Decision Tree 기반 Contextual Bandit 을 활용한 주문 접수 및 스케줄링 최적화: Delayed Reward Compensation 및 Counterfactual Feedback 통합

이찬규, 전성범, 강금석*
한국과학기술원 경영공학부, 동국대학교 산업시스템공학과, *한국과학기술원 경영공학부

changyu.lee@kaist.ac.kr, sbjun@dgu.ac.kr, *keumkang@kaist.ac.kr

A Decision Tree-Based Contextual Bandit Approach for Order Acceptance & Scheduling Optimization: Integrating Delayed Reward Compensation and Counterfactual Feedback

Lee Chan Gyu, Jun Sungbum, Kang Keum Seok*
KAIST Management Engineering, Dongguk Univ. Industrial System Engineering, *KAIST Management Engineering

요 약

본 논문에서는 다종류·다수의 주문이 실시간으로 유입되는 제조 생산 환경에서 제한된 생산 자원을 최적 활용하기 위한 Order Acceptance & Scheduling 문제를 다룬다. 제안하는 프레임워크는 복합적인 Context 를 의사결정나무로 세분화하여 각 분류 구간별로 Contextual Bandit 을 적용함으로써, 전체 Context 를 일괄 처리하는 방식 대비 더 정밀한 의사결정을 가능케 한다. 또한, 즉시 보상만으로는 포착하기 어려운 기회비용에 대한 피드백을 반영하기 위해 Delayed Reward Compensation 기법을 도입하고, 선택되지 않은 행동에도 학습 신호를 제공하는 Counterfactual Feedback 을 병합하여 탐색-활용 균형을 개선하였다. 병렬 기계 생산 환경을 모사한 시뮬레이션 환경에서 제안 모델은 MILP 최적해 대비 평균 76% 수준의 성능을 달성하였으며, 전통적인 Contextual Bandit 대비 약 10%p 이상의 성능 향상을 확인하였다.

Ⅰ. 서 론

최근 제조 산업은 급격히 변화하는 고객 요구와 글로벌 경쟁 심화로 인해 다종류·다수의 주문이 실시간으로 접수되는 환경에 직면하고 있다. 이러한 환경에서 제조 기업들은 제한된 자원을 최적으로 활용하여 수익성을 높이는 복합적인 의사결정을 신속히 내려야 하는 상황에 놓여 있다. 이러한 문제는 Order Acceptance & Scheduling 으로 알려져 있으며[1], 각 주문이 가지는 처리 시간, 수익 등의 특성과 생산 시스템의 현재 상태를 동시에 고려하여 최적의 결정을 도출해야 한다는 점에서 매우 어렵다. 따라서, 본 연구는 Decision Tree 로 복합적인 Context 를 세분화한 뒤 각 분류 구간에서 Contextual Bandit 을 적용해 실시간 환경에서도 적은 데이터로 신속하고 정확한 의사결정을 수행하는 접근법을 제안한다.

Ⅱ. 제안 프레임워크

Contextual Bandit 은 주문이 들어올 때마다 관측된 복잡한 Context 를 바탕으로 신속히 최적의 Action 을 결정할 수 있는 강점을 지닌다. 이는 실시간으로 변하는 환경에서도 빠르게 적응하며, 수학적으로 성능을 보장하고 적은 데이터로도 효과적인 정책 개선이 가능한 점에서 실시간 환경에 유리하다. 또한 본 연구에서 Bandit 의 성능 향상을 위해 Decision Tree 를 활용한다. Decision Tree 는 Context 를 세분화한 뒤, 각 분류 구간에서 Bandit 이 의사결정을 내리도록 활용하여 전체 구간을 한 번에 다루는 방식보다 더 정밀한 결정을 내리게 한다.

2.1 Contextual Bandit (CB)

주문이 도착할 때마다 관측되는 Context 벡터 $x_t \in R$에는 주문의 기본 특성뿐 아니라, 동적인 생산 시스템 정보와 이들로부터 추출된 통계적 특성이 포함된다. 이를 기반으로 각 주문마다 의사결정을 내리는데, 본 연구에서는 전통적인 수락, 거절 옵션 외에 유리한 주문 기회를 고려한 결정 유보와 생산 효율화를 위한 외주 비율 선택을 추가하여, K개 행동 집합 $A = \{a_1, \dots, a_K\}$ 중 하나를 선택하도록 설계하였다. 선택한 행동 a_t에 대해 실시간으로 순수익 r_t를 관측하고 목표는 전체 시간 T 동안의 누적 보상을 최대화하는 CB 의 수학적 정식화는 다음과 같다.

$$\max_{a_1,\dots,a_T} \sum_{t=1}^{T} E[r_t | x_t, a_t]$$

A. Delayed Reward Compensation (DRC)

기존 CB 모델은 행동 a_t를 시점 t에 선택한 뒤 즉시 얻는 보상 r_t를 학습에 활용한다. 그러나 실제로는 선택한 행동이 완료되는 시점 τ_t까지 다른 주문을 처리하거나 기

계를 점유함으로써 얻게 되는 기회비용과 추가 보상이 존재한다. 이를 반영하기 위해 본 연구에서는 DRC 메커니즘을 도입하여 주문의 처리 완료 시점까지의 잠재적인 보상을 반영하여 학습 신호를 제공한다. 이를 통해 더 장기적인 시각에서 행동의 가치를 평가할 수 있다.

B. Counterfactual Feedback (CF)

CB 학습에서는 선택된 행동에 대해서만 보상을 학습하기 때문에, 나머지 행동의 가치를 추정하기 어렵다. 본 연구에서는 CF 를 통해 선택되지 않은 행동들 또한 학습 신호를 부여한다. 이는 선택되지 않은 행동에도 학습 신호를 제공하여, 모델이 더 넓은 공간을 동시에 학습하도록 유도한다. 이를 통해 동시에 여러 학습 사례를 확보함으로써, 학습의 수렴 속도를 가속화하는 효과가 있다.

2.2 Decision Tree (DT)

DT 는 파라미터 조절 없이 Feature 만으로 분할 규칙을 학습하는 비모수 모델이다. 본 연구에서는 C4.5 로 DT 를 구축하고, 단일 DT 와 앙상블 방식을 통해 Tree 구조와 노드 확률의 불확실성을 모사한다. 각 노드는 특정 행동의 성공 확률을 추정하며 최종 노드에서 행동 별 성공 확률을 계산한다. 이를 통해 CB 의 탐색-활용 균형을 직관적이고 해석 가능하게 한다. 반면, C4.5 는 노드의 순수도 증가만을 기준으로 과도하게 분할하는 경향이 있어 과적합을 유발하여 일반화 성능을 저하시킬 수 있다. 본 연구에서는 Gain Ratio 에 균형화 가중치를 부여하여 불필요한 분기를 억제하고 예측력을 향상시킨다[3].

III. 시뮬레이션 및 실험 환경

제안 모델의 성능 검증을 위해 전체 시점 $T = 1000$ 에 걸쳐 각 시점마다 20% 확률로 0~5 개의 주문이 무작위로 발생하도록 시뮬레이션 데이터를 생성하였다. 각 주문의 처리 시간, 납기일, 수익 등은 무작위로 구성된다. 생산 시스템은 5 개의 병렬 기계 시스템으로 구성된다.

IV 절에서는 제안하는 모델의 성능을 시뮬레이션 환경을 통해 평가한다. 먼저, 단일 DT 와 Bootstrap 기반 앙상블을 각각 Context 분할 기법으로 적용하여[2], 분할 방식이 CB 의 의사결정 정확도에 미치는 영향을 정량적으로 비교한다. 수학적 최적해 기준으로 활용되는 MILP, Thompson Sampling-CB(CB), 무작위 행동 선택(Random)과 비교함으로써, 제안 모델이 실시간 주문 처리 환경에서 우수한 성능을 달성하는지 검증한다.

IV. 성능 비교 및 평가

본 절에서는 제안하는 모델의 유효성과 CF, DRC 의 효과를 검증하기 위하여, 제시한 시뮬레이션 환경하에서 다섯 가지 알고리즘을 비교하고 평가하였다. 각 실험은 30 회 반복 수행하였으며, 평가 지표로는 MILP 모델이 산출한 최적 해 대비 누적 순수익 비율을 사용하였다.

표 1 에 제시된 실험 결과에서, 제안한 Bootstrap 기반 CB(BS-CB)은 DRC 만 적용했을 때도 크게 향상되었고, CF를 병합함으로써 추가 성능 향상을 보였다. 단일 DT-CB 에서도 유사한 경향이 관찰되었다. CF 와 DRC 동시 적용 시 75.05%의 평균 성능을 기록하여, DRC 단독 대비 상승하였다.

표 1. 알고리즘 및 CF/DRC 적용 별 시뮬레이션 성능

	CF & DRC	Only CF	Only DRC	None
BS-CB	**75.8** ±4.66	71.64 ±4.96	**74.85** ±1.35	65.29 ±3.88
DT-CB	75.05 ±3.65	**71.76** ±5.43	73.24 ±1.95	**69.62** ±4.73
CB	65.33 ±4.2	65.24 ±4.39	64.32 ±0.59	65.71 ±3.86
Random	52.98 ±2.06	53 ±2.35	53.66 ±1.49	53.56 ±2.87
MILP	100	100	100	100

V. 결론 및 기여

제안한 모델이 최적해의 76% 수준에 근접하는 성과를 낼 수 있음을 확인할 수 있었다. 전반적으로, CF 와 DRC 를 통합 적용했을 때 가장 높은 평균 수익과 안정적인 분산을 확보하였으며, 특히 앙상블 기반 BS-CB 가 본 메커니즘을 가장 효과적으로 활용함을 보였다.

본 연구의 주요 기여는 다음과 같다. 첫째, 선택 가능한 행동으로 유보 옵션을 도입하여 유연성을 제고하였고 최적 외주 비율을 산출해 불필요한 외주를 방지하고 수익을 극대화하였다. 둘째, C4.5 기반 DT 알고리즘을 제안해 도메인 지식 없이 Bandit 문제에 적용 가능함을 보였고, Gain Ratio 계산 방식을 보정해 분할 품질 및 성능을 향상시켰다. 셋째, DRC 와 CF 기법을 통합 적용해 탐색-활용 균형을 강화하였다.

또한 통신 네트워크의 동적 자원 할당과 같이 복합적인 의사결정 문제를 해결하는 데 적용 가능하며, 이는 스마트 팩토리 등 자율 제어 시스템의 최적 정책 학습을 지원한다. 나아가 가상 환경 내 자율 최적화 작업을 효과적으로 구현할 수 있어, 제조·서비스·통신 분야뿐만 아니라 자율 제어 및 메타버스 자율 트윈 연구까지 아우르는 학제간 응용 가능성을 확보한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2021-II211816)

참 고 문 헌

[1] Wester, F. A. W., Wijngaard, J., & ZIJM, W. H. M. (1992). Order acceptance strategies in a production-to-order environment with setup times and due-dates. International Journal of Production Research, 30(6), 1313-1326.

[2] Elmachtoub, A. N., McNellis, R., Oh, S., & Petrik, M. (2017). A practical method for solving contextual bandit problems using decision trees. arXiv preprint arXiv:1706.04687.

[3] Leroux, A., Boussard, M., & Dès, R. (2018). Information gain ratio correction: improving prediction with more balanced decision tree splits. arXiv preprint arXiv:1801.08310.