창의융합프로젝트 #2

2021. 04. 16.

어문정, 서장원
Seoul National University
Graduate School of Convergence Science and Technology
Applied Data Science Lab.

목 차

- Project summary
- Data set information
- Several tips for the competition

Project Summary

❖ Goal

- 가스 센서 데이터를 활용하여 CO 가스 농도를 예측하는 모델을 구현.

❖ Description

- CO 농도는 16개의 센서로 측정되었으며, 최초 10시간 동안 측정한 데이터(train.csv)를 이용해 이후 2시간의 측정치 데이터 (test.csv)의 CO 농도를 예측하는 모델을 구현하는 것이 주요 과제.

CO concentration ↑ | sensor resistance

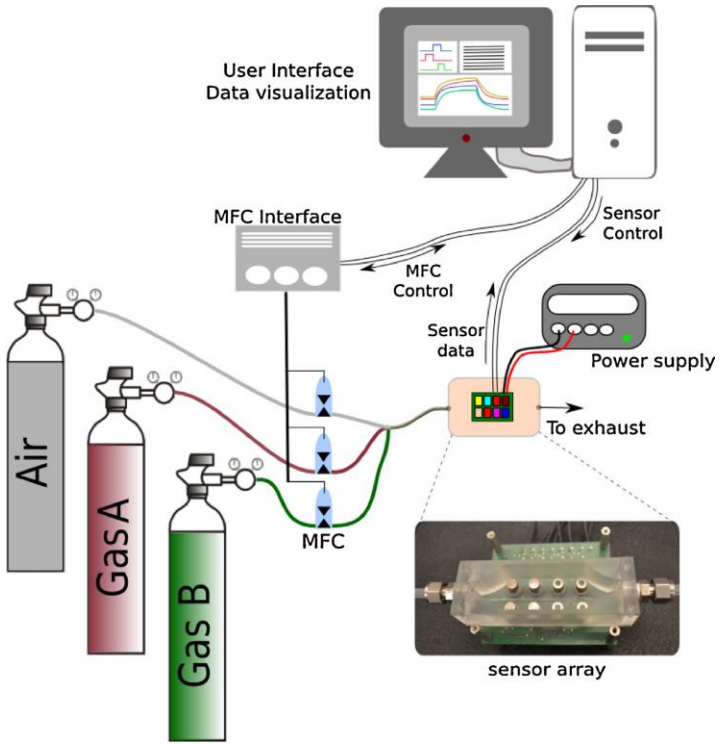
time	CO	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16
0	0	-50.85	-1.95	-41.82	1.3	-4.07	-28.73	-13.49	-3.25	55139.95	50669.5	9626.26	9762.62	24544.02	21420.68	7650.61	6928.42
0.01	0	-49.4	-5.53	-42.78	0.49	3.58	-34.55	-9.59	5.37	54395.77	50046.91	9433.2	9591.21	24137.13	20930.33	7498.79	6800.66
0.01	0	-40.04	-16.09	-27.59	0	-7.16	-42.14	-12.52	-5.86	53960.02	49299.3	9324.4	9449.81	23628.9	20504.94	7369.67	6697.47
0.03	0	-47.14	-10.57	-32.28	4.4	-11.22	-37.94	-7.16	-1.14	53047.71	48907	9170.64	9305.58	23101.66	20101.42	7285.13	6578.52
0.04	0	-33.58	-20.79	-33.25	6.03	3.42	-34.22	-14.46	8.31	52700.28	48330.96	9073.64	9163.47	22689.54	19694.07	7156.74	6468.32
0.05	0	-48.59	-11.54	-36.16	6.03	0.33	-29.05	-16.74	-1.14	51910.52	47609	8982.88	9021.08	22159.12	19332.57	7067.61	6385.31
0.06	0	-48.27	-9.11	-31.31	5.37	-7.97	-30.34	-8.62	7	51458.44	47047.36	8860.51	8966.48	21931.81	19027.69	6976.13	6300.97

※ train data : 337만개, test data : 84만개

❖ Evaluation

- 모델 평가 matrix는 MAE(Mean Absolute Error). $MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$

Data information



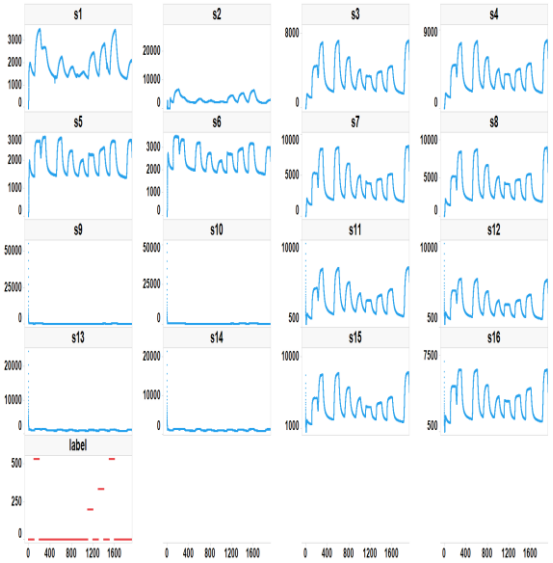
- **센서**
  - 금속산화물로 이루어진 4 종류의 public 화학센서
    (TGS-2600, TGS-2602, TGS-2610, TGS-2620).
  - 데이터 수집 시 각 센서는 네 개씩 사용됨 (총 16개의 센서).

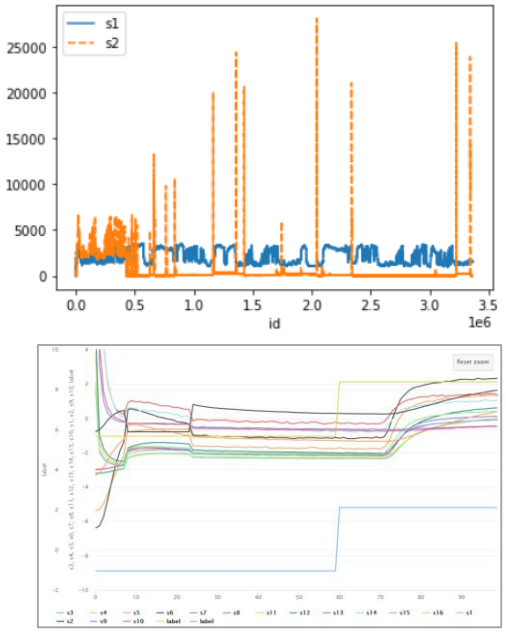
- **데이터 수집 특성**
  - 가스 주입 지속시간은 80~120s 의 범위를 갖는 한에서 random하게 주입, 총 12 시간 동안 실험 진행.
  - 가스 농도는 CO 0-600 **ppm** 농도 범위에서 random하게 변화 시킴.
  - 센서 온도는 일정하게 유지.
  - **100 Hz sampling frequency**.
  - 정답 y label로 사용될 농도 값은 chamber내에서의 가스 농도가 아닌 주입 농도임.

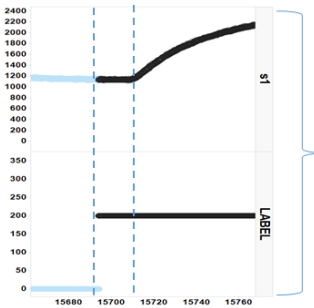
Reference:
Fonollosa et al. 'Reservoir Computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring'; Sensors and Actuators B, 2015

(1) 센서값의 특징 확인

- 센서값이 불안정한 구간은 없는가?
- Broken된 센서는 없는가?
- 센서가 가스주입 즉시 반응하는가?







Gas 주입 후

sensor 반응 까지 delay

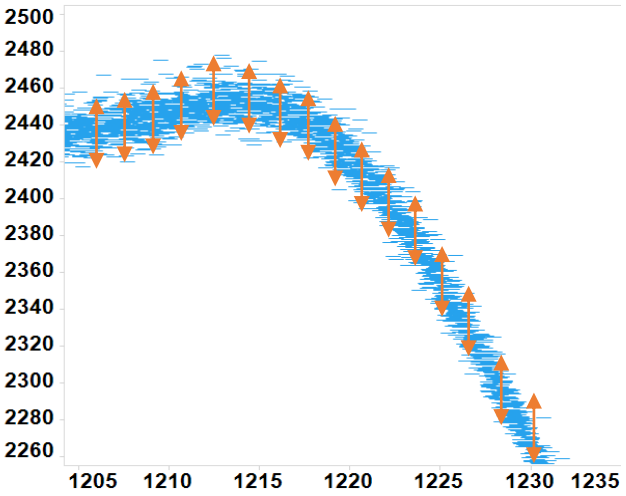
1645 rows count ≒16.45초

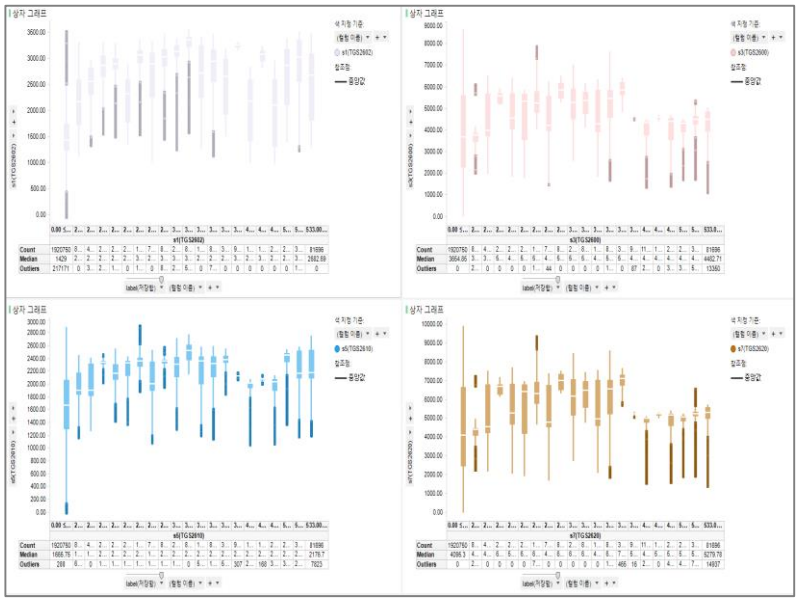
(2) Normalization

: 화학센서 데이터에 대해 대표적으로 min-max normalization이나 base normalization이 사용됨.

- min-max normalization
  : training data 의 min 값과 max 값으로 training dataset과 validation/test dataset을 (X-min)/(max-min) 으로 정규화.
- base normalization
  : training data의 특정 시간 t 동안의 평균값을 산출하여 전체 data에 대해 (X-평균값)/평균값 으로 정규화.

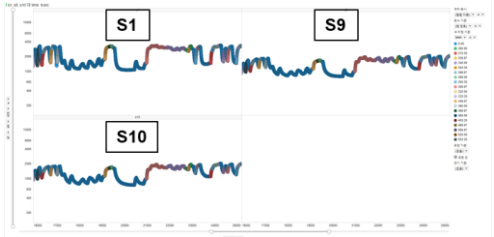
(3) Noise Reduction (eg. Moving Average, Down Sampling, outlier 제거)



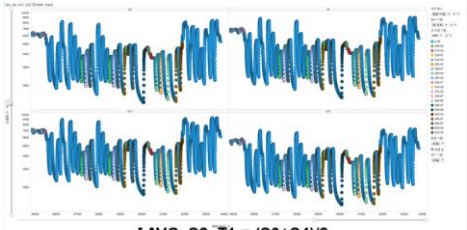


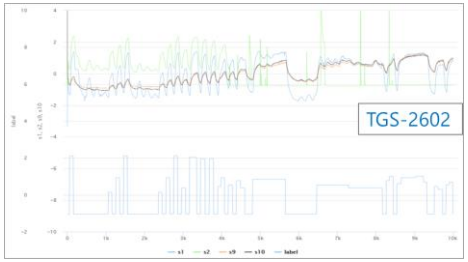
(1) 모든 센서값을 사용할 것인가?

- 동일한 종류의 센서가 4개씩 사용 되었는데 모두 필요한 정보일까?
- 같은 종류 센서 수집값들을 average 하는 등을 통해 새로운 feature 생성하면 도움이 되는가?
- 특정 종류의 센서가 CO 가스에 더 sensitive하게 반응하지는 않는가?



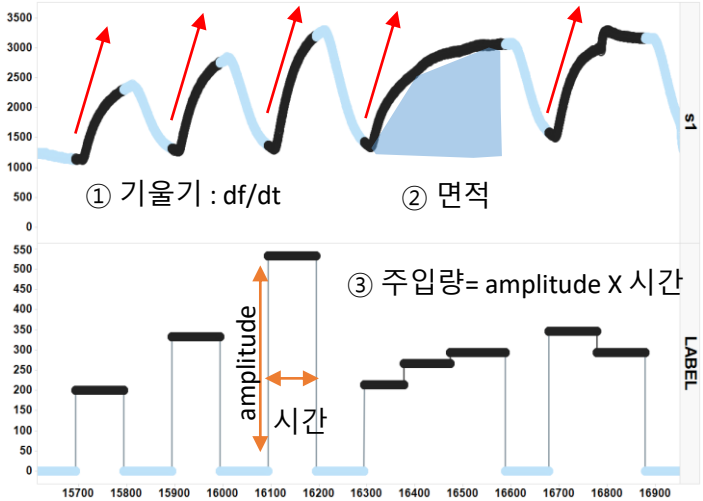
[AVG_S1_T2=(S9+S10)/2]


[AVG_S2_T1 = (S3+S4)/2,
AVG_S2_T2=(S11+S12)/2]



(2) 미분정보 혹은 적분정보 사용?

: 미분 등의 정보 추가는 센서 타입에 따라 다른 성능변화를 보일 수 있음.

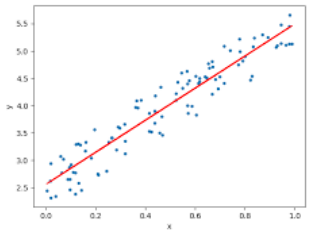


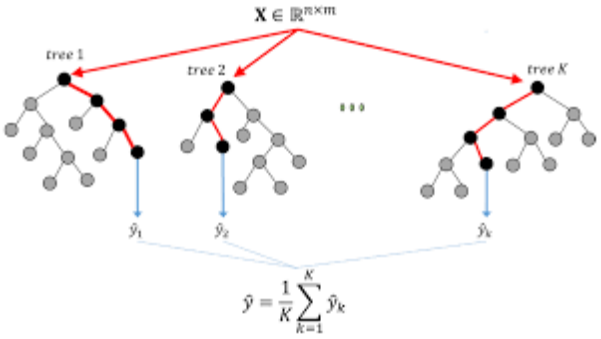
- 각 type의 센서값만 사용했을 때와 센서값과 미분값 (시간 t에서의 센서값 – (t-1)의 센서값)을 모두 사용했을 때 모델의 성능을 비교하면, 모델의 성능이 향상하는 경우도 있고 모델이 망가지는 경우도 있음.

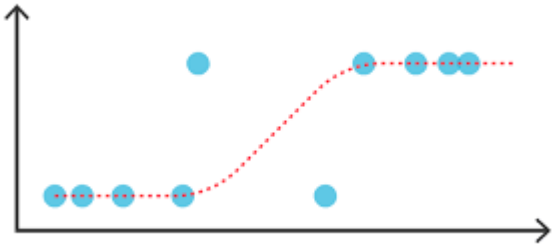
3. Network selection

Machine Learning

Linear Regression, Random Forest Regression, Logistic Regression.. Etc

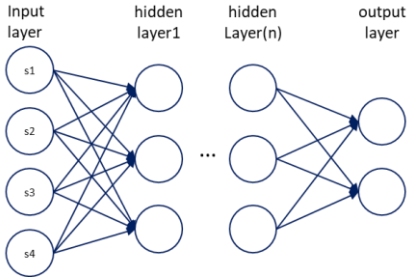


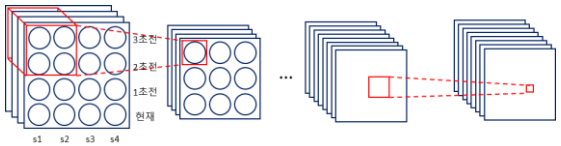


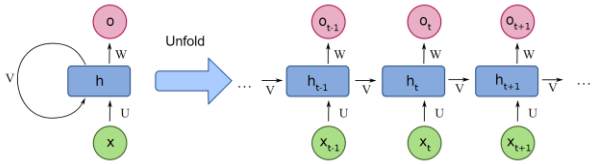


Deep learning

MLP (multi layer perceptron), CNN (convolutional neural network), RNN (Recurrent neural network).. etc





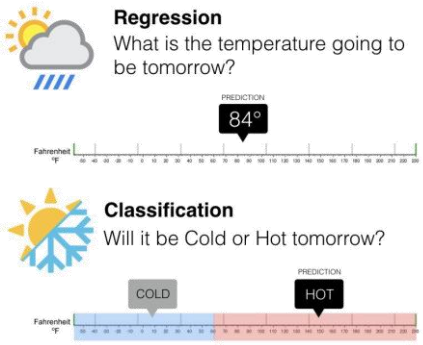


3. Task Selection

Regression vs. classification vs. hybrid tasks

Regression: 가스 농도값을 예측하는 것
Classification: 가스의 노출 여부 (0,1), 가스 농도의 10ppb 단위 구간 등 가스 농도의 class를 예측하는 것
Hybrid: regression과 classification을 동시에 수행하는 것

모델링의 환경과 목표에 맞게 알맞은 task를 선택해야 함.



4. Hyper parameter optimization

For deep networks

No feature engineering

But hyper-parameter optimization instead…