



(12)发明专利

(10)授权公告号 CN 102693262 B

(45)授权公告日 2017. 01. 18

(21)申请号 201210034651.3

(51)Int.Cl.

(22)申请日 2012.02.08

G06F 17/30(2006.01)

(65)同一申请的已公布的文献号

(56)对比文件

申请公布号 CN 102693262 A

WO 2008085706 A1, 2008.07.17,

(43)申请公布日 2012.09.26

US 2009144218 A1, 2009.06.04,

(30)优先权数据

审查员 胡斌

13/023181 2011.02.08 US

(73)专利权人 通用电气公司

地址 美国纽约州

(72)发明人 R·E·凯兰

(74)专利代理机构 中国专利代理(香港)有限公

司 72001

代理人 柯广华 朱海煜

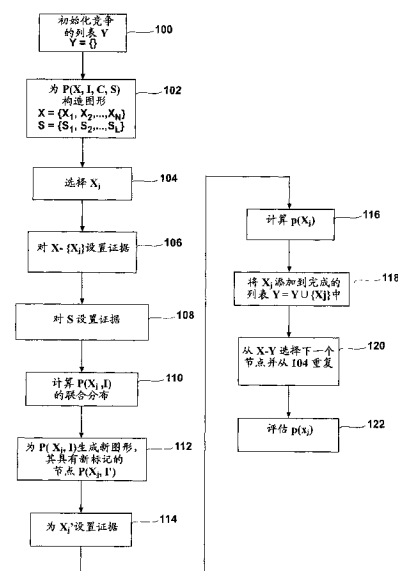
权利要求书1页 说明书6页 附图5页

(54)发明名称

确定现象中变量的影响的方法

(57)摘要

本发明的名称为“确定现象中变量的影响的方法”。一种确定现象中变量的影响的方法,包括:从非暂时介质中提取选择的变量以用于在处理器中分析和处理包括现象中的其他变量的图形操作的序列。计算选择的变量的变量影响指标并对其他选择的变量重复这些步骤,其能够实现选择的变量之间的评估以确定它们在现象中的影响。



1. 一种确定现象中变量的影响的方法,包括:

步骤a,在非暂时介质中以图形形式提供混合模型,所述混合模型包括模型成分、至少一个类节点和多个变量节点,所述至少一个类节点表示与所述模型成分关联的类,而所有所述多个变量节点表示遇到所述现象的系统内的物理数据,

步骤b,在处理器中,从所述非暂时介质中选择所述多个变量节点的至少一个,

步骤c,通过对所述多个变量节点中所选择的至少一个以外的多个变量节点设置证据来对所述图形形式执行操作,

步骤d,通过边缘化来计算所述多个变量节点中所选择的一个和所述至少一个类节点的联合分布,以生成新图形,

步骤e,由所述新图形计算所述多个变量节点中所述选择的一个的变量影响指标,

步骤f,对所述多个变量节点中其他选择的变量节点重复步骤b、c、d和e,以及

步骤g,彼此相对地评估所述多个变量节点的变量影响指标的量值。

2. 如权利要求1所述的方法,其中,所述新图形是通过 $f: P(X_1, I | e_{X-X_j}, e_S) \rightarrow P(X_1', I')$ 描述的变换,其中, I 表示所述模型成分, X_1 表示所述变量, S 表示类上的状态或分布, e 指示证据, e_{X-X_j} 为 X_j 以外的所有变量生成的证据,以及 e_S 为对属于 S 的变量设置的证据。

3. 如权利要求1所述的方法,其中,所述变量影响指标表示所述变量节点的值中的方向变化。

4. 如权利要求1所述的方法,其中,所述选择是与应用相关的。

5. 如权利要求1所述的方法,其中,所述选择包括所述多个变量节点的子集。

6. 如权利要求1所述的方法,其中,所述执行步骤包括按样式和按次序设置证据,以确定变量影响指标的类型。

7. 如权利要求1所述的方法,其中,所述系统是飞行器引擎,以及所述混合模型表示所述飞行器引擎的性能。

8. 如权利要求1所述的方法,其中,所述变量节点包括:表示连续参数的多个第一变量节点,以及表示与所述类内的变量关联的值或分布的多个第二变量节点,并且其中,所述选择步骤包括选择所述第一变量节点之一。

9. 如权利要求8所述的方法,其中,对所述第一变量节点和第二变量节点计算所述联合分布。

确定现象中变量的影响的方法

技术领域

[0001] 本文描述的技术涉及确定现象中给定变量的影响的方法。

背景技术

[0002] 检测与机器或观测的事件中的具体失灵或故障模式有关的样式可能非常具有挑战性。一般确定征兆(或测量)何时异常更为容易。获知情况异常可以是非常有价值的。但是,如果可以利用严重性评级标记异常和/或将异常与特定状况或故障模式关联,则更有价值。输入变量(例如,测量参数)与异常之间关联的样式中包含诊断信息。但是,此样式可能非常难以抽取。

[0003] 在加工行业内,经常使用主成分分析(PCA)来进行异常检测或故障诊断。可以计算残留成分(residual component)或主成分的变量贡献。此方法提供哪些变量对异常测量的贡献最大的指示。但是,PCA具有局限性。它是单模的,意味着当数据由复杂密度生成时它的效用受限,并且它未提供处理丢失数据的直观方法。

[0004] 检测变量贡献的另一个途径是计算残差(residual)。对于特定变量,使用回归技术来预测变量的值,然后用测量的值减去它以导出残差。残差的量值提供其对异常状态的贡献的测量。但是,直接比较不同的变量仍可能是困难的。并且,如果多个变量对异常有贡献,则来自这些残差的输出可能造成误导。回归技术往往是单模的并且将遇到与PCA类似的局限性。

发明内容

[0005] 在一个方面中,一种确定现象中变量的影响的方法,包括:以图形形式提供混合模型,混合模型包括模型成分、表示与该模型成分关联的类的至少一个类节点、以及表示与该类内的变量关联的值的多个变量节点,全部表示遇到现象的系统内的物理数据;选择变量节点的其中一个或子集;通过对选择的一个变量节点以外的变量节点设置证据来对图形形式执行操作;通过边缘化(marginalize)来计算选择的变量节点和一个或多个类节点的联合分布,以生成新图形;由新图形计算选择的变量节点的变量影响指标;对其他选择的变量节点重复选择步骤、执行步骤和计算步骤;以及彼此相对地评估变量节点的变量影响指标的量值。

[0006] 在另一个方面中,该新图形是通过 $f:P(X_1, I | e_{X-X_j}, e_S) \rightarrow P(X_1', I')$ 描述的变换,其中 I 表示模型成分, X 表示变量, S 表示类上的状态或分布,以及 e 指示证据。

[0007] 在进一步的方面中,该变量影响指标表示变量节点的值中的方向变化。同样,选择可以是与应用相关的。而且,执行步骤可以包括按样式和按次序设置证据以确定变量影响指标的类型。在一个实施例中,现象发生在飞行器引擎的系统中,以及混合模型表示飞行器引擎的性能。

附图说明

- [0008] 在附图中：
- [0009] 图1A示出给定现象中若干不同输入变量的数据绘图。
- [0010] 图1B是图1A中的输入变量的时间历史的似然率计分。
- [0011] 图2是现象的混合模型，其示出了高斯分布和用作滤波器的离散节点两者。
- [0012] 图3是基于鸢尾花数据的模型的示范对数似然率。
- [0013] 图4是描绘根据本发明一个实施例的、确定现象中变量的影响的方法的流程图。
- [0014] 图5是根据图4的方法对图1A的数据计算的变量影响指标的示例。

具体实施方式

[0015] 在下文描述中，为了解释的目的，阐述了多个特定细节，以便提供对本文描述的技术的透彻理解。但是对于本领域技术人员来说，显然在没有这些特定细节的情况下仍可以实施这些示范实施例。在其他实例中，以示意图形式示出结构和装置以便有助于描述这些示范实施例。

[0016] 下文参考这些附图来描述这些示范实施例。这些附图图示特定实施例中实现本文描述的模块、方法和计算机程序产品的某些细节。但是，这些附图不应解释为施加附图中可能存在的任何限制。该方法和计算机程序产品可以在任何机器可读介质上提供以便实现它们的操作。这些实施例可以使用现有计算机处理器、或通过为此目的或另一个目的并入的专用计算机处理器、或硬线连接的系统来实现。

[0017] 正如上文提到的，本文描述的实施例包括具有用于承载或其上存储有机器可执行指令或数据结构的非暂时机器可读介质的计算机程序产品。此类计算机可读介质可以是可由通用或专用计算机或具有处理器的其他机器访问的任何可用介质。通过举例，此类机器可读介质可以包括RAM、ROM、EPROM、EEPROM、CD-ROM或其他光盘存储装置、磁盘存储装置或其他磁存储装置或能够用于以机器可执行指令或数据结构的承载或存储期望的程序代码并且能够被通用或专用计算机或具有处理器的其他机器访问的任何其他介质。当通过网络或另一种通信连接（例如硬线连接、无线或硬线连接或无线的组合）向机器传送或提供信息时，该机器恰当地将该连接视为机器可读介质。因此，任何此类连接均恰当地称为机器可读介质。上文这些的组合也包含在机器可读介质的范围内。机器可执行指令包括，例如使通用计算机、专用计算机或专用处理机器执行某个功能或一组功能的指令和数据。

[0018] 实施例将在方法步骤的通用上下文中进行描述，这些方法步骤可以在一个实施例中通过包括例如采用联网环境中的机器执行的程序模块的形式的、如程序代码的机器可执行指令的程序产品来实现。一般地，程序模块包括，具有执行具体任务或实现具体抽象数据类型的技术效果的例行程序、程序、对象、组件和数据结构等。机器可执行指令、关联的数据结构和程序模块表示用于执行本文公开的方法步骤的程序代码的示例。此类可执行指令或关联的数据结构的具体序列表示用于实现此类步骤中描述的功能的对应动作的示例。

[0019] 实施例可以在联网环境中使用至具有处理器的一个或多个远程计算机的逻辑连接来实施。逻辑连接可以包括本文作为举例而非限制提出的局域网（LAN）和广域网（WAN）。此类联网环境在办公方面或企业方面计算机网络、内联网和因特网中是常见的，并且能够使用范围多样的不同通信协议。本领域技术人员将意识到，此类网络计算环境通常将涵盖多种类型的计算机系统配置，包括个人计算机、手持装置、多处理器系统、基于微处理器或

可编程的消费电子产品、网络PC、微计算机、主机计算机等。

[0020] 实施例还能够在分布式计算环境中实施,在这些分布式计算环境中,由通过(硬线连接的链路、无线链路或通过硬线连接的链路或无线链路的组合)经通信网络链接的本地和远程处理装置来执行任务。在分布式计算环境中,可以将程序模块放置在本地存储器存储装置和远程存储器存储装置两者中。

[0021] 用于实现这些示范实施例的整体或或多个部分的示范系统可以包括计算机形式的通用计算装置,这些通用计算装置包括处理单元、系统存储器和将包括系统存储器的多种系统组件耦合到处理单元的系统总线。该系统存储器可以包括只读存储器(ROM)和随机存取存储器(RAM)。该计算机还可以包括用于从磁硬盘读取以及向磁硬盘写入的磁硬盘驱动器、用于从可移动磁盘读取以及向可移动磁盘写入的磁盘驱动器以及用于从如CD-ROM的可移动光盘或其他光介质读取或向其写入的光盘驱动器。这些驱动器及其关联的机器可读介质提供用于该计算机的机器可执行的指令、数据结构、程序模块和其他数据的非易失性存储。

[0022] 这些实施例中公开的方法的技术效果包括更有效率地检测与机器中的具体失灵或故障模式有关的样式,减少诊断和故障处理时间并能够实现更好健康和维护规划。

[0023] 使用变量影响指标来提供变量的“兴趣”行为的指示。变量影响指标的一示例应用是确定哪些变量可为异常行为负责。使用一种称为混合模型的数据驱动构建的模型来计算变量影响指标。假定已使用历史数据以突出特定应用感兴趣的的行为的方式训练此模型。混合模型提供用于对范围广泛的物理现象建模的丰富资源,正如G. McLachlan和D. Peel在《Finite Mixture Models》, John Wiley&Sons, (2000)中所描述的。混合模型可用于对现象中的正常行为建模,并且因此也可用于检测异常行为。来自混合模型的似然率计分能够用于监视异常行为。基本上,变量影响指标是似然率计分。

[0024] 此上下文中的感兴趣行为意味着位于混合模型的密度边缘上的空间区域中的变量。该模型对于位于这些区域中的数据更敏感。对于如健康监视的许多应用,低密度空间的区域往往表示最感兴趣的区域,因为这些区域中操作的机器正在其设计极限外运行。

[0025] 似然率计分可以在数据过渡通过低密度区域时提供有用的诊断信息。似然率将往往展示趋势特征,这些趋势特征提供有关行为的信息(例如,健康正在恶化或随机和可能出现与差的仪器关联的感测)。图1A和图1B图示了此情况。图1A示出给定现象中八个不同变量的值的数据绘图。图1B是图1A中的全部输入变量的似然率计分的时间历史计分。此处,看到完整数据的似然率反映若干输入变量的形状—它提供一种融合的形式并概述所有输入变量上的行为(注意似然率总是在对数空间中示出)。如果所有输入变量的完整历史均位于高密度区域中,则在似然率计分中没有形状(向下趋势)。还有,似然率计分的量值取决于异常行为输入变量的数量。

[0026] 似然率计分展示混合模型对于与数据时间历史的最后部分关联的层面上运行的数据所具有的经验不多。如果混合模型被训练成表示正常行为,则似然率计分将展示渐增异常的行为。但是,虽然似然率计分显示异常行为,但是它未显示哪个组合的输入变量表现异常。而且,在直接利用这些输入变量处理时不易导出此信息。这是因为这些输入变量的标度(scale)和统计特性可能显著地不同。变量影响指标能够展示哪些变量显著地对异常有贡献。

[0027] 虽然变量影响指标是对数似然率计分,但是它们是以展示信息的特定方式来计算的。这意味着混合模型必须以展示感兴趣行为的方式来生成。此情况通常在以“图形形式”描述混合模型时便于解释。

[0028] 标准混合模型具有连接到表示混合成分的离散父节点(有时也称为“群集”)的高斯分布。图2图示此情况。在本发明的一个实施例中,用于计算变量影响指标的模型包含用作滤波器的附加节点。这些节点常常是离散,但是可能是连续的。这些滤波器可以设为在执行预测时改变模型成分的混合权重。例如,如果将不同的成分(或成分的组合)与个别的类关联,并且案例的类是已知的,则移除当前类的表示,并且获得从所有其他类的角度而言的当前案例的视图是可能的。图3中示出此类滤波对似然率计分的效果的特定示例。这是从公知的鸢尾花数据集(包含来自三种品种的鸢尾花(每个品种50个案例)的萼片和花瓣测量的集合的简单数据集)上构建的模型。示出对应于每个品种的使用所有输入变量的对数似然率。使用滤波器执行预测,滤波器确保计算中未使用与当前品种关联的成分。此类型的预测可以指示这些品种中哪个(如果有的话)是最不同的。图3示出是Setosa的品种,可以通过绘制散点图来容易地确认其如图这些的简单数据(这些似然率计分按品种排序,其中Setosa被首先绘制,然后是Versicolor,再后来是Virginica)。

[0029] 在图2中,I表示模型成分以及X是包括 $X_1, X_2, X_3 \dots X_N$ 的多变量高斯。节点C表示类变量。在一个实施例中,节点 S_L 表示类(即,个体类)内的变量值。节点C具有多个状态,等于类的数量(每个 S_L 对应于一个状态)。每个 S_L 的分布通常是二进制的,并且采用在当前类(对应 S_L)被去激活(即,从模型预测中移除)时所有其他类保持激活的方式将其进行编码。该分布还能够编码为执行此滤波的逆过程。在另一个实施例中,节点 S_L 可以是连续的节点,每个节点在节点C的值上编码一种“软”证据。

[0030] 图4中示出使用如图2所示的混合模型评估变量影响指标的方法的示范流程图。在此方法中,可以使用图形变换和推断(inference)来计算变量影响指标。变换是图形结构上的产生新图形结构的操作。推断包括输入证据(对一个或多个节点赋值)并计算联合概率或个体节点概率。不同的变换和相干步骤提供展示不同行为性质的变量影响指标的不同变体。例如,就飞行器引擎的行为的模型来考量。排放气体温度对于飞行中具体阶段将具有正常操作区域,并且甚高或甚低值可能表征异常行为。可以使用一种类型的变量影响指标来监视此“范围外”异常行为。当测量参数(例如,燃料流量和低压池速度)之间有相关性时,可以使用另一种类型的变量影响指标来监视不同样式的异常行为。尽管个体测量可能位于正常范围,但跨参数的样式可能是异常的(例如,当存在相关性丧失时)。

[0031] 在图4中,在100处,将列表Y初始化为空。此列表将保持跟踪已处理的测量节点。在102处,以图形方式定义如图2所述的混合模型。在104处选择变量节点 X_j 之一,并在106处为 X_j 以外的所有变量测量节点生成证据。证据仅在存在并被视为有效的(例如,可以将测量视为不可能的值)时才被输入。如果必要的话,在108处,对属于S的变量设置证据。然后,在110处计算 X_j 和I的联合分布。然后,在112处生成新图形,其包含编码110处计算的联合分布的新节点 X_j' 和 I' 。可以将新图形的示范变换指示如下:

[0032] $f: P(X_1, I | e_{X-X_j}, e_S) \rightarrow P(X_1', I')$,

[0033] 其中I表示模型成分,X表示变量,S表示类上的状态或分布,以及e指示证据。在114处,对 X_j' 设置证据(此证据指示为 x_j)以及在116处计算 $p(x_j)$ 。在118处,将 X_j 添加到完成的

列表Y中,然后选择新节点从104重复该过程。 $p(x_j)$ 的对数是 X_j 的基本变量影响指标。例如,就图2中的图形来考量。期望计算 X_2 、 X_3 和 X_4 的场合下 X_1 的变量影响指标。还知道,此案例来自类 S_2 (在本示例中, S 中所有节点均是离散的,但是它们可以是连续的或离散的与连续的组合)。将该案例的值指示如下:

[0034] $X_1=x_1, X_2=x_2, X_3=x_3, X_4=x_4$, 类= S_2

[0035] 输入证据,并在112处通过请求 (X_1, I) 的联合分布来生成新图形:

[0036] $f: P(X_1, I | x_2, x_3, x_4, S_2 = \text{true}) \rightarrow P(X_1', I')$

[0037] 函数 f 是指生成新图形的边缘化。上标'指示具有新分布的新变量。边缘化是应用于图形的标准方法,正如《Bayesian Networks and Decision Graphs》,Finn V.Jensen和Thomas D.Nielsen, Springer(2007)中教导的。实例化新图形能够实现执行进一步预测。在本示例中对 X_1 基本变量影响指标是

[0038] $p(x_1)$

[0039] 以及是在116处由新图形计算的。

[0040] 设置证据的过程确定所产生的变量影响指标的变体。例如,要计算对于单变量数据超出范围敏感的变量影响指标,不设置其他变量上的证据。但是,其他连续变量的证据仍可以用作确定载入新图形模型中的节点 I 的后验加权的证据。而且,当节点 X_N 被视为独立的时,计算后验加权的此证据设置可以是迭代的。此迭代包括输入证据节点的其中之一的证据,记录 I 上的分布,对所有其他证据节点重复,并然后计算 I 中每个状态的所记录的分布的积。因此,在步骤120中,重复对其他变量节点选择、生成、执行和计算变量影响指标的过程,以便在122处彼此相对地评估它们(如图绘制的)的量值以确定选择的变量节点的影响。可以通过与预定的准则进行比较来自动化122处的评估,或可以通过绘制的分布的可视化检查来人工执行。因此,可见到计算变量影响指标的变体中存在灵活性并且最适合的变体是与应用相关的。

[0041] 还可以对变量影响指标标符号,以使它们反映原始变量中的方向变化。例如,如果测量参数趋势向下,在变量影响指标中具有相同方向的趋势则可能是有用的。对变量影响指标标符号的一个简单方式是遵循证据设置的相同路径来生成新图形。然后将实际值(例如, x_1)与边缘分布的平均值比较。如果该值低于平均值,则变量影响指标具有负号,以及如果高于平均值,则具有正号。

[0042] 还可以相对于拟合度计分和模型阈值来换算变量影响指标。

[0043] 当将变量 X_N 视为相关集合时,异常变量可能对其他变量的变量影响指标有巨大影响。在这些情况中,可以对图4所示的数据流设置外环。然后,执行图4中的过程以检测具有最大影响的变量。将该变量设为NULL以处理为丢失,并重复图4中的过程。该过程在其余变量(即,未设为NULL的那些变量)具有视为正常的集合似然率计时终止。在重复图4中的过程的另一个变体中,可以将 X_N 的变量的不同子集(组合)设为NULL。当 N 小时,对于 X_N 的所有组合穷举地运行图4中的过程是可能的。小 N 的定义是与应用相关的,并且通过考量可用计算资源、数据负载和应用所需的系统响应时间来定义。

[0044] 图5中示出为图1A所示的输入数据计算的变量影响指标的示例。将理解不同类型的变量影响指标可以提供有关不同类型的异常(如单变量界外值和多变量界外值或去相关)的信息。可以通过输入的证据的样式和输入的证据的次序来确定变量影响指标的类型。

[0045] 本书面描述使用示例来公开包括最佳模式的本发明,以及还使本领域技术人员能制作和使用本发明。本发明可取得专利的范围由权利要求确定,且可包括本领域技术人员想到的其它示例。如果此类其它示例具有与权利要求字面语言无不同的结构要素,或者如果它们包括与权利要求字面语言无实质不同的等效结构要素,则它们规定为在权利要求的范围之内。

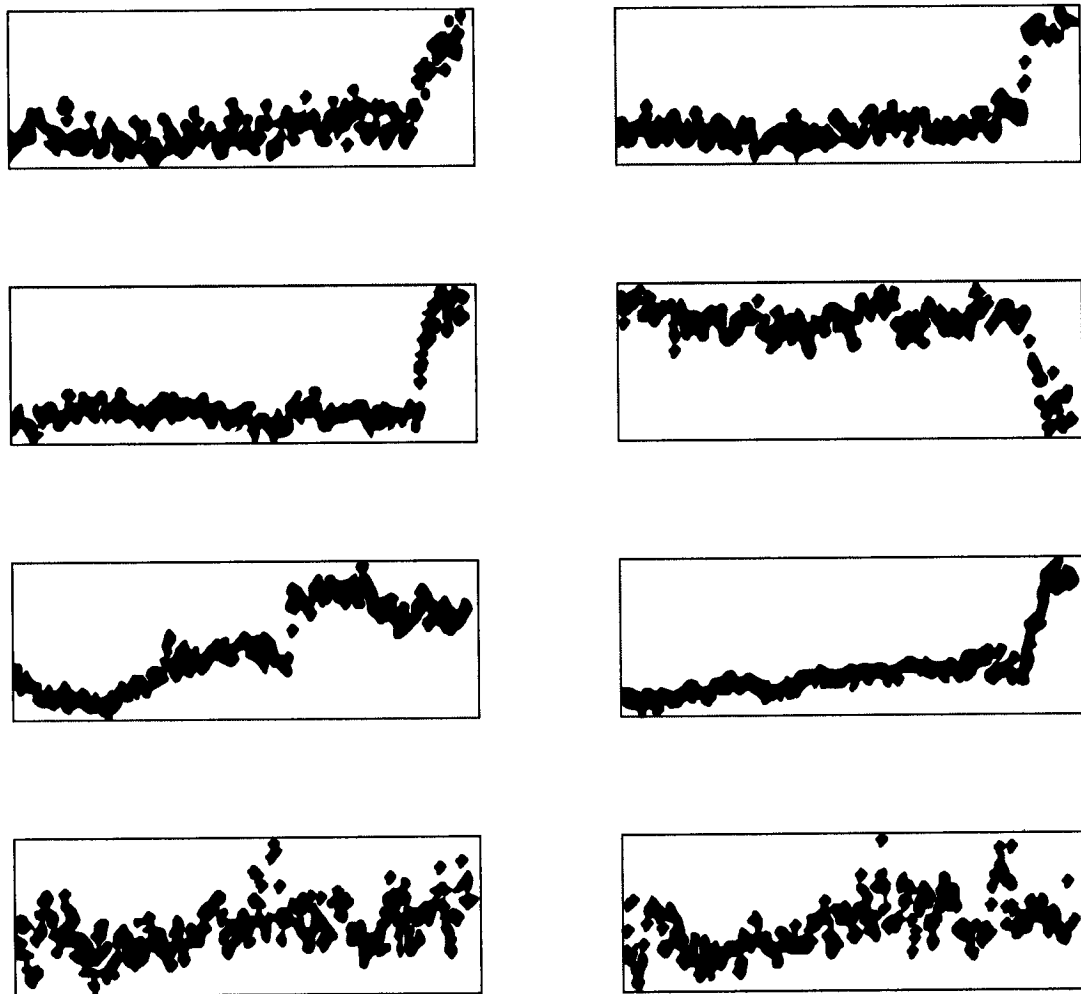


图1A

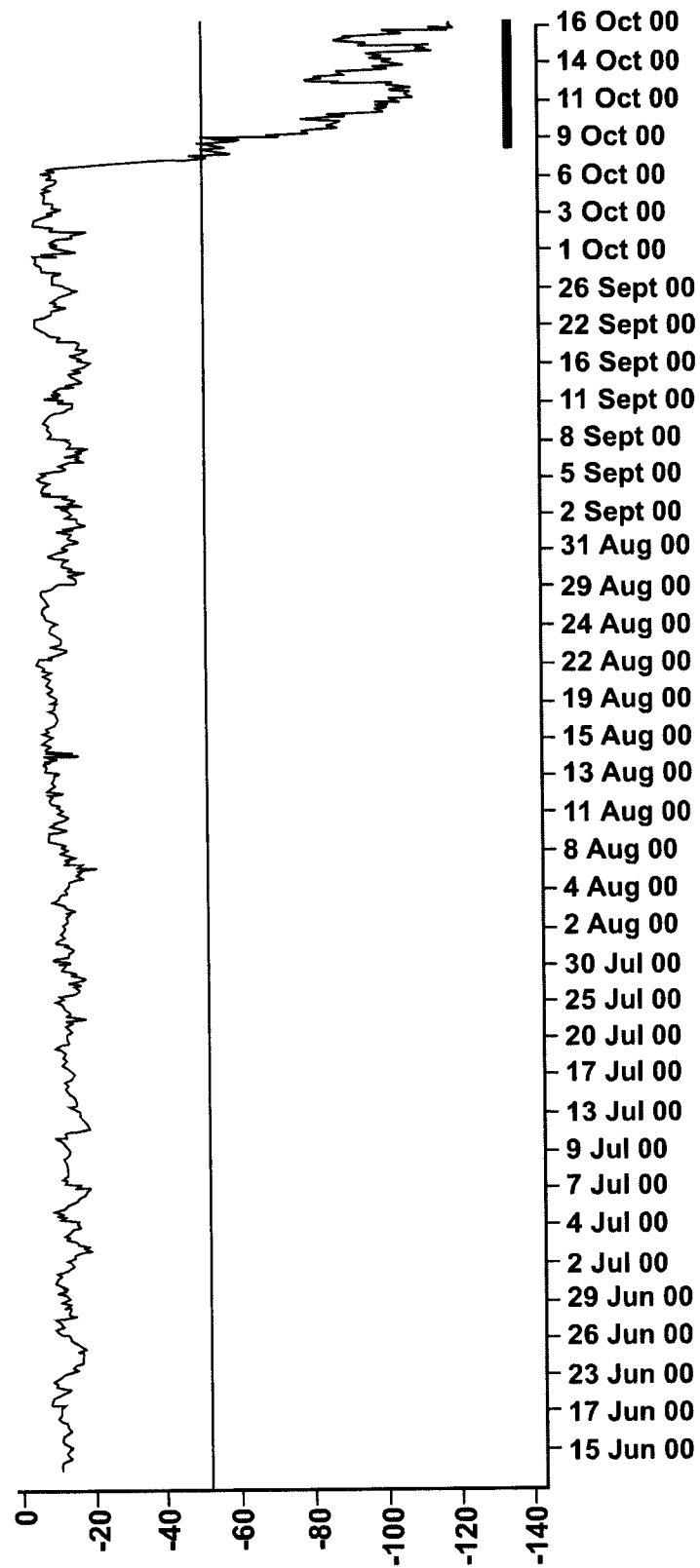


图1B

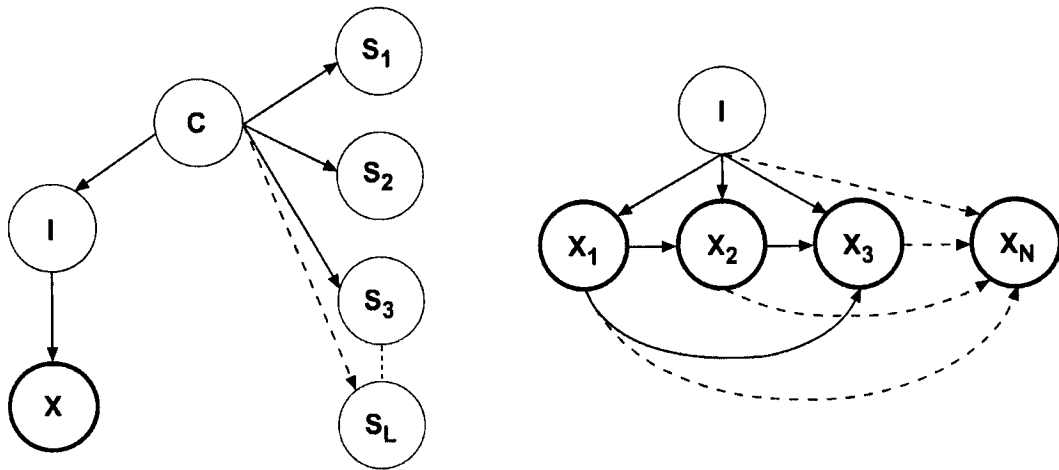


图2

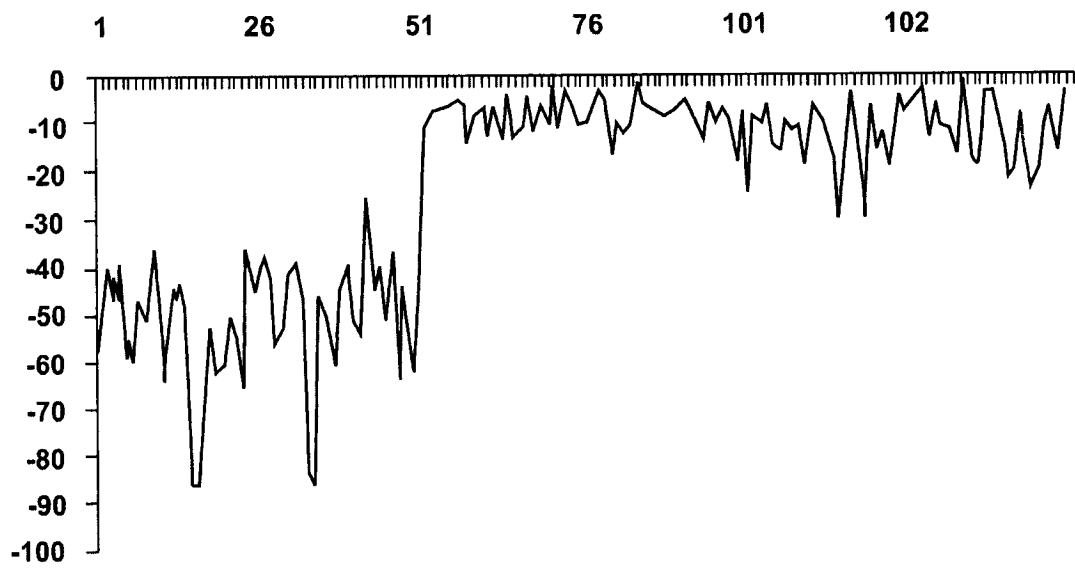


图3

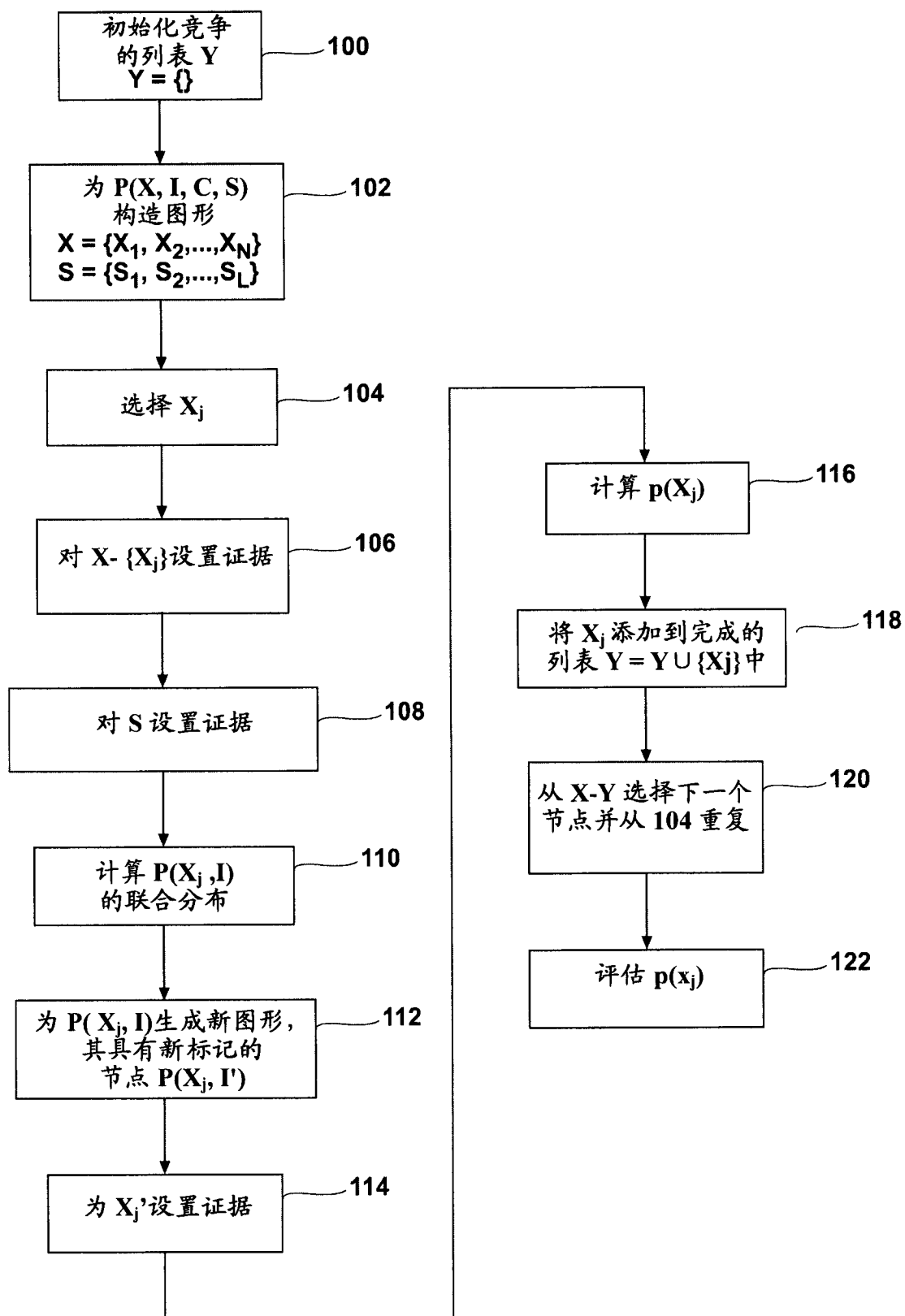


图4

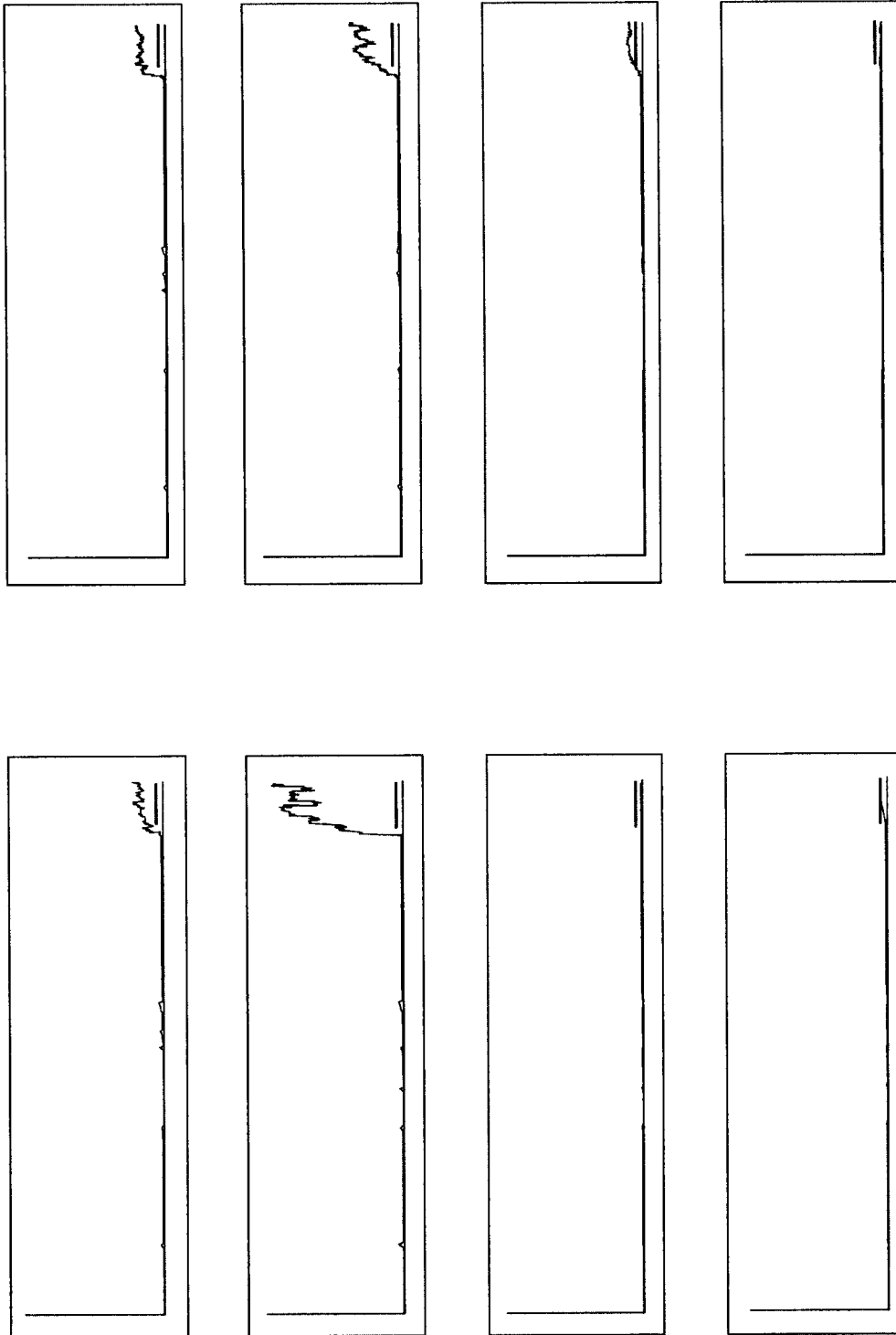


图5