(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 104584022 B
(45)授权公告日 2018.11.16

(21)申请号 201380039892.X
(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 104584022 A
(43)申请公布日 2015.04.29
(30)优先权数据
61/662,812 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/EP2013/062982 2013.06.21
(87)PCT国际申请的公布数据
WO2013/190085 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 向阳 朱丽娅·亨格
弗洛里安·马丁
(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 鲍进
(51)Int.Cl.
G06F 19/24(2006.01)
(56)对比文件
CN 1749988 A,2006.03.22,
刘昆宏.多分类器集成系统在基因微阵列数据分析中的应用.《中国博士学位论文全文数据库 信息科技辑》.2009,第2009年卷(第6期),I140-10.
Daniel Glez-Pena等.A simulated annealing-based algorithm for iterative class discovery using fuzzy logic for informative gene selection.《JOURNAL OF INTEGRATED OMICS》.2011,第1卷(第1期),第66-77页.
审查员 崔小利

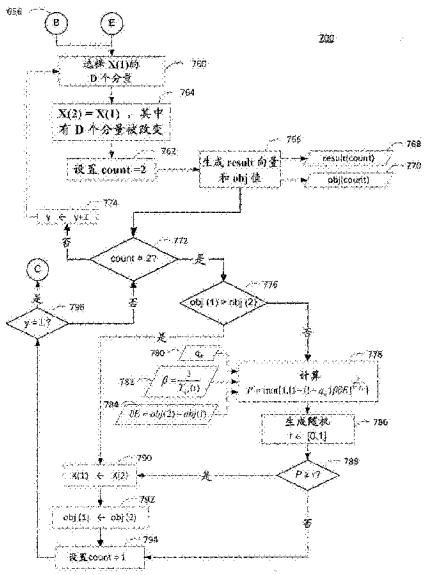
权利要求书2页 说明书16页 附图10页

(54)发明名称

一种生成生物标记签名的系统及方法

(57)摘要

本发明公开涉及利用集成的双融合和广义模拟退火技术生成生物标记签名的系统及方法。本文所描述的是用于利用融合分类技术分类数据集的系统和方法。通过把机器学习技术应用到训练数据集而迭代地生成分类器,并且训练类集是通过根据分类器分类训练数据集中的元素而生成的。目标值是基于训练类集计算的,并且与不同分类器相关联的目标值进行比较,直到达到期望的迭代数量,并且输出最终的训练类集。



CN 104584022 B

1.一种由处理器执行的、把数据集分类成两类或更多类的计算机实现的方法，其中所述数据集包括基因组数据，每个基因组数据对应于多种生物状态类之一，该方法包括：

(a)接收具有已知标签集的训练数据集，其中所述标签标识基因组数据的生物状态类；

(b)通过向所述训练数据集应用第一机器学习技术，生成用于所述训练数据集的第一分类器，其中所述第一机器学习技术识别第一分类方法集，其中每一分类方法对所述训练数据集进行投票；

(c)根据所述第一分类器分类所述训练数据集中的元素以获得用于所述训练数据集的第一预测标签集；

(d)根据所述第一预测标签集和所述已知标签集，计算第一目标值；

(e)对于多次迭代中的每一次，执行以下步骤(i)-(v)：

(i)通过向所述训练数据集应用第二机器学习技术，生成用于所述训练数据集的第二分类器，其中所述第二机器学习技术识别不同于第一分类方法集的第二分类方法集，其中每一分类方法对所述训练数据集进行投票；

(ii)根据所述第二分类器分类所述训练数据集中的元素以获得用于所述训练数据集的第二预测标签集；

(iii)根据所述第二预测标签集和所述已知标签集，计算第二目标值；

(iv)比较所述第一与第二目标值以确定所述第二分类器是否胜过所述第一分类器；

(v)当所述第二分类器胜过所述第一分类器时，用所述第二预测标签集替代所述第一预测标签集并且用所述第二目标值替代所述第一目标值，并且返回步骤(i)；及

(f)当已经达到期望次数的迭代时，输出所述第一预测标签集。

2.如权利要求1所述的方法，其中所述训练数据集是通过从聚合训练数据集中选择训练数据样本的子集而形成的，所述方法还包括：

自举所述聚合训练数据集以生成多个附加训练数据集，及

对每个附加训练数据集重复步骤(a)至(f)。

3.如权利要求2所述的方法，其中所述自举是用平衡的样本或者不用平衡的样本执行的。

4.如权利要求1-3中任何一项所述的方法，还包括：

识别产生输出的第一预测标签集的分类器；

在测试数据集中选择样本，所述测试数据集不同于所述训练数据集并且不具有已知标签集；及

利用识别出的分类器来为所选样本预测标签。

5.如权利要求1-3中任何一项所述的方法，其中：

通过应用第一随机向量来选择聚合分类方法集的子集来获得所述第一分类方法集；

所述第一随机向量包括与所述聚合分类方法集对应的一组二进制值；

每个二进制值指示所述聚合分类方法集中的对应分类方法是否包括在所述第一分类方法集中；以及

通过应用包括不同的一组二进制值的第二随机向量来获得所述第二分类方法集。

6.如权利要求5所述的方法，其中所述第一随机向量包括指示是否执行平衡的自举、多次自举、分类方法列表、基因列表或其组合的标志变量。

7.如权利要求1-3中任何一项所述的方法，其中所述第二目标值对应于Matthew相关系数，所述Matthew相关系数是根据所述第二预测标签集和所述已知标签集评定的。

8.如权利要求5所述的方法，其中计算所述第二目标值的步骤包括实现模拟退火方法。

9.如权利要求8所述的方法，其中所述模拟退火方法包括更新所述第一随机向量的一个或多个值以获得所述第二随机向量。

10.如权利要求9所述的方法，其中更新所述第一随机向量的一个或多个值包括：

随机地更新所述第一随机向量的每个元素以获得所述第二随机向量。

11.如权利要求1-3中任何一项所述的方法，还包括：(1)当所述第二目标值小于所述第一目标值时，及(2)如果所述第二目标值大于所述第一目标值，当随机值小于根据所述第一目标值和所述第二目标值计算的概率值时，确定所述第二分类器胜过所述第一分类器。

12.如权利要求11所述的方法，其中所述概率值是根据控制参数q、所述第一目标值、所述第二目标值和根据冷却公式计算的温度值计算的。

13.如权利要求1-3中任何一项所述的方法，其中所述第二分类器选自包括线性判别分析、基于支持向量机的方法、随机森林法和k-近邻法的组。

14.如权利要求5所述的方法，其中：

所述多次迭代包括第一组迭代和第二组迭代；并且

每个后一第二随机向量与前一第二随机向量相差的量对于第一组迭代而言比对于第二组迭代而言大。

15.如权利要求14所述的方法，其中对于第一组迭代和第二组迭代中的每次迭代，

后一第二随机向量的第一子集被选择为与前一第二随机向量的相应第一子集相同，

后一第二随机向量的第二子集被选择为与前一第二随机向量的相应第二子集不同，

用于第一组迭代的第一子集的大小比用于第二组迭代的第一子集的大小小，

用于第一组迭代的第二子集的大小比用于第二组迭代的第二子集的大小大。

16.如权利要求15所述的方法，其中用于第一组迭代的第二子集的大小是第二随机向量的长度的20％，而用于第二组迭代的第二子集的大小是1。

17.一种存储有计算机可读指令的计算机可读存储介质，所述计算机可读指令在包括至少一个处理器的计算机化系统中被执行时，使所述处理器执行权利要求1-16中任何一项的方法。

18.一种包括处理设备的计算机化系统，其中所述处理设备利用非临时性计算机可读指令来配置，所述计算机可读指令在被执行时，使所述处理设备执行权利要求1-16中任何一项的方法。

一种生成生物标记签名的系统及方法

[0001] 对相关申请的引用

[0002] 本申请按照美国法典第35章119条要求于2012年6月21日提交且标题为“Systems and Methods for Generating Biomarker Signatures with Integrated Dual Ensemble and Generalized Simulated Annealing Techniques(利用集成的双融合和广义模拟退火技术生成生物标记签名的系统及方法)”的美国临时专利申请No.61/662,812 的优先权，该申请的全部内容通过引用被结合于此。

技术领域

[0003] 本公开一般地涉及利用集成的双融合和广义模拟退火技术生成生物标记签名的系统及方法。

背景技术

[0004] 在生物医药领域，识别指示特定生物状态的物质，即，生物标记，是重要的。随着基因组学和蛋白质组学新技术的出现，生物标记在生物发现、药物开发和健康保健中变得越来越重要。生物标记不仅对许多疾病的诊断和预后有用，而且对理解用于治疗开发的基础也有用。生物标记的成功且有效识别可以加速新药物的开发过程。随着治疗与诊断和预后的结合，生物标记识别还将提高当前的医疗质量，因而在药理学、药物基因组学和药物蛋白质组学的使用中发挥重要作用。

[0005] 包括高吞吐量筛选的基因组和蛋白质组分析提供了关于在细胞中表达的蛋白质的数量和形式的大量信息，并且提供了为每个细胞识别所表达的特定细胞状态的蛋白质特征概要的可能性。在某些情况下，这种细胞状态可能是与疾病相关联的异常生理反应的特征。因此，识别并比较来自具有疾病的患者的细胞状态和来自正常患者的相应细胞的细胞状态能够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。研究人员已试图开发用于将这些数据集组织成对不同种群的个体可重复诊断的模式的方法。一种方法是从多个来源汇总数据，以形成合并的数据集，然后将数据集分成发现/训练集和测试/验证集。但是，转录概要数据和蛋白质表达概要数据常常都是由相对于可用数量样本的大量变量特征化的。

[0007] 在患者组或对照组的标本的表达谱之间观察到的差异通常被若干个因素所掩盖，包括疾病人群或对照人群内的生物变异性或未知的亚表型、由于研究协议中的差异所导致的特定于位点(site)的偏差、标本处理、由于仪器状态(例如，芯片批次等)的差异所导致的偏差，以及由于测量误差所导致的变化。一些技术试图对数据样本中的偏差 (这可能由于例如在数据集中表示的一类样本比另一类样本多)进行校正。

[0008] 已经开发了若干种基于计算机的方法来寻找最好地解释疾病和对照样本之间差异的一组特征(标记)。一些早期的方法包括统计测试，诸如LIMMA、逻辑回归技术和诸如支持向量机(SVM)的机器学习方法，其中LIMMA是FDA批准的用于识别与乳腺癌相关的生物标

记的MammaPrint技术。一般而言，从机器学习的角度来看，生物标记的选择通常是对分类任务的特征选择问题。但是，这些早期的解决方案面临几个缺点。由于受试者的包括和排除会导致不同的签名，因此，通过这些技术生成的签名常常是不可复制的。这些早期的解决方案还因为它们对具有小样本尺寸和高维度的数据集进行操作而产生许多假阳性签名并且是不健壮的。

[0009] 因此，需要用于识别用于临床诊断和/或预后的生物标记，并且更一般地，用于识别能够用来将数据集中的元素分类成两类或更多类的数据标记的改进技术。

发明内容

[0010] 本文所描述的是用于识别能够用来将数据集中的元素分成两类或更多类的数据标记的系统、计算机程序产品和方法。特别地，申请人已经认识到，方法与基因组数据的组合能够比单独的方法本身提供对测试数据的更好预测。本文所描述的计算机系统和计算机程序产品实现包括一种或多种用于将元素分成两个或更多类的此类技术的方法。特别地，生物标记签名利用集成的双融合(dual ensemble)和模拟退火 (simulated annealing)技术生成。该技术包括重新采样数据集并利用双融合方法预测表型。特别地，本文所描述的系统、计算机程序产品和方法包括形成指示一组分类方法和数据样本的随机向量。随机向量被迭代地扰动，并且计算对应于不同扰动的不同目标值。

[0011] 在某些方面，本文所描述的系统和方法包括由处理器执行的、用于将数据集分成两类或更多类的装置和方法。方法可以包括接收训练数据集。训练数据集可以通过将聚合数据集分成发现(训练)集和验证(测试)集来确定。例如，聚集数据集可以包括从多个来源汇集在一起的数据，并且聚集数据集可以被随机地分成训练和测试数据集。方法还可以包括通过向训练数据集应用第一机器学习技术而生成用于训练数据集的第一分类器。例如，机器学习技术可以对应于支持向量机(SVM)或者用于特征选择的任何合适的技术。通过根据第一分类器分类训练数据集中的元素而生成第一训练类集(class set)。特别地，第一分类器可以对应于将数据集中的每个样本分配给生理状态的 (诸如像患病的或无病的)分类规则。第一分类器可以组合多种分类方法，诸如SVN、基于网络的SVM、基于神经网络的分类器、逻辑回归分类器、基于决策树的分类器、利用线性判别分析技术的分类器、随机森林分析技术、任何其它合适的分类方法，或其任意组合。

[0012] 第一目标值基于训练类集来计算。特别地，可以使用二进制广义模拟退火法方法来计算该目标值。随机向量可以包括定义要使用的分类技术的一组参数作为其元素。由随机向量定义的技术被用来计算第一目标值。然后，对于多次迭代，第二机器学习技术被应用到训练数据集，以生成用于该训练数据集的第二分类器，并且通过根据该第二分类器分类训练数据集中的元素而生成第二训练类集。特别地，第二分类器可以通过随机扰动用来定义第一分类器的随机向量并利用随机向量的随机扰动定义第二分类器而生成。此外，计算基于第二训练类集的第二目标值，并比较第一和第二目标值。基于第一和第二目标值之间的比较，第一训练类集可以用第二训练类集替换，并且第一目标值可以被第二目标值替换，然后开始下一轮迭代。迭代一直重复到到达期望的迭代次数，然后输出第一训练类集。

[0013] 在上述方法的某些实施例中，该方法的步骤对多个训练数据集重复，其中多个训练数据集中的每个训练数据集都通过自举 (bootstrapping)聚合训练数据集而生成。自举

可以利用平衡的样本或者不利用平衡的样本执行。利用平衡的样本还是不利用平衡的样本进行自举可以由随机向量中的二进制元素确定，其值可以在随机向量被扰动时更新。其它的自举参数可以作为元素包括在随机向量中，诸如是否用替换或不用替换或多个自举从聚合样本集中采样样本的子集。在方法的某些实施例中，在测试数据集合中选择样本，并且对应于输出第一个训练类集合的分类器被用来预测与所选样本相关联的值。在方法的某些实施例中，第二分类器是通过应用随机向量以识别用于与第二分类器相关联的分类方案的参数而生成的，随机向量包括至少一个二进制值。在方法的某些实施例中，随机向量的参数包括指示是否执行平衡的自举、多个自举、分类方法列表、基因列表或其组合的标志变量。

[0014] 在方法的某些实施例中，计算第二目标值的步骤是基于 Matthew相关系数。特别地，目标值可以对应于1与结果的 Matthew相关系数之差。Matthew相关系数是可以用作复合性能分数的性能度量。在方法的某些实施例中，计算第二目标值的步骤包括实现二进制广义模拟退火方法。在方法的某些实施例中，二进制广义模拟退火方法包括局部扰动随机向量的一个或多个值，以识别用于分类方案的参数。在方法的某些实施例中，局部扰动随机向量的一个或多个值包括随机更新随机向量的每个元素以获得更新的随机向量、利用更新的随机向量计算更新的第二目标值，以及基于概率值和随机数之间的比较接受更新的第二目标值。在方法的某些实施例中，局部扰动随机向量的一个或多个值包括对每次迭代改变随机向量的一个元素。

[0015] 在方法的某些实施例中，用第二训练类集代替第一训练类集和用第二目标值代替第一目标值的步骤基于冷却公式。特别地，可能期望通过对随机向量执行大的扰动来减小二进制广义模拟退火方法中的目标值。在模拟退火中，人工温度值被逐渐降低，以模拟冷却。访问分配在模拟退火中用来模拟从一个点(即，用于随机向量的第一组值) 到另一个点(即，用于随机向量的第二组值)的试跳距离。基于第二目标值是否小于第一个目标值并基于接受概率来接受试跳。二进制广义模拟退火方法被用来识别最小化目标值的全局最小数。在方法的某些实施例中，第二分类器选自包括线性判别分析、基于支持向量机的方法、随机森林法，以及k-近邻方法的组。

[0016] 本发明的计算机系统包括用于实现如上所述的方法的各种实施例的装置。例如，描述了计算机程序产品，该产品包含计算机可读指令，当计算机可读指令在包括至少一个处理器的计算机化系统中执行时，使处理器执行上述任意方法的一个或多个步骤。在另一个例子中，描述了计算机化系统，该系统包括利用非临时性计算机可读指令配置的处理器，当计算机可读指令被执行时，使处理器执行上述任意方法。本文所描述的计算机程序产品和计算机化的方法可以在具有一个或多个计算设备的计算机化系统中实现，其中每个计算设备包括一个或多个处理器。一般而言，本文所描述的计算机化系统可以包括一个或多个引擎，引擎包括处理器或设备，诸如计算机、微处理器、逻辑装置或者利用硬件、固件和软件配置以便执行本文所述的一个或多个计算机化的方法的其它设备或处理器。这些引擎中的任何一个或多个都可以与任何一个或多个其它引擎物理地分离，或者可以包括多个物理上可分离的组件，诸如在共同或不同电路板上的独立的处理器。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。引擎有时可以互连，并且有时进一步连接到一个或多个数据库，包括扰动数据库、衡量标准(measurable)数据库、实验数据数据库和文献数据库。本文所描述的计算机化系统可以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。这种实现可能适合于经多个通信系统的分布式计算。

附图说明

[0017] 当结合附图考虑以下详细描述时，本公开内容进一步的特征、其性质和各种优点将显而易见，附图中相同的标号贯穿全文都指相同的部分，并且其中：

[0018] 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；

[0019] 图2是描绘数据样本的分类以及分类规则的确定的图；

[0020] 图3是双融合方法的流程图；

[0021] 图4是用于构建数据集的方法的流程图；

[0022] 图5是用于生成结果向量和目标值的方法的流程图；

[0023] 图6是用于初始化二进制广义模拟退火方法的方法的流程图；

[0024] 图7是用于减小二进制广义模拟退火方法中的目标值的方法的流程图；

[0025] 图8是用于进一步减小二进制广义模拟退火方法中的目标值的方法的流程图；

[0026] 图9是计算设备的框图，其中计算设备诸如图1的系统的任何组件；及

[0027] 图10是训练数据集中的基因签名的热图(heatmap)。

具体实施方式

[0028] 为了提供对本文所描述的系统和方法的全面理解，现在将描述某些说明性实施例，包括用于识别基因生物标记签名的系统和方法。但是，本领域普通技术人员应当理解，本文所描述的系统和方法可以针对其它合适的应用，诸如任何数据分类应用，而被调整和修改，并且这种其它的添加和修改将不背离其范围。一般而言，本文所描述的计算机化系统可以包括一个或多个引擎，引擎包括处理器或设备，诸如计算机、微处理器、逻辑设备或者利用硬件、固件和软件配置以便执行本文所描述的一种或多种计算机化方法的其它设备或处理器。

[0029] 本文所描述的系统和方法包括用于利用集成的双融合和模拟退火技术生成生物标记签名的技术。该技术涉及重新采样数据集并利用双融合方法预测表型。特别地，本文所描述的系统和方法包括构成指示一组分类方法、数据样本的随机向量，并且迭代地扰动随机向量并计算对应于不同扰动的不同目标值。

[0030] 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，其中可以实现本文所公开的分类技术。系统100包括生物标记生成器 102和生物标记整合器104。系统100还包括中央控制单元(CCU) 101，用于控制生物标记生成器102和生物标记整合器104的操作的某些方面。在操作过程中，诸如基因表达数据的数据在生物标记生成器102被接收。生物标记生成器102处理该数据，以生成多个候选生物标记和对应的误差率。生物标记整合器104接收这些候选生物标记和误差率并且选择具有最佳性能度量和尺寸的合适生物标记。

[0031] 生物标记生成器102包括若干个组件，用于处理数据并生成一组候选生物标记和候选误差率。特别地，生物标记生成器102包括用于将数据分成训练数据集和测试数据集的数据预处理引擎110。生物标记生成器102包括分类器114，用于接收训练数据集和测试数据集并将测试数据集分成两类或更多类中的一个(例如，疾病数据和非疾病的、易感和免疫，等等)。生物标记生成器102包括分类器性能监控引擎116，用于确定当分类器应用到由数据

预处理引擎110所选择的测试数据时分类器的性能。分类器性能监控引擎116基于分类器识别候选生物标记(例如,数据集中对分类最重要的元素的组件)并且为一个或多个候选生物标记生成性能度量,这可以包括候选误差率。生物标记生成器102还包括生物标记储存器118,用于存储一个或多个候选生物标记与候选性能度量。

[0032] 生物标记生成器可以由CCU 101进行控制,CCU 101又可以被自动控制或者被用户操作。在某些实施例中,生物标记生成器102可以操作成当每次将数据随机地分成训练和测试数据集时生成多个候选生物标记。为了生成这样的多个候选生物标记,生物标记生成器102 的操作可以被迭代多次。CCU 101可以接收一个或多个系统迭代参数,包括候选生物标记的期望数量,这又可以用来确定生物标记生成器102的操作可以被迭代的次数。CCU 101也可以接收其它的系统参数,包括期望的生物标记尺寸,这可以代表生物标记中的组件的数量(例如,生物标记基因签名中基因的数量)。生物标记尺寸信息可以被分类器性能监控引擎116使用,用于从训练数据中生成候选生物标记。生物标记生成器102以及尤其分类器114的操作将参考图2-8 更详细地进行描述。

[0033] 生物标记生成器102生成一个或多个候选生物标记和候选误差率,这被生物标记整合器104使用来产生健壮的生物标记。生物标记整合器104包括生物标记一致(consensus)引擎128,该引擎接收多个候选生物标记,并且生成具有跨多个候选生物标记最频繁出现的基因的新生物标记签名。生物标记整合器104包括误差计算引擎130,用于确定跨多个候选生物标记的总误差率。类似于生物标记生成器102,生物标记整合器104也可以由CCU 101控制,CCU 101又可以被自动控制或者被用户操作。CCU 101可以接收和/或确定用于最小生物标记尺寸的合适阈值,并且使用这个信息来确定操作生物标记生成器 102和生物标记整合器104两者的迭代次数。在一种实施例中,在每次迭代期间,CCU 101将生物标记的尺寸减一,并且迭代生物标记生成器102和生物标记整合器104两者,直到达到阈值。在这种实施例中,生物标记一致引擎128对每次迭代输出新的生物标记签名和新的总误差率。因此,生物标记一致引擎128输出一组新生物标记签名,每个生物标记签名都具有从阈值到最大生物标记尺寸变化的不同尺寸。生物标记整合器104还包括生物标记选择引擎126,该引擎检查这些新生物标记签名当中每一个的性能度量或误差率并且选择最优的生物标记用于输出。

[0034] 数据预处理引擎110接收一个或多个数据集。一般而言,数据可以表示样本中多个不同基因的表达值,和/或多种表型特征,诸如任何生物上显著的分析物的水平。在某些实施例中,数据集可以包括用于疾病状态和用于对照状态的表达水平数据。如本文所使用的,术语“基因表达水平”可以指由基因,例如RNA或多肽,编码的分子的量,或者miRNA的量。mRNA分子的表达水平可以包括mRNA的量(这是由编码该mRNA的基因的转录活性确定的)和mRNA的稳定性(这是由mRNA的半衰期确定的)。基因表达水平还可以包括对应于由基因编码的给定氨基酸序列的多肽的量。因而,基因的表达水平可以对应于从该基因转录的mRNA的量、由该基因编码的多肽的量、或者二者兼有。基因的表达水平可以通过不同形式的基因产物的表达水平被进一步分类。例如,由基因编码的RNA分子可以包括差异表达的剪接变异体,具有不同开始或停止位点的转录,和/或其它差异处理的形式。由基因编码的多肽可以涵盖多肽的裂解和/或修改的形式。多肽可以通过磷酸化、脂化、异戊二烯化、硫酸化、羟基化、乙酰化、核糖基化、法尼基化、添加碳水化合物等等进行修改。另外,可以存在具有给定

类型修改的多种形式的多肽。例如，多肽可以在多个位点被磷酸化并且表达差异磷酸化蛋白质的不同水平。每个这种修改的多肽的水平可以被单独地确定并在数据集中表示。

[0035] 分类器114从数据预处理引擎110接收一组或多组数据。在某些实施例中，分类器114生成分类规则来分类数据。图2用图形描绘了这种分类规则200。分类器114可以应用分类规则把数据集分配到两个类中的任意一个。例如，分类器114可以应用分类把数据集分配到疾病或对照。

[0036] 在某些实施例中，如关于图3-8中所描述的，分类器114使用与广义模拟退火方法相结合的双融合技术来产生分类规则。特别地，分类器114可以组合多种分类方法，诸如支持向量机(SVM)、基于网络的SVM、基于神经网络的分类器、逻辑回归分类器、基于决策树的分类器、采用线性判别分析技术的分类器、和/或随机森林分析技术、或者任何其它合适的分类方法。融合分类策略可以跨多种不同的分类方法使用投票过程来识别最优分类。通过结合多种分类方法，融合技术减少了过度拟合到小数据集的可能。以这种方式，与其它技术相比，通过利用融合技术可以更有效地使用小数据集。此外，与利用单一分类方法相比，利用多种分类方法的融合允许增强的分类，当融合中的多种分类方法彼此不同时尤其如此。

[0037] 此外，从数据预处理引擎110接收的数据可以被扰动，以便在提供更好的分类准确性的同时进一步增加整体多样性。数据扰动的例子联系图4，7和8更详细地描述。

[0038] 如本文所述，分类器114使用融合技术和广义模拟退火方法来生成分类规则并且是关于生物信息学中的应用来描述的。但是，本文所描述的系统和方法可以一般性地应用到任何大规模的计算技术，诸如特征选择或提取。

[0039] 分类器性能监控引擎116可以利用合适的性能度量来分析分类器 114的性能。特别地，当分析分类器114的性能时，分类器性能监控引擎116可以分析一个或多个候选生物标记的健壮性或性能。在某些实施例中，性能度量可以包括误差率。性能度量还可以包括正确预测的个数除以尝试的预测总数。在不背离本公开内容范围的情况下，性能度量可以是任何适当的测量。候选生物标记和对应的性能度量可以存储在生物标记储存器118中。

[0040] 在某些实施例中，细胞或组织中的基因表达水平可以由基因表达谱(gene expression profile)来表示。基因表达谱可以指在诸如细胞或组织的标本中的基因表达水平的特征表示。在来自个体的标本中基因表达谱的确定代表该个体的基因表达状态。基因表达谱反映细胞或组织中信使RNA或多肽或者其由一个或多个基因编码的形式的表达。表达谱一般可以指生物分子(核酸，蛋白质，碳水化合物)谱，它显示不同细胞或组织中的不同表达模式。表示基因表达谱的数据样本可以存储为表达水平的向量，向量中的每个条目对应于特定的生物分子或其它生物实体。

[0041] 在某些实施例中，数据集可以包括表示样本中多个不同基因的基因表达值的元素。在其它实施例中，数据集可以包括表示由质谱分析法检测到的峰值的元素。一般而言，每个数据集可以包括数据样本，其中每个数据样本对应于多种生物状态类之一。例如，生物状态类可以包括，但不限于：疾病在样本源(即，从其获得样本的患者)中存在/不存在；疾病的阶段；疾病的风险；疾病复发的可能性；在一个或多个基因座的共享基因类型(例如，常见的HLA单倍型；基因中的突变；基因的修改，诸如甲基化，等等)；暴露于药剂(例如，诸如有毒物质或潜在的有毒物质、环境污染物、候选药物，等等)或条件(温度、pH，等等)；人群统计特征(年龄、性别、体重；家族病史；先前存在的病史，等等)；对药剂的耐药性，对药剂的敏感性

(例如,对药物的反应性),等等。

[0042] 数据集可以是彼此独立的,以减少最终分类器选择中的收集偏差 (collection bias)。例如,它们可以从多个来源中收集,并且可以在不同的时间和从不同的位置利用不同的排除或包括标准进行收集,即,当考虑定义生物状态类的特征之外的特征时,数据集可以是相对异质的。有助于异质性的因素包括,但不限于:由于性别、年龄、种族导致的生物差异;由于饮食、运动、睡眠行为导致的个体差异;以及由于用于血液处理的临床方案导致的样本处理差异。但是,生物状态类可以包括一个或多个共同特征(例如,样品来源可以表示具有疾病和相同性别或者一个或多个其它共同人群特征的个体)。

[0043] 在某些实施例中,来自多个源的数据集通过在不同时间和/或在不同条件下从同一患者人群收集样本来生成。

[0044] 在某些实施例中,多个数据集从多个不同的临床试点获得,并且每个数据集包括在每个单独试点获得的多个患者样本。样本类型包括,但不限于:血液、血清、血浆、乳头抽取物、尿、眼泪、唾液、脊髓液、淋巴液、细胞和/或组织裂解物、激光显微切割的组织或细胞样本、嵌入的细胞或组织(例如,在石蜡块中或冷冻的);新鲜或存档的样本(例如,来自尸检)。样本可以从例如试管中的细胞或组织培养物得到。作为选择,样本可以从有生命的生物体或者从种群生物体,诸如单细胞生物体,得到。

[0045] 在一个例子中,当识别用于特定癌症的生物标记时,血液样本可以从由位于两个不同测试点的独立组选择的受试者中收集,由此提供将从其开发独立数据集的样本。

[0046] 在一些实现中,训练和测试集是由数据预处理引擎110生成的,其中数据预处理引擎110接收批量数据并将该批量数据分割成训练数据集和测试数据集。在某些实施例中,数据预处理引擎110随机地将数据分成这两个组。对于预测类别和产生健壮的基因签名,随机地分割数据会是期望的。在其它实施例中,数据预处理引擎110基于数据的类型或标签将数据分割成两组或更多组。一般而言,在不背离本公开内容范围的情况下,数据可以如所期望的那样以任何合适的方式分成训练数据集和测试数据集。训练数据集和测试数据集可以具有任何合适的尺寸并且可以是相同或不同的尺寸。在某些实施例中,数据预处理引擎110可以在将数据分成训练和测试数据集之前丢弃一块或多块数据。在某些实施例中,数据预处理引擎110可以在任何进一步的处理之前从训练数据集和/或测试数据集中丢弃一个或多块数据。

[0047] 分类器114可以从数据预处理引擎110接收一个或多个候选生物标记和一个或多个数据集。分类器114可以应用分类规则来将数据集分配给两种类别中的任何一个。例如,分类器114可以应用分类来将数据集分配给疾病或对照。在某些实施例中,分类器114可以包括支持向量机(SVM)分类器、基于网络的SVM、基于神经网络的分类器、逻辑回归分类器、基于决策树的分类器、采用线性判别分析技术的分类器、和/或随机森林分析技术。分类器114和其相应引擎的操作参考图2-8中更详细地描述。

[0048] 分类器性能监控引擎116可以利用合适的性能度量分析分类器 114的性能。特别地,当分析分类器114的性能时,分类器性能监控引擎116可以分析一个或多个候选生物标记的健壮性或性能。在某些实施例中,性能度量可以包括误差率。性能度量还可以包括正确预测的个数除以尝试的预测总数。在不背离本公开内容范围的情况下,性能度量可以是任何合适的测量。候选生物标记和对应的性能度量可以存储在生物标记储存器118中。

[0049] 如前面所指出的，CCU 101还可以控制生物标记整合器104的操作，用于基于在生物标记生成器102生成并存储的候选生物标记生成合适且健壮的生物标记。生物标记整合器104包括生物标记一致引擎128，该引擎从生物标记储存器118接收一个或多个候选生物标记。生物标记一致引擎128可以选择在一个或多个候选生物标记中频繁出现的基因用作新生物标记签名。新生物标记签名可以包括N个基因，其中N是生物标记的期望尺寸、生物标记的最大允许尺寸、生物标记的最小允许尺寸或者最大与最小尺寸之间的尺寸。在某些实施例中，数字N可以是用户可选择的并且可以是可按需调节的。

[0050] 图3是由分类器114用来利用投票方法预测表型类的方法300的流程图。如所示出的，方法300包括步骤：构建K个数据集(步骤 302)、识别M个分类方法(步骤306)、以及在K个数据集的每一个中识别G个样本(步骤312)。方法300还包括三个迭代循环，包括在K个数据集、M个分类方法、以及G个样本上进行迭代，其中 G是测试数据集的样本大小。特别地，在每次迭代中，分类方法j被应用到数据集i中的样本l，以预测表型(步骤318)，其中i＝1，2，...，K，j＝1，2，...，M，以及l＝1，2，...G。

[0051] 在步骤302，分类器114构建K个数据集。分类器可以使用在图 4中所描绘的方法来构建K个数据集。特别地，分类器114可以使用自举聚合(bootstrapping aggregation)方法来构成完整数据集的多个数据集。在步骤304，代表应用到数据集的标签的数据集迭代参数i被初始化为1。

[0052] 在步骤306，分类器114识别M个分类方法。分类器114可以从外部源接收分类方法，或者分类方法可以基于一些输入由分类器 114生成。作为例子，分类器114可以基于方法308的列表识别M 个分类方法。方法的例子包括线性判别分析、基于支持向量机的方法、随机森林法(Breiman，Machine Learning，45(1)：5-32(2001))、PAMR(Tibshirani等人，Proc Natl Acad Sci USA，99(10)：6567- 6572(2002))或k-近邻方法(Bishop，Neural Network for Pattern Recognition，ed.O.U.Press，1995)。任何数量的分类方法都可以使用并考虑。在步骤310，表示应用到分类方法的标签的方法迭代参数j被初始化为1。在步骤316，代表应用到数据样本的标签的样本迭代参数l被初始化为1。每个数据样本可以表示人、基因、或任何其它合适的数据点。

[0053] 在步骤312，分类器114选择测试数据集中的第l个样本，并且在步骤318，分类器114对数据集i应用分类方法j来构建分类器并预测测试数据中的样本l。样本l的预测可以对应于表型的预测。在一些实施例中，表型可以是标志变量(即，如果预测出人表达表型，则为1，否则为0)。但是，一般而言，表型可以取任何数量的值。特别地，表型预测可以存储为三维矩阵P(i，j，l)320中的值。

[0054] 在决定框322，分类器114确定最后的数据集是否已被考虑，或等效地，是否i＝K。如果i小于K，则分类器114在步骤324递增数据集迭代参数i并返回到步骤318，以便为新数据集预测表型。

[0055] 在所有K个数据集都被考虑之后，分类器114前进到决定块326，以确定最后的分类方法是否已被应用，或等效地，是否j＝M。如果j 小于M，则分类器114在步骤328递增方法迭代参数j并返回到步骤 318，以便为新分类方法预测表型。

[0056] 在所有K个数据集都被考虑并且所有M个分类方法都已被应用之后，分类器114具有用于当前数据样本l K×M个表型预测。这些表型预测可以被认为是选票，并且可以使用任

何种类的选票计数方法来得出代表K×M表型预测集合的复合选票。

[0057] 在决定框332，分类器确定是否所有G个数据样本都已被考虑，或等效地，是否l＝G。

[0058] 图4是用于构建数据集的方法400的流程图，并且可以在图3中的步骤302被分类器114使用。一般而言，方法400提供了生成作为更大数据集合的每个子集的多个数据集的方法。数据子集可以通过自举聚合(“装袋”)方法构成，这种方法涉及在大的数据集中随机选择样本的子集。样本的子集可以利用或者不利用替换来选择。如所示出的，方法400包括步骤：接收数据(步骤440)和确定是否期望不利用替换来执行自举(决定框454)。如果是这样，则可以从每个类中随机地选择W个样本(步骤456)来构成数据集。作为选择，可以利用替换从训练数据中随机地选择H个样本(步骤460和466)来构成数据集。用于H的值可以对应于训练数据集的样本大小。重复上述步骤，直到关于图3所描述的每个数据集i都已被考虑。

[0059] 在步骤440，分类器114接收数据。数据可以包括被分成两类的样本(即，类1样本442和类2样本444)、自举参数446、以及结果产生的数据集i(即，数据子集)的大小与类(即，类1或类2) 的大小之比s448。作为例子，自举参数446可以包括指示是利用还是不利用替换进行自举的变量以及自举数据集的数量(即，K)。数据442，444，446和448可以被分类器114用来构建这K个数据集。

[0060] 在步骤452，数据集迭代参数i被初始化为1。迭代参数i表示应用到数据集的标签。

[0061] 在决定框454，分类器114确定是否期望利用平衡的样本来自举。特别地，分类器114可以使用诸如自举参数446的变量来确定利用平衡的样本进行自举是否是期望的。一般而言，利用平衡的样本进行自举确保每个采样点跨所有K个数据集出现的总数是相同的。

[0062] 如果平衡的自举是期望的，则分类器114前进到步骤450，以确定数据集尺寸W。特别地，尺寸W可以依赖于比率s 448，例如，诸如W＝min{size(类1的样本)，size(类2的样本)}＊s。特别地，比率s可以是0和1之间的值。在步骤456，来自训练数据集的W个样本利用平衡的样本随机地选择，从而构成数据集i 458。当迭代参数i大于1时，在步骤456对W个样本的选择可以依赖于之前形成的数据集，使得自举被平衡。

[0063] 作为选择，如果利用平衡的样本进行自举不是期望的，则分类器 114前进到步骤460，以便利用替换从训练数据集中随机地选择H个样本。所选择的样本构成数据集i 464。

[0064] 如在图4中所描绘的，平衡的自举导致数据集具有尺寸W，而不利用平衡的样本自举数据导致数据集具有尺寸H。但是，一般而言，方法的任意合适组合都可以使用，诸如对具有尺寸W的数据集不利用平衡的样本进行自举，或对具有尺寸H的数据集进行平衡的自举。此外，也可以使用不利用替换方法的自举。

[0065] 在已经构成当前数据集i之后，分类器114前进到决定块470，以确定是否已经构成最后的数据集，或等效地，是否i＝K。如果不是，则在步骤472，递增数据集迭代参数i，并且分类器114前进到决定块454，以开始构成下一数据集。

[0066] 图5是用于生成结果向量和目标值的方法的流程图。一般而言，方法500提供了计算对应于随机向量X的目标值的途径。如在方法 500中所描绘的，随机向量X是二进制向量X并且包括关于是否利用替换进行自举的信息(506)、自举的次数(510)、分类方法列表(514)以及数据样本列表(518)。基于这些数据，构成预测矩阵 (步骤520)，并且确定主类(步骤524)。分类器114在数据样本上进行迭代，直到所有数据样本都已被考虑，并且基于为

数据样本确定的主类计算目标值(步骤532)。

[0067] 在步骤502,分类器114接收二进制随机向量X。在一个例子中,向量X可以是二进制值的列表。二进制值可以指示是否执行平衡的自举、自举的次数(即,K)、分类方法列表、和/或基因列表。特别地,自举的次数可以取零值或非零值(即,例如60)。在这种情况下,向量X中对应于自举次数的二进制值可以指示自举的此时是零还是非零。该随机值可以由随机值发生器或者用于生成随机值的任何其它合适的方法生成。如在本文所描述的,随机向量X是二进制向量,这意味着向量中的每个值是两个值中的一个(即,0或1)。但是,一般而言,随机向量X中的值可以是任何数量的值中的一个。分类器114基于向量X中的随机值识别各种参数。作为例子,分类器114识别用于标志506的值,该值在步骤504指示是否利用平衡的样本进行采样、在步骤508指示自举次数510、在步骤512指示分类方法列表514,以及在步骤516指示基因列表518。

[0068] 基于识别出的各种参数,在步骤520,分类器114生成预测矩阵。

[0069] 在步骤522,表示应用到数据样本的标签的样本迭代参数l被初始化为1。

[0070] 在步骤524,分类器114确定主类P(.,.,1)。特别地,分类器 114可以通过方法300中的步骤302-330解析,以识别K×M个表型预测,并对K×M个预测采用多数选票,以确定主类P(.,.,1)。一般而言,用于基于K×M预测集合生成复合预测的任何其它合适的方法都可以用来确定主类。主类可以存储为结果向量526中的条目。

[0071] 在决定框528,分类器114确定样本迭代参数l是否等于数据样本G的总数。如果不是,则迭代参数l在步骤530递增,并且为下一数据样本确定主类。

[0072] 在已经为G个样本的集合中的每个样本确定主类之后,分类器 114前进到步骤532,以计算目标值。目标值可以基于结果向量526 中条目的结果集合来计算。特别地,合成性能分数可以是性能度量的平均。如在方法500中所描绘的,目标值532被计算为1与结果的Matthew相关系数(MCC)之差。MCC是可以用作合成性能分数的性能度量。特别地,MCC是介于-1和+1之间的值,并且实质上是观察到的和预测的二进制分类之间的相关系数。MCC可以利用以下公式进行计算:

[0073] $$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

[0074] 其中,TP:真阳性;FP:假阳性;TN:真阴性;FN:假阴性。但是,一般而言,用于基于一组性能度量生成复合性能度量的任何合适技术都可以用来计算目标值。

[0075] 图6-8是用于通过二进制广义模拟方法的步骤进行解析的方法的流程图。一般而言,二进制广义模拟退火方法可以用来识别如图5 中所述的目标值的最优值(即,全局最小值)。如在本文所描述的,二进制广义模拟退火方法与图3中所描述的双融合方法结合使用。特别地,如图5中所描述的随机向量X以各种方式被扰动,以识别最优的目标值。图6是用于初始化二进制广义模拟退火方法的流程图。图7是用于随机扰动随机向量X的各个分量以减小目标值的流程图。图8是用于局部扰动随机向量X以进一步减小目标值的流程图。换句话说,在图7中所描绘的方法生成随机向量X的主要扰动,而在图8中所描绘的方法生成随机向量X的次要扰动。

[0076] 图6是用于初始化二进制广义模拟退火方法的方法600的流程图。方法600初始化

若干个参数并生成随机二进制向量X(1)。特别地，在步骤640、642和644，分类器114分别把参数t、y和count初始化为1。参数t对应于时间间隔，并且当合适的目标值被确定时递增，如关于图7和8所描述的。迭代参数y对应于被执行的主要扰动的数目，并且关于图7更详细地描述。参数count对应于用于跟踪是否已经生成当前向量X的扰动版本的参数，并且关于图7更详细地描述。在步骤646，分类器114产生随机二进制向量X。

[0077] 在步骤648，设置参数D。参数D对应于X中将被选择进行扰动的分量的个数。特别地，在步骤648，参数D被设置为0.2＊C，其中C对应于二进制向量X的长度。

[0078] 在步骤650，分类器114生成结果向量和目标值。特别地，分类器114可以使用在图5中所描绘的方法来生成结果向量526和目标值 534。但是，一般而言，确定表示复合性能度量的目标值的任何合适的方法都可以使用。在生成目标值之后，分类器114前进到图7中的步骤，以通过扰动随机向量X减小目标值。

[0079] 图7是用于通过对向量X执行主要扰动来减小二进制广义模拟退火方法中的目标值的方法的流程图。在模拟退火方法中，引入了人工温度(T(t＝1))并逐渐降低，以模拟冷却。在模拟退火中使用访问分配来模拟从一个点到另一个点(即，从一个随机向量X(1)到另一个随机向量X(2))的试跳距离。基于对应于随机向量X(2)的结果目标值是否小于对应于随机向量X(1)的目标值并且基于如以下定义的接受概率，来接受试跳。如本文所描述的，二进制广义模拟退火方法用来定位全局最小值(即，最小化目标值)。但是，一般而言，可以使用任何合适的算法，诸如最陡下降(steepest descent)、共轭梯度 (conjugate gradient)、单纯形(simplex)和蒙特卡罗(Monte Carlo)方法。

[0080] 在利用图6中所描绘的方法初始化模拟之后，分类器114在步骤 760开始选择向量X(1)的D个分量。向量X(1)的这D个分量可以随机地选择，或者可以执行选择向量X(1)的D个分量的任何其它合适的方法。在步骤762，count变量被设置为2。在步骤764，生成对应于有D个分量被改变的原始随机向量X(1)的第二随机二进制向量 X(2)。

[0081] 在步骤766，分类器114生成用于第二向量X(2)的结果向量768 和目标值770。特别地，分类器114可以使用在图5中所描绘的方法来生成结果向量和目标值。但是，一般而言，确定表示复合性能度量的目标值的任何合适的方法都可以使用。

[0082] 在生成第二结果向量和第二个目标值之后，分类器在决定框772 确定count变量等于2，并且前进到决定框776，以比较第一目标值 (即，对应于随机向量X(1))与第二目标值(即，对应于随机向量 X(2))。

[0083] 如果第二目标值不小于第一目标值，则这意味着第一向量X(1) 导致比第二向量X(2)更好或与其相等的相关性。在这种情况下，分类器前进到步骤778，以计算概率P。特别地，概率P对应于接受第二目标值的概率，并且基于等式：

[0084] $$P=\min\{1,[1-(1-q_e)\beta\partial E]^{\frac{1}{1-q_e}}\}$$

[0085] 其中$\partial E=obj(2)-obj(1)$

[0086] $$\beta=\frac{1}{T_{q_e}(t)}$$

[0087] q_a是用于接受概率P的控制参数，及

[0088] T_{qv}是温度值。

[0089] 如在本文所描述的，概率P对应于在广义模拟退火方法中使用的概率，但是，一般而言，任何合适的概率值都可以使用。在步骤 786，生成介于0和1之间的随机数r，其中包括0和1。随机数r可以从均匀分布或任何其它合适的分布中产生，并且在决定框788，r 与概率P进行比较。

[0090] 如果P大于或等于r，则这意味着接受第二目标值的概率高，即使第二目标值不小于第一目标值。在这种情况下，分类器114前进到步骤790和792，以便把第二向量X(2)和第二目标值分别存储为第一向量X(1)和第一目标值。

[0091] 作为选择，如果在决定框776分类器114确定第二目标值小于第一目标值，则这意味着向量X(2)导致更好的相关性或更好的性能。因此，分类器直接前进到步骤790，以便用向量X(2)更新向量X(1)，然后前进到步骤792，以便用第二目标值更新第一目标值。在步骤794，分类器114将count变量设置为等于1。

[0092] 作为选择，如果在决定框788分类器114确定r大于P，则这意味着接受第二个目标值的概率低，使得步骤790和792被绕过，并且向量X(1)和第一个目标值不被对应的第二个值覆盖。在这种情况下，分类器114前进到步骤794并将count变量设置成等于1。

[0093] 在将count变量重新设置为1之后，分类器114前进到决定框 796，在那里，迭代参数y与值L进行比较。值L对应于在前进到图 8所描绘的方法以执行小扰动之前要执行的大扰动的最大次数。如果迭代参数y不等于L，则分类器114前进到决定框772和步骤774来递增迭代参数y并在步骤760-764执行向量X的大扰动。重复上述步骤，直到已经执行了期望次数的大扰动L。如在图7中所描绘的，要执行的大扰动的次数是固定数目L。但是，用于L的值可以取决于任何数量的因素。例如，分类器114可以基于目标值的收敛来确定已经达到了大扰动的总数。在另一个例子中，如果在决定框776的固定数量的最近比较中没有发现小于第一目标值的第二目标值，则可能达到了大扰动的总数。一般而言，确定大扰动已经完成的任何合适的方法都可以使用，并且分类器114可以前进到图8去执行小扰动。

[0094] 图8是用于通过对向量X执行小扰动进一步减小二进制广义模拟退火方法中的目标值的方法的流程图。特别地，方法800在步骤 802开始并且设置变量C等于向量X(1)的长度。在步骤804，分类器 114将迭代参数c初始化为1并且将提高标志变量设置为假(false)。

[0095] 在步骤806，分类器114通过翻转X(1)的第c位以生成X_{temp}来对向量X(1)执行小扰动。特别地，X(1)是长度为C的二进制向量，并且除第c位之外X_{temp}几乎与X(1)完全相同。

[0096] 在步骤808，分类器114为临时向量X_{temp}生成结果向量810和目标值812。特别地，分类器114可以使用在图5中所描绘的方法来生成临时结果向量和临时目标值。但是，一般而言，确定代表复合性能度量的目标值任何合适的方法都可以使用。

[0097] 在决定框814，第一目标值与该临时目标值进行比较。如果临时目标值小于第一目标值，则这意味着被扰动的版本X_{temp}导致比原始向量X(1)更好的性能。在这种情况下，分类器114前进到步骤816，以便用被扰动的版本X_{temp}覆盖向量X(1)，前进到步骤818，以便用临时目标值覆盖第一目标值，然后前进到步骤819，以便将提高标志变量设置为真(true)。

[0098] 在决定框820，分类器114确定向量X(1)中的每一位是否都已经被翻转至少一次(即，在步骤806)，或等效地，迭代参数c是否等于X(1)的尺寸C。如果不是，则分类器114前进

到步骤822，以递增迭代参数c，然后前进到步骤806去翻转第c位。

[0099] 否则，如果分类器114在决定框820确定迭代参数c等于向量X(1)的长度C，则分类器114前进到决定框822，以确定是否期望进一步的提高。特别地，分类器114可以识别提高标志变量的值，以确定是否期望附加的位翻转。例如，如果提高标志变量为真(true)，则分类器114返回到步骤804，以便将迭代参数c重新初始化为1并且将提高标志变量重新初始化为假(false)。

[0100] 图8所描绘的方法使用提高标志变量来确定执行小扰动的过程 (即，位翻转)何时完成。但是，一般而言，任何其它合适的方法也可以用来确定小扰动何时完成。例如，分类器114可以要求目标值低于某个阈值，或者目标值与临时目标值之差低于某个阈值。如果这些需求没有被满足，则分类器114可以返回到步骤806去翻转向量X(1) 的另一位，以生成另一个临时目标值。

[0101] 在分类器114已确定最小目标值已被识别出之后，分类器114前进到步骤824和826，以分别递增参数t和递减参数D。

[0102] 在步骤828，分类器114利用在广义模拟退火中普遍使用的冷却公式计算温度T。但是，任何合适的公式都可以使用。

[0103] $$T_{qv}(t) \leftarrow T_{qv}(0)\frac{2^{t^{q_v-1}}-1}{(1+t)^{t^{q_v-1}}-1}$$

[0104] 其中q_v是定义分布函数的曲率的参数。

[0105] 在决定框830，分类器114确定$T_{qv}(t)$是否小于T_L。用于T_L的值表示阈值，其中，如果用于$T_{qv}(t)$的值低于T_L，则方法800结束，并且当前的随机向量X(1)被用作最优分类。

[0106] 本主题的实现可以包括，但不限于，包括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，以及包括可操作成使一个或多个机器(例如，计算机、机器人)产生本文所述操作的机器可读介质的制造品(articles)。本文所描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实现。这种多个计算系统可以经一个或多个连接——包括但不限于通过网络(例如，互联网、无线广域网、局域网、广域网、有线网络，等等)的连接，经这多个计算系统中一个或多个系统之间的直接连接，被连接并且可以交换数据和/或命令或其它指令等。

[0107] 图9是计算设备的框图，诸如图1的系统100的任何组件，包括用于执行参考图2-8所描述的过程的电路系统。系统100的每个组件可以在一个或多个计算设备900上实现。在某些方面，多个上述组件和数据库可以包括在一个计算设备900中。在某些实现中，组件和数据库可以跨若干个计算设备900实现。

[0108] 计算设备900包括至少一个通信接口单元、输入/输出控制器910、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 902)和至少一个只读存储器(ROM 904)。所有这些元件都与中央处理单元(CPU 906)通信，以便于计算设备900的操作。计算设备900可以以多种不同的方式配置。例如，计算设备900可以是常规的独立计算机或者，作为选择，计算设备900的功能可以跨多个计算机系统和体系架构分布。计算设备900 可以配置为执行数据分割、差分、分类、计分、排名及存储操作当中的一些或全部。在图9中，计算设备900经网络或局域网链接到其它的服务器或系统。

[0109] 计算设备900可以配置在分布式体系架构中，其中数据库和处理器被容纳在分开

的单元或位置中。一些这种单元执行主处理功能并且最少包含通用控制器或处理器和系统存储器。在这样一个方面，这些单元中的每一个都经通信接口单元908连接到通信集线器或端口(未示出)，其中通信集线器或端口充当与其它服务器、客户端或用户计算机和其它相关设备的主通信链路。通信集线器或端口自身可以具有最小的处理能力，主要充当通信路由器。多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。

[0110] CPU 906包括处理器，诸如一个或多个常规微处理器和一个或多个辅助协处理器，诸如用于从CPU 906卸载工作量的数学协处理器。CPU 906与通信接口单元1008和输入/输出控制器910通信，通过它们，CPU 906与其它设备，诸如其它服务器、用户终端或设备，通信。通信接口单元908和输入/输出控制器910可以包括多个通信信道，用于与例如其它处理器、服务器或客户终端同时通信。彼此通信的设备不需要连续地向对方发送。相反，这种设备只需要按需向对方发送，可以实际上在大多数时间避免交换数据，并且可以要求执行若干个步骤以在设备之间建立通信链接。

[0111] CPU 906还与数据存储设备通信。数据存储设备可以包括磁、光或半导体存储器的适当组合，并且可以包括，例如，RAM 902、ROM 904、闪存驱动器、诸如紧凑盘的光盘或者硬盘或驱动器。CPU 906和数据存储设备每个都可以例如完全位于单个计算机或其它计算设备中；或通过通信介质彼此连接，其中通信介质诸如USB端口、串行端口电缆、同轴电缆、以太网类型电缆、电话线、射频收发器或其它类似的无线或有线介质或前述的组合。例如，CPU 906可以经通信接口单元908连接到数据存储设备。CPU 906可以配置为执行一个或多个特定的处理功能。

[0112] 数据存储设备可以存储，例如，(i)用于计算设备900的操作系统1012；(ii)适于根据这里所描述的系统和方法，并且尤其是根据关于CPU 906详细描述的过程，指导CPU 906的一个或多个应用914 (例如，计算机程序代码或计算机程序产品)；或(iii)适于存储信息的数据库916，这些信息可以被用来存储程序所需的信息。在一些方面，数据库包括存储实验数据以及已发布的文献模型的数据库。

[0113] 操作系统912和应用914可以以例如压缩的、未编译的和加密的格式存储，并且可以包括计算机程序代码。程序的指令可以从除数据存储设备之外的计算机可读介质，诸如从ROM 904或从RAM 902，读到处理器的主存储器中。当程序中指令序列的执行使CPU 906执行本文所描述的处理步骤时，硬连线的电路系统可以用来代替用于本发明过程的实现的软件指令或者与其组合使用。因此，所描述的系统和方法不限于硬件和软件的任何特定组合。

[0114] 可以提供合适的计算机程序代码，用于执行与本文所描述的分类方法相关的一个或多个功能。程序还可以包括程序元素，诸如操作系统912、数据库管理系统以及允许处理器经输入/输出控制器910与计算机外围设备(例如，视频显示器、键盘、计算机鼠标等)接口的“设备驱动程序”。

[0115] 还提供了包括计算机可读指令的计算机程序产品。当计算机可读指令在计算机系统上被加载并执行时，使计算机系统根据上述方法或者上述方法的一个或多个步骤来操作。如本文所使用的，术语“计算机可读介质”指向计算设备900的处理器(或者本文所描述的设备的任何其它处理器)提供或参与提供要执行的指令的任何非临时性介质。这种介质可以采取多种形式，包括但不限于，非易失性介质和易失性介质。非易失性介质包括，例如，

光、磁、或光磁盘、或集成电路存储器，诸如闪存存储器。易失性介质包括通常构成主存储器的动态随机存取存储器(DRAM)。计算机可读介质的常见形式包括，例如，软盘、柔性盘、硬盘、磁带、任何其它磁介质、CD-ROM、DVD、任何其它光学介质、穿孔卡片、纸带、具有孔模式的任何其它物理介质、RAM、PROM、EPROM或EEPROM(电可擦除可编程只读存储器)、FLASH-EEPROM、任何其它存储器芯片或盒式磁带，或者计算机可以从中读取的任何其它非临时性介质。

[0116] 可以由各种形式的计算机可读介质把一条或多条指令的一个或多个序列运送到CPU 906(或者本文所描述的设备的任何其它处理器) 以供执行。例如，指令可以最初地在远程计算机(未示出)的磁盘上承载。远程计算机可以将指令加载到其动态存储器中，并通过以太网连接、电缆线、或者甚至利用调制解调器的电话线发送指令。在计算设备900(例如，服务器)本地的通信设备可以在各自的通信线路上接收数据，并且为了处理器而将数据放在系统总线上。系统总线将数据运送到主存储器，处理器从那里检索并执行指令。由主存储器接收的指令可以可选地在被处理器执行之前或之后存储在存储器中。此外，指令可以经通信端口作为电、电磁或光信号接收，这些信号是运送各种类型信息的无线通信或数据流的示例性形式。

[0117] 示例

[0118] 以下公开的数据集是从基因表达汇编(Gene Expression Omnibus,GEO)(http://www.ncbi.nlm.nih.gov/geo/)库中下载的：

[0119] a.GSE10106(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE10106)

[0120] b.GSE10135(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE10135)

[0121] c.GSE11906(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE11906)

[0122] d.GSE11952(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE11952)

[0123] e.GSE13933(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE13933)

[0124] f.GSE19407(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE19407)

[0125] g.GSE19667(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE19667)

[0126] h.GSE20257(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE20257)

[0127] i.GSE5058(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE5058)

[0128] j.GSE7832(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE7832)

[0129] k.GSE8545(www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc＝GSE8545).

[0130] 训练数据集是在Affymetrix平台(HGU-133+2)上。原始数据文件被属于R(R Development Core Team,2007)中Bioconductor (Gentleman,2004)的affy包(Gautier,2004)的ReadAffy函数读取，并且质量受控于：(利用affy包的AffyRNAdeg函数)生成RNA降解图、(利用函数affyPLM(Brettschneider,2008))生成 NUSE和RLE图，及计算MA(RLE)值；从训练数据集中排除在质量控制检查时降至低于一组阈值或在以上数据集中重复的数组；以及利用gcrma算法(Wu,2004)标准化通过质量控制检查的数组。对每个数据集，训练集样本分类从GEO数据库的系列矩阵文件中获得。输出由具有用于233个样本(28个COPD样本和205个对照样本)的54675个探测集的基因表达矩阵组成。为了产生平衡的数据集， COPD样本被自举多次，以便在应用如在共同未决的美国临时申请 61/662812中所描述的双融合方法之前获得224个COPD样本。利用包含205个对照和224个COPD患者的组合数据集，构建具有

409 个基因的基因签名。在随机向量中使用850个二进制值。在该方法中使用的分类方法包括以下R包：lda、svm、randomForest、knn、 pls.lda和pamr。最大迭代设为5000。训练数据集中交叉确认过程中的Matthew相关系数(MCC)和准确度分别是0.743和0.87。训练数据集中基因签名的热图在图10中示出。在图10的热图中，基因表达值按行居中。热图的颜色不能在灰度级中清楚地显示，但是图10 的数据显示对照数据在左侧示出，而COPD数据在右侧示出。测试数据集是从商业供应商(Genelogic)获得的未公布数据集，该数据集包含16个对照样本和24个COPD样本。在不应用本发明的变换不变量方法(transformation invariant method)的情况下，由双融合生成的基因签名正确地预测了全部40个样本中的29个样本。准确度为0.725，并且MCC是0.527。基因签名正确地预测了16个对照样本中的15个，并且正确地预测了24个COPD样本中的14个。

[0131] 虽然已经参考具体的例子特别地示出并描述了本发明的实现，但是本领域技术人员应当理解，在不背离本公开内容的主旨和范围的情况下，可以在其中进行各种形式和细节上的变化。

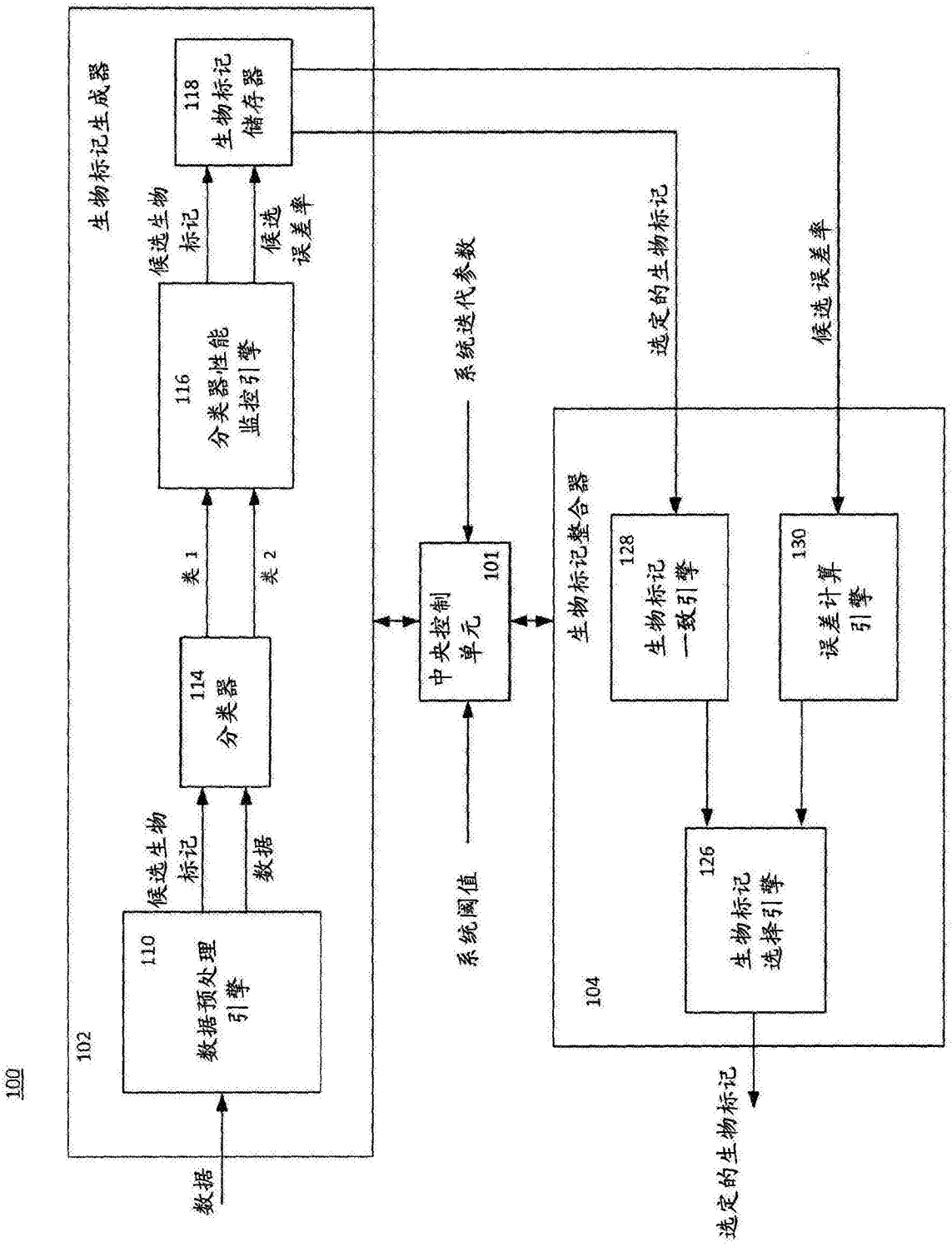

100
102
生物标记生成器
数据
110
数据预处理引擎
候选生物标记
数据
114
分类器
类 1
类 2
116
分类器性能监控引擎
候选生物标记
候选误差率
118
生物标记储存器
系统阈值
中央控制单元
101
系统迭代参数
104
生物标记整合器
126
生物标记选择引擎
128
生物标记一致引擎
130
误差计算引擎
选定的生物标记
候选误差率
选定的生物标记

图1

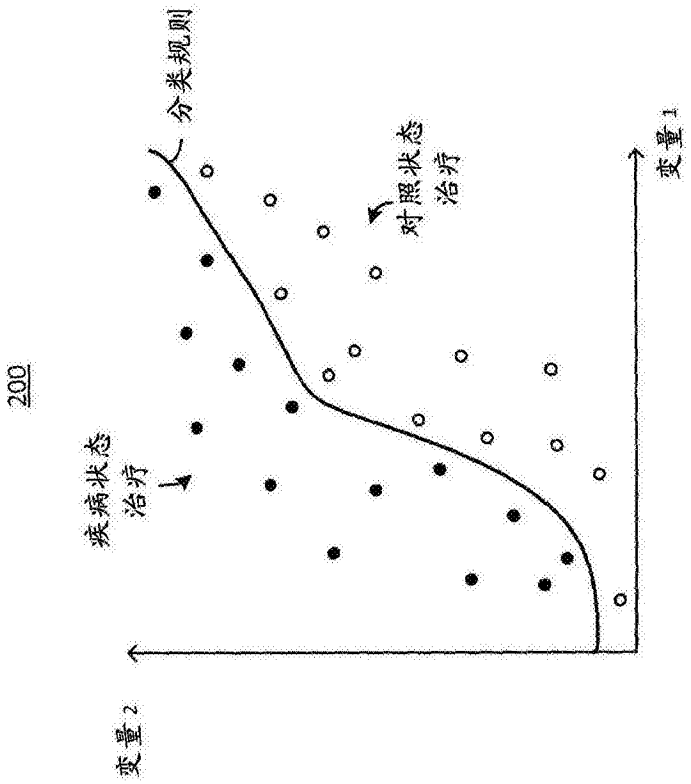


图2

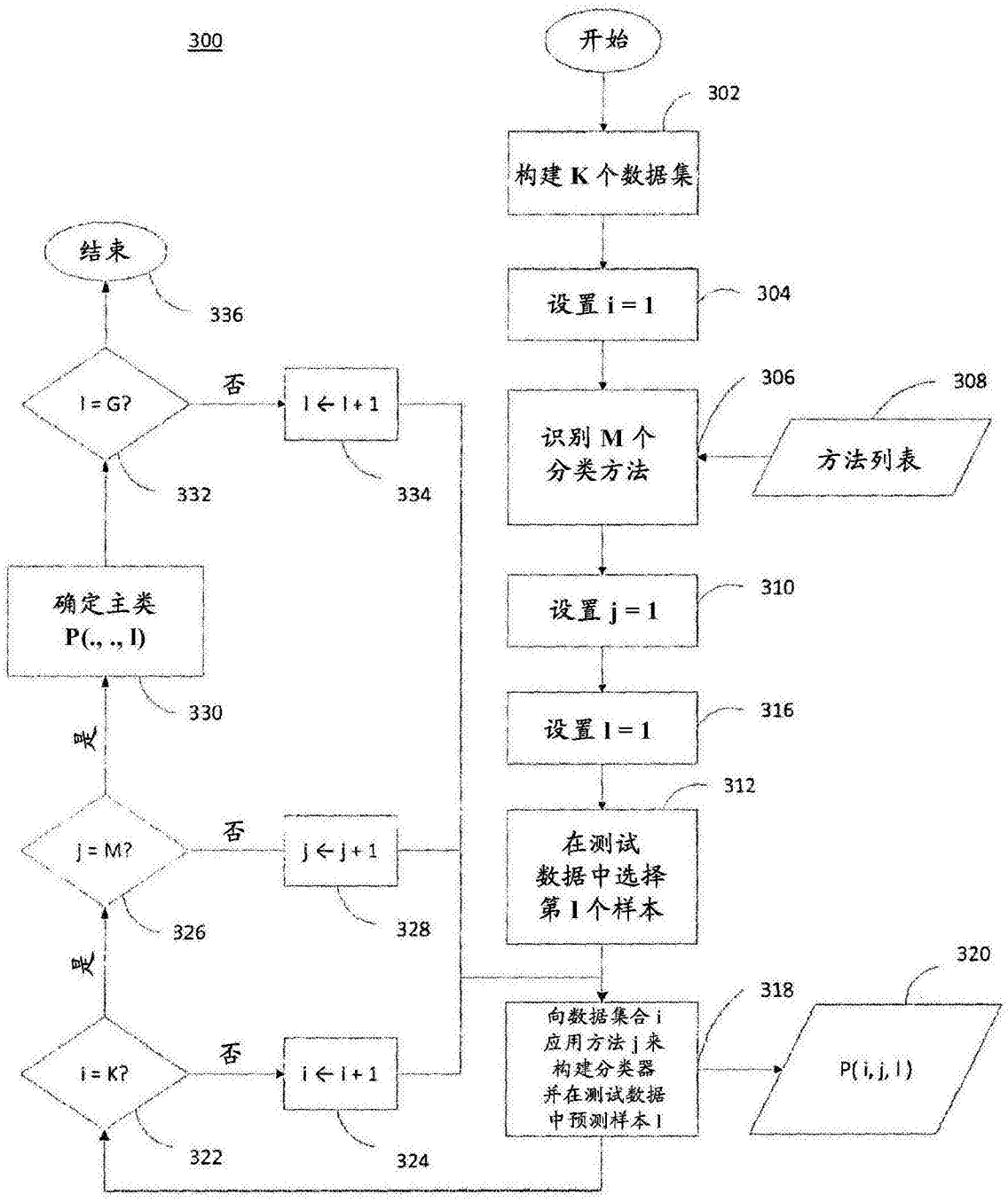

300
开始
302
构建 K 个数据集
设置 i = 1
304
306
识别 M 个
分类方法
308
方法列表
设置 j = 1
310
设置 l = 1
316
312
在测试
数据中选择
第 l 个样本
320
318
向数据集合 i
应用方法 j 来
构建分类器
并在测试数据
中预测样本 l
P(i, j, l)
i = K?
否
i ← i + 1
322
324
是
j = M?
否
j ← j + 1
326
328
是
l = G?
否
l ← l + 1
332
334
是
确定主类
P(., ., l)
330
结束
336

图3

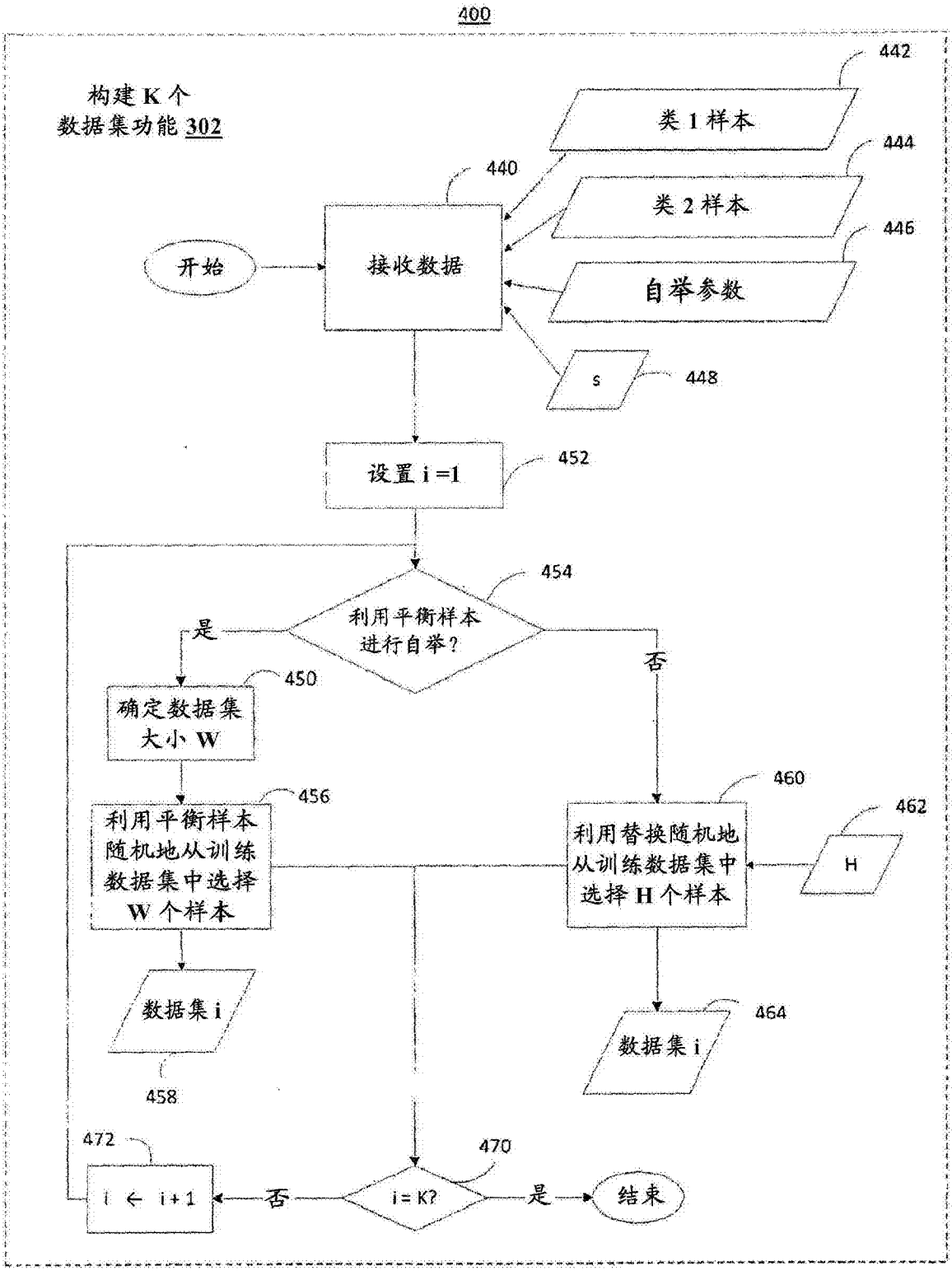

400
构建 K 个
数据集功能 302
442
类 1 样本
444
类 2 样本
446
自举参数
440
开始
接收数据
s
448
设置 i =1
452
454
利用平衡样本
进行自举?
是
否
450
确定数据集
大小 W
456
利用平衡样本
随机地从训练
数据集中选择
W 个样本
460
利用替换随机地
从训练数据集中
选择 H 个样本
462
H
数据集 i
458
464
数据集 i
472
i ← i+1
否
470
i = K?
是
结束

图4

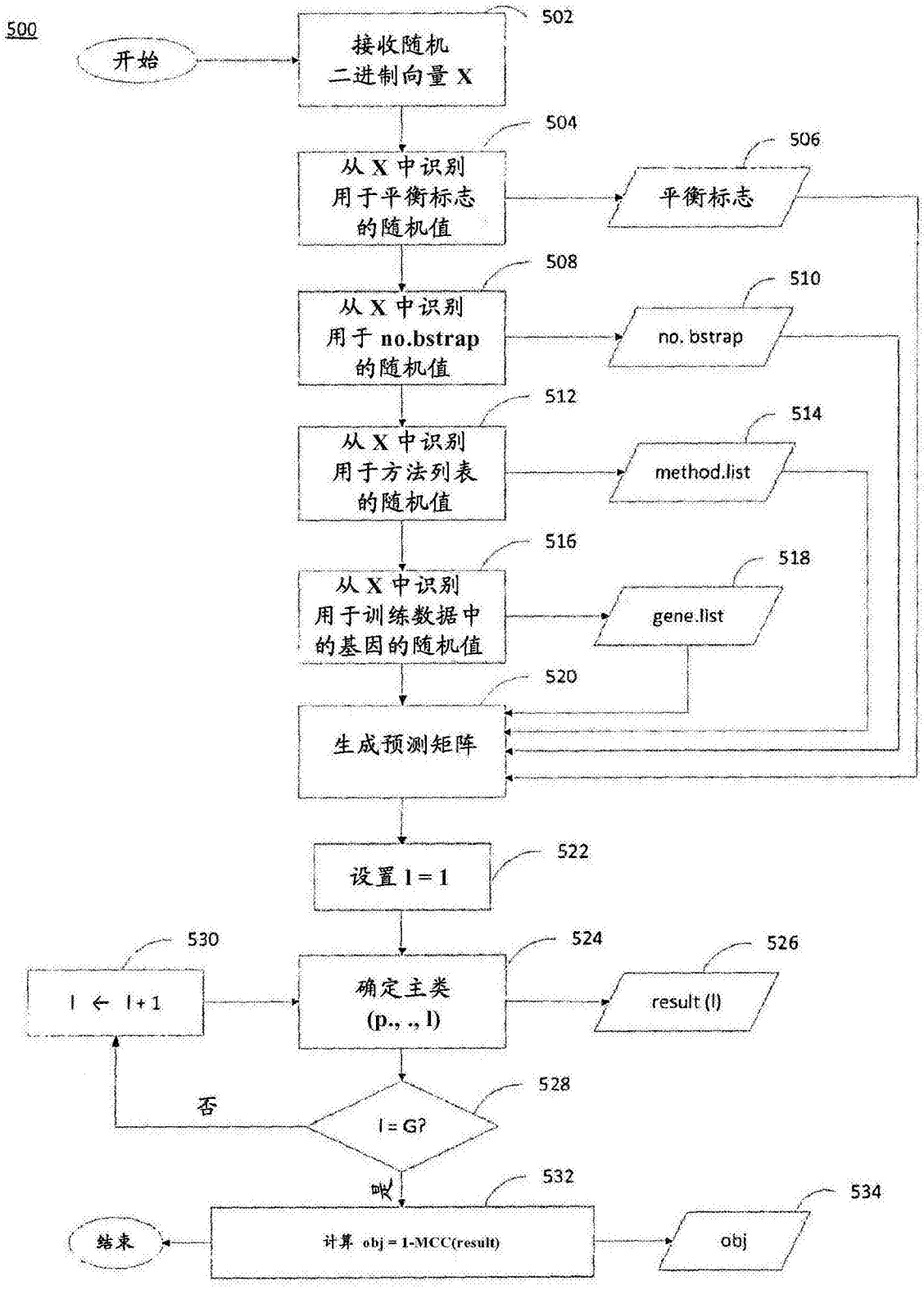

500
开始
接收随机
二进制向量 X
502
从 X 中识别
用于平衡标志
的随机值
504
平衡标志
506
从 X 中识别
用于 no.bstrap
的随机值
508
no. bstrap
510
从 X 中识别
用于方法列表
的随机值
512
method.list
514
从 X 中识别
用于训练数据中
的基因的随机值
516
gene.list
518
生成预测矩阵
520
设置 l = 1
522
l ← l+1
530
确定主类
(p., ., l)
524
result (l)
526
否
l = G?
528
是
计算 obj = 1-MCC(result)
532
obj
534
结束

图5

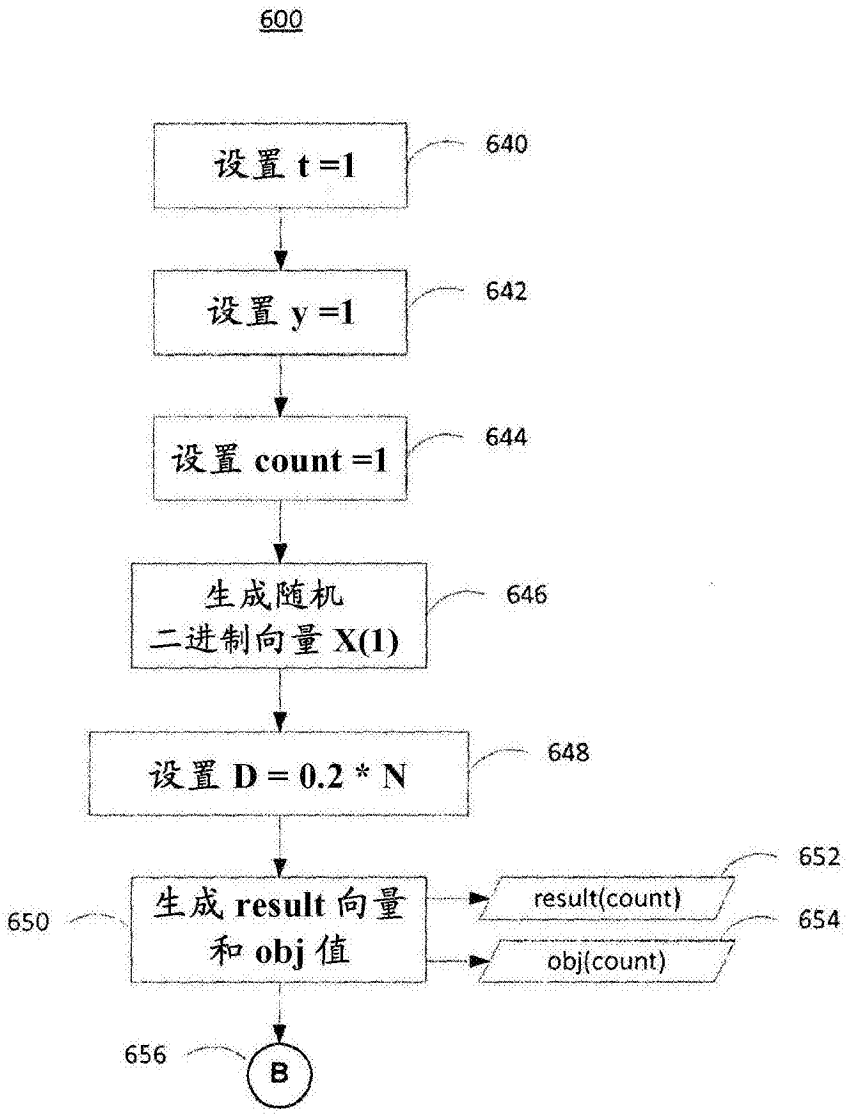

600
设置 t =1
640
设置 y =1
642
设置 count =1
644
生成随机
二进制向量 X(1)
646
设置 D = 0.2 * N
648
生成 result 向量
和 obj 值
650
result(count)
652
obj(count)
654
656
B

图6

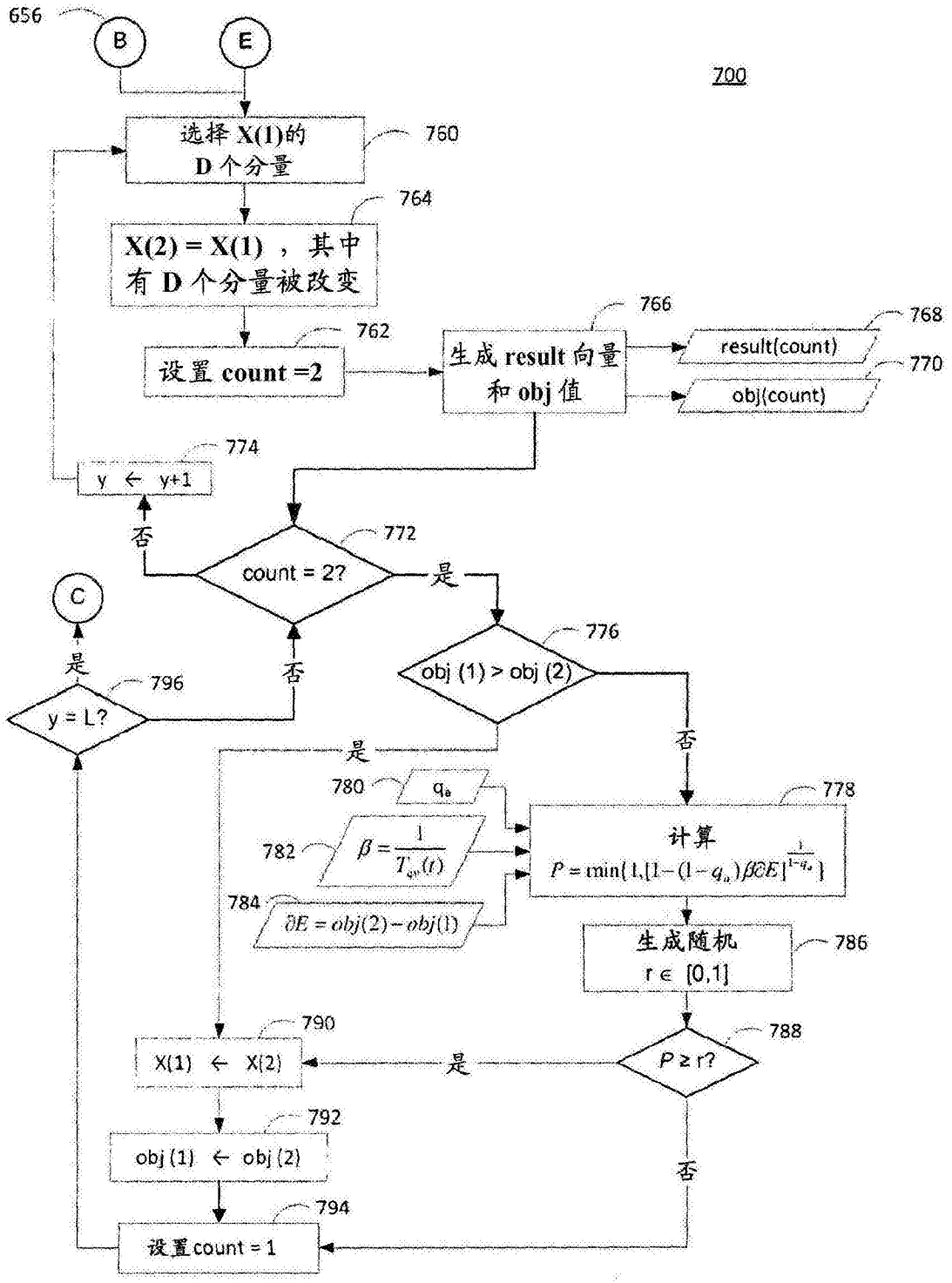

656
B
E
700
选择 X(1)的 D 个分量
760
764
X(2) = X(1) ，其中有 D 个分量被改变
762
设置 count =2
766
生成 result 向量和 obj 值
768
result(count)
770
obj(count)
774
y ← y+1
否
772
count = 2?
是
C
是
796
y = L?
否
776
obj (1) > obj (2)
否
是
780
q_e
782
$\beta = \frac{1}{T_{qv}(t)}$
784
$\partial E = obj(2) - obj(1)$
778
计算
$P = \min\{1, [1-(1-q_e)\beta\partial E]^{\frac{1}{1-q_e}}\}$
生成随机 r ∈ [0,1]
786
788
P ≥ r?
是
790
X(1) ← X(2)
792
obj (1) ← obj (2)
否
794
设置count = 1

图7

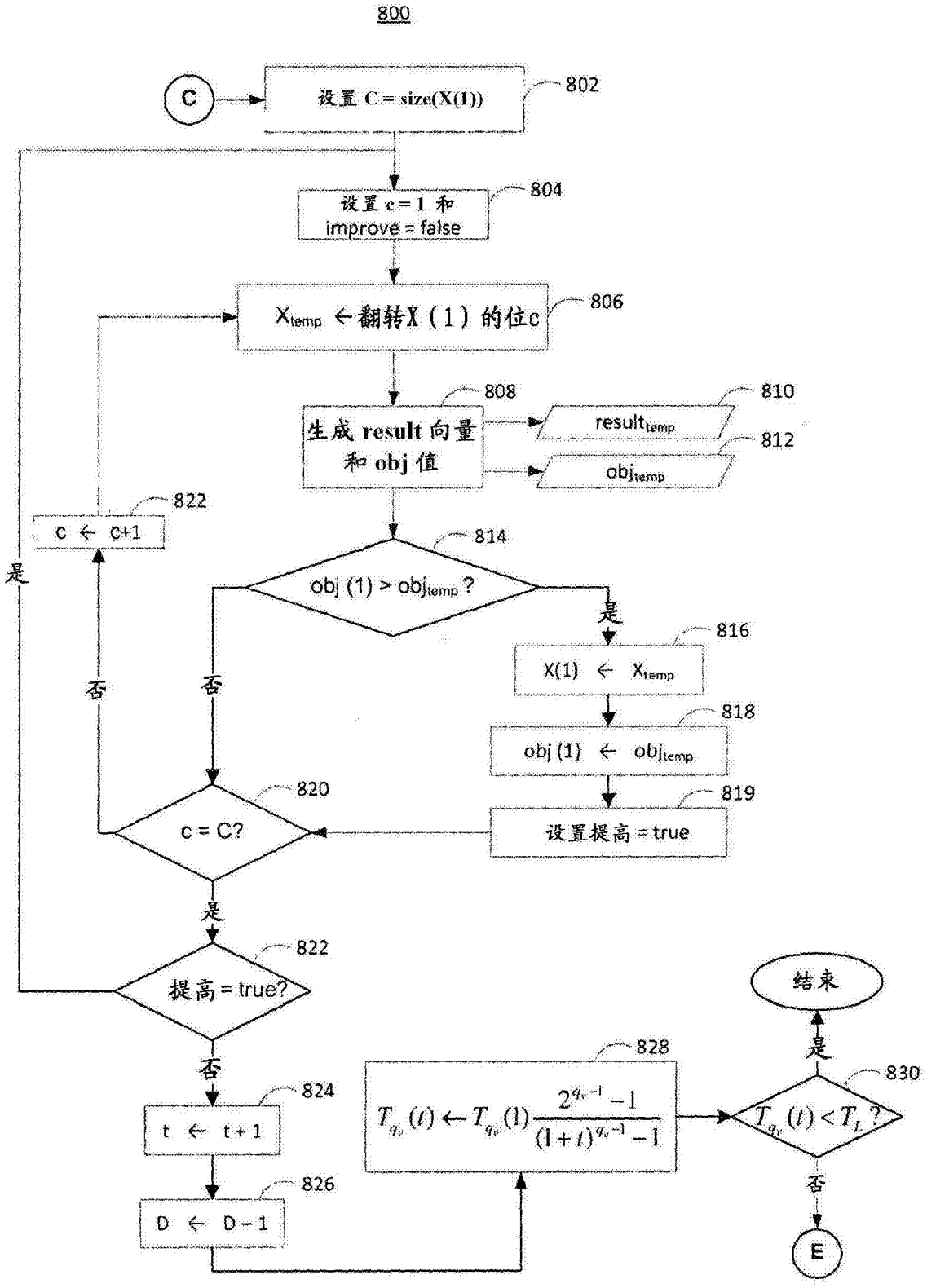

800
C
设置 C = size(X(1))
802
设置 c = 1 和 improve = false
804
X_{temp} ← 翻转X（1）的位c
806
生成 result 向量和 obj 值
808
$result_{temp}$
810
obj_{temp}
812
c ← c+1
822
obj (1) > obj_{temp} ?
814
是
否
X(1) ← X_{temp}
816
obj (1) ← obj_{temp}
818
设置提高 = true
819
c = C?
820
是
否
提高 = true?
822
是
否
t ← t+1
824
D ← D−1
826
$T_{q_v}(t) \leftarrow T_{q_v}(1)\frac{2^{q_v-1}-1}{(1+t)^{q_v-1}-1}$
828
$T_{q_v}(t) < T_L$?
830
是
否
结束
E

图8

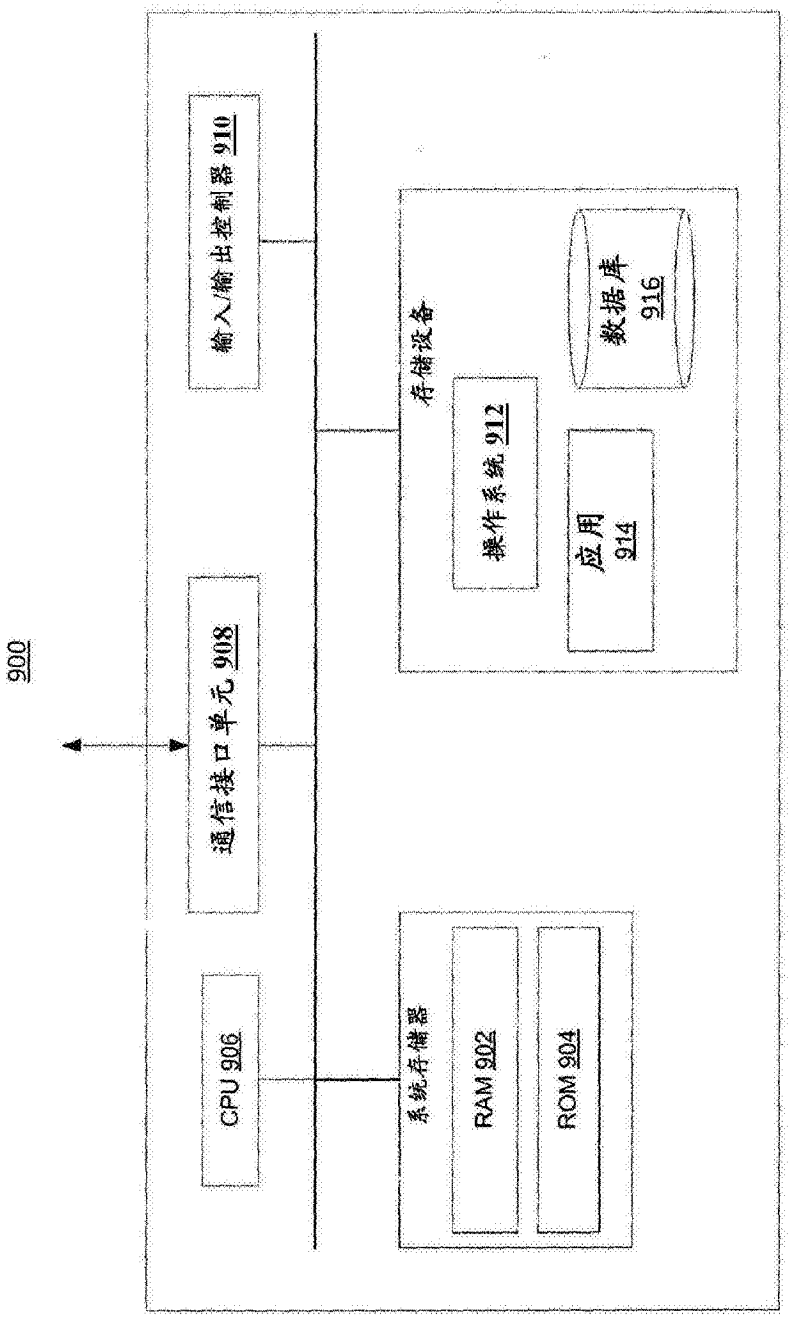

900
通信接口单元 908
输入/输出控制器 910
存储设备
操作系统 912
应用 914
数据库 916
CPU 906
系统存储器
RAM 902
ROM 904

图9

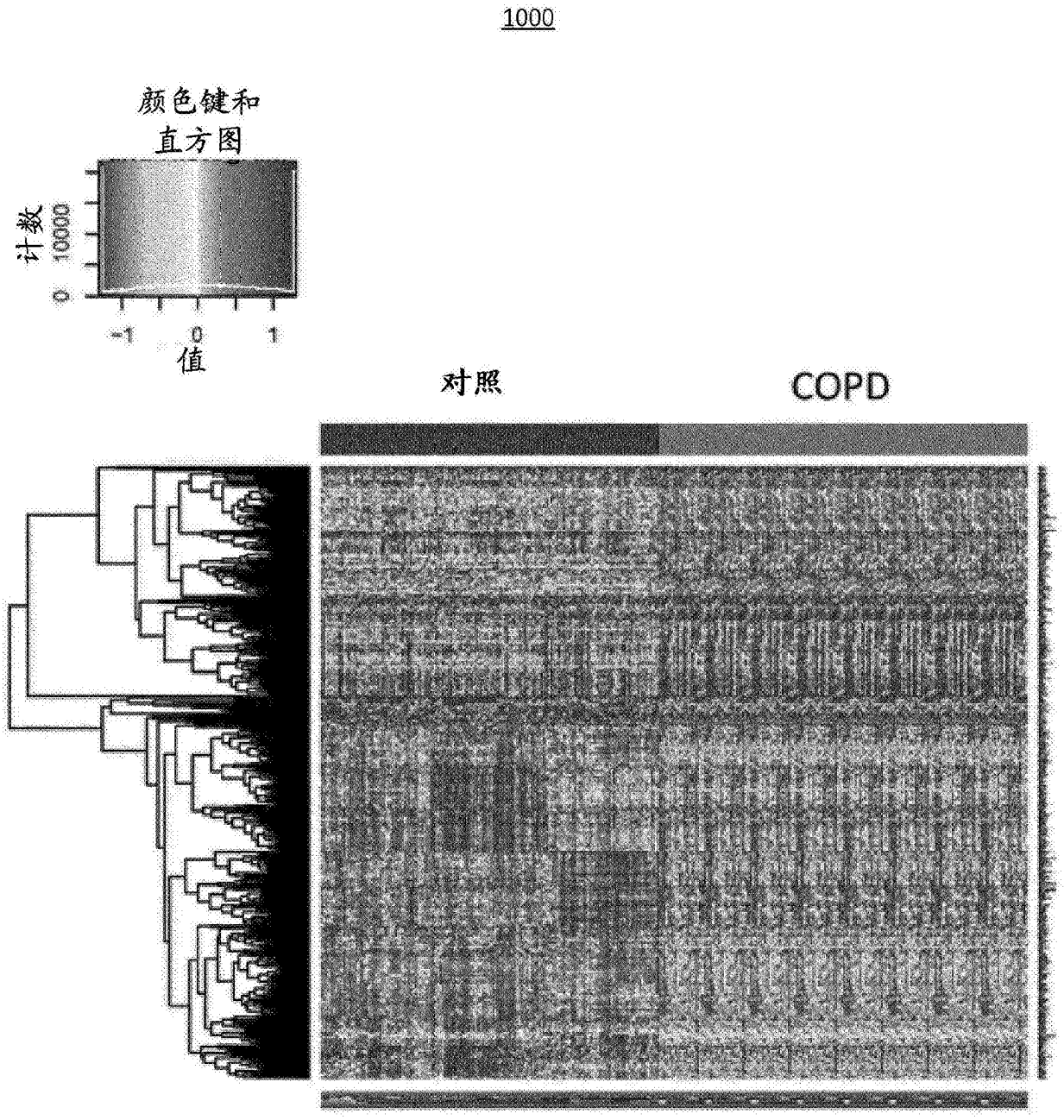


图10