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].