生物技术进展 2023 年 第 13 卷 第 5 期 671~680
Current Biotechnology ISSN 2095-2341

进展评述
Reviews

机器学习在肠道菌群宿主表型预测中的应用

曹海涛， 朱静*，马云鹏， 崔兴华

新疆农业大学计算机与信息工程学院，乌鲁木齐 830052

摘 要： 随着第二代DNA测序技术的发展，研究人员积累了大量的肠道菌群数据，研究表明肠道菌群与宿主健康状况存在密切联系，因此如何对复杂、高维的肠道菌群数据进行建模分析，是当前生物信息学研究中的重要挑战。人工智能的兴起为处理肠道菌群数据，揭示肠道菌群与宿主表型之间的复杂关系提供了可能。综述了现阶段肠道菌群与宿主表型之间的相关研究，重点介绍了常用的5种机器学习算法（线性回归、支持向量机、K-近邻、随机森林、人工神经网络）的理论原理及在相关研究中的应用，对预测宿主表型的机器学习算法选择提出了建议，并对该领域的未来发展进行了展望，以期为利用机器学习对肠道菌群宿主表型预测提供参考依据。

关键词： 肠道菌群；机器学习；建模预测

DOI：10.19586/j.2095-2341.2021.0201

中图分类号： Q93, TP181 **文献标志码：** A

Application of Machine Learning in Phenotypic Prediction of Gut Microbiota

CAO Haitao , ZHU Jing* , MA Yunpeng , CUI Xinghua

College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, *China*

Abstract: With the development of second-generation DNA sequencing technology, a large amount of gut microbiota data has been accumulated. The studies showed that gut microbiota were closely related to the health status of the host. Therefore how to model and analyze the complex and high-dimensional gut microbiota data has been an important challenge faced by bioinformatics at present. The rise of artificial intelligence had made it possible to process gut microbiota data and revealed the complex relationship between gut microbiota and host phenotypes. The paper summarized the present stage of gut microbiota and phenotypic correlation study among five kinds of machine learning algorithm (commonly used linear regression, support vector machine (SVM), K-nearest neighbor, random forests and artificial neural network), introduced five kinds of machine learning algorithms of theory and application in the related research, and to choose what kind of machine learning algorithms to predict recommendations to host phenotype. Finally, the future development of this field was prospected to provide a reference for predicting host phenotypes using machine learning based on gut microbiota data.

Key words: gut microbiota; machine learning; modeling and prediction

肠道菌群是指生活在宿主肠道内所有微生物的集合，包括细菌、病毒和真菌。越来越多的研究显示，宿主的健康状况与肠道菌群存在密切联系。高通量测序技术的应用及各个国家支持的大规模肠道菌群计划的实施，为揭示肠道菌群与宿主的健康状况提供了必要的数据支撑，同时也产生了大量的微生物组数据，如人类微生物组项目（human microbiome project，HMP）[1]、比利时弗莱明肠道菌群计划（Flemish gut flora project，FGFP）[2]和我国开展的广东省肠道菌群计划[3]等。随着人工智能的兴起，适用于复杂数据分析的机器学习受到了研究人员的青睐。例如，Najafabadi等[4]探究了深度学习在大数据分析中的应用和挑战；Hernández等[5]探究了机器学习和深度学习在

收稿日期： 2021-12-29；**接受日期：** 2022-05-16

基金项目： 新疆畜牧科学院畜牧研究所基础研究项目（2020BD1002-2-2-2）。

联系方式： 曹海涛 E-mail: 2232060551@qq.com；*通信作者 朱静 E-mail：zhujing@xjau.edu.cn

微生物组研究中的应用。利用微生物组数据+机器学习来进行医疗诊断已成为生物医学领域一个新兴的研究热点。

机器学习可作为微生物组数据的处理方法，如主成分分析、数据归一化、特征选择等。原始数据经过数据处理后可以消除冗余的数据，改变微生物组数据高维、稀疏的特点，并在一定程度上提升模型预测的精度；同时机器学习也可作为预测模型的核心建模算法，包括K近邻（K nearest neighbors，KNN）[6]、支持向量机（support vector machine，SVM）[7]、人工神经网络（artificial neutral network，ANN）等。Hacllar等[8]利用KNN构建炎症性肠病预测模型；Assegie等[9]使用K-近邻（KNN）算法和SVM构建了肝病分类模型；Liu等[10]使用SVM构建肥胖预测模型；Reiman等[11]使用ANN构建肝硬化预测模型；Nasser等[12]使用人工神经网络构建肺癌检测模型；Lyngdoh等[13]利用5种监督机器学习算法分析糖尿病模型的预测，使用KNN分类器实现了76%的稳定和最高准确度等。但这些预测模型都是基于特定的机器学习算法和微生物组数据，因此普遍存在在特定数据集表现良好，而泛化能力不足的情况。

本文综述了机器学习算法在基于肠道微生物组数据预测宿主表型方面中的应用，以及肠道微生物及微生物组中常用的5种机器学习算法（线性回归、支持向量机、K-近邻、随机森林、人工神经网络）的原理，重点归纳了机器学习算法在肠道菌群与宿主健康相关研究中的应用现状，应用机器学习算法构建预测模型的一般规律，以期为推动机器学习进行肠道菌群宿主表型预测提供参考依据。

1 肠道微生物概述

1.1 肠道菌群

人体肠道内含有大量的共生菌，由上千种微生物组成，包括古生菌、真菌、细菌、原生生物、病毒等，其中细菌是最主要的定殖菌[14]，因此肠道是人体微生物菌群最复杂的部位之一。目前，尚无研究证明肠道菌群中细菌种类的确切数目，一般认为肠道菌群中含有500~1 000种细菌[15]，但也有研究者发现肠道菌群中细菌的种类超过3 500种[16]，数量约为100万亿，总重量约为1~2 kg。由此可知，肠道菌群是人体免疫有机体的重要组成部分[17]，也被认为是人体肠道内的另一个“器官”[18]。

肠道中的微生物大多为专性厌氧菌，种类超过50个门[19]，如此庞大数量的细菌处于动态平衡状态中，具有高度的多样性、稳定性、抗逆性和耐药性，而肠道微生物菌群的紊乱则与多样性和共生性的丧失有关[20]。肠道菌群中主要有拟杆菌、乳杆菌、大肠杆菌、肠球菌4种细菌，其中拟杆菌属和犁头霉属在肠道微生物中的丰度最高，占肠道微生物总量的90%以上[21]。这些数量众多的肠道微生物主要通过自身的代谢产物或代谢产生的活性成分来调节宿主的新陈代谢，进而影响宿主的健康状况。

1.2 肠道菌群与宿主之间的关系

宿主表型是指为微生物菌群定殖以及其他寄生生物提供生存环境的生物体可观察的性状或特征，如生理、生化和行为方面的特性，是被定殖或寄生生物体所有性状的总和。肠道菌群可以提高宿主的免疫机能，促进营养物质吸收[22]，维持宿主免疫屏障的完整性[23]。研究发现，肠道菌群消化产物短链脂肪酸是宿主肠道上皮细胞的重要营养物质，可以促进宿主肠道上皮细胞的生长及分化，对维持肠道屏障的完整性具有重要作用[24]，可防止肠源性内毒素进入血液引起代谢性内毒素血症[25]；同时，宿主所处的地理环境、年龄、饮食习惯、服用药物史、疾病以及细菌之间的相互作用均会影响肠道菌群的丰度[26]。

肠道菌群会影响宿主免疫系统功能，而肠道菌群丰度和肠道微生态结构的改变可以引起肠道菌群失调[27]。一旦发生肠道菌群失调，肠道内的有益菌群（如双歧杆菌、乳酸菌、拟杆菌等）就会减少，而有害菌群（如产生毒素的拟杆菌，大肠杆菌、梭菌等）则会增加，且有害菌分泌的多种毒性因子会损伤肠道上皮细胞，导致多种疾病的发生，如肠易综合征（irritable bowel syndrome，IBS）[28]、结直肠癌（colorectal cancer，CRC）[29]、炎症性肠病（inflammatory bowel disease，IBD）[30]、自闭症（autism spectrum disorder，ASD）[31-32]、肥胖（obese）[33]、2型糖尿病（type 2 diabetes，T2D）[34]等。上述研究表明肠道菌群与宿主的多种疾病存在相关性，研究肠道菌群与宿主之间的关系，可为精准医疗提供可能[35-36]，进而使利用肠道菌群干预宿主的疾病治疗成为现代医学治疗的一种新兴手段[37]。

近年来的研究表明,肠道菌群与宿主的健康状态和疾病之间存在密切关联。这意味着肠道菌群的组成和丰度可能与宿主的疾病风险、发展和病程有关。这种关联不仅涵盖了消化系统相关的疾病,还包括了许多其他疾病,如免疫系统疾病、代谢性疾病和神经系统疾病等。

2 基于机器学习的研究进展

2.1 机器学习的发展

随着人工智能的兴起与发展,目前机器学习已应用于生命科学的各个领域,如癌症检测、药物开发、行为预测、人脸识别、语义分析、推荐个性化治疗等,且在复杂的微生物组学相关研究中应用效果显著[38]。第二代DNA测序技术的普及使微生物组学数据激增,传统的人工统计学方法已经无法适应这种高维、稀疏、数据量庞大的微生物组学分析,而机器学习可以从海量复杂的数据中,挖掘其内部潜在的信息,节省了大量人力和时间,提高了工作效率,已经逐渐成为微生物组学研究的主流方法[39]。而随着机器学习、计算机硬件及相关数学理论的发展,产生了一种新技术方法——深度学习(deep learning,DL)。该方法无需人工干预就可以自动捕捉到复杂数据中隐藏的数据结构,将其应用于肠道菌群数据分析中,可以揭示菌群与宿主健康之间的关系,从而对宿主的疾病及健康状况等方面进行决策[40]。尽管目前机器学习尚未普及到临床应用中,但这预示着未来有望充分利用机器学习技术来处理、分析和解释大规模的微生物组数据,从而深入理解微生物与宿主之间的相互作用,为医学、生态学和生物技术领域带来新的突破和创新。

2.2 基于微生物研究的相关机器学习算法选择

人工智能发展主要有机器学习、自然语言处理、基于规则的专家系统和机器人学习这4种类型[41]。机器学习可以在短时间内处理大量的数据,但是也受制于计算机的处理能力、数据量的大小及算法复杂性。截至目前,机器学习已成为微生物菌群领域中最常用的人工智能技术[42]。机器学习是一门多领域交叉学科,涉及统计学、概率论、最优化、凸分析等学科,其主要特点是模仿人类的学习行为,从复杂的数据规律或模式中获取新的知识,挖掘其中潜在的信息,是人工智能的核心。机器学习通常按照数据是否带有标签分为有监督学习和无监督学习[43]。按照数据是否为离散型,合为分类问题和回归问题[44]。宿主表型预测是利用带有标签的肠道菌群数据对机器学习模型进行训练,利用输入的肠道菌群数据预测宿主的健康情况,即为有监督的学习。常用于肠道菌群分析的5种机器学习算法有支持向量机(support vector machine,SVM)、K-近邻、线性回归、随机森林和人工神经网络。

2.2.1 *支持向量机* 支持向量机是一种二元分类模型,其目的是寻找一个超平面对数据进行划分,可以使用核函数进行非线性分类。对高维的肠道菌群数据具有很好的适用性,是肠道菌群领域应用较广泛的一种机器学习模型。2018年,Xu等[45]利用支持向量机构建预测模型,根据基因编码蛋白序列信息预测阿尔茨海默病(alzheimer disease,AD),准确率达到85.7%。有研究利用支持向量机和人类微生物组项目数据库构建微生物组分类器,结果发现分类精度、敏感性和特异性均较高[46]。SVM用于诊断皮肤病和预测心血管疾病,准确率分别达到95.39%和85%[47]。

如图1A所示,支持向量机的目标是在两个类别之间创建一个决策边界,从而能够在一个或多个特征向量中预测标签。该决策边界又称为超平面,以这样一种方式定向,其距离可能是从每个类别中最接近的数据点,而这些最近的点被称为支持向量。按公式(1)给定一个标记的训练数据集。

$$(x_1, y_1), \cdots, (x_n, y_n), x_i \in R^d, y_i \in (-1, +1) \quad (1)$$

式中,x_i是一个特征向量,y_i是训练化合物i的类别标签(负或正)。最优超平面可以定义为公式(2)。

$$wx^T + b = 0 \quad (2)$$

其中,w是权重向量,x是输入特征向量,b是偏差。

支持向量机的另一种用途是核方法,它使我们能够对高维的非线性模型建模。在非线性问题中,可以使用核函数向原始数据添加额外的维度,从而使其在高维空间中成为线性问题,如图1B所示,在二维数据无法线性划分时将二维上升到三维以成功创建超平面。

支持向量机的优点在于:①复杂性主要取决于支持向量的数目,而不是高维的样本空间,可以减轻高维的微生物数据所造成的影响;②对数据

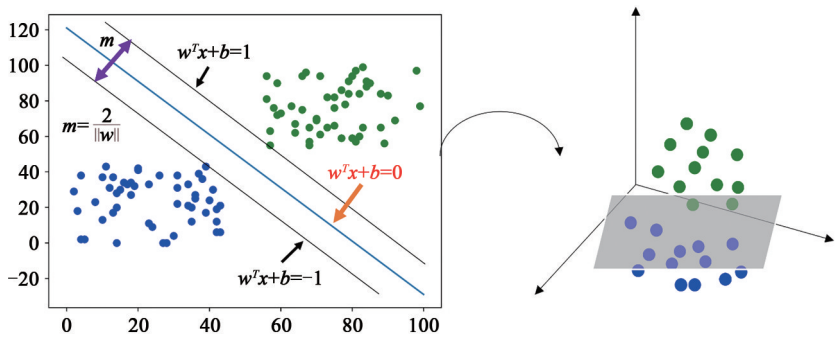

图1 支持向量机

Fig. 1 Support vector machine

的异常值不敏感，具有较好的鲁棒性；③可以使用凸优化找到全局最小值；④适用性较广泛。而支持向量机的缺点在于：①对多分类问题表现不够好；②对大数据量的计算周期较长；③对自身参数选择比较敏感。

2.2.2 K近邻 K近邻是根据距离选取K个样本点数据来推测预测点的类别。2018年，Wu等[48]利用K近邻证明了2型糖尿病（type 2 diabetes，T2D）、类风湿性关节炎（rheumatoid arthritis，RA）和肝硬化（liver cirrhosis，LC）等疾病的微生物组生物标志物与表型之间存在显著相关性。

如图2所示，测试样本应归入第一类的蓝色三角形或是第二类的五角星形。如果k=3（虚线圆圈）它被分配给第一类，那么有2个三角形和1个五角星形在内侧圆圈之内。如果k=11（实线圆圈）它被分配到第二类（5个三角形与6个五角星形在外侧圆圈之内），同样的方法也可以扩展到三维空间。

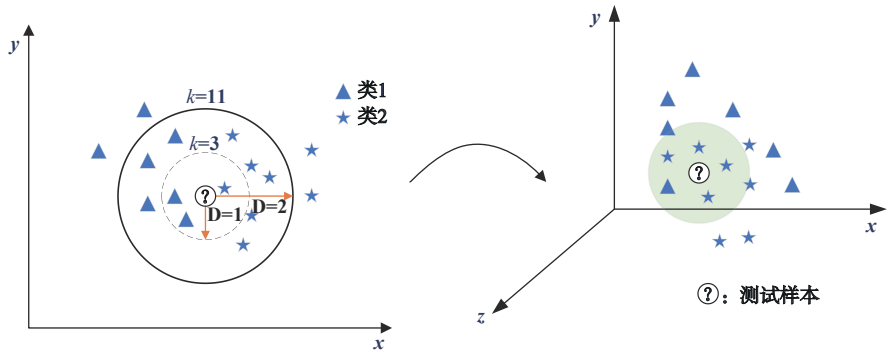

图2 K近邻

Fig. 2 K nearest neighbors

K近邻算法的优点在于：①容易理解，易实现；②适用于非线性分类；③算法调整方便，且便于调整K的数量以及距离；④对数量大的样本具有较好的适用性。K近邻算法的缺点在于：①对特征比较多的样本计算开销较大；②对样本不均衡的情况表现较差。

2.2.3 线性回归 线性回归指利用线性方程对数据进行拟合，是最常见的回归算法，其含有1个自变量和1个因变量，且二者存在线性关系，即可用一条直线表示，也被称为一元线性回归。肠道菌群数据通常含有2个以上的自变量，多采用多元线性回归，其最重要的2个变形是加入了L1正则化的Lasso回归和L2正则化的岭回归。Lasso回归的突出优势是加入了惩罚函数，使得相对不重要的特征项系数变为0，相当于进行了特征选择。岭回归则是将特征系数缩小到接近0，而不删除任何特征项，提高了预测精度，但也增加了解释复杂度。2021年，Yao等[49]利用线性回归观测到结直肠癌（colorectal cancer，CRC）患者微生物菌群多样性降低，且利用分辩微生物组方法可以有效检测结直肠癌。Li等[50]研究了基于线性回归的蛋白质中锌结合位点预测的整合方法，可以应用于

基于序列信息的锌结合位点识别，也可用于推断蛋白质功能，并且更有利于治疗某些疾病。

如图3所示，展示了一个横坐标表示真实值，纵坐标表示预测值的散点图，线性回归就是要找到一条直线（图中的红色线）来尽可能地拟合图中的数据点。

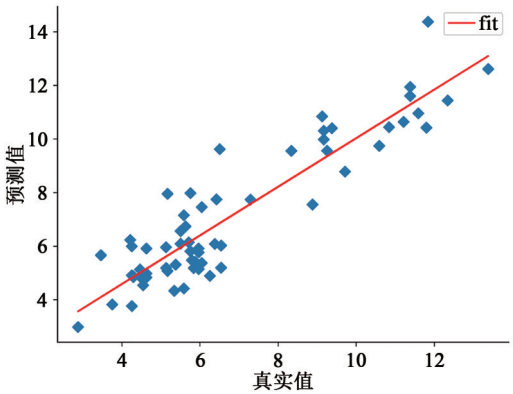

图3　线性回归

Fig. 3　Linear regression

线性回归的优点在于：①对小数据量、关系结构较为简单的样本效果较好；②算法较为基础，容易理解，可解释性较强。线性回归的缺点在于不能较好地拟合非线性数据。

2.2.4　*随机森林*　随机森林的本质是包含多个决策树的分类器的集合，而决策树的优势在于使数据形式易于理解[51]。决策树可以从众多不熟悉的数据集合中提取出一系列规则，创建规则的过程就是机器学习的过程。随机森林是一种在生物学和基因组学中应用越来越广泛的方法，其不仅适用于二分类，也适合多分类。Pasolli等[52]根据随机森林构建的炎症性肠病预测模型准确率达到0.89，肥胖预测模型准确率达到0.66。Yang等[53]采用多种方法构建华东地区心血管疾病模型，包括多元回归模型、分类和回归树、朴素贝叶斯、袋装树、Ada Boost和随机森林，实验结果表明随机森林优于其他方法，曲线下面积（area under curve，AUC）为0.787，且比基准有显著改善。

图4展示了随机森林的示例：首先对数据集使用Bootstrap方法对样本进行重抽样，然后将得到的每个样本输入决策树中进行分类，最后将若干个弱分类器的分类结果进行投票选择，根据投票决定最终结果。

随机森林算法的优点在于：①对复杂高维的数据展现出较好的适用性；②可用于筛选重要特征；③泛化能力较强；④可以处理样本的缺失特征。随机森林的缺点在于：①偏向选择投票最多的特征；②可能产生过度匹配的问题。

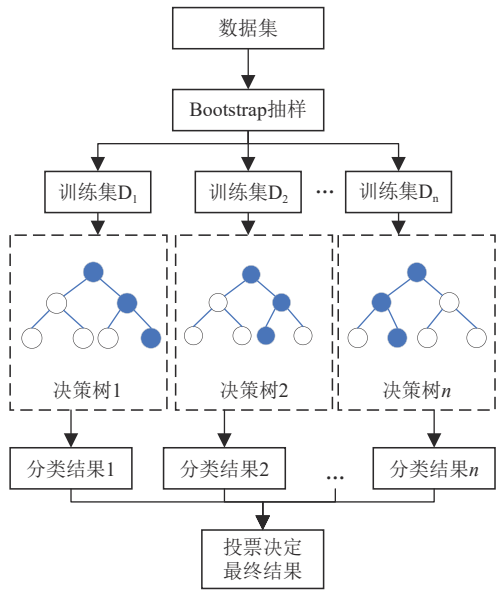

图4　随机森林

Fig. 4　Random forests

2.2.5　*人工神经网络*　人工神经网络作为一种运算模型，是对人脑神经元网络的抽象，由大量神经元节点相互连接而成，每个节点就是一种特定的激励函数。两个节点之间连接信号的加权值称为权重，相当于人工神经网络的记忆，其主要包含输入层、隐藏层、输出层3个部分，输入层接收外部的数据；隐藏层不能由系统外部观察；输出层实现结果的输出。使用人工神经网络作为预测模型时，通常对数据量有极高的要求，并且训练中参数的调参也更为严格，训练结果也更加不可预知和不可解释。2017年，Reiman等[11]利用卷积神经网络（convolution neural network，CNN）构建疾病预测模型，分类精度较传统方法更高。Tejamma等[54]使用卷积神经网络模型来预测心脏病，取得了非常好的效果。

图5展示人工神经网络模式：网络最左的一层为输入层，将多组数据（比如OTU_1到OTU_n）输入到输入层中的n个输入神经元中，输入层中的数据传输到隐藏层中，隐藏层会根据已经训练好的参数对数据进行处理，最后隐藏层将数据传输到输出层，并由输出层将结果输出。

人工神经网络的优点在于：①相较于传统机

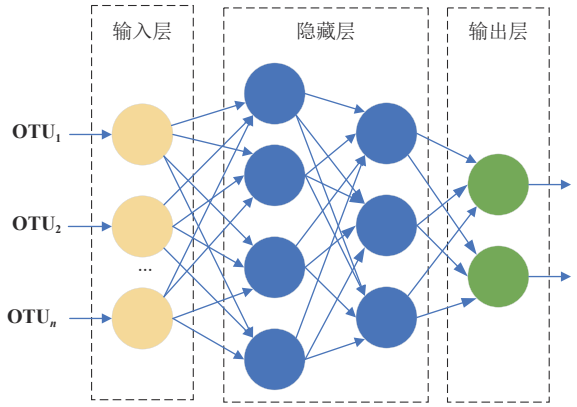

图5 人工神经网络

Fig. 5 Artificial neutral network

器学习，人工神经网络可以处理海量数据；②计算能力较强；③算法不断被优化。而人工神经网络的缺点包括：①“黑箱”操作，结果解释性不高；②计算耗时耗力；③模型训练需要更多的数据来满足。

3 机器学习在肠道菌群中的相关研究

3.1 基于肠道菌群的相关研究

1917年，Wehkamp等[55]首次分离出大肠杆菌，明确了微生物菌群在宿主中具有抵抗有害菌的作用。1965年，Schaedler等[56]首次将微生物菌群移植到无菌老鼠体内，揭示了微生物菌群对宿主健康发育的重要性，这创立了利用无菌宿主研究肠道菌群作用的新方法。1989年，研究发现微生物菌群对宿主的免疫系统具有调节作用[57]。2005年提出的第二代测序技术显著提升了基因测序深度，可以从分类层级上分析微生物菌群，有助于研究者深入了解微生态的功能与特征[58]。2007年实施的人类微生物组项目[59]以及2012年开始的美国肠道菌群计划[60]标志着微生物菌群研究从个体走向大规模人群。

第二代DNA测序技术可对人体皮肤、口腔、胃、肠道、腹腔等部位的微生物群落进行分析，这些微生物群落即为人类微生物群。研究发现，微生物群对人类健康有重要影响[61-62]，因此，对这些微生物菌群的研究，有利于研究人员开发新的诊断工具和治疗方法以判断人类身体健康状况和治疗相关疾病[63-64]，但不同的方法诊断和治疗结果可能存在明显的差异[65-66]。随着微生物组数据的不断增加，仅依靠传统的人工统计方法可能需要几个月甚至几年的时间，而人工智能为分析海量数据提供了一种快速高效的方式，目前已经广泛运用于微生物组学相关研究中。

3.2 机器学习在微生物对宿主疾病预测方面的应用

近年来，利用机器学习预测疾病的相关研究较多（表1），其具有良好的疾病预测能力，且可根据特征选择和特异性标记提高预测精度[67]。已有研究证实，唾液微生物群可以作为无创诊断胆管炎的标记物[68]和预测口腔异味（预测精度达97%），并且深度学习可以获得比传统机器学习更高的准确率[69]；Dadkhah等[70]研究发现监督式机器学习算法对复杂高维的微生物群数据有更好的适用性，并且进行特征选择可以有效地提高预测精度。以上研究证明，微生物菌群和宿主表型存在一定的关系，在这些疾病研究中算法的普遍预测精度可达到70%以上，甚至更高。利用微生物数据使用机器学习来预测宿主的健康状况一般为二分类问题，其中AUC值和F1分数（F1 score）可作为二分类模型的评价指标。F1分数为查准率和召回率的调和平均值，其中查准率（precision）表示预测正样本中的准确比例，召回率（recall）表示预测正确的正样本占所有正样本的比例。接收者操作特征（receiver operating characteristic，ROC）曲线也称为接受者工作特性曲线，其x轴为假阳性率（在所有真实值为负的样本中，预测错误所占的比例），y轴为真阳性率（即召回率）。AUC值是ROC曲线围成的一个面积值，理想的情况下AUC为1，即所有的样本都被正确分类；若AUC=0.5，则证明模型的性能和随机猜测相符；若AUC<0.5，则证明模型的性能不如随机猜测，几乎没有应用价值。一般选取AUC值在0.5~1之间具有研究价值。

在构建预测模型时，针对数据特点、应用场景及评价标准需要选择特定的机器学习算法，不同的算法有不同的特性与优势[71]，一般通过对比实验选取较优的算法（表1）。通过本文介绍的5种机器学习的特点以及在不同数据集上的性能表现，得出构建预测模型时选取机器学习算法的一般规律。①根据数据的特点来选择算法。数据特点包括数据形式（如数值型、文字型或布尔型）、数据量大小、数据冗余程度、缺失数据比例、数据均衡性等。在选取建模算法前将数据转变为数值型才能保证算法的运行；数据量较大可以选择适合大样

表1 机器学习不同疾病预测所使用算法及预测精度示例

Table 1 Examples of algorithms and prediction accuracy of different diseases predicted by machine learning

疾病类型	样本数	负样本数	正样本数	算法类型	评价标准	预测精度
2型糖尿病	344	170	174	随机森林	AUC	0.74
				支持向量机	AUC	0.66
				弹性网	AUC	0.70
				套索	AUC	0.71
	806	423	383	逻辑回归	F1分数	0.91
				支持向量机	F1分数	0.91
				自适应提升	F1分数	0.90
				梯度提升决策树	F1分数	0.87
				K近邻	F1分数	0.86
				随机梯度下降	F1分数	0.84
				随机森林	F1分数	0.83
肝硬化	232	118	114	随机森林	AUC	0.95
				支持向量机	AUC	0.92
				弹性网	AUC	0.91
				套索	AUC	0.88
结直肠癌	121	48	73	随机森林	AUC	0.87
				支持向量机	AUC	0.81
				弹性网	AUC	0.79
				套索	AUC	0.73
肥胖	253	164	89	随机森林	AUC	0.66
				支持向量机	AUC	0.65
				弹性网	AUC	0.64
				套索	AUC	0.60
炎症性肠病	110	25	85	随机森林	AUC	0.89
				支持向量机	AUC	0.86
				弹性网	AUC	0.83
				套索	AUC	0.81
胆管炎	48	24	24	随机森林	AUC	0.74
口臭	90	45	45	深度学习	AUC	0.97
				支持向量机	AUC	0.79
肠息肉	552	316	236	朴素贝叶斯	AUC	0.86
				人工神经网络	AUC	0.87

注：AUC—曲线下面积。

本学习的人工神经网络[72]，数据量较小则可以选择适合小样本学习的线性回归、支持向量机、K近邻、随机森林；数据冗余较大、不均衡、缺失比例高时可优先选择随机森林。②根据需求选择算法。需求包括运行的时空复杂度，模型的可解释性，分类或回归问题等，如依据预测的目标类型是数值变量或者类别变量选择是回归算法还是分类算法；要求较好的模型可解释性时可以选择线性回归和支持向量机；针对多分类问题可以选择随机森林、人工神经网络；对于时空复杂度要求较高的K近邻、人工神经网络算法则需要充分考虑计算机的硬件配置能否支撑起模型的运行。此外，在选取建模方法时应具体问题具体分析，综合考虑算法在时空复杂度、可解释性、普适性等方面的情况，结合前人的研究成果选取适合的算法，使得算法在预测模型中能够充分发挥自身优势。

4 展望

肠道微生物并不是仅依靠几种细菌就能够对宿主产生影响，而是大规模的微生物菌群协同作用的结果。当今机器学习应用于肠道菌群分析已较普遍，极大地推动了新型诊疗手段的发展。机器学习的应用有助于科研人员了解特定肠道菌群与宿主之间的关系，并挖掘它们深层次的特征，同时通过对筛选出来的特定靶点菌群进行机器学习预测及人工干预，用于临床辅助诊断和治疗。虽然科学技术的发展为人类提供了大量宿主与微生物菌群之间关系的信息[73]，促进了微生物学的发展，但仍存在机器学习预测精度不高、模型泛化能力不足、可解释性不强、模型容易过拟合、调动参数复杂等问题。因此，机器学习还需要在算法优化、特征提取、增加可解释性等方面进行改进，如利用仿生网络来进行算法优化及参数调整，以及使用融合方法代替单一方法来进行特征选择等。随着深度学习的兴起，对于大型的肠道菌群数据（$>10^4$），深度学习算法将会取得比传统机器学习更精确的预测结果[74]。本文为利用机器学习对肠道菌群宿主表型预测提供了一定的参考依据，而随着人工智能技术的飞速进步，机器学习正在逐渐渗透到生物信息学、生物医学和生物分类等领域，为这些领域带来了深刻的变革和创新。这种趋势对于加速科学研究、医学诊断和生物多样性研究都具有重要意义。

参 考 文 献

[1] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome[J]. Nature, 2012, 486(7402): 207-214.

[2] FALONY G, JOOSSENS M, VIEIRA-SILVA S, *et al*.. Population-level analysis of gut microbiome variation[J]. Science, 2016, 352(6285): 560-564.

[3] HE Y, WU W, ZHENG H M, *et al*.. Regional variation limits applications of healthy gut microbiome reference ranges and disease models[J]. Nat. Med., 2018, 24(10): 1532-1535.

[4] NAJAFABADI M M, VILLANUSTRE F, KHOSHGOFTAAR T M, *et al*.. Deep learning applications and challenges in big data analytics[J]. J. Big Data, 2015, 2(1): 1-21.

[5] HERNÁNDEZ MEDINA R, KUTUZOVA S, NIELSEN K N, *et al*.. Machine learning and deep learning applications in microbiome research[J]. ISME Commun., 2022, 2(1): 1-7.

[6] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transac. Inform. Theory, 1967, 13(1): 21-27.

[7] CORTES C, VAPNIK V. Support-vector networks[J]. Mach. Learn., 1995, 20(3): 273-297.

[8] HACLLAR H, NALBANTOĞLU O U, BAKIR-GÜNGÖR B. Machine learning analysis of inflammatory bowel disease-associated metagenomics dataset[C]//2018 3rd International Conference on Computer Science and Engineering (UBMK). IEEE, 2018: 434-438.

[9] ASSEGIE T A. Support vector machine and k-nearest neighbor based liver disease classification model[J]. Indonesian J. Electr. Engin. Med. Inform., 2021, 3(1): 9-14.

[10] LIU W, FANG X, ZHOU Y, *et al*.. Machine learning-based investigation of the relationship between gut microbiome and obesity status[J/OL]. Microbes Infect., 2022, 24(2): 104892[2022-05-04]. https://doi.org/10.1016/j.micinf.2021.104892.

[11] REIMAN D, METWALLY A, DAI Y. Using convolutional neural networks to explore the microbiome[J]. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2017, 2017: 4269-4272.

[12] NASSER I M, ABU-NASER S S. Lung cancer detection using artificial neural network[J]. Int. J. Engin. Inform. Systems, 2019, 3(3): 17-23.

[13] LYNGDOH A C, CHOUDHURY N A, MOULIK S. Diabetes disease prediction using machine learning algorithms[C]//2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, 2021: 517-521.

[14] GILL S R, POP M, DEBOY R T, *et al*.. Metagenomic analysis of the human distal gut microbiome[J]. Science, 2006, 312(5778): 1355-1359.

[15] XU J, GORDON J I. Honor thy symbionts[J]. Proc. Natl. Acad. Sci. USA, 2003, 100(18): 10452-10459.

[16] FRANK D N, AMAND A L, FELDMAN R A, *et al*.. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases[J]. FEMS Microbiol. Ecol., 2007, 104(34): 13780-13785.

[17] ZHANG X, ZHAO S, SONG X, *et al*.. Inhibition effect of glycyrrhiza polysaccharide (GCP) on tumor growth through regulation of the gut microbiota composition[J]. J. Pharmacol. Sci., 2018, 137(4): 324-332.

[18] O′HARA A M, SHANAHAN F. The gut flora as a forgotten organ[J]. EMBO Rep., 2006, 7(7): 688-693.

[19] SCHLOSS P D, HANDELSMAN J. Status of the microbial census[J]. Microbiol. Mol. Biol. Rev., 2004, 68(4): 686-691.

[20] MENG C, BAI C, BROWN T D, *et al*.. Human gut microbiota and gastrointestinal cancer[J]. Genom. Proteom. Bioinform., 2018, 16(1): 33-49.

[21] CARDING S, VERBEKE K, VIPOND D T, *et al*.. Dysbiosis of the gut microbiota in disease[J/OL]. Microb. Ecol. Health Dis., 2015, 26: 26191[2022-05-04]. https://doi.org/10.3402/mehd.v26.26191.

[22] HENNESSY A A, ROSS R P, FITZGERALD G F, *et al*.. Role of the gut in modulating lipoprotein metabolism[J/OL]. Curr. Cardiol. Rep., 2014, 16(8): 515[2022-05-04]. https://doi.org/10.1007/s11886-014-0515-2.

[23] CHELAKKOT C, GHIM J, RYU S H. Mechanisms regulating intestinal barrier integrity and its pathological implications[J]. Exp. Mol. Med., 2018, 50(8): 1-9.

[24] WALSH C J, GUINANE C M, O'TOOLE P W, *et al*.. Beneficial modulation of the gut microbiota[J]. FEBS Lett., 2014, 588(22): 4120-4130.

[25] WANG J, TANG H, ZHANG C, *et al*.. Modulation of gut microbiota during probiotic-mediated attenuation of metabolic syndrome in high fat diet-fed mice[J]. ISME J., 2015, 9(1): 1-15.

[26] RODRÍGUEZ J M, MURPHY K, STANTON C, *et al*.. The composition of the gut microbiota throughout life, with an emphasis on early life[J/OL]. Microb. Ecol. Health Dis., 2015, 26: 26050[2022-05-04]. https://doi.org/10.3402/mehd.v26.26050.

[27] LEPAGE P, COLOMBET J, MARTEAU P, *et al*.. Dysbiosis in inflammatory bowel disease: a role for bacteriophages?[J]. Gut, 2008, 57(3): 424-425.

[28] MÄTTÖ J, MAUNUKSELA L, KAJANDER K, *et al*.. Composition and temporal stability of gastrointestinal microbiota in irritable bowel syndrome: a longitudinal study in IBS and control subjects[J]. FEMS Immunol. Med. Microbiol., 2005, 43(2): 213-222.

[29] KEKU T O, DULAL S, DEVEAUX A, *et al*.. The gastrointestinal microbiota and colorectal cancer[J]. Am. J. Physiol. Gastrointest. Liver Physiol., 2015, 308(5): 351-363.

[30] ECK A, DE GROOT E F J, DE MEIJ T G J, *et al*.. Robust microbiota-based diagnostics for inflammatory bowel disease[J]. J. Clin. Microbiol., 2017, 55(6): 1720-1732.

[31] KANG D W, PARK J G, ILHAN Z E, *et al*.. Reduced incidence of *Prevotella* and other fermenters in intestinal microflora of autistic children[J/OL]. PLoS ONE, 2013, 8(7): e68322[2022-05-04]. https://doi.org/10.1371/journal.pone.0068322.

[32] SON J S, ZHENG L J, ROWEHL L M, *et al*.. Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the Simons simplex collection[J/OL]. PLoS ONE, 2015, 10(10): e0137725[2022-05-04]. https://doi.org/10.1371/journal.pone.0137725.

[33] ANGELAKIS E, ARMOUGOM F, MILLION M, *et al*.. The relationship between gut microbiota and weight gain in humans[J]. Future Microbiol., 2012, 7(1): 91-109.

[34] QIN J, LI Y, CAI Z, *et al*.. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.

[35] 张国庆,黄子琪,王明月,等. 大学生饮食习惯与唾液微生物多样性的关联[J]. 食品科学, 2019,40(1): 196-201.

[36] LI L, MINGLE D R. Mini review adv biotech & micro machine learning techniques on microbiome-based diagnostics[J]. Adv. Biotechnol. Microbiol., 2017, 6(4): 555695[2022-05-04]. https://doi.org/10.19080/AIBM.2017.06.555695.

[37] CAMMAROTA G, IANIRO G, AHERN A, *et al*.. Gut microbiome, big data and machine learning to promote precision medicine for cancer[J]. Nat. Rev. Gastroenterol. Hepatol., 2020, 17(10): 635-648.

[38] FRADKOV A. Early history of machine learning[J]. IFAC-Papers, 2020, 53(2): 1385-1390.

[39] ZHOU Y H, GALLINS P. A review and tutorial of machine learning methods for microbiome host trait prediction[J/OL]. Front. Genet., 2019, 10: 579[[2022-05-04]. https://doi.org/10.3389/fgene.2019.00579.

[40] ZHANG Y, YAN J, CHEN S, *et al*.. Review of the applications of deep learning in bioinformatics[J]. Curr. Bioinform., 2020, 15(8):1-14.

[41] DAVENPORT T, KALAKOTA R. The potential for artificial intelligence in healthcare[J]. Future Healthc. J., 2019, 6(2): 94-98.

[42] VUJKOVIC-CVIJIN I, SKLAR J, JIANG L, *et al*.. Host variables confound gut microbiota studies of human disease[J]. Nature, 2020, 587(7834): 448-454.

[43] CAMACHO D M, COLLINS K M, POWERS R K, *et al*.. Next-generation machine learning for biological networks[J]. Cell, 2018, 173(7): 1581-1592.

[44] MAHESH B. Machine learning algorithms-a review[J]. Int. J. Sci. Res., 2020, 9: 381-386.

[45] XU L, LIANG G, LIAO C, *et al*.. An efficient classifier for Alzheimer's disease genes identification[J/OL]. Molecules, 2018, 23(12): 3140[2022-05-04]. https://doi.org/10.3390/molecules23123140.

[46] KUNG H C, CHEN R M, TSAI J J P, *et al*.. Stratification of human gut microiome and building a SVM-based classifier[C]// 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2018: 14-17.

[47] ALTY S, MILLASSEAU S, CHOWIENCZYC P J, *et al*. Cardiovascular disease prediction using support vector machines[J]. Midwest Symp. Circuits Syst., 2004, 1: 376-379.

[48] WU H, CAI L, LI D, *et al*. Metagenomics biomarkers selected for prediction of three different diseases in Chinese population[J]. BioMed. Res. Int., 2018, 2018: 1-7.

[49] YAO Q, TANG M, ZENG L, *et al*.. Potential of fecal microbiota for detection and postoperative surveillance of colorectal cancer[J/OL]. BMC Microbiol., 2021, 21(1): 156[2022-05-04]. https://doi.org/10.1186/s12866-021-02182-6.

[50] LI H, PI D, WU Y, *et al*.. Integrative method based on linear regression for the prediction of zinc-binding sites in proteins[J/OL]. IEEE Access, 2017, PP(99): 1[2022-05-04]. https://doi.org/10.1109/ACCESS.2017.2731872.

[51] STATNIKOV A, HENAFF M, NARENDRA V, *et al*.. A comprehensive evaluation of multicategory classification methods for microbiomic data[J/OL]. Microbiome, 2013, 1(1): 11[2022-05-04]. https://doi.org/10.1186/2049-2618-1-11.

[52] PASOLLI E, TRUONG D T, MALIK F, *et al*.. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights[J/OL]. PLoS Comput. Biol., 2016, 12(7): e1004977[2022-05-04]. https://doi.org/10.1371/journal.pcbi.1004977.

[53] YANG L, WU H, JIN X, *et al*.. Study of cardiovascular disease prediction model based on random forest in Eastern China[J/OL]. Sci. Rep., 2020, 10(1): 5245[2022-05-04]. https://doi.org/10.1038/s41598-020-62133-5.

[54] TEJAMMA M, NAVEENKUMAR J P, PATIL S. A model based on convolutional neural network (CNN) to predict heart disease[J]. J. Algeb. Statist., 2022, 13(3): 2360-2367.

[55] WEHKAMP J, HARDER J, WEHKAMP K, *et al*.. NF-kappaB- and AP-1-mediated induction of human beta defensin-2 in intestinal epithelial cells by *Escherichia coli* Nissle 1917: a

novel effect of a probiotic bacterium[J]. Infect. Immun., 2004, 72(10): 5750-5758.

[56] SCHAEDLER R W, DUBOS R, COSTELLO R. The development of the bacterial flora in the gastrointestinal tract of mice[J]. J. Exp. Med., 1965, 122(1): 59-66.

[57] MAZMANIAN S K, LIU C H, TZIANABOS A O, *et al*.. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system[J]. Cell, 2005, 122(1): 107-118.

[58] 刘驰, 李家宝, 芮俊鹏, 等. 16S rRNA基因在微生物生态学中的应用[J]. 生态学报, 2015, 35(9): 2769-2788.

[59] CONSORTIUM H M P, HUTTENHOWER C, GEVERS D, *et al*.. Structure, function and diversity of the healthy human microbiome[J]. Nature, 2012, 486(7402): 207-214.

[60] MCDONALD D, HYDE E, DEBELIUS J W, *et al*.. American gut: an open platform for citizen science microbiome research[J]. Microorganisms, 2018, 3(3): e00031-e00018.

[61] PFLUGHOEFT K J, VERSALOVIC J. Human microbiome in health and disease[J]. Annu. Rev. Pathol., 2012, 7: 99-122.

[62] VALDES A M, WALTER J, SEGAL E, *et al*.. Role of the gut microbiota in nutrition and health[J/OL]. Brithish Med. J., 2018, 361: k2179[2022-05-04]. https://doi.org/10.1136/bmj.k2179.

[63] WONG A C, LEVY M. New approaches to microbiome-based therapies[J]. mSystems, 2019, 4(3): 119-122.

[64] SCHLABERG R. Microbiome diagnostics[J]. Clin. Chem., 2020, 66(1): 68-76.

[65] SCHLOSS P D. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research[J]. mBio, 2018, 9(3): 518-525.

[66] MCLAREN M R, WILLIS A D, CALLAHAN B J. Consistent and correctable bias in metagenomic sequencing experiments[J/OL]. eLife, 2019, 8: e46923[2022-05-04]. https://doi.org/10.7554/eLife.46923.

[67] QIN N, YANG F, LI A, *et al*.. Alterations of the human gut microbiome in liver cirrhosis[J]. Nature, 2014, 513(7516): 59-64.

[68] IWASAWA K, SUDA W, TSUNODA T, *et al*.. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker[J/OL]. Sci. Rep., 2018, 8(1): 5480[2022-05-04]. https://doi.org/10.1038/s41598-018-23870-w.

[69] SARWAR A, JAVED K, KHAN M J, *et al*.. Enhanced accuracy for motor imagery detection using deep learning for BCI[J]. Comp. Mater. Contin., 2021(9): 3825-3840.

[70] DADKHAH E, SIKAROODI M, KORMAN L, *et al*.. Gut microbiome identifies risk for colorectal polyps[J/OL]. BMJ Open Gastroenterol., 2019, 6(1): e000297[2022-05-04]. https://doi.org/10.1136/bmjgast-2019-000297.

[71] OSISANWO F Y, AKINSOLA J E T, AWODELE O, *et al*.. Supervised machine learning algorithms: classification and comparison[J]. Int. J. Comp. Trends Technol., 2017, 48(3): 128-138.

[72] LIVINGSTONE D J, MANALLACK D T, TETKO I V. Data modelling with neural networks: advantages and limitations[J]. J. Comput. Aided Mol. Des., 1997, 11(2): 135-142.

[73] LAMICHHANE S, SEN P, DICKENS A M, *et al*.. Gut metabolome meets microbiome: a methodological perspective to understand the relationship between host and microbe[J]. Methods, 2018, 149: 3-12.

[74] KUANG X, WANG F, HERNANDEZ K M, *et al*.. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN[J/OL]. Sci. Rep., 2022, 12(1): 2427[2022-05-04]. https://doi.org/10.1038/s41598-022-06449-4.