(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107292124 A
(43)申请公布日 2017.10.24

(21)申请号 201710490528.5

(22)申请日 2017.06.25

(71)申请人 广东国盛医学科技有限公司
地址 510000 广东省广州市高新技术产业开发区科学城开源大道11号C1栋第四层

(72)发明人 郑灏

(74)专利代理机构 北京科家知识产权代理事务所(普通合伙) 11427
代理人 李雪鹃

(51)Int.Cl.
G06F 19/18(2011.01)
G06F 19/24(2011.01)
G06F 19/20(2011.01)
G06N 3/02(2006.01)

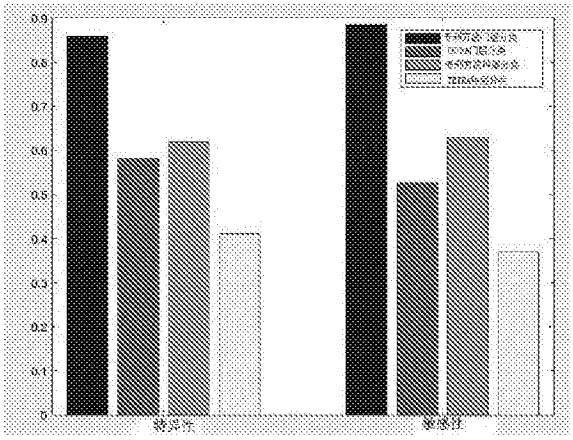
权利要求书1页 说明书3页 附图3页

(54)发明名称

基于分层主元深度学习的宏基因组操作分类单元识别方法

(57)摘要

本发明提供一种利用主元分析的特征向量结果去引导神经网络深度学习的初始化，属于宏基因组的操作分类单元识别技术领域，通过Relu开启函数和多次交叉验证学习，对预处理后的宏基因组特征来进行对宏基因组的分层OTU分类的方法，具有特异性和敏感性高的优点。



CN 107292124 A

1.基于分层主元深度学习的宏基因组操作分类单元识别方法，其特征在于：其包括如下步骤：

步骤S1样品处理：从样品中分离出存在于样品中的微生物，提取微生物中的所有DNA，并对提取的DNA进行高通量测序；

步骤S2数据预处理：对步骤S1中得到的reads、contigs和scaffold进行初步分析，将重复的DNA序列信息和已知的低质量区域的DNA序列信息剔除；

步骤S3基因特征分析：对DNA六聚体结构的混沌序列特征分析提取，确定并获得宏基因组特征信息；

步骤S4主元分析：录入宏基因组特征信息，通过统计检验筛选重要特征信息，对重要特征信息进行主元分析；

步骤S5建立神经网络分类模型：根据步骤S4主元分析结果作为初始化信息建立神经网络分类模型，再通过Relu开启函数$f(x) = max(0, x)$并进行多次交叉验证学习，对宏基因组进行分层操作分类单元分类。

2.根据权利要求1所述的基于分层主元深度学习的宏基因组操作分类单元识别方法，其特征在于：所述的步骤S2数据预处理还包括步骤S21保守序列分类，判断宏基因组是否存在保守区域序列，若存在保守区域序列则使用BLAST进行操作分类单元分类，不存在保守区域序列则在步骤S2结束后直接执行步骤S3。

3.根据权利要求1所述的基于分层主元深度学习的宏基因组操作分类单元识别方法，其特征在于：所述的步骤S4主元分析的具体操作如下：

对于高纬序列特征向量，$\{x\} \subset X$，通过奇异值分解得到一个酉矩阵$\Theta \in R^{M \times M}$，把每个高维空间中的x向量通过线性变换映射到 $y \equiv [y1, y2, \ldots, y^{M}]^{T}$，$Y = \Theta x - \Theta \mu x$；

其中μx空是{x}的均值；得到的 Θ 作为初始信息，引导建立神经网络分类模型。

基于分层主元深度学习的宏基因组操作分类单元识别方法

技术领域

[0001] 本发明属于宏基因组的操作分类单元识别技术领域，尤其涉及一种基于分层主元深度学习的宏基因组操作分类单元识别方法。

背景技术

[0002] 宏基因组学是一门新兴的生物信息和分子生物学研究，其技术避开传统的微生物分离培养方法直接从环境样品中提取总DNA，为科学家们研究环境微生物的种类和分布打开了一个新的篇章。

[0003] 操作分类单元(OTU)识别是宏基因组学中的一项核心技术，其目的在于研究宏基因组中的微生物种类和比例。随着最近下一代测序技术的大规模发展，使得深度研究宏基因组学成为可能，好的OTU分类算法更显得尤为重要。

[0004] 目前比较流行的操作分类单元(OTU)识别分类方法有TETRA和Phylopythia。TETRA利用四聚体结构序列特征对宏基因组进行OTU识别；Phylopythia利用已知的DNA序列基于支持向量机方法对宏基因组进行OTU识别，但上述两种方法的OTU识别的特异性和敏感性低，不能满足进一步的科学研究分析的需求。

发明内容

[0005] 基于现有技术存在上述问题，本发明提供一种利用主元分析的特征向量结果去引导神经网络深度学习的初始化，通过Relu开启函数和多次交叉验证学习，对预处理后的宏基因组特征来进行对宏基因组的分层OTU分类的方法，具有特异性和敏感性高的优点。

[0006] 本发明通过以下技术方案达到上述目的：

一种基于分层主元深度学习的宏基因组操作分类单元识别方法，其包括如下步骤：

步骤S1样品处理：从样品中分离出存在于样品中的微生物，提取微生物中的所有DNA，并对提取的DNA进行高通量测序；

步骤S2数据预处理：对步骤S1中得到的reads、contigs和scaffold进行初步分析，将重复的DNA序列信息和已知的低质量区域的DNA序列信息剔除；

步骤S3基因特征分析：对DNA六聚体结构的混沌序列特征分析提取，确定并获得宏基因组特征信息；

步骤S4主元分析：录入宏基因组特征信息，通过统计检验筛选重要特征信息，对重要特征信息进行主元分析；

步骤S5建立神经网络分类模型：根据步骤S4主元分析结果作为初始化信息建立神经网络分类模型，再通过Relu开启函数$f(x) = max(0, x)$并进行多次交叉验证学习，对宏基因组进行分层操作分类单元分类。

[0007] 其中，所述的步骤S2数据预处理还包括步骤S21保守序列分类，判断宏基因组是否存在保守区域序列，若存在保守区域序列则使用BLAST进行操作分类单元分类，不存在保守区域序列则在步骤S2结束后直接执行步骤S3。

[0008] 其中,所述的步骤S4主元分析的具体操作如下:

对于高纬序列特征向量,{x} X,通过奇异值分解得到一个酉矩阵Θ ∈ $R^{M\times M}$,把每个高维空间中的x向量通过线性变换映射到 y ≡ [y1,y2,...,y^M]^T,Y = Θx - Θμx;

其中μx空是{x}的均值;得到的 Θ 作为初始信息,引导建立神经网络分类模型。

附图说明

[0009] 图1,利用本发明提供的方法和TETRA对simHC数据集进行分析的结果对比图。

[0010] 图2,利用本发明提供的方法和Phylopythia对模拟合成数据进行分析的结果对比图。

[0011] 图3,本专利方法与TETRA、Phylopythia综合对比结果图。

具体实施方式

[0012] 下面结合具体实施例对本发明作进一步的描述。

[0013] 实施例一,使用simHC数据集进行OUT分类

利用广泛应用的simHC数据集进行OUT分类,SimHC 中含有113 个物种的基因组,DNA长度从130 到3,754 bps,其包括以下步骤:

一种基于分层主元深度学习的宏基因组操作分类单元识别方法,其包括如下步骤:

由于simHC数据集中的物种基因组均已经采用常规的分离提取方法完成提取,故省略步骤S1中的微生物分离,直接对simHC数据集中的基因组进行高通量测序。

[0014] 步骤S2数据预处理:对步骤S1中得到的reads、contigs和scaffold进行初步分析,将重复的DNA序列信息和已知的低质量区域的DNA序列信息剔除。

[0015] 步骤S21保守序列分类,判断宏基因组是否存在保守区域序列,存在保守区域序列,使用BLAST进行预先操作分类单元分类。

[0016] 步骤S3基因特征分析:对余下未分类的DNA六聚体结构的混沌序列特征分析提取,确定并获得simHC数据集特征信息,如结构信息、各功能序列位置信息、序列信息等特征。

[0017] 步骤S4主元分析:录入simHC数据集特征信息,通过统计检验筛选重要特征信息,对重要特征信息进行主元分析,具体如下:

对于高纬序列特征向量,{x} X,通过奇异值分解得到一个酉矩阵Θ ∈ $R^{M\times M}$,把每个高维空间中的x向量通过线性变换映射到 y ≡ [y1,y2,...,y^M]^T,Y = Θx - Θμx;

其中μx空是{x}的均值;得到的 Θ 作为初始信息,引导建立神经网络分类模型。

[0018] 步骤S5建立神经网络分类模型:根据步骤S4主元分析结果得到的酉矩阵Θ作为初始化信息建立神经网络分类模型,再通过Relu开启函数f(x) = *max*(0, x)并进行多次交叉验证学习,对simHC数据集进行分层操作分类单元分类,分类结果和TETRA对比结果如附图1。

[0019] 实施例二,使用模拟合成数据进行OUT分类

一种基于分层主元深度学习的宏基因组操作分类单元识别方法,其包括如下步骤:

由于模拟合成数据的物种基因组采用已知的物种基因组合成,故省略步骤S1中的微生物分离,直接对模拟合成数据中的基因组进行高通量测序。

[0020] 步骤S2数据预处理:对步骤S1中得到的reads、contigs和scaffold进行初步分析,

将重复的DNA序列信息和已知的低质量区域的DNA序列信息剔除。

[0021] 步骤S21保守序列分类，判断宏基因组是否存在保守区域序列，存在保守区域序列，使用BLAST进行预先操作分类单元分类。

[0022] 步骤S3基因特征分析：对余下未分类的DNA六聚体结构的混沌序列特征分析提取，确定并获得模拟合成数据特征信息，如结构信息、各功能序列位置信息、序列信息等特征。

[0023] 步骤S4主元分析：录入模拟合成数据特征信息，通过统计检验筛选重要特征信息，对重要特征信息进行主元分析，具体如下：

对于高纬序列特征向量，{x} X，通过奇异值分解得到一个酉矩阵 $\Theta \in R^{M \times M}$，把每个高维空间中的x向量通过线性变换映射到 $y \equiv [y1,y2,...,y^{M}]^{T}$，$Y = \Theta x - \Theta \mu x$；

其中μx空是{x}的均值；得到的 Θ 作为初始信息，引导建立神经网络分类模型。

[0024] 步骤S5建立神经网络分类模型：根据步骤S4主元分析结果得到的酉矩阵Θ作为初始化信息建立神经网络分类模型，再通过Relu开启函数 $f(x) = max(0, x)$ 并进行多次交叉验证学习，对模拟合成数据进行分层操作分类单元分类，分类结果和Phylopythia对比结果如附图2。

[0025] 实施例三，本专利方法与TETRA、Phylopythia综合对比

多次采用本专利方法对不同样品进行操作分类单元分类并与TETRA和Phylopythia对比，对比结果如附图3。

[0026] 从上述三个实施例可以看出本专利提供的方法具有更高的特异性和敏感性。

[0027] 以上所述实施例仅表达了本发明的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。因此，本发明专利的保护范围应以所附权利要求为准。

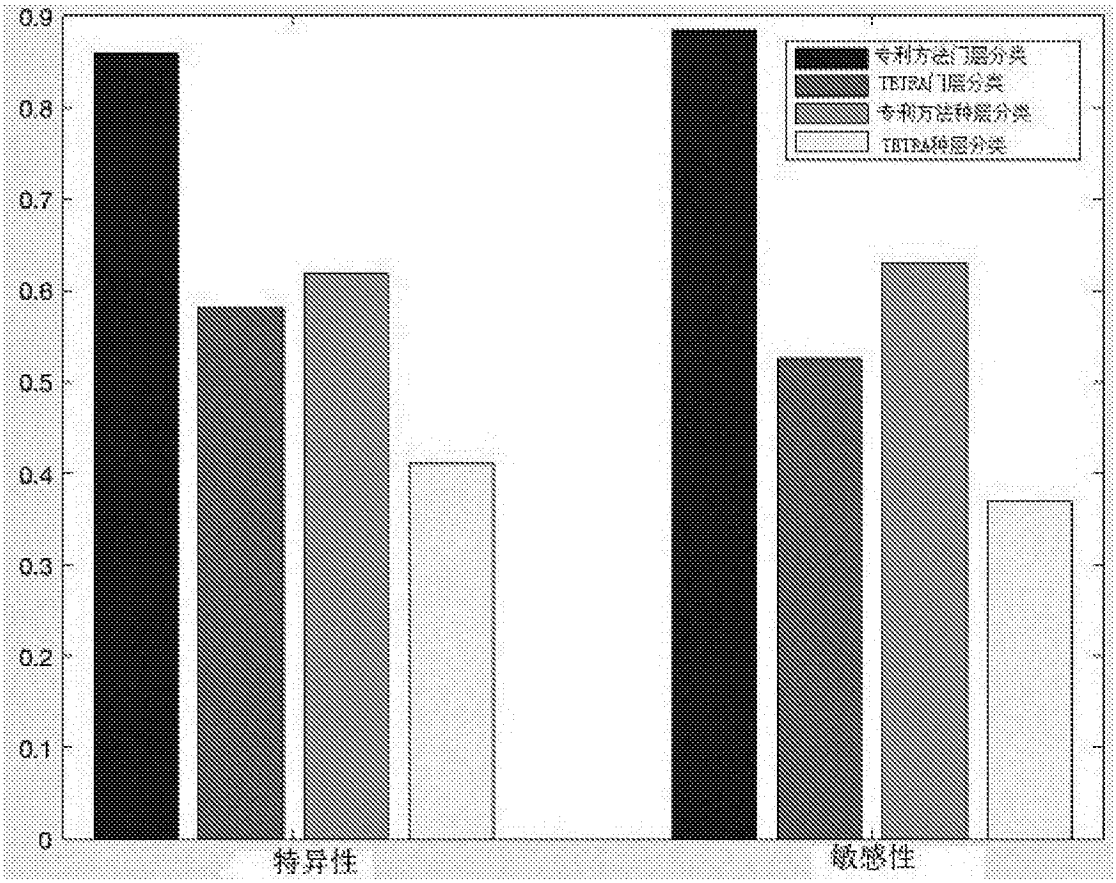

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
特异性
敏感性
专利方法门层分类
TETRA门层分类
专利方法种层分类
TETRA种层分类

图1

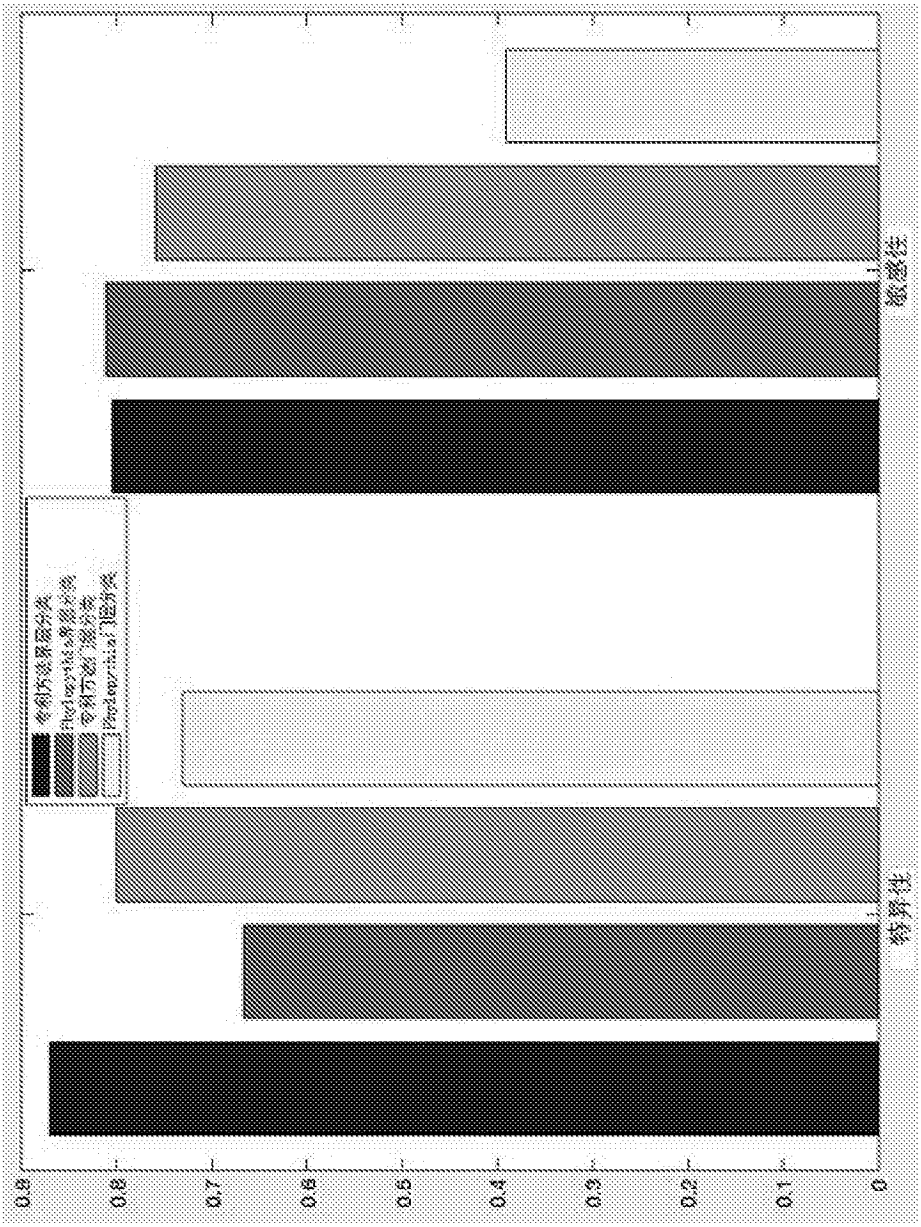


图2

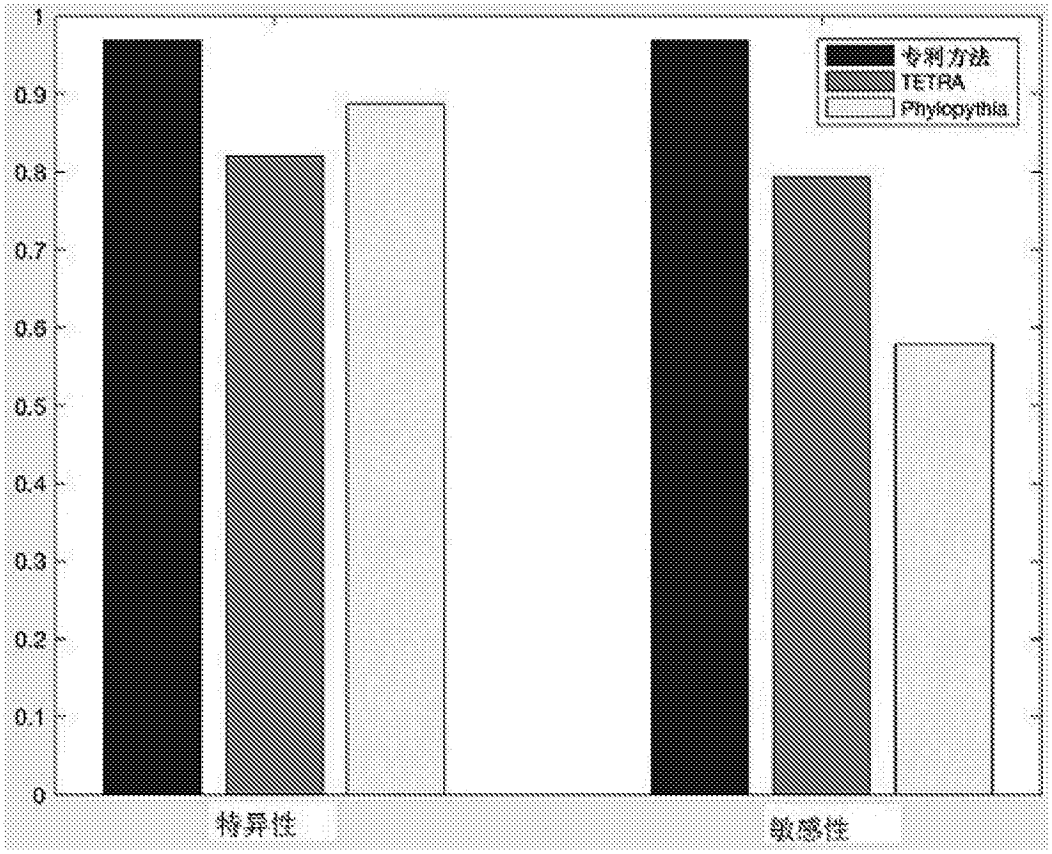


图3