DOI: 10.3724/SP. J. 1096. 2010. 00342

血糖近红外光谱分析的 Savitzky-Golay 平滑模式与偏最小二乘法因子数的联合优选

谢 军[1] 潘 涛*[1] 陈洁梅[2] 陈华舟[1,3] 任小焕[1]

[1](广东省高等学校光电信息与传感技术重点实验室(暨南大学) , 广州 510632)

[2](暨南大学生物工程学系, 广州 510632) [3](上海大学数学系, 上海 200444)

摘 要 利用偏最小二乘法(PLS) 和光谱 Savitzky-Golay(SG) 平滑方法,建立血清葡萄糖近红外光谱分析的优化模型。基于最优单波数模型的预测效果,提出划分校正集和验证集的一种新方法。采用 10000 ~ 5300 cm^{-1} 和 4920 ~ 4160 cm^{-1} 的组合波段,光谱经过 SG 平滑处理,利用 PLS 方法建立定标预测模型。将平滑点数扩充为 5,7,…,87(奇数) ,多项式次数扩充为 n = 2,3,4,5,6,得到包含 582 个平滑模式的 14 个平滑系数表。对所有平滑模式和 PLS 因子数(1 ~ 40) 分别建立 PLS 模型。按照预测效果进行优选,得到最优 SG 平滑模式为 1 阶导数平滑,3、4 次多项式类型,SG 平滑点数为 53,最优 PLS 因子数为 7,最优 RMSEP 达到 0. 376 mmol/L。所采用的划分校正集和验证集的方法、SG 平滑模式的扩充、SG 平滑模式和 PLS 因子数的联合大范围筛选能够有效地应用于近红外光谱分析的模型优化。

关键词 血糖; 近红外光谱; 偏最小二乘法; Savitzky-Golay 平滑; 校正集验证集划分

1 引 言

随着光谱技术和化学计量学的快速发展,近红外光谱以其分析效率高、速度快、成本低、非破坏性和易于在线分析等特点已广泛应用于农业、食品、烟草、医药等领域[1,2]。模型优化对于提高近红外光谱预测能力具有重要意义。偏最小二乘法(PLS) 是融合主成分分析和多元线性回归的一种有效的化学计量学方法[1~7],其中合理选用 PLS 因子数,对于充分利用光谱信息和消除噪声非常重要。在光谱预处理中,平滑可以保留光谱轮廓而消除噪声,求导则可以有效消除基线漂移、倾斜等噪声。Savitzky-Golay (SG) 方法是应用十分广泛而有效的平滑和求导预处理方法[8~11]。按照导数阶数(平滑看成 0 阶求导) 、多项式次数和平滑点数的不同,SG 平滑模式有很多种,计算公式也各不相同。其中平滑点数的设置非常重要,点数过少容易产生新误差,点数过多则容易使包含信息的光谱数据磨光丢失,都会造成模型精度下降。根据预测效果对 SG 平滑模式与 PLS 因子数联合筛选是很有必要的,但由于工作量庞大,既往的研究很少做到这一步。另一方面,考虑到有些实际测量体系可能需要更多的平滑点数,比如测量数据波长间隔小的情形,相邻波长点的数据过于相似,点数少的平滑效果往往不够好。为了拓宽适用范围,有必要按照原始论文的方法[8]扩充平滑系数表。

血糖近红外光谱分析及其模型优化是很重要的研究方向[3,4]。本实验以血糖近红外光谱分析为例,研究 SG 平滑模式与 PLS 因子数的联合优化设计在近红外光谱分析模型优化中的作用。为了改善模型预测能力,基于最优单波数模型提出了划分校正集和验证集的新方法。

2 实验部分

2.1 实验材料、仪器和测量方法

191 份血清样品由广州市某医院提供,样品葡萄糖的含量由全自动生化分析仪测定作为光谱分析的参考化学值。全体化学值范围 3. 53 ~ 6. 15 mmol/L,均值、标准偏差分别为 4. 90 和 0. 59 mmol/L。实验仪器为 5700 傅里叶变换型近红外光谱仪(美国 Necolet 公司) ,探测器为铟镓砷(InGaAs) 。用光程

2009-06-19 收稿; 2009-09-12 接受

本文系国家自然科学基金(No. 10771087) 、广东省自然科学基金(No. 7005948) 、广东省科技计划项目(Nos. 2007A020905001, 2007B030501008, 2007B020714001) 、广州市科技攻关项目(No. 2007Z3-E0281) 资助

* E-mail: tpan@ jnu. edu. cn

2 mm 的石英比色皿测量光谱,扫描谱区 10000 ~ 4000 cm^{-1}, 分辨率 4 cm^{-1}, 扫描次数 64。

2.2　校正集和验证集的划分方法

基于全体样品最优单波数模型的预测效果给出划分校正集验证集的一种新方法。根据比尔定律,考虑血清样品吸光度与葡萄糖化学值的单波数线性模型

$$A(v) = k(v)C + \varepsilon \tag{1}$$

其中,$A(v)$ 为样品在波数 v 的吸光度,$k(v)$ 为在波数 v 的葡萄糖单位浓度吸光系数,C 为样品的葡萄糖浓度化学值,ε 为其它未知干扰。在每个波数 v,利用全体样品的吸光度和化学值回归计算 $k(v)$,再利用 $k(v)$ 和样品吸光度计算样品 i 的预测值 $C''_i(v)$ ($i = 1,2,\cdots,N$) ,N 是全体样品个数。进一步计算预测值与化学值的均方根偏差(RMSE) ,设 C_i 为样品 i 的化学值,则

$$\mathrm{RMSE}(v) = \sqrt{\frac{\sum_{i=1}^{N}(C'_i(v) - (C_i)^2}{N-1}} \tag{2}$$

按 RMSE 值最小选出最优单波数模型和相应波数 v_{Optimal},根据最优单波数模型计算每个样品的浓度预测值与化学值的偏差,称为单波数预测偏差(Single wavenumber prediction bias, SWPB) 。

$$\mathrm{SWPB}_i = |C'_i(v_{\mathrm{Optimal}}) - C_i|,\ i = 1,2,\cdots,N \tag{3}$$

SWPB 是吸光度和化学值的一种关联指标,根据 SWPB 划分校正集检验集,利用计算程序筛选使两个集合的 SWPB 分布一致(均值和标准偏差相近,相对误差小于 1%) 。将化学值和光谱数据结合起来使校正集验证集具有相似性,从而具有建模代表性。为了使得校正集浓度范围能够涵盖验证集浓度范围,将化学值最大和最小的样品放在校正集,化学值次大次小的样品放在验证集。

2.3　SG 平滑方法

SG 平滑的参数包括导数阶数 s、多项式次数 n 和平滑点数 $2m+1$。SG 平滑把光谱区间的若干个连续点作为一个窗口,窗口内每点用多项式(以点的编号 0, ±1, ±2,…为变量) 来做实测数据的最小二乘拟合。拟合后,多项式在编号为 0(中心点) 的值就是 SG 平滑值,多项式对编号求导后在编号为 0(中心点) 的值就是 SG 导数值。按上述程序,窗口中心点的平滑值和各阶导数值都可以表示为窗口内各点实测数据的线性组合。线性组合的系数(即平滑系数) 由平滑点数(即窗口内的点数) 、多项式次数和导数阶数唯一确定。通过窗口移动,得到每个窗口中心点的平滑值和各阶导数值,从而得到原谱的 SG 平滑谱和 SG 导数谱。为了拓宽应用范围,本研究将平滑点数从原有的 5 ~ 25 之间奇数[8]扩充为 5 ~ 87 之间的奇数,多项式次数扩充为 $n = 2,3,4,5,6$(原为 $n = 2,3,4,5$) ,按照原方法[8]编写程序计算,得到 14 个涵盖原有平滑系数的平滑系数表,共有 582 个平滑模式(原有 117 个) ,是适用范围更宽的 SG 平滑预处理群。

2.4　模型的评价指标

模型评价指标主要包括预测均方根偏差(RMSEP) 和预测相关系数(R_{P})

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{M}(C'_{ip} - C_{ip})^2}{M-1}},\qquad R_{\mathrm{P}} = \frac{\sum_{i=1}^{M}(C_{ip} - C_{mp})(C'_{ip} - C'_{mp})}{\sqrt{\sum_{i=1}^{M}(C_{ip} - C_{mp})^2 \sum_{i=1}^{M}(C'_{ip} - C'_{mp})^2}} \tag{4}$$

其中,C'_{ip} 和 C_{ip} 分别为验证集中第 i 个样品的预测值和化学值,C'_{mp} 和 C_{mp} 分别为验证集样品的预测值均值和化学值均值,M 为验证集的样品个数。R_{P} 与 RMSEP 是有一定关联的,RMSEP 值低,R_{P} 一般也较高。本研究以 RMSEP 为优化目标来进行参数设计和模型优选。

3　结果与讨论

3.1　样品光谱、校正集和和验证集的划分

191 个血清样品的近红外光谱如图 1 所示。光谱在 6900 和 5200 cm^{-1} 附近有水分子的强烈吸收,除了水的吸收峰外没有其它显著的吸收峰,光谱重叠严重,吸收较弱。考虑到在 5200 和 4000 cm^{-1} 附近

吸收强烈，光谱能量低，信息含量差，噪音大，故把这两段（吸光度高于2的波段）光谱数据扣除后用于建模。用于建模的光谱波段是10000～5300 cm^{-1}和4920～4160 cm^{-1}两段的组合。

按照2.2节的方法，建立每个波数点的吸光度和化学值的单波数模型，按照RMSE最小找到最优波数$\nu_{Optimal}$为7232 cm^{-1}。根据7232 cm^{-1}对应的最优单波数模型计算每个样品的SWPB，全体样品的SWPB和化学值分布如图2所示。由图2可见，全体样品的化学值和SWPB分布均匀，无显著的异常样品。因此，全体样品都用于建模。按照大约2∶1的比例，校正集131个样品，验证集60个样品，按照2.2节方法划分校正集验证集，得到的校正集验证集的SWPB、化学值的均值和标准偏差如表1所示。表1和图2都表明，校正集验证集的化学值和SWPB分布都非常一致。

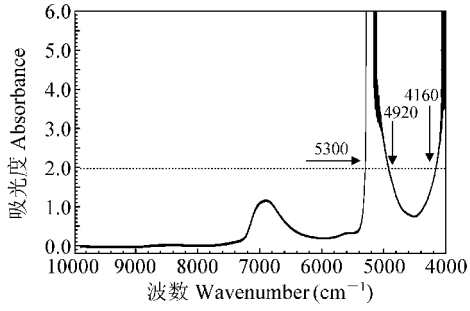

图1 191个血清样品的近红外光谱

Fig. 1 Near-infrared spectra of 191 serum samples

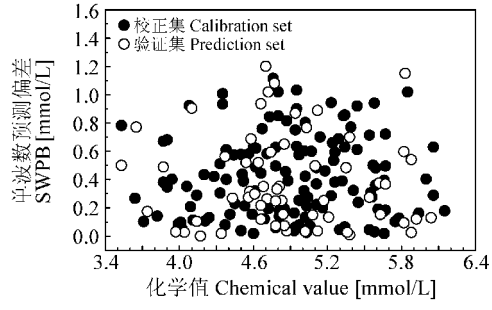

图2 SWPB与化学值的分布

Fig. 2 Distribution of single wavenumber prediction bias (SWPB) and chemical values

3.2 SG平滑模式与PLS因子数的联合优选

为了比较，在SG平滑前直接PLS方法建模。采用10000～5300 cm^{-1}和4920～4160 cm^{-1}组合波段，PLS因子数设置从1到40，按照RMSEP最小遴选最优因子数为8，最优RMSEP值为0.423 mmol/L。此结果优于既往的血清葡萄糖近红外光谱分析效果[3,4]。由此说明，所采用的组合波段（10000～5300 cm^{-1}和4920～4160 cm^{-1}）和校正集验证集的划分方法具有良好建模代表性和预测效果。

表1 校正集验证集SWPB、化学值的均值和标准偏差

Table 1 Mean and standard deviation of SWPB and chemical value in calibration set and prediction set

	化学值 Chemical value		单波数预测偏差 SWPB	
	均值 Mean	标准偏差 Standard deviation	均值 Mean	标准偏差 Standard deviation
校正集 Calibration set	4.91	0.59	0.430	0.294
验证集 Prediction set	4.87	0.60	0.428	0.295

建立计算机算法平台，把全部582种SG平滑模式和不同PLS因子数（1～40）组合分别建立PLS模型，按照预测效果优选SG平滑模式和PLS因子数。各阶导数平滑、各平滑点数的最优模型的RMSEP值（从不同多项式模型、不同PLS因子数中优选）如图3所示。各阶导数平滑（分开对应不同多项式）的最优模型的RMSEP值、最优平滑点数、最优PLS因子数如表2所示。未做SG平滑直接PLS方法建模的结果也列在表2中。全局最优的SG平滑模式为1阶导数平滑，3、4次多项式类型，53平滑点数，对应的最优因子数为7，最优RMSEP为0.376 mmol/L，预测相关系数R_P为0.781，预测效果明显优于未做SG平滑处理的结果。表2和图3表明，不同的导数平滑和不同的多项式次数类型对应的最优平滑点数、最优PLS因子数一般是不相同的。不同的导数平滑、不同的多项式次数类型和采用不同的平滑点数，对应的最优RMSEP值也是差别比较大的。如果根据既往文献或者其它研究对象所采用的平滑

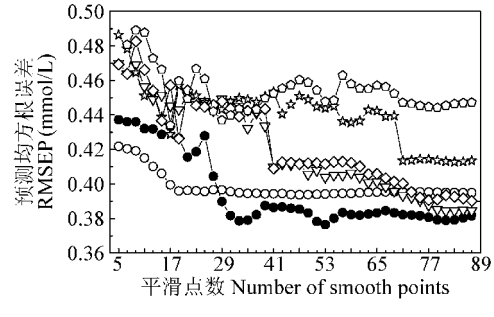

图3 各阶导数的平滑点数对应的最优RMSEP

Fig. 3 Optimal root mean square error of predivtion (RMSEP) corresponding to smooth points number for each order derivation

（○）零阶（0 Order）；（●）一阶（1 Order）；（▽）二阶（2 Order），（☆）三阶（3 Order）；（◇）四阶（4 Order）；（⬠）五阶（5 Order）。

模式的经验,不经过大范围比较筛选,很难得到最优的 SG 平滑模式和 PLS 因子数。另一方面,从表 2 和图 3 还可以看出,最优平滑点数一般都不在25 以内,如果采用 25 以内的平滑点数,就达不到现在的最优预测效果(25 点以内最优 RMSEP 为 0.396 mmol/L),这说明 SG 平滑模式的扩充是非常有必要的。

表 2　各阶导数平滑最优模型的预测效果

Table 2　Prediction effect of optimal model corresponding to each order derivation

	多项式次数 Polynominal degree	平滑点数 Smoothing points	PLS 因子数 PLS factor	预测均方根偏差 RMSEP
未平滑 No smoothing	–	–	8	0.423
零阶 0 Order	2,3	39	9	0.394
	4,5	47	9	0.394
	6	79	17	0.398
一阶 1 Order	2	33	7	0.379
	3,4	53	7	0.376
	5,6	81	8	0.379
二阶 2 Order	2,3	81	7	0.384
	4,5	87	7	0.409
	6	77	10	0.391
三阶 3 Order	3,4	87	7	0.415
	5,6	83	4	0.413
四阶 4 Order	4,5	15	6	0.437
	6	87	9	0.390
五阶 5 Order	5,6	17	7	0.434

为了观察 PLS 因子数对模型效果的影响,对不做 SG 平滑的直接 PLS 模型和最优 SG 平滑 PLS 模型(1 阶导数平滑,3、4 次多项式类型,53 平滑点数)分别给出 PLS 因子数对应的 RMSEP,如图 4 所示。直接 PLS 模型的最优因子数为 8,RMSEP 为 0.423 mmol/L。最优 SG 平滑 PLS 模型的最优因子数为 7,RMSEP 为 0.376 mmol/L。最优因子数不相同,平滑后预测效果提升较大,在 SG 平滑 PLS 模型中因子数的影响更为显著。

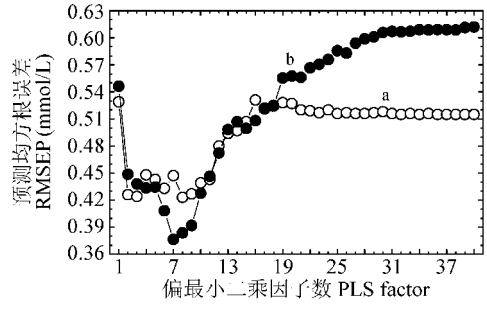

图 4　平滑前后的 PLS 因子数对应的 RMSEP

Fig. 4　RMSEP corresponding to PLS factor before and after smoothing

a. PLS 模型(PLS model); b. 最优 SG 平滑 PLS 模型(Optimal SG smoothing PLS model)。

4　结　论

实验结果表明,SG 平滑模式扩充以及 SG 平滑模式和 PLS 因子数的联合全局筛选都是非常有必要的。所提出的划分校正集验证集的方法、SG 平滑模式扩充、SG 平滑模式和 PLS 因子数的联合全局筛选能够有效地应用于近红外光谱分析的模型优化。

References

1　Burns D A, Ciurczak EW. *Handbook of Near-infrared Analysis*, 2*nd ed*, New York: Marcel dekker inc, **2001**: 633 ~ 647

2　CHU Xiao-Li(褚小立), XU Yu-Peng(许育鹏), LU Wan-Zhen(陆婉珍). *Chinese J. Anal. Chem.*(分析化学), **2008**, 36(5): 702 ~ 709

3　Kasemsumran S, Du Y P, Maruo K, Ozaki Y. *Chemometrics Intell. Lab. Syst.*, **2006**, 82(1-2): 97 ~ 103

4　CHEN Huo-Cai(陈华才), YANG Zhong-Guo(杨仲国), CHEN Xing-Dan(陈星旦). *Chinese J. Anal. Lab.*(分析试验室), **2005**, 4(7): 17 ~ 20

5　XU Hui-Rong(徐惠荣), WANG Hui-Sheng(汪辉君), HUANG Kang(黄 康), YING Yi-Bin(应义斌), YANG Cheng(杨 诚), QIAN Hao(钱 豪), HU Jun(胡 俊). *Spectroscopy and Spectral Analaysis*(光谱学与光谱分析), **2008**, 28(11): 2523 ~ 2526

6 CHEN Xue-Ying(陈雪英), LI Ye-Rui(李页瑞), CHEN Yong(陈 勇), WANG Long-Hu(王龙虎), Ding Ling(丁 玲). *Chinese J. Anal. Chem.*(分析化学), **2009**, 37(10): 1451 ~ 1456
7 YU Yan-Bo(于燕波), ZANG Peng(臧 鹏), FU Yuan-Hua(付元华), ZHANG Lu-Da(张录达), YAN Yan-Lu(严衍禄), CHEN Bin(陈 斌). *Spectroscopy and Spectral Analaysis*(光谱学与光谱分析), **2008**, 28(7): 1554 ~ 1558
8 Savitzky A, Golay M J E. *Anal. Chem.*, **1964**, 36(8): 1627 ~ 1637
9 CHU Xiao-Li(褚小立), YUAN Hong-Fu(袁洪福), LU Wan-Zhen(陆婉珍). *Prog. Chem.*(化学进展), **2004**, 16(4): 528 ~ 542
10 CHEN Jie-Mei(陈洁梅), PAN Tao(潘 涛), CHEN Xing-Dan(陈星旦). *Optics Preci. Eng.*(光学精密工程), **2006**, 14(1): 1 ~ 7
11 CAO Pu(曹 璞), PAN Tao(潘 涛), CHEN Xing-Dan(陈星旦). *Optics Preci. Eng.*(光学精密工程), **2007**, 15(12): 1952 ~ 1958

Joint Optimization of Savitzky-Golay Smoothing Models and Partial Least Squares Factors for Near-infrared Spectroscopic Analysis of Serum Glucose

XIE Jun[1], PAN Tao*[1], CHEN Jie-Mei[2], CHEN Hua-Zhou[1,3], Ren Xiao-Huan[1]
[1](*Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institutes, Jinan University, Guangzhou* 510632)
[2](*Department of Biological Engineering, Jinan University, Guangzhou* 510632)
[3](*Department of Mathematics, Shanghai University, Shanghai* 200444)

Abstract The optimal model for the near-infrared spectroscopic analysis of serum glucose was established by partial least squares(PLS) and Savitzky-Golay(SG) smoothing method. Based on the prediction effect of the optimal single wave number model, a new dividing method for calibration set and prediction set was given. The calibration and prediction models were established by PLS method adopting the combination bands of 10000 – 5300 cm^{-1} and 4920 – 4160 cm^{-1} with Savitizky-Golay(SG) smoothing. By extending the number of smoothing points to 5, 7, …, 87(odd) and polynomial degree to 2, 3, 4, 5, 6, fourteen smooth coefficient tables including 582 smooth modes were calculated. All PLS models corresponding to all smooth modes and all PLS factors(1 – 40) were constructed. The optimal model was selected by the prediction effect. And the derivation order was 1, the polynomial degree was 3 or 4, the number of smoothing points was 53, the optimal factor was 7 and the optimal RMSEP reach 0.376 mmol/L. The dividing method for calibration set and prediction set, the extending of SG smoothing modes, large-scale optimization combining SG smoothing modes and PLS factors can be effectively applied for the model optimization of near-infrared spectroscopic analysis.

Keywords Serum glucose; Near-infrared spectroscopy; Partial least squares; Savitzky-Golay smoothing; Dividing calibration set and prediction set

(Received 19 June 2009; accepted 12 September 2009)