Sensors and Actuators B 220 (2015) 115–124

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

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

Improving the transfer ability of prediction models for electronic noses

Ke Yan [a], David Zhang [b,*]

[a] Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[b] Biometric Research Centre, Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 11 December 2014
Received in revised form 18 April 2015
Accepted 21 May 2015
Available online 3 June 2015

Keywords:
Electronic nose
Calibration transfer
Device standardization
Breath analysis
Tikhonov regularization

ABSTRACT

Calibration transfer is attracting more and more attention in the field of electronic noses (e-noses). It aims at making the prediction model trained on one device transferable to other devices, which is important for the large-scale deployment of e-noses, especially when the cost of sample collection is high. In this paper, the transfer ability of prediction models is improved in two steps. First, windowed piecewise direct standardization (WPDS) is used to standardize the slave device, i.e. to transform the variables from the slave device to match the master one. Then, data from the master device are used to develop prediction models with a novel strategy named standardization error based model improvement (SEMI). Finally, the standardized slave data can be predicted by the models with a better accuracy. The proposed WPDS is a generalization of the widely used PDS algorithm. The main idea of SEMI is to make the trained models rely more on variables with small standardization errors, thus less sensitive to the inconsistency of the devices. It links the standardization step and the prediction step. To evaluate the algorithms, three e-noses specialized for breath analysis are adopted to collect a dataset, which contains pure chemicals and breath samples. Experiments show that WPDS outperforms previous methods in the sense of standardization error and prediction accuracy; SEMI consistently enhances the accuracy of the master model applied to standardized slave data. This study provides effective and extensible methods for model transfer of e-noses.

1. Introduction

Electronic noses (e-noses) have become effective tools in many areas, such as air quality monitoring [1,2], quality control of food [3], and clinical analysis [4–7]. As increasing number of e-nose systems are being deployed in real-life applications, the problem of calibration transfer is receiving more and more attention. When two e-noses of the same model are used to measure the same gas sample, their responses are usually not identical, which is due to the variations in the manufacture of gas sensors, e-nose devices, and the change in operational condition [8–10]. Therefore, if the prediction model trained on one device (master device) is applied to other devices (slave devices), there will be a degradation in accuracy. However, it is often impractical to collect a set of gas samples with each device to develop prediction models, especially when the cost of sample collection is high. This problem limits the popularization of e-noses.

In order to make prediction models more applicable on slave devices, researchers have presented various methods. Many of them were originally proposed for spectroscopic data [11–13], but can also be applied to e-noses. There are three typical ways of calibration transfer [9,13]: transforming the data from the slave device to match the master one; updating the prediction model of the master device according to the slave data; and transforming the predicted values of the slave data. In the field of e-nose, focuses have been paid on the first way [8,9,14–17], since it is feasible in most situations and easy to implement. This kind of methods are also known as device standardization methods, which essentially deal with a regression problem. Common categories include univariate direct standardization (UDS), direct standardization (DS), and piecewise direct standardization (PDS), which differ mainly in the number of input variables. Regression algorithms such as robust fitting [8,17], artificial neural network (ANN) [14,16], partial least squares (PLS) [15], ordinary least square (OLS), and principal component regression (PCR) [12] have been studied. Besides, in [18], standardization was performed on a subspace obtained by

* Corresponding author. Tel.: +852 27667271.
E-mail addresses: yank10@mails.tsinghua.edu.cn (K. Yan), csdzhang@comp.polyu.edu.hk (D. Zhang).

http://dx.doi.org/10.1016/j.snb.2015.05.060
0925-4005/

spectral regression. The method is better than DS when the number of transfer samples is not less than 20. Inspired by the connectivity strategy of the olfactory bulb, Polese et al. [19] developed a method with two self organizing map (SOM) layers. The method is effective for the calibration transfer of optical chemical sensors. A calibration transfer approach based on alternating trilinear decomposition (ATLD) was proposed in [20]. With the method, the correction coefficients of multiple devices can be simultaneously derived. But the method may only be suitable when the changes between devices are restricted to relative intensity.

In the widely used PDS method, one variable in the master device is fitted by a group of variables around the corresponding variable in the slave device. All input variables are given the same weight [11]. However, it is intuitive that the variables nearer to the corresponding variable should receive higher weights than the farther ones. With the constraint of the feature weights, the regression algorithm can be more stable. So we propose windowed piecewise direct standardization (WPDS) in this paper, which allows us to give different weights to the input variables by assigning different penalty parameters. Experimental results show that WPDS outperforms UDS, PDS, and DS in the sense of validation standardization error (the difference between standardized slave variables and the master variables) and prediction accuracy.

In current literatures, device standardization and prediction model training are always considered separately. One improves the transfer ability of prediction models only by minimizing the standardization error (SE). Nevertheless, we find that by incorporating some prior information obtained from standardization into the prediction models, the prediction accuracy of the slave data can also be enhanced. We call the strategy standardization error based model improvement (SEMI). The main idea is to make the models rely more on stable variables which have smaller SE. The strategy is combined with four popular prediction algorithms, i.e. logistic regression, support vector machine, ridge regression, and support vector regression. A weighted regularization term is included in the objective function of each algorithm. We impose larger penalty on the variables with larger SE, so as to reduce the weights of these variables in the trained model. Therefore, the model will be less sensitive to these unstable variables and have better transfer ability.

Calibration transfer is crucial in the application of clinical analysis because samples from patients are rather hard to collect. In our previous work, we introduced a portable e-nose specialized for breath analysis [7]. It achieves disease screening and monitoring through analyzing the biomarkers in breath, such as acetone, hydrogen, and ammonia. Three e-noses of this model are adopted to collect a gas sample dataset. Six pure chemical samples are chosen as transfer samples for device standardization. Several prediction tasks are designed to evaluate the transfer ability of the models, including classification or regression of pure chemicals or breath samples. Experimental results show that the SEMI strategy consistently enhances the accuracy of the master model applied to standardized slave data, especially when the inconsistency between devices is large. Despite its efficacy, SEMI can be easily extended to other prediction algorithms.

The paper is organized as follows. Section 2 describes WPDS and SEMI in detail. Section 3 introduces the experimental configurations, including the e-nose module, dataset, and related data analysis procedure. Section 4 presents the results of the calibration transfer experiments and provides some discussion. Section 5 concludes the paper.

2. Methods design

The calibration transfer process in the paper consists of two steps: (1) developing standardization models with WPDS to standardize the data from the slave device; (2) developing prediction models with the SEMI strategy to predict the standardized slave data. This section will describe the steps in detail.

2.1. Windowed piecewise direct standardization (WPDS)

The objective of standardization is to model the difference between two devices and reduce it. To achieve this, a set of transfer samples are measured on both devices. Then regression models are built based on these transfer samples, so as to transform each slave variable to match the corresponding master variable. Finally, the prediction models trained on master data can be applied to the standardized slave data and get a better accuracy.

In the simple univariate direct standardization (UDS) approach [14], each master variable is fitted using the corresponding slave variable and obtain two coefficients: the slope and the intercept. When the device variation is large, the univariate approach cannot always model the master variables well. The direct standardization (DS) proposed in [11] is a multivariate approach, which fits each master variable using all slave variables. Some researchers [14] reported that DS is better than UDS. However, when the number of variables is large and the number of transfer samples is limited, DS is prone to overfitting [13]. A trade-off approach between UDS and DS is piecewise direct standardization (PDS) [11]. In PDS, each master variable is related to only a subset of slave variables, for example, neighboring wavelengths in near-infrared spectroscopy data [11]. PDS is one of the most widely used standardization approaches in spectroscopic area. Its superiority is attributed to its local character and multivariate nature [13]. But it has not been well explored for e-nose data partially due to the feature extraction methods used in previous studies. Commonly, only one steady response feature is extracted from each sensor before standardization, hence there are no "neighboring" variables. If multiple transient features are extracted from each sensor's response curve, neighboring variables can be defined and PDS can be applied.

In PDS, the input variables are regarded as equally important. Intuitively, when fitting the kth master variable, the kth slave variable should be more important than the variables at some distance from k. Therefore, we propose a windowed PDS (WPDS) which gives different weights to input variables by assigning different penalty parameters in regression. The penalty parameters can be seen as a window around k. By changing the size and shape of the window, we can change the scope and weights of the input variables. Consequently, the original PDS turns out to be a special case of WPDS with a rectangular window (constant weights).

We adopt generalized ridge regression as the algorithm inside WPDS. Ridge regression [21,22] is a well-known shrinkage method for linear regression. Suppose the problem is to find proper $\boldsymbol{\beta}$ and β_0 in

$$y^{(i)} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}^{(i)} + \beta_0 + \varepsilon^{(i)}, \quad i = 1, 2, \ldots, N, \tag{1}$$

where y is the output variable; superscript (i) indicates the ith sample; N is the number of samples; $\boldsymbol{x}^{(i)} \in \mathbf{R}^M$ is a vector of M input variables; M is the window length of WPDS. $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_M]^{\mathrm{T}} \in \mathbf{R}^M$ and $\beta_0 \in \mathbf{R}$ are the regression coefficients to be estimated; and $\varepsilon^{(i)} \in \mathbf{R}$ is an error term. The problem formulation of ridge regression is

$$\min_{\boldsymbol{\beta},\beta_0} \left\{ \sum_{i=1}^{N} \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}^{(i)} + \beta_0 - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{M} \beta_j^2 \right\}. \tag{2}$$

The second term is a regularization term, which imposes a penalty on the coefficients' size. It forces the coefficients to shrink toward zero. $\lambda \geq 0$ is a parameter controlling the amount of shrinkage. The larger λ, the greater the shrinkage. Note that the intercept β_0 is not included in the regularization term [22].

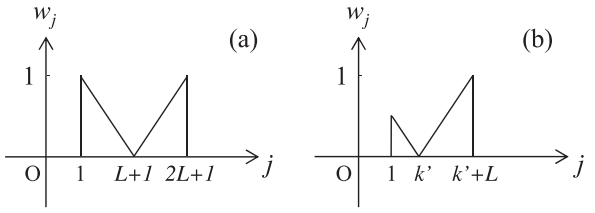

Fig. 1. Illustration of the triangular window used in WPDS. The *x*-axis is the index of the variables. The *y*-axis is the penalty parameter. Plot (a) shows a complete window; Plot (b) shows an example of a truncated window when the number of input variables is less than $2L+1$ (there are less than L variables on either the left or the right side of the kth variable).

In multivariate standardization approaches, the number of samples is limited and the input variables are often correlated. In such cases, the estimated coefficients can have large variance [22]. Ridge regression is particularly useful in such problems. By introducing the regularization term, ridge regression reduces the variance and stabilizes the regression model [22]. The ridge regression model can be generalized by adding a penalty parameter w for each coefficient in the regularization term:

$$\min_{\boldsymbol{\beta},\,\beta_0}\left\{\sum_{i=1}^{N}\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}^{(i)}+\beta_0-y^{(i)}\right)^2+\lambda\sum_{j=1}^{M}(w_j\beta_j)^2\right\}. \tag{3}$$

Large w_j brings large penalty to β_j and shrinks it. As a result, the regression model will depend less on variable j. A triangular window is designed for w_j. Suppose the window length $M=2L+1$, which means that when fitting the kth master variable, from the $(k-L)$th to the $(k+L)$th slave variables will form the input vector $\boldsymbol{x}$. So x_{L+1} is the slave variable which corresponds to the master variable to be fitted. As shown in Fig. 1(a), the penalty parameter w_{L+1} is set to be 0, and w grows linearly to 1 for the L adjacent variables on both sides. This means that we impose the minimum penalty to β_{L+1}; and the further the index is away from $L+1$, the larger the penalty. When $k\le L$ or $k>P-L$ (P is the total number of variables), the window should be truncated on one side since the number of input variables is less than $2L+1$, as shown in Fig. 1(b). It is worth noting that the proper shape of the window is dependent on the feature extraction method. The triangular window is suitable when the transient feature is utilized, either in time domain (the points on the sensors' response curves are used as variables) or frequency domain (the Fourier transform coefficients of the response curves are used). In such cases, neighboring variables are closely related. If other feature extraction methods are applied before standardization, the window should be adjusted based on the relationships among the input variables.

To solve Eq. (3), the input and output variables are first centered to eliminate the intercept term β_0 [22]. Then it can be derived that

$$\boldsymbol{\beta}=\left(X^{\mathrm{T}}X+\lambda W\right)^{-1}X^{\mathrm{T}}Y,\ W=\mathrm{diag}(w_1^2,\ldots,w_M^2), \tag{4}$$

where $X=[\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(N)}]^{\mathrm{T}}\in\mathbf{R}^{N\times M}$ and $Y=[y^{(1)},\ldots,y^{(N)}]^{\mathrm{T}}\in\mathbf{R}^{N}$.

2.2. Standardization error based model improvement (SEMI)

Compared with other calibration transfer methods, device standardization is widely used since it is easy to implement and feasible in most problems and prediction algorithms. Existing literatures have been treating standardization and prediction model training as two separate steps. In the latter step, the prediction algorithm focuses on fitting the master data, without trying to make the model adapt well to standardized slave data. In this section, we propose a strategy to link these two steps. By incorporating some prior information obtained from standardization, the prediction models can be improved and have better transfer ability.

To enhance the prediction accuracy on standardized slave data, a direct way is to improve the standardization algorithm and minimize the standardization error (SE) of each variable. However, in practice, some sensors' responses are more stable than others when measuring certain gases. The signal-to-noise ratios of different feature extraction methods are also different. Therefore, some variables will have less inter-device variance after standardization, thus have less SE. If the prediction model can rely more on these variables, its performance on standardized slave data will be better.

We define the SE of a variable x to be the root mean square deviation (RMSD) between the standardized slave variable x^{SS} and the master variable x^{Ma}:

$$\mathrm{SE}(x)=\mathrm{RMSD}(x)=\sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}\left(x_i^{\mathrm{Ma}}-x_i^{\mathrm{SS}}\right)^2}, \tag{5}$$

where the superscript "Ma" stands for "master" and "SS" stands for "standardized slave". N_t is the number of transfer samples.

To make the models more dependent on variables with less SE, we modify the objective functions of prediction algorithms to include a weighted regularization term. The method is based on Tikhonov regularization [23,24]. The regularization term has the form $\lambda\sum_{j=1}^{P}(w_j\beta_j)^2$, which is similar to the one in generalized ridge regression for WPDS. β_j is the coefficient of variable j to be estimated in the model; w_j is set to be the SE of variable j; λ is a positive constant controlling the weight given to the term. By minimizing the objective function with this regularization term, the variables with larger SE are given larger penalty parameters. Hence the coefficients (β_j) for these variables in the estimated model will be shrunk, and the model will be less dependent on them. We also tried to replace the L2-norm penalty with L1-norm ($\lambda\sum_{j=1}^{P}|w_j\beta_j|$). It can generate sparse models with some coefficients being exactly zero. But the estimated models are not stable [25] and show worse transfer ability than the L2-norm one.

This standardization error based model improvement (SEMI) strategy is applicable on various prediction algorithms. We combine it with four popular classification or regression algorithms, i.e. logistic regression, support vector machine, ridge regression, and support vector regression. For logistic regression, a regularization term is added into its objective function. In the objective functions of the other three algorithms, a regularization term already exists. So we turn the term into a weighted version. These algorithms will be briefly reviewed in the following subsections. Hereinafter each input vector $\boldsymbol{x}$ is assumed to contain a constant component ($x_0=1$), so the intercept coefficient β_0 is merged into the coefficient vector $\boldsymbol{\beta}$.

2.2.1. Logistic regression based classification

In binary-class cases, the decision function of logistic regression (LR) is a sigmoid function

$$h_{\boldsymbol{\beta}}(\boldsymbol{x})=\mathrm{sigmoid}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})=\frac{1}{1+\exp(-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}. \tag{6}$$

A test sample $\boldsymbol{x}$ is classified into the positive class if $h_{\boldsymbol{\beta}}(\boldsymbol{x})\ge 0.5$. Since $h_{\boldsymbol{\beta}}(\boldsymbol{x})$ is between 0 and 1, it can be viewed as the probability of the test sample belonging to the positive class, which is also the characteristic of LR. The coefficients $\boldsymbol{\beta}$ can be learned by maximum likelihood estimation, which seeks to maximize the log-likelihood function [22]:

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{N}y^{(i)}\log h_{\boldsymbol{\beta}}(\boldsymbol{x}^{(i)})+(1-y^{(i)})\log(1-h_{\boldsymbol{\beta}}(\boldsymbol{x}^{(i)})), \tag{7}$$

where $y^{(i)} = 1$ if $\boldsymbol{x}^{(i)}$ belongs to the positive class and 0 otherwise. Under the SEMI strategy, the LR problem can be formulated as

$$\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \frac{\lambda}{2} \sum_{j=1}^{P} (\mathrm{SE}_j \cdot \beta_j)^2 \right\}, \tag{8}$$

which can be solved using numerical optimization methods. In K-class cases, K LR models are trained using the one-vs-all strategy and $\boldsymbol{x}$ is classified into the class whose decision function has the largest value.

2.2.2. Support vector machine based classification

Support vector machine (SVM) is among the most popular techniques for classification. The main idea of the algorithm is finding a hyperplane to separate the training samples with a maximum margin. It has been proved to generalize well on test samples [26]. The detailed introduction of the algorithm can be found in [26]. The objective function of SVM has the form "loss + penalty" [22], so we modify the penalty term in an L2-loss SVM model and formulate the problem of SVM + SEMI as:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \xi_i^2 + \frac{\lambda}{2} \sum_{j=1}^{P} (\mathrm{SE}_j \cdot \beta_j)^2 \\ &\text{s.t. } y^{(i)} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}^{(i)} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i, \end{aligned} \tag{9}$$

where $y^{(i)} \in \{+1, -1\}$. The trust region Newton method in LIBLINEAR [27] is used in this paper to solve Eq. (9) in the primal form. In K-class cases, the one-vs-all strategy is used.

2.2.3. Ridge regression based regression

Ridge regression generates regression models which are more robust than ordinary least squares, especially in ill-conditioned problems [9,22]. Its formulation under the SEMI strategy and solution are similar to Eqs. (3) and (4):

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{P} (\mathrm{SE}_j \cdot \beta_j)^2 \right\}. \tag{10}$$

$$\boldsymbol{\beta} = (X^{\mathrm{T}} X + \lambda W)^{-1} X^{\mathrm{T}} Y, \quad W = \mathrm{diag}(0, \mathrm{SE}_1^2, \ldots, \mathrm{SE}_P^2). \tag{11}$$

The intercept coefficient β_0 is not penalized, so the first element of W is zero.

2.2.4. Support vector regression based regression

Support vector regression (SVR) is a frequently used regression algorithm with good generalization ability [9]. The details of the algorithm can be found in [28]. Similar to SVM, we modify the penalty term in an L2-loss SVR model and formulate the problem of SVR + SEMI as:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (\xi_i^2 + \xi_i^{*2}) + \frac{\lambda}{2} \sum_{j=1}^{P} (\mathrm{SE}_j \cdot \beta_j)^2 \\ &\text{s.t. } y^{(i)} - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}^{(i)} \leq \varepsilon + \xi_i \\ &\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}^{(i)} - y^{(i)} \leq \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0, \quad \forall i. \end{aligned} \tag{12}$$

The trust region Newton method in LIBLINEAR [27] is used in this paper to solve Eq. (12) in the primal form.

3. Experimental details

Breath analysis for disease screening and monitoring is an important application of e-noses. In this application, calibration transfer is a crucial step. In order to assess the proposed methods, we used three e-noses to collect a gas sample dataset. In this section, the e-nose module, the composition of the dataset, and the related data analysis procedure will be described.

3.1. E-nose module

In our previous work, we introduced a portable e-nose specialized for breath analysis [7,29]. It is equipped with an array of 11 sensors. Among them, we focus on 9 metal oxide semiconductor (MOS) sensors for the detection of volatile organic compounds (VOCs). They have diverse sensitivity spectrums. For example, some sensors (TGS826, TGS2602, GSBT11) are proved to be sensitive to acetone, which is a biomarker of diabetes [30,31]; some (TGS2610-D00, SP3S-AQ2, TGS822) are more sensitive to hydrogen, which has been used to detect functional intestinal disorders [32]; some (TGS826, TGS2602) are sensitive to ammonia, which is associated with renal failure [33]. There are also three sensors (TGS2600, TGS2602, WSP2111) operated under temperature modulation (TM). They are heated by a staircase voltage oscillated between 3V and 6V. In our experiments, we found that GSBT11 is unstable and has a very large SE. So we discarded the sensor. The following analysis is performed on the remaining 8 MOS sensors.

3.2. Dataset

Three e-noses of the same model were utilized to measure 7 groups of gas samples. Three groups of them are pure chemicals of different concentrations; one group is normal breath exhaled by healthy people; and the other three groups are chemicals blended with normal breath. As mentioned in the last section, three kinds of chemicals were considered, namely acetone, hydrogen, and ammonia, since they are typical breath biomarkers of certain diseases. In the last three groups, the chemicals were diluted with exhaled breath instead of clean air. The aim is to make the prediction tasks more challenging, since normal breath already contains the three chemicals of different concentrations [30,32,34], as well as many other interfering VOCs [35]. They are also used to simulate the breath samples of patients. Details of the dataset are listed in Table 1. The ranges of concentrations of the chemicals were determined by their typical concentrations in breath [30–33]. A visualization of the data in groups 1–3 can be found in the supplementary material, which also shows the sensitivity characteristics of the sensors.

Table 1
Composition of the dataset.

Group	Gas sample	#Samples	Notes
1	Acetone	16	8 concentrations (0.1, 0.2, 0.5, 1, 2, 5, 10, 20 ppm), 2 samples per concentration
2	Hydrogen	18	9 concentrations (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50 ppm), 2 samples per concentration
3	Ammonia	14	7 concentrations (0.1, 0.2, 0.5, 1, 2, 5, 10 ppm), 2 samples per concentration
4	Normal breath	80	Collected from 24 healthy subjects in different days, 2–6 samples per subject
5	Acetone + breath	40	8 concentrations (0, 0.2, 0.3, 0.7, 1.7, 3.3, 5.0, 6.7 ppm), 5 samples per concentration, mixed with normal breath from 5 subjects
6	Hydrogen + breath	40	8 concentrations (0, 0.4, 0.7, 1.7, 4.2, 8.3, 12.5, 16.7 ppm), 5 samples per concentration, mixed with normal breath from 5 subjects
7	Ammonia + breath	40	8 concentrations (0, 0.3, 0.8, 1.7, 2.5, 3.3, 4.2, 5.0 ppm), 5 samples per concentration, mixed with normal breath from 5 subjects

All 248 samples were measured using each of the three e-noses. The measurement procedure of each sample lasts for 144 seconds [7]. It includes 4 stages. In the baseline stage (1s), the baseline values of the sensors are recorded. In the injection stage (7s), the sample is drawn from a gas bag to the gas room. In the reaction stage (56s), the responses of the sensors approach their steady states. The gas room is purged with clean air in the purge stage (80s). The sampling frequency of the e-nose is 8 Hz.

Six pure chemical samples were empirically chosen as the transfer samples in this study. We chose two samples of each chemical, one with a low concentration and one with a high concentration, namely acetone (1 and 10 ppm), hydrogen (1 and 50 ppm), and ammonia (1 and 10 ppm). Experiments showed that with the proposed algorithms, these transfer samples can be used to standardize the breath samples with an acceptable accuracy.

3.3. Preprocessing and feature extraction

After a digitized gas sample is obtained, the baseline values are subtracted from the sensor responses to remove baseline drift. Then, discrete Fourier transform (DFT) features are extracted from the response curves. Compared with the traditional steady response feature, DFT contains the transient information of a curve, which will be helpful in prediction. Additionally, the energy of the typical response curve of a gas sensor lies mostly in low frequencies, so we can compress the curve using only a small number of variables in the frequency domain. In this paper, we extract the modulus of the first 30 DFT coefficients from each curve, which has 1152 data points in the time domain. As a result, each feature vector contains 30×8 sensors = 240 variables.

The variables are further normalized to eliminate additive and multiplicative variations among different devices [9]. For each device, the mean and standard deviation values of each variable are calculated from the transfer samples. Then all samples from the device are centered and scaled by these values. This normalization step can be viewed as an initial standardization of the variables.

3.4. Data analysis procedure

The entire data analysis procedure for device standardization and prediction is shown in Fig. 2. After preprocessing and feature extraction, the gas samples from master and slave devices are represented by matrices X_{Ma} and X_{Sl}, respectively. The transfer samples from both devices are adopted to build standardization models. Then X_{Sl} is standardized using the models and the standardized slave sample matrix X_{SS} is obtained. Next, training samples from X_{Ma} are used to develop master prediction models. When the SEMI strategy is used, the standardization error will also join the model training. The trained model is applied to predict the test samples in X_{Ma} and X_{SS}, obtaining the accuracy Acc_{Ma} and Acc_{SS}, respectively. We can also get Acc_{Sl} when the training and test samples are both from X_{Sl}. Generally, Acc_{Ma} and Acc_{Sl} should be better than Acc_{SS}, since their training and test samples are from the same devices. The final goal is to enhance Acc_{SS} to approach Acc_{Sl}, without degrading Acc_{Ma}. Note that for classification tasks, Acc is the classification precision (the larger the better); for regression tasks, Acc is the root mean square error (RMSE, the smaller the better).

Note that in the standardization step, if multivariate standardization methods are used, the input variables and the output variable should come from the same sensor. In the prediction step, 4 classification tasks and 6 regression tasks are executed to evaluate the algorithms. The classification tasks include a multi-class task (distinguishing the three chemicals) and 3 binary-class tasks (distinguishing normal breath samples from each group of blended ones). In each task, we randomly choose an equal number of samples from the each corresponding data group (class). Half of them are randomly chosen as training samples, leaving the rest as test ones. The two classification algorithms in Section 2.2 are applied. The process is repeated 20 times and an average accuracy is computed for each task. The regression tasks aim at predicting the concentration of chemicals in each data group except group 4. Within each group, the leave-one-out cross validation strategy is used. Ridge regression and support vector regression introduced in Section 2.2 are applied.

4. Results and discussion

4.1. Standardization

Among the three e-noses, we assigned device 1 as the master device, device 2 and 3 as the slave ones. Fig. 3 illustrates the relationship between the maximum responses of two devices. It is clear that the maximum responses of the two devices have a linear relationship.

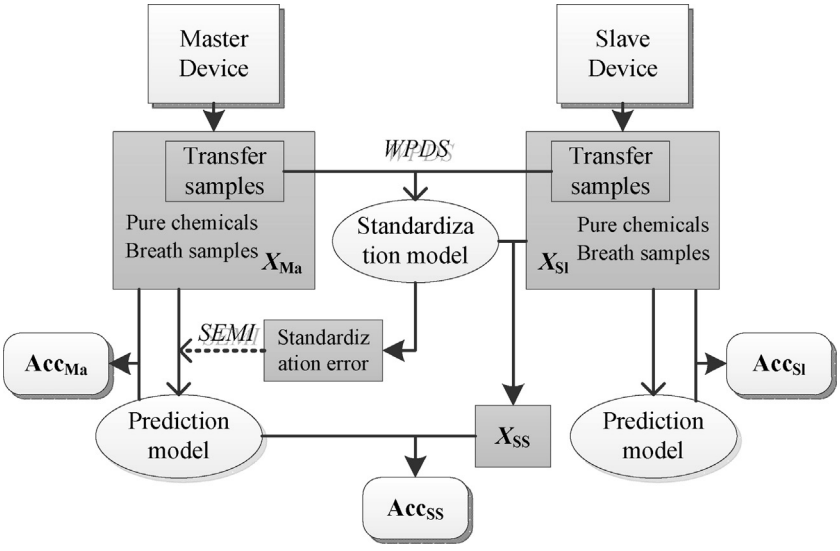

Fig. 2. Flowchart of the device standardization and prediction procedure. An unfilled arrow starts from an input, and a filled arrow points to an output. X_{Ma}, X_{Sl} and X_{SS} represent the sample matrices of master, slave, and standardized slave device, respectively. Acc stands for prediction accuracy.

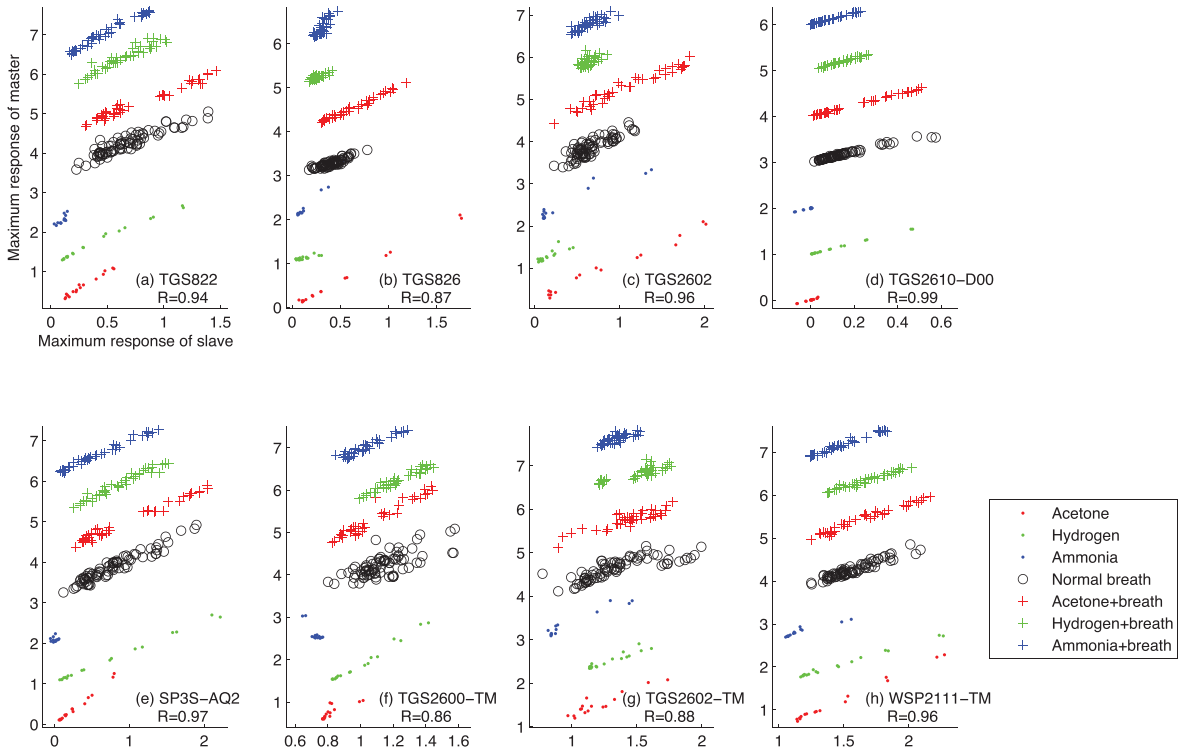

Fig. 3. Illustration of the relationship between the maximum responses of two devices. Plots (a)–(i) correspond to the 8 sensors. The *x*-axis and *y*-axis correspond to the responses of the slave (device 3) and master device (device 1), respectively. Note that there is a manual shift of 1 on the *y*-axis between different data groups. The *R* values are the correlation coefficients of *x* and *y* (without shift) for the samples in all groups.

Experiments were made to compare the performance of different standardization methods, including univariate direct standardization (UDS), direct standardization (DS), piecewise direct standardization (PDS), and the proposed windowed PDS (WPDS). The results are displayed in Table 2. In the method "only normalization", the slave variables are normalized (see Section 3.3) but not standardized. For UDS, the robust fitting algorithm in [8] was used. For DS and PDS, different regression algorithms were tested, such as OLS, PCR, PLS, and ridge regression (Eq. (2)). It was found that generally the order of performance is: ridge > PLS > PCR > OLS. So only the results of ridge regression are listed in Table 2 for clarity. For DS, all 30 variables of each sensor in slave device were used to fit each master variable. The window length of PDS and WPDS was 15. The shrinkage parameter λ in Eq. (2) and Eq. (3) was set to 10. The neural method was also tested with the same parameter setting as in [16].

In Table 2, the standardization error (SE) is not computed on transfer samples as in Eq. (5). Instead, it is computed on all chemical (groups 1–3) or breath (groups 4–7) samples and averaged over all variables. So we denote it as "validation SE". The Acc_{SS} is also listed, including the average accuracy of the 4 classification tasks and the average RMSE of the 6 regression tasks. The prediction algorithms are logistic regression and ridge regression without SEMI. WPDS has the lowest SE and best prediction accuracy. PDS and DS have the second lowest SE. It is found that good Acc_{SS} does not always occur together with low SE. It is possibly because high Acc_{SS} requires the variables important for prediction to have low SE. The displayed validation SE, however, is averaged over all variables. So it does not reflect the prediction results. This observation was also found in [12]. The performance of PDS (neural) is not good, for the neural network algorithm in [16] has nonlinear hidden neurons and is easy to overfit the transfer samples, especially when the number of input variables is large. The triangular window adopted in WPDS brings effective prior information to the problem, which makes it outperform other methods.

Fig. 4 is a visual illustration of the effect of the standardization process. Linear discriminant analysis (LDA) is used to reduce the master samples in three classes to a two-dimension subspace. Then the slave samples before and after standardization (WPDS) were projected onto the subspace. In Fig. 4, it is found that the raw samples from device 1 and 2 are similar. Meanwhile, the difference between device 1 and 3 is larger. It is probably because the

Table 2
Validation standardization error and prediction accuracy of different standardization methods.

Method	Device 2 as slave				Device 3 as slave			
	Validation SE		Acc_{SS}		Validation SE		Acc_{SS}	
	Chemical	Breath	Classifi. accuracy	Regress. RMSE	Chemical	Breath	Classifi. accuracy	Regress. RMSE
Only normalization	0.3668	0.6192	0.8125	3.4676	0.5492	0.6897	0.7045	3.6586
UDS	0.1905	0.4034	0.8762	1.8711	0.3776	0.5508	0.6159	2.7165
DS	0.1346	0.3139	0.7049	2.5188	0.2317	0.4280	0.6357	3.0510
PDS	0.1280	0.2848	0.7458	2.1887	0.2460	0.4235	0.6474	3.1444
PDS (neural)	0.7433	0.9093	0.6825	4.6332	0.8939	1.0148	0.6145	5.1032
WPDS	**0.0979**	**0.2713**	**0.8945**	**1.4658**	**0.2255**	**0.4189**	**0.7348**	**2.6188**

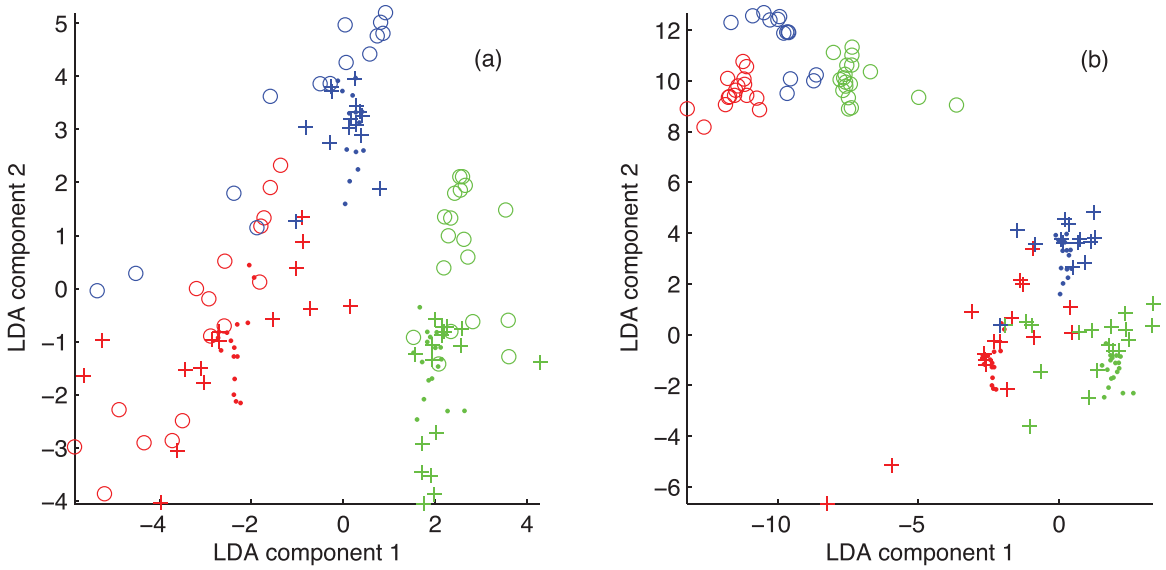

Fig. 4. Linear discriminant analysis (LDA) plot of the samples. The dots, circles, and plus signs represent the samples from the master device, the slave device before standardization, and the slave device after standardization, respectively. The red, green, and blue colors represent samples in three classes (acetone, hydrogen, and ammonia). The slave devices are device 2 and 3 in plots (a) and (b), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

gas route of device 3 is a bit different from the other two devices. Its sensor array is also more aged. These factors enlarge the difference between it and the master device. The standardization method reduces the difference, especially for device 3 (see Fig. 4(b)). It can make the prediction model trained on master data be applied to the standardized slave data without much accuracy loss.

It can be observed that the breath samples have larger SE than the chemical samples. Human breath is a complex mixture of gases. An average normal breath contains 204.2 VOCs [35]. It is not very similar to the transfer samples used in the paper. But since human breath is not reproducible, it cannot be directly used as transfer sample. Experiments in this paper show the possibility of using three chemicals to standardize breath samples. Further study is needed on the selection of transfer samples for breath analysis systems.

4.2. Prediction

The performance of the SEMI strategy is evaluated in this section. In the strategy, the SEs in Eq. (5) are used as penalty parameters in various prediction models. In Fig. 5(a), the SEs of the 240 variables are shown. Their mean is scaled to 1. Note that variables 1–30 are extracted from sensor 1, variables 31–60 are extracted from sensor 2, and so on. For each sensor, the first variable is the DFT feature of zero frequency and the frequency rises as the feature index increases. It can be seen that SEs of different sensors are different. Besides, the high frequency features generally have higher SEs. Fig. 5(b) and (c) shows the coefficients of a logistic regression model trained to distinguish three chemicals. Fig. 5(b) shows the model under the SEMI strategy. Compared with Fig. 5(c), its coefficients have larger magnitude if the corresponding SE is small (e.g. variables 90–120).

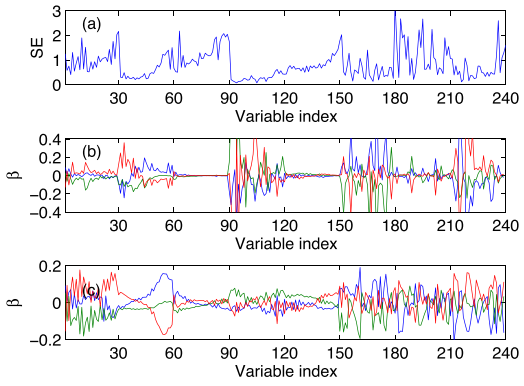

Fig. 5. (a) Standardization error of the variables. (b) The classification model trained by logistic regression with SEMI. The three curves are the coefficients of the decision functions of three chemicals. (c) The classification model trained by logistic regression without SEMI (the penalty parameters are all set to 1).

The classification accuracies based on different configurations are demonstrated in Fig. 6. Both the two slave devices and the two classification algorithms are investigated. The shrinkage parameter λ is varied. When the master models are applied to slave data which is normalized but not standardized, the performance is not good. The accuracy is enhanced after standardization with WPDS (the curve with triangles). The SEMI strategy further enhances it and makes it closer to Acc_{SI}, which is obtained by applying the slave model on slave data. The value of λ is related to Acc_{SS} with SEMI. If λ is too small, the weight of the regularization term is not enough to penalize the unstable variables. If λ is too large, the prediction algorithms will be too focused on penalization and cannot fit the training samples well. The results of the regression algorithms (Fig. 7) show similar trends.

Apart from the DFT feature, the traditional steady response feature was also tested. The SEMI strategy can also enhance its performance. But its overall performance is not as good as the DFT feature. It is probably because steady responses provide much less information than the DFT feature. In addition, SEMI is expected to work better if the dimension of the feature vector is relatively high. In such situations, the variables contain redundant information. When the coefficients of some variables are shrunk, the other variables can still provide sufficient information for prediction.

Detailed results for every prediction algorithm are listed in Tables 3–6. λ was searched among $\{10^{-4}, 10^{-3.5}, \ldots, 10^{4}\}$ for each result to find the best one. For device 2, the improvement of Acc_{SS} brought by WPDS is relatively large. For device 3, however, the effect of device standardization is not good enough. But SEMI significantly improves Acc_{SS} in this case. This proves that SEMI

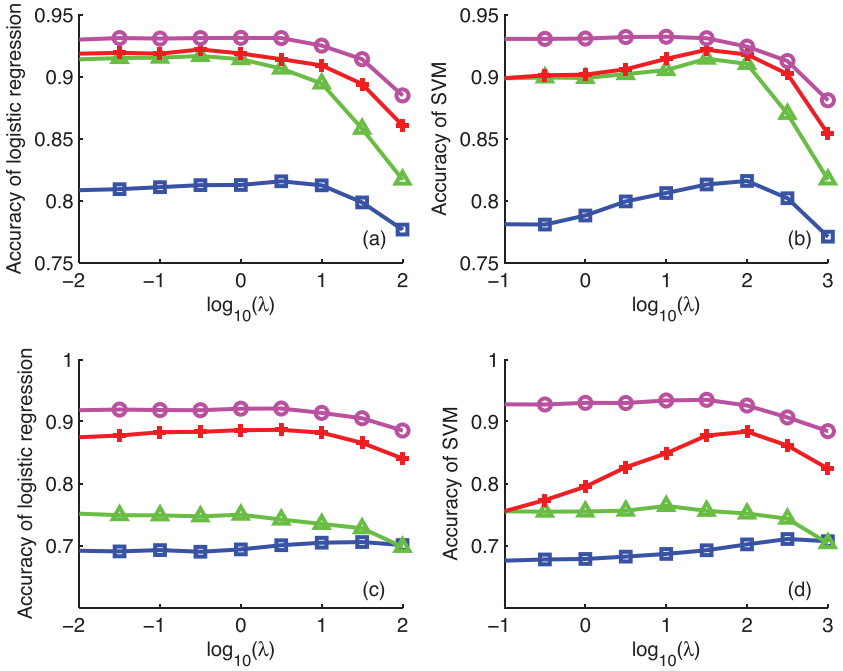

Fig. 6. Average accuracy of the classification tasks. Rows 1 and 2 correspond to device 2 and 3, respectively. Columns 1 and 2 correspond to logistic regression and SVM, respectively. Square: the master model is applied to slave data which is normalized but not standardized; triangle: Acc_{SS} without SEMI (the penalty parameters are all set to 1); plus: Acc_{SS} with SEMI (the mean of the penalty parameters (SEs) are scaled to 1); circle: Acc_{SI}.

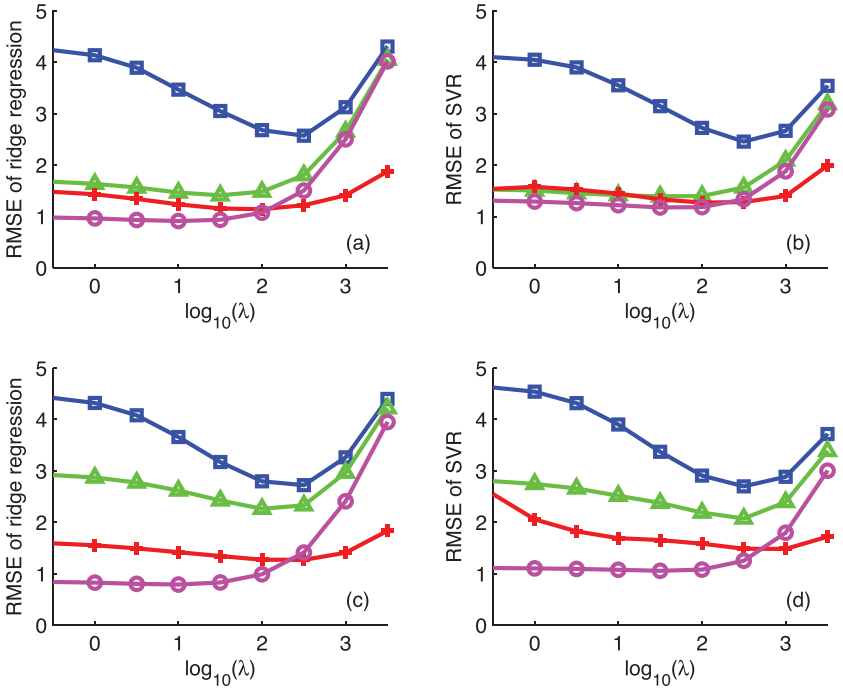

Fig. 7. Average RMSE of the regression tasks. Rows 1 and 2 correspond to device 2 and 3, respectively. Column 1 and 2 correspond to ridge regression and SVR, respectively. Square: the master model is applied to slave data which is normalized but not standardized; triangle: Acc_{SS} without SEMI (the penalty parameters are all set to 1); plus: Acc_{SS} with SEMI (the mean of the penalty parameters (SEs) are scaled to 1); circle: Acc_{SI}.

Table 3
Classification accuracy of logistic regression.

Slave device	Acc_{Ma}		Acc_{SI}		Acc_{SS}		
	Without SEMI	With SEMI	Without SEMI	With SEMI	Only normalization	WPDS	WPDS + SEMI
Device 2	0.9312	0.9324	0.9312	0.9410	0.8156	0.9167	0.9220
Device 3	0.9312	0.9336	0.9210	0.9314	0.7056	0.7532	0.8865

Table 4
Classification accuracy of SVM.

Slave device	Acc_{Ma}		Acc_{SI}		Acc_{SS}		
	Without SEMI	With SEMI	Without SEMI	With SEMI	Only normalization	WPDS	WPDS + SEMI
Device 2	0.9313	0.9331	0.9322	0.9395	0.8160	0.9144	0.9217
Device 3	0.9313	0.9343	0.9350	0.9364	0.7105	0.7637	0.8836

Table 5
Regression RMSE of ridge regression.

Slave device	Acc_{Ma}		Acc_{SI}		Acc_{SS}		
	Without SEMI	With SEMI	Without SEMI	With SEMI	Only normalization	WPDS	WPDS + SEMI
Device 2	0.8971	0.8810	0.9106	0.7716	2.5704	1.4119	1.1424
Device 3	0.8971	0.8012	0.7891	0.7873	2.7224	2.2629	1.2691

Table 6
Regression RMSE of SVR.

Slave device	Acc_{Ma}		Acc_{SI}		Acc_{SS}		
	Without SEMI	With SEMI	Without SEMI	With SEMI	Only normalization	WPDS	WPDS + SEMI
Device 2	1.0522	1.0739	1.1748	1.0732	2.4545	1.3904	1.2690
Device 3	1.0494	1.0157	1.0559	1.0558	2.7004	2.0737	1.4824

Table 7
Accuracy of logistic regression on device 2 of all classification tasks (distinguishing the samples in different data groups).

	Group 1 vs. 2 vs. 3	4 vs. 5	4 vs. 6	4 vs. 7	Average
Acc_{SS}: only normalization	0.7286	0.8941	0.8824	0.7574	0.8156
Acc_{SS}: WPDS + SEMI	0.9762	0.8941	0.9206	0.8971	0.9220
Acc_{SI}	0.9881	0.8779	0.9176	0.9412	0.9312

Table 8
RMSE of ridge regression on device 2 of all regression tasks (predicting the concentration of chemicals in different data groups).

	Group 1	2	3	5	6	7	Average
Acc_{SS}: only normalization	2.8982	5.2685	1.4605	0.6872	3.9549	1.1529	2.5704
Acc_{SS}: WPDS + SEMI	1.1109	1.4875	0.9921	0.6679	1.6318	0.9641	1.1424
Acc_{SI}	0.7763	0.7457	0.8403	0.3552	1.9783	0.7676	0.9106

is especially effective when the inconsistency between devices is large. Surprisingly, Acc_{Ma} and Acc_{SI} can also be slightly improved by applying the SEMI strategy (including the SE weighted regularization term in the prediction models), which implies that SEMI introduces helpful information about the variables (such as their stability) to the prediction algorithms. Among the 4 prediction algorithms, logistic regression and SVM are comparable; ridge regression performs relatively better than SVR. So we further list the results of logistic regression and ridge regression on all prediction tasks with the best λ settings in Tables 7 and 8. It can be observed that the proposed methods have made Acc_{SS} close to Acc_{SI}. Acc_{SS} is even better than Acc_{SI} in some classification and regression tasks due to the helpful information introduced by SEMI. The confusion matrix of logistic regression can be found in the supplementary material.

5. Conclusion

This paper is dedicated to making the prediction models of e-noses more transferable. Efforts were made in two aspects. First, the windowed piecewise direct standardization (WPDS) algorithm based on generalized ridge regression was proposed. Experiments showed that WPDS outperformed previous methods in the sense of validation standardization error (SE) and prediction accuracy. It was also found that validation SE is not proportional to prediction accuracy, which implies that standardization should be studied together with prediction in calibration transfer problems. Second, a novel strategy named standardization error based model improvement (SEMI) was applied in the prediction step. It incorporates a Tikhonov regularization term in the objective functions of prediction algorithms, so as to make the trained models more relied on stable variables, thus relatively not sensitive to the device inconsistency. It fills the gap between standardization and prediction by effectively combining the information obtained from the former step with the latter step. Experiments confirmed that it could enhance the accuracy of the master model applied to standardized slave data, especially when the inconsistency between devices is large.

The proposed methods are easily extensible. WPDS is also applicable for spectroscopic data. Its window can be adjusted to adapt to different feature extraction algorithms. The SEMI strategy can also be combined with various prediction algorithms other than the four explored in this paper. Future works may include extending the strategy to nonlinear algorithms and subspace algorithms.

Clinical analysis is an important application of e-noses. However, literatures about the calibration transfer in this application are rare. In this paper, we investigated the use of only six chemical samples as transfer samples in a breath analysis system. Experiments showed that with the algorithms proposed in this paper, the accuracy after calibration transfer is largely improved. Further study is still needed to choose an optimized set of transfer samples and test the algorithms with real patients' breath under different temperature and humidity.

Acknowledgements

The work is partially supported by the GRF fund from the HKSAR Government, the central fund from Hong Kong Polytechnic University, the NSFC fund (61332011, 61020106004, 61272292, 61271344), Shenzhen Fundamental Research fund (JCYJ2013 0401152508661), and Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China. The authors would like to thank Yujing Ning and the colleagues in HIT for their help in the collection of the dataset. They would also like to thank Dr. Zhenhua Guo and the anonymous reviewers for their valuable comments.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2015.05.060

References

[1] E. Martinelli, E. Zampetti, S. Pantalei, F.L. Castro, M. Santonico, G. Pennazza, R. Paolesse, G. Mascetti, V. Cotronei, F. Giannini, C. Di Natale, A. D'Amico, Design and test of an electronic nose for monitoring the air quality in the international space station, Microgravity Sci. Technol. 19 (2007) 60–64.
[2] L. Zhang, F. Tian, H. Nie, L. Dang, G. Li, Q. Ye, C. Kadri, Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine, Sens. Actuators B: Chem. 174 (2012) 114–125.
[3] L.P. Pathange, P. Mallikarjunan, R.P. Marini, S. O'Keefe, D. Vaughan, Non-destructive evaluation of apple maturity using an electronic nose system, J. Food Eng. 77 (2006) 1018–1023.
[4] F. Röck, N. Barsan, U. Weimar, Electronic nose: current status and future trends, Chem. Rev. 108 (2008) 705–725.
[5] A. D'Amico, C. Di Natale, R. Paolesse, A. Macagnano, E. Martinelli, G. Pennazza, M. Santonico, M. Bernabei, C. Roscioni, G. Galluccio, R. Bono, E. Finazzi Agrò, S. Rullo, Olfactory systems for medical applications, Sens. Actuators B: Chem. 130 (2008) 458–465.
[6] D. Guo, D. Zhang, N. Li, L. Zhang, J. Yang, A novel breath analysis system based on electronic olfaction, IEEE Trans. Biomed. Eng. 57 (2010) 2753–2763.
[7] K. Yan, D. Zhang, D. Wu, H. Wei, G. Lu, Design of a breath analysis system for diabetes screening and blood glucose level prediction, IEEE Trans. Biomed. Eng. 61 (2014) 2787–2795.
[8] L. Zhang, F. 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, Sens. Actuators B: Chem. 160 (2011) 899–909.
[9] S. Marco, A. Gutiérrez-Gálvez, Signal and data processing for machine olfaction and chemical sensing: a review, IEEE Sens. J. 12 (2012) 3189–3214.
[10] M. Bruins, J.W. Gerritsen, W.W. van de Sande, A. van Belkum, A. Bos, Enabling a transferable calibration model for metal-oxide type electronic noses, Sens. Actuators B: Chem. 188 (2013) 1187–1195.
[11] Y. Wang, D.J. Veltkamp, B.R. Kowalski, Multivariate instrument standardization, Anal. Chem. 63 (1991) 2750–2756.
[12] K.-S. Park, Y.-H. Ko, H. Lee, C.-H. Jun, H. Chung, M.-S. Ku, Near-infrared spectral data transfer using independent standardization samples: a case study on the trans-alkylation process, Chemometr. Intell. Lab. 55 (2001) 53–65.
[13] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferr, Transfer of multivariate calibration models: a review, Chemometr. Intell. Lab. 64 (2002) 181–192.
[14] M. Balaban, F. Korel, A. Odabasi, G. Folkes, Transportability of data between electronic noses: mathematical methods, Sens. Actuators B: Chem. 71 (2000) 203–211.
[15] O. Tomic, H. Ulmer, J.-E. Haugen, Standardization methods for handling instrument related signal shift in gas-sensor array measurement data, Anal. Chim. Acta 472 (2002) 99–111.
[16] L. Zhang, F. Tian, X. Peng, L. Dang, G. Li, S. Liu, C. Kadri, Standardization of metal oxide sensor array using artificial neural networks through experimental design, Sens. Actuators B: Chem. 177 (2013) 947–955.
[17] S. Deshmukh, K. Kamde, A. Jana, S. Korde, R. Bandyopadhyay, R. Sankar, N. Bhattacharyya, R. Pandey, Calibration transfer between electronic nose systems for rapid in situ measurement of pulp and paper industry emissions, Anal. Chim. Acta 841 (2014) 58–67.
[18] J. Peng, S. Peng, A. Jiang, J. Tan, Near-infrared calibration transfer based on spectral regression, Spectrochim. Acta A 78 (2011) 1315–1320.
[19] D. Polese, E. Martinelli, G. Magna, F. Dini, A. Catini, R. Paolesse, I. Lundstrom, C. Di Natale, Sharing data processing among replicated optical sensor arrays, Sens. Actuators B: Chem. 179 (2013) 252–258.
[20] Y. Liu, W. Cai, X. Shao, Standardization of near infrared spectra measured on multi-instrument, Anal. Chim. Acta 836 (2014) 18–23.
[21] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, New York, 2009.
[23] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, 1977.
[24] J.H. Kalivas, G.G. Siano, E. Andries, H.C. Goicoechea, Calibration maintenance and transfer using Tikhonov regularization approaches, Appl. Spectrosc. 63 (2009) 800–809.
[25] M.Y. Park, T. Hastie, L1-regularization path algorithm for generalized linear models, J. R. Stat. Soc. Ser. B: Stat. Methodol. 69 (2007) 659–677.
[26] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[27] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[28] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[29] K. Yan, D. Zhang, Blood glucose prediction by breath analysis system with feature selection and model fusion, in: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, IEEE, 2014, pp. 6406–6409.
[30] C. Deng, J. Zhang, X. Yu, W. Zhang, X. Zhang, Determination of acetone in human breath by gas chromatography–mass spectrometry and solid-phase microextraction with on-fiber derivatization, J. Chromatogr. B 810 (2004) 269–275.
[31] C. Turner, C. Walton, S. Hoashi, M. Evans, Breath acetone concentration decreases with blood glucose concentration in type I diabetes mellitus patients during hypoglycaemic clamps, J. Breath Res. 3 (2009) 046004.
[32] A. Eisenmann, A. Amann, M. Said, B. Datta, M. Ledochowski, Implementation and interpretation of hydrogen breath tests, J. Breath Res. 2 (2008) 046002.
[33] S. Davies, P. Spanel, D. Smith, Quantitative analysis of ammonia on the breath of patients in end-stage renal failure, Kidney Int. 52 (1997) 223–228.
[34] B. de Lacy Costello, R. Ewen, N.M. Ratcliffe, A sensor system for monitoring the simple gases hydrogen, carbon monoxide, hydrogen sulfide, ammonia and ethanol in exhaled breath, J. Breath Res. 2 (2008) 037011.
[35] M. Phillips, J. Herrera, S. Krishnan, M. Zain, J. Greenberg, R.N. Cataneo, Variation in volatile organic compounds in the breath of normal humans, J. Chromatogr. B 729 (1999) 75–88.

Biographies

Ke Yan received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. His research interests include data analysis and machine learning in biomedical engineering and machine olfaction.

David Zhang (F'09) graduated in Computer Science from Peking University. He received his M.Sc. in Computer Science in 1982 and his Ph.D. in 1985 from the Harbin Institute of Technology (HIT). From 1986 to 1988 he was a postdoctoral fellow at Tsinghua University and then an associate professor at the Academia Sinica, Beijing. In 1994 he received his second Ph.D. in Electrical and Computer Engineering from the University of Waterloo, Ontario, Canada. Currently, he is a Chair professor at the Hong Kong Polytechnic University where he is the Founding Director of the Biometrics Technology Centre (UGC/CRC) supported by the Hong Kong SAR Government in 1998. He also serves as Visiting Chair professor in Tsinghua University, and Adjunct professor in Peking University, Shanghai Jiao Tong University, HIT, and the University of Waterloo. He is the Founder and Editor-in-Chief, International Journal of Image and Graphics (IJIG); Book Editor, Springer International Series on Biometrics (KISB); Organizer, the International Conference on Biometrics Authentication (ICBA); Associate Editor of more than ten international journals including IEEE Transactions and Pattern Recognition; and the author of more than 10 books and 200 journal papers. Professor Zhang is a Croucher senior research fellow, Distinguished Speaker of the IEEE Computer Society, and a fellow of both IEEE and IAPR.