DOI:10.13652/j.issn.1003-5788.2020.09.017

基于近红外光谱技术的小米产地溯源研究

Geographic origin determination of millet based on near infrared spectroscopy technique

李 楠[1] 杨春杰[2]

LI Nan[1] *YANG Chun-jie*[2]

(1. 运城学院生命科学系,山西 运城 044000;2. 运城学院机电工程系,山西 运城 044000)

(1. *Department of Life Science, Yuncheng University, Yuncheng, Shanxi* 044000, *China*;
2. *Department of Mechanical and Electrical Engineering, Yuncheng, Shanxi* 044000, *China*)

摘要:采用便携式近红外光谱仪结合主成分分析(PCA)、费舍尔线性判别(FLDA)及多层感知器神经网络(MLP-NN)模型,探讨近红外光谱技术应用于小米产地溯源的可行性。PCA 分析显示,除山西、河南、黑龙江 3 省的样品差异较小难以区分外,其余 8 个省份的样品均能清晰区分产地。FLDA 和 MLP-NN 分析均能识别出小米样品产地,但 MLP-NN 识别效果优于 FLDA,两个模型对预测集的识别正确率分别为 92.3%,84.6%。以上结果表明,近红外光谱技术可有效应用于小米的产地溯源。

关键词:小米;产地溯源;近红外;主成分分析;线性判别;神经网络

Abstract: In order to provide a scientific method for the geographic origin determination of millet, a portable near-infrared (NIR) spectrometer was used in combination with other methods, including principal component analysis (PCA), linear discriminant analysis (LDA) and multi-layer perceptron neural network (MLP-NN). The results showed that, for PCA model, samples from different places except Shanxi, Henan and Heilongjiang provinces clustered into different groups significantly. Millet samples from different origins could be effectively discriminated by FLDA and MLP-NN. The MLP-NN model was better than FLDA model in recognition rate. The recognition accuracy of the two models for the predication set were 92.3% and 84.6% respectively. Therefore, the NIR spectroscopy technique can be used for the geographic origin determination of the millet.

Keywords: millet; geographic origin; near-infrared spectrometer; principal component analysis; linear discriminant analysis; neural network

小米又称粟(米),禾本科狗尾草属[1]。在中国,作为五谷之一的小米有着悠久的食用历史,数千年来一直作为传统主食养育了中国北方文明,在现代仍是主要杂粮之一[2]。小米的产地来源与其品质密切相关,地域特色小米具有反映该区域自然环境的特有品质,中国已出现许多名优原产地域小米产品群落。假冒产地不仅损害消费者和企业利益,同时也增加了食品安全问题追溯与风险管理难度[3]。农产品产地溯源方法主要包括特定化学成分(如矿物元素、同位素、多酚、糖、氨基酸等)差异分析的破坏性溯源方法和光谱、仿生(电子鼻、电子舌)、介电特性、核磁共振检测等无损溯源方法[4]。其中,近红外光谱技术因其高效、无损、环保等优点已成为近几十年来发展最迅速的无损检测技术之一,也被认为是当前经济性最高的食品溯源技术[4-8]。近红外光谱技术已被应用于茶叶[9-10]、橄榄油[11]、肉类[12-14]、酒类[15-17]等食品的产地溯源研究,显现出较好的应用前景。目前,采用近红外、高光谱、拉曼光谱等光谱技术的小米产地溯源研究较少,同时这些研究多存在样本数量较少[18-20]、样本产地来源偏少[18-22]、模型预测准确率偏低[18]等问题。此外,相关研究多使用体积大、价格高的精密型近红外光谱仪,在实际应用方面存在一定局限。因此,研究拟以来源于 11 个主产省份的 150 份小米样品为研究对象,采用便携式近红外光谱仪检测样品,同时为了提高模型准确率和泛化能力,采用鲁棒主成分分析(rPCA)识别异常数据,并采用 DUPLEX 方法划分样本集,进而比较主成分分析、线性判别、人工神经网络等模式识别方法的分类结果,为利用近红外光谱技术实现小米产地溯源提供参考。

1 材料与方法

1.1 材料

小米样品:采集于中国 11 个省份,涵盖所有国内小米

基金项目:山西省重点学科建设经费资助(编号:FSKSC);山西省"1331"工程重点学科项目(编号:098-091704);运城学院院级科研项目(编号:XK-2018010)

作者简介:李楠,女,运城学院讲师,硕士。

通信作者:杨春杰(1985—),男,运城学院讲师,硕士。
E-mail: 592846672@qq.com

收稿日期:2020-03-15

主产地(见表 1)。小米采集后铝箔袋真空密封,4 ℃保存。

1.2 仪器与设备

便携式近红外谷物分析仪:ZX-50IQ 型,美国 Zeltex 公司;

真空封口机:RS-BZ11A 型,合肥荣事达电子电器集团有限公司;

电子天平:FA1604 型,上海舜宇恒平科学仪器公司。

1.3 方法

1.3.1 光谱采集　样品预先放置于 25 ℃环境自然升温至室温。近红外分析仪开机预热 30 min 后校准。每次测量时,称取 50.0 g 样品,均匀置于样品杯,然后放于样品室关闭遮光罩进行测量。每个样品测量 3 次,取平均值作为最终分析光谱值。

1.3.2 数据分析　采用鲁棒主成分分析(rPCA)方法剔除样品光谱数据异常值后,使用 DUPLEX 方法将数据按 3∶1 划分为训练集和预测集,最后对数据进行主成分分析(PCA)、费舍尔线性判别分析(LDA)及多层感知器神经网络(MLP-NN)建模识别分析。建模分析采用 SPSS20 软件;异常值检测、训练集及预测集划分采用 Matlab R2015b 软件。

2 结果与分析

2.1 样品近红外光谱

由图 1 可知,不同样本的光谱形状具有相似性,但吸收峰的位置均存在一定的差异性,说明不同产地小米的组成成分存在差异,这些差异通过近红外光对含氢基团(C—H、N—H、O—H、S—H 等)振动的倍频和组合频吸收不同而生成差异化的红外光谱图。为保证光谱数据具有代表性,每个样品测量 3 次,取平均值作为最终分析光谱值。由于试验使用的便携式近红外分析仪只有 14 个近红外波长,且小米样品在不同波长下的吸光值均有差异,因此将全部波长数据用于后续分析。

表 1 小米样品产地及数量

Table 1 Origin and quantity of millet samples

序号	产地	样本数	序号	产地	样本数
1	陕西	14	7	吉林	20
2	山西	20	8	黑龙江	15
3	山东	15	9	河南	15
4	宁夏	10	10	河北	10
5	内蒙古	10	11	甘肃	11
6	辽宁	10	合计		150

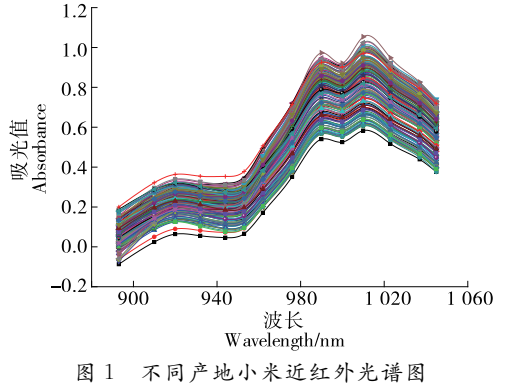

图 1 不同产地小米近红外光谱图

Figure 1 Near infrared spectra of millet samples from different origins

2.2 异常值检测

异常值会影响模型的可靠性,甚至会导致模型严重失真,因此在模型建立前需要识别并剔除异常值。鲁棒主成分分析(rPCA)被用于识别异常值,该方法能够高效识别出异常值[23]。首先计算每个样品的主成分得分距离(SD)和正交距离(OD),然后将样品分为 4 组:常规组(低 SD 低 OD)、良好主成分转换组(高 SD 低 OD)、正交异常值组(低 SD 高 OD)、不良主成分转换组(高 SD 高 OD),后两组样品不利于模型分析,被识别为异常值。如图 2 所示,产自陕西的 14 个样品中,2、7、9、13 号共 4 个样品异常值被检出。同样的方法识别其余样品组异常值,最终得到 131 组数据用于进一步分析。

2.3 主成分分析(PCA)

主成分分析是一种无监督的分析方法,在尽可能保证原有信息的前提下将多元数据降维转化为少数新变量,减少数据冗余,进而方便理解和展示原有变量差异。依据前两个主成分得到的 PCA 得分图,可以直观地表现原始数据所代表的样本状态,样品点的聚集、离散程度反映出样品间的差异大小。如图 3 所示,第一主成分的方差贡献率为 95.48%,第二主成分的方差贡献率为4.22%,合计为 99.70%,因此前两个主成分可以充分反映原始数据信息。山西、河南、黑龙江 3 省内的样品点分布较为分散,其余省内样品点分布相对集中,说明来自于山西、河南、黑龙江 3 省样品的省内差异较大,其余省份样品的省内差异较小。同时,山西、河南、黑龙江 3 省样品点与其余 8 个省份样品点部分重叠,但 8 个省份样品点分布相对独立且界限清晰,说明除了山西、河南、黑龙江3省份

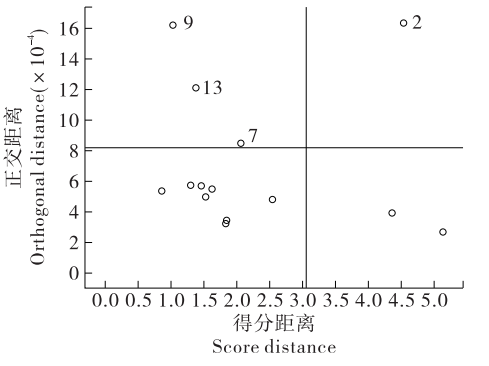

图 2 小米样品鲁棒主成分分析检测异常值

Figure 2 The outlier diagnosis obtained by rPCA for millet samples

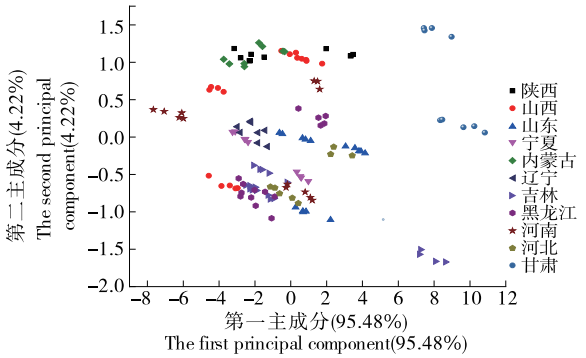

图 3　不同产地小米样品主成分分析二维图

Figure 3　PCA plot of millet samples with PC1 and PC2

样品与部分省份样品差异较小之外,其余省份样品省间差异明显。上述结果说明在主成分分析中,除山西、河南、黑龙江 3 省的样品省内差异较大导致难以区分省间差异外,其余省份的样品均能依据省间差异区分省份产地。

2.4　模型建立

2.4.1　训练集及预测集划分　双向数据分组(DUPLEX)方法是一种计算机训练集识别方法,该方法能保证训练集中样本按照空间距离均匀分布,保证训练集样本的代表性[24]。该方法的选取过程:① 选择样本组中欧式距离最大的两个样本划入训练集;② 在余下的样本组中,选择欧式距离最大的两个样本划入预测集[25]。重复上述操作,直到满足预测集所需的样本数,余下的样本全部划入训练集。使用该方法最终由 92 个样品组成训练集,39 个样品组成预测集,具体结果见表 2。

2.4.2　费舍尔线性判别分析(FLDA)　FLDA 是一种有监督的线性分类方法,将高维模式样本投影到最佳鉴别矢量空间,降维的同时保证样本有最大的类间距离和最小的类内距离,使得各类样品能够更好的区分。在 SPSS 软件中,将训练集作为 FLDA 的变量输入,产地信息作为判别输出,利用 Fisher 函数、wilks'lambda 变量选择,采用步进判别法进行分析,结果见表3。训练集中对不同样品

表 2　利用 DUPLEX 方法的分组结果

Table 2　The grouped results based on DUPLEX

产地	训练集样本数	预测集样本数	产地	训练集样本数	预测集样本数
陕西	8	2	吉林	16	3
山西	11	6	黑龙江	10	5
山东	8	7	河南	9	4
宁夏	8	1	河北	6	2
内蒙古	5	3	甘肃	6	3
辽宁	5	3			

表 3　训练集和测试集的费舍尔线性判别分析结果

Table 3　The training and prediction results obtained by FLDA

样本		预测										正确率/%
		陕西	山西	山东	宁夏	内蒙	辽宁	吉林	黑龙江	河南	河北	甘肃	
训练集	陕西	8											100.0
	山西		11										100.0
	山东			8									100.0
	宁夏				8								100.0
	内蒙					5							100.0
	辽宁						5						100.0
	吉林							16					100.0
	黑龙江								10				100.0
	河南									9			100.0
	河北										6		100.0
	甘肃											6	100.0
测试集	陕西	2											100.0
	山西		6										100.0
	山东		1	5					1				71.4
	宁夏				1								100.0
	内蒙	1	1			0				1			0.0
	辽宁						2		1				66.7
	吉林							3					100.0
	黑龙江								5				100.0
	河南									4			100.0
	河北										2		100.0
	甘肃											3	100.0

产地溯源的平均正确率为 100.0%，预测集中对不同样品产地溯源的平均正确率为 84.6%，其中来源于内蒙古的 3 个样品产地预测全部错误。以上结果表明建立的 FLDA 模型在训练集上表现良好，但对测试集数据表现一般，模型的泛化能力较差，有可能是训练集样本数量不足或特征波长选择不合适导致了模型的过拟合。

2.4.3　多层感知器神经网络分析(MLP-NN)　MLP-NN 是一种前馈式有监督神经网络，由一个输入层、一个输出层以及一个或多个隐藏层组成。作为神经网络方法中最有影响的方法之一，MLP-NN 具有从训练数据中学习复杂非线性映射的能力，能够发现数据间复杂的关系。利用训练集数据构建 MLP-NN 模型，隐藏层和输出层的激活函数分别为双曲正切和 Softmax，隐藏层层数为 1，单位数为 50，优化算法为调整的共轭梯度。结果见表 4，训练集中对样品产地溯源的平均正确为 95.7%，预测集中对样品产地溯源的平均正确率为 92.3%。以上结果表明建立的 MLP-NN 模型具有较高的准确度和可靠性，因此，相较于建立的 FLDA 判别模型，基于 MLP-NN 判别模型的近红外光谱技术可有效应用于小米的产地溯源。

3　结论

以产地相对全面的小米样品为研究对象，采用便携式近红外光谱仪检测样品，建立了基于近红外光谱技术的小米产地多层感知器神经网络、费舍尔线性判别模型。结果显示：多层感知器神经网络模型优于费舍尔线性判别模型，费舍尔线性判别模型准确度高，但泛化能力一般(测试集正确率为 84.6%)；多层感知器神经网络模型具有较高的准确度和可靠性(测试集正确率为 92.3%)。因此，基于多层感知器神经网络模型的近红外光谱技术可有效应用于小米的产地溯源。

研究中检测近红外波长以及算法模型都较少，后续研究应该扩展近红外波长范围(780～2 500 nm)，优选新的数据算法(数据预处理、特征波长选择、建模方法等)，进而深入揭示小米近红外光谱数据、产地以及组成成分之间的关系。另外，小米的品质除了受地域环境(如气候、土壤等)影响外，还与基因(品种)、种植、管理和加工等因素相关，这些因素均能影响产地溯源的准确性。实际应用时需要考虑并克服这些因素，因此实际应用建模

表 4　训练集和测试集的多层感知器神经网络分析结果

Table 4　The training and prediction results obtained by MLP-NN

样本		预测											正确率/%
		陕西	山西	山东	宁夏	内蒙	辽宁	吉林	黑龙江	河南	河北	甘肃	
训练集	陕西	7	1										87.5
	山西		10			1							90.9
	山东			6							2		75.0
	宁夏				8								100.0
	内蒙					5							100.0
	辽宁						5						100.0
	吉林							16					100.0
	黑龙江								10				100.0
	河南									9			100.0
	河北										6		100.0
	甘肃											6	100.0
测试集	陕西	2											100.0
	山西		6										100.0
	山东			7									100.0
	宁夏				1								100.0
	内蒙	1				2							66.7
	辽宁						3						100.0
	吉林				1			2					66.7
	黑龙江			1					4				80.0
	河南									4			100.0
	河北										2		100.0
	甘肃											3	100.0

样本的数量及来源会远远超过研究中的样本,甚至需要建立规模庞大的样本数据库并持续完善以降低模型的预测风险。

参考文献

[1] 刘三才,朱志华,李为喜,等. 谷子品种资源微量元素硒和蛋白质含量的测定与评价[J]. 中国农业科学,2009,42(11):3 812-3 818.

[2] 李顺国,刘斐,刘猛,等. 我国谷子产业现状、发展趋势及对策建议[J]. 农业现代化研究,2014,35(5):531-535.

[3] 管骁,古方青,杨永健. 近红外光谱技术在食品产地溯源中的应用进展[J]. 生物加工过程,2014,12(2):77-82.

[4] 魏益民,郭波莉,魏帅,等. 食品产地溯源及确证技术研究和应用方法探析[J]. 中国农业科学,2012,45(24):5 073-5 081.

[5] 张勇,王督,李雪,等. 基于近红外光谱技术的农产品产地溯源研究进展[J]. 食品安全质量检测学报,2018,9(23):6 161-6 166.

[6] 马奕颜,郭波莉,魏益民,等. 植物源性食品原产地溯源技术研究进展[J]. 食品科学,2013,35(5):246-250.

[7] 魏益民,郭波莉,赵海燕,等. 论食品溯源技术研究方法与应用原则[J]. 中国食品学报,2012,12(11):8-13.

[8] 马冬红,王锡昌,刘利平,等. 近红外光谱技术在食品产地溯源中的研究进展[J]. 光谱学与光谱分析,2011,31(4):877-881.

[9] 周健,成浩,曾建明,等. 基于近红外的多相偏最小二乘模型组合分析实现茶叶原料品种鉴定与溯源的研究[J]. 光谱学与光谱分析,2010,30(10):2 650-2 653.

[10] 任广鑫. 基于近红外分析技术的红茶成分分析与产地识别的研究[D]. 合肥:安徽农业大学,2012:18-21.

[11] SINELLI N, CERRETANI L, EGIDIO V D, et al. Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity[J]. Food Research International, 2010, 43(1): 369-375.

[12] 孙潇,史岩. 近红外光谱技术对加工后鸡肉产地溯源的研究[J]. 现代食品科技,2015(6):322-328.

[13] SUN Shu-min, GUO Bo-li, WEI Yi-min, et al. Classification of geographical origins and prediction of $\delta^{13}C$ and $\delta^{15}N$ values of lamb meat by near infrared reflectance spectroscopy[J]. Food Chemistry, 2012, 135(2): 508-522.

[14] XICCATO G, TROCINO A, TULLI F, et al. Prediction of chemical composition and origin identification of European sea bass (*Dicentrarchus labrax* L.) by near infrared reflectance spectroscopy (NIRS)[J]. Food Chemistry, 2004, 86(2): 275-281.

[15] LIU Liang, COZZOLINO D, CYNKAR W U, et al. Preliminary study on the application of visible-near infrared spectroscopy and chemometrics to classify Riesling wines from different countries[J]. Food Chemistry, 2008, 106(2): 781-786.

[16] EGIDIO V D, OLIVERI P, WOODCOCK T, et al. Confirmation of brand identity in foods by near infrared transflectance spectroscopy using classification and class-modelling chemometric techniques: The example of a Belgian beer[J]. Food research international, 2011, 44(2): 544-549.

[17] LIU Liang, COZZOLINO D, CYNKAR W U, et al. Geographic classification of Spanish and Australian tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis[J]. Journal of Agricultural and Food Chemistry, 2006, 54(18): 6 754-6 759.

[18] 吉海彦,任占奇,饶震红. 基于高光谱成像技术的不同产地小米判别分析[J]. 光谱学与光谱分析,2019,39(7):2 271-2 277.

[19] 张石定,卢全伟,王涛,等. 拉曼光谱用于小米品种和产地的鉴别与化学成分关联研究[C]// 第二十届全国光散射学术会议(CNCLS 20)论文摘要集. 苏州:中国物理学会光散射专业委员,2019:264.

[20] 李佳洁,吴建虎,张海波. 利用可见/近红外光谱技术对小米产地进行溯源研究[J]. 食品安全质量检测学报,2017,8(8):209-215.

[21] 钱丽丽,宋雪健,类彦波,等. 近红外漫反射光谱技术对小米产地的快速检测[J]. 食品工业,2018,39(6):257-261.

[22] 宋雪健,钱丽丽,周义,等. 近红外漫反射光谱技术对小米产地溯源的研究[J]. 食品研究与开发,2017,38(11):143-148.

[23] XU Lu, SHI Peng-tao, YE Zi-hong, et al. Rapid analysis of adulterations in Chinese lotus root powder (LRP) by near-infrared (NIR) spectroscopy coupled with chemometric class modeling techniques[J]. Food Chemistry, 2013, 141(3): 2 434-2 439.

[24] SNEE R D. Validation of regression models: Methods and examples[J]. Technometrics, 1977, 19(4): 415-428.

[25] XU Lu, YAN Si-min, CAI Chen-bo, et al. Untargeted detection and quantitative analysis of poplar balata (PB) in Chinese propolis by FT-NIR spectroscopy and chemometrics[J]. Food Chemistry, 2013, 141(4): 4 132-4 137.