강화학습을 이용한 IP TV 상품 추천시스템 연구

데이터 지식서비스 공학과
이원욱

Since 2015
Dankook University Machine Learning Lab

ML Lab
DANKOOK UNIVERSITY

INDEX

요약

추천시스템은 온라인 산업에 중요한 부분으로 차지하고 있으며, 다양한 온라인 플랫폼에서 사용되고 있다. 근래에는 기술이 발달함에 따라 새로운 온라인 플랫폼인 IP TV가 등장하였다. IP TV에서 생활 서비스를 제공하던 위젯을 통해 각종 상품 판매 및 쿠폰을 제공하며 고객에게 상품 구매를 유도하고 있다. 이에 따라 본 연구에서는 로그 데이터에 최적화된 새로운 Model-Based 강화학습 방법을 통해 IP TV 상품 추천 시스템을 개발하였다. 새로운 강화학습 방법을 개발한 이유는 기존 강화학습은 보상 및 환경이 명확하게 정의되지 않아, 적용하는 것이 어렵기 때문이다. 따라서 이를 해결할 수 있는 2가지 방법을 적용하였다. 첫째로, GAN의 Mini-Max 최적화 방식을 적용하여 고객의 행동을 예측하는 모델을 개발하였다. 이를 통해 1차로 고객이 관심을 가질 만한 상품을 추천할 수 있었다. 둘째로, 1차 추천 상품 중에 고객의 관심이 가장 높은 상품을 추천하기 위해 계단식 DQN 방법을 적용하였다. 그리고 3개의 추천시스템을 비교한 결과, 강화학습을 적용한 모델이 가장 높은 성능을 보이며 최적의 상품을 추천할 수 있었다. 향후 이 방법론은 로그데이터 기반의 새로운 온라인 플랫폼에 새로운 지표가 될 수 있을 것이다.

*새로운 Model-Based 강화학습 방법을 통해 IP TV 상품 추천시스템을 개발하는 것이 목표

1. GAN의 Mini-Max 최적화 방식을 이용하여 고객 행동 모델 구축
2. 고객의 관심이 가장 높은 상품을 추천하기 위해 계단식 DQN 알고리즘 개발

I.서론

1.초개인화 기술과 추천시스템

*초개인화 기술

-트렌트 코리아에서 선정한 10대 키워드 중 하나는 '초개인화 기술'

-데이터를 통해 고객을 세분화 하는 기술로, 근래에는 초개인화 기술과 추천 시스템을 결합하여 고객의 감정을 분석하여 광고를 추천해주는 시스템이 등장

*로그 데이터 기반의 추천시스템

-온라인 환경에서 추천시스템과 고객의 상호 작용은 추천시스템이 추천한 상품을 고객이 클릭하는 것을 통해 이루어짐

-일반적인 추천시스템은 모델이 예측한 상품과 실제 클릭한 상품의 차이를 최소화하는 방향으로 학습하여 상품을 추천함

I.서론

1.초개인화 기술과 추천시스템

*강화학습을 사용해야 하는 이유
-일반적인 추천시스템은 장기간 동안 고객의 만족도를 고려하지 않음

-사용자의 관심은 시간이 지남에 따라 변할 수 있으며, 사용자의 행동 또한 마찬가지로 변화함

*강화학습에 추천시스템 적용이 어려운 이유
1)강화학습에서는 고객의 행동을 유도하는 관심을 아는 것이 중요함. 하지만, 추천시스템에서 고객의 관심을 이끌어 내는 것은 다양하고 복잡하여 알기 어려움.

2) 일반적으로 사용하는 Model-Free RL은 좋은 정책을 배우기 위해 환경과 상호작용이 많이 필요로 함. 하지만, 추천시스템에서 이는 비실용적이다 → 추천시스템이 쓸데없는 상품을 추천할 경우, 고객은 빠르게 서비스를 버리기 때문

*따라서 1차로 학습하고 강화학습을 진행하는 Model-Based 기반의 새로운 강화학습 방법론 개발이 필요.

I.서론

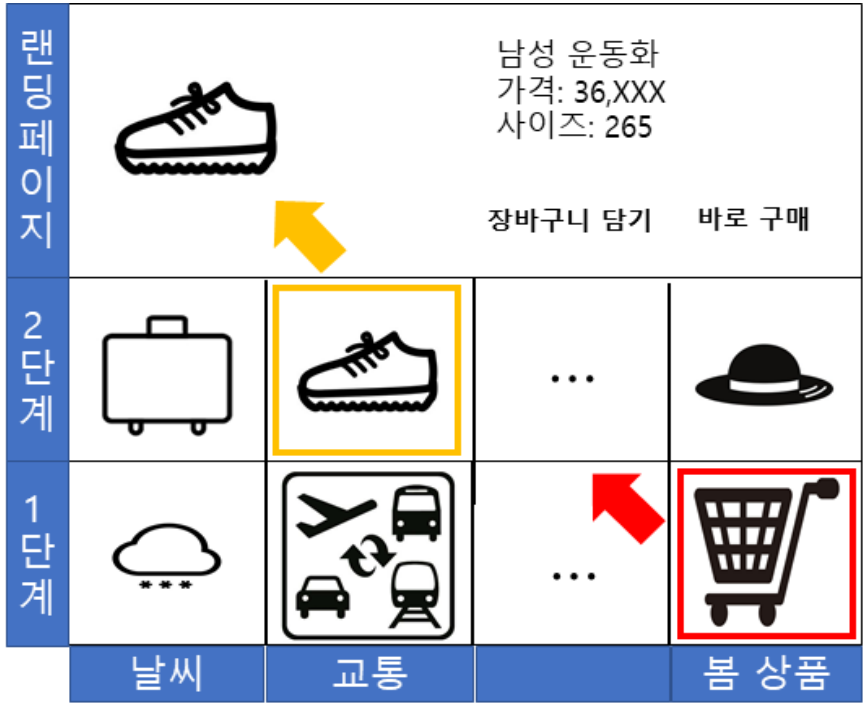
2. IP TV 상품 추천시스템

*로그 데이터
-연구 목표: 추천시스템을 통해 IP TV의 위젯에서 광고하는 상품을 추천하여 고객의 상품 구매를 유도하는 것.

-위젯은 1단계, 2단계, 랜딩페이지 단계로 이루어져 있음

-1단계: 생활정보 서비스 및 테마별 상품(대분류)
-2단계: 1단계에서 상품을 클릭하였을 경우, 세분화된 상품(중분류)
-랜딩 페이지: 2단계에서 선택한 상품의 상세 정보

*위젯의 2단계에서 랜딩페이지로 클릭하는 것을 상품의 관심도라고 판단하여 상품을 추천함



<그림 1> 위젯의 단계별 그림

I.서론

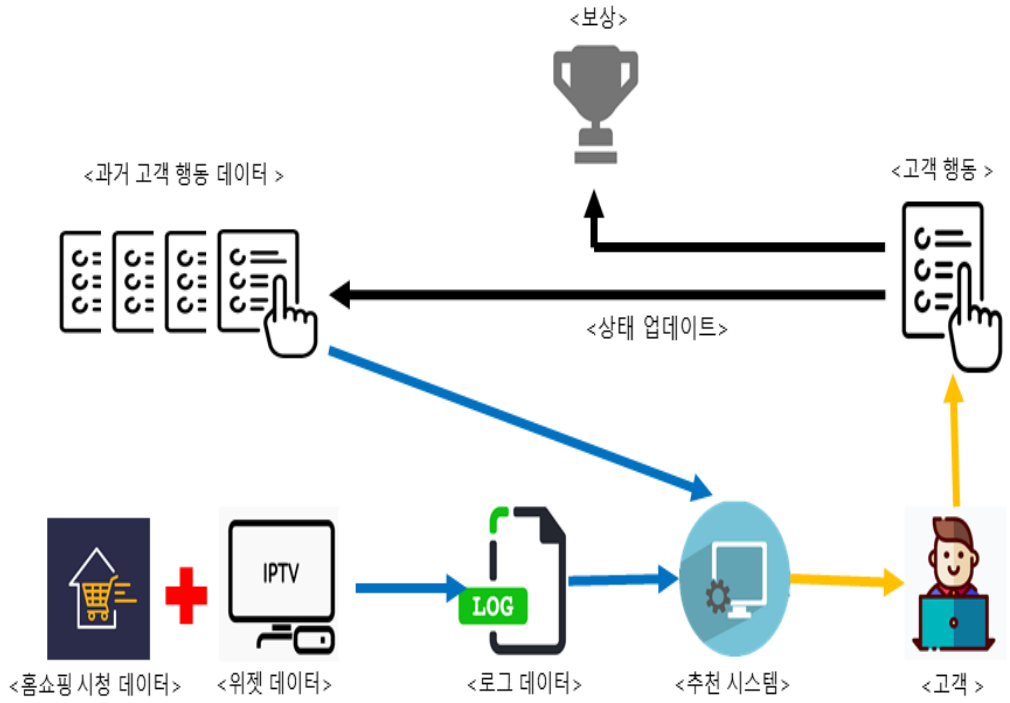
2. IP TV 상품 추천시스템

*홈쇼핑 시청 데이터
-통신사에서 실제로 수집한 로그데이터의 2가지 문제점
1)로그데이터의 수 부족
-수동적인 데이터 수집 방법, 데이터 건수 부족(5,000개)

2)고객이 클릭한 상품의 데이터 부족
-고객이 클릭한 상품과 클릭하지 않은 상품의 데이터 비율이 1:5 였으며, 결측 값이 전체 데이터에 30% 차지.

*고객이 시청한 홈쇼핑 채널 데이터
-고객이 본 채널의 상품 기록 및 시청 시간, 날짜 등을 활용

*홈쇼핑 시청 데이터를 로그 데이터화 한뒤 추가

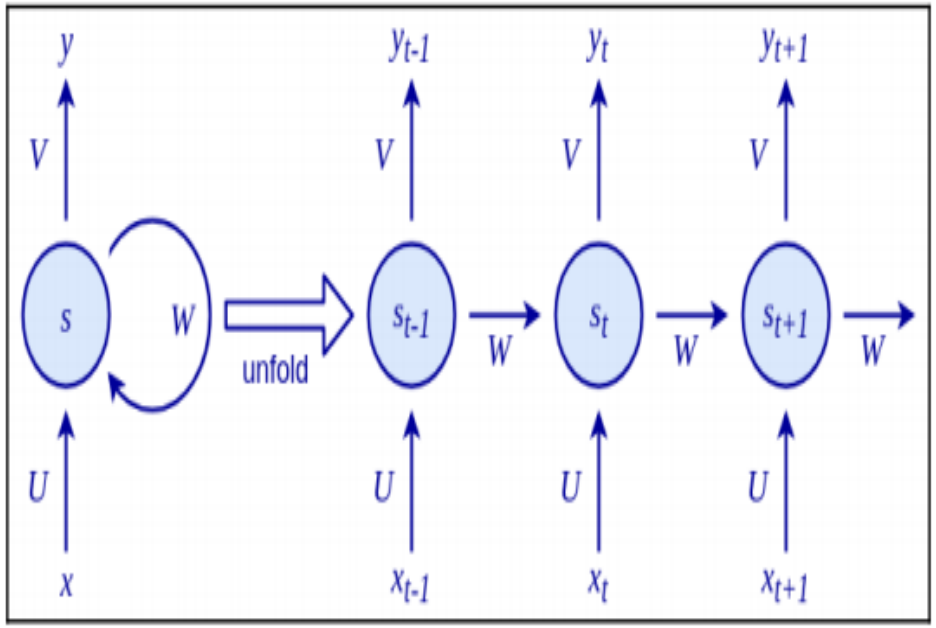


<그림 2> 강화 학습 설계도

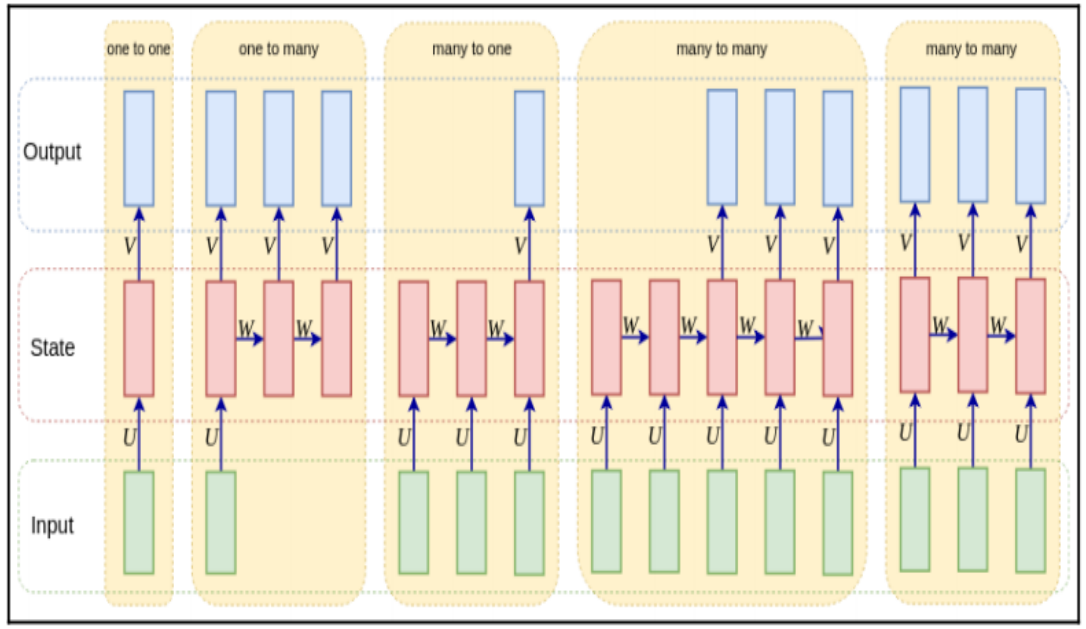
II. 방법론

1. 딥러닝

*RNN(순환 신경망)
-시계열 데이터를 다룰 수 있는 대표적인 딥러닝 기법.
-RNN은 각 State의 정보를 기억하여 다음 State 의 정보를 예측할 수 있음.
-본 연구에서는 강화학습을 위해 Many to Many 형태의 RNN 유형을 사용



<그림 3> RNN의 구조



<그림 4> RNN의 5가지 유형

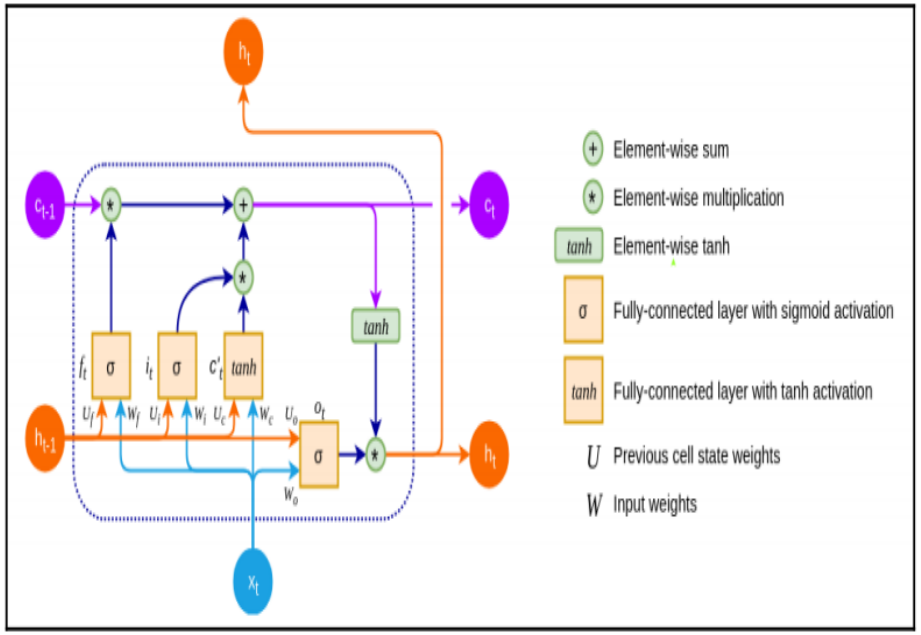
II. 방법론

1. 딥러닝

*LSTM(Long Short-term memory)
-RNN은 State 정보를 저장하는 것에 한계가 있으며, 역전파 과정에서 기울기 소실(Vanishing Gradient) 문제가 생길 수 있음.

-이를 해결하기 위해 LSTM을 사용.

-LSTM은 Input Gate, Forget Gate, Output Gate로 이루어져 있으며, 각 Gate들이 컨베이어 벨트같은 역할을 하며 선별된 정보를 받음.



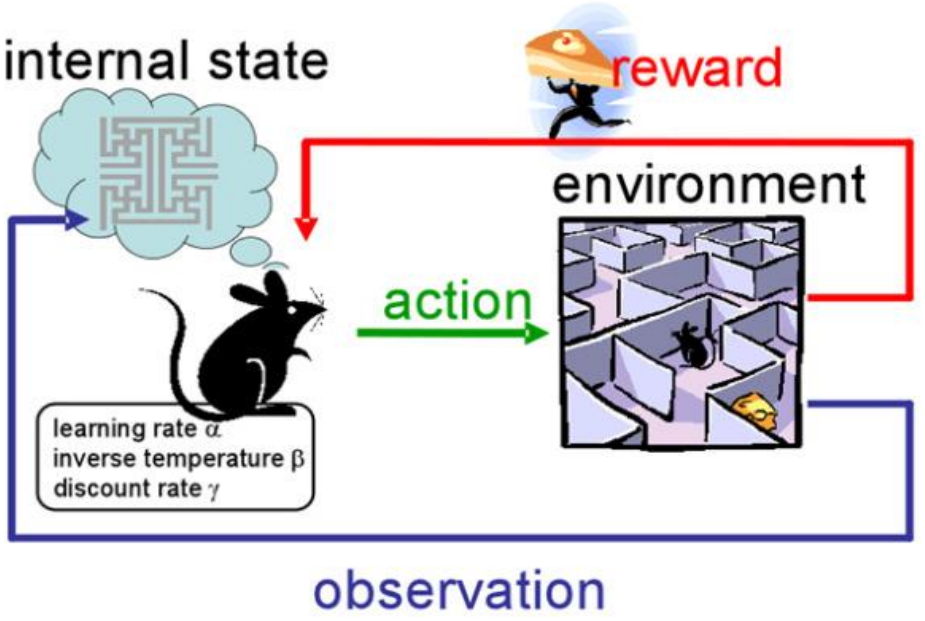
<그림 5> LSTM의 구조

II. 방법론

2. 강화학습

*강화학습 용어

1) **에이전트**(Agent): 학습 및 행동의 주체.
2) **환경**(Environment): 에이전트가 속해 있는 세계
3) **상태**(State): 에이전트가 현재 처한 상황.
4) **행동**(Action): 에이전트가 할 수 있는 행동.
5) **보상**(Reward): 에이전트가 행동한 뒤 환경으로부터 얻을 수 있는 피드백을 뜻함.
6) **정책**(Policy): 에이전트 행동의 연속 또는 집합. 누적 보상의 합을 최대화하는 최적의 정책을 찾는 것이 강화학습의 목표.



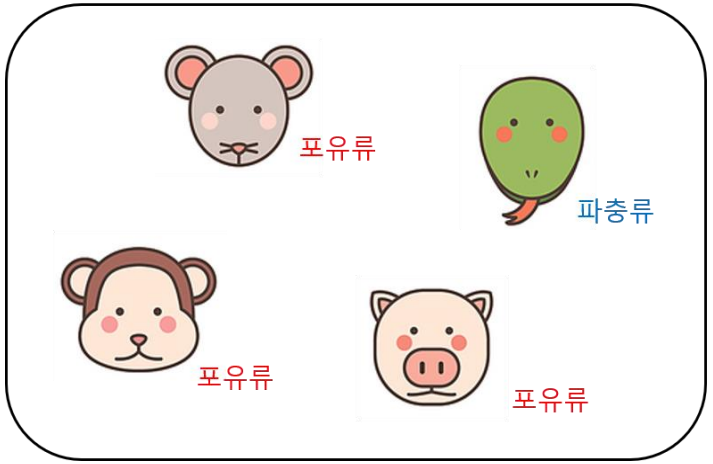
<그림 6> 강화학습의 구조

II. 방법론

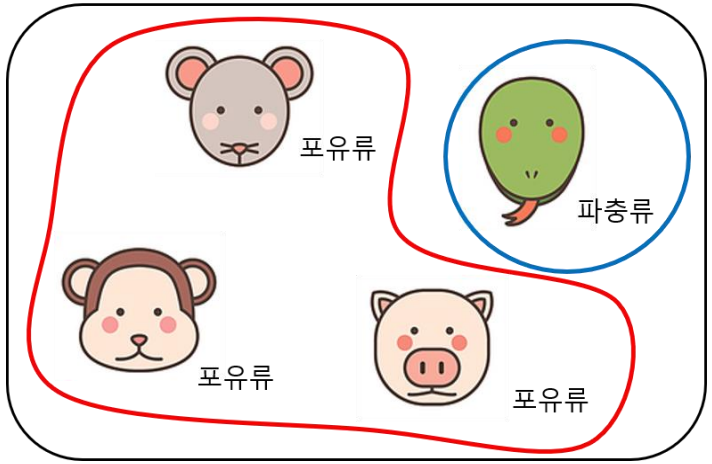
2. 강화학습

* 강화학습과 기계학습

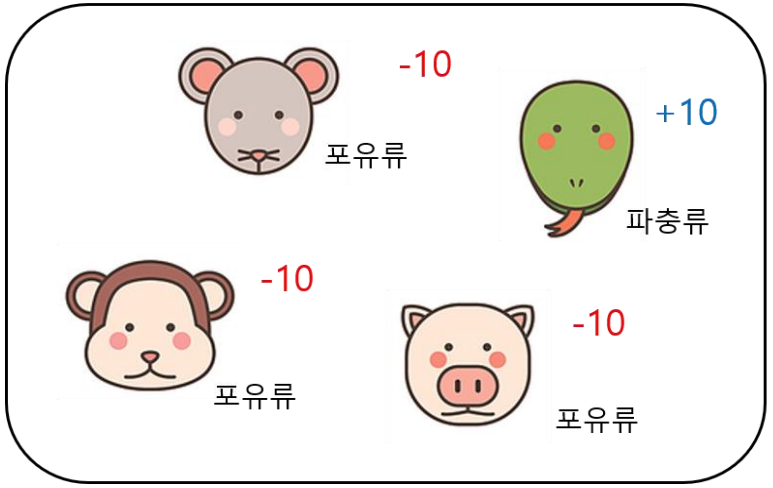
7-1. 지도학습



7-2. 비지도학습



7-3. 강화학습



1) 지도학습(Supervised Learning): 정답이 있으며, 주어진 데이터를 통해 학습하는 방법(Ex. 예측, 분류)

2) 비지도학습(Unsupervised Learning): 정답이 따로 없이 학습하는 방법 (Ex. 군집분석)

3) 강화학습(Reinforcement Learning): 보상과 벌을 주며, 보상을 최대화하도록 학습하는 방법

II. 방법론

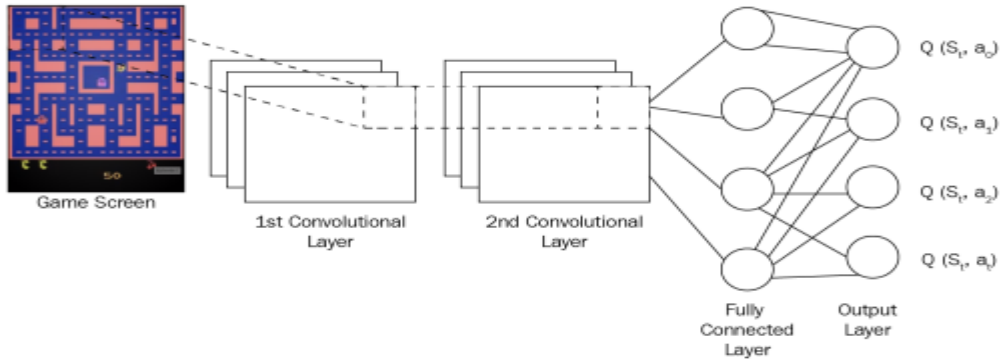
2. 강화학습

*DQN(Deep Q-Network)

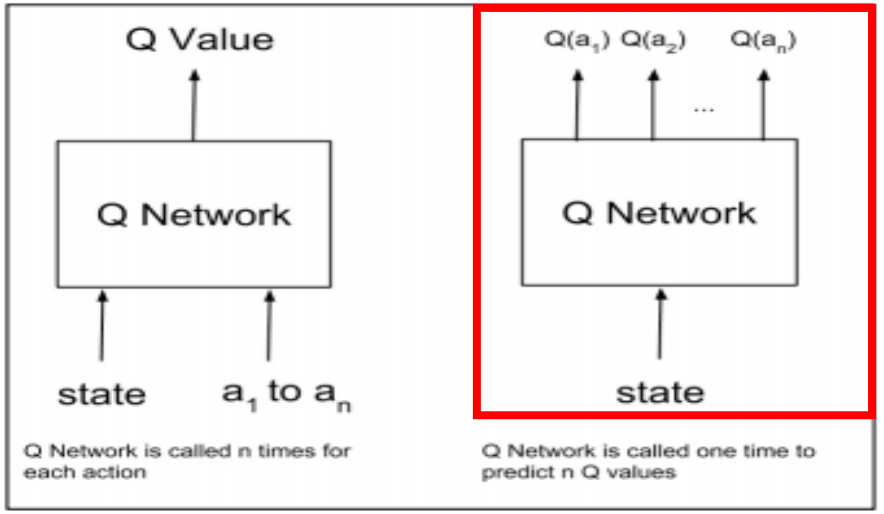
-Q-Function: 강화학습의 행동-가치 함수를 뜻하며, 현재 상태에서 에이전트가 행동을 할 때, 받는 누적 보상의 기댓값. 강화학습은 Q-함수 값을 최대화 시킬 수 있도록 최적의 정책을 학습하는 것이 목표

*Q-Learning: Q-Function을 학습하기 위해 사용하는 방법.
*Q-Network: Q-Learning 신경망에 접목한 방법

*DQN
-Q-Learning에 Deep Learning을 접목한 방법
-Q-Learning의 단점을 보완할 수 있음
^



<그림 9> DQN의 학습 과정



<그림 10> Q-Network의 2가지 학습 유형

II. 방법론

2. 강화학습

*DQN(Deep Q-Network)

*Q-Network의 문제

1)타겟 값의 변화: 강화학습에서 타겟 값은 학습 과정에 따라 계속해서 바뀜.(모델 불안정)

2)데이터 간의 강한 상관관계: 강화학습은 시계열 데이터를 사용하며, t시점 값에 따라 t+1 시점의 값이 결정되기 때문에 데이터들 간에 강한 상관관계가 있음

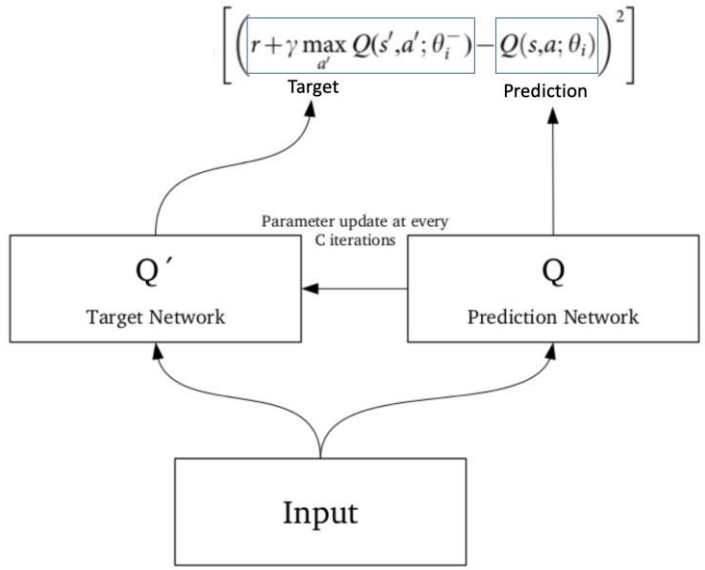
*DQN의 해결방법

1) 2개의 Network 사용(Target Network, Prediction Network)

- Prediction Network는 학습을 위한 네트워크로 Target Network의 고정된 타겟 값을 기준으로 학습
- 일정한 학습횟수가 지나면 Target Network는 Prediction Network의 weight를 복사
- 모델의 불안정 문제 해소

2) Experience Replay Memory

- Experience Replay Memory에 데이터를 저장하고 랜덤하게 데이터를 샘플링 하여 학습을 진행
- 데이터들의 상관 관계가 줄어들고 에이전트는 다양한 경험을 할 수 있음.



III. 실험

1.데이터

고객 ID	세션 ID	시간	상품 ID	클릭 여부
A	A_1	0	구두	0
A	A_1	0	신발	1
A	A_2	1	티셔츠	0
A	A_2	1	운동화	0
A	A_2	1	코트	1
B	B_1	2	구두	0
B	B_1	2	신발	0
B	B_1	2	코드	1

<표 1> 로그 데이터 예시

1)로그 데이터

-2019년 11월 13일 ~ 12월 4일(3주간 데이터), 약 5000건

-IP TV의 위젯의 상품 클릭정보가 담긴 노출-반응 로그 데이터

-클릭여부 변수: 랜딩 페이지로 넘어가기 위해 상품을 클릭한 경우 1, 이전에 세션이 종료되었거나 클릭하지 않았을 경우 0으로 구분

-고객 ID, 세션 ID, 로그 시간, 상품 ID등의 변수로 구성

2)홈쇼핑 데이터

-2019년 11월 13일 ~ 12월 4일(3주간 데이터),약 1.8억 건

-고객 ID, 상품 ID, 상품 시청 시간, 시청한 날짜 등의 변수로 구성

III. 실험

1.데이터

고객 ID	상품 ID	시청 시간 (초)	시청 날짜
A	구두	30	20191101
A	신발	20	20191101
A	티셔츠	100	20191102
A	운동화	150	20191103
A	코트	10	20191103
B	구두	5	20191101
B	신발	10	20191101
B	코드	30	20191101

<표 1> 홈쇼핑 데이터 예시

1)로그 데이터

-2019년 11월 13일 ~ 12월 4일(3주간 데이터), 약 5000건

-IP TV의 위젯의 상품 클릭정보가 담긴 노출-반응 로그 데이터

-클릭여부 변수: 랜딩 페이지로 넘어가기 위해 상품을 클릭한 경우 1, 이전에 세션이 종료되었거나 클릭하지 않았을 경우 0으로 구분

-고객 ID, 세션 ID, 로그 시간, 상품 ID등의 변수로 구성

2)홈쇼핑 데이터

-2019년 11월 13일 ~ 12월 4일(3주간 데이터),약 1.8억 건

-고객 ID, 상품 ID, 상품 시청 시간, 시청한 날짜 등의 변수로 구성

III. 실험

2.데이터 전처리

*홈쇼핑 데이터의 로그 데이터화 과정

1) 고객 ID 기준으로 데이터가 5개 이하일 경우 제거
2) <그림 11-1>처럼 시청 시간의 편차가 크기 때문에, 시청시간에서 IQR*3의 범위를 벗어나면 이상치라 판단하여 제거
3) 로그 데이터는 상품의 클릭 정보가 있으며, 이것은 상품에 대한 고객의 관심도라고 할 수 있음. 시청 데이터에는 시청 시간을 관심도로 볼 수 있으며 Binary 형태로 표현하기위해 관심 변수 생성

(시청시간의 상위 25% 이내에 속할 경우 1, 아니면 0. (84초 이상))

4) 시청 시간을 주 단위로 나누고, 주별로 고객이 관심도가 있는 상품이 한 개도 없을 경우 제거(관심 변수: 1)
5) 데이터의 중복 제거 및 데이터 분할(Train, Validation, Test, 6:2:2)

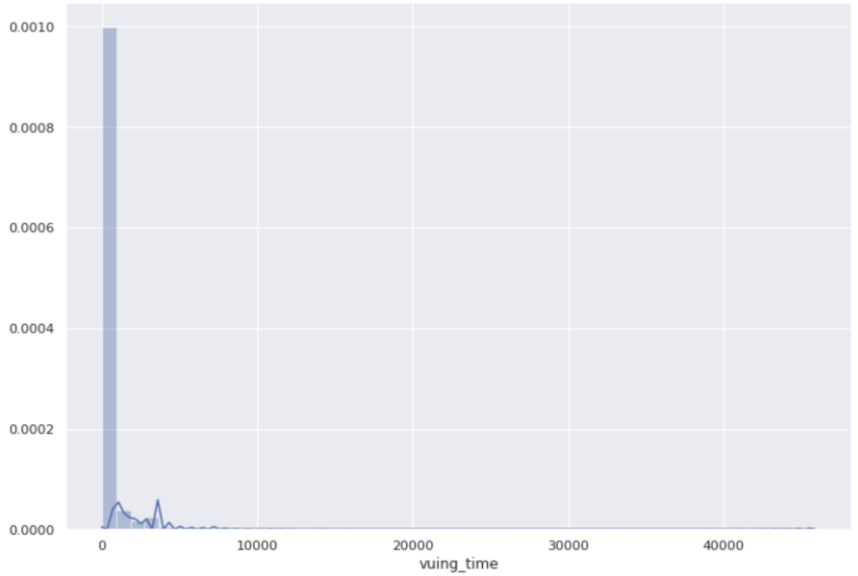


그림 11-1 홈쇼핑 시청 시간

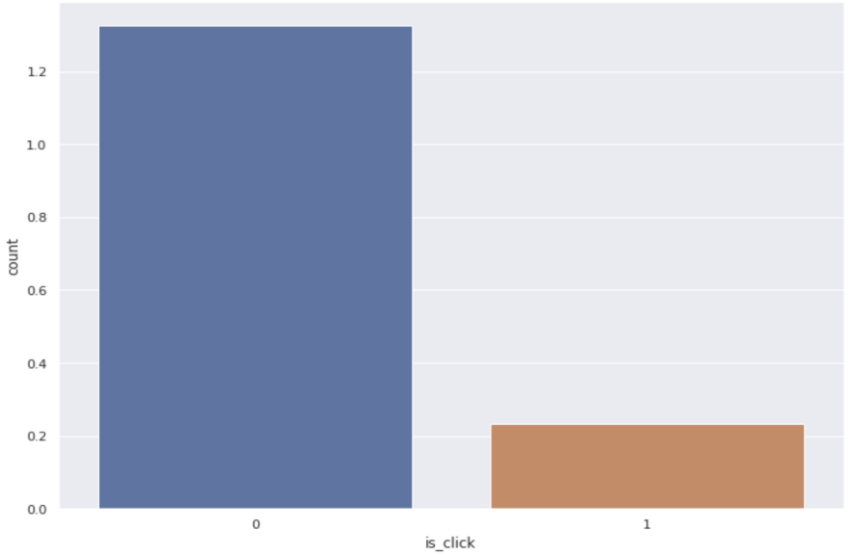


그림 11-2 고객의 관심 여부

III. 실험

3.사용 모델

*강화학습의 모델 환경

1)추천시스템은 I개의 전체 상품 중 k 개의 상품을 추천.

$$\pi^* = \arg\max_{\pi(s^t, I^t)} E[\sum_{t=0}^{\infty} \gamma^t r(s^t, a^t)]$$

2)강화학습의 각 요소

1. **Environment**: IP TV 위젯 또는 고객이 시청한 홈쇼핑 TV 채널

2. **State**: 고객이 클릭한 상품의 기록 또는 시청한 홈쇼핑 채널 상품의 기록

3. **Action**: 추천시스템이 추천한 상품

4. **Reward**: t시점의 State에서 Action을 하고 난 뒤 고객의 만족도(보상)에 해당한다. 본 연구에서 보상은 추천시스템의 유용성과 같음

5. **Policy**: t시점의 State에서 아이템을 고르는 Action들의 집합

* Environment와 State는 고객과 관련이 있으며, Action와 Policy는 추천시스템과 관련이 있다. Reward는 고객과 추천시스템 모두와 관련이 있다.

III. 실험

3.사용 모델

*고객 행동 모델: 고객의 행동을 모방(예측)하는 모델

1. 가정

1) 고객은 수동적인 사람이 아니다. 고객은 자신의 이익(보상)을 최대화하는 방향으로 상품을 선택한다.

2) 보상은 추천시스템의 상품 추천 뿐만 아니라 고객이 상품을 선택하는 것에도 영향을 준다.

Ex.)유튜브와 같이 추천시스템이 추천한 영상을 고객이 시청함으로써, 고객의 선택이 바뀔 수 있음.

2. 강화학습의 목표

$$\Phi^*(s^t, A^t) = \underset{\Phi \in \Delta^{k-1}}{\operatorname{argmax}} E_{\Phi}[r(s^t, a^t)] - R(\Phi)/\eta$$

$\underset{\Phi \in \Delta^{k-1}}{\operatorname{argmax}} E_{\Phi}[r(s^t, a^t)]$ 는 최적의 정책을 구하는 식이며, $R(\Phi)/\eta$ 는 에이전트가 탐험적으로 행동하기위한 정규화 함수

3.정규화 함수

-Negative Shannon Entropy를 사용. η 는 정규화 함수의 영향력을 통제하는 역할. η 값이 작을 수록 모델은 더욱 탐험적으로 행동함

III. 실험

3.사용 모델

*모델 파라미터

1. 실제 State: t시점 전까지 클릭한 상품의 기록을 Embedding 한 값.

$$s^t = h(F^{1:t-1} := [f_*^1, ..., f_*^{t-1}])$$

•f_x^T는 t시점에서 클릭한 상품 기록의 Feature를 뜻함

2.Reward의 파라미터

$$r(s^t, a^t) := v^T \sigma(V[(s^t)^T, (f_{a^t}^t)^T]^T + b$$

- V는 가중 행렬 이며, b는 bias 벡터를 뜻함.

3.고객 행동 모델의 파라미터

$$\Phi(s, A^t) \propto \exp(v'^T \sigma(V'[(s^t)^T, (f_{a^t}^t)^T]^T + b'))$$

*보상 기능에 대한 모든 파라미터는 θ 로, 고객 행동 모델에 대한 모든 파라미터는 α 로 표현.

III. 실험

3.사용 모델

*모델 학습

Generative Adversarial Training:

$$\min_{\theta}\max_{\alpha}\left(\mathbb{E}_{\phi_\alpha}\left[\sum_{t=1}^{T} r_\theta(s_{true}^t, a^t)\right] - R(\phi_\alpha)/\eta\right) - \sum_{t=1}^{T} r_\theta(s_{true}^t, a_{true}^t), \quad (5)$$

실제 수식

$$\begin{cases} \alpha \leftarrow \alpha + \gamma_1 \nabla_\alpha \mathbb{E}_{\phi_\alpha}\left[\sum_{t=1}^{T} r_\theta\right] - \gamma_1 \nabla_\alpha R(\phi_\alpha)/\eta; \\ \theta \leftarrow \theta - \gamma_2 \mathbb{E}_{\phi_\alpha}\left[\sum_{t=1}^{T} \nabla_\theta r_\theta\right] + \gamma_2 \sum_{t=1}^{T} \nabla_\theta r_\theta. \end{cases}$$

<모델의 파라미터>

*행동 모델 Φ는 Reward를 최대화하기 위해
실제 사용자가 제공하는 행동 순서를 모방하려고 함.

Φ : 고객의 행동을 생성하는 Generator 역할

R: Generator와 사용자의 실제 행동을 구별하는
Discriminator 역할을 함.

-Generator: 실제행동과 생성한 행동의 차이를 적게 만들어 Reward를 많이 받는 것이 목표

-Discriminator: 모델과 실제 행동을 구분하여 Generator에 최대한 Reward 적게 주는 것이 목표

III. 실험

3.사용 모델

*Cascading DQN

-추천시스템의 최적의 정책을 Q-Function으로 표현하면 아래와 같음

$$Q^*(s^t, A^t) = E[r(s^t, A^t, a^t) + \gamma \max_{A' \subset I'} Q^*(s^{t+1}, A^t)]$$

$$\pi^*(s^t, I^t) = \arg\max_{A^t \subset I^t} Q^*(s^t, A^t)$$

- $I^t \subset I$ 이며, t시점에서 추천가능한 전체 상품들이다. 추천시스템은 각 고객이 본 K개의 상품 후보군 중에서 k개 상품을 추천.($\binom{K}{k}$) Ex.) 고객이 본 상품이 1000개라면 그 중에서 5개의 상품 추천.
- Q-Function에서 각 시점에서 전체 상품과 고객이 본 상품, 그 중에서 상품을 추천하는 행동은 공간이 매우 커지며, 계산량이 많아지게 됨.
- 고객에게 최적의 상품을 추천하기 위해 추천한 상품들 간의 경쟁을 필요함.

III. 실험

3.사용 모델

*Cascading DQN

1)추천시스템의 최적의 행동: $A^* = a^*_{1:k} = \arg\max_A Q^*(s, A)$

2)추천시스템의 정책: $\max_{a_{1:k}} Q^*(s, a_{1:k}) = \max_{a1}(\max_{2:k} Q^*(s, a_{1:k}))$

3)Cascading DQN 알고리즘

Cascading Q-Networks:

$$a_1^* = \arg\max_{a_1}\{Q^{1*}(\boldsymbol{s}, a_1) := \max_{a_{2:k}} Q^*(\boldsymbol{s}, a_{1:k})\},$$
$$a_2^* = \arg\max_{a_2}\{Q^{2*}(\boldsymbol{s}, a_1^*, a_2) := \max_{a_{3:k}} Q^*(\boldsymbol{s}, a_{1:k})\},$$
$$\ldots$$
$$a_k^* = \arg\max_{a_k}\{Q^{k*}(\boldsymbol{s}, a_{1:k-1}^*, a_k) := Q^*(\boldsymbol{s}, a_{1:k})\}.$$

-각 행동에 마다 이전 행동을 고려하여 Q-Values 값이 가장 높은 행동을 하도록 학습. 예를 들어, a_2^* 는 전단계에서 선택한 최적의 행동 a_1^*과 새롭게 선택한 a_2 행동 중 Q-values가 더 높은 행동을 취함.

III. 실험

3.사용 모델

*전체 모델의 학습 과정

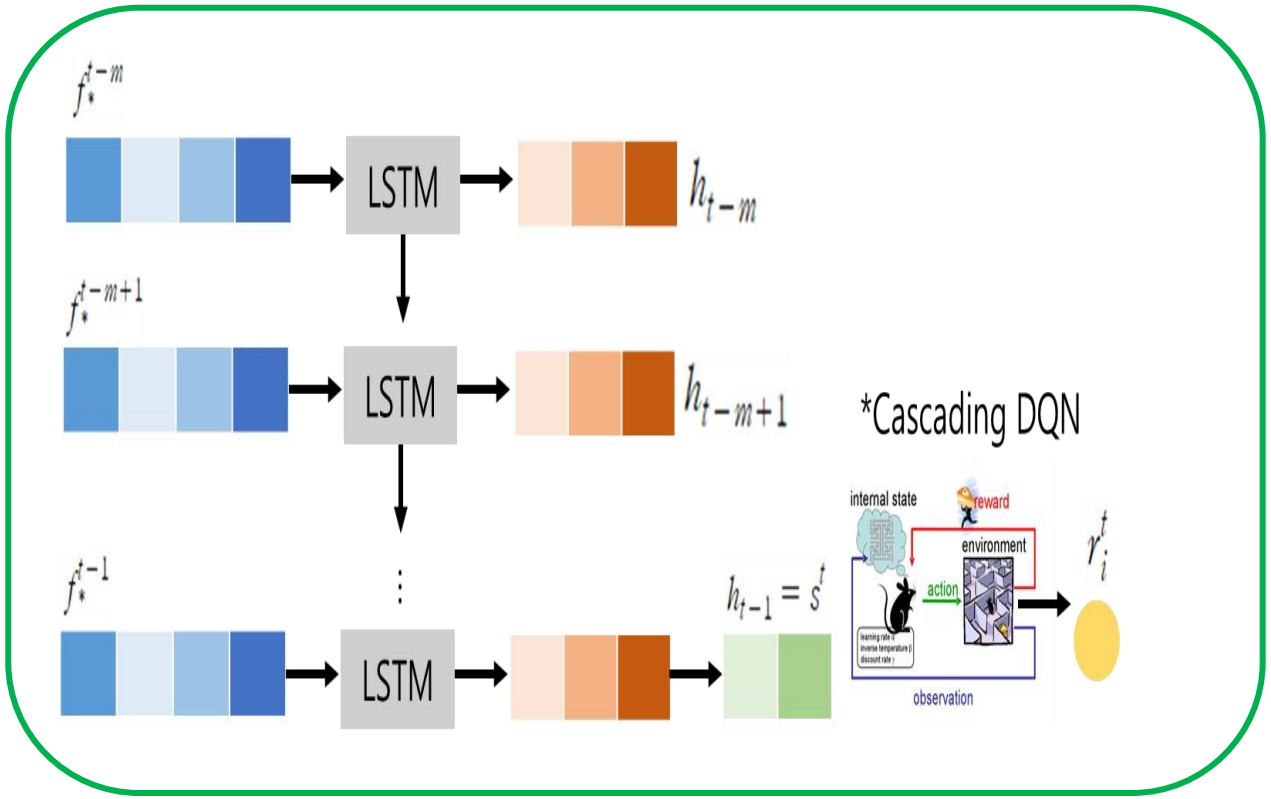


그림 12-1 전체 구조

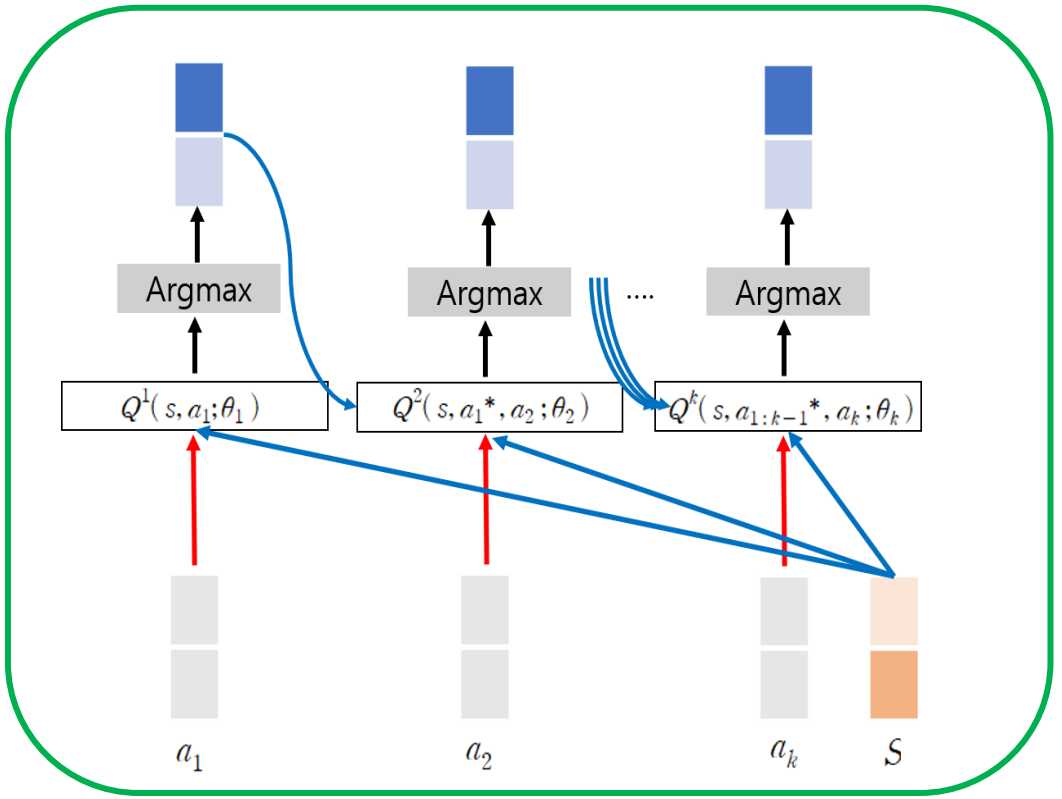


그림 12-2 Cascading DQN

IV. 성능 평가

*모델의 평가척도

1) Accuracy: 전체 고객 중 추천시스템이 추천한 상품과 실제로 고객이 클릭한 상품의 정확도

2) Precision @K : 추천한 상품 중에 유저가 실제로 클릭한 K개 상품의 비율

*각 모델의 hyper-parameter

Epoch-10000, Batch_size- 25, Learning_late- 0.0005

*RNN hidden dim-200

*Train, validation, Test를 나누는 과정을 5번 반복 진행하여 Mean과 Std 값을 구함

Model	loss	Accuracy		Precision@1		Precision@2	
		Mean	Std	Mean	Std	Mean	Std
TensorRec	1.815	0.4048	0.1414	0.2816	0.1479	0.4425	0.1416
LSTM	1.1112	0.7323	0.1064	0.5422	0.0987	0.7731	0.1314
Cascading RL	0.7300	0.7739	0.0408	0.5822	0.0523	0.8288	0.1739

V. 결론 및 향후 과제

* 결론

- 본 연구에서는 IP TV 상품 추천시스템을 위한 새로운 Model-Based 강화학습 방법론을 제안하였음

1) 고객 행동 모델 구축을 위해 GAN의 Mini-Max 최적화 사용.

2) 1차로 추천한 상품 중에서 고객의 관심도가 높은 최적의 상품을 추천하기 위해 Cascading DQN 적용.

- 3개의 모델 성능평가 결과 강화학습의 Accuracy와 Precision이 가장 높았음.

*향후 과제

- 로그 데이터의 부족 문제

-두 변수사이에 연관성이 있었지만, 두 데이터의 수집방법이 다름. 따라서 로그 데이터만 이용하였을 때 상품 추천은 다른 결과가 나올 수 있음.

-고객이 시청한 상품 및 클릭한 상품이 적을 경우, 추천시스템의 추천 정확도가 떨어질 수 있음.

* 강화학습은 새로운 데이터가 추가되면 자동으로 파라미터를 조절할 수 있다는 장점이 있음.

- 즉, 고객의 변화된 관심도를 반영한 상품을 추천할 수 있음. 비록 실험은 오프라인으로 진행되었지만, 더 좋은 결과는 현재 진행하고 있는 온라인 A/B 테스트를 통해 얻을 수 있을 것으로 기대됨.

감사합니다.

단국대학교 기계학습연구실
이원욱

Since 2015
Dankook University Machine Learning Lab