红外光谱数据融合对美味牛肝菌产地鉴别

胡翼然1,李杰庆1,刘鸿高2,范茂攀1*,王元忠3*

1. 云南农业大学资源与环境学院,云南昆明 650201

2. 云南农业大学农学与生物技术学院,云南昆明 650201

3. 云南省农业科学院药用植物研究所,云南昆明 650200

摘 要 近年来食品安全问题频发,消费者愈加重视食品原产地的环境安全,导致地理标志产品的需求增加。美味牛肝菌(Boletus edulis)作为一种健康食品,其产品品质受原产地环境影响较大,为保护消费者的身体健康,防止假冒伪劣产品进入市场,急需一种高效、廉价的美味牛肝菌产地鉴别技术。采用数据融合策略结合偏最小二乘判别(PLS-DA)模型对美味牛肝菌的产地进行鉴别。扫描来自8个产地(昆明、楚雄、玉溪、迪庆、大理、保山、文山和曲靖)141个样品的傅里叶变换近红外光谱和傅里叶变换中红外光谱。使用 Kennard-Stone 算法将所有样品划分为2/3 的训练集和1/3 的预测集,利用三种融合策略(低级、中级和高级)对4个单一光谱矩阵:近红外的菌柄(N-b)、近红外的菌盖(N-g)、中红外的菌柄(M-b)、中红外的菌盖(M-g),建立偏最小二乘判别(PLS-DA)模型。用交叉验证均方根误差(RMSECV)和预测均方根误差(RMSEP)评价模型稳定性,非错误率(NER)、训练集正确率和预测集正确率评价模型分类性能,综合多种评价指标,找出美味牛肝菌产地鉴别的最佳方法。结果表明:(1)近红外和中红外光谱均能鉴别美味牛肝菌产地;(2)中红外光谱所建立的模型优于近红外光谱所建立的模型;(3)三种融合策略均可提高美味牛肝菌的产地鉴别效果,产地鉴别效果优劣依次为中级融合、高级融合、低级融合、单一光谱模型。通过融合近红外和中红外光谱使用 PLS-DA 进行基于特征值 LV 的中级融合策略,建立不同产地美味牛肝菌鉴别模型,有最少的变量数(49),最高的产地训练集正确率(100%),最高的产地预测集正确率(100%),最高的产地预测集正确率(100%),最低的 RMSEP(0.133),实现了美味牛肝菌产地的快速、准确鉴别,可以作为美味牛肝菌产地溯源的一种可靠方法。

关键词 美味牛肝菌;产地鉴别;数据融合;傅里叶变换中红外光谱;傅里叶变换近红外光谱 中图分类号:O433.4 文献标识码:A DOI:10.3964/j.issn.1000-0593(2020)04-1276-07

引 言

近年来由于环境污染造成的食品安全问题频繁发生,食品质量安全引起全国高度重视^[1]。地理标志产品来源于特定地区,有特定的加工方法,解决了消费者和生产者之间信息不对等的问题,其产品质量安全受到消费者的信任^[2]。作为牛肝菌之王的美味牛肝菌(*Boletus edulis* Bull.; Fr.)富含维生素、膳食纤维、矿质元素、人体必需的氨基酸、不饱和脂肪酸等,是高蛋白、低脂、低钠的健康食品同时还具有抗氧化、抗疲劳、抗突变、增强免疫力等药用活性^[3]。美味牛肝菌品质受生长环境(温度、降水等)影响巨大^[4],导致不同地区之间的美味牛肝菌营养成分有显著差异^[5-6],进而导致食

药用价值的差异。我国云南省地形复杂,晴雨交错的气候生态环境特征有利于美味牛肝菌的生长,美味牛肝菌产量丰富,约占全国出口量的70%。根据各地气候、地理、文化等特点,可以建立相应的野生食用菌地理标志,形成独具特色的品牌。为此,现阶段急需一种高效、廉价的美味牛肝菌产地鉴别技术,以确定美味牛肝菌的真实产地,防止假冒伪劣产品进入市场,保护消费者身体健康和权益。

近红外和中红外光谱具有无损、快速、便捷、廉价等优 点,广泛应用于食品质量控制中^[78]。但利用单一来源的光 谱信息无法完全表征样品中的化学信息,产地鉴别效果弱。 而数据融合可以使不同来源的光谱信息相互补充,从而增强 模型性能,实现样品的精确鉴别。近年来融合近红外和中红 外光谱在食品质量控制方面有许多成果,如Li等^[9]融合近

e-mail: huyiran94@126.com

* 通讯联系人 e-mail: boletus@126.com; mpfan@126.com

作者简介:胡翼然,1994年生,云南农业大学资源与环境学院硕士研究生

收稿日期: 2019-08-15,修订日期: 2019-12-22

基金项目:国家自然科学基金项目(31660591),云南省农业基础研究联合专项基金项目(2018FG001-033)资助

红外与中红外的光谱数据,使用偏最小二乘判别(partial least squares discriminant analysis, PLS-DA)模型进行低级融合、中级融合、高级融合,成功鉴别3种植物油掺假6个不同来源橄榄油。Li等^[10]融合近红外与中红外的光谱数据,使用随机森林进行低级融合、中级融合、高级融合,成功鉴别三七地理位置。

现阶段,野生牛肝菌的产地鉴别以中红外光谱为主,基 于近红外光谱对野生牛肝菌产地鉴别未见报道。本研究的主 要目的是探讨融合近红外光谱与中红外光谱鉴别不同产地美 味牛肝菌的可行性,对4个单一信息矩阵:近红外的菌柄 (N-b)、近红外的菌盖(N-g)、中红外的菌柄(M-b)、中红外 的菌盖(M-g),结合 PLS-DA 进行低级、中级、高级数据融 合,选出美味牛肝菌产地鉴别最佳方法,为野生牛肝菌产地 鉴别提供参考。

1 实验部分

1.1 材料

141 份美味牛肝菌采自云南 8 个产地,分别为昆明、楚 雄、玉溪、迪庆、大理、保山、文山、曲靖(表 1)均由云南农 业大学刘鸿高教授鉴定。样品采集后去除土样等杂质,用纯 净水清洗干净,置于 50 ℃烘箱烘干至恒重,高速粉碎机 (FW-100型)粉碎,过 80 目标准筛盘,分别储存于自封袋 中,保存于避光处。

表 1 美味牛肝菌产地信息 Table 1 The specific geographical origin

information of Boletus edulis								
产地	样本数	经纬度	海拔/m					
云南省昆明市安宁县温水村	9	N24°34'34.92", E102°23'16.15"	1 958					
云南省昆明市安宁县八街镇	10	N24°42′51.28″, E102°18′15.60″	1 939					
云南省昆明市安宁县八街镇凤仪村	8	N24°41′47.27″, E102°19′39.62″	2 602					
云南省昆明市石林县蓑衣山	10	N24°38′50.75″, E103°20′38.70″	1 917					
云南省楚雄州南华县龙川镇	10	N25°11′39.86″, E101°16′33.78″	1 857					
云南省楚雄州南华县沙桥镇	4	N25°14′28.75″, E101°08′36.81″	1 920					
云南省玉溪市易门县普贝乡	11	N24°36′55.77″, E102°10′51.78″	1 980					
云南省玉溪市易门县铜厂	27	N24°43'19.80", E102°2'0.52"	2 180					
云南省迪庆州香格里拉县普达措	5	N27°51′6.50″, E99°46′39.15″	3 541					
云南省迪庆州维西县保和镇	7	N27°12′12.77″, E99°17′56.21″	2 320					
云南省大理州弥渡县弥城镇	10	N25°21′27.93″, E100°31′44.58″	1 668					
云南省大理州鹤庆县松桂镇	7	N26°21′15.54″, E100°11′8.17″	2 150					
云南省保山市隆阳区	9	N25°6′16.79″, E99°7′44.17″	1 578					
云南省文山市东山乡	7	N23°24′28.12″, E104°16′15.23″	1 600					
云南省曲靖市马龙区旧县镇	7	N25°23'5.93", E103°24'43.25"	2 300					

1.2 光谱信息采集

1.2.1 近红外光谱

近红外光谱是由 Antaris Ⅱ型傅里叶变换近红外光谱仪 (Thermo Fisher 公司, USA)用漫反射显微镜采集。称取 20 g 样品混合均匀后,置于玻璃器皿中压缩,扫描。扫描波数 范围 10 000~4 000 cm⁻¹,分辨率 4 cm⁻¹,信号扫描累计 64 次。每个样本重复扫描 3 次,取平均光谱。

1.2.2 中红外光谱

中红外光谱是由 Frontier 型傅里叶变换红外光谱仪 (Perkin Elmer 公司, USA)采集。取(1.5 ± 0.2)mg 美味牛肝 菌样品和(150 ± 20)mg KBr 粉末在研钵中磨细混匀,再将细 粉倒入磨具中压成薄片,扫描。扫描波数范围 4 000~400 cm⁻¹,分辨率 4 cm⁻¹,信号扫描累计 16 次,每个样本重复 扫描 3 次,取平均光谱。

1.3 偏最小二乘判别分析及评价

PLS-DA 是在标准 PLS 回归算法的基础上建立的线性监督分类方法,从变量矩阵(x)中寻找与分类矩阵(y)有最大协方差的变量,其中 y 分为 2 类, y=1 代表样本属于特定分

类, y=0 代表样品不属于特定分类,最终得出每个样品分类 到每一类的概率。该算法解释样本观测数目少且可以降低样 本间多重共线性的影响。使用 Kennard-Stone 算法划分数据 集(141)的 2/3 为训练集(97)和 1/3 为预测集(44)以提高模 型识别能力,避免随机选择的不可重复性。

根据交叉验证均方根误差(root mean square error of cross validation, RMSECV)和预测均方根误差(root mean square error of prediction, RMSEP)两个重要参数评价模型稳定性, RMSECV≥RMSEP 且数值越小,表明模型稳定且预测能力越好。综合训练集和预测集的正确率再加上非错误率(non-error rate, NER)评价模型分类性能。NER 是灵敏度和特异性的平均值, NER 越高模型分类性能越强。

1.4 数据融合

低级融合又名数据级融合,直接将多个数据矩阵串联得 到一个新的数据矩阵,再建立判别模型;中级融合又名特征 级融合,将多个特征值数据矩阵串联得到一个新的数据矩 阵,再建立判别模型。潜在变量(latent variable, LV)是一种 常见的特征提取方法,根据 Q² 第一次到达最大值时的因子 数确定最优 LV 个数,最大程度的表征样品信息;高级融合 又名决策级融合,各数据源进行分析获得独立结果,再根据 一定准则对结果进行融合,最终得到整体一致的决策。本文 基于"模糊集合论",首先提取各原始数据的 LV 结合 PLS-DA 建模,把模糊现象(同一样品在不同光谱信息来源下有不 同分类结果)通过最小值(Min)、最大值(Max)、平均值 (Avg)和乘积(Prod)这4种运算符连接,再进行多数投票, 表决出最终样品分类结果^[11]。

2 结果与讨论

2.1 光谱分析

图 1 为八个不同产地美味牛肝菌的近红外和中红外平均 光谱图,从图 1(a)和(b)可以看出近红外光谱在 8 272~ 8 474,5 782~5 790,5 155~5 169,4 751~4 840,4 583~ 4 580,4 331~4 339 cm⁻¹处明显可见。8 272~8 474 cm⁻¹与 C—H的第二泛音有关;5 782~5 790 cm⁻¹与 C—H的第一 泛音有关;5 155~5 169 cm⁻¹与 C—O 的第二泛音有关可 能与水、蛋白质有关; 4 751~4 840 cm⁻¹与 C=O 和 O-H 组合带的第一泛音有关; 4 583~4 586 cm⁻¹与 N-H 的组合 带有关; 4 331~4 339 cm⁻¹是关于-CH₂ 多糖的反对称伸缩 振动^[12]。

从图 1(c)和(d)可以看出中红外的 14 个特征峰在 3 380 ~3 396, 2 922~2 933, 1 639~1 642, 1 550~1 560, 1 445 ~1 521, 1 405~1 410, 1 078~1 082, 1 032~1 036, 873, 840, 798, 776~771, 696~698 和 615~620 cm⁻¹ 处明显可 见。3 380~3 396 cm⁻¹与O—H 和 N—H 的伸缩振动有关可 能与糖、纤维素有关; 2 922~2 933 cm⁻¹与 N—H 和 C—H 的伸缩振动有关; 1 642~1 639 cm⁻¹与酰胺 I 带和酰胺 II 带 的 C—O 伸缩振动有关; 1 550~1 560 cm⁻¹与亚硝基的 N—O 伸缩振动有关; 1 445~1 521 cm⁻¹与烷烃的 C—H 弯曲振动有关; 1 405~1 410 cm⁻¹与酸的 O—H 变形振动有 关; 1 078~1 082 cm⁻¹与醚、酯等含氧化合物的 C—O 伸缩 振动有关可能为糖类、蛋白质; 1 032~1 036 cm⁻¹与醇、酚 的 C—O 伸缩振动有关可能为寡糖、蛋白质; 873~776 cm⁻¹ 与多糖结构有关^[8,13]。





(c):8个不同产地菌柄的中红外平均光谱图;(d):8个不同产地菌盖的中红外平均光谱图

Fig. 1 Near-infrared and mid-infrared average spectra of Boletus edulis from different sampling places

(a): Near-infrared average spectra of stipes from 8 different origins;(b): Near-infrared average spectra of caps 8 different origins;(c): Mid-infrared average spectra of stipes from 8 different origins;(d): Mid-infrared average spectra of caps from 8 different origins;

不同地区美味牛肝菌的光谱之间有相同趋势,且峰出现 的位置也相近,代表不同产地间美味牛肝菌所含化学成分相 似,但吸光度差异明显,代表不同产地之间化学成分含量不同。中红外光谱的特征峰数量远多于近红外光谱的特征峰数量

1279

量,且吸光度值也更大,表明中红外光谱能更好的表征美味 牛肝菌化学信息。从图1(a)和(c)比较(b)和(d)可以看出菌 柄的吸光度高于菌盖的吸光度,代表菌柄积累的化学物质更 多。杨天伟等利用聚类分析发现美味牛肝菌的菌柄和菌盖之 间有明显差异^[14],从光谱图证明了这一点。

2.2 单一光谱分析

表 2 为所建模型的主要参数。其中,N-b 代表 141× 3 112(141 个样本×3 112 个变量)的数据矩阵,提取前 10 个 潜在变量;有第二高的训练集正确率为 80.4%和第三高的预 测集正确率为 75%。N-g 代表 141×3 112 的数据矩阵,提取 前 6 个 LV,有最低的训练集正确率 66%和最低的预测集正 确率 70.5%。M-b 代表 141×1 867(141 个样本×1 867 个变 量)的数据矩阵,提取前 7 个 LV,有第三高的训练集正确率 76.3%和第二高的预测集正确率 79.6%。M-g 代表 141× 1 867的数据矩阵,提取前 13 个 LV,有最高的训练集正确率 91.8%和最高的预测集正确率 88.6%。

表 2 中 4 个单一光谱(N-b, N-g, M-b, M-g)所建立的模型的预测集正确率在 70.5%~88.6%之间。预测效果优劣依次为 M-g(88.6%), M-b(79.6%), N-b(75%), N-g(70.5%)。中红外光谱的预测效果优于近红外光谱的预测效果,表明中红外光谱相对于近红外光谱有更好的预测能力。 其原因可能是近红外光谱的吸光度弱且谱带重叠严重,中红外光谱相对于近红外光谱有更高的灵敏度,能表征更多的样本化学信息。

表 2 单一光谱模型主要参数

Table	2	The	major	parameters	of	single	spectral	model
-------	---	-----	-------	------------	----	--------	----------	-------

单一	L Va		训练集		预测集			
光谱	LVS	正确率	NER	RMSECV	正确率	NER	RMSEP	
N-b	10	0.804	0.75	0.24	0.75	0.831	0.22	
N-g	6	0.66	0.762	0.25	0.705	0.763	0.23	
M-b	7	0.763	0.808	0.26	0.796	0.799	0.23	
M-g	13	0.918	0.921	0.25	0.886	0.897	0.2	

2.3 数据融合分析

2.3.1 低级融合

将四个单一光谱矩阵((N-b),(N-g),(M-b)和(M-g)) 进行低级融合形成一个141×9958(141个样本×9958个变 量)的矩阵,其中N-b提供3112个变量,N-g提供3112个 变量,M-b提供1867个变量,M-g提供1867个变量。如表 3所示,变量数为9958,训练集正确率为97.9%,RMSECV 为0.238;预测集正确率为93.2%,RMSEP为0.188。相较 于单一光谱,低级融合策略有更多的信息量,对产地的预测 正确率更高。 232 中级融。

2.3.2 中级融合

提取四个单一光谱((N-b),(N-g),(M-b),(M-g))的 LV形成一个141×49(141个样本×49个变量)的矩阵,其 中N-b提供13个LV,N-g提供15个LV,M-b提供8个 LV,M-g提供13个LV。如表3所示,变量数为49,训练集 正确率为100%,RMSECV为0.193预测集正确率为100%, RMSEP为0.133。相较于单一光谱和低级融合策略,中级融 合策略效果更好。低级融合不仅融合了有效信息,还融合了 很多干扰信息。而中级融合策略在提取特征值的过程中去除 了无效信息,不仅降低运算成本,而且提高了产地预测正确 率。

2.3.3 高级融合

提取四个单一光谱((N-b), (N-g), (M-b)和(M-g))的 LV 进行高级融合,如表 3 所示,预测集正确率为 97.5%。 表 4 为高级数据融合统计表。其中,总共 141 组样品中有 55 组样品需要进行高级融合,其中有2组分类错误(88,90),3 组分类歧义(91,92,114),49组分类正确。表为5部分样品 高级融合结果表。其中, 37 号样品被 N-b, N-g 错误分类为 class1, M-b, M-g 正确分类为 class3, 经高级融合后正确分 类到 class3, 表明高级融合有很好的纠错效果, 可以有效提 高产地鉴别效果。92 号样品被 N-b 错误分类为 class1, N-g, M-b, M-g 正确分类为 class7, 经高级融合后分类为 class1 和 class7,表明基于 LV 的高级融合策略受错误分类结果影响, 混淆了最终决策结果。高级融合策略相较于单一光谱和低级 融合策略有更高的产地预测正确率,其原因可能是提取特征 值的过程中去除了无效信息。高级融合策略汲取了中级融合 策略的优点,再利用"模糊集合论"进行决策,理论上高级融 合策略应当取得比中级融合策略更好的产地鉴别效果,但在 本研究过程中,有3.5%的模型受错误分类结果影响了最终 决策结果,其原因可能是高级融合策略不擅长处理异常值。

2.4 数据分析

7个数据矩阵(N-b),(N-g),(M-b),(M-g),(Low-level),(Mid-level),(High-level)构建的模型预测正确率依次 为 88.6%,79.6%,75%,70.5%,93.2%,100%和 97.5%。三种数据融合策略模型预测集正确率均高于单一光 谱模型预测集正确率,结果表明数据融合策略可以提高美味 牛肝菌产地鉴别模型分类效果。比较三种数据融合策略 ((Low-level)),(Mid-level),(High-level)),基于特征值LV的中级融合策 略在产地鉴别效果上优势最明显,用最少的变量数(49)取得

表 3 数据融合主要参数 Table 3 Major parameters of data fusion

			-	-					
融合方式	亦昌粉	· · · · · · · · · · · · · · · · · · ·				预测集			
	受重数 -	正确率	NER	RMSECV	正确率	NER	RMSEP		
Low-level	9 958	0.979	0.989	0.238	0.932	0.971	0.188		
Mid-level	49	1	1	0.193	1	1	0.133		

光谱学与光谱分析

0.97

0.975

High-level

—

_

第 40 卷

—

		Table 4	The result	statistics of	of high-leve	l data fusio	n		
True Origin	N-b	N-g	M-b	M-g	Min	Max	Avg	Prod	Result
1, Class1	1	3	3	1	1	1	1	1	Class1
4, Class1	1	3	3	1	1	1	1	1	Class1
6, Class1	1	1	4	1	1	1	1	1	Class1
8, Class1	1	3	1	3	1	1	1	1	Class1
12, Class1	1	3	1	1	1	1	1	1	Class1
14, Class1	1	3	3	1	1	1	1	1	Class1
16, Class1	3	1	1	1	1	1	1	1	Class1
21, Class1	1	1	3	1	1	1	1	1	Class1
24, Class1	1	1	3	1	1	1	1	1	Class1
30, Class2	2	2	3	2	2	2	2	2	Class2
35, Class2	2	1	2	2	2	2	2	2	Class2
37, Class3	1	1	1	3	1	3	3	3	Class3
39, Class3	4	3	3	3	3	3	3	3	Class3
44, Class3	3	3	5	3	3	3	3	3	Class3
54, Class3	3	3	3	2	3	3	3	3	Class3
65, Class4	1	4	4	4	4	4	4	4	Class4
70, Class5	5	5	3	5	5	5	5	5	Class5
71, Class5	5	5	3	5	5	5	5	5	Class5
73, Class5	5	3	5	5	5	3	5	5	Class5
77, Class5	5	5	5	4	5	5	5	5	Class5
78, Class5	5	5	1	5	5	5	5	5	Class5
82, Class6	6	6	3	6	6	6	6	6	Class6
83, Class6	6	6	2	6	6	6	6	6	Class6
84, Class6	6	6	2	6	6	6	6	6	Class6
85, Class6	6	6	2	6	6	6	6	6	Class6
86, Class6	6	6	3	6	6	6	6	6	Class6
87, Class6	6	6	2	6	6	6	6	6	Class6
88, Class7	1	7	7	5	1	1	1	1	Class1
89, Class7	1	7	7	5	7	7	7	7	Class7
90, Class7	1	7	7	5	1	7	1	1	Class1
91, Class7	1	7	7	5	1	1	7	7	Class1, 7
92, Class7	1	7	7	7	1	1	7	7	Class1, 7
94, Class8	8	8	1	8	8	8	8	8	Class8
95, Class8	8	8	1	8	8	8	8	8	Class8
96, Class8	8	8	1	8	8	8	8	8	Class8
97, Class8	8	8	1	8	8	8	8	8	Class8
99, Class1	5	1	1	1	1	1	1	1	Class1
103, Class1	1	3	1	1	1	1	1	1	Class1
104, Class1	1	3	1	1	1	1	1	1	Class1
105, Class1	1	3	1	1	1	1	1	1	Class1
106, Class1	3	1	1	1	1	1	1	1	Class1
112, Class2	2	2	2	3	2	2	2	2	Class2
114, Class3	3	1	3	3	1	3	3	1	Class1, 3
124, Class3	3	3	1	3	3	3	3	3	Class3
126, Class3	4	4	3	4	4	4	4	4	Class4
127, Class4	4	4	4	5	4	4	4	4	Class4
128, Class4	4	4	4	5	4	4	4	4	Class4
129, Class4	4	4	4	5	4	4	4	4	Class4
130, Class5	5	3 C	1	5	5	5	5	5	Class5
135, Class6	6	6	Z	6	6	6	6	b	Class6
130, Classo	0	0	2	0 C	0	0	0	0 E	Classo
137, Classo	0	0	3	Ø	0	0	0	0	Classo

表 4 高级数据融合结果统计

第	4	期	

138, Class7	1	7	7	7	7	1	7	7	Class7
139, Class7	1	7	7	5	7	1	7	7	Class7
141, Class8	8	8	1	8	8	8	8	8	Class8

True	Origin identification											
Origin	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	- Result			
37, Class3												
N-b	0.449	-0.006	0.412	0.038	-0.098	0.046	0.094	0.064				
N-g	0.294	0.018	0.254	0.017	-0.009	-0.005	0.284	0.147				
M-b	0.233	0.160	0.445	0.094	0.079	0.040	-0.065	0.015				
M-g	0.332	0.042	0.547	0.032	-0.034	-0.063	0.133	0.012				
Min	0.232	-0.006	0.254	0.017	-0.098	-0.063	-0.065	0.012	Class3			
Max	0.449	0.160	0.547	0.094	0.079	0.046	0.284	0.147	Class3			
Avg	0.327	0.054	0.415	0.045	-0.016	0.004	0.112	0.059	Class3			
Prod	0.010	0.000	0.025	0.000	0.000	0.000	0.000	0.000	Class3			
									Class3			
92, Class7												
N-b	0.561	0.106	0.088	0.153	-0.171	0.058	0.162	0.043				
N-g	0.195	0.0540.037	-0.049	0.254	0.096	0.505	-0.091					
M-b	0.204	0.051	-0.113	0.263	-0.057	0.131	0.476	0.045				
M-g	0.194	-0.039	0.216	0.114	0.247	0.012	0.270	-0.014				
Min	0.194	-0.039	-0.113	-0.049	-0.171	0.012	0.162	-0.091	Class1			
Max	0.561	0.106	0.216	0.263	0.254	0.131	0.505	0.045	Class1			
Avg	0.288	0.043	0.057	0.120	0.068	0.074	0.353	-0.004	Class7			
Prod	0.004	0.000	0.000	0.000	0.001	0.000	0.011	0.000	Class7			
									Class1, 7			

表 5 部分样品高级融合结果

Table 5 The results of high-level fusion of some samples

最高的产地鉴别正确率(100%)。其次是高级融合策略,其 建立的 PLS-DA 模型产地鉴别正确率为 97.5%,有少量异常 数据无法鉴别。最后是低级融合策略模型有最多的信息量 (9958)但受干扰因素影响其产地鉴别正确率仅为 93.2%。与 姚森等鉴别绒柄牛肝菌产地溯源研究结果相似^[15]。

3 结 论

探讨了美味牛肝菌不同部位近红外光谱和中红外光谱及 三种数据融合策略(低、中、高)对产地溯源的可行性。结果 表明:(1)近红外和中红外光谱均能鉴别美味牛肝菌产地。

References

(2)中红外光谱所建立的模型优于近红外光谱所建立的模型。
(3)三种融合策略均可提高美味牛肝菌的产地鉴定效果,产地鉴别效果优劣依次为中级融合、高级融合、低级融合、单一光谱。

通过融合近红外和中红外光谱结合 PLS-DA 进行基于特 征值 LV 的中级融合策略,建立不同产地美味牛肝菌鉴别模 型,有最少的变量数(49),最高的产地训练集正确率 (100%),最高的产地预测集正确率(100%),最低的 RM-SEP(0.133),实现了美味牛肝菌产地的快速、准确鉴别,可 以作为美味牛肝菌产地溯源的一种可靠方法。

- [1] Lu Y, Song S, Wang R, et al. Environment International, 2015, 77: 5.
- [2] Cei L, Stefani G, Defrancesco E, et al. Land Use Policy, 2018, 75: 620.
- [3] Wang X, Zhang J, Wu L, et al. Food Chemistry, 2014, 151: 279.
- [4] Falandysz J, Saba M, Liu H, et al. Environmental Science and Pollution Research, 2016, 23(23): 23730.
- [5] MA Zhang-zhong, XU Jin-hua, LUO Zhang, et al(马长中,徐锦华,罗章,等). The Food Industry(食品工业), 2017, 38(3): 311.
- [6] GUI Ming-ying, XU Jun, ZHANG Xin, et al(桂明英,徐 俊,张 鑫,等). The Food Industry(食品工业), 2015, 36(1): 196.
- [7] Quelal-Vásconez M A, Lerma-García M J, Pérez-Esteve É, et al. Food Control, 2019, 99: 68.
- [8] YANG Tian-wei, LI Tao, LI Jie-qing, et al(杨天伟,李 涛,李杰庆,等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2016, 36(7): 2117.
- [9] Li Y, Xiong Y, Min S. Vibrational Spectroscopy, 2019, 101: 20.

- [10] Li Y, Zhang J, Wang Y. Analytical and Bioanalytical Chemistry, 2018, 410(1): 91.
- [11] Márquez C, López M I, Ruisánchez I, et al. Talanta, 2016, 161: 80.
- [12] Ribeiro J S, Ferreira M M C, Salva T J G. Talanta, 2011, 83(5): 1352.
- [13] MENG Ling-zhi, GONG Shu-ling (孟令芝, 龚淑玲). Organic Spectral Analysis(有机波谱分析). 2nd ed(第2版). Wuhan: Wuhan University Press(武汉:武汉大学出版社), 2003. 213.
- [14] YANG Tian-wei, LIU Hong-gao, ZHANG Ji, et al(杨天伟, 刘鸿高,张 霁, 等). Hubei Agricultural Science(湖北农业科学), 2016, 55(9): 2362.
- [15] YAO Sen, LI Tao, LIU Hong-Gao, et al(姚 森,李 涛,刘鸿高,等). Food Science(食品科学), 2018, 39(8): 212.

The Origin Identification Study of *Boletus Edulis* Based on the Infrared Spctrum Data Fusion Strategy

HU Yi-ran¹, LI Jie-qing¹, LIU Hong-gao², FAN Mao-pan^{1*}, WANG Yuan-zhong^{3*}

- 1. College of Resources and Environment, Yunnan Agricultural University, Kunming 650201, China
- 2. College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming 650201, China
- 3. Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming 650200, China

Abstract In recent years, food safety problems happened frequently, and consumers pay more and more attention to the environmental safety of food origin, which leads to an increase in demand for geographical indication products. As a healthy food, the quality of Boletus edulis is greatly affected by the environment of its origin. In order to protect consumers' health and prevent fake and inferior products from entering the market, it is urgent to develop an efficient and low-cost identification technology of the origin of delicious Boletus edulis. Data fusion strategy and partial least squares discrimination (PLS-DA) model were used to identify the origin of Boletus edulis. In this paper, Fourier transform near infrared and Fourier transform middle infrared spectra of 141 samples from 8 Origin (Kunming, Chuxiong, Yuxi, Diqing, Dali, Baoshan, Wenshan and Qujing) were scanned. Kennard-stone algorithm was used to divide all samples into 2/3 training set and 1/3 prediction set. Three fusion strategies (lowlevel, mid-level, high-level) were used to analyze four single spectral matrices spectra: near-infrared average spectra of stipes (N-b), near-infrared average spectra of caps (N-g), mid-infrared average spectra of stipes (M-b), mid-infrared average spectra of caps (M-g) and to establish a partial least squares discriminant (PLS-DA) model. In which root mean square error of cross validation (RMSECV) and the root mean square prediction error (RMSEP) are used to evaluate model stability. The purpose of the non-error ratio (NER), training set classification accuracy and forecast set classification accuracy evaluation model classification performance. It contributes to find out the best way to geographic origin identification of Boletus edulis. The results showed that: (1) near infrared and middle infrared spectra can identify the origin of Boletus edulis; (2) the model established by middle infrared spectrum is better than that in near infrared spectrum; (3) all the three fusion strategies can improve the identification effect of origin of Boletus edulis, and the identification results of producing area from good to bad are in order of mid fusion, high fusion, low fusion and single spectral model. By using PLS-DA intermediate fusion strategy to fuse in near infrared and Mid-infrared spectrum, different origin Boletus edulis identification models are established, with the least number of variables (49), the highest accuracy of training set in producing area (100%), the highest accuracy of prediction set of origin (100%), the lowest RMSEP (0.133). As a reliable method, it can identify the geographical origin of Boletus edulis fast and accurately.

Keywords Boletus edulis; Geographic origin identification; Data fusion; Fourier transform mid-infrared spectrum; Fourier transform near infrared spectrum

(Received Aug. 15, 2019; accepted Dec. 22, 2019)

* Corresponding authors