中文稿件要求

1题目和摘要

来稿题目应简明扼要，提纲挈领，精准地概括全文主旨，字数不超过20个。

中文摘要须超过250字，英文摘要须超过200个单词，均须包含四个要素：目的、方法、结果、结论，其中一定要有数据和量化的东西出现。英文摘要要以第三人称撰写，句型简单、语句通畅、意思明确，中英文要相互对应。

2 中图分类号和关键词

中图分类号请参阅http://www.ztflh.com/。

关键词3~5个为宜，其中第一个关键词须体现文章主旨或列出文章主要工作所属的学科名称。

3 基金项目和作者简介

文章页脚处标注基金项目和作者简介。

受资助的基金须为省部级及以上级别，注明基金项目名称及批准号。

作者简介（第一作者为学生的须再写出导师简介）格式如下：姓名（出生年-），性别，学历，职称，研究方向，Email。

4 图、表和公式

图、表和公式在文章中按照先后顺序编排序号，并排在正文的相应位置。每个图、表要有简明的中、英文题名，图字、表文须以英文写出，有分图时用(a)、(b)、(c)等标号。特别指出：全文中图、表总数不能超过10个。

坐标图要特别注意：必须标注横纵坐标的“物理量、物理量规定的符号、单位”，这三者只有在不必标明（无量纲等情况）时方可省略。

5 名词术语、物理量符号、计量单位和缩略词

请使用国家标准名词术语。

容易混淆的大小写，上、下角标，算符等须清晰注明。

物理量符号用斜体，物理量单位用正体，矩阵、向量、矢量符号用黑斜体，注意物理量符号要全文统一。

计量单位应严格执行国家标准规定，已经废止的计量单位，必须按照现行标准折算，如（英寸（in）须折算为厘米（cm），1in=2.54cm）。

文章中首次出现的符号和缩略词须给出解释或者全称。

面粉灰分近红外光谱检测的异常样本分析

吴胜男，刘翠玲，吴静珠，孙晓荣，董秀丽

（北京工商大学 计算机与信息工程学院，北京 100048）

摘要：利用近红外光谱仪检测面粉灰分时，用马氏距离法和蒙特卡洛交叉验证法（MCCV）分别对异常样本进行剔除，并用偏最小二乘法（PLS）进行建模，最终，用马氏距离法剔除异常样本，当权重系数为 1.5，剔除样本数为 3 时，得到最好结果，相关系数(R^2)为 92.67，交互验证均方差(RMSECV)为 0.0485；MCCV 法剔除异常样本，剔除样本数为 3，得到最好结果， R^2 为 94.64，RMSECV 为 0.0411。可见，马氏距离法和 MCCV 剔除异常样本后模型的准确度都有提高，而 MCCV 法剔除异常样品后所建的面粉灰分模型更加稳定，准确度更加提高。

关键词：近红外光谱； 异常样本； 面粉灰分； 马氏距离法； MCCV

Outlier sample analysis on near infrared spectroscopy determination for flour ash

WU Shengnan, LIU Cuiling, WU Jingzhu, SUN Xiaorong, DONG Xiuli

（Beijing Technology and Business University, Beijing 100048,China）

Abstract: The flour ash was determined by near infrared(NIR)apparatus. Mahalanobis distance method and Monte Carlo cross validation (MCCV) method for abnormal samples removed，and using partial least squares (PLS) modeling, Eventually, Mahalanobis distance method for abnormal samples removed, when the weight factor of 1.5, excluding the number of samples is 3 to get the best results, the correlation coefficient (R^2) is 92.67, cross validation mean square error (RMSECV) is 0.0485，MCCV method for abnormal samples removed, excluding the number of samples 3 to get the best results, R^2 is 94.64, RMSECV is 0.0411. Mahalanobis distance method and MCCV method for abnormal samples removed have improved the accuracy of the model ，and MCCV method for flour ash model is more stable and more to improve the accuracy.

Key words: Near infrared spectroscopy; Unusual samples; Mahalanobis distance; MCCV; Flour ash

0 引 言

随着化学计量学的发展，近红外光谱技术被广泛应用于药品、石油、食品等的品质检测中[1-3]。传统的面粉灰分的检测有方法 550℃灼烧法和 850℃乙酸镁法等这些方法存在操作繁琐、耗时长、费时费力和检测效率低等问题[4]，而近红外光谱分析技术可以实现快速、无损地对面粉中的灰分的含量进行检测，这

基金编号：******

作者简介：吴胜男(1987-)，女，河北唐山人，硕士，智能检测技术及仪器的研究，Emai: wsn_0601@126.com

导师简介：刘翠玲(1963-),女，河北唐山人，博士，智能检测技术及仪器的研究，Email:sim688@163.com

会成为面粉品质分析的主要趋势。

近红外光谱分析技术是一种间接分析技术，分析结果的可靠性主要取决于预测模型的准确性和稳定性[5-7]。面粉样品的原始数据，即样品的近红外光谱图和化学值的相关性直接影响模型的预测能力，而异常样品的干扰是影响分析模型的重要因素，因此异常样品的判别与处理是提高模型预测能力的一个重要步骤。产生异常样品的原因：(1)测量仪器、测量方法及环境等客观因素影响；(2)技术人员主观因素的作用；(3)样品的复杂性、多样性。建立定量分析模型时，样品中常常混有异常样本，需要在训练模型之前，把异常样本进行剔除[8]。本文主要探讨马氏距离法以及蒙特卡罗交叉验证法，对异常近红外光谱的判别方法和准则，通过试 验来验证两种方法的有效性，并对两种方法进行比较，从而提高近红外光谱面粉灰分检测模型的准确性和可靠性。

1 试验材料、仪器与方法

1.1 样品的准备

试验所用面粉样本，是从合作单位古船面粉厂取得的不同日期、不同生产线生产的不同种类的面粉，共60个。

1.2 样品化学值的测量

本次实验采用国标法850℃乙酸镁法，准确测量面粉样本的灰分含量，所测值作为建模时的化学值。

1.3 样品近红外光谱的采集

本次试验使用傅立叶变换近红外光谱仪VERTEX 70，将上述面粉样品放置在漫反射样品台的样品杯中，进行近红外光谱采集。大样品杯旋转采样，环境温度23～25℃，波数范围12000～4000cm^{-1}，分辨率8cm^{-1}，扫描次数64次。60个面粉样本的近红外漫反射光谱图，如图1所示。

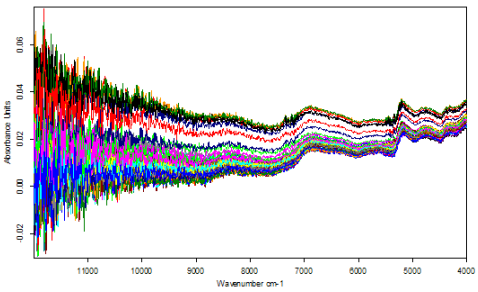

图1 面粉样本的近红外光谱图

Fig.1 Near infrared spectra of flour

1.4 马氏距离与蒙特卡罗交叉验证算法

1.4.1 马氏距离算法

光谱数据为A($n\times k$)矩阵，n代表样品数，k代表选出的主成分数(具有代表性的波长点)。

计算出n个样品的平均光谱:

$$\overline{A}_j=\sum_{i=1}^{m}A_{ij}/n \tag{1}$$

式中:A——光谱矩阵；n——样品数；j——波长序号；$\overline{A}$ —— 平均光谱。

把光谱矩阵进行中心化处理:

$$A(i,:)=A(i,:)-\overline{A} \tag{2}$$

然后计算出马氏矩阵:

$$M=\left(\frac{A'\mathrm{g}A}{n-1}\right) \tag{3}$$

此时的选出的波长数为k，所以M是$k\times k$维矩阵。

计算校正集样品到平均光谱的马氏矩阵:

$$D^2=(A_i-\overline{A})\mathrm{g}M^{-1}\mathrm{g}(A_i-\overline{A})' \tag{4}$$

根据上述计算出的n个马氏距离设置一个阈值范围来检验这n个样品中的奇异样品的存在。

其阈值范围计算方法为:

$$D_t=\overline{D}+e\mathrm{g}\sigma_D \tag{5}$$

式中: $\overline{D}$ ——马氏距离的平均值；σ_D ——马氏距离的标准差；e——调整闭值范围的参数。

如果当$D_i\le D_t$，则认为样品i与样品平均光谱非常接近，称i为平均样品的邻近样品。令N_i为i的邻近样品个数，N_i值越大，样品i的邻近样品数目就越多，在空间上就越密集。设置不同的阈值范围参数e，从而调节N_i值的大小，当e值越大，N_i值越大，阈值范围越宽，邻近样品在空间上的密集度就越高；反之亦然。针对以上不同的e值所选取的阈值范围，分别使用PLS建模回归预测，选取最合适的阈值范围[9-10]。

1.4.2 蒙特卡罗交叉验证算法

蒙特卡罗交叉验证算法(Monte Carlo Cross Validation ，MCCV)的异常样本筛选法是一种新近提出的筛选异常样本的方法，能够降低由掩蔽效应带来的风险，有效检出光谱阵和性质阵方向的奇异点，与传统筛选方法相比具有较高的识别异常样本的能力[11]。利用蒙特卡洛随机取样(Monte Carlo sampling，

MCS)法选取 80%的样本作为校正集建立 PLS 回归模型，剩余部分作预测集，多次循环，从而可以得到各样本的一组预测残差，求出各样本预测残差均值(MEAN)与方差(STD)，从而判断异常样本[12]。

通过校正集相关系数(R^2)、交叉验证均方差（RMSECV)、预测均方差(RMSEP)对模型进行评价，从而验证剔除异常样本是否有利于模型精度的提高。

2 结果与讨论

2.1 含异常样品的面粉近红外光谱分析

将 60 个样本应用于近红外定量分析，通过 Kennard-Stone(KS)方法，确定校正集 50 个样本，剩余 10 个样本用于模型验证。通过 OPUS6.5 软件的分析和优化，选择最优处理算法，寻找面粉的吸收光谱较丰富的波段。分析表明，面粉对光谱信息贡献量最大的谱区范围是 4848.4～4246.7cm^{-1}，维数为 6，利用 PLS 方法进行建模，可得相关系数(R^2)为 85.69，交互验证均方差(RMSECV)为 0.0672，50 个面粉样本近红外光谱图交叉验证后灰分的近红外计算值与化学分析值，如图 2 所示，其中，横坐标为灰分化学分析值，纵坐标为近红外计算值。

参照以往面粉灰分近红外定量研究，相关系数相对较低，模型预测效果不够理想，可能由于异常样品的存在影响了模型的预测精度，所以必须把异常样本剔除。

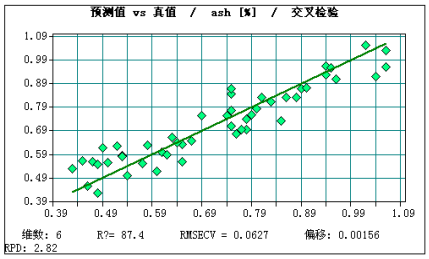

图 2 近红外光谱交叉验证计算值与化学分析值

Fig.2 Cross-validation of the near-infrared spectra calculated with the chemical analysis values

2.2 马氏距离法剔除异常样品

对 50 个校正集样本的近红外光谱进行马氏距离计算，可得到马氏距离分布图，如图 3 所示。

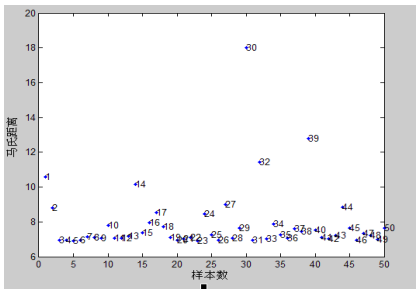

图 3 校正集的马氏距离分布图

Fig.3 Mahalanobis distance maps of Calibration set

从图 3 中能够很明显地看出一些异常样品如 30,39 等,为了进一步对异常样品进行判断分析, 设定 6 个不同权重系数 e (3,2.5,1.5,1.2,1.0, 0.5) ,分别剔除异常样本为：30（e=3），30、39（e=2.5），30、32、39（e=1.5），1、30、32、39（e=1.2），1、14、30、32、39（e=1.0），1、14、27、30、32、39（e=0.5），剔除异常样品后,对光谱信息贡献量最大的谱区范围 4848.4～4246.7cm-1 分别采取 PLS 方法建模,其主成分数的选择采用交互验证(cross validation)方法来选取，所得结果如表 1 所示，马氏距离法剔除异常样品后交叉验证计算值与化学分析值，如图 4 所示。

表1 不同阈值剔除异常样本后PLS校正模型交互校验结果

Tab.1 Different threshold excluding abnormal samples cross validation results of PLS calibrationmodel

权重系数	剔除个数	主成分数	R^2	RMSECV
∞	0	6	85.69	0.0672
3	1	6	88.47	0.0605
2.5	2	7	91.78	0.0516
1.5	3	8	92.67	0.0485
1.2	4	8	92.48	0.0495
1.0	5	8	92.35	0.0491
0.5	6	8	91.60	0.052

由表 1 可知，当权重系数为 1.5，主成分数为 8，剔除异常样本数为 3 时，得到最好结果，相关系数(R2)为 92.67，交互验证均方差(RMSECV)为 0.0485。

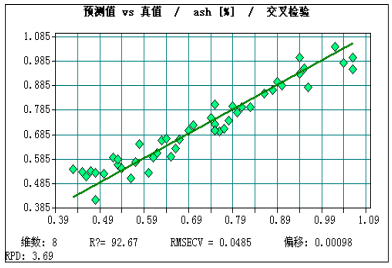

图 4　马氏距离法剔除异常样品后交叉验证计算值与化学分析值

Fig.4　Cross-validation calculated after Mahalanobis distance method for abnormal samples removed and chemical analysis

2.3　蒙特卡罗交叉验证算法剔除异常样本

在 50 个校正集样本中，用蒙特卡洛随机取样(Monte Carlo sampling，MCS)法选取 80%的样本作校正集建立 PLS 回归模型，剩余部分作预测集，循环 2000 次，得到各样本的一组预测残差，得到各样本预测残差的均值(MEAN)与方差(STD)的 MEAN-STD 图，如图 5 所示，为了确定异常样本，绘制误差的火柴梗图，如图 6 所示。

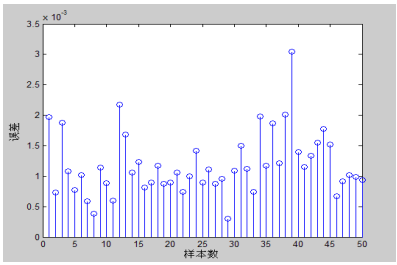

图 5　均值方差分布图

Fig.5　Mean Variance Distribution

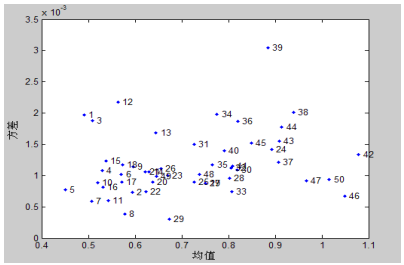

图 6　预测误差的火柴梗图

Fig.6　Stick Figure of forecast error

从图中可知，某些样本明显偏离主体样本，如 39，12 这些样本可视为奇异样本，应该剔除，由 MEAN-STD 图和火柴梗图确定出需要剔除异常样本，由表 2 可知，奇异样本剔除前后 PLS 校正模型的 RMSECV 的变化情况。MCCV 剔除异常样品后交叉验证计算值与化学分析值，如图 7 所示。

表2 剔除异常样本前后PLS校正模型交互校验结果

Tab.2 Cross-validation results of PLS calibration model after excluding abnormal samples

剔除样本编号	主成分数	R^2	RMSECV
	6	85.69	0.0672
39	8	93.42	0.0457
12、39	8	93.97	0.0434
1、12、39	8	94.64	0.0411
1、12、34、39	8	94.44	0.0423
1、3、12、34、39	8	93.90	0.0443

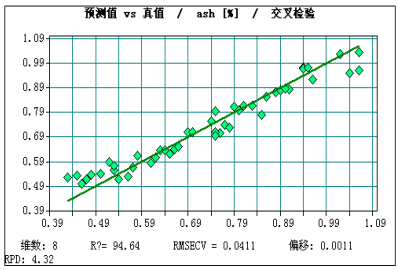

图 7　MCCV 法剔除异常样品后交叉验证计算值与化学分析值

Fig.7　Cross-validation calculated after MCVV method for abnormal samples removed and chemical analysis

由表 2 可知，剔除异常样品个数为 3，得到最好结果，相关系数(R2)为 94.64，交互验证均方差(RMSECV)为 0.0411。

2.4　预测模型的精度比较

为了验证剔除异常样本的准确性，对预测集的 10 个样本进行预测，预测结果，如图表 3 所示。真实值与预测值之间的相关图，如图 8-10 所示。

表3　剔除异常样本后校正模型的预测结果

Tab.2 Correction model predictions after excluding abnormal samples

剔除方法	剔除个数	主成分数	R^2	RMSECV	RMSEP
不剔	0	6	85.69	0.0672	0.0205
马氏	3	8	92.67	0.0485	0.0151
MCCV	3	8	94.64	0.0411	0.0127

由表 3 可知，用马氏距离法和 MCCV 法剔除异常样本后校正模型的精度和预测精度确实有所提高，MCCV 法剔除异常样本模型精度和预测精度提高的相对明显。

3 结 语

本次试验用马氏距离法和蒙特卡洛采样法分别对异常样本进行了剔除，用马氏距离法剔除异常样本，当权重系数为 1.5，剔除样本数为 3 时，得到最好结果，相关系数(R2)为 92.67，交互验证均方差(RMSECV)为 0.0485。MCCV 法剔除异常样本，剔除异常样本数为 3 时，得到最好结果，相关系数(R^2)为 94.64，交互验证均方差(RMSECV)为 0.0411。结果表明：马氏距离法剔除异常样本确实能提高校正模型的精度和预测精度，但 MCCV 法剔除异常样本模型精度和预测精度提高的相对明显。

未来与展望：在本次的试验中发现，虽然两种异常样本剔除方法都使模型精度得到提高并且剔除异常样本的个数相同，但是剔除的样本并不同，可能存在以下问题：一、马氏距离剔除异常样本只需要光谱数据而不需要样本的化学值，MCCV 法不仅需要光谱数据而且需要样本的化学值，可能存在由于人为误差导致化学值测量不准确，从而导致两种方法剔除不同的样本。二、两种方法的原理不同：马氏距离法，是通过光谱数据验证样本间的距离，MCCV 方法是经过 2000 次的 PLS 建模验证得到的结果，所以得到结果不同。由于是刚刚对剔除异常样本进行研究，所以做的都是验证工作，下一步的工作目标是找到问题存在的原因，并且寻找更好的异常样本剔除方法，从而提高预测模型的准确性和稳定性。

参考文献:

[1] 陆婉珍,袁洪福,徐广通.现代近红外光谱分析技术[M].北京: 中国石油化工出版社,2000:37-45.

[2] 倪永年.化学计量学在分析化学中的应用[M].北京:科学出版社,2004:304-310.

[3] 刘建学.实用近红外光谱分析技术[M].北京:科学出版社,2008:168-186.

[4] 白剑侠,赵欣.制粉厂的生产检验[J].技术粮食工程,2010,52(3):53-54.

[5] Atul D. Karande, Heng PaulWan Sia, Celine Valeria Liew. In-line quantification of micronized drug and excipients in tablets by near infrared (NIR) spectroscopy: Real time monitoring of tabletting process, International Journal of Pharmaceutics, vol.396, pp.63-74, 2010.

[6] CHEN Quansheng, JIANG Pei, ZHAO Jiewen. Measurement of total flavones content in snow lotus (Saussurea involucrate) using near infrared spectroscopy combined with interval PLS and genetic algorithm, Spectrochimica Acta Part A, vol .76, pp.50-55, 2010.

[7] QU Nan, ZHU Mingchao, MI Hong, et al. Nondestructive determination of compound amoxicillin powder by NIR spectroscopy with the aid of chemometrics, Spectrochimica Acta Part A, vol .70, pp. 1146–1151, 2008.

[8] 陈斌,邹贤勇,朱文静. PCA结合马氏距离法剔除近红外异常样品[J].江苏大学学报,2008,29(4):277-279.

[9] SHAO Yongni, HE Yong. Measurement of soluble solids and pH of Yogurt using visible/near infrared spectroscopy and chemometrics[J].Food Bioprocess Techno, 2009, 2(2): 229-233.

[10] Edword J. Graphical modelling and the mahalanobis distance [J]. Journal of Applied Statistics, 2005, 32 (9):959-967.

[11] 李水芳,单杨,范伟,等. 基于MCCV奇异样本筛选和CARS变量选择法对蜂蜜pH值和酸度的近红外光谱检测[J].食品科学学,2011,32(8):182-184.

[12] LIU Yande, Ying Yibin, JIANG Haiyan. Rapid determination of maturity in apple using outlier detection and calibration model optimization [J]. Transactions of the ASAB E, 2006, 49 (1):91-95.