基于激光诱导击穿光谱技术结合机器学习算法的3种干腌火腿产地识别

郭 茂[1]，黄忠宇[1]，汪 杰[2]，周卫东[1*]
([1]浙江省光信息检测与显示技术研究重点实验室 浙江金华 321001
[2]浙江师范大学数学与计算机科学学院 浙江金华 321001)

摘要 火腿种类多，产地不同。本文采用激光诱导击穿光谱技术(LIBS)结合机器学习算法鉴别火腿原产地。采集16个火腿切片样品(4个如皋火腿样品、5个金华火腿样品、7个宣威火腿样品)的LIBS光谱数据，应用K近邻(KNN)、支持向量机(SVM)以及全连接神经网络(DNN)对火腿样品产地进行分类。采用主成分分析(PCA)对火腿样品的光谱数据降维处理，然后，分别结合KNN和SVM算法对样品进行分类，并研究对建模速度和预测准确率的影响。结果表明：KNN和SVM在结合PCA后，建模分析时间大幅减少；KNN、PCA+KNN、SVM、PCA+SVM 4种分类方法的平均正确率分别为70.53%，73.50%，79.53%，80.42%，使用PCA结合KNN和SVM时分类准确率有小幅度的提高；使用DNN对火腿样品进行分类，正确率最高达85.56%，相比于KNN和SVM，DNN对火腿LIBS光谱数据具有更高的分类正确率。结论：用LIBS结合机器学习算法区分不同产地的火腿样品是可行的.

关键词 火腿；激光诱导击穿光谱；机器学习；分类
文章编号 1009-7848(2022)10-0279-07 **DOI**：10.16429/j.1009-7848.2022.10.030

干腌火腿是我国的传统肉制品，历史悠久，风味独特，种类繁多，受到广大消费者的喜爱，这其中以浙江金华火腿、云南宣威火腿、江苏如皋火腿最为著名。火腿品质的好、坏与产地有着紧密的联系，如余功雄[1]指出：传统的金华火腿之所以长盛不衰，最主要的因素是：以中国名猪种“金华两头乌”的后腿为原料，加上金华地区特殊的气候条件和民间千年世代相传的腌制工艺，使其具有独特的地域性，离开这个特定的地域，是腌制不出真正的金华火腿的。说明对干腌火腿进行产地识别具有重要意义。

迄今为止，国内外已有一些科研人员将多种技术手段应用于干腌火腿的分类鉴别上，如：宋雪[2]分别基于电子鼻(一种电化学传感器阵列)和电子舌(一种味觉传感器阵列)对金华火腿和宣威火腿进行产地识别与品级评定，取得较好的结果。吕晓雷等[3]采用气-质谱联用技术区分不同年份的金华火腿。高韶婷[4]通过红外三级鉴定法分析不同产地和等级间干腌火腿的红外谱图，为不同产地干腌火腿及肉制品的鉴别提供了一种新的方法。姚璐[5]利用电子鼻采集试验数据，分别结合PCA和LDA的方法较好地区分特级、一级和二级的金华火腿，其中PCA区分效果好，LDA显示一级品和二级品之间有少部分重叠，之后，将高光谱成像系统获得的数据结合数据分析软件建立基于高光谱的金华火腿判别模型，训练集和验证集的总体识别率分别为96.19%和89.52%。Laureati等[6]采用理化分析、电子鼻分析、仪器质构分析、图像分析、感官评定、统计分析等多种方法将帕尔马火腿、圣丹尼尔火腿和托斯卡纳火腿3种火腿区分开来。Santos等[7]设计一种氧化锡多传感器系统，在结合人工神经网络后可以鉴别用不同饲料饲养的猪制作的火腿以及不同成熟时间的猪制作的火腿。然而，这些技术方法需对样品进行预处理，难以实现在线实时检测，具有一定的局限性。

激光诱导击穿光谱(Laser-induced breakdown spectroscopy，LIBS)是一种元素分析技术，由一束高能激光激发样品表面使之产生等离子体，其会发射表征样品组分信息的元素特征谱线。LIBS技术具有适用范围广(可用于固态、液态、气态)，对物质损伤小，检测速度快，样品无需预处理或处理简单等优点。目前LIBS技术广泛应用于钢铁制造

收稿日期：2021-10-12
基金项目：国家自然科学基金项目(61975186)
作者简介：郭茂(1997—)，男，硕士生
通信作者：周卫东 E-mail：wdzhou@zjnu.com

及加工[8]、环境监测[9]、生物医疗[10]、深空探测[11]、食品安全[12-13]等领域，具有极大的发展前景。

目前，LIBS 技术在物质的分类鉴别上也得到很好的应用。冯中琦等[14]将 LIBS 技术与化学计量学方法结合，可以快速、准确识别航空合金牌号。陈兴龙等[15]在用 LIBS 技术取得试验数据后，以主成分作为自组织映射神经网络的输入变量，可对火山灰岩、砂岩、白云岩实现 100%的准确分类。於筱岚等[16]利用 LIBS 技术结合 LDA 判别分析模型，鉴别了不同厂家生产的抹茶和不同杀青方式制成的绿茶粉。Bilge 等[17]利用 LIBS 技术对牛肉、猪肉和鸡肉进行研究，在结合 PCA 算法后，对 3 种肉的识别率达 83.37%。目前，将激光诱导击穿光谱技术应用于干腌火腿的分类鉴别还未见报道。

本文使用激光诱导击穿光谱技术对国内不同产地的 3 种著名干腌火腿进行分类，探究激光诱导击穿光谱技术结合机器学习算法区分产地的可行性，为后续干腌火腿的快速区分和检测提供新技术。

1 材料与方法

1.1 样品制备

浙江金华火腿，金字火腿股份有限公司；江苏如皋火腿，南通今天食品有限公司；云南宣威火腿，宣威市浦记火腿食品有限公司。

通过人工切片的方式将火腿切成 30 mm×30 mm×3 mm 的薄片，尽量选取瘦肉部分。获得如皋火腿样品 4 片，金华火腿样品 5 片，宣威火腿 7 片。

1.2 光路图与试验仪器

采用装置如图 1 所示，激光光源为 Plite 200-Ⅱ型双脉冲激光器（北京中科思远光电科技有限公司），输出激光波长 1 064 nm，激光能量 50 mJ，脉冲频率 1 Hz，脉冲宽度 8 ns。激光脉冲经反射镜反射后，由 1 块焦距为 100 mm 的透镜聚焦在样品表面产生等离子体。等离子体发射的特征光由光纤探头收集并由光纤传递至光谱仪（AvaSpec-2048-USB2 型光纤光谱仪）中，光谱仪采集范围为 196~510 nm，积分时间 2 ms，光谱分辨率 0.09~0.10 nm。激光脉冲和光谱仪采集间的延时由数字信号发生器控制，经试验条件优化，采集延时设置为 650 ns。待测火腿切片样品被放在二维移动平台上，可实现激光烧蚀位置的实时调节。

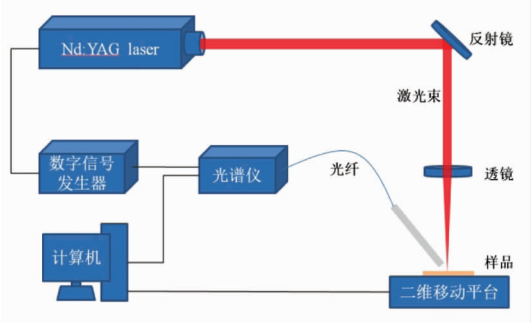

图 1 LIBS 试验装置示意图

Fig.1 The diagram of LIBS experimental device

1.3 试验步骤

在相同的试验条件下，先、后采集 4 片如皋火腿、5 片金华火腿和 7 片宣威火腿的光谱。每片火腿正反面各采集 150 个点，最终获得如皋火腿光谱 1 200 个，金华火腿光谱 1 500 个，宣威火腿光谱 2 100 个。其中，金华火腿和宣威火腿需采集更多数据的原因：相比于如皋火腿，这两种火腿更难获得有效的 LIBS 光谱，因此增加了样本数量。

2 结果与讨论

2.1 数据预处理

为了减少试验数据的波动，尽量排除激光能量不均和样品表面不平整带来的影响，采用最大最小归一化方法（Min-max normalization）对光谱数据进行预处理。最大最小归一化是将原始数据线性映射到[0，1]区间，归一化公式如下：

$$x^{*}=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{1}$$

式中，x——原光谱数据；$x_{\max}$——原光谱数据中强度最大值；$x_{\min}$——原光谱数据中强度最小值；x^{*}——最大最小归一化后的光谱数据。

在采集光谱过程中发现 3 种火腿在 422.752 nm 处都有较强的信号。依据此处的信噪比，每种火腿筛选出信噪比最高的 100 个光谱，作为之后机器学习的数据集。

每种火腿筛选出的 100 个光谱经平均得到的光谱图见图 2。可以发现，3 种火腿的 LIBS 光谱比较相似。

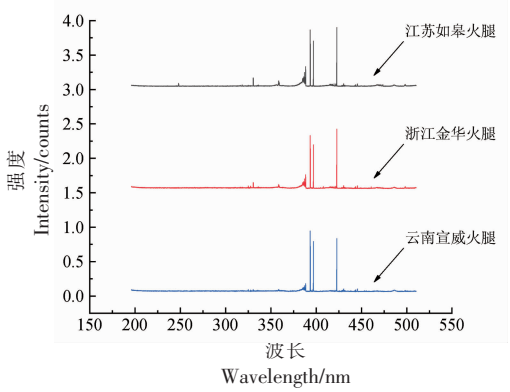

图 2　3 种火腿样品的 LIBS 光谱图

Fig.2　LIBS spectra of three ham samples

2.2　训练集与测试集的选取

训练集与测试集的划分见表 1。随机选取每种样品 70%的光谱数据作为训练集，剩下的 30%作为测试集。

在此必须指出：训练集与测试集的选取是随机的，当选取不同的训练集和测试集时，最终得到的预测正确率也是不同的。在试验中，将算法独立重复多次，以平均正确率作为衡量标准。

表 1　训练集与测试集的划分

Table 1　The division of training set and test set

样本名称	样本总数/个	训练集数量/个	测试集数量/个
如皋火腿	100	70	30
金华火腿	100	70	30
宣威火腿	100	70	30
总计	300	210	90

2.3　KNN(K 近邻)

K 近邻(K-Nearest Neighbor,KNN)是机器学习中一种基本的分类方法，分类时，对于待预测样本，根据 K 个最近的训练样本的类别，通过多数表决的方式进行预测。光谱数据中，每一个波长点对应一个特征。

正如 2.2 节所述，训练集和测试集的选取是随机的，由于不同的组合搭配会得出不同的正确率，因此将 KNN 算法独立试验了 100 000 次，每次都随机选取每种火腿光谱数据的 70%作为训练集，30%作为测试集。结果如图 3 所示。

图 3 中，横轴代表 90 个预测样本中预测正确的个数，纵轴代表 100 000 次独立试验中某个预测正确个数出现的次数。可以看出，结果基本符合正态分布。对其进行高斯曲线拟合，结果显示均值为 63.48，平均正确率为 70.53%。其中最优异的一次预测正确 77 个样本，最高正确率为 85.56%。

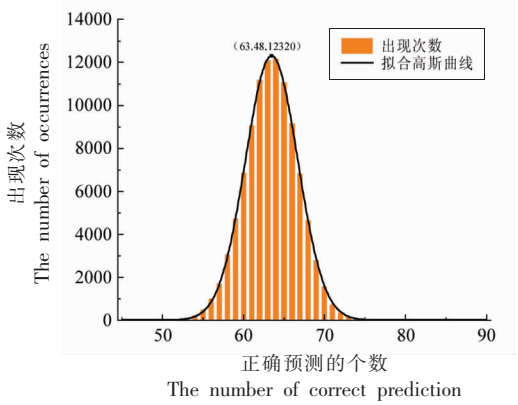

图 3　KNN 算法循环 100 000 次结果图

Fig.3　The results figure of KNN loops 100 000 times

2.4　支持向量机

支持向量机(Support vector machine，SVM)最早在上世纪 90 年代由 Cortes 等[18]提出，是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器。

将 SVM 用于多分类问题有多种思路。本文采用一对一(One-versus-one)，即每次只针对其中的某两类构建分类器模型。当有 N 个类别时，需要构建的分类器模型数量为 $N(N-1)/2$。最终由这些分类器投票，由票数决定最终的类别。该思路不适用于类别数量较多的情况，这是因为会导致需要构建的分类器数量急剧上升。本试验共有 3 个类别，只需要构建 3 个分类器。

核函数(Kernel function)是 SVM 模型的一个重要参数。用与 KNN 相同的方法将 SVM 用于火腿测试样本的预测。不同核函数的 SVM 模型平均预测正确率见表 2。

本试验的 SVM 模型采用线性核，其独立试验 100 000 次结果见图 4。

拟合出的高斯曲线均值为 71.58，平均正确率为 79.53%。其中，最优异的一次预测正确 85 个样本，最高正确率为 94.44%

表 2 SVM 模型不同核函数的平均预测正确率

Table 2 The accuracy rate of different kernel functions of the SVM model

核函数	平均预测正确率/%
线性核(linear)	79.53
二次多项式核(ploy2)	79.51
三次多项式核(ploy3)	76.40
高斯核(gaussian)	66.84

2.5 PCA

主成分分析 (Principal component analysis, PCA)是一种常用的无监督学习方法,可以实现对数据的降维。光谱数据常有数千个光谱点,数据量大。利用 PCA 处理数据可在有效保留光谱数据信息的同时, 减少数据量, 增加模型的建立分析速度。

用 PCA 处理 3 种火腿样品共计 300 个光谱数据后, 以前 3 个主成分得分绘制三维空间散点图(图 5)。

其中,第 1 主成分(PC1)、第 2 主成分(PC2)和第 3 主成分(PC3)分别包含了 70.32%,3.68%和 1.27%的方差信息。

从图 5 可以看出,3 种火腿存在大量的重叠,难以直接区分。

选取不同个数的主成分, 在该条件下分别结合 KNN 和 SVM,独立重复 1 000 次试验,计算平均预测正确率,PCA+KNN 和 PCA+SVM 结果如图 6 所示。

PCA+KNN 的平均预测正确率在主成分达到 22 个时到达最大。PCA+SVM 的平均预测正确率则在主成分达到 23 个后开始趋于稳定,在主成分达到 79 个时达到最大。

综上, 为保证分类正确率同时提升建模分析速度,对 PCA+KNN 和 PCA+SVM 分别选取主成分 22 个和 79 个。随后独立重复 100 000 次试验,比较最终结果。前 22 个主成分包含的方差信息为 80.79%, 前 79 个主成分包含的方差信息为 89.13%。

将 KNN、SVM、PCA+KNN、PCA+SVM 4 种方法的结果汇总,见表 3。

KNN 在结合 PCA 算法后,不论是平均正确率

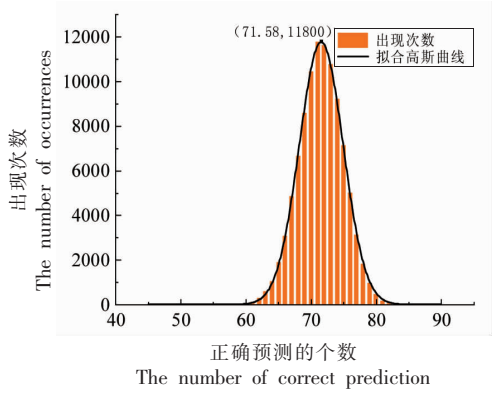

图 4 SVM 算法循环 100 000 次结果图

Fig.4 The results figure of SVM loops 100 000 times

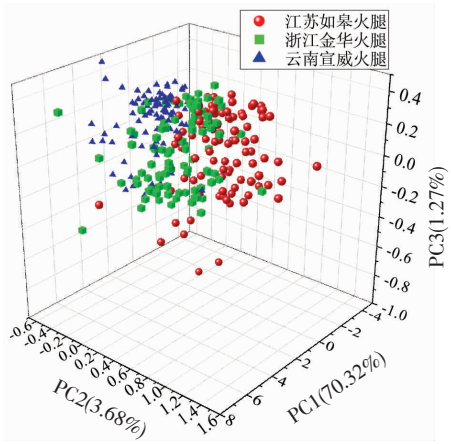

图 5 3 种火腿样品的 PCA 三维散点图

Fig.5 PCA three-dimensional scatter plots of three ham samples

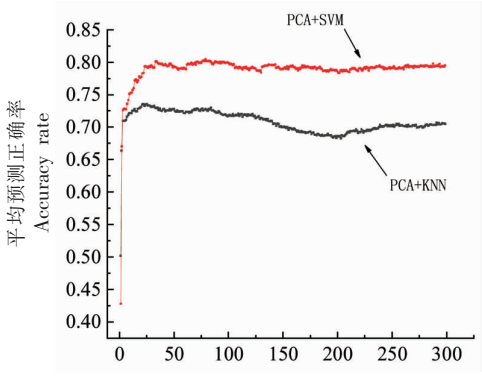

图 6 不同主成分数量下的平均预测正确率

Fig.6 The accuracy rate of different number of principal components

表 3　4 种方法结果汇总表

Table 3　Summary of the results of the four methods

算法	平均预测正确率/%	最高预测正确率/%	10 万次试验所需时间/s
KNN	70.53	85.56	64 537
PCA+KNN	73.50	88.89	4 424
SVM	79.53	94.44	12 847
PCA+SVM	80.42	94.44	2 775

还是最高正确率，均有小幅提升;PCA+SVM 相比 SVM 提升 0.89%的平均正确率。这是因为 PCA 算法降低了计算的复杂度，避免了过拟合现象。此外,PCA 实现了对数据的降维,从而大大加快了建模分析速度，这在物质的快速鉴别中具有重要意义。

2.6　全连接神经网络

全连接神经网络 (Deep neural networks, DNN)是最朴素的神经网络,也是当前广为运用的神经网络之一。相比于传统的神经网络模型,DNN 更强调其隐藏层的深度。DNN 的基本原理如图 7 所示。

根据采集到的光谱数据的实际情况，建立由输入层、隐藏层和输出层构成的全连接神经网络。处理初始数据并获取属性标签，取学习速率为 0.0005,通过 ReLU 和 Softmax 激活函数完成多分类任务。该网络运行结果见图 8。

通过该网络的预测结果可看出，在完成 150 次训练后，结果收敛到期望的误差，准确率达 85.56%,取得较好的结果。

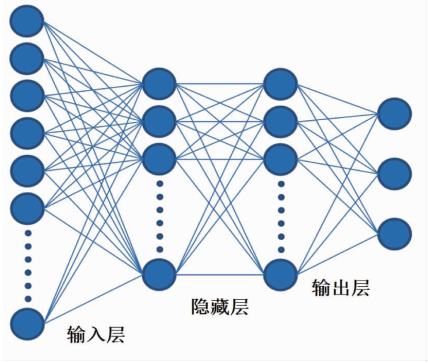

图 7　DNN 原理示意图

Fig.7　The Schematic diagram of DNN

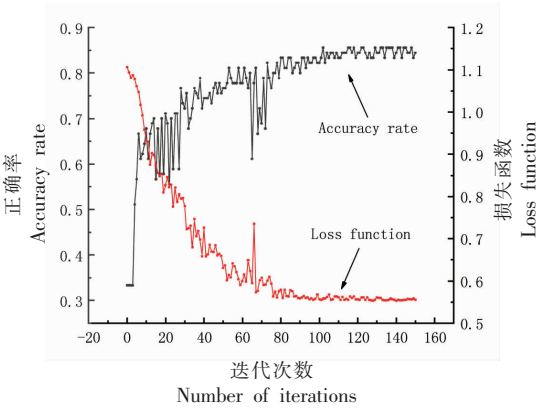

图 8　DNN 运行结果

Fig.8　The result of DNN

3　结论

利用激光诱导击穿光谱技术结合 4 种机器学习算法区分 3 种产地不同的火腿。KNN 的平均正确率为 70.53%,SVM 为 79.53%。对于干腌火腿的 LIBS 光谱数据,SVM 算法比 KNN 算法具有更高的分类正确率。用 PCA 对 3 种火腿的光谱数据进行预处理，分别取前 22 个主成分和前 79 个主成分作为 KNN 和 SVM 的输入变量,PCA+KNN 和 PCA+SVM 的平均正确率分别为 73.50%和 80.42%,与直接使用 KNN 和 SVM 相比,分类正确率均有提升,并且建模分析速度大幅提升。利用全连接神经网络构建的分类器，在 150 次训练后仍具有最高的分类正确率,为 85.56%。本研究结果为干腌火腿的产地快速区分和检测提供了新技术手段。

参　考　文　献

[1]　余功雄. 产地、原产地和商标——金华火腿工业产权保护的对策与思考[J]. 浙江师大学报，2000，25(3)：66-70.
YU G X. Place of production, country of origin and trademark - Countermeasures and thinking on industrial property protection of Jinhua ham[J]. Journal of Zhejiang Normal University, 2000, 25(3): 66-70.

[2]　宋雪. 金华火腿和宣威火腿风味品级研究[D]. 上海：上海海洋大学，2015.
SONG X. Study on the flavor and quality grade of Jinhua ham and Xuanwei ham[D]. Shanghai: Shang-

hai Ocean University, 2015.

[3] 吕晓雷, 韩剑众, 王彦波, 等. 金华火腿品质特征的指纹图谱研究[J]. 中国食品学报, 2013, 13(9): 201-206.
LV X L, HAN J Z, WANG Y B, et al. Finger prints of quality characters from Jinhua ham[J]. Journal of Chinese Institute of Food Science and Technology, 2013, 13(9): 201-206.

[4] 高韶婷. 基于多指纹图谱技术的我国三大干腌火腿风味品级评价研究[D]. 上海: 上海海洋大学, 2016.
GAO S T. Lavor and quality grade evaluation of three main dry-cured hams in China by multiple fingerprints technology[D]. Shanghai: Shanghai Ocean University, 2016.

[5] 姚璐. 金华火腿品质检测技术与分级方法研究[D]. 杭州: 浙江大学, 2012.
YAO L. The quality testing and grade discrimination of Jinhua ham[D]. Hangzhou: Zhejiang University, 2012.

[6] LAUREATI M, BURATTI S, GIOVANELLI G, et al. Characterization and differentiation of Italian Parma, San Daniele and Toscano dry-cured hams: A multi-disciplinary approach[J]. Meat Science, 2014, 96(1): 288-294.

[7] SANTOS J P, GARCÍA M, ALEIXANDRE M, et al. Electronic nose for the identification of pig feeding and ripening time in Iberian hams[J]. Meat Science, 2004, 66(3): 727-732.

[8] CUI M C, DEGUCHI Y, WANG Z Z, et al. Remote open-path laser-induced breakdown spectroscopy for the analysis of manganese in steel samples at high temperature[J]. Plasma Science and Technology, 2019, 21(3): 56-63.

[9] 李科学, 周卫东, 沈沁梅, 等. 激光烧蚀-快脉冲放电等离子体光谱技术分析土壤中的 Sn[J]. 光谱学与光谱分析, 2011, 31(8): 2249-2252.
LI K X, ZHOU W D, SHEN Q M, et al. Laser ablation and fast pulse discharge plasma spectroscopy analysis of Sn in soil[J]. Spectroscopy and Spectral Analysis, 2011, 31(8): 2249-2252.

[10] 杨思博. 基于 LIBS 技术的乳腺癌组织元素成像和聚类分析研究[D]. 哈尔滨: 哈尔滨工业大学, 2019.
YANG S B. Elemental imaging and cluster analysis of breast cancer tissues using laser-induced breakdown spectroscopy[D]. Harbin: Harbin Institute of Technology, 2019.

[11] DYAR M D, TUCKER J M, HUMPHRIES S, et al. Strategies for mars remote laser-induced breakdown spectroscopy analysis of sulfur in geological samples[J]. Spectrochimica Acta Part B: Atomic Spectroscopy, 2010, 66(1): 39-56.

[12] 林雨青, 田野, 陈倩, 等. 基于激光诱导击穿光谱技术分析鳕鱼中 8 种元素含量[J]. 食品科学, 2020, 41(14): 247-254.
LIN Y Q, TIAN Y, CHEN Q, et al. Analysis of eight elements in cod by laser induced breakdown spectroscopy[J]. Food Science, 2020, 41(14): 247-254.

[13] 徐聪, 范爽, 徐琢频, 等. 基于激光诱导击穿光谱检测水稻叶片镉的研究[J]. 量子电子学报, 2020, 37(3): 363-369.
XU C, FAN S, XU Z P, et al. Investigation of detection of cadmium in rice leaves based on laser-induced breakdown spectroscopy[J]. Chinese Journal of Quantum Electronics, 2020, 37(3): 363-369.

[14] 冯中琦, 张大成, 崔敏超, 等. 激光诱导击穿光谱技术识别航空合金牌号[J]. 冶金分析, 2020, 40(12): 99-104.
FENG Z Q, ZHANG D C, CUI M C, et al. Recognition of aerial alloy grades by laser-induced breakdown spectroscopy[J]. Metallurgical Analysis, 2020, 40(12): 99-104.

[15] 陈兴龙, 董凤忠, 陶国强, 等. 激光诱导击穿光谱在地质录井岩性快速识别中的应用[J]. 中国激光, 2013, 40(12): 243-248.
CHEN X L, DONG F Z, TAO G Q, et al. Fast lithology by laser-induced breakdown spectroscopy[J]. Chinese Journal of Laser, 2013, 40(12): 243-248.

[16] 於筱岚, 彭继宇, 刘飞, 等. 激光诱导击穿光谱技术用于抹茶和绿茶粉的快速鉴别[J]. 光谱学与光谱分析, 2017, 37(6): 1908-1911.
YU X L, PENG J Y, LIU F, et al. Fast identification of matcha and green tea powder with laser-induced breakdown spectroscopy[J]. Spectroscopy and Spectral Analysis, 2017, 37(6): 1908-1911.

[17] BILGE G, VELIOGLU H M, SEZER B, et al. Identification of meat species by using laser-induced breakdown spectroscopy[J]. Meat Science, 2016, 119(9): 118-122.

[18] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

Origin Identification of Three Kinds of Dry-cured Ham Based on Laser-induced Breakdown Spectroscopy Technology Combined with Machine Learning Algorithm

Guo Mao[1], Huang Zhongyu[1], Wang Jie[2], Zhou Weidong[1*]

([1]*Key Laboratory of Optical Information Detecting and Display Technology of Zhejiang Province, Jinhua 321001, Zhejiang*

[2]*College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321001, Zhejiang*)

Abstract There are many types of hams and their origins are different. This article uses laser-induced breakdown spectroscopy (LIBS) combined with machine learning algorithms to carry out research on the identification of ham origins. The LIBS spectrum data of 16 ham slice samples (4 Rugao ham samples, 5 Jinhua ham samples, 7 Xuanwei ham samples) were collected in the experiment, using K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Deep Neural Network (DNN) classifies the origin of ham samples, and studies the dimensionality reduction processing of the spectrum data of the ham samples using Principal Component Analysis (PCA), and then combines the KNN and SVM algorithms to classify the samples and the speed of modeling And the impact of forecast accuracy. The research results show that: KNN and SVM combined with PCA, the modeling and analysis time is greatly reduced; the average accuracy of the four classification methods of KNN, PCA+KNN, SVM, and PCA+SVM are 70.53%, 73.50%, 79.53%, and 80.42%, when using KNN and SVM combined with PCA, the classification accuracy is improved slightly; Using DNN to classify the ham samples, the classification accuracy rate can reach 85.56%, compared with KNN and SVM, DNN has a higher classification accuracy rate for ham LIBS spectrum data. The above show that LIBS combined with machine learning algorithm is feasible to distinguish ham samples from different origins.

Keywords ham; laser-induced breakdown spectroscopy; machine learning; classification