Journal of Convergence for Information Technology
Vol. 11. No. 11, pp. 57-65, 2021
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2021.11.11.057

심층신경망 모델을 이용한 대기오염망 자료확정 알고리즘 연구

이선우[1], 양호준[1], 이문형[2], 최정무[2], 윤세환[2], 권장우[3*], 박지훈[4], 정동희[4], 신혜정[4]
[1]인하대학교 전기컴퓨터공학과 학생, [2]인하대학교 컴퓨터공학과 학생,
[3]인하대학교 컴퓨터공학과 교수, [4]국립환경과학원 대기환경연구과 환경연구원

A Study on the Air Pollution Monitoring Network Algorithm Using Deep Learning

Seon-Woo Lee[1], Ho-Jun Yang[1], Mun-Hyung Lee[2], Jung-Moo Choi[2], Se-Hwan Yun[2], Jang-Woo Kwon[3*], Ji-Hoon Park[4], Dong-Hee Jung[4], Hye-Jung Shin[4]
[1]Student, Electric Computer Engineering, Inha University
[2]Student, Computer Engineering, Inha University
[3]Professor, Computer Engineering, Inha University
[4]Researcher, Air Quality Research Department, Air Quality Research Division

요 약 본 논문은 딥 러닝(Deep Learning)을 이용하여 대기오염측정망 데이터 중 특정 증상이 나타나는 이상 데이터를 탐지하는 방법을 제시한다. 기존 방법들은 일반적으로 시계열 데이터 내에서 기존과는 다른 특이한 패턴이 나타나는 데이터를 탐지하여 이상치로 분류하며, 이는 특정 증상만을 탐지하기에는 적합하지 않다. 본 논문에서는 주로 이미지의 전경 분리(Sementic Segmentation)에 사용되는 DeepLab V3+ 모델의 2차원 합성곱 신경망 구조를 1차원 구조로 변형하여 이미지 대신 여러 센서의 시계열 측정값을 입력받고 특정 증상이 나타나는 데이터를 탐지하도록 하는 방법을 제시한다. 또한, 데이터에 '조각별 집계 근사법(Piecewise Aggregate Approximation)'을 적용하여 잡음이 많은 대기오염측정망 데이터의 복잡도를 줄임으로써 성능을 높인다. 실험 결과를 통해 준수한 성능으로 이상치 탐지를 수행할 수 있음을 확인할 수 있다.

주제어 : 대기질, 딥러닝, 이상탐지, 대기오염측정망, 기계학습

Abstract We propose a novel method to detect abnormal data of specific symptoms using deep learning in air pollution measurement system. Existing methods generally detect abnomal data by classifying data showing unusual patterns different from the existing time series data. However, these approaches have limitations in detecting specific symptoms. In this paper, we use DeepLab V3+ model mainly used for foreground segmentation of images, whose structure has been changed to handle one-dimensional data. Instead of images, the model receives time-series data from multiple sensors and can detect data showing specific symptoms. In addition, we improve model's performance by reducing the complexity of noisy form time series data by using 'piecewise aggregation approximation'. Through the experimental results, it can be confirmed that anomaly data detection can be performed successfully.

Key Words : Air Quality, Deep Learning, Abnormal Detection, Air Pollution Monitoring Network, Machine Learning.

1. 서론

최근 미세먼지, 온난화 및 이상기후에 대한 관심이 높아지면서, 전국의 국가대기오염측정망 숫자도 2010년 290개소에서 2020년 505개소로 지속적으로 늘어

*This work was supported by a grant from the National Institute of Environmental Research (NIER), funded by the Ministry of Environment (MOE) of the Republic of Korea (NIER-RP-2020-04-02-118).
*Corresponding Author : Jang-Woo Kwon(jwkwon@inha.ac.kr)
Received September 3, 2021 Revised October 20, 2021
Accepted November 20, 2021 Published November 28, 2021

나고 있는 추세이다.

국가대기오염측정망의 개별 측정소에서 관측된 데이터는 기록 장치의 결함이나 자연재해 등의 이유로 결측치나 이상치가 포함될 수 있다. 이러한 이상치나 결측치가 많아질 경우, 정보량이 정보의 질을 담보하지 못한다는 점에서 신뢰성을 잃어버리고 통계적 분석의 오류와 정보의 질적 하락 문제를 가져올 수 있다. 이러한 문제는 데이터 분석에 있어서 중요한 문제를 일으킨다. 그 이유는 정보량이 많더라도 분석하고자 하는 데이터의 오류가 많다면, 해당 정보를 이용한 통계적 추론의 신뢰성을 보장할 수 없고, 통계자료가 가지는 표본으로서의 조건을 충족하지 못하기 때문이다.

이러한 이유로 국가대기오염측정망 측정소에서는 데이터의 신뢰성을 확보하고자 매월 마지막 주에는 데이터의 오류를 구분하여 결측치나 이상치를 표시(Labeling)해주는 작업을 수행하고 있다.

그러나 이러한 작업은 시간과 비용이 많이 들 뿐만 아니라 각 측정소에서의 환경적인 요인과 주변 측정소와의 연계성 및 오염물질 간의 상관관계 등 여러 영향 요소를 종합적으로 고려해야 하므로 전문적인 지식을 가지고 있는 인력이 수행하여야 한다. 또한 여러 작업자에 의해 판정 작업이 이루어지기 때문에 수행하는 인원마다 이상치 판정 시 주관적인 판단이 개입될 가능성이 있다. 이러한 요인들로 인해 현 이상치 판정 방식에는 많은 전문인력이 필요하지만 막상 이렇게 판정된 이상치는 일관성이 부족하여 신뢰성이 보장되지 않는 실정이다.

본 논문에서는 기존에 사람이 해오던 대기오염측정망 자료의 이상치 판별 작업을 인공지능 방법론의 하나인 심층신경망 모델을 이용한 이상 탐지 모델을 적용하여 자동으로 탐지하고 이상치로 추정 가능한 영역을 표시해주는 모델을 제안하고자 한다. 제안하고자 하는 이상탐지 방법은 지도학습(Supervised Learning) 기반의 전경 분리(Semantic Segmentation)[1,2] 모델 중 DeepLab V3+[2] 모델을 기반으로 한 이상 탐지 모델이다. 본 논문에서는 DeepLab V3+에서 제안하는 특징도(Feature Map) 부분의 ResNet[3] 합성곱 필터를 2차원 구조에서 1차원[4]) 구조로 변형하였고, '조각별 집계 근사법'[6-8]을 적용하여 잡음이 많은 형태의 시계열 데이터를 더욱 단순하게 보기 위하여 입력 데이터의 복잡도를 줄이는 방법론을 제안하였다. 제안하는 모델의 학습 및 테스트 데이터는 국가대기오염측정망 중 46개의 측정소에서 측정된 대기오염물질 8개 항목의 측정 자료이며, 총 두 기간의 데이터(2016~2018년, 2018~2020년)를 기반으로 학습 및 테스트하여 결과를 도출하였다.

본 논문의 구성은 2장에서 시계열 데이터의 이상치 탐지 및 처리를 위해 사용되는 기존 알고리즘 및 기법과 한계점을 분석하고, 3장에서 이상치 판별을 위한 신경망 모델과 학습 기법을 제안한다. 4장에서는 제안된 모델의 학습 결과를 분석하고 5장에선 결론 및 추후 연구 방향에 대하여 논의한다.

2. 관련 연구

대기질 데이터와 같은 시계열 데이터에는 기기 이상, 주변 환경의 급격한 변화 등의 여러 요인으로 인해 데이터 통계의 신뢰성을 해치는 이상치가 발생할 수 있으며, 이러한 이상치를 탐지 및 처리하기 위한 연구들이 진행되어 왔다. Ren, Hansheng, et al. [9]은 시계열 데이터의 이상탐지 방안으로써 SR-CNN(Spectral Residual Convolution Neural network)를 제시하였다. 스펙트럼 잔차(Spectral Residual) 알고리즘을 통해 시계열 데이터를 특징 맵(Saliency Map)으로 변환시키며, 특징 맵에서는 기존 패턴과 다른 특이한 패턴의 데이터가 주변값에 비해 높은 값을 띄게 되므로 데이터의 이상 여부를 보다 쉽게 판별이 가능하다. 최종적으로 이러한 특징 맵을 다수의 1D 합성곱 필터로 구성된 신경망에 입력하여 이상값과 정상값의 특징을 학습할 수 있는 구조를 제안하였다. Zhou, Bin, et al. [10]은 시계열 데이터에서 레이블링의 어려움을 극복하고 데이터의 비정상적 패턴을 시각화하기 위하여 정상 데이터로만 이루어진 학습 데이터를 통해 비지도학습 모델인 BeatGAN 모델을 학습시킨 뒤 비정상 데이터를 입력하여 정상 데이터와 다른 패턴을 보이는 데이터를 찾아내는 방식을 제안하였다.

위 방법들은 시계열 데이터에서 기존에 나타나지 않던 특이한 패턴이나 값을 가지는 데이터들을 이상치로 판정하는 구조이다. 그러나 본 연구에서는 '베이스라인 이상'의 특정한 증상이 나타나는 데이터의 탐지를 목표로 하기 때문에 이러한 방법들을 적용하기에는 부적절하다. 다음 장부터는 이상치 분류 및 이상 증상 판별을 수행하기 위해 본 연구에서 제안하는 기법을 상세히 설명한다.

3. 실험 방법

3.1 기존 데이터 확정 프로세스

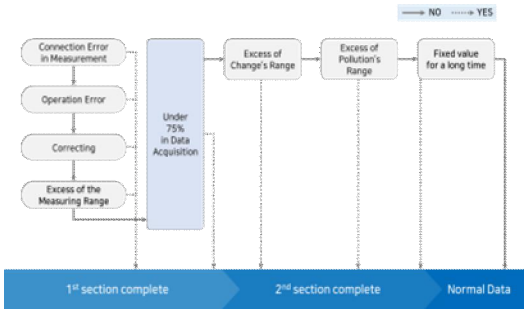

Fig. 1. The process of determining the air pollution measurement data

Fig. 1은 대기오염측정망의 자료 수집 및 확정 과정을 나타낸다. 각지의 측정소에서는 1시간 단위로 데이터를 수집하며, 측정된 데이터에서 기기 이상 및 오작동으로 인한 이상 데이터를 1차적으로 선별한다. 이후 취합된 데이터에 대해 전문 인력들에 의한 2차 선별 및 최종 데이터 확정 과정이 진행되는데, 이 때 '베이스라인 이상' 증상의 판정 또한 이루어진다. 이후 연구 내용에서는 1차 선별이 완료된 데이터를 다룬다.

3.2 데이터 분석 및 학습 데이터 선정

본 논문의 연구는 국립환경과학원에서 제공한 472개소 측정소에서 '16년~19년의 기간 동안 측정된 8개 대기오염물질($SO_2, NO, NO_2, NO_x, O_3, CO, PM_{2.5}, PM_{10}$) 데이터를 기반으로 진행되었다.

대기오염측정망 내에서 '베이스라인 이상'이라고 분류할 수 있는 측정상의 오류를 검출하기 위한 연구를 진행하였다. 베이스라인 이상이란 Fig. 2과 같이 데이터의 평균, 진폭 등을 통해 정해지는 기준선(베이스라인)이 평상시와 다른 패턴을 보이는 현상을 말한다.

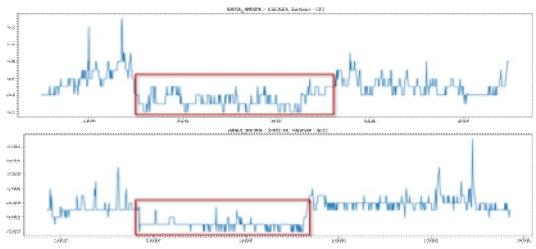

Fig. 2. Example of 'Abnormal Baseline'.

Table 1은 대기오염물질 데이터의 각 측정 대상별 '베이스라인 이상' 증상이 나타나는 비율과 결측치 비율을 나타낸다.

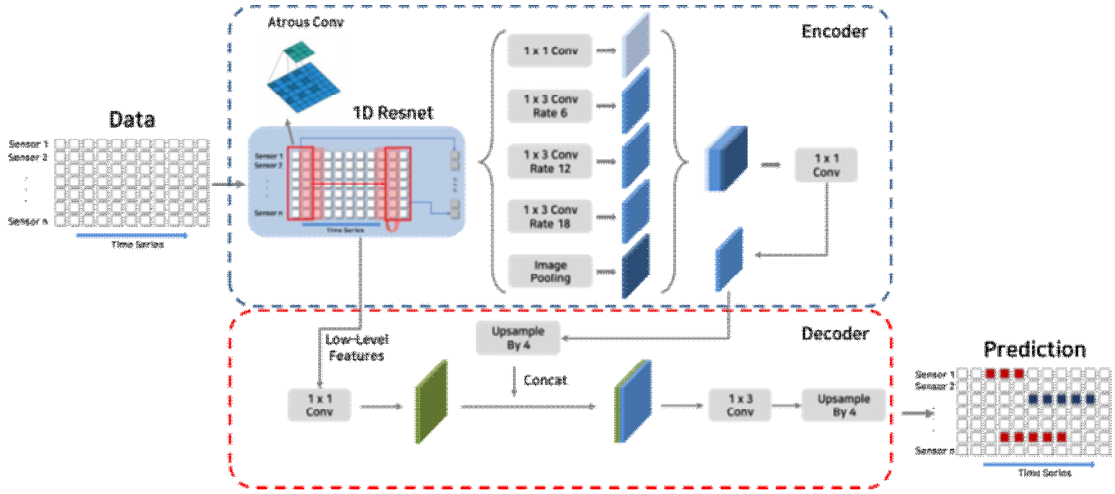

Fig. 3. The proposed model is based on the DeepLab V3+ model. The peculiarity is that the existing DeepLab V3+ model was based on a 2D CNN filter, while the model proposed in this paper was based on a 1D CNN filter.

Table 1. Rate of abnormal and missing data for each elements

Elements	Rate of Abnormal data	Rate of Missing data
SO_2	0.289%	1.35%
CO	0.375%	1.46%
PM_{10}	0.001%	0.31%
$PM_{2.5}$	0.015%	3.17%
NO	0.195%	0.72%
NO_2	0.195%	0.71%
NO_x	0.195%	0.72%
O_3	0.022%	0.53%

NO, NO_2, NO_x의 세 측정 대상은 같은 장비에서 측정되며, 이 중 하나의 측정 대상에서 발생한 이상은 해당 장비의 이상을 의미한다. 따라서 이후 내용에서는 세 측정 대상을 한 개의 측정 대상 물질인 NO_x로 취급한다.

3.3 제안하는 모델

3.3.1 DeepLab V3+

Fig. 3는 제안하는 모델의 구성도로써, 전경분리(Semantic Segmantation) 모델 중 DeepLab V3+ 모델을 기반으로 구성하였다. 전경분리 알고리즘은 이미지 처리 분야에서 주로 사용된다. 합성곱 필터는 본래의 2차원 필터를 1차원 필터로 변경하였고, 특성도(Feature Map)부분의 ResNet34를 1D-CNN[4,5])으로 변경하여 구성하였다. 또한, 각 측정 데이터의 확정은 한 달 단위로 수행된다는 점에 착안하여 학습 데이터 크기를 1,440시간 단위로 설정하여 실시하였다.

3.3.2 조각별 집계근사법

기존의 대기오염측정망 데이터는 진폭이 크고 데이터 변동이 매우 심하며, 이러한 특징으로 인해 인공지능 모델이 데이터의 중요한 특성을 제대로 학습하지 못할 우려가 존재한다.

인공지능 모델이 보다 수월하게 데이터의 특성을 학습하도록 하기 위해, 수식 1과 같이 시계열 데이터 x에 대하여 일정한 시간 간격 M마다 평균값을 산출 및 대푯값으로 근사하는 방식인 조각별 집계근사법(Piecewise Aggregate Approximation)[6-8]을 적용하였고, Fig. 4와 같이 잡음을 줄임으로써 입력 데이터의 복잡도를 줄일 수 있었다.

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j \quad (1)$$

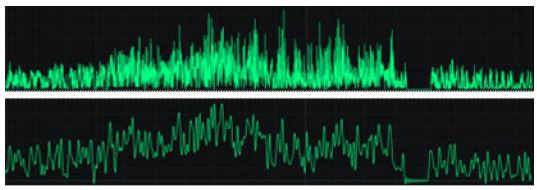

Fig. 4. Before Piecewise Aggregate Approximation (above), after(below).

3.3.3 하이퍼 파라미터

본 논문에서 제안하는 모델의 하이퍼 파라미터는 Table 2와 같이 구성하였다. 에폭은 총 700 에폭으로 구성하였으며 배치 사이즈는 8로 구성하였다. 학습 속도(Learning Rate)는 0.0001로 구성하였으며, 학습 감쇠는 0.0005로 오버슈팅(Over-shooting)이 일어나는 것을 방지하고자 하였다. 최적화 함수(Optimizer)는 SGD로 설정하였다.

Table 2. Hyperparameter of Model

Hyperparameter	Size
Epoch	700
Batch Size	8
Learning Rate	0.0001
Weight Decay	0.0005
Optimizer	SGD

3.4 평가방법

일반적으로 데이터에 대한 이상 진단 성능 평가방식은 전체 데이터 구간 중에서 모델이 정상으로 판정한 구간과 비정상으로 판정한 구간으로 분리 후, 정상 영역과와 비정상 영역에 대해 평가를 진행한다. 기존의 시계열 평가방식은 전체 데이터에서 시간별 이상 판정을 얼마나 맞췄는지 비율을 평가지표로 사용하기 때문에 비정상 영역을 얼마만큼 완벽하게 맞추었는지에 초점이 맞추어져 있으며, 보다 많은 이상 영역을 부분적으로라도 탐지하는 것은 평가에 반영되지 않는다.

예를 들어, Fig 5의 두 모델은 서로 다른 결과를 산출하였지만 기존의 방식대로는 Table 3과 같이 Model 1과 Model 2의 정확도가 같게 나오는 현상을 확인할 수 있다. 시계열 데이터에서는 단순하게 이상을 맞추는 Model 1보다, 부분적으로라도 영역을 잡는 Model 2가 이

상 진단 분야의 시계열 평가에서 더 적합한 모델로 평가되어야 한다. 따라서 모델이 얼마나 많은 '이상 영역에 해당하는 곳을 이상 판정하였는가'를 고려하여 평가할 필요가 있다.

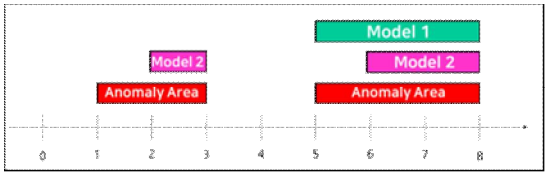

Fig. 5. Example of commercial evaluation

Table 3. Comparison of the commercial evaluation method and the proposed evaluation method

Model	$Precision$	$Precision_r$	$Recall$	$Recall_r$
Model 1	1.0	0.5	0.6	0.5
Model 2	1.0	1.0	0.6	0.79

따라서 이상 영역에서의 이상 판정 여부, 영역 내부에서 이상 판정된 위치, 실제 이상 영역과 판정영역의 범위 차이 등을 고려한 '영역기반 시계열 데이터 평가 방식'[11,12]을 사용하였다.

영역기반 재현율(Range-base Recall)은 모델이 예측한 영역이 실제 이상 판정영역과 얼마나 오버랩(Overlab)되는 지를 나타낸다. 식은 아래와 같다.

$$Recall_T(R_i,P) = \alpha \times ExistenceReward(R_i,P) + (1-\alpha) \times OverlapReward(R_i,P) \quad (2)$$

$$Precision_T(R,P_i) = CardinalityFactor(P_i,R) \times \sum_{j=1}^{N_r} \omega(P_i,P_i \cap R_j,\delta) \quad (3)$$

수식 2와 수식 3은 각각 기존의 Recall, Precision을 산출해내는 방식이다. R은 이상 구간 데이터의 전체 집합을 의미하며, 는 i번째 이상 구간을 의미한다. P는 모델이 예측한 이상 구간 데이터의 전체 집합을 의미하며, 는 i번째로 예측한 이상 구간 데이터를 의미한다.

수식 2에서의 ER(Exsistence Reward) 함수는 예측한 이상구간 중 실제 이상구간이 많이 겹칠수록 그에 상응하는 보상값을 주며 값을 통해 ER 함수와 OR(OverlapReward) 함수의 trade-off를 조절한다. OR 함수는 예측한 이상 영역과 실제 이상 영역이 겹치는 부분에 대해 값을 산출하는 함수와 그 함수값을 조절해주는 CF(Cardinality Factor)함수로 이루어져 있다. 함수의 는 사용자의 선호도에 따라 실제 이상 영역의 앞단을 예측했을 시 더 높은 가중치를 부여할지, 뒷단을 예측했을 시 더 높은 가중치를 부여할지 등에 대한 점수부여 방식을 설정할 수 있는 함수이다. 도출된 함수값은 CF 함수를 통해 보상 정도가 결정된다.

실제 실험을 진행할 때 제안하는 평가방법의 OR(OverlapReward) 함수를 성능지표로 사용하기 위해 값을 0으로 설정하였는데, 이는 전문가에 의한 자료의 2차 검토 시 효율성 측면에서 중요할 것으로 판단되기 때문이다.

4. 연구결과

본 논문에서 제시한 전처리 방식의 검증을 위하여 조각별 집계 근사법 적용 전과 적용 후 데이터를 모델에 적용하였다. 또한, 모델의 성능평가를 위하여, '18년 ~'19년, '16 ~'18년 두 그룹의 테스트 데이터를 모델에 적용하였다.

4.1. 실험환경

본 연구에서는 472개 측정소에서 1시간 단위로 '16 ~'18년의 기간 동안 수집 및 1차 선별이 완료된 데이터에 대한 실험을 수행하였다. 대상 측정소 중에는 해당 기간 동안 신설되거나 폐쇄된 측정소가 포함된다. 총 데이터 개수는 각 8개의 대기오염물질 당 12,819,158개이며, '베이스라인 이상' 증상에 대한 레이블이 존재하는 '18 ~ '19년, 6,386,583개 데이터에 대하여 모델의 학습을 진행하였다. 학습 데이터 중 '베이스라인 이상' 증상이 나타나는 데이터는 총 67,603개로 약 0.17% 비율로 나타났다.

4.2 조각별 집계 근사법 적용 결과

조각별 집계 근사법을 적용하기 전의 결과는 Table 4와 같으며, 적용 후 결과는 Table 5와 같다. 전체적으로 모델의 성능이 향상되었으며, 특히 SO_2정밀도의 경우 0.24의 증가율을 보였다.

분석결과 전체적으로 낮은 수치를 가지는 성분인 SO_2의 경우, 전처리과정에서 값이 많이 증가하는 이상치의 영향을 크게 받았다. 조각별 집계 근사법 방식을

이용하여 데이터의 잡신호를 최소화함으로써 데이터 전처리 전인 Table 4와 비교하여 향상된 결과가 나왔다.

(W-Code: 이상코드, Pre: 정밀도, Rec: 재현률, F1: 조화평균, Acc: 정확도)

Table 4. Before data pre-processing.

Air	W-Code	Pre	Rec	F1	Acc
CO	Normal	0.99	0.98	0.99	0.98
	Abnormal	0.64	0.79	0.70	
O_3	Normal	1.00	0.99	0.99	0.99
	Abnormal	0.76	0.94	0.84	
SO_2	Normal	0.99	0.94	0.97	0.94
	Abnormal	0.51	0.87	0.65	

Table 5. After data pre-processing

Air	W-Code	Pre	Rec	F1	Acc
CO	Normal	0.99	0.99	0.99	0.98
	Abnormal	0.84	0.87	0.85	
O_3	Normal	1.00	1.00	1.00	0.99
	Abnormal	0.80	0.92	0.90	
SO_2	Normal	1.00	0.99	0.99	0.98
	Abnormal	0.75	0.89	0.82	

4.3 이상치 판별 실험 1

학습데이터와 기간이 같은 '18년~'19년도의 데이터에 적용한 이상치 판별 결과는 Table 6 와 같았다. CO의 재현율이 타 성분과 비교하여 0.78로 낮게 나왔다. CO의 경우 차량의 유해 배기가스 같은 요인의 영향을 크게 받는데, 도로변 대기 측정망과 같이 차량의 이동량이 많은 측정소의 경우 위치적 특성상 CO의 값이 다른 측정소에 비하여 평균적으로 높게 관측된다. 하지만 학습 시에는 이러한 측정소의 위치적 특성이 해당 모델의 학습에서는 반영되지 않으며, 이로 인해 이와 같은 현상이 발생하는 것으로 판단된다.

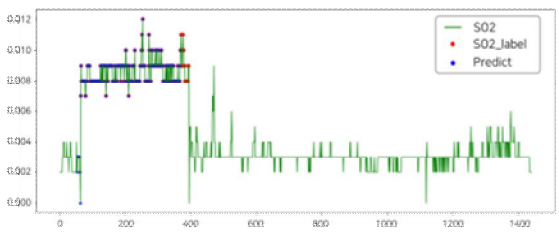

Fig. 6. Example of Experimental Result

Fig. 6은 인공지능 모델의 베이스라인 이상 분류 결과 예시로써, 붉은색 점이 베이스라인 이상으로 레이블된 영역, 파란색 점이 인공지능이 베이스라인 이상으로 분류한 영역을 의미한다. 기존의 레이블과 미묘한 차이는 존재하지만, 대부분의 베이스라인 이상 영역을 성공적으로 분류해낸 것을 확인할 수 있다.

Table 6. Experimental Results for 2018 ~ 2019

Air	W-Code	Pre	Rec	F1	Acc
CO	Normal	0.99	0.99	0.99	0.98
	Abnormal	0.91	0.78	0.84	
O_3	Normal	1.00	1.00	1.00	0.99
	Abnormal	0.93	0.93	0.93	
SO_2	Normal	1.00	0.99	0.99	0.98
	Abnormal	0.84	0.88	0.86	
NO_x	Normal	0.93	1.00	1.00	0.99
	Abnormal	0.96	0.92	0.94	
PM_{10}	Normal	1.00	0.99	0.99	0.99
	Abnormal	0.88	0.98	0.93	
$PM_{2.5}$	Normal	1.00	0.99	0.99	0.99
	Abnormal	0.91	0.95	0.93	
Avg	Normal	0.98	0.99	0.99	0.98
	Abnormal	0.905	0.90	0.90	

4.4 이상치 판별 실험 2

인공지능 모델의 검증을 위해 '16년~'19년 데이터를 적용한 결과는 Table 6과 같이 나타났으며, 전체적으로'18 ~ '19년 데이터를 적용한 Table 7의 결과 대비 다소 낮은 성능을 보였다. 대기오염측정망 데이터는 기후의 영향을 크게 받으며, 이에 따라 계절 단위의 패턴이 나타난다는 특징을 가지고 있는데, 본 모델의 학습에 사용된 데이터는'18년 4월~ '19년 12월까지의 데이터로써 계절별 특성을 학습하기에는 데이터의 양이 부족하여 이러한 현상이 나타난 것으로 판단된다.

Table 7. Experimental Results for 2016 ~ 2019

Air	W-Code	Pre	Rec	F1	Acc
CO	Normal	0.99	0.99	0.99	0.98
	Abnormal	0.87	0.69	0.75	
O_3	Normal	1.00	1.00	1.00	0.99
	Abnormal	0.80	0.75	0.76	
SO_2	Normal	1.00	0.99	0.99	0.98
	Abnormal	0.74	0.79	0.78	
NO_x	Normal	0.93	1.00	1.00	0.99
	Abnormal	0.58	0.61	0.60	
PM_{10}	Normal	1.00	0.99	0.99	0.99
	Abnormal	0.70	0.78	0.74	
$PM_{2.5}$	Normal	1.00	0.99	0.99	0.99
	Abnormal	0.90	0.64	0.75	
Avg	Normal	0.98	0.99	0.99	0.98
	Abnormal	0.765	0.71	0.73	

Table 7. Data analysis per year 2016 ~ 2019

Air	Year	mean	70%	max
CO	2016	0.577	0.645	2.119
	2017	0.557	0.619	2.084
	2018	0.559	0.619	1.984
	2019	0.581	0.645	2.102
O_3	2016	0.020	0.026	0.118
	2017	0.021	0.027	0.122
	2018	0.021	0.026	0.140
	2019	0.023	0.030	0.145
SO_2	2016	0.0049	0.0054	0.0186
	2017	0.0046	0.0050	0.0169
	2018	0.0042	0.0046	0.0150
	2019	0.0038	0.0043	0.0141
NO_x	2016	0.072	0.082	0.453
	2017	0.068	0.076	0.460
	2018	0.063	0.071	0.445
	2019	0.060	0.066	0.420
PM_{10}	2016	50.55	59.04	373.39
	2017	48.34	56.24	328.74
	2018	43.01	50.63	357.87
	2019	43.82	49.25	245.69
$PM_{2.5}$	2016	16.41	19.52	72.41
	2017	18.69	22.30	98.91
	2018	23.36	27.87	144.76
	2019	25.23	27.89	174.60

Table 8은 모델의 성능 저하 원인을 분석하기 위해 각 연도별 정상 데이터의 평균과 하위 70% 농도값 그리고 최대값을 도출한 결과이다. '16 ~ '17년도와 '18 ~'19년도 사이에 값의 변동이 크지 않았던 CO, O3, SO2에 대한 예측 성능은 CO의 재현율이 낮은 것을 제외하면 안정적이지만, '16 ~ '17년도와 '18 ~'19년도 사이에 비교적 큰 변동이 일어난 NO_x, PM_{10}, $PM_{2.5}$의 예측 성능은 전체적으로 낮거나 해당 연도의 예측 성능이 극단적으로 낮게 나타났다. 이러한 현상은 학습 데이터의 부족으로 연 단위의 값 변동에 대한 모델의 대응성이 부족하여 나타나는 것으로 추정된다. 특히 농도의 변동성이 큰 측정값의 입력이 다른 성분의 예측 성능에도 영향을 미치는 것으로 판단된다.

또한, 본 연구에서는 최대 최소 정규화(Min-Max Regularization)를 통해 데이터를 정규화하였는데, 미세먼지 성분의 최댓값이 유독 다른 성분에 비해 변동성이 크기 때문에 이러한 데이터에 대해서는 다른 전처리 수단을 취해 볼 필요가 있을 것으로 판단된다.

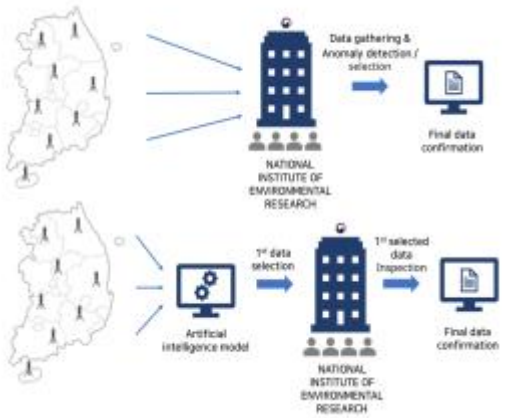

Fig. 7. Our new process of determining the air pollution measurement data with deeplearning network

5. 결론

본 논문에서는 국립환경과학원의 국가대기오염측정망 데이터로부터 '베이스라인 이상' 증상을 판정하는 연구를 진행하였다. 이러한 증상은 기계적인 방법으로는 판정이 어렵기 때문에 판정 과정에 반드시 전문 지식을 가진 인력을 필요로 한다. Fig. 7은 인공지능 모델이 적용된 이상치 판정 프로세스를 나타내며, 본 연구를 통해 이러한 작업을 인공지능 모델로 일정 부분 대체함으로써 전문인력의 업무 부담이 낮아지고 더욱 생산적인 일에 집중할 수 있을 것으로 전망된다.

초기 2018~2019년 학습 데이터로 학습을 수행한 모델에 2018~2019년 2년치 테스트데이터와 2016~2019년 4년치 테스트데이터를 각각 적용하였을 경우, 2016~2019년 4년치 데이터의 적용결과가 2018~2019년 2년치 데이터 적용결과에 비하여 다소 낮은 성능을 보였다. 이는 학습 데이터의 확충 및 데이터의 추가 전처리 등을 통해 개선될 수 있을 것으로 전망된다.

현재 국립환경과학원의 국가대기오염측정망 자료는 규칙 기반의 이상 자료 레이블 분류 후 규칙만으로 선별할 수 없는 기타 요인(지역적 특성, 장비 교정상태)을 고려하여 최종 확정된 자료로써, 비정형화된 기타 요인에 따라 학습에 악영향을 끼칠 가능성이 존재한다. 이처럼 비정형화된 기타 요인에 따라 레이블링 된 레이블

을 노이즈 레이블(Loise Label)이라 한다. 현장의 데이터는 이러한 노이즈 레이블이 존재할 수밖에 없으므로 이러한 레이블을 포함한 데이터를 이용해 인공지능 모델을 학습하는 기법들에 관한 연구가 활발하게 진행 중이며, 본 연구에서는 레이블의 일관성 부족 문제를 완화하기 위해 잡신호 레이블이 존재하는 데이터의 학습 기법을 적용하는 방안을 검토 중이다.

레이블 문제를 해결하기 위한 추가적인 방안으로써, 별도의 레이블 없이 데이터만으로 학습이 가능한 오토인코더(AutoEncoder)[13], 적대적 생성 신경망(Generative Adversarial Network, GAN)[14] 등과 같은 비지도학습 모델을 적용하는 방안 또한 검토 중이다. 지도학습 모델보다 정확도는 다소 떨어질 수 있지만, 레이블이 필요하지 않기 때문에 현행 데이터에 적용하였을 때 더 개선된 결과를 얻을 수 있을 것으로 전망된다.

REFERENCES

[1] L. C. Chen, G. Papandreou, F. Schroff & H. Adam。(2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587.*

[2] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy & A. L. Yuille。(2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40(4),* 834-848.
DOI : 10.1109/TPAMI.2017.2699184

[3] K. He, X. Zhang, S. Ren & J. Sun. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
DOI : 10.1109/CVPR.2016.90

[4] H. I. Fawaz et al. (2020). Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery, 34(6),* 1936-1962.
DOI : 10.1007/s10618-020-00710-y

[5] W. Tang, G. Long et al. (2021). Rethinking 1D-CNN for Time Series Classification: A stronger baseline. *arXiv preprint arXiv:2002.10061.*

[6] N. A. Zainuri, A. A. Jemain & N. Muda. (2015). A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana, 44(3),* 449-456.

[7] J. Faouzi & H. Janati, (n.d.). pyts: A Python Package for Time Series Classification. 6.

[8] Y. Kim & H. Park. (2019). Comparision of Missing Imputaion Methods In fine dust data. *The Journal of Bigdata, 4(2),* 105-114.

[9] H. Ren et al. (2019). Time-series anomaly detection service at microsoft. In Proceedings of the 25th ACM SIGKDD *International Conference on Knowledge Discovery & Data Mining* (pp. 3009-3017).

[10] B. Zhou, S. Liu, B. Hooi, X. Cheng & J. Ye. (2019, August). BeatGAN: Anomalous Rhythm Detection using Adversarially Generated Time Series. *In IJCAI* (pp. 4433-4439).

[11] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam & J. Gottschlich. (2019). Precision and Recall for Time Series. *arXiv preprint arXiv:1803.03639.*

[12] H. Ren. (2019). Time-Series Anomaly Detection Service at Microsoft. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3009-3017).
DOI : 10.1145/3292500.3330680

[13] D. Bank, N. Koenigstein & R. Giryes. (2020). Autoencoders. *arXiv preprint arXiv:2003.05991.*

[14] I. J. Goodfellow et al. (2014). Generative Adversarial Networks. *ArXiv:1406.2661.* http://arxiv.org/abs/1406.2661

이 선 우(Seon-Woo Lee) [정회원]

· 2017년 2월 : 인하대학교 컴퓨터정보학과 (공학사)
· 2019년 2월 : 인하대학교 컴퓨터공학과 (공학석사)
· 2019년 3월 ~ 현재 : 인하대학교 전기컴퓨터공학과 박사과정

· 관심분야 : 머신러닝, 딥러닝, 이상진단, 데이터분석
· E-Mail : x21999@inha.edu

양 호 준(Ho-Jun Yang) [정회원]

· 2021년 2월 : 인하대학교 컴퓨터공학과 (공학사)
· 2021년 3월 ~ 현재 : 인하대학교 전기컴퓨터공학과 석사과정
· 관심분야 : 머신러닝, 임베디드 인공지능, 데이터분석
· E-Mail : hjyang@inha.edu

이 문 형(Lee-Mun Hyung) [정회원]

· 2016년 2월 ~ 현재 : 인하대학교 컴퓨터공학과 재학
· 관심분야 : 인공지능
· E-Mail : mun0659@inha.edu

최 정 무(Jung-Mu Choi) [정회원]

· 2015년 2월 ~ 현재 : 인하대학교 컴퓨터공학과 재학
· 관심분야 : 인공지능, 열화상
· E-Mail : cjm4788@inha.edu

윤 세 환(Se-Hwan Yun) [정회원]

·2018년 3월 ~ 현재 : 인하대학교 컴퓨터공학과 재학
·관심분야 : 머신러닝, 통계
·E-Mail : 12181638@inha.edu

권 장 우(Jang-Woo Kwon) [정회원]

· 1990년 2월 : 인하대학교 전자공학과 졸업
· 1992년 2월 : 인하대학교 전자공학과석사
· 1996년 8월 : 인하대학교 전자공학과박사
· 1998년 2월 : 특허청 사무관
· 2009년 12월 : 동명대학교 컴퓨터공학과 부교수
· 2012년 2월 : 정보통신산업진흥원 인재양성단장
· 2012년 ~ 현재 인하대학교 컴퓨터공학과 교수
· 관심분야 : 인공지능, 인간과 컴퓨터 상호작용, 딥러닝
· E-Mail : jwkwon@inha.ac.kr

박 지 훈(Ji-Hoon Park) [정회원]

· 2006년 2월 : 인하대학교 환경토목공학부 졸업
· 2008년 2월 : 인하대학교 환경공학과(공학석사)
· 2008년 2월 ~ 현재 : 국립환경과학원 대기환경연구과 연구원

· 관심분야 : 모델링, 대기오염측정망
· E-Mail : pjhdo80@korea.kr

정 동 희(Dong-Hee Jung) [정회원]

· 2012년 2월 : 영남대학교 환경공학과 졸업
· 2014년 2월 : 영남대학교 환경공학과 석사 졸업
· 2014년 4월 ~ 현재 : 국립환경과학원 대기환경연구과 전문연구원 재직중

· 관심분야 : 미세먼지, VOCs
· E-Mail : ehdgml6869@korea.kr

신 혜 정(Hye-Jung Shin) [정회원]

· 1999년 2월 : 이화여자대학교 환경공학과 환경공학과 졸업(공학사)
· 2001년 2월 : 이화여자대학교 환경학과 대기환경(석사) 졸업
· 2011년 8월 : 이화여자대학교 환경학과 대기환경(박사) 졸업
· 2003년 ~ 2006년 대구지방환경청 환경연구사
· 2006년 ~ 2007년 국립환경과학원 환경연구사
· 2009년 9월 ~ 2009년 11월 캐나다 환경청 환경연구사
· 2007년 3월 ~ 현재 국립환경과학원 환경연구관
· 관심분야 : 인공지능, 미세먼지, 대기오염측정망 구축
· E-Mail : shjoung@korea.kr