(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 104508671 B
(45)授权公告日 2018.10.19

(21)申请号 201380039806.5
(22)申请日 2013.06.21
(65)同一申请的已公布的文献号
申请公布号 CN 104508671 A
(43)申请公布日 2015.04.08
(30)优先权数据
61/662,792 2012.06.21 US
(85)PCT国际申请进入国家阶段日
2015.01.27
(86)PCT国际申请的申请数据
PCT/EP2013/062980 2013.06.21
(87)PCT国际申请的公布数据
WO2013/190084 EN 2013.12.27
(73)专利权人 菲利普莫里斯生产公司
地址 瑞士纳沙泰尔
(72)发明人 弗洛里安·马丁 向阳
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 宋岩
(51)Int.Cl.
G06F 19/24(2006.01)
(56)对比文件
CN 1749988 A,2006.03.22,
CN 101944122 A,2011.01.12,
CN 102135979 A,2011.07.27,
CN 102214213 A,2011.10.12,
TIBSHIRANI R ET AL.Diagnosis of multiple cancer types by shrunken centroids of gene expression.《PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES-PNAS》.2002,第99卷(第10期),全文.
审查员 贾云杰

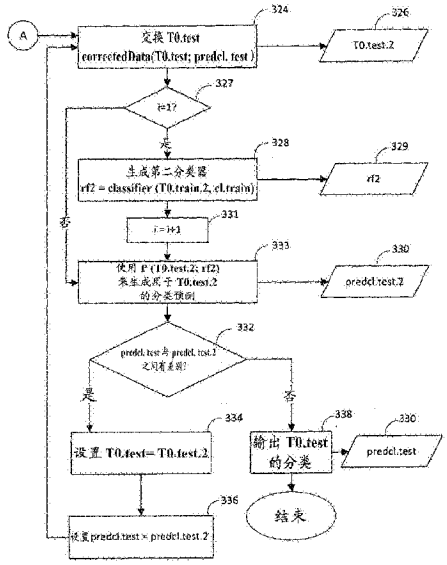
权利要求书2页 说明书12页 附图6页

(54)发明名称

通过偏差校正和分类预测生成生物标记签名的系统和方法

(57)摘要

本文详述了用于按集成方式校正数据集并对数据集进行分类的系统和方法。训练数据集、训练分类集和测试数据集被接收。对于所述训练数据集,通过将机器学习技术应用到训练数据集和训练分类集来生成第一分类器,并且通过根据第一分类器对测试数据集中的元素进行分类来生成第一测试分类集。对于多次迭代中的每一次,训练数据集被变换,测试数据集被变换,并且通过将机器学习技术应用到经变换的训练数据集来生成第二分类器。根据第二分类器来生成第二测试分类集,并且将第一测试分类集与第二测试分类集相比较。



CN 104508671 B

1.一种由处理器运行的将数据集分类到两个或更多个分类的计算机实现的方法，包括：

(a)接收训练数据集和训练分类集，所述训练数据集的元素表示患病病人、对疾病有抵抗力的病人或未患病病人的基因表达数据，所述训练分类集包括已知标签的集合，各已知标签标识与所述训练数据集中的每个元素相关联的分类；

(b)接收测试数据集；

(c)通过将第一机器学习技术应用到所述训练数据集和所述训练分类集来生成用于所述训练数据集的第一分类器；

(d)通过根据所述第一分类器对所述测试数据集中的元素进行分类来生成第一测试分类集；

(e)通过将所述训练数据集中的元素移动与训练分类质心的集合的中心相对应的量来变换所述训练数据集，其中各训练分类质心代表所述训练数据集中的元素的子集的中心；以及

(f)对于多次迭代中的每一次：

(i)通过将所述测试数据集中的元素移动与测试分类质心的集合的中心相对应的量来变换所述测试数据集，其中各测试分类质心代表所述测试数据集中的元素的子集的中心；

(ii)通过根据第二分类器对经变换的测试数据集中的元素进行分类来生成第二测试分类集，其中所述第二分类器是通过将第二机器学习技术应用到经变换的训练数据集和所述训练分类集而生成的；

(iii)当所述第一测试分类集与所述第二测试分类集不同时，将所述第二测试分类集存储为所述第一测试分类集并将所述经变换的测试数据集存储为所述测试数据集并且返回步骤(i)。

2.根据权利要求1所述的方法，还包括当所述第一测试分类集与所述第二测试分类集并非不同时，输出所述第二测试分类集。

3.根据权利要求1-2中任一个所述的方法，其中所述训练数据集是从总数据集中的样本的随机子集形成的，所述测试数据集是从所述总数据集中的样本的剩余子集形成的。

4.根据权利要求1-2中任一个所述的方法，其中步骤(e)处的移动包括对所述训练数据集应用旋转、剪切、线性变换或非线性变换来获得所述经变换的训练数据集。

5.根据权利要求1-2中任一个所述的方法，其中步骤(i)处的移动包括对所述测试数据集应用旋转、剪切、线性变换或非线性变换来获得所述经变换的测试数据集。

6.根据权利要求1-2中任一个所述的方法，其中：

所述测试数据集包括已知标签的测试集合，各已知标签标识与所述测试数据集中的每个元素相关联的分类；

所述第一测试分类集包括用于所述测试数据集的预测标签的集合；并且

所述第二测试分类集包括用于所述经变换的测试数据集的预测标签的集合。

7.根据权利要求1-2中任一个所述的方法，还包括对于所述多次迭代中的每一次，将所述第一测试分类集与所述第二测试分类集进行比较。

8.根据权利要求1-2中任一个所述的方法，其中第一机器学习技术和第二机器学习技术是相同的机器学习技术。

9.根据权利要求1-2中任一个所述的方法，其中步骤(e)处的变换是通过应用与步骤(i)的变换相同的变换来执行的。

10.根据权利要求1-2中任一个所述的方法，还包括将所述第二测试分类集提供到显示设备、打印设备或存储设备。

11.根据权利要求1-2中任一个所述的方法，其中如果所述第一测试分类集中的任一元素与所述第二测试分类集中的相应元素不同，则所述第一测试分类集和所述第二测试分类集不同。

12.根据权利要求1-2中任一个所述的方法，其中所述第二测试分类集包括用于所述经变换的测试数据集的预测标签的集合，所述方法还包括：通过计算代表所述第二测试分类集中的正确预测标签的数目除以预测标签的总数的性能度量来评价所述第二分类器。

13.一种包含计算机可读指令的计算机可读介质，当在包括至少一个处理器的计算机化系统中被运行时，所述计算机可读指令使得所述至少一个处理器执行根据权利要求1-12中任一个所述的方法。

14.一种包含配备有非暂时性计算机可读指令的至少一个处理器的计算机化系统，所述非暂时性计算机可读指令当被运行时使得处理器执行根据权利要求1-12中任一个所述的方法。

通过偏差校正和分类预测生成生物标记签名的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请根据35 U.S.C§119要求2012年6月21日递交的、题为“SYSTEMS AND METHODS FOR GENERATING BIOMARKER SIGNATURES WITH INTEGRATED BIAS CORRECTION AND CLASS PREDICTION”的美国临时专利申请No.61/662,792的优先权，该美国临时专利申请被完整结合于此。

技术领域

[0003] 本发明一般地涉及用于通过集成的偏差校正和分类预测生成生物标记签名的系统和方法。

背景技术

[0004] 在生物医学领域，识别表明特定生物状态的物质即生物标记 (biomarker)很重要。随着基因组学和蛋白质组学的新技术出现，生物标记对于生物发现、药品研发和健康护理正变得越来越重要。生物标记不仅对于很多疾病的诊断和预后有用，而且对于理解疗法的发展基础有用。生物标记的成功和有效识别可以加速新药品研发过程。通过疗法与诊断和预后的结合，生物标记识别还将增强当前医疗的质量，因而在药物遗传学、药物基因组学和药物蛋白质组学的使用中扮演重要角色。

[0005] 包括高吞吐量筛选在内的基因组学和蛋白质组学分析提供了关于细胞中表达的蛋白质的数量和形式的丰富信息并提供了针对每个细胞识别特定细胞状态的被表达的蛋白质特性的谱的潜力。在某些情况下，该细胞状态可能是与疾病相关的异常生理反应的特征。结果，识别患病病人的细胞状态并与正常病人的相应细胞状态进行比较可以提供诊断和治疗疾病的机会。

[0006] 这些高吞吐量筛选技术提供了基因表达信息的大数据集。研究人员已尝试研发用于将这些数据集组织到可再现地诊断不同数量的个体的模式中的方法。一种方法是汇集来自多个源的数据以形成组合的数据集，然后将数据集划分成发现/训练集和测试/验证集。然而，相对于可用数量的样本，转录谱数据和蛋白质表达谱数据二者常常由大量变量来表征。

[0007] 来自控制或病人群的试样的表达谱之间的观察到的差异通常被若干因素掩盖，包括疾病或控制群体内的未知子表型或生物学差异、由研究方案的差别造成的依位置而定的偏差、试样处理、由仪器条件 (例如，芯片批次等)的差别造成的偏差以及由测量误差造成的变化。一些技术尝试针对数据样本中的偏差进行校正(所述偏差例如可能源于数据集中表示的一类样本多于另一类)。

[0008] 若干基于计算机的方法已被研发以找出最佳地解释疾病和控制样本之间的差别的一组特征(标记)。某些早期方法包括诸如LIMMA 之类的统计测试、用于识别与乳腺癌有关的生物标记的FDA批准的 mammaprint技术、诸如支撑向量机(SVM)之类的逻辑回归技术和机器学习方法。一般地，从机器学习的角度，生物标记的选择通常是分类任务的特征选择

问题。然而，这些早期方案面临若干缺点。由这些技术生成的签名(signature)常常是不能再现的，因为对象的包含与排除可能导致不同的签名。这些早期方案还生成很多假阳性签名并且不鲁棒，因为它们是在具有小样本尺寸和高维度的数据集上操作的。

[0009] 因此，需要用于识别用于临床诊断和/或预后的生物标记的改进的技术，并且更具体地，需要用于识别能够用来将数据集中的元素分类到两个或更多个分类中的数据标记的改进的技术。

发明内容

[0010] 申请人已认识到现有的基于计算机的方法不利地与分类预测技术分开应用偏差校正技术。本文描述的计算机系统和计算机程序产品实现将集成方法应用到偏差校正和分类预测的方法，其可在生物标记和其他数据分类应用中实现改进的分类性能。具体地，本文公开的计算机实现的方法采用迭代方法进行偏差校正和分类预测。在计算机实现的方法的各种实施例中，系统中的至少一个处理器接收训练数据集和训练分类集，训练分类集标识与训练数据集中的每个元素相关的分类。系统中的处理器还接收测试数据集。处理器通过将机器学习技术应用到训练数据集和训练分类集来生成用于训练数据集的第一分类器，并通过根据第一分类器对测试数据集中的元素进行分类来生成第一测试分类集。对于多次迭代中的每一次，处理器：基于训练分类集和测试分类集中的至少一个来变换训练数据集，通过应用前一步的变换来变换测试数据集，通过将机器学习技术应用到经变换的训练数据集和训练分类集来生成用于经变换的训练数据集的第二分类器，并且通过根据第二分类器对经变换的测试数据集中的元素进行分类来生成第二测试分类集。处理器还将第一测试分类集与第二测试分类集相比较，并且当第一测试分类集与第二测试分类集不同时，处理器将第二分类集存储为第一分类集，将经变换的测试数据集存储为测试数据集并返回到迭代开头。本发明的计算机系统包括用于实现如上所述的方法及其各种实施例的装置。

[0011] 在如上所述方法的某些实施例中，该方法还包括当第一测试分类集与第二测试分类集并非不同时输出第二分类集。具体地，如上所述的迭代可重复至第一测试分类集和第二测试分类集收敛并且预测的分类之间没有差别。在如上所述方法的某些实施例中，训练数据集的元素表示患病病人、对疾病有抵抗力的病人或未患病病人的基因表达数据。训练分类集的元素可对应于训练数据集中的数据样本的已知分类标识。例如，分类标识可包括诸如“疾病阳性”、“疾病免疫”或“无疾病”之类的类别。

[0012] 在如上所述方法的某些实施例中，训练数据集和测试数据集是通过将总数据集中的样本随机指派到训练数据集或测试数据集而生成的。将总数据集随机地分裂成训练数据集和测试数据集可能是预测分类和生成鲁棒基因签名所需要的。另外，总数据集的样本可在分裂之前被丢弃，或者训练数据集或测试数据集的样本可在分裂之后被丢弃。在如上所述方法的某些实施例中，变换训练数据集的步骤、变换测试数据集的步骤、或者变换训练数据集和变换测试数据集的步骤二者包括通过基于数据集的质心调整数据集的元素来执行偏差校正技术。变换是根据变换函数来执行的，变换函数可基于训练数据集来定义变换。在如上所述方法的某些实施例中，偏差校正技术包括从数据集的每个元素中减去质心的分量。例如，偏差校正技术的结果可以是训练数据集、测试数据集或者训练和测试数据集二者的每个元素通过将数据集中表示的每个分类的质心考虑在内而“回到中心”(recenter)。

在如上所述方法的某些实施例中，变换训练数据集的步骤、变换测试数据集的步骤、或者变换训练数据集和变换测试数据集的步骤二者包括应用旋转、剪切、移动、线性变换或非线性变换。

[0013] 在如上所述方法的某些实施例中，该方法还包括对于多次迭代中的每一次，将第一测试分类集与第二测试分类集相比较。作为比较结果，如果第一测试分类集中的任一单个元素与第二测试分类集中的相应元素不同，则第一测试分类集和第二测试分类集可被认为不同。一般地，阈值可被设置以使得如果第一测试分类集中的预定数目的元素与第二测试分类集中的相应元素不同，则第一测试分类集和第二测试分类集被认为不同。

[0014] 在如上所述方法的某些实施例中，该方法还包括对于多次迭代中的每一次，通过将机器学习技术应用到经变换的训练数据集和训练分类集来生成用于经变换的训练数据集的第二分类器。在如上所述方法的某些实施例中，测试数据集的变换涉及与变换训练数据集的变换相同的变换。在如上所述方法的某些实施例中，该方法还包括将第二测试分类集提供到显示设备、打印设备或存储设备。在如上所述方法的某些实施例中，该方法还包括基于误差率计算第二分类器的性能度量。在某些实施例中，诸如但不限于线性判别分析(LDA)、逻辑回归、支撑向量机、朴素贝叶斯分类器之类的线性分类器是优选的。

[0015] 本发明的计算机系统包括用于实现如上所述方法的各种实施例的装置。例如，计算机程序产品被描述，该产品包括计算机可读指令，当在包括至少一个处理器的计算机化系统中运行时，所述计算机可读指令使得该处理器实现如上所述方法中的任一方法的一个或多个步骤。在另一示例中，计算机化系统被描述，该系统包括配备了非暂时性计算机可读指令的处理器，所述非暂时性计算机可读指令当被运行时使得该处理器实现如上所述方法中的任一方法。本文描述的计算机程序产品和计算机化方法可在具有一个或多个计算设备的计算机化系统中实现，每个计算设备包括一个或多个处理器。一般地，本文描述的计算机化系统可包括一个或多个引擎，所述引擎包括处理器或设备，如配备了硬件、固件和软件以实现本文描述的一个或多个计算机化方法的计算机、微处理器、逻辑设备或其他设备或处理器。这些引擎中的任一个或多个可与任一个或多个其他引擎物理上可分离，或者可包括多个物理上可分离的部件，如公共或不同电路板上的分离的处理器。本发明的计算机系统包括用于实现如上所述的方法及其各种实施例的装置。引擎可不时被互连，并且不时被进一步连接到一个或多个数据库，包括扰动数据库、可测量量数据库、实验数据数据库和文献数据库。本文描述的计算机化系统可包括具有通过网络接口通信的一个或多个处理器和引擎的分布式计算机化系统。该实现方式可以适合于多个通信系统上的分布式计算。

附图说明

[0016] 考虑以下结合附图的详细说明后，将明了本公开的其他特征、性质和各种优点，附图中相同的标号指代各处相同的部分，其中：

[0017] 图1描绘了用于识别一个或多个生物标记签名的示例性系统；

[0018] 图2例示了数据集中元素的分类；

[0019] 图3是用于对数据集进行分类的示例性处理的流程图；

[0020] 图4是诸如图1的系统的部件中的任一部件之类的计算设备的框图；

[0021] 图5是训练数据集中的基因签名的热图。

具体实施方式

[0022] 为了提供本文描述的系统和方法的整体理解，现在将描述某些例示性实施例，包括用于识别基因生物标记签名的系统和方法。然而，本领域普通技术人员将理解，本文描述的系统、计算机程序产品和方法可针对例如任何数据分类应用之类的其他合适应用被改编和修改，并且这类其他补充和修改将不脱离其范围。一般地，本文描述的计算机化系统可包括一个或多个引擎、处理器或设备，如配备了硬件、固件和软件以实现本文描述的一个或多个计算机化方法的计算机、微处理器或逻辑设备。

[0023] 图1描绘了用于识别一个或多个生物标记签名的示例性系统 100，其中可实现本文公开的分类技术。该系统100包括生物标记发生器102和生物标记巩固器(consolidator)104。系统100还包括用于控制生物标记发生器102和生物标记巩固器104的操作的某些方面的中央控制单元(CCU)101。操作期间，诸如基因表达数据之类的数据在生物标记发生器102处被接收。生物标记发生器102处理该数据以生成多个候选生物标记和相应的误差率。生物标记巩固器104接收这些候选生物标记和误差率并选择具有最优性能量度和尺寸的合适生物标记。

[0024] 生物标记发生器102包括用于处理数据并生成一组候选生物标记和候选误差率的若干部件。具体地，生物标记发生器包括用于将数据分到训练数据集和测试数据集中的数据预处理引擎110。生物标记发生器102包括用于接收训练数据集和测试数据集并将测试数据集的元素分类到两个或更多个分类之一中的分类引擎114(例如，患病的和未患病的，易感的、免疫的和患病的，等等)。生物标记发生器102 包括用于确定应用到数据预处理引擎110选择的测试数据的分类器的性能的分类器性能监视引擎116。分类器性能监视引擎116基于分类器来识别候选生物标记(例如，对分类最重要的数据集的元素的分量)并为一个或多个候选生物标记生成可包括候选误差率的性能量度。生物标记发生器102还包括用于存储一个或多个候选生物标记和候选性能量度的生物标记存储器118。

[0025] 生物标记发生器可由CCU 101控制，CCU 101进而可被自动控制或由用户操作。在某些实施例中，生物标记发生器102可操作来在每次将数据随机分到训练和测试数据集中时生成多个候选生物标记。为了生成该多个候选生物标记，生物标记发生器102的操作可被迭代多次。CCU 101可接收包括候选生物标记的期望数目的一个或多个系统迭代参数，所述候选生物标记的期望数目进而可用于确定生物标记发生器102的操作可被迭代的次数。CCU 101还可接收包括期望的生物标记尺寸的其他系统参数，所述期望的生物标记尺寸可表示生物标记中分量的数目(例如，生物标记基因签名中基因的数目)。生物标记尺寸信息可被分类性能监视引擎116用于从训练数据生成候选生物标记。生物标记发生器102和分类引擎114的操作具体地被参考图2-4更详细地描述。

[0026] 生物标记发生器102生成被生物标记巩固器104用于生成健壮的生物标记的一个或多个候选生物标记和候选误差率。生物标记巩固器 104包括生物标记共识(consensus)引擎128，生物标记共识引擎 128接收多个候选生物标记并生成在多个候选生物标记中具有最频繁出现的基因的新生物标记签名。生物标记巩固器104包括用于确定多个候选生物标记上的总体误差率的误差计算引擎130。类似于生物标记发生器102，生物标记巩固器104也可由CCU 101控制，CCU 101 进而可被自动控制或由用户操作。CCU 101可接收和/或确定

用于最小生物标记尺寸的合适阈值，并使用该信息来确定操作生物标记发生器102和生物标记巩固器104二者的迭代次数。在一个实施例中，每次迭代期间，CCU 101将生物标记尺寸减小一并迭代生物标记发生器102和生物标记巩固器104二者，直至达到阈值。在该实施例中，对于每个迭代，生物标记共识引擎128输出新的生物标记签名和新的总体误差率。生物标记共识引擎128因而输出各自具有从阈值变化至高达最大生物标记尺寸的不同尺寸的新的生物标记签名的集合。生物标记巩固器104还包括生物标记选择引擎126，生物标记选择引擎 126查阅这些新的生物标记签名的每一个的性能量度或误差率，并选择最优的生物标记来输出。生物标记巩固器104及其各个引擎的操作被参考图2-4更详细地描述。

[0027] 图3是用于对数据集进行分类的示例性处理的流程图。在步骤 302，分类引擎114接收训练数据和测试数据。如下所述，分类引擎 114使用训练数据来开发一个或多个分类器，然后将这一个或多个分类器应用于测试数据。如图3所例示，训练数据包括训练数据集T0.train 304和训练分类集cl.train 306。训练数据集T0.train 304中的每个元素代表数据样本(例如，来自特定病人的表达数据的向量) 并对应于训练分类集cl.train 306中的已知分类标识。例如，在三分类的情形中，训练数据集T0.train 304中的第一元素可代表患有特定疾病的病人的基因表达数据，并可对应于训练分类集cl.train 306中的第一元素“疾病阳性”；训练数据集T0.train 304中的第二元素可代表对该特定疾病有抵抗力或免疫的病人的基因表达数据，并可对应于训练分类集cl.train 306中的第二元素“疾病免疫”；训练数据集 T0.train 304中的第三元素可代表没有该特定疾病的病人的基因表达数据，并可对应于训练分类集cl.train 306中的第三元素“无疾病”。步骤302处接收的测试数据包括测试数据集T0.test 308，测试数据集T0.test 308代表与训练数据集T0.train 304中的数据样本相同潜在类型的数据，但可代表例如从不同病人或不同实验取得的样本。可选地，分类引擎114还接收包括用于测试数据集中的数据样本的已知分类标识的测试分类集cl.test 310，所述已知分类标识可用于在分类引擎114生成的分类器被应用于测试数据集T0.test 308时评价该分类器的性能。在某些实现方式中，测试数据集T0.test 308中没有数据样本的已知分类可用，因此测试分类集cl.test 310不被提供到分类引擎114。

[0028] 一般地，步骤302处接收的数据可代表从中可提取分类的任何实验的或以其他方式获得的数据，例如样本中的多个不同基因的表达值，和/或诸如任何生物上重要的分析物的级别之类的各种表型特征。在某些实施例中，数据集可包括用于疾病条件和用于控制条件的表达水平数据。如本文所使用的，术语“基因表达水平”可指代例如 RNA或多肽之类的基因所编码的分子的量。mRNA分子的表达水平可包括mRNA的量(由编码mRNA的基因的转录活动决定)和 mRNA的稳定性(由mRNA的半衰期决定)。基因表达水平还可包括与基因所编码的给定氨基酸序列相对应的多肽的量。相应地，基因的表达水平可以对应于从基因转录的mRNA的量、基因所编码的多肽的量或者它们二者。基因的表达水平还可按不同形式的基因产品的表达水平来归类。例如，基因所编码的RNA分子可包括不同地表达的剪接变异体、具有不同起始位置或终止位置的转录和/或其他不同地处理的形式。基因所编码的多肽可涵盖裂开和/或修改形式的多肽。多肽可因磷酸化、脂化、异戊烯化、硫酸盐化、羟基化、乙酰化、核糖基化、法尼基化、增加糖类等而被修改。另外，具有给定类型的修改的多种形式的多肽可以存在。例如，多肽可以在多个位置处被磷酸化并且表达被不同地磷酸化的蛋白质的不同水平。

[0029] 在某些实施例中，细胞或组织中的基因表达水平可用基因表达谱来表示。基因表达谱可指代诸如细胞或组织之类的试样中的基因的表达水平的特征表示。来自个体的试样中的基因表达谱的确定代表该个体的基因表达状态。基因表达谱反映由细胞或组织中的一个或多个基因编码的信使RNA或多肽的表达或其形式。表达谱一般可指代示出不同细胞或组织中的不同表达模式的生物分子(核酸、蛋白质、碳水化合物)的谱。代表基因表达谱的数据样本可被存储为表达水平的向量，向量中的每个条目对应于特定生物分子或其他生物实体。

[0030] 在某些实施例中，数据集可包括代表样本中的多个不同基因的基因表达值的元素。在其他实施例中，数据集可包括代表质谱分析法检测到的峰的元素。一般地，每个数据集可包括各自对应于多个生物状态分类之一的数据样本。例如，生物状态分类可包括但不限于：样本的源(即，从其获得样本的病人)中是否存在疾病；疾病的阶段；疾病的风险；疾病复发的可能性；一个或多个基因座处的共享基因型 (例如，常见HLA单体型；基因突变；基因修改，如甲基化等)；接触药剂(例如，有毒物或潜在有毒物，环境污染物，候选药物等) 或条件(温度，pH等)；人口学特征(年龄，性别，体重；家族史；已有状况史等)；对药剂的耐受性，对药剂的敏感性(例如，对药物的反应性)等等。

[0031] 数据集可彼此独立以减少最终分类器选择中的收集偏差。例如，它们可能从多个源被收集并且可能使用不同的排除或包含标准在不同时间从不同地点被收集，即，当考虑定义生物状态分类的特征以外的特征时，数据集可能相对异质(heterogeneous)。造成异质的因素包括但不限于：由性别、年龄、种族造成的生物学差异；由饮食、运动、睡眠行为造成的个体差异；以及由用于血液处理的临床方案造成的样本处理差异。然而，生物状态分类可包括一个或多个常见特征 (例如，样本源可代表具有疾病和相同性别或一个或多个其他常见人口学特征的个体)。在某些实施例中，来自多个源的数据集是通过来自不同时刻和/或不同条件下的同一病人群体的样本的收集而生成的。

[0032] 在某些实施例中，多个数据集是从多个不同的临床试验位置获得的并且每个数据集包括在每个单独的试验位置获得的多个病人样本。样本类型包括但不限于：血液，血清，血浆，乳头抽取液，尿液，泪液，唾液，脊髓液，淋巴液，细胞和/或组织溶解物，激光显微切割组织或细胞样本，(例如石蜡块中的或冷冻的)包埋细胞或组织；(例如来自尸检的)新鲜或档案样本。样本例如可以从试管内的细胞或组织培养物获得。作为替代，样本可以从活的有机体或从诸如单细胞有机体之类的有机体群体获得。在一个示例中，当识别用于特定癌症的生物标记时，血液样本可从位于两个不同的测试位置的独立群所选择的对象被收集，从而提供将从中开发独立的数据集的样本。

[0033] 在某些实现方式中，训练和测试集由数据预处理引擎110(图 1)生成，数据预处理引擎110接收批量数据并将批量数据分到训练数据集和测试数据集中。在某些实施例中，数据预处理引擎110随机地将数据分到这两个群中。随机地分数据可能是预测分类和生成健壮的基因签名所需要的。在其他实施例中，数据预处理引擎110基于数据的类型或标签来将数据分到两个或更多个群中。一般地，数据可以按所需的任何合适方式被分到训练数据集和测试数据集中，而不脱离本公开的范围。训练数据集和测试数据集可具有任何合适的尺寸并且可具有相同或不同的尺寸。在某些实施例中，数据预处理引擎110可在将数据分到训练和测试数据集之前丢弃一条或多条数据。在某些实施例中，数据预处理引擎110可在任何

进一步处理之前从训练数据集和/或测试数据集丢弃一条或多条数据。

[0034] 在步骤311，分类引擎114将计数器变量i设为等于1。在步骤 312，分类引擎114基于训练数据集T0.train 304和训练分类集 cl.train 306生成第一分类器rf 314。图2例示了数据集中的元素的分类。分类引擎114可在步骤312使用任一种或多种已知的机器学习算法，包括但不限于：支撑向量机技术，线性判别分析技术，随机森林技术，k最邻近邻居技术，偏最小二乘技术(包括将最小二乘与线性判别分析特征相结合的技术)，逻辑回归技术，基于神经网络的技术，基于决策树的技术以及缩小质心技术(例如，Tibshirani，Hastle，Narasimhan和Chu在"Diagnosis of multiple cancer types by shrunken centroids of gene expression，"PNAS，v.99，n.10，2002中所描述的)。许多这类技术可作为R编程语言包来获得，包括与线性判别分析、支撑向量机、随机森林(Breiman，Machine Learning， 45 (1)：5-32(2001))、k最邻近邻居(Bishop，Neural Networks for Pattern Recognition，ed.O.U.Press，1995)、偏最小二乘判别分析和 PAMR(Tibshirani等人，Proc Natl Acad Sci USA，99(10)：6567-6572 (2002))相对应的lda，svm，randomForest，knn，pls.lda和pamr。分类引擎114可在步骤312中将第一分类器rf 314存储在存储器中。

[0035] 在步骤316中，分类引擎114通过将(步骤312处生成的)第一分类器rf 314应用于测试数据集T0.test 308来生成一组预测的测试分类predcl.test 318。分类引擎114可在步骤316中将预测的分类 predcl.test 318存储在存储器中。

[0036] 在步骤320中，分类引擎114变换训练数据集T0.train 304。该变换根据变换函数correctedData进行，该变换函数基于训练分类集 cl.train 306来变换训练数据集T0.train 304。步骤310的变换结果为分类引擎114可存储到存储器中的经变换的训练数据集T0.train.2 322。在某些实现方式中，分类引擎114在步骤320处执行的变换包括偏差校正技术。例如，变换可通过相对于作为一个整体来看的数据集的质心或数据集中表示的每个分类的质心调整训练数据集T0.train 304的元素来使训练数据集T0.train 304"回到中心"。

[0037] 一个具体的回到中心技术涉及基于不同群的质心的中心来将训练数据集T0.train 304的元素置于中心。如果训练数据集T0.train 304 中存在n个数据样本，并且每个数据样本是具有p个条目的向量 (例如，表示p个不同基因的表达水平)，则令xij表示数据样本j 的第i个条目。如果训练分类集cl.train 308表示K个不同的分类，则令Ck表示分类k中的nk个样本的索引。分类引擎114可将分类 k的质心的第i个分量计算为

[0038] $$\overline{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k} \quad (1)$$

[0039] 并且可将分类质心的中心的第i个分量计算为

[0040] $$\overline{x}_i^c = \sum_{k=1}^{K} \frac{\overline{x}_{ik}}{K} \quad (2)$$

[0041] 分类引擎114还可将总体质心的第i个分量计算为

[0042] $$\overline{x_i} = \sum_{j=1}^{n} \frac{x_{ij}}{n} \quad (3)$$

[0043] 分类引擎114随后可执行变换，所述变换包括通过添加由下式给出的差来调整训练数据集T0.train 304的每个元素中的第i个条目：

[0044] $$\Delta = -\overline{x_i^c} \quad (4)$$

[0045] 在某些实现方式中，在步骤320处执行的变换包括与上面参考公式1-4描述的移动不同的移动、旋转、剪切、这些变换的组合或任何其他线性或非线性变换。

[0046] 在步骤324，分类引擎114变换测试数据集T0.test 308。应用到测试数据集T0.test 308的变换correctedData与步骤320处应用到训练数据集T0.train 304的变换类型相同，但是是针对变元T0.test 308 和predcl.test 318而非T0.train 304和predcl.train 314应用的。例如，如果训练数据集T0.train 304的元素在步骤320处被调整了相对于训练数据集T0.train 304的分类的质心计算的由公式4给出的值Δ，则测试数据集T0.test 308的元素在步骤324处被调整相对于测试数据集T0.test308的分类的质心计算的由公式4给出的值Δ。步骤324的变换结果是分类引擎114可存储到存储器中的经变换的测试数据集T0.test.2 326。

[0047] 在步骤327，分类引擎114确定迭代计数器i的值是否等于1。如果是，则分类引擎114继续运行步骤328，其中分类引擎114使用经变换的训练数据集T0.train.2 322和训练分类集cl.train 306来生成第二分类器rf2 329。如上面参考步骤332并参考步骤336所描述的，任何机器学习技术可被应用以在步骤328生成分类器。第二分类器rf2 329可以与第一分类器rf 314具有相同类型(例如，都为SVM 分类器)或者具有不同类型。

[0048] 在步骤331，分类引擎114递增迭代计数器i，然后继续运行步骤333，其中分类引擎114将第二分类器rf2 329应用到(分类引擎 114在步骤324生成的)经变换的测试数据集T0.test.2 326。步骤333的输出是用于经变换的数据集T0.test.2 326的预测的分类的集合 predcl.test.2 330。分类引擎114可将该预测的分类输出到显示设备、打印设备、存储设备、跨网络与分类引擎114通信的另一设备或系统 100内部或外部的任何其他设备。

[0049] 在步骤332，分类引擎114确定(步骤316处生成的)预测的分类集predcl.test 318和(步骤328处生成的)预测的分类集 predcl.test.2 330的分类之间是否存在任何差别。如果预测的分类集一致(即，对于测试数据集T0.test 308中的每个数据样本，该数据样本的预测的分类在两个预测的分类集之间相同)，则分类引擎114 进行到步骤338并输出预测的分类集predcl.test.2 330(等同地，预测的分类集predcl.test 318)作为测试数据集T0.test 308的最终分类。

[0050] 如果分类引擎114在分类数据集predcl.test 318和分类数据集 predcl.test.2 330之间识别出差别，则分类引擎114进行到步骤334 并用(步骤324的变换生成的)经变换的测试数据集T0.test.2 326的值替换测试数据集T0.test 308的先前存储的值。结果，测试数据集 T0.test 308具有经变换的测试数据集T0.test.2 326的值。分类引擎 114进行到步骤336并用(步骤328处生成的)预测的分类集 predcl.test.2 330的值替换(步骤316处生成的)预测的分类集 predcl.test 318的先前存储的值。结果，预测的分类集

predcl.test 318 具有预测的分类集predcl.test.2 330的值。

[0051] 一旦测试数据集T0.test 308的值已被利用经变换的测试数据集 T0.test.2 326的值进行了更新并且预测的分类集predcl.test 318已被利用预测的分类集predcl.test.2 330的值进行了更新，分类引擎114 就返回到步骤324以执行新的变换并迭代该过程直到分类引擎114 (在步骤332)确定预测的分类之间不存在差别。

[0052] 分类器性能监视引擎116可使用合适的性能度量来分析图3的处理完结时分类引擎114产生的最终分类的性能。在某些实施例中，性能度量可包括误差率。性能度量还可包括正确的预测除以尝试的全部预测所得的数。在不脱离本公开的范围的情况下，性能度量可以是任何合适的量度。

[0053] 本发明的实现方式可包括但不限于：包含如本文所述的一个或多个特征的系统方法和计算机程序产品以及包含可操作来令一个或多个机器(例如，计算机、机器人)引起本文所述的操作的机器可读介质的物品。本文所述的方法可以由驻留在单个计算系统或多个计算系统中的一个或多个处理器或引擎来实现。这类多个计算系统可以被连接并且可以经由一个或多个连接来交换数据和/或命令或其他指令等，包括但不限于网络上的连接(例如，因特网，无线广域网，局域网，广域网，有线网络等等)，经由多个计算系统中的一个或多个之间的直接连接。

[0054] 图4是诸如包括用于执行参考图1-3描述的处理的电路的图1的系统100的部件中的任一个之类的计算设备的框图。系统100的部件中的每一个可在一个或多个计算设备400中实现。在某些方面，多个上述部件和数据库可被包含在一个计算设备400中。在某些实现方式中，部件和数据库可跨若干计算设备400来实现。

[0055] 计算设备400包含至少一个通信接口单元、输入/输出控制器 410、系统存储器以及一个或多个数据存储设备。系统存储器包括至少一个随机存取存储器(RAM)402和至少一个只读存储器 (ROM)404。这些元件全部都与中央处理单元(CPU)406通信以辅助计算设备400的操作。计算设备400可以以许多不同方式配置。例如，计算设备400可以是传统的独立计算机或者作为替代的，计算设备400的功能可分布在多个计算机系统和架构上。计算设备400可被配置成执行数据分裂、区分、分类、评分、排名和存储操作中的一些或全部。在图4中，计算设备400经由网络或本地网络链接到其他服务器或系统。

[0056] 计算设备400可被配置成分布式架构，其中数据库和处理器被装在分开的单元或地点中。某些这类单元执行主要处理功能并至少包含通用控制器或处理器和系统存储器。在该方面，这些单元中的每一个经由通信接口单元408被附接到通信集线器或端口(未示出)，所述通信集线器或端口用作与其他服务器、客户端或用户计算机和其他相关设备的主要通信链接。通信集线器或端口本身可具有最小的处理能力，主要用作通信路由器。各种通信协议可以是系统的一部分，包括但不限于：以太网、SAP、SAS™、ATP、蓝牙™、GSM和 TCP/IP。

[0057] CPU 406包含处理器(如一个或多个传统微处理器)和一个或多个补充的协同处理器(如用于从CPU 406卸载工作负荷的算术协同处理器)。CPU 406与通信接口单元408和输入/输出控制器410通信，其中CPU 406通过输入/输出控制器410与诸如其他服务器、用户终端或设备之类的其他设备通信。通信接口单元408和输入/输出控制器410可包括用于同时与例如其他处理器、服务器或客户终端通信的多个通信信道。彼此通信的设备无需连续向

彼此进行发送。相反，这类设备只需按需要向彼此进行发送，可实际上避免大部分交换数据，并且可要求执行若干步骤以建立设备之间的通信链路。

[0058] CPU 406还与数据存储设备通信。数据存储设备可包括磁、光或半导体存储器的适当组合，并且例如可包括RAM 402、ROM 404、闪速驱动器、诸如致密盘之类的光盘或硬盘或驱动器。CPU 406和数据存储设备各自可以例如完全位于单个计算机或其他计算设备内；或者可通过通信介质彼此连接，所述通信介质例如是USB端口、串行端口电缆、同轴电缆、以太网型电缆、电话线、射频收发机或其他类似的无线或有线介质或前述的组合。例如，CPU 406可经由通信接口单元408连接到数据存储设备。CPU 406可被配置成执行一个或多个特定的处理功能。

[0059] 数据存储设备例如可存储(i)计算设备400的操作系统412；(ii)适于根据这里描述的系统和方法，特别是根据针对CPU 406详细描述的处理，来指导CPU 406的一个或多个应用414(例如，计算机程序代码或计算机程序产品)；或(iii)可被用来存储程序所要求的信息的适于存储信息的(一个或多个)数据库416。在某些方面，(一个或多个)数据库包括存储实验数据和公开的文献模型的数据库。

[0060] 操作系统412和应用414可被以例如压缩的、非编译的和加密的格式存储，并且可包括计算机程序代码。程序的指令可被从非数据存储设备的计算机可读介质，例如从ROM 404或从RAM 402读取到处理器的主存储器中。虽然运行程序中的指令序列使得CPU 406执行本文所述的处理步骤时，但是硬布线电路可取代或结合用于实现本发明的处理的软件指令而被使用。因此，描述的系统和方法不限于硬件和软件的任一特定组合。

[0061] 合适的计算机程序代码可被提供以执行与本文所述的建模、评分和聚合有关的一个或多个功能。程序还可包括诸如操作系统412、数据库管理系统和允许处理器经由输入/输出控制器410与计算机外围设备(例如，视频显示器、键盘、计算机鼠标等)相接口的“设备驱动器”之类的程序元件。

[0062] 包含计算机可读指令的计算机程序产品也被提供。计算机可读指令当被加载并运行在计算机系统上时，使得计算机系统根据上述方法或方法的一个或多个步骤来操作。本文使用的术语“计算机可读介质”指提供或参与向计算设备400的处理器(或本文所述的设备的任何其他处理器)提供指令以供运行的任何非暂时性介质。该介质可采取多种形式，包括但不限于：非易失性介质和易失性介质。非易失性介质例如包括光、磁或磁光盘，或诸如闪存之类的集成电路存储器。易失性介质包括动态随机存取存储器(DRAM)，通常构成主存储器。常见形式的计算机可读介质例如包括柔性盘、软盘、硬盘、磁带、任何其他磁介质、CD-ROM、DVD、任何其他光介质、打孔卡、纸带、任何其他具有孔图案的物理介质、RAM、PROM、EPROM或EEPROM(电可擦除可编程只读存储器)、FLASH- EEPROM、任何其他存储芯片或盒、或者计算机可读取的任何其他非暂时性介质。

[0063] 将一个或多个指令的一个或多个序列运送到CPU 406(或本文所述设备的任何其他处理器)以供运行时可涉及各种形式的计算机可读介质。例如，指令可最初诞生于远程计算机(未示出)的磁盘上。远程计算机可以将指令加载到它的动态存储器中并通过以太网连接、电缆线路或者甚至是使用调制解调器的电话线来发送指令。计算设备 400(例如服务器)本地的通信设备可以在相应的通信线路上接收数据并将数据置于处理器的系统总线上。系统总线将数据运送到主存储器，其中处理器从主存储器获取并运行指令。可选地，主

存储器接收的指令可在被处理器运行之前或之后被存储在存储器中。另外，指令可作为电、电磁或光信号经由通信端口被接收，其中电、电磁或光信号是承载各种类型的信息的无线通信或数据流的示例性形式。

[0064] 示例

[0065] 以下公开数据集是从Gene Expression Omnibus(GEO) (http://www.ncbi.nlm.nih.gov/geo/)库下载的：

[0066] a.GSE10106(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10106)

[0067] b.GSE10135(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10135)

[0068] c.GSE11906(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1 1906)

[0069] d.GSE11952(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1 1952)

[0070] e.GSE13933(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13933)

[0071] f.GSE 19407(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE 19407)

[0072] g.GSE19667(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19667)

[0073] h.GSE20257(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20257)

[0074] i.GSE5058(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSES058)

[0075] j.GSE7832(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7832)

[0076] k.GSE8545(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8545).

[0077] 训练数据集位于Affymetrix平台(HGU-133+2)。元数据文件由属于R(R Development Core Team,2007)中的Bioconductor (Gentleman,2004)的affy包(Gautier,2004)的ReadAffy函数读取，并且质量是这样控制的：生成RNA降解图(利用affy包的AffyRNAdeg函数)、NUSE和RLE图(利用affyPLM函数(Brettschneider,2008))并计算MA(RLE)值；从训练数据集排除低于质量控制检查的阈值集或在上述数据集中重复的数组；使用gcrma 算法(Wu,2004)来标准化通过质量控制检查的数组。训练集样本分类是从用于每个数据集的GEO数据库的系列矩阵文件获得的。输出包括具有用于233个样本(28个COPD样本和205个控制样本) 的54675个探针组的基因表达矩阵。为了得到平衡的数据集，COPD 样本是多个时间的，以在共同待决的美国临时申请61/662812中描述的Duel Ensemble方法被应用之前获得224个COPD样本。利用包含205个控制和224个COPD病人的组合的数据集，具有409个基因的基因签名被建立。850个二进制值在随机向量中被使用。方法中使用的分类方法包括以下R包：lda，svm，randomForest，knn，pls.lda 和pamr。最大迭代次数被设为5000。训练数据集中的交叉验证处理中的Matthew相关系数(MCC)和准确度分别被设为0.743和0.87。训练数据集中的基因签名的热图被示于图5。在图5的热图中，基因表达值按行位于中心。热图的颜色可能未被清楚地以灰度示出，但是图5的数据表明控制数据在左侧被示出，COPD数据在右侧被示出。测试数据集是从商业供应商(Genelogic)获得的未公开的数据集，包括16个控制样本和24个COPD样本。不应用本发明的变换不变量方法，Dual Ensemble生成的基因签名正确地预测了共40 个样本中的29个样本。准确度为0.725，MCC为0.527。在16个控制样本中，基因签名正确地预测了15个作为控制，而错误地预测了 1个作为COPD。在24个COPD样本中，基因签名正确地预测了14 个作为COPD样本，而错误地预测了10个作为控制。

[0078] 然而，当在根据两个或更多个分类的中心进行移动并且最大迭代次数设为100的情况下应用变换不变量方法时，相同的基因签名正确地预测了共40个样本中的30个样本。

准确度为0.75,MCC为 0.533。在16个控制样本中,基因签名正确地预测了14个作为控制,而错误地预测了2个作为COPD。在24个COPD样本中,基因签名正确地预测了16个作为COPD样本,而错误地预测了8个作为控制。

[0079] 虽然已参考特定示例具体示出并描述了本发明的实现方式,但是本领域技术人员应理解,可在不脱离本公开的精神和范围的情况下对本发明做出各种形式和细节的改变。

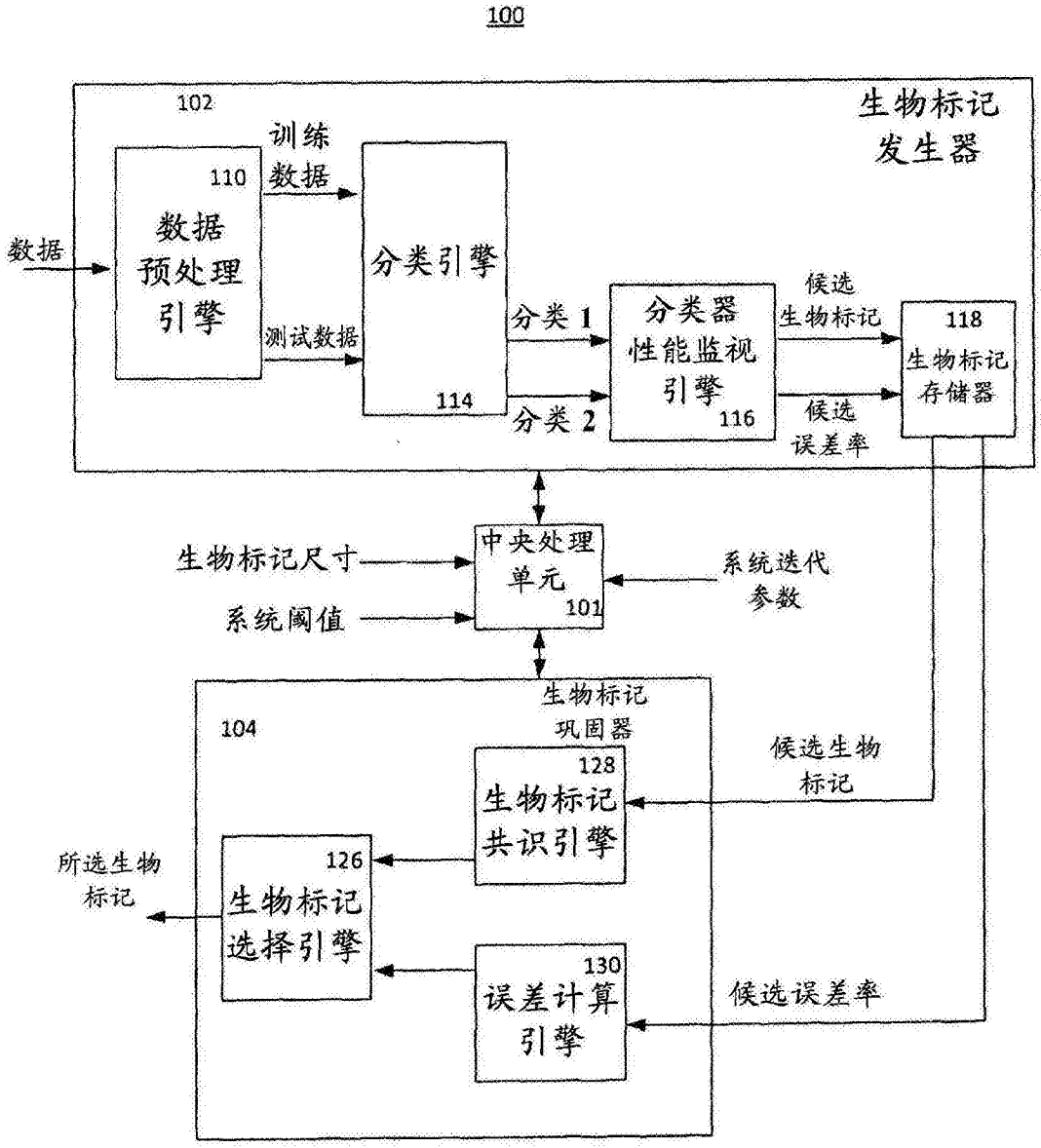

100
102
生物标记发生器
训练数据
110
数据预处理引擎
数据
分类引擎
测试数据
114
分类 1
分类 2
分类器性能监视引擎
116
候选生物标记
候选误差率
118
生物标记存储器
中央处理单元
101
生物标记尺寸
系统阈值
系统迭代参数
104
生物标记巩固器
128
生物标记共识引擎
候选生物标记
126
生物标记选择引擎
所选生物标记
130
误差计算引擎
候选误差率

图1

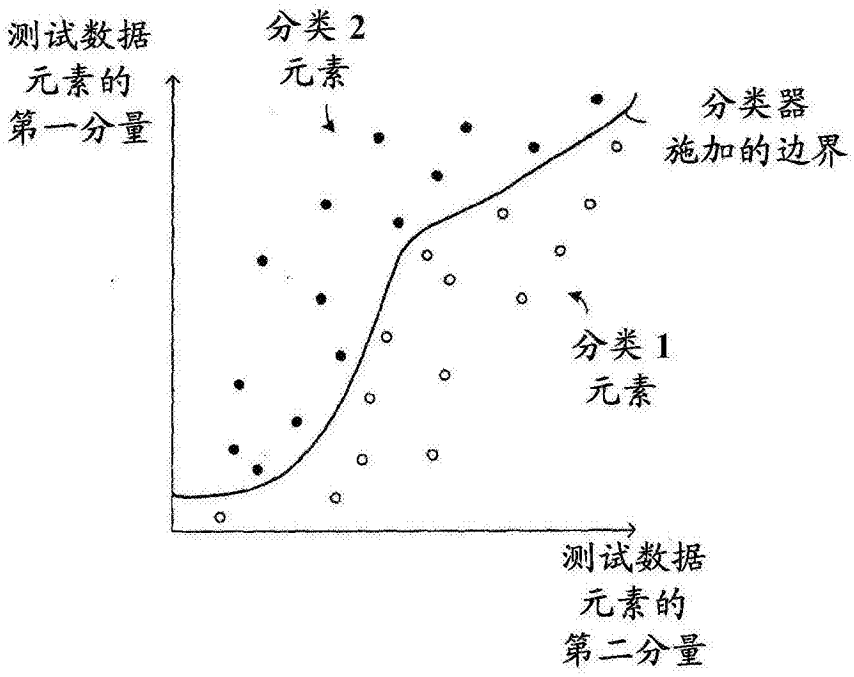


图2

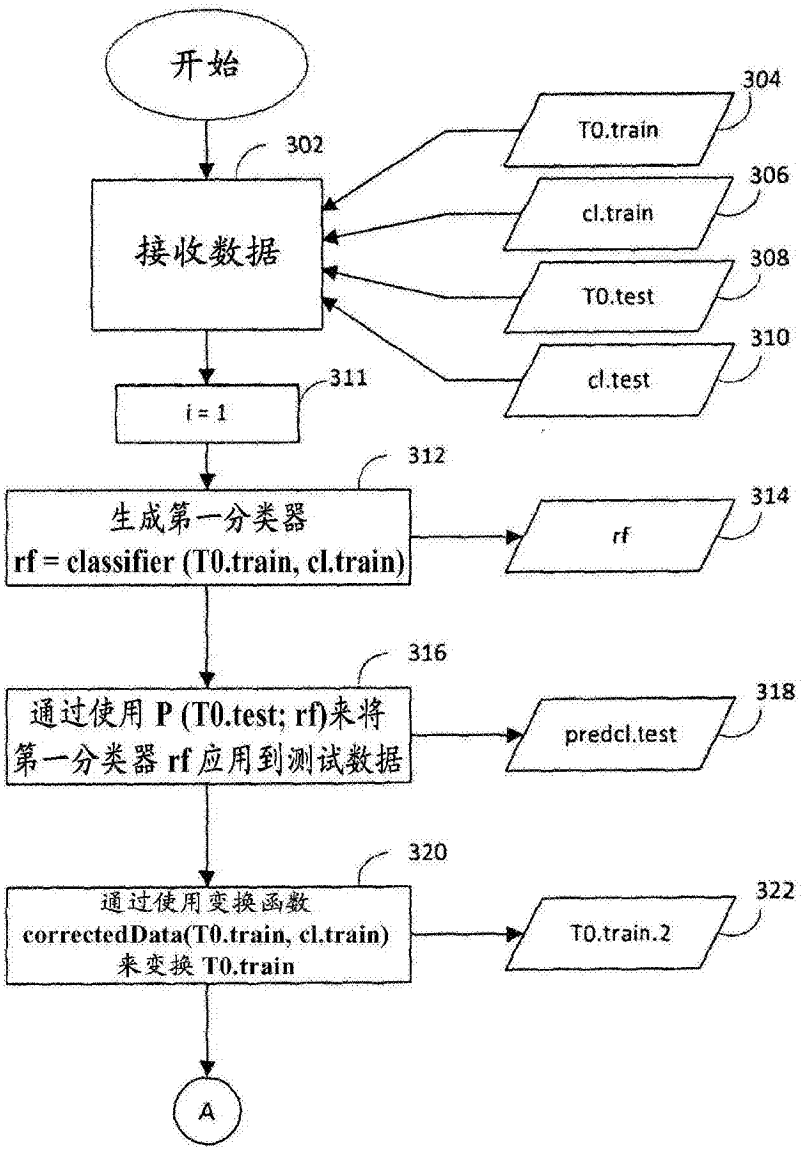

开始
302
接收数据
304
T0.train
306
cl.train
308
T0.test
310
cl.test
311
i = 1
312
生成第一分类器
rf = classifier (T0.train, cl.train)
314
rf
316
通过使用 P (T0.test; rf)来将
第一分类器 rf 应用到测试数据
318
predcl.test
320
通过使用变换函数
correctedData(T0.train, cl.train)
来变换 T0.train
322
T0.train.2
A

图3,第1页

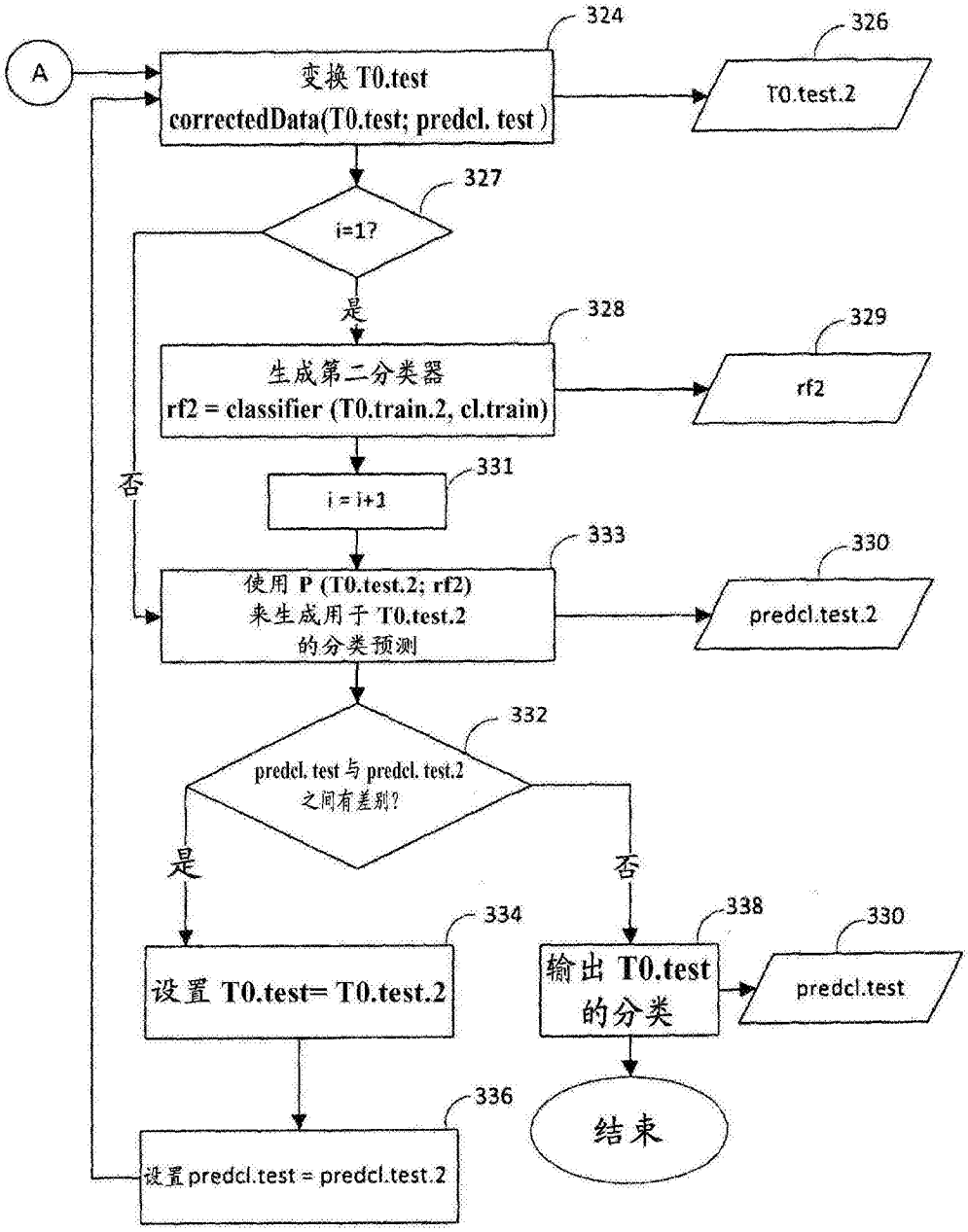

A
变换 T0.test
correctedData(T0.test; predcl. test)
324
T0.test.2
326
i=1?
327
是
否
生成第二分类器
rf2 = classifier (T0.train.2, cl.train)
328
rf2
329
i = i+1
331
使用 P (T0.test.2; rf2)
来生成用于 T0.test.2
的分类预测
333
predcl.test.2
330
predcl. test 与 predcl. test.2
之间有差别?
332
是
否
设置 T0.test= T0.test.2
334
输出 T0.test
的分类
338
predcl.test
330
结束
设置predcl.test = predcl.test.2
336

图3,第2页

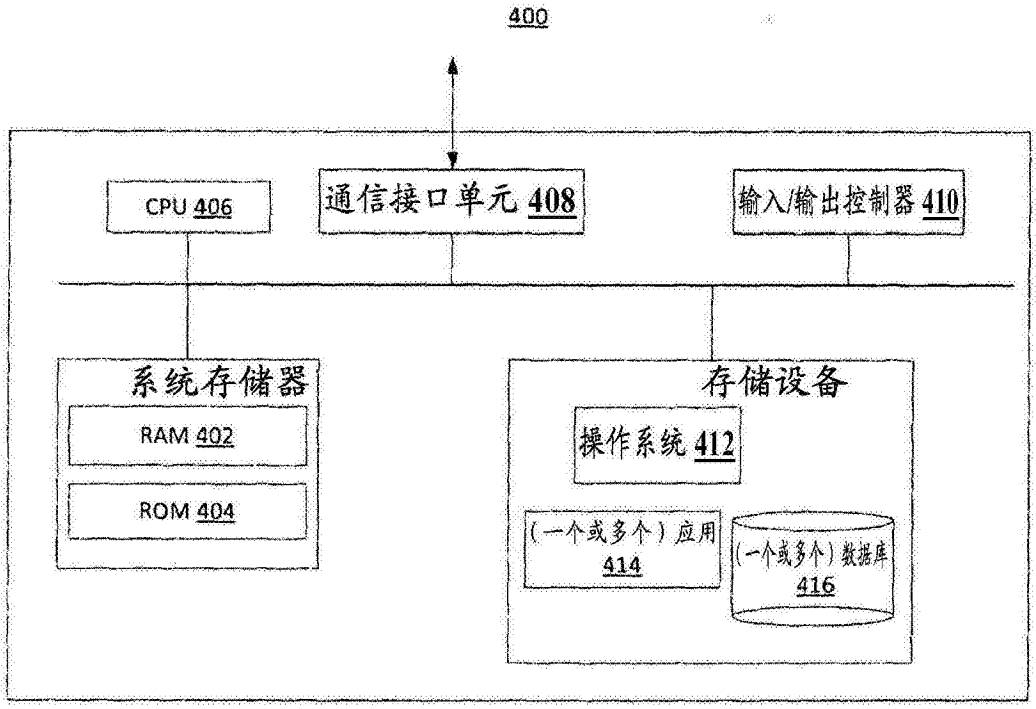

400
CPU 406
通信接口单元 408
输入/输出控制器 410
系统存储器
RAM 402
ROM 404
存储设备
操作系统 412
(一个或多个)应用 414
(一个或多个)数据库 416

图4

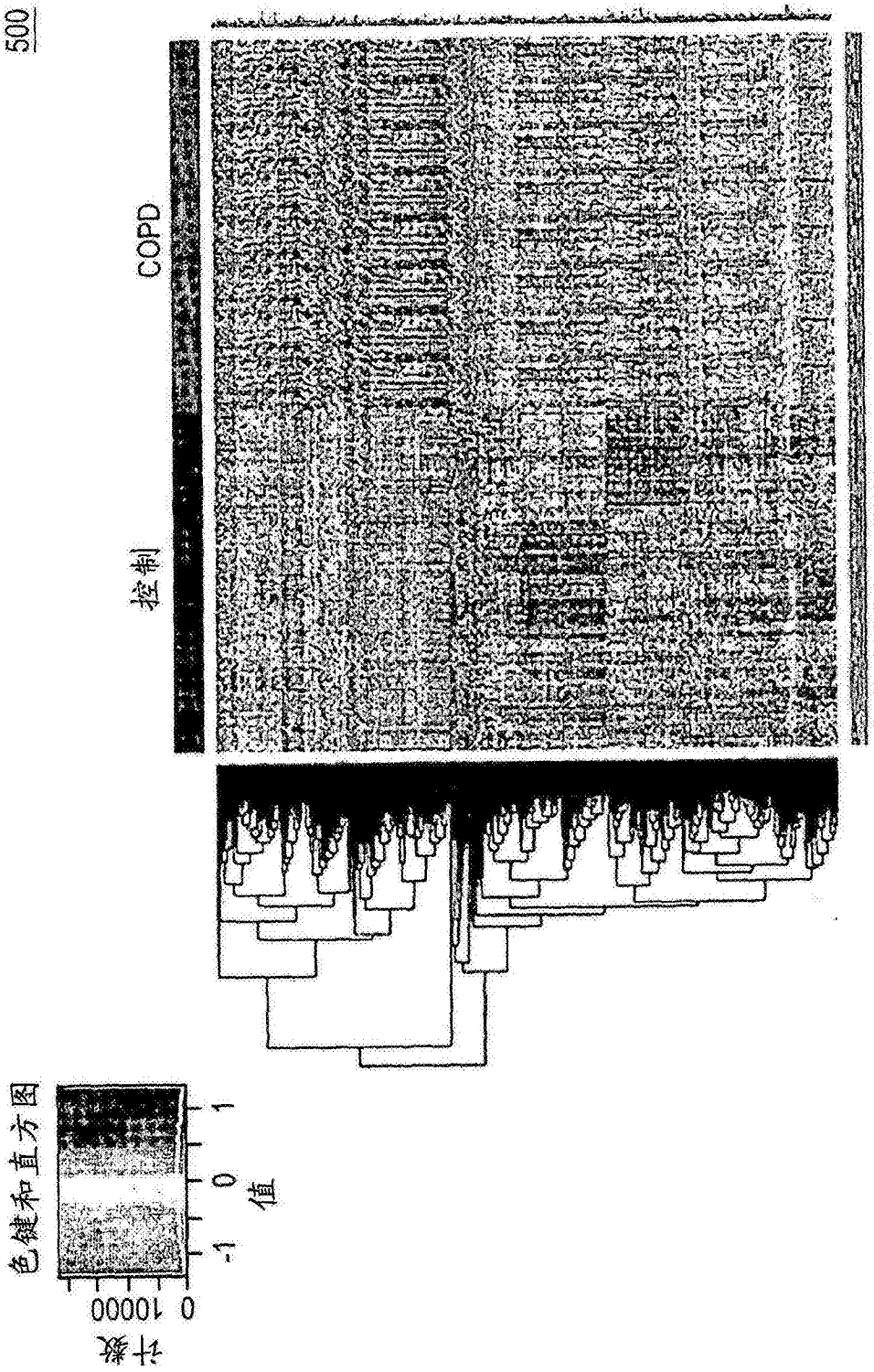


图5