

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0014744

(43) 공개일자 2022년02월07일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)

(21) 출원번호 10-2020-0094758

(22) 출원일자 2020년07월29일

심사청구일자 2021년07월02일

(71) 출원인
서강대학교산학협력단

서울특별시 마포구 백범로 35 (신수동, 서강대학교)

(72) 발명자

최서린

전라남도 순천시 삼산로 103-13 현대아파트 109동 903호

양지훈

서울특별시 은평구 백련산로2길 19 백련산힐스테이트1차 105동 603호

(74) 대리인

이지연

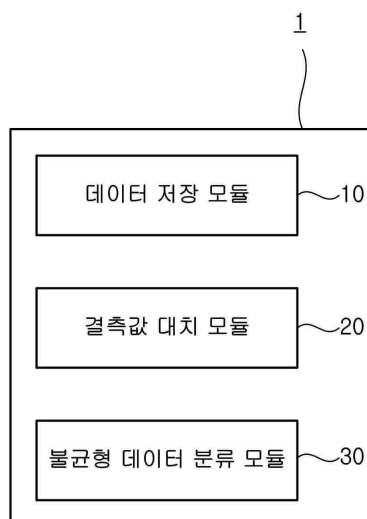
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 강화 학습을 기반으로 한 데이터 전처리 시스템 및 방법

(57) 요약

본 발명은 강화 학습을 이용한 데이터 전처리 장치 및 방법에 관한 것이다. 상기 데이터 전처리 장치는, 컨트롤러 모델과 차일드 모델을 구비하여, 결측값을 대치하는 결측값 대치 모듈; 및 DQN(Deep Q-Network) 알고리즘을 이용한 강화 학습을 통해 불균형 데이터를 분류하는 불균형 데이터 분류 모듈;을 구비한다. 상기 결측값 대치 모듈은, 컨트롤러 모델에 의해 결측값에 대치할 값들의 벡터를 샘플링하고, 차일드 모델을 통해 상기 샘플링된 값들을 결측값에 대치시킨 후 성능을 계산하고, 상기 차일드 모델에 의해 계산된 보상값을 기반으로 하여 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 차일드 모델의 정확도가 높아지도록 상기 컨트롤러 모델을 강화 학습시키게 된다. 상기 불균형 데이터 분류 모듈은 DQN 알고리즘을 통하여 각 클래스 샘플 수에 따라 보상을 줌으로써 소수 범주에 있는 데이터가 더 잘 분류될 수 있게 된다.

대표도 - 도1



명세서

청구범위

청구항 1

강화 학습을 이용한 데이터 전처리 장치에 있어서,
컨트롤러 모델과 차일드 모델을 구비하여, 결측값을 대치하는 결측값 대치 모듈;을 포함하며,
상기 컨트롤러 모델은 결측값에 대치할 값들의 벡터를 샘플링해주는 것을 특징으로 하며,
상기 차일드 모델은 상기 컨트롤러 모델에 의해 샘플링된 값들을 결측값에 대치시킨 후 성능을 계산하는 것을 특징으로 하며,
상기 차일드 모델에 의해 계산된 성능을 이용하여 상기 컨트롤러 모델을 학습시키는 것을 특징으로 하는 강화 학습을 이용한 데이터 전처리 장치.

청구항 2

제1항에 있어서,
상기 컨트롤러 모델에 의해 샘플링된 벡터는 결측값들에 대치될 값들의 분포의 평균으로 이루어진 벡터와 표준편차로 이루어진 벡터인 것을 특징으로 하는 데이터 전처리 장치.

청구항 3

제1항에 있어서, 상기 컨트롤러 모델은,
데이터의 결측값에 이전 액션의 값인 대치값을 채워 넣은 후 1차원 배열로 변환시킨 벡터가 입력되며, 입력된 데이터들에 대하여 학습하여 결측값에 대치할 값들을 샘플링하는 신경망 구조; 및
상기 신경망 구조로부터 제공된 샘플링된 값들에 대하여 Fully-Connected layer를 거쳐 가우시안 분포의 평균(μ)과 표준편차(σ)를 생성하여 제공하는 연속적인 액션 공간에서의 정책 생성 네트워크;
를 구비하는 것을 특징으로 하는 데이터 전처리 장치.

청구항 4

제1항에 있어서, 상기 차일드 모델은,
결측값이 대치된 데이터들 중 학습 데이터로 학습시키고 시험 데이터로 정확도를 계산하며,
현재 상태에서 계산된 정확도와 이전 상태에서 계산된 정확도의 차이를 보상값으로 정의하는 것을 특징으로 하며,
상기 컨트롤러 모델은 상기 차일드 모델에 의해 계산된 보상값을 기반으로 하여 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 강화 학습시키는 것을 특징으로 하는 데이터 전처리 장치.

청구항 5

제1항에 있어서, 상기 컨트롤러 모델은 PPO 알고리즘을 이용한 정책 기반 강화 학습을 통해 업데이트시키는 것을 특징으로 하며,
상기 정책 기반 강화 학습의 목표는 정책을 근사하는 정책 함수($J(\theta)$)를 최대화시키는 정책 함수 파라미터(θ)를 찾는 것을 특징으로 하는 데이터 전처리 장치.

청구항 6

제1항에 있어서, 상기 데이터 전처리 장치는,

DQN(Deep Q-Network) 알고리즘을 이용한 강화 학습을 통해 불균형 데이터를 분류하는 불균형 데이터 분류 모듈;을 더 구비하고,

상기 DQN 알고리즘은 CNN(Convolution Neural Network)을 이용하여 Q 함수를 학습하는 것을 특징으로 하

상기 Q 함수는 강화 학습에서 상태-액션 조합의 질을 계산하는 함수인 것을 특징으로 하는 데이터 전처리 장치.

청구항 7

제6항에 있어서, 상기 DQN 알고리즘은,

입력 데이터가 들어오면 입력 데이터가 속하는 클래스를 예측하고, 옳게 예측하면 양의 보상을 주고 옳지 않게 예측하면 음의 보상을 주는 것을 특징으로 하며,

보상을 주기 위하여, 입력 데이터마다 클래스별 샘플수를 계산하고, 소수 클래스는 전체 데이터수를 가장 많은 클래스별 샘플수로 나누어준 값을 보상으로 주고, 다수 클래스는 전체 데이터수를 가장 작은 클래스별 샘플수로 나누어준 값을 보상으로 줌으로써, 소수 범주에 있는 데이터가 더 잘 분류되도록 구성된 것을 특징으로 하는 데이터 전처리 장치.

청구항 8

강화 학습을 이용한 데이터 전처리 방법에 있어서,

(a1) 컨트롤러 모델에 의해 결측값에 대치할 값들의 벡터를 샘플링해주는 단계;

(a2) 차일드 모델에 의해 상기 컨트롤러 모델에 의해 샘플링된 값들을 결측값에 대치시킨 후 성능을 계산하는 단계;

(a3) 상기 차일드 모델에 의해 계산된 성능을 이용하여 상기 컨트롤러 모델을 학습시키는 단계;

를 포함하여, 결측값을 대치하는 결측값을 대치하는 것을 특징으로 하는 강화 학습을 이용한 데이터 전처리 방법.

청구항 9

제8항에 있어서, 상기 (a1) 단계는,

데이터의 결측값에 이전 액션의 값인 대치값을 채워 넣은 후 1차원 배열로 변환시킨 벡터가 신경망 구조로 입력되고,

신경망 구조로 입력된 데이터들에 대하여 학습하여 결측값에 대치할 값들을 샘플링하고,

상기 신경망 구조로부터 제공된 샘플링된 값들에 대하여 Fully-Connected layer를 거쳐 가우시안 분포의 평균(μ)과 표준편차(σ)를 생성하여 제공하는 것을 특징으로 하는 데이터 전처리 방법.

청구항 10

제8항에 있어서, 상기 (a2) 단계는,

결측값이 대치된 데이터들 중 학습 데이터로 학습시키고 시험 데이터로 정확도를 계산하며,

현재 상태에서 계산된 정확도와 이전 상태에서 계산된 정확도의 차이를 보상값으로 정의하는 것을 특징으로 하며,

상기 컨트롤러 모델은 상기 차일드 모델에 의해 계산된 보상값을 기반으로 하여 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 강화 학습시키는 것을 특징으로 하는 데이터 전처리 방법.

청구항 11

제8항에 있어서, 상기 데이터 전처리 방법은,

(b) 결측값이 대치된 데이터들에 대하여, DQN(Deep Q-Network) 알고리즘을 이용한 강화 학습을 통해 불균형 데이터를 분류하는 단계;를 더 구비하고,

상기 DQN 알고리즘은 CNN(Convolution Neural Network)을 이용하여 Q 함수를 학습하는 것을 특징으로 하며,
상기 Q 함수는 강화 학습에서 상태-액션 조합의 질을 계산하는 함수인 것을 특징으로 하는 데이터 전처리 방법.

청구항 12

제11항에 있어서, 상기 DQN 알고리즘은,

입력 데이터가 들어오면 입력 데이터가 속하는 클래스를 예측하고, 옳게 예측하면 양의 보상을 주고 옳지 않게 예측하면 음의 보상을 주는 것을 특징으로 하며,

보상을 주기 위하여, 입력 데이터마다 클래스별 샘플수를 계산하고, 소수 클래스는 전체 데이터수를 가장 많은 클래스별 샘플수로 나누어준 값을 보상으로 주고, 다수 클래스는 전체 데이터수를 가장 작은 클래스별 샘플수로 나누어준 값을 보상으로 줌으로써, 소수 범주에 있는 데이터가 더 잘 분류되도록 구성된 것을 특징으로 하는 데이터 전처리 방법.

발명의 설명

기술 분야

[0001] 본 발명은 강화 학습을 기반으로 한 데이터 전처리 시스템 및 방법에 관한 것으로서, 더욱 구체적으로는 PPO(Proximal Policy Optimization) 알고리즘을 이용한 강화 학습을 통해 학습 데이터의 결측값에 대한 대처값을 산출하고, DQN(Deep Q-Network) 알고리즘을 이용하여 불균형 데이터를 분류하여 학습의 성능을 향상시킬 수 있도록 하는 강화학습을 기반으로 한 데이터 전처리 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 4차 산업혁명이 대두된 이래, 데이터에 알고리즘을 적용해 분석하는 방식은 여러 분야에서 사용되고 있다. 하지만 많은 수의 데이터들은 데이터가 생성되면서 값이 누락되는 경우가 빈번하다. 이렇게 누락된 데이터를 '결측치'라 한다. 이와 같이, 결측치가 많이 존재하는 경우 이에 대한 제대로 된 분석이 어려울 수 밖에 없다.

[0003] 데이터가 누락되어 결측값이 존재하는 데이터의 경우, 정확한 데이터 분석을 위하여, 데이터 분석의 전처리 단계에서 이를 처리하거나 대체시킨다. 결측값 처리 방법으로는, 대표적으로 결측값 자체를 제거하는 방법, 평균값, 중앙값, 최빈값 등으로 대체하는 통계적 방법과 K-최근접 이웃 탐색(K-Nearest Neighbor)를 활용하는 기계 학습 방법이 사용된다.

[0004] 한편, 기계학습 실험을 진행하기 위한 학습 데이터로는 범주 별로 데이터의 수가 비슷하게 구성되어 있는 것을 사용하는 것이 일반적이다. 그러나, 많은 실제 데이터들은 클래스 별로 데이터의 수가 비슷하지 않기 때문에 데이터 불균형 문제를 갖게 되며, 이와 같이 데이터 불균형 문제가 있는 경우, 소수 범주에 속한 데이터들은 오 분류되어 성능이 낮게 나올 가능성이 높다. 이는 기계학습 알고리즘의 설계 특성상 전반적인 성능을 최적화시키려 하기 때문에 발생하는 것이다.

[0005] 불균형 데이터 분류 문제를 해결하기 위한 대표적인 방법으로는 데이터 수준 접근 방법이 있다. 데이터 수준의 접근 방법에는 언더샘플링(Under-Sampling), 오버샘플링(Over-Sampling)이나 이들을 동시에 사용하여 데이터의 균형을 조정하는 방법이 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국등록특허공보 제 10-2093079호

(특허문헌 0002) 한국공개특허공보 제 10-2020-0027834호

발명의 내용

해결하려는 과제

[0007] 전술한 문제점을 해결하기 위한 본 발명의 목적은 PPO 알고리즘을 이용한 강화 학습을 통해 결측치에 대한 대치값을 추출하고, 대치값을 추출한 후 DQN 알고리즘을 통해 다중 클래스에 대한 분류를 수행하여 성능을 향상시킬 수 있도록 하는 PPO와 DQN을 이용한 강화학습을 기반으로 한 데이터 전처리 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0008] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치는, 컨트롤러 모델과 차일드 모델을 구비하여, 결측값을 대치하는 결측값 대치 모듈;을 포함하며,

[0009] 상기 컨트롤러 모델은 결측값에 대치할 값들의 벡터를 샘플링해주는 것을 특징으로 하며, 상기 차일드 모델은 상기 컨트롤러 모델에 의해 샘플링된 값들을 결측값에 대치시킨 후 성능을 계산하는 것을 특징으로 하며, 상기 차일드 모델에 의해 계산된 성능을 이용하여 상기 컨트롤러 모델을 학습시키도록 구성된다.

[0010] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, 상기 컨트롤러 모델에 의해 샘플링된 벡터는 결측값들에 대치될 값들의 분포의 평균으로 이루어진 벡터와 표준 편차로 이루어진 벡터인 것이 바람직하다.

[0011] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, 상기 컨트롤러 모델은, 데이터의 결측값에 이전 액션의 값인 대치값을 채워 넣은 후 1차원 배열로 변환시킨 벡터가 입력되며, 입력된 데이터들에 대하여 학습하여 결측값에 대치할 값들을 샘플링하는 신경망 구조; 및 상기 신경망 구조로부터 제공된 샘플링된 값들에 대하여 Fully-Connected layer를 거쳐 가우시안 분포의 평균(μ)과 표준편차(σ)를 생성하여 제공하는 연속적인 액션 공간에서의 정책 생성 네트워크;를 구비하는 것이 바람직하다.

[0012] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, 상기 차일드 모델은, 결측값이 대치된 데이터들 중 학습 데이터로 학습시키고 시험 데이터로 정확도를 계산하며, 현재 상태에서 계산된 정확도와 이전 상태에서 계산된 정확도의 차이를 보상값으로 정의하는 것을 특징으로 하며, 상기 컨트롤러 모델은 상기 차일드 모델에 의해 계산된 보상값을 기반으로 하여 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 강화 학습시키는 것이 바람직하다.

[0013] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, 상기 컨트롤러 모델은 PPO 알고리즘을 이용한 정책 기반 강화 학습을 통해 업데이트시키는 것을 특징으로 하며, 상기 정책 기반 강화 학습의 목표는 정책을 근사하는 정책 함수($J(\theta)$)를 최대화시키는 정책 함수 파라미터(θ)를 찾는 것이 바람직하다.

[0014] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, DQN(Deep Q-Network) 알고리즘을 이용한 강화 학습을 통해 불균형 데이터를 분류하는 불균형 데이터 분류 모듈;을 더 구비하고, 상기 DQN 알고리즘은 CNN(Convolution Neural Network)을 이용하여 Q 함수를 학습하는 것을 특징으로 하며, 상기 Q 함수는 강화 학습에서 상태-액션 조합의 질을 계산하는 함수인 것이 바람직하다.

[0015] 전술한 제1 특징에 따른 강화 학습을 이용한 데이터 전처리 장치에 있어서, 상기 DQN 알고리즘은, 입력 데이터가 들어오면 입력 데이터가 속하는 클래스를 예측하고, 옳게 예측하면 양의 보상을 주고 옳지 않게 예측하면 음의 보상을 주는 것을 특징으로 하며, 보상을 주기 위하여, 입력 데이터마다 클래스별 샘플수를 계산하고, 소수 클래스는 전체 데이터수를 가장 많은 클래스별 샘플수로 나누어준 값을 보상으로 주고, 다수 클래스는 전체 데이터수를 가장 작은 클래스별 샘플수로 나누어준 값을 보상으로 줌으로써, 소수 범주에 있는 데이터가 더 잘 분류되도록 구성된 것이 바람직하다.

[0016] 본 발명의 제2 특징에 따른 강화 학습을 이용한 데이터 전처리 방법은, (a1) 컨트롤러 모델에 의해 결측값에 대치할 값들의 벡터를 샘플링해주는 단계; (a2) 차일드 모델에 의해 상기 컨트롤러 모델에 의해 샘플링된 값들을 결측값에 대치시킨 후 성능을 계산하는 단계; (a3) 상기 차일드 모델에 의해 계산된 성능을 이용하여 상기 컨트롤러 모델을 학습시키는 단계;를 포함하여, 결측값을 대치한다.

[0017] 전술한 제2 특징에 따른 강화 학습을 이용한 데이터 전처리 방법에 있어서, 상기 (a1) 단계는, 데이터의 결측값에 이전 액션의 값인 대치값을 채워 넣은 후 1차원 배열로 변환시킨 벡터가 신경망 구조로 입력되고, 신경망 구조로 입력된 데이터들에 대하여 학습하여 결측값에 대치할 값들을 샘플링하고, 상기 신경망 구조로부터 제공된

샘플링된 값들에 대하여 Fully-Connected layer를 거쳐 가우시안 분포의 평균(μ)과 표준편차(σ)를 생성하여 제공하는 것이 바람직하다.

[0018] 전술한 제2 특징에 따른 강화 학습을 이용한 데이터 전처리 방법에 있어서, 상기 (a2) 단계는, 결측값이 대치된 데이터들 중 학습 데이터로 학습시키고 시험 데이터로 정확도를 계산하며, 현재 상태에서 계산된 정확도와 이전 상태에서 계산된 정확도의 차이를 보상값으로 정의하는 것을 특징으로 하며, 상기 컨트롤러 모델은 상기 차일드 모델에 의해 계산된 보상값을 기반으로 하여 PPO(Proximal Policy Optimization) 알고리즘을 적용하여 강화 학습시키는 것이 바람직하다.

[0019] 전술한 제2 특징에 따른 강화 학습을 이용한 데이터 전처리 방법에 있어서, (b) 결측값이 대치된 데이터들에 대하여, DQN(Deep Q-Network) 알고리즘을 이용한 강화 학습을 통해 불균형 데이터를 분류하는 단계;를 더 구비하고, 상기 DQN 알고리즘은 CNN(Convolution Neural Network)을 이용하여 Q 함수를 학습하며, 상기 Q 함수는 강화 학습에서 상태-액션 조합의 질을 계산하는 함수인 것이 바람직하다.

[0020] 전술한 제2 특징에 따른 강화 학습을 이용한 데이터 전처리 방법에 있어서, 상기 DQN 알고리즘은, 입력 데이터가 들어오면 입력 데이터가 속하는 클래스를 예측하고, 옳게 예측하면 양의 보상을 주고 옳지 않게 예측하면 음의 보상을 주는 것을 특징으로 하며, 보상을 주기 위하여, 입력 데이터마다 클래스별 샘플수를 계산하고, 소수 클래스는 전체 데이터수를 가장 많은 클래스별 샘플수로 나누어준 값을 보상으로 주고, 다수 클래스는 전체 데이터수를 가장 작은 클래스별 샘플수로 나누어준 값을 보상으로 줌으로써, 소수 범주에 있는 데이터가 더 잘 분류되도록 구성된 것이 바람직하다.

발명의 효과

[0021] 본 발명에 따른 데이터 전처리 장치는 결측값에 알맞은 대치값을 찾고 데이터 불균형 문제를 해결하기 위하여 강화 학습을 사용함으로써, 분류의 정확성을 향상시킬 수 있게 된다.

[0022] 본 발명에 따른 데이터 전처리 장치의 결측값 대치 모듈은, 결측값 대치를 위하여 컨트롤러 모델이 결측치에 알맞은 대치값을 추출하도록 구성되고, 컨트롤러 모델에서 나온 대치값을 이용하여 차일드 모델이 성능을 평가하도록 구성된다. 차일드 모델의 성능을 측정하고, 현재 상태의 성능과 이전 상태의 성능의 차이를 기반으로 하여 보상값을 계산해서, PPO 알고리즘을 통해 차일드 모델의 정확도가 높아지도록 컨트롤러 모델을 학습시킴으로써, 더 적절한 대치값을 산출할 수 있게 된다.

[0023] 또한, 본 발명에 따른 데이터 전처리 장치의 상기 불균형 데이터 분류 모듈은, 상태는 데이터 각각을 의미하고 행동은 입력으로 들어오는 데이터가 어떤 클래스에 속하는지 분류하는 것으로 정의하고, 보상은 각 클래스에 속하는 샘플의 수에 따라 계산함으로써, DQN 알고리즘을 통해 각 클래스 샘플수에 따라 보상을 줘서 소수 범주에 있는 데이터가 더 잘 분류될 수 있도록 구성된다. 특히, 보상은 전체 데이터 수에 대하여 소수 클래스에는 전체 데이터 수에서 다수 클래스의 샘플수를 나눈 값으로 정의되고, 다수 클래스에는 전체 데이터 수에서 소수 클래스의 샘플수를 나눈 값으로 정의된다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 바람직한 실시예에 따른 강화 학습을 기반으로 한 데이터 전처리 장치의 구성을 개략적으로 도시한 블록도이다.

도 2는 강화 학습의 구조를 도시한 것이다.

도 3은 오토인퓨테이션 구조를 도시한 것이다.

도 4는 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, 결측값 대치 모듈(10)의 컨트롤러 모델을 도시한 구조도이다.

도 5는 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, 결측값 대치 모듈(10)의 컨트롤러 모델의 연속적인 액션 공간에서의 정책 생성 네트워크를 도시한 구조도이다.

도 6은 Clip의 효과를 설명하기 위하여 도시한 그래프들이다.

도 7은 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, PPO 알고리즘을 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 강화학습을 기반으로 한 데이터 전처리 장치 및 방법에 대하여 구체적으로 설명한다.
- [0026] 도 1은 본 발명의 바람직한 실시예에 따른 강화 학습을 기반으로 한 데이터 전처리 장치의 구성을 개략적으로 도시한 블록도이다.
- [0027] 도 1을 참조하면, 본 발명에 따른 데이터 전처리 장치(1)는 데이터 저장을 위한 데이터 저장 모듈(10), 결측값 대치 모듈(20) 및 불균형 데이터 분류 모듈(30)을 구비한다. 상기 데이터 저장 모듈(10)은 사전 준비된 학습 데이터들이 저장 및 관리된다. 상기 결측값 대치 모듈(20)은 PPO 알고리즘을 적용한 강화 학습을 통해 데이터들의 결측값에 대한 대치값을 찾고, 상기 불균형 데이터 분류 모듈(30)은 DQN 알고리즘을 이용하여 불균형 데이터를 분류한다. 본 발명에 따른 데이터 전처리 장치는 컴퓨터 등에 의해 구성되며, 데이터 전처리 장치를 구성하는 결측값 대치 모듈 및 불균형 데이터 분류 모듈은 컴퓨터 등의 중앙처리장치 등에 의해 실행될 수 있는 소프트웨어로 구성될 수 있다. 이하, 상기 결측값 대치 모듈 및 불균형 데이터 분류 모듈에 대하여 보다 구체적으로 설명한다.
- [0028] 강화 학습은 레이블되어 있지 않은 데이터에 대해 학습하는 비지도 학습(Unsupervised Learning)의 일종으로 환경과의 상호 작용을 통해 보상을 최대화하도록 행동을 취해서 학습하는 방법으로서, 학습하고자 하는 문제에 맞게 주어진 환경, 행동 및 평가에 의해 학습이 이루어진다. 즉, 행동을 취할 때마다 외부 환경에서 보상이 주어지는데, 이러한 보상을 최대화시키는 방향으로 강화 학습이 진행된다. 강화 학습은 에이전트와 환경으로 이루어지며, 환경은 상태, 행동, 보상으로 이루어진다. 에이전트는 환경에서 상태를 관찰하고 행동을 선택하는 학습을 수행하는 주체이며, 에이전트가 행하는 행동으로 보상을 받게 되고, 에이전트는 이러한 보상을 최대화하는 방향으로 행동을 취한다. 환경은 에이전트를 제외한 나머지로써, 확률적이며 행동을 한 후에 환경의 변화와 보상은 일정하지 않다. 특정한 상태에서 수행해야 할 행동을 선택하는 규정을 정책(Policy)라고 한다.
- [0029] 도 2는 강화 학습의 구조를 도시한 것이다. 도 2를 참조하면, 강화 학습은 시간을 t 라 할 때, 시간에 따른 환경의 상태를 S_t 로 표현하며, 에이전트가 취하는 행동을 A_t 로 표현하며, 보상은 R_t 로 표현한다.
- [0030] 본 발명에 따른 데이터 전처리 장치의 결측값 대치 모듈(10)은 강화 학습을 이용한 결측값 대체 방법 중 하나인 오토임пут레이션(Autoimputation) 기법에서 액터-크리틱을 적용한 컨트롤러의 학습에 PPO 학습 방법을 적용한 것을 특징으로 한다.
- [0031] 도 3은 오토임пут레이션 구조를 도시한 것이다. 도 3을 참조하면, 오토 임пут레이션 기법은 컨트롤러 모델과 차일드(Child) 모델로 구성되며, 컨트롤러 모델은 결측값에 대치할 값을 샘플링하는 역할을 하고, 차일드 모델은 컨트롤러 모델에 의해 샘플링된 값을 기반으로 대치한 데이터를 이용하여 학습 성능을 계산하는 역할을 하며, 차일드 모델에 의해 계산된 성능을 토대로 하여 보상값을 계산하고, 이를 이용하여 컨트롤러 모델의 파라미터를 업데이트하는 과정을 반복한다. 이때, 컨트롤러 모델의 학습은 Actor-Critic을 활용하게 된다.
- [0032] 컨트롤러 모델의 입력은 데이터의 결측값에 이전 액션의 값인 대치값을 채워 넣은 후, 1차원 배열로 변환시킨 벡터이다. 처음 에피소드에서는 결측값에 대한 대치값으로 0을 넣고, 컨트롤러에서는 결측값에 대치할 값들의 벡터를 샘플링한다. 샘플링한 벡터는 결측값의 각각에 대치될 값의 분포의 평균(μ)으로 이루어진 벡터와 표준편차(σ)로 이루어진 벡터이다. 이때, 벡터의 크기는 결측치의 개수와 같게 된다. 이렇게 컨트롤러에서 샘플링한 평균과 표준 편차를 통해 대치값을 선택할 확률 분포(π)를 수학적 1과 같이 가우시안 분포로 고려한다.

수학적 1

$$\pi(a, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(a - \mu)^2}{2\sigma^2}\right)$$

- [0033]
- [0034] 도 4는 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, 결측값 대치 모듈(10)의 컨트롤러 모델을 도시한 구조도이다. 도 4를 참조하면, 컨트롤러 모델은 2개의 신경망을 쌓아서 이루어진다. 컨트롤러 모델의 신경망은 고정되어 있으며, 학습 알고리즘의 성능 계산을 통해 얻은 보상값을 통해 컨트롤러 모델을 업데이트한다.
- [0035] 결측값 대치에서 액션은 컨트롤러에서 샘플링한 값들을 결측값에 대치하는 것으로 결측값에 적절한 대치값을 찾

는 문제를 연속적인 액션 공간 문제(Continuous Action Space)로 정의해준다. 따라서, 알고리즘을 적용하는 환경도 연속적이라고 가정한다. 강화 학습에서 액션 공간은 주로 이산 액션 공간으로 다루어진다. 이산 액션 공간에서의 강화 학습은 유한한 크기의 이산 액션 집합으로부터 임의의 액션들이 다른 결합없이 확률적으로 선택된다. 하지만, 본 발명에 따른 모델에서는 액션을 연속적인 액션 공간 문제로 정의해주었기 때문에 기존의 이산 액션 공간과는 다르게 구성되어져야 한다. 연속적인 액션 공간에서 액션은 수학적 1과 같이 가우시안 분포 함수의 확률 밀도를 증가시키는 방향으로 선택 확률을 키워간다. 즉, 연속적인 액션 공간 문제에서는 정책 생성 네트워크에서 가우시안 분포의 평균과 표준 편차를 생성한다.

[0036] 도 5는 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, 결측값 대체 모듈(10)의 컨트롤러 모델의 연속적인 액션 공간에서의 정책 생성 네트워크를 도시한 구조도이다. 도 5를 참조하면, 정책 생성 네트워크에서, 완전 결합 레이어(Fully-Connected Layer)를 거쳐서 가우시안 분포의 평균(μ)을 생성하고, Fully-Connected Layer와 SoftPlus 함수를 거쳐서 가우시안 분포의 표준편차(σ)를 생성한다.

[0037] 차일드 모델(child Model)은 XGBoost(Extreme Gradient Boosting)을 사용한다. 컨트롤러 모델에서 샘플링된 값을 통해 결측값에 대체한 후, 해당 대체값이 적절한지를 판단하기 위해 차일드 모델을 통해 데이터의 성능을 계산한다. 대체된 데이터는 학습 데이터와 시험 데이터로 나누어 주며, 이 경우 시험 데이터는 대체된 데이터가 존재하지 않게 설정한다. 학습 데이터로 차일드 모델을 학습시킨 후 시험 데이터로 정확도를 계산하며, 현재 상태에서 계산된 정확도와 이전 상태에서 계산된 정확도와의 차이를 보상값(R)로 정의한다. 이전 상태보다 정확도가 높아졌다면 양의 보상값을 갖게 되며, 이전 상태보다 정확도가 낮아졌다면 음의 보상값을 갖게 된다.

[0038] XGBoost는 기본적으로 부스팅(Boosting)이라 불리는 기술을 사용하며, XGBoost는 의사 결정 트리(Decision Tree) 모델에 부스팅 기법을 적용한 모델이다. 부스팅 기법은 앙상블 기법 중 하나로서, 단순한 분류가 가능한 약한 예측 모델들을 결합해서 강한 예측 모델을 만드는 알고리즘이며, 주어진 데이터를 약한 분류기를 통해서 학습후 학습된 결과에서 나타나는 오차를 또 다른 약한 분류기에서 학습시켜 오차를 줄여나가는 것이다. 따라서, 첫번째 학습을 통해 생성된 모델에서 오류를 발생시키는 데이터들을 다음 모델을 이용하여 오류를 줄이고, 또 다시 발생된 오류 데이터들은 그 다음 모델을 이용하여 오류를 줄이는 방식을 순차적으로 적용하는 것이다. XGBoost는 부스팅 기법을 통해 모델들을 결합할 때, 모델별로 서로 다른 가중치를 부여해서 중요도가 높은 트리 모델에 높은 점수를 부여하게 된다.

[0039] 컨트롤러의 학습에 적용한 액터-크리틱은 강화 학습의 정책 기반(Policy Gradient) 학습 중 하나로 액터와 크리틱을 각각 학습하는 알고리즘이다. 액터 모델은 상태에 따라 어떤 행동을 할지를 결정하고, 크리틱 모델은 행한 행동이 얼마가 가치있는가를 판단함으로써, 정책을 평가한다. 액터-크리틱은 액터와 크리틱의 학습을 반복하면서 학습이 지속되는데, 액터는 크리틱을 바탕으로 누적 보상 R을 높이는 방향으로 학습을 하고, 크리틱은 액터로 인해 얻은 R을 학습한다.

[0040] 본 발명에 따른 결측값 대체 모듈(20)의 컨트롤러 모델의 강화 학습에는 PPO 알고리즘을 사용한다. PPO는 강화 학습의 정책 기반 강화 학습 중 하나이며, 여기서 정책 기반 학습이란 최적의 정책을 찾기 위하여 정책 기반을 학습하는 방법이다. 여기서, 정책 기반 학습이란 최적의 정책을 찾기 위해서 정책 기반을 학습하는 방법이다. 강화 학습의 목적은 최적의 보상을 얻기 위해서 에이전트에게 최적의 행동 전략을 찾는 것이다. 정책 기반 학습은 정책을 직접적으로 모델링하고 최적화하는데 주력한다. 이 방법은 가치 기반(Value-based) 강화 학습을 통해 학습하면 발생하는 최적의 정책이 무한히 발견되지 못하는 단점을 보완하게 된다. 정책 기반 학습은 정책을 근사하는 정책 함수 $J(\theta)$ 를 최대화하는 θ 를 찾는 것이다. 정책 함수 파라미터(θ)는 수학적 2와 같이 업데이트 된다.

수학적 2

$$\theta_{i+1} = \theta_i + \alpha [\nabla_{\theta} \log \pi_{\theta}(a|s) G_i]$$

[0041]

[0042] PPO는 정책 기반 강화학습 중 하나인 TRPO(Trust Region Policy Optimization) 알고리즘에서 나온 것으로서, TRPO의 장점을 가지면서도 학습이 훨씬 더 간단하다. TRPO에서는 정책 갱신(Policy Update)을 제한해서 목적 함수(Objective Function)를 최대화한다. 제한을 두는 것으로 과도하게 정책이 갱신되는 것을 방지할 수 있다. 수학적 3은 TRPO의 surrogate 목적 함수를 나타낸다.

수학식 3

$$L^{CPI}(\theta) = \widehat{E}_t \left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \widehat{A}_t \right] = \widehat{E}_t[r_t(\theta) \widehat{A}_t]$$

여기서, $\widehat{A}_t$ 는 가치를 평가하는 네트워크에 의해 t의 시점에서 추정되는 이점(Advantage)을 말한다. TRPO에서는 KL-Divergence를 이용해서 제한을 두는데 PPO에서는 계산적으로 효율적인 제한을 두고 과도하게 큰 정책 갱신을 방지하기 위해 클리핑(Cliping) 기법을 사용한다. 이는 수학식 4로 정의된다.

수학식 4

$$L^{CLIP}(\theta) = \widehat{E}_t[\min(r_t(\theta) \widehat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \widehat{A}_t)]$$

여기서, $r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ 이고, ϵ 은 하이퍼파라미터, $r_t(\theta) \widehat{A}_t$ 는 원래 손실(Loss)이고, $\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \widehat{A}_t$ 는 잘린 손실(Clipped Loss)이다. $r_t(\theta)$ 가 1 이내의 값은 간격인 $[1-\epsilon, 1+\epsilon]$ 내에서 유지될 수 있게 제한을 걸어준다. $\text{Clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)$ 은 $[1-\epsilon, 1+\epsilon]$ 내에서 비율(Ratio)을 다듬는 역할을 한다. 따라서, PPO의 목적 함수는 원래 값과 잘린 값들 중 작은 값을 취하는 형태로 되어 있는데, 이를 통해서 좀 더 나은 보상을 얻기 위해서 극도로 정책을 갱신하는 동작이 없어지게 된다.

도 6은 Clip의 효과를 설명하기 위하여 도시한 그래프들이다. 도 6의 (a)는 A의 값이 양수일 경우, $r_t(\theta)$ 의 변화에 따라 L^{CLIP} 의 변화를 나타낸 것이다. A가 양수라는 것은 가치가 현재보다 높다는 것이며, 파라미터를 양의 방향으로 업데이트해야 한다. 상태(s)에서 행동(a)을 선택할 확률인 $\pi(a_t|s_t)$ 가 증가하도록 업데이트하며, $r_t(\theta)$ 가 아무리 커지더라도 ϵ 으로 자름으로써 신뢰구간 내에서 파라미터를 업데이트한다. 도 6의 (b)는 A의 값이 음수일 경우, $r_t(\theta)$ 의 변화에 따라 L^{CLIP} 의 변화를 나타낸 것이다. 도 6의 (a)와는 반대로, A가 음수라는 것은 가치가 현재보다 낮다는 것이며, 상태(s)에서 행동(a)을 선택할 확률 $\pi(a_t|s_t)$ 를 감소시키는 방향으로 ϵ 으로 자름으로써, 신뢰구간 내에서 업데이트한다.

도 7은 본 발명의 바람직한 실시예에 따른 데이터 전처리 장치에 있어서, PPO 알고리즘을 도시한 것이다.

본 발명에 따른 결측값 대체 모듈(20)은 컨트롤러 모델에서 샘플링된 액션을 통해 결측값을 대체시킨 후, 차일드 모델에 넣어서 성능을 측정한다. 성능을 통해 계산된 보상값을 기반으로 하여 PPO를 적용하여 컨트롤러 모델을 학습시킨다.

이하, 본 발명에 따른 데이터 전처리 장치의 불균형 데이터 분류 모듈(40)의 구성 및 동작에 대하여 구체적으로 설명한다.

본 발명에 따른 데이터 전처리 장치의 불균형 데이터 분류 모듈(40)은 강화학습 알고리즘 중에서 DQN 알고리즘을 기반으로 하여 보상 함수를 새롭게 정의해서 다수 클래스가 존재하는 불균형 데이터들을 분류하는 것을 특징으로 한다. DQN은 CNN(Convolution Neural Network)을 이용하여 Q 함수를 학습하는 강화학습 기법이다. Q 함수는 강화학습에서 상태-액션 조합의 질을 계산하는 함수를 말한다. 수학식 5는 Q 함수를 나타낸다.

수학식 5

$$Q^{\pi}(s,a) = E_{\pi}[g_t | s_t = s, a_t = a]$$

[0052]

[0053]

여기서, π 는 정책이며, $Q(t)$ 는 상태(s_t)에서 에이전트에 의해 행동(a_t)이 취해졌을 경우 행동으로 맵핑해주는 함수이며, g_t 는 누적된 보상을 나타낸다. 누적된 보상은 수학식 6으로 나타낼 수 있다.

수학식 6

$$g_t = \sum_{k=0}^{\infty} \gamma^{k+1}$$

[0054]

[0055]

DQN은 Q-learning의 한계점을 극복하기 위하여 출현되었다. Q-learning은 에이전트가 특정 상황에서 특정 행동을 하라는 최적의 정책을 배우는 것으로, 현재 상태로부터 시작하여 모든 연속적인 단계들을 거쳤을 때 전체 보상의 예측값을 극대화시킨다. 하지만, 이러한 Q-learning은 실제로 실행시켜 보면 잘 동작하지 않는 경우가 빈번하다. 순차적인 샘플 데이터 간의 상관관계(Correlation)로 인해 학습이 어렵고 예측한 값을 비교해야 하는 값(Target)이 안정되어 있지 않다는 문제점이 있다. 이러한 이유들로 인하여 실제로 Q-learning이 잘 동작하지 않게 된다.

[0056]

DQN은 이러한 Q-learning의 한계점을 극복하기 위하여 다음과 같은 기법들을 사용한다. 먼저, 경험 재생 버퍼(Experience Replay Buffer)를 사용하며, 목표값 신경망(Target Neural Network)과 예측값 신경망(Predict Neural Network)으로 네트워크를 분리한다. 경험 재생 버퍼는 순차적인 샘플 데이터간의 상관관계를 해결하기 위하여 사용된다. 에이전트의 상태가 변경된 즉시 훈련시키지 않고 일정 수의 샘플이 수집될 동안 기다리게 된다. 후에 일정한 수의 샘플이 버퍼에 쌓이게 되면, 무작위로 샘플을 추출해서 미니배치를 이용해 학습을 한다. 이때 하나의 샘플에는 상태, 행동, 보상, 다음 상태가 저장된다. 여러 개의 샘플로 학습을 수행한 결과들을 모두 수렴해서 결과를 내기 때문에 상관 관계 문제를 해결할 수 있다. 또한, 하나의 네트워크를 사용하면 목표 Q값이 계속 변경되므로 목표값이 변하는 것을 막기 위해 목표값을 출력으로 도출하는 목표값 신경망과 Q값을 예측하는 신경망을 분리한다. 이때 두개의 네트워크는 가중치 파라미터를 제외한 모든 것이 같은 네트워크로 정의해 준다. 예측한 Q값의 안정된 수렴을 위하여, 목표값 신경망은 계속 업데이트하는 것이 아니라 주기적으로 한번씩 업데이트시키는 것이 바람직하다.

[0057]

본 발명에서는 다수 클래스가 존재하는 불균형 데이터를 더 잘 분류할 수 있도록 하기 위하여 DQN을 사용한다. 본 발명에 따른 데이터 전처리 장치의 데이터 불균형 분류 모델에서는, DQN을 3개의 층을 쌓아서 구현하였다. DQN은 입력 데이터가 들어오면, 입력 데이터가 어떤 클래스에 속하는지 예측하며, 만약 예측이 맞으면 양의 보상을 주고, 예측이 틀리면 음의 보상을 줌으로써, 올바르게 예측하는 경우가 증가하도록 한다.

[0058]

DQN에 있어서, 상태는 훈련 데이터에 의해서 결정된다. 초기 상태(s_1)에서 에이전트는 먼저 데이터 샘플(x_1)을 받게 된다. 행동은 에이전트가 받은 데이터 샘플에 대해서 어떤 클래스 라벨에 속하는지를 예측하는 것이다. 보상은 행동의 결과로서 반환되는 것으로 에이전트의 행동의 성공 또는 실패에 대한 환경의 피드백이다. 보상은 만약 A, B, C 3개의 클래스가 존재하는 데이터를 가지고 실험하는 경우, 각 클래스가 각각 a, b, c 개의 샘플 수를 가지고 있고, 그 수가 $a > b > c$ 로 불균형하다고 가정한다. 알고리즘이 소수 범주에 있는 데이터에 민감할 수 있도록 C 클래스가 가장 적은 샘플 수를 가지고 있으므로 전체 데이터 수 N을 가장 많은 데이터 샘플수인 a로 나누어 준 a/N 을 보상으로 하여, 소수 클래스를 옳게 분류했을 경우 더 큰 보상을 주고 소수 클래스를 옳지 않게 분류했을 경우 더 큰 처벌을 주는 것이다. 이런 방식으로 다음으로 데이터 수가 적은 B 클래스는 b/N 을, 다수 범주에 속하는 A 클래스는 c/N 의 보상을 준다. 이렇게 입력 데이터마다 클래스별 샘플 수를 계산하고, 다수 클래스에는 전체 데이터에서 가장 샘플이 적은 클래스의 샘플 수를 나눠주고, 반대로 소수 클래스에는 전체 데이터에서 가장 샘플이 많은 클래스의 샘플 수를 나눠주는 것을 보상으로 하는 것이다. 위의 설명을 수학식 7로 표현할 수 있다.

수학식 7

$$R(s_t, a_t, l_t) = \begin{cases} \frac{c}{N} > \frac{b}{N} > \frac{a}{N}, & a_t = l_t \\ -\frac{c}{N} > -\frac{b}{N} > -\frac{a}{N}, & a_t \neq l_t \end{cases}$$

[0059]

[0060]

여기서, l_t 는 클래스 라벨을 말한다. 본 발명에 따른 불균형 데이터 분류 모듈은, 에이전트가 불균형 데이터에서 최적의 분류 정책을 학습할 수 있도록 소수 클래스에서 샘플에 대한 절대적 보상을 다수 클래스보다 더 높게 정의해준다.

[0061]

전술한 바와 같이, 본 발명에 따른 데이터 전처리 장치는, PPO를 이용한 강화 학습을 통해 결측값을 대체한 후, DQN-Multiclass를 이용한 강화 학습을 통해 불균형 데이터를 분류함으로써, 기존의 다른 방법들보다 월등히 높은 분류 정확도를 얻을 수 있게 된다.

[0062]

이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0063]

1 : 데이터 전처리 장치

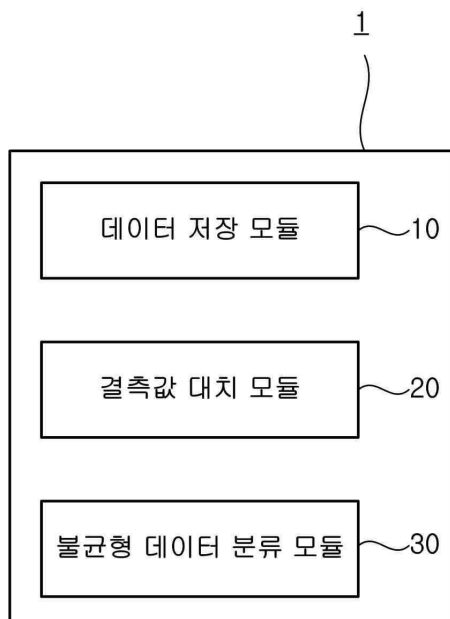
10 : 데이터 저장 모듈

20 : 결측값 대체 모듈

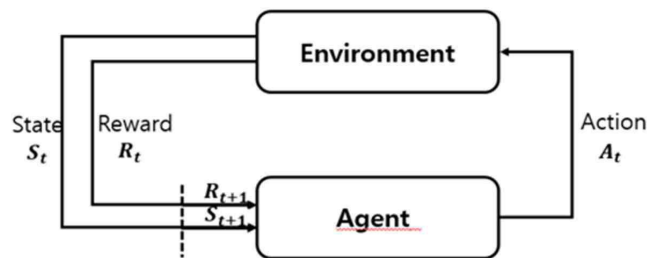
30 : 불균형 데이터 분류 모듈

도면

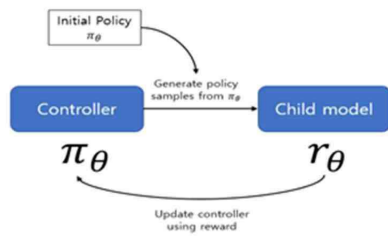
도면1



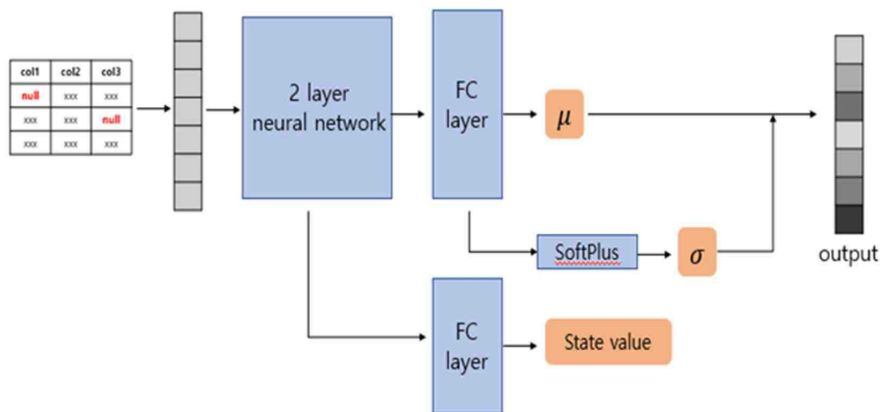
도면2



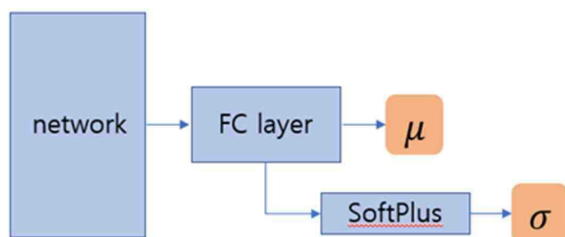
도면3



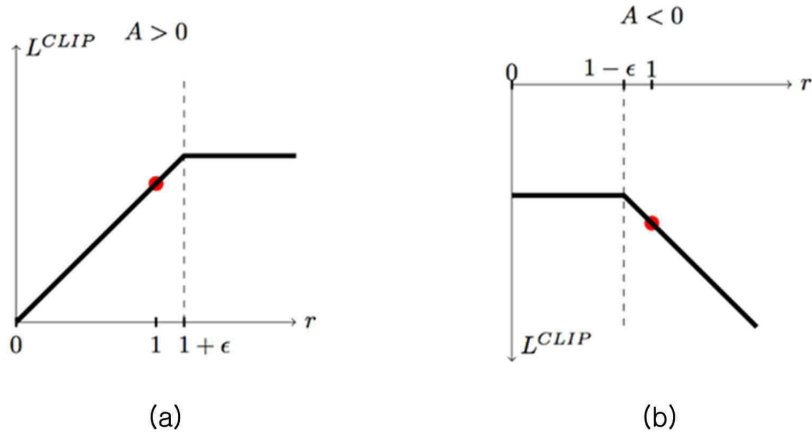
도면4



도면5



도면6



도면7

```

Algorithm 1. PPO, Actor-Critic Style
for iteration=1,2,... do
  for actor =1,2,...,N do
    Run policy  $\pi_{\theta_{dd}}$  in environment for T timesteps
    Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$ 
  end for
  Optimize surrogate L wrt  $\theta$ , with K epochs and minibatch size
   $M \leq NT$ 
   $\theta_{old} \leftarrow \theta$ 
end for
  
```