경영정보학연구
제26권 제2호, 2024년 5월 http://dx.doi.org/10.14329/isr.2024.26.2.083

DTW 거리 기반 kNN을 활용한 시계열 데이터 정보 추출 및 회귀 예측

Exploring Time Series Data Information Extraction and Regression using DTW based kNN

양 현 준 (Hyeonjun Yang) 광주과학기술원 전기전자컴퓨터공학부 학부생
임 채 국 (Chaeguk Lim) CJ올리브네트웍스 AI연구소 연구원
정 우 혁 (Woohyuk Jung) CJ올리브네트웍스 AI연구소 연구원
우 지 환 (Jihwan Woo) CJ올리브네트웍스 AI연구소 연구소장, 고려대학교 기술경영전문대학원 겸임교수, 교신저자

요 약

본 연구는 도금욕 공정의 완성도 예측을 위한 시계열 데이터의 효과적인 표현을 목표로, Dynamic Time Warping(DTW) 및 k-Nearest Neighbors(kNN) 기반의 전처리 방법론을 제안한다. 제안된 DTW 기반 kNN 전처리 방법을 다양한 회귀 모델에 적용하여 비교한 결과, 기존 결정 나무(Decision tree) 대비 최대 RMSE에서 43%과 MAE에서 24% 개선된 성능 향상을 보였으며, 신경망 구조를 갖는 회귀 모델과 결합했을 때 성능 향상이 두드러졌다. 본 논문에서 제안하는 전처리 방법과 회귀 모델을 결합한 구조는 길이가 긴 시계열 데이터와 제한된 데이터 샘플이 있는 상황에서 적합할 것으로 사료되며, 데이터가 부족한 상황에서도 과적합의 위험을 감소시키며, 합리적인 예측을 가능하게 함을 시사한다. 그러나 DTW 및 kNN 알고리즘은 데이터 샘플이 많아질수록 연산량이 늘어난다는 한계가 존재하며, 향후 연구를 통해 이러한 계산 효율성의 문제를 개선할 수 있는 연구가 필요할 것으로 보인다.

키워드 : 가상 계측, 시계열 데이터, 특징 추출

Ⅰ. 서 론

1.1 연구 배경

제조 분야에서는 설비 센서 장비를 통해서 데이터의 수집 공유가 활발하게 일어나고 있다. 센서 설비를 통해 수집된 데이터를 일정한 시간별로 누적, 축적하여 시각화나 변수 추출을 통해 제조 설비를 효율적으로 운영하는 것이 제조 분야의 핵심이 되어가고 있다(Oh *et al.*, 2020). 제조 분야에서는 여러 설비에 부착된 센서로 부터 수집한 시계열 데이터를 활용해 설비의 현재 상태를 실시간으로 확인할 수 있도록 한다. 설비 센서 데이터를 통해 설비의 고장을 조기에 진단 할 수 있으며, 이는 정비 시 원인을 파악하고 이에 대한 대응을 좀 더 효율적으로 운용 가능하게 한다(Yang *et al.*, 2019). 제조 분야에서 활용되고 있는 스마트 팩토리 시스템에는 각 설비 센서에서 수집된 데이터를 통해

공정의 불량 여부를 사전에 확인할 수 있으며 제조 분야의 다양한 센서에서 시계열 데이터 분석이 요구되고 필요해짐에 따라 다양한 방법론이 제안되어 왔다. 특히 GPU의 컴퓨팅 리소스의 비약적인 발전과 Cloud 스토리지의 보편화에 따라 저장할 수 있는 데이터의 양이 증가하고 데이터를 저장하는 비용이 절약되면서 많은 시계열 데이터 분석 방법론과 알고리즘들이 제안되고 있다(Ismail Fawaz *et al.*, 2020). 우선 시계열 간 연관성을 통계적 기법으로 분석하는 연구들이 있다. Jeong *et al.*(2011)은 두 개의 시계열 데이터의 거리 및 유사도를 계산하여 최적화된 index를 매칭하여 추론한 후 이를 바탕으로 길이가 짧은 데이터에 distance 기반 분석을 수행한 연구를 진행하였으며, Ahn *et al.*(2017)은 시계열 사이의 관련성을 계산하기 위해 카이 제곱 통계량, t-검정 통계량, 왈드 통계량 등의 통계적 지표 등을 활용하였으며, filtering을 활용해 중복성이 낮은 특징을 선택하고 시계열 데이터를 분류하는 연구를 진행하였다. 시계열 데이터 특성상 데이터 포인트가 방대한 경우가 많고 특히 센서 데이터와 같은 경우는 초단위, 밀리세컨즈 단위의 데이터 간격을 가지는 경우도 있기 때문에 보통 분석에서 분류 및 예측은 통계적 방법론과 선형 모델과 같은 비교적 간단한 방법론들을 위주로 사용되고 있다. 하지만 최근에는 딥러닝 알고리즘을 활용해 시계열 데이터의 분류 및 예측 연구가 활발하게 진행되고 있다. Smirnov and Nguifo(2018)은 Multi-Layer Perceptron(MLP), Fully Convolutional Neural Network (FCN)을 시계열 데이터 분류에 사용하였으며, Wenninger *et al.*(2019)은 Residual Network(ResNet)를 사용한 연구를 진행하였으며, MLP, FCN, ResNet 모두를 사용해 센서 시계열 데이터를 분류한 연구도 있다.

하지만 위와 같은 연구들은 보통 빅데이터 기반의 알고리즘에 초점을 맞춘 연구들이 대부분이며, 데이터가 적은 상황에서 시계열 데이터를 분석하는 새로운 연구는 상대적으로 찾기 힘든 상황이다. 따라서 본 논문에서는 별도의 학습이 필요하지 않는 kNN(k-Nearest Neighbor)과 다양한 회귀 모델을 결합하여 적은 데이터에서도 DNN(Deep Neural Network)과 같은 신경망 모델을 활용할 수 있는 하이브리드 방법론을 제시하고자 한다.

1.2 도금욕 공정 품질 연구 배경

이노징크 및 세라믹 아연도금 처리의 첫 번째 단계인 산제 전처리는 제품 표면의 산화층을 제거하고, 공정들의 밀착성을 향상시키는 전처리 방법이다. 이 과정에서, 용액 농도 및 수소 가스 침투로 인해 품질 저하가 발생할 수 있으며, 이에 따라 KAMP 제조 플랫폼에서는 도금 공정에서 발생하는 데이터를 분석하여 산제 전처리 공정 운영 최적화 모델을 개발하고 있다.

본 논문에서 데이터를 수집한 KAMP(Korea AI Manufacturing Platform)는 중소벤처기업을 위한 제조 데이터 플랫폼으로, 중소 벤처기업의 인공지능 개발 및 활용을 위해 데이터 셋을 구축하며, 인공지능 모델을 제공한다. KAMP에서 선행적으로 진행되었던 기존 연구에서는 결정 나무(Decision Tree) 모델을 사용하였으며, 평균 제곱근 오차(Root Mean Square Error) 3.16의 오차율의 성능을 보였다.

본 연구에서는 시계열 데이터가 수집되는 공정을 평가하는 단일 종속 변수를 예측하는 상황에서, 시계열 데이터의 전반적인 특성을 고려할 때 더 높은 성능을 보인다는 것을 밝힌다. 또한, 시계열 데이터의 특성을 단일 수치 값(Numeric Value) 으로 축소 치환하는 방안으로 DTW(Dynamic Time Warping)을 거리 함수로 한 kNN(k-Nearest Neighbor) 모듈을 제시한다.

kNN 모듈을 통해 축소 치환된 시계열 데이터의 특성은 Regression Layer의 입력으로 사용되며, 이 특성을 통해 종속 변수를 예측하는 방법론을 제시한다. 이는 기존의 시계열 데이터의 특성을 부각시키며 기존의 단일 방법론들보다 높은 성능을 가짐을 보였다는 데에 그 기여점을 갖는다.

본 논문은 총 5장으로 이루어져 있다. 제Ⅱ장에

서는 관련 연구를 상세히 소개하고, 제Ⅲ장에서는 본 연구의 프로세스와 활용한 방법론, 제Ⅳ장에서는 실험 세팅과 모델 결과를 서술하였다. 마지막으로 제Ⅴ장에서 결론 및 시사점과 추후 연구를 서술하였다.

Ⅱ. 문헌 연구

산업에서 진행된 공정 가상 예측 방법에 관한 선행연구와 본문에서 제안하는 모델의 핵심 알고리즘인 kNN(k-Nearest Neighbor)과 DTW(Dynamic Time Warping)의 선행연구에 대해서 정리한다.

2.1 가상 계측 관련 선행연구

가상 계측이란, 제조 공정에서 발생하는 다양한 센서 데이터 및 장비 데이터, 공정 이력을 기반으로 제품의 품질을 평가하는 분야로, 전통적인 하드웨어 계측 장비를 대체하거나 보완하기 위해 컴퓨터를 사용하여 계측 및 제어 작업을 수행하는 방법을 가리킨다. 특히, 계측에 있어 시간적, 재화적 비용이 큰 분야에서 관련 연구가 많이 진행되었다. 송세리, 박상철(2019)은 LCD 공정의 데이터를 이용하여 5개의 은닉층(hidden layer)과 250개 노드를 최종 선택한 DNN(Deep Neural Network) 예측 모델을 구축하였으며, 다중 회귀 모델(Multiple Linear Regression), 랜덤 포레스트(Random Forest), DNN (Deep neural Network)의 성능을 비교 분석하여, Feature selection을 통한 DNN에서의 성능이 같은 조건에서의 다중 회귀 모델, 랜덤 포레스트 모델에 비해 각각 11.18%p, 3.09%p 높은 성능을 보인 것을 확인하였다. 한정석, 김형근(2022)의 연구에서는 반도체 공정에 XGBoost 알고리즘을 적용하여 제품의 양품과 불량을 구분하는 모델을 활용하였다. 결과적으로 정확성과 검출력은 높지만 재현성이 낮은 이유로, 제조 공정의 도메인 지식을 바탕으로 한 주요 변수의 필요성을 시사하였으며, 훈련 데이터 밖의 새로운 데이터를 예측하는 데에 트리 구조의 한계점이 존재한다는 것 또한 확인할 수 있었다.

최근에는 시계열 데이터에도 머신 러닝 및 AI를 적용하는 연구가 많이 진행되고 있다. Jung *et al.*(2020)에서는 시계열 데이터를 AI를 활용해서 분석하였으며, 이상우 등(2021)의 연구에서는, 머신러닝 적용을 위한 센서 시계열 데이터 전처리 과정으로, 센서서 값의 시계열 값과 시계열 값의 1차 미분 데이터, 센서 응답의 상승 구간 평균값과 피크 값 그리고 센서 응답의 하강 구간 평균값, 센서 응답의 피크 값 총 3가지 방식의 전처리 방식을 활용하였다.

2.2 KNN 관련 선행연구

kNN(k-Nearest Neighbor) 알고리즘은 Cover, and Hart(1967)로부터 처음 고안되었으며, 이미 다양한 분야에서 성공적으로 활용되어 검증 받은 기계학습 방법 중 하나이다. kNN 알고리즘은 인스턴스 기반 학습 알고리즘으로, 주어진 훈련 데이터를 저장하고, 새로운 관측치에 대한 예측을 수행할 때, 가장 유사한 k개의 이웃 관측치를 선택하여 그들의 레이블을 기반으로 예측한다. 이 알고리즘은 일반적으로 분류 문제에서 이산적인 클래스를 예측하는 데에 사용되지만, 연속형 속성을 가진 회귀 문제에서도 효과적으로 활용된다. 분류 방식과 유사한 방법으로, 특성이 유사한 k개의 이웃 관측치를 선택하여 그들의 연속형 값의 평균으로 회귀를 진행한다. 실제로 Lora *et al.*(2007)의 연구에서 kNN 회귀 기법은 전기 가격 예측, 전기 부하량 예측과 같은 응용 분야에서 수치 예측을 목적으로 한 연구에서 활발하게 활용되어 왔다. 기존 선형 회귀(Linear regression), 의사 결정 트리(Decision Tree), 서포트 벡터 머신(Support Vector Machine)과 달리 모델을 훈련하지 않고 학습 데이터 자체를 메모리에 보관하고 예측할 때마다 모든 학습 데이터와의 거리를 계산해야 하므로 학습 데이터가 증가할수록 예측 속도가 상대적으로 느릴 수 있다는 한계점이 존재한다. 따라서 대량의 데이터에 대해서는 계산 복잡성이 높고, 데이터에

노이즈가 많거나 차원이 높은 경우 성능이 저하될 수 있다는 우려도 존재한다.

2.3 DTW 관련 선행연구

DTW(Dynamic Time Warping) 알고리즘은 Vintsyuk (1968)로부터 제안되었으며, 두 시계열 데이터 간의 유사성을 비교하는 데에 활발히 사용된다. 기존 유클리드 거리(Euclidean distance) 방식이 시간적 불일치를 고려하지 않고 데이터의 각 시간 단계를 독립적으로 처리했다면, DTW는 시간적 불일치를 고려하여, 패턴의 확대, 축소, 시간적 왜곡을 모두 고려하여 유사성을 비교한다. 이 과정에서 동적 프로그램을 사용하기 때문에, 계산이 비교적 복합할 수 있으며 최적 정렬 경로를 찾기 위해 추가적인 계산이 필요하다는 단점이 있다.

그럼에도 시간 차가 존재하는 두 시계열 패턴의 유사성을 잘 검출한다는 점에 있어서, DTW는 음성 인식, 음악 분석, 패턴 인식, 바이오 인포매틱스 등 다양한 분야에서 활용되고 있다(장민석 등, 2015). 이 연구에서는 전력 신호 패턴을 기반으로 가전 기기를 분류하는 과정에서 DTW를 적용하여 여러 전력 신호 패턴 사이의 거리를 효율적으로 계산하였으며, 사전 입력된 기준 데이터 패턴에 포함 되어있지 않는 전력 패턴이 발생한다면 인식을 잘 수행하지 못한다는 단점 또한 실증하였다.

이환철, 허선(2020)의 연구에서는 길이가 다른 두 신호를 DTW을 활용하여, 유사도를 보존한 상태로 시계열 데이터의 길이를 같게 변환하는 방법을 제시하고 있으며, 이를 기반으로 서포트 벡터 머신(Support Vector Machine; SVM), 랜덤 포레스트(Random Forest; RF), 나이브 베이즈(naïve Bayes; NB), 길이를 같게 변환한 시계열 데이터는 인공신경망(Artificial Neural Network; ANN) 등 다양한 분류 알고리즘에 적용하여 기존보다 높은 성능을 보였다는 것을 밝혔다.

또한 김준석 등(2021)의 연구에서는 추진 전동기 잔여 수명을 예측하는 과제에서, DTW을 kNN의 거리함수로 활용한 연구가 진행 되었으며, 기존의 유클리드 거리 함수에 비해 DTW 거리 함수가 시계열 데이터의 유사성을 비교하는 데에 효과적이며, kNN의 회귀 성능에 큰 영향을 미쳤다고 밝혔다.

본 논문에서는 선행 연구에서 시계열 데이터의 유사도를 파악하는 데에 높은 성능을 보였던 DTW 거리 기반의 kNN 모듈을 활용한다. 하지만, kNN으로 직접적으로 회귀를 하는 것이 아닌, 시계열 데이터의 크기를 축소 치환하는 시계열 데이터 전처리 과정에서 활용한다.

Ⅲ. 제안 모델

본문에서는 DTW 기반의 kNN을 활용한 시계열 데이터 전처리 모듈을 제안하며, 총 네 가지 종류의 회귀 모델을 결합하여 본문에서 제안하는 전처리 모듈의 성능에 대해서 간접적으로 평가한다.

3.1 데이터 수집 및 설명

본 연구를 수행하기 위해 활용한 데이터셋은 이노징크 전기도금 공정 중에 발생한 데이터로, 산제 전처리 설비 내에 설치된 센서를 통해 5초 주기로 측정되었다.

각 5분 동안 진행된 총 726개의 공정은 크게 공정별 메타 데이터와 시계열 센서 데이터로 구분된다. 메타 데이터에는 공정의 순서를 의미하는 “LoT”과 예측 변수(Target)인 최종 공정의 완성도를 수치상으로 평가한 “Process Rate”가 있으며, 시계열 데이터에는 공정 과정 동안 센서를 통해 수집한 “Temp(온도)”와 “pH(산성도)” 시계열 데이터로 이루어져 있다. 시계열 데이터는 한 공정 내에서, 69의 시간 길이(Time Length)의 데이터 형식을 지닌다.

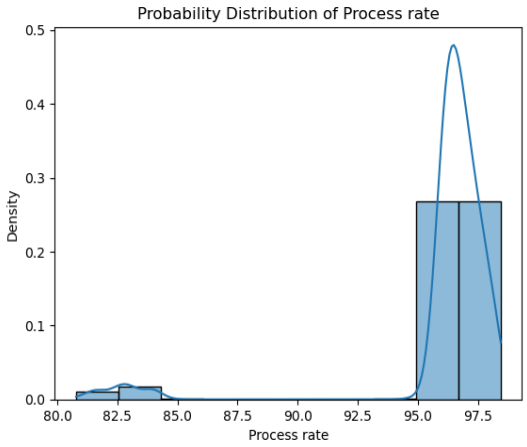
본 연구에서는 총 726개의 공정 샘플에 대해서 LoT(공정 순서), Temp(온도) 시계열 데이터, pH (산성도) 시계열 데이터를 기반으로 Process Rate

(공정 완성도)를 예측한다. 726개의 각 공정 별 LoT과 Process Rate 데이터 형식은 <그림 1>과 같다.

LoT 변수는 하루에 진행되는 22개의 공정을 구분하는 단위로, 1부터 22까지의 정수를 순차적으로 할당한다. Process Rate는 최종 공정 결과물의 완성도를 수치상으로 평가한 것이며, <그림 2>는 Process Rate의 확률 분포를 나타낸다.

	Date	LoT	Process Rate
0	2021-09-06	1	96.38
1	2021-09-06	2	97.4
2	2021-09-06	3	95.4
3	2021-09-06	4	96.35
4	2021-09-06	5	94.77
...	...	...	...
721	2021-10-27	18	97.29
722	2021-10-27	19	97.21
723	2021-10-27	20	98.38
724	2021-10-27	21	98.36
725	2021-10-27	22	96.03

〈그림 1〉 726개의 공정 샘플의 LoT, Process Rate 데이터



〈그림 2〉 Process Rate 확률 분포

하나의 공정 동안, 온도(Temp) 센서와 산성도(pH) 센서는 시간 길이가 69인 시계열 데이터를 수집하며, <그림 3>은 726개의 공정 샘플에서 수집된 모든 시계열 센서 데이터를 표현하며, <표 1>은 시계열 데이터 통계를 나타낸다.

Time	LoT	pH	Temp
2021-09-06 09:01:18	1	1.02	47.18
2021-09-06 09:01:23	1	1.05	47.34
2021-09-06 09:01:28	1	1.09	48.45
2021-09-06 09:01:33	1	1.12	48.46
2021-09-06 09:01:38	1	1.15	48.47
...	...	...	...
2021-10-27 11:14:41	22	2.79	51.83
2021-10-27 11:14:46	22	3.62	42.2
2021-10-27 11:14:51	22	3.4	41.88
2021-10-27 11:14:56	22	3.59	40.62
2021-10-27 11:15:01	22	3.82	42.08

〈그림 3〉 726개의 공정에서 수집된 pH, Temp 센서 데이터

〈표 1〉 전 공정 시계열 센서 데이터 통계

	pH	Temp
Count	50094	50094
Mean	2.006	49.876
Std	0.552	1.345
Min	1.010	38.020
Max	3.90	54.190

3.2 시계열 데이터 전처리

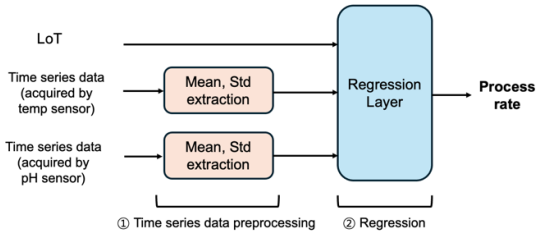
품질 평가는 공정이 모두 완료된 후에 한 차례 이루어지기 때문에, 공정 과정에서 측정된 시계열 데이터를 그대로 회귀 모델의 입력으로 활용하는 것보다, 이를 대표할 수 있는 특성 값으로 축소 치환하여, 회귀 모델에 입력하여 공정 완성도(Process

rate)를 예측하는 접근 방식을 도입하게 되었다.

시계열 데이터를 축소 치환하는 전처리 방법으로 평균과 표준편차를 활용하는 방법과 본문에서 제안하는 DTW based kNN 방법으로 나누어 실험을 진행하였다. 그 후 Regression Layer를 결합하여, 최종적으로 제안하는 DTW based kNN 기반의 시계열 데이터 축소 치환 방식의 성능을 간접적으로 평가하고자 한다.

3.2.1 평균, 표준 편차

시계열 데이터가 정규 분포를 따른다고 가정할 때, 시계열 데이터를 수치 값(Numeric Value)으로 표현할 수 있는 대표적인 방법은 평균과 표준편차이다. <그림 4>는 평균과 표준편차를 시계열 데이터의 특성 값으로 활용하여 회귀 층(Regression Layer)의 입력으로 활용하여 공정 완성도(Process rate)를 예측하는 모델의 전반적인 개략도를 표현한다.

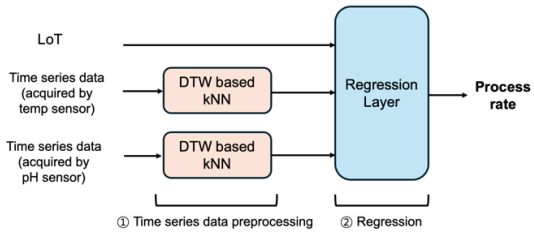


〈그림 4〉 시계열 데이터의 평균, 표준편차를 활용한 회귀 모델 구조

3.2.2 DTW based kNN

입력으로 들어온 두 종류의 시계열 데이터는 서로 다른 DTW 거리 기반의 kNN층을 거치게 된다. 입력으로 들어온 센서와 동일한 종류의 센서 시계열 데이터가 모여 있는 공간에서 kNN 과정이 이루어지며, 입력 데이터와 그래프 개형이 유사한 k개의 시계열이 데이터가 선택되며, 해당 데이터가 속한 샘플 공정의 Process Rate 평균을 취하게 된다.

해당 과정을 통해, 입력으로 들어온 시계열 데이터는 추가적인 학습 없이도, 학습 데이터를 척도로 한 상대적인 값을 갖게 된다. 본문에서는 실험을 통해 k값을 7로 설정하였으며, 이 과정에서 69개의 데이터로 이루어진 시계열 데이터는 크기가 1인 값으로 축소 치환된다. <그림 5>는 DTW 거리 기반의 kNN으로 추출한 특성을 활용하는 회귀 모델의 전반적인 개략도를 표현한다.



〈그림 5〉 DTW based kNN 시계열 데이터 전처리를 활용한 회귀 모델 구조

3.3 회귀 모델 성능 평가 지표

3.3.1 MAE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

MAE의 수식적 전개는 위와 같으며, 실제 값과 예측값 사이에서 발생한 오차를 절댓값으로 변환한 성능지수이다. 오차를 직관적으로 표현한다는 장점이 있지만, 큰 오차에 덜 민감하다는 한계점이 있다.

3.3.2 RMSE

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{n}$$

RMSE의 수식적 전개는 위와 같으며, 오차를 제곱하여 더하기 때문에, 크게 발생한 오차에 대해 더 많은 페널티를 부여한다. 따라서, 크게 벗어난 이상치를 평가하는 데에 유용하며, MAE의 단점을 보완하는 성능 지수로 활용할 수 있다.

Ⅳ. 실험 및 결과

본 연구에서는 726개 공정 샘플에 대한 두종류의 센서에서 발생하는 시계열 데이터와 각 공정별로 할당된 LoT 번호와 최종 결과물의 완성도를 정량 지표화 한 Process Rate 데이터를 활용하였다. 그 중 8:2의 비율로 학습 데이터와 평가 데이터로 분할하여 성능 평가를 진행하였다.

본문에서는 1) 시계열 데이터를 전처리 없이 회귀 모델을 적용한 성능, 2) 평균과 표준편차를 활용한 시계열 데이터 전처리에 회귀 모델을 결합한 성능, 3) DTW based kNN 모듈을 통한 시계열 데이터 전처리에 회귀 모델을 결합한 성능으로 나누어서 비교하였다. 회귀 모델로는 Decision tree, Linear regression, DNN, XGBoost으로 구성하였으며, DNN은 2-hidden layer 구조로, learning_rate(0.01)의 Adam optimizer로 학습을 진행하였다.

최종 회귀 성능을 통해, 본문에서 제안하는 DTW based kNN을 통한 시계열 데이터 전처리 성능을 간접적으로 평가하고자 하였다.

〈표 2〉 시계열 데이터 전처리 없이 적용한 회귀 모델 성능

Regression Layer	RMSE	MAE
Decision tree(Baseline)	3.16	1.34
Linear regression	2.94	1.60
DNN	4.27	1.87
XGBoost	3.19	1.41

〈표 3〉 시계열 데이터의 평균, 표준편차를 활용한 회귀 모델 성능

Preprocess	Regression Layer	RMSE	MAE
Mean & Std	Decision tree	3.04	1.23
	Linear Regression	2.08	1.34
	DNN	2.17	1.47
	XGBoost	4.99	1.05

〈표 4〉 DTW based kNN 시계열 데이터 전처리를 적용한 회귀 모델 성능(ours)

Preprocess	Regression Layer	RMSE	MAE
DTW based kNN(k=7)	Decision tree	2.87	1.14
	Linear regression	1.79	1.02
	DNN	1.94	1.08
	XGBoost	2.16	1.04

4.1 시계열 데이터 전처리 효과

시계열 데이터 전처리 없이 회귀 모델을 단독으로 사용했을 때와 비교했을 때, 두 방식의 시계열 데이터 전처리(평균과 표준편차, DTW based kNN)를 거친 회귀 성능이 전반적으로 개선되었음을 확인할 수 있었다. 특히 이러한 개선은 회귀 모델이 신경망 구조 DNN일 때 더 두드러졌는데, 기존에 큰 길이의 시계열 데이터를 신경망에 적용했을 때에는 학습 데이터에 과적합되어 테스트 데이터에서 오차가 많이 발생하는 것을 예상할 수 있었다. 반면, 전처리 과정을 수반한 회귀 모델은 시계열 데이터를 축소 치환하는 전처리를 통해 신경망 구조의 입력 차원을 줄일 수 있게 됨으로써, 복잡한 신경망을 단순화하여 과적합 현상을 완화시키는 효과가 있었고 해석할 수 있었다.

4.2 DTW based kNN 전처리 효과

또한, 전반적으로 DTW based kNN 모듈을 통해 시계열 전처리를 거친 회귀 모델이 평균과 표준편차를 활용한 회귀 모델보다 더 개선된 성능을 보이는 것을 확인할 수 있었다. Decision Tree, Linear Regression, DNN, XGBoost를 결합했을 때 RMSE에서 각각 6%, 14%, 11%, 57%만큼 개선된 성능을 보였으며, MAE에서는 7%, 24%, 39%, 1%만큼 개선된 모습을 보였다. 이는 DTW based kNN 전처리 방식이 기존에 길이가 69인 시계열 데이터를 더 낮은 차원의 크기가 1인 단일 값으로 축소함과 동시에, 시계열 데이터의 대표성을 포함하고 있다는

주요한 지표로 인식하였다.

본 실험을 통해, 시계열 데이터를 DTW 거리 기반의 kNN을 통해 축소 치환하는 전처리 과정이 회귀 모델의 성능에 직접적인 영향을 미친다는 것을 확인할 수 있었다. 시계열 데이터를 전처리 과정에서는 Instance 기반의 kNN을 활용하여 시계열 데이터를 축소 치환하고, 그 위에 회귀 모델을 쌓는 하이브리드 방법론이 효과가 있음을 보였다. 부가적인 효과로, 시계열 데이터 전처리 과정을 통해 방대한 양의 시계열 데이터가 단일 값으로 축소 치환됨으로써, 최종 회귀 모델의 입력 차원이 줄어들어, 적은 학습 데이터로도 신경망 구조의 회귀 모델을 학습할 수 있다는 여지 또한 남긴다.

본 실험에서는, 제안하는 시계열 데이터 전처리 성능을 확인하기 위해, 여러 회귀 모델을 결합하여, 성능을 간접적으로 측정한다. 최근에는 높은 성능을 보이는 다양한 딥러닝 방법론들이 등장하고 있지만, 샘플이 적은 실험 데이터 특성상 제한적인 회귀 모델만 사용했다는 한계점 또한 존재한다.

Ⅴ. 결 론

본 연구에서는 도금욕 공정 중 발생하는 시계열 데이터의 정보를 활용하여 공정의 완성도를 예측하는 가상 계측 분야에 DTW(Dynamic Time Warping)를 기반으로 한 kNN(k-Nearest Neighbors) 시계열 데이터 전처리 방법을 제안한다. 이 방법론은 다양한 회귀 모델들과 결합하여 성능을 비교했을 때, 기존에 사용된 결정 나무(Decision Tree) 모델에 비해 RMSE(Root Mean Square Error)는 최대 43%, MAE(Mean Absolute Error)는 24%까지 성능이 개선되었음을 보여주며, 특히 신경망 구조의 회귀 모델과 결합될 때 성능 개선이 두드러졌다.

본문에서 제안한 전처리 방법과 회귀 모델을 결합한 구조는, 시계열 데이터의 길이가 길고 학습에 사용할 수 있는 데이터 샘플이 부족한 상황에서 적합할 것으로 보이며, 인스턴스 기반의 kNN을 활용하여 시계열 데이터를 축소 치환하는 방식은 뒤따르는 회귀 모델의 과적합을 방지하며, 제한된 데이터로도 품질을 예측하는 데 효과적일 것으로 보인다. 이러한 방법론은 공정 완성도 예측뿐만 아니라, 다른 시계열 데이터를 다루는 다양한 분야에서도 활용 가능성을 제시한다.

하지만, DTW와 kNN 알고리즘은 상대적으로 많은 연산량을 필요로 하며, 데이터 샘플이 많아질수록 추론 과정에서 DTW를 거리 함수로 사용하는 kNN 모듈의 연산 시간이 증가한다는 한계 역시 존재한다. 이러한 한계를 보완하기 위해, 연산 최적화를 비롯한 추가적인 연구와 노력이 필요할 것으로 보인다.

또한, 최근에 높은 성은 성능을 보이는 다양한 딥러닝 기반 방법론이 등장하고 있지만, 본 실험에서는 샘플이 적은 데이터 특성상, 비교적 단순한 회귀 모델을 결합하여 실험 하였다. 따라서, 진행되는 다음 연구에서는 데이터 확보 및 kNN 연산 최적화를 통해, 대량의 데이터에서 최근 등장한 딥러닝 기반의 모델들과 결합한 성능에 대한 평가가 필요할 것으로 보인다.

참 고 문 헌

[1] 김준석, 이강복, 황회선, 안지수, 오정림, 장명훈, 전홍배, “DTW 기반 추진 전동기 잔여수명 예측 알고리즘 개발 사례연구”, *한국CDE학회 논문집*, 제26권, 제4호, 2021, pp. 386-397.

[2] 송세리, 박상철, “LCD 검사 공정에서 가상계측을 위한 머신 러닝 기반 예측 모델”, *한국CDE학회논문집*, 제24권, 제3호, 2019, pp. 329-338.

[3] 이상우, 김병희, 서영호, “계단응답 데이터 전처리 방식에 따른 머신러닝 기반 화학물질분류 시스템의 분류특성평가”, *한국정밀공학회학술발표대회논문집*, 2021, pp. 416-416.

[4] 이환철, 허선, “효과적인 시계열 데이터 분류를

위한 동적시간왜곡 기반의 시계열 길이 변환", *대한산업공학회지*, 제46권, 제4호, 2020, pp. 356-364.

[5] 장민석, 공성배, 고락경, 정주영, 주성관, "Dynamic Time Warping(DTW)기법을 이용한 가전기기별 부하 패턴 분류 기초연구", *대한전기학회학술대회논문집*, 제2015권, 제7호, 2015, pp. 45-46.

[6] 한정석, 김형근, "반도체 공정에서 가상계측 위한 XGBoost 기반 예측모델", *한국정보처리학회 학술대회논문집*, 제29권, 제1호, 2022, pp. 477-480.

[7] Ahn, G. S., H. C. Lee, and S. Hur, "Feature selection method for multivariate time series data classification", *Journal of the Korean Institute of Industrial Engineers*, Vol.43, No.6, 2017, pp. 413-421.

[8] Cover, T. and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol.13, No.1, 1967, pp. 21-27.

[9] Ismail Fawaz, H., B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, Geoffrey, I. Webb, L. Idoumghar, P. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification", *Data Mining and Knowledge Discovery*, Vol.34, No.6, 2020, pp. 1936-1962.

[10] Jeong, Y. S., M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification", *Pattern Recognition*, Vol.44, No.9, 2011, pp. 2231-2240.

[11] Jung, S. H., G. J. Gu, D. Kim, and J. W. Kim, "Predicting stock prices based on online news content and technical indicators by combinatorial analysis using CNN and LSTM with self-attention", *Asia Pacific Journal of Information Systems*, Vol.30, No.4, 2020, pp. 719-740.

[12] KAIST(ABH, Impix), AI Dataset for Process Operation Optimization, KAMP(Korea AI Manufacturing Platform), Korea, 2022, Available at https://www.kamp-ai.kr/.

[13] Lora, A. T., J. C. Riquelme, J. L. M. Ramos, J. M. R. Santos, and A. G. Expósito, "Influence of kNN-Based load forecasting errors on optimal energy production", *Progress in Artificial Intelligence*, Vol.2902, 2003, pp. 189-203.

[14] Lora, A. T., J. M. R. Santos, A. G. Expósito, J. L. M. Ramos, and J. C. R. Santos, "Electricity market price forecasting based on weighted nearest neighbors techniques", *IEEE Transactions on Power Systems*, Vol.22, No.3, 2007, pp. 1294-1301.

[15] Oh, C., S. Han, and J. Jeong, "Time-series data augmentation based on interpolation", *Procedia Computer Science*, Vol.175, 2020, pp. 64-71.

[16] Smirnov, D. and E. M. Nguifo, "Time series classification with recurrent neural networks", *Advanced Analytics and Learning on Temporal Data*, Vol.8, 2018.

[17] Vintsyuk, T. K., "Speech discrimination by dynamic programming", *Cybern Syst Anal*, Vol.4, 1968, pp. 52-57.

[18] Wenninger, M., S. P. Bayerl, J. Schmidt, and K. Riedhammer, "Timage-A robust time series classification pipeline", *International Conference on Artificial Neural Networks*, 2019, pp. 450-61.

[19] Yang, C. L., Z. X. Chen, and C. Y. Yang, "Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images", *Sensors*, Vol.20, No.1, 2019, p. 168.

Information Systems Review
Volume 26 Number 2
May 2024

Exploring Time Series Data Information Extraction and Regression using DTW based kNN

Hyeonjun Yang* · Chaeguk Lim** · Woohyuk Jung*** · Jihwan Woo****

Abstract

This study proposes a preprocessing methodology based on Dynamic Time Warping (DTW) and k-Nearest Neighbors (kNN) to effectively represent time series data for predicting the completion quality of electro-plating baths. The proposed DTW-based kNN preprocessing approach was applied to various regression models and compared. The results demonstrated a performance improvement of up to 43% in maximum RMSE and 24% in MAE compared to traditional decision tree models. Notably, when integrated with neural network-based regression models, the performance improvements were pronounced. The combined structure of the proposed preprocessing method and regression models appears suitable for situations with long time series data and limited data samples, reducing the risk of overfitting and enabling reasonable predictions even with scarce data. However, as the number of data samples increases, the computational load of the DTW and kNN algorithms also increases, indicating a need for future research to improve computational efficiency.

Keywords: Virtual Instrumentation, Time Series Data, Feature Extraction

* Undergraduate Student, School of EECS, Gwangju Institute of Science and Technology
** Researcher, CJ OliveNetworks AI Research
*** Researcher, CJ OliveNetworks AI Research
*** Corresponding Author, Head, CJ OliveNetworks AI Research, Adjunct Professor, School of Management of Technology, Korea University

◐ 저 자 소 개 ◑

양 현 준 (jjun8030@gm.gist.ac.kr)

현재 광주과학기술원 전기전자컴퓨터 소속 학부생으로, 시계열 데이터 분석과 자연어 처리에 관한 연구에 관심이 있다.

임 채 국 (chaeguk.lim@cj.net)

현재 CJ올리브네트웍스 AI연구소에서 및 데이터 분석 연구를 수행중이다. 주요 연구 분야로는 Computer Vision, Data Analytics 등이다.

정 우 혁 (wh.jung2@cj.net)

현재 CJ올리브네트웍스 AI연구소에서 Natural language processing 및 데이터 분석 연구를 수행중이다. 주요 연구 분야로는 NLP, Data Analytics, HCI, Bio signal processing 등이다.

우 지 환 (jihwan_woo@korea.ac.kr)

삼성전자 삼성 리서치, 미국 카네기멜론대학 로봇연구소 등을 거쳐 현재 CJ 올리브네트웍스 AI연구소에서 연구소장으로 재직 중이며, 고려대학교 기술경영 대학원의 겸임교수이다. 컴퓨터 비전 기반 로보틱스 및 AI 연구 개발 경력이 있으며, 이와 함께 기술 전략 수립 및 가치 평가 등에 관심이 있다.

논문접수일 : 2023년 12월 01일 **게재확정일 : 2024년 03월 04일**
1차 수정일 : 2024년 01월 22일 **2차 수정일 : 2024년 02월 22일**