doi:10.3969/j.issn.1007-7146.2015.01.006

奇异数据筛选法在玉米籽粒蛋白质近红外光谱检测中的应用

梁秀英[1,3], 李小昱[1], 杨万能[1,2,3,*]

(1. 华中农业大学工学院, 湖北 武汉 430070; 2. 华中农业大学作物遗传改良国家重点实验室, 湖北 武汉 430070;
3. 华中农业大学农业生物信息湖北省重点实验室, 湖北 武汉 430070)

摘 要: 相对于传统生化测定方法,基于近红外光谱(Near infrared spectroscopy, NIRS)玉米籽粒蛋白质含量检测是一种快速、非破坏、且适用于多组分同时检测的新方法。但在建模过程中,由于奇异数据(异常值)的存在会影响近红外光谱模型的预测精度和稳定性,我们采用奇异数据筛选法剔除了玉米籽粒近红外光谱中的奇异数据并建立了玉米籽粒蛋白质含量的偏最小二乘支持向量机(Least squares support vector machine, LS-SVM)模型。本文分别采用杠杆值法(Leverage)、半数重采样法(Resampling by Half-Mean, RHM)和蒙特卡洛采样法(Monte-Carlo Sampling,MCS)剔除了玉米籽粒蛋白质光谱数据中的奇异数据并对模型结果进行比较。在剔除奇异数据的基础上,采用偏最小二乘回归法(Partial least squares regression, PLSR)提取主成分,并基于小生境蚁群算法(Niche ant colony algorithm, NACA)优化偏最小二乘支持向量机(LS-SVM)模型参数(γ 和 σ^2),建立基于 LS-SVM 的玉米籽粒蛋白质定量分析模型。结果表明,采用 3 种奇异数据筛选法剔除奇异数据后所建 LS-SVM 模型的预测结果都优于采用原光谱数据所建模型,相比较而言,蒙特卡洛采样法为基于近红外光谱检测玉米籽粒蛋白质的最佳奇异数据筛选法。

关键词: 玉米籽粒蛋白质; 奇异数据筛选法; 偏最小二乘支持向量机 (LS-SVM); 小生境蚁群算法 (NACA)

中图分类号: O657

文献标识码: A

文章编号: 1007-7146(2015)01-0038-08

Outlier Detection for Measurement of Protein Content in Maize Kernels Based on Near-infrared Reflectance Spectroscopy

LIANG Xiuying[1,3], *LI Xiaoyu*[1], *YANG Wanneng*[1,2,3*]

(1. College of Engineering, Huazhong Agricultural University, Wuhan 430070, Hubei, China;
2. National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, Hubei, China;
3. Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, Hubei, China)

Abstract: As the classical chemical analysis of protein content in maize kernel was slow and destructive, and the exist-

收稿日期: 2015-12-02; **修回日期:** 2015-01-02

基金项目: 国家高技术研究发展计划(863 计划,2013AA102403);华中农业大学博士启动基金(项目编号:2662014BQ038);Fundamental Research Funds for the Central Universities (2012ZYTS023)

作者简介: 梁秀英(1976 -),女,汉族,浙江绍兴人,讲师,主要从事水稻表型组学研究。(手机)18062065806;(电子邮箱)nancy@mail.hzau.edu.cn

* **通讯作者:** 杨万能(1984 -),男,汉族,湖北南漳人,副教授,主要从事植物表型组学研究。(手机)15871800820;(电子邮箱)ywn@mail.hzau.edu.cn

ence of the outliers in the near infrared (NIR) spectra would affect the accuracy and stability of the NIR models, we applied outlier detection methods for measuring protein content in maize kernel based on near infrared spectroscopy. 3 outlier screening methods, leverage method, resampling by half-mean method (RHM), leverage method, and monte-carlo sampling method (MCS), were compared to detect outliers in the protein spectra and the least squares support vector machine (LS-SVM) models were built with using partial least squares regression (PLSR) method to extract the optimal component scores and using niche ant colony algorithm (NACA) to optimize the parameters (γ and σ^2) of the LS-SVM model. The results showed that the performances of the LS-SVM models with those samples removed the outliers were better than the LS-SVM model with all samples. The prediction results of the validation set also showed that the MCS method was optimal for detecting outliers in the spectra of the protein of the whole maize kernel based on NIRS.

Key words: protein content in maize kernel; outlier screening methods; the least squares support vector machine (LS-SVM); niche ant colony algorithm (NACA)

0　引言

玉米是世界三大作物之一,是我国重要的粮食和饲料作物[1,2]。玉米籽粒中蛋白质含量是评价玉米品质和用途的重要指标之一。传统的玉米籽粒蛋白质含量检测主要采用生化测定方法,比如双缩脲法。但传统生化测定方法存在速度慢,且破坏了玉米籽粒不能用于后续的播种。因此,需要寻求一种快速且非破坏的玉米籽粒蛋白质检测方法。

近红外光谱(NIRS)是一种快速、非破坏、且适用于多组分同时检测的技术,目前已广泛应用于农产品的检测中,如小麦[3,4]、大米[5,6]、玉米籽粒[7,8,9,10]和大豆[11,12]。但奇异数据的存在会影响近经外光谱模型的预测精度和稳定性。Abookasis 和 Workman [13] 开发了一种奇异数据筛选法用于区分"好的"和"坏的"近红外光谱,试验结果表明剔除奇异数据后,验证集的预测误差降低。刘蓉等[14]剔除了牛奶光谱中的奇异数据,结果表明所建立的 PLS 模型的预测误差比采用所有样本所建模型的预测误差小。因此,奇异数据的剔除能有效地提高模型的稳定性和预测能力[15]。

本文分别采用杠杆值法(Leverage)、半数重采样法(Resampling by Half-Mean, RHM)和蒙特卡洛采样法(Monte-Carlo Sampling, MCS)剔除玉米籽粒蛋白质光谱数据中的奇异样本,并基于三种方法对模型结果进行比较分析;采用偏最小二乘法(PLS)提取主成分、小生境蚁群算法(niche ant colony algorithm, NACA)优化偏最小二乘支持向量机(LS-SVM)模型的参数(γ 和 σ^2),建立基于 LS-SVM 的玉米籽粒蛋白质含量定量分析模型,并确定基于近红外光谱玉米籽粒蛋白质含量检测的最佳奇异数据剔除法。

1　材料与方法

1.1　试验材料

试验用玉米样品共 4 个品种(饲料行业用硬粒浅色型,饲料行业用硬粒深色型,华玉 4 号和 HZ06148,样本数量分别为 26, 24, 30, 30),共计 110 份样品(每个样品 40 粒玉米籽粒)。图 1 为 4 个品种的部分玉米图片。玉米籽粒蛋白质含量人工测量值采用双缩脲法测定。

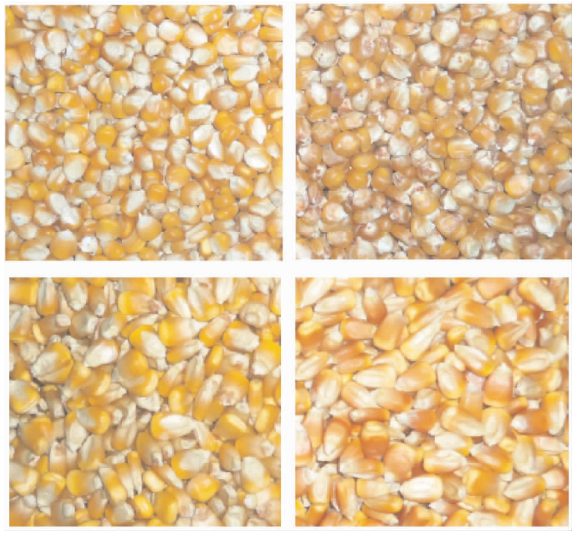

图 1　四个品种部分玉米样本图片

Fig. 1　The four accessions of maize samples

1.2　近红外光谱的测定

试验用的近红外光谱仪器为赛默飞世尔科技(原热电公司)公司生产的 Antaris II 傅里叶变换近红外(FT-NIR)光谱仪,集成透射、反射、漫透射、光纤探头等检测模块,如图 2 所示。样品采集方法如下:光谱类型为漫反射,扫描谱区为 3 999 ~ 10 001 cm^{-1},

分辨率为 4 cm^{-1},扫描次数为 32 次,光谱数据点数为 1 557 个[16],仪器配有旋转样品池(大约能装 40 粒玉米籽粒),每个样品均重复装样三次,每次在不同部位采集二次光谱,计算其平均光谱并存入计算机,最终得到 110 个样品光谱数据,图 3 为 110 个玉米样品的光谱图,其中纵坐标为吸光度,横坐标为波数。光谱数据处理在 Matlab 7.0 平台实现。

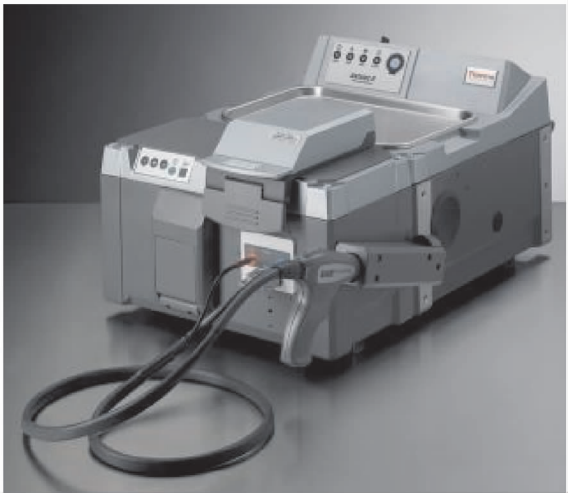

图 2 近红外光谱仪

Fig. 2 The near infrared spectroscopy instrument

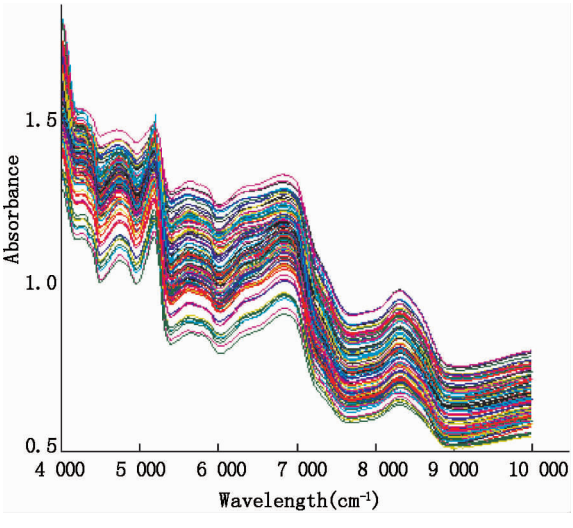

图 3 110 个玉米样品的光谱

Fig. 3 110 maize samples spectra

1.3 奇异数据剔除方法

由于测量仪器、测量方法、测量人员的主观性、样品来源的多样性等因素的影响,使样品的光谱数据或化学测量值存在异常。异常样本的存在会在一定程度上影响模型的适应性和准确性[14,17],论文采用 3 种奇异数据剔除法剔除玉米籽粒蛋白质近红外光谱中的奇异值。

1.3.1 *杠杆值法* 用杠杆值剔除奇异数据是一种传统的方法,具有简单直观的优点,但计算过程中需要计算矩阵的逆运算。第 i 个样品的杠杆值(Leverage)的计算公式如公式(1)所示[14]:

$$h_{ii} = diag(X(X^TX)X^T) = x_i(X^TX)^1x_i^T \quad (1)$$

其中,h_{ii}为第 i 个样品的杠杆值;x_i 为第 i 个样品的光谱; X 为样品的光谱矩阵。

1.3.2 *半数重采样法*(Resampling by Half-Mean, RHM) 从原始光谱矩阵 X 中随机选择原始样本数的一半组成新矩阵 $X(i)$(表示第 i 个采样矩阵),计算 $X(i)$的均值 $\mu(i)$和方差 $\sigma^2(i)$,对原始光谱矩阵 X 进行标准化处理。第 i 个采样矩阵中的样本的向量长度 $L(i)$定义为[14]:

$$L(i) = \sqrt{\sum_{k=1}^{p}\left(\frac{X_k - m_k(i)}{\sigma_i(i)}\right)^2} \quad (2)$$

式中,p 为光谱波长点数;$m_k(i)$为 X_k 的平均值;$\sigma^2(i)$为 X_k 的方差。

对全部样本的向量长度进行排序,向量长度最大的 6% 的样本分布认为是奇异点,进行剔除。

1.3.3 *蒙特卡洛采样法*(Monte-Carlo Sampling, MCS) 基于蒙特卡洛采样的奇异数据剔除法是基于预测误差对奇异数据的敏感特性,以降低多个奇异数据的掩蔽效应[18]。其方法如下:利用偏最小二乘(PLS)法确定最佳隐变量数;用 MCS 法按一定比例(通常训练集占样本总数的 70%-90%)将原始光谱数据随机分为校正集和验证集;建立近红外光谱分析模型,计算验证集中每一个样本的预测误差,循环足够多次以保证每个样本均被预测到;计算每个样本预测残差的均值(μ_i)和标准偏差(σ_i),对所有样本以 σ_i 对 μ_i 作图,得各样本的方差-均值分布图,那些位于高均值或高标准偏差区域的样本最有可能是奇异数据。

1.4 SPXY 样本集划分法

论文采用基于联合 x-y 距离的样本集划分(SPXY)将剔除奇异数据后的样本集划分为校正集和验证集。SPXY 法是由 Galvao 等首先提出的[19],它是在 KS 算法的基础上发展而来的,SPXY 在样品间距离计算时将 x 变量和 y 变量同时考虑在内,其距离公式如下[20]:

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}[x_p(j) - x_q(j)]^2}; p,q\epsilon[1,N] \quad (3)$$

$$d_y(p,q)=\sqrt{(y_p-y_q)^2}=|y_p-y_q|;p,q\epsilon[1,N] \quad (4)$$

公式(3)和(4)中,$x_p(j)$ 和 $x_q(j)$ 分别为样本 p 和 q 在第 j 个光谱波长上的吸光度。J 为光谱波长点数,N 为总的样本个数。

SPXY 的逐步选择的过程和 KS 法相似,但用 $d_{xy}(p,q)$ 代替了 $d_x(p,q)$,为了让样本在 x 和 y 空间具有相同的权重,将 $d_x(p,q)$ 和 $d_y(p,q)$ 分别除以它们在数据集中的最大值,因此标准化的 x-y 的距离公式为[20]:

$$d_{xy}(p,q)=\frac{d_x(p,q)}{\max_{p,q\epsilon[1,N]}d_x(p,q)}+\frac{d,y(p,q)}{\max_{p,q\epsilon[1,N]}d_y(p,q)};\ p,q\epsilon[1,N] \quad (5)$$

1.5　近红外光谱建模方法

1.5.1　偏最小二乘回归法(Partial Least Squares Regression, PLSR)　PLS 是一种类似于主成分回归(PCR)的光谱定量分析方法,它同时将校正集浓度矩阵 X 和相应的响应量矩阵 Y 进行主成分分解,在构造校正模型时充分利用了 X 和 Y 阵中的信息,是比较完善的基于因子分析原理的校正方法[21],可降低噪声对校正模型的影响,比较适用于处理变量多而样本数少的问题[22]。

1.5.2　偏最小二乘支持向量机(Least squares support vector machine, LS-SVM)　支持向量机(Support vector machine,SVM)是基于统计学习理论的小样本学习方法,具有很好的泛化性能,在解决小样本、非线性及高维数据中具有许多特有的优势[23, 24]。偏最小二乘支持向量机(LS-SVM)是 SVM 改进后的算法,对非线性问题有更强的建模能力[25]。LS-SVM 常用的非线性核函数有:多项式、径向基、Sigmoid 型。其中径向基函数可以将非线性样本数据映射到高维特征空间,可处理具有非线性关系的样本数据[26]。论文采用径向基 RBF 核函数,采用径向基 RBF 核函数的 LS-SVM 的参数主要 γ 和 σ^2(γ 是控制对错分样本惩罚的程度的可调参数,σ^2 是径向基核函数的参数)[27]采用实数编码的小生境蚁群算法进行优化。

1.5.3　小生境蚁群算法(Niche ant colony algorithm,NACA)　蚁群算法,又称蚂蚁算法,是一种求解组合最优化问题的新型通用启发式方法。小生境蚁群算法是蚁群算法的一种,具有较强的局部搜索能力,对蚁群算法的后期进行局部的搜索,可找到更优解。采用实数编码的小生境蚁群算法优化 LS-SVM 参数 γ 和 σ^2 的具体算法步骤如下[28]:

(1)随机产生蚂蚁的初始位置(限制在可行域内);

(2)计算每个蚂蚁的初始信息素(正比于目标函数值);

$$T[i]=k\times f(X[i]) \quad (6)$$

式中,T 为蚂蚁的信息素,$f(x)$ 为目标函数,k 为比例常数。

(3)根据蚂蚁留下的信息素大小,确定每个蚂蚁的下一步转移概率:

$$P[i]=(\text{T\_Best-}T[i])/\text{T\_Best} \quad (7)$$

式中,T_Best 为最大信息素;$P[i]$ 为第 i 个蚂蚁的转移概率;T 第 i 个蚂蚁的信息素。

若蚂蚁的转移概率小于 P_0(全局转移选择因子),则进行局部搜索,否则,进行全局搜索。

(4)更新信息素。每个蚂蚁信息素的更新规则如下:

$$T[i]=(1-P_1)\times T[i]+k\times f(X[i]) \quad (8)$$

式中,P_1 为信息素蒸发系数。

(5)保存每代最优解,在每次迭代过程中,将信息素最大的蚂蚁保存下来,返回第(1)步,进行迭代循环。

(6)得到全局最优解。如果迭代次数满足开始设置的要求,则搜索完成,从而得到最佳蚂蚁,将最佳蚂蚁转换成 LS-SVM 参数 γ 和 σ^2。

1.6　近红外光谱分析模型的评价

论文采用相关系数、验证集均方根误差来评价玉米籽粒蛋白质含量的定量分析模型。

(1) 相关系数(R)

相关系数(R)的公式如下所示:

$$R=\frac{\sum_{i=1}^{N}\left(\hat{y}_i-\bar{\hat{y}}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i-\bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}} \quad (9)$$

式中,N 为样本数,$\hat{y}_i$为样本 i 的预测值,$\bar{\hat{y}}_i$ 为样本预测值的平均值,y_i 为样本 i 的实测值,$\overline{y_1}$为样本实测值的平均值。

(2) 验证集均方根误差(Root mean squared error of prediction, RMSEP)

$$RMSEP=\sqrt{\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{N-1}} \quad (10)$$

式中,N 为验证集样本数,$\hat{y}_i$为验证集样本 i 的预

测值，y_i为验证集样本 i 的实测值。

2 结果与讨论

2.1 玉米籽粒蛋白质含量检测的奇异数据剔除法

在全谱范围(3 999-10 001 cm^{-1})内，采用杠杆值法(Leverage)、半数重采样法(RHM)和蒙特卡洛采样法(MCS)分别计算出 110 个样本原始光谱数据的杠杆值、RHM 值、预测残差的均值(μ_i)和标准偏差(σ_i)。采用杠杆值法计算的杠杆值与样品序号分布结果如图 4 所示，样本 7、11、15、27、33、35、44 的杠杆值较大，即这 7 个样本最有可能为奇异数据点。用 RHM 法(循环 1000 次)计算得到的每个样本被选为异常值次数如图 5 所示，RHM 法剔除奇异数据非常直观，绝大多数样本的 RHM 得分为 0，得分最大 6% 的样本是样本 7，15，33，82，86，93 和 104，即这 7 个样本最有可能为奇异数据点。图 6 为用 MCS 法循环 2500 次测得的每个样本预测残差的方差-均值分布图，样本 73 预测残差的平均值较大，样本 21、30、72、74 预测残差的标准偏差较大，即这 4 个样本最有可能为异常样本点。

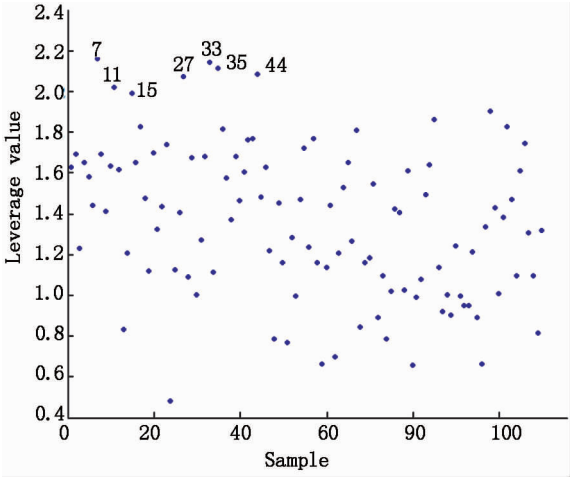

图 4 采用杠杆值法每个样本杠杆值与样品序号分布图

Fig. 4 The leverage values of maize samples

2.2 基于 LS-SVM 和 NACA 的玉米籽粒蛋白质定量分析

分别采用杠杆值法、RHM 和 MCS 法剔除原始光谱数据中的奇异数据点后，用 SPXY 法将剔除奇异数据后的样本集按 2∶1划分成校正集和验证集，将原始光谱进行‘Autoscale’预处理后，用 PLS 法对光谱数据进行降维并提取偏最小二乘因子作为 LS-SVM 模型的输入。LS-SVM 的核函数采用径向基核函数，并用小生境蚁群算法优化 LS-SVM 模型的参数 γ 和 σ^2。采用 3 种不同奇异数据剔除方法所建的 LS-SVM 模型的结果如表 1 所示。从验证集的相关系数和均方根误差可以看出，剔除奇异数据后所建的 LS-SVM 模型优于所有样本所建的模型，采用 MCS 法剔除奇异数据后所建模型的验证集均方根误差(RMSEP)较小，且与校正集均方根误差(RMSEC)较接近，模型较稳定，所以 MCS 为玉米蛋白质近红外光谱检测中的最佳奇异数据剔除法。

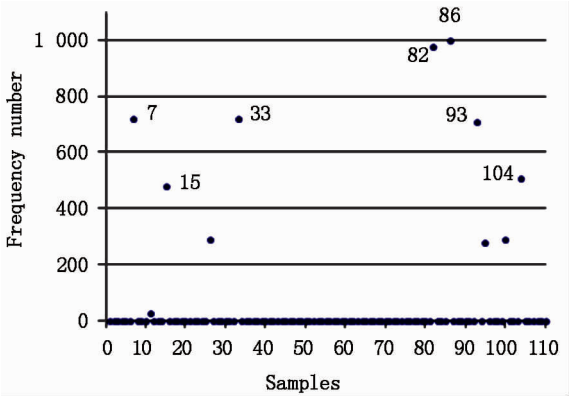

图 5 用 RHM 法计算得到的各样本被选为异常值次数

Fig. 5 The frequency number of each sample selected as outliers by RHM method

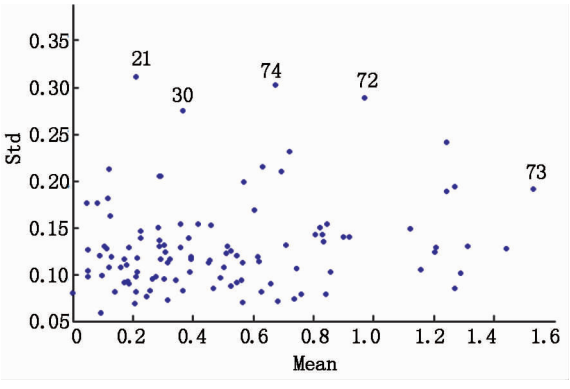

图 6 用 MCS 法求得的每个样本预测残差的方差-均值分布图

Fig. 6 The distribution diagram of residual mean and variance based on MCS method

用 SPXY 划分样本校正集和验证集时玉米蛋白质含量化学值的基本统计数据如表 2 所示。

采用 MCS 法剔除奇异数据后所建模型，校正集预测值与人工测量化学值的散点图如图 7(a)所示，

验证集预测值与人工测量化学值的散点图如图 7(b)所示。玉米籽粒蛋白质含量验证集样本绝对相对误差的频数图所图 8 所示,所有样本的绝对相对误差都小于 8%。

表 1　基于不同奇异数据剔除法的 LS-SVM 模型结果

Tab. 1　The predicted results of LS-SVM model based on different outlier screening methods

方法 Method	预处理方法 Preprocessing method	样本集 Sample subset	隐变量数 Number of latent variables	LS-SVM 模型参数 Parameters of LS-SVM model		校正集 Calibration set		验证集 Validation set	
				γ	σ^2	R	RMSEC	R	RMSEP
无	Autoscale	所有样本	6	1.015	1.680	0.952	0.299	0.646	0.494
杠杆值	Autoscale	剔除样本 7, 11, 15, 27, 33, 35 和 44	9	1.000	11.103	0.857	0.345	0.729	0.434
RHM	Autoscale	剔除样本 7, 15, 33, 82, 86, 93 和 104	7	64.751	82.386	0.813	0.339	0.672	0.489
MCS	Autoscale	剔除样本 21, 30, 72, 73, 和 74	6	3.577	12.849	0.818	0.363	0.775	0.393

表 2　用 SPXY 划分校正集和验证集时玉米样品蛋白质含量化学值的统计数据

Tab. 2　Statistics of maize protein content for calibration set and validation set with SPXY subsets partition method

样本集划分方法 Subsets partition method	校正集 Calibration set				验证集 Validation set			
	样本数 Sample size	蛋白质含量范围(%) Range of protein content	平均值 Mean	标准差 Standard deviation	样本数 Sample size	蛋白质含量范围(%) Range of protein content	平均值 Mean	标准差 Standard deviation
SPXY	70	8.541-11.807	10.311	0.760	35	9.445-11.455	10.535	0.585

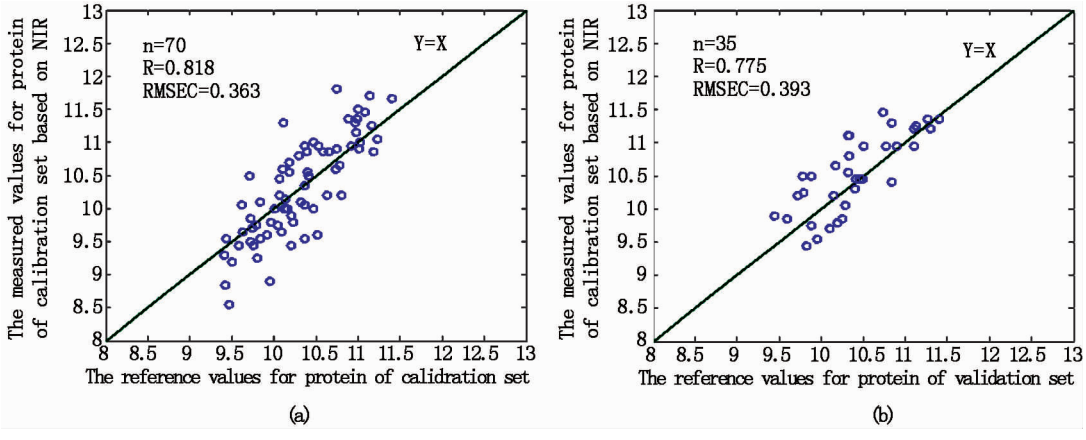

图 7　玉米籽粒蛋白质含量校正集和验证集散点图(a)校正集预测值与人工测量值的散点图;(b)验证集预测值与人工测量值的散点图

Fig. 7　Scatter plots of predicted values versus measured values of maize kernel protein content based on LS-SVM (a) Calibration set (b) Validation set

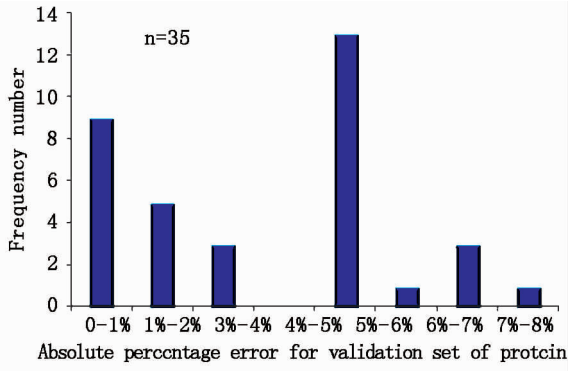

图 8 验证集绝对相对误差的频数图

Fig. 8 The frequency number plot of absolute relative error of the predicted values versus measured values for validation set

3 结论

基于近红外光谱测量技术,本文比较了 3 种奇异数据剔除法对玉米籽粒蛋白质含量分析模型的影响,并确定 MCS 法为玉米籽粒蛋白质含量近红外光谱检测的最佳奇异数据剔除法。采用 NACA 算法优化 LS-SVM 模型的参数,并建立了基于 LS-SVM 的玉米籽粒蛋白质含量分析模型。校正集和验证集的相关系数分别为 0.818 和 0.775,校正集均方根误差(RMSEC)和验证集均方根误差(RMSEP)分别为 0.363 和 0.393,验证集所有样本相对误差都小于 8%。由此可见,本文实现一种基于近红外光谱玉米籽粒蛋白质无损测量技术,并比较得出最优奇异数据筛选法,通过模型验证取得了较为稳定的建模和预测效果,为后续玉米籽粒蛋白质含量无损、高通量、数字化提取提供了重要的技术保障。

参考文献

[1] PRASANNA B M, PIXLEY K, WARBURTON M L, *et al*. Molecular marker-assisted breeding options for maize improvement in Asia[J]. Mol Breeding, 2010, 26:339-356.

[2] YANG X, GAO S, XU SH, *et al*. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize[J]. Mol Breeding, 2011, 28:511-526.

[3] SINELLI N, PAGANI M A, LUCISANO M, *et al*. Prediction of semolina technological quality by FT-NIR spectroscopy [J]. Journal of Cereal Science, 2011, 54:218-223.

[4] 王卫东, 谷运红, 秦广雍, 等. 近红外漫反射光谱法测定整粒小麦单株蛋白质含量[J]. 光谱学与光谱分析, 2007, 27(4):697-701.
WANG Weidong, GU Yunhong, QIN Guangyong, *et al*. Prediction of protein content of intact wheat seeds with near infrared reflectance spectroscopy(NIRS) [J]. Spectroscopy and Spectral Analysis, 2007, 27 (4):697-701.

[5] LIU J, WU S, FANG R. Rapid measurement of rice protein content by near infrared spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2001, 32:68-70.

[6] LIU M, TANG Y, LI X, *et al*. Study on NIR spectral detection model of rice protein content[J]. Chinese Agricultural Science Bulletin, 2013, 29:212-216.

[7] TALLADA J G, PALACIOS-ROJAS N, ARMSTRONG P R. Prediction of maize seed attributes using a rapid single kernel near infrared instrument[J]. Journal of Cereal Science, 2009, 50:381-387.

[8] FASSIO A, FERNANDEZ E G, RESTAINO E A, *et al*. Predicting the nutritive value of high moisture grain corn by near infrared reflectance spectroscopy[J]. Computers and Electronics in Agriculture, 2009, 67:59-63.

[9] 卢宝红, 张俊, 张义荣, 等. 玉米完整籽粒近红外品质分析模型的比较及改进[J]. 中国粮油学报, 2005, 20 (4): 44-49.
LU Baohong, ZHANG Jun, ZHANG Yirong, *et al*. Comparison and improvement of quality analytical models for whole maize kernels using near infrared spectroscopy [J]. Journal of the Chinese Cereals and Oils Association, 2005, 20 (4):44-49.

[10] 魏良明, 严衍禄, 戴景瑞. 近红外反射光谱测定玉米完整籽粒蛋白质和淀粉含量的研究[J]. 中国农业科学, 2004, 37(5):630-633.
WEI Liangming, YAN Yanlu, DAI Jingrui. Determining protein and starch contents of whole maize kernel by near infrared reflectance spectroscopy (NIRS) [J]. Scientia Agricultura Sinica, 2004, 37(5):630-633.

[11] PLANS M, SIMÓJ, CASAИAS F, *et al*. Characterization of common beans (Phaseolus vulgaris L.) by infrared spectroscopy:comparison of MIR, FT-NIR and dispersive NIR using portable and benchtop instruments[J]. Food Research International, 2013, 54:1643-1651.

[12] FERREIRA D S, GALãO O F, PALLONE J A L. Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples[J]. Food Control, 2014, 35:227-232.

[13] ABOOKASIS D, WORKMAN J J. Application of spectra cross-correlation for Type II outliers screening during multivariate near-infrared spectroscopic analysis of whole blood[J]. Chemometrics and Intelligent Laboratory Systems, 2011, 107:303-311.

[14] 刘蓉, 陈文亮, 徐可欣, 等. 奇异点快速检测在牛奶成分近红外光谱测量中的应用[J]. 光谱学与光谱分析, 2005, 25 (2) :207-210.
LIU Rong, CHEN Wenliang, XU Kexin, *et al*. Fast outlier detection for milk near-infrared spectroscopy analysis [J]. Spectroscopy and Spectral Analysis, 2005, 25 (2):207-210.

[15] LILLHONGA T, GELADI P. Replicate analysis and outlier de-

tection in multivariate NIR calibration, illustrated with biofuel analysis[J]. Analytical Chimica Acta, 2005, 544:177-183.

[16] ZHU X, LI S, SHAN Y, *et al*. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics [J]. Journal of Food Engineering, 2010, 101:92-97.

[17] 王龙. 葡萄糖溶液浓度检测的预测模型研究[D]. 武汉:华中科技大学硕士研究生学位论文, 2006.
WANG Long. The Research of Prediction Model In Detecting Glucose Solutions Concentration [D]. Wuhan: Huazhong University of Science and Technology, 2006.

[18] 张华秀. 近红外光谱法快速检测牛奶中蛋白质与脂肪含量[D]. 中南大学硕士研究生学位论文, 2010.
ZHANG Huaxiu. Determination of Protein and Fat in Milk by Near Infrared Spectroscopy [D]. Central South University, 2010.

[19] 程志颖, 孔浩辉, 张俊, 等. 粒子群算法结合支持向量机回归法用于近红外光谱建模[J]. 分析测试学报, 2010, 29(12):1215-1219.
CHENG Zhiying, KONG Haohui, ZHANG Jun, *et al*. Application of particle swan optimization-Least square support vector machine regression to modeling of near infrared spectra [J]. Journal of Instrumental Analysis, 2010, 29(12):1215-1219.

[20] GALVAO R K H, ARAUJO M C U, JOSÉG E, *et al*. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67:736-740.

[21] 瞿海斌. 基于 PLS 的建模方法[J]. 浙江大学学报(工学版), 1999, 33(5):471-474.
QU Haibin. Modeling method based on PLS [J]. Journal of Zhejiang University (Engineering Science), 1999, 33(5): 471-474.

[22] 张军, 郑咏梅, 王芳荣, 等. 谷物近红外光谱分析中常用数据处理方法讨论[J]. 吉林大学学报(信息科学版), 2003, 21(1):4-9.
ZHANG Jun, ZHENG Yongmei, WANG Fangrong, *et al*. Discussion on some regular methods for cereal near infrared spectra analysis [J]. Journal of Jilin University (Information Science Edition), 2003, 21(1):4-9.

[23] LIANG X Y, LI X Y, LEI T W, *et al*. Study of sample temperature compensation in the measurement of soil moisture content [J]. Measurement, 2011, 44:2200-2204.

[24] FERNWáNDEZ PIERNA J A, LECLER B, CONZEN J P, *et al*. Comparison of various chemometric approaches for large near infrared spectroscopic data of feed and feed products[J]. Analytica Chimica Acta, 2011, 705:30-34.

[25] 牛晓颖, 周玉宏, 邵利敏. 基于 LS-SVM 的草莓固酸比和可滴定酸近红外光谱定量模型[J]. 农业工程学报, 2013, 29(增刊):270-274.
NIU Xiaoying, ZHOU Yuhong, SHAO Limin. Improved NIR quantitative model of soluble solids titratable acid ratio and titratable acidity in strawberry based on LS-SVM[J]. Transactions of the CSAE, 2013, 29(Supp. 1):270-274.

[26] 赵杰文, 呼怀平, 邹小波. 支持向量机在苹果分类的近红外光谱模型中的应用[J]. 农业工程学报, 2007, 23(4): 149-152
ZHAO Jiewen, HU Huaiping, ZOU Xiaobo. Application of support vector machine to apple classification with near-infrared spectroscopy [J]. Transaction of the CSAE, 2007, 23(4): 149-152.

[27] PELCKMANS K, SUYKENS J A K, GESTEL T V, *et al*. LS-SVMlab Toolbox User's Guide (Version 1.5)[G]. ESAT-SCD-SISTA Technical Report 02-145, 2003.

[28] 李彦苍, 索娟娟. 基于熵的小生境蚁群算法及其应用[J]. 四川大学学报(工程科学版), 2007, 39 Supp:229-232.
LI Yancang, SUO Juanjuan. Niche ACO based on entropy and its application [J]. Journal of Sichuan University (Engineering Science Edition), 2007, 39 Supp:229-232.

奇异数据筛选法在玉米籽粒蛋白质近红外光谱检测中的应用

作者： 梁秀英， 李小昱， 杨万能， LIANG Xiuying， LI Xiaoyu， YANG Wanneng

作者单位： 梁秀英,LIANG Xiuying(华中农业大学工学院，湖北 武汉430070；华中农业大学农业生物信息湖北省重点实验室，湖北 武汉430070)， 李小昱,LI Xiaoyu(华中农业大学工学院,湖北 武汉,430070)， 杨万能,YANG Wanneng(华中农业大学工学院，湖北 武汉430070；华中农业大学作物遗传改良国家重点实验室，湖北 武汉430070；华中农业大学农业生物信息湖北省重点实验室，湖北 武汉430070)

刊名： 激光生物学报 ISTIC

英文刊名： Acta Laser Biology Sinica

年，卷(期)： 2015(1)