

正交信号校正用于偏最小二乘建模过拟合现象的研究

张 娴, 袁洪福*, 郭 峥, 宋春风, 李效玉, 谢锦春

北京化工大学材料科学与工程学院, 北京 100029

摘 要 通过用近红外光谱和 PLS 建立测定二元、三元调合食用油中花生油含量模型以及二甲亚砷水溶液浓度模型, 使用交互验证统计参数, 包括决定系数 RC , 标准偏差 SEC , 预测值和实际值线性拟合方程的斜率 a 和截距 b , 和外部验证统计参数, 包括决定系数 RV , 标准偏差 SEP , 预测值和实际值线性拟合方程的斜率 a 和截距 b , 评价建模效果, 比较了分别使用原始光谱和正交信号校正 (OSC) 处理光谱的 PLS 建模结果。研究表明: OSC 校正可明显改善近红外光谱与被测物质浓度的线性相关性。虽然 OSC 预处理可以改善 PLS 模型的交互验证结果, 但是外部验证结果变差, 即正交信号校正用于 PLS 建模将导致过拟合现象。通过分析算法原理, 认为 OSC 与 PLS1 对光谱中与性质无关的信息作了重复剔除, 误删除了部分有效信息, 导致过拟合效应。

关键词 正交信号校正; 近红外光谱; PLS; 过拟合

中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2011)06-1688-04

引 言

近红外光谱 (NIR) 分析技术具有同时测量多种性质、速度快、无损测量等优点, 被广泛用于工农业生产诸多领域。近红外光谱分析基于比尔定律, 即光谱与浓度 (或性质) 须呈线性相关关系, 但是, 由于样品成分中的不同化学基团近红外谱带之间存在严重重叠, 以及光谱测量过程引入的多种噪声, 在一定程度上使样品近红外光谱与被测性质偏离线性关系。通常, 近红外光谱分析采用多元校正建模方法, 尤其是偏最小二乘法 (PLS1) 建模对克服这些非线性影响具有明显作用。尽管如此, 上述各种干扰都会对近红外光谱模型预测能力产生不良影响, 有时还比较严重。因此, 近红外光谱建模前, 必须对光谱信号进行数据预处理, 降低其对模型预测能力的影响。

正交信号校正 (orthogonal signal correction, OSC) 是近红外光谱分析信号预处理方法之一^[1,2], 其思想是通过剔除光谱中与被测性质无关的信号, 改善校正后光谱与被分析性质之间的线性相关关系。我们前期的研究工作表明, 当 OSC 用于 PLS1 建模的光谱预处理, 可显著改善交互验证结果, 但不能有效改善 PLS1 模型的预测能力, 有时甚至会更差。本文以近红外光谱法测定调合食用油中花生油含量以及二甲亚砷溶液浓度为研究对象, 通过比较有、无 OSC 校正用于

PLS1 建模的效果, 以期确证和揭示 OSC 校正用于 PLS 建模将产生过拟合现象的原因。

1 实验部分

1.1 数据采集

调合食用油: 使用定量移液管配制二组分调合油样 (花生油、大豆油) 以及三组分调合油样 (花生油、大豆油、葵花籽油), 三种原料食用油均为从市场购买的国家标准一级食用油。二甲亚砷水溶液: 用电子天平称量一定比例的二甲亚砷与纯水, 放入锥形瓶摇匀后静置一天进行测量。所有样品浓度数据如表 1 所示。

Table 1 Concentration statistics of all samples

名称	最大值	最小值	平均值	样品数	精度
二甲亚砷含量/w%	77.923	4.973	41.511	60	0.001
二组分花生油含量/v%	100.00	0	50.00	61	0.05
三组分花生油含量/v%	100.00	0	33.04	78	0.05

在 SupNIR-5500 型近红外光谱仪 (聚光科技公司) 上使用光程为 2 mm 的比色皿测量样品近红外光谱, 分辨率 6 nm, 扫描范围 1000~1800 nm, 扫描次数为 64 次, 样品恒温 60 ℃。连续重复测定 10 张光谱取其平均值作为样品光

收稿日期: 2010-07-19, 修订日期: 2010-10-15

基金项目: 国家 (863 计划) 项目 (2009AA04Z129) 资助

作者简介: 张 娴, 女, 1986 年生, 北京化工大学材料科学与工程学院硕士研究生 * 通讯联系人 e-mail: hfyan@mail.buct.edu.cn

©1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

谱。

1.2 OSC 处理

由上述样本光谱和浓度数据分别组成二元食用油, 三元食用油和二甲亚砷水溶液的样本集。使用各样本集, 按校正: 验证 = 80: 20, 以随机方式产生各自的校正集和验证集。使用 O-PLS 方法对光谱信号进行 OSC 预处理。根据我们关于 OSC 最佳主成分数确定方法的研究(近期将在光谱学与光谱分析上刊登), 对上述 3 个样本集选择的最佳 OSC 成分数依次为 3, 8, 4。使用 MATLAB 7.0 软件, 对校正集进行 OSC 校正, 再使用由校正集 OSC 处理产生的 OSC 权重 (W) 和主成分光谱 (P) 对验证集光谱进行 OSC 校正, 得到 OSC 校正后的验证样本集。

1.3 模型建立

对调合食用油中花生油浓度和二甲亚砷水溶液中二甲亚砷浓度分别建立定量校正模型。分别使用原始光谱和 OSC 校正后的光谱文件, 使用 CM2000 化学计量学软件(聚光科技公司) 建立模型。对于使用原始光谱建立 PLS1 模型, 将样品集随机分成校正集/ 验证集(80/20), 进行均值中心化处理后, 采用 CARS 方法^[3] 选择波长组合建立模型(如表 2 所示), 根据 PRESS 曲线以及 RMSEP 值选择最佳主因子数。对于 OSC 校正光谱, 使用 OSC 校正集采用与原始光谱建模相同的波长分别建立 PLS1 模型, 使用 OSC 验证集对模型进行外部验证。

1.4 模型评价参数

以交互验证和外部验证的决定系数 (RC , RV)、预测值 (y) 与真实值 (x) 之间的线性拟合方程 $y = ax + b$ 中的斜率 a 、截距 b 及预测标准误差 (SEC , SEP) 作为模型优劣的判定依

据。对于理想模型, a 和 b 应分别为 1 和 0; RC 和 RV 应最接近于 1, 且预测标准误差 (SEC , SEP) 应最接近于参考方法的误差。

Table 2 Wavelengths used in modeling	
性质	波长范围/ nm
花生油(二元调合油)	1 203~ 1 218 1 686~ 1 783
花生油(三元调合油)	1 281 1 321 1 322 1 324 1 447 1 449 1 452 1 453 1 454 1 697 1 708 1 711 1 726
二甲亚砷	1 010 1 015 1 016 1 029 1 042 1 053 1 137 1 144 1 168 1 188 1 227 1 238 1 282 1 299 1 327 1 407 1 420 1 488 1 520 1 545 1 566 1 709 1 776

2 结果与讨论

在多种正交信号校正(OSC) 算法中, O-PLS 算法的计算效率和被剔除信号与被测性质的正交性保持相对较好^[4], 本文选择 O-PLS 算法进行光谱校正。前期研究工作发现, OSC 用于 PLS1 建模时只能改善交互验证结果, 不能改善外部预测结果, 由此产生了 OSC 与 PLS1 结合使用时会产生过拟合效应的质疑。通过比较使用原始光谱与 OSC 校正后的光谱建立的 PLS1 模型评价参数, 就可以验证 OSC 和 PLS 联用是否产生过拟合现象。关于调合食用油中花生油浓度模型和二甲亚砷水溶液中二甲亚砷浓度模型, 有、无 OSC 处理的校正模型交互验证结果和预测能力的对比数据分别列于表 2 和表 3, 以下将对结果进行分析和讨论。

Table 3 Cross validation results of PLS1 models using and without using OSC

建模方法及结果	花生油(二元调和油)		花生油(三元调和油)		二甲亚砷水溶液	
采用方法	PLS	PLS+ OPLS	PLS	PLS+ OPLS	PLS	PLS+ OPLS
主因子数	4	4	6	1	5	2
剔除校正样本数*	0	0	4	4	3	3
决定系数 RC	1.000	1.000	1.000	1.000	1.000	1.000
交互验证 a	1.000	1.000	0.999	1.000	1.000	1.000
交互验证 b	0.006	0.003	0.025	0.016	0.001	0.001
SEC	0.36	0.27	0.93	0.69	0.134	0.134

* 对花生油含量(三元调合油) 进行建立模型时, 剔除的样本均为相同的 4 个样本, 均为性质残差过高; 二甲亚砷剔除样本均为相同的三个样本, 理由为性质残差过高。

2.1 有、无 OSC 校正的 PLS1 模型交互验证结果对照

由表 3 可见, 无论食用油还是二甲亚砷水溶液, 所有定量模型的交互验证相关系数 RC 都为 1, 由交互验证结果获得的 a 均趋近于 1, 说明光谱与被测性质之间具有良好的线性关系, 而且光谱数据和浓度数据均具有良好的测量精度。

经过 OSC 预处理后, 光谱与性质之间线性相关性得到明显改善, 如图 1 所示。由模型预测值和参考值计算的参数 a , b , 和 SEC 均得到了明显改善, 这些结果都表明 O-PLS 处理可以有效剔除光谱中的无用信息, 使交互验证结果获得改善。由于交互验证的预测结果与校正集样本有关, 因此, OSC 处理能否真正有效改善模型预测能力则需要外部验证

结果(外部验证样本与校正集样本无关) 进行验证。

2.2 有、无 OSC 校正对 PLS1 模型外部验证结果对照

值得指出的是, 近红外光谱模型准确度评价是以参考测量方法测定结果为基准, 因此, 使用 SEP 数据评价模型时, 不仅仅考察 SEP 数据大小变化, 还必须考虑参考性质数据的测量精度。如果 SEP 数值大于参考性质数据的测量精度, 由此作出的有关评价是有意义的, 反之亦然。花生油浓度和二甲亚砷浓度的参考测量方法的测量精度分别为 0.05 和 0.001。二甲亚砷浓度是称重方法获得的, 建模性质参考数据比较准确。表 4 中所有 SEP 均大于其相应参考测量方法的精度误差。比较表 3 和表 4 的数据不难看出: 交互验证结

果优于外部验证结果,这符合一般建模规律。外部验证结果表明:经过 OSC 校正后,对于二甲亚砷浓度模型外部验证结果的 a 、 b 和 SEP 均变差;食用油含量模型外部验证结果的 a 和 b 变化没有规律,其原因尚不清楚。但是,所有三种模型

的 SEP 都变差了,这就验证了一个基本事实,即如果 OSC 处理和 PLS 校正联用,将产生过拟合现象,使模型预测能力变差。

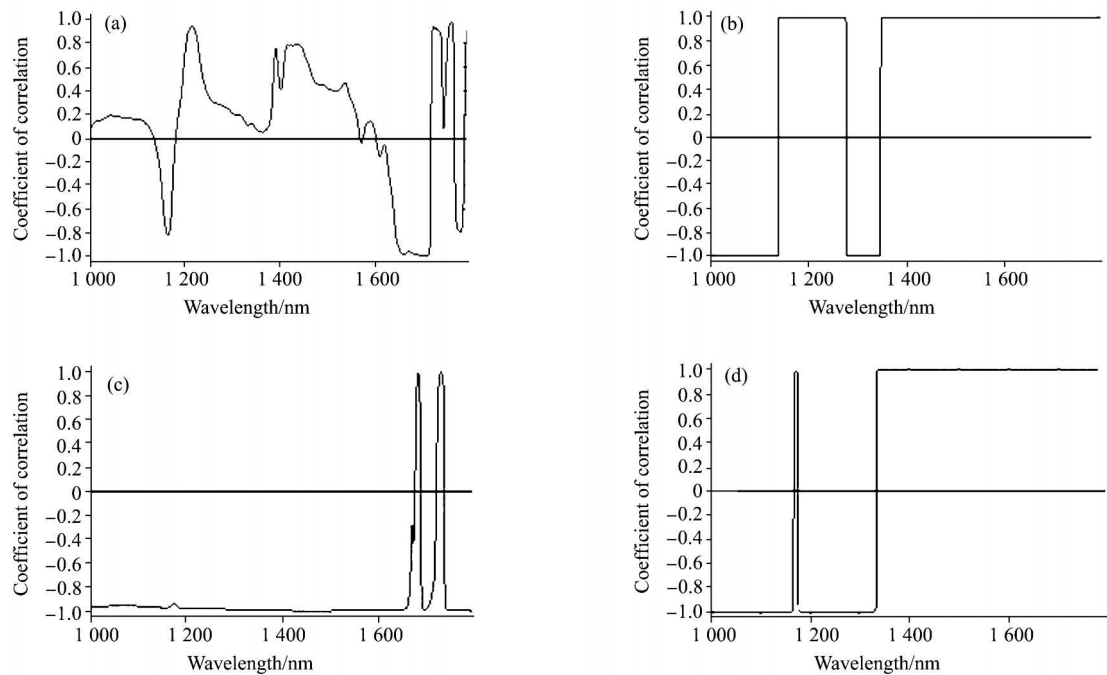


Fig 1 Correlation curve of raw spectra and of OSC corrected spectra

(a): Correlation curve of raw oil spectra; (b): Correlation curve of OSC corrected oil spectra
(c): Correlation curve of raw dimethylsulfoxide solution spectra;
(d): Correlation curve of OSC corrected dimethylsulfoxide solution spectra

Table 4 Predication results of PLS1 models

建模方法及结果	花生油(二元调和油)		花生油(三元调和油)		二甲亚砷水溶液	
采用方法	PLS	PLS+ O-PLS	PLS	PLS+ O-PLS	PLS	PLS+ O-PLS
决定系数 R^2	1.000	1.000	1.000	0.999	1.000	1.000
外部验证 a	1.003	0.999	0.999	0.996	0.999	0.987
外部验证 b	-0.104	0.112	0.504	0.488	0.103	0.775
SEP	0.61	0.63	0.66	0.69	0.105	0.583

2.3 过拟合原因分析

从算法分析 OSC 用于 PLS1 建模导致过拟合现象,可以看出,OSC 通过内置 PLS 迭代计算得到光谱中与被测性质无关的信息部分,再将其从原始光谱中扣除,从而改善了校正后光谱与被测性质之间的线性相关关系。而 PLS1 通过迭代计算,求出与被测性质相关的主成分光谱,通过选取一定数目(从第一主成分开始计)与性质相关的主成分光谱重构光谱进行建模,也起到隔离原始光谱中与性质无关信息的作用。两种方法目标是一致的,但是其技术路线不同。虽然 OSC 校正可以提高拟合光谱与性质之间的相关程度,但当 OSC 和 PLS1 一起使用时,在整个计算过程中,对与性质无关的信息做了两次剔除(即重复剔除),导致光谱中部分有用信息被误剔除,从而使模型预测能力变差。因此,建议不要

OSC 和 PLS 联用进行建模。但 OSC 对剔除光谱中与被测性质无关的信号的确具有很好的能力,如果与其他不存在重复剔除信息的校正算法,如多元线性回归(MLR)等,联用,则可避免重复剔除导致过拟合现象,有可能改善模型预测能力。今后,我们将进一步研究和证实这一观点。

3 结 论

OSC 校正可以有效地剔除光谱中与被测性质无关的信号,与 PLS1 联用建模时,可明显改善模型的交互验证结果,但是并不能改善模型的预测能力,反而使 PLS1 模型预测能力变差,即存在过拟合现象。因此,建模避免将 OSC 和 PLS1 联用建模。

References

- [1] Wold S, Antti I, Lindgren R, et al. *Chemometrics and Intelligent Laboratory Systems*, 1998, 44: 175.
- [2] Bertran E, Iturriaga H, Maspocho S, et al. *Analytica Chimica Acta*, 2001, 431: 303.
- [3] Li Hongdong, Liang Yizeng, Xu Qingsong, et al. *Journal of Computational Chemistry*, 2010, 3(3): 592.
- [4] Trygg J, Wold S. *Journal of Chemometrics*, 2002, 16: 119.

Study of the Over-Fitting in Building PLS Model Using Orthogonal Signal Correction

ZHANG Xian, YUAN Hong-fu*, GUO Zheng, SONG Chun-feng, LI Xiao-yu, XIE Jir-chun

College of Materials Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Abstract In the present paper, the over fitting phenomenon in building PLS model using orthogonal signal correction (OSC) was studied through establishment of quantitative calibration models for the peanut oil content in blending edible oils, and for the dimethylsulfoxide concentration in water solution. The cross validation results and the predication results of PLS models using OSC and without using OSC were compared to evaluate the effectiveness of OSC for improving the performance of PLS1 model. The results show that the application of OSC to PLS modeling will lead to an over fitting phenomenon. According to the principles of their algorithms, when OSC and PLS are used together, the signals which are not correlated to the interested property are removed twice from the raw spectra. This leads to deleting the parts of useful information in spectra, and to spoiling the predictive ability of PLS models to some extent.

Keywords Orthogonal signal correction; Near infrared spectroscopy; PLS; Over fitting

(Received Jul. 19, 2010; accepted Oct. 15, 2010)

* Corresponding author