(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109448842 B
(45)授权公告日 2019.09.24

(21)申请号 201811357592.7
(22)申请日 2018.11.15
(65)同一申请的已公布的文献号
申请公布号 CN 109448842 A
(43)申请公布日 2019.03.08
(73)专利权人 苏州普瑞森基因科技有限公司
地址 215000 江苏省苏州市工业园区星湖街218号生物纳米园A5楼505单元
(72)发明人 朱永亮 陆敏 穆延召 陈倩 张水龙 史文阳
(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 史晶晶
(51)Int.Cl.
G16H 50/20(2018.01)

审查员 丁小汀

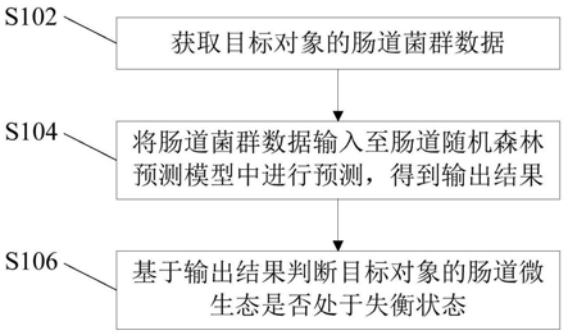
权利要求书6页 说明书17页 附图6页

(54)发明名称

人体肠道微生态失衡的确定方法、装置及电子设备

(57)摘要

本发明提供一种人体肠道微生态失衡的确定方法、装置及电子设备，涉及微生物生态学技术领域，人体肠道微生态失衡的确定方法包括：获取目标人体的肠道菌群数据，其中，肠道菌群数据为包含多个物种标记物的一个样本数据；将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果；肠道随机森林预测模型对应有用于判断样品表型的阈值，肠道随机森林预测模型包括：随机森林OTU模型或Shannon指数随机森林模型；基于输出结果判断目标人体的肠道微生态是否处于失衡状态。本发明能够通过预测模型判断目标人体是否处于肠道微生态失衡状态。



CN 109448842 B

1.一种人体肠道微生态失衡的确定方法,其特征在于,包括:

获取目标人体的肠道菌群数据,其中,所述肠道菌群数据为包含多个物种标记物的样本数据;

将所述肠道菌群数据输入至肠道随机森林预测模型中进行预测,得到输出结果;所述肠道随机森林预测模型对应有用于判断样品表型的阈值,所述肠道随机森林预测模型包括:随机森林OTU模型或Shannon指数随机森林模型;

基于所述输出结果判断所述目标人体的肠道微生态是否处于失衡状态;

所述肠道随机森林预测模型的数量为多个;

将所述肠道菌群数据输入至肠道随机森林预测模型中进行预测,得到输出结果,包括:将所述肠道菌群数据输入至每个所述肠道随机森林预测模型中进行预测,得到多个输出结果;

基于所述多个输出结果判断所述目标人体的肠道微生态是否处于失衡状态,包括:基于每个所述输出结果确定每个所述肠道随机森林预测模型的预测结果,得到多个预测结果;并基于所述多个预测结果中目标预测结果的比例确定所述目标人体的肠道微生态是否处于失衡状态,其中,所述目标预测结果为所述多个预测结果中用于表征所述目标人体的肠道微生态处于所述失衡状态的预测结果;

基于所述多个预测结果中目标预测结果的比例确定所述目标人体的肠道微生态是否处于失衡状态,包括:

若所述多个预测结果中,所述目标预测结果所占的比例大于第一预设比例阈值,则确定出所述目标人体的肠道微生态处于失衡状态;

若所述多个预测结果中,所述目标预测结果所占的比例大于第二预设比例阈值,且小于所述第一预设比例阈值,则基于所述多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定所述目标人体的肠道微生态是否处于失衡状态,所述目标肠道随机森林预测模型为所述多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型,所述第二预设比例阈值小于所述第一预设比例阈值。

2.根据权利要求1所述的方法,其特征在于,基于每个所述输出结果确定每个所述肠道随机森林预测模型的预测结果,包括:

将所述肠道菌群数据分别输入肠道随机森林预测模型Ai中,得到输出结果Bi,其中,i依次取1至I,I为所述多个肠道随机森林预测模型的数量;

若所述输出结果Bi大于或者等于所述肠道随机森林预测模型Ai所对应的阈值,则得到用于表征所述目标人体的肠道微生态处于失衡状态的预测结果;

若所述输出结果Bi小于所述肠道随机森林预测模型Ai所对应的阈值,则得到用于表征所述目标人体的肠道微生态处于正常状态的预测结果。

3.根据权利要求1所述的方法,其特征在于,通过如下步骤构建所述随机森林OTU模型,具体包括:

获取第一样本集;所述第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵,所述第一人体群体中各个人体的肠道微生态处于正常状态,所述第二人体群体中各个人体的肠道微生态处于所述失衡状态,所述目标矩阵为M*N的矩阵,所述第一样本集中包含M个样品,N表示每个样品对应N个物种标记物;所述目标矩阵

由M个样品中每个样品对应的N个物种标记物的用于表示物种标记物数量的丰度值组成；

基于所述第一样本集，建立肠道随机森林模型；

基于所述肠道随机森林模型，从所述第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集；

基于所述第二样本集构建所述随机森林OTU模型。

4.根据权利要求3所述的方法，其特征在于，基于所述第二样本集构建所述随机森林OTU模型，包括：

基于所述第二样本集，采用留一交叉验证法和目标建模方法建立所述随机森林OTU模型，其中，所述目标建模方法为用于构建所述肠道随机森林模型的方法。

5.根据权利要求4所述的方法，其特征在于，所述目标建模方法，包括：

按照有放回抽样方式，对所述第一样本集进行M次抽样，得到M个训练集，并为每个所述训练集建立决策树，得到M棵决策树；

对所述M棵决策树中每棵决策树进行分裂处理，得到M棵最大限度生长的决策树，以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征；

将所述M棵最大限度生长的决策树进行组合，得到所述肠道随机森林模型。

6.根据权利要求5所述的方法，其特征在于，对所述M棵决策树中每棵决策树进行分裂处理，包括：

在所述第一样本集中的N个物种标记物中，为每棵决策树中每个分裂节点选择对应的分裂特征，从而实现对每棵决策树进行分裂处理。

7.根据权利要求3所述的方法，其特征在于，基于所述肠道随机森林模型，从所述第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集，包括：

通过所述肠道随机森林模型中的重要性importance函数，输出所述N个物种标记物的重要性的排序结果；

将所述排序结果中前预设个数最大重要性所对应的物种标记物作为所述重要物种标记物；

基于所述重要物种标记物的序列分类单位OTU表，组成所述第二样本集。

8.根据权利要求4所述的方法，其特征在于，基于所述第二样本集，采用留一交叉验证法和目标建模方法建立所述随机森林OTU模型，包括：

基于所述第二样本集中的每个样本构建M个测试集和M个训练集，其中，M为所述第二样本集中样本的数量，当所述M个测试集中的第m个测试集为第m个样本时，所述M个训练集中第m个训练集为所述第二样本集中除所述第m个样本之外的其他样本；

基于所述M个训练集及所述目标建模方法，建立M个第一子模型，作为所述随机森林OTU模型。

9.根据权利要求8所述的方法，其特征在于，所述方法还包括：

利用所述M个测试集中与所述M个第一子模型相对应的测试集，对每个所述第一子模型进行预测，得到M个第一输出结果；

根据所述M个第一输出结果绘制第一ROC特征曲线；

基于所述第一ROC特征曲线，计算所述随机森林OTU模型的阈值，其中，所述随机森林OTU模型的阈值用于判断样品表型。

10.根据权利要求3所述的方法，其特征在于，在基于所述肠道随机森林模型，从所述第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集之后，还包括：

根据所述重要物种标记物的丰度值，计算所述第二样本集中每个样品的Shannon指数；

将所述Shannon指数添加至所述重要物种标记物的OTU表中，得到第三样本集；

基于所述第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，其中，所述目标建模方法为用于构建所述肠道随机森林模型的方法。

11.根据权利要求10所述的方法，其特征在于，根据所述重要物种标记物的丰度值，计算所述第二样本集中每个样品的Shannon指数，包括：

利用下式，计算所述第二样本集中每个样品的Shannon指数：

$S=-(P_1 \times \ln P_1+P_2 \times \ln P_2+...+P_n \times \ln P_n)$；

其中，P_1、P_2...P_n为样品第1个、第2个...第n个物种标记物的丰度值，S为所述样品的Shannon指数。

12.根据权利要求10所述的方法，其特征在于，基于所述第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，包括：

基于所述第三样本集中的每个样本构建M个测试集和M个训练集，其中，M为所述第三样本集中样本的数量，当所述M个测试集中的第m个测试集为第m个样本时，所述M个训练集中第m个训练集为所述第三样本集中除所述第m个样本之外的其他样本；

基于所述M个训练集和所述目标建模方法建立M个第二子模型，作为所述Shannon指数随机森林模型。

13.根据权利要求12所述的方法，其特征在于，所述方法还包括：

利用所述M个测试集中与M个第二子模型相对应的测试集，对每个所述第二子模型进行预测，得到M个第二预测结果；

根据所述M个第二预测结果绘制第二ROC特征曲线；

基于所述第二ROC特征曲线，计算所述Shannon指数随机森林模型的阈值，其中，所述Shannon指数随机森林模型的阈值用于判断样品表型。

14.一种人体肠道微生态失衡的确定装置，其特征在于，包括：

数据获取模块，用于获取目标人体的肠道菌群数据，其中，所述肠道菌群数据为包含多个物种标记物的一个样本数据；

模型预测模块，用于将所述肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果；所述肠道随机森林预测模型对应有用于判断样品表型的阈值，所述肠道随机森林预测模型包括：随机森林OTU模型或Shannon指数随机森林模型；

结果判断模块，用于基于所述输出结果判断所述目标人体的肠道微生态是否处于失衡状态；

所述肠道随机森林预测模型的数量为多个；

所述模型预测模块，还用于：将所述肠道菌群数据输入至每个所述肠道随机森林预测模型中进行预测，得到多个输出结果；

所述结果判断模块，还用于：基于所述多个输出结果确定每个所述肠道随机森林预测模型的预测结果，得到多个预测结果；并基于所述多个预测结果中目标预测结果的比例确

定所述目标人体的肠道微生态是否处于失衡状态，其中，所述目标预测结果为所述多个预测结果中用于表征所述目标人体的肠道微生态处于所述失衡状态的预测结果；

所述结果判断模块，还用于：

若所述多个预测结果中，所述目标预测结果所占的比例大于第一预设比例阈值，则确定出所述目标人体的肠道微生态处于失衡状态；

若所述多个预测结果中，所述目标预测结果所占的比例大于第二预设比例阈值，且小于所述第一预设比例阈值，则基于所述多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定所述目标人体的肠道微生态是否处于失衡状态，所述目标肠道随机森林预测模型为所述多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型，所述第二预设比例阈值小于所述第一预设比例阈值，每个所述肠道随机森林预测模型对应一个阈值。

15.根据权利要求14所述的装置，其特征在于，所述结果判断模块，还用于：

将所述肠道菌群数据分别输入肠道随机森林预测模型Ai中，得到输出结果Bi，其中，i依次取1至I，I为所述多个肠道随机森林预测模型的数量；

若所述输出结果Bi大于或者等于所述肠道随机森林预测模型Ai所对应的阈值，则得到用于表征所述目标人体的肠道微生态处于失衡状态的预测结果；

若所述输出结果Bi小于所述肠道随机森林预测模型Ai所对应的阈值，则得到用于表征所述目标人体的肠道微生态处于正常状态的预测结果。

16.根据权利要求14所述的装置，其特征在于，还包括：模型构建模块；所述模型构建模块包括：

第一样本集获取模块，用于获取第一样本集；所述第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵，所述第一人体群体中各个人体的肠道微生态处于正常状态，所述第二人体群体中各个人体的肠道微生态处于所述失衡状态，所述目标矩阵为M*N的矩阵，所述第一样本集中包含M个样品，N表示每个样品对应N个物种标记物；所述目标矩阵由M个样品中每个样品对应的N个物种标记物的用于表示物种标记物数量的丰度值组成；

基础模型建立模块，用于基于所述第一样本集，建立肠道随机森林模型；

第二样本集获取模块，用于基于所述肠道随机森林模型，从所述第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集；

预测模型构建模块，用于基于所述第二样本集构建所述随机森林OTU模型。

17.根据权利要求16所述的装置，其特征在于，所述预测模型构建模块，还用于：

基于所述第二样本集，采用留一交叉验证法和目标建模方法建立所述随机森林OTU模型，其中，所述目标建模方法为用于构建所述肠道随机森林模型的方法。

18.根据权利要求17所述的装置，其特征在于，所述基础模型建立模块，还用于：

按照有放回抽样方式，对所述第一样本集进行M次抽样，得到M个训练集，并为每个所述训练集建立决策树，得到M棵决策树；

对所述M棵决策树中每棵决策树进行分裂处理，得到M棵最大限度生长的决策树，以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征；

将所述M棵最大限度生长的决策树进行组合，得到所述肠道随机森林模型。

19.根据权利要求18所述的装置,其特征在于,所述基础模型建立模块,还用于:

在所述第一样本集中的N个物种标记物中,为每棵决策树中每个分裂节点选择对应的分裂特征,从而实现对每棵决策树进行分裂处理。

20.根据权利要求16所述的装置,其特征在于,所述第二样本集获取模块,还用于:

通过所述肠道随机森林模型中的重要性importance函数,输出所述N个物种标记物的重要性的排序结果;

将所述排序结果中前预设个数最大重要性所对应的物种标记物作为所述重要物种标记物;

基于所述重要物种标记物的序列分类单位OTU表,组成所述第二样本集。

21.根据权利要求17所述的装置,其特征在于,所述预测模型构建模块,还用于:

基于所述第二样本集中的每个样本构建M个测试集和M个训练集,其中,M为所述第二样本集中样本的数量,当所述M个测试集中的第m个测试集为第m个样本时,所述M个训练集中第m个训练集为所述第二样本集中除所述第m个样本之外的其他样本;

基于所述M个训练集及所述目标建模方法,建立M个第一子模型,作为所述随机森林OTU模型。

22.根据权利要求21所述的装置,其特征在于,还包括:第一阈值确定模块,用于:

利用所述M个测试集中与所述M个第一子模型相对应的测试集,对每个所述第一子模型进行预测,得到M个第一输出结果;

根据所述M个第一输出结果绘制第一ROC特征曲线;

基于所述第一ROC特征曲线,计算所述随机森林OTU模型的阈值,其中,所述随机森林OTU模型的阈值用于判断样品表型。

23.根据权利要求16所述的装置,其特征在于,所述模型构建模块,还包括:

第三样本集获取模块,用于根据所述重要物种标记物的丰度值,计算所述第二样本集中每个样品的Shannon指数;将所述Shannon指数添加至所述重要物种标记物的OTU表中,得到第三样本集;

所述预测模型构建模块,还用于基于所述第三样本集,采用留一交叉验证法和目标建模方法,建立Shannon指数随机森林模型,其中,所述目标建模方法为用于构建所述肠道随机森林模型的方法。

24.根据权利要求23所述的装置,其特征在于,所述第三样本集获取模块,还用于:

利用下式,计算所述第二样本集中每个样品的Shannon指数:

$S=-(P_1 \times \ln P_1+P_2 \times \ln P_2+\ldots+P_n \times \ln P_n)$;

其中,P_1、P_2...P_n为样品第1个、第2个...第n个物种标记物的丰度值,S为所述样品的Shannon指数。

25.根据权利要求23所述的装置,其特征在于,所述预测模型构建模块,还用于:

基于所述第三样本集中的每个样本构建M个测试集和M个训练集,其中,M为所述第三样本集中样本的数量,当所述M个测试集中的第m个测试集为第m个样本时,所述M个训练集中第m个训练集为所述第三样本集中除所述第m个样本之外的其他样本;

基于所述M个训练集和所述目标建模方法建立M个第二子模型,作为所述Shannon指数随机森林模型。

26.根据权利要求25所述的装置,其特征在于,还包括:第二阈值确定模块,用于:

利用所述M个测试集中与M个第二子模型相对应的测试集,对每个所述第二子模型进行预测,得到M个第二预测结果;

根据所述M个第二预测结果绘制第二ROC特征曲线;

基于所述第二ROC特征曲线,计算所述Shannon指数随机森林模型的阈值,其中,所述Shannon指数随机森林模型的阈值用于判断样品表型。

27.一种电子设备,包括存储器、处理器,所述存储器上存储有可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现上述权利要求1至13任一项所述的方法的步骤。

人体肠道微生态失衡的确定方法、装置及电子设备

技术领域

[0001] 本发明涉及微生物生态学技术领域，尤其是涉及一种人体肠道微生态失衡的确定方法、装置及电子设备。

背景技术

[0002] 由于不良饮食习惯、环境污染及各种农药、化肥、抗生素、激素的不合理使用最终影响到人类的健康。这些影响中有些可能是一个累积的过程，不会立即导致人体患病，但可能导致人体微生态失衡，成为促成人体亚健康的重要因素。微生态学是近三、四十年来发展起来的一门新兴生物学科，他是研究宿主(人、动物、植物)与其体内的正常微生物和内外环境相互关系的学科。从生态学的角度，研究人体健康与内外环境的关系，揭示各种内外环境对人体的影响规律。微生态失衡是人体亚健康一个重要标志。出现微生态失衡预示人体即将患病或已经患病。纠正微生态失衡或者说纠正人体的亚健康，不能靠医疗部门去解决如此众多的人体群体问题，必须靠宣传和普及公众健康和营养知识来实现。

[0003] 几乎所有引起亚健康的诱因都可能导致肠道微生态的失衡。肠道微生态的失衡既是亚健康的结果，同时也可能加重亚健康，导致疾病的发生。肠道微生态是机体最重要、最庞大，尤为特殊的生态系统。肠道内大量微生物菌时刻处在动态平衡和相对稳定之中。众多因素影响这个平衡。人体亚健康的发生、发展和治疗转归均伴随着肠道微生态正常菌群的变化或失衡。但是，到目前为止，还没有很好的判断人体肠道微生态失衡的方法。

发明内容

[0004] 有鉴于此，本发明的目的在于提供一种人体肠道微生态失衡的确定方法。

[0005] 第一方面，本发明实施例提供了一种人体肠道微生态失衡的确定方法，包括：获取目标人体的肠道菌群数据，其中，肠道菌群数据为包含多个物种标记物的一个样本数据；将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果；肠道随机森林预测模型对应有用于判断样品表型的阈值，肠道随机森林预测模型包括：随机森林OTU模型或Shannon指数随机森林模型；基于输出结果判断目标人体的肠道微生态是否处于失衡状态。

[0006] 结合第一方面，本发明实施例提供了第一方面的第一种可能的实施方式，其中，肠道随机森林预测模型的数量为多个；将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果，包括：将肠道菌群数据输入至每个肠道随机森林预测模型中进行预测，得到多个输出结果；基于多个输出结果判断目标人体的肠道微生态是否处于失衡状态，包括：基于每个输出结果确定每个肠道随机森林预测模型的预测结果，得到多个预测结果；并基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态，其中，目标预测结果为多个预测结果中用于表征目标人体的肠道微生态处于失衡状态的预测结果。

[0007] 结合第一方面，本发明实施例提供了第一方面的第二种可能的实施方式，其中，基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态，包

括:若多个预测结果中,目标预测结果所占的比例大于第一预设比例阈值,则确定出目标人体的肠道微生态处于失衡状态;若多个预测结果中,目标预测结果所占的比例大于第二预设比例阈值,且小于第一预设比例阈值,则基于多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定目标人体的肠道微生态是否处于失衡状态,目标肠道随机森林预测模型为多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型,第二预设比例阈值小于第一预设比例阈值,每个肠道随机森林预测模型对应一个阈值。

[0008] 结合第一方面,本发明实施例提供了第一方面的第三种可能的实施方式,其中,基于每个输出结果确定每个肠道随机森林预测模型的预测结果,包括:将肠道菌群数据分别输入肠道随机森林预测模型Ai中,得到输出结果Bi,其中,i依次取1至I,I为多个肠道随机森林预测模型的数量;若输出结果Bi大于或者等于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于失衡状态的预测结果;若输出结果Bi小于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于正常状态的预测结果。

[0009] 结合第一方面,本发明实施例提供了第一方面的第四种可能的实施方式,其中,通过如下步骤构建随机森林OTU模型,具体包括:获取第一样本集;第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵,第一人体群体中各个人体的肠道微生态处于正常状态,第二人体群体中各个人体的肠道微生态处于失衡状态,目标矩阵为M*N的矩阵,第一样本集中包含M个样品,N表示每个样品对应N个物种标记物;目标矩阵由M个样品中每个样品对应的N个物种标记物的用于表示物种标记物数量的丰度值组成;基于第一样本集,建立肠道随机森林模型;基于肠道随机森林模型,从第一样本集的N个物种标记物中,筛选出预设个数的重要物种标记物,组成第二样本集;基于第二样本集构建随机森林OTU模型。

[0010] 结合第一方面,本发明实施例提供了第一方面的第五种可能的实施方式,其中,基于第二样本集构建随机森林OTU模型,包括:基于第二样本集,采用留一交叉验证法和目标建模方法建立随机森林OTU模型,其中,目标建模方法为用于构建肠道随机森林模型的方法。

[0011] 结合第一方面,本发明实施例提供了第一方面的第六种可能的实施方式,其中,目标建模方法,包括:按照有放回抽样方式,对第一样本集进行M次抽样,得到M个训练集,并为每个训练集建立决策树,得到M棵决策树;对M棵决策树中每棵决策树进行分裂处理,得到M棵最大限度生长的决策树,以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征;将M棵最大限度生长的决策树进行组合,得到肠道随机森林模型。

[0012] 结合第一方面,本发明实施例提供了第一方面的第七种可能的实施方式,其中,对M棵决策树中每棵决策树进行分裂处理,包括:在第一样本集中的N个物种标记物中,为每棵决策树中每个分裂节点选择对应的分裂特征,从而实现对每棵决策树进行分裂处理。

[0013] 结合第一方面,本发明实施例提供了第一方面的第八种可能的实施方式,其中,基于肠道随机森林模型,从第一样本集的N个物种标记物中,筛选出预设个数的重要物种标记物,组成第二样本集,包括:通过肠道随机森林模型中的重要性importance函数,输出N个物种标记物的重要性的排序结果;将排序结果中前预设个数最大重要性所对应的物种标记物作为重要物种标记物;基于重要物种标记物的序列分类单位OTU表,组成第二样本集。

[0014] 结合第一方面，本发明实施例提供了第一方面的第九种可能的实施方式，其中，基于第二样本集，采用留一交叉验证法和目标建模方法建立随机森林OTU模型，包括：基于第二样本集中的每个样本构建M个测试集和M个训练集，其中，M为第二样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第二样本集中除第m个样本之外的其他样本；基于M个训练集及目标建模方法，建立M个第一子模型，作为随机森林OTU模型。

[0015] 结合第一方面，本发明实施例提供了第一方面的第十种可能的实施方式，其中，方法还包括：利用M个测试集中与M个第一子模型相对应的测试集，对每个第一子模型进行预测，得到M个第一输出结果；根据M个第一输出结果绘制第一ROC特征曲线；基于第一ROC特征曲线，计算随机森林OTU模型的阈值，其中，随机森林OTU模型的阈值用于判断样品表型。

[0016] 结合第一方面，本发明实施例提供了第一方面的第十一种可能的实施方式，其中，在基于肠道随机森林模型，从第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集之后，还包括：根据重要物种标记物的丰度值，计算第二样本集中每个样品的Shannon指数；将Shannon指数添加至重要物种标记物的OTU表中，得到第三样本集；基于第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，其中，目标建模方法为用于构建肠道随机森林模型的方法。

[0017] 结合第一方面，本发明实施例提供了第一方面的第十二种可能的实施方式，其中，根据重要物种标记物的丰度值，计算第二样本集中每个样品的Shannon指数，包括：利用下式，计算第二样本集中每个样品的Shannon指数：S＝-(P1x lnP1+P2x lnP2+...+Pn x lnPn)；其中，P1、P2...Pn为样品第1个、第2个...第n个物种标记物的丰度值，S为样品的Shannon指数。

[0018] 结合第一方面，本发明实施例提供了第一方面的第十三种可能的实施方式，其中，基于第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，包括：基于第三样本集中的每个样本构建M个测试集和M个训练集，其中，M为第三样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第三样本集中除第m个样本之外的其他样本；基于M个训练集和目标建模方法建立M个第二子模型，作为Shannon指数随机森林模型。

[0019] 结合第一方面，本发明实施例提供了第一方面的第十四种可能的实施方式，其中，方法还包括：利用M个测试集中与M个第二子模型相对应的测试集，对每个第二子模型进行预测，得到M个第二预测结果；根据M个第二预测结果绘制第二ROC特征曲线；基于第二ROC特征曲线，计算Shannon指数随机森林模型的阈值，其中，Shannon指数随机森林模型的阈值用于判断样品表型。

[0020] 第二方面，本发明实施例提供一种人体肠道微生态失衡的确定装置，包括：数据获取模块，用于获取目标人体的肠道菌群数据，其中，肠道菌群数据为包含多个物种标记物的一个样本数据；模型预测模块，用于将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果；肠道随机森林预测模型对应有用于判断样品表型的阈值，肠道随机森林预测模型包括：随机森林OTU模型或Shannon指数随机森林模型；结果判断模块，用于基于输出结果判断目标人体的肠道微生态是否处于失衡状态。

[0021] 结合第二方面，本发明实施例提供了第二方面的第一种可能的实施方式，其中，肠

道随机森林预测模型的数量为多个;模型预测模块,还用于:将肠道菌群数据输入至每个肠道随机森林预测模型中进行预测,得到多个输出结果;结果判断模块,还用于:基于多个输出结果确定每个肠道随机森林预测模型的预测结果,得到多个预测结果;并基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态,其中,目标预测结果为多个预测结果中用于表征目标人体的肠道微生态处于失衡状态的预测结果。

[0022] 结合第二方面,本发明实施例提供了第二方面的第二种可能的实施方式,其中,结果判断模块,还用于:若多个预测结果中,目标预测结果所占的比例大于第一预设比例阈值,则确定出目标人体的肠道微生态处于失衡状态;若多个预测结果中,目标预测结果所占的比例大于第二预设比例阈值,且小于第一预设比例阈值,则基于多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定目标人体的肠道微生态是否处于失衡状态,目标肠道随机森林预测模型为多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型,第二预设比例阈值小于第一预设比例阈值。

[0023] 结合第二方面,本发明实施例提供了第二方面的第三种可能的实施方式,其中,结果判断模块,还用于:将肠道菌群数据分别输入肠道随机森林预测模型Ai中,得到输出结果Bi,其中,i依次取1至I,I为多个肠道随机森林预测模型的数量;若输出结果Bi大于或者等于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于失衡状态的预测结果;若输出结果Bi小于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于正常状态的预测结果。

[0024] 结合第二方面,本发明实施例提供了第二方面的第四种可能的实施方式,其中,还包括:模型构建模块;模型构建模块包括:第一样本集获取模块,用于获取第一样本集;第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵,第一人体群体中各个人体的肠道微生态处于正常状态,第二人体群体中各个人体的肠道微生态处于失衡状态,目标矩阵为M*N的矩阵,第一样本集中包含M个样品,N表示每个样品对应N个物种标记物;目标矩阵由M个样品中每个样品对应的N个物种标记物的用于表示物种标记物数量的丰度值组成;基础模型建立模块,用于基于第一样本集,建立肠道随机森林模型;第二样本集获取模块,用于基于肠道随机森林模型,从第一样本集的N个物种标记物中,筛选出预设个数的重要物种标记物,组成第二样本集;预测模型构建模块,用于基于第二样本集构建随机森林OTU模型。

[0025] 结合第二方面,本发明实施例提供了第二方面的第五种可能的实施方式,其中,预测模型构建模块,还用于:基于第二样本集,采用留一交叉验证法和目标建模方法建立随机森林OTU模型,其中,目标建模方法为用于构建肠道随机森林模型的方法。

[0026] 结合第二方面,本发明实施例提供了第二方面的第六种可能的实施方式,其中,基础模型建立模块,还用于:按照有放回抽样方式,对第一样本集进行M次抽样,得到M个训练集,并为每个训练集建立决策树,得到M棵决策树;对M棵决策树中每棵决策树进行分裂处理,得到M棵最大限度生长的决策树,以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征;将M棵最大限度生长的决策树进行组合,得到肠道随机森林模型。

[0027] 结合第二方面,本发明实施例提供了第二方面的第七种可能的实施方式,其中,基础模型建立模块,还用于:在第一样本集中的N个物种标记物中,为每棵决策树中每个分裂节点选择对应的分裂特征,从而实现对每棵决策树进行分裂处理。

[0028] 结合第二方面，本发明实施例提供了第二方面的第八种可能的实施方式，其中，第二样本集获取模块，还用于：通过肠道随机森林模型中的重要性importance函数，输出N个物种标记物的重要性的排序结果；将排序结果中前预设个数最大重要性所对应的物种标记物作为重要物种标记物；基于重要物种标记物的序列分类单位OTU表，组成第二样本集。

[0029] 结合第二方面，本发明实施例提供了第二方面的第九种可能的实施方式，其中，预测模型构建模块，还用于：基于第二样本集中的每个样本构建M个测试集和M个训练集，其中，M为第二样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第二样本集中除第m个样本之外的其他样本；基于M个训练集及目标建模方法，建立M个第一子模型，作为随机森林OTU模型。

[0030] 结合第二方面，本发明实施例提供了第二方面的第十种可能的实施方式，其中，还包括：第一阈值确定模块，用于：利用M个测试集中与M个第一子模型相对应的测试集，对每个第一子模型进行预测，得到M个第一输出结果；根据M个第一输出结果绘制第一ROC特征曲线；基于第一ROC特征曲线，计算随机森林OTU模型的阈值，其中，阈值用于判断样品表型。

[0031] 结合第二方面，本发明实施例提供了第二方面的第十一种可能的实施方式，其中，模型构建模块，还包括：第三样本集获取模块，用于根据重要物种标记物的丰度值，计算第二样本集中每个样品的Shannon指数；将Shannon指数添加至重要物种标记物的OTU表中，得到第三样本集；预测模型构建模块，还用于基于第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，其中，目标建模方法为用于构建肠道随机森林模型的方法。

[0032] 结合第二方面，本发明实施例提供了第二方面的第十二种可能的实施方式，其中，第三样本集获取模块，还用于：利用下式，计算第二样本集中每个样品的Shannon指数：S＝-(P1x lnP1+P2x lnP2+...+Pn x lnPn)；其中，P1、P2...Pn为样品第1个、第2个...第n个物种标记物的丰度值，S为样品的Shannon指数。

[0033] 结合第二方面，本发明实施例提供了第二方面的第十三种可能的实施方式，其中，预测模型构建模块，还用于：基于第三样本集中的每个样本构建M个测试集和M个训练集，其中，M为第三样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第三样本集中除第m个样本之外的其他样本；基于M个训练集和目标建模方法建立M个第二子模型，作为Shannon指数随机森林模型。

[0034] 结合第二方面，本发明实施例提供了第二方面的第十四种可能的实施方式，其中，还包括：第二阈值确定模块，用于：利用M个测试集中与M个第二子模型相对应的测试集，对每个第二子模型进行预测，得到M个第二预测结果；根据M个第二预测结果绘制第二ROC特征曲线；基于第二ROC特征曲线，计算Shannon指数随机森林模型的阈值，其中，阈值用于判断样品表型。

[0035] 第三方面，本发明实施例还提供一种电子设备，包括存储器、处理器，存储器上存储有可在处理器上运行的计算机程序，处理器执行计算机程序时实现上述第一方面中任一种可能的实施方式所述的方法的步骤。

[0036] 第四方面，本发明实施例还提供一种具有处理器可执行的非易失的程序代码的计算机可读介质，程序代码使处理器执行第一方面中任一种可能的实施方式所述的方法。

[0037] 本发明实施例带来了以下有益效果：

[0038] 本发明实施例提供的人体肠道微生态失衡的确定方法中，首先获取目标人体的肠道菌群数据，该肠道菌群数据为包含多个物种标记物的一个样本数据，将该肠道菌群数据输入至预先训练好的肠道随机森林预测模型中进行预测，得到该预测模型对应的输出结果；其中，肠道随机森林预测模型可以是多个也可以是一个，每个肠道随机森林预测模型对应有判断样品表型的阈值，且肠道随机森林预测模型可以包括：随机森林OTU模型或Shannon指数随机森林模型；最后基于上述输出结果判断目标人体的肠道微生态是否处于失衡状态。通过上述方案，可以快速相对准确地判断目标人体是否处于肠道微生态失衡状态。

[0039] 本发明的其他特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

[0040] 为使本发明的上述目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0041] 为了更清楚地说明本发明具体实施方式或现有技术中的技术方案，下面将对具体实施方式或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施方式，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1为本发明实施例提供的一种人体肠道微生态失衡的确定方法的流程图；

[0043] 图2为本发明实施例提供的另一种人体肠道微生态失衡的确定方法的流程图；

[0044] 图3为本发明实施例提供的另一种人体肠道微生态失衡的确定方法的流程图；

[0045] 图4为本发明实施例提供的一种人体肠道微生态失衡的确定方法中的第一样本集中的M*N矩阵示意图；

[0046] 图5为本发明实施例提供的一种人体肠道微生态失衡的确定方法中用于确定随机森林OTU模型阈值的ROC曲线图；

[0047] 图6为本发明实施例提供的另一种人体肠道微生态失衡的确定方法的流程图；

[0048] 图7为本发明实施例提供的一种人体肠道微生态失衡的确定方法中用于确定Shannon指数随机森林模型阈值的ROC曲线图；

[0049] 图8为本发明实施例提供的一种人体肠道微生态失衡的确定方法中测序片段到物种丰度分析步骤示意图；

[0050] 图9为本发明实施例提供的一种人体肠道微生态失衡的确定装置的示意图；

[0051] 图10为本发明实施例提供的另一种人体肠道微生态失衡的确定装置的示意图；

[0052] 图11为本发明实施例提供的一种电子设备的示意图。

具体实施方式

[0053] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合附图对本发明的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提

下所获得的所有其他实施例，都属于本发明保护的范围。

[0054] 目前，还没有很好的判断人体肠道微生态失衡的方法。基于此，本发明实施例提供一种人体肠道微生态失衡的确定方法，能够通过预测模型判断目标人体是否处于肠道微生态失衡状态。

[0055] 为便于对本实施例进行理解，首先对本发明实施例所公开的一种人体肠道微生态失衡的确定方法进行详细介绍。

[0056] 本发明实施例提供了一种人体肠道微生态失衡的确定方法，参见图1所示，该方法包括以下步骤：

[0057] 步骤S102，获取目标人体的肠道菌群数据。

[0058] 其中，肠道菌群数据为包含多个物种标记物的一个样本数据，比如物种标记物的个数为N，那么该肠道菌群数据就是一个1*N的矩阵，该矩阵由N个物种标记物的丰度值组成。

[0059] 步骤S104，将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果。

[0060] 具体实施中，上述肠道随机森林预测模型可以是一个，如：随机森林OTU模型或Shannon指数随机森林模型，也可以是多个，如：通过不同数量的物种标记物的样本集，所训练出来的多个随机森林OTU模型和/或多个Shannon指数随机森林模型。其中，每个肠道随机森林预测模型对应有用于判断样品表型的阈值，样品表型包括正常和失衡。

[0061] 对于一个预测模型的情况，将肠道菌群数据输入至肠道随机森林预测模型中进行预测后，得到一个输出结果。对于多个预测模型的情况，将肠道菌群数据输入至多个肠道随机森林预测模型中进行预测后，得到多个输出结果。上述输出结果均为模型的预测概率结果。

[0062] 步骤S106，基于输出结果判断目标人体的肠道微生态是否处于失衡状态。

[0063] 对于一个预测模型的情况，在得到一个输出结果后，通过该输出结果与模型对应阈值的比较，判断目标人体的肠道微生态是否处于失衡状态。

[0064] 具体的判断过程为：

[0065] 判断上述输出结果是否大于肠道随机森林预测模型所对应的阈值；如果是，则确定出所述目标人体的肠道微生态处于失衡状态；如果否，则确定出所述目标人体的肠道微生态处于正常状态。

[0066] 对于多个预测模型的情况，在得到多个输出结果后，首先基于多个输出结果确定每个肠道随机森林预测模型的预测结果，然后基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态，其中，目标预测结果为多个预测结果中用于表征目标人体的肠道微生态处于失衡状态的预测结果。

[0067] 通过上述两种情况的模型预测，均可以实现对目标人体的肠道微生态是否处于失衡状态进行判断。通过一种模型对目标人体的肠道微生态是否处于失衡状态进行判断，可以实现快速较准确的判断过程，而通过上述多种预先训练好的模型进行预测，并基于多个预测结果进行判断，能够提高判断的准确性，其准确性相对来说会更高。

[0068] 在一种可能的实施方式中，上述基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态具体包括以下两种情况：

[0069] (1)若多个预测结果中,目标预测结果所占的比例大于第一预设比例阈值,则确定出目标人体的肠道微生态处于失衡状态。

[0070] 这里,目标预测结果为多个预测结果中用于表征目标人体的肠道微生态处于失衡状态的预测结果,也就是预测结果为失衡。在实际应用中,上述第一预设比例阈值一般设为80%,即在多个预测结果中,有大于80%的预测结果为失衡的时候,就可以确定出上述目标人体的肠道微生态处于失衡状态。当然,上述第一预设比例阈值越大,对目标人体的肠道微生态的判断结果越准确。

[0071] (2)若多个预测结果中,目标预测结果所占的比例大于第二预设比例阈值,且小于第一预设比例阈值,则基于多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定目标人体的肠道微生态是否处于失衡状态,目标肠道随机森林预测模型为多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型,其中,第二预设比例阈值小于第一预设比例阈值。

[0072] 实际应用中,第二预设比例阈值一般设为20%,也就是说,当多个预测结果中,有大于20%小于80%的预测结果为失衡的时候,进一步根据多个肠道随机森林预测模型中,阈值最大的那个肠道随机森林预测模型的预测结果,确定目标人体的肠道微生态是否处于失衡状态。如果该预测结果是失衡,则确定目标人体的肠道微生态处于失衡状态,如果该预测结果是正常,则确定目标人体的肠道微生态处于正常状态。

[0073] 上述多个肠道随机森林预测模型分别对应有一个阈值,基于多个肠道随机森林预测模型的输出结果确定每个肠道随机森林预测模型的预测结果的过程包括以下步骤,参见图2所示:

[0074] 步骤S202,将肠道菌群数据分别输入肠道随机森林预测模型Ai中,得到输出结果Bi,其中,i依次取1至I,I为多个肠道随机森林预测模型的数量;

[0075] 步骤S204,若输出结果Bi大于或者等于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于失衡状态的预测结果;

[0076] 步骤S206,若输出结果Bi小于肠道随机森林预测模型Ai所对应的阈值,则得到用于表征目标人体的肠道微生态处于正常状态的预测结果。

[0077] 其中,肠道随机森林预测模型所对应的用于判断样品表型的阈值是通过复杂的过程从众多的阈值中所筛选出来的最佳阈值(具体筛选过程见下文所述),通过上述输出结果与该阈值的比较所得到的预测结果更加准确,再基于多个较为准确的预测结果中失衡状态所占的比例,确定出目标人体的肠道微生态是否失衡的结果,这种方式所确定出的结果的准确性更高。

[0078] 下面对肠道随机森林预测模型中的随机森林OTU模型的构建过程进行详细阐述,参见图3所示,该预测模型的构建方法包括以下步骤:

[0079] 步骤S302,获取第一样本集。

[0080] 其中,第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵,第一人体群体中各个人体的肠道微生态处于正常状态,第二人体群体中各个人体的肠道微生态处于失衡状态,目标矩阵为M*N的矩阵,第一样本集中包含M个样品,N表示每个样品对应N个物种标记物;目标矩阵由M个样品中,每个样品对应的N个物种标记物的丰度值组成。

[0081] 以人为例，第一人体群体为肠道微生态正常的人群，第二人体群体为肠道微生态失衡的人群，比如：第一样本集由500个肠道微生态正常的人和500个肠道微生态失衡的人的肠道微生态物种标记物的数据组成，一般来说，可以将每个人的肠道微生态物种标记物数据看作一个样品，每个样品包含N(如1000)个物种标记物，并且每个样品标记为一个表型，如正常或失衡。如图4所示，M*N矩阵由M个样品中，每个样品对应的N个物种标记物的用于表示物种标记物数量的丰度值组成。

[0082] 步骤S304，基于第一样本集，建立肠道随机森林模型。

[0083] 具体的，包括以下步骤：

[0084] (1)按照有放回抽样方式，对第一样本集进行M次抽样，得到M个训练集，并为每个训练集建立决策树，得到M棵决策树。

[0085] (2)对M棵决策树中每棵决策树进行分裂处理，得到M棵最大限度生长的决策树，以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征。

[0086] 对M棵决策树中每棵决策树进行分裂处理包括：

[0087] 从第一样本集中的N个物种标记物中，为每棵决策树中每个分裂节点选择对应的分裂特征，从而实现对每棵决策树进行分裂处理。

[0088] (3)将M棵最大限度生长的决策树进行组合，得到肠道随机森林模型。

[0089] 具体的，假设M表示训练样本个数，N表示样本中的特征数目，随机森林的构建过程如下：

[0090] 1)输入特征数目N，用于确定决策树上一个节点的决策结果；其中n应远小于N。

[0091] 2)从M个训练样本中以有放回抽样的方式，取样M次，形成一个训练集，并用未抽到的样本作预测，评估其误差。

[0092] 3)对于每一个节点，随机选择n个特征，决策树上每个节点的决定都是基于这些特征确定的。根据n个特征，计算其最佳的分裂方式。

[0093] 4)每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用。

[0094] 5)重复上述步骤，构建另外一棵棵决策树，直到达到预定数目的一群决策树为止，即构建好了随机森林。

[0095] 具体实施中，对于样品，采用有放回的方式。对于物种标记物，从N个物种标记物中，选择n个(n<N)，即：当每个样本有N个属性，在决策树的每个节点需要分裂时，随机从这N个属性中选取出n个属性。对采样之后的数据使用完全分裂的方式建立出决策树。分裂的办法是：采用上面说的列采样的过程从这n个属性中采用某种策略(比如说某个物种标记物)来选择1个属性作为该节点的分裂属性。决策树形成过程中每个节点都要按完全分裂的方式来分裂，一直到不能够再分裂为止。

[0096] Mtry参数是随机森林建模中，构建决策树分支时随机抽样的变量个数。选择合适的Mtry参数值可以降低随机森林模型的预测错误率。例如有n个物种标记物，可通过遍历设定Mtry参数为1至m进行m次建模，并得出每次建模的错误率，选择错误率最低的Mtry参数进行随机森林模型建模。

[0097] 步骤S306，基于肠道随机森林模型，从第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集。

[0098] 具体筛选过程包括以下步骤：

[0099] (1)通过肠道随机森林模型中的重要性importance函数，输出N个物种标记物的重要性的排序结果。

[0100] (2)将排序结果中前预设个数最大重要性所对应的物种标记物作为重要物种标记物。

[0101] (3)基于重要物种标记物的序列分类单位OTU表，组成第二样本集。

[0102] 在实际应用中，通过实验数据证明，选取排序结果中前15个物种标记物作为重要物种标记物，组成样本进行模型训练，最终得到的模型的预测精度较高。通过上述过程所筛选的重要物种标记物如下表所示：

[0103]

ID	Markers
1	Bifidobacterium adolescentis
2	Bifidobacterium longum
3	Faecalibacterium prausnitzii
4	Akkermansia muciniphila
5	Bifidobacterium
6	Lactobacillus
7	Streptococcus thermophilus
8	Faecalibacterium
9	Dorea
10	Pseudobutyrivibrio
11	Roseburia
12	Prevotella
13	Escherichia coli
14	Parabacteroides distasonis
15	Bacteroides fragilis

[0104]

[0105] 步骤S308，基于第二样本集构建随机森林OTU模型。

[0106] 具体的，基于第二样本集，采用留一交叉验证法和目标建模方法建立随机森林OTU模型，其中，目标建模方法为用于构建肠道随机森林模型的方法。

[0107] 在具体实施中，基于第二样本集中的每个样本构建M个测试集和M个训练集，其中，M为第二样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第二样本集中除第m个样本之外的其他样本；

[0108] 基于M个训练集及目标建模方法，建立M个第一子模型，作为随机森林OTU(Operational Taxonomic Unit)模型。

[0109] 上述留一交叉验证法中，假设样本数据集中有M个样本数据。将每个样本单独作为测试集，其余M-1个样本作为训练集，这样得到了M个分类器或模型，用这M个分类器或模型的分类准确率的平均数作为此分类器的性能指标。由于每一个分类器或模型都是用几乎所有的样本来训练模型，最接近样本，这样评估所得的结果比较可靠。实验没有随机因素，整个过程是可重复的。

[0110] 通过阈值和模型的输出结果的比较，才能得到预测结果，因此，上述方法还包括计算模型阈值的过程，如下：

[0111] (1)利用M个测试集中与M个第一子模型相对应的测试集,对每个第一子模型进行预测,得到M个第一输出结果。

[0112] (2)根据M个第一输出结果绘制第一ROC(Receiver Operating Characteristic Curve)特征曲线。

[0113] (3)基于第一ROC特征曲线,计算随机森林OTU模型的阈值,其中,阈值用于判断样品表型。

[0114] 该方法是在M个样品(如500个肠道微生态正常的人和500个肠道微生态失衡的人)中,随机选择一个人作为测试集,剩余人群的数据作为训练集按照上述肠道随机森林模型建立方法进行建模,并对测试集进行预测。上述过程反复执行,直至每个个体都被选择一次作为测试集,统计每个个体预测概率结果。根据统计预测概率结果绘制ROC曲线(受试者工作特征曲线)及AUC(Area Under Curve)值(AUC值为ROC曲线所覆盖的区域面积,显然,AUC越大,分类器分类效果越好),进而计算特异度,敏感度,最终选择出最佳的判断样品表型(正常或肠道菌群失衡)的阈值。图5为本发明实施例所提供的一种ROC曲线,在ROC曲线上,最靠近坐标图左上方的点为敏感度和特异度均较高的临界值,即最佳阈值。为了获取ROC曲线的最佳阈值,需要使用一个指标--约登指数,也称正确指数。约登指数是灵敏度与特异度之和减去1。从图5中确定出的阈值为0.628,模型区别正常样品与肠道微生态失衡的样品的灵敏度是0.737(73.7%),专一性是0.957(95.7%)。

[0115] 在另一种可能的实施方式中,在基于肠道随机森林模型,从第一样本集的N个物种标记物中,筛选出预设个数的重要物种标记物,组成第二样本集之后,还包括以下步骤,参见图6所示:

[0116] 步骤S602,根据重要物种标记物的丰度值,计算第二样本集中每个样品的Shannon指数。

[0117] 具体实施中,利用下式,计算第二样本集中每个样品的Shannon指数:

[0118] $S=-(P_1 x\ lnP_1+P_2 x\ lnP_2+...+P_n\ x\ lnP_n)$;

[0119] 其中,P_1、P_2...P_n为样品第1个、第2个...第n个物种标记物的丰度值,S为样品的Shannon指数。

[0120] Chao1,ACE,Shannon,npShannon,Simpson和Good's Coverage为六大类常用的α多样性指数。总体来说,Chao1/ACE指数主要关心样本的物种丰富度信息;Good's Coverage反映样本的低丰度OTU覆盖情况;Simpson/Shannon/npShannon主要综合体现物种的丰富度和均匀度。Shannon指数与Chao1,ACE不一样,Chao1和ACE主要用于计算物种的丰富度(Richness),更在乎样本是否有这个物种。而Shannon指数不只关心物种丰富度,而且同时关心物种的均匀度(Evenness),所以是对群落结构的更综合性的反应。

[0121] 计算物种多样性指数的公式有很多,形式各异,而实质是差不多的。大部分多样性指数中,组成群落的生物种类越多,其多样性指值越大。

[0122] 步骤S604,将Shannon指数添加至重要物种标记物的OTU表中,得到第三样本集。

[0123] 步骤S606,基于第三样本集,采用留一交叉验证法和目标建模方法,建立Shannon指数随机森林模型,其中,目标建模方法为用于构建肠道随机森林模型的方法。

[0124] 具体Shannon指数随机森林模型的建模过程同上述随机森林OTU模型的建立过程,具体如下:

[0125] 基于第三样本集中的每个样本构建M个测试集和M个训练集，其中，M为第三样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第三样本集中除第m个样本之外的其他样本。

[0126] 基于M个训练集和目标建模方法建立M个第二子模型，作为Shannon指数随机森林模型。

[0127] 同样，上述方法还包括Shannon指数随机森林模型的阈值的计算过程，如下：

[0128] (1)利用M个测试集中与M个第二子模型相对应的测试集，对每个第二子模型进行预测，得到M个第二预测结果。

[0129] (2)根据M个第二预测结果绘制第二ROC特征曲线。

[0130] (3)基于第二ROC特征曲线，计算Shannon指数随机森林模型的阈值，其中，阈值用于判断样品表型。

[0131] 在本实施方式中，每个样品的Shannon指数当作一个新的“OTU”，加到前述的15个重要物种标记物OTU表里，形成一个新的16个OTU组成的数据表(15个OTU+Shannon指数)。对这新的OTU表，采用留一交叉验证法(Leave one out cross validation)建立Shannon指数随机森林模型。该方法是对500个正常人群和500个肠道微生态失衡人群随机选择一个人作为测试集，剩余人群作为训练集按照上述肠道随机森林模型建立方法进行建模，并对测试集进行预测。上述过程直至每个个体都被选择一次作为测试集，每个个体均只能被选择一次，统计每个个体预测概率结果。根据统计的预测概率结果绘制ROC曲线(受试者工作特征曲线)及AUC值，计算特异度，敏感度，从而选择出最佳的判断阈值。图7表示本例的ROC曲线，判断阈值为0.717，这个模型区别正常样品与肠道微生态失衡的样品的灵敏度是0.868(86.8%)，专一性是0.826(82.6%)。与上一模型比较，AUC值提高，灵敏度提高。

[0132] 下面对目标人体的肠道菌群数据的获取过程，即OUT表生成过程进行详细的阐述：

[0133] 1.基因测序仪得到的数据的初步处理：

[0134] 高通量测序产生出了海量的DNA序列数据，如何对这些数据进行生物信息学处理，提取有生物学意义的信息是整个项目的重要一环。本实施例采用的是16S rRNA高通量测序数据处理的生物信息平台。此平台整合了公共数据库资源及一些开源软件，有自主在Linux操作系统编写的Perl、Python、R语言程序，对16S rRNA数据进行分析。

[0135] 数据质量控制：高通量测序中通常会出现一些点突变等测序错误，而且序列末端的质量比较低，为了得到更高质量及更准确的生物信息分析结果，需要对测序原始数据进行优化处理。高通量测序数据一般包含序列信息和测序质量数据，对测序结果原始图像数据利用软件CASAVA(v1.8.2)进行图像碱基识别(Base calling)，初步质量分析，得到测序原始数据rawdata，结果以FASTQ文件格式存储，其中包含测序序列信息(FASTQ格式第二列)与其对应的测序质量信息(FASTQ格式第四列)。

[0136] 优化步骤及参数：

[0137] 1)去除引物及接头序列，去除两端质量值低于25的碱基，划窗法去除平均质量低于25的碱基。

[0138] 2)使用pandaseq软件将两条序列进行比对，根据比对的末端重叠区进行拼接，去除拼接结果中含有N的序列，去除拼接结果中大于480bp小于400bp的序列。

[0139] 3)将上面拼接过滤后的序列与数据库进行比对，去除其中的嵌合体序列，得到最

终的有效数据。

[0140] 质量检查还包括去除质量不好(读长异常、碱基识别模糊等)的序列,然后进行嵌合体检查,如果有叶绿体和线粒体序列的污染,应该先把这部分序列剔除,高通量测序的数据中会存在一些嵌合体,数据处理时也要去除这部分序列。最后根据OTU矩阵和赋予的种属关系进行统计分析,将选用软件如MAFFT、Blastn或系统性的开源软件如RDP、QIIM等。

[0141] 2.OTU计算:

[0142] 一个样品里可能有几百个或上千个微生物,16S rRNA测序测定每个微生物16SrRNA基因的可变区片段,可以简单理解为一个菌种在一个样品里只有一个个体,机器就测出这个种的一个片段,另一个菌种在这个样品里比较多(丰度高),比方说有100个个体,机器就测出100个该菌种的片段。理论上机器测出来的序列相同的片段是代表研究微生物种类的。在对机器出来的序列进行分析时,先把这些片段按序列的相似度进行归类分析(Clustering),将其归类为许多集,一般97%相似的就归为一个集,这个集就是一个OTU,OTU是Operational Taxonomic Units的简称,可称为序列分类单位。一般一个OTU对应一个微生物种,通过与标准微生物基因库如序列比较,就可知道是哪个物种。而每一个OTU在一个样品里测到的次数就是该微生物种在这个样品里的丰度。

[0143] 样本中OTU数量和相对丰度代表实际观察到的多样性,因此计算不同样品中OTU的分布,是高通量测序数据处理中一个非常重要的步骤。计算OTU之前,序列要进行比对(alignment)。比对的算法有很多,聚类(clustering)的方法也有多种,不同的软件所选择的算法各有不同,如ESPRIT平台应用成对比对法(pairwisealignment),使用Qiime平台分析方法:使用UCLUST方法进行OTU聚类,OTU中序列相似性设为97%,得到OTU列表及OTU代表性序列。使用RDP classifier贝叶斯算法对97%相似水平的OTU代表序列进行分类学分析,并在各个水平统计每个样本的群落组成,如图8所示。

[0144] 图9示出了本发明实施例提供的一种人体肠道微生态失衡的确定装置,包括:数据获取模块91、模型预测模块92和结果判断模块93。

[0145] 其中,数据获取模块91,用于获取目标人体的肠道菌群数据,其中,肠道菌群数据为包含多个物种标记物的一个样本数据;模型预测模块92,用于将肠道菌群数据输入至肠道随机森林预测模型中进行预测,得到输出结果;肠道随机森林预测模型对应有用于判断样品表型的阈值,肠道随机森林预测模型包括:随机森林OTU模型或Shannon指数随机森林模型;结果判断模块93,用于基于输出结果判断目标人体的肠道微生态是否处于失衡状态。

[0146] 当肠道随机森林预测模型的数量为多个时,模型预测模块92,还用于:将肠道菌群数据输入至每个肠道随机森林预测模型中进行预测,得到多个输出结果;结果判断模块93,还用于:基于多个输出结果确定每个肠道随机森林预测模型的预测结果,得到多个预测结果;并基于多个预测结果中目标预测结果的比例确定目标人体的肠道微生态是否处于失衡状态,其中,目标预测结果为多个预测结果中用于表征目标人体的肠道微生态处于失衡状态的预测结果。

[0147] 在具体实施中,结果判断模块93,还用于:若多个预测结果中,目标预测结果所占的比例大于第一预设比例阈值,则确定出目标人体的肠道微生态处于失衡状态;若多个预测结果中,目标预测结果所占的比例大于第二预设比例阈值,且小于第一预设比例阈值,则基于多个肠道随机森林预测模型中目标肠道随机森林预测模型的预测结果确定目标人体

的肠道微生态是否处于失衡状态，目标肠道随机森林预测模型为多个肠道随机森林预测模型中最大阈值所对应的肠道随机森林预测模型，第二预设比例阈值小于第一预设比例阈值，每个肠道随机森林预测模型对应一个阈值。

[0148] 在具体实施中，结果判断模块93，还用于：将肠道菌群数据分别输入肠道随机森林预测模型Ai中，得到输出结果Bi，其中，i依次取1至I，I为多个肠道随机森林预测模型的数量；若输出结果Bi大于或者等于肠道随机森林预测模型Ai所对应的阈值，则得到用于表征目标人体的肠道微生态处于失衡状态的预测结果；若输出结果Bi小于肠道随机森林预测模型Ai所对应的阈值，则得到用于表征目标人体的肠道微生态处于正常状态的预测结果。

[0149] 在另一种实施方式中，上述人体肠道微生态失衡的确定装置还包括：模型构建模块94，参见图10所示，模型构建模块94具体包括：第一样本集获取模块941、基础模型建立模块942、第二样本集获取模块943和预测模型构建模块944。

[0150] 其中，第一样本集获取模块941，用于获取第一样本集；第一样本集为基于从第一人体群体和第二人体群体中采集到的肠道菌群数据组成的目标矩阵，第一人体群体中各个人体的肠道微生态处于正常状态，第二人体群体中各个人体的肠道微生态处于失衡状态，目标矩阵为M*N的矩阵，第一样本集中包含M个样品，N表示每个样品对应N个物种标记物；目标矩阵由M个样品中每个样品对应的N个物种标记物的丰度值组成；基础模型建立模块942，用于基于第一样本集，建立肠道随机森林模型；第二样本集获取模块943，用于基于肠道随机森林模型，从第一样本集的N个物种标记物中，筛选出预设个数的重要物种标记物，组成第二样本集；预测模型构建模块944，用于基于第二样本集构建随机森林OTU模型。

[0151] 在具体实施中，预测模型构建模块944，还用于：基于第二样本集，采用留一交叉验证法和目标建模方法建立随机森林OTU模型，其中，目标建模方法为用于构建肠道随机森林模型的方法。

[0152] 上述基础模型建立模块942，还用于：按照有放回抽样方式，对第一样本集进行M次抽样，得到M个训练集，并为每个训练集建立决策树，得到M棵决策树；对M棵决策树中每棵决策树进行分裂处理，得到M棵最大限度生长的决策树，以确定每棵决策树的分裂节点和每个分裂节点所对应的分裂特征；将M棵最大限度生长的决策树进行组合，得到肠道随机森林模型。

[0153] 在另一种实施方式中，基础模型建立模块942，还用于：在第一样本集中的N个物种标记物中，为每棵决策树中每个分裂节点选择对应的分裂特征，从而实现对每棵决策树进行分裂处理。

[0154] 在具体实施中，第二样本集获取模块943，还用于：通过肠道随机森林模型中的重要性importance函数，输出N个物种标记物的重要性的排序结果；将排序结果中前预设个数最大重要性所对应的物种标记物作为重要物种标记物；基于重要物种标记物的序列分类单位OTU表，组成第二样本集。

[0155] 在另一种实施方式中，预测模型构建模块944，还用于：基于第二样本集中的每个样本构建M个测试集和M个训练集，其中，M为第二样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第二样本集中除第m个样本之外的其他样本；基于M个训练集及目标建模方法，建立M个第一子模型，作为随机森林OTU模型。

[0156] 在另一种可能的实施方式中，上述人体肠道微生态失衡的确定装置还包括：第一

阈值确定模块946，用于：利用M个测试集中与M个第一子模型相对应的测试集，对各个第一子模型进行预测，得到M个第一输出结果；根据M个第一输出结果绘制第一ROC特征曲线；基于第一ROC特征曲线，计算随机森林OTU模型的阈值，其中，阈值用于判断样品表型。

[0157] 在另一种可能的实施方式中，上述模型构建模块94，还包括：第三样本集获取模块945，用于根据重要物种标记物的丰度值，计算第二样本集中每个样品的Shannon指数；将Shannon指数添加至重要物种标记物的OTU表中，得到第三样本集；预测模型构建模块944，还用于基于第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，其中，目标建模方法为用于构建肠道随机森林模型的方法。

[0158] 在具体实施中，第三样本集获取模块945，还用于：利用下式，计算第二样本集中每个样品的Shannon指数：

[0159] $S=-(P_1 x\ lnP_1+P_2 x\ lnP_2+...+P_n\ x\ lnP_n)$；

[0160] 其中，P_1、P_2...P_n为样品第1个、第2个...第n个物种标记物的丰度值，S为样品的Shannon指数。

[0161] 上述预测模型构建模块944，还用于：基于第三样本集中的每个样本构建M个测试集和M个训练集，其中，M为第三样本集中样本的数量，当M个测试集中的第m个测试集为第m个样本时，M个训练集中第m个训练集为第三样本集中除第m个样本之外的其他样本；基于M个训练集和目标建模方法建立M个第二子模型，作为Shannon指数随机森林模型。

[0162] 在另一种可能的实施方式中，上述人体肠道微生态失衡的确定装置还包括：第二阈值确定模块947，用于：利用M个测试集中与M个第二子模型相对应的测试集，对各个第二子模型进行预测，得到M个第二预测结果；根据M个第二预测结果绘制第二ROC特征曲线；基于第二ROC特征曲线，计算Shannon指数随机森林模型的阈值，其中，阈值用于判断样品表型。

[0163] 本发明实施例所提供的人体肠道微生态失衡的确定装置中，各个模块与前述人体肠道微生态失衡的确定方法具有相同的技术特征，因此，同样可以实现上述功能。本装置中各个模块的具体工作过程参见上述方法实施例，在此不再赘述。

[0164] 图11示出了本发明实施例所提供的一种电子设备，该电子设备包括：处理器110，存储器111，总线112和通信接口113，所述处理器110、通信接口113和存储器111通过总线112连接；处理器110用于执行存储器111中存储的可执行模块，例如计算机程序。处理器执行计算机程序时实现如方法实施例所述的方法的步骤。

[0165] 其中，存储器111可能包含高速随机存取存储器(RAM，RandomAccessMemory)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。通过至少一个通信接口113(可以是有线或者无线)实现该系统网元与至少一个其他网元之间的通信连接，可以使用互联网，广域网，本地网，城域网等。

[0166] 总线112可以是ISA总线、PCI总线或EISA总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图11中仅用一个双向箭头表示，但并不表示仅有一根总线或一种类型的总线。

[0167] 其中，存储器111用于存储程序，所述处理器110在接收到执行指令后，执行所述程序，前述本发明实施例任一实施例揭示的流过程定义的装置所执行的方法可以应用于处理器110中，或者由处理器110实现。

[0168] 处理器110可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可以通过处理器110中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器110可以是通用处理器，包括中央处理器(Central Processing Unit，简称CPU)、网络处理器(Network Processor，简称NP)等；还可以是数字信号处理器(Digital Signal Processing，简称DSP)、专用集成电路(Application Specific Integrated Circuit，简称ASIC)、现成可编程门阵列(Field-Programmable Gate Array，简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实现或者执行本发明实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本发明实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器111，处理器110读取存储器111中的信息，结合其硬件完成上述方法的步骤。

[0169] 本发明实施例所提供的人体肠道微生态失衡的确定方法的计算机程序产品，包括存储了处理器可执行的非易失的程序代码的计算机可读存储介质，所述程序代码包括的指令可用于执行前面方法实施例中所述的方法，具体实现可参见方法实施例，在此不再赘述。

[0170] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置及电子设备的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0171] 附图中的流程图和框图显示了根据本发明的多个实施例方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0172] 在本发明的描述中，需要说明的是，术语“中心”、“上”、“下”、“左”、“右”、“竖直”、“水平”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0173] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，又例如，多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0174] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显

示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0175] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0176] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个处理器可执行的非易失的计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0177] 最后应说明的是:以上所述实施例,仅为本发明的具体实施方式,用以说明本发明的技术方案,而非对其限制,本发明的保护范围并不局限于此,尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

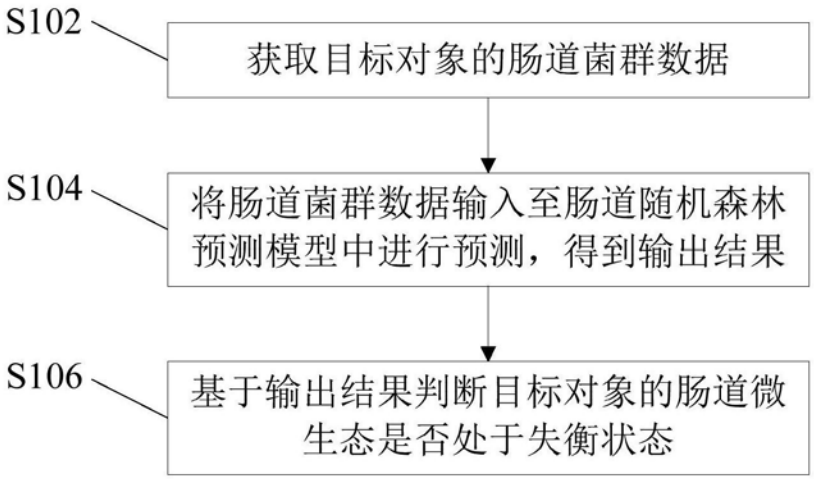

S102
获取目标对象的肠道菌群数据
S104
将肠道菌群数据输入至肠道随机森林预测模型中进行预测，得到输出结果
S106
基于输出结果判断目标对象的肠道微生态是否处于失衡状态

图1

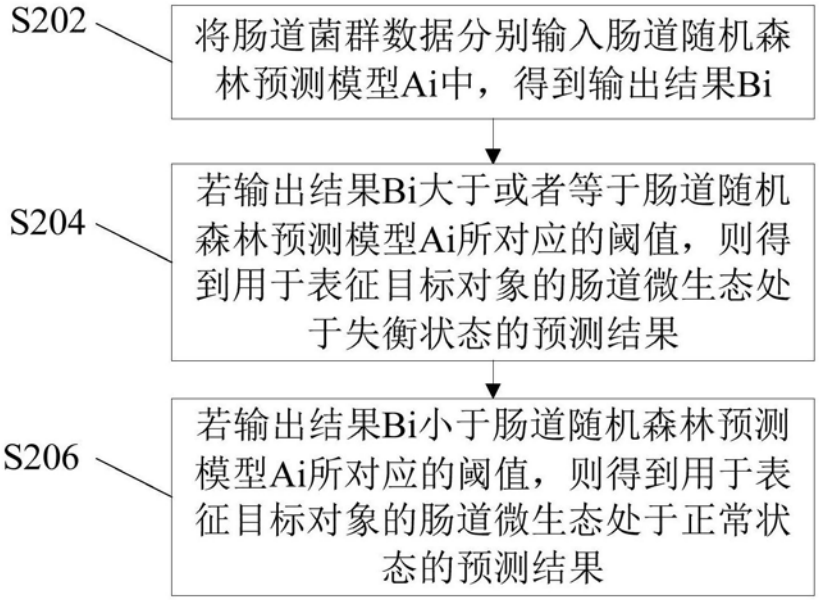

S202
将肠道菌群数据分别输入肠道随机森林预测模型Ai中，得到输出结果Bi
S204
若输出结果Bi大于或者等于肠道随机森林预测模型Ai所对应的阈值，则得到用于表征目标对象的肠道微生态处于失衡状态的预测结果
S206
若输出结果Bi小于肠道随机森林预测模型Ai所对应的阈值，则得到用于表征目标对象的肠道微生态处于正常状态的预测结果

图2

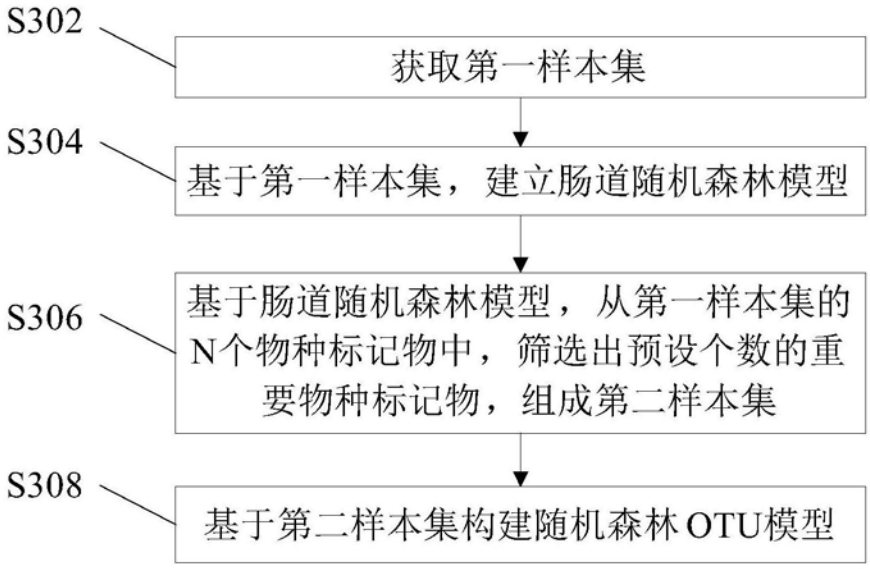


图3

	种1	种2	种3	种4	…	种N
样1	0.01	0.03	0.03	0.06	…	n1
样2	0.07	0.01	0.03	0.18	…	n2
…..						
样M	0.02	0.10	0.07	0.02	…	n

图4

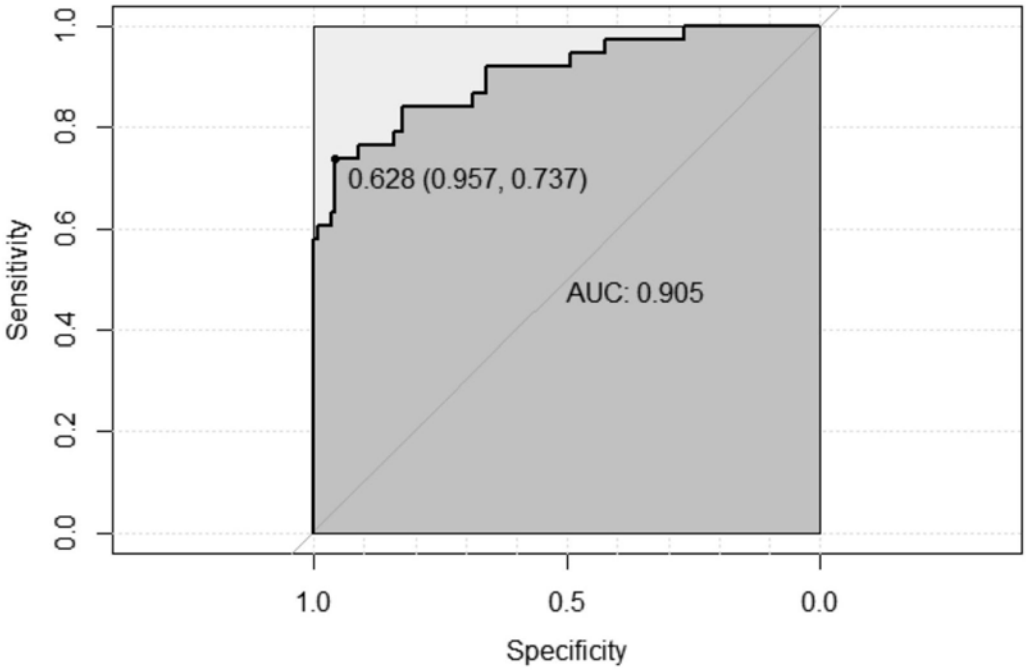

Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.628 (0.957, 0.737)
AUC: 0.905
1.0
0.5
0.0
Specificity

图5

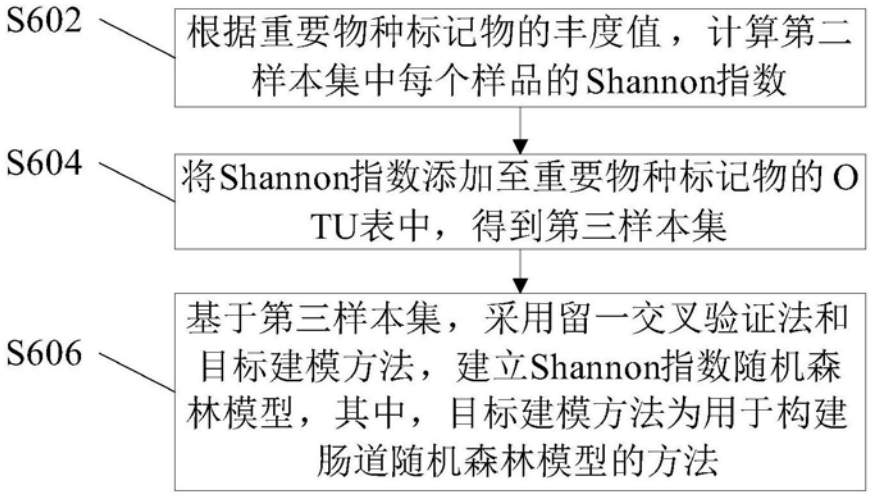

S602
根据重要物种标记物的丰度值，计算第二样本集中每个样品的Shannon指数
S604
将Shannon指数添加至重要物种标记物的OTU表中，得到第三样本集
S606
基于第三样本集，采用留一交叉验证法和目标建模方法，建立Shannon指数随机森林模型，其中，目标建模方法为用于构建肠道随机森林模型的方法

图6

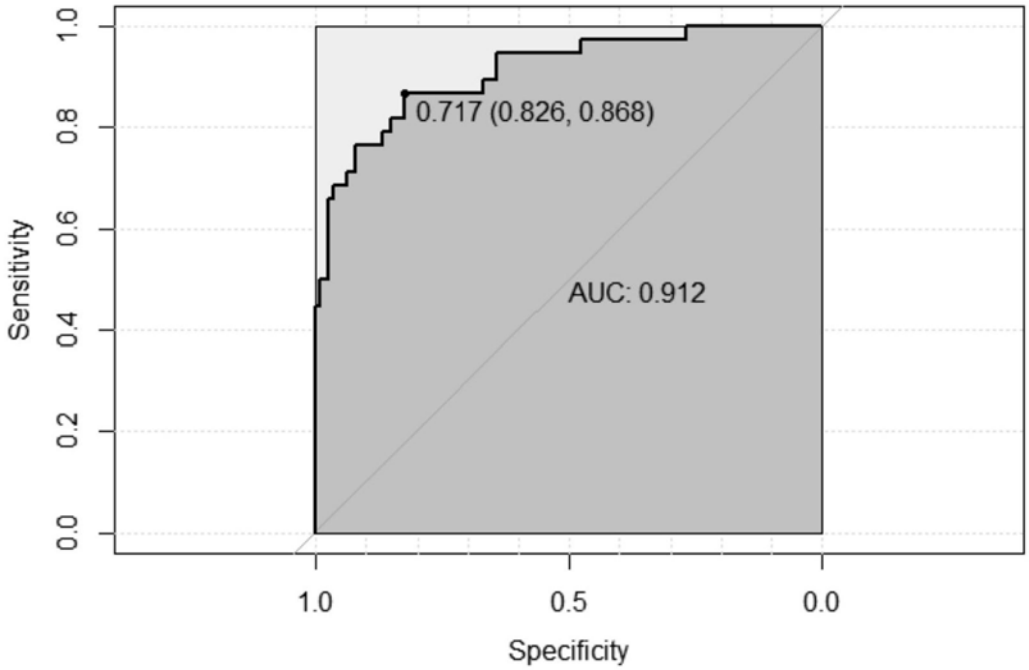

Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.717 (0.826, 0.868)
AUC: 0.912
1.0
0.5
0.0
Specificity

图7

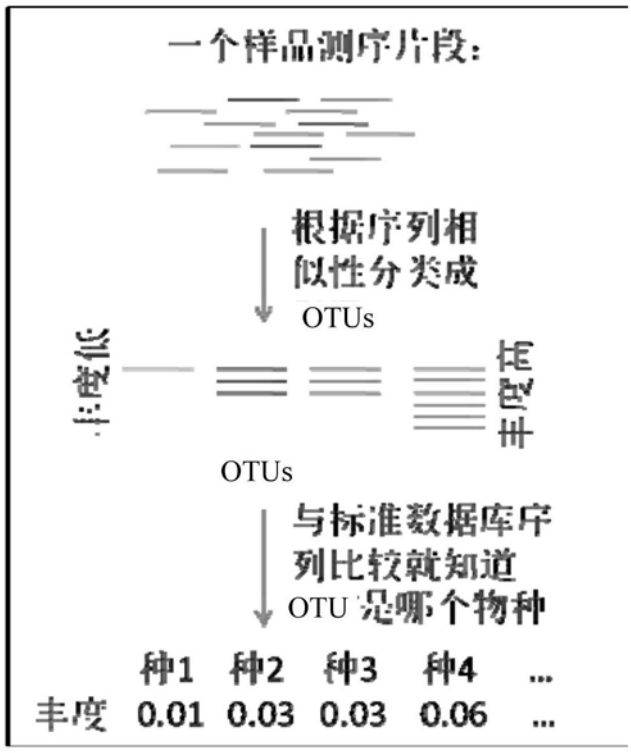

一个样品测序片段:
根据序列相似性分类成OTUs
丰度
丰度
OTUs
与标准数据库序列比较就知道OTU是哪个物种
种1 种2 种3 种4 ...
丰度 0.01 0.03 0.03 0.06 ...

图8

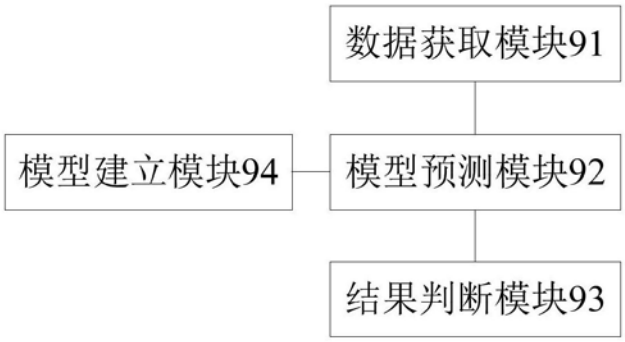

数据获取模块91
模型建立模块94
模型预测模块92
结果判断模块93

图9

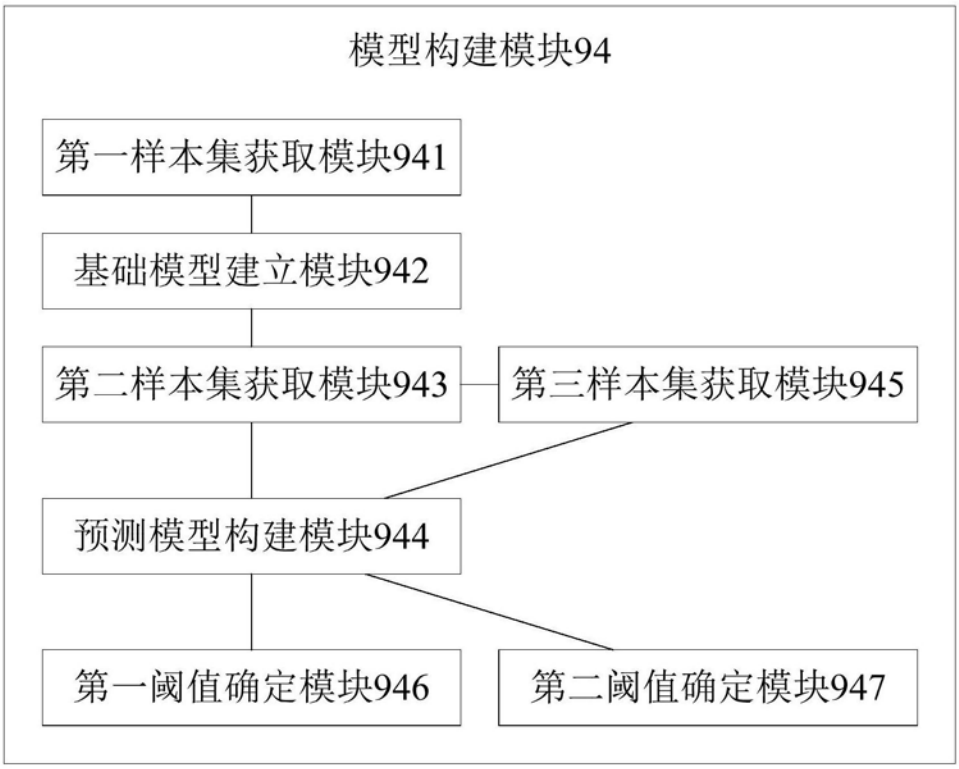

模型构建模块94
第一样本集获取模块941
基础模型建立模块942
第二样本集获取模块943
第三样本集获取模块945
预测模型构建模块944
第一阈值确定模块946
第二阈值确定模块947

图10

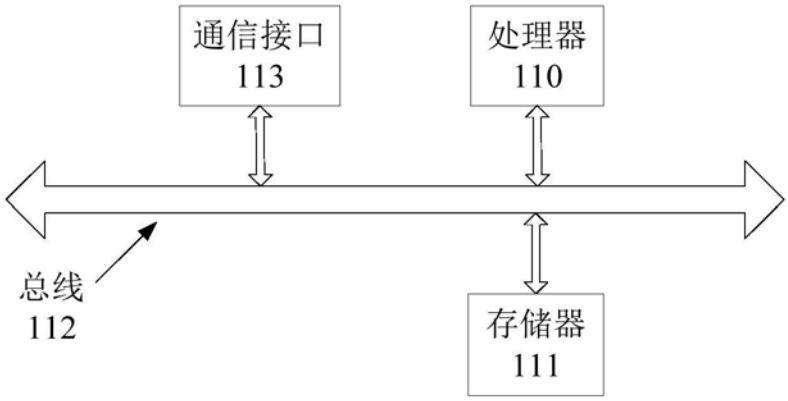


图11