微生物学报
Acta Microbiologica Sinica
2021, 61(9): 2581–2593
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20200632

Review 综 述

机器学习在微生物组宿主表型预测中的应用

李高磊，黄玮，孙浩，李余动*

浙江工商大学食品与生物工程学院，浙江 杭州 310018

摘要：随着大数据时代的到来，如何将生物组学海量数据转化为易理解及可视化的知识是当前生物信息学面临的重要挑战之一。为了处理复杂、高维的微生物组数据，目前机器学习算法已被应用于人体微生物组研究，以揭示疾病背后的复杂机制。本文首先简述了微生物组数据处理方法及常用的机器学习算法，如支持向量机(SVM)、随机森林(RF)和人工神经网络(ANN)等，然后对机器学习的工作流程及其要点进行阐述，并探讨了机器学习算法在基于微生物组数据预测宿主表型方面的应用。最后以唾液微生物组数据预测口腔异味为例，实现了机器学习算法的模型构建与评估分析，并提供了可用于微生物组研究实践的 R/Python 代码(https://github.com/LiLabZSU/microbioML)。

关键词：机器学习，微生物组，大数据，宿主表型，预测

人体微生物群(microbiota)是所有生活在人体内部和表面的微生物的集合，由多种微生物组成，包括细菌、古菌、真菌和病毒等，主要分布在口腔、皮肤、胃肠道和泌尿生殖道等部位。例如人类口腔中定殖了超过 600 种不同的细菌、病毒、真菌及衣原体等微生物物种[1]。这些微生物群的遗传信息的总和称为人类微生物组(microbiome)，是人体本身基因组外的第二基因组(second genome)。随着人类微生物组计划的开展，越来越多的研究表明，人体微生物群落结构的变化与许多疾病相关，如炎症性肠病(IBD)、糖尿病、肥胖、肿瘤等，而菌群失调(dysbiosis)的干预手段将可用于治疗疾病[2]。因此，了解人体微生物组的多样性和分布，特别是在不同生理和疾病状态下发生的变化，或发现微生物群落组成与疾病临床特征之间的关联性，将有助于疾病的临床检测、诊断和治疗[3–4]。目前基于微生物组数据进行准确的疾病诊断与治疗已成为生物医学领域研究的热点方向之一。

随着基因测序技术的进步和测序成本不断下降，大样本量的微生物组学研究激增。传统的统计方法已经不再适用于极度高维、稀疏的微生

基金项目：国家自然科学基金(31671836)

*通信作者。Tel：+86-571-28008900；E-mail：lyd@zjsu.edu.cn

收稿日期：2020-10-10；**修回日期**：2021-03-08；**网络出版日期**：2021-03-26

物组数据分析，而适用于复杂数据分析的机器学习逐渐成为微生物组学数据分析的首选方法[5]。机器学习(machine learning，ML)让计算机能够自主“学习” (拟合训练数据)，通过神经网络、决策树等算法，从数据中自动分析寻找规律，从而生成经验模型，用于新数据的预测或分类。机器学习算法的主要理论成果在 20 世纪 60 年代就已产生，伴随高性能计算和大数据技术的发展，又产生深度学习(deep learning，DL)等人工智能技术。深度学习能够很好地处理组学序列数据，自动确定关键特征(features)用于构建模型，从而对疾病进行准确的预测和诊断等[6]。目前机器学习已被广泛应用于生物医学与临床研究，例如 DeepMind Health 团队致力于开发有效的深度学习技术帮助医生鉴定有急性肾损伤风险的患者(https://github.com/LiLabZSU/microbioML)。欧盟于2019年启动了ML4Microbiome项目(https://www.ml4microbiome.eu/)，旨在优化和规范机器学习在微生物组研究中的应用。

机器学习方法往往需要结合统计学、概率论、线性代数和算法复杂性理论等数学知识，对于生物医学研究者是一个巨大的挑战。本文综述了机器学习方法在基于微生物组数据预测宿主表型方面的应用，简要介绍了微生物组研究中常用的机器学习方法、数据处理步骤以及性能评价指标，并提供了一个基于唾液微生物组数据预测口腔异味的研究案例。

1 微生物组数据

微生物组研究一般可分 16S rRNA 基因扩增子(amplicon)测序与全基因组鸟枪法(shotgun)测序两种策略。根据不同的研究目的，可从唾液、皮肤、粪便或血液等中收集微生物样本。微生物样品的 16S rRNA 基因测序数据需要使用微生物组分析流程(如 QIIME 或 mothur 等)处理，按序列相似度聚类为不同簇(cluster)，即可操作分类单位(OTU)，再用于后续微生物多样性分析。OTU 聚类一般有 3 种方法：(1) 通过直接与参考序列数据库进行序列比对来确定 OTU (closed-reference)；(2) 如果没有参考序列数据库，可从头开始聚类来获得 OTU，这种是从头聚类法(*de novo* clustering)；(3) 先基于参考序列来获得 OTU 聚类，剩下不能与参考序列匹配的序列，再重新进行从头聚类(open-reference)。一旦确定了 OTU 簇，就可以通过 RDP Classifier 等方法为每个 OTU 的代表序列指定物种分类信息。16S rRNA 数据已经被证实具有较高分类准确性，关于从实验样品中获得微生物组序列的方法的更多细节可参见文献[7]。

OTU 丰度数据是最常用的微生物群落与宿主性状关联分析的输入特征(features)。如表 1 所示，OTU 表为菌群计数数据的 n×m 矩阵 X[n, m]，其中 n 为 OTU 特征数，m 为样本数。Y(m)为长度为 m 的宿主性状(traits)向量，性状可能是一个分类数据，如受试者有无口臭症状，或者是一个连续数据，如身体重量指数(BMI)。OTU 表根据序列相似性聚类后得到的 OTU 簇数量表示特定分类单元的丰度，避免了只对相同序列进行分组而导致的过度稀疏。由于微生物组研究的样本数一般远少于 OTU 或物种数，造成数据集的维度较高，即每一个样本有很多特征。OTU 表的高维特性使其难以用传统统计学方法进行数据处理，机器学习作为分析高维数据的一种有效手段，可以用于阐明微生物类群(或其他宏基因组特征)与宿主或环境属性之间的联系[8]。

表 1. OTU 表(部分)[a]

Table 1. OTU table (partial)

#OTU_ID	Sample1	Sample2	Sample3	Sample4	Sample5
OTU3	18	10	166	16	66
OTU7	29	15	350	35	50
OTU35	29	24	112	17	92
OTU78	76	57	0	62	0
OTU115	22	36	182	0	0
Trait	P	N	P	P	N

[a] Each row represents an OTU, and each column represents a sample. The last row represents a host trait (P stands for positive and N stands for negative).

2 机器学习算法

根据训练数据是否带有标签(tag)，机器学习算法可分为两大类：监督学习与无监督学习(图 1)。监督学习算法利用带有标签的训练集来构建数学模型(model)，以描述如何通过特征(如基因表达量)的组合与目标变量(如对癌症进展有无影响)进行关联，然后此模型可被用于新数据的预测。标签可以是离散的，比如是否患病，适用于分类算法；也可以是连续的，比如存活率，适用于回归算法。与监督学习相反，无监督学习方法不需要预先标注的数据作为训练集。这一大类算法可以帮助发现或探索数据的结构，其在生物学中广泛应用的例子是数据聚类(clustering)，根据观察变量的属性对观察变量进行分组，例如在不同癌症样本中根据基因表达量发现共表达(co-expressed)的基因。

虽然在大数据时代获得数据往往比较容易，但由于获取带标签数据的成本与难度较大，因此大规模数据中往往仅有部分带有标签，而大多数样本是无标签数据，例如大型人群队列检测样本中只有小部分是经诊断的癌症患者样本。此时就需要通过一部分有标签数据对其余数据进行预测。半监督(semi-supervised)学习就是可应用在有标签数据与无标签数据混合的训练数据中的机器学习算法。常见的半监督学习有简单自训练(self-training)、标签传播算法和半监督深度学习等。

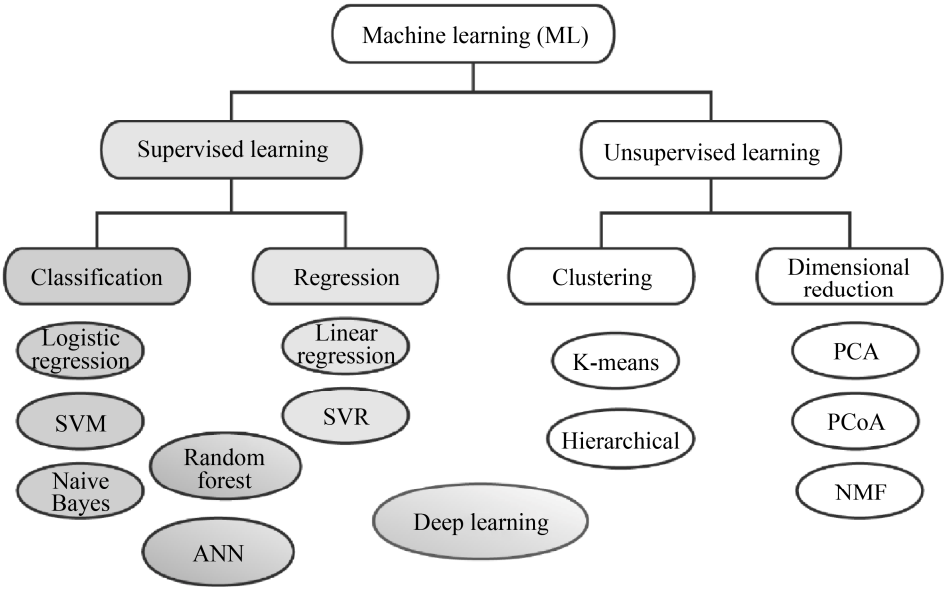

图 1. 机器学习算法分类[5]

Figure 1. Classification of machine learning algorithms[5].

2.1 监督学习算法

监督学习算法将从带标签的训练集提供的事实中学习，以建立一个通用的模型，用于预测新数据的标签。目前许多著名的监督学习算法(如支持向量机、随机森林、贝叶斯网络)已经应用于微生物组学研究。下面介绍 4 种常见的监督学习算法(图 2)。

2.1.1 线性回归：在监督学习方法中，线性回归算法使用线性方程对数据集进行拟合，是一种常见的回归算法。惩罚回归(penalized regression)适用于与样本数量相比包含大量特征的高维数据的回归方法，包括 LASSO (least absolute shrinkage and selection operator)、岭回归(ridge regression)和弹性网络。为了避免在变量之间存在相关性时过高估计回归系数，惩罚回归将惩罚函数应用于回归系数。LASSO 回归也叫线性回归的 L1 正则化，该方法最突出的优势在于通过对所有变量系数进行回归惩罚，使得相对不重要的独立变量系数变为 0，从而被排除在建模之外。因此，它在拟合模型的同时进行特征选择。岭回归也叫线性回归的 L2 正则化(平方根函数)，它将系数值缩小到接近零，但不删除任何变量。岭回归可以提高预测精准度，但在模型的解释上会更加的复杂化。Zeller 等[9]报道了利用 LASSO 模型与人类肠道微生物组作为 CRC (colorectal cancer)筛选工具的可能性，将 CRC 患者与正常组区别开来。

2.1.2 支持向量机：支持向量机(support vector machine，SVM)是一种二分类模型，它的目的是寻找一个超平面来对样本进行分割，分割的原则是从数据点到超平面的最小距离之和最大化，最终转化为一个凸二次规划问题来求解。一般情况

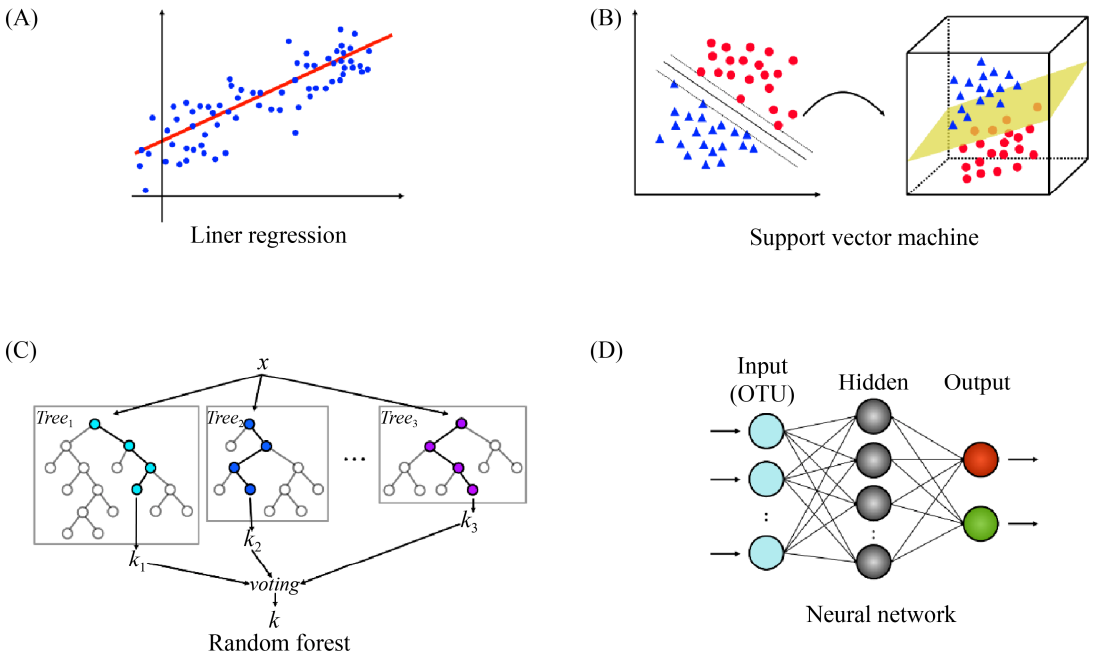

图 2. 四种算法原理图解

Figure 2. Schematic diagram of four algorithms. A: linear regression; B: support vector machine; C: random forest; D: neural network.

下，当数据分布在二维平面线性可分时，用来确定中间直线(即超平面)只需要虚线上几个点即可(图 2-B)，而在三维空间中这些点表现为向量，通过这些向量来支持超平面的生成，距离超平面最近的这些点被称为支持向量。对于更加复杂的数据，往往需要引入核函数，如常用的线性核与高斯核函数，将样本数据从原始空间映射到一个更加高维度的空间，使得样本数据在该空间内是线性可分的。SVM 被广泛应用于生物医学多个领域，例如 Jie 和 Robert 等[10]引入不同核函数构建 SVM 分类器，对 2702 例口腔样本进行分类，探究口腔微生物群落的分类。

2.1.3 随机森林：随机森林(random forest，RF)整合了自举聚合(bootstrap aggregating，bagging)和随机特征选择两种思想。Bagging (装袋)是一种集成学习方法，其使用自助采样法训练多个基础分类器，再将训练好的基础分类器整合起来，获得最终结果。随机森林是 Bagging 的变体之一，通过随机选择变量的一个子集，并使用每一个子集构建决策树，然后以投票方式集成多棵决策树得到最终决策结果(图 2-C)。随机森林可以用来做分类、回归等问题，也可用于执行特征选择或降维，它被公认为是微阵列分析和其他高维数据领域中表现最好的分类器之一。Yatsunenko 等[11]基于肠道菌群 16S 测序数据，应用随机森林算法对样本进行分类，分析了人类饮食与遗传因素对微生物群落的组成与功能的影响。

2.1.4 人工神经网络与深度学习：人工神经网络(artificial neural network，ANN)最初的设计思想是模仿人脑的工作机制，它是一种图计算模型，被称为神经元(neuron)的计算单元通过分层、相互连接从而相互传递信息(图 2-D)。输入层的所有神经元都连接到第一个隐藏层的所有神经元，每个连接用权值表示。微生物组数据常用 OTU 矩阵表作为输入层。隐藏层利用反向传播优化输入变量的权值，提高模型的预测能力。这个过程一直持续到最后一个隐藏层连接到输出层为止。输出层是基于来自输入层和隐藏层的数据的预测。深度神经网络或称深度学习(deep learning)指具有多层的神经网络建模技术，通过添加多个隐藏层或层内的节点数，网络可以表现高度复杂的功能。如果给足够大的模型与足够多的带标签的样本，深度学习将能获得优异的性能。目前常见的深度学习框架有 TensorFlow (https://www.tensorflow.org)与 PyTorch (https://pytorch.org)，他们都提供复杂网络结构建模的工具。深度学习已被广泛应用于有监督或无监督(如自动编码器)的机器学习问题，如 Lo 与 Marculescu 等[12]利用卷积神经网络(CNN)与微生物组数据预测宿主的疾病状态。

2.2 无监督学习算法

与监督学习相反，当标签信息不可用或在建模过程中不使用的情况下称为无监督学习，无监督学习模型可以推断出数据的一些内在结构，主要用于聚类(clustering)与降维(dimension reduction)两个方面。

2.2.1 聚类分析：聚类有助于从已知的数据中探索未知的信息，当样本无标签信息时，基于数据间的相似度将样本分组(簇)，划分依据组内距离最小化而组间距离最大化的原则。例如微生物组数据常按 16S rDNA 序列>97%相似度确定为相同物种(species)。常用的聚类方法有 K 均值(K-means)聚类与层次聚类。聚类最重要的步骤是得到一个好的距离矩阵，图 3 列出常见的不同类别之间距

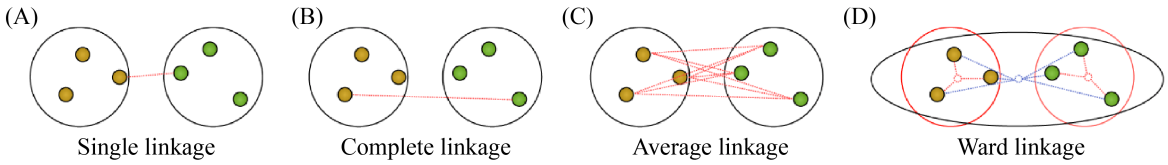

图 3.　常见簇间距离计算方法[5]

Figure 3.　Types of linkage as a method of defining inter-cluster distance[5]. A: single linkage; B: complete linkage; C: average linkage; D: ward linkage.

离的计算方法。微生物组研究中广泛使用的距离度量是 Bray-Curtis 和 UniFrac 差异指标。Bray-Curtis 差异指标是量化两个样本之间菌群组成计数的差异，而 UniFrac 是基于系统发育关系的距离度量指标[7]。对于具有高维特征向量的数据聚类结果，一般通过以下几个指标来评价聚类效果：(1) 聚类中心的距离：距离过大，再分出新类；(2) 聚类域内样本数：聚类数少且离中心点远，考虑为噪音；(3) 聚类域内样本方差：方差过大考虑是否属于该类。

2.2.2　数据降维：微生物组数据通常都是包含大量 OTU 变量的高维数据，降维为我们可视化高维数据提供了一种新思路。常用的降维分析方法主要是主成分分析(PCA)、主坐标轴分析(PCoA)和无度量多维尺度法(NMDS)。PCA 基于方差分解的原理对多维数据进行降维处理，从而提取出数据中的主要因子和结构，称为主成分(PCs)，并以反映数据差异性最大和次大的主成分 PC1 和 PC2 用于高维多变量数据的可视化。PCoA 与 PCA 相似，但利用 Bray-Curtis 或 UniFrac 距离计算样本间的物种差异度，并选取贡献率最大的主坐标组合进行数据可视化，如果样品距离越接近，表示物种组成越相似。Jiang 等[13]基于属水平和 OTU 水平对植物根际沉积物样本的细菌群落进行 PCoA 分析，发现细菌群落动态变化主要受根际效应的影响。

3　机器学习流程

在生物学领域，通过机器学习方法进行数据分析的流程主要分为 4 个方面(图 4)：(1) 数据准备(data preparation)，包括收集基因序列数据、探索数据及数据预处理，如确定数据是否稀疏及数据缺失程度等，从而确定使用的机器学习算法。(2) 特征工程(feature engineering)，对数据进行特征提取、特征选择并将数据分割成训练集与测试集等，准备好数据以输入算法[14]。(3) 模型建立(model generation)，采用训练集进行模型训练，并采用测试集对模型进行验证。

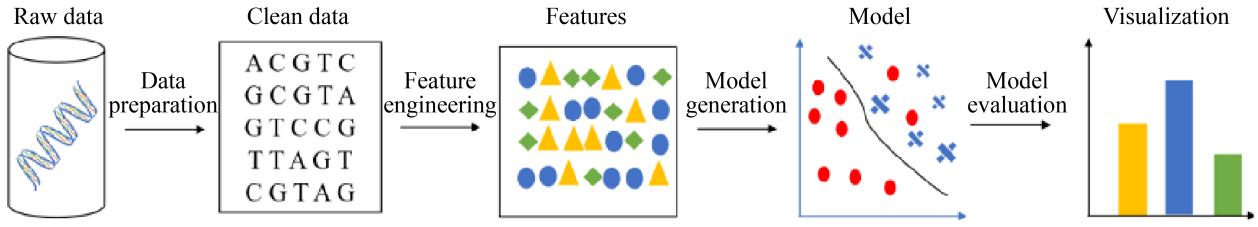

图 4.　机器学习分析流程

Figure 4.　Workflow of machine learning analysis.

(4) 模型评估(model evaluation)，选择 K 折交叉验证法(K-fold cross validation)、自助聚合法(bagging)或留一交叉验证法(leave-one-out cross validation，LOOCV)等对模型的性能进行验证。

下面介绍基于微生物组数据的机器学习流程中的一些关键要点。

3.1 数据预处理

数据预处理主要包括：缺失值处理，即对数据集的错误及缺失数据进行修改或填补；特征提取，即在原始数据中提取更有用的特征或生成新特征；数据降维，用于降低数据的维度，进而降低数据集的复杂度；特征变换，主要包括对数据特征的归一化或标准化等。

在微生物组研究中，根据采集到的样品，获取相关的 OTU 表数据是微生物组研究的重要步骤[15]。由于微生物组数据的样本量一般较少且 OTU 特征量较多，导致数据维度大，难以直接进行分析，所以会使用一些减少特征变量的方法，如 LASSO、主成分分析等剔除不相关或冗余的特征，从而提高模型准确度。最近 Oudah 与 Henschel 提出多层级特征工程(hierarchical feature engineering，HFE)方法，利用 OTU 特征分层结构的相关性与分类信息进行特征提取，取得更好的预测性能[14]。另外，在大数据样本中，不同属性的量纲往往不一致，数值间的差别也比较大，在机器学习过程中会导致“大数吃小数”等问题。因此，需要对数据进行规范化处理，使之归一化在特定的数值区间，以提高机器学习性能。OTU 特征数据一般将每个 OTU 的值除以每个样本的总读数计算其相对丰度用于机器学习。

3.2 算法选择

研究者们提出了各种各样的机器学习算法，算法选择是与预测模型性能优劣有关的重要因素。机器学习算法的选择取决于数据类型、建模目的及应用场景等。没有标签的数据集可用无监督学习算法(如 K-means，PCA)，而有标签数据集可用监督学习算法(如 SVM，RF)。如果标签是类别变量，则可建立分类模型；如果是连续变量，则可建立回归模型。目前多层神经网络的应用日渐成熟，深度学习方法成了热门选择。复杂的模型往往预测性能较好，但模型的可解释性(interpretability)较差[16]。在实际应用中，研究人员希望得到一个可解释性更好的模型用于鉴定与疾病相关的微生物群，而临床医生则希望得到一个预测性能更强的模型。我们通常会选择大家普遍认同的合适算法，如 SVM、RF、ANN 等。最后，集成方法(ensemble methods)可将多个 ML 模型整合成一个预测模型，与单个 ML 模型相比，可以获得更好、更稳健的预测。表 2 中总结了一些常见的机器学习算法在预测宿主表型中的应用，更多应用参见 https://github.com/LiLabZSU/microbioML 中的内容。

3.3 数据抽样与交叉验证

机器学习中，通常会将数据集通过重抽样方法(re-sampling)划分为训练集(training set)和测试集(test set)或验证集(validation set)。用于训练模型的数据称为训练集，而对模型进行测试的数据称为测试集。验证集则用于验证模型的泛化性能(generalizability)，即可用于其他相似数据的能力。现实中的许多数据的特征存在不平衡的问题，进行机器学习时，测试集数据的分布可能与训练集分布存在差异，将对预测结果产生巨大误差。因此，抽取的样本类别应该大致相等，使得

表 2. 机器学习算法在预测宿主表型中的应用

Table 2. Applications of Machine learning algorithms in predicting host phenotype

Algorithm	Trait[a]	Samples (P/N)	Performance[b]	References
SVM	Liver cirrhosis	118/114	AUC (0.980)	Pasolli et al., 2016[17]
RF	PSC	24/24	AUC (0.743)	Iwasawa et al., 2018[18]
ANN	Colorectal polyps	316/236	AUC (0.870)	Dadkhah et al., 2019[19]
Linear regression	IBD	56/56	Accuracy (0.780)	Eck et al., 2017[20]
Naive bayes	Colorectal polyps	316/236	AUC (0.860)	Dadkhah et al., 2019[19]
Boosting	IBD	500/500	F1-macro (0.650)	Lo et al., 2019[12]
K-nearest neighbor	T2D	423/383	F1-score (0.860)	Wu et al., 2018[21]
Deep learning	Oral malodour	45/45	Accuracy (0.967)	Nakano et al., 2018[22]

[a]: PSC: primary sclerosing cholangitis; IBD: inflammatory bowel disease; T2D: diabetes mellitus, type 2. [b]: The performance metrics were described in the model evaluation section below.

到的样本数据集较为平衡。关于不平衡数据处理方法可参考相关文献[5]。

为了充分利用数据集和提高模型的泛化能力，机器学习引入了 K-折交叉验证方法。K-折交叉验证(图 5)通过将数据集 D 分成 K 个子集，编号 D_1、D_2 ...D_k，并依次将其中 1 个子集作为测试集，剩下的 K–1 个子集作为训练集，使得每一条数据都有均等的几率被选中于训练集和测试集，每个训练集得到预测结果 E_k。最后将模型训练和测试后的结果取平均，提高模型的泛化能力。

3.4 模型评估

回归模型的性能评估指标有皮尔斯相关系数(R)、误差平方和(SSE)和均方根误差(RMSE)等。分类模型一般通过计算样本预测值和真实值之间的对应关系来评估模型的性能，评价指标包括预测灵敏度(sensitivity)、精确率(precision)、特异性(specificity)和准确率(accuracy)。灵敏度也称真阳性率或召回率(recall)，表示在所有阳性类别的样本中，被正确预测为阳性的比例(公式 1)；而精确度表示在被预测为阳性的样本中，真正是阳性的比例。特异性也称真阴性率，表示在所有阴性的样本中，被正确预测为阴性的比例(公式 2)。准确率表示预测为阳性与阴性样本的总体准确比例(公式 3)。

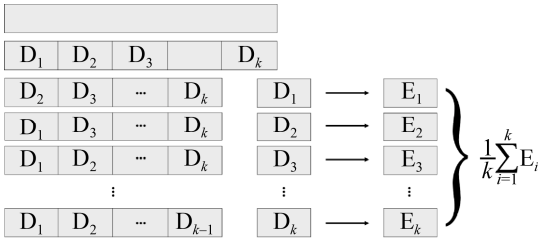

图 5. K-折交叉验证

Figure 5. K-fold cross validation.

$$\text{Sensitivity}(Sn)=\frac{TP}{TP+FN} \quad \text{公式(1)}$$

$$\text{Specificity}(Sp)=\frac{TN}{TN+FP} \quad \text{公式(2)}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \quad \text{公式(3)}$$

其中，TP：true positive；FN：false negative；TN：true negative；FP：false negative。F1 值(F1-score)是精确率和召回率的加权调和平均值(最大值是 1，最小值是 0)。它有两种计算方式分别为 F1-micro 与 F1-macro，在二分类问题中两者计算方式完全一致，但在多分类问题中有差异。

ROC 曲线(receiver operating characteristic curve)，又称“受试者工作特性曲线”，最直观的应用就是能反映模型在选取不同阈值的时候其敏感性和特异性的趋势(图 6)。ROC 曲线的横坐标轴(1-specificity)表示假阳性率(FPR)，即真阴性条件下，预测为阳性；纵坐标表示敏感性(sensitivity)或真阳性率(TPR)。根据 ROC 曲线位置，曲线下方部分的面积被称为 AUC (area under curve)，用来表示预测准确性，AUC 值越高，说明预测准确率越高，即曲线越接近左上角，预测准确率越高。一般 AUC 在 0.5 到 1.0 之间，当 AUC<0.5 时，说明模型没有意义。Ai 等[23]通过 AUC 显示不同模型的预测效果，得出 Bayes 和随机森林模型具有较好的预测能力。

4 案例分析

本案例来自 Yoshio 等[22]通过机器学习基于口腔微生物群进行口腔异味预测的研究，基于 90 例受试者唾液样本的 16S rRNA 基因扩增子数据分析所得到的 OTU 矩阵，对口腔异味和健康呼吸进行机器学习分类预测。本例将通过开源免费软件 R 中的随机森林、支持向量机等机器学习算法得到分类预测模型，并绘制 ROC 曲线以评价各算法的预测性能。

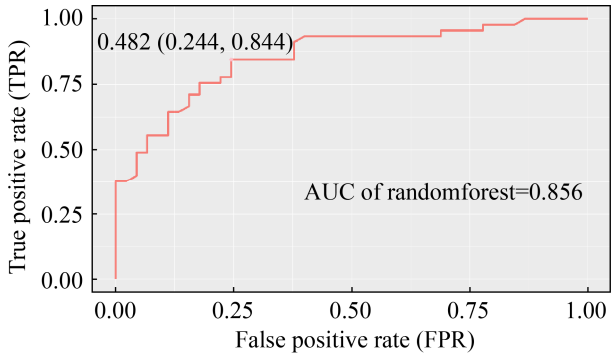

图 6. 口臭和健康呼吸分类的 ROC 曲线

Figure 6. ROC curve for classification of malodourous and healthy breath.

4.1 硬件与软件

4.1.1 硬件：运行 Windows、Linux 或 macOS 操作系统的 64 位计算机；计算机的处理器与内存需求依赖分析数据的大小与分析模型。

4.1.2 软件：使用开源免费软件 R (https://www.r-roject.org)，想要了解更多基于 R 实现的机器学习库，可以参考 CRAN 中机器学习的任务视图(https://cran.r-project.org/web/views/MachineLearning.html)。

通过以下命令安装本案例需要的 R 包：install.packages (c(“randomForest”, “pROC”, “ggplot2”))。

4.2 数据集(dataset)

Yoshio 等使用的研究对象包括 90 名患者(男性 37 名，女性 53 名，平均年龄 50.0±14.7 岁)，他们于 2011 年 8 月至 2016 年 10 月期间在福冈牙科医学院和牙科医院口腔异味诊所就诊，主述口臭。3 个月内未服用抗生素，无耳鼻喉或代谢性疾病。在这 90 位病人中，45 位没有或有微弱的口腔异味，45 位有明显的口腔异味。Nakano Y 等[22]的文章中文件 12903_2018_591_MOSESM1_ESM.csv (此处重命名 S1_table.csv)是一个 90×109 的 OTU 数据集(https://github.com/LiLabZSU/microbioML)，第二列“Malodour”表示呼吸为有口臭(P，阳性)或正常(N，阴性)。

4.3 建立模型(model)

下面介绍如何在 R 中基于微生物组数据构建口腔异味预测模型的分析过程。关于模型构建更详细代码参考 https://github.com/LiLabZSU/microbioML 中的内容。

4.3.1 加载数据：在 R 中设置本机工作目录在“D:\microbioML”，并载入所需要的数据集：S1_table.csv(样本中 OTU 丰度数据集)。

```
setwd(“D:\microbioML”) # 按数据实际路径修改
data <- read.csv(“S1_table.csv”, header=TRUE, row.names=1, sep=“,”)
data$Malodour<-as.factor(data$Malodour)
head(data) #显示部分数据
```

4.3.2 数据分次交叉验证：进行交叉检验首先要对数据分组，数据分组要符合随机且平均的原则。K-折交叉验证(K-fold cross-validation)是交叉验证方法里一种，数据分折(k-fold)函数如下：

```
CVgroup <- function(k, datasize, seed){
cvlist <- data.frame()
  set.seed(seed)
  n <- rep(1:k, ceiling(datasize/k))[1:datasize] #将数据分成 k 份，并生成的完整数据集 n
  temp <- sample(n, datasize) #随机化
  x <- 1:k
  dataseq <- 1:datasize
  cvlist <- lapply(x, function(x) dataseq[temp == x]) #随机生成 k 个随机有序数据列
  return(cvlist)
} #定义分折函数
k <- 10      #k-fold 设为 10
datasize <- 90 #样本数为 90
cvlist <- CVgroup(k = k, datasize = datasize, seed = 123) #代入参数
```

4.3.3 随机森林模型：随机森林预测模型使用 R 包 randomForest 构建，变量 pred 显示实际值与预测初步结果概率的数据框，方便绘制 ROC 曲线。

```
library(randomForest)   #载入模型库
pred <- data.frame() #存储预测结果
for (i in 1:10) {
  train <- data[-cvlist[[i]],] #训练集
  test <- data[cvlist[[i]],] #测试集
  #建立 randomforest 模型，ntree 指定树数
  rf.model <- randomForest(Malodour ~ ., data = train, ntree=100)
  summary(rf.model)
  rf.pred <- predict(rf.model, test, type = "prob")[,2] #预测
  kcross <- rep(i, length(rf.pred)) #i 为第几次循环交叉，共 K 次
  temp <- cbind(Malodour = test$Malodour, rfPredict = as.data.frame(rf.pred), kcross)
  pred <- rbind(pred, temp)
}#循环计算 K 次模型
head(pred)
```

4.3.4 绘制 ROC 曲线：通过 pROC 包里的 roc 函数和 ggplot2 里的 ggroc 函数绘制 ROC 曲线。

```
library(pROC)
library(ggplot2)
rf.roc <- roc(pred$Malodour, as.numeric(pred$rf.pred))
rf.roc$auc
ggroc(rf.roc, alpha=0.5, colour=“red”, linetype=1, size=2, legacy.axes = TRUE) +
annotate(“text”, x = .75, y = .15, label = paste(“AUC of Random Forest =”, round(rf.roc$auc,3)))
```

4.4 模型性能评价

与文献[22]中报道的 SVM 机器学习结果类似，本研究对唾液样本的 16S rRNA 基因扩增子测序所得的微生物组成数据进行随机森林分类器预测的正确率(accuracy)也为 80%左右(图 6)。

ROC 曲线显示随着真阳性率(sensitivity)的上升，假阳性率(1-specificity)也相应增加。最接近 ROC 曲线左上角的那一点为最佳临界点(0.244，0.844)，对应分类阈值是 0.482，模型的灵敏度和

特异度达到较好平衡。ROC 曲线下面积(AUC)体现模型准确性,随机森林模型的 AUC 值为 0.856,说明模型预测准确性较高,即当输入一组新样本数据时可以较好地判断出样本来源受试者是否有口腔异味。

5 展望

微生物对其宿主表型的影响往往不是通过某一两种细菌来决定,而是通过微生物群的协同作用来决定。因此,一旦我们确定了所研究宿主的微生物组特征,可以使用机器学习方法来探索微生物组与宿主表型之间的相互作用关系。微生物组机器学习研究一般首先以 16S rRNA 基因序列聚类后的 OTU 表作为微生物概况数据,进行数据预处理与特征选择;然后使用合适的机器学习算法构建预测模型,并利用测试数据集评估模型性能。虽然机器学习等人工智能方法已成功应用于生物医学领域,但目前还面临许多问题,如模型的精度有待提高、泛化能力还需加强等,尤其微生物组数据往往存在研究样本量少、特征多导致过拟合的问题。为了应对这些挑战,机器学习未来还需要在特征工程与算法等方面有所改进。目前大多微生物组研究基于 16S rRNA 基因测序数据,而宏基因组测序(shotgun metagenomic)数据能提供菌种(species)或菌株(strains)水平的分类信息[24],并可采用 K-mers 代替 OTUs 作为特征,由于 K-mers 分析不需要序列比对,将大大加快分析速度。另外,随着大样本队列微生物组研究的增加,深度学习算法将能应用于构建预测模型,而对于大样本量数据(如>1 万例),深度学习会有更精确的预测结果[6]。最后,为衡量 ML 算法模型的可靠性,还需一个由全球科学社区共同建设并维护的公共微生物组数据库,如 Microbiome Learning Repo (ML-repo)[25]。总之,随着基因组测序与人工智能技术的快速发展,机器学习将成为基于微生物组数据预测宿主性状或疾病诊断的重要工具。

参 考 文 献

[1] Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu WH, Lakshmanan A, Wade WG. The human oral microbiome. *Journal of Bacteriology*, 2010, 192(19): 5002–5017.

[2] Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, Gasbarrini A, Tortora G. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology & Hepatology*, 2020, 17(10): 635–648.

[3] Zhang GQ, Huang ZQ, Wang MY, Kong JH, Li YD, Chen JS. Association between dietary habits and salivary microbial diversity in college students. *Food Science*, 2019, 40(1): 196–201. (in Chinese)
张国庆, 黄子琪, 王明月, 孔俊豪, 李余动, 陈建设. 大学生饮食习惯与唾液微生物多样性的关联. 食品科学, 2019, 40(1): 196–201.

[4] Mingle D. Machine learning techniques on microbiome-based diagnostics. *Advances in Biotechnology & Microbiology*, 2017, 6(4): 97–99.

[5] Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Frontiers in Genetics*, 2019, 10: 579.

[6] Li Y, Huang C, Ding LZ, Li ZX, Pan YJ, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, 2019, 166: 4–21.

[7] Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu QY, Caporaso JG, Dorrestein PC. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 2018, 16(7): 410–422.

[8] Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 2019, 50: 71–91.

[9] Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 2014, 10: 766.

[10] Ning J, Beiko RG. Phylogenetic approaches to microbial community classification. *Microbiome*, 2015, 3(1): 1–13.

[11] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Gregory Caporaso J, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. Human gut microbiome viewed across age and geography. *Nature*, 2012, 486(7402): 222–227.

[12] Lo C, Marculescu R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics*, 2019, 20(Suppl 12): 314.

[13] Jiang XT, Peng X, Deng GH, Sheng HF, Wang Y, Zhou HW, Tam NFY. Illumina sequencing of 16S rRNA tag revealed spatial variations of bacterial communities in a mangrove wetland. *Microbial Ecology*, 2013, 66(1): 96–104.

[14] Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics*, 2018, 19(1): 227.

[15] Namkung J. Machine learning methods for microbiome studies. *Journal of Microbiology*, 2020, 58(3): 206–216.

[16] Topçuoğlu B, Lesniak N, Ruffin M, Wiens J, Schloss P. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *mBio*, 2020, 11(3): 1–13.

[17] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Computational Biology*, 2016, 12(7): e1004977.

[18] Iwasawa K, Suda W, Tsunoda T, Oikawa-Kawamoto M, Umetsu S, Takayasu L, Inui A, Fujisawa T, Morita H, Sogo T, Hattori M. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. *Scientific Reports*, 2018, 8: 5480.

[19] Dadkhah E, Sikaroodi M, Korman L, Hardi R, Baybick J, Hanzel D, Kuehn G, Kuehn T, Gillevet PM. Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterology*, 2019, 6(1): e000297.

[20] Eck A, Zintgraf LM, de Groot EFJ, de Meij TGJ, Cohen TS, Savelkoul PHM, Welling M, Budding AE. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinformatics*, 2017, 18(1): 1–13.

[21] Wu HL, Cai LH, Dongfang L, Wang XY, Zhao SC, Zou FH, Zhou K. Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *BioMed Research International*, 2018, 2018: 2936257.

[22] Nakano Y, Suzuki N, Kuwata F. Predicting oral malodour based on the microbiota in saliva samples using a deep learning approach. *BMC Oral Health*, 2018, 18(1): 128.

[23] Ai LY, Tian HY, Chen ZF, Chen HM, Xu J, Fang JY. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget*, 2017, 8(6): 9546–9556.

[24] Carrieri AP, Rowe WP, Winn M, Pyzer-Knapp EO. A fast machine learning workflow for rapid phenotype prediction from whole shotgun metagenomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33: 9434–9439.

[25] Vangay P, Hillmann BM, Knights D. Microbiome Learning Repo (ML Repo): a public repository of microbiome regression and classification tasks. *Giga Science*, 2019, 8(5): giz042.

Applications of machine learning in predicting host phenotype based on microbiome

Gaolei Li, Wei Huang, Hao Sun, Yudong Li*

School of Food Science and Biotechnology, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang Province, China

Abstract: With the advent of the era of big data, how to transform the omics data into easy-to-understand and visualized knowledge is one of the important challenges in bioinformatics. Recently, machine learning techniques had been utilized to analyze the complicated, high-dimensional microbiome data to address the complex mechanisms of human diseases. Here, we firstly summarized microbiome data procession approaches and the most commonly used machine learning algorithms, such as support vector machine (SVM), random forest (RF), and artificial neural networks (ANN). Then, the workflow of machine learning studies was described, and the application of ML algorithms in predicting host phenotypes based on microbiome data was evaluated. Finally, the model construction and validation of machine learning algorithms were demonstrated by using saliva microbiome data to predict oral malodour as an example, and R/Python code for practical data analysis was provided (https://github.com/LiLabZSU/microbioML).

Keywords: machine learning, microbiome, big data, host phenotype, prediction

(本文责编：张晓丽)

Supported by the National Natural Science Foundation of China (31671836)

*Corresponding author. Tel: +86-571-28008900; E-mail: lyd@zjsu.edu.cn

Received: 10 October 2020; Revised: 8 March 2021; Published online: 26 March 2021