Reducing BCI Calibration Effort in RSVP Tasks Using Online Weighted Adaptation Regularization with Source Domain Selection

Dongrui Wu
DataNova
Clifton Park, NY 12065
Email: drwu09@gmail.com

Vernon J. Lawhern
U.S. Army Research Laboratory
University of Texas at San Antonio
Email: vernon.j.lawhern.civ@mail.mil

Brent J. Lance
Translational Neuroscience Branch
U.S. Army Research Laboratory
Email: brent.j.lance.civ@mail.mil

***Abstract*—Rapid serial visual presentation based brain-computer interface (BCI) system relies on single-trial classification of event-related potentials. Because of large individual differences, some labeled subject-specific data are needed to calibrate the classifier for each new subject. This paper proposes an online weighted adaptation regularization (OwAR) algorithm to reduce the online calibration effort, and hence to increase the utility of the BCI system. We show that given the same number of labeled subject-specific training samples, OwAR can significantly improve the online calibration performance. In other words, given a desired classification accuracy, OwAR can significantly reduce the number of labeled subject-specific training samples. Furthermore, we also show that the computational cost of OwAR can be reduced by more than 50% by source domain selection, without a statistically significant sacrifice of classification performance.**

***Keywords*—Brain-computer interface (BCI), EEG, event-related potentials (ERP), domain adaptation, transfer learning, single-trial classification, online calibration**

I. Introduction

Single-trial classification of event-related potentials (ERPs) is used in many real-world brain-computer interface (BCI) applications. For example, in a rapid serial visual presentation (RSVP) based BCI system, an analyst is shown a sequence of images in rapid succession (e.g. 2-10 Hz) [6], [19], and he/she needs to detect sparsely-appearing target images that appear in a series of non-target or distractor stimuli. Each image is classified based on the EEG response of the analyst. A target image usually evokes the P300 ERP, which can be detected by the BCI system [18] and triaged for further inspection. RSVP-based BCI systems enable image analysts to detect targets in large aerial photographs faster and more accurately than traditional standard searches [17], [23], [30].

However, because of large individual differences, it is very difficult to build a generic classifier whose parameters fit all subjects. Usually some labeled subject-specific data are needed to calibrate the classifier for each new subject. Reducing this calibration effort, i.e., minimizing the number of labeled subject-specific data required in calibration, would greatly increase the utility of the BCI system.

Generally there are two types of calibration in BCI:

1) *Offline calibration*, in which a pool of unlabeled EEG epochs have been obtained *a priori*, and a user is queried to label some of these epochs, which are then used to train a classifier to label the remaining epochs in the pool.
2) *Online calibration*, in which some labeled EEG epochs are obtained on-the-fly, and then a classifier is trained from them to classify future (unseen) EEG epochs.

The major different between offline calibration and online calibration is that, in the former, the unlabeled EEG epochs can be used to help design the classifier, whereas in online calibration there are no unlabeled EEG epochs. Additionally, in offline calibration we can query any epoch in the pool for the label, but in online calibration usually the sequence of the epochs is pre-determined and the subject has little control on which epochs he/she will see next.

Transfer learning (TL) [16], which makes use of auxiliary data to improve the learning performance of a new task, has started to find applications in both offline and online BCI calibration [1], [12], [13], [24], [28]. The auxiliary data could be legacy data from the same subject in the same task or similar tasks, or legacy data from different subjects in the same task or similar tasks.

A TL approach, weighted adaptation regularization with source domain selection (wARSDS), was proposed in [29] for *offline* single-trial classification of ERPs in a visually-evoked potential task. It showed that wARSDS significantly outperformed the baseline algorithm, which did not use TL, and a simple TL approach proposed in [28]. This paper extends wARSDS to *online* BCI calibration, and shows that online wARSDS (OwARSDS) can significantly reduce the online calibration effort, compared with a baseline approach that does not use TL, and the simple TL approach proposed in [28]. As individual differences are also a well-known and pervasive phenomenon in psychology and affective computing [4], [10], [11], [21], [25], we believe OwARSDS will also be applicable to these areas.

The rest of the paper is organized as follows: Section II introduces the details of the OwARSDS algorithm. Section III describes experimental results and performance comparisons of different algorithms. Section IV draws conclusions.

II. Online Weighted Adaptation Regularization with Source Domain Selection (OwARSDS)

This section introduces the OwARSDS algorithm, which extends the offline wARSDS algorithm [29] to online BCI calibration. Similar to wARSDS, OwARSDS also copes well with class-imbalance problem and multiple source domains. For simplicity, we only consider 2-class classification here.

OwARSDS consists of two parts: source domain selection (SDS) to select the closest source domains, and online weighted adaptation regularization (OwAR) for each selected source domain. We will introduce OwAR first, and then SDS, because SDS relies on the results of OwAR.

A. OwAR: Problem Definition

A domain [14], [16] $\mathcal{D}$ in TL consists of a d-dimensional feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{x})$, i.e., $\mathcal{D} =$

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

$\{\mathcal{X}, P(\mathbf{x})\}$, where $\mathbf{x} \in \mathcal{X}$. Two domains $\mathcal{D}_s$ and $\mathcal{D}_t$ are different means $\mathcal{X}_s \neq \mathcal{X}_t$, and/or $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$.

A task [14], [16] $\mathcal{T}$ in TL consists of a label space $\mathcal{Y}$ and a conditional probability distribution $Q(y|\mathbf{x})$. Two tasks $\mathcal{T}_s$ and $\mathcal{T}_t$ are different means $\mathcal{Y}_s \neq \mathcal{Y}_t$, or $Q_s(y|\mathbf{x}) \neq Q_t(y|\mathbf{x})$.

Given a source domain $\mathcal{D}_s$ with n labeled samples, $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, and a target domain $\mathcal{D}_t$ with m labeled samples, $\{(\mathbf{x}_{n+1}, y_{n+1}), ..., (\mathbf{x}_{n+m}, y_{n+m})\}$, domain adaptation TL aims to learn a target prediction function $f : \mathbf{x}_t \mapsto y_t$ with low expected error on $\mathcal{D}_t$, under the assumptions $\mathcal{X}_s = \mathcal{X}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$, and $Q_s(y|\mathbf{x}) \neq Q_t(y|\mathbf{x})$.

For example, in a RSVP task, EEG epochs from a new subject are in the target domain, while EEG epochs from an existing subject (usually different from the new subject) are in the source domain. There could be more than one source domain, but in OwAR we consider each source domain separately. A single data sample would consist of the feature vector for a single EEG epoch from a subject, collected as a response to a specific visual stimulus. Though the features in source and target domains are computed in the same way, generally their marginal and conditional probability distributions are different, i.e., $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$ and $Q_s(y|\mathbf{x}) \neq Q_t(y|\mathbf{x})$, because the two subjects usually have different neural responses to the same visual stimulus. As a result, the auxiliary data from a source domain cannot represent the primary data in the target domain accurately, and must be integrated with some labeled data in the target domain to induce the target predictive function.

B. OwAR: The Learning Framework

Because

$$f(\mathbf{x}) = Q(y|\mathbf{x}) = \frac{P(\mathbf{x}, y)}{P(\mathbf{x})} = \frac{Q(\mathbf{x}|y)P(y)}{P(\mathbf{x})}, \tag{1}$$

to use the source domain data in the target domain, we need to make sure[1] $P_s(\mathbf{x}_s)$ is close to $P_t(\mathbf{x}_t)$, and $Q_s(\mathbf{x}_s|y_s)$ is also close to $Q_t(\mathbf{x}_t|y_t)$.

Let the classifier be $f = \mathbf{w}^T\phi(\mathbf{x})$, where $\mathbf{w}$ is the classifier parameters, and $\phi : \mathcal{X} \mapsto \mathcal{H}$ is the feature mapping function that projects the original feature vector to a Hilbert space $\mathcal{H}$. The learning framework of OwAR is formulated as:

$$\begin{aligned} f = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} & \sum_{i=1}^{n} w_{s,i}\ell(f(\mathbf{x}_i), y_i) + w_t \sum_{i=n+1}^{n+m} w_{t,i}\ell(f(\mathbf{x}_i), y_i) \\ & + \sigma\|f\|_K^2 + \lambda[D_{f,K}(P_s, P_t) + D_{f,K}(Q_s, Q_t)] \end{aligned} \tag{2}$$

where n is the number of labeled samples in the source domain, m is the number of labeled samples in the target domain, ℓ is the loss function, $K \in R^{(n+m)\times(n+m)}$ is the kernel function induced by ϕ such that $K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$, and σ and λ are non-negative regularization parameters. w_t is the overall weight for target domain samples, which should be larger than 1 so that more emphasis is given to target domain samples than source domain samples. $w_{s,i}$ is the weight for the i^{th} sample in the source domain, and $w_{t,i}$ is the weight for the i^{th} sample in the target domain, i.e.,

$$w_{s,i} = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{D}_{s,1} \\ n_1/(n - n_1), & \mathbf{x}_i \in \mathcal{D}_{s,2} \end{cases} \tag{3}$$

$$w_{t,i} = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{D}_{t,1} \\ m_1/(m - m_1), & \mathbf{x}_i \in \mathcal{D}_{t,2} \end{cases} \tag{4}$$

in which $\mathcal{D}_{s,c} = \{\mathbf{x}_i|\mathbf{x}_i \in \mathcal{D}_s \wedge y_i = c\}$ is the set of samples in Class c of the source domain, $\mathcal{D}_{t,c} = \{\mathbf{x}_j|\mathbf{x}_j \in \mathcal{D}_t \wedge y_j = c\}$ is the set of samples in Class c of the target domain, $n_c = |\mathcal{D}_{s,c}|$ and $m_c = |\mathcal{D}_{t,c}|$. The goal of $w_{s,i}$ ($w_{t,i}$) is to balance the number of samples from difference classes in the source (target) domain.

Briefly speaking, the first two terms in (2) minimize the loss on fitting the labeled samples in the source domain and target domain, respectively. The 3rd term minimizes the structural risk of the classifier. The last term minimizes the distance between the marginal probability distributions in the two domains, and also the distance between the conditional probability distributions.

By the Representer Theorem [2], [14], the solution of (2) is:

$$f(\mathbf{x}) = \sum_{i=1}^{n+m} \alpha_i K(\mathbf{x}_i, \mathbf{x}) = \boldsymbol{\alpha}^T K(X, \mathbf{x}) \tag{5}$$

where $X = [\mathbf{x}_1, ..., \mathbf{x}_{n+m}]^T$, and $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_{n+m}]^T$ are coefficients to be computed.

Note that the formulation and derivation of OwAR closely resemble the adaptation regularization - regularized least squares (ARRLS) algorithm proposed by Long et al. [14]; however, there are some major differences:

1) OwAR assumes labeled samples from the target domain can be obtained on-the-fly, whereas ARRLS assumes all samples in the target domain are available and unlabeled, i.e., OwAR considers online calibration, whereas ARRLS considers offline calibration.
2) OwAR explicitly considers the class imbalance problem in both source and target domains by introducing the weights on samples from different classes, whereas ARRLS does not consider that.
3) ARRLS also includes manifold regularization [2]. We investigated it but was not able to achieve improved performance in our application, so we excluded it in this paper.

Additionally, with the help of SDS, OwARSDS can effectively handle multiple source domains, whereas ARRLS only considers one source domain.

The OwAR algorithm is also very similar to the wAR algorithm in [29]. The major difference is that wAR considers offline calibration, which assumes there are m_u unlabeled target domain samples, in addition to the m labeled target domain samples, whereas OwAR considers online calibration, which only has access to the m labeled target domain samples. As a result, the procedure for solving $\boldsymbol{\alpha}$ in (5) is different.

C. OwAR: Loss Functions Minimization

The squared loss for regularized least squares

$$\ell(f(\mathbf{x}_i), y_i) = (y_i - f(\mathbf{x}_i))^2 \tag{6}$$

is considered in this paper. Let

$$\mathbf{y} = [y_1, ..., y_{n+m}]^T \tag{7}$$

where $\{y_1, ..., y_n\}$ are known labels in the source domain, and $\{y_{n+1}, ..., y_{n+m}\}$ are known labels in the target domain. Define $E \in R^{(n+m)\times(n+m)}$ as a diagonal matrix with

$$E_{ii} = \begin{cases} w_{s,i}, & 1 \leq i \leq n \\ w_t w_{t,i}, & n+1 \leq i \leq n+m \end{cases} \tag{8}$$

Then, substituting (5) and (6) into the first two terms in (2), it follows that

$$\begin{aligned} & \sum_{i=1}^{n} w_{s,i}\ell(f(\mathbf{x}_i), y_i) + w_t \sum_{i=n+1}^{n+m} w_{t,i}\ell(f(\mathbf{x}_i), y_i) \\ = & \sum_{i=1}^{n} w_{s,i}(y_i - f(\mathbf{x}_i))^2 + w_t \sum_{i=n+1}^{n+m} w_{t,i}(y_i - f(\mathbf{x}_i))^2 \\ = & \sum_{i=1}^{n+m} E_{ii}(y_i - f(\mathbf{x}_i))^2 = \| (\mathbf{y}^T - \boldsymbol{\alpha}^T K)E^{\frac{1}{2}} \|^2 \end{aligned} \tag{9}$$

[1]Strictly speaking, we should also make sure $P_s(y)$ is also close to $P_t(y)$. However, in this paper we assume all subjects conduct similar RSVP tasks, so $P_s(y)$ and $P_t(y)$ are intrinsically close. Our future research will consider the general case that $P_s(y)$ and $P_t(y)$ are different.

D. OwAR: Structural Risk Minimization

As in [14], [29], we define the structural risk as the squared norm of f in $\mathcal{H}_K$, i.e.,

$$\begin{aligned}\|f\|_K^2 &= \sum_{i=1}^{n+m} f(\mathbf{x}_i) \times \sum_{j=1}^{n+m} f(\mathbf{x}_j) \\ &= \sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \end{aligned} \tag{10}$$

E. OwAR: Marginal Probability Distribution Adaptation

Similar to [14], [20], we compute $D_{f,K}(P_s,P_t)$ using the projected maximum mean discrepancy (MMD):

$$\begin{aligned} D_{f,K}(P_s,P_t) &= \left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{1}{m}\sum_{i=n+1}^{n+m} f(\mathbf{x}_i)\right]^2 \\ &= \boldsymbol{\alpha}^T K M_0 K \boldsymbol{\alpha} \end{aligned} \tag{11}$$

where $M_0 \in R^{(n+m)\times(n+m)}$ is the MMD matrix:

$$(M_0)_{ij} = \begin{cases} \frac{1}{n^2}, & 1 \le i \le n, 1 \le j \le n \\ \frac{1}{m^2}, & n+1 \le i \le n+m, \\ & n+1 \le j \le n+m \\ \frac{-1}{nm_l}, & \text{otherwise} \end{cases} \tag{12}$$

F. OwAR: Conditional Probability Distribution Adaptation

Let $\mathcal{D}_{s,c} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{D}_s \wedge y_i = c\}$ be the set of samples in Class c of the source domain, $\mathcal{D}_{t,c} = \{\mathbf{x}_j | \mathbf{x}_j \in \mathcal{D}_t \wedge y_j = c\}$ be the set of samples in Class c of the target domain, $n_c = |\mathcal{D}_{s,c}|$, and $m_c = |\mathcal{D}_{t,c}|$. Then, the distance between the conditional probability distributions in source and target domains is computed as:

$$\begin{aligned} &D_{f,K}(Q_s,Q_t) \\ =&\sum_{c=1}^{2}\left[\frac{1}{n_c}\sum_{\mathbf{x}_i\in\mathcal{D}_{s,c}} f(\mathbf{x}_i) - \frac{1}{m_c}\sum_{\mathbf{x}_j\in\mathcal{D}_{t,c}} f(\mathbf{x}_j)\right]^2 \end{aligned} \tag{13}$$

Substituting (5) into (13), it follows that

$$\begin{aligned} &D_{f,K}(Q_s,Q_t) \\ =&\sum_{c=1}^{2}\left[\frac{1}{n_c}\sum_{\mathbf{x}_i\in\mathcal{D}_{s,c}} \boldsymbol{\alpha}^T K(X,\mathbf{x}) - \frac{1}{m_c}\sum_{\mathbf{x}_j\in\mathcal{D}_{t,c}} \boldsymbol{\alpha}^T K(X,\mathbf{x})\right]^2 \\ =&\sum_{c=1}^{2} \boldsymbol{\alpha}^T K M_c K \boldsymbol{\alpha} = \boldsymbol{\alpha}^T K M K \boldsymbol{\alpha} \end{aligned} \tag{14}$$

where

$$M = M_1 + M_2 \tag{15}$$

in which M_1 and M_2 are MMD matrices computed as:

$$(M_c)_{ij} = \begin{cases} 1/n_c^2, & \mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}_{s,c} \\ 1/m_c^2, & \mathbf{x}_i,\mathbf{x}_j \in \mathcal{D}_{t,c} \\ -1/(n_c m_c), & \mathbf{x}_i \in \mathcal{D}_{s,c}, \mathbf{x}_j \in \mathcal{D}_{t,c}, \text{or} \\ & \mathbf{x}_j \in \mathcal{D}_{s,c}, \mathbf{x}_i \in \mathcal{D}_{t,c} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

G. OwAR: The Closed-Form Solution

Substituting (9), (10), (11) and (14) into (2), it follows that

$$\begin{aligned} f =& \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} ||(\mathbf{y}^T - \boldsymbol{\alpha}^T K)E^{\frac{1}{2}}||^2 + \sigma\boldsymbol{\alpha}^T K\boldsymbol{\alpha} \\ &+ \lambda\boldsymbol{\alpha}^T K(M_0+M)K\boldsymbol{\alpha} \end{aligned} \tag{17}$$

Setting the derivative of the objective function above to 0 leads to

$$\boldsymbol{\alpha} = [(E + \lambda M_0 + \lambda M)K + \sigma I]^{-1} E\mathbf{y} \tag{18}$$

H. Source Domain Selection (SDS)

SDS is mainly used to reduce the computational cost, because when there are many source domains, performing OwAR for each source domain and then aggregating the results would be very time-consuming. However, SDS may also potentially improve classification performance, because source domains that are very noisy or very far away from the target domain may contaminate the results.

The SDS procedure in OwARSDS is very close to the one proposed in [29]. The only difference is that SDS in [29] also makes use of the unlabeled target domain samples, whereas here SDS only makes use of the m labeled target domain samples (because there are no unlabeled target domain samples in online calibration).

Assume there are Z different source domains. For the z^{th} source domain, we first compute $\mathbf{m}_{z,c}$ $(c = 1, 2)$, the mean vector of each class. Then, we also compute $\mathbf{m}_{t,c}$, the mean vector of each class in the target domain, from the m labeled samples. The distance between the two domains is:

$$d(z,t) = \sum_{c=1}^{2} ||\mathbf{m}_{z,c} - \mathbf{m}_{t,c}|| \tag{19}$$

We next cluster the Z numbers, $\{d(z,t)\}_{z=1,\ldots,Z}$, by k-means clustering, and finally choose the cluster that has the smallest centroid, i.e., the source domains that are closest to the target domain. In this way, on average we only need to performing OwAR for Z/k source domains. We used $k = 2$ in this paper.

I. The Complete OwARSDS Algorithm

The pseudo code for the complete OwARSDS algorithm is described in Algorithm 1. We first use SDS to select the closest source domains, and then perform OwAR for each selected source domain separately. The final classification is a weighted average of these individual classifiers, with the weight being the training accuracy of the corresponding OwAR.

III. Experiments and Discussions

RSVP experimental results are presented in this section to demonstrate the performance of OwARSDS.

A. Experiment Setup

Short 500-ms video clips were used in this RSVP experiment [15], [22]. Each video clip consisted of five consecutive 100-ms images. Targets are people or vehicles, and non-targets are background scenes. Participants were instructed to make a manual button press with their dominant hand only when they detected a target. 64-channel 512 Hz EEG signals were recorded using a BioSemi Active Two system. They were referenced offline to the average of the left and right earlobes. External leads were placed on the outer canthi and below the orbital fossa of both eyes to record electrooculography. The experiments were approved by U.S. Army Research Laboratory (ARL) Institutional Review Board [26], [27].

B. Preprocessing and Feature Extraction

Data from 15 subjects (9 male, ages 18-57, average 39.5) were used in this analysis. We used EEGLAB [7] for EEG signal preprocessing and feature extraction. Of the 64 BioSemi EEG channels, we only used 21 channels (Cz, Fz, P1, P3, P5, P7, P9, PO7, PO3, O1, Oz, POz, Pz, P2, P4, P6, P8, P10, PO8, PO4, O2) mainly in the parietal and occipital areas. We first band-passed the EEG signals to $[1, 50]$ Hz, then downsampled them to 64 Hz, and next epoched them to the $[0, 0.7]$ second interval timelocked to stimulus onset. We removed mean baseline from each channel in each epoch and removed epochs with incorrect button press responses. The final numbers of epochs from the 15 subjects are shown in Table I. Observe that there is significant class imbalance for every subject; that's why we need to use $w_{s,i}$ and $w_{t,i}$ in (2) to balance the two classes in both domains.

TABLE I
NUMBERS OF EPOCHS AND TARGETS FOR EACH SUBJECT AFTER PREPROCESSING.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Epochs	2,114	1,669	2,176	2,234	2,223	1,866	1,996	2,244	2,196	2,251	2,043	2,217	2,018	1,957	2,259
Number of Targets	338	266	349	355	354	301	320	357	347	358	327	352	325	317	357

Algorithm 1: The OwARSDS algorithm.

Input: Z source domains, where the z^{th} $(z = 1, ..., Z)$ domain has n_z labeled samples $\{\mathbf{x}_i^z, y_i^z\}_{i=1,...,n_z}$; m labeled target domain samples, $\{\mathbf{x}_j^t, y_j^t\}_{j=1,...,m}$; Parameters σ, λ, and k in k-means clustering.

Output: The OwARSDS classifier.

```
// SDS starts
```

if $m == 0$ **then**
  Select all Z source domains;
  Go to OwAR.
else
  for $z = 1, 2, ..., Z$ **do**
    Compute $d(z, t)$, the distance between the target domain and the z^{th} source domain, by (19).
  end
  Cluster $\{d(z, t)\}_{z=1,...,Z}$ by k-means clustering;
  Retain only the Z' source domains that belong to the cluster with the smallest centroid.
end

```
// SDS ends; OwAR starts
```

Choose a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$;
for $z = 1, 2, ..., Z'$ **do**
  Construct feature matrix $\{\mathbf{x}_j\}_{j=1,...,n_z+m}$, where the first n_z rows are the samples from the z^{th} source domain, and the next m rows are labeled samples from the target domain;
  Compute kernel matrix K_z from $\{\mathbf{x}_j\}_{j=1,...,n_z+m}$;
  Construct $\mathbf{y}$ in (7), E in (8), M_0 in (12), and M in (15);
  Compute $\boldsymbol{\alpha}$ by (18) and record it as $\boldsymbol{\alpha}_z$;
  Use $\boldsymbol{\alpha}_z$ to classify the $n_z + m$ labeled samples from both domains and record the accuracy as a_z;
end
Return $f(\mathbf{x}) = \sum_{z=1}^{Z'} a_z \boldsymbol{\alpha}_z K_z(X, \mathbf{x})$.

Each [0, 0.7] second epoch contains 21×45 raw EEG magnitude samples. To reduce the dimensionality, in each iteration of OwARSDS, we performed a simple principal component analysis to take the first 20 principal components as features. We then normalized each feature dimension separately to $[0, 1]$.

C. Evaluation Process and Performance Measure

Although the data was collected offline and we know the labels of all EEG epochs for all 15 subjects, we simulate a different and realistic incremental online calibration scenario: we have labeled EEG epochs for 14 subjects, but initially no labeled samples at all from the 15th subject; we generate labeled samples from the 15th subject iteratively and sequentially on-the-fly, which are used to train a classifier to label the remaining epochs from that subject. The performance measure is an average of the accuracy on target class and the accuracy on non-target class. We repeat this procedure 15 times so that each subject has a chance to be the "15th" subject.

More specifically, assume the 15th subject had M sequential epochs in the offline RSVP experiment, and in simulated incremental online calibration we want to label p successive epochs in each iteration, starting from zero labeled epoch. We first generate a random number $m_0 \in [1, M]$, representing the starting position in the RSVP sequence. Then, in the first iteration, we add Epochs m_0, $m_0 + 1$, ..., $m_0 + p - 1$ to the training dataset, build different classifiers, and test their performances on the remaining $M - p$ epochs. In the next iteration, we continue adding Epochs $m_0 + p$, $m_0 + p + 1$, ..., $m_0 + 2p - 1$, build different classifiers, and test their performances on the remaining $M - 2p$ epochs. We continue the iterations until the maximum number of iterations is reached. When the end of the sequence is reached, we rewind to the beginning of the sequence, e.g., if $m_0 = M$, then Epoch $m_0 + 1$ would be the 1st sample in the RSVP sequence, Epoch $m_0 + 2$ would be the 2nd, etc.. To obtain statistically meaningful results, this process was repeated 10 times, each time with a random starting point m_0.

D. Algorithms

We compared the performances of OwARSDS with five other algorithms:

1) Baseline 1 (BL1), which assumes we know all labels of the samples from the new subject, and uses 5-fold cross-validation and SVM to find the highest classification accuracy. This usually represents an upper bound of the classification performance we can get, by using the data from the new subject only.
2) Baseline 2 (BL2), which is a simple iterative procedure: in each iteration we sequentially select five unlabeled training samples from the new subject, label and add them to the labeled training dataset, and then train an SVM classifier by 5-fold cross-validation. We iterate until the maximum number of iterations is reached.
3) The TL algorithm introduced in [28], which simply combines the labeled samples from the new subject with samples from each existing subject and train a SVM classifier. The final classification is a weighted average of all individual classifiers, with weight being the cross-validation accuracy of the corresponding classifier.
4) TLSDS, which is the above TL algorithm with SDS.
5) OwAR, which uses data from all 14 existing subjects as auxiliary data, instead of performing SDS.

Weighted libSVM [5] with RBF kernel was used as the classifier in BL1, BL2, TL and TLSDS. We chose $w_t = 2$ in OwAR and OwARSDS to give the labeled target domain samples more emphasis, and $\sigma = 0.1$ and $\lambda = 10$, following the practice in [14], [29].

E. Experimental Results

The performances of the six algorithms, which are averaged across the 10 runs for each subject, are shown in Fig. 1, where each subfigure represents a different "15th" subject. The average performance of the six algorithms across the 15 subjects is shown in Fig. 2. Observe that:

1) Generally the performances of all algorithms (except BL1, which is not iterative) increase as more subject-specific training samples are labeled and added, which is intuitive.

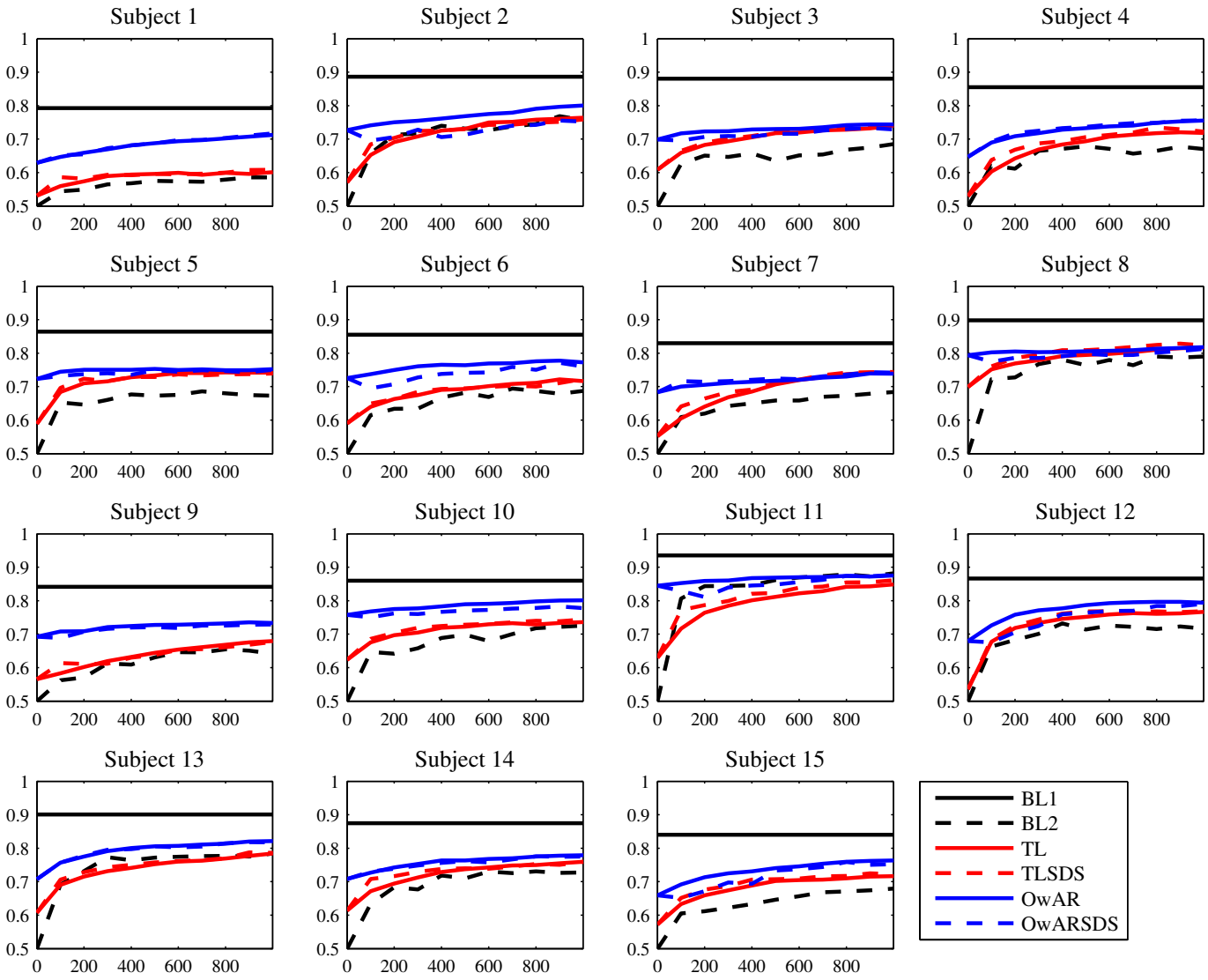

Fig. 1. Performance of the six algorithms for each individual subject, averaged over 10 runs. Horizontal axis: m, number of labeled subject-specific training samples; vertical axis: classification accuracy.

2) BL2 cannot build a model when there are no labeled samples at all from the new subject (observe that the first point on the BL2 curve in Fig. 1 is always .5, representing random guess), but all TL based algorithms can, because they can borrow information from other subjects. Moreover, without any labeled samples from the new subject, OwAR and OwARSDS can build a model with an average classification accuracy of 71.17%, which is much better than random guess.
3) On average TL outperforms BL2, because it can borrow useful information from other subjects.
4) TLSDS always outperforms TL, especially when m is small. This is because TL uses a very simple way to combine the samples from the new and existing subjects, and hence an existing subject whose ERPs are significantly different from the new subject's would have a negative impact on the final classification performance. SDS removes (some of) such subjects, and hence benefits the performance.
5) On average OwAR and OwARSDS significantly outperform BL2, TL, and TLSDS. This is because a sophisticated domain adaptation algorithm is used in OwAR and OwARSDS, which explicitly considers class imbalance, and is optimized not only for high classification accuracy, but also for small structural risk and close similarity of the features.
6) On average OwARSDS has slightly worse performance than OwAR, but instead of using 14 subjects in OwAR, OwARSDS uses on average 5.8 subjects in our experiment, which represents a 58.57% computational cost saving.
7) SDS selects the most similar source domains according to the distance on the class-conditional feature means, which performs well for the majority of the 15 subjects. However, Fig. 1 also shows that for Subjects 2 and 11, the performance of OwARSDS may be worse than BL2 sometimes. This suggests that SDS still has room for improvement. Alternatively, it may be possible to develop some heuristics on when the current SDS approach may not work well. These are some of our future research directions.

Non-parametric multiple comparison tests using Dunn's procedure [8], [9] was used to determine if the difference between any pair of

algorithms is statistically significant, with a p-value correction using the FDR method by [3]. The overall performance of each of the five algorithms (BL1 was not included because it is not iterative) was measured by the area-under-performance-curve, the area under the curve of the accuracies obtained at each of the 10 runs, normalized to $[0, 1]$. The results showed that the performances of OwAR and OwARSDS are statistically significantly different from BL2, TL, and TLSDS ($p = .0000$ in all cases). There is no statistically significant performance difference between OwAR and OwARSDS ($p = .0263$).

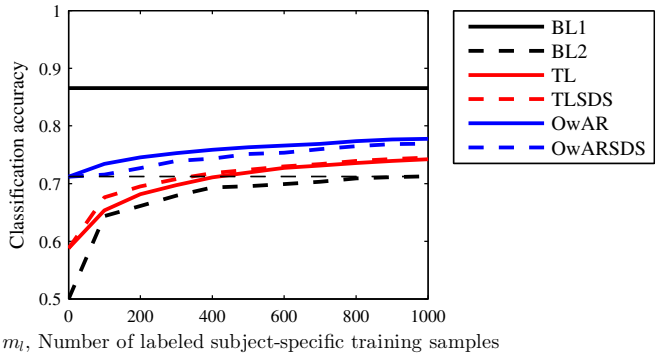

Fig. 2. Average performance of the six algorithms across the 15 subjects.

In summary, we have demonstrated that given the same number of labeled subject-specific training samples, OwAR and OwARSDS can significantly improve online calibration performance. In other words, given a desired classification accuracy, OwAR and OwARSDS can reduce the number of labeled subject-specific training samples. For example, in Fig. 2, the average classification accuracy of BL2 is 71.25%, given 1000 labeled subject-specific training samples. However, OwAR and OwARSDS can achieve average accuracy 71.17% without using any subject-specific training samples.

IV. Conclusions

RSVP-based BCI system relies on single-trial classification of ERPs. Because of large individual differences, some labeled subject-specific data are needed to calibrate the classifier for each new subject. This paper proposed a novel OwAR algorithm to reduce the online calibration effort, and hence to increase the utility of the BCI system. We showed that given the same number of labeled subject-specific training samples, OwAR can significantly improve the online calibration performance. In other words, given a desired classification accuracy, OwAR can significantly reduce the number of labeled subject-specific training samples. Furthermore, we also showed that the computational cost of OwAR can be reduced by more than 50% by source domain selection, without a statistically significant sacrifice of classification performance. As individual differences are also a well-known and pervasive phenomenon in psychology and affective computing, we believe our approaches are also applicable there.

Acknowledgment

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S Government.

References

[1] M. Ahn, H. Cho, and S. C. Jun, "Calibration time reduction through source imaging in brain computer interface (BCI)," *Communications in Computer and Information Science*, vol. 174, pp. 269–273, 2011.

[2] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, pp. 289–300, 1995.

[4] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappa, Eds., *The Oxford Handbook of Affective Computing*. Oxford University Press, 2014.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm.

[6] M. M. Chun and M. C. Potter, "A two-stage model for multiple target detection in rapid serial visual presentation," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 1, pp. 109–127, 1995.

[7] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[8] O. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, pp. 62–64, 1961.

[9] ——, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, pp. 214–252, 1964.

[10] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience," *Journal of personality and social psychology*, vol. 69, no. 1, pp. 153–166, 1995.

[11] T. Ito and J. Cacioppo, "Variations on a human universal: Individual differences in positivity offset and negativity bias," *Cognition & Emotion*, vol. 19, pp. 1–26, 2005.

[12] P.-J. Kindermans and B. Schrauwen, "Dynamic stopping in a calibration-less P300 speller," in *Proc. 5th Int'l. Brain-Computer Interface Meeting*, Pacific Grove, CA, June 2013.

[13] P.-J. Kindermans, H. Verschore, D. Verstraeten, and B. Schrauwen, "A P300 bci for the masses: Prior information enables instant unsupervised spelling," in *Proc. Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, December 2012.

[14] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2014.

[15] A. Marathe, V. Lawhern, D. Wu, D. Slayback, and B. Lance, "Improved neural signal classification in a rapid serial visual presentation task using active learning," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 2015, submitted.

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

[17] L. Parra, C. Christoforou, A. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. Philiastides, and P. Sajda, "Spatiotemporal linear decoding of brain state," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 107–115, 2008.

[18] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S.-F. Chang, and P. Sajda, "Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases," *Journal of Neural Engineering*, vol. 8, no. 3, 2011.

[19] M. C. Potter, "Short-term conceptual memory for pictures," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 2, no. 5, pp. 509–522, 1976.

[20] B. Quanz and J. Huan, "Large margin transductive transfer learning," in *Proc. 18th ACM Conf. on Information and Knowledge Management (CIKM)*, Hong Kong, November 2009.

[21] W. Revelle, J. Wilt, and D. M. Condon, "Individual differences and differential psychology: A brief history and prospect," in *Handbook of Individual Differences*, T. Chamorro-Premuzic, A. Furnham, and S. von Stumm, Eds. Wiley-Blackwell, 2011.

[22] A. J. Ries and G. B. Larkin, "Stimulus and response-locked P3 activity in a dynamic rapid serial visual presentation (RSVP) task," DTIC, Tech. Rep. ARL-TR-6314, 2013. [Online]. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA579452

[23] P. Sajda, E. Pohlmeyer, J. Wang, L. Parra, C. Christoforou, J. Dmochowski, B. Hanna, C. Bahlmann, M. Singh, and S.-F. Chang, "In a blink of an eye and a switch of a transistor: Cortically coupled computer vision," *Proc. of the IEEE*, vol. 98, no. 3, pp. 462–478, 2010.

[24] W. Samek, F. Meinecke, and K.-R. Muller, "Transferring subspaces between subjects in brain-computer interfacing," *IEEE Trans. on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[25] K. R. Scherer, T. Banziger, and E. Roesch, Eds., *A Blueprint for Affective Computing*. Oxford University Press, 2010.

[26] US Department of Defense Office of the Secretary of Defense, "Code of federal regulations protection of human subjects," *Government Printing Office*, vol. 32 CFR 19, 1999.

[27] US Department of the Army, "Use of volunteers as subjects of research," *Government Printing Office*, vol. AR 70-25, 1990.

[28] D. Wu, B. J. Lance, and V. J. Lawhern, "Active transfer learning for reducing calibration data in single-trial classification of visually-evoked potentials," in *Proc. IEEE Int'l. Conf. on Systems, Man, and Cybernetics*, San Diego, CA, October 2014.

[29] D. Wu, V. J. Lawhern, and B. J. Lance, "Reducing offline BCI calibration effort using weighted adaptation regularization with source domain selection," in *Proc. IEEE Int'l. Conf. on Systems, Man and Cybernetics*, Hong Kong, October 2015.

[30] T. O. Zander and C. Kothe, "Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, 2011.