Smart Media Journal / Vol.13, No.8 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2024.13.8.58

온라인 학습법을 활용한 석탄화력 발전소의 가스 터빈 내 질소산화물(NOx) 배출량 예측

(Nitrogen Oxide (NOx) Emissions Prediction of Gas Turbine in Coal-Fired Power Plant Using Online Learning Method)

박 진*, 고창완, 정영선*****

(Jin Park, Changwan Ko, Young-Seon Jeong)

요 약

질소산화물(NOx)은 대기 오염의 주요 원인 물질로, 오존과 초미세먼지의 형성을 유발하여 건강에 해로운 영향을 준다. 석탄화력 발전소에서는 NOx 등 다양한 유해 물질이 발생하고 있고 그에 대한 정확한 예측은 매우 중요하다. 지금까지는 오프라인 학습법에 기반한 연구가 주류를 이루었고, 또한 초기 데이터가 부족한 상황을 고려한 연구는 존재하지 않았다. 본 연구는 온라인 학습 방법을 활용하여 화력발전소의 NOx 배출량을 제안한다. 온라인 학습법은 새로운 관측치가 발생할 때마다 모델을 학습하여 초기 데이터가 부족한 상황에서도 높은 예측 정확도를 보이는 모델이다. 오픈 데이터를 사용하여 훈련 데이터가 적은 상황을 가정해 실험을 진행하였으며, 오프라인 방법론과 비교한 결과 본 연구에서 적용한 온라인 학습법이 가장 우수한 예측 성능을 보였다.

■ 중심어 : 온라인 학습법 ; 질소산화물 ; 회귀 예측 ; 석탄화력발전소

Abstract

Nitrogen oxides(NOx) in coal-fired power plants are significant contributors to air pollution, influencing the formation of ozone and fine particulate matter, thereby adversely affecting health. Therefore, accurate prediction of NOx emissions is essential. Existing researches have mainly performed based on off-line learning methods, leading to poor prediction performance with the limited training dataset. This paper proposes the online learning model of online support vector regression to predict NOx emissions from coal-fired power plants. Online learning model, which updates a model whenever new observations come out, demonstrates high prediction accuracy even when initial data is scarce. The experimental results showed that the performance of online learning prediction was better than existing off-line learning methods. The results indicated online learning method is a valuable tool for predicting NOx emissions, especially in situations where initial data is limited and data is continuously updated in real-time.

■ keywords : Online Learning Method ; Nitrogen Oxides ; Regression Prediction ; Coal-Fired Power Plant

Ⅰ. 서 론

미세먼지는 농도가 10$\mu g/m^3$ 증가할 경우 전체 질병의 사망률을 4% 증가시키는 주요 대기오염 물질이며, 그중 초미세먼지의 대표적인 발생 원인은 질소산화물(NOx)이다[1]. NOx는 대기 중에서 광화학적 반응을 일으켜 오존을 생성하고, 미세먼지의 주요 성분인 초미세먼지(PM2.5)를 생성한다[2]. 발생한 오존과 초미세먼지는 인간의 호흡기 질환을 악화시키고, 심장 질환과 같은 다양한 건강 문제를 유발하고 있다[3, 4]. 또한

* 학생회원, 전남대학교 산업공학과; ** 정회원, 전남대학교 산업공학과; *** 정회원, 전남대학교 산업공학과 및 아트&디자인 테크놀로지 협동과정 교수
이 논문은 전남대학교 학술연구비(중견일반연구) 지원을 받아 수행된 연구임 (No. 2023-1123-01). 또한 한국연구재단의 지원을 일부 받아 수행함 (No. NRF-2022R1F1A1063174).
접수일자 : 2024년 07월 15일
수정일자 : 2024년 08월 05일
게재확정일 : 2024년 08월 20일
교신저자 : 정영선 e-mail : young.jeong@jnu.ac.kr

NOx는 산성비와 광화학 스모그의 주요 원인 물질이다[5, 6]. 환경부는 대기오염물질 저감을 통해 대기질이 개선된다면, 대기오염으로 인한 조기 사망자의 수가 약 52.2% 감소할 것으로 추정하였다. 또한 지역 주민의 건강 개선으로 5조 668억 원의 편익이 생길 것으로 계산하였다[7]. 이를 위하여 NOx의 배출량 저감이 중요하다.

2024년 6월 기준, 국내에서는 발전을 위한 원료로 LNG와 유연탄이 전체 56% 이상의 가장 큰 비중을 차지하고 있어, 화력발전이 국내 발전에서 중요한 역할을 하고 있다[8]. 그러나 이 때문에 화력발전의 결과로 NOx와 같은 물질이 발생한다. 환경부는 석탄 화력 발전소의 미세먼지 및 유해 물질 저감을 위한 환경설비 시설에 투자를 늘리고 NOx, 황산화물(SOx), 오존(O3), 탄화수소(HC), 먼지 등 미세먼지 생성물질에 배출허용기준을 강화하며 대기오염 물질의 절대적인 배출량을 감축하고자 앞장서고 있다. 그 예시로 대기오염 물질 배출량에 따라 납부해야하는 '대기배출 부과금'이 존재한다[9]. 사업체는 배출량을 정확히 예측하면 부과금을 예측할 수 있어 지출을 관리할 수 있다. 결과적으로 초과 부과금 발생을 막아 비용을 줄일 수 있으며 배출량을 저감할 수도 있어 정확한 배출량 예측이 필요하다.

화력발전소의 NOx 배출량 예측은 주로 메커니즘 계산을 통한 방법과 통계적 방법이 있다[10]. 최근 주로 사용되는 통계적 방법은 데이터를 바탕으로 알고리즘을 만들어 배출량을 예측한다. 그렇기에 통계적 방법은 물리, 화학과 관련된 지식 없이 데이터 간의 상관관계를 분석하는 데에 의의가 있다. 다만 기존 데이터를 바탕으로 예측하는 과정이기에, 새로운 상황, 갑작스러운 변화에 대응하기가 어려워 위 상황에서 정확도가 떨어질 수 있다.

통계적 방법의 정확도를 높이고자, 기계학습 기법을 통해 NOx의 배출량을 예측하는 연구가 다양하게 진행되었다. 그러나 대부분의 기계학습 연구 방법은 오프라인 학습법에 기반한 방법론으로 대량의 학습 데이터가 필요하며, 데이터가 부족한 경우에는 예측의 정확도가 떨어졌다[11]. 이는 기존 데이터가 없는 경우 초기 배출량을 예측하기 어렵고 실시간으로 변화하는 데이터를 정확히 예측하기 어렵다는 문제가 존재한다.

본 연구에서는 온라인 학습법 방법론인 온라인 서포트 벡터 회귀(Online Support Vector Regression, Online SVR)를 도입하고자 한다. Online SVR은 서포트 벡터 회귀(Support Vector Regression, SVR)기반 온라인 학습법으로 새로운 관측치에 대해 실시간으로 모델에 반영하는 모델이다. 새로운 관측치가 발생할 때마다 지지 벡터(support vector) 여부를 평가하여 모델을 갱신한다. 따라서, 데이터가 적은 상황과 갑작스러운 변화에 대한 유연성이 높으며 예측 성능과 정확도가 높다는 장점이 있다. 본 연구에서는 Online SVR을 활용한 화력발전소 NOx 배출량 예측 모델을 제안한다.

본 논문은 아래와 같이 구성되었다. 2장에서는 국내외에서 진행된 선행 연구에 대해 알아본다. 본 연구와 같은 데이터를 사용하였거나 NOx 배출량을 예측하고자 한 다른 연구와 비교하며 연구의 차별성을 확인한다. 3장은 실험 계획을 서술한다. 실험에 사용된 데이터 세트와 방법론 및 알고리즘 모델에 대해, 4장은 실험 과정을 서술하였다. 5장에서는 실험 결과를 다루고 6장 결론에서 알고리즘의 성능을 보이며 향후 연구 발전 방향성을 제시하였다.

II. 선행연구 고찰

통계적 기법을 활용한 NOx 배출량 예측 성능 분석에 관한 연구는 국내외적으로 활발하게 진행되고 있다.

국내 연구의 경우 진일봉(2018)[12]이 500MW 표준화력발전소의 석탄 연소 보일러에서 NOx 배출량 예측을 위해 주성분 분석(Principal Component Analysis, PCA)과 칼만 필터를 사용

해 데이터 전처리를 하였으며 부분 최소제곱법(Partial Least Squares, PLS)을 사용한 모델과 역전파 신경망(Back Propagation Neural Network, BPNN) 모델을 사용하였다. 칼만 필터로 데이터의 잡음을 없애 데이터의 경향성을 높였으며 주성분 분석을 통해 선발한 8개의 독립변수로 예측을 진행하였고, 결과적으로 BPNN 모델이 PLS 모델보다 더욱 정확한 성능을 보여주었다. 두 모델 모두 전처리 이전에는 예측을 전혀 못 하였으나 전처리 이후 정확하게 예측할 수 있었다.

서미연(2021)[13]에서는 60초와 120초 샘플링 간격으로 황산화물 배출 농도를 예측하였다. Long Short Term Memory network(LSTM) 최적화 모델이 Nonlinear Auto Regressive Neural Network with Exogenous Input (NARX) 모델보다 더 예측 정확도가 높았다. 또한 공정 특성에 맞는 변수와 학습 데이터의 기간 설정 등 최적화 방안이 중요함을 시사하였다.

Leandro(2024)[14]에서는 CO 및 NOx 배출량 예측을 위하여 변수의 가공과 최적 회귀 모델을 제안하였다. 변수 가공은 가스 터빈에서 측정된 데이터의 변수들을 새롭게 가공하여 회귀 모델에 도입하는 방법이다. 변수 가공을 위하여 Uniform Manifold Approximation and Projection (UMAP), t-SNE, PCA라는 방식을 사용하였고, 최종적으로 UMAP 방식으로 가공된 변수를 선택하여 모델에 도입하였다. 두 번째는 배출량 예측을 위한 회귀 모델링 파이프라인을 제시하였다. K겹 교차 검증 (K-fold cross validation) 방식을 통해 매개변수를 선택한 후, Ridge 회귀, K-최근접 이웃 기법, 랜덤포레스트 등 다양한 방법론을 선택하여 CO 및 NOx 배출량을 예측하였다. 위 결과로 매개변수의 선택이 모델링에 큰 영향을 미치며, 앙상블 모델을 활용한 회귀모델을 사용할 시 더욱 예측 결과가 향상될 것을 시사하였다. 또한 제안된 딥 포레스트 회귀(Deep Forest Regression, DFR)모델은 CO 및 NOx 배출 예측 성능이 뛰어났으며 정확도가 더 높은 온라인 모니터링 도구에 사용할 경우 더 좋은 성능을 보일 것임을 시사하였다.

David A.(2023)[15]에서는 복합 사이클 가스 터빈(Combined cycle gas turbines, CCGT) 발전소에서 CO 및 NOx 배출량을 예측한 연구이다. 12개의 다층 퍼셉트론(Multi-Layer Perceptron, MLP) 및 머신러닝 모델을 사용하여 배출량을 예측하였다. 연구 결과로 KNN 알고리즘과 XGBoost(XGB) 모델에서 가장 높은 예측 성능을 보였다.

선행 연구는 모두 오프라인 방법론으로 모델의 학습이 끝나면 더 이상 모델을 수정하지 않는다. 다시 말해서, 데이터의 변동성을 고려하지 않았으며 정적 데이터 환경만을 고려한 연구가 진행되었다. 하지만, 변동성이 큰 환경에서는 새로운 관측치가 발생할 때마다 이를 모델에 반영하는 것이 예측력 향상에 매우 중요하다. 따라서, 본 연구에서는 모델을 실시간으로 수정할 수 있는 온라인 기반 예측 모델, Online SVR을 제안한다.

III. 연구 설계

1. 실험 데이터

본 연구에 사용된 데이터는 터키에서 수집된 데이터 셋인 'Gas Turbine CO and NOx Emission Data Set'이다[16]. 2011년 1월 1일부터 2015년 12월 31일까지 1시간 간격으로 측정된 11개의 센서 데이터를 평균 또는 합계를 낸 데이터로, 총 36,733개가 존재한다[16]. 데이터 셋은 9개의 독립 변수와 2개의 종속 변수로 구성되었다. 독립 변수는 가스 터빈 내 공기 환경 조건(AH, AP, AT)과 CCGT 운전 조건(AFDP, GTEP, TIT, TAT, TEY, CDP)과 관련된 변수가 있으며, 종속 변수로는 CO와 NOx이다. 본 연구에서는 이 중 NOx만을 종속변수로 삼았으며 변수에 관한 정보는 [표 1]을 통해 확인할 수 있다.

표 1. 데이터 변수 설명

변수명	표기	단위	최대	최소	평균	편차
Ambient temperature	AH	℃	37.10	−6.23	17.71	7.45
Ambient pressure	AP	mbar	1036.56	985.85	1013.07	6.46
Ambient humidity	AT	%	100.20	24.08	77.87	14.46
Air filter difference pressure	AFDP	mbar	7.61	2.09	3.93	0.77
Gas turbine exhaust pressure	GTEP	mbar	40.72	17.70	25.56	4.20
Turbine inlet temperature	TIT	℃	1100.89	1000.85	1081.43	17.54
Turbine after temperature	TAT	℃	550.61	511.04	546.16	6.84
Turbine energy yield	TEY	MWH	179.50	100.02	133.51	15.62
Compressor discharge	CDP	mbar	15.16	9.85	12.06	1.09
Carbon oxides	CO	mg/ m3	44.10	0	2.37	2.26
Nitrogen oxides	NOX	mg/ m3	119.91	25.90	65.29	11.68

2. 실험 모델

가. SVR

SVR은 사용자가 설정한 예측 오차 값의 범위 안에 실제 데이터가 포함되도록 하는 최적의 회귀선과 예측값을 찾는 것이 목표이다. Vapnik[17]에 의해 제안된 SVR은 주로 비선형 문제에서 강력한 성능을 보이며, 데이터의 패턴을 학습하고 예측하는 지도학습 회귀 문제에 사용된다.

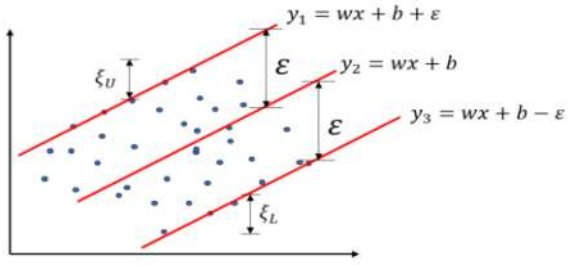


그림 1. 선형 SVR 개념

[그림 1]은 SVR을 시각화한 자료로 파란 점은 데이터이며, y_1, y_3선은 중앙 회귀선의 오차범위를 수용한 상한선과 하한선, y_2선은 예측한 회귀선이다. SVR은 예측값과 실제값 사이의 오차인 ε의 범위 안에서 가능한 많은 데이터가 포함하도록 하는 회귀 모델이다. 즉, 오차가 특정 범위 안에 있으면, 그 오차를 허용한다는 특징이 있다. 데이터를 고차원 공간으로 대응(mapping)하고, 그 공간에서 구조적 위험을 최소화하는 초평면 식(1)과 같은 $H(x)$을 찾아내어 예측을 수행한다. 이때 $H(x)$는 독립 변수의 개수와 같은 차원의 특징 공간 F에 존재한다.

$$H(x) = W^T\phi(x) + b \tag{1}$$

W는 F에서의 벡터이고 x는 데이터의 독립 변숫값, b는 편향(bias)이다. x를 독립 변수의 초평면에서 F로 매핑시키는 함수 $\phi(x)$를 통하여 초평면을 찾아낸다. SVR은 회귀계수의 크기와 실제값과 예측값의 차이를 최소화하는 식을 찾아내는 것이 목표이며 식(2)과 같이 나타낼 수 있다.

$$\min(\|w\|, b) \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^U + \xi_i^L) \tag{2}$$

$$s.t. \quad y_i - (W^T\phi(x) + b) \le \varepsilon + \xi_i^U$$
$$-y_i + (W^T\phi(x) + b) \le \varepsilon + \xi_i^L$$
$$\xi_i^U, \xi_i^L \ge 0, i = 1...l.$$

w는 가중치 벡터이자 회귀계수이다. x_i와 y_i는 i번째에 해당하는 데이터의 독립변수와 종속변수 값을 의미한다. ξ_i^U와 ξ_i^L는 각 마진의 상하 슬랙 변수, ε은 오차 범위, C는 정규화 매개변수이며 사용자가 직접 설정해야 하는 상수이다. SVR은 주어진 오차 범위 ε내에서 가능한 한 많은 데이터를 포함하고자 한다. 이 범위에 벗어나는 데이터에 대해서는 슬랙 변수를 사용하여 허용 오차를 더한다. 이때, C는 이러한 오차에 대한 페널티(penalty)를 부과한다. 페널티란 오차범위 밖에 있는 데이터에게 C값의 배율로 값을 부여하는 것이다. 이를 통해 회귀선의 상한과 하한을 정한다.

나. Online SVR

기존 SVR은 학습 데이터 전체를 한 번에 처리하는 배치학습 방식이다. 그러나 데이터가 매우 크거나 끊임없이 새로운 데이터가 유입되는 상황에서는 이 방식이 비효율적일 수 있다. 이 문제를 해결하기 위하여 데이터를 순차적으로 한

개 또는 작은 배치 단위로 처리하는 방식인 온라인 방식을 도입한 Online SVR이다[18]. Online SVR은 초기 데이터가 적거나 변동성이 큰 상황에 적합하게 설계된 모델이다. 훈련 데이터가 극히 적은 상황일수록 타 모델보다 우수한 정확도를 보여 손해를 줄일 수 있으며, 실시간으로 크게 변화하는 데이터에 대한 예측 성능도 향상할 수 있다. Online SVR의 진행 과정은 다음과 같다[19-21].

$$\begin{aligned} A_{O,B} &= \{\mathrm{X(t)},\mathrm{y(t)}\}_{\mathrm{t=B}}^{0-1} \\ \mathrm{X(t)} &= [x(t),...,x(t-B+1)]^T \\ \mathrm{y(t)} &= x(t+1) \\ t &= 1,2,3,...,O \end{aligned} \tag{3}$$

진행을 위한 시계열 데이터 $x(t)$, 예측 시작 시점 O, 시간 t가 사전에 있으며, 이를 통해 학습 데이터 세트 $A_{O,B}$가 식(3)과 같이 있다. $A_{O,B}$는 과거 시점의 데이터와 다음 시점의 데이터를 포함하는 쌍들로 구성된다. B는 $A_{O,B}$의 임베딩 차원이다. $A_{O,B}$를 통해 예측기 $P(A_{O,B};X)$를 학습한다. 학습한 예측기를 활용하여 예측된 다음 데이터 값은 식(4)과 같다.

$$\hat{x}(O+1) = P(A_{O,B};X(O)) \tag{4}$$

이후 다시 새로운 데이터 $x(O+1)$가 나타날 수 있기에, 예측 시작점을 $O=O+1$로 바꿔준다. 위 과정을 마지막 데이터에 도달할 때까지 반복하며, 반복을 통해 실시간으로 추가되는 데이터를 모델에 반영한다. $P(A_{O,B};X)$는 데이터가 더욱 늘어나면서 모델을 갱신 및 개선해 정확도를 올린다[15]. 만약 새로운 데이터가 예측기 내 SVR 속 서포트 벡터가 될 수 있는 데이터라면, 이를 모델에 반영하여 학습시킨다. 같은 의미로, 데이터가 서포트 벡터가 될 수 없는 데이터라면 모델에는 영향을 끼치지 않는다. 이에 대한 여부는 새로운 데이터가 카루시-쿤-터커(Karush - Kuhn - Tucker conditions, KKT) 조건으로 현재 서포트 벡터에 영향을 미치는지에 따라 달라진다. 조건에 해당하면 모델에 영향을 끼치지 않거나 최소한의 업데이트만이 진행되나, 조건에 해당하지 않을 경우 서포트 벡터에 추가되고 매개변수를 조정한다. 이러한 방식으로 모델 학습의 효율성을 높이고 계산 비용을 낮추었다. 본 논문에서는 SVR과 Online SVR 모두 식(5)과 같은 Radial Basis Function(RBF) 커널을 채택하였다. 데이터의 유사도를 측정하는 RBF 커널은 일반 커널 함수에 비해 파라미터만 결정하면 되는 간단한 구조의 커널이다. 현재 가장 널리 사용되는 커널 함수로, 선행 연구와 본 논문의 사전 실험에서도 타 커널보다 우수한 예측 성능을 보여주어 본 실험에서 해당 커널을 선택하였다[20].

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2),\ \gamma \geq 0 \tag{5}$$

식(5)에서 exp는 지수함수를 뜻한다. γ는 커널 내 지수함수에 쓰이는 매개변수이며 데이터를 뜻하는 x_i와 x_j간의 거리를 조절하여 얼마나 영향을 줄 것인지를 조절하는 역할을 한다. SVR, Online SVR의 매개변수는 γ, C, ε가 있으며, 메타 매개변수 선택을 위한 선행 연구의 공식에 따라 설정하였다[20-22].

Ⅳ. 실험방법 설계

본 연구는 다중 선형 회귀(Multiple linear Regression, MLR), MLP, SVR, 온라인 선형 회귀(Online Linear Regression, OLR), Online SVR을 활용하여 데이터 세트에 대한 예측 성능을 비교하였다[23-25]. 분류하면, 오프라인 모델 3개와 온라인 모델 2개를 선정하였다. 데이터 세트는 이전에 설명한 'Gas Turbine CO and NOx Emission Data Set'를 사용하였다. 그중에서 2013년 데이터 세트를 활용하였으며 데이터 총 수는 7,152개이다. 화력 발전소를 새로 지었거나, 데이터 오염이 발생하여 새롭게 데이터를 수집하는 등 초기 데이터가 5~100개뿐인 상황에서, 이후 생성되는 데이터 500개의 배출량을 예측할 때, 가장 예측 성능이 높은 모델을 찾고자 한다. 실험을 위해 훈련 데이터를 3, 5, 10, 15, 20, 100개로 설정하였으며 평가 데이터는 500개로 고정

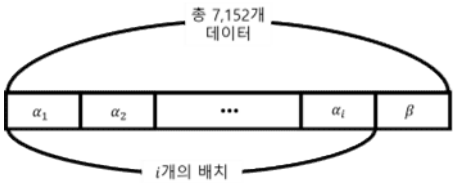


그림 2. 데이터 세트 설명

하였다. [그림 2]와 같이, 전체 데이터를 훈련 데이터의 개수와 평가 데이터 500개를 합한 개수만큼 배치로 나누고, 각 배치에서 훈련 세트와 테스트 세트를 분리하였다. 이때 α는 훈련 데이터와 평가 데이터를 합친 배치이며 시계열 데이터이기에 순차적으로 묶였다. 이때 생성되는 배치의 개수는 i개이다. β는 i개의 α를 만들고 남은 데이터로, 실험에 사용되지 않는다. 성능 지표는 평가 데이터의 실제값과 예측값 사이의 결정계수(R-Squared, R2)값과 평균 제곱근 오차(Root Mean Squared Error, RMSE)값을 계산하여 평가하였다.

V. 실험결과

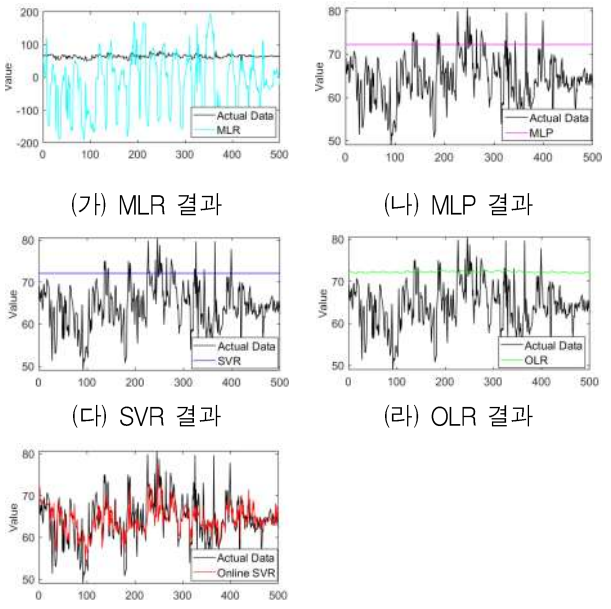
정량적인 결과로 Online SVR은 모든 훈련 데이터 크기에서 다른 모델보다 우수한 성능을 보였다. 특히, 훈련 데이터가 매우 적은 상황에서도 높은 정확도를 유지하였다. 모든 실험의 평균 RMSE 값과 R2값을 각 [표 2]와 [표 3]에 나타냈다. [그림 3]은 3,031번째 데이터부터 3,035번째 데이터를 훈련 데이터로 사용하고 3,036번째 데이터부터 3,535번째 데이터까지 평가 데이터로 사용한 7번째 배치의 예측 결과를 시각화하여 나타낸 그림이다. [그림 3]의 (가), (나), (다), (라), (마)는 각각 MLR, MLP, SVR, OLR, Online SVR 모델의 예측 결괏값을 시각화한 자료이다. [그림 3]의 (다), (마)의 매개변수로

표 2. 각 모형별 RMSE 값

학습 데이터	MLR	MLP	SVR	OLR	Online SVR
3개	36.521	13.861	13.359	11.525	6.644
5개	53.455	13.193	12.035	7.126	6.877
10개	66.803	13.051	11.326	9.392	7.214
15개	37.685	16.609	10.743	14.246	8.557
20개	39.278	84.445	10.351	13.873	8.161
100개	11.866	66.012	7.591	13.657	7.252

표 3. 각 모형별 R2

학습 데이터	MLR	MLP	SVR	OLR	Online SVR
3개	0.648	0.951	0.958	0.974	0.991
5개	0.489	0.962	0.967	0.990	0.990
10개	0.546	0.961	0.971	0.983	0.989
15개	0.631	0.898	0.973	0.963	0.983
20개	0.635	0.759	0.977	0.966	0.986
100개	0.933	0.457	0.988	0.966	0.989



(가) MLR 결과 (나) MLP 결과

(다) SVR 결과 (라) OLR 결과

(마) Online SVR 결과

그림 3. 훈련 데이터가 5개일 때 7번째 배치 결과

γ, C, ε가 있으며 각각 '5', '75.0395', '1.6747'으로 설정하였다. 이는 본 연구의 반복 실험을 통해 도출되어 가장 우수한 성능이 나온 값이다. 각 그림 속 직선은 평가 데이터에 대한 값으로, 검은색 선이 실제값, 색 선이 예측값이다.

MLR은 다중의 독립변수와 종속변수 사이에 선형의 관계가 있다는 전제로 관계식을 수립하여 예측하는 모델이다[21]. [그림 3]의 (가)를 보면, MLR은 배출량을 음수로 예측하거나 무작위의 수를 선정하는 모습을 보인다. 다만 [표 2], [표 3]에서, 훈련 데이터가 100개가 되는 시점에서 성능이 개선되어 모델이 예측할 수 있음을 보여주었다. 이는 MLR이 초기 데이터가 부족한 상황에서 불안정하다는 것을 알 수 있다.

SVR과 MLP은 [그림 3]의 (나), (다)와 같이 두 모델 모두 실험 결과에서 초기 데이터가 적어 학습을 따라가지 못하여 하나의 값으로만 예측하였다. 그러나 평균 성능 지표상으로는 준수한 성능을 보여준다. 수렴되는 하나의 예측값은 대체로 모델의 학습 과정 중 예측값의 평균에 가까우며, 이는 전체 실제값에서 크게 벗어나지 않은 값이다. 성능 지표의 결과는 좋아 보일 수 있지만, 실제로는 한 값으로 수렴하기에 예측 모델로 부족하다. SVR은 초기 데이터가 100개가 되자 준수하게 예측을 진행하였으나, MLP는 더욱 큰 편차로 예측에 실패하였다.

[그림 3]의 (라)는 Online SVR과 마찬가지로 온라인 예측을 진행한 모델이다. 실시간으로 데이터가 들어오면 선형회귀 예측을 진행하며 예측마다 계산을 위한 매개변수를 갱신한다[24, 25]. 같은 선형회귀 모델인 MLR와 비교했을 때, 더 안정적으로 모델이 진행됨을 알 수 있다. 그러나 SVR, MLP와 같이 한 값으로 수렴하려는 경향을 보인다.

[그림 3]의 (마)를 보았을 때, Online SVR 모델은 다른 비교 모델보다 실시간으로 들어오는 데이터를 모델에 반영하여 학습하며, 초기 데이터가 부족한 상황에서도 높은 예측 정확도를 유지함을 보인다. 이는 학습 데이터에서 구성한 지지 벡터를 통해, 새로운 데이터가 나타날 때마다 지지 벡터 갱신 여부를 판단하며 모델을 개선해왔기 때문이다. KKT 조건을 통한 판단은 갑작스러운 변화에도 잘 적응하여 좋은 예측 성능을 보였다. 이러한 결과는 데이터가 부족한 상황을 가정하였을 때, Online SVR이 NOx 배출량 예측에 있어 오프라인 모델보다 더 효율적이고 신뢰할 수 있는 모델임을 시사한다. 같은 온라인 모델인 OLR과 비교했을 때도, 비선형 데이터 처리에 강점이 있는 Online SVR이 유의미한 예측을 보여주었다.

Ⅵ. 결 론

본 논문에서는 온라인 학습 방법론을 활용한 화력발전소 질소산화물(NOx) 배출량 예측 모델을 제안하였다. 이 모델은 실시간으로 발생하는 데이터를 반영하며, 초기 데이터가 부족한 상황에서도 우수한 예측 성능을 발휘하도록 설계되었다. 제안한 모델은 초기 데이터의 개수가 3, 5, 10, 15, 20, 100개인 상황에서 500개의 평가 데이터를 예측하였고, 이를 전체 데이터에 맞추어 평균 13번 반복 수행하였다. 실험 결과는 결정계수 값과 평균 제곱근 오차값을 기준으로 다중 선형 회귀, 다층 퍼셉트론, 서포트 벡터 회귀(Support Vector Regression, SVR), 온라인 선형 회귀, 온라인 서포트 회귀 모형(Online Support Vector Regression, Online SVR) 모델과 비교하여 성능을 측정하였다. 그 결과, 오프라인 모델에서는 초기 데이터가 부족한 경우 학습에 필요한 조건과 매개변수들이 부족하여 학습되지 않았으며 Online SVR 모델이 다른 모델들보다 우수한 예측 성능을 보였다.

본 연구에서는 Online SVR 모델이 실시간 데이터를 반영하며, 초기 데이터가 부족한 상황에서 꾸준한 모델 개선을 통해 높은 예측 성능을 보인다는 것을 기존 오프라인 모델과 비교하여 증명하였고, 같은 온라인 예측일지라도 단순 회귀분석보다 비선형 데이터에 강점이 있는 SVR 방식이 더 우수함을 확인할 수 있었다. 이는 기존 데이터가 매우 부족한 상황에서 정확한 예측 성능을 보인다는 점과 시사하는 바가 크다. 위와

같은 점이 NOx 배출량 예측에 있어 실시간 변화에 빠르게 적응할 수 있는 모델의 장점을 보여주었다.

본 연구에서는 특정 데이터 세트를 사용하였으나, 다양한 발전소와 환경에서 수집된 데이터 세트를 적용하여 모델의 일반화 성능을 평가할 필요가 있다. 또한 Online SVR의 성능을 더욱 향상하기 위해 매개변수 최적화, 다른 머신러닝 알고리즘과의 결합 등 다른 방안을 모색할 수 있다. 위와 같은 향후 연구를 통해 현장에서 더 효율적이고 정확한 예측이 가능할 것이라 기대한다.

REFERENCES

[1] 이재호, 임성호, 김진호, 송호영, “미세먼지 저감 동향, *전자통신동향분석*, 제34권, 제2호, 83-91쪽, 2019년 4월

[2] 김미연, 김형근, 박진철, “초미세먼지 저감을 위한 광촉매 외장도료의 질소산화물 제거 성능 현장 실험 분석,” *설비공학논문집*, 제32권, 제12호, 585-592쪽, 2020년 12월

[3] S.Maji, S.Ahmed, W.A.Siddigui, S.Ghosh, “Short term effects of criteria air pollutants on daily mortality in Delhi,” *tmospheric Environment*, vol.150, pp.210-219, Feb. 2017.

[4] 김선미, 임명진, 신주현, “미세먼지와 진료과목의 상관관계 분석을 통한연관성 예측 방법,”*스마트미디어저널*, 제7권, 제3호, 22-28쪽, 2018년 1월

[5] S.Sudalma, P.Purwanto, L.W.Santoso, “The effect of SO2 and NO2 from transportation and stationary emissions sources to SO42− and NO3 − in rain water in Semarang“, *Procedia Environmental Sciences,* vol. 23, pp. 247-252, Jan. 2015.

[6] 정용승, 정재섭, “서울 수도권 지역의 광화학오존에 관한 연구,“ *한국대기환경학회지*, 제7권, 제3호, 169-179쪽, 1991년 12월

[7] 환경부, “2차 수도권 대기환경관리 기본계획(2015-2024)“, (2015-2024) 수정계획, 2022년

[8] 발전설비 전력통계정보시스템(2024), https://epsis.kpx.or.kr/epsisnew/selectEkpoBftChart.do?menuId=020100, (accessed Jun., 13, 2024).

[9] 대기환경보전법, 법률 제18905호

[10] P.Tüfekci, “Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods”. *Int. J. Electr. Power Energy Syst.*, vol .60, pp. 126-140, Sep. 2014.

[11] 김상하, “학습 데이터 부족 환경에서 데이터 중요도가 반영된 유사 데이터 활용을 통한 학습 성능 개선 방안 연구,” *서강대학교 정보통신대학원*, 2022년 8월

[12] 진일봉, “PCA기반 데이터 전처리 방법을 이용한 500MW 표준 석탄 화력 발전소에서의 NOx 배출량 예측방법,”*한양대학교*, 2018년 2월

[13] 서미연, “석탄화력 발전소 탈황공정 운전 Data를 활용한 성능예측모델 적용 연구: Prediction Model of Desulfurization Efficiency of Coal-Fired Power Plants Based on RNN,” *서울대학교 공학전문대학원*, 2021년 2월

[14] L.D.S. Coelho, H.V.H. Ayala, V.C. Mariani, “CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering,” *Fuel*, vol. 355, Jan. 2024.

[15] D.A.Wood, “Long-term atmospheric pollutant emissions from a combined cycle gas turbine: Trend monitoring and prediction applying machine learning,” *Fuel*, vol. 343, Jul. 2023.

[16] H.Kaya, P.Tüfekci, E.Uzun, “Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS,” *Turk. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 6, pp. 4783-4796, Jan. 2019.

[17] H.Drucker, C.J.Burges, L.Kaufman, A.Smola, V.Vapnik, “Support vector regression machines“, *Adv Neural Inf Process Syst*, vol.9, December 1996.

[18] J.Ma, J.Theiler, S.Perkins, “Accurate on-line support vector regression,” *Neural computation*, vol. 15, no. 11, pp. 2683-2703, Nov. 2003.

[19] L.J.Tashman, “Out-of-sample tests of forecasting accuracy: an analysis and review,” *Int. J. forecast*, vol. 16, no. 4, pp. 437-450, Oct. 2000.

[20] H.Wang, D.Xu, “Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function,” *J. control sci. eng. (Online)*, pp. 183-193, Nov. 2017.

[21] V.Cherkassky, Y.Ma, “Selection of meta parameters for support vector regression“, *Artificial Neural Networks−ICANN 2002*, pp. 687-693, Jan. 2002.

[22] V.Cherkassky, Y.Ma, “Practical selection of SVM parameters and noise estimation for SVM regression,” *Neural networks*, vol.17, no.1, pp. 113-126, Jan. 2004.

[23] 채철주, 이현조, 김용기, 구현정, “농업 공공 빅데이터를 이용한 머신러닝 기반 생산량 및 판매 수익금 예측,” *스마트미디어저널*, 제11권, 제4호, 19-29쪽, 2022년 5월

[24] F.Zhdanov, V.Vovk, “Competitive Online Generalized Linear Regression under Square Loss”, *in ECML PKDD, Machine Learning and Knowledge Discovery in Databases*, pp. 531-546, Barcelona, Spain, Sep. 2010.
[25] V.Vovk, “Competitive on line statistics,” *International Statistical Review*, vol. 69, no. 2, pp. 213-248, Aug. 2001.

저 자 소 개

박 진(학생회원)
2024년 전남대학교 산업공학과 학사 졸업
2024년~현재 전남대학교 산업공학과 석사과정

<주관심분야 : 데이터마이닝, 스트리밍 데이터, 연속 학습>

고창완(학생회원)
2019년 전남대학교 산업공학과 학사 졸업
2021년 전남대학교 산업공학과 석사 졸업
2022년~현재 전남대학교 산업공학과 박사과정

<주관심분야 : 통계적 데이터마이닝, 불확실성 데이터 분석, 머신러닝>

정영선(정회원)
1997년 전남대학교 산업공학과 학사 졸업
2001년 고려대학교 산업공학과 석사 졸업
2011년 뉴저지주립대학교 산업시스템공학과 박사 졸업
2014년~현재 전남대학교 산업공학과 교수

<주관심분야 : 통계적 데이터마이닝, 반도체 공정 자동화>