머신러닝 기법 활용 선박 블록 제조 공정 지연 예측 모델 개발

제15회 한국대학생 산업공학 프로젝트경진대회 2020.10.15

참여 연구원

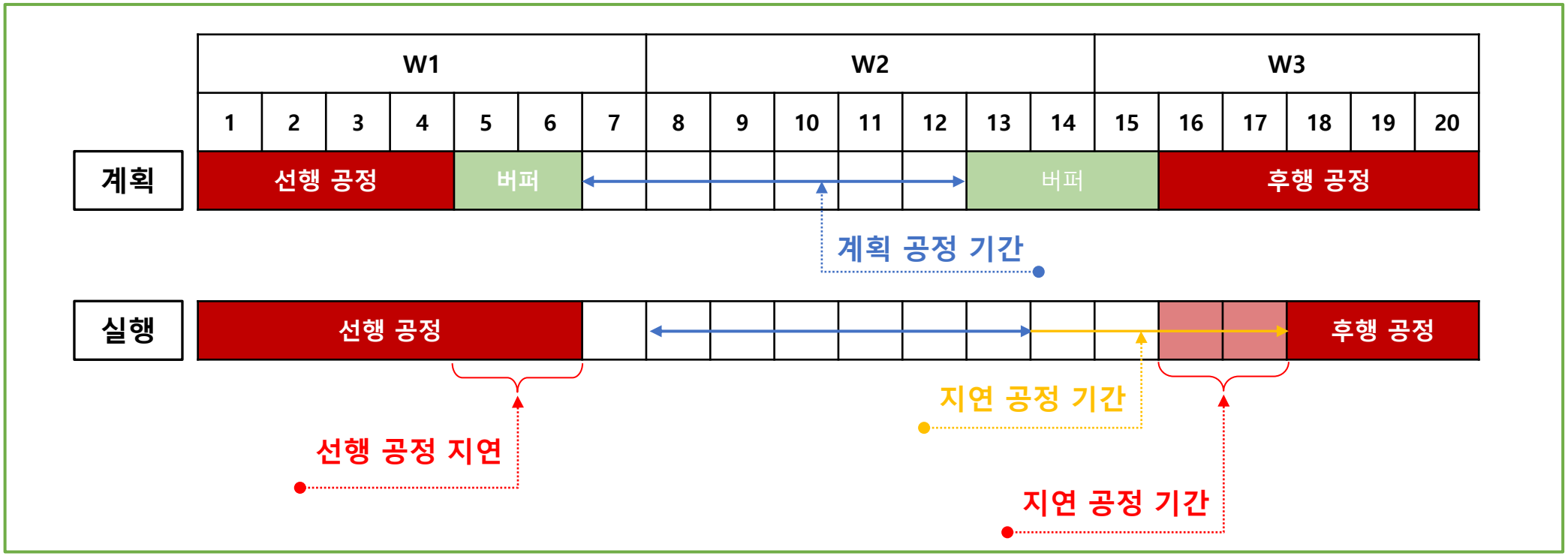
유상현, 곽동훈

목차

연구 배경

- 조선소 관리의 문제점
  - 계획과 실적의 큰 격차
  - 공정 예측의 어려움으로 지속적 손실 발생

- 빈번한 공정 지연에 따른 손실
  - 배원계획의 불균형에 따른 인력 수급 문제 → 야근, 외주 등 추가 지출
  - 납기 지연 → 고객 신뢰도 하락, 패널티 발생



연구 목표

문제 정의

- 노동 집약적 산업으로 계획과 실적의 불가피한 격차 발생
- 격차를 줄이기 위한 부가 비용의 지출

해결 방안

- 공정 데이터 기반 기계학습
- 지연 공정 사전 예측 모델 개발

기대 효과

- 지연 공정 사전 예측을 통한 지연 방지 대응 시간 확보
- 공정 기간 단축을 위한 배원 계획 및 인력 재분배
- 지속적 학습 체계 구축을 통한 알고리즘 정교화

학습 데이터

◆ 학습을 위한 입력 데이터 분석

블록 특성
- ○ 자재, 공정 기간, Process수 등
- ○ 23개 특성
- ○ PCA 분석 → 대표 특성 5개 도출

시계열 특성
- ○ 공정 기간 (%)
- ○ 자재 불출율
- ○ 투입 공수율
- ○ 입력 진도율

학습 데이터

Block ID	ID_1	ID_1	ID_1	…	ID_n	…
특성 1	-2748	-2748	-2748		$X_{1,n}$	
⋮	⋮	⋮	⋮		⋮	
특성 5	151	151	151		$X_{5,n}$	
공정 기간	0%	42.86%	57.14%		$t_{n,m}$	
자재 불출율	100	100	100		$X_{6,n,t_{n,m}}$	
투입 공수율	29.78	39.84	56.8		$X_{7,n,t_{n,m}}$	
입력 진도율	52.29	65.38	75.63		$X_{8,n,t_{n,m}}$	

분석 알고리즘 선정

◆ 모든 알고리즘을 학습, 모델 평가 지표로 비교 & 분석

연구 Specification	Algorithm		
	Classification	NN	
		LSTM	DNN
예상 공정 지연 유무로 블록 단위 공정을 분류하는 알고리즘	Y	Y	Y
시계열 데이터 예측의 용이성	W	S	W
공정 기간, 자재 등 블록 특성 고려의 용이성	W	W	S
적은 양의 데이터에 따른 상대적 기대 학습 수준	S	W	W

(Classification : SGD, Logistic Regression, SVM, Decision Tree, Random Forest)
(Y : Yes, W : Weak, S : Strong)

모델 평가 지표

◆ 평가 지표

- **Accuracy**
  - 가장 직관적인 모델의 성능을 평가하는 지표
- **Precision**
  - 모델이 True로 예측한 데이터 중 실제 True 비율
  - <u>손실 최소화의 관점</u>에서 모델이 True로 예측했으나 False인 데이터가 가장 치명적인 오류

$$Acc = \frac{TN + TP}{TN + FP + FN + TP}$$

$$Pre = \frac{TP}{TP + FP}$$

		예측	
		False	True
실제	False	TN	FP
	True	FN	TP

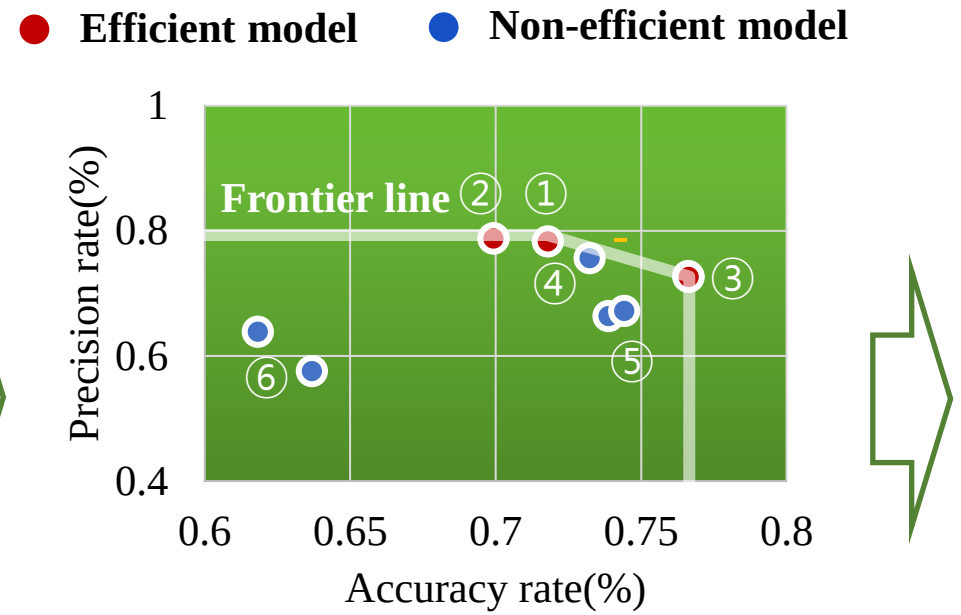
◆ 평가 방법 : DEA(Data Envelopment Analysis)

최적 모델 선정

◆ DEA 분석을 통한 학습 변수 최적화

- 알고리즘 별 hyper-parameter tuning 결과를 DEA로 알고리즘 별 효율적인 모델 집합 도출
- 효율적인 모델 집합에서 Test Data로 성능 지표 재평가

Acc	Pre	param
		①
		④
		②
		⑤
		③
		⑥



<DEA Frontier Graph>

Model	Acc	Pre	param
LSTM			①
LSTM			②
LSTM			③
DNN			
DNN			
DTC			
RF			
⋮			

차원 축소 트릭

- Classification과 DNN의 시계열 예측 용이성 극복 방안
  - 특성의 조합과 레이블의 상관관계 : 시계열 데이터와 Time step 조합
  - Time step에 대해 Linear Synchronization하여 시계열 데이터 특성 데이터로 변환

$$X_{n,m} = \left(X_{1,n}, \cdots, X_{5,n}, X_{6,n,t_{n,m}}, X_{7,n,t_{n,m}}, X_{8,n,t_{n,m}}, t_{n,m}\right)$$

Linear Synchronization of time step

$$Let,\ X'_{6,n,t_{n,m}} = \frac{X_{6,n,t_{n,m}}}{t_{n,m}},\ X'_{7,n,t_{n,m}} = \frac{X_{7,n,t_{n,m}}}{t_{n,m}},\ X'_{8,n,t_{n,m}} = \frac{X_{8,n,t_{n,m}}}{t_{n,m}}$$

Dimension Reduction

$$X'_{n,m} = \left(X_{1,n}, \cdots, X_{5,n}, X'_{6,n,t_{n,m}}, X'_{7,n,t_{n,m}}, X'_{8,n,t_{n,m}}\right)$$

◆ 알고리즘 DEA 성능 지표

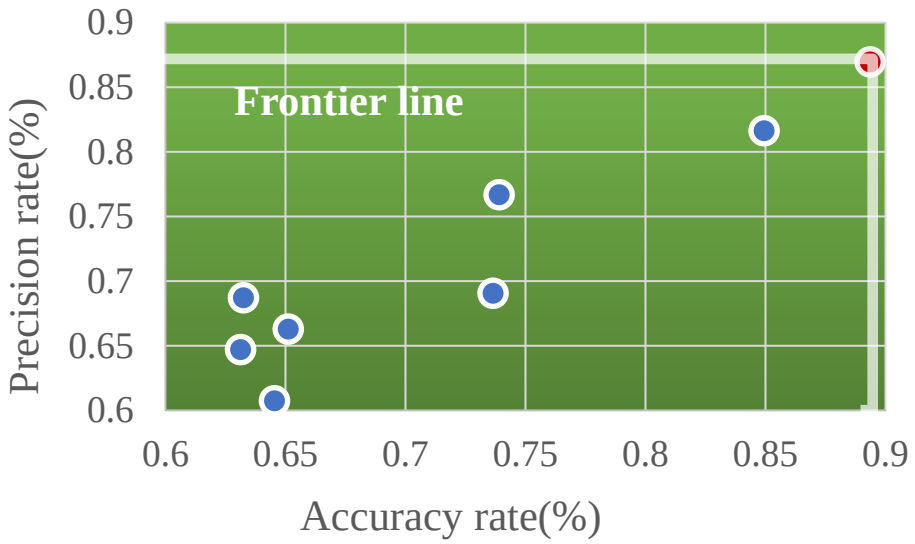
$X_{n,m}$ 학습 알고리즘

Model	Acc	Pre	Efficiency
SGD	0.67359	0.62156	0.75363
Linear SVM	0.67333	0.6436	0.75334
Kernel SVM	0.6388	0.74302	0.84497
Kernel SVM	0.66658	0.6902	0.78490
Kernel SVM	0.69411	0.67007	0.77659
Kernel SVM	0.70995	0.66636	0.79431
Logistic Regression	0.67203	0.61723	0.75189
Decision Tree	0.82342	0.76636	0.92127
Random Forest	**0.89379**	**0.87935**	**1**
DNN	0.71191	0.73736	0.83853

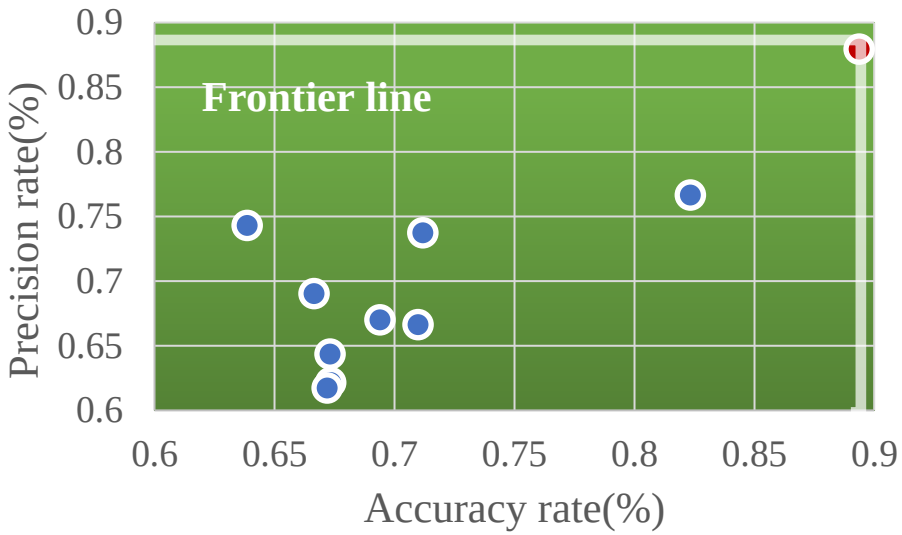
$X'_{n,m}$ 학습 알고리즘

Model	Acc	Pre	Efficiency
SGD	0.63256	0.68702	0.790169
Linear SVM	0.63136	0.64706	0.744209
Kernel SVM	0.65122	0.66276	0.762266
Kernel SVM	0.73668	0.69045	0.824239
Logistic Regression	0.6455	0.60748	0.722222
Decision Tree	0.84953	0.81609	0.950502
Random Forest	**0.89377**	**0.86946**	**1**
DNN	0.739165	0.766667	0.881773

◆ DEA Efficiency Frontier 시각화



<$X_{n,m}$ Algorithm Frontier Graph>



<$X'_{n,m}$ Algorithm Frontier Graph>

공정 구간 별 성능 지표 평가

◆ 최적화 모델의 공정 구간 별 성능 및 한계

Time step	0~20%	20~40%	40~60%	60~80%	80~100%
Accuracy	0.8476	0.92332	0.92432	0.93164	0.86853
Precision	0.82972	0.9177	0.91633	0.93304	0.84971

<공정 기간 별 랜덤 포레스트 알고리즘 성능 지표>

- 종료일에 가까워질수록 정확도가 높아질 것이라는 일반적인 견해
- 랜덤 포레스트 지도학습은 공정 중간에 더 높은 Accuracy와 Precision 측정
- 종료일에 가까울수록 precision의 오류에 의한 손실 극대화의 한계

결론 및 향후 연구

결론

- 알고리즘 내부 hyper-parameter tuning 시 발생하는 성능 평가 기준인 DEA 기법과 상관없이 두 성능 모두 최상의 지표를 가진 랜덤 포레스트 알고리즘 개발
- 효과적인 사전 예측과 더불어 손실 최소화 기대

한계 및 향후 연구

- 산업 특성 상의 데이터 잡음에 의한 정교한 분석의 한계 존재
- 인공 신경망과의 큰 차이는 빅 데이터의 부재와 함께, 데이터 고유의 특성의 결과로 추정
- 데이터 축적과 함께 지속적인 알고리즘 평가 및 비교 요구
- 공정 후반 Accuracy와 Precision의 하락에 따른 산업에 실제 손실 평가의 필요성

참고문헌

참고문헌

- 권세혁, 머신러닝 기법을 활용한 대용량 시계열 데이터 이상 시점탐지 방법론 : 발전기 부품신호 사례 중심, 산업경영시스템학회지 43, no. 2 (2020): 33-38
- Malhotra, Pankaj, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long Short Term Memory Networks for Anomaly Detection in Time Series. Paper presented at the Proceedings, 2015.
- Géron, Aurélien. Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.