第30卷, 第8期
2010年8月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 30, No. 8, pp2088-2092
August, 2010

一种多模型融合的近红外波长选择算法

洪明坚[1,2], 温志渝[1]

1. 重庆大学微系统研究中心, 重庆 400030
2. 重庆大学软件学院, 重庆 400030

摘 要 针对近红外光谱数据的特点, 在分析了单模型波长选择方法的基础上, 提出了一种多模型融合的变量选择方法。它融合多个模型的回归系数, 以提高波长选择的准确性和稳定性。并用3个业界标准的近红外光谱数据集对提出的方法进行了验证, 同时与UVE-PLS和GA-PLS算法进行了比较。实验结果表明, 经该方法选择变量后, 提高了模型的预测能力, 降低了复杂度, 达到甚至优于UVE-PLS和GA-PLS, 而且具有算法简单、效率高的优点, 具有广泛的实用价值。

关键词 融合; 波长选择; 光谱分析; 近红外
中图分类号: O657.3 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2010)08-2088-05

引 言

近红外光谱分析作为一种微弱信号的检测与分析方法, 在农业、化工、制药和环境监测等领域获得了广泛的应用[1]。随着仪器技术的进步, 近红外光谱的波长点(即变量)个数已经达到几百甚至上千。但是, 这些变量存在严重的共线性(collinear)问题[2]。偏最小二乘(partial least squares, PLS)[3]作为一种全谱分析工具, 有效地解决了共线性问题。然而, 并非所有的变量都对模型的建立做出了积极的贡献。有些变量不仅对建立模型没有任何帮助, 甚至还会影响所建立的建模质量, 导致复杂的模型和较大的预测误差。剔除这些变量, 对于提高模型的可理解性和预测能力具有重要的意义。

常见的变量选择方法可以粗略地分成两大类: 一类是对所有变量逐一考察, 并剔除或者保留。传统的方法如后向剔除法(Backward elimination)、前向选择法(Forward selection)和逐步回归法(Stepwise regression)等[2]都属于这一类。文献[4]提出了BVSPLS, 它对每一个变量剔除一次(leave-one-variable-out), 并用剩余的变量建立模型并评估预测能力, 选择预测能力最强的模型, 并将对应的变量真正剔除; 对剩余的变量重复上述过程, 直到模型的预测能力开始下降为止。文献[5-7]等将变量选择问题作为全局优化问题, 采用了随机搜索算法, 如遗传算法(genetic algorithm, GA)、退火算法(simulated annealing, SA)和Tabu算法等进行搜索, 得到(局部)最优变量组合。至于第二类方法, 它们首先将变量进行排序, 再确定阈值, 最后剔除低于阈值的变量或保留高于阈值的变量。文献[8]首次提出基于PLS回归系数的变量选择, 将回归系数的绝对值作为排序指标。而文献[9]通过对各种变量的排序指标进行研究和实验, 发现PLS回归系数对于变量选择是最为有效的指标。文献[10]提出了UVE-PLS, 基本思想是如果一个变量的重要性比人工引入的随机变量小, 那么这个变量应该被剔除。它采用的指标是PLS回归系数的稳定性。基于类似的思想, 文献[11]提出了RT-PLS方法, 它通过随机测试(randomization test, RT)建立一系列随机回归模型并获取回归系数, 然后保留回归系数能够显著地大于随机模型回归系数所对应的变量。除此之外, 最近人们也开始研究的基于非线性建模方法的变量选择[12,13]。

然而, 前述基于回归系数的变量选择方法有一个共同的缺点, 即它们在度量变量的重要性时, 只是根据单个模型的回归系数做出判断。本文首先对回归系数进行了分析, 总结了单个模型的回归系数用于变量选择的存在的问题, 进而提出了一种基于多模型融合的变量选择算法。该算法首先利用蒙特卡洛交叉验证[14](MCCV)选择最佳PLS成分, 再对多个模型的回归系数进行融合, 得到变量选择的依据。然后, 确定最佳阈值, 选择最终参与建模的变量。最后, 通过3个典型的数据集进行了测试, 验证了方法的有效性。

收稿日期: 2009-10-09, 修订日期: 2010-01-16
基金项目: 国家科技部国际合作项目(2007DFC00040)和国家(863计划)重点项目(2007AA042101)资助
作者简介: 洪明坚, 1977年生, 重庆大学光电工程学院博士研究生 e-mail: hongmingjian@gmail.com, hmj@cqu.edu.cn
© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1　方　法

1.1　PLS回归系数分析

一般地，线性回归模型具有如下形式

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{E} \quad (1)$$

其中 $\boldsymbol{X}_{n\times p}$ 是光谱矩阵，$\boldsymbol{y}_{n\times 1}$ 一般是浓度向量，$\beta_{p\times 1}$ 是回归系数，$\boldsymbol{E}$ 是残差项，n 是样本个数，p 是波长点(变量)个数。对于近红外光谱数据，一般有 $n<<p$，因此 $\boldsymbol{X}_{n\times p}$ 的各个变量之间存在严重的共线性问题。此时，β 显然不是唯一的。偏最小二乘(partial least squares，PLS)是求解(1)最常用的方法，它用 NIPALS[3] 或 LBD[15] 算法对矩阵 $\boldsymbol{X}_{n\times p}$ 进行分解 $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{U}^{\mathrm{T}}$，其中 $\boldsymbol{W}_{n\times a}$ 和 $\boldsymbol{U}_{p\times a}$ 的列向量互为正交，$\boldsymbol{R}_{a\times a}$ 为上 3 对角矩阵。从而，得到 $\boldsymbol{X}_{n\times p}$ 的广义逆矩阵 $\boldsymbol{X}^{-} = \boldsymbol{U}\boldsymbol{R}^{-1}\boldsymbol{W}^{\mathrm{T}}$，最后得到 $\beta_a = \boldsymbol{X}^{-}\boldsymbol{y}$，其中 $1 \leqslant a \leqslant \mathrm{rank}(\boldsymbol{X})$ 称为成分(principal component，PC)个数，它度量了模型的复杂度，值越大越复杂。如果 a 太小，PLS 没有充分提取样本的有效信息，导致模型欠拟合(under fitting)；如果 a 太大，则可能提取了样本中的噪声、背景等无效信息，导致过拟合(over fitting)。一般通过交叉验证(cross validation，CV)求得最优的 a，例如最常用的留一法交叉验证(leave one out CV，LOOCV)。但是，LOOCV 确定的 a 仍然导致过拟合[16]，许青松[14]等提出的 Monte Carlo 交叉验证(MCCV)方法一定程度上克服了这个问题，已被广泛使用。

回归系数 β 也被称为回归权重，它的每一个元素的绝对值大小给出了对应的变量在模型中的重要程度。文献[8]通过 LOOCV 获得最优成分个数 $a=A$，然后基于 β_A 进行变量选择。但是，基于单个 β_A 进行变量选择存在下面两个问题：(1) A 的选择比较困难，因为通过交叉验证主要验证模型的预测能力，得到的 A 只适合于建立预测能力较强的模型，不一定适合于选择变量；(2)即使选定了 A，变量对于模型的贡献并不是仅仅通过单个 β_A 就可以衡量，也就是说，变量对各个模型的贡献可能不同。下面用一个例子来说明这两个问题。

这里采用后面第 2 节中的 meat 数据集来做实验。首先，通过 MCCV 确定 $A=5$，且最小的均方预测误差(root mean square error of prediction，RMSEP)也是在 $A=5$ 时得到的。图 1(a)分别给出了用 β_4 和 β_5 做变量选择的结果。图 1(b)表明基于 β_4 选择的变量所建立的模型更加紧凑、具有更小的 RMSEP。因此，尽管 $A=5$ 适合于建立模型，但却不适合用于变量选择。其次，从图 2 可以看出，在 850 和 960 nm 附近回归系数 β_4 和 β_5 差别比较大，如果用 β_5 来衡量变量的重要性并不能提高模型的预测性能，甚至会造成更复杂的模型。

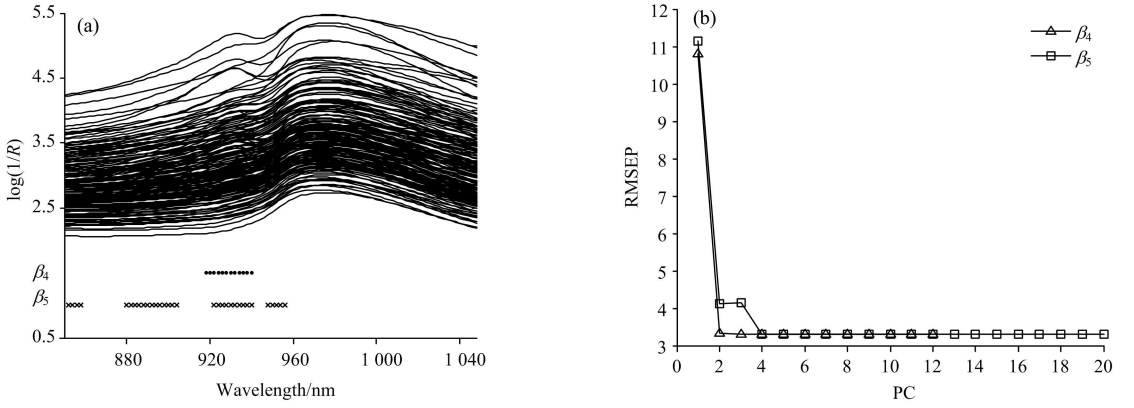

Fig 1　(a) Variables selected by β_4 and β_5 for meat dataset; (b) Comparison of RMSEP

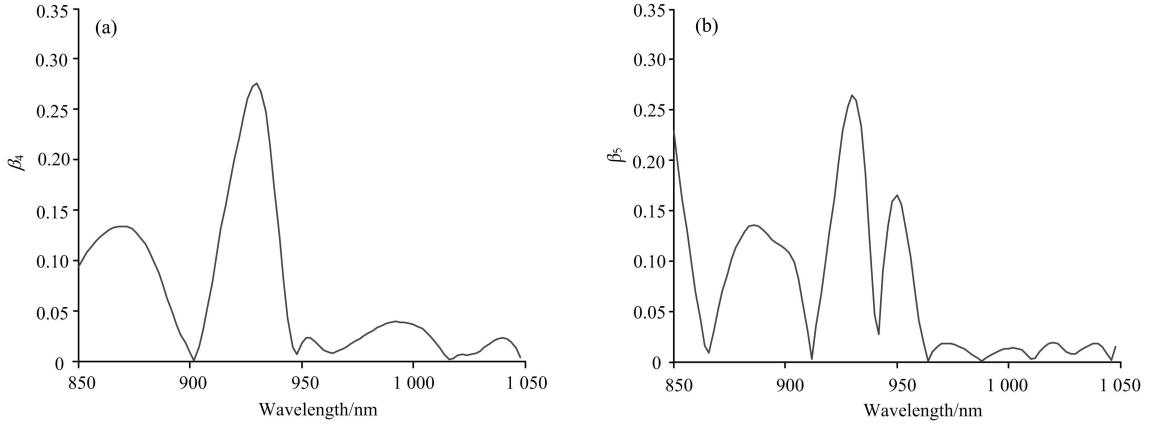

Fig 2　PLS regression coefficients

(a): β_4; (b): β_5

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

1.2 多模型融合波长选择方法

针对上面分析的问题，提出一种新的变量选择算法，基本思想是融合多个模型的回归系数，提高变量选择的准确性和稳定性。考虑到变量对各个模型的贡献不同，首先用 MCCV 得到最佳成分个数 K，然后将 1 至 K 的各个模型的回归系数归一化并求和，作为变量选择的依据。最后，确定阈值，选择所有大于阈值的变量。具体算法流程如下:

Step 1: 对 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 扣均值，并对 $\boldsymbol{X}$ 方差归一化;

Step 2: 用 MCCV 选择成分个数 K;

Step 3: 建立 K 个成分的 PLS 模型，得到 β_a (a= 1, …, K)，求绝对值并归一化 $\beta_a = \frac{|\beta_a|}{\|\beta_a\|}$;

Step 4: 对 β_a 求和，得到 $\boldsymbol{b}_K = \sum_{a=1}^{K}\beta_a$;

Step 5: 确定阈值 b_{cut}，选择所有{$i|\boldsymbol{b}_{K,i} > b_{cut}$, i= 1, 2, …, p}的变量;

Step 6: 用所选择的变量建立回归模型并计算均方预测误差。

三点说明:

(1) 该算法有两个参数，分别是 K 和 b_{cut}。K 的选择主要使得 $\boldsymbol{b}_K$ 尽可能正确反映各个变量真正的重要性。如果 K 太小，有些变量的重要性没有体现出来; 反之，如果 K 太大，将在 $\boldsymbol{b}_K$ 中引入大量的噪声，误导变量选择。因此，这里仍然采用 MCCV 来选择 K。b_{cut} 的优化采用下面的方法: 先计算的 $\boldsymbol{b}_K$ 均值 $\mu_K = \frac{1}{p}\sum_{i=1}^{p}\boldsymbol{b}_{K,i}$ 和标准差 $\sigma_K = \frac{1}{p-1}\sum_{i=1}^{p}(\boldsymbol{b}_{K,i} - \mu_K)^2$，然后选择变量{$i|\boldsymbol{b}_{K,i} > b_{cut} = \mu_K + r\sigma_K$}，其中 $r \in [0, 3]$。为了确定 r，以 0.1 为间隔等分$[0, 3]$，通过均方预测误差来确定最优的 r。

(2) 该算法没有对回归系数相乘进行融合，因为相乘将选择对所有模型都重要的变量，导致过多的变量丢失，而用求和进行融合将选择对某一个或几个模型有贡献的变量，尽可能不遗漏有贡献的变量。

(3) 该算法有 3 个特点: ①融合了多个模型的回归系数，不依赖于单个模型的回归系数，使得选择的变量更加准确和稳定; ②所选择的变量由一些连续的区域组成，容易解释和理解，而 GA，SA 和 Tabu 等算法常常选择一些离散的点; ③与 GA 等全局搜索算法相比，该算法计算简单、效率高。

2 实验部分

为了验证算法的有效性，采用了 4 个业界标准的数据集对算法进行了测试，限于篇幅这里给出其中 3 个的测试结果。所有数据集的样本都采用 DUPLEX 方法[17] 分成训练集、验证集和测试集，其中训练集用于建立模型和选择 K，验证集用于选择 r，测试集用于做最终测试。

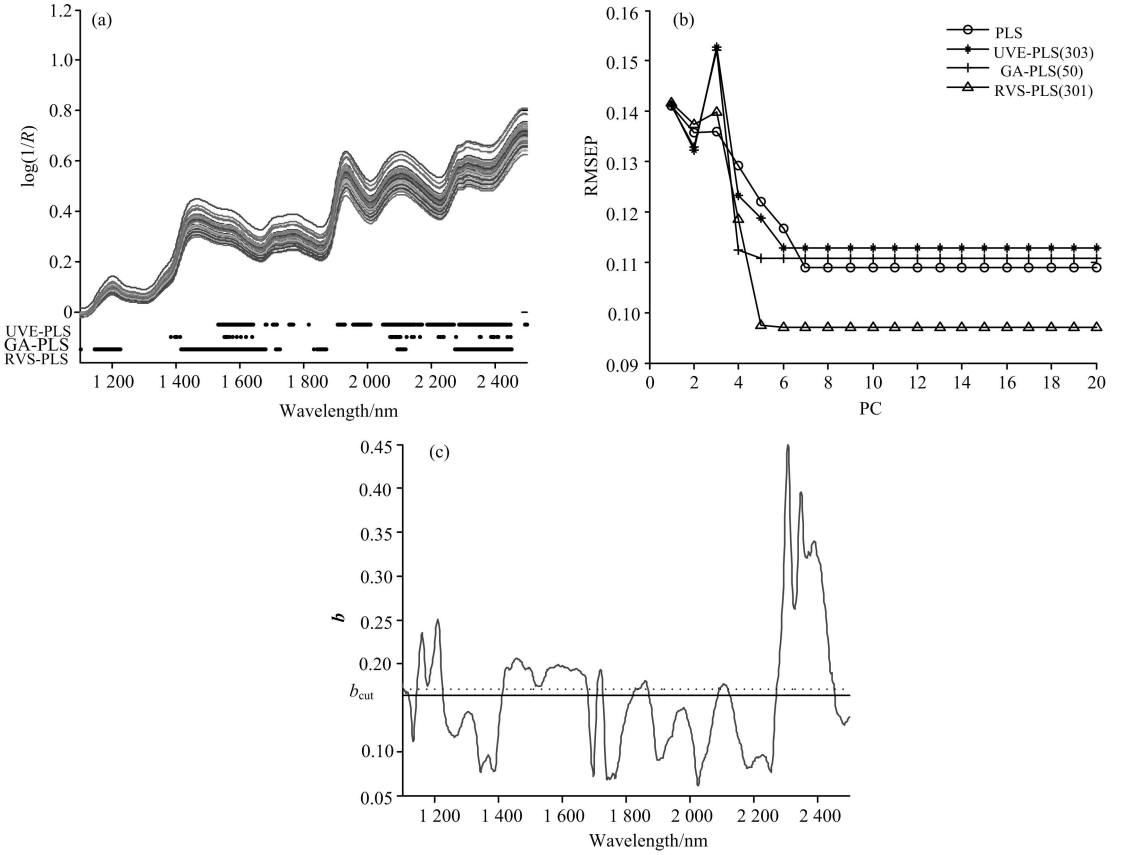

Fig 3 (a) Variables selected for corn dataset; (b) Comparison of RMSEP, the number in the parenthesis is the number of variables selected; (c) b_K and b_{cut} of RVS PLS

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.1 数据集

(1) 玉米(corn)的近红外光谱与植物油(oil)含量。光谱采集范围 1 100~ 2 500 nm, 步长 2 nm。训练集 50 个样本, 验证集和测试集各 15 个。该数据集可以从 http://www.eigenvector.com 下载。

(2) 肉类(meat)的近红外光谱和脂肪(fat)含量。光谱采用 Tecator Infratec Food and Feed Analyzer 记录, 采集范围 850~ 1 050 nm, 步长 2 nm。训练集 120 个样本, 验证集和测试集各 60 个。该数据集可以从 http://lib.stat.cmu.edu/datasets/tecator 下载。

(3) 汽油[18](gasoline)的近红外光谱和辛烷(octane)值。光谱采集范围是 900~ 1 700 nm, 步长 2 nm。前 100 个变量(即 900~ 1 100 nm)几乎没有信息, 事先把它们删除了。训练集 40 个样本, 验证集和测试集各 10 个。

2.2 结果及分析

为了叙述方便, 本文提出的方法记为 RVS-PLS。实验选择了 UVE-PLS 和 GA-PLS 两个典型的算法与 RVS-PLS 进行了比较。UVE-PLS 和 GA-PLS 的实现分别来自 chemoAC (http://vub.vub.ac.be/~fabi/research/chemoac) 和 genpls (http://www.models.kvl.dk/source/GAPLS)。实验过程中, UVE-PLS 和 GA-PLS 没有用到验证集样本。

2.2.1 Corn 数据集

从图 3(a) 可以看出, 尽管 3 种方法所选择的变量不尽相同, 但是还是有明显的重叠区域。UVE-PLS 和 GA-PLS 选择的变量比较分散, 没有形成连续的区域。RVS-PLS 得到了几个连续的变量区域, 比其他两种方法多出了 1 142~ 1 224 nm, 显然这个区域对应于 C—H 键的二级倍频[1]。根据图 3(b), RVS-PLS 明显地提高了模型的预测性能, 但是 RVS-PLS 选择的变量个数(301 个)多了一些, 与 UVE-PLS (303 个)相当。此外, 尽管 GA-PLS 选择的变量最少(50 个), 但是预测误差表明它可能漏选了一些重要的变量。

2.2.2 Meat 数据集

图 4(a) 给出了 3 种方法变量选择的结果。但是, 3 种方法都没有显著提高模型的预测性能, 只是降低了模型的复杂度(从 5 降到 4), 如图 4(b) 所示。但是, RVS-PLS 选择的变量个数最少(12 个)。

2.2.3 Gasoline 数据集

对于 gasoline 数据集, 3 种算法达到了几乎一样的预测性能[见图 5(b)]。但是, RVS-PLS 模型的复杂度(4 PCs)比 UVE-PLS 和 GA-PLS(2 PCs)要高。这说明, 尽管 RVS-PLS 选择了最少的变量个数, 但是误选了一部分变量, 导致了更复杂的模型。

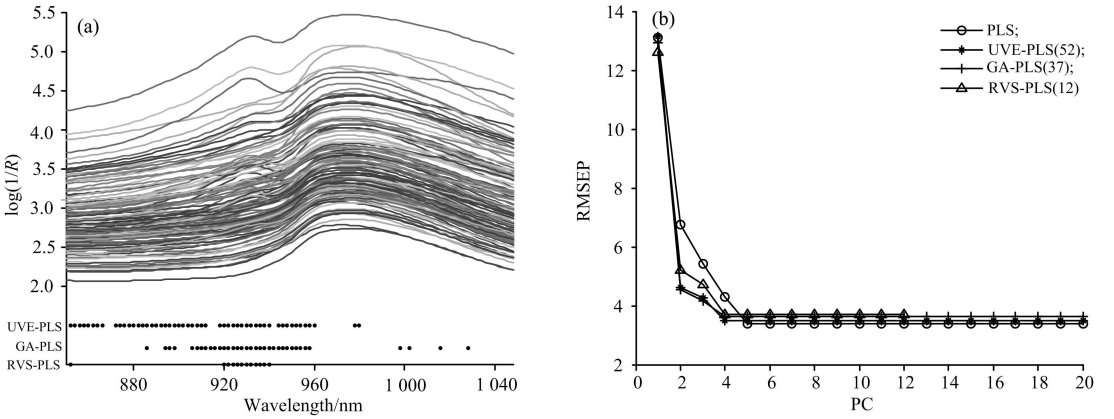

Fig 4　(a) Variables selected for meat dataset; (b) Comparison of RMSEP, the number in the parenthesis is the number of variables selected

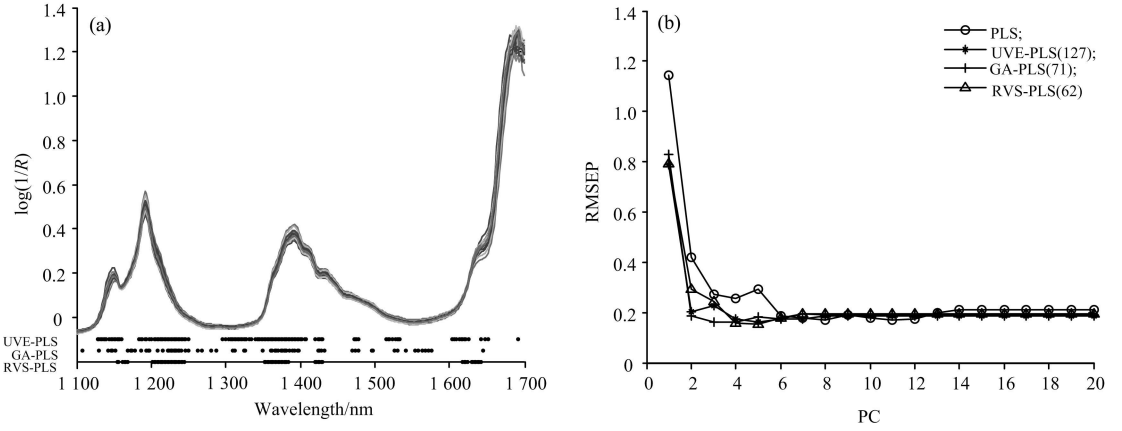

Fig 5　(a) Variables selected for gasoline dataset; (b) Comparison of RMSEP, the number in the parenthesis is the number of variables selected

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

3 结 论

本文分析了基于单个模型的波长选择方法，针对存在的两个问题，提出了一种新的波长选择方法，它融合了多个模型的回归系数，得到了更加合理的波长点变量重要性的度量，使得波长选择更加准确和稳定。最后，通过实验验证了方法的有效性，并与两种典型的方法进行了比较。实验结果表明，本文提出的方法能够取得一致的实验结果，甚至能得到更好的预测性能、更少的波长点个数或更加紧凑的模型。

参 考 文 献

[1] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍, 袁洪福, 徐广通, 等). Modern Analysis Techniques for Near Infrared Spectroscopy(现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2000.

[2] Martens H, Naes T. Multivariate Calibration, Chichester. UK: John Wiley & Sons, Inc., 1989.

[3] Geladi P, Kowalski B R. Analytica Chimica Acta, 1986, 185: 1.

[4] Pierna J A F, Abbas O, Baeten V. Analytica Chimica Acta, 2009, 642: 89.

[5] Leardi R. Journal of Chemometrics, 2000, 14: 643.

[6] Sutter J M, Kalivas J H. Microchemical Journal, 1993, 47: 60.

[7] Hageman J A, Streppel M, Wehrens R. Journal of Chemometrics, 2003, 17: 427.

[8] Frenich A G, Jouan-Rimbaud D, Massart D L. Analyst, 1995, 120: 2787.

[9] Teófilo R F, Martins J A P A, Ferreira M M C. Journal of Chemometrics, 2009, 23: 32.

[10] Centner V, Massart D, Noord O E D. Analytical Chemistry, 1996, 68: 3851.

[11] Xu H, Liu Z, Cai W. Chemometrics and Intelligent Laboratory Systems, 2009, 97: 189.

[12] Benoudjit N, Cools E, Meurens M. Chemometrics and Intelligent Laboratory Systems, 2004, 70: 47.

[13] CHEN Xiao-jing, WU Di, YU Jian-jia(陈孝敬, 吴 迪, 虞佳佳). Acta Optica Sinica(光学学报), 2008, 28: 2153.

[14] Xu Q, Liang Y. Chemometrics and Intelligent Laboratory Systems, 2001, 56: 1.

[15] Eldén L. Computational Statistics & Data Analysis, 2004, 46: 11.

[16] Martens H A, Dardenne P. Chemometrics and Intelligent Laboratory Systems, 1998, 44: 99.

[17] Snee R D. Technometrics, 1977, 19: 415.

[18] Kalivas J H. Chemometrics and Intelligent Laboratory Systems, 1997, 37: 255.

A New Wavelength Selection Algorithm Based on the Fusion of Multiple Models

HONG Ming-jian[1,2], WEN Zhi-yu[1]

1. Micro-Electromechanical System Research Center of Chongqing University, Chongqing 400030, China

2. School of Software Engineering, Chongqing University, Chongqing 400030, China

Abstract NIR spectroscopy makes a feature of a large number of wavelengths with a much smaller set of samples. However, some of the wavelengths contribute no information to the modeling. Even worse, they may contain the irrelevant information such as noise and background, which may result in a complex model and/or bad predictive ability of the model. So, it's important to do research in-depth to eliminate these wavelengths and improve the quality of the final model. The present paper firstly summarizes the variable selection methods based on a single PLS regression model and concludes that (1) the cross-validation can be used to select optimal model with good predictive ability, but the resulting model may be not suitable for selecting variables; (2) selecting variables based on a single regression model is inaccurate and instable because a single vector of regression coefficients may not measure the importance of the variables correctly and may vary with models of different complexity. On basis of this analysis, this paper proposed a new method for variable selection based on the fusion of multiple PLS models. This method fuses the multiple PLS regression coefficients to form a vector, then a threshold is determined to eliminate the variables whose corresponding element in the vector is lower than this threshold. Finally, this method is verified by 3 well-known NIR datasets and compared with the UVE-PLS and GA-PLS algorithms. The experiments show that this method may result in a model with less complexity and/or better predictive ability. Moreover, the proposed method is elegant and efficient and therefore can be put in practical use.

Keywords Fusion; Wavelength selection; Spectra analysis; NIR

(Received Oct. 9, 2009; accepted Jan. 16, 2010)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net