Vol. 28 高等学校化学学报 No. 2
2007 年 2 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 246 ~ 249

基于多模型共识的偏最小二乘法用于近红外光谱定量分析

李艳坤, 邵学广, 蔡文生
(南开大学化学系, 天津 300071)

摘要 建立了多模型共识偏最小二乘(cPLS)建模方法, 并应用于烟草样品近红外(NIR)光谱与常规成分氯含量之间的建模研究, 探讨了建模参数对预测结果的影响. 结果表明, cPLS 方法与传统的偏最小二乘算法(PLS)相比, 所建模型更稳定可靠, 预测结果也可得到了明显改善.

关键词 多模型共识; 偏最小二乘法; 近红外光谱; 烟草样品; 定量分析
中图分类号 O65 **文献标识码** A **文章编号** 0251-0790(2007)02-0246-04

近年来, 近红外光谱(NIR)法以其快速、简便及无损等特点[1,2], 在复杂样品化学成分的测定中占有重要地位. 由于近红外光谱产生于分子振动, 吸收较弱, 吸收峰严重重叠, 且多组分复杂样品的近红外光谱往往不是各组分光谱的简单叠加, 必须借助于化学计量学方法才能进行定性定量分析. 因此, 化学计量学方法已成为近红外光谱分析中的研究热点. 各种多元校正技术, 如多元线性回归(MLR)[3]、主成分回归(PCR)[4]、偏最小二乘回归(PLS)[5]和人工神经网络(ANN)[6]等方法在近红外光谱分析中已得到了广泛应用.

在近红外光谱分析中, 建立可靠的定性和定量模型是对未知样品的类别、组成或性质做出准确预测的前提. 因此, 建模方法的研究一直是其核心内容之一. 传统的多元校正技术(如 PLS 和 PCR)一般采用单一模型, 即首先采用一定的训练集建立一个最优模型, 然后用于测定. 但是, 当训练集样本数目有限或者存在较大误差时, 模型的预测精度与稳定性往往不能令人满意. 共识策略(Consensus strategy)[7~13]采用同一训练集中的不同子集建立多个模型同时进行预测, 将多个预测结果通过简单平均或加权平均作为最终的预测结果, 从而获得更高的预测精度和稳定性.

共识策略与 PCR[7]、ANN[8,9]或 Decision Tree[10,11]、PLS[12]算法结合在定量构效关系(QSAR)等[13,14]研究中的应用均已取得较好结果. 本文采用基于多模型共识的偏最小二乘法(Consensus-PLS, 简称为 cPLS)对烟草样品中氯的含量与近红外光谱的模型进行了研究, 结果表明, cPLS 较 PLS 算法的模型更加稳定可靠, 预测结果也得到了明显改善.

1 原理与算法

共识策略的基本思想是采用随机或组合的方式利用同一训练集中的不同子集建立多个模型同时进行预测, 将多个预测结果的均值作为最终结果. 其突出特点是通过多次使用训练集中不同子集样本的信息, 降低了预测结果对某一样本的依赖性. 因此, 该方法有望解决过拟合问题, 从而提高模型的预测稳定性.

多模型共识算法的预测误差可表示为[8,9]

$$e(\vec{x}) = \overline{\varepsilon}(\vec{x}) - \overline{\alpha}(\vec{x}) \quad (1)$$

收稿日期: 2006-03-23.
基金项目: 国家自然科学基金(批准号: 20325517, 20575031)和教育部博士学科点基金(批准号: 20050055001)资助.
联系人简介: 蔡文生(1965 年出生), 女, 博士, 教授, 博士生导师, 主要从事计算机化学与化学信息学研究.
E-mail: wscai@ nankai. edu. cn

$$\bar{\varepsilon}(\vec{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(\hat{y}_i - y)^2 \tag{2}$$

$$\bar{\alpha}(\vec{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(\hat{y}_i - \hat{y})^2 \tag{3}$$

式中，N_m 为成员模型的总数，$\vec{x}$ 为某样品的光谱，y 为某样品浓度的实际测量值，$\hat{y}_i$ 为第 i 个成员模型的预测结果，$\hat{y}$ 为某样品浓度的最终预测结果. $\bar{\varepsilon}(\vec{x})$ 为所有成员模型的平均误差，$\bar{\alpha}(\vec{x})$ 表示成员模型相对于整体模型的方差. 本文中最终预测结果采用 Nm 个成员模型的简单平均值，即

$$\hat{y} = \frac{1}{N_m}\sum_{i=1}^{N_m}\hat{y}_i \tag{4}$$

从式(1)可以看出，影响整体模型误差 $e(\vec{x})$ 大小的因素有两个：一是成员模型的平均误差 $\bar{\varepsilon}(\vec{x})$，二是成员模型预测结果的方差. 由此可见，成员模型的个数及其多样性可以提高整体模型的预测精度和稳定性.

本文采用的 cPLS 建模过程如下：(1) 设置 cPLS 的有关参数，如训练集与检验集的样本数、成员模型数及成员模型的接纳标准等；(2) 将训练集样本随机地分为训练子集和检验集；(3) 以训练子集的样本建立 PLS 模型；(4) 用所建立的模型预测检验集；(5) 根据检验集预测结果与相应实验值之间的平均相对误差作为成员模型的接纳标准，确定该模型是否接纳为 cPLS 的成员模型；(6) 重复(2)～(5)步至 cPLS 的模型数达到预定值.

2　实验部分

2.1　仪器与样品

Vector 22/N FT-NIR System(Bruker)；Auto Analyzer Ⅲ型连续流动分析仪(BRAN + LUBBE).

样品为 58 个牌号的烤烟型卷烟产品，按照 YC/T31-1996 制备成粉末样品，平均粒度为 0.45 mm.

2.2　实验与计算方法

样品中氯的含量采用 AAⅢ型连续流动分析仪按照标准方法测定. NIR 光谱采用 FT-NIR 光谱仪测定，波数范围为 4000～9000 cm^{-1}，采样间隔为一个波数(共 5001 点).

从 58 个样品的数据集中随机取出 20 个样本作为预测集，其余样本作为训练集和检验集用于 cPLS 的建模. 为了进行比较，在 cPLS 和 PLS 的计算中采用了相同的训练集、检验集(PLS 不使用)、预测集和主成分数. 检验集样本数和主成分数分别采用 6 和 20. 参数的优化判据为预测值的均方根误差(Root mean square error of prediction，RMSEP). 程序采用 MATLAB 语言编写.

3　结果与讨论

3.1　成员模型的筛选

如上所述，cPLS 由一组模型构成，每一个模型称为成员模型(Member model). 由于 cPLS 的模型是根据随机选择的训练集而建立的，必须符合一定的条件才能接纳为成员模型. 本文采用检验集样本的预测值与真实值(实验值)的相对误差作为接纳成员模型的判据，即只有误差小于某一阈值的模型才能成为 cPLS 的成员模型.

由公式(1)可知，cPLS 的预测误差不仅取决于各成员模型预测误差的均值[$\bar{\varepsilon}(\vec{x})$]，而且还取决于成员模型的差异[$\bar{\alpha}(\vec{x})$]. 二者数值相当时，cPLS 的预测结果最佳. 显然，成员模型的接纳标准越高(即误差要求越小)，每个成员模型的预测值越准确，$\bar{\varepsilon}(\vec{x})$ 和 $\bar{\alpha}(\vec{x})$ 均会减小，但由于固有误差的存在，前者的减小是有一定限度的；反之，$\bar{\varepsilon}(\vec{x})$ 和 $\bar{\alpha}(\vec{x})$ 则可能有所增加，但当接纳标准合适时，二者具有相近的数值. 因此，接纳成员模型的误差判据是 cPLS 的重要参数.

分别采用不同的相对误差判据[即建模过程(5)中定义的成员模型接纳标准]进行统计，结果表明，采用 10%，20%，30%，50% 和 100% 作为判据时运行 40 次得到的平均 RMSEP 分别为 0.0447，0.0436，0.0436，0.0437 和 0.0438. 可见，尽管 RMSEP 的差别并不很大，但与式(1)的理论模型具有

较好的一致性，即随着相对误差的提高，RMSEP 先小幅度下降，然后稍有升高. 对计算过程的跟踪结果表明，只有很少数随机模型的相对误差大于 30%. 因此，本文采用 30% 作为误差判据的阈值.

3.2 成员模型总数的确定

从 cPLS 的原理可以看出，多模型共识算法的优势在每个成员模型给出不同的预测结果时才能体现出来，cPLS 算法中应该包括尽可能多的不同预测结果的成员模型[14]. 因此，成员模型的总数是另一个重要参数，对预测结果的稳定性和准确性起着关键作用.

本文选取模型数为 1 ~ 100 进行了计算，预测集 RMSEP 随模型数的变化曲线如图 1. 由图 1 可以看出，模型数较少(<40)时 RMSEP 较大且不稳定，但模型数增至 60 以上时，RMSEP 趋于稳定. 故本文采用了 60 个成员模型.

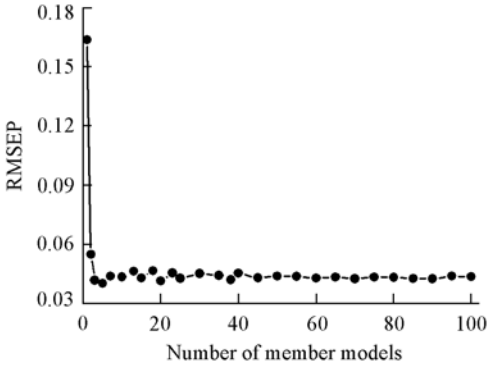

Fig. 1 Variation of RMSEP with the change of number of member models in cPLS

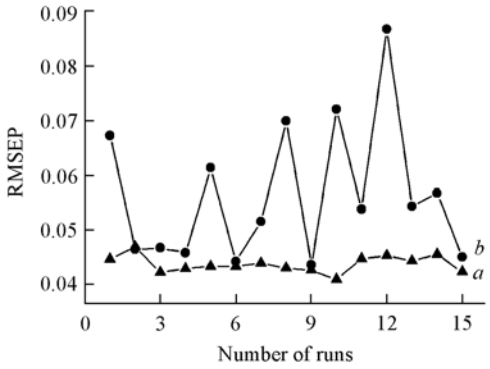

Fig. 2 RMSEPs obtained by cPLS(*a*) and PLS(*b*) in 15 runs of prediction

3.3 cPLS 与 PLS 的比较

因为 cPLS 采用了多个模型的平均，所以预测稳定性是其重要特点之一.

图 2 是样品分别经 15 次 cPLS 和 PLS 运算的预测结果比较. 可见，用 PLS 算法 15 次预测的 RMSEP 之间相差较大，表现出模型的稳定性较差；而用 cPLS 算法 15 次预测结果的波动很小，表现出非常好的稳定性. 从图 2 还可以看出，尽管有时 PLS 与 cPLS 的预测误差相近，但总体来说，cPLS 的预测误差要优于 PLS.

为进一步对 cPLS 和 PLS 的预测结果进行比较，图 3 分别给出了两种方法预测结果与实验值的相关关系. 由于 PLS 的预测结果波动较大，故采用 15 次预测结果的平均值. 同时，图 3 中还给出了相关系数、标准偏差和 15 次运算中的最低及最高回收率. 可见，cPLS 的预测结果与实验值之间的相关系数和标准偏差均优于 PLS，cPLS 的预测回收率范围(87.25% ~ 111.12 %)也同样优于 PLS(62.65% ~ 120.34%).

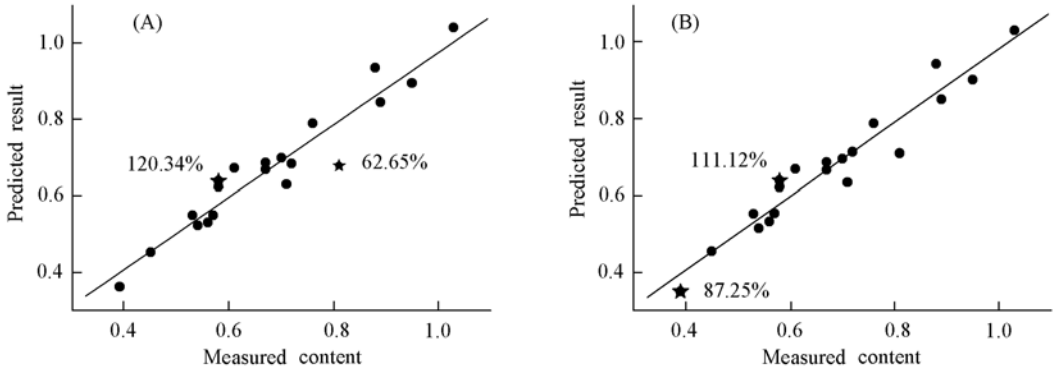

Fig. 3 Relationship between the measured contents and the average results predicted by PLS(A) and cPLS(B) in 15 runs

(A) $R = 0.9570$, SD = 0.0494; (B) $R = 0.9653$, SD = 0.0448.

4 结 论

基于多模型共识的基本思想，建立了一种用于近红外(NIR)光谱建模的多模型共识偏最小二乘(cPLS)方法，并对建模参数进行了讨论. 由于 cPLS 基于多模型进行预测，与传统的 PLS 相比所建立的模型更加稳定、可靠，预测结果也明显得到改善. 因此，cPLS 为近红外光谱定量分析提供了一种新的建模方法，对于克服 PLS 建模方法在样品复杂且校正集样品较少时的不稳定性具有一定意义.

参 考 文 献

[1] Bruno-soares A. M., Murray I., Paterson R. M., *et al*.. Animal Feed Sci. Tech.[J], 1998, **75**: 15—25

[2] Borjesson T., Stenberg B., Linden B., *et al*.. Plant and Soil[J], 1999, **214**: 75—83

[3] Ben-Gera I., Norris K. H.. J. Food Sci.[J], 1968, **33**(1): 64—67

[4] Thomas E. V., Haaland D. M.. Anal. Chem.[J], 1990, **62**: 1091—1099

[5] Geladi P., Kowalski B. R.. Anal. Chim. Acta[J], 1986, **185**: 1—17

[6] Borggard C., Thodberg H.. Anal. Chem.[J], 1992, **64**: 545—551

[7] Aleksander J., Kazimierz W., Hanna W.. Electroanalysis[J], 2005, **17**: 1477—1485

[8] Tesauro G., Touretzky D. S., Leen T. K.. Advances in Neural Information Processsing Systems 7[M], Cambridge MA: MIT Press, 1995: 231—238

[9] Navone H. D., Granitto P. M., Verdes P. F., *et al*.. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial[J], 2001, **12**: 70—74

[10] Drucker H., Cortes C.. Boosting Decision Trees[M], Cambridge MA : MIT Press, 1996, **8**: 479—485

[11] Tong W. D., Hong H. X., Fang H., *et al*.. J. Chem. Inf. Comput. Sci.[J], 2003, **43**: 525—531

[12] Geladi P., Esbensen K. H.. Journal of Chemometrics[J], 1990, **4**(5): 337—354

[13] Govindan S., Douglas B. K.. J. Comput. Aided Mol. Des.[J], 2003, **17**(10): 643—664

[14] Nicolas B., Jean C. M., Eric A., *et al*.. J. Chem. Inf. Comput. Sci.[J], 2004, **44**: 276—285

Partial Least Squares Regression Method Based on Consensus Modeling for Quantitative Analysis of Near-Infrared Spectra

LI Yan-Kun, SHAO Xue-Guang, CAI Wen-Sheng*

(*Department of Chemistry*, *Nankai University*, *Tianjin* 300071, *China*)

Abstract Consensus modeling averages the results of multiple independent models to obtain a single prediction, which avoids the instability of a single model. Based on the philosophy of consensus modeling, a consensus partial least squares regression(cPLS) method was proposed and applied to building the quantitative model of NIR spectra of tobacco samples. Through an investigation of the parameters involved in the modeling, a satisfied model was achieved for predicting the content of chlorine in tobacco samples. With repeated independent runs, cPLS model was found to be more robust and credible than PLS model. Furthermore, compared with PLS method, cPLS model gives more stable and accurate prediction results.

Keywords Consensus modeling; Partial least squares; Near-infrared spectroscopy; Tobacco sample; Quantitative analysis

(Ed.: K, G)