(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0016859
(43) 공개일자 2025년02월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/092 (2023.01) *G06N 3/045* (2023.01)
G06N 3/047 (2023.01)
(52) CPC특허분류
G06N 3/092 (2023.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2023-0097385
(22) 출원일자 2023년07월26일
심사청구일자 없음

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이병준
서울특별시 광진구 아차산로36길 39, 701동 1513호
안희준
경기도 부천시 상오정로 16, 301호
(74) 대리인
김준석, 박민욱

전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **다중 할인율을 적용한 오프라인 강화학습 방법**

(57) 요 약

본 발명의 일 기술적 측면에 따른 오프라인 강화 학습 방법은 할인율을 코사인 엠베딩 기법을 이용하여 고차원 벡터로 전환하고, 할인율과 데이터셋을 입력 데이터로 설정하는 단계; 할인율과 데이터셋을 오프라인 강화학습 알고리즘을 적용하여 Q-함수와 할인율에 대응하는 예측값을 출력하도록 학습하는 단계;를 포함한다.

대 표 도 - 도1

Algorithm 1 학습 알고리즘

Require: 데이터셋 $\mathcal{D}$, M개의 앙상블 Q-함수 네트워크 Q_k, 정책 네트워크 π_θ, 배치 사이즈 B.

할인율 리스트 Γ : 사전에 정해진 최소/최대 할인율과 유효 지평선 차이에 따라서 생성
할인율에 코사인 엠베딩 적용
모든 네트워크 초기화
for iteration t= 1, 2, ... n **do**
  B 크기만큼 데이터셋 $\mathcal{D}$로 부터 배치 샘플링
  for Batch samples **do**
    N개의 앙상블 Q_k를 [수학식 5] 손실함수를 이용하여 업데이트
    정책 네트워크 π_θ를 [수학식 7] 손실함수를 이용하여 업데이트
  end for
end for
Return π_θ, Q_k

(52) CPC특허분류
G06N 3/047 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193973
과제번호 2019-0-00079-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능대학원지원(고려대학교)
기 여 율 50/100
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711193771
과제번호 2022-0-00311-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 사람중심인공지능핵심원천기술개발
연구과제명 일상적 물건들의 다접촉 로봇 조작을 위한 목적지향강화학습 기술 개발
기 여 율 50/100
과제수행기관명 한국과학기술원
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

할인율을 코사인 엠베딩 기법을 이용하여 고차원 벡터로 전환하고, 상기 할인율과 데이터셋을 입력 데이터로 설정하는 단계; 및

상기 할인율과 상기 데이터셋을 오프라인 강화학습 알고리즘을 적용하여 Q-함수와 상기 할인율에 대응하는 예측값을 출력하도록 학습하는 단계;를 포함하는 오프라인 강화학습 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 다중 할인율을 적용한 오프라인 강화학습 방법에 관한 것이다.

배 경 기 술

[0002] 강화학습은 마르코프 결정 과정(Markov Decision Process)을 따르고, 의사결정과정을 모델링하는 수학적인 틀이다.

[0003] 마르코프 결정 과정은 의사결정 과정에 의한 다음 상태의 확률이 오로지 전 상태에만 영향을 받는 상황을 가정한다. 이러한 가정은 실제 환경에서도 잘 준수되는 편이기에, 강화학습이 실제 환경에서도 잘 적용되는 이유 중 하나이기도 하다.

[0004] 강화학습 알고리즘은 이러한 마르코프 결정과정을 따르고 있다고 가정하면, 학습하고자 하는 환경으로부터 상태 정보를 입력으로 받으면 학습 목표(goal)을 달성하기 위한 행동(action)을 출력하는 정책(policy)을 학습하는 것을 주요 목적으로 하고 있다.

[0005] 일반적으로, 강화학습의 학습 목표는 에이전트가 환경과의 상호작용을 통하여 얻는 보상(reward)의 할인된 총합(discounted sum)을 최대화하는 것으로 한다.

[0006] 가장 일반적인 오프라인 강화학습 방법은 보수적인 정책을 학습하는 방법이며, 이러한 방법들은 데이터셋 외부 분포(Out of Data distribution)의 상태-행동에 패널티를 가함으로써 정책이 데이터셋 상태-행동 분포를 따르도록 하였다.

[0007] 패널티를 가하는 방법은 크게 두가지로 나뉘는데, 첫번째로는 데이터셋 상태-행동 분포와의 분포적 거리를 추정하는 정규화 함수를 이용하는 방법이 있으며, 대표적으로 CQL(Conservative Q-learning)이 있고, 두 번째로는 앙상블(ensemble) 네트워크를 이용하여 상태-행동의 불확실성을 측정하고 이를 패널티로 가하는 방법이 있고, 대표적인 알고리즘은 SAC-N(Soft Actor-Critic-N)이 있다.

[0008] 종래의 알고리즘들은 이러한 에이전트를 학습하는 과정에서 하나의 할인율(discount factor)만을 사용하였다.

[0009] 그러나 대규모 환경에서 한정된 데이터셋으로 학습할 때 비평가 네트워크의 예측값은 필연적으로 에러를 내포하고 있으며, 이로 인해, 높은 할인 요인을 적용하여 이펙티브 호라이즌을 넓힌상황에서는 예측값이 많은 에러를 담게되어 학습에 악영향을 끼칠 수 있다.

[0010] 또한, 종래의 오프라인 강화학습을 위한 방법론들은 데이터셋이 아우르는 상태-행동 분포가 한정됨으로 인해 야기되는 불확실성(unceratainty)를 해결하는 것을 관건으로 하였다.

[0011] 오프라인 강화학습에서 Q함수는 한정된 데이터셋을 이용하여 학습하기 때문에 데이터셋이 다루고 있는 상태-행동 분포가 넓지 않을 경우, 학습된 Q 함수가 데이터셋 분포 밖 샘플(Out of distribution)에 대해서 예측하는 가치가 실제의 가치를 모사하지 못할 가능성이 높아지는 문제가 있다.

[0012] 따라서, 오프라인 강화학습 연구자들은 Q-함수가 부정적(pessimisstic)이고 보수적(conservative)으로 학습되도록 하는 방법을 연구했었다.

[0013] 여기서, 부정적인 Q 함수는 데이터셋을 통해 추정하고 있는 마르코프 의사결정 과정과 실세계와의 차이로 인해 생기는 에러를 감안하여 예상범위내의 가장 낮은 가치값을 도출하는 Q-함수를 의미하며, 보수적인 Q 함수는 데이터셋 분포내의 상태-행동에 대한 가치를 높게 보는 Q-함수를 의미한다.

[0014] Q-함수의 부정성(pessimism)과 보수성(conservatism)을 올리기 위해, 기존의 오프라인 강화학습 알고리즘은 학습을 위한 목표함수에 정규화(regularization) 기법을 적용해 왔었다.

[0015] 다만, 이러한 기법을 적용할 경우 학습을 통해 최적화한 정책의 최적성(optimality) 손실이 수반된다는 단점이 있으며, 이 단점을 완화시키기 위한 연구가 활발히 이루어져 왔다.

[0016] 이러한 연구 흐름에서, 오프라인 강화학습 연구에서 낮은 할인율이 Q함수의 부정성과 보수성을 조절할 수 있음을 보이는 논문이 발표되었고, 이와 관련된 연구들이 많은 주목을 받았다.

[0017] 이 연구들은 결과적으로 기존의 많은 연구들에서 일상적으로 적용되었던 높은 값의 할인율(0.99)보다 낮은 할인율을 적용함으로써, 정규화항 없이도 충분히 부정적이고 보수적이며, 최적성 손실이 덜한 에이전트를 학습할 수 있다는 것을 강조하였으나, 낮은 할인율의 장점을 활용하기 위한 최적화된 학습방법에 대해서는 아직 제시된 방안이 없다는 한계가 있다.

선행기술문헌

특허문헌

[0018] (특허문헌 0001) 대한민국 등록특허공보 제10-1990326호("감가율 자동 조정 방식의 강화 학습 방법", 공개일자: 2019.06.18)

발명의 내용

해결하려는 과제

[0019] 본 발명은 한정된 데이터셋을 이용하여 학습하는 오프라인 강화학습 알고리즘이 다양한 할인율에 대해서 동시에 학습할 수 있도록 하고, 다양한 할인 요인 별로 계산하며, 다양한 할인 요인을 종합적으로 최적화할 수 있는 에이전트를 만드는 오프라인 강화학습 방법을 제공하고자 한다.

과제의 해결 수단

[0020] 본 발명의 일 기술적 측면은 할인율을 코사인 엠베딩 기법을 이용하여 고차원 벡터로 전환하고, 상기 할인율과 데이터셋을 입력 데이터로 설정하는 단계; 및 상기 할인율과 상기 데이터셋을 오프라인 강화학습 알고리즘을 적용하여 Q-함수와 상기 할인율에 대응하는 예측값을 출력하도록 학습하는 단계;를 포함한다.

발명의 효과

[0021] 본 발명의 일 실시형태에 따르면, 높은 상태 행동 space를 가지는 환경에서 더욱더 강건한 강화학습 에이전트를 학습할 수 있고, 종래의 기술들에 비해 적은수의 앙상블 네트워크를 사용하더라도 좋은 성능을 낼 수 있다.

[0022] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0023] 도 1은 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 오프라인 강화학습 방법을 설명하기 위한 순서도이다.

도 3은 하이퍼파라미터(Hyperparameter)를 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘으로 학습한 평가결과를 SAC-N과 비교한 결과를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0024] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0025] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.

[0026] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0027] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에서 사용되는 "부", "모듈", "유닛" 등의 용어는 적어도 하나의 기능 또는 동작을 처리하는 단위를 의미하며, 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소, 또는 소프트웨어와 하드웨어의 결합으로 구현될 수 있다. 그렇지만 "부", "모듈", "유닛" 등의 용어가 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부", "모듈", "유닛" 등은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 ""부", "모듈", "유닛" 등의 용어는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.

[0028] 아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략한다.

[0029] "제1", "제2" 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 구성 요소들은 용어들에 의해 한정되지는 않는다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는" 이라는 용어는 복수의 관련된 항목들의 조합 또는 복수의 관련된 항목들 중의 어느 하나의 항목을 포함한다.

[0031] 이하, 도면을 참조하여 본 발명의 일 실시예에 따른 다중 할인율을 적용한 오프라인 강화학습 방법을 설명하도록 한다.

[0032] 도 1은 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘을 설명하기 위한 도면이고, 도 2는 본 발명의 일 실시예에 따른 오프라인 강화학습 방법을 설명하기 위한 순서도이다.

[0033] 본 발명의 일 실시예에 따른 오프라인 강화학습 과정에서는 종래의 오프라인 강화학습 알고리즘이 단일 할인율만을 사용하여 학습하던 것과는 다르게, 복수의 할인율을 선정하고, 각각에 대한 Q-함수의 예측값 및 정책의 행동 예측값을 도출할 수 있도록 네트워크를 구성하여 학습시키는 오프라인 강화학습 알고리즘을 사용한다.

[0034] 본 발명의 일 실시예에 따른 오프라인 강화학습 방법은 보상(reward)의 할인된 총합 최대화를 환경과의 추가 상호작용 없이 사전에 확보되어 있는 한정된 수량의 데이터셋만을 사용할 수 있다.

[0035] 본 발명의 일 실시예에 따른 오프라인 강화학습 방법은 데이터셋의 상태에서 특정 행동을 실행하였을 때 미래에 예상되는 할인된 총합을 예측하는 네트워크인 행동-가치 함수(action value function)에 해당하는 Q 함수(Q function)를 학습한다.

[0036] 오프라인 강화학습 방법은 Q함수를 사용하여 배우 네트워크가 데이터셋의 상태-행동 쌍 중에서 미래 가치가 높은 행동을 취한 데이터에 대해서 높은 확률을 부여할 수 있도록 해준다.

[0037] 도 1에서 도시한 바와 같이, 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘은 학습에 사용할 복수의 할인율을 설정할 때 할인율의 유효 지평선(effective horizon)값이 동일 간격을 가지도록 설정한다.

[0038] 이때, 유효지평선은 γ를 할인율으로 표현할 때, $\frac{1}{1-\gamma}$ 값을 의미하며, 이 값은 할인율에 의해 계산되는 할인된 보상합이 평균적으로 반영하는 미래 시간의 길이를 나타낸다.

[0039] 할인율을 입력 데이터로써 취급할 때, 하나의 실수값에 해당하는 할인율을 그대로 입력값으로 사용하면 네트워크의 고차원적인 추론이 어려워지므로, 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘에서는 할인율에 다음의 수학식 1로 표현되는 코사인 엠베딩(cosine embedding)기법을 사용하여 고차원 벡터로 전환하여 입력데이터로 설정한다.

수학식 1

[0040] $$\text{Embedding}(\gamma) = \left[\cos(\frac{\pi}{1-\gamma}), \cos(\frac{2\pi}{1-\gamma}), ..., \cos(\frac{N\pi}{1-\gamma})\right]$$

[0041] 여기서, N은 엠베딩 벡터의 차원수를 나타낸다.

[0042] 즉, 본 발명의 일 실시예에 따른 오프라인 강화학습 방법에서는 과 데이터셋과 고차원 벡터로 전환된 할인율을 입력데이터로 설정할 수 있다.

[0043] 도 2에서 도시한 바와 같이, 본 발명의 일 실시예에 따른 오프라인 강화학습 방법은 데이터셋과 할인율을 기초로 변경된 네트워크 구조를 이용하여 Q-함수와 행위자 네트워크가 여러 개의 할인율에 대응되는 예측값을 출력할 수 있도록 하는 학습한다.

[0044] 기존 네트워크 구조는 데이터셋 샘플만을 입력데이터로 받아서 여러 은닉층(Hidden layer)을 거친 후 나온 출력값을 사용하고, 출력값을 이용하여 단일 할인율이 적용된 Q-함수의 손실함수를 계산한 후 손실함수를 최소화하는 방향으로 업데이트를 진행한다.

[0045] 손실함수와 관련하여, 본 발명의 일 실시예에 따른 오프라인 강화학습 방법은 데이터셋 외부분포의 상태-행동에 대한 패널티를 가하기 위해 앙상블 네트워크 기법을 적용한 SAC-N 알고리즘을 벤치마크한다.

[0046] SAC-N 알고리즘에서는 앙상블 네트워크중 하나의 Q 함수의 할인율 에 대한 손실함수를 다음의 수학식 2와 같이 연산할 수 있다.

수학식 2

[0047] $$Loss(Q_k, \gamma) = \mathbb{E}_{s,a,s' \sim D}\left[\frac{1}{2}[Q_k(s,a) - \left(r(s,a) + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}(\min_i Q_i(s',a') - \alpha \log \pi(a'|s'))\right)\right]^2$$

[0048] 따라서, 기존 알고리즘들은 다음의 수학식 3과 같이 전체 손실함수를 연산하여 업데이트를 진행한다.

수학식 3

[0049] $$\text{Total Loss} = \frac{1}{M}\sum_{k}^{M} Loss(Q_k, \gamma)$$

[0050] 여기서, M은 전체 앙상블의 개수를 나타낸다.

[0051] 본 발명의 일 실시예에 따른 오프라인 강화학습 방법에서는 할인율을 입력데이터로 사용할 수 있으므로, 상기 수학식 1에 대응하는 손실함수를 다음의 수학식 4를 이용하여 연산할 수 있다.

수학식 4

[0052] $$Loss(Q_k,\gamma)=\mathbb{E}_{s,a,s'\sim D}\left[\frac{1}{2}[Q_k(s,a,\gamma)-\left(r(s,a)+\gamma\mathbb{E}_{a'\sim\pi(\cdot|s',\gamma)}(\min_i Q_i(s',a',\gamma)-\alpha\log\pi(a'|s',\gamma)\right)\right]$$

[0053] 이와 같은 방법으로 연산된 손실함수들을 이용하여 앙상블 네트워크 수와 할인율의 개수(N)에 대하여 평균을 구한 것을 전체 손실함수로 정의할 수 있다.

[0054] 전체 손실함수는 다음의 수학식 5와 같이 나타낼 수 있다.

수학식 5

[0055] $$\text{Total Loss}=\frac{1}{MN}\sum_k^M\sum_i^N Loss(Q_k,\gamma_i)$$

[0056] 또한, 본 발명의 일 실시예에 따른 오프라인 강화학습 방법은 손실함수들을 이용하여 학습한 Q-함수를 통해 에이전트의 정책을 학습할 수 있다.

[0057] 정책 손실함수은 기존의 알고리즘인 SAC-N을 벤치마크한 다음의 수학식 6을 사용하여 연산할 수 있다.

수학식 6

[0058] $$Loss(\pi,\gamma_i)=\mathbb{E}_{\substack{a\sim\pi(\cdot|s,\gamma)\\ s\sim D}}\left[\alpha\log(a|s,\gamma)-\min_i Q_i(s,a,\gamma))\right]$$

[0059] 여기서, 정책은 Q-함수와 다르게 앙상블 기법을 적용하지 않았다.

[0060] 따라서, 최종 정책 손실함수는 할인율들에 대한 평균으로만 표현된 다음의 수학식 7로 나타낼 수 있다.

수학식 7

[0061] $$TotalLoss=\frac{1}{N}\sum_i^N Loss(\pi,\gamma_i)$$

[0062] 다시 도 1을 참조하면, 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘에서는 상기 연산된 손실함수들을 사용하여 학습한다.

[0063] 오프라인 강화학습 알고리즘은 매 iteration마다 손실함수를 계산하기 위해 데이터셋으로부터 배치 샘플링을 진행하고 이 샘플들에 대한 손실함수의 평균을 구함으로써, 상기 수학식 4 및 수학식 6의 데이터셋에 대한 평균값을 추정한다.

[0065] 이하, 도면을 참조하여 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘의 테스트를 설명하도록 한다

[0066] 도 3은 하이퍼파라미터(Hyperparameter)를 설명하기 위한 도면이고, 도 4는 본 발명의 일 실시예에 따른 오프라인 강화학습 알고리즘으로 학습한 평가결과를 SAC-N과 비교한 결과를 설명하기 위한 도면이다.

[0067] 본 발명의 일 실시예에서는 알고리즘 테스트를 강화학습 벤치마크로서, 널리 사용되고 있는 MuJoCo-벤치마크의 환경중 HalfCheetah/Hopper/Walker2d 세가지 환경에서 진행하였다.

[0068] 도 3에서 도시한 바와 같이, 환경별로 에이전트에게 요구되는 보수성의 정도가 다르기 때문에, 앙상블 개수는

다르게 설정하였고, 할인율은 0.8부터 0.99까지 유효지평선이 5가 차이나도록 계산하여 총 20개의 할인율에 대한 학습이 되도록 하였다.

[0069] 도 3에서 결합 전 층수는 도 2에서 표현한 데이터셋 샘플 및 할인율 엠베딩이 각각 통과하는 네트워크 은닉층 층수를 의미하며, 두가지 은닉층 모두 동일한 층수로 설정하였다. 동일한 흐름으로, 결합 후 층수는 출력값 결합 후의 층수를 의미한다.

[0070] 도 4에서 도시한 바와 같이, 본 발명의 일 실시예에서 제공한 강화학습 알고리즘으로 MuJoCo 벤치마크에서 학습한 평가결과를 기존 알고리즘에 해당하는 SAC-N과 비교한 결과를 나타낸 것으로, 세로축의 값은 알고리즘이 에피소드당 평균적으로 얻는 보상합에 해당하는 성능수치를 MuJoCo 벤치마크에서 제공하는 전문가 평균성능-무작위 정책 평균성능으로 나눠서 얻은 백분율 값을 의미하며, 가로축 값은 20개의 할인율이고 오른쪽일수록 높은값을 의미한다.

[0071] SAC-N은 적용할 수 없는 부분을 제외하면, 도 3에서 제시한 하이퍼파라미터와 동일하게 적용하여 학습하였고, 할인율 20개에 대해서 각각 학습을 하여 성능치를 측정하였다.

[0072] 도 3의 주황색 그래프는 본 발명에서 제시한 알고리즘의 성능 트렌드이고, 파란색 그래프는 기존 알고리즘에 해당하는 SAC-N의 성능 트렌드이다. 도 3의 결과에 따르면, 본 발명의 알고리즘이 모든 할인율값에 대해서 기존 알고리즘에 비해 개선된 성능을 보여주고 있으며 그중에서 최적의 성능을 보이는 할인율을 선택할 수 있는 자유도 또한 가지고 있기 때문에 기존 알고리즘에 비해 우월한 성능을 보일 수 있음을 확인하였다.

[0074] 이상에서 설명한 바와 같이, 높은 상태 행동 space를 가지는 환경에서 더욱더 강건한 강화학습 에이전트를 학습할 수 있고, 종래의 기술들에 비해 적은수의 앙상블 네트워크를 사용하더라도 좋은 성능을 낼 수 있다.

[0076] 본원 발명의 실시예들과 관련된 기술 분야에서 통상의 지식을 가진 자는 상기 기재의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로, 개시된 방법들은 한정적인 관점이 아닌 설명적 관점에서 고려되어야 한다. 본 발명의 범위는 발명의 상세한 설명이 아닌 특허청구 범위에 나타나며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

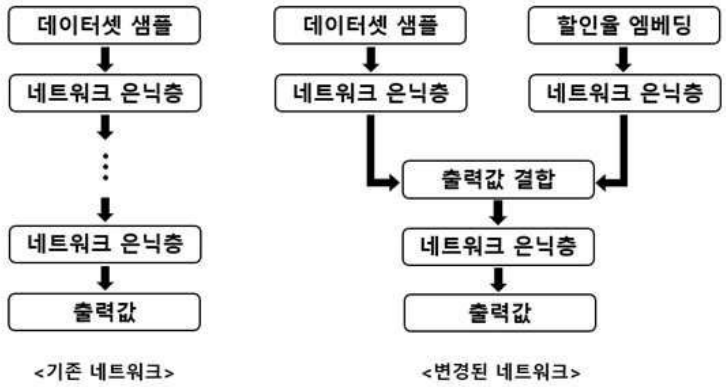
도면1

Algorithm 1 학습 알고리즘

Require: 데이터셋 $\mathcal{D}$, M개의 앙상블 Q-함수 네트워크 Q_k, 정책 네트워크 π_θ, 배치 사이즈 B.

```
할인율 리스트 Γ : 사전에 정해진 최소/최대 할인율과 유효 지평선 차이에 따라서 생성
할인율에 코사인 엠베딩 적용
모든 네트워크 초기화
for iteration t= 1, 2, ... n do
    B 크기만큼 데이터셋 D로 부터 배치 샘플링
    for Batch samples do
        N개의 앙상블 Q_k를 [수학식 5] 손실함수를 이용하여 업데이트
        정책 네트워크 π_θ를 [수학식 7] 손실함수를 이용하여 업데이트
    end for
end for
Return π_θ, Q_k
```

도면2



도면3

Hyperparameter		Value
앙상블 개수	Halfcheetah	10
	Hopper	50
	Walker2d	20
할인율 설정	최솟값	0.8
	최댓값	0.99
	유효 지평선 차잇값	5
코사인 엠베딩 차원수		32
배치 샘플수		256
네트워크 학습율(learning rate)		3e-4
은닉층 크기(hidden size)		256
은닉층 층수	결합 전 층수	2
	결합 후 층수	2
전체 iteration		3e6

도면4

