



基于近红外光谱大豆蛋白质、脂肪快速无损检测模型的优化构建

王翠秀, 曹见飞, 顾振飞, 徐明雪, 吴泉源

(山东师范大学 地理与环境学院, 山东 济南 250358)

摘要:为实现大豆蛋白质、脂肪含量的快速无损检测, 采集 350 ~ 2 500 nm 光谱范围内的大豆近红外光谱。运用经典 Kennard-Stone 算法选取建模样本及验证样本, 对近红外原始光谱进行卷积平滑 (savitzky and golay, SG) + 一阶微分、变量标准化 (standard normal variate, SNV) + 去趋势算法 (de-trending, DT)、正交信号校正 (orthogonal signal correction, OSC) 处理; 然后通过竞争性自适应重加权采样方法 (competitive adaptive reweighted sampling, CARS) 筛选出特征波长, 比较偏最小二乘法 (partial least squares, PLS)、BP 神经网络法所建模型, 最终获得对于大豆蛋白质、脂肪含量的快速、无损检测的最佳模型。结果表明: (1) 经 CARS 特征波段挑选后, 波长的变量个数由 1 981 个减少为 100 个以下, 变量压缩率大于 94.95%; (2) CARS 波段选择能够提高建模精度, 基于挑选的特征波段所建立模型的决定系数均 > 0.9; (3) OSC + CARS + PLS 与 OSC + CARS + BP 该类数据处理组合方式在一定程度上能够实现大豆蛋白质、脂肪的快速、无损检测。优化构建的该模型能够精准快速无损的检测大豆蛋白质、脂肪含量, 对大豆品质评估以及作物改良具有重要意义。

关键词:大豆; 蛋白质; 脂肪; 近红外光谱; 竞争性自适应重加权采样法; BP 神经网络

Rapid Nondestructive Test of Soybean Protein and Fat by Near Infrared Spectroscopy Combined with Different Model Methods

WANG Cui-xiu, CAO Jian-fei, GU Zhen-fei, XU Ming-xue, WU Quan-yuan

(School of Geography and Environment, Shandong Normal University, Shandong 250358, China)

Abstract: In order to realize the rapid nondestructive testing of soybean protein and fat content, the near infrared spectrum of soybean was collected in the range of 350 – 2 500 nm. The classical kennard-stone algorithm was used to select modeling samples and verification samples, SG + first-order differential, SNV + DT, OSC processing. Then, the characteristic wavelength was selected by the competitive adaptive weighted sampling method (CARS), the detection model of soybean protein, fat content established by partial least squares (PLS) and BP neural network method was compared. Finally, the detection of soybean protein and fat content was achieved. It was demonstrated that: (1) After the CARS method was preferred, the number of wavelength variables was reduced from 1 981 to less than 100, and the variable compression ratio was greater than 94.95%. (2) The CARS band selection improved the modeling accuracy, and the determination coefficient of the model based on the selected feature bands was higher than 0.9. (3) The combination of OSC + CARS + PLS and OSC + CARS + BP could achieve rapidly and non-destructive detection of soy protein and fat to a certain extent. The optimized model can accurately, rapidly and nondestructively detect the accurate estimation of soybean protein and fat content, which is of great significance for soybean quality evaluation and crop improvement.

Keywords: Soybean; Protein; Fat; Near infrared spectroscopy; Competitive adaptive weighted sampling method; BP neural network

大豆蛋白质、脂肪含量是评价大豆品质的重要指标之一, 对疾病的预防也有重要的作用^[1-4]。目前蛋白质和油脂测量通常是采用凯氏定氮法和索氏提取法, 该类化学方法具有应用范围广、灵敏度高、仪器价钱适中的优点但也存在操作复杂、破坏样品、费时费力、污染环境等缺点^[5]。而采用 VNIR 可以实现蛋白质和油脂的快速、低成本、绿色无污染测量。但是原始光谱往往会存有大量噪声, 且波

段起伏不明显, 需要用到不同的预处理方式, 而不同的预处理方式, 对模型预测结果有很大的影响, 张松等^[6]以冬小麦籽粒为材料, 利用不同的预处理方法, 建立冬小麦籽粒蛋白质含量的无损检测模型, 模型精度高达 97.6%。洛曲等^[7]对藏区酥油脂肪和蛋白质含量采用 SNV、数据归一化、二阶导数、S-G 滤波法以及多种方式相组合对光谱进行预处理, 建立脂肪和蛋白质的定量模型具有良好的预测

收稿日期: 2019-06-02

基金项目: 国家自然科学基金 (41371395, 41601549)。

第一作者简介: 王翠秀 (1995-), 女, 硕士, 主要从事农业遥感研究。E-mail: 1632691999@qq.com。

通讯作者: 吴泉源 (1959-), 男, 博士, 教授, 主要从事农业遥感研究。E-mail: wqy6420582@163.com。

能力。为解决近红外光谱波峰宽,吸收重叠严重的问题,石岩等^[8]利用 CARS 成功提取人工牛黄特征波长变量,使得变量数量减少,模型评价参数更佳。吴建中^[9]利用 CARS 对模型进行了优化,准确测定不同酶水解样品抗氧化性能力。李路等^[10]利用 CARS 提取大米蛋白质、脂肪、总糖、含水量中近红外光谱的敏感波段,比较偏最小二乘法 and BP 神经网络法,取得了较好的效果。不同的回归模型能获得不同的模型预测精度,匡静云等^[11]应用反向传播神经网络和主成分分析建立原料乳中蛋白质与脂肪的预测模型,结果表明反向传播神经网络建立的检测模型效果最好。

以上研究表明,近红外光谱技术在食品检测方面效果理想,但是不同产品的理化特性各不相同,因此选择可以应用于大豆蛋白质以及脂肪含量的建模组合方式,能够最大程度的优化大豆近红外光谱检测模型,具有一定的实际应用价值。SG + 一阶微分可以有效的消除光谱白噪声,降低信噪比、SNV 通常与 DT 一起使用可以有效的消除光程差及表面散射的影响、OSC 可以剔除无用的光谱信息^[12]。偏最小二乘回归是在近红外光谱反演建模中最为常用的一类方法,模型精度高,适用性强^[13];BP 神经网络模型可以实现任何复杂非线性映射,自学习能力强^[14]。本研究将不同的预处理方法与建模方式组合,并运用竞争性自适应重加权变量提取特征波段。以期寻找大豆蛋白质、脂肪快速无损检测模型的最优模型组合方式。

1 材料与方法

1.1 样品获取与化验

1.1.1 大豆样本获取 研究材料为 2017 年山东省粮食与物资储备局黄大豆储存库中随机抽取的 191 份大豆样品。

1.1.2 大豆蛋白质、脂肪含量化学测定 大豆蛋白质含量化学测定方法为凯氏定氮法(GB5009.5 - 2010),使用 FOSS 公司的全自动凯氏定氮仪。准确称取 0.1 g 经过光谱测量并粉碎处理的待测样品,置于 50 mL 消煮管中,加 5 mL 硫酸加热消化,加入过氧化氢,等待样品溶液清亮,定容至 50 mL。吸取适当样品溶液,放入自动定氮仪测量含氮量。根据以下公式求出大豆蛋白质(soy protein content,SPC)含量:

$$SPC(\%) = c \times 6.25 \times 100 \quad (1)$$

式中: c 为所测得的氮含量($\text{mg} \cdot \text{L}^{-1}$);6.25 为氮含量换算成蛋白质的平均系数。

大豆脂肪的检测采用的是索氏抽提法(GB/

T5009.6 - 2003)。准确称取 0.3 g 经过光谱测量并粉碎处理的待测样品放在纸套筒内,用 FOSS 公司的 Soxtec Avanti 2055 脂肪检测仪侧定得到油脂含量。

1.2 光谱采集与处理

1.2.1 光谱采集 采用美国 ASD 公司 FieldSpec 3 光谱仪(测定范围为 350 ~ 2 500 nm)对完整大豆粒进行室内光谱测定。为了避免室外的不确定因素(光强、水汽等)影响原始光谱反射强度,光谱的采集过程选择在暗室内进行,其光源为 14.5 W 卤素灯。首先,将大豆样品中的杂质去除,将大豆样本均匀平铺于整个玻璃皿,并将表面压平,覆盖完整。光谱仪探头垂直朝下,视场角 25°,与样品间隔 20 cm。每次光谱测试之前进行白板校正,以保证采集到光谱的准确性。将每个大豆样品扫描 10 次,取其平均值作为大豆反射率,得到最终的大豆实际反射光谱数据。

1.2.2 异常样本剔除 光谱测量时不可避免的会受到人为读数误差和由光谱仪本身误差、样本前处理不当、温度和湿度等外界因素的干扰,引起光谱数据的误差^[15]。因此在建模所采用的样品中将异常值剔除可以大幅度提高模型的预测精度。本研究在 Unscrambler X 10.4 中用 PCA 对样本的光谱数据异常值进行判断,根据其得分共剔除 5 个异常值。

1.2.3 光谱变换 光谱测量过程中会受到噪声、样品背景和测量仪器等因素的影响,所以在进行建模分析之前对光谱数据预处理是必不可少的。本研究采用卷积平滑(SG) + 一阶微分、变量标准化(SNV) + 去趋势算法(DT)、正交信号校正(OSC),3 种光谱预处理方法对原始光谱数据进行优化。

1.3 竞争性自适应重加权采样

近红外光谱冗余信息较多,若采用全光谱建模,不仅计算工作量大,也会导致模型精度、稳定性降低。但是光谱波段过少同样会影响到最终模型的结果,如果能挑选出与目标组分近红外吸收谱图高相关的波段,过滤掉无关的冗余波段,能够有效提高模型准确度,改善模型质量。本研究采用竞争性自适应重加权采样法特征波段挑选方式。

竞争性自适应重加权采样法是一种模仿达尔文进化理论中的“适者生存”的原则,将蒙特卡罗采样与 PLSR 模型回归系数结合起来的特征变量选择方法。在 CARS 算法中,保留权重较大的点作为新的子集并去掉权重较小的点,然后利用新的子集建立 PLS 模型,通过多次计算,最终选择 PLS 模型交互验证均方根误差最小的子集中的波长作为特征波长。

CARS 算法的具体过程如下^[16-17]。

(1)利用蒙特卡罗采样法,在校正集中随机选择 80% 的样本作为建模集,其余 20% 进入预测集,建立 PLS 模型。提前设定好蒙特卡罗采样次数 (N),记录采样过程 PLS 模型中的回归系数绝对值权重:

$$w_i = |b_i| / \sum_{i=1}^m |b_i| \tag{2}$$

式中: $|b_i|$ 为第 i 个变量的回归系数绝对值; w_i 为第 i 个变量的回归系数绝对值权重; m 为每次采样中剩余的变量数。

(2)利用指数衰减函数 (exponentially decreasing function, EDF) 去掉绝对值权重较小的波长。在第 i 次基于 MC 采样建立 PLS 模型时,保留的波长点的比例 R_i 为:

$$R_i = \mu e^{-ki} \tag{3}$$

式中: μ 与 k 是常数,可以按照以下两种情况进行计算。①在一次采样并进行相应计算时,所有的波长都参与了建模分析,因此此时保留的波长点的比例为 1。②在最后一次采样 (第 N 次) 完成并进行相应计算时,只剩下两个波长参与 PLS 建模,此时保留的波长点的比例为 $2/n$,其中 n 是原始波长点数。由以上最初及最后一次采样的情况可知, μ 与 k 的计算公式为:

$$\mu = \left(\frac{n}{2}\right)^{\frac{1}{N-1}} \tag{4}$$

$$k = \frac{\ln \frac{n}{2}}{N-1} \tag{5}$$

(3)在每一次采样时,都从上一次采样时的变量数中采用自适应加权采样 (CARS) 选择数量为 $R_i \times n$ 个的波长变量,进行 PLS 建模,采用交叉验证计算 $RMSECV$ (root mean square error of cross validation)。

(4)在采样完成之后,得到了 N 组特征波长子集,以及对应的 $RMSECV$ 值,选择 $RMSECV$ 最小值所对应的波长变量子集为特征波长。

1.4 模型构建方法

偏最小二乘法 (PLS) 是光谱数据分析中常用的一种多元统计数据分析方法。PLS 算法在建模时对 X 和 Y 进行同时分解,并且同时考虑到了光谱信息和其对应的理化性质信息,以此来探究 X 和 Y 之间的线性关系。通过线性变换,将原始的数据转换为相互正交且不相关的新变量。新变量尽可能多的包含了原始数据的有效信息。它可以较好地解决普通多元线性回归面临的共线性问题^[18]。

反向传播神经网络 (BPNN) 是较为常见的神经网络模型,主要采用误差反向传播算法。在 Matlab 中进行网络结构的设计及运行 (图 1),3 层网络的拓扑结构由输入层、隐含层和输出层组成。通过对设定的训练样本部分进行建模,通过误差反向传播不断修正网络连接权值,使实际输出值与预测输出值之间的误差最小^[19-20]。

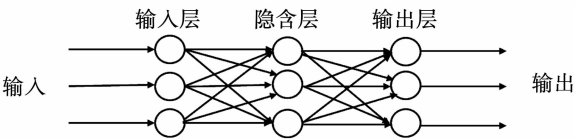


图 1 BPNN 的拓扑结构
Fig. 1 BPNN topology

1.5 模型精度评价

使用 Unscrambler X 10.4 对光谱数据进行预处理,使用 Matlab 2016a 对光谱进行特征波段提取以及模型建立。采用建模均方根误差 ($RMSE_C$)、预测均方根误差 ($RMSE_P$) 和决定系数 (R^2) 对模型进行评价。决定系数 (R^2) 值反映反射率和成分含量之间相关的密切程度, R^2 的取值范围为 0 ~ 1,越接近 1,表明模型的准确度越高; $RMSE_C$ 、 $RMSE_P$ 越小,模型精度越高、预测能力越强。

2 模型建立与验证

2.1 样本统计与分析

用 K-S (Kennard-Stone) 算法计算各样本间的欧氏距离,可以保证样本集按空间距离均匀分布,以 3:1 的比例选取 138 个大豆作为建模样本构建模型,47 个样本作为验证样本。从表 1 的统计特征看,建模集与预测集的变异系数差异均保持在 0.3% 以内,差异较小,表明样品挑选合理,且具有一定的代表性。

2.2 光谱分析

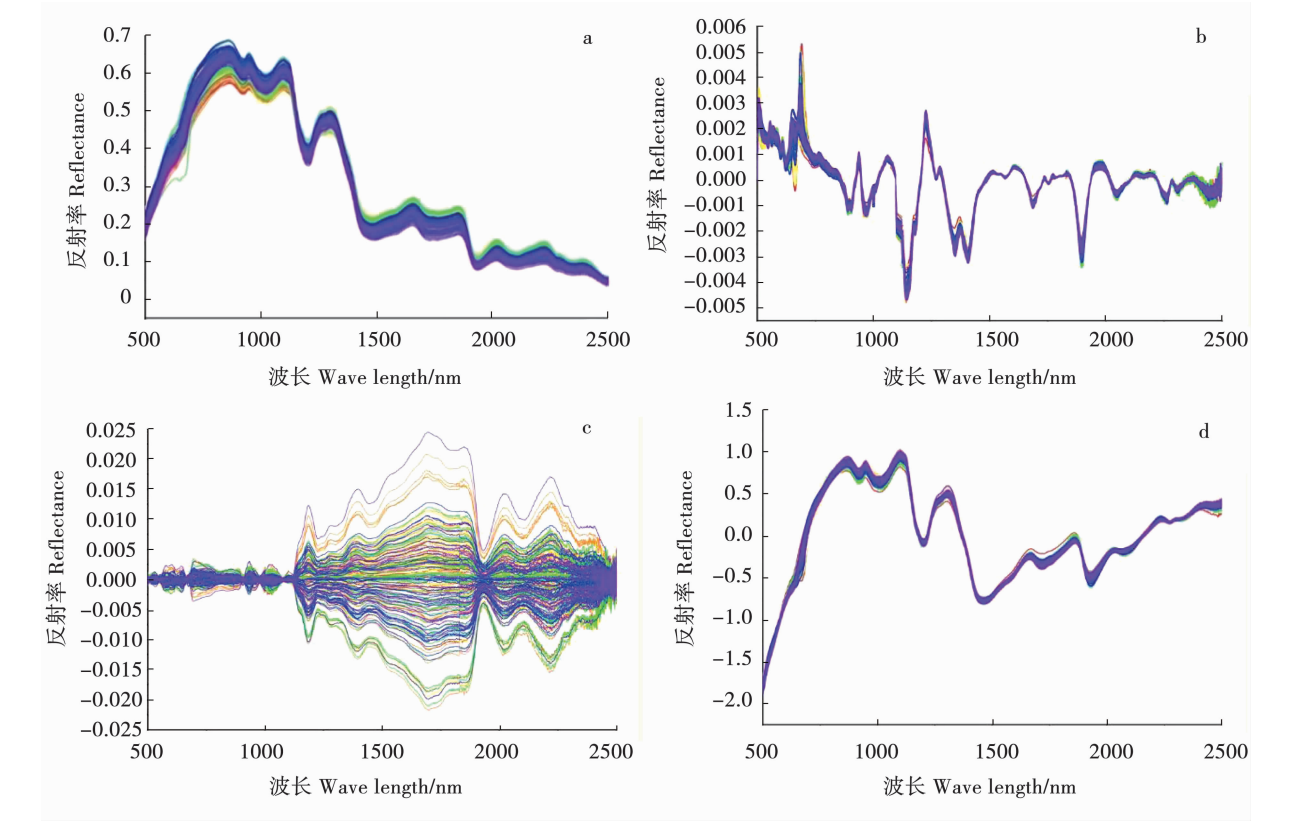
由于波段边缘会受到仪器噪声的影响,因此去除边缘波段 (350 ~ 499 nm),保留 500 ~ 2 500 nm 波段作为光谱数据进行处理。原始高光谱经变换后可以消除背景噪声,增强差别,突出光谱特征。1 500 nm 左右属于 O-H 基团伸缩振动的一级倍频和 N-H 基团振动的一级倍频吸收带,1 600 nm 左右反映了 O-H 基团伸缩振动的一级倍频吸收带,而在 1 700 nm 左右反映的是 C-H 和 O-H 基团伸缩振动的一级倍频吸收带^[21-25]。利用 3 种光谱预处理,可以对原始光谱数据优化、减弱或是消除背景噪声、增强差别、突出光谱特征。

表 1 大豆蛋白质、脂肪含量的描述性统计分析

Table 1 Descriptive statistical analysis of soybean protein and fat content

项目 Item	样品集 Sample set	样品数 Sample number	均值 Mean	最小值 Minimum /%	最大值 Maximum /%	标准偏差 Standard deviation	变异系数 Coefficient of variation/%
蛋白质 Protein	建模集 Calibration set	138	40. 85	36. 51	47. 96	2. 52	6. 21
	预测集 Prediction set	47	41. 25	36. 32	48. 12	2. 71	6. 56
脂肪 Fat	建模集 Calibration set	138	21. 44	17. 79	25. 02	1. 48	6. 93
	预测集 Prediction set	47	21. 47	17. 54	25. 09	1. 65	7. 20

由图 2 可知,原始光谱图(2a)各样品的谱线具有基本相同的变化趋势,但又各具差异,这主要是由于不同样本各成分含量的差异产生的;SG + 一阶微分(2b)方法可以对原始光谱进行平滑和求导,消除基线和其它背景的干扰,提高光谱分辨率以及灵敏度,还可有效的分辨重叠峰;正交信号校正处理(2c)可以筛选出有用信息,删除与因变量无关的信息,具备较强的特征提取能力;变量标准化 + 去趋势化分析(2d)的光谱形状总体上与原始光谱相似,但是曲线更加平滑,增强了光谱的反射特征。



a:原始光谱;b:卷积平滑 + 一阶微分;c:正交信号校正;d:变量标准化 + 去趋势。

a:Raw spectral reflectance;b:Savitzky and Golay + First order differential reflectance;c:Orthogonal signal correction reflectance;d:Standard normal variate + Detrending reflectance.

图 2 大豆原始及变换形式后的光谱反射率

Fig. 2 Spectral reflectance of soybean original and its transformed form

2.3 模型构建

2.3.1 全波段建模 将经过预处理之后的数据建立 PLS 与 BP 神经网络回归模型,以此计算大豆蛋白质、脂肪含量,其结果如表 2 所示。在对大豆蛋白质、脂肪的回归预测中,PLS 法与 BP 神经网络法均得到了较高的 R^2 、较低的 $RMSE_c$ 、 $RMSE_p$,说明化学测定值与近红外光谱检测值之间具有良好的线性关系,可用于实际的检测。建模结果显示,基于 OSC 变换的 PLSR 函数模型对蛋白质含量的建模结果最优($R_c^2=0.98$, $RMSE_c=0.35$),而基于 OSC 变换的 BP 神经网络函数模型对脂肪含量的建模结果最优

($R_c^2=0.99$, $RMSE_c=1.28$),神经网络模型所预测出的模型预测值与样本实测值之间存在极高的相关性,模型预测结果具有具有较好的拟合优度以及极高的可信度。模型验证结果表明,基于 OSC 变换的 PLSR 函数模型对蛋白质含量的反演精度最佳($R_p^2=0.96$, $RMSE_p=0.51$),与建模结果一致获得最高精度(图 3);而基于 SNV + DT 变换的 BP 神经网络函数模型则是对脂肪含量反演的最佳估算模型($R_p^2=0.93$, $RMSE_p=0.49$),与建模集所获得的最佳组合方式不同,说明模型不稳定。

表 2 不同校正与建模方法对大豆蛋白质、脂肪近红外检测模型的影响

Table 2 Effects of different calibration and modeling methods on soy protein and fat near infrared detection model							
项目 Item	建模方法 Modeling method	预处理方法 Pretreatment method	R_c^2	$RMSE_c$	R_p^2	$RMSE_p$	变量个数 Number of variables
蛋白质 Protein	PLS	SG + DI	0.92	0.69	0.89	0.97	1981
		OSC	0.98	0.35	0.96	0.51	1981
		SNV + DT	0.84	0.99	0.82	1.21	1981
	BP	SG + DI	0.98	1.18	0.90	1.18	1981
		OSC	0.96	1.28	0.84	1.21	1981
		SNV + DT	0.97	0.80	0.93	0.86	1981
脂肪 Fat	PLS	SG + DI	0.94	0.34	0.90	0.50	1981
		OSC	0.93	0.39	0.90	0.50	1981
		SNV + DT	0.95	0.34	0.84	0.64	1981
	BP	SG + DI	0.96	1.18	0.73	0.67	1981
		OSC	0.99	1.28	0.91	0.53	1981
		SNV + DT	0.98	0.80	0.93	0.49	1981

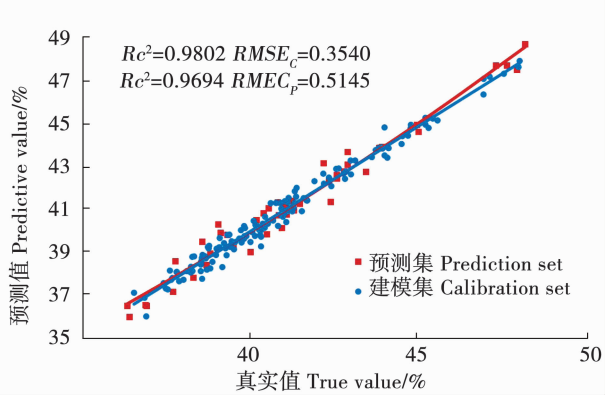


图 3 OSC + PLSR 大豆蛋白质模型和预测结果
Fig. 3 OSC + PLSR soybean protein model and prediction results

一般来说,建模集精度与预测集模型预测精度差异越小则说明模型的泛化能力越强。在对大豆蛋白质、脂肪的回归预测中,PLSR 函数模型比 BP 神经网络函数模型建模集精度与预测集精度相差小,因此,以 PLSR 函数模型为基础建立的模型具有较高的稳定性。

2.3.2 特征波段建模 图 4 为经过卷积平滑、一阶微分处理的大豆蛋白质、脂肪特征光谱的筛选结果。CARS 进行变量筛选时,蒙特卡罗的采样次数为 200 次。图 4a 为波段筛选过程中挑选出变量的变化趋势,变量数随着运行次数的增加而减少,两者间呈指数递减关系。图 4b 为波长变量筛选过程中,采用交叉验证得到的 $RMSECV$ 的变化趋势,

$RMSECV$ 值减小,说明光谱变量中无关变量的数量减少, $RMSECV$ 值增大,说明剔除了有效变量;在 1 ~ 140 次采样过程中 $RMSECV$ 呈现递减趋势,140 次后开始波动递增,在此时保留的有效波长变量数为 16 个。图 4c 中各线表示随着运行次数增加各波长变量回归系数的趋势,“*”所对应位置处的 RM -

$SECV$ 值最小,即第 140 次采样。遵循 $RMSECV$ 值最小原则,第 140 次采样获得的波长变量子集为最优结果。最终选择的波长变量数为 16 个(图 4a)。波长变量数由 1 981 个减少为 16 个,变量压缩率高达 99.19%,效果理想。

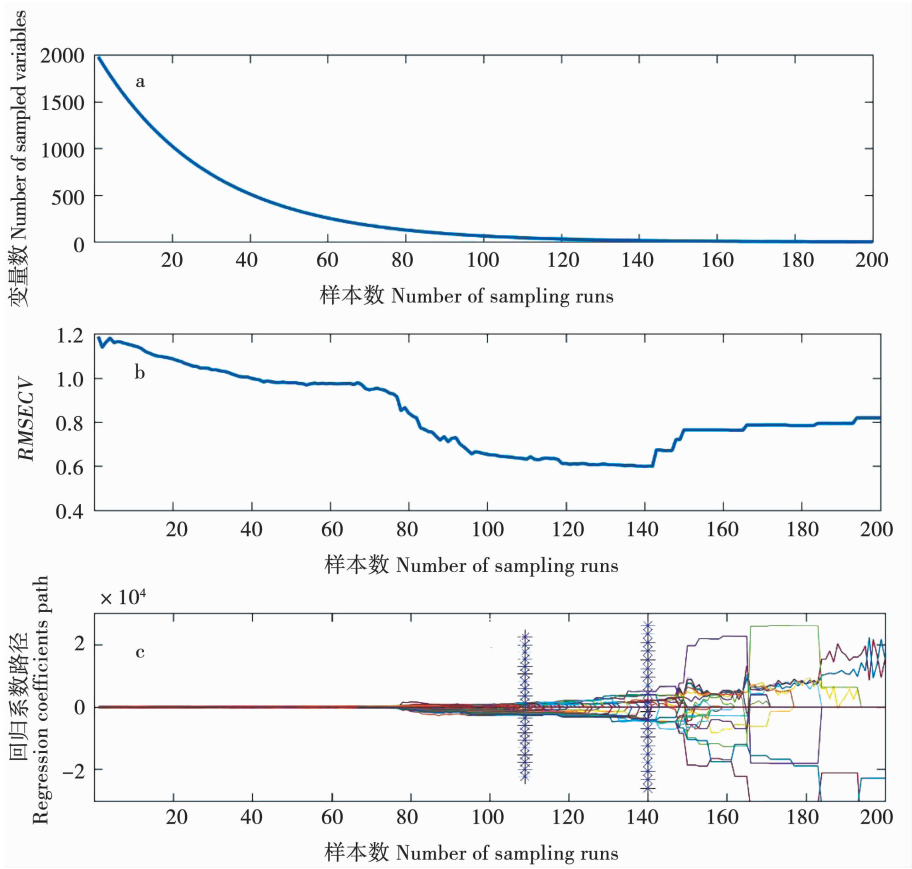


图 4 CARS 挑选特征波段

Fig. 4 CARS selected feature bands

利用 CARS 选取特征波长建立的回归模型结果如表 3 所示。可以看出,经过波长优选后,CARS 特征波段挑选后变量个数大大减少,模型稳定性也有所提高。但不同预处理方式下,CARS 特征波段挑选的结果不同。总体来说,经 CARS 优选后,波长变量数由 1 981 个减小到 100 个以下,变量压缩率大于 94.95%,大大减少了波长变量数。经过特征波段挑选过后,利用 BP 神经网络建模所获得模型稳定性最好,在面对不同的预处理方式时所获得模型精度相差很小。OSC + CARS + PLS 建模方式依旧是大豆蛋白质的最佳回归组合方式。($R^2_c = 0.98$, $RMSE_c = 0.29$, $R^2_p = 0.98$, $RMSE_p = 0.32$)。大豆脂肪建模稳定性得到提高,建模集与预测集的最优组

合方式都是 OSC + CARS + BP 神经网络建模方式 ($R^2_c = 0.99$, $RMSE_c = 0.46$, $R^2_p = 0.98$, $RMSE_p = 0.57$)。

对比表 2 和表 3 结果可知,经过 CARS 特征波段挑选后定标模型的建模集相关系数更高, $RMSE_c$ 、 $RMSE_p$ 更低,整体模型性能更优。SNV + DT + CARS + PLS 模型精度提高最大, R^2_c 由 0.84 提高到 0.94, R^2_p 由 0.82 提高到 0.93, $RMSE_c$ 由 0.99 降低到 0.59, $RMSE_p$ 由 1.21 降低到 0.72,模型复杂性降低,稳健性提高。由此说明,CARS 方法不仅可以有效筛选大豆蛋白质、脂肪的特征变量及相关影响变量,而且可以剔除冗余及噪声变量,从而有效降低基体元素的影响,提高定标模型的预测精度。

表3 CARS对大豆蛋白质、脂肪近红外检测模型的影响

Table 3 Effect of CARS on soy protein and fat near infrared detection model

项目 Item	建模方法 Modeling method	预处理方法 Pretreatment method	R_C^2	$RMSE_C$	R_p^2	$RMSE_p$	变量个数 Number of variables
蛋白质 Protein	CARS + PLS	SG + DI	0.95	0.55	0.94	0.70	16
		OSC	0.98	0.29	0.98	0.32	47
		SNV + DT	0.94	0.59	0.93	0.72	27
	CARS + BP	SG + DI	0.97	0.58	0.96	0.49	16
		OSC	0.98	0.46	0.98	0.57	47
		SNV + DT	0.97	0.62	0.95	0.53	27
脂肪 Fat	CARS + PLS	SG + DI	0.94	0.34	0.90	0.70	97
		OSC	0.95	0.32	0.96	0.30	62
		SNV + DT	0.96	0.29	0.96	0.30	82
	CARS + BP	SG + DI	0.97	0.58	0.96	0.24	97
		OSC	0.99	0.46	0.98	0.57	62
		SNV + DT	0.97	0.62	0.95	0.53	82

2.4 模型验证

采用仿真 sim 函数对模型进行精度验证,其结果如图 5、图 6 所示。模型验证结果表明,模型的预测结果具有同真实值相似的起伏趋势,可以较好地预测蛋白质和脂肪的含量。数据分析结果显示,BP 神经网络模型预测脂肪结果的均方根误差为 0.57,模型本身的拟合优度与预测结果均比较理想,精度达到 97.7%;蛋白质预测结果的均方根误差为 0.32,模型拟合精度达到 98.4%。对蛋白质和脂肪含量的预测值和真实值进行配对 t 检验,得到 P 值分别为 0.693 和 0.833,均大于显著性水平 0.05,说明预测值与真实值之间没有显著性差异,OSC + CARS + PLS 和 OSC + CARS + BP 神经网络建模对大豆蛋白质和脂肪具有良好的预测效果。

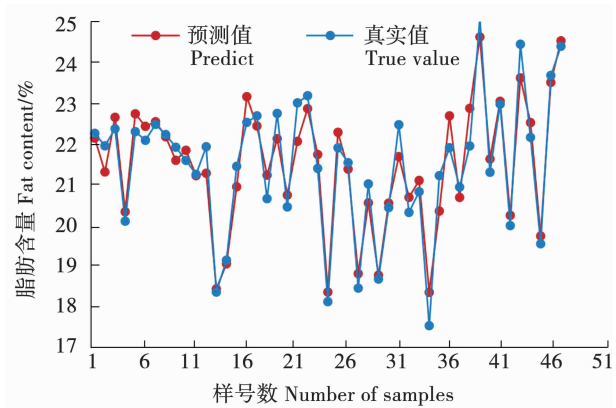


图5 预测集脂肪含量预测结果对比

Fig. 5 Comparison of prediction results of fat content in prediction set

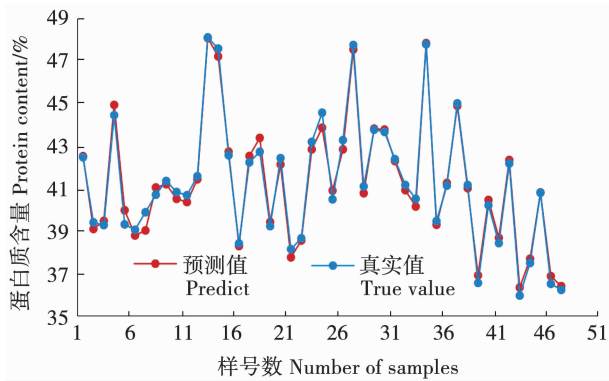


图6 预测集蛋白质含量预测结果对比

Fig. 6 Comparison of prediction results of protein content in prediction set

3 讨论

研究中所采用的研究近红外光谱法与化学检测的蛋白质、脂肪含量差异较小,可以确定近红外光谱法可以将大豆中的蛋白质与脂肪含量准确分析出来,这与朱贞映等^[26]、刘波等^[27]的研究结果一致。研究将大豆划分为建模的校正集和预测集,对大豆完整籽粒样品进行光谱扫描,可在很大程度上减少浪费、提高效率,进一步提高预测模型的精度以及适用度。

本研究在李琳琳等^[28]、王丽萍等^[29]研究的基础上进行了特征波段的挑选,压缩了变量,选用偏最小二乘回归(PLSR)、BP 神经网络 2 种预测模型,运用不同的预处理和竞争性自适应重加权挑选特征波段的方法下,变量压缩率大于 94.95%,筛选的特征波长变量作为输入变量的定量校正模型鲁棒

性强、预测能力更好;对比偏最小二乘回归(PLSR)、BP神经网络和2种不同的预测方法模型,OSC+CARS+PLS获得的建模集与预测集精度均属最高,均方根误差最小,建模方式是大豆蛋白质的最佳回归方式,OSC+CARS+BP神经网络建模方式是大豆脂肪含量检测的最佳回归方式。

研究中利用BP神经网络建模的均方根误差存在偏高现象,考虑是由于数据离散程度大、组分值动态范围大,今后需要对模型进行改进,使之适用于组分值范围宽广的数据;研究样品为山东省粮食储备局的大豆,若再补充不同大豆品种以及不同地区的大豆品种,所建模型将会更加可靠,适用范围将更加广泛。

4 结 论

本研究利用经过不同预处理方式的大豆近红外光谱建立蛋白质、脂肪含量的检测模型,分析了经过CARS特征波段挑选后与已有的模型相比的精度差距,比较了2类建模方法在大豆蛋白质、脂肪含量近红外检测中的适用性:

(1)CARS特征波段筛选可以降低无效变量,压缩率大于94.95%,在保证精度的同时降低了计算量。

(2)3种预处理方式与两种建模方式相组合,均能获得良好的检测效果。

(3)OSC+CARS+PLS建模方式是大豆蛋白质的最佳回归方式, $RMSE_c$ 和 R_c^2 分别为0.29和0.98, $RMSE_p$ 和 R_p^2 分别为0.32和0.98;OSC+CARS+BP神经网络建模方式是大豆脂肪含量检测的最佳回归方式, $RMSE_c$ 和 R_c^2 分别为0.46和0.99, $RMSE_p$ 和 R_p^2 分别为0.57和0.98。这两类组合方式降低了模型复杂度,提高了建模精度,适用于大豆蛋白质、脂肪的近红外快速检测。

参考文献

[1] Xiao C W , Wood C , Huang W , et al. Tissue-specific regulation of acetyl-CoA carboxylase gene expression by dietary soya protein isolate in rats [J]. British Journal of Nutrition, 2006, 95 (6):1048.

[2] Aoki H , Kimura K , Igarashi K , et al. Soy protein suppresses gene expression of acetyl-CoA carboxylase alpha from promoter PI in rat liver [J]. Bioscience, Biotechnology and Biochemistry, 2006, 70(4):843-849.

[3] 刘新旗,涂丛慧,张连慧,等.大豆蛋白的营养保健功能研究现状[J].食品科学学报,2012,30(2):1-6. (Liu X Q, Tu C H, Zhang L H, et al. Research on nutrition and health benefits of soy protein[J]. Journal of Food Science and Technology,2012,

30(2):1-6.)

[4] 张亚楠.大豆营养成分研究进展[J].现代农村科技,2017(10):64-65. (Zhang Y N. Research progress on nutritional components of soybean [J]. Modern Rural Science and Technology, 2017(10):64-65.)

[5] 赵影,王文和,滕娇琴,等.两种仪器测定国产大豆粗蛋白含量的比较[J].粮食储藏,2018(4):37-39,44. (Zhao Y, Wang W H, Teng J Q, et al. Comparison on determination of content of crude protein in domestic soybean by near-infrared grain analyzer and Kjeldahl nitrogen determinator [J]. Grain Storage, 2018(4):37-39,44.)

[6] 张松,冯美臣,杨武德,等.基于近红外光谱的冬小麦籽粒蛋白质含量检测[J].生态学杂志,2018,37(4):1276-1281. (Zhang S, Feng M C, Yang W D, et al. Detection of grain protein content in winter wheat based on near infrared spectroscopy [J]. Chinese Journal of Ecology,2018,37(4):1276-1281.)

[7] 洛曲,于修烛,张建新,等.基于近红外光谱的藏区酥油脂肪和蛋白质含量快速检测分析[J].中国油脂,2018,43(3):136-140. (Luo Q, Yu X Z, Zhang J X, et al. Rapid determination of fat and protein contents in ghee using near - infrared spectroscopy [J]. China Oils and Fats,2018,43(3):136-140.)

[8] 石岩,孙冬梅,熊婧,等.近红外光谱结合竞争性自适应重加权采样算法用于人工牛黄的质量分析研究[J].中国药学杂志,2018,53(14):1216-1221. (Shi Y, Sun D M. Xiong J, et al. A-nalysis of artificial cow-bezoar by near-infrared spectroscopy coupled with competitive adaptive reweighted sampling method [J]. Chinese Pharmaceutical Journal,2018,53(14):1216-1221.)

[9] 吴建中.大豆蛋白的酶法水解及产物抗氧化活性的研究[D].广东:华南理工大学,2003. (Wu J Z. Study on the enzymatic hydrolysis of soy protein and antioxidative activity of its hydrolysate [D]. Guangdong:South China University of Technology,2003.)

[10] 李路,黄汉英,赵思明,等.大米蛋白质、脂肪、总糖、水分近红外检测模型研究[J].中国粮油学报,2017,32(7):121-126. (Li L, Huang H Y, Zhao S M, et al. NIR spectra selection model of protein, fat, total sugar and moisture in rice [J]. Journal of the Chinese Cereals and Oils Association,2017,32(7):121-126.)

[11] 匡静云,管骁,刘静.原料乳中蛋白质与脂肪的近红外光谱快速定量研究[J].分析科学学报,2015,31(6):783-786. (Kuang J Y, Guan X, Liu J. Rapid determination of protein and fat contents in raw milk by near infrared spectroscopy analysis [J]. Journal of Analytical Science,2015,31(6):783-786.)

[12] 王骏超,葛俊锋. NIRS 数据样本选择与预处理方法综述 [J]. 国外电子测量技术,2019,38(3):1-7. (Wang J C, Ge J F. Overview of NIRS data sample selection and pretreatment methods [J]. Foreign Electronic Measurement Technology,2019,38(3):1-7.)

[13] 田永超,张娟娟,姚霞,等.基于近红外光谱的土壤有机质含量定量建模方法[J].农业工程学报,2012,28(1):145-152. (Tian Y C, Zhang J J, Yao X. et al. Quantitative modeling method of soil organic matter content based on near-infrared photoacoustic spectroscopy [J]. Transactions of the Chinese Society of Agricultural Engineering,2012,28(1):145-152.)

[14] 毕京翠,张文伟,肖应辉.等.应用近红外光谱技术分析稻米蛋白质含量[J].作物学报,2006(5):709-715. (Bing J C, Zhang W W, Xiao Y H. Analysis for protein content in rice by near infrared reflectance spectroscopy (NIRS) technique [J]. Acta Agro-

nomica Sinica,2006(5):709-715.)

[15] Wu S N,Liu C L,Wu J Z. et al. Outlier sample analysis on near infrared spectroscopy determination for flour ash[C]//Fifth International Conference on Measuring Technology & Mechatronics Automation. IEEE Computor Society,2013.

[16] Liang R,Bao Z, Su B. et al. Feasibility of ionic liquids as extractants for selective separation of vitamin D-3 and tachysterol(3) by Solvent extraction[J]. Journal of Agricultural and Food Chemistry, 2013, 61(14):3479-3487.

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

[18] 成忠. PLSR 用于化学化工建模的几个关键问题的研究[D]. 杭州:浙江大学,2005. (Research on several key technologies of partial least squares regression in chemistry and chemical process modeling[D]. Hangzhou: Zhejiang University, 2005.)

[19] Bishop T F A ,Mcbratney A B . Creating field extent digital elevation models for precision agriculture[J]. Precision Agriculture, 2002, 3(1):37-46.

[20] 章德宾,徐家鹏,许建军,等. 基于监测数据 BP 神经网络的食品安全预警模型[J]. 农业工程学报,2010,26(1):221-226. (Zhang D B,Xu J P,Xu J J, et al. Model for food safety warning based on inspection data and BP neural network[J]. Transactions of the Chinese Society of Agricultural Engineering,2010,26(1):221-226.)

[21] Chang K C . A neural network approach to geographical analysis of population pattern change [J]. Ann Arbor Michigan University Microfilms International, 1994, 18(95):657-662.

[22] Burns D A,Ciurczak E W. Handbook of near-infrared analysis [M]. Newyork: Crc Press, 2001.

[23] 杰尔·沃克曼,洛伊斯·文依. 近红外光谱解析实用指南 [M]. 褚小应,许有鹏,田高友,译. 北京:化学工业出版社,2009. (Workman J, Weyer J. Practial guide to interpretive near-infrared spectroscopy[M]. Zhu X L, Xu Y P, Tian G Y, transtate. Beijing:Chemical Industry Press,2009.)

[24] 李栓明,郭银巧,王克如,等. 小麦籽粒蛋白质光谱特征变量筛选方法研究[J]. 中国农业科学,2015,48(12):2317-2326. (Li S M,Guo Y Q, Wang K R, et al. Researchon variable selection of wheat kernel protein content with near-infrared spectroscopy[J]. Scientia Agricultura Sinica,2015,48(12):2317-2326.)

[25] 江艳,武培怡. 大豆蛋白的中红外和近红外光谱研究[J]. 化学进展,2009, 21(4):705-714. (Jiang Y, Wu P Y. Study of soy protein by mid-infrared spectroscopy and near-infrared spectroscopy[J]. Progress in Chemistry, 2009, 21(4):705-714.)

[26] 朱贞映,袁建,鞠兴荣,等. 傅立叶变换近红外光谱在大豆蛋白质和粗脂肪检测中的研究[J]. 食品安全质量检测学报,2015(12):4924-4931. (Zhu Y Z,YuanJ, Ju X R, et al. Detection of soybean protein and crude fat by Fourier transform-near infrared spectroscopy[J]. Journal of Food Stafety and Qualiyy, 2015(12):4924-4931.)

[27] 刘波,张丽娟,苗保河,等. 近红外光谱法与国标法测定大豆蛋白质和脂肪的比较[J]. 山东农业科学,2007(1):109-111. (Liu B,Zhang L J,Miao B H,et al. Near infrared spectroscopy compared with national standard method was developed for the determination of soybean protein and fat[J]. Shandong Agricultural Sciences,2007(1):109-111.)

[28] 李琳琳,金华丽,崔彬彬,等. 基于近红外透射光谱的大豆蛋白质和粗脂肪含量快速检测[J]. 粮食与油脂,2014,27(12):57-60. (Li L L,Jin H L,Cui B B,et al. Rapid determination of soybean protein and crude fat content by near-infrared transmittance spectroscopy[J]. Cereals and Oils,2014,27(12):57-60.)

[29] 王丽萍,陈文杰,赵兴忠,等. 基于近红外漫反射光谱法的大豆粗蛋白和粗脂肪含量的快速检测[J]. 大豆科学,2019,38(2):280-285. (Wang L P,Chen W J,Zhao X Z,et al. Rapid determination of crude protein and crude oil content of soybean based on near infrared diffuse reflectance spectroscopy[J]. Soybean Science,2019,38(2):280-285.)