Data Article

Synthetic benchmarks for machine olfaction: Classification, segmentation and sensor damage ☆

Andrey Ziyatdinov [a,b,*], Alexandre Perera [a,b]

[a] B2SLab, Department of ESAII, Universitat Politenica de Catalunya, Pau Gargallo 5, Barcelona, Spain
[b] Centro de Investigacion Biomedica en Red en Bioingenieria, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain

ARTICLE INFO

ABSTRACT

The design of the signal and data processing algorithms requires a validation stage and some data relevant for a validation procedure. While the practice to share public data sets and make use of them is a recent and still on-going activity in the community, the synthetic benchmarks presented here are an option for the researches, who need data for testing and comparing the algorithms under development. The collection of synthetic benchmark data sets were generated for classification, segmentation and sensor damage scenarios, each defined at 5 difficulty levels. The published data are related to the data simulation tool, which was used to create a virtual array of 1020 sensors with a default set of parameters [1].

The data presented here are publicly available at the web server of Polytechnic University of Catalonia on the following link ⟨http://neurochem.sisbio.recerca.upc.edu/public/datasets/benchmarks⟩.

Specifications Table

Subject area	*Chemistry, Engineering*
More specific subject area	*Chemometrics, Machine Olfaction, Electronic Nose, Chemical Sensing, Machine Learning*
Type of data	*Table*
How data was acquired	*Data simulation tool*
Data format	*Raw*
Experimental factors	*Statistical models used in the data simulation tool were fitted to a reference data set.*
Experimental features	*An array of 1020 virtual sensors was created by the data simulation tool with the default parameters. The sensor signals were generated in response to a rectangular gas pulse of 60 time units.*
Data source location	*Barcelona, Spain*
Data accessibility	*The data sets are publicly available at the web server of Polytechnic University of Catalonia on the following link* http://neurochem.sisbio.recerca.upc.edu/public/datasets/benchmarks.

Value of the data

- The benchmark material in the field of machine olfaction was published for the first time.
- The proposed definitions of scenarios combined with the data simulation tool can be used as a reference workflow for other scenarios in machine olfaction.
- The generated data sets have concentration profiles of mixtures of analytes, a considerably large number of sensors and realistic noise in the data.

1. Materials and methods

Synthetic benchmarks were an alternative to the real measurements at the middle stage of the Neurochem project, when the main sensor array of the project was under development [2]. The realization of the synthetic experiments required a model of an array of gas sensors. That model needed to capture the main features shown by polymer sensors (the reference data set was measured with an array of conducting polymer sensors) and be simple enough so that it could be included in the system software. The model was implemented in the data simulation tool (the R package chemosensors) [1,3].

The synthetic benchmarks produced for the three scenarios classification, segmentation and sensor damage possess a particular feature of the large number of sensors (1020). This feature will particularly suit for examination of the role of diversity and redundancy among the sensors at large scale. Recent examples of the data analysis based on real large sensor arrays include an array of 96 metal-oxide sensors combined with 10 different sensor families modulated in temperature [5], and an array of 16,384 conducting polymer sensors based on 24 different kinds of polymer materials [6] (both arrays are products of the Neurochem project).

1.1. Scenarios

Ten scenarios for machine olfaction – classification, quantification, segmentation, habituation, event detection, novelty detection, drift compensation I, drift compensation II, sensor replacement I and sensor replacement II – were designed and formalized in the framework of the data simulation tool [3, Supporting Information, File S1]. For three of these scenarios – classification, segmentation, and sensor damage (adopted from sensor replacement scenario) – synthetic benchmark data sets at different difficulty levels were generated.

General definitions of the three scenarios are the following.

1. Classification scenario: John has three vessels with three odors A, B, C. The system is trained with all three compounds separately. John approaches the vessel B to the system. The machine identifies correctly odor B. The difficulty is the similarity between the odors to be identified.