(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0015284
(43) 공개일자 2023년01월31일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
(21) 출원번호 10-2022-0088713
(22) 출원일자 2022년07월19일
심사청구일자 없음
(30) 우선권주장
10 2021 207 868.0 2021년07월22일 독일(DE)

(71) 출원인
로베르트 보쉬 게엠베하
독일 데-70442 슈투트가르트 포스트파호 30 02 20
(72) 발명자
플린, 하미쉬
독일 70182 슈투트가르트 슈티첸부르크슈트라쎄 7
립, 데이비드
독일 71272 레닝엔 넬켄슈트라쎄 80
(뒷면에 계속)
(74) 대리인
양영준, 이민호, 백만기

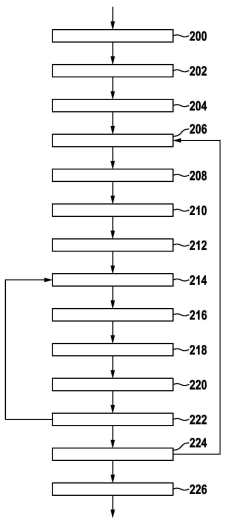
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **머신 학습을 위한 디바이스, 컴퓨터 프로그램 및 컴퓨터로 구현된 방법**

(57) 요 약

머신 학습을 위한 디바이스, 컴퓨터 프로그램 및 컴퓨터로 구현된 방법으로서, 이 방법은 멀티-암드 밴딧 문제 또는 컨텍추얼 밴딧 문제의 액션 공간 및 액션들에 따라 조건 지정되는 보상들에 대한 분포를 포함하는 작업을 제공하는 단계(200); 하이퍼프라이어를 제공하는 단계(202) - 하이퍼프라이어는 액션 공간에 대한 분포임-; 하이퍼프라이어에 따라, 하이퍼파스티리어로부터 샘플링된 프라이어들을 이용할 때, 미래 밴딧 작업들에 대한 예상 보상에 대한 하한이 가능한 한 큰 값을 갖는 하이퍼파스티리어를 결정하는 단계(224)를 포함하고, 하이퍼파스티리어는 액션 공간에 대한 분포이다.

대 표 도 - 도2



(72) 발명자

페터스, 얀

독일 64342 제하임-유겐하임(오티 말헨) 임 슈타인 빙거트 7

칸데미르, 멜리

덴마크 5240 오덴세 브예르가드슬로켄 22

명 세 서

청구범위

청구항 1

머신 학습을 위한 컴퓨터로 구현된 방법으로서,

상기 방법은,

멀티-암드 밴딧 문제 또는 컨텍추얼 밴딧 문제의 액션 공간 및 액션들에 따라 조건 지정되는 보상들에 대한 분포를 포함하는 작업을 제공하는 단계(200);

하이퍼프라이어를 제공하는 단계(202) - 상기 하이퍼프라이어는 상기 액션 공간에 대한 분포임-;

상기 하이퍼프라이어에 따라, 하이퍼파스티리어로부터 샘플링된 프라이어들을 이용할 때, 미래 밴딧 작업들에 대한 예상 보상에 대한 하한이 가능한 한 큰 값을 갖는 하이퍼파스티리어를 결정하는 단계(224) - 상기 하이퍼파스티리어는 상기 액션 공간에 대한 분포임-;

을 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 2

제1항에 있어서,

상기 하이퍼파스티리어를 결정하는 단계(224)가 상기 예상 보상에 대한 하한을 최대화하는 상기 하이퍼파스티리어를 결정하는 단계를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 3

제1항 또는 제2항에 있어서,

상기 방법은,

특히 센서 데이터를 분류하기 위해, 상기 하이퍼파스티리어로부터 샘플링된 프라이어에 따라, 센서 데이터, 특히 디지털 이미지 데이터 또는 오디오 데이터를 프로세싱하는 단계(226);

상기 센서 데이터에서 오브젝트들의 존재를 검출하거나 또는 상기 센서 데이터에 대해 시맨틱 세그멘테이션을 수행하는 단계; 또는

상기 하이퍼파스티리어로부터 프라이어들을 샘플링할 때, 상기 머신 학습의 강건성에 대한 척도, 특히 다음 작업에 대한 예상 에러가 미리 결정된 값을 초과하지 않을 확률을 결정하는 단계(226); 또는

상기 하이퍼파스티리어로부터 샘플링된 프라이어에 따라 센서 데이터에서 비정상을 검출하는 단계(226); 또는

물리 시스템을 제어하기 위한 정책을 학습하고 상기 하이퍼파스티리어로부터 샘플링된 프라이어에 따라 상기 물리 시스템을 제어하기 위한 제어 신호를 결정하는 단계

를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 방법은

다수의 반복들에서 상기 하이퍼파스티리어를 결정하는 단계(224); 및

이전 반복의 상기 하이퍼파스티리어로부터 상기 반복의 프라이어를 샘플링하는 단계(208)

를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 5

제4항에 있어서,

상기 방법은 작업들에 대한 분포로부터 상기 반복의 상기 작업을 샘플링하는 단계(206)를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 6

제1항 내지 제5항 중 어느 한 항에 있어서,

상기 방법은 상기 하이퍼파스티리어를 상기 하이퍼프라이어로 초기화하는 단계(204)를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서,

상기 방법은

파스티리어를 상기 프라이어로 초기화하는 단계(210);

상기 작업의 행동 정책들의 세트로부터 상기 작업 파스티리어와 연관된 행동 정책을 결정하는 단계(214) - 상기 행동 정책이 확률 질량에 의한 액션들에 대한 분포를 포함함-;

상기 확률 질량으로부터 액션을 무작위로 샘플링 또는 선택하는 단계(216);

상기 액션에 따라 상기 보상들에 대한 분포로부터 보상을 샘플링하는 단계(218);

상기 액션 및 상기 보상을 포함하는 데이터세트를 결정하는 단계(220);

상기 작업 데이터세트를 포함하도록 상기 파스티리어를 업데이트하는 단계(222)

를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 8

제7항에 있어서,

상기 작업을 제공하는 단계(200)는 상태 공간 및 초기 상태들에 대한 분포를 포함하는 상기 작업을 제공하는 단계를 포함하며, 상기 방법은 상기 초기 상태들에 대한 분포로부터 초기 상태를 무작위로 샘플링(218) 또는 선택하는 단계를 더 포함하며, 상기 보상들에 대한 분포는 상기 상태 공간의 상기 액션들 및 상태들에 따라 조건 지정되는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 9

제1항 내지 제8항 중 어느 한 항에 있어서,

상기 방법이 상기 데이터세트를 빈 세트로 초기화(212)한 후 상기 파스티리어를 미리 결정된 수의 라운드들에서 업데이트하는 단계를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 10

제1항 내지 제9항 중 어느 한 항에 있어서,

상기 하이퍼파스티리어를 결정하는 단계(224)는 상기 하이퍼파스티리어 및 상기 하이퍼프라이어의 Kullback-Leibler 다이버전스에 따라 상기 예상 보상을 결정 및 근사하는 단계를 포함하는 것을 특징으로 하는, 머신 학습을 위한 컴퓨터로 구현된 방법.

청구항 11

머신 학습을 위한 디바이스(100)로서,

상기 디바이스(100)는 제1항 내지 제10항 중 어느 한 항에 따른 방법의 단계들을 실행하도록 구성되는 것을 특

징으로 하는, 머신 학습을 위한 디바이스(100).

청구항 12

컴퓨터 프로그램으로서,

컴퓨터 상에서 실행될 때, 상기 컴퓨터로 하여금, 제1항 내지 제10항 중 어느 한 항에 따른 방법을 수행하게 하는 컴퓨터-판독가능 명령들을 포함하는 것을 특징으로 하는, 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 머신 학습을 위한 디바이스, 컴퓨터 프로그램 및 컴퓨터로 구현된 방법에 관한 것이다.

배 경 기 술

[0002] 문헌(A PAC-Bayesian bound for Lifelong Learning, Anastasia Pentina, Christoph H. Lampert, arXiv:1311.2838)은 머신 학습을 위한 평생 학습 설정의 양태들을 개시한다.

발명의 내용

[0003] 독립항들에 따른 컴퓨터로 구현된 방법, 디바이스 및 컴퓨터 프로그램은 향상된 머신 학습을, 특히, 평생 학습 시스템을 훈련하기 위한 목적 함수에 기초하여 제공한다.

[0004] 머신 학습을 위한 컴퓨터로 구현된 방법은 멀티-암드 밴딧(multi-armed bandit) 문제 또는 컨텍추얼 밴딧(contextual bandit) 문제의 액션 공간 및 액션들에 따라 조건 지정되는 보상들에 대한 분포를 포함하는 테스크를 제공하는 단계; 하이퍼프라이어(hyperprior)를 제공하는 단계 - 하이퍼프라이어는 액션 공간에 대한 분포임-; 하이퍼프라이어에 따라, 하이퍼파스티리어로부터 샘플링된 프라이어들(priors)을 이용할 때 미래 밴딧 작업들에 대한 예상 보상에 대한 하한이 가능한 한 큰 값을 갖는 하이퍼파스티리어(hyperposterior)를 결정하는 단계 - 하이퍼파스티리어는 액션 공간에 대한 분포임-;을 포함한다. 이는 작업들이 멀티-암드 밴딧 또는 컨텍추얼 밴딧 문제임을 의미한다. 하한은 PAC-베이지안 일반화 한계이며, 하한들은 하이퍼파스티리어로부터 샘플링된 프라이어들을 이용할 때 미래 밴딧 작업들에 대한 예상 보상이다. 하한은 평생 학습을 위한 목적 함수로서 사용되며, 작업을 관측한 후, 하이퍼파스티리어가 대략적으로 하한을 최대화하도록 업데이트된다. 하한은 관측가능한 양들을 포함하며 결과적으로 관측된 작업들로부터의 훈련 데이터를 이용하여 계산될 수 있다. 이렇게, 관측된 멀티-암드 밴딧 또는 컨텍추얼 밴딧 작업들의 세트로부터의 사전 지식이 새로운 멀티-암드 밴딧 또는 컨텍추얼 밴딧 작업들로 전달된다.

[0005] 하이퍼파스티리어를 결정하는 단계는 예상 보상에 대한 하한을 최대화하는 하이퍼파스티리어를 결정하는 단계를 포함할 수도 있다. 하한을 최대화하는 하이퍼파스티리어는 미래 작업들에서 낮은 에러를 초래하는 프라이어들에 높은 확률을 할당할 것으로 예상된다.

[0006] 본 방법은 관측된 멀티-암드 밴딧 또는 컨텍추얼 밴딧 작업들의 세트로부터의 사전 지식을 새로운 멀티-암드 밴딧 또는 컨텍추얼 밴딧 작업들에 전달하는데 이용될 수 있다. 많은 문제들은 멀티-암드 밴딧 문제들 또는 컨텍추얼 밴딧 문제들로 나타낼 수 있다. 따라서, 본 방법은 사전 지식을 다수의 유형들의 작업들로 전달하는데 이용될 수 있다.

[0007] 본 방법은 특히 센서 데이터를 분류하기 위해, 하이퍼파스티리어로부터 샘플링된 프라이어에 따라, 센서 데이터, 특히 디지털 이미지 데이터 또는 오디오 데이터를 프로세싱하는 단계; 센서 데이터에서 오브젝트들의 존재를 검출하거나 또는 센서 데이터에 대한 시맨틱 세그멘테이션(semantic segmentation)을 수행하는 단계; 또는 하이퍼파스티리어로부터 프라이어들을 샘플링할 때, 머신 학습의 강건성(robustness)에 대한 척도, 특히 다음 작업에 대한 예상 에러가 미리 결정된 값을 초과하지 않을 확률을 결정하는 단계; 또는 하이퍼파스티리어로부터 샘플링된 프라이어에 따라 센서 데이터에서 비정상을 검출하는 단계; 또는 물리 시스템을 제어하기 위한 정책을 학습하고 하이퍼파스티리어로부터 샘플링된 프라이어에 따라 물리 시스템을 제어하기 위한 제어 신호를 결정하는 단계를 포함할 수도 있다. 따라서, 본 방법은 새로운 문제들, 특히 새로운 분류 문제들, 새로운 강건성 문제들, 새로운 비정상 검출 문제들, 및/또는 새로운 제어 문제들에 사전 지식을 전달하는데 이용된다.

[0008] 하한은 또한 머신 학습의 강건성에 대한 척도이다. 하이퍼파스티리어는 또한 프라이어들이 하이퍼파스티리어로

부터 샘플링될 때 다음 작업에 대한 예상 에러에 관한 정보를 제공한다. 예를 들어, 예상 에러는 프라이어들이 특정의 하이퍼파스티리어로부터 샘플링될 때 특정의 값을 초과하지 않는다.

[0009] 본 방법은 다수의 반복들(iterations)에서 하이퍼파스티리어를 결정하는 단계; 및 이전 반복의 하이퍼파스티리어로부터 반복의 프라이어를 샘플링하는 단계를 포함할 수도 있다. 따라서, 머신 학습의 반복으로, 사전 지식이 이전에 반복들로부터 전달된다.

[0010] 본 방법은 작업들에 대한 분포로부터 반복의 작업을 샘플링하는 단계를 포함할 수도 있다. 특히 많은 수의 상이한 작업들을 프로세싱하면 머신 학습의 결과를 향상시킨다.

[0011] 본 방법은 바람직하게는 하이퍼프라이어로 하이퍼파스티리어를 초기화하는 단계를 포함한다. 모든 가용 하이퍼프라이어는 머신 학습을 위한 우수한 시작점이다.

[0012] 본 방법은 파스티리어(posterior)를 프라이어로 초기화하는 단계; 작업의 행동 정책들의 세트로부터 작업 파스티리어와 연관된 행동 정책을 결정하는 단계 - 행동 정책은 확률 질량에 의한 액션들에 대한 분포를 포함함-; 확률 질량으로부터 액션을 무작위로 샘플링 또는 선택하는 단계; 액션에 따라 보상들에 대한 분포로부터 보상을 샘플링하는 단계; 액션 및 보상을 포함하는 작업 데이터세트를 결정하는 단계; 작업 데이터세트를 포함하도록 작업 파스티리어를 업데이트하는 단계를 포함할 수도 있다. 이는 생성된 파스티리어가 이전 반복의 지식으로 작업 데이터세트를 보존하는 멀티-암드 밴딧 또는 컨텍추얼 밴딧 작업들에 대한 매우 효율적인 머신 학습을 제공한다.

[0013] 작업을 제공하는 것은, 바람직하게는 상태 공간 및 초기 상태들에 대한 분포를 포함하는 작업을 제공하는 것을 포함하며, 본 방법은 초기 상태들에 대한 분포로부터 초기 상태를 무작위로 샘플링 또는 선택하는 단계를 더 포함하며, 보상들에 대한 분포는 상태 공간의 액션들 및 상태들에 따라 조건 지정된다. 이는 컨텍추얼 밴딧 작업들에 대한 매우 효율적인 머신 학습을 제공한다.

[0014] 본 방법은 빈 세트로 작업 데이터세트를 초기화한 후 작업 파스티리어를 미리 결정된 수의 라운드들에서 업데이트하는 단계를 포함할 수도 있다. 따라서, 하한이 관측된 작업들로부터의 훈련 데이터로부터 계산된다.

[0015] 하이퍼파스티리어를 결정하는 것은 하이퍼파스티리어 및 하이퍼프라이어의 Kullback-Leibler 다이버전스에 따라 하한을 결정하는 것을 포함할 수도 있다. 프라이어가 새로운 작업에서와 같이 관측된 작업에 대해 효과적이도록 최적화될 때, 목표, 즉 향상된 머신 학습이 제공된다. 하이퍼파스티리어 및 하이퍼프라이어의 Kullback-Leibler 다이버전스는 관측된 작업들로부터의 훈련 데이터로부터 하한을 결정하는 우수한 방법을 제공한다.

[0016] 머신 학습을 위한 디바이스이며, 이 디바이스는 제1항 내지 제9항 중 하나에 따른 방법의 단계들을 실행하도록 구성된 것을 특징으로 한다.

[0017] 컴퓨터 프로그램이며, 이 컴퓨터 프로그램은 컴퓨터 상에서 실행될 때, 컴퓨터로 하여금, 제1항 내지 제9항 중 하나에 따른 방법을 수행하게 하는 컴퓨터-판독가능 명령들을 포함하는 것을 특징으로 한다.

도면의 간단한 설명

[0018] 추가적인 유리한 실시형태들은 다음 설명 및 도면으로부터 도출될 수 있다. 도면에서:

도 1은 머신 학습을 위한 디바이스의 부분을 개략적으로 도시한다.

도 2는 머신 학습을 위한 방법의 단계들을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0019] 도 1은 머신 학습을 위한 디바이스(100)의 부분을 개략적으로 도시한다. 디바이스(100)는 적어도 하나의 프로세서(102) 및 적어도 하나의 스토리지(104)를 포함한다. 적어도 하나의 스토리지(104)는 컴퓨터 상에서 실행될 때, 컴퓨터로 하여금, 도 2를 참조하여 아래에서 설명되는 방법을 수행하게 하는 컴퓨터-판독가능 명령들을 포함하는 컴퓨터 프로그램을 포함할 수도 있다. 디바이스(100)는 방법의 단계들을, 특히 적어도 하나의 프로세서(102)가 컴퓨터 프로그램의 명령들을 실행할 때, 실행하도록 구성된다.

[0020] 본 예에서 머신 학습의 방법은 기본 학습 알고리즘 Q=Q(D,P)를 이용하며, 단계 200에서 시작한다. 기본 학습 알고리즘은 파스티리어 분포 Q를 반환한다. 기본 학습 알고리즘이 데이터세트 D 및 프라이어 P를 이용하므로, 파스티리어는 D 및 P에 대한 의존성을 명확하게 하기 위해 설명에서는 Q(D, P)로 기재된다.

[0021] 단계 200에서, 본 방법은 작업 Ti를 제공하는 단계를 포함한다.

[0022] 작업 Ti는 멀티-암드 밴딧 문제의 액션 공간 A, 즉 액션들 a의 세트, 및 액션들 a에 따라 조건 지정되는 보상들 r에 대한 분포 $\rho_i(r|a)$를 포함하는, 커플

[0023] $$T_i = (A, \rho_i)$$

[0024] 로서 표현되는 멀티-암드 밴딧 문제의 작업일 수도 있다.

[0025] 작업 Ti는 액션 공간 A, 즉 액션들 a의 세트, 및 상태 공간 S, 및 액션들 a 및 상태들 s에 따라 조건 지정되는 보상들 r에 대한 분포 $\rho_i(r|s,a)$, 및 컨텍추얼 밴딧 문제의 초기 상태들 $\mu(s)$에 대한 분포를 포함하는, 쿼드러플

[0026] $$T_i = (S, A, \mu_i, \rho_i)$$

[0027] 로서 표현되는 컨텍추얼 밴딧 문제의 작업일 수도 있다.

[0028] 보상들 r은 0과 1 사이로 가정되며 작업들 T_i는 환경 $\mathbf{T}$로부터 샘플링되는 것으로 가정된다.

[0029] 이후, 단계 202가 실행된다.

[0030] 단계 202에서, 본 방법은 하이퍼프라이어 $\mathcal{P}$를 제공하는 단계를 포함한다.

[0031] 하이퍼프라이어 P는 가능한 프라이어 분포들의 세트에 대한 분포이다. 각각의 프라이어 분포는 액션 공간 A에 대한 분포이다.

[0032] 이후, 단계 204가 실행된다.

[0033] 단계 204에서, 본 방법은 하이퍼파스티리어 $\mathcal{Q}$를 하이퍼프라이어 $\mathcal{P}$로 초기화하는 단계를 포함한다. 하이퍼파스티리어 $\mathcal{Q}$는 프라이어 분포들에 대한 다른 분포이다.

[0034] 이후, 단계 206이 실행된다.

[0035] 단계 206에서, 본 방법은 작업들 T에 대한 분포 $\mathbf{T}$로부터 반복 i의 작업 Ti를 샘플링하는 단계를 포함한다.

[0036] 작업 T_i와 연관된 것은 m_i개의 행동 정책들을 포함하는 작업 Ti의 행동 정책들 $B_{ij} = (b_{i1}, \ldots, b_{im_i})$의 세트이다.

[0037] 행동 정책 b_{ij}는 멀티-암드 밴딧 문제에 대해 확률 질량 $b_{ij}(a)$에 의한 액션들 a에 의한 분포를 포함한다. 본 예에서, 행동 정책 b_{ij}는 이전에 관측된 훈련 데이터

[0038] $$D_1, \ldots, D_{1-i}, D_i^{ij-1}$$

[0039] - 여기서, 멀티-암드 밴딧 작업의 경우

[0040] $D_i^{ij-1} = \left((a_{i1}, r_{i1}), \ldots, (a_{ij-1}, r_{ij-1})\right)$ 임 -

[0041] 에 대한 반복 i에 의존하는 액션들 a에 대한 분포이다.

[0042] 훈련 데이터세트의 엘리먼트들은 다음과 같이 분산된 멀티-암드 밴딧 작업에 있다:

[0043] $$a_{ij} \sim b_{ij}(a|D_1, \ldots, D_{1-i}, D_i^{ij-1}), r_{ij} \sim \rho_i(r|a_{ij})$$

[0044] 행동 정책 b_{ij}는 컨텍추얼 밴딧 문제에 대해 확률 질량 $b_{ij}(a|s_{ij})$에 의한 상태들 s에 따라 조건 지정된 액션들 a에 대한 분포를 포함한다. 본 예에서, 행동 정책 b_{ij}는 이전에 관측된 훈련 데이터

[0045] $$D_i^{ij-1} = \left((s_{i1}, a_{i1}, r_{i1}), \ldots, (s, a_{ij-1}, r_{ij-1})\right)$$

[0046] 에 대해 반복 i에 의존하는 액션들 a에 대한 분포이다.

[0047] 훈련 데이터세트의 엘리먼트들은 다음과 같이 분산된 컨텍추얼 밴딧 작업에 있다

[0048] $$a_{ij} \sim b_{ij}(a|s_{ij}, D_1, \ldots, D_{1-i}, D_i^{ij-1}), r_{ij} \sim \rho_i(r|s_{ij}, a_{ij})$$

[0049] 이후, 단계 208이 실행된다.

[0050] 단계 208에서, 본 방법은 이전 반복 i-1의 하이퍼파스티리어 Q로부터 반복 i의 프라이어 P를 샘플링하는 단계를 포함한다.

[0051] 이후, 단계 210이 실행된다.

[0052] 단계 210에서, 본 방법은 파스티리어 Q를 프라이어 P로 초기화하는 단계를 포함한다.

[0053] 이후, 단계 212가 실행된다.

[0054] 단계 212에서, 본 방법은 데이터세트 D_i를 빈 세트로 초기화하는 단계를 포함한다.

[0055] 이후, 단계 214가 실행된다.

[0056] 단계 214에서, 본 방법은 작업 T_i의 행동 정책들 B_{ij}의 세트로부터, 작업 파스티리어 Q와 연관되는 행동 정책 b_{ij}를 결정하는 단계를 포함한다.

[0057] 이후, 단계 216이 실행된다.

[0058] 단계 216에서, 본 방법은 행동 정책 b_{ij}의 확률 질량으로부터 액션 a_{ij}를 무작위로 샘플링 또는 선택하는 단계를 포함한다.

[0059] 이후, 단계 218이 실행된다.

[0060] 단계 218에서, 본 방법은 멀티-암드 밴딧 문제에 대해 보상들 r에 대한 분포 $\rho_i(r|a)$로부터의 액션 a_{ij}에 따라 보상 r_{ij}를 샘플링하는 단계를 포함한다.

[0061] 단계 218에서, 본 방법은 컨텍추얼 밴딧 문제에 대해 초기 상태들 $\mu(s)$에 대한 분포로부터 초기 상태 s_{ij}를 무작위로 샘플링 또는 선택하는 단계 및 상태 공간 S의 액션들 a 및 상태들 s에 따라 조건 지정된 보상들 r에 대한 분포 $\rho_i(r|s,a))$로부터의 액션 a_{ij}에 따라, 보상 r_{ij}를 샘플링하는 단계를 포함한다.

[0062] 이후, 단계 220이 실행된다.

[0063] 단계 220에서, 본 방법은 액션 aij 및 보상 rij를 포함하는 데이터세트 D_i를 결정하는 단계를 포함한다.

[0064] 데이터세트 D_i는 멀티-암드 밴딧 문제에 대해 m_i개의 관측가능한 훈련 데이터 쌍들을 포함하는 훈련 세트이다:

[0065] $$D_i = \left((a_{i1}, r_{i1}), \ldots, (a_{im_i}, r_{im_i})\right)$$

[0066] 데이터세트 D_i는 컨텍추얼 밴딧 문제에 대해 m_i개의 관측가능한 훈련 데이터 트리플을 포함하는 훈련 세트이

다:

[0067] $$D_i = \left((s_{i1}, a_{i1}, r_{i1}), \ldots, (s, a_{ij-1}, r_{ij-1})\right)$$

[0068] 이후, 단계 222가 실행된다.

[0069] 단계 222에서, 본 방법은 데이터세트 D_i를 포함하도록 파스티리어 Q를 업데이트하는 단계를 포함한다.

[0070] 이를 위해서, 결정론적인 기본 학습 알고리즘은 데이터세트 D_i 및 프라이어 P를 입력들로 취하여 파스티리어 $Q = Q(D_i; P)$를 생성한다.

[0071] 작업 T_i의 목표는 예상 보상을 최대화하는 파스티리어 Q를 찾는 것이다. 멀티-암드 밴딧 문제에 대한 예상 보상은 다음과 같다:

[0072] $$L_i(a) = \underset{r\sim\rho_i(r|a)}{E}[r]$$

[0073] 여기서, E는 기대 값이다.

[0074] 데이터세트 D_i를 이용하여, 예상 보상이 멀티-암드 밴딧 문제의 예에서 다음과 같이 추정된다:

[0075] $$\hat{L}_i(a, D_i) = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{b_{ij}(a_{ij})} r_{ij} I\{a_{ij} = a\}$$

[0076] $I\{a_{ij} = a\}$는 $a_{ij} = a$이면 1을 반환하고 그렇지 않으면 0을 반환하는 표시자 함수(indicator function)이다. 컨텍추얼 밴딧 문제에 대한 예상 보상은 다음과 같다:

[0077] $$L_i^{(cb)}(\pi) = \underset{s\sim\mu_i}{E}\underset{a\sim\pi(a|s)}{E}\underset{r\sim\rho_i(r|s,a)}{E}[r]$$

[0078] 여기서, E는 개별 기대 값이고, 프라이어 P 및 파스티리어 Q는 정책들의 공간 Π에 대한 분포들이고, $\pi \in \Pi$는 상태들 s에 따라 조건 지정된 액션들 a에 대한 분포 $\pi(a|s)$ 이다.

[0079] 데이터세트 D_i를 이용하여, 예상 보상이 컨텍추얼 밴딧 문제의 예에서 다음과 같이 추정된다:

[0080] $$\hat{L}_i^{(cb)}(\pi, D_i) = \frac{1}{m_i}\sum_{j=1}^{m_i}\frac{\pi(a_{ij}|s_{ij})}{b_{ij}(a_{ij}|s_{ij})} r_{ij}$$

[0081] 본 예에서, 본 방법은 파스티리어 Q를 미리 결정된 수의 라운드들 m에서 업데이트하는 단계를 포함한다. 이는 본 방법이 m번의 라운드들 동안 단계 214로 속행한다는 것을 의미한다.

[0082] 이후, 단계 224가 실행된다.

[0083] 단계 224에서, 본 방법은 하이퍼프라이어 $\mathcal{P}$에 따라, 하이퍼파스티리어 $\mathcal{Q}$로부터 샘플링된 프라이어들 P를 이용할 때, 미래 밴딧 작업들에 대한 예상 보상 L에 대한 하한이 가능한 한 큰 값을 갖는 하이퍼파스티리어 $\mathcal{Q}$를 결정하는 단계를 포함한다.

[0084] 본 예에서 하이퍼파스티리어 $\mathcal{Q}$를 결정하는 단계는 예상 보상 L에 대한 하한(lower bound)을 최대화하는 하이퍼파스티리어 $\mathcal{Q}$를 결정하는 단계를 포함한다.

[0085] $\mathcal{Q} \leftarrow \text{argmax}_{\mathcal{Q}'} \text{lower bound}(\mathrm{L}(\mathcal{Q}'))$

[0086] 본 예에서, 멀티-암드 밴딧 문제에 대한 함수 lower bound$^{(\mathrm{L}(\mathcal{Q}'))}$은 다음 목적 함수들 중 하나이다:

$$LIKL_a(\mathcal{Q}) = -\frac{1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) + \frac{1}{\kappa_{max}} kl_{-}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i \hat{L}_i(Q(D_i,P_i),D_i) \,\middle\|\, \frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \sum_{i=1}^{n} D_{\mathrm{KL}}(Q(D_i,P_i)||P_i) + n\,\ln(\xi(m)) - \ln(\delta/2)}{nm}\right)$$

$$LIKL_b(\mathcal{Q}) = -\frac{1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) + \frac{1}{\kappa_{max}} kl_{-}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i \hat{L}_i(Q(D_i,P_i),D_i) \,\middle\|\, \frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \sum_{i=1}^{n} D_{\mathrm{KL}}(Q(D_i,P_i)||P_i) + n\,\ln(\xi(m)) - \ln(\delta/2)}{nm}\right)$$

$$LP_a(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\kappa_i}{\kappa_{max}}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) - \frac{\sqrt{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \sum_{i=1}^{n} D_{\mathrm{KL}}(Q(D_i,P_i)||P_i) + n\ln(\xi(m)) - \ln(\delta/2)}}{\kappa_{max}\sqrt{2nm}}$$

$$LP_b(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\kappa_i}{\kappa_{max}}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) - \frac{\sqrt{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \sum_{i=1}^{n} D_{\mathrm{KL}}(Q(D_i,P_i)||P_i) + n\ln(\xi(m)) - \ln(\delta/2)}}{\kappa_{max}\sqrt{2nm}}$$

[0087]

$$LHA_a(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) - \frac{\beta}{n}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)}{\sqrt{-\ln(\delta/2)}} + 2\sqrt{-\ln\left(\frac{\delta}{2}\right)}\right)$$

$$LHA_b(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) - \frac{\beta}{n}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)}{\sqrt{-\ln(\delta/2)}} + 2\sqrt{-\ln\left(\frac{\delta}{2}\right)}\right)$$

$$CLIKL_a(\mathcal{Q}) = -\frac{\beta_1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) + \frac{1}{\kappa_{max}}kl^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i\hat{L}_i(Q(D_i,P_i),D_i)\,\middle\|\,\beta_2\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)+n\ln(\xi(m))-\ln(\delta/2)}{nm}\right)$$

$$CLIKL_b(\mathcal{Q}) = -\frac{\beta_1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) + \frac{1}{\kappa_{max}}kl^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i\hat{L}_i(Q(D_i,P_i),D_i)\,\middle\|\,\beta_2\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)+n\ln(\xi(m))-\ln(\delta/2)}{nm}\right)$$

$$CLP_a(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\kappa_i}{\kappa_{max}}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{\beta_1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) - \frac{\beta_2\sqrt{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)+n\ln(\xi(m))-\ln(\delta/2)}}{\kappa_{max}\sqrt{2nm}}$$

$$CLP_b(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\kappa_i}{\kappa_{max}}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{\beta_1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) - \frac{\beta_2\sqrt{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)+n\ln(\xi(m))-\ln(\delta/2)}}{\kappa_{max}\sqrt{2nm}}$$

$$CLHA_a(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{\beta_1}{\sqrt{n}}\left(D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P}) + \frac{1}{8} - \ln\left(\frac{\delta}{2}\right)\right) - \frac{\beta_2\beta}{n}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)}{\sqrt{-\ln(\delta/2)}} + 2\sqrt{-\ln\left(\frac{\delta}{2}\right)}\right)$$

$$CLHA_b(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(Q_i(D_i,P_i),D_i) - \frac{\beta_1}{\sqrt{8n}}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})}{\sqrt{-\ln(\delta/2)}} + \sqrt{8} + 2\sqrt{-\ln(\delta/2)}\right) - \frac{\beta_2\beta}{n}\left(\frac{D_{\mathrm{KL}}(\mathcal{Q}||\mathcal{P})+\sum_{i=1}^{n}D_{\mathrm{KL}}(Q(D_i,P_i)||P_i)}{\sqrt{-\ln(\delta/2)}} + 2\sqrt{-\ln\left(\frac{\delta}{2}\right)}\right)$$

[0088]

[0089] 여기서, β_1 및 β_1은 상수들 $0 \le \beta_1, \beta_2 \le 1$이며,

[0090] 여기서,

$$\varepsilon(m) = \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^{k}\left(1-\frac{k}{m}\right)^{m-k}$$

[0091]

[0092] 그리고 여기서, $0 < \delta \le 1$는 신뢰도 파라미터이므로, 멀티-암드 밴딧 문제에 대해 다음 부등식들이 각각 훈련 데이터 $D_1, \ldots, D_n$ 및 파스티리어들 $Q_1, \ldots, Q_n$에 대해 확률 $1-\delta$를 유지한다:

$$\frac{1}{n}\sum_{i=1}^{n} L_i(Q_i) \geq \frac{1}{\kappa_{max}} kl^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_i \hat{L}_i(Q_i, D_i) \middle\| \frac{D_{KL}(Q^n \| P^n) + n\,ln\big(\varepsilon(m)\big) - \ln(\delta)}{nm}\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} L_i(Q_i) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{\kappa_i}{\kappa_{max}}\hat{L}_i(Q_i, D_i) - \frac{1}{\kappa_{max}\sqrt{2nm}}\sqrt{D_{KL}(Q^n \| P^n) + n\,ln\big(\varepsilon(m)\big) - \ln(\delta)}$$

[0093]

[0094] 여기서, Q^n은 액션들 $a_1, \ldots, a_n$에 대한 조인트 분포를 나타내며, 여기서 $a_i \sim Q_i$이며, P^n는 액션들 $a_1, \ldots, a_n$에 대한 조인트 분포를 나타내며 여기서 $a_i \sim P_i$이며 $\kappa_i = \min_{a \in A, j=1,\ldots,m}\left(b_{ij}(a)\right)$이며, 그리고 $\kappa_{max} = \max_i \kappa_i$이며,

[0095] $\beta = \sqrt{\frac{1}{8}\sum_{i=1}^{n}\frac{1}{{m_i}^2}\sum_{j=1}^{m_i}\frac{1}{\left(\kappa_{i,j}\right)^2}}$ 이고,

[0096] 여기서, $\kappa_{i,j} = \min_{a \in A, j=1,\ldots,m_i}\left(b_{ij}(a)\right)$이며,

[0097] 여기서, m_i는 작업 i에 대한 훈련 세트 샘플들의 개수이다.

[0098] 컨텍추얼 밴딧 문제에 대해, 이들 목적 함수들은 L_i를 $L_i^{(cb)}$로 그리고 $\hat{L}_i$를 $\hat{L}_i^{(cb)}$로 대체함으로써 적용될 수도 있으며, Q_i 및 P_i는 액션들 a_i이 아닌 정책들 π_i에 대한 분포들이며, $\kappa_i = \min_{a \in A, j=1,\ldots,m}\left(b_{ij}(a|s_{ij})\right)/\pi_{max}$라 하고, 여기서, π_{max}는 임의의 정책 π 및 임의의 액션 상태에 대한 가능한 최대 확률 질량이다.

[0099] 하이퍼파스티리어 $\mathcal{Q}$의 성능은 예를 들어, 프라이어 P~$\mathcal{Q}$가 탑재된 기본 학습 알고리즘 $Q(D_{n+1}, P)$를 이용할 때, 새로운 작업 T_{n+1}에 대한 예상 보상인, 마진 전달 보상(marginal transfer reward)에 의해 측정된다.

[0100] 멀티-암드 밴딧 문제에 대해, 마진 전달 보상은 다음과 같다:

$$\mathrm{L}(\mathcal{Q}) = \underset{\mathrm{P} \sim \mathrm{Q}}{E}\left[\underset{(T_{n+1}, D_{n+1})}{E}\left[L_{n+1}\big(Q(D_{n+1}, P)\big)\right]\right]$$

[0101]

[0102] 컨텍추얼 밴딧 문제의 경우 마진 전달 보상은 다음과 같다:

$$L^{(cb)}(\mathcal{Q}) = \underset{\mathrm{P} \sim \mathrm{Q}}{E}\left[\underset{(T_{n+1}, D_{n+1})}{E}\left[{L_{n+1}}^{(cb)}\big(Q(D_{n+1}, P)\big)\right]\right]$$

[0103]

[0104] 상기 함수들을 이용하여, 마진 전달 보상을 최대화하는 대신, 본 방법은 마진 전달 보상에 대한 PAC-베이지안 하한들을 최대화한다. 한계들 중 일부는 2진 Kullback-Leibler, KL, 다이버전스의 역을 포함한다.

[0105] 이는 하이퍼파스티리어 $\mathcal{Q}$를 결정하는 단계가 하이퍼파스티리어 $\mathcal{Q}$ 및 하이퍼프라이어 $\mathbf{P}$의 Kullback-Leibler 다이버전스 $\mathrm{DKL}(\mathcal{Q} \| \mathbf{P})$에 따라 예상 보상 L을 결정 및 근사하는 단계를 포함한다는 것을 의미한다.

[0106] 하이퍼파스티리어 $\mathcal{Q}$는 본 예에서 n번의 반복들 i에서 결정된다.

[0107] 이는 본 방법이 n번 반복들 동안 단계 206에서 속행한다는 것을 의미한다. 이후, 선택적으로, 단계 226이 실행된다.

[0108] 단계 226에서, 본 방법은 본 예에서, 하이퍼파스티리어 $\mathcal{Q}$로부터 샘플링된 프라이어에 따라, 센서 데이터, 특히 디지털 이미지 데이터 또는 오디오 데이터를 프로세싱하는 단계를 포함한다. 센서는 마이크로폰, 비디오, 레이더, 라이다, 초음파, 모션, 열 이미지, 또는 소나 센서일 수도 있다.

[0109] 센서 데이터는 예를 들어, 센서 데이터를 분류하거나, 센서 데이터에서 오브젝트들의 존재를 검출하거나 또는 센서 데이터에 대해 시맨틱 세그멘테이션을 수행하도록 프로세싱된다. 센서 데이터 분류는 예를 들어, 컨텍추얼 밴딧 문제로서 프레이밍(frame)된다. 이미지들은 예를 들어, 교통 표지들, 도로 표면들, 보행자들, 운송체들의 존재 또는 부재 또는 유형에 따라 분류될 수도 있다.

[0110] 대신, 또는 추가적으로, 단계 226은 물리 시스템을 제어하기 위한 정책을 학습하고 하이퍼파스티리어 $\mathcal{Q}$로부터 샘플링된 프라이어에 따라 물리 시스템을 제어하기 위한 제어 신호를 결정하는 단계를 포함할 수도 있다.

[0111] 일 예로서, 하이퍼파스티리어 $\mathcal{Q}$로부터 프라이어들을 샘플링할 때, 다음 작업에 대한 예상 에러가 미리 결정된 값을 초과하지 않을 확률이 결정될 수도 있다.

[0112] 대신, 또는 추가적으로, 단계 226은 머신 학습의 강건성에 대한 척도를 결정하는 단계를 포함할 수도 있다. 본 예에서, 강건성에 대한 척도는 하이퍼파스티리어 $\mathcal{Q}$로부터 샘플링된 프라이어들에 대한 예상 에러이다.

[0113] 예컨대, 예상 에러가 임계치 미만일 때 프라이어가 센서 데이터를 프로세싱하는데 이용될 수도 있으며, 그렇지 않으면, 다른 프라이어가 센서 데이터를 프로세싱하거나 또는 제어 신호를 결정하기 위해 샘플링될 수도 있다.

[0114] 대신, 또는 추가적으로, 단계 226은 하이퍼파스티리어 $\mathcal{Q}$로부터 샘플링된 프라이어에 따라 센서 데이터에서 비정상을 검출하는 단계를 포함할 수도 있다. 본 양태에서, 비정상 검출 문제는 멀티-암드 밴딧 문제로서 프레이밍된다.

[0115] 이후, 본 방법은 본 예에서 종료한다.

[0116] 멀티-암드 밴딧 문제들의 머신 학습을 위한 방법에서 컴퓨터에 의해 취해지는 단계들에 대한 예가 아래에 제공된다:

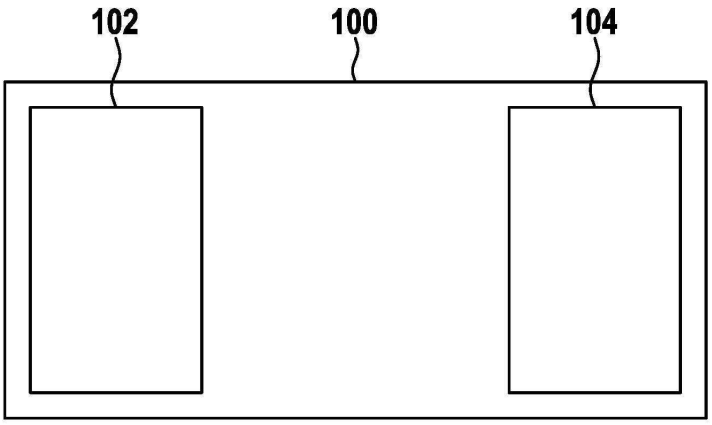
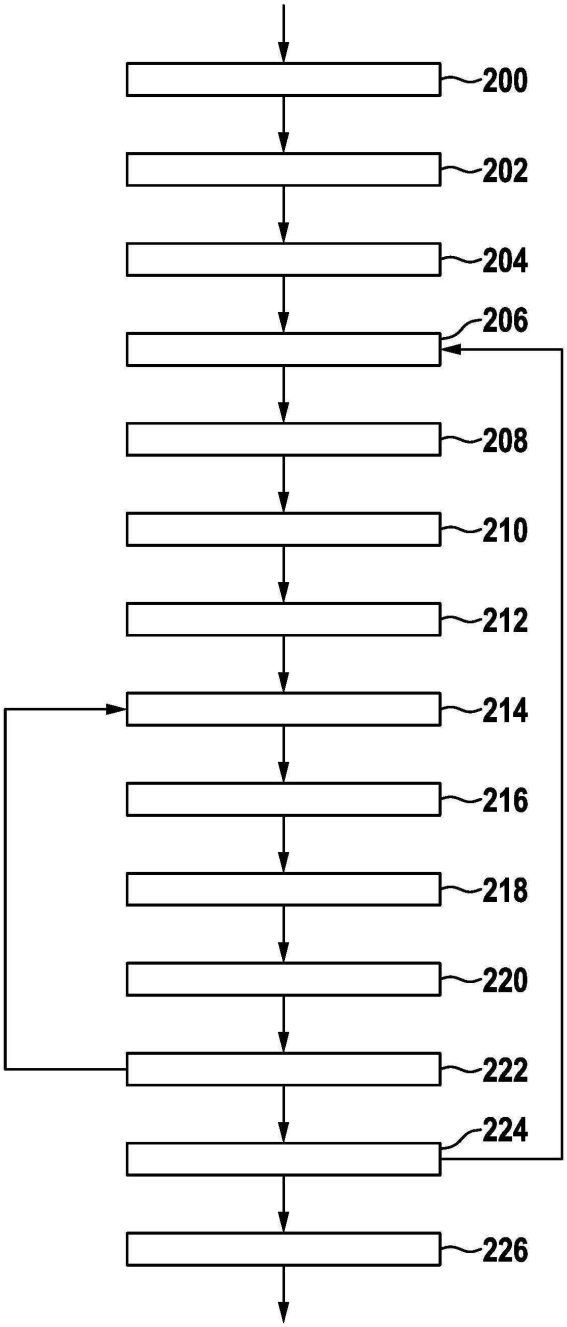
Algorithm 1 PAC-Bayesian Lifelong MABs

Input: Task distribution $\mathcal{T}$, base learning algorithm $Q = Q(D, P)$, hyperprior $\mathcal{P}$
Initialise hyperposterior $\mathcal{Q} \leftarrow \mathcal{P}$
for $i = 1$ **to** n **do**
  Sample task $T_i \sim \mathcal{T}$
  Sample prior $P \sim \mathcal{Q}$
  Initialise task posterior $Q \leftarrow P$
  Initialise task dataset $D_i \leftarrow ()$
  for $j = 1$ **to** m **do**
    Get behaviour policy $b_{ij} \leftarrow B(Q)$
    Sample $a_{ij} \sim b_{ij}(a)$ and $r_{ij} \sim \rho_i(r|a_{ij})$
    Append (a_{ij}, r_{ij}) to D_i
    Update task posterior $Q \leftarrow Q(D_i, P)$
  end for
  $\mathcal{Q} \leftarrow \text{argmax}_{\mathcal{Q}'}(\text{lower bound}(\mathcal{L}(\mathcal{Q}')))$
end for

[0117]

도면

도면1



도면2