DOI:10.13652/j.issn.1003-5788.2016.01.010

基于拉曼光谱的大米快速分类判别方法

Method for rapid discrimination of varieties rice by using Raman spectroscopy

孙 娟　张 晖　王 立　钱海峰　齐希光

SUN Juan　ZHANG Hui　WANG Li　QIAN Hai-feng　QI Xi-guang

(江南大学食品学院,江苏 无锡　214122)

(*School of Food Science and Technology, Jiangnan University, Wuxi, Jiangsu* 214122, *China*)

摘要:以拉曼光谱技术为手段,结合化学计量学方法,对来自黑龙江、江苏、湖南 3 个产地共 123 份大米样品的光谱数据进行采集,并对得到的拉曼图谱进行主成分分析(PCA)和偏最小二乘判别分析(PLSDA),建立大米快速分类判别方法。应用主成分分析对不同种类、产地和品种的大米进行粗分类鉴别;选择不同种类、品种和产地的稻米样本建立相应的偏最小二乘判别分析模型,其中 2/3 的样本作为建模训练集,1/3 的样本作为建模校正集,按照种类、产地、品种建立的模型其训练集样本正确判别率均为 100%,校正集样本正确判别率分别为 100%,100%,94.12%。因此,研究所建立的拉曼光谱技术结合化学计量学方法可以快速、有效地鉴别大米种类、品种及产地。

关键词:大米;拉曼光谱;分类;主成分分析;偏最小二乘判别分析

Abstract: The aim of this research was to establish the fast detection method for distinguishing the rice samples from different geographical origins and different breeds in China by using Raman micro-spectroscopy. A total of 123 rice samples from Heilongjiang, Jiangsu and Hunan province were analyzed by Raman spectra, and the data were statistic analyzed by chemometrics. The data was subjected to principal component analysis (PCA) and partial least squares discriminant analysis (PLSDA) to distinguish differences among samples from different geographical origins and different species. PCA could classify the rice samples preliminarily and classification model developed by PLSDA was used to classify and predict different rice samples. The rice samples (2/3) were as a training set of modeling, and the rest of samples were as a prediction set of modeling. The correct classification rates in the training set according to the varieties of rice samples were 100%, and those prediction set were 100%, 100% and 94.12%, respectively. The results in this research indicated it is a quickly efficacious method to identify rice from different geographical origins and species by Raman spectroscopy with chemometrics.

Keywords: rice; Raman spectroscopy; classification; PCA; PLSDA

大米是世界上最重要的谷物粮食作物之一,不仅是中国人的传统主食,更是世界一半以上人口的主食[1]。因品种、产地、生长条件的不同,大米的营养成分含量存在很大的差别。近年来,由于人们对大米的营养价值和口感品质的追求不断提高,市场上出现了非优质大米冒充优质大米、以次充好、品牌冒充、产地冒充等现象,严重损害了消费者利益。中国大米种植区域广范、品种繁多,因此市场监管困难。传统的鉴别方法主要包括感官检测和化学检测,主观性强且费时费力,不能满足市场监管中快速判别的需求。

拉曼光谱(Raman spectrum)作为一种物质结构鉴定分析测试手段,具有灵敏度高、检测速度快的优点,不仅能够满足无损、痕量检测的需求,而且可以适应不同的工作环境。在生物化学、法医学、制药、食品等领域得到了广泛应用[2-7]。目前拉曼光谱在谷物领域的应用报道较少,主要应用在成分结构分析[8]、安全质量控制[9-10]、谷物产地检测[11]等定性检测方面,以及成分含量的定量检测[12]等少数几个方面。作为一种快速、无损、简便的分析检测手段,拉曼光谱技术为大米种类、品种、产地的快速判别以及市场的有序监控提供了可能。

本试验拟以拉曼光谱技术为主要手段,结合化学计量学方法中的主成分分析,对大米进行聚类,并应用偏最小二乘判别分析建立大米种类、品种和产地的分类判别模型,旨在实现粳米、籼米及品种和产地的快速分类判别。

1　材料与方法

1.1　原料与仪器

1.1.1　原料

采集 3 个产地 10 个品种共 123 份大米样品,其中包含粳米 70 份,籼米 53 份。具体的大米样品信息见表 1。将样

基金项目:国家"十二五"科技支撑计划(编号:2012BAD37B08)

作者简介:孙娟(1991—),女,江南大学在读硕士研究生。
E-mail:1159925343@qq.com

通讯作者:张晖

收稿日期:2015-09-12

表 1 大米样品品种与产地信息

Table 1 Information of samples of rice varieties and geographical origins

种类	产地	品种	样本数
籼米	湖南	华两优 285	13
籼米	湖南	淮两优 608	13
籼米	湖南	金建软粘 1 号	14
籼米	湖南	玉针香	13
粳米	江苏	金穗 999	13
粳米	江苏	原稻 3 号	13
粳米	黑龙江	垦稻 17	11
粳米	黑龙江	垦稻 19	11
粳米	黑龙江	垦稻 23	11
粳米	黑龙江	龙粳 25	11

品用去离子水洗净后自然晾干,不需进一步处理直接进行拉曼光谱检测。

1.1.2 仪器

显微共焦激光拉曼光谱仪:LabRAM HR Evolution 型,法国 HORIBA Jobin Yvon S. A. S. 公司。

1.2 试验方法

1.2.1 拉曼光谱采集方法 使用 LabRAM HR Evolution 型显微共聚焦拉曼光谱仪,扫描范围为 200～1 600 cm^{-1},扫描时间为 30 s,扫描 2 次,激发波长为 632.8 nm,样品采集时直接将样品置于 50 倍镜头下进行检测。测试条件为室温,相对湿度< 60%。每个样品重复扫描 3 次并计算平均光谱代表样品信息,以消除样品不均匀性带来的干扰。

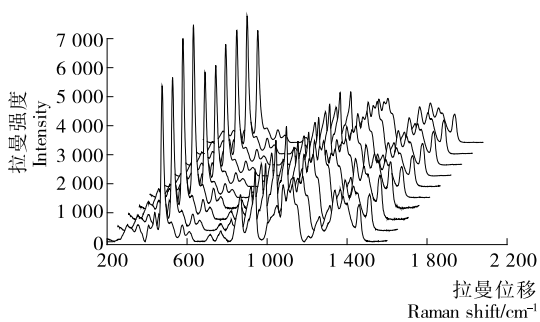
1.2.2 光谱处理与数据分析 拉曼光谱检测会遇到荧光背景的干扰,同时由于仪器本身系统稳定性的限制,会产生背景噪声并出现基线漂移现象,对分析结果会产生很大影响,需要对数据进行预处理以减少上述影响。在全光谱范围内,考察了基线校正(Baseline)、平滑(Smoothing)、中心化(Mean Center)、归一化(Normalize)及支持向量机(SNV)等 5 种数据预处理方法对分类结果的影响,最终确定了基线校正、平滑作为光谱预处理的方法。因此,所有拉曼图谱在进行数据分析前都需进行基线校正及平滑处理,以消除基线漂移和噪声的影响。同时,引入化学计量学方法对大米种类和产地进行定性分类判别,并建立判别模型对未知的大米样品进行预测。利用 The Unscrambler 9.7(CAMO, USA)软件对大米样品数据进行主成分分析(PCA),并对采集得到的样本随机选取 1/3 做校正集,2/3 的样本做训练集建立偏最小二乘判别分析(PLSDA)模型。利用 OriginPro 8.0 软件对数据分析结果作图。

1.2.3 偏最小二乘判别分析 偏最小二乘判别分析是用一组已知类别的样本作为训练集,即用已知的样本进行训练,建立识别模型,对未知样本进行分类和预测。偏最小二乘判别分析是一种稳健的判别分析方法,适合于变量数多且存在多重共线性的情况[13]。该方法首先将样本类别作为反应量处理,即对不同类别的样品进行数值变量标定,然后,运用偏最小二乘回归建立解释变量与反应量之间的关系模型。最后,通过比较模型的反应变量预测值大小,来确定各样本的类别[14]。在应用偏最小二乘判别法进行大米的分类鉴别时,对不同种类、品种和产地的大米进行赋值作为类别标志,预测类别值与给定类别值之间的差异小于 0.50 表示分类正确。

2 结果与讨论

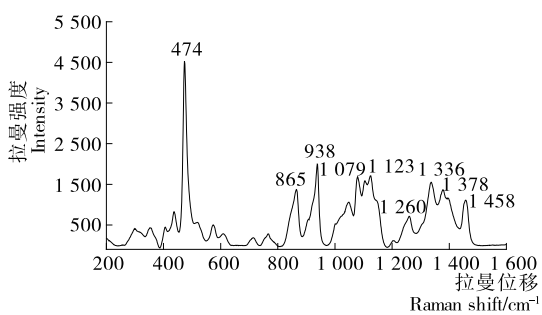
2.1 拉曼光谱分析

图 1(a)为大米样本经基线校正后的拉曼光谱图,(b)为经过数据预处理后的大米在 200～1 600 cm^{-1}的大米拉曼光谱信息图。由图 1(a)可知,不同大米样品的拉曼光谱在出峰位置上没有明显的区别,且峰形相似,粳米、籼米的拉曼光谱间的差异直观上难以识别;江苏、黑龙江不同产地的大米样品以及同一产地不同品种的大米样品间的拉曼光谱差异甚微。在图 1(b)中,大米最基本的拉曼吸收峰在 475,865,938,1 079,1 123,1 260,1 336,1 378,1 458 cm^{-1}处。结合陈健、Hoonsoo 等[15－16]的研究可以得知,475 cm^{-1}的强吸收峰为淀粉的主链特征峰,938,1 079,1 123 cm^{-1}处为淀粉骨架的指纹图谱;865,1 260 cm^{-1}处为 CH_2中的 C—H 摇摆振动,1 336 cm^{-1}处是 CH_2中的 C—H 平面形变振动,1 458 cm^{-1}处为 C—H 形变振动。但总体上,不同品种和产地的大米的拉曼光谱很相近,难以从肉眼上识别区分,因此,需引入化学计量学方法对大米分类作进一步的研究。



图中曲线从上至下依次为华两优285、淮两优608、金建软粘1号、玉针香、金穗999、原稻3号、垦稻17、垦稻19、垦稻23、龙粳25

(a) 拉曼光谱图



(b) 信息图

图 1 大米样品的拉曼光谱图和大米信息图(经预处理)

Figure 1 Rice samples Raman spectra and rice information (baseline correction)

2.2　粳米、籼米的快速分类鉴别

图 2 是对所有大米的拉曼光谱进行主成分分析,以前 2 个主成分为坐标建立的样本 PCA 得分图。由于前两个主成分对光谱矩阵的累积方差贡献达 94.73%(图 3),因此,样本在得分图中的分布可大体反应其特征。由图 2 可知,在 123 份大米样品中,70 份粳米样品聚集在图中左侧部分(PC1＜0),而籼米样品聚集在图中右侧部分(PC1＞0),可以实现粳米与籼米之间初步的种类分类。而在粳米、籼米各类别内样本的分布又存在明显的聚类趋势,这可能是由于样本的品种、产地不同导致样本按照相似程度再次聚类。而根据 X-加载图(图 3)中表征的各波段对模型前 10 个主成分的贡献大小,得出大米粳米、籼米种类分类的特征波段主要为 450～500,830～940,1 066～1 306 cm^{-1},根据淀粉、蛋白质和脂肪的吸收特征,可分析出 3 个特征波段主要反映了不同大米营养成分的差异[17]。而对粳米、籼米分类结果贡献最大的波段为 450～500 cm^{-1},这反映了淀粉在粳米、籼米两种大米中的差异显著,对分类起主要的影响作用。总体而言,主成分分析结果表明应用拉曼光谱技术可以将粳米和籼米分为两类,方便、快速、有效。

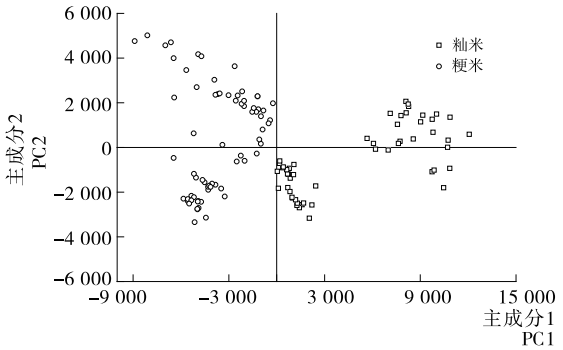
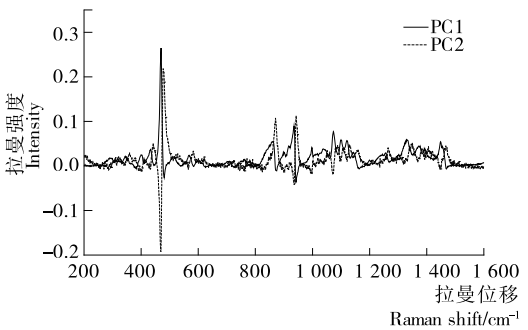
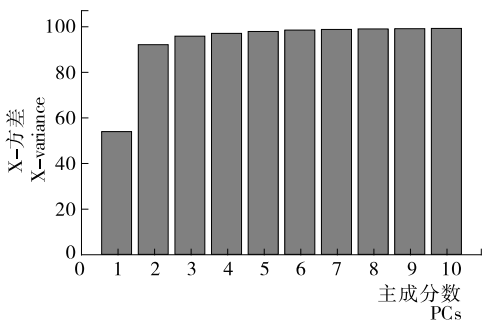


图 2　粳米、籼米的 PCA 得分图

Figure 2　Principal component analysis score plots (PC1×PC2) of japonica and indica rice samples



(a) 前2个主成分的X-加载图



(b) 前10个主成分的累计方差图

图 3　主成分分析 X-加载图和累计方差图

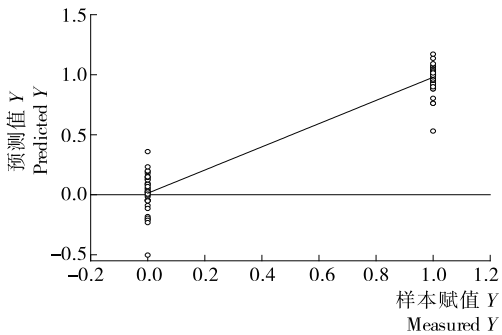
Figure 3　Principal component analysis X-loading plot and cumulative variance plot

将选定的 82 份训练集样本和 41 份校正集样本导入 The Unscrambler 9.7 软件进行偏最小二乘判别分析。其中,将籼米赋值为 0,粳米赋值为 1,计算数值大于 0.50 视为粳米,小于 0.50 视为籼米,模型结果见图 4。得到的偏最小二乘判别分析模型的相关系数为 0.95,预测均方根误差(RMSEP)为 0.15,模型的正确判别率为 100%;而未知样品的预测值误差均在±0.30 内,样本全部预测正确,校正集样本判别正确率为 100%,由此可见该模型的预测精度较高,对于粳米、籼米具有很好的判别预测能力。

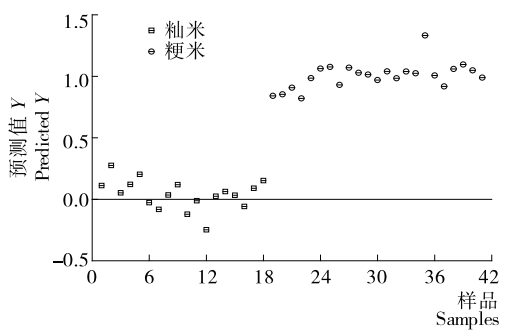
2.3　不同产地大米的快速分类鉴别

大米的品质和口感不仅受其自身遗传因素的影响,还受地理环境和气候条件的影响,因此不同品种和产地的大米的品质差异显著。选择来自黑龙江、江苏 2 个省份 6 个品种共 70 份粳米为样本进行大米产地快速分类判别的研究。由于两省份的大米同属于粳米,在外观上十分相似,难以从肉眼上区别。

图 5 是不同产地的粳米拉曼图谱的 PCA 得分图。其中,江苏省的大米样品集中在图中左侧(PC1＜−1 000),而黑



(a) 偏最小二乘判别分析模型



(b) 偏最小二乘判别分析模型预测结果

图 4　大米粳米、籼米偏最小二乘判别分析结果

Figure 4　Results of partial least squares discriminant analysis of indica and japonica rice samples

龙江省的样品除有两个样本被错误分类入江苏省外，其余均分布在图中右侧(PC>－1 000)，说明应用主成分分析可以将不同产地大米进行快速分类且效果较好。由图 5 可知，虽

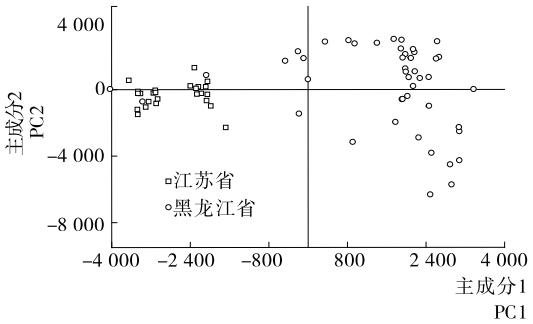
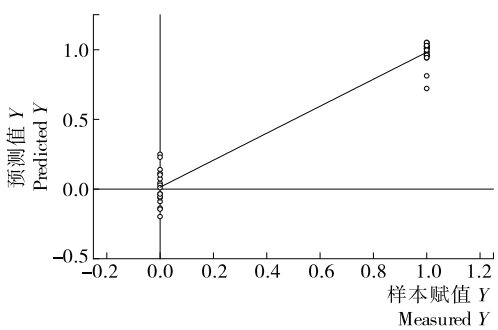


图 5 不同产地的粳米 PCA 得分图

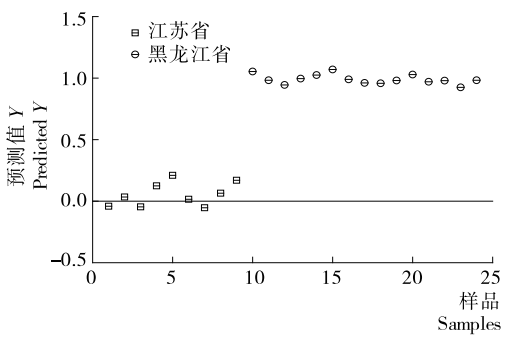
Figure 5 Principal component analysis score plots (PC1×PC2) of rice samples from different geographical origins

然同属于粳米，但黑龙江省的 42 个样本聚为一类，江苏省的 26 个样本聚为另一类，说明不同产地的大米存在差异，大米产地对其分类有很重要的影响。产地不同，土壤肥沃程度不同，所产大米营养成分和含量会存在差异，这种差异会反映在拉曼光谱上，从而直接影响化学计量学分析结果，从而可以将大米按照其产地进行分类。

图 6 为不同产地的粳米偏最小二乘判别分析结果，其中江苏省的大米样本赋值为 0，黑龙江省的大米赋值为 1。建立的不同产地大米的偏最小二乘判别分析模型的相关系数为 0.98，预测均方根误差为 0.094，各参数训练模型的预测精度较高，样本全部判断正确，训练集样本的判别正确率为 100%。用未知产地信息的大米样本对该偏最小二乘判别分析模型进行判别和验证，9 份江苏大米、15 份黑龙江大米全部预测正确，即用该模型对校正集样品进行判别分析的正确判别率为 100%，说明建立的大米产地偏最小二乘分析模型的预测精度较高，对大米的产地具有很好的判别预测能力。



(a) 偏最小二乘判别分析模型



(b) 偏最小二乘判别分析预测结果

图 6 不同产地大米的偏最小二乘判别分析结果

Figure 6 Results of partial least squares discriminant analysis of rice samples from different geographical origins

2.4 不同品种大米的快速分类鉴别

为进一步研究同种类、同产地但品种不同大米之间的分类效果，以湖南省的金建软粘 1 号、华两优 285、淮两优 608、玉针香共 4 个品种 53 份大米样品为例，对大米品种间的分类判别进行探究。首先利用 PCA 对大米品种进行分类，结果见图 7。

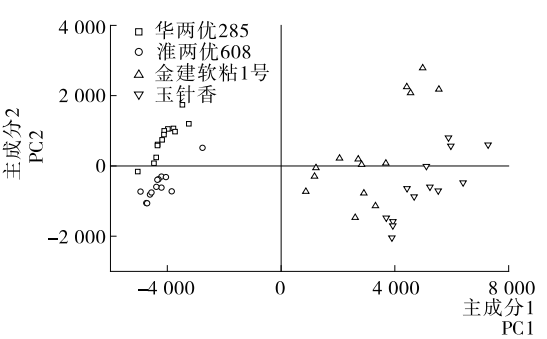


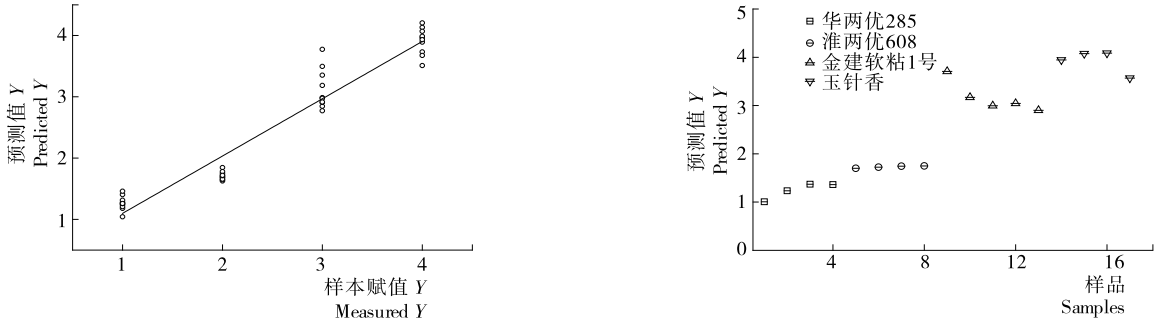
图 7 不同品种的大米 PCA 得分图

Figure 7 Principal component analysis score plots (PC1×PC2) of different rice breeds samples

由图 7 可知，4 个不同品种的大米在得分图中有初步的聚类趋势，但品种间距离较近，不能完全分开，其中淮两优 608、华两优 285 分布在得分图中的左侧 (PC1<0)；金建软粘 1 号分布在右侧 (PC1>0)，而玉针香分布在右下侧 (PC1>0)。导致该结果的原因可能是不同品种的大米遗传基因不同，但由于此次检测的样本产地均为湖南省，样本之间的差异较小。图 8 为利用偏最小二乘判别分析建立的不同品种间大米的分类判别模型，将华两优 285 赋值为 1、淮两优 608 赋值为 2、金建软粘 1 号赋值为 3、玉针香赋值为 4，将计算值与真实值进行比较从而可以得知未知样品的品种归类。建立的不同品种大米的偏最小二乘判别分析模型的相关系数为 0.94，预测均方根误差为 0.29，训练集模型样本品种判别正确率为 100%。利用该模型预测未知品种的大米样本，得到的结果中属于金建软粘 1 号的 1 个样本误判为玉针香品种，其余品种样本均预测判别正确，校正集样本识别正确率为 94.12%，模型正确率较高，可以用于大米品种的分类判别及未知品种样品的预测。

3 结论

本试验采用拉曼光谱技术结合化学计量学方法中的主成



（a）偏最小二乘判别分析模型　　（b）偏最小二乘判别分析预测结果

图 8　不同品种大米的偏最小二乘判别分析结果

Figure 8　Results of partial least squares discriminant analysis of rice samples from different breeds

分分析和偏最小二乘判别分析,对粳米和籼米的种类以及不同品种和产地的大米进行快速分类判别,并建立了相应的判别模型。其中应用主成分分析法可以区别粳米、籼米,建立的偏最小二乘判别分析模型的相关系数为 0.95,预测均方根误差为 0.15,训练集及校正集样本判别正确率均为 100%。黑龙江、江苏两个产地的大米在主成分分析得分图中各自按产地聚类,聚类趋势明显,建立的大米产地偏最小二乘判别分析模型的相关系数为 0.98,预测均方根误差为 0.094,模型预测效果很好,训练集和校正集样本的正确判别率均为 100%。对来自湖南省的金建软粘 1 号、华两优 285、淮两优 608、玉针香 4 个品种按照品种的不同进行主成分分析,样本按照品种在得分图中有相应的聚类趋势,但不能完全分开;偏最小二乘判别分析模型的相关系数为 0.94,预测均方根误差为 0.29,训练集样本正确判别率为 100%,校正集样本正确判别率为 94.12%;因此,可以认为采用拉曼光谱结合主成分分析和偏最小二乘判别分析方法可以准确地对大米样品按照种类、品种和产地进行简便、快速、有效地分类鉴别。

参考文献

[1] 夏立娅，申世刚，刘峥颢，等. 基于近红外光谱和模式识别技术鉴别大米产地的研究[J]. 光谱学与光谱分析，2013，33(1)：102-105.

[2] Alan G Ryder，Gerard M O'Connor，Thomas J Glynn. Quantitative analysis of cocaine in solid mixtures using Raman Spectroscopy and chemometric methods[J]. Journal of Raman Spectroscopy，2000，31(3)：221-227.

[3] De Beer T R M，Baeyens W R G，Vermeire A，et al. Raman spectroscopic method for the determination of medroxyprogesterone acetate in a pharmaceutical suspension：validation of quantifying abilities，uncertainty assessment and comparison with the high performance liquid chromatography reference method[J]. Analytica Chimica Acta，2007，589(2)：192-199.

[4] Sylwester Mazurek，Roman Szostak. Quantification of atorvastatin calcium in tablets by FT-Raman spectroscopy[J]. Journal of Pharmaceutical and Biomedical Analysis，2009，49(1)：168-172.

[5] Sergio Armenta，Salvador Garrigues，Miguel De La Guardia，et al. Sweeteners determination in table top formulations using FT-Raman spectrometry and chemometric analysis[J]. Analytica Chimica Acta，2004，521(2)：149-155.

[6] Niculina Peica. Identification and characterization of the E951 artificial food sweetener by vibrational spectroscopy and theoretical modeling[J]. Journal of Raman Spectroscopy，2009，40(12)：2 144-2 154.

[7] 刘燕德，施宇，蔡丽君，等. 拉曼光谱在重金属分析中的研究进展[J]. 食品与机械，2012，28(4)：1-4.

[8] 黄卫宁，张君慧，Hoseney R C. FT-Raman 光谱与谷物化学研究的最新进展[J]. 食品科学，2004，25(11)：411-414.

[9] Feng Xin-wei，Zhang Qing-hua，Cong Pei-sheng，et al. Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis[J]. Talanta，2013(115)：548-555.

[10] 欧阳思怡，叶冰，刘燕德，等. 表面增强拉曼光谱法在农药残留检测中的研究进展[J]. 食品与机械，2013，29(1)：243-246.

[11] Hwang J，Kang S，Lee K. Enhanced Raman spectroscopic discrimination of the geographical origins of rice samples via transmission spectral collection through packed grains[J]. Talanta，2012(101)：488-494.

[12] 宋瑜，孙晓荣，刘翠玲，等. 拉曼光谱和近红外光谱在小麦粉品质定量分析中的应用[J]. 食品科学技术学报，2014，32(2)：24-27.

[13] Lidia E A，David D E，Susan D，et al. Feasibility of near infrared spectroscopy for analyzing corn kernel damage and viability of soybean and corn kernels[J]. Journal of Cereal Science，2012，55(2)：160-165.

[14] 黄亚伟，张令，王若兰. 新陈玉米的拉曼光谱快速判别研究[J]. 现代食品科技，2014，30(12)：149-152.

[15] 陈健，肖凯军，林福兰，等. 拉曼光谱在食品分析中的应用[J]. 食品科学，2007，28(12)：554-558.

[16] Hoonsoo L，Cho B K，Kim M S，et al. Prediction of crude protein and oil content of soybeans using Raman spectroscopy[J]. Sensors and Actuators B：Chemical，2013，185(8)：694-700.

[17] 成明华，关东胜，张慧敏，等. 10 种稻米的品质分析[J]. 粮油食品科技，2001，9(6)：13-16.