데이터 시각화 기반의 UCI Sensor Data 분석

*장일식 **최희조 ***박구만

*서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공

**서울과학기술대학교 일반대학원 미디어IT공학과

***서울과학기술대학교 전자미디어IT공학과

*foreverme@naver.com

UCI Sensor Data Analysis based on Data Visualization

*Chang, Il-Sik **Choi, Hee-jo ***Park, Goo-man

*Dept. of Information Technology and Media Engineering

**Dept. of Media IT Engineering

***Dept. of Electronics and IT Media Engineering

Seoul National University of Science and Technology

요약

대용량의 데이터를 시각적 요소를 활용하여 눈으로 볼 수 있도록 하는 데이터 시각화에 대한 관심이 꾸준히 증가하고 있다. 데이터 시각화는 데이터의 전처리를 거쳐 차원 축소를 하여 데이터의 분포를 시각적으로 확인할 수 있다. 공개된 데이터 셋은 캐글(kaggle), 아마존 AWS 데이터셋(Amazon AWS datasets), UC 얼바인 머신러닝 저장소(UC irvine machine learning repository)등 다양하다. 본 논문에서는 UCI의 화학 가스의 데이터셋을 이용하여 딥러닝을 이용하여 다양한 환경 및 조건에서의 학습을 통한 데이터분석 및 학습 결과가 좋을 경우와 그렇지 않을 경우의 마지막 레이어의 특징 벡터를 시각화하여 직관적인 결과를 확인 가능 하도록 하였다. 또한 다차원 입력 데이터를 시각화 함으로써 시각화 된 결과가 딥러닝의 학습결과와 연관이 있는지를 확인 한다.

1. 서론

시각화는 측정된 대량의 데이터를 분석하여, 특징추출, 정보 결집 등을 통해 시각적 표현을 수행한다. 많은 양의 다양한 데이터를 단순하게 보유하는 게 아니고 어떻게 활용할 것인가에 대한 부분이 중요해 지고 있다. 활용하는 분야로 데이터 분석과 시각적 스토리텔링이 있다. 전자는 기술적으로 데이터를 수집하고 정제하는 데이터 가공 단계부터 분석 기법을 활용한 데이터 분석에 이르는 과정에 필요한 능력을 말한다. 후자는 데이터 분석 결과를 시각적으로 표현해 스토리텔링을 하는 능력이다. 과거에는 이 두 역량 간 경계가 뚜렷했던 반면, 최근 '이것'의 활용으로 인해 두 역량을 아우르는 사람들이 등장, 이들의 역할에 대한 중요성이 강조되고 있다. '이것'은 무엇일까? 바로 '데이터 시각화'이다. 본 논문은 UCI의 화학 가스의 데이터셋[1]을 이용하여 센서 데이터의 특징을 세 가지로 나누어 학습을 하여 데이터 분석 및 학습 결과가 좋을 경우와 그렇지 않을 경우의 마지막 레이어의 특징 벡터를 시각화하여 학습결과와 비교하여 확인이 가능 하도록 하였다. 또한 다차원 입력 데이터를 시각화함으로써 시각화 된 결과가 딥러닝의 학습결과와 서로 연관이 있는지를 확인하였다. 본 논문의 다수의 표와 그림은 On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines[2]에서 가져왔다. 데이터 분석 및 딥러닝은 Python을 이용하여 구현하였으며, 데이터 시각화의 차원 축소는 LDA(Linear Discriminant Analysis) 방법을 사용하고, 그림 및 학습 결과는 Matplotlib를 사용하여 표현하였다.

2. 가스 분류를 위한 DNN 방법 및 시각화

UCI Sensor의 종류는 표 1과 같다. MOX 센서는 8개 센서 어레이를 포함한다. 각각의 센서는 화학 물질에 따라 다르게 반응한다.

테스트 베드는 그림1과 같이 각각의 위치별로 MOX 센서가 위치한다. 각각의 P1~P6 위치에 9개의 MOX 센서가 위치 한다. 총 54 위치에서의 센서 값을 수집하여 확인할 수 있다. MOX 센서는 총 8개의 값을 전달하므로, 54 X 8 의 값을 확인할 수 있다.

표 1. Sensor 종류

Sensor type	Number of units	Target gases
TGS2611	1	Methane
TGS2612	1	Methane, propane, butane
TGS2610	1	Propane
TGS2600	1	Hydrogen, carbon monoxide
TGS2602	2	Ammonia, H_2S, volatile organic compounds (VOC)
TGS2620	2	Carbon monoxide, combustible gases, VOC

테스트 베스에서 fan 속도는 1500rpm, 3900rpm, 5500rpm 이고, 화학 센서의 내장 발열체에 적용되는 전압 값은 4V, 4.5V, 5V, 5.5V, 6V 으로 구분 된다. Fan 속도, 발열체 전압의 변화를 통한 다양한 센서 데이터 측정가능 하다.

그림 2는 메탄 Gas 주입 시 Sensor 값을 나타낸다.

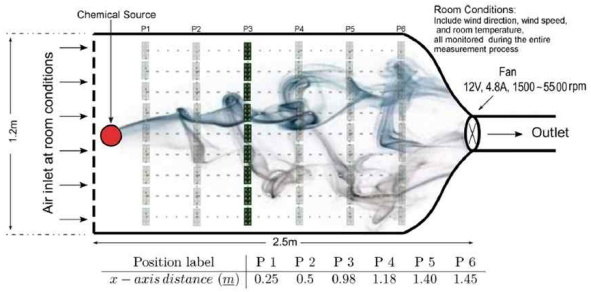

그림 1. 테스트 베드

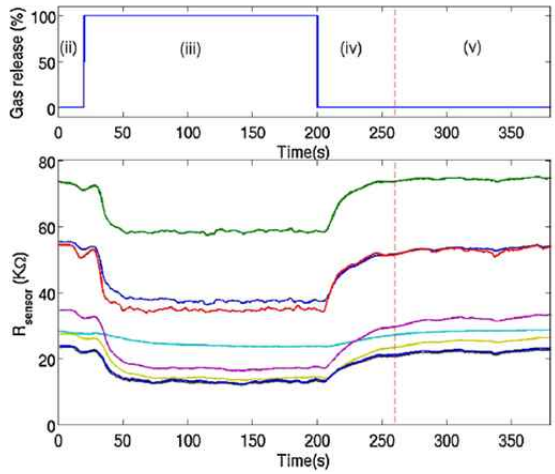

그림 2. 메탄 Gas 주입 시 Sensor 값

(ii) 20 초 동안 화합물이 방출되지 않은 상태에서 센서 신호의 기준선을 측정한다.

(iii) 3 분 동안 화합물의 방출

(iv) 1 분 동안 깨끗한 공기를 순환시켜 센서의 복구 신호를 획득한다.

(v) 테스트 필드를 청소하기 위해 2 분 동안 최대 속도로 잔류가스를 제거한다.

Class는 표 2와 같이 총 11 가지로 구분할 수 있다.

표 2. 가스 주입에 따른 Class

Chemical analyte name	Molecular formula	Concentration
Acetone	C_3H_6O	2500 ppm
Acetaldehyde	C_2H_4O	500 ppm
Ammonia	NH_3	10,000 ppm
Butanol	C_4H_9OH	100 ppm
Ethylene	C_2H_4	500 ppm
Methane	CH_4	1000 ppm
Methanol	CH_4O	200 ppm
Carbon monoxide	CO	1000 ppm/4000 ppm
Benzene	C_6H_6	200 ppm
Toluene	C_7H_8	200 ppm

실제 UCI Dataset은 20msec 단위로 0 ~ 260 Sec 동안 저장되어 있다. P1 ~ P6 위치 별 MOX 센서 값이 측정된다. 본 논문에서는 안정적인 가스 화합물 배출시간인 70초 ~ 200초까지의 데이터를 이용한다. 또한 1~20초에 하나의 센서 값을 데이터로 사용하고, Heater 및 Fan 속도는 상관하지 않는다.

가스 분류를 위한 학습 방법은 총 3가지로 구분하여 테스트를 진행한다. 첫 번째는 DNN 방법으로 8개의 센서 값을 특징으로 하여 학습한다. 두 번째는 DNN 방법으로 9개의 위치별 센서(9X8:72개)를 특징으로 하여 학습한다. 마지막 세 번째는 1D Convolution Network 방법으로 9개의 위치별 센서값을 공간적인 상관관계를 이용하여 학습한다.

그림 3은 첫 번째 방법의 DNN 모델[3]을 나타낸다.

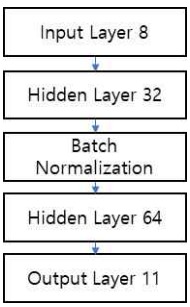

그림 3. 첫 번째 DNN 모델

x_train: 276166, x_valid: 30686, x_test: 34095

test_loss: 0.5580711434551445

test_accuracy: 0.7710221409797668

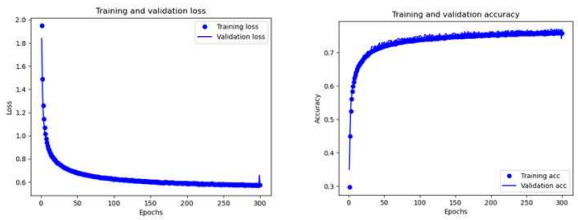

그림 4. 첫 번째 DNN 손실, 정확도

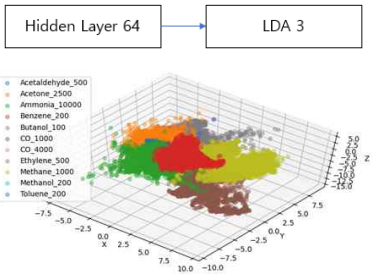

그림 5. 첫 번째 DNN 마지막 레이어 시각화

첫 번째 8개의 센서 값을 특징으로 하여 학습한 결과는 정확도가 0.77로 좋지 못한 성능을 나타내고 있다. 학습 후 마지막 레이어를 시각화하여 그림 5와 같이 확인하였다. 시각화 결과 역시 분류가 잘 되지 않음을 알 수 있다.

그림 6은 두 번째 방법의 DNN 모델[3]을 나타낸다.

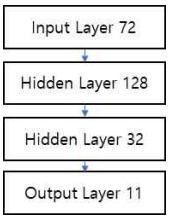

그림 6. 두 번째 DNN 모델

x_train: 7670, x_valid: 853, x_test: 948
test_loss: 0.0260751939099675
test_accuracy: 0.9936708807945251

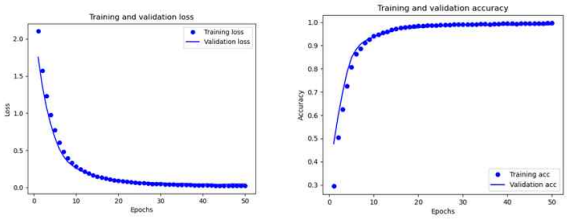

그림 7. 두 번째 DNN 손실, 정확도

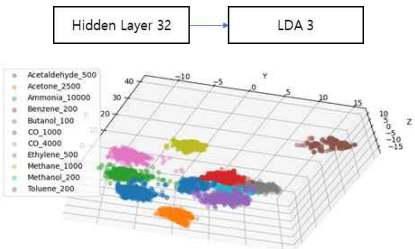

그림 8. 두 번째 DNN 마지막 레이어 시각화

두 번째 9개의 위치별 센서(9X8:72개)를 특징으로 하여 학습한 결과는 정확도가 0.99로 좋은 성능을 나타내고 있다. 학습 후 마지막 레이어를 시각화하여 그림 8와 같이 확인하였다. 시각화 결과 역시 분류가 잘 됨을 확인할 수 있다.

그림 9는 세 번째 방법의 DNN 모델을 나타낸다.

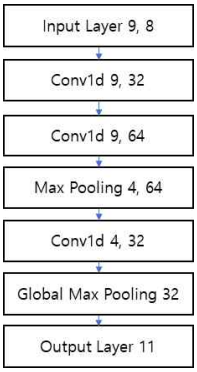

그림 9. 세 번째 DNN 모델

x_train: 7670, 72, x_valid: 853, 72, x_test: 948, 72
test_loss: 0.03113413624119658
test_accuracy: 0.9915611743927002

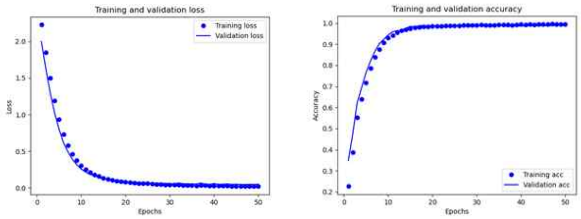

그림 10. 세 번째 DNN 손실, 정확도

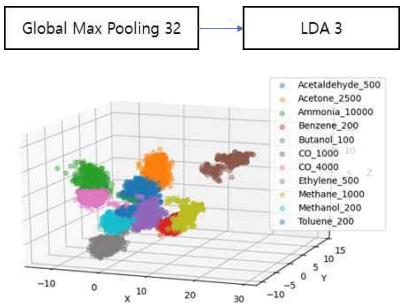

그림 11. 세 번째 DNN 마지막 레이어 시각화

세 번째 1D Convolution Network[3] 방법으로 9개의 위치별 센서 값을 공간적으로 상관관계를 이용하여 학습한 결과는 정확도가 0.99로 좋은 성능을 나타내고 있다. 학습 후 마지막 레이어를 시각화하여 그림 11과 같이 확인하였다. 시각화 결과 역시 분류가 잘 됨을 확인할 수 있다.

그림 12는 각각의 위치(P1~P6) 에서의 9개의 위치 별 센서 입력 값을 차원 축소하여 시각화한 그림이다. 72개의 특징을 사용하였으며 P4 위치에서의 데이터를 시각화 하였다.

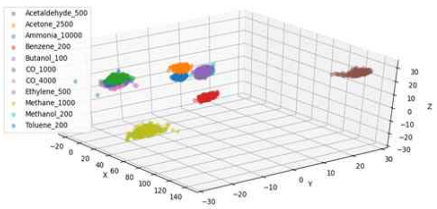

그림 12. 각 위치에서의 위치 별 센서 값 시각화

그림 13은 전체 위치에서 센서 입력 값을 차원 축소하여 시각화한 그림이다. 8개의 특징을 사용하였으며 P4 위치에서의 데이터를 시각화 하였다.

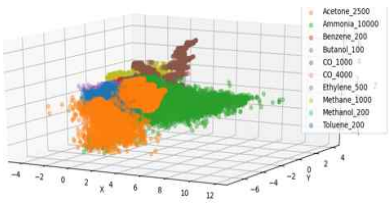

그림 13. 전체 센서 값 시각화

그림 12, 13을 통해 입력 센서 데이터의 시각화를 통해 데이터의 분포를 확인하고 해당 데이터를 딥러닝 학습에 사용할 경우 그림 4, 그림 7에서 확인할 수 있듯이 정확도와 상관관계가 있음을 알 수가 있다.

3. 결론

본 논문에서는 UCI 센서 데이터셋을 이용하여 센서의 입력 특징을 다양하게 선택 했을 때 학습의 결과를 보고, 입력 데이터의 특징을 시각화 하여 학습의 정확도와 상관관계가 있음을 보았다. 단순 9개의 특징을 입력으로 한 DNN은 특징 정보의 부적합성으로 인하여 정확성이 떨어짐을 볼 수 있다. DNN 은 각각의 위치에서의 입력 값을 모두 특징으로 잡아 처리한 결과 좋은 성능을 보였다. 1D CNN 은 각각의 위치에서 위치별 상관관계도 포함하여 학습을 한 결과 좋은 성능을 보였다. 고차원의 경우 데이터 분포를 직관적으로 알 수 없다. 차원 축소를 통해 실제 데이터의 분포를 시각적으로 확인할 수 있다. 또한 학습의 마지막 레이어 구간을 차원 축소하여 시각화 할 경우 시각적으로 구분이 잘 되면, 분류 성능이 좋은 결과 나타냄을 볼 수 있다. 입력 데이터를 선택 할 경우 데이터 시각화를 통한 불필요한 입력 특징을 제거하는데 판단의 근거로 삼을 수 있을 것이다. 좀 더 복잡한 구조 및 실시간 센서 데이터를 통한 사용자 친화적 데이터 시각화 부분의 연구를 진행 할 예정이다.

Acknowledgement

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01106, Sub-ppb급 가스성분 감지를 위한 후각지능 기술 개발)

참고문헌

[1] Gas sensor arrays in open sampling settings Data Set http://archive.ics.uci.edu/ml/datasets/Gas+sensor+arrays+in+open+sampling+settings?ref=datanews.io

[2] Alexander Vergara, Jordi Fonollosa, Jonas Mahiques, Marco Trincavelli, Nikolai Rulkov, Ramon Huerta, On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines, Sensors and Actuators B 185 (2013)

[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks NIPS, 2012