제지 공정 강화학습에서 GMM 기반 OOD 감지와 계층적 안전 스위칭 정책의 CVaR 기반 위험 관리 효과 분석

유지오, 백민우, 길기훈, 이서영, 장여진, 이상금*
*국립한밭대학교,

{uzo7383, bmw5779, minegihun, syoung2353, jyeoj251}@gmail.com *sangkeum@hanbat.ac.kr

A Study on CVaR-based Risk Management using GMM-based OOD Detection and Hierarchical Safety Switching Policy in Paper Manufacturing Reinforcement Learning

Jio Yoo, Minu Baek, Gihun Gil, Seoyoung Lee, Yeojin Jang, Sangkeum Lee*
*Hanbat National Univ.

요 약

본 논문은 제지 공정 강화학습 기반 공정 제어에서 평균 성능뿐 아니라 드물게 발생하는 최악 상황의 손실, 즉 꼬리 위험(tail risk)을 함께 고려하기 위해 GMM(Gaussian Mixture Model) 기반 OOD(Out-of-Distribution) 감지와 계층적 안전 스위칭 정책을 제안한다. GMM 을 활용하여 학습 분포에서 벗어난 이상 상태(OOD)를 감지하고, 압력이 안전 경계 근처인 위험 구간을 식별하며, 정상 상태에서는 압력 기반으로 PPO(Proximal Policy Optimization)와 SAC(Soft Actor-Critic) 정책을 선택한다. OOD 상태이거나 압력 위험 구간에서는 보수적인 SAC 정책을 강제하여 안전성을 확보한다. 실험 결과, 제안 방법은 기존 GMM OOD 단독 방법 대비 Mean Return 을 유지하면서 CVaR(Conditional Value-at-Risk)을 -58.71 에서 -24.37 로 34 점 개선하였으며, 표준편차도 37.23 에서 24.89 로 감소하여 정책의 안정성이 향상되었다.

Ⅰ. 서 론

최근 강화학습이 공정 제어에 적용되고 있으나, 대부분 평균 보상 중심으로 평가하여 드물게 발생하는 꼬리 위험에 대한 고려가 부족하다[1]. 특히 실제 공정에서는 학습 시 경험하지 못한 새로운 운전 조건이 발생할 수 있으며, 이러한 OOD(Out-of-Distribution) 상태에서 공격적인 정책을 적용하면 심각한 성능 저하나 안전 위반이 발생할 수 있다. 또한 압력이 안전 범위의 경계에 근접한 상황에서는 작은 제어 오차도 안전 위반으로 이어질 수 있어 별도의 안전 장치가 필요하다. 본 논문은 이러한 문제를 해결하기 위해 GMM(Gaussian Mixture Model) 기반 OOD 감지와 압력 경계 안전 메커니즘을 결합한 계층적 안전 스위칭 정책을 제안하며, CVaR(Conditional Value-at-Risk) 기반 위험 지표로 정책의 꼬리 위험 관리 효과를 분석한다.

PPO(Proximal Policy Optimization)는 클리핑 기반 on-policy 알고리즘으로 안정적인 정책 업데이트를 제공한다[2]. SAC(Soft Actor-Critic)는 최대 엔트로피 기반 off-policy 알고리즘으로 탐색과 활용의 균형을 유지하고, 확률적 정책을 통해 제어 입력이 과도하게 치우치는 것을 억제하는 특성이 있다[3]. GMM 은 데이터 분포를 여러 개의 가우시안 분포의 혼합으로 모델링하는 방법으로, 밀도 추정 및 이상 감지에 널리 활용된다[4]. 새로운 데이터 포인트의 log-likelihood 를 계산하여 학습 분포에서 벗어난 정도를 정량화할 수 있다. CVaR 은 VaR(Value-at-Risk) 이하 구간의 평균으로 정의되며, 꼬리 위험을 정량화하는 일관된 위험 측도로 활용되기 때문에[5], 하위 10% 에피소드의 평균 Return 을 CVaR(10%)로 정의한다.

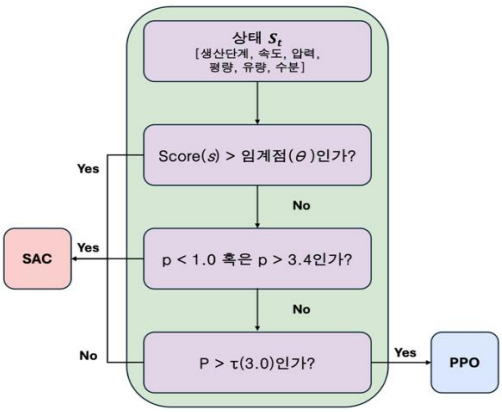
Ⅱ. 제안 방법

2.1 문제 정의

제지 공정은 MDP(Markov Decision Process)로 모델링되며, 상태 s 는 [생산단계, 속도, 압력, 평량, 유량, 수분]으로 구성한다. 행동 a는 [Δ속도, Δ압력, Δ유량]의 변화량이며, 보상은 생산량, 수분, 평량, 에너지 항의 가중 합으로 정의한다. 압력 안전 범위는 [0.7, 3.7]로 정의하며, 범위 이탈 시 위반(high pressure violation, highV)으로 집계한다.

2.2 계층적 안전 스위칭 정책

그림 1 에서 제안 방법은 세 단계의 계층적 의사결정 구조를 갖는다.



[그림 1] 제안 방법 계층적 의사결정 구조

학습 데이터로 K 개(K=5)의 가우시안 분포를 혼합한 GMM 을 학습한다. 새로운 상태 s 에 대해 OOD 점수는 다음과 같다.

$$score(s) = -log\,\Sigma_k\,\pi_k \cdot N(s\,|\mu_k, \Sigma_k)$$

상태 s 가 각 가우시안 컴포넌트 k 에서 관측될 확률밀도($\mathcal{N}(s\mid\mu_k,\Sigma_k)$)를 혼합 가중치 π_k 로 가중합 해, 전체 확률밀도를 계산한다. OOD 점수는 score(s) 로 정의하며, 학습 분포에서의 likelihood(데이터가 분포에서 나왔을 확률)가 낮을수록 score 가 커져 OOD 정도를 정량화한다. 학습 데이터에서 계산한 OOD 점수의 95 백분위수를 임계값(Θ)으로 설정하고, score(s) > Θ 인 경우 OOD 로 판단한다. 또한 압력이 안전 범위의 경계에 근접하면 위험 구간으로 판단한다.

$$OOD = (p < p_{min} + m) \vee (p > p_{max} - m)$$

$$= (p < 1.0) \vee (p > 3.4)$$

OOD 도, 위험 구간도 아닌 정상 상태에서는 압력 임계값 τ=3.0 을 기준으로 정책을 선택한다.

2.3 설계 근거

PPO 는 성능 최적화를 위해 압력을 한계까지 밀어붙이는 경향이 있다. 안전 한계값 주변에 버퍼 구간을 추가해 경계 구간에서 안전 정책으로 변환시키는 파라미터(Margin)를 두고 미리 SAC 로 전환하여 안전 위반을 예방한다. 실험 결과 고압 구간(p > 3.0)에서는 PPO 가, 정상 구간에서는 SAC 가 더 나은 성능을 보여 이를 반영한다.

2.4 검증 프로토콜

데이터 누수 방지를 위해 LOT 단위로 Train(80%)과 Holdout(20%)을 분리한다. GMM 은 Train 데이터로 학습하며, 하이퍼파라미터(τ, m)는 Train 에서 그리드 탐색으로 선정하고 Holdout 에서 최종 성능을 평가한다.

Ⅲ. 실험 결과

3.1 실험 설정

2022 년 제지 공정 데이터를 LOT 단위로 분리하고 50 에피소드, 최대 100 step 으로 평가를 수행하며, 평가 지표는 Mean Return(에피소드 평균 리워드), CVaR(10%), Std, highV 이다. 또한 GMM 은 5 개의 가우시안 컴포넌트로 구성하며, 학습 데이터 10,000 개 상태로 학습한다.

3.2 비교 분석

Method	Mean	CVaR	Std	highV
GMM only	32.06	-58.71	37.23	286
Proposed	32.16	-24.37	24.89	300

[표 1] GMM 만 썼을 때와 제안 방법 성능 비교

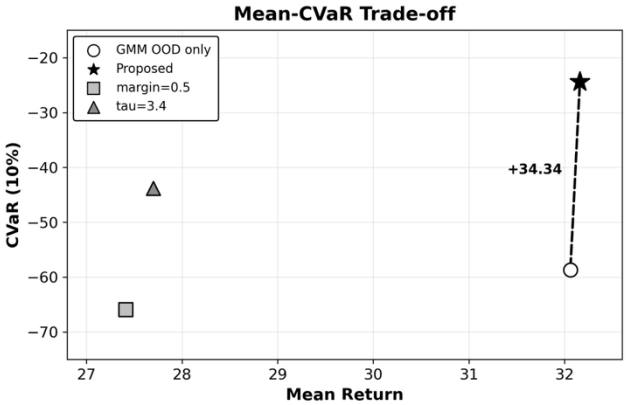
표 1 에서 제안 방법은 GMM 만 쓰는 학습 방법 대비 ΔMean +0.1, ΔCVaR +34.34, Δstd -12.34, ΔhighV +14 인 결과를 보인다.

Margin	Mean	CvaR	highV
0.00	2.06	-58.71	286
0.3	32.16	-24.37	300
0.5	27.41	-65.92	520

[표 2] Margin 에 따른 성능 비교

표 2 에서 Margin 에 따른 Mean, CVaR, highV 값을 비교한 결과, Margin 이 너무 크면 안전 구간이 지나치게 좁아져 성능이 저하되기 때문에, 0.3 일 때 최적의 균형을 보인다.

3.3 위험-수익 분석



[그림 2] 제안 정책 위험-수익 성능 개선

그림 2 에서 Margin 을 0.5 와 τ 를 3.4 로 설정한 정책은 상대적으로 낮은 위험-수익 성능을 보이며, GMM 만 쓰는 정책에 비해 제안 방법은 Mean-CVaR 공간에서 우상단으로 +34.34 이동하여 평균 성능을 유지하면서 꼬리 위험을 효과적으로 감소시킨다.

Ⅳ. 결 론

본 논문은 제지 공정 강화학습에서 꼬리 위험을 관리하기 위해 GMM 기반 OOD 감지와 압력 경계 안전 메커니즘을 결합한 계층적 안전 스위칭 정책을 제안한다. 제안 방법은 세 단계의 의사결정 구조를 통해 학습 분포 이탈 상태, 압력 위험 구간, 정상 운전 구간을 구분하고, 각 상황에 적합한 정책을 선택한다.

실험 결과, 평균 성능을 유지하면서 CVaR(10%)을 34 점 개선하고 표준편차를 12 점 감소시켜, 정책의 안정성과 꼬리 위험 관리 효과를 입증하였다. 특히 규칙 기반의 해석 가능한 구조를 통해 왜 특정 정책이 선택되었는지 설명할 수 있어 실제 공정 적용 시 운전자의 신뢰를 확보한다. 향후 연구에서는 CVaR 을 직접 최적화하는 risk-sensitive 강화학습 또는 분포형 강화학습을 적용하여 학습 단계에서 꼬리 위험을 직접 제어하는 방향으로 확장할 계획이다.

참 고 문 헌

[1] J. García and F. Fernández, "A Comprehensive Survey on Safe Reinforcement Learning," J. Mach. Learn. Res., vol. 16, pp. 1437-1480, 2015.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv:1707.06347, 2017.

[3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in Proc. 35th Int. Conf. Mach. Learn. (ICML), pp. 1861-1870, 2018.

[4] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[5] R. T. Rockafellar and S. Uryasev, "Conditional Value-at-Risk for General Loss Distributions," J. Banking & Finance, vol. 26, no. 7, pp. 1443-1471, 2002.