第39卷,第12期
2019年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 39, No. 12, pp3783-3787
December, 2019

SVDD的近红外光谱定性分析光谱质量判定方法

李浩光[1,2], 于云华[1,2], 沈学锋[1,2], 逄 燕[1]

1. 中国石油大学胜利学院, 山东 东营 257061
2. 中国石油大学(华东)控制科学与工程学院, 山东 东营 257061

摘 要 近红外光谱属微弱信号, 其质量易受被测物体自身状态及各种外界因素干扰, 具体而言, 在近红外光谱定性分析中, 影响光谱质量的因素主要有光谱仪状态改变、光谱采集人员错误操作、奇异样本干扰等。建模时若混入质量较差的光谱易影响所建模型的稳健性与适用性, 因此光谱质量判定是确保模型预测能力的一项重要工作。目前用于定量分析的光谱质量判定研究较多, 而用于定性分析的光谱质量判定研究较少, 为此, 提出一种基于支持向量机数据描述的近红外光谱定性分析光谱质量判定方法, 采用自制漫透射近红外光谱装置采集单籽粒玉米光谱, 以正常状况下采集的某品种玉米单籽粒漫透射光谱作为正常样本, 而人为漏光、近红外探测器窗口覆盖玉米表皮碎屑、光源强度改变、光源与被测玉米籽粒距离改变、相近品种玉米籽粒混入等几种情况下所采集光谱作为异常样本, 在此数据集基础上研究了基于支持向量机数据描述的定性分析光谱质量判定模型建立的原理与方法, 其后将支持向量机数据描述方法与常用的马氏距离法、局部异常因子法等光谱质量判定方法进行了对比, 并以正常样本正确识别率与异常样本正确拒识率的均值作为评价标准, 对实验结果进行分析, 由实验结果可以看出相比其他两种方法, 基于支持向量机数据描述的光谱质量判定方法具有最优判定能力, 建模集正常样本数目会影响光谱质量判定能力, 在实际使用光谱质量判定方法时, 建模集应包含足量样本。在近红外定性分析时可以将该方法作为剔除异常光谱的手段, 在预处理、特征提取, 模式分类等近红外光谱定性分析步骤前首先进行基于支持向量机的光谱质量判定步骤, 并剔除异常光谱, 可有效提高近红外光谱定性分析模型的可靠性, 亦为近红外光谱定性分析光谱质量判定提供新的方法参考。

关键词 近红外光谱; 定性鉴别; 质量判定; 支持向量数据描述

中图分类号: O657.3 **文献标识码**: A **DOI**: 10.3964/j.issn.1000-0593(2019)12-3783-05

引 言

近红外光谱分析由于无损、快捷、低成本等优点在食品、药品、纺织、石油石化、农业等领域取得了广泛应用[1]。近红外光谱分析据其用途可分定量和定性分析两种, 两方法所建模型预测能力均取决于光谱质量, 因此光谱质量判定对于提升模型性能至关重要[1-2], 而近红外光谱属于微弱信号, 其质量易受被测物自身及多种外界因素干扰。

近红外定性分析光谱质量主要受如下因素影响: (1)测量仪器状态改变: 通常近红外光谱仪在出厂时经过校验, 性能有保证, 但随仪器使用时间延长, 光谱仪中某些易损部件出现老化, 影响光谱质量。如光谱仪中的卤素光源强度发生变化, 导致光谱质量下降。(2)操作人员错误操作: 操作人员因疲劳或错误操作, 采集光谱时装样方式发生明显改变造成光谱异常。如整杯漫反射采集方式中, 正常采样时应装满测试杯进行测量, 若装样时仅浅层覆盖测试杯底部, 则所采光谱出现异常。(3)奇异样本干扰: 若某品种中包含部分与其他品种近似的个体, 这些个体光谱与本品种个体光谱有一定差异, 则属奇异样本, 易影响模型性能。上述情况下采集的近红外光谱质量较差, 易对近红外定性分析模型的预测性能、鲁棒性与适用性产生较大影响, 因此, 在定性分析特征提取及模式分类步骤前, 应首先判定近红外光谱质量, 并剔除质量较差的光谱。

目前针对近红外定量分析光谱质量判定研究相对较多[3], 而针对定性分析模型的光谱质量判定方法研究较少。

收稿日期: 2019-01-22, **修订日期**: 2019-05-04

基金项目: 国家重大科学仪器设备开发专项(2014YQ470377), 山东省教育厅科技计划项目(J18KA329), 中国石油大学胜利学院科技计划项目(KY2017006)资助

作者简介: 李浩光, 1981年生, 中国石油大学胜利学院副教授 e-mail: lihaoguang@upc.edu.cn

定量分析中常用光谱质量判定方法只有马氏距离法(Mahalanobis distance，MHD)及局部异常因子方法(local outlier factor，LOF)可用于定性分析，其他方法大多需用到定量分析时的真值(定标值)，并不适用于近红外光谱定性分析方法，因此研究用于近红外定性分析的光谱质量判定方法对于提升近红外定性分析的模型性能具有重要意义。故提出了一种基于支持向量机数据描述的光谱质量判定方法，并以玉米单籽粒漫透射光谱数据为例，通过向正常光谱中掺杂实际可能出现的异常光谱，并对所提出的方法与其他光谱质量判定方法进行对比实验研究。

1 近红外光谱质量判定模型设计

1.1 模型原理

支持向量数据描述方法(support vector data description，SVDD)本质是一种单分类方法，SVDD 方法通过核映射在高维空间构建涵盖目标光谱样本并拒绝非目标光谱样本的最小超球实现异常检测。若训练光谱样本数据集为$\{x_i\}_{i=1}^n$，则求解包含大多数光谱样本最小超球如式(1)所示

$$\min\left[R^2+C\sum_{i=1}^{n}\xi_i\right] \tag{1}$$

$$\text{s. t. } \| x_i-c \|^2 \leqslant R^2+\xi_i,\ i=1,2,\cdots,n$$

式(1)中，c 为超球中心，R 为超球半径，松弛变量 $\xi_i\geqslant 0$，C 为正则化系数，用于限制错分样本的惩罚程度。式(1)可转化为拉格朗日极值问题，即

$$L(c,R,\xi,\alpha,\beta)=R^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(R^2+\xi_i-\| x_i-c \|^2)-\sum_{i=1}^{n}\beta_i\xi_i \tag{2}$$

式(2)中 α 与 β 为拉格朗日乘子，将式(2)转化为对偶问题，即可得式(3)

$$L(c,R,\xi,\alpha,\beta)=\sum_{i}^{n}\alpha_i K(x_i,x_j)-\sum_{i}^{n}\alpha_i\alpha_j K(x_i,x_j) \tag{3}$$

$$\text{s. t. } \sum_{i}^{n}\alpha_i=1,\ 0\leqslant\alpha_i\leqslant 1$$

式(3)中，核函数 $K(x_i,x_j)=(\phi(x_i),\phi(x_j))$，设其核函数参数为 σ，易知式(3)是个二次规划问题，求得其最小值，即可求得 α_i 的最优解 α_i^*。

若 $\| x_i-a \|^2<R^2$，则 $\alpha_i=0$，说明此时样本 x_i 位于超球的表面或者内部；若 $\| x_i-a \|^2>R^2$，则 $\alpha_i=C$，此时样本 x_i 位于超球的表面或外侧。其中大部分的 α_i^* 都是零，仅仅有一小部分的 α_i^* 不为零，通常将不为零的 α_i^* 所对应样本 x_i 称为支持向量，并可表示成 x_{sv}。

此时最小超球体半径可通过式(4)求得，

$$R^2=K(x_{sv},x_{sv})-2\sum_{i=1}^{n}\alpha_i^*K(x_{sv},x_i)+\sum_{i,j=1}^{n}\alpha_i^*\alpha_j^*K(x_i,x_j) \tag{4}$$

包裹正常光谱样本的超球球心可由式(5)求得

$$c=\sum_{i=1}^{n}\alpha_i^*x_i \tag{5}$$

对于未知样本 x_i，其与超球体球心的距离可由(6)式求得

$$\| x_i-c \| =K(y,y)-2\sum_{i=1}^{n}\alpha_i^*K(y,x_i)+\sum_{i,j=1}^{n}\alpha_i^*\alpha_j^*K(x_i,x_j) \tag{6}$$

若光谱样本满足 $\| x_i-c \|^2>R^2$，可将该样本判定为异常样本；若光谱样本满足 $\| x_i-c \|^2\leqslant R^2$，则可将该样本判定为正常样本。SVDD 通过对光谱数据超球以及半径的学习，并优化正则化参数 C 以获取最优光谱质量判定效果。

图 1 是基于 SVDD 的光谱质量判定方法流程图。首先在仪器正常状态下，按照规范操作采集一定数量的被测样品正常光谱样本，将所有光谱样本归一化后，进行主成分特征提取，其后使用 SVDD 方法建立光谱质量判定模型。使用模型时，首先采集待判定的光谱样本，经过归一化、PCA 特征提取后，调用基于 SVDD 的光谱质量判定模型，并对特征提取后的向量进行质量判定。

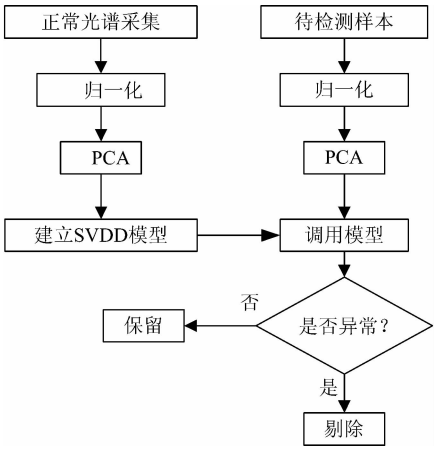

图 1 基于 SVDD 的光谱质量判定流程图

Fig. 1 Flow chart of spectral quality determination based on SVDD

若判定为异常光谱则将其从光谱数据集中剔除，否则，将其保留在光谱数据集中，光谱数据集经过光谱质量判定后即可用于后期分类鉴别。

2 实验验证及结果分析

用自制实验装置[4-6]依序采集各玉米籽粒的漫透射近红外光谱，每颗籽粒采集 10 条光谱，并对 10 条光谱取平均，经过光谱仪自带软件预处理后得到 125 维光谱向量，后期数据分析处理使用 Matlab2016a。

以单籽粒漫透射采集方式为例，导致光谱异常的原因有漏光、光源强度改变、光源与被测物之间距离改变、奇异样本干扰等，为检验光谱质量判定方法的有效性，可通过人为改变光源强度、光源与被测物距离等手段来产生实际测量中

可能出现的异常光谱。正常光谱样本采集时玉米籽粒需完全覆盖光阑小孔，确保光谱仪采集所得为贯穿玉米籽粒的近红外光，而非光源直接照射至检测器窗口的近红外光或者杂散光。

实验中采集的正常光谱与异常光谱如表 1 所示，为叙述方便，正常光谱与异常光谱均使用代码表示。使用上述数据作为实验数据集，将正常光谱的一半作为训练集，正常光谱的另一半与所有异常光谱作为测试集，分别使用 SVDD 法、LOF 法、MHD 方法建立判定模型对测试集光谱样本进行质量判定对比实验。

表 1　光谱质量判定实验数据集说明表

Table 1　Data set description table for spectral quality determination

光谱类型	光谱代码	数量/条
正常光谱(农华 205)	N1	600
探测器窗口覆盖碎屑	AN1	20
光源强度减弱(卤钨灯电压升高 0.5 V)	AN2	20
光源强度增大(卤钨灯电压降低 0.5 V)	AN3	20
光源距离升高 0.5 cm	AN4	20
光源距离降低 0.5 cm	AN5	20
品种奇异光谱 1(农华 206)	AN6	20
品种奇异光谱 2(农华 98)	AN7	20
漏光	AN8	20

表 2 为采用 SVDD，LOF 和 MHD 法对上述实验数据进行光谱质量判定所得结果。

表 2　三种质量判定方法对比

Table 2　Comparison of three quality determination methods

类型	数量	SVDD		MHD		LOF	
		正确个数	CRR/%	正确个数	CRR/%	正确个数	CRR/%
AN1	20	19	95	17	85	18	90
AN2	20	18	90	17	85	18	90
AN3	20	20	100	18	90	18	90
AN4	20	19	95	18	90	17	85
AN5	20	18	90	16	80	15	75
AN6	20	19	95	10	50	11	55
AN7	20	18	90	9	45	12	60
AN8	20	20	100	20	100	20	100

图 2 为含异常光谱的光谱曲线图，图中使用不同颜色区分异常度较为明显的漏光(绿)、光源距离改变(蓝)、光源强度改变(红)等几种类型的异常光谱，图中黑色曲线为正常光谱。

由表 2 可知，SVDD 对异常光谱剔除能力高于其他两种方法，分析如下：

(1)传感器窗口覆盖碎屑 AN1、光源强度改变 AN2—AN3、光源距离改变 AN4—AN5：SVDD 方法的正确拒识率可达 90%以上，剔除效果较好。LOF 及 MHD 对于 AN1—AN5 异常光谱识别率在 80%～90%之间波动。说明 LOF 及 MHD 方法能够剔除上述类型的大部分异常光谱，但仍有部分 AN1—AN5 类型的异常光谱无法判定剔除。

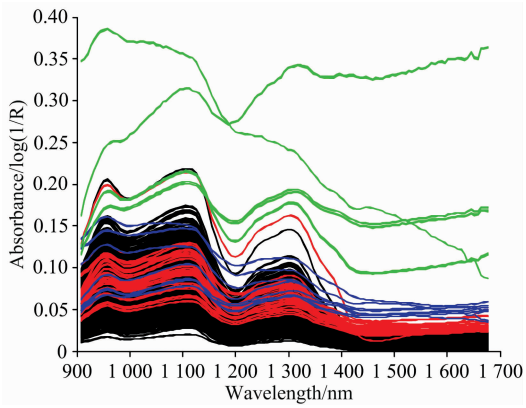

图 2　含异常光谱的光谱曲线

Fig. 2　Spectral curves with abnormal spectra

(2)近似品种模拟品种奇异光谱 AN6—AN7：SVDD 方法正确拒识率达到 95%以上，LOF 方法正确拒识率为 55%～60%，而 MHD 方法正确拒识率为 50%左右。可见，SVDD 方法对品种奇异样本的质量判定能力显著高于其他两种判定方法。

(3)漏光异常光谱 AN8：3 种光谱质量判定方法均可准确剔除该类型异常光谱。因该种情形下，玉米籽粒摆放未能完全覆盖光阑小孔，导致近红外光源透过缝隙直接照射传感器窗口，导致光谱异常，从图 2 可以看出该类型异常光谱与正常光谱差异明显，因此 LOF 法及 MHD 方法也能够对该类型异常光谱进行有效判定，正确拒识率能够达到 100%。

上述结果说明：LOF 法及 MHD 法能较有效地判定 AN1—AN5 及 AN8 类型异常光谱，但无法剔除 AN6—AN7 类型的品种奇异光谱，而 SVDD 方法通过非线性变换构建包含正常样本的最小超球，实现了对 AN1—AN8 类型异常光谱的有效判定剔除。

图 3 和图 4 是建模集中包含正常样本数量逐级增加时，SVDD 及 LOF 法、MHD 方法对正常光谱的正确识别率与对异常光谱的正确拒识率变化曲线，建模集中初始正常样本数量设置为 100 条。由图 3 和图 4 可见，3 种方法在建模集包

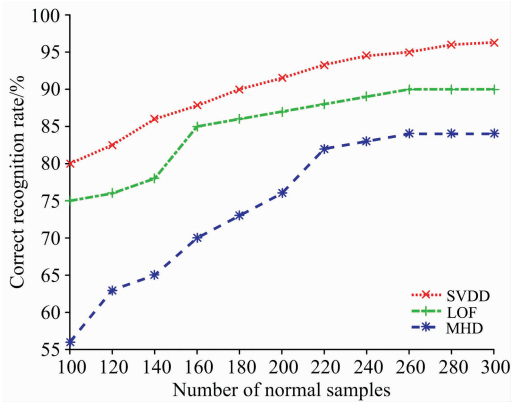

图 3　正确识别率变化曲线

Fig. 3　Correct recognition rate curves

含正常光谱数量递增时，光谱质量判定模型对异常光谱拒识能力与对正常光谱识别能力均有上升趋势，SVDD 方法对正常光谱正确识别率与对异常光谱正确拒识率均高于 MHD 与 LOF 方法。

当正常光谱数量大于 200 条时，3 种方法所得正确拒识率与正确识别率均趋于稳定，SVDD 的正确拒识率与正确识别率都在 95%左右，LOF 方法的正确识别率为 90%，其拒识率在 79%左右，而 MHD 方法的正确识别率为 85%，其拒识率为 70%左右。

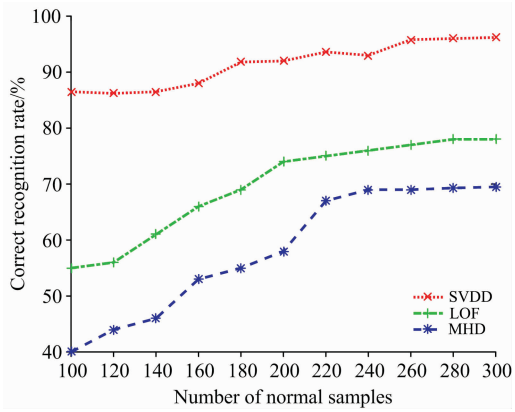

图 4 正确拒识率变化曲线

Fig. 4 Correct rejection rate curves

综上，建模集正常样本数目会影响光谱质量判定能力，在实际使用光谱质量判定方法时，建模集应包含足量样本。足量样本能够使模型具有较好鲁棒性，也能够在剔除异常样本时使得光谱质量判定模型具有更强判定性能，在仪器正常情况下，所采光谱中异常光谱占比相对较小，随着采集光谱总量增加，正常光谱的所占比例也会相应增大，此时，SVDD 法能够更精准感知正常光谱在高维空间的分布，并构建包含正常样本的超球，实现对异常光谱的判定。

经 SVDD 光谱质量判定模型判定并剔除异常光谱的光谱曲线如图 5 所示，从图 5 中已无法观测到明显异常光谱，说明 SVDD 对实验中所采集的几种异常光谱具有较强判定剔除能力。

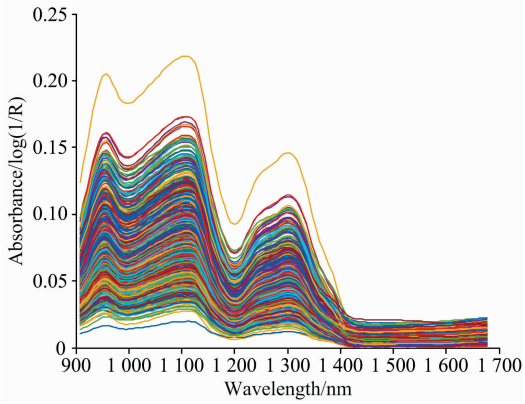

图 5 剔除异常光谱后的近红外光谱曲线图

Fig. 5 Near infrared spectral curves after removing abnormal spectra

3 结 论

在近红外光谱定性分析时，光谱质量判定是保证模型期性能的重要环节。以正常情况下采集的某品种玉米单籽粒漫透射光谱作为训练集正常样本；以人为漏光、近红外探测器窗口覆盖玉米碎屑、光源强度改变、采集距离改变、近似品种玉米籽粒模拟的品种奇异样本等几种情况下所采光谱作为异常样本，通过 SVDD 等光谱质量判定方法进行异常光谱样本判定与剔除，并进行对比实验，实验结果表明，SVDD 方法具有最优光谱质量判定能力。

References

[1] YAN Yan-lu, CHEN Bin, ZHU Da-zhou(严衍禄，陈 斌，朱大洲). Near Infrared Spectroscopy Analytical-Principles, Technology and Application(近红外光谱分析的原理、技术与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社)，2007.

[2] SHI Bo-lin, ZHAO Lei, LIU Wen, et al(史波林，赵 镭，刘 文，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报)，2010，41(2)：132.

[3] SHI Lu-zhen, ZHANG Jing-chuan, WANG Yan-qun(石鲁珍，张景川，王彦群). Journal of Chinese Agricultural Mechanization(中国农机化学报)，2016，36(6)：99.

[4] QIN Hong, MA Jing-yi, CHEN Shao-jiang, et al(覃 鸿，马竞一，陈绍江，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析)，2015，25(11)：1807.

[5] QIN Hong, MA Jing-yi, CHEN Shao-jiang, et al(覃 鸿，马竞一，陈绍江，等). Infrared Technology(红外技术)，2015，1(37)：78.

[6] LI Hao-guang, LI Wei-jun, QIN Hong, et al(李浩光，李卫军，覃 鸿，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报)，2016，47(6)：259.

Research on NIR Spectra Quality Detection Method Based on Support Vector Data Description

LI Hao-guang[1,2], YU Yun-hua[1,2], SHEN Xue-feng[1,2], PANG Yan[1]

1. Shengli College, China University of Petroleum, Dongying 257061, China

2. College of Information and Control Engineering, China University of Petroleum, Dongying 257061, China

Abstract Near infrared spectroscopy (NIR) is a weak signal, and its spectral quality is easily disturbed by the state of the measured object and various external factors. Specifically, the spectral quality in the qualitative analysis of NIR is mainly affected by the state change of measuring instrument, wrong operation, and the interference of singular samples. The robustness and applicability of the model are easily affected by the incorporation of poor quality spectra, so spectral quality determination is of vital importance to ensure the model prediction ability. At present, there are many studies on the determination of spectral quality for quantitative analysis, but few studies on the determination of spectral quality for qualitative analysis. In this paper, a method for the determination of spectral quality for near-infrared qualitative analysis based on data description of support vector is proposed. A self-made diffuse reflectance NIR acquisition device is used to collect the spectra of single-grain maize as an experimental object, and under normal conditions, the diffuse transmission spectra of a maize single grain were collected as normal samples, while the collected spectra were used as abnormal spectra under the conditions of artificial light leakage, near infrared detector window covering maize epidermis debris, intensity change of light source, distance change between light source and tested maize grain, and mixture of similar maize seeds. On this basis, the determination based on support vector data description (SVDD) was studied. The principle and method of establishing spectral quality judgment model were analyzed. Because the parameters of kernel function and regularization have important influence on the performance of spectral quality judgment model based on SVDD, the combination of grid search and cross validation was used to optimize the parameters of kernel function and regularization, and the optimal parameters of Gauss kernel were determined through experiments. Then, the SVDD method was compared with other spectral quality determination methods such as Mahalanobis distance and local anomaly factor. The average of correct recognition rate of normal samples and correct rejection rate of abnormal samples were used as evaluation criteria. The experimental results show that the spectral quality determination method based on support vector data description has the best performance. In near infrared qualitative analysis, this method can be used as a means of eliminating abnormal spectra before feature extraction and pattern classification, and the spectra quality determination step based on SVDD can effectively improve the reliability of the qualitative analysis.

Keywords Near infrared spectroscopy; Qualitative analysis; Quality determination; Support vector machines data description

(Received Jan. 22, 2019; accepted May 4, 2019)