Learning from feedback training data at a self-paced brain–computer interface

2011 J. Neural Eng. 8 046035

(http://iopscience.iop.org/1741-2552/8/4/046035)

IOP Publishing Journal of Neural Engineering

J. Neural Eng. **8** (2011) 046035 (12pp) doi:10.1088/1741-2560/8/4/046035

Learning from feedback training data at a self-paced brain–computer interface

Haihong Zhang[1], **Sidath Ravindra Liyanage**[2], **Chuanchu Wang**[1] **and Cuntai Guan**[1]

[1] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR),
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
[2] National University of Singapore, Singapore

E-mail: hhzhang@i2r.a-star.edu.sg, sidath@nus.edu.sg, ccwang@i2r.a-star.edu.sg and ctguan@i2r.a-star.edu.sg

Received 12 April 2011
Accepted for publication 22 June 2011
Published 19 July 2011
Online at stacks.iop.org/JNE/8/046035

Abstract

Inherent changes that appear in brain signals when transferring from calibration to feedback sessions are a challenging but critical issue in brain–computer interface (BCI) applications. While previous studies have mostly focused on the adaptation of classifiers, in this paper we study the feasibility and the importance of the adaptation of feature extraction in a self-paced BCI paradigm. First, we conduct calibration and feedback training on able-bodied naïve subjects using a new self-paced motor imagery BCI including the idle state. The online results suggest that the feature space constructed from calibration data may become ineffective during feedback sessions. Hence, we propose a new supervised method that learns from a feedback session to construct a more appropriate feature space, on the basis of the maximum mutual information principle between feedback signal, target signal and EEG. Specifically, we formulate the learning objective as maximizing a kernel-based mutual information estimate with respect to the spatial-spectral filtering parameters. We then derive a gradient-based optimization algorithm for the learning task. An experimental study is conducted using offline simulation. The results show that the proposed method is able to construct effective feature spaces to capture the discriminative information in feedback training data and, consequently, the prediction error can be significantly reduced using the new features.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Inherent changes in brain signals, either between calibration sessions or from calibration to feedback application, pose a critical challenge to EEG-based brain–computer interface (BCI) research [1–3] and have recently attracted a surge of attention in the field [4–12]. In particular, there has been much interest in BCIs using motor imagery (MI) [2, 13, 14]—the imagination or mental rehearsal of a motor action without any real motor output.

The underlying non-stationarity of the EEG signal accounts for many of the changes, where the distribution of electrical fields on the scalp is subject to large variations over time. The non-stationarity can be caused by shifts in background brain activities, varying mental states or individual users changing their strategy for BCI control [4]. Especially in feedback applications, more brain functions can be activated to further complicate the changes in EEG, giving rise to complex EEG phenomena such as error potentials [15] or rhythmic power shifts over the scalp [5]. Consequently, the feature extraction and prediction models (e.g. a classifier) built on data from past BCI sessions may become ineffective. Therefore, there is a strong need for new mathematical models capable of accurately predicting a user's intentions from his/her brain signals in session-to-session transfer. The adaptive BCI that can learn from new data in a supervised, semi-supervised or unsupervised manner is a viable approach to solve this problem.

So far most of the works on the adaptive BCI have focused on the adaptation of the classifiers. In [5], three supervised adaptation methods using labeled data were investigated. These included a simple bias adjustment technique, a linear discriminant analysis (LDA) retraining technique and a technique which retrains both LDA and common spatial pattern (CSP)-based feature extraction [16]. It was reported that overall the LDA-retraining approach yielded the lowest error rate. In [17], a covariance shift algorithm was introduced for an unsupervised adaptation of the linear classifier. In particular, the covariance shift algorithm is able to perform with neither labeling data nor predicting labels. In [18], the method for adaptation was further developed and combined with a bagging approach that resulted in improved stability. More recently, in [8] different types of adaptation method were extensively studied using multiple BCI datasets, and the result was in favor of a bias adjustment method rather than generic covariance shift adaptation.

Another interesting online BCI was presented in [7], where a quadratic discriminative analysis classifier was adapted in every cue-based feedback trial. It showed that the distribution of EEG features shifted significantly from one session to another. The BCI was further studied in [10]. Different from those systems using CSP features mentioned earlier, the BCI basically used adaptive autoregressive features or band powers, or the combination of the two. In [6], a classifier with band power features as input was updated continuously, where only non-feedback (i.e. calibration) sessions were used for offline study.

However, a few works have been devoted to the adaptation of feature extraction models, especially for exploring feedback training data including the idle state. As indicated in experimental results in [7] and [8], it appears that the non-stationarity may not be solved by adapting classifiers alone. Rather, possible significant brain signal changes from calibration to feedback training sessions may render the feature space derived from calibration data ineffective where little discriminative information can then be recovered.

Therefore, the primary purpose of this paper is to validate the feasibility and the importance of adapting feature extraction models, especially for a self-paced MI BCI that allows continuous feedback control [19–24]. It seems that adapting feature extraction models can be a challenging issue, in view of the unsatisfactory performance of retrained CSP models in [5].

First, we develop and test a new self-paced BCI, and study calibration and feedback training on three able-bodied naïve subjects. The empirical result poses questions on the efficacy of applying the feature space derived from calibration data to feedback sessions.

Hence, we propose a new supervised method that learns from a feedback session to construct a more appropriate feature space. In particular, the method tries to account for the underlying complex relationships between feedback signal, target signal and EEG, using a mutual information formulation. The learning objective is formulated as maximizing kernel-based mutual information estimation with respect to the spatial-spectral filters. We then derive a gradient-based optimization algorithm for the learning task.

An experimental study is conducted using offline simulation. The results show that the proposed method is capable of constructing effective feature spaces that capture more discriminative information in the feedback sessions. Consequently, the prediction errors can also be significantly reduced by using the new features.

The rest of the paper is organized as follows. Section 2 describes the data collection with a self-paced BCI, as well as the online training result. Section 3 elaborates the new method for learning effective spatial and spectral features from feedback session data. Section 4 presents an extensive analysis, followed by discussion in section 5. Section 6 concludes the paper.

2. Materials

2.1. Feedback training data collection

Three BCI-naïve adults participated as BCI subjects in the data collection. All gave informed consent, which was reviewed and approved by the Institutional Review Board of the National University of Singapore. The subjects were seated comfortably in an armchair, with their hands resting on the chair's arms or on the table in front of them. A 20 inch widescreen LCD monitor was placed on the table at a distance of approximately 1 m from the subject. The subjects were asked to remain still and comfortable to avoid movement artifacts.

EEG was recorded using a Neuroscan NuAmps 40-channel data acquisition system, with electrodes placed according to an extended international 10-20 system and a sampling frequency of 500 Hz. A total of 30 channels were used, including F7, F3, Fz, F4, F8, FT7, FC3, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O1, Oz, O2, PO1 and PO2. The reference electrode was attached to the right ear. A high-pass filter at 0.05 Hz was applied in the Neuroscan's data acquisition setting.

The subjects faced a graphic user interface displayed on the LCD monitor, as illustrated in figure 1, which guided them through the following sessions.

- *Calibration session.* This session consisted of 40 MI tasks; each was 4 s long and followed by a 6 s idle state. The MI tasks were evenly and pseudo-randomly distributed into left- and right-hand MI tasks. A graphic user interface illustrated in the left panel of figure 1 guided the subjects through the session, where a red circle in the middle served as the eye fixation point. In the background, a sequence of rectangular shapes was scrolling upward, representing left/right-hand MI tasks with blue color boxes on the left/right side and idle state tasks with gray bars. Specifically, when the red circle was in a gray color bar, the subject should relax while minimizing physical movements; otherwise, the subject should imagine left-/right-hand movement, if a blue color box was on the left/right side of the circle.

  The filter bank CSP (FBCSP) method [25–27], which was the first winner of BCI Competition IV Dataset I [28], was employed to build subject-specific MI detection

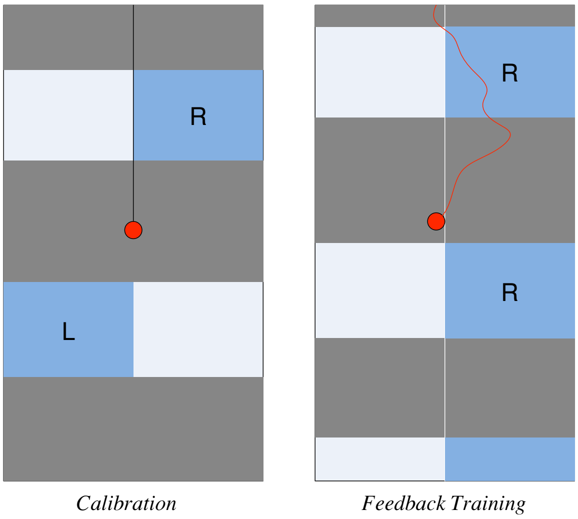

Figure 1. The graphic user interface for calibration (left panel) and for self-paced feedback training (right panel). The gray and blue color blocks scroll smoothly upward in the background, and the red circle in the center serves as the eye-fixation point. During feedback training, the horizontal position of the red circle serves as the feedback signal that updates every 40 ms, while its trajectory over the background blocks is depicted by a red curve.

models. The method constructed two separate models from the calibration data, one for differentiating between the left-hand MI and idle state (hereafter referred to as the L-model), and the other for differentiating between the right-hand MI and idle state (hereafter the R-model). For the L-model (or the R-model), each 2.5 s long shift window of EEG with a step of 0.5 s was mapped to the label of the data: 0 if the time window ended in an idle state time period, 1 (or -1) if it ended in a left-hand (or right-hand) MI period. The mapping parameters were obtained using the linear least-mean-square method.

Since a user's mental state could be uncertain and variable during the transition period from one state to another, we defined the gray region as $[-1\ 1]$ s with respect to the boundary of each idle/MI task, and excluded from FBCSP learning any EEG segments with centers in this gray region.

- *Feedback training sessions.* After calibration, each subject participated in four sessions of feedback training, i.e. two sessions of left-hand MI BCI training using the L-model and two sessions of right-hand MI training using the R-model. This arrangement allowed a subject to concentrate in each session on a particular MI task. A training session consisted of 20 MI tasks, where each lasted 5 s and was followed by a 6 s idle state. A graphical user interface illustrated on the right panel of figure 1 guided the user through the session. The meaning of the graph was similar to that for calibration, except that the red circle was moving horizontally as a feedback signal: its horizontal position was determined by the FBCSP output updated every 40 ms.

  During the feedback training, the subjects tried to move the red circle to the left/right side as far as possible during left-hand/right-hand MI tasks. We would like to emphasize that the subjects were requested not to voluntarily control the feedback signal by any means during periods of the idle state. This is because voluntary control of the feedback signal would spoil the idle state data.

In between sessions were short breaks. The first feedback training session started within 5 min after the calibration session. And the interval between consecutive feedback sessions was from 1 to 5 min. Note that a special tryout session was in place after the calibration, where every subject tried online feedback for a short while so as to get a feeling for the feedback and also to prepare for the actual training sessions. The tryout session was not included in the analysis.

We would like to briefly introduce the FBCSP method used in the online experiment as it will also be compared with the proposed learning method later. FBCSP was introduced in [25] as a feature selection algorithm that combines a filter bank framework with the spatial filtering technique CSP. More specifically, it decomposes EEG data into an array of pass-bands, performs CSP in each band and selects a reduced set of features from all the bands. Its efficacy was demonstrated in the latest BCI Competition [28], where it served as the basis of all the winning algorithms in the EEG categories. FBCSP was improved in [26] by employing a robust maximum mutual information criterion for feature selection.

2.2. *Data screening*

The recorded EEG data during feedback training sessions were inspected visually using MATLAB by the authors. Any EEG segments identified with EOG and EMG contamination [29] were rejected and excluded from the analysis. Again, we defined the gray regions in a similar way to the calibration method described above. Therefore, any EEG segments centered within $[-1\ 1]$ s with respect to any task boundary were excluded from the analysis.

2.3. *Online performance and initial data analysis*

Online performance was assessed using the mean-square-error (MSE) measure between the feedback signal and the target signal. Figure 2 plots the bar graph of MSE in each feedback training session. The error was apparently comparable between the first and the second training session in most cases. This actually indicates that online feedback training in the BCI can be a difficult task since it was anticipated that the subjects should have gained better control of the BCI over training sessions. Again, this indicates the necessity of adapting models during session-to-session transfers.

To further understand the feedback training data, we plot in figure 3 the distribution of EEG feature vector samples produced by FBCSP. Note that for clarity of presentation, we used evenly re-sampled feature vector samples because the original samples count up to thousands. As expected, the MI class samples and the idle class samples were easily separable in the calibration data, but the discriminative information had disappeared in the same feature space in most feedback training sessions. As a consequence, either there was no

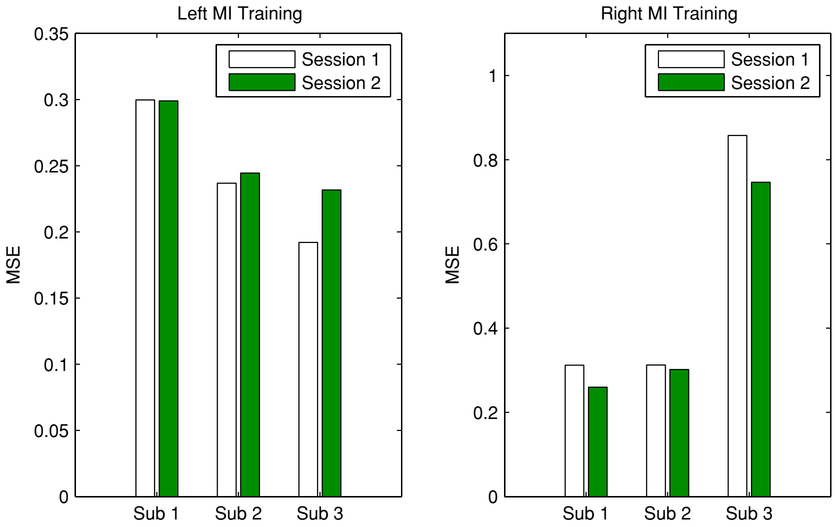

Figure 2. Online performance of subjects in terms of MSE between the feedback signal and the target. There is a strong bias shift (from calibration to feedback) in right MI sessions in subject 3, which explains his particularly large error.

effective separation between the two classes, or the separation hyper-plane was severely altered (similar to some cases in [7, 8]).

Therefore, it is advisable to first look into the issue of ineffective feature space before trying to adapt a classifier/regressor. To address this issue, we propose a new method to learn an effective feature space from feedback data. We would also like to note that, compared with calibration data, online feedback training data pose more challenges to effective feature extraction, because the feedback may involve more brain functions and produce more complex EEG phenomena [5, 15].

3. The new learning method

3.1. *Spatio-spectral features*

The primary phenomenon of MI EEG is event-related desynchronization (ERD) or event-related synchronization (ERS) [2, 13], the attenuation or increase of the rhythmic activity over the sensorimotor cortex generally in the μ (8–14 Hz) and β (14–30 Hz) rhythms. The ERD/ERS can be induced by both imagined movements in healthy people or intended movements in paralyzed patients [21, 30, 31]. It is noteworthy that another neurological phenomenon called the Bereitschaftspotential is also associated with MI EEG but is non-oscillatory [14]. In this paper, we consider ERD/ERS features only.

The feature extraction of ERD/ERS is, however, a challenging task due to its poor low signal-to-noise ratio. Therefore, spatial filtering in conjunction with frequency selection (via processing in either temporal domain or spectral domain) in multi-channel EEG has been highly successful in increasing the signal-to-noise ratio [16, 27, 32–34].

Let us consider the spatial-spectral filtering in the spectral domain, where each n_c-channel EEG segment with a sampling rate of F_s Hz can be described by an $n_c \times n_f$ matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n_f} \\ \vdots & \ddots & \vdots \\ x_{n_c1} & \cdots & x_{n_cn_f} \end{bmatrix}, \tag{1}$$

where x_{ij} denotes the discrete Fourier transform of the ith channel at frequency $\omega_j = \frac{j-1}{2n_f}F_s$.

A joint spatial-spectral filter on $\mathbf{X}$ can be essentially represented by a spatial filtering vector $\mathbf{w} \in \mathbb{R}^{n_c\times 1}$ and a spectral filter vector $\mathbf{f} \in \mathbb{R}^{n_f\times 1}$. The feature y_0 is the energy of the EEG segment after filtering:

$$y_0 = \mathrm{diag}\{\widetilde{\mathbf{w}^T\mathbf{X}}(\mathbf{w}^T\mathbf{X})\}\mathbf{f}, \tag{2}$$

where the wave line $\tilde{}$ on the right-hand side denotes the conjugate of a complex value and the diag( ) function stands for the diagonal vector of a matrix.

In this paper, we consider a general case in which multiple spatial filters are associated with one particular spectral filter. Therefore, the feature extraction model is determined by the matrix $\mathbf{f}$ and a vector $\mathbf{W}$, the latter being the collection of spatial filters in columns:

$$\mathbf{W} = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_{n_w}]. \tag{3}$$

Suppose the spectral filters in $\mathbf{F}$ are given (see the last paragraph of section 3.3 for details), we can use the following shorthand for the auto-correlation matrix of EEG processed by the kth spectral filter:

$$\hat{\mathbf{X}}_k = \sum_i^{n_f} f_i \mathbf{X}\tilde{\mathbf{X}}, \tag{4}$$

and express the logarithmic feature vector by

$$\mathbf{y} = \left[\log\left(\mathbf{w}_1\hat{\mathbf{X}}_1\mathbf{w}_1^T\right), \ldots, \log\left(\mathbf{w}_{n_w}\hat{\mathbf{X}}_{n_w}\mathbf{w}_{n_w}^T\right)\right]^T. \tag{5}$$

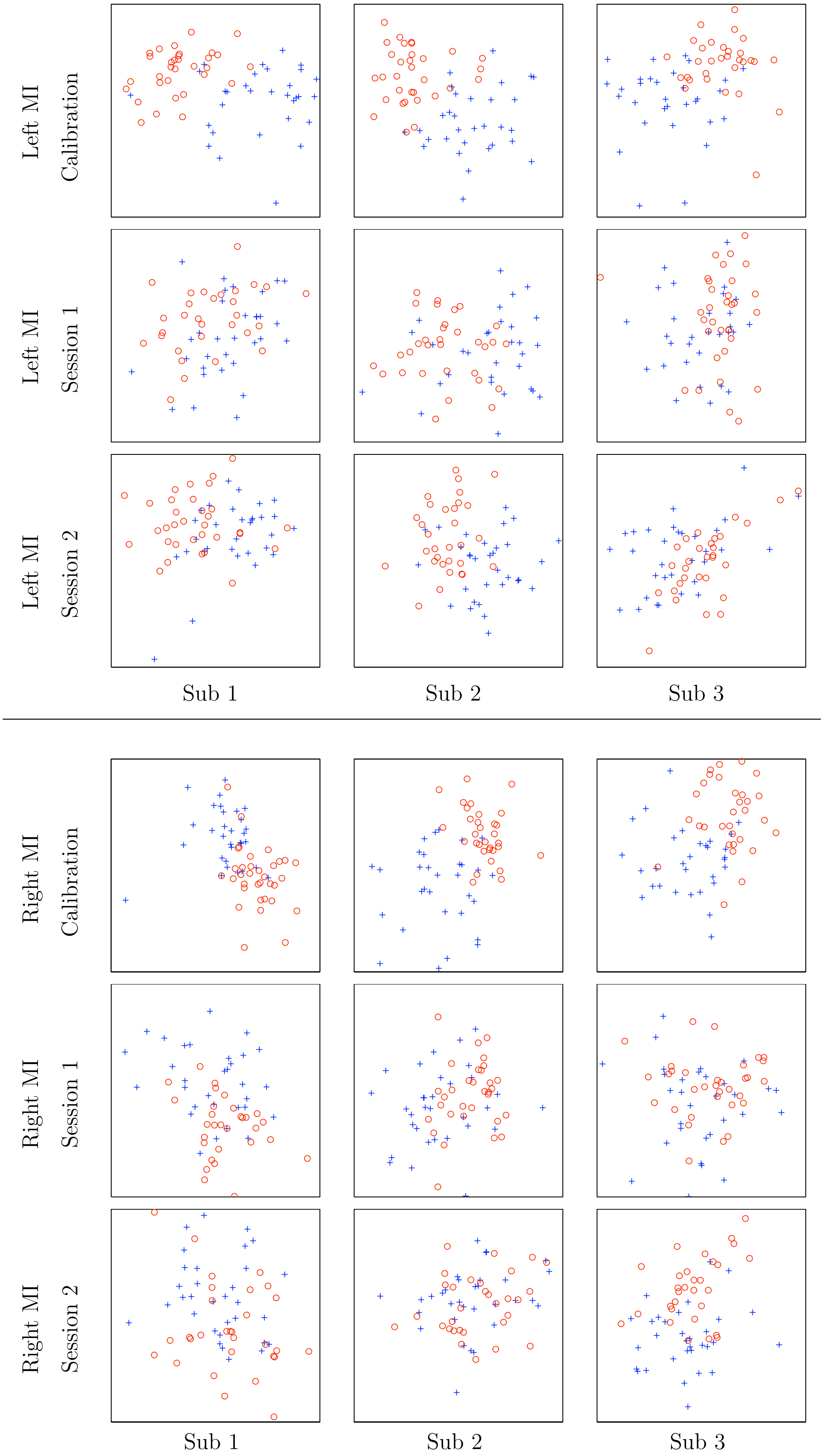

Figure 3. Feature distributions during MI calibration and feedback training sessions, for left MI in the upper three rows and for right MI in the lower three rows. The horizontal axis and the vertical axis are the first and the second FBCSP features. The axis range is made consistent in each column (i.e. each subject). The red circles represent MI samples, while the black crosses denote idle state samples. Note the especially significant change in the distribution of MI samples.

3.2. Formulation of the objective function for learning

To capture the underlying complex structure of spatio-spectral data in ERD/ERS, we would like to design a mutual information-based objective function for learning **W** and **F**. Mutual information [35], which stemmed from information theory, basically measures the reduction of uncertainty about class labels due to the knowledge of the features. Readers interested in mutual information-based feature extraction/selection may find related works in [36–41].

For feedback training data, we consider a mutual information measure $\hat{I}$ between the class labels and the EEG features as well as the feedback signal. Specifically, mutual information is between the class label (i.e. the variable to be predicted) and the observations, including both the feedback signal and the EEG feature vector. Let the random variables of the label, the EEG feature vector and the feedback signal be $\mathcal{C}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively. There is

$$\hat{I}(\{\mathcal{Y},\mathcal{Z}\},\mathcal{C}) = \hat{H}(\mathcal{Y},\mathcal{Z}) - \sum_{c} P(c)\hat{H}(\mathcal{Y},\mathcal{Z}|c), \tag{6}$$

where $\hat{H}$ denotes the entropy measure of a random variable.

Like [39, 41], we resort to a non-parametric approach for mutual information estimation since it does not rely on the underlying distributions.

Suppose the feedback training data comprise l samples of EEG to be represented by the feature vectors $\mathbf{y}_i$s and the concurrent feedback signal z_is ($i \in [1,\ldots,l]$). The non-parametric approach computes each entropy in equation (6) separately, e.g. $\hat{H}(\mathcal{Y},\mathcal{Z})$ by

$$\hat{H}(\mathcal{Y},\mathcal{Z}) = -\frac{1}{l}\sum_{i=1}^{l}\log\left\{\frac{1}{l}\sum_{j=1}^{l}\varphi_y(\mathbf{y}_i,\mathbf{y}_j)\varphi_z(z_i,z_j)\right\}, \tag{7}$$

and φ_y and φ_z are kernel functions and usually take a Gaussian form. For example,

$$\varphi(\mathbf{y},\mathbf{y}_i) = \alpha\exp\left(-\tfrac{1}{2}(\mathbf{y}-\mathbf{y}_i)^T\Psi^{-1}(\mathbf{y}-\mathbf{y}_i)\right). \tag{8}$$

The coefficient α is discarded hereafter because it will be canceled out when equation (8) is substituted into equation (7) and then substituted into equation (6). It should be noted that the kernel size matrix Ψ is diagonal, and each diagonal element is determined by

$$\psi_{k,k} = \zeta\frac{1}{l-1}\sum_{i=1}^{l}(y_{ik}-\bar{y}_k)^2, \tag{9}$$

where $\bar{\mathbf{y}}_k$ is the empirical mean of $\mathbf{y}_k$, and we set the coefficient $\zeta = \left(\frac{4}{3l}\right)^{0.1}$ according to the normal optimal smoothing strategy [42].

The conditional entropy $\hat{H}(\mathcal{Y}|c)$ in equation (6) can also be estimated similar to equation (7), but using samples from class-c only.

Using the maximum mutual information principle [36], we now define the learning task as searching for the optimum spatial and spectral filters **W** and **F** that satisfies

$$\{\mathbf{W},\mathbf{F}\}_{\text{opt}} = \underset{\{\mathbf{W},\mathbf{F}\}}{\operatorname{argmax}}\,\hat{I}(\{\mathcal{Y},\mathcal{Z}\},\mathcal{C}). \tag{10}$$

The above formulation describes the inter-dependency between the target signal, the feedback signal and the EEG signal as a function over the feature extraction parameters in spatial-spectral filters. It basically aims to maximize the information about the target signal to be predicted, contained in the extracted features in conjunction with feedback. Refer to section 5 for a further discussion on this formulation.

3.3. Gradient-based solution to the learning problem

Here, we propose a numerical solution to equation (10) by devising a gradient-based optimization algorithm. We consider a spatial filter vector $\mathbf{w}_k$, and note that the gradient of the objective function $\hat{I}$ with respect to $\mathbf{w}_k$ is

$$\nabla_{\mathbf{w}_k}\hat{I}(\{\mathcal{Y},\mathcal{Z}\},\mathcal{C}) = \nabla_{\mathbf{w}_k}\hat{H}(\mathcal{Y},\mathcal{Z}) - \sum_{c\in\mathcal{C}} P(c)\nabla_{\mathbf{w}_k}\hat{H}(\mathcal{Y},\mathcal{Z}|c). \tag{11}$$

From equation (7), we have

$$\nabla_{\mathbf{w}_k}\hat{H}(\mathcal{Y},\mathcal{Z}) = -\frac{1}{l}\sum_{i=1}^{l}\beta_i\frac{1}{l}\sum_{j=1}^{l}\varphi_z(z_i,z_j)\frac{\partial\varphi_y(\mathbf{y}_i,\mathbf{y}_j)}{\partial\mathbf{w}_k}, \tag{12}$$

where

$$\beta_i = \left(\frac{1}{l}\sum_{j=1}^{l}\varphi_z(z_i,z_j)\varphi_y(\mathbf{y}_i,\mathbf{y}_j)\right)^{-1}. \tag{13}$$

From equation (8), we have

$$\frac{\partial\varphi_y(\mathbf{y}_i,\mathbf{y}_j)}{\partial\mathbf{w}_k} = -\frac{1}{2}\varphi_y(\mathbf{y}_i,\mathbf{y}_j)\frac{\partial(\mathbf{y}_i-\mathbf{y}_j)^T\Psi^{-1}(\mathbf{y}_i-\mathbf{y}_j)}{\partial\mathbf{w}_k}. \tag{14}$$

Let us denote the quadratic function $(\mathbf{y}_i-\mathbf{y}_j)^T\Psi^{-1}(\mathbf{y}_i-\mathbf{y}_j)$ by ϑ_{ij}, which can be further decomposed into

$$\vartheta_{ij} = \sum_{k_1=1}^{d_o}\sum_{k_2=1}^{d_o}\psi_{k_1k_2}^{-1}(y_{ik_1}-y_{jk_1})(y_{ik_2}-y_{jk_2}). \tag{15}$$

Hence, the gradient of ϑ_{ij} is

$$\frac{\partial\vartheta_{ij}}{\partial\mathbf{w}_k} = \sum_{k_1=1}^{d_o}\sum_{k_2=1}^{d_o}\left[\frac{\partial\psi_{k_1k_2}^{-1}}{\partial\mathbf{w}_k}(y_{ik_1}-y_{jk_1})(y_{ik_2}-y_{jk_2}) + \psi_{k_1k_2}^{-1}\frac{\partial(y_{ik_1}-y_{jk_1})(y_{ik_2}-y_{jk_2})}{\partial\mathbf{w}_k}\right]. \tag{16}$$

Consider that $(y_{ik_1}-y_{jk_2})^2$ is a function of $\mathbf{w}_k$ if and only if $k_1 = k$ and/or $k_2 = k$, and $\psi_{k_1k_2}^{-1}$ is a function of $\mathbf{w}_k$ if and only if $k_1 = k_2 = k$. Furthermore, $\psi_{k_1k_2}^{-1} = 0$ only if $k_1 \neq k$ or $k_2 \neq k$. The expression of the gradient above can be written as

$$\frac{\partial\vartheta_{ij}}{\partial\mathbf{w}_k} = \frac{\partial\psi_{kk}^{-1}}{\partial\mathbf{w}_k}(y_{ik}-y_{jk})^2 + \psi_{kk}^{-1}\frac{\partial(y_{ik}-y_{jk})^2}{\partial\mathbf{w}_k}. \tag{17}$$

From equation (9), we have

$$\frac{\partial\psi_{k,k}^{-1}}{\partial\mathbf{w}_k} = -\frac{2\zeta}{\psi_{k,k}^2(l-1)}\sum_{i'=1}^{l}(y_{i'k}-\bar{y}_k)\frac{\partial\,(y_{i'k}-\bar{y}_k)}{\partial\mathbf{w}_k}, \tag{18}$$

where $\bar{y}_k$ denotes the mean value of $y_{i'k}$s, and its partial derivative w.r.t. $\mathbf{w}_k$ can be expressed by

$$\frac{\partial \bar{y}_k}{\partial \mathbf{w}_k} = \frac{1}{l}\sum_{i''}^{l} \frac{\partial y_{i''k}}{\partial \mathbf{w}_k}. \tag{19}$$

We further note that $\hat{\mathbf{X}}_{ki}$ (the auto-correlation matrix for the *i*th EEG sample processed by the *k*th spectral filter, see equation (4)) is conjugate symmetric, and

$$\frac{\partial y_{ik}}{\partial \mathbf{w}_k} = \frac{\left(\hat{\mathbf{X}}_{ki} + \hat{\mathbf{X}}_{ki}^T\right)\mathbf{w}_k}{y_{ik}} = \frac{2\,\mathrm{Re}(\hat{\mathbf{X}}_{ki})\mathbf{w}_k}{y_{ik}}, \tag{20}$$

where Re( ) denotes the real part of a complex matrix. The derivatives of $y_{i'k}$ and y_{jk} can be computed the same way as above.

We can summarize the above steps as follows:

$$\nabla_{\mathbf{w}_k} \hat{H}(\mathcal{Y}) = \mathbf{A}\mathbf{w}_k, \tag{21}$$

where

$$\begin{aligned}\mathbf{A} = \frac{2}{l^2}\sum_{i=1}^{l}\beta_i \sum_{j=1}^{l}\varphi_z(z_i, z_j)\varphi_y(\mathbf{y}_i, \mathbf{y}_j)\Bigg[&\frac{-\zeta(y_{ik} - y_{jk})^2}{\psi_{k,k}^2(l-1)} \\ &\times \sum_{i'=1}^{l}(y_{i'k} - \bar{y}_k)\left(\frac{\mathrm{Re}(\hat{\mathbf{X}}_{ki'})}{y_{i'k}} - \frac{1}{l}\sum_{i''}^{l}\frac{\mathrm{Re}(\hat{\mathbf{X}}_{ki''})}{y_{i''k}}\right) \\ &+ \psi_{kk}^{-1}(y_{ik} - y_{jk})\left(\frac{\mathrm{Re}(\hat{\mathbf{X}}_{ki})}{y_{ik}} - \frac{\mathrm{Re}(\hat{\mathbf{X}}_{kj})}{y_{jk}}\right)\Bigg].\end{aligned} \tag{22}$$

There will be, for each conditional entropy $\hat{H}(\mathcal{Y}|c)$, an equation similar to equation (21). Then the gradient of the objective function **I** with respect to the spatial filter $\mathbf{w}_k$ is

$$\nabla_{\mathbf{w}_k}\hat{I}(\{\mathcal{Y}, \mathcal{Z}\}, \mathcal{C}) = \left(\mathbf{A} - \sum_{c} P(c)\mathbf{A}_c\right)\mathbf{w}_k. \tag{23}$$

We would like to note that the above equation does not suggest that the gradient is a linear function over $\mathbf{w}_k$ since the multiplier term $\left(\mathbf{A} - \sum_c P(c)\mathbf{A}_c\right)$ itself is a rather complicated function over $\{\mathbf{y}_i\}$, which in turn is a function of **W**.

With the gradient information, our iterative optimization algorithm updates a spatial filter by

$$\mathbf{w}_k^{(\mathrm{iter}+1)} = \mathbf{w}_k^{(\mathrm{iter})} + \lambda\nabla_{\mathbf{w}_k}\hat{I}(\{\mathcal{Y}^{(\mathrm{iter})}, \mathcal{Z}\}, \mathcal{C}), \tag{24}$$

where λ is the step size. In this paper, we utilize a line search procedure to determine the step size in each of the iteration. Note that all spatial filter vectors in **W** are updated together.

In our implementation, the line search procedure tests a number of (tentatively 16) λ values in the range of $[-0.05\ 0.10]\times\xi$, and decreases ξ in the logarithm scale until a local maximum of **I** is found but not at $\lambda = 0$. The λ for the local maximum is then used to update all the spatial filters $\mathbf{w}_k$s in equation (24), and then the optimization procedure proceeds to the next iteration.

The iterations will terminate when a convergence criterion is met. In this paper, we use a simple criterion: mutual information gain less than 1×10^{-5}. Since the iterative algorithm is a typical gradient-based greedy optimization method, the pseudo-code is omitted to save space.

The initial values for $\mathbf{w}_k$ can be learned by the CSP method [16] that maximizes the Rayleigh coefficient

$$\frac{\mathbf{w}_k \sum_{i=1}^{l_1}\hat{\mathbf{X}}_{ki}\mathbf{w}_k}{\mathbf{w}_k \sum_{j=1}^{l_0}\hat{\mathbf{X}}_{kj}\mathbf{w}_k}, \tag{25}$$

where $\hat{\mathbf{X}}_{ki}$ denotes the *i*th sample of MI EEG while $\hat{\mathbf{X}}_{kj}$ the *j*th sample of idle state EEG.

Finally, we describe how to select the spectral filters for **F**. Like FBCSP, we can also create a set of candidate spectral filters consisting of band-pass filters that cover the MI EEG spectrum. For instance, in the experimental study to be introduced in the following section, we borrowed the filter banks configuration from [26] that had eight band-pass filters with central frequency ranging from 4 to 32 Hz. After band-pass filtering in the spectral domain, we trained CSP according to equation (25) to extract discriminative energy features. Then we selected the optimum n_w features from all, using the method in [26]. The spectral filters associated with the optimum features then comprised the matrix **F**.

4. Results

We conducted an offline simulation of the self-paced BCI using the online feedback training data. The simulation was carried out in MATLAB, and the proposed method was implemented in hybrid MATLAB and C code so as to improve computation and programming efficiency. The EEG features together with the feedback signal z served the inputs to a regressor (refer to section 5 for a related discussion), in order to predict the target value of 0 (idle state), −1 (right-hand MI) or 1 (left-hand MI). We employed a linear support vector regression using the LibSVM toolbox [43]. Note that we had attempted other regression methods such as Gaussian-kernel nonlinear support vector regression and linear MSE regression. But no significant difference was found in the results, and we will only show the linear support vector regression results here.

Similar to the online feedback training described in section 2, the offline simulation tested left-hand MI BCI and right-hand MI BCI separately. For example, for the left-hand MI BCI, the first left-hand MI training session was used to learn the optimum spatial-spectral filtering and then the linear support vector regressor was trained. Next, the feature extraction and regression was tested on the second left-hand MI training session. The simulation used a 2 s long shift window with a step of 0.4 s.

For a comparative analysis with the state of the art, we also tested FBCSP using the same setting.

4.1. Convergence of the optimization algorithm

We studied the convergence of the optimization algorithm. First, we considered a simple scenario that only included three EEG channels (CP3, CPz, CP4) and one spatial filter. We would like to note that similar findings were also obtained in our extensive tests that used different selection of channels around the sensorimotor cortex regions, e.g. C3, Cz, C4.

Since the mutual information measure is always invariant to the non-zero norm of the spatial filter, we set the norm of

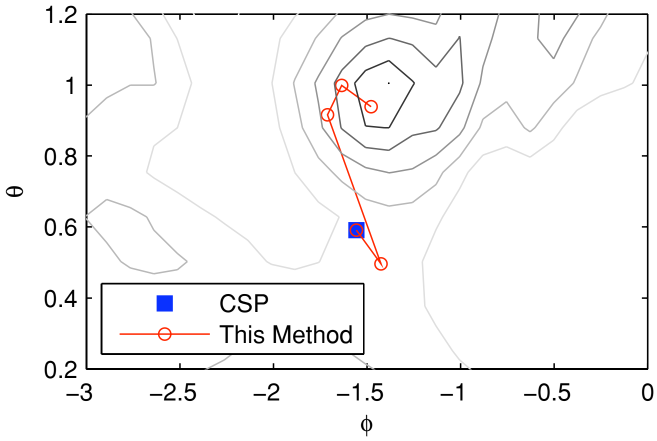

Figure 4. Optimization on the mutual information surface: an example with a spatial filter vector for three-channel EEG. See section 4.1 for details.

the spatial filter to 1 without loss of generality. Therefore, the spatial filter can be represented by two variables in the spherical coordinate system: $\theta = \text{acos}(w_3)$ and $\phi = \text{atan}(\frac{w_2}{w_1})$. This should not be confused with the Euclidean space where the actual optimization takes place. The two-variable representation is just meant for visualization.

Figure 4 shows a typical example from the left-hand MI learning in subject 2. The spatial filter solution migrated in four steps from the initial point (generated by CSP) to approximately a local maximum where the iteration converged (mutual information gain $<1\times10^{-5}$).

The algorithm was initialized using the method described in the previous section, and then in most cases the optimization algorithm converged within seven iterations. We also tested random spatial filters for initialization, and the iteration procedure generally became longer but converged within 50 iterations in all 100 test runs.

4.2. Feature distributions

We used the first feedback training session to learn two spatial-spectral filters by the proposed method and extracted EEG features from the second feedback session. Figure 5 plots the distribution of the features (as the original samples amount to thousands, we used evenly re-sampled feature vector samples for a clear presentation).

Comparing with those features produced by calibration models in figure 3 (especially in the bottom row for the same training session), the new features appear to be more separable between the MI classes and the idle states. To verify this, we assess the separability in terms of classification accuracy by a linear support vector machine (using the same LibSVM toolbox from [43]). The results for the original and the new features are compared in table 1.

The table clearly indicates that the proposed method, which adapted both the classifier and the feature extraction model, produced significantly better performance in terms of class separability than when only the classifier was adapted. This verifies our argument in the introduction that the non-stationarity in EEG may not be solved by adapting classifiers alone. Rather, it is advisable to adapt both the feature extraction model and the classifier so as to accurately capture the variation of EEG over time.

Table 1. Class separability: new feature space ('This method') versus original feature space ('Original'). Class separability is measured as the classification accuracy by a linear support vector machine that is adapted to the data (feedback training session 2). Note 'Original' uses the adaptation of the classifier only, while 'This method' adapts both the classifier and the feature extraction model. The higher accuracy rates between the two feature spaces are shown in bold. See section 4.2 for related description.

	Features	Sub 1	Sub 2	Sub 3
Left MI	Original	73.7%	79.0%	66.9%
	This method	**85.0%**	**84.8%**	**81.0%**
Right MI	Original	67.9%	59.7%	78.1%
	This method	**80.0%**	**69.6%**	**84.0%**

4.3. Accuracy of feedback control prediction

We investigate whether the new features can generate better prediction of the user's state. We would also like to test the adaptation of the regressor since the classification hyper-plane may have shifted from the first feedback session to the second. Therefore, we tested a supervised adaptation, which used a portion (called adaptation data which started from the beginning of the session) of the second feedback session, and re-trained the regressor (using both the adaptation data and the first feedback session data); we also tested the models on the remainder of the second feedback session. We examined different sizes for the adaptation data in terms of the percentage of the whole session, ranging from 0 (i.e. no adaptation) to 0.45.

FBCSP was also evaluated using the same method for comparison. And the comparative results are illustrated in figure 6. Apparently, both FBCSP and the proposed method can learn a much more accurate predictor from the first feedback session than the original BCI that used only the calibration data. Furthermore, the prediction error was also effectively reduced by the supervised adaptation. But, this improvement is not as significant as the improvement observed from the original BCI to the proposed method. Furthermore, the proposed method also consistently outperformed FBCSP, significantly in most cases.

We examined the impact of the new method on the feedback signal curves. Figure 7 illustrates a graph comparing the new feedback signal to the original feedback signal, for subject 2. Clearly, the new feedback signal curve followed the target curve much more accurately.

We also investigated if the new method works with a reduced set of channels. In particular, we tested 15, 9 and 6 channels (see table 2 for the channel names), and carried out the proposed method and FBCSP, using the same method as described above (see figure 8), and performed the *t*-test to check if our method produced lower MSE with statistical significance compared with FBCSP and the original feedback training result.

The result indicates that the new method improved the performance in terms of MSE with statistical significance in

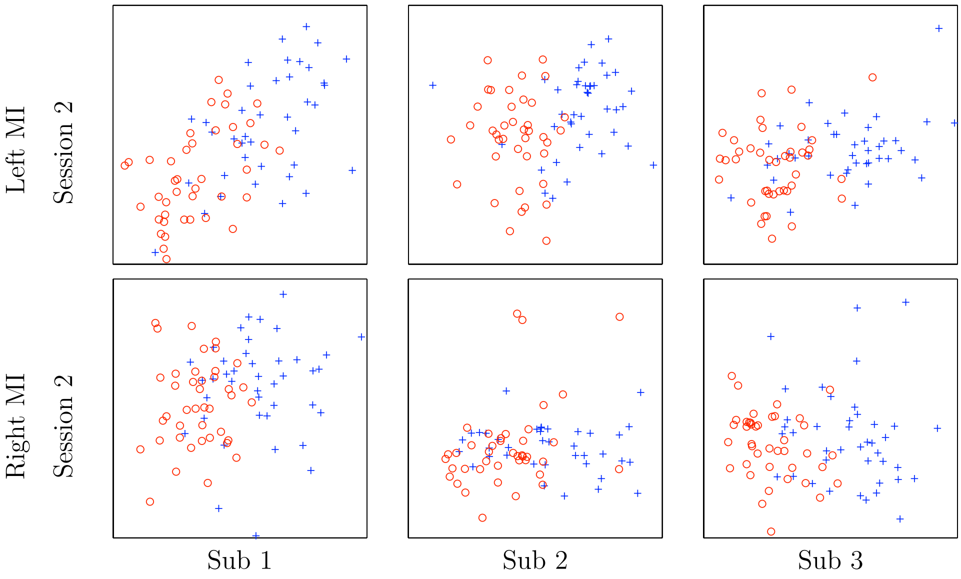

Figure 5. Feature distributions by the proposed learning method for the left/right MI feedback training session 2. The horizontal axis and the vertical axis are, respectively, the first and the second features learned by the learning method. The graphs in the upper row are generated from left MI training data, while those in the lower row are from right MI training data. The red circles represent MI samples, while the black crosses denote idle state samples. See figure 3 (especially the bottom row for the same session) for a comparison.

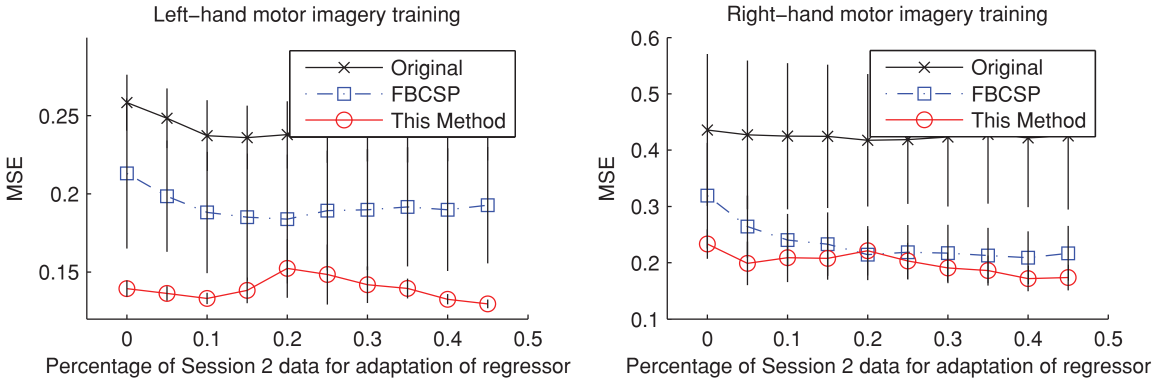

Figure 6. Comparison of the prediction error in terms of MSE by different methods. The horizontal axis denotes the percentage of the second feedback session being used for re-training the support vector regression machine that maps EEG features to the target signal. For the original online feedback, there is no re-training but MSE is computed at each percentage point using the same test set. The test set is the second feedback session excluding the part for regressor re-training. The curves plot the average of MSE over the three subjects, while the vertical line centered at each point represents the standard deviation by its length. See section 4.3 for related description.

Table 2. Statistical paired *t*-test (*p*-value shown here) of comparing the new method's MSE with that of FBCSP or the original feedback training result, using a different number of channels. Significant results with *p*-value <0.05 are shown in bold.

		p-value		
No of channels	Data	This versus FBCSP	This versus Original	Channel names
All	Left MI Right MI	<**0.01** <**0.04**	<**0.01** <**0.01**	All 30 channels (see section 2)
15	Left MI Right MI	<**0.01** 0.09	<**0.01** <**0.01**	F3,F4,FC3,FCz,FC4,T3,Cz, C4,T4,CP3,CPz,CP4,P3,P4
9	Left MI Right MI	<**0.01** 0.86	<**0.01** <**0.01**	FC3,FCz,FC4,C3,Cz,C4,CP3, CPz,CP4
6	Left MI Right MI	0.48 0.93	<**0.01** <**0.01**	FC3,FC4,C3,C4,CP3,CP4

all the channel sets being tested. While if we compare the new method with FBCSP, it still yielded significant lower MSE with as few as nine channels. In the case of six channels, the method and FBCSP produced comparable results, while both significantly outperformed the original model constructed from calibration only.

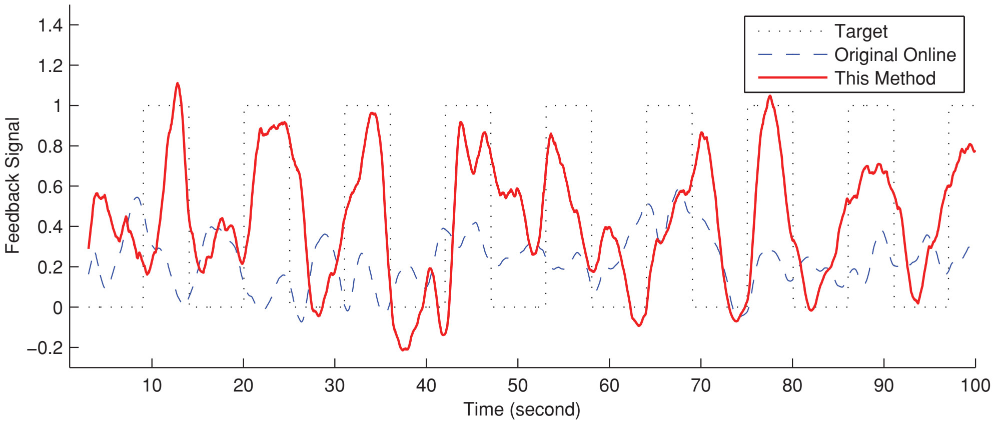

Figure 7. Comparison between the target, the original feedback signal and the new prediction by the proposed method. Here is an example from the left MI training session of subject 2. The timing is in alternation between approximately 5 s MI (*target* = 1) and 6 s idle state (*target* = 0), except the first idle state period, which is slightly longer.

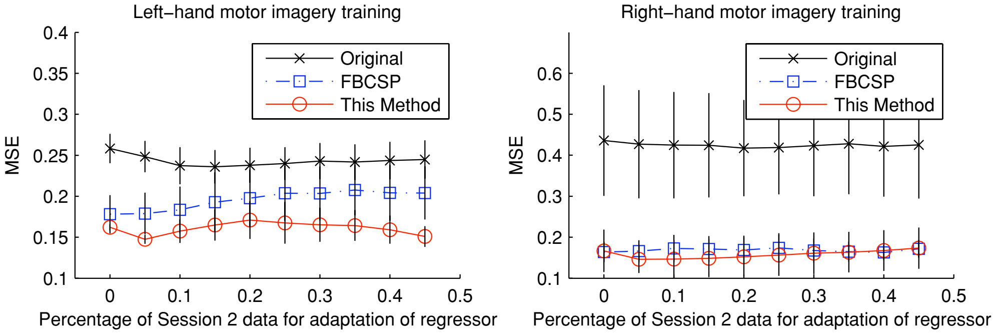

Figure 8. Comparison of the prediction error in MSE by different methods using nine EEG channels only. See figure 6 and section 4.3 for descriptions.

5. Discussions

Figure 6 gives clear evidence that the proposed method of using the new spatial-spectral learning algorithm can significantly increase the prediction accuracy. The MSE for left (or right) MI feedback training was effectively reduced from approximately 0.3 (or 0.5) to a slightly lesser value of 0.2 (or 0.25). The improved accuracy can also be seen in the prediction curves in the example case of figure 7, which actually showcases a reduction of MSE from 0.24 to 0.13.

The increased accuracy can be largely attributed to the improved feature space shown in figure 5 in contrast to the original feature spaces in figure 3. The original feature space that was used in feedback training was built using the calibration data. The changes in feature distributions in the original feature space have highlighted the effect of session-to-session transfer, which is generally consistent with prior studies on the adaptive BCI. Thus, during feedback sessions, the MI EEG and idle state EEG were predominantly non-separable. Even if they were separable it was subject to distribution shift. On the other hand, the new feature space was learned from feedback training data comprised of three sources of information; namely, EEG, target signal and feedback signal. Therefore, it has been able to capture essential information for user state prediction during online feedback training.

It is also worthwhile to mention again that the new model uses a non-parametric formulation for learning, which aims to account for arbitrary dependences among EEG, target and feedback signals. Section 4.1 has shown that our optimization algorithm, derived through the new formulation, has good convergence properties. Figure 4 has shown that the objective function surface for the three-channel EEG data is smooth, which is a favorable condition for the greedy algorithm. However, we expect that the mutual information surface can become far more complicated, especially for EEG data with a large number of channels. Therefore, future research may investigate more advanced optimization techniques. However, such techniques would usually incur much heavier computational costs.

While this paper has focused on the development and validation of a new learning method for the adaptive BCI, it would be interesting to investigate its performance during online training. Even though it is beyond the scope of this paper, it is within the scope of our ongoing research. Generally, a large number of subjects would be required in order to draw statistically significant comparisons between adaptive and non-adaptive BCI systems.

It is also interesting to look back into the formulation of objective formulation in section 3.2. As stated earlier, the goal is to maximize the information about the target signal to be predicted, contained in the EEG features in conjunction with the feedback. It is therefore advisable to include both the new EEG features and the prediction outputs of the current model as inputs to the classifier or regression machine in the new model. Importantly, the feedback serves two purposes: not only does it serve as a visual 'stimulus' to the subject, but it also represents the current prediction model that contains essential information extracted from earlier calibration/feedback sessions. The first rationale is that the feedback and its relative position to the target signal may have an effect on brain activations to complicate MI EEG. The second function gives rise to multiple implications, as explained below. First, the formulation considers only the output of the current BCI model, but not the internal mechanism of the model. Thus, it can work with any BCI model and adapt it during new feedback training sessions. Secondly, if a user with a prediction model can control the feedback signal to match the target signal satisfactorily during a feedback session, further re-adaptation of the prediction model can be unnecessary, as co-adaption of the user and machine has already been achieved. This can also be viewed as a special case of the objective function (10): if the feedback variable $\mathcal{Z}$ in the objective function already carries essential information about the target signal $\mathcal{C}$, re-adaptation of the BCI by including new EEG features would produce no significant gain in the objective function.

We would like to emphasize again that the proposed method works in a supervised learning fashion. In other words, it requires the data labels for adaptative learning. Unlike the unsupervised or semi-supervised online learning approach, this enables the learning system to measure the compliance of a subject to the BCI tasks, so as to ensure the stability of the adaptation process.

The proposed method with the current solution may be better suited for offline adaptation than for online adaptation. In online adaptation both user training and machine adaptation take place at the same time. While in offline adaptation, machine adaptation is performed after the user finishes a training session. Although this method is applicable to online adaptation, the expensive computation can be a serious concern for practical online use. We estimate that the computational complexity of computing the gradient by equations (23) and (22) is of the order of $O\left(l^2 n_c^2\right)$ and that of evaluating the objective function by equations (7) and (6) is $O\left(l^2 n_c\right)$. Here l denotes the number of samples and n_c the number of channels. In our experimental setup for the results presented in section 4, we implemented a learning code using hybrid MATLAB and C coding without multi-threading. On our test computer with a Xeon CPU at 2.93 GHz, the code took approximately 130 s to complete one iteration for $n_c = 30$-channel EEG data, or 18 s for $n_c = 6$-channel EEG data, both of $l = 2230$ time segment samples. The primary cause of the high computational complexity is the non-parametric (kernel-based) nature of the method that requires computation in each pair of samples. Therefore, a possible solution to this problem will be to reduce the number of samples for adaptation but without losing useful information.

6. Conclusion

In this paper we have studied and addressed the critical issue of session-to-session transfer at a brain–computer interface (BCI). While previous studies have often focused on the adaptation of classifiers, we have shown the importance of and the feasibility of adapting feature extraction models within a self-paced BCI paradigm. First, we conducted calibration and feedback training on able-bodied naïve subjects using a new self-paced MI BCI including the idle state. The online results suggested that the feature extraction models built from calibration data may not generalize well to feedback sessions. Hence, we have proposed a new supervised adaptation method that learns from feedback data to construct a more accurate model for feedback training. Specifically, we formulated the learning objective as the maximization of kernel-based mutual information estimation with respect to spatial-spectral filters, and derived a gradient-based optimization algorithm for the learning task. We have conducted an experimental study through offline simulations and the results suggest that the proposed method can significantly increase prediction accuracies for feedback training sessions.

References

[1] Wolpaw J R, Birbaumer N, MacFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interface for communication and control *Clin. Neurophysiol.* **113** 767–91

[2] Pfurtscheller G, Neuper C, Flotzinger D and Pregenzer M 1997 EEG-based discrimination between imagination of right and left hand movement *Electroencephalogr. Clin. Neurophysiol.* **103** 642–51

[3] Nijholt A and Tan D 2008 Brain–computer interfacing for intelligent systems *IEEE Intell. Syst.* **23** 72–9

[4] del R Millan J *et al* 2007 Adaptation in brain–computer interfaces *Towards Brain–Computer Interfacing* ed G Dornhege, J del R Millan, T Hinterberger, D McFarland and K R Mueller (Cambridge, MA: MIT Press)

[5] Shenoy P, Krauledat M, Blankertz B, Rao R P N and Muller K-R 2006 Towards adaptive classification for BCI *J. Neural Eng.* **3** 13–23

[6] Buttfield A, Ferrez P W and Millan J d R 2006 Towards a robust BCI: error recognition and online learning *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 164–8

[7] Vidaurre C, Schlogl A, Cabeza R, Scherer R and Pfurtscheller G 2006 A fully online adaptive BCI *IEEE Trans. Biomed. Eng.* **53** 1214–9

[8] Vidaurre C, Kawanabe M, Bunau P von, Blankertz B and Muller K R 2011 Toward unsupervised adaptation of LDA for brain–computer interfaces *IEEE Trans. Biomed. Eng.* **58** 587–97

[9] Lenhardt A, Kaper M and Ritter H J 2008 An adaptive P300-based online brain–computer interface *IEEE Trans. Neural Syst. Rehabil. Eng.* **16** 1–11

[10] Vidaurre C, Schlogl A A, Cabeza R, Scherer R and Pfurtscheller G 2007 Study of on-line adaptive discriminant analysis for EEG-based brain–computer interfaces *IEEE Trans. Biomed. Eng.* **54** 550–6

[11] Yuanqing Li and Guan C 2006 An extended EM algorithm for joint feature extraction and classification in brain–computer interfaces *Neural Comput.* **18** 2730–61

[12] Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Nikulin V and Müller K-R 2008 Invariant common spatial patterns: alleviating nonstationarities in brain–computer

interfacing *Advances in Neural Information Processing Systems* (Cambridge, MA: MIT Press) pp 113–120

[13] Muller-Gerking J, Pfurtscheller G and Flyvbjerg H 1999 Designing optimal spatial filtering of single trial EEG classification in a movement task *Clin. Neurophys.* **110** 787–98

[14] Blankertz B, Dornhege G, Schafer C, Krepki R, Kohlmorgen J, Müller K-R, Kunzmann V, Losch F and Curio G 2003 Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 127–31

[15] Schalk G, Wolpaw J R, McFarland D J and Pfurtscheller G 2000 EEG-based communication: presence of an error potential *Clin. Neurophys.* **111** 2138–44

[16] Ramoser H, Muller-Gerking J and Pfurtscheller G 2000 Optimal spatial filtering of single trial EEG during imagined hand movement *IEEE Trans. Rehabil. Eng.* **8** 441–6

[17] Sugiyama M, Krauledat M and Mueller K R 2007 Covariance shift adaptation by importance weighted cross validation *J. Mach. Learn. Res.* **8** 985–1005

[18] Li Y, Kambara H, Koike Y and Sugiyama M 2010 Application of covariate shift adaptation techniques in brain–computer interfaces *IEEE Trans. Biomed. Eng.* **57** 1318–24

[19] Zhang H and Guan C 2010 A maximum mutual information approach for constructing a 1D continuous control signal at a self-paced brain–computer interface *J. Neural Eng.* **7** 056009

[20] Mason S G and Birch G E 2000 A brain-controlled switch for asynchronous control applications *IEEE Trans. Rehabil. Eng.* **47** 1297–307

[21] Kübler A, Nijboer F, Mellinger J, Vaughan T M, Pawelzik H, Schalk G, McFarland D J, Birbaumer N and Wolpaw J R 2005 Patients with ALS can use sensorimotor rhythms to operate a brain–computer interface *Neurology* **64** 1775–7

[22] Zhang H, Guan C and Wang C 2008 Asynchronous p300-based brain–computer interfaces: a computational approach with statistical models *IEEE Trans. Biomed. Eng.* **55** 1754–63

[23] Blankertz B, Dornhege G, Krauledat M, Müller K-R and Curio G 2007 The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects *NeuroImage* **37** 539–50

[24] Galan F, Nuttin M, Lew E, Ferrez P W, Vanacker G, Philips J and Millán J R 2008 A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots *Clin. Neurophys.* **119** 2159–69

[25] Ang K K, Chin Z Y, Zhang H and Guan C 2008 Filter bank common spatial pattern (FBCSP) in brain–computer interface *IJCNN 2008: Int. Joint Conf. on Neural Networks* pp 2391–8

[26] Zhang H, Guan C and Wang C 2009 Spatio-spectral feature selection based on robust mutual information estimate for brain–computer interfaces *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 2391–8

[27] Zhang H, Chin Z Y, Ang K K, Guan C and Wang C 2011 Optimum spatio-spectral filtering network for brain–computer interface *IEEE Trans. Neural Netw.* **22** 52–63

[28] BCI Competition IV http://www.bbci.de/competition/

[29] Fatourechi M, Fatourechi A, Ward R K and Birch G E 2007 EMG and EOG artifacts in brain–computer interface systems: a survey *Clin. Neurophys.* **118** 480–94

[30] Dornhege G, Blankertz B, Curio G and Müller K-R 2004 Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms *IEEE Trans. Biomed. Eng.* **51** 993–1002

[31] Grosse-Wentrup M and Buss M 2008 Multiclass common spatial patterns and information theoretic feature extraction *IEEE Trans. Biomed. Eng.* **55** 1991–2000

[32] Blankertz B, Tomioka R, Lemm S, Kawanabe M and Müller K-R 2008 Optimizing spatial filters for robust EEG single-trial analysis *IEEE Signal Process. Mag.* **25** 41–56

[33] Dornhege G, Blankertz B, Krauledat M, Losch F, Curio G and Müller K-R 2006 Combined optimization of spatial and temporal filters for improving brain–computer interfacing *IEEE Trans. Biomed. Eng.* **53** 2274–81

[34] Lemm S, Blankertz B, Curio G and Müller K-R 2005 Spatio-spectral filters for improving the classification of single trial EEG *IEEE Trans. Biomed. Eng.* **52** 1541–8

[35] Cover T M and Thomas J A 2006 *Elements of Information Theory* 2nd edn (New York: Wiley)

[36] Petridis S and Perantonis S J 2004 On the relation between discriminant analysis and mutual information for supervised linear feature extraction *Pattern Recognit.* **37** 857–74

[37] Ben-Bassat M 1982 User of distance measures, information measures and error bounds in feature evaluation *Handbook of Statistics* ed P Krishnaiah and L Kanal (Amsterdam: North-Holland) pp 773–91

[38] Last M, Kander A and Maimon O 2001 Information-theoretic algorithm for feature selection *Pattern Recognit. Lett.* **22** 799–811

[39] Sotoca J M and Pla F 2010 Supervised feature selection by clustering using conditional mutual information-based distances *Pattern Recognit.* **43** 2068–81

[40] Estevez P A, Tesmer M, Perez C A and Zurada J M 2009 Normalized mutual information feature selection *IEEE Trans. Neural Netw.* **20** 189–201

[41] Kwak N and Choi C-H 2002 Input feature selection by mutual information based on Parzen window *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1667–71

[42] Bowman A W and Azzalini A 1997 *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* (New York: Oxford University Press)

[43] Chang C-C and Lin C-J 2001 LIBSVM: a library for support vector machines http://www.csie.ntu.edu.tw/∼cjlin/libsvm