가스 센서 데이터셋 시각화를 위한 데이터 전처리 기법

*김준수 *박경원 *임태범 **박구만
*한국전자기술연구원 **서울과학기술대학교
jesus0084@keti.re.kr

Data Preprocessing Techniques for Visualizing Gas Sensor Datasets

*Kim, Junsu *Park, Kyungwon *Lim, Taebum **Gooman Park
*Korea Electronics Technology Institute
**Seoul National University of Science and Technology

요약

최근 AI(Artificial Intelligence)를 기반으로 정밀한 가스 성분 감지를 위한 후각지능(Olfactory intelligence) 기술에 연구가 활발히 진행 중이다. 후각지능 학습데이터는 다른 감지 방식의 가스 센서들이 동시에 적용되는 멀티모달리티의 특성을 지니며 또한, 공간상에 분포된 센서 배열을 통해 획득된 다차원의 시계열 특성을 지닌다. 따라서 대량의 다차원 데이터에 대한 정확한 이해와 분석을 위해서는 데이터를 전처리하고 시각화할 수 있는 기술이 필요하다. 본 논문에서는 후각지능 학습을 위한 다차원의 복잡한 가스 데이터의 시각화를 위해 잡음 등의 불필요한 값을 제거하고, 데이터가 일관성을 가지도록 하며, 데이터의 차원을 시각화 가능하도록 축소하기 위한 전처리 방법을 제시한다.

1. 서론

최근 IT를 비롯한 여러 분야에서는 딥러닝(Deep Learning)을 이용한 데이터 해석을 통해 필요한 정보를 추출하고 분석하여 인공지능 및 의사결정 분야에 적용이 이루어지고 있다. 특히, 기체 성분의 위험물 검출 분야에서는 AI(Artificial Intelligence)를 기반으로 정밀한 가스 성분 감지를 위한 후각지능(Olfactory intelligence) 기술에 연구가 활발히 진행 중이다. 후각지능 학습데이터는 다른 센서 데이터와 다르게 다른 감지 방식의 가스 센서들이 동시에 적용되는 멀티모달리티(Multi-modality)의 특성을 지닌다. 또한 시간에 따라 확산되는 가스 데이터의 포집을 위해 공간상에 센서 배열을 배치하고 시-공간상의 데이터를 취득하기 때문에 고차원의 데이터 구조를 지닌다. 이와 같은 대량의 복잡한 구조의 데이터 시각화는 사람과 기계가 데이터의 의미와 연관성 등을 분석하고 이해하는 데에 있어 중요한 기술이며, 다차원의 복잡한 데이터에 대한 시각화를 수행하기 위해서는 수집된 데이터들에 일관성을 부여하고 사람이 직관적으로 이해할 수 있는 데이터로 변환하는 전처리(Preprocessing) 과정이 필요하다.

후각지능을 위한 가스 데이터의 전처리 과정에는 잡음 등의 불필요한 값을 제거하고, 데이터가 일관성을 가지도록 센서간에 동기화를 수행하며, 데이터를 시각화에 접한 차원 크기로 변환하는 과정 등이 포함된다. 본 논문에서는 여러 방식의 가스 센서들을 동시에 적용하여 가스 데이터를 수집한 UCI Gas Sensor Dataset 중 개방형 가스 센서 데이터 샘플링 환경 데이터셋을 분석하고, 이를 바탕으로 하여 멀티모달 환경에서 수집된 데이터를 시각화하기 위한 전처리 과정의 방법에 대해 제시한다[1].

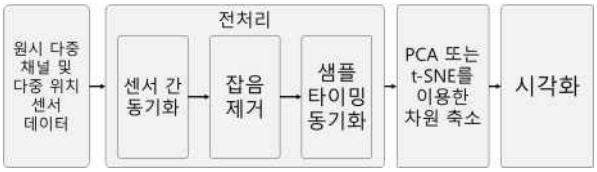
2. 본론

UCI 가스 데이터셋은 2.5m × 1.2m × 0.4m 크기의 평평한 바닥을 가지며 가스 인입구와 노출된 가스를 배기하기 위한 팬 시설이 장착되며, 컴퓨터에 의해 제어되는 3개의 디지털 유체 흐름 제어기가 설치된 개방형 풍동 시설에서 수집되었다. 각 제어기는 지속적으로 가스를 터널 앞쪽에서 입력하고 터널 뒤쪽으로 배출한다. 해당 풍동 시설 내에는 54개의 금속산화물(MOX) 화학 센서 모듈(각 모듈은 8개의 센서로 구성)이 설치되어 있으며, 9개씩 6열로 배열되어 총 432개의 센서가 설치된다. 전체 데이터셋은 10개의 클래스 (아세톤(C_3H_6O), 아세트알데히드(C_2H_4O), 암모니아(NH_3), 부탄올($C4H_9OH$), 에틸렌(C_2H_4), 메탄(CH_4), 일산화탄소(CO), 벤젠(C_6H_6), 톨루엔(C_7H_8))와 18,000개의 인스턴스, 1,950,000개의 속성 ($75 \times 260 \times 100 = 1{,}950{,}000$, 72개의 가스 센서 데이터, 시간, 온도, 상대습도를 포함한 총 75개의 시계열 데이터가 260초간 100Hz로 샘플링 된 데이터)로 이루어져 있다[2]. 이 데이터셋은 주로 각 가스를 감지하기 위하여 종류을 분류(Classification)하는 데에 주로 사용되지만, 가스가 발생한 위치를 예측하는 것과 같은 응용에 적용될 수 있다[3].

2.1 데이터셋의 전처리

앞서 설명한 바와 같이 UCI 개방형 가스 센서 데이터는 큰 차원의 데이터셋으로 구성되기 때문에, 데이터셋의 특성을 시각화를 위해서는 <그림 1>에 나타난 것과 같이 데이터셋을 전처리하고 차원을 축소하는

작업이 필요하다.



<그림 1> 개방형 가스 데이터 샘플링 환경 데이터셋의 시각화 과정

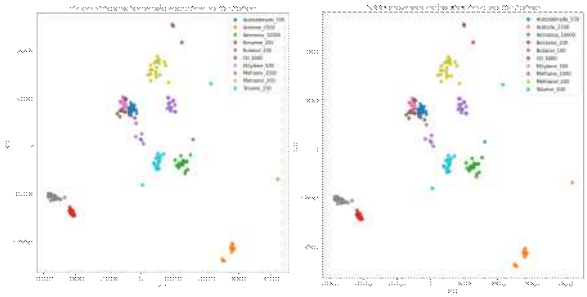
전처리 과정에는 센서 간 동기화, 잡음 제거, 샘플 타이밍 동기화 등의 작업이 포함되며, 고차원의 데이터셋을 사람이 이해할 수 있는 2차원 또는 3차원 공간에 시각화하기 위해 차원 축소 과정을 거친다. 센서로부터 취득된 데이터는 다양한 외부 잡음원에 노출되어 오염되기 때문에 필터 등을 잡음을 제거한다. 또한, 각 가스 센서와 샘플링 ADC(Analog to Digital Converter)간에 완벽한 동기화가 이루어지지 않아 시작 지점이 다르거나, 값이 누락되거나, 값이 정확한 시간에 입력되지 않아 데이터 수집 주기가 어긋나는 등의 오류를 가지고 있다. 이러한 상태에서는 데이터들을 동일한 조건에서 관측하고 처리하기 어렵기 때문에 전처리 과정에서 이들 오류를 보상한다. 먼저 센서 간의 데이터 동기화를 위해 가스가 풍동에 유입되기 시작한 시점을 검출하여 해당 시점 후에 기록된 값만을 사용함으로 센서간에 데이터들의 시작점을 동기화하고 실제 가스가 방출되었을 때의 데이터만을 취득하여 불필요한 데이터를 제거한다. 그 다음 중간값 필터(median filter) 또는 저역통과 필터(low-pass filter) 등을 사용해 잡음을 제거하고, 불균일하게 취득된 데이터를 보간(interpolation)한 뒤 센서의 동작 주파수인 100Hz에 맞춰 재샘플링(resampling)을 수행하였다. 보간한 데이터를 그대로 차원 축소 단계에 사용할 수도 있으며, 필요에 따라 서브샘플링을 추가로 적용하여 데이터셋의 크기를 더 줄이는 것이 가능하다.

2.2 데이터셋의 차원 축소 및 시각화

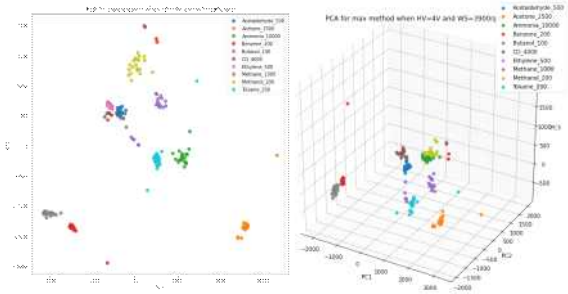
다음 <그림 2>는 전처리를 통해 100Hz로 재샘플된 데이터에 추가적으로 서브샘플링 방식과 최대값을 적용한 후, PCA(Principal Component Analysis) 기법으로 투사하여 시각화한 결과를 보여준다. 그림 (a)는 100Hz로 재샘플이된 데이터를 모두 적용한 것이고, (b)는 1/100로 서브샘플링을 수행 후 PCA를 적용한 것이다. 그림 (c)는 각 센서별로 시각영역에서 최대값을 추출하여 특성으로 추출하고 PCA를 적용한 결과를 나타낸다. 그림 (a)와 (b)의 두 결과를 비교해 보면 결과 상에 거의 차이가 없고, 특성을 유지하며 시간영역 샘플을 수를 크게 줄일 수 있음 알 수 있다. 또한, (c)와 비교하면 투사 결과가 (a)나 (b)와 차이가 있음을 확인할 수 있다.

3. 결론

본 논문에서는 복잡한 다차원 멀티모달 특성을 지닌 후각지능 학습 데이터인 가스 데이터에 대해 분석하고 시각화를 위한 전처리 기법을 제시하였다. 멀티 모달로 구성된 가스 센서 배열을 적용하는 경우 취득 프로토콜에 따라 센서 간 및 데이터 샘플간에 동기화 필요하며, 잡음 제거를 필터링 과정이 필요하다. 본 논문에서는 불균일하게 취득된 데이터에



(a) 전체 데이터 투영 (b) 서브샘플링(1/100) 데이터 투영



(c) 최대값 투영 (d) 3D 투영

<그림 2> PCA 기반 2D 및 3D 시각화

보간 기법을 적용하여 센서간에 균일한 샘플시간이 유지되도록 하였으며, 가스 데이터의 고유 특성을 유지하며 데이터 차원을 줄일 수 있는 서브샘플링 기법을 전처리 과정에 적용하였다.

ACKNOWLEDGMENT

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 settigns Data set", *UCI Machine Learning Repository*, last modified June 6, 2013, accessed Oct. 20, 2021, http://archive.ics.uci.edu/ml/datasets/gas+sensor+arrays+in+open+sampling+settings.

[2] Alexander Vergara, Jordi Fonollosa, Jonas Mahiques, Marco Trincavelli, Nikolai Rulkov, Ramón Huerta, "*On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines*", Sensors and Actuators B: Chemical, Vol. 185, 2013.

[3] Javier Burgués, Santiago Marco, "*Feature Extraction for Transient Chemical Sensor Signals in Response to Turbulent Plumes: Application to Chemical Source Distance Prediction*", Sensors and Actuators B: Chemical, Volume 320, 2020.