FAST VARIABLE SELECTION FOR MASS SPECTROMETRY ELECTRONIC NOSE APPLICATIONS

RAPIDA SELECCION DE VARIABLES PARA APLICACIONES DE NARICES ELECTRONICAS BASADAS EN ESPECTROMETRIA DE MASAS

PhD. Oscar Eduardo Gualdrón Guerrero, PhD. Ivaldo Torres Chávez

Universidad de Pamplona
Ciudadela Universitaria. Pamplona, Norte de Santander, Colombia.
Tel.: 57-7-5685303 Ext. 144, Fax: 57-7-5685303.
E-mail: {oscar.gualdron, ivaldo.torres}@unipamplona.edu.co

Abstract: High dimensionality is inherent to MS-based electronic nose applications where hundreds of variables per measurement (m/z fragments) — a significant number of them being highly correlated or noisy — are available. Feature selection is, therefore, an unavoidable pre-processing step if robust and parsimonious pattern classification models are to be developed. In this article, a new strategy for feature selection has been introduced and its good performance demonstrated using two MS e-nose databases. The feature selection is conducted in three steps. The first two steps are aimed at removing noisy, non-informative and highly collinear features (i.e., redundant), respectively. These two steps are computationally inexpensive and allow for dramatically reducing the number of variables (near 80% of initially available features are eliminated after the second step). The third step makes use of a stochastic variable selection method (simulated annealing) to further reduce the number of variables. For example, applying the method to an Iberian ham database has resulted in the number of features being reduced from 209 down to 14. Using the surviving m/z fragments, a fuzzy ARTMAP classifier was able to sort ham samples according to producer and quality (11-category classification) with a 97.24% success rate. The whole feature selection process runs in a few minutes in a Pentium IV PC platform.

Resumen: Una alta dimensionalidad es inherente en aplicaciones de narices electrónicas basadas en MS, donde se pueden encontrar cientos de variables por medida, un número significativo de ellas proporcionan ruido o una alta correlación entre ellas. En este articulo, una nueva estrategia de selección de variables es desarrollada con buenos resultados usando dos bases de datos de narices electrónicas basadas en MS. El proceso se realizo en tres pasos. En los dos primeros pasos el objetivo es eliminar ruido e información altamente colineal (redundancia), respectivamente. El tercer paso se utiliza el método de selección estocástico (*simulated annealing*) para reducir significativamente el número de variables. El proceso de selección total se ejecuto en pocos minutos en una plataforma Pentium IV.

Keywords : Feature selection, Mass spectrometry, Electronic nose, Simulated annealing, Neural networks, Iberian ham.

1. INTRODUCCIÓN

In the last few years, with the use of mass spectrometry (MS), a new branch within electronic nose research has developed and gained importance. Unlike in classical gas chromatography/mass spectrometry systems (GC/MS), in MS based electronic noses the sample delivery unit directly injects complex volatile mixtures (such as the ones generated in the headspace of foodstuffs or beverages) into a ionisation chamber, without a previous separation step (provided by GC). This results in very complex ionisation patterns that are recorded at the detector side. These ionisation patterns are then processed by pattern recognition engines to perform tasks associated to electronic nose systems such as classification, recognition and, to a limited extent, quantification. [1-4]. Although the detector of a MS gives a signal that depends linearly, at least within a range, on the abundance of any given mass to charge ratio, the complexity of the ionisation patterns that are analysed in some particular applications justifies the need of using non-linear pattern recognition methods, including neural networks [4].

In MS-based electronic noses, every mass to charge ratio (m/z) in the mass spectra can be thought of as a sensor. In accordance with the electronic nose philosophy, a priori knowledge of the components present in the headspace being analysed should not been required. This is why most applications developed using this approach consider spectra consisting of a wide range of m/z ratios (e.g. from m/z 35 to m/z 300), which cover the fragmentation of volatile molecules. This implies that over two hundred features are going to be available for the pattern recognition analysis. Therefore, it is not uncommon that the number of features exceeds the number of measurements available to train the pattern recognition methods and this is a dangerous situation because there is a high risk of overfitting [5]. Actually, a significant number of sensors (i.e., m/z ratios) can be irrelevant (i.e., noisy) for the application considered, while other sensors can show highly correlated responses. A step of dimensionality reduction seems, therefore, imperative prior to attempt the building of pattern recognition methods.

Different strategies have been reported for the reduction of dimensionality. These basically consist of either choosing directly among the variables available [6,7] (e.g., m/z ratios) or to compute new variables called factors (e.g., by performing a principal component analysis or a linear discriminant analysis, etc.) and selecting among the factors [8]. Selecting from the full spectrum of mass to charge ratios is challenging because there is considerable overlapping among the spectra and distinctive features can be almost imperceptible. Furthermore, spectra are affected by noise. However, methods based on the selection of m/z ratios are interesting because the variables chosen carry relevant chemical information. Therefore, these methods are expected to be robust toward the experimental conditions of each specific application. Unlike in m/z selection, factor selection uses the full spectrum (e.g. including noisy or redundant m/z ratios) to compute the factors before selecting from among them. The selection of an optimal subset of factors is not necessarily straightforward because the magnitude of an eigenvalue is not always a measure of its significance for the calibration [9]. Furthermore, unlike m/z ratios, factors have no chemical meaning.

Once determined that selecting among m/z ratios is the more interesting approach, it should be pointed out that an exhaustive search is out of question, given the high number of variables considered for selection. Several methods that avoid being exhaustive, the so-called greedy methods, have been reported as useful. These include deterministic methods such as branch and bound, sequential forward selection, sequential backward selection or stepwise selection and, stochastic methods such as genetic algorithms or simulated annealing [10-19]. Deterministic methods can make a fair selection with relatively few operations but are prone to get trapped in a local optimum of the search space. On the other hand, stochastic methods such as genetic algorithms or simulated annealing are more likely to find a global optimum at the cost of lengthy computation. For example, a genetic algorithm for variable selection running in a Pentium IV PC platform can take as long as several days to converge to a good solution when the number of variables for selection is above two hundred. Therefore, applying a stochastic method to select among the features found in MS-based electronic nose applications can easily turn to be unpractical.

In this paper we introduce a new method for an effective feature selection especially suitable for applications where the dimension of feature space is high, a significant degree of correlation exists between features and some of them are affected by

noise, such as in MS electronic nose applications. The method is efficient in the sense that after the selection process, only those features that are important for the application considered are retained to build the pattern recognition models and all the process is conducted at a very low computational cost. The usefulness of the new feature selection method is assessed using two different MS-based electronic nose databases.

2. EXPERIMENTAL

2.1 Solvent Database

The first database consisted of measurements taken from samples with a well-characterised headspace. The samples were 4 different solutions of pure ethanol containing added impurities (trichloroethylene, 1-butanol, ethylbenzene and toluene). The exact composition of the different samples is shown in Table 1. Six aliquots of 10 ml of each original sample were placed into 20-ml glass vials and sealed hermetically with silicone septa and caps (i.e. 24 vials in total). Sampling based on solid-phase micro extraction (SPME) was performed with a 75-µm Carboxen/PDMS fiber purchased from Supelco (Supelco Park, Bellefonte, PA). Prior to any extraction, the fibre was conditioned following the manufacturer's recommendations. In each measurement, the fibre was pushed out of its stainless steel housing and exposed to the sample headspace for 20 min at room temperature. The SPME holder assembly scale was adjusted to 3.0 scale units to ensure that the fibre was positioned in the headspace above the sample in exactly the same way from run to run.

Table 1: Composition of the different samples in database 1. Quantities are expressed in % dissolved in ethanol.

	Compounds			
Sample #	TCE	1-B	EB	TOL
S1	1	1	-	1
S2	1	1	1	1
S3	1	1	1	-
S4	1	-	1	1

A Shimadzu QP 5000 GC/MS (Shimadzu Corp., Tokyo, Japan) was used to implement a MS-based electronic nose. The instrument was equipped with a capillary column (Supelcowax, 30m × 0.25 mm i.d., × 0.25 mm coating thickness). The volatile compounds trapped on the SPME fibre were subsequently desorbed for 3 min at 280ºC into the glass-lined injection port of the GC, actuated in the splitless mode. The carrier gas was helium 99.995% set to 1.0 ml/min. The temperature of the GC oven and of the GC/MS interface was held constant at 250ºC so chromatographic separation was avoided. For any given measurement each unresolved peak was integrated and the resulting averaged mass spectrum gave a fingerprint that was characteristic of the headspace of the sample under analysis. Mass spectra were recorded at a rate of 2 scans/s over m/z ratios that ranged between 40 to 150 amu, operating the MS in the electron impact (EI) mode (70 eV). Further data analysis was performed on the relative mass spectra (i.e., normalised by the amplitude of the highest peak).

2.2 Iberian ham database

Eleven types of Spanish Iberian dry-cured hams were analysed. Samples were obtained directly from five producers and they differed in the type of food the pigs fed on during their fattening period (i.e., either acorn or fodder) and in their quality (type of pigs). Table 2 gives more details on the hams used.

Table 2: The 11 types of Spanish Iberian dry-cured hams analysed. The hams differ in producer, type of pigs and pigs' feeding.

Ham brand	Short name	# Ham types	Pig feeding on
Extremadura	EX	4	acorn
Guijuelo #1	G1	1	acorn
Huelva	HU	1	acorn
Guijuelo #2	G2	2	fodder
Guijuelo #3	G3	3	fodder

Samples were prepared as follows: three grams of ham (taken from biceps femoris) were crushed and introduced in 10 ml glass vials, which were then sealed with a septum and an aluminium cap. For each type of ham, 10 samples were prepared (exception made of one type from Extremadura with nine only). This gave a total of 109 ham samples to be analysed. Sampling was based on static headspace. A headspace autosampler Agilent 7694 was used. Oven, loop and transfer line temperatures were set to 90, 100 and 110ºC, respectively. The times for vial equilibration, pressurisation, loop filling, loop equilibration and injection were 30, 0.4, 0.15, 0.2 and 1 minutes, respectively. Reproducible headspace samples

were injected into the injection port of a Hewlett-Packard 6890 series II gas chromatograph coupled to a mass selective detector (Hewlett-Packard HP 5973; Wilmington, DE, USA). The injection port was used in splitless mode and maintained at 280ºC. The system was equipped with a HP 19091J-215 (50m × 0.32mm id, film thickness 1.05 µm) column, kept at 200ºC in isothermal conditions. In this way, chromatographic separation was avoided and the column merely acted as a transfer line delivering volatiles to the mass detector. The column flow rate was set to 1.5 ml/min. Volatile compounds were co-eluted into the mass spectrometer, where mass spectra were obtained using an electronic impact mass selective detector at 70 eV, a multiplier voltage of 2706 V, and collecting data at a rate of 1 scan s^{-1} over the m/z range 45–250 amu.

3. FEATURE SELECTION

The feature selection introduced here consists of three steps that are run consecutively. The first step helps detecting and removing non-informative, noisy features and is conducted in a supervised way. The second step is aimed at detecting collinearity between features in an unsupervised way. As a result, highly collinear features can be removed. Finally, in the third step, a greedy search method (e.g. a stochastic one) is applied to the reduced feature set, which results from applying the first two steps. With this approach, the whole variable selection process is time efficient since the first two steps are able to dramatically reduce the number of features at very low computational cost.

3.1 Removal of non-informative and noisy features

In electronic nose applications, the outcome sought after the system has been trained is an automated recognition or classification of new unknown samples. During the training phase, the pattern recognition ability of the system is built by using calibration samples. In the first step of feature selection, a criterion was used to rate the discrimination ability of each feature (i.e. m/z ratio). Measurements used for training were grouped in categories (e.g. measurements of the same type of ham were grouped in a category, etc.). For each m/z ratio, intra-category and inter-category variances were computed. Intra-category variance was defined as the variance of an m/z ratio considered within a given category of measurements. Therefore, the intra-category variance of the j-th m/z ratio, was defined as:

$$s^2_{\text{intra},j} = \frac{\sum_{i=1}^{n} \left(m/z_{ji} - m_j\right)^2}{n-1} \qquad (1)$$

Where n is the number of measurements within the category, m/z_{ji} is the value of mass to charge ratio j for measurement i and μ_j is the mean of mass to charge ratio j over the measurements within the category.

In a similar way, for every mass to charge ratio, a inter-category variance was defined as the variance within the category means. Therefore, the inter-category variance was defined as:

$$s^2_{\text{be},j} = \frac{\sum_{i=1}^{d} \left(m_{ji} - \overline{m}_j\right)^2}{d-1} \qquad (2)$$

Where μ_{ji} is the mean of mass to charge ratio j over the measurements within group i, d is the number of different categories and $\overline{m}_j$ is the mean over the μ_{ji}.

The discrimination ability of the j-th m/z ratio was defined as follows:

$$DA_j = \frac{s^2_{\text{be},j}}{s^2_{\text{intra},j}} \qquad (3)$$

The higher the discrimination ability for a given m/z ratio is, the more important is this m/z ratio to correctly discriminate between the categories. In other words, noisy or non-informative mass to charge ratios will have associated low discrimination abilities. Therefore, a set of m/z ratios, which comprises those that have the higher figure of merit, is selected for further analysis. This method would be equivalent to compute Fisher's linear discriminant if the number of categories to sort measurements within was $d = 2$. This process is univariate and there is a risk of eliminating those synergetic variables that have low discrimination ability when considered individually. To minimise this problem the process described by equations 1, 2 and 3 is repeated considering all the possible combinations between two m/z ratios. Figure 1 illustrates this process. As a result, a new list of figures of merit, $DA_{i,j}$, i.e., the discrimination ability when m/z ratios i and j are used simultaneously, is obtained. This allows for re-selecting variables that had been removed previously, if a synergistic effect is revealed.

However, it is important to notice that this method does no prevent redundant features (i.e., highly collinear) from being selected. This will be the task of the second step of feature selection.

3.2 Detection and removal of redundant features

Let be R the calibration matrix resulting from the first step of feature selection. R is a ($n \times p$) matrix. Its number of columns, p, corresponds to the number of features selected in the first step and, its number of rows, n, corresponds to the number of measurements within the calibration set. If R^t denotes the transpose of R:

$$R^t = \begin{bmatrix} m/z_{1,m1} & m/z_{1,m2} & \cdots & m/z_{1,mp} \\ m/z_{2,m1} & m/z_{2,m2} & \cdots & m/z_{2,mp} \\ \vdots & \vdots & \cdots & \vdots \\ m/z_{n,m1} & m/z_{n,m2} & \cdots & m/z_{n,mp} \end{bmatrix} \quad (4)$$

Where $m/z_{i,mj}$ corresponds to the intensity of the i-th mass to charge ratio for measurement j. For any mass to charge ratio, a unity-norm response vector can be defined as follows:

$$m/z_i = \left(\frac{m/z_{i,m1}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}}, \frac{m/z_{i,m2}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}}, \cdots, \frac{m/z_{i,mp}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}} \right) \text{ for } i = 1 \text{ to } n. \quad (5)$$

Equation 5 shows the unity-norm response vector for the i-th mass to charge ratio. Now, the degree of collinearity existing in the calibration set between two different mass to charge ratios can be assessed by computing the scalar product of their unity-norm response vectors as shown below:

$$P_{i,k} = \sum_{q=1}^{p} \left(\frac{m/z_{i,mq}}{\sqrt{\sum_{j=1}^{p} m/z_{i,mj}^2}} \times \frac{m/z_{k,mq}}{\sqrt{\sum_{j=1}^{p} m/z_{k,mj}^2}} \right) \quad (6)$$

$P_{i,k}$ is the scalar product between the normalised response vectors associated to features i and k. $P_{i,k}$ ranges between 0 and 1. The closer to unity $P_{i,k}$ is, the higher the collinearity between masse to charge ratios i and k is. Since n is the number of features, the collinearity of which is to be checked, the number of scalar products to be computed is $\sum_{i=1}^{n-1}(n-i)$.

After these scalar products have been obtained, a collinearity threshold is set and used to determine which features are redundant and should be removed. This second step of variable selection is non-supervised since, unlike in the previous step, there is no need to classify training samples according to their category.

After the removal of noisy, irrelevant and redundant features, the set of surviving features is ready for the last step of feature selection.

3.3 Stochastic feature selection

Stochastic methods such as genetic algorithms (GA) or simulated annealing (SA) are more likely to find a global optimum in the optimisation problem. These methods represent a trade off between the simple sequential methods (prone to get trapped in a local optimum) and the burden of exponential methods [15-17]. Genetic algorithms and simulated annealing solve the optimisation problem by exploring all regions of the potential solutions. Because explored points in a solution space are chosen by stochastic rather than deterministic rules, stochastic methods do not need to make assumptions about the characteristics of the problem to be solved and, therefore, apply generally. In the particular case of feature selection, these methods explore different subsets of the original set of features. Both GA and SA make use of a cost function, which in the case reported here, is an estimate of the prediction error of a neural network classifier (e.g. fuzzy ARTMAP) computed using the training measurements. This cost function is used to rank the fitness of solutions (i.e., combinations of features) during the process of stochastic feature selection. Since in most MS-based electronic nose applications a high number of variables are highly collinear or non-informative, about 80% of the original variables are eliminated by the first to steps. Therefore, the last step is aimed at fine tuning the selection process. Although stochastic feature selection methods are time- consuming, run to select among a reduced set of features that result from the two previous steps is fast.

4. RESULTS AND DISCUSSION

4.1 Analysis of the solvent database

A priori, the main challenge to correctly identify these compounds is due to the presence of ethylbenzene and toluene in the mixtures as these two species show some similarities between their mass spectra fragmentation pattern. Table 3 shows

which are the most intense fragments found in the mass spectra of the different compounds used. This database is a good benchmark for the feature selection method introduced here, since looking at table 3, one could select a set of mass to charge ratios to discriminate the different mixtures.

Table 3: Most intense m/z fragments found in the mass spectra of the different compounds used in the solvent database. Fragments appear sorted by decreasing intensity.

Compound	10 more intense m/z fragments
Tricloroetilene	132, 130, 95, 97, 60, 134, 47, 62, 59, 94
1-Butanol	56, 41, 43, 42, 55, 45, 40, 57, 44, 53
Ethylbenzene	91, 106, 51, 65, 77, 78, 50, 92, 52, 63
Toluene	91, 92, 65, 51, 63, 45, 50, 46, 62, 89

Therefore, the main objective sought with this database is to assess whether the 3-step feature selection was able to correctly determine, out of the 111 features available, the few ones that would enable a classifier to correctly identify the mixtures.

Before performing variable selection, a fuzzy ARTMAP was trained and validated using the leave-one-out cross validation approach. A description of the fuzzy ARTMAP algorithm can be found elsewhere [22]. The classifier made use of 111 inputs and the number of categories was set to 4 since this was the number of different mixtures analysed. The success rate in classification was 95.83%, which corresponded to one sample S2 being identified as S4 (see table 1).

The process of feature selection was conducted as follows. The 24 measurements available were split in 6 different feature selection datasets, which contained 20 measurements (5 replicates per solvent mixture) and their corresponding 6 validation datasets, which contained the remaining measurements (i.e. one measurement per solvent mixture not used in the corresponding feature selection dataset). Then the process of variable selection was performed 6 times on each feature selection dataset. The first step of feature selection was applied to eliminate noisy and irrelevant features. By setting to 0.5 the threshold value of the discrimination ability (both univariate and multivariate methods), between 31 and 35 out of 111 features were initially selected, depending on the feature selection dataset used. The second step was then applied to eliminate collinear variables. By setting to 0.15 the values of the collinearity threshold, between 17 and 20 features were retained. Computing the first two steps required about 6 minutes in a Pentium 4 PC platform. Finally a simulated annealing feature selection was run to select among the remaining features [23]. The SA algorithms were run for 50 different annealing temperatures and the number of iterations per temperature was set to 17. More details on the simulated annealing algorithm used can be found in the annex. In the end, only 3 features were selected (no matter the feature selection dataset used). These were the m/z ratios 46, 56 and 106. Using these three features as inputs, 6 fuzzy ARTMAP classifiers were trained employing the 6 feature selection datasets, and their performance in classification estimated using the corresponding validation datasets. The success rate in solvent mixture classification, estimated over the 6 training/validation sets was 100%. Figure 2 shows a block diagram of the feature selection and validation processes. It is important to keep in mind that for every fuzzy ARTMAP classifier, the validation implies using measurements that have not participated in the feature selection process and are, therefore, new. Considering table 3, it can be derived why the method has selected these specific features:

- m/z=46, which is the most relevant mass to charge ratio for Toluene not found for the other compounds in the solvent mixture.
- m/z=56, which is the most relevant mass to charge ratio for 1-Butanol.
- m/z=106, the second more relevant mass to charge ratio for Ethylbenzene.

It is important to notice that the different feature selection processes have disregarded using m/z=91. This is the most frequent ion for ethylbenzene and toluene, which would not help discriminating between these two compounds. Finally, no mass to charge ratio that is characteristic of trichloroethylene has been selected. This is correct because trichloroethylene is present in all the different samples to be discriminated and, therefore, no information about this compound is needed for a good discrimination among the samples analysed (see table 3). These results show that the three-step feature selection process introduced here is able to find the essential information needed to solve the discrimination problem considered. The whole process of feature selection took 15 minutes to complete in a Pentium 4 PC platform.

4.2 Analysis of the Iberian ham database

Initially an 11-category classification was attempted using a fuzzy ARTMAP classifier

without a previous step of feature selection. Because in this database the number of measurements available was higher, a different method of cross validation was employed. A 5-fold validation was implemented, which consists of defining 5 training and validation datasets. A training dataset comprised 8 replicate measurements (out of the 10 available) per type of ham (i.e., 87 measurements in total, since one type of ham had 9 replicate measurements instead of 10). The corresponding validation dataset comprised the 2 measurements per ham sample that had been left out (i.e. 22 measurements). Therefore, the fuzzy ARTMAP classifier was trained and validated 5 times using the 5 training and validation sets and the success rate in ham classification was averaged over the 5 tests. The success rates for the 5 folds were 63.63%, 95.45%, 100%, 100% and 81.81%, which gave an overall classification success rate of 88.18% (the standard deviation was 15.61%). This corresponds, in average, to 13 samples out of 110 being misclassified. Confusions occur between samples belonging to different producers and different quality hams within a producer.

The process of feature selection was performed using the 5 training and validation sets described above. For every pair of training and validation sets, feature selection was conducted on the training set, then a fuzzy ARTMAP classifier was trained using the features selected and, finally, its success rate in ham classification was estimated using measurements in the validation dataset. The first and second steps of feature selection were applied to remove noisy, irrelevant and redundant variables in every training and validation fold. The threshold values used were 0.5 and 0.8, respectively. This resulted in 42 features being selected in average.

Then, the third step of feature selection, which consisted in selecting among the surviving features using a simulated annealing procedure, was performed. As previously, the process was conducted independently for the 5 folds available. The SA algorithms were run for 50 different annealing temperatures and the number of iterations per temperature was set to 40. Fuzzy ARTMAP classifiers (one per fold) were trained and validated using as inputs the features that remained selected after the last step. The success rates in sample classification were 81.18%, 95.45%, 90.90%, 100% and 90.90% and that gave an overall success rate of 91.68% in ham classification (the standard deviation was 6.98%). This corresponds, in average, to 9 samples out of 110 being misclassified. Confusions occur between samples belonging to different producers but never between different quality hams within a producer. The average number of features selected after the three-step variable selection was 19 (see table 4), i.e., near 8% of the features initially available. Table 4 shows that a high number of features are shared by the different folds. This demonstrates the robustness of the variable selection method applied.

Table 4: m/z fragments selected for each selection/validation fold after the three-step feature selection process. The last row shows the most frequent m/z fragments.

Fold #	Selected m/z fragments
1	45, 47, 49, 56, 58, 59, 64, 70, 71, 73, 77, 79, 80, 81, 83, 85, 94, 100, 104, 111, 114
2	45, 47, 49, 53, 56, 57, 58, 60, 61, 64, 71, 72, 77, 81, 83, 84, 94, 100, 114, 208
3	47, 48, 55, 61, 64, 67, 77, 81, 82, 83, 84, 104, 138
4	45, 49, 56, 58, 64, 71, 77, 81, 82, 83, 84, 104, 138
5	45, 47, 51, 53, 56, 57, 58, 60, 61, 64, 67, 69, 70, 71, 72, 73, 79, 81, 82, 83, 84, 85, 93, 94, 101, 105, 108, 114, 133, 138
Most frequent m/z fragments	45, 47, 49, 56, 58, 64, 71, 77, 81, 83, 84, 94, 114, 138

Finally, an eleven-category classification was envisaged using a fuzzy ARTMAP classifier using the outcome of the previous variable selection steps. Only the 14 most frequently selected m/z ratios were used as inputs of the classifier (see table 4 for details). Its performance in the correct classification of Iberian hams was estimated to be 97.25% by using a leave one out cross-validation approach. Only 3 out of 109 ham samples were misclassified. Furthermore, a correct discrimination between hams from pigs fed on acorn or fodder was found to be possible. These results compare favourably to the 94.49% classification success rate reached with a fuzzy ARTMAP classifier that used all the features available (i.e., 209).

A short discussion on the fragments selected by the feature selection method and used to build the ham classification models is as follows. Differences in pig feeding lead to different volatile profiles obtained from crushed samples of subcutaneous fat and meat in Iberian hams. The levels of hexanal and pentanal, which arise mainly from the oxidation of linoleic acid, are rather similar regardless of pig feeding. On the other hand,

nonanal, the most important aldehyde derived from oleic acid is found in significantly larger quantities in pigs fed on acorn than in pigs fed on fodder [24,25]. The m/z fragment 114 selected in he model is present in the mass spectrum of nonanal, and therefore, helps discriminating between acorn and fodder fed pigs. Other fragments selected such as m/z 77, may arise from aromatic volatiles, m/z 71 from esters, alkanes, propylketones and butanoate and m/z 45 could be due to the presence carboxylic acids or alcohols. Finally, the presence of pentylketones and methylketones is revealed by m/z 56 and 58, respectively. All these compounds have been reported to be characteristic of the headspace of dry cured Iberian hams [24,25].

4. CONCLUSIONS

A new strategy for feature selection has been introduced and its good performance demonstrated using different MS e-nose databases. The feature selection consists of three steps, the first two being aimed at eliminating non-informative and highly collinear features, respectively. The removal of noisy and redundant features is computationally inexpensive and allows for dramatically reducing the number of variables (near 80% of initially available features are eliminated after the second step). This is especially interesting to solve MS-based electronic nose problems where the number of features (m/z fragments) available per measurement is high. The third step makes use of simulated annealing, which is a stochastic search method to further reduce the number of variables (fine-tuning of the feature selection process).

The strategy has been applied initially to a database consisting of synthetic mixtures of volatile compounds. This simple database has been used to show that the feature selection process is able to identify a minimal set of fragments that enables the correct discrimination between mixtures using a simple fuzzy ARTMAP classifier. Furthermore, given the simple nature of the problem envisaged, it was possible to show that the fragments selected 'made sense', that is, were characteristic ionisation fragments of the species present in the mixtures to be discriminated.

Once the correct performance of the feature selection method was demonstrated, it was applied to an additional database (Iberian hams).

Applying the method to the Iberian ham database resulted in the number of features being reduced from 209 down to 14. Using the surviving features, a fuzzy ARTMAP classifier was able to discriminate ham samples according to producer and quality (11-category classification) with a 97.24% success rate. It was also possible to identify, with a 100% success rate, whether the pigs had been fed on acorn or fodder.

For the different databases studied, performing variable selection results in a dramatic decrease in dimensionality and an increase in classification performance. The methods introduced here are useful not only to solve MS-based electronic nose problems, but are of interest for any electronic nose application suffering from high-dimensionality problems, no matter the sensing technology employed.

REFERENCES

[1]. R. Marsili, SPME-MS-MVA as an electronic nose for the study of off-flavors in milk, *J. Agric. Food Chem.*, 47 (1999) 648-654.

[2]. E. Fallik, S. Alkali-Tuvia, B. Horev, A. Copel, V. Rodov, Y. Aharoni, D. Ulrich, H. Schulz, Characterisation of 'Galia' melon aroma by GC and mass spectrometric sensor measurements after prolonged storage, *Postharvest Biol. Technol.*, 22 (2001) 85-91.

[3]. C. Peres, F. Begnaud, L. Eveleigh, J.L. Berdagué, Fast characterization of foodstuff by headspace mass spectrometry (HS-MS), *Trends Anal. Chem.*, 22 (2003) 858-866.

[4]. M. Vinaixa, S. Marín, J. Brezmes, E. Llobet, X. Vilanova, A. Ramos, V. Sanchís, Early detection of fungal growth in bakery products by use of an electronic nose based on mass spectrometry, *J. Agric. Food Chem.*, 52 (2004) 6068-6074.

[5]. P. Jonsson, J. Gullberg, A. Nordström, M. Kusano, M. Kowalczyk, M. Sjöstrolm, T. Moritz, A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS, *Anal. Chem.*, 76 (2004) 1738-1745.

[6]. B. Dittmann, S. Nitz, Strategies for the development of reliable QA/QC methods when working with mass spectrometry-based chemosensory systems, *Sensors and Actuators B*, 69 (2000) 253-257.

[7]. C.B. Lucasius, M.L.M Beckers, G. Kateman, Genetic algorithms in wavelength selection: a

comparative study, *Anal. Chim. Acta* 286 (1994) 135-153.

[8]. U. Depczynski, V.J. Frost, K. Molt, Genetic algorithms applied to the selection of factors in principal component regression, *Anal. Chim. Acta* 420 (2000) 217-227.

[9]. J. Sun, A correlation principal component regression analysis of NIR data, *J. Chemom.* 9 (1995) 21-29.

[10]. T. Eklöv, P. Mårtensson, I. Lundström, Selection of variables for interpreting multivariable gas sensor data, *Anal. Chim. Acta*, 381 (1999) 221-232.

[11]. Lu Xu, Wen-Jun Zhang, Comparison of different methods for variable selection, *Anal. Chim. Acta*, 446 (2001) 477-483.

[12]. N. Paulsson, E. Larson, F. Winquist, Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose, *Sensors and Actuators A*, 84 (2000) 187-197.

[13]. J. Brezmes, P. Cabré, S. Rojo, E. Llobet, X. Vilanova, X. Correig, Discrimination between different samples of olive using variable selection techniques and modified fuzzy ARTMAP neural networks, Proceedings of the 9th International Symposium on Olfaction and Electronic Nose, ISOEN'02, Rome, Italy, Vol. 1, 188-190, 2002.

[14]. T. Artursson, M. Holmberg, Wavelet transform of electronic tongue data, *Sensors and Actuators B*, 87 (2002) 379-391.

[15]. J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, *Microchem. J.*, 47 (1993) 60-66.

[16]. D. Broadhurst, R. Goodacre, A. Jones, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, *Anal. Chim. Acta*, 348 (1997) 71-86.

[17]. L. Nolle, D.A. Armstrong, A.A. Hopgood, J. A. Wware, Simulated annealing and genetic algorithms applied to finishing mill optimisation for hot rolling of wide steel strip, *Int. J. of Know.-based Intell. Engin. Sys.*, 6, (2002) 104-111.

[18]. E. Llobet, J. Brezmes, O. Gualdrón, X. Vilanova, X. Correig, Building parsimonious fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm: application to multisensor systems for gas analysis, *Sensors and Actuators B*, 99 (2004) 267-272

[19]. J.W. Gardner, P. Boilot, E.L. Hines, Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach, *Sensors and Actuators B*, 106 (2005) 114-121.

[20]. The Mathworks Inc., Matlab User's Guide, 2004. http:// www.mathworks.com

[21]. B. Wise, N.B. Gallager, PLS ToolBox 2.0, Eigenvector Research, http://www.eigenvector.com

[22]. R. Ionescu, E. Llobet, X. Vilanova, J. Brezmes, J.E. Sueiras, J. Calderer, X. Correig, Quantitative analysis of NO_2 in the presence of CO using a single tungsten oxide semiconductor sensor and dynamic signal processing, *The Analyst*, 127 (2002) 1237-1246.

[23]. M. Shen, A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn, A. Tropsha, Quantitative structure-activity relationship analysis of unctionalized amino acid anticonvulsant agents using *k*-nearest neighbor and simulated annealing PLS Methods, *J. Med. Chem.*, 45 (2002) 2811-2823.

[24]. I. González-Martín, J.L. Pérez-Pavón, C. González-Pérez, J. Hernández-Méndez, N. Álvarez-García, Differentiation of products derived from Iberian breed swine by electronic olfactometry (electronic nose), *Anal. Chim. Acta*, 424 (2000) 279–287.