学研制了滤光片型漫透射近红外谷物品质分析仪等。但由于对近红外光谱仪器的研制起步较晚，在近红外仪器制造的核心技术及仪器普及方面，我国仍滞后于一些发达国家，目前国内近红外仪器市场依旧大量依靠进口，只有小部分企业单位购买了近红外光谱仪器，未来的市场增长空间还非常大[1-2]。在国产化和低成本化的光栅型近红外光谱仪上研究光栅型光谱数据的处理方法以提高国产仪器性能，对国产近红外仪器的普及、提高国内食品和生产酿造类中小型企业的效益以及推动国内近红外技术产业的发展具有重要的现实意义。

收稿日期：2018-06-30，**修订日期**：2018-10-25

基金项目：国家重点研发计划课题(2016YFD0700304，2004BA210A03)资助

作者简介：张甜甜，女，1994年生，中国农业大学信息与电气工程学院硕士研究生 e-mail：1521958103@qq.com

*通讯联系人 e-mail：caunir@cau.edu.cn；libing@lengguang.com

本文基于国产 S450 型近红外高密度光栅光谱仪，以小麦和烟草为实验样品，通过建立小麦粗蛋白、烟草烟碱及总糖的定量分析模型，研究适用的数学方法对高密度光谱进行预处理，以最大限度的滤除噪声提高光谱质量，使其模型的性能满足实际应用需求。

1 实验部分

1.1 材料

采用的 72 份小麦粉末样品由中国农业科学院提供，并已使用国标凯氏定氮法测定其粗蛋白含量，随机从样本中选取 52 个小麦样品本用于模型的建立，剩余 20 个样品用于评估模型的预测能力；41 份烤烟粉末样品由上海烟草集团北京卷烟厂提供，并按照烟草行业标准 YC/T 468-2013 和 YC/T 159-2002 测定其总糖、烟碱含量，随机选取 30 个烟草样品用于模型的建立，其余 11 个样本用于评估模型的预测能力[5, 9, 11]。

1.2 仪器与光谱采集

实验所用 2 台国产仪器均为 S450 型光栅积分球漫反射近红外光谱仪，由上海棱光公司与中国农业大学合作开发研制，仪器上配备中国农业大学近红外建模软件(CAUNIR6.0)。

S450 扫描条件：波长扫描，波长范围为 900～2 500 nm，分辨率 12 nm，扫描次数为 4 次，每隔 30 min 扫描一次背景，同时在 S450 的两台机器上扫描所有样品[10, 12-14]。

1.3 方法

1.3.1 数据预处理方法

(1)S.G. 平滑：S.G. 平滑是一种在时域内基于局域多项式的最小二乘法(partial least squares，PLS)拟合算法，基本原理是利用多项式来对移动窗口内的原始光谱数据进行多项式分解并运用最小二乘法进行数据拟合，用拟合值代替原始数值，以达到去除高频噪声点平滑原数据序列的作用[4]。

(2)FCMWS：固定窗口组合滑动窗口平滑(fixed window combine moved window smoothing，FCMWS)是本文针对高密度光谱数据量大的特点，提出的一种新型平滑方法。其核心思想是：第一层使用固定窗口平滑可极大地减少数据点数提高建模速度，第二层使用滑动窗口平滑可进一步降低噪声分量。图 1 为 FCMWS 算法流程图，其涉及的参数 K_1 为第一层固定窗口的大小，K_2 为第二层滑动窗口的大小。

(3)一阶导数(first derivative，FD)：导数光谱既可以消除基线漂移或平缓背景干扰的影响，也可以提供比原光谱更高的分辨率和更清晰的光谱轮廓变化[2,4]。采用一阶导数配合上述方法对原光谱进行预处理。

1.3.2 仪器台间差异的评价方法

为获取仪器台间差异进行仪器稳定性研究，将所有样品分为建模样品与模型外部检验样品，分别在 1 号与 2 号两台仪器上进行扫描，对采集得光谱数据进行预处理后再进行如图 2 所示的操作：1 号仪器测量得到数据建立模型 M1 对 2 号仪器测量的外部检验样品进行预测，得到预测集 P1-2，2 号仪器测量的数据建立模型 M2 对 1 号仪器测量的外部检验样品进行预测，得到预测集 P2-1；对比仪器间模型的预测数据，获取差异后，参考外部检验样品化学含量的真值(参考值 T)对两台仪器性能差异及模型传递进行研究[2,6-7]。

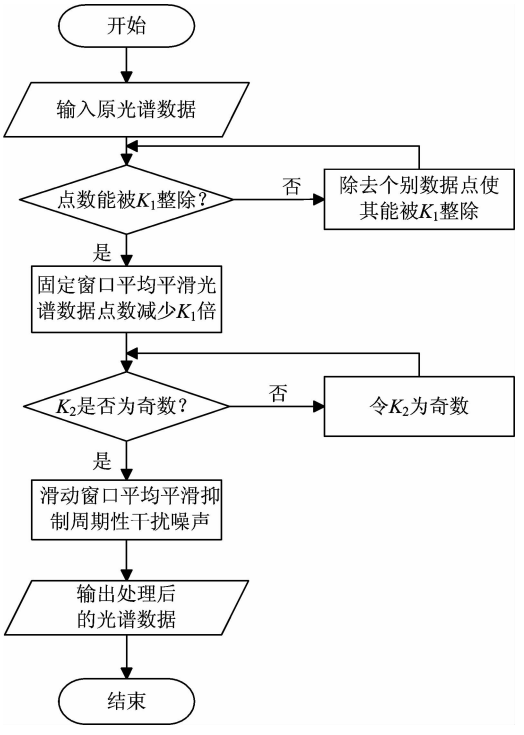


图 1 FCMWS 算法流程图

Fig. 1 Flow chart of FCMWS algorithm

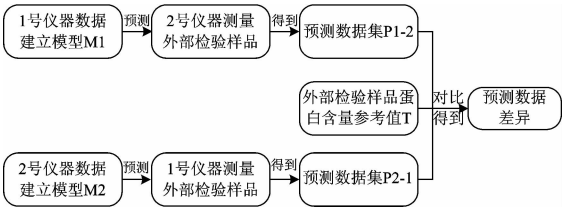


图 2 获取台间数据差异的流程评价方法

Fig. 2 Flow chart for getting data of instrument differences

1.3.3 模型评价方法

评价模型的指标性参数为交叉验证集与预测集的决定系数(R^2)、标准偏差(standard error of cross-validation/prediction，SECV/SEP)和相对标准差(relative standard deviation，RSD/%)[2-3,8]。

1.3.4 样品吸光度噪声计算方法

样品吸光度噪声(absorbance noise of sample，Ans)通过差谱法获得，计算公式为

$$\mathrm{Ans}=\frac{\sum_{1}^{n}|A_2-A_1|}{n} \tag{1}$$

式(1)中 n 为采集的光谱点数，A_1 和 A_2 分别为一个样品在同一波长点前后两次测量的吸光度值。

2　结果与讨论

2.1　S. G. 平滑与 FCMWS 方法的参数优化

(1)S. G. 平滑方法的参数优化

S. G. 平滑效果受平滑点数 N 的影响较大，点数设置过少容易引进新误差，点数过多则容易磨光丢失包含样品信息的光谱数据，都可能使光谱的质量下降影响模型精度[6]，考虑到低点数平滑已不适用于数据点量大且密集的高密度光谱，本工作选择拟合阶次 M 为 2，重点对平滑点数 N 进行了优化研究。

小麦粗蛋白、烟草烟碱及总糖模型的交叉验证集标准偏差(SECV)与预测集标准偏差(SEP)随参数 N 的变化趋势如图 3 所示：随着平滑点数的增加，小麦粗蛋白、烟草烟碱及总糖的模型效果均先呈上升趋势，后在 201～801 点之间呈平稳趋势，当平滑点数超过 801 点后模型效果愈来愈不理想，数据结果虽存在一定的统计波动但不影响整体的变化趋势。

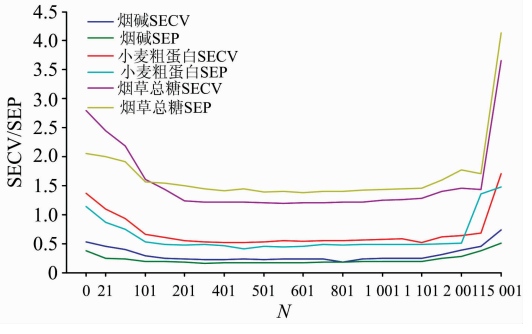


图 3　小麦、烟草模型的 SECV 与 SEP 随参数 N 的变化趋势

Fig. 3　Trends of SECV and SEP in wheat and tobacco models with N

(2)FCMWS 方法的参数优化

对于 FCMWS 方法，优化了其包含的参数 K_1 和 K_2。小麦粗蛋白、烟草烟碱及总糖模型的 SECV 与 SEP 随参数 K_1 和 K_2 的变化趋势如图 4 所示：当第一层平滑窗口 K_1 的大小一定时，改变第二层滑动窗口 K_2 的大小，模型效果先呈变优趋势，在两层参数之积处于 150～310 区间时模型效果较优且呈稳定状态，当两层参数之积大于 310 时模型效果变差。

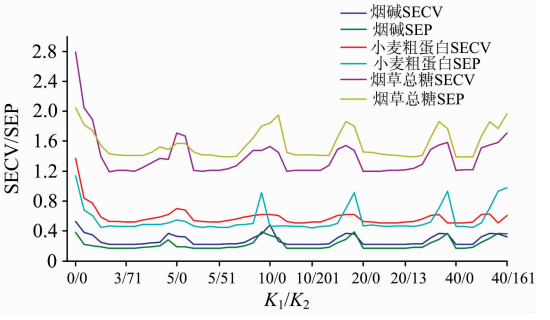
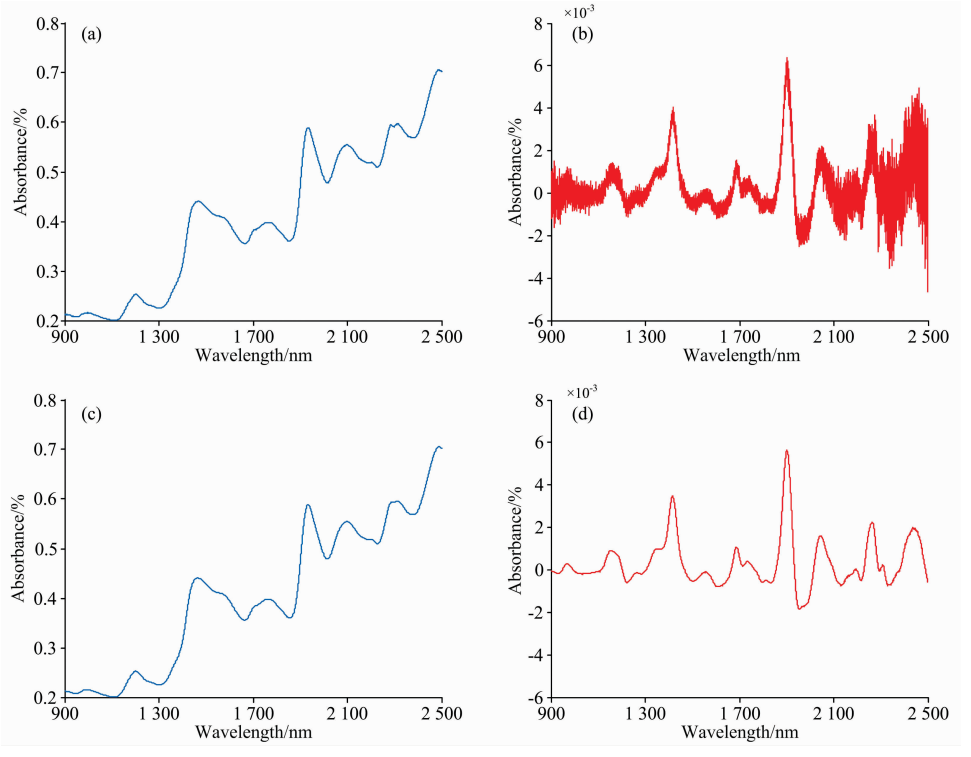


图 4　小麦、烟草模型的 SECV 与 SEP 随参数 K_1 和 K_2 的变化趋势

Fig. 4　Trends of SECV and SEP in wheat and tobacco models with K_1, K_2

2.2　光谱数据处理前后的信号分析

以一条小麦原光谱为例进行处理前后的信号分析，对原光谱分别进行 S. G. 平滑和 FCMWS 方法处理，处理前后的对比光谱图像如图 5(a—f)所示。



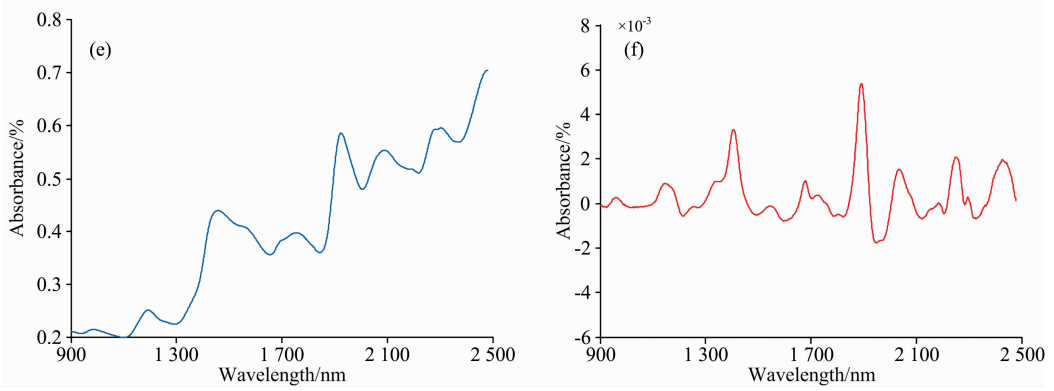


图 5 处理前后的小麦光谱对比图

(a)：原光谱；(b)：原光谱的一阶导数光谱；(c)：S. G. 平滑后的光谱；(d)：S. G. 平滑后的一阶导数光谱；(e)：FCMWS 处理后的光谱；(f)：FCMWS 处理后的一阶导数光谱

Fig. 5 Comparison of wheat Spectra before and after processing

(a): Raw spectra; (b): Raw spectra processed by the first derivatives; (c): Spectra processed by SG smoothing; (d): Spectra processed by SG smoothing and the first derivatives; (e): Spectra processed by FCMWS; (f): Spectra processed by FCMWS and the first derivatives

在小麦和烟草样品中各挑选 5 个样品，每个样品分别扫描两次，通过 1.3.4 中式(1)计算得到光谱处理前后的样品吸光度噪声汇总见表 1，其中 S. G. 平滑后波长点数与原光谱一致，波长点不变，而 FCMWS 方法由于第一层为固定窗口的平滑处理，当窗口大小取 10 时，处理后波长点数减少至 1/10，此时将各个窗口平滑后的值赋给其窗口内的第一个波长点。

表 1 小麦、烟草样品经不同方法处理前后的样品吸光度噪声对比

Table 1 Comparison of absorbance noise of wheat and tobacco samples before and after processing by different methods

样品对象	处理光谱	样品 1	样品 2	样品 3	样品 4	样品 5
小麦	原光谱	0.000 439	0.000 243	0.000 256	0.000 215	0.000 255
	S. G. 平滑	0.000 374	0.000 179	0.000 159	0.000 150	0.000 196
	FCMWS	0.000 365	0.000 174	0.000 155	0.000 146	0.000 193
烟草	原光谱	0.000 262	0.000 281	0.000 274	0.000 417	0.000 245
	S. G. 平滑	0.000 192	0.000 234	0.000 231	0.000 369	0.000 166
	FCMWS	0.000 197	0.000 226	0.000 231	0.000 360	0.000 166

从图 5(a,b)中可以看出，处理前后光谱图像的波形基本保持一致，且 S. G. 平滑和 FCMWS 方法处理后的一阶导数光谱图像均变得尤为平滑清晰。表 1 的结果表明，相比原光谱 S. G. 平滑后小麦样品吸光度噪声平均降低 26.51%、烟草样品吸光度噪声平均降低 20.59%，FCMWS 方法处理后相比原光谱，小麦样品吸光度噪声平均降低 28.27%、烟草样品吸光度噪声平均降低 21.21%，说明无论 S. G. 平滑还是 FCMWS 方法均可大幅滤除噪声，提高原光谱的信噪比。

2.3 光谱数据处理前后的建模结果分析

表 2 列出了 S450 型仪器测得的高密度近红外光谱数据在经 SG 平滑、FCMWS 方法处理后的建模结果，结果表明对小麦样品和烟草样品，S. G. 平滑与 FCMWS 方法对原高密度光谱的去噪效果均很显著，使模型的性能大幅提高，可满足实际应用中一般质量检测的精度要求。

表 2 处理前后不同对象的模型对比

Table 2 Comparison between models of different objects before and after processing

建模对象	预处理方法	Cross-validation		Prediction	
		R_c^2	SECV	R_p^2	SEP
小麦蛋白	None	0.50	1.37	0.73	1.14
	SG (M=2, N=451)	0.91	0.52	0.97	0.41
	FCMWS (K_1=10, K_2=21)	0.91	0.52	0.97	0.44
烟草总糖	None	0.74	2.79	0.83	2.05
	SG (M=2, N=601)	0.95	1.20	0.92	1.38
	FCMWS (K_1=5, K_2=61)	0.95	1.23	0.92	1.39
烟草烟碱	None	0.46	0.53	0.70	0.38
	SG (M=2, N=451)	0.92	0.23	0.90	0.17
	FCMWS (K_1=10, K_2=17)	0.92	0.22	0.89	0.17

2.4 不同仪器间的模型传递

在 72 份小麦样品中选取 52 个作为建模样品，剩余 20 个作为外部检验样品，分别在 S. G. 平滑(N=451)与 FCMWS(K_1=10，K_2=21)方法下，预处理 S450 型 1 号、2 号仪器测得的原光谱，按图 2 操作流程进行建模预测。表 3 为预测数据集与参考值以及预测数据集之间的数据关系结果。

表 3　S. G. 平滑与 FCMWS 方法下台间模型预测差异

Table 3　Prediction deviation of models based on different instruments by SG and FCMWS

数据集	SG			FCMWS		
	R^2	STD	RSD/%	R^2	STD	RSD/%
P1-2 与 T	0.97	0.42	2.86	0.97	0.44	2.95
P2-1 与 T	0.98	0.39	2.65	0.97	0.41	2.77
P1-2 与 P2-1	0.99	0.21	1.43	0.99	0.27	1.74

表 3 数据表明两台仪器模型的相互预测数据集间，相对偏差小于 2.00%，远小于其各自与参考值间的相对偏差，原高密度光谱数据经 S. G. 平滑或 FCMWS 方法预处理后所建模型的预测性能无明显差异，国产新型高密度光栅近红外仪 S450 具有较高稳定性，整体性能良好。

3　结　论

国产新型高密度光栅光谱仪采用高速采集技术，可采集到间隔为 0.1 nm 的高密度近红外光谱，从提高光谱信噪比的角度出发，结合不同滤波算法各自的优势，兼顾模型稳定性、预测性，提出了适用于高密度光谱预处理的两种方法，即 S. G. 平滑和 FCMWS 方法，并对涉及参数进行了优化。实验结果表明所提出的这两种方法对高密度光谱平滑效果显著，且 FCMWS 方法可极大地提高建模速度，通过比较处理前后的样品吸光度噪声，表明提出的两种方法均提高了原光谱的信噪比，仪器及模型性能良好，该工作对国产近红外仪器的推广具有积极意义。

References

[1] DING Ying(丁　莹). Infrared(红外), 2012, 33(7): 1.

[2] YAN Yan-lu(严衍禄). Principle, Technology and Application of NIR Spectra Analysis(近红外光谱分析的原理、技术与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社), 2013.

[3] CHU Xiao-li, et al(褚小立，等). Practical Handbook for Near Infrared Spectroscopy(近红外光谱分析技术实用手册). Beijing: China Machine Press(北京：机械工业出版社), 2016. 3.

[4] Brad Swarbrick. NIR News, 2016, 27(1): 39.

[5] Lin C, Chen X, Jian L, et al. Food Chemistry, 2014, 162: 10.

[6] Liudmil Antonov. Journal of Near Infrared Spectroscopy, 2017, 25(2): 145.

[7] CHEN Hua-zhou, PAN Tao, CHEN Jie-mei(陈华舟，潘　涛，陈洁梅). Computer and Applied Chemistry(计算机与应用化学), 2011, 28(5): 518.

[8] FU Yi, ZHANG Yong-jun, CHEN Hua-cai, et al(傅　谊，张拥军，陈华才，等). Food Science and Technology(食品科技), 2012, 37(5): 42.

[9] CAI Jian-hua, XIAO Yong-liang, LI Xiao-qin(蔡剑华，肖永良，黎小琴). China Tobacco(中国烟草学报), 2017, 23(4): 9.

[10] Jerome J Workman. Applied Spectroscopy, 2018, 72(3): 340.

[11] ElMasry G, Sun D W, Allen P, et al. Journal of Food Engineering, 2012, 100(1): 127.

[12] Ana Garrido-Varo. NIR News, 2017, 28(5): 2.

[13] Ba 瘟堵 lar Mehmet, Ertugay Mustafa Fatih. Turkish Journal of Agriculture and Forestry, 2011, 35(2): 139.

[14] SUN Jia-yin, LI Chun, LIU Ying, et al(孙佳音，李　淳，刘　英，等). Infrared and Laser Engineering(红外与激光工程), 2016, 45(7): 148.

Study on Spectral Data Processing Methods of New Type High-Density Grating Spectrometer Made in China

ZHANG Tian-tian[1], LI Bing[2*], CAI Gui-min[2], LI Jun-hui[1*], MA Yan-jun[3], MA Li[3], ZHAO Long-lian[1], WU Shu-en[2]

1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
2. Shanghai Lengguang Technology Co., Ltd., Shanghai 200023, China
3. Beijing Cigarette Factory of Shanghai Tobacco Group, Beijing 101121, China

Abstract In this paper, we used the S450 near-infrared high-density grating spectrometer with technology of high-speed acquisition developed by Shanghai Lengguang Technology Co., Ltd. and China Agricultural University, took wheat and tobacco as the experimental object, and aimed at the high-density spectra (wavelength range is 900～2500nm, interval of wavelength is 0.1 nm, contains 16 001 data points). By adapting processing methods such as S. G. (Savitzky-Golay) smooth, FCMWS (Fixed window combine moved window smoothing) and the First Derivative, Partial Least Squares (PLS) was also used to model and predict the content of crude protein in wheat, nicotine and total sugar in tobacco, evaluate performance of the spectrometer, and optimize the parameters of processing methods. The results show that: (1) The performance of the models was greatly improved after the high density spectrum was processed by S. G. and the first derivative. Optimizing the parameter M (fitting order) and N(number of smoothing point), if M is a fixed number, N can be selected from a wider range, and when $M=2$, N is in the interval of 201～801, the performance of models is ideal and stable; (2) The FCMWS was designed for smoothing layers of two, fixed window size of the first layer K_1 and second layer K_2, and it was concluded that the performance of models is better and superior when the multiplication of K_1 and K_2 is about 150～310, moreover the FCMWS algorithm is speedy in modeling. (3) In order to analyze instrument differences, only took wheat as the object, which was measured by two S450 spectrometers, experimentally, whether the spectrum is processed by S. G. or FCMWS, the relative deviation of the predicted data from different models between instruments is less than 2.00%, which is far lower than the relative deviation between the predicted and reference values. It indicates that the above two methods can reduce the instrument differences and models can transfer stably among instruments. For wheat, tobacco and other agricultural products, the results of this study reflect that the domestic high-density grating spectrometer S450 combined with de-noising methods, can meet the actual requirements of quality detection and model transfer, and the grating instrument is relatively low-cost, which is significant for popularizing application of the rapid detection technology of near infrared in the agricultural field.

Keywords Near infrared; Grating spectrometer; Smoothing de-noising; Model transfer

(Received Jun. 30, 2018; accepted Oct. 25, 2018)

* Corresponding authors