Sensors and Actuators B 187 (2013) 471–480

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

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

Optimal feature selection for classifying a large set of chemicals using metal oxide sensors

Thomas Nowotny[a,b,*], Amalia Z. Berna[b], Russell Binions[c], Stephen Trowell[b]

[a] CCNR, School of Engineering and Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, UK
[b] CSIRO Ecosystem Sciences and Food Futures Flagship, GPO Box 1700 Canberra, ACT 2601, Australia
[c] School of Engineering and Materials Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Available online 13 February 2013

Keywords:
Feature selection
Metal oxide sensors
Classification
Support vector machines
Electronic nose

ABSTRACT

Using linear support vector machines, we investigated the feature selection problem for the application of all-against-all classification of a set of 20 chemicals using two types of sensors, classical doped tin oxide and zeolite-coated chromium titanium oxide sensors. We defined a simple set of possible features, namely the identity of the sensors and the sampling times and tested all possible combinations of such features in a wrapper approach. We confirmed that performance is improved, relative to previous results using this data set, by exhaustive comparison of these feature sets. Using the maximal number of different sensors and all available data points for each sensor does not necessarily yield the best results, even for the large number of classes in this problem. We contrast this analysis, using exhaustive screening of simple feature sets, with a number of more complex feature choices and find that subsampled sets of simple features can perform better. Analysis of potential predictors of classification performance revealed some relevance of clustering properties of the data and of correlations among sensor responses but failed to identify a single measure to predict classification success, reinforcing the relevance of the wrapper approach used. Comparison of the two sensor technologies showed that, in isolation, the doped tin oxide sensors performed better than the zeolite-coated chromium titanium oxide sensors but that mixed arrays, combining both technologies, performed best.

1. Introduction

Feature selection is one of the more important issues in the field of machine learning and bioinformatics. In general, the goals of feature selection are to reduce data dimensionality and to build robust classification models. The method of feature selection can affect the results of both classification and clustering. A good feature subset should strongly support classification and clustering. Filter and wrapper methods are two well-known feature selection techniques for high dimensional data sets. In the filter method, features are selected on the basis of feature separability of samples, which is independent of the learning algorithm. The separability only takes into account the relations between the features, so the selected features may not be optimal. Wrapper methods search for critical features based on the learning algorithm to be employed, and often lead to better results than filter methods [1].

In the field of chemical sensing, when using sensor arrays, an important consideration is the type and the number of sensors to use. Further choices apply to how to sample data from the sensors and how to pre-process the collected raw data (see [2] for a recent review). It is well known in machine learning that this process of feature selection is very important for the eventual success of the overall classification (recognition) system. From the perspective of maximizing information it may seem that using more sensors can only improve performance, as long as the sensors are not fully redundant (identical) or fully uncorrelated with the problem (equivalent to noise). Moreover, from a practical point of view, solid-state chemistry using combinatorial synthetic approaches [3] or biosensor design, based on natural genetic diversity [4], are now capable of generating an almost limitless repertoire of potential chemical sensors. Identifying optimal or at least efficient subsets of these sensors and their secondary features for incorporation into chemical sensor arrays is becoming increasingly important.

Here, we systematically investigate the feature definition (extraction) and selection problem for fully classifying a set of 20 chemicals using metal oxide sensors (MOS) [5] and linear support vector machines (SVMs) [6], as in [7,8], in a wrapper approach

Abbreviations: SVM, support vector machine; FS, feature selection; MOS, metal oxide sensors.

* Corresponding author at: School of Engineering and Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, UK. Tel.: +44 1273 678593; fax: +44 1273 877873.
E-mail addresses: t.nowotny@sussex.ac.uk (T. Nowotny), Amalia.Berna@csiro.au (A.Z. Berna), r.binions@qmul.ac.uk (R. Binions), Stephen.Trowell@csiro.au (S. Trowell).

http://dx.doi.org/10.1016/j.snb.2013.01.088

[9,10], similar to the approach in [11] but using an exhaustive search as in [12,13].

In the development of electronic noses, classical feature choices have been the maximum response and the area under the response curve. More recently, the area under a phase–space embedding of the sensor response [14,15] and exponential moving averages of the derivative of the sensor response have been proposed as suitable feature sets [16]. Other authors have put forward methods ranging from subspace projection methods with biomimetic inspiration [17], to spectral methods such as Fast Fourier Transform [18] and Discrete Wavelet Transform [19–22], classical statistical methods such as linear discriminant analysis and principle component analysis [23], and parametric methods such as curve fitting [24–27].

Here, we take a step back and investigate the use of very simple potential features – a subset of the measured data points (resistances at different measurement times). We contrast this analysis with some of the more popular feature choices discussed above. In contrast to many other works on enoses, we evaluate our methods on a quite large set of 20 analytes that are classified all-against-all. With this approach we complement the related works on smaller classification problems with only a few classes.

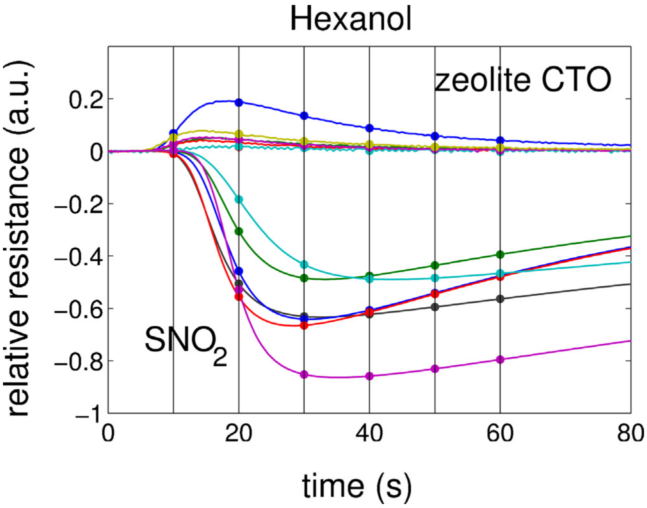

Fig. 1. Example of responses from the FOX Enose fitted with the twelve-sensor array. Responses of SnO_2 sensors are drawn downwards and responses of CTO sensors upwards. The vertical lines mark the chosen available sampling times. Note that the SnO_2 sensors are more sensitive overall than the CTO sensors.

2. Materials and methods

2.1. Electronic nose measurements

The measurements were performed with a FOX 3000 Enose (Alpha M.O.S., Toulouse, France). Originally, the instrument was equipped with a modified array of 12 semiconducting sensors comprising an array of six standard, doped tin dioxide (SnO_2) and six chromium titanium oxide (CTO) and tungsten oxide (WO_3) sensors. We removed the CTO and WO_3 sensors and replaced them with an array of six novel CTO based sensors, five of these being zeolite-coated [28] and one an uncoated CTO sensor.

The basic concept behind the modified CTO sensors is the addition of a transformation layer over the porous chromium titanium oxide sensing element. A transformation element is designed to modify or restrict the composition of gases that contact the sensing element. In this case, the transformation layer comprises the acid (or sodium) forms of zeolites A, ZSM-5 and ZSM22 MCM-41. Zeolites are ideal for this purpose due to their porous nature, having pore and channel structures of molecular dimensions (see Table 1). They are able to restrict the size and shape of gas phase molecules reaching the sensor through pore size control and selective permeability [29,30]. They also act as selective cracking and partial oxidation catalysts [31] with molecular size- and shape-specificity. Furthermore, the zeolite's catalytic behaviour can be modified, and thus tuned, by insertion of metal ions or various nanoparticles. Modification may be by either ion-exchange or lattice substitution on either their internal or external surfaces.

Due to the physical characteristics of the sensors, the two arrays were housed in different chambers. The set of chemical compounds analysed consisted of five chemicals each from four chemical groups: alcohols, aldehydes, esters and ketones (Table 2). Chemicals were chosen from a larger set of chemicals used in a comparison of metal oxide with biological sensors [32]. Each chemical compound was diluted using paraffin oil to give a final concentration in the range 1.22–8.03 $\times 10^{-5}$ M (Table 2). In total, ten replicates of each sample were prepared. Standard concentrations were chosen for each chemical class, such that the absolute values of the responses for most chemicals of that class were towards the higher end of the scale for any of the two sensor types, i.e. the ratio of the maximal resistance change to the baseline (R/R_0) was between ±0.8 and ±1 (Fig. 1). Samples of 1 ml were presented in a 20 ml glass vial using the static headspace method. The instrument was equipped with an autosampler (HS50, CTC Analytics, Switzerland), which allows reproducible injections. The injection port of the Fox was set to 30 °C and the headspace volume taken for analysis was 500 μl. Samples were analysed in groups based on chemical family with the analysis of the 200 samples being completed over four days. Dry zero grade air (flow rate 150 ml min^{-1}) was used to sweep the sample through the two sensor chambers. The sensor response was recorded for a total of 300 s at 2 Hz, see Fig. 1 for a typical measurement. A 240 s delay was imposed between samples to allow

Table 1
Overview of the sensors in the electronic nose, together with a brief description of each sensor.

#	Sensor	Description
1	CTO	Chromium–titanium oxide sensor without coating. general VOC sensor.
2	CTO-HZSM-5	Chromium–titanium oxide sensor with H-ZSM-5 zeolite overlayer, pore size 5.1–5.5 Å.
3	CTO + NaZSM-5	Chromium–titanium oxide sensor with Na-ZSM-5 zeolite overlayer, pore size 5.1–5.5 Å.
4	CTO + HLTA	Chromium–titanium oxide sensor with H-LTA zeolite overlayer, pore size 3.5 Å.
5	CTO + MCM-41	Chromium–titanium oxide sensor with MCM-41 overlayer, pore size 30–100 Å.
6	CTO + HZSM-22	Chromium–titanium oxide sensor with H-ZSM-22 zeolite overlayer, pore size 4.6–5.7 Å.
7	T30/1	SnO_2 sensor for detecting solvents.[a]
8	P10/1	SnO_2 sensor for detecting hydrocarbons and methane.[a]
9	P10/2	SnO_2 sensor for detecting methane, propane and aliphatic non polar molecules.[a]
10	P40/1	SnO_2 sensor for detecting chlorinated and fluorinated compounds.[a]
11	T70/2	SnO_2 sensor for detecting alcohol vapours and aromatic compounds.[a]
12	PA/2	SnO_2 sensor for detecting low concentration of hydrogen, ammonia, amines.[a]

[a] *Source*: Alpha M.O.S. (1995). FOX 2000–4000 Electronic Nose User Manual. Alpha M.O.S., Toulouse.

Table 2
Concentrations of analytes used according to their chemical classes.

Alcohols (1.22×10^{-5} M)	Aldehydes (8.03×10^{-5} M)	Esters (3.70×10^{-5} M)	Ketones (3.79×10^{-5} M)
1-Pentanol	Acetaldehyde	Ethylhexanoate	Acetone
1-Hexanol	Butanal	Ethylacetate	2-Butanone
Z2-Hexen-1-ol	Hexanal	Isopentylacetate	2-Pentanone
1-Octen-3-ol	E2-hexenal	Methylacetate	2-Heptanone
3-Methylbutanol	Furfural	Ethylbutyrate	2,3-Butanedione

the sensors to return to baseline, this cleaning procedure was performed at a flow rate of 150 ml min^{-1} using dry zero grade air. Data were captured and pre-analysed using AlphaSoft v.8 (Toulouse, France).

2.2. Feature sets

To define a reasonably sized set of possible features, for each sensor, we extracted six candidate time points, namely 20, 40, 60, 80, 100 and 120 s (Fig. 1), from the 600 data points available (2 Hz/300 s). The global population of candidate feature sets comprised these 1–6 time points in all permutations and, for each of the time point permutations, all possible permutations of 1–12 available sensors. Note that, in order to reduce the complexity of the problem and make it computationally manageable, we did not include feature sets where the selected time point(s) varied among sensors.

In a second set of numerical experiments we considered six features that are commonly used in the enose literature instead:

1. The absolute maximal response, $\left|R_{\max} - R_0\right|$.
2. The area under the full response curve, $\int_0^T R(t)dt$, where T is our total measurement time, $T = 300$ s.
3. The phase space area [14], $\int_0^{R_{\max}} (dR(t)/dt)dR = \int_0^{T_{\max}} (dR(t)/dt)^2 dt$, where we made use of the fact that $R(t)$ is (approximately) smooth and strictly monotonic, hence invertible to derive the right hand side version.

4–6. Exponential moving averages [16] of the derivative of the sensor response, $E_\alpha(R) = \max_k \text{ema}_\alpha R(k)$, where the discretely sampled exponential moving average $y(k) = \text{ema}_\alpha R(k)$ is define recursively as $y(k) = (1-\alpha)y(k-1) + \alpha(x(k) - x(k-1))$ with smoothing factors $\alpha = 0.005, 0.05, 0.5$ which are the equivalent values for our sampling frequency of 2 Hz to the values in [16] for sampling at 10 Hz.

2.3. Classification algorithm and cross-validation

We used the libsvm [33] library for linear support vector machines [6] to perform the cross-validation experiments. We performed classification using linear C-SVC (SVM classification with cost parameter C) with four C values, C = 1024, 4096, 16,384, 65,536. We observed consistent performance for all tested C values and report only results for C = 65,536 in the remainder of the paper.

All results reported were obtained from 10-fold balanced cross-validation: the data set was split into a training and test set by randomly choosing one measurement of the ten measurements available for each chemical to form the test set. The remaining 20 × 9 measurements form the training set. This procedure was repeated ten times, excluding all previously chosen test samples from the choice until all measurements have been used in a test set once. We performed ten repetitions of this entire procedure, so that the performance measurements reported below are the average performance of training and testing 100 classifiers.

2.4. Clustering quality and Mahalanobis distance

For the purpose of comparing the classification results with the structure of the data, given each particular feature set, we defined the quality of clustering as the quotient $d_{\text{inter}}/d_{\text{intra}}$ of the average Euclidean distance between average class vectors,

$$d_{\text{inter}} = \left\langle \left\| \left\langle \vec{x} \right\rangle_i - \left\langle \vec{x} \right\rangle_j \right\| \right\rangle_{i,j} \tag{1}$$

and the average Euclidean distance of vectors within a class to the average class vector,

$$d_{\text{intra}} = \left\langle \left\| \vec{x}_{i,k} - \left\langle \vec{x} \right\rangle_i \right\| \right\rangle_k . \tag{2}$$

Here, $\vec{x}_{i,k}$ denotes the kth measurement of chemical (class) i, $\langle\cdot\rangle$ denotes taking an average and $\|\cdot\|$ denotes the Euclidean norm in the space defined by the feature set of interest.

As a second measure of data set structure we employed the average pairwise Mahalanobis distance suggested by Muezzinoglu et al. [34]. The squared Mahalanobis distance of two classes was estimated by

$$D^2(i,j) = \left(\left\langle \vec{x} \right\rangle_i - \left\langle \vec{x} \right\rangle_j \right)^T \hat{S}_{ij}^{-1} \left(\left\langle \vec{x} \right\rangle_i - \left\langle \vec{x} \right\rangle_j \right) \tag{3}$$

where $\hat{S}_{ij}^{-1}$ denotes the inverse of the weighted average of the estimated covariance matrices $\hat{S}_i$ and $\hat{S}_j$ of the vectors in classes i and j and $\left\langle \vec{x} \right\rangle_i$ and $\left\langle \vec{x} \right\rangle_j$ the average class vectors for class i and j, respectively. For assessing the overall structure of the data set using the Mahalanobis distance, we calculated the mean pairwise Mahalanobis distance MD of all pairs of classes, defined by

$$MD^2 = \sum_{i \neq j}^{N} D^2(i,j). \tag{4}$$

2.5. Mean absolute pairwise correlation of features

To calculate the mean absolute pairwise correlation of features we used for the pairwise correlation of two features x and y the estimator

$$r_{x,y} = \frac{1/(n-1)\sum_k (x_k - \bar{x})(y_k - \bar{y})}{s_x s_y} \tag{5}$$

here x_k and y_k are the values the two features take across the full data set, $k = 1, \ldots, n$, where n = 200 and s_x and s_y denote the standard estimators for the standard deviations of x and y. For example, x may be the values of sensor i at time r and y the values of sensor j at time s, where $i, j \in \left\{1, \ldots, 12\right\}$ and $s, t \in \left\{20, 40, \ldots, 120\right\}$. To obtain the mean absolute pairwise correlation of a feature set S, we average over the absolute value of all so estimated pairwise correlations of features x and y in S:

$$r(S) = \frac{2}{N(N+1)} \sum_{x \le y \in S} \left| r_{x,y} \right| \tag{6}$$

where the sum is over all features in the feature set S, but counting pairs (x,y) and (y,x) only once (as the correlation is symmetric). Note that we included x = y to obtain a meaningful measure for feature sets that have only one feature. We take the absolute value as we are interested in how linearly dependent features are, no matter whether they are correlated or anti-correlated.

3. Results

3.1. Classification performance of feature sets

We first investigated the overall performance of all potential feature sets for classifying the 20 chemicals against each other. We measure and report performance as the fraction (or percentage) of correctly classified measurements in the test sets (see Section 2), ranging from 0 (no examples classified correctly) to 1 (all examples classified correctly). Fig. 2A shows the observed performance for the feature sets formed of six candidate time points and all 12 sensors. The performance is reported for feature sets of different size constraint, e.g. (2,3) denotes a feature set with two time points each from three sensors. We note that the best feature sets lead to much better classification performance than previously reported classifiers based on this set of measurements [35]. There are several feature sets that lead to 100%, i.e. error-free, classification in the ten repetitions of 10-fold cross-validation. There are quite a few feature sets that enable this optimal performance (3368) but it is a small fraction of all sets tested (1.3% of the tested 257,985 sets). The best performance is not achieved with the naïvely expected maximal sensor- and data-use (12 sensors, six time points, top line of Fig. 2). Neither is the classical approach of using just one time point for all sensors (e.g. the time when the maximum signal is observed) particularly successful (second line of Fig. 2), confirming earlier findings in other enose applications [13,36].

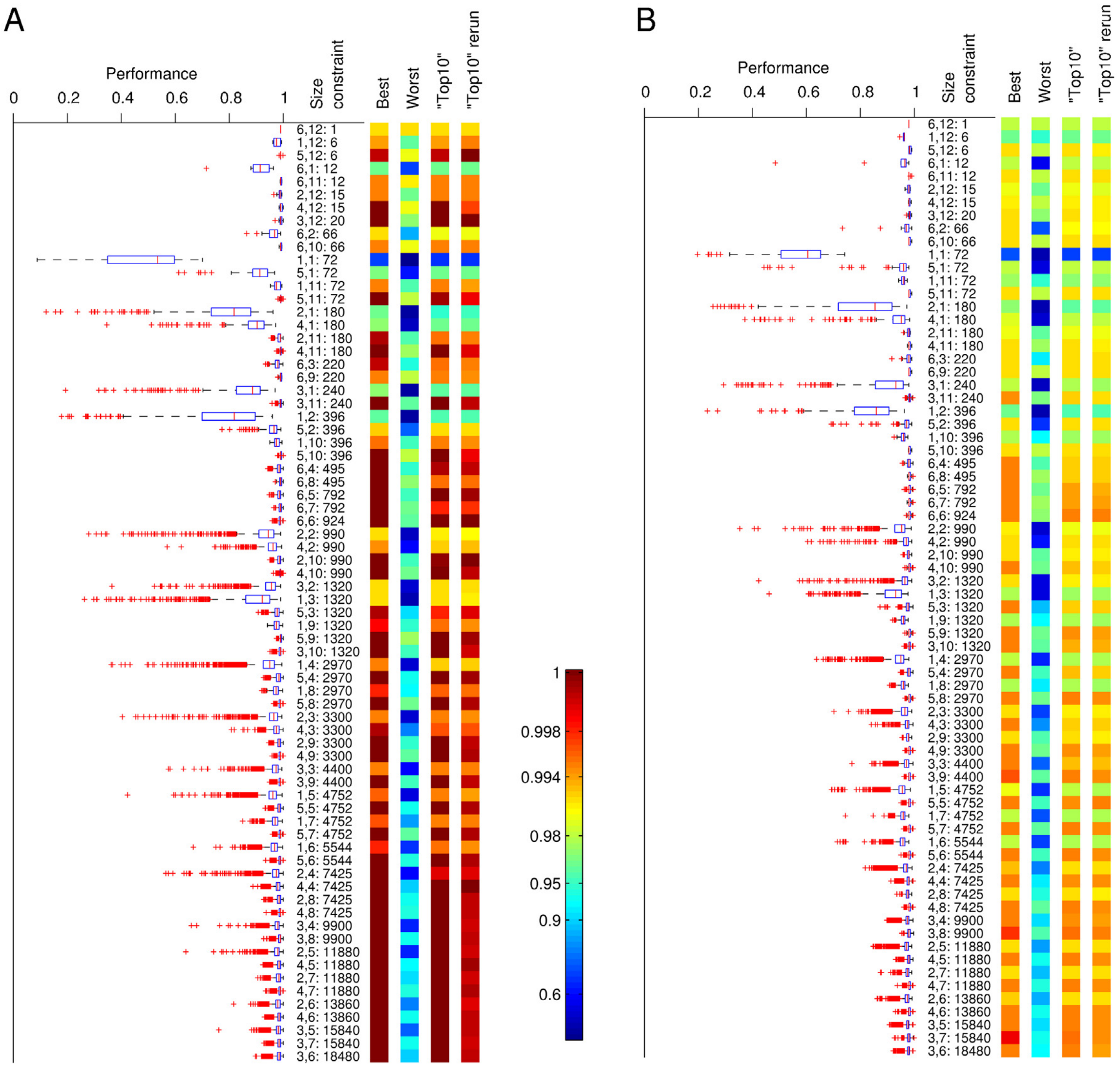

Fig. 2. (A) Fractional prediction accuracy of 10-fold cross-validation using linear SVMs for all allowed combinations of six time points and the 12 available sensors. The box plots show the median and 25% and 75% quantiles, the estimated overall range (whiskers) and identified outliers (red crosses) of the observed performances for each given sub-family, constrained to have *x* time points and *y* sensors (first and second numerical columns). The third numerical column indicates the resulting number of combinations that was tested. The coloured columns indicate the best observed performance, the worst performance, the performance of the "top 10" group of feature choices (see Sections 2 and 3.1) and the performance of this group in a repeat 10-fold cross-validation. The prevalence of excellent performance at the bottom of the graph indicates that sub-families with many different choices typically contain choices that lead to excellent performance. However, there are also examples of excellent performance for other situations, e.g. the (3,12), (6,6) and (2,10) groups. Note the highly non-linear, logarithmic colour code for the classification performance in which high performances close to 1 are resolved in many colour gradations from dark red to cyan and weak performances closer to 0 are compressed into a few blue colours. (B) The same analysis applied to the same array of 12 sensors and set of 20 chemicals using six different popular feature sets from the enose literature: absolute maximal response, the area under the response, the area of the phase–space curve of the response [14], and exponential moving averages of the derivative of the response curve [16]. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of the article.)

To control for selection biases, we defined a group of well-performing feature sets (column "top10" in Fig. 2) and repeated cross-validation for this group (column "top10 rerun" in Fig. 2). As expected, the performance in the rerun is typically a bit worse than in the original run, as we had selected the most successful feature sets on this first run and hence might have collected those where we just "got lucky" with the partitions of the ten times 10-fold cross-validation. The rerun demonstrates that this selection bias is noticeable but that it is a minor effect compared to the overall variation in performance due to feature selection: the results in the rerun, albeit slightly less good than in the original run, are consistently better for feature sets that were optimal in the original run than for feature sets that were not. There are even rare examples where the classifier performs better in the rerun (for the (5,12) feature set) than in the original selection run. We particularly note that the superior performance of many of the smaller feature sets over the full (6,12) choice remains intact. Also note, however, that for smaller feature sets, classification success depends critically on the choice of the used feature sets from the pool of all potential sets of a given size. This is illustrated by the much lower worst performance (see column "worst" and the low outliers in the left column of Fig. 2).

The data is presented ordered by the number of possible feature sets for each given size constraint, ranging from a single (6,12) feature set on the top to 18,480 possible (3,6) feature sets at the bottom. The prevalence of excellent "best", "top 10" and "top 10 rerun" performance at the bottom of the graph illustrates that size constraints with many different feature set choices are more likely to have well-performing sets, even though this is not an absolute rule, see e.g. the 2970 (1,4) feature sets that are much less successful than the 20 (3,12) sets.

Fig. 2(B) illustrates the same analysis for an alternative feature set comprising six features suggested in the enose literature: absolute maximal response, the area under the response, the area of the phase–space curve of the response [14], and three different exponential moving averages of the derivative of the response curve [16] (see Section 2 for details). In the remainder of this section we will refer to the latter set as "derived features" whereas we will refer to the six representative data points that are the focus of this study as "simple features". The best-performing feature sets based on subsets of derived features perform consistently worse than the best sets using subsets of simple features. Note, however, that the performance of the derived feature set was still quite high and comparable, if not superior, to the performance based on EMA features, previously reported for this data set [35].

It is also noteworthy that the size constraints (i.e. the number of data points and number of sensors used) that lead to the better or the rather poorly performing feature sets are the same for both scenarios: the pattern of better-performing and worse-performing feature set sizes, in terms of best performance, is identical in Fig. 2(A) (simple features) and (B) (derived features), cf column "Best". We quantified this observation by calculating the correlation between the best performance of each size constraint in the two scenarios and obtained a correlation coefficient of 0.975. Furthermore, the full distributions of performances are also similar, as illustrated by the boxplots in Fig. 2(A) and (B), which share many properties. For example, the (1,1) size constraint has the same very broad and very low performance distribution in both cases and the performance distribution for feature sets of five data points and nine sensors is particularly narrow and high. Interestingly, the derived features used in Fig. 2(B) lead to a slightly improved worst performance indicated by the shorter tail of outliers in the boxplots, in particular for the lower half of the plot that contains size constraints that allow large numbers of feature choices.

3.2. Relationship to clustering

To identify possible explanations for the improved performance of some feature sets over others we compared the classification performance for each set with the quality of clustering given this feature choice. For this purpose we defined the quality of clustering as the quotient $d_{\text{inter}}/d_{\text{intra}}$ of the average Euclidean distance between average class vectors and the average Euclidean distance of vectors within a class (see Section 2). The results are illustrated in Fig. 3. We notice a positive correlation between clustering performance and classification performance, in particular in the form of the absence of examples of good performance with very low clustering quality (upper left corner in Fig. 3) and of (very) low performance with high clustering quality (lower right corner in Fig. 3). The clustering quality, however, clearly does not fully explain the classification results. The overall correlation coefficient between classification performance and clustering quality as defined here is positive but only 0.205. The most relevant cases are the best-performing feature sets, for which clustering performance is above average but not necessarily maximal, and vice versa, the very best feature sets in terms of clustering quality, which do not necessarily perform optimally (Fig. 3, inset). One possible explanation for the failure to predict classification performance for the best performances is that in these cases it is the worst case of clustering quality between classes that is important and not necessarily the mean. If we analyse the minimal $d_{\text{inter}}/d_{\text{intra}}$ ratio for all pairs of classes instead of the mean, the relationship becomes more clear (Supplemental Fig. 1) and the overall correlation reaches 0.427. However, it still does not fully predict the classification performance, in particular in terms of the maximal performance that is of most interest.

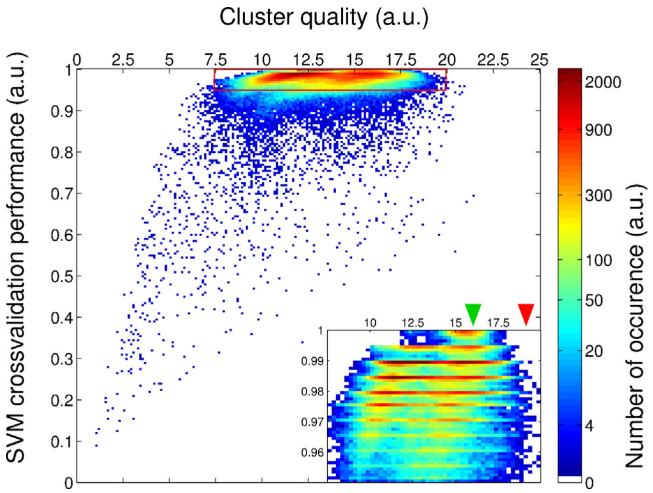

Fig. 3. Classification performance of linear SVMs for all possible feature choices in tenfold cross-validation plotted against the clustering quality (ratio of inter-class to average intra-class Euclidean distance, see Section 3.2). The displayed colour indicates the number of occurrences of each particular combination of clustering quality and performance (white represents 0). A clear correlation is noticeable, in particular the absence of points with low clustering quality and high performance (upper left corner) or low performance and high clustering quality (lower right corner). The inset shows the region marked by the red rectangle on top. While the correlation overall is minor (Pearson correlation coefficient 0.205) it is noticeable that the overwhelming majority of the very top performers (green arrow head) have above-average clustering quality. The reverse, however, is not true: the feature sets that show the very highest clustering quality do not necessarily lead to the highest performance (red arrow head). The horizontal stripes visible in the enlarged plot are not an artefact but reflect that if the classifiers fail for one specific measurement, then they tend to do so consistently in all 10 repetitions of cross-validation. The distance between stripes (0.5% = 1/200) is consistent with this interpretation. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of the article.)

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2013.01.088.

We then performed the same analysis using the mean pairwise Mahalanobis distance as suggested by Muezzinoglu et al. [34] with similar results (Supplemental Fig. 2). The correlation coefficient of mean pairwise Mahalanobis distance (MD) and classification performance is 0.189, or, if we use the logarithm of MD due to its wide range of values, the correlation coefficient is 0.392. As with the more basic clustering quality measure above, the Mahalanobis distance of classes is related to classification performance but cannot fully predict it. When considering the minimal Mahalanobis distance between mean class vectors rather than the mean, a similar picture emerges (Supplemental Fig. 3) and the correlation coefficients between minimal Mahalanobis distance and performance and logarithm of minimal Mahalanobis distance are 0.295 and 0.548, respectively.

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2013.01.088.

3.3. Sensor correlation

Another commonly held belief is that the degree of correlations of the responses of different sensors is strongly related to, if not an explicit predictor of (lack of) performance in classification. Previous work has shown that MOS sensor responses can be highly correlated in spite of different doping of the individual sensor types [32]. This has been used as an explanatory construct to justify why biological chemical receptors, which appear to be much less correlated in their responses, may outperform MOS sensor arrays. With the full assay of the performance of feature sets in SVM classification in this work we can directly test this hypothesis by comparing the classification performance of feature sets to the mean absolute pairwise correlation of the features in the sets. We calculated the average Pearson correlation coefficient of each feature in a set with each other feature in the same set across all pairs of inputs of two distinct chemicals and plotted these values against the observed performance in 10-fold cross-validation (Fig. 4). Overall there is a weak trend where less correlated features in a feature group are correlated with better performance, i.e. there is a weakly negative correlation between mean pairwise correlation of features and the classification performance of the feature set (Pearson correlation coefficient −0.284). Similarly as when comparing to the clustering quality (Section 3.2), the best performing feature sets have typically a comparably low mean pairwise correlation of their features (Fig. 4, inset), but minimal mean pairwise correlation of features does not necessarily predict optimal performance. Also, due to

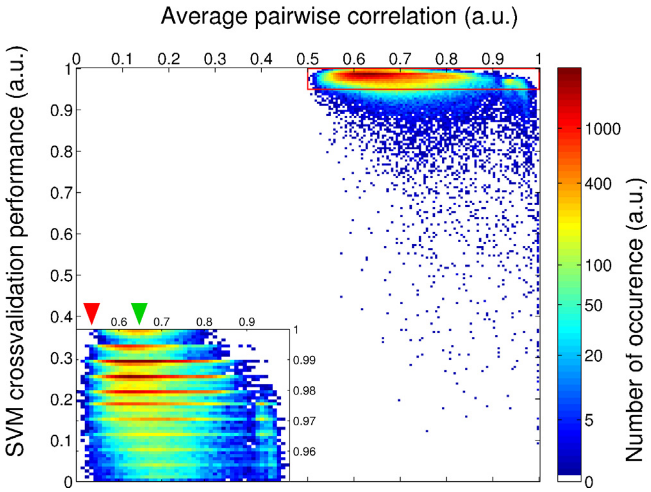

Fig. 4. Average pairwise correlation plotted against the performance of the linear SVM in tenfold cross-validation. The colour represents the number of occurrences, white representing 0. The inset shows an expanded plot of the area in the red rectangle on the top. There is a small negative correlation between the mean pairwise correlation of sensor responses in a feature set and the resulting classification performance overall (Pearson correlation −0.284) but a low pairwise correlation of features is by no means a direct predictor of good performance. When inspecting the top performers, we notice as for the clustering quality in Fig. 3 that top performing feature sets in the majority have a (in this case low) typical mean pairwise correlation of features (green arrow) but the feature sets with the lowest mean pairwise correlation of features do not necessarily lead to top classification performance. Note the highly non-linear colour scale in this figure. The stripes in the inset have the same origin as the ones in Fig. 3 (see Fig. 3 caption). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of the article.)

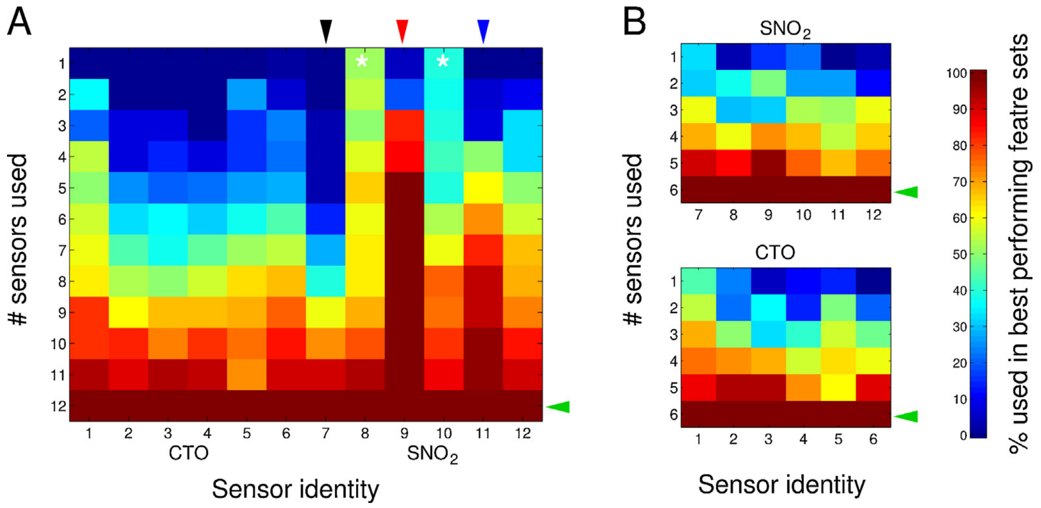

Fig. 5. Frequency of appearance of individual sensors in optimal feature sets. For each type of feature set, in terms of the number of sensors used (y-axis), the colour indicates the percentage of feature sets in the top 10 groups that contain the sensor indicated on the x-axis, for the whole array (A) and for the separate 6-sensor arrays of SnO_2 sensors and CTO sensors (B). In the full array, sensor 9 is used in almost every feature set, except those with very few sensors (red arrowhead). The second most used sensor is sensor 11 (blue arrowhead). Some other sensors are almost never used, in particular sensor 7 (black arrowhead). For feature sets of only 1 or 2 sensors the SnO_2 sensors 8 or 10 (white stars) appear often, for optimal sets of two sensors they are often found in combination with the CTO sensor 1. For the separate arrays of SnO_2 sensors and CTO sensors (B), the picture is less clear. While sensor 9 seems to be used more often than average, sensor 7, which was under-represented in the full array, is also used quite frequently. By design, when all sensors are used, they must be used equally often (i.e. always, green arrows). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of the article.)

the nature of measuring pairwise correlations, feature sets with a single feature are maximally correlated but still can lead to performances that are clearly above chance levels. The absolute values of the mean pairwise correlation of sensors are all above 0.5 which is fairly high, in particular when compared to biological chemical sensors [32].

3.4. *Composition of and relationships between optimal feature sets*

An important question for understanding the nature of our results is what relationship the optimal feature sets have to each other, what they have in common and where they differ. First we asked whether individual sensors appear preferentially in these best performing groups. Fig. 5A illustrates the number of occurrences of each sensor (*x*-axis) in the “top10” sensor groups of different size constraints (*y*-axis). While all sensors appear to some degree in the feature sets, sensor 9 appears in almost all “top10” feature sets of size three and larger, while sensors 8 (P10/1) and 10 (P40/1) are preferentially used in the very small feature sets of two or only one sensor. Sensor 7 (T30/1), on the other hand, is almost never used.

When performing the same analysis for the individual sensor technologies separately (Fig. 5B), the picture is less clear even though there is preferential use of sensor 9 (P10/2) for the SnO_2 sensors. Interestingly, when only using SnO_2 sensors, sensor 7 (T30/1) is quite well used, being the most-used sensor for feature sets with a single sensor. One could think that this discrepancy arises because sensor 7 (T30/1) may have response properties that are unusually redundant with the responses of CTO sensors. However, when calculating the average absolute correlation of sensor 7 (T30/1) with all CTO features we find the same level of correlations (0.121) as when using sensor 9 (0.120). Also the profile of correlations with individual CTO sensors is similar (data not shown).

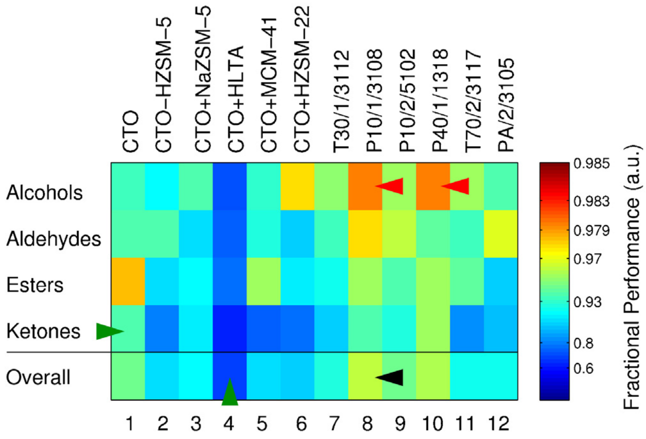

Fig. 6. Classification performance based on data from individual sensors, using all six data points. Classification performance is reported in a colour code, separately for the four chemical groups (rows 1–4) and overall (row 5). In the overall performance, sensor 8 (P10/1) performs best, black arrow head, and the most successful classification for a single chemical group occurs for alcohols using either sensor 8 (P10/1) or sensor 10 (P40/1), red arrow heads. Overall, the most challenging chemical group for classification are the ketones (green horizontal arrow head) and the worst performing sensor, when used on its own is sensor 4 (CTO + HLTA). The likely reason for this poor performance is the very small pore size of the H-LTA zeolite (~3.5 Å), which is smaller than the width of most of the molecules investigated here. Hence, most analytes will not reach the sensor surface and responses are small and unspecific [30]. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of the article.)

3.5. *Performance of individual sensors*

Not surprisingly, individually, sensors do not perform particularly well in classifying the 20 chemicals. Even so, some individual sensors can partially discriminate certain chemicals and there are significant differences in how they perform for the whole set and for each chemical class. Fig. 6 illustrates the performance of individual sensors, using all six data points, resolved for the four chemical classes of analytes. The performance reported for each class of analytes is the fraction of correct recognitions of the members of this class when compared to all analytes from all chemical classes. Sensors 8 (P10/1) and 10 (P40/1) are the best sensors individually

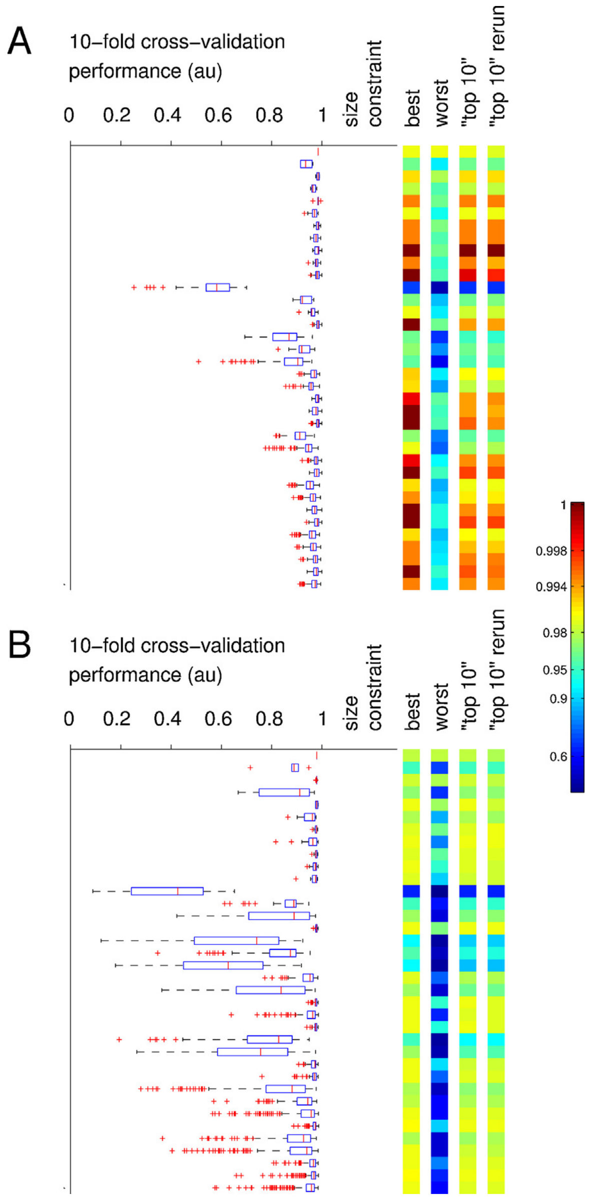

Fig. 7. Classification performance of feature sets that use only one of the two available technologies. The plots are as in Fig. 2. (A) Performance if subsets of the 6 SnO_2 sensors are used. Perfect 100% classification is still possible, notably most robustly for the (4,6) size constraint. Overall, however, the optimal performance is achieved less frequently and is less robust within the “top10” groups. (B) Performance if subsets of the 6 CTO sensors are used. Here 100% performance is not achieved and performance levels are measurably lower for all available feature choices than for the SnO_2 sensors. The SnO_2 sensors also perform much better in terms of the distribution of performance across all feature sets: the boxplots in (A) are much tighter than those in (B).

and work particularly well for alcohols (red arrowheads). It was interesting to note that sensor 11 (T70/2), which is more sensitive to alcohols, was not the best sensor to recognize them. Closer inspection of the average ratio of inter-class and intra-class distances of responses in individual sensors for alcohols revealed that indeed T70/2 has the best signal-to-noise ratio for the alcohols in this sense (Supplemental Fig. S4A). When inspecting the matrix of recognition success and failure (Supplemental Fig. S4B) and a PCA plot of sensor responses using the 6 data points of T70/2 (Supplemental Fig. S4C), we find that the problem is not distinguishing the alcohols from each other, which is done flawlessly, but distinguishing 1-pentanol (input 1) from hexanal (input 8) and E2-hexenal (input 9), see arrowheads in Supplemental Fig. S4C and the corresponding clusters in Fig. 8 D for the better-performing sensor 8 (P10/1).

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2013.01.088.

Returning to the results in Fig. 6, ketones are by far the hardest of the chemical classes to classify (row marked by the green arrowhead) and sensor 4 (CTO + HLTA) works particularly poorly for all analytes (column marked by the green arrowhead). The likely reason for the latter is the very small pore size of the H-LTA zeolite (~3.5 Å), which is too small to allow passage of most of the analytes used in this study, hence the responses of the sensor are of low amplitude and quite unspecific [30].

The overall best performing sensor is number 8 (P10/1, black arrowhead).

3.6. Comparison of the two sensor technologies

One of the novelties of the data analysed in this work is the use of the family of recently introduced zeolite-coated CTO sensors [28]. The best feature sets typically contain sensors of both the standard SnO_2 and zeolite-coated CTO types. When examining the sensor combinations that allowed 100% performance for the whole sensor array, SnO_2 sensors are always used but 99.55% of the optimal feature sets use sensors from both technologies. In this subsection we compare the performance of each of the sensor technologies when used on their own. Fig. 7 shows the performance profile of feature sets that contain either only SnO_2 sensors (Fig. 7A) or exclusively CTO sensors (Fig. 7B). As expected, the overall classification performance levels are lower than for feature sets taken from the combined array combining both sensor technologies, in correspondence with earlier findings [12,37]. Comparing the performances of the two technologies, the SnO_2 sensors appear to have higher performance, achieving 100% success for a few feature sets whereas the CTO sensors never reach 100% performance. The poorer performance of the CTO sensors may be a reflection of the lower amplitude and the resulting smaller signal-to-noise ratio in these sensors (Fig. 1).

4. Discussion

In designing a chemical sensor array, such as an artificial nose or tongue, choices concerning the number and properties of the sensors and the number and identity of data points are very important practical considerations. Engineering limitations and computational demand preclude the use of all available sensors and data points and selection of the minimal optimal set would be an efficient strategy. As Marco and Gutiérrez-Gálvez have recently pointed out [2] there is a disconnection between practitioners of machine olfaction and those developing the computational tools to process data from chemical sensor arrays. The results illustrated in Fig. 2 suggest that it may be beneficial to design a sensor array specifically for each envisioned application domain and, if doing so, that a few well-chosen sensors and data sampling times may outperform simply using the maximal array and many data points. However, it is worth noting that choosing the correct sensors and data-sampling times is critical. For example, the median performance of (three data points, six sensors) feature sets (0.985) is actually worse than the performance of the single comprehensive choice (6 data points, 12 sensors). This implies that just taking an arbitrary feature set (three data points, six sensors) would likely not improve overall success.

We notice that a large number of classification results are almost optimal and some of the differences we base our conclusions on amount to discrepancies of a single error in classifying 200 measurements of chemicals. This indicates that the array we used is well capable of this quite challenging classification problem. In future work we intend to extend our analysis to even more challenging applications, including lower or multiple concentrations, and measurements taken over an extended period of time, where sensor drift may become limiting.

As pointed out above, from the perspective of maximizing information we would have expected that classification performance can only increase when additional features are added. In the worst case one would have expected unchanged performance if the data provided by the additional features was not useful. Here, however, we saw that adding additional features can decrease the accuracy of classification. The likely explanation of this phenomenon is overfitting of relatively noisy data. The additional data may provide additional information for the training data, but this can lead to overly specific classifiers that may not generalize as well to new testing data as “less informed” ones. This trade-off between optimal classification on the training data and optimal ability to generalize to new test data, the so-called over-fitting problem, is a classic topic in machine learning. Future work will focus on unravelling what the optimal solutions are for given practical problems and the degree to which these are generalizable within or beyond a given problem set.

The work reported here was conducted with a specific classification method, i.e. a linear support vector machine. One could argue that the observed phenomenon of better classification with smaller feature sets may be specific to this particular method. While we cannot fully exclude this possibility, the effects of over-fitting are known to affect all approaches to classification. While the details may differ for other classification methods, the principal results are likely to apply to a variety of such methods and similar results have been observed for other applications [13,36].

Finally, the wrapper approach to feature selection used here led in many cases to error-free classification, in contrast to earlier work where features were chosen based on different criteria [35]. Our analysis of the relationship between the performance in classification with the clustering quality of the data and the correlation of the sensor responses in Sections 3.2 and 3.3, respectively demonstrates that there is a relationship between these properties of the data and feature choice and the eventual classification success based on the chosen features. However, the relationships are not particularly strong, suggesting that a filter approach to feature selection based on either sensor correlation or data clustering would likely be less successful. This observation is reinforced by close inspection of the structure of the data after dimensionality reduction, using PCA, for particular exemplary feature choices (Fig. 8). While there is a clear difference in the quality of clustering in more successful feature sets and it is straightforward to identify the inputs that lead to classification errors in the less successful cases (arrowheads in Fig. 8), there are also many points where errors could have occurred but did not. This figure illustrates

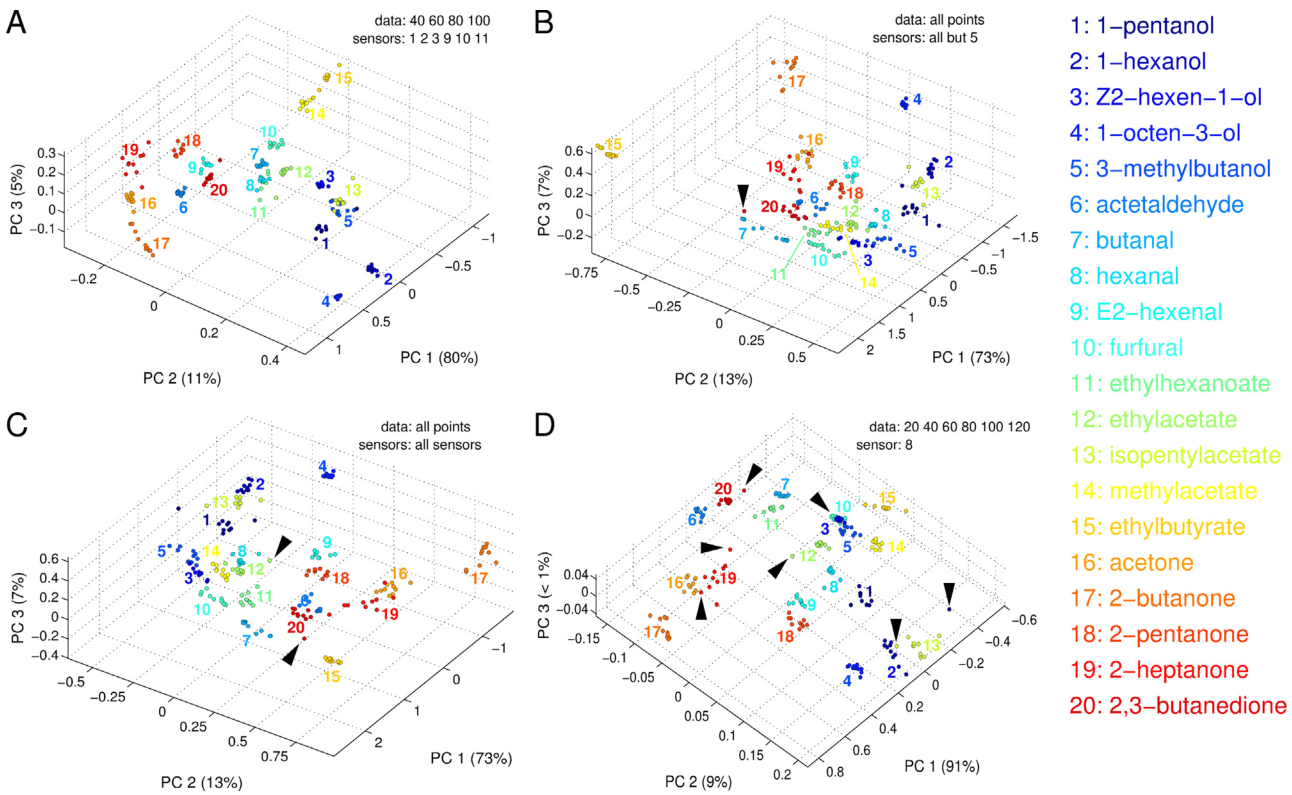

Fig. 8. PCA plots for four exemplary feature choices. (A) PCA plot of the inputs obtained with one of the optimal feature sets of four data points and six sensors (100% performance). (B) PCA plot of the inputs obtained when using the slightly less well-performing best feature set of six data points and eleven sensors (99.5% performance). (C) PCA plot of inputs when all available six data points and twelve sensors are used (99% performance). (D) PCA plot of inputs when using the best sensor selection of six data points and a single sensor (96.4% performance). The arrowheads mark inputs that lead to classification errors, e.g. in (B) the marked input of class 20 (2,3-butanedione) is misidentified as class 7 (butanal). The percentages at the axes give the fraction of the variance that is explained by the corresponding principle component. While the less well-performing feature set used for panel (D) is low dimensional and its classes are harder to separate, the best-performing feature set in (A) is not the one with highest dimensionality (in the sense of being the least well-captured by the first three principal components) between the 4 examples shown here (the sets in panels (B) and (C) have less of their overall variance explained by their first three principal components). Dimensionality in this sense is hence not a direct predictor of the classification success.

once more that there are no simple correlates to predict classification performance, re-emphasizing the relevance of the wrapper approach.

5. Conclusions

We set out systematically to assess the question of feature selection for arrays of metal oxide sensors in a classification task, using standard machine learning methods. We found that feature selection can improve classification performance and that the best-performing feature sets are not necessarily the naively expected ones.

In future work we plan to analyse in depth why particular combinations of sensors are very successful, whether this information can be translated to future novel measurements with these sensors and whether our results translate to classification methods other than linear support vector machines.

Acknowledgement

This work was partially supported by an OCE Distinguished Scientist Award of CSIRO to TN.

References

[1] Q.J. Liu, Z.M. Zhao, Y.X. Li, Y.Y. Li, Feature selection based on sensitivity analysis of fuzzy ISODATA, Neurocomputing 85 (2012) 29–37.
[2] S. Marco, A. Gutierrez-Galvez, Signal and data processing for machine olfaction and chemical sensing: a review, Sensors Journal, IEEE 12 (2012) 3189–3214.
[3] H. Koinuma, I. Takeuchi, Combinatorial solid-state chemistry of inorganic materials, Nature Materials 3 (2004) 429–438.
[4] H. Dacres, J. Wang, V. Leitch, I. Horne, A.R. Anderson, S.C. Trowell, Greatly enhanced detection of a volatile ligand at femtomolar levels using bioluminescence resonance energy transfer (BRET), Biosensors and Bioelectronics 29 (2011) 119–124.
[5] A. Berna, Metal oxide sensors for electronic noses and their application to food analysis, Sensors 10 (2010) 3882–3910.
[6] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[7] M. Pardo, G. Sberveglieri, Classification of electronic nose data with support vector machines, Sensors and Actuators B: Chemical 107 (2005) 730–737.
[8] C. Distante, N. Ancona, P. Siciliano, Support vector machines for olfactory signals recognition, Sensors and Actuators B: Chemical 88 (2003) 30–39.
[9] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.
[10] R.K.G. John, K. Pfleger, Irrelevant features and the subset selection problem, in: Fifth International Conference on Machine Learning, New Brunswick, NJ, Morgan Kaufmann, Los Altos, 1994, pp. 121–129.
[11] E. Phaisangittisagul, H.T. Nagle, Sensor selection for machine olfaction based on transient feature extraction, IEEE Transactions on Instrumentation and Measurement 57 (2008) 369–378.
[12] M. Pardo, L.G. Kwong, G. Sberveglieri, K. Brubaker, J.F. Schneider, W.R. Penrose, J.R. Stetter, Data analysis for a hybrid sensor array, Sensors and Actuators B: Chemical 106 (2005) 136–143.
[13] M. Pardo, G. Sberveglieri, Comparing the performance of different features in sensor arrays, Sensors and Actuators B: Chemical 123 (2007) 437–443.
[14] E. Martinelli, C. Falconi, A. D'Amico, C. Di Natale, Feature extraction of chemical sensors in phase space, Sensors and Actuators B: Chemical 95 (2003) 132–139.
[15] M. Falasconi, M. Pardo, G. Sberveglieri, I. Ricco, A. Bresciani, The novel EOS835 electronic nose and data analysis for evaluating coffee ripening, Sensors and Actuators B: Chemical 110 (2005) 73–80.
[16] M.K. Muezzinoglu, A. Vergara, R. Huerta, N. Rulkov, M.I. Rabinovich, A. Selverston, H.D.I. Abarbanel, Acceleration of chemo-sensory information processing using transient features, Sensors and Actuators B: Chemical 137 (2009) 507–512.

[17] A. Perera, T. Yamanaka, A. Gutierrez-Galvez, B. Raman, R. Gutierrez-Osuna, A dimensionality-reduction technique inspired by receptor convergence in the olfactory system, Sensors and Actuators B: Chemical 116 (2006) 17–22.

[18] A. Heilig, N. Barsan, U. Weimar, M. Schweizer-Berberich, J.W. Gardner, W. Gopel, Gas identification by modulating temperatures of SnO_2-based thick film sensors, Sensors and Actuators B: Chemical 43 (1997) 45–51.

[19] C. Distante, M. Leo, P. Siciliano, K.C. Persaud, On the study of feature extraction methods for an electronic nose, Sensors and Actuators B: Chemical 87 (2002) 274–288.

[20] R. Ionescu, E. Llobet, Wavelet transform-based fast feature extraction from temperature modulated semiconductor gas sensors, Sensors and Actuators B: Chemical 81 (2002) 289–295.

[21] X.J. Huang, Y.K. Choi, K.S. Yun, E. Yoon, Oscillating behaviour of hazardous gas on tin oxide gas sensor: Fourier and wavelet transform analysis, Sensors and Actuators B: Chemical 115 (2006) 357–364.

[22] K.E. Kramer, S.L. Rose-Pehrsson, M.H. Hammond, D. Tillett, H.H. Streckert, Detection and classification of gaseous sulfur compounds by solid electrolyte cyclic voltammetry of cermet sensor array, Analytica Chimica Acta 584 (2007) 78–88.

[23] F.J. Acevedo, S. Maldonado, E. Dominguez, A. Narvaez, F. Lopez, Probabilistic support vector machines for multi-class alcohol identification, Sensors and Actuators B: Chemical 122 (2007) 227–235.

[24] J. Samitier, J.M. Lopezvillegas, S. Marco, L. Camara, A. Pardo, O. Ruiz, J.R. Morante, A new method to analyze signal transients in chemical sensors, Sensors and Actuators B: Chemical 18 (1994) 308–312.

[25] R. Gutierrez-Osuna, H.T. Nagle, S.S. Schiffman, Transient response analysis of an electronic nose using multi-exponential models, Sensors and Actuators B: Chemical 61 (1999) 170–182.

[26] S. Wlodek, K. Colbow, F. Consadori, Signal-shape analysis of a thermally cycled tin-oxide gas sensor, Sensors and Actuators B: Chemical 3 (1991) 63–68.

[27] L. Carmel, S. Levy, D. Lancet, D. Harel, A feature extraction method for chemical sensors in electronic noses, Sensors and Actuators B: Chemical 93 (2003) 67–76.

[28] R. Binions, H. Davies, A. Afonja, S. Dungey, D. Lewis, D.E. Williams, I.P. Parkin, Zeolite-modified discriminating gas sensors, Journal of the Electrochemical Society 156 (2009) J46–J51.

[29] P. Varsani, A. Afonja, D.E. Williams, I.P. Parkin, R. Binions, Zeolite-modified WO_3 gas sensors – enhanced detection of NO_2, Sensors and Actuators B: Chemical 160 (2011) 475–482.

[30] R. Binions, A. Afonja, S. Dungey, D.W. Lewis, I.P. Parkin, D.E. Williams, Discrimination effects in zeolite modified metal oxide semiconductor gas sensors, IEEE Sensors Journal 11 (2011) 1145–1151.

[31] N.Y. Chen, T.F. Degnan, C. Morris Smith, Molecular Transport and Reaction in Zeolites: Design of Shape Selective Catalysts, VCH, New York, 1994.

[32] A.Z. Berna, A.R. Anderson, S.C. Trowell, Bio-benchmarking of electronic nose sensors, PLoS One 4 (2009).

[33] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines in http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[34] M.K. Muezzinoglu, A. Vergara, R. Huerta, M.I. Rabinovich, A sensor conditioning principle for odor identification, Sensors and Actuators B: Chemical 146 (2010) 472–476.

[35] A.Z. Berna, A. Vergara, M. Trincavelli, R. Huerta, A. Afonja, I.P. Parkin, R. Binions, S. Trowell, Evaluating zeolite-modified sensors: towards a faster set of chemical sensors, in: P. Gouma (Ed.), AIP Conf. Proc, Amer Inst Physics, Melville, 2011, pp. 50–52.

[36] A. Pardo, S. Marco, C. Calaza, A. Ortega, A. Perera, T. Sundic, J. Samitier, Methods for sensors selection in pattern recognition, in: J.W. Gardner, K.C. Persaud (Eds.), Electronic Noses and Olfaction, IoP Publishing, 2000, pp. 83–88.

[37] H. Ulmer, J. Mitrovics, U. Weimar, W. Gopel, Sensor arrays with only one or several transducer principles? The advantage of hybrid modular systems, Sensors and Actuators B: Chemical 65 (2000) 79–81.

Biographies

Thomas Nowotny received his PhD in theoretical physics from the University of Leipzig, Germany and was a postdoctoral researcher and assistant research scientist at UCSD. He is now a Reader in Informatics at the University of Sussex. He has authored over 30 peer-reviewed publications in international journals. His broad research interests include problems in computational neuroscience, biological and artificial olfaction and cognition and learning in brains and machines.

Amalia Z. Berna received her engineering degree in food technology from the Agrarian National University, Peru and PhD in bioscience engineering from Catholic University of Leuven, Belgium. She is currently working at a CSIRO as Team Leader of the Chemometric team and Research Scientist. She has experience in sensor characterization and the application of multivariate statistical methods to sensors and sensor arrays. She currently leads CSIRO's research into diagnosis of disease in expired breath volatiles and detection of microbial contaminants in the food chain. Amalia has 21 peer-reviewed papers in international journals and 16 refereed conference papers.

Russell Binions holds a degree in chemistry from the University of Durham and a PhD in Chemistry from University College London. He is currently a Lecturer in Functional Materials in the School of Engineering and Materials at Queen Mary, University of London and an Honorary Senior Research Associate at UCL. He is the author of over 50 peer reviewed journal papers, 4 book chapters and 1 book. His research interests encompass new chemical vapour deposition techniques, metal oxide semiconductor materials, gas sensors, photocatalysis, chromogenic materials, nanocomposite films and energy efficient building materials.

Stephen Trowell holds a natural sciences degree from Cambridge University, majoring in biochemistry, and a PhD in visual neuroscience from the Australian National University. Stephen has 100 publications, 40 of them refereed international journals and book chapters, and is inventor on 12 patent families. Stephen's practical achievements include leading the team that developed The LepTon^{TM} Test Kit, an immunodiagnostic kit used to manage insecticide resistance. He is a Senior Principal Research Scientist at CSIRO and he currently leads a team developing a bioelectronic nose to detect explosives and other analytes. He finds the multidisciplinary nature of this challenge particularly rewarding.