第42卷，第2期
2022年2月

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

Vol.42，No.2，pp549-554
February，2022

# 基于偏最小二乘法判别分析与随机森林算法的牛肝菌种类鉴别

陈凤霞[1]，杨天伟[2]，李杰庆[1]，刘鸿高[3]，范茂攀[1*]，王元忠[4*]

1. 云南农业大学资源与环境学院，云南 昆明 650201
2. 云南省热带作物科学研究所，云南 景洪 666100
3. 云南农业大学农学与生物技术学院，云南 昆明 650201
4. 云南省农业科学院药用植物研究所，云南 昆明 650200

**摘 要** 牛肝菌作为一种著名的野生食用菌，具有较高的食用价值和经济价值。牛肝菌种类繁多，不易区分，建立一种有效、快速、可信的种类鉴别技术，可为牛肝菌提高品质提供一种方法。本研究采集云南不同地区7种野生牛肝菌共计683株，获取样品中红外光谱和紫外光谱，分析不同种类牛肝菌平均光谱图特征。基于多种预处理组合(SNV+SG，2D+MSC+SNV，1D+MSC+SNV+SG，MSC+2D)的单一光谱数据结合两种特征值提取法(PCA，LVs)构建了偏最小二乘法判别分析与随机森林算法并结合数据融合策略对牛肝菌进行种类鉴别，有一定的创新性。结果表明：(1)中红外光谱和紫外光谱的不同种类牛肝菌平均光谱吸收峰差异较小，吸光度具有细微差异。(2)合适的预处理可提高光谱数据信息，偏最小二乘法判别分析和随机森林算法模型的中红外光谱数据和紫外光谱数据最佳预处理组合为2D+MSC+SNV，SNV+SG，2D+MSC+SNV，1D+MSC+SNV+SG。(3)单一光谱模型中，中红外光谱模型优于紫外光谱模型，中红外光谱最佳预处理组合2D+MSC+SNV的偏最小二乘法判别分析模型正确率训练集99.78%，验证集99.12%；随机森林模型正确率训练集93.20%，验证集99%。(4)数据融合策略提高了分类正确率，低级融合的偏最小二乘法判别分析模型训练集和验证集正确率为100%，99.12%。随机森林模型训练集和验证集正确率为92.32%，99.14%。(5)随机森林算法中级数据融合Latent variable(LVs)正确率为训练集92.76%，验证集96.04%；中级数据融合Principal components analysis(CPA)正确率为训练集97.15%，验证集100%。(6)偏最小二乘法判别分析中级数据融合(LVs)正确率为训练集100%，验证集99.56%；中级数据融合(CPA)训练集和验证集正确率均能达到100%。基于偏最小二乘法判别分析和随机森林算法结合数据融合策略对牛肝菌进行种类鉴别，鉴别效果理想。偏最小二乘法判别分析中级数据融合(CPA)可作为一种低成本高效率的牛肝菌种类鉴别技术。

**关键词** 牛肝菌；中红外光谱；紫外光谱；偏最小二乘法判别分析；随机森林；数据融合

**中图分类号**：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)02-0549-06

## 引 言

牛肝菌(Boletus)属担子菌亚门，伞菌目牛肝菌科(Boletaceae)和松塔牛肝菌科(Strobilomycetaceae)统称[1]。中国已知牛肝菌种类390种，其中3/5可食用。牛肝菌是著名的野生食用菌，具有较高的食用价值和药用价值[2]，我国凭借得天独厚的地理条件成为食用菌主要生产和出口大国[3]。野生牛肝菌种类混杂，难以分辨，建立一种快速有效的牛肝菌鉴别方法，可提高牛肝菌品质，保障消费者健康。

数据融合将多源数据进行联合以便获得更多信息[4]。近年来，光谱数据融合结合机器学习进行食用菌鉴别研究较为普遍[5]。胡翼然等[6]采用随机森林结合红外光谱数据融合法成功找到绒柄牛肝菌产地鉴别技术。Yao等[7]的研究表明红外光谱和紫外光谱结合数据融合用于牛肝菌产地及种类区分具有可靠性。光谱指纹图谱技术具有低成本、快捷、易获取等特点，常用于质量检测及食用菌研究[8]。光谱指纹图谱结合数据融合策略是将同一研究物的多个不同来源数据信息用

**收稿日期**：2021-01-04，**修订日期**：2021-02-01

**基金项目**：国家自然科学基金地区项目(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**：陈凤霞，1994年生，云南农业大学资源与环境学院硕士研究生 e-mail：cfx0909@126.com

*通讯作者 e-mail：mpfan@126.com；boletus@126.com

化学计量学方法进行优化重组。现目前，数据融合策略在药品、食品等领域研究较为广泛[9]。Li 等[10]利用偏最小二乘法判别分析和随机森林进行光谱数据融合技术，寻找到灵芝种类鉴别分析法。Yao 等[11]研究了 232 株食用菌，采用支持向量机和随机森林结合数据融合分析鉴别野生食用菌和栽培食用菌，为食用菌分类提供了准确有效的方法。偏最小二乘法判别分析与随机森林基于样本数据建立数学模型，在菌类研究中使用广泛且效果显著。

光谱数据准确性受各种背景影响，如：噪音、基线漂移、随机误差等，进行光谱数据预处理，可提高分辨率和灵敏度。本工作采用预处理后的中红外光谱和紫外光谱结合偏最小二乘法判别分析与随机森林算法对牛肝菌种类进行单一光谱和数据融合数据模型分析，寻找出最优牛肝菌种类鉴别方法，为牛肝菌质量控制和保障消费者食用安全提供借鉴。

## 1 实验部分

### 1.1 样品

采集云南各地 7 种成熟牛肝菌共计 683 份，其中灰褐牛肝菌 98 株、美味牛肝菌 221 株、栗色牛肝菌 110 株、小美牛肝菌 32 株、皱盖疣柄牛肝菌 132 株、双色牛肝菌 30 株、绒柄牛肝菌 60 株，详见图 1 和表 1，均由云南农业大学刘鸿高教授鉴定。样品初始处理用陶瓷刀刮去泥土，自来水清洗干净后用超纯水润洗三遍；晾干表面水分后置于恒温烘箱 50 ℃烘干至恒重。不同种类牛肝菌分类粉碎后使用 100 目标准筛过筛保存。

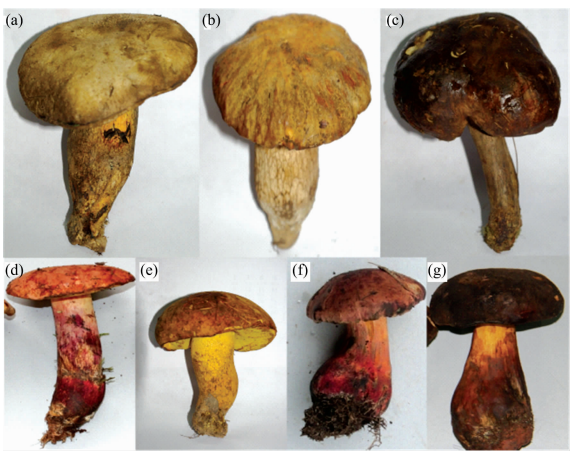



**图 1 牛肝菌样品**

(a)：灰褐牛肝菌；(b)：美味牛肝菌；(c)：栗色牛肝菌；(d)：小美牛肝菌；(e)：皱盖疣柄牛肝菌；(f)：双色牛肝菌；(g)：绒柄牛肝菌

**Fig. 1 Boletus samples**

(a)：*Boletus griseus* Frost；(b)：*Boletus edulis* Bull. : Fr；(c)：*Boletus umbriniporus* Hongo；(d)：*Boletus speciosus* Forst；(e)：*Leccinum rugosicepes* (Perk) Sing；(f)：*Boletaceae bicolor* Peck；(g)：*Boletus tomentipes* Earle

### 1.2 仪器与试剂

Frontier 型傅里叶变换红外光谱仪，美国 Perkin Elmer 公司；TU-1901 紫外-可见分光光度计，日本岛津公司；AR1140 型电子分析天平，上海升隆电子科技有限公司；100 目标准筛盘，浙江上虞市道墟五四仪器厂；FW-100 型高速粉碎机，天津市华鑫仪器厂；UPT-I-10 超纯水机，优谱科技有限公司。分析纯溴化钾，天津风船化工科技有限公司；氯仿，西陇化工股份有限公司。

**表 1 牛肝菌样品信息**

**Table 1 Boletus samples information**

| 种类 | 编号 | 产地 | 数量 |
|---|---|---|---|
| 灰褐牛肝菌 | a | 保山市隆阳区、玉溪市江川区、大理市弥渡县、昆明市晋宁区、曲靖市马龙区、昆明市安宁区 | 98 |
| 美味牛肝菌 | b | 楚雄市南华县、迪庆州香格里拉市、大理市弥渡县、迪庆州维西县、保山市隆阳区、昆明市安宁区、玉溪市红塔区、大理市鹤庆县、文山市东山乡、昆明市石林县 | 221 |
| 栗色牛肝菌 | c | 昆明市石林县、曲靖市马龙区、保山市隆阳区、大理市弥渡县、红河州石屏县、楚雄市元谋县、玉溪市红塔区、红河州个旧市 | 110 |
| 小美牛肝菌 | d | 楚雄市南华县、玉溪市红塔区、保山市隆阳区 | 32 |
| 皱盖疣柄牛肝菌 | e | 曲靖市麒麟区、红河州石屏县、大理市弥渡县、昆明市安宁区、迪庆州维西县、保山市隆阳区、楚雄市元谋县、玉溪市红塔区 | 132 |
| 双色牛肝菌 | f | 楚雄市南华县、大理市苍山、曲靖市麒麟区 | 30 |
| 绒柄牛肝菌 | g | 红河州个旧市、迪庆州香格里拉市、红河州石屏县、大理市鹤庆县 | 60 |

### 1.3 光谱采集

中红外光谱：采用电子分析天平称取牛肝菌样品粉末(1.2±0.2) mg，溴化钾粉末(150±0.5) mg 置于玛瑙研钵中混合研磨成均匀细粉于压片机制成薄片，光谱仪用溴化钾片进行空白扫描，扣除背景影响，每个样本重复扫描三次，信号累计扫描 16 次，范围在 4 000～400 $cm^{-1}$，分辨率为 4 $cm^{-1}$。

紫外光谱：称取牛肝菌样品粉末(150±0.5) mg，将称取好的样品置于 25 mL 试管加入 10 mL 氯仿封口，超声提取 40 min，三层滤纸过滤取清液，光谱仪预热 1 h，扣除背景后进行光谱扫描，采样间隔 0.3 min，每个样本重复三次，取平均光谱。

### 1.4 数据处理

将整理好的光谱数据进行数据预处理，使用 OMNIC 软件进行自动基线校正和吸光度转换，SIMCA 软件进行 Savitzky-Golay(SG)、Standard normal variables(SNV)、Multiple scattering correction(MSC)、First derivative(1D)、Second derivatives(2D)联合预处理，不同的光谱预处理具有不同的作用和优势，结合实际合理选择最佳的预处理方法，将不同预处理组合后的数据用 Kennard-Stone 算法分为 2/3(456)的训练集合 1/3(227)的验证集，用于模型建立。先进

行单一光谱数据建模，寻找到最佳预处理建模结果后，将最佳数据集组合用于建立数据融合模型。将两种单一光谱数据进行联合分析，以增加信息量。研究中使用两种算法(偏最小二乘法判别分析和随机森林)进行牛肝菌样品模型建立，寻找最佳种类分类模型，达到牛肝菌种类鉴别目的。

## 2　结果与讨论

### 2.1　光谱分析

图 2 为中红外和紫外光谱图，图中为 7 种牛肝菌样品的平均光谱。由图 2(a)可知，波数范围在 3 700～2 750 $cm^{-1}$之间为主要强吸收峰波段，不同种类牛肝菌吸收峰差异较小但具有细微差异，表明化学成分大致相同。波数在 3 385 $cm^{-1}$附近与糖 O—H；纤维素 N—H 伸缩振动有关，波数在 2 994～2 927 $cm^{-1}$范围表征炔烃类基团 CH 伸振动。1 637～1 695 $cm^{-1}$与烯烃类 C═C 伸缩振动有关。1 079～1 059 $cm^{-1}$与醚类、酯类等含氧化合物 C—O—C 伸缩振动，主要成分为蛋白质，糖类。846～532 $cm^{-1}$为醇类、酚类 O－H 伸缩振动，表征多糖结构。由图 2(b)可知，牛肝菌紫外光谱有效吸收峰范围在 225～350 $cm^{-1}$，不同种类牛肝菌紫外光谱存在指纹差异性，较强吸收峰在 296～263 $cm^{-1}$，主要组成为蛋白质、氨基酸类。

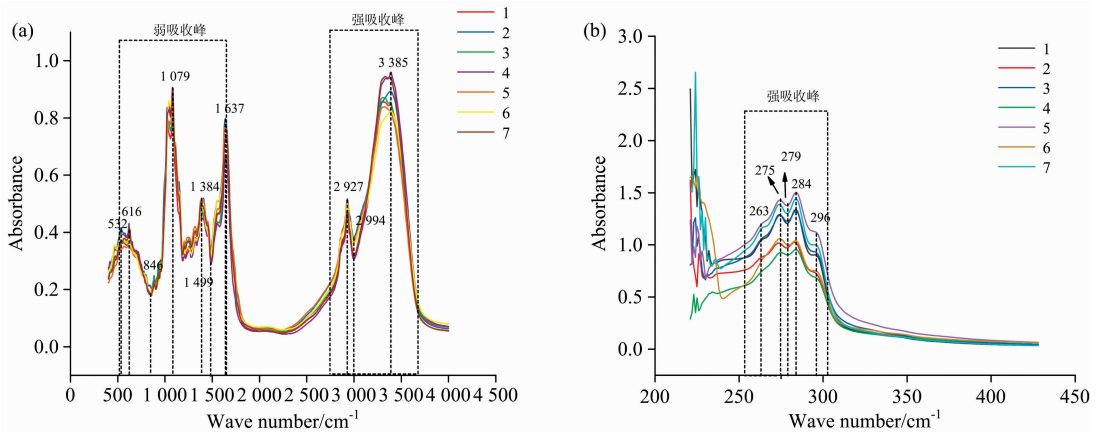



**图 2　7 种牛肝菌平均光谱**

(a)：中红外光谱；(b)：紫外光谱

**Fig. 2　Average spectra of 7 species of Boletus**

(a)：Mid-infrared spectrum；(b)：UV-Vis spectrum

### 2.2　偏最小二乘法判别分析

偏最小二乘法判别分析(partial least squares discriminantion analysis)，是一种与主成分有关的统计学方法，将数据降维后建立回归模型并对结果进行判别分析[12]。模型通过 $Y$ 轴累积解释率($R^{2Y}_{cum}$)、模型累积预测率($Q^{2}_{cum}$)、校正均方根误差(root mean square error of calibration，RMSEC)、交叉验证均方根误差(root mean square error of cross validation，RMSECV)建立模型参数，每个主成分数均能反映出变量的大量信息。$R^{2Y}_{cum}$与 $Q^{2}_{cum}$值差距越小且接近 1；RMSEC 与 RMSECV 值小于 1 且接近 0，表示模型效果越好。如表 2 所示，单一光谱的 $R^{2Y}_{cum}$和 $Q^{2}_{cum}$效果相比数据融合后的 $R^{2Y}_{cum}$和 $Q^{2}_{cum}$效果差。紫外数据模型与中红外数据模型相比，中红外数据模型效果较好，其中采用 2D＋MSC＋SNV 预处理后的数据模型训练集和验证集正确率为 99.78％和 99.12％。中红外模型效果从差到好为 1D＋MSC＋SNV＋SG、SNV＋SG、2D＋MSC＋SNV；紫外模型效果从差到好为 MSC＋2D、2D＋MSC＋SNV、1D＋MSC＋SNV＋SG、SNV＋SG。

**表 2　偏最小二乘法判别分析模型主要参数与正确率**

**Table 2　The main parameters and accuracy of the discriminant analysis model of partial least squares**

| 光谱类型 | 预处理方法 | 主成分数 | $R^{2Y}_{cum}$ | $Q^{2}_{cum}$ | RMSEC | RMSECV | 训练集/％ | 验证集/％ |
|---|---|---|---|---|---|---|---|---|
| 中红外光谱 | SNV＋SG | 24 | 0.693 | 0.555 | 0.172 449 | 0.211 937 | 97.59 | 98.24 |
| | 2D＋MSC＋SNV | 18 | 0.778 | 0.651 | 0.149 082 | 0.193 025 | 99.78 | 99.12 |
| | 1D＋MSC＋SNV＋SG | 17 | 0.673 | 0.575 | 0.178 289 | 0.210 069 | 97.37 | 97.80 |
| 紫外光谱 | MSC＋2D | 7 | 0.038 2 | 0.024 7 | 0.320 725 | 0.329 83 | 34.65 | 32.16 |
| | SNV＋SG | 16 | 0.426 | 0.281 | 0.247 129 | 0.282 096 | 80.70 | 80.18 |
| | 2D＋MSC＋SNV | 10 | 0.351 | 0.262 | 0.265 682 | 0.284 48 | 67.32 | 65.64 |
| | 1D＋MSC＋SNV＋SG | 16 | 0.364 | 0.257 | 0.262 953 | 0.284 499 | 76.10 | 73.13 |
| 低级融合 | 2D＋MSC＋SNV，SNV＋SG | 19 | 0.8 | 0.676 | 0.140 484 | 0.185 499 | 100 | 99.12 |
| 中级融合(LVs) | | 6 | 0.834 | 0.812 | 0.131 507 | 0.148 948 | 100 | 99.56 |
| 中级融合(CPA) | | 11 | 0.927 | 0.882 | 0.086 528 | 0.152 926 | 100 | 100 |

低级融合通过单一光谱数据模型，选取最佳预处理(红外：2D+MSC+SNV，紫外：SNV+SG)数据进行合并建模。模型选择特征值大于 1 的成分作为提取主成分的个数。根据表 2 所示，低级融合选取主成分 19 个，$R^{2Y}_{cum}$ 为 0.8，$Q^{2}_{cum}$ 为 0.676；RMSEC 与 RMSECV 分别为 0.140 484 和 0.185 499，模型正确率达到训练集 100%，验证集 99.12%。低级融合效果相比单一光谱最佳预处理(2D+MSC+SNV)模型，其验证集正确率均为 99.12%，训练集低级融合正确率为 100%。中级融合则选取最佳预处理后的数据结合 LVs 和 CPA 方法提取特征值组合成新矩阵进行建模。中级模型效果最佳，其 LVs 特征值提取法模型主成分 6 个，$R^{2Y}_{cum}$ 和 $Q^{2}_{cum}$ 值为 0.834 与 0.812，参数值差距小且与 1 接近，训练集为 100%，验证集为 99.56%。CPA 特征值提取法模型主成分 11 个，RMSEC(0.086 528)、RMSECV(0.152 926)、$R^{2Y}_{cum}$(0.927)、$Q^{2}_{cum}$(0.882)，正确率训练集与验证集均为 100%。

### 2.3 随机森林分析

随机森林(random forest，RF)属于机器学习中集成学习方法的一种，包含了多个 Bagging 集成学习技术的决策树，通过集成学习思维将多棵决策树作为分类器的集成算法[13]。对于一个输入样本，将输出 N 个分类结果的原理，在多个分类器输出的分类结果中筛选投票次数最多的类别作为结果。模型建立中参数 Ntree 和 Mtry 的选择决定模型质量。Ntree 指定随机森林所包含的决策树树目，Ntree 越大，决策树棵数越多，模型训练工作量越大。Mtry 确定了每次迭代变量抽样数值，用于二叉树变量个数。为避免过拟合风险增加，模型建立分为两步，先进行初始筛选，计算出袋外错误率(out-of-bag error，OOB)，再进行 Ntree 和 Mtry 精确筛选，以达到高质量模型建立的目的。如表 3 所示，在单一光谱数据矩阵中，中红外光谱采用了三种预处理组合方法建立随机森林模型，分别是 SNV+SG，2D+MSC+SNV，1D+MSC+SNV+SG。初始筛选 Ntree：2 000，1 800 和 2 100，Mtry：40，50 和 42；精确筛选 Ntree：1 412，1 149 和 1 537，Mtry：3，5 和 5。三种不同预处理得出不同的正确率，其中 2D+MSC+SNV 预处理方法结果最佳，训练集和验证集分别为 93.20%和 99%。紫外光谱采用了 4 种预处理方法，由表 3 可知，紫外光谱模型结果正确率均低于中红外光谱，其中 1D+MSC+SNV+SG 精确筛选后的 Ntree 和 Mtry 值为 1 562 和 4，训练集和验证集分别为 62.28%与 76.14%，模型较佳。

低级融合是将最佳的单一光谱预处理(红外：2D+MSC+SNV，紫外：1D+MSC+SNV+SG)矩阵进行简单的数据合并，组合成新的数据集，Kennard-Stone 分类后再进行模型建立，低级融合效果相比单一光谱，正确率有所提高，其中训练集和验证集分别为 92.32%和 99.14%。结果表明数据融合策略用于随机森林算法具有可行性，中级融合方法为两个单一光谱(中红外，紫外)数据特征提取值组合矩阵。本工作使用了两种特征值提取方法，潜在变量和主成分提取法。不同的特征值提取法结合 OOB 选择最佳决策树树目。图 3 为两种特征值提取法模型的 Ntree 选择信息；虽决策树选取越多越好，但为避免过拟合风险，在图 3(a)，图 3(b)预测结果中均选用平稳段 Ntree(865，1 249)，和 Mtry 随时抽样变量个数 5；图 3(c)—(f)分别为模型正确率矩阵，其中蓝色部分为正确数，黄色部分为错误数，通过矩阵可确定 7 类牛肝菌训练集和验证集模型的正确率。由图 3(c)和(e)可知，中级融合(LVs)训练集和验证集正确率总个数为 423 与 218；由图 3(d)和(f)可知，中级融合(CPA)训练集和验证集正确率总个数为 443 与 227。中级融合 LVs 预测集与验证集为 92.76%和 96.04%，中级融合 CPA 预测集与验证集为 97.15%和 100%。

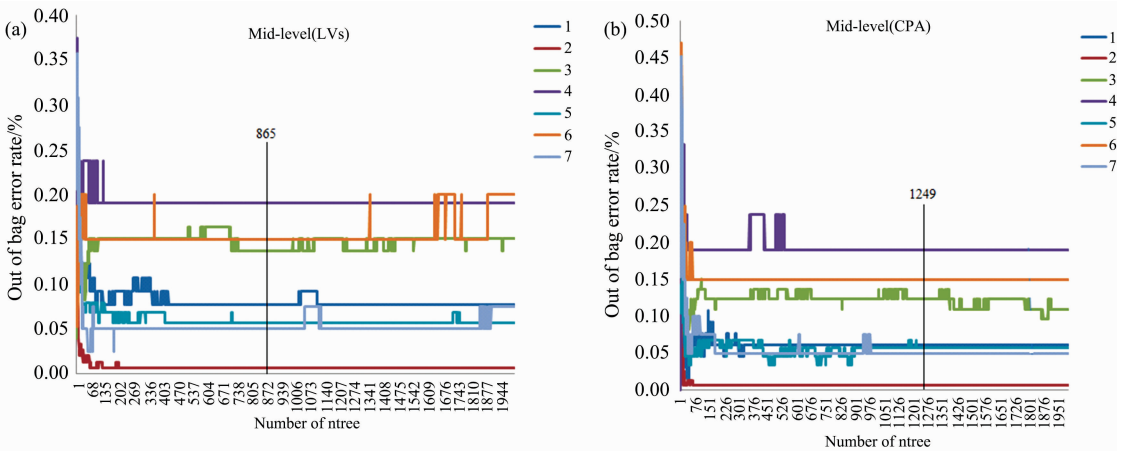
**表 3 随机森林模型主要参数与正确率**

**Table 3 The main parameters and accuracy of the random forest model**

| 光谱类型 | 预处理方法 | Ntree | Mtry | 训练集/% | 验证集/% |
|---|---|---|---|---|---|
| 中红外光谱 | SNV+SG | 2 000，1 412 | 40，3 | 76.75 | 85 |
| | 2D+MSC+SNV | 1 800，1 149 | 50，5 | 93.20 | 99 |
| | 1D+MSC+SNV+SG | 2 100，1 537 | 42，5 | 87.72 | 96.29 |
| 紫外光谱 | 2D+MSC | 2 000，1 821 | 40，4 | 62.06 | 64.30 |
| | SNV+SG | 1 700，1 327 | 35，2 | 56.80 | 70.71 |
| | 2D+MSC+SNV | 2 000，1 549 | 40，5 | 64.47 | 73.10 |
| | 1D+MSC+SNV+SG | 1 800，1 562 | 30，4 | 62.28 | 76.14 |
| 低级融合 | 2D+MSC+SNV，1D+MSC+SNV+SG | 2 000，1 184 | 40，5 | 92.32 | 99.14 |
| 中级融合(LVs) | | 2 000，865 | 40，5 | 92.76 | 96.04 |
| 中级融合(CPA) | | 2 000，1 249 | 40，5 | 97.15 | 100 |

## 3 结 论

研究了 7 种牛肝菌两种光谱谱数据融合结合偏最小二乘法判别分析和随机森林算法建立数学模型分析牛肝菌种类的准确性。结果表明：7 种牛肝菌红外平均光谱和紫外平均光谱吸收峰大致相同，吸光度具有细微差异，不同种类牛肝菌化学成分相同但含量具有差异性。偏最小二乘法判别分析和随机森林算法模型的中红外光谱数据和紫外光谱数据最佳预处理组合为 2D+MSC+SNV 和 SNV+SG，2D+MSC+SNV




(c)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 正确率(%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 1 | 4 | 0 | 0 | 0 | 0 | 92.31 |
| 2 | 1 | 148 | 0 | 0 | 0 | 0 | 0 | 99.33 |
| 3 | 3 | 0 | 62 | 3 | 4 | 0 | 1 | 84.93 |
| 4 | 1 | 0 | 1 | 17 | 0 | 2 | 0 | 80.95 |
| 5 | 1 | 0 | 3 | 0 | 83 | 0 | 1 | 94.32 |
| 6 | 1 | 1 | 2 | 0 | 0 | 16 | 0 | 80 |
| 7 | 1 | 1 | 1 | 0 | 0 | 0 | 37 | 92.5 |
| 合计 | | | | | | | | 92.76 |

(d)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 正确率(%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 63 | 2 | 0 | 0 | 0 | 0 | 0 | 96.92 |
| 2 | 0 | 149 | 0 | 0 | 0 | 0 | 0 | 100 |
| 3 | 0 | 2 | 70 | 0 | 1 | 0 | 0 | 95.89 |
| 4 | 1 | 1 | 0 | 18 | 1 | 0 | 0 | 85.71 |
| 5 | 0 | 0 | 0 | 0 | 88 | 0 | 0 | 100 |
| 6 | 0 | 3 | 0 | 0 | 0 | 17 | 0 | 85 |
| 7 | 0 | 2 | 0 | 0 | 1 | 0 | 38 | 92.68 |
| 合计 | | | | | | | | 97.15 |

(e)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 正确率(%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 29 | 0 | 3 | 0 | 0 | 0 | 1 | 87.88 |
| 2 | 0 | 71 | 0 | 0 | 0 | 1 | 0 | 98.61 |
| 3 | 1 | 0 | 35 | 0 | 1 | 0 | 0 | 94.59 |
| 4 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 100 |
| 5 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 100 |
| 6 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 100 |
| 7 | 0 | 1 | 0 | 0 | 1 | 0 | 18 | 90 |
| 合计 | | | | | | | | 96.04 |

(f)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 正确率(%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| 2 | 0 | 72 | 0 | 0 | 0 | 0 | 0 | 100 |
| 3 | 0 | 0 | 37 | 0 | 0 | 0 | 0 | 100 |
| 4 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 100 |
| 5 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 100 |
| 6 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 100 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 100 |
| 合计 | | | | | | | | 100 |

**图 3　Ntree 选择图与正确率矩阵**

(a)：中级融合(LVs)Ntree 最佳选择图；(b)：中级融合(CPA)Ntree 最佳选择图；(c)：中级融合(LVs)训练集正确率矩阵；(d)：中级融合(CPA)训练集正确率矩阵；(e)：中级融合(LVs)验证集正确率矩阵；(f)：中级融合(CPA)验证集正确率矩阵

**Fig. 3　Ntree selection diagram and Correct rate matrix**

(a)：Mid-level (LVs) Ntree best choice map；(b)：Mid-level (CPA) Ntree best choice map；(c)：Mid-level (LVs) Training set correct rate matrix；(d)：Mid-level (CPA) Training set correct rate matrix；(e)：Mid-level (LVs) Validation set correct rate matrix；(f)：Mid-level (CPA) Validation set correct rate matrix

和 1D+MSC+SNV+SG。单一光谱模型结合偏最小二乘法判别分析，单一光谱模型结合随机森林效果不佳，中红外光谱数据模型正确率大于紫外光谱数据模型。数据融合策略可提高牛肝菌种类鉴别模型的正确率，随机森林、偏最小二乘法判别分析结合数据融合分类效果明显。其中随机森林算法中级数据融合(CPA)训练集为 97.15%，验证集为 100%；偏最小二乘法判别分析中级数据融合(CPA)训练集和验证集正确率均能达到 100%，可作为牛肝菌种类分类方法。

## References

[1] GU Ke-fei, ZHOU Chang-yan, SHAO Yi, et al(顾可飞，周昌艳，邵　毅，等). Food Research and Development(食品研究与开发), 2017, 38(17): 129.

[2] Qi L, Liu H, Li J, et al. Sensors, 2018, 18(1): 241.
[3] Mleczek M, Rzymski P, Budka A, et al. Journal of Food Composition and Analysis, 2018, 66: 168.
[4] Li Y, Zhang J, Wang Y. Analytical and Bioanalytical Chemistry, 2018, 410(1): 91.
[5] Gao R, Chen C, Wang H, et al. PLOS ONE, 2020, 15(8): e238149.
[6] HU Yi-ran, LI Jie-qing, LIU Hong-gao, et al(胡翼然, 李杰庆, 刘鸿高, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(5): 1495.
[7] Yao S, Li T, Liu H, et al. Journal of the Science of Food and Agriculture, 2018, 98(6): 2215.
[8] YAO Sen, ZHANG Ji, LIU Hong-gao, et al(姚 森, 张 雳, 刘鸿高, 等). Food Science(食品科学), 2018, 39(1): 305.
[9] Li Y, Zhang J, Wang Y. Analytical and Bioanalytical Chemistry, 2018, 410(1): 91.
[10] Li X, Li J, Liu H, et al. International Journal of Food Properties, 2020, 23(1): 227.
[11] Yao S, Li J, Duan Z, et al. Analytical Letters, 2019, 53(7): 1019.
[12] Ríos-Reina R, Elcoroaristizabal S, Ocaña-González J A, et al. Food Chemistry, 2017, 230: 108.
[13] Probst, Philipp, Boulesteix, et al. Journal of Machine Learning Research, 2018, (18): 1.

# Identification of Boletus Species Based on Discriminant Analysis of Partial Least Squares and Random Forest Algorithm

CHEN Feng-xia[1], YANG Tian-wei[2], LI Jie-qing[1], LIU Hong-gao[3], FAN Mao-pan[1*], WANG Yuan-zhong[4*]
1. College of Resources and Environmental Sciences, Yunnan Agricultural University, Kunming 650201, China
2. Yunnan Institute for Tropical Crops Research, Jinghong 666100, China
3. College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming 650201, China
4. Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming 650200, China

**Abstract** As a famous wild edible mushroom, boletus has great edible and economic value. There are many kinds of boletus, and it is not easy to distinguish. An effective, rapid and credible species identification technology can be established to improve the quality of boletus. In this study, a total of 683 strains of 7 species of wild bolete from different regions of Yunnan were collected, the infrared and ultraviolet spectra of the samples were obtained, and the average spectral characteristics of different kinds of bolete were analyzed. Based on the single spectral data of multiple preprocessing combinations (SNV+SG, 2D+MSC+SNV, 1D+MSC+SNV+SG, MSC+2D) combined with two feature value extraction methods (PCA, LVs), the partial least squares discrimination analysis and random forest algorithm combined with data fusion strategy to identify the species of boletus. There is a certain degree of innovation. The results show: (1) The average spectral absorption peaks of different types of boletus in the mid-infrared spectrum and the ultraviolet spectrum have small differences, and the absorbance has subtle differences. (2) Appropriate preprocessing can improve spectral data information. The best preprocessing combination of mid-infrared spectral data and ultraviolet spectral data for partial least square discriminant analysis and random forest algorithm model is 2D+MSC+SNV, SNV+SG, 2D +MSC+SNV, 1D+MSC+SNV+SG. (3) The mid-infrared spectroscopy model is better than the ultraviolet spectroscopy model in the single spectrum model. The partial least squares discriminant analysis model of the best preprocessing combination of mid-infrared spectroscopy 2D+MSC+SNV has a correct rate of 99.78% in the training set and 99.12% in the validation set. The accuracy of the random forest model is 93.20% on the training set and 99% on the validation set. (4) The data fusion strategy improves classification accuracy. The accuracy of the low-level fusion partial least squares discriminant analysis model training set and validation set is 100%, 99.12%. The accuracy of the random forest model's training set and validation set are 92.32% and 99.14%. (5) Random Forest Algorithm Intermediate Data Fusion latent variable (LVs) training set 92.76%, validation set 96%; Intermediate Data Fusion principal components analysis (CPA) training set 97.15%, validation set 100%. (6) Partial Least Squares Discriminant Analysis Intermediate Data Fusion (LVs) training set is 100%, and validation set is 99.56%; the accuracy of intermediate data fusion (CPA) training set and validation set can reach 100%. Based on the discriminant analysis of the partial least squares method and random forest algorithm combined with data fusion strategy, the species identification of boletus is satisfactory. Partial Least Squares Discriminant Analysis Intermediate Data Fusion (CPA) can be used as a low-cost and high-efficiency technology for identifying boletus species.

**Keywords** Boletus; Mid-infrared spectroscopy; Ultraviolet spectroscopy; Discriminant analysis by partial least squares; Random forest; Data fusion

* Corresponding authors
(Received Jan. 4, 2021; accepted Feb. 1, 2021)