Data Analysis for Chemical Sensors

Santiago Marco

Signal and Information Processing for Sensing Systems Lab, Institute for Bioengineering of Catalonia
Universitat de Barcelona

Outline

- Introduction:
  - The research team and activities
- Introduction to Signal and Data Processing for Chemical Sensing
  - Motivation
  - Introduction to Pattern Recognition
  - Some open problems in e-nose data analysis
- Performance estimation and validation
- Further insight in basic building blocks
- Summary
The Team

- IP: Santiago Marco, PhD in Physics
- Antonio Pardo, PhD in Physics
- Agustín Gutierrez, PhD in Computer Science

- Post-doc
  - Juan Mª Jiménez Soto, Analytical Chemist

- PhD Students
  - Erola Pairo, MSc Computational Physics
  - Victor Pomareda, MSc Electronic Engineering
  - Ariadna Bartra, MSc Biomedical Engineering
  - L. Fernández, MSc Electronic Engineering
  - Ana Guamán, MSc Biomedical Engineering
  - Milad Avazbeigi, MSc Computer Science
  - Sergi Oller, MSc Computational Physics

Mission:

- Develop smart chemical sensor systems based on micro-nano technologies embedding advanced signal and data processing
**Artificial Olfaction**

- An e-nose is an instrument which combines (Gardner, S&B,B, 1994)
  - an array of chemical sensors with partial and overlapping specificities
  - a pattern-recognition system capable of processing the multivariate response across sensors

- Main technological drivers:
  - Decrease the analysis times (less 1 min is the target)
  - Decrease the size: Handheld versions
  - Decrease the cost: less than 1000 euros.

---

**Low Selectivity continuous patterns**

- MOX Sensing

---
Expertise: Signal and Data Processing

- Complex Chemical Measurements need additional signal & data processing for detection, identification and quantification.

Introduction to Data Analysis for Chemical Sensors

Optimization feedback

[ Gutierrez-Osuna, 1998 ]

MPCA scores for initial

[ Padilla, 2010 ]

MPCA scores for data corrected by OSC

Introduction to Data Analysis for Chemical Sensors
Introduction

- Some words for motivation
- Basics of pattern recognition for AO
- E-nose robustness
- Major problems in e-nose that can be addressed by data processing
  - Drift – Time stability
  - Sensitivity to environmental and (or) operational conditions
  - Background changes
  - Calibration transfer
- Performance estimation and validation

- Basic building blocks of a pattern recognition system
  - Signal pre-processing
  - Dimensionality reduction
  - Classifiers

Cites

"If we believe that the
Purpose of Computing is Insight, Not Numbers,
Then it follows that
the man (woman) who is to get the insight must understand the computing"

R.W. Hamming (1973)

(Inventor of the Hamming correcting codes
and proposer of the Hamming window in Spectral Analysis)
Cites II

- “It can be proven that most claimed research findings are false” (Ioannidis, 2005).

- “‘what the data say’ is often obscured by questionable answers to unanswerable questions (Cornfield, 1966)”

- “Left to our own devices, ...we are all too good at picking out non-existent patterns that happen to suit our purposes” (Efron and Tibshirani, 1993)

Introduction

- Signal processing and data analysis are of increasing importance in analytical chemistry and chemical sensing/detection.

- Modern Chemical Instrumentation offers enormous capabilities for signal recording and storage

- Then signal / data analysis and interpretation may become the bottleneck in the process.

- “We are drowning in information (data) and starving for knowledge”. Rutherford D. Roger (American Librarian)
**Applications that need Signal/Data Processing**

- **Problems we have addressed**
  - Quantify components in simple chemical mixtures
  - Identify chemical products by a chemical fingerprint
  - Monitor chemical / biochemical processes from volatile emission patterns
  - Analysis of biomedical fluids (sweat, urine, breath, etc)
  - Alarm when a chemical sensor is giving erroneous measurements
  - Correct erroneous readings from chemical sensors from redundant information
  - Make the instrument more robust by rejecting drift and cross-sensitivities to environmental parameters.
  - Locate odour sources with mobile vehicles

- **Higher-order Chemical Instrumentation**
  - Sensor arrays
  - IMS/DMA
  - GC/MS
  - Mid-IR Microspectrometers

---

**Components of a pattern recognition system**

- **A basic pattern classification system contains**
  - sensors (in our case, gas sensor array)
  - A preprocessing mechanism
  - A feature extraction mechanism
  - A predictive model
  - A calibration procedure to provide a set of examples (training set) already classified or described

---

Adapted from R. Gutierrez-Osuna, TAMU
Functional Diagram of Electronic Nose

Cond. Electron. DAC/ADC sys

Manifold

Pump

In 1

In 2

In N

Conductive Polymers sensor die

Exposure to 3 ppm Ammonia
Temperature Modulated MOX array

Exposure to increasing steps of acetoone: 20 to 100 ppm

Responses of CP and MOX chemosensors

Temperature modulation produces distinctive patterns for different analytes

Diversity of sensitivities for two analytes
Features and patterns

- **Pattern**
  - A composite of traits or features characteristic of an individual/sample
  - In classification tasks, a pattern is a pair of variables \(\{x, \omega\}\) where
    - \(x\) is a collection of observations or features (feature vector)
    - \(\omega\) is the concept behind the observation (label)

![Fruits and patterns example](Image)

[Gutierrez-Osuna, TAMU]

---

Features, patterns and classifiers

- **Feature**
  - The combination of \(d\) features is represented as a \(d\)-dimensional column vector called a **feature vector**
    - The \(d\)-dimensional space defined by the feature vector is called **feature space**
    - Objects are represented as points in feature space. This representation is called a **scatter plot**

![Feature vector, feature space, scatter plot](Image)

[Gutierrez-Osuna, TAMU]
**Features, patterns and classifiers**

- **What makes a “good” feature vector?**
  - The quality of a feature vector is related to its ability to discriminate examples from different classes
    - Examples from the same class should have similar feature values
    - Examples from different classes have different feature values

  ![“Good” features](image1) ![“Bad” features](image2)  

  [Gutierrez-Osuna, TAMU]

---

**Features, patterns and classifiers**

- **More feature properties**

  ![Linear separability](image3) ![Non-linear separability](image4) ![Multi-modal](image5) ![Highly correlated features](image6)

- **Classifiers**
  - The goal of a classifier is to partition feature space into class-labeled **decision regions**
  - Borders between decision regions are called **decision boundaries**
**Introduction to Data Processing: Classification**

- **Given:**
  - input/output \((x, t)\) data pairs, \((x\) continuous, \(t\) categorical)

- **Question:**
  - What's the best label \(t\) given \(x_{\text{new}}\)?
  - What's the probability of \(t\) given \(x_{\text{new}}\)?

---

**Introduction to Data Analysis: Regression**

- **Given:**
  - input/output \((x, t)\) data pairs
  - \((x, t\) continuous)
  - \((x\) will be in general a vector)

Generated by an underlying \(t = f(x) + \text{noise}\)

- **Question:**
  - How the function \(f(x)\) look like?
  - What’s the best value of \(t\) given \(x\)?
  - What’s the probability of \(t\) given \(x\)\? \(p(t/x)\)
Given some data points $z_i$ (dataset)
- We need a model that captures the important structure
- Typically the model is a probability distribution $p(z)$
- This is an Unsupervised Learning problem

Some open problems in e-nose data analysis

Can signal processing help to make enoses more robust?
**E-nose Robustness**

- **Robustness definition (World Health Organization, 2007):**
  - “Robustness (or Ruggedness) is the ability of the procedure to provide analytical results of acceptable accuracy and precision under a variety of conditions. The results from separate samples are influenced by changes in operational or environmental conditions”

- **Additional remarks**
  - When considering e-noses robustness has to be evaluated in the full set (hardware + software)

**Drift**

- **While instruments provide good initial results, these results rapidly deteriorate in time requiring frequent recalibration.**

M. Padilla, Chemolab, 2010
**Change in environmental/sampling conditions**

- Instruments are calibrated in a range of environmental and operational conditions.
- Instruments operate in an extended range of environmental conditions (temperature and humidity).
- Sample input flow shifts from calibration conditions (e.g. due to filters).

---

**Background/sample change**

Instrument is able to detect fungi growing in cardboard material. However the instrument is unable to detect the fungi in wallpaper.

Odours in a bioreactor change with the feeding material. Plant control models from one feeding material fail if the feeding material composition change.

---

M. Kuske, S&A.B, 2006

Gilles, Isoen, 2011
**Calibration transfer**

- An e-nose company has developed a successful application by extensively calibrating an e-nose instrument.
- The company wants to transfer the predictive model to a number of identical instruments.
- Direct application of the master model fails.

---

**Performance Estimation and Validation**

*Emphasis on binary classifiers*
Outline

- Discriminant functions
- Figures of merit from the confusion matrix:
  - Uncertainties due to finite sample effects
- Signal Detection Basics: The ROC curve

- Most common errors in binary classifiers!!
  - Ignoring the prevalence of the condition in the normal population
  - Neglecting cross-validation leading to overfitting.
  - Confusing p-value with predictive power.
  - Bias

---

Discriminant functions

- A convenient way to represent a pattern classifier is in terms of a family of discriminant functions \( g_i(x) \) with a simple MAX gate as the classification rule.

\[
\text{Class assignment} \\
\text{Select max} \\
\begin{align*}
g_1(x) & \quad g_2(x) \\
x_1 & \quad x_2 \\
& \quad x_3 \\
& \quad x_4
\end{align*}
\]

Assign \( x \) to class \( \omega_i \) if \( g_i(x) > \text{threshold} \)

- How do we choose the discriminant functions \( g_i(x) \) leads to different classification algorithms
- For probabilistic interpretation \( 0 \leq g_i(x) \leq 1 \) and \( \sum g_i(x) = 1 \)

[Gutierrez-Osuna, TAMU]
Evaluation of Binary Classifiers: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>True-negatives (TN)</td>
<td>False positives (FP)</td>
</tr>
<tr>
<td>Alarm</td>
<td>False negatives (FN)</td>
<td>True-positives (TP)</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{TP + TN}{TP + TN + FP + FN} \)

Sensitivity (Recall) = \( TP / (TP + FN) \)

Specificity = \( TN / (TN + FP) \)

Positive Predictive Power = \( TP / (TP + FP) \)

Negative Predictive Power = \( TN / (TN + FN) \)

Distribution of proportions (confidence limits)

- Among \( n \) samples, there are \( n_g \) correctly classified and \( n-n_g \) not.
- Proportion is assumed to follow a binomial distribution

\[
\hat{p} = \frac{n_g}{n} \\
\hat{q} = \frac{n - n_g}{n} = 1 - \hat{p} \\
\]

Condition: \( n\hat{p} = n_g > 5 \) \( n\hat{q} = n - n_g > 5 \)

Example 1: 120 samples, 110 correct

\[
\hat{p} = \frac{110}{120} = 0.916 \pm z_{\alpha/2} \cdot 0.025 \\
\hat{p}(\alpha = 0.05) = 0.916 \pm 1.96 \cdot 0.025 = 0.92 \pm 0.05 
\]

Method 1 provides 90% classification rate

Method 2 provides 93% classification rate

Surely enough the difference is not statistically significant

Gardner and Altman, British Medical J., 1989
**Distribution of proportions (confidence limits)**

- **ROC analysis** does not depend on selected at random.
- The ROC curve explores the trade-off for all possible values of the threshold.
- The ROC curve does not depend on the underlying pdfs.
- ROC analysis does not depend on the prevalence of the condition. Final threshold is decided by the clinical staff.

Area Under Curve (AUC) is an estimation of the probability that a member of the group condition is larger than a member of the group control, when both have been selected at random.
**Receiver Operating Characteristics**

- Effect of a larger Signal/Noise Ratio (better separation between Disease/Control)

---

**Main errors in the evaluation of binary classifiers**

- **Ignoring the prevalence of the condition**: Classifiers applied to very low prevalence conditions could provide large false positives. Bayesian approaches using priors are required.

- **"Overfitting"**: typically by failing to use independent/blind ‘test’ samples which are held back from model optimization and used only to test the robustness of prediction in the final phase of the study.

- **Confusing statistical significance with predictive accuracy**: A number of features (biomarkers) could be statistically different and still provide very low predictive accuracy.

- **Bias**, in which a feature is differentially distributed between the classes ‘case’ and ‘control’ but happens to be correlated with another uncontrolled variable that truly underlies the variance of the feature, such confounding variables from the individuals including smoking status, gender, diet, but also instrumental: different time, different location, different operator, different temperature/humidity, different instrument ....
Confusing p-values and predictive power

- Binary classifier with 286 features & 174 samples. Univariate test of individual features by p-value and by an independent test set (Kenny, 2005).

P-values and ROC

Bonferroni-correction
Sample size effects

- The larger the sample size, the smaller has to be the difference (relative to the standard deviation) to have enough statistical power.

“Small difference of no real interest can be statistically significant with large sample sizes, whereas clinically important effects maybe statistically non-significant only because the number of subjects studied was small” (Gardner and Altman, 1989).

Bias / Confounding factors

- “Bias is the most important threat to validity”, Ransohoff (2005).
- “All observational research should be presumed guilty of bias, until proven innocent”, Ransohoff (2005).

- Confounding factors are not equally present in the two groups and can be the underlying reason for the observed differences.

- Common confounding factors: Instrumental
  - Time of the measurement (all instruments shift in time, particularly enoses)
  - Sample handling differences (e.g. time from collection to analysis, temperature at collection)
  - Laboratory, Instrument, Operator

- Common confounding factors: Clinical
  - Gender, Age, Diet, Medication....
Hypothesis test dealing with confounding factors

- Use (whenever possible) hypothesis test that are able to deal with multiple factors (e.g. Multi-way ANOVA), Cox (1958).
- Be skeptical if the observed differences are statistically significant for potential confounding effects.

**Recommendation:**
- Consider before hand all the potential confounding factors.
- Block them during experimental design
- Always consider time as confounding factor when working with enoses

**Example of a bad experimental design:**
- Month 1: 10 Lung cancer subjects are tested for breath analysis
- Month 3: 10 control subjects are tested for breath analysis
- Unless further evidences are shown, it is impossible to attribute differences to time or to condition.

**Good experimental design:**
- Month 1: 5 cancer + 5 controls tested
- Month 3: 5 cancer + 5 controls tested

---

**GC/MS detection of prostatic cancer in urine headspace**

**Sample distribution:**

<table>
<thead>
<tr>
<th>Samples</th>
<th>Cancer</th>
<th>Control</th>
<th>Blank</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blank</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum (-Blanks)</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status distribution for different dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>14/05/09</td>
</tr>
<tr>
<td>11/06/09</td>
</tr>
<tr>
<td>06/07/09</td>
</tr>
<tr>
<td>27/11/09</td>
</tr>
<tr>
<td>19/07/11</td>
</tr>
<tr>
<td>31/01/12</td>
</tr>
<tr>
<td>09/02/12</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
</tr>
</tbody>
</table>

**Main issues:**
- Unbalanced sampling on several days → Status ~ Date confusion
- 2010 samples discarded due to different instrumentation settings
- Sample sequences: 1 Blank, 5 Control, 1 Blank, 5 Cancer → not randomized
- Experimental design in chemistry: A tutorial, R. Leardi ACA 2009
Validation techniques

Adapted from Lecture Notes by R. Gutierrez-Osuna,
Dept. Computer Science, TAMU

Motivation

- Validation techniques are motivated by two fundamental problems in data analysis: model selection and performance estimation
- Model selection / algorithm optimization
  - Almost invariably, all data processing techniques have one or more free parameters
    - The number of neighbors in a kNN classification rule
    - The network size, learning parameters and weights in MLPs
  - How do we select the "optimal" parameter(s) or model for a given classification problem?
- Performance estimation
**Complexity Control (Algorithmic selection)**

- Too simple model:
  - Low complexity

- Too complex model:
  - High complexity

- Large training errors
- Large test errors

- Zero training errors
- Large test errors
- Poor generalization

Duda, Hart, Stork, 2001

---

**Motivation: Complexity Control**

- Some optimal classifier exist

- The complexity of the model has to be controlled for good performance

Duda, Hart, Stork, 2001
Motivation: Complexity Control

- Regression example: polynomial fitting
- Question: What is the best polynomial order to fit the data?
  - A 10th order polynomial predicts perfectly the training data but fits also the noise, producing large errors for the prediction of new samples.

![Graph showing polynomial fitting](image)

Duda, Hart, Stork, Pattern Classification, 2001

Motivation: Complexity Control

- Large Samples/Dim ratio
- Small Samples/Dim ratio
The holdout method

- Split dataset into two groups
  - Training set: used to train the classifier
  - Test set: used to estimate the error rate of the trained classifier

- Fundamental trade-off
  - The larger the training set, the more accurate the classifier will be
  - The larger the test-set, the more accurate the error estimation

- However in hold-out
  - Large training set implies small test set or,
  - Large test set implies small training set

The holdout method has two basic drawbacks

- In problems where we have a sparse dataset we may not be able to afford the “luxury” of setting aside a portion of the dataset for testing
- Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an “unfortunate” split

The limitations of the holdout can be overcome with a family of resampling methods at the expense of more computations

- Cross Validation
  - Random Subsampling
  - K-Fold Cross-Validation
  - Leave-one-out Cross-Validation
- Bootstrap
**K-Fold Cross-validation**

- Create a K-fold partition of the dataset
  - For each of K experiments, use K-1 folds for training and the remaining one for testing

![Diagram of K-Fold Cross-validation](image)

- The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing
- As before, the true error is estimated as the average error rate
  \[
  E = \frac{1}{K} \sum_{i=1}^{K} E_i
  \]

**Random Subsampling**

- Random Subsampling performs K data splits of the dataset
  - Each split randomly selects a (fixed) no. examples without replacement
  - For each data split we retrain the classifier from scratch with the training examples and estimate \( E_i \) with the test examples

![Diagram of Random Subsampling](image)

- The true error estimate is obtained as the average of the separate estimates \( E_i \)
  - This estimate is significantly better than the holdout estimate
  \[
  E = \frac{1}{K} \sum_{i=1}^{K} E_i
  \]
**Leave-one-out Cross Validation**

- Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples
  - For a dataset with N examples, perform N experiments
  - For each experiment use N-1 examples for training and the remaining example for testing

  ![Diagram of Leave-one-out Cross Validation](image)

- As usual, the true error is estimated as the average error rate on test examples
  
  $$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

**Bootstrap**

- The bootstrap is a resampling technique with replacement
  - From a dataset with N examples
    - Randomly select (with replacement) N examples and use this set for training
    - The remaining examples that were not selected for training are used for testing
      - This value is likely to change from fold to fold
      - Repeat this process for a specified number of folds (K)
  - As before, the true error is estimated as the average error rate on test examples
Three-way data splits

- If model selection and true error estimates are to be computed simultaneously, the data needs to be divided into three disjoint sets
  - **Training set**: a set of examples used for learning: to fit the parameters of the classifier
    - In the MLP case, we would use the training set to find the “optimal” weights with the back-prop rule
  - **Validation set**: a set of examples used to tune the model complexity of classifier
    - In the MLP case, we would use the validation set to find the “optimal” number of hidden units or determine a stopping point for the back propagation algorithm
  - **Test set**: a set of examples used only to assess the performance of a fully-trained classifier
    - In the MLP case, we would use the test to estimate the error rate after we have chosen the final model (MLP size and actual weights)
    - After assessing the final model with the test set, YOU MUST NOT further tune the model

Three-way data splits

- Why separate test and validation sets?
  - The error rate estimate of the final model on validation data will be biased (smaller than the true error rate) since the validation set is used to select the final model
  - After assessing the final model with the test set, YOU MUST NOT tune the model any further

- Procedure outline

1. Divide the available data into training, validation and test set
2. Select architecture and training parameters
3. Train the model using the training set
4. Evaluate the model using the validation set
5. Repeat steps 2 through 4 using different architectures and training parameters
6. Select the best model and train it using data from the training and validation set
7. Assess this final model using the test set
Three-way data splits

In the presence of REPLICAS from the same individual, they have to be treated as an INDIVISIBLE BLOCK when doing the partition in training, data. Random sampling produces always OVEROPTIMISTIC results since neglects instrumental shift.

In the presence of CONFOUNDING FACTORS, perform “LEAVE OUT ONE FULL FACTOR” level to test the prediction capabilities. A robust system will preserve prediction capabilities for confounding factor levels never seen. If this not the case, suspect confounding factors play a role.

Some recommendations on validation

- If possible use ALWAYS the three-way split

- Independent samples have to be always in the FUTURE of the training data.
  - Random sampling produces always OVEROPTIMISTIC results since neglects instrumental shift.

- In the presence of REPLICAS from the same individual, they have to be treated as an INDIVISIBLE BLOCK when doing the partition in training, validation and independent test sets.

- Use Bootstrap instead of Random Subsampling / K-fold
Summary

- E-nose Research has been accused often of unreliability and lack of reproducibility.
- Most lab findings have never reached the industry or the clinic.
- Extreme methodological care is needed when dealing with e-nose experiences and data.
- Strict validation is a MUST!!!
- Final confirmation is only given by the application of the technology in the real scenario.
- Data Analysis should not be concerned today with finding the best predictor since usually the best predictor is not reliable.
- Robustness improvement is a major goal in data analysis for AO systems. This includes rejecting drift, environmental perturbations, sensor replacement shifts, etc.