(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 112836117 A
(43)申请公布日 2021.05.25

(21)申请号 202011336985.7

(22)申请日 2020.11.25

(30)优先权数据
62/940,179 2019.11.25 US

(71)申请人 渊慧科技有限公司
地址 英国伦敦

(72)发明人 C.弗纳德 A.吉奥吉 T.A.曼恩

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 金玉洁

(51)Int.Cl.
G06F 16/9535(2019.01)
G06N 7/00(2006.01)
G06N 20/00(2019.01)

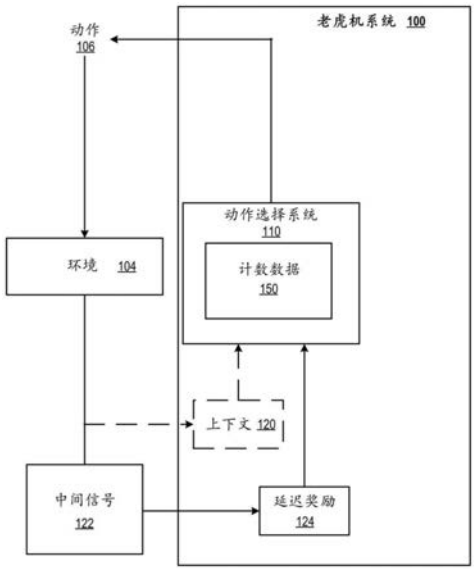
权利要求书2页 说明书9页 附图4页

(54)发明名称

具有中间信号的非固定延迟老虎机

(57)摘要

从要在环境中执行的动作集合中选择动作的方法、系统和装置，包括编码在计算机存储介质上的计算机程序。方法之一包括，在每个时间步：维护计数数据；对于每个动作，确定相应当前转移概率分布，该相应当前转移概率分布包括中间信号中的每一个的相应当前转移概率，该当前转移概率表示如果该动作被执行则中间信号将被观察到的当前可能性的估计；对于每个中间信号，确定相应的奖励估计，该相应的奖励估计是作为观察到中间信号的结果将接收到的奖励的估计；从相应当前转移概率分布和相应的奖励估计确定每个动作的相应动作得分；以及基于相应动作得分来选择要执行的动作。



CN 112836117 A

1.一种从要在环境中执行的动作集合中选择动作的方法,所述方法包括,在多个时间步的序列中的每个时间步:

维护计数数据,所述计数数据:

(i)对于每个动作以及对于中间信号的离散集合中的每个中间信号,指定响应于所述动作被执行而已经观察到所述中间信号的次数的计数,以及

(ii)对于每个中间信号,指定对于其响应于在所述时间步执行的动作所述中间信号已经被观察到的时间步的已经接收到的奖励的计数,

其中,所述离散集合中的每个中间信号描述动作已经被执行之后但是已经接收到执行的动作的奖励之前所述环境的对应状态,并且

其中,每个奖励是度量响应于其而观察到所述中间信号的动作的质量的数值;

从所述计数数据中针对每个动作确定包括所述中间信号中的每一个的相应当前转移概率的相应当前转移概率分布,所述相应当前转移概率表示如果所述动作被执行则所述中间信号将被观察到的当前可能性的估计;

从所述计数数据中针对每个中间信号确定相应的奖励估计,所述相应的奖励估计是作为所述中间信号被观察到的结果将接收到的奖励的估计;

从所述相应当前转移概率分布和所述相应的奖励估计确定每个动作的相应动作得分;以及

基于所述相应动作得分来选择要在所述环境中执行的动作。

2.根据权利要求1所述的方法,其中,所述环境是内容项推荐设定,其中,所述动作对应于内容项,并且其中,在所述内容项推荐设定中向用户推荐对应于所选动作的内容项。

3.根据权利要求1所述的方法,其中,选择动作包括:

选择具有最高动作得分的动作。

4.根据权利要求1所述的方法,还包括:

接收响应于所选动作被执行而观察到中间信号的指示;以及

作为响应,更新所述计数数据。

5.根据权利要求1所述的方法,还包括:

接收作为在更早的时间步观察到的先前中间信号的结果的奖励;以及

作为响应,更新所述计数数据。

6.根据权利要求1所述的方法,其中,对于每个动作以及对于中间信号的离散集合中的每个中间信号,响应于所述动作被执行而已经观察到所述中间信号的次数的计数为:

对在包括固定数量的最新近时间步的最近时间窗期间响应于所述动作被执行而已经观察到所述中间信号的次数进行计数的窗口化计数。

7.根据权利要求6所述的方法,其中,从所述计数数据中针对每个动作确定包括所述中间信号中的每一个的相应当前转移概率的相应当前转移概率分布包括:

基于以下确定所述中间信号中的每一个的所述相应当前转移概率:(i)对在所述最近时间窗期间响应于所述动作被执行而已经观察到所述中间信号的次数进行计数的窗口化计数与(ii)对在所述最近时间窗期间所述动作已经被执行的次数进行计数的窗口化计数的比率。

8.根据权利要求1所述的方法,其中,对于每个中间信号,作为所述中间信号被观察到

的结果而已经接收到的奖励的计数为：

对在更长时间窗期间响应于所述动作被执行而已经观察到所述中间信号的时间步接收的奖励进行计数的奖励计数，所述更长时间窗不包括所述最近时间窗中的一些或所有最新近时间步。

9.根据权利要求8所述的方法，其中，所述计数数据还指定：

对于每个中间信号，在所述更长时间窗期间已经观察到所述中间信号的次数的延迟计数，所述更长时间窗不包括所述最近时间窗中的一些或所有最新近时间步。

10.根据权利要求9所述的方法，其中，从所述计数数据中针对每个中间信号确定相应的奖励估计，所述相应的奖励估计是作为所述中间信号被观察到的结果将接收到的奖励的估计，包括：

基于以下确定所述相应的奖励估计：(i)所述中间信号的奖励计数与(ii)所述中间信号的延迟计数的比率。

11.根据权利要求1所述的方法，其中，从所述相应当前转移概率分布和所述相应的奖励估计确定每个动作的相应动作得分包括：

使用随机老虎机技术从所述相应当前转移概率分布和所述相应的奖励估计确定所述相应动作得分。

12.一种系统，包括一个或多个计算机以及存储指令的一个或多个存储设备，所述指令在由所述一个或多个计算机执行时，使所述一个或多个计算机执行任何前述权利要求的相应方法的操作。

13.一个或多个计算机可读存储介质，存储指令，所述指令在由一个或多个计算机执行时，使所述一个或多个计算机执行权利要求1-11中的任一项的相应方法的操作。

具有中间信号的非固定延迟老虎机

技术领域

[0001] 本说明书涉及多臂老虎机(multi-armed bandit)。

背景技术

[0002] 在多臂老虎机场景中,智能体(agent)从可能动作集合中迭代地选择要在环境中执行的动作。响应于每个动作,智能体接收度量所选动作的质量的奖励。智能体试图选择最大化响应于执行所选动作而接收的预期奖励的动作。

发明内容

[0003] 本说明书描述了实施为在一个或多个位置的一个或多个计算机上的计算机程序的系统,该系统使用非固定延迟老虎机方案来选择要执行的动作。

[0004] 可以实施本说明书中描述的主题的特定实施例,以便实现以下优点中的一个或多个。

[0005] 在线推荐系统在接收反馈时经常面临长延迟,尤其是在针对一些长期指标进行优化时。具体地,当度量由推荐系统选择的动作的质量的奖励仅在动作已经被选择之后的许多时间步可用时,延迟发生。

[0006] 虽然缓解学习中延迟的影响可以在固定环境中得到补偿,但当环境随时间而改变时,即,当响应于接收到任何给定动作而预期接收到的奖励的分布随时间而改变时,问题变得更具挑战性。

[0007] 事实上,如果改变的时间尺度与接收奖励的延迟相当,则许多现有技术不可能了解环境,因为一旦接收到奖励,可用的观察已经过时。

[0008] 本说明书中描述的技术通过利用在没有延迟或有相对于以其接收奖励的延迟来说小的延迟的情况下可用的中间信号解决了这些缺陷,并允许在具有延迟奖励的动态环境中进行有效的学习(从而进行有效的动作选择)。具体地,所描述的技术利用了这样一个事实,即给定那些信号,系统的长期行为是固定的或非常缓慢地改变。具体地,通过将动作选择问题分解成(i)估计响应于给定动作而接收任何给定中间信号的改变的概率和(ii)估计在接收到给定中间信号之后接收给定奖励的固定概率,即使在存在延迟奖励和非固定环境的情况下,系统也能够有效地选择动作。

[0009] 本说明书中描述的主题的一个或多个实施例的细节在附图和以下描述中阐述。根据描述、附图和权利要求,本主题的其他特征、方面和优点将变得显而易见。

附图说明

[0010] 图1A示出了示例老虎机系统。

[0011] 图1B示出了具有中间信号和延迟奖励的环境的示例。

[0012] 图2是用于在给定时间步选择动作的示例过程的流程图。

[0013] 图3是用于计算动作的动作得分的另一示例过程的流程图。

具体实施方式

[0014] 本说明书一般描述了重复地选择要在环境中执行的动作的系统。

[0015] 每个动作是从预定的动作集合中选择的，并且系统试图最大化响应于所选动作而接收的奖励来选择动作。

[0016] 通常，奖励是度量所选动作的质量的数值。在一些实施方式中，每个动作的奖励或为零或为一，而在其他实施方式中，每个奖励是从下限奖励值和上限奖励值之间的连续范围中抽取的值。

[0017] 更具体地，针对任何给定动作而接收的奖励相对于该动作被选择(并且在环境中执行)的时间在时间上延迟。例如，奖励可以度量一些长期目标，这些长期目标只能或通常仅在动作被执行之后的相当长时间量才被满足。

[0018] 然而，在动作被执行之后，可以从环境中观察到中间信号。

[0019] 中间信号是描述在动作被执行之后相对不久(例如，在同一时间步或紧接着的时间步)接收的环境状态、并且提供了对动作选择的奖励可能是什么的指示的数据。

[0020] 具体地，在动作被执行之后，环境假设中间状态，该中间状态可以由中间信号的离散集合中的一个来描述。在一段时间的延迟之后，接收到取决于中间信号是什么的奖励。

[0021] 在一些情况下，动作是内容项(例如，书籍、视频、广告、图像、搜索结果或其他多条内容)的推荐。

[0022] 在这些情况下，奖励值度量如由长期目标度量的推荐的质量，并且中间信号可以是与内容项的初始短期交互的指示。

[0023] 例如，当内容项是书籍时，奖励值可以基于用户的电子阅读器应用是否指示用户阅读了书籍多于阈值量。另一方面，中间信号可以指示用户是否下载了电子书。

[0024] 作为另一示例，当内容项是广告时，奖励值可以基于转换事件是否作为正在呈现广告的结果而发生。另一方面，中间信号可以指示是否发生了点击事件，即，用户是否点击或以其他方式选择了所呈现的广告。

[0025] 作为另一示例，当内容项是软件应用(例如，移动应用)时，奖励值可以基于在相当长时间量(例如，一周或一个月)之后用户多频繁地使用该软件应用的度量。另一方面，中间信号可以指示用户是否从应用商店下载了该软件应用。

[0026] 图1示出了示例老虎机系统100。老虎机系统100是实施为在一个或多个位置的一个或多个计算机上的计算机程序的系统的示例，其中实施了下面描述的系统、组件和技术。

[0027] 系统100重复地(即，在多个时间步中的每一个时间步)选择环境104中要(例如，由系统100或由另一系统)执行的动作106。例如，如上所述，动作可以是例如在网页上或在软件应用中要向环境中(即，在针对内容项推荐的设定中)的用户做出的内容项推荐。

[0028] 在一些情况下，系统100响应于接收到的上下文输入120(例如，表征当前时间步的特征向量或其他数据)来选择动作。

[0029] 在内容项推荐设定中，数据通常包括描述其中内容项将被推荐的情形的数据，例如，当前时间、将向其显示推荐的用户的用户设备的属性、已经被推荐给用户的先前内容项的属性和对那些先前内容项的用户响应、以及内容项将被放置在其中的设定的属性中的任何一种。

[0030] 每个所选动作106的执行通常使得系统100从环境104接收奖励124。

[0031] 通常,奖励124是表示所选动作106的质量的数值。

[0032] 在一些实施方式中,每个动作106的奖励124或为零或为一,即,指示动作是否成功,而在其他实施方式中,奖励124是从下限奖励值和上限奖励值之间的连续范围中抽取的值,即,将动作106的质量表示为来自连续范围的值,而不是二元值(binary value)。

[0033] 具体地,动作选择系统110试图最大化响应于所选动作而接收的奖励而选择动作。

[0034] 然而,环境104是提供具有显著延迟(即,在对应动作106已经被执行之后的多个时间步的延迟)的奖励124的环境。因此,奖励124被称为"延迟奖励"。

[0035] 相反,在动作106被执行之后,系统100从环境104接收(或"观察")中间信号122。中间信号124是(i)在动作106被执行之后但没有延迟或者有相对于在接收奖励之前发生的延迟来说小的延迟而(即,在动作106被执行的阈值数量的时间步内,例如,在同一时间步或紧接着的时间步)接收的,并且(ii)提供了对动作选择的奖励可能是什么的指示的数据。换句话说,响应于给定动作选择而接收的奖励124可以相对于动作选择在时间上延迟,但是取决于在做出动作选择之后没有延迟或者以相对小延迟接收的中间观察122。

[0036] 图1B示出了具有中间信号和延迟奖励的环境的示例。

[0037] 在图1B的示例中,在时间步t,执行动作A_t,并且随后观察到中间信号S的离散集合中的一个。

[0038] 具体地,在执行动作A_t之后,中间信号可以被认为是从取决于A_t的时变概率分布p_t采样的,该时变概率分布p_t向离散集合中的每个中间信号分配(assign)相应的转移概率。

[0039] 在图1B的示例中,观察到中间信号S_t。

[0040] 多个时间步之后,接收到针对动作A_t的奖励R_t。给定中间信号S_t,可能奖励上的概率分布B近似独立于动作A_t。换句话说,一旦观察到中间信号S_t,相同的概率就被分配给每个可能的奖励,而不管选择了什么动作使得中间信号S_t被观察到。

[0041] 此外,如上所述,环境是非固定的。具体地,针对任何给定动作的中间信号上的概率分布p_t可以随时间而改变,因为环境的某些方面(例如,用户对系统选择的动作如何反应)随时间而改变。

[0042] 然而,概率分布B是固定的,并且不随时间而改变。也就是说,一旦观察到中间信号S_t,尽管系统可能不知道实际的概率分布β,但它不会随时间而改变或者非常缓慢地改变。

[0043] 尽管在任何给定时间系统100不知道确切的概率分布p_t和B,系统100试图通过估计这些分布并使用估计选择动作来最大化预期奖励而选择动作。

[0044] 返回到图1A的描述,系统100选择动作来应对(account for)(i)中间信号122的非固定性质和(ii)延迟奖励124。

[0045] 具体地,动作选择引擎110维护计数数据150,并使用维护的计数数据150来选择最优化预期奖励(即,最优化响应于给定当前转移概率分布和固定奖励分布执行动作而要接收的预期延迟奖励124)的动作122。

[0046] 更具体地,动作选择引擎110在计数数据150中针对动作集合中的每个动作维护中间信号122中的每一个响应于动作被执行而已经被接收到的频率的计数。引擎110还在计数数据150中针对可能的中间信号122中的每一个维护在观察到中间信号之后已经接收到的奖励的计数。

[0047] 然后,动作选择引擎110使用计数数据150来估计中间信号的转移概率,并估计中

间信号的奖励分布，并使用这些估计来选择动作。

[0048] 选择动作将在下面参考图2和图3更详细地描述。

[0049] 图2是用于在当前时间步选择动作的示例过程200的流程图。为了方便起见，过程200将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如，适当编程的老虎机系统(例如，图1的老虎机系统100)可以执行过程200。

[0050] 具体地，系统可以在时间步的序列中的每个时间步执行过程200，以重复地选择要在环境中执行的动作。

[0051] 系统维护计数数据(步骤202)。

[0052] 如上所述，计数数据包括两个不同种类的计数：中间信号计数和奖励计数。

[0053] 具体地，如上所述，转移概率是非固定的。因此，对于每个动作，系统维护对中间信号中的每一个的相应窗口化计数。

[0054] 对于给定动作，对任何给定中间信号的窗口化计数是(i)响应于给定动作被执行和(ii)在当前时间步的最近时间窗内(即，在最新近的W个时间步内，其中W是固定常数)给定中间信号被观察到的次数的计数。

[0055] 通过维护仅跟踪“最近”动作选择的窗口化计数，系统可以应对转移概率的非固定性质，如将在下面更详细地描述的。

[0056] 如上所述，以一些延迟观察到奖励，且奖励的分布是(i)独立于给定中间信号的动作和(ii)固定的。

[0057] 因此，对于每个特定的中间信号和可能奖励集合中的每一个，系统维护在已经观察到该特定的中间信号之后已经接收到的奖励的相应计数，即满足如下条件的奖励的相应计数：作为动作被执行的结果接收奖励，该动作被执行也导致该特定的中间信号被观察到。换句话说，如果奖励作为动作选择的结果而被接收，该动作选择也导致该特定的中间信号被观察到，则奖励满足条件。

[0058] 因为奖励是固定的，所以不需要对该计数加窗，并且计数是在更长的时间窗上，该时间窗通常包括比用于中间信号的最近时间窗计数多得多的时间步。例如，更长时间窗可以包括直到满足以下条件的最新近时间步的所有更早的时间步：针对在该时间执行的动作已经接收到奖励。也就是说，因为奖励被延迟，所以在最新近时间窗中的至少一些时间步将没有数据可用，即，因为还没有接收到响应于针对在那些时间步选择的动作所观察到的中间信号的奖励。

[0059] 系统还针对每个中间信号维护延迟计数，该延迟计数是在更长时间窗中已经观察到中间信号的次数的计数。注意，如上所述，因为奖励被延迟，并且更长时间窗不包括最新近时间步，所以该延迟计数通常将小于在所有更早的时间步上已经观察到中间信号的总次数。

[0060] 在一些情况下，为了播种(seed)计数数据，系统可以在使用下面描述的技术选择动作之前执行每个动作阈值次数，例如，通过均匀随机不替代(replacement)地选择动作，直到每个动作被选择一次。

[0061] 系统针对每个动作从计数数据中确定中间信号上的当前转移概率分布的估计(步骤204)。当前转移概率分布的估计包括每个中间信号的相应当前转移概率估计。

[0062] 具体地，对于集合中的每个中间信号和每个动作，系统确定该动作的当前转移概

率的估计，该当前转移概率表示如果在给定时间步选择该动作则中间信号将被观察到的可能性。

[0063] 具体地，对于任何特定的动作，系统可以将特定中间信号的转移概率估计计算为：(i)给定该特定中间信号被观察到的情况下接收的奖励的计数，和(ii)在更长时间窗期间该特定中间信号被观察到的次数的总计数(即，给定特定动作被执行的情况下所有中间信号的窗口化计数之和)的比率。

[0064] 系统针对每个中间信号确定奖励估计，该奖励估计是如果观察到中间信号将接收到的奖励的估计(步骤206)。

[0065] 具体地，对于任何特定的中间信号，系统可以将特定中间信号的奖励估计计算为：(i)更长时间窗内特定中间信号的奖励计数，和(ii)特定中间信号的延迟计数的比率。

[0066] 系统从转移概率估计和奖励估计中确定每个智能体的动作得分(步骤208)。具体地，系统使用随机老虎机技术(stochastic bandit technique)将转移概率估计和奖励估计映射到每个智能体的相应动作得分，该动作得分估计响应于动作被执行而将接收的(延迟)奖励。虽然可以使用任何适当的随机技术，但是这种技术的具体示例在下面参考图3进行描述。

[0067] 系统基于动作得分选择动作中的一个(步骤210)。例如，系统可以选择具有最高动作得分的动作，或者可以根据一些探索策略选择动作。探索策略的示例是ε(epsilon)贪婪策略，其中以概率ε从集合中选择随机动作，并且以概率1减去ε选择具有最高动作得分的动作。

[0068] 系统接收响应于所选动作被执行而观察到的中间信号(步骤210)。如上所述，中间信号在没有显著延迟的情况下被观察到。

[0069] 系统更新计数数据(步骤212)。具体地，系统更新针对所选动作的窗口化计数，即，以从所有中间信号的窗口化计数中移除最近时间窗中最旧的时间步，并且仅向观察到的中间信号的窗口化计数添加一。

[0070] 系统接收奖励(步骤214)。因为奖励被延迟，所以接收到的奖励是响应于在更早的时间步采取的动作的，并且作为在该更早的时间步观察到的中间信号的结果。

[0071] 系统更新计数数据(步骤212)。具体地，对于在更早的时间步观察到的中间信号，系统更新该信号的奖励计数和延迟计数，而不需要更新其他中间信号的计数。

[0072] 图3是用于执行随机老虎机技术以生成特定动作的动作得分的示例过程300的流程图。为了方便起见，过程300将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如，适当编程的老虎机系统(例如，图1的老虎机系统100)可以执行过程300。

[0073] 系统可以对集合中的所有动作执行过程300，以生成所有动作的相应动作得分。

[0074] 系统针对每个中间信号计算信号的奖励估计的置信上限(步骤302)。

[0075] 具体地，系统可以通过向奖励估计增加基于已经发生的时间步数、可能的中间信号的总数和更长时间窗内中间信号的延迟计数的奖金(bonus)来计算乐观的奖励估计。

[0076] 作为特定示例，信号s的奖金可以满足：

[0077] $$\sqrt{\frac{2\log\left(\frac{2TS}{\delta}\right)}{N_t^D(s)}},$$

[0078] 其中,T是系统的固定时间范围(fixed time horizon),S是中间信号的总数,δ是固定常数,并且$N_t^D(s)$是中间信号s的延迟计数。

[0079] 然后系统可以将置信上限计算为以下中的最小值:(i)最大可能奖励,和(ii)乐观奖励估计。

[0080] 系统计算动作的容差参数(步骤304)。容差参数基于最近时间窗的大小W、动作的总数K、在最近时间窗期间已经执行的动作的总次数的窗口化计数$N_t^W(a)$以及已经发生的时间步的总数。

[0081] 作为特定示例,动作a的容差参数可以满足:

[0082] $\sqrt{\frac{2S\log\left(\frac{KWT}{\delta}\right)}{N_t^W(a)}}$。

[0083] 系统从动作的当前转移概率分布估计、中间信号的置信上限和动作的容差参数中计算动作的动作得分(步骤306)。

[0084] 具体地,系统将动作得分计算为在给定在当前估计的转移概率分布的容差参数内的任何转移概率分布的情况下的最大预期奖励。

[0085] 对于每个中间信号,任何转移概率分布的预期奖励的乐观估计是信号的转移概率和信号的置信上限的相应乘积之和。

[0086] 换句话说,动作得分满足:

[0087] $\max\{q^T U_t : \|\hat{p}_t(a) - q\|_1 \leq TP\}$

[0088] 其中,q是可能的转移概率分布的集合Δ_S中的转移概率分布,U_t是中间信号的置信上限的向量,$\hat{p}_t(a)$是当前转移概率分布,并且TP是容差参数。

[0089] 用于计算这种最大预期奖励的示例技术在Jaksch,T.、Ortner,R.和Auer,P.的Near-optimal regret bounds for reinforcement learning.Journal of Machine Learning Research,11(51):1563-1600,2010中描述。

[0090] 本说明书结合系统和计算机程序组件使用术语“被配置为”。对于被配置为执行特定操作或动作的一个或多个计算机的系统来说,意味着该系统已经在其上安装了在操作中使得该系统执行这些操作或动作的软件、固件、硬件或它们的组合。对于被配置为执行特定操作或动作的一个或多个计算机程序来说,意味着该一个或多个程序包括指令,当该指令被数据处理装置执行时,使得该装置执行这些操作或动作。

[0091] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路中、在有形体现的计算机软件或固件中、在包括本说明书中公开的结构及其结构等同物的计算机硬件中、或者在它们中的一个或多个的组合中实施。本说明书中描述的主题的实施例可以被实施为一个或多个计算机程序,即,编码在有形非暂时性存储介质上的计算机程序指令的一个或多个模块,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质可以是机器可读存储设备、机器可读存储基底、随机或串行存取存储器设备、或者是它们中的一个或多个的组合。替代地或附加地,程序指令可以被编码在人工生成的传播信号上,例如,机器生成的电、光或电磁信号,生成该信号是为了对信息进行编码,以便传输到合适的接收装

置用于由数据处理装置执行。

[0092] 术语“数据处理装置”是指数据处理硬件，并且包括用于处理数据的所有种类的装置、设备和机器，包括例如可编程处理器、计算机或多个处理器或计算机。装置也可以是或进一步包括专用逻辑电路，例如，FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外，装置可以可选地包括为计算机程序创建执行环境的代码，例如，构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组合的代码。

[0093] 计算机程序，也可以被称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码，可以用任何形式的编程语言编写，包括编译或解释语言，或者声明性或过程性语言；并且它可以以任何形式部署，包括作为独立程序或作为模块、组件、子例程或适合在计算环境中使用的其他单元。程序可以但不是必须对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分中(例如，存储在标记语言文档中的一个或多个脚本)、存储在专用于所讨论的程序的单个文件中或存储在多个协调文件中(例如，存储一个或多个模块、子程序或代码部分的文件)。计算机程序可以被部署为在一个计算机上或在位于一个站点处或跨多个站点分布并通过数据通信网络互连的多个计算机上执行。

[0094] 在本说明书中，术语“数据库”被广泛用于指代任何数据集：数据不需要以任何特定的方式被结构化，或者根本不需要被结构化，并且它可以被存储在一个或多个位置的存储设备上。因此，例如，索引数据库可以包括多个数据集，每个数据集可以被不同地组织和访问。

[0095] 类似地，在本说明书中，术语“引擎”被广泛用于指代基于软件的系统、子系统或被编程为执行一个或多个具体功能的过程。通常，引擎将被实施为安装在一个或多个位置的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下，一个或多个计算机将专用于特定的引擎；在其他情况下，可以在相同(多个)计算机上安装并运行多个引擎。

[0096] 本说明书中描述的过程和逻辑流程可以由一个或多个可编程计算机执行，该一个或多个可编程计算机执行一个或多个计算机程序，以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路(例如，FPGA或ASIC)来执行，或者由专用逻辑电路和一个或多个编程计算机的组合来执行。

[0097] 适于执行计算机程序的计算机可以基于通用或专用微处理器或两者，或者任何其他种类的中央处理单元。通常，中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的元件是用于执行或运行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。中央处理单元和存储器可以由专用逻辑电路来补充或并入专用逻辑电路。通常，计算机还将包括用于存储数据的一个或多个大容量存储设备(例如，磁盘、磁光盘或光盘)，或可操作地耦合以从用于存储数据的一个或多个大容量存储设备接收数据或向其传送数据或接收数据和传送数据两者。然而，计算机不需要具有这样的设备。此外，计算机可以嵌入到另一设备中，例如，移动电话、个人数字助理(Personal Digital Assistant，PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(Global Positioning System，GPS)接收器或便携式存储设备(例如，通用串行总线(Universal Serial Bus，USB)闪存驱动器)，仅举几例。

[0098] 适于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备，包括例如半导体存储器设备，例如，EPROM、EEPROM和闪存设备；磁

盘,例如,内部硬盘或可移动磁盘;磁光盘;以及CD ROM和DVD-ROM盘。

[0099] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在具有用于向用户显示信息的显示器设备(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)和用户可以通过其向计算机提供输入的键盘和定点设备(例如,鼠标或轨迹球)的计算机上实施。也可以使用其他种类的设备来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感官反馈,例如,视觉反馈、听觉反馈或触觉反馈;并且可以以任何形式接收来自用户的输入,包括声音、语音或触觉输入。此外,计算机可以通过向用户使用的设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从用户的设备上的web浏览器接收的请求将网页发送到web浏览器。此外,计算机可以通过向个人设备(例如,正在运行消息传递应用的智能手机)发送文本消息或其他形式的消息以及反过来接收来自用户的响应消息来与用户交互。

[0100] 用于实施机器学习模型的数据处理装置还可以包括例如用于处理机器学习训练或生产的公共和计算密集型部分(即,推理、工作负荷)的专用硬件加速器单元。

[0101] 机器学习模型可以使用机器学习框架(例如,TensorFlow框架、微软认知工具包(Microsoft Cognitive Toolkit)框架、Apache Singa框架或Apache MXNet框架)来实施和部署。

[0102] 本说明书中描述的主题的实施例可以在包括后端组件(例如,作为数据服务器)、或者包括中间件组件(例如,应用服务器)、或者包括前端组件(例如,具有用户通过其可以与本说明书中描述的主题的实施方式交互的图形用户界面、web浏览器或app的客户端计算机)、或者包括一个或多个这样的后端组件、中间件组件或前端组件的任意组合的计算系统中实施。系统的组件可以通过任何形式或介质的数字数据通信(例如,通信网络)来互连。通信网络的示例包括局域网(Local Area Network,LAN)和广域网(Wide Area Network,WAN),例如,互联网。

[0103] 计算系统可以包括客户端和服务器。客户端和服务器通常彼此远离,并且通常通过通信网络进行交互。客户机和服务器的关系是通过运行在相应的计算机上并且彼此具有客户端-服务器关系的计算机程序产生的。在一些实施例中,服务器例如为了向与作为客户端的设备交互的用户显示数据并从用户接收用户输入的目的,向用户设备发送例如HTML页面的数据。可以在服务器处从设备接收在用户设备处生成的数据,例如,用户交互的结果。

[0104] 虽然本说明书包含许多具体的实施方式细节,但是这些不应被解释为对任何发明的范围或对可能要求保护的范围的限制,而是对特定于特定发明的特定实施例的特征的描述。本说明书中在分离的实施例的上下文中描述的特定特征也可以在单个实施例中组合实施。相反,在单个实施例的上下文中描述的各种特征也可以在多个实施例中分开实施或者以任何合适的子组合实施。此外,尽管特征可以在上文中被描述为以特定组合起作用,并且甚至最初被如此要求保护,但是在一些情况下,来自所要求保护的组合的一个或多个特征可以从该组合中删除,并且所要求保护的组合可以指向子组合或子组合的变型。

[0105] 类似地,虽然在附图中以特定顺序描绘了操作并且在权利要求中以特定顺序记载了操作,但是这不应该被理解为需要以所示的特定顺序或连续顺序执行这些操作,或者需要执行所有示出的操作,来获得期望的结果。在特定情形下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应该被理解为在所有实施例中都

需要这种分离，并且应当理解，所描述的程序组件和系统通常可以一起集成在单个软件产品中或者封装到多个软件产品中。

[0106] 已经描述了主题的特定实施例。其他实施例在所附权利要求的范围内。例如，权利要求中记载的动作可以以不同的顺序执行，并且仍然获得期望的结果。作为一个示例，附图中描绘的过程不一定需要所示的特定顺序或连续顺序来获得期望的结果。在一些情况下，多任务和并行处理可能是有利的。

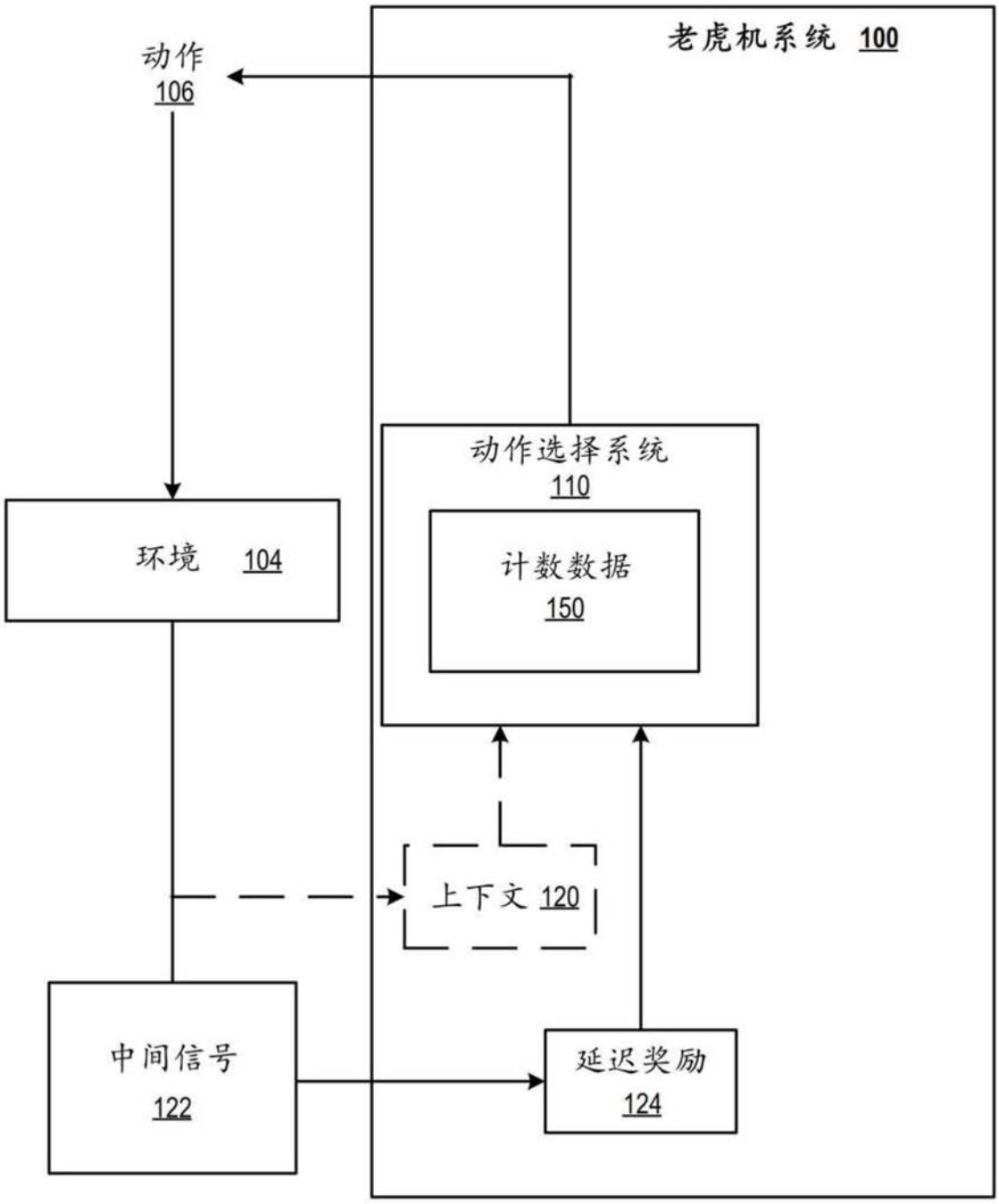

老虎机系统 100
动作
106
动作选择系统
110
计数数据
150
环境 104
上下文 120
中间信号
122
延迟奖励
124

图1A

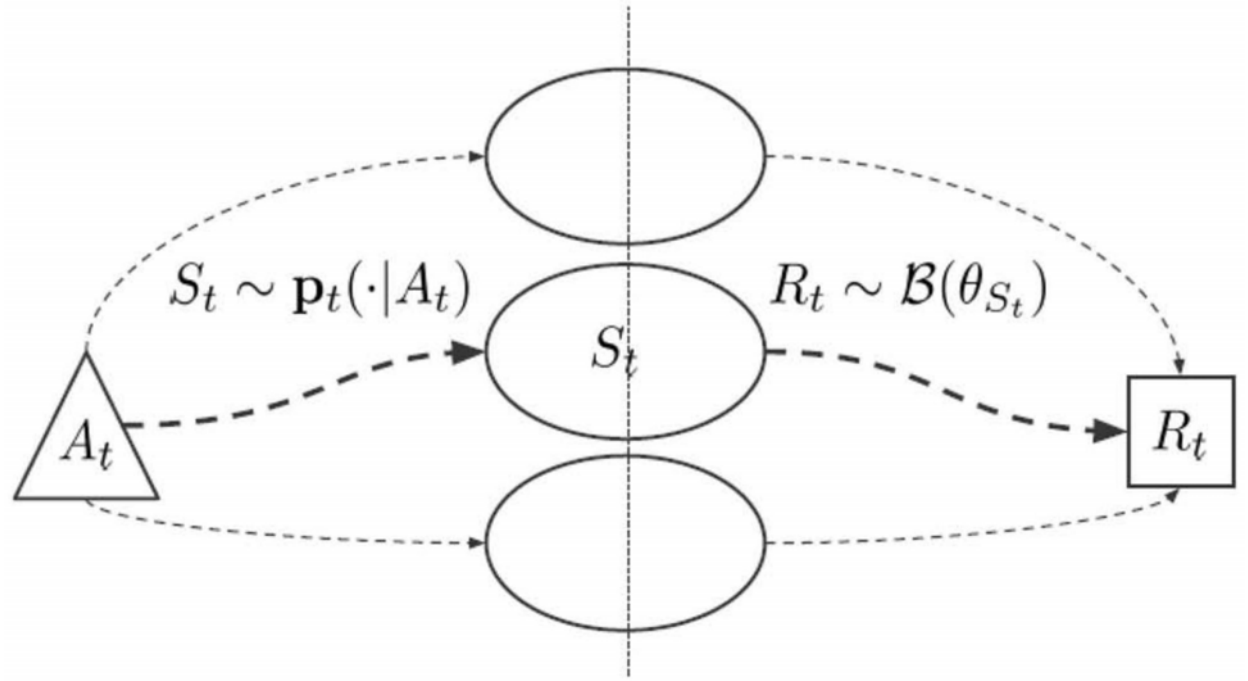


图1B

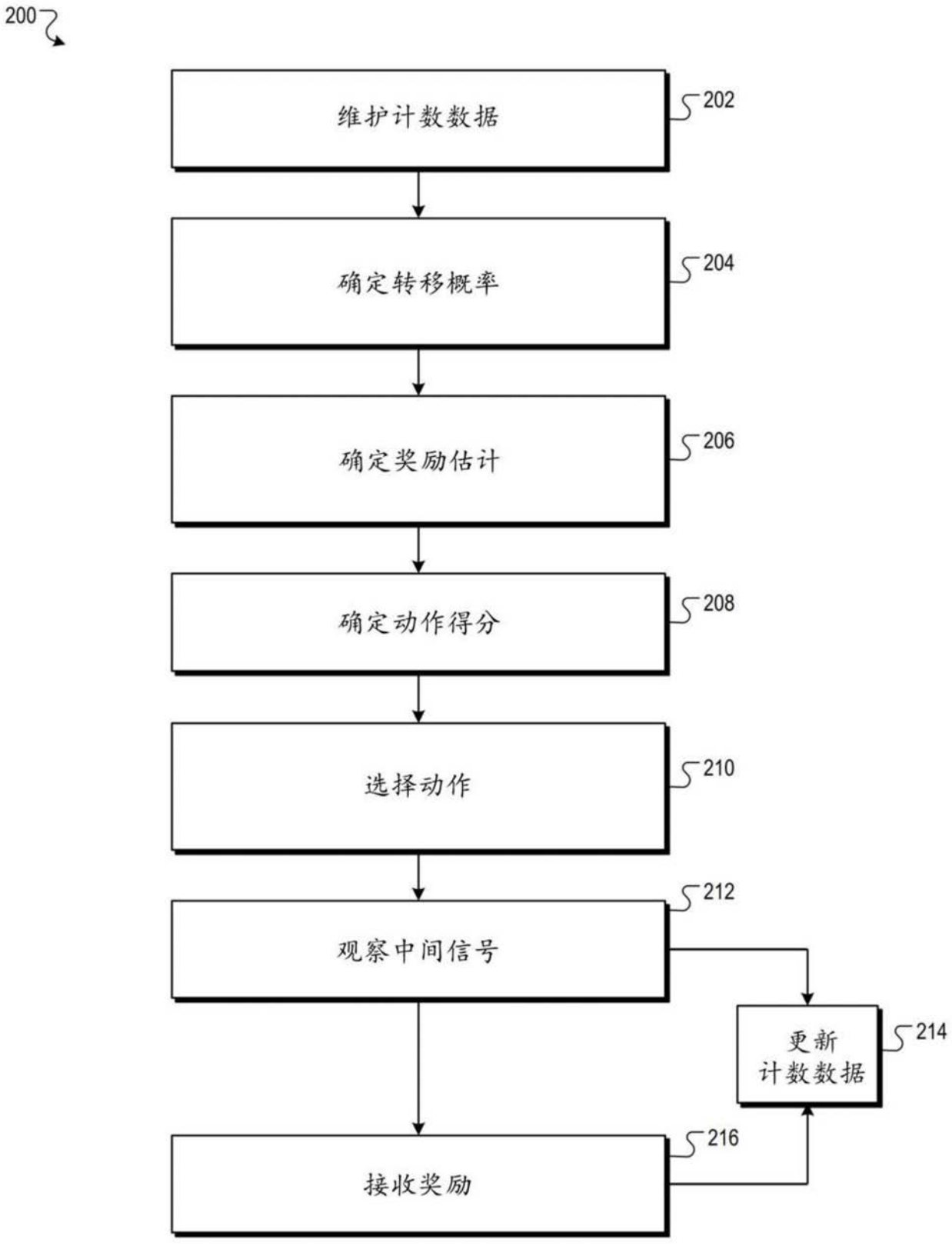

200
维护计数数据
202
确定转移概率
204
确定奖励估计
206
确定动作得分
208
选择动作
210
观察中间信号
212
更新
计数数据
214
接收奖励
216

图2

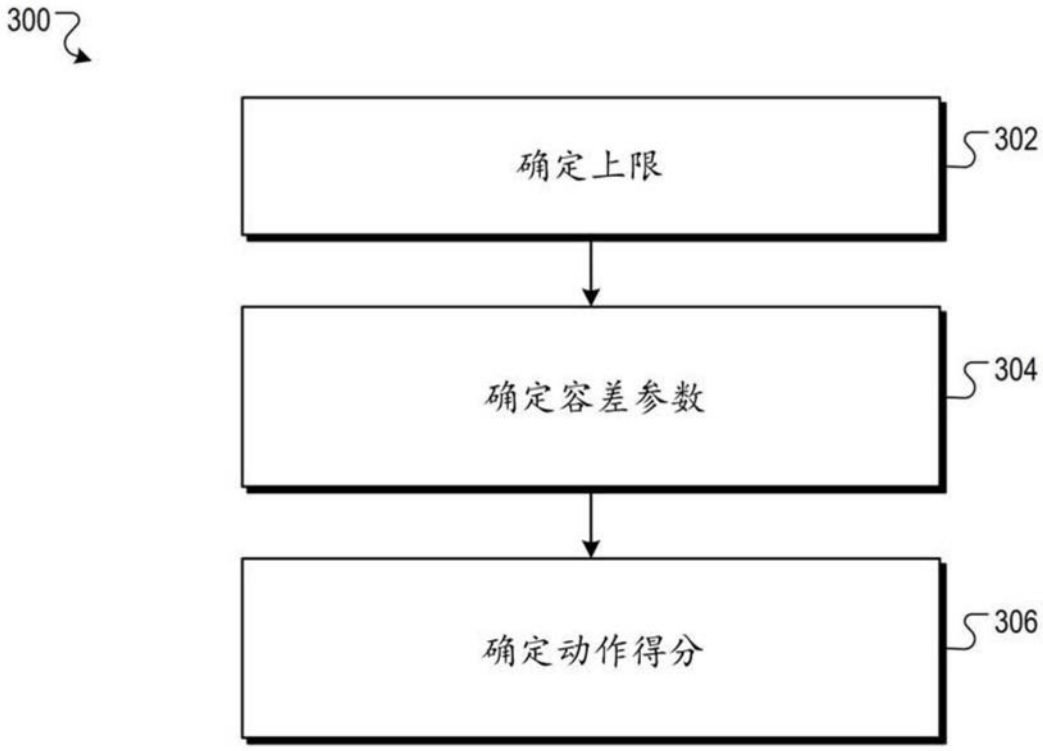


图3