第40卷，第5期
2020年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 40, No. 5, pp1495-1502
May, 2020

红外光谱的随机森林算法与数据融合策略对绒柄牛肝菌产地鉴别

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 绒柄牛肝菌(*Boletus tomentipes* Earle)是一种健康食品，受广大消费者的青睐，其子实体营养物质积累量受生长环境(海拔、气候等)影响，不同产地间营养物质含量差异显著，为去劣存优，急需建立一种准确、快速、廉价的产地鉴别技术。采用数据融合策略结合随机森林算法(RF)对绒柄牛肝菌的产地进行鉴别，比较了多种特征值提取方法对RF模型分类效果的影响。扫描来自4个产地(北亚热带、北温带、南亚热带、中亚热带)87个样品不同部位的傅里叶变换近红外光谱和傅里叶变换中红外光谱，分析其光谱特征。通过Kennard-Stone算法将所有样品划分为2/3的训练集(58)和1/3的验证集(29)，基于4种红外光谱(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))与三种数据融合策略(低级融合、中级融合、高级融合)的数据，结合RF建立产地鉴别模型，比较了不同方法提取的特征值(投影重要性指标值、Boruta、潜在变量)对模型分类效果的影响。其中，根据袋外错误率(oob)选择最优ntree和mtry；以特异性、灵敏度、训练集正确率和验证集正确率评价模型分类性能，综合多种评价指标，找出绒柄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反映不同产地绒柄牛肝菌间存在的细微差异。(2)单一光谱结合RF建立判别模型效果不理想。(3)三种融合策略均可提高绒柄牛肝菌的产地鉴定效果，产地鉴别效果优劣依次为高级融合、中级融合、低级融合。通过扫描绒柄牛肝菌近红外和中红外光谱，采用基于特征值LV的高级融合策略，结合RF建立不同产地绒柄牛肝菌鉴别模型，有高验证集正确率(99.6%)，高灵敏度(0.969)，高特异性(0.986)，实现了绒柄牛肝菌产地的准确、快速、廉价鉴别，可以作为绒柄牛肝菌产地溯源的一种可靠方法。

关键词 绒柄牛肝菌；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)05-1495-08

引 言

绒柄牛肝菌(*Boletus tomentipes* Earle)隶属于牛肝菌科(Boletaceae)，又名黑牛肝、毛脚牛肝菌，是中国西南地区常见的野生食用牛肝菌，其子实体富含人体必需的蛋白质、维生素及矿质元素等营养元素，是一种健康食品[1]。野生食用菌子实体中营养元素的积累量，受海拔、温度、降水等影响巨大[2]。云南地形多样，气候复杂，野生牛肝菌资源丰富，是绒柄牛肝菌的主产区之一，但不同地区间生长环境差异大，导致各地绒柄牛肝菌品质优劣不一。鲁永新等[3]的研究表明云南各地野生食用菌生长环境差异显著，且不同产地间差异显著。Falandysz等[4]的研究也表明不同产地间绒柄牛肝菌的品质差异显著。为防止劣质产地的绒柄牛肝菌流入市场，确保优质产地的绒柄牛肝菌不被混淆，促进野生食用牛肝菌市场稳健发展，急需建立一种准确、快速、廉价的绒柄牛肝菌产地鉴别技术。

传统形态分类学鉴别技术，鉴别准确率低、受主观影响大；现代分子生物学鉴别技术，虽然分类准确，但成本昂贵、操作复杂、样本损耗大；化学指纹图谱结合化学计量学鉴别产地，因具有准确、快速、廉价的特点而迅速发展。近年来，野生牛肝菌产地鉴别以单一化学指纹图谱为主如中红外光谱法、紫外光谱法[5]、高效液相色谱法[6]、电感耦合等离子体-原子发射光谱法[7]等，然而野生牛肝菌化学组成复杂，单一

收稿日期：2019-10-08，**修订日期**：2020-01-20

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

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 e-mail：huyiran94@126.com

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

指纹图谱无法完全表征样品化学信息。现阶段，利用数据融合策略串联不同仪器取得了相较于单一光谱更精准的鉴别效果，成为食品质量控制领域的热门研究方向。近红外与中红外光谱由于波段不同，反映的化学信息也不同，可以起到互补作用，更全面的表征样品的化学信息。如 Li 等[8]在三七产地鉴别研究中，扫描三七粉的近红外和中红外光谱，结合随机森林建立鉴别模型，研究表明，利用高级融合策略与中级融合策略有效提高模型分类性能，验证集正确率均达 100%。Wang 等[9]在石斛种类鉴别研究中，扫描石斛粉的近红外和中红外光谱，结合偏最小二乘判别、支持向量机、随机森林建立判别模型，结果表明，初级融合策略有效提高模型分类性能，验证集正确率达 100%。

迄今为止，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系统报告。本研究挖掘不同部位绒柄牛肝菌近红外和中红外的光谱信息，结合随机森林建立判别模型，鉴别 4 个产地的绒柄牛肝菌，根据分类效果选出绒柄牛肝菌产地鉴别方法，为野生常见食用牛肝菌鉴别和质量控制提供参考。

1 实验部分

1.1 材料

87 份绒柄牛肝菌采自云南 4 个气候带，分别为北亚热带、北温带、南亚热带、中亚热带(图 1，表 1)均由云南农业大学刘鸿高教授鉴定。样品采集后用纯水清洁表面，置于 50 ℃烘箱烘干至恒重，研磨成粉过 80 目标准筛盘，分别保存于聚氯乙烯自封袋中，储存于干燥避光处。

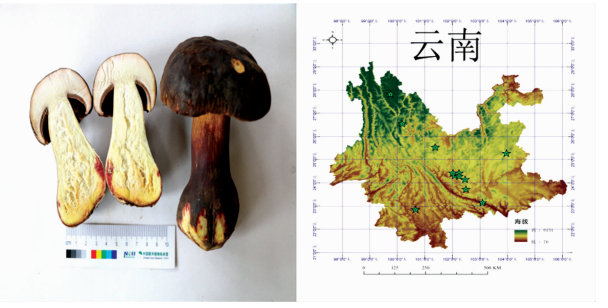

图 1 绒柄牛肝菌地理位置

Fig. 1 The geographic location of *Boletus tomentipes*

表 1 绒柄牛肝菌产地信息

Table 1 The specific geographical origin information of *Boletus tomentipes*

气候带	样本数	经纬度	海拔/m
北亚热带	23		
云南省楚雄市姚安县前场镇	7	E101°37′66.52″，N25°53′83.84″	2 388
云南省楚雄市南华县沙桥镇	4	E 102°24′34.41″，N24°42′63.91″	2 039
云南省曲靖市麒麟区桂花树	5	E 103°96′24.32″，N 25°27′89.52″	2 033
云南省大理市鹤庆州	7	E 100°17′64.44″，N 26°56′17.56″	2 236
北温带	8		
云南省迪庆州香格里拉县	8	E 99°7′52.17″，N 27.83′61.52″	3 281
南亚热带	22		
云南省普洱市南邦河村	7	E 100°65′89.33″，N 22°85′55.64″	958
云南省个旧市乍甸镇	7	E 103°09′56.37″，N 23°15′09.14″	1 430
云南省红河州石屏县	8	E 102°49′39.72″，N 23°72′6.9″	1 430
中亚热带	34		
云南省玉溪市峨山县富良棚乡	8	E 102°08′42.52″，N 24°31′65.79″	2 080
云南省玉溪市峨山县小街镇	9	E 102°46′32.82″，N 24°15′77.72″	1 600
云南省玉溪市易门县铜厂乡	10	E 102°02′01.93″，N 24°43′13.55″	2 187
云南省玉溪市峨山县岔河乡	7	E 102°24′31.92″，N 24°28′78.73″	1 450

1.2 仪器与试剂

UPT-I-10L 型超纯水处理器(四川成都优越科技有限公司)；101A-1 型电热鼓风恒温干燥箱(上海崇明实验仪器厂)；AR1140 型电子分析天平(上海升隆电子科技有限公司)；Antaris Ⅱ型傅里叶变换近红外光谱仪(Thermo Fisher 公司，USA)，配置漫反射模块；Frontier 傅里叶变换中红外光谱仪(Perkin Elmer 公司，USA)；FW-100 型高速粉碎机(浙江华鑫仪器厂)；YP-2 型压片机(上海山岳科学仪器有限公司)；80 目标准筛盘(浙江绍兴道墟五四仪器厂)；分析纯级溴化钾(天津风船化工科技有限公司)。

1.3 光谱信息采集

1.3.1 中红外指纹图谱采集

中红外光谱是由 Frontier 型傅里叶变换红外光谱仪采集。取(1.5±0.2) mg 绒柄牛肝菌样品和(150±20) mg 溴化钾粉末在玛瑙研钵中磨细混匀，再将细粉倒入压片机中制成薄片，扫描。扫描波数范围 4 000～400 cm^{-1}，分辨率 4 cm^{-1}，信号累计扫描次数 16 次。每个样本重复扫描 3 次，取平均光谱。

1.3.2 近红外指纹图谱采集

近红外光谱由 Antaris Ⅱ型傅里叶变换近红外光谱仪用漫反射显微镜采集。称取 20 g 绒柄牛肝菌粉末，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}，信号累计扫描次数 64 次。每个样本重复扫描 3 次，取平均光谱。

1.4　随机森林原理及评价

随机森林(random forest，RF)是一种基于决策树和自助采样法的集成学习方法。本研究在 RStudio(3.5.3)中使用 randomForest 包构建 RF 模型，原理如图 2 所示，具体步骤如下：(1) 总样本数为 Y，利用自助采样法提取 y(约 2/3 的 Y)构建决策树；(2) 每个样本有 M 个变量，随机取其中 m 个样本；(3) 重复(1)和(2)过程 n 次，建立 n 棵决策树；(4) 每棵决策树自由生长产生一个决策结果，n 棵决策树进行投票，分类结果取决于 RF 中所有决策树的多数表决。

其中，RF 模型分类性能取决于 2 个参数的选择，mtry(m)是随机变量数，决定单棵树的分类性能，ntree(n)是决策树的数量，它决定了 RF 的规模。m 和 n 越大，决策树之间的相关性和分类能力随之增强，模型过拟合风险增加；m 和 n 越小，决策树之间的相关性和分类能力随之减弱，模型欠拟合风险增加，因此计算袋外错误率(out-of-bag error，OOB)对模型进行内部评估，选出最优 m 和 n。根据不同 ntree 数的 oob 图，选择 OOB 低且稳定的区间内任意一点为最优 ntree。为避免默认 $m(\sqrt{M})$陷入局部最优解，因此在$\sqrt{M}\pm10$ 范围内搜索最优 mtry[8]。

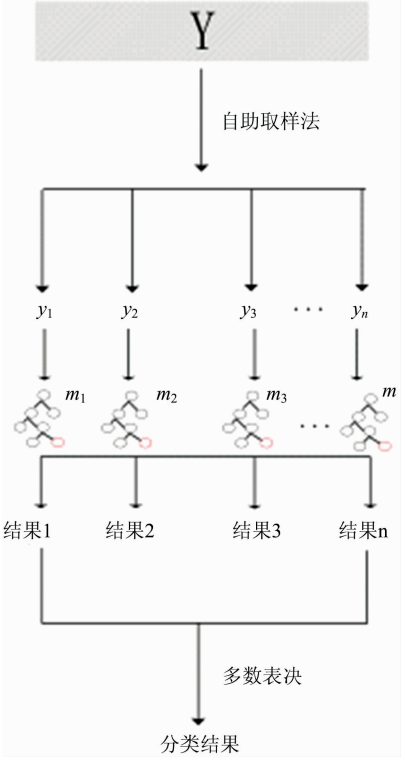

图 2　随机森林原理

Fig. 2　Schematic of random forest

1.5　提取特征值

红外光谱表征样品化学信息全面，但也带来维数诅咒的问题，同时红外光谱中含有大量的噪声和干扰变量，使得其预测性能不可靠。因此，要得到一个拟合良好的模型，筛选特征变量是一种有效方法。本工作使用的 3 种特征变量筛选方法在食品领域鉴别研究中已有广泛应用：

投影重要性指标值(variable importance in projection，VIP)表示自变量对模型拟合的重要性，VIP 值越高，波长点对标签的解释能力越强[10]。根据 VIP 用 10 折交叉验证对各波长点进行迭代筛选，选出有效波长点作为特征变量。Boruta 算法是围绕 RF 算法构建的包装器，通过创建混合副本，重新排列原始特征，使每个波长点有对应的阴影特征，比较真实样本与最佳阴影特征的排列精度重要性，将每个变量划分为确定、暂定、拒绝这 3 个标签[11]。该特征提取方法，可以评估所有波长点的重要性，去除负面变量，得到一个最小最优的特征子集，提高模型分类性能。提取标签为确定、暂定的波长点作为特征变量。潜在变量(latent variable，LV)类似于作成分，基于偏最小二乘关联算法将数据正交变换为互不线性相关的多组 LV，提取对数据解释能力强的 LV 代替原始数据。根据 Q2(累计预测能力)第一次到达最大值时的因子数确定提取 LV 个数。

1.6　数据融合策略

数据融合分为 3 个层次低级融合，中级融合，高级融合。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，再建立鉴别模型；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，再建立鉴别模型；高级融合又名决策级融合，提取各指纹图谱特征值建立判别模型获得分类结果，再根据一定准则对各分类结果进行融合，最终得到整体一致的决策。本工作根据中级融合分类结果选出最合适的特征值，再基于“模糊集合论”，把各独立模型的模糊现象(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。

2　结果与讨论

2.1　光谱指纹图谱分析

图 3 为不同产地绒柄牛肝菌的近红外和中红外平均光谱图，从图 3(a)和(b)可以看出近红外光谱有 6 处特征峰。其中，8 240～8 530 cm^{-1}附近与 C—H，N—H，O—H 的二倍频峰有关；5 775～5 785 cm^{-1}附近与 C—H 的基频峰有关；5 160 cm^{-1}附近可能与水、蛋白质 C═O 的二倍频峰有关；4 770～4 774 cm^{-1}附近与 C═O 和 O—H 组合带的基频峰有关；4 596 cm^{-1}附近与 N—H 的合频峰有关；4 337～4 346 cm^{-1}是关于—CH_2 多糖的反对称伸缩振动[9]。

从图 3(c)和(d)可以看出中红外光谱有 8 处特征峰。其中，3 310～3 390 cm^{-1}附近的可能与糖、纤维素 O—H 和 N—H 的伸缩振动有关；2 922～2 929 cm^{-1}附近与 N—H 和 C—H 的伸缩振动有关；1 636 cm^{-1}附近与酰胺Ⅰ带和酰胺Ⅱ带的 C═O 伸缩振动有关；1 382～1 384 cm^{-1}附近与酸类的 O—H 变形振动有关；1 083～1 090 cm^{-1}附近与酯类、醚类等含氧化合物的 C—O 伸缩振动有关，可能为糖类、蛋白质；1 038～1 041 cm^{-1}附近与酚类、醇类的 C—O 伸缩振

动有关，可能为寡糖、蛋白质；883～891 cm^{-1}附近与多糖结构有关[5, 12]。

不同地区绒柄牛肝菌间有相似的峰位、峰型，代表不同产地间绒柄牛肝菌所含化学成分相似，但吸光度差异明显，代表不同产地间化学成分含量不同。图 3(a)和(c)与(b)和(d)比较可以看出绒柄牛肝菌菌柄与菌盖的吸光度差异不明显，代表绒柄牛肝菌菌柄和菌盖积累的化学物质相当。从光谱图中可以反映样品间存在细微差异，但仅靠光谱图无法实现产地的精准鉴别，因此需进一步结合化学计量学鉴别产地。

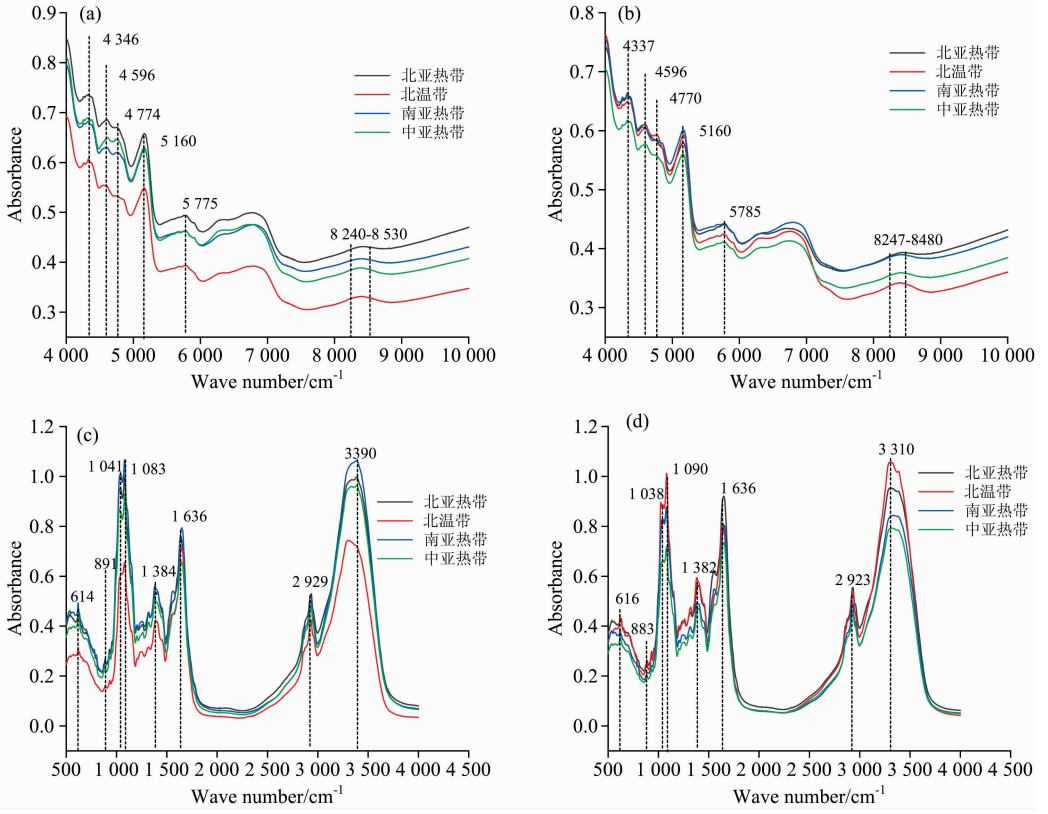

图 3 不同产地绒柄牛肝菌的近红外和中红外平均光谱

(a)：菌柄的近红外平均光谱；(b)：菌盖的近红外平均光谱；

(c)：菌柄的中红外平均光谱；(d)：菌盖的近红外平均光谱

Fig. 3 Near-infrared and mid-infrared average spectra of *Boletus tomentipes* from different sampling places

(a): Near-infrared awerage spectra of stipes; (b): Near-infrared average spectra of caps;

(c): Mid-infrared average spectra of stipes; (d): Mid-infrared average spectra of caps

2.2 单一光谱分析

使用 Kennard-Stone 算法将数据集(87)分为 2/3 的训练集(58)和 1/3 的验证集(29)。如图 4 随机森林参数选择图所示，根据 OOB 选出 ntree 和 mtry，如表 2 单一光谱模型主要参数图，其中 4 个单一光谱(N-c，N-g，M-b，M-g)所建立的模型，验证集正确率在 72.4%～86.2%之间，预测效果优劣依次为 N-g(86.2%)，N-b(86.1%)，M-b(82.8%)，M-g(72.4%)。近红外光谱的预测效果优于中红外光谱的预测效果，表明近红外光谱相对于中红外光谱在绒柄牛肝菌产地鉴别上有更好的预测能力。但单一光谱模型训练集正确率与验证集正确率之间相差超过 20%，欠拟合风险大，结合 RF 用于对绒柄牛肝菌产地鉴别效果不理想，原因可能是光谱中的噪音影响了模型拟合能力。

2.3 特征值提取

图 5(a)为 Boruta 算法筛选的波数，标签 0 代表拒绝，标签 1 代表暂定，标签 2 代表确定。其中，从 N-b 的 3 112 个变量中筛选出 6 个确定标签，23 个暂定标签；从 N-g 的 3 112 个变量中筛选出 1 个确定标签，28 个暂定标签；从 M-b 的 1 867个变量中筛选出 1 个确定标签，56 个暂定标签；从 N-b 的 3 112 变量中筛选出 1 个确定标签，31 个暂定标签。图 5(b)为根据 VIP 排列的变量，迭代 10 次后进行交叉验证的错误率，当交叉验证错误率最低时，其变量数为最优变量数。其中，筛选 N-b 的前 22 个变量为最优变量数；筛选 N-g 的前 92 个变量为最优变量数；筛选 M-b 的前 427 个变量为最优变量数；筛选 M-g 的前 247 个变量为最优变量数。图 5(c)为根据 Q2 确定最优 LV 数，当 Q2 第一次到达最大值或趋于稳定时，其 LV 数为最优 LV 数。其中，N-b 的 LV 数在11时 Q2 趋于稳定；N-g 的 LV 数在 10 时 Q2 达到最大；N-b 的 LV 数在 10 时 Q2 第一次达到最大；N-b 的 LV 数在 12 时 Q2 达到最大。

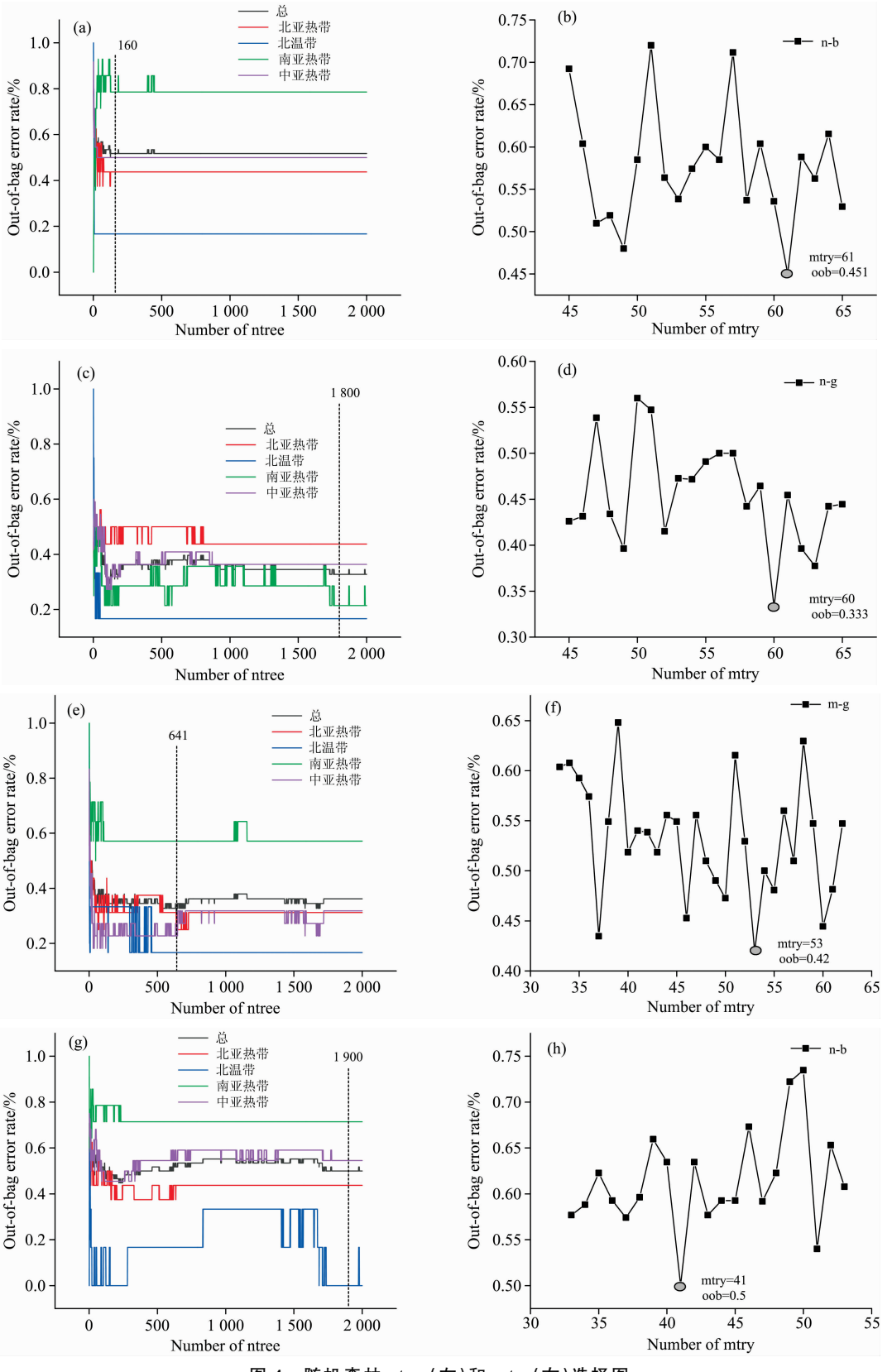

图 4　随机森林 ntree(左)和 mtry(右)选择图

(a)，(b)：N-b；(c)，(d)：N-g；(e)，(f)：M-b；(g)，(h)：N-g

Fig. 4　The selection diagram of random forest ntree (left) and mtry (right)

(a)，(b)：N-b；(c)，(d)：N-g；(e)，(f)：M-b；(g)，(h)：N-g

表 2 单一光谱模型主要参数

Table 2 The major parameters of single spectral model

单一光谱	ntree	mtry	训练集			验证集		
			正确率	灵敏度	特异性	正确率	灵敏度	特异性
N-b	160	61	0.5	0.539	0.752	0.861	0.906	0.95
N-g	1 800	60	0.638	0.669	0.842	0.862	0.876	0.943
M-b	641	53	0.638	0.756	0.927	0.828	0.756	0.927
M-g	2 000	43	0.5	0.576	0.765	0.724	0.762	0.885

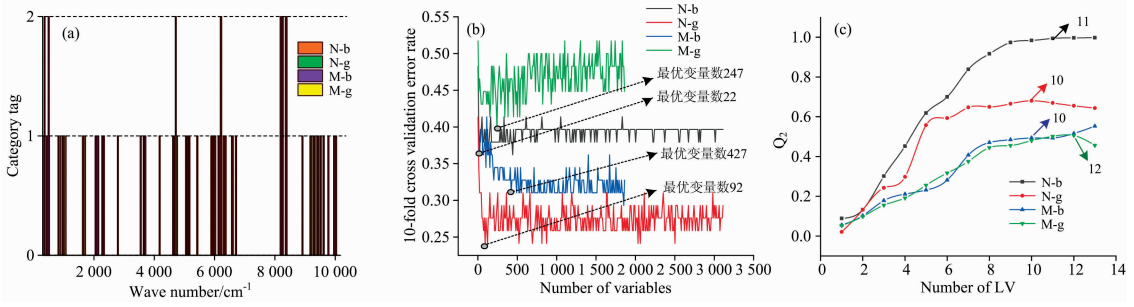

图 5 特征选择图

(a)：Boruta 算法；(b)：VIP；(c)：LV

Fig. 5 Feature selection diagram

(a)：Boruta algorithm；(b)：VIP；(c)：LV

2.4 数据融合分析

2.4.1 低级融合

将四个单一光谱矩阵[(N-b)，(N-g)，(M-b)，(M-g)]进行低级融合形成一个 87 个样本×9 958 个变量的矩阵，其中 N-b 提供 3 112 个变量，N-g 提供 3 112 个变量，M-b 提供 1 867个变量，M-g 提供 1 867 个变量。

2.4.2 中级融合

筛选 VIP 提取四个单一光谱[(N-b)，(N-g)，(M-b)，(M-g)]的特征值形成一个 87 个样本×788 个变量的矩阵，其中 N-b 提供 22 个变量，N-g 提供 92 个变量，M-b 提供 427 个变量，M-g 提供 247 个变量。筛选 Boruta 提取四个单一光谱[(N-b)，(N-g)，(M-b)，(M-g)]的特征值形成一个 87 个样本×147 个变量的矩阵，其中 N-b 提供 29 个变量，N-g 提供 29 个变量，M-b 提供 57 个变量，M-g 提供 32 个变量。提取四个单一光谱[(N-b)，(N-g)，(M-b)，(M-g)]的 LV 形成一个 87 个样本×43 个变量的矩阵，其中 N-b 提供 11 个 LV，N-g 提供 10 个 LV，M-b 提供 10 个 LV，M-g 提供 12 个 LV。

2.4.3 高级融合

基于特征值 LV 进行高级融合。提取四个单一光谱[(N-b)，(N-g)，(M-b)，(M-g)]的 LV 结合 RF 建立鉴别模型，其中，N-b 有 12 个错误、N-g 有 11 个错误、M-b 有 17 个错误、M-g 有 18 个错误，对 4 个模型的分类结果进行决策。总共 87 组样品中有 45 组样品需要进行高级融合，如表 3 所示，其中有 2 组分类错误(6，26)，2 组分类歧义(8，52)，43 组分类正确。其中，6 号样品被 N-g 和 M-b 错误分类为 class4，M-g 错误分类为 class3，N-b 正确分类为 class1，经高

表 3 未正确分类样品高级融合结果

Table 3 The results of high-level fusion of misclassification samples

True Origin	Origin identification				Result	True Origin	Origin identification				Result
	Class1	Class2	Class3	Class4			Class1	Class2	Class3	Class4	
6，Class1						8，Class1					
N-b	**0.504**	0.059	0.098	0.339		N-b	**0.720**	0.053	0.090	0.138	
N-g	0.196	0.007	0.018	**0.778**		N-g	0.250	0.000	0.053	**0.697**	
M-b	0.385	0.039	0.129	**0.447**		M-b	0.376	0.073	0.143	**0.408**	
M-g	0.031	0.000	**0.671**	0.297		M-g	0.299	0.011	0.034	**0.657**	
Min	0.031	0.000	0.018	**0.297**	Class4	Min	**0.250**	0.000	0.034	0.138	Class1
Max	0.504	0.059	0.671	**0.778**	Class4	Max	**0.720**	0.073	0.143	0.697	Class1
Avg	0.279	0.026	0.229	**0.465**	Class4	Avg	0.411	0.034	0.080	**0.475**	Class4
Prod					Class4	Prod					Class4
	0.001	0.000	0.000	**0.035**			0.020	0.000	0.000	**0.026**	
					Class4						**Class1，4**

级融合后，错误分类为 class4。

2.4.4　小结

基于数据融合策略建立鉴别模型的主要参数如表 4 所示，低级融合策略相较于单一光谱使模型表现出更强的拟合能力增强和分类效果，表明近红外结合中红外光谱对分类性能起协同作用；基于 VIP 的中级融合策略相较于单一光谱和低级融合策略模型，数据量小，分类能力提高，但模型拟合能力变弱，欠拟合风险增加，原因可能为其特征变量受异常值影响，导致模型过拟合；基于 Boruta 的中级融合策略相较于单一光谱、低级融合策略和基于 VIP 的中级融合策略，数据量小，模型拟合性能良好，表明该方法可提高模型分类性能；基于 LV 的中级融合策略相较基于 VIP 和 Boruta 的中级融合策略，模型拟合能力优秀，分类性能高，数据量小，原因可能为其特征变量解释样品的大部分信息，充分挖掘样品信息。

高级融合策略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。中级融合策略在提取特征值的过程中去除样品无效信息，不仅降低运算成本，而且增加了有效信息，提高了模型分类性能。高级融合策略汲取了中级融合策略的优点，再加上“模糊集合论”的对分类结果决策，更进一步提高了模型分类性能。研究表明，提取特征值 LV 与数据融合策略组合挖掘绒柄牛肝菌红外光谱信息，可以大幅提高模型分类效果，与 Li[8] 等鉴别三七产地研究结果相似。

表 4　数据融合主要参数

Table 4　Major parameters of data fusion

融合方式	ntree	mtry	训练集			验证集		
			正确率	灵敏度	特异性	正确率	灵敏度	特异性
Low-level	800	107	0.759	0.783	0.9	0.862	0.871	0.941
Mid-level								
(vip)	1 150	38	0.69	0.701	0.866	0.931	0.933	0.971
(boruta)	900	4	0.81	0.802	0.923	0.862	0.871	0.941
(lv)	1 500	30	0.948	0.951	0.979	0.966	0.964	0.985
High-level	\	\	\	\	\	**0.966**	**0.969**	**0.986**
(N-b)	1 000	3	0.914	0.919	0.967	0.813	0.899	0.856
(N-g)	1 500	5	0.81	0.823	0.923	1	1	1
(M-b)	1 800	2	0.741	0.689	0.889	0.931	0.943	0.974
(M-g)	1 300	12	0.828	0.852	0.933	0.724	0.756	0.878

3　结　论

研究了绒柄牛肝菌不同部位近红外光谱和中红外光谱及数据挖掘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反映不同产地绒柄牛肝菌间的微小差异；(2)单一光谱结合 RF 建立判别模型不理想，平均正确率仅 81.9%；(3)三种数据融合策略均可提高绒柄牛肝菌的产地鉴定效果，产地鉴别效果优劣依次为高级融合、中级融合、低级融合。

通过扫描绒柄牛肝菌近红外和中红外光谱，使用基于特征变量 LV 的高级融合策略，结合 RF 建立不同产地绒柄牛肝菌鉴别模型，有高产地验证集正确率(99.6%)，高灵敏度(0.969)，高特异性(0.986)，实现了绒柄牛肝菌产地的准确、快速、廉价鉴别，可以作为绒柄牛肝菌产地鉴别的一种可靠方法。

References

[1] Wang X, Zhang J, Wu L, et al. Food Chemistry, 2014, 151: 279.

[2] YANGBAI Qiu-xiu, CHEN Xun, LIU Xiao-fei(杨白秋秀，陈　旭，刘晓飞). Edible Fungi of China(中国食用菌), 2017, 36(5): 13.

[3] LU Yong-xin, TIAN Hou-ming, YANG Hai-shu, et al(鲁永新，田侯明，杨海抒，等). Chinese Journal of Eco-Agriculture(中国生态农业学报), 2015, 23(6): 748.

[4] Falandysz J, Zhang J, Wiejak A, et al. Ecotoxicology and Environmental Safety, 2017, 142: 497.

[5] YANG Tian-wei, CUI Bao-kai, ZHANG Ji, et al(杨天伟，崔宝凯，张　霁，等). Mycosystema(菌物学报), 2014, 33(2): 262.

[6] Chen Y, Yan Y, Xie M, et al. Journal of Pharmaceutical and Biomedical Analysis, 2008, 47(3): 469.

[7] Wang X, Zhang J, Li T, et al. Journal of Analytical Methods in Chemistry, 2015, 2015: http://dx.doi.org/10.1155/2015/165412.

[8] Li Y, Zhang J, Wang Y. Analytical and Bioanalytical Chemistry, 2018, 410(1): 91.

[9] Wang Y, Zuo Z T, Huang H Y, et al. Royal Society Open Science, 2019, 6(5): 190399.

[10] He P, Xu X, Zhang B, et al. Estimation of Leaf Chlorophyll Content in Winter Wheat Using Variable Importance for Projection (VIP) with Hyperspectral Data. Proceedings of SPIE, 2015, 9637: 963708.

[11] CHEN Yi-jie, TANG Jia-shan(陈逸杰，唐加山). Software Guide(软件导刊), 2019, 18(4): 69.

[12] Mellado-Mojica E, López M G. Food Chemistry, 2015, 167: 349.

Infrared Spectral Study on the Origin Identification of *Boletus Tomentipes* Based on the Random Forest Algorithm and Data Fusion Strategy

HU Yi-ran[1], LI Jie-qing[1], LIU Hong-gao[2], 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 *Boletus tomentipes* Earleas a kind of healthy food is favored by the majority of consumers. The nutrient accumulation of the fruiting body is affected by the growth environment (altitude, climate, etc.). There is a significant difference in the content of nutrient between different regionsIt is urgent to establish an accurate, rapid and cheap origin identification technology. In this paper, a data fusion strategy combined with random forest algorithm (RF) was used to identify the origin of *B. tomentipes*, and the effects of various eigenvalue extraction methods on the classification of RF models were compared. Fourier transform near infrared and Fourier transform mid-infrared spectra of 87 samples from 4 producing areas (north subtropics, north temperate zones, south subtropical zones and middle subtropical zones) were scanned to analyze their spectral characteristics. All the sampleswere divided into two thirds of the training set (58) and a third of the validation set (29) by the kennard-stone algorithm. Based on 4 kinds of infrared 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 three data fusion strategies (low-level fusion strategies, mid-level fusion strategies, high-level fusion strategies) of data, combining with the RF building identification model, the effects of different characteristic value (variable importance in projection, Boruta, latent variables) on the classification results of the model are compared. Among them, the optimal ntree and mtrywere selected according to oob. The classification performance of the model was evaluated with specificity, sensitivity, training set correctness, and validation set accuracy. Finally, the best method to identify the origin of *B. tomentipes* was found by multiple evaluation indicators. The results showed that (1) near infrared and middle infrared spectra could identify the origin of *B. tomentipes*. (2) It is not ideal for establish a discriminant model with a single spectrum combined with RF. (3) All three fusion strategies can improve the origin identification effect of *B. tomentipes*. Theresults of origin identification from good to bad are in order of high-level fusion, mid-level fusion, low-level fusion. By scanning the near infrared and middle infrared spectra of *B. tomentipes*, a high-level fusion strategy based on characteristic value LV was adopted, and the identification model of *B. tomentipes* from different regions was established with RF, which has high verification set accuracy (99.6%), high sensitivity (0.969) and high specificity (0.986). As a reliable method, it can identify the geographical origin of *B. tomentipes* quickly and accurately.

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

(Received Oct. 8, 2019; accepted Jan. 20, 2020)

* Corresponding authors