Personalizing EEG-based Affective Models with Transfer Learning

Wei-Long Zheng[1] and **Bao-Liang Lu**[1,2,3,*]
[1]Center for Brain-like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2]Key Laboratory of Shanghai Education Commission for
Intelligent Interaction and Cognitive Engineering
[3]Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China
{weilong, bllu}@sjtu.edu.cn

Abstract

Individual differences across subjects and non-stationary characteristic of electroencephalography (EEG) limit the generalization of affective brain-computer interfaces in real-world applications. On the other hand, it is very time consuming and expensive to acquire a large number of subject-specific labeled data for learning subject-specific models. In this paper, we propose to build personalized EEG-based affective models without labeled target data using transfer learning techniques. We mainly explore two types of subject-to-subject transfer approaches. One is to exploit shared structure underlying source domain (source subject) and target domain (target subject). The other is to train multiple individual classifiers on source subjects and transfer knowledge about classifier parameters to target subjects, and its aim is to learn a regression function that maps the relationship between feature distribution and classifier parameters. We compare the performance of five different approaches on an EEG dataset for constructing an affective model with three affective states: positive, neutral, and negative. The experimental results demonstrate that our proposed subject transfer framework achieves the mean accuracy of 76.31% in comparison with a conventional generic classifier with 56.73% in average.

1 Introduction

Affective brain-computer interfaces (aBCIs) [Mühl *et al.*, 2014a] introduce affective factors into conventional brain-computer interfaces [Chung *et al.*, 2011]. aBCIs provide relevant context information about users affective states in brain-computer interfaces (BCIs) [Zander and Jatzev, 2012], which can help BCI systems react adaptively according to users' affective states, rather in a rule-based fashion. It is an efficient way to enhance current BCI systems with an increase of information flow, while at the same time without additional cost. Therefore, aBCIs have attracted increasing interests in both research and industry communities, and various studies have presented their efficiency and feasibility [Mühl *et al.*, 2014a; 2014b; Jenke *et al.*, 2014; Eaton *et al.*, 2015]. Although the large progress has been obtained about development of aBCIs in recent years, there still exist some challenges such as the adaptation to changing environments and individual differences.

Until now, most of studies emphasized choices of features and classifiers [Singh *et al.*, 2007; Jenke *et al.*, 2014; Abraham *et al.*, 2014] and neglected to consider individual differences in target persons. They focus on training subject-specific affective models. However, these approaches are practically infeasible in real-world scenarios since they need to collect a large number of labeled data. In addition, the calibration phase is time-consuming and annoying. An intuitive and straightforward way to dealing with this problem is to train a generic classifier on the collected data from a group of subjects and then make inference on the unseen data from a new subject. However, the existing studies indicated that following this way, the performance of generic classifiers were dramatically degraded due to the structural and functional variability between subjects as well as the non-stationary nature of EEG signals [Samek *et al.*, 2013; Morioka *et al.*, 2015]. Technically, this issue refers to the covariate-shift challenges [Sugiyama *et al.*, 2007]. The alternative way to dealing with this problem is to personalize a generic classifier for target subjects in an unsupervised fashion with knowledge transfer from the existing labeled data in hand.

The problem mentioned above has motivated many researchers from different fields in developing transfer learning and domain adaptation algorithms [Duan *et al.*, 2009; Pan and Yang, 2010; Chu *et al.*, 2013]. Transfer learning methods try to transfer knowledge from source domain to target domain with few or no labeled samples available from subjects of interest, which refer to inductive and transductive setups, respectively. Figure 1 illustrates the covariate-shift

*Corresponding author

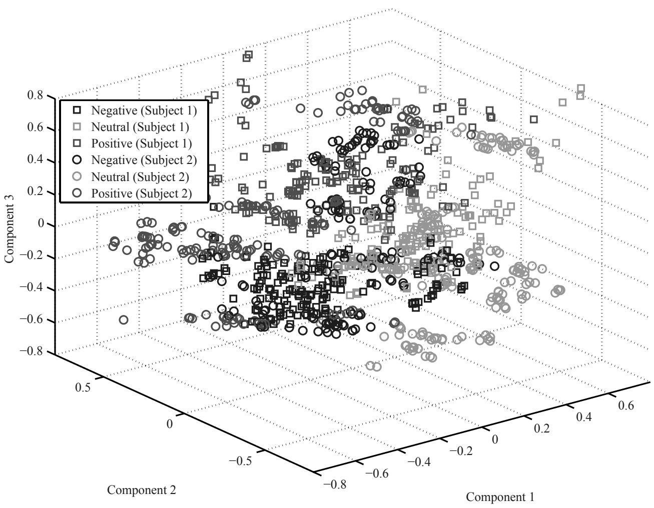

Figure 1: Illustration of the covariate-shift challenges of constructing EEG-based affective models. Here, two sample subjects (subjects 1 and 2) are with three classes of emotions and EEG features are different in conditional probability distribution across subjects.

challenges of constructing EEG-based affective models. Traditional machine learning methods have a prior assumption that the distributions of training data and test data are independently and identically distributed (i.i.d.). However, due to the variability from subject to subject, this assumption can not be always satisfied in aBCIs.

In this work, we adopt transfer learning algorithms in a transductive setup (without any labeled samples from target subjects) to tackle the subject-to-subject variability for building EEG-based affective models. Let $X \in \mathcal{X}$ be the EEG recording of a sample (X, y), here $y \in \mathcal{Y}$ represents the corresponding emotion labels. In this case, $\mathcal{X} = \mathbb{R}^{C \times d}$, C is the number of channels, and d is the number of time series samples. Let $P(X)$ be the marginal probability distribution of X. According to [Pan and Yang, 2010], $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is a domain, which in our case is a given subject from which we record the EEG signals. The source and target domains in this paper share the same feature space, $\mathcal{X}_S = \mathcal{X}_T$, but the respective marginal probability distributions are different, that is, $P(X_S) \neq P(X_T)$. The key assumption in most domain adaptation methods is that $P(Y_S|X_S) = P(Y_T|X_T)$.

Recently, Jayaram and colleges made a timely survey on current transfer learning techniques for BCIs [Jayaram *et al.*, 2016]. Morioka *et al.* proposed to learn a common dictionary shared by multiple subjects and used the resting-state activity of a previously unseen target subject as calibration data for compensating for individual differences, rather than task sessions [Morioka *et al.*, 2015]. Krauledat and colleges proposed a zero-training framework for extracting prototypical spatial filters that have better generalization properties [Krauledat *et al.*, 2008]. Although most of these methods are based on the variants of common spacial patterns (CSP) for motor-imagery paradigms, some studies focused on passive BCIs [Wu *et al.*, 2013] and EEG-based emotion recognition [Zheng *et al.*, 2015].

In this paper, we propose to personalize EEG-based affective models by adopting two kinds of domain adaptation approaches in an unsupervised manner. One is to find a shared common feature space between source and target domains. We apply Transfer Component Analysis (TCA) and Kernel Principle Analysis (KPCA) based methods proposed in [Pan *et al.*, 2011]. These methods are adopted to learn a set of common transfer components underlying both the source domain and the target domain. When projected to this subspace, the difference of feature distributions of both domains can be reduced. The other is to construct individual classifiers and learn a regression function that maps the relationship between data distribution and classifier parameters, which refers to Transductive Parameter Transfer (TPT) [Sangineto *et al.*, 2014]. We evaluate the performance of these approaches on an EEG dataset, SEED[1], to personalize EEG-based affective models.

2 Methods

2.1 TCA-based Subject Transfer

Transfer component analysis (TCA) proposed by Pan *et al.* learns a set of common transfer components between source domain and target domain [Pan *et al.*, 2011] and finds a low-dimensional feature subspace across source domain and target domain, where the difference of feature distributions between two domains can be reduced. The aim of this transfer learning algorithm is to find a transformation $\phi(\cdot)$ such that $P(\phi(X_S)) \approx P(\phi(X_T))$ and $P(Y_S|\phi(X_S)) \approx P(Y_T|\phi(X_T))$ without any labeled data in target domain (target subjects). An intuitive approach to find the mapping $\phi(\cdot)$ is to minimize the Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2006] between the empirical means of the two domains,

$$MMD(X'_S, X'_T) = ||\frac{1}{n_1}\sum_{i=1}^{n_1}\phi(x_i^s) - \frac{1}{n_2}\sum_{i=1}^{n_2}\phi(x_i^t)||^2_{\mathcal{H}}, \quad (1)$$

where n_1 and n_2 represent the sample numbers of source domain and target domain, respectively. However, $\phi(\cdot)$ is usually highly nonlinear and a direct optimization with respect to $\phi(\cdot)$ can easily get stuck in poor local minima [Pan *et al.*, 2011].

TCA is a dimensionality reduction based domain adaptation method. It embeds both the source and target domain data into a shared low-dimensional latent space using a mapping ϕ. Specially, let the Gram matrices defined on the source domain, target domain and cross-domain data in the embedded space be $K_{S,S}$, $K_{T,T}$, and $K_{S,T}$, respectively. The kernel matrix K is defined on all the data as

$$K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}. \quad (2)$$

By virtue of kernel trick, the MMD distance can be rewritten as $tr(KL)$, where $K = [\phi(x_i)^\top \phi(x_j)]$, and $L_{ij} = 1/n_1^2$ if $x_i, x_j \in X_S$, else $L_{ij} = 1/n_2^2$ if $x_i, x_j \in X_T$, otherwise,

[1] http://bcmi.sjtu.edu.cn/~seed/index.html

$L_{ij} = -(1/n_1 n_2)$. A matrix $\widetilde{W} \in \mathbb{R}^{(n_1+n_2)\times m}$ transforms the empirical kernel map K to an m-dimension space (where $m \ll n_1 + n_2$). The resultant kernel matrix is

$$\widetilde{K} = (KK^{-1/2}\widetilde{W}(\widetilde{W}^{\top} K^{-1/2}K)) = KWW^{\top}K, \quad (3)$$

where $W = K^{-1/2}\widetilde{W}$. With the definition of $\widetilde{K}$ in Eq.(3), the MMD distance between the empirical means of the two domain X'_S and X'_T can be rewritten as

$$Dist(X'_S, X'_T) = tr((KWW^{\top}K)L) = tr(W^{\top}KLKW). \quad (4)$$

A regularization term $tr(W^{\top}W)$ is usually added to control the complexity of W, while minimizing Eq.(4).

Besides reducing the difference of the two distributions, ϕ should also preserve the data variance that is related to the target learning task. From Eq.(3), the variance of the projected samples is $W^{\top}KHKW$, where $H = I_{n_1+n_2} - (1/(n_1 + n_2))\mathbf{1}\mathbf{1}^{\top}$ is the centering matrix, $\mathbf{1} \in \mathbb{R}^{n_1+n_2}$ is the column vector with all 1's, and $I_{n_1+n_2} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ is the identity matrix.

Therefore, the objective function of TCA is

$$\begin{aligned} \min_{W} \quad & tr(W^{\top}KLKW) + \mu tr(W^{\top}W) \\ s.t. \quad & W^{\top}KHKW = I_m, \end{aligned} \quad (5)$$

where $\mu > 0$ is a regularization parameter, and $I_m \in \mathbb{R}^{m\times n}$ is the identity matrix. According to [Pan *et al.*, 2011], the solutions W are the m leading eigenvectors of $(KLK + \mu I)^{-1}KHK$, where $m \le n_1 + n_2 - 1$. The algorithm of TCA for subject transfer is summarized in Algorithm 1. We recommend the readers to refer to [Pan *et al.*, 2011] for the detailed descriptions of TCA. After obtaining the transformation matrix W, standard machine learning methods can be used in this feature subspace.

Algorithm 1 TCA-based Subject Transfer

input : Source domain data set $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_1}$, and target domain data set $\mathcal{D}_T = \{x_j^t\}_{j=1}^{n_2}$.
output : Transformation matrix W.

1: Compute kernel matrix K from $\{x_i^s\}_{i=1}^{n_1}$ and $\{x_j^t\}_{j=1}^{n_2}$, matrix L, and the centering matrix H.
2: Eigendecompose the matrix $(KLK + \mu I)^{-1}KLK$ and select the m leading eigenvectors to construct the transformation matrix W.
3: **return** tranformation matrix W.

2.2 KPCA-based Subject Transfer

Kernel PCA [Schölkopf *et al.*, 1998] projects the original D-dimensional feature space into an M-dimensional feature space with a nonlinear transformation $\phi(x)$, where $M \gg D$. For KPCA-based subject transfer, we concatenate the source and target data as the training data and construct the kernel matrix. The kernel principal components are then computed using singular value decomposition. Each sample x_i is projected to a point $\phi(x_i)$ in lower dimensional subspaces. The algorithm of KPCA-based subject transfer is summarized in Algorithm 2.

Algorithm 2 KPCA-based Subject Transfer

input : Source domain data set $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_1}$, and target domain data set $\mathcal{D}_T = \{x_j^t\}_{j=1}^{n_2}$.
output : The kernel principal components p_k.

1: Concatenate the source and target domain data sets as the training data set, $\{x_i\}_{i=1}^{n_1+n_2} = [\{x_i^s\}_{i=1}^{n_1}; \{x_j^t\}_{j=1}^{n_2}]$.
2: Construct the kernel matrix from the training data set $\{x_i\}_{i=1}^{n_1+n_2}$.
3: Compute the vectors $\mathbf{a}_k$.
4: Compute the kernel principal components p_k.
5: **Return** the kernel principal components p_k.

2.3 Transductive Parameter Transfer

The transductive parameter transfer (TPT) approach is firstly proposed by Sangineto and colleges for action units detection and spontaneous pain recognition [Sangineto *et al.*, 2014]. The TPT approach consists of three main steps as illustrated in Figure 2. First, multiple individual classifiers are learned on each training dataset D_i^s. Second, a regression function is trained to learn the relation between the data distributions and classifiers' parameter vectors. Finally, the target classifier is obtained using the target feature distribution and the distribution-to-classifier mapping.

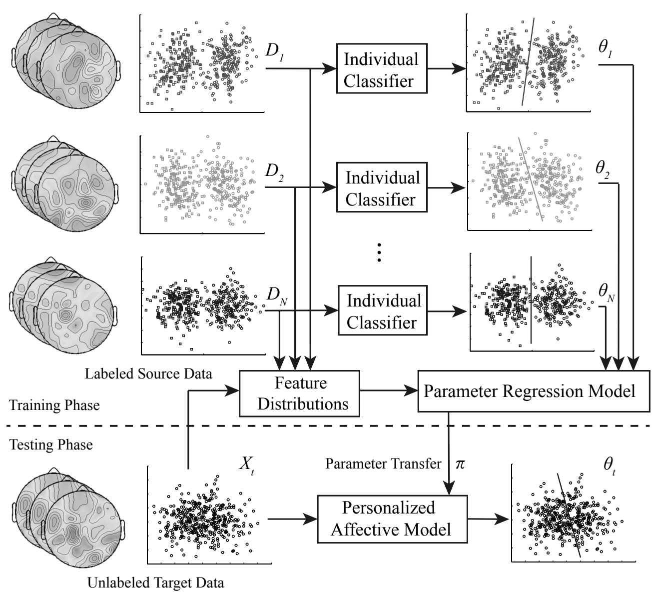

Figure 2: The framework of the transductive parameter transfer (TPT) approach adopted in this work.

We adopt the TPT approach to personalize EEG-based affective models in this paper. In the first phase, we train multiple individual classifiers on each source dataset D_i^s. Here, we use linear support vector machine (SVM) as a classifier. $\boldsymbol{\theta}_i = [\boldsymbol{w}'_i, b_i]$ defines the hyperplane in the feature space. The objective function is as follows,

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \lambda_L \sum_{j=1}^{n_i^s} l(\boldsymbol{w}'\boldsymbol{x}_j^s + b, y_j^s), \quad (6)$$

where $l(\cdot)$ is the hinge loss.

In the second step, a regression function f is learned for the mapping: $D \rightarrow \Theta$ with the source data distributions. Since the data distributions and the optimal corresponding hyperplanes are relevant, we can predict the hyperplane and construct the classifier on target data without any label information via learning this mapping. To quantify the similarity between pairs of datasets X_i and X_j, a kernel function $\kappa(X_i, X_j)$ is adopted. In this implement, we use the density estimation kernel [Blanchard *et al.*, 2011], which is defined as follows,

$$\kappa(X_i, X_j) = \frac{1}{nm} \sum_{p=1}^{n} \sum_{q=1}^{m} \kappa_{\mathcal{X}}(\boldsymbol{x}_p, \boldsymbol{x}_q), \tag{7}$$

where n, m are cardinality of X_i, X_j, respectively, and $\kappa_{\mathcal{X}}(\cdot)$ is a Gaussian kernel. After computing the kernel matrix, the mapping function f can be learned using the Multioutput Support Vector Regression framework [Tuia *et al.*, 2011].

Finally, the parameter vector of the target classifier can be predicted by $\boldsymbol{\theta}_t = f(X^t)$ without any label information from target subjects. Given the target features $\boldsymbol{x}$ and the classifier parameters $\boldsymbol{\theta}$, the label can be predicted by the decision function: $y = sign(\boldsymbol{w}_t'\boldsymbol{x} + b_t)$. The algorithm of TPT-based subject transfer is summarized in Algorithm 3. For more details about the TPT algorithm, we recommend the readers to refer to [Sangineto *et al.*, 2014].

Algorithm 3 TPT-based Subject Transfer

input : Source domain data sets $\mathcal{D}_1^s, ..., \mathcal{D}_N^s$, target domain data set X^t, and some regularization parameters for SVMs.
output : The parameter vector of target classifier: $\boldsymbol{w}_t, b_t$.

1: Construct individual classifiers: $\{\boldsymbol{\theta_i} = (\boldsymbol{w}_i, b_i)\}_{i=1}^N$.
2: Create a training set $\mathcal{T} = \{X_i^s, \boldsymbol{\theta_i}\}_{i=1}^N$.
3: Compute the kernel matrix $\boldsymbol{K}$, $\boldsymbol{K}_{ij} = \kappa(X_i^s, X_j^s)$.
4: Given K and $\mathcal{T}$, learn $f(\cdot)$ using multioutput support vector regression.
5: Compute $(\boldsymbol{w}_t, b_t) = f(X^t)$
6: **Return** $\boldsymbol{w}_t, b_t$.

3 Experiment Setup

3.1 EEG Dataset for Constructing Affective Models

We adopt film clips as stimuli to elicit emotions in the laboratory environment, since film clips contains both visual and auditory stimuli. A preliminary study is conducted using scores (1-5) to select a pool of film clips to elicit three emotions: positive, neutral, and negative. Each clip is well edited for its coherent context for corresponding emotion and time length of about four minutes. Finally, 15 emotional film clips with high ratings are chosen from the materials. For the experimental protocol, each experiment consists of 15 sessions with 15 different film clips. For each session, there is a 5 s hint for starting before each clip, a 45 s self-assessment, and a 15 s rest after each clip. There are totally 15 subjects (8 females, mean: 23.27, std: 2.37) participated in our experiments. They are all informed about the experiments and instructress before the experiments. They are required to elicit their own corresponding emotions while watching the film clips. Only the data with right elicited emotions are used in further analysis. EEG data are recorded simultaneously with a 62-electrode cap according to the international 10-20 system using ESI Neuroscan system. The original sampling rate is 1000 Hz. The impedance of each electrode is lower than 5 kΩ.

3.2 Data Preprocessing and Feature Extraction

For data preprocessing, since there is often contamination of electromyography (EMG) signals from facial expressions and Electrooculogram (EOG) signals from eye movements in EEG data [Fatourechi *et al.*, 2007], the raw EEG data is processed with a bandpass filter between 1 Hz and 75 Hz and the data with serious noise and artifacts are discarded. The 62-channel EEG signals are further down-sampled to 200 Hz and EEG features are extracted with each segment of the preprocessed EEG data with a non-overlapping 1 s time window.

For feature extraction, we employ differential entropy (DE) features [Duan *et al.*, 2013; Zheng and Lu, 2015], which show superior performance than conventional power spectral density (PSD) features. According to [Zheng and Lu, 2015], for a fixed length EEG sequence, the DE feature is equivalent to the logarithm of PSD in a certain frequency band. Therefore, the DE features can be calculated in five frequency bands (δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz), which are widely used in EEG studies using Short-term Fourier transform. The total dimension of a 62-channel EEG segment is 310.

3.3 Evaluation Details

We adopt a leave-one-subject-out cross validation method for the evaluation. Each time, we separate the data from one subject as the target domain and the resting data from other 14 subjects as the source domain. For the baseline method, we concatenate data from all available subjects as training data and train a generic classifier with linear SVM. A variant of SVM called Transductive SVM (T-SVM) is also developed to learn a decision boundary and maximize the margin with unlabeled data [Collobert *et al.*, 2006]. We use the implement of T-SVM in SVM^{light} [Joachims, 1999].

For TCA and KPCA, it is practically infeasible to include all the data from available subjects due to limits of memory and time cost for singular value decomposition. Therefore, we randomly select a subset of samples (5000 samples) from 14 subjects as training data. For kernel functions, we employ linear kernels. We will evaluate how the performance varies with respect to the dimensionality of new feature space. The regularization parameter μ is set to 1, the same as [Pan *et al.*, 2011]. After TCA and KPCA project the original features into low-dimensional subspace using transfer components, standard linear SVMs are trained on the new extracted features. The value of the regularization parameter for TPT is 0.1. To deal with multi-class classification task, we adopt the one vs one strategy to avoid label unbalance problem. All the algorithms are implemented in MATLAB.

4 Experiment Results

In this section, we carry out experiments to demonstrate the effectiveness of the proposed methods for personalizing EEG-based affective models. All results are conducted using a leave-one-subject-out evaluation scheme. First, we evaluate how the performance of KPCA and TCA approaches varies with the dimensionality of the low-dimensional feature subspace and choose the best dimensions. Figure 3 depicts the accuracy curve with respect to varying dimensions. TCA achieves better performance than KPCA in lower dimensions (less than 30) and reaches the peak accuracy with 63.64% in the 30-dimensional subspace. In contrast, the accuracies of KPCA approximately increase with the increasing dimensions. The saturation is reached at about 35-dimension point. Regarding the accuracy, TCA outperforms KPCA slightly (63.64% vs 61.28%).

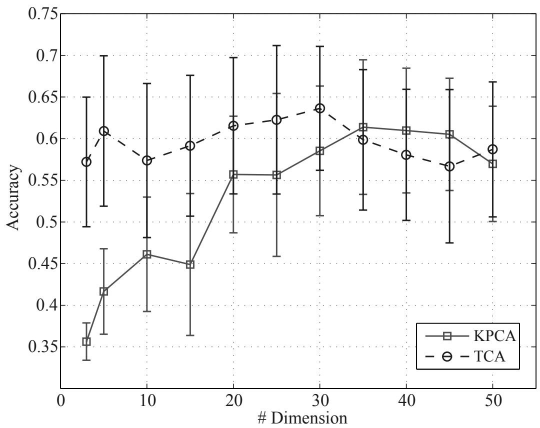

Figure 3: Comparison of KPCA and TCA approaches for different dimensionality of the subspace.

We compare the performance of different subject transfer methods. Figure 4 shows the accuracies of different methods (Generic classifiers, KPCA, TCA, T-SVM, and TPT) for total 15 subjects and Table 1 presents the mean accuracies and standard deviations. The generic classifiers perform poorly with a mean accuracy of only 56.73% due to the fact that this method directly includes all source samples for the training. Since there exist some individual differences across subjects and some training samples from irrelevant subjects included, all these factors may bias the optimal hyperplane and dramatically degrade the performance of subject transfer, which refers to some negative transfer. TCA and KPCA learn a set of transfer components underlying both the source and target distributions. The feature distributions of both domains are similar in the new low-dimensional subspace. Both TCA and KPCA methods outperform the generic classifiers, indicating the efficient knowledge transfer through feature reduction. T-SVM archives comparative accuracies among these methods with the mean accuracy and standard deviation of 72.53%/14.00%, respectively. T-SVM learns the decision boundary in a semi-supervised manner and weights all the training instances equally, which may still introduce some irrelevant source data during training.

TPT method outperforms the other four approaches with the highest accuracy of 76.31% and achieves a significant improvement in comparison with generic classifiers (one way

Stats.	Generic	KPCA	TCA	T-SVM	TPT
Mean	0.5673	0.6128	0.6364	0.7253	**0.7631**
Std.	0.1629	0.1462	0.1488	**0.1400**	0.1589

Table 1: The mean accuracies and standard deviations of the five different approaches.

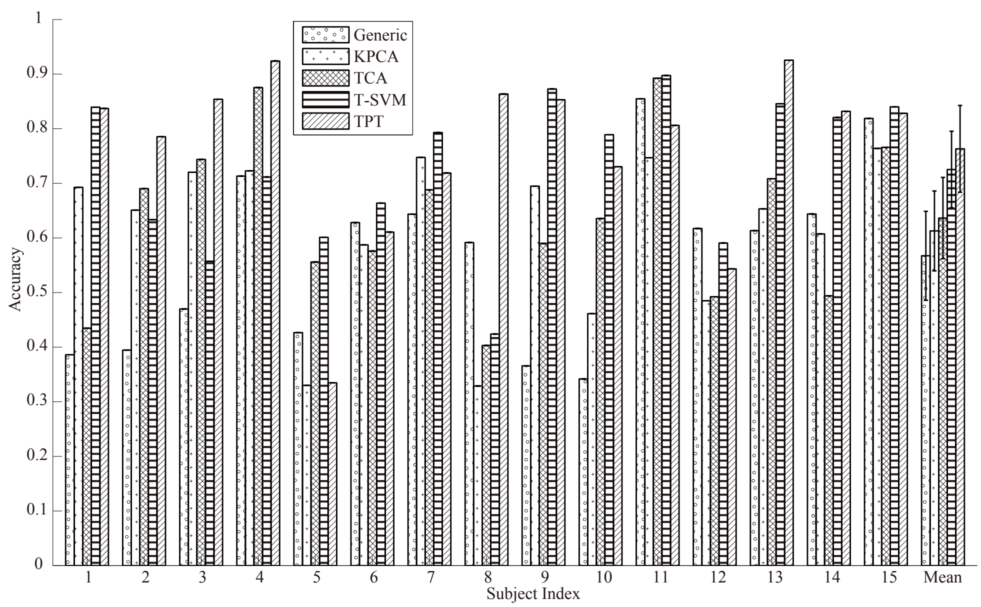

Figure 4: The accuracies of the five different methods (Generic classifiers, KPCA, TCA, T-SVM, and TPT) for each subject.

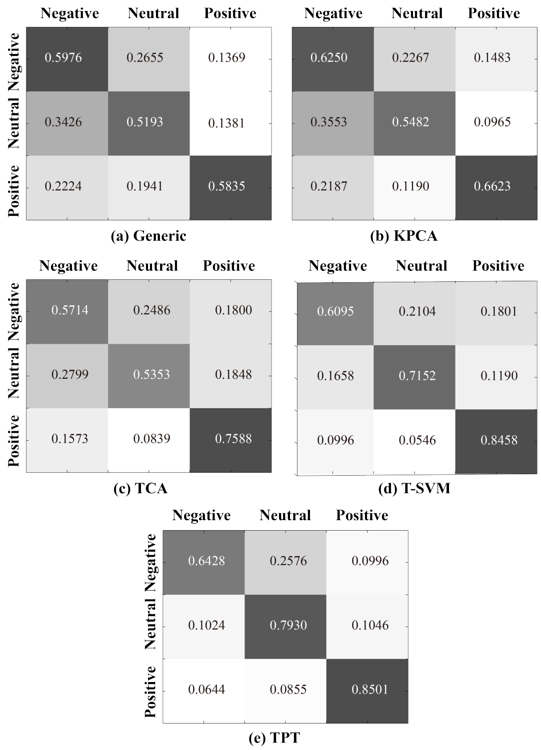

Figure 5: The confusion matrices of the five methods.

analysis of variance, $p < 0.01$). TPT method can measure the similarity between pairs of data distributions from different subjects and learn the mapping function from data distributions and classifier parameters. Therefore, it can extract the relevant information to determine the decision function and bypass the bias caused by irrelevant information. Besides accuracy, we compare the generalization performance of KPCA, TCA, and TPT approaches. The KPCA and TCA need to construct the kernel matrix using source and target domains and find the manifold subspaces where the differences between them can be reduced. Their limitation is high memory and time cost for training. They can not constantly benefit from the increasing samples in the source domain. In contrast, TPT keeps individual classifiers for different source subjects. The new regression function can be trained only on the kernel matrix. It is incremental while data from a new training subject is available. Therefore, TPT approach is more feasible in practice regarding both the performance and incremental learning property.

The confusion matrices of the five approaches are shown in Figure 5. Each row of the confusion matrix represents the target class and each column represents the predicted class. The element (i, j) is the percentage of samples in class i that is classified as class j. For generic method, the accuracies for three emotions are almost similar. For TCA and KPCA, they have an improvement for classifying positive emotions with the accuracies of 75.88% and 66.23%, respectively. Besides the improved performance of recognizing positive emotions, T-SVM has a significant increase in performance for recognizing neutral emotions (71.52%). TPT has highest accuracies for classifying all of the three emotions among these approaches. Comparing the accuracies of different emotions, we can find that positive emotions can be more easily recognized using EEG with the comparatively high accuracy of 85.01%. Negative emotions are often confused with neutral emotions (25.76%) and vice versa (10.24%). These results indicate that the neural patterns of negative and neutral emotions are similar. In summary, the experimental results demonstrate the efficiency of the TPT approach for constructing personalized affective models from EEG.

5 Conclusion and Future Work

In this paper, we have proposed a novel method for personalizing EEG-based affective models with transfer learning techniques. The affective models are personalized and constructed for a new target subject without any label information. We have compared the performance of five different methods: TPT, T-SVM, TCA, KPCA, and conventional generic classifiers. The experimental results have demonstrated that the transductive parameter transfer approach significantly outperforms the other approaches in terms of the accuracies, and a 19.58% increase in recognition accuracy is achieved. TPT can capture the similarity between data distributions taking advantages of kernel functions and learn the mapping from data distributions to classifier parameters with the regression framework. For the future work, we will evaluate the performance of our proposed approaches for more categories of emotions as well as the publicly available EEG datasets.

Acknowledgments

This work was supported in part by the grants from the National Natural Science Foundation of China (Grant No.61272248), the National Basic Research Program of China (Grant No.2013CB329401), and the Major Basic Research Program of Shanghai Science and Technology Committee (15JC1400103).

References

[Abraham *et al.*, 2014] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014.

[Blanchard *et al.*, 2011] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances In Neural Information Processing Systems*, pages 2178–2186, 2011.

[Chu *et al.*, 2013] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Selective transfer machine for personalized facial action unit detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522. IEEE, 2013.

[Chung *et al.*, 2011] Mike Chung, Willy Cheung, Reinhold Scherer, and Rajesh PN Rao. A hierarchical architecture for adaptive brain-computer interfacing. In *IJCAI'11*, volume 22, page 1647, 2011.

[Collobert *et al.*, 2006] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Large scale transductive SVMs. *The Journal of Machine Learning Research*, 7:1687–1712, 2006.

[Duan *et al.*, 2009] Lixin Duan, Ivor W Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *26th Annual International Conference on Machine Learning*, pages 289–296. ACM, 2009.

[Duan *et al.*, 2013] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering*, pages 81–84. IEEE, 2013.

[Eaton *et al.*, 2015] Joel Eaton, Duncan Williams, and Eduardo Miranda. The space between us: Evaluating a multi-user affective brain-computer music interface. *Brain-Computer Interfaces*, 2(2-3):103–116, 2015.

[Fatourechi *et al.*, 2007] Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118(3):480–494, 2007.

[Gretton *et al.*, 2006] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

[Jayaram *et al.*, 2016] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

[Jenke *et al.*, 2014] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

[Joachims, 1999] Thorsten Joachims. Making large scale SVM learning practical. Technical report, Universität Dortmund, 1999.

[Krauledat *et al.*, 2008] Matthias Krauledat, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Mller. Towards zero training for brain-computer interfacing. *PloS One*, 3:e2967, 08 2008.

[Morioka *et al.*, 2015] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 2015.

[Mühl *et al.*, 2014a] Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84, 2014.

[Mühl *et al.*, 2014b] Christian Mühl, Camille Jeunet, and Fabien Lotte. EEG-based workload estimation across affective contexts. *Frontiers in Neuroscience*, 8, 2014.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[Samek *et al.*, 2013] Wojciech Samek, Frank C Meinecke, and Klaus-Robert Muller. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Transactions on Biomedical Engineering*, 60(8):2289–2298, 2013.

[Sangineto *et al.*, 2014] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *ACM International Conference on Multimedia*, pages 357–366. ACM, 2014.

[Schölkopf *et al.*, 1998] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[Singh *et al.*, 2007] Vishwajeet Singh, Krishna P Miyapuram, and Raju S Bapi. Detection of cognitive states from fMRI data using machine learning techniques. In *IJCAI'07*, pages 587–592, 2007.

[Sugiyama *et al.*, 2007] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research*, 8:985–1005, 2007.

[Tuia *et al.*, 2011] Devis Tuia, Jochem Verrelst, Luis Alonso, Fernando Pérez-Cruz, and Gustavo Camps-Valls. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geoscience and Remote Sensing Letters*, 8(4):804–808, 2011.

[Wu *et al.*, 2013] Dongrui Wu, Brent J Lance, and Thomas D Parsons. Collaborative filtering for brain-computer interaction using transfer learning and active class selection. *PloS One*, 8(2), 2013.

[Zander and Jatzev, 2012] Thorsten O Zander and S Jatzev. Context-aware brain–computer interfaces: exploring the information space of user, technical system and environment. *Journal of Neural Engineering*, 9(1):016003, 2012.

[Zheng and Lu, 2015] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[Zheng *et al.*, 2015] Wei-Long Zheng, Yong-Qi Zhang, Jia-Yi Zhu, and Bao-Liang Lu. Transfer components between subjects for EEG-based emotion recognition. In *International Conference on Affective Computing and Intelligent Interaction*, pages 917–922. IEEE, 2015.