Datasheet for 'Air Quality dataset'*

Shuyuan Zheng

March 28, 2024

1 Motivation

This datasheet documents the Air Quality dataset, detailing aspects such as motivation, composition, collection process, and recommended uses. The dataset comprises hourly averaged responses from an array of 5 metal oxide chemical sensors within an Air Quality Chemical Multisensor Device, situated in a heavily polluted urban area in Italy. The data spans from March 2004 to February 2005, offering a comprehensive year-long record of urban air pollution.

2 Composition

The dataset contains 9,358 instances, each representing hourly averaged sensor responses from the deployed device. The following parameters are included:

Date (DD/MM/YYYY) Time (HH.MM.SS) Five distinct metal oxide chemical sensor responses Hourly averaged concentrations of CO, Non-Methane Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2) from a certified reference analyzer Temperature Relative Humidity

Missing values are indicated with a -200 value. The dataset captures various pollution levels, providing a detailed temporal record of urban air quality.

[*] Code and data are available at: https://github.com/EAsUpluckYX/-Air_Quality, reviewed by Frank Meng

3 Collection process

The data for each instance was acquired through real-time measurements from chemical sensor devices, alongside hourly averaged concentrations from a co-located reference certified analyzer(Vito 2016).The chemical sensors detected various pollutant gases, while the reference analyzer provided precise measurements for validation purposes. The dataset represents continuous monitoring data rather than a sample. It encompasses a comprehensive year-long record, capturing a wide range of environmental conditions and pollution levels.

4 Preprocessing/cleaning/labeling

This include calibration of sensor readings, normalization of data to account for environmental variables (e.g., temperature, humidity), and handling of missing values. Missing values, especially common in long-term environmental monitoring due to sensor malfunctions or maintenance periods, were tagged with a specific code (-200) to indicate absence of data.

5 Uses

The dataset has been explicitly utilized for on-field calibration of an electronic nose aimed at benzene estimation within an urban pollution monitoring context. (Vito 2016)This involves correlating sensor responses from the electronic nose with reference measurements to accurately estimate benzene concentrations in the air, a crucial task for understanding urban air quality and its health implications.

This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

6 Distribution

The Air Quality dataset is freely available for research purposes under the condition that it is not used for commercial purposes. Users are required to cite the original source(Vito 2016)when using the dataset in their work.

7 Maintenance

Maintenance involve collaboration with the UCI Machine Learning Repository. This dataset only captures part of the 2004-2005 data and will not be updated in the future.

References

Vito, Saverio. 2016. “Air Quality.” UCI Machine Learning Repository.