第 48 卷 第 3 期/2021 年 2 月/中国激光 研究论文

中国激光

基于随机森林算法的食源性致病菌拉曼光谱识别

王其[1]，曾万聃[1*]，夏志平[2**]，李志萍[2]，曲晗[2]

[1]上海应用技术大学计算机科学与信息工程学院，上海 201418；

[2]军事兽医研究所，吉林 长春 130062

摘要 药品食品的安全问题一直是人们关注的重点。相比于传统的食源性致病菌光谱检测方法，拉曼光谱法具有检测范围广、检测灵活、光谱特征突出等特点。本文以常见的食源性致病菌为研究对象，利用拉曼光谱仪采集了 11 种食源性致病菌样品的 132 个拉曼光谱数据，提出了一种基于主成分分析和随机森林算法的分类模型。实验结果表明，主成分分析结合随机森林算法的分类模型可以将食源性致病菌区分开，且分类准确度可达到 91.36%。

关键词 光谱学拉曼光谱；机器学习；食源性致病菌检测；主成分分析；随机森林

中图分类号 TP391 **文献标识码** A **doi**：10.3788/CJL202148.0311002

1 引 言

食源性致病菌的检测是保证公共卫生安全的关键步骤。目前，食源性致病菌的检测方法主要包括直接接种分离法、增菌培养分离法、直接实时荧光定量聚合酶链式反应(PCR)法和增光后实时 PCR 法等[1-2]。这些方法的操作过程较为复杂而且检测周期较长，可达数小时甚至数天，不能满足食源性致病菌检测预防控制的需求[3]。拉曼光谱反映的是分子内部的振动和转动能级[4-6]，是物质的指纹谱，可以用来鉴定分子中存在的官能团。拉曼光谱法具有无损、快速、准确等特点，是物质成分判别的有力工具。

将拉曼光谱法与机器学习算法结合起来对物质的成分进行识别和分类是目前光谱分析中常用的方法。利用拉曼光谱结合计算机算法进行识别分类可以缩短食源性致病菌的检测周期，大大降低人工识别拉曼峰的误判率。张燕君等[7]提出了一种结合激光拉曼光谱和人工蜂群支持向量机回归(ABC-SVR)快速定量检测三组分调和油中脂肪酸含量的方法；吴承炜等[8]提出了一种基于拉曼光谱和 Siamese 网络的相似性学习方法，该方法能够对矿物进行识别；Žuvela 等[9]基于自然遗传算法实现了拉曼诊断平台(鼻咽癌临床鼻内镜)在分子水平上的实时活体检测；de Souza Lins Borba 等[10]采集了 14 种不同品牌、不同型号的商用蓝色圆珠笔墨水在 A4 亚硫酸盐纸上墨线的拉曼光谱，建立了基于偏最小二乘判别分析(PLS-DA)的层次分类模型，这说明拉曼光谱结合计算机科学的方法是一种很有前途的快速无损的工具，可以区分文档中非常相似的墨水的类型。

随机森林(RF)算法[11-13]是一种集成学习(ensemble learning)方法。Vigneau 等[14]将随机森林应用在感官分析中，结果发现，随机森林模型比偏最小二乘(PLS)回归模型具有更好的预测能力；Lin 等[15]利用随机森林算法建立了重症监护病房(ICU)内急性肾损伤(AKI)患者的死亡率预测模型，并将模型的预测结果与其他两种机器学习模型和定制的简化急性生理评分(SAPS)II 模型进行了比较，结果表明，随机森林模型有助于 ICU 临床医生及时做出临床干预决策，对降低 AKI 患者的院内死亡率具有重要意义。Huang 等[16]采用随机森林算法对 T 细胞表位和非 T 细胞表位进行了分类，结果表明，基于特征和随森林相结合的 T 细胞表位预测方法是有效的。

史如晋等[17]构建了一种基于 Stacking 集成学习方法的食源性致病菌分类模型，成功地将大肠杆

收稿日期：2020-07-06；**修回日期**：2020-08-12；**录用日期**：2020-09-14

* **E-mail**：zengwd@sit.edu.cn；** **E-mail**：zipxia@126.com

菌 O157:H7 以及布鲁氏 S2 株分离开;但他们研究的食源性致病菌的类别数只有 2 个,并不能满足实际需求。本文在此基础上将食源性致病菌的类别数增加到 11 个(样本数为 132 个),这样虽然增加了训练和分类的难度,但更加符合实际。

本文提出了一种基于主成分分析[18-19](PCA)结合随机森林算法的拉曼光谱识别模型。本文采用拉曼光谱仪收集拉曼数据。在光谱预处理阶段,使用 min-max 进行归一化处理,使用 Savitzky-Golay 算法[20-21]进行平滑去噪处理;对于具有高维特征的样本数据,使用主成分分析进行特征降维。在模型评估阶段,本文使用 K 折交叉验证[22](K-CV)对模型进行评估;评估结果表明,本文提出的基于随机森林算法的拉曼光谱识别模型能够将收集到的食源性致病菌样本区分开。

2 数据收集与预处理

2.1 实验样本的采集

本实验所用食源性致病菌样本均购于中国工业微生物菌种保藏管理中心(CICC, http://cicc.china-cicc.org/),11 种食源性致病菌的 CICC 编号与名称如表 1 所示。利用拉曼光谱仪采集 11 种食源性致病菌样品的 132 个拉曼光谱数据,测量的拉曼偏移范围为 500~1600 cm^{-1}。所有的菌株都可以根据标准菌株编号在 CICC 网站查询。

表 1 11 种食源性致病菌的 CICC 编号与名称

Table 1 CICC numbers and names of eleven food-borne pathogenic bacteria

Number	Latin name
10869	*Yersinia enterocolitica*
10870	*Klebsiella pneumoniae*
21482	*Salmonella enterica* subsp. *enterica* serovar Infantis
21530	*Escherichia coli* EHEC O157:H7
21534	*Shigella flexneri*
21560	*Cronobacter sakazakii*
21600	*Staphylococcus aureus*
21617	*Vibrio parahaemolyticus*
22933	*Acinetobacter baumannii*
22956	*Salmonella enterica* subsp. *enterica* serovar Typhimurium
23794	*Vibrio cholerae*

拉曼光谱仪采集到的原始拉曼光谱数据特征数为 604 个,采集到的食源性致病菌特征数较大但种类相对较少,所以人工识别拉曼光谱的难度相对较大。本文以阪崎氏年轻泰坦杆菌(阪崎克罗诺杆菌)为样本进行数据预处理。图 1 为原始拉曼光谱图。

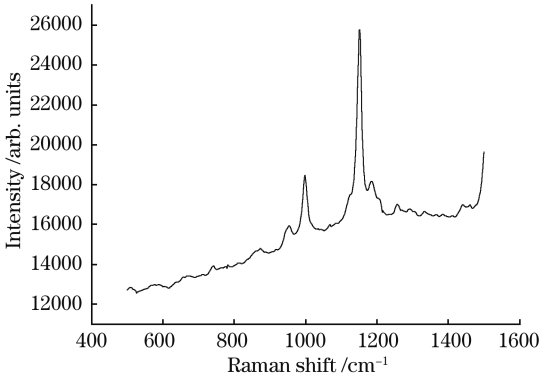

图 1 原始拉曼光谱

Fig. 1 Original Raman spectra

2.2 数据归一化

通过观察拉曼光谱仪测量得到的整体数据可以发现:不同拉曼偏移值对应的强度差异比较大。当把不同的特征列在一起时,由于特征本身表达方式的原因,绝对值大的数据的重要性大于绝对值小的数据。这时就需要对抽取出来的特征向量进行归一化处理,以保证每个特征被分类器平等地对待,使数据的处理保持一致。下面采用 min-max 准则对原始数据进行归一化处理,并对数据进行可视化。进行归一化处理的公式为

$$x_{\text{normalization}}=\frac{x-x_{\min}}{x_{\max}-x_{\min}},\tag{1}$$

式中:$x_{\max}$ 为样本数据的最大值;$x_{\min}$ 为样本数据的最小值;$x_{\text{normalization}}$ 为归一化后的样本值。对图 1 所示拉曼光谱进行归一化处理后的结果如图 2 所示。可见,将强度均映射到[0,1]之间,便于比较。

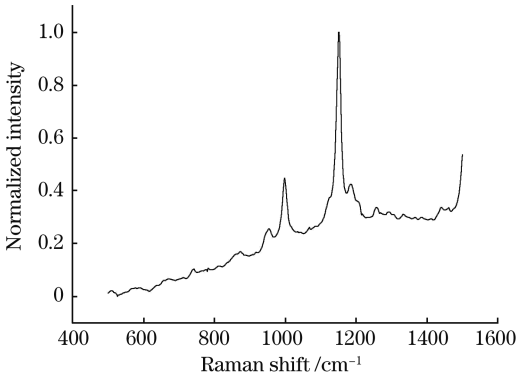

图 2 归一化后的拉曼光谱

Fig. 2 Raman spectra after normalization

2.3 Savitzky-Golay 平滑去噪

拉曼光谱仪在采集拉曼光谱数据时,会受到采

集环境的光照以及样品本身纯度等诸多因素的影响，因此收集到的拉曼光谱数据总会带有一些噪声和荧光干扰，这会在一定程度上影响光谱的质量。Savitzky-Golay 滤波算法是拉曼光谱中常用的去噪方法之一。本文也采用 Savitzky-Golay 算法进行平滑去噪处理。

本文选择的窗口宽度为 27，多项式阶数为 2。对图 2 所示光谱进行 Savitzky-Golay 平滑去噪处理，结果如图 3 所示。可以看到，拉曼光谱图上的部分毛刺在一定程度上得到了平滑处理。

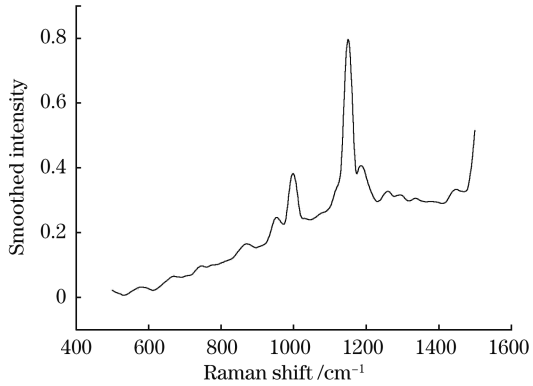

图 3　Savitzky-Golay 处理后的拉曼光谱

Fig. 3　Raman spectrum after Savitzky-Golay processing

2.4　光谱特征降维

本实验使用拉曼光谱仪采集数据时，拉曼光谱的拉曼偏移范围为 500～1600 cm^{-1}。拉曼光谱数据具有波段范围广、数据冗余度高等特点。如果对原始高维数据直接进行定量与定性分析，就很有可能使分析结果的误差比较大。主成分分析可以将数据从 N 维降低到 M 维，此时需要找到 M 个向量用于投影原始数据，使投影误差（投影距离）最小。因此，可以对原始数据进行主成分分析，这样就可以使用具有较少维度且不相关的数据来取代原始的高维数据，然后用变换后的数据进行建模。投影误差表达式为

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\| x^{(i)}-x_{\text{approx}}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\| x^{(i)}\|^2}\leqslant 0.01, \tag{2}$$

式中：m 表示特征的个数。

对经归一化处理和平滑去噪处理后的 132 个拉曼光谱数据进行主成分分析降维，得到帕累托图（Pareto chart），如图 4 所示。从帕累托图中可以看到，当保留 9 个主成分时，特征贡献率为 99.058%，之后每增加一个主成分，其贡献率增加不足 0.5%。所以，本文在计算中采用前 9 个主成分。

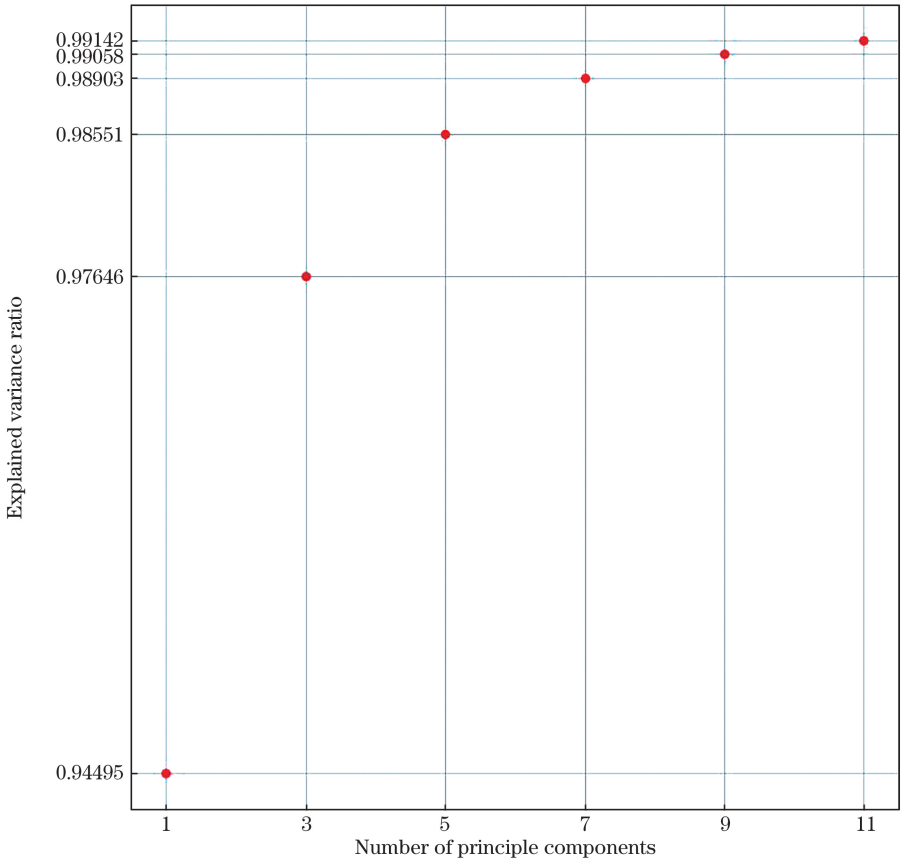

图 4　主成分帕累托图

Fig. 4　Pareto chart of principle components

3 实验与讨论

3.1 随机森林算法

集成学习就是构建并结合多个机器学习器来完成学习任务，从而减小单个分类器的误差，使分类的准确率较高。主要的集成学习算法有 Stacking、随机森林及 Adaboost 等。随机森林是一种集成学习算法，属于集成学习算法中弱学习器之间不存在依赖关系的一种算法，它利用多棵决策树对样本进行训练，每一棵决策树相当于一个专家。该算法就相当于若干个专家对某个任务进行决策分类[23]。

随机森林算法的架构如图 5 所示。

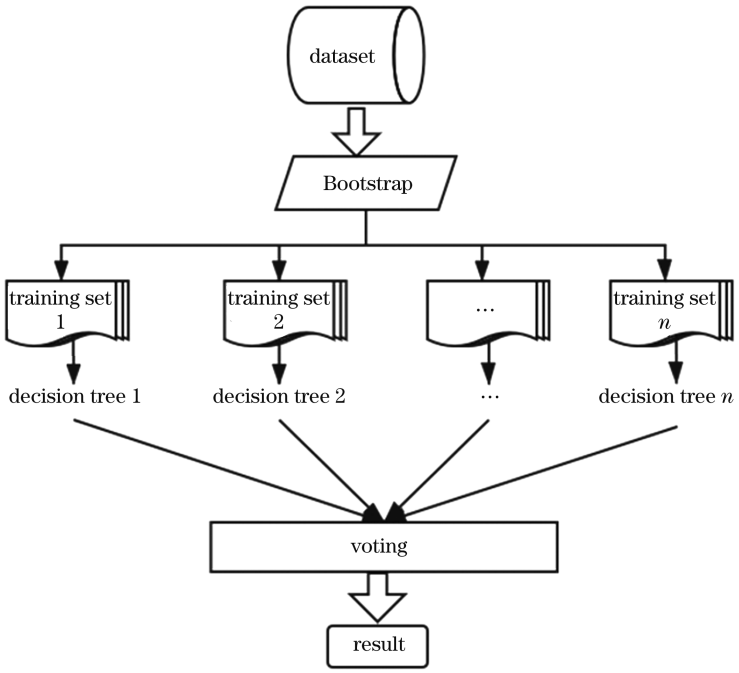

图 5 随机森林算法架构图

Fig. 5 Frame of random forest algorithm

食源性致病菌分类模型的构建步骤如下：

1)使用拉曼光谱仪检测拉曼偏移范围为 500～1600 cm^{-1} 的样本，并使用 LabSpec6.0 软件进行光谱数据的采集，将数据保存在 CSV 文件中；

2)加载原始数据并利用 Bootstrap(它是一种从给定训练集中有放回的均匀抽样)进行有放回的随机重抽样，产生独立同分布子集；

3)计算每个特征属性的 Gini 值，对节点进行排序并分配节点权重；

4)随机抽取特征并计算特征蕴藏的信息量，从随机抽取的特征中选择最具分裂能力的特征进行分裂；

5)根据决策树算法构建多棵决策树(为了防止过拟合，根据 Gini 值寻找特征分割点，并根据特征分割点划分数据子集)；

6)将生成的决策树组成随机森林，使用组成的森林进行决策分类，最终的结果使用投票法(voting)决定；

7)对生成的随机森林模型进行网格搜索和交叉验证，以获得模型的最佳参数。

对于决策树算法，为了达到分类预测的目的，需要对目标进行多个预测并计算其出现的概率。因此，在决策树中将叶节点的不纯度(即 Gini 值)作为二元分割的标准。

假设有 K 个类别，第 K 个类别的概率为 p_k，概率分布的 Gini 系数表达式为

$$\mathrm{Gini}(p)=\sum_{k=1}^{K}p_k(1-p_k)。\tag{3}$$

根据 Gini 指数对样本进行分割，最后将样本分成不同的子节点，每个叶节点对应一个预测结果。这就是构建决策树算法的思想与流程。决策树算法的具体流程如表 2 所示。

表 2 决策树算法的工作流程

Table 2 Work process of decision tree algorithm

Decision tree algorithm

Input: sample X, sample numbers N, feature counts M
Output: Decision Tree model
$X\rightarrow$for bagging//processing X with bagging cycles
end for
while extracting n_{try} ($n_{try}=N$)$\rightarrow X_{train}$ do
$M\rightarrow m_{try}$ ($m_{try}\ll M$)// random selection of m_{try} attributes
$m_{try}\rightarrow$the best node
$X\rightarrow X_{samples}$// build samples using Bootstrap
end while
for ($i_{tree}=0$; $1<i_{tree}\leqslant N_{tree}$; $i_{tree}++$)
// node splitting by optimal attributes to generate decision trees
end for
end procedure

随机森林算法集成了多棵决策树，能够对测试数据进行分类，比单一的弱分类器具有更强的分类效果和泛化能力。集成决策树之后，每个决策树对目标进行独立的预测，然后对决策树的预测结果进行投票，得到最终的预测结果。食源性致病菌的预测算法如表 3 所示。

表 3 食源性致病菌预测算法的流程

Table 3 Process of algorithm used for prediction of food-borne pathogenic bacteria

Food-borne pathogenic bacteria prediction algorithm

Input: sample X, training set X_{train}, test set X_{test}
Output: K trees, prediction result r
for all $i = 1$ to K do
while $j\leqslant N$ do
$row_{sample}=row_{sample}+select(X_{train})$
$j++$
end while
while stop condition not true do
$col_{sample}-select(row_{sample})$
split_Attribute$-\min\{Gini(col_{sample})\}$
// classification attributes are determined by the minimum Gini value
tree$-$AddNode(split_Attribute)
end while
leaf_node$\leftarrow$node
end for
for all $i-1$ to K do
$R_i=T_i$_Predict(D_{test})
$R=$MostCommon(R_i)
end for
end procedure

3.2 模型的构建与训练

本文将随机森林与拉曼光谱结合起来构建食源性致病菌分类模型，主要包括如下几个核心思想：

1）对原始拉曼光谱数据进行数据预处理；

2）将多个准确率较低的决策树模型进行集成；

3）对决策树输出的类别标签进行投票，决定输出的类别标签；

4）用 pyhton 调用随机森林库，自动多线程运行 CPU；

5）使用 GridSearchCV 进行网格搜索，选择最优参数。

在模型训练过程中使用网格搜索进行参数调优。模型中的参数有 criterion、n_estimators 和 max_depth。其中：criterion 是划分决策树时对特征的评价标准，默认是系数 Gini；n_estimators 表示弱学习器的最大迭代次数，若值太小容易欠拟合，过大则容易过拟合；max_depth 表示决策树的最大深度。

将处理好的数据按照 3∶7 的比例划分为测试集和训练集。n_estimators 的范围设定为[0，200]，max_depth 范围设定为[0，100]。将上述参数作为网格搜索参数来训练模型。图 6 和图 7 主要展示了对

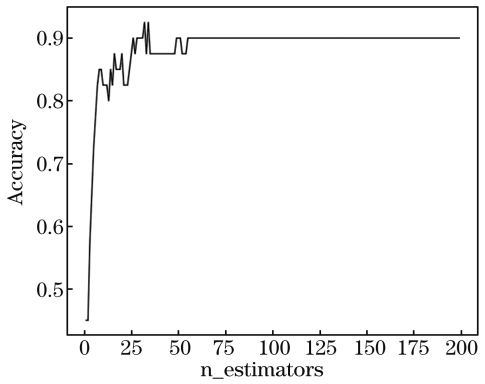

图 6 模型准确率随 n_estimators 的变化

Fig. 6 Model accuracy change with n_estimators

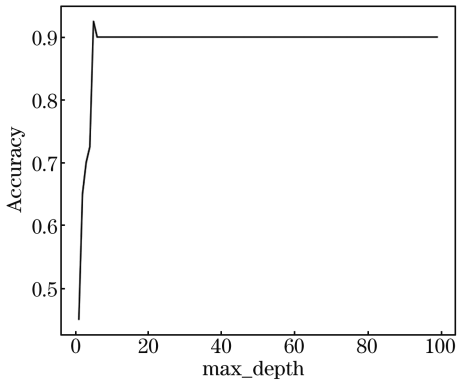

图 7 模型准确率随 max_depth 的变化

Fig. 7 Model accuracy change with max_depth

n_estimators 和 max_depth 这两个参数的优化。可以看出，当参数 n_estimators 取 65，参数 max_depth 取 8 时，模型的准确率最高，且模型较为稳定。

3.3 模型效果评估

为了对模型的精确性进行评估，本文对随机森林模型进行 K 折交叉验证。K 折交叉验证可以有效地避免过拟合与欠拟合的发生，最终得到的结果也比较具有说服性。

K 折交叉验证就是将数据集进行分层取样，将数据集划分为 K 个大小相似的互斥子集，并将其中 $K-1$ 个子集作为训练集，剩下的 1 个子集作为测试集进行试验。这样做就可以得到 K 个训练/测试集，每一组测试均可得到一个结果，从而可以得到 K 个结果，对这 K 个结果取平均值就可得到 K 折交叉验证的最终结果。本实验选取的 K 值为 10。图 8 为 10 折交叉验证示意图。

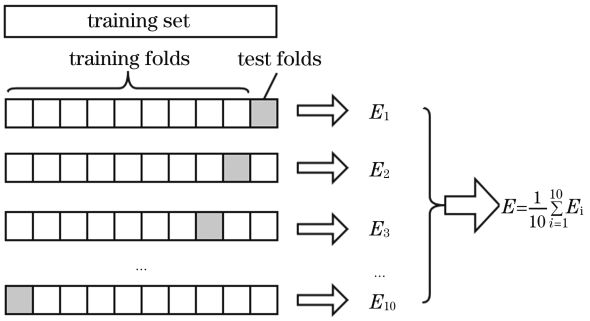

图 8 10 折交叉验证示意图

Fig. 8 10-fold cross-validation diagram

确定随机森林模型的参数后，可以建立定性鉴别模型。本文将本次实验训练得到的随机森林模型和一些常见的传统分类算法的精度进行了比较，如表 4 所示。

表 4 各模型的精度

Table 4 Accuracy of each model

Model	Accuracy /%
PCA + KNN(K-nearest neighbors)	88.19
PCA + logistic regression	88.25
PCA + SVM(support vector machines)	83.86
PCA + decision tree	82.63
PCA + RF(our)	**91.36**

从表 4 中可以看出，在 10 折交叉验证的模型中，随机森林模型的准确率要优于传统的机器学习算法。决策树模型的表现最差，准确率为 82.63%。这是因为与随机森林算法相比，决策树为单一的弱学习器，而随机森林则是将多个决策树模型进行组合投票形成了强学习器，所以随机森林较单一的决策树分类器具有更高的分类能力。

与传统的机器学习算法相比，随机森林算法在模型构建部分加入了两种随机性，即随机性抽取样本和随机性选择特征。由于随机森林由决策树组成，所以决策树的相关性越大，错误率就越大。随机抽取样本决定了随机森林中每棵树的相关性较小。随机森林中的每棵树随机选用部分特征，在少量的特征中选择最优分裂能力的特征作为决策树左右子树划分的依据，将随机性的效果扩大，进一步增强了模型的鲁棒性。两种随机性的引入对于降低模型的方差具有重要作用，故随机森林一般不需要额外进行剪枝即可获得较好的泛化能力和抗过拟合能力。另外，由于拉曼光谱数据预处理阶段使用 Savitzky-Golay 滤波算法进行去噪，故模型的抗干扰能力较强。

4 结 论

本文利用随机森林算法对拉曼光谱仪采集到的 11 种食源性致病菌的 132 个光谱数据样本进行了分类预测，达到了预期的效果。

本研究构建了适用于对食源性致病菌拉曼光谱进行鉴定分析的方法，结果表明：本文提出的主成分分析结合随机森林算法的分类模型具有比传统的单一机器学习方法更高的准确性，这为食源性致病菌的检测提供了新的检测方法，提高了人工识别拉曼光谱的速度。

随机森林模型处理噪声比较大的样本集时会很容易出现过拟合。为了提高模型的精度，可以对数据预处理阶段的去噪声处理进行优化，或者在随机森林算法中对数据特征的选择进行优化。本研究仅使用了 11 种食源性致病菌的样本，在后期的模型构建中拟引入更多的样本，以构建更加完整的拉曼光谱数据库。

参 考 文 献

[1] Gao Y, Yin X B, Wang T. PCR assay for enteropathogenic bacteria and evaluation of its application value[J]. Capital Food Medicine, 2019, 26(22): 101.
高扬, 尹啸冰, 王彤. 肠道致病菌 PCR 检测及应用价值评估[J]. 首都食品与医药, 2019, 26(22): 101.

[2] Pannetier C. PCR[J]. Immunology Today, 1996, 17(12): 590.

[3] Vinner L, Fomsgaard A. Inactivation of orthopoxvirus for diagnostic PCR analysis[J]. Journal of Virological Methods, 2007, 146(1/2): 401-404.

[4] Zhang M. Rapid identification of species′ blood based on Raman spectroscopy[D]. Nanchang: Nanchang University, 2018.
张铭. 基于拉曼光谱实现物种血液的快速鉴别研究[D]. 南昌: 南昌大学, 2018.

[5] McLaughlin G, Doty K C, Lednev I K. Raman spectroscopy of blood for species identification[J]. Analytical Chemistry, 2014, 86(23): 11628-11633.

[6] Wang S, Haishan Z. Real-time *in vivo* Raman spectroscopy and its clinical applications in early cancer detection[J]. Chinese Journal of Lasers, 2018, 45(2): 0207002.
王爽, Haishan Zeng. 实时拉曼光谱分析技术及其在临床早期癌症检测中的应用[J]. 中国激光, 2018, 45(2): 0207002.

[7] Zhang Y J, Zhang F C, Fu X H, et al. Detection of fatty acid content in mixed oil by Raman spectroscopy based on ABC-SVR algorithm[J]. Spectroscopy and Spectral Analysis, 2019, 39(7): 2147-2152.
张燕君, 张芳草, 付兴虎, 等. 基于 ABC-SVR 算法的拉曼光谱检测混合油脂肪酸含量[J]. 光谱学与光谱分析, 2019, 39(7): 2147-2152.

[8] Wu C W, Shi R J, Zeng W D. Mineral Raman spectral recognition based on Siamese network[J]. Laser & Optoelectronics Progress, 2020, 57(9): 093301.
吴承炜, 史如晋, 曾万聃. 基于 Siamese 网络的矿物拉曼光谱识别[J]. 激光与光电子学进展, 2020, 57(9): 093301.

[9] Žuvela P, Lin K, Shu C, et al. Fiber-optic Raman spectroscopy with nature-inspired genetic algorithms enhances real-time *in vivo* detection and diagnosis of nasopharyngeal carcinoma[J]. Analytical Chemistry, 2019, 91(13): 8101-8108.

[10] de Souza Lins Borba F, Saldanha Honorato R, de Juan A N. Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks[J]. Forensic Science International, 2015, 249: 73-82.

[11] Fang K N, Wu J B, Zhu J P, et al. A review of technologies on random forests[J]. Statistics & Information Forum, 2011, 26(3): 32-38.
方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[12] Ma L. Research on optimization and improvement of random forests algorithm[D]. Guangzhou: Jinan University, 2016.
马骊. 随机森林算法的优化改进研究[D]. 广州: 暨南大学, 2016.

[13] Xie J F, Luo J, Xu M, et al. Study on identifi cation of 100% cotton textile by Raman spectroscopy and random forest method[J]. China Fiber Inspection, 2014(22): 76-78.
谢剑飞, 罗峻, 许敏, 等. 拉曼光谱结合随机森林方法应用于全棉纺织品真伪鉴别的研究[J]. 中国纤检, 2014(22): 76-78.

[14] Vigneau E, Courcoux P, Symoneaux R, et al. Random forests: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception[J]. Food Quality and Preference, 2018, 68: 135-145.

[15] Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model[J]. International Journal of Medical Informatics, 2019, 125: 55-61.

[16] Huang J H, Xie H L, Yan J, et al. Using random forest to classify T-cell epitopes based on amino acid properties and molecular features[J]. Analytica Chimica Acta, 2013, 804: 70-75.

[17] Shi R J, Xia F Z, Zeng W D, et al. Raman spectroscopic classification of foodborne pathogenic bacteria based on PCA-stacking model[J]. Laser & Optoelectronics Progress, 2019, 56(4): 043003.
史如晋, 夏钒曾, 曾万聃, 等. 基于 PCA-Stacking 模型的食源性致病菌拉曼光谱识别[J]. 激光与光电子学进展, 2019, 56(4): 043003.

[18] Han X H, Zhang Y H, Sun F J, et al. Method for determining index weight based on principal component analysis[J]. Journal of Sichuan Ordnance, 2012, 33(10): 124-126.
韩小孩, 张耀辉, 孙福军, 等. 基于主成分分析的指标权重确定方法[J]. 四川兵工学报, 2012, 33(10): 124-126.

[19] Li X R. Compare and application of principal component analysis, factor analysis and clustering analysis[J]. Journal of Shandong Education Institute, 2007, 22(6): 23-26.
李新蕊. 主成分分析、因子分析、聚类分析的比较与应用[J]. 山东教育学院学报, 2007, 22(6): 23-26.

[20] Lei L P. Curve smooth denoising based on Savitzky-Golay algorithm[J]. Computer and Information Technology, 2014, 22(5): 30-31.
雷林平. 基于 Savitzky-Golay 算法的曲线平滑去噪[J]. 电脑与信息技术, 2014, 22(5): 30-31.

[21] Zhu L L, Feng A M, Jin S Z, et al. Fluorescence suppression methods in Raman spectroscopy detection and their application analysis[J]. Laser & Optoelectronics Progress, 2018, 55(9): 090005.
朱磊磊, 冯爱明, 金尚忠, 等. 拉曼光谱检测中荧光抑制方法及其应用分析[J]. 激光与光电子学进展,

2018, 55(9): 090005.

[22] Hu J X, Zhang G J. K-fold cross-validation based selected ensemble classification algorithm [J]. Bulletin of Science and Technology, 2013, 29(12): 115-117.
胡局新, 张功杰. 基于 K 折交叉验证的选择性集成分类算法[J]. 科技通报, 2013, 29(12): 115-117.

[23] Bao Q L, Ding J L, Wang J Z. Prediction of soil moisture content by selecting spectral characteristics using random forest method [J]. Laser & Optoelectronics Progress, 2018, 55(11): 113002.
包青岭, 丁建丽, 王敬哲. 利用随机森林方法优选光谱特征预测土壤水分含量[J]. 激光与光电子学进展, 2018, 55(11): 113002.

Recognition of Food-Borne Pathogenic Bacteria by Raman Spectroscopy Based on Random Forest Algorithm

Wang Qi[1], Zeng Wandan[1*], Xia Zhiping[2**], Li Zhiping[2], Qu Han[2]

[1] *College of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai* 201418, *China*;

[2] *Military Veterinary Institute, Changchun, Jilin* 130062, *China*

Abstract

Objective Food and drug safety is of great concern to society. Food pathogenic bacteria are pathogenic bacteria that can cause food poisoning or bacteria that use food as the vector of transmission. Therefore, quick and effective detection of food-borne pathogenic bacteria in food is crucial to protect public health. The culture separation method, which is traditionally used to examine microorganisms, depends on the medium used for culturing, separation, and biochemical identification. Detection of food-borne pathogenic bacteria generally requires five to seven days and includes a series of detection procedures such as pre-enrichment, selective enrichment, microscopic examination and serological verification. Therefore, traditional detection methods are insufficient for preventing and controlling food-borne pathogenic bacteria. However, Raman spectroscopy is a nondestructive method that can be used to rapidly and accurately identify molecules existing in the functional groups. In this study, 11 food-borne pathogenic bacteria samples were used to construct a recognition and classification model based on a random forest algorithm and Raman spectra. This model was then used to build a classification and recognition model to resolve the problems of low classification accuracy and long detection time required by traditional methods used to detect food-borne pathogenic bacteria. The results of this study will help to ensure public health safety by rapidly and effectively detecting pathogens in food and drugs.

Methods All of the food-borne pathogenic bacteria in this study were purchased from China Center of Industrial Culture Collection. First, a sample of food-borne pathogenic bacteria was detected by Raman spectrometry in a shift range of 500–1600 cm^{-1}. LabSpec 6.0 software was used for spectral collection, and each sample was collected 15 times. After screening, 132 Raman spectral data were obtained. Min-max normalization was performed on the Raman spectral data in the spectral preprocessing stage, and the intensity was mapped to a range of [0, 1] for comparison. The Savitzky-Golay algorithm was used for smooth denoising to remove noise and fluorescence interference. Principal component analysis (PCA) was used for feature dimensionality reduction for sample data with high-dimensional characteristics to avoid problems caused by excessively high dimensions. In the model evaluation stage, *K*-fold cross-validation was used to verify whether the model balanced underfitting and overfitting phenomena and to evaluate the model stability. According to these criteria, the Raman spectral recognition model based on the random forest algorithm proposed in this study was able to effectively distinguish different food-borne pathogenic bacteria among the collected samples.

Results and Discussions In this study, *K*-nearest neighbors (KNN), logistic regression, support vector machine (SVM), decision tree, and random forest models were used for classification prediction of the pre-treated Raman spectral data of the food-borne pathogenic bacteria (Table 4). Among the 10-fold cross-validation models, the accuracy of the random forest model was better than that of the traditional machine learning algorithms. The decision tree model presented the worst results, with an accuracy rate of 82.63%. This is because the decision tree results in a single weak learner, whereas the random forest model includes multiple votes that are combined to form strong learning (Fig. 5). Therefore, the classification ability of the random forest algorithm is higher than that of a single

decision tree classifier. Compared with traditional machine learning algorithms, the random forest algorithm adds two randomness elements in the model construction: sampling randomness and feature selection randomness (Table 2). Because the random forest is composed of decision trees, a higher correlation of decision trees results in a higher error rate. Random sampling determines the decrease degree in the correlation of each tree in the random forest. Among a small number of features selected randomly by each tree in the random forest, the features of optimal splitting ability are chosen as the left and right subtrees of the decision tree. This expands the effect of randomness and further enhances the robustness of the model. Because the introduction of the two randomness elements has a strong effect on reducing the variance of the model, the random forest generally does not need additional pruning. That is, it can achieve better generalization and a stronger ability to avoid overfitting, resulting in low variance. In addition, the Savitzky-Golay filtering algorithm was used for denoising in the preprocessing stage of the Raman spectral data (Fig. 3) to ensure good anti-interference ability in the model.

Conclusions Raman spectroscopy is a mature technology that has a significant effect on the detection and classification of food-borne pathogenic bacteria. In this study, a Raman spectrometer was used to detect the spectral data of 11 food-borne pathogens. According to the spectral properties, the spectral data were normalized, smoothed, and denoised in the preprocessing stage, which facilitated the model construction and training. In addition, a method was developed for identification and analysis of food-borne pathogenic bacteria by using Raman spectroscopy. The experimental results show that the classification model of PCA combined with the random forest algorithm proposed in this study has higher accuracy for Raman spectral data than that of the single machine learning method used conventionally for detecting food-borne pathogens. In addition, the new method improves the speed of manual identification of the Raman spectra. However, the random forest model was prone to overfitting in the sample sets with large noise processing. Future research to improve the accuracy of the model will show that denoising can be optimized in the data pretreatment stage and that the data feature selection algorithm can be optimized using the random forest algorithm. Only 11 samples of food-borne pathogenic bacteria were used in this study. Additional samples could be introduced in the construction of a later model to build a more complete Raman spectral database.

Key words Raman spectroscopy; machine learning; food-borne pathogen detection; principle component analysis; random forest Spectroscopy

OCIS codes 330.6230; 300.6450; 240.6695