(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 102693265 A
(43)申请公布日 2012.09.26

(21)申请号 201210041495.3

(22)申请日 2012.02.15

(30)优先权数据
13/027,829 2011.02.15 US

(71)申请人 通用电气公司
地址 美国纽约州

(72)发明人 R·E·凯兰 B·拉登

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 何欣亭 朱海煜

(51)Int.Cl.
*G06F 17/30*(2006.01)
*G06N 7/00*(2006.01)

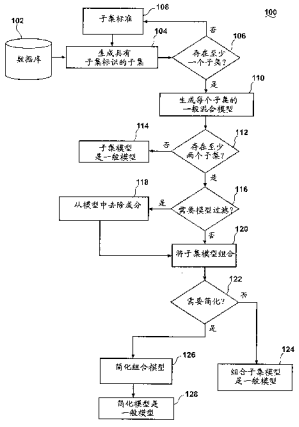
权利要求书 1 页 说明书 6 页 附图 4 页

(54)发明名称

构造混合模型的方法

(57)摘要

一种构造数据集的一般混合模型(100)的方法包括按照预定义的标准(108)将数据集划分为至少两个子集(104),生成至少两个子集的每个的子集混合模型(110),然后组合来自每个子集的混合模型,以便生成一般混合模型(120)。




CN 102693265 A

1. 一种生成非暂时介质 (102) 中存储的数据集的一般混合模型 (100) 的方法，包括下列步骤：

提供用于定义所述数据集的子集的子集标准 (108)；

在处理器中基于所述子集标准 (108) 将所述数据集划分为至少两个子集 (108)；

生成所述至少两个子集的每个的子集混合模型 (110)；以及

将所述至少两个子集的每个的所述子集混合模型组合为所述一般混合模型 (120)。

2. 如权利要求 1 所述的方法，其中，所述子集标准包括在关系数据库中被定义以及根据至少一维来过滤所述数据集的一个。

3. 如权利要求 1 或 2 所述的方法，其中，所述生成步骤包括识别子集的至少一个成分 (104)、将函数拟合到子集的至少一个成分、根据换算因子来换算拟合函数以及对换算拟合函数求和中的至少一个。

4. 如权利要求 3 所述的方法，其中，所述函数是概率密度函数。

5. 如权利要求 4 所述的方法，其中，所述概率密度函数是正态分布函数。

6. 如权利要求 3 所述的方法，其中，所述换算因子是标量值。

7. 如权利要求 4 所述的方法，其中，与子集的所述拟合函数的每个对应的所有所述换算因子之和为 1。

8. 如权利要求 1 或 2 所述的方法，其中，所述组合步骤 (120) 包括连接所述至少一个子集的每个的所述子集混合模型，独立换算所述至少一个子集的每个的所述子集混合模型，然后连接所述换算子集混合模型，并且在组合所述子集混合模型之前去除一个或更多成分函数 (150)。

9. 如权利要求 8 所述的方法，其中，在组合所述子集混合模型之前去除一个或更多成分函数 (150) 包括，选择成分并且确定所述所选成分与来自除了对应于所述所选成分的所述子集之外的子集的所有所述成分之间的距离 (144)。

10. 如权利要求 9 所述的方法，其中，在组合所述子集混合模型之前去除一个或更多成分函数 (150) 还包括去除具有最大距离的所述成分。

# 构造混合模型的方法

## 技术领域

[0001] 本发明涉及构造混合模型的方法。

## 背景技术

[0002] 数据挖掘是用于从数据中提取信息和值的技术。数据挖掘算法用于许多应用中，例如预测购物者对目标市场的消费习惯、检测信用卡诈骗交易、预测顾客的网站的浏览路线、机器的故障检测等。数据挖掘使用多年来由人工智能 (AI) 和统计建模团体所开发的一系列算法。存在许多不同类的算法，但是它们全部共有一些共同特征，例如 (a) 表示（隐式地或显式地）数据域的知识的模型，(b) 使用训练数据来构造模型的模型构建或学习阶段，以及 (3) 获得新数据并且对数据应用模型以进行预测的推理工具。已知的示例是线性回归模型，其中通过对第二变量的值进行加权并且对加权值与常数值进行求和，由第二变量来预测第一变量。加权和常数值是模型的参数。

[0003] 混合模型是学术研究团体中的数据挖掘应用的常用模型，如 G McLachlan 和 D Peel 在有限混合模型 (Finite Mixture Models)(John Wiley&Sons, (2000)) 中所述。存在混合模型的类的变化，例如混合专家 (Mixture of Experts) 和分层混合专家 (Hierarchical Mixture of Experts)。还存在用于构建混合模型的有大量文献证明的算法。一个示例是期望最大化 (EM)。一般通过识别数据中的簇或成分并且将适当数学函数拟合每个簇，来构造这类混合模型。

## 发明内容

[0004] 在一个方面，生成非暂时介质中存储的数据集的一般混合模型 (general mixture model) 的方法包括下列步骤：提供用于定义数据集的子集的子集标准；在处理器中基于子集标准将数据集划分为至少两个子集；生成至少两个子集的每个的子集混合模型；以及将至少两个子集的每个的子集混合模型组合为一般混合模型。

## 附图说明

[0005] 附图包括：

[0006] 图 1 是示出按照本发明的一个实施例的生成一般混合模型的方法的流程图。

[0007] 图 2 是示出作为图 1 所示方法的一部分的从子集混合模型过滤成分的方法的流程图。

[0008] 图 3 是示出按照图 1 的生成一般混合模型的方法的数据集的过滤的示例的图表。

[0009] 图 4 是示出第一子集的子集混合模型的图表。

[0010] 图 5 是示出第二子集的子集混合模型的图表。

[0011] 图 6 是示出通过图 1 公开的方法而构造的一般混合模型的图表。

[0012] 附图标记说明

[0013] 100 生成一般混合模型的方法；102 数据库；104 生成具有子集标识的子集；106

存在至少一个子集？ ;108 子集标准 ;110 构建每个子集的混合模型 ;112 存在至少两个子集？ ;114 子集模型是一般模型 ;116 需要模型过滤？ ;118 从模型中去除成分 ;120 将子集模型组合 ;122 需要简化？ ;124 组合子集模型是一般模型 ;126 简化组合模型 ;128 简化模型是一般模型 ;140 接收来自所有子集的所有成分 ;142 选择下一个成分 ;144 计算所选成分与其它子集中的所有其它成分的距离 ;146 距离大于预定值？ ;148 所选成分是最后一个成分？ ;150 去除所选成分 ;152 识别最终成分集 ;180 第一子集数据 ;190 第二子集数据 ;G1 第一簇拟合函数 ;G2 第二簇拟合函数 ;G3 第三簇拟合函数 ;G4 第四簇拟合函数 ;G5 第五簇拟合函数 ; $\mu_1$ 第一簇的均值 ; $\mu_2$ 第二簇的均值 ; $\mu_3$ 第三簇的均值 ; $\mu_4$ 第四簇的均值 ; $\mu_5$ 第五簇的均值。

## 具体实施方式

[0014] 为便于说明，以下描述中提出了大量具体细节，以便提供对本文所述技术的透彻理解。然而，本领域的技术人员将会清楚地知道，没有这些具体细节也可实施示范实施例。在其它情况下，结构和装置以简图形式示出，以便于示范实施例的描述。

[0015] 下面参照附图来描述示范实施例。这些附图示出实现本文所述的模块、方法和计算机程序产品的具体实施例的某些细节。但是，附图不应当被理解为施加附图中可存能在的任何限制。方法和计算机程序产品可在任何机器可读介质上提供以用于实现其操作。实施例可使用现有的计算机处理器、或者通过为这个或另一个目的而结合的专用计算机处理器、或者通过硬连线系统来实现。

[0016] 如上所述，本文所述的实施例包括计算机程序产品，其中包括用于携带或其上存储了机器可执行指令或数据结构的机器可读介质。这类机器可读介质能够是能够由通用或专用计算机或者具有处理器的其它机器来访问的任何可用介质。举例来说，这类机器可读介质能够包括 RAM、ROM、EPROM、EEPROM、CD-ROM 或其它光盘存储装置、磁盘存储装置或者其它磁存储装置、或者能够用于携带或存储采取机器可执行指令或数据结构形式的预期的程序代码并且能够由通用或专用计算机或者具有处理器的其它机器来访问的其它任何介质。当信息通过网络或另一种通信连接（硬连线、无线或者硬连线或无线的组合）传递或提供给机器时，机器将连接适当地看作是机器可读介质。因此，任何这种连接都适当地称作机器可读介质。以上的组合也包含在机器可读介质的范围之内。机器可执行指令例如包括使通用计算机、专用计算机或者专用处理机来执行某个功能或某组功能的指令和数据。

[0017] 将在方法步骤的一般上下文中描述实施例，方法步骤在一个实施例中可通过程序产品来实现，其中程序产品包括例如采取由组网环境中的机器所执行的程序模块形式的机器可执行指令、如程序代码。一般来说，程序模块包括具有执行特定任务或者实现特定抽象数据类型的技术效果的例程、程序、对象、组件、数据结构等。机器可执行指令、关联数据结构和程序模块表示用于运行本文所公开方法的步骤的程序代码的示例。这类可执行指令或关联数据结构的特定序列表示用于实现在这类步骤中所述的功能的相应动作的示例。

[0018] 实施例可在使用到具有处理器的一个或更多远程计算机的逻辑连接的组网环境中实施。逻辑连接可包括在此作为举例而不是限制来提供的局域网 (LAN) 和广域网 (WAN)。这类组网环境是办公室范围或企业范围的计算机网络、内联网和因特网中常见的，并且可使用很多种不同的通信协议。本领域的技术人员将会理解，这类网络计算环境通常包含许

多类型的计算机系统配置，其中包括个人计算机、手持装置、多处理器系统、基于微处理器或者可编程的消费电子产品、网络 PC、小型计算机、大型计算机等等。

[0019] 实施例还可在分布式计算环境中实施，在这些环境中，任务由通过通信网络链接（通过硬连线链路、无线链路或者通过硬连线或无线链路的组合）的本地和远程处理装置来执行。在分布式计算环境中，程序模块可位于本地和远程存储器存储装置中。

[0020] 一种用于实现示范实施例的整体或部分的示范系统可包括采取计算机形式的通用计算装置，其中包括处理单元、系统存储器以及将包括系统存储器在内的各种系统组件耦合到处理单元的系统总线。系统存储器可包括只读存储器 (ROM) 和随机存取存储器 (RAM)。计算机还可包括用于对磁硬盘进行读取和写入的磁硬盘驱动器、对可移动磁盘进行读取或写入的磁盘驱动器以及用于对可移动光盘、如 CDROM 或其它光介质进行读取或写入的光盘驱动器。驱动器及其关联的机器可读介质提供对机器可读指令、数据结构、程序模块和计算机的其它数据的非易失性存储。

[0021] 实施例中公开的方法的技术效果包括更有效地提供用于挖掘预测模式的复杂数据集的精确模型。该方法引入用于实质上使用单一算法从不同观点来探索数据的高度灵活性，其中单一算法被分派为解决不同问题。因此，技术效果包括更有效的数据探索、异常检测、用于预测值和取代缺失数据的回归以及数据的分段。如何能够使用所公开方法有效地探索这类数据的示例包括基于顾客的购买习惯的目标市场、通过识别有风险的信贷申请人来降低信贷风险以及通过了解飞行器的健康状态的预测维护。

[0022] 本发明涉及生成数据集的一般混合模型。更具体来说，数据集划分为两个或更多子集，为各子集生成子集混合模型，然后将子集混合模型组合以生成数据集的一般混合模型。

[0023] 现在参照图 1，公开生成一般混合模型 100 的方法。首先提供数据库 102 中包含的数据集连同子集标准 108，用于生成具有子集标识 104 的子集。具有组成数据集的数据库能够存储在电子存储器中。数据集能够包含多个维或参数，其中各维具有与其关联的一个或更多值。这些值能够是离散值或连续值。例如，数据集能够包括具有离散值 CFM56、CF6、CF34、GE90 和 GEnx 的燃气涡轮发动机的维。离散值表示通用电气公司制造和销售的燃气涡轮发动机的各种型号。数据集还能够包括称为机身的具有离散值 B737-700、B737700ER、B747-8、B777-200LR、B777-300ER 和 B787 的另一个维，表示数据集的燃气涡轮发动机维的燃气涡轮发动机能够安装到其上的各种机身。继续这个示例，数据集还可包括称为推力的具有诸如 18000 磅力至 115000 磅力 (80kN-512kN) 的范围中的值之类的连续值的维。

[0024] 子集标准 108 能够是能够用于过滤数据集的数据集的一个或更多维的一个或更多值。子集标准能够存储在关系数据库中或者通过任何其它已知方法来指定。一般来说，子集标准 108 由数据集的用户基于用户想要从数据集了解的内容来制定。子集标准 108 能够包含用于过滤和划分数据集中的数据的任何数量的单独标准。继续上面的示例，子集标准 108 可包括三个不同元素，例如安装在 B747-8 上的 GE90 发动机、安装在 B777-300ER 上的 GEnx 发动机以及安装在 B787 上的 GEnx。虽然这是具有三个元素的二维子集标准的示例，但是子集标准可包括一直到数据集中的维数的任何数量的维，并且可包含任何数量的元素。

[0025] 生成子集和子集标识 104 包括滤过数据集并且识别每个子集中的每个元素。子集的数量相当于选择标准中的元素的数量。过滤过程可通过运行于可访问包含数据库 102 的

电子存储器的处理器的计算机软件单元来实现。在过滤之后或同时，为每个子集指配子集标识符，以便区分子集及其组成元素与其它子集的每个及其组成元素。子集标识符能够是文本串或者识别在 104 生成的子集的任何其它已知方法。

[0026] 随后在 106 评估是否存在至少一个子集。如果不存在至少一个子集，则方法 100 返回到 108，以便接受产生至少一个子集的新的子集标准。如果存在至少一个子集，则方法 100 在 110 为每个子集生成混合模型。混合模型的生成在数据挖掘领域中又通常称作训练。每个子集的混合模型能够通过任何已知方法并且作为任何已知类型的混合模型来生成，一个非限制性示例是使用期望最大化 (EM) 所训练的高斯混合模型。为每个子集生成混合模型的过程产生表示子集密度的数学泛函。在对连续随机向量进行建模的示例中，每个子集的数学泛函表示是概率密度函数 (pdf) 的换算和 (scaled summation)。每个 pdf 对应于对其生成混合模型的子集中的数据元素的成分或簇。换言之，为每个子集生成混合模型的方法 110 由运行于处理器的软件单元来进行，其中软件单元考虑子集中的所有数据元素，将数据元素群集为一个或更多成分，将 pdf 拟合到每个成分，并且将换算因子归于每个成分，以便生成数据的数学泛函表示。混合模型的一个非限制性示例是如下形式的高斯或正态分布混合模型 ：

[0027] $$p(X)=\sum_{k=1}^{K}\pi_k N(X|\mu_k,\sum_k)$$

[0028] 其中 p(X) 是子集的数学泛函表示，

[0029] X 是变量的多维向量表示，

[0030] k 是表示子集中的每个成分的索引，

[0031] K 是子集中的成分的总数，

[0032] $\pi_k$ 是与簇 k 对应的标量换算因子，其中所有 K 簇的所有 $\pi_k$ 之和等于 1，

[0033] $N(X|\mu_k,\Sigma_k)$ 是成分均值 $\mu_k$ 和协方差 $\Sigma_k$ 的向量 X 的正态概率密度函数。

[0034] 如果向量 X 具有单维，则 $\Sigma_k$ 是 X 的方差，而如果 X 具有两个或更多维，则 $\Sigma_k$ 是 X 的协方差矩阵。

[0035] 在 110 生成每个子集的混合模型之后，在 112 确定是否存在至少两个子集。如果不存在至少两个子集，则在 110 生成的单一子集混合模型是一般混合模型。但是，如果在 112 确定存在至少两个子集，则随后在 116 确定是否需要模型成分的过滤。如果在 116 需要过滤，则在 118 从模型中去除一个或更多成分。结合图 2 更详细地描述 118 的过滤方法。一旦在 118 进行过滤，或者如果在 116 不需要过滤，则方法 100 进入 120，在 120 将子集模型组合。

[0036] 在 120 将子集模型组合能够包括连接对每个子集生成的混合模型，以便生成组合模型。备选地，将子集模型组合能够包括在连接每个混合模型以生成组合模型之前独立换算单独子集的每个混合模型。

[0037] 在 122，确定是否需要模型的简化。如果在 122 不需要简化，则在 124 组合子集模型是一般模型。如果在 122 需要简化，则在 126 执行组合模型的简化，并且在 128 认为经简化的组合模型是一般模型。简化 126 能够包括组合来自两个或更多不同子集的一个或更多簇。简化 126 还能够包括从子集的组合混合模型中去除一个或更多成分。

[0038] 现在参照图 2，描述在将子集混合模型组合之前在 118 过滤单独子集混合模型的

成分的方法。首先，在 140 清除用于将每个成分以及与其它成分的关联距离制成表格的完整列表。随后，在 142，由处理器和关联电子存储器接收来自所有子集的所有成分。在 144 选择来自所有成分的一个成分，并且在 146 确定所选成分与其它子集中的所有其它成分的距离。换言之，将所选成分与具有与所选成分的子集标识符不同的子集标识符的所有其它成分进行比较。距离能够通过任何已知方法来计算，包括但不限于库尔贝克 - 莱布勒 (Kullback Leibler) 发散。在 148，将成分以及与其它子集的所有其它成分的关联距离制成表格，并且附加到完整列表。换言之，完整列表包含从该成分到其它子集的所有成分的距离。在 150，确定所选成分是否为最后一个成分。如果不是的话，则方法 118 返回到 144 以选择下一个成分。但是，如果在 150 确定所选成分是最后一个成分，则对所有子集的所有成分来更新完整列表，并且方法进入 152，在 152 完整列表按照在 146 所计算的距离的降序来存储。在 154，去除或滤除完整列表上的顶部成分或者具有与所有其它子集的所有其它成分的最大距离的成分。在 156，确定是否满足过滤标准。例如，过滤标准能够是待过滤成分的预定总数。备选地，过滤标准能够是成分总数的预定百分比的过滤。如果在 156 满足过滤标准，则在 160 识别最终成分集。但是，如果在 156 不满足过滤标准，则在 158 确定是否需要迭代过滤。对迭代过滤的需要能够由方法 118 的用户来设置。如果在 158 不需要迭代过滤，则该方法返回到 154，以便从其余成分中去除具有与其它子集的所有其它成分的最大距离的成分。在 158，如果确定需要迭代过滤，则方法 118 返回到 140。

[0039] 迭代过滤意味着方法 118 重新计算每个成分与每另一个成分的距离，并且通过每次从混合模型中去除成分时运行 140 至 152，来生成新的完整列表。成分之间的距离能够发生变化，并且因此在从混合模型中去除成分时，完整列表上的成分的相对顺序能够发生变化。因此，通过运行迭代过滤，能够更可信地确保被去除的成分是具有与每另一个子集的成分的最大距离的成分。但是，在一些情况下，可能不希望运行迭代过滤，因为迭代过滤是更加计算密集的并且因此是更费时的。换言之，当运行本文所公开的过滤方法 118 时，可评估过滤性能与进行过滤所需时间之间的权衡以在 158 确定是否需要迭代过滤。

[0040] 图 3-6 示出运行生成一般混合模型的上述方法 100 的示例。图 3 中，相对变量 $x_1$ 来绘制来自数据集的数据 180 和 190。按照结合方法 100 的 104 所述的过程，将数据进一步分为在图表上示为空心圆的第一子集 180 以及在图表上示为实心三角形的第二子集 190。虽然方法 100 能够适用于具有许多子集的多变量分析，但是为了可视化方法 100 中的简洁起见，在本示例中示出仅具有两个子集的单变量数据相关性。

[0041] 图 4 和图 5 示出在步骤 110 分别对第一子集 180 和第二子集 190 的混合模型的生成。在第一子集 180 的情况下，识别三个成分，并且将各成分拟合到分别具有均值 $\mu_1$、$\mu_2$ 和 $\mu_3$ 的换算高斯分布 G1、G2 和 G3。在第二子集 190 的情况下，识别两个成分，并且将各成分拟合到分别具有均值 $\mu_4$ 和 $\mu_5$ 的换算高斯分布 G4 和 G5。因此，第一子集 180 的混合模型由组成成分 G1、G2 和 G3 的换算拟合函数的包络来表示。类似地，第二子集 190 的混合模型由组成成分 G4 和 G5 的换算拟合函数的包络来表示。图 6 中，示出过滤之后在方法 100 的步骤 120 的一般混合模型的组合组成换算拟合函数。在本示例中能够看到，在过滤步骤 118 中，发现具有拟合函数 G3 的成分处于超过某个预定值（未示出）的离开其它子集 G4 和 G5 的成分的距离，并且因此从图 6 的一般混合模型中去除成分 G3。

[0042] 本书面描述使用示例来公开包括最佳模式的本发明，并且还使本领域的技术人员

能够进行和使用本发明。本发明的可取得专利的范围由权利要求来定义，并且可包括本领域的技术人员想到的其它示例。如果这类其它示例具有与本权利要求的文字语言相同的结构单元，或者如果它们包括具有与本权利要求的文字语言的非实质差异的等效结构单元，则它们意在落入本权利要求的范围之内。

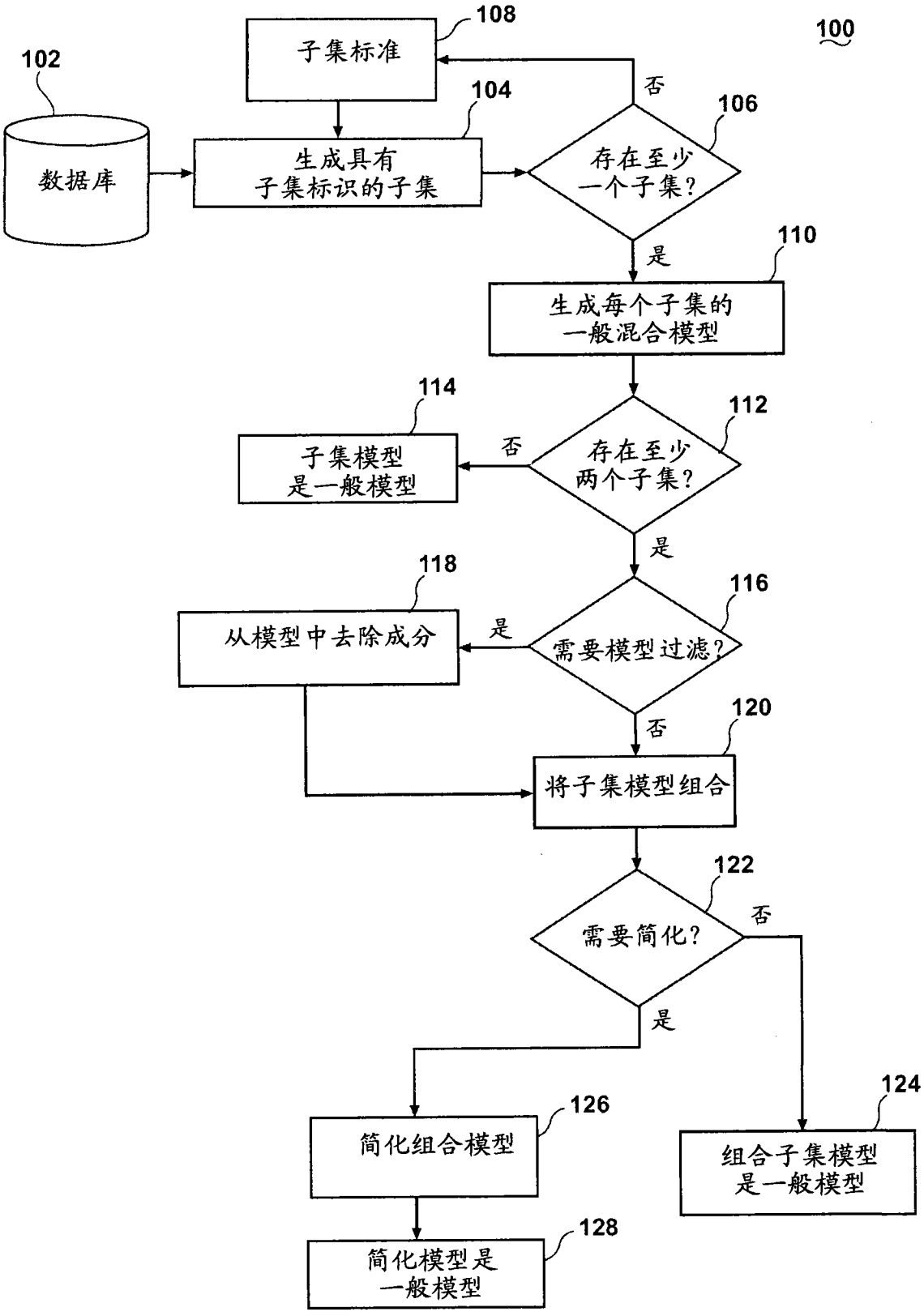

100
102
数据库
104
生成具有
子集标识的子集
106
存在至少
一个子集?
否
108
子集标准
是
110
生成每个子集的
一般混合模型
112
存在至少
两个子集?
否
114
子集模型
是一般模型
是
116
需要模型过滤?
是
118
从模型中去除成分
否
120
将子集模型组合
122
需要简化?
否
是
124
组合子集模型
是一般模型
126
简化组合模型
128
简化模型是
一般模型

图 1

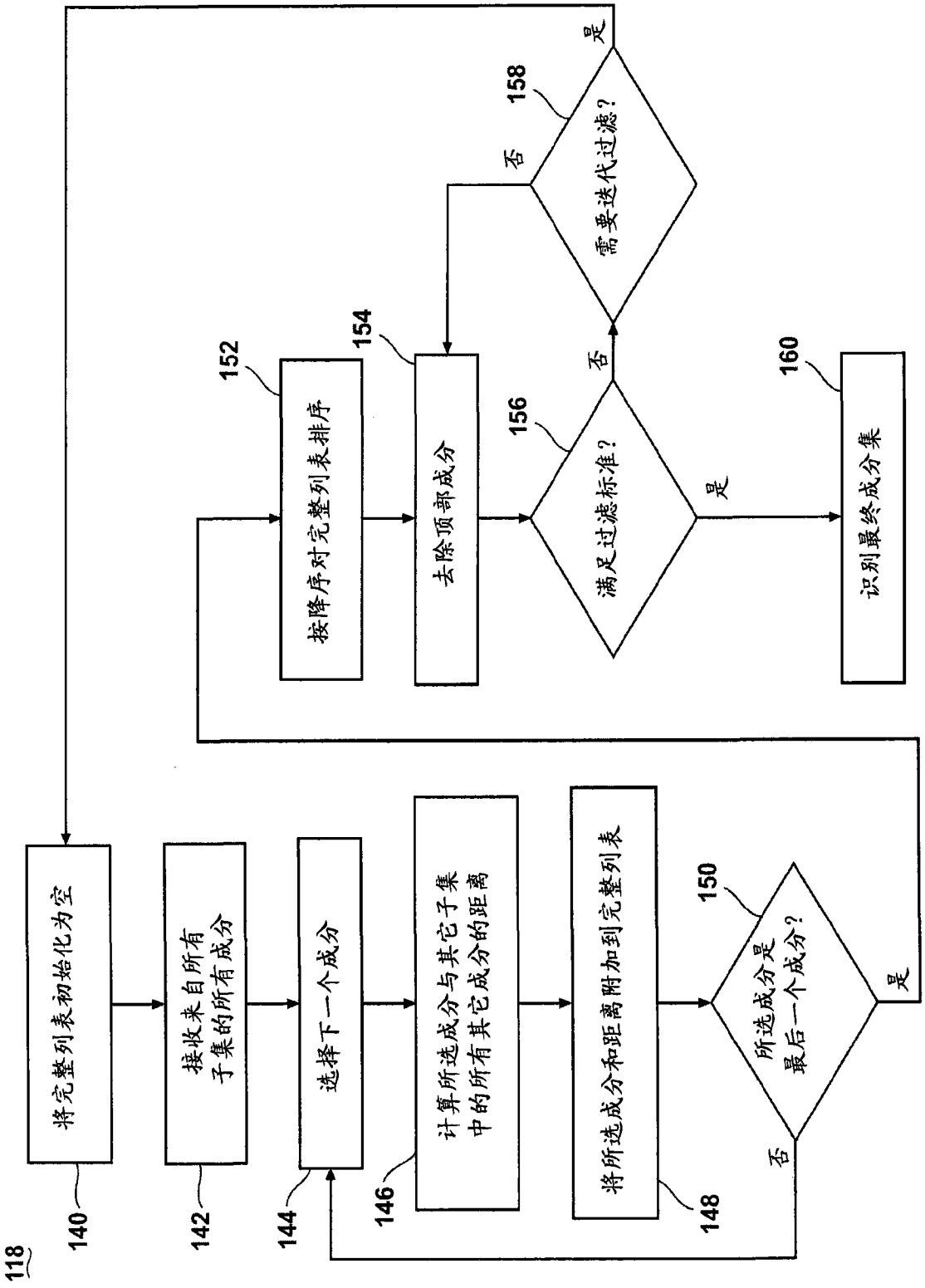

118
140
将完整列表初始化为空
142
接收来自所有子集的所有成分
144
选择下一个成分
146
计算所选成分与其它子集中的所有其它成分的距离
148
将所选成分和距离附加到完整列表
150
所选成分是最后一个成分?
否
是
152
按降序对完整列表排序
154
去除顶部成分
156
满足过滤标准?
否
是
158
需要迭代过滤?
否
是
160
识别最终成分集

图 2

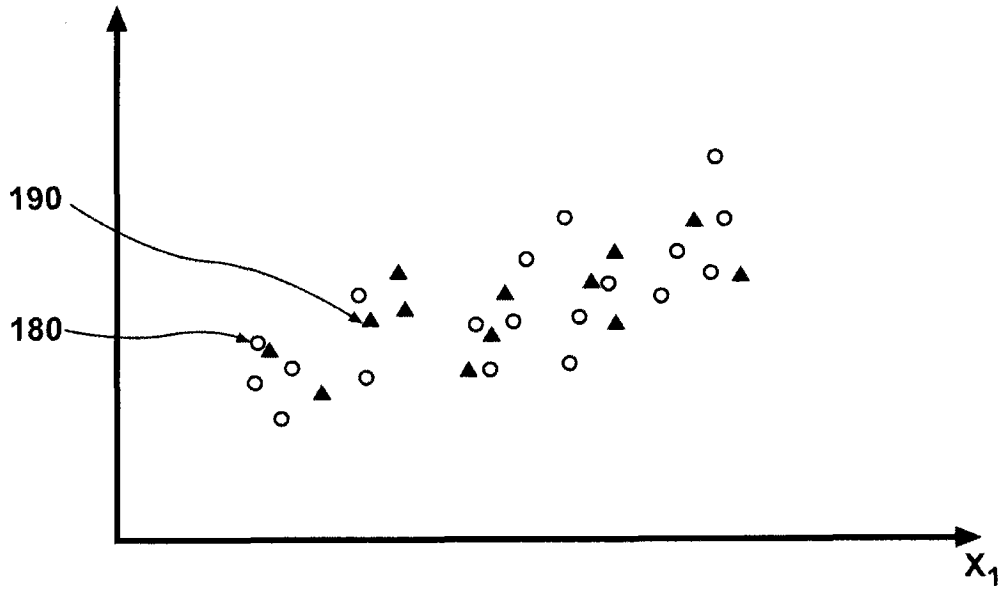



图 3

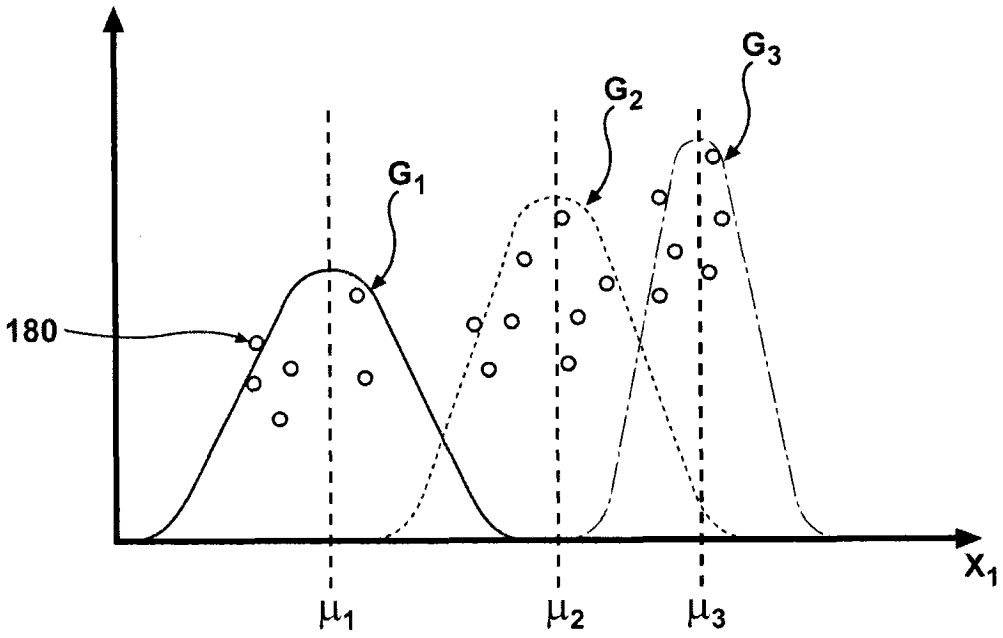



图 4

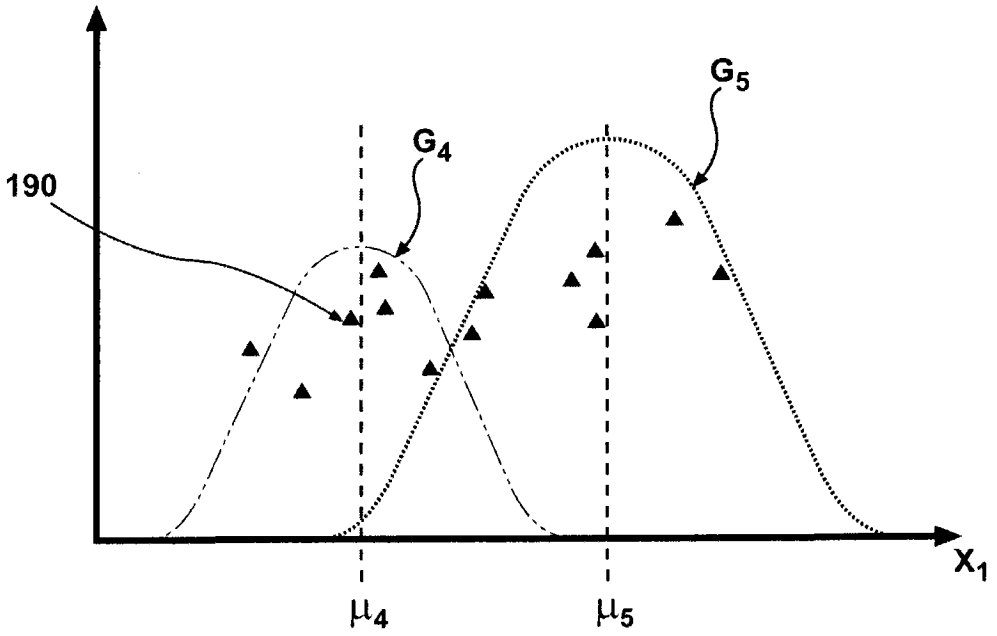



图 5

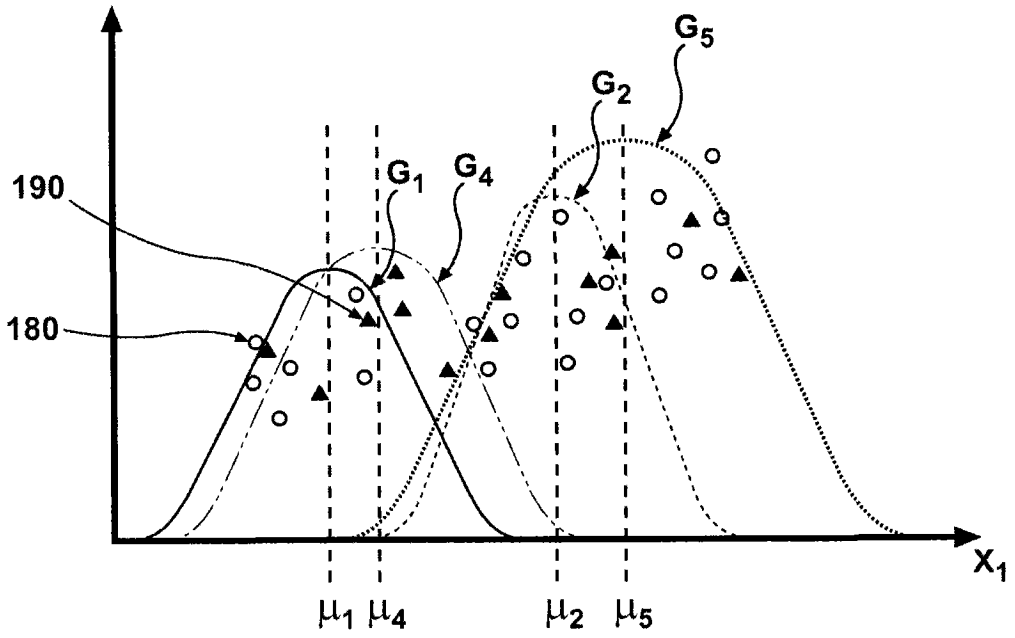



图 6