저가형 초미세먼지 센서 정확도 향상을 위한 인공지능 기반 센서 데이터 예측 기법

서 경 덕*

Artificial Intelligence-Based Sensor Data Prediction Technique for Improving the Accuracy of Low-cost Ultra-fine Dust Sensor

Suh KyungDuk*

요 약

컴퓨팅 속도가 발전함에 따라 빠른 컴퓨팅 속도를 이용한 머신러닝과 딥러닝을 이용한 센서 데이터의 분석 사례가 늘고 있다. 이 논문에서는 여러 인공지능 분석기법을 이용해 저가형 센서 데이터가(PM2.5) 고가형 센서와 근접한 성능을 낼 수 있는 방법에 대해 제안하며, 각각의 분석기법을 통한 예측값을 비교해 봄으로로써 인공지능 기법의 특성에 관해 기술하였다.

Abstract

As computing speed develops, there are more and more cases of machine learning using fast computing speed and analysis of sensor data using deep learning. This paper suggests how low-cost sensor data(PM2.5) can perform close to high-end sensors by using various artificial intelligence analysis methods. This paper also describes the characteristics of artificial intelligence techniques by comparing the predicted values through each analysis method.

Key words
Sensor Data Etimation, Machine Learning, Recurrent Neural Network,

Ⅰ. 서 론

최근 반도체 장비의 발전과 5G 통신의 발전으로 IoT 제품이 많이 개발되고 있으며, 그와 더불어서 많은 센서가 모바일 기기에 부착되어 사용되고 있다. 예를 들면 로봇청소기에는 자이로센서, 카메라 센서, LDS와 같은 센서가 들어가며, 공기청정기에는 공기질 측정 센서가 장착되어 있다. 특히 부가적

* 광운대학교 로봇학부 소속

※ "본 연구는 조석헌 박사님의 도움을 받아 Qualcomm Institute에서 수행되었음."
※ "본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 "(2017-0-00096)
※ "This research was supported by the MIST(Ministry of Science and ICT), under the National Program for Excellence in SW(2017-0-00096), supervised by the IITP(Institute for Information & communications Technology Promotion)"

인 기능이 많아질 수 록 다양한 종류의 센서가 사용되게 된다. 이렇게 센서는 주변 환경을 파악하고, 사용자를 인식해 사용자 맞춤형 서비스를 제공할 수 있게 하는 등 여러 기능을 수행하는 데 도움이 되며, IoT 제품에는 필수적인 요소라고도 할 수 있다. 센서 의존도가 높은 IoT 제품의 경우 센서의 품질에 따라 가격이 상이하며, 제품에 고가형 센서를 사용했는지, 저가형 센서를 사용했는지에 따라 성능의 차이가 나게 된다.

본 논문에서는 고가형 센서의 데이터를 학습하여, 저가형 센서의 데이터를 보정해주는 인공지능 기법에 대해 소개하며, 각각의 기법은 R^2와 Root Mean Squared Error(이하 RMSE)를 이용해 성능 평가하였다. 사용한 기계학습 방법으로는 선형 회귀 기법, 다항식 회귀 기법, 랜덤 포레스트 기법, Gradient Boost 기법과 딥러닝인 Long Short-Term Memory(LSTM) 방법을 이용해 보았다. 데이터의 80%는 학습에 사용하였으며, 20%는 평가에 사용하였다. 사용된 데이터는 상대적 고가의 센서 데이터인 federal equivalent method(FEM) Approved TEOM 1405-DF 장비의 Data와 상대적 저가의 센서 데이터인 California Air Resources Board(CARB)의 데이터이며 측정 지역은 미국에 위치한 BakersField이다.

II. 선형(다항식) 회귀 기법

선형 회귀 기법은 데이터를 선형화시키기 위해 종속 변수 y와 한 개 이상의 독립변수 x와의 상관관계를 모델링하는 기법이다. 여기서 다항식 선형 회귀 기법은 둘 이상의 설명 변수를 이용해 상관관계를 모델링 하는 경우를 말한다. 선형 회귀 모델은 손실 함수(loss function)를 최소화하는 방식으로 세웠다. 단순 선형 회귀 모델을 식 (1)이라고 정의할 수 있고,

$$h_\beta = \beta_0 + \beta_1 x \tag{1}$$

단순 선형 회귀 모델의 손실함수 $J(\beta_0, \beta_1)$는 (x, y) 데이터 m개가 있을 때 식 (2)로 정의할 수 있다.

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\beta(x^i) - y^i)^2 \tag{2}$$

마찬가지로 (x, y) 데이터 m개와 n개의 독립변수 X가 있을 때, 다항식 회귀 모델을 식 (3)라고 정의할 수 있으며,

$$h_\beta = \beta_0 + \beta_1 x + \ldots + \beta_n x_n \tag{3}$$

손실함수 $J(\beta)$는 식 (4)라고 정의할 수 있다.

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\beta(x_1^i, x_2^i, \ldots x^{i_n}) - y^i)^2 \tag{4}$$

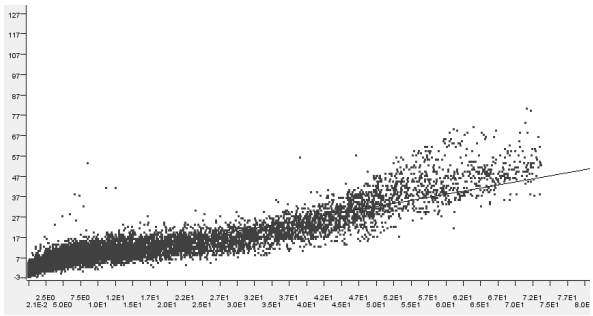

그림 1. 선형 회귀
Fig. 1. Linear regression

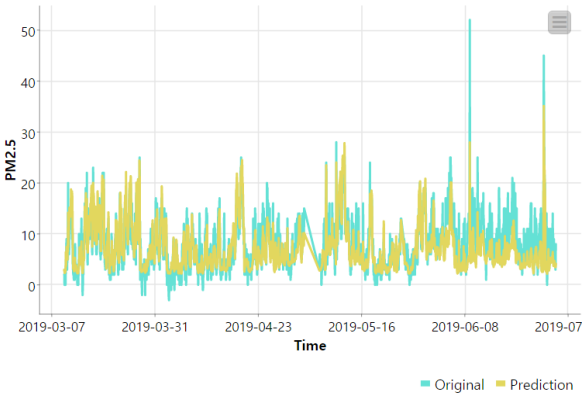

그림 2. 선형 회귀 예측값과 실제값의 비교
Fig. 2. Comparison of Prediction value and Actual Value after linear regression

데이터의 X축 값은 저가형 센서의 데이터로, Y축 값은 고가형 센서 데이터로 설정해 측정해본 결과, $y = 0.603x + 2.154$의 선형 회귀 모델이 나왔다.

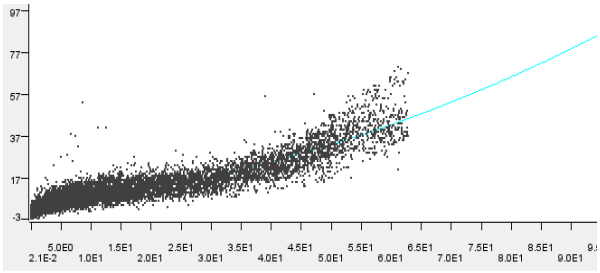

그림 3. 2차 다항식 회귀
Fig. 3. Quadratic polynomial regression

반면에 2차 다항식 회귀 기법을 이용한 경우는 다음과 같았다.

$y = 0.0065x^2 + 0.2393x + 4.6625$의 2차 다항식 회귀 모델이 나왔고, 두 모델을 남은 20%의 평가 데이터로 평가해보았다. 1차 선형 회귀 기법을 썼을 경우 R^2=0.613, RMSE = 2.982였고, 2차 다항식 회귀 기법을 사용했을 경우 R^2=0.593, RMSE = 3.059였음을 확인했다.

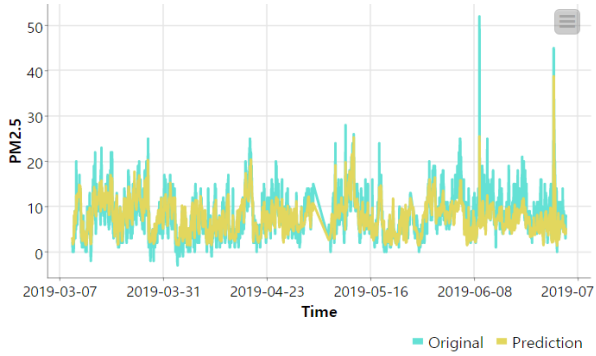

그림 4. 2차 다항식 회귀 예측값과 실제값의 비교
Fig. 4. Comparison of Prediction value and Actual Value after quadratic polynomial regression

III. 앙상블 학습 - 랜덤 포레스트 기법과 Gradient Boosting 기법

랜덤 포레스트 기법은 회귀 분석 등에 쓰이는 앙상블 학습 방법의 일종으로 다수의 결정 트리로부터 분류 또는 평균 예측치를 출력함으로써 동작하는 기계학습 종류이다.

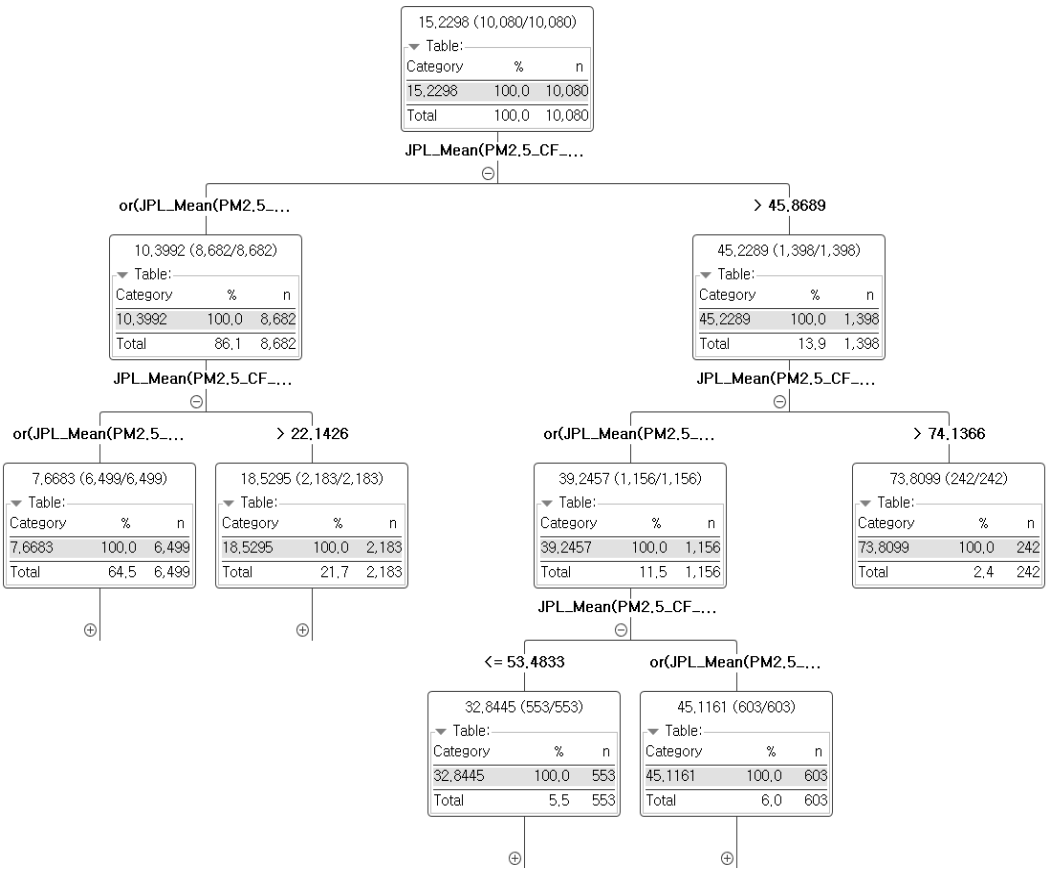

그림 5. 랜덤 포레스트 결정트리 (일부)
Fig. 5. Random Forest Decision Tree (Partial)

본 논문에서는 주어진 훈련 데이터에서 중복을 허용하여 원 데이터셋과 같은 크기의 데이터 셋을 만드는 Bootstrap 과정을 통해 각각의 훈련 데이터를 훈련하며, 후에 결합시키는 방법을 사용했다. 구체적인 훈련 방법으로는

1. Bootstrap aggregating (Bagging) 방법을 통해 100개의 결정 트리를 생성한다.
2. 생성된 100개의 결정트리를 훈련한다.
3. 투표방식을 사용해 결정 트리들을 하나의 분류기(랜덤 포레스트)로 결합한다.

결정트리의 투표 결과값으로 도출된 예측 모델을 이용한 평가치는 R^2 = 0.457, RMSE = 3.482로 출력되었다.

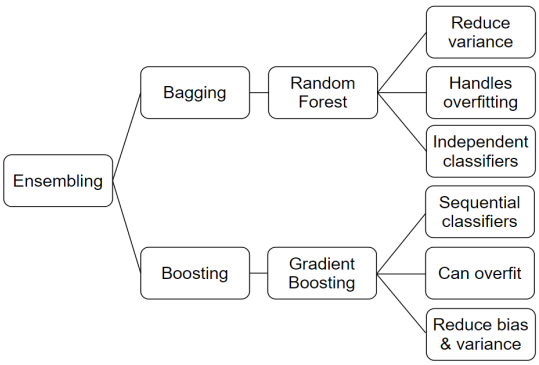

그림 6. 랜덤 포레스트와 Gradient Boosting의 특징
Fig. 6. Features of Random Forest and Gradient Boosting

랜덤 포레스트 기법과는 달리 Gradient Boosting 기법은 Boosting 계열에 속하는 앙상블 방법론 중 하나이며, Random Forest 기법에 비해 과적합이 일어날 수도 있다는 단점이 있지만, 현재 주어진 Data는 시계열 데이터로 연속적으로 이루어져 있고, Tabular format 데이터의 형태이므로 예측 성능이 랜덤 포레스트 기법보다 더 우수한 성능을 보여준다.

Gradient Boosting 기법을 이용한 결과, 학습된 예측모델을 이용한 평가치는 R^2 = 0.654, RMSE = 2.794 이었다. 랜덤 포레스트 기법에 비해 상당한 성능 향상이 있었음을 알 수 있다.

Ⅳ. 딥러닝 - LSTM

LSTM은 Long short term memory network의 줄임말로 순환신경망(이하 RNN)의 일종이다. RNN은 계층의 출력이 순환하는 인공신경망이며, 앞뒤 신호의 상관도가 있는 경우 많이 쓰인다. RNN은 출력된 신호가 계속 순환하면 활성화 함수를 반복적으로 거치게 되어서 경사값을 구하기 힘들다는 문제가 있기 때문에 학습이 제대로 이뤄지지 않을 수 있다. 위의 문제를 해결하기 위해 LSTM을 이용한 딥러닝 기법을 이용하였다. LSTM은 입력 조절 벡터, 망각 벡터, 출력 조절 벡터를 이용해 입력과 출력 신호를 게이팅 한다[5].

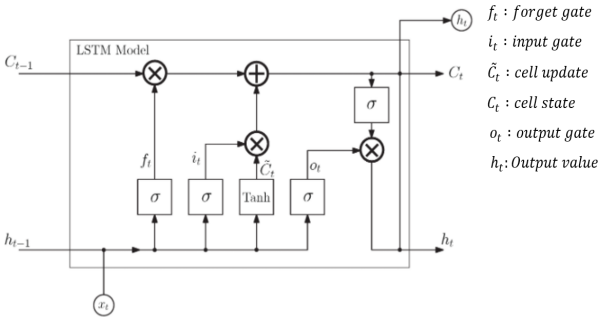

그림 7. LSTM 모델 구조
Fig. 7. LSTM model structure

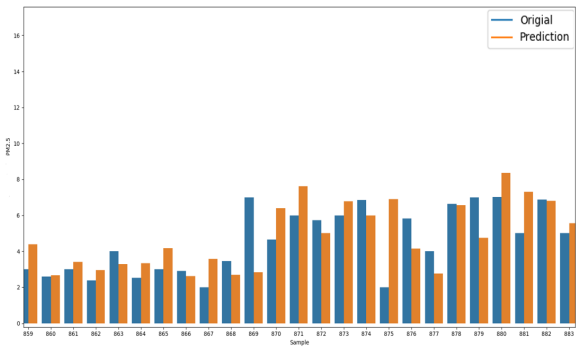

그림 8. 예측값과 실제값 비교를 위한 막대그래프 (일부)
Fig. 8. Bar graph for predicted and actual values (partial)

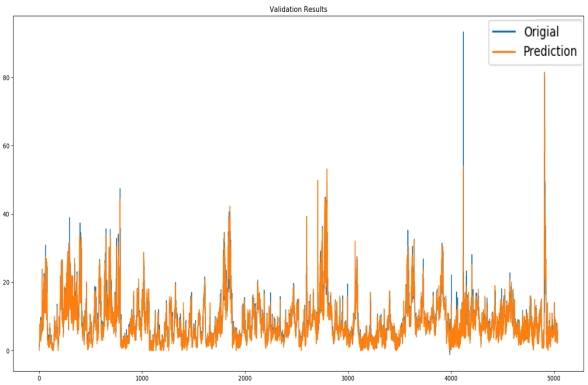

그림 9. LSTM 학습 후 예측값과 실제값 비교
Fig. 9. Comparison of Prediction value and Actual Value

망각 게이트 레이어는 시그모이드 계층에 의해 결정되며 h_{t-1}과 x_t를 입력으로 받아 C_{t-1}에 0과 1사이의 값을 보내주는 역할을 한다. 수식으로 표현하면 아래의 식 (5) 로 정의할 수 있다.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

입력 게이트 레이어는 시그모이드 계층에 의해 업데이트할 정보를 정하게 되며, tanh 계층에 의해 만들어진 $\widetilde{C_t}$ 벡터가 더해지게 된다. 수식으로 표현하면 아래의 식 (6), (7)로 정의할 수 있다.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

새로운 cell state인 C_t는 망각의 게이트 레이어의 f_t 와 입력 게이트 레이어의 $\widetilde{C_t}, i_t$ 에 의해 정해지게 된다.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \quad (8)$$

마지막 출력 값 h_t은 cell state와 출력 게이트 값인 o_t의해 결정되며, 수식으로 표현하면 아래의 식 (9),(10) 로 정의할 수 있다.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (10)$$

학습에 쓰인 데이터는 -1과 1사이의 값으로 활성화를 수행하는 LSTM 인공신경망을 충분히 활용하기 위해 평균화를 진행하였다 [4]. 또한 10개의 LSTM 노드를 사용하였으며 이를 처리하는 파라미터를 480개로 설정, 최종 출력 계층의 노드는 하나이며, LSTM과 출력계층 사이에 적용되는 가중치 수는 11개이다. 11개중 10개는 입력값에 대한 가중치로 설정, 1개는 평균값을 조절하는 가중치로 설정했다.

학습 곡선을 그려본 결과, 대략 100회 정도가 지나면 1차 학습이 이뤄졌으며 300회가 지나면 학습이 거의 완료 되었다. 예측값을 평가해본 결과 R^2 = 0.811, RMSE = 2.869 이었음을 확인했다.

V. 결 론

본 논문에서는 선형 회기, 다항식 회귀, 앙상블 학습, 순환신경망을 이용해 저가형 PM2.5 센서 데이터를 보정해 보았다.

표 1. 전체 학습 데이터 평가
Table 1. Overall Scorer Results

평가 기법 / 학습 기법	R^2	RMSE
선형 회귀	0.613	2.982
다항식 회귀	0.593	3.059
랜덤 포레스트	0.457	3.482
Gradient boost	0.654	2.794
순환신경망(LSTM)	0.811	2.869

R^2에서 가장 우수한 성능을 보여준 학습기법은 LSTM이었고, RMSE에서 가장 우수한 성능을 보여준 학습기법은 Gradient boost 기법이었다. 전반적으로 LSTM 성능이 우수했으며, PM2.5의 시계열 데이터 예측에 가장 우수했음을 보였다.

참 고 문 헌

[1] 양영권. "인공지능 창호 환기시스템의 성능평가에 따른 미세먼지 농도 예측 모델 개발." 국내박사학위논문 中央大學校 大學院, 2019. 서울

[2] 송정현. "기온 예측 정확도 향상을 위한 순환신경망 기반 기온 예측 기법." 국내석사학위논문 경희대학교 대학원, 2018. 경기도

[3] 김경도. "딥러닝을 이용한 시계열 관측 데이터 예측 및 보정." 국내석사학위논문 광운대학교 대학원, 2018. 서울

[4] 김성진 (2018).「코딩셰프의 3분 딥러닝 케라스 맛」,한빛미디어.pp.166-195

[5] Yan, K., Wang, X., Du, Y., Jin, N., Huang, H. and Zhou, H. Multi-Step Short-Term Power Consumption Forecasting with a Hybrid Deep Learning Strategy. Energies, 11(11), p.3089, 2018.