Sensors and Actuators B 174 (2012) 114–125

Contents lists available at SciVerse ScienceDirect

Sensors and Actuators B: Chemical

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

# Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine

Lei Zhang[a,*], Fengchun Tian[a], Hong Nie[b], Lijun Dang[a], Guorui Li[a], Qi Ye[a], Chaibou Kadri[a]

[a] College of Communication Engineering, Chongqing University, 174 ShaZheng Street, ShaPingBa District, Chongqing 400044, China
[b] Academy of Metrology and Quality Inspection, Chongqing 401123, China

ARTICLE INFO

*Article history:*
Received 9 February 2012
Received in revised form 19 June 2012
Accepted 3 July 2012
Available online xxx

*Keywords:*
Electronic nose
Classification
Multi-class problem
Hybrid support vector machine
Fisher linear discrimination analysis

ABSTRACT

This paper presents a laboratory study of multi-class classification problem for multiple indoor air contaminants which belongs to a completely linear-inseparable case. Six kinds of indoor air contaminations (formaldehyde, benzene, toluene, carbon monoxide, ammonia and nitrogen dioxide) were recognized as indicators of air quality in this project. The effectiveness of the proposed HSVM model has been rigorously evaluated on the experimental E-nose data sets. In addition, we have also compared it with existing five methods including Euclidean distance to centroids (EDC), simplified fuzzy ARTMAP network (SFAM), multilayer perceptron neural network (MLP) based on back-propagation learning rule, individual FLDA and single SVM. Experimental results have demonstrated that the HSVM model outperforms other classifiers in general. Also, HSVM classifier preliminarily shows its superiority in solution to discrimination in various electronic nose applications.



## 1. Introduction

Electronic nose (E-nose) system, which imitates the perceptional mechanisms of biological olfactory using a chemical sensor array, is designed to detect and discriminate complex odors. The sensor array in an E-nose system consists of several non-specific sensors, and when exposed to an odorant stimulus, a characteristic pattern from the sensor array would be generated. Patterns from known odorants are employed to construct a database and train a pattern recognition model through some learning rules, so that unknown odorants can be classified and discriminated subsequently [1]. In recent years, E-nose technology has been widely employed in diverse fields such as environmental controls [2–4], medical areas [5,6], agriculture [7], food and pharmaceutical industries [8–10].

Metal oxide semiconductors gas sensors array with cross sensitivity toward different components have been widely applied in E-nose system. In pattern analysis, one or more features in steady state responses were selected and a vector which can be seemed as the pattern of one observation was obtained. In discrimination, a classification model is first developed on the training patterns; then, the performance of the model is evaluated by means of the independent testing samples; the final classification accuracy can be calculated by comparing their predicted categories with their own true categories. So far, many pattern recognition models based on intuitive, linear and nonlinear supervised techniques have been explored in E-nose data. In this paper, E-nose technology was used to discriminate six kinds of indoor air contaminants by developing a hybrid classification model. Compared with those previous studies, we have systematically studied different linear and nonlinear tools and try to find an optimal model for gases classification. Among a large number of classification models, we select five representative methods for comparisons. They are Euclidean distance to centroids (EDC) [11], fuzzy ARTMAP network [12,13], multilayer perceptron neural network (MLP) [14–16], fisher linear discrimination analysis (FLDA) [17], and support vector machine (SVM) [18,19].

EDC, which assigns samples to the class with the minimum distance, is a very intuitive classification method. For each class and each variable, the centroid is calculated over all samples in that class. It is assumed that the distribution of samples around the centroid is symmetrical in the original variable space for each class. However, it cannot make use of the full discriminatory power available in all the variables so that this method actually obtains worse classification. Artificial neural networks (ANNs), especially fuzzy ARTMAP and MLP based on back-propagation learning rule, have been recognized to be successful in pattern recognition system (PARC). Fuzzy ARTMAP is a constructive neural network model developed upon adaptive resonance theory and fuzzy set theory [20,21]. It allows knowledge to be added during training if necessary so that it has also been used for pattern recognition. Back-propagation multilayer perceptron neural network, which is

* Corresponding author. Tel.: +86 13629788369; fax: +86 23 65111745.
*E-mail address:* leizhang@cqu.edu.cn (L. Zhang).

0925-4005/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.snb.2012.07.021

a non-linear, non-parametric and supervised method, performed well in a variety of application [22,23]. When it comes to the drawbacks of MLP, back-propagation algorithm has a limited capability to compensate for undesirable characteristics of the sensor system (e.g. temperature, humidity variations and drift) and it is trained "off-line" and unable to adapt autonomously to the changing environment. Consequently, recalibration is still necessary in different periods. Although ARTMAP can realize "on-line" training through testing the new measurements, the problem is that it does not know the specific component or category in each new measurement. And also, the robustness and real-time characteristic of ARTMAP will be lost when compared with MLP in real applications. LDA, as a supervised method, has been used for feature extraction and variable selection [24] in a dataset like the unsupervised principle component analysis (PCA). Both of them extract features by transforming the original parameter vectors into a new feature space through a linear projection. Besides, LDA has also been used for discrimination. However, when the actual problem becomes completely nonlinear (e.g. the sensor array system), it will become unqualified. SVM, which was first introduced by Vapnik, is a relatively new machine learning technique [25,26]. It has been proven advantageous in handling classification tasks with excellent generalization performance and robustness. For improvement of SVM, LDA as feature extraction method has been combined with SVM for fault diagnosis and hepatitis disease diagnosis [27]. Unfortunately, the sensor array produces a response vector for each observation, but not a matrix or dataset in real-time E-nose monitoring. In other words, a certain sampling time for a dataset collection should be needed for easy analysis by LDA or PCA which would make an online/real-time use of an E-nose impossible. Since the feature extraction by LDA or PCA cannot operate in real-time processing, the hybrid classification model would also become meaningless.

Particularly, most classification models can successfully solve a simple two-class problem. However, in this paper we devote to solving a complex multi-class problem. An electronic nose can be a better alternative to conventional methods for continuous and real-time monitoring of air quality indoor in dwellings or in a car as a portable E-nose instrument. Four classes of contaminants (physical, chemical, biological and radioactive) were reported in indoor air quality standard. Chemical contaminants including sulfur dioxide, nitrogen dioxide, carbon monoxide, carbon dioxide, ammonia, ozone, formaldehyde, benzene, toluene, inhalable particle, and volatile organic compounds were recognized as harmful substances to public health indoor [28]. The common contaminants in people's dwellings which we aim to employ in our project using E-nose technology contain formaldehyde, benzene, toluene, carbon monoxide, ammonia and nitrogen dioxide. These odorants have been mostly investigated for their potential harms to public health as pollutants of indoor air quality from numerous studies [28–33]. By conclusions of these publications, we find that the fixed six gases in our project were widely studied in dwellings. The emissions from new furniture, oil paint, and building materials of residuals often contain formaldehyde, benzene, toluene, and ammonia [28]. Besides, carbon monoxide and nitrogen dioxide are often produced from the smoking of cigarettes, wood burning stoves and car exhaust. A detail comparison research of indoor air pollutants in urban dwellings in Japan and Sweden has been developed in [29] and provided new data concerning the concentrations of formaldehyde and nitrogen dioxide.

In this work, we present a laboratory study of a multi-class problem for classification of six contaminants using a hybrid discrimination model based on fisher linear discrimination analysis (FLDA) and support vector machine (SVM) for monitoring and realizing a real-time gases category decision in people's dwellings by an E-nose. The role of FLDA is equivalent to a pre-classification by transforming the original data into a new feature space with more linearly independent variables correlated with each classifier, and more prior information about each class in the new feature space would be obtained. Thus, it makes SVM easier for final discrimination in the new feature space. For clarity, the hybrid model of FLDA and SVM in this paper is called HSVM. The comparison results with EDC, simplified fuzzy ARTMAP (SFAM), MLP, individual FLDA and single SVM demonstrate the potential ability of HSVM in E-nose.

## 2. Classification methodologies considered

In this section, we illustrate the basic principle, mathematical formulas and some important details about the related classification methodologies. An in-depth description of the classification theory is beyond the scope of this paper. For clarity, we have referred readers with interest to related references.

### 2.1. Euclidean distance to centroids (EDC)

EDC is a very intuitive classification method through assigning the nearest samples with the centroid to the corresponding class [11]. For each class, the mean (centroid) is calculated over all samples in that class. The Euclidean distance between sample $i$ and the class $k$ centroid is calculated as

$$d_{ik} = \sqrt{(\boldsymbol{x}_i - \bar{x}_k) \cdot (\boldsymbol{x}_i - \bar{x}_k)^{\mathrm{T}}} \quad (1)$$

where $\boldsymbol{x}_i$ is the $i$th sample described by a row vector with $n$ variables ($n$ denotes the number of sensors), $\bar{x}_k$ is the centroid of class $k$, and T denotes the transpose of a vector. The sample with the minimum distance $d$ will be assigned to a specific class.

### 2.2. Simplified fuzzy ARTMAP network (SFAM)

ARTMAP consists of two modules (fuzzy ART and inter-ART) that create stable recognition categories in response to the input patterns. Fuzzy ART receives a stream of input features representing the pattern map to the output classes in the category layer. Fuzzy ART module has three layers: $F_0$, $F_1$, and $F_2$. Inter ART module works by increasing the small vigilance parameter $\varepsilon$ of fuzzy ART for updating the prediction error in the output category layer. We refer interested readers to [34] for the basic mathematical descriptions of SFAM.

For parameter settings, the related parameters in [34] such as vigilance $\rho = 0.9$, $\alpha = 0.2$, learning rate $\beta = 1$, and $\varepsilon = 0.001$ are used in this paper; the number of maximum categories and training times are set to 100, respectively.

### 2.3. Multilayer perceptron neural network (MLP)

A typical multilayer perceptron consists of an input layer, one hidden layer and one output layer. The input and output elements denote the observations composed of six variables (sensor) and the known category (labels) of each observation. Detailed description of MLP is out of the scope of this present study; for that, we refer the readers to [35]. In this paper, we use three-bit binary codes to represent the identities of six categories. The identities of formaldehyde, benzene, toluene, carbon monoxide, ammonia and nitrogen dioxide were labeled as $(0, 0, 1)^{\mathrm{T}}$, $(0, 1, 0)^{\mathrm{T}}$, $(0, 1, 1)^{\mathrm{T}}$, $(1, 0, 0)^{\mathrm{T}}$, $(1, 0, 1)^{\mathrm{T}}$ and $(1, 1, 0)^{\mathrm{T}}$, respectively. The number of nodes in input layer, hidden layer and output layer were set as 6, 35 and 3, respectively. The output value $p$ for each node should be adjusted as if $p \geq 0.5$, $p = 1$; else $p = 0$. The activation functions of the hidden layer and output layer we have used in classification are "logsig" and "purelin". The training goal and training times are set to 0.05 and 1000, respectively.

### 2.4. Fisher linear discrimination analysis (FLDA)

Fisher linear discrimination analysis easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set and thereby guaranteeing maximum separability and also producing a linear decision boundary between two classes. A brief mathematical description for a two-class problem is shown as follows.

Assume we have a set of $n$-dimensional dataset $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2\}$, where $\mathbf{X}_1$ belongs to class 1 which contains $N_1$ column vectors and $\mathbf{X}_2$ belongs to class 2 which contains $N_2$ column vectors. The centroid of each class is calculated by

$$\mu_i = \frac{1}{N_i} \cdot \sum X_i, \quad i = 1, 2 \tag{2}$$

The within-scatter matrix of class $i$ is shown by

$$S_i = \sum_{j=1}^{N_i} (X_{i,j} - \mu_i)(X_{i,j} - \mu_i)^{\mathrm{T}}, \quad i = 1, 2 \tag{3}$$

Then the within-class scatter matrix $S_\mathrm{w}$ and the between-class scatter matrix $S_\mathrm{b}$ can be calculated by

$$S_\mathrm{w} = \sum_{i=1}^{2} S_i \tag{4}$$

$$S_\mathrm{b} = \sum_{i=1}^{2} N_i \cdot (\mu_i - \bar{X}) \cdot (\mu_i - \bar{X})^{\mathrm{T}} \tag{5}$$

where $\bar{X}$ denotes the centroid of the total dataset $\boldsymbol{X}$.

Finally, the fisher criterion in terms of $S_\mathrm{w}$ and $S_\mathrm{b}$ is expressed as

$$J(\boldsymbol{W}) = \frac{W^T S_\mathrm{b} W}{W^T S_\mathrm{w} W} \tag{6}$$

where $\boldsymbol{W}$ is the transformation matrix which can be calculated by solving the eigenvalue problem

$$W^* = \mathrm{argmax}\{J(W)\} = S_w^{-1} \cdot (\mu_1 - \mu_2) \tag{7}$$

### 2.5. Support vector machine (SVM)

Support vector machines perform structural risk minimization in the framework of regularization theory. For linearly inseparable cases SVM applies a non-linear kernel function to transform the input space to a higher dimensional feature space so that the classes may be linearly separable prior to calculate the separating hyperplane. This kernel function can be polynomial, Gaussian radial basis function (RBF) or sigmoid function. In this work, a linearly inseparable case is considered, and only Gaussian RBF kernel function was attempted for classification due to its good generalization and without the guidance from those prior experiences. Therefore, this problem aims to solving a quadratic optimization in a higher dimensional feature space. The Lagrangian function is shown by

$$L_{\mathrm{LSSVM}}(\alpha) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \cdot \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i)^{\mathrm{T}} \phi(x_j) \tag{8}$$

which needs to be minimized under the constraints: $\alpha_i > 0$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

By introducing a kernel function

$$K(x_i, x_j) = \phi(x_i)^{\mathrm{T}} \phi(x_j) \tag{9}$$

the Lagrangian function can be rewritten by

$$L_{\mathrm{LSSVM}}(\alpha) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \cdot \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{10}$$

The Gaussian RBF kernel function can be represented as

$$K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{\sigma^2}\right) \tag{11}$$

where $\sigma^2$ is the kernel parameter which determines the bandwidth of RBF. The decision function can be expressed as

$$f(x) = sgn\left(\sum_{i=1}^{N} \alpha_i K(x_i, x) + b\right) \tag{12}$$

where $\alpha$ and $b$ are the optimal decision parameters.

## 3. Application of classification models to experimental E-nose data

### 3.1. The E-nose system with sensor array

The details of the E-nose module have been illustrated in experimental part of our previous publication [36]. Briefly, four metal oxide semiconductor gas sensors (TGS2602, TGS2620, TGS2201A and B from Figaro company) and an extra module (HTD2230-I$^2$C) with two auxiliary sensors for temperature and humidity compensations were used in our E-nose. The sensors were mounted on a custom designed printed circuit board (PCB), along with associated electrical components. A 12-bit analog-digital converter (A/D) is used as interface between the FPGA (Field Programmable Gate Array) processor and the sensors. The system can be connected to the PC via a JTAG (Joint Test Action Group) port. The sensor array will produce a group of odorant pattern with six variables including temperature, humidity, TGS2620, TGS2602, TGS2201A and TGS2201B in each observation. The reasons for selection of these four gas sensors can be concluded as two aspects. First, they have a good sensitivity to indoor air contaminants. The sensitivity can be indicated as the ratio of sensor resistance ($R_\mathrm{s}$) at various concentrations and sensor resistance ($R_\mathrm{o}$) in fresh air or 300 ppm of ethanol. The corresponding parameters including the basic circuits, heater voltage, heater current, standard test curves, etc. for each sensor of this work have been provided with the datasheet (.pdf) in the supplementary data. The species monitored by the sensor array contain carbon monoxide, nitric oxide, nitrogen dioxide, ammonia, toluene, ethanol, hydrogen, methane, hydride and VOCs. Second, they have a long-term stability and good reproducibility. Also, we refer readers to the sensors' datasheets available in http://www.figaro.co.jp/en/product/index.php?mode=search&kbn=1 for more information on the other TGS sensors. For visualization of our E-nose system, the experimental platform was simplified and presented in Fig. 1. From the experimental platform, we can find that five ports (from port_1 to port_5) are used in the chamber. For clarity, port_1 is used for injection of contaminants, port_2 is used to clean the chamber after each experiment through injection of fresh air (nitrogen), port_3 is set to control the relative humidity in the chamber by using a humidifier with a valve, port_4 is for data collection by connecting the PC to the sensor array board with a JTAG and port_5 is set to sampling by a gas sampler for true concentrations.

### 3.2. E-nose data

Six familiar chemical contaminants indoor including formaldehyde (HCHO), benzene ($C_6H_6$), toluene ($C_7H_8$), carbon monoxide (CO), ammonia ($NH_3$) and nitrogen dioxide ($NO_2$) are investigated

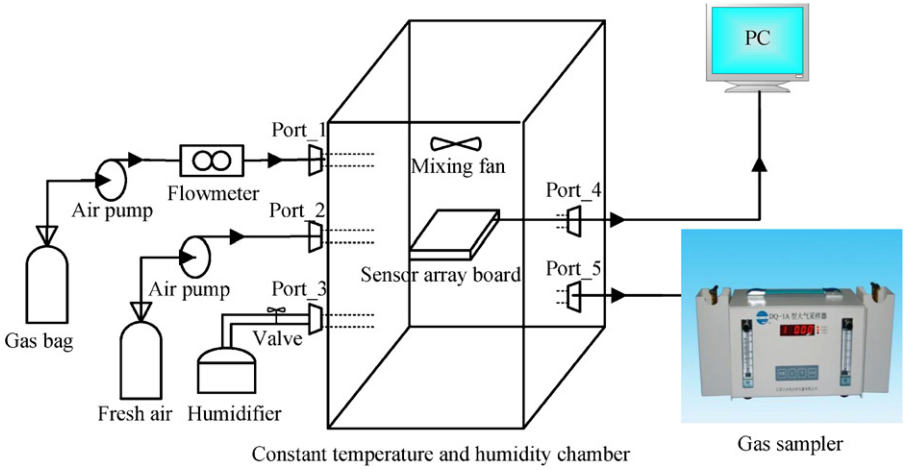


**Fig. 1.** Systematic experimental platform in this work.

in this work. The experiments were employed by an E-nose in the constant temperature and humidity chamber whose type is LRH-150S. The accuracy for temperature and humidity of the chamber is ±0.5 °C and ±5%. In gases preparation, HCHO, $C_6H_6$ and $C_7H_8$ are liquor, and CO, $NH_3$ and $NO_2$ are standard gas. In each gas measurements, a gas bag collected with target gas and nitrogen ($N_2$) was prepared for injection into the chamber. Note that $N_2$ is used to dilute the gas concentration in the gas bag, and we get various concentrations by setting different injection times (injection speed to 5 l/min). The true concentrations for HCHO and $NH_3$ were measured using spectrophotometer, $C_6H_6$ and $C_7H_8$ were employed using Gas Chromatography, and the true concentrations of CO and $NO_2$ were obtained using the reference instruments whose measurement accuracy are within ±3%. For each experiment, 12 min (e.g. 2 min for baseline and 10 min for response) were consumed and extra 15 min were also needed for cleaning the chamber by injecting pure air. Totally, 260, 164, 66, 58, 29 and 30 samples with target temperature, humidity and various concentrations were collected for HCHO, $C_6H_6$, $C_7H_8$, CO, $NH_3$, and $NO_2$, respectively. These samples were measured with different combinations of the target temperatures 15, 25, 30, 35 °C and relative humidity (RH) of 40%, 60%, 80% which can approximately simulate the indoor temperature and humidity for improving the classifier robustness of the E-nose. The conditions including temperature, humidity and concentration of the experimental samples were presented in Table 1. 12 combinations {(15, 60), (15, 80), (20, 40), (20, 80), (25, 40), (25, 60), (25, 80), (30, 40), (30, 60), (30, 80), (35, 60)} in manner of ($T$, RH), in which $T$ denotes temperature and RH (%) denotes

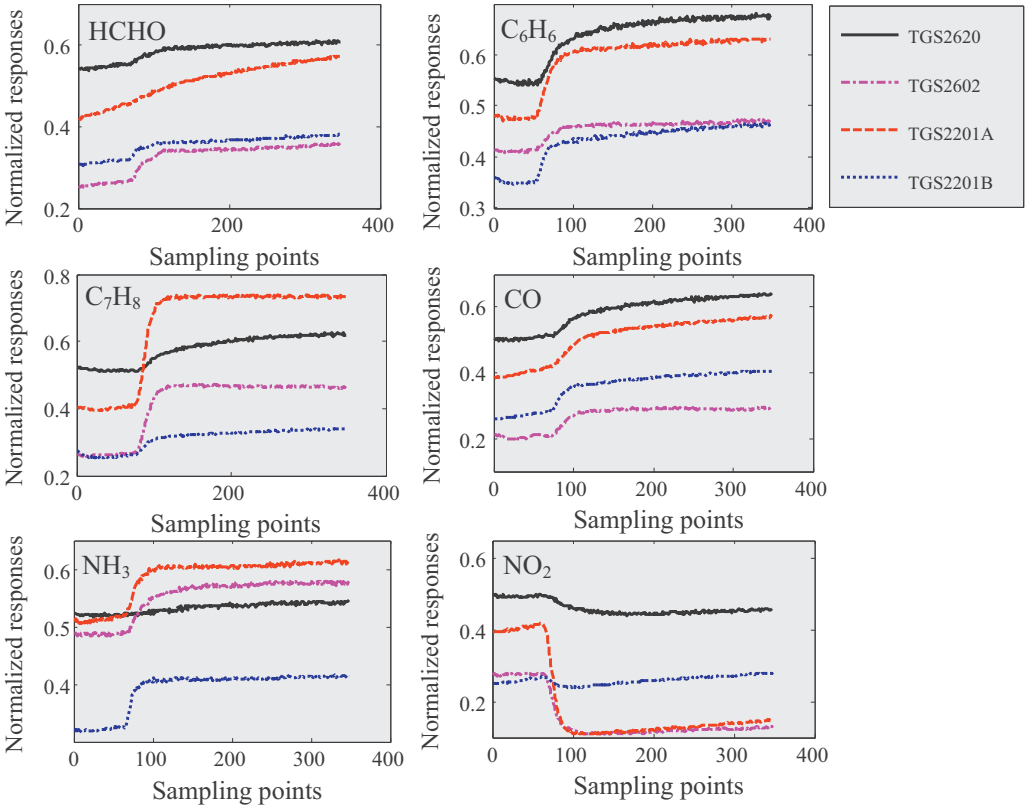


**Fig. 2.** Response curves of four gas sensors for different odorants.

**Table 1**
Concentration (ppm) condition of each experimental sample in different combinations ($T$, RH) in which $T$ denotes temperature and RH (%) denotes relative humidity.

| (15, 60) | (15, 80) | (20, 40) | (20, 60) | (20, 80) | (25, 40) | (25, 60) | (25, 80) | (30, 40) | (30, 60) | (30, 80) | (35, 60) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Conditions of HCHO samples | | | | | | | | | | | |
| 0.04 | 0.14 | 0.07 | 0.10 | 0.08 | 0.13 | 0.24 | 0.06 | 0.23 | 0.13 | 0.09 | 0.04 |
| 0.05 | 0.16 | 0.07 | 0.07 | 0.06 | 0.45 | 0.10 | 0.25 | 0.39 | 0.15 | 0.02 | 0.09 |
| 0.08 | 0.72 | 0.16 | 0.15 | 0.11 | 0.31 | 0.26 | 1.04 | 1.37 | 0.22 | 0.25 | 0.81 |
| 0.12 | 0.10 | 1.32 | 0.09 | 0.28 | 0.52 | 2.11 | 0.02 | 0.58 | 2.06 | 0.09 | 0.16 |
| 0.64 | 0.34 | 0.60 | 0.23 | 1.10 | 0.22 | 0.37 | 0.11 | 0.52 | 0.08 | 0.13 | 0.58 |
| 0.21 | 1.22 | 0.61 | 0.17 | 0.13 | 0.49 | 0.05 | 0.04 | 0.02 | 0.23 | 0.43 | 0.04 |
| 0.25 | 0.13 | 0.16 | 0.16 | 0.15 | 0.07 | 0.01 | 0.27 | 0.09 | 0.27 | 0.76 | 0.12 |
| 0.06 | 0.26 | 2.62 | 0.20 | 0.25 | 0.30 | 0.11 | 0.33 | 0.12 | 0.29 | 1.15 | 0.45 |
| 0.05 | 0.52 | 0.45 | 0.21 | 0.35 | 0.26 | 0.17 | 0.79 | 0.56 | 0.31 | 2.42 | 0.39 |
| 0.18 | 0.69 | 0.05 | 0.22 | 0.59 | 0.23 | 0.24 | 1.01 | 0.68 | 0.56 | 2.01 | 0.61 |
| 0.22 | 2.29 | 0.08 | 0.24 | 1.16 | 0.04 | 0.17 | 1.29 | 0.92 | 1.01 | 2.30 | 1.62 |
| 0.12 | 2.45 | 1.16 | 0.24 | 1.88 | 1.01 | 0.27 | 1.93 | 1.31 | 1.06 | | 0.22 |
| 0.19 | 0.12 | 3.13 | 0.60 | 1.17 | 1.09 | 0.12 | 2.17 | 2.37 | 1.65 | | 0.31 |
| 0.20 | 1.06 | 0.48 | 0.77 | 1.83 | 2.62 | 0.17 | 0.26 | 0.01 | 1.84 | | 0.32 |
| 0.52 | 1.44 | 0.60 | 1.23 | | 0.06 | 0.24 | 1.01 | 0.09 | 0.08 | | 0.39 |
| – | – | – | – | | – | – | – | – | – | | – |
| Conditions of $C_6H_6$ samples | | | | | | | | | | | |
| 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 |
| 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 |
| 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 |
| 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 |
| 0.11 | 0.06 | 0.09 | 0.08 | 0.20 | 0.19 | 0.15 | 0.15 | 0.19 | 0.15 | 0.20 | 0.14 |
| 0.18 | 0.20 | 0.07 | 0.15 | 0.06 | 0.10 | 0.13 | 0.06 | 0.18 | 0.18 | 0.13 | |
| 0.25 | 0.21 | 0.25 | 0.21 | 0.14 | 0.21 | 0.14 | 0.14 | 0.08 | 0.10 | 0.24 | |
| 0.18 | 0.11 | 0.26 | 0.36 | 0.16 | 0.18 | 0.19 | 0.16 | 0.24 | 0.19 | 0.22 | |
| 0.24 | 0.30 | 0.18 | 0.42 | 0.21 | 0.33 | 0.20 | 0.16 | 0.25 | 0.30 | | |
| 0.32 | 0.22 | 0.06 | 0.43 | 0.21 | 0.16 | 0.10 | 0.20 | 0.18 | 0.41 | | |
| 0.11 | 0.26 | 0.11 | | | | 0.17 | 0.21 | 0.24 | | | |
| Conditions of $C_7H_8$ samples | | | | | | | | | | | |
| 0.05 | 0.05 | 0.05 | 0.05 | 0.06 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| 0.08 | 0.06 | 0.08 | 0.06 | 0.08 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |
| 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| 0.06 | 0.08 | 0.06 | 0.08 | 0.05 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 | 0.08 |
| Conditions of CO samples | | | | | | | | | | | |
| 6 | 4 | 6 | 5 | 5 | 6 | 4 | 5 | 5 | 14 | 4 | 5 |
| 11 | 23 | 12 | 22 | 22 | 24 | 8 | 23 | 8 | 29 | 16 | 13 |
| 43 | 43 | 41 | 43 | 44 | 46 | 10 | 45 | 23 | 49 | 48 | 20 |
| 23 | 12 | 22 | 11 | 12 | 14 | 21 | 33 | 37 | 55 | 13 | 29 |
| | | 13 | 9 | 20 | 10 | 12 | 48 | 6 | 25 | 16 | 20 |
| Conditions of $NH_3$ samples | | | | | | | | | | | |
| 0.10 | 0.28 | 0.34 | 0.80 | 0.98 | 0.09 | 0.33 | 0.27 | 0.66 | 0.79 | 0.20 | 0.28 |
| 0.50 | | 1.72 | 0.79 | 0.44 | 0.53 | 0.79 | 0.73 | 0.09 | 0.92 | 0.36 | 2.15 |
| 0.25 | | 0.80 | | | 0.12 | 0.55 | | | 1.18 | | 0.27 |
| Conditions of $NO_2$ samples | | | | | | | | | | | |
| 0.09 | | 0.03 | 0.16 | 0.10 | 0.12 | 0.15 | 0.03 | | 0.21 | | |
| 0.20 | × | 0.92 | 0.84 | 0.54 | 0.31 | 0.22 | 0.05 | × | 0.61 | × | × |
| 1.62 | | 0.77 | 0.28 | 0.18 | 0.20 | 0.70 | 0.87 | | 1.36 | | |
| 0.66 | | | | | | 0.02 | 1.59 | | | | |
| | | | | | | | 0.07 | | | | |
| | | | | | | | 0.17 | | | | |

relative humidity, were employed for covering the indoor conditions. Due to the large data space, we presented the basic data structure which can cover a majority of the indoor conditions in Table 1. In visualization of sensor response curve in one measurement, we randomly select one sample in the combination of (15, 60) for each gas from Table 1 and present the whole sensor response curves from the baseline to the steady state response (12 min) in Fig. 2. Note that the sensor responses have been normalized, and the measured concentrations of each selected sample for HCHO, $C_6H_6$, $C_7H_8$, CO, $NH_3$, and $NO_2$ were 0.18 ppm, 0.28 ppm, 0.14 ppm, 6.0 ppm, 0.50 ppm and 1.62 ppm, respectively. The normalization is that sensor responses were directly divided by 4095. It is worthy noting that the digit of 4095 (that is, $2^{12}-1$) is the maximum value of the 12-bit A/D output for each sensor. In feature extraction, one value at the steady state response (the 240th point in Fig. 2) for each sensor was selected as the corresponding feature in such a simple way. Therefore, a vector with 6 variables including temperature and relative humidity is extracted as the feature vector of that sample for subsequent pattern recognition.

### 3.3. Multi-class HSVM discrimination of E-nose data

The common used two methods for solving multi-class problems are "one-against-all" and "one-against-one" [37]. In this work, the "one-against-one" strategy (OAO) is used in HSVM to build the $k=6$ classes classifier for the recommendation that it would be a better choice for $k \leq 10$ [19]. Thus, this strategy builds $k\cdot(k-1)/2=15$ sub-classifiers (FLDA classifier or SVM classifier) trained using input patterns of two classes. Consequently, a complex multi-class problem can be untied through solutions of multiple two-class classifiers with a voting scheme in decision that if the indicator function of each sub-classifier says that $x$ belongs to

**Table 2**
Distribution of training set and testing set.

| Training-testing proportion (%) | Number of samples in the subset | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Training set | | | | | | Testing set | | | | | |
| | HCHO | $C_6H_6$ | $C_7H_8$ | CO | $NH_3$ | $NO_2$ | HCHO | $C_6H_6$ | $C_7H_8$ | CO | $NH_3$ | $NO_2$ |
| 30–70 | 78 | 49 | 20 | 18 | 9 | 9 | 182 | 115 | 46 | 40 | 20 | 21 |
| 50–50 | 130 | 82 | 33 | 29 | 15 | 15 | 130 | 82 | 33 | 29 | 14 | 15 |
| 80–20 | 208 | 131 | 53 | 46 | 23 | 24 | 52 | 33 | 13 | 12 | 6 | 6 |

class $i$, then the vote for class $i$ is increased by one, otherwise, the vote for class $j$ is increased by one.

In terms of OAO strategy, 15 FLDA and 15 SVM classifiers in the HSVM model should be designed separately in a 6-category classification problem. Here, FLDA transforms the original data into a new feature space composed of more linearly independent variables which can be recognized as the new variables with more characteristic of linear separability and prior information of each sub-classifier which are easier for SVM classification. The implementation process of HSVM in E-nose data can be illustrated as follows.

Assume that the number of HCHO, $C_6H_6$, $C_7H_8$, CO, $NH_3$, and $NO_2$ training samples is $n_1$, $n_2$, $n_3$, $n_4$, $n_5$, and $n_6$, respectively. Thus, the original training input data matrix $\mathbf{X}_{\text{original}}$ can be constructed in order by

$$\mathbf{X}_{\text{original}} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6\} \tag{13}$$

where the $i$th matrix $\mathbf{X}_i$ is $6 \times n_i$, $i = 1, \ldots, 6$; thus, $\mathbf{X}_{\text{original}}$ is a matrix of $6 \times \sum_{i=1}^{6} n_i$, and each column denotes one observation vector.

The training goal (category label) is constructed in order as

$$\text{Label} = \{\overbrace{1, \ldots, 1}^{n_1}, \overbrace{2, \ldots, 2}^{n_2}, \overbrace{3, \ldots, 3}^{n_3}, \overbrace{4, \ldots, 4}^{n_4}, \overbrace{5, \ldots, 5}^{n_5}, \overbrace{6, \ldots, 6}^{n_6}\} \tag{14}$$

Assume that the total transformation matrix of the 15 FLDA classifiers is expressed by

$$\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{15}\} \tag{15}$$

where $\boldsymbol{w}_j$ ($j = 1, \ldots, 15$) is a column vector of $6 \times 1$ representing the transformation of each sub-classifier between two classes which can be directly used for classification. Therefore, $\mathbf{W}$ is a matrix with a size of $6 \times 15$. Then the input data $\mathbf{X}_{\text{HSVM}}$ of HSVM can be reconstructed by projection

$$\mathbf{X}_{\text{HSVM}} = \mathbf{W}^{\text{T}}\mathbf{X}_{\text{original}} \tag{16}$$

Similarly, the original testing input data matrix should also be reconstructed in terms of the principle of training input data. Consequently, the training and test input pattern of SVM has now been correlated with the initial FLDA classification, and more linearly independent variables correlated with sub-classifiers were produced through a linear projection without shifting the number of original patterns. The new patterns preprocessed by FLDA will then be used for SVM classification with structural risk minimization.

### 3.4. Data analysis

The performance of E-nose data classification was assessed in terms of the classification accuracy of test samples. The classification accuracy is defined as a percentage of correct classifications in all test samples. Also, the average accuracy of training and test samples were calculated for insight of the whole data. To validate the robustness and generalization of all classifiers considered in this work, three proportions 30–70%, 50–50% and 80–20% of training and test samples were analyzed, respectively. For selection of training set in terms of some proportion, a Kennard–Stone sequential (KSS) algorithm [38] based on the multivariate Euclidean distance was used, and the remaining samples were recognized as test samples. The distribution with three proportions of training set and test set for each class is represented in Table 2.

Note that, all the classification models were only performed on the training sets, then, the trained parameters were applied on the testing sets. All algorithms for multi-class discrimination were implemented in MATLAB 2009a, operating on a laboratory computer equipped with Inter Core (TM) i3 CPU 530, 2.93 GHz processors and 2 GB of RAM.

## 4. Results and discussion

### 4.1. Experimental results

To evaluate the effectiveness of the hybrid model HSVM, the E-nose data were analyzed by using all the classifiers considered in our project. We first presented the PCA (principle component analysis) results of the original training sets. Fig. 3 illustrates three 2D scatter sub-plots (PC-1 vs PC-2, PC-1 vs PC-3 and PC-2 vs PC-3) and a 3D scatter sub-plot of the first three principal components when perform PCA program on the 80% training set. From the 2D and 3D PCA plots, we can get that the first three PCs can totally account for 92.59% information of the training data. Obviously, the multi-class problem in this work belongs to a completely linear-inseparable case because of the serious overlaps among all classes. Especially, the patterns of HCHO and $C_6H_6$ as indoor air contaminants are completely inseparable with other gas patterns from the PCA results. It is noteworthy that PCA is an unsupervised method which transforms the original data into the space of the principal components through a linear projection. Namely PCA is a multi-dimensional signal analysis method in statistical learning by projecting correlated variables into another orthogonal feature space and thus a group of new variables with the largest variance (global variance maximization) were obtained [39]. A key feature of PCA is its ability to reduce large multivariate data to a few orthogonal principal components, which still contain the majority of information held in the raw data. The analysis of the PCA results confirms the necessity of nonlinear classifiers employment due to that they can make the linearly inseparable problem separable in a high dimensional space through a non-linear transform.

To visualize the magnitude and sign of each variable's contribution to the first two and three principal components, Fig. 4 illustrates the 2D and 3D PCA results on the 80% training set. From Fig. 4, we can find that TGS2620 and TGS2201B have the same direction and similar contribution. Humidity works in reverse direction compared with other variables. It means that humidity is a key variable in data analysis and cannot be neglected in sensor array. In fact, if one variable has no use in multivariate data, it should be in the origin point of the PC space in Fig. 4. That also means the important role of temperature in the sensor array. In addition, we can also find that humidity shows the significance in PC-2, while temperature shows the significance in PC-1 and PC-3. Therefore,

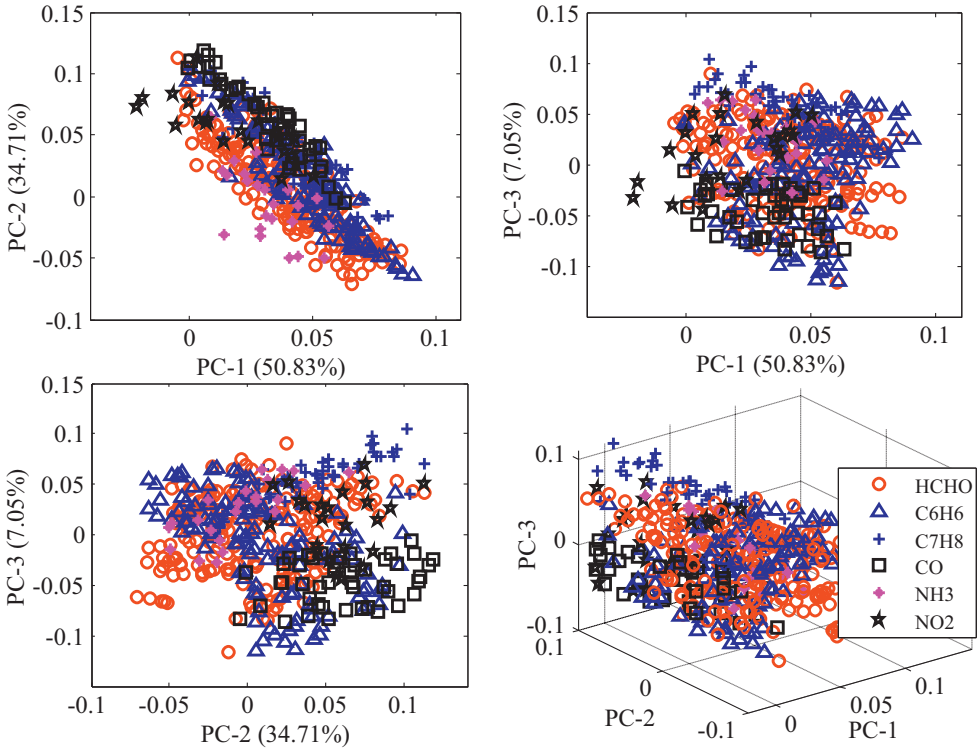


**Fig. 3.** PCA results with the first three principal components of 80% training set.

both of them integrated together can be more effective for classification.

Besides the macro-analysis of the total gases, we have also presented the PCA results in Fig. 5 for organic and inorganic contaminants, respectively. Seen from the left subfigure, three organic contaminants have same direction in the first two principles, and they are crossed with each other; in the right subfigure, we can find the three inorganic gases could be linearly separable with each other. The reasons can be concluded in three aspects. First, the patterns of organic gases are similar with each other, but different with the patterns of inorganic gases because of their different chemical characteristics. Second, the selected sensors are more sensitive to inorganic gases than organic gases. Third, more sample points of HCHO, $C_6H_6$ and $C_7H_8$ will cover a larger PC space which may result in an easier overlap among them.

Tables 3–5 present the discrimination results of 70%, 50% and 20% of testing samples using the HSVM classification model developed on the remaining 30%, 50% and 80% of training samples, respectively. The digits with bold type in diagonal line denote the number of correctly classified samples, while others denote the number of misclassified samples.

For quantification of classification accuracy and present the comparisons with other classification models, Table 6 shows the classification accuracy including the train set and test set with a

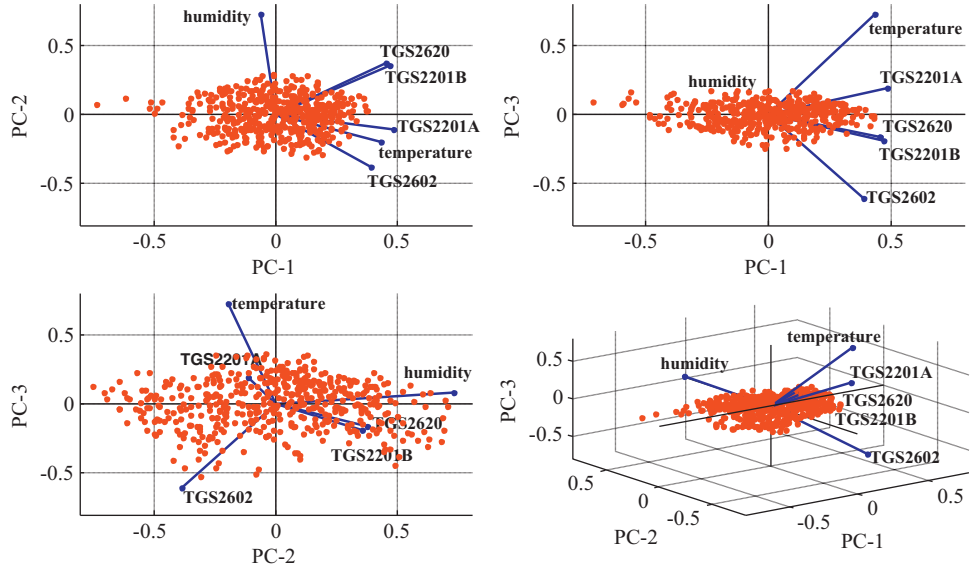


**Fig. 4.** Relations between each variable and the first three principle components.

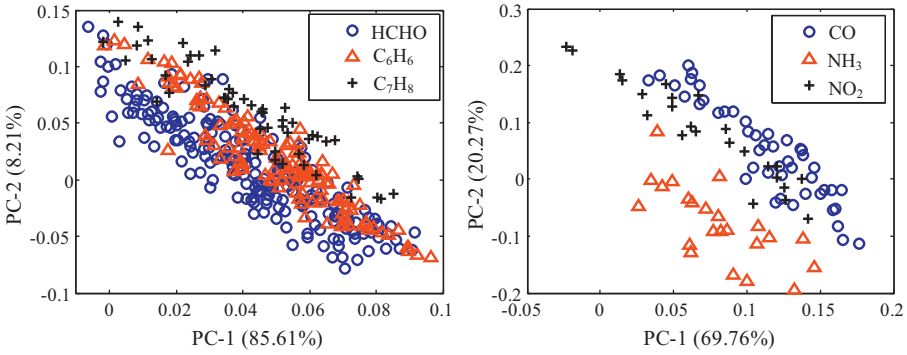


**Fig. 5.** PCA results of organic and inorganic contaminants.

**Table 3**
Multi-classification results of *N* testing samples which occupy 70% of the total samples using HSVM classification method.

| Class | *N* | Classified as | | | | | |
|---|---|---|---|---|---|---|---|
| | | HCHO | $C_6H_6$ | $C_7H_8$ | CO | $NH_3$ | $NO_2$ |
| HCHO | 182 | **166** | 6 | 8 | 1 | 0 | 1 |
| $C_6H_6$ | 115 | 16 | **80** | 11 | 1 | 7 | 0 |
| $C_7H_8$ | 46 | 0 | 0 | **46** | 0 | 0 | 0 |
| CO | 40 | 1 | 3 | 0 | **36** | 0 | 0 |
| $NH_3$ | 20 | 2 | 0 | 3 | 0 | **15** | 0 |
| $NO_2$ | 21 | 2 | 0 | 8 | 2 | 1 | **8** |

**Table 4**
Multi-classification results of *N* testing samples which occupy 50% of the total samples using HSVM classification method.

| Class | *N* | Classified as | | | | | |
|---|---|---|---|---|---|---|---|
| | | HCHO | $C_6H_6$ | $C_7H_8$ | CO | $NH_3$ | $NO_2$ |
| HCHO | 130 | **116** | 5 | 6 | 1 | 2 | 0 |
| $C_6H_6$ | 82 | 9 | **66** | 5 | 1 | 1 | 0 |
| $C_7H_8$ | 33 | 0 | 0 | **33** | 0 | 0 | 0 |
| CO | 29 | 0 | 0 | 0 | **29** | 0 | 0 |
| $NH_3$ | 14 | 0 | 0 | 1 | 0 | **13** | 0 |
| $NO_2$ | 15 | 2 | 5 | 0 | 0 | 0 | **8** |

**Table 5**
Multi-classification results of *N* testing samples which occupy 20% of the total samples using HSVM classification method.

| Class | *N* | Classified as | | | | | |
|---|---|---|---|---|---|---|---|
| | | HCHO | $C_6H_6$ | $C_7H_8$ | CO | $NH_3$ | $NO_2$ |
| HCHO | 52 | **49** | 3 | 0 | 0 | 0 | 0 |
| $C_6H_6$ | 33 | 5 | **28** | 0 | 0 | 0 | 0 |
| $C_7H_8$ | 13 | 0 | 0 | **13** | 0 | 0 | 0 |
| CO | 12 | 0 | 0 | 0 | **12** | 0 | 0 |
| $NH_3$ | 6 | 0 | 0 | 0 | 0 | **6** | 0 |
| $NO_2$ | 6 | 0 | 0 | 1 | 0 | 0 | **5** |

**Table 6**
Classification accuracy with the proportion 30–70% of training and testing samples.

| Class | Classification accuracy (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train set | | | | | | Test set | | | | | |
| | EDC | SFAM | MLP | FLDA | SVM | HSVM | EDC | SFAM | MLP | FLDA | SVM | HSVM |
| HCHO | 17.95 | 100.0 | 89.31 | 69.23 | 83.33 | 96.15 | 23.63 | 65.38 | 84.62 | 66.48 | 84.07 | **91.21** |
| $C_6H_6$ | 48.98 | 100.0 | 89.80 | 67.35 | 65.31 | 97.96 | 71.30 | 57.39 | 75.65 | 72.17 | 72.17 | 69.57 |
| $C_7H_8$ | 60.00 | 100.0 | 80.00 | 85.00 | 90.00 | 95.00 | 54.35 | 86.96 | 89.13 | 93.48 | 89.13 | **100.0** |
| CO | 66.67 | 100.0 | 77.78 | 88.89 | 77.78 | 77.78 | 70.00 | 60.00 | 87.50 | 97.50 | 70.00 | **90.00** |
| $NH_3$ | 55.56 | 100.0 | 77.78 | 77.78 | 77.78 | 100.0 | 70.00 | 50.00 | 35.00 | 65.00 | 60.00 | **75.00** |
| $NO_2$ | 38.09 | 100.0 | 66.67 | 88.89 | 55.56 | 66.67 | 38.09 | 33.33 | 57.38 | 52.38 | 47.62 | 38.10 |
| Mean | 47.88 | 100.0 | 80.22 | 79.52 | 74.96 | 88.93 | 54.56 | 58.84 | 71.55 | 74.50 | 70.49 | **77.31** |
| Total | 38.49 | 100.0 | 85.61 | 73.77 | 77.05 | 93.44 | 47.17 | 62.73 | 79.26 | 73.11 | 77.12 | **82.79** |

**Table 7**
Classification accuracy with the proportion 50–50% of training and testing samples.

| Class | Classification accuracy (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train set | | | | | | Test set | | | | | |
| | EDC | SFAM | MLP | FLDA | SVM | HSVM | EDC | SFAM | MLP | FLDA | SVM | HSVM |
| HCHO | 17.69 | 100.0 | 93.08 | 71.54 | 98.41 | 100.0 | 15.38 | 69.23 | 91.54 | 71.54 | 86.92 | **89.23** |
| $C_6H_6$ | 42.68 | 100.0 | 80.49 | 64.63 | 93.90 | 100.0 | 57.32 | 71.95 | 69.51 | 79.27 | 82.93 | 80.48 |
| $C_7H_8$ | 69.70 | 100.0 | 90.91 | 90.91 | 96.67 | 96.97 | 54.55 | 69.70 | 90.91 | 93.94 | 96.97 | **100.0** |
| CO | 68.97 | 100.0 | 82.76 | 89.66 | 86.21 | 89.66 | 65.52 | 79.31 | 86.21 | 100.0 | 89.66 | **100.0** |
| $NH_3$ | 66.67 | 100.0 | 66.67 | 73.33 | 93.33 | 93.33 | 78.57 | 57.14 | 50.00 | 71.43 | 71.43 | **92.86** |
| $NO_2$ | 26.67 | 100.0 | 73.33 | 86.67 | 86.67 | 86.67 | 26.67 | 80.00 | 46.67 | 46.67 | 66.67 | 53.33 |
| Mean | 48.73 | 100.0 | 81.21 | 79.46 | 92.53 | 94.44 | 49.67 | 71.22 | 72.47 | 77.14 | 82.43 | **85.98** |
| Total | 37.83 | 100.0 | 86.19 | 74.34 | 95.01 | 97.70 | 39.27 | 70.96 | 80.86 | 77.56 | 85.47 | **87.46** |

**Table 8**
Classification accuracy with the proportion 80–20% of training and testing samples.

| Class | Classification accuracy (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train set | | | | | | Test set | | | | | |
| | EDC | SFAM | MLP | FLDA | SVM | HSVM | EDC | SFAM | MLP | FLDA | SVM | HSVM |
| HCHO | 25.96 | 100.0 | 90.87 | 73.56 | 95.40 | 98.07 | 7.69 | 78.85 | 86.54 | 67.31 | 90.38 | **94.23** |
| $C_6H_6$ | 51.91 | 100.0 | 81.68 | 70.99 | 91.30 | 98.47 | 69.70 | 69.70 | 78.79 | 72.73 | 90.91 | 84.85 |
| $C_7H_8$ | 69.81 | 100.0 | 90.57 | 88.68 | 95.65 | 96.23 | 38.46 | 69.23 | 92.31 | 100.0 | 100.0 | **100.0** |
| CO | 71.74 | 100.0 | 84.78 | 91.30 | 86.96 | 100.0 | 75.00 | 75.00 | 83.33 | 100.0 | 91.67 | **100.0** |
| $NH_3$ | 73.91 | 100.0 | 78.26 | 82.61 | 95.00 | 95.65 | 66.67 | 83.33 | 33.33 | 33.33 | 83.33 | **100.0** |
| $NO_2$ | 50.00 | 100.0 | 58.33 | 70.83 | 90.00 | 91.67 | 50.00 | 66.67 | 50.00 | 50.00 | 66.67 | **83.33** |
| Mean | 57.22 | 100.0 | 80.75 | 79.66 | 92.39 | 96.68 | 51.25 | 73.80 | 70.72 | 70.56 | 87.16 | **93.74** |
| Total | 45.57 | 100.0 | 85.60 | 76.49 | 93.23 | 97.83 | 39.34 | 74.59 | 80.33 | 72.95 | 90.16 | **92.62** |

proportion of 30–70%. The average accuracy of the 6 classes for each model and the classification accuracy for the total train set and test set separately are also given. Similarly, Tables 7 and 8 present the classification accuracy with proportion of 50–50% and 80–20%, respectively.

From Tables 6–8, we can clearly find that with the increasing number of training samples, the classification accuracy increases also. For each model, the discrimination of $NO_2$ was not that successful, two reasons may explain it. The first one is the smaller number of samples. Totally, 30 samples were collected, and unbalanced samples may also influence the classification. Second, the sensitivity of gas sensors to $NO_2$ (oxidizing gas) is negative which is contrary to other five contaminants. The sample of $NO_2$ is easier to be classified as HCHO and $C_6H_6$ from Table 4. Concluded from the digits in bold in Tables 6–8, the HSVM classification with FLDA is always better than other models for HCHO, $C_7H_8$, $NH_3$ and CO. The 100% classification accuracy of $C_7H_8$, $NH_3$ and CO can be obtained on the testing samples by using the HSVM model. Also, the highest 94.23% classification accuracy of HCHO on the testing samples is obtained.

For visualization, Fig. 6 illustrates the classification accuracy of the test samples with three different proportions based on the presented 6 classifiers. We can see that HSVM model performs the best multi-class discrimination. Note that each node (from number 1 to 6) in Fig. 6 denotes one kind of classifier and three kinds of symbols ("square", "circle" and "triangle") represent three different proportions, respectively. The single SVM classifier performs the second best when the training-testing proportion is 50–50% and 80–20%. However, with 30% training samples SVM performs worse than SFAM, MLP and FLDA models. While HSVM is obviously superior to all the models considered. It also confirms that with small number of samples HSVM can still show the best classification performance.

To study the classification performance of different models using only four metal oxide semiconductor gas sensors without considering temperature and humidity in feature space, we perform all the classification procedure on the features with only four variables on the training and testing samples with proportion of 80–20%. Table 9 presents the classification accuracy of training and testing samples separately without temperature and humidity integration. We can find that HSVM still performs the best discrimination. Another finding is that the accuracy of $C_6H_6$ decreased for all the models, which means that temperature and humidity are also important as classification features of $C_6H_6$. Fig. 4 also demonstrates that temperature and humidity play an important role in gas sensor array from the first three principal components. Therefore, both temperature and humidity are key features in pattern recognition for improving the classification performance of E-nose. From the datasheets of the sensors, we can also find that temperature and humidity have a great influence to the sensitivity of metal oxide

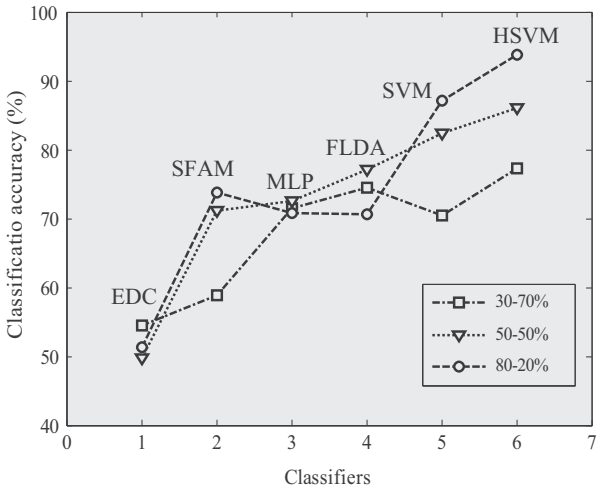


**Fig. 6.** Total classification accuracy on test samples with three proportions using six classifiers labeled as EDC, SFAM, MLP, FLDA, SVM and HSVM (from number 1 to 6).

**Table 9**
Classification accuracy of the train and test set without temperature and humidity integration.

| Class | Classification accuracy (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train set | | | | | | Test set | | | | | |
| | EDC | SFAM | MLP | FLDA | SVM | HSVM | EDC | SFAM | MLP | FLDA | SVM | HSVM |
| HCHO | 21.63 | 100.0 | 88.94 | 61.54 | 90.53 | 94.37 | 28.84 | 86.53 | 94.06 | 69.23 | 92.47 | **94.23** |
| $C_6H_6$ | 45.03 | 100.0 | 78.63 | 60.30 | 65.65 | 63.71 | 51.51 | 75.75 | 49.77 | 66.67 | 64.35 | **69.69** |
| $C_7H_8$ | 75.47 | 100.0 | 92.45 | 94.34 | 98.11 | 100.0 | 38.46 | 84.61 | 96.15 | 100.0 | 96.09 | **100.0** |
| CO | 78.26 | 100.0 | 80.43 | 73.91 | 93.47 | 100.0 | 91.67 | 91.67 | 97.77 | 91.67 | 95.53 | **100.0** |
| $NH_3$ | 56.52 | 100.0 | 73.91 | 78.26 | 78.26 | 87.00 | 60.00 | 84.00 | 92.89 | 88.00 | 87.94 | **96.00** |
| $NO_2$ | 50.00 | 100.0 | 95.83 | 62.50 | 95.83 | 81.40 | 50.00 | 66.67 | 66.57 | 66.67 | 67.02 | **100.0** |
| Mean | 54.49 | 100.0 | 85.03 | 71.81 | 86.97 | 87.75 | 53.41 | 81.54 | 82.87 | 80.37 | 83.90 | **93.32** |
| Total | 42.26 | 100.0 | 85.36 | 66.80 | 84.60 | 86.25 | 46.81 | 82.97 | 82.83 | 76.60 | 84.60 | **90.07** |

semiconductor sensors as the key environmental elements. Therefore, it is better to integrate both two variables in data treatments for classification.

### 4.2. Discussion

This paper mainly investigated a multi-class problem of E-nose data using linear and nonlinear classification methods. In the hybrid HSVM discrimination model, the FLDA is developed as a pre-classification which uses a transformation matrix to reconstruct new patterns with more variables associated with each sub-classifier while not for dimension reduction referred in the previous study. Concretely, this work aims to obtain variables correlated with each sub-classifier through a projection matrix $\mathbf{W}_{6\times15}$ of FLDA, where 6 denotes the number of variables and 15 denotes the number of sub-classifiers in a six-classes classification problem in terms of "one-against-one" strategy. Note that the projection matrix **W** should be obtained through the original data set beforehand, thus, the pre-classification cannot influence the characteristic of real-time classification. From the results of PCA (see Fig. 3), we know that the E-nose data of contaminants in our project belongs to a linearly inseparable case. The HCHO and $C_6H_6$ data have completely overlapped in the data space and hardly been discriminated with other odorants. Unlike a simple two-class problem, a linear decision method may be enough to solve a practical classification. Therefore, we believe that nonlinear method like SVM should be considered in complex discrimination problems and employ gases classification in a higher dimensional feature space transformed by a nonlinear kernel function. Due to the correlations among the variables in original data space, the classification task will also become difficult, thus, FLDA was used to project the original data space onto a new feature space with more linear independent variables related with each classifier and enhance the discriminatory power. Consequently, the linearly independent variables in the new data space after projection on the 15 classifiers can help to implement better classification of multi-class SVM. From the results of the six classifiers considered, EDC performs badly, which is consistent with its principle that the nearest sample vector apart from the centroid of class $k$ was automatically assigned to this class, in which the centroid of each class is the mean of all training vectors in that class. Seen from Fig. 3 we can know that the high misclassification rate is believable because of the completely confused E-nose data sets. That is, Euclidean distance cannot effectively differentiate two odorants in such data confusion without using modern pattern recognition methods, and false discrimination become possible. As we thought before, SFAM performs perfectly in the training process, but unpromising in the testing process. During SFAM training, the weights $\boldsymbol{w}$ are updated in terms of the new training samples, and the $\boldsymbol{w}$ would be well fitted with the training samples. However, when a new test sample was tested on the well trained weights, the weights may not be fitted with the new sample (the value of the matching function is less than the vigilance parameter $\rho$) because the new sample is not trained, and thus gives a false discrimination. In addition, SFAM was developed mostly for a fast on-line training based on a vigilance parameter which controls the update of weights $\boldsymbol{w}$. In real-time application, this method should be developed further for online use in the future. With knowledge that even though MLP neural network based on BP learning rule performs better, it also has the risk of overfitting for its empirical risk minimum criterion. In MLP classification, we use binary codes $(0, 0, 1)^T$, $(0, 1, 0)^T$, $(0, 1, 1)^T$, $(1, 0, 0)^T$, $(1, 0, 1)^T$ and $(1, 1, 0)^T$ with 3 bits to represent the label of each gas and expect to achieve a much better recognition performance than simple representation of decimal numbers from 1 to 6. Note that the results using latter decimal way were not presented because it was not the key study of this paper.

From the chemical characteristic of gases, HCHO, $C_6H_6$, and $C_7H_8$ belong to the organic class, CO, $NH_3$, and $NO_2$ belong to the inorganic class, and the three inorganic gases can be linearly separable from the PCA results (see Fig. 5) which is superior to the organic class. The reasons have been concluded from two angles: chemical properties of the odorants and the sensor selection. The used sensors are more sensitivity to CO, $NH_3$ and $NO_2$. The sensor selection plays an important role in classification of E-nose data. However, unlike electrochemical sensor, metal oxide semiconductor sensor has weak selectivity and is widely used based on the cross-sensitivity of the sensor array combined with modern pattern recognition techniques. For classification of the six contaminants indoor, simple or single pattern recognition methods will result in high misclassification rate because of the serious overlapping between the organic and inorganic classes. Besides, seen from Tables 3 and 4, the inorganic gases are easily misclassified as organic classes. The tendency of classification may result from the uneven samples for organic and inorganic classes. That is, the class with fewer samples is easier to be discriminated as the class with more samples. The solution of uneven samples in classification should also be employed in the future.

Thus, nonlinear discrimination method such as SVM is very necessary for solution of a complex multi-class problem in E-nose application for its structural risk minimization principle. Even though the sensor selection may help E-nose to realize accurate classification, a good pattern recognition method would also be necessary due to the cost of sensor array in the future application. It is noteworthy that in the training process of SVM, the parameter optimization (e.g. regularization and kernel parameters) adopts LOO-CV cross-validation by using a grid-searching method for model selection in this paper. For uniformity of comparison, we adopt the common used grid-search method in SVM and the HSVM. Here, we refer interested readers to the LS-SVMlab Toolbox version 1.8 for study including its details [40].

## 5. Conclusions

In this paper, we studied the potential applicability of an E-nose in classification of air contaminants indoor using different data treatment methods. Six classification models including the EDC, SFAM, MLP, individual FLDA, single SVM and the HSVM model have been developed on experimental electronic nose data sets measured using six kinds of air contaminations indoor including formaldehyde, benzene, toluene, carbon monoxide, ammonia and nitrogen dioxide. The experimental results demonstrate that HSVM has the best classification performance in terms of three different training-testing proportions (30–70%, 50–50% and 80–20%) compared with other classifiers in detection of indoor air contaminants. Take proportion of 80–20% as an example, the average and total test accuracy for classification of the six contaminants achieves 93.74% and 92.62% which are higher than 87.16% and 90.16% obtained using SVM, respectively.

The results of this study demonstrate that the HSVM model has better performance than the ordinary methods in classification. HSVM model may be more effective and applicable in indoor air contaminants monitoring by an E-nose in the future.

## Acknowledgements

We would like to express our sincere appreciation to the anonymous reviewers for their insightful comments, which have greatly improved the quality of the paper.

This work was supported by the Key Science and Technology Research Program (Nos. CSTC2010AB2002, CSTC2009BA2021) and Chongqing University Postgraduates' Science and Innovation Fund (No. CDJXS12160005).

## References

[1] M. Peris, L. Escuder-Gilabert, A 21st century technique for food control: electronics noses, Analytica Chimica Acta 638 (2009) 1–15.
[2] J.W. Gardner, H.W. Shin, E.L. Hines, C.S. Dow, An electronic nose system for monitoring the quality of potable water, Sensors and Actuators B 69 (2000) 336–341.
[3] Q. Ameer, S.B. Adeloju, Polypyrrole-based electronic noses for environmental and industrial analysis, Sensors and Actuators B 106 (2005) 541–552.
[4] A. Lamagna, S. Reich, D. Rodrigurez, A. Boselli, D. Cicerone, The use of an electronic nose to characterize emissions from a highly polluted river, Sensors and Actuators B 131 (2008) 121–124.
[5] C. Di Natale, A. Macagnano, E. Martinelli, R. Paolesse, G. D'Arcangelo, C. Roscioni, A. Finazzi-Agro, A.D'Amico, Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors, Biosensors and Bioelectronics 18 (2003) 1209–1218.
[6] M. Bernabei, G. Pennazza, M. Santortico, C. Corsi, C. Roscioni, R. Paolesse, C. Di Natale, A. D'Amico, A preliminary study on the possibility to diagnose urinary tract cancers by an electronic nose, Sensors and Actuators B 131 (2008) 1–4.
[7] H.M. Zhang, J. Wang, Detection of age and insect damage incurred by wheat with an electronic nose, Journal of Stored Products Research 43 (2007) 489–495.
[8] E.A. Baldwin, J. Bai, A. Plotto, S. Dea, Electronic noses and tongues applications for the food and pharmaceutical industries, Sensors 11 (2011) 4744–4766.
[9] A.H. Gomez, J. Wang, G.X. Hu, A.G. Pereira, Monitoring storage shelf life of tomato using electronic nose technique, Journal of Food Engineering 85 (2008) 625–631.
[10] H.M. Zhang, M.X. Chang, J. Wang, S. Ye, Evaluation of peach quality indices using an electronic nose by MLR, QPST and BP network, Sensors and Actuators B 134 (2008) 332–338.
[11] S.J. Dixon, R.G. Brereton, Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure, Chemometrics and Intelligent Laboratory Systems 95 (2009) 1–17.
[12] E. Llobet, E.L. Hines, J.W. Gardner, P.N. Bartlett, Fuzzy ARTMAP based electronic nose data analysis, Sensors and Actuators B 61 (1999) 183–190.
[13] Z. Xu, X. Shi, L. Wang, J. Luo, C.J. Zhong, S. Lu, Pattern recognition for sensor array signals using fuzzy ARTMAP, Sensors and Actuators B 141 (2009) 458–464.
[14] P. Ciosek, W. Wroblewski, The analysis of sensor array data with various pattern recognition techniques, Sensors and Actuators B 114 (2006) 85–93.
[15] Q. Chen, J. Zhao, Z. Chen, H. Lin, D.A. Zhao, Discrimination of green tea quality using the electronic nose technique and the human panel test, comparison of linear and nonlinear classification tools, Sensors and Actuators B 159 (2011) 294–300.
[16] B. Debska, B. Guzowska-Swider, Application of artificial neural network in food classification, Analytica Chimica Acta 705 (2011) 283–291.
[17] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D.L. Massart, S. Heuerding, F. Erni, Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis, applied to NIR data, Analytica Chimica Acta 329 (1996) 257–265.
[18] K. Brudzewski, S. Osowski, T. Markiewicz, Classification of milk by means of an electronic nose and SVM neural network, Sensors and Actuators B 98 (2004) 291–298.
[19] L.H. Chiang, M.E. Kotanchek, A.K. Kordon, Fault diagnosis based on Fisher discriminant analysis and support vector machines, Computers and Chemical Engineering 28 (2004) 1389–1401.
[20] G.A. Carpenter, S. Grossberg, N. Marcuzon, J.H. Reinolds, D.B. Rosen, Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps, IEEE Transactions on Neural Networks 3 (1992) 698–713.
[21] G.A. Carpenter, S. Grossberg, N. Marcuzon, D.B. Rosen, Fuzzy ART: fast stable learning and categorization of analogue patterns by an adaptive resonance system, Neural Networks 4 (1991) 759–771.
[22] J.W. Gardner, E.L. Hines, M. Wilkinson, The application of artificial neural networks in an electronic nose, Measurement Science and Technology 1 (1990) 446–451.
[23] E. Llobet, J. Brezmes, X. Vilanova, J.E. Sueiras, X. Correig, Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick film tin oxide gas sensor array, Sensors and Actuators B 41 (1997) 13–21.
[24] C. Maugis, G. Celeux, M.L. Martin-Magniette, Variable selection in model-based discriminant analysis, Journal of Multivariate Analysis 102 (2011) 1374–1387.
[25] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[26] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[27] H.L. Chen, D.Y. Liu, B. Yang, J. Liu, G. Wang, A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis, Expert Systems with Applications 38 (2011) 11796–11803.
[28] A.P. Jones, Indoor air quality and health, Atmospheric Environment 33 (1999) 4535–4564.
[29] K. Sakai, D. Norbäck, Y. Mi, E. Shibata, M. Kamijima, T. Yamada, Y. Takeuchi, A comparison of indoor air pollutants in Japan and Sweden: formaldehyde, nitrogen, dioxide, and chlorinated volatile organic compounds, Environmental Research 94 (2004) 75–85.
[30] S.C. Lee, M. Chang, Indoor and outdoor air quality investigation at schools in Hong Kong, Chemosphere 41 (2000) 109–113.
[31] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B 129 (2008) 750–757.
[32] M. Blaschke, T. Tille, P. Robertson, S. Mair, U. Weimar, H. Ulmer, MEMS gas-sensor array for monitoring the perceived car-cabin air quality, IEEE Sensors Journal 6 (2006) 1298–1308.
[33] W.Y. Chung, S.C. Lee, An air quality sensor system with a momentum back propagation neural network, Journal of the Korean Physical Society 49 (2006) 1087–1091.
[34] C.K. Loo, A. Law, W.S. Lim, M.V.C. Rao, Probabilistic ensemble simplified fuzzy ARTMAP for sonar target differentiation, Neural Computing and Applications 15 (2006) 79–90.
[35] S. Haykin, Neural Networks, a Comprehensive Foundation, Macmillan, New York, 2002.
[36] L. Zhang, F.C. Tian, C. Kadri, B. Xiao, H. Li, L. Pan, H. Zhou, On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality, Sensors and Actuators B 160 (2011) 899–909.
[37] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks 13 (2002) 415–425.
[38] F. Sales, M.P. Callao, F.X. Rius, Multivariate standardization for correcting the ionic strength variation on potentiometric sensor arrays, Analyst 125 (2000) 883–888.
[39] J. Karhunen, Generalization of principal component analysis optimization problems and neural networks, Neural Networks 8 (1995) 549–562.
[40] J. Vandewalle, J.A.K. Suykens, http://www.esat.kuleuven.be/sista/lssvmlab/, 2011.

## Biographies

**Lei Zhang** received his bachelor degree in electrical/electronics engineering in 2009 from the Nanyang Institute of Technology, China; from September 2009 to December 2010, he studied for a MS degree in signal and information processing. He is presently with Chongqing University, pursuing his Ph.D. degree in circuits and systems. His research interests include computational intelligence, artificial olfactory system, and nonlinear signal processing in electronic nose.

**Fengchun Tian** received Ph.D. degree in 1997 in electrical engineering from Chongqing University. He is currently a professor with the College of Communication Engineering of Chongqing University. His research interests include electronic nose technology, artificial olfactory systems, pattern recognition, chemical sensors,

signal/image processing, wavelet, and computational intelligence. In 2006 and 2007, he was recognized as a part-time professor of GUELPH University, Canada.

**Hong Nie** studied in major of analytical chemistry from 1981 to 1984 in institute of industry, Chongqing. In 2006, she has become a senior engineer of Academy of Metrology and Quality Inspection, Chongqing. Her research interest was mainly analytical chemistry.

**Lijun Dang** received her Bachelor degree in School of Electronic and Information Engineering in 2011 from the Dalian University of Technology, China; from September 2011 to June 2012, she studied for a MS degree in circuits and system. Her research interests include circuits and system design in electronic nose technology.

**Guorui Li** received his bachelor degree in College of Communication Engineering in 2010 from the Chongqing University, China; from September 2010 to June 2012, he studied for a MS degree in circuits and system. His research interests include signal processing in electronic nose technology.

**Qi Ye** received his bachelor degree in College of Communication Engineering in 2009 from the Chongqing University, China; from September 2010 to June 2012, he studied for a MS degree in circuits and system. His research interests include circuits and system design in electronic nose technology.

**Chaibou Kadri** received his bachelor degree in electrical/electronics engineering in 2001 from the Federal University of Technology Bauchi, Nigeria; his MS degree in communication and information system in 2009, from Chongqing University China. He is presently with Chongqing University, pursuing his Ph.D. degree in circuits and systems. His research interests include signal processing for gas sensors array instruments, and machine learning.