DOI: 10.19756/j.issn.0253-3820.231441

基于近红外光谱与高斯过程的高粱单宁含量快速检测

赵瑾熠[1] 陈争光*[1] 衣淑娟[2]
(黑龙江八一农垦大学, 信息与电气工程学院[1], 工程学院[2], 大庆 163000)

摘 要 高粱是我国主要的酿酒原料之一,也是重要的饲料原料。在酿酒过程中,高粱籽粒中的单宁含量对酒品品质具有决定性作用; 作为饲料原料时,高粱的单宁含量对饲料的利用率有重要影响,因此高粱中的单宁含量对其品质和用途具有重要影响。传统方法检测高粱中单宁含量时存在耗时长和成本高等问题,本研究利用近红外光谱结合化学计量学方法实现了高粱单宁含量的快速无损检测。在对光谱进行预处理的基础上,使用蒙特卡洛交叉验证法(MCCV)结合高斯过程回归(GPR)进行异常样本剔除; 然后,将样本集随机划分为建模集和预测集,使用无信息变量消除法(UVE)进行特征波长选择; 最后,建立 GPR 模型,并与偏最小二乘回归(PLSR)模型和支持向量机回归(SVR)模型进行性能对比。结果表明, GPR 模型的性能全面优于 PLSR 和 SVR 模型,经去趋势组合 Savitzky-Golay 卷积平滑进行预处理,剔除异常样本并进行特征波长选择后,建立的 GPR 模型为最优模型,其建模集决定系数(R_C^2)、预测集决定系数(R_P^2)和相对分析误差(RPD)分别为 0.9979、0.9529 和 4.8453。本研究结果表明,采用近红外光谱结合化学计量学方法建立的 GPR 模型可用于高粱单宁含量的快速无损检测。

关键词 近红外光谱; 化学计量学; 高粱单宁; 高斯过程法; 快速无损检测

高粱是世界第五大粮食作物[1],也是我国重要的粮食作物之一[2]。高粱富含多种营养物质,在我国主要用作酿酒原料,而在国际市场主要用于饲料行业[3-4]。高粱作为酿酒原料时,籽粒中的单宁含量对产品品质具有决定性作用。这主要是由于单宁在酿酒过程中会产生丁香酸和丁香醛等风味物质,赋予白酒独特的风味。同时,在发酵过程中,单宁还具有抑制有害微生物生长和提高出酒率的功效[5]。然而,单宁也是一种抗营养因子,味苦涩,具有收敛性。单宁可与蛋白质、糖类和金属离子形成难以吸收的复合物,从而降低动物的摄食率,并影响营养物质的吸收利用率,但适量的单宁可以改善禽畜的生长性能,提高饲料的利用率[6-7]。因此,快速、高效和低成本地检测高粱中的单宁含量对于高粱农业生产和质量控制至关重要。目前,高粱中单宁含量的检测方法主要有人工经验判别法和实验室化学方法[8],人工经验判别容易受主观影响,效率低,难以形成统一的标准; 实验室化学方法操作繁琐、费时费力,并且需要对样品进行破坏性处理。近红外光谱技术通过测量样品在近红外光谱范围内的吸收和反射特性获取样品的化学信息,无需对样品进行破坏性处理[9]。作为一种高效无损的检测技术,近红外光谱具有快速、无污染、无损伤和低成本等优点,并可以实现在线检测[10]。

目前,研究人员已基于近红外光谱构建了多种谷物养分预测模型,余松柏等[11]利用偏最小二乘回归(Partial least squares regression, PLSR)模型对高粱的多个成分进行预测,采用多元散射校正(Multiplicative scatter correction, MSC)对采集的高粱光谱数据进行预处理,使用蒙特卡洛无信息变量消除法(Monte carlo-elimination of uninformative variables, MCUVE)选择特征波长,建立的基于高粱完整籽粒的单宁 PLSR 回归模型的预测集决定系数(Prediction set determination coefficient, R_P^2)为 0.8841; 使用 MSC 结合 Z-Score 标准化进行预处理,使用反向区间偏最小二乘法选择特征波长,建立的基于高粱粉末的单宁含量的 PLSR 回归模型的 R_P^2 为 0.9414。尽管基于高粱粉末样本的单宁含量预测模型精度高于基于高粱籽粒的模型,但是检测过程需要破坏样本。刘敏轩等[12]测定了 60 份高粱籽粒的 4 个部位以及整粒种子所组成

2023-11-16 收稿; 2024-05-20 接受
国家自然科学基金项目(No. 52275246)和黑龙江八一农垦大学“三纵”支持项目(No. ZRCPY202214) 资助。
* E-mail: ruzee@byau.edu.cn

的 300 份样本的 6 种酚类物质的含量，其中，缩合单宁采用 MSC 结合一阶导数法（First derivative，FD）对光谱数据进行预处理，在此基础上建立 PLSR 模型，模型的 R_P^2 为 0.9558。Dykes 等[13]对高粱籽粒的总酚、缩合单宁和 3-脱氧花青素建立模型并进行预测，其中，缩合单宁的改进 PLSR 模型的 R_P^2 仅为 0.81。Zhang 等[14]对葡萄果皮和种子中的单宁进行建模预测，分别使用 MSC 结合支持向量机和 Savitzky-Golay 卷积平滑（Savitzky-golay smoothing，SG 平滑）结合 PLS 建立果皮和种子的单宁预测模型，R_P^2 分别为 0.8960 和 0.9243。

高斯过程回归（Gaussian process regression，GPR）是一种非参数的统计建模方法，基于高斯过程（Gaussian process，GP）的概念，将数据点视为随机变量，并假设数据点服从多元正态分布，通过对已观测到的数据点进行建模，可以预测未观测到的数据点的值，并估计其不确定性，还可通过选择合适的核函数而适应不同类型的数据和问题，从而提高模型的预测性能。GPR 对高维度、小样本的数据具有较强的处理能力，并具有容易实现、收敛性好和超参数自适应性等特点[15]。GPR 在许多实际问题中都表现出色，已被应用于机器学习、统计学和工程学等领域。李元等[16]提出了一种基于 GPR 的绝缘纸老化分析算法，并获得了较高的准确率。张韬等[17]提出使用蜻蜓算法优化 GPR 对锂电池健康状态进行预测，结果表明，模型预测精度高，运算速度快，尤其在处理小样本方面更具优势。以上研究表明，基于 GPR 建立高粱单宁预测模型具有可行性。

基于近红外光谱分析技术的单宁含量快速检测已有大量的研究报道，但这些研究的建模方法相对单一，多采用 PLS 建模方法，模型在预测集上的性能仍有提升可能。为了建立高粱单宁的快速检测模型，本研究利用近红外光谱技术采集高粱光谱，使用多种预处理方法，过滤光谱中的噪声信息。在预实验的基础上，选用无信息变量消除法（Elimination of uninformative variables，UVE）选择特征波长，提取光谱中的有效信息。在优选核函数基础上，建立 GPR 回归预测模型，并与 PLSR 和支持向量机回归（Support vector machine regression，SVR）等模型对比。通过计算模型的决定系数（Coefficient of determination，R^2）、均方根误差（Root mean square error，RMSE）和相对分析误差（Relative percent deviation，RPD），选择最优方案，建立高粱单宁含量的高性能预测模型，为高粱中单宁的快速检测提供了技术支持。

1　实验部分

1.1　仪器与试剂

TANGO FT-NIR 近红外光谱仪（德国 Bruker 公司）；UV-1800 紫外可见分光光度计（AOE 翱艺仪器上海有限公司）；WH-71 电热恒温干燥箱（天津市泰斯特仪器有限公司）；DM-50g 粉碎机（南京东迈科技仪器有限公司）；双杰 JJ224BC 电子分析天平（常熟市双杰测试仪器厂）；MK-60 低速台式离心机（湖南迈克尔实验仪器有限公司）；VM-210 漩涡振荡器（群安科学仪器浙江有限公司）；HJ-1 磁力搅拌器（金坛区西城新瑞仪器厂）。实验用水为蒸馏水。

单宁酸和柠檬酸铁铵（分析纯，福晨天津化学试剂有限公司）；8.0 g/L 氨溶液（分析纯，以达科技泉州有限公司）；75%二甲基甲酰胺溶液（分析纯，中国石化公司）。

2 g/L 单宁酸溶液：称取 0.2 g 单宁酸溶于蒸馏水中，定容至 100 mL；3.5 g/L 柠檬酸铁铵：称取 0.35 g 柠檬酸铁铵溶于蒸馏水中，定容至 100 mL。

1.2　样品采集与处理

本研究选取的高粱样本为 2022 年黑龙江八一农垦大学农学院收获的高粱，包含 65 个品种，共计 305 个样本。利用 TANGO FTNIR 近红外光谱仪先测得每个高粱样本完整籽粒光谱数据后，将其粉碎，过 40 目筛（筛孔直径 0.425 mm），采用柠檬酸铁铵法[18]测定单宁含量。

1.3　单宁含量测定

按国标（GBT 15686—2008）方法[18]对单宁含量进行测定。称取适量高粱粉碎后的样本，采用二甲基甲酰胺溶液提取高粱单宁，经离心后，取两份上清液，其中一份加水、柠檬酸铁铵溶液和氨溶液，另一份只加水和氨溶液（柠檬酸铁铵溶液替换成等体积水），显色后，以水为空白对照，采用分光光度计测定 525 nm 处吸光度值，采用单宁酸标准品绘制标准曲线。

单宁含量(X)以干基中单宁酸的质量分数(%)表示,按式(1)计算。

$$X=\frac{2C}{M}\times\frac{100}{100-H} \tag{1}$$

其中,C 为从标准曲线中读取的试样提取液中单宁酸的浓度(g/L);M 为试样的质量(g);H 为试样的水分含量(%)。

1.4 水分测定

按国标 GB 5009.3—2016[19]中的直接干燥法测定水分含量,通过干燥前后的称量数值计算出水分的含量。水分含量按式(2)计算:

$$H=\frac{M_1-M_2}{M_1}\times 100 \tag{2}$$

其中,H 为试样中水分的含量(%);M_1 和 M_2 分别为干燥前和干燥后试样的质量(g)。

1.5 光谱采集

采用 TANGO FT-NIR 近红外光谱仪采集 305 份高粱籽粒在 11542~3940 cm^{-1} 范围的近红外光谱,测量方式为漫反射和透射,分辨率为 8 cm^{-1}。扫描 32 次获得平均光谱。

1.6 数据处理

数据处理软件为 The Unscrambler X(10.4 版)、Matlab(R2021a 版本)和 Microsoft Office Excel。

1.7 定量模型的构建

以高粱的单宁作为分析指标,分别采用去趋势(Detrending, Det)、标准正态变换(Standard normal variate transformation, SNV)、去趋势组合标准正态变换和去趋势组合 SG 平滑对光谱数据进行预处理。使用蒙特卡洛交叉验证法(Monte Carlo cross-validation, MCCV)结合 GPR 建模对原始光谱进行异常样本剔除。采用随机法(Random selection, RS)按照 8∶2 的比例将样本集划分为建模集和预测集。采用 UVE 选择特征波长。建立基于 GP 的高粱中单宁含量预测模型,并与偏最小二乘法(Partial least squares, PLS)和支持向量机(Support vector machine, SVM)回归模型进行对比分析。通过对比不同模型的 R^2、RMSE 和 RPD 评价模型的性能。

1.8 建模方法

GPR 作为一种非参数的回归方法,用于建模和预测数据的连续函数关系,适于处理高维数、小样本和非线性等复杂问题。在给定样本光谱数据分布的前提下,GPR 用于推断对应样本高粱单宁值的分布。所得分布函数的数学期望即为 GPR 模型的预测结果。假设单宁值 y 是高斯分布函数 $f(x)$,服从均值为 $m(x)$、协方差为 $k(x, x')$的高斯过程分布,如式(3)所示:

$$y=f(x)\sim GP[m(x),k(x,x')] \tag{3}$$

其中,均值函数 $m(x)$和协方差核函数 $k(x, x')$定义如下:

$$m(x)=E[f(x)] \tag{4}$$

$$k(x,x')=\mathrm{cov}[f(x),f(x')] \tag{5}$$

GPR 的关键是定义一个核函数,用于衡量不同数据点之间的相似性。核函数的选择可根据问题的特点进行调整,常用的核函数有指数核函数、平方指数核函数、有理二次核函数和 matern 核函数等[20]。通过核函数可以计算出观测数据点之间的协方差矩阵,进而得到预测结果的均值和方差,在预实验的基础上,本研究采用的核函数为 matern32 核,由 matern 核函数的参数 v=3/2 得到[21],v 的大小影响函数的光滑性,matern52 核由 v=5/2 得到,matern32 核函数公式见式(6):

$$k(x,x')=\sigma_f^2\left(1+\frac{\sqrt{3}\left\|x-x'\right\|}{\sigma_l}\right)\exp\left(-\frac{\sqrt{3}\left\|x-x'\right\|}{\sigma_l}\right) \tag{6}$$

其中,x 和 x'是输入变量;σ_f^2 是信号方差参数,用于控制局部相关性的程度;σ_l 是特征长度尺度参数,用于调节输入变量之间的间隔;$\|x-x'\|$表示输入变量之间的欧氏距离。

1.9　模型评价指标

1.9.1　决定系数

R^2 是一种用于衡量模型对观测数据拟合程度的指标，通常用于比较不同模型的性能。R^2 值越大，模型性能越好，建模集和预测集的 R^2 的比值应控制在 0.9~1.1 之间，小于 0.9 表明模型存在欠拟合，大于 1.1 表明模型存在过拟合[22]。当 R^2=1.0 时，表明模型完美地预测了数据。在回归分析过程中，R^2 可评估模型的预测能力，并解释变量对因变量的影响程度。R^2 的计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{\sum_{i=1}^{n}(y_{i,\text{actual}} - \overline{y}_{\text{actual}})^2} \tag{7}$$

其中，$y_{\text{i,actual}}$ 和 $y_{i,\text{predicted}}$ 分别为第 i 个样本的真实值和预测值，y_{actual} 为真实值的平均值，n 为样本数。

1.9.2　均方根误差

RMSE 表示预测值和实际观测值之间的差值，用于测量模型的预测精度。RMSE 越小，模型精度越高，表明预测值与实际观测值之间的差异越小。计算公式见式(8)：

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_{i,\text{actual}} - y_{i,\text{predicted}})^2}{n}} \tag{8}$$

1.9.3　相对分析误差

RPD 用于评估模型的预测能力。RPD 值越大，模型预测能力越好，当 RPD<2.0 时，模型效果不理想；当 2.0<RPD<3.0 时，模型具有较高的可靠性，可用于大量样品筛选工作[23]；当 RPD>3.0 时，模型能精准预测所测成分含量，可用于预测分析[24]。计算公式见式(9)：

$$\text{RPD}=\frac{1}{\sqrt{(1-R^2)}} \tag{9}$$

2　结果与讨论

2.1　高粱化学值的测定结果

高粱中的单宁(干基)含量和水分含量的统计特征见表 1。本研究中，单宁含量最低为 0.47%，最高为 3.21%，可以认为高粱单宁含量分布具有一定代表性。该结果与文献[25]测得的高粱单宁含量分布(0.24%~4.76%)有一定的差异，这可能与高粱品种、土壤肥水和种植模式的差异有关[5]。

表1　高粱中的单宁和水分含量统计特征

Table 1　Statistical characteristics of tannin and water content in sorghum

指标 Index	样品个数 Number of sample	平均值 Mean/%	最大值 Max/%	最小值 Min/%	标准差 Standard deviation/%
单宁 Tannin	305	1.75	3.21	0.47	0.40
水分 Moisture	305	6.79	8.81	5.06	0.78

2.2　高粱的近红外光谱图

高粱完整籽粒的近红外漫反射光谱如图 1 所示，在 10075、8316、6821、5766、5186、4700、4321 和 4008 cm^{-1} 处出现吸收峰。其中，10075 cm^{-1} 处为酚中 O—H 的二级倍频吸收峰；8316 cm^{-1} 附近的吸收峰与甲基和亚甲基中的 C—H 的二级倍频有关；6821 cm^{-1} 附近的吸收峰与醇中氢键键合的 O—H 的一级倍频相关；5766 cm^{-1} 处的吸收峰为与芳环相连的甲基 C—H 的反对称和对称伸缩振动一级倍频的吸收峰；5186 cm^{-1} 处为酚和醇中 O—H 的合频和 C═O 的二级倍频的组合频吸收峰；4700 cm^{-1} 处的吸收峰为芳烃 C—C 伸缩振动和 C—H 伸缩振动的组合频吸收峰；4321 cm^{-1} 处为甲基、亚甲基和芳烃 C—H 的组合频吸收峰；4050 cm^{-1} 处为苯环 C—H 的伸缩振动和弯曲振动的组合频吸收峰。

2.3 预处理方法的选择

光谱中除了有用的化学信息外,还包含着大量的噪声和无关信息,在前期预实验基础上,本研究采用多种光谱预处理方法,包括 Det、SNV、去趋势组合标准正态变换(Det+SNV)以及去趋势组合 SG 平滑(Det+SG)分别对光谱进行预处理。Det 主要用于消除光谱的基线漂移,通过使用原始光谱减去多项式拟合出一条趋势线,对光谱进行去趋势处理; SG 平滑适用于消除不规则的随机噪声,通过选择合适的窗口大小和多项式阶数对光谱数据进行平滑处理,本研究选择窗口大小为 21,多项式阶数为 2; SNV 用于消除样本颗粒大小、表面散射光以及光程变化等对近红外光谱的影响,通过单个样本光谱的标准偏差修正光谱的变化[26]。为了消除光谱的基线漂移和噪声的影响,将 Det 与 SNV 和 SG 结合,得到 SNV、Det、Det+SNV 和 Det+SG 4 个光谱预处理结果(图 2)。经预处理的光谱图像相比原始光谱呈现出较高的光谱平滑度和集中性, Det+SG(图 2D)较 Det(图 2B)的噪声明显减少, 7000~7500 cm^{-1} 和 5000~5500 cm^{-1} 处的光谱噪声波动消失,呈现出更高的平滑度。

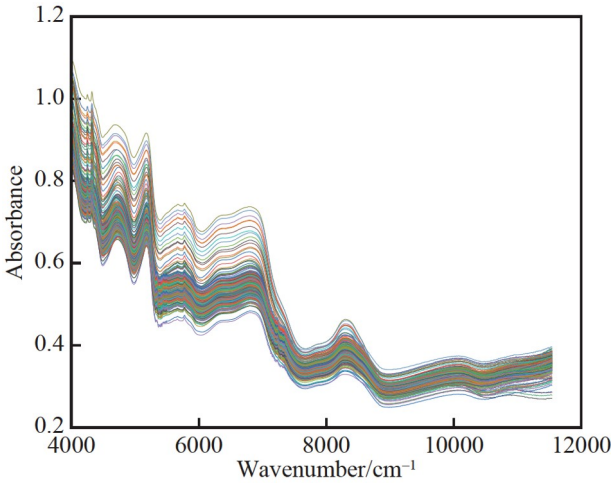

图1 高粱的近红外漫反射光谱图

Fig.1 Near infrared diffuse reflectance spectra of sorghum

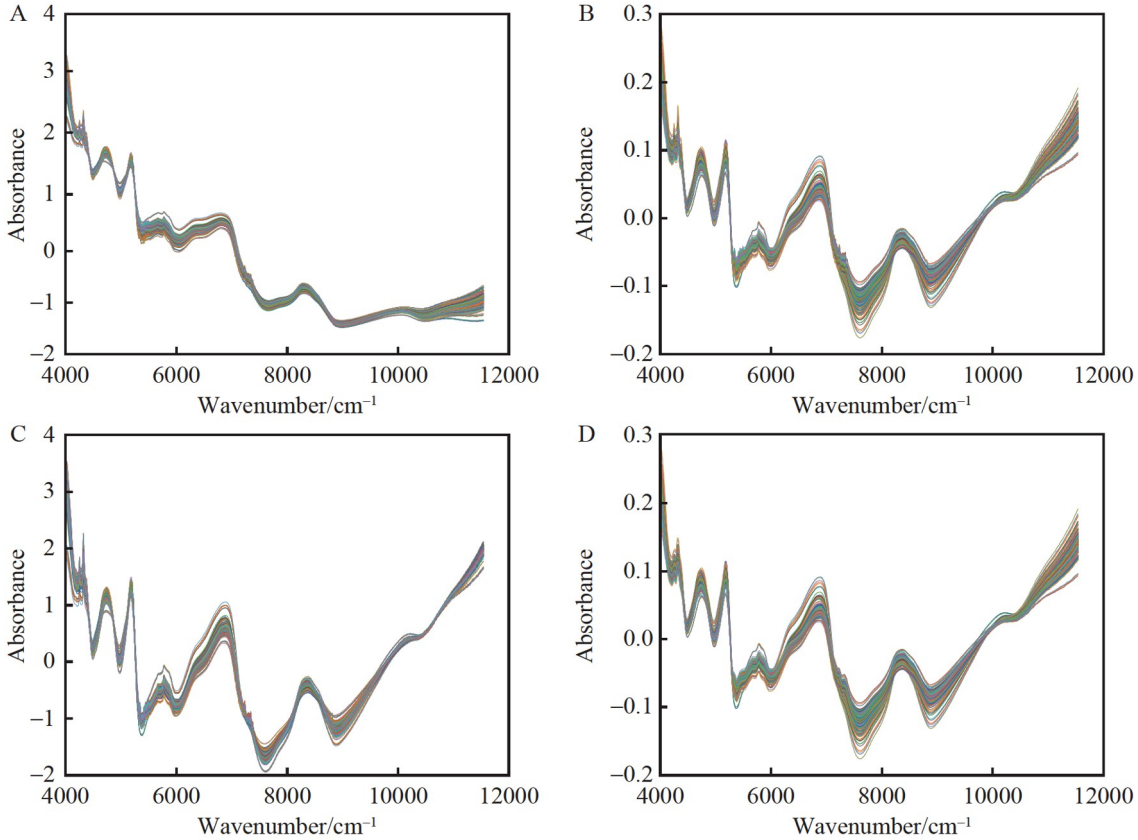

图2 采用不同方法预处理后的近红外光谱图:(A) 标准正态变换(SNV);(B) 去趋势(Det);(C) Det+SNV;(D) Det+Savitzky-Golay(SG)平滑

Fig.2 Near-infrared spectra after preprocessed by different methods: (A) Standard normal variation (SNV); (B) Detrending (Det); (C) Det+SNV; (D) Det and Savitzky-Golay (SG) smoothing

2.4　异常值的剔除

为了避免在实验过程中因操作误差造成的光谱异常或单宁含量检测结果异常对模型的不利影响，以预处理后的光谱为输入，使用 MCCV 结合 GPR 模型对样品进行异常值剔除，在建模 1000 次后，得到残差方差-均值图(图 3)。本研究取残差均值和方差最大的 15%的样本的残差均值的平均值和残差方差的平均值作为阈值，将样本残差均值或残差方差大于阈值的样本识别为异常样本。

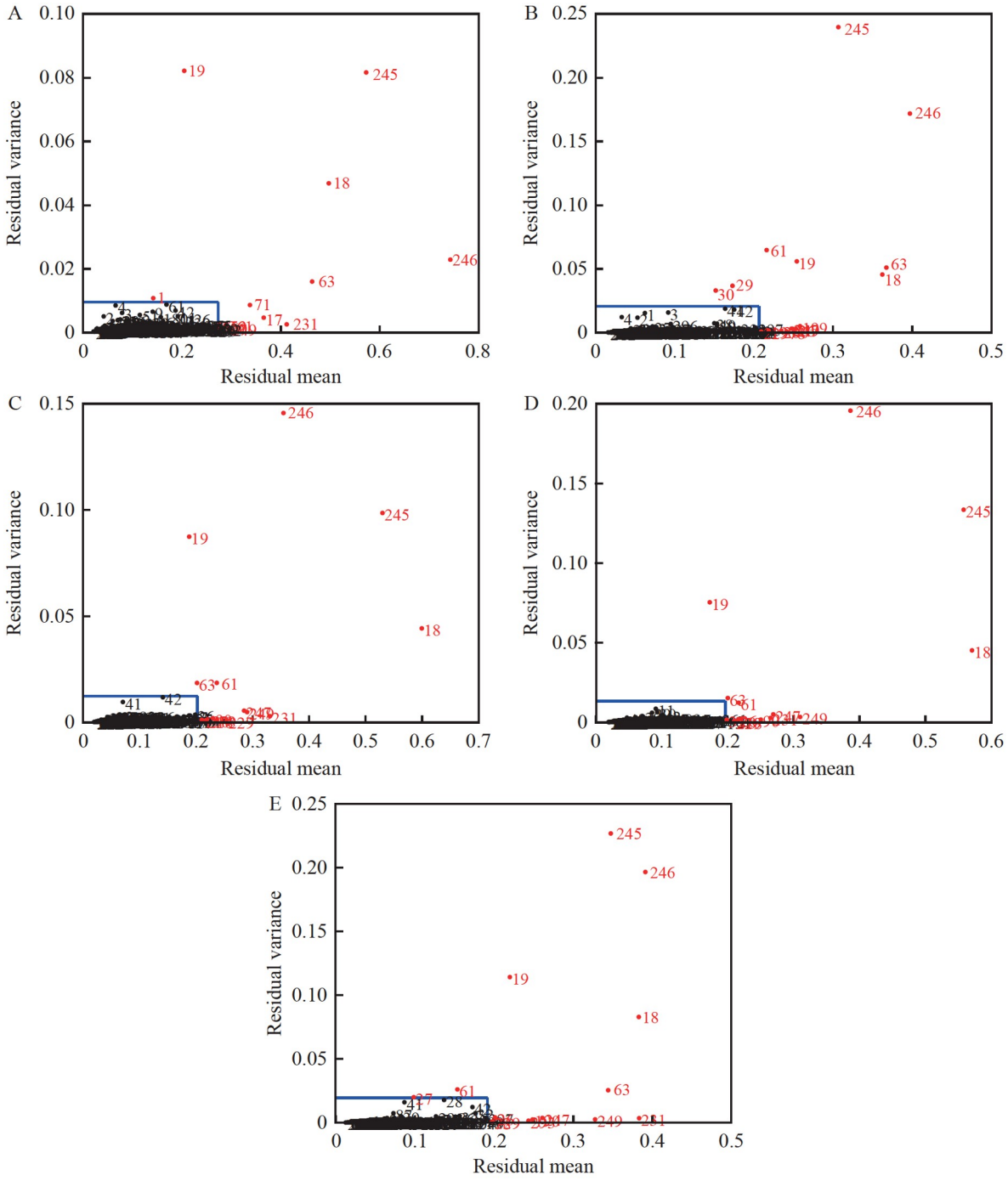

图3　(A) 无预处理、(B) SNV 预处理、(C) Det 预处理、(D) Det+SG 平滑预处理和(E) Det+SNV 的残差方差-均值图

Fig.3　Residual variance-mean plot of (A) no preprocessing, (B) SNV preprocessing, (C) Det preprocessing, (D) Pet+SG smoothing preprocessing, and (E) Det-SNV preprocessing

2.5 特征波长选择

近红外光谱数据维度较高,包含有大量冗余信息。建模前进行特征选择,可以减少冗余信息,降低维度,减少计算复杂度,缩短模型训练时间,从而更好地利用近红外光谱数据进行分析和预测。

UVE 最初由 Centner 等[27]提出,是一种基于回归系数建立的波长变量选择算法,将回归系数的稳定性值作为波长变量重要性的衡量指标。首先对光谱数据中波长变量的稳定性值进行评价,然后根据每个波长的稳定性值(均值/方差)剔除对模型没有贡献的变量。相比于其它特征波长选择算法, UVE 能够更好地处理存在异常值和噪声的数据,提高特征选择的稳定性和可靠性。该方法不依赖于具体的统计分布假设,适应性强,并具有较好的鲁棒性。

UVE 建模方法采用 PLS 回归, PLS 模型的主成分数设置为 15,阈值为 0.995,采用留一法交叉验证。图 4 为各变量回归系数统计分布,蓝色实线为光谱变量矩阵的稳定性值,红色实线为随机噪声,红色水平虚线为通过阈值 0.995 选出的临界值,绝对值大于临界值的波长为所选择的波长。图 5 为 UVE 所选波长(经 Det-SG 和剔除异常样本的光谱),共选取了 662 个特征波长,占总波长的 35.88%。每个波峰附近都有所选择的波长,在 4 个波峰(6821、5766、4700 和 4321 cm^{-1})附近尤为聚集,其中, 6821 cm^{-1} 附近的吸收峰与醇中氢键键合的 O—H 的一级倍频相关, 5766 cm^{-1} 处为与芳环相连的甲基 C—H 的反对称和对称伸缩振动一级倍频的吸收峰; 4700 cm^{-1} 处的吸收峰为芳烃 C—C 伸缩振动和 C—H 伸缩振动的组合频吸收峰; 4321 cm^{-1} 处为甲基、亚甲基和芳烃 C—H 的组合频吸收峰。

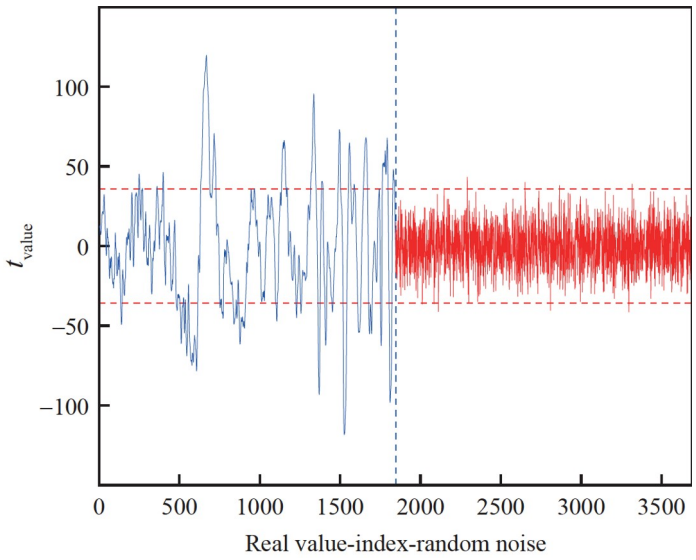

图4 无信息变量消除法(UVE)各变量回归系数统计分析

Fig.4 Statistical analysis of regression coefficients for various variables in elimination of uninformative variables (UVE)

2.6 模型建立

数据预处理、剔除异常样本和特征波长选择后建立 GPR 模型,核函数为 matern32 核,重复建模 1000 次取平均值,以保证评价指标的准确性,建模结果如表 2 所示。

表 2 中 GPR 模型的 RPD 值均大于 3,因此,模型均可用于预测单宁含量。其中, Det-SG 和 Det-SNV 最优秀,并且两个模型的 R_P^2 和 RPD 值十分接近。Det-SG 的决定系数 R_P^2 略高于 Det-SNV,但 RPD 略低于 Det-SNV,这表明 Det-SG 相对于 Det-SNV 更加稳定,模型预测性能的波动较小。在多次建模过程中, Det-SG 比 Det-SNV 更加稳定,能够提供更一致的预测精度,故建立在 Det-SG 预处理和 UVE 进行特征波长选择结果的 GPR 为最优模型,其建模集决定系数(R_C^2)和 R_P^2 分别为 0.9979 和 0.9529,建模集和预测集 R^2 的比值为 1.05,不存在过拟合和欠拟合, RPD 值为 4.8453,大于 3,说明模型可以精确预测单宁含量。

对比表 2 可知,未经过光谱预处理的模型有轻微的过拟合现象,因此,光谱预处理能明显提升模型性

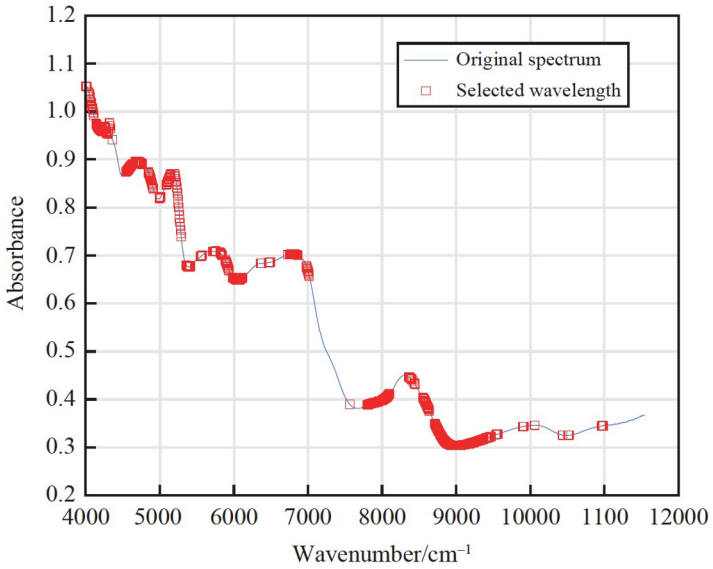

图5　UVE 所选波长图

Fig.5　Selected wavelength of UVE

表2　高斯过程回归(GPR)建模结果

Table 2　Gaussian process regression (GPR) modeling results

预处理方法 Preprocessing method	建模集决定系数 R_C^2	建模集均方根误差 RMSEC	预测集决定系数 R_P^2	预测集均方根误差 RMSEP	相对分析误差 RPD
无 Without	0.9643	0.0694	0.8817	0.1221	3.0042
SNV	0.9985	0.0138	0.9471	0.0823	4.6057
Det	0.9993	0.0092	0.9472	0.0811	4.7165
Det+SG	0.9979	0.0164	0.9529	0.0779	4.8453
Det+SNV	0.9992	0.0090	0.9516	0.0776	4.8895

注(Note)：R_C^2, 建模集决定系数 (Model set determination coefficient)；RMSEC, 建模集均方根误差 (Root mean square error of modeling set)；R_P^2, 预测集决定系数 (Prediction set determination coefficient)；RMSEP, 预测集均方根误差 (Root mean square error of prediction set)；RPD, 相对分析误差 (Relative percent deviation)。

能[28]，经过光谱预处理后，模型的 RPD 大于 4，模型的过拟合现象明显改善。建立在两种预处理方法上的模型性能与单一预处理方法基础上的模型相比，性能仅略有提升[29]。4 种不同预处理方法对单宁含量预测模型影响不大。文献[11]建立的基于高粱籽粒的单宁含量预测模型 MSC-MCUVE-PLSR 的 R_P^2 为 0.8841，而本研究建立的 Det-SG-UVE-GPR 模型的 R_P^2 为 0.9529，对于整粒高粱单宁的预测性能有明显提升，可以更精准地预测整粒高粱中的单宁含量。

2.7　GPR 模型与其它模型的比较

PLS 是化学计量学的经典方法，在近红外光谱建模过程中广泛应用[30-32]；此外，近年来，SVM 在近红外光谱建模方面取得了较好的效果[33-34]。为了说明 GPR 模型的优势，本研究选取在近红外光谱和机器学习算法方面使用较多的 PLS 和 SVM 建立回归模型，并与 GPR 模型进行对比。其中，PLSR 模型通过计算 RMSECV 的最小值选择最优主成分数，并进行建模预测；SVR 模型使用径向基核函数，通过网格寻优法寻找最优参数，并进行建模预测，重复建模 200 次取平均值，建模结果见表 3。

对比表 2 和表 3 可知，PLSR 模型的 RPD 值在 2.40~2.75 之间，均小于 3，最优 PLSR 模型的 R_P^2 和 RPD 分别为 0.8595 和 2.7493，对于预测高粱单宁含量略显牵强。SVR 模型的性能全面优于 PLSR 模型，并且性能接近 GPR 模型(表 3)。预处理可以明显提升模型的预测精度，经预处理后，SVM 模型 RPD 值在 3.31~3.41 之间，均大于 3，因此，建立在预处理基础上的 SVM 模型可用于预测高粱单宁含量。在 SVM 模型性能参数方面，在 4 种不同的预处理方法中，Det-SG 略显优势，其 R_P^2 和 RPD 分别为 0.9022 和

表3 不同预处理方法下偏最小二乘回归(PLSR)和支持向量机回归(SVR)模型结果

Table 3 Results of partial least squares regression (PLSR) and support vector machine regression (SVR) models under different preprocessing methods

预处理方法 Preprocessing methods	建模方法 Modeling methods	建模集决定系数 R_C^2	建模集均方根误差 RMSEC	预测集决定系数 R_P^2	预测集均方根误差 RMSEP	相对分析误差 RPD
无 Without	PLSR	0.8930	0.1207	0.8177	0.1501	2.4078
SNV	PLSR	0.9067	0.1141	0.8225	0.1524	2.4503
Det	PLSR	0.9190	0.1058	0.8300	0.1494	2.5016
Det+SG	PLSR	0.9221	0.1037	0.8595	0.1364	2.7493
Det+SNV	PLSR	0.9140	0.1084	0.8352	0.1436	2.5458
无 Without	SVR	0.9145	0.1059	0.8237	0.1492	2.5172
SNV	SVR	0.9520	0.0815	0.9001	0.1139	3.3181
Det	SVR	0.9540	0.0801	0.9001	0.1128	3.3486
Det+SG	SVR	0.9487	0.0839	0.9022	0.1130	3.3746
Det+SNV	SVR	0.9526	0.0808	0.8930	0.1132	3.2016

3.3746,但其预测精度仍低于 GPR 模型,建模过程耗时也远长于 GPR 模型。

综合表 2 和表 3 可知,在相同的光谱预处理情况下,3 种模型性能从高到低依次为 GPR>SVR>PLSR,这与文献[16-17]的研究结果类似。基于相同预处理方法(Det-SG)的 GPR 模型的 RPD 比 PLSR 模型提升了 76.24%,比 SVR 模型提升了 43.58%。这种预测准确度的不同可能是由于高粱近红外光谱受到多种因素(如复杂的化学成分、光谱峰重叠、仪器和环境等)的影响,这些因素会导致光谱数据间存在复杂的非线性关系和复杂数据分布,而 GPR 模型在处理非线性关系和复杂数据分布方面更具有优势,采用核函数适应数据的非线性特征,不对数据进行特定假设。相对于 PLSR 和 SVR,GPR 模型可以提供对预测的不确定性估计,可更好地处理非相关的数据,提高预测性能。相比之下,PLSR 的建模能力受限于其线性假设,对于非线性关系的建模能力有限。SVR 虽然可用于解决小样本、非线性和高维数据空间模式识别等问题,但对数据的分布和特征的敏感度较高,如果数据的分布不符合其假设或者特征不显著,可能会影响其预测效果。所有 PLSR 模型的 RPD 均小于 3,并且多种预处理方法相差不大,由此可见,预处理方法对于高粱单宁的 PLSR 模型性能提升不明显,而对 GPR 和 SVR 性能均有较大提升,这再次表明了 PLSR 模型建模能力有限。

3 结论

采用近红外光谱仪器结合化学计量学方法,建立了基于高粱完整籽粒的单宁的 GPR 预测模型,对比 PLSR 和 SVR 两种常用建模方法,本研究建立的 GPR 高粱单宁预测模型准确度最高,建模结果有显著优势,其 RPD 值较 PLSR 和 SVR 分别提升了 76.24%和 43.58%,模型性能大幅提升,可用于高粱单宁含量的快速检测。相比于传统的人工经验判别法和实验室化学方法,本方法检测方便快捷且准确,能够更好地为高粱中单宁的快速检测提供技术支持。

References

[1] ALFIERI M, CABASSI G, HABYARIMANA E, QUARANTA F, BALCONI C, REDAELLI R. J. Near Infrared Spectrosc., 2019, 27(1): 46-53.

[2] LI Shun-Guo, LIU Meng, LIU Fei, ZOU Jian-Qiu, LU Xiao-Chun, DIAO Xian-Min. Chin. Agric. Sci., 2021, 54(3): 471-482.
李顺国, 刘猛, 刘斐, 邹剑秋, 陆晓春, 刁现民. 中国农业科学, 2021, 54(3): 471-482.

[3] JING Xiao-Lan, LIU Qing-Shan, PING Jun-Ai, CHENG Qing-Jun, BAI Wen-Bin, ZHANG Fu-Yue. J. Shanxi Agric. Sci., 2014, 42(6): 621-624.
景小兰, 柳青山, 平俊爱, 程庆军, 白文斌, 张福跃. 山西农业科学, 2014, 42(6): 621-624.

[4] FENG Hai-Zhi, LI Long, WANG Dong, ZHANG Kai, FENG Miao, SONG Hai-Jiang, LI Rong, HAN Ping. Spectrosc. Spectral Anal. , 2023, 43(1): 16-24.
冯海智, 李龙, 王冬, 张凯, 冯淼, 宋海江, 李荣, 韩平. 光谱学与光谱分析, 2023, 43(1): 16-24.
[5] NAN Huai-Lin, CAO Xiong, LIANG Xiao-Hong, LIU Jing, ZHANG Rui-Dong, WANG Song-Yu. Bull. Agric. Sci. Technol. , 2021, 592(4): 7-9.
南怀林, 曹雄, 梁晓红, 刘静, 张瑞栋, 王颂宇. 农业科技通讯, 2021, 592(4): 7-9.
[6] BHATT R S, SARKAR S, SHARMA P, SONI L, SAHOO A. Small Ruminant Res. , 2023, 218: 106876.
[7] BUTLER L G, RIEDL D J, LEBRYK D G, BLYTT H J. J. Am. Oil Chem. Soc. , 1984, 61(5): 916-920.
[8] ZHENG Xue-Ling, WANG Xu-Juan, HAN Xiao-Xian. J. Henan Univ. Technol. (Nat. Sci. Ed.), 2020, 41(2): 132-140.
郑学玲, 王旭娟, 韩小贤. 河南工业大学学报(自然科学版), 2020, 41(2): 132-140.
[9] CHENG J, SUN J, YAO K, XU M, DAI C. Meat Sci. , 2023, 201: 109196.
[10] XU Guang-Tong, YUAN Hong-Fu, LU Wan-Zhen. Spectrosc. Spectral Anal. , 2000, 20(2): 134-142.
徐广通, 袁洪福, 陆婉珍. 光谱学与光谱分析, 2000, 20(2): 134-142.
[11] YU Song-Bai, HUANG Zhang-Jun, WU Qi-Xiao, JIA Jun-Jie, WANG Hong-Mei, WANG Song-Tao, SHEN Cai-Hong. Sci. Technol. Food Ind. , 2023, 44(10): 311-319.
余松柏, 黄张君, 吴奇霄, 贾俊杰, 王红梅, 王松涛, 沈才洪. 食品工业科技, 2023, 44(10): 311-319.
[12] LIU Min-Xuan, WANG Yun-Wen, HAN Jian-Guo. Chin. J. Anal. Chem. , 2009, 37(9): 1275-1280.
刘敏轩, 王赟文, 韩建国. 分析化学, 2009, 37(9): 1275-1280.
[13] DYKES L, HOFFMANN L, PORTILLO-RODRIGUEZ O, ROONEY W L, ROONEY L W. J. Cereal Sci. , 2014, 60(1): 138-142.
[14] ZHANG N, LIU X, JIN X, LI C, WU X, YANG S, NING J, YANNE P. Food Chem. , 2017, 237: 811-817.
[15] HUANG Yin-Yan, YU Chao, HUANG Wen-Xin, QIN Zhi-Jun, BI Le-Ming, YANG Lin. Mod. Electr. Power, 2021, 38(6): 664-673.
黄银燕, 于超, 黄文新, 覃智君, 毕乐明, 杨琳. 现代电力, 2021, 38(6): 664-673.
[16] LI Yuan, ZHANG Wen-Bo, CHEN Xiao-Lin, LI Han, ZHANG Guan-Jun. Spectrosc. Spectral Anal. , 2022, 42(10): 3073-3078.
李元, 张文博, 陈晓琳, 李含, 张冠军. 光谱学与光谱分析, 2022, 42(10): 3073-3078.
[17] ZHANG Tao, WANG Yang, WANG Yan-Zi, ZHANG Jian, WANG Yu-Hang, MA Rui. J. Chongqing Univ. Technol. (Nat. Sci.), 2023, 37(11): 1-9.
张韬, 王阳, 王言子, 张健, 王宇航, 马瑞. 重庆理工大学学报(自然科学), 2023, 37(11): 1-9.
[18] GB/T 15686—2008. Sorghum-Determination of Tannin Content. National Standards of the People's Republic of China.
高粱 单宁含量的测定. 中华人民共和国国家标准. GB/T 15686—2008.
[19] GB 5009.3—2016. National food Safety Standard Determination of Moisture in Foods. National Standard of the People's Republic of China.
食品安全国家标准 食品中水分的测定. 中华人民共和国国家标准. GB 5009.3—2016.
[20] XIAO H, CHEN Z, YI S, LIU J. Vib. Spectrosc. , 2023, 129: 103595.
[21] DAI X, ANDANI H T, ALIZADEH A, ABED A M, SMAISIM G F, HADRAWI S K, KARIMI M, SHAMSBORHAN M, TOGHRAIE D. Eng. Appl. Artif. Intel. , 2023, 122: 106107.
[22] CHEN Yue-Yao, XIA Jing-Jing, WEI Yun, XU Wei-Xin, MAO Xin-Ran, CHEN Yue-Fei, MIN Shun-Geng, XIONG Yan-Mei. Chin. J. Anal. Chem. , 2023, 51(3): 454-462.
陈玥瑶, 夏静静, 韦芸, 徐惟馨, 毛欣然, 陈月飞, 闵顺耕, 熊艳梅. 分析化学, 2023, 51(3): 454-462.
[23] CHEN G L, ZHANG B, WU J G, SHI C H. Anim. Feed Sci. Technol. , 2011, 165(1-2): 111-119.
[24] QUINTELAS C, RODRIGUES C, SOUSA C, FERREIRA E C, AMARAL A L. Food Chem. , 2024, 435: 137607.
[25] PAN L, LI W, GU X M, ZHU W Y. Anim. Feed Sci. Technol. , 2022, 292: 115419.
[26] ZHU Si-Cong, GAO Xi-Ya, ZHANG Zhu-Shan-Ying, CAO Hui-Min, ZHENG Dong-Yun, ZHANG Li, XIE Qin-Lan, SA Ji-Ming. Chin. J. Anal. Chem. , 2022, 50(9): 1415-1429.
朱思聪, 高西娅, 张朱珊莹, 曹汇敏, 郑冬云, 张莉, 谢勤岚, 撒继铭. 分析化学, 2022, 50(9): 1415-1429.
[27] CENTNER V, MASSART D L, NOORD O E D, JONG S D, VANDEGINSTE B M, STERNA C. Anal. Chem. , 1996, 68(21): 3851-3858.
[28] WANG Chao, LIU Yan, XIA Zhen-Zhen, WANG Qiao, DUAN Shuo. Spectrosc. Spectral Anal. , 2023, 43(1): 156-161.
王超, 刘言, 夏珍珍, 王桥, 段烁. 光谱学与光谱分析, 2023, 43(1): 156-161.
[29] LI Quan-Lun, CHEN Zheng-Guang, JIAO Feng. Spectrosc. Spectral Anal. , 2023, 43(4): 1030-1036.
李泉伦, 陈争光, 焦峰. 光谱学与光谱分析, 2023, 43(4): 1030-1036.
[30] LI Z, SONG J, MA Y, YU Y, HE X, GUO Y, DOU J, DONG H. Food Chem.: X, 2023, 17: 100539.

[31] QIAN L, LI D, SONG X, ZUO F, ZHANG D. J. Food Compos. Anal. , 2022, 105: 104203.
[32] ROSSI G B, LOZANO V A. LWT-Food Sci. Technol. , 2020, 126: 109290.
[33] ZHONG Y Q, LI J Q, LI X L, DAI S Y, SUN F. Vib. Spectrosc. , 2023, 127: 103556.
[34] LI J, DENG J, BAI X, DA GRACA NSELEDGE MONTEIRO D, JIANG H. Spectrochim. Acta, Part A, 2023, 303: 123208.

Rapid Detection of Sorghum Tannin Content Based on Near-Infrared Spectroscopy and Gaussian Process

ZHAO Jin-Yi[1], CHEN Zheng-Guang[*1], YI Shu-Juan[2]
[1](*College of Information and Electrical Engineering*, *Heilongjiang Bayi Agricultural University*, *Daqing* 163000, *China*)
[2](*Engineering College*, *Heilongjiang Bayi Agricultural University*, *Daqing* 163000, *China*)

Abstract The tannin content of sorghum seeds had a significant impact on the wine's quality during the brewing process. Additionally, when used as a feed ingredient, the tannin content had a major impact on feed consumption. Thus the tannin content of sorghum has a substantial impact on its quality and application. To quickly and nondestructively determine the tannin content of sorghum, near-infrared spectroscopy was combined with chemometrics in this study, which eliminated the need for time-consuming and costly conventional approaches. Following the spectra's preprocessing, anomalous samples were removed by using a combination of Gaussian process regression (GPR) and Monte Carlo cross-validation (MCCV). The sample set was then randomly divided into a modeling set and a prediction set, with feature wavelength selection carried out using the elimination of uninformative variables (UVE) method. Subsequently, a GPR model was developed, and its performance was compared with partial least squares regression (PLSR) and support vector machine regression (SVR) models. The results indicated that the GPR model outperformed the PLSR and SVR models in all aspects. The optimized GPR model, generated following pre-processing process such as Detrending and Savitzky-Golay smoothing, elimination of anomalous samples, and selection of feature wavelengths, demonstrated superior performance, with model set determination coefficient (R_c^2), prediction set determination coefficient (R_p^2), and relative percent deviation (RPD) values of 0.9979, 0.9529, and 4.8453, respectively. These findings validated the effectiveness of the GPR regression model, which integrated near-infrared spectroscopy with chemometrics, for the rapid and non-destructive detection of sorghum tannins.

Keywords Near-infrared spectroscopy; Chemometrics; Sorghum tannin; Gaussian process; Rapid non-destructive testing

(Received 2023-11-16; accepted 2024-05-20)

Supported the by National Natural Science Foundation of China (No. 52275246) and the Basic Research in "Sanzong" Program of Heilongjiang Bayi Agricultural University (No. ZRCPY202214).