17种分类算法在牛肝菌种类鉴别研究中的应用

张 钰[1,2]，李杰庆[1]，李 涛[3]，刘鸿高[1*]，王元忠[2*]

1. 云南农业大学农学与生物技术学院，云南 昆明 650201
2. 云南省农业科学院药用植物研究所，云南 昆明 650200
3. 玉溪师范学院资源环境学院，云南 玉溪 653100

摘 要 由于部分毒菌与野生食用菌形态和生物学特征相似，农民仅凭经验采集，难免将两者混淆，从而导致严重的食品安全事故。云南省作为国内野生食用菌产量最高、出口量最大的省份，野生食用菌产业发展为云南农村经济发展做出了突出贡献，对不同种类野生食用菌进行快速鉴别，有利于野生食用菌产业的健康发展；分析食用菌亲缘关系，对食用菌育种工作具有积极作用。七种牛肝菌样品，采自云南及周边七个产地，利用FTIR光谱仪分别采集菌柄和菌盖红外指纹图谱，基于低级与中级数据融合策略，将预处理后的菌柄和菌盖FTIR光谱数据进行融合，结合Decision Trees，Discriminant Analysis，Logistic Regression Classifiers，Support Vector Machines，Nearest Neighbor Classifiers和Ensemble Classifiers中的17种算法，分别建立菌柄、菌盖、低级数据融合和中级数据融合模型，每个分类模型连续进行10次运算，通过比较训练集分类正确率平均值，确定牛肝菌种类鉴别最佳分类算法。中级数据融合数据集进行系统聚类分析(HCA)，对推测不同种类牛肝菌样品的亲缘关系进行鉴定。结果显示：(1)菌柄、菌盖和低级数据融合模型最佳分类算法均为Linear Discriminant，训练集分类正确率分别为92.8%，96.4%和97.6%。中级数据融合模型最佳分类算法为Subspace Discriminant，训练集分类正确率为100%；(2)菌柄、菌盖、低级数据融合和中级数据融合最佳分类模型，全部样品分类正确率平均值分别为93.61%，95.54%，96.99%和99.88%，中级数据融合模型优于其他三种模型，表明中级数据模型可以将相似度较高的样品区分开，且减少了产地对种类鉴别的影响；(3)中级数据融合模型数据集进行HCA，华丽牛肝菌和美味牛肝菌聚类距离最小，表明这两种牛肝菌化学信息较相似，亲缘关系较近；(4)华丽牛肝菌与皱盖疣柄牛肝菌聚类临界值距离最大，表明样品化学信息差异较大，亲缘关系较远。综上表明，基于中级融合策略将不同部位FTIR光谱数据融合，结合Subspace Discriminant与HCA，可以准确鉴别不同种类牛肝菌和快速推测样品亲缘关系，可作为野生食用菌种类鉴别与亲缘关系推测的一种新方法。

关键词 牛肝菌；FTIR；种类鉴别；不同部位；数据融合

中图分类号：TS227 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)02-0448-06

引 言

毒蘑菇亦称毒菌，全世界有1 000种以上，我国至少有435种，其中39种为牛肝菌，约占我国已知牛肝菌种类的11%[1-3]。目前，多数野生牛肝菌无法人工栽培，主要来源于农民野外采集。由于许多毒菌与野生食用菌形态和生物学特征相似，农民仅凭经验采集，难免将两者混淆，从而导致严重的食品安全事故。例如，2012年四川一次婚宴上，200多人出现集体食品中毒症状，研究后发现华丽牛肝菌、毒牛肝菌和中华牛肝菌三种毒菌的混入是导致此次事故的罪魁祸首[4]。云南省作为国内野生食用菌产量最高、出口量最大的省份，野生食用菌产业发展为云南农村经济发展做出了突出贡献。因此，对不同种类野生食用菌进行快速鉴别，有利于野生食用菌产业的健康发展。另一方面随着野生食用菌驯化、栽培等过程，菌种出现异种同名或同种异名的现象，以

收稿日期：2017-12-09，**修订日期**：2018-05-21

基金项目：国家自然科学基金项目(31660591，21667031)，云南省教育厅科学研究基金项目(2016ZZX106)，云南省高校食用菌资源开发与利用重点实验室建设项目资助

作者简介：张 钰，1992年生，云南农业大学农学与生物技术学院硕士研究生 e-mail：m15343842322@163.com

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

及品种退化、混杂、病毒感染等问题[5]，分析食用菌亲缘关系，对食用菌育种工作具有积极作用。

对遗传背景、有机成分、矿质元素等特征性指纹信息差异进行分析，可用于鉴别不同种类野生食用菌[6-7]。经过加工后失去原有形态特征的样品，凭借传统形态学鉴别很难区分开。采用交配亲和试验分类[8]、同工酶分析[9]、分子生物学鉴别分析等方法[10-11]，操作繁琐、检测时间长且费用高等，难以对大量的商品野生食用菌进行快速鉴别。傅里叶变换红外光谱(Fourier transform infrared spectrometer，FTIR)鉴别分析技术，是有机成分指纹分析中常用的一种方法，具有快速、经济、可靠、简便等特点，现在已广泛应用于中草药、食品等领域[12-13]。

数据融合通过数据获取、预处理、特征筛选，将两个或两个以上光谱数据进行融合，得到的新数据集进行模型训练，从不同的信息层面反映样品间差异，更加全面解释样品属性[14]。Márquez 等[15]基于数据融合策略将拉曼光谱和近红外光谱融合，对榛子酱中掺入杏仁酱与鹰嘴豆酱的食品欺诈行为进行鉴别，结果显示数据融合分类判别结果优于单一信息。Reis 等[16]采用傅里叶变换衰减全反射与漫反射红外光谱数据的融合，用来鉴别咖啡掺假，对四种掺假方式进行判别分析，结果显示，数据融合模型分类效果优于采用单一数据模型。以上研究结果表明数据融合较单一信息，显示更多化学指纹信息差异，有利于准确的样品表征。

食用菌样品特征性指纹信息不仅受到遗传信息影响，还可能受产地、储存年限、采集年份等多种因素干扰，为了在复杂的牛肝菌背景信息中，探讨产地因素对不同种类样品鉴别的影响，本研究 7 种牛肝菌样品，采自云南及周边 7 个产地，采用 FTIR 光谱仪测定牛肝菌的菌柄和菌盖红外指纹图谱。基于低级与中级数据融合策略，将预处理后的菌柄和菌盖 FTIR 光谱数据进行融合，结合 Decision Trees，Discriminant Analysis，Logistic Regression Classifiers，Support Vector Machines，Nearest Neighbor Classifiers 和 Ensemble Classifiers 中 17 种算法，分别建立菌柄、菌盖、低级数据融合、中级数据融合模型。通过比较分类鉴别的结果，确定最佳分类模型。最佳分类数据矩阵进行系统聚类分析(hierarchical cluster analysis，HCA)，探讨不同种类牛肝菌的亲缘关系，以期为野生食用菌质量控制提供一种有效的新方法。

1　实验部分

1.1　样品

246 份七种牛肝菌的不同部位(菌柄、菌盖)的样品，于 2012 年采自云南省及周边七个产地(每个州或市代表一个产地)见图 1，材料来源见表 1。

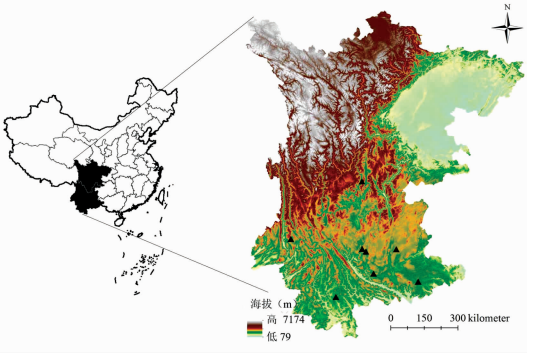

图 1　牛肝菌样品地理信息

Fig. 1　The geographical location of Boletaceae samples with different species

表 1　不同种牛肝菌材料来源

Table 1　Information of Boletaceae samples with different species

样品名	拉丁名	采集地点	编号
华丽牛肝菌	*B. magnificus*	云南省玉溪市易门县铜厂乡	1-1—1-10
华丽牛肝菌	*B. magnificus*	云南省普洱市思茅区	1-11—1-20
小美牛肝菌	*B. speciosus*	云南省玉溪市易门县铜厂乡	2-1—2-10
小美牛肝菌	*B. speciosus*	云南省保山市隆阳区	2-11—2-20
砖红绒盖牛肝菌	*Xerocomusspadiceus*	云南省玉溪市易门县铜厂乡	3-1—3-7
砖红绒盖牛肝菌	*X. spadiceus*	云南省昆明市石林县维则乡	3-8—3-15
皱盖疣柄牛肝菌	*Leccinumduriusculum*	云南省保山市隆阳区	4-1—4-7
皱盖疣柄牛肝菌	*L. duriusculum*	云南省玉溪市易门县普贝乡	4-8—4-14
栗色牛肝菌	*B. umbriniporus*	云南省红河州石屏县陶村乡	5-1—5-8
栗色牛肝菌	*B. umbriniporus*	云南省玉溪市易门县铜厂乡	5-9—5-18
绒柄牛肝菌	*B. tomentipes*	云南省玉溪市易门铜厂乡	6-1—6-10
绒柄牛肝菌	*B. tomentipes*	四川凉山州德昌县马安乡	6-11—6-19
美味牛肝菌	*B. edulis*	云南省玉溪市易门县铜厂乡	7-1—7-10
美味牛肝菌	*B. edulis*	云南省文山市东山乡	7-11—7-17

1.2　仪器与试剂

傅里叶变换红外光谱仪(美国珀金埃尔默公司，配有 DTGS 检测器)、奥豪斯电子分析天平(梅特勒-托多仪器有限公司)、YP-2 型压片机(上海市山岳科学仪器有限公司)、101A-1 型电热鼓风恒温干燥箱(上海市崇明实验仪器厂)、60 目不锈钢筛盘(浙江上虞市道墟五四仪器厂)、溴化钾(分

析纯；天津市风船化学试剂科技有限公司)。

1.3 样品处理及 FTIR 光谱采集

样品采集后去除杂质，清水洗净，50 ℃烘干至恒重。牛肝菌样品分为菌盖和菌柄两部分粉碎后过 60 目筛，分别装于自封袋中避光保存备用。称取样品粉末(1.5±0.2) mg 及溴化钾粉末(100±0.2) mg，在玛瑙研钵中充分混合并研磨成细粉，放入压片模具压成厚度均匀的薄片。仪器预热 1 h 后测定光谱，光谱扫描范围 4 000～400 cm^{-1}，扫描前扣除压片背景的干扰，每个样品重复测定 3 次，取平均光谱作为样品测量光谱。

1.4 光谱数据预处理

Omnic8.2 软件对样品 FTIR 原始光谱进行吸光度转换、纵坐标归一化处理。SIMCA-P^+ 13.0 对 FTIR 原始光谱，进行标准正态变换(standard normal variate，SNV)、二阶导数(second-derivative，2D)预处理，Origin8.0 软件绘制菌柄、菌盖 FTIR 平均光谱图，以及对 FTIR 光谱数据进行 PLS-DA 降维，提取特征变量。利用 MATLAB2017a 软件中 Classification Learnertoolbox 对数据集进行模型训练。

2 结果与讨论

2.1 FTIR 光谱分析

图 2 为经过吸光度转换、纵坐标归一化预处理后，不同种类牛肝菌 FTIR 平均光谱。由图可知，不同种类牛肝菌 FTIR 光谱在 3 291，2 929，1 640，1 548，1 401，1 082，1 027，619，533 和 469 cm^{-1} 等附近有明显的特征吸收峰。3 291 cm^{-1} 附近吸收峰主要为多糖和蛋白质中羟基的 O—H 伸缩振动以及 N—H 伸缩振动，2 929 cm^{-1} 附近吸收峰主要为多糖、蛋白质中甲基、亚甲基伸缩振动，1 640 cm^{-1} 附近主要为芳香环中 C═C 伸缩振动，1 548 cm^{-1} 附近主要为蛋白质酰胺Ⅱ带的 N—H 变形和 C═N 伸缩振动，1 401 cm^{-1} 附近可能为羧酸根离子中 C—O 弯曲振动和 O—H 变形，1 239 cm^{-1} 附近可能为羧酸中的 C—O 伸缩振动和 O—H 变形或者是酚类 C—O 伸缩振动，1 082 cm^{-1} 附近为脂肪族 C—OH 伸缩振动，1 082～1 027 cm^{-1} 范围内吸收峰为多糖或类多糖 C—O 伸缩振动，1 027～619 cm^{-1} 附近主要是糖类的异构体[17-19]。吸收峰强度的差异，表明不同种类牛肝菌内蛋白质、脂肪酸、多糖、芳香族类等化学成分含量可能存在差异。

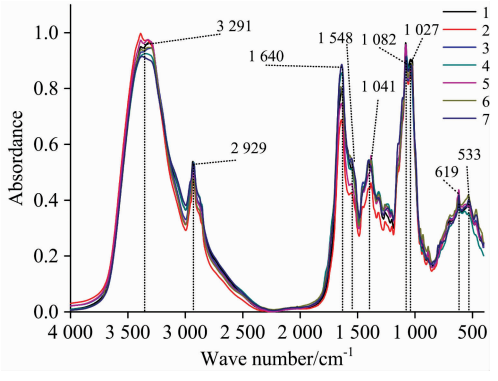

图 2 不同种牛肝菌的 FTIR 平均光谱

Fig. 2 FTIR Average spectra of Boletaceae samples with different species

2.2 分类算法的选择

模式识别(pattern recognition)是通过机器处理和分析各种信息，以对事物进行描述、解释和分类[20]，其过程为数据获取、预处理、特征筛选和分类决策，现在广泛应用于食品安全监控、计算机、生物信息学、海洋探测、分析化学等领域[21-25]。支持向量机(support vector machine，SVM)由 Vapnik 于 1995 年提出，是一种基于风险最小化与 VC 维理论建立的机器学习方法，其优势是解决小样本、高维数据、非线性、局部极小点等问题[26]，K-最近邻分类算法(K-nearest neighbor，KNN)是将最邻近的一个样品扩展为 K 个，当 K 个样品中属于哪一类样品较多，就归为这一类。决策树(decision tree)从一组无规则的事例推理出决策树表示形式的分类规则，具有计算量较小、分类准确率较高等优点。传统的 SVM，KNN，Tree 等方法解决二分类问题存在一些局限，为了克服不足，对其进行了改进，例如 Linear SVM，Quadratic SVM，Cosine KNN 和 Cubic KNN 等算法。

柄和盖 FTIR 光谱预处理后，分别建立柄和盖的独立决策模型。将归一化后柄和盖 FTIR 光谱数据串联，得到新的数据集进行模型训练，建立低级数据融合模型。柄和盖 FTIR 光谱数据进行 PLS-DA 降维，筛选特征值大于 1 的前几个主成分，分别提取柄和盖模型的前 16 与前 18 个主成分得分进行融合，建立中级数据融合模型。不同模型的数据集采用 Kennard-Stone 算法，选择 83 个样品作为训练集，其余 40 个样品为预测集，Complex Tree，Medium Tree，Simple Tree，Linear Discriminant，Linear SVM，Quadratic SVM，Cubic

表 2 不同分类器的训练集预测结果(%)

Table 2 Predicting results of training set with different classifier (%)

模型	柄	盖	低级数据融合	中级数据融合
Complex Tree	45.8	48.2	75.5	77.1
Medium Tree	45.8	48.2	75.9	77.1
Simple Tree	41	42.2	59	50.6
Linear Discriminant	92.8	96.4	97.6	97.6
Linear SVM	79.5	75.9	92.8	98.8
Quadratic SVM	77.1	71.1	91.6	98.8
Cubic SVM	77.1	67.5	90.4	98.8
Fine Gaussian SVM	13.3	18.1	13.3	13.3
Medium Gaussian SVM	73.5	62.7	89.2	44.6
Fine KNN	66.3	63.7	89.2	95.2
Medium KNN	68.7	55.4	79.5	91.6
Cosine KNN	68.7	55.4	80.7	97.6
Cubic KNN	66.3	55.4	81.9	90.4
Weighted KNN	71.1	60.2	81.9	94
Bagged Trees	63.9	56.6	89.2	92.8
Subspace Discriminant	90.4	94	96.4	100
Subspace KNN	67.5	61.4	88	96.4

SVM，Fine Gaussian SVM，Medium Gaussian SVM，Fine KNN，Medium KNN，Cosine KNN，Cubic KNN，Weighted KNN，Bagged Trees，Subspace Discriminant 和 Subspace KNN 算法进行模型训练，比较训练集正确率(表 2)，确定不同模型的最佳分类算法。结果显示，菌柄、菌盖和低级数据融合模型最佳分类算法均为 Linear Discriminant，训练集正确率分别为 92.8%，96.4%和 97.6%。中级数据融合模型为 Subspace Discriminant，训练集正确率为 100%。表明 Linear Discriminant 与 Subspace Discriminant 算法能改善模型分类效果。

2.3　不同模型最佳分类算法比较

Linear Discriminant 与 Subspace Discriminant 对模型数据集进行运算，模型随机选择有代表性的数据进行训练，运算结果可能出现偏差，需要对数据进行多次运算[27-28]。本次研究对菌柄、菌盖、低级数据融合和中级数据融合模型进行 10 次运算，训练集预测结果混淆矩阵见图 3，横坐标表示样品的真实标签，纵坐标表示预测标签。菌柄、菌盖、低级数据融合和中级数据融合最佳分类模型，训练集(83 个样品)进行 10 次运算，平均每次分类正确数分别为 77.6，79.3，80.5 和 82.8，表明中级数据融合模型分类效果最佳。对不同模型进行 10 次运算，模型样品数为 830(10×83)个，柄模型分类效果最差，第 1 类中 10 和 10 个样品被错误分类为第 3 和 7 类。第 2 类中 1，1 和 4 个样品，分别被错误分类为第 1，5 和 7 类。第 3 类中 25、1 和 2 个样品，分别被错误分类为第 1，2 和 6 类。中级数据融合模型分类效果最佳，仅第 1 类样品中 2 个样品被错误分类为第 7 类；不同模型进行 10 次运算，模型预测集正确率均为 100%，全部样品运算结果见表 3，菌柄、菌盖、低级数据融合和中级数据融合模型，全部样品分类正确率平均值分别为 93.61%，95.54%，96.99%和 99.88%。结果表明基于中级数据融合策略，将不同部位 FTIR光谱数据进行融合，应用 Subspace Discriminant 算法，可以将相似度较高的样品区分开，且减少了产地对种类鉴别的影响，是一种快速、准确鉴别不同种类牛肝菌的新方法。

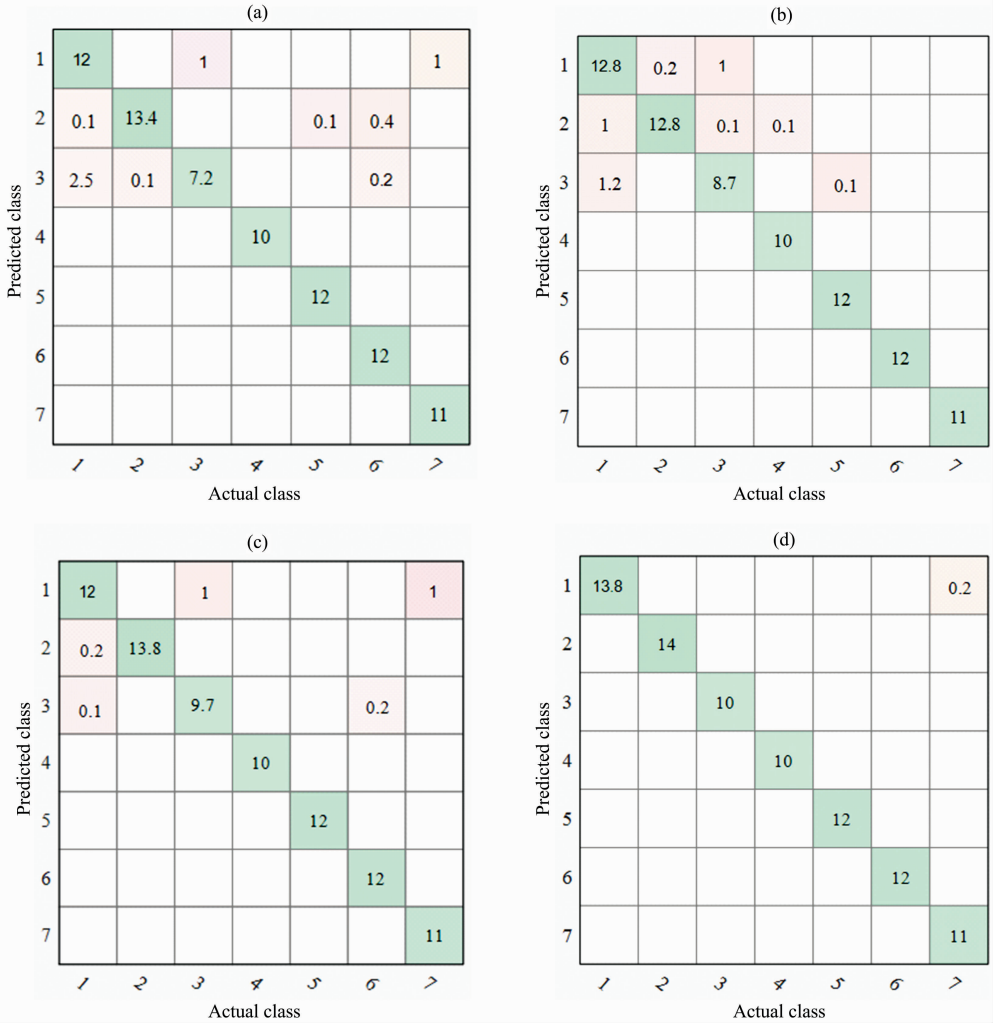

图 3　训练集预测结果混淆矩阵

(a)：菌柄；(b)：菌盖；(c)：低级数据融合；(d)：中级数据融合

Fig. 3　The confusion matrix of training set

(a)：UV-Vis；(b)：FTIR；(c)：Low-level data fusion；(d)：Mid-level data fusion

表 3 最佳分类模型预测结果(%)

Table 3 Prediction results of the optimal model(%)

次数	柄	盖	低级数据融合	中级数据融合
1	93.98	95.18	97.60	100
2	93.98	96.39	95.20	100
3	92.77	93.98	96.40	100
4	92.77	95.18	97.60	100
5	92.77	96.39	97.60	100
6	95.18	96.39	95.20	100
7	95.18	95.18	96.40	100
8	93.98	95.18	97.60	100
9	92.77	96.39	96.40	98.80
10	92.77	95.18	97.60	100.00
平均值	93.61	95.54	96.99	99.88

2.4 系统聚类分析

系统聚类分析(hierarchical cluster analysis，HCA)依据样品特征性指纹信息相似程度，将比较相似样品聚为一类[29]。采用不同部位 FTIR 中级数据融合数据进行 HCA，首先对不同部位 FTIR 光谱中有利于分类的化学信息进行挖掘，比较化学信息差异，推测样品亲缘关系。不同种类牛肝菌中级数据融合 HCA 见图 4，图中相同颜色的样品代表同一种类，横坐标为样品编号，纵坐标为不同种类间临界值的距离，距离越小，聚为一类的样品越相似。聚类分析树状图显示，当临界值距离为 2 时，不同种类牛肝菌样品分类全部正确，七种牛肝菌样品被分为七组，第一到七组分别为 1 号(华丽牛肝菌)、7 号(美味牛肝菌)、2 号(小美牛肝菌)、5 号(栗色牛肝菌)、3 号(砖红绒盖牛肝菌)、6 号(绒柄牛肝菌)和 4 号(皱盖疣柄牛肝菌)样品，聚类临界值距离从第一组到第七组逐渐变大，表明华丽牛肝菌与美味牛肝菌样品化学信息较相似，这两种牛肝菌亲缘关系可能比较近。华丽牛肝菌与皱盖疣柄牛肝菌样品化学信息差异最大，这两种牛肝菌亲缘关系可能较远。

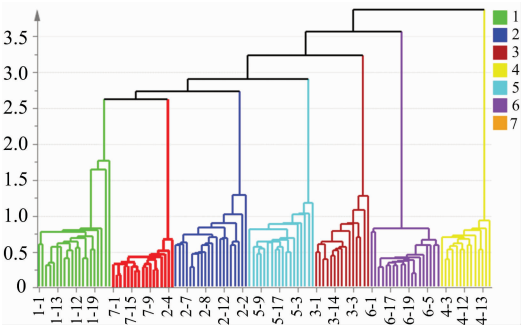

图 4 中级数据融合模型系统聚类分析图

Fig. 4 The HCA plots of mid-level data fusion

3 结 论

用 FTIR 光谱仪分别采集七种牛肝菌，菌柄和菌盖样品红外光谱指纹图谱，基于低级与中级数据融合策略，将预处理后的菌柄和菌盖 FTIR 光谱数据进行融合，结合 Complex Tree，Medium Tree，Simple Tree，Linear Discriminant 等 17 种算法，分别建立菌柄、菌盖、低级数据融合和中级数据融合分类模型。结果显示，菌柄、菌盖和低级数据融合模型最佳分类算法为 Linear Discriminant，中级数据融合模型为 Subspace Discriminant。中级数据融合分类结果优于菌柄、菌盖和低级数据融合，表明中级数据模型可以将相似度较高的样品区分开，且减少了产地对种类鉴别的影响。HCA 结果显示，华丽牛肝菌和美味牛肝菌聚类临界值距离最小，表明样品化学信息较相似，这两种牛肝菌亲缘关系可能比较近，华丽牛肝菌与皱盖疣柄牛肝菌聚类临界值距离最大，表明样品化学信息差异最大，这两种牛肝菌亲缘关系可能比较远。综上表明，基于中级融合策略将不同部位 FTIR 光谱数据融合，结合 Subspace Discriminant 与 HCA，可以准确鉴别不同种类牛肝菌和快速推测样品亲缘关系，可作为野生食用菌种类鉴别和推测亲缘关系的一种新方法。

References

[1] MAO Xiao-lan(卯晓岚). Mycosystema(菌物学报), 2006, 25(3): 345.
[2] BAU Tolgor, BAO Hai-ying, LI Yu(图力古尔，包海鹰，李 玉). Mycosystema(菌物学报), 2014, 33(3): 517.
[3] MAO Xiao-lan(卯晓岚). The Macrofungi in China(中国大型真菌). Zhengzhou: Henan Science and Technology Press(郑州：河南科学技术出版社), 2000.
[4] WEN Hua-an, YANG Zhu-liang, LI Tai-hui, et al(文华安，杨祝良，李泰辉，等). Science World(科学世界), 2013, (10): 56.
[5] ZHOU Zu-fa, LU Na, SONG Ji-ling, et al(周祖法，陆 娜，宋吉玲，等). Journal of Fungal Research(菌物研究), 2017, (3): 188.
[6] Juma I, Mshandete A, Tibuhwa D, et al. Tanzania Journal of Science, 2016, 42(1): 109.
[7] Zhao R L, Li G J, Sánchez-Ramírez S, et al. Fungal Diversity, 2017, 84(1): 43.
[8] Avin F A, Bhassu S, Shin T Y, et al. Journal of Animal & Plant Sciences, 2014, 24(1): 89.
[9] Fan X Z, Zhou Y, Xiao Y, et al. Microbiological Research, 2014, 169(5): 453.
[10] Lu T, Bau T. Biotechnology & Biotechnological Equipment, 2017, 31(7): 1.
[11] Yadav M K, Chandra R, Singh H B, et al. International Journal of Current Microbiology and Applied Sciences, 2017, 6(5): 1260.
[12] YANG Tian-wei, ZHANG Ji, LI Tao, et al(杨天伟，张 霁，李 涛，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(11): 3510.
[13] Qi L M, Zhang J, Zhao Y L, et al. Analytical Letters, 2017, 50(9): 1497.
[14] Ouyang Q, Zhao J W, Chen Q S. Analytica Chimica Acta, 2014, 841(23): 68.
[15] Márquez C, López M I, Ruisánchez I, et al. Talanta, 2016, 161: 80.

[16] Reis N, Botelho B G, Franca A S, et al. Food Analytical Methods, 2017, 10(8): 2700.
[17] SUN Su-qin(孙素琴). Analysis of Traditional Chinese Medicine by Infrared Spectroscopy(中药红外光谱分析与鉴定). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2010.
[18] He X S, Xi B D, Wei Z M, et al. Chemosphere, 2011, 82(4): 541.
[19] Silva S D, Feliciano R P, Boas L V, et al. Food Chemistry, 2014, 150: 489.
[20] Sergios Theodoridis. Pattern Recognition(模式识别). Translated by LI Jing-jiao(李晶皎, 译). Beijing: Publishing House of Electronics Industry(北京: 电子工业出版社), 2006.
[21] Zhang L, Li L D, Yang A Q, et al. Pattern Recognition, 2017, 69: 199.
[22] Zhang Y, Zhou G X, Jin J, et al. Neurocomputing, 2017, 225: 103.
[23] Fang J W, Wang L P, Wang Y, et al. Molecular Bio Systems, 2017, 13(8): 1575.
[24] Lu W, Dong X, Qiu L L, et al. Journal of Hazardous Materials, 2017, 326: 130.
[25] Li Y, Zhang J, Li T, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017, 177: 20.
[26] Vapnik V N. The Nature of Statistical Learning Theory Springer, 1995.
[27] Balcázar J, Dai Y, Osamu Watanabe. Algorithmic Learning Theory Washington, DC, 2001. 119.
[28] Lee Y J, Mangasarian O L. RSVM: Reduced Support Vector Machines, Proc of the First SLAM International on Data Mining, Chicago, 2001.
[29] Sârbu C, Naşcu-Briciu R D, Kot-Wasik A, et al. Food Chemistry, 2012, 130(4): 994.

Application of 17 Classification Algorithms for Authentication Research of Various Boletus

ZHANG Yu[1,2], LI Jie-qing[1], LI Tao[3], LIU Hong-gao[1*], WANG Yuan-zhong[2*]
1. College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming 650201, China
2. Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming 650200, China
3. College of Resources and Environment, Yuxi Normal University, Yuxi 653100, China

Abstract Many wild nocuous fungi are similar to the edible in morphology and biological characteristic, which easily leads to serious food safety incident because it is difficult for farmers to distinguish them just by experience. The progress of wild edible production makes a great contribution to rural economy of Yunnan province where the yield and export volume are highest in China. Rapid authentication of wild edible fungi variety is beneficial for wild edible industry towards healthy development. Meanwhile, the authentication also contributes to the analysis of the genetic relationship between edible mushroom and their breeding. Seven kinds of fungi were collected from Yunnan and other seven origins around Yunnan. Fingerprint of caps and stipe were obtained with Fourier transforms infrared (FTIR) spectrometer, respectively. Cap model, stipe model, low-level data fusion model and mid-level data fusion were established using prepressed spectra according to low- and mid-level fusion strategy combined with decision trees, discriminant analysis, logistic regression classifiers, support vector machines, nearest neighbor classifiers and ensemble classifiers that every model was computed 10 times. The optimal classification algorithm was selected based on the accuracy of training set. Hierarchical cluster analysis (HCA) was executed using the mid-level fusion dataset to judge genetic relationship between seven fungi. The results indicated: (1) The best algorithm of caps, stipe and low-level fusion is linear discrimination that accuracy is 92.8%, 96.4%, and 97.6%, respectively. Subspace discriminant is the most optimal in mid-level fusion that accuracy is 100%. (2) The average accuracy of all samples is 93.61%, 95.54%, 96.99% and 99.88% based on the best model of stipe, cap, low-level data fusion and mid-level data fusion. The performance of mid-level fusion is better than other three models, which indicated that the model could distinguish the highly -similar samples by reducing the influence caused by their origins. (3) The result of HCA based on mid-level fusion dataset displayed that the distance between *Boletus magnificus* and *B. edulis* was very close, which showed their chemical information were similar and genetic relationship was close. (4) The result of HCA based on mid-level fusion dataset displayed that the distance between *Boletus magnificus* and *Leccinum duriusculum* was very long, which showed their chemical information were different and genetic relationship was inferior. In a word, mid-level data fusion strategy combining FTIR spectra of different parts, subspace discriminant and HCA could effectively distinguish different kinds of edible fungi and judge the genetic relationship, which is a novel method used for variety authentication and genetic relationship judgment of wild edible fungi.

Keywords Boletaceae; FTIR; Species identification; Different parts; data fusion

* Corresponding authors
(Received Dec. 9, 2017; accepted May 21, 2018)