微生物学报
Acta Microbiologica Sinica
2020, 60(5): 841–855
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20190403

Review 综 述

机器学习在 MALDI-TOF MS 鉴定微生物中的应用

刘宏生[1,2,3#*]，冯华炜[1#]，张力[1,2,3]，孟金蕙[1]，董雪[4]

[1]辽宁大学生命科学院，辽宁 沈阳 110036
[2]辽宁省生物大分子模拟计算与信息处理工程研究中心，辽宁 沈阳 110036
[3]辽宁省药物分子模拟与设计工程实验室，辽宁 沈阳 110036
[4]沈阳市疾病预防控制中心，辽宁 沈阳 110031

摘要：基质辅助激光解吸/电离飞行时间质谱(matrix-assisted laser desorption/ionization time-of-flight mass spectrometry，MALDI-TOF MS)是一种新兴的高通量技术，已广泛应用于临床微生物、食品微生物和水产微生物的快速鉴定。如何进一步提高 MALDI-TOF MS 在微生物鉴定中的分辨率是该技术当前面临的一大挑战。为了高效处理大量高维微生物 MALDI-TOF MS 数据，各种机器学习算法得到了应用。本文综述了机器学习在微生物 MALDI-TOF MS 鉴定中的应用。首先，本文在介绍机器学习在微生物 MALDI-TOF MS 分类中的工作流程后，进一步对 MALDI-TOF MS 的数据特征、MALDI-TOF MS 数据库、数据的预处理和模型的性能评估进行了描述。然后讨论了典型的机器学习分类算法和集成学习算法的应用。简单的机器学习算法很难满足微生物 MALDI-TOF MS 分类的高分辨率的需求，而组合不同机器学习算法和集成学习算法可以获得更好的微生物分类性能。在 MALDI-TOF MS 数据的预处理方面，小波算法和遗传算法的应用最广，它们结合分类算法可以有效提高 MALDI-TOF MS 的分类性能。随着微生物 MALDI-TOF MS 数据量的不断增加，在未来的研究工作中应更重视分类算法的改进、不同算法的选择或组合以及预处理算法的改进。

关键词：微生物鉴定，MALDI-TOF MS，机器学习算法，预处理算法

基金项目：辽宁省高等学校国(境)外培养项目(2018LNGXGJWPY-YB006)；中国科协优秀中外青年交流计划(2018CASTQNJL50)；辽宁省重点研发计划(2019JH2/10300041)；沈阳市科技计划项目(18-014-4-34，F16-205-1-51，17-65-7-00，17-231-1-04)

[#]并列第一作者。
[*]**通信作者**。Tel/Fax：+86-24-62202280；E-mail：liuhongsheng@lnu.edu.cn

收稿日期：2019-09-01；**修回日期**：2019-12-10；**网络出版日期**：2020-03-20

基质辅助激光解吸/电离飞行时间质谱(matrix-assisted laser desorption/ionization time-of-flight mass spectrometry，MALDI-TOF MS)是一种直接从完整的微生物细胞表面检测蛋白的质谱(mass spectrometry，MS)方法，这种方法不仅可以得到与 16S rRNA 基因序列分析相似的鉴定结果，还具有速度快、成本低的特点，适用于细菌、古细菌和真菌的快速可靠鉴定[1–2]。大量研究表明，与传统的表型和生化测试相比，MALDI-TOF MS 的鉴定结果更为准确[3–4]。但是，对常规的 MALDI-TOF MS 和 16S rRNA 基因测序技术来讲，某些特定的分类群，如蜡状芽孢杆菌复合群(*Bacillus cereus* complex)、洋葱伯克霍尔德复合群(*Brukholderia cepacia* complex)、阴沟肠杆菌复合群(*Enterobacter cloacae* complex)、亲缘性高的大肠杆菌(*Escherichia coli*)与志贺氏菌(*Shigella*)以及恶臭假单胞菌复合群(*Pseudomonas* putida complex)的鉴定仍然具有挑战性[5–6]。通过改进算法和开发分析软件可以提高 MALDI-TOF MS 的分辨率。例如，ClinPro Tools 软件已用于大肠杆菌与志贺氏菌的快速分类[5,7]。此外，尽管 MALDI-TOF MS 已用于微生物的分型和抗性分类，如肠杆菌科(*Enterobacteriaceae*)和金黄色葡萄球菌(*Staphylococcus aureus*)的分型，但得到的结果仍然不能令人满意，鉴定方法的普适性也较弱[8–9]。大量的研究证明，微生物 MALDI-TOF MS 分类的成功率通常与其所使用的计算机算法直接相关[9]。显然，传统的统计方法已不能满足快速、准确地对大量 MALDI-TOF MS 数据进行分类的要求，因此在微生物 MALDI-TOF MS 的分类中需要进一步应用机器学习算法。

目前，机器学习已应用于 MALDI-TOF MS 的微生物分类。鉴于已有许多研究探讨了微生物 MALDI-TOF MS 数据的特征提取和特征选择的算法[10–11]，本文将以微生物 MALDI-TOF MS 鉴定中分类算法的应用情况为重点，结合一些常用的预处理算法，拟从以下 3 个方面进行综述：(1) 介绍了机器学习在微生物 MALDI-TOF MS 分类中的工作流程，并对 MALDI-TOF MS 数据的特征、MALDI-TOF MS 数据库的建设情况、数据的预处理和模型的性能评估进行了重点描述。(2) 重点阐述了几种具有代表性的机器学习算法，包括支持向量机(support vector machines，SVM)、随机森林(random forest，RF)、人工神经网络(artificial neural networks，ANN)、遗传算法(genetic algorithm，GA)和朴素贝叶斯算法(naïve bayes algorithm，NB)等；重点探讨了基于集成策略的集成学习(ensemble learning，EL)算法的应用；(3) 对小波算法(wavelet algorithm，WA)和 GA 等典型的预处理算法的应用进行了探讨。

1 基于机器学习的微生物 MALDI-TOF MS 分类

机器学习是一种多元的分析算法。通过学习训练集中的多元数据的模式，机器学习模型可以对未知的数据作出预测。通常，基于机器学习方法开发微生物 MALDI-TOF MS 分类模型的过程包括以下 4 个主要步骤(图 1)：(1) 数据的收集。微生物 MALDI-TOF MS 数据集的主要来源为数据库，其次为已发表论文中的共享数据，也可采用研究人员自己得到的实验数据作为模型开发的数据集。(2) 数据的预处理。包括原始光谱数据的预处理、特征提取、数据分割和特征选择。(3) 模

型建立。采用训练集进展模型的训练，采用验证集进行模型的测试。(4) 模型的评估。选择交叉验证法、留出法、自助法等模型评估方法对训练的模型进行评估。也可采用外部测试集对模型的性能进行验证。在模型的性能评估中，通常以准确率(accuracy，ACC)和受试者工作特性曲线下的面积(area under the receiver operating characteristic curve，AUC)作为评价指标。

1.1 MALDI-TOF MS 数据的特征

在 MALDI-TOF MS 中，样品与基质共结晶，从而产生大量的原始数据。对于一个典型的病原菌来说，MALDI-TOF MS 谱图包含质荷比(mass to charge ratio，*m/z*)、峰高、峰面积、质谱数等数千个测量值，形成 10–30 个峰，其质量范围通常为 2–20 kDa。在 MALDI-TOF MS 谱图中，分布于 *X* 轴的质荷比(取决于样品中检测到的分子质量)和分布于 *Y* 轴的强度(intensity values，通常为 0–100000，取决于样本中检测到的分子数量)共同构成了微生物的二维分类信息。总体而言，微生物的 MALDI-TOF MS 产生的数据噪声较高，维度也远大于样本的数量，因此这种复杂而庞大的生物大数据给基于 MALDI-TOF MS 的微生物分类算法的开发带来了巨大的挑战。

1.2 MALDI-TOF MS 数据库

开发基于机器学习的微生物分类预测模型的第一步就是获得高质量的实验数据。数据库是机器学习中微生物 MALDI-TOF MS 数据集的一个重要来源。目前，MALDI-TOF MS 数据库可分为 3 种类型[10]。第一类为收录物种种类最多的商业数据库，包括 Bruker-Biotyper (Bruker Daltonics，Bremen，德国)、Vitek-MS (bioMérieux，Marcy L'Etoile，法国)、Axima Assurance (Shimadzu，Kyoto，日本)和 Andromas (Andromas SAS，Paris，法国)等四大数据库[1]。第二类数据库为基于商业数据库扩展微生物参考光谱的内部数据库。这些由用户自己建立的内部数据库，包括疏螺旋体

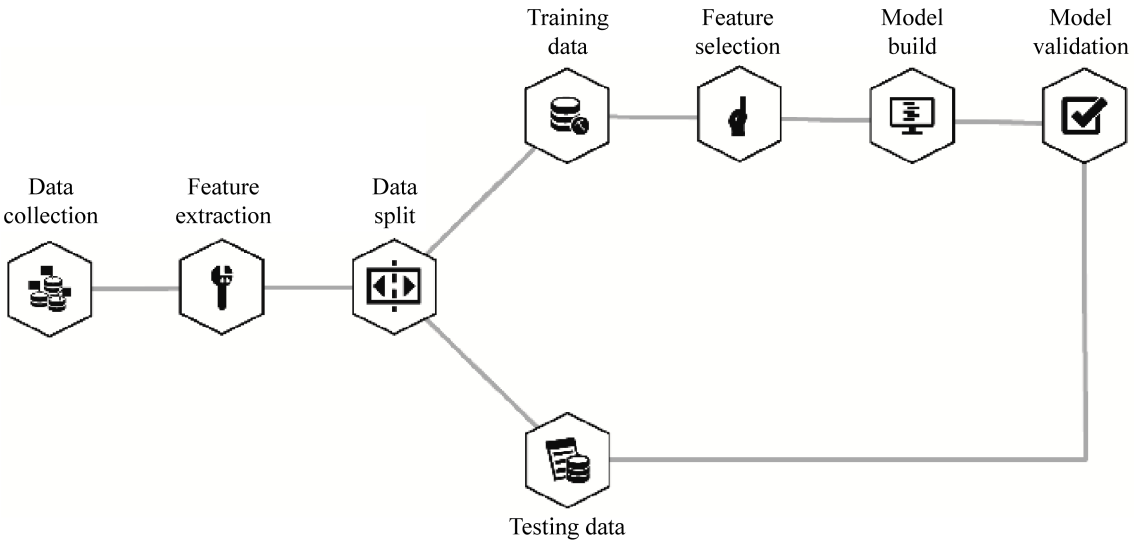

图 1. 机器学习在微生物 MALDI-TOF MS 分类中的工作流程

Figure 1. Workflow of machine learning in microbial MALDI-TOF MS classification.

属(*Borrelia*)、螺菌属(*Spirillum*)、慢生根瘤菌属(*Bradyrhizobium*)、钩端螺旋体(*Leptospira*)、诺卡氏菌(*Nocardia*)、根瘤菌(*Rhizobium*)、布鲁氏菌属(*Brucella*)和马拉色菌属(*Malassezia*)数据库[1,12–15]。第三类为由研究人员自己建立的开放性微生物MALDI-TOF MS数据库(表1)。

总体而言，微生物MALDI-TOF MS数据集的来源仍然是一个大问题。鉴于收录多种微生物MALDI-TOF MS原始数据建立的公共数据库尚未开发，因此，在微生物分类模型的开发中，模型的训练数据可以从上述的数据库中获取，或已发表论文的共享数据中获取，也可采用自己的实验数据作为训练集或验证集[16]。然而，尽管现有的数据库数据量较大，但其商业性质使得研究人员无法自由访问，而开放的数据库又存在收录物种单一、物种数量偏小的缺陷，这也对分类算法的开发造成了限制。已发表的论文也很少共享数据，导致研究人员只能采用本实验室自行检测的有限种类和数量的菌株数据进行算法的开发，因此分类模型的性能还不能令人满意。

1.3 数据的预处理

通常，MALDI-TOF MS的数据可能会存在数据噪声高、数据缺失、分布不均衡及存在异常等诸多数据不规范的问题。因此需要对收集的原始数据进行预处理，并选择有意义的特征进行模型训练。在基于机器学习的微生物MALDI-TOF MS分类算法中，数据的预处理包括四个步骤：(1) 原始光谱数据的预处理。原始数据的预处理包括数据的归一化、谱线平滑和基线校正，用于消除光谱中的噪声，消除微生物的个体差异，识别并去除MALDI-TOF MS数据的基本强度值[21]。数据预处理后得到包含 *m/z* 和相对强度的峰值列表，该列表为微生物分类模型开发时的标准输入数据[22]。(2) 特征提取。通过对齐同类型分离株的 *m/z* 值，计算峰值出现的发生概率及信号强度产生代表性的峰值特征列表。(3) 数据分割。为了避免分类模型的过拟合，将数据集拆分为训练集和测试集。训练集在进行特征选择后可用于模型的建立，测试集则直接用于模型性能的验证。(4) 特征选择。尽管峰值列表可直接作为分类模型的输入数据，但是这些数据中仍然包含会降低分类器准确性的噪音以及不相关或者冗余的峰。因此，在建立分类模型之前，可使用过滤式、包裹式和嵌入式等特征选择方法从代表性的峰值特征中选择最具判别性的特征峰。

1.4 模型性能的评估

在机器学习模型泛化能力进行评估时，性能度量指标是衡量一个模型好坏的关键。一些性能度量指标被用于分类模型的评估。准确率(accuracy，ACC)是最直观的模型性能评估指标，

表 1. 开放性的微生物 MALDI-TOF MS 数据库

Table 1. Open access MALDI-TOF MS database

Database name	Organisms	URL	References
FoodBIMS	26 foodborne pathogens	http://bioinformatica.isa.cnr.it/Descr_Bact_Dbase.htm	[17]
SpectraBank	70 bacterials	http://www.usc.es/gl/investigacion/grupos/lhica/spectrabank/Database.html	[18]
URMS	*Bartonella*	http://ifr48.timone.univ-mrs.fr/portail2/index.php?option=com_content&task=view&id=97&Itemid=54	[19]
VibrioBase	*Vibrio*	https://doi.org/10.1016/j.syapm.2014.10.009	[20]

表示预测为阳性与阴性微生物的总体预测准确率。敏感性(sensitivity，SEN)也被称为查全率或召回率(recall)，它表示阳性的微生物预测准确率。特异性(specificity，SPE)表示预测为阴性微生物的预测准确率。精确率(precision，PRE)又称为查准率，它表示在所有被分类为阳性的微生物样本中，真正是阳性的比例。敏感性和精确率是一对相对矛盾的度量。在模型评估时，精确率高时，敏感性偏低，而敏感性高时，精确率又偏低。采用 *F*1 值(*F*-Score，即精确率和敏感性的调和均值)可以解决二者的矛盾，当精确率和敏感性接近时，*F*1 值最大。受试者工作特性曲线(receiver operating characteristic curve，ROC)是测试中所有可能的截断点的真阳性率(*Y* 轴-敏感性)对假阳性率(*X* 轴-特异性)的曲线图。AUC 是常用的性能度量指标，它代表了模型区分正样本与负样本的整体能力。AUC 值越接近 1 说明模型预测结果越理想，AUC 值为 0.5 代表模型与随机预测相同。

准确率、敏感性、特异性、准确率和 *F*1 值的定义分别如下：

$$ACC=\frac{TP+TN}{TP+TN+FN+FP}\times 100\% \quad (1)$$

$$SEN=\frac{TP}{TP+FN}\times 100\% \quad (2)$$

$$SPE=\frac{TN}{TN+FP}\times 100\% \quad (3)$$

$$PRE=\frac{TN}{TP+FP}\times 100\% \quad (4)$$

$$F1=2*\frac{SEN*PRE}{SEN+PRE}\times 100\% \quad (5)$$

其中，*TP*、*TN*、*FP* 和 *FN* 分别代表真阳性(true positives)、真阴性(true negatives)、假阳性(false positives)和假阴性(false negatives)的数量。

2 分类算法

2.1 典型的机器学习算法

2.1.1 支持向量机(SVM)：SVM 通过非线性变换把原空间映射到高维空间，然后在这个高维空间构造线性分类器。在变换后的高维空间中，边界是线性的，但在原始尺度上，它们通常是非线性的，这使得 SVM 比其他线性分类器具有更高的灵活性。在微生物的分类中，SVM 还可以结合其他机器学习算法对数据进行预处理来提高泛化性能，进而提高 SVM 的分类准确率[11]。由于 SVM 可以很好地表达小样本高维特征空间，现已成为微生物 MALDI-TOF MS 分类中常用的机器学习算法。

2014 年，Almasoud 等[23]分别使用线性 SVM 和 Jaccard kernel SVM 对 34 株(7 个种)杆菌属(*Bacillus*)和短杆菌属(*Brevibacillus*)分别进行种水平和菌株水平的半定量和定性分类。结果显示，线性 SVM 和 Jaccard kernel SVM 在种水平都具有良好的分类性能，准确率分别为 89.27%和 88.92%，优于贝叶斯分类器(77.69%)。但是这两种 SVM 在菌株水平的分类性能还很差，准确率仅为 45.88%–54.04%。Montaudo 等[24]的研究发现，微生物分类的主要影响因素之一是蛋白质含量的定性信息，而非细菌细胞中蛋白质的定量表达水平。2015 年，Lafolie 等[25]分别用 SVM 和快速分类器(quick classifier，QC)对产 β-内酰胺酶的 109 株(94 株为临床株，15 株为环境分离株) ST131 大肠杆菌进行了分类。基于 8496 *m/z*、9713 *m/z*、9738 *m/z* 和 104744 *m/z* 等 4 个峰标记物，SVM 的敏感性为 100%，略高于 QC 法的 99.75%。2016 年，Mather 等[26]开发了一种基于 R 语言的

SVM 算法，对抗万古霉素的金黄色葡萄球菌进行分类。结果显示，SVM 识别万古霉素中介金黄色葡萄球菌(vancomycin intermediate *Staphylococcus aureus*，VISA)和万古霉素敏感金黄色葡萄球菌(vancomycin susceptible *Staphylococcus aureus*，VSSA)分离株的准确率分别为 100%和 97%，总的分类准确率为 98%。在 SVM 分类模型中添加异质性万古霉素中介金黄色葡萄球菌(heterogeneou vancomycin intermediate *Staphylococcus aureus*，hVISA)菌株后，SVM 模型识别 hVISA、VISA 和 VSSA 分离株的敏感性分别降至 76%、100%和 89%，总的分类准确率降低了 9%。但是，在构建 SVM 模型时，该研究似乎未对数据集进行分割，采用整个数据集进行特征选择和模型的优化，分类模型可能存在过度拟合问题。另外，由于 SVM 的分类规则通常无法轻易解释，因此 Wang 等[27]引入径向基核函数(radial basis function，RBF kernel)构建 SVM 二分类模型，对 ST5、ST45、ST59 和 ST239 耐甲氧西林金黄色葡萄球菌(Methicillin resistant *Staphylococcus aureus*，MRSA)进行序列分型，结果显示 RBF kernel SVM 的性能优于决策树(decision tree，DT)和 K-最近邻法(K-nearest neighbor algorithm，KNN)，AUC 为 0.919–0.991。但是，随着类别数量的增加，基于 RBF kernel SVM、DT 和 KNN 算法建立的 MRSA 多分类模型的准确率显著降低。此外，Wang 等[16]还分别使用 SVM、DT、KNN 和 RF 等机器学习算法对 hVISA 和 VISA 分离株进行分类，结果显示，基于 1132 *m/z*、2855 *m/z*、3176 *m/z* 和 6591 *m/z* 等 4 个标记峰建立的 SVM 模型产生了最佳的分类性能，平均敏感性和特异性分别为 77%和 81.4%。

2.1.2 随机森林(RF)：RF 是一个基于树的非参数组合分类器，适用于处理高维和非线性可分离的 MALDI-TOF MS 数据。2011 年，De Bruyne 等[28]采用 RF 和 SVM 对明串珠菌属(*Leuconostoc*)和嗜果糖乳酸细菌属(*Fructobacillus*)的细菌在种水平上进行分类，结果显示 RF 的准确率为 98.4%，高于 SVM 的 94.1%。在对 MALDI-TOF MS 法难以鉴定的超出 MS 仪测量范围的 MRSA 和甲氧西林敏感金黄色葡萄球菌(methicillin susceptible *Staphylococcus aureus*，MSSA)的分类方面，Dai 等[29]采用改良的 RF (在预处理步骤对数据集进行装箱和滑动窗口，然后再进行 RF 模型训练)对 345 株 MSSA 和 382 株 MRSA 菌株进行分类。结果显示，改良的 RF 克服了样本量过小的问题，其准确率、敏感性、特异性和精确率均在 90%以上，比传统的 RF 更为可靠和稳定。2018 年，Asakura 等[30]采用 RF 对 129 株 VISA、VSSA 和 hVISA 分离株进行分类，模型的敏感性和特异性分别为 99%和 88%。值得一提的是，由于 Asakura 等[30]的 RF 分类器是在自动选择光谱中的峰值组合后建立的，因此模型的敏感性比 Mather 等[26]采用 SVM 构建的分类器高 23%。基于 RF 算法，Asakura 等[30]还开发了一个“一体化”的在线软件，该软件允许用户使用他们开发的分类器对个人上传的原始数据进行分析。但是，由于该研究仅选择了 1 个 hVISA 菌株中的多个菌落进行光谱分析，生成的模型也可能存在过拟合问题。

2.1.3 人工神经网络(ANN)：ANN 是由大量处理单元互联组成的非线性、自适应信息处理系统。ANN 试图模拟大脑神经网络处理、记忆信息的方式来处理信息。神经网络算法对噪声具有很强的稳健性和容错性，能够逼近复杂的非线性关系。在微生物的 MALDI-TOF MS 鉴定中，α-溶血性

的草绿色链球菌群(viridans group *Streptococci*, VGS)内的肺炎链球菌(*Streptococci pneumoniae*)、缓征链球菌(*Streptococci mitsi*)、口腔链球菌(*Streptococci oralis*)和假肺炎链球菌(*Streptococci pseudopneumoniae*)具有高度的亲缘性，经常被错误识别。2013 年, Ikryannikova 等[31]分别采用 GA、ANN 和 QC 算法对 62 株不同表型和遗传特征的 VGS 菌株(25 株肺炎链球菌、34 株缓征链球菌和 3 株口腔链球菌)进行分类，ANN 采用 6 个峰，在外部验证集中敏感性和特异性均为 100%, 与其他两个算法的性能无显著差异，可以很好地区分 VGS 内的肺炎链球菌和缓征链球菌。但是，Lasch 等[32]的研究认为 ANN 还无法对屎肠球菌(*Enterococcus faecium*)和金黄色葡萄球菌进行分型。尽管在该研究中,ANN 的准确率达到了 87%，但是在峰值特征选择时，并未发现屎肠球菌和金黄色葡萄球菌的特异性生物标志物峰,使得 ANN 难以对这两类微生物的克隆和克隆复合体进行分型。Angeletti 等[33]分别用 GA、ANN 和 QC 对 25 株连续的非重复的临床分离的耐碳青酶烯类肺炎克雷伯杆菌(*Klebsiella pneumoniae*)进行分类，ANN 采用 2 个标记峰，对耐碳青酶烯类肺炎克雷伯杆菌的分类敏感性为 100%，显著优于其他算法。为了对 MRSA 进行分型，Camoez 等[34]分别用 ANN、GA 和 QC 对属于 4 个克隆群的 82 株 MSRA 进行分类，结果显示 ANN 性能最佳，敏感性和特异性分别为 100%和 99.11%。2017 年，Marí-Almirall 等[35]采用 ANN、GA 和 QC 对 5 个抗体组的 78 株鲍氏不动杆菌(*Acinetobacter baumannii*)进行分型，其中 ANN 的敏感性为 100%，可以把 38 株分离株的验证集中的大部分菌株分类到不动杆菌属(*Acinetobacter*)菌株，其敏感性可达 96.8%。

2.1.4 遗传算法(GA)：GA 是一种基于群体的元启发式全局优化技术，用于处理超大搜索空间的复杂问题。Boggs 等[36]使用 GA 对 47 株 USA300 金黄色葡萄球菌(*Staphylococcus aureus*)和 77 株非 USA300 金黄色葡萄球菌进行分类。结果显示，GA 可采用 5932 *m/z*、6423 *m/z* 和 6592 *m/z* 三个标记峰进行 USA300 金黄色葡萄球菌的分类，对 224 个测试分离株的验证结果显示，GA 的准确率为 87.95%，敏感性为 87%，特异性为 89%。鉴于金黄色葡萄球菌会持续受到宿主和抗生素的压力，USA300 家族菌株会不断进化，进而导致模型的准确率降低，因此该模型对于 USA300 金黄色葡萄球菌分类的适用性还需进一步确认。此外，GA 算法还可用于 1 型和 2 型的肺炎支原体(*Mycoplasma pneumoniae*)的分型，GA 在外部验证集中的特异性和敏感性均为 100%。但是该模型的测试集(25 株分离株)和验证集(43 株分离株)数据较少，可能存在过拟合的问题[37]。Khot 等[5]采用 GA 对亲缘性高的 138 株志贺氏菌属(66 株)与大肠杆菌(72 株)革兰氏阴性菌进行分类时，在种水平上采用 11 个生物标记峰建模，种水平的敏感性可达 90%。Fisher 等[38]的研究也证明了 GA 的混合模型在大肠杆菌和志贺氏菌的分类及其血清分型上均具有良好的性能。总体而言，上述研究大多只用了一半的分离株构建模型，使得分类模型不能涵盖所有受试分离株，进而导致模型性能受到限制。在血清分型上，Nakano 等[39]分别采用 GA、ANN 和 QC 对 574 株肺炎链球菌的 3、6B、15A、15C、19A、19F、23A、24F、35B 和 38 等 10 种血清型进行分类，结果显示，在这三种算法中，GA 可以更好地识别这 10 种血清型，平均敏感性

在 90%以上。

2.1.5 朴素贝叶斯算法(NB)：NB 是基于贝叶斯定理的一种简单的概率分类器。NB 逻辑简单易于实现，分类过程中时空开销小，对于不同特点数据的分类性能差别不大，具有较强的稳健性。依据微生物蛋白质序列的分型策略，Tomachewski 等[40]开发了一种基于核糖体蛋白的质荷比(*m/z*)进行细菌分类的在线工具——Ribopeaks (http://www.ribopeaks.com)，该工具采用 NB 建立分类模型，可以实现 28500 种细菌的分类。对环境来源的 116 株细菌的分类结果显示，Ribopeaks 在属水平和种水平上的敏感性分别为 90.51%和 87.93%[41]。

2.2 基于集成学习(EL)的分类算法

尽管有很多的强分类器(如 SVM、KNN 和 RF)用于微生物的分类。但是，不同的分类器对不同类型的数据分类具有倾向性。分类器的性能取决于多个性能指标，如准确率、敏感性、特异性和 AUC 等，因此很难确定某个算法对某个特定类型数据的分类是最优的。在此背景下，基于集成策略的分类算法在微生物 MALDI-TOF MS 分类中得以应用。2007 年，Assareh 等[42]提出了 EL 方法，该方法使用不同的机器学习算法作为基分类器，如 KNN、SVM、DT 和线性判别分析(linear discriminant analysis，LDA)等，并用不同的训练集训练不同的学习算法。该研究使用 Bhanot 等[43]的 SELDI-TOF MS 前列腺癌数据集进行 EL 的性能评价，结果显示 EL 的性能显著优于基分类器，其敏感性和特异性分别可达 92.55%和 96.86%。2010 年，Datta 等[44]提出了一种新的自适应集成分类器，该分类器通过组合装袋(bagging)和排序聚合(rank aggregation)算法，能够根据被分类的数据类型自适应地改变其性能。模型的验证结果显示，EL 模型的分类性能优于其他的基分类器和简单的集成分类器。2018 年，Ribeiro 等[45]利用 NB、Logistic 回归、DT 和 RF 集成 EL 分类模型对土壤细菌中的 30 个属进行分类，结果显示 EL 的准确率可达 88.89%，优于 RF 的 80.61%、Logistic 回归的 80.29%、NB 的 68.96%以及 DT 的 60.95%，精确率、敏感性和 F1 值等其他性能指标也均高于单个机器学习模型。由此可见，基于集成策略的 EL 大大提升了分类模型的稳定性和预测能力，可以作为微生物 MALDI-TOF MS 分类的首选算法。

尽管 EL 的泛化性能比单一的分类器更加优越，但是由于 EL 融合了多个基分类器，其性能很容易受到弱的基分类器的影响。在建立微生物 MALDI-TOF MS EL 分类模型后，可通过观察 EL 模型的 AUC 值与基分类器数目之间的变化关系，选择具有最大 AUC 值的模型作为最佳的 EL 模型。此外，EL 还存在训练和预测的计算成本高、模型难以解释等缺陷。在集成之后通过集成修剪可以减小模型的存储开销和计算时间开销。在可解释性方面，通过将集成转化为单模型、从集成中抽符号规则等策略衍生的“二次学习技术”和可视化技术，可改善 EL 模型的可解释性。

3 预处理算法

微生物 MALDI-TOF MS 数据的预处理是将原始数据转换为合适的输入，为进一步构建分类模型奠定基础。合适的数据预处理算法不仅可以防止数据结构不兼容导致的机器学习算法无法工作的问题，还可以加快机器学习算法的训练速度，提高算法的精度。目前，机器学习算法已在 MALDI-TOF MS 数据的预处理中得到应用。基于

这些算法，一些开源的数据预处理工具进一步得到开发(表 2)。例如，MALDIquant 已成功应用于鉴定鱼类的致病菌黏着杆菌属(*Tenacibaculum*)[46]、嗜血杆菌(*Haemophilus influenzae*)的荚膜分型[47]、产志贺毒素大肠杆菌的血清分型[48]。本部分将重点介绍 WA 和 GA 在 MALDI-TOF MS 数据预处理中的应用。

3.1 小波算法(WA)

WA 可以用于 MS 数据的去噪和特征选择[54]。在 MALDI-TOF MS 数据中，噪声信号通常来自仪器的干扰、测量和基线失真。WA 不但可以处理化学噪声和仪器噪声，还可以很好地处理非均匀噪声。在特征选择方面，与传统的主成分分析(principal component analysis，PCA)和 LDA 方法相比，WA 更能够保持时间特性，且通过检测局部特征，大大减少了 MALDI-TOF MS 数据的计算量。为了解决 MS 光谱中噪声导致的高误报率问题，Du 等[55]采用基于连续小波变化的峰值检测算法(continuous wavelet transform，CWT)直接对原始的 MALDI-TOF MS 数据去噪，该算法将光谱变换到小波空间，简化了模式匹配问题，可以从峰值噪声和有色噪声中更好地识别和分离信号。与 Bioconductor PROcess 包中的峰值检测算法和 Coombes 等采用的峰值去噪算法相比，CWT 的敏感性更高而误报率更低[56]。但是，计算成本较高是 CWT 算法需要解决的问题。2015 年，Murugesan 等[57]采用对偶树复小波变换(dual tree complex wavelet transform，DTCWT)进行 MS 原始数据的去噪，结果显示，与离散小波变换(discrete wavelet transformation，DWT)和平稳小波变换(stationary wavelet transform，SWT)相比，DTCWT 的性能更佳，计算载荷更低。2016 年，Zheng 等[58]将 CWT 与疯狂爬坡算法(crazy climbing algorithm，CCA)结合对 CWT 进行了改良。对模拟噪声光谱和真实光谱的评估显示，改良后的方法在识别重叠峰方面效果更好。2017 年，Gutiérrez 等[59]采用 CWT 对来自葡萄园和酒厂 109 个酿酒酵母菌株(*Saccharomyces cerevisiae*)和 107 个非酵母菌分离株的 MALDI-TOF MS 数据进行预处理。结果显示，结合 MALDI-TOF MS 数据采集算法，CWT 可以产生高质量的数据集，属水平的准确率为 95.4% (206/216)，种水平的准确率为 100% (216/216)，高于 Ge 等的 77.8% (63/81)和 93.8% (76/81)[60]。尽管如此，该研究尚不能实现酵母菌菌株水平的分类。

3.2 遗传算法(GA)

考虑到 MALDI-TOF MS 数据具有的高维度和小样本量特征[61]，常用 GA 对 MALDI-TOF MS 数据进行特征选择。1997 年，Broadhurst 等[62]将 GA 用作热解 MS 的特征选择算法。GA 可以用于

表 2. 开源的 MALDI-TOF MS 数据预处理工具

Table 2. Open access tools for preprocessing MALDI-TOF MS data

Tool name	URL	References
MALDIquant	http://strimmerlab.org/software/maldiquant/	[49]
Mass-Up	http://sing.ei.uvigo.es/mass-up	[50]
SPECLUST	http://bioinfo.thep.lu.se/speclust.html	[18]
BIOSPEAN	http://biochemie.upol.cz/index.php/cs/vyzkum/odkazy	[51]
MALDIrppa	https://CRAN.R-project.org/package=MALDIrppa	[52]
HABase	https://uhcl-habase.shinyapps.io/habase_web-based_spectra_analysis/	[53]

寻找多元线性回归(multiple linear regression，MLR)和偏最小二乘法(partial least squares，PLS)回归等模型中回归变量的最优子集，从而将变量从 150 降至20以下。2011年，为了实现芽孢杆菌(*Bacillus*)的鉴定和种水平分类，Correa 等[63]将 GA 与贝叶斯网络算法(bayesian network，BN)相结合，依据数据子集的不同，将变量从 150 个降低到 22–39 个。此外，GA-BN 在芽孢的生物标记物挖掘方面也具有优越的性能。2017 年，Bai 等[64]提出了一种基于 wrapper 的改良 GA 对 MRSA 和 MSSA 的 MS 进行特征选择，在采用 SVM 进行分类后，改良的 GA 算法明显优于传统方法，平均准确率为 72%，平均敏感性为 71% (比传统算法高 1.6%)。

4 展望

MALDI-TOF MS 是一种功能强大、经济有效、快速且稳健的微生物分类技术，现已成功应用于细菌、真菌和古生菌的分类鉴定。为了进一步区分亲缘关系较近和难以分型的微生物，划分和扩展数据库变得越来越重要。然而，现有的微生物 MALDI-TOF MS 数据库的数量和大小还不能满足微生物的分类需求，而缺乏提高 MALDI-TOF MS 分辨率的有效算法更是当前该技术面临的一大挑战。在此背景下，本文研究了典型的机器学习分类算法、集成学习分类算法和数据预处理算法在微生物 MALDI-TOF MS 分类中的应用情况。从目前的应用状况来看，简单而基本的机器学习算法很难满足微生物 MALDI-TOF MS 分类的需求。随着机器学习算法研究的不断深入，组合不同的机器学习算法和基于集成策略的 EL 显示出优越的分类性能，有效提高了基于微生物 MALDI-TOF MS 鉴定的分辨率。在 MALDI-TOF MS 数据的预处理中，特征峰的确定主要取决于特征选择算法和机器学习算法。WA 和 GA 常用于去噪、特征选择和特征提取，通过组合其他机器学习分类算法可以进一步提高 MALDI-TOF MS 在微生物分类中的分辨率[57,64]。此外，卷积神经网络已经用于 MS 峰值的检测和基线校正，这表明未来深度学习算法可用于微生物 MALDI-TOF MS 分类任务[65]。总而言之，MALDI-TOF MS 的数据集庞大而复杂，未来还需在加大微生物 MALDI-TOF MS 数据库建设的基础上，将分类算法的选择、不同机器学习算法的组合和预处理算法的改进作为研究的重点。

此外，在采用机器学习算法对微生物 MALDI-TOF MS 进行分类的实际工作中，数据预处理是影响分类准确性一个主要原因，采用统计容差法计算光谱数据的容差值进行峰值对齐更有助于增加模型的鲁棒性[66]。在混合微生物的分类中，采用基于核糖体蛋白的生物标记物作为输入特征是一种优选的策略。另外，对于微生物分类算法的开发，基于蛋白质序列数据库比对的策略也是一个良好的选择。该策略采用微生物基因组预测的蛋白质分子量来匹配微生物产生的 MALDI-TOF MS 的质荷比，不仅可以提高微生物 MALDI-TOF MS 的分辨率，还可解决算法开发时 MALDI-TOF MS 数据缺乏的问题。

参考文献

[1] Rahi P, Prakash O, Shouche YS. Matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry (MALDI-TOF MS) based microbial identifications: challenges and scopes for microbial ecologists. *Frontiers in Microbiology*, 2016, 7: 1359.

[2] Bellanger AP, Gbaguidi-Haore H, Liapis E, Scherer E, Millon L. Rapid identification of *Candida* sp. by MALDI-TOF mass spectrometry subsequent to short-term incubation on a solid medium. *APMIS*, 2019, 127(4): 217–221.

[3] Bessède E, Solecki O, Sifré E, Labadi L, Mégraud F. Identification of *Campylobacter* species and related organisms by matrix assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry. *Clinical Microbiology and Infection*, 2011, 17(11): 1735–1739.

[4] Carbonnelle E, Grohs P, Jacquier H, Day N, Tenza S, Dewailly A, Vissouarn O, Rottman M, Herrmann JL, Podglajen I, Laurent R. Robustness of two MALDI-TOF mass spectrometry systems for bacterial identification. *Journal of Microbiological Methods*, 2012, 89(2): 133–136.

[5] Khot PD, Fisher MA. Novel approach for differentiating *Shigella* species and *Escherichia coli* by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology*, 2013, 51(11): 3711–3716.

[6] Almuzara M, Barberis C, Traglia G, Famiglietti A, Ramirez MS, Vay C. Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry for species identification of nonfermenting gram-negative Bacilli. *Journal of Microbiological Methods*, 2015, 112: 24–27.

[7] Paauw A, Jonker D, Roeselers G, Jonathan MH, Mars-Groenendijk RH, Trip H, Molhoek EM, Jansen HJ, van der Plas J, de Jong AL, Majchrzykiewicz-Koehorst JA, Speksnijder AGCL. Rapid and reliable discrimination between *Shigella* species and *Escherichia coli* using MALDI-TOF mass spectrometry. *International Journal of Medical Microbiology*, 2015, 305(4/5): 446–452.

[8] Li P, Xin WW, Xia SS, Luo Y, Chen ZW, Jin DZ, Gao S, Yang H, Ji B, Wang HH, Yan Y, Kang L, Wang JL. MALDI-TOF mass spectrometry-based serotyping of *V. parahaemolyticus* isolated from the Zhejiang province of China. *BMC Microbiology*, 2018, 18(1): 185.

[9] Culebras DE. Application of MALDI-TOF MS in bacterial strain typing and taxonomy//Cobo F. The Use of Mass Spectrometry Technology (MALDI-TOF) in Clinical Microbiology. Amsterdam: Academic Press, 2018: 213–233.

[10] Tsuchida S. Application of MALDI-TOF for bacterial identification//Cobo F. The Use of Mass Spectrometry Technology (MALDI-TOF) in Clinical Microbiology. Amsterdam: Academic Press, 2018: 101–112.

[11] Datta S, Pihur V. Feature selection and machine learning with mass spectrometry data//Matthiesen R. Bioinformatics Methods in Clinical Research. New York: Humana Press, 2010: 205–229.

[12] Lohmann C, Sabou M, Moussaoui W, Prévost G, Delarbre JM, Candolfi E, Gravet A, Letscher-Bru V. Comparison between the Biflex III-Biotyper and the Axima-SARAMIS systems for yeast identification by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology*, 2013, 51(4): 1231–1236.

[13] Sonthayanon P, Jaresitthikunchai J, Mangmee S, Thiangtrongjit T, Wuthiekanun V, Amornchai P, Newton P, Phetsouvanh R, Day NPJ, Roytrakul S. Whole cell matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for identification of *Leptospira* spp. in Thailand and Lao PDR. *PLoS Neglected Tropical Diseases*, 2019, 13(4): e0007232.

[14] Mesureur J, Arend S, Cellière B, Courault P, Cotte-Pattat PJ, Totty H, Deol P, Mick V, Girard V, Touchberry J, Burrowes V, Lavigne JP, O'Callaghan D, Monnin V, Keriel A. A MALDI-TOF MS database with broad genus coverage for species-level identification of *Brucella*. *PLoS Neglected Tropical Diseases*, 2018, 12(10): e0006874.

[15] Honnavar P, Ghosh AK, Paul S, Shankarnarayan SA, Singh P, Dogra S, Chakrabarti A, Rudramurthy SM. Identification of *Malassezia* species by MALDI-TOF MS after expansion of database. *Diagnostic Microbiology and Infectious Disease*, 2018, 92(2): 118–123.

[16] Wang HY, Chen CH, Lee TY, Horng JT, Liu TP, Tseng YJ, Lu JJ. Rapid detection of heterogeneous vancomycin-intermediate *Staphylococcus aureus* based on matrix-assisted laser desorption ionization time-of-flight: using a machine learning approach and unbiased validation. *Frontiers in Microbiology*, 2018, 9: 2393.

[17] Mazzeo MF, Sorrentino A, Gaita M, Cacace G, Di Stasio M, Facchiano A, Comi G, Malorni A, Siciliano RA. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for the discrimination of food-borne microorganisms. *Applied and Environmental Microbiology*,

2006, 72(2): 1180–1189.

[18] Böhme K, Fernández-No IC, Barros-Velázquez J, Gallardo JM, Cañas B, Calo-Mata P. SpectraBank: an open access tool for rapid microbial identification by MALDI-TOF MS fingerprinting. *Electrophoresis*, 2012, 33(14): 2138–2142.

[19] Fournier PE, Couderc C, Buffet S, Flaudrops C, Raoult D. Rapid and cost-effective identification of *Bartonella* species using mass spectrometry. *Journal of Medical Microbiology*, 2009, 58(9): 1154–1159.

[20] Erler R, Wichels A, Heinemeyer EA, Hauk G, Hippelein M, Reyes NT, Gerdts G. VibrioBase: A MALDI-TOF MS database for fast identification of *Vibrio* spp. that are potentially pathogenic in humans. *Systematic and Applied Microbiology*, 2015, 38(1): 16–25.

[21] López Fernández H, Reboiro-Jato M, Pérez Rodríguez JA, Fdez-Riverola F, Glez-Peña D. Implementing effective machine learning-based workflows for the analysis of mass spectrometry data. *Journal of Integrated OMICS*, 2016, 6(1): 23–27.

[22] Esener N, Green MJ, Emes RD, Jowett B, Davies PL, Bradley AJ, Dottorini T. Discrimination of contagious and environmental strains of *Streptococcus uberis* in dairy herds by means of mass spectrometry and machine-learning. *Scientific Reports*, 2018, 8(1): 17517.

[23] Almasoud N, Xu Y, Nicolaou N, Goodacre R. Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS) for the characterization of *Bacillus* and *Brevibacillus* species. *Analytica Chimica Acta*, 2014, 840: 49–57.

[24] Montaudo G, Montaudo MS, Puglisi C, Samperi F. Characterization of polymers by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: molecular weight estimates in samples of varying polydispersity. *Rapid Communications in Mass Spectrometry*, 1995, 9(5): 453–460.

[25] Lafolie J, Sauget M, Cabrolier N, Hocquet D, Bertrand X. Detection of *Escherichia coli* sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies? *Journal of Hospital Infection*, 2015, 90(3): 208–212.

[26] Mather CA, Werth BJ, Sivagnanam S, Sengupta DJ, Butler-Wu SM. Rapid detection of vancomycin-intermediate *Staphylococcus aureus* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Journal of Clinical Microbiology*, 2016, 54(4): 883–890.

[27] Wang HY, Lee TY, Tseng YJ, Liu TP, Huang KY, Chang YT, Chen CH, Lu JJ. A new scheme for strain typing of methicillin-resistant *Staphylococcus aureus* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach. *PLoS One*, 2018, 13(3): e0194289.

[28] De Bruyne K, Slabbinck B, Waegeman W, Vauterin P, De Baets B, Vandamme P. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Systematic and Applied Microbiology*, 2011, 34(1): 20–29.

[29] Dai YL, Fan ZC, Zhang LP, Xu XY, Zhang ZL. Improved random forest algorithm to classify methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* on mass spectra//Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology. Lisbon, Portugal: ACM, 2017: 64–69.

[30] Asakura K, Azechi T, Sasano H, Matsui H, Hanaki H, Miyazaki M, Takata T, Sekine M, Takaku T, Ochiai T, Komatsu N, Shibayama K, Katayama Y, Yahara K. Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant *Staphylococcus aureus* by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *PLoS One*, 2018, 13(3): e0194212.

[31] Ikryannikova LN, Filimonova AV, Malakhova MV, Savinova T, Filimonova O, Ilina EN, Dubovickaya VA, Sidorenko SV, Govorun VM. Discrimination between *Streptococcus pneumoniae* and *Streptococcus mitis* based on sorting of their MALDI mass spectra. *Clinical Microbiology and Infection*, 2013, 19(11): 1066–1071.

[32] Lasch P, Fleige C, Stämmler M, Layer F, Nübel U, Witte W, Werner G. Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of *Enterococcus faecium* and *Staphylococcus aureus* isolates. *Journal of Microbiological Methods*, 2014, 100: 58–69.

[33] Angeletti S, Dicuonzo G, Lo Presti A, Cella E, Crea F, Avola A, Vitali MA, Fagioni M, de Florio L. MALDI-TOF mass

spectrometry and bla_{kpc} gene phylogenetic analysis of an outbreak of carbapenem-resistant *K. pneumoniae* strains. *New Microbiologica*, 2015, 38(4): 541–550.

[34] Camoez M, Sierra JM, Dominguez MA, Ferrer-Navarro M, Vila J, Roca I. Automated categorization of methicillin-resistant *Staphylococcus aureus* clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. *Clinical Microbiology and Infection*, 2016, 22(2): 161.e1–161.e7.

[35] Marí-Almirall M, Cosgaya C, Higgins PG, van Assche A, Telli M, Huys G, Lievens B, Seifert H, Dijkshoorn L, Roca I, Vila J. MALDI-TOF/MS identification of species from the *Acinetobacter baumannii* (Ab) group revisited: inclusion of the novel *A. seifertii* and *A. dijkshoorniae* species. *Clinical Microbiology and Infection*, 2017, 23(3): 210.e1–210.e9.

[36] Boggs SR, Cazares LH, Drake R. Characterization of a *Staphylococcus aureus* USA300 protein signature using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of Medical Microbiology*, 2012, 61(5): 640–644.

[37] Xiao D, Zhao F, Zhang HF, Meng FL, Zhang JZ. Novel strategy for typing *Mycoplasma pneumoniae* isolates by use of matrix-assisted laser desorption ionization-time of flight mass spectrometry coupled with ClinProTools. *Journal of Clinical Microbiology*, 2014, 52(8): 3038–3043.

[38] Fisher MA. Differentiation of closely related organisms using MALDI-TOF MS//Shah HN, Gharbia SE. MALDI-TOF and Tandem MS for Clinical Microbiology. West Sussex: John Wiley & Sons Ltd, 2017: 147–165.

[39] Nakano S, Matsumura Y, Ito Y, Fujisawa T, Chang B, Suga S, Kato K, Yunoki T, Hotta G, Noguchi T, Yamamoto M, Nagao M, Takakura S, Ohnishi M, Ihara T, Ichiyama S. Development and evaluation of MALDI-TOF MS-based serotyping for *Streptococcus pneumoniae*. *European Journal of Clinical Microbiology & Infectious Diseases*, 2015, 34(11): 2191–2198.

[40] Tomachewski D, Galvão CW, de Campos Júnior A, Guimarães AM, Ferreira Da Rocha JC, Etto RM. Ribopeaks: a web tool for bacterial classification through *m*/*z* data from ribosomal proteins. *Bioinformatics*, 2018, 34(17): 3058–3060.

[41] Ziegler D, Pothier JF, Ardley J, Fossou RK, Pflüger V, de Meyer S, Vogel G, Tonolla M, Howieson J, Reeve W, Perret X. Ribosomal protein biomarkers provide root nodule bacterial identification by MALDI-TOF MS. *Applied Microbiology and Biotechnology*, 2015, 99(13): 5547–5562.

[42] Assareh A, Moradi MH, Esmaeili V. A novel ensemble strategy for classification of prostate cancer protein mass spectra//Proceedings of 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Lyon, France: IEEE, 2007: 5987–5990.

[43] Bhanot G, Alexe G, Venkataraghavan B, Levine AJ. A robust meta-classification strategy for cancer detection from MS data. *Proteomics*, 2006, 6(2): 592–604.

[44] Datta S, Pihur V, Datta S. An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data. *BMC Bioinformatics*, 2010, 11: 427.

[45] Ribeiro LG, Da Rocha JCF, Fedacz GL, Dos Santos F, Tomachewski D, Etto RM. Um modelo Ensemble discriminativo para classificação de bactérias do Solo. *Anais SULCOMP*, 2018, 9: 1–10.

[46] Fernández-Álvarez C, Torres-Corral Y, Saltos-Rosero N, Santos Y. MALDI-TOF mass spectrometry for rapid differentiation of *Tenacibaculum* species pathogenic for fish. *Applied Microbiology and Biotechnology*, 2017, 101(13): 5377–5390.

[47] Månsson V, Gilsdorf JR, Kahlmeter G, Kilian M, Kroll JS, Riesbeck K, Resman F. Capsule typing of *Haemophilus influenzae* by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Emerging Infectious Diseases*, 2018, 24(3): 443–452.

[48] Mclean K, Palarea-Albaladejo J, Currie CG, Imrie LHJ, Manson EDT, Fraser-Pitt D, Wright F, Alexander CJ, Pollock KGJ, Allison L, Hanson M, Smith DGE. Rapid and robust analytical protocol for *E. coli* STEC bacteria subspecies differentiation using whole cell MALDI mass spectrometry. *Talanta*, 2018, 182: 164–170.

[49] Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 2012, 28(17): 2270–2271.

[50] López-Fernández H, Santos HM, Capelo JL, Fdez-Riverola F, Glez-Peña D, Reboiro-Jato M. Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry

knowledge discovery. *BMC Bioinformatics*, 2015, 16: 318.

[51] Raus M, Šebela M. BIOSPEAN: a freeware tool for processing spectra from MALDI intact cell/spore mass spectrometry. *Journal of Proteomics & Bioinformatics*, 2013, 6(12): 283–287.

[52] Palarea-Albaladejo J, Mclean K, Wright F, Smith DGE. MALDIrppa: quality control and robust analysis for mass spectrometry data. *Bioinformatics*, 2018, 34(3): 522–523.

[53] LaMontagne M, Shetty T, Gajjar T, Kayyuru C, Sriram S, Zhang CL, Buddharaju P. HABase: A web-application for the analysis of protein spectra and identification of microbial species//Proceedings of the International Conference on Bioinformatics and Computational Biology. Las Vegas, Nevada, USA: CSREA Press, 2017: 77–78.

[54] Liu YH. Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in Biology and Medicine*, 2009, 39(9): 818–823.

[55] Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 2006, 22(17): 2059–2065.

[56] Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 2005, 5(16): 4107–4117.

[57] Murugesan S, Tay DBH, Cooke I, Faou P. Application of dual tree complex wavelet transform in tandem mass spectrometry. *Computers in Biology and Medicine*, 2015, 63: 36–41.

[58] Zheng Y, Fan RL, Qiu CL, Liu Z, Tian D. An improved algorithm for peak detection in mass spectra based on continuous wavelet transform. *International Journal of Mass Spectrometry*, 2016, 409: 53–58.

[59] Gutiérrez C, Gómez-Flechoso MÁ, Belda I, Ruiz J, Kayali N, Polo L, Santos A. Wine yeasts identification by MALDI-TOF MS: optimization of the preanalytical steps and development of an extensible open-source platform for processing and analysis of an in-house MS database. *International Journal of Food Microbiology*, 2017, 254: 1–10.

[60] Ge MC, Kuo AJ, Liu KL, Wen YH, Chia JH, Chang PY, Lee MH, Wu TL, Chang SC, Lu JJ. Routine identification of microorganisms by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: success rate, economic analysis, and clinical outcome. *Journal of Microbiology, Immunology and Infection*, 2017, 50(5): 662–668.

[61] Li YF, Liu YH, Bai L. Genetic algorithm based feature selection for mass spectrometry data//Proceedings of 2008 8th IEEE International Conference on BioInformatics and BioEngineering. Athens, Greece: IEEE, 2008: 1–6.

[62] Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 1997, 348(1/3): 71–86.

[63] Correa E, Goodacre R. A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: application to the rapid identification of *Bacillus* spores and classification of *Bacillus* species. *BMC Bioinformatics*, 2011, 12: 33.

[64] Bai J, Fan ZC, Zhang LP, Xu XY, Zhang ZL. Classification of methicillin-resistant and methicillin-susceptible *Staphylococcus aureus* using an improved genetic algorithm for feature selection based on mass spectra//Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology. Lisbon, Portugal: ACM, 2017: 57–63.

[65] Schmidt MN, Alstrøm TS, Svendstorp M, Larsen J. Peak detection and baseline correction using a convolutional neural network//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, United Kingdom: IEEE, 2019: 2757–2761.

[66] Chung CR, Wang HY, Lien F, Tseng YJ, Chen CH, Lee TY, Liu TP, Horng JT, Lu JJ. Incorporating statistical test and machine intelligence into strain typing of *Staphylococcus haemolyticus* based on matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Frontiers in Microbiology*, 2019, 10: 2120.

Application of machine learning in MALDI-TOF MS identification of microorganisms

Hongsheng Liu[1,2,3#*], Huawei Feng[1#], Li Zhang[1,2,3], Jinhui Meng[1], Xue Dong[4]

[1] School of Life Sciences, Liaoning University, Shenyang 110036, Liaoning Province, China
[2] Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning Province, Liaoning University, Shenyang 110036, Liaoning Province, China
[3] Engineering Laboratory of Molecular Modeling and Design for Drug of Liaoning Province, Liaoning University, Shenyang 110036, Liaoning Province, China
[4] Shenyang Centre for Disease Control and Prevention, Shenyang 110031, Liaoning Province, China

Abstract: Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a novel high-throughput technology widely used in rapid identification of clinical microorganisms, food microorganisms and aquatic microorganisms. Currently, however, how to further improve the resolution of MALDI-TOF MS in microbial identification is a major challenge for this technology. To effectively deal with the large amounts of high-dimensional microbial MALDI-TOF MS data, a variety of machine learning algorithms have been applied. This paper reviews the applications of machine learning in MALDI-TOF MS identification of microorganisms. Herein, the workflow of machine learning in the classification of microbial MALDI-TOF MS is introduced. Then, the characteristics of MALDI-TOF MS data, MALDI-TOF MS database, the preprocessing of the MALDI-TOF MS data, and the performance evaluation of the model are further described. The applications of typical machine learning classification algorithms and ensemble learning algorithms are also discussed.

Keywords: microbiological identification, MALDI-TOF MS, machine learning algorithm, preprocessing algorithm

(本文责编：李磊)

Supported by the High-level Innovation Team Foreign Training Project (2018LNGXGJWPY-YB006), by the Excellent Chinese and Foreign Youth Exchange Plan Project from China Association for Science and Technology (2018CASTQNJL50), by the Liaoning Province Key R&D Program (2019JH2/10300041) and by the Shenyang Science and Technology Plan Project (18-014-4-34, F16-205-1-51, 17-65-7-00, 17-231-1-04)

[#]These authors contributed equally to this work.
[*]Corresponding author. Tel/Fax: +86-24-62202280; E-mail: liuhongsheng@lnu.edu.cn

Received: 1 September 2019; Revised: 10 December 2019; Published online: 20 March 2020