食品工业科技
Science and Technology of Food Industry
ISSN 1002-0306,CN 11-1759/TS

《食品工业科技》网络首发论文

题目：基于主成分分析和人工神经网络的近红外光谱大豆产地识别
作者：田琼，马新华，袁俊杰，龙阳，洪武兴，卢韵宇
DOI：10.13386/j.issn1002-0306.2020060271
网络首发日期：2020-10-26
引用格式：田琼，马新华，袁俊杰，龙阳，洪武兴，卢韵宇．基于主成分分析和人工神经网络的近红外光谱大豆产地识别．食品工业科技.
https://doi.org/10.13386/j.issn1002-0306.2020060271

网络首发时间：2020-10-26 15:28:00
网络首发地址：https://kns.cnki.net/kcms/detail/11.1759.TS.20201026.1442.010.html

作者简介：田琼（1978.7），男，高级工程师，研究方向：商品产地溯源，E-mail：13542047202@163.com。

基金项目：海关总署科研项目（2019HK052）。

基于主成分分析和人工神经网络的近红外光谱大豆产地识别

田琼，马新华，袁俊杰，龙阳，洪武兴，卢韵宇

（湛江海关技术中心，广东湛江 524022）

摘　要：为了准确、快速地识别大豆产地，通过近红外光谱技术(NIRS)结合主成分分析（PCA）和人工神经网络技术(ANN)研究不同国家大豆内含特征，建立进口大豆产地识别模型。采用箱型图校正法，剔除阿根廷、巴西、乌拉圭、美国等 4 个国家 166 组大豆样本中 12 组异常样本。采用多元散射校正（MSC）、标准正态变量（SNV）、Savitzky-Golay(SG)平滑滤波等方法进行光谱数据预处理，结果表明,采用 SG（3）平滑结合 MSC 预处理效果最好。主成分分析表明，前 10 个主成分的累积贡献率达到 99.966%。选取主成分分析得到前 10 个主成分为输入向量，4 个产地作为目标向量，分别采用支持向量机（SVM），邻近算法（KNN）与人工神经网络法（ANN）建立识别模型。结果表明，采用 BP-ANN 建模效果最好，总体测试集准确率为 95.65%，其中阿根廷准确率为 100%，巴西准确率为 100%，乌拉圭准确率为 80%，美国准确率为 100%，该模型能够现实对进口大豆生产国别的识别。

关键词：近红外光谱，主成分分析，人工神经网络，大豆产地识别

Study on Method of Soybean Origin Identification Based on Near-Infrared Spectrum of Principal Component Analysis and Artificial Neural Network Model

TIAN Qiong, MA Xin-hua, YUAN Jun-jie, LONG Yang, HONG Wu-xing, LU Yun-yu

(Technology Center of Zhanjiang Customs District, Zhanjiang, 524022, China)

Abstract: The study demonstrated that near infrared spectrum based on artificial neural can be used as an accurate and rapid technique for identification of origin of soybean. Near infrared Spectrum（NIRS) in combination with principal component analysis(PCA) and artificial neural network(ANN) was used to identify the geographical origin of soybean. The near-infrared reflectance spectrums of the 166 soybeans from Argentina, Brazil, Uruguay and the United States had been collected, then the twelve outliers of NIRS were eliminate by using a box-plot graph. The original spectral data was processed by means of multiplicative scatter correction(MSC),standard normal variate(SNV), Savitzky-Golay(SG),etc. It was gotten the optimal result by the preprocessing method based on the smoothing treatment in SG (3 points) with MSC. The PCA was used to compress the NIRS, the analysis results showed that the cumulative variance contribution of PC1 to PC10 (the first ten principal components) were

99. 966%. The first ten principal components of the samples were applied as inputs and the origin of soybeans were applied as the outputs. Support vector machine(SVM），K-nearest neighbor(KNN) and ANN were respectively developed for pattern recognition. The ability of identify of ANN model is the best of SVM model and KNN model. The overall discrimination accuracy for the test set was 95.65% by ANN model , and the discrimination accuracy for soybean samples from Argentina ,Brazil, Uruguay and the United States were 100% , 100%,80% and 100%, respectively. The ANN model can identify the origin of soybeans imported from different countries
Key words: Near infrared spectrum (NIRS); Principal component analysis (PCA); Artificial neural network (ANN); Soybean origin identification
中图分类号：TS202.1　　文献标志码：　A
doi:10.13386/j.issn1002-0306. 2020060271

大豆，又称黄豆、黄大豆，为一年生草本植物原产于中国，在新石器时代已有栽培。大豆是一种重要的粮油兼用农产品，其植物蛋白质含量为 35%~45%，富含多种人体必备的氨基酸[1]。同时，大豆油富含亚油酸，有降低血清胆固醇含量、预防心血管疾病的功效[2]。我国是世界上最大的大豆进口国，据海关统计,2019 年我国大豆进口总量 8551.1 万吨,近年来我国进口大豆总量和对外依存度一直维持高位，各大豆来源国的大豆质量参差不齐，其安全问题越来越受到关注[3-6]。因此,研究快速、有效的进口大豆产地识别方法,对加强原产地管理和进口大豆质量安全监管具有重要意义。

近红外光谱(Near Infrared Spectrum）技术具有操作简单、快速、非破坏、无污染等特点，被广泛地应用食品、农产品的品质分析[7-10]和产地识别[11-14]。赵海燕等[15]应用近红外光谱仪对中国小麦主产区河北省、河南省、山东省和陕西省共 240 份小麦籽粒样品,采用偏最小二乘判别分析法建模型，总体准确率达到 80%以上。李勇等[16]利用近红外分析仪对江苏、辽宁、湖北、黑龙江 4 个省份的 169 个大米样品进行了检测，利用蒙特卡罗模拟方法(Monte Carlo method)判别 4 个省份的大米产地，预测准确率达 90%以上。

在大豆产地识别研究方面，研究人员主要基于大豆的有机组分[17,18]和矿物质含量[19,20]，对国内不同产区大豆，进行产地鉴别。鹿保鑫等[21] 测定了黑龙江嫩江及北安共 168 份大豆中的矿物元素含量及蛋白质、脂肪、可溶性总糖和灰分含量。采用步进式方法筛选出 10 种特征指标，建立判别模型对训练集大豆产地整体准确率达 95%以上。结果表明 7 种矿物元素（Mn、As、Sr、La、Nd、Tb、Hf）和 3 种有机成分（蛋白质、脂肪、可溶性总）是用于大豆产地判别的主要特征指标,携带了充分的产地判别信息。目前，对于进口大豆产地识别相关报道较少，主要有国内的张勇等[22]采用气相色谱-质谱法（gas chromatography mass spectrometry）测定 48 份进口大豆（美国 15 份、巴西 15 份、加拿大 10 份、阿根廷 4 份、乌拉圭 4 份）和 48 份国产大豆的脂肪酸组成，采取随机森林方法建立了进口大豆与国产大豆间的判别模型，交互检验预测准确率达 95%以上,可有效区分进口大豆与国产大豆；日本的 Akiko Otaka 等[23]采用能量色散 X 射线荧光光谱仪（Energy Dispersive X-Ray Fluorescence Spectrometer）测定 46 个大豆样本（日本 23 个，美国 8 个，中国 7 个，加拿大 7 个，美国加拿大混合样 1 个）的 8 个元素（Mg，P，Cl，K，Mn，Cu，Br 和 Ba），采取主成分分析，能够地区分日本和非日本的大豆样本。 这两篇报道，研究内容仅局限于本国大豆与国外大豆两者之间的识别，由于收集的不同国别大豆样本量有限，均未进一步对不同国别大豆进行产地识别。

本研究基于主成分分析和人工神经网络，采用近红外光谱技术建立了进口大豆的产地识别模型，收集 166 组 2017 — 2019 年间，进口大豆的近红外光谱数据,经数据预处理后，采用主成分分析、人工神经网络等算法,建立进口大豆（阿根廷、巴西、乌拉圭、美国）产地

识别模型，为加强进口大豆质量安全管理及海关原产地管理，提供有效技术支持。

1 材料与方法

1.1 材料与仪器

试验样品：采集 166 组大豆样本，分别来自湛江、黄埔、张家港、泉州、深圳、南沙、汕头、阳江等口岸在 2017—2019 年入境的留存样品，去除杂质、破碎粒后，得到阿根廷 14 组、巴西 90 组、乌拉圭 26 组、美国 36 组，合计 166 组大豆试验样本。

Infraxact 近红外分析仪（检测器：硅 570 ~ 1100 nm，铟镓砷 1100 ~ 1850 nm） 瑞典 Foss 公司

1.2 红外光谱测定

参考 GB/T 24870-2010[24]测定大豆样品的近红外光谱。在室温下，将约 200 g 的整粒大豆样品，采用自然装样方式，扫描波长范围 570~1848 nm，光谱采样间隔（波段宽）2 nm，波数据采集频率 3 s/次，扫描 20 次取平均值。

1.3 光谱数据预处理

采用箱型图校正法，剔除异常样本。采用多元散射校正（multiplicative scatter correction，MSC）、标准正态变量（standard normal variate，SNV）、Savitzky-Golay(SG)平滑滤波及它们之间相互组合，进行光谱数据预处理。

1.4 识别模型及评价

首先主成分分析法进行数据降维，在此基础上，采用常见三种模式识别算法，支持向量机（support vector machine, SVM），邻近算法（K-nearest neighbor,KNN）与人工神经网络法（artificial neural network，ANN）分别建立识别模型。随机选取 70%的样品为训练集，30%的样品为测试集（ANN 分别选取 15%的样品为验证集、测试集）。识别模型的效果评价，以测试集准确率为依据。

准确率表示模型预测结果的准确程度，准确率用公式（1）表示：

准确率（%）=（正确预测样品数量/总样品数量）×100 式（1）

1.5 分析软件

使用 MATLAB R2020a 进行箱型图校正、MSC、SNV、SG 平滑滤波、主成分分析，采用分类学习、神经网络模式识别工具箱建立产地识别模型。使用 Office Excel 2016 绘制图形。

2 结果与分析

2.1 大豆原始光谱分析

近红外光谱是有机分子中与氢相连化学键振动的合频和倍频吸收，如图 1 所示：在 920 nm 处的吸收峰与脂肪族烃中 C-H_2 振动有关；在 1200 nm 处的吸收峰与水分和脂肪中的 C-H 键、O- H 键及 C=O 键的振动有关；在 1464 nm 处的吸收峰与蛋白质中的 N-H 键伸缩振动一级倍频吸收有关，1760 nm 处的吸收峰与脂肪中的 C-H 键振动有关，1788 nm 处的吸收峰与水分中的 O-H 键振动倍频吸收有关。结果表明，由于不同国别大豆的脂肪、蛋白质、水分含量不同，造成近红外光谱存在一定差异，但走势大体一致，需要对光谱数据做进一步处理。

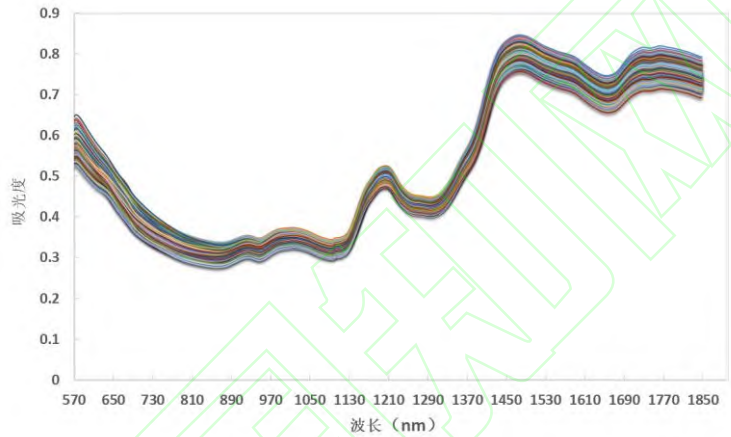

图 1 大豆近红外原始光谱图

Fig.1 The original NIR spectrum of soybeans

2.2 光谱数据预处理

对 166 组大豆样本近红外原始谱图，做箱型图分析。如图 2 所示，十字图形数据为离群值（范围±1.5 倍四分位距以外的值），被视为异常值。删除了光谱数据异常的 12 组样本，其中阿根廷 2 组，巴西 5 组，乌拉圭 5 组，剔除完成后，得到最终使用的 154 组样本近红外光谱数据集，其中阿根廷 12 组、巴西 85 组、乌拉圭 21 组、美国 36 组。

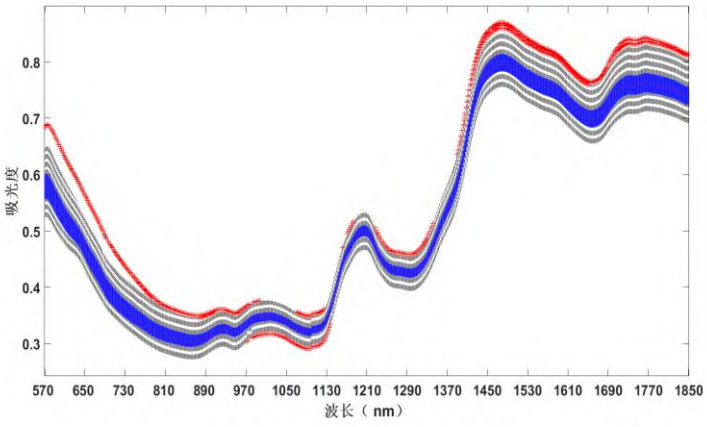

图 2 近红外原始光谱数据的箱型图

Fig.2 Box plot graph of the original NIR spectrum

为消除在采集光谱信息过程中，基线漂移、 样本差异、环境光线等其他因素的干扰，对 154 组样本近红外光谱数据集，采用 MSC、SNV、SG 平滑滤波及它们之间相互组合等手段预处理光谱。以 ANN 建模，用总体测试集准确率评价光谱预处理效果，如表 1 所示。

表 1　不同预处理方法对建模的影响

Table 1　Effects of different pretreatment methods on modeling

预处理方法	总体测试集准确率（%）
原始光谱（剔除异常值）	86.96
MSC	91.30
SNV	91.30
SG(3 点)	86.96
SNV+MSC	91.30
SG(3 点)+MSC	95.65
SG(3 点)+SNV	91.30

由表 1 可见，选取采用 SG 平滑滤波法，选用平滑窗口为 3，再进行 MSC 预处理，得到预处理后的大豆近红外光谱数据集，建模效果最好。结合图 3 所示，光谱数据经过预处理，能够较好地消除高频噪声、基线漂移的影响。

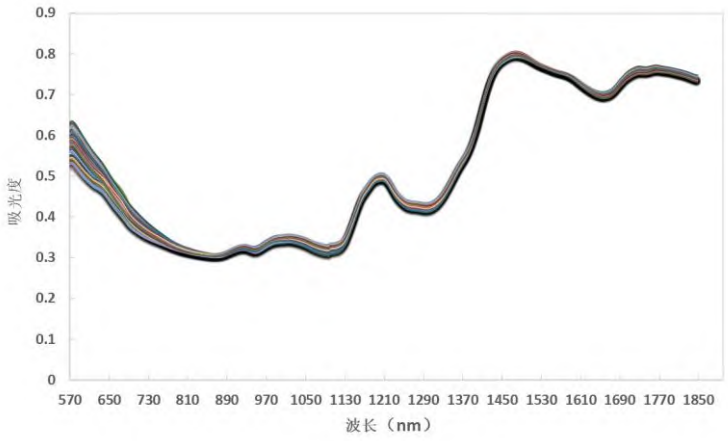

图 3 预处理后的大豆光谱

Fig.3 NIR spectrum after SG and MSC

2.3 主成分分析

主成分分析（Principal Component Analysis，PCA）是一种多元统计分析方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。主成分分析能够有效地降维数据，并消除众多信息相互重叠的信息部分。实验采集的大豆样品图谱从 570~1848 nm 共有 640 数据点，数据量大，冗余信息多。如表 2 所示，前 10 个主成分的累积方差贡献率已达到 99.966%，说明前 10 个变量很好地表征了原始数据集的主要特征信息，进而数据集从 154×640 减少到 154×10（10 个主成分）。

表 2 主成分分析累积方差贡献率

Table 2 The cumulative variance contribution of principal component analysis

主成分	方差贡献率（%）	累积方差贡献率（%）
1	93.392	93.392
2	2.763	96.155
3	1.691	97.846
4	0.779	98.625
5	0.602	99.227
6	0.464	99.691
7	0.160	99.851
8	0.050	99.901
9	0.045	99.946
10	0.020	99.966

如图 4 所示，在 PC1 和 PC2 的主成分得分图中，巴西样本主要分布在第 1 象限、第 2 象限和第 4 象限；美国样本主要分布在第 1 象限、第 2 象限和第 3 象限；阿根廷样本主要分布在第 3 象限和第 4 象限，乌拉圭样本主要分布在第 3 象限和第 4 象限。四个国家的样本分布过于分散，且重叠严重。在 PC2 和 PC3 的主成分得分图中，阿根廷样本分布在第三象限，

与分布在第二象限和第四象限的乌拉圭样本能够明显区分开来，但乌拉圭样本与美国、巴西样本都有重叠，且四个产地样本都有一定程度的重叠，故在主成分分析的基础上，需进一步优化算法，提高识别模型的准确率。

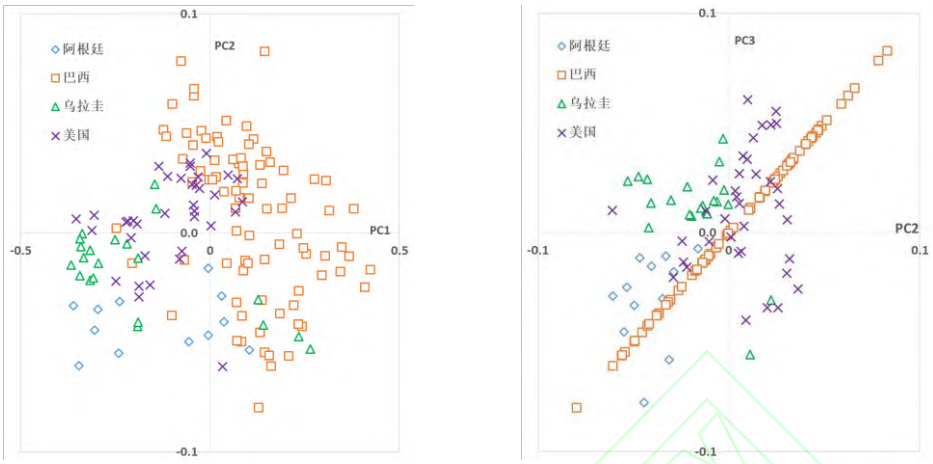

图 4　PC1、PC2、PC3 的主成分得分图

Fig.4 Principal component scores of PC1 ,PC2 and PC3

2.4　识别模型的建立与结果评价

选取主成分分析得到前 10 个主成分（154×10 维矩阵）为输入向量；设置“1”代表阿根廷大豆样本，“2”代表巴西大豆样本，“3”代表乌拉圭大豆样本，“4”代表美国大豆样本，作为目标向量，建立识别模型。训练集，测试集分别按照 70%、30%的比例随机选取，即从 154 个样本中随机抽取 108 个样本为训练集，46 个样本为测试集。分别采用 SVM，KNN，BP-ANN 建立识别模型。

SVM 是一种基于核的算法，其原理是把输入数据映射到一个高阶的向量空间，在这些高阶向量空间里，有些分类能够更容易的解决。采用 SVM 算法，优化参数，设定核函数为线性，核尺度为自动，框约束级别为 1，多类方法为一对一，建立识别模型。

KNN 是一种基于实例的算法，其原理是选取一定量的样本数据，然后根据特征近似性把新数据与样本数据进行比较。通过这种方式来寻找最佳的匹配，进而实现分类。采用 KNN 算法，优化参数，设定邻点个数为 1，距离度量为欧几里得（Eulidean），距离权重为等距离，建设识别模型。

ANN 是一种强有力的模仿人脑神经细胞的结构和功能的学习系统，能够实现输入与输出之间的高度非线性映射。采用 ANN 算法，训练集，验证集、测试集分别按照 70%、15%和 15%的比例随机选取，即从 154 个样本中随机抽取 108 个样本为训练集， 23 个样本为验证集，23 个样本为测试集。训练集用于网络训练时权值的调整，验证集用于提前停止训练以避免过拟合，测试集用于网络训练结束后，测试网络训练效果时使用。为了提高预测准确率与计算效率，优化模型参数，隐藏层节点数设定为 10，学习速率为 0.01，动量因子设定为 0.95 时,经过 23 次迭代，建立识别模型。

三种算法的建模效果，如表 3 所示，采用 SVM 建模，总体测试集准确率 89.13%，其中乌拉圭准确率较低为 66.67%；采用 KNN 建模,识别效果有所改善，总体测试集准确率为 91.30%，其中乌拉圭准确率提高为 70.00；采用 BP-ANN 建模效果最好，总体测试集准确率最高为 95.65%，其中阿根廷准确率为 100%，巴西准确率为 100%，乌拉圭准确率为 80%，

美国准确率为 100%。

表 3 不同算法建模的总体测试集准确率

Table 3 Accuracy of total test set by different algorithms modeling

国别	SVM		KNN		BP-ANN	
	错误个数	准确率（%）	错误个数	准确率（%）	错误个数	准确率（%）
阿根廷	0	100.00	0	100	0	100
巴西	2	92.31	1	95.65	0	100
乌拉圭	2	66.67	3	70.00	1	80.00
美国	1	90.91	0	100	0	100
总体	5	89.13	4	91.30	1	95.65

采用 ANN 建立的模型具有良好的鉴别能力，能够准确识别大豆样本的产地国别信息。如表 4 所示，全部数据集准确率为 98.70%，预测结果的错误数为 2，具体为 1 个“阿根廷”被误判为“乌拉圭”，1 个“美国”被误判为“乌拉圭”，说明所建立的 ANN 模型具有一定泛化能力。

表 4 人工神经网络识别结果

Table 4 Results of artificial neural network prediction

国别	测试集		验证集		训练集	
	错误个数	准确率（%）	错误个数	准确率（%）	错误个数	准确率（%）
阿根廷	0	100	0	100	0	100
巴西	0	100	0	100	0	100
乌拉圭	1	80.00	1	66.66	0	100
美国	0	100	0	100	0	100
总体	1	95.65	1	95.65	0	100

3 结论

采用箱型图校正法，剔除异常样本，经 SG 平滑滤波、MSC 处理原始红外光谱。采用主成分分析降维，将前 10 个主成分作为 ANN 模型输入向量，在学习速率为 0.01，动量因子设定为 0.95，输出层节点为 4，隐藏层节点为 10 时，模型识别准确率为 95.65%，全部数据集准确率为 98.7%，只有 2 个样本判断错误。由此可见，基于 ANN 模型的近红外光谱检测技术能够快速、准确地鉴别大豆产地,为加强进口大豆质量安全管理和原产地管理提供科学依据。

参考文献：

[1]韩立德,盖钧镒,张文明.大豆营养成分研究现状[J].种子,2003(5):58-60.
[2]马新华, 赵菊鹏, 龙阳,等.进口大豆检验检疫[M]. 北京: 中国农业出版社, 2018:1-2.
[3]潘广,杨帆,章桂明,等.进口转基因大豆品系检测及分析[J].中国油料作物学报,2020,42(2):298-305.
[4]龙沈飞,潘龙,吴阳,等.国产和进口大豆常规化学成分的综合分析[J].饲料工业,2020,41(7):55-59.
[5]袁俊杰,魏霜,龙阳,等.转基因大豆 MON87701 和 MON87708 双重实时荧光 PCR 检测技术的建立与应用[J].农业生物技术学报,2020,28(2):342-348.
[6]金俊,姜秋水,刘新.进口大豆热损贬值的因素分析与评估[J].中国油脂,2019,44(9):133-137.
[7]薛雅琳,王雪莲,赵会义,等.利用近红外分析技术测定大豆水分含量方法的研究[J].中国油脂,2009,34(7):69-71.
[8]王丽萍,陈文杰,赵兴忠,等.基于近红外漫反射光谱法的大豆粗蛋白和粗脂肪含量的快速检测[J].大豆科学,2019,38(2):280-285.
[9]宋雪健,王洪江,钱丽丽等.基于近红外光谱技术对草莓品质的快速检测研究[J].黑龙江八一农垦大学学报,2020,32(3):35-43.
[10]田翔,刘思辰,王海岗等.近红外漫反射光谱法快速检测谷子蛋白质和淀粉含量[J].食品科学,2017,38(16):140-144.
[11]Riccardo N, Simone C, Federico M,et al.Geographical characterization by MAE-HPLC and NIR methodologies and carbonic anhydrase inhibition of Saffron components[J]. Food Chemistry, 2017, 221:855-863.
[12]Haroon E T, Muhammad A, Gustav K M, et al.Authentication of the geographical origin of Roselle (*Hibiscus sabdariffa L*) using various spectroscopies: NIR, low-field NMR and fluorescence[J].Food Control,2020,114：1-8.
[13]Patrizia F, Silvia D L, Remo B,et al. Near infrared (NIR) spectroscopy-based classification for the authentication of Darjeeling black tea[J].Food Control,2019，100：292-299.
[14]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.
[15]赵海燕,郭波莉,魏益民,等.近红外光谱对小麦产地来源的判别分析[J].中国农业科学,2011,44(7):1451-1456.
[16]李勇,严煌倩,龙玲等.化学计量学模式识别方法结合近红外光谱用于大米产地溯源分析[J].江苏农业科学,2017,45(21):193-195.
[17]刘文静. 基于大豆异黄酮特征的大豆产地溯源研究[D].黑龙江八一农垦大学,2018.
[18]卢锡纯.基于脂肪酸含量的大豆产地溯源的研究[J].食品研究与开发,2018,39(16):55-59.
[19]Lai H Q, Xi J L ,Sun J C, et al. Multi-elemental Analysis by Energy Dispersion X-ray Fluorescence Spectrometry and Its Application on the Traceability of Soybean Origin[J]. Atomic Spectroscopy, 2020, 41(1): 20-28.
[20]鹿保鑫,马楠,王霞,等.基于电感耦合等离子体质谱仪分析矿物元素含量的大豆产地溯源[J].食品科学,2018,39(8):288-294.
[21]鹿保鑫,马楠,王霞，等.大豆有机成分辅助矿物元素指纹特征产地溯源[J].食品科学,2019,40(4):338-344.
[22]张勇,李雪,汪雪芳,等.基于脂肪酸组成的进口大豆鉴别技术研究[J].食品安全质量检测学报,2020,11(8):2375-2379.
[23]Akiko O, Akiko H, Izumi N.Determination of trace elements in soybean by X-ray fluorescence analysis and its application to identification of their production areas[J],Food Chemistry,2014,147: 318-326.
[24]全国粮油标准化委员会,GB/T 24870-2010 粮油检验 大豆粗蛋白质、粗脂肪含量的测定 近红外法[S].北京:中国标准出版社, 2010.