Chemometrics and Intelligent Laboratory Systems 130 (2014) 123–134

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

On the calibration of sensor arrays for pattern recognition using the minimal number of experiments

Irene Rodriguez-Lujan [a,*], Jordi Fonollosa [a], Alexander Vergara [b], Margie Homer [c], Ramon Huerta [a]

[a] BioCircuits Institute, University of California, La Jolla, San Diego, CA 92093, USA
[b] Biomolecular Measurement Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8362, USA
[c] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form 18 October 2013
Accepted 19 October 2013
Available online 26 October 2013

Keywords:
Electronic nose
Gas sensing
Support Vector Machines
Fast calibration
Pattern recognition
Gas discrimination
Active learning

ABSTRACT

We investigate optimal experiment selection to train a classifier on gas sensor arrays to get the maximal possible performance in a limited number of experiments. In gas sensing, while collecting data for a particular sensor array, one has to choose what gas and concentration level is going to be presented in the next experiment. It is an active decision by the operator selecting the experiments and training the classifiers. Can the algorithm be trained sooner rather than later? Can we minimize the costs of collecting the data in terms of the man-hour of the operator and the expenses of the experiment itself? Active control sampling provides a way to deal with the challenge of minimizing the calibration costs and is applicable to any situation where experimental selection is parameterized by an external control variable. Our results indicate that active sampling strategies can only improve a random selection of experiments over a wide range of concentration of gasses. However, random or uninformed selection is fairly close. Additionally, our active sampling methodology reveals that, when there is no prior knowledge about the range of concentrations to which the sensor will be exposed during real operation, sensor must be calibrated over the entire working range, not just high concentrations. In fact, our results show that it is especially important to include low concentrations in the calibration since the lack of these values would dramatically decrease the performance of the system.

1. Introduction

The calibration of gas sensor arrays is an expensive, but necessary, process to establish the functional relationship between measured values and analytical quantities. Traditionally, calibration includes first, the selection of the functional form of a computational model; second, the estimation of the corresponding model parameters and the errors based on a training dataset; and third, the model validation [1]. The resulting computational model is then used to deconvolve new measurements and predict the analyte amount/class. However, after a certain period of time, the performance of the model degrades due to the changing characteristics of the sensing elements, and the system needs to be recalibrated.

In particular, the calibration (or recalibration) of solid state gas sensor arrays has been investigated for decades using nonlinear multivariate techniques [2,3]. During these years a large variety of calibration techniques has been investigated for chemical detection systems, including, but not limited to, artificial neural networks, linear discriminants, multilayer perceptrons, k-NN classifiers, partial least square regressors, and more recently, Support Vector Machines [4–7]. However, irrespective of the selected data processing technique, a training dataset needs to be collected to perform the calibration of an analytical system. The generation of the training dataset represents a significant cost in terms of time and budget due to the expenses of the experiment itself and the dedication of technicians. This situation is especially critical in applications where the acquisition of new samples is costly, such as air quality control in space ships, environmental monitoring of public spaces, and industrial leak detection, among others. Additionally, systems based on Metal Oxide (MOX) gas sensors, which still are a common choice for chemical detection applications due to their sensitivity, low cost, ease of operation, ability to detect large number of chemicals, and robustness [8,9], are dynamic systems that show a transient response when exposed to a constant stimulus [10]. Hence, in order to collect a thorough training dataset, it is necessary to capture the complete transient response of the sensors for each training example, thereby making the calibration of a MOX gas sensor array an extensively expensive, laborious, and time consuming operation [11,12].

In order to reduce the frequency of the recalibrations and the associated costs, methodologies aimed at extending the time between recalibrations have been presented that attenuate the effects of sensor drift [10,13–15], sensor failure [16–18], or sensor poisoning [19,20]. However, the selection of the training examples utilized to reduce the cost of the calibration process has been overlooked. The practitioner

* Corresponding author at: BioCircuits Institute, University of California, San Diego, 9500 Gilman Dr., Mail Code 0402, La Jolla, CA 92093-0402, USA. Tel.: +1 858 534 6876.
E-mail addresses: irenerodriguez@ucsd.edu (I. Rodriguez-Lujan), fonollosa@ucsd.edu (J. Fonollosa), alexander.vergaratinoco@nist.gov, vergara@ucsd.edu (A. Vergara), margie.l.homer@jpl.nasa.gov (M. Homer), rhuerta@ucsd.edu (R. Huerta).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.10.012

has to design an experimental protocol to collect the training dataset to maximize the accuracy after the calibration, while reducing the number of training examples and calibration costs.

The challenge of the calibration process is to sample the space of admissible control parameters to achieve superior and stable performance after the operation. When performing a system calibration, each new training example (measurement) has an associated class label (gas type) and a control parameter (gas concentration). It is generally believed that gasses presented at higher concentrations are easier to classify than those at lower concentrations. However, when a chemical detection system is deployed in its operating environment, it is responsible for identifying various gasses from a wide range of concentrations. When selecting a new training example, the question that the experimentalist faces is: what is the next training example (gas class and concentration) that maximizes learning?

This idea of selecting intelligently the next experiment to calibrate the sensor array is in line with the active sampling algorithms proposed in machine learning. The goal of active learning is to improve the performance of a preexisting classifier on-the-fly while accelerating the learning process by incorporating and labeling samples to the learning process, taking into account the information provided by the previous examples. While most theories and methods in machine learning assume independent and identically distributed observations, they make use of *passive learning* strategies based on non-adaptive (usually uniform) sampling. However, there are several theoretical and empirical works in the literature demonstrating the superiority of active sampling over passive learning in terms of generalization error, uncertainty, and stability [21–29]. Unfortunately, the application of the proposed criteria to determine the suitability of active sampling for a real-world problem is unfeasible due to their strong assumptions about the distribution of the data and the noise.

Different paradigms can be found in the active sampling literature namely, *adaptive sampling* or *query learning* [30], *instance selection* or *selective sampling* [31,32] and *pool-based* learning [33]. Adaptive sampling assumes that the learning algorithm actively creates or selects unlabeled samples to be labeled by an expert. An alternative to query learning is selective sampling that, under the hypothesis that obtaining unlabeled instances is inexpensive, decides whether or not to request the label for samples drawn according to the data distribution. Finally, pool-based learning assumes that there exists a big pool of unlabeled instances while there is a small subset of labeled data. Then, the sampling strategy chooses the best query after ranking all the instances in the pool according to a certain measure. Thus, the main difference between pool-sampling and selective sampling is that the former needs to evaluate all the instances in the pool while the latter considers instances individually. For a more detailed review on active sampling, the reader is referred to Ref. [21]. Therefore, the active control sampling strategy suggested in this work is in line with the adaptive sampling approach as the learning algorithm creates queries by selecting the gas concentration level (control parameter) from a set of feasible values.

In contrast to active sensing methodologies that actively adapt the sensor properties or exposure conditions to optimize the system performance [34–36], in this paper we propose a methodology to select the best training examples to calibrate the system without modifying the configuration of the sensor array. Our methodology will allow us to determine (i) whether an active sampling strategy outperforms a uniform selection of experiments; and (ii) the range of concentrations to be used to calibrate the sensor array, especially when there is no prior knowledge about the range of concentrations to which the sensor will be exposed during real operation. The approach is based on Support Vector Machines, a technique that was successfully introduced to classify e-nose data [7,37–39] with actual measurement recordings from a 16-element MOX gas sensor array.

The paper is organized as follows: In Section 2 we describe the experimental setup and dataset. Then, we introduce the formal description of the problem in Section 3, followed by the results and discussion (Section 4) and the conclusions of this work (Section 5).

2. Experimental details

2.1. Dataset and measurement collection procedure

We implemented the suggested active sampling methodology to a dataset recorded over 1 month utilizing sixteen screen-printed MOX gas sensors commercialized by Figaro Inc. [40].[1] The resulting dataset comprises 1800 recordings of three distinct pure gaseous substances, namely ethanol, ethylene and acetaldehyde, each dosed at concentration values ranging from 2.5 μmol/mol (ppm) to 300 μmol/mol (ppm). The distribution of the recordings as a function of the gas type and its concentration is given in Fig. 1. It shows that the distributions for the three gasses are similar, with the concentrations of ethanol being slightly biased toward the lower values. Therefore, the acquired dataset is suitable to explore the calibration strategies since it is well balanced for all the classes and concentrations.[2]

In order to visualize the distribution of the data, Fig. 2 shows the data projected into the 2-dimensional and 3-dimensional spaces defined by the first two and three principal components, respectively, and obtained from Principal Component Analysis using the correlation matrix [41]. The first principal component retains 63.59% of the variance, while second and third components represent 16.50% and 11.51% of the variance, respectively. The three principal components represent 91.60% of the total variance of the data but they do not provide a good representation to clearly classify the three gasses.

As previously described in [10], to construct our dataset, we placed the sensor array – a set of chemical sensors tagged by the manufacturer as TGS2600, TGS2602, TGS2610, and TGS2620 (four of each) – into a 60 ml volume polytetrafluoroethylene/stainless steel air-tight test chamber, a vapor flow cell into which the gaseous substances are injected at a constant flow and in a random order. The test chamber is attached in series to a vapor delivery system that provides the selected concentrations of the chemical substances by means of three digital mass flow controllers (Bronkhorst High-Tech B.V. [42]) and the calibrated gas cylinders (Air gas [43]). The entire measurement system setup is fully operated by a computerized environment and provides versatility for setting the concentrations with high accuracy and in a highly reproducible manner (see Fig. 3).

To generate the dataset, we adopted a measurement procedure consisting of the following three steps. First, in order to stabilize the sensors and measure the baseline of the sensor response, we circulated synthetic dry air (10% R.H. measured at 25 ± 1 °C) through the sensing chamber during 50 s. Second, we randomly added one of the analytes of interest to the carrier gas (in our case synthetic dry air) and made it circulate through the sensor chamber during 100 s. Finally, we re-circulated clean dry air for the subsequent 200 s to clear up the sensors and the test chamber to have the system ready for a new measurement. The measurements were performed at a constant flow rate of 200 ml/min (sccm).

To capture the sensors' responses and control their operating temperature (indirectly controlled by the sensor operating voltage), we adapted a PC platform fitted with the appropriate data acquisition and serial communication cards with a National Instruments data acquisition board controlled by a customized LabVIEW code [44]. The dynamic response of each sensor was recorded at a sample rate of 100 Hz while the gas sensor array was kept at a stable operating temperature (400 °C).[3]

[1] Certain commercial equipment, instruments or materials are identified in this report to specify adequately the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

[2] The complete acquired dataset is freely available at the UCI repository at http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations.

[3] We do not have access to the actual sensing surface temperature due to packaging, but a look-up table relating it to the voltage applied to the sensor's embedded heating element can be found upon request in Ref. [40]. Note that the effect of varying the sensors' operating temperature was investigated in a previous work, by following information theoretic optimization formalisms [45].