基于模型集群的东北/非东北大米产地高光谱鉴别方法研究

林 珑[1]，吴静珠[1*]，刘翠玲[1*]，于重重[1]，刘 志[2]，袁玉伟[2]

1. 北京工商大学食品安全大数据技术北京市重点实验室，北京 100048
2. 浙江省农业科学院农业部农产品信息溯源重点实验室，浙江 杭州 310021

摘 要 采集东北和非东北产地大米样本高光谱图像，筛选多个特征波长图像并提取图像特征，结合模式识别方法建立判别模型，并联合多个模型构成模型集群快速、准确判别东北/非东北大米产地。东北大米以粳米为主，主要涵盖长粒香，圆粒香，稻花香和小町米4个品种。为建立实用性强、适用范围广的东北/非东北大米产地判别模型，实验主要收集了国内粳米代表性产区且以上述4个品种为主的样本，构成原始样本集：其中东北产地5份，包括黑龙江(1)、吉林(2)、辽宁(2)，非东北产地5份，包括河北(1)、浙江(1)、江苏(2)、安徽(1)。每个产地样本随机选取100粒，共计100×10粒大米样本。采用芬兰Specim公司的SisuCHEMA高光谱成像系统采集样本900～1 700 nm高光谱图像。按照大米轮廓选取感兴趣区域提取出单粒大米样本的平均光谱，采用Kennard-Stone法按照4∶1划分训练集和测试集。应用连续投影算法筛选得到原始样本集光谱的8个特征波长：1 460.30，1 400.20，1 424.92，945.98，1 315.62，1 220.87，1 705.91和942.53 nm；采用方向梯度直方图分别提取8个波长下的图像特征，结合支持向量机建立基于单特征波长图像的东北/非东北大米产地鉴别模型，识别准确率分别为85.5%，77.5%，76.5%，73.5%，71%，68.5%，67%和65.5%；鉴于单模型识别率不高的现状，提出建立基于特征波长图像模型集群综合判别大米产地的策略，即按照单模型识别率从高到低排序后分别联合3个、5个和7个特征波长图像模型的预测结果，当预测样本判定为真的比率＞50%，则判定样本为真，反之则为假。联合1 460.30，1 400.20，1 424.92，945.98，1 315.62，1 220.87和1 705.91 nm七个波段的模型集合对测试集样本的识别率可达90.5%。实验结果表明高光谱结合模型集群策略可为建立性能稳健、适用范围广的东北/非东北大米产地快速检测模型提供切实可行的思路和方法参考。

关键词 高光谱；模型集合；东北大米；产地鉴别；方向梯度直方图

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)03-0905-06

引 言

我国近2/3以上居民以大米为主食[1]。随着国民生活水平的提高，优质高端大米日趋受到青睐，其中尤以东北大米为代表。东北大米种植于我国东北辽宁省、黑龙江省，吉林省平原地区。土壤肥沃、光照时间长、昼夜温差大为东北大米提供了良好的生长条件。真正的东北大米生长周期长，营养价值高，口感好，价格高。但是由于目前我国农产品市场准入制度和溯源体系尚不完善，不法商贩受经济利益驱动销售假冒或是掺假东北大米的事件频发，严重影响了消费者的权益。

传统的大米品质检测方法大多以感官判别和化学分析为主：感官判别受检测人员主观影响较大；化学分析方法检测精度高、但是检测繁琐、周期长，且对样本具有破坏性等弊端。现有传统检测技术[2]无法满足我国市场监督和大米流通行业日益增长的快速、无损检测需求。

高光谱技术以其图谱合一，信息量丰富，兼具外观和内观分析技术于一体的特点日趋成为大米品质快速检测领域的新兴热点。孙俊等[3]将从市场购买的东北长粒香大米和江苏溧水大米按照5种不同掺假水平制备掺伪东北大米样本，采用PCA分别对大米样本高光谱图像和高光谱数据(390～

收稿日期：2019-05-13，**修订日期**：2019-08-30

基金项目：农业部农产品信息溯源重点实验室开放课题，国家重点研发计划子课题(2018YFD0101004-03)，北京工商大学2019年基本科研业务费(PXM2019_014213_000007)资助

作者简介：林 珑，女，1994年生，北京工商大学食品安全大数据技术北京市重点实验室硕士研究生 e-mail：404223716@qq.com

*通讯联系人 e-mail：pubwu@163.com；liucl@btbu.edu.cn

1 050 nm)进行处理，建立了基于特征波长的 SVM 模型用于判别东北大米是否掺伪，其识别率最高可达 98%。王朝晖等[4]应用高光谱成像技术和 SPA 降维后，对梅河大米理化指标含量进行相关性分析，选出 9 个特征波长用于区分梅河大米和柳河县大米样品，识别率达 95%。表明高光谱结合模式识别方法用于大米产地鉴别、掺伪识别等具有较为光明的应用前景。但是市场上东北大米造假的情况极为复杂：有直接假冒产地的，有单种大米掺伪的，也有多种大米同时掺伪等情况。再加上东北大米产区辽阔，品种较多，即使同为东北大米，自然环境和品种的不同都会导致东北大米个体成分及组成、形态存在显著差异。这都为应用高光谱建立适应范围广、稳健性能好的东北大米产地鉴别模型带来了极大的干扰和困难。

以大米产地鉴别模型的适用范围和稳健性为出发点，选取主流东北大米品种和多个东北/非东北产区的大米样本构建模型适用范围广的样本集，通过高光谱特征提取方法结合模型集群策略来提高大米产地鉴别模型的稳健性，为建立符合市场需求的东北/非东北大米产地快速鉴别高光谱模型提供可行性探索。

1 实验部分

1.1 样本制备

东北大米产区辽阔，涵盖黑、辽、吉三省，主流品种以长粒香，圆粒香，稻花香和小町米 4 种为主。自然环境的不同会导致不同产区的大米的组成存在细微差异，如直链淀粉和支链淀粉的含量，尤其不同品种的大米，其形态、透明度等更是在外观上存在显著差异，如长粒香大米外观呈细长型，而圆粒香为圆短型。因此即使同为东北大米，个体也会因产区和品种存在较大差别。

东北大米以粳米为主。粳米产区主要分布在东北、江苏、安徽、浙江和河北产区，而籼米主要分布在湖南、湖北、广东、广西、江西和四川等地[5]。根据市场掺伪的实际情况，本实验选取样本均为粳米，产地及品种信息如表 1 所示，共收集 10 个产地样本。实验样本由浙江省农业科学院、北京古船米业有限公司分别于 2018 年 6 月和于 2018 年 11 月提供。

表 1 大米样本信息

Table 1 Rice samples information

类别	产地	品种	样本数
东北大米	黑龙江	长粒香	1
	吉林	稻花香	1
		圆粒香	1
	辽宁	小町米	2
非东北大米	江苏	长粒香	1
		小町米	1
	浙江	圆粒香	1
	安徽	小町米	1
	河北	小町米	1

1.2 高光谱图像采集

采用芬兰 Specim 公司 SisuCHEMA 高光谱成像系统采集大米样本高光谱图像。采集参数如下：相机型号为 FX17，波长范围 900～1 700 nm，光谱分辨率为 8 nm，共包括 224 个波段，曝光时间为 5 μs，帧频为 40 Hz。

大米颗粒相对较小且表面圆滑，易在扫描过程中由于载物台的移动出现晃动和偏移导致成像质量差。因此将大米样本置于 10×10 的数粒板上，将数粒板置于移动载物台进行成像实验，如图 1 所示。每种产地大米样本，随机选取 100 粒进行高光谱成像实验，共计采集 100×10 个大米样本的高光谱图像。

图 1 大米高光谱图像采集实验

Fig. 1 Rice hyper-spectral image acquisition experiment

1.3 数据处理

1.3.1 光谱特征提取

选用一种使矢量空间的共线性达到最小化的连续投影算法(successive projections algorithm，SPA)[6-8]作为光谱特征提取方法。连续投影算法通过向前循环，计算在 224 个波段中的某一波长对剩余波长的投影，选取投影最大的波长，之后选取的波长都与该波长线性最小，以消除高光谱数据中的冗余信息。

1.3.2 图像特征提取

选用方向梯度直方图(histograms of oriented gradients，HOG)作为图像特征提取方法，它是将一副图像分割成很多“细胞”再从中提取出特征。因为 HOG 是对图像的局部单元进行操作，所以它对图像几何和光学的形变都能保持很好的不变性[9-10]。

1.3.2 支持向量机分类原理

支持向量机(support vector machine，SVM)遵循结构风险最小化的学习过程，最小化了对未知数据的分类错误，是受监督的非参数统计学习模型[11-12]。SVM 在训练过程中避免了过拟合问题，解决了调参难和收敛慢的问题，并且保证找到的极值解就是全局最优解[13]。

2　结果与讨论

2.1　光谱特征波长提取

在 ENVI 4.8 软件中对大米样本高光谱进行黑白板校正后，按照大米轮廓选取感兴趣区域提取出每粒大米样本的平均光谱。根据样本集光谱信息，采用 KS 法按照 4∶1 划分训练集样本(800 个)和测试集样本(200 个)。图 2 所示为样本集中 10 个产地的大米平均光谱。由于大米化学成分相似，因此其光谱曲线轮廓非常相似，无法直接从谱图上分辨出东北和非东北大米产地的差异。

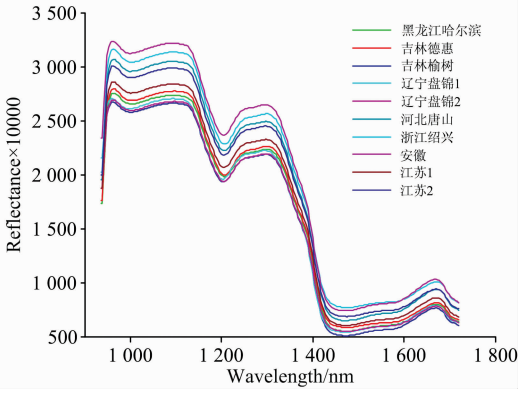

图 2　不同产地大米样本平均光谱

Fig. 2　Average spectra of rice from different origins

采用 SPA 法挑选出 8 个近红外特征波长为 942.52，945.98，1 220.87，1 315.62，1 400.20，1 424.92，1 460.30 和 1 705.91 nm，如图 3 所示。其中 942.52 和 945.98 nm 附近主要反映了游离水的 O—H 伸缩振动的二级倍频信息；1 220.87和 1 315.62 nm 则集中反映了 C—H 第二组合频的信息，淀粉、蛋白等成分中含有丰富的 C—H 基团；1 400.20，1 424.92 和 1 460.30 nm 附近信息量较为集中，既有游离水的 O—H 一级倍频信息，也有 C—H 的组合频信息，还有酰胺的 N—H 一级一级倍频信息；1 705.91 nm 主要反映了—CH_3 和—CH_2 的一倍频信息。因此采用 SPA 法筛选得到的特征波长与大米成分如水分、淀粉、蛋白等紧密相关[14]。

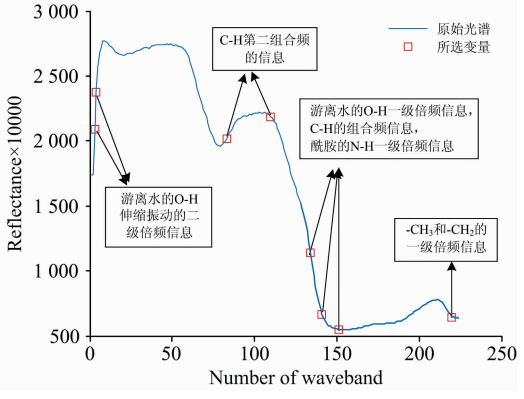

图 3　SPA 筛选特征波长结果图

Fig. 3　Characteristic wavelength selected by SPA

2.2　单波长图像 HOG 特征提取

针对上述 8 个特征波长，提取相应波长处的的图像，采用 HOG[15] 提取图像特征，首先将图像缩放至 256×256 后，采用 Gamma 校正对图像进行颜色空间的归一化，降低图像局部阴影和光照变化所产生的影响，抑制噪音干扰，并对图像每个像素的梯度方向和大小进行计算。再将图像分成 8×8 的细胞单元，统计梯度直方图，应用梯度的幅值进行投票，然后将相邻的细胞组成块并对重叠部分进行直方图归一化。最后将所有块中的梯度方向直方图合并组成特征向量，具体步骤如图 4 所示。

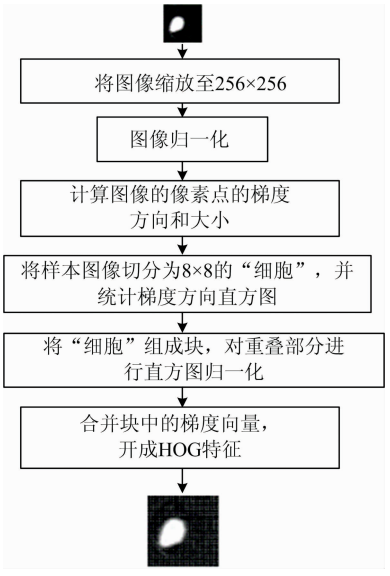

图 4　HOG 特征提取流程图

Fig. 4　Flow chart of HOG feature extraction

2.3　基于单波长图像特征的大米产地鉴别模型的建立

实验采用 SVM(线性核函数)分别建立了基于 8 个单波长图像 HOG 特征的东北/非东北大米产地模型。单波长模型的训练集识别率可以达到 100%，测试集识别率如表 2 所示。根据识别率高低排序可得，在 1 460.30，1 400.20 和 1 424.92 nm 波长下建立的分类模型识别率相对较好，分析其原因主要由于该区间反映的信息极为丰富，涵盖了 O—H，N—H 和 C—H 基团，与大米成分所反映出的特征信息紧密相关。其中尤以 1 460.30 nm 处所建模型识别率最高，而该波长附近正是反映伯酰胺中 N—H 对称和反对称伸缩振动的组合频谱带。该基团反映出了东北大米和非东北大米在蛋白质成分上有显著差异。但是总体而言，基于单特征波长图像的模型识别率不高，有进一步提升的空间。

表 2　基于单波长图像 HOG 特征的大米产地鉴别模型识别率

Table 2　Rice recognition rate based on single model

波长/nm	识别率/%	波长/nm	识别率/%
1 460.30	85.5	1 315.62	71.0
1 400.20	77.5	1 220.87	68.5
1 424.92	76.5	1 705.91	67.0
945.98	73.5	942.53	65.5

2.4 基于多波长图像特征的大米产地鉴别模型集群的建立

为建立适用范围广的判别模型，本实验中收集的样本来源差异较大，如品种和产地的相互交叉等，因此同一样本在不同的特征波长处反映的光谱信息也存在显著差异，直接导致同一样本在不同的单波长模型中存在截然不同的识别结果。鉴于上述单特征波长图像模型识别率不高的实验结果，提出采用多模型共识判别策略，即联合多个单特征波长图像模型，通过模型集群来综合判别大米产地。判别流程如图 5 所示。假设子模型个数为 n，采用 n 个子模型预测同一样本可以得到 n 个识别结果，当识别结果中识别为真的比率＞50%，则判定样本为真，反之则为假。

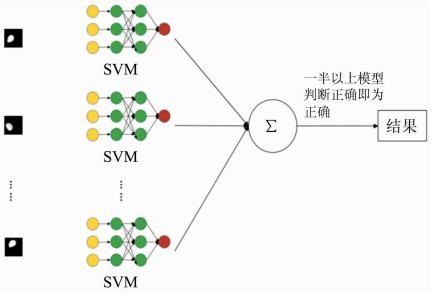

图 5 模型集群共识判别流程

Fig. 5 Multi-model discrimination diagram

为了保证综合判别的结果不会出现同一个样本判别为真和假的识别率相同，本实验确定联合子模型个数为奇数 3，5 和 7。为了精简组合个数，首先根据表 2 中单波长子模型的识别率从高到低进行排序，然后依次选取子模型进行组合判别。以联合 3 个波长建立模型集群为例，如表 3 所示。以单波长下模式识别率最高的 1 460.30 和 1 400.20 nm 两个子模型为基准，依次顺序选取剩余的 5 个单波长子模型进行联合判别，则有如表 3 所示的 6 种组合可能。从表 3 中可知，联合 3 个模型后模型识别率均有了一定程度的提高。其中联合 1 315.62 nm 波长的模型识别率最高，达 88%。1 315.62 nm 处反映了 C—H 第二组合频的信息，淀粉、蛋白等成分中含有丰富的 C—H 基团。而东北大米和非东北大米在淀粉组成和蛋白质含量确实存在显著差异。

表 3 三波长联合模型识别率

Table 3 Recognition rate based on three combined models

固定波长/nm	联合波长/nm	识别率/%
	1 424.92	87.0
	945.98	87.5
1 460.30	1 315.62	88.0
1 400.20	1 220.87	85.5
	1 705.91	85.5
	942.53	86.5

同理固定表 2 中前 4 个识别率最高的 1 460.30，1 400.20，1 424.92 和 945.98 nm 波长的子模型，依次顺序选取剩余的 4 个单波长子模型进行联合判别，则有如表 4 所示的 4 种组合可能。从表 4 中可知，分别联合 1 315.62 和 1 705.91 nm 处模型，模型识别率得到了进一步提高。而该两个波段同样反映了淀粉、蛋白质等的 C—H 和—CH3 基团信息。

表 4 五波长联合模型识别率

Table 4 Recognition rate based on five combined models

固定波长/nm	联合波长/nm	识别率/%
1 460.30	1 315.62	88.5
1 400.20	1 220.87	87.0
1 424.92	1 705.91	88.5
945.98	942.53	88.0

固定表 2 中前 6 个识别率最高的 1 460.30，1 400.20，1 424.92，945.98，1 315.62 和 1 220.87 nm 波长的子模型，依次顺序选取剩余的 2 个单波长子模型进行联合判别，则有如表 5 所示的 2 种组合可能。模型识别率最高可达 90.5%。综合表 2—表 5 可得关键波长处的子模型对模型集群判别结果起主要作用，如 1 460.30 和 1 400.20 nm 处的子模型；联合模型个数越多，模型集群识别率也越高，但是模型识别率的提高速度较为缓慢。

表 5 七波长联合模型识别率

Table 5 Recognition rate based on seven combined models

固定波长/nm	联合波长/nm	识别率/%
1 460.30		
1 400.20	1 705.91	90
1 424.92		
945.98		
1 315.62	942.53	90.5
1 220.87		

3 结 论

采集了 10 个产地、4 个品种共计 1 000 粒大米样本的高光谱图像，采用 SPA 法针对样本集光谱筛选出 8 个特征波长，分别提取 8 个特征波长对应图像的 HOG 特征，建立基于单波长图像特征的 SVM 模型。将单特征波长图像模型的识别率高低排序后，联合 3 个、5 个、7 个单波长模型对大米产地进行共识判别，可将东北/非东北大米产地的识别率从单模型的 85.5%显著提高到 90.5%。实验结果表明基于高光谱技术和机器学习算法的模型集群共识策略有望为建立稳健、切实可行的大米产地溯源模型提供思路和方法参考。

References

[1] YING Xing-hua, JIN Lian-deng, XU Xia, et al(应兴华, 金连登, 徐 霞, 等). Quality and Safety of Agricultural Products(农产品质量与安全), 2010, (6): 40.

[2] FU Long-sheng, Elkamil Tola, Ahmad Al-Mallahi, et al. Biosystems Engineering, 2019, 183: 184.

[3] SUN Jun, JIN Xia-ming, MAO Han-ping, et al(孙 俊, 金夏明, 毛罕平, 等). Journal of Agricultural Engineering Research(农机化研究), 2014, 30(21): 301.

[4] WANG Zhao-hui, YANG Jun-zhou, WANG Yan-hui, et al(王朝晖, 杨郡洲, 王艳辉, 等). Journal of Chinese Cereals and Oils Association(中国粮油学报), 2019, 34(11): 113.

[5] ZHANG Min, MIAO Jing, SU Hui-min, et al(张 敏, 苗 菁, 苏慧敏, 等). Food Science(食品科学), 2017, 38(16): 110.

[6] ZHAO Liu, QI Hai-jun, JIN Xiu, et al(赵 刘, 齐海军, 金 秀, 等). Jiangsu Agricultural Sciences(江苏农业科学), 2018, 46(17): 235.

[7] Filho H A D, Galvão R K H, Araùjo M C U, et al. Chemometrics and Intelligent Laboratory Systems, 2004, 72(1): 83.

[8] Araùjo M C U, Saldanha T C B, Galvão R K H, et al. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65.

[9] XU Xiao-yu, YAO Peng(徐笑宇, 姚 鹏). Computer Engineering and Applications(计算机工程与应用), 2016, 52(11): 175.

[10] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection, Computer Vision and Pattern Recognition, 2005. CVPR2005. IEEE Computer Society Conference on, 2005. 886.

[11] Giorgos Mountrakis, Jungho Im, Caesar Ogole. Elsevier, 2010, 66(3): 247.

[12] Baassou Belkacem, He Mingyi, Mei Shaohui. An Accurate SVM-Based Classification Approach for Hyperspectral Image Classification, IEEE Computer Society (Geoinformatics 2013), 2013.

[13] WANG Hai-yan, LI Jian-hui, YANG Feng-lei(汪海燕, 黎建辉, 杨风雷). Application Research of Computers(计算机应用研究), 2014, 31(5): 1281.

[14] SONG Xue-jian, QIAN Li-li, ZHANG Dong-jie, et al(宋雪健, 钱丽丽, 张东杰, 等). Food Science(食品科学), 2018, 39(16): 321.

[15] DING Fei, LIU Gui-hua, GAO Jun-qiang(丁 飞, 刘桂华, 高军强). Automation Instrument(自动化仪表), 2017, 38(3): 45.

Study on Hyperspectral Identification Method of Rice Origin in Northeast/Non-Northeast China Based on Conjunctive Model

LIN Long[1], WU Jing-zhu[1*], LIU Cui-ling[1*], YU Chong-chong[1], LIU Zhi[2], YUAN Yu-wei[2]

1. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China
2. Key Laboratory of Information Traceability of Agricultural Products, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China

Abstract Hyperspectral images of rice from northeast/non-northeast regions were collected, and spectral images at characteristic wavelengths were screened. The clustering combination of image features and pattern recognition method was established to quickly and accurately identify northeast/non-northeast rice origin. Northeast rice is mainly japonica rice,and the typical northeastern rice varieties include long-grain, round-grain, rice flower and Xiaoding rice. Considering the practicability and applicability of rice origin identification model,samples of 10 origins and 4 varieties above were collected to form the original sample set. Among them, there are five northeastern origins, including Heilongjiang (1), Jilin (2), Liaoning (2), and five non-northeastern origins, including Hebei (1), Zhejiang (1), Jiangsu (2) and Anhui (1). 100 samples were selected randomly from each producing area. Hyperspectral images of 100×10 rice samples were collected using SisuCHEMA hyperspectral imaging system (Specim, Finland)in the range of 900～1 700 nm. Extracting the average spectra of a single rice sample by selecting the region of interest according to the rice contour,Kennard-Stone method was used to divide training set and test set according to the ratio of 4∶1. Eight characteristic wavelengths were screened by Successive Projections Algorithm (SPA): 1 460.30, 1 400.20, 1 424.92, 945.98, 1 315.62, 1 220.87, 1 705.91, 942.53 nm. The eight models were built respectively by HOG features extracted from single characteristic wavelength Image and SVM to identify the rice origin whether it was from northeast or non-northeast China. The recognition accuracy was as follows: 85.5%, 77.5%, 76.5%, 73.5%, 71%, 68.5%, 67%, 65.5%. In view of the low recognition rate of single model, a strategy of establishing model cluster based on single characteristic wavelength

image model to synthetically discriminate rice origin was proposed. According to the recognition rate of single model from high to low, the cluster models were established by respectively combining three, five and seven the signal models above. While the probability of the sample judged to be true predicted by the conjunctive model is greater than 50%, the sample will be judged to be true, otherwise it will be false. The experimental results showed that the recognition rate of the test set samples can reach 90.5% by combining the model sets of 1 460.30, 1 400.20, 1 424.92, 945.98, 1 315.62, 1 220.87 and 1 705.91 nm bands. This study shows that hyperspectral technology combined with the strategy of conjunctive model consensus can provide feasible and effective methods to establish a robust and wide applicability model to recognize the rice origin (northeast/non-northeast) rapidly.

Keywords Hyperspectral image; Conjunctive model; Northeast rice; Origin identification; HOG

(Received May 13, 2019; accepted Aug. 30, 2019)

* Corresponding authors