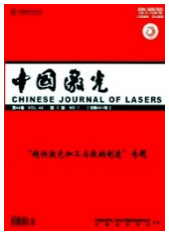

中国激光
Chinese Journal of Lasers
ISSN 0258-7025,CN 31-1339/TN

《中国激光》网络首发论文

题目：基于随机森林算法的食源性致病菌拉曼光谱识别
作者：王其，曾万聃，夏志平，李志萍，曲晗
收稿日期：2020-07-06
网络首发日期：2020-09-24
引用格式：王其，曾万聃，夏志平，李志萍，曲晗．基于随机森林算法的食源性致病菌拉曼光谱识别[J/OL]．中国激光．
https://kns.cnki.net/kcms/detail/31.1339.TN.20200923.1208.006.html

网络首发时间：2020-09-24 11:07:26
网络首发地址：https://kns.cnki.net/kcms/detail/31.1339.TN.20200923.1208.006.html

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，拉曼光谱具有检测范围广、检测灵活、光谱的特征突出等特点。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 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

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。当前，食源性致病菌检测的方法主要包括：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）和增光后实时 PCR 法等[1,2]。这些方法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]。拉曼光谱反映分子内部振动和转动能级[4-6]，是物质指纹谱，可以用来鉴定分子中存在的官能团，具有无损、快速、准确等特点，是物质成分判别的有力工具。

拉曼光谱方法结合机器学习算法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检测周期，大大降低了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）的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]提出一种基于拉曼光谱和 Siamese 网络的相似性学习方法，能够对矿物进行识别。Žuvela Petar 等[9]使用自然遗传算法开发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 Flávia 等[10] 研究了 14 种不同品牌、不同型号的商用蓝色圆珠笔墨水，在 A4 亚硫酸盐纸上的墨线上获得了拉曼光谱，建立了基于偏最小二乘判别分析（PLS-DA）的层次分类模型。证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工具，可以区分文档中非常相似的墨水类型。

随机森林算法[11-13]（Random Forest，RF）是一种集成学习（Ensemble Learning）方法。E. Vigneau 等[14]将随机森林应用在感官分析中，实验观察到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。Ke Lin 等[15]利用随机森林（RF）算法建立重症监护病房（ICU）急性肾损伤（AKI）患者的死亡率预测模型，并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 模型进行比较，实验证明 RF 模型有助于 ICU 临床医生及时做出 AKI 患者的临床干预决策，对降低 AKI 患者的院内死亡率具有重要意义。Jian-Hua Huang 等[16]采用随机森林（RF）算法对 T 细胞表位和非 T 细胞表位进行分类。结果表明，目前基于特征和 RF 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，成功的将大肠杆菌 O157:H7 以及布鲁氏菌 S2 株分离开。但是所研究的食源性致病菌类别数只有 2 个，并不能满足实际的工业生产需求。本文在此基础上将类别数增加到 11 个，样本数 132 个，平均每类样本数有 10 个左右。这样就增加了训练和分类的难度，但是更加符合实际的生产需要。

本文提出了一种基于主成分分析[18,19]（Principal Component Analysis，PCA）结合随机森林算法的拉曼光谱识别模型。对使用拉曼光谱仪收集到的拉曼数据，光谱预处理阶段，本实验中使用 Min-Max 对其进行归一化处理；Savitzky-Golay 算法[20,21]进行平滑去噪；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。模型评估阶段，本文使用 K 折交叉验证[22]（K-fold cross validation，K-CV）对模型进行评估。基于此，证明了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 实验样本采集

本文的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[1]（China Center of Industrial Culture Collection, CICC），如表 1 所示。本实验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，测量的拉曼偏移范围为 500-2000cm^{-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 CICC 编号与名称

Table 1 11 foodborne pathogenic bacteria CICC numbers and names

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是种类相对较少，所以人工识别拉曼光谱的难度相对较大。本文将以肠沙门氏菌肠亚种鼠伤寒血清型(Cronobacter sakazakii)为样本进行数据预处理。图 1 为原始拉曼光谱图。

[1] http://cicc.china-cicc.org/

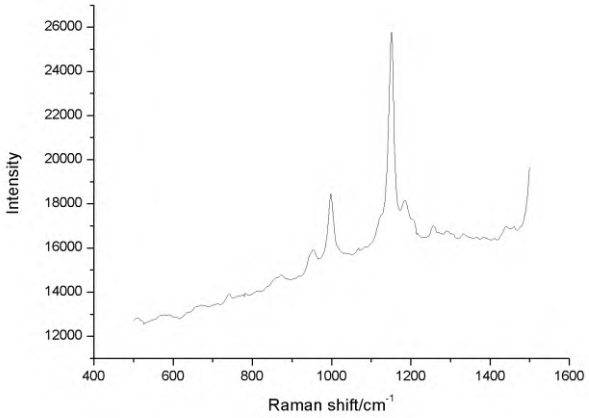

图 1 原始拉曼光谱

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发现：不同拉曼偏移值所对应的强度差异化比较大。当把不同的特征列在一起的时候，由于特征本身表达方式的原因会导致绝对数值大的数据的重要性大于绝对数值小的数据。这时我们需要对抽取出来的特征向量进行归一化处理，用来保证每个特征被分类器平等对待，让数据的处理保持一致。下面对原始数据使用 Min-Max 归一化处理，并对数据进行可视化。

本文采用 Min-Max 标准化：

$$x_{normalization} = \frac{x - Min}{Max - Min}, \qquad (1)$$

其中 Max 为样本数据的最大值， Min 为样本数据的最小值， $x_{normalization}$ 为归一化后的样本值。归一化处理后的图 1 所示拉曼光谱如图 2 所示，将强度均映射到[0,1]之间，便于比较。

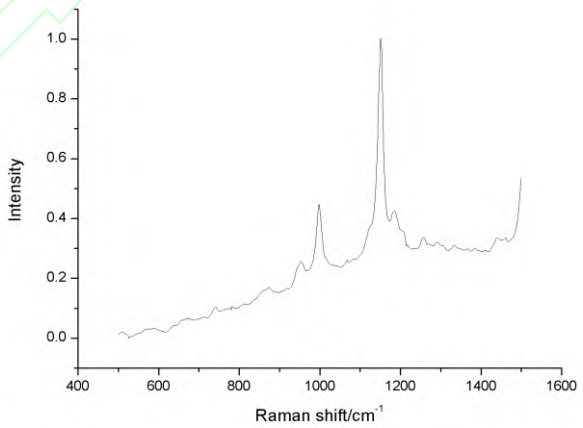

图 2 Min-Max 标准化处理后的拉曼光谱

Fig. 2 Raman spectra after Min-Max 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影响。收集到的拉曼光谱数据总会带有一些噪声和荧光干扰，会在一定程度上影响到光谱的质量。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 去噪。

本文选择的窗口宽度为 27，多项式阶数为 2。图 3 为图 2 经过 Savitzky-Golay 平滑去噪拉曼光谱图像，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。

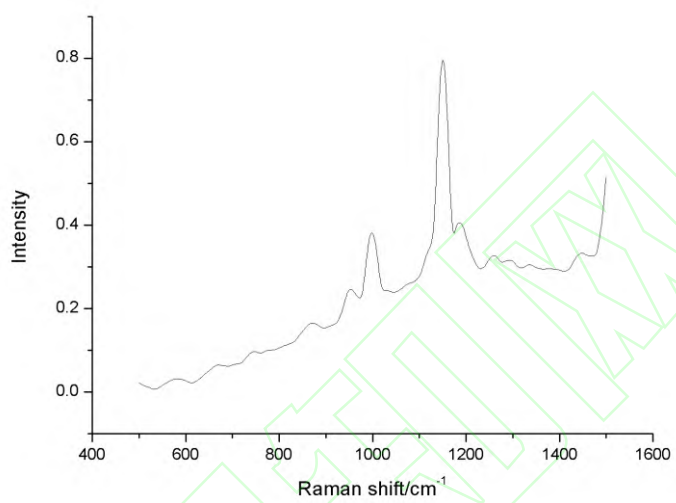

图 3 Savitzky-Golay 处理后的拉曼光谱

Fig. 3 Raman spectra after Savitzky-Golay smooth denoising

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 500～2000cm^{-1}。拉曼光谱数据具有波段范围广、数据具有很高的冗余度等特点。如果对原始高维数据直接进行定量与定性分析则很可能使得分析结果误差比较大。主成分分析是为了将数据从 N 维降低到 K 维，需要找到 K 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，使用较少的维度并且不相关的数据来取代原始的高维数据，使用变换后的数据进行建模。式（2）为投影误差表达式。

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

其中 m 表示特征个数。

对归一化处理和去噪平滑处理后的 132 组拉曼光谱数据进行主成分分析降维，得到其帕累托图（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%，所以计算中采用前 9 个主成分。

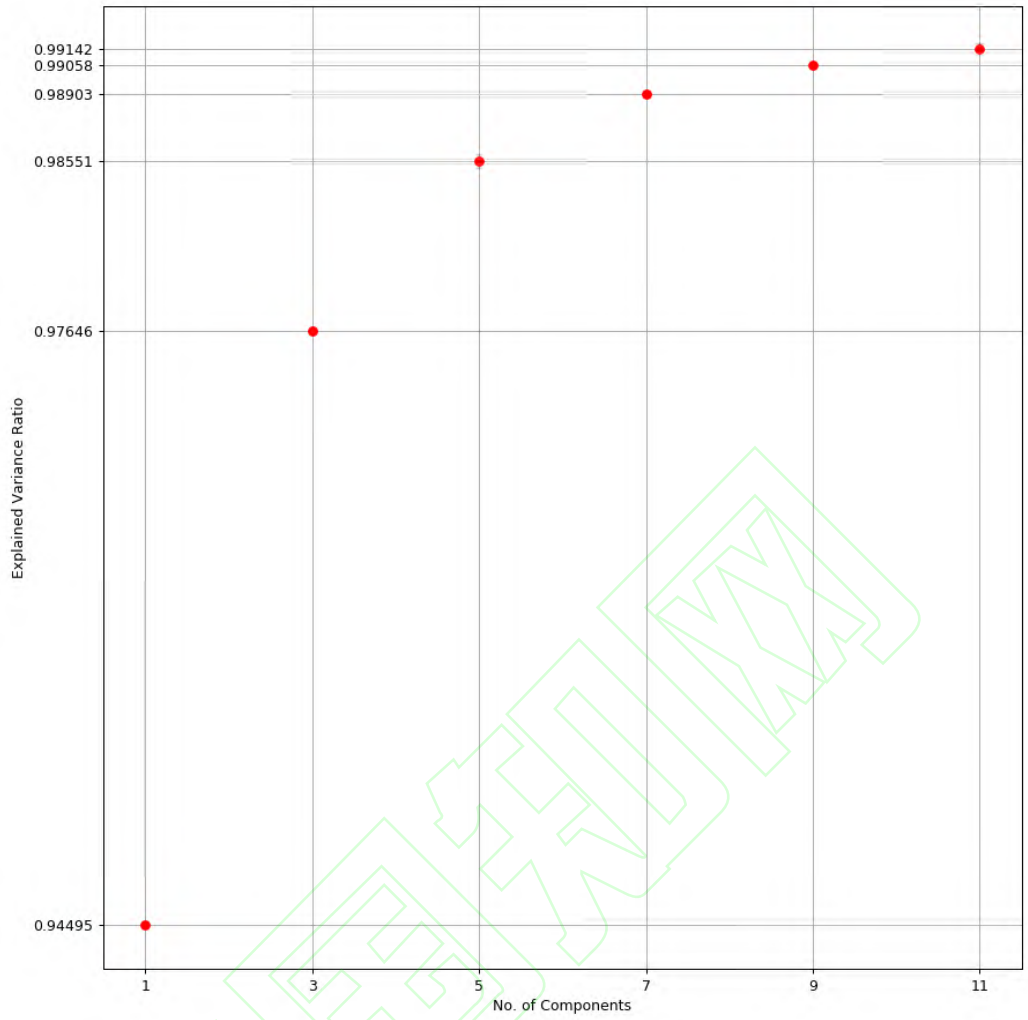

图 4 主成分帕累托图

Fig. 4 Preto Chart of the components

3 实验与讨论

3.1 随机森林算法

集成学习是通过构建并结合多个机器学习器来完成学习任务，其中主要有 Stacking、随机森林以及 Adaboost 等，可以减少单个分类器的误差，分类所得到的准确率较高。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部分。它利用多颗决策树对样本进行训练，每一颗决策树相当于一个专家，随机森林就是若干个专家对某个任务进行决策分类[23]。

其算法架构如图 5 所示：

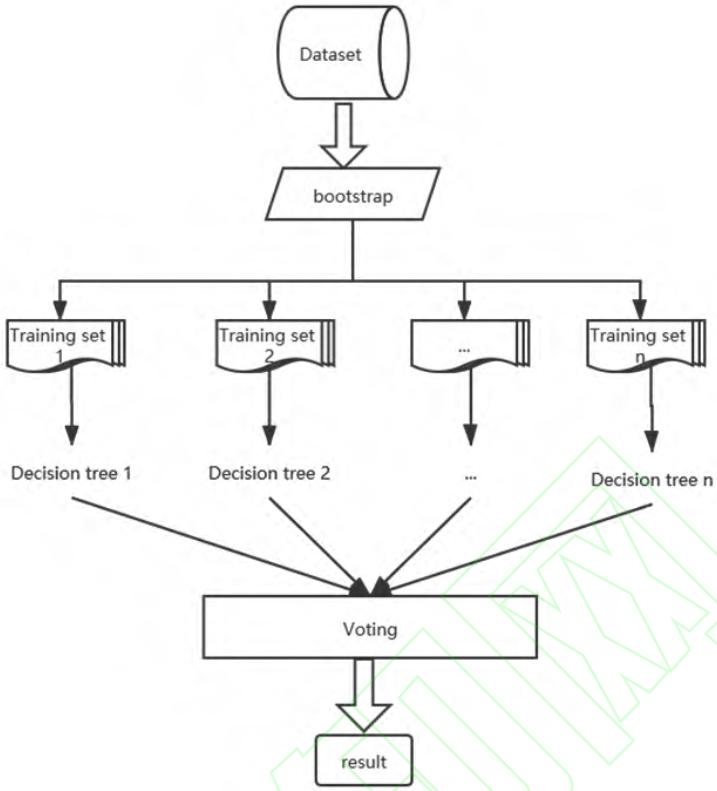

图 5 随机森林算法架构图

Fig. 5 Frame of Random Forest

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 500-2000cm^{-1} 的样本并使用 LabSpec6.0 软件进行光谱数据采集，将数据保存在 CSV 文件里。

2．加载原始数据并利用 Bootstrap（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，特征属性采样。

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有分裂能力的特征进行分裂。

5．根据决策树算法构建多个决策树，并且为了防止过拟合，采用控制 Gini 值变化的大小和节点样本数量等方法。

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来获得模型的最佳参数。

其中对于决策树算法，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。因此在决策树中定义叶节点的不纯度即 Gini 值来作为二元分割的标准。

假设 K 个类别，第 K 个类别的概率为 p_k，概率分布的基尼系数表达式：

$$Gini(p)=\sum_{K=1}^{K} p_k(1-p_k), \qquad (3)$$

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。由此提出决策树算法的建立。

表 2 决策树算法工作流程

Table 2 Process of Decision Tree

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 $(i_{tree} = 0; 1 < i_{tree} \leq N_{tree};\ i_{tree}++)$

// Node splitting by optimal attributes to generate decision trees

end for

end procedure

随机森林集成了多个决策树，从而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。本文的食源性致病菌预测算法如表 3 所示。

表 3 食源性致病菌预测算法工作流程

Table 3 Process of foodborne pathogenic bacteria prediction algorithm

Foodborne 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 $j \leq N$ **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**

$$R_i = T_i\_\Pr edict(D_{test})$$

$$R = MostCommon(R_i)$$

end for

end procedure

3.2 模型构建与训练

本文将随机森林与拉曼光谱结合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. 对于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多数投票决定输出的类别标签；
4. 使用 pyhton 调用 Random Forest 库自动运行 CPU 多线程；
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。模型中参数 criterion 即决策树做划分时对特征的评价标准，默认是系数 Gini。n_estimators 表示弱学习器最大迭代次数，若值太小容易欠拟合，过大又容易过拟合。max_depth 表示决策树最大深度。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 范围设定为[0,200]，max_depth 范围设定在[0,100]。将上述参数作为网格搜索参数训练模型。如图 6、图 7 所示，主要展示对 n_estimators 和 max_depth 的优化。在参数 n_estimators 取 65，参数 max_depth 取 8 是模型达到最高的准确率，且模型较为稳定。

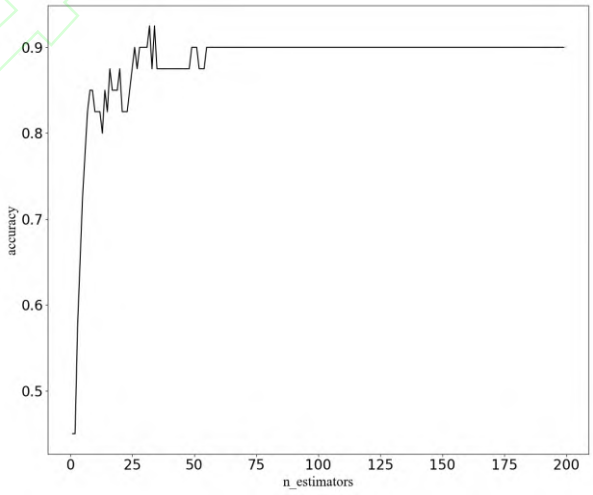

图 6 模型准确率随 n_estimators 变化图

Fig. 6 Model accuracy change with n_estimators

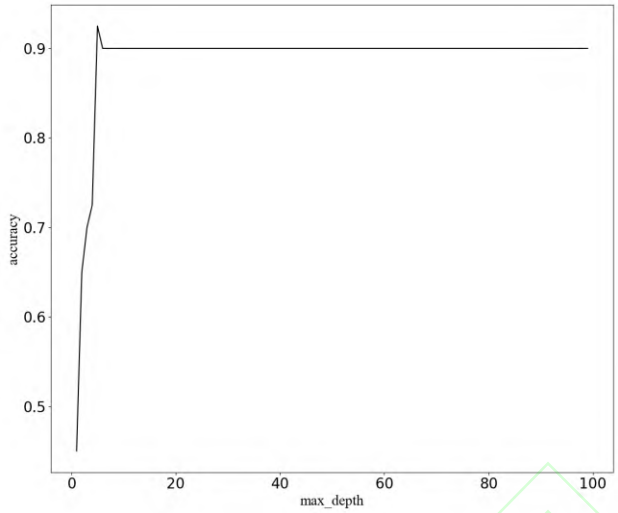

图 7 模型准确率随 max_depth 变化图

Fig. 7 Model accuracy change with max_depth

3.3 模型效果评估

对于训练出的随机森林模型进行 K 折交叉验证，可对模型精确性进行评估。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 折交叉验证是将数据集进行分层取样，划分为 *K* 个大小相似的互斥子集。将 *K*-1 个子集作为训练集，剩下的 1 个作为测试集进行试验，这样就可以得到 *K* 个训练/测试集，每一组测试可得到一个结果，可得到 *K* 个结果，对其取平均值可得到 *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 折交叉验证的示意图。

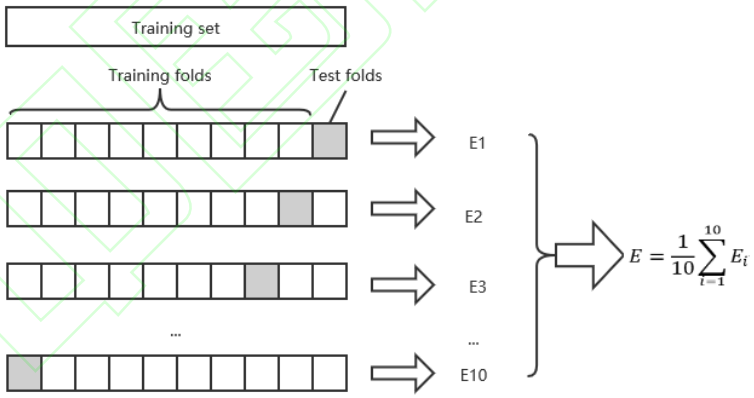

图 8 10 折交叉验证示意图

Fig. 8 10 fold cross validation diagram

确定随机森林的参数之后，可以建立定性鉴别模型。经过 10 折交叉验证，本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相比，分类结果如表 4 所示。

表 4 分类结果

Table 4 Result of classification

Model	Accuracy/%
PCA+KNN	88.19
PCA+Logistic Regression	88.25
PCA+SVM	83.86
PCA+Decision Tree	82.63
PCA+RF	**91.36**

从表4中可以看出，本研究对预处理后的食源性致病菌拉曼光谱数据，分别使用了KNN、Logistic Regression、SVM、决策树和随机森林模型进行分类预测。10 折交叉验证的模型中随机森林模型准确率要优于传统的机器学习算法。决策树模型表现的结果最差，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的部分加入了两个随机性：抽取样本随机和特征选择随机。由于随机森林由决策树组成，所以决策树的相关性越大，错误率越大。随机抽取样本决定了随机森林的每棵树的相关性的减小。随机森林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的左右子树划分，将随机性的效果扩大，这进一步增强了模型的鲁棒性。由于两个随机性的引入，对于降低模型的方差很有作用，故随机森林一般不需要额外做剪枝，即可以取得较好的泛化能力和抗过拟合能力（Low Variance）。另外由于拉曼光谱数据预处理阶段使用了 Savitzky-Golay 滤波算法进行去噪，使得模型较好的抗干扰能力。

4 结　　论

本文对使用拉曼光谱仪采集到的 11 种食源性致病菌的 132 个光谱数据利用随机森林算法对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食源性致病菌拉曼光谱鉴定分析的方法，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对于拉曼光谱数据具有比常见的传统单一的机器学习方法更高的准确性，这为食源性致病菌检测提供了另一种检测方法，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上会很容易出现过拟合的状态。在之后的研究中，数据预处理阶段我们可以对其去噪声处理进行优化，随机森林算法中可以对数据特征选择进行优化以提高模型的精确度。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可引入更多的样本用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 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] Christophe Pannetier. PCR[J]. Immunology Today,1996,17(12).

[3] Lasse Vinner,Anders Fomsgaard. Inactivation of orthopoxvirus for diagnostic PCR analysis[J]. Journal of Virological Methods,2007,146(1).
[4] Zhang M. Rapid identification of species' blood based on Raman spectroscopy[D]. Nanchang University,2018:09-30.张铭. 基于拉曼光谱实现物种血液的快速鉴别研究[D]. 南昌大学, 2018:09-30.
[5] McLaughlin Gregory,Doty Kyle C,Lednev Igor K. Raman spectroscopy of blood for species identification.[J]. Analytical chemistry,2014,86(23).
[6] Wang S,Zeng H. Real-time Raman spectroscopy and its application in early cancer detection.[J]. Chinese Journal of Lasers, 2018, 45(002):0207002.王爽, Zeng H . 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(002):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(07):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(07):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(09):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(09):093301.
[9] Žuvela Petar,Lin Kan,Shu Chi,Zheng Wei,Lim Chwee Ming,Huang Zhiwei. 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).
[10] de Souza Lins Borba Flávia,Honorato Ricardo Saldanha,de Juan Anna. Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks.[J]. Forensic science international,2015,249.
[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(03):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(03):32-38.
[12] Ma L. Research on Optimization and Improvement of Random Forest Algorithm[D]. Jinan University,2016:05-19.马骊. 随机森林算法的优化改进研究[D]. 暨南大学, 2016:05-19.
[13] Xie J F,Luo J,Xu M,et al. Study on Identification 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] E. Vigneau,P. Courcoux,R. Symoneaux,L. Guérin,A. Villière. Random Forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception[J]. Food Quality and Preference,2018,68.
[15] Ke Lin,Yonghua Hu,Guilan Kong. 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.
[16] Jian-Hua Huang,Hua-Lin Xie,Jun Yan,Hong-Mei Lu,Qing-Song Xu,Yi-Zeng Liang. Using Random Forest to classify T-cell epitopes based on amino acid properties and molecular features[J]. Analytica Chimica Acta,2013,804.
[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(04): 043003-1-043003-6. 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(04): 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, 王少华. 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(06):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(06): 23-26.
[20] Lei L P. Curve Smooth Denoising Based on Savitzky-Golay Algorithm[J]. Computer Information Technology,2014,22(05):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(05): 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(09): 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(09): 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，Din J L，Wang J Z. Prediction of Soil Moisture Content by Selecting Spectral Characteristics Using Random Forest Method[J]. Laser & Optoelectronics Progress, 2018, 055(011):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， 055(011): 113002.

Recognition of Foodborne 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]*Shanghai Institute of Technology, College of Computer Science and Information Engineering, Shanghai 201418*

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

Abstract

Objective： Food and drug safety is an issue of great concern to the current society. Food pathogenic bacteria are pathogenic bacteria that can cause food poisoning or take food as the vector of transmission. Therefore, it is an urgent public health problem to quickly and effectively detect foodborne pathogenic bacteria in food. The traditional microorganism examination method is the culture separation method, depends on the medium to carry on the culture, the separation and the biochemical identification. The detection of foodborne pathogenic bacteria generally takes a long time, including a series of detection procedures such as progeria, selective germplasm, microscopic examination and serological verification, which take 5-7 days. Therefore, the traditional detection methods cannot meet the needs of prevention and control of foodborne pathogenic bacteria detection. Raman spectroscopy can be used to identify molecules that exist in the functional groups, has the characteristics of nondestructive, rapid and accurate. In this study, 11 foodborne pathogenic bacteria samples were used to construct a recognition and classification model based on Random Forest algorithm and Raman spectrum. This model aims to build a classification and recognition model different from traditional foodborne pathogen detection methods, so as to solve the problems of low classification accuracy and long detection time of foodborne pathogenic bacteria, and make contributions to food and drug detection and public health safety.

Methods： All the foodborne pathogenic bacteria in this study were purchased from China Center of Industrial Culture Collection. Firstly, the sample of foodborne pathogenic bacteria was detected by Raman spectrometer, and the measured Raman shift range was 500-2000cm^{-1}. LabSpec6.0 software was used for spectral collection, and each sample was collected 15 times. After screening, 132 Raman spectral data were obtained. For the Raman spectral data of foodborne pathogenic bacteria collected by the Raman spectrometer, in the spectral preprocessing stage, Min-Max was used in this experiment to normalize the data, and the intensity was mapped between [0,1] for comparison. Savitzky-Golay algorithm is used for smooth denoising to remove noise and fluorescence interference. Principal component analysis(PCA) is used for feature dimensionality reduction for sample data with high-dimensional characteristics to avoid dimensional disaster caused by excessively high dimension. In the model evaluation stage, this paper uses K-fold Cross Validation (K-CV) to evaluate the model to verify whether the model has underfitting and overfitting phenomenon and to verify the stability of the model. Based on this, it is proved that the Raman spectral recognition model based on Random Forest algorithm proposed in this paper can distinguish the collected samples of foodborne pathogenic bacteria.

Results and discussion： In this study, KNN, Logistic Regression, SVM, decision tree and Random Forest model were used for classification prediction of the pre-treated Raman spectral data of foodborne pathogenic bacteria (**Table 4**). Among the 10 fold cross-validation models, the accuracy of Random Forest model is better than that of traditional machine learning algorithms. The decision

tree model presented the worst results with an accuracy rate of 82.63%. This is because compared with the Random Forest algorithm, the decision tree for a single weak learning, decision tree model and Random Forest multiple vote are combined to form the strong learning (**Fig. 5**). Therefore, Random Forest has higher classification ability than single decision tree classifier. Compared with the traditional machine learning algorithm, the Random Forest algorithm adds two randomness in the part of model construction: sampling randomness and feature selection randomness (**Table 2**). Since the Random Forest is composed of decision trees, the higher the correlation of decision trees, the higher the error rate. Random sampling determines the decrease in the correlation of each tree in the Random Forest. Each tree in the Random Forest randomly selects some features, and selects the features of optimal splitting ability as the left and right subtrees of the decision tree among a small number of features, which expands the effect of randomness and further enhances the robustness of the model. Since the introduction of two randomness has a great effect on reducing the Variance of the model, the Random Forest generally does not need additional pruning, that is, it can achieve better generalization ability and anti-over-fitting ability (Low Variance). In addition, Savitzky-Golay filtering algorithm is used for denoising in the pre-processing stage of Raman spectral data (**Fig. 3**), so that the model has a good anti-interference ability.

Conclusion： Raman spectroscopy is a more and more mature technology, which has a significant effect on the detection and classification of foodborne pathogenic bacteria. In this study, Raman spectrometer was used to detect the spectral data of 11 foodborne pathogens. According to the spectral properties, the spectral data are normalized, smoothed and de-noised in the pre-processing stage, which is of great help to the model construction and training. This study constructed a method for identification and analysis of foodborne pathogenic bacteria by Raman spectroscopy. The experimental results show that the classification model of principal component analysis combined with random forest algorithm proposed in this paper has a higher accuracy for Raman spectral data than the conventional single machine learning method, which provides another detection method for foodborne pathogens and improves the speed of manual identification of Raman spectrum. The random forest model is prone to overfitting in the sample sets with large noise processing. In the later research, in the data pretreatment stage, we can optimize the de-noise processing, and in the RANDOM forest algorithm, the data feature selection algorithm can be optimized to improve the accuracy of the model. Only 11 samples of food-borne pathogenic bacteria were used in this study, and more samples could be introduced in the later model construction to build a more complete Raman spectral database.

Key words： Raman spectroscopy; Machine learning; Foodborne pathogen detection; PCA; Random Forest

OCIS codes 330.6230; 300.6450; 240.6695

网络首发：

标题：基于随机森林算法的食源性致病菌拉曼光谱识别
作者：王其,曾万聃,夏志平,李志萍,曲晗
收稿日期：2020-07-06
录用日期：2020-09-14
DOI：10.3788/cjl202148.0311002

引用格式：
王其,曾万聃,夏志平,李志萍,曲晗. 基于随机森林算法的食源性致病菌拉曼光谱识别[J].中国激光,2021,48(03):0311002.
