(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 111095232 A
(43)申请公布日 2020.05.01

(21)申请号 201880060638.0

(22)申请日 2018.07.18

(30)优先权数据
15/653,441 2017.07.18 US

(85)PCT国际申请进入国家阶段日
2020.03.18

(86)PCT国际申请的申请数据
PCT/IB2018/000929 2018.07.18

(87)PCT国际申请的公布数据
WO2019/016608 EN 2019.01.24

(71)申请人 生命分析有限公司
地址 加拿大安大略省

(72)发明人 保罗·格鲁希 蒂莫西·伯顿
阿里·侯索斯 阿比那夫·多姆拉
萨尼·古普塔

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
代理人 邬玥 方挺

(51)Int.Cl.
G06F 15/16(2006.01)
G06N 3/12(2006.01)

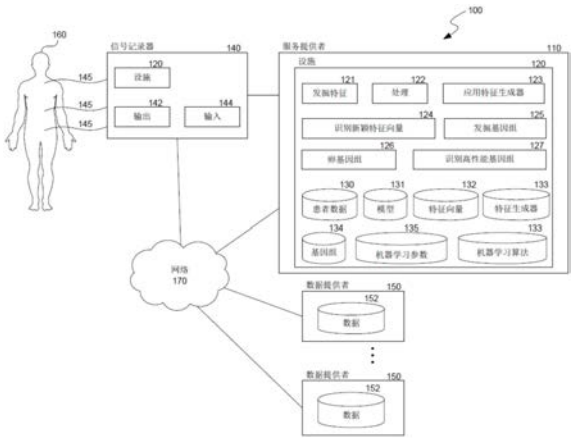
权利要求书4页 说明书15页 附图8页

(54)发明名称

发掘用于机器学习技术中的基因组

(57)摘要

公开了一种用于识别特征和机器学习算法参数的组合的设施,其中每个组合可以与一个或多个机器学习算法组合以训练模型。该设施基于使用该基因组和机器学习算法训练的模型当在应用于验证数据组时生成准确结果的能力来评估每个基因组,例如,通过为训练的模型和相应的用于训练模型的基因组生成匹配或验证分数。选择生成匹配分数超过匹配阈值的基因组进行突变,突变,然后重复该过程。然后,可以将这些训练的模型应用于新数据,以生成对基本主题的预测。



CN 111095232 A

1.一种具有存储器和处理器的系统,用于发掘机器学习基因组,所述系统包括:

第一组件,被配置为生成多个基因组,其中每个基因组识别至少一个特征和用于至少一个机器学习算法的至少一个参数,其中,生成所述多个基因组的第一基因组包括:

从一组特征中随机选择一个或多个所述特征,

从至少一个机器学习算法的一组参数中随机选择一个或多个所述参数,以及

为每个选择的参数分配至少一个随机值;

第二组件,被配置为为每个生成的基因组,

使用生成的基因组训练一个或多个模型,以及

对于使用所述生成的基因组训练的每个模型,

通过将训练的模型应用于验证数据组,至少部分地为训练的模型计算匹配分数,以及

至少部分地基于为使用所述生成的基因组训练的所述模型所生成的匹配分数来生成所生成的基因组的匹配分数;

第三组件,被配置为从所生成的基因组中识别具有超过匹配阈值的匹配分数的多个基因组;和

第四组件,被配置为针对每个识别的基因组,使所述识别的基因组突变,

其中所述组件中的至少一个包括存储在存储器中以供所述系统执行的计算机可执行指令。

2.根据权利要求1所述的系统,还包括:

第五组件,被配置为对于包括第一组特征的第一基因组,至少部分地通过以下方式从所述第一组特征中识别相关的特征:

对于所述第一组特征中的每个特征,

将与所述特征关联的特征生成器应用于数据的训练组,以生成所述特征的特征向量,

对于至少一对特征向量,

计算所述一对特征向量中每个特征向量之间的距离,

确定计算的距离小于距离阈值,

响应于确定所述计算的距离小于距离阈值,从所述第一基因组中去除与所述一对特征向量对中的至少一个特征向量相对应的特征,

其中,对于多个患者中的每个患者,每个特征向量包括通过将第一特征生成器应用于代表所述患者的生理数据的至少一个表示而生成的单个值。

3.根据权利要求2所述的系统,其中,从所述第一基因组中去除与第一对特征向量中的至少一个特征向量相对应的至少一个特征包括:

随机选择所述第一对特征向量的一个特征向量,

从所述第一基因组的特征中识别与随机选择的特征向量相对应的特征;以及

从所述第一个基因组中去除所识别的特征。

4.根据权利要求1所述的系统,还包括:

第五组件,被配置为对于包括第一组特征的所述第一基因组,生成包括用于所述第一组特征的每个特征的顶点的图;

第六组件,被配置为生成顶点之间的边,所述顶点的对应特征具有超过相关阈值的相关值,或者具有小于距离阈值的距离值;和

第七组件，被配置为从图中去除顶点，直到图中没有连接的顶点为止。

5.根据权利要求1所述的系统，还包括：

机器，被配置为从至少一个患者接收生理信号数据；

第五组件，被配置为对于每个患者，

将训练的模型中的至少一个应用于由机器为患者接收的所述生理信号数据的至少一部分，并且

至少部分地基于将所述训练模型中的至少一个应用至所接收的生理信号的至少一部分，为所述患者生成预测。

6.一种由具有存储器和处理器的计算系统执行的用于发掘机器学习基因组的方法，所述方法包括：

由所述处理器生成多个基因组，其中每个基因组识别至少一种特征和用于至少一种机器学习算法的至少一个参数；

对于每个生成的基因组，

使用所述生成的基因组训练至少一个模型，以及

至少部分地基于训练的至少一个模型来生成所述基因组的匹配分数；

从所述生成的基因组中识别至少一个具有超过匹配阈值的匹配分数的基因组；以及

突变每个识别的基因组。

7.根据权利要求6所述的方法，其中，生成所述多个基因组中的第一基因组包括：

从一组特征中随机选择一个或多个特征；

从至少一种机器学习算法的一组参数中随机选择一个或多个所述参数；以及

为每个所选参数分配至少一个值。

8.根据权利要求7的方法，其中生成所述第一基因组进一步包括：

对于随机选择的特征中的每个特征，

至少部分地基于与所述特征相关联的特征生成器和数据的训练组，检索所述特征的特征向量；

从所生成的特征向量中识别相关特征向量对；和

对于每个识别的相关特征向量对，

识别所述相关特征向量对中的一个特征向量，

从所述第一基因组中去除与用于生成所识别的特征向量的特征生成器相关的特征；

从一组特征中随机选择要添加到所述第一个基因组的特征，以及

将随机选择的特征添加到所述第一个基因组。

9.根据权利要求8所述的方法，其中，识别相关特征向量对包括：

对于每对特征向量，

计算特征向量对的距离度量，以及

确定为所述特征向量对计算的距离度量是否小于距离阈值，

其中，至少部分地基于针对每对特征向量确定的所述计算的距离度量来确定所述距离阈值。

10.根据权利要求6所述的方法，其中生成第一基因组的匹配分数包括：

识别通过将使用所述第一个基因组训练的模型应用于两个或多个验证数据组而生成

的多个假阳性;和

识别通过将使用所述第一个基因组训练的模型应用于两个或多个验证数据组而生成的多个假阴性。

11.根据权利要求6所述的方法,其中生成第一基因组的匹配分数包括:

对于至少一个使用所述第一基因组训练的模型,生成接收器工作特性曲线;和

计算生成的接收器工作特性曲线下的面积。

12.根据权利要求6所述的方法,其中,生成第一基因组的匹配分数包括:对于使用所述第一基因组训练的至少一个模型,计算选自以下的组中的误差中的一个或多个:均方预测误差,均值绝对误差,四分位数误差,和对数损失误差,接收器-操作员特征曲线误差,和f分数错误。

13.根据权利要求6所述的方法,其中突变第一识别的基因组包括:

选择所述第一识别的基因组的至少一个特征;以及

从所述第一识别的基因组中去除所述第一识别的基因组的每个选择的特征。

14.根据权利要求6所述的方法,其中突变所述第一识别的基因组还包括:

从一组特征中随机选择多个所述特征;以及

向所述第一识别的基因组中添加随机选择的多个特征中的每个特征。

15.根据权利要求6所述的方法,其中突变第一识别的基因组包括:

修改所述第一识别的基因组的至少一个特征。

16.根据权利要求6所述的方法,其中突变第一识别的基因组包括:

修改所述第一识别的基因组的至少一个机器学习算法参数。

17.一种存储指令的计算机可读介质,如果该指令由具有存储器和处理器的计算系统执行,则使该计算系统执行用于发掘机器学习基因组的方法,所述方法包括:

生成多个基因组,其中每个基因组识别至少一个特征;

对于每个生成的基因组,

使用生成的基因组训练至少一个模型,以及

至少部分地基于训练的至少一个模型来生成所述基因组的匹配分数;和从所述生成的基因组中识别一个或多个具有超过匹配阈值的匹配分数的基因组。

18.根据权利要求17所述的计算机可读介质,其中,每个基因组进一步识别用于至少一种机器学习算法的至少一个参数。

19.根据权利要求17所述的计算机可读介质,所述方法还包括:

突变每个识别的具有超过匹配阈值的匹配分数的基因组。

20.根据权利要求17所述的计算机可读介质,所述方法还包括:

至少部分地通过基于为每个所述生成的基因组生成的匹配分数确定总体匹配分数来计算所述匹配阈值。

21.根据权利要求17所述的计算机可读介质,所述方法还包括:

至少部分地通过基于为每个所述生成的基因组生成的匹配分数确定匹配分数的第n个百分点数来计算所述匹配阈值。

22.根据权利要求17所述的计算机可读介质,所述方法还包括:

至少部分地通过从为每个所述生成的基因组生成的所述匹配分数中确定第n个最高匹

配分数来计算所述匹配阈值。

23.根据权利要求17所述的计算机可读介质,还包括:

对于使用所述第一基因组训练的每个模型,

计算使用所述第一基因组训练的模型的匹配分数;和

合计使用所述第一基因组训练的模型计算的所述匹配分数。

24.根据权利要求23所述的计算机可读介质,其中,合计使用所述第一基因组训练的模型计算的匹配分数包括计算针对使用所述第一基因组训练的模型计算的匹配分数的平均值。

25.根据权利要求17所述的计算机可读介质,所述方法还包括:

对于从所述生成的基因组中识别出的所述多个基因组中的每个,

突变识别的基因组,

使用所述突变的基因组训练至少一个模型,以及

至少部分地基于使用所述突变基因组训练的至少一种模型来生成所述突变的基因组的匹配分数。

26.一个或多个共同存储基因组数据结构的计算机存储器,其中,所述基因组数据结构包括:

多个特征,每个特征识别至少一个特征生成器,以及

用于至少一种机器学习算法的多个参数,

其中,所述基因组数据结构被配置为用于根据所述多个特征和至少一种机器学习算法来训练至少一个模型。

发掘用于机器学习技术中的基因组

[0001] 相关申请的交叉引用

[0002] 本申请要求于2018年7月18日提交的美国专利申请号15/653,441的优先权，在此通过引用整体并入本文。

[0003] 相关应用

[0004] 本申请涉及2013年8月19日提交的标题为“用于表征心血管系统的无创方法和系统”的美国专利申请号13/970,580，现在是美国专利号9,289,150；2016年3月4日提交的题为“用于表征心血管系统的无创方法和系统”的美国专利申请号15/061,090号；2017年5月5日提交的标题为“用于表征心血管系统的无创方法和系统”的美国专利申请号15/588,148；2012年9月6日提交的标题为“用于评估电生理信号的系统和方法”的美国专利申请号13/605,364，现在为美国专利号8,923,958；2013年8月19日提交的标题为“用于表征心血管系统的全因死亡率和突发性心脏病死亡风险的无创方法和系统”的美国专利申请13/970,582，现在为美国专利号9,408,543；2016年7月11日提交的标题为“用于表征心血管系统的全因死亡率和突发性心脏病死亡风险的无创方法和系统”的美国专利申请号15/207,214；2014年6月4日提交的标题为“用于估算哺乳动物心脏腔大小和机械功能的无创心电图方法”的美国专利申请号14/295,615；2013年11月12日提交的标题为“用于估计哺乳动物心脏腔大小和机械功能的无创心电图方法”的美国专利申请号14/077,993；2015年1月14日提交的标题为“用于估计葡萄糖，糖化血红蛋白和其他血液成分的无创方法”的美国专利申请号14/596,541，现为美国专利9,597,021；2017年3月16日提交的标题为“用于估计葡萄糖，糖基化血红蛋白和其他血液成分的无创方法”的美国专利申请号15/460,341；2015年2月12日提交的标题为“用于从单通道数据表征心血管系统的方法和系统”的美国专利申请号14/620,388；2016年6月24日提交的标题为“使用数学分析和机器学习来诊断疾病的方法和系统”的美国专利申请号15/192,639；2016年8月26日提交的标题为“生物信号获取装置”的美国专利申请号15/248,838；2016年9月21日提交的标题为“用于心动相空间层析成像的图形用户界面”的美国临时专利申请号62/397,895；2017年6月26日提交的标题为“用于测量心肌缺血，狭窄识别，局部化和分流储备的无创方法和系统”的美国专利申请号15/633,330；以及与本文同时提交的标题为“发掘可用于机器学习技术的新特征，例如用于诊断医疗状况的机器学习技术”的美国专利申请号15/653,433。上述申请和已授权的专利中的每一个均通过引用整体并入本文。

背景技术

[0005] 机器学习技术基于输入数据组来预测结果。例如，机器学习技术正用于预测天气模式，地质活动，提供医学诊断等。机器学习技术依赖于使用训练数据组(即观测数据组，在其中每个待预测的结果均已知的情况下)生成的一组特征，其中每个特征代表观测数据的某些可测量方面，以生成并调整一个或多个预测模型。例如，可以分析观察到的信号(例如，来自多个对象的心跳信号)以收集频率，平均值，以及关于这些信号的其他统计信息。机器学习技术可以使用这些特征来生成和调整将这些特征与一个或多个状况，例如某种形式的

心血管疾病(CVD),包括冠状动脉疾病(CAD))相关的模型,然后将该模型应用于具有未知结果的数据来源,例如未诊断的患者或未来的天气模式,等等。通常,这些功能是由数据科学家与领域专家一起手动选择和组合的。

[0006] 附图的简要说明

[0007] 图1是示出在一些示例中设施在其中操作的环境的框图。

[0008] 图2是示出在一些示例中的发掘特征部件的处理的流程图。

[0009] 图3是示出一些示例中的处理组件的处理的流程图。

[0010] 图4是示出一些示例中的应用特征生成器组件的处理的流程图。

[0011] 图5是示出了在一些示例中的识别新颖特征向量组件的处理的流程图。

[0012] 图6是示出根据一些示例的发掘基因组组件的处理的流程图。

[0013] 图7是示出根据一些示例的卵(spawn)基因组组件的处理的流程图。

[0014] 图8是示出根据一些示例的识别高性能基因组组件的处理的流程图。

具体实施方式

[0015] 因为机器学习技术依赖于特征和/或特征的组合,所以特征选择和组合的过程通常是机器学习过程的重要组成部分。此外,由于存在大量多样的机器学习算法(例如决策树,人工神经网络(ANN),深度ANN,遗传(和元遗传)算法等),因此算法的选择和任何相关参数也很重要。例如,不同的机器学习算法(或机器学习算法系列)可能最适合于不同类型的数据和/或要进行的预测类型。此外,不同的机器学习算法可以权衡资源(例如,存储器,处理器利用率),速度,准确性等方面。通常,使用基于个体的偏好和/或个体指定的条件由个体选择的机器学习算法,特征,和参数来训练模型。发明人已经认识到,手动识别特征,机器学习算法,和相应的参数可能是昂贵且费时的,更加困难的是生成特征,机器学习算法,和相应的参数以生成更准确的模型,从而得到更准确的预测。因此,发明人已经构思并实践用于执行特征,机器学习算法,和/或机器学习参数的组合的自动发掘的设施。

[0016] 在一些示例中,该设施作为机器学习流水线的一部分操作,该机器学习流水线基于时间序列和/或其他信号,例如,生理信号,来构建和评估预测模型,例如,用于疾病诊断的预测模型。机器学习过程使用特征来识别训练数据组内的模式,并基于这些模式生成预测模型。可以使用验证数据组(即,结果已知但未用于训练模型的数据组)验证这些预测模型,并将其应用于新的输入数据,以便从输入数据预测结果,例如提供对医疗状况的诊断等。当生成或获取新数据和新特征时,机器学习过程通过合并新特征,并在某些情况下舍弃其他特征,例如那些被确定与其他功能过于相似的其他特征,来改善这些模型的预测能力。

[0017] 特别地,该设施寻求识别特征和机器学习算法参数的组合,其中每种组合都可用于训练一个或多个模型。特征和/或机器学习参数的组合在本文中有时被称为“基因组”。该设施基于使用机器学习算法和该基因组训练的模型当将该基因组应用于验证数据组时,以生成准确的结果的能力来评估每个基因组,例如,通过为训练的模型和用于训练模型的基因组生成匹配或验证分数。在某些情况下,该设施将验证分数用作匹配分数,而在其他情况下,验证分数是匹配分数的元素(例如,匹配分数=训练评分+验证评分)。在某些情况下,可以使用基因组来训练多个模型,并且可以合计所得的匹配分数以生成基因组的合计匹配分数。

[0018] 作为例子，用于识别特征和机器学习算法参数的组合的设施可以用于医学诊断预测建模任务。在该例子中，该设施针对多个患者或受试者接收一组或多组生理数据，该一组或多组生理数据与一段时间内(例如，小于一秒，几秒钟，大约十秒钟，大约30秒以及最多大约五分钟，大约一个小时或更长的时间等)的某种生理输出或患者的状况有关，例如脑电图等。这些数据可以与设施的操作实时地或接近实时地或几乎同时地被接收，或者它们可以在更早的时间被接收。在某些情况下，设施会丢弃信号的某些部分，以确保来自每个患者的信号以稳定且一致的初始状态开始。此外，可以将数据归一化以去除潜在的误导性信息。例如，该设施可以归一化信号数据的幅度(例如，变换为z分数(z-score))，以解决由传感器接触或其他非生理数据引起的信号强度的变化。作为另一例子，在心脏信号的情况下，设施可以执行峰值搜索并丢弃信号中识别出的第一个心跳之前和信号中识别出的最后一个心跳之后的任何数据。

[0019] 在一些示例中，该设施将一组特征生成器应用于一组信号，以针对信号和特征生成器的每种组合生成该信号的特征值。因此，每个特征值代表基础信号数据的某些属性。在一个示例中，设施接收1000个患者中的每个患者的患者数据，并将一个或多个特征生成器应用于该数据，以针对特征生成器对单个患者的数据的每次应用生成特征值(或一组特征值)。该设施在“特征向量”中收集由单个特征生成器生成的特征值，以使特征向量为每个患者存储一个特征值。一旦生成了特征向量，就可以将它们进行比较以确定每个特征向量相对于其他特征向量的每一个如何不同。该设施为每个特征向量计算距离度量，以评估相应特征生成器的新颖性。基于评估的新颖性，设施(1)提供特征生成器，这些特征生成器将新颖的特征向量生成到机器学习过程中，以使得新的预测模型以特征生成器为基础；以及(2)修改这些特征生成器以创建新一代的特征生成器。该设施重复此进化过程，以识别甚至更多的新颖功能，供机器学习过程使用。

[0020] 在一些示例中，对于每个接收到的数据组，设施计算或识别来自数据的一个或多个值的分离组。例如，在将数据作为心电图的一部分生成的情况下，设施识别数据中的全局和局部最大值和最小值，从数据中计算频率/周期信息，在特定时间段内计算数据的平均值(例如，在QRS复合波期间生成的平均持续时间和值)，等等。在某些情况下，设施变换接收到的数据并从变换后的数据中提取多组一个或多个值。该设施可以通过多种方式变换接收到的信号数据，例如获取数据的一个或多个(连续)导数，获取数据的一个或多个偏导数，对数据进行积分，计算数据的梯度，对数据应用函数，应用傅立叶变换，应用线性或矩阵变换，生成拓扑度量/特征，生成计算几何度量/功能，生成差分流形度量/特征等。以这种方式，该设施生成数据的多个方面(perspective)，以生成多种特征的不同的组。尽管通过例子的方式提供了这些变换，但是本领域的普通技术人员将认识到可以以多种方式来变换数据。

[0021] 在一个例子中，该设施接收多个输入信号(例如，由连接到患者的不同电极或导线收集的输入信号，多峰信号，例如来自宽带生物电势测量设备导线和S_pO_2(血氧饱和度)通道的信号等等和/或变换后的信号，并通过为每个信号计算采样周期内信号的平均值，从信号数据中提取值。在该例子中，表示了每个患者四个信号，但是本领域的普通技术人员将认识到，可以监视和/或接收任何数量的信号以用于设施的处理和进一步分析。因此，在此例子中，每个患者的提取数据可以表示为这些随时间推移的一组平均值，例如：

[0022]

患者	*A*	*B*	*C*	*D*
1	0.24	0	0	30
2	0.2	0.6	4.2	5
…				
n	.32	2	4	.02

[0023] 表1

[0024] 表1表示n位患者中每位患者的一组平均信号值(A,B,C和D)。尽管此处使用平均值,但是本领域普通技术人员将认识到,可以从基础数据信号中提取或计算任何类型的数据,例如信号超过阈值的时间量,一个信号的值而另一个信号的值超过阈值,依此类推。

[0025] 在一些示例中,在已经从接收到的信号中提取数据之后,设施将一个或多个特征生成器应用于接收到的或生成的数据,诸如提取的数据,原始或预处理的信号数据,变换的数据等等。特征生成器接收信号数据的至少一部分或表示作为输入,并生成相应的输出值(或一组值)(即"特征")。一组特征包括以下等式:

[0026] F1＝A+C-D, (等式1)

[0027] $F2 = \frac{A*S(4)*B}{D} + C + \sqrt{D}$, 以及 (等式 2)

[0028] F3＝S(1)*D, (等式3)

[0029] 其中,A,B,C和D分别表示从特定患者的数据中提取的值,而S(t)表示每个信号在时间t处的信号值。例如,在等式1中,F1代表特征的名称,而等式A+C-D代表相应的特征生成器。在某些情况下,该设施使用复合特征生成器,其中一个特征生成器充当另一个特征生成器的输入,例如:

[0030] $F4 = \frac{F1*F2}{\sqrt[3]{F3}} + .057$ (等式 4)

[0031] 在此示例中,该设施将特征生成器应用于表1中表示的每个患者的提取数据,以为每个特征生成器生成三个值的特征向量(每个患者一个),例如下面的表2中所示:

[0032]

患者	*F1*	*F2*	*F3*
1	-29.76	5.48	905.83
2	-0.6	6.67	9.57
…			
n	4.3	185.74	0.04

[0033] 表2

[0034] 在此例子中,设施已将每个特征生成器F1,F2,和F3应用于表1中所示的提取数据,以为每个特征生成器生成包括每个患者的值的相应特征向量。例如,通过将特征生成器F1

应用于提取的数据而生成的特征向量包括患者1的值-29.76，患者2的值-0.6，等等。因此，对于每个特定特征生成器，每个特征向量基于每个患者的生理数据的至少一部分(即，特征生成器所应用的生理数据中表示的患者)代表相应特征生成器的签名(不一定是唯一的)。在一些示例中，特征生成器使用不同的结构或模型来表达，例如表达树，神经网络等。本领域的普通技术人员将认识到，该设施可以在特征向量的生成中采用任何数量的特征生成器和任何数量的生理学数据组(或其一部分)。在一些示例中，设施随机地选择多个先前生成的特征生成器以用于生成特征向量，而不是采用每个可用的特征生成器。在一些示例中，设施通过例如随机地生成表达树，将权重随机分配给神经网络内的连接等，来创建和/或修改特征生成器。

[0035] 在一些示例中，在设施生成多个特征向量之后，设施使用某种形式的新颖性搜索来识别所生成的特征向量中最“新颖”的特征向量。新颖性对应于特定特征向量与其他特征向量的比较组(由当前迭代过程中设施生成的任何特征向量以及任何早期迭代中选择的特征生成器生成的特征向量组成)之间的差如何；与比较组的特征向量的差越大，新颖性就越大。该设施使用距离的形式作为新颖性的度量(即，每个特征向量与其他特征向量之间的距离多远)。在这种情况下，对于每个生成的特征向量，设施都会计算该特征向量与其他每个生成的特征向量之间的距离，并对生成的距离值进行合计，例如计算特征向量的平均值或均值(例如算术，几何，调和等)距离值，或特征向量与其他每个生成的特征向量之间的总(和)距离，确定特征向量的模式距离值，中值距离值，最大距离值，等等。例如，使用表2的特征向量(针对患者1，2，和n)，可以按以下方式计算每组特征向量的距离：

[0036] $F1$-$F2$ 距离: $\sqrt{(-29.76-5.48)^2+(-0.6-6.67)^2+(4.3-185.74)^2}=184.97.$

[0037] $F1$-$F3$ 距离: $\sqrt{(-29.76-905.83)^2+(-0.6-9.57)^2+(4.3-0.04)^2}=936.23$

[0038] $F2$-$F3$ 距离: $\sqrt{(5.48-905.83)^2+(6.67-9.57)^2+(185.74-0.04)^2}=919.70.$

[0039] 在该例子中，已经计算了每个特征向量之间的总欧几里德距离，作为用于计算两个向量中的每个向量之间的差的手段。除了由特征生成器的当前组(即，当前代)生成的特征向量之外，该设施还包括由较早代的特征生成器生成的特征向量。在一些例子中，设施在比较之前向每个特征向量施加权重，例如随机生成的权重，和/或归一化每组特征向量。因此，此例子中每个特征向量的距离测量如下：

[0040]

特征生成器	到$F1$的距离	到$F2$的距离	到$F3$的距离	平均距离	最大距离
$F1$	–	184.97	936.23	560.60	936.23
$F2$	184.97	–	919.70	552.34	919.70
$F3$	936.23	919.70	–	927.97	936.23

[0041] 表3

[0042] 在该例子中，设施基于所计算的距离来识别最“新颖的”特征向量，其对于每个特

征向量起着“新颖性得分”或“匹配得分”的作用。设施识别与其他向量的平均距离最大的特征向量(例如,由F3生成的特征向量),最大距离最大的特征向量(例如,由F1和F3生成的特征向量)等等。在一些例子中,所识别的新颖特征向量的数量被固定(或封顶)为预定数量,例如5,10,100,500等。在其他例子中,要识别的新颖特征向量的数量是动态确定的,例如,基于新颖性分数分析的特征向量的前10%,具有新颖性分数大于超过所分析特征向量的平均新颖性分数的预定数量的标准偏差的任何特征向量,等等。然后,可以将生成每个已识别的新颖特征向量的特征生成器添加到可用的特征组中,以用作由机器学习流水线构建和评估的模型的输入。这些模型可以应用于患者数据,例如用于诊断,预测,治疗,或其他分析,科学,健康相关或其他目的。

[0043] 在一些示例中,除了提供用于生成所识别的新颖特征向量以供机器学习过程使用的特征生成器之外,该设施还随机地突变或修改用于生成所识别的新颖特征向量的特征生成器。每个突变都会影响相应特征生成器中的某些更改,并创建新版本的特征生成器,可用于为新一代特征生成器做出贡献。该设施使用此新特征生成器生成新特征向量,然后评估新特征向量的新颖性。此外,可以进一步对相应的特征生成器进行突变,以继续进行特征向量和特征代的创建过程。例如,可以通过随机选择等式中的一个或多个元素(例如,随机选择的元素)并将选定的元素替换为其他元素,来使以等式形式表示的特征生成器,例如$F1_0$=A+C-D)发生突变。在此示例中,可以通过用B替换A以创建$F1_1$=B+C-D或用$\sqrt[3]{C-B^2}$替换C-D来创建$F1_1 = B + \sqrt[3]{C-B^2}$来更改等式。在这种情况下,已包括下标0和1以表示每个特征生成器的代标记或计数。换句话说,$F1_0$代表第0代(即第一代)的以上(等式1)的F1,$F1_1$代表了第1代(即第二代)的F1的突变版本,依此类推。在某些情况下,较早的一代(或其变换)作为元素包含在后续的一代中,例如$F2_1 = \sqrt{F2_0} + C^2$或$F2_n = \sqrt{F2_{n-1}} + C^2\ (n \neq 0)$。

[0044] 在一些示例中,设施以不同的方式获得特征。例如,设施可以从诸如领域专家的用户接收用户已经识别为最佳的和/或用户期望被测试的一组特征(和相应的特征生成器)。作为另一例子,可以从一个或多个特征存储器中编辑地选择特征。在某些情况下,可以将设施自动生成的特征与其他特征合并以创建各种混合特征。甚至可以使用出处不明的特征。

[0045] 在一些示例中,该设施识别用于训练模型的基因组,从这些基因组中识别“最佳”(最高评分)的基因组,并对所识别的基因组进行突变以生成甚至更多的可用于训练模型的基因组。在使用基因组训练一个或多个模型之后,该设施将每个训练的模型应用于验证数据组,以便对训练的模型进行评分(例如,训练的模型在基础验证数据组中正确识别和/或分类对象的程度如何)。该设施对生成最佳结果(例如,具有最高的验证或匹配分数)的基因组进行突变,使用这些突变的基因组训练新模型,并重复此过程,直到满足一个或多个终止标准(例如,预定的代的数量,在先前代的预定或动态数量(例如1、5、8、17等)期间不会生成额外的高得分(高于预定或动态生成的阈值)的基因组),其组合等)。

[0046] 在一些示例中,该设施使用先前识别或生成的基因组作为第一组基因组(即第一代),从中发掘用于机器学习算法的基因组。在其他例子中,该设施通过以下方式自动生成第一代基因组:对于每个基因组,从一个或多个先前生成的一组特征向量(例如,由将特征生成器应用于一组训练数据生成的特征向量)随机(具有或不具有置换)选择一个或多个特征向量。基因组还可以包括机器学习算法的一个或多个机器学习算法参数,例如预测器的

数量(例如,回归器,分类器,用于机器学习算法的决策树的数量和/或最大数量,等)用于与该算法关联的基础全体方法,机器学习算法的最大深度(例如,决策树的最大深度)等等。在基因组被配置为与一种特定的机器学习算法一起使用的情况下,基因组可以被配置为为与该机器学习算法相关联的每个机器学习参数定义值。在其他情况下,基因组的元素之一在不同的机器学习算法中进行选择,并且可以进行突变,以便将基因组及其相应的参数值与不同的机器学习算法一起使用,以在进化过程中训练模型。例如,在第一代中,基因组可以识别依赖决策树的机器学习算法,而同一基因组的突变版本可以识别使用一个或多个支持向量机,线性模型等的机器学习算法。在这种情况下,基因组可以为每个可以与基因组结合以训练模型的机器学习算法指定建模参数。因此,单个基因组可以包括用于多种机器学习算法的机器学习参数。然而,基因组不需要包括用于相应机器学习算法的每个建模参数。如果要使用特定的机器学习算法和不包含该机器学习算法的机器学习参数值的基因组来训练模型,则设施可以从例如机器学习参数存储器检索这些参数的默认值。

[0047] 例如,一组基因组可以表示为:

[0048]

$G1_1$		MLA=4	F23	F78798	F32	F55	F453	F234
$G2_1$		MLA=9	F9701	F223	F1	F63	F349	P9:1=7
$G3_1$		MLA=2	F823	F525	F732	F525	F125	
$G4_1$		MLA=6	F597	F135	F404	F31	P6:1=5	P6:2=150
	…							
$G20_1$		MLA=1	F43	F65	P1:1=8	P1:2=218	P1:3=0.3	

[0049] 表4

[0050] 其中每一行对应于第一代选定或生成的基因组中的不同的基因组(在左侧的第一列中命名),并识别了用于使用基因组训练模型的机器学习算法("MLA";左侧的第二列),例如进入机器学习算法存储器的索引。例如,基因组$G3_1$在机器学习算法存储器中指定与索引2对应的机器学习算法(MLA=2)。在此例子中,每个非阴影区域(在第二列的右侧)标识不同的特征。基因组还可以包括相应的特征生成器或对相应特征生成器的引用,例如到特征生成器存储器的链接。如上所述,这些特征可以由设施自动生成和/或从另一源检索。

[0051] 此外,表4中的每个阴影区域代表特定机器学习参数的值。在该例子基因组集中,机器学习参数由指示符或参考(例如,P6:1)表示,后跟等号和相应的值。例如,机器学习算法参数P6:1在基因组$G20_1$中的对应值为8。在此示例基因组集中,每个机器学习参数均以二维数组中的索引的形式呈现,例如"P6:1"代表"第六个"机器学习算法的"第一个"机器学习参数(即具有索引为6的机器学习算法且具有索引为1的机器学习参数)。如上所述,基因组可以为任何或所有机器学习参数指定值,这些参数可以用于使用基因组(或该基因组的突变版本)训练模型。而且,从表4可以清楚地看出,基因组的长度可以变化。例如,基因组$G1_1$包括六个特征的值和零个机器学习参数,而基因组$G2_1$包括两个特征的值和三个机器学习参数。因此,该设施可以在机器学习过程中采用可变长度的基因组。

[0052] 在一些示例中,该设施可以从基因组内过滤特征和/或过滤基因组本身以避免每个基因组之间的冗余。为了过滤特征和/或基因组,该设施为每对生成相关值,并丢弃该对中的一项。为了从基因组中识别和过滤相关特征,该设施通过将与特征相关联的特征生成

器应用于一组训练数据以生成一组值，为每个特征生成特征向量。该设施将每个生成的特征向量与其他生成的特征向量进行比较，以确定是否有任何特征向量是“高度”相关的(即在所选的特征向量集中不“新颖”)。例如，组件可以计算相对于其他特征向量的每个所生成的特征向量的距离值(如以上关于识别新颖特征生成器所讨论的)，并且如果任意一对(两个的集合)之间的距离小于或等于距离阈值(即“高度”相关或不“新颖”)，则丢弃对应于一对特征向量之一的特征。此外，该设施可以用诸如随机选择的特征之类的新特征代替丢弃的特征。类似地，该设施可以通过为基因组的每个特征生成特征向量，基于所生成的特征向量计算每对(两个的集合)基因组的距离度量，并识别计算的距离不超过基因组距离阈值的基因组对，来识别和丢弃冗余基因组。对于每个识别出的基因组对，该设施可以丢弃或突变一个或两个基因组，以减少一组基因组之间的相关性和冗余度。尽管在该例子中将距离用作确定两个向量或一组向量之间的相关性的度量，但是本领域的普通技术人员将认识到，可以以其他方式来计算两个或一组向量之间的相关性，例如归一化交叉-相关性，等等。在一些示例中，该设施可以采用额外或其他技术来过滤基因组，例如生成图，其中特征表示图中的顶点，该顶点通过图中的边连接。例如，如果两个特征之间的相关值超过预定的相关阈值和/或两个特征之间的距离小于预定的距离阈值，则生成两个特征之间的边。生成图后，该设施将从图形中删除连接的顶点(特征)，直到图中没有边(在删除连接的顶点时边被删除)为止，然后选择其余未连接的顶点(特征)包含在“过滤的”基因组中。在某些情况下，设施可能会随机选择要删除的连接顶点。此外，该设施可以针对一组顶点(特征)多次执行该过程，然后选择优选的“过滤的”基因组，例如去除了最多或最少顶点(特征)的基因组。

[0053] 为了测试每个基因组的匹配或有效性，该设施使用该基因组的特征，机器学习参数，和/或机器学习算法来训练至少一个模型。例如，该设施可以使用AdaBoost(“自适应增强”)技术来使用相应的特征，机器学习参数，机器学习算法，和数据的训练组来训练模型。然而，本领域普通技术人员将认识到，可以使用许多不同的技术来训练给定基因组或一组基因组的一个或多个模型。在训练模型之后，设施将训练的模型应用于一组或多组验证数据，以评估训练的模型识别和/或分类在验证数据组中的先前确定或分类的对象的程度如何。例如，可以生成基因组以训练模型以识别数据组中表示的可能患有糖尿病的患者。一旦使用这些基因组之一对模型进行了训练，就可以将训练的模型应用于验证数据组，以确定验证分数，该分数反映训练的模型从验证组中识别出已知或现在患有糖尿病的患者的程度如何；为每个正确的确定(例如，真阳性和真阴性)得分(加)一个“点”，为每个错误的确定(例如，假阳性和假阴性)失去(减去)一个“点”。因此，可以基于当将训练的模型应用于一组或多组验证数据时基于训练的模型得分的多少“点”来确定训练的模型的总体得分。本领域普通技术人员将认识到，可以使用多种技术来为训练的模型生成匹配分数，例如计算相应的接收器工作特性(ROC)曲线下方的面积，计算均方预测误差，f得分，敏感性，特异性，阴性和阳性预测值，诊断比值比等。在该例子中，在使用基因组训练单个机器学习算法的情况下，所生成的匹配分数可以类似地归因于基因组。在其他情况下，基因组可用于训练多个机器学习算法，并且那些训练的机器学习算法中的每一个可应用于多个验证组，以针对用于训练机器算法的每个基因组生成多个匹配分数。在这些情况下，该设施通过合计为使用基因组训练的机器学习算法生成的每个匹配分数，来生成相应基因组的匹配分数。在某些情况下，可以在合计之前合计和/或过滤所生成的匹配分数。

[0054] 在一些示例中，在设施已经为每个基因组生成了匹配分数之后，设施基于这些匹配分数识别“最佳的”基因组。例如，该设施可以基于所生成的匹配分数来建立匹配阈值，并且将“最佳”基因组识别为其得到的匹配分数超过匹配阈值的那些基因组。可以以多种方式来生成或确定匹配阈值，诸如从用户接收匹配阈值，基于匹配分数的集合计算匹配阈值(例如，平均值，平均值加15％，前15，前n个百分点(其中n由用户提供或由设施自动生成)，依此类推。然后设施将每个基因组与其对应的匹配分数相关联地存储，并选择被标识为“最佳”的基因组用于突变(即，匹配分数超过匹配阈值的基因组)。

[0055] 在一些示例中，该设施通过添加，去除，或改变基因组的任何一个或多个特征向量或机器学习参数来突变基因组。例如，下表5代表了以上表4中代表的基因组的许多突变。

[0056]

$G1_2$	MLA=5	F23	F78798	F32	~~F55~~	~~F453~~	~~F234~~	
$G2_2$	MLA=9	~~F9701~~	F223	F1	F63	F349	F584	P9:1=12
$G4_2$	MLA=6	F597	F135	F404	F31	F24	F982	P6:1=5	P6:2=150
…								
$G20_2$	MLA=1	F43	F65*F14	P1:1=8	P1:2=218	P1:3=0.3		

[0057] 表5

[0058] 在该示例中，每一行对应于从第二代选择用于突变的基因组中的不同的基因组(在左侧的第一列中命名)。在此例子中，基于其低匹配分数，该设施未选择基因组$G3_1$进行突变，因此，表5不包括该基因组突变版本的相应条目。此外，通过删除三个特征向量(以删除线表示)并将参考机器学习算法索引从4更改为5，已对基因组$G1_1$进行了突变(表示为$G1_2$)。此外，该设施通过以下对基因组$G2_1$进行了突变：1)删除了特征向量F9701，2)添加特征向量F584，以及3)将机器学习参数$P9_1$从7调整为12；通过添加特征F24和F982获得基因组$G4_1$；通过将F65生成的值乘以F14生成的值来获得基因组Gn_1。然后，这些突变的基因组可用于训练一种或多种机器学习算法，通过将训练的机器学习算法应用于一个或多个验证数据组进行评分(scored)，选择用于突变，突变等。设施执行该过程，直到到达终点为止，例如当已经生成预定数目的代(例如6，30，100，000等)时，等等。

[0059] 图1是说明根据所公开技术的一些示例的设施在其中操作的环境100的框图。在该例子中，环境100包括服务提供者110，信号记录器140(例如，宽带生物电势测量设备)，数据提供者150，患者160，和网络160。在该例子中，服务提供者包括设施120，设施120包括发掘组件121，处理组件122，应用特征生成器组件123，识别新颖特征向量组件124，发掘基因组组件125，卵基因组组件126，识别高性能基因组组件127，患者数据存储器130，模型存储器131，特征向量存储器132，和特征生成器存储器133。设施调用发掘特征组件121以基于接收到的数据来识别和突变特征生成器。发掘特征组件121调用处理组件122，以处理和变换患者信号数据，例如来自信号记录器140(例如，一个或多个用于收集基础数据的测量设备和/或系统，例如宽带生物电势测量设备，等)的原始信号数据，3-D图像数据等。应用特征生成器组件123由发掘特征组件调用，以将一组一个或多个特征生成器应用于已处理和变换后的患者信号数据。由发掘特征组件调用识别新颖特征向量组件124，以从例如由一个或多个特征生成器生成的一组特征向量中识别出最新颖的特征向量。设施120调用发掘基因组组

件125以生成,分析,和突变基因组,以供机器学习算法使用。卵基因组组件126由发掘基因组组件调用以生成包含任何数量的特征向量和/或机器学习参数的基因组。识别高性能基因组组件127由发掘基因组组件调用,以从一组基因组中识别具有超过匹配阈值的相应匹配分数的基因组。患者数据存储器130包括生理患者数据,例如原始生理数据(包括但不限于通过例如信号记录器140获得的数据),变换的生理数据,个人(biographical)信息,人口统计信息等。这些数据可以匿名存储以保护每个相应患者的隐私,并且可以被处理和加密为确保其传输和存储符合任何管辖法律及其实施法规,例如1996年美国健康保险可移植性和责任法案(经修订),欧洲数据保护指令,加拿大个人信息保护和电子文档法案,1988年的澳大利亚隐私法,2015年的日本个人信息保护法(经修订),州和省法律法规等。模型存储器131存储有关通过将机器学习技术应用于训练数据而生成的模型的信息,例如Christopher M.Bishop在《模式识别和机器学习(2006)》(国会图书馆控制编号:2006922522;ISBN-10:0-387-31073-8)中描述的机器学习技术,其全部内容通过引用合并于此。特征向量存储器132存储通过将一个或多个特征生成器应用于一组生理数据而生成的特征向量组。特征生成器存储器133存储可以应用于患者生理数据并且可以包括特征生成器的多代的特征生成器的组。基因组存储器134存储由设施和/或其他来源生成的和/或突变的基因组。机器学习参数存储器135对于许多机器学习算法中的每一个,存储可以用作该机器学习算法的输入的一组参数以及与该参数有关的其他信息,例如,对应参数的最大值,对应参数的最小值,对应参数的默认值等。机器学习算法存储器133存储用于多个机器学习算法中的每一个的逻辑,每个机器学习算法的逻辑可以由设施选择性地训练和验证。在该例子中,信号记录器140经由电极145连接至患者160,并且包括设施120,一个或多个输出装置142,例如显示器,打印机,扬声器等,以及一个或多个输入装置144,例如,设置控件,键盘,生物特征数据读取器等。因此,如本例子所示,该设备可以配置为从患者和其他诊断设备远程操作和/或与诸如宽带生物电势测量设备(即配置为捕获未经过滤的电生理信号的任何装置,包括那些频谱成分未改变的信号)的诊断设备一起使用或作为诊断设备的一部分。因此,该设施可以被配置为在读取生理数据时实时操作和/或可以被应用于先前记录的生理数据。数据提供者150,每个数据提供者包括数据存储器152,可以提供信息以供设施分析或使用,例如工作场所之外的记录的生理患者数据(例如,在无法访问房屋设施的医院或诊所,第三方数据提供者等),在其他地方生成或生成的特征向量和/或特征生成器,等等。网络170表示通信链路,环境100的多种元件可以通过该通信链路进行通信,例如互联网,局域网等。

[0060] 在多种示例中,这些计算机系统和其他装置可以包括服务器计算机系统,台式计算机系统,膝上型计算机系统,上网本,平板电脑,移动电话,个人数字助理,电视,照相机,汽车计算机,电子媒体播放器,电器,可穿戴装置,其他硬件和/或类似物。在一些示例中,该设施可以在专用计算系统上运行,例如宽带生物电势测量设备(或配置为捕获未过滤的电生理信号,包括具有不变频谱成分的电生理信号的任何装置),脑电图设备,放射学设备,声音录音设备,等等。在多种示例中,计算机系统和装置包括以下一个或多个:被配置为执行计算机程序的中央处理单元("CPU");计算机存储器,其被配置为当程序和数据被使用时,存储程序和数据,包括正在测试的多线程程序,调试器,设施,包括内核的操作系统,以及设备驱动器;持久性存储装置,例如配置为持久性存储程序和数据(例如,固件等)的硬盘驱动器或闪存驱动器;计算机可读存储介质驱动器,例如软盘,闪存,CD-ROM,或DVD驱动器,配置

为读取存储在计算机可读存储介质，例如软盘，闪存设备，CD-ROM，或DVD中的程序和数据；以及配置为将计算机系统连接到其他计算机系统以发送和/或接收数据的网络连接，例如通过互联网，局域网(LAN)，广域网(WAN)，点对点拨号连接，手机网络，或其他网络及在多种示例中，包括路由器，交换机，和多种类型的发射器，接收器，或计算机可读传输介质的其他网络的网络硬件。尽管可以将如上所述配置的计算机系统用于支持设施的操作，但是本领域技术人员将容易认识到，可以使用多种类型和配置并且具有多种组件的装置来实现该设施。可以在由一个或多个计算机或其他装置执行的计算机可执行指令，例如程序模块，的一般情境中描述设施的元件。通常，程序模块包括被配置为执行特定任务或实现特定抽象数据类型并且可以被加密的例程，程序，对象，组件，数据结构，和/或类似物。此外，在多种例子中，可以根据需要组合或分布程序模块的功能。此外，显示页面可以以多种方式中的任何一种来实现，例如以C++或以XML(可扩展标记语言)，HTML(超文本标记语言)，JavaScript，AJAX(异步JavaScript和XML)技术中的网页，或创建可显示数据的任何脚本或方法，例如无线访问协议(WAP)来实现。典型地，程序模块的功能可以在多种示例中根据需要进行组合或分布，包括基于云的实现，Web应用，用于移动装置的移动应用，等。

[0061] 以下讨论提供了可以在其中实现所公开的技术的合适的计算环境的简要，一般的描述。尽管不是必需的，但是在计算机可执行指令的一般情境中描述了所公开技术的多个方面，例如，由通用数据处理装置执行的例程，诸如服务器计算机，无线设备，或个人计算机。相关领域的技术人员将理解，可以用其他通信，数据处理，或计算机系统配置来实践所公开技术的多个方面，包括：互联网或具有其他网络功能的电器，手持式装置(包括个人数字助理(PDA))，可穿戴计算机(例如，面向健身的可穿戴计算装置)，各种形式的蜂窝电话或移动电话(包括在IP上的语音(VoIP)电话)，非智能终端(dumb terminal)，媒体播放器，游戏装置，多处理器系统，基于微处理器或可编程的消费类电子产品，机顶盒，网络PC，小型计算机，大型计算机等。实际上，术语“计算机”，“服务器”，“主机”，“主机系统”等在本文中通常可互换使用，并且是指任何上述装置和系统，以及任何数据处理器。

[0062] 所公开技术的多个方面可以体现在专用计算机或数据处理器中，例如专用集成电路(ASIC)，现场可编程门阵列(FPGA)，图形处理单元(GPU)，多核处理器，等等，它们被特别地编程，配置，或构造为执行在此详细解释的一个或多个计算机可执行指令。尽管所公开的技术的某些方面，例如某些功能，被描述为仅在单个设备上执行，但是所公开的技术也可以在分布式计算环境中实践，在分布式计算环境中功能或模块在不同的处理设备之间共享，这些功能或模块通过通信网络，例如局域网(LAN)，广域网(WAN)或互联网连接。在分布式计算环境中，程序模块可以位于本地和远程存储装置备中。

[0063] 所公开技术的多个方面可以被存储或分布在有形计算机可读介质上，该有形计算机可读介质包括磁性或光学可读计算机磁盘，硬接线或预编程的芯片(例如，EEPROM半导体芯片)，纳米技术存储器，生物存储器，或其他计算机可读存储介质。可替代地，在所公开技术的方面下的计算机实现的指令，数据结构，屏幕显示，和其他数据可以在一段时间内在互联网上或在其他网络(包括无线网络)上，在传播介质(例如，电磁波，声波等)上的传播信号上分布，或它们也可以在任何模拟或数字网络(分组交换，电路交换，或其他方案)上被提供。此外，术语计算机可读存储介质不包括信号(例如，传播信号)或瞬态介质。

[0064] 图2是示出了根据所公开技术的一些示例的发掘特征组件121的处理的流程图。设

施调用发掘特征组件以基于所选患者数据来识别新颖特征向量。在框205中，该组件接收生理信号数据，例如从信号记录器直接接收的原始信号数据，从另一设备或站点先前生成的生理信号等。存在几种用于从患者收集和分析生理信号(例如，电生理信号，生物信号)的技术用于诊断和其他目的，包括，例如，活动跟踪器，超声心动图，宽带生物电势测量设备，脑电图，肌电图，眼电图，皮肤电反应，心率监测器，磁共振成像，脑磁图，肌力图，可穿戴设备技术装置(例如FITBIT)等。虽然这些系统提供的数据有助于识别医疗问题和诊断医疗状况，但它们通常只是诊断过程的起点。此外，鉴于大多数此类系统的特定性质，通常会对其进行分析的数据进行过度过滤，以降低系统本身或技术人员，医师，或其他医疗保健提供者的复杂性(在这种情况下，以降低视觉复杂性等)，从而消除可能具有未开发诊断价值的数据。在框210中，该组件调用过程信号数据组件以处理和变换接收到的信号数据，这可以生成多组数据和变换后的数据。在框215中，该组件将生成值设置为等于0。在框220中，该组件通过例如随机生成表达式树，随机生成神经网络的一组权重，随机突变一组先前生成的特征生成器中的一个或多个，依此类推，来生成一个或多个特征生成器。在框225中，该组件调用应用特征生成器组件以将所生成的特征生成器应用于一组或多组处理信号数据以生成一组特征向量。在框230中，组件调用识别新颖特征向量组件以从特征生成器所生成的一组特征向量中识别出最新颖的特征向量。在框235中，组件将生成所识别的特征向量的特征生成器存储在例如特征生成器存储器中。在框240中，该组件增加代变量。在判定框245中，如果代变量大于或等于代阈值，则该组件完成，否则该组件在框250处继续。该组件还可以使用其他停止条件，例如不会生成至少阈值数量的新颖特征向量的特征生成器的几个代(a number of generations)。在框250中，组件复制并突变所识别的特征生成器，然后循环回到框225，以将经突变的特征生成器应用于一组或多组经处理的信号数据。如上所述，组件可以将任何一种或多种类型的突变应用于特征生成器，例如将多点突变和/或随机重组应用于一个或多个表达树，随机生成神经网络的一组连接权重，等等。

[0065] 图3是示出了根据所公开技术的一些示例的处理组件122的处理的流程图。处理组件由发掘特征调用以处理和变换患者信号数据。在框305至365中，该组件循环遍历一组接收信号(或数据组的组)中的每个信号(或数据组)，每个信号代表从患者接收的生理数据。在框310中，该组件对接收到的信号进行预处理，例如对信号应用一个或多个信号滤波器，对数据执行峰值搜索并丢弃无关信息，对接收信号进行下采样，对接收信号进行上采样，对接收信号进行子采样，将模拟信号变换为数字信号，将图像数据变换为信号数据等。在框315中，组件将预处理的信号存储在例如患者数据存储器中。信号数据可以匿名存储(即，没有显式或隐式地识别相应的患者等)。然而，与同一患者相关联的信号数据的不同实例可以与匿名的唯一标识符相关联，使得来自单个患者的多个信号可以结合用于训练和诊断目的。在框320中，该组件从所存储的信号数据中提取一个或多个值。在框325中，组件存储一个或多个提取的值。在框330中，该组件识别要应用于信号的任何变换。例如，设施可以存储对一组变换或变换函数(例如，傅里叶变换，应用于信号的函数，导数，偏导数等)的指示，以应用于特定信号。作为另一个例子，该设施可以从变换目录中随机选择一个或多个变换以应用于信号数据。在框335至360中，该组件循环遍历每个变换并将该变换应用于信号数据。在框340中，组件将变换应用于信号(例如，相对于特定变量计算三阶导数，计算通过将一个函数应用于信号数据而生成的复合函数的结果(即，表示信号的函数)等)。在框345中，该组

件将变换后的信号数据存储在例如患者数据存储器中;在框350中,该组件从变换后的信号数据中提取一个或多个值;在框355中,该组件存储一个或多个提取的值。在框360中,如果有任何识别的变换要应用,则该组件选择下一个变换并循环回到框335,以将该变换应用于信号数据,否则该组件在框365继续进行。在框365,如果有任何信号尚待分析,则组件选择下一个信号并循环回到框305以处理下一个信号,否则组件完成。

[0066] 图4是示出了根据所公开技术的一些示例的应用特征生成器组件123的处理的流程图。发掘特征组件121调用应用特征生成器组件,以将一组一个或多个特征生成器应用于信号数据,例如预处理和变换后的信号数据,建模信号数据等。在框410至470中,该组件循环遍历接收到的每个特征生成器组,并将特征生成器应用于接收到的一组信号数据中的每个信号。例如,所接收的信号数据可以包括用于多个患者中的每个患者的多个信号数据组,该数据的多个变换等等。在框420至450中,组件循环遍历每个信号以将特征生成器应用于信号数据。在框430中,组件将当前选择的特征生成器应用于当前选择的信号数据。例如,组件可以将特征生成器应用于当前选择的数据信号的预处理版本和该数据的任何变换版本中的每一个。作为另一个例子,组件将由建模信号数据生成的系数“插入”或替换为具有一组变量的特征生成器,以生成输出特征值。作为另一例子,该组件可以将建模信号数据的一个或多个元素应用于神经网络以生成输出特征值。在框440中,组件存储输出值。在框450中,如果存在任何待分析的信号,则该组件选择下一个信号并循环回到框420以处理下一个信号,否则该组件在框460处继续。在框460中,该组件生成包含每个生成的特征值的特征向量,并将与特征生成器关联的特征向量存储在例如特征向量存储器中。例如,特征向量可以包括特征阵列以及到相应特征生成器的链接或标识符。组件还可以将特征向量与用于生成特征向量的信号数据相关联。在框470中,如果有任何特征生成器尚待处理,则该组件选择下一个特征生成器并循环回到框410以处理该特征生成器,否则该组件完成。

[0067] 图5是示出了根据所公开技术的一些示例的识别新颖特征向量组件124的处理的流程图。在该例子中,设施接收一组特征向量,并且对于每个特征向量,接收与相应特征生成器有关的信息,例如特征生成器的标识符。在框505中,组件收集特征向量的比较组,该比较组包括例如由发现为新颖的较早一代的特征生成器生成的特征向量和由当前一代的特征向量生成的特征向量。例如,组件可以从特征存储器中随机选择一组新颖的特征向量。在某些情况下,检索特征向量的请求包括要检索的每个特征向量的特征值数量的上限和下限,例如不小于50(下限阈值)和不大于5000(上限阈值)。在框510至540中,组件循环遍历当前一代特征生成器的每个特征向量,以确定它们的每个相应特征向量与特征向量的比较组的每个特征向量有多不同。在框515至530中,组件循环遍历特征向量的比较组的每个特征向量,以将每个特征向量与当前选择的特征生成器的特征向量进行比较。在框520中,组件计算比较组的当前选择的特征向量与当前选择的特征生成器的特征向量之间的差值。例如,组件可以计算每个特征向量之间的距离值。在框525中,组件存储计算出的差值。在框530中,如果存在任何要比较的特征向量,则组件选择下一个特征向量并循环回到框515以处理特征向量,否则该组件在框535处继续。在框535中,该组件基于所存储的差值,例如平均或最大距离计算当前选定特征生成器的新颖性分数,并将该新颖性分数与特征生成器相关联地存储(例如,在特征生成器存储器中)。在框540中,如果有任何特征生成器尚待评估,则该组件选择下一个特征生成器,并循环回到框515以处理特征生成器,否则该组件在框

545继续。在框545至560中，组件根据计算出的新颖性分数测试每个特征向量是否新颖，并识别任何相应的特征生成器。在判定框550中，如果当前选择的特征生成器的新颖性分数大于新颖性阈值，则该组件在框555继续，否则该组件在框560继续。可以以任意数量的方式生成或确定新颖性阈值，例如从用户处接收新颖性阈值，基于一组新颖性分数(例如，平均值，平均值加25％，前n个(其中n由用户提供或由设施自动生成)，前十个百分点)，计算新颖性阈值，等等。以这种方式，新颖性阈值可以基于例如没有超过当前新颖性阈值的新特征生成器的代的数量动态地变化(例如，一代又一代)，以确保设施正在生成和测试新特征生成器和相应的特征。在框555中，组件将当前选择的特征向量识别为新颖的。在框560中，如果有任何特征向量尚待处理，则组件选择下一个特征向量，并循环回到框545以处理特征向量，否则该组件完成。

[0068] 图6是示出了根据所公开技术的一些示例的发掘基因组组件126的处理的流程图。该设施调用发掘基因组组件来生成和分析基因组，以供机器学习算法使用。在框610，设施初始化代变量等于0。在框620，组件基于例如用户输入，系统参数，或随机地确定要生成的基因组数量(n)。在框630中，该组件n次调用卵基因组组件以生成(spawn)适当数量的卵基因组。在框640中，该组件调用识别高性能基因组组件，以从卵(spawned)基因组中识别具有超过匹配阈值的匹配分数的基因组。在框650中，该组件增加代变量。在判定框660中，如果代变量大于或等于代阈值，则该组件的处理完成，否则该组件在框670继续。在框670中，该组件突变高性能基因组，然后循环返回步骤640，以从突变的基因组(例如，具有超过匹配阈值的匹配分数的突变的基因组)中识别出高性能的基因组。组件可通过添加，改变，或去除(或其任何组合)可变长度基因组的一个或多个元素来突变基因组。例如，该组件可以通过将一个特征替换为另一特征并将新特征添加到突变的基因组来突变一个基因组。在另一个例子中，组件可以选择新的机器学习算法来与基因组关联。在这种情况下，该组件还可以删除或突变任何不相关的机器学习算法参数和/或用新选择的机器学习算法的机器学习参数值替换它们。作为另一个例子，基因组可以通过随机选择多个基因组的元素并结合这些元素以形成新的基因组而使用有性生殖技术作为突变的一种形式。此外，可以配置基因组的一个或多个元素，使得它们在本文所述的进化过程中保持固定(即，不改变)。

[0069] 图7是示出了根据所公开技术的一些示例的卵基因组组件126的处理的流程图。发掘基因组组件125调用卵基因组组件以生成识别任何数量的特征，机器学习参数，和/或机器学习算法的基因组。在框710中，组件识别一组可用特征，例如在一个或多个特征生成器存储器中引用的特征。在框720中，组件确定要包括在待生成的基因组中的特征的数量。例如，组件可以基于用户输入，系统参数，或随机地确定要包括在待生成的基因组中的特征的数量。在框730中，组件从识别出的特征中随机选择确定数量的特征。在框740中，该组件用随机选择的特征替换所选特征中的相关特征。在框750中，组件识别一组可用的机器学习参数。例如，组件可以针对设施可用的每种机器学习算法，识别与该机器学习算法相关联的一组参数，这些参数可以存储在该组件可用的列表或其他数据结构中(例如，机器学习参数存储器)。在某些情况下，可以为单个机器学习算法(或固定的一组机器学习算法)生成基因组。在这种情况下，组件可以仅识别与单个或固定的一组机器学习算法相关联的机器学习参数(或其适当子集)。在其他情况下，基因组可以包括识别机器学习算法并且可以被突变的元素。在这种情况下，组件可以识别在此突变范围内的任何或所有机器学习算法的机器

学习参数(即,与基因组及其后代关联的机器学习算法的参数,以在本文所述的进化过程期间训练模型)。在框760中,组件确定要包括在待生成的基因组中的机器学习参数的数量。例如,组件可以基于用户输入,系统参数,或随机地确定要包括在待生成的基因组中的机器学习参数的数量。例如,一个基因组可以包括与特定机器学习算法或一组机器学习算法关联的每个机器学习参数,而另一个基因组仅包括与特定机器学习算法相关联的机器学习参数的适当子集。在框770中,组件从识别出的机器学习参数中随机选择确定数量的机器学习参数,并基于任何相关联的约束,例如在最小值和最大值之间随机选择一个与参数关联的值,将值分配给该参数。在框780中,组件将每个选择的特征和机器学习参数存储在基因组数据结构中,然后返回基因组数据结构。

[0070] 图8是示出了根据所公开技术的一些示例的识别高性能基因组组件127的处理的流程图。发掘基因组组件调用识别高性能基因组组件,以从一组基因组中识别具有超过匹配阈值的相应匹配分数的基因组(即"高性能"基因组)。在框810至850中,该组件循环通过提供给该组件的一组基因组,例如第一代基因组,突变的基因组,或其某种组合。在框820中,组件使用当前选择的基因组训练一个或多个模型,包括其特征,机器学习参数,和任何指定的机器学习算法。如果基因组的机器学习参数与用于训练模型的机器学习算法不相关联,则可以忽略该机器学习参数。类似地,如果特定的机器学习算法要求在当前选择的基因组中不包括的特定机器学习参数作为输入,则设施(或机器学习算法本身)可以提供从例如机器学习参数存储器中检索到的默认值。在框830中,该组件例如通过将训练的模型应用于一组验证数据并评估训练的模型从验证数据中正确识别或分类对象的能力,来生成用于当前选择的基因组的验证或匹配分数。在框840中,该组件将针对训练的模型所生成的分数和/或其生成的合计与当前选择的基因组相关联地存储。在框850中,如果尚有任何基因组需要评分,则该组件选择下一个基因组并循环回到框810,否则该组件在框860处继续。在框860至890中,该组件评估为每个基因组生成的分数,并选择"最佳"基因组进行突变。在这个例子中,"最佳"基因组是那些生成超过匹配阈值的验证或匹配分数的基因组。在判定框870中,如果为当前选择的基因组生成的分数超过匹配阈值,则该组件在框880继续,否则该组件在框890继续。在框880,该组件标记当前选择的基因组用于突变。在一些示例中,该组件可以基于除匹配分数以外,或额外于匹配分数的标准选择用于突变的基因组。例如,该组件可以使用新颖性分数或其他分数来选择用于突变的基因组。在某些情况下,该组件可以采用比赛选择(tournament selection)过程,其中从种群中随机选择多个基因组,并选择该"比赛"中得分最高的基因组进行复制(reproduction)。在此例子中,如果仅低分基因组出现在比赛中,则将选择低分基因组进行复制。在框890中,如果有任何基因组尚待处理,则该组件选择下一个基因组并循环回到框860,否则该组件返回标记的基因组,并且该组件的处理完成。

[0071] 根据前述内容,将理解的是,出于说明的目的,本文已经描述了所公开的技术的特定示例,但是在不背离所公开的技术的范围的情况下可以进行多种修改。例如,所公开的技术可以应用于医学领域之外的领域,例如预测天气模式,地质活动,或基于采样的输入数据在其中进行预测的任何其他领域。为了减少权利要求的数量,下面以某些权利要求的形式呈现了所公开技术的某些方面,但是申请人考虑了以任何数量的权利要求形式的所公开的技术的多个方面。因此,除了所附权利要求书外,所公开的技术不受限制。

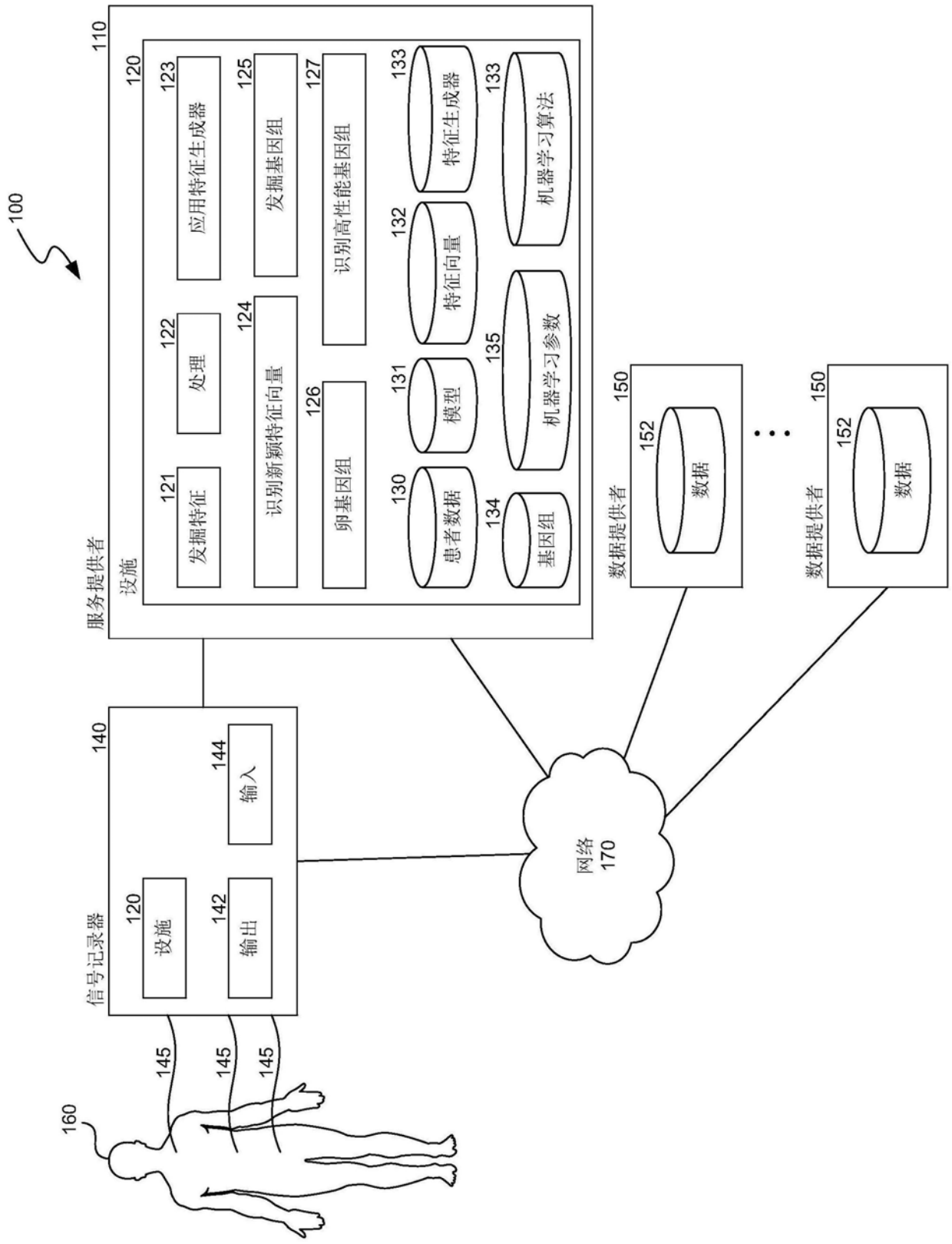

100
110
服务提供者
设施
120
121
发掘特征
122
处理
123
应用特征生成器
124
识别新颖特征向量
125
发掘基因组
126
卵基因组
127
识别高性能基因组
130
患者数据
131
模型
132
特征向量
133
特征生成器
134
基因组
135
机器学习参数
133
机器学习算法
140
信号记录器
120
设施
142
输出
144
输入
145
145
145
160
网络
170
150
数据提供者
152
数据
150
数据提供者
152
数据

图1

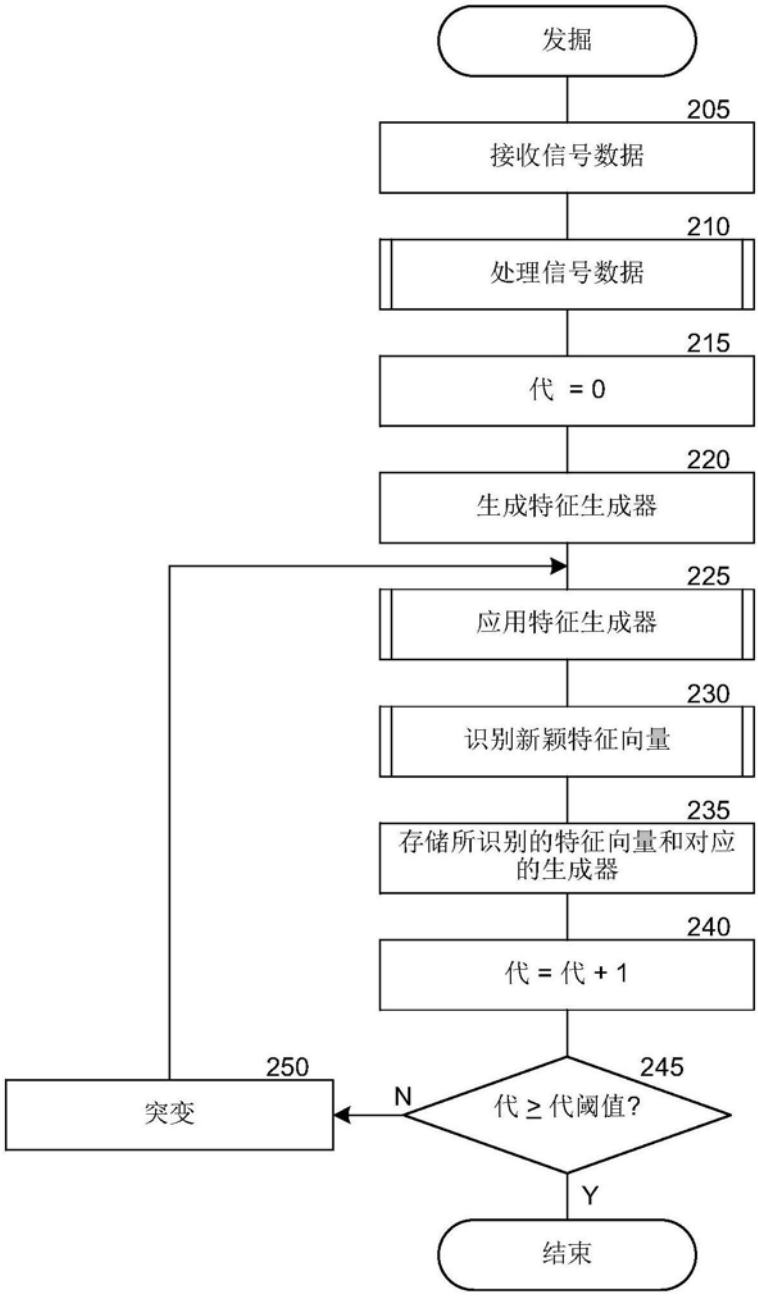

发掘
205
接收信号数据
210
处理信号数据
215
代 = 0
220
生成特征生成器
225
应用特征生成器
230
识别新颖特征向量
235
存储所识别的特征向量和对应的生成器
240
代 = 代 + 1
245
代 ≥ 代阈值?
N
250
突变
Y
结束

图2


处理
305
对于每个信号
310
预处理
315
存储信号数据
320
提取值
325
存储值
330
识别变换
335
对于每个变换
340
对信号进行变换
345
存储变换后的信号数据
350
提取值
355
存储值
360
下一个变换
365
下一个信号
返回

图3

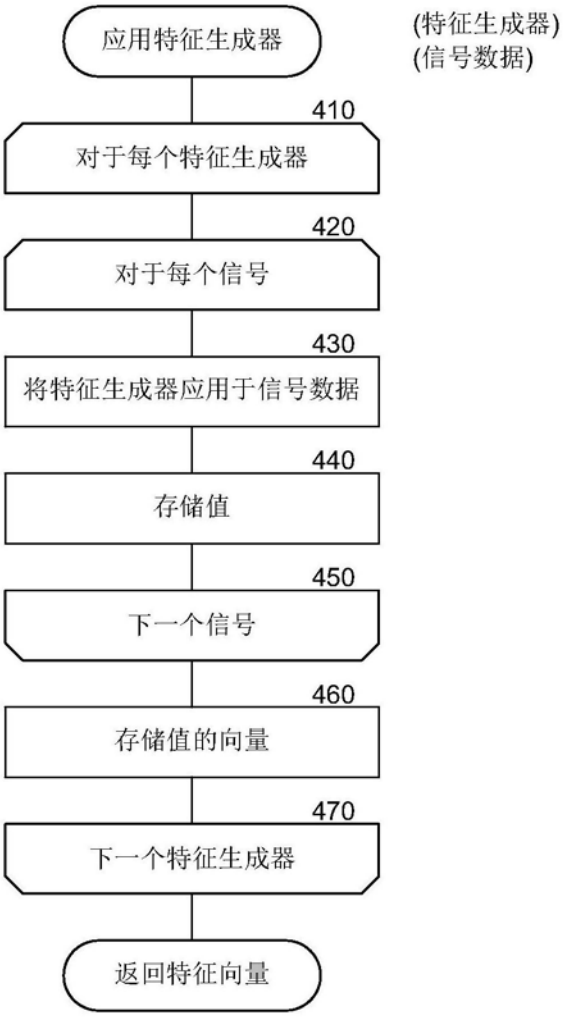

应用特征生成器
(特征生成器)
(信号数据)
410
对于每个特征生成器
420
对于每个信号
430
将特征生成器应用于信号数据
440
存储值
450
下一个信号
460
存储值的向量
470
下一个特征生成器
返回特征向量

图4

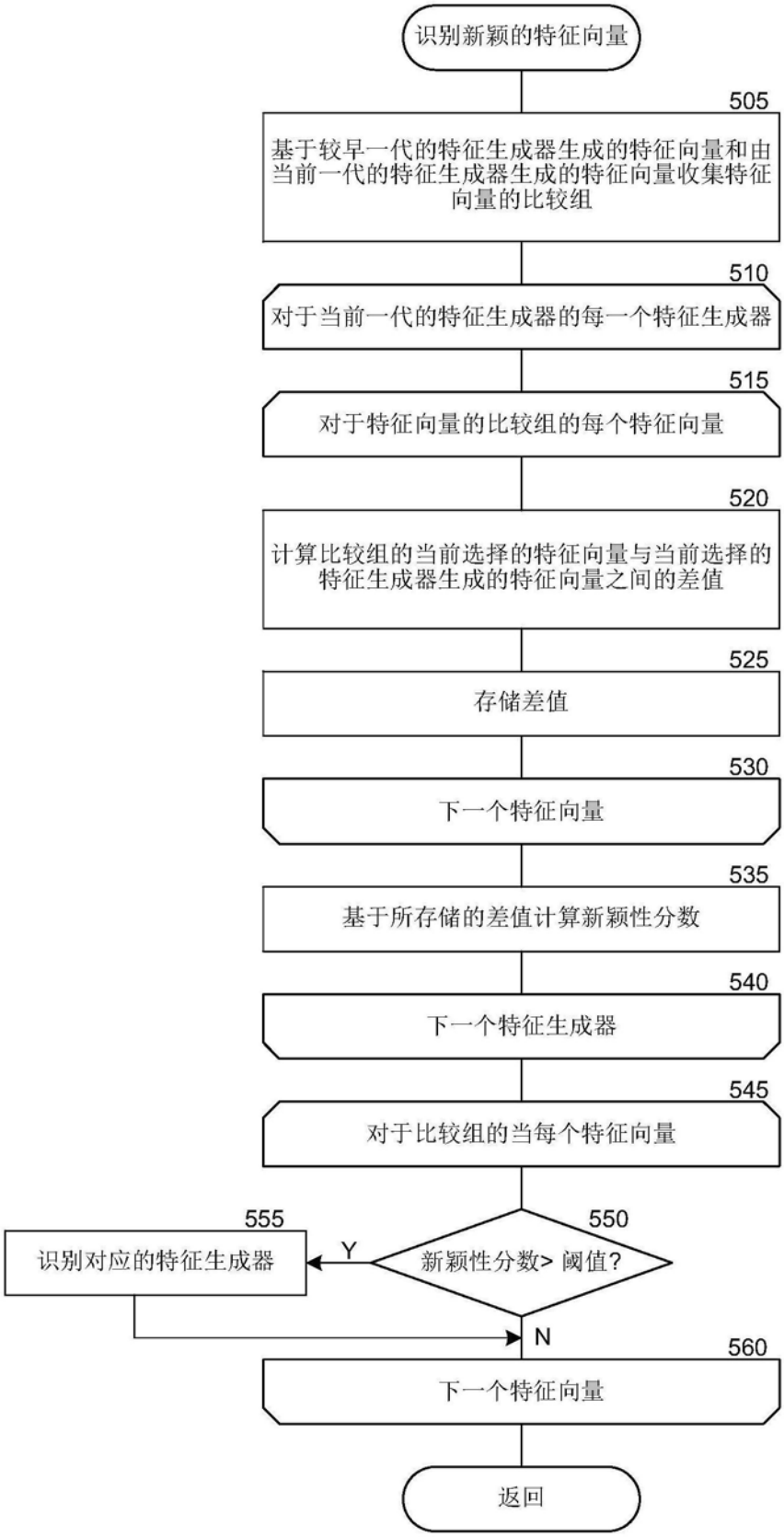

识别新颖的特征向量
505
基于较早一代的特征生成器生成的特征向量和由当前一代的特征生成器生成的特征向量收集特征向量的比较组
510
对于当前一代的特征生成器的每一个特征生成器
515
对于特征向量的比较组的每个特征向量
520
计算比较组的当前选择的特征向量与当前选择的特征生成器生成的特征向量之间的差值
525
存储差值
530
下一个特征向量
535
基于所存储的差值计算新颖性分数
540
下一个特征生成器
545
对于比较组的当每个特征向量
550
新颖性分数> 阈值?
Y
555
识别对应的特征生成器
N
560
下一个特征向量
返回

图5

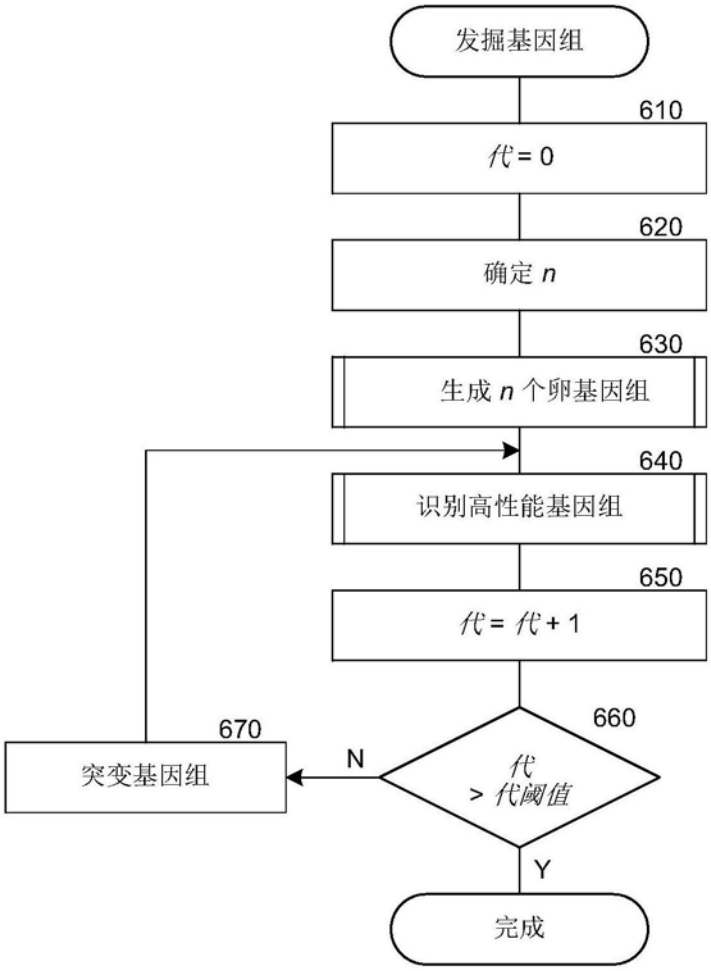

发掘基因组
610
代 = 0
620
确定 n
630
生成 n 个卵基因组
640
识别高性能基因组
650
代 = 代 + 1
660
代
> 代阈值
N
670
突变基因组
Y
完成

图6


生成基因组
710
识别特征
720
确定待包含的特征的数量
730
选择特征
740
替换相关的特征
750
识别机器学习参数
760
确定参数的数量
770
选择参数的数量
780
构建基因组
返回基因组

图7

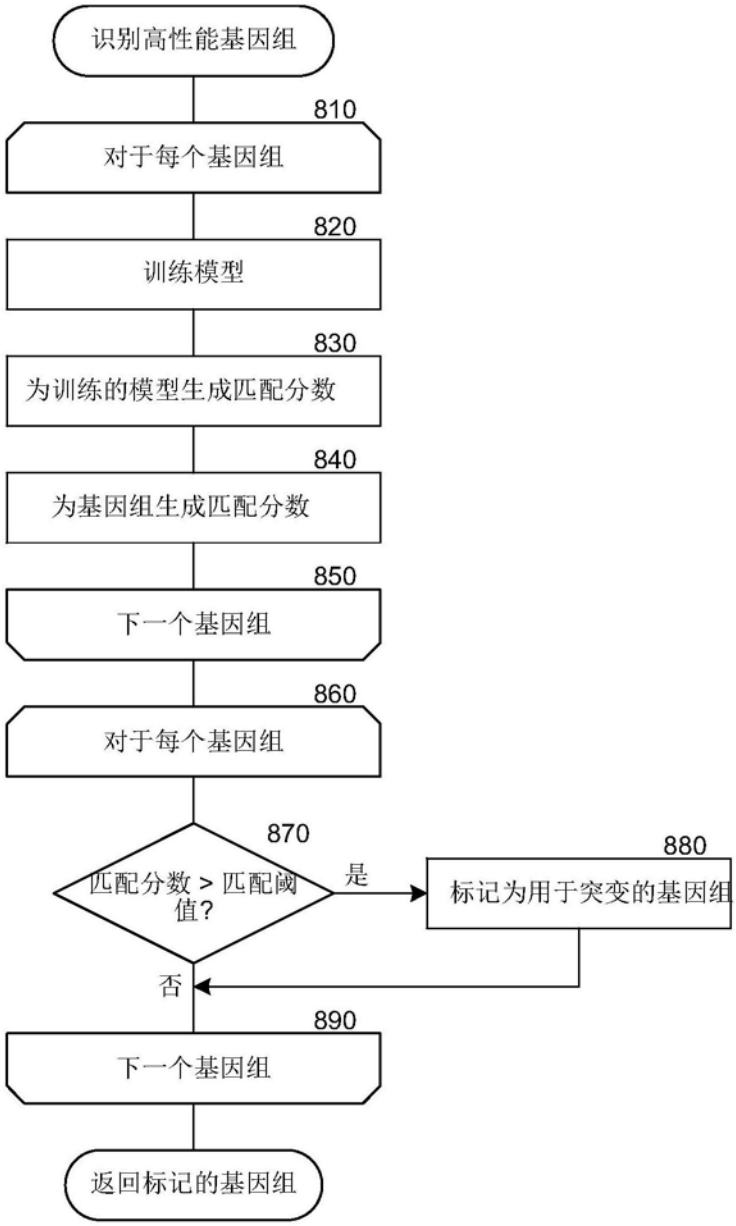

识别高性能基因组
810
对于每个基因组
820
训练模型
830
为训练的模型生成匹配分数
840
为基因组生成匹配分数
850
下一个基因组
860
对于每个基因组
870
匹配分数 > 匹配阈值?
是
880
标记为用于突变的基因组
否
890
下一个基因组
返回标记的基因组

图8