第46卷第3期 中国科学技术大学学报 Vol. 46, No. 3
2016年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2016

文章编号:0253-2778(2016)3-0208-07

近红外光谱分析中的一种基于 *XY* 变量联合的异常样本剔除算法

尹宝全[1,2],史银雪[1],孙瑞志[1]

(1. 中国农业大学 农业部农业信息获取技术重点实验室,北京 100083; 2. 中国农业大学烟台研究院,山东烟台 264000)

摘要:在近红外光谱分析中,异常样本的存在会影响所建预测模型的性能.为了剔除异常样本,提高预测模型的预测能力,首先提出并证明了 *XY* 距离关系定理;在此基础上,设计了一种新型的基于 *XY* 变量联合的 ODXY 异常样本剔除算法.本次研究对 102 个羊肉样本的近红外光谱及其含水率进行了测定,在此样本集上分别采用常用的马氏距离剔除法、蒙特卡洛采样法和本文提出的 ODXY 算法对异常样品进行判别和剔除,并用剔除后的样本建立偏最小二乘预测模型;然后采用预测均方差 RMSEP 和决定系数 R^2 来检验模型的性能;最后,通过重新分配训练集和验证集检验算法的泛化能力.实验结果表明,在利用 ODXY 算法剔除预测样本的基础上建立的预测模型性能最佳,且具有更好的泛化能力.

关键词:异常样本;预测模型;近红外光谱;马氏距离;蒙特卡洛采样法

中图分类号:S132;O657.3 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.03.005

引用格式:YIN Baoquan, SHI Yinxue, SUN Ruizhi. An outlier sample eliminating algorithm based on joint *XY* distances for near infrared spectroscopy analysis[J]. Journal of University of Science and Technology of China, 2016,46(3):208-214.

尹宝全,史银雪,孙瑞志. 近红外光谱分析中的一种基于 *XY* 变量联合的异常样本剔除算法[J]. 中国科学技术大学学报,2016,46(3):208-214.

An outlier sample eliminating algorithm based on joint *XY* distances for near infrared spectroscopy analysis

YIN Baoquan[1,2], SHI Yinxue[1], SUN Ruizhi[1]

(1. *Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, China Agricultural University, Beijing* 100083, *China*;
2. *Yantai Academy of China Agricultural University, Yantai* 264000, *China*)

Abstract: Outlier samples in near infrared spectroscopy analysis can strongly influence on the performance of the prediction model. To detect and eliminate the outlier samples, a new outlier sample eliminating algorithm base on joint *XY* distances (ODXY) was presented, and the relation of *XY* distances of NIR is proposed and proved. In this research, 102 lamb samples were collected and the data of NIR spectroscopy and moisture content was measured and analyzed. Initially, Mahalanobis distances method, Monte-Carlo

收稿日期:2015-08-27;**修回日期**:2015-12-01

基金项目:新疆生产建设兵团科技支疆计划课题(2014AB037)资助.

作者简介:尹宝全,男,1972年生,博士生/讲师.研究方向:数据挖掘算法、农业信息化.

通讯作者:孙瑞志,博士/教授. E-mail: sunrz_cn@sina.com.cn.

sampling method and ODXY method to were employed to eliminate the outlier samples and built the PLS prediction model based on the processed samples. Then, the predictive mean square error (RMSEP) and the coefficient of determination (R^2) were used to test the performance of the prediction model. Finally, the generalization of the eliminating algorithm was tested by new calibration and validation sets. The experiments show that ODXY method has better performance and better generalization ability than the other methods tested in our experiments.

Key words: outlier samples; prediction model; near infrared spectroscopy; Mahalanobis distance; Monte-Carlo sampling method

0　引言

在利用数据挖掘进行多元分析预测过程中，预测模型的性能很大程度上依赖于原始数据的准确性[1]. 异常样本的存在影响了数据集的整体分布，降低了输出变量 Y 和输入变量 X 之间的相关性，最终将影响预测模型的精度和泛化能力，因此检测并剔除异常样本是建立可靠预测模型的重要步骤.

近红外光谱(near infrared spectroscopy, NIR)分析是典型的多元分析过程，近年来随着数据挖掘技术的发展，NIR 因其测量时快速、无损等特点被广泛地应用于农产品、化工、医药等领域[2-4]. 由于 NIR 光谱具有谱带重叠、吸收强度较低、光谱信噪比低以及光谱测定时易受样本状态及环境的影响而造成波动等特点，导致 NIR 光谱中的有效信息率较低，所以剔除异常样本在 NIR 分析中显得尤为重要.

本文在基于 NIR 的羊肉含水率无损检测的建模过程中，首先研究了目前常用的异常样本剔除算法，提出并证明了 XY 距离关系定理，在此基础上设计了一种基于 XY 变量联合的异常样本剔除算法，提高了预测模型的性能和可靠性.

1　相关工作

在 NIR 光谱分析中，异常样本通常来源于输入变量 X 值(光谱数据)的显著异常和输出变量 Y(在 NIR 分析中又称为化学值，在本文研究中即为被测羊肉的含水率)的显著异常. 同时，异常样本还应包括光谱数据 X 和输出变量 Y 之间的相关性显著异常.

为方便描述，首先定义如下：

n：样本数；

m：每条光谱的数据点数，即光谱矩阵的维数；

$Y(n)$：输出变量，为一维向量；

$\boldsymbol{X}(n\times m)$：光谱矩阵；

$\bar{\boldsymbol{X}}$：n 个样本的平均光谱，通过对光谱矩阵 $\boldsymbol{X}$ 按列求平均进行计算.

通常每条光谱有上百或上千个采样点构成，各点间具有较强的相关性，因此在对光谱数据进行分析之前，通常先对光谱数据进行主成分分析(principal component analysis，PCA)，设经 PCA 降维后得到的光谱矩阵为 $Xpca(n\times k)$，其中 k 为选出的主成分数.

NIR 光谱分析中异常样本剔除的方法有：马氏距离法[5]、蒙特卡洛采样法[6]、“二审”剔除法[7]，另外还有半数重采样(resampling by half-mean, RHM)法[8]、邻域等级差[9]、预测残差法以及聚类分析等方法.

1.1　马氏距离法

马氏距离(Mahalanobis distance，MD)是一种有效的计算两个未知样本集相似度的方法，与欧氏距离不同的是，它考虑了样本各维属性之间的关联关系，依赖于样本的总体分布. 在计算马氏距离时，要求总体样本数大于样本的维数. 由于 NIR 数据的维数 m 很大，所以在计算马氏距离前，通常先采用 PCA 对光谱数据进行降维.

第 $i(i=1,\cdots,n)$ 个样本到平均光谱的马氏距离 MD_i 计算如下[8]：

$$\mathrm{MD}_i=\left[\sum_{i=1}^{n}(x_i-\bar{x})^{\mathrm{T}}\boldsymbol{V}^{-1}(x_i-\bar{x})\right]^{\frac{1}{2}} \tag{1}$$

式中，$\boldsymbol{V}$ 为样本的协方差矩阵，且

$$\boldsymbol{V}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}} \tag{2}$$

利用马氏距离剔除异常样本的方法就是在计算出所有样本的马氏距离的基础上，通过设定阈值，将马氏距离超过阈值的样本判定为异常样本.

算法 1.1　MD 异常样本剔除算法.

输入：光谱矩阵 $\boldsymbol{X}(n\times m)$，阈值 threshold_$\boldsymbol{X}$

输出:异常样本号 OutlierNo

1 对光谱矩阵 X(n×m)利用 PCA 降维,得到 X_{PCA}(n×k);

2 对 Xpca 按列求平均得到平均光谱$\bar{X}$(n);

3 利用公式(1)、(2)求出所有样本的马氏距离 MD(n);

4 计算马氏距离 MD(n)的平均值 mean_MD 和标准差 std_MD;

5 FORi = 1 to n

6 IF abs(MD(i)-mean_MD)/std_MD) > threshold_X

7 该样本为异常样本,将 i 添加到 OutlierNo

8 ENDIF

9 END

10 返回 OutlierNo

MD 算法剔除异常样本只需要对光谱数据进行计算,算法比较简单快速,可以剔除由于光谱测量时人为误差造成的异常光谱,是 NIR 分析中剔除异常样本的常用算法. 其缺点是没有考虑输出变量 Y 的异常情况,可能导致:①某些光谱正常但 Y 异常的样本没有被剔除;②某些光谱由于样本化学值不同引起马氏距离超出阈值的正常样本被错判为异常样本. 针对以上情况的一种简单改进方法是在 MD 算法异常样本的判定条件中加入对输出变量 Y 异常的判定,只有当①光谱数据正常但输出变量 Y 异常;②输出变量 Y 正常但光谱数据异常时,才判定该样本为异常样本. 对光谱数据和输出变量同时异常的样本不作处理.

1.2 蒙特卡洛采样法

针对马氏距离法未考虑输出变量 Y 与输入变量 X 之间的关联性的缺点,蒙特卡洛采样(Monte-Carlo sampling, MCS)法通过随机选取一定比例(通常为 80%)的样本作为训练集,利用偏最小二乘法(partial least squares regression, PLS)建立预测模型,剩余样本作为验证集对模型进行验证,计算验证集中样本的预测误差,循环足够多次(如 2 000 次),保证每个样本均被预测过,得到每个样本的预测误差分布. 计算每个样本的预测残差的均值和方差,绘制样本的均值—方差分布图. 预测残差具有高均值或高方差的样本被认为是异常样本.

算法 1.2 MCS 异常样本剔除算法.

输入:光谱矩阵 $\boldsymbol{X}(n\times m)$,循环次数 numLoop

输出:异常样本号 OutlierNo

1 FOR i=1 to numLoop

2 随机抽样,将 X(n×m)按比例分成训练集 X_train 和验证集 X_test;

3 对 X_train 利用 PLS 建立预测模型;

4 利用建立的 PLS 预测模型;

5 计算验证集各样本的预测残差;

6 将预测残差按样本序号添加到预测残差矩阵 predictDiff 中;

7 END

8 按行计算预测残差矩阵中每个样本的 predictDiff 的均值 mean_Sample 和方差 Var_Sample;

9 绘制均值方差分布图;

10 根据均值方差分布图确定异常样本号 OutlierNo;

11 返回 OutlierNo

MCS 算法是一种基于预测残差的异常样本识别方法,它利用预测残差对异常样本敏感的特性,能够在一定程度上降低异常样本带来的掩蔽效应,与传统方法相比具有较高的识别异常样本的能力[10]. 缺点是算法运行时间太长.

2 基于 XY 变量联合的异常数据检测算法

NIR 光谱分析中的异常样本检测方法通常采用输入变量的值或输出变量的值是否超出正常范围来判断. 这类方法没有考虑输入变量和输出变量之间存在的关联性,易造成异常样本的误判或漏判;另外一类方法是利用部分样本构建预测模型,通过预测残差来判断样本是否异常. 该类方法考虑输出变量和输入变量之间的关联关系,但在构建验证模型时,训练集中不可避免地掺有异常样本,在此情况下所计算的预测残差并非样本的真实误差,导致判断错误. 此外,该类算法需要对样本进行一次或多次拟合,运行时间相对较长. 针对以上问题,本文从理论推导出发,找出近红外光谱分析中输入变量 X 和输出变量 Y 之间的关系,并在此基础上提出一种新的基于 XY 变量联合的异常样本检测和剔除(outlier detection base on joint X-Y distances, ODXY)算法.

定理 2.1 XY 距离关系定理

NIR 光谱分析中,任意样本的光谱到平均光谱的距离与该样本的化学值到化学平均值的距离成正比.

证明 首先假设第 i 个样本的光谱为一维向量 $X(i)$;其化学值为 $Y(i)$,且存在以下关系:

$$Y(i) = AX(i) + b \tag{3}$$

对所有样本求平均:

$$\frac{\sum_{i=1}^{n}Y(i)}{n}=\frac{\sum_{i=1}^{n}[AX(i)+b]}{n} \tag{4}$$

$$\frac{\sum_{i=1}^{n}Y(i)}{n}=A\frac{\sum_{i=1}^{n}[X(i)]}{n}+b \tag{5}$$

$$\bar{Y}=A\bar{X}+b \tag{6}$$

由式(3)～式(6)可得

$$Y(i)-\bar{Y}=A(X(i)-\bar{X}) \tag{7}$$

由于 $Y(i)-\bar{Y}$ 是一个数,故有

$$\frac{A(X(i)-\bar{X})}{Y(i)-\bar{Y}}=1 \tag{8}$$

式(8)两边左乘 A^{-1},得

$$\frac{X(i)-\bar{X}}{Y(i)-\bar{Y}}=A^{-1} \tag{9}$$

对式(9)两边取 $L2$ 范数,并设

$\mathrm{d}Y(i)=\|Y(i)-\bar{Y}\|$;$\mathrm{d}X(i)=\|X(i)-\bar{X}\|$,有

$$\frac{\mathrm{d}X(i)}{\mathrm{d}Y(i)}=\|A^{-1}\|=C \tag{10}$$

式中,C 为常数.

因为 $X(i)-\bar{X}$ 的 $L2$ 范数等效于 $X(i)$到 $\bar{X}$ 的欧氏距离,因此式(10)的几何意义为:任意一个样本的光谱到所有样本的平均光谱之间的距离与该样本的化学值到所有样本化学值的平均值之间的距离的比值是一个常数,即二者成正比.

补充说明:①为消去因 $\mathrm{d}X$,$\mathrm{d}Y$ 权重不同造成的影响,设计算法时分别对其作除以各自最大值的处理. ②光谱分析中,由于异常值的影响,通常用中位数来代替平均值,所以在判定 $\mathrm{d}X/\mathrm{d}Y$ 是否超出正常范围时,采用中位数来替代平均值.

根据定理 2.1,设计基于 XY 变量联合的异常样本剔除算法如下:

算法 2.1　ODXY 异常样本剔除算法.

输入:光谱矩阵 $\boldsymbol{X}(n\times m)$,输出变量 $Y(n)$,阈值 threshold

输出:异常样本号 OutlierNo

1 对光谱矩阵 X(n×m)利用 PCA 降维,得到 X_{PCA}(n×k);

2 对 Xpca 按列求平均得到平均光谱$\bar{X}$(n);

3 对 Y(n)求平均,得到$\bar{Y}$;

4 FOR i = 1 to n

5 dX(i) = ‖X_{PCA}(i)−$\bar{X}$‖;

6 dY(i) = ‖Y(i)−$\bar{Y}$‖;

7 END

8 dX_normal = dX / max(dX);

9 dY_normal = dY. / max(dY);

10 dX_dY= dX_normal / dY_normal;

11 计算 dX_dY 的中位数 median 和标准偏差 std;

12 FORi = 1 to n

13 IF abs(dX_dY(i)−median)/std > threshold

14 该样本为异常样本,将 i 添加到 OutlierNo;

15 ENDIF

16 END

17 返回 OutlierNo

3　实验

由于 NIR 分析具有快速、无损检测等优点,基于 NIR 技术的冷鲜肉含水率的无损检测越来越受到关注,但用于羊肉含水率无损检测的研究仍然较少. 本研究的目的是确定羊肉含水率与近红外光谱之间的相关关系,建立 NIR 光谱与羊肉含水率之间的预测模型,以期找到一种基于 NIR 技术的快速无损检测羊肉含水率的方法.

3.1　实验仪器

光谱数据的获取方法:本研究使用自主研发的近红外光谱仪进行光谱采集,仪器光谱范围 900～1688nm,128 个采样点. 光谱扫描时,每个样本扫描 3 次求平均作为该样本的光谱数据.

羊肉含水率的获取方法:在光谱照射区域取小片羊肉,采用电热恒温鼓风干燥箱和电子天平进行测量并计算该样本的含水率,重复 4 次求平均值.

样品来源:实验用样本均为当天从当地超市购买的冷鲜羊肉,并放冰箱冷藏室保鲜,实验时从冰箱取出采样.

3.2　实验步骤及方法

本次实验对 2015 年 4 月到 5 月期间采集的 102 个样本进行分析. 样本剔除、数据预处理及建模均采用 Matlab 2013 编程实现.

(Ⅰ)数据预处理:近红外光谱通常伴随着许多高频随机噪声、基线漂移、光散射等噪声信息,直接影响所建模型的可靠性和稳定性,因此需对羊肉样本的原始光谱进行预处理,以减少噪声,提高信号的分辨率和灵敏度. 为达到对三种异常样本剔除算法进行比较的目的,预处理算法均采用 S-G 平滑＋标

准正态变量变换(SNVT)+去趋势化处理.

(Ⅱ)异常样本的剔除算法分别采用NIR分析中常用的MD算法和MCS算法以及本文提出的ODXY算法.实验时还曾采取MD+输出变量阈值判断、XY拟合剔除等方法,但对本次实验样本集效果并不明显.因此,本文中实验对比结果中只对MD算法、MCS算法和ODXY算法的实验结果进行对比.

(Ⅲ)训练集和验证集划分算法选择.分别采用随机(RS)法、KS法和SPXY法进行划分训练集和验证集.SPXY样本划分算法[11]同本文提出的ODXY样本剔除算法类似,均是在联合考虑X/Y的基础上进行判断.实验对比结果显示,SPXY法划分出的训练集的含水率范围及光谱范围更能覆盖整个样品集,建模结果更为合理.由于样本集划分方法不是本文讨论的重点,因此后面的对比实验中样本集的划分方法均直接采用SPXY算法.

(Ⅳ)建模算法.PLS克服了光谱数据间的高度相关性,较好地解决了样本个数少于变量个数等问题,是光谱分析中最常用的建模方法[12-14].实验中选用Matlab 2013提供的PLS算法.

(Ⅴ)模型评价.为对比三种剔除算法对建模的影响,本文首先对数据预处理后,利用训练集样本建立PLS预测模型;然后利用已建立的预测模型对验证集样本进行预测,比较验证集中羊肉含水率的预测值和实测值之间的关系,计算预测均方差RMSEP和预测决定系数R^2.通过对比所建预测模型的RMSEP和R^2来实现对剔除算法性能的评价.

通常,一个好的预测模型应该具有高的R^2值和较低的RMSEP值.决定系数R^2越接近1,说明预测值和实测值的拟合程度越好.

3.3 结果

实验中所采用的羊肉样本近红外光谱如图1所示.

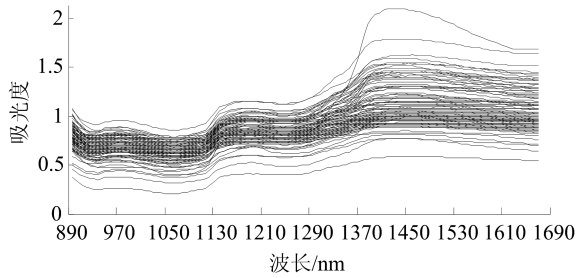

图1 羊肉样本的原始光谱

Fig. 1 The original spectra of mutton samples

首先,对所用训练集样本与验证集样本按3∶1的比例进行划分.

然后,在不剔除任何样本的情况下,对所有样本利用3.2节的建模方法进行建模预测,得到验证集样本的测量值与预测值的相关图如图2所示.预测均方差RMSEP为0.541,决定系数R^2为0.847.

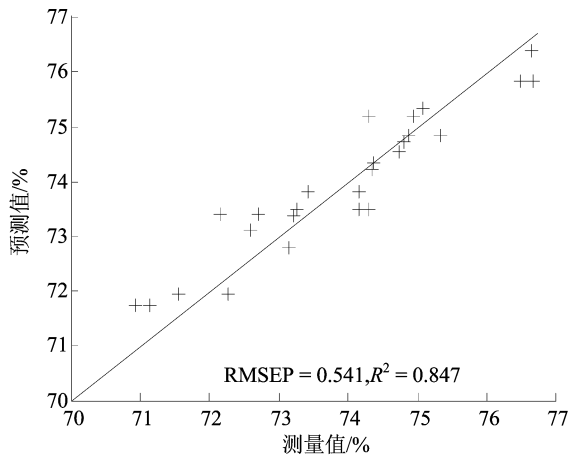

图2 未剔除样本的模型测量值与预测值的相关图

Fig. 2 Measured values and predicted values of the model without eliminating samples

利用MD算法剔除样本时,建立的预测模型的测量值与预测值的相关图如图3所示.预测均方差RMSEP为0.483,决定系数R^2为0.866.

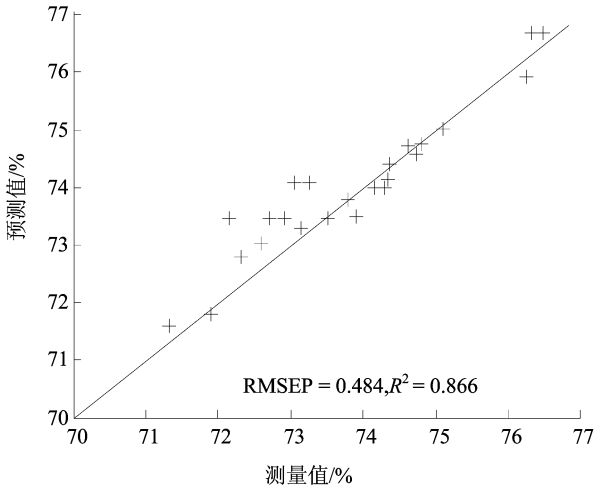

图3 MD算法剔除样本的模型测量值与预测值的相关图

Fig. 3 Measured values and predicted values of the model in which samples eliminated by MD algorithm

利用MCS算法剔除样本时,建立的预测模型的测量值与预测值的相关图如图4所示.预测均方差RMSEP为0.516,决定系数R^2为0.871.

利用ODXY算法剔除样本时,建立的预测模型

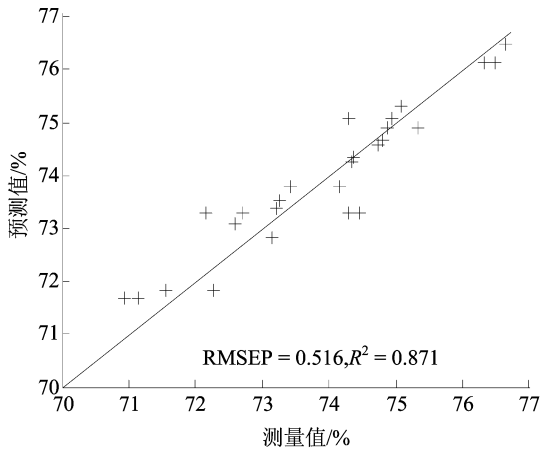

图 4　MCS 算法剔除样本的模型测量值与预测值的相关图

Fig. 4　Measured values and predicted values of the model in which samples eliminated by MCS algorithm

的测量值与预测值的相关图如图 5 所示. 预测均方差 RMSEP 为 0.354,决定系数 R^2 为 0.924.

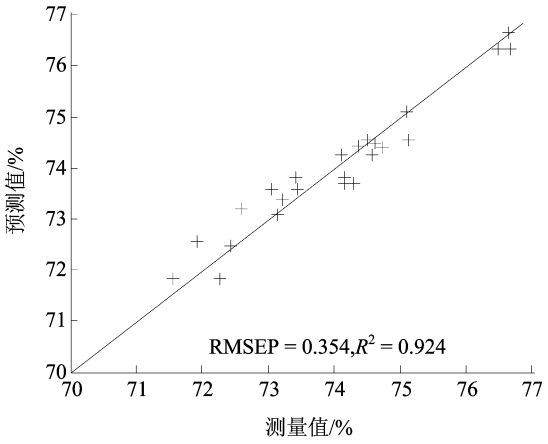

图 5　ODXY 算法剔除样本的模型测量值与预测值的相关图

Fig. 5　Measured values and predicted values of the model in which samples eliminated by ODXY algorithm

结果表明,MD 算法、MCS 算法和 ODXY 算法三种异常样本剔除算法均能提高所建模型的性能. 利用 ODXY 法剔除异常样本所得到的预测模型具有更低的预测均方差 RMSEP 和更高的决定系数 R^2,即利用 ODXY 法剔除异常样本较其他两种算法更能提高预测模型的性能.

为了进一步检验 ODXY 剔除算法的泛化性能,通过重新划分训练集和验证集,对以上三种异常样本剔除算法进行比较. 因为已确定采用 SPXY 方法作为样本划分算法,为获得不同的训练集和验证集,实验采用改变训练集和验证集的比例的方式实现,同时综合考虑验证集样本较多和较少两种情况,训练集和验证集的比例分别取 3∶1、4∶1、9∶1 和 10∶1,对比结果如表 1 所示.

表 1　MD/MCS/ODXY 剔除算法比较

Tab. 1　Comparison of MD, MCS and ODXY algorithms

样本划分	性能指标	MD	MCS	ODXY
3∶1	RMSEP	0.483	0.516	0.354
	R^2	0.866	0.871	0.924
4∶1	RMSEP	0.392	0.570	0.351
	R^2	0.886	0.840	0.917
9∶1	RMSEP	0.501	0.500	0.398
	R^2	0.852	0.851	0.858
10∶1	RMSEP	0.510	0.566	0.312
	R^2	0.866	0.817	0.908

本次实验表明,ODXY 异常样本剔除算法较 MD 法、MCS 法在以上所有样本划分比例的情况下所得到的预测模型都具有更小的 RMSEP 和更大的 R^2 值,说明 ODXY 算法较 MD、MCS 算法均有更好的泛化性能.

ODXY 算法较实验中的对比算法具有更好结果,原因可能是:较 MD 方法,ODXY 不仅考虑了输出变量 Y 的异常,同时考虑了输出变量 Y 与输入变量 X 之间的关系;MCS 方法是基于 X,Y 拟合后的模型基础上进行分析,但由于异常样本的存在导致拟合模型本身存在误差,从而可能出现误判的情况.

4　结论

本文经推导得出一个 NIR 分析中光谱与化学值之间关系的 XY 距离关系定理,并在此定理的基础上提出了 ODXY 异常样本剔除算法. 实验表明,本文提出的 ODXY 异常样本算法对异常样本具有更好的判别能力,且对不同样本集的具有更好的泛化性能.

此外,ODXY 算法具有一定的理论基础,算法简单,在运算速度方面,与 MD 算法运算速度相当,较 MCS 算法运行速度快.

下一步工作,将 ODXY 算法在更多数据集中进一步验证.

参考文献(References)

[1] KOTEESWARAN S, VISU P, JANET J. A review on clustering and outlier analysis techniques in data mining[J]. American Journal of Applied Sciences, 2012, 9(2): 254-258.

[2] ZOU X B, ZHAO J W, POVEY M J W, et al. Variables selection methods in near-infrared spectroscopy[J]. Analytica Chimica Acta, 2010, 667(1-2):14-32.

[3] MOUROT B P, GRUFFAT D, DURAND D, et al. Breeds and muscle types modulate performance of near-infrared reflectance spectroscopy to predict the fatty acid composition of bovine meat[J]. Meat Science, 2015, 99(99):104-112.

[4] TALENS P, MORA L, MORSY N, et al. Prediction of water and protein contents and quality classification of Spanish cooked ham using NIR hyperspectral imaging[J]. Journal of Food Engineering, 2013, 117(3): 272-280.

[5] REEVES J B, VAN KESSEL J S. Near-infrared spectroscopic determination of carbon, total nitrogen, and ammonium-N in dairy manures[J]. Journal of Dairy Science, 2000, 83(8): 1829-1836.

[6] 刘智超，蔡文生，邵学广. 蒙特卡洛交叉验证用于近红外光谱奇异样本的识别[J]. 中国科学(B辑:化学)，2008，38(4):316-323.

[7] 祝诗平，王一鸣，张小超,等. 近红外光谱建模异常样品剔除准则与方法[J]. 农业机械学报，2004，35(4): 115-119.
ZHU Shiping, WANG Yiming, ZHANG Xiaochao, et al. Outlier sample eliminating criterions and methods for building calibration model of near infrared spectroscopy analysis[J]. Transactions of the Chinese Society for Agricultural Machinery, 2004, 35(4): 115-119.

[8] 赵振英，林君，张怀柱. 近红外光谱法分析油页岩含油率中异常样品识别和剔除方法的研究[J]. 光谱学与光谱分析，2014，34(6)：1707-1710.
ZHAO Zhenying, LIN Jun, ZHANG Huaizhu. Research on outlier detection methods for determination of oil yield in oil shales using near-infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2014, 34(6): 1707-1710.

[9] BHATTACHARYA G, GHOSH K, CHOWDHURY A S. Outlier detection using neighborhood rank difference[J]. Pattern Recognition Letters, 2015, 60(C): 24-31.

[10] NURUNNABI A, WEST G, BELTON D. Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data[J]. Pattern Recognition, 2015, 48(4): 1404-1419.

[11] REIS M M, ROSENVOLD K. Early on-line classification of beef carcasses based on ultimate pH by near infrared spectroscopy [J]. Meat Science, 2014, 96(2): 862-869.

[12] ALAMPRESE C, CASALE M, SINELLI N, et al. Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy[J]. LWT-Food Science and Technology, 2013, 53(1): 225-232.

[13] KAMRUZZAMAN M, SUN D W, ELMASRY G, et al. Fast detection and visualization of minced lamb meat adulteration using NIR hyperspectral imaging and multivariate image analysis[J]. Talanta, 2013, 103(2): 130-136.

[14] GALVAO R K, ARAUJO M C U, JOSÉ G E, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736-740.