정보통신기술 기반 제조 시계열 데이터의 자동 변수 선택 기법과 효율성 평가

장재석, 정찬영, 김균엽, 정하일*

인터엑스, *인터엑스

jay.jang@interxlab.com, cy.jung@interxlab.com, gy.kim@interxlab.com, *hail.jung@interxlab.com

Jaeseok Jang, Chanyoung Jung, Gyun Yeop Kim, Hail Jung*

INTERX, *INTERX

요 약

본 연구는 제조 공정 자율화 및 자동 모델링을 위한 접근으로, STL, PCA, DTW, 클러스터링 기반의 시계열 패턴 분석을 통해 불필요한 변수를 제거하고, 이를 통해 모델의 예측 정확도와 추론 속도를 동시에 향상시키는 자동 변수 선택 기법을 제안합니다.

Ⅰ. 서 론

제조 산업에서 인공지능(AI)의 도입은 생산성 향상, 품질 고도화, 비용 절감 등을 실현할 수 있는 핵심 전략으로 주목받고 있으며, 특히 불량률 저감과 공정 최적화를 위한 실시간 예측 및 품질 분석 시스템은 스마트 제조의 기반 기술로 빠르게 확산되고 있습니다[1-3].

그러나 제조 공정은 설비 구성, 공정 조건, 원재료 등 다양한 요인에 따라 공정마다 상이한 특성을 보이며, 이로 인해 생성되는 시계열 데이터 역시 각기 다른 구조와 복잡한 패턴을 갖습니다. 이러한 데이터는 다변량 구조, 추세와 계절성, 급격한 변동성, 노이즈 등 복합적인 특성을 내포하고 있어, 단순한 통계 요약이나 기존 피처 추출 기법으로는 중요한 정보를 충분히 반영하기 어렵고, 이를 그대로 활용할 경우 예측 모델 성능 저하로 이어질 수 있습니다[4-5]. 더불어 시계열 모델의 높은 차원과 복잡한 구조는 학습 및 추론 속도를 늦춰 실시간 대응이 요구되는 제조 환경과 충돌할 수 있습니다. 특히 초 단위 생산 주기를 가진 공정에서는 예측 속도가 이를 초과할 경우 실효성이 떨어지게 됩니다. 따라서 정확도 향상 및 추론 속도 개선을 위해 공정별로 특화된 모델 설계가 필요합니다[6-7].

이에 본 연구에서는 제조 시계열 데이터의 특성을 효과적으로 반영할 수 있는 패턴 기반 자동 변수 선택 기법을 제안하고자 합니다. 구체적으로는 STL(Seasonal-Trend decomposition using LOESS) 기법을 활용하여 데이터의 계절성과 추세를 분리하고, 결함 제품과 정상 제품 간의 주요 차이를 분석하기 위해 주성분 분석(PCA)을 적용합니다. 이어서, 시간 축의 주요 패턴 차이를 정량적으로 파악하기 위해 Dynamic Time Warping(DTW) 기반 유사도 계산과 클러스터링 기법을 활용함으로써, 예측에 실질적으로 기여하는 핵심 변수들을 자동으로 선별합니다.

이러한 접근 방식은 전체 시계열 정보를 모두 사용하는 기존의 고비용 모델들과 달리, 예측에 필요한 정보만을 선별적으로 활용함으로써 연산 효율성을 극대화할 수 있습니다. 또한, 시계열 데이터 내의 의미 있는 패턴을 기반으로 변수 선택이 이루어지기 때문에, 도메인 지식이 부족하더라도 모델이 데이터를 효과적으로 학습하고 일반화할 수 있는 구조를 갖추게 됩니다.

따라서 본 연구는 단순한 예측 성능 향상을 넘어, 실시간 대응성과 추론 속도, 그리고 모델 구축 비용 측면에서 제조 산업의 실질적인 요구를 충족시킬 수 있는 실용적인 해법을 제시하고자 합니다.

Ⅱ. 본론

본 연구에서는 양품(정상 제품)과 불량 제품 간의 시계열 패턴 차이를 기반으로 주요 변수를 자동으로 선별하는 시계열 특화 Feature Selection 방법을 제안합니다. 본 방법론은 STL 기반 시계열 분해, 주성분 분석(PCA), Dynamic Time Warping(DTW) 기반 유사도 분석, 클러스터링 기반 주요 변수 추출의 네 단계로 구성됩니다.

첫 번째 단계에서는 원본 제조 데이터를 양품과 불량 데이터로 분리한 후, 각 센서의 시계열 데이터를 STL 기법을 통해 계절성과 추세성, 잔차로 분해합니다. STL은 반복적인 로컬 회귀 기반 분해 방식을 통해 데이터의 장기 추세와 주기성을 효과적으로 분리할 수 있으며, 시계열의 구조적 패턴을 정밀하게 분석하는 데 유용합니다. 제조 데이터는 센서별로 시계열 길이가 상이한 경우가 많기 때문에, 분석의 일관성을 확보하기 위해 zero-padding을 적용하여 길이를 통일합니다. 이를 통해 모든 시계열 데이터를 동일한 기준으로 비교할 수 있도록 전처리를 수행합니다.

두 번째 단계에서는 STL로 분해된 양품 및 불량 데이터 각각의 Trend 및 Residual 성분에 대해 주성분 분석(PCA)을 수행합니다. PCA는 고차원 시계열 데이터를 저차원 공간으로 투영하여 주요 정보를 유지하면서 데이터의 차원을 줄이는 기법입니다. 이를 통해 각 센서별로 양품과 불량 데이터를 대표하는 주성분을 각각 1개씩 추출하며, 이는 이후 유사도 비교를 위한 핵심 입력으로 사용됩니다.

세 번째 단계에서는 추출된 주성분 기반의 양품/불량 시계열 간 유사도를 측정하기 위해 Dynamic Time Warping(DTW) 알고리즘을 적용합니다. DTW는 서로 다른 길이나 시간 축이 어긋난 시계열 데이터를 비선형적으로 정렬하여 가장 유사한 경로를 찾아내는 알고리즘으로, 시계열 데이터 간 정렬된 패턴 유사도를 정량적으로 계산할 수 있습니다. 특히, zero-padding으로 인한 노이즈 영향을 최소화할 수 있어 제조 환경에 적합합니다. DTW는 센서별 양품과 불량 데이터의 Trend 및 Residual 성분에 대해 각각 유사도를 계산하며, 최종 누적 비용을 유사도의 기준값으로 사용합니다.

네 번째 단계에서는 앞서 계산된 DTW 유사도를 기반으로 모든 센서를 클러스터링하여 유사도 수준에 따라 세 개의 그룹으로 분류합니다. DTW는 상대적 거리값이기 때문에 절대적인 판단 기준으로 사용하기 어렵습니다. 따라서 유사도가 가장 높은 1군(Class 1)은 제거하고, 나머지 Class 2와 Class 3에 해당하는 센서를 유의미한 변수(중요 변수)로 정의합니다.

이와 같은 과정은 Trend와 Residual 각각에 대해 개별적으로 수행되며, 두 구성 요소에서 모두 Class 2 또는 3으로 분류된 센서만을 최종 Feature로 선택합니다. 이 과정을 통해 추출된 변수들은 실제로 양품과 불량 간 패턴 차이를 명확히 반영하고 있는 변수들이며, 시계열 예측 모델 학습 시 불필요한 변수 제거로 인해 계산 효율성과 예측 성능을 동시에 확보할 수 있습니다.

아래 표는 Kaggle 오픈소스 데이터인 Tabular Playground Series – April 2022를 활용한 모델링 결과를 나타낸 것입니다. 해당 데이터셋은 13개의 센서로 구성된 다변량 시계열 데이터로, 총 26,000개의 샘플을 포함하며 이진 분류 문제를 위한 데이터입니다.

표 1 Tabular Playground Series – April 2022를 활용한 모델링 결과

Method	Model	Accuracy	Recall	F1-Score
AutoGluon.TS	DirectTabular	0.6645	0.6484	0.658
AutoGluon.TS	AutoETS	0.5023	0	0
AutoGluon.TS	DeepAR	0.4974	0.967	0.6569
AutoGluon.Tabular	Weighted -Ensemble	0.7925	0.8673	0.8062
AutoGluon.Tabular	NeuralNet	0.7918	0.8717	0.8065
AutoGluon.Tabular	CatBoost	0.7965	0.8461	0.8054
PyCaret	Best model	0.728	0.755	0.734
TPOT	Best model	0.761	0.715	0.751
TS DL	Bi-LSTM	0.8369	0.8199	0.8421
Our TS DL	**Bi-LSTM**	**0.8919**	**0.9169**	**0.895**

표에 따르면, 기존의 자동화 모델링 툴(AutoGluon, PyCaret, TPOT)은 시계열 데이터를 활용한 예측에서 충분한 정확도를 보여주지 못했으며, 일반적인 시계열 모델이 오히려 더 우수한 성능을 나타냈습니다. 더 나아가 제안된 Feature Selection 기법을 적용한 결과, 전체 13개 센서 중 10개의 센서만이 의미가 있는 Feature로 추출 되었습니다. 이를 활용하여 동일한 모델 구조에 모델링한 결과 F1 Score가 기존 대비 5% 이상 향상되었고, 10개 데이터 기준 추론 시간도 14초에서 12.7초로 줄어 약 9.3%의 연산 시간 절감 효과를 확인할 수 있었습니다.

이처럼 본 방법론은 공정별 패턴의 차이를 정량적으로 분석하고, 정보량이 높은 변수만을 선별적으로 활용함으로써 기존의 전체 시계열 입력 방식에 비해 연산 효율성과 실시간 예측 성능을 크게 향상시킵니다. 또한 도메인 지식 없이도 시계열 분석과 모델링이 가능하다는 점에서, 제조 현장에서의 AI 도입 장벽을 낮추는 데 실질적인 기여할 수 있습니다.

Ⅲ. 결론

본 연구에서는 제조 산업에서 발생하는 시계열 데이터를 보다 효과적으로 처리하기 위해, 양품과 불량 제품 간의 시계열 패턴 차이를 기반으로 주요 변수를 선별하는 자동화된 Feature Selection 기법을 제안합니다. 시계열 데이터는 일반적으로 고차원, 다변량 구조를 가지며, 시간에 따른 추세와 계절성, 센서 간 상호작용이 복잡하게 얽혀 있어, 모든 변수를 그대로 활용하는 기존 모델은 예측 정확도뿐만 아니라 추론 시간 측면에서도 한계를 가집니다. 특히 실시간 품질 판별이 요구되는 제조 환경에서는 모델의 성능뿐 아니라 예측 속도 또한 매우 중요한 요소로 작용합니다.

이를 해결하기 위해 본 논문이 제안한 방법은 예측에 불필요한 변수를 제거함으로써 모델 입력의 차원을 줄이고, 복잡한 시계열 분석에 필요한 계산량을 줄이는 동시에, 모델이 보다 중요한 정보에 집중하도록 유도합니다.

실제 실험 결과에서도 제안한 시계열 Feature Selection 기법의 유효성이 입증되었습니다. Kaggle 오픈소스 이진 분류 데이터셋을 활용한 실험에서는, 전체 센서를 사용하는 모델 대비 제안된 Feature Selection 기반 시계열 딥러닝 모델(TS DL)의 성능과 추론 속도가 모두 향상되었습니다. 같은 구조의 모델에서 제안된 방법을 적용한 모델의 경우 기존 모델 대비 정확도와 F1 점수는 대략 5% 상승하였으며, 추론 시간은 약 9.3% 단축되었습니다. 이 결과는 의미가 없는 변수만 제거하더라도 모델 복잡도와 연산량이 의미 있게 줄어들 수 있음을 보여주며, 시계열 데이터 내 불필요한 정보의 제거가 실시간 예측 성능 향상에 기여함을 시사합니다.

또한 이러한 방법은 도메인 전문가의 수작업 없이도 주요 변수를 자동으로 추출할 수 있어, 제조 AI 시스템의 도입 및 확산에 있어 진입 장벽을 낮추고 실용성을 높여줍니다. 특히 생산 주기가 수 초 단위인 고속 제조 라인에서는 1초 이내의 예측 수행이 필수적인데, 제안된 시계열 Feature Selection 기반 모델은 경량화된 구조를 통해 이러한 요구를 충족할 수 있는 잠재력을 보여줍니다.

향후 연구에서는 현재의 Feature Selection 기반 구조를 더욱 경량화된 딥러닝 모델(GRU, CNN, Transformer 등)과 결합하여, 정확도 손실 없이 예측 속도를 더욱 개선할 수 있는 방안을 모색할 예정입니다. 이를 통해 다양한 공정 조건과 제품 특성 변화에 신속하게 대응하고, 실시간 의사결정이 가능한 지능형 품질 관리 시스템 구축에 실질적으로 기여할 수 있을 것으로 기대됩니다.

ACKNOWLEDGMENT

본 논문은 울산시-ETRI 2차 공동협력사업의 일환으로 수행되었음. [25AB1600, 제조 혁신을 위한 주력산업 지능화 기술 개발 및 산업현장에서의 사람-이동체-공간 자율협업지능 기술 개발]

참 고 문 헌

[1] Cioffi R. et al., "AI and machine learning in smart production," *Internal Report on Sustainable Manufacturing*, pp. 492, 2020.

[2] Peres R. S. et al., "Industrial AI in Industry 4.0: Review and outlook," *Technical Report*, pp. 220121–220139, 2020.

[3] Wang B., "The future of manufacturing," *Engineering White Paper*, pp. 722–728, 2018.

[4] Fatima S. S. W. and Rahimi A., "Forecasting algorithms for industrial systems," *Machines Internal Document*, p. 380, 2024.

[5] Gerling A. et al., "AutoML for error analysis in manufacturing," *Proc. ICEIS*, 2022.

[6] Hsu C.-Y. and Liu W.-C., "CNN for fault diagnosis in semiconductor manufacturing," *Semiconductor Industry Workshop*, pp. 823–836, 2021.

[7] Farahani M. A. et al., "Time-series classification in smart manufacturing," *RCIM Working Draft*, p. 102839, 2025.

[8] Cleveland R. B. et al., "STL: A seasonal-trend decomposition," *Workshop on Statistical Decomposition*, pp. 3–73, 1990.

[9] Pearson K., "On closest fit to systems of points in space," *Philosophical Magazine*, pp. 559–572, 1901.

[10] Müller M., "Dynamic time warping," *Tutorial Notes on Information Retrieval*, pp. 69–84, 2007.