ELSEVIER

Sensors and Actuators B 72 (2001) 149–159

SENSORS AND ACTUATORS B CHEMICAL

www.elsevier.nl/locate/sensorb

Assessing the ability to predict human percepts of odor quality from the detector responses of a conducting polymer composite-based electronic nose

Michael C. Burl[b], Brett J. Doleman[a], Amanda Schaffer[a,b], Nathan S. Lewis[a,*]

[a] *127-72 Noyes Laboratory, Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[b] *126-347 Jet Propulsion Laboratory, 4800 Oak Grove Blvd, Pasadena, CA 91109, USA*

Received 30 May 2000; received in revised form 25 August 2000; accepted 2 September 2000

Abstract

The responses of a conducting polymer composite ''electronic nose'' detector array were used to predict human perceptual descriptors of odor quality for a selected test set of analytes. The single-component odorants investigated in this work included molecules that are chemically quite distinct from each other, as well as molecules that are chemically similar to each other but which are perceived as having distinct odor qualities by humans. Each analyte produced a different, characteristic response pattern on the electronic nose array, with the signal strength on each detector reflecting the relative binding of the odorant into the various conducting polymer composites of the detector array. A ''human perceptual space'' was defined by reference to English language descriptors that are frequently used to describe odors. Data analysis techniques, including standard regression, nearest-neighbor prediction, principal components regression, partial least squares regression, and feature subset selection, were then used to determine mappings from electronic nose measurements to this human perceptual space. The effectiveness of the derived mappings was evaluated by comparison with average human perceptual data published by Dravnieks. For specific descriptors, some models provided cross-validated predictions that correlated well with the human data (above the 0.60 level), but none of the models could accurately predict the human values for more than a few descriptors.

Keywords: Electronic nose; Conducting polymer composites; Odor quality; Human perception; Olfaction; Regression models; Feature selection

1. Introduction

Arrays of broadly responsive vapor detectors are attracting increasing interest as ''artificial noses'' [1–3]. Like the receptors in the mammalian olfactory system [4], each detector in an ''artificial nose'' responds to more than one analyte, and each analyte elicits a response from more than one detector [1–3]. Pattern recognition algorithms are then used to classify, identify, and in some cases quantify, an analyte in the vapor phase. One motivation for studying such arrays is eventually to learn enough about the process of olfaction to construct a man-made, functional analogue of a mammalian olfactory system [5].

Perhaps the ultimate challenge for an artificially-constructed olfactory system is to mimic faithfully the mapping of an odorant-induced detector response pattern to the quality of an odor, e.g. to its ''minty-ness'', as perceived by a human. This task is difficult because the human olfactory system is highly nonlinear in many respects. For example, perceived odor intensity is a nonlinear function of analyte concentration [6]. In addition, qualitatively different human percepts are often produced by varying the concentration of a given odorant. Cross-adaptation, masking, and other processes involved with the human perception of odor mixtures [7] further complicate the signal processing involved in olfaction [8]. A further level of complexity results because humans are variable genetically in their perception of many odorants [9]. Thus, any static, generically-constructed, artificial olfactory device could at best capture some average human perceptual processes for a representative set of odorants. Of course, this does not eliminate the possibility of a ''trainable'' device that could be tuned to match the perceptual profile of a specific individual, however, developing such a system poses yet another set of challenges.

The artificial nose implementation that was used in this study consists of an array of conducting polymer composites, in which each detector material of the array has regions of a conductive material interspersed into regions of an insulating organic polymer [10,11]. The conductive material is typically carbon black although it could also be an

* Corresponding author. Tel.: +1-626-395-6335; fax: +1-626-795-7487.
E-mail address: nslewis@caltech.edu (N.S. Lewis).

0925-4005/01/$ – see front matter
PII: S0925-4005(00)00645-6

inorganic metal or an organic electrically-conductive polymer. Sorption of an odorant into the polymer produces a swelling of the polymer film, which in turn leads to an increase in the dc electrical resistance of the detector. The electrical output signals from an array of such detectors are then transferred to a central processing unit for odorant analysis and classification. This implementation of an electronic nose was chosen for study because it is readily investigated experimentally [11], allows inclusion of a chemically diverse set of detectors [12], has been shown to parallel mean human olfactory detection threshold behavior for several classes of organic vapors [10], and has been shown in selected test cases to parallel human and monkey olfaction in the positive correlation between its discrimination ability and the chemical dissimilarity between members of a pair of odorants [13].

The specific focus of the current study was to investigate whether the responses of an array of such detectors could be used to predict accurately the perceived quality of an odorant as reported by human panelists. Only chemically pure single-component analytes were investigated, due to the further complications described above relating to the human olfactory perception of odorant mixtures. Data on human perception of odor quality for a variety of odorants were obtained from tabulations available in the literature [14]. Electronic nose responses were collected for a selected subset of the same compounds. The odorants investigated in our work included molecules that are chemically similar but which are perceived as being different by humans, as well as molecules that are chemically quite distinct from each other. Successful odor quality prediction is critical not only for meeting the intellectual challenge of constructing an artificial olfactory system, but also for many industrial quality control applications of an artificial nose in which product assessments (good/bad) must be made with respect to human perception rather than with respect to changes in the chemical composition of the odors of concern [15–17].

2. Experimental

2.1. Chemicals and data collection

Twenty-one odorants (Table 1) were evaluated in this work. All chemicals were obtained from Aldrich Chemical Corp. and were used as received. Sets of chemically homologous odorants (for example, a series of straight-chain alcohols, a series of aliphatic esters, a series of straight-chain aliphatic acids, a series of benzene derivatives, etc.) were chosen such that the odors were associated with common, but not identical, human odor descriptors both within a set and between sets of odorants. A total of 20 insulating polymers were used to form the carbon black/polymer composite detectors in the electronic nose (Table 2). Detectors were fabricated as described previously [12].

Table 1
Odorants used in this study

1	1-Butanol
2	1-Hexanol
3	1-Heptanol
4	1-Octanol
5	Ethyl propionate
6	Ethyl butyrate
7	Propyl butyrate
8	Amyl butyrate
9	Isopentyl acetate
10	Pentanoic acid
11	Hexanoic acid
12	Toluene
13	Anisole
14	Phenyl ethanol
15	Phenyl acetylene
16	Tetrahydrothiophene
17	Thiophene
18	Butanoic acid
19	Pyridine
20	Citral
21	Limonene

All odorant exposures were performed using a computer-controlled vapor generation and control system that regulated the identity, concentration, exposure time, and flow rate of the analyte above the detectors [18]. The experimental protocol for each odorant exposure was 5 min of clean air flow, followed by 5 min of air flow containing the odorant at a partial pressure corresponding to 5% of its vapor pressure, followed by another 5 min of clean air flow. The detectors were exposed to each odorant a minimum of 10 times. Analyte identities were varied in random order between each exposure.

Table 2
Polymers contained in the detectors of the carbon black/polymer composite electronic nose array

Detector	Polymer
1	Poly(4-vinyl phenol)
2	Poly(*N*-vinylpyrrolidone)
3	Poly(sulfone)
4	Poly(methyl methacrylate)
5	Poly(caprolactone)
6	Poly(ethylene-*co*-vinyl acetate), 82% ethylene
7	Poly(ethylene oxide)
8	Poly(ethylene)
9	Poly(vinylidene fluoride)
10	Poly(ethylene glycol)
11	Poly(vinyl acetate)
12	Poly(styrene)
13	Poly(butadiene)
14	Poly(styrene-*co*-allyl alcohol)
15	Poly(α-methylstyrene)
16	Hydroxypropyl cellulose
17	Poly(styrene sulfonic acid)
18	Poly(carbonate bisphenol A)
19	Poly(epichlorohydrin)
20	Poly(styrene-*co*-butadiene)

Only the steady-state response data were used in analysis of the electronic nose array signals. Specifically, the data were reduced to produce a $\Delta R/R_b$ value for each detector, where R_b is the drift-corrected baseline response of the detector during the analyte exposure period and ΔR is the steady-state differential resistance response of the detector with respect to the value of R_b. The data for each exposure were then expressed as a response vector, with each component of the vector corresponding to the $\Delta R/R_b$ value of a particular detector. The results from individual exposures to a given odorant were averaged to produce a single twenty-dimensional vector that described the response of the detector array to each odorant. The electronic nose measurements were, thus, reduced to a 21×20 matrix, $\boldsymbol{M}$, whose rows corresponded to different odorants and whose columns corresponded to different detectors.

The measurement data can be visualized to some extent by performing principal components analysis (PCA) on the raw twenty-dimensional measurement space and projecting the data onto the two leading principal component directions. Fig. 1 shows all of the odorants in this PCA space. The numerical label next to each point can be translated into a chemical name using Table 1. Note, for example, that points 1–4, which correspond to straight chain aliphatic alcohols, are well-clustered in the principal components space.

2.2. Perceptual odor quality

Perceptual odor quality values for humans were obtained from Dravnieks' *Atlas of Odor Character Profiles*. [14]. In Dravnieks' study, over 100 people of both sexes, spanning a wide range of ages, and including smokers as well as non-smokers, were evaluated. The rationale for using such a diverse group of panelists was apparently to insure that the reported percepts would be consistent with those of the population at large. Each participant was asked to smell a collection of odorants and was instructed to assign a score from 0 through 5 to each of 146 different descriptors (adjectives) that are used in the English language to describe odors. For example, a panelist could give an odorant a score of 3 in the ''etherish, anesthetic'' category, 4 in the ''minty'' category, and 0 in each of the remaining categories. Scores are intended to reflect the degree to which the panelist believes that a descriptor is appropriate for a given odorant, with a value of 0 meaning not appropriate. As described by Dravnieks, care was exercised in the experimental procedure to insure that artificial biases were not introduced into the results. Of the 146 descriptors considered by Dravnieks, the seventeen descriptors listed in Table 3 were selected for use in our study based on the frequency and extent to which they were used by the panelists to describe our selected set of test odorants (Table 1).

For the purposes of our study, a limitation with Dravnieks' *Atlas* is that only averages across the entire group of panelists are provided, so score profiles for individual participants are not available. Also, the variance (or the distribution) of scores given to a particular odorant–descriptor pair was not reported. Instead, the available data for each odorant–descriptor pair consist of two quantities: percentage of usage and percentage of applicability. The usage indicates

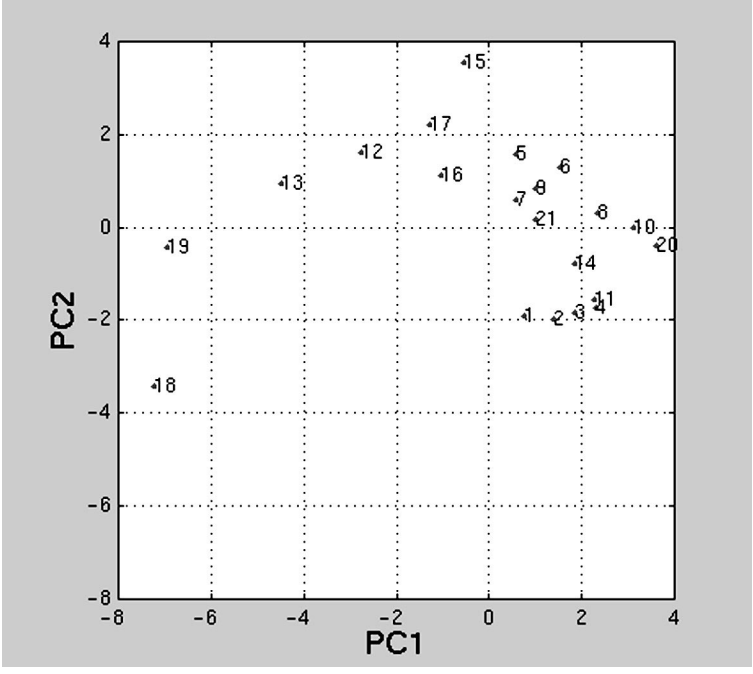

Fig. 1. Conducting polymer composite detector response data for the 21 odorants in two-dimensional principal component space. The numerical labels associated with each point correspond to the analytes listed in Table 1.

Table 3
Scent descriptors used in this study[a]

1	Fruity (citrus)
2	Fruity (non-citrus)
3	Floral
4	Minty
5	Etherish
6	Gasoline
7	Sharp, pungent
8	Oily
9	Putrid, foul, decayed
10	Woody
11	Sweet
12	Herbal
13	Musty
14	Medicinal
15	Sour
16	Paint
17	Sweaty

[a] From [14].

the percentage of panelists who assigned a non-zero score to the descriptor for the given odorant. This quantity can be interpreted as a probability value. The percentage of applicability is the geometric mean of the usage and the average score level assigned by the panelists.

Dravnieks suggests that the percentage of applicability is "the most equitable indicator of the descriptor applicability" to human perception. However, we have observed that the usage, score level, and applicability are all highly correlated. For example, Fig. 2 shows a plot of the score level versus usage for the chemicals and descriptors used in our study. A clear functional relationship is apparent between the score level and usage. However, because the relationships between the quantities are nonlinear, it is not clear which, if either, quantity will be more easily predicted from the electronic nose measurements. Feature vectors for the electronic nose response data and values of the organoleptic descriptors for the 21 odorants of Table 1 are available at http://www-aig.jpl.nasa.gov/mls/home/burl/data/odor_-quality/.

2.3. Data analysis

The goal is to assess the degree to which the response of the electronic nose to a given odorant can quantitatively predict the usage, score level, and/or applicability that would be provided on average by a group of human panelists for each of the seventeen descriptors. The different approaches that have been explored towards this goal are described below.

2.3.1. Standard regression and nearest-neighbor approaches

2.3.1.1. Standard linear regression. With the electronic nose measurements expressed as a 21×20 matrix $\boldsymbol{M}$ and the human panelist ratings for a particular descriptor/quantity expressed as a 21×1 vector $\boldsymbol{h}$, the problem reduces to finding a vector-valued function f such that the prediction $\boldsymbol{h}_p = f(\boldsymbol{M})$ is approximately equal to $\boldsymbol{h}$. Of the many possible classes of functions f, we restricted our attention to linear ($f(\boldsymbol{M}) = \boldsymbol{M}\boldsymbol{w}$) and affine ($f(\boldsymbol{M}) = \boldsymbol{M}\boldsymbol{w} + \boldsymbol{w}_0$) models, as well as "clipped" linear and affine models

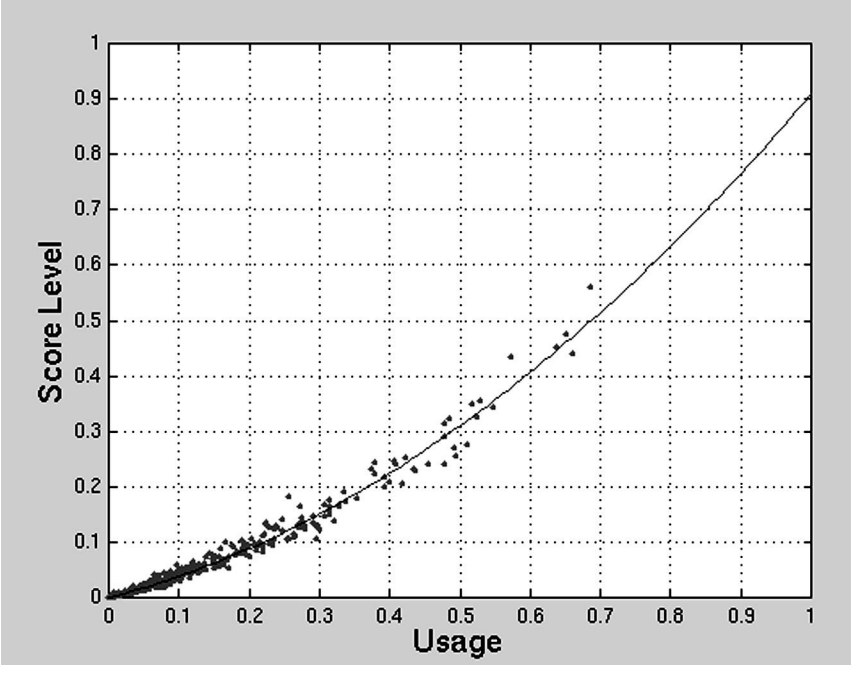

Fig. 2. Plot of the average human-assigned score level vs. percent usage for each chemical/descriptor pair in our study. The two quantities are clearly highly correlated, with s approximately equal to $0.33u + 0.58u^2$.

that incorporate a nonlinearity on the output to confine predictions to the range [0, 1], the same range as for the human perceptual data. Note that the affine form can be reduced to the linear case by augmenting the measurement matrix $\boldsymbol{M}$ with an additional column of ones and increasing the dimensionality of the weight vector by one. Hence, the linear form can be focused on without loss of generality.

The weight vector that minimizes the mean squared error between the predictions and the targets is well known:

$$\boldsymbol{w} = (\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{h} \quad (1)$$

Given a 1×20 measurement vector $\boldsymbol{m}$ for an odorant whose quality descriptors are to be predicted, the regression model predicts: $\boldsymbol{h}_p = \boldsymbol{m}\boldsymbol{w}$. Note that for numerical calculations the weight vector $\boldsymbol{w}$ is generally obtained by solving the equation $(\boldsymbol{M}'\boldsymbol{M})\boldsymbol{w} = \boldsymbol{M}'\boldsymbol{h}$ via LU decomposition and backsubstitution, rather than by computing the inverse of $(\boldsymbol{M}'\boldsymbol{M})$ [19].

2.3.1.2. Nearest neighbor models. A different type of predictor is based on the idea of nearest neighbors. If two odorants have similar electronic nose patterns, then it might be expected that the descriptors provided by human observers for these two odorants would also be very similar. The nearest neighbor model makes a prediction for a test odorant having an electronic nose signature $\boldsymbol{m}$ by first finding the row in $\boldsymbol{M}$ that is most similar to $\boldsymbol{m}$ (e.g. using a Euclidean distance metric). The human perceptual value for this nearest neighbor is then taken as the prediction for the test odorant. The reference library of electronic nose measurements imparts a partitioning of the measurement space into distinct regions in which a single library example will be the nearest neighbor of any example that falls into the region. The predictions for any example falling in this region will be the same as for the library example.

2.3.2. Basic testing procedure

A leave-one-out cross-validation procedure was used to evaluate the performance of the predictors. In this approach, the electronic nose signals for one odorant were withheld for use in testing. The remaining $N - 1(=20)$ odorants were then used to train the model (i.e. to determine the weight vector $\boldsymbol{w}$ in the regression approach or to serve as the reference library in the nearest neighbor approach). This process was repeated N times, with each odorant serving a turn as the holdout example. The predictions of both models for all odorants were then compared to the actual human target data for the odorants of interest.

3. Results

3.1. Performance of standard regression and nearest neighbor approaches

Fig. 3a shows the correlation coefficients between the predictions of the clipped linear regression model and the odor targets. The integer-valued x-coordinates (values from 1 to 17) represent the particular descriptors indicated in Table 3. The y-axis shows the correlation coefficient achieved for each descriptor/quantity (u = usage, s = score, a = applicability). Overall, the regression predictor performed poorly. Only 3 of the 17 descriptors (''floral'', ''sour'', and ''paint'') have at least one quantity predicted above the 0.60 level. The median correlation coefficient across all descriptors/quantities for this model is 0.21. The predictability of usage, score, or applicability were all fairly similar, having median values of 0.21, 0.21, and 0.22, respectively.

Fig. 3b presents the correlation coefficients between the predictions of the nearest neighbor model and targets. Overall, this predictor also performed poorly. Only 3 of the 17 descriptors (''fruity non-citrus'', ''woody'', and ''sweet'') have at least one quantity predicted above the 0.60 level. It is interesting to note that the descriptors that are predicted well by the regression model do not intersect with the descriptors that are predicted well by the nearest-neighbor model. The median correlation coefficient for the nearest neighbor model across all descriptors/quantities is 0.25.

Several variations in the preprocessing of the electronic nose measurements were explored, including normalization of each chemical signature to remove concentration information, and auto-scaling the detector values to remove means and equalize variances. Concentration normalization resulted in a slight increase in the performance of the regression model (median = 0.32) and a modest decrease in the performance of the nearest neighbor model (median = 0.07; only one descriptor above 0.60). Auto-scaling does not affect the regression model, but resulted in degraded performance for the nearest neighbor model (median = 0.07).

3.2. More sophisticated regression and feature selection approaches

The poor predictive abilities of the standard linear regression and nearest neighbor models should not be surprising given that the number of examples (odorants) is comparable to the number of dimensions (detectors). In fact, under the leave-one-out cross validation procedure, 20 examples and 20 dimensions were present. Assuming that the 20×20 measurement matrix is non-singular, the regression model, thus, provides a unique weight vector that *exactly* maps the training measurements to the training targets. However, since the measurements (and targets) include noise, this will clearly lead to overfitting and poor generalization on new examples. There are several approaches available for such situations: (1) ridge regression to improve the conditioning of the $\boldsymbol{M}'\boldsymbol{M}$ matrix; (2) principal components regression to orthogonalize the measurements and discard noise; (3) partial least squares regression (PLS), also used to ignore redundant detector measurements and to discard

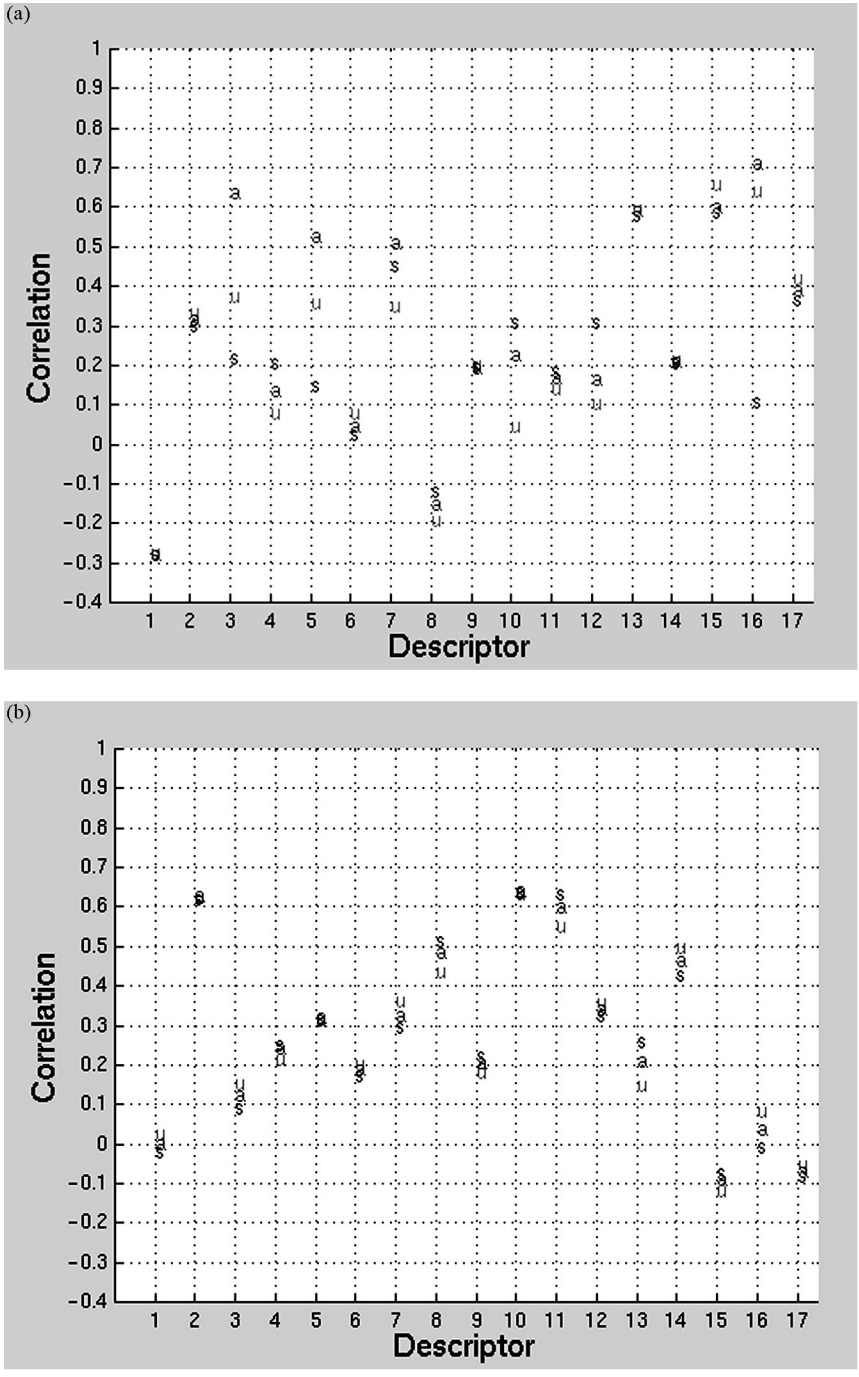

Fig. 3. (a) Correlation between clipped linear regression predictions and human target values. The symbols "*u*", "*s*", and "*a*" represent usage, score, and applicability, respectively. Only descriptors 3, 15, and 16 (floral, sour, paint) are predicted above the 0.6 level. (b) Correlation between nearest neighbor predictions and human target values. Only descriptors 2, 10, and 11 (fruity non-citrus, woody, sweet) are predicted above the 0.6 level.

noise; and (4) feature subset selection algorithms to reduce the number of features used by the predictor.

3.2.1. Ridge regression

One view of standard regression is that the matrix $\boldsymbol{M}'\boldsymbol{M}$ is a covariance matrix (if the mean value of each detector is first removed from $\boldsymbol{M}$). The standard regression solution, therefore, attempts to estimate the covariance between detectors based on data from a very limited number of examples. Such estimates can often be improved by shrinking toward the identity matrix [20]. Mathematically, $\boldsymbol{M}'\boldsymbol{M}$ is replaced by $(1-\gamma)\boldsymbol{M}'\boldsymbol{M} + \gamma\,\mathrm{tr}[\boldsymbol{M}'\boldsymbol{M}]/\mathrm{ns}\boldsymbol{I}$, where ns is the number of detectors and γ a free parameter that controls the amount of regularization.

To select the proper value of γ, the following nested cross-validation procedure was used:

```
for x = 1: nx (% holdout example x for final testing)
  % Perform selection of γ*
  for γ = 0: dγ:_1
    for y = 1: nx−1
        Perform cross-validated evaluation with
        parameter γ using all examples excluding x.
    end
  end
  Choose γ* = _best γ_value based on inner loop
  results.
  Train regularized regression model using γ* and all
  examples excluding x.
  Use regularized regression model to predict human
  value for example x.
end
Compare predicted values with targets.
```

Experimentally, a value of $d\gamma = 0.01$ was used. Due to the small number of examples available in the inner loop, the γ loop was started at $\gamma = d\gamma$ rather than at 0. The inner loop cross-validation generally selected γ^* values of 0.01 or 0.02, with one value as high as 0.04; however, the outer loop performance using the γ^* regularized regression model turned out to be slightly worse than using the standard regression model with $\gamma = 0$.

3.2.2. Principal components regression

When the measurements from different detectors are highly correlated or are noisy, the presence of the inverse in Eq. (1) often precludes obtaining a good weight vector through standard linear regression. One approach to resolve this problem is to perform a principal components analysis on $\boldsymbol{M}'$ to determine the directions that have the most variance. The data are then projected onto this reduced dimensional subspace, and directions with smaller variance are presumed to correspond to noise and are discarded. The target values are then predicted from the projected subspace rather than from the original data. In the chemometrics literature, this approach is known as principal components regression (PCR), and the projected data are commonly referred to as the ''score matrix''. In many cases, PCR provides an alternative solution to the regression problem that may be better-behaved than standard regression.

For the data of concern in this work, however, the PCR approach performed poorly. In fact, the median cross-validated correlation coefficient across all descriptors/quantities was worse than for the raw 20-dimensional data for all values of k, except $k = 18$, for which the result obtained using PCR was better by a small amount than that obtained using standard regression.

3.2.3. Partial least squares regression

Partial least squares regression (PLS) is another method that provides an alternative solution to the regression problem. The PLS method is similar to PCR, except that both the target vector and the measurements are used to determine a lower-dimensional subspace from which the predictions will be made. Determination of the subspace is accomplished through an iterative procedure as described in the literature [21].

The same leave-one-out cross-validation testing scheme described above was used to evaluate the effectiveness of PLS regression in predicting the values of the human-provided descriptors from the conducting polymer composite vapor detector measurements. On average, the PLS predictor performed slightly better than standard regression, producing median correlation values for usage, score, and applicability of 0.30, 0.29, and 0.30, respectively. However, as shown in Fig. 4, only one descriptor (#10, ''woody'') was predicted above the 0.60 level.

3.2.4. Feature subset selection

A different approach to possibly improve predictions based on models derived from limited data is to consider subsets of the raw detectors. For 20 detectors, there are $n\text{fs} = (2\hat{}20) - 1$ (slightly over 1 million) possible subsets. Various feature subset selection algorithms that use heuristics to guide the search for good subsets have been developed in the machine learning and pattern recognition communities [22,23]. In our study, however, we have avoided the use of such techniques and instead have used large amounts of computation to exhaustively evaluate *every* subset of features. By doing so, we avoid the uncertainty inherent in not knowing whether the search heuristics lead to significant degradations in achievable performance. In other words, exhaustive subset evaluation is an academic approach that will enable us to study the prediction problem without adding confusion from approximations. However, for significantly larger array sizes and/or with limited computational resources, the heuristic approaches may be the only feasible way to proceed.

To evaluate a particular subset of features, the following nested cross-validation procedure was considered:

```
for x = 1: nx (% holdout example x for final testing)
  % Perform feature selection
  for fs = 1: nfs (%)
    for y = 1: nx−1 (%)
      Perform cross-validated evaluation of feature
      set f using all examples excluding x.
    end
  end
  Choose feature set fs* based on some selection
  criteria.
  Train regression model using feature set fs* and all
  examples excluding x.
  Use regression model to predict human value for
  example x based on feature set fs*.
end
Compare predicted values with targets.
```

Clearly, the computational requirements of this procedure are quite demanding. To make the execution feasible, the

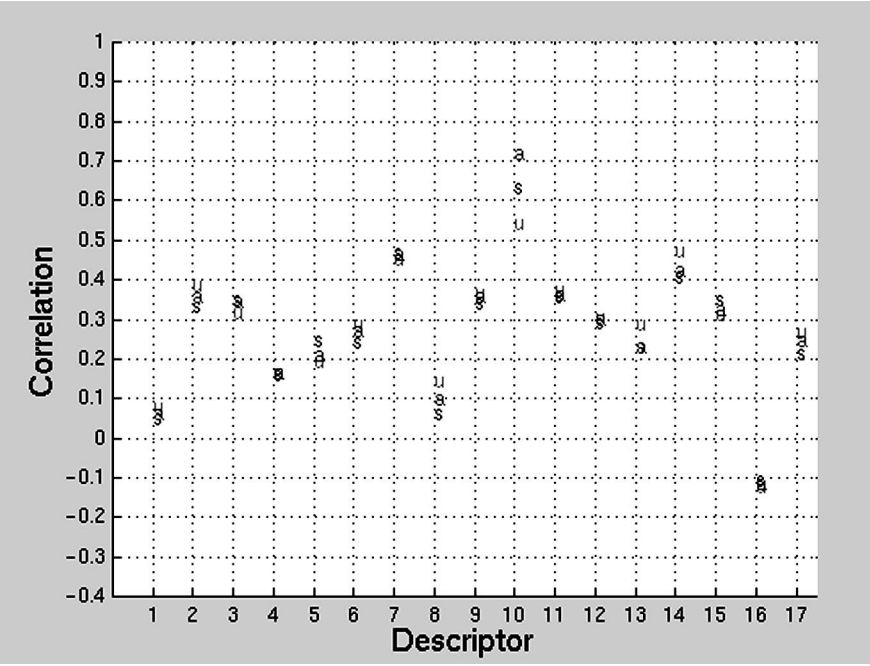

Fig. 4. Correlation between partial least squares predictions and human target values. The symbols "*u*", "*s*", and "*a*" represent usage, score, and applicability, respectively. Only descriptor 10 (woody) is predicted above the 0.6 level.

outer *x*-loop was parallelized across a dozen or so SUN Ultra 60 (dual CPU) and SUN Ultra 2 (single CPU) workstations, with each processor handling one pass through the body of the loop. Results were saved to disk at the end of the fs-loop and later merged.

The inner-most cross-validation loop (the *y*-loop) requires the computation of $(\boldsymbol{M}'_{xy}\boldsymbol{M}_{xy})^{-1}\boldsymbol{M}'_{xy}\boldsymbol{H}_{xy}$, where $\boldsymbol{M}_{xy}$ consists of all the measurements excluding examples *x* and *y* and $\boldsymbol{H}_{xy}$ consists of the corresponding human values. $\boldsymbol{H}_{xy}$ is a matrix with each column containing the human data for a particular descriptor/quantity; thus $\boldsymbol{H}_{xy}$ has $51 = 17 \times 3$ columns. Each pass through the *y*-loop appears to require an inversion of the matrix $\boldsymbol{M}'_{xy}\boldsymbol{M}_{xy}$, requiring work proportional to the third power of the number of features involved. However, the Sherman–Morison–Woodbury formula, which specifies how a rank one correction to a matrix affects the inverse, can be applied so that only one inversion is required, rather than $nx-1$ inversions. The main idea is to compute the inverse of $\boldsymbol{M}'_x\boldsymbol{M}_x$ (measurements excluding example *x*) once before the *y*-loop and then note that $\boldsymbol{M}'_{xy}\boldsymbol{M}_{xy}$ is related to $\boldsymbol{M}'_x\boldsymbol{M}_x$ by a rank one correction involving example y. A correction can also applied to the quantity $\boldsymbol{M}'_x\boldsymbol{H}_x$ to obtain $\boldsymbol{M}'_{xy}\boldsymbol{H}_{xy}$ with less work than recomputing on each pass through the loop. This procedure greatly reduced the computational complexity.

A number of possible selection criteria could be applied at the conclusion of the loop over feature sets. The most straightforward criterion is to choose for each descriptor/quantity the feature set fs^* that results in the best inner-loop (*y*-loop) cross-validated correlation. Other inner-loop assessment metrics such as mean square error, maximum absolute error, and maximum relative error can also be considered. Intuition suggests, however, that models based on a smaller number of features may generalize better to new examples, so it seems reasonable to bias the feature selection strategy towards feature sets having smaller numbers of features. For example, rather than choosing the feature set providing the best inner loop performance, the best feature set having exactly (or at most) a predetermined number of features was chosen. Similar biasing rules such as selecting the smallest-sized feature set whose performance on the inner loop exceeds a given threshold, or whose performance on the inner loop is "close enough" to the best performance achieved by any feature set, may also be considered.

Table 4 shows the results of the feature selection experiment. For each iteration of the outer loop, the inner loop cross-validated correlation scores were used to select the best set of *k* detectors. Thus, *nx* best detector sets could potentially be identified for a given *k* value. The median inner-loop correlation values over these best detector sets are shown in column 2 of the table, where the median is taken over all descriptors/quantities as well as over the best detector sets. The score of the best inner loop models when applied to the outer loop holdout examples is shown in column 3 (median across all descriptors/quantities).

The best cross-validated correlation score that could have been achieved by a single feature set of size *k* is shown in column 4. Note that these scores were obtained by evaluating all feature sets of size *k* on the outer loop task; then the set that produced the best results was chosen and its results were reported in column 4 (i.e. feature subset selection was based on the test set).

Table 4
Results of feature selection experiments

Number of detectors (k)	Median inner loop score	Outer loop score	Idealized outer loop score
1	0.37	0.19	0.40
2	0.48	0.10	0.50
3	0.55	0.10	0.57
4	0.61	0.10	0.63
5	0.65	0.13	0.65
6	0.70	0.12	0.74
7	0.74	0.11	0.77
8	0.76	0.13	0.77
9	0.77	0.15	0.78
10	0.82	0.21	0.82
11	0.86	0.25	0.87
12	0.88	0.30	0.90
13	0.90	0.26	0.90
14	0.91	0.25	0.91
15	0.92	0.26	0.88
16	0.93	0.17	0.86
17	0.96	0.21	0.82
18	0.99	0.08	0.84
19	0.39	0.09	0.63
20	0.02	0.18	0.25

The fact that the column 4 numbers are large is interesting, because it indicates that certain linear regression models using subsets of features can accurately predict the human data in cross-validation tests. For example, Table 5 shows the smallest models that yielded cross-validated correlation above 0.6 for the *usage* quantity of each descriptor. However, in Table 4 the large margin between column 2 and column 3 casts doubt on whether the models that produced column 4 will generalize well to new data. More specifically, column 2 represents the expected performance based on the models that worked best in the inner-loop cross-validation (train on $nx-2$ examples, test on 1). Column 3 indicates the performance of these ''good'' inner-loop models when applied to the outer-loop holdout example. Clearly, the performance of the models was highly degraded under these conditions. Even though a good inner-loop model could be obtained, the existence of such a model did not insure that the same model would work well in the outer-loop cross-validation. Analogously, some models worked well on the outer-loop cross-validation, but this does not insure that the same models will work well for new data.

Table 5
Smallest models yielding a cross-validated correlation above 0.6 for the descriptor usage quantity

	Descriptor	k	Model[a]
1	Fruity citrus	15	5, 7, 11, 16, 19
2	Fruity non-citrus	5	5, 12, 13, 15, 18
3	Floral	8	2, 3, 5, 10, 12, 14, 17, 19
4	Minty	9	2, 4, 7, 9, 10, 11, 14, 18, 20
5	Etherish	4	1, 5, 6, 13
6	Gasoline	3	1, 5, 13
7	Sharp, pungent	2	4, 14
8	Oily	2	16, 20
9	Putrid, foul, decayed	5	1, 6, 15, 18, 20
10	Woody	4	11, 13, 15, 19
11	Sweet	4	7, 12, 15, 17
12	Herbal	2	6, 14
13	Musty	2	6, 20
14	Medicinal	7	3, 4, 5, 7, 8, 11, 12
15	Sour	6	1, 2, 5, 15, 16, 18
16	Paint	4	3, 6, 13, 14
17	Sweaty	2	2, 20

[a] Underlined values indicate excluded detectors.

4. Discussion

This study considered single-component odorants consisting of simple organic vapors without significant aroma activity (arguably the simplest case). For specific descriptors, some models provided cross-validated predictions that correlated well with the human data (above the 0.60 level), but none of the models could accurately predict the human values for more than a few descriptors.

The models based on feature subset selection were especially intriguing. Relatively small subsets of detectors (e.g. as listed in Table 5) in some cases provided good cross-validated predictions of most of the human descriptors. However, because these subsets could not be identified through a rigorous model selection procedure, the results may not generalize to new data. Further evidence for this conclusion is given by the model selection procedure itself. The large discrepancy between the inner loop cross-validated correlation values and the resulting outer loop cross-validated correlation values (i.e. comparison of column 2 and column 3 in Table 4) indicates that at least some ''good'' inner-loop models perform poorly on new data.

The term ''overfitting'' is typically applied to describe the situation in which models or parameters are adjusted to fit a training set to the best degree possible. In our case, the feature subset selection procedure uses the criteria ''find the best subset of features smaller than size k'', which provides a bias toward smaller models, i.e. the decision as to which feature subset to choose is not based solely on optimizing the predictions to the targets. Hence, we are not strictly overfitting during the subset selection. The poor generalization results, however, may indicate that the bias is inadequate to lead the model selection procedure to good subsets that will generalize.

Once a feature subset is selected, the standard regression solution attempts to find the best set of weights that fits the training set, i.e. it is overfitting. It does not appear that this weight overfitting is catastrophic, however, because it also is present for column 4 of Table 4.

There are several plausible explanations for why human perception of odor quality could not be predicted reliably from the conducting polymer composite electronic nose

signals. First, the number and diversity of receptors in the artificial nose was significantly (1–2 orders of magnitude) more limited than in humans. Second, linear and affine models may be too simplistic to enable human percepts of even single component organic vapor odorants to be predicted well from conducting polymer composite electronic nose signals. This may reflect the fact that these data models have no physically-based relationship to the neuronal connections and signal processing involved in olfactory perception. A third possibility is that the form of the data models may be adequate for the limited task considered, but the parameters of these models could not be estimated reliably enough from the amount of data available. A fourth possibility is that the feature subset selection procedure did not have enough data to identify feature subsets that would provide good generalization ability.

In any case, it is clear that the situation would be significantly worse for odorants that are mixtures of pure compounds, because human odor perception of mixtures is often not linearly related to the mole fraction of the individual components of the mixture. Similarly, for mixtures of odorants that contain aroma-active compounds which are detectable at very low concentration levels in the human nose but which would not even produce signals above the noise level of the current conducting polymer composite detectors, no correlation would be expected between the conducting polymer composite vapor detector response and that of human odor perception.

Given the complex, nonlinear characteristics of human olfaction, it is not surprising that the analogy between the electronic nose and the human olfactory system only extends to the design principle that both systems utilize arrays of broadly responding, cross-reactive detectors. Hence, the primary contribution of the present study is to advance the idea of formulating, developing, and testing models of olfactory perceptual processing using artificial data sets such as those generated by the electronic nose in place of spike train data measured only with great difficulty on biological olfactory receptors. In this respect, one advantage of the conducting polymer composite-based electronic nose used in this study is that characteristic response patterns that are essentially independent of the concentration of the analyte presented to the detectors can be obtained. Thus, one degree of freedom (choice of analyte concentration) can be eliminated, unlike the situation for metal-oxide detectors, dye-impregnated polymers on optical fibers, and other detectors that exhibit response patterns that are a function of concentration of the analyte of interest [3]. A secondary contribution is that the results based on simple numerical models provide a benchmark against which more sophisticated models can be judged, and strongly suggest that significant system architecture changes and highly increased data analysis algorithm complexity are needed before any artificial model could provide a robust predictive ability for human odor quality descriptors on even a simple set of test odorants. The correlation between the predictions for certain odor descriptors and the human perceptual data are definitely interesting, but more extensive experiments would be required to assess the validity of these predictors for other odorants.

Acknowledgements

We acknowledge a MURI supported through the Army Research Office for support of this work, and thank W. Cain of UCSD and D. DeCoste for numerous helpful discussions.

References

[1] H.V. Shurmer, An electronic nose — a sensitive and discriminating substitute for a mammalian olfactory system, IEEE Proc.-G Circ. Dev. Syst. 137 (1990) 197–204.
[2] J.W. Gardner, P.N. Bartlett, A brief-history of electronic noses, Sens. Actuators B 18 (1994) 211–220.
[3] K.J. Albert, N.S. Lewis, C.L. Schauer, G.A. Sotzing, S.E. Stitzel, T.P. Vaid, D.R. Walt, Cross-reactive chemical sensor arrays, Chem. Rev. 100 (2000) 2595–2626.
[4] B. Malnic, J. Hirono, T. Sato, L.B. Buck, Combinatorial receptor codes for odors, Cell 96 (1999) 713–723.
[5] T.C. Pearce, Computational parallels between the biological olfactory pathway and its analogue 'The Electronic Nose'. 2. Sensor-based machine olfaction, Biosystems 41 (1997) 69–90.
[6] F.T. Schiet, W.S. Cain, Odor intensity of mixed and unmixed stimuli under environmentally realistic conditions, Perception 19 (1990) 123–132.
[7] D.G. Laing, H. Panhuber, M.E. Willcox, E.A. Pittman, Quality and intensity of binary odor mixtures, Physiol. Behav. 33 (1984) 309–319.
[8] J.D. Pierce Jr., X.N. Zeng, E.V. Aronov, G. Preti, C.J. Wysocki, Cross-adaptation of sweaty-smelling 3-methyl-2-hexenoic acid by a structurally-similar, pleasant-smelling odorant, Chem. Senses 20 (1995) 401–411.
[9] S. Ayabe-Kanamura, I. Schicker, M. Laska, R. Hudson, H. Distel, T. Kobayakawa, S. Saito, Differences in perception of everyday odors: a Japanese–German cross-cultural study, Chem. Senses 23 (1998) 31–38.
[10] B.J. Doleman, E.J. Severin, N.S. Lewis, Trends in odor intensity for humans and electronic noses: relative roles of odorant vapor pressure vs. molecularly specific odorant binding, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 5442–5447.
[11] M.C. Lonergan, E.J. Severin, B.J. Doleman, S.A. Beaber, R.H. Grubbs, N.S. Lewis, Array-based vapor sensing using chemically sensitive, carbon black-polymer resistors, Chem. Mater. 8 (1996) 2298–2312.
[12] B.J. Doleman, M.C. Lonergan, E.J. Severin, T.P. Vaid, N.S. Lewis, Quantitative study of the resolving power of arrays of carbon black-polymer composites in various vapor-sensing tasks, Anal. Chem. 70 (1998) 4177–4190.
[13] B.J. Doleman, N.S. Lewis, Comparison of odor detection thresholds and odor discriminablities of a conducting polymer composite electronic nose vs. mammalian olfaction, Sens. Actuators B 72 (2001) 41–50.
[14] A. Dravnieks, Atlas of Odor Character Profiles, American Society for Testing and Materials, Baltimore, 1985.
[15] C. Di Natale, A. Macagnano, R. Paolesse, R. Mantini, E. Tarizzo, F. Sinesio, et al., Electronic nose and sensorial analysis: comparison of performance in selected cases, Sens. Actuators B: Chem. 50 (1998) 246–252.

[16] B.-I.L. Willing, A. Brunders, I. Lundstrom, Odour analysis of paperboard, the correlation between human senses and electronic sensors using multivariate analysis, Packaging Technol. Sci. 11 (1998) 59–67.
[17] Y. Blixt, E. Borch, Using an electronic nose for determining the spoilage of vacuum-packaged beef, Int. J. Food Microbiol. 46 (1999) 123–134.
[18] E.J. Severin, B.J. Doleman, N.S. Lewis, An investigation of the concentration dependence and response to analyte mixtures of carbon black-insulating organic polymer composite vapor detectors, Anal. Chem. 72 (2000) 658–668.
[19] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes in C, Cambridge University Press, Cambridge, 1988.
[20] J.H. Friedman, Regularized discriminant analysis, J. Am. Stats. Assoc. 84 (405) (1989) 165–175.
[21] P. Geladi, B.R. Kowalski, Partial least squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[22] F. Ferri, P. Pudil, M. Hatef, J. Kittler, Comparative study of techniques for large-scale feature selection. IV. Pattern Recognition in Practice, in: E.S. Gelsema, L.N. Kanal (Eds.), Elsevier, Amsterdam, 1994, pp. 403–413.
[23] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.