

(12) 发明专利申请

(10) 申请公布号 CN 102985923 A

(43) 申请公布日 2013. 03. 20

(21) 申请号 201180032445. 2

代理人 罗朋 周建华

(22) 申请日 2011. 06. 09

(51) Int. Cl.

(30) 优先权数据

G06F 17/30 (2006. 01)

12/824, 849 2010. 06. 28 US

(85) PCT申请进入国家阶段日

2012. 12. 28

(86) PCT申请的申请数据

PCT/US2011/039750 2011. 06. 09

(87) PCT申请的公布数据

W02012/009071 EN 2012. 01. 19

(71) 申请人 阿尔卡特朗讯

地址 法国巴黎市

(72) 发明人 陈爱友 雄明

(74) 专利代理机构 北京汉昊知识产权代理事务

所（普通合伙） 11370

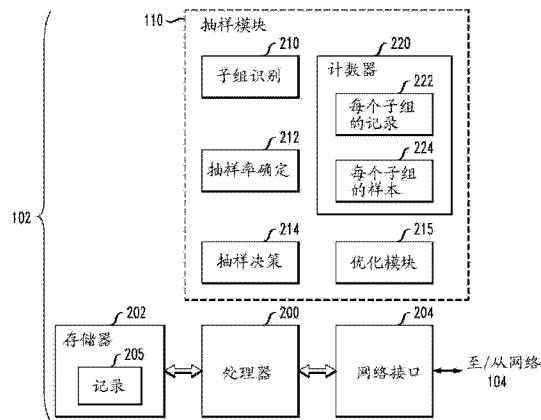
权利要求书 2 页 说明书 8 页 附图 5 页

(54) 发明名称

高维分层抽样

(57) 摘要

在一个方面中，信息处理系统的处理装置用于执行数据库的高维分层抽样，所述数据库包括多条布置在交叠子组中的记录。对于给定记录，处理装置确定给定记录与哪个子组相关联，对于与给定记录相关联的每个子组，检查子组的抽样率是否小于指定抽样率。如果每个子组的抽样率都小于指定抽样率，处理装置对所述给定记录抽样，否则不对给定记录抽样。针对其他记录重复确定、检查和抽样操作，处理抽样操作获得的样本以产生表征数据库的信息。本发明的其他方面涉及通过迭代地优化目标函数来确定对哪些记录抽样，目标函数可以基于例如被抽样记录的似然函数。



1. 一种设备，包括：

处理装置，包括具有关联存储器的处理器；

其中，所述处理装置用于：

对于给定记录，确定所述给定记录与多个子组中的哪个子组相关联；

对于与所述给定记录相关联的每个子组，检查所述子组的抽样率是否小于指定抽样率；

如果每个子组的抽样率都小于所述指定抽样率，对所述给定记录抽样，否则，不对所述给定记录抽样；以及

对于多个其他记录中的每个，重复所述确定、检查和抽样操作；

其中，由所述抽样操作所获得的样本被处理以产生表征包括所述子组的数据库的信息。

2. 根据权利要求 1 所述的设备，其中，所述处理装置包括具有抽样模块的控制器，所述抽样模块被配置成，对于所述给定记录和所述多个其他记录，执行所述确定、检查和抽样操作。

3. 根据权利要求 1 所述的设备，其中，所述子组包括所述数据库记录的交叠集合。

4. 根据权利要求 3 所述的设备，其中，所述处理装置还用于为每个子组维持第一计数器和第二计数器，所述第一计数器指示与所述子组相关联的记录数量，所述第二计数器指示来自所述子组的所述记录被抽样的次数。

5. 根据权利要求 4 所述的设备，其中，所述处理装置还用于为每个子组确定抽样率，所述抽样率根据为所述子组维持的所述第一计数器的值和为所述子组维持的所述第二计数器的值确定。

6. 根据权利要求 5 所述的设备，其中，所述处理装置还用于基于所述给定记录是否被抽样，为与所述给定记录相关联的每个子组更新所述第一和第二计数器中的至少一个。

7. 一种包括根据权利要求 1 所述的设备的集成电路。

8. 一种处理器实现的方法，包括如下步骤：

对于给定记录，确定所述给定记录与多个子组中的哪个子组相关联；

对于与所述给定记录相关联的每个子组，检查所述子组的抽样率是否小于指定抽样率；

如果每个子组的抽样率都小于所述指定抽样率，对所述给定记录抽样，否则，不对所述给定记录抽样；

对于多个其他记录中的每个，重复所述确定、检查和抽样步骤；以及

处理由所述抽样步骤获得的样本，以产生表征包括所述子组的数据库的信息。

9. 一种包括计算机可读存储介质的产品，所述计算机可读存储介质中包含可执行程序代码，在由处理装置的处理器执行时，所述代码令所述装置执行根据权利要求 8 所述方法的步骤。

10. 一种设备，包括：

处理装置，包括具有关联存储器的处理器；

其中，所述处理装置用于：

通过迭代地更新指定多个记录的各自记录是否被抽样的二进制指示符的分量，优化目

标函数,所述目标函数表征在交叠的记录子组中多个记录的那个记录被抽样;以及

基于所述二进制指示符的已更新分量优化所述目标函数的值,对所述多个记录的特定记录抽样;

其中,由所述抽样操作所获得的样本被处理以产生表征包括所述记录子组的数据库的信息。

11. 一种处理器实现的方法,包括如下步骤:

通过迭代地更新指定多个记录的各自记录是否被抽样的二进制指示符的分量,优化目标函数,所述目标函数表征在交叠的记录子组中多个记录的那个记录被抽样;

基于所述二进制指示符已更新分量优化所述目标函数的值,对所述多个记录的特定记录抽样;以及

处理由所述抽样步骤所获得的样本,以产生表征包括所述多个记录的数据库的信息。

高维分层抽样

技术领域

[0001] 本发明总体上涉及信息处理领域,更具体地,涉及对与信息处理系统数据库相关联的记录进行分层抽样的技术。

背景技术

[0002] 大型数据库常常包括数百万或更多记录,每条记录都具有很多属性。可以利用抽样技术对这样的数据库进行统计操作,一般涉及到从数据库随机选择记录。然后可以分析所选的记录以产生表征该数据库中完整记录集合的统计值。为了确保所得的统计值精确表征该数据库,可以使用分层抽样技术。在分层抽样中,数据库记录被分成子组或“层”,然后随机从每个子组中选择一个或多个记录加以分析。在题为“Stratified Sampling of Data in a Database System”的美国专利申请公开 No. 2002/0198863 中描述了常规分层抽样技术的范例。

[0003] 常规分层抽样技术的问题是这种技术通常试图将记录分成互相排斥的子组,因此可能仅考虑了有限数量的属性。通常每条记录的属性数量被称为该数据库的“维数”,常规的分层抽样技术实际上仅处于低维领域。不过,很多现代的数据库,例如用于跟踪电信应用中连接数据的那些数据库,具有非常高的维数。

[0004] 例如,考虑存储了 N 条记录的数据库,每条记录有 K 个属性,其中每个属性取 m_k 个离散值, $1 \leq k \leq K$ 。如果 K 很小,就能够简单地连锁这些属性,以便将数据库划分成互相排斥的子组。在这种情况下,由 $\prod_{k=1}^K m_k$ 给出子组的数目。不过,随着 K 变大,这种方法就变得不切实际了。例如,如果 $m_k = 5$ 且 $K = 10$,那么有接近 10^7 个子组,其中很多将不包含记录或仅包含少量记录。在这种高维上下文中,常规的分层抽样技术不能针对 K 个属性的每个提供适当的分层样本。在众多信息处理应用中,包括大规模数据库集成和维护、数据挖掘、数据仓储、查询处理、电信网络流量分析、意见调查等,这个问题都很明显。

发明内容

[0005] 本发明的例示性实施例提供了高维分层抽样技术,适用于每条记录的记录数量 N 和属性数量 K 都大的应用中。这些实施例包括顺序和最优高维分层抽样算法。前者对于在线抽样尤其有用,而后者对于离线或周期性抽样尤其有用,但两者都还可以用于各种其他抽样应用中。

[0006] 根据本发明的一个方面,信息处理系统的处理装置用于执行数据库的高维分层抽样,所述数据库包括多条布置在交叠子组中的记录。对于给定记录,处理装置确定给定记录与哪个子组相关联,对于与给定记录相关联的每个子组,检查子组的抽样率是否小于指定抽样率。如果每个子组的抽样率都小于指定抽样率,处理装置对所述给定记录抽样,否则,不对给定记录抽样。针对其他记录重复确定、检查和抽样操作,处理抽样操作获得的样本以产生表征数据库的信息。

[0007] 根据本发明的另一方面,信息处理系统的处理装置通过优化表征要对多条记录的

哪条抽样的目标函数来执行数据库的高维分层抽样，所述数据库包括多条布置在交叠子组中的记录。目标函数可以基于例如被抽样记录的似然函数，更具体而言，可以基于被抽样记录的似然函数的二项式 - 正态近似。通过迭代地更新二进制指示符的分量来进行目标函数的优化，二进制指示符指定是否对多条记录的相应记录抽样。处理装置基于二进制指示符的已更新分量优化目标函数的值对多条记录的特定记录抽样，处理所得的样本以产生表征包括记录子组的数据库的信息。

[0008] 例示性实施例相对于常规方法提供了显著优点。例如，可以使用例示性实施例中的顺序和最优高维分层抽样过程产生计算和存储要求最小的可靠无偏样本。

[0009] 从附图和以下详细描述，本发明的这些和其他特征和优点将变得更加明显。

附图说明

[0010] 图 1 是本发明例示性实施例中实施高维分层抽样的信息处理系统的方框图。

[0011] 图 2 示出了图 1 系统的处理装置的更详细视图。

[0012] 图 3 是本发明例示性实施例中的顺序高维分层抽样过程的流程图。

[0013] 图 4 是本发明例示性实施例中的最优高维分层抽样过程的流程图。

[0014] 图 5 示出了可以应用图 3 或 4 的高维分层抽样过程的网络流量应用中一组连接记录的简单范例。

[0015] 图 6 是比较顺序和最优高维分层抽样作为抽样率的函数的估计误差与常规随机抽样估计误差的一组曲线。

[0016] 图 7 示出了多组曲线，每组曲线比较了顺序和最优高维分层抽样作为子组数量的函数的估计误差与常规随机抽样估计误差。

具体实施方式

[0017] 这里将结合示范性信息处理系统、处理装置和高维分层抽样技术例示本发明。不过要理解，本发明不限于用于所公开的特定类型的系统、装置和技术。例如，可以利用除结合例示性实施例描述的那些之外的处理装置和工序在各种各样的其他信息处理系统配置中实现本发明的各方面。

[0018] 图 1 示出了信息处理系统 100，包括通过网络 104 耦合到数据库系统 105 的控制器 102，数据库系统包括多个服务器 106-1、106-2、……106-N，也表示为服务器 1、服务器 2、……服务器 N。服务器 106 的每个都具有关联的数据库 108。这些数据库存储记录或其他数据对象，供控制器 102 经由网络 104 访问。本实施例中的控制器 102 包括抽样模块 110，配置成实施下文更详细描述的一种或多种高维分层抽样技术。抽样模块 110 利用高维分层抽样技术处理记录集合，记录集合被分成未必互相排斥的子组。由抽样模块 110 处理的记录可以从数据源 112 接收或从数据库系统 105 的一个或多个数据库 108 检索。可以由控制器 102 在样本数据库 114 中存储得到的分层样本。尽管在图中被示为与数据库系统 105 分开，但也可以在数据库系统 105 之内实现诸如控制器 102 和样本数据库 114 的系统元件。

[0019] 控制器 102 可以包括适于通过网络 104 与数据库系统 105 通信的计算机或任何其他类型处理装置的至少一部分。例如，控制器可以包括便携式或膝上型计算机、移动电话、个人数字助理 (PDA)、无线电子邮件装置、电视机顶盒 (STB) 或其他通信装置。

[0020] 网络 104 可以包括诸如因特网的广域网、城域网、局域网、有线电视网、电话网、卫星网络以及这些或其他网络的部分或组合。

[0021] 在其他实施例中,可以在一个或多个服务器 106 中或其关联的数据库 108 中,或在耦合到这些元件中一个或多个的独立集中式控制器中实现抽样模块 110。也可以通过分布式方式实现抽样模块,模块的各部分分布置于装置 102、106 或 108 的相应装置或其子集中。

[0022] 数据库 108 不必是任何特定配置,因此这里使用的术语“数据库”意在被宽泛解释为涵盖存储记录的任何数量的不同布置。

[0023] 现在参考图 2,示出了系统 100 的控制器 102 的一种可能实施方式。在这一实施例中,控制器包括耦合到存储器 202 的处理器 200,还包括网络接口电路 204。假设存储器 202 存储记录 205 或其部分以供抽样模块 110 处理。存储的记录 205 可以从数据源 112 接收或通过网络 104 从数据库系统 105 检索。在这种实施方式中,控制器 102 的抽样模块 110 包括子组识别模块 210、抽样率确定模块 212、抽样决策模块 214、优化模块 215 和一组计数器 220,包括统计每个子组的记录数目的计数器 222 和统计每个子组的样本数目的计数器 224。下文将结合图 3 和 4 更详细地描述这些模块和计数器的操作。

[0024] 可以将处理器 200 实现为微处理器、微控制器、专用集成电路 (ASIC) 或其他类型的处理装置,以及这种装置的部分或组合。存储器 202 可以包括电子随机存取存储器 (RAM)、只读存储器 (ROM)、基于磁盘的存储器或其他类型的存储装置以及这种装置的部分或组合。可以在存储和执行用于高维分层抽样,以及用于执行相关操作,例如与记录存储和处理相关联的那些操作的一种或多种软件程序时使用处理器和存储器。因此可以至少部分利用这样的软件程序实现模块 210、212、214 和 215。可以将存储器 202 看做这里更一般地称为计算机程序产品或更一般地称为计算机可读存储介质(其中包含可执行程序代码)的范例。计算机可读存储介质的其他范例可以包括处于任何组合形式的磁盘或其他类型的磁性或光学介质。

[0025] 处理器 200、存储器 202 和接口电路 204 可以包括经适当修改以工作于这里所述方式下的公知常规电路。而且,可以将图 2 中所示的各种模块视为用于实现关联功能的电路范例。例如,这样的电路的部分可以包括矩阵乘法电路或其他类型的运算逻辑电路。这种电路的常规方面是本领域的技术人员公知的,因此这里将不会详细描述。

[0026] 要认识到,可以利用除图 1 和 2 的示范性布置中具体示出的那些之外的部件和模块实现这里公开的信息处理系统和关联控制器。

[0027] 现在将参考图 3 和 4 的流程图描述例示性实施例中的系统 100 的操作。这些流程图例示了相应的顺序和最优高维分层抽样技术。对于这些实施例,将假设将抽样技术用于存储 N 条记录的数据库,每条记录有 K 个属性,其中每个属性取 m_k 个离散值, $1 \leq k \leq K$ 。这个被抽样的数据库例如可以包括数据库系统 105 中的一个或多个数据库 108 或整个数据库系统 105。应当指出,在这种语境中,N 是指所存储记录的总数,不是如图 1 语境中那样的服务器 106 和数据库 108 的数量。

[0028] 一般由感兴趣字段及其组合的类别预定义子组。在以下描述的部分中,不带限制性地假设,记录的每个子组都针对一个属性取 m_k 个离散值或类别值(对于连续的属性,可以将它们离散化或分类成 m_k 个值)的特定一个,从而共有 $J = \sum_{k=1}^K m_k$ 个子组或层。因此,在这些实施例中子组可以具有很多交叠的记录。这与常规的分层抽样不同,如前所述,常规的

分层抽样将记录分成互相排斥的子组。应当指出,对于大规模的复杂数据库 J 可能非常大。
[0029] 而且,子组的数量 J 可能大于上述假设下得到的 $\sum_{k=1}^K m_k$ 个子组:每个子组针对一个属性取 m_k 个离散或类别值的特定一个。例如,可能通过取超过一个属性的组合来定义子组。在很多典型实际应用中,多个属性的这种组合都可能很重要。因此,J 可能大于 $\sum_{k=1}^K m_k$ 但小于 $\prod_{k=1}^K m_k$ 。

[0030] 可以如下用公式表示记录和子组之间的关系。令 A 是 $N \times J$ 的二元矩阵,其中 A_{ij} 表示第 i 条记录是否是第 j 个子组的部分, $i = 1, \dots, N, j = 1, \dots, J$ 。为简单起见,假设每条记录都属于至少一个子组,于是 A 的每排必须包含至少一个 1。令 $c \in \{0, 1\}^N, \sum_{i=1}^N c_i = n$, 其中 n 是被抽样的记录数量,N 是要从其抽样的记录数量,使得 c_i 表示第 i 条记录是否被抽样。令

$$[0031] n_j = \sum_{i=1}^N A_{ij}$$

$$[0032] s_j = \sum_{i=1}^N c_i A_{ij}$$

[0033] 分别是对于第 j 个子组而言的记录数量和被抽样记录数量。由于 J 可能很大,可以将当前语境中高维分层抽样的目标表征为选择,使得 $s_j \approx n_j p, j = 1, \dots, J$ 。如前所述,这里提到的用于高维分层抽样的两种不同技术为顺序和最优高维分层抽样,并分别结合图 3 和 4 描述。

[0034] 重要的是指出上述 $N \times J$ 的二元矩阵 A 通常非常稀疏,从而 A 可能被存储于紧凑的存储空间中。而且,可以利用本领域的技术人员公知的稀疏矩阵运算有效率地进行利用了 A 稀疏性的计算。

[0035] 图 5 中示出了网络流量应用中一组连接记录的简单范例,其中可以应用图 3 或 4 的高维分层抽样处理。在本范例中,示出了三条连接记录,每条记录都包括用于开始时间、终止时间、连接类型、失败呼叫尝试 (FCA) 和最强导频的字段。不过要认识到,可以将这里公开的技术应用于任何类型的记录,不要求使用任何特定的记录格式。因此,这里使用的术语“记录”应当做宽泛的解释,以便涵盖所存储数据或其他数据对象的众多不同布置。

[0036] 现在参考图 3,示出了流程图,例示了在图 1 的系统 100 中实施的顺序高维分层抽样过程。本实施例中的抽样过程一般涉及顺序处理记录,例如,在从数据源 112 接收这样的记录时。这是一种“在线”抽样,因为可以在每条新记录可用时以顺序方式实时对记录抽样。该过程包括图示的步骤 300 到 312。

[0037] 在步骤 300 中,获得要考虑进行抽样的下一记录。如前所述,这条记录可以是从数据源 112 之一接收的新记录,要存储于数据库 108 之一中。在一些实施例中,可以随机变更考虑抽样的记录次序,以便确保抽样不会被诸如本地存储器结构的因素影响。

[0038] 在步骤 302 中,确定这一特定记录属于 J 个子组的哪个。在这一实施例中,假设子组是以上述方式预定的。在其他实施例中,可以利用诸如关联规则采掘算法的技术确定子组。

[0039] 在步骤 304 中,确定针对记录所属的每个子组的抽样率是否小于指定的抽样率 p。利用每个子组计数器 222 记录的对应一个和每个子组计数器 224 样本的对应一个为给定子

组确定抽样率。每个子组计数器的记录给出了以作为该子组一部分的记录数衡量的子组大小度量。每个子组计数器的样本给出了子组被抽样的次数。确定子组被抽样次数除以作为子组一部分的记录数目作为子组的抽样率。针对包括考虑要抽样的记录的每个子组独立确定这一抽样率。

[0040] 如果针对记录所属的每个子组的抽样率小于指定抽样率 p , 如步骤 306 中所示对记录抽样。否则, 如步骤 308 中所示, 不对记录抽样。于是, 当且仅当对于记录所属的每个子组, 由指定抽样率 p 界定所实现抽样率的上限时, 才对考虑中的给定记录抽样。

[0041] 然后该过程前进到步骤 310, 以为被抽样或未抽样记录所属的子组更新适当的计数器。然后在应用于下一条考虑抽样的记录时在过程的下一次迭代中使用更新的计数器。如果在步骤 306 中对样本抽样, 对于该记录所属的每个子组, 更新每个子组记录的计数器 222 的对应一个和每个子组样本的计数器 224 的对应一个。不过, 如果不对考虑的记录抽样, 使得过程通过步骤 308 到达步骤 310, 仅需要更新每个子组计数器的记录, 因为每个子组的样本数量将不变。

[0042] 在步骤 312 中, 确定是否有更多记录要处理。如果有更多记录, 该过程返回到步骤 300 以获得考虑进行抽样的下一记录。否则, 如图所示结束该过程。

[0043] 一旦已经利用图 3 的过程产生了给定一组记录的适当样本, 样本就可以被存储在样本数据库 114 中或系统 100 中的别处, 并用于执行回归分析、数据采掘或其他功能。更一般地, 进一步处理样本操作的结果以产生表征包括记录子组的数据库的信息。

[0044] 现在参考图 4, 示出了流程图, 例示了在图 1 的系统 100 中实施的最优高维分层抽样过程。本实施例中的抽样过程一般涉及按组处理记录, 例如, 在从一个或多个数据库 108 检索时或在从一个或多个数据源 112 接收时。可以将此视为一种“离线”或周期抽样, 因为不是像图 3 实施例那样顺序处理记录。该过程包括图示的步骤 400 到 410。

[0045] 在最优抽样过程中, 优化目标函数获得期望的抽样方案。一种可能的目标函数是根据 c 使函数 $\sum_j^J (s_j - n_j p)^2$ 最小化。这是一个二次范数, 容易忽略 n_j 小的层, 因此在特定应用中不适宜。另一种可能性是使相对误差 $\sum_{j=1}^J \left(\frac{s_j - n_j p}{n_j} \right)^2$ 最小化, 这更多注重小层。不过, 作为这两个可能目标函数的替代, 下面将描述在大层和小层之间实现良好折衷的目标函数。

注意, 如前所述, 针对每个子组的样本大小 s_j 遵循二项式分布。通过独立处理每个子组, 能够如下表示由样本的似然函数给出的二项式目标函数 :

$$[0046] \prod_{j=1}^J \binom{n_j}{s_j} p^{s_j} (1-p)^{n_j - s_j}$$

[0047] 其中 n_j 是第 j 个子组的大小。注意, 对子组的独立性假设并不表示子组不交叠。相反, 仅仅意味着每个子组可能涉及到记录的任意子集, 与什么记录与其他子组相关联无关。因此, 这暗中假设与不同子组相关联的记录间有随机交叠。似然函数的最大化将获得关于要对哪些记录抽样的方案。

[0048] 基于二项式 - 正态近似, 即, s_j 大致遵守正态分布 $N(n_j p, n_j p(1-p))$, 可以如下用公式表示对应的正常目标函数 :

$$[0049] L(c) = \sum_j \frac{(s_j - n_j p)^2}{n_j p(1-p)}$$

[0050] 这是基于正态近似的 $\{s_j : 1 \leq j \leq J\}$ 的对数似然函数（直到常数）。注意，二项式和正常目标函数之间有两个主要差别。首先，正常目标函数是加权的平方和，其中针对由其大小加权，使小子组降级的子组，由 $n_i^{-1} |s_i p^{-1} - n_i|$ 定义相对估计误差。因此，它比二项式目标函数更加直观。 $L(c)$ 的小值意味着小的相对估计误差。第二，由于 $s_j = A_j^T c$ ，其中 $c \in \{0, 1\}^n$ 是二元矢量，未知参数，表示是否对记录抽样，正常目标函数是 c 的二次形式，使得正常目标函数的优化比在二项式情况下更简单。由于这些优点，图 4 的实施例利用了正常目标函数 $L(c)$ 。可以将 $L(c)$ 的二次形式写为：

$$[0051] L(c) = \sum_j \frac{(A_j^T c - n_j p)^2}{n_j p(1-p)}.$$

[0052] 然后，相对于 c 对 $L(c)$ 进行最小化获得抽样方案，这里称为优化的抽样。

[0053] 应当指出，这里使用的术语，例如“最优”和“优化”不需要实现任何特定绝对最小值或绝对最大值，而是应做宽泛的解释，以涵盖例如在指定界限之内或有指定剩余误差的情况下实现最小值或最大值。

[0054] 在图 4 的抽样过程的步骤 400 中，用公式表示前述 $N \times J$ 的二元矩阵 A ，其中 A_{ij} 表示第 i 条记录是否是第 j 个子组的部分， $i = 1, \dots, N, j = 1, \dots, J$ 。同样，假设每条记录属于至少一个子组，从而 A 的每排必须包含至少一个 1。

[0055] 在步骤 402 中，将 c_i 指定为第 i 条记录是否被抽样的二进制指示符。如上所述， $c \in \{0, 1\}^N, \sum_{i=1}^N c_i = n$ 。

[0056] 在步骤 404 中，基于如前所述的二项式 - 正态近似，用公式表示上述正常目标函数 $L(c)$ 。

[0057] 在步骤 406 中，优化目标函数 $L(c)$ ，更具体地，相对于 c 进行最小化，以提供期望的抽样方案。这一特定的最小化问题是一种二元二次优化问题，通常是 NP 困难的。解决这种的优化问题的已知算法包括模拟退火和禁忌搜索，但困难非常耗时。相反，步骤 406 中实施的优化利用了迭代过程，对于 $i = 1, \dots, n$ ，固定除 c_i 之外 c 的所有成分，并根据 $c_i = 1$ 或 $c_i = 0$ 是否给出更小的值来更新 $L(c)$ 。在目标函数单调减小时，迭代步骤收敛于局部解。可以迅速达到局部收敛，即，每个 c_i 将仅需要更新几次，这不会导致很大的计算负担。应当指出，可以向记录组应用图 3 的高维顺序抽样过程或常规随机抽样以提供 c 的良好初始化点，对 $L(c)$ 进行最小化。还可以使用替代技术，例如频谱近似，以获得 c 的初始值。

[0058] 在步骤 408 中，基于在优化步骤 406 中确定的 c 中的值对记录抽样。

[0059] 在步骤 410 中，确定是否有更多记录要处理。如果有更多记录，该过程返回到步骤 400 以获得考虑进行抽样的其他记录。否则，如图所示结束该过程。

[0060] 像在图 3 的过程中那样，一旦已经利用图 4 的过程产生了给定一组记录的适当样本，样本就可以被存储在样本数据库 114 中或系统 100 中的别处，并用于执行回归分析、数据采掘或其他功能。这样的函数是样本操作的结果可以被进一步处理以产生表征包括记录子组的数据库的信息的方式范例。

[0061] 在很多实际应用中，记录通常是相继到达的， N 可能极大。因此，可以定期应用图 4

的最优过程以获得样本,然后将那些样本与使用图 3 的顺序过程获得的一个或多个较早样本合并。于是,本发明的其他实施例可以组合图 3 和 4 的顺序和最优抽样过程或这种过程的部分。

[0062] 现在将描述图 3 和 4 的抽样过程的性能模拟。在这些模拟中,性能是采样率 p 和记录子组矩阵 A 的函数。使用参数 $r \in (0, 1)$ 表征两个子组之间的相关,其中矩阵 A 的每项 A_{ij} 被产生为独立的伯努利随机数,即 $P(A_{ij} = 1) = r$ 。然后由取值 1 的 A 项确定子组。每个子组的大小具有期望值 $m = nr$,两个子组具有等于 $nr^2 = mr$ 的交叠记录的期望值。因此两个子组之间的余弦相关大约为 r 。在 $r = 0$ 时,子组是互相排斥的,在 r 接近 1 时,每对子组有很多交叠的记录。不过, k 个不同子组间交叠记录的期望数为 nr^k ,随着 k 增大,其呈指数形式衰减。

[0063] 图 6 示出了比较顺序和最优高维分层抽样作为抽样率 p 的函数的估计误差与常规随机抽样估计误差的一组曲线。在这些曲线中, $n = 10^6$, $J = 10000$, $r = .001$,且 p 从 .001 变化到 .5。从这些曲线可以看出,常规随机抽样的估计误差不会随抽样率改变。不过,顺序和最优高维分层抽样的估计误差都远好于随机抽样,除了极小的抽样率之外。此外,随着抽样率增大,顺序和最优高维分层抽样的估计误差都迅速减小,相对于 p^{-1} 稍快于线性。还可以看到,最优抽样的性能一致好于顺序抽样,尤其对于小抽样率而言。

[0064] 图 7 示出了比较顺序和最优高维分层抽样作为子组数量 J 的函数的估计误差与常规随机抽样估计误差的多组曲线。在这些曲线中, $n = 10^6$, p 从 .001 变到 .1, r 从 2^{-14} 变化到 2^{-6} ,对应于从低于 100 到 20000 的子组大小。可以看出,就一切情况而论,随机抽样和最优抽样的估计误差都随着 J 几乎线性地增大,在对数刻度下斜率几乎为 1,而顺序抽样的斜率稍大,即,在 J 增大时,衰减得比其他两者更快。就一切情况而论,除了 p 和 r 都小时,从随机抽样到最优抽样都有显著的误差减小。顺序抽样的性能通常好于随机抽样,但不如最优抽样,除了 p 和 r 都非常小时或 J 大时。

[0065] 如前所述,可以在各种应用中实施这里公开的高维分层抽样技术。例如,可以在涉及连接记录的数据库查询和维护应用中使用这些技术,连接记录是针对无线网络中的每次呼叫产生的。这种网络中的连接记录数据库可以包括几百个属性。需要定期更新数据库,因为新的记录以每天几百万左右的速率到达。典型地,由于容量大,不能在数据库中长时间保持记录。因此,有益的是有一种样本数据库能够覆盖更长的记录历史(例如几个月)而且代表完整的数据库。在这样的应用中,可能希望对记录抽样,从而代表在每个时段(例如 5 分钟时段)和每个位置(例如城市的分区)做出的连接,并基于它们在完整记录中的比例对每类失败连接抽样。样本记录还应当代表与呼叫失败的根源相关的因素,例如会话建立的类型、会话建立阶段中的信号特征、连接已建立阶段中的信号特征、话务量、导频数量等。同样重要的是表示多个因素之间的相关,例如表示连接失败但信号强度很强且邻近基本基站的记录。这些变量的组合能够获得数以万计的交叠子组。其他示范性应用包括以界定的精确度高效率地处理对指定数据立方的查询并在大的群体获取的意见测验中产生无偏样本。

[0066] 可以使用上述例示性实施例中的顺序和最优高维分层抽样过程产生计算和存储要求最小的可靠样本。在不能访问完整的记录集合(例如在民意测验中不能从所有客户收集信息)或完整记录集合太大,系统不能为所有查询给出精确回答(例如,大的综合性数据库或网络数据)时,这样能够高效率地集成不同的信息源并产生成本可承受的样本。所得

的样本大致是不偏的，允许进行精确的后期分析。

[0067] 如前所述，可以至少部分以一个或多个软件程序的形式实现本发明的实施例，软件程序存储于存储器或信息处理系统的处理装置的其他计算机可读介质中。可以至少部分利用软件程序实现诸如模块 210、212、214 和 215 的系统部件。当然，在根据本发明实现这些和其他系统元件时可以利用任何组合中的硬件、软件或固件的众多替代布置。例如，可以在一个或多个现场可编程门阵列 (FPGA)、ASIC、数字信号处理器或其他类型的集成电路装置中，通过任何组合实现本发明的实施例。这样的集成电路装置以及其部分或组合，是“电路”的范例，如这里使用后一术语那样。

[0068] 再次应该强调，上述实施例仅仅出于例示的目的，不应被解释为以任何方式进行限制。其他实施例可以根据特定分层抽样应用的需要使用不同类型和布置的系统部件。因此，替代实施例可以在希望为记录集合实施精确和有效率抽样的其他语境中利用这里描述的技术。而且，还应当指出，在描述例示性实施例的语境中做出的特定假设不应被解释为本发明的要求。可以在这些特定假设不成立的其他实施例中实现本发明。在所附权利要求的范围之内的这些和众多其他替代实施例对于本领域技术人员而言将是显而易见的。

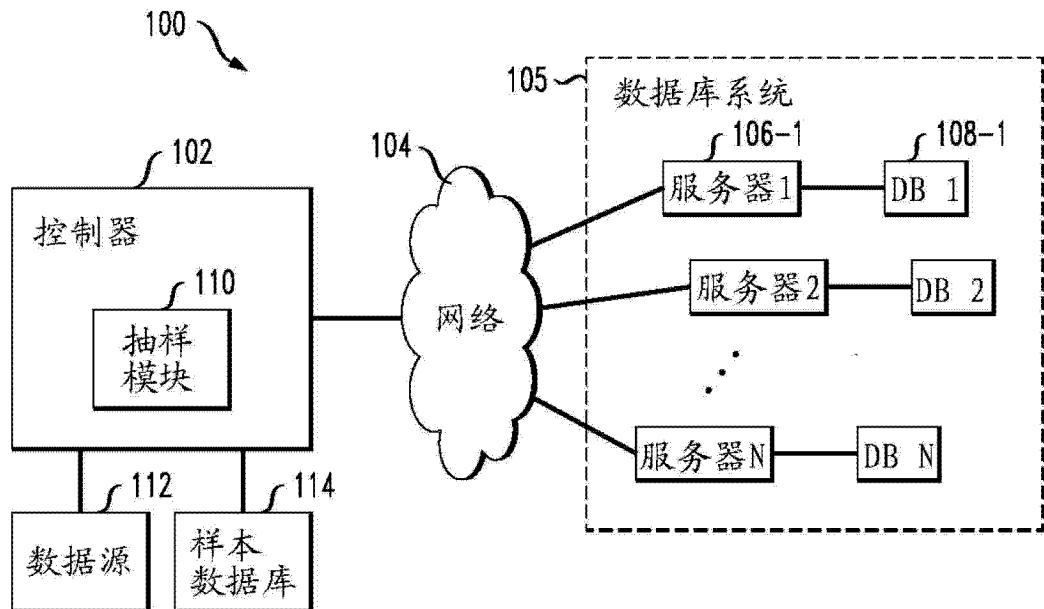


图 1

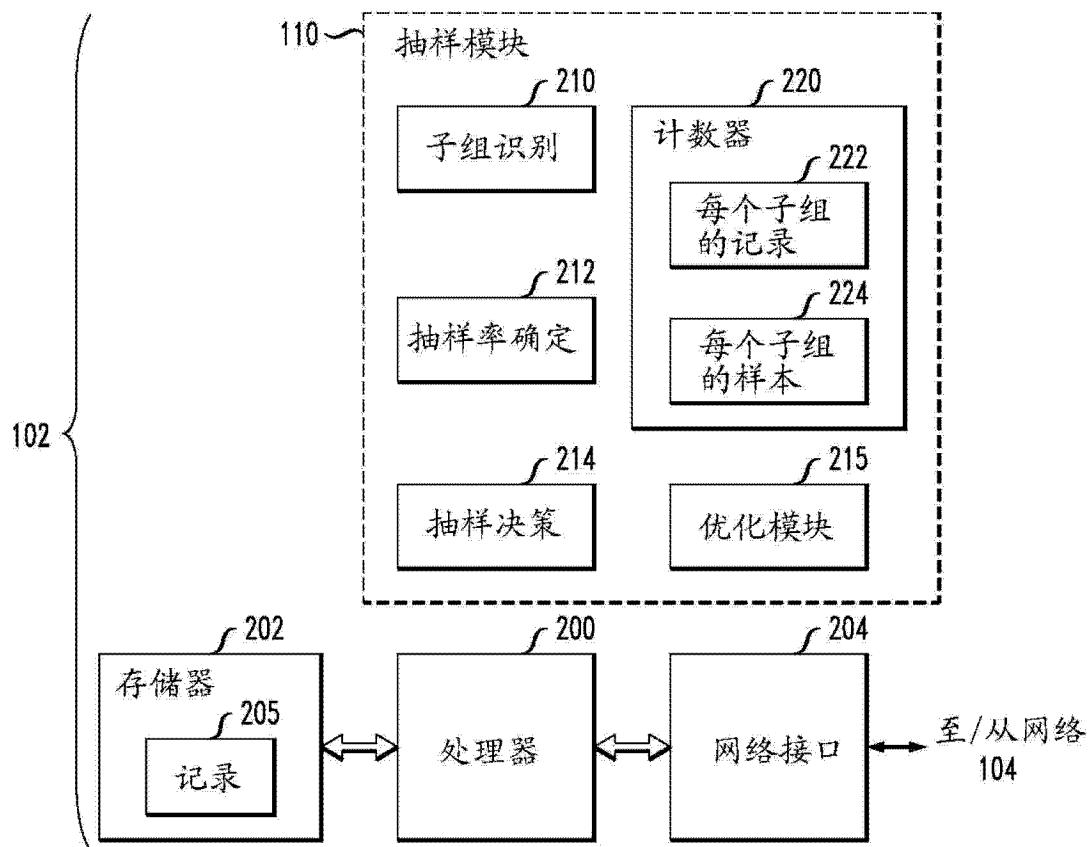


图 2

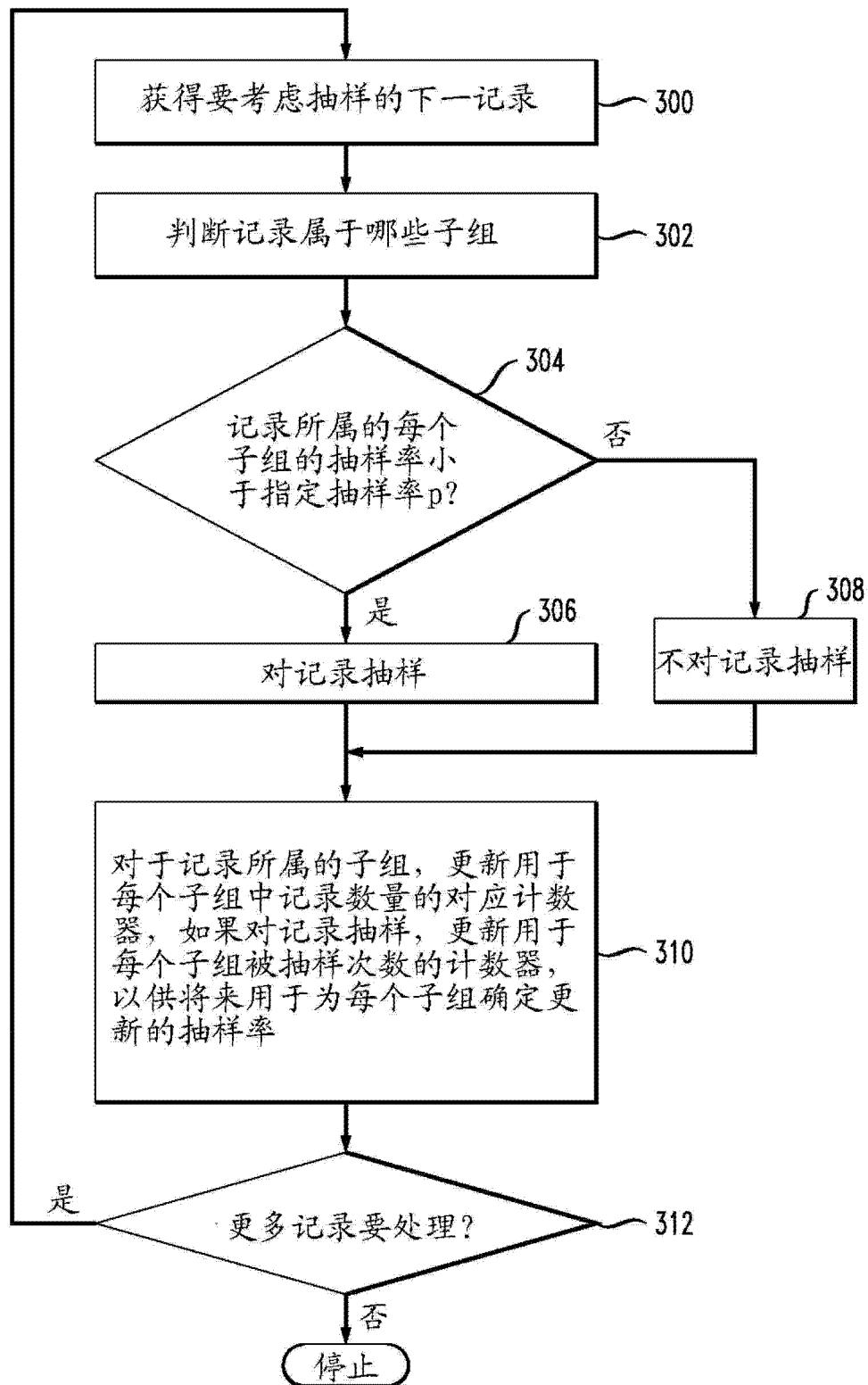


图 3

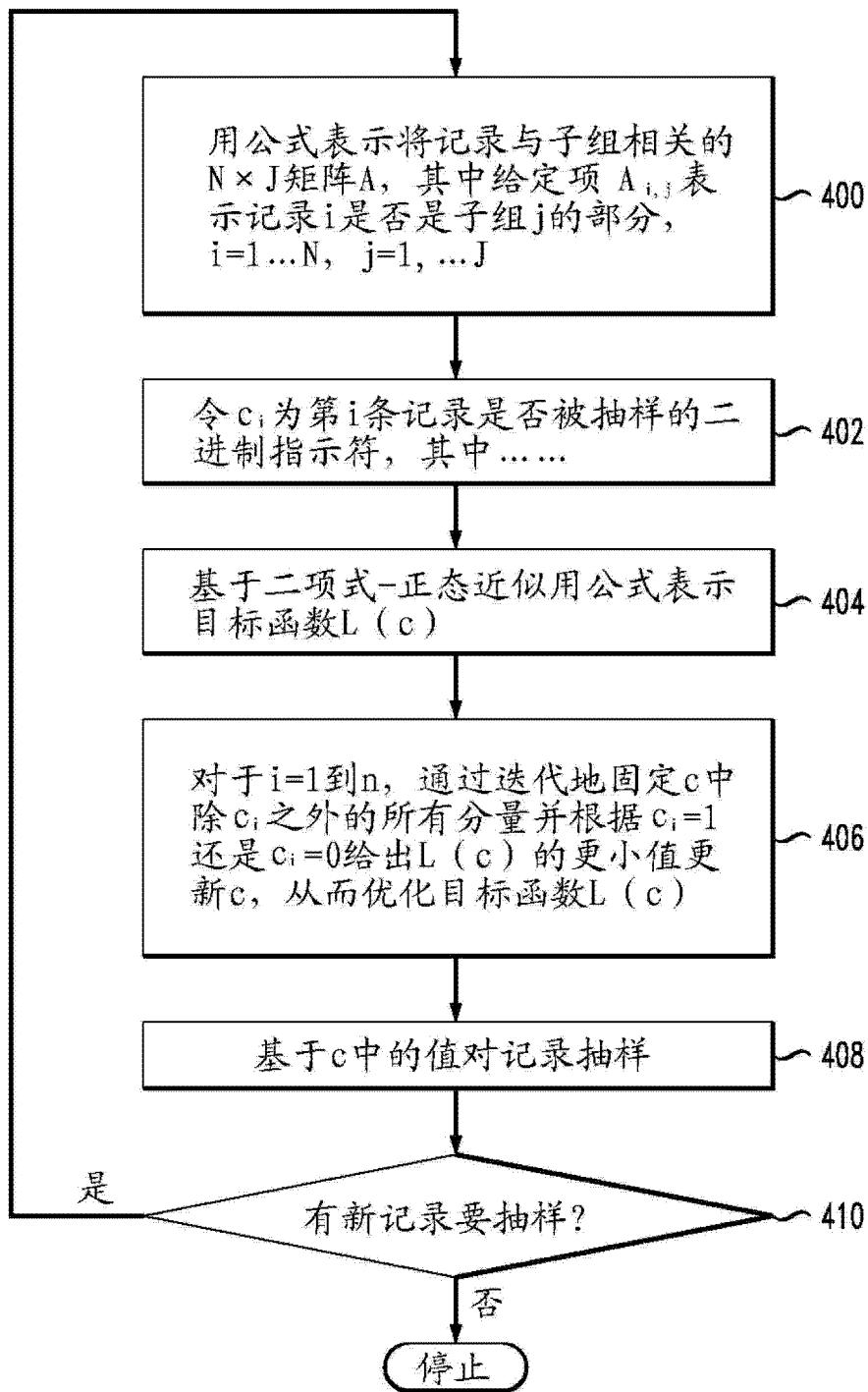


图 4

连接记录	开始时间	终止时间	类型	FCA	最强导频	...
1	05/10/2009 03:20:10	05/20/2009 03:20:10	会话建立	0	-3 dB	
2	05/10/2009 03:20:12	05/10/2009 03:20:13	会话建立	1	-9 dB	
3	05/10/2009 03:20:15	05/10/2009 03:21:20	用户初始	0	-1 dB	

图 5

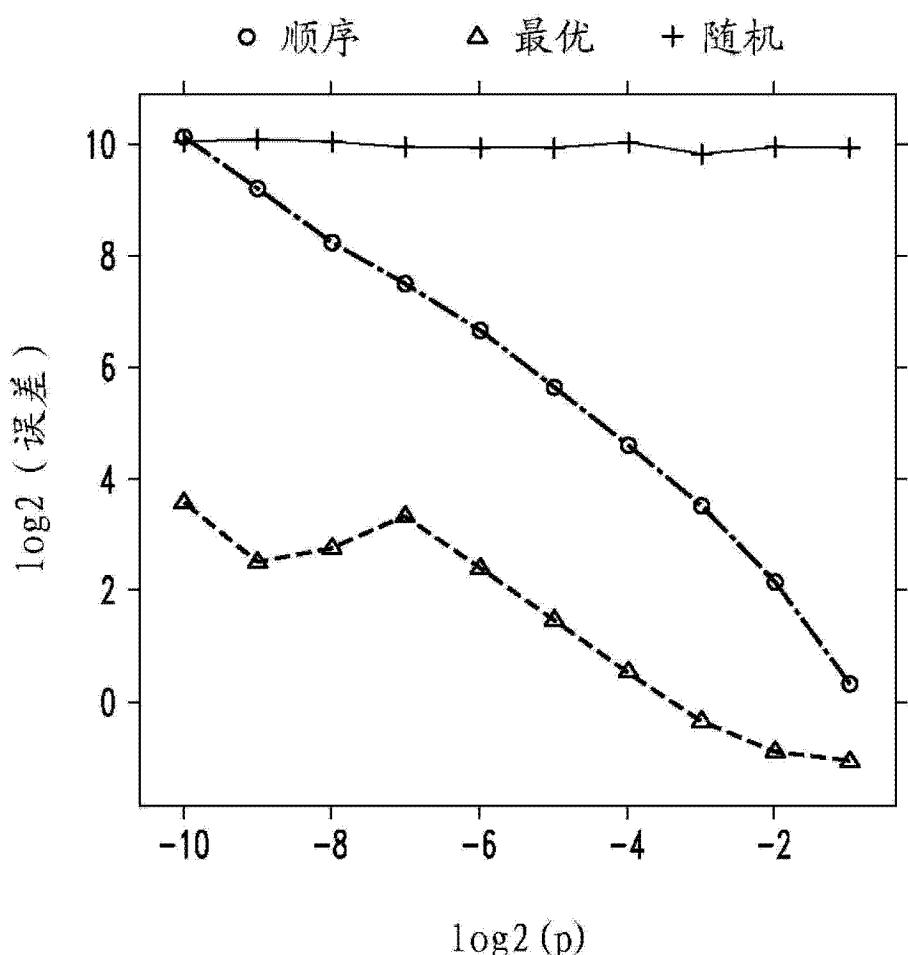


图 6

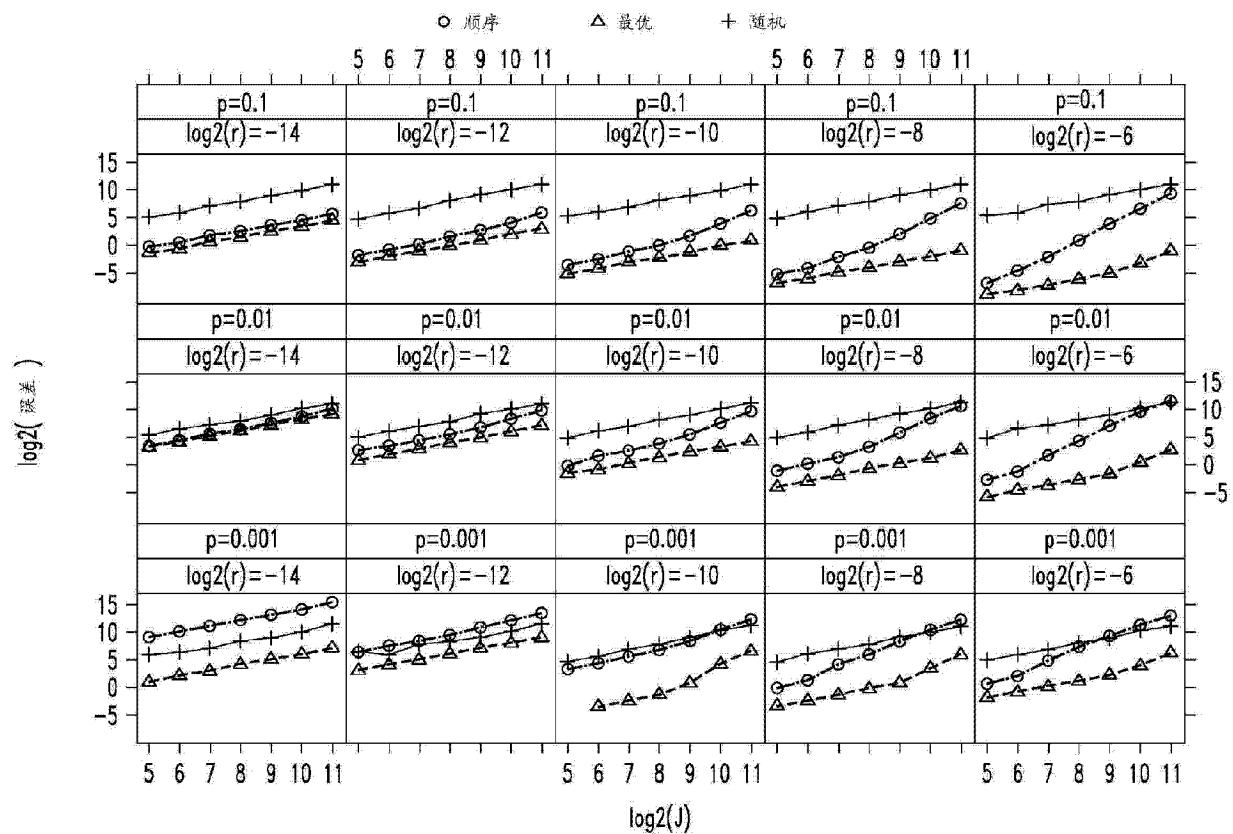


图 7