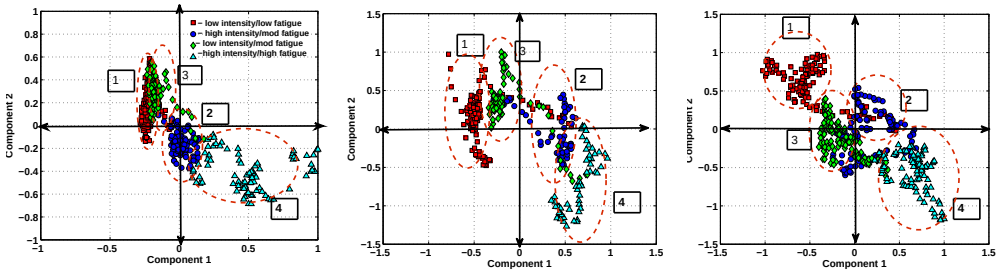
Multi-Source Domain Adaptation and Its Application to Early Detection of Fatigue

RITA CHATTOPADHYAY, Arizona State University
QIAN SUN, Arizona State University
JIEPING YE, Arizona State University
SETHURAMAN PANCHANATHAN, Arizona State University
WEI FAN, IBM T.J.Watson Research
IAN DAVIDSON, University of California

We consider the characterization of muscle fatigue through noninvasive sensing mechanism such as surface electromyography (SEMG). While changes in the properties of SEMG signals with respect to muscle fatigue have been reported in the literature, the large variation in these signals across different individuals makes the task of modeling and classification of SEMG signals challenging. Indeed, the variation in SEMG parameters from subject to subject creates differences in the data distribution. In this paper, we propose two transfer learning frameworks based on the multi-source domain adaptation methodology for detecting different stages of fatigue using SEMG signals, that addresses the distribution differences. In the proposed frameworks, the SEMG data of a subject represent a domain; data from multiple subjects in the training set form the multiple source domains and the test subject data form the target domain. SEMG signals are predominantly different in conditional probability distribution across subjects. The key feature of the first framework is a novel weighting scheme that addresses the conditional probability distribution differences across multiple domains (subjects) and the key feature of the second framework is a two-stage domain adaptation methodology which combines weighted data from multiple sources based on marginal probability differences (first stage) as well as conditional probability differences (second stage), with the target domain data. The weights for minimizing the marginal probability differences are estimated independently, while the weights for minimizing conditional probability differences are computed simultaneously by exploiting the potential interaction among multiple sources. We also provide a theoretical analysis on the generalization performance of the proposed multi-source domain adaptation formulation using the weighted Rademacher complexity measure. We have validated the proposed frameworks on Surface Electromyogram signals collected from 8 people during a fatigue-causing repetitive gripping activity. Comprehensive experiments on the SEMG dataset demonstrate that the proposed method improves the classification accuracy by 20% to 30% over the cases without any domain adaptation method and by 13% to 30% over the existing state-of-the-art domain adaptation methods.

1. INTRODUCTION



(a) Subject 1 (b) Subject 2 (c) Subject 4

Fig. 1. Three sample subjects (subjects 1, 2, 4) with four classes (four physiological stages) in our SEMG dataset: SEMG signals are predominantly different in conditional probability distribution across subjects.

Daily life activities such as typing on the keyboard, dusting, brooming, ironing, as well as the use of hand tools such as scissors and knives, repetitive work in assembly lines, repetitive lifting, involve repetitive movements of the different parts of the body. It has been proved that repetitive task makes work particularly hazardous, as it is the primary cause of muscle fatigue [Silverstein et al. 1986; Higgs 1992; Young et al. 1995]. According to the US Bureau of Labor Statistics, in 2002, there

were more than 345,000 on the job back injuries, due to muscle fatigue, which required time off from work. According to the Bureau of Labor Statistics (2004), annual direct cost of occupational injuries due to slip and fall caused due to muscle fatigue is expected to exceed $43.8 billion by the year 2020 in the US.

These accidents and the consequential loss in work hours and lifes, besides the high medical cost, can be avoided if one can intervene such fatigue inducing repetitive activities at an early stage by intelligent devices having the capability for detecting different stages of fatigue. Technologies for detecting muscle fatigue at an early stage can also be used to remove the cause of fatigue by altering the environmental ergonomics where possible

There are a number of techniques that can be used to objectively determine the level of fatigue in a subject. Electromyography (EMG) is a method for biosignal recording of skeletal muscle activity. Surface Electromyography (SEMG) allows for noninvasive recording of these biosignals. Researchers have observed that certain aspects of SEMG signals change as a muscle becomes fatigued. Localized muscle fatigue has been correlated with a shift in the power spectral density of SEMG signals, root mean square (rms), instantaneous frequency, zero crossing rate, mean-frequency, median-frequency, etc. However, there is a large variation in the values of these measures across different subjects, due to variances in their SEMG power spectrum and their shifts. These generally unpredictable and wide variations make the task of modeling SEMG difficult, and the task of automating the process of signal classification as a generalized tool challenging. The variation in SEMG parameters from subject to subject creates differences in the data distribution. Figure 1 shows the distribution of the data over four stages of a fatigue-causing activity, done with varying speed, for three different subjects (subjects 1, 2, 4). The data distribution shown in Figure 1 is of factor scores obtained as a result of factor analysis applied on the twelve dimensional feature vectors derived from raw SEMG signals[1]. The four physiological stages corresponding to four classes, shown in the figure, are (l) *low intensity of activity and low fatigue*, (2) *high intensity of activity and moderate fatigue*, (3) *low intensity of activity and moderate fatigue* and (4) *high intensity of activity and high fatigue*. We observe that the data distribution during each stage or class varies from subject to subject. This variation leads to predominantly conditional probability differences across subjects.

Traditional data mining algorithms assume that training data and test data are drawn from the same distribution, and they may not be effective if the assumption is violated as in the case of SEMG data over multiple subjects. One effective approach is domain adaptation which enables transfer of knowledge between the source and target domains [Pan and Yang 2009]. It has been applied successfully in various applications [Blitzer et al. 2007; Duan et al. 2009; Daumé III 2007; Pan et al. 2008] including text classification (parts of speech tagging, webpage tagging, etc), video concept detection across different TV channels, sentiment analysis (identifying positive and negative reviews across domains), WiFi Localization (locating device location depending upon the signal strengths from various access points.

In this paper we present a successful case study of application of multi-source domain adaptation techniques for detecting different stages of fatigue based on the Surface Electromyogram signals across multiple subjects. To the best of our knowledge, this is the first systematic analysis of subject based variability in SEMG signals. The proposed frameworks address the subject based variability, predominantly the distribution differences in conditional probabilities in Surface Electromyogram signals. Specifically, a classifier is learnt to distinguish the four classes as shown in Figure 1 on the basis of some labeled and unlabeled data from the target domain (or subject).

In the first proposed framework named as *Conditional Probability based Multi-source Domain Adaptation* (CP-MDA) the unlabeled data are labeled using a weighting scheme that measures the similarities in conditional probabilities between the source and target domain data; the key of this proposed weighting scheme is a joint optimization framework based on smoothness assumption on the probability distribution of the target domain data.

[1]More details on the twelve features derived, the factor analysis results and a real time deployment of fatigue grading framework can be found in our earlier papers [Chattopadhyay et al. 2009; Chattopadhyay et al. 2010].

The second multi-source domain adaptation framework, named as *Two Stage Weighting Framework for Multi-source Domain Adaptation* (2SW-MDA) computes weights for the data samples from multiple sources to reduce both marginal and conditional probability differences between the source and target domains. In the first stage, we compute weights of the source domain data samples to reduce the marginal probability differences, using Maximum Mean Discrepancy (MMD) [Borgwardt et al. 2006; Huang et al. 2007] as the measure. The second stage computes the weights of multiple sources to reduce the conditional probability differences; the computation is based on the smoothness assumption on the conditional probability distribution of the target domain data. Finally, a target classifier is learned on the re-weighted source domain data. A novel feature of our weighting methodologies is that no labeled data is needed from the target domain, thus widening the scope of their applicability. The proposed framework is easily extendable to the case where a few labeled data may be available from the target domain.

In addition, we present a detailed theoretical analysis on the generalization performance of our proposed framework. The error bound of the proposed target classifier is based on the weighted Rademacher complexity measure of a class of functions or hypotheses, defined over a weighted sample space [Bartlett and Mendelson 2002; Koltchinskii 2001]. The Rademacher complexity measures the ability of a class of functions to fit noise. The empirical Rademacher complexity is data-dependent and can be measured from finite samples. It can lead to tighter bounds than those based on other complexity measures such as the VC-dimension. Theoretical analysis of domain adaptation has been studied in [Ben-David et al. 2010; Mansour et al. 2009a]. In [Ben-David et al. 2010], the authors provided the generalization bound based on the VC dimension for both single-source and multi-source domain adaptation. The results were extended in [Mansour et al. 2009a] to a broader range of prediction problems based on the Rademacher complexity; however only the single-source case was analyzed in [Mansour et al. 2009a]. We extend the analysis in [Ben-David et al. 2010; Mansour et al. 2009a] to provide the generalization bound for our proposed two-stage framework based on the weighted Rademacher complexity; our generalization bound is tighter than the previous ones in the multi-source case.

We have applied the proposed algorithms to Surface Electromyogram signals collected from 8 people during a fatigue-causing repetitive gripping activity. Our extensive experiments on the SEMG dataset demonstrate that the proposed methods improves the subject independent classification accuracy by 20% to 30% over the cases without any domain adaptation method and by 13% to 30% over the existing state-of-the-art domain adaptation methods.

2. PROPOSED FRAMEWORKS

The proposed domain adaptation frameworks focuses on learning from multiple auxiliary sources related to the target data, e.g., multiple subject data having different distributions, collected under similar physiological conditions. Specifically, we consider the problem of detecting different stages of fatigue in a subject for whom we have very few labeled samples available in the training data. The training data also includes data from many other subjects collected under similar physiological conditions. The test subject data forms target domain data and the multiple subject data in the training domain form multiple auxiliary sources.

2.1. Problem Setting and Motivation

Assume that there are k subjects in the source domain. The s-th subject in the source domain is characterized by a sample set $D^s = (x_i^s, y_i^s)|_{i=1}^{n_s}$, where x_i^s is the feature vector, y_i^s is the corresponding label, and n_s is the total number of samples for the subject s. The target domain consists of a few labeled data $D_l^T = (x_i^T, y_i^T)|_{i=1}^{n_l}$ and plenty of unlabeled data $D_u^T = x_i^T|_{i=n_l+1}^{n_l+n_u}$ where n_l and n_u are numbers of labeled and unlabeled target domain samples respectively, $D^T = D_l^T \bigcup D_u^T$, and $n_T = n_l + n_u$. The goal is to develop a target classifier f^T that can predict the labels of the unlabeled data in the target domain, using the multi-source domain data and a few labeled target domain data.

Table I. Some of the notations used in the paper.

Notation	Explanation
k	Total number of source domains
D^s	s-th subject in source domain
D_l^T	Labeled data from target domain
D_u^T	Unlabeled data from target domain
D^T	Total data from target domain
n_l	Number of labeled target domain data
n_u	Number of unlabeled target domain data
n_T	Number of total target domain data
β	$k \times 1$ weight vector, based on conditional probability distribution difference
α_i^s	Weight of i-th data in s-th source domain based on marginal probability distribution difference
$\hat{\epsilon}_{\alpha,\beta}(h)$	Empirical error function on (α, β)-weighted source domain data
$\hat{\epsilon}_T(h)$	Empirical error function on target domain data
$\hat{E}^S_{\alpha,\beta}(h)$	Empirical joint error function on (α, β)-weighted source and the target domain data
$E^S_{\alpha,\beta}(h)$	True joint error function
$\epsilon_T(h)$	True error function on target domain data

One simple approach for predicting the labels of the target domain data is to combine the training samples from all subjects and build a single classifier based on the pooled training samples. However, this simple approach will not work well in our application as there are significant conditional probability differences across subjects in the SEMG data. A better alternative is to learn individual models for each subject in the source domain, and then combine the hypothesis generated by each of these source models, on the basis of some similarity measures between the source and target domains. The similarity measures are commonly computed by considering each source domain data separately. This procedure has two potential limitations. First, it minimizes the loss with respect to the probability distribution $P_s(x, y|D^s)$ on the source domain which will not generally coincide with the minimal loss on the distribution $P_T(x, y|D^T)$ on the target domain. Second, it assumes all sources are independent when computing the similarity measures, thus it does not fully exploit the interaction among multiple sources.

Thus one of the important issues is to choose the right similarity measure between the source and target domains depending upon the nature of differences in the distribution.

In this paper, we present two multi-source domain adaptation methodologies CP-MDA and 2SW-MDA. CP-MDA addresses predominantly conditional probability differences between the source and target domain, where as 2SW-MDA addresses both marginal and conditional probability differences.

We observe there are significant conditional probability differences in our multi-subject SEMG data as shown in Figure 1. In addition, we observe from the figure that different classes vary differently over subjects. Hence we present here a conditional probability based weighting scheme that computes the weights of each source in a joint optimization framework that takes care of mutual dynamics between the subjects. Table I summarizes a few of the notations used in the paper.

2.2. Conditional Probability based Multi-source Domain Adaptation (CP-MDA)

We learn a classifier f^T for the target domain data, using a few labeled samples and a large number of unlabeled samples from the target domain. The key of this proposed approach is a novel weighting scheme that integrates multiple source domain data using a set of weights, one for each source domain. We use these weights to compute the labels of the unlabeled target domain data, called "pseudo labels". The target domain prediction model is then learned from both labeled and pseudo labeled target domain samples in a regularized framework. Specifically, the proposed multi-source domain adaptation framework is given as follows:

$$\min_{f^T \in H_K} \gamma_A \|f^T\|_K^2 + \frac{1}{n_l} \sum_{i=1}^{n_l} V(x_i, y_i, f^T) + \Omega_r(f_u^T) + \Omega_m(f^T) \tag{1}$$

The first term controls the complexity of the classifier f^T in the Reproducing Kernel Hilbert Space (RKHS) H_K, γ_A controls the penalty factor, the second term is the empirical error of the target classifier f^T on the few labeled target domain data D_l^T, and n_l is the number of labeled target domain data. The empirical error on the unlabeled target data, labeled using a conditional probability based weighting scheme, forms the third term. This regularizer enforces the target classifier f^T to have similar decision values to the auxiliary source which has similar conditional probability distribution, explained in detail in Subsection 2.2.1. The fourth term is a manifold based regularizer based on the smoothness assumption [Belkin et al. 2006] on target domain data: if two points x_i and x_j are close to each other in the intrinsic geometry of marginal distribution then they are most likely to have similar conditional probabilities, i.e., $f^T(x_i)$ should be similar to $f^T(x_j)$. The *manifold based regularizer* is defined as in [Belkin et al. 2006]:

$$\Omega_m(f^T) = \frac{\gamma_I}{n_T^2} f^{T'} L f^T. \tag{2}$$

where L is the graph Laplacian matrix constructed on D^T, $f^T = [f^T(x_1), \cdots, f^T(x_{n_T})]$, γ_I controls the complexity of the function f^T in the intrinsic geometry of the marginal probability of x and the normalizing coefficient $\frac{1}{n_T^2}$ is the natural scale factor for the empirical estimate of the Laplace operator, and the symbol $'$ is used to represent the matrix or vector *transpose* operation.

2.2.1. Multi-Source Weighting. Let $f_u^T = [f_{n_{l+1}}^T \cdots f_{n_T}^T]'$ be the decision values of the target classifier f^T for the unlabeled target domain data and let $f_u^s = [f_{n_{l+1}}^s \cdots f_{n_T}^s]'$ be the decision values of the s-th auxiliary classifier for the same unlabeled target domain data. Let β^s be the measure of relevance or similarity between the distributions of the s-th source and the target data, and let $f_j^T = f^T(x_j)$ be the decision value of target classifier on the target domain data x_j and $f_j^s = f^s(x_j)$ be the decision value of the s-th auxiliary source classifier on x_j. We use a weighted combination of the k source domain classifiers f^s to estimate the target classifier. Specifically, the estimated label ($\hat{y}_j$) of the unlabeled target data x_j based on the k source domain classifiers f^s is given by

$$\hat{y}_j = \sum_{s=1}^{k} \beta^s f_j^s, \tag{3}$$

where $\beta^s > 0$ is the weight for the s-th source. We assume that the weights are normalized, that is, $\sum_s \beta^s = 1$. The auxiliary classifier f^s for the s-th source is pre-computed based on its respective data. The auxiliary classifiers f^s and the target classifier f^T can be trained using different kernels or even different learning methods. The resulting regularizer, $\Omega_r(f_u^T)$, named as *relevance based regularizer* measures the difference between the target classifier decision value and the estimation based on multiple source data, and is defined as follows:

$$\Omega_r(f_u^T) = \frac{\theta}{2} \sum_{j=n_l+1}^{n_T} \|f_j^T - \sum_{s=1}^{k} \beta^s f_j^s\|^2, \tag{4}$$

where $\theta > 0$ is a constant. θ is used to control the relative importance of true labels and psuedo labels. The weight β^s which provides a measure of relevance between the s-th auxiliary source domain and the target domain is computed on the basis of a *Conditional Probability based Weighting Scheme*, which evaluates the similarities in distributions between the source and target domains predominantly based on conditional probability differences.

Next, we show how to estimate the weights β^s's. The proposed weighting scheme evaluates the similarities between auxiliary source data and the target domain data considering the similarities in their conditional probabilities.

Let $F_i^S = [f_i^1 \cdots f_i^k]$ be the $1 \times k$ vector of predicted labels of k auxiliary source models for the i-th sample of target domain data. Let $\beta = [\beta_1 \cdots \beta_k]^T$ be the $k \times 1$ weight vector, where β^s is

the weight corresponding to the s-th auxiliary source. Following (3), the predicted label for the i-th sample of target domain data is

$$\hat{y}_i = \sum_{s=1}^{k} \beta^s f_i^s = F_i^S \beta. \tag{5}$$

This motivates us to estimate the weight vector β based on the smoothness assumption on the conditional probability distribution: we compute the optimal weight vector β by minimizing the difference in predicted labels between two nearby points in the target domain. Specifically, the proposed weighting framework solves the following problem:

$$\min_{\beta:\beta' e=1,\beta\geq 0} \sum_{i,j=n_l+1}^{n_l+n_u} (F_i^S\beta - F_j^S\beta)^2 W_{ij} \tag{6}$$

where $F_i^S\beta$ and $F_j^S\beta$ are the predicted labels for i-th and j-th samples of target domain data and W_{ij} is the edge weight between the i-th and j-th samples given by $e^{\frac{-(x_i-x_j)^2}{2\sigma^2}}$. We can rewrite the minimization problem as follows:

$$\min_{\beta:\beta' e=1,\beta\geq 0} \beta' \left(F^S\right)' L_u F^S \beta \tag{7}$$

where F^S is an $n_u \times k$ matrix with each row of F^S being the $1 \times k$ vector of k predicted labels for a sample of target domain data and L_u is normalized graph Laplacian associated with the target domain data D_u^T, given by $L_u = I - D^{-0.5} W D^{-0.5}$, where I is the identity matrix of size n_u, W is the adjacency graph defining edge weights between the n_u unlabled samples in the target domain data, and D is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{n_u} W_{ij}$.

The minimization problem in (7) is a standard quadratic problem and can be solved by applying many existing solvers. We simply use the 'quadprog' function in MATLAB. With the computed weights, the labels for the unlabeled target domain data, called *psuedo labels*, are computed using (3), and are substituted into the regulariser in (4).

Intuitively, by enforcing that nearby points in the marginal distribution of the target data have similar class labels (or conditional probability) via the optimization in (7), the proposed weighting scheme is likely to give higher weights to those sources with the conditional probability distribution similar to the target data. This is verified in our empirical study on both SEMG and synthetic data. If a source has a conflicting conditional distribution as the target, it is likely to get a low or even zero weight. In addition, different from many existing weight schemes which compute the weights by considering each source independently, the proposed weighting scheme computes the optimal value of β or the optimal weights of all the k sources simultaneously, thus taking the potential interaction among multiple subjects in the source domain into account.

2.2.2. Proposed Algorithm. Using the least square error and substituting the regularizers we can rewrite (1) as follows:

$$\begin{aligned} &\min_{f^T \in H_K} \gamma_A \|f^T\|_K^2 + \frac{1}{n_l} \sum_{i=1}^{n_l} (f_i^T - y_i^T)^2 \\ &+\frac{\theta}{2} \sum_{j=n_l+1}^{n_T} \|f_j^T - \sum_{s=1}^{k} \beta^s f_j^s\|^2 + \frac{\gamma_I}{n_T^2} f^{T'} L f^T \end{aligned} \tag{8}$$

By the Representer theorem [Schlkopf and Smola 2002], we can find an optimal solution of (8), which is a linear expansion of the kernel function K, over both the labeled D_l^T and the pseudo

labeled target domain data D_u^T given as follows:

$$f^T(x) = \sum_{i=1}^{n_l+n_u} \alpha_i K(x_i, x). \tag{9}$$

Substituting this into (8), we can obtain the optimal $\alpha = [\alpha_1 \cdots\cdots\cdots \alpha_{n_l+n_u}]^T$ by solving the following optimization problem:

$$\begin{aligned} \min_{\alpha} \frac{1}{n_l + \theta n_u} (Y - K\alpha)^{'} J (Y - K\alpha) \\ + \gamma_A \alpha^{'} K \alpha + \frac{\gamma_I}{(n_u + n_l)^2} \alpha^{'} K L K \alpha \end{aligned} \tag{10}$$

where K is the $(n_l + n_u) \times (n_l + n_u)$ kernel Gram matrix over the target domain data, Y is the label vector over labeled and pseudo labeled target domain data points given by:

$$\left[y_1 \cdots y_{n_l} \sum_s \beta^s_{(n_l+1)} f^s_{(n_l+1)} \cdots \sum_s \beta^s_{(n_l+n_u)} f^s_{(n_l+n_u)} \right] \tag{11}$$

L is the graph Laplacian defined over labeled and pseudo labeled target domain data, and J is a diagonal matrix of size $(n_l + n_u) \times (n_l + n_u)$ given by $J = diag(1, \cdots 1, \theta, \cdots \theta)$ with the first n_l diagonal entries as 1 and the rest as θ. θ is assigned a number between 0 and 1, thus the pseudo labels of the target domain data get smaller weights compared to the labels of the labeled target domain data. From (10), the optimal α^* is given by:

$$\alpha^* = \left(JK + \gamma_A (n_l + \theta n_u) I + \frac{\gamma_I (n_l + \theta n_u)}{(n_u + n_l)^2} LK \right)^{-1} JY.$$

With the computed α^*, the prediction of any unseen test data x is given by:

$$f^T(x) = \sum_{i=1}^{n_l+n_u} \alpha_i^* K(x_i, x). \tag{12}$$

Since the proposed domain adaptation framework is based on multiple sources whose similarities to target domain data or weights are computed based on a conditional probability based weighting scheme, we refer the proposed framework as *Conditional Probability based Multi-Source Domain Adaptation (CP-MDA)*.

2.3. Two Stage Weighting Framework for Multi-source Domain Adaptation (2SW-MDA)

The second proposed approach consists of two stages. In the first stage, we compute the weights of source domain data based on the marginal probability difference; in the second stage, we compute the weights of source domains based on the conditional probability difference, as described Section 2.2.1. A target domain classifier is learned on these re-weighted data.

2.3.1. Re-weighting data samples based on marginal probability differences. The difference between the means of two distributions after mapping onto a reproducing kernel Hilbert space, called Maximum Mean Discrepancy, has been shown to be an effective measure of the differences in their marginal probability distributions [Borgwardt et al. 2006]. We use this measure to compute the weights α_i^s's of the s-th source domain data by solving the following optimization problem [Huang

et al. 2007]:

$$\min_{\alpha^s} \quad \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \alpha_i^s \Phi(x_i^s) - \frac{1}{n_T} \sum_{i=1}^{n_T} \Phi(x_i^T) \right\|_H^2 \qquad (13)$$
$$\text{s.t.} \quad \alpha_i^s \geq 0$$

where $\Phi(x)$ is a feature map onto a reproducing kernel Hilbert space H [Steinwart 2001], n_s is the number of samples in the s-th source domain, n_T is the number of samples in the target domain, and α^s is the n_s dimensional weight vector. The minimization problem is a standard quadratic problem and can be solved by applying many existing solvers. We simply use the 'quadprog' function in MATLAB.

2.3.2. Re-weighting Sources based on Conditional probability differences. In the second stage the proposed framework modulates the α^s weights of a source domain s obtained on the basis of marginal probability differences in the first stage, with the weighting factor β^s computed as described in Section 2.2.1. The weight β^s reflects the similarity of a particular source domain s to the target domain with respect to conditional probability distributions.

To illustrate the proposed two-stage framework, we demonstrate the effect of re-weighting data samples in source domains D1 and D2 of the toy dataset (shown in Figure 2), based on the computed weights, in the appendix. Figure

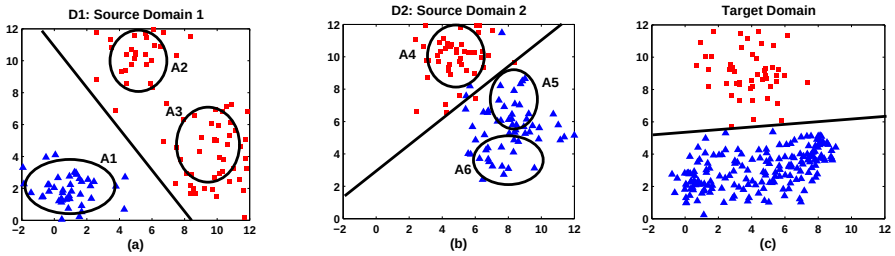


Fig. 2. Two source domains D1 and D2 and target domain data with different marginal and conditional probability differences, along with conflicting conditional probabilities (the red squares and blue triangles refer to the positive and negative classes).

2.3.3. Learning the Target Classifier. The target classifier is learnt based on the re-weighted source data and the few labeled target domain data (if available). We also incorporate an additional weighting factor μ to provide a differential weight to the source domain data with respect to the labeled target domain data. Mathematically, the target classifier $\hat{h}$ is learnt by solving the following optimization problem:

$$\hat{h} = \underset{h}{\text{argmin}} \quad \mu \left(\sum_{s=1}^{k} \frac{\beta^s}{n_s} \sum_{i=1}^{n_s} \alpha_i^s \mathcal{L}(h(x_i^s), y_i^s) \right) + \sum_{j=1}^{n_l} \frac{1}{n_l} \mathcal{L}(h(x_j^T), y_j^T) \qquad (14)$$

where n_l is the number of labeled data from the target domain.

We refer to the proposed framework as *2-Stage Weighting framework for Multi-Source Domain Adaptation (2SW-MDA)*. Algorithm 1 below summarizes the main steps involved in 2SW-MDA.

We now present a theoretical analysis of the joint loss function given in Equation 14 and present an upper bound on the error of the target classifier h (learned by minimizing 14) on target domain data. To do this, we first prove an upper bound on the empirical joint error with respect to the true joint error and then prove an upper bound on the error on the target domain data only.

ALGORITHM 1: 2SW-MDA

Input μ, k source domain data $\{D^s\}_{s=1}^k$, unlabeled target domain data D_u^T and labeled target domain data D_l^T (if available)
Output Target classifier h
for $s = 1, \ldots, k$ **do**
    Compute α^s by solving (13)
    Learn a hypothesis h^s on the α^s weighted source data
end for
Form the $n_u \times k$ prediction matrix H^S as in Section 2.3.2
Compute matrices W, D and L using the unlabeled target data D_u^T
Compute β^s by solving (7)
Learn the target classifier $\hat{h}$ by solving (14)

2.3.4. Theoretical Analysis. For convenience of presentation, we rewrite the empirical joint error function on (α, β)-weighted source domain and the target domain defined in (14) as follows:

$$\hat{E}^S_{\alpha,\beta}(h) = \mu\hat{\epsilon}_{\alpha,\beta}(h) + \hat{\epsilon}_T(h) = \mu\left(\sum_{s=1}^{k}\frac{\beta^s}{n_s}\sum_{i=1}^{n_s}\alpha_i^s\mathcal{L}(h(x_i^s), f_s(x_i^s))\right) + \sum_{i=1}^{n_l}\frac{1}{n_l}\mathcal{L}(h(x_i^0), f_0(x_i^0)) \tag{15}$$

where $y_i^s = f_s(x_i^s)$ and f_s is the labeling function for source s, $\mu > 0$, (x_i^0) are samples from the target, $y_i^t = f_0(x_i^0)$ and f_0 is the labeling function for the target domain, and $S = (x_i^s)$ include all samples from the target and source domains. The true (α, β)-weighted error $\epsilon_{\alpha,\beta}(h)$ on weighted source domain samples is defined analogously. Similarly, we define $E^S_{\alpha,\beta}(h)$ as the true joint error function. For notational simplicity, denote $n_0 = n_l$ as the number of labeled samples from the target, $m = \sum_{s=0}^{k} n_s$ as the total number of samples from both source and target, and $\gamma_s^i = \mu\beta^s\alpha_i^s/n_s$ for $s \geq 1$ and $\gamma_s^i = 1/n$ for $s = 0$. Then we can re-write the empirical joint error function in (15) as:

$$\hat{E}^S_{\alpha,\beta}(h) = \sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\mathcal{L}(h(x_i^s), f_s(x_i^s)).$$

Next, we bound the difference between the true joint error function $E^S_{\alpha,\beta}(h)$ and its empirical estimate $\hat{E}^S_{\alpha,\beta}(h)$ using the weighted Rademacher complexity measure [Bartlett and Mendelson 2002; Koltchinskii 2001] defined as follows:

DEFINITION 1. *(Weighted Rademacher Complexity) Let $\mathbb{H}$ be a set of real-valued functions defined over a set X. Given a sample $S \in X^m$, the empirical weighted Rademacher complexity of $\mathbb{H}$ is defined as follows:*

$$\hat{\Re}_S(H) = E_\sigma\left[\sup_{h\in\mathbb{H}}|\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s h(x_i^s)|\,\middle|\, S = (x_i^s)\right].$$

The expectation is taken over $\sigma = \{\sigma_i^s\}$ where $\{\sigma_i^s\}$ are independent uniform random variables taking values in $\{-1, +1\}$. The weighted Rademacher complexity of a hypothesis set $\mathbb{H}$ is defined as the expectation of $\hat{\Re}_S(H)$ over all samples of size m:

$$\Re_m(H) = E_S\left[\hat{\Re}_S(H)\,\middle|\, |S| = m\right].$$

Our main result is summarized in the following lemma, which involves the estimation of the Rademacher complexity of the following class of functions:

$$\mathbb{G} = \{x \mapsto \mathcal{L}(h'(x), h(x)) : h, h' \in \mathbb{H}\}.$$

LEMMA 1. *Let $\mathbb{H}$ be a family of functions taking values in $\{-1,+1\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for $h \in \mathbb{H}$:*

$$\left|E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right| \leq I\!R_S(\mathbb{H}) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}.$$

Furthermore, if $\mathbb{H}$ has a VC dimension of d, then the following holds with probability at least $1-\delta$:

$$\left|E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right| \leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\left(\sqrt{2d\log\frac{em}{d}}+1\right),$$

where e is the natural number.

The proof is provided in Section A of the supplemental material.

2.3.5. Error bound on target domain data . In the previous section we presented an upper bound on the difference between the true joint error function and its empirical estimate and established its relation to the weighting factors γ_i^s. Next we present our main theoretical result, i.e., an upper bound of the error function on target domain data, i.e., an upper bound of $\epsilon_T(\hat{h})$. We need the following definition of divergence for our main result:

DEFINITION 2. *For a hypothesis space $\mathcal{H}$, the symmetric difference hypothesis space $d_{\mathbb{H}\Delta\mathbb{H}}$ is the set of hypotheses*

$$g \in \mathbb{H}\Delta\mathbb{H} \Leftrightarrow g(x) = h(x) \oplus h^{'}(x)\ for\ some\ h, h^{'} \in \mathcal{H},$$

where $\oplus$ is the XOR function. In other words, every hypothesis $g \in \mathbb{H}\Delta\mathbb{H}$ is the set of disagreements between two hypotheses in $\mathcal{H}$.
The $\mathbb{H}\Delta\mathbb{H}$-divergence between any two distributions D_S and D_T is defined as

$$d_{\mathbb{H}\Delta\mathbb{H}}\left(D_S, D_T\right)) = 2\sup_{h,h'\in\mathbb{H}}\left|Pr_{x\backsim D_S}[h(x)\neq h'(x)] - Pr_{x\backsim D_T}[h(x)\neq h'(x)]\right|.$$

THEOREM 1. *Let $\hat{h} \in \mathbb{H}$ be an empirical minimizer of the joint error function on similarity weighted source domain and the target domain:*

$$\hat{h} = arg\min_{h\in\mathbb{H}}\hat{E}_{\alpha,\beta}(h) \equiv \mu\hat{\epsilon}_{\alpha,\beta}(h) + \hat{\epsilon}_T(h)$$

for fixed weights μ, α, and β and let $h_T^* = \min_{h\in\mathbb{H}}\epsilon_T(h)$ be a target error minimizer. Then for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:

$$\begin{aligned}\epsilon_T(\hat{h}) \;\leq\; & \epsilon_T(h_T^*) + \frac{2\Re_S(H)}{1+\mu} + \frac{2}{1+\mu}\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \\ & + \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right), \end{aligned} \tag{16}$$

if $\mathbb{H}$ has a VC dimension of d, then the following holds with probability at least $1-\delta$:

$$\begin{aligned}\epsilon_T(\hat{h}) \;\leq\; & \epsilon_T(h_T^*) + \frac{2}{1+\mu}\left(\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\left(\sqrt{2d\log\frac{em}{d}}+1\right)\right) \\ & + \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right), \end{aligned} \tag{17}$$

where $\lambda_{\alpha,\beta} = \min_{h\in\mathbb{H}}\{\epsilon_T(h) + \epsilon_{\alpha,\beta}(h)\}$*, and* $d_{\mathbb{H}\Delta\mathbb{H}}(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T))$ *is the symmetric difference hypothesis space for* (α, β)*-weighted source and target domain data.*

The proof as well as a comparison with the result in [Ben-David et al. 2010] is provided in the supplemental material.

We observe that μ and the divergence between the weighted source and target data play significant roles in the generalization bound. Our proposed two-stage weighting scheme aims to reduce the differences between the source and target domain hypothesis by learning the source domain hypothesis based on re-weighted instances. The re-weighted instances tend to have a distribution which is similar ti target domain both in marginal and conditional probability distributions. Next, we analyze the effect of μ. When $\mu = 0$, the bound reduces to the generalization bound using the n_l training samples in the target domain only. As μ increases, the effect of the source domain data increases. Specifically, when μ is larger than a certain value, for the bound in (17), as μ increases, the second term will reduce, while the last term capturing the divergence will increase. In the extreme case when $\mu = \infty$, the second term in (17) can be shown to be the generalization bound using the weighted samples in the source domain only (the target data will not be effective in this case), and the last term equals to $2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T)$. Thus, effective transfer is possible in this case only if the divergence is small. We also observed in our experiments that the target domain error of the learned joint hypothesis follows a bell shaped curve; it has a different optimal point for each dataset under certain similarity and divergence measures.

3. RELATED WORK

Most of the existing methods measure the similarity between a particular source and the target domain based on the similarity of their marginal probabilities. Shimodaira *et al.* [Shimodaira 2000] biased the training samples by their test-to-training ratio to match the marginal distribution of the test data. Sugiyama *et al.* [Sugiyama et al. 2008] proposed to reduce the gap in marginal probabilities by minimizing the KL-divergence between test and weighted training data and Bickel *et al.* [Bickel et al. 2009] discriminated training against test data with a probabilistic model that accounts for the marginal probability difference between training and test distribution. There are several other methods which are also based on marginal probability differences using Maximum Mean Discrepancy [Borgwardt et al. 2006] as a measure such as Kernel Mean Matching [Huang et al. 2007] and Transfer Component Analysis [Pan et al. 2008]. The proposed domain adaptation frameworks CP-MDA and 2SW-MDA, differ from all these methods in two ways: (1) they are predominantly based on conditional probability differences, with 2SW-MDA being marginal probability based as well and (2) they are based on multiple source domains.

Several algorithms have been developed in past to combine knowledge from multiple sources. Luo *et al.* used consensus maximization as the basis of combining multiple source data [Luo et al. 2008]. Mansour *et al.* based the transferability of knowledge on a distribution weighted combination of the hypothesis generated by the independent sources [Mansour et al. 2009b]. The theoretical proof of both frameworks are based on strong assumptions on the predictive power of the individual source domains on the target domain data. In [Shi et al. 2009], a clustering based knowledge transfer was proposed for applications with different class labels across source and target domains, unlike the application addressed in this paper.

The proposed frameworks CP-MDA and 2SW-MDA are related to two multi-source domain adaptation frameworks including Domain Adaptation Machine (DAM) [Duan et al. 2009] and Locally Weighted Ensemble (LWE) [Gao et al. 2008]. The proposed framework differs from DAM in the way the weights are computed for different auxiliary sources. In DAM, the weight assigned to each auxiliary source is obtained by measuring the marginal probability distribution difference between the target domain and the particular auxiliary source only, using an empirical estimate of the difference based on the Maximum Mean Discrepancy measure [Borgwardt et al. 2006]. The proposed frameworks however computes weights for the auxiliary source data considering predominantly conditional probability distribution of the target data. The weights for all sources are computed in

a joint optimization framework, which takes the interaction among multiple auxiliary sources into account.

The proposed frameworks differ from LWE [Gao et al. 2008] in that in LWE, the label y of an unlabeled target domain data x is computed using a local weighting ensemble (LWE) scheme:

$$P(y|x) = \sum_{i=1}^{k} w_{M_i,x} P(y|M_i, x) \tag{18}$$

where $P(y|M_i, x)$ is the prediction made by one of the k models M_i for target data point x and $w_{M_i,x}$ is the weight of the model M_i at point x computed by comparing the similarity graphs of the source and target data around point x. Different from the proposed weighting scheme where we compute all weights in a joint framework, the weight for each auxiliary classifier is computed independently [Gao et al. 2008].

We also compare our frameworks with representative single-source domain adaptation algorithms such as Kernel Mean Matching (KMM) proposed by Huang *et al.* [Huang et al. 2007], Transfer Component Analysis (TCA) proposed by Pan *et al.* [Pan et al. 2009] and KMapEnsemble (KE) proposed by Zhong *et al.* [Zhong et al. 2009]. KMM re-weights the samples in the source domain so as to minimize the marginal probability difference between the source and target domain using Maximum Mean Discrepancy (MMD) as the measure. TCA is based on feature mapping so as to reduce the marginal probability differences between the source and target distributions again using MMD as the measure. KE differs from the first two algorithms, in which it addresses the conditional probability differences by sample selection after performing a feature mapping step to reduce the marginal probability differences.

There was some classification work dealing with physiological signals using neural networks [leon et al. 2007] and linear discriminant analysis [Kim and Andre 2008]; they achieved moderate generalization performance across subjects. To the best of our knowledge we report the first systematic empirical analysis of domain adaptation methods to address the distribution differences due to the subject based variability in physiological signals.

4. EXPERIMENTS

The proposed algorithms have been evaluated on multi-dimensional feature vectors extracted from SEMG (Surface electromyogram) signals collected from 8 subjects during a fatiguing exercise.

4.1. Experimental Setup

4.1.1. SEMG data. The SEMG data was collected during a repetitive gripping action performed by the forearm. Figure 3 shows the subject with surface EMG differential electrodes on the extensor carpi radialis muscle to record the SEMG signal. The subject performs a cycle of flexion-extension of forearm as shown in Figure 3 at two different speeds, i.e., low speed (1 cycles/sec) and high speed (2 cycles/sec) repetitively for about 4 minutes. The cycles of low and high speed are alternated after every minute to form four phases (or classes) as discussed in the introduction.

The raw SEMG activity was recorded by Grass Model 8-16C at 1000Hz and passed through a band pass filter of 20Hz to 500Hz. The data was collected and saved by the LabView software (from National Instruments) running on a PC. Data of the order of 1.92 Million samples ($1000 \times 4 \times 60 \times 8$), was collected from 8 subjects including male and female of the age group of 25 years to 45 years. A set of twelve amplitude and frequency domain features including mean frequency, median frequency, spectral energy, spectral entropy, root mean square, number of zero crossings, to mention a few are derived from running windows of 1000 time samples with 50% overlap [Knaflitz and Bonato 1999].

Each subject data consists of around 280 to 400 samples of 12 dimensional feature vectors, belonging to four classes with around 70 to 100 samples per class (some subjects who got fatigued sooner and hence could not maintain the required uniform speed for 1 minute the time period was reduced to 30 to 45 secs per phase, hence the number of samples varies between different subjects).

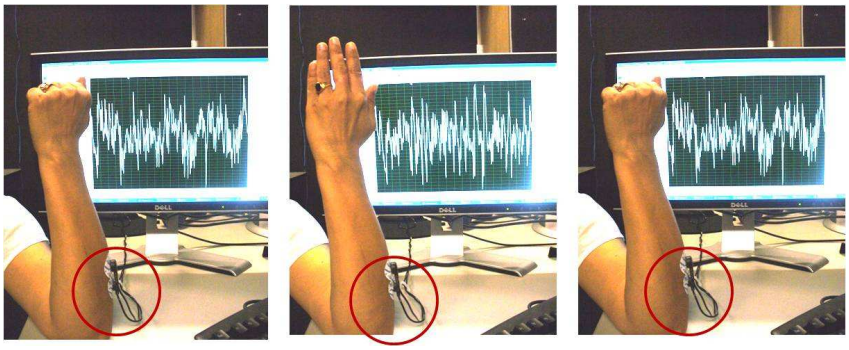


Fig. 3. SEMG data collection during a repetitive gripping activity

4.1.2. Experimental Procedure. To evaluate the effectiveness of the proposed methods, we compare the results with four baseline methods, including SVM-C, SVM-M, SMA, and TSVM (Transductive SVM), and two recently proposed multi-source learning methods, including Locally Weighted Ensemble (LWE) [Gao et al. 2008] and Domain Adaptation Machine (DAM) [Duan et al. 2009].

SVM-C refers to *all but one* method where the training data comprises of data from seven subjects and the test data is the data from the remaining subject. SVM-M, refers to the *majority voting* based ensemble framework. The class y assigned to each unlabeled test data x is $\max_y NV(y|x)$ where $NV(y|x)$ is the number of votes given for class y for a particular test sample x by the seven auxiliary sources. SMA refers to *simple model averaging*, which provides equal weight to all the classifiers learned on each auxiliary source domain in an weighted ensemble framework used to generate the label for the target domain data. TSVM refers to Transductive SVM [TSV] implemented in the svmlight package. It is a semi-supervised method where the training data consists of labeled data from all seven subjects from the source domain and unlabeled data from the target subject.

We vary the number of labeled samples per class in the target domain. DAM(1) and DAM(7) refer to the DAM framework with 1 and 7 target domain labeled samples per class respectively. The proposed CP-MDA method is also implemented using 1 and 7 labeled data from target domain, referred as CP-MDA(1) and CP-MDA(7) respectively. For both cases the unlabeled data from the target domain, is fixed at 10% of the target domain data. The weights of the auxiliary sources computed by the proposed method are also based on this 10% unlabeled target domain data. The rest of the target domain data is treated as unseen target domain data. All the methods are tested on the same pool of *unseen* unlabeled target domain data. The accuracies are computed in a subject independent manner.

We mention here briefly some of the parameters used in implementing the existing and the proposed methods. The values of γ_A and γ_I were kept as 0.014 and 0.01 respectively, as suggested in [Belkin et al. 2006]. The Laplacian graph matrix used in calculating the weights was set as 'binary' type based on the N nearest neighbors with $N = 10$. The value of Θ was estimated via 5-fold cross validation on the set $\{i10^{-2}|i = 0, 1, \cdots, 100\}$.

5. RESULTS AND ANALYSIS

We first present the comparative performance of CP-MDA and then compare the performance of CP-MDA with 2SW-MDA, followed by a discussion on the relative performance of the two proposed methods.

5.1. Comparative Performance of CP-MDA

We compare different methods including SVM-C, SVM-M, SMA, TSVM, LWE, DAM and the proposed CP-MDA. The results are summarized in Table II. The first column of the table indicates the subject data under test (target domain). The training data (source domain) consists of the data from

Table II. Comparative performance of CP-MDA on SEMG data - Accuracy (%)

Test Sub	SVM-C	SVM-M	SMA	TSVM	LWE	DAM(1)	DAM(7)	CP-MDA(1)	CP-MDA(7)
1	70.76	33.9	44.96	49.09	67.44	74.83	77.43	81.93	85.25
2	43.69	50.76	44.61	55.68	77.54	81.36	83.35	84.73	87.7
3	50.11	56.85	56.84	65.09	75.55	74.77	78.99	82.45	85.06
4	59.65	47.93	49.67	56.98	81.22	80.63	84.32	81.27	86.4
5	40.37	44.79	50.15	62.5	52.48	76.74	81.14	80.74	86.62
6	59.21	61.45	60.33	71.32	65.77	59.21	74.28	83.12	88.09
7	47.13	46.91	45.76	60.73	60.32	74.27	83.31	81.57	86.4
8	69.85	64.53	74.46	68.55	72.81	84.55	86.6	88.5	90.56
Average	55.09	50.85	53.34	61.24	69.14	75.79	81.18	**83.04**	**87.01**

Table III. Comparison of SVM-T, DAM, and CP-MDA on Subject 6 (top) and Subject 7 (bottom) in terms of accuracy (%) when the number of labeled target domain data per class varies.

	Number of labeled data per class					
Method	1	2	3	4	6	7
SVM-T	4.26	4.26	49.09	73.63	84.69	85.67
DAM	59.12	59.21	59.35	59.59	65.03	74.28
CP-MDA	**83.12**	83.12	85.45	87.58	87.77	88.09
SVM-T	10.59	45.5	77.79	83.25	85.48	84.97
DAM	74.27	75.11	79.10	81.19	82.43	83.3
CP-MDA	**81.57**	83.99	85.81	86.24	86.32	86.4

the remaining seven subjects. Similar to the results obtained in the case of synthetic data we see that SVM-C, SVM-M, SMA, and TSVM perform very poorly. We observe significant improvement in classification accuracy when domain adaptation methodologies are employed. The proposed method CP-MDA(1) provides a 20% to 30% improvement over the baseline methods including SVM-C, SVM-M, SMA and TSVM. The classification accuracies of the proposed method are in average 13% higher than LWE. It is also observed as in the case of synthetic data that CP-MDA(1) performs not only better than DAM(1) but also better than DAM(7) in 5 out of 8 cases. These results verify the effectiveness of the proposed method.

Next, we evaluate the performance of CP-MDA when the number of labeled target domain data varies. We compare CP-MDA with DAM and SVM-T. SVM-T refers to an SVM classifier trained on the labeled target domain data. The results for two subjects are summarized in Table III; we obtain similar results for the other six subjects and the results are omitted. We can observe from the table that when the number of labeled target domain data per class is small, e.g., 1 to 4 samples per class, both domain adaptation methods perform much better than SVM-T. But with an increasing number of labeled data from the target domain the accuracies become comparable. However the proposed method always performs better than the other two methods. This result demonstrates that domain adaption is especially useful when the amount of labeled target domain data is small.

We also compare the performance of the weighting schemes used in LWE, DAM and CP-MDA. Table IV summarizes the results for different test cases. We observe that CP-MDA-WE performs better than the other methods in 6 out of 8 cases, and LWE performs better in the remaining 2 cases. Recall that like CP-MDA-WE, LWE computes weights for the auxiliary source domain based on the conditional probability differences between the source and target domains, while MMD-WE computes the weights based on the marginal probability differences only. Since SEMG data has significant conditional probability differences, CP-MDA-WE and LWE are expected to outperform DAM-WE.

The proposed algorithm CP-MDA computes weights for each class for each of the auxiliary source domain data, thus exploiting the similarities and dissimilarities at the class level. Table V shows the weights for four different classes assigned to each training subject in the source domain for test subject 1 in the target domain. We observe that the proposed weighting scheme assigns different weights to different auxiliary source domain data (subject data) for different classes. Subjects 5 and 8 get higher weights for class 1, subject 7 gets a higher weight for class 2, subject 5 gets a

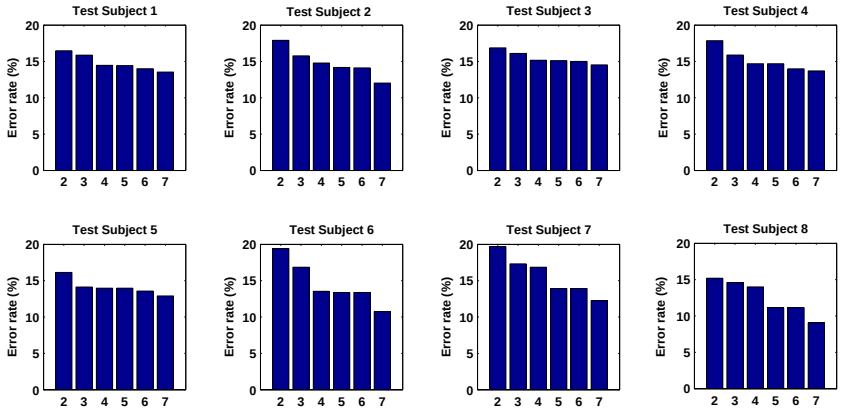


Fig. 4. The effect of the number of auxiliary source domains (horizontal axis) in the training set on the proposed CP-MDA algorithm in terms of the classification error rates (%) for all eight subjects.

higher weight for class 3, and for class 4 subject 4 gets a higher weight. We observe from Figure 1 that the data distribution of class 4 of subject 1 is very similar to that of class 4 of subject 4.

One of the key advantage of the proposed algorithm is that it exploits the information from multiple source domains for classifying the target data. It will be interesting to study how the number of sources used in the training set affects the classification. Figure 4 presents the error rates obtained for each test subject when the number of subjects in the source domain varies (from 1 to 7); we simply keep adding subjects to the training set in increasing order of the subject number. We observe that for all subjects, the error rate decreases monotonically when the number of subjects increases. These results demonstrate the effectiveness of the proposed algorithm for extracting useful information from multiple sources.

To evaluate the benefit of a multi-source domain adaption framework for addressing subject based variability, we compare the proposed algorithm with three representative single-source domain adaption algorithms Table VI summarizes the classification accuracies obtained by different

Table IV. Comparison of different weighting schemes for different test subjects - Accuracy (%).

Test Sub	LWE	MMD-WE	CP-MDA-WE
1	67.44	68.27	75.12
2	77.54	69.48	83.23
3	75.55	71.84	75.68
4	81.22	62.65	81.09
5	52.48	68.32	78.16
6	65.77	58.91	76.11
7	60.32	67.75	75.07
8	72.81	66.11	78.71
Average	69.14%	66.66%	**77.89**%

Table V. Weights computed by CP-MDA for four different classes for each of the source domain subjects 2-8 for test target subject 1.

Class	Target subject						
	2	3	4	5	6	7	8
1	0	0	0.02	0.50	0.48	0	0
2	0	0.01	0.03	0	0.11	0.74	0.11
3	0	0.02	0.12	0.75	0	0.01	0.11
4	0.09	0.02	0.66	0.11	0.11	0.01	0

Table VI. Comparison of CP-MDA with three single source domain adaptation algorithms (KMM, TCA, and KE) - Accuracy(%).

Test Sub	CP-MDA(7)	KMM	TCA	KE
1	85.25	65.15	45.15	71.85
2	87.7	46.96	68.93	74.62
3	85.06	59.55	56.78	74.79
4	86.4	73.38	52.68	69.35
5	86.62	45.31	60.15	73.44
6	88.09	70.62	76.92	83.92
7	86.4	51.13	55.64	77.97
8	90.56	42.79	67.24	79.48
Average	**87.01**	56.86	53.84	75.67

methods for each of the test subjects. The target data is from the subject shown in column 1 and the source data consists of the combined data from the remaining seven subject. Classification results were averaged over 10 runs with different sets of randomly selected 7 labeled samples per class from the target domain data. We can observe from the table that combining all the subject data and forming a single domain degrades the performance. We also observe that among the three single domain adaption algorithms, KMM [Huang et al. 2007] or TCA [Pan et al. 2008] which consider the marginal probability differences only perform worse than KE [Zhong et al. 2009]. These results are expected as SEMG data has significant conditional probability differences. Our results demonstrate the effectiveness of the proposed multi-domain framework for dealing with subject based variability in SEMG data.

5.2. Comparative Performance of CP-MDA and 2SW-MDA

Table VII. Comparison of different methods on SEMG dataset - Accuracy(%) (%).

Test Sub	SVM-C	LWE	KE	KMM	TCA	DAM	CP-MDA	2SW-MDA
1	70.76%	67.44%	63.55%	64.94%	66.35%	74.83%	81.93%	83.03%
2	43.69%	77.54%	74.62%	63.63%	59.94%	81.36%	84.73%	87.96%
3	50.11%	75.55%	62.50%	64.06%	56.78%	74.77%	82.54%	88.96%
4	59.65%	81.22%	69.35%	52.68%	73.38%	80.63%	81.27%	88.49%
5	40.37%	52.48%	65.61%	49.77%	57.48%	76.74%	80.74%	86.14%
6	59.21%	65.77%	83.92%	70.62%	76.92%	59.21%	83.12%	87.10%
7	47.13%	60.32%	77.97%	51.13%	55.64%	74.27%	81.57%	87.08%
8	69.85%	72.81%	79.48%	67.24%	42.79%	84.55%	88.50%	93.01%
Toy data	60.05%	75.63%	81.40%	68.01%	64.97%	84.27%	93.21%	98.54%

Comparative Studies. Table VII shows the classification accuracies of different methods on the SEMG and the toy datasets. We observe that SVM-C performs poorly for all cases. This may be attributed to the distribution difference among the multiple source and target domains. The physiological signals, such as SEMG are predominantly different in conditional probability distributions due to the high subject based variability in the power spectrum of these signals and their variations as fatigue sets in [Contessa et al. 2009; Georgakis et al. 2003; Gerdle et al. 2000]. We observe that the proposed CP-MDA and 2SW-MDA methods outperform other domain adaptation methods and achieve higher classification accuracies in most cases. However 2SW-MDA performs better than CP-MDA, this can be attributed to the fact that 2SW-MDA addresses both marginal and conditional probability differences, where as CP-MDA addresses only conditional probability differences. Also the conditional probability based weights are computed with source hypothesis learned on re-weighted source instances (as per marginal probability differences) , thus increasing the accuracy of computed weights. An average unsupervised classification accuracy of 81.56% was obtained using re-weighted instances versus 77.89% obtained using source instances without re-weighting (Table IV).

The accuracies of an SVM classifier, on the toy dataset, when learned only on the source domains D1, D2 individually and on the combined source domains, are 60.67% and 71.84% and 60.05%

respectively, while 2SW-MDA achieves an accuracy of 98.54%. More results are provided in the appendix.

6. CONCLUSIONS

Domain adaptation is an important problem that arises in a variety of modern applications where limited or no labeled data is available for a target application. We presented here two novel multi-source domain adaptation frameworks. The proposed frameworks are based on a weighting scheme that computes the weights of each source in a joint optimization framework. CP-MDA predominantly addresses conditional probability differences between the domains, where as 2SW-MDA follows a two-step procedure in order to reduce both marginal and conditional probability distribution differences between the source and target domain. Besides, 2SW-MDA is based on instance re-weighting, where as CP-MDA is based on hypothesis weighting. The psuedo labels in CP-MDA are computed based on hypothesis combination. Both the proposed methods perform better than state-of-the-art single and multi-source domain adaptation methods on all three datasets. We also presented a theoretical error bound on the target classifier learned on re-weighted data samples from multiple sources as in 2SW-MDA framework. Empirical comparisons with existing state-of-the-art domain adaptation methods demonstrate the effectiveness of the proposed approach. As a part of the future work we plan to extend the proposed multi-source framework to applications involving other types of physiological signals for developing generalized models across subjects for emotion and health monitoring [leon et al. 2007; Kim and Andre 2008]. We would also like to extend our framework to video and speech based applications, which are commonly affected by distribution differences [Duan et al. 2009].

7. ACKNOWLEDGEMENTS

We acknowledge Dr. Arthur Gretton, Dr. Yishay Mansour, Dr. Lixin Duan, Dr. Jing Gao, Dr. Erheng Zhong, Dr. Luo Ping, Dr. Sinno Jialin Pan, and Dr. Masashi Sugiyama for providing codes and other supporting materials. This research is sponsored in part by NSF IIS-0953662, CCF-1025177, and ONR N00014-11-1-0108.

APPENDIX

A: Proof of Lemma 1

PROOF. Define $\Phi(S) = \sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)$. Changing the i-th point in the s-th source affects $\Phi(S)$ by at most $\gamma^s_i = \mu\beta^s\alpha^s_i$, while changing a point in the target affects $\Phi(S)$ by at most $\gamma^s_i = 1/n$ ($s = 0$). Applying McDiarmid's inequality [McDiarmid 1989] to $\Phi(S)$, the following holds with probability at least $1 - \delta/2$:

$$\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, using standard techniques used in [Bartlett and Mendelson 2002], we bound the expectation as follows:

$$\begin{aligned}
E_S[\Phi(S)] &= E_S\left[\sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_S\left[\sup_{h\in\mathbb{H}} E_{\bar{S}}[\hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)]\right] \\
&\leq E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&= E_{\sigma,S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&\leq 2E_{\sigma,S}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\mathcal{L}(h(x^s_i), f_s(x^s_i))\right] \leq 2\Re_S(G) = \Re_S(H),
\end{aligned}$$

where the last step follows from the standard techniques for relating the Rademacher complexities [Kakade and Tewari 2008], and $\mathbb{G}$ is a class of functions given by:

$$\mathbb{G} = \{x \mapsto \mathcal{L}(h'(x), h(x)) : h, h' \in \mathbb{H}\}.$$

Thus, for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$E^S_{\alpha,\beta}(h) \leq \hat{E}^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Similarly, by defining $\Phi'(S) = \sup_{h\in\mathbb{H}} \hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)$ and bounding the expectation of $\Phi'(S)$, we can show that for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$\hat{E}^S_{\alpha,\beta}(h) \leq E^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Thus, with probability at least $1 - \delta$:

$$\left|\hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)\right| \leq \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, we bound $\Re_S(H)$ as follows [Kakade and Tewari 2008]:

$$\begin{aligned}
\Re_S(H) &= E_{S,\sigma}\left[\sup_{h\in\mathbb{H}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s h(x_i^s)| \,\middle|\, S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s| \,\middle|\, S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s| \,\middle|\, S=(x_i^s)\right] \\
&\leq E_S\left[\max_{u\in\mathbb{H}_{|S}} ||u||\sqrt{2\log|\mathbb{H}_{|S}|}\right] \text{ (Massart's Lemma [Massart 2000])} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} E_S\left[\sqrt{2\log|\mathbb{H}_{|S}|}\right] \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \sqrt{2\log|\prod_{\mathbb{H}}(m)|} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \sqrt{2d\log\frac{em}{d}},
\end{aligned}$$

where $\mathbb{H}_{|S}$ is the restriction of $\mathbb{H}$ on S, $\prod_{\mathbb{H}}(m)$ is the growth function for $\mathbb{H}$ given by the maximum number of ways m points can be classified by $\mathbb{H}$, and e is the natural number. □

B: Proof of Theorem 1

PROOF. Let $h^* = \arg\min_{h\in\mathbb{H}}\{\epsilon_T(h) + \epsilon_{\alpha,\beta}(h)\}$. By the triangle inequality, we have

$$\begin{aligned}
|\epsilon_{\alpha,\beta}(h) - \epsilon_T(h)| &\leq |\epsilon_{\alpha,\beta}(h) - \epsilon_{\alpha,\beta}(h,h^*)| + |\epsilon_{\alpha,\beta}(h,h^*) - \epsilon_T(h,h^*)| + |\epsilon_T(h,h^*) - \epsilon_T(h)| \\
&\leq \epsilon_{\alpha,\beta}(h^*) + |\epsilon_{\alpha,\beta}(h,h^*) - \epsilon_T(h,h^*)| + \epsilon_T(h^*) \\
&\leq \lambda_{\alpha,\beta} + \frac{1}{2} d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right).
\end{aligned}$$

Next, we bound $(1+\mu)\epsilon_T(\hat{h})$ as follows:

$$
\begin{aligned}
& (1+\mu)\epsilon_T(\hat{h}) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(\hat{h}) + \epsilon_T(\hat{h}) + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(\hat{h}) + \hat{\epsilon}_T(\hat{h}) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(h_T^*) + \hat{\epsilon}_T(h_T^*) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(h_T^*) + \epsilon_T(h_T^*) + 2\Re_S(H) + 2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \\
\leq\ & (\mu+1)\epsilon_T(h_T^*) + 2\Re_S(H) + 2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right)
\end{aligned}
$$

Thus,

$$
\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + \frac{2\Re_S(H)}{1+\mu} + \frac{2}{1+\mu}\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \tag{19}
$$

□

Note that our proof follows a similar procedure in [Ben-David et al. 2010]. The main differences include (1) we employ the weighted Rademacher complexity, which provides a tighter bound than the one in [Ben-David et al. 2010] based on the VC dimension; (2) the empirical minimizer $\hat{h}$ of our joint error function includes two terms involving both source and target domain data with a differential weight μ, while the one in [Ben-David et al. 2010] involves one term only. For the special case when $\mu = 1$ and α_i^s's are given a uniform weight, i.e., $\alpha_i^s = 1/n_s$, our bound in (17) is strictly tighter than the one in [Ben-David et al. 2010] (due to the $1/2$ factor in the last term). In the general case with different choices of μ and α_i^s's, our bound can be further improved.

C: More details on parameters used for the implementation of different methods

A Gaussian kernel with $\sigma = 10$ was used to compute the α values for each source. The weighted hypothesis for each source was learned using Support Vector Machines implemented in the LibSVM package, with a linear kernel and a regularization penalty $C = 10$. The β weights were computed based on a binary similarity matrix, i.e., $W_{ij} = 0$ if the i-th data point is among the N nearest neighbors of the j-th data point or the j-th data point is among the N nearest neighbors of the i-th data point; we set $N = 10$. We implemented TCA with a linear kernel and KMM with a Gaussian kernel as they gave the best results. All parameters were tuned using 10-fold cross-validation.

D: Additional empirical results

Figure 5 shows the α-weighted data samples in both source domain D1 and source domain D2 of the toy data shown in Figure 2. We observe that data samples having similar marginal probabilities in both the domains get higher weight, shown by the size of the points. The size of the points are proportional to their weights. We also observe that since at this stage the source data is re-weighted based only on marginal probability distribution difference, hence some of the data samples from

source domain D1 having conflicting conditional probabilities with target domain data also get higher weight as they share similar marginal probability distributions.

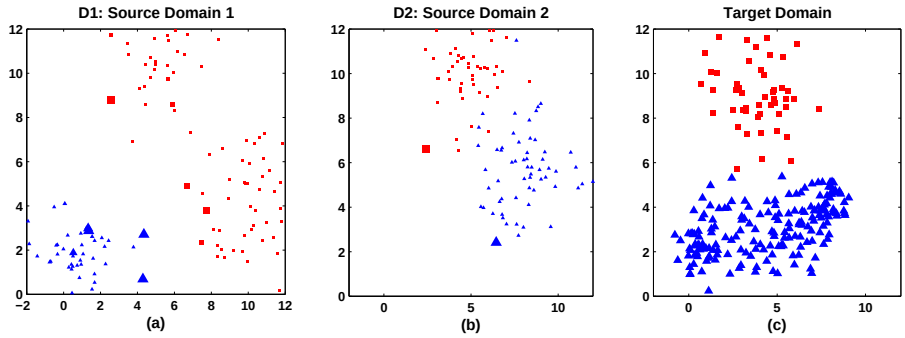


Fig. 5. Data samples in source domains D1 and D2 re-weighted by α_i^s. We can observe that points from source domain D1 also get large weights due to the similarity in marginal probabilities (the size of a point is proportional to its weight).

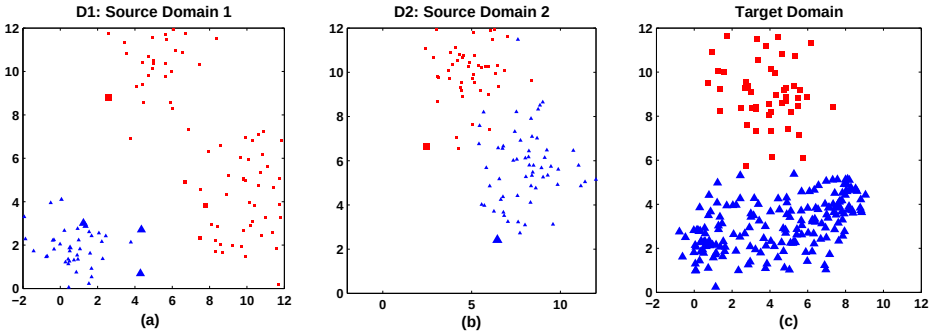


Fig. 6. Data samples in the source domains D1 and D2 re-weighted by both α_i^s and β^s. We observe that the points with conflicting conditional probabilities get moderated by β^s (the size of a point is proportional to its weight).

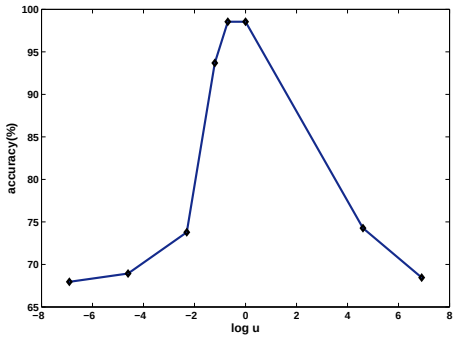


Fig. 7. Performance of proposed 2SW-MDA method on the toy dataset shown in Figure 2 with varying μ - Accuracy (%).

Figure 6 shows the results of applying β-weights to the data samples in both source domain D1 and source domain D2 of the toy data. We observe that the data samples in source domain D1 with conflicting conditional probabilities get reduced when moderated with β weights, as source domain D2 is more similar to target data in conditional probability distribution than the source domain D1.

Figure 7 shows the performance of 2SW-MDA on toy dataset shown in Figure 2 with varying μ. The result is consistent with the theoretical result established in this paper.

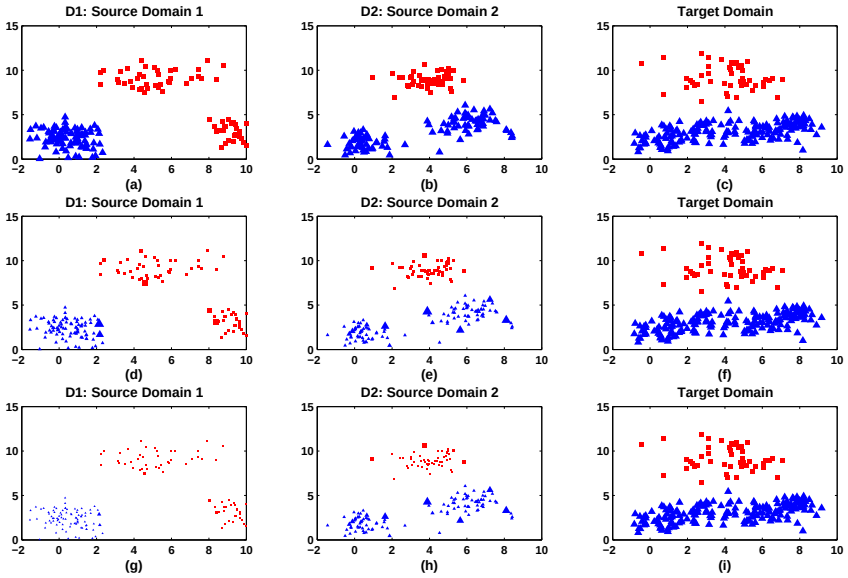


Fig. 8. Results on another toy dataset: First row shows the original distribution of two source domains D1 and D2 and a target domain. The second and third rows show the results of applying α and β weights, respectively. We observe that source domain data samples with similar marginal and conditional probabilities get higher weight. The β values for D1 and D2 are 0.17 and 0.83 respectively, individual accuracies being 61.65% and 89.51% and proposed method gives 98.51%.

Figure 8 shows the results of applying the proposed 2SW-MDA method on another set of toy dataset consisting of two source domains and a target domain with different marginal and conditional probability differences. We observe that the distribution D1 which has conflicting conditional probabilities with target domain data gets under-weighted by the proposed weighting scheme and hence transfer happens mostly from the source distribution D2, which shares similar marginal and conditional probability differences with the target domain. We get β value of 0.17 for D1 and 0.83 for D2, individual accuracies being 61.65% and 89.51% and proposed method gives 98.51%.

REFERENCES

http://svmlight.joachims.org/.

BARTLETT, P. AND MENDELSON, S. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR 3*, 463–482.

BELKIN, M., NIYOGI, P., AND SINDHWANI, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR 7*, 2399–2434.

BEN-DAVID, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F., AND VAUGHAN, J. 2010. A theory of learning from different domains. *Journal of Mach Learn 79*, 151–175.

BICKEL, S., BRÜCKNER, M., AND SCHEFFER, T. 2009. Discriminative learning under covariate shift. *JMLR 10*, 2137–2155.

BLITZER, J., M.DREDZE, AND PEREIRA, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

BORGWARDT, K., GRETTON, A., RASCH, M., KRIEGEL, H., SCHÖLKOPF, AND SMOLA, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics 22,* 14, 49–57.

CHATTOPADHYAY, R., GAURAV, P., AND SETHURAMAN, P. 2009. A generalized machine learning framework for continuous monitoring of physiological conditions based on fatigue and intensity of activity in daily living. In *WIML, NIPS*.

CHATTOPADHYAY, R., SETHURAMAN, P., AND GAURAV, P. 2010. Towards fatigue and intensity measurement framework during continuous repetitive activities. In *I2MTC*.

CONTESSA, P., ADAM, A., AND LUCA, C. J. D. 2009. Motor unit control and force fluctuation during fatigue. *Journal of Applied Physiology*.

DAUMÉ III, H. 2007. Frustratingly easy domain adaptation. In *ACL*.

DUAN, L., TSANG, I., XU, D., AND MAYBANK, S. 2009. Domain transfer svm for video concept detection. In *CVPR*.

DUAN, L., TSANG, I. W., XU, D., AND CHUA, T. S. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*. 289–296.

GAO, J., FAN, W., JIANG, J., AND HAN, J. 2008. Knowledge transfer via multiple model local structure mapping. In *KDD*. 283–291.

GEORGAKIS, A., STERGIOULAS, L., AND GIAKAS, G. 2003. Fatigue analysis of the surface EMG signal in isometric constant force contractions using the averaged instantaneous frequency. *Biomedical Engineering, IEEE Transactions on 50,* 2, 262–265.

GERDLE, B., LARSSON, B., AND KARLSSON, S. 2000. Criterion validation of surface EMG variables as fatigue indicators using peak torque: a study of repetitive maximum isokinetic knee extensions. *Journal of Electromyography and Kinesiology 10,* 4, 225–232.

HIGGS, P. 1992. Upper extremity impairment in workers performing repetitive tasks. *Plastic and reconstructive surgery 90*, 614.

HUANG, J., SMOLA, A. J., GRETTON, A., BORGWARDT, K., AND SCHÖLKOPF, B. 2007. Correcting sample selection bias by unlabeled data. In *NIPS*.

KAKADE, S. AND TEWARI, A. 2008. Lecture notes of CMSC 35900: Learning theory, Toyota Technological Institute at Chicago.

KIM, J. AND ANDRE, E. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30,* 12, 2067–2083.

KNAFLITZ, M. AND BONATO, P. 1999. Time-frequency methods applied to muscle fatigue assessment during dynamic contractions. *9,* 5, 337–350.

KOLTCHINSKII, V. 2001. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory 47,* 5, 1902–1914.

LEON, E., CLARKE, G., CALLAGHAN, V., AND SEPULVEDA, F. 2007. A user independent real time emotion recognition system for software agents in domestic environment. *Engineering Applications of Artificial Intelligence 20,* 3, 337–345.

LUO, P., ZHUANG, F., XIONG, H., XIONG, Y., AND HE, Q. 2008. Transfer learning from multiple source domains via consensus regularization. In *CIKM*.

MANSOUR, Y., MOHRI, M., AND ROSTAMIZADEH, A. 2009a. Domain adaptation : Learning bounds and algorithms. *abs/0902.3430*.

MANSOUR, Y., MOHRI, M., AND ROSTAMIZADEH, A. 2009b. Domain adaptation with multiple sources. In *NIPS*.

MASSART, P. 2000. Some applications of concentration inequalities to statistics. *Annales de la Faculte des sciences de ToulouseSciences de Toulouse IX,* 2, 245–303.

MCDIARMID, C. 1989. *On the method of bounded differences.* Vol. 5. Cambridge University Press, Cambridge.

PAN, S. J., KWOK, J. T., AND YANG, Q. 2008. Transfer learning via dimensionality reduction. In *AAAI*.

PAN, S. J., TSANG, I. W., KWOK, J. T., AND YANG, Q. 2009. Domain adaptation via transfer component analysis. In *IJCAI*.

PAN, S. J. AND YANG, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.

SCHLKOPF, B. AND SMOLA, A. J. 2002. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press.

SHI, X., FAN, W., YANG, Q., AND REN, J. 2009. Relaxed transfer of different classes via spectral partition. In *KDD*.

SHIMODAIRA, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. In *JSPI*.

SILVERSTEIN, B., FINE, L., AND ARMSTRONG, T. 1986. Hand wrist cumulative trauma disorders in industry [1351-0711]. *Occupational and Environmental Medicine 43*, 779.

STEINWART, I. 2001. On the influence of the kernel on the consistency of support vector machines. *JMLR 2*, 67–93.

SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. V., AND KAWANABE, M. 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*.

YOUNG, V. L., SEATON, M. K., FEELY, C. A., ARFKEN, C., EDWARDS, D. F., BAUM, C. M., AND LOGAN, S. 1995. Detecting cumulative trauma disorders in workers performing repetitive tasks. In *AJIM*.

ZHONG, E., FAN, W., PENG, J., ZHANG, K., REN, J., TURAGA, D., AND VERSCHEURE, O. 2009. Cross domain distribution adaptation via kernel mapping. In *KDD*.