文章编号:1001-9014(2020)03-0318-06 DOI:10.11972/j.issn.1001-9014.2020.03.009

一种基于变量稳定性与集群分析相结合的近红外波长的选择方法

张 峰, 汤晓君*, 仝昂鑫, 王 斌, 王经纬
(西安交通大学 电力设备电气绝缘国家重点实验室, 陕西 西安 710049)

摘要:为了提高分析模型的效率与性能,提出了一种基于变量稳定性与集群分析相结合(VSPA)的波长选择方法。该算法将变量分为样本空间与变量空间,在样本空间里计算变量的稳定性,根据稳定性值,利用加权自举采样技术将变量划分为有用变量与无用变量;在变量空间中,统计每个变量出现的频率,利用指数衰减函数在无用变量中去掉变量频率较低的变量。将算法应用在近红外光谱玉米数据集中来预测玉米中淀粉的含量,其预测集均方根(RMSEP)与相关系数(R_p)分别为0.0409和0.9974,筛选后的特征变量仅为原始光谱数据的2.7%,说明提出的变量选择方法能够提高模型的运算效率与预测能力,是一种有效的变量选择方法。

关 键 词:波长选择; 加权自举采样; 近红外光谱; 偏最小二乘
中图分类号:Q657.33 **文献标识码**: A

A near infrared wavelength selection method based on the variable stability and population analysis

ZHANG Feng, TANG Xiao-Jun*, TONG Ang-Xin, WANG Bin, WANG Jing-Wei
(Xi'an Jiaotong University State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an 710049, China)

Abstract: In order to improve the efficiency and performance of the analysis model, a wavelength selection method based on variable stability and population analysis (VSPA) is proposed. Firstly, the variables are divided into sample space and variable space, and the stability of variables is calculated in the sample space. According to the stability value, the variables are divided into useful variables and useless variables by weighted bootstrap sampling technology. Then, in the variable space, the frequency of each variable is calculated, and the exponential decline function is used to remove the variables with lower frequency from the useless variables. Finally, the proposed algorithm is applied to corn NIR data set to predict the starch content. The predicted root root-mean-square (RMSEP) and predicted correlation coefficient (R_p) is 0.0409 and 0.9974, respectively. The variables after selection are only 2.7% of the original spectral data. It shows that the proposed variable selection method can improve the operational efficiency and prediction accuracy of the model, and is proved to be an effective variable selection method.

Key words: wavelength selection, weighted bootstrap sampling, near infrared spectral, partial least squares
PACS: 07.05.Kf

收稿日期:2019-12-19,**修回日期**:2020-05-09 **Received date**:2019-12-19,**Revised date**:2020-05-09
基金项目:国家重点研发计划(2016YFF0102805)
Foundation items:Supported by the National Great Program on scientific instrument and equipment research of China (2016YFF0102805)
作者简介(Biography): 张峰(1988-),男,西安交通大学博士生,主要研究领域为光谱分析、智能传感器. E-mail: 774149296@qq.com;
*通讯作者(Corresponding author): E-mail: xiaojun_tang@mail.xjtu.edu.cn

引言

红外光谱定性与定量分析技术由于其响应速度快、分析成分多、预测准确等特点,广泛应用于电力设备故障诊断、石油天然气勘探、煤矿灾害预警等领域[1-3]。通常,直接由光谱仪获得的谱线多达成千上万条,然而光谱仪主要由光学部件构成,这些元器件易受到周围环境的影响,导致扫描获得的谱线不可避免了包含干扰谱线,无用谱线。如果将获得的全部光谱数据用来建立分析模型,不仅会增加模型的运行时间,甚至还会降低模型的预测性能[4-7]。因此,在建立分析模型之前对光谱数据进行特征提取具有重要的意义。

偏最小二乘法(partial least square, PLS)由于其操作简单与较高的预测精度,广泛应用于线性模型中[8]。在PLS模型中,回归系数$\boldsymbol{\beta}$是一个重要指标参数,基于该回归系数,产生了几种波长变量筛选方法。这些方法中具有代表的有蒙特卡洛无信息变量消除法(Monte Carlo non-information variable elimination, MCUVE)[9]、竞争自适应重加权抽样法(competitive adaptive reweighted sampling, CARS)[10]、自举柔性收缩法(bootstrapping soft shrinkage, BOSS)[11]。其中MCUVE通过蒙特卡洛产生大量的样本空间,对每个样本空间建立PLS模型,获取PLS回归系数的平均值与标准差,计算每个变量的稳定性。当变量的稳定性小于设定的阈值时,认为该变量为无用变量,进行剔除处理,然而应用该方法进行变量选择时,筛选的变量交多,且阈值的大小对变量选择的结果影响很大。CARS算法在进行变量筛选时,根据回归系数的大小来对变量进行强制剔除,然而,回归系数的大小会随着样本空间的改变而变化,这样导致剔除的变量可能包含有用变量。BOSS算法是一种基于变量空间的谱线选择方法,通过加权自举采样方法集合蒙特卡洛算法产生大量的变量空间子模型,对这些子模型分别建立PLS模型,计算模型的回归系数,对这些回归系数取绝对值并归一化处理,来更新变量的权重,这样权重值较大的变量在下次迭代过程中会有更大的机会被选中。但是BOSS算法仅仅考虑了变量空间中回归系数这个指标特征,筛选的变量未必是最优的。

针对上述波长选择方法存在的问题,本文提出了一种基于变量稳定性与集群分析相结合的变量选择方法对红外光谱变量进行筛选,该方法将变量分为样本空间与变量空间,在样本空间中计算变量的稳定性,在变量空间中计算变量的频率,根据变量稳定性利用加权自举采样将变量分为有用变量与无用变量,利用指数衰减函数在无用变量中强制剔除在变量空间在频率较低的变量,实现对变量的筛选。为了评价提出方法的性能,将现有MCUVE、CARS、BOSS与提出的VSPA四种变量选择方法,分别应用于玉米的近红外光谱数据集中,并对四种方法筛选的变量分别建立PLS模型来预测玉米中淀粉的含量。结果表明,本文提出的VSPA算法选择的变量个数与BOSS算法相当,预测结果最好,是一种有效的变量筛选方法。

1　算法介绍

1.1　样本空间变量稳定性定义

假设样本的红外光谱矩阵为$\mathbf{X}_{n\times p}$,$\mathbf{X}$为n个样本扫描获得的p个谱线,通常n<<p, $\mathbf{y}$为分析样本的浓度信息矩阵,当样本包含的成分为1时,$\mathbf{y}$为浓度信息向量,$\mathbf{e}$为随机误差。建立PLS回归模型时光谱矩阵与浓度信息之间的关系可以表示为如下:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e} \quad , \quad (1)$$

式(1)中,$\boldsymbol{\beta}$为回归系数向量,$\boldsymbol{\beta}=[\beta_1, \beta_2, \cdots, \beta_p]^T$,这里只考虑样本成分为1。应用蒙特卡洛算法,从n个样本中随机选择n1个样本作为样本空间,该过程循环M次,这样获得了M个样本空间,对每个样本空间建立PLS回归模型,可以得到回归系数矩阵$\boldsymbol{\beta}_{p\times M}$,计算每个变量回归系数的平均值与标准差,则第i个变量的稳定性可以由下式计算:

$$S_i = \left| \frac{\mathrm{mean}(\beta_{i1},\beta_{i2}\cdots\beta_{iM})}{\mathrm{std}(\beta_{i1},\beta_{i2}\cdots\beta_{iM})} \right| \quad . \quad (2)$$

1.2　VSPA特征波长选择方法

VSPA算法以变量回归系数的稳定性进行粗选,根据变量的频率指标进行精选。当变量的稳定性值越小时,有很大概率被认为是无用变量,此时如果变量获得的频率越小,则该变量会被强制剔除,当循环到达设定的次数时,迭代停止。VSPA算法具体的实现步骤如下所示:

Step1:循环开始,变量的初始长度等于全部光谱变量,记为p。从n个样本中随机选择n1个样本作为样本空间,计算每个变量的稳定性S_i,根据S_i的值,利用加权自举采样技术,将p个变量划分为有用变量,与无用变量。值得一提的是,有用变量的个数约为变量长度的0.632倍[12];

Step2：保持样本的数量n不变，应用蒙特卡洛算法随机从p个变量中选择p1个变量，该步骤循环W次，获得W个变量空间，分别为每个变量空间建立PLS模型，每个模型会获得相应的均方根误差值（RMSE），从W个模型里面选择RMSE值较小的αW个模型，统计每个变量出现的频率f_i；

Step3：利用指数衰减函数，确定每次迭代后剩余的变量，指数衰减函数可以表示如下：

$$r_i = ae^{-ki} \quad , \quad (3)$$

式(3)中r_i为当前迭代后剩余的变量数，$a = (p/2)^{(1/(N-1))}$，$k = \ln(p/2)/(N-1)$，i为第i次循环，N为循环的运行次数。当剩余变量的个数大于step1中有用变量的个数时，从无用变量中删除step2中频率低的变量，当剩余变量小于有用变量个数时，删除全部的无用变量，并且从有用变量中删除稳定性值低的变量；

Step4：对每次循环中剩余的变量建立PLS模型，记录每次循环中获取的RMSE值，同时更新p的值，使得p=r，循环次数i=i+1；

Step5：若i≤N，执行step1，否则，执行Step6；

Step6：选择最小的RMSE值对应的变量组合作为最终选择的变量。

2 实验部分

2.1 数据来源

所用的近红外光谱数据集为玉米数据集，来源于benchmark红外光谱数据库，可以在网站http://www.eigenvector.com/data/Corn/index.html上免费下载。该数据集是由同一批80个玉米样本分别在编号为M5、MP5、MP6三台光谱仪上采样获得的。文中选择编号为M5光谱仪获得的光谱建立分析模型，波长扫描范围为1100~2498nm，波长分辨率为2nm，因此，每个样本可以获得700个谱线变量，每个样本中包含四个指标含量，分别为水分、油、蛋白质、淀粉。本文以淀粉含量作为模型评价指标。在变量选择之前，利用SG平滑算法[13]对玉米数据集进行了降噪处理，并应用SPXY方法将80个玉米样本分为60个训练样本与20个验证样本。

2.2 VSPA参数确定

VSPA算法中需要确定的参数有：(1)VSPA循环的次数N；(2)从样本中随机选择的样本个数n1；(3)利用蒙特卡洛生成的样本空间的个数M；(4)计算变量频率时蒙特卡洛生成次数W；(5)选择的最优模型占全部变量空间模型的比例系数α，可以分两步来进行参数确定。

首先，确定样本空间的参数，固定变量空间参数，α设置为0.1，W设置为1000。固定样本空间中循环次数，N设置为50，因为验证集的样本数为60，因此n1的值可以设置从20到55，间隔设置为5，为了求取样本空间的稳定性，M取值从40到320之间取值即可，间隔为20，N的大小为50，共经历120次VSPA运算后，获得120个RMSE值，选择RMSE最小的值作为最终的n1与M。图1为RMSE随着n1与M变化的趋势图。从图1中可以看出，当M的值为180，n1为45时，得到的RMSE值最小，为0.055；将获得的n1与M作为常数量，进行循环次数N的确定，因为玉米光谱的变量为700个，可以设置VSPA循环次数N的取值范围为20到200，间隔为设置为20，N每次取值时，VSPA算法重复运行30次，计算30次RMSE的平均值，记作mRMSE，于是，共经历300次VSPA运算后，可以获得10组mRMSE值，如图2所示。从图2中可知，当N为100时，mRMSE的值最低。

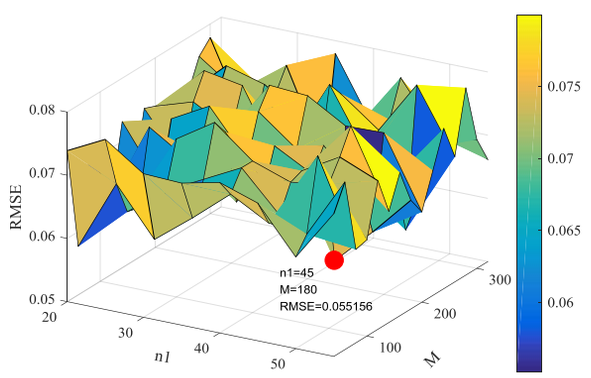


图1 VSPA算法中参数n1与M的优化选择

Fig 1 The optimization and selection of parameters n1 and M of the VSPA algorithm

然后固定样本空间中确定的参数，进行变量空间子模型个数W与最优模型比例α参数的选取。为了保证每个变量都有机会分配到对应的子模型，W的值设置较大，设置W的初始值为500，在500到5000范围内以间隔为500取值。α的值应该尽量小，通常要小于0.5，保证了所选择的变量空间预测结果更好，α的设置范围为0.05到0.5，间隔为0.05，经过100次VSPA计算后，获得100个RMSE值，图3为RMSE的值随W与α之间的关系图。由图3可知，当W等于2500，α取值0.05时，获得的RMSEV值最小。因此，最终获得的参数值为：算法

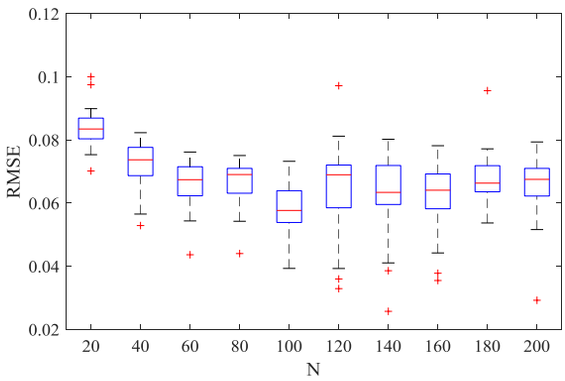


图2　VSPA算法中参数N的优化选择

Fig 2　The optimization and selection of parameter N of the VSPA algorithm

循环的次数N为100、随机选择的样本个数n1为45，样本空间采样次数M为100，蒙特卡洛生成变量空间的采样次数为2500，最优模型占全部变量空间模型的比例系数α为0. 05。

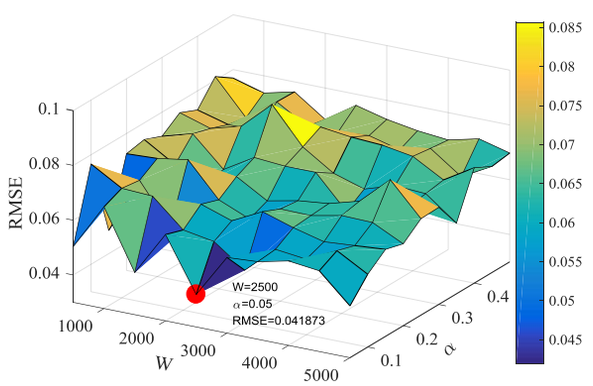
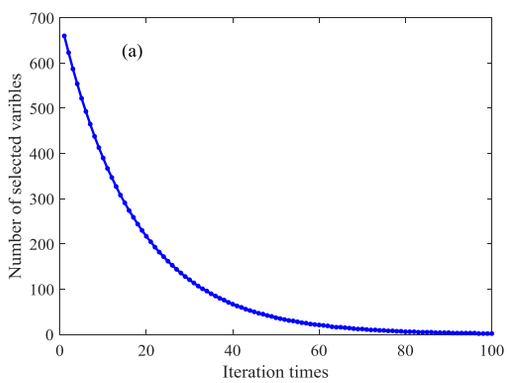


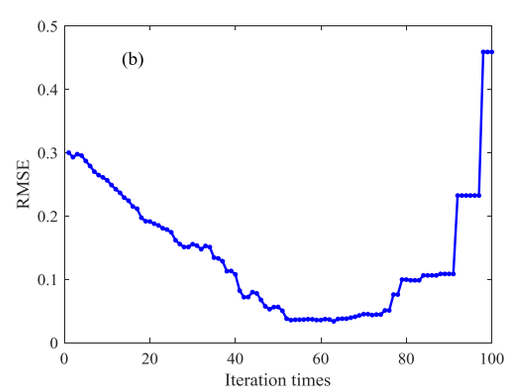
图3　VSPA算法中参数W与α的优化选择

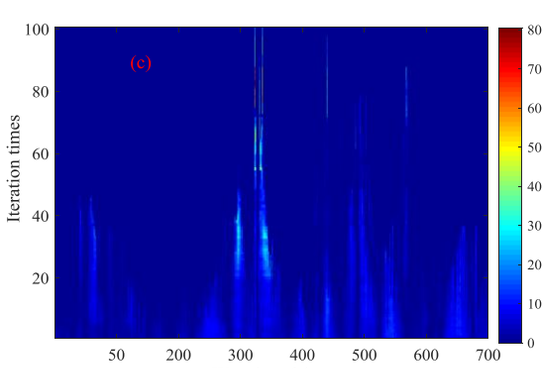
Fig 3　The optimization and selection of parameter W and α of the VSPA algorithm

2. 3　变量选择过程

根据2. 2小节所确定的参数，采用VSPA算法对玉米数据集的700个谱线变量进行选择。700个变量选择过程如图4所示。图4(a)为剩余变量的变化趋势，可以看出，剩余变量在迭代的初期衰减很快，而在迭代的末期，衰减很慢，体现了指数递减函数的粗选与精选的特点，经过100次迭代后，剩余的变量为2。图4(b)为变量选择过程中RMSE值随着迭代次数增加的变化趋势，从图中可以看出，RMSE的值在第63次迭代后，获得的值最低，所对应的变量个数为19。表明在前面迭代过程中剔除的变量为无用变量或者干扰变量，在第63次迭代后，RMSE的值呈现上升的趋势，说明剔除的变量含有有用的信息变量，选择第63次迭代后剩余的变量作为最终筛选的变量子集。图4(c)样本空间中变量稳定性随着迭代次数的变化图，可以看出，当变量的稳定性值较大时，变量会保留到下一代迭代中，随着迭代次数的增加，1700~1900nm处的变量始终保持着较大的稳定性，表明该部分变量为有用的信息变量。与图4(d)变量空间中变量频率随着迭代次数的变化图，可以得出，随着迭代的进行，一些变量的频率呈现逐渐上升的趋势，这是因为随着变量的减少，有用变量被选中的概率越来越大。值得注意的是，迭代次数大于69时，剔除的变量为样本空间中全部的无用变量与有用变量中稳定性较低的变量。







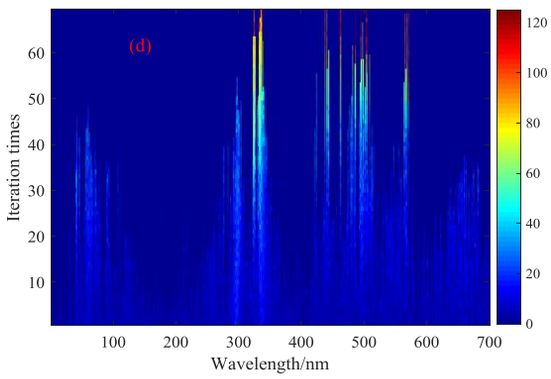


图4 VSPA算法变量选择过程（a）选择的变量数与迭代次数的关系；(b) RMSE与迭代次数的关系；(c) 样本空间中变量稳定性随着迭代次数的变化图；(d) 变量空间中变量频率随着迭代次数的变化图

Fig 4 Spectra variables selection procedure with VSPA algorithm: (a) Relationship between selected variables and iteration times; (b) Relationship between RMSE and iteration times; (c) The change of stability with the number of iterations in sample space; (d) The change of variable frequency with the number of iterations in variable space

2.4 方法对比

为了评价提出波长选择方法的性能，应用MCUVE、CARS、BOSS与本文提出的VSPA四种方法进行变量选择。图5为四种方法在玉米数据集中的变量分布图。从图中可以看出，MCUVE选择的变量最多，并且分布比较散；CARS选择的变量比MCUVE少，BOSS与VSPA选择的变量相当。四种方法均选择了1700~1800nm处的变量，这部分区域对应了C-H官能团的吸收区域。除此之外，MCUVE与CARS还选择了2400nm附近的变量，然而这些变量由于模型的预测效果差被认为是无用的变量。

利用上述四种方法提取的变量分别建立PLS模型来预测玉米数据集中淀粉的含量，四种方法的预测结果如表1所示。从表中可以看出，无论是在校准集还是在测试数据集上，上述四种变量选择方法建立的PLS模型均比未进行变量选择建立的PLS模型预测效果好，表明波长变量选择方法能够提高模型的预测精度。在上述四种方法中，VSPA获得的R_c与R_p值分别为0.9957、0.9974，预测均方根RMSEP为0.0409，而MCUVE、CARS、BOSS三种方法R_c与R_p值分别为0.9353、0.9677、0.9842与0.9724、0.9598、0.9870，RMSEP分别为0.1328、0.1603、0.0911，相比而言，VSPA建立的模型性能评价指标

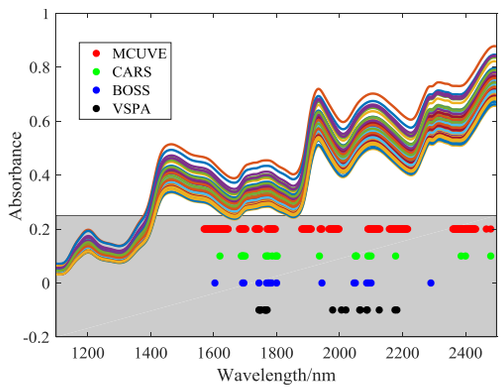


图5 四种方法选择的变量分布图

Fig 5 The variables selected by MCUVE, CARS, BOSS and VSPA

最好，表明了利用VSPA进行波长变量选择后，建立的模型效率更高、预测性能更好。VSPA算法选择的变量数目与BOSS方法相近，但是VSPA的预测结果比BOSS好，这是因为BOSS算法是在变量空间中根据变量的回归系数大小利用加权自举采样技术实现对波长的选择，忽略了变量在样本空间的稳定性，而文中提出的VSPA算法不但考虑了变量在样本空间的稳定性，还考虑了在变量空间中频率这个重要评价指标，能够更好的对变量进行筛选；MCUVE选择的变量最多，预测结果最差，这是因为MCUVE仅仅考虑了变量在样本空间的稳定性，忽略了在变量空间中频率评价指标，当变量之间的共线性较高时，易导致该变量的回归系数值降低，进一步导致变量稳定性降低；CARS选择的变量居中，预测结果比VSPA、BOSS差，主要原因是CARS根据变量空间中回归系数这个指标来进行波长筛选的，没有考虑变量在样本空间的稳定性。

表1 五种不同方法的淀粉预测结果

Table 1 Result of five different methods on the starch dataset

模型	变量数	校正集		预测集	
		R_c	RMSECV	R_p	RMSEP
PLS	700	0.8664	0.2999	0.9335	0.2062
MCUVE-PLS	198	0.9353	0.2018	0.9724	0.1328
CARS-PLS	25	0.9677	0.1475	0.9598	0.1603
BOSS-PLS	20	0.9842	0.1032	0.9870	0.0911
VSPA-PLS	19	0.9957	0.0536	0.9974	0.0409

为了更加直观的评价VSPA-PLS模型的预测性能，绘制了玉米数据集中淀粉含量的真实值与预测值之间的散点图，从图6中可以看出，真实值与预测

值的散点集中在1:1线上,表明文中提出的变量选择方法建立的模型预测性能好,其校正集的相关系数 R_c 与均方根误差 RMSECV 分别为 0.9957 和 0.0536,测试集的相关系数 R_p 和均方根误差 RMSEP 分别为 0.9974 与 0.0409。

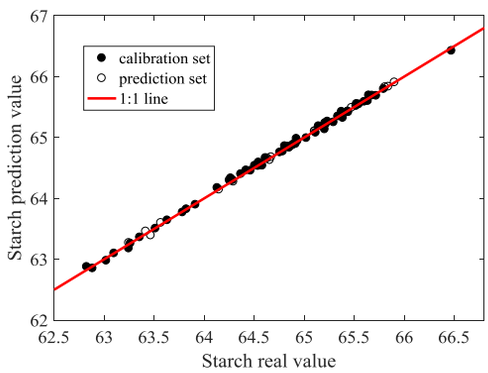


图6　玉米中淀粉含量的真实值与预测值的散点图

Fig 6　Scatter diagram of real and predicted value of starch content

3　结论

分析了近年来基于偏最小二乘模型中回归系数的波长选择方法,提出了一种基于变量稳定性与集群分析的变量选择方法,将该方法与 MCUVE、CARS、BOSS 四种变量选择方法分别应用于玉米近红外光谱数据集中,对四种方法筛选的变量分别建立 PLS 模型来预测玉米中淀粉含量。结果表明,本文提出的波长变量选择方法预测性能最好,在波长选择数量上与 BOSS 方法相当,是一种有效的波长变量选择方法,具有实际的应用价值。

References

[1] Shen X C, Xu L, Ye S B, *et al*. Automatic baseline correction method for the open-path Fourier transform infrared spectra by using simple iterative averaging[J]. *Optics Express*, 2018, **26**(10): A609-A614

[2] Tang X J, Liang Y T, Dong H Z, *et al*. Analysis of index gases of coal spontaneous combustion using Fourier Transform Infrared Spectrometer [J]. *Journal of Spectroscopy*, 2014, **2014**: 1-8

[3] Tang X J, Li Y J, Zhu L J, *et al*. On-line multi-component alkane mixture quantitative analysis using Fourier transform infrared spectrometer[J]. *Chemometrics and Intelligent Laboratory Systems*, 2015, **146**: 371-377

[4] Chen J M, Yang C H, Zhu H Q, *et al*. A novel variable selection method based on stability and variable permutation for multivariate calibration [J]. *Chemometrics and Intelligent Laboratory Systems*. 2018, **182**: 188-201

[5] Yun Y H, Wang W T, Deng B C, *et al*. Using variable combination population analysis for variable selection in multivariate calibration. *Analytica Chimica Acta*, 2015, **140**: 14-23

[6] SHI Ji-Yong, ZOU Xiao-Bo, ZHAO Jie-Wen, Mao Han-Ping. Selection of wavelength for strawberry NIR spectroscopy based on BiPLS combined with SAA. [J]. *J. Infrared Millim. Waves*(石吉勇,邹小波,赵杰文,等.BiPLS结合模拟退火算法的近红外光谱特征波长选择研究.红外与毫米学报),2011, **30**(5): 458-462

[7] Deng B C, Yun Y H, Ma P, *et al*. A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals [J]. *Analyst*, 2015,**140**: 1876-1885

[8] LIU Guo-hai, XIA Rong-sheng, JIANG Hui, *et al*. A Wavelength Selection Approach of Near Infrared Spectra Based on SCARS Strategy and Its Application [J]. *Spectroscopy and Spectral Analysis* (刘国海,夏荣盛,江辉,等.一种基于SCARS策略的近红外特征波长选择方法及其应用.光谱学与光谱分析). 2014, **34**(8): 2094-2097

[9] Han Q J, Wu H L, Cai C B, *et al*. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Analytica Chimica Acta*, 2008, **612**: 121-125

[10] Li H D, Liang Y Z, Xu Q S, *et al*. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. *Analytica Chimica Acta*, 2009, **648**(1): 77-84

[11] Deng B C, Yun Y H, Cao D S, *et al*. A bootstrapping soft shrinkage approach for variable selection in chemical modeling[J]. *Analytica Chimica Acta*, 2016, **908**: 63-74

[12] Song X Z, Huang Y, Yan H, *et al*. A novel algorithm for spectral interval combination optimization [J]. *Analytica Chimica Acta*, 2016, **948**: 19-29

[13] ZHAO An-xin, TANG Xiao-jun, ZHANG Zhong-hua, *et al*. Optimizing Savitzky-Golay Parameters and Its Smoothing Pretreatment for FTIR Gas Spectra [J]. *Spectroscopy and Spectral Analysis*,(赵安新, 汤晓君, 张钟华,等.优化Savitzky-Golay滤波器的参数及其在傅里叶变换红外气体光谱数据平滑预处理中的应用.光谱学与光谱分析). 2016, **36**(05):1340-1344