딥러닝 회귀 모델 개발을 위한 센서 데이터 윈도우 사이즈 최적화 기법 1)

최민서 [1], 유동연 [2], 이정원 [1,2]
[1] 아주대학교 전자공학과
[2] 아주대학교 AI 융합네트워크학과
minseo24@ajou.ac.kr, dongs0125@ajou.ac.kr, jungwony@ajou.ac.kr

Optimization of Sensor Data Window Size for Deep Learning Regression Model

Min-Seo Choi[1], Dong-Yeon Yoo[2], Jung-Won Lee[1,2]
[1]Dept. of Electrical and Computer Engineering, Ajou University
[2]Dept. of AI Convergence Network, Ajou University

요 약

센서 데이터의 중요성이 커지면서 센서 데이터 처리 연구의 수요가 증가하고 있다. 센서 데이터 기반의 딥러닝 모델 개발 시, 센서 데이터 단일 값에 의한 출력이 아닌 시계열적인 특성을 반영하여 연속적인 데이터 간의 연관성을 파악할 수 있는 슬라이딩 윈도우 기법을 통해 효율적으로 데이터를 분석하고 처리할 수 있다. 하지만, 기존의 방법들은 학습 성능(학습 시간 및 모델 성능)에 미치는 영향을 평가하는 기준 없이 입력 데이터의 윈도우 사이즈를 임의로 설정하여 데이터를 처리해 왔다. 따라서, 본 논문은 학습 시간과 모델 성능을 기준으로 센서 데이터의 윈도우 사이즈 최적화 기법을 제안한다. 제안한 방법은 전류를 이용하여 스위치와 다이오드 온도를 추정하는 가상 센서(virtual sensor) 실험 테스트베드에 적용하여, 학습 시간 중심으로는 5%의 윈도우 사이즈를, 모델 성능 중심으로는 R2 SCORE 의 값을 0.9295 로 갖는 8%의 윈도우 사이즈가 최적으로 도출되었다.

1. 서론

센서 데이터 시대가 도래하면서 방대한 양의 센서 데이터 처리 방법 연구의 수요가 지속적으로 증가하고 있다[1]. 센서 데이터는 로봇, 설비 기기 등에 탑재된 온도 센서, 자이로 센서 등 다양한 센서를 통해 얻을 수 있으며, 전류, 온도, 주파수 등 주기성을 가진 데이터를 얻을 수 있다. 주기성을 갖는 데이터가 학습의 입력 데이터로 사용되는 경우 주기적인 특성을 통해 예측한 출력 데이터는 정확도가 높고 데이터의 흐름을 효율적으로 파악할 수 있다는 점에서 주기성을 갖는 입력 데이터의 출력 예측 연구는 필수적이다[2].

주기성을 갖는 데이터의 예측을 위한 방법으로는 슬라이딩 윈도우 기법이 있다. 슬라이딩 윈도우 기법은 사용자가 설정한 윈도우 사이즈 크기만큼의 연속적인 데이터를 처리한다. 이 기법은 많은 양의 연속된 데이터를 효율적으로 처리할 수 있으며, 데이터 간의 연관성이 매우 큰 주기성 입력 데이터가 정확하게 출력 데이터를 예측할 수 있도록 효율적인 전처리가 가능하다. 또한, 데이터에 대한 중요성이 점차 커지는 사회에서 제한된 메모리양 내에서 효율적인 데이터 처리는 필수적이다[3-4].

[5]의 연구는 슬라이딩 윈도우 기법을 통해 입출력 데이터 간의 연관성이 큰 지점을 찾음으로써 윈도우 사이즈를 채택한다. 이때, 출력 데이터 예측 모델에 대해 R-square 는 0.9706 이라는 매우 높은 성능으로 보아 슬라이딩 윈도우 기법을 통한 주기성 입력 데이터 전처리는 매우 효율적이다. 하지만, R-square 와 연산량을 비교하여 임의로 최적의 윈도우 사이즈를 채택했다는 점에서 한계점이 있다. 또한, [4]의 연구는 윈도우 사이즈에 따른 성능 평가를 한 가지의 성능 평가 메트릭을 통해 보

1) 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1A2C1007400)

여주고 임의로 윈도우 사이즈를 지정하여 실험하고, 채택했다는 점에서 채택된 윈도우 사이즈가 최적이라는 것을 증명하기에는 한계가 있다. 본 논문은 입력 데이터로 전류, 출력 데이터로 온도를 사용하여 앞선 슬라이딩 윈도우 기법의 이점을 근거로 윈도우 사이즈를 변경하며, 슬라이딩 윈도우 기법을 통해 입력 데이터를 전처리하고 출력을 예측한다. 이후, R2 SCORE, MSE, MAE 세 가지의 평가 메트릭을 통해 윈도우 사이즈에 따른 모델의 성능을 평가한다. 성능 데이터는 학습 시간이 강조되는 경우와 성능이 강조되는 경우, 2 가지의 경우로 나누어 각 경우에 최적화된 성능지수 평가식을 제안한다. 제안된 식은 학습 시간, 연산량, 학습 성능을 반영한 식으로 경우에 따라 학습 시간과 FLOPs, R2 SCORE 의 상관 비율을 조절하여 효과적으로 최적의 윈도우 사이즈를 채택할 수 있도록 한다.

2. 센서 데이터 기반 회귀 모델 개발을 위한 윈도우 사이즈 최적화 방법

2.1 슬라이딩 윈도우를 통한 데이터 전처리

2.1 절에서는 슬라이딩 윈도우를 통한 데이터 분석에 대해 설명한다. 본 논문의 입력 데이터는 1s 의 주기를 가지며, 주기의 일정 비율(%)로 윈도우 사이즈를 채택한다. 주기에 의해 채택한 윈도우 사이즈를 통해 데이터를 전처리하여 전처리된 데이터셋을 생성하는데, 방법은 [그림 1]과 같다.

[그림 1] 슬라이딩 윈도우 기반 데이터 전처리 방법

[그림 1]을 살펴보면, 채택한 윈도우 사이즈를 통해 전처리된 데이터셋에 포함될 입력 데이터를 결정한다. 이후, 선택된 입력 데이터와 연관성을 갖는 출력 데이터를 전처리된 데이터셋에 포함하여 모델 학습에 필요한 데이터셋을 생성한다. 본 논문은 윈도우 사이즈에 따른 효과적인 성능 분석을 위해 주기의 1%부터 100%까지 총 100 가지의 윈도우 사이즈를 통해 모델의 성능을 분석하였다.

2.2 윈도우 사이즈 최적화 방법

2.2 절에서는 최적의 윈도우 사이즈 채택을 위한 성능지수 평가식에 대해 설명한다. 본 논문은 학습 시간이 중요한 경우와 성능이 중요한 경우 2 가지로 나누어 성능지수 평가식을 제안하며 이 식을 통해 최적의 윈도우 사이즈를 채택한다. 성능지수 평가식은 식의 형태는 같지만 식의 파라미터 β, γ 의 범위를 달리하여 2 가지 경우를 구분한다. 성능지수 평가식은 수식 (1)과 같으며, 각 경우에 대한 파라미터 범위는 <표 1>과 같다. 파라미터 범위 기준값은 실험을 통해 최적의 윈도우 사이즈를 채택할 수 있는 값으로 선정하였다.

$$\text{성능지수} = \left(\text{학습 시간} + (\text{FLOPs} * \alpha)\right) * \beta + \frac{1}{\text{R2 SCORE}} * \gamma \quad (1)$$

<표 1> 학습 목표에 따른 평가 식 파라미터 범위

학습 시간이 중요한 경우	성능이 중요한 경우
$\alpha \leq 0.01$	
$\beta \geq 0.001$	$\beta \leq 0.001$
$\gamma \leq 10$	$\gamma \geq 100$

성능지수 평가식은 학습 시간, FLOPs, R2 SCORE 를 통해 성능지수를 계산하며, 결과 값이 작을수록 더 좋은 성능을 의미한다. 학습 시간은 FLOPs 에 의해 크게 영향을 받아 정해지는 값이다[6]. 이를 보완하기 위해 성능지수 평가식의 학습 시간과 FLOPs 를 더하는 과정에서 FLOPs 와 α 를 곱하여 FLOPs 의 영향을 감소시켜 식을 도출하였다. 학습 시간이 중요한 경우 β 의 값을 0.001 이상으로, γ 의 값은 10 이하로 하여 성능지수를 계산하며, 성능이 중요한 경우에는 β 의 값을 0.001 이하로, γ 의 값은 100 이상으로 하여 성능지수를 계산한다. 가장 좋은 성능지수를 도출한 윈도우 사이즈를 최적의 윈도우 사이즈로 채택한다.

3. 실험 및 평가

3.1. 실험 데이터 및 성능 평가 지표

본 논문은 입력 데이터로 전류를 출력 데이터로 스위치의 온도와 다이오드의 온도를 사용한다. AC 전압은 15V, 20V, 25V, 30V 를 사용하였으며, DC 전압은 각각의 AC 전압에 대해 5V, 6V, 7V, 8V, 9V, 10V 를 사용하여 데이터를 수집하였다. 각 전압 조합당 20 개의 데이터셋을 보유하며, 전류 변화를 통해 각 데이터셋당 20,000 개의 데이터를 수집하였다. 각 데이터셋의 데이터 수집 주기의 단위 시간은 0.001s 이며, 해당 단위 시간을 통해 수집한 데이터의 주기는 1s 로 각 데이터셋에 대해 1,000 개의 데이터가 하나의 데이터 주기를 의미한다. 이와 같은 방법으로 수집한 데이터의 수는 총 9,600,000 개로 실험을 위해 사용하였다.

본 논문은 R2 SCORE, MAE, MSE 총 세 가지의 학습 성능 평가 메트릭을 통해 학습 모델의 성능을 평가한다. 세 가지의 학습 성능 평가 메트릭은 각각 수식 (2)~(4)를 통해 정의된다.

$$R2\ \ \mathrm{SCORE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (2)$$

$$MAE = \frac{\sum |y - \hat{y}|}{n} \quad (3)$$

$$MSE = \frac{\sum (y_i - \hat{y})^2}{n} \quad (4)$$

수식 (2)는 R2 SCORE 의 정의식으로 y_i은 실제 값, $\bar{y}$은 실제 값의 평균, $\hat{y_i}$은 예측값을 의미한다. R2 SCORE 는 분산 기반 성능 평가 메트릭으로 값이 1 에 근접할수록 데이터와 학습된 회귀 모델의 연관성이 높아 성능이 좋은 모델로 평가한다. 수식 (3)은 MAE 의 정의식으로 y 는 실제 값, $\hat{y}$은 예측값을 의미한다. MAE 는 실제 값과 예측값의 차의 절댓값의 평균으로 값이 0 에 근접할수록 정확도가 높은 모델로 평가한다. 수식 (4)는 MSE 의 정의식으로 y 는 실제 값, $\hat{y}$은 예측값을 의미한다. MSE 는 실제 값과 예측값의 차의 제곱의 평균으로 값이 0 에 근접할수록 정확도가 높은 모델로 평가한다.

3..2 모델 학습

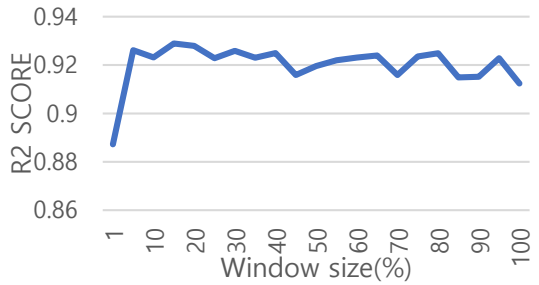
본 논문은 Artificial Neural Network(ANN) 구조를 통해 모델을 학습하였으며 실험에서 사용한 학습 모델의 구조는 <표 2>과 같다.

<표 2> 학습 모델 구조

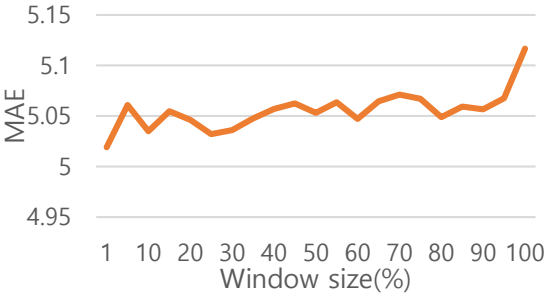
layer	unit(s)
Input layer	64
Two Hidden layers	64
Output layer	1

<표 2>의 설명과 같이 64 개의 뉴런을 갖는 입력층과 각각 64 개의 뉴런을 갖는 은닉층, 1 개의 뉴런을 갖는 출력층으로 ANN 을 구성하였다. 입력층의 입력 데이터는 2.1 절에서 언급한 전처리 방식을 통해 생성된 입력 데이터이며, 출력층은 하나의 출력 데이터를 출력한다. 모델의 학습 파라미터는 Epoch 20, Batch size 64 로 설정하였으며, 윈도우 사이즈를 1%씩 증가시키며 모델을 생성하고 학습하였다. 윈도우 사이즈에 따른 학습 모델의 성능 평가 메트릭 그래프는 아래 [그림 2]~[그림 4]와 같다.

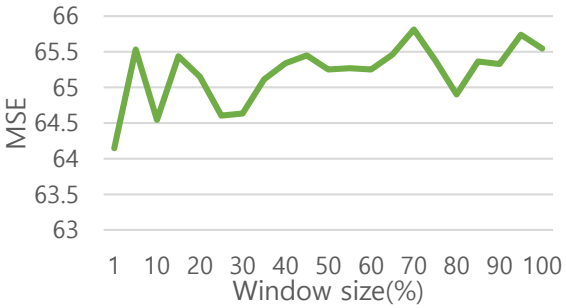
[그림 2]는 윈도우 사이즈에 따른 R2 SCORE 의 성능 그래프이다. 학습 결과 1%의 윈도우 사이즈를 시점으로 8%의 윈도우 사이즈 구간까지는 미세한 변동이 있지만 R2 SCORE 가 증가하는 양상을 보였으며, 8%의 윈도우 사이즈에서 0.9295 의 R2 SCORE 값으로 가장 좋은 성능을 보였다. 8% 이후의 윈도우 사이즈는 윈도우 사이즈를 증가할수록 R2 SCORE 가 감소하는 양상을 보였다.



[그림 2] R2 SCORE 성능 그래프



[그림 3] MAE 성능 그래프



[그림 4] MSE 성능 그래프

MAE 와 MSE 는 0 에 근접한 값일수록 예측값이 실제 값에 근접한 값이므로, 높은 정확도를 갖는 모델이다. [그림 3]은 윈도우 사이즈에 따른 MAE 의 성능 그래프이다. 학습 결과 윈도우 사이즈가 증가할수록 MAE 가 증가하는 양상을 보인다. 하지만, 가장 작은 MAE 의 값이 4% 윈도우 사이즈의 5.0206, 가장 큰 MAE 의 값이 100% 윈도우 사이즈의 5.1166 으로 MAE 의 값이 가장 작은 경우와 큰 경우의 값의 차이가 0.1 로 매우 작아 윈도우 사이즈를 변경하여도 MAE 값의 변동이 매우 미미한 양상을 보였다. [그림 4]에 의해 설명된 MSE 또한 가장 작은 MSE 의 값이 4% 윈도우 사이즈의 64.4457, 가장 큰 MSE 의 값이 70% 윈도우 사이즈의 65.8113 으로 MSE 의 값이 가장 작은 경우와 큰 경우의 값 차이가 1.66 으로 매우 작아 윈도우 사이즈를 변경하더라도 MSE 값의 변동이 매우 미미한 양상을 보였다.

3.2 평가

본 논문에서 제안한 성능지수 평가식을 통해 최적의 윈도우 사이즈를 채택한다. 두 가지 경우에 대한 성능지수 평가를 위해 설정한 파라미터는 <표 3>과 같다. 학습 시간이 우선인 경우의 성능지수는

<표 4>의 (a)와 같으며, 성능이 우선인 경우의 성능지수는 <표 4>의 (b)와 같다.

<표 3> 성능지수 파라미터

학습 시간 우선	성능 우선
$\alpha = 0.01$	
$\beta = 0.01$	$\beta = 0.001$
$\gamma = 10$	$\gamma = 100$

<표 4> 경우에 따른 성능지수 평가

(a) 학습 시간이 우선인 경우

Window size	성능지수	R2 SCORE	학습 시간 (s)
2%	41688.5595	0.9183	4006
3%	41291.3907	0.9190	4073.3871
4%	41636.5073	0.9232	4027.2703
5%	**40425.44729**	**0.9261**	**3927.8768**
6%	43044.3178	0.9199	4183.3631
7%	44324.47137	0.9273	4304.9793
8%	44007.2655	0.9295	4266.8590

(b) 성능이 우선인 경우

Window size	성능지수	R2 SCORE
2%	113.0715	0.9183
3%	112.9873	0.9190
4%	112.4838	0.9232
5%	112.0175	0.9261
6%	113.0095	0.9199
7%	112.2711	0.9273
8%	**111.9909**	**0.9295**
9%	112.5099	0.9251
10%	112.5096	0.9231

성능지수 평가식을 통해 성능지수를 계산한 결과 5%의 윈도우 사이즈에서 가장 낮은 성능지수를 보였으며, 이를 통해 학습 시간이 우선인 경우에는 5% 윈도우 사이즈를 가장 최적의 윈도우 사이즈로 평가한다. 성능지수 평가식을 통해 성능지수를 계산한 결과 8%의 윈도우 사이즈에서 가장 낮은 성능지수를 보였으며, 이를 통해 성능이 우선인 경우에는 8% 윈도우 사이즈를 가장 최적의 윈도우 사이즈로 평가한다.

4. 결론

본 논문은 임의로 윈도우 사이즈를 채택하여 학습하는 한계점을 보완하기 위하여, 학습 시간과 성능에 따른 성능지수 평가식을 제안함으로써, 최적의 윈도우 사이즈를 채택할 수 있는 기법을 제안한다. 실험 결과, 학습 시간이 우선인 경우에는 3927.8768s 의 학습 시간을 갖는 5%의 윈도우 사이즈가 최적으로 도출되었으며, 성능이 우선인 경우에는 R2 SCORE 의 값을 0.9295 로 갖는 8%의 윈도우 사이즈가 최적으로 도출되었다. 본 논문의 제안 기법을 통해 경우에 따른 최적의 윈도우 사이즈를 채택함으로써 센서 데이터 기반의 효율적인 데이터 전처리가 가능하며, 최고 성능의 딥러닝 회귀 모델을 개발할 수 있다.

참고문헌

[1] 김삼근, 오택일, IoT 스트리밍 센서 데이터 기반 실시간 PM10 농도 예측 LSTM 모델, 한국산학기술학회논문지, 19, 11, 310-318, 2018

[2] Yisha Lin, Zongxiang Lu, Ying Qiao, Mingjie Li, Zhifent Liang, Medium and long-term wind energy forecasting method considering multi-scale periodic pattern, 2020 10th International Conference on Power, Energy and Electrical Engineering (CPEEE 2020), E3S Web of Conferences, 2020, 1~5

[3] Mahmood Deypir, Mohammad Hadi Sadreddini, Sattar Hashemi, Towards a variable size sliding window model for frequent itemset mining over data streams, Computers & Industrial Engineering, 63, 1, 161-172, 2012

[4] 김태언, 김성환, 안용덕, IoT 환경에서의 슬라이딩 윈도우 기반 센서 데이터 처리, Journal of Digital Contents Society(JDCS), 32, 4, 825-832, 2020

[5] 김혜진, 박예슬, 이정원, 주기성을 갖는 센서 데이터의 학습 안정성을 위한 하이퍼 파라미터 회귀-앙상블 학습 방법, Korea Comouter Congress 2021(KCC2021), 한국정보과학회, 2021, 2128-2130

[6] Yanliang He, Junmin Liu, Peipei Wang, Wenjie Xiong, Yuexiang Wu, Xinxing Zhou, Yun Cheng, Yanxia Gao, Ying Li, Shuqing Chen, Dianyuan Fan, Detecting Orbital Angular Momentum Modes of Vortex Beams Using Feed-Forward Neural Network, IEEE, 37, 23, 5848-5855, 2019