심층강화학습 기반 반도체 공정 데이터 결과 예측 모델 연구

최우석[O1], 장병탁[1,2]
[1]서울대학교 협동과정 뇌과학전공, [2]서울대학교 컴퓨터공학부
{wschoi, btzhang}@bi.snu.ac.kr

Prediction model for semiconductor industrial process data based on Deep Reinforcement Learning

Woo Suk Choi[O1], Byoung-Tak Zhang[1,2]
[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University

요 약

강화학습(Reinforcement Learning)은 현재 로보틱스(Robotics), 자연어 처리(Natural Language Processing), 컴퓨터 비전(Computer Vision) 등에서 널리 활용되고 있는 학습 방법이며 실제로 광범위한 분야 (공정, 금융, 의료, 화학 및 예술)에 적용되어 사용되고 있다. 이러한 실 사례들을 모티브로 삼아 본 논문에서는 강화학습 알고리즘을 활용하여 반도체 공정 데이터 분석 및 결과를 예측하는 모델을 소개한다. 해당 모델은 Google DeepMind에서 제안한 심층강화학습인 Deep Q-Network(DQN)를 공정데이터에 적용시켜 환경(Environment), 에이전트(Agent), 상태(State)를 정의하고 행동(Action)을 수행하여 보상(Reward) 및 벌칙(Penalty)을 얻음으로써 결과를 예측한다. 또한 예측한 결과 값을 기반으로 범용적인 자동화 공정 시스템으로 확장하기 위한 방법을 논의한다.

1. 서 론

인간을 포함한 모든 동물들은 어떠한 환경에 적응할 때 무수히 많은 행동과 경험을 통해 옳고 그름을 알게 된다. 그리고 그 옳고 그름을 바탕으로 주변 환경이 다양하게 바뀌더라도 이에 맞춰 적응하고 행동하게 된다. 이러한 모든 과정이 어떤 경험 또는 행동에 대해 학습을 하는 단계라고 볼 수 있다. 강화학습의 목표도 인간 또는 동물이 환경에 대해 학습하는 과정과 비슷하다. 강화학습이란 크게는 인공지능(Artificial Intelligence)의 한 영역이고 세부적으로는 기계학습(Machine Learning)의 한 영역이며, 이상적인 학습 과정을 통해 환경에 대한 행동으로부터 목표를 찾아가는 것을 말한다. 이를 통해 목표에 얼마나 빠르고 효율적으로 찾아가는지가 주요 관건이다.

일반적으로 강화학습의 구조에는 행동(Action)을 하는 주체인 에이전트(Agent)와 그 행동에 대한 대응하는 환경(Environment)이 있다. 초기에 에이전트는 환경에 대한 정보가 없어 현재의 환경에 대한 주어진 정보인 상태만을 받고 이에 대한 행동을 취하게 되며 환경은 에이전트의 행동에 대해서 판단 후 보상(Reward)과 벌칙(Penalty)을 준다. 보상과 벌칙을 주는 과정 중에도 환경은 에이전트가 취한 행동에 대해 새롭게 업데이트된 상태를 다시 에이전트에게 알려준다. 이러한 과정을 반복함으로써 최종적으로 에이전트가 가장 높은 보상 값을 갖도록 학습된다.

이상적인 학습 과정이란 Q-학습(Q-learning)[1] 또는 Deep Q-Network(DQN)[2] 같이 인간이 만든 알고리즘을 의미하며 본 논문에서는 이상적인 학습 및 행동을 반도체 공정에 적용시켜 보고자 Q-학습 기반 강화학습 알고리즘 중 하나인 Deep Q-Network(DQN)을 활용하여 반도체 공정 결과 예측 모델을 연구하였고, 예측한 결과 값을 분석하고 이를 기반으로 범용적인 자동화 공정 시스템으로 확장하기 위한 방법을 논의한다.

2. 연구방법

Q-학습은 주어진 상태에서 주어진 행동을 수행 하는 것이 가져다줄 효용의 기댓값을 예측하는 Q-함수(Q-function)를 학습함으로써 최적의 정책을 학습한다. 최적의 정책을 위해 behavior 정책은 Epsilon-greedy(ε -greedy), 타겟 정책은 greedy를 사용한다. Q-학습은 현재 상태(S_t)에서 탐욕정책에 따라 행동(A_t)을 선택한 다음 Q-함수를 이용하여 다음 상태(S_{t+1})에서의 행동(a')은 타겟 정책에 따라 선택 되고 이에 따라 Q-함수를 업데이트 한다. Q-함수(Q)의 업데이트 식은 다음과 같다.

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(R_{t+1} + \gamma \max Q(S_{t+1},a') - Q(S_t,A_t)) \quad (1)$$

이 업데이트 식은 동적 프로그래밍의 value iteration을 기반으로 한 것이며, 최적의 행동-값 함수(Optimal action-value function)로 수렴하게 된다. 에이전트는 실제로 다음 상태에서 어떤 행동을 했는지와 상관없이 현재 상태 s의 Q-함수를 업데이트할 때 다음 상태의 최대 Q-함수를 사용한다.

DQN에는 경험 리플레이(Experience Replay)가 있는데, 경험 리플레이는 에이전트가 환경에서 탐험하며 얻은 샘플(s,a,r,s')을 메모리에 저장하는 방식으로 샘플의 효율성을 높여준다. 기존 DQN에서는 탐험(Exploration)을 통해 새로운 상황을 학습하지만 공정에서는 탐험이 시스템에 큰 영향을 미칠 수 있기에 이를 방지하기 위해 데이터를 기반으로 상태를 입력으로 넣는다. 그리고 심층신경망을 학습하도록 설정하고 학습된 Q-함수는 행동 값을 추출하여 환경에 행동하도록 구축하였다. 이 때,

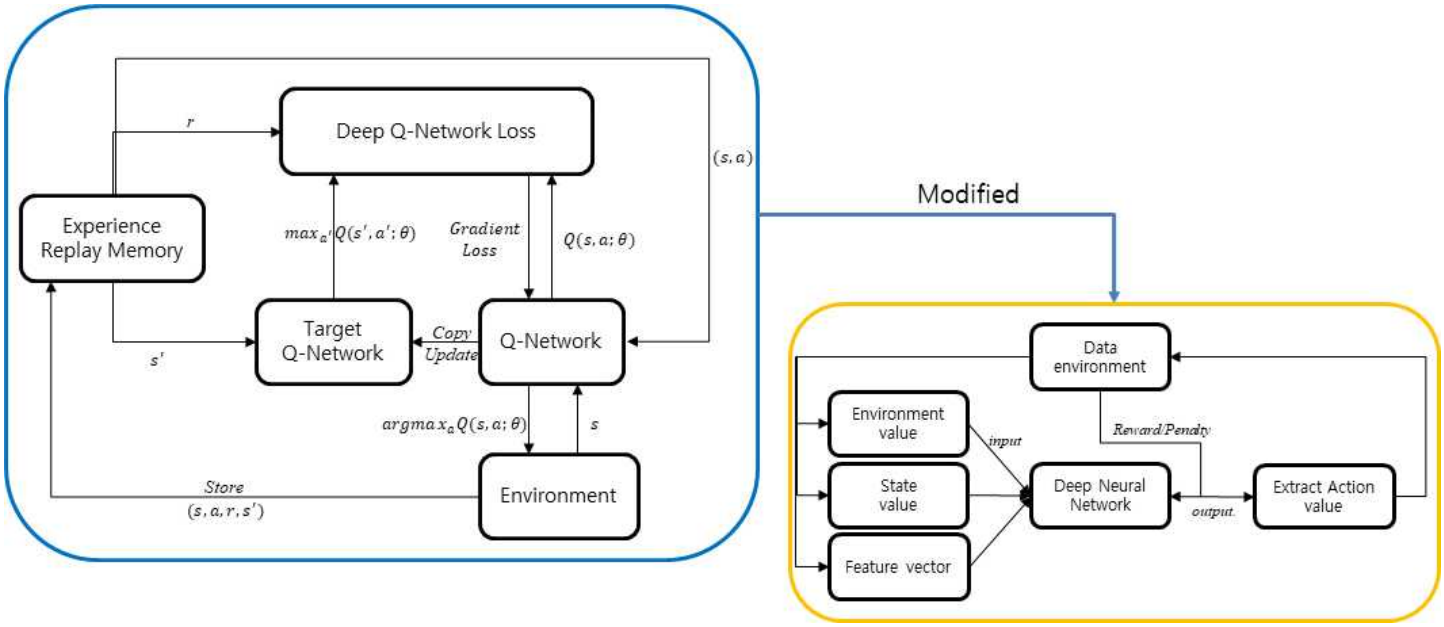

그림 1. DQN 기반 공정 데이터 결과 예측 모델

정답을 내는 Q-Network가 계속 업데이트 되면 업데이트의 목표가 되는 정답이 계속 변화하기 때문에 방지하기 위해 타겟 Q-Network(Target Q-network)에서 일정 시간 간격으로 유지하며 update를 복사한다. 제안된 모델은 그림 1과 같다. 학습을 위한 DQN의 오류 함수는 Mean Square Error(MSE)와 동일하며 다음과 같이 정의하였다.

$$MSE = (Answer - \mathrm{Prediction})^2 \quad (2)$$
$$= (R_{t+1} + \gamma \max Q(s', a'; \theta) - Q(s, a; \theta))^2$$

에이전트가 이상적인 학습을 하도록 데이터를 기반으로 출력 매개변수에 최소/최댓값을 설정하여 행동이 범위 안의 값으로 행동하면 보상(Reward)을 주고 출력 값의 최소/최댓값에서 벗어나면 벌칙(Penalty)을 주도록 설정하였다. 에이전트의 상태 값은 에이전트가 행동한 값을 입력 매개변수의 최소/최댓값 기준 10개의 색인(index)에 대입하여 얻은 값이며, 그에 대한 기울기 값을 feature vector로 만들어 심층 신경망의 input으로 넣도록 하였다.

3. 실험구성 및 결과

3.1 실험구성

실험은 이산화규소 식각데이터(SiO_2 etching Data)를 사용하였다. 이산화규소 식각데이터의 데이터 개수는 50개이며 실제 etching process 에서의 장비(입력 값) 매개변수와 결과(출력 값)로 구성되어있다.

장비 매개변수 속성은 총 69개로 크게 분광법(OES line intensities), 13개의 장비(Equipment), 1개의 현상학 모델링(PI) 매개변수로 구성되어있다. 장비 매개변수 전체를 입력 값으로 넣는 것은 결과 예측에 비효율적일 것으로 판단되어 각 변수간의 선형적 및 비선형적 관계를 알아보기 위해 상관 분석(Correlation Analysis)을 수행하였다. 결과는 그림 2와 같다. 그 결과, PI 매개변수가 OES와 Equipment에 비해 결과 매개변수

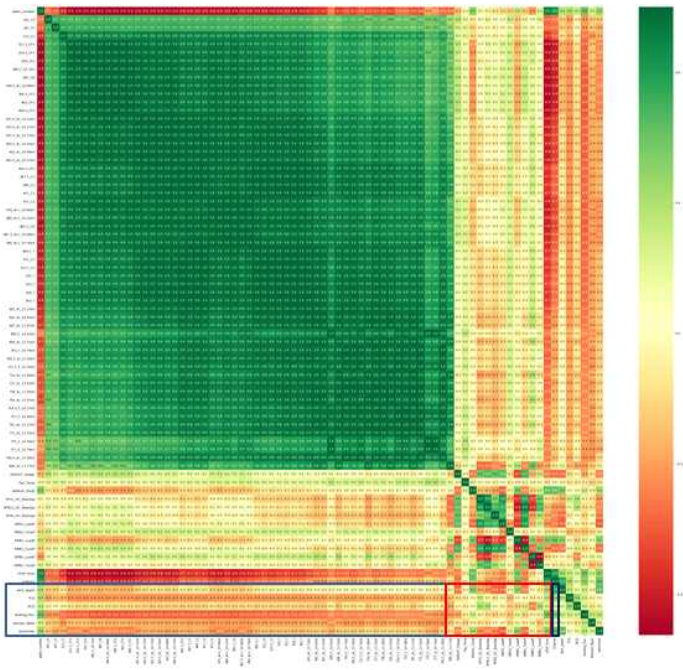

그림 2. 매개변수 간의 상관 분석 히트맵. 파란 박스: OES, 빨간 박스: Equipment, 검은 박스: PI

값들과 상관관계가 상대적으로 높은 것을 알 수 있었다.

결과 매개변수의 경우, 총 6개의 특성으로 식각 깊이(Etch depth), 윗 선폭(Top Critical Dimension), 보잉 선폭(Bowing Critical Dimension), 보잉 위치(Bowing Position), 잔여 마스크(Remaining mask), 선택비(Selectivity)로 구성되어있다. 결과 매개변수 중 식각 깊이가 식각 공정에서 가장 중요한 변수로 작용하기 때문에 식각 깊이를 예측할 결과 값으로 설정하였다.

총 50개의 데이터 중 40개 에피소드를 학습에 사용하였고 나머지 10개 에피소드는 예측 결과 값과 Ground-truth를 비교하기 위해 사용되었다.

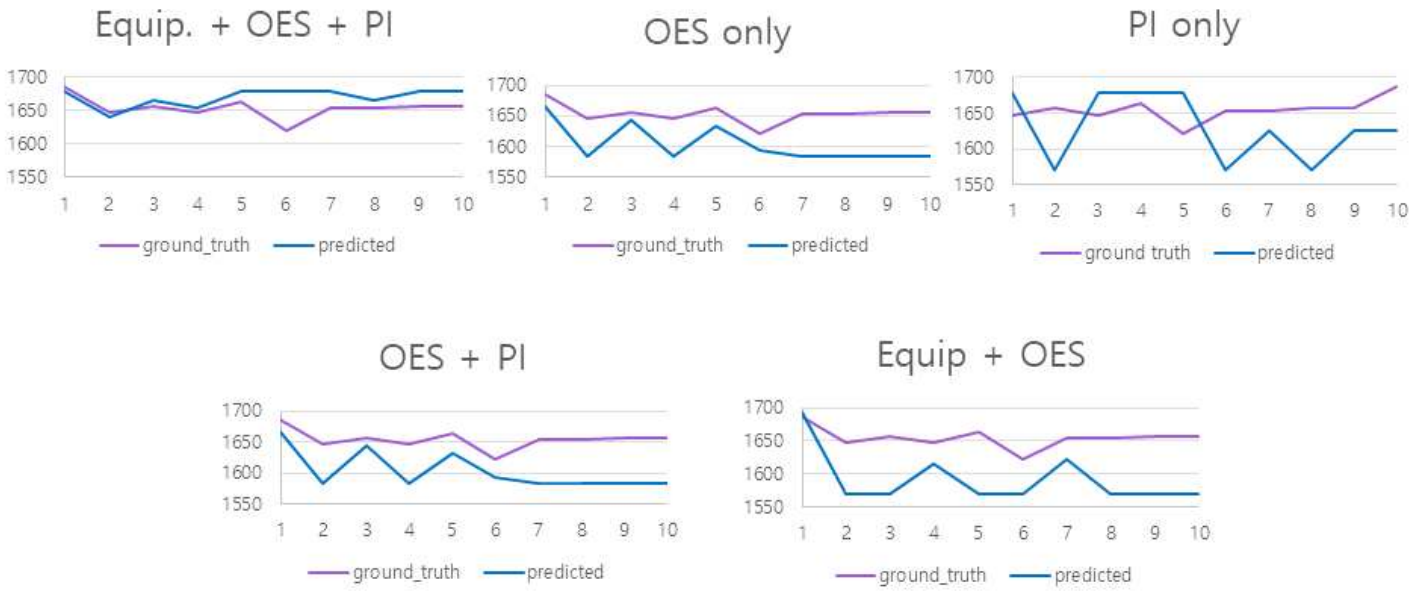

그림 3. 다양한 입력 값 조합으로 식각 깊이(etch depth) 예측 결과 값을 실제 값(Ground-Truth)과 비교한 그래프

3.2 실험결과

상관 분석 결과를 기반으로 먼저 현상학 모델링(PI) 하나만을 입력 값으로 하여 식각 깊이 결과를 예측하도록 실험하였고, 이후 입력 값에 OES와 Equipment를 포함하여 여러 가지 다양한 조합으로 실험을 진행하였다. 모든 조합의 예측 결과 값을 실제 값(Ground-truth)과 비교한 결과는 그림 3에 나와 있다.

그림 3와 같이 입력 매개변수 조합 중 한 가지 또는 두 가지 조합만 사용했을 때의 예측 결과 값이 Equipment, OES, PI 매개변수 세 가지 모두 입력 값으로 썼을 때보다 오차 범위가 크고 값이 일정하지 않았다. 각 입력 값 조합별 실제 값과 예측 결과값 사이의 평균 MSE를 표 1에 작성했다.

입력 값 조합	평균 MSE
Equip + OES + PI	551.5
OES only	1426.9
PI only	3326.5
OES + PI	3081.6
Equip. + OES	4880.7

표 1. 입력 값 조합별 실제 값 및 예측 결과 값의 평균 MSE

4. 결론

공정에 강화학습을 적용하는데는 충분한 학습을 위한 충분한 데이터를 요구한다. 하지만 공정 데이터는 기본적으로 산업 데이터이기에 많은 수의 데이터를 공개하기에는 쉽지 않다.

충분한 데이터를 확보하여 기계가 환경 및 상태를 인식하고 보상과 벌칙을 통해 학습한다면 더 나은 예측 결과 값을 출력할 수 있을 것이다. 또한 본 논문에서 사용한 Deep Q-Network 이외에도 현재 많은 강화학습 알고리즘들(Double DQN[3], Dueling DQN[4])이 나오고 있어 현재 4차 산업혁명을 통해 진행 중인 스마트 팩토리[5] 또는 스마트 사물인터넷(IoT) 연구 발전에 기여할 수 있을 것이다.

5. 감사의 글

이 논문은 삼성전자의 지원을 받아 이루어진 연구이며, 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind)와 한국산업기술진흥원 (P0006720-GENKO)의 지원을 일부 받았음.

참고문헌

[1] C. Watkins, C.J.C.H., “Learning from Delayed Rewards” (PhD Thesis), 220-228, 1989

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D.Wierstra, and M. Reidmil;er, “Playing Atari with Deep Reinforcement Learning,” arXiv preprint arXiv: 1313.5602, 2013

[3] H. Hasselt, A. Guez, D. Silver, “Deep Reinforcement Learning with double Q-learning”, arXiv: 1509.06461

[4] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, “Dueling Network Architectures for Deep Reinforcement Learning”, arXiv:1511.06581

[5] A. Radziwon, A. Bilberg, E. S. Madsen, “The Smart Factory: Exploring adaptive and Flexible Manufacturing Solutions”, Elsevier, Procedia Engineering Vol. 69, 1184-1190, 2014