基于 L1-L2 联合范数约束的中药近红外光谱波长选择

任 真 李四海*
(甘肃中医药大学信息工程学院 甘肃 兰州 730000)

摘 要 针对近红外光谱分析中存在的高维数据降维、多重共线性及模型稀疏性问题,提出一种基于正则偏最小二乘 RPLS(Regularization Partial Least Squares)的近红外光谱波长变量选择方法。该方法在偏最小二乘回归模型中同时引入 L1 和 L2 范数罚正则项,使模型产生稀疏性,通过交替迭代算法求解主成分载荷系数的稀疏解,实现光谱数据降维和重要波长变量的自动选择。对当归近红外光谱进行正则偏最小二乘波长选择实验。结果表明,与 CARS(Competitive Adaptive Reweighted Sampling)随机蛙跳等变量选择方法相比,正则偏最小二乘方法在选择变量数及模型的预测能力方面均具有一定的优势。

关键词 近红外光谱 偏最小二乘回归 正则化 变量选择

中图分类号 TP181 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.12.019

A WAVELENGTH SELECTION METHOD FOR NEAR INFRARED SPECTROSCOPY OF CHINESE MEDICINE BASED ON L1-L2 NORM SIMULTANEOUS CONSTRAINT

Ren Zhen Li Sihai*
(*College of Information Engineering*, *Gansu University of Chinese Medicine*, *Lanzhou 730000*, *Gansu*, *China*)

Abstract In the view of the problems of dimensionality reduction, multiple collinearity and model sparsity existed in the near infrared spectroscopy analysis, we proposed a wavelength variable selection method for near infrared spectroscopy based on regularization partial least squares (RPLS). The L1 and L2 norm penalty regularization terms were introduced in the partial least squares regression (PLSR) model to make the model sparse. We utilized the alternative and iterative algorithm to obtain the sparse solution of principal component load coefficients so as to achieve dimensionality reduction of the spectral data and the automatic selection of the important wavelength variables. Regular partial least squares wavelength selection experiments were conducted for near infrared spectra of Angelica sinensis. The experimental results show that compared with variable selection methods such as CARS and random-frog, RPLS method has certain advantages in terms of selected variable number and prediction accuracy of model.

Keywords Near infrared spectroscopy (NIRS) Partial least squares regression (PLSR) Regularization Variable selection

0 引 言

中药具有物质成分多、作用机理复杂等特点,保证中药产品质量的可靠稳定是实现中药现代化必须要解决的关键问题。近红外光谱分析技术是 20 世纪 90 年代以后迅速发展起来的一种新型在线检测技术,具有简便、低成本、不破坏样品等优点,已在农业、食品、石油化工、药物质量控制等领域得到广泛应用。在中药质量控制方面,近红外光谱能够快速、准确地鉴别中药材的真伪、种类和产地,并且能够快速测定中药材中有效成分的含量以及中药辅料的品质[1-2]。

近红外光谱信号具有维度高、变量多重共线性严重等特点[3]。文献[4]用随机蛙跳算法(Random-frog)对生物柴油近红外光谱进行波长选择,分别建立了生物柴油含水量的偏最小二乘和支持向量机预测模型,

收稿日期:2018-04-16。甘肃省自然科学基金项目(1508RJZA008);甘肃省科技计划项目(1506RJZA046)。**任真**,副教授,主研领域:数据挖掘。**李四海**,副教授。

取得了较好的效果。文献[5]建立了金银花醇沉过程中绿原酸含量的偏最小二乘回归模型,通过竞争自适应抽样 CARS 变量选择方法,提高了模型的预测精度。以上研究表明,基于波长变量选择的近红外光谱建模方法能够有效提高模型的预测能力。

现有的波长变量选择方法大多是封装式的:通过遍历搜索,找到变量空间的最优特征子集,使模型的均方根误差最小。由于特征子集的搜索采用前向、后向或蒙特卡洛随机搜索策略,对每一个特征子集的评价都需要重新训练模型,计算开销较大。嵌入式特征选择是机器学习中一类重要的特征选择方法,其通过最小化目标函数,使得特征选择和模型训练融为一体,同步完成,由于 L1 范数的稀疏特性,嵌入式特征选择能够自动实现变量选择,降低模型的过拟合风险,提高模型的可解释性。

本文通过在偏最小二乘回归模型中同时引入 L1 和 L2 正则项,建立一种嵌入式的光谱特征变量选择方法。将光谱变量选择和预测模型的建立融合在一起,解决光谱数据存在的多重共线性问题,提高了偏最小二乘回归模型的可解释性,实现了对当归中藁本内酯含量的快速、准确检测。

1 正则偏最小二乘回归算法设计

1.1 L1 正则与稀疏性

正则化具有产生稀疏模型的能力[6-9],Tibshirani[10]提出的线性回归 Lasso 模型通过将岭回归中的 L2 正则项替换为 L1 正则项,使模型具有变量选择和数据降维能力。

假设自变量矩阵 $\boldsymbol{X}\in R^{n\times p}$,因变量 $\boldsymbol{Y}\in R^{n\times 1}$,线性回归的 Lasso 模型为:

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\frac{1}{2}\|Y-X\beta\|_2^2+\lambda\|\beta\|_1 \tag{1}$$

式中:$\|\cdot\|_2$表示 L2 范数,β 为回归系数向量,$\|\beta\|_1$ 表示所有回归系数的绝对值之和,为 L1 范数罚,λ 为惩罚系数。Lasso 模型的解可通过坐标下降算法求得[11]:

对于 $i=1,2,\cdots,p$

$$\hat{\beta}_i=sign(\hat{\beta}_i^{\mathrm{OLS}})(|\hat{\beta}_i^{\mathrm{OLS}}|-\lambda)_+=\begin{cases}\hat{\beta}_i^{\mathrm{OLS}}-\lambda & \hat{\beta}_i^{\mathrm{OLS}}>\lambda\\ 0 & -\lambda\leqslant\hat{\beta}_i^{\mathrm{OLS}}\leqslant\lambda\\ \hat{\beta}_i^{\mathrm{OLS}}+\lambda & \hat{\beta}_i^{\mathrm{OLS}}<-\lambda\end{cases}$$

式中:$\hat{\beta}_i^{\mathrm{OLS}}$ 为普通最小二乘 OLS(Ordinary Least Square)的解。可以看出,Lasso 通过对回归系数施以相同程度的惩罚,将$|\hat{\beta}_p^{\mathrm{OLS}}|<\lambda$ 的回归系数置为0,使回归系数为 0 的变量不参与模型的拟合,从而实现了变量选择和模型稀疏化效果[12-13]。

1.2 偏最小二乘回归

偏最小二乘回归(PLSR)是主成分分析(PCA)和典型相关分析(CCA)的有效结合,其对 CCA 方法进行了进一步拓展。PLSR 能有效解决高维变量之间的多重共线性问题,在近红外光谱的定量分析中得到了广泛应用[14]。

假设 X 和 Y 分别为光谱数据矩阵和待测含量矩阵,$X\in R^{n\times p}$,$Y\in R^{n\times q}$,X 和 Y 的每一列均为零均值且标准差为 1。

PLSR 首先从 X 和 Y 中提取第一对主成分 u_1和 t_1,满足:$Var(u_1)\rightarrow max$,$Var(t_1)\rightarrow max$,$Corr(u_1,t_1)\rightarrow max$。即 u_1和 t_1最大化且二者的相关性最大化。然后计算 X 和 Y 残差并根据残差继续提取下一个主成分。由于各投影方向之间相互正交,抽取的特征位于不同的投影方向,因此偏最小二乘能够有效去除变量之间的多元共线性。PLSR 的最优主成分个数一般通过交叉验证方法确定。

1.3 正则偏最小二乘回归

近红外光谱数据具有高维度、小样本的特点。PLSR 的投影向量是所有原始波长变量的线性组合,一方面,将部分噪声变量也纳入投影向量参与模型拟合会使有效变量的回归系数产生衰减,导致预测精度下降;另一方面,当 $p>>n$,即 p 远大于 n 时,投影向量包含所有的原始波长变量导致模型可解释性不强。针对上述问题,在偏最小二乘回归模型中引入正则项,建立正则偏最小二乘回归 RPLS 算法,RPLS 可形式化为如下的最优化问题:

$$\begin{aligned}\min_{\alpha,c}\ &-\eta\alpha^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\alpha+(1-\eta)(c-\alpha)^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}(c-\alpha)+\\ &\lambda_1\|c\|_1+\lambda_2\|c\|_2^2\\ &\text{s.t. }\alpha^{\mathrm{T}}\alpha=1\end{aligned} \tag{2}$$

式中:η、λ_1和 λ_2均为常量,用于初始化算法,c 为主成分载荷系数向量 α 的副本且与 α 取值相近。式(2)可通过交替迭代算法求解[15]:

1) 固定 c,求解 α。对固定的 c,式(2)变为:

$$\begin{aligned}\min_{\alpha}\ &-\eta\alpha^{\mathrm{T}}M\alpha+(1-\eta)(c-\alpha)^{\mathrm{T}}M(c-\alpha)\\ &\text{s.t. }\alpha^{\mathrm{T}}\alpha=1\end{aligned} \tag{3}$$

令 $\boldsymbol{Z}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,$\eta'=(1-\eta)/(1-2\eta)$,则式(3)可重写为:

$$\min_{\alpha}(\boldsymbol{Z}^{\mathrm{T}}\alpha-\eta'\boldsymbol{Z}^{\mathrm{T}}c)^{\mathrm{T}}(\boldsymbol{Z}^{\mathrm{T}}\alpha-\eta'\boldsymbol{Z}^{\mathrm{T}}c) \tag{4}$$

s. t. $\alpha^T\alpha=1$

式中的 α 可通过拉格朗日乘子法求解。

2）固定 α,求解 c。对固定的 α,式(2)变为:

$$\min_c(\boldsymbol{Z}^{\mathrm{T}}c-\boldsymbol{Z}^{\mathrm{T}}\alpha)^{\mathrm{T}}(\boldsymbol{Z}^{\mathrm{T}}c-\boldsymbol{Z}^{\mathrm{T}}\alpha)+\lambda_1\|c\|_1+\lambda_2\|c\|_2^2 \quad (5)$$

问题转化为因变量为 $\boldsymbol{Z}^{\mathrm{T}}\alpha$ 的弹性网问题[16],c 可以通过 LARS 算法求解[17]。

令 $\hat{\beta}^{\mathrm{PLS}}$ 为 RPLS 算法回归系数,X_A 为活动变量集,K 为主成分个数。RPLS 算法如下:

Step1　初始化 $\hat{\beta}^{\mathrm{PLS}}=0$,$A=\{\}$,$k=1$,$Y_1=Y$

$\hat{c}=(Z_1-\lambda_1/2)_+sign(Z_1)$

式中:$\boldsymbol{Z}_1=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}/\|X^{\mathrm{T}}\boldsymbol{Y}\|$ 为第一投影方向单位向量。

Step2　当 $k\leqslant K$ 时

(1) 求解式(4),得到 $\hat{\alpha}$;求解式(5),更新 $\hat{c}$

(2) 根据 $\{i:\hat{\alpha}_i\neq0\}\cup\{i:\hat{\beta}_i^{\mathrm{PLS}}\neq0\}$,得到非零系数组成的活动变量集 X_A

(3) Y_1 对 X_A 做 PLS 回归,更新 $\hat{\beta}_i^{\mathrm{PLS}}$

(4) 计算残差 $Y_1=Y-X\hat{\beta}^{\mathrm{PLS}}$。$k=k+1$,重复执行以上步骤直至得到最优的 k 个主成分。

1.4　算法参数选择

对于单因变量回归问题,一般取 $\eta=1/2$,$\lambda_2\to\infty$。因此,算法的关键参数有两个:L1 正则项系数 λ_1 和最优主成分个数 k。

λ_1 用于控制选择的波长变量个数,其值越大,选择的波长变量数越少,模型可解释性越好,但参与建模的变量数过少会导致模型预测能力下降。因此,λ_1 的选取要权衡稀疏度和模型预测能力,通过实验选择最优的 λ_1 值。

最优主成分的个数 k 通过留一交叉验证法计算得到的 Q_k^2 值来确定[18]。

$$Q_k^2=1-PRESS_k/SS_{(k-1)} \quad (6)$$

式中:$PRESS_k$ 为使用前 k 个主成分对预留样本预测误差的平方和,$SS_{(k-1)}$ 为使用前 $k-1$ 个主成分对所有样本拟合误差的平方和。当 $Q_k^2\geqslant0.0975$ 时,认为第 k 个主成分作用显著。

2　模型的建立

2.1　实验数据

采集甘肃不同产地的当归样本 76 个,使用美国 Thermo 公司的 Nicolet-6700 型近红外光谱仪扫描得到所有样本的近红外光谱,光谱范围为 4 000 cm^{-1} ~ 10 000 cm^{-1},全谱共包括 1 557 个波数变量。76 个当归样本的近红外光谱如图 1 所示。

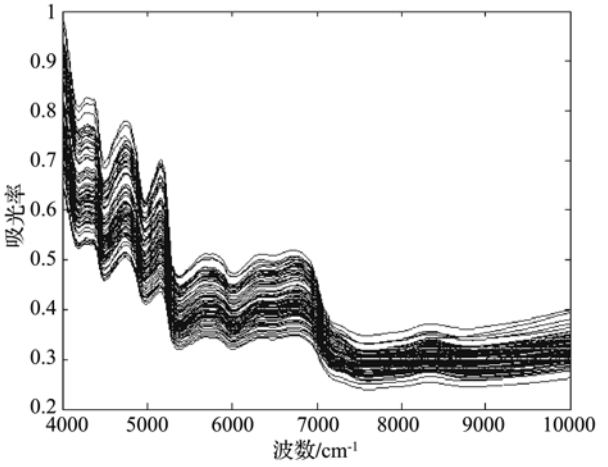

图 1　当归样本的近红外光谱

为消除基线漂移并提高光谱信号的信噪比,对原始光谱进行一阶导数结合正交信号校正预处理。当归中藁本内酯的含量采用高效液相色谱法(HPLC)测定[19]。

2.2　波长变量选择

将样本划分为训练集和测试集,训练集样本 56 个,测试集样本 20 个。使用训练集建立正则偏最小二乘回归模型对当归近红外光谱进行波长变量选择,因变量为当归中的藁本内酯含量。训练集和测试集中藁本内酯含量的分布情况见表 1。

表 1　训练集和测试集中藁本内酯含量分布

样本集	样本编号	最小值/(mg·g^{-1})	最大值/(mg·g^{-1})	平均值/(mg·g^{-1})	RSD
训练集	1 ~ 56	2.690 1	12.385 9	5.886 5	0.393 9
测试集	57 ~ 76	2.592 1	11.788 2	6.821 6	0.398 2

实验中用 eta 参数控制变量选择个数和模型稀疏度,对 eta 和 k 的最优组合采用网格寻优法确定。设定主成分 k 的搜索范围为[2,8],步长为 1,eta 的搜索范围为[0.7,0.9],步长为 0.1。最优波长变量子集通过权衡训练集上的预测均方根误差和波长变量个数来最终确定。

3　实验结果与分析

3.1　波长变量选择结果

根据主成分数 k 和参数 eta 的搜索范围,共得到 21 个波长变量子集,分别建立以上变量子集在 56 个训练样本上的偏最小二乘回归模型。不同变量子集在测试集上的预测结果对比如图 2 所示。

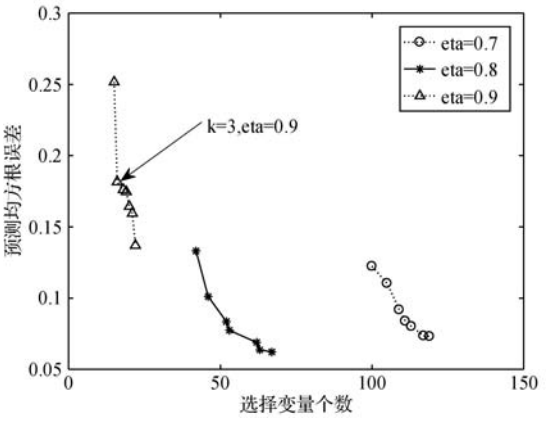

图 2 k 和 *eta* 对选择波长变量个数的影响

从图 2 可知,主成分个数 k 和 *eta* 对选择的波长变量个数都会产生影响,变量个数随着 k 的增加而增加,其中 *eta* 参数对变量个数的影响较大,*eta* 值越大,选择的波长变量个数越少。当 *eta* 为 0.7 时,选择变量个数最少为 100 个,最多为 119 个,模型的拟合效果很好,但参与建模的变量过多,模型的可解释性不强。当 *eta* 为 0.8 时,选择变量个数最少为 42 个,最多为 67 个,此时选择的变量仍然过多。当 *eta* 为 0.9 时,选择变量个数最少为 15 个,最多为 22 个。综合考虑选择变量个数和模型的预测能力,本文选择 $eta = 0.9$,$k = 3$ 为最优参数,此时共选择 16 个波长变量。分别为 5 164.4 cm^{-1}、5 218.4 cm^{-1}、5 222.3 cm^{-1}、5 226.1 cm^{-1}、5 230.0 cm^{-1}、5 233.9 cm^{-1}、5 237.7 cm^{-1}、5 241.6 cm^{-1}、5 245.4 cm^{-1}、5 249.3 cm^{-1}、5 253.1 cm^{-1}、5 257.0 cm^{-1}、5 260.9 cm^{-1}、5 704.4 cm^{-1}、5 708.3 cm^{-1}、5 712.1 cm^{-1}。根据近红外光谱的特征峰理论,所选择的波数大多位于 $C=O$ 基团的伸缩振动吸收峰附近,与当归中的内酯类化合物有关,这说明所选择的波长变量具有较好的化学意义。

3.2 不同预测方法结果对比

使用 RPLS 选择的 16 个波长变量,在 56 个训练样本上建立偏最小二乘回归模型,取前 3 个主成分。对 20 个测试样本中的藁本内酯含量的预测结果见图 3。

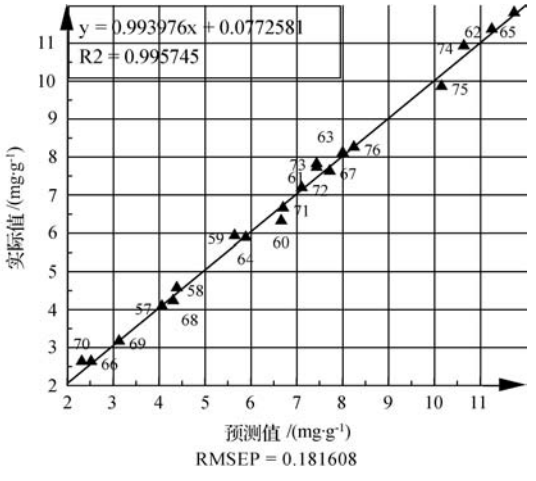

图 3 藁本内酯预测值和实际值对比

从 3 可以看出,预测值和真实值之间非常接近,预测均方根误差 RMSEP(root mean square error of prediction)为 0.181 6,决定系数为 0.995 7。RMSEP 的计算公式如下:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{7}$$

式中:n 为测试样本个数,y_i 为第 i 个测试样本中藁本内酯含量的实际值,$\hat{y}_i$ 为第 i 个测试样本中藁本内酯含量的预测值。由于对 20 个测试样本的预测均方根误差值较小,决定系数接近于 1,说明基于正则偏最小二乘波长变量选择的预测模型能够实现对当归中藁本内酯含量的快速、准确检测。

为进一步说明本文波长变量选择方法的有效性,将 RPLS 与目前近红外光谱分析中常用的 CARS、Random-frog 等波长变量选择方法进行了对比分析。CARS 算法迭代次数为 1 200 次,5 折交叉验证,从 1 200个子模型中选择交叉验证均方根误差最小的模型。Random-frog 算法变量迭代次数为 1 000 次,实验中迭代达到 200 次时选择的变量结果基本稳定。根据三种方法所选择的变量,分别建立 PLSR 模型,表 2 对比了不同的波长变量选择方法的预测性能。

表 2 不同变量选择方法的预测结果对比

方法	变量数	主成分个数	RMSEP	决定系数
全谱	1 557	4	0.230 1	0.994 1
CARS	22	5	0.236 5	0.990 1
Random-frog	30	6	0.241 7	0.985 3
RPLS	16	3	0.181 6	0.995 7

从表 2 可以看出,与全谱建模相比,根据 CARS 和随机蛙跳选出的重要变量建模,能够获得与全谱建模相当的预测性能。但随机蛙跳需要预先设置种子变量,然后采用蒙特卡洛抽样技术,将其他波长变量依次添加到种子变量中形成不同的变量子集,根据均方根误差确定最优的波长变量子集,其迭代次数和变量选择结果均依赖于种子变量初始值,算法的计算开销较大。与 CARS 和随机蛙跳相比,RPLS 波长变量选择方法通过设置 *eta* 和 k 的值来控制变量选择个数,并不是以变量子集的预测均方根误差来衡量变量子集的优劣,而是在变量稀疏性和预测能力之间取得折中,方法更为稳健。RPLS 方法选择的波长变量数最少,预测性能优于 CARS 和随机蛙跳,且所选择的变量具有较好

的化学意义,模型可解释性较好。

4 结　语

通过在偏最小二乘回归模型中引入 L1 和 L2 范数罚正则项,建立了正则偏最小二乘波长选择方法。该方法能够将近红外光谱中噪声变量在主成分上的载荷系数置为 0,保留有效变量,达到选择重要变量的目的。与 CARS 和随机蛙跳变量选择方法相比,RPLS 变量选择方法在选择的波长数、模型的预测精度及可解释性等方面均具有一定优势。

本文提出的正则偏最小二乘波长选择方法对噪声变量和有效变量施加相同的惩罚,对主成分载荷系数的估计是有偏估计。如何减弱甚至消除对有效变量的惩罚,得到载荷系数的近似无偏估计,提高近红外光谱波长变量选择的针对性将是下一步的研究方向。

参 考 文 献

[1] 褚小立,陆婉珍. 近五年我国近红外光谱分析技术研究与应用进展[J]. 光谱学与光谱分析,2014,34(10): 2595-2605.

[2] 涂瑶生,柳俊,张建军. 近红外光谱技术在中药生产过程质量控制领域的应用[J]. 中国中药杂志,2011,36(17): 2433-2436.

[3] 林晏清, 胡澍芃, 刘晶,等. 一种基于 Java 平台的近红外光谱实时分析系统设计与开发[J]. 计算机应用与软件,2017,34(12):8-12,106.

[4] 陈立旦,赵艳茹. 可见-近红外光谱联合随机蛙跳算法检测生物柴油含水量[J]. 农业工程学报,2014,30(8): 168-173.

[5] 陈昭,吴志生,史新元,等. Bagging 偏最小二乘和 Boosting 偏最小二乘算法的金银花醇沉过程近红外光谱定量模型预测能力研究[J]. 分析化学,2014,42(11):1679-1686.

[6] 孔康,汪群山,梁万路. L1 正则化机器学习问题求解分析[J]. 计算机工程,2011,37(17):175-177.

[7] 邵言剑,陶卿,姜纪远,等. 一种求解强凸优化问题的最优随机算法[J]. 软件学报,2014,25(9):2160-2171.

[8] Cunningham J P, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations[J]. Journal of Machine Learning Research,2015,16(1):2859-2900.

[9] Gui J, Sun Z, Ji S, et al. Feature selection based on structured sparsity: a comprehensive study[J]. IEEE transactions on neural networks and learning systems, 2017, 28(7): 1490-1507.

[10] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(3): 273-282.

[11] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of statistical software,2010,33(1):1-22.

[12] 刘建伟,崔立鹏,刘泽宇,等. 正则化稀疏模型[J]. 计算机学报,2015,38(7):1307-1325.

[13] Rasmussen M A, Bro R. A tutorial on the Lasso approach to sparse modeling[J]. Chemometrics and Intelligent Laboratory Systems,2012,119:21-31.

[14] 张新玉,王颖杰,刘若西,等. 近红外光谱技术应用于玉米单籽粒蛋白质含量检测分析的初步研究[J]. 中国农业大学学报,2017,22(5):25-31.

[15] Chun H, KeleS S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2010,72(1):3-25.

[16] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2005,67(2):301-320.

[17] Efron B, Hastie T, Johnstone I, et al. Least angle regression [J]. The Annals of statistics,2004,32(2):407-499.

[18] Lee D, Lee W, Lee Y, et al. Sparse partial least-squares regression and its applications to high-throughput data analysis [J]. Chemometrics and Intelligent Laboratory Systems,2011, 109(1):1-8.

[19] 李四海, 陈建国, 任国瑾. 近红外光谱技术快速测定当归中藁本内酯含量[J]. 传感器与微系统, 2017,34(12):114-117.

(上接第 98 页)

[3] 黄淑玲. 基于 Actionscript 3.0 自定义路径动画类的设计[J]. 计算机时代,2016(4):68-69.

[4] 叶凤华,叶欢. 基于 Flash 和 3D 动画渲染技术的育苗机器人设计[J]. 农机化研究,2018,40(3):189-192.

[5] 张宇. 化工生产技术管理与化工安全生产的关联性[J]. 化工管理,2016(16):123-123.

[6] 孙晓波,王荣浩,张鹏举. 基于 Flash 的远程工业监控系统设计[J]. 电子设计工程,2012,20(9):64-67.

[7] 冯梅,姜联成,苏会忠. 高温高压反应釜自控系统的设计[J]. 化工自动化及仪表,2017,44(8):730-733.

[8] 孟祥增,徐振国,刘瑞梅. 基于内容结构的网络 Flash 动画检索方法[J]. 中国图书馆学报,2016,42(1):83-95.

[9] 罗立宏,谭夏梅. 基于外部 Flash 的 Web3D 虚拟场景二维导航[J]. 计算机应用与软件,2018,35(2):156-160.