강화학습 기반 제지 공정 자율 운전: 초기 안정화 가속화 전략

백민우[1], 유지오[1], 길기훈[1], 이서영[1], 도윤미[2], 최진영[2], 이상금[1]*

*국립한밭대학교[1], 한국전자통신연구원[2]
{bmw5779, uzo7383, minegihun, syoung2353} @gmail.com, {ydoh, choij0}@etri.re.kr,
*sangkeum@hanbat.ac.kr

Reinforcement Learning-Based Autonomous Control for Papermaking Process: An Acceleration Strategy for Initial Stabilization

Minu Baek[1], Jio Yoo[1], Gihun Gil[1], Seoyoung Lee[1], Yoonmee Doh[2], Jinyoung Choi[2], Sangkeum Lee[1]*
*Hanbat National University[1], Electronics and Telecommunications Research Institute[2]

요 약

본 논문은 연속 제조 공정의 효율성을 저해하는 핵심 병목 현상인 생산 LOT(공정 단위) 초기의 불안정한 과도 상태(transient state)를 안정화시키는 새로운 강화학습(RL) 프레임워크를 제안한다. 먼저, 동일 지종 내에서도 달라지는 펄프 배합비 특성을 반영하고자 K-Means 클러스터링으로 생산 LOT 을 5 개 그룹으로 군집화하고 해당 군집 ID 를 핵심 상태(state) 변수로 사용한다. 또한, 공정을 초기-중기-후기 세 구간으로 나누어 보상 함수를 동적으로 조절한다. 초기에는 전문가 궤적을 앵커(anchor)로 삼아 안정성을 확보하고 [1], 중기에는 생산성을 극대화해, 후기에는 품질을 유지하도록 설계한다. 실험 결과, 제안된 에이전트는 과거 운전 데이터 대비 전체 구간에서 생산량 1.41% 향상, 후건조기 압력 27.17% 감소, 그리고 보상 변동성 5.3 배 감소라는 복합적인 성과를 달성하며 제안한 프레임워크의 효과를 입증한다.

Ⅰ. 서론

제지 공정과 같은 에너지 집약적 고속 공정은 생산성, 품질, 에너지 효율 간의 상충 관계로 인해 최적 운전이 어렵고, 운전자 경험에 의존하는 기존 방식은 일관성 확보에 한계가 있다 [2]. 이러한 문제를 해결하기 위해 강화학습(RL)은 불확실성 하에서 순차적 의사결정 문제를 해결하는 강력한 패러다임으로 주목받고 있다. 하지만 실제 산업 데이터는 LOT 의 시작 및 종료 구간에서 변동성이 매우 커서 강건한 제어 정책을 학습하는 데 어려움이 따른다. 본 연구는 실시간 공정 변화에 동적으로 적응할 수 있는 폐쇄 루프(closed-loop) 제어 정책 개발을 목표로, LOT 교체 시 발생하는 초기 안정화 구간의 비효율 문제를 해결하기 위해 구간별 동적 보상과 전문가 궤적 앵커링이라는 새로운 RL 프레임워크를 제안한다.

Ⅱ. 본론

1. 공정 안정화를 위한 강화학습 프레임워크

제안하는 방법론은 제어 문제를 MDP(Markov Decision Process)로 정립하고, 에이전트는 물리 모델과 데이터 기반 모델이 결합된 하이브리드 디지털 트윈 환경 내에서 학습한다. 이러한 하이브리드 시뮬레이션은 순수 데이터 기반 모델의 예측 불확실성과 물리 모델의 부정확성을 상호 보완하여 실제 공정과 유사한 환경을 제공하는 데 효과적이다 [3].

2. 펄프 배합비 기반 공정 군집화

동일 제지 공정에서 원료인 펄프의 배합비는 생산성과 최종 품질을 결정하는 핵심 인자이지만, 원가 및 수급 상황에 따라 동적으로 변동하여 공정 제어의 불확실성을 높이는 주요 원인이 된다. 이러한 배합비 변화에 따른 공정의 동적 특성을 에이전트가 명시적으로 인지하고 적응적 제어 정책을 수립할 수 있도록, 본 연구에서는 주요 펄프 태그인 배합 비율을 핵심 변수 5 개로 사용하여 K-Means 클러스터링 분석을 수행한다. 이에 본 연구에서는 펄프 배합비의 뚜렷한 차이를 기반으로 K-Means 클러스터링을 수행하여 전체 생산 LOT 을 5 개 군집으로 분류하고, 이 군집 ID 를 상태 변수에 포함시켜, 에이전트가 현재 공정의 특성을 식별한다.

3. 위상 적응형 보상 함수 설계

에이전트의 학습을 안내하는 보상 함수 R_t는 공정의 진행 단계(초기, 중기, 후기)에 따라 동적으로 가중치가 변하는 구조를 가진다. 즉, 공정 단계별로 보상 가중치와 페널티를 조절함으로써, 해당 구간에서 에이전트가 따라야 할 허용 가능한 정책의 범위를 효과적으로 정의하고 안내한다.

3.1. 초기 구간: 전문가 궤적 앵커링

LOT 시작 구간(~10%)에 해당하는 비효율적인 탐색을 방지하기 위해 전문가의 과거 운전 데이터를 앵커로 사용한다 [1]. 에이전트의 현재 상태 (s_t) 가 전문가의 과거 상태 (s_t^{LOT}) 뿐 아니라 식 (1)과 같이 페널티(P_{anchor})를 부과해서 자유로운 전략을 도출한다.

$$P_{anchor}(s_t) = \lambda_{anchor} \cdot d_t \cdot ||(s_t - s_t^{LOT}) \odot m||^2 \quad (1)$$

여기서 λ_{anchor} 는 가중치, d_t 는 시간 감쇠 항, m 은 마스크이다. 본 연구의 궤적 앵커링은 상태 기반 규제(state-based regularization)를 통해 안전한 영역 내에서 전문가를 능가하는 전략 탐색을 가능하게 한다.

3.2. 중기 구간: 생산성 극대화

공정이 안정화된 중기 구간(10%~90%)에서는 보상 함수의 가중치를 생산성 극대화로 전환한다. 에이전트는 안정적인 품질 범위 내에서 공정 속도를 높여 생산량을 최대화하도록 보상을 받는다.

3.3. 후기 구간: 품질 및 안정성 유지

LOT 종료 직전의 후기 구간(90%~)에서는 다시 품질 및 공정 안정성 유지에 보상의 초점을 맞춘다. 또한, 에이전트가 목표 시간 안에 공정을 안정시키면 안정화 달성 보너스를 부여하여 신속한 안정화를 명시적으로 장려한다.

4. 실험 결과 및 분석

제안된 에이전트의 성능을 검증하기 위해, 학습에 사용되지 않은 30 개의 LOT 에 대한 과거 운전 데이터와 성능을 비교한다.

4.1. 생산성 및 에너지 효율성 분석

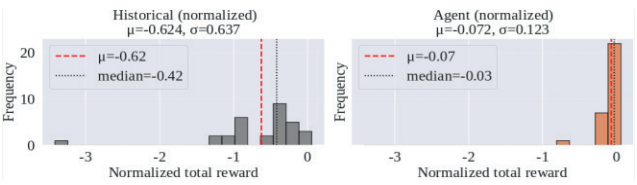
구분	과거 평균	에이전트 평균	변화율(%)
생산량	14.545	14.749	**+1.41%**
공정 속도 (m/min)	753.814	754.079	+0.04%
후건조기 압력 (kg/cm^2)	1.653	1.204	**-27.17%**
수분(%)	4.664	4.746	+1.76%

[표 1] 전체 구간 핵심성과지표(KPI) 비교

에이전트는 과거 운전 대비 생산성을 1.41% 향상시키면서 후건조기 압력은 27.17% 절감했으며, 이는 허용 규격 내에서 수분(+1.76%)을 소폭 높여 에너지 효율을 극대화한 트레이드오프 전략의 학습 결과이다.

4.2. 안정성 및 리스크 분석

에이전트는 과거 운전 방식의 가장 큰 문제였던 변동성과 리스크를 개선한다.



[그림 1] 과거 운전(Historical)과 에이전트(Agent)의 정규화된 보상 분포 비교

(1) 전반적 성능 및 신뢰도

에이전트는 전체 테스트 LOT 중 90%에서 과거 운전보다 높은 성과(Win rate)를 달성했으며, 평균 보상 차이(Δ mean)는 +0.645 로 통계적으로 유의미한 우위를 보인다.

(2) 변동성 감소

[그림 1]에서 볼 수 있듯, 과거 운전(좌)의 보상 분포는 넓게 흩어져 있고 표준편차(σ)가 0.745 에 달하는 반면, 에이전트(우)의 분포는 평균 -0.086 을 중심으로 매우 좁게 밀집되어 있으며 표준편차는 0.140 에 불과하다. 이는 성능의 변동성이 5.3 배 이상 감소했음을 의미하며(σ ratio: 0.188), 에이전트가 매우 일관되고 예측 가능한 운전을 수행함을 보인다.

(3) 최악 시나리오 개선

에이전트의 가장 큰 성과는 치명적인 실패를 회피하는 능력에 있다. 과거 운전에서 최하위 10%의 평균 보상은 -2.257 이었으나, 에이전트는 이를 -0.400 으로 크게 개선하여 최악의 경우에도 안정적인 성능을 유지한다. 특히, 과거 운전에서 최악의 성과(-3.995)를 보인 LOT 의 경우, 에이전트는 보상 점수를 -0.704 로 크게 개선하여(Δ +3.291) 문제 상황 해결 능력을 명확히 입증한다.

Ⅲ. 결론

본 논문은 제지 공정의 초기 안정화 문제를 해결하기 위해, 구간별 동적 보상 함수와 전문가 궤적 앵커링을 결합한 강화학습 프레임워크를 제안했다. 실험을 통해 증명된 성능 및 안정성 향상은 자율 운전 기반의 지능형 산업 시스템으로 나아가는 진전을 보여준다. 향후 과제로는 시뮬레이션과 실제 공정 간의 차이(sim-to-real gap)를 극복하기 위한 파일럿 테스트가 있다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구과제임. (No. 20202020900290)

참 고 문 헌

[1] K. Banihashem, et al., "Admissible Policy Teaching through Reward Design," In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 4830-4838, 2022.

[2] S. Lee, et al., "Factory Energy Management by Steam Energy Cluster Modeling in Paper-Making," in Proc. 11th IEEE Int. Conf. Smart Grid Commun. (SmartGridComm), 2023, pp. 1-5.

[3] A. A. G. Requena, et al., "Hybrid physics-based and data-driven modeling forprobabilistic failure prognosis," In *Annual Conference of the Prognostics and Health Management Society*, vol. 9, 2017.