A FEATURE SELETION STRATEGY FOR THE DEVELOPMENT OF A NEW DRUG SENSING SYSTEM

Ulrici A[1,3] Calderisi M[2,3], Seeber R[2,3] Uotila J[4], Secchi A[5], Fiorello AM[5], Dispenza M[5]

1 Dipartimento di Scienze Agrarie e degli Alimenti, Università di Modena e Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy (alessandro.ulrici@unimore.it)

2 Dipartimento di Chimica, Università di Modena e Reggio Emilia, Via Campi 183, 41125 Modena, Italy

3 Consorzio INSTM, Via G. Giusti 9, 50121 Firenze

4 Gasera Ltd., Tykistökatu 4, 20520 Turku, Finland

5 Selex-SI, Via Tiburtina, Km 12,400, 00131 Rome, Italy

Abstract

In order to efficiently detect four drug precursor molecules in presence of interfering species and background air, using a EC-QCLPAS sensor operating in the mid-infrared region, a complex strategy of spectral response simulation has been developed. In this context, spectra of gases from literature databases have been collected, denoised by means of the Wavelet Transform and mixed together according to a concentration matrix, which was specifically designed to represent a comprehensive combination of possible realistic cases. To scale database spectra to the appropriate concentration levels, an *ad-hoc* algorithm based on a sigmoidal transfer function has been used. In this way the baseline shape and intensity is preserved. Afterwards, a preliminary wavelength selection has been carried out to exclude noisy regions. The optimal range has finally been defined by maximizing the classification efficiency for all the target gases by means of Partial Least Squares-Discriminant Analysis.

1. Introduction

In the context of EU FP7 project CUSTOM (Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques) a new sensor system for the detection of drug precursors in gaseous samples is being developed. It makes use of two integrated systems: an External Cavity-Quantum Cascade Laser Photo Acoustic Sensor (EC-QCLPAS) and a Led Induced Fluorescence Optochip (FLUO). In particular, as for the EC-QCLPAS system, the optimal wavenumber values in a

200cm^{-1} range must be defined in the mid-infrared region, in order to achieve optimal detection of the drug precursor (target) molecules in presence of interfering species (pollutants) and at variable composition of the air components.

To this aim, using a wide set of EC-QCLPAS simulated spectra obtained by proper elaboration of FT-IR literature spectra, a suitable complex strategy has been developed. It essentially involves preprocessing of spectral data and experimental design techniques for the estimation of the concentrations of the various gaseous species involved in the study, together with proper feature selection algorithms for the definition of optimal spectral range. Using this approach, even in absence of true experimental spectra, we were able to identify the optimal spectral region for the identification of 4 target species in presence of 20 possible pollutants and 9 air components. The choice was the guide for the fabrication of the proper laser source.

2. Data base build up and algorithms implementation

The first step of the database build up consisted in denoising the database spectra using a Wavelet Transform (WT) [1] based algorithm developed ad-hoc, in order to consider only the relevant spectral information and to filter off the stochastic variation, i.e. the noise associated with the database spectra. Then, an algorithm was developed to import spectral data with different file formats, wavenumber ranges and resolutions, sampled at either constant or varying rates, in order to obtain uniform datasets at constant concentration of 1 ppm in the 1000-2500 cm^{-1} spectral range.

In order to build a representative concentration matrix of the mixtures to simulate, different combinations of experimental design techniques were implemented. The mixtures of targets and pollutants were planned using different combinations of Full Factorial Designs; the overall number of considered factors and levels is reported for target molecules and pollutants in Tables 1 and 2, respectively, together with the corresponding number of simulated mixtures. Moreover, for the air components 1000 mixtures were generated using, for each component, random values taken from a lognormal distribution, whose parameters were derived from the mean and maximum observed gas concentrations.

The noise structure of the ECQCL-PAS signals was estimated by means of WT, using two sample signals measured with a prototype instrument and applied to the simulated spectra of gas mixtures.

A sigmoidal transfer function was used in order to multiply correctly spectra by concentrations, preserving the background shape and intensity and operating only on the spectral regions where the absorption bands are located.

The final matrix is composed by 499 mixtures of targets, 750 mixtures of pollutants (sampled randomly from the whole pollutants concentration matrix) and 1000 air components mixtures. This procedure was iterated twice, in order to build a final data matrix with 2000 gas mixture spectra representing a wide variety of possible combinations.

Table 1. Target mixtures: parameters of the Full Factorial Design

N. of factors (targets)	Concentration levels	Mixtures
1	40	160
2	5	150
3	3	108
4	3	81

Table 2. Pollutants mixtures (20 gases - mixtures up to 3 gases)

N. of factors(pollutants)	Concentration levels	Mixtures
1	3	260
2	4	3040
3	3	30780

Following previous literature suggestions [2], a thoughtful pre-selection of the spectral range was made using a WT-based procedure (named SMARTGRID), which allows the exclusion of noisy spectral regions due to small interfering molecules leading to sharp absorption peaks (Fig.1).

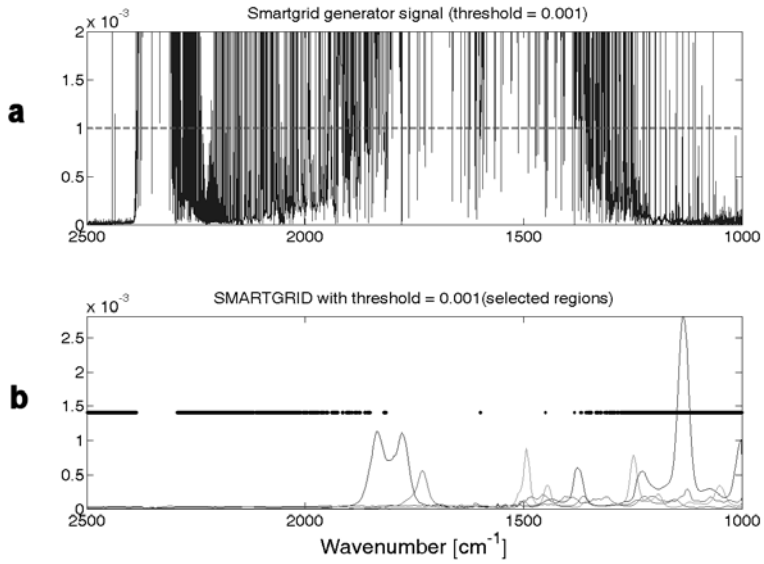

Figure 1. The SMARTGRID procedure allows us to exclude noisy regions. (a) Estimate of the spectral regions containing the sharp peaks and definition of a threshold value. (b) Selected regions and target spectra.

The spectra of the final mixtures at the pre-selected wavenumbers were used to finally choose the optimal range, by maximizing the Classification Efficiency estimated by Partial Least Squares-Discriminant Analysis [3]. Data were pre-

processed using autoscaling (AUTO) and Pareto scaling (PARETOMNCN). Models were validated using both Cross-Validation (CV) and an external Test Set (TS). The overall results are satisfactory (Figure 2) and generally converge as for the wavenumber range selected for the different targets. The optimal range was selected on the basis of the Global Classification Efficiency (GCE), which expresses the mean of the Classification Efficiency for each target, which is, in turn, the geometric mean of sensitivity and specificity. In particular, the window centered at 1182 cm^{-1} leads to the best results.

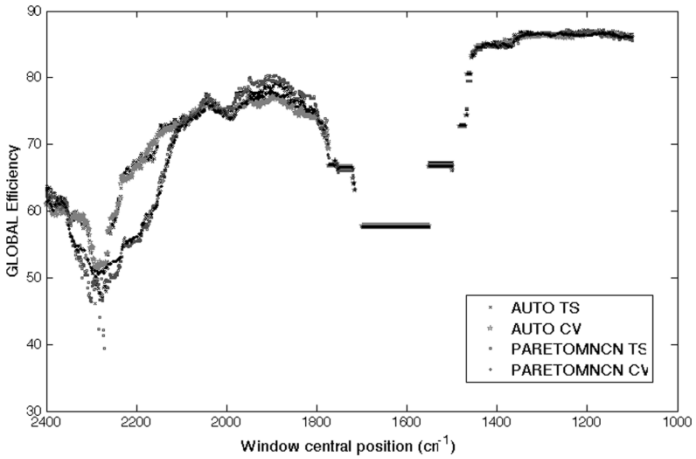

Figure 2. GCE values for the 4 target molecules as a function of the wavenumber corresponding to the central position of the 200 cm^{-1} moving window.

3. Conclusions

The optimal range chosen leads to the best results in crossvalidation. This window spans from 1281.5 to 1082.5 cm-1 (spectral resolution = 0.5 cm-1) and comprises 364 variables (out of a total of 399 variables, some being pre-deleted by application of the SMARTGRID, due to the presence of sharp peaks of interfering molecules).

References

1. Walczak B (2000) Wavelets in Chemistry. Elsevier, Amsterdam
2. Dunayevskiy I et al. (2007) High-sensitivity detection of triacetone triperoxide (TATP) and its precursor acetone. Appl. Optics, 46 (25), 6397-6404. http://dx.doi.org/10.1364/AO.46.006397
3. Pigani L et al. (2011) PEDOT modified electrodes in amperometric sensing for analysis of red wine samples. Food Chem. 129 (1), 226-233. doi: 10.1016/j.foodchem.2011.04.046