

基于近红外光谱的大豆水分和粗脂肪含量的快速检测

郭东升, 张志勇*, 武志明, 席 前, 袁 凯, 伍莹芮, 何国康

(山西农业大学农业工程学院, 太谷 030801)

摘要: 目的 基于近红外光谱(near infrared spectroscopy, NIRS)技术快速检测大豆中水分和粗脂肪含量。
方法 首先采集 350~2500 nm 光谱范围的大豆近红外光谱, 采用光谱-理化值共生距离(sample set partitioning based on joint x-y distance, SPXY)算法将大豆样本划分为校正集样本与测试集样本, 然后对原始光谱分别采用多元散射校正、标准正态变量交换、归一化等 9 种方法进行预处理, 最后使用偏最小二乘回归分析方法建立模型对样本进行定量分析。**结果** 原始光谱经过多元散射校正后建立的偏最小二乘回归模型对水分的预测精度最高, 其校正集和测试集的相关系数分别为 0.8964 和 0.9055, 均方根误差分别为 0.4211 和 0.5933; 原始光谱经过归一化处理后建立的偏最小二乘回归模型对粗脂肪的预测精度最高, 其校正集和测试集的相关系数分别为 0.9084 和 0.9295, 均方根误差分别为 0.6897 和 0.6462。**结论** 近红外光谱结合预处理及偏最小二乘回归法, 可以快速、准确的检测大豆水分和粗脂肪含量。

关键词: 大豆; 水分; 粗脂肪; 近红外光谱; 偏最小二乘回归

Rapid detection of moisture and crude fat content in soybean based on near infrared spectroscopy

GUO Dong-Sheng, ZHANG Zhi-Yong*, WU Zhi-Ming, XI Qian, YUAN Kai,
WU Ying-Rui, HE Guo-Kang

(College of Engineering, Shanxi Agricultural University, Taigu 030801, China)

ABSTRACT: Objective To detect moisture and crude fat content in soybean rapidly based on near infrared spectroscopy(NIRS). **Methods** Firstly, the near infrared spectrum of soybean in the spectral range of 350-2500 nm was collected, and the soybean samples were divided into calibration set samples and test set samples by using the sample set partitioning based on joint x-y distance(SPXY) algorithm, and then the original spectrum was pretreated by nine methods, such as multiplicative scatter correction, standard normal variable exchange, and normalization, etc. Finally, the partial least square regression analysis method was used to establish a model for quantitative analysis of samples. **Results** The partial least square regression model established after the original spectrum was corrected by multiple scattering has the highest prediction accuracy for water content, and the correlation coefficients of the correction set and the test set are 0.8964 and 0.9055 respectively. The root mean square error were 0.4211 and 0.5933

基金项目: 山西省自然科学基金项目(201701D121103)、国家重点研发计划项目(2016YFD0701801)

Fund: Supported by Shanxi Natural Science Foundation (201701D121103), National Key R & D Program (2016YFD01801)

***通讯作者:** 张志勇, 博士, 副教授, 硕士生导师, 主要研究方向为农产品检测技术。E-mail: zzyzy1979@163.com

***Corresponding author:** ZHANG Zhi-Yong, Ph.D, Associated Professor, Shanxi Agricultural University, Taigu 0308001, China. E-mail: zzyzy1979@163.com

respectively. The partial least square regression model established after the original spectrum was corrected by normalization has the highest prediction accuracy for crude fat, and the correlation coefficients of the correction set and the test set were 0.9084 and 0.9295 respectively, the root mean square error were 0.6897 and 0.6462 respectively.

Conclusion NIRS combined with pretreatment and partial least square regression can detect the content of water and crude fat in soybean rapidly and accurately.

KEY WORDS: soybean; water; crude fat; near infrared spectroscopy; partial least square regression

1 引言

大豆作为我国主要的经济农作物,具有分布范围广、营养价值高、市场需求量大等特点。大豆中水分与粗脂肪含量的不同决定了其能够储存时间的长短以及收购价格的差异,其含量是评价大豆品质的重要指标,因此在大豆的收购及存储过程中快速检测其成分格外重要^[1]。传统的化学检测方法精确度高,但其过程繁琐、耗时长、成本高^[3]。近红外光谱法依据被测物质所含 C-H、O-H、N-H 基团的振动信息来进行成分检测,具有采样速度快、操作简单的特点,已成为农产品无损检测的重要手段之一^[4]。但光谱数据易受到杂散光、噪声、基线漂移等因素的干扰,从而影响建模效果。因此,建模前对光谱数据进行预处理就显得格外重要。Qiao 等^[6]采用了 6 种不同的预处理方法对原始光谱进行处理,结合偏最小二乘回归方法建立了土壤有机质预测模型;张新玉等^[7]对原始光谱依次进行了平滑、一阶差分、标准化等预处理,分别采用偏最小二乘回归法和支持向量机回归法建立了玉米单籽粒蛋白质含量检测模型,所建立的模型均取得了较好的效果;郝勇等^[8]采用 Norris-Williams 平滑求导和多元散射校正方法对光谱进行预处理,建立了饲料中粗脂肪和粗纤维含量快速检测模型,结果表明经过多元散射校正预处理后的光谱模型优于其他预处理方法的光谱模型。

本文利用近红外光谱实现对大豆水分含量及粗脂肪含量的检测,在利用建立近红外光谱检测模型前,采用不同的光谱预处理方法对原始光谱进行处理,通过比较经过不同预处理后的近红外光谱建立的偏最小二乘回归(partial least square regression, PLSR)模型,建立最优的光谱检测模型。以期更好地应用于大豆水分及粗脂肪含量检测。

2 材料与方法

2.1 材料与仪器

大豆样本采集于全国不同省份不同地区。

FiledSpec3 光谱仪(美国 ASD 公司); JJ-2 型高速组织捣碎匀浆机(江苏省金坛市荣华仪器制造有限公司)。

2.2 实验方法

2.2.1 样本获取与化验

选取来自全国不同省份不同地区的大豆样本共 64 个,利用光合高速组织捣碎匀浆机将大豆粉碎,过 80 目标准分样筛,制成粉末状样本。大豆的水分含量与粗脂肪含量按照国家标准 GB5009.3-2016《食品安全国家标准食品中水分的测定》^[9]和 GB5009.5-2016《食品安全国家标准食品中蛋白质的测定》^[10]测定。

2.2.2 近红外光谱的采集

使用光谱仪对大豆粉末进行光谱测定,其波长范围为 350~2500 nm。将大豆粉末均匀平铺于培养皿,逐层压实,最后用直尺将表面刮平。为了保证测量结果的准确性,光谱仪开机后进行预热,每次光谱测试前进行白板校正,并且整个测量过程在室内封闭的暗箱内进行,对每个样本的 9 个不同位置进行光谱采集,每个位置采集 2 次光谱,取平均值作为该样本的原始光谱数据。

2.2.3 近红外光谱预处理及样本划分

常见的光谱预处理方法包括很多种,根据预处理的效果可以分为基线校正、散射校正、平滑处理和尺度缩放 4 类,每一类又包括多种方法。基线校正包括一阶导数(first derivative, FD)、二阶导数(second derivative, SD)、连续小波变换(continuous wavelet transform, CWT)等;散射校正包括多元散射校正(multiplicative scatter correction, MSC)、标准正态变量(standard normal variate, SNV)等;平滑处理包括(savitzky-golay, SG)卷积平滑法、移动窗口平滑等;尺度缩放包括数据中心化(mean centering)、Pareto 尺度化、归一化(normalization, Nor)、标准化(autoscales)等^[11]。本文为了探究针对大豆水分及粗脂肪含量检测的最佳预处理方法,选取涵盖以上 4 大类共 9 种预处理方法,包括:一阶导数、二阶导数、多元散射校正、标准正态变量、SG 卷积平滑、移动窗口平滑、中心化、归一化、标准化。

光谱-理化值共生距离(sample set partitioning based on joint x-y distance, SPXY)算法在样品集的划分中应用广泛,优于随机算法(random sampling, RS), Kennard-Stone(KS)算法,双向算法(duplex)等^[14]。在建立大豆水分和粗脂肪含量定量识别模型时,采用光谱-理化值共生距离算法分别以水分含量与粗脂肪含量为 Y 变量,近红外光谱值为 X 变量,

为使得样本集划分合理即校正集样本含量范围涵盖测试集样本含量范围,按照 3:1 比例对 64 个样本光谱进行划分,校正集有 48 个样本,测试集有 16 个样本^[15]。

2.2.4 模型构建方法

偏最小二乘法回归(partial least squares regression, PLSR)分析方法是光谱数据分析中常用的一种多元统计数据分析方法,该方法在建模时对光谱数据矩阵 X 和被测成分矩阵 Y 进行同时分解,综合考虑了光谱信息和其对应的理化性质信息,使从两者中提取的信息相关性达到最大,然后用所获得的成分建立多元回归分析模型^[16]。偏最小二乘回归分析方法集主成分分析、典型相关分析和多元线性回归分析 3 种分析方法的优点于一身,能够更好地建立检测模型。

2.2.5 模型的评价

光谱经过预处理后,利用校正集样本建立大豆水分与粗脂肪含量检测模型,再利用测试集样本对模型进行验证,确定所建立模型是否可以有效实现对大豆中水分与粗脂肪含量的分析。采用校正集相关系数 R_c 、测试集相关系数 R_p 、交叉验证均方根误差(RMSECV)、测试集均方根误差(RMSEP)对模型进行评价。相关系数越高、均方根误差越低,则 PLS 模型的预测效果越好^[19]。

3 结果与分析

3.1 光谱样本划分结果

SPXY 算法划分后的大豆水分含量模型校正集和测试集统计信息如表 1 所示。

SPXY 算法划分后的大豆粗脂肪含量模型校正集和测试集统计信息如表 2 所示。

从以上表中可以看出,校正集样本的成分含量涵

盖了测试集样本的成分含量范围,且变异系数均处于合理范围,样品集划分合理,有利于建立更加稳定、可靠的模型。

3.2 不同预处理方法的模型结果

本研究采用多元散射校正、标准正态变量、移动窗口平滑、Savitzky-Golay 卷积平滑法、归一化(normalize)、数据中心化(mean centering)、标准化(autoscales)、一阶导数、二阶导数共 9 种预处理方法对原始光谱进行处理。如表 3 所示为不同预处理方法的大豆水分和粗脂肪含量模型结果。

通过表 3 可以看出,相比于未经预处理的原始光谱构建的 PLSR 模型,经过移动窗口平滑、SG 平滑、数据中心化、一阶导数、二阶导数这几种方法处理后的大豆水分模型精度较差,而 MSC、SNV、Nor 这 3 种方法对水分模型的精度有明显提升,其中 MSC 预处理的水分模型精度最好,校正集的 RMSECV 由 0.5393 下降到 0.4211, R_c 由 0.8037 提高到 0.8964,测试集的 RMSEP 由 0.7194 下降到 0.5933, R_p 由 0.8258 提高到 0.9055。图 1 为大豆的原始近红外光谱,图 2(a)为经过 MSC 处理后的光谱图,可见处理后修正了光谱间的相对基线漂移,使得光谱更加紧密。经过 Nor、标准化、数据中心化这几种方法处理后的大豆粗脂肪模型相比原始光谱构建的模型有了明显的改善,其中 Nor 这种方法建立粗脂肪模型稳定性更好,校正集的 RMSECV 由 1.0818 下降到 0.6897, R_c 由 0.8072 提高到 0.9084,测试集的 RMSEP 由 1.4607 下降到 0.6462, R_p 由 0.7642 提高到 0.9295。图 2(b)为经过 Nor 处理后的光谱,光谱都落入了一个特定的区间之内,去除了尺寸差异。

表 1 大豆水分含量模型校正集和测试集统计信息表(%)

Table 1 Statistical information table of correction set and prediction set of soybean water content model (%)

数据集	样本数	最小值	最大值	平均值	标准偏差	变异系数
校正集	48	8.46	14.70	10.68	0.91	8.52
测试集	16	9.29	13.04	10.52	0.89	8.49
总计	64	8.46	14.70	10.64	0.92	8.65

表 2 大豆粗脂肪含量模型校正集和测试集统计信息表(%)

Table 2 Statistical information table of correction set and test set of soybean crude fat content model (%)

数据集	样本数	最小值	最大值	平均值	标准偏差	变异系数
校正集	48	14.60	23.15	19.28	0.86	4.47
测试集	16	17.39	20.94	19.56	0.87	4.46
总计	64	14.60	23.15	19.35	0.89	4.60

表 3 不同预处理方法的大豆水分和粗脂肪含量模型结果
Table 3 Results of soybean water content and crude fat content model with different pretreatment methods

预处理方法	水分				粗脂肪			
	校正集		测试集		校正集		测试集	
	R_c	RMSECV	R_p	RMSEP	R_c	RMSECV	R_p	RMSEP
原始光谱	0.8037	0.5393	0.8258	0.7194	0.8072	1.0818	0.7642	1.4607
MSC	0.8964	0.4211	0.9055	0.5933	0.9004	0.5685	0.7914	1.0450
SNV	0.8850	0.4305	0.8649	0.6628	0.9007	0.5677	0.7948	1.0382
移动窗口平滑	0.8165	0.5381	0.7985	0.7781	0.7972	0.5472	0.8399	0.7055
SG 平滑	0.8149	0.5389	0.7230	0.8982	0.4612	1.0792	0.5790	1.4480
Nor	0.8936	0.5188	0.8452	0.7047	0.9084	0.6897	0.9295	0.6462
数据中心化	0.8014	0.5407	0.8246	0.7238	0.8326	0.5044	0.8023	0.7248
标准化	0.8742	0.5072	0.8374	0.7105	0.8968	0.4203	0.8990	0.6262
一阶导数	0.9999	0.0156	0.7915	0.8007	0.9963	0.1309	0.9101	0.6924
二阶导数	0.9983	0.0701	0.5475	1.0726	0.9948	0.1283	0.6361	1.3929

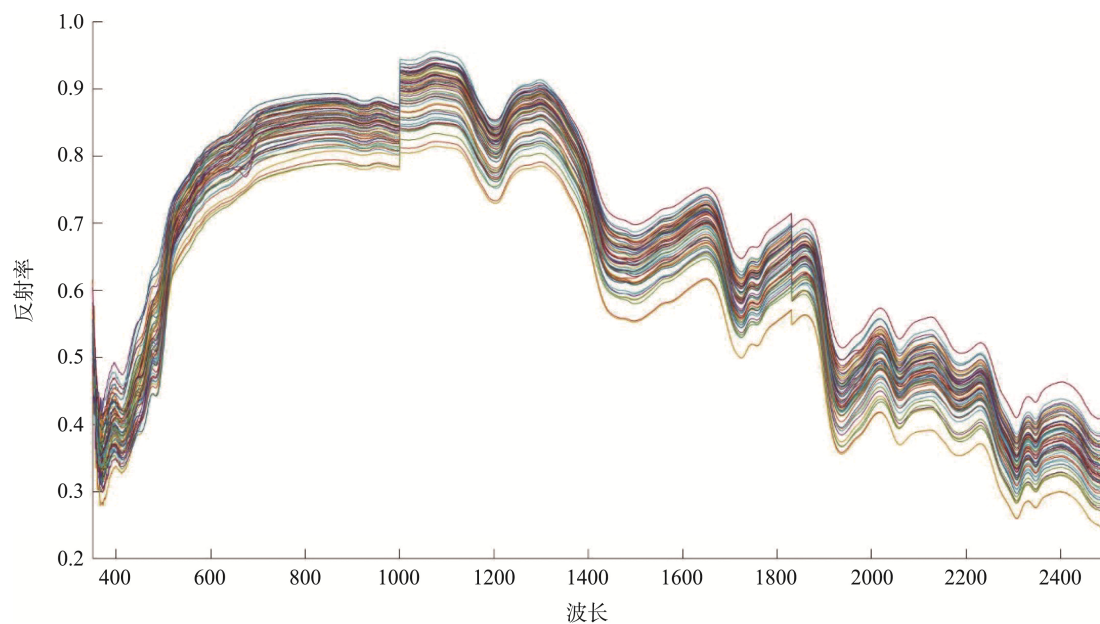
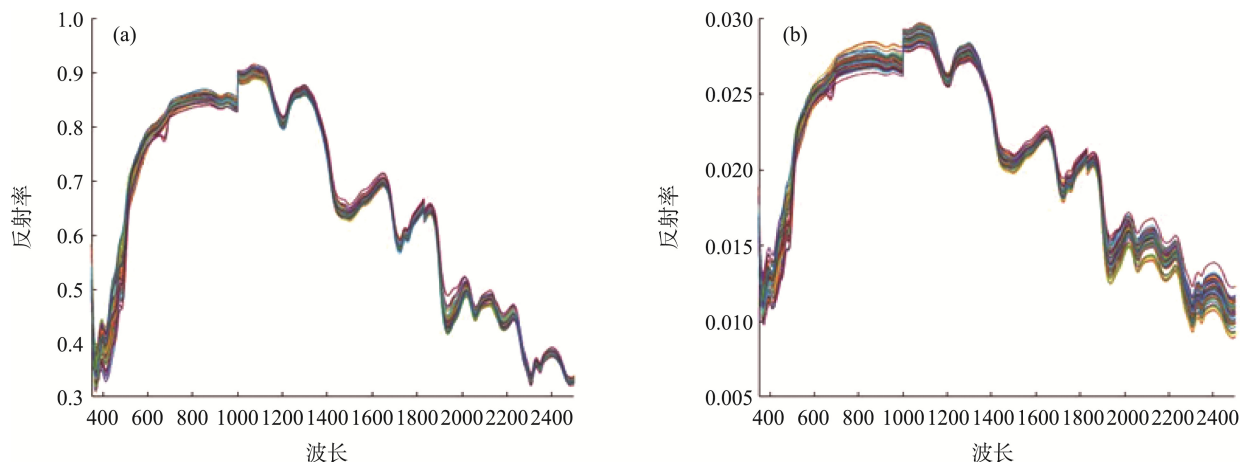


图 1 大豆的原始近红外光谱
 Fig.1 Original NIRS of soybean



注: (a)经过 MSC 预处理后的光谱图; (b)经过 Nor 预处理后的光谱图。

图 2 经过 MSC 预处理后及 Nor 预处理后的光谱图

Fig.2 Spectra after MSC pretreatment and Nor pretreatment

3.3 模型分析

图 3 为利用经过 MSC 预处理后的光谱建立的 PLSR 模型相关图, 图 3(a)为校正集预测结果, 图 3(b)为测试集预测结果。

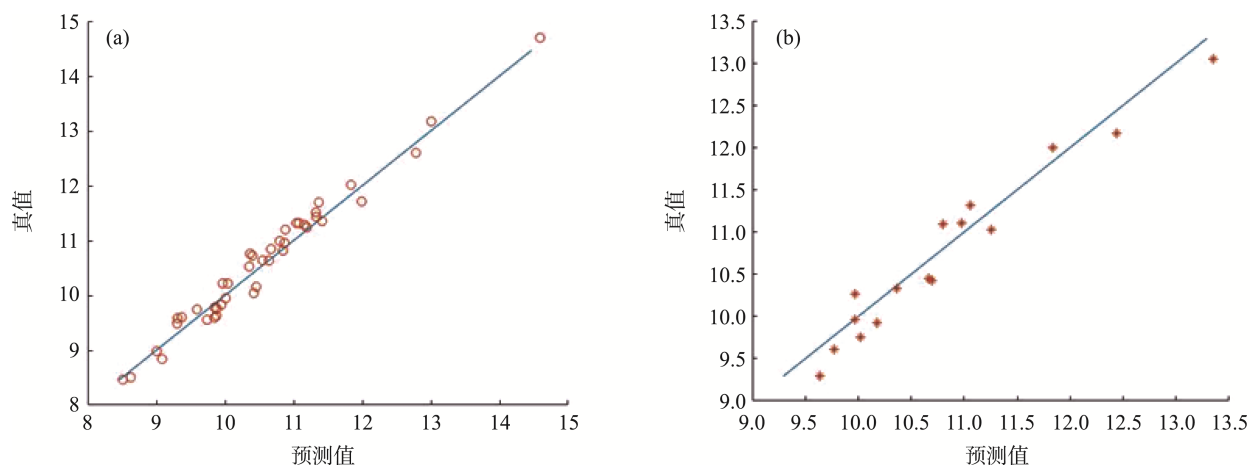
利用经过 Nor 预处理后的光谱建立的 PLSR 模型相关图见图 4, 图 4(a)为校正集预测结果, 图 4(b)为测试集预测结果。

从图 3 和图 4 的散点图可以得出, 经过 MSC 预处理的光谱建立的大豆水分含量检测模型与经过 Nor 预处理的光谱建立的大豆粗脂肪含量检测模型的相关性都比较好, 预测精度较高。

4 结论

本文将近红外光谱分析方法用于大豆中水分和粗

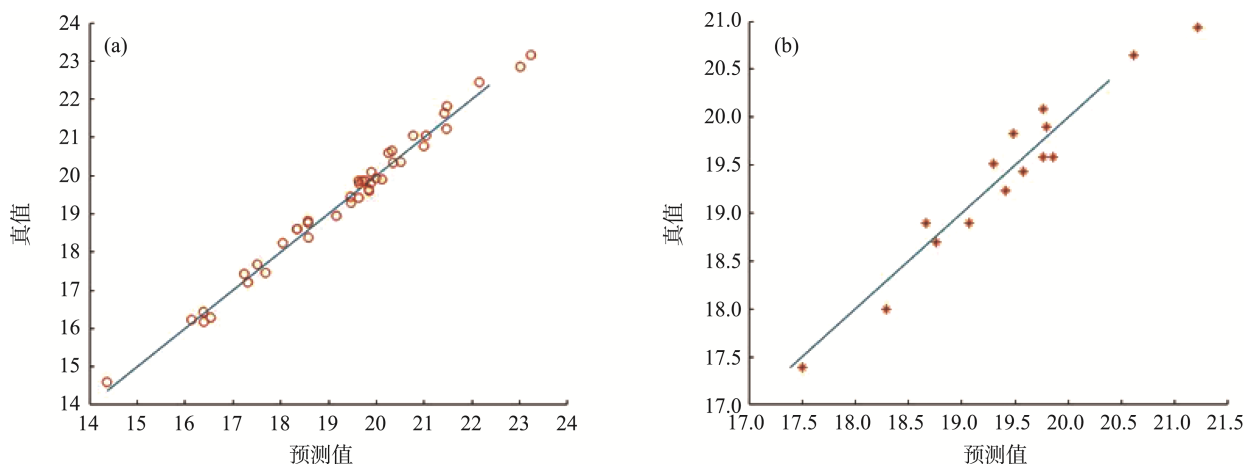
脂肪含量的定量分析, 采用不同的预处理方法对光谱模型进行优化。结果表明, 在对大豆水分定量分析中, MSC 预处理方法可以建立最优水分检测模型, 模型的 R_p 和 R_c 分别为 0.9055 和 0.8964, RMSEP 和 RMSECV 分别为 0.5933 和 0.4211; 在对大豆粗脂肪定量分析中, 归一化预处理方法可以建立最优粗脂肪含量检测模型, 模型的 R_p 和 R_c 分别为 0.9295 和 0.9084, RMSEP 和 RMSECV 分别为 0.6462 和 0.6897。因此, 基于近红外光谱建立的 MSC-PLSR 水分模型和 Nor-PLSR 粗脂肪模型可以有效的进行大豆中水分和粗脂肪含量的定量分析。通过本文的研究, 找到了针对建立大豆水分及粗脂肪含量快速检测模型的最佳预处理方法, 较以往采用通用的预处理方法建立的模型更具有针对性, 模型更加适合应用于大豆水分及粗脂肪含量检测。



注: (a)水分校正集的预测值与真值的关系(b)水分测试集的预测值与真值的关系。

图 3 多元散射校正后的大豆水分含量分析 PLSR 模型校正集和测试集相关图

Fig.3 Correlation diagram of PLSR model correction set and test set of soybean moisture content analysis after multiplicative scatter correction



注: (a)脂肪校正集的预测值与真值的关系; (b)脂肪测试集的预测值与真值的关系。

图 4 归一化(Nor)后的大豆粗脂肪含量分析 PLSR 模型校正集和测试集相关图

Fig.4 Correlation diagram of PLSR model correction set and test set of soybean crude fat content analysis after normalization (Nor)

参考文献

- [1] 褚丽琼. 储存条件对大豆品质变化的影响[J]. 现代食品, 2017, 11: 76-78.
Chu LQ. Effects of storage conditions on soybean quality [J]. Mod Food, 2017, 11: 76-78.
- [2] 肖青青. 近红外光谱用于玉米水分和粗脂肪的快速分析[D]. 广州: 暨南大学, 2015.
Xiao QQ. NIR spectroscopy applied in rapid analysis of moisture and fat in corn [D]. Guangzhou: Jinan University, 2015.
- [3] 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 LP, Chen WJ, Zhao XZ, et al. Rapid determination of crude protein and crude oil content of soybean based on near infrared diffuserelectance spectroscopy [J]. Soybean Sci, 2019, 38(2): 280-285.
- [4] 王多加, 周向阳, 金同铭, 等. 近红外检测技术在农业和食品分析上的应用[J]. 光谱学与光谱分析, 2004, 24(4): 447-450.
Wang DJ, Zhou XY, Jin TM, et al. Application of near-infrared Spectroscopy to agriculture and food analysis [J]. Spectrosc Spect Anal, 2004, 24(4): 447-450.
- [5] 谢玉荣, 李强, 王娇. 红外光谱技术在食品检测中的应用[J]. 食品安全质量检测学报, 2019, 10(22): 7773-7778.
Xie YR, Li Q, Wang J. Application of infrared reflectance spectroscopy on food detection [J]. J Food Saf Qual, 2019, 10(22): 7773-7778.
- [6] Qiao XX, Wang C, Feng MC, et al. Hyperspectral estimation of soil organic matter based on different spectral preprocessing techniques [J]. Spectrosc Letters, 2017, 50(3): 156-163.
- [7] 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.
Zhang XY, Wang YJ, Liu RX, et al. application of near-infrared spectroscopy technology to analyze protein content in single kernel maize seed [J]. J Chin Agric Univ, 2017, 22(5): 25-31.
- [8] 郝勇, 吴文辉, 商庆园. 饲料中粗脂肪和粗纤维含量的近红外光谱快速分析[J]. 光谱学与光谱分析, 2020, 40(1): 215-220.
Hao Y, Wu WH, Shang QY, et al. The research on quantitative analysis of feed crude fat and corase fiber based on near infrared spectroscopy and variables selection methods [J]. Spectrosc Spect Anal, 2020, 40(1): 215-220.
- [9] GB 5009. 3-2016 食品安全国家标准食品中水分的测定[S].
GB 5009. 3-2016 National food safety standard-Determination of moisture in foods [S].
- [10] GB 5009. 5-2016 食品安全国家标准食品中蛋白质的测定[S].
GB 5009. 5-2016 National food safety standard-Determination of protein in foods [S].
- [11] 褚小立. 近红外光谱分析技术实用手册[M]. 北京: 机械工业出版社, 2016.
Chu XL. Practical Handbook of near infrared spectroscopy [M]. Beijing: China Machine Press, 2016.
- [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 JC, Ge JF. Overview of NIRS data sample selection and preprocessing methods [J]. Foreign Elec Meas Technol, 2019, 38(3): 1-7.
- [13] 第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, 39(9): 2800-2806.
Di WPY, Bian XH, Wang ZF, et al. study on the selection of spectral preprocessing methods [J]. Spectrosc Spect Anal, 2019, 39(9): 2800-2806.
- [14] 王世芳, 韩平, 崔广禄, 等. SPXY 算法的西瓜可溶性固形物近红外光谱检测[J]. 光谱学与光谱分析, 2019, 39(3): 738-742.
Wang SF, Han P, Cui GL, et al. The NIR detection research of soluble solid content in watermelon based on SPXY algorithm [J]. Spectrosc Spect Anal, 2019, 39(3): 738-742.
- [15] 刘伟, 赵众, 袁洪福, 等. 光谱多元分析校正集和验证集样本分布优选方法研究[J]. 光谱学与光谱分析, 2014, 34(4): 947-951.
Liu W, Zhao Z, Yuan HF, et al. An optimal selection method of calibration set and validation set for spectral multivariate analysis [J]. Spectrosc Spect Anal, 2014, 34(4): 947-951.
- [16] 王翠秀, 曹见飞, 顾振飞, 等. 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[J]. 大豆科学, 2019, 38(6): 968-976.
Wang CX, Cao JF, Gu ZF, et al. Rapid nondestructive test of soybean protein and fat by near infrared spectroscopy combined with different

model methods [J]. *Soybean Sci*, 2019, 38(6): 968–976.

- [17] 郭成, 马月, 梁梦醒, 等. 基于近红外光谱结合波长优选检测单颗葡萄的 SSC 含量[J]. *食品与机械*, 2016, 32(9): 39–43.

Guo C, Ma Y, Liang MX, *et al.* Determination of SSC content in single grape based on NIR combined with wavelength selection [J]. *Food Mach*, 2016, 32(9): 39–43.

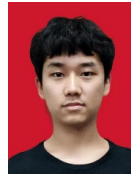
- [18] 邹涛, 兰树明, 阎巍, 等. 基于便携式近红外光谱仪的大豆蛋白波长优选[J]. *分析仪器*, 2019, 3: 94–99.

Zhou T, Lan SM, Yan W, *et al.* Soybean protein wavelength optimization based on portable NIR spectrometer [J]. *Anal Inst*, 2019, 3: 94–99.

- [19] Pedro SS, Andreia S, Ana C, *et al.* Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms [J]. *Food Chem*, 2018, 242: 196–204.

(责任编辑: 王 欣)

作者简介



郭东升, 硕士研究生, 主要研究方向为食品检测技术。

E-mail: gdsgds4568@163.com



张志勇, 博士, 副教授, 主要研究方向为农产品检测技术。

E-mail: zzyzyy1979@163.com