江苏农业学报(*Jiangsu J.of Agr.Sci.*),2019,35(4):960~964
http://www.jsnyxb.com

张津源,张德贤,张　苗. 基于连续投影算法的小麦蛋白质含量近红外光谱预测分析[J].江苏农业学报,2019,35(4):960-964.
doi:10.3969/j.issn.1000-4440.2019.04.030

基于连续投影算法的小麦蛋白质含量近红外光谱预测分析

张津源,　张德贤,　张　苗
(河南工业大学信息科学与工程学院,河南 郑州 450001)

摘要:　为了快速无损检测分析小麦蛋白质含量,构建近红外光谱最优小麦蛋白质定量检测分析模型。利用一阶 S-G 平滑算法+SNV 算法对光谱进行预处理。使用连续投影算法(Successive projections algorithm, SPA)提取光谱中的特征波段点,使全谱图的 141 个波段点降低到 17 个特征波段点。在选择的 17 个特征波段点基础上分别建立偏最小二乘回归(Partial least regression, PLSR)模型、支持向量机(Support vector machine, SVM)模型、多元线性回归(Multiple linear squares regression, MLR)模型和主成分回归(Principal component regression, PCR)模型。在构建的 4 种小麦蛋白质含量预测模型中,MLR 预测分析模型的验证集均方根误差(*RMSEV*)和校正集均方根误差(*RMSEC*)最小,验证集相关系数(r_v)和校正集相关系数(r_c)最大,其r_v=0.968,r_c=0.976,*RMSEV*=0.300,*RMSEC*=0.275。因此,相比于其他 3 种检测模型,建立的 MLR 小麦蛋白质含量检测模型最优,稳定性和精确性最高。

关键词:　小麦;蛋白质含量;近红外光谱;检测模型;特征波段点
中图分类号:　S126　　**文献标识码**:　A　　**文章编号**:　1000-4440(2019)04-0960-05

Prediction and analysis of wheat protein content by near-infrared spectroscopy based on successive projections algorithm

ZHANG Jin-yuan,　ZHANG De-xian,　ZHANG Miao
(*College of Information Science and Engineering*, *Henan University of Technology*, *Zhengzhou* 450001, *China*)

Abstract:　In order to detect the wheat protein content quickly and non-destructively, an optimal quantitative analysis model of wheat protein content was constructed. The first derivative S-G smoothing algorithm and the standard normal variable (SNV) were used to preprocess the spectrum. The successive projections algorithm (SPA) was used to extract the characteristic band points in the spectrum, so that 141 band points of the full spectrum were reduced to 17 characteristic band points. Partial least square regression (PLSR) model, support vector machine (SVM) model, multiple linear regression (MLR) model and principal component regression (PCR) model were established on the basis of 17 selected characteristic band points. In the four wheat protein content prediction models, the MLR model had the smallest root-mean-square error of the validation set (*RMSEV*), the smallest root-mean-square error of the calibration set (*RMSEC*), the largest correlation coefficient of the validation set (r_v) and the largest correlation coefficient of calibration set (r_c). The correlation coefficients of validation set and calibration set were 0.968 and 0.976. The root mean square error of validation set and calibration set were 0.300 and 0.275. Compared with the other three detection models, the MLR detection model is the best, and the stability and accuracy are the highest.

Key words:　wheat; protein content; near infrared spectroscopy; detection model; characteristic band points

收稿日期:2018-07-04
基金项目:国家科技支撑计划项目(2013BAD17B04);河南省科技厅自然科学项目(172106000013);粮食信息处理与控制教育部重点实验室开放基金课题(KFJJ2016102)
作者简介:张津源(1992-),男,河南周口人,硕士研究生,主要研究方向为模式识别与智能信息处理。(E-mail)15303816835@163.com
通讯作者:张德贤,(E-mail)zdx@haut.edu.cn

在小麦生产、交易、食品加工及储藏过程中,为了更好地监控和掌握小麦品质,需要检测小麦蛋白质含量。现有的小麦蛋白质含量检测技术因分析时间长,需要大量化学试剂和昂贵的化学仪器,无法满足大批量小麦品质检测的要求。吕程序等[1]利用多次统计自适应加权算法对小麦原始光谱进行优选,选出12个小麦蛋白质含量敏感波段点,利用偏最小二乘算法建立小麦蛋白质含量检测模型,得到预测集相关系数(r_p)为0.961,验证集均方根误差(*RMSEV*)为0.369。张松等[2]利用SVM算法对经连续投影后的光谱信息进行建模分析,实现了对小麦蛋白质含量的快速检测。宋雨宸等[3]用蒙特卡洛算法结合变量组合的方法对小麦光谱信息进行变量提取,对提取的10个小麦蛋白质含量特征变量进行建模分析。毛晓东等[4]结合近红外光谱技术对水稻品种进行判别分析,识别率达到100%。为了更好地得到近红外光谱的矫正检测模型,钱海波等[5]利用连续投影算法提取敏感波段点。吴静珠[6]等建立的Si-cPLS模型相关系数(r)为0.967,预测集均方根误差(*RMSEP*)为0.08,说明Si-cPLS模型的稳定性好,预测精度高。王赋腾等[7]利用近红外光谱仪对125个小麦面粉样品进行了近红外光谱扫描,对得到的125个光谱数据分别用12种方法进行预处理,其中用矢量归一法+S-G滤波平滑法得到的效果最好,说明光谱预处理在对小麦面粉光谱数据分析中具有很大的影响。张平平等[8]利用近红外光谱技术检测了小麦谷蛋白大聚体含量。本研究利用一阶S-G平滑算法[9]+SNV算法[10]对光谱进行预处理,用连续投影算法(SPA)[11-13]提取光谱中的特征波段点,建立稳定和准确的小麦蛋白质含量检测模型。

1 材料与方法

1.1 试验材料

小麦样品主要来源于河南省农业科学院、中储粮周口直属库、陕西咸阳面粉厂等。选取50个品种共212份小麦样品,去除杂质,晾晒并编号,存放在25 ℃恒温箱中。

1.2 仪器

DA7250近红外光谱分析仪,瑞典Perten公司产品,波长范围为950~1 650 nm,1 s扫描100次,带宽5 nm。全自动凯氏定氮仪,济南海能仪器股份有限公司产品。

1.3 样品处理

每份小麦样品取2 g进行制粉,样品消化后,使用全自动凯氏定氮仪测定蛋白质含量。在25~30 ℃下,用近红外光谱分析仪对每份小麦籽粒样品进行非接触旋转扫描,测出其光谱矩阵,并对每份小麦样品进行第2次扫描,作为平行校正,输出近红外光谱数据。

1.4 样本集的划分

用SPXY(Sample set partitioning based on joint X-Y distances)算法划分样本集。SPXY算法是在K-S算法的基础上改进而来的,在K-S的基础上引入了物化指标y,同时也考虑了样品的光谱反射率x,有效地覆盖了光谱矩阵的多维空间。其中校正集样品数量为144,蛋白质含量最大值为17.60%,最小值9.13%,均值12.80%,标准差1.42%;预测集样品数量为68,蛋白质含量最大值为17.20%,最小值9.16%,均值12.91%,标准差1.32%。

1.5 光谱数据预处理

对小麦样品的近红外光谱信息和用化学试剂方法测得的小麦蛋白质含量化学值进行定量分析建模,其中小麦蛋白质含量测定值作为参考值。由于光谱采集过程中存在小麦样品装盘不均匀,背景干扰,以及近红外光的散射、白噪声、随机噪声等因素[14-15],建模前需要对原始近红外光谱进行预处理。利用一阶S-G平滑算法+SNV算法对光谱数据进行预处理。通过对光谱数据求一阶导数和二阶导数,解决正常量级和正线型基线漂移与平移问题,进而提高光谱的分辨率与信噪比,将重叠峰区分开,降低信息的复杂度。先利用一阶导数S-G平滑算法对采集到的小麦光谱数据进行预处理,设置的差值宽度为2,平滑点数为5。为了更好地突出峰与峰之间的差异性,增强其特征性,降低甚至消除固体颗粒散射的干扰,用SNV算法对光谱数据进行再运算。

1.6 连续投影算法

连续投影算法(SPA)是一种新的选择特征波段点方法。对于光谱矩阵$X_{N\times M}$(N为样本数,M为光谱变量数),选取最大的特征变量数为K,SPA步骤如下:(1)在初始迭代$t=1$时,在$X_{N\times M}$中任意选择一列向量x_j,记作$x_{l(0)}$,$l(0)$为任意选择性变量的初始位置[$l(0)=j, 1\leqslant j\leqslant M$],定义$W$为其他剩余变量位置的集合,则$W=\{j, 1\leqslant j\leqslant M, j\neq[l(0),\cdots,l(K-$

1)]}。(2) 计算剩余列向量x_j在选择向量$x_{l(t-1)}$构成的正交向量空间中的投影:$x_j = \{E - x_{l(t-1)}[x_{l(t-1)}]^T/[x_{l(t-1)}]^T x_{l(t-1)}\}x_j$,$E$为单位矩阵。(3) 提取最大向量空间投影值的变量 $\arg[\max(\|x_j\|)]$,$j \in W$,添加到选择的变量集。(4) $t=t+1$,如果$t<K$,返回(2)循环计算,直到满足条件,跳出循环,输出选择出的集合 W。

为了求得最优选择,K、$l(0)$的选择至关重要。由于变量之间存在共线性,所以 K 值一般情况下相对较小,如果设定的 K 很大,其光谱值投影将趋于 0 甚至等于 0。$l(0)$的每次选择时,用 PLS 进行循环验证分析,当交叉验证均方根误差(*RMSECV*)最小时,$l(0)$和 K 为最优选择。

1.7 数据分析

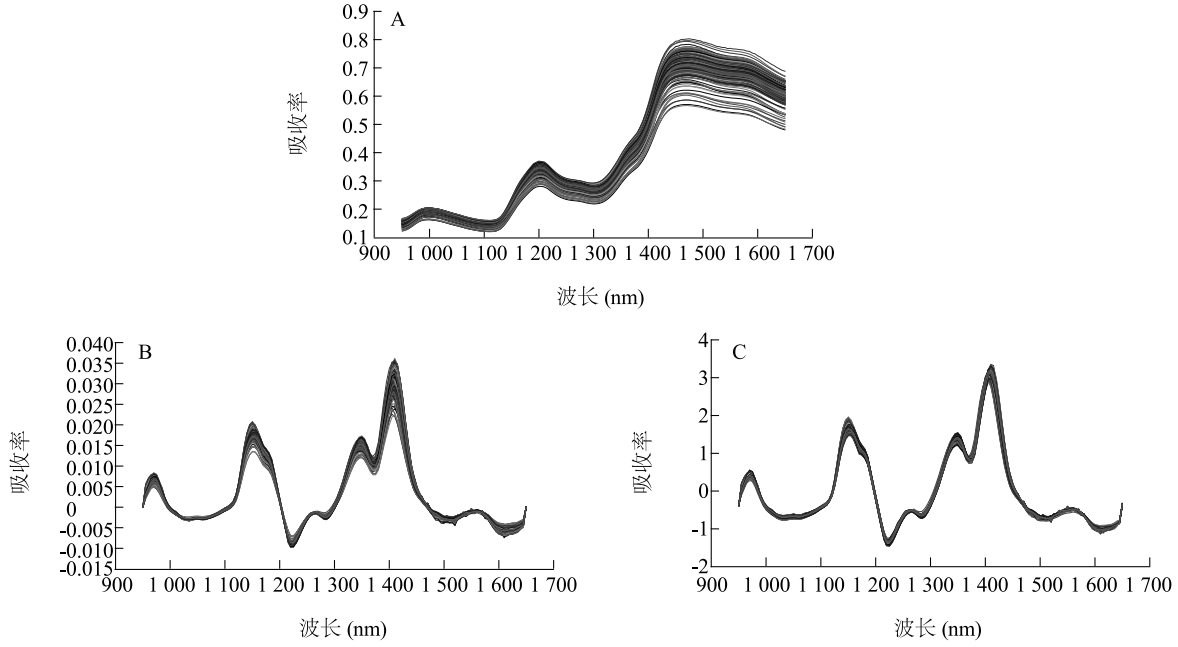
对经过连续投影算法优选后的波段点分别建立偏最小二乘回归(PLSR)、支持向量机(SVM)、多元线性回归(MLR)和主成分回归(PCR)的定量预测模型。利用交叉验证优选出最佳的波段点数。用 Matlab 2016a 对原始光谱进行特征波段点的选择和预测模型的建立,使用验证集均方根误差(*RMSEV*)、预测集均方根误差(*RMSEP*)、验证集相关系数(r_v)和校正集相关系数(r_c)对定量预测模型进行评价。

2 结果与分析

采用连续投影法(SPA)选择经一阶 S-G 平滑算法+SNV 算法预处理后的小麦近红外光谱全谱图数据,选取敏感波段点数据建立 PLSR 检测分析模型。图 1 为预处理后的全谱图。利用 SPA 法对数据进行选择,提取有效的特征光谱段,即可满足样品光谱所含信息,降低光谱信息中变量的奇异性和共线性,增强小麦蛋白质含量检测模型的精确性和稳定性,提高建模的速度和效率。经 SPA 法提取的特征波段点有 973 nm、1 005 nm、1 011 nm、1 026 nm、1 111 nm、1 122 nm、1 140 nm、1 153 nm、1 161 nm、1 173 nm、1 202 nm、1 230 nm、1 256 nm、1 362 nm、1 365 nm、1 407 nm、1 443 nm、1 474 nm、1 493 nm、1 548 nm 和 1 585 nm,使全谱图的 141 个波段点降低到 21 个特征波段点。在这个基础上再次采用SPA 法对这 21 个光谱特征波段点进行处理,得到1 026 nm、1 154 nm、1 200 nm、1 367 nm、1 407 nm、1 446 nm、1 473 nm、1 495 nm、1 548 nm 和1 586 nm,在第 1 次的基础上又减少了 11 个特征波段点,余下 10 个波段点。分别对提取的 21 个光谱波段点和余下的 10 个光谱波段点进行 PLSR 建模分析(图 2、图 3)。由图 2 可知,分析结果中校正集 *RMSEC* 为 0.312,相关系数 r_c 为 0.965,验证集 *RMSEV* 为 0.336,相关系数 r_v 为 0.960。由图 3 可知,分析结果中校正集 *RMSEC* 为 0.411,相关系数 r_c 为 0.942,验证集 *RMSEV* 为 0.433,相关系数 r_v 为 0.935。对比图 2 和图 3 可知,21 个波段点的 PLSR 模型的验证集 *RMSEV* 和校正集 *RMSEC* 都小于 10 个波段点模型的验证集 *RMSEV* 和校正集 *RMSEC*,说明 21 个光谱波段点所建模型的稳定性和精确性比 10 个光谱波段点的高;21 个光谱波段点模型的验证集相关系数和校正集的相关系数都大于 10 个波段点模型的验证集相关系数和校正集的相关系数,说明 21 个光谱波段点模型的相似度高。从全谱图提取的 21 个波段点所建模型和 10 个波段点所建模型的验证集均方根误差和校正集均方根误差都小于 0.45,说明这 2 种方法得到的 PLSR 预测分析模型都存在很高的检测能力和稳定性。

由图 2、图 3 可以看出,21 个波段点的模型要好于 10 个波段点的模型。提取的波段点数量不同,所建立模型的精确性和稳定性也不同。根据全谱中提取的2~21 个波段点的验证集均方根误差和校正集均方根误差,确定最优模型的波段点数量。从图 4 可以看出,当选择 17 个波段点建立 PLSR 预测模型时,其验证集 *RMSEV*、校正集 *RMSEC* 均达到最小值,分别为0.324 06、0.303 55,说明选择 17 个波段点所建模型最优。图 5 为选择 17 个波段点建立的 PLSR 模型,这 17 个波段点分别为 972 nm、1 007 nm、1 008 nm、1 025 nm、1 108 nm、1 125 nm、1 143 nm、1 153 nm、1 160 nm、1 169 nm、1 204 nm、1 230 nm、1 258 nm、1 360 nm、1 365 nm、1 410 nm 和 1 443 nm。

为了构建最优模型,在 17 个波段点 PLSR 模型的基础上,在同等条件下,对这 17 个特征波段点分别建立偏支持向量机(SVM)模型、多元线性回归(MLR)模型和主成分回归(PCR)模型并进行对比分析。结果(表 1)表明,与其他 3 种模型相比,MLR 模型的验证集均方根误差和校正集均方根误差均最小,验证集和校正集相关系数均最大,故建立的 MLR 模型相比于其他 3 种模型最优,稳定性和精确性最高。



A:原始光谱;B:一阶导数 S-G 平滑算法预处理后的光谱;C:一阶 S-G 平滑算法+SNV 算法预处理后的光谱。

图 1 小麦粉近红外光谱的一阶 S-G 平滑算法+SNV 算法预处理

Fig.1 Preprocessing of near infrared spectroscopy of wheat flour by first derivative S-G smoothing algorithm and standard normal variable (SNV) algorithm

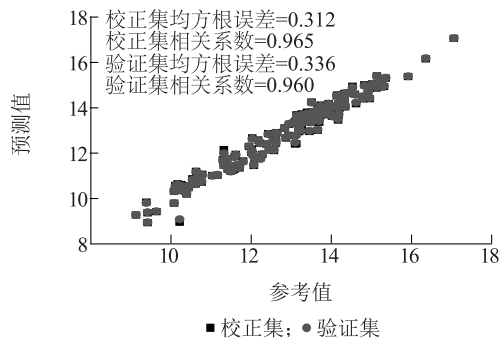


图 2 21 个特征波段点的偏最小二乘回归(PLSR)模型

Fig.2 Partial least square regression (PLSR) model of 21 characteristic band points in the spectrum

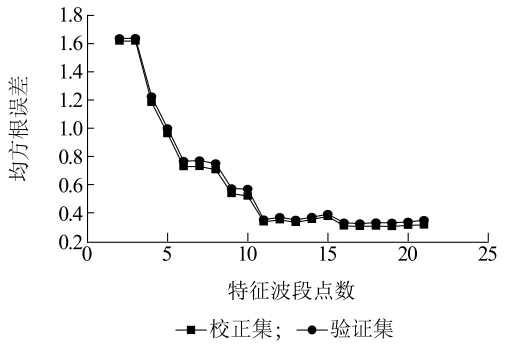


图 4 2~21 个特征波段点的均方根误差(*RMSE*)变化趋势

Fig.4 Variation diagram of root-mean-square error (*RMSE*) of 2−21 characteristic band points in the spectrum

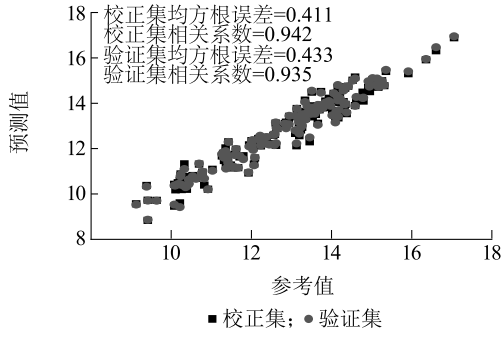


图 3 10 个特征波段点的偏最小二乘回归(PLSR)模型

Fig.3 Partial least square regression (PLSR) model of 10 characteristic band points in the spectrum

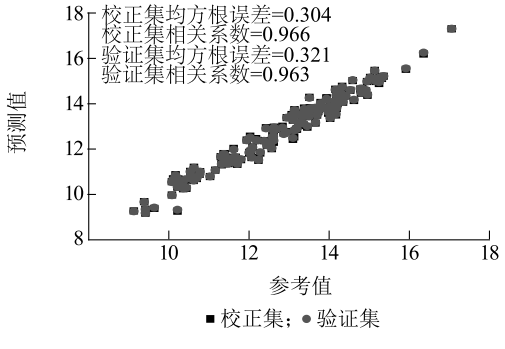


图 5 17 个特征波段点的偏最小二乘回归(PLSR)模型

Fig.5 Partial least square regression (PLSR) model of 17 characteristic band points in the spectrum

表 1 不同模型预测效果比较

Table 1 Comparison of prediction effects between different models

模型	校正集		验证集	
	校正集均方根误差(*RMSEC*)	校正集相关系数(r_c)	验证集均方根误差(*RMSEV*)	验证集相关系数(r_v)
偏最小二乘回归(PLSR)	0.304	0.966	0.321	0.963
主成分回归(PCR)	0.319	0.963	0.337	0.959
支持向量机(SVM)	0.303	0.972	0.346	0.962
多元线性回归(MLR)	0.275	0.976	0.300	0.968

3 结 论

采用 SPA 算法从小麦粉近红外光谱中提取 17 个特征波段点,以这 17 个特征波段点建立的 MLR 小麦蛋白质含量预测模型的预测效果最好,其验证集均方根误差(*RMSEV*)和校正集均方根误差(*RMSEC*)分别为 0.300 和 0.275,验证集相关系数(r_v)和校正集相关系数(r_c)分别为 0.968 和 0.967。由此可见选取不同数量的特征波点对建模存在影响,不同的建模方法所建模型的效果也不同。

参考文献:

[1] 吕程序,蒋训鹏,张银桥,等. 基于变量选择的小麦粗蛋白质含量的近红外光谱检测[J].农业机械学报,2016,47(S1):266-270.

[2] 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质含量检测[J].生态学杂志,2018,37(4):1276-1281.

[3] 宋雨宸,宦克为,韩雪艳,等.基于蒙特卡洛变量组合集群分析法的小麦蛋白质质近红外光谱变量选择[J].长春理工大学学报,2017,40(5):29-34.

[4] 毛晓东,孙来军,戴长军,等.基于近红外光谱的小麦品质分类研究[J].中国农学通报,2013,29(36):386-390.

[5] 钱海波,孙来军,王乐凯,等.基于连续投影算法的小麦湿面筋近红外校正模型优化[J].中国农学通报,2011,27(18):51-56.

[6] 吴静珠,董文菲,董晶晶,等.基于 Si_cPLS 的小麦种子发芽率近红外模型优化研究[J].光谱学与光谱分析,2017,37(4):1114-1117.

[7] 王赋腾,孙晓荣,刘翠玲,等. 光谱预处理对便携式近红外光谱仪快速检测小麦粉灰分含量的影响[J].食品工业科技,2017(10):58-61,66.

[8] 张平平,张 瑜,唐 果,等. 近红外光谱技术检测小麦谷蛋白大聚体含量[J].江苏农业学报,2017,33(6):1207-1211.

[9] 惠广艳,孙来军,王佳楠,等.可见-近红外光谱的小麦硬度预测模型预处理方法研究[J].光谱学与光谱分析,2016,36(7):2111-2116.

[10] 王 冬,李 安,靳欣欣,等.基于 2 原理的近红外光谱仪对辐射花生的快速鉴别比较[J].食品科学,2016,37(8):212-215.

[11] 吴 迪,吴洪喜,蔡景波,等.基于无信息变量消除和连续投影算法的可见-近红外光谱技术白蚁种分类方法研究[J].红外与毫米波学报,2009,28(6):423-427.

[12] CHENG Z, ZHANG L Q, LIU H Y, et al. Successive projections algorithm and its application to selecting the wheat near-infrared spectral variables[J]. Spectroscopy and Spectral Analysis, 2010, 30(4): 949-952.

[13] 孙旭东,郝 勇,蔡丽君,等.基于抽取和连续投影算法的可见近红外光谱变量筛选[J]. 光谱学与光谱分析,2011,31(9):2399-2402.

[14] 任 东,瞿芳芳,陆安祥,等.近红外光谱分析技术与应用[M].北京:科学出版社,2017:31-42.

[15] 展晓日,朱向荣,史新元,等.SPXY 样本划分及蒙特卡洛交叉结合近红外光谱永远橘叶中橙皮苷的含量测定[J].光谱学与光谱分析,2009,29(4):964-968.

(责任编辑:张震林)