Chapter 15
Methods and Graphical Tools for Exploratory Data Analysis of Artificial Olfaction Experiments

Matteo Falasconi
University of Brescia, Italy

Matteo Pardo
Institute of Applied Mathematics and Information Technology, Italy

Giorgio Sberveglieri
University of Brescia, Italy

ABSTRACT
Visualization and initial examination of the Electronic Nose data is one of the most important parts of the data analysis cycle. This aspect of data investigation should ideally be performed iteratively together with data collection in order to optimize experimental protocols and final results. Once exploration has been completed, a complete supervised data analysis on a full dataset can be run, leading to prediction and thereby to e-nose performance evaluation. Exploratory Data Analysis (EDA) comprises three tasks: checking the quality of the data, calculating summary statistics, and producing plots of the data to get a feel of their structure. Graphical visualization of data allows checking for instrumental malfunctioning, discovering human errors, removing outliers, understanding the influence of experimental parameters, verifying the ability of the machine in discriminating the examined samples, and eventually formulating new hypotheses. A number of different techniques have been developed for data visualization, including multivariate statistical analysis, non-linear mapping, and clustering techniques. This chapter will present an overview of methods, tools, and software for EDA of artificial olfaction experiments. These will cover visualization and data mining tools for both raw and preprocessed data, such as: histograms, scatter plots, feature and box plots, Principal Component Analysis (PCA), Cluster Analysis (CA), and Cluster Validity (CV). Some case studies that demonstrate the application of the methods to specific chemical sensing problems will be illustrated.
INTRODUCTION

Since almost three decades of experience with chemical sensing devices (Persaud, 1982) it is known that the applicative success of this technology strongly depends on proper selection and subsequent optimization of crucial experimental parameters. The experimental outcomes depend on a tremendously high number of variables, such as: sensor type (Röck, 2008; Pardo, 2004) and variables selection (Pardo, 2007; Roussels, 1998), odour sampling approach (Šetkus, 2010; Roussels, 1999), humidity level and/or presence of interfering species (Vezzoli, 2008), time progression and sensor stability (Padilla, 2010; Sharma, 2001; Ionescu, 2000). Therefore, the successful design of an Electronic Nose (EN) requires a careful consideration of the various issues involved in the experimental procedure (Gutierrez-Osuna, 2002).

The visualization and initial examination of the data—also called Exploratory Data Analysis (EDA) (Tukey, 1977)—is one of the most important parts of the data analysis cycle (Webb, 2002). Indeed, EDA is a necessary step in which the user can interact with the machine to check the quality of experimental results before embarking in successive, more automated steps, thus saving lot of unnecessary efforts.

The aims of EDA are manifold: maximize insight into a data set, uncover underlying structure, extract important features, and detect outliers. A most valuable outcome of EDA is to check for prior assumptions, understand how they affect the EN response, and determine the optimal experimental settings.

EDA includes three relevant aspects:

1. **Checking the quality of the data.** A first look at the sensor responses serves to control the correct functioning of the equipment. For example, in the case of the responses of chemical sensors to controlled as mixtures in ENs, the expected form of the response is known. Malfunctioning of the equipment (sampling system, sensors, electronics) can be spotted by plotting the sensor dynamic response vs the acquisition time.

2. **Calculating summary statistics.** Summary statistics can be used to characterize the data: few numbers can convey the fundamental properties of the data set, e.g., by calculating (sample) mean and variance for each feature and for each class it is possible to detect the (more obvious) outliers and to get clues about the variables important for discriminating particular classes.

3. **Producing plots of the data in order to get a feel of their structure.** This aspect of data investigation should ideally be performed iteratively together with data collection in order to adjust the experimental conditions for maximizing the system performance (e.g. samples classification by the EN).

Graphical views of the data (Chernoff faces, radar plots, histograms, scatter plots, multivariate data projections (linear or nonlinear), and cluster analysis) are the most useful tools for providing an insight into the nature of multivariate data. Visual data mining is especially useful at the initial stage of the process when little is known about the data and the exploration goals are vague. Since the user is directly involved in the process, shifting and adjusting the experimental design and the data acquisition can be easily implemented.

This Chapter will provide an overview of single and multiple variable displays techniques and of cluster analysis and cluster validity approaches. A set of Matlab functions (a toolbox) including several graphical tools has been developed in the SENSOR Lab to interactively implement such functionalities for EN data examination. The presentation of the techniques is accompanied by some case studies that demonstrate the practical application of these tools.