Correcting Instrumental Variation and Time-Varying Drift Using Parallel and Serial Multitask Learning

Ke Yan, David Zhang, *Fellow, IEEE*, and Yong Xu, *Senior Member, IEEE*

Abstract—When instruments and sensor systems are used to measure signals, the posterior distribution of test samples often drifts from that of the training ones, which invalidates the initially trained classification or regression models. This may be caused by instrumental variation, sensor aging, and environmental change. We introduce transfer-sample-based multitask learning (TMTL) to address this problem, with a special focus on applications in machine olfaction. Data collected with each device or in each time period define a domain. Transfer samples are the same group of samples measured in every domain. They are used by our method to share knowledge across domains. Two paradigms, parallel and serial transfer, are designed to deal with different types of drift. A dynamic model strategy is proposed to predict samples with known acquisition time. Experiments on three real-world data sets confirm the efficacy of the proposed methods. They achieve good accuracy compared with traditional feature-level drift correction algorithms and typical labeled-sample-based MTL methods, with few transfer samples needed. TMTL is a practical algorithm framework which can greatly enhance the robustness of sensor systems with complex drift.

Index Terms—Drift correction, machine olfaction, multitask learning (MTL), transfer learning, transfer sample.

I. Introduction

IN THE field of sensors and measurements, sometimes the training and test samples are collected under different conditions. For example, suppose a company has produced a batch of devices of the same model to classify two kinds of signals. Generally, one would collect training data with one device, train prediction models using pattern recognition algorithms, and wish the models applicable to all the other devices. However, because of the variations in the fabrication of sensors

and devices, the responses to the same signal source may not be identical for different devices. In addition, the sensing characteristics of the sensors, the operating condition, or even the signal source itself, can change over time. These factors lead to a drift of posterior distribution of the measured variables [1], which will degrade the prediction accuracy.

A typical application plagued by this problem is machine olfaction [2], which senses gas using electronic noses (e-noses). An e-nose comprises an array of chemical sensors and a pattern recognition system [3]. It is capable of predicting the type or concentration of odors. E-noses have been adopted in areas such as food, environmental monitoring, and disease analysis [4]–[6]. Nevertheless, the responses of gas sensors are often influenced by the instrumental variation and time-varying drift as mentioned above [2], which have greatly affected the robustness of e-noses and hindered their popularization. Besides, instrumental variation also affects spectroscopy [9], [12].

Transfer learning can be used to address this challenging problem. Assuming the training data are from a source domain where labeled samples are sufficient, and the test data are from a target domain where labeled samples are scarce or not available, transfer learning aims to improve the prediction accuracy in the target domain by leveraging the knowledge from both domains [1]. Multitask learning (MTL), a type of inductive transfer learning method, has been successfully applied in several fields [7]–[10]. Besides, when dealing with the time-varying drift problem, there is a class of algorithms known as concept drift adaptation [11]. However, most MTL and concept drift adaptation algorithms rely on labeled samples in the target domain, which are sometimes hard to acquire in real-world applications. For instance, when a breath analysis system based on an e-nose [6] is produced in batch, it is impractical to collect patients' breath samples with each new device to update its model. In this situation, it is a good idea to use transfer samples to obtain knowledge from the target domain [2], [5], [12]–[14]. In the field of machine olfaction, transfer samples often consist of standard gases, which are reproducible and easy to acquire. A group of transfer samples can be measured in both source and target domains (e.g., old and new devices). Then, the mapping information between domains can be obtained by analyzing the correspondence relationship between transfer sample groups. Finally, the target data can be transformed to match those of the source [2], [5], [12]. This frequently used method is known as variable standardization.

While this method is easy to implement, its accuracy is not promising when the drift is complex. In this paper, we present

a novel method named transfer-sample-based MTL (TMTL) to predict data with drift. It combines MTL with transfer samples, thus has strength in both accuracy and practical convenience. In the proposed algorithm, labeled source data and a group of transfer samples are leveraged to learn the source and (multiple) target models jointly. The type of drift determines the relationship between domains, so we designed a parallel and a serial transfer paradigm for different drifts. To predict the sample measured in a specific time and handle noise in transfer samples, a dynamic model strategy using a combination of neighboring models is proposed.

Three real-world data sets with different types of drifts are used to evaluate the proposed algorithms. Experimental results show that TMTL achieves better prediction accuracy on data with drift compared with other typical algorithms in the fields of machine olfaction and MTL. The rest of this paper is organized as follows. Section II briefly reviews the related works in MTL and machine olfaction. Section III describes the proposed TMTL in detail. Section IV presents the experimental configurations and results, along with some analysis. Section V concludes this paper.

II. RELATED WORK

Following the terms in transfer learning [1], we refer to the data without drift as data from the source domain, and data with drift as data from the target domain. Generally, some samples from the target domain are needed to obtain knowledge about the domain. According to the type of the target samples, we classify drift correction algorithms into three categories, i.e., those based on labeled target samples, unlabeled target samples, and transfer samples.

In the setting of labeled-sample-based methods, some labeled data from the target domain are available, but not sufficient to retrain a target model. In this case, one intuitive idea is to use the source and the target data together to train a model, meantime increase the weights of the target samples to ensure the model's feasibility in the target domain. For instance, Zhang and Zhang [15] combined e-nose data before and after drift into the objective function of an extreme learning machine. Although easy to implement, this kind of method often needs many target samples to capture the variance in the target domain. In the case of time-varying drift, drifted data come in the form of streams. Concept drift adaptation methods make use of newly arrived labeled data to update the prediction models [11], [16]. As an example, Vergara *et al.* [17] adopted an ensemble strategy to cope with time-varying drift in e-noses. Samples collected in different time were split into several batches. Then, a prediction model was trained on each batch. Finally, for a test sample in batch k, the outputs of models 1 to $k-1$ were fused by weighted majority voting, with the weights estimated from the prediction accuracy of the model on batch $k-1$. The method requires all samples in prior batches to be labeled, which is often impractical.

MTL uses a different strategy to fuse knowledge from different domains. Models of all domains are learned jointly. In the objective function of an MTL method, the prior knowledge about the relationship of the models and the features can be specified. Consequently, information can be shared properly among the tasks, so as to enhance the generalization ability of all models, especially for the target domain which has less labeled samples. Regularized MTL (RMTL) was proposed in [7], in which a regularization term was introduced to penalize the deviation among multiple models. Binfeng *et al.* [9] applied RMTL to transfer models between near-infrared spectra measured in different conditions (e.g., multiple devices) and achieved good results. In [8], Zhou *et al.* formulated disease progress prediction as a multitask regression problem, with learning the model at each time period as a task. Models of neighboring time periods were required to be close to capture the intrinsic temporal smoothness. Group Lasso regularization was also employed for feature selection.

The second category of methods is unlabeled-sample-based ones, whose main advantage is that unlabeled target samples are much easier to acquire in practice. Transductive transfer learning [1] and semisupervised learning algorithms can be adopted in this setting. A transfer learning approach based on weighted geodesic flow kernel and a semisupervised classifier based on manifold regularization were used in [18] to address sensor drift in e-noses. Liu *et al.* [30] tried to apply deep learning methods to extract robust features from unlabeled gas samples and tackle sensor drift implicitly. The maximum independence domain adaptation algorithm proposed in [19] can handle different types of drift flexibly. Generally, unlabeled-sample-based methods are less accurate because of the limited information contained in the unlabeled samples.

Transfer samples are widely used in machine olfaction [2] and spectroscopy [12]. They are commonly a group of standard gases with selected types and concentrations. After being measured in both source and target domains, they can be used to estimate the mapping between domains. They are more informative than unlabeled target samples, meanwhile more convenient to obtain than labeled target samples in many real-world applications. Most existing transfer-sample-based methods concentrate on feature-level correction. Algorithms based on variable standardization build regression models using the transfer samples. Each variable in the source domain is fit with one or multiple variables in the target domain using regression algorithms such as robust fitting and ridge regression [5], [13], so as to transform the target data to the source domain. Then, the corrected data can be predicted by the source models. Algorithms based on component correction (CC) are also popular. CC-PCA [20] finds the drift-related direction in the feature space by applying principal component analysis to the transfer samples. Then, the component on the direction can be removed from all data. Orthogonal signal correction (OSC) [21] is a CC-like method that relies on labeled target samples. It pools samples with and without drift and finds the undesired components by calculating the subspace that is orthogonal to the labels. One drawback of CC-like methods is that when the drift is complex, it may be difficult to accurately separate the directions of useful information and drift [22]. Recently, Yan and Zhang [14] proposed a novel deep learning method to learn drift-corrected features. However, it is computationally intensive and performs better with enough training samples.

III. TRANSFER-SAMPLE-BASED MULTITASK LEARNING

In this section, we will first consider the situation with only one source and one target domain. Transfer-sample-based coupled task learning (TCTL), the basic form of TMTL, will be introduced for this situation. Note that TCTL was described in our previous work [23], but we have improved its model similarity term in this paper. Then, we will extend TCTL to TMTL which involves multiple domains, and describe a parallel paradigm and a serial one to deal with different interdomain relationships. Finally, we will propose a combination of the two paradigms and a dynamic model strategy.

A. Transfer-Sample-Based Coupled Task Learning

In order to depict the problem setup more concretely, we take calibration transfer as an example. The calibration transfer is the term used in machine olfaction and spectroscopy for transferring the model of one device to another. Suppose an e-nose (the source device) has been utilized to collect some training samples. A classification model was trained on these data. Now, we have made a new e-nose (the target one) of the same model. A set of standard gas samples has been measured with both the old and the new e-noses. Then, TCTL can be used to learn the classification model of the new device.

Denote $X_S \in \mathbf{R}^{n \times p}$ as the matrix of source training data with each row as a feature vector; n is the number of labeled source samples; p is the number of variables; $y_S \in \mathbf{R}^n$ is the label vector; $T_S \in \mathbf{R}^{n_t \times p}$ and $T_T \in \mathbf{R}^{n_t \times p}$ are matrices of the source and the target transfer samples, respectively; n_t is the number of transfer samples; $\boldsymbol{\beta}_S, \boldsymbol{\beta}_T \in \mathbf{R}^p$ are the parameter vectors of source and target prediction models to be estimated, respectively. The objective function of TCTL is presented as the following:

$$\min_{\boldsymbol{\beta}_S, \boldsymbol{\beta}_T} \ell(X_S, y_S, \boldsymbol{\beta}_S) + \lambda_1 \|T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T\|_2^2$$
$$+ \lambda_2 \|X_S \boldsymbol{\beta}_S - X_S \boldsymbol{\beta}_T\|_2^2 + \mu \sum_{j=1}^{p} w_j^2 (\beta_{S,j}^2 + \beta_{T,j}^2). \quad (1)$$

In (1), the first term represents the empirical loss function for the source training samples. $\|T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T\|_2^2$ is the transfer sample term. It requires the corresponding source and target transfer samples to be close after they are, respectively, projected by the source and target parameter vectors. The term $\|X_S \boldsymbol{\beta}_S - X_S \boldsymbol{\beta}_T\|_2^2$ encourages similar source and target parameter vectors by requiring that they project the source training samples to similar values. The last term is a weighted shrinkage term. $\beta_{S,j}$ stands for the jth element of $\boldsymbol{\beta}_S$. The weights are defined as

$$w_j = \sqrt{\sum_{i=1}^{n_t} (t_{S,ij} - t_{T,ij})^2}, \quad (2)$$

where $t_{S,ij}$ means the element in the ith row (sample) and jth column (variable) of T_S. The shrinkage term penalizes the variables that have large deviation between the source and target transfer samples. λ_1, λ_2 and $\mu \geq 0$ are regularization parameters controlling the strength of the terms.

The transfer sample term is key for information transfer between domains. It aligns the transfer samples of the two domains in their respective projected spaces, so as to reduce the interdomain drift. Thus, the discriminative information of the labeled source samples can be used in the target domain. However, if we rely solely on the transfer sample term to infer $\boldsymbol{\beta}_T$ from $\boldsymbol{\beta}_S$, the control over $\boldsymbol{\beta}_T$ will be too weak. Because the number of transfer samples is often small, there will be infinite solutions to $\boldsymbol{\beta}_T$ that can minimize the transfer sample term and make it zero. Therefore, we add the model similarity term $\|X_S \boldsymbol{\beta}_S - X_S \boldsymbol{\beta}_T\|_2^2$ to introduce an inductive bias reflecting the prior belief that the models resemble each other. To reduce the interdomain difference before applying TCTL, one can preprocess the source and target data separately with standard normal variate (SNV) [2], i.e., each variable is centered and scaled by the mean and standard deviation calculated from the transfer samples of its domain. Additionally, many MTL algorithms [7]–[9] simply penalize the deviation between two parameter vectors, e.g., minimizing $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|_2^2$. This requirement is too strict when the interdomain difference is large. Our model similarity term relaxes this requirement. The two parameter vectors may not be identical, but their difference should be orthogonal to the space spanned by the source training samples. Experimental results show that the model similarity term in this form is better than that in the traditional form.

In [13], we proposed a strategy to improve the transfer ability of prediction models of e-noses, namely, standardization error-based model improvement (SEMI). The motivation is that some variables inherently contain more drift that cannot be eliminated by simple standardization methods such as SNV. Hence, it will be beneficial to make the model depend less on such variables. The amount of drift of a variable can be measured by the SE, namely, the L_2 norm of the difference between the variable in the transfer samples of two domains [see (2)]. The mean of SEs can be scaled to 1. SEMI shrinks the variables with large SEs, making the trained models less sensitive to interdomain drift.

1) Classification: Logistic Loss: The proposed framework can be coupled with various loss functions. Logistic loss function is demonstrated in this paper because logistic regression (LR) is a popular and effective classifier. We denote $\boldsymbol{x}^{(i)} \in \mathbf{R}^p$ as the ith training sample and $y^{(i)} \in \{0, 1\}$ as its label $X = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}]^T$, $\boldsymbol{y} = [y^{(1)}, \ldots, y^{(n)}]^T$. In binary-class cases, the decision function of LR is a sigmoid function $h_{\boldsymbol{\beta}}(\boldsymbol{x}) = 1/(1 + e^{-\boldsymbol{\beta}^T \boldsymbol{x}})$. A test sample $\boldsymbol{x}$ is classified into the positive class, if $h_{\boldsymbol{\beta}}(\boldsymbol{x}) \geq 0.5$. The logistic loss function can be written as

$$\ell_L(X, \boldsymbol{y}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \log h_{\boldsymbol{\beta}}(\boldsymbol{x}^{(i)})$$
$$+ (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\beta}}(\boldsymbol{x}^{(i)})). \quad (3)$$

Combining (3) with (1), we formulate the objective function of TCTL with logistic loss as

$$J_L(\boldsymbol{\beta}_S, \boldsymbol{\beta}_T) = \ell_L(X_S, y_S, \boldsymbol{\beta}_S) + \frac{\lambda_1}{2n_t} \|T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T\|_2^2 + \frac{\lambda_2}{2n}$$
$$\times \|X_S \boldsymbol{\beta}_S - X_S \boldsymbol{\beta}_T\|_2^2 + \frac{\mu}{2} \sum_{j=1}^{p} w_j^2 (\beta_{S,j}^2 + \beta_{T,j}^2)$$
$$(4)$$

whose gradient is given by

$$
\frac{\partial J_L}{\partial \boldsymbol{\beta}_S} = \frac{1}{n} X_S^T (h_{\boldsymbol{\beta}}(X_S) - \boldsymbol{y}_S) + \frac{\lambda_1}{n_t} T_S^T (T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T)
$$
$$
+ \frac{\lambda_2}{n} X_S^T X_S (\boldsymbol{\beta}_S - \boldsymbol{\beta}_T) + \mu W \boldsymbol{\beta}_S
$$
$$
\frac{\partial J_L}{\partial \boldsymbol{\beta}_T} = -\frac{\lambda_1}{n_t} T_T^T (T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T) - \frac{\lambda_2}{n} X_S^T X_S (\boldsymbol{\beta}_S - \boldsymbol{\beta}_T)
$$
$$
+ \mu W \boldsymbol{\beta}_T
$$
$$
W = \mathrm{diag}(w_1^2, \ldots, w_p^2). \tag{5}
$$

The problem above can be solved using numerical optimization methods such as conjugate gradient. In K-class cases, K LR models are trained using the one-versus-all strategy and $\boldsymbol{x}$ is classified into the class whose decision function has the largest value.

2) Regression: Squared Loss: For regression problems, the squared loss function is adopted in this paper. The objective function of TCTL with squared loss is

$$
J_S(\boldsymbol{\beta}_S, \boldsymbol{\beta}_T) = \frac{1}{2n} \|X_S \boldsymbol{\beta}_S - \boldsymbol{y}_S\|_2^2 + \frac{\lambda_1}{2n_t} \|T_S \boldsymbol{\beta}_S - T_T \boldsymbol{\beta}_T\|_2^2
$$
$$
+ \frac{\lambda_2}{2n} \|X_S \boldsymbol{\beta}_S - X_S \boldsymbol{\beta}_T\|_2^2
$$
$$
+ \frac{\mu}{2} \sum_{j=1}^p w_j^2 (\beta_{S,j}^2 + \beta_{T,j}^2). \tag{6}
$$

By setting its gradient to zero, the closed-form solution to $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_T$ can be derived as follows:

$$
\begin{pmatrix} \boldsymbol{\beta}_S \\ \boldsymbol{\beta}_T \end{pmatrix} = (A_1 + A_2 + A_3)^{-1} \boldsymbol{b}, \tag{7}
$$

where

$$
A_1 = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \frac{\lambda_1}{n_t} \begin{pmatrix} T_S^T T_S & -T_S^T T_T \\ -T_T^T T_S & T_T^T T_T \end{pmatrix}
$$
$$
A_3 = \begin{pmatrix} \lambda_2 P + \mu W & -\lambda_2 P \\ -\lambda_2 P & \lambda_2 P + \mu W \end{pmatrix}, \quad \boldsymbol{b} = \frac{1}{n} \begin{pmatrix} X_S^T \boldsymbol{y}_S \\ 0 \end{pmatrix}
$$
$$
P = \frac{1}{n} X_S^T X_S, \quad W = \mathrm{diag}(w_1^2, \ldots, w_p^2).
$$

B. TMTL-Parallel and TMTL-Serial

TCTL only exploits information from two domains. In reality, there are situations of multiple domains. If a number of new devices have been manufactured, each new device can be regarded as a target domain which is different but related to each other and the old device (source domain). In another situation, a device may have been used to collect data for a long time. Knowing its slow and irregular time-varying drift, we have collected transfer samples periodically. In this case, each period can be viewed as a target domain which has relatively small intradomain drift. Each domain is different but related with its previous domain, i.e., the time period prior to it. TMTL shares information across many domains, which would probably be superior to TCTL. For instance, transfer samples in one domain may contain noises or outliers due to the uncertainty in the measurement process. In TCTL, the noises and outliers will mislead the model transfer process.

However, in TMTL, the influence of noises and outliers to one model can be mitigated owing to the similarity requirements with all the other models.

Considering the relationship between domains, we have designed two paradigms, namely, TMTL-parallel and TMTL-serial. TMTL-parallel is suitable for situations such as calibration transfer, where multiple domains are similar to each other. Here, we use a subscript k to denote the variable in the kth target domain, and a subscript 0 to denote the variable in the source domain for simplicity. The total number of target domains is d. The objective function of TMTL-parallel is expressed as

$$
\min_{\boldsymbol{\beta}_S, \boldsymbol{\beta}_T^{(1)}, \ldots, \boldsymbol{\beta}_T^{(d)}} \ell(X_0, \boldsymbol{y}_0, \boldsymbol{\beta}_0) + \lambda_1 \sum_{k=1}^d \|T_0 \boldsymbol{\beta}_0 - T_k \boldsymbol{\beta}_k\|_2^2
$$
$$
+ \lambda_2 \sum_{k=0}^d \left\| X_0 \left(\boldsymbol{\beta}_k - \frac{1}{d+1} \sum_{r=0}^d \boldsymbol{\beta}_r \right) \right\|_2^2
$$
$$
+ \mu \sum_{k=0}^d \sum_{j=1}^p w_{k,j}^2 \beta_{k,j}^2. \tag{8}
$$

It is a natural extension of TCTL to multiple target domains. The transfer samples of each target domain are aligned to those in the source domain in their respective projected spaces. Each parameter vector is encouraged to resemble an average parameter vector [7]. In the SEMI term, the shrinkage weight for variable j in the source domain ($w_{0,j}$) is the average of those in the target domain. By minimizing (8), we can obtain the prediction parameter vectors for all devices efficiently.

TMTL-serial is specialized for situations such as time-varying drift. The difference between the parallel and serial TMTL is that the latter one encourages each parameter vector to be similar to its previous parameter vector

$$
\min_{\boldsymbol{\beta}_S, \boldsymbol{\beta}_T^{(1)}, \ldots, \boldsymbol{\beta}_T^{(d)}} \ell(X_0, \boldsymbol{y}_0, \boldsymbol{\beta}_0) + \lambda_1 \sum_{k=1}^d \|T_0 \boldsymbol{\beta}_0 - T_k \boldsymbol{\beta}_k\|_2^2
$$
$$
+ \lambda_2 \sum_{k=1}^d \|X_0 (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1})\|_2^2
$$
$$
+ \mu \sum_{k=0}^d \sum_{j=1}^p w_{k,j}^2 \beta_{k,j}^2. \tag{9}
$$

The intuition is to capture the temporal smoothness prior as in [8]. Note that there are two typical modes to analyze data streams. In the offline mode, data in all time periods are analyzed together, which implies that transfer samples collected in later periods can aid the model transfer process of former periods. In this mode, models of all periods can be obtained simultaneously by optimizing (9). The online mode, on the other hand, requires data in the current period to be analyzed in real time. This means that only the transfer samples collected before can be used. In this mode, whenever a new group of transfer samples are collected, we can re-optimize (9) to get the latest parameter vector ($\boldsymbol{\beta}_d$), then use it to predict the recent samples.
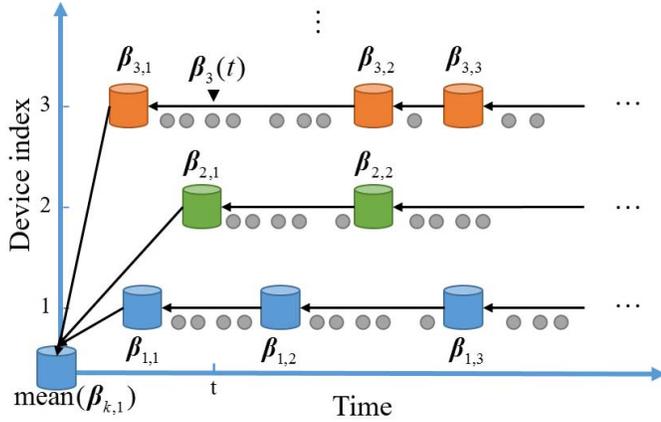
Fig. 1.   Illustration of the sample collection process in the most general case. The jth cylinder located in the ith row represents the jth group of transfer samples measured by device i, which also corresponds to a parameter vector $\boldsymbol{\beta}_{i,j}$. The circles are ordinary samples measured by the device at a specific time. The arrows indicate the model similarity relationships: the model at the beginning of an arrow should resemble the model at the end.

C. TMTL-General and the Dynamic Model Strategy

In the most general case, samples can be collected by several devices in a long period of time, as illustrated in Fig. 1. So we can go one step further and combine the parallel and serial TMTL to simultaneously learn all models. In TMTL, each group of transfer samples corresponds to a model. First, the initial group of transfer samples measured by the oldest device (denoted as device 1) is selected as the overall reference. All the other groups should be aligned with it in their respective projected spaces, which form the transfer sample term in the objective function. Second, each model is expected to be similar to its previous model of the same device, while the first model of each device should resemble their average [mean$(\boldsymbol{\beta}_{k,1})$], as shown in Fig. 1.

To deal with time-varying drift, the data stream of a device is split into discrete batches in the most previous studies [8], [17] and the discussions above. Each batch corresponds to one fixed model. This strategy loses the information carried in the exact acquisition time of the samples in the same batch. The drift within a batch cannot be modeled. Therefore, we propose a dynamic model strategy to exploit the information. Assuming that the time-varying drift of a device is smooth, it is intuitive to also let the model change smoothly over time. We set the model of device i at time t to be a function of all models of the same device. A straightforward method is to interpolate between neighboring models. We find it better to use a weighted combination as follows:

$$\boldsymbol{\beta}_i(t) = \sum_j c_{i,j}(t)\boldsymbol{\beta}_{i,j} \Big/ \sum_j c_{i,j}(t)$$
$$c_{i,j}(t) = \exp(-\sigma(t - t_{i,j})^2). \qquad (10)$$

where $t_{i,j}$ is the acquisition time of the jth group of transfer samples of device i. The closer t is to $t_{i,j}$, the larger the weight $c_{i,j}(t)$ will be. σ is the window size parameter. The $\sum_j c_{i,j}(t)$ in the denominator is a normalization term. When using this dynamic model strategy, the model for every training

and test sample should be calculated using (10). The mean and standard deviation values used to normalize variables in SNV should also be modified according to (10), which we will call dynamic SNV. These dynamic strategies can make the models more accurate. Another important function of the strategies is to deal with noises and outliers in transfer samples. They can smooth the noise contained in individual models, which has a similar insight to the ensemble strategy [17]. Details about TMTL with logistic or squared loss can be extended from (5) to (7), thus will not be presented here for brevity.

Compared with TCTL [23], TMTL does not limit the number of models, so that knowledge can be transferred between multiple domains. Different knowledge transfer paradigms are defined according to the nature of drift, so that prior knowledge can be incorporated to improve the accuracy. With the dynamic model strategy, samples do not have to be split into discrete batches and noise in neighboring models can be smoothed.

IV. Experiments

In this section, we will conduct four experiments on three data sets to evaluate the performance of the proposed algorithms. The three data sets contain time-varying drift, instrumental variation, and both, respectively. Comparison will be established between our methods and other typical methods in the fields of machine olfaction and MTL. Different strategies in our methods will also be explored and analyzed.

A. Gas Sensor Array Drift Data Set

The gas sensor array drift data set is a public dataset[1] introduced by Vergara *et al.* [17], [24]. An e-nose with 16 gas sensors was utilized to collect the data set over a course of 36 months. Six kinds of gases (ammonia, acetaldehyde, acetone, ethylene, ethanol, and toluene) at different concentrations were measured. The total number of samples is 13 910. Each sample is represented by a feature vector of 128 variables extracted from the sensors' response curves [17]. The data set is split into 10 batches in chronological order. The period of collection and the number of samples in each batch can be found in Table I. In this section, the goal is to classify the type of gases, despite their concentrations. We choose batch 1 (source domain) as the training set and test on batches 2–10 (target domains). This evaluation strategy was also used in [15], [17], and [18] and resembles the situation in real-world applications.

Fig. 2 shows a scatter map for visual inspection of the time-varying drift across batches. The samples are projected to a 2-D subspace using principal component analysis (PCA). It can be found that the ammonia samples drift roughly to the $+x$-direction, whereas the drift of acetaldehyde is small. There are also some samples that do not follow the general trend of drift, which implies that the drifting pattern of the samples is complex and it is hard to directly compensate it [15].

To explain the principle of proposed transfer-sample-based strategy, we depict the effect of TCTL in Fig. 3. An experiment

TABLE I

PERIOD OF COLLECTION AND NUMBER OF SAMPLES OF THE GAS SENSOR ARRAY DRIFT DATA SET [17]

Batch ID	Month	Ammonia (2.5-1000)[a]	Acetaldehyde (2.5-300)	Acetone (10-1000)	Ethylene (2.5-300)	Ethanol (2.5-600)	Toluene (1-230)	# Total
1	1, 2	83	30	70	98	90	74	445
2	3, 4, 8-10	100	109	532	334	164	5	1244
3	11-13	216	240	275	490	365	0	1586
4	14,15	12	30	12	43	64	0	161
5	16	20	46	63	40	28	0	197
6	17-20	110	29	606	574	514	467	2300
7	21	360	744	630	662	649	568	3613
8	22, 23	40	33	143	30	30	18	294
9	24, 30	100	75	78	55	61	101	470
10	36	600	600	600	600	600	600	3600

[a] Numbers in the parenthesis show the range of concentration in parts per million (ppm).
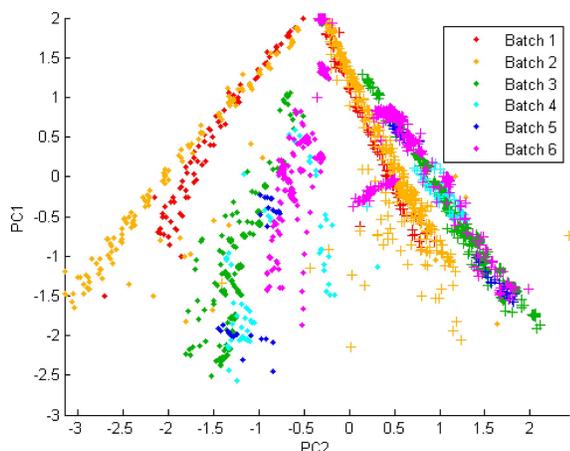
Fig. 2. Example of the drift across batches 1–6 in the gas sensor array drift data set. Dots and plus signs represent ammonia and acetaldehyde samples, respectively. Different colors indicate different batches.
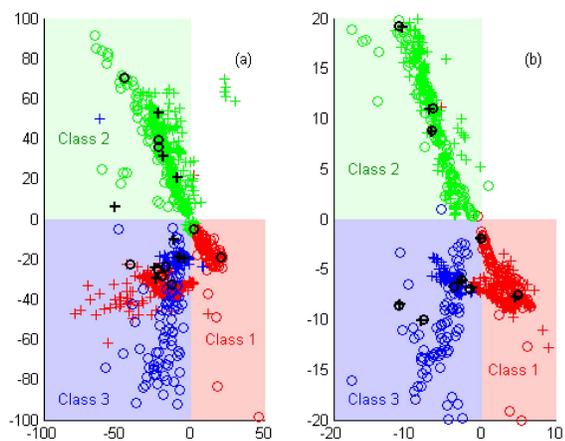
Fig. 3. Illustration of the effect of TCTL. Markers in different colors are samples from different classes, except the black ones, which represent the transfer samples. Circles are samples from batch 1 (source); Plus signs are those from batch 2 (target). (a) Samples from both batches are projected by the source parameter vectors learned by LR. (b) Source samples are projected by the source parameter vectors learned by TCTL with logistic loss, whereas the target samples are projected by the target parameter vectors learned by it.

was made with samples of three classes in two batches. Suppose a LR model is trained to distinguish class 1 from other classes. The *x*-axis of Fig. 3 is the coordinate of the samples

projected by the model's parameter vector. Another LR model is trained to distinguish class 2 from other classes, which corresponds to the *y*-axis. Therefore, if correctly classified, all samples from classes 1, 2, and 3 should be in the red, green, and blue regions, respectively. In Fig. 3(a), the LR models were trained on the source domain and applied on both source and target domains. We can find that some samples from the target domain (plus signs) are not in the correct region. This is because the source model cannot adapt well on the target domain. In Fig. 3(b), models for the source and target domains were simultaneously learned using TCTL. The transfer samples (black circles and plus signs) have been aligned, which is the principle of TCTL. In this situation, most samples in both domains lie in the correct regions.

The first step of our methods is choosing transfer samples. They are not directly provided in the data set, hence need to be selected from a candidate set. The candidate set of batch k ($k = 2, \ldots, 10$) was defined as the overlapping samples in batch 1 and k, namely, the samples of the same gas and concentration. Then, we used locally linear reconstruction (LLR) [25], an unsupervised active learning method, to choose n_t transfer samples for each candidate set. In our experiments, LLR achieved better accuracy than the Kennard–Stone selection algorithm [26] often used in machine olfaction and spectroscopy literatures [5], [12], [15]. After that, the samples in each batch were preprocessed by SNV. The models for batches 2–10 were learned using TCTL or TMTL. For TMTL, the serial paradigm and the online analysis mode were adopted. For a target batch k, the labeled training samples in batch 1 and the transfer sample groups of batches 1 to k were fed into (9) with logistic loss. After prediction, an average classification accuracy was computed.

Fig. 4 shows the average accuracy of TMTL-serial when parameters λ_1 and λ_2 are varied in $\{2^{-8}, 2^{-7}, \ldots, 2^2\}$. μ was fixed to 10^{-3}. The number of transfer samples in each batch was 10. We notice that the accuracy is the highest when λ_1 is neither too small nor too large. λ_1 controls the weight of the transfer sample term. If it is too small, the transfer samples cannot be aligned well. Meanwhile, putting too much emphasis on the transfer samples will cause overfitting. The accuracy degrades when λ_2 is large, indicating that the source and target models cannot be too similar because of the drift.

TABLE II

CLASSIFICATION ACCURACY OF VARIOUS METHODS ON THE GAS SENSOR ARRAY DRIFT DATA SET

Target batch ID	2	3	4	5	6	7	8	9	10	Average
No transfer	88.59	66.96	40.99	54.82	43.22	44.40	31.63	45.74	39.11	50.61
CC-PCA [20]	90.92	40.86	47.20	59.39	56.74	56.71	36.39	45.32	37.72	52.36
OSC [21]	88.10	66.71	54.66	53.81	65.13	63.71	36.05	40.21	40.08	56.50
Ensemble [17]	74.36	87.83	93.79	95.43	69.17	69.72	91.84	76.38	65.50	80.45
SMIDA [19]	83.68	82.28	73.91	75.63	93.00	63.49	79.25	62.34	45.50	73.23
DAELM-S [15]	87.98	95.74	85.16	95.99	94.14	83.51	86.90	100.00	53.62	87.00
TCTL	97.35	95.46	90.68	98.48	93.22	93.91	89.12	87.02	69.97	90.58
TMTL (sim2)	97.51	98.74	93.79	96.95	95.04	90.51	90.14	92.55	69.72	91.66
TMTL (no SEMI)	96.46	97.35	95.65	97.97	95.04	84.83	82.31	93.19	70.78	90.40
TMTL-parallel	97.35	97.16	93.17	97.46	95.96	91.70	90.14	90.85	73.28	91.90
TMTL-serial	97.35	98.80	90.06	98.48	95.35	91.50	91.84	96.38	71.56	**92.37**
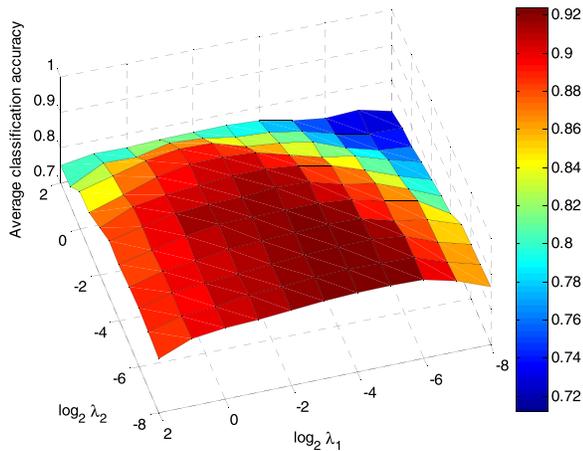
Fig. 4. Impact of the parameters λ_1 and λ_2 on the average classification accuracy of TMTL-serial.
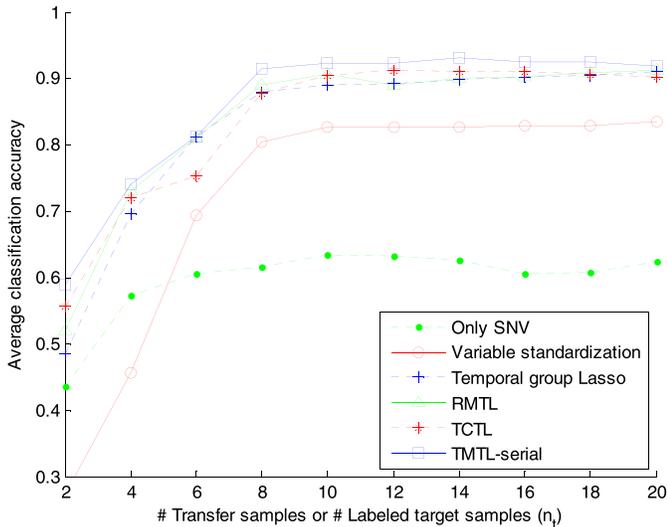
Fig. 5. Performance comparison on the gas sensor array drift data set with varying n_t.

Fig. 5 compares TCTL and TMTL-serial with several other methods, including only preprocessing the features with SNV (Only SNV) [2], variable standardization [23], MTL based on temporal group Lasso [8], [27], and RMTL [7]. LLR was used to select n_t transfer samples from the source domain or labeled target samples from the target domain. The first two methods standardize each target variable based on the transfer samples and then use the source models learned by LR to predict the standardized target samples. Their performances are not promising possibly because the drift is complex and the capacity of the feature-level correction methods is limited. The latter two are MTL methods (with logistic loss function and linear kernel) based on labeled target samples. The parameters were tuned by grid search for each result. Their performances are comparable with TCTL. TMTL-serial has the best accuracy for each n_t. Moreover, TCTL and TMTL have the advantage of not having to select and label the target samples.

More results of existing methods are listed in Table II. For "no transfer," data in batches 2–10 were directly predicted by the classification model trained on batch 1. Its accuracy is poor especially for batches with large IDs, which proves the influence of drift. The results of ensemble, domain adaptation extreme learning machine source (DAELM-S), and semi-supervised maximum independence domain adaptation (SMIDA) are copied from the original papers. Although the ensemble method and DAELM-S achieve good results, they both need relatively large amount of auxiliary target samples. DAELM-S requires 30 selected labeled samples in each target batch. The ensemble method requires all samples in batches 1 to $k-1$ to be labeled when predicting batch k. SMIDA needs only unlabeled target samples, but its accuracy is still not satisfactory.

In order to assess the strategies adopted in our methods, we have tested some possible alternatives, whose results are listed in the last five rows of Table II. For "TMTL (sim2)," the proposed model similarity constraint ($\|X_S\boldsymbol{\beta}_1 - X_S\boldsymbol{\beta}_2\|_2^2$) is replaced by $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2$, which occurs in many MTL papers. For "TMTL (no SEMI)," the proposed weighted shrinkage term is replaced by an ordinary shrinkage term with uniform weights. TMTL-serial outperforms the two alternatives, indicating the superiority of the proposed strategies. Besides, TMTL-serial is slightly better than TMTL-parallel in this problem. We also ran pairwise t-tests at the 5% significance level with results shown in the supplementary materials. The improvement of TMTL-serial is significant compared with others' methods, but not significant compared with other strategies of TMTL and TCTL.

TABLE III

AVERAGE RMSE ON THE CORN DATA SET WITH DIFFERENT NUMBER OF AUXILIARY SAMPLES

	Mp5 as target device				Mp6 as target device			
# Auxiliary samples	7	10	15	20	7	10	15	20
No transfer		1.242				1.347		
Only SNV	0.220	0.216	0.227	0.224	0.231	0.224	0.237	0.231
Variable standardization [13]	0.219	0.215	0.220	0.214	0.231	0.225	0.230	0.224
DAELM-S [15]	0.213	0.217	0.200	0.206	0.222	0.227	0.207	0.216
RMTL (squared loss)	0.201	0.197	0.213	0.206	0.204	0.193	0.207	0.204
RMTL (SVR) [9]	-				0.210	0.202	**0.181**	**0.177**
TCTL	0.196	0.189	0.196	**0.186**	0.194	0.184	0.190	0.181
TMTL-parallel	**0.186**	**0.183**	**0.190**	0.189	**0.188**	**0.182**	0.191	0.191
Trained on target device		0.185				0.189		
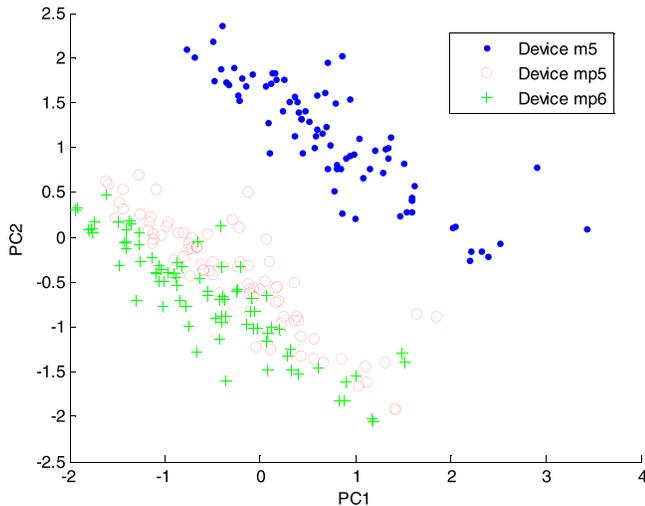
Fig. 6.　Scatter plot of the samples measured by the three spectrometers. Samples are projected to a 2-D subspace using PCA.

B. Corn Data Set

The corn data set is a publicly available data set in spectroscopy.[2] Three near-infrared spectrometers designated as m5, mp5, and mp6 were involved. Each device was adopted to measure the moisture, oil, protein, and starch contents of 80 corn samples. The ranges of the measured values are 9.377 to 10.993, 3.088 to 3.832, 7.654 to 9.711, and 62.826 to 66.472, respectively. The wavelength range is 1100–2498 nm at 2 nm intervals, resulting in 700 variables for each sample. Fig. 6 illustrates the variation in distribution of the same samples measured by the three devices.

We follow the experimental setting in [9] and study the calibration transfer from m5 to the other two devices. A four-fold cross validation was made by assigning every 4th sample to the test set. In each fold, the transfer samples or labeled target samples were selected by LLR from the training samples. Before training, each spectrum was first down-sampled to form a feature vector with 234 variables, followed by preprocessing with SNV. The four measured values were predicted separately and an average root mean squared error (RMSE) was computed. Table III lists the results on the two target devices when different number of transfer samples/labeled target samples were used. The parameters were tuned by

[2]http://www. eigenvector .com/data/Corn/
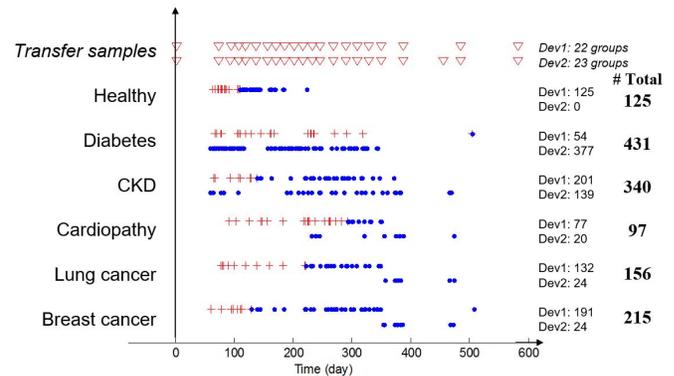
Fig. 7.　Overview of part of the breath analysis data set. Each point denotes a sample (or a group of transfer samples) collected in a specific time. The two rows of each class represent samples measured by the two devices, with the sample sizes labeled on the right. Red plus signs denote the training samples.

grid search for each result. The results of "RMTL (support vector regression (SVR))" are copied from [9], which only provided the results on mp6. TMTL-parallel was applied in the experiment. It achieves the best performance when the number of auxiliary samples (n_t) is small. Pairwise t-test results show that TMTL is significantly better than other methods except RMTL (SVR). The latter method has smaller RMSE on mp6 when the n_t is larger than 15, which is probably because the labeled-sample-based method can extract more information from extra additional labeled samples, whereas information brought by extra transfer samples is marginal when n_t is large (can also be observed from Fig. 5). RMTL (SVR) also benefits from an ε-insensitive loss function with radial basis function kernel. It will be our future work to equip our methods with more powerful loss functions and kernels. For "trained on target device," regression models were trained and tested on the same device. It can be regarded as an ideal result for calibration transfer. We find that with the help of only 10 transfer samples, TMTL can actually outperform it.

C. Breath Analysis Data Set

Breath analysis with e-noses is attracting increasing attention [6], [28], [29]. Researchers found that some diseases are related with biomarkers at abnormal concentrations in exhaled breath [29]. For example, the concentration of acetone in breath of diabetics is often higher than that of healthy subjects. With the progress of sensor technology, it is possible

TABLE IV

Sensitivity/Specificity on the Breath Analysis Data Set

	Task 1	2	3	4	5	Average
No transfer	79.23 / 56.98	62.03 / 76.74	68.75 / 77.91	65.77 / 56.98	75.93 / 58.14	70.34 / 65.35
MDC [20]	68.72 / 80.23	83.77 / 63.95	82.81 / 88.37	85.23 / 36.05	92.13 / 38.37	82.53 / 61.40
CC-PCA [20]	74.36 / 66.28	64.35 / 72.09	81.25 / 72.09	79.87 / 54.65	81.02 / 69.77	76.17 / 66.98
TMTL-parallel	60.26 / 55.81	66.38 / 65.12	84.38 / 66.28	72.48 / 44.19	73.61 / 66.28	71.42 / 59.53
TMTL-general	66.92 / 88.37	88.99 / 80.23	93.75 / 82.56	74.50 / 81.40	89.81 / 76.74	82.79 / 81.86
Random train	84.32 / 90.70	77.70 / 86.51	83.59 / 88.88	76.09 / 85.67	77.70 / 82.93	79.88 / 86.94
Random train + TMTL	90.30 / 94.86	85.07 / 89.42	87.69 / 87.84	77.28 / 89.44	79.42 / 85.09	83.95 / 89.33
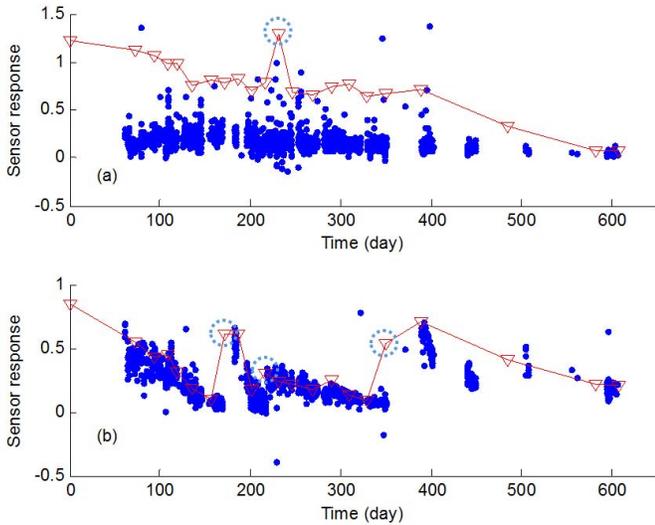
Fig. 8. Responses of two sensors in all breath samples (blue dots) and one transfer sample (red triangles). Each point represents the steady response of the sensor in one sample. Dashed circles mark the (a) outlier in transfer samples or (b) replacement of the sensor.

to diagnose and monitor diseases by measuring breath samples with e-noses. It has the advantage of being noninvasive, convenient, cheap, and fast. One big obstacle of this technique is the drift problem. Solutions based on labeled target samples are impractical in this case because of the difficulty in collecting breath samples from patients. Therefore, only transfer-sample-based methods will be tested in this section.

A breath analysis data set was collected using two e-noses of the same model [6]. The collection process lasted for about 500 days starting from 2014. From the data set, we select five diseases that have been proved to be related with certain biomarkers in breath, namely, diabetes, chronical kidney disease, cardiopathy, lung cancer, and breast cancer [29]. Their sample sizes and days of collection are illustrated in Fig. 7, together with those of the healthy samples and transfer samples. Transfer sample groups were measured periodically, with eight preselected standard gas samples in each group.

This real-world data set suffers from a number of factors that will cause drift in data distribution, e.g., instrumental variation, sensor aging, temperature and humidity change, sensor damage and replacement, and so on. As an example, we draw the steady-state responses of two sensors in Fig. 8. The sensitivity of the sensor, shown in Fig. 8(a), gradually decayed over time, as can be observed from the trend of breath

and transfer samples. For the sensor in Fig. 8(b), however, the decay was much faster, so we replaced it three times. It is worth noting that the transfer samples contain noise and outliers [Fig. 8(a)], which cannot precisely reflect the true distribution of the data, thus will degrade the accuracy if we transfer knowledge based on them. One solution is to detect the outliers according to some prior knowledge. In this paper, we use the dynamic model strategy in (10) to deal with it.

The experimental settings are as follows. Five binary-class classification tasks (healthy versus disease) were executed. Sensitivity and specificity [6] were adopted as the accuracy metrics in this medical application. To simulate real-world applications, we used only the first 50 samples collected with device 1 in each class as training samples (see Fig. 7), others as test ones. Considering the complexity of the drift and the noise in transfer samples, we utilized the offline analysis mode, namely, all groups of transfer samples were used to learn all models simultaneously. The 9-D feature vector consists of steady-state responses of nine gas sensors, followed by dynamic SNV described in Section III-C.

Experimental results are listed in Table IV. The parameters of each method were tuned by grid search. For methods except TMTL and "random train + TMTL," LR was adopted as the classifier. Multiplicative drift correction (MDC) is a simplified version of variable standardization which corrects each variable with a multiplicative factor. It performed better than variable standardization in this data set. However, the two transfer-sample-based feature-level correction methods, MDC and CC-PCA, showed a little improvement over "no transfer." The accuracy of TMTL-parallel is not good because time-varying drift is severe in this data set. In TMTL-general, the dynamic model strategy was applied since the exact acquisition time of each sample is known. Forty five models were learned simultaneously, as there were 45 groups of transfer samples altogether. The time-specific model for each training or test sample is a combination of neighboring models. The window size parameter in (10) was empirically set to 10^{-4}. We find that this strategy is important for the data set. If it is not used and each sample is predicted by an individual adjacent model, the accuracy will be poor. The noise in transfer samples could be the major cause. The combined model can smooth the noise. A minor drawback is that it cannot deal with "abrupt drift," e.g., sensor replacement. The accuracy of TMTL is close to "random train," in which the 50 training samples of each class were randomly selected from all devices and time periods to include the information of drift

in the model. If we use TMTL with randomly select training samples, the accuracy can be further improved, indicating that TMTL can reduce the influence of drift effectively with the information contained in the transfer samples. Pairwise *t*-test results show that TMTL-general is significantly better than half of the other methods. However, the insignificant results are partial because there are only 5 tasks to compare in *t*-test.

V. CONCLUSION

We propose TMTL to address the drift problem of sensors and devices, with a special focus on machine olfaction. By drift, we refer to the change of posterior data distribution caused by instrumental variation, sensor aging, and environmental change. Instead of correcting the drifted signals as in conventional methods, our method handles drift under the framework of transfer learning and MTL. The key idea is to reduce the influence of drift in the target domains by aligning the transfer samples at the model level.

Different from existing MTL methods depending on labeled samples, TMTL leverages transfer samples to transfer knowledge from the source domain to the target ones. In our experiments, it achieved better results, and the number of transfer samples needed for effective transfer was usually small (about 10). Besides, transfer samples are not required to be of the same type with the training and test samples. Thus, the proposed method is more convenient to use in many real-world applications. The second feature of TMTL is that it can simultaneously handle the "discrete" instrumental variation and the "continuous" time-varying drift. By using the parallel and serial transfer paradigm and the dynamic model strategy, the models in different domains are linked to reflect the prior knowledge about the drift.

Overall, TMTL is a practical algorithm framework to predict data with complex drift caused by various factors. The robustness of sensor systems (e.g., e-noses) can be greatly enhanced. Future works may include improving the objective functions by making more sophisticated assumptions on the structures of the models and features.

REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[2] S. Marco and A. Gutierrez-Gálvez, "Signal and data processing for machine olfaction and chemical sensing: A review," *IEEE Sensors J.*, vol. 12, no. 11, pp. 3189–3214, Nov. 2012.

[3] J. W. Gardner and P. N. Bartlett, "A brief history of electronic noses," *Sens. Actuators B, Chem.*, vol. 18, nos. 1–3, pp. 210–211, Mar. 1994.

[4] K. Brudzewski, S. Osowski, and A. Dwulit, "Recognition of coffee using differential electronic nose," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 6, pp. 1803–1810, Jun. 2012.

[5] L. Zhang *et al.*, "On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality," *Sens. Actuators B, Chem.*, vol. 160, no. 1, pp. 899–909, 2011.

[6] K. Yan, D. Zhang, D. Wu, H. Wei, and G. Lu, "Design of a breath analysis system for diabetes screening and blood glucose level prediction," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 11, pp. 2787–2795, Nov. 2014.

[7] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. ACM SIGKDD*, Seattle, WA, USA, 2004, pp. 109–117.

[8] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. ACM SIGKDD*, San Diego, CA, USA, 2011, pp. 814–822.

[9] Y. Binfeng and J. Haibo, "Near-infrared calibration transfer via support vector machine and transfer learning," *Anal. Methods*, vol. 7, no. 6, pp. 2714–2725, Mar. 2015.

[10] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.

[11] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, Apr. 2014.

[12] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferré, "Transfer of multivariate calibration models: A review," *Chemometrics Intell. Lab. Syst.*, vol. 64, no. 2, pp. 181–192, Nov. 2002.

[13] K. Yan and D. Zhang, "Improving the transfer ability of prediction models for electronic noses," *Sens. Actuators B, Chem.*, vol. 220, pp. 115–124, Dec. 2015.

[14] K. Yan and D. Zhang, "Correcting instrumental variation and time-varying drift: A transfer learning approach with autoencoders," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2012–2022, Sep. 2016.

[15] L. Zhang and D. Zhang, "Domain adaptation extreme learning machines for drift compensation in E-nose Systems," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 7, pp. 1790–1801, Jul. 2015.

[16] P. Kadlec, R. Grbić, and B. Gabrys, "Review of adaptation mechanisms for data-driven soft sensors," *Comput. Chem. Eng.*, vol. 35, no. 1, pp. 1–24, Jan. 2011.

[17] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," *Sens. Actuators B, Chem.*, vols. 166–167, pp. 320–329, May 2012.

[18] Q. Liu, X. Li, M. Ye, S. S. Ge, and X. Du, "Drift compensation for electronic nose by semi-supervised domain adaption," *IEEE Sensors J.*, vol. 14, no. 3, pp. 657–665, Mar. 2014.

[19] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybern.*, to be published.

[20] T. Artursson, T. Eklöv, I. Lundström, P. Mårtensson, M. Sjöström, and M. Holmberg, "Drift correction for gas sensors using multivariate methods," *J. Chemometrics*, vol. 14, nos. 5–6, pp. 711–723, 2000.

[21] S. Wold, H. Antti, F. Lindgren, and J. Öhman, "Orthogonal signal correction of near-infrared spectra," *Chemometrics Intell. Lab.*, vol. 44, nos. 1–2, pp. 175–185, Dec. 1998.

[22] A. C. Romain and J. Nicolas, "Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview," *Sens. Actuators B, Chem.*, vol. 146, no. 2, pp. 502–506, Apr. 2010.

[23] K. Yan and D. Zhang, "Calibration transfer and drift compensation of e-noses via coupled task learning," *Sens. Actuators B, Chem.*, vol. 225, pp. 288–297, Mar. 2016.

[24] I. Rodriguez-Lujan, J. Fonollosa, A. Vergara, M. Homer, and R. Huerta, "On the calibration of sensor arrays for pattern recognition using the minimal number of experiments," *Chemometrics Intell. Lab.*, vol. 130, pp. 123–134, Jan. 2014.

[25] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang, "Active learning based on locally linear reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2026–2038, Oct. 2011.

[26] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969.

[27] J. Zhou, J. Chen, and J. Ye. (2011). *MALSAR: Multi-Task Learning Via Structural Regularization.* http://www.public.asu.edu/~jye02/Software/MALSAR

[28] F. Röck, N. Barsan, and U. Weimar, "Electronic nose: Current status and future trends," *Chem. Rev.*, vol. 108, no. 2, pp. 705–725, Jan. 2008.

[29] A. D. Wilson and M. Baietto, "Advances in electronic-nose technologies developed for biomedical applications," *Sensors*, vol. 11, no. 1, pp. 1105–1176, Jan. 2011.

[30] Q. Liu, X. Hu, M. Ye, X. Cheng, and F. Li, "Gas recognition under sensor drift by using deep learning," *Int. J. Intell. Syst.*, vol. 30, no. 8, pp. 907–922, 2015.

Ke Yan received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China.

He is currently a Post-Doctoral Researcher in medical image analysis with the National Institutes of Health, Bethesda, MD, USA. His current research interests include computer vision, machine learning, and their biomedical applications.

Dr. Yan was a recipient of the 2016 Tsinghua University Excellent Doctoral Dissertation Award.

David Zhang (F'09) received the B.Sc. degree in computer science from Peking University, Beijing, China, the M.Sc. degree in computer science and the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University, Beijing. He was an Associate Professor with the Academia Sinica, Beijing. He is currently the Chair Professor with The Hong Kong Polytechnic University, Hong Kong, 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 a Visiting Chair Professor with Tsinghua University and an Adjunct Professor with Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He has authored more than 10 books and 200 journal papers.

Prof. Zhang is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of IAPR. He is the Founder and an Editor-in-Chief of the *International Journal of Image and Graphics*, a Book Editor of *Springer International Series on Biometrics*, an Organizer for the *International Conference on Biometrics Authentication*, and an Associate Editor of more than 10 international journals including the IEEE TRANSACTIONS AND PATTERN RECOGNITION.

Yong Xu (M'06–SM'15) was born in Sichuan, China, in 1972. He received the B.S. and M.S. degrees from the Air Force Institute of Meteorology, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2005.

He is currently with the Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China. His current research interests include pattern recognition, biometrics, machine learning, video analysis, and bioinformatics.