doi:10.6041/j.issn.1000-1298.2021.06.038

基于近红外光谱和机器学习的大豆种皮裂纹识别研究

汪六三　黄子良　王儒敬
(中国科学院合肥物质科学研究院, 合肥 230031)

摘要: 针对目前大豆种皮裂纹检测主要依靠人工、检测效率低、误差大的问题,提出一种基于近红外光谱技术和机器学习的大豆种皮裂纹自动识别方法。采用 FT-NIR 光谱仪采集 150 粒大豆样品(裂纹大豆 75 粒,正常大豆 75 粒)的近红外光谱,采用原始光谱、标准正态变量变换(Standard normal variate, SNV)、多元散射校正(Multiple scatter correction, MSC)、一阶导数结合 SG 平滑、二阶导数结合 SG 平滑等 5 种方法对获得的光谱进行预处理,分别采用偏最小二乘判别分析法(Partial least squares discriminant analysis, PLS-DA)、k-近邻法(k-nearest neighbor, KNN)、支持向量机法(Support vector machine, SVM)、随机森林法(Random forest, RF)、随机梯度提升法(Stochastic gradient boosting, SGB)、极端梯度提升法(Extreme gradient boosting, XGBoost)等 6 种机器学习方法建立了大豆种皮裂纹识别模型,研究了不同光谱预处理方法对 6 种机器学习方法分类效果的影响,对比分析了不同建模方法的分类效果。结果表明,光谱预处理方法对不同机器学习方法的分类效果差别较大。在合适的光谱预处理条件下,6 种不同的机器学习算法的验证集准确率均不低于 80.00%。PLS-DA 的分类效果最好,验证集最优准确率达到 90.00%;XGBoost 的分类效果次之,验证集最优准确率达到 86.67%,接下来依次是 SVM、KNN、SGB 和 RF。利用近红外光谱技术和机器学习方法识别大豆种皮裂纹是可行的,在原始光谱条件下,PLS-DA 是大豆种皮裂纹识别的最佳方法。

关键词: 大豆种皮; 裂纹识别; 近红外光谱; 机器学习
中图分类号: O657.33; S123　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)06-0361-08　　**OSID**:

Identification of Soybean Seed Coat Crack Based on Near Infrared Spectroscopy and Machine Learning

WANG Liusan　HUANG Ziliang　WANG Rujing
(*Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031, *China*)

Abstract: At present, the detection of soybean seed coat crack mainly depends on visual inspection, which has low detection efficiency and large error, a method for automatic identification of soybean seed coat cracks based on near infrared spectroscopy and machine learning was proposed. The near infrared spectra of 150 soybean samples (75 cracked and 75 normal) were collected by FT-NIR spectrometer. The original spectra, standard normal variable (SNV), multiple scatter correction (MSC), the first derivative and the second derivative with SG smoothing were used to process the obtained spectra. Then partial least squares discriminant analysis (PLS-DA), k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), stochastic gradient boosting (SGB) and extreme gradient boosting (XGBoost) were used to establish soybean seed coat crack identification models. The effects of different spectral preprocessing methods on the classification results of the six machine learning methods were compared and analyzed. Under the appropriate spectral preprocessing conditions, the accuracy of validation set of six different machine learning algorithms was not less than 80.00%. PLS-DA had the best classification result, and the optimal accuracy rate of validation set reached 90.00%; the next was XGBoost, the optimal accuracy rate of validation set reached 86.67%, followed by SVM, KNN, SGB and RF. The results showed that near infrared spectroscopy combined with machine learning was feasible to identify soybean seed coat cracks, and PLS-DA was the best method to identify soybean seed coat cracks under the original spectral conditions. The research result can provide a method for automatic identification of soybean seed coat cracks.

Key words: soybean seed coat; crack identification; near infrared spectroscopy; machine learning

收稿日期: 2020-07-28　修回日期: 2020-08-30
基金项目: 国家重点研发计划项目(2018YFD0101004)
作者简介: 汪六三(1982—),男,副研究员,博士,主要从事光谱技术及其应用研究,E-mail: lswang@iim.ac.cn

0 引言

大豆具有很高的营养与应用价值,既可用作动物饲料的原料,也可用于其他食品行业[1]。大豆在种植生产、收获、加工运输、贮藏等过程中容易产生种皮裂纹。大豆种皮裂纹性状不仅容易使籽粒和幼苗遭受微生物及害虫的侵害,还会导致籽粒蛋白质及油脂产量和品质的降低,使部分大豆不能满足市场等级的要求,进而影响大豆的经济价值[2]。

传统的大豆种皮裂纹检测方法有目测法和化学法。目视检查方法具有主观性,其检测结果不一致和过程缓慢等;化学方法具有破坏性,且检测耗时[3-4]。利用机器视觉技术可检测大豆种皮裂纹,其分类准确率最高达到了 96.00%[5-6]。虽然获得了很高的分类结果,但是利用机器视觉技术只能提供大豆的物理性质,无法提供其化学性质。在大豆分级中,物理性质和化学性质均是大豆的重要质量指标,如裂纹、蛋白质含量、油脂含量等。近红外光谱技术能同时检测被检测对象的物理和化学性质[7],且具有快速、无损等优点。在种子领域,主要利用近红外光谱技术检测种子的内部品质,如蛋白质、脂肪、水分含量和损伤等[8-15],利用近红外光谱技术检测种子裂纹的研究鲜见报道。

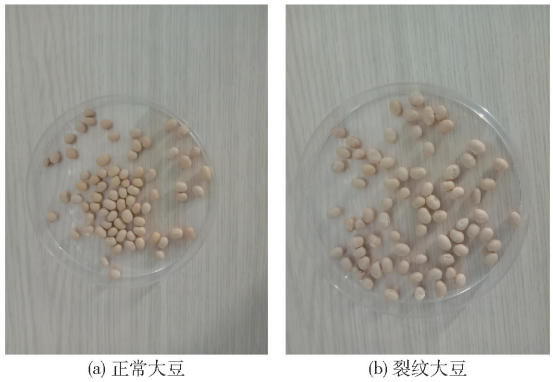
近红外光谱是一种重叠、宽而弱的信号,没有单个成分的明显特征[16],故有必要使用适当的机器学习建模方法来提取细微、有价值的信息。近年来,偏最小二乘判别法、支持向量机、k-近邻法等机器学习分类算法在近红外光谱上应用较多。LIDIA 等[14]利用近红外光谱结合偏最小二乘判别法、簇类独立软模式法、k-近邻法、最小二乘支持向量机 4 种分类算法,对大豆、玉米种子损伤和活力的判别进行了可行性研究,结果表明,对玉米热损伤判别是可行的,采用偏最小二乘判别法对玉米热损伤判别正确率达到了 99.00%。彭彦昆等[15]研究了近红外光谱用于热损伤番茄种子快速无损识别的可行性,结果表明,采用偏最小二乘判别法和支持向量机判别模型的验证集总体正确率均大于 96%,均可用于热损伤番茄种子的判别。BEVILACQUA 等[17]利用近红外光谱结合偏最小二乘判别分析法对不同产地的特级初榨橄榄油进行了分类研究,取得了较好的结果。王亚丽等[18]利用近红外光谱结合偏最小二乘判别法分析法对玉米种子活力进行了无损判别,判别准确率达到了 96.00%。高彤等[19]利用近红外光谱结合支持向量机对单粒玉米种子成熟度进行了无损识别,判别模型的分类准确率达到了 92.00%。然而,随机梯度提升法、极端梯度提升法等算法在近红外光谱分类中的应用报道较少。

本研究以 150 粒大豆籽粒(75 粒裂纹和 75 粒正常大豆)为研究对象,利用 FT-NIR 光谱仪采集大豆光谱,对比分析原始光谱(为了对照,当作预处理方法的一种)、标准正态变量变换(SNV)、多元散射校正(MSC)、一阶导数结合 SG 平滑、二阶导数结合 SG 平滑 5 种光谱预处理方法对偏最小二乘判别分析法、k-近邻法、支持向量机法、随机森林法、随机梯度提升法、极端梯度提升法 6 种机器学习方法分类效果的影响,研究这 6 种机器学习方法的分类效果,探索近红外光谱技术结合机器学习方法识别大豆种皮裂纹的可行性。

1 材料与方法

1.1 样品收集和制备

在种子市场上购买大豆种子(皖豆 15,安徽绿雨种业股份有限公司)进行实验。选取 75 粒种皮完好无损的大豆种子放入水中(24℃)浸泡 15 min,然后使用干燥箱在 105℃条件下对浸种大豆种子干燥 2 h。大豆样品干燥后和浸泡前质量接近。由于大豆在干燥过程中水分快速变化,产生大豆种皮裂纹。另外,从未处理的大豆种子中筛选出 75 粒种皮完好无损的大豆种子样品,共同组成 150 粒大豆种子样品,见图 1。



(a) 正常大豆 (b) 裂纹大豆

图 1 150 粒大豆种子样品

Fig. 1 150 soybean samples

1.2 光谱采集

用 FT-NIR 光谱仪(布鲁克公司,德国)对 150 粒大豆种子进行了光谱采集。图 2 为 FT-NIR 光谱仪。将大豆种子放在样品台上,盖上带有收集探头的盖子,用近红外卤素灯照射,光透过种子传播到检测器,采集每粒大豆种子的透射光谱。在波段 5 793 ~ 12 489 cm^{-1} 范围内,以 16 cm^{-1} 的光谱分辨率分别测定了裂纹大豆种子和正常大豆种子。对每粒大豆种子进行 32 次连续扫描,共测量两次,得到平均光谱。光谱仪开机预热 0.5 h,待卤素灯稳定再进行测量。样品光谱采集前首先测量背景,然后每

隔 3 h 左右测量一次背景。

图 2　FT－NIR 光谱仪

Fig. 2　FT－NIR spectrometer

1.3　光谱预处理

大豆种子测量状态和光谱仪工作状态的改变可能导致随机噪声、基线漂移和多元散射[20－21]。为了降低这些效应对模型稳健性的影响，提高模型预测精度，需要对光谱数据进行预处理。本研究采用了原始光谱、标准正态变量变换、多元散射校正、一阶导数结合 SG 平滑和二阶导数结合 SG 平滑 5 种方法对大豆种子光谱进行预处理。一阶导数结合 SG 平滑和二阶导数结合 SG 平滑均使用二次多项式拟合和 15 点窗口平滑。

1.4　模型建立与评价

本研究中采用 R 语言 cart 包的均衡取样函数[22]选取 80% 大豆种子(120 粒)和 20% 大豆种子(30 粒)分别作为校正集和验证集，分别采用偏最小二乘判别分析法[23－24]、k-近邻法[25]、支持向量机法[26－27]、随机森林法[28－30]、随机梯度提升法[31]和极端梯度提升法[32]建立大豆裂纹识别模型。在建立模型前，对每种方法均采用前面所述的 5 种光谱预处理方法对光谱数据进行处理，比较不同的预处理方法对 6 种机器学习建模方法性能的影响。最后比较 6 种机器学习建模方法的性能差异。

使用特异性(Specificity，Sp)、灵敏度(Sensitivity，Sn)和准确率(Accuracy，Ac)3 个指标评价模型性能[33]。一般来说，特异性、灵敏度、准确率越高，模型性能越好。

1.5　数据分析软件

采用 Unscrambler X 10.4.1 预处理光谱数据。大豆裂纹识别模型均采用 R 语言 3.6.1 构建。采用 Matlab R2012b 和 Microsoft Excel 2010 作图。

2　结果与讨论

2.1　光谱特征分析

图 3 显示了 75 粒裂纹大豆种子和 75 粒正常大豆种子的原始光谱和经过 SNV、MSC、一阶导数结合 SG 平滑和二阶导数结合 SG 平滑处理的平均光谱。图 3a 的原始光谱中裂纹和正常大豆种子光谱差异不明显，在 6 827～7 136 cm^{-1}波段附近近似重合，其它波段有偏移，偏移可能是由散射造成的。经过 SNV 处理(图 3b)和 MSC 处理(图 3c)后的裂纹和正常大豆种子光谱近似，均表现在 5 809～7 020 cm^{-1}和 8 239～8 432 cm^{-1}波段附近差异明显，其它波段接近重合，图 3a 中的偏移被消除了。图 3d 中一阶导数结合 SG 平滑处理的裂纹和正常大豆光谱在 6 102～

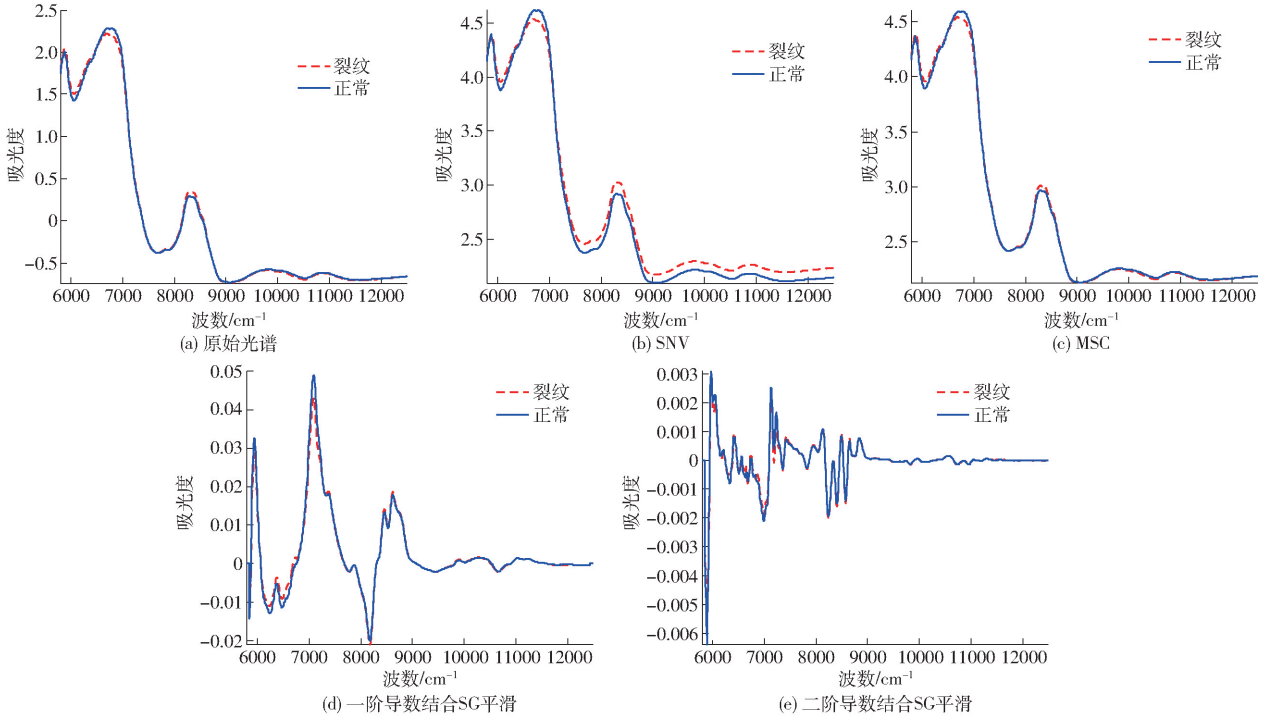


图 3　不同预处理条件下裂纹大豆种子和正常大豆种子的平均光谱

Fig. 3　Average spectra of cracked and normal soybeans under different pretreatment conditions

6 858 cm^{-1}和7 035 ~ 7 406 cm^{-1}波段差异明显,其它大部分波段近似重合。图 3e 二阶导数结合 SG 平滑处理的裂纹和正常大豆光谱在 5 971 ~ 6 889 cm^{-1}和 7 128 ~ 7 552 cm^{-1}波段有差异,但差异较小,在其它大部分波段与图 3d 相同,两光谱近似重合。总体来看,SNV、MSC 和一阶导数结合 SG 平滑对于裂纹和正常大豆种子的区分明显。图 3a ~ 3d 中在近红外光谱各个区域呈现出吸收峰,这是由于分子振动的组合和泛频所导致,如—CH、—OH 和—NH 等官能团的拉伸和弯曲。6 750 cm^{-1}附近的吸收峰可能属于水分吸收和无定形纤维素吸收[34]。8 347 cm^{-1}附近的吸收峰属于 CH 拉伸的 3 倍泛频。在 9 797 cm^{-1}和 10 880 cm^{-1}附近的吸收峰分别属于 NH 拉伸的 3 倍泛频和 CH 拉伸的 4 倍泛频,这两个吸收峰可能对应于蛋白质[35]。

2.2 裂纹识别模型建立

表 1 为不同预处理条件下 PLS - DA 分类结果。从表 1 中可以看出,潜变量数(Latent variable,LV)除多元散射校正(LV 为 9)和二阶导数结合 SG 平滑(LV 为 4)外都为 5,校正集中特异性在 86.67% ~ 100% 之间,灵敏度在 91.67% ~ 98.33% 之间,准确率除标准正态变量变换(90.00%)外均大于 90.00%;验证集中特异性(66.67% ~ 86.67%)均低于对应的校正集特异性,灵敏度除原始光谱(100%)外均低于对应的校正集灵敏度,准确率最大为 90.00%,最小为 63.33%。原始光谱取得了最好的分类结果,其验证集准确率达到 90.00%。一阶导数结合 SG 平滑光谱的分类结果次于原始光谱,其验证集准确率为 83.33%。SNV 光谱校正集准确率达到了 90.00%,验证集准确率达到了 80.00%,分类结果次于一阶导数结合 SG 平滑光谱。MSC 光谱校正集准确率为 99.17%,优于 SNV,但其验证集准确率只有 73.33%,低于 SNV 的验证集准确率(80.00%)。二阶导数结合 SG 平滑光谱分类结果最差,验证集准确率只有 63.33%。综上所述,5 种预处理方法对 PLS - DA 的大豆裂纹分类效果优劣顺序依次为原始光谱、一阶导数结合 SG 平滑、SNV、MSC、二阶导数结合 SG 平滑。

表 2 为不同预处理条件下 KNN 分类结果。K 是 KNN 算法中的重要可调参数。从表 2 中可以看出,K 为 3 和 5 时,校正集中特异性在 53.33% ~ 91.67% 之间,灵敏度在 78.33% ~ 95.00% 之间,准确率在 40.00% ~ 90.83% 之间;验证集中特异性在 66.67% ~ 93.33% 之间,灵敏度在 53.33% ~ 80.00% 之间,准确率在 56.67% ~ 83.33% 之间。SNV 和 MSC 校正集取得了最好的分类结果,且分类结果相同,校正集和验证集准确率分别为 90.83% 和 83.33%,校正集和验证集特异性分别为 91.67% 和 93.33%,校正集和验证集灵敏度分别为 90.00% 和 73.33%。原始光谱的分类结果次之,其校正集和验证集准确率分别达到 86.67% 和 73.33%。一阶导数结合 SG 平滑的校正集和验证集准确率均低于对应的原始光谱分类准确率。二阶导数结合 SG 平滑光谱分类结果最差。综上所述,5 种预处理方法对 KNN 的大豆种皮裂纹分类效果优劣顺序依次为 SNV(MSC)、原始光谱、一阶导数结合 SG 平滑、

表 1 不同预处理条件下 PLS - DA 分类结果

Tab. 1 Classification results of PLS - DA under different pretreatment conditions

光谱预处理方法	LV	校正集			验证集		
		Sp/%	Sn/%	Ac/%	Sp/%	Sn/%	Ac/%
原始光谱	5	90.00	91.67	90.83	80.00	100	90.00
SNV	5	86.67	93.33	90.00	80.00	80.00	80.00
MSC	9	100	98.33	99.17	66.67	80.00	73.33
一阶导数结合 SG 平滑	5	91.67	95.00	93.33	86.67	80.00	83.33
二阶导数结合 SG 平滑	4	98.33	96.67	97.50	73.33	53.33	63.33

表 2 不同预处理条件下 KNN 分类结果

Tab. 2 Classification results of KNN under different pretreatment conditions

光谱预处理方法	K	校正集			验证集		
		Sp/%	Sn/%	Ac/%	Sp/%	Sn/%	Ac/%
原始光谱	5	78.33	95.00	86.67	66.67	80.00	73.33
SNV	3	91.67	90.00	90.83	93.33	73.33	83.33
MSC	3	91.67	90.00	90.83	93.33	73.33	83.33
一阶导数结合 SG 平滑	5	83.33	86.67	85.00	73.33	60.00	66.67
二阶导数结合 SG 平滑	5	53.33	78.33	40.00	73.33	53.33	56.67

二阶导数结合 SG 平滑。

表 3 为不同预处理条件下 SVM 分类结果。γ 是核函数的参数，控制分割超平面的形状。γ 越大，支持向量越少。c 表示犯错成本，成本大意味着模型对误差的惩罚大，对应训练集中的误差小。表 3 中参数 γ 和 c 为优化值。从表 3 中可以看出，校正集中特异性在 80.00% ~100% 之间，灵敏度在 91.67% ~ 100% 之间，准确率在 85.83% ~ 100% 之间；验证集中特异性在 40.00% ~80.00% 之间，灵敏度在 73.33% ~93.33% 之间，准确率在 56.67% ~83.33% 之间。原始光谱取得了最好的分类结果，校正集和验证集准确率分别达到 99.17% 和 83.33%。SNV 光谱分类结果次之，校正集和验证集准确率分别达到 98.33% 和 80.00%。MSC 光谱校正集准确率(100%)大于 SNV，但其验证集准确率(76.67%)小于 SNV。一阶导数结合 SG 平滑光谱的校正集和验证集准确率均低于对应的 MSC。二阶导数结合 SG 平滑分类结果最差，虽然其校正集准确率为 94.17%，但验证集准确率仅为 56.67%。综上所述，5 种预处理方法对 SVM 的大豆种皮裂纹分类效果优劣顺序依次为原始光谱、SNV、MSC、一阶导数结合 SG 平滑、二阶导数结合 SG 平滑。

表 4 为不同预处理条件下 RF 分类结果。在本研究中，使用了程序默认的参数，决策树数(ntree)为 500，随机抽取的特征数(mtry)为 29。从表 4 中可以看出，校正集中特异性在 76.67% ~86.67% 之间，灵敏度在 75.00% ~88.33% 之间，准确率在 77.50% ~ 86.67% 之间。验证集中特异性在 66.67% ~80.00% 之间，灵敏度在 73.33% ~ 86.67% 之间，准确率在 76.67% ~80.00% 之间。一阶导数结合 SG 平滑光谱分类结果最好，校正集和验证集准确率分别达到 86.67% 和 80.00%。除原始光谱外其余 3 种预处理方法的校正集准确率不小于 80.00%，但验证集准确率均为 76.67%，说明各种预处理方法对随机森林分类的影响相对较小。依据表 4 中校正集、验证集准确率，5 种预处理方法对 RF 的大豆裂纹分类效果优劣顺序依次为一阶导数结合 SG 平滑、SNV、二阶导数结合 SG 平滑、MSC、原始光谱。

表 3　不同预处理条件下 SVM 分类结果

Tab. 3　Classification results of SVM under different pretreatment conditions

光谱预处理方法	γ	c	校正集			验证集		
			Sp/%	Sn/%	Ac/%	Sp/%	Sn/%	Ac/%
原始光谱	1×10^{-6}	1×10^{6}	98.33	100	99.17	73.33	93.33	83.33
SNV	1×10^{-6}	10 000	100	96.67	98.33	80.00	80.00	80.00
MSC	1×10^{-5}	1×10^{5}	100	100	100	73.33	80.00	76.67
一阶导数结合 SG 平滑	10	10	80.00	91.67	85.83	73.33	73.33	73.33
二阶导数结合 SG 平滑	10	1 000	90.00	98.33	94.17	40.00	73.33	56.67

表 4　不同预处理条件下 RF 分类结果

Tab. 4　Classification results of RF under different pretreatment conditions　%

光谱预处理方法	校正集			验证集		
	Sp	Sn	Ac	Sp	Sn	Ac
原始光谱	76.67	78.33	77.50	66.67	86.67	76.67
SNV	83.33	80.00	81.67	80.00	73.33	76.67
MSC	80.00	80.00	80.00	80.00	73.33	76.67
一阶导数结合 SG 平滑	85.00	88.33	86.67	80.00	80.00	80.00
二阶导数结合 SG 平滑	86.67	75.00	80.83	80.00	73.33	76.67

表 5 为不同预处理条件下 SGB 分类结果。从表 5 中可以看出，校正集中特异性在 98.33% ~100% 之间，灵敏度在 95.00% ~100% 之间，准确率在 96.67% ~100% 之间；验证集中特异性在 66.67% ~80.00% 之间，灵敏度在 66.67% ~86.67% 之间，准确率在 66.67% ~80.00% 之间。一阶导数结合 SG 平滑光谱分类结果最好，校正集和验证集准确率分别达到了 100% 和 80.00%。原始光谱、SNV、MSC 的验证集准确率均为 76.67%，但 SNV 和 MSC 校正集的准确率为 100%，优于原始光谱的准确率(96.67%)。SNV 验证集的灵敏度为 80.00%，大于 MSC 的灵敏度(73.33%)，SNV 光谱分类效果略优于 MSC 光谱。二阶导数结合 SG 平滑光谱分类结果最差，验证集准确率为 66.67%。综上所述，5 种预处理方法对 SGB 的大豆裂纹分类效果优劣顺序依次为一阶导数结合 SG 平滑、SNV、

表 5 不同预处理条件下 SGB 分类结果

Tab. 5 Classification results of SGB under different pretreatment conditions %

光谱预处理方法	校正集			验证集		
	Sp	Sn	Ac	Sp	Sn	Ac
原始光谱	98.33	95	96.67	66.67	86.67	76.67
SNV	100	100	100	73.33	80.00	76.67
MSC	100	100	100	80.00	73.33	76.67
一阶导数结合 SG 平滑	100	100	100	80.00	80.00	80.00
二阶导数结合 SG 平滑	100	98.33	99.17	66.67	66.67	66.67

MSC、原始光谱、二阶导数结合 SG 平滑。

表 6 为不同预处理条件下 XGBoost 分类结果。其中 D 为最大深度，e 为学习率，表中参数 D 和 e 为优化值。从表 6 中可以看出，校正集中特异性在 95.00% ~100% 之间，灵敏度在93.33% ~100% 之间，准确率在 94.17% ~100% 之间；验证集中特异性在 60.00% ~86.67% 之间，灵敏度在73.33% ~93.33% 之间，准确率在 76.67% ~86.67% 之间。各预处理方法在校正集中均取得了很好的分类结果，准确率大于 94.00%，但验证集分类结果没有校正集好，准确率均小于 90.00%。SNV 光谱取得了最好的分类结果，校正集和验证集准确率分别达到 100% 和 86.67%。MSC 和一阶导数结合 SG 平滑次之，校正集和验证集准确率分别达到 100% 和 83.33%。MSC 在验证集灵敏度(86.67%)大于一阶导数结合 SG 平滑的灵敏度(80.00%)，MSC 光谱分类效果略优于一阶导数结合 SG 平滑。二阶导数结合 SG 平滑光谱和原始光谱验证集准确率均为 76.67%，但二阶导数结合 SG 平滑光谱校正集准确率(100%)大于原始光谱准确率(94.17%)。综上所述，5 种预处理方法对 XGBoost 的大豆种皮裂纹分类效果优劣顺序依次为 SNV、MSC、一阶导数结合 SG 平滑、二阶导数结合 SG 平滑、原始光谱。

表 6 不同预处理条件下 XGBoost 分类结果

Tab. 6 Classification results of XGBoost under different pretreatment conditions

光谱预处理方法	D	e	校正集			验证集		
			Sp/%	Sn/%	Ac/%	Sp/%	Sn/%	Ac/%
原始光谱	1	0.6	95.00	93.33	94.17	60.00	93.33	76.67
SNV	3	0.1	100	100	100	80.00	93.33	86.67
MSC	1	0.6	100	100	100	80.00	86.67	83.33
一阶导数结合 SG 平滑	3	0.4	100	100	100	86.67	80.00	83.33
二阶导数结合 SG 平滑	1	0.4	100	100	100	80.00	73.33	76.67

2.3 建模方法对比分析

图 4 给出了 6 种机器学习方法在最优光谱预处理方法下的模型性能。从图 4 中可以看出，PLS-DA 的分类结果最好，其校正集特异性、灵敏度和准确率均不小于 90.00%，验证集特异性、灵敏度均不小于 80.00%，准确率达到 90.00%；XGBoost 的分类结果次之，其校正集特异性、灵敏度和准确率均达到 100%，验证集特异性、灵敏度均不小于 80.00%，准

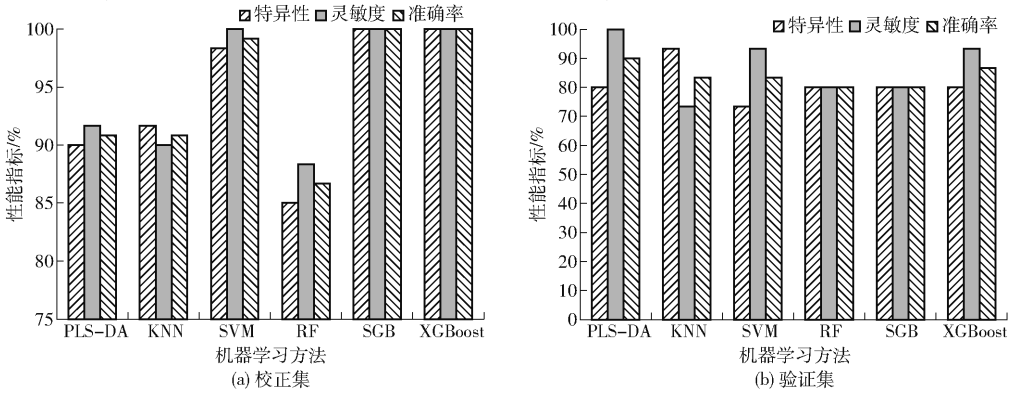


图 4 6 种建模方法的性能比较

Fig. 4 Performance comparison of six modeling methods

确率达到 86.67%；其次是 SVM 和 KNN 算法，验证集准确率相同，均为 83.33%，但 SVM 的校正集特异性、灵敏度和准确率（均大于 98.00%）优于 KNN 校正集（均大于 90.00%）。最后是 RF 和 SGB 算法，验证集准确率相同，均为 80.00%，但 SGB 的校正集特异性、灵敏度和准确率均达到 100%，优于 RF 校正集（均不小于 85.00%）。综上所述，在本研究中，依据分类结果，分类算法的优劣顺序依次为 PLS－DA、XGBoost、SVM、KNN、SGB、RF。

本研究的大豆种皮裂纹识别准确率（80.00% ~ 90.00%）低于 GUNASEKARAN 等[5]利用图像技术识别大豆种皮裂纹的准确率（96.00%），与 LIU 等[6]图像识别大豆种皮裂纹结果（准确率为 85.60%）一致，导致这种结果的原因可能是本文的样品裂纹相对于文献[5]的样品复杂而与文献[6]接近。与近红外光谱技术识别大豆种皮裂纹相比较，本研究的大豆种皮裂纹识别准确率与 WANG 等[7]的识别准确率（84.00%）相一致。

3　结束语

利用 FT－NIR 光谱仪采集裂纹和正常大豆种子近红外光谱，使用偏最小二乘判别分析法、k-近邻法、随机森林法、支持向量机法、随机梯度提升法、极端梯度提升法 6 种机器学习方法建立了大豆种皮裂纹识别模型，研究了 5 种光谱预处理方法对 6 种机器学习建模方法分类效果的影响。结果表明，不同光谱预处理方法对 6 种机器学习方法的分类效果差别较大，需要根据具体的应用情况测试并确定适宜的预处理方法。对比分析了 6 种机器学习方法的分类效率，在合适的预处理条件下，6 种方法验证集准确率均不低于 80.00%。PLS－DA 分类效果最好，验证集准确率最高达到 90.00%，XGBoost 次之，其余依次为 SVM、KNN、SGB、RF。研究表明，利用近红外光谱技术和机器学习方法识别大豆种皮裂纹具有可行性。在原始光谱条件下，近红外光谱技术结合 PLS－DA 是大豆种皮裂纹识别的最佳方法。

参　考　文　献

[1] QUTOB D, MA F, PETERSON C A, et al. Structural and permeability properties of the soybean seed coat[J]. Botany, 2008, 86(3): 219－227.

[2] 崔晓培，郑金焕，胡冬梅. 大豆种皮裂纹的研究进展[J]. 大豆科学, 2019, 38(4): 656－663.
CUI Xiaopei, ZHENG Jinhuan, HU Dongmei. Research progress on cracks in soybean seed coat [J]. Soybean Science, 2019, 38(4): 656－663. (in Chinese)

[3] NEETHIRAJAN S, KARUNAKARAN C, JAYAS D S, et al. Detection techniques for stored-product insects in grain[J]. Food Control, 2007, 18(2): 157－162.

[4] SINGH C B, JAYAS D S, PALIWAL J, et al. Identification of insect damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging[J]. Computers and Electronics in Agriculture, 2010, 73(2): 118－125.

[5] GUNASEKARAN S, COOPER T M, BERLAGE A G. Soybean seed coat and cotyledon crack detection by image processing [J]. Journal of Agricultural Engineering Research, 1988, 41(2): 139－148.

[6] LIU D J, NING X F, LI Z M, et al. Discriminating and elimination of damaged soybean seeds based on image characteristics [J]. Journal of Stored Products Research, 2015, 60: 67－74.

[7] WANG D, RAM M S, DOWELL F E. Classification of damaged soybean seeds using near-infrared spectroscopy [J]. Transactions of the ASAE, 2002, 45(6): 1943－1948.

[8] CHENG F, YING Y B, LI Y B. Detection of defects in rice seeds using machine vision[J]. Transactions of the ASABE, 2006, 49(6): 1929－1934.

[9] DELWICHE S R, SOUZA E J, KIM M S. Limitations of single kernel near-infrared hyperspectral imaging of soft wheat for milling quality[J]. Biosystems Engineering, 2013, 115(3): 260－273.

[10] FERREIRA D S, GALAO O F, PALLONE J A L, et al. Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples[J]. Food Control, 2014, 25(1): 227－232.

[11] FERREIRA D S, PALLONE J A L, POPPI R J. Fourier transform near-infrared spectroscopy (FT－NIRS) application to estimate Brazilian soybean [*Glycine max* (L.) Merril] composition [J]. Food Research International, 2013, 51(1): 53－58.

[12] ARMSTRONG P R, TALLADA J G, HURBURGH C, et al. Development of single seed near-infrared spectroscopic predictions of corn and soybean constituents using bulk reference values and mean spectra[J]. Transactions of ASABE, 2011, 54(4): 1529－1535.

[13] 姚鑫淼，张瑞英，李霞辉，等. 近红外透射光谱法(NITS)分析大豆品质的研究[J]. 大豆科学, 2006, 25(4): 417－420, 424.
YAO Xinmiao, ZHANG Ruiying, LI Xiahui, et al. Study on analysis of soybean quality by near infrared transmittance spectroscopy[J]. Soybean Science, 2006, 25(4): 417－420, 424. (in Chinese)

[14] LIDIA E A, DAVID D E, SUSAN D, et al. Feasibility of near infrared spectroscopy for analyzing corn kernel damage and viability of soybean and corn kernels[J]. Journal of Cereal Science, 2012, 55(2): 160－165.

[15] 彭彦昆，赵芳，李龙，等. 利用近红外光谱与 PCA－SVM 识别热损伤番茄种子[J]. 农业工程学报, 2018, 34(5): 159－165.

PENG Yankun, ZHAO Fang, LI Long, et al. Discrimination of heat-damaged tomato seeds based on near infrared spectroscopy and PCA - SVM method[J]. Transactions of the CSAE, 2018, 34(5): 159 - 165. (in Chinese)
[16] TAN C, WU T, XU Z H, et al. A simple ensemble strategy of uninformative variable elimination and partial least-squares for near-infrared spectroscopic calibration of pharmaceutical products[J]. Vibrational Spectroscopy, 2012, 58: 44 - 49.
[17] BEVILACQUA M, BUCCI R, MAGRA D, et al. Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: a case study[J]. Analytica Chimica Acta, 2012, 717: 39 - 51.
[18] 王亚丽,彭彦昆,赵鑫龙,等. 玉米种子活力逐粒无损检测与分级装置研究[J/OL]. 农业机械学报, 2020,51(2):350 - 356.
WANG Yali, PENG Yankun, ZHAO Xinlong, et al. Design and experiment of non-destructive testing and grading device for corn seed vigor[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(2):350 - 356. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=2002038&flag=1&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2020.02.038. (in Chinese)
[19] 高彤,吴静珠,毛文华,等. 单粒玉米种子成熟度快速判别方法统[J/OL]. 农业机械学报,2019,50(增刊):399 - 403.
GAO Tong, WU Jingzhu, MAO Wenhua, et al. Method for quickly determining maturity of single corn seed[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019,50(Supp.):399 - 403. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=2019s061&flag=1&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2019.S0.061. (in Chinese)
[20] SHI T, CHEN Y, LIU Y, et al. Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals[J]. Journal of Hazardous Materials, 2014, 265: 166 - 176.
[21] RINNAN A, VAN DEN BERG F, ENGELSEN B. Review of the most common pre-processing techniques for near-infrared spectra[J]. TRAC Trends in Analytical Chemistry, 2009, 28(10):1201 - 1222.
[22] KUHN M. Building predictive models in R using the caret package[J]. J. Stat. Softw., 2008, 28(5): 1 - 26.
[23] BARKER M, RAYENS W. Partial least squares for discrimination[J]. Journal of Chemometrics, 2003, 17(3):166 - 173.
[24] BALLABIO D, CONSONNI V. Classification tools in chemistry. Part 1: linear models. PLS - DA[J]. Analytical Methods, 2013, 5(16): 3790 - 3798.
[25] TAN C, QIN X, LI M. Comparison of chemometric methods for brand classification of cigarettes by near-infrared spectroscopy [J]. Vibrational Spectroscopy, 2009, 51(2): 276 - 282.
[26] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5 - 32.
[27] SPEISER J L, MILLER M E, TOOZE J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Expert Systems with Applications, 2019, 134: 93 - 101.
[28] LIAW A, WIENER M. Classification and regression by random forest[J]. R News, 2002, 2(3): 18 - 22.
[29] VAPNIK V. The nature of statistical learning theory[M]. Springer, 2000.
[30] POCHET N, DE SMET F, SUYKENS J A, et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction[J]. Bioinformatics, 2004, 20: 3185 - 3195.
[31] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4):367 - 378.
[32] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016.
[33] SMITH B R, BAKER M J, PALMER D S. PRFFECT: a versatile tool for spectroscopists[J]. Chemometrics and Intelligent Laboratory, 2018, 172:33 - 42.
[34] AGELET L E, ARMSTRON P R, TALLADA J G, et al. Differences between conventional and glyphosate tolerant soybeans and moisture effect in their discrimination by near infrared spectroscopy[J]. Food Chemistry, 2013, 141(3): 1895 - 1901.
[35] TAJUDDIN T, WATANABE S, MASUDA R, et al. Application of near infrared transmittance spectroscopy to the estimation of protein and lipid contents in single seeds of soybean recombinant inbred lines for quantitative trait loci analysis[J]. Journal Near Infrared Spectroscopy, 2002, 10(4): 315 - 325.