Spatial Filter Adaptation Based on the Divergence Framework for Motor Imagery EEG Classification

Xinyang Li[1,2], Cuntai Guan[2], Kai Keng Ang[2], Haihong Zhang[2], and Sim Heng Ong[3]

***Abstract*— To address the nonstationarity issue in EEG-based brain computer interface (BCI), the computational model trained using the training data needs to adapt to the data from the test sessions. In this paper, we propose a novel adaptation approach based on the divergence framework. Cross-session changes can be taken into consideration by searching the discriminative subspaces for test data on the manifold of orthogonal matrices in a semi-supervised manner. Subsequently, the feature space becomes more consistent across sessions and classifiers performance can be enhanced. Experimental results show that the proposed adaptation method yields improvements in classification performance.**

I. INTRODUCTION

Signal nonstationarity in EEG is one of the most critical issues faced by brain computer interface (BCI) systems that are driven based on EEG. Independent of voluntary muscle control, a BCI-based rehabilitation system helps patients to restore their motor functions as an important alternative to labor-intensive and expensive traditional physical therapy [1], [2], [3], [4], [5]. However, EEG patterns generated by BCI users or patients could vary drastically due to task-unrelated mental conditions and different experimental setups [6], [7]. Such significant nonstationarity in the data causes failures of BCI in detecting the correct mental conditions from EEG signals, which undermines the effectiveness of the rehabilitation process. Because the calibration procedure is tedious and time-consuming, usually only the computational model that is obtained from the calibration session is available for all the following rehabilitation sessions [1].

Common spatial pattern (CSP) is one of the most successful feature extraction methods in discriminating EEG in BCI [8], [9]. However, as a supervised method, CSP is sensitive to data variation across sessions. Efforts have been made to enhance its performance by considering the shifts in the CSP feature space [6], [10]. It is shown in [6] that the two-class motor imagery EEG classification accuracy could increase significantly by a bias adaptation of the classifier in the CSP feature space. The shortcoming of this kind of methodology is that the classifier adaptation is not effective when the test features are inseparable. To address the issue of feature separability, some works investigate the adaptation of the feature extraction model [11], [12]. In particular, since the solution of the spatial filter in CSP is based on the joint diagonalization of the average covariance matrices, variations of EEG data across sessions can be taken into consideration by incorporating data from test sessions to update the projection matrix in CSP. Another approach assumes that there is a domain-invariant subspace, where the classifier trained by training data could be equally effective to test data [13], [14]. In [13], this domain-invariant subspace is assumed to be the whitened subspace, where the whitened training data and test data have the same (or similar) marginal distributions, and the posterior distributions of the labels are the same across domains. Therefore, the whitening part in the spatial filter is updated based on test data, which is equivalent to projecting both training data and test data to the whitened space. As pointed in [13], this domain-invariant assumption of the whitened space holds only when the linear transformation between the two domains is symmetric.

Due to significant cross-session data variation, the discriminative subspaces vary from the training data to the test data. The major challenge is adapting discriminative subspaces for the test data while keeping the feature spaces consistent from session to session. To solve this problem, we develop a model adaptation method by formulating the divergence of distributions in different subspaces based on the framework proposed in [15]. The adaptation objective is to maximize inter-class divergence between the test data distribution in the adapted subspaces and the training data distribution in the original subspaces. By adding a regularization term, within-class divergence could also be taken into consideration. In this way, although different projection matrices are applied to training data and test data, the feature space is more consistent and the performance of the classifier can be improved.

This paper is organized as follows. In Section II, the divergence-based framework of the spatial filter design and adaptation are presented. In Section III, the validity of the proposed method is verified by experimental studies on two-class motor imagery classification. Concluding remarks are given in Section IV.

II. SPATIAL FILTER ADAPTATION BASED ON THE DIVERGENCE FRAMEWORK

A. Divergence-Based CSP

It is showed in [15] that spatial filters W in CSP project the EEG data into subspaces where the Kullback-Leibler divergence (KL-divergence) between the data distributions

[1]X. Li is with the NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 119613 a0068297@nus.edu.sg

[2]C. Guan, K. K. Ang, and H. Zhang are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 {ctguan,kkang,hhzhang}@i2r.a-star.edu.sg

[3]S. H. Ong is with the Department of Electrical and Computer Engineering, and Department of Biomedical Engineering, National University of Singapore, Singapore 119613 eleongsh@nus.edu.sg

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

from two classes is maximized. Thus, the objective function of the divergence-based CSP (divCSP) is in the form

$$\mathcal{L}_0(W) = (1-\lambda)\tilde{D}_{kl}(W^T R^+ W || W^T R^- W) - \lambda\Delta \quad (1)$$

where $R^{+/-} \in \mathbb{R}^{n_c \times n_c}$ is the average covariance matrix of class + or - with n_c being the number of channels used to measure the EEG data. In (1), $\tilde{D}_{kl}(W^T R^+ W || W^T R^- W)$ is the CSP objective in the form of symmetric KL-divergence, Δ is the regularization term, and λ is the regularization parameter. Δ is also based on KL-divergence and it is defined according to the type of nonstationarity to be minimized. The regularization used in the proposed adaptation objective function will be introduced in the next section.

The solution of minimizing (1) is in the form

$$W = (P^T U)^T \quad (2)$$

where $P \in \mathbb{R}^{n_c \times n_c}$ is the whitening matrix, and $U \in \mathbb{R}^{n_c \times n_c}$ is the rotation matrix, of which each column is orthogonal to each other. Considering the training stage, we use W_{tr}, P_{tr} and U_{tr} to denote the matrices obtained using the training data. Thus, we have

$$P_{tr}(R^+_{tr} + R^-_{tr})P^T_{tr} = I \quad (3)$$

where $R^{+/-}_{tr}$ is the average covariance matrix of training data from class $+$ or $-$, and I is the identity matrix. In [13], it has been established that the projection matrix can be adapted by replacing the whitening part, which is also regarded as a normalization approach. With $P_{te} = R_{te}^{-\frac{1}{2}}$, where R_{te} is the average covariance matrix of test data regardless of labels, the updated projection matrix W_n is

$$\begin{aligned} W_n &= (P^T_{te} U_{tr})^T \\ &= W_{tr} P^{-1}_{tr} P_{te} \end{aligned} \quad (4)$$

By only updating the whitening part, the orthogonal part U_{tr} in W_{tr} is maintained. It is also pointed out in [13] that the orthogonal part U_{tr} is kept constant across sessions if and only if the cross-session data projection is symmetric. To address a more general case of adaptation, we propose to adapt the rotation matrix U using the divergence-based framework [15].

B. Spatial Filter Adaptation through Subspace Tracking

To ensure that the test features are in the same space with the classifier, we propose the following objective function for adaptation

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{csp} + \lambda\Delta \quad (5)$$

where

$$\begin{aligned} \mathcal{L}_{csp} &= \tilde{D}_{kl}(W^T_{te} R^+_{te} W_{te} || W^T_{tr} R^-_{tr} W_{tr}) + \\ &\quad \tilde{D}_{kl}(W^T_{te} R^-_{te} W_{te} || W^T_{tr} R^+_{tr} W_{tr}) \quad (6) \\ \Delta &= D_{kl}(W^T_{te} R^+_{te} W_{te} || W^T_{tr} R^+_{tr} W_{tr}) + \\ &\quad D_{kl}(W^T_{te} R^-_{te} W_{te} || W^T_{tr} R^-_{tr} W_{tr}) \quad (7) \end{aligned}$$

Instead of the distribution divergence of the test data between two classes, the distribution divergence between the test data and training data is formulated in (6) and (7). In this way, inter-class and within-class divergence between the test data in the adapted subspaces and the training data in the original subspaces could be maximized and minimized, respectively. This is to guarantee that the classifier trained by training features and the test features are in the same space.

Given P_{te}, the covariance matrix of test data after whitening is

$$\tilde{R}^{+/-}_{te} = P R^{+/-}_{te} P^T \quad (8)$$

Note that for the adaptation without test labels, $R^{+/-}_{te}$ is estimated using the predicted labels. Based on (8), (5)-(7) can be rewritten as functions of U

$$\begin{aligned} \mathcal{L}(U) &= (1-\lambda)\mathcal{L}_{csp}(U) + \lambda\Delta(U) \quad (9) \\ \mathcal{L}_{csp}(U) &= \tilde{D}(I_d U \tilde{R}^+_{te} U^T I^T_d || W^T R^-_{tr} W) \\ &\quad + \tilde{D}(I_d U \tilde{R}^-_{te} U^T I^T_d || W^T R^+_{tr} W) \quad (10) \\ \Delta(U) &= D_{kl}(I_d U \tilde{R}^+_{te} U^T I^T_d || W^T R^+_{tr} W) + \\ &\quad D_{kl}(I_d U \tilde{R}^+_{te} U^T I^T_d || W^T R^-_{tr} W) \quad (11) \end{aligned}$$

where $I_d \in \mathbb{R}^{d \times n_c}$ is the identity matrix truncated to the first d rows, and d is the number of spatial filters. Therefore, the adapted rotation matrix is

$$U_{te} = \arg\min_U \mathcal{L}(U) \quad (12)$$

C. Semi-Supervised Gradient Descent Searching

To solve (12), we adopt a subspace approach based on gradient descent on the manifold of orthogonal matrices [15], [16]. In the training stage, subspace searching can be performed with labels and stopped by the convergence of the loss function $\mathcal{L}_0$. For the adaptation without test labels, convergence of the loss function $\mathcal{L}$ could be problematic. Adaptation until the convergence of $\mathcal{L}$ is more prone to the overfitting due to the limited number of available test trials and performance drops caused by the incorrect predicted labels. To avoid these problems, in the proposed adaptation design, the objective function $\mathcal{L}$ used to update U is calculated based on a subset of available test trials. During the gradient descent search, the loss function for all available test trials, denoted as $\mathcal{L}_a$, is also evaluated at each iteration step. Stopping of the iterations is subject to both the change of $\mathcal{L}$ and $\mathcal{L}_a$. In this way, some of the trials used to evaluate the change of loss function are independent of the adaptation. Details of the semi-supervised adaptation is summarised in Algorithm 1. After U_{te} is obtained, the projection matrix W_{te} can be calculated as

$$W_{te} = (P^T_{te} U_{te})^T \quad (13)$$

III. Experimental Study

A. Experimental Setup

EEGs from the full 27 channels were obtained using Nuamps EEG acquisition hardware with unipolar Ag/AgCl electrodes channels. The sampling rate was 250 Hz with a resolution of 22 bits for the voltage range of $\pm$ 130 mV. A

Algorithm 1: Subspace searching based on gradient descent

Input: training data and adaptation data;
Output: U_{te}.
begin
  Compute P_{te};
  Initialize $U = U_{tr}$;
  repeat
    Compute the gradient matrix M of $\mathcal{L}(U)$ with respect to U;
    Compute

$$H = \begin{pmatrix} 0 & M \\ -M^T & 0 \end{pmatrix} \tag{14}$$

    Determine the optimal step size;
    Update the rotation matrix

$$U_{k+1} = \exp(tH)U_k \tag{15}$$

    Compute $\mathcal{L}_a(U_{k+1})$;
    if $\mathcal{L}_a(U_{k+1}) > \mathcal{L}_a(U_k)$ **then**
      break.
    end
  until *Convergence of* $\mathcal{L}$;
  $U_{te} = U$;
end

bandpass filter of 0.05 to 40 Hz was set in the acquisition hardware.

In the experiment, the training and test sessions were recorded on different days with the subjects performing motor imagery. During the EEG recording process, the subjects were asked to avoid physical movement and eye blinking. Additionally, they were instructed to perform kinaesthetic motor imagery of the chosen hand in two runs. During the rest state, they did mental counting to make the resting EEG signal more consistent. Each run lasted for approximately 16 minutes and comprised 40 trials of motor imagery and 40 trials of rest state. Each training session consisted of 2 runs while the test session consisted of 2-3 runs. Details of the experimental setup can be found in [17].

B. Data Processing and Feature Extraction

First, we train a CSP model and the naive bayesian parzen window (NBPW) classifier with the training data as in [18], [19]. Then, as described in Section II-C, with the predicted labels of a batch of the test data from the new session, the projection matrix W_{te} is calculated and applied to test data for feature extraction. The number of the spatial filters d is 6. Finally, test features are classified by the original classifier. For convenience of presentation, we refer the batch of test data used to update the training set as the adaptation batch and the rest of test data as the evaluation batch. In this work, we use the first 1/5 of the test data as adaptation batch and the remaining 4/5 as the evaluation batch. The subset of the adaptation trials used to calculate $\mathcal{L}(U)$ is chosen as 50% of the adaptation trials with higher posterior probabilities given by the NBPW classifier.

C. Results

Figure 1 summarizes the results of the proposed adaptation method, denoted by W_{te}, compared with the normalization approach without adapting the orthogonal part in the projection matrix, denoted by W_n. Note that all classification accuracies are based on the evaluation batch. As shown by Figure 1, for most of the subjects the proposed adaptation method yields improvements with very few drops compared to the normalization approach in [13]. Besides, the average accuracy of the proposed method using W_{te} is 68.58%, which is higher than that of using W_n, i.e., 67.50%.

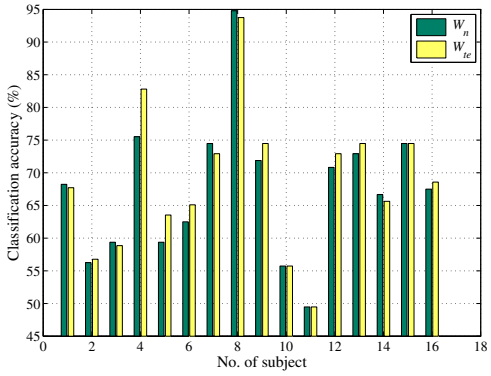
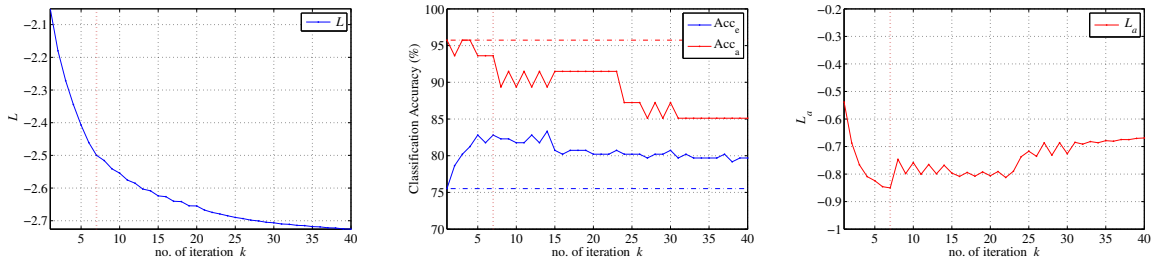


Fig. 1. Accuracy comparison

The change in $\mathcal{L}$, the classification accuracy, and $\mathcal{L}_a$ with respect to the iteration number k are shown in Figures 2 (a)-(c), respectively. In Figure 2(a), $\mathcal{L}$ decreases with respect to k until the convergence at $k = 40$. In Figure 2(b), for both the adaptation batch and the evaluation batch, the classification accuracies first increase and then decrease. If the adaptation is stopped upon the convergence of $\mathcal{L}$, the classification accuracy of neither batch is optimal. As illustrated in Figure 2(c), $\mathcal{L}_a$ decreases first and begins to increase at $k = 7$, which means that the adaptation could no longer benefit the unselected trial after $k = 7$. Thus, in the proposed method, the adaptation is stopped when $\mathcal{L}_a(k) > \mathcal{L}_a(k-1)$, i.e., $k = 7$ in this case. As shown in Figure 2(b), the classification accuracies when $k = 7$ of both batches are higher than that when $k = 40$, i.e., the convergence of $\mathcal{L}$. This shows the effectiveness of the stop criterion in the proposed adaptation design. As shown in Figure 1, the proposed method fails to increase the performance for some subjects. To investigate the underlying reason, we perform similar analysis of the change in the loss function and the classification accuracy for the subjects with little improvement in performance. We find that for subjects 10, 11 and 15 the adaptation is stopped at the very beginning of the iteration, which yields results very similar to the baseline. A possible reason for subjects 10 and 11 is that the classification accuracy of the adaptation batch is similar to that obtained purely by chance. Hence, with very few correct predicted labels it is difficult to find



(a) $\mathcal{L}$ (b) Classification accuracy (c) $\mathcal{L}_a$

Fig. 2. The x-axis represents the value of k, and the y-axis represents $\mathcal{L}$ in (a), classification accuracy in (b), and $\mathcal{L}_a$ in (c). In (a)-(c), the stop point of the iteration, $k = 7$, according to Algorithm 1 is denoted by a vertical dotted line. In (b), Acc_a and Acc_e represent the classification accuracy of adaptation batch and test batch, respectively, and the baselines of the normalization approach are denoted with dotted-dashed lines.

a right adaptation direction and the iteration stops at the beginning. The benefit of this result is that the adaptation toward a wrong direction is avoided. In our future work, we will focus on solving the problem with a better search strategy.

IV. CONCLUSIONS

This study investigates the feasibility of adapting the spatial filters based on the divergence-based framework. In the proposed adaptation method, the change of the discriminative subspaces is addressed by searching new discriminative subspaces for test data on the manifold of orthogonal matrices in a semi-supervised manner. In this way, the rotation part could be adapted along with the update of the whitening part in the spatial filters, and, a more general nonstationary case with the asymmetric data transformation could be investigated. To account for the risk in the semi-supervised learning, the adaptation trials are divided into two subsets. Only one subset is used to obtain the adaptation direction, while the search is stopped by the change of loss function of all adaptation trials. The advantage of this cross-validation-like design is to have independent validation of the adaptation and to avoid possible over-fitting. Experimental studies show that the proposed method further enhances the BCI performance based on the normalization adaptation approach. In our future work, we will improve the search strategy and investigate its performance in a sequential mode.

REFERENCES

[1] K. K. Ang, C. Guan, K. S. G. Chua, B. T. Ang, C. W. K. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, and H. Zhang, "A large clinical study on the ability of stroke patients to use EEG-based motor imagery brain-computer interface," *Clinical EEG and Neuroscience*, vol. 42, no. 4, pp. 253–258, 2011.

[2] P. von Bunau, F. C. Meinecke, F. J. Kiraly, and K.-R. Muller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, p. 214101, 2009.

[3] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughanr, "Brain-computer interface technology: A review of the first international meeting," *IEEE Transactions on Rehabilitaion Engineering*, vol. 8, no. 2, pp. 164–173, 2000.

[4] M. D. Fox, A. Z. Snyder, J. L. Vincent, and M. E. Raichle, "Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior," *Neuron*, vol. 56, pp. 171–184, 2007.

[5] F. de Pasquale, S. D. Penna, A. Z. Snyder, C. Lewis, D. Mantini, L. Marzetti, P. Belardinelli, L. Ciancetta, V. Pizzella, G. L. Romani, and M. Corbetta, "Temporal dynamics of spontaneous MEG activity in brain networks," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 6040–6045, 2010.

[6] C. Vidaurre, M. Kawanabe, P. von Bunau, B. Blankertz, and K. R. Muller, "Toward unsupervised adaptation of LDA for brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2011.

[7] X. Li, H. Zhang, C. Guan, S. H. Ong, K. K. Ang, and Y. Pan, "Discriminative learning of propagation and spatial pattern for motor imagery EEG analysis," *Neural Computation*, vol. 25, no. 10, pp. 2709–2733, 2013.

[8] Z. J. Koles, "The quantitative extraction and toporraphic mapping of the abnormal components in the clinical EEG," *Electroencephalography and Clinical Neurophysiology*, vol. 79, pp. 440–447, 1991.

[9] J. M. Gerkinga, G. Pfurtscheller, and H. Flyvbjergc, "Designing optimal spatial filters for single-trial EEG classiffication in a movement task," *Clinical Neurophysiology*, vol. 110, pp. 787–798, 1999.

[10] C. Vidaurre, A. Schlogl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 550–556, 2007.

[11] Y. Li and C. Guan, "An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces," *Neural Computation*, vol. 18, pp. 2730–2761, 2006.

[12] A. Bamdadian, C. Guan, K. K. Ang, and J. Xu, "Online semi-supervised learning with KL distance weighting for motor imagery-based BCI," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2732–2735, 2012.

[13] R. Tomioka, J. Hill, B. Blankertz, and K. Aihara, "Adapting spatial filtering methods for nonstationary BCIs," *2006 Workshop on Information-Based Induction Sciences*, pp. 65–70, Octorber 2006.

[14] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "EEG data space adaptation to reduce inter-session non-stationarity in brain-computer interface," *Neural Computation*, vol. 25, pp. 2146–2171, August 2013.

[15] W. Samek, M. Kawanabe, and K. Muller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2013.

[16] M. D. Plumbley, "Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras," *Neurocomputing*, vol. 67, pp. 161 – 197, 2005.

[17] K. K. Ang, C. Guan, C. Wang, K. S. Phua, A. H. G. Tan, and Z. Y. Chin, "Calibrating EEG-based motor imagery brain-computer interface from passive movement," *2011 Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, EMBC*, pp. 4199–4202, 2011.

[18] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, no. 39, 2012.

[19] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs," *Pattern Recognition*, vol. 45, no. 6, pp. 2137–2144, 2012.