Learning Classification and Regression Models for Data with Drift based on Transfer Samples

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

***Abstract*—When devices 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, environmental change, etc. 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. TMTL reduces the influence of drift in the target domains by aligning the transfer samples at the model level. Two paradigms, parallel and serial transfer, are designed to reflect different relationships between domains, which are further dependent on the cause of drift. A dynamic model strategy is proposed to predict samples with known acquisition time and to handle noise in transfer samples. Classification and regression experiments on three real-world datasets confirm the efficacy of the proposed methods. They achieve good accuracies 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*—Transfer sample, drift, transfer learning, multitask learning, machine olfaction, calibration transfer**

I. Introduction

In many real-world pattern recognition problems, the training and test data may have different distributions [1]. In the field of sensors and measurement, for example, suppose a company has produced a batch of devices of the same model which can 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 cause a drift of posterior distribution of the measured variables, which will degrade the accuracy of the initially trained models.

A typical application plagued by this problem is machine olfaction [2], which senses gas using electronic noses in an analogue to the human nose. An electronic nose (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 ranging from agriculture to environmental monitoring, robotics, and disease analysis [4-8]. Nevertheless, the responses of gas sensors are always influenced by the factors mentioned above [2], which has greatly affected the robustness of e-noses and hindered their popularization.

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) is a type of inductive transfer learning method which has been successfully applied in several fields [9-13]. It treats model learning in different domains as different but related tasks. Multiple models are learned simultaneously, so that information can be shared across them during the learning process to improve their accuracies [14]. Besides, when dealing with the time-varying drift problem, i.e. the relationship between data and label changes over time, there is a class of algorithms known as concept drift adaptation [15]. They update the prediction model based on newly arrived labeled data to react to concept drifts.

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 example, when a breath analysis system based on an e-nose [5] 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, 16, 17]. In the field of

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, 61272292, 61271344), Shenzhen Fundamental Research fund (JCYJ20140508160910917, JCYJ20150403161923528), and Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China.

K. Yan is with the Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: yank10@mails.tsinghua.edu.cn).

*D. Zhang is with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Biometrics Research Centre, Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (email: yongxu@ymail.com).

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, 16-18]. This commonly used method is known as variable standardization.

While this method is easy to implement, it is not accurate when the drift is complex. In this paper, we present a novel method named transfer-sample-based multitask learning (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 exploited 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 the noise in transfer samples, a dynamic model strategy that uses a combination of neighboring models is proposed. TMTL is a framework that can be implemented using various loss functions. Two popular classification/regression loss functions, i.e. logistic and squared loss, are demonstrated in this paper. We also compared three algorithms to select representative transfer samples.

Two real-world classification problems and two regression ones with different types of drift 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. We mainly focus on machine olfaction in this paper, but the proposed methods have potential in fields such as spectroscopy [12], indoor localization [11], and color correction of digital cameras [19]. For these problems, data measured by sensors or devices contain drift, meanwhile transfer samples can be collected (e.g. the spectra of the same chemical, the WiFi signals in the same location, and the images of the same color).

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. The transfer sample selection algorithms adopted in this paper are introduced in Section IV. Section V presents the experimental configurations and results, along with some comprehensive analysis. Section VI concludes the paper.

II. Related Works

The main interest of this paper is to improve the prediction accuracy on data with drift, with a focus on machine olfaction systems. Typical techniques related to this problem will be discussed in this section, including those in the field of pattern recognition and machine olfaction. 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 methods 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 is 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 et al. [20] 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 comes in the form of streams. Concept drift adaptation methods make use of newly arrived labeled data to update the prediction models [15, 21, 22]. As an example, Vergara et al. [23] 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 accuracies of the models on batch k-1. The method requires all samples in prior batches to be labeled, which is often impractical.

Multitask learning (MTL) uses a different strategy to fuse knowledge from different domains. Models for 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 [9], in which a regularization term was introduced to penalize the deviation among multiple models. Yu et al. [12] applied RMTL to transfer models between near-infrared spectra measured in different conditions (e.g. multiple devices) and achieved good results. In [10], 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 are unlabeled-sample-based ones, whose main advantage is that unlabeled target samples are much easier to acquire in practice. Transductive transfer learning [1] and semi-supervised learning algorithms can be adopted in this setting. A transfer learning approach based on weighted geodesic flow kernel and a semi-supervised classifier based on manifold regularization were used in [24] to address sensor drift in e-noses. On the dataset introduced in [23], the prediction accuracy on drifted data was improved.

Transfer samples are widely used in machine olfaction [2] and spectroscopy [16]. 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 fitted with one or multiple variables in the target domain using regression algorithms such as robust fitting and ridge regression [7, 18], so as to transform the target data to the source domain. Then, the corrected data can be predicted by the source models. This strategy has also been utilized for color correction of digital cameras to reduce instrumental variation [19]. Algorithms based on component correction (CC) are also popular. CC-PCA [17] 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) [25] is a CC-like method that relies on labeled target samples. It pools samples with and without drift and finds the undesired component 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 [26].

III. Transfer-sample-based Multitask Learning (TMTL)

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, is introduced for this situation. Then, we will extend TCTL to TMTL which involves multiple domains, and describe a parallel paradigm and a serial one to deal with different inter-domain relationships. Finally, we propose a combination of the two paradigms and a dynamic model strategy.

A. Transfer-sample-based Coupled Task Learning (TCTL)

A preliminary version of TCTL was introduced in our previous work [27]. In order to depict the problem setup more concretely, we take calibration transfer as an example. 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) was utilized to collect 50 breath samples from healthy subjects and 50 from diabetes patients. A classification model was trained on these data. Now we have made a new e-nose (the target device) of the same model for diabetes screening. A set of standard gas samples have been be measured with both the old and the new e-nose. Then, TCTL can be used to learn the classification model of the new device.

Denote $X_S \in \mathbb{R}^{n\times p}$ 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; $\boldsymbol{y}_S \in \mathbb{R}^n$ is the label vector; $T_S \in \mathbb{R}^{n_t\times p}$ and $T_T \in \mathbb{R}^{n_t\times p}$ are the matrices of the source and target transfer samples, respectively; n_t is the number of transfer samples; $\boldsymbol{\beta}_S, \boldsymbol{\beta}_T \in \mathbb{R}^p$ are the source and target prediction models to be estimated, respectively. The objective function of TCTL is presented as the following:

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

In (1), the first term represents the empirical loss function for the source training samples. $\left\| T_S\boldsymbol{\beta}_S - T_T\boldsymbol{\beta}_T \right\|_2^2$ 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 models. The term $\left\| X_S\boldsymbol{\beta}_S - X_S\boldsymbol{\beta}_T \right\|_2^2$ encourages similar source and target models 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 inter-domain drift. Thus, the discriminative information of the labeled source samples can be used in the target domain. The transfer sample term can also be regarded as an improvement on the conventional variable standardization (VS) method. In linear cases, the latter method is essentially estimating a matrix $M \in \mathbb{R}^{p\times p}$ to transform the target variables to the source space, i.e. to make $T_T M \approx T_S$. In the prediction step, the transformed target samples are projected by $\boldsymbol{\beta}_S$. So the goal actually boils down to reducing the difference between the two domains in the projected direction, in other words, minimizing $\left\| T_S\boldsymbol{\beta}_S - T_T M\,\boldsymbol{\beta}_S \right\|_2^2$. It is exactly the transfer sample term in TCTL if we set $\boldsymbol{\beta}_T = M\,\boldsymbol{\beta}_S$. M no longer needs to be estimated, which makes TCTL more efficient and less prone to overfitting compared with VS.

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 inter-domain 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 [9-13] simply penalize the deviation between two models, e.g. minimizing $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_T\|_2^2$. This requirement is too strict when the inter-domain difference is large. Our model similarity term relaxes this requirement. The two models 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 [18], 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 sensors inherently produce more drift, which cannot be eliminated by simple standardization methods like 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 standardization error (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 inter-domain drift.

1) Classification: Logistic Loss

The proposed framework can be tied with various loss functions. Logistic loss function is demonstrated in this paper because logistic regression (LR) is a popular and effective classifier. We denote $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$ as the *i*th training sample and $y^{(i)} \in \{0,1\}$ as its label. $X = [\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)}]^\top$, $\boldsymbol{y} = [y^{(1)},\ldots,y^{(n)}]^\top$. In binary-class cases, the decision function of LR is a sigmoid function $h_{\boldsymbol{\beta}}(\boldsymbol{x}) = 1/(1+e^{-\boldsymbol{\beta}^\top \boldsymbol{x}})$. 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:

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

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

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

whose gradient is given by:

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

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-vs-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:

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

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

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

where

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

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 intra-domain 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:

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

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 model is encouraged to resemble an average model [9]. 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 models 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 model to be similar to its previous model:

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

The intuition is to capture the temporal smoothness prior as in [10]. 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, we can optimize (9) each time a new group of transfer samples are collected, and use the latest model ($\boldsymbol{\beta}_d$) obtained to predict recent samples.

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 forms 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 $\text{mean}(\boldsymbol{\beta}_{k,1})$, as shown in Fig. 1.

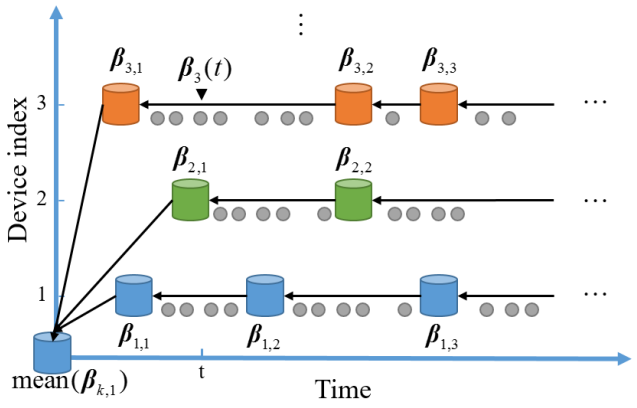

Fig. 1. Illustration of the sample collection process in the most general case. The jth cylinder located in the ith row represents the jth group of transfer samples measured by device i, which also corresponds to a model $\boldsymbol{\beta}_{i,j}$. The circles are ordinary samples measured by the device 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.

To deal with time-varying drift, the data stream of a device is split into discrete batches in most previous studies [10, 20, 23, 24] 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:

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

$t_{i,j}$ is the acquisition time of the *j*th 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. Note that $c_{i,j}(t)$ should be normalized to keep a sum of 1, i.e. $c_{i,j}(t) = c_{i,j}(t) / \sum_j c_{i,j}(t)$. 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 probably make the models more accurate. Another important function of the strategies is to deal with noises and outliers in transfer samples. They can further smooth the noise contained in individual models, which has similar insight to the ensemble strategy [23]. Details about TMTL with logistic or squared loss can be extended from (5) and (7), thus will not be presented here for brevity.

IV. Selection of Transfer Samples

The selection of transfer samples is also an important issue in TMTL. Transfer samples should be capable of representing one domain in order to effectively transfer knowledge between domains. Meanwhile, the number of transfer samples should be as small as possible to ease the burden of collecting them repeatedly [28]. Because only the source samples are accessible in the training stage, one often gathers a sufficient set of candidates from the source domain, then select a compact and representative group from them. The selected transfer samples can then be measured by every new device and in each time period.

In the field of machine olfaction and spectroscopy, the Kennard-Stone (KS) algorithm is the most intensively used [7, 16, 20, 29]. Given a set of candidate transfer samples, KS aims to sequentially select the samples that capture the most variance of the candidates. First, the two samples farthest apart from each other are picked. The next sample selected should have the largest nearest distance from the existing selections. This procedure is repeated until desired number of samples have been chosen. One disadvantage of KS is that the selected samples may contain outliers.

Active learning methods [30, 31] are suitable for this problem. In this paper, we explore two methods in this category. For easy controlling of the sample size, only sequential selection algorithms are considered. Transductive experimental design (TED) [30] selects samples that can be used to reconstruct the whole data set most precisely [31]. Locally Linear Reconstruction (LLR) [31] further takes into account the local manifold structure. It requires that a data point can only be linearly reconstructed from its neighbors, then selects the samples that best reconstruct the whole data set. We will compare the three methods mentioned above in the next section. Another related issue is to choose the proper time to collect transfer samples. For this issue, one can refer to the change detection algorithms in [15].

V. Experiments

In this section, we will conduct four experiments on three datasets to evaluate the performance of the proposed algorithms. The three datasets 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. Classification: Gas Sensor Array Drift Dataset

The gas sensor array drift dataset is a public dataset (http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations) introduced by Vergara et al. [23, 32]. An e-nose with 16 gas sensors was utilized to collect the dataset 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 13910. Each sample is represented by a feature vector of 128 variables extracted from the sensors' response curves [23]. The dataset 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 [20, 23, 24] 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 2D subspace using 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 [20].

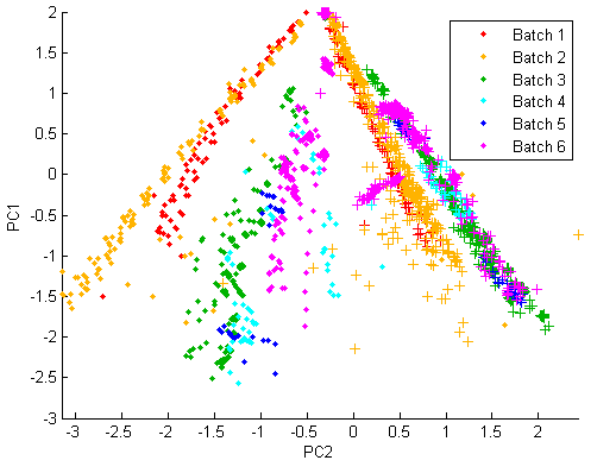

Fig. 2. Example of the drift across batches 1-6 in the gas sensor array drift dataset. Dots and plus signs represent ammonia and acetaldehyde samples, respectively. Different colors indicate different batches.

TABLE I
PERIOD OF COLLECTION AND NUMBER OF SAMPLES IN THE GAS SENSOR ARRAY DRIFT DATASET [23]

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).

To explain the principle of proposed transfer-sample-based strategy, we depict the effect of TCTL in Fig. 3, in which transfer samples are leveraged to align the drifted samples in the projected subspace. An experiment was made with samples of three classes in two batches. Two classification models were trained to distinguish class 1 or 2 from the other two classes. Then, the samples were projected by the two models. The colored areas suggest the correct regions for the samples in each class. In plot (a), samples from both batches are projected by the source model. Therefore, some target samples (plus signs) fall into the wrong region because of the drift, thus will not be correctly classified. In plot (b), TCTL is applied to learn the source and the target models simultaneously. With the transfer samples in both domains aligned (black points), the drift is reduced at the model level and the target samples fall into correct regions.

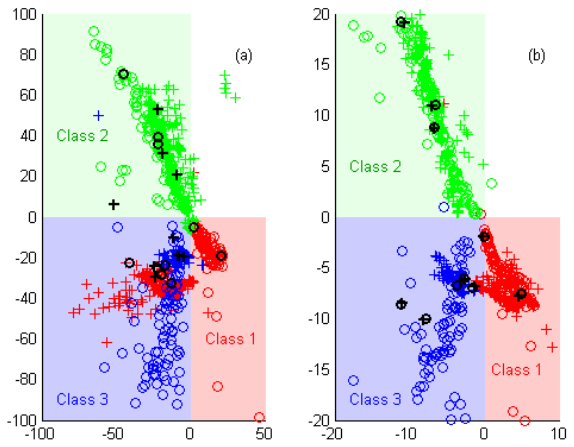

Fig. 3. Illustration of the effect of TCTL. Markers in different colors are samples from different classes, except the black ones, which represent the transfer samples. Circles are samples from batch 1 (source); Plus signs are those from batch 2 (target). In plot (a), samples from both batches are projected by the source model learned by LR. In plot (b), the source samples are projected by the source model learned by TCTL with logistic loss, whereas the target samples are projected by the target model learned by it.

The first step of our methods is choosing transfer samples. They are not directly provided in the dataset, hence need to be selected from a candidate set. The candidate set of batch k (k = 2, …, 10) was defined as the overlapping samples in batch 1 and k, namely the samples of the same gas and concentration. Then, we used the three selection algorithms introduced in Section IV to choose n_t transfer samples for each batch. 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.

Comparison of the sample selection algorithms is displayed in Fig. 4. The parameters for TED and LLR were set to be the same with those in the original papers [30, 31]. The parameters of TCTL and TMTL were optimized by grid search for each result, except that μ was empirically fixed to 10^{-3}. It can be found that the accuracy improves as n_t increases, until n_t reaches about 10. The overall order of performance is LLR > KS > TED. The effectiveness of the locally linear reconstruction strategy of LLR is proved. The traditional KS algorithm, although simple, shows performances close to LLR. Besides, TMTL is generally better than TCTL, which is because TMTL involves more tasks that can help each other. It makes use of all k groups of transfer samples for batch k, whereas TCTL only uses two groups. Furthermore, TMTL-serial is able to capture the temporal smoothness prior of the data. The influence of noises and outliers in transfer samples is better mitigated.

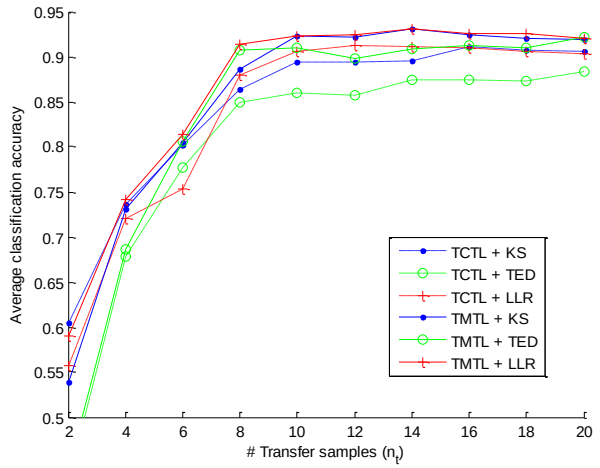

Fig. 4. Comparison of the three sample selection methods on the classification task of the gas sensor array drift dataset.

Fig. 5 shows the average accuracy of TMTL-serial when parameters λ_1 and λ_2 are varied in $\{2^{-8},2^{-7},\ldots,2^{2}\}$. LLR was used to select 10 transfer samples in this experiment. μ was still fixed to 10^{-3}. 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.

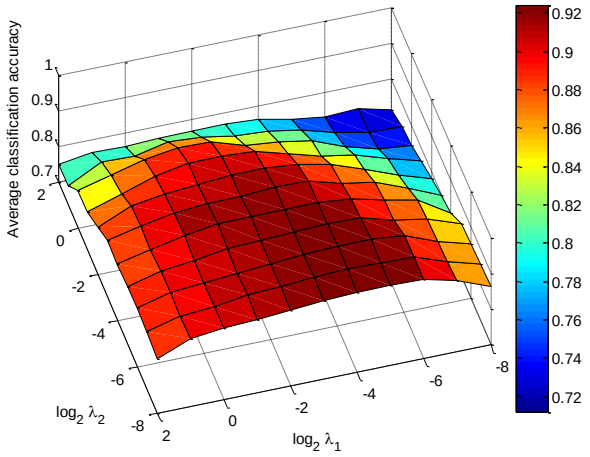

Fig. 5. The impact of the parameters λ_1 and λ_2 on the average classification accuracy of TMTL-serial.

Fig. 6 compares TCTL and TMTL-serial with several other methods, including only preprocessing the features with SNV (Only SNV) [2], variable standardization [27], MTL based on temporal group Lasso (TGL) [10, 33], and regularized MTL (RMTL) [9]. LLR was used to select 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, 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.

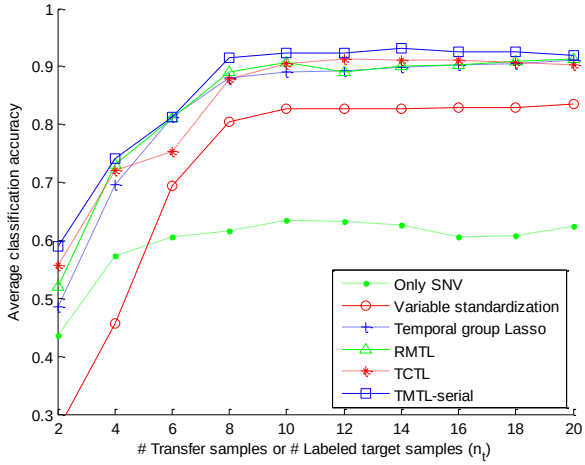

Fig. 6. Performance comparison on the classification task of the gas sensor array drift dataset.

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, DAELM-S, and ML-comGFK 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*. ML-comGFK 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.

TABLE II
CLASSIFICATION ACCURACY OF VARIOUS METHODS AND DIFFERENT STRATEGIES IN OUR METHODS

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 [17]	90.92	40.86	47.20	59.39	56.74	56.71	36.39	45.32	37.72	52.36
OSC [25]	88.10	66.71	54.66	53.81	65.13	63.71	36.05	40.21	40.08	56.50
Ensemble [23]	74.36	87.83	93.79	95.43	69.17	69.72	91.84	76.38	65.50	80.45
ML-comGFK [24]	80.25	74.99	78.79	67.41	77.82	71.68	49.96	50.79	53.79	67.28
DAELM-S [20]	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**

B. Regression: Gas Sensor Array Drift Dataset

In this section, we predict the concentration of the gases in the gas sensor array drift dataset. The range of concentration of each kind of gas can be found in Table I. A regression model was trained for each kind of gas. After prediction, the root mean squared error (RMSE) for each target batch was computed. In practice, we found that the RMSE of batch 2 is much higher than the other batches, which is possibly because there are many outliers in it, so we only report the results in batches 3-10 and their average. Fig. 7 compares the three transfer sample selection algorithms using TCTL and TMTL with the squared loss. TMTL-serial and the online analysis mode were adopted. For each result, the parameters λ_1 and λ_2 were searched in $\{2^{-8},2^{-7},\ldots,2^{2}\}$ and μ was empirically fixed to 10^{-1}. Other experimental configurations were similar to those in Section A. According to Fig. 7, the order of performance is still LLR > KS > TED and TMTL > TCTL.

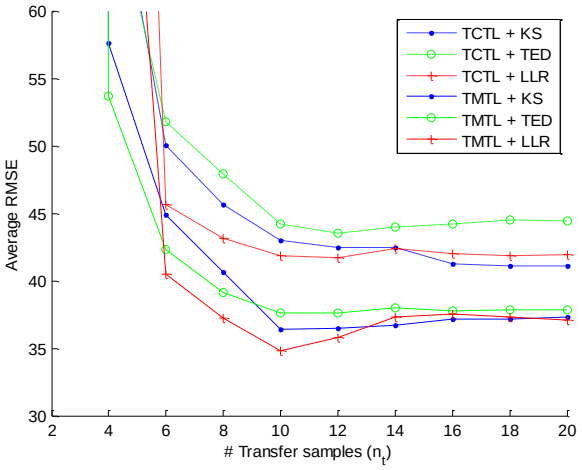

Fig. 7. Comparison of the three sample selection methods on the regression task of the gas sensor array drift dataset.

Several methods in the fields of machine olfaction and MTL were tested on the problem. Their results are demonstrated in Fig. 8. LLR was used to select transfer samples or labeled target samples. The parameters were tuned by grid search for each result. Ridge regression was used after feature normalization with SNV or variable standardization. The results of DAELM-S are average values over 10 repetitions considering the random factor in the algorithm. Squared loss and linear kernel were used in DAELM-S, TGL, and RMTL. Note that these three methods are based on labeled target samples, which need to be selected separately for each gas. On the contrary, TCTL and TMTL can use the same group of transfer samples for all gases. Therefore, according to Fig. 8, TCTL and TMTL outperform the three methods and largely save the effort to collect auxiliary target samples. We also find that the RMSE of TCTL and TMTL is poor when n_t is less than 4. It implies that if the transfer samples are too few, the source and target domains cannot be properly aligned, probably due to the noise in transfer samples and overfitting. More results are listed in Table III. It can be observed from the last five rows that the proposed TMTL-serial surpasses other alternative strategies in this problem.

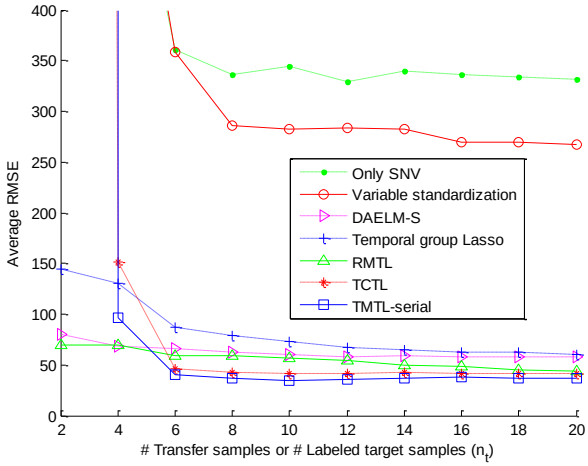

Fig. 8. Performance comparison on the regression task of the gas sensor array drift dataset.

TABLE III
REGRESSION RMSE OF VARIOUS METHODS AND DIFFERENT STRATEGIES IN OUR METHODS

Target batch ID	3	4	5	6	7	8	9	10	Average
No transfer	142.68	288.83	189.03	189.29	128.05	115.04	132.77	337.09	190.35
CC-PCA [17]	108.38	178.21	139.08	108.77	125.01	93.79	77.60	165.72	124.57
OSC [25]	235.14	168.72	158.36	121.01	106.25	90.51	106.24	272.06	157.29
TCTL	32.82	44.70	38.14	31.52	44.40	40.64	28.39	74.48	41.89
TMTL (sim2)	28.39	34.52	25.14	24.31	42.42	41.30	49.88	79.11	40.63
TMTL (no SEMI)	34.22	35.98	38.64	32.89	46.70	39.03	68.00	108.60	50.51
TMTL-parallel	28.44	35.24	25.12	23.54	41.97	40.68	49.60	76.70	40.16
TMTL-serial	22.43	29.83	20.31	26.59	38.15	36.95	33.70	70.30	**34.78**

C. *Regression: Corn Dataset*

The corn dataset is a publicly available dataset in spectroscopy (http://www.eigenvector.com/data/Corn/). 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. 9 illustrates the variation in distribution of the same samples measured by the three devices.

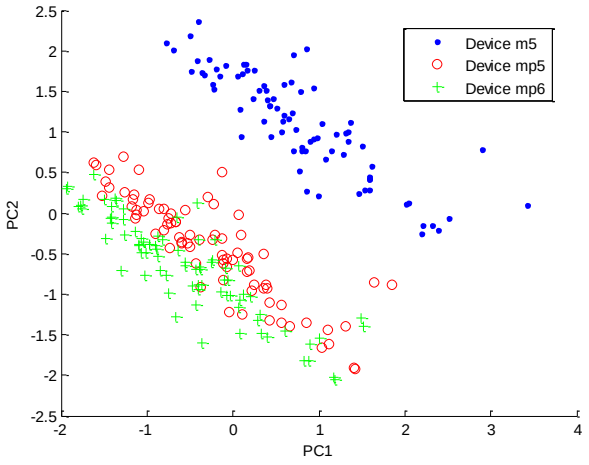

Fig. 9. Scatter plot of the samples measured by the three spectrometers. The samples are projected to a 2D subspace using PCA.

We follow the experimental setting in [12] and study the calibration transfer from m5 to the other two devices. A four-fold cross validation was made by assigning every fourth 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 spectra 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 RMSE was computed. Table IV lists the results on the two target devices when different number of transfer sample / labeled target samples were used. The parameters were tuned by grid search for each result. The results of "RMTL (SVR)" are copied from [12], 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. RMTL (SVR) 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 the additional labeled samples, whereas information brought by the additional transfer samples is marginal when n_t is large (can also be observed from Fig. 6 and 8). RMTL (SVR) also benefits from an ε-insensitive loss function with RBF 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 objective result for calibration transfer. We find that with the help of only 10 transfer samples, TMTL can actually outperform it.

TABLE IV
AVERAGE RMSE ON THE CORN DATASET 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 [18]	0.219	0.215	0.220	0.214	0.231	0.225	0.230	0.224
DAELM-S [20]	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) [12]	-				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			

D. *Classification: Breath Analysis Dataset*

Breath analysis with e-noses are attracting increasing attention [4-6]. Researchers found that some diseases are related with biomarkers at abnormal concentrations in exhaled breath [6]. 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 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 dataset was collected using two e-noses of the same model [5]. The collection process lasted for about 500 days starting from 2014. From the dataset, we select five diseases that have been proved to be related with certain biomarkers in breath, namely diabetes, chronical kidney disease (CKD), cardiopathy, lung cancer, and breast cancer [6]. Their sample sizes and days of collection are illustrated in Fig. 10, together with those of the healthy samples and transfer samples. Transfer sample groups were measured periodically, with eight pre-selected standard gas samples in each group.

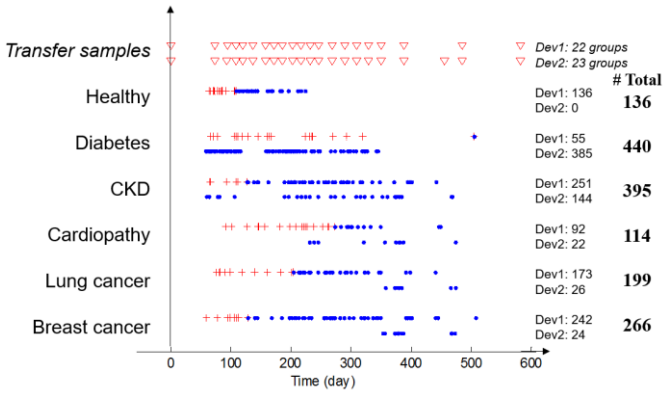

Fig. 10. Overview of part of the breath analysis dataset. Each point denotes a sample (or a group of transfer samples) collected in a specific time. The two rows of each class represent samples measured by the two devices, with the sample sizes labeled on the right. Red plus signs denote the training samples.

This real-world dataset 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, etc. As an example, we draw the steady-state responses of two sensors in Fig. 11. The sensitivity of the sensor in plot (a) gradually decayed over time, as can be observed from the trend of breath and transfer samples. For the sensor in plot (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 (e.g. in plot (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.

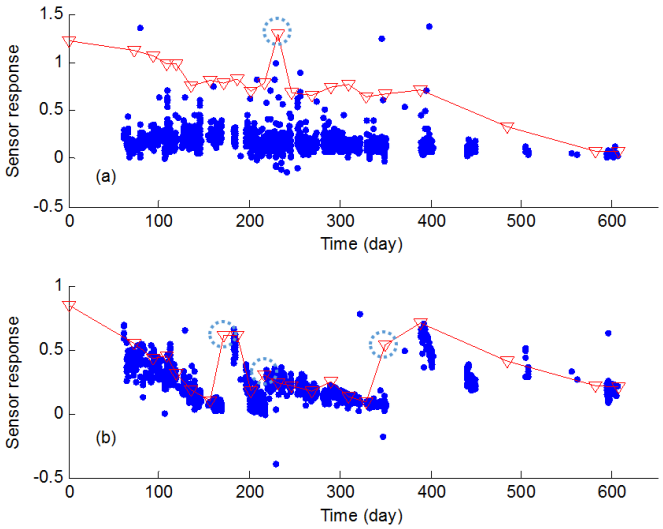

Fig. 11. Responses of two sensors in all breath samples (blue dots) and one transfer sample (red triangles). Each point represents the steady response of the sensor in one sample. Dashed circles mark the outlier in transfer samples (plot (a)) or the replacement of the sensor (plot (b)).

The experimental settings are as follows. Five binary-class classification tasks (healthy vs. disease) were executed. Because the classes are imbalanced, F-score was adopted as the accuracy metric. To simulate real-world applications, we used only the first 50 samples collected with device 1 in each class as training samples (see Fig. 10), 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 9D feature vector consists of steady-state responses of nine gas sensors, followed by dynamic SNV described in Section III.C.

TABLE V
F-SCORE OF THE CLASSIFICATION TASKS

	Task 1	2	3	4	5	Average
No transfer	66.29	68.61	73.04	61.06	65.85	66.97
MDC [17]	74.03	72.53	85.50	50.67	54.18	67.38
CC-PCA [17]	70.09	68.00	76.40	64.90	74.97	70.87
TMTL	76.17	84.38	87.80	77.79	82.77	**81.78**
Random train	87.45	81.87	86.30	80.35	80.16	83.22
Random train + TMTL	95.80	87.70	89.11	83.88	84.84	88.27

Experimental results are listed in Table V. 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 dataset. However, the two transfer-sample-based feature-level correction methods, MDC and CC-PCA, showed little improvement over "no transfer." For TMTL, TMTL-general with the dynamic model strategy was applied since the exact acquisition time of each sample is known. 45 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 dataset. 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.

VI. Conclusion

We propose transfer-sample-based multitask learning (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, environmental change, etc. Different from conventional methods that correct the drifted signals, our method handles drift under the framework of transfer learning and MTL. The key idea of our method is to reduce the influence of drift in the target domains by aligning the transfer samples at the model level. In this paper, we have three observations:

1) Different from existing MTL methods depending on labeled or unlabeled target samples, TMTL leverages transfer samples to transfer knowledge from source to target domains. 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.

2) TMTL learns models for multiple target domains jointly. It is always better than its basic version, TCTL, which only involves one target domain. This confirms that TMTL has organized the models in a proper way so that they can improve each other.

3) In the cases of time-varying drift, the serial transfer paradigm is better because it can capture the temporal smoothness prior. The dynamic model strategy is feasible when the acquisition time of each sample is known and when the transfer samples contain noise.

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 making more sophisticated assumptions on the structures of the models and features. It will also be beneficial to further exploit the information contained in unlabeled target samples.

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. Gutiérrez-Gálvez, "Signal and data processing for machine olfaction and chemical sensing: a review," *IEEE Sens. 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, no. 1, pp. 210-211, Mar. 1994.

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

[5] 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.

[6] 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.

[7] L. Zhang, F. Tian, C. Kadri, B. Xiao, H. Li, L. Pan, and H. Zhou, "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.

[8] A. Loutfi, S. Coradeschi, A. J. Lilienthal, and J. Gonzalez, "Gas distribution mapping of multiple odour sources using a mobile robot," *Robotica*, vol. 27, no. 02, pp. 311-319, Mar. 2009.

[9] T. Evgeniou, and M. Pontil, "Regularized multi-task learning," in Proc. ACM SIGKDD, Seattle, Washington, 2004, pp. 109-117.

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

[11] S. J. Pan, J. T. Kwok, Q. Yang, and J. J. Pan, "Adaptive Localization in a Dynamic WiFi Environment through Multi-view Learning," in Proceedings of the national conference on artificial Intelligence, 2007, pp. 1108.

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

[13] X. Gu, F.-L. Chung, H. Ishibuchi, and S. Wang, "Multitask Coupled Logistic Regression and Its Fast Implementation for Large Multitask Datasets," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1953-1966, Sep. 2014.

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

[15] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 44, Apr. 2014.

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

[17] T. Artursson, T. Eklöv, I. Lundström, P. Mårtensson, M. Sjöström, and M. Holmberg, "Drift correction for gas sensors using multivariate methods," *J. Chemometr.*, vol. 14, no. 5‐6, pp. 711-723, 2000.

[18] 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.

[19] X. Wang, and D. Zhang, "An optimized tongue image color correction scheme," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 6, pp. 1355-1364, Nov. 2010.

[20] L. Zhang, and D. Zhang, "Domain Adaptation Extreme Learning Machines for Drift Compensation in E-Nose Systems," *IEEE Trans. Instrum. Meas.*, Jul. 2015.

[21] 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.

[22] Y.-R. Yeh, and Y.-C. F. Wang, "A rank-one update method for least squares linear discriminant analysis with concept drift," *Pattern Recogn.*, vol. 46, no. 5, pp. 1267-1276, May 2013.

[23] 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.*, vol. 166, pp. 320-329, May 2012.

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

[25] S. Wold, H. Antti, F. Lindgren, and J. Öhman, "Orthogonal signal correction of near-infrared spectra," *Chemometr. Intell. Lab.*, vol. 44, no. 1, pp. 175-185, Dec. 1998.

[26] 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.
[27] K. Yan, and D. Zhang, “Calibration transfer and drift compensation of e-noses via coupled task learning,” accepted by *Sens. Actuators B: Chem.*, Paper SNB_19316.
[28] O. Y. Rodionova, and A. L. Pomerantsev, “Subset selection strategy,” *J. Chemometr.*, vol. 22, no. 11‐12, pp. 674-685, Jul. 2008.
[29] R. W. Kennard, and L. A. Stone, “Computer aided design of experiments,” *Technometrics*, vol. 11, no. 1, pp. 137-148, Feb. 1969.
[30] K. Yu, J. Bi, and V. Tresp, “Active learning via transductive experimental design.” pp. 1081-1088.
[31] 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.
[32] 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,” *Chemometr. Intell. Lab.*, vol. 130, pp. 123-134, Jan. 2014.
[33] J. Zhou, J. Chen, and J. Ye, “MALSAR: Multi-tAsk Learning via StructurAl Regularization,” http://www.public.asu.edu/~jye02/Software/MALSAR, 2011.