



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0159913
(43) 공개일자 2023년11월23일

(51) 국제특허분류(Int. Cl.)
G06F 30/27 (2020.01) G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 30/27 (2020.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2022-0059281
(22) 출원일자 2022년05월16일
심사청구일자 2022년05월16일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이동현
대구광역시 남구 효성중앙길 38,205동 401호 (봉덕동, 효성타운)
(74) 대리인
윤귀상

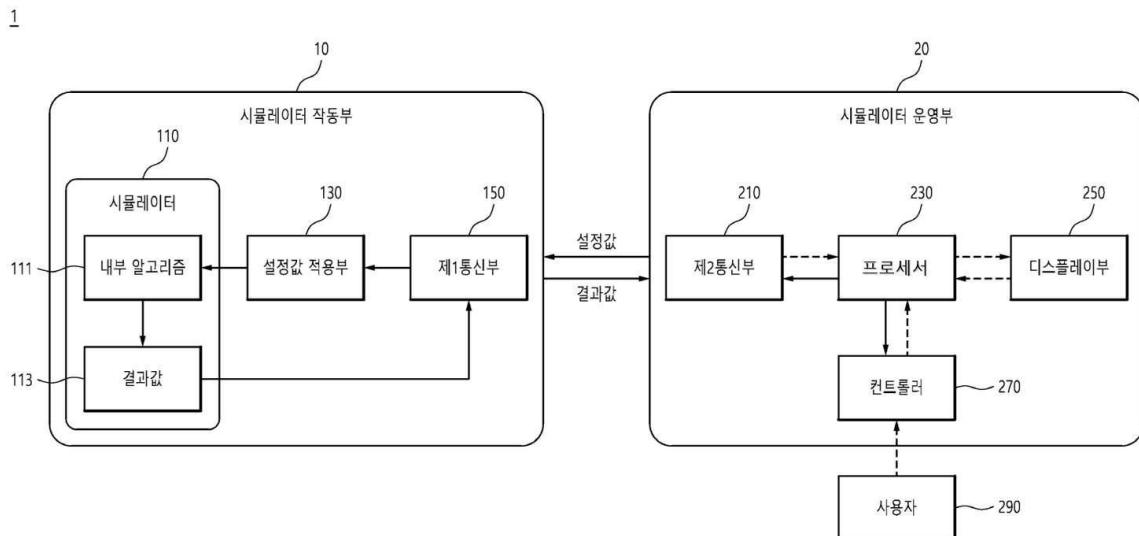
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 강화학습을 이용한 시뮬레이터 자동 운영 방법, 이를 수행하기 위한 장치 및 기록 매체

(57) 요약

강화학습을 이용한 시뮬레이터 자동 운영 장치로서, 프로세서; 상기 프로세서에 연결되는 메모리를 포함하되, 상기 메모리는, 입력값을 수신하고, 상기 입력값에 기반하여 설정값을 결정하고, 시뮬레이터에 상기 설정값을 적용하여 결과값을 도출하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치이다. 이에 따라, 시뮬레이션 시스템의 운용 단계에서 현실 환경과의 격차가 발생할 때, 이에 대응할 수 있는 강화학습 알고리즘에 기반한 시뮬레이션 시스템의 세부 조정 및 운용을 통하여, 시뮬레이션 시스템이 상시 성능을 유지할 수 있다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1711168748
과제번호	2020R1G1A1102828
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	생애첫연구
연구과제명	실적용을 목적으로 하는 강인한 인공지능용 학습 알고리즘 연구
기 여 율	1/1
과제수행기관명	고려대학교 산학협력단
연구기간	2020.09.01 ~ 2023.02.28

명세서

청구범위

청구항 1

강화학습을 이용한 시뮬레이터 자동 운영 장치로서,
프로세서;
상기 프로세서에 연결되는 메모리를 포함하되,
상기 메모리는,
입력값을 수신하고,
상기 입력값에 기반하여 설정값을 결정하고,
시뮬레이터에 상기 설정값을 적용하여 결과값을 도출하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 2

제 1항에 있어서,
상기 메모리는,
상기 결과값에 기반하여 상기 설정값을 업데이트하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 3

제1항에 있어서,
상기 프로세서로부터 결정된 설정값으로 상기 시뮬레이터를 작동하는 시뮬레이터 작동부를 포함하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 4

제3항에 있어서,
상기 시뮬레이션 작동부는 상기 프로세서로부터 상기 설정값을 전달받아 상기 시뮬레이터에 적용하는 설정값 적용부를 포함하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 5

제4항에 있어서,
상기 프로세서를 포함하는 시뮬레이터 운영부를 포함하되,
상기 시뮬레이터 운영부는 상기 결과값을 디스플레이하는 디스플레이부를 포함하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 6

제4항에 있어서,

상기 시뮬레이터 작동부와 상기 시뮬레이터 운영부는 상기 설정값과 상기 결과값을 송수신하는, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서,

상기 설정값을 결정하는 것은 하기의 수학적식을 만족하고,

$$x_t = \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$$

여기서 x_t 는 설정값이고, $\mathcal{X}$ 는 모든 가능한 결정을 포함하는 집합이고, $\bar{\mu}_t^x$ 는 평균 파라미터 세트이고, ρ_t 는 계수이며, $\nu_t^{RKG, x}$ 는 시간 t에서의 동작 x의 정규화된 놀로지 그래디언트 (Knowledge Gradient)인, 강화학습을 이용한 시뮬레이션 자동 운영 장치.

청구항 8

입력값을 수신하는 단계;

상기 입력값에 기반하여 설정값을 결정하는 단계; 및

시뮬레이터에 상기 설정값을 적용하여 결과값을 도출하는 단계;를 포함하는, 강화학습을 이용한 시뮬레이션 자동 운영 방법.

청구항 9

제8항에 있어서,

상기 결과값에 기반하여 상기 설정값을 업데이트하는 단계를 포함하는, 강화학습을 이용한 시뮬레이션 자동 운영 방법.

청구항 10

제8항에 있어서,

상기 설정값을 결정하는 단계는 하기의 수학적식을 만족하고,

$$x_t = \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$$

여기서 x_t 는 설정값이고, $\mathcal{X}$ 는 모든 가능한 결정을 포함하는 집합이고, $\bar{\mu}_t^x$ 는 평균 파라미터 세트이고, ρ_t 는 계수이며, $\nu_t^{RKG, x}$ 는 시간 t에서의 동작 x의 정규화된 놀로지 그래디언트 (Knowledge Gradient)인, 강화학습을 이용한 시뮬레이션 자동 운영 방법.

청구항 11

제8항 내지 제10항 중 어느 하나의 항에 따른 상기 강화학습을 이용한 시뮬레이터 자동 운영 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 저장 매체.

발명의 설명

기술 분야

- [0001] 본 발명은 시뮬레이터 자동 운영 시스템, 방법 및 이를 수행하기 위한 기록 매체에 관한 것이며, 더욱 상세하게는 시뮬레이터 운영시 현실 환경과의 차이가 발생하는 경우, 강화학습 알고리즘에 기반한 시뮬레이션 시스템의 세부 조정을 통하여 시뮬레이션 시스템이 상시 성능을 유지할 수 있게 하는 강화학습을 이용한 시뮬레이터 자동 운영 시스템 및 방법에 관한 것이다.

배경 기술

- [0002] 실제의 상황을 간단하게 축소한 모형을 통해서 실험을 하고 그 실험결과에 따라 행동이나 의사결정을 하는 기법을 시뮬레이션이라고 한다. 어떤 복잡한 시스템에 대하여 직접적인 방법으로 실험이나 분석을 하여 결과를 도출해 내는 것이 가장 바람직스러운 일이지만 경제적, 시간적, 효율적 측면의 문제점들이 발생될 우려가 있는 경우에 실제 시스템을 모형화하여 그 모형을 통해서 실제와 똑같은 여러 요인들을 동적으로 변화시켜서 시스템의 성격을 이해하거나 주어진 범위내에서 여러가지 실험을 실시해 본 후 그 실험 결과를 가지고 대안(代案)들을 평가하기도 하고 시스템 설계를 위한 목적에 사용하기도 한다.
- [0003] 시뮬레이션하기 위해서 설정한 모델을 시뮬레이션 모델이라 한다. 시뮬레이션 모델은 보통 수식으로 정형화되어 있지만 컴퓨터나 여러가지의 제어요소 등의 장치로 구체적으로 조립한 것을 시뮬레이터라고 한다.
- [0004] 기존 시뮬레이션 모델 또는 시뮬레이터는 특정 사용례에 한정하여 해당 실제 환경을 모사하기 위한 시뮬레이션 모델을 최적화하고, 이렇게 최적화된 시뮬레이션 모델을 통하여 생산된 분석결과를 어떻게 가공해서 사용할 것인지에 대한 내용이 주를 이룬다.
- [0005] 이는 특정 사용예인 현실 환경을 최대한 정확하게 모사하는 노력과, 이를 통해 도출된 분석결과를 효율적으로 사용하여 활용하기 위한 방법에 중점을 두어서 설계된다. 이를 위해 해당 시뮬레이터의 환경에 대한 전문적인 특화설계 및 데이터처리에 대한 가정사항들이 포함되어 실제적으로 시뮬레이터가 구현되게 된다. 통상적으로 이러한 구현물들은 운용 과정 중에 원 설계와 맞지 않는 실제상황을 접하게 되면, 이에 유동적으로 대응할 수 있는 구조적인 요소가 제한적이며, 대응하는 방법 또한 해당 시뮬레이터 및 시뮬레이션의 대상이 되는 실제 환경에 대한 전문 지식을 요한다.
- [0006] 이 경우 시뮬레이터가 실제 환경과 차이가 발생하는 경우, 시뮬레이터 내부구조와 시뮬레이션 대상분야 양쪽에 전문지식이 필요하여 능동적인 대응이 쉽지 않다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) KR 10-2008-0113348 A
(특허문헌 0002) US 2022/0091948 A1
(특허문헌 0003) KR 10-2142205 B1

발명의 내용

해결하려는 과제

- [0008] 이에, 본 발명의 기술적 과제는 시뮬레이터가 실제 환경과 차이가 발생하는 경우, 전문인력에 대한 의존도를 낮추기 위해 강화학습을 이용한 자동으로 시뮬레이터를 운영하는 시스템 및 방법을 제공하는 것이다.

[0009] 본 발명의 다른 목적은 자동으로 시뮬레이터를 운영하는 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 기록 매체를 제공하는 것이다.

과제의 해결 수단

[0010] 상기한 본 발명의 목적을 실현하기 위한 일 실시예에 따른 강화학습을 이용한 시뮬레이터 자동 운영 장치는, 프로세서; 상기 프로세서에 연결되는 메모리를 포함하되, 상기 메모리는, 입력값을 수신하고, 상기 입력값에 기반하여 설정값을 결정하고, 시뮬레이터에 상기 설정값을 적용하여 결과값을 도출하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장한다.

[0011] 본 발명의 실시예에서, 상기 메모리는, 상기 결과값에 기반하여 상기 설정값을 업데이트하도록, 상기 프로세서에 의해 실행되는 프로그램 명령어들을 저장할 수도 있다.

[0012] 본 발명의 실시예에서, 상기 프로세서로부터 결정된 설정값으로 상기 시뮬레이터를 작동하는 시뮬레이터 작동부를 포함할 수도 있다.

[0013] 본 발명의 실시예에서, 상기 시뮬레이션 작동부는 상기 프로세서로부터 상기 설정값을 전달받아 상기 시뮬레이터에 적용하는 설정값 적용부를 포함할 수도 있다.

[0014] 본 발명의 실시예에서, 상기 프로세서를 포함하는 시뮬레이터 운영부를 포함하되,

[0015] 상기 시뮬레이터 운영부는 상기 결과값을 디스플레이하는 디스플레이부를 포함할 수도 있다.

[0016] 본 발명의 실시예에서, 상기 시뮬레이터 작동부와 상기 시뮬레이터 운영부는 상기 설정값과 상기 결과값을 송신할 수도 있다.

[0017] 본 발명의 실시예에서, 상기 설정값을 결정하는 것은 하기의 수학적식을 만족할 수도 있으며,

$$x_t = \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$$

[0018] 여기서 x_t 는 설정값이고, $\mathcal{X}$ 는 모든 가능한 결정을 포함하는 집합이고, $\bar{\mu}_t^x$ 는 평균 파라미터 세트이고, ρ_t 는 계수이며, $\nu_t^{RKG, x}$ 는 시간 t에서의 동작 x의 정규화된 놀로지 그래디언트 (Knowledge Gradient)이다.

[0020] 상기한 본 발명의 다른 목적을 실현하기 위한 강화학습을 이용한 시뮬레이션 자동 운영 방법은, 입력값을 수신하는 단계; 상기 입력값에 기반하여 설정값을 결정하는 단계; 및 시뮬레이터에 상기 설정값을 적용하여 결과값을 도출하는 단계를 포함할 수도 있다.

[0021] 본 발명의 실시예에서, 상기 결과값에 기반하여 상기 설정값을 업데이트하는 단계를 포함할 수도 있다.

[0022] 본 발명의 실시예에서, 상기 설정값을 결정하는 단계는 하기의 수학적식을 만족할 수도 있으며,

$$x_t = \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$$

[0023] 여기서 x_t 는 설정값이고, $\mathcal{X}$ 는 모든 가능한 결정을 포함하는 집합이고, $\bar{\mu}_t^x$ 는 평균 파라미터 세트이고, ρ_t 는 계수이며, $\nu_t^{RKG, x}$ 는 시간 t에서의 동작 x의 정규화된 놀로지 그래디언트 (Knowledge Gradient)이다.

[0025] 상기한 본 발명의 다른 목적을 실현하기 위한 일 실시예에 따른 컴퓨터로 판독 가능한 저장 매체에는, 강화학습을 이용한 시뮬레이터 자동 운영 방법을 수행하기 위한 컴퓨터 프로그램이 기록되어 있다.

발명의 효과

[0026] 본 발명은 시뮬레이션 시스템의 운용 단계에서 현실 환경과의 격차가 발생할 때, 이에 대응할 수 있는 강화학습 알고리즘에 기반한 시뮬레이션 시스템의 세부 조정 및 운용을 통하여, 시뮬레이션 시스템이 상기 성능을 유지할

수 있게 하는 효과를 지닌다.

[0027] 종래 시뮬레이터들은 운용단계의 세부조정을 하는 데 전문인력이 필요하고, 특히 시뮬레이터 내부구조와 시뮬레이션 대상분야 양쪽에 전문지식이 필요한 경우가 많아서, 실제 운용시에 현실과 괴리가 발생할 경우 능동적인 대응이 쉽지 않다. 본 발명을 사용한다면, 이러한 괴리 상황에 전문인력에 대한 의존도를 획기적으로 낮추고, 시뮬레이터 자체 수준에서 괴리를 줄이기 위하여 자동화된 능동 대응을 비숙련 사용자도 쉽게 시작할 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 강화학습을 이용한 시뮬레이터 자동 운영 시스템의 개념도이다.

도 2는 본 발명의 강화학습을 이용한 시뮬레이터 자동 운영 방법에 대한 순서도이다.

도 3은 본 발명의 실 실시예에 따라 가우스 MAB에서 κR 에 대한 ORKG의 감도를 나타내는 그래프이다.

도 4는 본 발명의 실 실시예에 따라 가우스 MAB에서 δ 에 대한 ORKG의 감도를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0029] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예에 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

[0030] 이하, 도면들을 참조하여 본 발명의 바람직한 실시예들을 보다 상세하게 설명하기로 한다.

[0031] 도 1은 본 발명의 시뮬레이터 자동 운영 시스템의 개념도이다. 본 발명에 따른 본 발명의 시뮬레이터 자동 운영 시스템은 시뮬레이터 운영시 현실 환경과의 차이가 발생하는 경우, 강화학습 알고리즘에 기반한 시뮬레이션 시스템의 세부 조정을 통하여 시뮬레이션 시스템이 상시 성능을 유지할 수 있게 하는 기술이다.

[0032] 도 1을 참조하면, 본 발명의 시뮬레이터 자동 운영 시스템은 시뮬레이터 작동부 (10) 및 시뮬레이터 운영부 (20)로 구성된다.

[0033] 시뮬레이터 작동부 (10)는 시뮬레이터 (110), 설정값 적용부 (130) 및 제1통신부 (150)을 포함한다. 시뮬레이터 (110)는 시뮬레이션 모델로서, 시뮬레이션 모델은 보통 수식으로 정형화되어 있지만 컴퓨터나 여러가지의 제어 요소 등의 장치로 구체적으로 조립한 것을 시뮬레이터라고 한다. 본 발명의 시뮬레이터 (110)는 시뮬레이터를 동작하는 내부 알고리즘(111)과 내부 알고리즘(111)으로부터 도출되는 결과값 (113)을 포함한다.

[0034] 설정값 적용부 (130)는 시뮬레이터(110)를 동작하게 하는 설정값을 내부 알고리즘 (111)에 전달한다. 설정값은 하나 이상의 파라미터를 포함하며, 시뮬레이터 (110)가 현실 환경조건을 가장 잘 반영할 수 있도록 결정된 값이다.

[0035] 설정값 적용부(130)로부터 설정값을 수신한 내부 알고리즘(111)은 알고리즘을 수행 후 결과값(113)을 도출한다. 도출된 결과값 (113)은 제1통신부(150)로 전달된다. 제1통신부(150)는 수신한 결과값을 시뮬레이터 운영부(20)의 제2통신부(210)에 전달한다. 또한, 제2통신부(210)로부터 설정값을 수신하여 설정값 적용부 (130)에 전달한다.

[0036] 시뮬레이터 운영부 (20)는 제2통신부 (210), 프로세서(230), 디스플레이부(250) 및 컨트롤러 (270)를 포함한다.

[0037] 제2통신부 (210)는 시뮬레이터 작동부(10)의 제1통신부(150)와 설정값 및 결과값을 송수신한다. 구체적으로는 시뮬레이터 (110)의 결과값 (113)을 전달받은 제1통신부(150)로부터 결과값(113)을 수신하여 프로세서(230)에 전달한다. 또한, 프로세서(230)로부터 수신한 설정값을 제1통신부(150)에 전달한다.

[0038] 프로세서 (230)는 컨트롤러 (270)로부터 수신한 입력값에 기반하여 설정값을 결정한다. 프로세서 (230)는 자율적으로 취합된 시뮬레이터 내외부의 정보와, 사용자를 통해 공급된 정보인 입력값을 사용하여, 시뮬레이터의 운

용방식을 변경하는 행동을 취하는 결정을 내리는 데이터 기반의 소프트웨어로 구현될 수 있다. 또한, 소프트웨어의 기능을 모사하는 하드웨어 (ASIC, FPGA 등)로 일부 또는 전체가 구현되는 경우도 포함한다

[0039] 프로세서 (230)에 취합되는 정보는 3가지 형태로 저장될 수도 있다. 1) 지금까지 수집되는 정보들의 일부 또는 전체를 기록 또는 그 기반으로 계산된 값의 메모리 2) 특정 상황 이후에 수집되는 정보들을 기반으로 계산된 값의 메모리 3) 특정 상황에 특정 행동을 취했을 때, 그 이후에 수집되는 정보들을 기반으로 계산된 값의 메모리.

[0040] 프로세서 (230) 행동 결정은 위와 같이 취합된 정보들을 기반으로 정해지고, 예를 들어 다음과 같은 수식으로 결정되는 방법도 있다.

[0041] 가능한 행동들을 $\mathcal{X} = \{1, 2, \dots, K\}$ 로 나타낼 때 (K 가지가 있다고 가정)

[0042] 프로세서(230)가 결정하는 시뮬레이터의 조정 행동값 x_t 는 다음과 같은 수식으로 결정된다

$$x_t = \operatorname{argmax}_{x \in \mathcal{X}} \{\bar{\mu}_t^x + \rho_t v_t^x\}$$

[0044] 여기서, t 는 시뮬레이터가 가동된 시점부터의 지난간 시간이다.

[0045] 또한, $\bar{\mu}_t^x$ 는 동작 x 를 취했을 때 지금(t)까지 수집된 정보들의 일부 또는 전체를 기반으로 계산된 결과값의 평균이다.

$$\rho_t = \sqrt{2 \ln \left(\frac{2K}{\delta \pi_t} \right)} \cdot \frac{1}{\max\{\kappa_R, \min_{x \in \mathcal{X}} \kappa_t^x\}}$$

[0046] 또한,

$$\pi_t = \frac{1}{(t+1)^2} \cdot \frac{6}{\pi^2} \quad (\pi \text{는 원주율}), \quad \delta = 0.01, \quad \kappa_R = 0.01$$

[0048] κ_t^x 는 아래에서 더욱 상세히 설명한다.

[0049] 또한 $v_t^x = \bar{\sigma}_t^x \max\{\kappa_R, \kappa_t^x\}$ 이며,

[0050] 여기서 $\bar{\sigma}_t^x$ 는 행동 x 를 취했을 때 지금(t)까지 수집된 정보들의 일부 또는 전체를 기반으로 계산된 결과값의 표준편차이며,

[0051] 또한 $\kappa_t^x = \frac{KG_t^x}{\bar{\sigma}_t^x}$ 이고, KG_t^x 는 행동 x 를 취하는 것에 대한 놀리지 그래디언트 (참고문헌: Frazier, P.I., Powell, W.B., Dayanik, S.: A Knowledge-Gradient Policy for Sequential Information Collection. SIAM Journal on Control and Optimization 47(5), 2410-2439)이다.

[0052] 프로세서(230)가 시뮬레이션 운영부(20)에서 작동하는 방식의 예시는 다음과 같다.

[0053] 첫번째, 본 발명의 실시예로서 메타버스 (일상, 재난 등) 시뮬레이터이다.

실시예 1

[0054] 시뮬레이터 운영부(20) 작동 이후, 매 시간단위마다, 시뮬레이션 운영부(20)는 이전 시간에 정해진 행동을 그대로 취하거나, 프로세서(230)에서 결정된 새로운 행동을 취할 수 있다. 프로세서(230) 매 시간단위, 또는 어느 정도의 시간단위들이 지난 뒤에 새로운 행동을 결정해서 시뮬레이터 작동부(10)에 알려줄 수 있다.

[0055] 시뮬레이터 운영부(20)는 과거의 시뮬레이터 운영 동작들과 현재 및 미래의 예정된 동작을 사용자가 모니터링할 수 있게 디스플레이할 수 있다. 시뮬레이터 운영부(20)는, 프로세서(230)에 새로운 동작을 결정해서 통보해달라고 요청할 수 있다. 시뮬레이터 운영부(230)는, 시뮬레이터 작동부(10)로부터 전달받은 결과값을 저장할 수 있고, 저장 전 또는 후에 프로세서(230)에 상시 또는 임의대로 전달할 수 있다. 이렇게 전파된 정보는, 프로세서(230)에 의해서 가공되어 저장될 수 있다.

[0056] 입력값은 사용자(290)가 현실 환경조건에 부합하는 하나 이상의 파라미터이다. 이는 후술할 믿음 상태 (Bt : belief state)가 될 수 있다.

[0057] 또한 프로세서 (230)는 제2통신부(210)로부터 수신한 결과값 (113)을 참고하여 설정값을 업데이트할 수도 있다.

구체적인 업데이트 과정은 후술한다. 업데이트된 설정값은 제2통신부(210)를 통해 제1통신부(150)를 거쳐 설정값 적용부(130)로 전달된다. 설정값 적용부(130)는 업데이트된 설정값을 시뮬레이터(110)에 전달하고, 시뮬레이터(110)의 내부 알고리즘 (111)은 업데이트된 설정값에 따른 결과값(113)을 도출한다. 도출된 결과값(113)은 다시 제1통신부(150) 및 제2통신부(210)를 통해 프로세서(230)로 전달된다.

- [0058] 프로세서(230)는 업데이트된 설정값에 의해 도출된 결과값(113)을 재 업데이트 할 수도 있으며, 이 경우 상기한 과정을 반복한다.
- [0059] 디스플레이부(250)는 프로세서(230)로부터 수신한 결과값(1130)을 디스플레이한다. 디스플레이부(250)는 전기적으로 전송되는 화상신호를 인간이 인식할 수 있는 형태로 하는 것으로서 브라운관, 액정, 플라스마 디스플레이 등이 될 수도 있다. 사용자(290)는 디스플레이부(250)를 통해 결과값(113)을 관찰할 수도 있다.
- [0060] 본 발명의 적용되는 실시예 중 하나는 메타버스 (일상, 재난 등) 시뮬레이터이다. 즉, 시뮬레이터 작동부(10)에 일상생활 메타버스 (관광, 부동산 탐방) 및 특수상황 메타버스 (공연/스포츠행사, 재난상황)의 시뮬레이터(110)를 사용하는 실시예이다.
- [0061] 내부 알고리즘(111)은 위 상황에 상응하는 메타버스 시뮬레이터가, 컴퓨터 등의 설비를 사용해서 가동되는 상황이다. 해당 시뮬레이터(110)는 작동완료한 뒤에 결과값(113)이 자동으로 출력되지만, 해당 시뮬레이터(110)가 작동을 완료하기 전, 진행중인 상황에도 본 발명이 적용될 수 있다.
- [0062] 해당 시뮬레이터(110)를 사용하고자 하는 사용자(290)는 원할 때 시뮬레이터 운영부(20)에서 시뮬레이터의 동작 상황을 알려주는 정보를 디스플레이부(250)를 통해 모니터링 할 수 있다.
- [0063] 설정값 적용부(130)에서 설정하는 값은, 시뮬레이터(110)에서 모사되는 단위체인 에이전트들의 행동양태 모델의 파라미터값 (단위시간당 이동속도, 방향전환 빈도, 시야범위, 반응의 신속성, 결정과정의 합리성 등) 들을 포함한다.
- [0064] 제1통신부(150)는 위의 값들 일부 또는 전체를 제2통신부(210)와 송수신한다. 제1통신부(150)는 시뮬레이터 작동부(10)의 내부 상황의 일부 또는 전체를 제2통신부(210)와 송수신하며, 이 때 전달되는 내부 정보의 예로서는 에이전트 개개의 상태값 (위치, 내부적 행동판단값, 상태를 모사하는 수치데이터 등), 에이전트들이 존재하는 것으로 모사되는 환경의 상태값 (날씨, 온도, 풍향 등), 그리고 위 값들의 2차가공값 (에이전트들의 평균적인 상태값, 시뮬레이션 환경의 특정 영역 안의 에이전트 수, 온도의 시간당 변화값, 최종 결과물에 포함될 결과값의 중간 진행값 등) 등이 있다.
- [0065] 컨트롤러(270) 부분을 통해서 사용자(290)는 시뮬레이터의 운영 조작을 위한 변경 요구사항값들을 직접 입력할 수 있다. 이 때 입력되는 값들은 시뮬레이터의 구동과 관련된 값 (시뮬레이션의 연산 정밀도, 시뮬레이터를 구동하는 하드웨어 자원의 사용량 조정방향 등), 시뮬레이션 환경설정에 관련된 값 (시뮬레이션의 물리적 범주의 규모, 시뮬레이션 되는 에이전트들의 수, 시뮬레이션 단위시간의 규모 등), 시뮬레이션 결과물에 관련된 값 (최종 결과값 중 특정 종목들의 상대적/절대적 중요도 값, 최종 결과에서 정성적으로 중요한 결과값들의 목록, 정량적 결과물의 요구목표값 등) 등이 있다.
- [0066] 제2통신부(210)는 제1통신부(150)로부터 수신받은 값들과, 사용자(290)가 컨트롤러(270)를 통해 입력한 값들, 그리고 자체적으로 보유하고 있거나 외부 네트워크 (인터넷 등)에서 가져온 값 등을 기반으로 프로세서(230)의 구동양상을 조정해줄 수 있다.
- [0067] 제1통신부(150)와 제2통신부(210)의 통신은 정기적 또는 비정기적으로 일어날 수 있고, 이를 위해 업데이트가 필요한 값을 포함하여 교신되는 전체 값들의 일부만을 송신할 수도 있다.
- [0068] 프로세서(230)는 상기 통신과정을 통해 시뮬레이터(110)의 다른 부분에 내부값들로 있던 정보를 인지하며, 이렇게 내재화된 외부정보들을 사용하여 시뮬레이터(110)의 다른 부분의 행동양상을 변화시킬 수 있다
- [0069] 위의 과정을 통하여 (2)의 모니터링 부분에 표시되는 정보를 업데이트하여, user가 추가적인 조작을 내리는 데 영향을 줄 수 있는 최신 정보들을 시뮬레이터 외부로 공급할 수 있다.
- [0070] 전술한 과정을 통하여 제1통신부(150)와 제2통신부(210)의 통신과정을 거쳐 시뮬레이터(110)의 다른 부분의 내부값들을 변경하여 시뮬레이터의 동작 양상을 변경할 수 있다. 프로세서(230)는 과거에 제1통신부(150)와 제2통신부(210)의 통신과정을 통하여 받은 값들을 내부적으로 저장하고 있을 수 있으며, 이를 기반으로 과거 내용도 최신 정보들과 함께 고려하여 행동양상을 변화시킬 수 있다.

[0071] 또한, 본 발명의 또 다른 실시예로서 자동차 주행 시뮬레이터가 있다.

실시예 2

[0072] 본 실시예는 시뮬레이터의 목적이 자동차의 주행을 시뮬레이션하는 것으로서, 주행시 주변환경 등과 함께, 에이전트들인 운전자, 동승자, 보행자 및 주변 차량들을 시뮬레이션 하는 장치인 경우이다.

[0073] 해당 시뮬레이터(110) 안에서 구동되는 에이전트들 중 하나 이상에는 인공지능적인 알고리즘이 포함되어 있을 수 있다. 예를 들어, 운전자 에이전트는 인공지능적으로 감/가속, 조향, 지시등 등을 조절하는 소프트웨어가 에이전트의 결정을 담당하고 있는 상황을 포함한다.

[0074] 실시예1에서 추가적으로, 설정값으로 쓰일 수 있는 것에는, 에이전트들의 행동양태 모델의 파라미터값 (단위시간당 이동속도, 방향전환 빈도, 시야범위, 반응의 신속성, 결정과정의 합리성 등) 들에는 인공지능적인 에이전트의 행동양태를 결정하는 값 (예를 들어, 인공신경망모델, 학습알고리즘 및 그 설정값 등)들이 포함된다.

[0075] 또한, 본 발명의 또 다른 실시예로서 설비/장치 구동 시뮬레이터가 있다.

실시예 3

[0076] 실시예 1 대비하여, 시뮬레이터(110)의 목적은 특정한 설비 또는 장치가 가동하는 것을 시뮬레이션하는 것으로서, 상기 설비 또는 장치가 실물로 구동되고, 이에 대응하는 주변환경을 실물장비의 센서 등을 통하여 시뮬레이션으로 제공하는 장치이다.

[0077] 예를 들어, 하드디스크나 SSD, NAS, 데이터서버 등의 데이터 저장장치가 실제로 구동되고 있으며, 본 장치에 실제로 데이터를 저장하고 회수하는 환경을 해당 시뮬레이터가 제공하고 있는 상황을 포함한다.

[0078] 또 다른 예로는, 특정 생산설비 또는 시설/건물이 실제로 사용되고 있으며, 이 설비/시설의 센서나 조작부 등을 통하여 시뮬레이터가 값을 기록하는 등의 방식으로 구동되는 실물 시설/건물이 마치 해당 환경에 처한 것 처럼 시뮬레이션하고 있는 상황을 포함한다.

[0079] 실시예 1대비하여, 제1통신부(150)가 기록하는 설정값들 및, 수집하는 상태값들은, 각각 실물 설비의 외부 조작부의 값과 센서의 상태값들 뿐만 아니라, 실물 구동되는 설비 또는 장치의 내부값들도 포함할 수 있다.

[0080] 도 2는 본 발명의 시뮬레이터 자동 운영 방법이다. 도 2를 참조하면, 사용자는 컨트롤러에 입력값을 입력한다(S10). 프로세서(230)는 입력값에 기반하여 설정값을 결정한다(S20). 이하 프로세서(230)가 설정값을 결정하는 단계를 더욱 구체적으로 설명한다.

[0081] 이 과정은 의사 결정자가 일련의 결정에서 반복적으로 선택하고 제한된 수량화 보상이 도출될 수 있는 결과를 관찰하는 불확실성에 직면하여 가능한 최선의 결정을 내리는 문제를 고려한다. 에이전트는 가능한 결정 집합을 알고 있다고 가정한다. 이는 유한하며 에이전트가 선택하고 학습하는 시간 범위 동안 동일하게 유지된다. 그러한 에이전트가 학습하는 동안 선택한 선택에 의해 발생하는 보상을 무시하고 어떤 선택이 가장 좋은 보상을 얻을 것인지를 얼마나 잘 찾아내는지로 평가된다면 에이전트는 순위 및 선택(R&S) 문제에 직면하게 된다.

[0082] 이 과정의 새로운 기여는 다음과 같이 요약된다. 놀로지 그래디언트 (KG : knowledge gradient)를 사용하는 새로운 온라인 학습 알고리즘인 독립적인 가우스 믿음을 가진 ORKG(Online Regularized Knowledge Gradients) 알고리즘을 사용한 온라인 학습을 제시한다. 확률적 MAB 문제에서 $O(\sqrt{KT \ln(KT)})$ 의 ORKG의 리그렛(regret) 상한은 K 경계의 독립 가우시안 암 (arms)을 사용하여 KG기반 알고리즘에 대한 첫 번째 하위 선형 리그렛 경계이다. 또한 ORKG의 이론적 특성에 대한 실증적 검증과 ORKG의 주요 하이퍼파라미터에 대한 경험적 민감도 분석을 수행한다. 마지막으로, 검증 가능한 리그렛 범위를 가진 다른 잘 알려진 MAB 알고리즘에 대한 가우스 확률적 MAB 문제에서 ORKG의 경험적 성능을 검증한다.

[0083] 본 기술에서 의사 결정 에이전트가 부분 정보 확률론적 MAB 문제, 특히 제한된 가우시안 확률적 암과 결정할 수 없는 총 결정 횟수에 직면하는 "온라인" 순차 결정 문제를 고려한다. 유한 시간 호라이즌 T 를 알 수 없는 각 시간 인덱스 $t \in \{0, 1, \dots, T-1\}$ 에 대해 에이전트는 인덱싱할 수 있는 $K < \infty$ 상호 독립적인 암 중에서 $\mathbf{x}$ 로 표시되는 결정을 내려야 한다. $i \in \{1, 2, \dots, K\}$ 에 의해, 그 후 에이전트에게 알려지지 않은 평균 μ_i 및 표준 편차 σ_i 를 사용하여, 각 암의 분포에서 제한된 무작위 보상/기여도 C_i 를 관찰한다. 우리는 시간 t에서 내

린 결정에 x_t 를 사용하고 모든 가능한 결정을 포함하는 집합으로 $\mathcal{X}$ 를 사용한다. 즉 $\forall t : x_t \in \mathcal{X}$, and $|\mathcal{X}| = K$ 이다.

[0084] 에이전트의 목표는 두 가지이다. 1) 과거 결정에 의해 초래된 관찰 결과를 사용하여, 보상 분포의 평균이 가장 큰 경우 (즉, $\mu^{i*} = \max_i \{\mu^i\} =: \mu^*$) 인 최고의 암을 배우는 것 2) 최고의 암을 사전에 알지 못한 채 최고의 암이 아닌 암을 선택함으로써 야기되는 불가피한 차선택의 영향을 제어하는 것이다. "온라인"이라는 용어는 온라인 볼록(convex) 최적화와 동일하지 않지만 학습 에이전트의 목표의 두 번째 측면과 관련이 있다. R&S 문제와 같은 일괄 학습과 달리 에이전트가 학습하는 동안(즉, "온라인") 에이전트의 성능이 중요하다. 여기서 최종 성능만 중요하다.

[0085] 독립 가우시안 믿음 모델을 가진 KG 기반 알고리즘에 대해, 믿음 상태 (Bt : belief state)는 각 행동 $x \in X$ 에 의해 발생하는 가우스 보상을 모델링하기에 충분한 정보로 정의된다. 따라서 $B_t := \{(\bar{\mu}_t^x, \bar{\sigma}_t^x) | x \in \mathcal{X}\}$ 로 정의한다. 모든 $x \in X$ 에 대해 평균 파라미터 세트가 $\bar{\mu}_t^x$ 를 추정하고 표준편차 파라미터가 $\bar{\sigma}_t^x$ 를 추정한다.

[0086] 독립적인 가우스 믿음 모델에서 시간 t에서 x를 선택하는 KG는 다음과 같은 클로즈 형태 공식 [수학식 1]을 사용하여 효율적으로 계산될 수 있다.

수학식 1

[0087]
$$\nu_t^{KG,x} := \tilde{\sigma}_t^x (\xi_t^x \Phi(\xi_t^x) + \phi(\xi_t^x))$$

[0088] 여기서, $\Phi(\cdot)$ 및 $\phi(\cdot)$ 는 각각 표준가우스분포의 누적분포함수 및 확률밀도함수이며, 여기서 ξ_t^x 는 아래의 [수학식 2]로 정의된다.

수학식 2

[0089]
$$\xi_t^x := -\frac{|\bar{\mu}_t^x - \max_{x' \neq x} \bar{\mu}_t^{x'}|}{\tilde{\sigma}_t^x}$$

[0090] 여기서 $\tilde{\sigma}_t^x := \bar{\sigma}_t^x / \sqrt{1 + (\sigma^\epsilon / \bar{\sigma}_t^x)^2}$ 이며, σ^ϵ 는 모든 $x \in \mathcal{X}$ 에 대해 모든 관찰된 보상 $C(x)$ 에서 발견된 것으로 간주되는 제로-민 (zero-mean) 가우시안 측정 노이즈의 표준편차이다. 대부분의 KG 기반 알고리즘은 하이퍼파라미터로 사용되는 σ^ϵ 를 가진다

[0091] KG를 그대로 사용하여 온라인 학습 문제를 해결하는 것은 실패할 것으로 예상된다. R&S 문제가 고정된 알려진 수의 선택으로 인한 보상을 무시하고, 이를 학습 프로세스로 간주하기 때문이다. 이러한 관점에서 온라인 학습 문제(OKG)에 대한 KG 알고리즘이 제안된다. OKG 알고리즘은 아래 [수학식 3]과 같이 시간 t에서 동작 x_t 를 선택한다.

수학식 3

[0092]
$$x_t = \begin{cases} \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + (T - t) \nu_t^{KG,x} \right\} & (t < T) \\ \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x \right\} & (t \geq T) \end{cases}$$

[0093] 여기서 T는 온라인 학습 문제에서 선택할 수 있는 총 수이다. 당연히 OKG 알고리즘은 실제 시간 호라이즌 T를 알아야 하며, 그 후에 학습된 정보를 활용하고 가장 기대되는 평균 보상이 있는 작업을 선택한다.

[0094] 본 기술은 독립적인 가우시안 믿음이 있는 정규화된 KG를 사용한 온라인 학습(ORKG)이며, 이는 놀로지 그래디언트가 있는 새로운 온라인 학습 알고리즘 (이하 설명할 알고리즘 1)이다. OKG 알고리즘과 비교하여 ORKG는 혁신에 두 가지 핵심을 도입한다. 1) 놀로지 그래디언트의 표준화 및 정규화 2) 탐색 파라미터 ρ_t 를 적응적으로 학습하는 것이다. ORKG는 2개의 핵심 하이퍼파라미터 $\kappa_R > 0$ 및 $0 < \delta < 1$ 을 포함하며 $\kappa_R = 0.01$, $\delta = 0.01$ 을 기본값으로 사용한다.

[0095] 하기의 알고리즘 1은 독립적인 가우스 믿음을 가진 ORKG이며, 하기의 순서로 진행된다. 특히 6 단계의 [수학식 4]와 같이, ORKG 알고리즘은 시간 t 에서 동작 x 를 결정한다.

수학식 4

[0096]
$$x_t = \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$$

[0097] [알고리즘 1]

[0098] 1단계 : 믿음 상태 초기화 : $\{\bar{\mu}_0^x, \bar{\sigma}_0^x\}_{x \in \mathcal{X}}$

[0099] 2단계 : 시간 $t = 0, 1, 2, \dots$ 동안 진행

[0100] 3단계 : 표준화된 KG 연산 : $\kappa_t^x \leftarrow \frac{\nu_t^{KG, x}}{\bar{\sigma}_t^x}$

[0101] 여기서 $\nu_t^{KG, x}$ 는 [수학식 1]에 의해 연산한다.

[0102] 4단계 : 정규화된 KG 연산 : $\nu_t^{RKG, x} \leftarrow \bar{\sigma}_t^x \max(\kappa_R, \kappa_t^x)$

[0103] 5단계 : 계수 연산 : $\rho_t \leftarrow \sqrt{2 \ln \left(\frac{2|\mathcal{X}|}{\delta \pi_t} \right)} \frac{1}{\max\{\kappa_R, \min_{x \in \mathcal{X}} \kappa_t^x\}}$

[0104] 6단계 : 동작 결정 : $x_t \leftarrow \arg \max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + \rho_t \nu_t^{RKG, x} \right\}$

[0105] 7단계 : 보상 $C_{t+1} \sim C(x_t)$ 관찰

[0106] 8단계 : C_{t+1} 을 이용하여 $x = x_t$ 에 대해 $\bar{\mu}_{t+1}^x, \bar{\sigma}_{t+1}^x$ 업데이트

[0107] 여기서 $\rho_t := \sqrt{2 \ln \left(\frac{2|\mathcal{X}|}{\delta \pi_t} \right)} \frac{1}{\max\{\kappa_R, \min_{x \in \mathcal{X}} \kappa_t^x\}}$ 이며, 여기서 $\delta \in (0, 1)$ 이고, π_t 는

$\sum_t^\infty \pi_t = 1$ 을 만족하는 시퀀스 (asequence) 이다. 예를 들어, $\pi_t = \frac{1}{(t+1)^2} \frac{6}{\pi^2}$ 이다. 이 ρ_t 를

사용하여, ORKG는 현재 동작 x에 의해 발생한 평균 보상의 현재 추정치인, $\bar{\mu}_t^x$ 를 최대화하는 익스플로잇 (exploitation) 동작과, 시간 t에서의 동작 x의 정규화된 놀로지 그래디언트 $\nu_t^{RKG, x}$ 를 최대화하는 익스플

로레이션 (exploration) 동작을 밸런싱 (balance)한다. ORKG는 시간 호라이즌 T를 알 필요가 없다. 반면 OKG 알고리즘은 참 (true) T를 알아야 합니다. 이 속성을 사용하면 ORKG를 온라인 학습 문제에 쉽게 적용할 수 있다. 명시적인 호라이즌의 끝은 알 수 없거나 시간이 지남에 따라 변경된다.

[0108] 이하 본 기술에 대한 이론적 분석을 더욱 구체적으로 설명한다.

[0109] ORKG 알고리즘에서 사용되는 KG의 정규화를 정의하고, ORKG의 준선형 리그렛 경계가 의존하는 정규화된 KG의 이론적 속성을 분석한다. ORKG 알고리즘에서 KG의 정규화에 대한 개념적 요약은 1) KG를 단위 없는 값 (unitless value)으로 "표준화"하고, 2) 아래에서 고정된 균일 경계를 갖도록 강제하고, 3) KG와 일치하도록 단위를 되돌려주는 것이다. 1단계는 표준화된 KG를 계산하여 이루어지며, 2, 3단계는 표준화된 KG에서 정규화된 KG를 계산하여 수행된다.

[0110] 시간 t에서의 동작 $x \in X$ 의 표준화된 놀로지 그래디언트 κ_t^x $x \in X$ 에 대해 아래 [수학식 5]에 의해 정의된다.

수학식 5

$$\kappa_t^x := \frac{\nu_t^{KG,x}}{\bar{\sigma}_t^x}$$

[0112] 여기서 놀로지 그래디언트 $\nu_t^{KG,x}$ 는 믿음 상태 B_t 로부터 연산된다.

[0113] κ_t^x 는 [아래이 수학식 6]의 ξ_t^x 와 같은 단위를 갖는다는 점에서 "표준화된" KG이다.

수학식 6

$$\kappa_t^x = \frac{\bar{\sigma}_t^x}{\underbrace{\sqrt{(\bar{\sigma}_t^x)^2 + (\sigma^\epsilon)^2}}_{\text{unitless}}} \underbrace{(\xi_t^x \Phi(\xi_t^x) + \phi(\xi_t^x))}_{\text{same unit as } \xi_t^x}$$

[0115] 여기서 ξ_t^x 는 [수학식 2]에서 정의한 바이며, Φ 는 누적 분포 함수이고 ϕ 는 표준 정규 분포의 확률 밀도 함수이다.

[0116] 이하 ORKG의 리그렛의 하위 선형 상한에 대해 필요한 속성을 달성함과 동시에 해석하기 쉽게 설계된 다음과 같은 정규화 방법을 소개한다.

[0117] 주어진 믿음 상태 B_t 와 시간 t에서의 동작 x를 결정하기 위한 정규화된 KG인 $\nu_t^{RKG,x}$ 는 아래의 [수학식 7]에 의해 정의된다.

수학식 7

$$\nu_t^{RKG,x} := \bar{\sigma}_t^x \max \{ \kappa_R, \kappa_t^x \}$$

[0119] 여기서 $\kappa_R > 0$ 은 모든 x, t 에 대한 κ_t^x 에 대한 작은 임의의 일정한 균일 하한인 정규화 파라미터이며, κ_t^x 는 [수학식 5], [수학식 6]에 따라 주어진 믿음 상태 B_t 에서의 시간 t 에서 연산된 표준화된 KG이다.

[0120] 이 정규화에서 ORKG 알고리즘의 두 하이퍼파라미터 중 하나인 κ_R 이 발생한다. κ_R 은 κ_t^x 가 모든 x, t 에 대해 얻을 수 있는 작은 범위에 대한 균일한 하한을 나타낸다.

[0121] 이하 본 발명에 따른 성능에 대한 다각적 실증적 검증을 제시한다. 온라인 학습에서 ORKG 알고리즘의 Python 패키지 smpybandit[1]을 사용하여 64GB RAM이 장착된 AMD Ryzen 3900x CPU에서 모든 확률적 MAB(Multi-armed bandit) 벤치마크를 구현한다. 벤치마크는 무작위화되어 100번 반복되며 모든 반복에서 표본 평균과 표준 편차가 보고된다.

[0122] 첫째, 제한된 가우스 확률적 MAB 문제에서 KG 기반 알고리즘의 경험적 성능을 비교하여 ORKG의 이론적 개선이 실현되는 방법을 보여준다. ORKG 알고리즘을 독립적인 가우시안 믿음 알고리즘(KG)이 있는 KG[6]와 온라인 학습 문제 알고리즘(OKG)의 일반 클래스에 대한 KG[13]와 비교한다. 또한 애플리케이션에서 빈번히 보이는 널리 알려진 벤치마크 알고리즘으로 상수 $\epsilon(t) = 0.01$ 을 사용하여 ϵ -greedy 알고리즘을 테스트한다. 알고리즘의 주요 차이점은 [표 1]에 요약되어 있다. 공정한 비교를 위해 KG 알고리즘 간의 공통 하이퍼파라미터 값으로 $\sigma_\epsilon = 0.1$ 을 사용한다.

표 1

	Decision Rule	Belief State	Hyperparameters	Regret Bound
ϵ -greedy	$\bar{\mu}_t^x$ w.p. $1 - \epsilon$	$\bar{\mu}_t^x$	$\epsilon(t)$	N/A
KG	$\nu_t^{KG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	σ_ϵ	N/A
OKG	$\bar{\mu}_t^x + (T - t)\nu_t^{KG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	σ_ϵ, T	N/A
ORKG	$\bar{\mu}_t^x + \rho_t \nu_t^{RK G,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	$\sigma_\epsilon, \delta, \kappa_R$	$O\left(\sqrt{ \mathcal{X} T \ln \mathcal{X} T}\right)$

[0123]

[0124] 확률적 MAB 벤치마크 문제에 대한 알고리즘을 가우스 보상을 생성하는 5, 10 및 20 암으로 테스트한다. 평균 파라미터 μ_x 는 $[-5, 5]$ 에서 동일한 거리로 샘플링되고 모든 동작 x 에 대해 $\sigma_x^2 = 0.1$ 의 낮은 분산 시나리오와 높은 분산 시나리오 $\sigma_x^2 = 1$ 를 갖는다.

[0125] $t = 1, \dots, 10000$ 에서 관찰된 리그렛을 합산하고 [표 2]에 100개의 독립적 반복에서 평균 및 표준 편차를 나타낸다.

표 2

MAB Setting		Algorithms			
Arms	Variance	ORKG	OKG	KG	ϵ -greedy
5	High	215 ± 102	204 ± 96	33100 ± 256	8830 ± 8200
5	Low	17 ± 9	12 ± 12	33200 ± 235	6570 ± 8110
10	High	1060 ± 85	2580 ± 3210	39600 ± 355	14700 ± 11100
10	Low	40 ± 9	1020 ± 2840	40600 ± 241	17400 ± 11900
20	High	2210 ± 105	5950 ± 3900	39900 ± 774	19600 ± 10500
20	Low	96 ± 10	6210 ± 4690	44400 ± 264	21100 ± 14500

[0126]

[0127] ORKG는 테스트된 모든 설정에서 누적 리그렛을 제어하는 예상 동작을 보여 주는 반면, 하위 선형 리그렛 경계가

없는 다른 KG 알고리즘은 대부분 큰 리그렛을 나타낸다.

- [0128] OKG는 실제 시간 범위 $T = 10000$ 에 대한 추가 정보가 제공되더라도 10개와 20개 암이 있는 더 어려운 설정이 아닌 단지 5개의 암에서 ORKG에 필적하는 결과를 얻는다. R&S 문제를 해결하려는 KG는 예상대로 최악의 실적이나 아쉬움을 남겼다.
- [0129] 실제로 널리 사용되는 알고리즘인 ϵ -greedy는 매우 큰 표준편차를 보여 온라인 학습에서 승부차기 (hit-or-miss) 성능을 나타낸다.
- [0130] RKG는 [표 1]과 같이 다른 KG 알고리즘과 비교하여 새로운 하이퍼파라미터 δ 및 κR 을 도입한다. 이러한 하이퍼파라미터는 ORKG의 하위선형 리그렛 경계에서 중요한 역할을 하기 때문에 본 발명은 δ 및 κR 에 대한 ORKG의 경험적 민감도를 하나씩 분석하여 테스트하였다.
- [0131] 우선, $\delta = 0.01$ 로 고정하면서 $\kappa R \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ 을 변경하고 [도 3]에 표시된 궤적과 같이 100회 반복을 통해 평균한 t 에 대한 누적 리그렛의 시간 변화를 나타낸다.
- [0132] 도 3은 본 발명의 실 실시예에 따라 가우스 MAB ($K = 10$, $\sigma^2 = 1$, $\delta = 0.01$)에서 κR 에 대한 ORKG의 감도를 나타내는 그래프이다.
- [0133] ORKG는 광범위한 κR 에 관계없이 견고한 리그렛 제어를 보여주고, KG 및 OKG에 비해 견고한 이점을 유지한다는 것을 나타낸다. KG의 하한을 적용하고 KG 표면의 부드러움 (smoothness)을 정규화하는 ORKG에서의 κR 의 직관적인 역할을 고려할 때 κR 에 대한 미묘한 감도는 이론적으로 예상되며 다음과 같이 해석될 수 있다. κR 을 변경하면 탐색이 발생할 때 ORKG 결정 규칙 [수학식 4]의 값을 일시적으로 변경할 수 있다. 이는 도 3에서 다른 κR 값을 가진 ORKG의 초기 단계 궤적($T < 10^3$)의 사소한 차이로 나타난다.
- [0134] KG의 직관적인 의미는 단일 보상에서 기대되는 개선을 의미하기 때문에, KG의 낮은 경계가 될 정도로 그 값이 충분히 작아야 한다는 이해를 기반으로 기본값 $\kappa R = 0.01$ 을 권장한다. 또한 κR 은 선행적 정보로 또는 문제 공식화 단계에서 조정할 수 있다. 보상의 최대 평균과 최소 평균 사이의 간격이 알려져 있거나 보상을 클리핑하여 적용할 수 있는 경우 κR 은 간격보다 최소한 충분히 작도록 세팅될 수 있다.
- [0135] 다음으로, $\kappa R = 0.01$ 로 고정하면서 $\delta \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ 를 변경하고 도 4에 표시된 궤적과 같이 100회 반복에 걸쳐 평균한 t 에 대한 누적 후회의 시간 변화를 나타낸다.
- [0136] 도 4는 본 발명의 실 실시예에 따라 가우스 MAB($K = 10$, $\sigma^2 = 1$, $\kappa R = 0.01$)에서 δ 에 대한 ORKG의 감도를 나타내는 그래프이다.
- [0137] δ 가 중간 범위 $10^2 < T < 10^4$ 에서 ORKG의 행동에 영향을 미치는 것은 다양하며 대부분의 경우 $\delta \leq 0.1$ 의 경우에 리그렛이 제어되기 때문에 영향이 일시적인 것으로 나타난다. $\delta = 0.9$ 의 경우 ORKG의 극적으로 다른 행동이 관찰되며 이는 ORKG에서 δ 의 역할에 따라 예상된다. 보상을 만날 확률 δ 는 추정된 평균과 p_t 로 스케일링된 탐사 보너스 조건보다 더 많이 벗어난다([수학식 4]에서 주어진 것처럼). δ 가 ORKG [수학식 4]에 따라 행동을 선택한 후 예기치 않은 보상에 직면하는 드문 이벤트의 확률이기 때문에, 직관적으로 δ 가 크면 탐욕스럽게 익스플로잇하기 전에 ORKG를 더 신중하게 한다. 이는 도 4에서 $\delta = 0.9$ 인 경우 ORKG에 의해 경험적으로 나타난다. 따라서 $\delta \in (0, 1)$ 가 이론적으로 허용되더라도 ORKG의 δ 를 비교적 작은 값으로 설정하는 것이 합리이며, δ 는 ORKG가 얼마나 희귀 이벤트를 발생할지를 예상하는지를 조정한다. 1%는 "희귀한" 이벤트가 발생하기 위한 좋은 기준점이므로 $\delta = 0.01$ 의 기본값을 권장한다. 더 빈번한 놀라움이 예상되는 경우 더 큰 δ 가 권장된다.
- [0138] 이상 프로세서(230)가 설정값 X_t 를 결정하는 단계 (S20)을 설명하였다.
- [0139] 그 후 설정값은 시뮬레이터 작동부(10)에 전달한다 (S30). 더욱 구체적으로는 설정값은 제2통신부(210)로 전송되고 시뮬레이터 작동부(10)의 제1통신부(150)로 전달된다. 그 후 제1통신부는(150)는 설정값 적용부(130)에 설정값을 전달한다. 설정값 적용부(130)는 시뮬레이터(110)의 내부 알고리즘에 (111)에 설정값을 적용한다.
- [0140] 그 후 시뮬레이터(110)는 설정값에 따라 시뮬레이터를 동작한다(S40). 구체적으로는 시뮬레이터(110)의 내부 알고리즘(111)이 구동하여 결과값(113)을 도출한다.
- [0141] 그 후 시뮬레이터(110)의 결과값(113)을 시뮬레이터 운영부(20)에 전달한다(S50). 구체적으로는 결과값(113)은 시뮬레이터 작동부(10)의 제1통신부(150)를 통해 시뮬레이터 운영부(20)의 제2통신부(210)로 전달된다. 그 후

결과값(113)은 시뮬레이터 운영부(20)의 프로세서(230)로 전달된다.

[0142] 프로세서(230)는 결과값에 기반하여 설정값을 업데이트한다 (S60). 여기서 결과값은 전술한 [알고리즘 1]의 8단계의 과정을 통한다. 관찰된 결과값 C_{t+1} 을 사용하여 $\bar{\mu}_{t+1}^x$, $\bar{\sigma}_{t+1}^x$ 를 업데이트 하고, 시간 $t+1$ 에서의 X_{t+1} 를 결정한다.

[0143] 이와 같은, 강화학습을 이용한 시뮬레이터 자동 운영 방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.

[0144] 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0145] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CDROM, DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-opticalmedia), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0146] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용하여 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0147] 이상에서는 실시예들을 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

산업상 이용가능성

[0148] 4차 산업혁명의 핵심은 컴퓨터의 발전과 활용이다. 특히 컴퓨터 시뮬레이터는 열 흐름, 화학반응과 다상(고상·액상·기상) 유체 흐름 등 전 공학 분야에 걸쳐 안정적이고 효율적인 시스템의 연구개발을 위해 널리 활용되는 전문 기술이다.

[0149] 이 기술이 상용화된다면 어렵고 복잡한 컴퓨터 시뮬레이션에 대한 부담을 획기적으로 줄일 수 있고, 그 목적에 따라 누구나 쉽고 간단하게 접근 가능하다.

부호의 설명

[0150] 1 : 시뮬레이터 자동 운영 시스템

10 : 시뮬레이터 작동부

20 : 시뮬레이터 운영부

110 : 시뮬레이터

111 : 내부 알고리즘

113 : 결과값

130 : 설정값 적용부

150 : 제1통신부

210 : 제2통신부

230 : 프로세서

250 : 디스플레이부

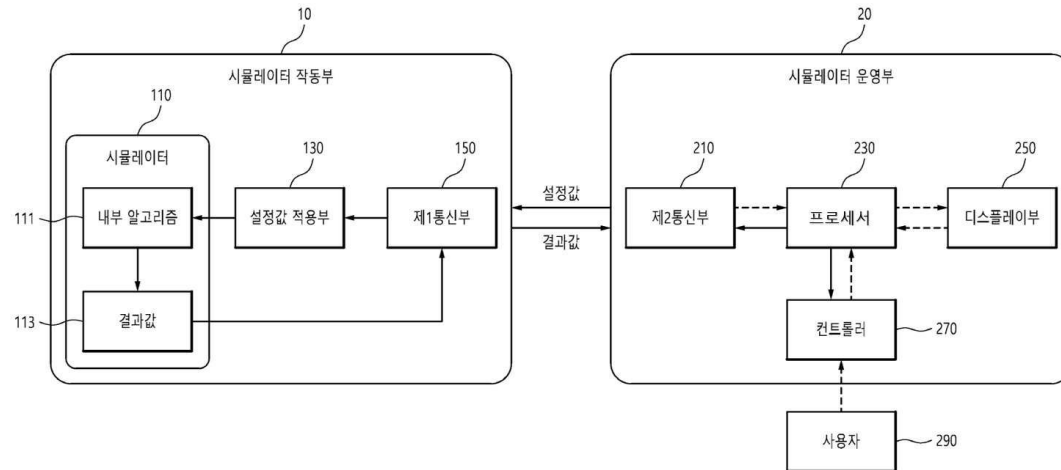
270 : 컨트롤러

290 : 사용자

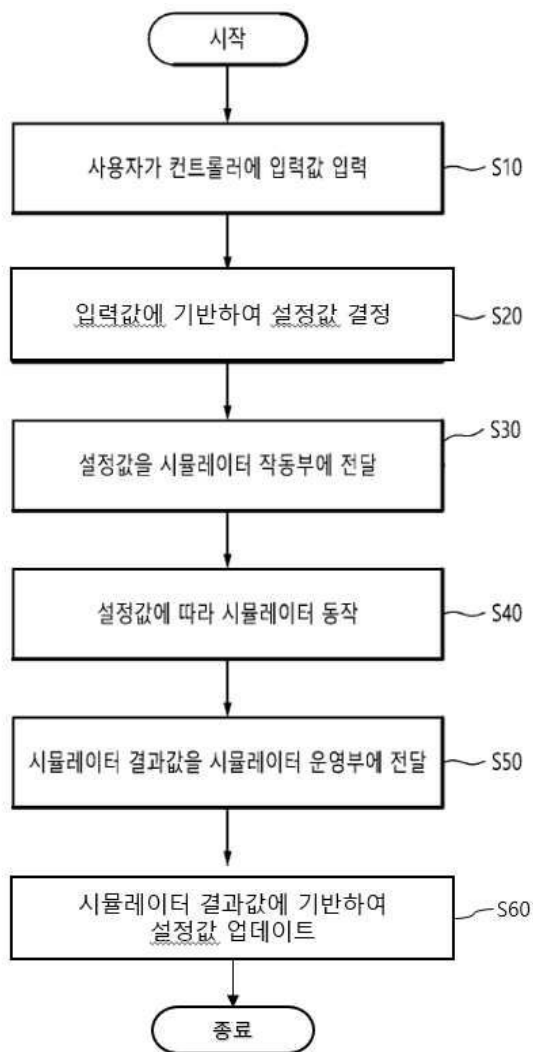
도면

도면1

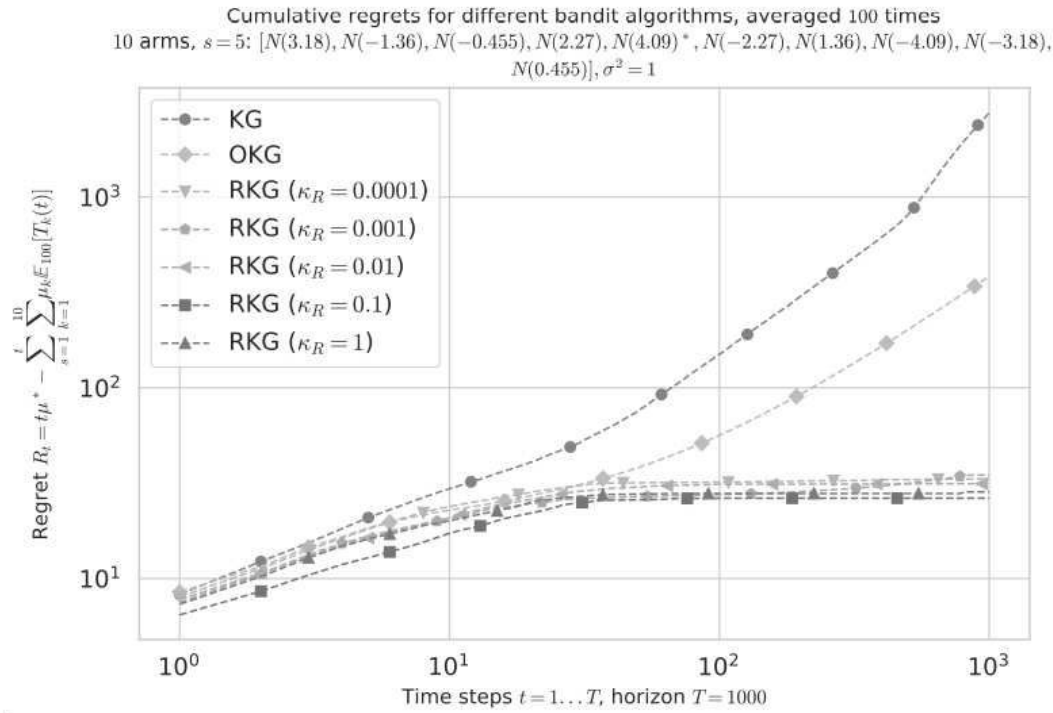
1



도면2



도면3



도면4

