DOI: 10.11895/j.issn.0253-3820.140868

近红外光谱与化学计量学方法用于镉污染稻米的定性鉴别

朱向荣[1,2] 李高阳[1,2] 黄绿红[1] 苏东林[1,2] 刘 伟[1,2] 单 杨*[2]
[1](湖南省农业科学院 湖南省食品测试分析中心,长沙 410125)
[2](中南大学研究生院隆平分院,长沙 410125)

摘 要 采用近红外光谱漫反射模式结合化学计量学方法对稻米镉含量是否超标进行可行性鉴别分析。本研究收集了 120 个样本,测定其镉含量值(合格 49 个,不合格 71 个) 。对光谱数据预处理方法优化,确定了平滑,一阶导数以及自归一化后的数据作为输入变量。采用竞争性自适应重加权算法筛选了 45 个关键变量,并对上述变量的光谱吸收带进行归属。比较了主成分分析-判别分析法、偏最小二乘识别分析、线性判别分析、*K*-最近邻法与簇类独立软模式法 5 种模式识别方法。确定采用偏最小二乘识别分析建模效果最好,模型训练集与预测集鉴别准确率分别达到 98.8% 与 91.7% 。结果表明, 近红外光谱作为初筛方法可用于鉴别稻米中镉含量是否超标。

关键词 近红外光谱; 化学计量学; 稻米; 镉; 污染; 定性鉴别

1 引 言

水稻是我国主要的粮食作物,65% 的中国人以稻米为主食,稻米品质的优劣直接关系到人体健康水平。镉(Cd) 是自然界中广泛存在的重金属元素,具有较强的毒性。水稻被认为是镉吸收最强的大宗谷类作物,镉容易被水稻吸收并积累[1] 。稻田镉污染通过通过土壤-植物-人体的食物链途径传递,严重威胁到食用者健康[2] 。石墨炉原子吸收光谱法(GF-AAS) 与电感耦合等离子体质谱(ICP-MS) 等分析方法被广泛用于农产品中镉含量的测定。这些方法虽然灵敏度高、准确性好,但也存在着需要专业人员操作、繁杂的样品前处理以及消耗大量的强酸试剂等缺点。

近红外(Near infrared,NIR) 光谱具有快速、便捷、无损等优点,在稻米各种营养指标检测中均有应用。虽然无机元素在 NIR 光谱区并没有吸收,但是有机物质通过螯合或络合的方式与无机元素形成螯合物或络合物,从而这些物质在近红外光谱区有相应的响应与吸收[3,4] 。研究者采用 NIR 光谱方法对动物样品[5~8] 、植物样品[9~11] 以及水样[12~14] 中的重金属元素进行定性与定量分析,但对镉污染稻米的研究鲜有报道。由于稻米待检的 Cd 含量低,特征变量较难识别,这为 NIR 光谱分类带来困难。本研究采用化学计量学方法结合模式识别方法构建最优分类器,建立稻米镉含量是否超标的定性模型,实现类间差异微小的 NIR 光谱有效分类。并对光谱预处理方法与光谱变量筛选做了系统考察,取得了满意效果。

2 实验部分

2.1 仪器与试剂

AA-6800 石墨炉原子吸收光谱仪(日本岛津公司) 。Nicolet Antaris Ⅱ傅里叶变换近红外光谱仪(美国 Thermo 公司) ,配有积分球漫反射采样系统,InGaAs 检测器,Omnic7.3 光谱采集软件,TQ Analyst v6.2.1 分析软件,采用 Matlab 7.1 软件(Mathwork Inc.) 进行数据处理。实验用镉标准贮备液(1000 mg/L) 与大米粉中镉成分分析标准物质(GBW08511) (0.504 mg/L) 均由国家标准物质中心提供。其它试剂均为分析纯。

2.2 实验方法

2.2.1 样本收集 本实验所用的 120 个稻米样本均采自湖南省长沙县北山镇农田。

2014-09-24 收稿; 2015-01-29 接受
本文系科技部“十二五”国家科技支撑计划(No. 2012BAK17B17) ,农业部农业科研杰出人才培养计划,湖南省科技重大专项(No. 2011FJ1002-4) 资助项目
* E-mail: shanyang_jgs@163.com

2.2.2 原子吸收光谱法 采用《GB/T 5009/15-2003 食品中镉的测定》方法进行测定。准确称量0.5 g样品于250 mL锥形瓶中，加入20 mL HNO_3，采用沙浴加热进行消化，加热至锥形瓶中的溶液变澄清停止加热，冷却后，用去离子水溶解并定容至25 mL，待测。每批均采用含镉稻米标准物质进行质控，并以空白样品(仅试剂)消除背景。

2.2.3 近红外光谱方法 NIR光谱采用漫反射检测系统，扫描波数10000～4000 cm^{-1}，扫描次数32次，分辨率8 cm^{-1}，增益为2。内置背景为参照。每批样品3次平行实验，取其平均光谱，以消除样品不均匀性带来的干扰。

2.2.4 化学计量学方法 采用基于马氏距离的Kennard-Stone(KS)[15]法，选择有代表性校正集的样本。归一化主要采用均值中心化(Mean centering，MC)、自归一化(Autoscaling)与Pareto归一化3种方法。采用竞争性自适应权重取样法(Competitive adaptive reweighted sampling，CARS)对变量进行筛选。采用主成分分析识别分析(Partial component analysis-discriminant analysis，PCA-DA)、偏最小二乘识别分析(Partial square least discriminant analysis，PLS-DA)、线性判别分析(Linear discriminant analysis，LDA)、*K*-最近邻法(*K*-Nearest Neighbor，KNN)与簇类独立软模式法(Soft independent modeling class analog，SIMCA)进行定性建模。

3 结果与讨论

3.1 镉含量测定

将上述已经优化好的样品预处理方法以及原子吸收光谱条件用于测定120个稻米样本重金属镉含量。根据《GB2762-2012 食品中污染物限量》规定的稻米镉限量为0.2 mg/kg，判定所有样本的真实属性，49个为合格稻米，71个为镉超标稻米。

3.2 校正集的选择和光谱预处理优化

对NIR光谱数据进行建模，选择有代表性的训练集，不但可以减少建模的工作量，而且可提高模型的适用性和准确性。通过KS法从收集的120个样本中依次挑选出84个作为训练集，余下的36个作为测试集。

图1为样本的近红外光谱图，记录了10000～4000 cm^{-1}波数样本的NIR光谱曲线，光谱1和2分别表示Cd含量合格与超标稻米样本，二者的NIR光谱无明显差异，用肉眼很难辨别，必须采用化学计量学方法进行数据预处理和建立模型。

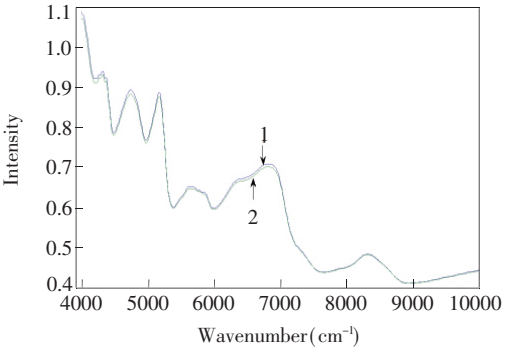

图1 代表性样本的近红外光谱

Fig. 1 NIR spectra of the representative samples

1. 合格样本; 2. 镉超标样本。

1. Qualified Cd; 2. Excessive Cd.

采集的NIR全光谱中，原始光谱数据中存在着随机噪声与基线漂移现象，测定环境背景、光程变化和光散射等因素也对建模的准确性产生影响。因此，本研究以训练集准确率(Accuracy of training set，ATRS)与预测集准确率(Accuracy of testing set，ATES)为指标，考察了Savitzky-Golay平滑(SG smoothing)、一阶导数(1^{st} derivative，1. der)、二阶导数(2^{nd} dervative，2. der)、MC、Autoscaling、Pareto以及上述光谱方法组合等9种数据预处理方法的分类效果，如表1所示。通过优化，采用平滑、一阶导数与均值中心化对光谱进行处理。

3.3 光谱变量选择

提高模型的预测能力是NIR光谱分析中的研究热点和难点问题。通过有效的变量筛选可以剔除无信息或者不相关变量，提高模型的预测能力与泛化能力[16]。竞争性自适应权重取样(Competitive adaptive reweighted sampling，CARS)是近年提出的一种新型变量选择方法[17]。此算法模拟达尔文进化论中“适者生存”原则，将每个变量看成一个个体，对变量实行逐步淘汰的选择过程。同时，引入了指数衰减函数来控制变量的保留数，具有较高的计算效率，适合于高维数据的变量选择。

表 1　不同的光谱预处理方法对模型鉴别准确率影响

Table 1　Effect of different pretreating methods on the accuracy rate of the models

序号 No.	光谱预处理方法 Spectral pretreatment methods	训练集准确率 ATRS (%)	预测集准确率 ATES (%)
1	Smoothing + autoscaling	76.2	91.7
2	Smoothing + MC	79.8	91.7
3	Smoothing + pareto	76.2	83.7
4	Smoothing + 1. der + autoscaling	81.0	81.7
5	**Smoothing +1. der + MC**	**100**	**86.2**
6	Smoothing + 1. der + pareto	97.6	83.4
7	Smoothing + 2. der + autoscaling	85.8	91.7
8	Smoothing + 2. der + mean centering	78.6	83.4
9	Smoothing +2. der + pareto	81.0	91.7

ATRS: accuracy of training set; ATES: accuracy of testing set; MC: mean centering; der: derivative.

CARS 法筛选过程为: 采用 10 折交互检验后,找到 RMSECV 最低点,在 1 ~ 28 次间残差呈递减趋势,表明筛选过程剔除了与样本性质无关的变量,28 次后开始递增,则可能剔除关键变量,从而导致残差增大。经过 CARS 方法筛选,从 1557 个变量筛选了 45 个特征光谱变量。上述变量 NIR 吸收谱带解析为[18]:4312.1,4339.1,4346.8,4350.6,4431.6,4439.3,4516.5,4597.5 与 4643.7 cm^{-1}为脂肪族烃中 C—H 键第一组合频; 7405.3 与 7455.5 cm^{-1}为 C—H 键第二组合频; 9090.8,9121.6,9129.4,9341.5,9549.8 与 9734.9 cm^{-1}为 C—H 键第三组合频; 5565.6,5770.0,5789.3,6082.4,6140.2,6221.2与6232.8 cm^{-1}为 C—H 键一级倍频; 8327.1,8415.8,8736.0,8990.5 与 9090.8 cm^{-1}为 C—H 键二级倍频; 6082.4,6140.2,6221.2 与 6232.8 cm^{-1}为 = C—H 键的一级倍频区间。7031.2 cm^{-1}为醇羟基 O—H 伸缩振动的一级倍频; 5160.6 cm^{-1}为 O—H 伸缩和弯曲振动的组合频; 6734.2 cm^{-1}为 N—H 键伸缩振动的一级倍频; 5160.6 cm^{-1}为胺基 N—H 键伸缩振动与弯曲振动的组合频; 6734.2 和 6938.6 cm^{-1}分别为 N—H 键的对称伸缩一级倍频与反对称伸缩振动; 9823 cm^{-1}为 N—H 键的对称伸缩振动的二级倍频。

3.4　模型的建立与预测

采用PCA-DA法建立定性分类模型,结果如表2所示。36个预测集样本中,33个样本的属性预测

表 2　样本测试集的预测结果

Table 2　Prediction result of testing set

样品号 Sample No.	含量 Content (mg/kg)	真实属性 True nature	预测值* Predictive value	样品号 Sample No.	含量 Content (mg/kg)	真实属性 True nature	预测值* Predictive value
1	0.301	1	1	19	0.404	1	1
2	0.411	1	1	20	0.420	1	1
3	0.280	1	1	21	0.390	1	1
4	0.221	1	1	22	0.563	1	1
5	0.559	1	1	23	0.069	–1	–1
6	0.647	1	1	24	0.105	–1	–1
7	0.463	1	1	25	0.101	–1	–1
8	0.267	1	1	26	0.123	–1	–1
9	0.520	1	1	27	0.058	–1	–1
10	0.428	1	1	28	0.218	1	–1
11	0.685	1	1	29	0.131	–1	–1
12	0.213	1	1	30	0.111	–1	–1
13	0.391	1	1	31	0.103	–1	–1
14	0.097	–1	–1	32	0.243	1	–1
15	0.356	1	1	33	0.076	–1	1
16	0.339	1	1	34	0.231	1	1
17	0.374	1	–1	35	1.082	1	1
18	0.284	1	1	36	0.555	1	1

* 以“1”: 表示镉含量超标样本(excessive content sample); “–1”: 表示合格样本(qualified sample)。

值与其真实属性相符，预测正确。17#、28#与 36#共 3 个样本的属性预测值与实际值不符，预测错误。预测集准确率达到 91.7%。其中，28#样本的 Cd 含量为 0.218 mg/kg，接近限量阈值。33#样本镉含量为 0.076 mg/kg，含量太低，超出检出限。

图 2 为 120 个稻米样本分类图，横坐标代表样本号，纵坐标代表标准变量得分。纵坐标小于 0 的区域内的小三角形表示镉含量合格稻米，纵坐标大于 0 的区域内的小方框代表镉含量超标的稻米。训练集的 84 个稻米样本中，83 个样本预测正确，准确率为 98.8%；测试集的 36 个稻米样本中，33 个样本预测正确，准确率为 91.7%，总体正确识别率能够达到 95.2%，分类效果良好。

3.5 模型方法的比较

采用 PLS-DA、LDA、KNN 与 SIMCA 方法建模，并与 PCA-DA 对比，结果如表 3 所示。PCA-DA 的鉴别结果最好，LDA 与 SIMCA 方法预测结果较差。原因在于，在 SIMCA 模式识别中，模型的建立是利用 LDA 的方法，SIMCA 方法在建立模型时，没有考虑其它的类，因此，在每个类的模型中，有些因素在获取类中明显的变化时只能反映出有限的鉴别信息[19]。稻米样本本身包含的分类信息不够丰富以及镉超标稻米与合格稻米的性质差异太少。因此，当多维数据两类中的子空间都非常接近时，由于类之间不必要的重叠，从而存在产生非优化鉴别模型的危险。而 PCA-DA 与 PLS-DA 方法分别是基于主成分回归(PCR)与 PLS 回归的判别分析方法[20]，在构建因素时考虑到了辅助矩阵以代码形式提供的类成员信息，因此具有高效的鉴别能力。因此，PCA-DA 与 PLS-DA 这两种方法的收敛能力与全局寻优能力比其余 3 种方法更强。

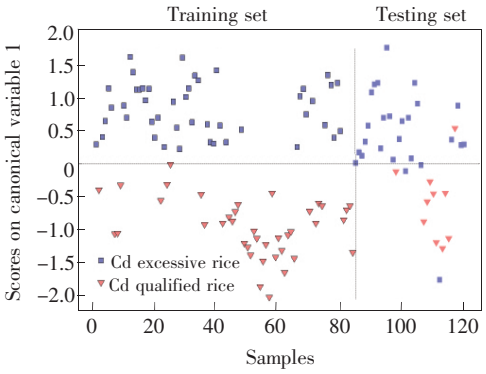

图 2 采用 PCA-DA 方法得到的镉超标与合格两类样本分类图

Fig. 2 Classifications of Cd excessive and qualified rice sample by principal component anlaysis-discriminant analysis (PCA-DA) method

表 3 5 种建模方法的比较

Table 3 Comparison of five modeling methods

方法 Methods	训练集准确率 ATRS (%)	预测集准确率 ATES (%)	总准确率 Total accuracy (%)
PCA-DA	98.8	91.1	95.2
PLS-DA	100	86.1	93.0
LDA	100	72.2	86.1
KNN	94.0	86.1	90.0
SIMCA	100	66.6	83.3

LDA: linear discriminant analysis; KNN: K-nearest neighbor; SIMCA: soft independent modeling class/analog.

4 结 论

采用近红外漫反射光谱结合化学计量学方法，初步实现了镉污染稻米是否超标的定性鉴别。本研究将进一步扩大样本数量，优化 NIR 光谱数据预处理方法，提高模型的稳健性与准确性。本研究结果将为近红外光谱技术在稻米镉快速识别上提供初步依据，有利于保障稻米质量安全。

致 谢 感谢中南大学中药现代化研究中心提供的 CARS 软件。

References

1 Chaney R L, Reeves P G, Ryan J A, Simmons R W, Welch R M, Angle J S. *Biometals*, **2004**, 17(5): 549 - 553

2 Lindén A, Olsson I M, Bensryd I, Lundh T, Skerfving S, Oskarsson A. *Ecotox. Environ. Safe.*, **2003**, 55(2): 213 - 222

3 Chen G P, Mei Y, Tao W, Zhang C, Tang H R, Iqbal J, Du Y P. *Anal. Chim. Acta*, **2011**, 670(1-2): 39 - 43

4 Kumagai M, Ohisa N, Amano T, Ogawa N. *Anal. Sci.*, **2003**, 19: 1553 - 1555

5 González-Martín I, González-Pérez C, Hernández-Méndez J, Villaescusa-García V. *Anal. Chim. Acta*, **2002**, 468(2): 293 - 301

6 Morón A, Cozzolino D. *The J. Agric. Sci.*, **2002**, 139(4): 413 - 423

7 Font R, Del Río-Celestino M, Vélez D, de Haro-Bailón A, Montoro R. *Anal. Chem.*, **2004**, 76(14): 3893 - 3898

8 González-Martín I, Alvarez-García N, González-Pérez C, Villaescusa-García V. *Anal. Chim. Acta*, **2008**, 75(2): 351 - 355
9 Font R, Del Río-Celestino M, Vélez D, Montoro R, De Haro A. *Sci. Total. Environ.*, **2004**, 327(1-3): 93 - 104
10 Font R, Vélez D, Del Río-Celestino M, De Haro-Bailón A, Montoro R. *Microchim. Acta*, **2005**, 151(3-4): 231 - 239
11 Moros J, Llorca I, Cervera M L, Pastor A, Garrigues S, De La Guardia M. *Anal. Chim. Acta*, **2008**, 613(2): 196 - 206
12 Huang Z X, Tao W, Fang J J, Wei X M, Du Y P. *Chemometr. Intell. Lab.*, **2009**, 98(2): 195 - 200
13 Ning Y, Li J H, Cai W S, Shao X G. *Spectrochim. Acta A*, **2012**, 96: 289 - 294
14 Liu F X, Cai W S, Shao X G. *Vib. Spectrosc.*, **2013**, 68: 104 - 108
15 Kennard R W, Stone L A. *Technometrics*, **1969**, 11(1): 137 - 148
16 ZHU Xiang-Rong, LI Na, SHI Xin-Yuan, QIAO Yan-Jiang, ZHANG Zhuo-Yong. *Chinese J. Anal. Chem.*, **2008**, 36(6): 770 - 774
朱向荣, 李 娜, 史新元, 乔延江, 张卓勇. 分析化学, 2008, 36(6): 770 - 774
17 Li H D, Liang Y Z, Xu Q S, Cao D S. *Anal. Chim. Acta*, **2009**, 648(1): 77 - 84
18 LU Wan-Zhen. *Modern Near Infrared Spectroscopy Analytical Technology* [Second Edition], China Petrochemical Press, **2006**: 29 - 31
陆婉珍. 现代近红外光谱分析技术(第二版), 中国石化出版社, 2006: 29 - 31
19 HAO Yong, SUN Xu-Dong, GAO Rong-Jie, PAN Yuan-Yuan, LIU Yan-De. *Trans. Chin. Soc. Agric. Eng.*, **2012**, 26(12): 373 - 376
郝 勇, 孙旭东, 高荣杰, 潘媛媛, 刘燕德. 农业工程学报, 2012, 26(12): 373 - 376
20 YANG Zhong, REN Hai-Qing, JIANG Ze-Hui. *Spectroscopy Spectral Analysis*, **2008**, 28(4): 793 - 796
杨 忠, 任海青, 江泽慧. 光谱学与光谱分析, 2008, 28(4): 793 - 796

Near Infrared Spectroscopy Combining with Chemometrics for Qualitative Identification of Cadmium-Polluted Rice

ZHU Xiang-Rong[1,2], LI Gao-Yang[1,2], HUANG LÜ-Hong[1], SU Dong-Lin[1,2], LIU Wei[1,2], SHAN Yang*[2]
[1](*Hunan Food Test and Analysis Centre, Hunan Academy of Agricultural Science, Changsha* 410125, *China*)
[2](*Longping College, Graduate School, Central South University, Changsha* 410125, *China*)

Abstract Near-infrared (NIR) diffuse reflectance spectroscopy and chemometrics method were used to discriminate cadmium-polluted rice. The samples set contained 120 spectra of qualified (n = 49) and excessive (n = 71) was collected and scanned. After optimization, a combination (smoothing coupled with first derivative and mean centering) was utilized as a spectral pretreatment method. Competitive adaptive reweighed sampling (CARS) was adapted to selected 45 key variables, and each band of the variables was assigned. Five modeling methods including partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), K-nearest neighbor (KNN), soft independent modeling class analog (SIMCA) and principal component analysis-discriminant analysis (PCA-DA) were used and compared. PCA-DA was finally selected as the optimal qualitative model. The accuracy rate of training set and testing set for PCA-DA method was 98.8% and 91.7%, respectively. The results showed that NIR spectroscopy could be used as a rapid, non-destructive and convenient analytical method for primary screening and detecting cadmium-polluted rice.

Keywords Near infrared spectroscopy; Chemometrics; Rice; Cadmium-polluted; Qualitative identification
(Received 24 September 2014; accepted 29 January 2015)
This work was supported by the National Twelfth Five-year Science and Technology Support Project (No. 2012BAK17B17), the Agricultural Scientific and Research Outstanding Talents Cultivation Plan of the Ministry of Agriculture, Hunan Provincial Key Scientific and Technological Special Project (No. 2011FJ1002-4)