디 YNAMIC 피 RICING ON 이자형- 상업 피 LATFORM WITH 디 EEP 아르 자형 EINFORCEMENT 엘 적립

익명의 저자
이중 맹검의 검토중인 용지

에이 BSTRACT

본 논문에서는 전자 상거래 플랫폼에서 주소 동적 가격 문제에 깊은 강화 학습 (DRL)을 기반 접근 방식을 개발합니다. 우리 마르코프 의사 결정 프로세스와 같은 모델 실시간 세계 전자 상거래 동적 가격 문제. 환경성 상태는 드 Fi를 다른 비즈니스 데이터의 네 그룹으로 네드된다. 우리는 최첨단 DRL 기반의 동적 가격에 대한 몇 가지 주요 개선을 접근 : 1. 우리 Fi를 처음, 연속적 조직 구성 단위 가격 활동 공간에 동적 가격의 응용 프로그램을 확장. 2. 우리는 다른 보상 기능을 설계함으로써 미지의 수요 함수 문제를 해결. 3. 콜드 시작 문제는 과거 판매 데이터를 사용하여 사전 교육 및 평가를 도입하여 해결됩니다. 필드 전직 periments 설계 및 실제 전자 상거래 플랫폼 개월 동안 지속되는 제품의 SKU pric- 보내고 수천 실시하고 있습니다. 전자 상거래 플랫폼에서 수익 전환율 (DRCR)의 차이는 이전 연구에서 결론 상이한 만 수익보다 적절한 보상 함수의 실험 결과를 나타낸다. 한편, 더 나은 이산보다 연속 동작 모델 수행을 제안했다.

1 명 I 서론

종종뿐만 아니라 수익 관리라고 동적 가격은, 가격이 왼쪽과 수요 반응은 수익을 극대화 수시로 관찰 인 - ventories에 따라 조정하는 것입니다. 그것은 1970 년대에 항공 산업의 규제 완화 이후 지난 수십 년 동안 큰 주목을 받고있다. 웨더 및 신체 (1992)와 Talluri 및 반 Ryzin (2006)는 과학 수율 관리의 결합 영역을 초과 예약하는 것이 ELD, 가격은 부패하기 쉬운 자산 수익 관리의 ELD Fi를,에서 수행 된 연구의 개요를했다.

사업의 최근 개발하는 동안, 많은 산업은 수익 관리에보다 적극적이되었다. 둥네 짱처럼 타고 공유 플랫폼은 '서지'가격과로 알려진, 동적 가격 전략을 구현했습니다 첸 & 셀던 (2016)이 더 운전 시간에 대한 동기 부여에 유의미한 영향을 미치는 것으로 나타났다. 자라와 같은 소매 업체들은 체계적인 동적 인하 가격 전략을 구현했습니다. 자라가 판매하는 많은 수의 항목을 유지하면서 각 항목에 의해 생성 된 수익을 증대하기 위해 오목 & (2012) GALLIEN은 통관 가격을 공부했다. 크로 거 지금 켄터키 (Nicas (2015))의 한 가게에서 전자 가격 태그를 테스트하고있다.

온라인 소매 업체로 인해 더 복잡한 opera- TIONS 동적 가격 전략에 대한 강한 욕망을 가지고있다. Amazon.com은 **356,000,000 제품 (현재 562,000,000)를 판매하고있다. Walmart.com는 2017 추정에 따라 420 만 개 제품을 판매** [1]. **Taobao.com,** 중국에서 가장 큰 전자 상거래 플랫폼은 현재 제품의 수십억을 판매하고있다. 이러한 항목은 항목의 수가이 높은 갈 때 임무 불가능할 것이다, 경쟁력을 유지하기 odically 수술 주변에 대한 운영 전문가는 세트 가격이 있습니다. 그 결과, 아마존은 자동 **가격 책정 시스템을 구현하고이를 Amazon.com 가격을 15 분 간격으로 변경할 수 있다는보고** [2]. **첸 둥. (2016) Amazon.com** 경험적의 가격 전략을 공부하고 가격에 대한 유의 한 인자를 산출했다.

본 논문에서는 주소로 접근에게 온라인 소매 업체에 의한 수치 렘 가격 역동적 인 학습 보강을 제안했다. 우리가 생각 시나리오는 Tmall.com, 중국에서 가장 큰 사업에 소비자 소매에 다른 제품에 대한 동적 가격에, Taobao.com에서 분사 한 방법이다. 그곳에

[1] https://www.scrapehero.com/how-many-products-does-walmart-com-sell-vs-amazon-com

[이] https://www.whitehouse.gov/sites/default/ Fi를 레 / 문서 / 빅 데이터 보고서 Nonembargo의 V2 _ _ _ _ PDF

이러한 전자 상거래 플랫폼에서 가격에 대한 많은 DIF Fi를 culties입니다. 첫째, 시장 환경은 정량화 될 수 없다. 같은 제품에 대한 수익으로 인해 트래픽 Fi를 C, 다른 제품의 가격 또는 이전 구매자도 의견의 변화 매일 고객의 예측할 수없는 플로리다 uctuation에 극적으로 변경 될 수 있습니다. 보상 기능이 같은 복잡한 환경에서 제대로 설정되지 않은 경우 둘째, 그것은 비 융합 정책으로 이어질 것입니다. 셋째, 신속하게 대규모 자본 손실이 발생할 수있는 온라인 약간 부적절 가격 때문에, 직접 온라인 가격 모델을 적용 할 적용 할 수 없습니다. 이 냉간 시동 문제를 목표로 따라서, 사전 교육 및 평가가 필요하게된다. 넷째, 추천 시스템과는 달리 온라인 A / B 테스트를 수행하는 것은 불가능합니다, 불법이기 때문에 다른 고객에게 동시에 서로 다른 가격을 노출. 즉, 과학 ELD의 실험 기간 동안 서로 다른 가격 정책의 성과를 평가하는 장애물이다. 이러한 DIF Fi를 culties을 극복하기 위해, 우리는 최적화 장기 수익 DRL 동적 가격 결정을위한 프레임 워크를 제안한다. 본 논문은 여러 가지 공헌이있다. 우선, 우리는 Fi를 이산 및 연속 가격 문제 실제 전자 상거래 플랫폼보다는 Q-학습을 사용하여 기존의 이산 한 모두 (에스 트레 외. (2018)) 또는 내 시뮬레이션 환경을 DRL을 적용 RST된다. 둘째, 수익 전환율과 차이가 오히려 수익보다 보상 함수로서 더 적합하다는 것을 발견 (에스 트레 외. (2018), Kim 등. (2016), SCHWIND 및 벤트 (2002))로 인해 볼록 자연. 셋째, 거의 가정을 만들어 : 수요 함수는 드 미리 네드 Fi와 반드시시 불변 아니다; 수익 기능은 반드시 볼록 또는 동굴 CON-되지 않는다; 고객의 행동 전략이 될 수 있습니다. 마지막으로, 우리는 우리의 DRL 모델이 가격의 SKU 제품의 수천 개월 동안 지속되는 대규모 과학 ELD 실험을 실시하고 있습니다. 우리 Fi를 실험으로 ments는 성취와 이런 종류의 작업을 처음 Fi를입니다 전자 상거래 플랫폼에서 역동적 가격 프레임 워크의 효과를 증명 ELD.

본 논문은 다음과 같이 남아있는이 구성되어 다음 섹션 목록 동적 가격 문제에서 일부 관련 작업을. 3 절을 소개합니다 우리가 문제가 마르코프 의사 결정 프로세스 모델로 모델링 동적 가격 책정을 위해 설계 접근한다. 두 이산 가격 행동 모델과 지속적인 가격 행동 모델을 제안한다. 섹션 4에서, 폴로리다 오프라인 및 온라인 실험 모두에서 결과는 문제의 우리의 접근 방식을 검증하는 소개합니다. 결론 및 향후 작업 방향은 섹션 5에 요약되어있다.

2 L ITERATURE 검토

많은 연구는 수십 년 동안 역동적 인 가격에 완료되었습니다. 우리는 최근의 발전에 대한 종합적인 검토를 위해 (2015) 보어 덴을 참조하십시오. (1) 통계적 학습, 구체적인 캘리 추정 수요와 (2) 가격 최적화에 문제에 적용 :이 두 연구 과학 필드들을 결합한다. 이전 연구의 대부분은 가격과 수요 사이의 함수 관계가 의사 결정자에게 알려진 것으로 가정하는 경우에 초점을 맞추고있다. 쿠 르노 (1897) 수학적으로 제품의 가격 수요의 관계를 설명하고 최적의 수익을 달성하기 위해 수학 문제를 해결하기 위해 처음 Fi를로 인정 받고 있습니다. 그러나, 관계는 일반적으로 현실에서 진정한 보유하지 않는 정적 인 이상 시간이라고 가정. (1924) 에반스, 수요가의 가격의 함수뿐만 아니라 시간 유도체 가격뿐만 아니라 것으로 시간이 지남에 따라 가격의 동적 수요 함수로 이어지는. Kamrad 등. 가격을 최적화하는 동시에 (2005) 캡처 수요의 불확실성을 확률 모델을 소개했다. 레고 & 반 Ryzin (1994) 제한 재고와 과학 무한 계획 기간 같은 고려 제약.

실제로, 종종 사전에 수요를 설명하는 DIF Fi를 승배입니다. 많은 최근의 연구는 알 수없는 수요 함수와 동적 가격에 초점을 맞추고 있습니다. 연구원 좋은 첫 번째 매개 변수 접근하여 문제를 해결. Bertsimas 및 Perakis (2006) 수요 func-의 TIONS의 파라 메트릭 가족이 시간에 배운 것으로 가정. 파리 아스 & (2010) 반 로이 역사 구매 데이터로부터 배울 수있는 방법을 제안했다. 해리슨 등. (2012) AD- 드레스 수요 불확실성 베이지안 동적 가격 정책을 이용했다. 그러나 수익은 최적의 인해 잘못 지정하는 수요 가족에 출발 할 수있다. 따라서, 많은 최근의 연구는 주로 주변 비모수 AP-proaches을 돌아 가지. 베스 베스 및 지비 (2009), 및 베스 베스 지비 (2015), 왕 등. 수입이 접근하면서 (2014 년) 깊숙이 학습을 보였다; 왕 및 등. (2014) 지금까지 가장 작은 차이를 가지고 발표했다. 그러나 그들은 모두 수익 기능은 3.1에서와 같이 전자 상거래 소매 업계에서 진정한 보유하지 수있는, 엄격하게 오목 및 미분이다 가정.

계산의 발전에 따라 강화 학습 (RL)이 문제 NAMIC 따라 동적 주소로 도입된다. Kephart 등. (2000) 표현 Q는 학습 사용할 수있는 가능성을 보여 주었다

소위 pricebot을 형성 가능한 가격에 대한 예상되는 미래의 할인 프로 파이 TS는 시장 상황의 변화에 따라 가격을 조정합니다. SCHWIND & (2002) 벤트는 수율 관리보기에서 정보 제품의 동적 가격에 대한 시간적 차이를 이용했다. 라주 등. (2003) formu- lated 하나의 판매자와이 판매자 시뮬레이션 된 상황에서 가격 문제와 고용 다른 RL 알고리즘 동적. Kutschinski 등. (2003)은 시장 시나리오에서 경쟁력있는 가격 전략을 결정하기 위해 비동기 멀티 에이전트 RL 방법의 다른 유형을 사용했다. Vengerov (2007)과 김 등. (2016) 활용 강화는 에너지 시장에 최적화 가격에 학습. 에스 트레 등. (2018)는 시뮬레이션 환경의 공평성을 향상시키면서 수익 유지할 신경망 근사치와 Q가 학습 제안. 이러한 모든 이전의 작품, 그러나 Fi를 드 보상이 아니라 수익 워크 아웃과 네드 단순화 Fi를 에드 시장 설정과 시뮬레이션을 수행하고 있습니다. 그리고 DNNs는 이산 가격 approximators로 사용된다. 그것은 실제 시장의 경우에는 해당되지 않습니다.

3 M ETHODOLOGY

우리는 지금 우리가 위에서 설명한 동적 가격 문제에 대한 의사 결정 모델을 구축하는 방법을 고려한다. 우리 Fi를 첫 번째는 마르코프 결정 프로세스 (MDP) 등의 문제를 나타냅니다. 에이전트는 정기적으로 환경 상태를 관찰 한 후 그 작업으로 제품의 가격을 변경합니다. 새로운 환경 상태는 관찰 할 수 있고 보상도받을 수있다. 제품이 품절 인 경우 각각의 가격 에피소드의 끝에 도달. 이 모델은 역사의 판매 데이터도 FL 오프라인 평가에 사용되는 이전 전문가 '가격 액션에 의해 사전 훈련을한다. 프레임 워크는도 1에 도시되어있다.

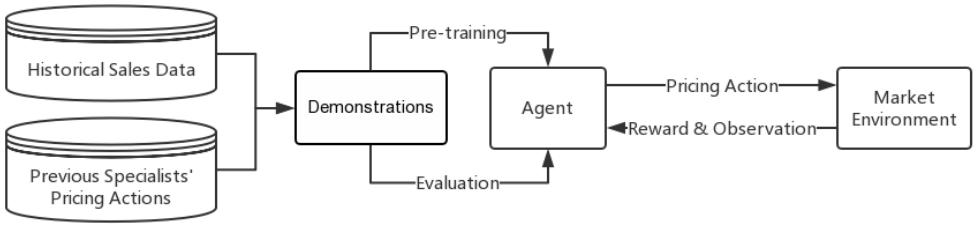


그림 1 : 전자 상거래 플랫폼에서 시위 DRL을 사용하여 동적 가격 프레임 워크.

3.1 P ROBLEM 에프 ORMULATION

우리는 주로 이름, 전자 상거래 플랫폼을위한 동적 가격 응용 프로그램 두 종류의 고려
인하 가격 과 매일 가격 이 작업의 나머지 부분입니다. 가격 애플리 양이온이 두 종류의 전자 상거래 플랫폼에서 가장 역동적 인 가격 책정 시나리오를 커버 할 수있다. 인하 가격 및 일일 가격 모두를 들면, 우리 드 Fi를 네브라스카 *엔* 제품에서 표시 *난* = 1, 2, ..., *엔* 개별적으로. 가격은 수정 된 과학 에드로 결정 또는 이산 시간 단계에서 유지 *t* = 1, 2, ..., *T,* 이 시간 단계 사이에 가격이 Fi를 될 것입니다 동안 Kutschinski 등 시장의 모델을 참조하여 고정 된. (2003), 마드 하반 (2000)와 오하라 (1995). 두 시간 **단계 사이의 거리는 hyperparameter 의해 드 인터넷이다 NE *디.***

인하 가격을 위해 공급은 밖으로 재고의 경우 매일 가격에 공급이 무제한으로 간주되는 동안 특정 제품에 대한 가격 결정 과정, 그 끝에 도달 있도록 제한됩니다. Kim 등의 전자 상거래 플랫폼에서 동적 가격 결정 과정은 마르코프 결정 프로세스 **(MDP)로 형성된다 참조 작품. (2016) 등 Vengerov (2007), 라주. (2003) 등의 각각의 시간 스텝 *티,* 가격 에이전트는 관찰 *에스*그것 제품의 상태를 설명 *나는,* 하고 조치를 취합니다 *에이*그것. 그러면 상담원은 보상 받는다 *아르 자형*그것새로운 상태에 대한 해당 작업뿐만 아니라 관찰 *에스*$_{i,t+1}$. 이 네 개의 요소 (전이 형성 *에스*그것, *에이*그것, *아르 자형*그것, *에스*$_{i,t+1}$), 이는 (인터넷과 같은 단순화 될 수 ED *S, A, R, S*′).**

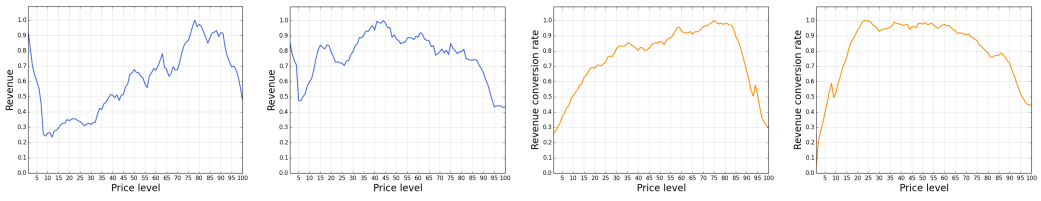
직관적으로 우리는 작은 원하는 *디*/ 시간에 반응 또는 연속 가격을 보관 가격 책정 작업을 확인합니다. 그러나 정확하게 환경의 변화를 설명하기위한 소정의 관찰 기간을 필요로 할 수있다. 빠르게 가격을 변경하면 깨질 수 *가격 이미지*/ 제품에 대한, 심지어 전자 상거래 플랫폼에서 신용 문제의 원인. 우리의 실험은 온라인 가격을 바꿀 것 때문에, 우리는주의 깊게 전문 가격 **매니저와의 토론 후이 기간을 설정합니다. 이 작업의 나머지 부분에서, 기간 가격 *디*/ 일일으로 정착된다. 따라서, 시간 간격 *티*/ 또한 일을 나타냅니다 *티*/.**

국가 공간. 우리의 모델에서 각 제품 여기에 *나는* 시간 단계에서 기능의 네 가지 그룹으로, 개별적으로 가격이 책정됩니다 *티*/상태를 설명하는 *에스 그것:* 가격 기능, 판매 기능, 고객 트래픽 Fi를 다 기능과 경쟁력 기능을 제공합니다. 가격 기능은 판매 기능 Fi를 C 기능은 시간을 제품의 페이지를 포함하는 트래픽 판매량, 매출 등 고객을 포함하는이 제품에 대한 실제 지급, 할인 요금, 쿠폰 **등을 포함 *나는* (PV) 확인되었습니다, 순 방문자 수는 제품 (UV), 제품에 대한 구매자의 수를 볼 *나는*, 코멘트와 유사한 제품의** 상태 등 경쟁력 기능에 기여한다. 일부 핵심 기능에 대한 설명은 (부록) 표 1에 제시되어있다.

작업 공간. 각 제품에 대한 작업 공간 NE 우리는 또한 드 Fi를 *나는* 갈라져. 우리는 최대 가격을 사용 *피난, 최대*최소 가격 *피 나는 ~ 많이있다* 제품의 *나는* 드 인터넷 NE에 기록 기간들 중 특정 수 동안은 상한과 하한. 그것은 가정이 가격 프레임 워크이 지역 밖으로 출력하지 가격해야한다. 가격 책정 공간은 분리 또는 다른 응용 프로그램의 연속이 될 수 있습니다. 이 분리 된 경우, 각 작업은 가격 범위를 의미합니다.

보상 기능. 우리는 드 Fi를 NE 즉각적인 보상에 여러 가지 방법을 비교 *아르 자형 그것* 수익은 우리가 위에서 언급 한 바와 같이, 트래픽 Fi를 C 고객이 분명히 forewarn- ING없이 변경 후 매출 영향력 플로리다 수 적합하지 않습니다. 따라서, 기존의 소매 업계에서 같은 가격과 수익 사이에 명확하고 설명 할 관계가되지 않을 수 있습니다. 드 Fi를 닝 보상 기능을 사용하여 수익은 **에이전트를 오해하고 비 융합으로 끝낼 심각 할 수있다. 그러나 전자 상거래 플랫폼, 가격 및 수익 전환율 사이의 링크가있다 *아르 자형t = 수익 그것 자외선 그것,* 어디 *자외선 그것*재현 된 하느님 분개는 순 방문자 수는 제품을 볼 *나는* 시간 단계 사이 *티*과 *티* - 1. 이 작품의 일부, 우리는 또한로 나누어, 프로 파이 t 전환율을 사용하여 *유 그것,* 우리는 재고 비용의 지식을 가지고있는 경우.**

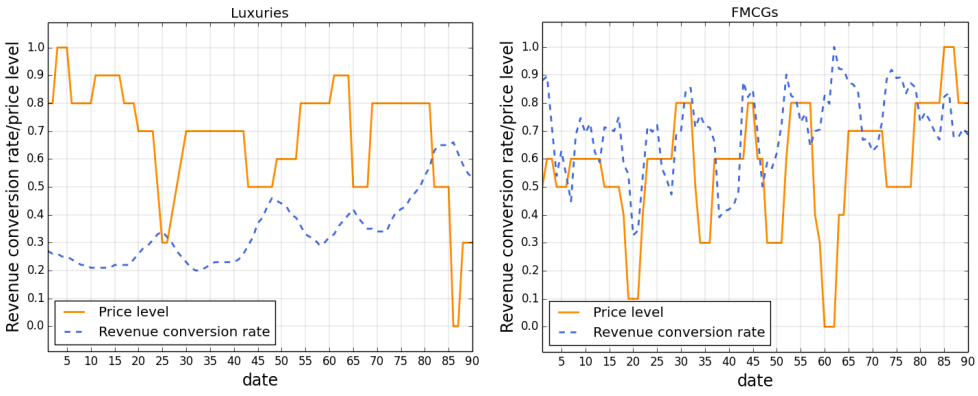
이 아이디어를 증명하기 위해, 우리는 수입과 다른 가격에 자신의 전환율의 분포에 대한 제품의 다른 종류를 분석 할 수 있습니다. 여기, 우리는 SKU를 샴푸의 그림이 3 개월 이내에 다른 가격 수준에서의 평균 수익 및 수익 전환율에 대해, Tmall.com에 판매하는 사탕의 SKU를 4800 이상에 걸쳐 3300의 결과를 보여줍니다.



(에이) (비) (씨) (디)

도 2 (a) 및 (b)는 3 개월 이내 1 내지 100 다른 가격 레벨에서 각각의 SKU 샴푸 3300 사탕 4800 개 SKU에 대한 평균 수익을 나타낸다. 가격 레벨 1은 가장 낮은 3 개월 이내 가격과 가격 수준 최고 100 스탠드를 의미합니다. (c)와 (d)는 각각 샴푸 사탕 평균 수익 전환율을 나타낸다. 매출 수익 전환율에 대한 수치 값은 최대 값으로 나누어 rescalled된다.

수익 전환율 수익 자체보다 더 볼록이고,도 2에 도시 된 바와 같이. 온라인 실험에서 수익 전환율을 사용하여 보상 기능이 작동하기 때문에 Fi를 북동에서 가격 인하 가격 AP-주름 성형술 때 가격 결정 프로세스의 라이프 사이클을 결정하는 매우 명확하고 정확한 재고가 1); 2) 가격 인하 제품의 대부분은 낮은 판매 볼륨 사치품 갖는 낮은하지만 가격 sensi-적인 수익 전환율이다. 그러나이 작품의 또 다른 가격 애플리케이션에서 매일 빠른 고객 제품 (FMCGs)를 이동 가격, 공급이 적절하고 주가는 무제한으로 간주 될 수있다. 각각의 가격 결정 과정에 대한 명확한 평가 기준은 거의 드 Fi를 네드 수 없었다. 더 메신저 -는 portantly이 FMCGs의 평균 판매량은 가격 인하 가격 사치에 비해 훨씬 높다. 이 경우, 우리 Fi를 차 일정 기간을 통해 가격과 평균 수익 전환율의 관계에도 불구하고,이 관계는이 기간에 다른 시간에 꾸준히 유지하지 않을 수 있습니다. (그림 3) 90 일 수익 전환율과의 가격 수준의 동향에 대한 제품의 두 가지 다른 종류의 또 다른 연구는이 현상을 보여준다. 우리는 (a)는 즉, 이러한 사치품, 가격이 떨어질 때, 수익 전환율의 (b)에 FMCGs을 위해, 특히 하루에 25, 45, 65, 그러나 85 주위에 상승을 위해, 더이 그림 3에서 볼 수있다 (그림 3) 90 일 수익 전환율과의 가격 수준의 동향에 대한 제품의 두 가지 다른 종류의 또 다른 연구는이 현상을 보여준다. 우리는 (a)는 즉, 이러한 사치품, 가격이 떨어질 때, 수익 전환율의 (b)에 FMCGs을 위해, 특히 하루에 25, 45, 65, 그러나 85 주위에 상승을 위해, 더이 그림 3에서 볼 수있다 (그림 3) 90 일 수익 전환율과의 가격 수준의 동향에 대한 제품의 두 가지 다른 종류의 또 다른 연구는이 현상을 보여준다. 우리는 (a)는 즉, 이러한 사치품, 가격이 떨어질 때, 수익 전환율의 (b)에 FMCGs을 위해, 특히 하루에 25, 45, 65, 그러나 85 주위에 상승을 위해, 더이 그림 3에서 볼 수있다



(에이) (비)

도 3 : 사치품 2000 SKU에 대한 평균 수익 전환율 및 가격 레벨 (하위의 인터넷 gure (a))과 각각 구십일 통해 (서브 인터넷 gure (b)에서) FMCGs 4000 개의 SKU. 0 수익 전환율 가격 수준 모두 90 일이며 최대 값 1 스탠드를 통해 최소치를 나타낸다. 상관 관계 COEF Fi를 사치품에 대한 cients 및 FMCGs는 각각 -0.57과 0.15이다.

같은 관계지만, 일주일에 주위 개별 주파수 수익 전환율 FL uctuates. 따라서 이러한 FMCGs에 대한 보상 기능으로 사용 수익 전환율, 모델의 융합을 보장 할 수 없습니다. 도 2 및도 3에 두 현상을 비교하면, 우리 드 인터넷 NE 다른 보상 기능 식으로 수익 전환율 (DRCR)의 차이를 이용. (1).

$$\text{아르 자형}_{t,T} = \frac{\text{수익}_{\text{그것}}}{\text{자외선}_{\text{그것}}} - \frac{\text{수익}_{\text{그것}-T}}{\text{자외선}_{\text{그것}-T}} \tag{1}$$

어디 T 수익 전환율을 비교하는 시간을 나타냅니다. 이 정의상 뒤에 아이디어는 우리가 성공적으로 가격 행동과 수익 전환율을 올릴 경우, 에이전트에게 긍정적 인 신호를 줄 수 있도록 노력하겠습니다 것입니다. 보상 기능을 해결할 수있는 문제의이 정의상 FMCGs 매일 판매 수렴 문제는 잠시 또한 가격 인하 가격에 과학 NE를 해결할 수 있습니다.

3.2 D ISCRETE 피 RICING 에이 CTION 엠 ODELS

이 부분에서 우리는 우리가 우리 드 Fi를 NED 위의 동적 가격 MDP를 해결하기 위해 사용하는 모델을 소개하기 시작합니다. 우리 Fi를 처음 사용 Q-학습 (왓킨스 (1989)) Fi를 차 최적의 정책. Q 학습 최적 정책을 계산하는 값 반복 방법이다. 그것은 **무작위로 이니셜로 시작 $큐$ 값 전환을 사용하여 다시 반복 cursively $t = (S, A, R, S')$ 최적를 얻을 수 $큐^*$뿐만 아니라 최적의** 정책으로

$$큐_{t+1}(S, A) \leftarrow (1 - \alpha) \cdot 큐_{타}(S, A) + \alpha \cdot [R + \gamma \cdot \underset{에이 \cdot 큐_{타}(에스', 에이')}{\text{최대}}] \tag{2}$$

이리 $\alpha \in (0, 1]$ 학습 속도입니다. γ 할인 요인이다. 인해 상태 공간의 높이 치수에, 우리는 깊은 Q-네트워크 (DQN, Mnih 외. (2015))의 개념을 따르는 상태 공간에서 Q-값을 매핑 깊은 네트워크를 사용한다. 액션 값 네트워크를 업데이트하려면, 한 단계 오프 정책 평가는 손실 함수를 최소화하는 데 사용됩니다 :

$$L(\theta) = \text{이자형}_{(S,A,R,S')\sim 다}[R + \gamma \cdot \underset{에이' Q(S', 에이', \theta') - Q(S, A \mid \theta)]^{이}}{\text{최대}}] \tag{삼}$$

이리 D/리플레이 버퍼 작업에 포함 된 전환을 통해 분배된다. θ 질문 네트워크의 파라미터이며 θ'타겟을 계산하는 데 사용되는 네트워크 파라미터이다. 타겟 네트워크 파라미터 θ'오직 Q-네트워크 파라미터마다 단계 C로 갱신된다. 이러한 DQN 경험 재생 분리 타겟 네트워크 fromMnih 등을위한 두 가지 기술이있다. (2015). 이 별도의 조치 방법을 적용하기 위해, **우리는에서 가격 공간을 분할 $피_{나는 \sim 안에있다}$에 $피_{난, 최대}$으로 $케이$**

분리 영역뿐만 $케이$개별 행동. 가격 $피_{난\ 케이}$satis 좋은 불평등을 말이지 :

$$피_{나, 분 + (피_{난, 최대} - 피_{나는 \sim 안에있다})/케이} \cdot (케이 - 1) \leq 피_{나, K} < 피_{나, 분 + (피_{난, 최대} - 피_{나는 \sim 안에있다})/케이} \cdot 케이 \tag{4}$$

를 선택하는 것으로 간주됩니다 $케이$일 ($케이 \in [1, 케이]$) 제품에 대한 조치를 가격 $나는$.

3.3 C ONTINUOUS 피 RICING 에이 CTION 엠 ODEL

개별 작업 공간에서의 가격은 명백한 갈등 이산 AC-TIONS의 수를 설정하면 hyperparameter 발생 *케이.* 만약 *케이* 큰 가격 영역은 동일한 가격으로 너무 작은 간주 될 것입니다. 동시에, 별도의 작업 공간 알고리즘 있었던 총 작업 공간이 몇몇 부분 **공간으로 분할 될 수 있다면 불분명 가격 작업 공간의 출력 서브 스페이스. 한편, *케이* 행동의 많은 역사에서 너무 큰 탐구되지** 않습니다되고 미래의 탐사도 INEF 과학 효율적인 수 있습니다. 따라서, 우리는 출력에 계속하지만 uous 공간에 정확한 가격 대신에 가격 영역 모델의 가격을 구축하는 것이 좋습니다. 우리는 가치 반복 방법과 정책 반복 방법을 결합하고 제안 된 다른 문제에서 잘 수행 된 (위튼 (1977)), (Vamvoudakis & 루이스 (2010), Mnih 외. (2016) 배우 - 비평가 알고리즘을 적용). 캘리 구체적인, 우리는 우리 배우 비평가 방법으로 깊은 결정 정책 그라데이션 (DDPG, Lillicrap 외. (2015))에 적용됩니다. 이 **모델의 배우 부분은 정책 네트워크를 유지하고 *π (A | S, θμ)* 입력 및 출력으로 연속 작업 환경 상태를 고려 *A = μθ(에스).* 그리고 비평 부 입력으로 상태 및 동작 모두를 받아 동작 값 함수를 추정 *Q (S, A | θ큐). θμ*과 *θ큐* 네트워크 매개 변수입니다. 그래서,** 손실 함수는 다음과 같습니다

$$L(\theta) = \text{이자형}_{s,A,R,s_1 \sim D}[R + \gamma \cdot Q(S'; \mu(\text{들}_1 \theta_{\mu_1}), \theta_{\text{큐}_1}) - Q(S, A \mid \theta_{\text{큐}})]^{\text{이}} \quad (5)$$

그래서, 우리는 또한 경험의 재생과 별도의 대상 네트워크 기술을 적용 이곳까지 *θμ*·과 *θ큐*·각각 배우 비평가위한 타깃 네트워크 파라미터이다. 그리고 네트워크 정책을 갱신하기 위해 Q 값의 기울기를 취

$$\nabla_{\theta_\mu} \mu \approx \text{이자형}_{\mu_1} \nabla \text{에이} Q(S, A \mid \theta_Q)|_{A=\mu(\text{들})} \nabla_{\theta_\mu} \mu(\text{들} \mid \theta_\mu)] \quad (6)$$

아이디어의 뒤에은 매개 변수를 조정하는 것입니다 *θμ* perfor- mance 구배 이러한 알고리즘되는 기본 정책 그라데이션 정리 방향의 정책 네트워크 (서튼 외. (2000)). 이러한 방식으로, 우리의 모델의 배우 것이라고 출력 SPECI 아니라 출력 가격 영역보다 연속 가격 작업 공간 Fi를 C 가격, 모델의 비평가는 DRL 모델의 효율성과 정확성을 향상시킬 수있는이 구체적인 C 조치를 평가하는 반면 .

3.4 P RE-교육

우리가 직접 전자 상거래 동적 가격에 강화 학습 알고리즘을 적용하면, 그들은 매우 저조한 실적으로 시작, 콜드 스타트 문제를 충족하고 자본 손실이 발생할 수 있습니다. 로봇 레빈 등과 같은 다른 지역에서는. (2016) 및 게임 실버 등. (2016), 에이전트가 정책을 배울 수있는 내 정확한 시뮬레이터,있을 수 있습니다. 그러나 동적 가격 문제에 대한 이러한 시뮬레이터는 없다. 대신, 우리는 충분한 환경의 데이터와 일부 이전 컨트롤러에 의한 가격 결정을해야합니다. 이 컨트롤러는 일부 전문가 또는 몇 가지 규칙이 될 수 있으며, 그 가격 결정의 일부는 합리적인 될 수 있습니다. 환경에 직면 그 결정의 레코드 Sendonaris 및 Dulac-아놀드 (2017)에서 에이전트가 사전 훈련 효과적인 것으로 입증되었다 데모로 간주 될 수있다.

앞서 언급 한 바와 같이, 가격 책정 작업을 주기적으로 수행되었다. 따라서, 내 환경, 표준 상태 잘 보상은이 기간 BE- 트윈 **동작에서 수집 된 데이터에 의해 표현 될 수있는만큼. 따라서, 우리는 튜플의 데모를 형성 *< 에스 티, 에이 티, 아르 자형 티, 에스 t +1>* 사전** trianing 위해 그들을 사용합니다. 구체적인는 캘리, 우리는 데모에서 DQN과 깊은 결정적 정책 그라데이션에 대한 우리의 사전 교육 방법 (DDPGfD) Vecer'ık 등으로 깊은 Q-학습 데모에서 (DQfD) Sendonaris 및 Dulac - 아놀드 (2017)의 아이디어를 참조하십시오. (2017)에 대한 DDPG.

3.5 O FFLINE 이자형 평가 엠 ETHODOLOGY

우리가 위에서 논의한 바와 같이, 우리는 온라인 가격하기 전에 미리 훈련 도중 시위와 모델을 평가해야합니다. 온라인 평가를위한 방법은 4.2 절에서 자세히 논의 될 것이다 동안 WWE는이 부분에서 플로리다 오프라인 평가를위한 방법론을 소개합니다. 우리는 Fi를 최신을 사용하여 실온
***티*/튜플에서 기록 기간에 *< 에스 티, 에이 티, 아르 자형 티, 에스 t +1> 티 ∈ [1, T].* 그리고, 우리는 두 부분으로 이러한 튜플을 분할 : D 튜플 처음 Fi를이 사전 교육에 사용됩니다 *티 ∈ [1, 디].* 그리고**
D <t <T, 튜플는 평가를 위해 사용됩니다. 일부 처음 fi의 다른 모델에 대한 보상을받을 수있는 능력을 평가하는 것입니다.
아이디어는 우리가 보상을 요약이다 *아르 자형* 경우에만 조치 *에이*/정책의 출력에 가까운 *에이* - *< π (S) <A + ?.* 의 세부 알고리즘 알고리즘 1 스케치 평가.

알고리즘 1 데모 튜플과 정책 평가.

입력: $T > 0$: 데모의 수는 평가를 위해 튜플; π : 정책 평가합니다;
산출: 아르 자형_{π} : 평균 보상 양식 정책 π;

```
1: R = 0, N의 = 0
2 : ...에 대한 단계 티 ∈ {1, 2, ..., 타} 하다
삼:     반복
4 :         다음 튜플을 가져 오기 < S, A, R, S >
5 :     ...까지 에이 - < π (S) <A + ?
6 :     아르 자형 ← R + R
7 :     앤 ← N + 1
8 : 대한 종료
9 : 만약 N> 0 그때 아르 자형π = R / N
10 : 그밖에 아르 자형π = 0
```

그런 다음 우리는 모델의 정확도를 평가한다. 분리 작업 공간 방법, DQN, 우리는 식을 사용한다. (2) 두 개의 상태 사이에서 **예상 즉시 보상을 계산할** $\text{에스}_{티}$**과** 에스_{t+1} **모델에서 다음 실제 보상과 비교** $\text{아르 자형}_{티}$**오류 속도를 얻을 수 있습니다** *이자형:*

$$E = | \text{이자형}[\text{아르 자형}_{타} + \text{에}[Q(\text{에스}_{t+1}, \text{에이}_{타}) - \gamma \cdot \text{최대}_{에이} Q(S_{t+1}, 에이) - \text{아르 자형}_{타}}{\text{아르 자형}_{티} \qquad \text{아르 자형}_{티}} | \tag{7}$$

. DDPG 식 용 동안 (7) 식으로 변경한다 (8). :

$$E = | Q(들_{타}, \text{에이}_{타}) - \gamma \cdot Q(S_{t+1}, \mu(들_{t+1})) - \text{아르 자형}_{타}}{\text{아르 자형}_{티}} | \tag{8}$$

우리는 일정한 임계 값을 설정 공지 사항 것을 $0 < \text{아르 자형}_{c} \ll 1$ **여기에 오류를 계산합니다. | 만약** $\text{아르 자형}_{티} < \text{아르 자형}_{씨}$**실제 보상이 너무 작, 다음 오류가 드 파이와 네드입니다** $E = | \text{이자형}[\text{아르 자형}_{타} - \text{아르 자형}_{t}$.

4 E XPERIMENTAL 아르 자형 ESULTS

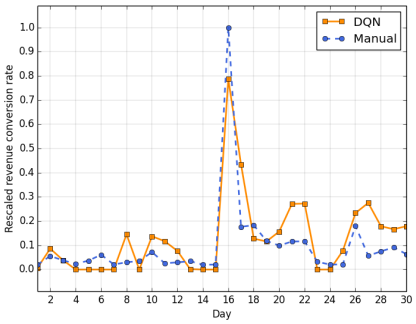
DQN 및 DDPG : 우리는 주로 동적 가격에 대해 위에서 소개 한 두 DRLmethods을 평가 하였다. 우리 Fi를 첫 번째는 Tmall.com의 데이터를 사용하여, 플로리다 오프라인의 실험을 소개합니다. 그리고 우리는 우리가 인하 시나리오와 매일에 Tmall.com에 제품에 대한 온라인 가격을 변경 온라인 실험 결과를 소개합니다.

4.1 O FFLINE 이자형 XPERIMENTS

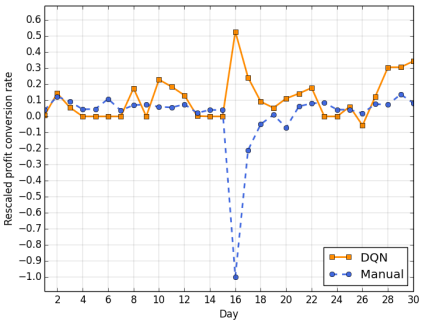
FL 오프라인의 정책 평가에서, 우리는 빠른 육십일는 총 2,400,000 튜플에 대해 형성하는 기록을 판매하는 소비자 제품을 이동하는 40,000의 SKU를 선택했다. 우리는 사전 교육과 마지막 날의 평가에 대한 시위로 오십구일 '기록 첫 Fi를 사용했다.
우리는 설정 $K의 = 10$, $\alpha = 0.01$, $\tau = 1$
점차 증가 γ **0.5 내지 0.99. DDPG 정책 평가를 위해, 우리는 = 설정 0.05. 그 결과를도 8 (부록) (a)에 도시되어있다. 우리는** 시위의 특정 번호 사전이-trainied 후 더 나은 DQN보다 DDPG 실시하는 것을 볼 수 있었다. 그런 다음 우리는 사전 훈련 기간 동안 서로 다른 모델의 정확성을 테스트하기위한 실험을했다. 3.5 절에 논의 된 정밀도에 대한 실험 결과 (부록)도 8 (b)에 나타낸다. 그것은 그 표시, DDPG 사전 훈련 도중 DQN보다 낮은 오류율을 가지고있다.

4.2 O NLINE 이자형 XPERIMENTS

인하 시즌 가격. Fi를 처음 우리는 주변에 고급 제품의 500 개 SKU를 Markdown을 시즌 동안 (주로 핸드백과 옷을) 가격에 대한 DQN을 적용했다. 각 제품은 재고 10 개 항목 주위에 가지고 있으며, 가격 인하 시즌의 목표는 프로 Fi를 t 전환율 및 매출 전환율을 모두 극대화하는 것입니다. 환경 FL 영향력에서 제외하는 기준으로 간주 될 수있는 같은 기간에 수동으로 가격
유사한 제품, 2000 SKU를 가진 다른 그룹이있다. 다음은이 실험의 우리의 세트 $D = 90$**, 이는 우리가 사전에 훈련 90 일을 기록** 데이터를 사용하는 것을 의미한다. 실험의 결과는이 온라인 실험 15 일 첫 Fi를 그림 4에 표시됩니다,이 제품은 일상 가격에 있었다. 그것은 두 개의 그룹을 보여줍니다



(에이)



(비)

그림 4 (a)와 (b)는 재 규격화 수익 전환율과 직접 가격 DQN에 의해 가격 제품과 제품을 비교하는 배율을 조정하고 프로 Fi를 t 전환율 플롯이다. 1 일부터 주 15, 제품은 정상 판매 가격에 있었다. DQN 그룹 및 수동 그룹 모두 0.04의 수익 전환율, 0.06의 프로 Fi를 t 전환율을했다. 일 30 DQN 그룹에서 하루 15에서 시작 및 종료 마크 다운 가격은 평균 및 수동 그룹의 전환율에 0.22의 수익 전환율 (1 슬캘링 일 16 수동 그룹의 수익 전환율)이 기간 동안 0.16을 달성했다. (b)에서, 인하 시즌, DQN 그룹의 프로 파이 t 전환율을 수득

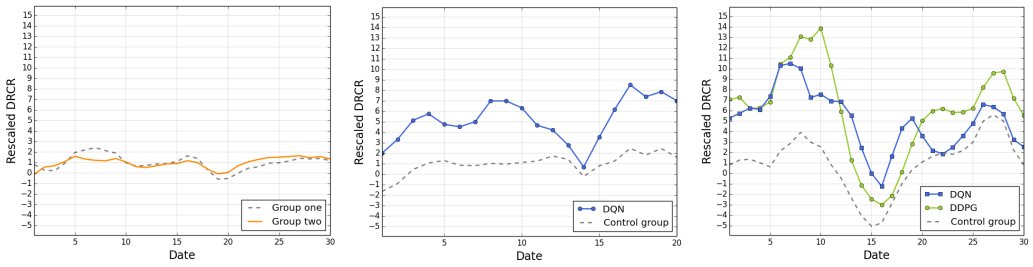
수동 그룹의 프로 파이 t의 전환율이 평균 -0.04로 하락하면서 0.16은 (주 16 수동 그룹의 프로 파이 t 변환 속도 배율을 조정 -1).

제품의 하루 매출이 모두 수행한다. 그런 다음 인하 시즌 16 일에 시작하여 15 일 동안 지속되었다. 우리는 두 그룹은 16 **일에서 수익 전환율을 밀어하고 다음과 같은 두 가지 일 급감 것을 볼 수 있었다. 그것은 소위에 의해 야기 될 수있다** ***위상 열 전*** *:*
하루나 이틀 제품은 활동의 가격으로 보여주는하지만 여전히 자신의 일상 가격에 판매되는 동안 활동 전에. 따라서, 일 **(16)라는 E- 상거래 활동에 매우 일반적인 현상을 보여** ***폭발성 위상*** **일반적으로 주요 수익 기여 활동의 시작 부분에. 그런 다음** 잘 날이 25에서 30으로 하루에 21, 22, DQN 그룹이 성공적으로 인해 총 수익 전환율을 극대화를 목표로 정책에 다시 수동 그룹을 제치고 매출 전환율을 끌어. 프로 Fi를 t 전환율 그래프 그림 4B와 비교, 더 그것을 분명하다, 수동 가격 메소드는 DQN의 가격 정책이 성공적으로 대부분하시오 긍정적 프로 파이 t 동안 수익을 증폭하면서 부정적인 프로 파이 t을 일으키는 원인이되는 가격이 비용보다 감소하여 수익을 뽑아 Markdown을 계절. 그것은 하루에 26에서, 그 재미, DQN는 부정적인 프로 파이 t 속도로 제품을 가격, 음의 즉각적인 보상을 얻었다. 그러나이 조치는 가격 인하 시즌의 나머지 수익 및 프로 Fi를 t 전환율을 모두 뽑아.

매일 판매 가격. 식의 보상 기능의 효과를 확인하려면. (1), 우리는 가격 공급 무제한 FMCGs에 AN-다른 온라인 실험을 설정합니다. 우리 Fi를 처음 드 Fi를 NED ***시미-제품*** **전자 상거래 플랫폼, 같은 브랜드, 같은 카테고리와 유사한 판매 행위와** 함께 제품. 우리는 시미 - 제품의 두 그룹이하는 서로 다른 총 수익을 경우에도, 같은 가격 정책을 사용하는 것을 발견, 자신의 DRCR 인해 시즌 FL uctuation와 플로리다에서 영향력의 제거에 (그림 5의 (a)) 매우 가까운 수 경영 전략에서. 따라서, 우리는 그들의 DRCR을 비교하여 서로 다른 pric- ING 정책을 평가할 수 있습니다. 우리는 또한 플랫폼에 판매 시미 - 제품의 (그림 5에서 그룹이 (가)) 우리의 실험 그룹으로, Tmall.com, 주로 식품, 스낵, 매일 화학 물질 FMCGs의 다음 일치하는 3000 개 SKU는 약 1000 개의 SKU를 선택 대조군 (도 5 (a)의 그룹 중 하나).

실험의 이십일 첫 Fi를, 우리는 대조군을 보내고 compar-하여 DQN의 가격 정책의 동작을 조사 하였다. 우리는 설정 ***D*** **= 30 일 FMCGs의** 행동은 고급 제품보다 더 빠르게 변화 '로 인하 가격보다 짧은. 다른 매개 변수는 오프라인 평가와 동일한 설정했다. 두 그룹의 20 일 이내 DRCR는도 5의 (b)에 나타낸다. 우리는 그것을 볼 수 있었다, DQN 그룹은 대조군보다 실적.

그런 다음 우리는 DQN 및 DDPG을 테스트하기 위해 두 그룹으로 무작위로 실험 그룹을 나누어 오프라인 평가와 같은 매개 변수를 사용했다. 결과는도 5의 (c)에 도시된다. 이 부분에서



(에이) (비) (씨)

그림 5 : 유사 제품의 다른 그룹에 대한 DRCR 비교. (a)는 30 일 이내 유사한 DRCR가 비슷한 제품의 두 그룹. 그룹 하나에 대한 평균은 배율을 조정한다
1.00 그룹 두 평균 스케일링 후에 0.99가된다. (b)이 20 일을 각각 DQN 대조군으로 저렴한 유사한 제품 그룹의 DRCR를 나타낸다. DQN 기의 평균은 1.00 재 스케일링 대조군의 평균으로 5.24이다. (c)에 도시 한 30 일 DDPG, DQN 대조군으로 저렴한 유사한 제품 그룹 DRCR. DDPG 및 DQN 가격표 그룹의 평균은 6.07이며, 배율을 조정 대조군의 평균은 각각 5.03

1.00.

(일부 쿠폰 제공과) 실험, 우리는 몇 가지 일상적인 경영 활동을 발생했습니다. 이 활동은보다 작 발생 10 % 총 수익 및 우리가 위에서 언급 한 DRCR의 현상 플로리다 uctuation은 여전히 관찰 할 수있다. 두 DRL 방법을 모두 상회 대조군 DDPG 더 나은 수행하는 동안.

(2013) 이브라힘를 참조하면, 우리는 훈련 신경망의 입력 상태 기능의 중요성을 평가했다. 우리는 우리가 함께 절 3 절에 도입 된 기능의 네 그룹에 대한 중요도 점수가 그림 6에 나타나있다 (3)에 도입 된 기능 모두에 대한 상대적 중요도를 계산하기 위해 연결 가중치 알고리즘 (옛날 & 잭슨 (2002)) 사용 1. 배율을 조정하고 가장 높은 점수 우리는 가장 낮은 구매자의 현재 판매량, 매출과 번호가있는 동안 가장 높은 상대적 중요성과 기능, 자외선 관련 기능 및 일부 관련 가격, 즉 볼 수 있었다. 판매 기능은 낮은 점수를 얻을 때 일반적으로 고객 트래픽 Fi를 다 기능과 가격 기능은 높은 점수를 얻을. 3 장에서는 조사와 비교, 그것은 우리에게 가격 결과에 FL 영향력의 기능에 대한 전자 상거래 플랫폼에서 동적 가격 문제에 대한 통찰력을 제공합니다. 흥미로운 현상은 매주 의견은 매월 의견보다 높은 중요성을 느끼는 동안 즉, 경쟁력 기능의 그룹에서 제품에 대한 월별 평균 점수는 주간 평균 점수보다 높은 중요성을 얻을 수있다. 페이지에 제품 전시 연계 점수가 매월 계산되기 때문에 고객의 대부분이 최신 의견을 읽을 수 있습니다하면서이다. 도 6의 입력 기능에 대한 설명은 (부록) 표 1에 제시되어있다. 페이지에 제품 전시 연계 점수가 매월 계산되기 때문에 고객의 대부분이 최신 의견을 읽을 수 있습니다하면서이다. 도 6의 입력 기능에 대한 설명은 (부록) 표 1에 제시되어있다. 페이지에 제품 전시 연계 점수가 매월 계산되기 때문에 고객의 대부분이 최신 의견을 읽을 수 있습니다하면서이다. 도 6의 입력 기능에 대한 설명은 (부록) 표 1에 제시되어있다.

5 C ONCLUSIONS 및 고찰

이 작품에서 우리는 E- 상거래 플랫폼에서 동적 가격에 대한 깊은 강화 학습 프레임 워크를 제안 하였다. 우리 마르코프 의사 결정 프로세스로 드 Fi를 NED 가격 결정 과정과는 드 Fi를 서로 다른 가격 응용 프로그램에 대한 상태 공간, 이산 및 연속 작업 공간 및 다른 보상 기능을 네드. 우리는 정책을 가격에 대한 우리의 방법을 적용하고 실시간으로 온라인 가격에 적용. 우리 Fi를 첫 번째는 가격 인하 시즌 제품 가격에 대한 깊은 보강 학습 방법을 적용 할 수 있습니다. 과학 ELD 실험은 수동 인하 가격 전략을 능가 것으로 나타났다. FMCGs 매일 가격으로, 우리는 다른 가격 전략을 테스트 A / B의 법적 문제를 해결하기 위해 온라인 가격 정책 평가를위한 체계적인 메커니즘을 설계합니다. 우리는 DDPG 및 DQN에서 가격 정책이 cantly 다른 가격 정책 유의 Fi를 상회 것으로 나타났다. 이 작품은 Fi를 실시간으로 제품의 SKU를 수천 가격, E- 상거래 플랫폼에서 동적 가격 문제에 대한 깊은 강화 학습을 사용하는 첫 번째입니다. 본 연구에서는 몇 가지 제약이 제거 될 수있다. 첫째, 우리의 가격 프레임 워크는 별도로 각 제품을 훈련한다. 그 결과, 낮은 판매 볼륨 제품은 SUF 과학 효율적인 훈련 데이터가 없을 수 있습니다. 이는 유사한 제품을 클러스터링과 같은 클러스터의 제품 가격을 학습 전송을 사용하여 해결할 수 있습니다. 메타 학습은이 문제에 대한 도움도 할 수있다. 둘째, 우리의 프레임 워크 출력에 대한 정책을 가격 그 결과, 낮은 판매 볼륨 제품은 SUF 과학 효율적인 훈련 데이터가 없을 수 있습니다. 이는 유사한 제품을 클러스터링과 같은 클러스터의 제품 가격을 학습 전송을 사용하여 해결할 수 있습니다. 메타 학습은이 문제에 대한 도움도 할 수있다. 둘째, 우리의 프레임 워크 출력에 대한 정책을 가격 그 결과, 낮은 판매 볼륨 제품은 SUF 과학 효율적인 훈련 데이터가 없을 수 있습니다. 이는 유사한 제품을 클러스터링과 같은 클러스터의 제품 가격을 학습 전송을 사용하여 해결할 수 있습니다. 메타 학습은이 문제에 대한 도움도 할 수있다. 둘째, 우리의 프레임 워크 출력에 대한 정책을 가격

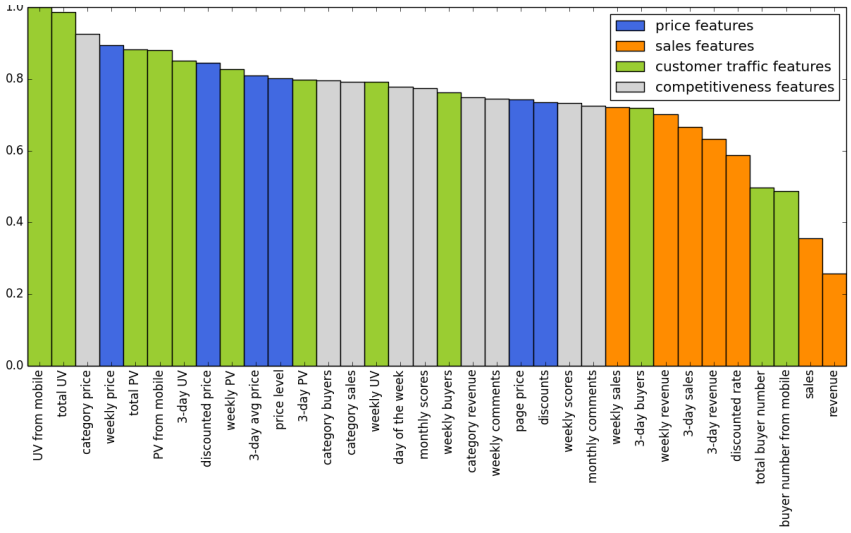


그림 6 : 일부 기능에 대한 슬캘링 중요성을 기록했다. 점수의 평균은 0.81, 0.56,
가격 기능, 판매 기능, 고객 트래픽 Fi를 다 기능과 경쟁력을위한 0.79와 0.70은 각각 있습니다.

각 제품은 개별적으로, 그러나 때때로 우리는 어떤 마케팅 전략을 형성하기 위해 함께 다양한 제품 가격을
책정하도록하겠습니다. 이는 조합 활동 공간에 의해 해결 될 수있다. 셋째, 우리의 가격 프레임 워크에, 우리는 환경 상태를
설명하는 데에만 제품과 관련된 기능을 가지고. 미래에, 우리는 더 많은 구체적인 C 시나리오, 예를 들어, 프로모션 가격이나
회원 가격에서 가격에 대한 고려 기능의 종류 이상을 시도합니다.

아르 자형 EFERENCES

디미트리 Bertsimas 조지아 Perakis. 동적 가격, • 학습 방법. 에서 수학
정체에 대한 계산 모델은 쪽을 충전. 45-79을. 스프링, 2006 년 오마르 베스 베스와 아사 프 지비. 수요 함수를 모르고 동적

가격 : 위험 경계
가까운 최적 알고리즘. 운영 연구, 57 (6) : 1407-1420, 2009 년 오마르 베스 베스와 아사 프 지비. 동적 가격에 대한 선형

모델 (놀라게) SUF Fi를 결핍증에
수요 학습과. 경영 과학, 61 (4) : 723-739, 2015 년 펠리페 오목 및 J'

er'emie GALLIEN. 패스트 패션 소매 업체 재고 정리 가격 최적화.
운영 연구, 60 (6) : 1,404에서 1,422 사이, 2012.

르 첸, AlanMislove 및 ChristoWilson. 아마존에 알고리즘 가격의 실증 분석
시장. 25 월드 와이드 웹, PP에 대한 국제 회의의 절차에서. 1339-
위원회 2016 핸들 1349 국제 월드 와이드 웹 컨퍼런스.

M 키스 첸과 마이클 셀던. 노동 시장에서의 동적 가격 : 서지 가격과 플로리다 융통성이
둥네 짱 플랫폼에서 작동합니다. EC, PP. 455, 2016 년.

앙투안 오거스틴 쿠 르노. 웰스의 이론의 수학적 원리에 대한 연구.
맥밀란, 1897.

아 누드 V 덴 보어. 동적 가격 및 학습 : 역사적 기원, 현재의 연구, 새로운
지도. 작업의 조사 연구 및 경영 과학, 20 (1) : 1-18, 2015.

GC 에반스. 독점의 역학. AmericanMathematical 월, 31 (2) : 77-83, 1924 ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2300113.

비벡 F 파리 아스와 베냐민 반 로이. 시장의 반응에 앞서 동적 가격. 운영
연구, 58 (1) : 16-29, 2010.

기예르모 레고와 가렛 반 Ryzin. 확률과 재고의 최적의 동적 가격
과학 무한 지평에 걸쳐 요구한다. 경영 과학, 40 (8) : 999-1020, 1994 J 마이클 해리슨, N 보라 Keskin, 그리고 아사 프 지비.

베이지안 동적 가격 정책 : 학습
이진 사전 분포에 따라 적립. 경영 과학, 58 (3) : 570-586, 2012 OM 이브라힘. 입력 변수의 상대적 중요성을 평가하기위한

방법의 비교
인공적인 신경망. 응용 과학 연구 논문집, 9 (11) : 5692-5700, 2013 Bardia Kamrad, Shreevardhan S 렐레, 악 타르 시디,

로버트 J 토마스. 혁신 확산
불확실성, 광고 및 가격 정책. 운영 연구의 유럽 저널, 164 (3) : 829-850, 2005.

제프리 O Kephart, 제임스 E 핸슨, 에이미 R는 Greenwald. 소프트웨어 에이전트에 의해 동적 가격.
컴퓨터 네트워크, 32 (6) : 731-752, 2000.

이 병 노랭 김, 유 장, 미하엘 반 데르 SCHAAR, 그리고 장 - 원 리. 동적 가격과
강화 학습과 에너지 소비 예약. 2,187에서 2,198 사이, 2016 : 스마트 그리드, 7 (5)에 IEEE 거래.

에리히 Kutschinski, 토마스 Uthmann, 다니엘 Polani. 경쟁력있는 가격 전략을 학습
멀티 에이전트 강화 학습에 의해. 2,207에서 2,218 사이, 2003 년 경제 역학 및 제어, 27 (11 ~ 12)의 저널.

세르게이 레빈, 첼시 핀, 트레버 대럴, 그리고 피터 Abbeel. 깊은 visuo-의 엔드 - 투 - 엔드 교육
모터 정책. 기계 학습 연구 논문집, 17 (1) : 1,334에서 1,373 사이, 2016 년 디모데 P Lillicrap, 조나단 J 헌트, 알렉산더

Pritzel, 니콜라스 Heess, 톰 Erez, Yuval 교수 Tassa,
데이비드 실버 및 Daan Wierstra. 깊은 강화 학습과 연속 제어 할 수 있습니다. arXiv 프리 프레스 arXiv : 1509.02971 2015
년.

Ananth 마드 하반. 시장 미세 구조, • 조사. 과학 재무 시장의 학회지, 3 (3) : 205-258,
2000.

로베르토 에스 트레, 후안 듀크, 알베르토 루비오, 후안 Ar' evalo. 강화 공정을위한 학습
동적 가격. arXiv 프리 프레스 arXiv : 1803.09967, 2018.

볼로디미르 Mnih, Koray Kavukcuoglu, 데이비드 실버, 안드레이 루수, 조엘 Veness, 마크 G Belle-
암말, 알렉스 그레이브스, 마틴 Riedmiller, 안드레아스 K Fidjeland, 게오르그 Ostrovski는 등. 깊은 강화 학습을 통해 인간
수준의 제어. 자연, 518 (7540) : 529, 2015 볼로디미르 Mnih, 아드라아 Puigdomenech 바디 아, 메디 미르자, 알렉스
그레이브스, 디모데 Lillicrap, 팀
할리, 데이비드 실버 및 Koray Kavukcuoglu. 깊은 보강 학습을위한 비동기 방법. 기계 학습, PP. 1928-1937, 2016 JACK
Nicas에 관한 국제 회의에서. 지금 가격은 분에서 분으로 변경할 수 있습니다. 월스트리트 저널, 2015 년 모린 오하라. 시장
미시 구조 이론, 볼륨 (108) 블랙웰 출판사 캠브리지,
MA, 1995.

줄리안 D 옛날과 도널드 잭슨. 무작위 접근 : 블랙 박스 조명
인공적인 신경 네트워크의 이해 변수 기여. 생태 모델링, 154 (1-
2) : 135-150, 2002.

CVL 라주, Y Narahari 및 K Ravikumar. 강화 동적 가격이 응용 프로그램을 학습
소매 시장. 전자 상거래에서 2003 년 CEC, PP에 2003 IEEE 국제 회의. 339-
(346) IEEE 2003.

마이클 SCHWIND와 올리버 벤트. reinforce-에 따라 정보 제품의 동적 가격
, 표준 학습 : 수율의 관리 방법. 인공적인 지능, PP. 51-66에 연례 회의에서. 스프링 2002.

앤드류 Sendonaris 및 COM 가브리엘 Dulac - 아놀드. 현실 세계에 대한 시위에서 학습
강화 학습. arXiv 프리 프레스 arXiv : 1704.03732 2017.

데이비드 실버, 아자 황, 크리스 J Maddison, 아서 Guez, 로랑 Sifre, 조지 반 덴 Driessche,
줄리안 Schrittwieser, 요안 Antonoglou, 베다 Panneershelvam, 마크 Lanctot, 등. 깊은 신경 네트워크와 트리 검색과 이동의
게임을 마스터. 자연을, 529 (7587) : 484, 2016 년 리처드 S 서튼, 데이비드 McAllester, Satinder P 싱, 그리고 Yishay
만수르. 정책 gradi-
기능 근사치 강화 학습을위한 ENT 방법. 신경 정보 처리 시스템의 발전에 쪽. 1,057에서 1,063 사이, 2000.

칼리 T Talluri과 개렛 J 반 Ryzin. 이론과 수익 관리의 실천, 부피
매화 (68) 스프링 과학 비즈니스 미디어, 2006.

키리아 코스 G Vamvoudakis와 프랭크 L 루이스. 온라인 배우 비평가 알고리즘은 continuous-를 해결하기 위해
과학 무한의 지평선 최적 제어 문제의 시간. AUTOMATICA, 46 (5) : 878-888, 2010 마테이 Vecer'ık, 토드 헤스터, 조나단
솔츠, Fumin 왕, 올리비에 Pietquin, 빌랄 Piot, 니콜라스
Heess, 토마스 로스 ORL, 토마스 램프;, andMartin ARiedmiller. 대한 활용 데모
스파 스 보상과 로봇 공학 문제에 학습 깊은 강화. CORR, ABS / 1707.08817,
2017.

데이비드 Vengerov. 부분 - 동적 가격에 그라데이션 기반 강화 학습 방법
관찰 환경. 2007.

Zizhuo 왕, SHIMING 덩 샤오핑, 그리고 인유 예. 닫기 간격 : 학습-동안-하고 알고리즘
단일 제품 매출 관리 문제. 운영 연구, 62 (2) : 318-331, 2014 년 크리스토퍼 존 콘월어 Hellaby 왓킨스. 지연 보상에서 학습.

박사 학위 논문, 왕의
대학, 캠브리지, 1989.

로렌스 R 웨더 사무엘 E 신체. perishable-의 분류 및 연구 개요
자산 수익 관리 : 수율 관리, 초과 예약 및 가격. 831-844, 1992 : 조작, 40 (5) 연구.

이안 H 위튼. 이산 시간 마르코프 환경을위한 최적의 적응 제어기. 정보
및 제어, 34 (4) : 286-295, 1977.

AA 뉴 부가 이자형 XPERIMENT 아르 자형 ESULTS

여기에서 우리는 4.1 절에서 소개 된 실험 결과를 제시한다.

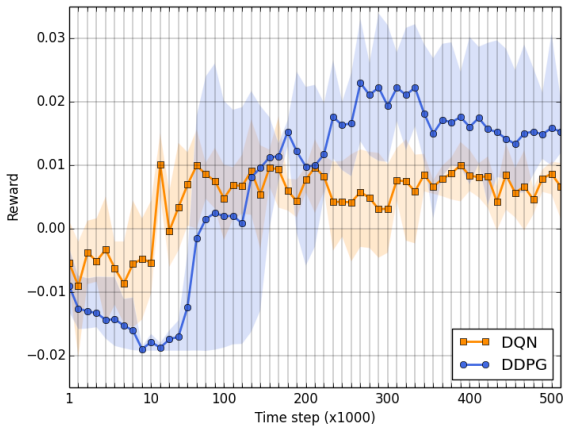


그림 7 : Tmall.com 역사적인 판매 데이터와 DQN 및 DDPG에 대한 FL 오프라인 정책 평가.

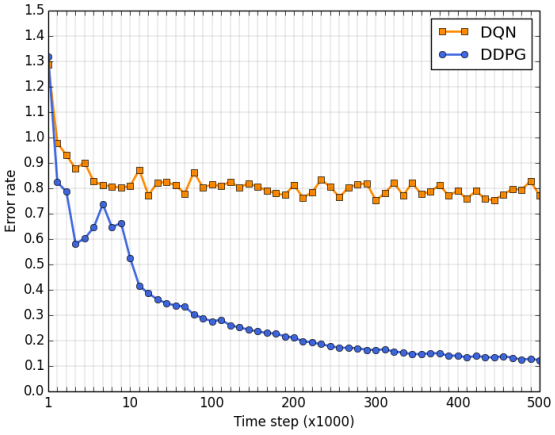


그림 8 : Tmall.com 역사적인 판매 데이터와 사전 훈련 도중 DQN 및 DDPG에 대한 DRCR의 오류율.

BA 뉴 부가 이자형 □ : FOR 에스 TATE 에프 EATURES

우리는 여기서 3 절에서 소개 된 상태 기능에 대한 몇 가지 설명을 제공합니다. 공지 사항 그것은 단지 우리는 단지이 작품에서 전자 상거래 플랫폼에 대한 몇 가지 주로 사용되는 데이터를 표현하기 위해 이러한 기능과 그들 사이의 비교를 선택 그림 6에 보여 주었다 기능을 나열 것이다.

표 1 : 일부 국가의 기능에 대한 설명

특색	설명
모바일에서 UV	고유 방문자는 제품 형태의 모바일 기기를 볼
층 UV	고유 방문자의 층 수는 제품을 볼
범주 가격	범주에있는 모든 제품에 대한 평균 가격
주 가격	일주일에 제품에 대한 평균 가격
층 PV	제품의 상세 페이지가 보여졌습니다 층 시간
모바일에서 PV	제품의 상세 페이지가 이동 장치에 의해 시청 된 시간
3 일 UV	지난 3 일 동안의 평균 UV
할인 가격	가격은 실제로 구매자에 의해 지불 하 고 살만한
주간 PV	지난 주 평균 PV
3 일 가격	지난 3 일 동안 평균 가격
가격 수준	상대적으로 가격이 0에서 1로 만들었
3 일 PV	지난 3 일 동안의 평균 PV
카테고리 구매자	범주에 대한 구매자의 평균 수
카테고리 판매	범주의 평균 판매
주간 UV	지난 주 평균 UV
매달 점수	지난 달에 구매자로부터 제품에 주어진 평균 점수
매주 구매자	지난 주에 대한 구매자의 평균 수
카테고리 수익	범주에 대한 평균 수익
매주 의견	지난 주 동안 구매자로부터 제품에 대한 주어진 댓글의 수
페이지 가격	세부 정보 페이지에서 할인 된 표시하지 않고 가격
할인	제품에 주어진 할인의 평균
주간 점수	지난 주에 구매자로부터 제품에 주어진 평균 점수
매월 의견	지난 달 동안 구매자로부터 제품에 대한 주어진 댓글의 수
주간 판매	지난 주 평균 판매
3 일 구매자	지난 3 일간 구매자의 평균 수
주간 수익	지난 주에 평균 수익
3 일 판매	지난 3 일 동안의 평균 판매
3 일 수익	지난 3 일 동안 평균 수익
구매자 수	제품에 대한 구매자의 수
매상	제품의 판매 볼륨
수익	제품에 대한 수익