

基于近红外光谱的大米蛋白质含量快速检测

Rapid detection of protein content in rice based on near infrared spectroscopy

殷 坤¹ 刘金明² 张东杰^{1,3,4} 张爱武^{1,3}

YIN Kun¹ LIU Jin-ming² ZHANG Dong-jie^{1,3,4} ZHANG Ai-wu^{1,3}

(1. 黑龙江八一农垦大学食品学院, 黑龙江 大庆 163319; 2. 黑龙江八一农垦大学信息与电气工程学院, 黑龙江 大庆 163319; 3. 黑龙江八一农垦大学国家杂粮工程技术研究中心, 黑龙江 大庆 163319; 4. 黑龙江省农产品加工与质量安全重点实验室, 黑龙江 大庆 163319)

(1. College of Food Science, Heilongjiang Bayi Agriculture University, Daqing, Heilongjiang 163319, China; 2. College of Information and Electrical Engineering, Heilongjiang Bayi Agriculture University, Daqing, Heilongjiang 163319, China; 3. National Coarse Cereals Engineering Research Center, Heilongjiang Bayi Agricultural University, Daqing, Heilongjiang 163319, China; 4. Key Laboratory of Agro-Products Processing and Quality Safety of Heilongjiang Province, Daqing, Heilongjiang 163319, China)

摘要: 针对国标化学检测方法耗时耗力、成本昂贵的问题, 分析了近红外光谱(NIRS)结合化学计量学进行大米蛋白质含量检测的可行性。基于变量选择、特征提取和非线性建模的策略, 将反向区间偏最小二乘(BiPLS)与主成分分析(PCA)和支持向量机(SVM)相结合, 构建了BiPLS-PCA-SVM模型, 用于提高蛋白质回归模型的性能。在BiPLS-PCA-SVM模型中, 将蒙特卡罗交叉验证与预测残差平方和相结合进行最佳主成分个数的选择, 通过遗传模拟退火算法对模型参数进行优化。为了评估BiPLS-PCA-SVM模型的性能, 建立了Full-PLS、BiPLS和BiPLS-SVM 3种模型, 并系统分析了上述模型的预测精度和鲁棒性。BiPLS-PCA-SVM模型在预测蛋白质含量方面显示的性能高于其他模型, 使用最佳主成分和优化后的SVM参数建立的模型具有更高的鲁棒性和准确性。对于BiPLS-PCA-SVM模型, 验证集的决定系数、均方根误差和剩余预测偏差分别为0.928 9、0.196 7%和4.024 6。结果表明, NIRS与BiPLS-PCA-SVM模型相结合, 可作为替代策略实现大米中蛋白质含量的快速检测。

关键词: 大米; 蛋白质; 近红外光谱; 反向区间偏最小二乘; 主成分分析; 支持向量机

Abstract: In view of the time-consuming, labor-intensive, and costly problem of chemical detection methods in the national standard, the feasibility of near-infrared spectroscopy (NIRS) combined with chemometrics for rapid detection of rice protein was investigated. Based on strategies of variable selection, feature extraction and nonlinear modeling, BiPLS-PCA-SVM was constructed by combining reverse interval partial least squares (BiPLS) with principal component analysis (PCA) and support vector machine (SVM) to improve the performance of the protein regression model. In BiPLS-PCA-SVM, the optimal number of principal components (PCs) was selected by combining Monte Carlo cross validation with the predicted residual sum of squares, and the model parameters were optimized by genetic simulated annealing algorithm. To evaluate the performance of BiPLS-PCA-SVM, three different models, including Full-PLS, BiPLS and BiPLS-SVM, were established, and the prediction accuracy and model robustness of all models were systematically analyzed. The performance of BiPLS-PCA-SVM model in predicting protein content was higher than that of other models, and the model established by using the optimal number of PCs and optimized SVM parameters had higher robustness and accuracy. For BiPLS-PCA-SVM, the determination coefficient, root-mean square error and residual predictive deviation of the validation set were 0.928 9, 0.196 7% and 4.024 6, respectively. The results showed that NIRS combined with BiPLS-PCA-SVM model could be used as a reliable alternative strategy to realize the rapid detection of

基金项目: 国家重点研发计划(编号:2018YFE0206300); 黑龙江省自然科学基金研究团队项目(编号:TD2020C003); 中央引导地方项目(编号:ZY18B01); 黑龙江省博士后基金项目(编号:LBH-Z19087); 黑龙江八一农垦大学三横三纵支持计划(编号:ZRCQC202007)

作者简介: 殷坤, 男, 黑龙江八一农垦大学在读硕士研究生。

通信作者: 张爱武(1964—), 女, 黑龙江八一农垦大学高级实验师。E-mail: 1940720058@qq.com

收稿日期: 2020-12-31

protein content in rice.

Keywords: rice; protein; near infrared spectroscopy; backward interval PLS; principal component analysis; support vector machine

大米蛋白质是公认的优质植物蛋白质之一^[1],其氨基酸组成均衡合理,不会产生过敏反应^[2]。研究^[3]表明,大米中蛋白质含量与大米的食味品质呈负相关,蛋白质含量增加,会导致米饭的硬度增大、弹性降低、色泽变差,大米的食味品质变差。大米蛋白质含量的检测是评价大米质量的重要手段之一。目前,蛋白质含量的测定方法主要依据 GB 5009.5—2016《食品安全国家标准 食品中蛋白质的测定》中的凯氏定氮法、分光光度法和燃烧法。该国标检测方法具有准确性好、灵敏度高的特点,但操作繁琐、检测时间长且破坏样品。

近红外光谱(near infrared spectroscopy, NIRS)分析技术因其简便、快速、低成本、无损以及多组分同步检测的优点被广泛应用于大米蛋白质、脂肪等成分含量的快速检测。苗雪雪等^[4]使用偏最小二乘(partial least squares, PLS)、组合区间 PLS 和移动窗口 PLS 3 种方法对 327 份大米样本的 NIRS 和水分含量进行建模,结果表明移动窗口 PLS 法建立的模型的预测集相关系数达到了 0.961 7,可以用于大米水分含量的快速检测。刘文丽等^[5]使用 OPUS 定量分析软件分别对大米蛋白质和蛋白质(干基)进行 PLS 建模,得到的最佳近红外模型决定系数 R^2 分别为 0.956 5 和 0.974 1,剩余预测偏差(residual predictive deviation, RPD)分别为 4.79 和 6.21,说明两种模型都有较好的精确性,可以用于预测大米中蛋白质含量。俞法明等^[6]对同一批样品的稻谷、糙米、精米、糯米粉分别建立了蛋白质含量 NIRS 检测模型,发现精米和糯米粉回归模型的校正决定系数大于 0.90,可用于进行大米蛋白质含量的快速检测。

上述大米组分 NIRS 快速检测模型主要采用 PLS 线性建模方法建立相应的校正模型,但 PLS 难以体现光谱数据与待测目标组分之间的非线性关系。因此,相关学者开始研究使用支持向量机(support vector machine, SVM)和神经网络等非线性多元统计方法建立 NIRS 定量回归模型^[6]。其中,SVM 是基于统计学习理论和结构风险最小化原则的机器学习方法^[7],与传统的神经网络相比结构简单、泛化能力强,对求解小样本、非线性、高维数问题有更好的适应性。但是在应用 SVM 进行光谱数据非线性建模时,因 NIRS 原始数据维数过高,若直接以全部的波长数据作为 SVM 的输入变量,将导致模型过于复杂、运行时间长,而且冗余光谱数据还会影响模型的建模精度^[8]。李子文等^[9]分别使用 PLS 和最小二乘 SVM 建立了 372 个核桃露样品脂肪含量 NIRS 定量检测模型,

结果表明两种模型均有较高的精度,且最小二乘 SVM 建立的模型性能更高。孙丽萍等^[10]使用 SVM 建立了东北黑木耳多糖含量的 NIRS 分类模型,模型精确率为 85.7%,召回率为 87.3%,说明 SVM 可以用来建立黑木耳多糖含量分类模型。

相关学者将主成分分析(principal component analysis, PCA)与 SVM 相结合构建 PCA-SVM 用于建立 NIRS 快速检测模型。张晓冬等^[11]使用 PCA-SVM 建立了 4 种含铁矿物药的 NIRS 快速鉴别模型,结果表明所建的模型预测能力强,可以用于 4 种含铁矿物药的快速鉴别。魏从师等^[12]利用 NIRS 法结合 PCA 和 SVM 建立了 6 味中药的识别模型,所建的模型对训练集和验证集样品预测的正确率均达到 100%。上述研究均使用网格搜索(grid search, GS)对 SVM 建模参数进行优化,但 GS 法搜索步长过大时结果不准确,搜索步长小则运行时间较长^[13]。而且前人的研究大多基于全谱进行建模,在特征波长优选与 PCA-SVM 结合方面少有涉猎。研究拟基于遗传模拟退火算法(genetic simulated annealing algorithm, GSA)对 PCA-SVM 的参数进行优化,探讨将 PCA-SVM 与反向区间偏最小二乘(backward interval partial least squares, BiPLS)相结合建立大米蛋白质含量 NIRS 快速检测模型的可行性,以 BiPLS 优选后的谱区作为 PCA-SVM 的输入参数,建立大米蛋白质含量快速检测模型。

1 材料与方法

1.1 材料与仪器

1.1.1 材料与试剂

大米:采集自黑龙江省查哈阳、五常、方正、响水和建三江 5 个地区 150 个样品,垄谷 3 次、碾米 2 次后粉碎,装入密封袋于阴凉干燥处保存;

硫酸、硫酸钾、硼酸、乙醇:分析纯,辽宁泉瑞试剂有限公司;

硫酸铜:分析纯,哈尔滨市新达化工厂;

氢氧化钠:分析纯,天津市博华通化产品销售中心;

甲基红、溴甲酚绿:分析纯,天津市恒兴试剂制造有限公司。

1.1.2 主要仪器设备

垄谷机:THU35C 型,佐竹机械(苏州)有限公司;

碾米机:TM05C 型,佐竹机械(苏州)有限公司;

粉碎机:CT193 型,福斯分析仪器(苏州)有限公司;

电子分析天平:AR323CN 型,奥豪斯仪器(上海)有限公司;

全自动凯氏定氮仪:K9860 型,济南海能仪器股份有限公司;

石墨消解仪:SH220N 型,济南海能仪器股份有限公司;

傅里叶变换近红外光谱仪: TANGO 型, 德国 BRUKER 公司。

1.2 方法

1.2.1 蛋白质含量测定 按 GB 5009.5—2016 的凯氏定氮法执行。

1.2.2 光谱数据采集 粉碎后的大米样品填满样品杯, 振荡均匀后使用近红外光谱仪进行积分球漫反射光谱扫描, 光谱范围 $3\,946 \sim 11\,542\text{ cm}^{-1}$ ($866 \sim 2\,534\text{ nm}$), 分辨率为 8.0 cm^{-1} , 每个样品扫描 32 次, 背景每小时扫描一次。保持室内温湿度基本不变, 每个样品装样 3 次, 取 3 次扫描平均值作为样品原始光谱。

1.2.3 BiPLS BiPLS 是在间隔偏最小二乘法 (interval partial least square, iPLS)^[14] 的基础上发展而来的一种特征谱区优选算法。BiPLS 在 iPLS 的基础上依次去除交叉验证均方根误差 (root mean squared error of cross-validation, RMSECV) 值最大的区间, 当 RMSECV 最小时所对应的子区间组合为优化后的特征谱区域^[15]。

1.2.4 PCA-SVM PCA 作为一种多元统计方法, 能够消除各指标间的相互关联影响, 在数据降维方面应用广泛^[16]。SVM 非线性回归的核心思想是通过选取的核函数将非线性问题映射为高维特征空间的线性问题, 并在这个空间中进行线性回归^[17]。

PCA-SVM 融合了 PCA 光谱数据降维方法和 SVM 非线性回归思想, 既解决了 PLS 在处理光谱数据多重共线性时效果不佳的问题, 又弥补了以全谱波长变量直接建模效率低的不足。在采用径向基 (radial basis function, RBF) 核函数进行 PCA-SVM 建模过程中, 主成分 (principal components, PCs) 个数、惩罚参数 C 、RBF 核函数参数 γ 和不敏感损失函数参数 ϵ 对建模性能具有重要影响。采用 MCCV 结合 PLS 回归模型进行最佳 PCs 个数的选取。通过比较 PLS 回归模型的预测残差平方和 (prediction residual error sum of squares, PRESS), 选取 PRESS 值最小的 PCs 个数作为最佳 PCs 个数^[18]。为了有效提高 PCA-SVM 的学习和泛化能力, 采用 GSA 算法^[19] 对回归模型的参数 C 、 γ 和 ϵ 进行优化。GSA 以 PCA-SVM 回归模型的 k 折 RMSECV 为目标函数, 结合温度参数进行适应度优化设计如下:

$$f_{it}(x) = \exp \left[-\frac{f(x) - f_{\min}}{t} \right], \quad (1)$$

式中:

$f_{it}(x)$ ——适应度函数;

$f(x)$ ——当前染色体目标函数值;

$f_{\min}$ ——当前代种群中的最小目标函数值;

t ——当前代温度值。

1.2.5 模型建立及评价 为了改善光谱特征、修正基线

漂移和光谱散射、消除仪器产生的随机噪声, 有效提高光谱数据的信噪比, 采用 Savitzky-Golay (SG) 平滑、多元散射校正 (multivariate scattering correction, MSC)、标准正态变换、一阶导数、二阶导数及其组合对光谱数据进行预处理。采用 Kernard-Stone (KS) 法^[20] 将预处理后的样本光谱数据按 4 : 1 的比例划分为校正集和验证集, 建立全谱 PLS 回归模型, 并基于校正集的 RMSECV 确定采用的光谱预处理方法为 SG 平滑结合 MSC。选定预处理方法后, 使用 BiPLS 进行大米蛋白质特征谱区优选, 以优选后的特征谱区建立 PCA-SVM 定量分析模型, 并采用校正决定系数 (R_c^2)、验证决定系数 (R_p^2)、校正均方根误差 (root mean squared error of calibration, RMSEC)、预测均方根误差 (root mean squared error of prediction, RMSEP) 和 RPD 对校正模型的性能进行评价。 R^2 (R_c^2 和 R_p^2) 和 RPD 的值越大, 均方根误差 (root mean squared error, RMSE) (RMSEC 和 RMSEP) 的值越小, 模型的精度越高。同时, R_c^2 和 R_p^2 或 RMSEC 和 RMSEP 之间的差异越小越好, RMSEC 和 RMSEP 之间很小的不同表明模型具有很好的鲁棒性。 R^2 、RMSE 和 RPD 的计算公式为:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2)$$

$$R_{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad (3)$$

$$R_{\text{PD}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}, \quad (4)$$

式中:

R^2 ——决定系数;

R_{MSE} ——均方根误差;

R_{PD} ——剩余预测偏差;

y_i ——第 i 个样本的测量值;

$\hat{y}_i$ ——第 i 个样本的预测值;

$\bar{y}$ ——测量平均值;

n ——样本个数。

研究所有算法, 包括光谱预处理、样本集划分、特征谱区优选、主成分分析、模型参数优化和回归模型构建等全部在 Matlab R2016b 软件平台中实现, 其中 BiPLS 和 SVM 分别基于 iToolBox^[21] 和 LibSVM^[22] 工具箱实现。

2 结果与分析

2.1 采集数据分析

大米样品光谱数据如图 1 所示。由图 1(a) 可知, 原始光谱数据存在基线漂移现象。由图 1(b) 可知, 通过 SG 平滑和 MSC 多元散射校正相结合对原始光谱进行预处理, 修正了因扫描时间、周边环境、设备状态不同而导致

的基线漂移、噪声干扰和光谱散射等问题,有效提高了光谱数据的分辨率和信噪比。其中,SG 主要用于消除随机噪声的干扰,MSC 主要用于修正基线漂移和光谱散射。对 NIRS 数据进行预处理后,使用 KS 法按 4 : 1 的比例进行样本集划分,得到校正集样本 120 个、验证集样本 30 个。样品中蛋白质含量测量结果如表 1 所示。

对预处理后的 NIRS 数据进行 PCA 计算,第一、第二和第三 PCs 的贡献率分别为 82.72%,7.31%,6.07%,前 3 个 PCs 的累积贡献率达 96.10%。校正集和验证集的三维 PCs 空间分布情况如图 2 所示。

由表 1 和图 2 可知,校正集基本涵盖了验证集,且校正集和验证集样本在 PCs 空间上分布均匀,可以用该样本划分方法进行 NIRS 建模。

2.2 特征谱区优选

在使用 BiPLS 算法进行 NIRS 特征谱区优选时,区间划分个数的选择很重要。为寻找最佳区间划分个数,

首先将预处理后的样品光谱数据划分为 k 个子区间 ($k=10,20,30,40,50,60,70,80,90$),并采用 BiPLS 计算每个 k 值下的最佳特征谱区,完成划分区间个数的初选(如表 2 所示)。由表 2 可知, $k=70$ 时,蛋白质 BiPLS 回归模型的 RMSECV 最小; $k=80$ 时,其 RMSECV 比 $k=70$ 时略大,因此可以推测蛋白质最佳谱区划分个数为 70~80。

为获得蛋白质最佳谱区划分个数,采用 BiPLS 计算区间个数从 71 到 79 对应的最佳特征谱区(如表 3 所示)。由表 3 可知,区间划分个数为 79 时,BiPLS 优选出 42 个蛋白质特征谱区对应的 RMSECV 最小(0.211 0),对应区间列表为[1 3 4 7 8 9 10 12 15 17 19 20 22 23 25 26 27 30 31 32 33 34 36 37 39 40 44 46 47 48 50 57 58 63 64 68 69 70 72 74 76 77],选中的特征波长变量共 983 个。

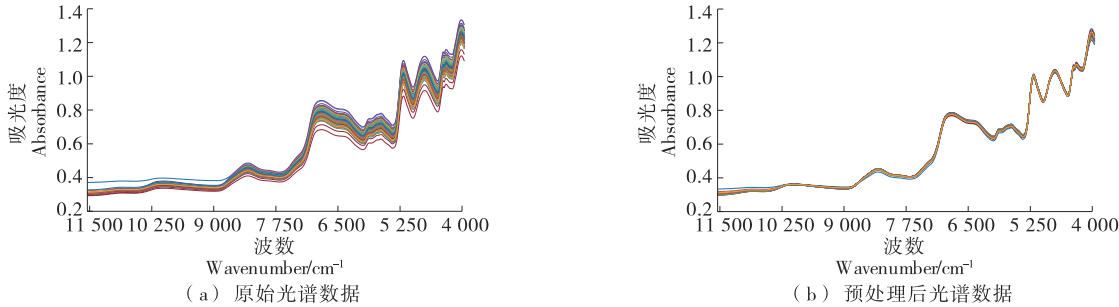


图 1 样品光谱数据
Figure 1 Spectral data of samples

表 1 样品中蛋白质含量
Table 1 Protein content in samples

项目	平均值/%	最大值/%	最小值/%	标准差/%	变异系数/%
校正集($n=120$)	6.755 0	9.350 6	4.682 6	0.957 1	14.168 5
验证集($n=30$)	6.676 5	9.548 8	5.263 4	0.824 1	12.344 1

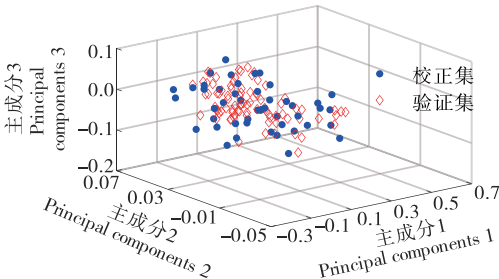


图 2 样本在主成分空间中的分布
Figure 2 Distribution of the samples in principal components space

表 2 BiPLS 优化蛋白质光谱特征区间初选结果

Table 2 Primary results of BiPLS optimization of protein spectral feature interval

区间个数	选中区间个数	波长变量个数	RMSECV/%
10	8	1 475	0.247 2
20	5	461	0.243 3
30	17	1 043	0.233 5
40	15	692	0.229 0
50	14	514	0.225 4
60	21	642	0.224 2
70	20	525	0.213 9
80	33	763	0.216 3
90	39	800	0.284 4

表 3 BiPLS 最佳特征谱区筛选结果

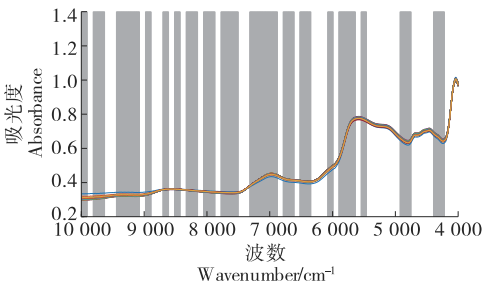
Table 3 Results of BiPLS optimal feature spectrum screening

区间个数	选中区间个数	选中区间列表	波长变量个数	RMSECV/%
71	15	[7 11 28 30 41 42 45 52 57 61 64 65 68 69 70]	390	0.220 0
72	22	[3 7 10 11 23 24 29 30 35 37 43 45 52 58 59 62 65 66 67 69 70 71]	562	0.218 1
73	17	[1 7 10 18 27 30 43 53 59 60 65 66 67 70 71 72 73]	429	0.222 3
74	19	[1 3 8 9 11 12 13 28 31 44 54 55 60 64 66 69 71 72 73]	472	0.218 2
75	39	[1 3 7 8 9 10 11 12 13 14 15 18 19 22 23 24 25 27 29 31 32 34 37 38 43 44 46 47 52 54 55 60 61 66 68 70 72 73 74]	962	0.221 3
76	46	[1 2 3 6 8 9 10 12 13 14 16 17 18 19 22 23 24 25 26 27 29 30 31 32 33 34 35 42 44 45 47 48 52 53 55 61 62 65 66 67 69 70 71 73 74 75]	1 118	0.223 0
77	24	[1 2 7 8 9 10 13 15 26 27 30 31 32 46 48 54 56 63 66 68 70 71 74 76]	575	0.214 2
78	32	[1 2 7 8 9 10 13 15 16 17 19 20 22 25 26 29 30 32 33 38 40 46 53 56 62 67 69 71 73 75 76 77]	758	0.211 6
79	42	[1 3 4 7 8 9 10 12 15 17 19 20 22 23 25 26 27 30 31 32 33 34 36 37 39 40 44 46 47 48 50 57 58 63 64 68 69 70 72 74 76 77]	983	0.211 0

绘制 BiPLS 优选的蛋白质特征谱区与光谱数据对比图(如图 3 所示)。图 3 中,特征谱区波数 9 075 ~ 9 450 cm⁻¹ 区域对应着蛋白质—RNH₂基团的三级倍频;波数 6 900~7 300 cm⁻¹区域对应着蛋白质—CONH₂基团的二级倍频;波数 5 640~5 880 cm⁻¹对应着蛋白质—CH 和—CH₂基团的一级倍频,波数 4 220 ~ 4 370 cm⁻¹区域对应着蛋白质—CH、—CH₂和—CH₃基团的组合频。

2.3 回归模型建立及评价

最佳 PCs 个数的选择是建立 PCA-SVM 模型的第一



灰色阴影区域为蛋白质特征谱区

图 3 蛋白质光谱特征谱区

Figure 3 Spectral characteristic intervals of protein

步,合理的 PCs 个数既可以提高模型的预测精度,又可以提高模型的稳定性。采用 MCCV 结合 PLS 回归模型进行最佳 PCs 个数的选取。设 PCs 个数的搜索范围为 1~20,步长为 1,分别建立不同 PCs 个数下的 PLS 回归模型,并计算每个 PLS 回归模型的 PRESS,选取 PRESS 值最小时对应的 PCs 个数作为最佳 PCs 个数。以 BiPLS 优选的 983 个蛋白质特征波长变量为输入,采用 MCCV 结合 PLS 回归模型分析 PRESS 随 PCs 个数变化情况如图 4 所示。由图 4 可知,随着 PCs 个数的增加,PRESS 呈先迅速减少,后缓慢增加的趋势。图 4 中箭头所示红色圆点的 PRESS 值最小(2.182 6),PCs 的个数为 11,说明这 11 个 PCs 的解释能力最强。选取此点的 PCs 个数作为大米蛋白质 BiPLS 特征波长优选后的最佳 PCs 个数建立 PCA-SVM 模型。

遗传算法(genetic algorithm,GA)因其具有强大的搜索能力,在 SVM 参数优化方面得到了广泛应用^[23]。但 GA 存在早熟问题,且进化后期搜索效率较低。因此,采用 GSA 算法对 PCA-SVM 回归模型的参数 C、γ 和 ε 进行优化。参数 C、γ 和 ε 的寻优范围分别是[0,100]、[0,100]和[0.001,1.000],染色体码长为 60(每个参数对应 20 位基因),种群规模为 50,遗传代数为 200,初始温度参

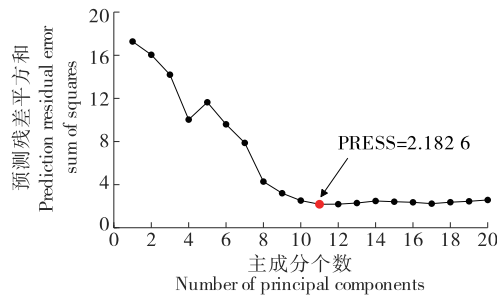


图 4 PRESS 与 PCs 关系图

Figure 4 Relationship between PRESS and the number of PCs

数 K 为 200, 退温系数 α 为 0.95, 交叉概率为 0.7, 变异概率为 $0.7/L$ (L 为染色体码长), 采用 10 折 RMSECV 为目标函数。以 BiPLS 优选的 983 个大米蛋白质特征波长变量按最佳 PCs 个数为 11, 使用校正集数据进行 PCA-SVM 回归模型参数优化。参数优化的过程如图 5 所示。由图 5 可知, 在高温时(进化前期)GSA 求得的目标函数值与最佳目标函数值差异很大, 而低温时(进化后期)平均目标函数值更接近于最佳目标函数值。原因在于 GSA 算法融合了 GA 和模拟退火算法的优点, 结合温度参数设计适应度函数和 Metropolis 选择复制策略, 既提高了算法高温时(进化前期)种群的多样性, 又扩大了低温时(进化后期)优良染色体的适应度函数值, 能够有效克服 GA 早熟收敛和进化后期搜索效率低的不足。

为了分析 BiPLS 与 PCA-SVM 相结合建立大米蛋白质含量 NIRS 快速检测模型的性能, 分别建立全谱 PLS 和 BiPLS 优选谱区的 PLS、SVM 和 PCA-SVM 回归模型, 依次记为 Full-PLS、BiPLS、BiPLS-SVM 和 BiPLS-PCA-SVM。不同回归模型的评价指标如表 4 所示。

由表 4 可知, PCA-SVM 回归模型的泛化能力最强, 针对蛋白质 BiPLS 优选谱区建立的回归模型性能优于 SVM 和 PLS 回归模型。同时, SVM 回归模型的建模性能高于 PLS 回归模型。说明当波长变量数量较多且存在一定冗余波长变量时, SVM 与 PLS 相比, 能够充分发挥非线性映射的优势, 建立性能较好的校正模型。PCA-SVM 回归模型使用 PCA 对光谱数据进行简化, 通过选取贡献率大的最佳 PCs 个数, 弱化了不相关和共线性冗余波长变量对建模性能的影响, 通过其良好的泛化能力

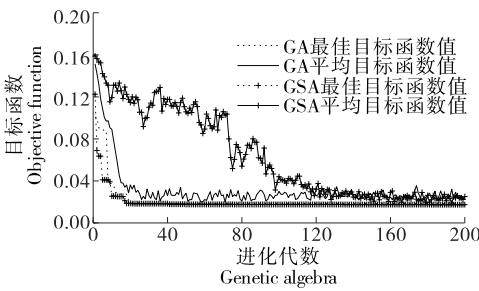


图 5 基于特征波长的蛋白质 PCA-SVM 的参数优化过程

Figure 5 Optimization process of parameters for PCA-SVM based on characteristic wavelengths for protein

解决了高维数据的复杂性, 有效提高了建模精度和效率。BiPLS 通过搜索 RMSECV 最小的光谱区间组合, 能够获取蛋白质的敏感建模波段。蛋白质的 BiPLS-PCA-SVM 回归模型的 R_p^2 为 0.928 9, RMSEP 为 0.196 7%, RPD 为 4.024 6。模型的 $R_p^2 > 0.9$, $RPD > 3.0$, 说明 BiPLS-PCA-SVM 回归模型能够满足蛋白质含量快速检测的需求^[31]。以 BiPLS 优选的 983 个特征波长变量建立 SVM, 由于波长变量个数较多, BiPLS-SVM 模型的建模效率较低。但以 BiPLS 优选的 983 个特征波长变量建立 PCA-SVM 回归模型时, 983 个波长变量经 PCA 降维处理后选用 11 个 PCs 进行 SVM 建模, 有效缩短了模型运行时间。表 4 中 BiPLS-SVM 和 BiPLS-PCA-SVM 回归模型的校正集评价指标略低于验证集, 可能是由于采用 KS 法按 4:1 比例划分校正集和验证集, 得到 120 个校正集样本和 30 个验证集样本, 校正集样本较多且完全涵盖了验证集造成的。采用 BiPLS-PCA-SVM 建立蛋白质校正模型并绘制测量值与预测值空间分布图, 如图 6 所示。

由图 6 可知, 蛋白质的测量值与预测值点基本呈对角线分布, 样本点位于对角线两侧, 说明蛋白质回归模型的预测效果较好。经 t 检验计算, 得到样本蛋白质含量实测值与预测值的 t 值为 1.058 2, 小于相应的临界值 $t_{(0.05, 149)} = 1.976$, Matlab ttest 函数返回检验结果均为 0, 显著性水平概率为 0.928 2, 说明蛋白质测量值与预测值无显著性差异, 一致性较好。图 6 中存在个别偏离对角线较严重的样本点, 原因在于部分样品消化过程中因淀

表 4 不同回归模型的评价指标

Table 4 Evaluation indicators of different regressive models

模型	波长变量个数	PCs	C	γ	ϵ	R_c^2	R_p^2	RMSEC/%	RMSEP/%	RPD
Full-PLS	1 557	12	—	—	—	0.982 3	0.914 2	0.126 2	0.270 4	3.047 8
BiPLS	983	11	—	—	—	0.954 7	0.914 7	0.210 9	0.234 3	3.378 4
BiPLS-SVM	983	—	99.651 4	0.009 3	0.016 9	0.927 8	0.935 7	0.244 1	0.196 8	4.021 5
BiPLS-PCA-SVM	983	11	1.709 1	0.121 6	0.075 1	0.911 4	0.928 9	0.264 5	0.196 7	4.024 6

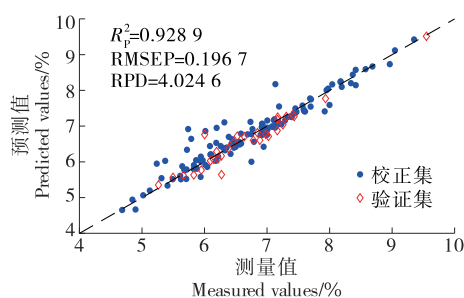


图 6 测量值和预测值分布图

Figure 6 Distribution of measured and predicted values

粉含量过高,导致在消化温度 300 °C 左右时产生大量黑色气泡粘壁,消化温度提高到 420 °C 后部分附着在消化管壁的气泡无法被消化,从而形成未消化完全的黑色颗粒物而产生的误差。因此在建立 NIRS 快速检测模型时,一定要保证前期测量的基础目标属性化学含量值的准确。BiPLS-PCA-SVM 融合了 BiPLS 特征谱区优选、PCA 光谱数据降维、MCCV 的最佳主成分选取、SVM 非线性建模、GSA 参数优化,既减少了 SVM 模型的输入数据量,又提高了 SVM 模型的回归精度,还有效解决了线性方法建模的多重共线性问题。在建立基于 BiPLS-PCA-SVM 的大米蛋白质含量 NIRS 快速检测模型过程中,并未对图 6 中偏离对角线的异常样本进行剔除,但建立的模型仍能满足大米蛋白质含量的快速检测需求。说明使用 BiPLS-PCA-SVM 建立 NIRS 检测模型时,在保证足够多样本的前提下,少量样本化学含量测定存在一定误差对最终模型的影响不大,充分体现了文中提出的 BiPLS-PCA-SVM 建模方法具有较高的鲁棒性。

3 结论

试验探讨了近红外光谱结合化学计量学方法进行大米蛋白质含量快速检测的可行性。为了提高近红外光谱回归模型的检测精度和效率,将反向区间偏最小二乘(Bi-PLS)、主成分分析(PCA)和支持向量机(SVM)相结合构建 BiPLS-PCA-SVM 模型,所建立回归模型的验证决定系数为 0.928 9,预测均方根误差为 0.196 7%,剩余预测偏差为 4.024 6。结果表明,BiPLS-PCA-SVM 模型通过融合了特征谱区优选、光谱数据降维和非线性建模策略,其建模性能优于全谱偏最小二乘模型和反向区间偏最小二乘优选谱区建立的偏最小二乘和支持向量机回归模型,能够满足大米蛋白质含量的实际检测需求。BiPLS-PCA-SVM 与近红外光谱相结合为建立简单、高效、低成本的大米蛋白质含量快速检测方法提供了新途径。由于时间有限,试验仅将 BiPLS-PCA-SVM 应用于大米蛋白质含量的测定,后期将进一步研究其在大米直链淀粉、脂肪等其他营养成分含量测定方面的应用。

参考文献

- [1] 刘芳,许宙,陈茂龙,等.碱与热处理对大米蛋白质结构与功能性质的影响[J].食品与机械,2018,34(8):10-15.
- [2] 刘文丽,严虞虞,吴东慧,等.近红外光谱技术无损检测大米中蛋白质[J].食品工业,2019,40(1):205-209.
- [3] 吴长明,孙传清,陈亮,等.稻米营养品质性状的 QTL 及其与食味品质的关系研究[J].吉林农业科学,2002,27(4):3-7.
- [4] 苗雪雪,苗莹,龚浩如,等.不同偏最小二乘法的近红外光谱技术测定大米中水分的研究[J].分析科学学报,2019,35(5):643-649.
- [5] 俞法明,陆艳婷,严文潮,等.用近红外光谱技术快速测定籼稻品种的蛋白质含量[J].中国粮油学报,2009,24(5):134-138.
- [6] YANG Mei-hua, XU Dong-yun, CHEN Song-chao, et al. Evaluation of machine learning approaches to predict soil organic matter and pH using vis-NIR spectra [J]. Sensors (Basel), 2019, 19(2): 263-277.
- [7] 刘金明,谢秋菊,王雪,等.基于 GSA-SVM 的畜禽舍废气监测缺失数据恢复方法[J].东北农业大学学报,2015,46(5):95-101.
- [8] 刘金明,郭坤林,甄峰,等.基于近红外光谱的沼液挥发性脂肪酸含量快速检测[J].农业工程学报,2020,36(18):188-196.
- [9] 李子文,李宗朋,买书魁,等.最小二乘支持向量机的核桃露饮品中脂肪成分的定量分析[J].光谱学与光谱分析,2019,39(12):3 916-3 920.
- [10] 孙丽萍,张希萌,何睿,等.基于 SVM 的近红外黑木耳多糖含量分类[J].电子科技,2019,32(8):16-21.
- [11] 张晓冬,陈龙,黄必胜,等.基于 NIRS 和 PCA-SVM 算法快速鉴别 4 种含铁矿物药[J].中成药,2018,40(2):404-410.
- [12] 魏从师,雷福汉,艾伟霞,等.基于 NIRS 技术和 PCA-SVM 算法 6 种树脂及其他类中药的快速鉴别[J].中国实验方剂学杂志,2017,23(9):17-23.
- [13] 顾凯成.改进网格搜索的支持向量机参数优化研究及应用[D].兰州:兰州理工大学,2016:18-19.
- [14] NORGAARD L, SAUDLAND A, WAGNER J, et al. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy[J]. Applied Spectroscopy, 2000, 54(3): 413-419.
- [15] LEARDI R, NØRGAARD L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions[J]. Journal of Chemometrics, 2010, 18(11): 3 222-3 235.
- [16] AMELIA P, WEIN A S, BANDEIRA A S, et al. Optimality and sub-optimality of PCA I: Spiked random matrix models[J]. Annals of Statistics, 2018, 46(5): 2 416-2 451.
- [17] YANG Yue, LIU Xue-song, LI Wei-li, et al. Rapid measurement of epimedin A, epimedin B, epimedin C, icariin, and moisture in Herba Epimedii using near infrared spectroscopy [J]. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 2017, 171: 351-360.

(下转第 175 页)

- [6] 薛月萃, 周英, 康颖倩, 等. 姜黄挥发油的研究进展[J]. 华西药学杂志, 2016, 31(2): 212-215.
- [7] LI Shi-you, WEI Yuan, DENG Guang-rui, et al. Chemical composition and product quality control of turmeric (*Curcuma longa* L.)[J]. Pharmaceutical Crops, 2011, 5(1): 28-54.
- [8] 刘英, 张霄程, 陈启文. 姜黄素对结直肠癌细胞增殖、凋亡的作用机制研究[J]. 现代消化及介入诊疗, 2020, 25(11): 1 475-1 479.
- [9] HAML D H A, MUTAZAH R, YUSOFF M M, et al. Comparative analysis of antitumor and antiproliferative activities of *Rhodymenia tomentosa* extracts prepared with various solvents[J]. Food and Chemical Toxicology, 2017, 108: 451-457.
- [10] GONG Hai-yan, HE Li-juan, LI Shao-yu, et al. Antimicrobial, antifungal and antitumor activities of essential oil of *Agastache rugosa* from Xinjiang, China[J]. Saudi Journal of Biological Sciences, 2016, 23(4): 524-530.
- [11] JAN S, KHAN M R. Antipyretic, analgesic and anti-inflammatory effects of *Curcuma zanzibar* [J]. Journal of Ethnopharmacology, 2016, 182: 90-100.
- [12] 邵佳, 邹俊波, 史亚军, 等. 姜黄挥发油热稳定性研究及其抗氧化剂筛选[J]. 中国药理学杂志, 2019, 54(15): 1 237-1 244.
- [13] 方月娟, 夏道宗, 王思为, 等. 高良姜总黄酮对营养性肥胖合并高脂血症大鼠减脂降脂作用研究[J]. 中华中医药杂志, 2015, 30(8): 2 907-2 910.
- [14] CHEN Bin, ZHANG You-zhi, WANG Yang, et al. Curcumin inhibits proliferation of breast cancer cells through Nrf2-mediated down-regulation of Fen1 expression[J]. The Journal of Steroid Biochemistry and Molecular Biology, 2014, 143: 11-18.
- [15] 张婧菲, 韩红丽, 沈明明, 等. 姜黄素类化合物的体外抗氧化活性及其对鸡红细胞氧化损伤的保护作用[J]. 食品科学, 2020, 41(13): 96-105.
- [16] 赵欣, 王爱里, 袁园, 等. 姜黄中姜黄素, 去甲氧基姜黄素, 双去甲氧基姜黄素的光稳定性分析[J]. 中草药, 2013, 44(10): 1 338-1 341.
- [17] 羊青, 晏小霞, 王茂媛, 等. 不同产地姜黄挥发油的化学成分及其抗氧化活性[J]. 中成药, 2016, 38(5): 1 188-1 191.
- [18] 强悦越, 韦航, 方灵, 等. 福建姜黄挥发油化学成分的 HS-SPME-GC-MS 分析[J]. 中国食品添加剂, 2020, 31(1): 147-153.
- [19] 康显杰, 杜伟锋, 凌珏, 等. 3 种干燥方法对片姜黄挥发油的影响[J]. 中成药, 2017, 39(9): 1 900-1 903.
- [20] 白秀蓉. 不同干燥方法对荆芥药材中 5 种有效成分含量的影响和评价[J]. 中国药师, 2020, 23(12): 2 476-2 479.
- [21] 韩刚, 崔静静, 毕瑞, 等. 姜黄素、去甲氧基姜黄素和双去甲氧基姜黄素稳定性研究[J]. 中国中药杂志, 2008, 33(22): 2 611-2 614.
- [22] 李会伟, 刘培, 钱大玮, 等. 不同干燥方法及其影响因子对玄参药材初加工过程品质形成的影响[J]. 中国中药杂志, 2015, 40(22): 4 417-4 423.
- [23] 张玉秀, 黄嫫, 冯剑, 等. GC-MS 分析不同干燥程度温莪术根茎挥发油的动态变化[J]. 中国实验方剂学杂志, 2018, 24(3): 70-74.
- [24] 吴妙鸿, 强悦越, 吴艺杰, 等. 姜黄中姜黄素类化合物提取工艺研究[J]. 食品安全质量检测学报, 2019, 10(13): 4 328-4 334.
- [25] 刘江, 樊钰虎, 王泽秀, 等. 响应面法优化姜黄挥发油提取工艺及其化学成分研究[J]. 西南师范大学学报(自然科学版), 2012, 37(10): 70-76.
- [26] 卢彩会. 姜黄挥发油的成分分析及性能研究[D]. 石家庄: 河北科技大学, 2018: 6-7.
- [27] 孙秀燕, 李秀琴, 王金辉, 等. 姜黄挥发油抗癌活性成分研究[J]. 中草药, 2006, 37(7): 982-983.
- [28] 蒋建兰, 丁洪涛, 苏鑫, 等. 基于组效关系的姜黄挥发油抗肿瘤活性成分辨析研究[J]. 分析化学, 2012(10): 1 488-1 493.

(上接第 88 页)

- [18] LIANG Long, WEI Lu-lu, FANG Gui-gan, et al. Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection [J]. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 2019, 225: 117515.
- [19] LIU Jin-ming, LI Nan, ZHEN Feng, et al. Rapid detection of carbon-nitrogen ratio for anaerobic fermentation feedstocks using near-infrared spectroscopy combined with BiPLS and GSA[J]. Applied Optics, 2019, 58(18): 5 090.
- [20] STONE R W K A. Computer aided design of experiments[J]. Technometrics, 1969, 11(1): 137-148.
- [21] LEARDI R, NREGAARD L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions[J]. Journal of Chemometrics, 2004, 18(11): 486-497.
- [22] CHANG Chih-hung, LIN Chih-jen. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.
- [23] CAO Jie, ZHU Shi-jie, LI Chao, et al. Integrating support vector regression with genetic algorithm for hydrate formation condition prediction[J]. Processes, 2020, 8(5): 519.
- [24] STOCKL A, LICHTI F. Near-infrared spectroscopy (NIRS) for a real time monitoring of the biogas process[J]. Bioresource Technology, 2018, 268(1): 1 249-1 252.