Neurocomputing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface

Jianjun Meng, Gan Huang, Dingguo Zhang, Xiangyang Zhu *

The State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 3 December 2011
Received in revised form
30 August 2012
Accepted 3 November 2012
Communicated by S. Hu

Keywords:
Brain–computer interface (BCI)
Common spatial pattern (CSP)
Spatial spectral pattern
Time segment
Channel configuration
Feature selection

ABSTRACT

The power of common spatial pattern (CSP) has been widely validated in electroencephalogram (EEG) based brain–computer interface (BCI). However, its effectiveness is highly dependent on subject-specific time segment, channel configuration and frequency band. Hence, the preprocessing procedure of CSP algorithm is critical to enhance the performance of BCI system. This paper proposes a feature extraction and selection method based on common spatial and spectral pattern for motor imagery brain–computer interface (BCI). We formulate the optimization of spatial spectral patterns, channel configuration and time segment as maximizing the proposed criterions including mutual information algorithm, Fisher ratio algorithm and wrapper method. The proposed method is evaluated on single trial EEG from dataset IVa of BCI competition III. The results show that best features are selected by a wrapper method and these features in cross-validation yield better performance compared to most of the reported results.

1. Introduction

Brain–computer interface (BCI) technique aims to translate humans' thoughts into commands. In recent years there is continuous progress in both invasive and noninvasive BCI technology [1]. The study of electroencephalogram (EEG) based BCI is attracting more and more researchers because of its relatively simple and inexpensive equipment [2]. Moreover, the open accessed datasets in BCI competition provide a comparable platform to evaluate algorithms [3]. In this paper, we focus on noninvasive, EEG based BCI systems.

It is reported that both actual movement and imaginary movement of different body parts can cause mu and beta rhythms: event-related desynchronization/synchronization (ERD/ERS) [4]. ERD/ERS has been used widely in motor imagery BCI. One of the biggest challenges to such a system is the low signal-to-noise ratio (SNR) [5]. Since raw EEG recordings have a poor spatial resolution due to volume conduction, it has motivated applications of spatial filters, which extract spatial information of humans' intention as much as possible. The common spatial pattern algorithm is one of the frequently used methods for this purpose. It finds the directions simultaneously diagonalizing two covariance matrices, which are associated with two classes of motor imagery conditions [6,7].

The CSP algorithm focuses only on spatial information. However, the performance of CSP is sensitive to subject-specific parameters such as the time segment, channel configuration and temporal frequency band-pass filtering of the EEG signals. To solve the problem of manually tuning the subject-specific frequency band for the CSP algorithm, several extensions of CSP have been proposed. The solutions can be divided into three categories. The first one can be viewed as embedding one or several time-delay signals into the CSP procedure, such as common spatial spectral pattern (CSSP) [8] and its improved version common sparse spectral spatial pattern (CSSSP) [9]. The second one optimizes the temporal filters equivalently in frequency domain. The spectrally weighted common spatial pattern (SWCSP) proposed by Tomioka et al. [10] and iterative spatial spectral pattern learning (ISSPL [11]) fall into this category. The third one utilizes several narrow band filters and select a reduced set of features from all the narrow bands. The sub-band common spatial pattern and filter bank common spatial pattern [12,13] belong to this category. All the solutions have its own advantages and disadvantages in optimizing the frequency information. Moreover, all the above methods do not consider the channel configuration, which is another important factor to influence the CSP performance.

Since multi-channel EEG signals are highly correlated, and signals from different channels do not contain the same amount

* Corresponding author.
E-mail address: mexyzhu@sjtu.edu.cn (X. Zhu).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.004

of discriminative information. Applying a large number of EEG channels may include more noisy and redundant signals in some channels [14,15], which in turn degrades the BCI performance. Another additional problem associated with CSP is its tendency to overfit with a large number of channels because of inaccurate estimation of the covariance matrix [16]. While too few channels may not include enough information for high performance. How to choose an appropriate number of most discriminant channels is still an open question. To the best of our knowledge, only a few researchers try to solve the channel selection problem for common spatial pattern algorithm. It includes regularized CSP for sparse solution [17,18], which shrinks the coefficient of some unimportant channels to be zeros. However, these algorithms project the signals in the most discriminative direction at the expense of decreasing the accuracy. Recently, another extension of sparse algorithm is proposed [19], which reports better classification accuracy than regularized CSP [17,18]. It solves a more complicated quadratically constrained quadratic programming problem, while the computational time may be a bottleneck for real application. Based on the fact that it is a lack of efficient algorithm for CSP to select proper channels. We learn from the study method provided by optimal subject-independent channel configuration for sensorimotor rhythms (SMR) based brain–computer interfaces [20]. Twelve different fixed channel configurations are compared to the full dense placement of electrodes. The aim is to find the most proper channel configuration for common spatial spectral patterns, hence, classification accuracy might be further improved.

To address the problem of optimizing time segment, channel configuration and temporal frequency band-pass filtering for EEG signals jointly, the spectrally weighted common spatial pattern (SWCSP) is calculated on different time segments and channel configurations. These SWCSP features are highly correlated and feature selection techniques are used to select the best time segment and channel configuration for each individual subject. Feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods [21]. Filter methods based on mutual information and Fisher ratio, wrapper method based on SVM classifier are used in this study.

The main contribution of this paper includes firstly the optimization of spatial spectral pattern, channel and time configuration is arranged in an optimization framework for each subject with few parameters to tune. Secondly, we show that the channel configuration is equivalently important to the BCI performance as temporal or frequency information optimization. The features of spatial spectral patterns are extracted by learning spatial filters and spectral filters separately but with close relationship. The best features are selected from multiple channel configurations and time segments according to maximize mutual information, Fisher ratio or classification accuracy by the wrapper method.

The rest of paper is organized as follows. We describe the method in Section 2 and show the description of datasets and experimental setup in Section 3. The results and discussion are given in Sections 4 and 5, respectively. Section 6 concludes the paper.

2. Methodology

2.1. Problem definition

Denote the original short segment of EEG signals as $X_{(i)} \in \Re^{N\times T}$, $i=1,2,\ldots,L$, in the training set $\mathcal{D}$ and the corresponding class labels as $y_{(i)} \in \{+1,-1\}$ in the class label set Ω e.g. right or left hand imaginary movement. The $X_{(i)}$ corresponds to one trial of imaginary movement in a specific time window, N is the number of channels, T is the number of time points in the time window, and L is the total number of trials. Throughout this paper, we assume that $X_{(i)}$ has been centered, which means $X_{(i)}=X_{(i)}(I_T-1/T\times \mathbf{1_T}\cdot\mathbf{1_T^\dagger})$, where † means the conjugate transpose of a matrix and I_T is the $T\times T$ identity matrix, $\mathbf{1_T}=[1,1,\ldots,1]^\dagger$ is a $T\times 1$ vector. The time segment sets used in this paper are denoted as $\mathcal{T}$, channel configuration sets as $\mathcal{C}$. Let $(n_t,n_c)\in(\mathcal{T}\times\mathcal{C})$, $n_t=1,\ldots,|\mathcal{T}|$, $n_c=1,\ldots,|\mathcal{C}|$ to be a specific setting of time segment and channel configuration, where $(\mathcal{T}\times\mathcal{C})$ is the cartesian product, $|\cdot|$ is the cardinality of set.

The central task in this binary classification is to assign single trial EEG data $X_{(i)}$ to one of the predefined class labels $y_{(i)}\in\{+1,-1\}$. To achieve this, a discriminant feature extractor is essential to reduce the observation data space $\mathcal{D}$ to suitable lower dimensional subspace Φ. That means for any $X_{(i)}\in\mathcal{D}$, the goal is to find a transform $f^*:\mathcal{D}\rightarrow\Phi$ while preserving discriminability as much as possible. For simplicity of notation, the subscript i which denotes the trial number is omitted. We make it explicit whenever confusion may arise.

In this study, we use spatial spectral feature extractor, w.r.t. X

$$\phi_j(X;\mathbf{w_j},B_j)=\mathbf{w_j^\dagger}XB_jB_j^\dagger X^\dagger\mathbf{w_j}, \tag{1}$$

$\{\mathbf{w_j},B_j\}$ is a pair of spatial and temporal filters, respectively. We use the total number of J pairs of spatial and temporal filters, denoted as $\{\mathbf{w_j},B_j\}_{j=1,\ldots,J}$. The temporal filter B_j is optimized equivalently in the frequency domain in this paper and its equivalent spectral filter is denoted as $\boldsymbol{\beta}_j$. Hence the spatial temporal filters $\{\mathbf{w_j},B_j\}$ are also called spatial spectral filters. We do not make explicit distinction between the two terms. This feature extractor consists of multiple pairs of spatial spectral filters. In each pair of filters, the temporal filter aims to capture the most discriminative frequency information, while the associated spatial filter projects the multi-channel data into surrogate channel.

Assuming total number of J spatial spectral features are used, we denote the feature vector by

$$\boldsymbol{\phi}=[\phi_1(X;\mathbf{w_1},B_1),\ldots,\phi_J(X;\mathbf{w_J},B_J)]^\dagger. \tag{2}$$

Then for each pair of time segment, channel configuration (n_t,n_c), a feature vector $\boldsymbol{\phi}^{(n_t,n_c)}$ is generated from this setting. Mutual information and Fisher ratio between feature sets and class label are computed, respectively. For wrapper method, classification accuracy is computed by SVM classifier using extracted features. Then feature sets with maximum Mutual information, Fisher ratio or classification accuracy are chosen to be the best subject-dependent channel configuration and time segment.

The method comprises four stages: multi-time segment and channel configuration of original EEG data, spatial spectral learning, best setting selection (feature selection) and classification. The best setting for time segment and channel configuration, the most discriminative spatial spectral filters are computed from the training data. These parameters computed from training phase are then used to evaluate the test data.

2.2. Multi-time segment and channel configurations

The first stage employs various channel configurations to perform the spatial spectral feature extraction on the multi-time segments of EEG. The time segments are 0.5–2.5, 1.0–3.0, 1.5–3.5, and 0.5–3.5 s from the onset of the visual cue given to the subject. The first 0.5 s period after the cue-onset is excluded as it may contain the spontaneous responses to the visual stimulus. The channel configurations are according to the 12 channel configurations defined in

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

[20] combined with the 118 full channels (see Fig. 2). Note that the last channel configuration 12 is modified as (9ch) which is different from the last one (8ch) in [20].

2.3. Spatial spectral pattern learning

2.3.1. Learning spatial filters

Common spatial pattern algorithm [6] is used to generate spatial filters. Using a specific temporal filter B_j, $(j=1,\ldots,J)$,[1] the filtered EEG signal is written as XB_j. Denote the mean spatial covariance matrices for the two classes by $\Sigma_j^{(+)}$ and $\Sigma_j^{(-)}$, respectively, with

$$\Sigma_j^{(+/-)} = \langle XB_jB_j^{\dagger}X^{\dagger}\rangle_{(+/-)}, \tag{3}$$

where angle brackets denote expectation within a specific class. The CSP is formulated to maximize the power of projected signal for one class and minimize that for the other class simultaneously:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\frac{\mathbf{w}^{\dagger}\Sigma_j^{+}\mathbf{w}}{\mathbf{w}^{\dagger}\Sigma_j^{-}\mathbf{w}} \quad \text{and} \quad \underset{\mathbf{w}}{\operatorname{argmax}}\frac{\mathbf{w}\dagger\,\Sigma_j^{-}\mathbf{w}}{\mathbf{w}^{\dagger}\Sigma_j^{+}\mathbf{w}}. \tag{4}$$

The above optimization problem is easily solved by the following generalized eigenvalue problem:

$$\Sigma_j^{+}\mathbf{w} = \lambda\Sigma_j^{-}\mathbf{w} \tag{5}$$

The eigenvectors correspond to the largest and smallest eigenvalue of Eq. (5) are solutions of problems (4), respectively. The maximum values of problems (4) are denoted as $\lambda_j^{(+)}$ and $\lambda_j^{(-)}$, respectively. The discriminability of each eigenvector is measured by the corresponding eigenvalue. Previous studies show that the second or the third eigenvectors are helpful to improve the classification. Therefore, we use the m pairs of eigenvectors corresponding to the largest and smallest m eigenvalues as the set of spatial filters, which means $J=2$ m in Eq. (2). Denote $W_j^{(+)}$ as the set of m eigenvectors that satisfy the first problem of Eq. (4), $W_j^{(-)}$ as the set of m eigenvectors that satisfy the second problem of Eq. (4). Set $W^{(+)} := W_{j^*}^{(+)}$ with $j^* = \arg_{j=1,\ldots,J}\max\lambda_j^{(+)}$ and $W^{(-)} := W_{j^*}^{(-)}$ with $j^* = \arg_{j=1,\ldots,J}\max\lambda_j^{(-)}$. and $W=\{W^{(+)}, W^{(-)}\}$.

2.3.2. Learning spectral filters

With the derived spatial filters $\{\mathbf{w_j}\}_{j=1,\ldots,J}\in W$, we continue to learn its corresponding temporal filter B_j. The temporal filter is convenient to be derived in frequency domain as $\boldsymbol{\beta}_j$. A linear time-invariant temporal filter B_j $(j=1,\ldots,J)$, which is a circulant matrix, can be diagonalized by the discrete Fourier transform [22]

$$F^{\dagger}B_j = diag([\beta_j^{(1)},\ldots,\beta_j^{(T)}])F^{\dagger}, \tag{6}$$

where $F\in R^{T\times T}$ is the Fourier matrix, $\boldsymbol{\beta}_j=[\beta_j^{(1)},\ldots,\beta_j^{(T)}]^{\dagger}$ is the equivalent spectral filter of B_j. It then follows

$$\begin{aligned} F^{\dagger}B_jB_j^{\dagger}F &= diag([\beta_j^{(1)},\ldots,\beta_j^{(T)}])F^{\dagger}F\ diag([\beta_j^{(1)},\ldots,\beta_j^{(T)}]) \\ &= diag([\alpha_j^{(1)},\ldots,\alpha_j^{(T)}]), \end{aligned} \tag{7}$$

where $\alpha_j^{(i)}=(\beta_j^{(i)})^2$, $i=1,\ldots,T$. Here we use the fact $F^{\dagger}F=I_T$ for the second equality. Then the feature $\phi_j(X;\mathbf{w_j},B_j)$ in Eq. (1) can be rewritten as

$$\begin{aligned} \phi_j(X;\mathbf{w_j},B_j) &= \mathbf{w_j}^{\dagger}XB_jB_j^{\dagger}X^{\dagger}\mathbf{w_j} = \mathbf{w_j}^{\dagger}XFF^{\dagger}B_jB_j^{\dagger}FF^{\dagger}X^{\dagger}\mathbf{w_j} \\ &= \mathbf{w_j}^{\dagger}\widehat{X}\ diag([\alpha_j^{(1)},\ldots,\alpha_j^{(T)}])\widehat{X}^{\dagger}\mathbf{w_j} \\ &= \sum_{i=1}^{T}\alpha_j^{(i)}\mathbf{w_j}^{\dagger}\widehat{\mathbf{x}}^{(\mathbf{i})}\widehat{\mathbf{x}}^{(\mathbf{i})\dagger}\mathbf{w_j} = 2\sum_{i=1}^{T'}\alpha_j^{(i)}\mathbf{w_j}^{\dagger}Re[\widehat{\mathbf{x}}^{(\mathbf{i})}\widehat{\mathbf{x}}^{(\mathbf{i})\dagger}]\mathbf{w_j} \\ &= \sum_{i=1}^{T'}\alpha_j^{(i)}\widehat{z}_j^{(i)} = \boldsymbol{\alpha}_j^{\dagger}\widehat{\mathbf{z}}_\mathbf{j} \quad (j=1,\ldots,J), \end{aligned} \tag{8}$$

where $\widehat{X}=XF$ denote the discrete Fourier transformed data matrix, and $\widehat{\mathbf{x}}^{(\mathbf{i})}$ is its ith column. The third equality holds from Eq. (7) and the fifth equality due to the spectral coefficient of a temporal filter are symmetric around $T'=\lceil(T+1)/2\rceil$ ($\lceil\rceil$ means rounding to the nearest integer towards zero). We denote $2\mathbf{w_j}^{\dagger}Re[\widehat{\mathbf{x}}^{(\mathbf{i})}\widehat{\mathbf{x}}^{(\mathbf{i})\dagger}]\mathbf{w_j}$ by $\widehat{z}_j^{(i)}$, and define $\widehat{\mathbf{z}}_\mathbf{j}=[\widehat{z}_j^{(1)},\ldots,\widehat{z}_j^{(T')}]$. The variable $\widehat{z}_j^{(i)}$ is the power at the ith frequency bin of spatially filtered data, hence, $\widehat{\mathbf{z}}_\mathbf{j}$ is the power spectrum. The CSP takes a homogeneous weighting of the spectrum, i.e. $\boldsymbol{\alpha}_\boldsymbol{j}=\mathbf{1}\in R^{T'\times 1}$. Different frequency bands might make different role in discriminating ERD/ERS of two classes. In order to achieve good discrimination, we formulate the problem of optimizing $\boldsymbol{\alpha}_\boldsymbol{j}$ as follows [10]:

$$\begin{aligned} &\max_{\alpha}\frac{\boldsymbol{\alpha}^{\dagger}(\langle\widehat{\mathbf{z}}\rangle_{+}-\langle\widehat{\mathbf{z}}\rangle_{-})}{\sqrt{Var[\boldsymbol{\alpha}^{\dagger}\widehat{\mathbf{z}}]_{+}+Var[\boldsymbol{\alpha}^{\dagger}\widehat{\mathbf{z}}]_{-}}} \\ &\text{s.t. } \alpha^{(i)}\geq 0 \quad (\forall i=1,2,\ldots,T'). \end{aligned} \tag{9}$$

Here, we omit the subscript j of $\boldsymbol{\alpha}$ and $\widehat{\mathbf{z}}$ for simplicity of notation. The angle brackets denote expectation and $Var[\cdot]$ denote the variance for the specific class. Eq. (9) takes nonhomogeneous weighting of power spectrum in order to discriminate signals from two classes as much as possible. This optimization can be viewed as Fisher discriminant analysis with an additional constraint that all coefficients must be positive. The solution of FDA might not be right for the constrained problem (9). We use the assumption that the signal is a stationary Gaussian process like in [10]. Then the frequency components of $\widehat{\mathbf{z}}_j$ can be seen independent to each other for a given class label.

Without loss of generality, we assume that the first m spatial temporal filters correspond to the "+1" class, and the latter m ones correspond to the "−1" class. It is reasonable to assume the following relationship [11]:

$$\left\langle\sum_{i=1}^{T'}\alpha_j^{(i)}\widehat{z}_j^{(i)}\right\rangle_{+} > \left\langle\sum_{i=1}^{T'}\alpha_j^{(i)}\widehat{z}_j^{(i)}\right\rangle_{-} \quad (j=1,\ldots,m), \tag{10}$$

for the first m spatial temporal filters according to Eq. (4). Then the solution of Eq. (9) is given by

$$\alpha_j^{(i)} = \begin{cases} \Sigma_{z^i}^{-1}(\langle\widehat{z}_j^{(i)}\rangle_{+}-\langle\widehat{z}_j^{(i)}\rangle_{-}) & \text{if } (\langle\widehat{z}_j^{(i)}\rangle_{+}-\langle\widehat{z}_j^{(i)}\rangle_{-})\geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad i=1,\ldots,T' \tag{11}$$

where $\Sigma_{z^i}=Var[\widehat{z}_j^{(i)}]_{+}+Var[\widehat{z}_j^{(i)}]_{-}$. Having solved $\alpha_j^{(i)}$, it is direct to get $\beta_j^{(i)}$. Note that the labels "+" and "−" are exchanged for the latter m spatial temporal filters which correspond to "−1" class. Because the convergence of SWCSP, ISSPL is not proved, and several iterations of spatial and spectral learning also take more computational time. We simply iterate the spatial and spectral learning twice.

2.4. Feature selection

The spatial spectral features of the ith trial for the EEG from the n_tth time segment and n_cth channel configuration are given by

$$\boldsymbol{\phi}_i^{(n_t,n_c)} = [\log(\phi_1),\ldots,\log(\phi_J)]^{\dagger}. \tag{12}$$

The logarithm operation is favored for its purpose as "to approximate normal distribution of the data" [6]. Another positive effect

[1] For the initial temporal filter B_{init}, we set its equivalent spectral filter to be $\boldsymbol{\beta}_{init}=\mathbf{1}$. After the first iteration, there are J spectral filters, learned by 'learning spectral filters' automatically. Refer to Table 1 for more details.

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

of the logarithm operation is the reduced dynamic range, which facilitates the subsequent processing [23]. For a specific setting $(n_t,n_c)\in(\mathcal{T}\times\mathcal{C})$, the feature sets from all the training data $X_{(i)}\in\mathcal{D}$ form a feature vector $F_{n_t,n_c}=[\boldsymbol{\phi}_1^{(n_t,n_c)\dagger},\ldots,\boldsymbol{\phi}_L^{(n_t,n_c)\dagger}]^\dagger$. The corresponding class labels form a vector $Y=[y_1,\ldots,y_L]^\dagger$. In this stage, the spatial spectral features in the most discriminative setting (n_t^*,n_c^*) are chosen for a specific subject. Various feature selection algorithms can be used. Based on the reported study of BCI competition III in [24], the mutual information based best individual feature (MIBIF) [25] yielded better results than others. While because features in different channel configurations and time segments might be correlated with each other, the features cannot be seen as mutually independent. Hence only the estimation of mutual information is used as a criterion to evaluate the efficiency of feature selection. In the study, we also find feature selection based on Fisher ratio yielded comparable results, sometimes even better results than MIBIF. Therefore, it is also used in this paper.

Mutual information algorithm

Step 1: For each n_tth time segment, n_cth channel configuration, compute the set of CSP features or spatial spectral (SWCSP) features F_{n_t,n_c} from the training dataset $\mathcal{D}$.

Step 2: Calculate the mutual information of each feature set F_{n_t,n_c} with the class label set $\Omega=\{+1,-1\}$

$$I(F_{n_t,n_c},\Omega)=H(\Omega)-H(\Omega|F_{n_t,n_c}), \tag{13}$$

where $H(\Omega)=-\sum_{\omega\in\Omega}P(\omega)\log(P(\omega))$; and the conditional entropy of Ω given obtained feature vector F_{n_t,n_c} is

$$\begin{aligned}H(\Omega|F_{n_t,n_c})&=-\sum_{\omega\in\Omega}P(\omega|F_{n_t,n_c})\log(P(\omega|F_{n_t,n_c}))\\&=-\frac{1}{L}\sum_{\omega\in\Omega}\sum_{i=1}^{L}P(\omega|\boldsymbol{\phi}_i^{(n_t,n_c)})\log(P(\omega|\boldsymbol{\phi}_i^{(n_t,n_c)})),\end{aligned} \tag{14}$$

where $\boldsymbol{\phi}_i^{(n_t,n_c)}$ is the feature vector of ith trial in the setting of $(n_t,n_c)\in(\mathcal{T}\times\mathcal{C})$.

The probability function $P(\omega|\boldsymbol{\phi}_i^{(n_t,n_c)})$ can be estimated from the samples using the Bayes rule given in the following two equations:

$$P(\omega|\boldsymbol{\phi}_i^{(n_t,n_c)})=\frac{p(\boldsymbol{\phi}_i^{(n_t,n_c)}|\omega)P(\omega)}{p(\boldsymbol{\phi}_i^{(n_t,n_c)})}, \tag{15}$$

where $p(\boldsymbol{\phi}_i^{(n_t,n_c)}|\omega)$ is the conditional probability density function of $\boldsymbol{\phi}_i^{(n_t,n_c)}$ given class ω, which is also called likelihood. $P(\omega)$ is the prior probability of class ω; and the evidence factor

$$p(\boldsymbol{\phi}_i^{(n_t,n_c)})=\sum_{\omega\in\Omega}p(\boldsymbol{\phi}_i^{(n_t,n_c)}|\omega)P(\omega). \tag{16}$$

The conditional probability density function $p(\boldsymbol{\phi}_i^{(n_t,n_c)}|\omega)$ can be estimated from samples using kernel density estimation [26]

$$\widehat{p}(\boldsymbol{\phi}_i^{(n_t,n_c)}|\omega)=\frac{1}{n_\omega}\sum_{l\in I_\omega}\varphi(\boldsymbol{\phi}_i^{(n_t,n_c)}-\boldsymbol{\phi}_l^{(n_t,n_c)}), \tag{17}$$

where I_ω is the set of indices of the training data belonging to class ω, $n_\omega=|I_\omega|$. In this paper, we use Gaussian kernel function φ for density estimation. A multivariate Gaussian function is given by

$$\varphi(\mathbf{r})=(2\pi)^{-n_l/2}|\psi|^{-1/2}e^{(-(1/2)\mathbf{r}^\dagger\psi^{-1}\mathbf{r})}, \tag{18}$$

where n_l is the dimensionality of vector $\mathbf{r}$, $\mathbf{r}$ denote the term $\boldsymbol{\phi}_i^{(n_t,n_c)}-\boldsymbol{\phi}_l^{(n_t,n_c)}$, ψ usually takes a diagonal matrix form. The diagonal elements of ψ are given by

$$\begin{aligned}\psi_{m,m}&=\zeta^2\sigma^2\\&=\frac{(4/3n_\omega)^{2/5}}{n_\omega-1}\sum_{i=1}^{n_\omega}(\phi_{im}^{(n_t,n_c)}-\overline{\phi}_m^{(n_t,n_c)})^2\quad(m=1,\ldots,n_l),\end{aligned} \tag{19}$$

where $\overline{\phi}_m^{(n_t,n_c)}$ is the empirical mean of $\{\phi_{im}^{(n_t,n_c)}\}$. The normal optimal smoothing strategy as in [24,27] is used to set coefficient, i.e. $\zeta=(4/3n_\omega)^{(1/5)}$.

Step 3: Select the setting (n_t^*,n_c^*) which satisfies

$$(n_t^*,n_c^*)=\arg\max_{(n_t,n_c)}I(F_{n_t,n_c},\Omega) \tag{20}$$

to be the best time segment and channel configuration for the subject. The $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}$ is the corresponding spatial spectral filters.

Fisher ratio algorithm

Step 1: This step is the same as step 1 of the mutual information algorithm.

Step 2: Calculate the fisher ratio [26] of each feature set F_{n_t,n_c} with the class label set $\Omega=\{+1,-1\}$

$$R_F(F_{n_t,n_c},\Omega)=\frac{\|\langle\boldsymbol{\phi}^{(n_t,n_c)}\rangle_+-\langle\boldsymbol{\phi}^{(n_t,n_c)}\rangle_-\|^2}{\mathrm{tr}(Var[\boldsymbol{\phi}^{(n_t,n_c)}]_++Var[\boldsymbol{\phi}^{(n_t,n_c)}]_-)}. \tag{21}$$

Step 3: Select the setting (n_t^*,n_c^*) which satisfies

$$(n_t^*,n_c^*)=\arg\max_{(n_t,n_c)}R_F(F_{n_t,n_c},\Omega) \tag{22}$$

to be the best time segment and channel configuration for the subject. The $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}$ is the corresponding spatial spectral filters.

Wrapper method

In the wrapper approach, the feature subset selection algorithm exists as a wrapper around the induction (classification) algorithm. The feature subset selection algorithm conducts a search for a good subset using the induction (classification) algorithm itself as part of the function evaluating feature subsets [28]. The SVM classifier is used for wrapper method in this paper. The best features are selected by maximizing the classification accuracy. Cross-validation on the training dataset is used to avoid overfitting. The steps of wrapper method are shown in the following:

Step 1: For each n_tth time segment, n_cth channel configuration, compute the set of CSP features or spatial spectral (SWCSP) features F_{n_t,n_c} from the training dataset $\mathcal{D}$.

Step 2: Split the training dataset into m-fold subsets for cross-validation.

Step 3: Repeat the m-fold cross-validation n times and calculate the average accuracy of cross-validations. Denote the average accuracy as

$$R_F(n_t,n_c)=Ave(ACC) \tag{23}$$

Step 4: Select the setting (n_t^*,n_c^*) which satisfies

$$(n_t^*,n_c^*)=\arg\max_{(n_t,n_c)}R_F(F_{n_t,n_c},\Omega) \tag{24}$$

to be the best time segment and channel configuration for the subject.

2.5. Classification

The best time segment and channel configuration for a subject is chosen according to the training dataset $\mathcal{D}$. Let (n_t^*,n_c^*) be the best setting, $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}_{j=1,\ldots,J}$ is the corresponding spatial spectral filters. Then for the test data set $\mathcal{E}$, the test feature vector is computed according to the spatial spectral filters $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}_{j=1,\ldots,J}$ in the best setting. Table 1 summarizes the steps of feature extraction and selection algorithm for optimal spatial spectral filters from various settings. Denote the test feature vectors as $\overline{F}_{n_t,n_c}$, we use the SVM as the classifier. Note that the SVM is used as the classifier to verify the classification accuracy of test feature vectors $\overline{F}_{n_t,n_c}$ for both filter ones (including mutual information and Fisher ratio feature selection methods) and wrapper method. Furthermore, the SVM is also used as the classifier for nested cross-validation (in wrapper method).

Support vector machine (SVM) has broad applications in classification. A lot of good results using SVM have been reported

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

Table 1
Learning algorithm for optimal spatial spectral filters jointly with best setting.

Input: The training dataset $\mathcal{D}=X_{(1)},\ldots,X_{(L)}$, corresponding class labels $y_{(1)},\ldots,y_{(L)}$ and the initial temporal filter B_{init}, its corresponding spectral filter $\boldsymbol{\beta}_{init}=\mathbf{1}$. J, the used total number of spatial spectral filters. Time segment set $\mathcal{T}$ and channel configuration set $\mathcal{C}$.
Output: The best setting includes time segment n_t^* and channel configuration n_c^* and the corresponding spatial and spectral filters $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}, j=1,\ldots,J$.
Step 1: Construct the set of time segments and channel configurations ($\mathcal{T}\times\mathcal{C}$).
Step 2: The EEG signals for training are configured as $X_{(i)}^{(n_t,n_c)}$, $i=1,\ldots,L$. For simplicity, the superscript (n_t,n_c) is omitted.
(1) Compute the discrete Fourier transformed data $\widehat{X}_{(i)}=X_{(i)}F$, $i=1,\ldots,L$.
(2) Repeat spatial and spectral pattern learning twice.
(a) Compute the empirical covariance matrix $\Sigma_j^{(+/-)}$ by Eq. (3) for each set of spectral coefficients. The initial spectral coefficients are given by $\boldsymbol{\beta}_{init}$.
(b) Select $J=2\,m$ eigenvectors that correspond to the largest and smallest m eigenvalues of Eq. (5). Denote $W_j^{(+)}$ as the set of m eigenvectors that satisfy the first problem of Eq. (4), $W_j^{(-)}$ as the set of m eigenvectors that satisfy the second problem of Eq. (4).
(c) Set $W^{(+)}:=W_{j^*}^{(+)}$ with $j^*=\arg_{j=1,\ldots,J}\max\lambda_j^{(+)}$ and $W^{(-)}:=W_{j^*}^{(-)}$ with $j^*=\arg_{j=1,\ldots,J}\max\lambda_j^{(-)}$.
(d) For each spatial filters $\mathbf{w}_j\in\{W^{(+)},W^{(-)}\}$ ($j=1,\ldots,J=2\,m$) calculate the spectral filters $\boldsymbol{\beta}_j$ according to Eq. (11) for the first m ones and exchange the labels '+' and '-' in Eq. (11) for the last m spectral filters.
(3) Compute the spatial spectral features ϕ according to Eqs. (1) and (2) by $\{\mathbf{w}_j,\boldsymbol{\beta}_j\}_{j=1,\ldots,J}$ and denote the feature vector as $F_{n_t,n_c}=[\boldsymbol{\phi}_1^{(n_t,n_c)^\dagger},\ldots,\boldsymbol{\phi}_L^{(n_t,n_c)^\dagger}]^\dagger$.
Step 3: Calculate mutual information or Fisher ratio of each feature set F_{n_t,n_c} with the class label set $\Omega=\{+1,-1\}$ by Eq. (13) or (21), respectively. Calculate the average accuracy of cross-validations for features set F_{n_t,n_c} from the training dataset $\mathcal{D}$ (by Eq. (23)).
Step 4: Select the setting (n_t^*,n_c^*) that satisfies Eq. (20), (22) or (24) to be the best setting. The $\{\mathbf{w}_j^*,\boldsymbol{\beta}_j^*\}$, $j=1,\ldots,J$ is the corresponding spatial spectral filters.

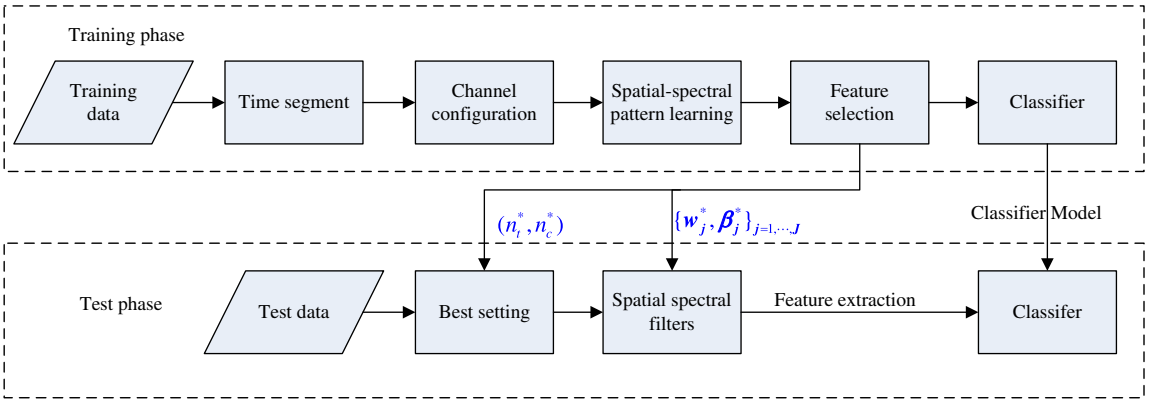


Fig. 1. Methodology of the proposed feature extraction and selection of optimal spatial spectral filters from various settings ($\mathcal{T}\times\mathcal{C}$) for the training and test phase.

in the BCI system [29,11]. In this paper, we use the SVM Matlab packet [30] as the classification tool. The basic idea is to separate feature vector $\mathbf{x}\in\Re^{n\times1}$ from two classes by finding a weight vector $\boldsymbol{\beta}\in\Re^{n\times1}$ and an offset $\beta_0\in\Re$ of a hyperplane

$$H:\mathbf{x}\mapsto\text{sign}(\boldsymbol{\beta}^\dagger\cdot\mathbf{x}+\beta_0), \tag{25}$$

with the largest possible margin [31]. One variant of the algorithm is to solve the following optimization problem:

$$\begin{aligned}&\min_{\boldsymbol{\beta}\in\Re^n}\frac{1}{2}\|\boldsymbol{\beta}\|_2^2+C\sum_{i=1}^{N}\xi_i\\&\text{s.t. } y_i(\boldsymbol{\beta}\cdot\mathbf{x}^{(\mathbf{i})}+\beta_0)\geq1-\xi_i\quad\forall i.\\&\quad\xi_i\geq0\end{aligned} \tag{26}$$

The parameters ξ_i are called slack variables and ensure that the problem has a solution in case the data are not linear separable. C is the penalty parameter of the error term. In this study, the Gaussian radial basis function (RBF) $K(\mathbf{x_i},\mathbf{x_j})=\exp(-\gamma\|\mathbf{x_i}-\mathbf{x_j}\|^2)$, $\gamma>0$ is chosen as the kernel function. The cost value C and γ in kernel function are fixed as experience values 10 and 0.17, respectively, although grid search to optimize these parameters for individual person might further improve the performance. The main purpose of this study is to evaluate the efficiency of optimizing spatial spectral patterns jointly with channel and time segment configurations, hence, the optimization of SVM classifier is out of the scope of this research.

As a summary, the proposed optimal spatial spectral patterns selected by feature selection algorithm is illustrated in Fig. 1.

3. Application

3.1. Data description

The EEG data used here are dataset IVa in BCI competition III. They are provided by Fraunhofer FIRST (Intelligent Data Analysis Group) and Campus Benjamin Franklin of the Charité—University Medicine Berlin (Neurophysics Group) [32]. The EEG data were recorded from five healthy subjects (aa, al, av, aw and ay) and 118 electrodes were placed for each subject with a sampling rate of 1000 Hz. These datasets contain data from four initial sessions without feedback. Subjects sat in a comfortable chair with arms resting on armrests. In each trial, the subject was given visual cues for 3.5 s ($t\in[0,3.5]$ s), during which one of the three motor imageries should be performed: left hand, right hand and right foot. The presentation of target cues was intermitted by periods of random length, 1.75–2.25 s, in which the subject could relax. Only EEG trials for right-hand and right-foot movements were provided for analysis. A total of 280 trials were performed by each subject and the number of trials for each task is equal.

3.2. Data preprocessing

As described before, the time segment set is $\mathcal{T}=\{[0.5,2.5]$ s, $[1.0,3.0]$ s, $[1.5,3.5]$ s, $[0.5,3.5]$ s$\}$. Channel configuration set $\mathcal{C}$ is shown in Fig. 2. The 118 full channel configuration is not showed on the figure for better visualization. The cardinality of set $\mathcal{T}$ and $\mathcal{C}$ is $|\mathcal{T}|=4$, $|\mathcal{C}|=13$, respectively. Firstly, the EEG data were

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

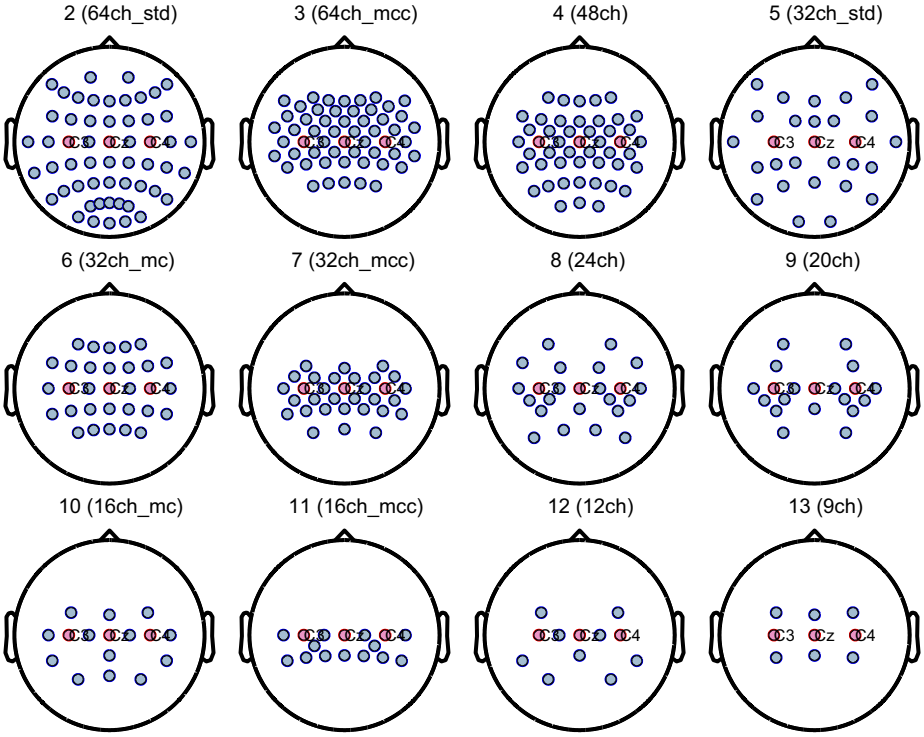


Fig. 2. Predefined channel configurations. Channels 'C3', 'Cz', 'C4' are labelled and filled with red color, labels of other channels are omitted for a better visualization. The first 118 full channel configuration "1(118ch)" is omitted. The other 12 channel configurations are sequenced according to the number of channels in descending order. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
10×10-fold cross-validation results of average classification accuracy $\pm$ standard deviation (%) of SWCSP and CSP features in different time segments, with 118 full channel configuration. The best performance for each subject is highlighted in bold and bold italic for SWCSP and CSP, respectively.

Subject	[0.5,2.5] s		[1.0,3.0] s		[1.5,3.5] s		[0.5,3.5] s	
	SWCSP	CSP	SWCSP	CSP	SWCSP	CSP	SWCSP	CSP
aa	**85.8 ± 6.3**	78.4 ± 7.8	83.2 ± 6.5	***79.1 ± 7.4***	75.8 ± 7.2	71.7 ± 7.4	83.9 ± 6.6	78.8 ± 7.1
al	**97.9 ± 2.6**	97.6 ± 2.9	97.2 ± 3.2	97.5 ± 2.9	96.6 ± 3.2	96.8 ± 3.5	97.7 ± 2.7	***97.6 ± 2.8***
av	**75.6 ± 8.5**	73.7 ± 8.4	72.5 ± 8.3	71.6 ± 8.5	67.9 ± 8.6	69.0 ± 9.4	73.7 ± 9.0	***73.8 ± 8.4***
aw	95.5 ± 4.3	93.1 ± 5.2	97.8 ± 2.6	93.9 ± 5.8	96.5 ± 3.8	91.1 ± 5.9	**98.1 ± 2.5**	***95.7 ± 3.6***
ay	**95.7 ± 4.2**	***94.4 ± 4.5***	94.4 ± 4.3	94.0 ± 4.2	92.2 ± 4.6	92.0 ± 5.0	95.6 ± 4.0	94.2 ± 4.2
Average	**90.1 ± 5.2**	87.4 ± 5.7	89.0 ± 4.9	87.2 ± 5.8	85.8 ± 5.5	84.1 ± 6.3	89.8 ± 5.0	***88.0 ± 5.2***

down-sampled to 100 Hz for use. The initial temporal filter B_{init} ($\boldsymbol{\beta}_{init}$) is set to be (7–32 Hz) band-pass filter (the fifth order Butterworth bandpass filter was used in this study). Hence, the frequency bins in $\boldsymbol{\beta}_{init}$ that belong to (7–32 Hz) are all ones, other frequency bins are set to be zeros. Ordinary CSP with (7–32 Hz) band-pass filtered signals are also computed with optimal feature selection algorithms for comparison. Usually more than one pair of CSP features are used, we choose three pairs of spatial spectral patterns, i.e., $m = 3$, $J = 2m = 6$ in this study.

4. Results

In this paper, we focus on optimizing several parameters for spatial spectral patterns rather than to deal with small training dataset problems. Hence we use all the 280 trials of dataset IVa to perform cross-validation rather than to use the splitting of dataset IVa in competition for convenient of analysis. Ten-fold cross-validation is used to assess the performance of extracted features. In each fold of this procedure, each nine parts are used as the training dataset, the remaining one part is used as the test dataset. The 10-fold cross-validation is then repeated 10 times to get the 10×10 cross-validation results.

In order to investigate the effect of feature selection, the cross-validation results with and without feature selection are reported separately. Firstly we show the cross-validation results without feature selection in Section 4.1. The SWCSP and CSP features are extracted from fixed setting, respectively. The corresponding classification results are reported in this section. Secondly, the feature selection methods including filter ones and wrapper method are applied to SWCSP and CSP features. The details are reported in Section 4.2.

4.1. Cross-validation without feature selection

Firstly, we test whether the time segments make any difference in the performance of extracted features. Table 2 shows the 10×10 cross-validation classification accuracy of SWCSP and CSP features in different time segments with 118 full channels. The best performance for each subject by SWCSP and CSP are highlighted in bold and bold italic font, respectively. From the table, we can see that the most active time segment by SWCSP is

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

[0.5,2.5] s for all the subjects except subject 'aw', whose most active one is [0.5,3.5] s. While the most active time segment by CSP is [0.5,3.5] s for most of the subjects. The results suggest the best time segment might be subject specific, it might be helpful to choose most active time segment for different subjects. Also we can see that SWCSP features is superior than CSP features in most situations, especially for subjects 'aa' and 'aw'. The broad band CSP features are not the best one for each subject, despite it might be a convenient choice.

Next, we use fixed time segment [0.5,2.5] s and [0.5,3.5] s for all the subjects to test the performance of SWCSP and CSP, respectively, with different channel configurations based on the results of Table 2. Although, the most active time segment for each subject and method is different, we use the average best performance to choose the most reactive time segment for SWCSP and CSP. The average performance of all the five subjects with SWCSP and CSP features is shown in Fig. 3. Clearly, the 118 full channel configuration is not the best choice. The average performance of SWCSP and CSP features with channel configuration 7 (SWCSP: $92.1 \pm 4.6\%$, CSP: $89.3 \pm 5.2\%$) is better than all the others, however, the situation is different for every single subject. The results in the next subsection show the performance of SWCSP and CSP with feature selection by the wrapper method will further improve the accuracy nearly 1% on average for the dataset (SWCSP: $93.0 \pm 4.3\%$, CSP: $90.7 \pm 5.0\%$).

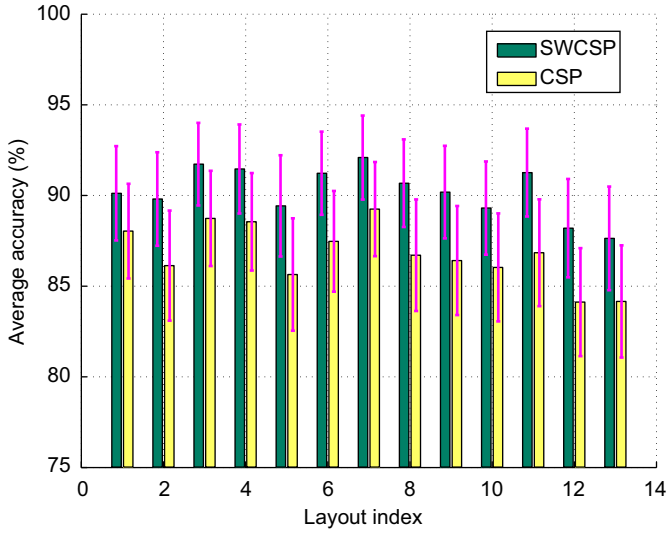


Fig. 3. Average accuracy (five subjects) comparison between SWCSP and CSP features with 13 different channel configurations, fixed time segment [0.5,2.5] s, [0.5,3.5] s for SWCSP and CSP, respectively. The horizontal axis shows the index of channel configuration, the vertical axis is the 10×10 cross-validation average classification accuracy of five subjects.

4.2. Cross-validation with feature selection

The proposed various feature selection algorithms are performed on the training dataset, the remaining test dataset is kept unseen and used to evaluate the efficiency of the feature selection algorithms. In each fold of this procedure, the SWCSP and CSP features are extracted for each setting (n_t,n_c), respectively. For example, the SWCSP and CSP features for subject 'aa' are extracted from each of the 13 channel configurations and 4 time segments. Totally, 52 SWCSP and CSP feature sets denoted as $F^{SWCSP}_{n_t,n_c}$ and $F^{CSP}_{n_t,n_c}$ are extracted from each training dataset. Then the mutual information algorithm, Fisher ratio algorithm and SVM based wrapper method are used to select the best setting (n_t^*,n_c^*), respectively. The classification accuracy of the proposed feature selection algorithms is evaluated on the remaining test dataset by SWCSP and CSP features in the best setting.

Table 3 shows the results of 10×10-fold cross-validation performed on dataset IVa. The best performance for each subject is highlighted in bold. In the column of "CSP", the results are derived from the CSP features with selection of best setting by mutual information (MI), Fisher ratio (Fs) and wrapper method (wrapper), respectively. The results for optimal spatial spectral patterns are those columns corresponding to SWCSP. Note that the mutual information (MI) and Fisher ratio (Fs) are computed for every training dataset in each fold of the 10×10 cross-validation. Take subject 'aa' with channel configuration 2 and time segment 1 as an example. In each 10-fold cross-validation, there are 252 trials for training, the SWCSP features are generated by signals in the channels shown in Fig. 2 (channel configuration 2 (64ch_std)) and in the time segment [0.5,2.5] s after the visual cue onset. The spatial spectral filters $\{\mathbf{w_j},\boldsymbol{\beta}_j\}_j = 1,\ldots,J$ contain three pairs of spatial filter $\mathbf{w_j}$, which is a 64-dimensional vector and spectral filter $\boldsymbol{\beta}_j$ which is a 200-dimensional vector in this situation. The mutual information between the training dataset and class label is computed and denoted as $I(F^{SWCSP}_{2,1},\Omega)$. Also the Fisher ratio is computed as $R_F(F^{SWCSP}_{2,1},\Omega)$. The best setting (n_t^*,n_c^*) is chosen as the maximum among all the $I(F^{SWCSP}_{n_t,n_c},\Omega)$ or $R_F(F^{SWCSP}_{n_t,n_c},\Omega)$. Then the classification accuracy is evaluated by SWCSP features in best setting $(F^{SWCSP}_{n_t^*,n_c^*})$ on the remaining 28 test trials. The wrapper one is the method of selecting features by the SVM classifier, where the best setting is selected by 2×10 nested cross-validation on the training dataset. This feature selection is only performed on the first training dataset, since the computational burden increases greatly if all the 10×10 training dataset are used to perform feature selection by nested 2×10 cross-validation. To alleviate the unfairness between the wrapper method and the other two filter ones, we perform the above 10×10 cross-validation by the wrapper method another 10 times. Then we sort the average performance of all the five subjects in descending order and report the median performance of this method.

Table 3
Average 10×10 cross-validation classification accuracies $\pm$ standard deviation (%) of CSP, SWCSP with various feature selection algorithms. The best performance for each subject is highlighted in bold.

Subject	CSP			SWCSP		
	MI	Fs	Wrapper	MI	Fs	Wrapper
aa	84.6 ± 6.4	83.8 ± 6.6	86.6 ± 6.2	93.1 ± 5.3	90.9 ± 5.3	$\mathbf{94.2 \pm 4.5}$
al	98.6 ± 2.4	97.3 ± 3.0	98.7 ± 2.1	98.8 ± 2.1	99.0 ± 1.9	$\mathbf{99.2 \pm 1.8}$
av	75.3 ± 8.5	75.3 ± 9.1	76.2 ± 9.3	73.9 ± 9.9	74.6 ± 9.3	$\mathbf{78.0 \pm 8.5}$
aw	94.8 ± 5.2	92.5 ± 5.9	95.7 ± 3.6	97.8 ± 2.6	$\mathbf{98.3 \pm 2.3}$	97.7 ± 2.9
ay	95.1 ± 4.3	95.1 ± 4.4	$\mathbf{96.2 \pm 3.6}$	93.3 ± 4.7	94.3 ± 4.5	95.6 ± 3.8
Average	89.7 ± 5.3	88.8 ± 5.8	90.7 ± 5.0	91.4 ± 4.9	91.4 ± 4.6	$\mathbf{93.0 \pm 4.3}$

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

From Table 3, we can see that the group of CSP with feature selection yields a little improvement over CSP ones in Table 2. The SWCSP with feature selection is superior than CSP according to the average classification accuracy. But this is not the truth for subjects 'av' and 'ay' with MI and Fs feature selection. That might be caused by overfitting or underfitting on these two subjects. The SWCSP with feature selection by the wrapper method yields the best average results. Since the best spatial spectral patterns are selected among the whole settings, the best average accuracy can be expected. In this study, the feature selection by the filter method performs inferior than the wrapper method. The reason might be that the features behave more likely non-stationary and non-Gaussian in the motor imagery based BCI. Hence, the filter approach might perform inferior than the wrapper method. In the next section, we show that the best time segment and channel configuration is subject dependent.

In Table 4, we compare the average classification errors with several typical reported results on dataset IVa. Because several papers [12,33,34] report classification errors and paper [11] reports classification accuracy, we transform classification accuracy (the last column of Table 3) into classification errors. Our method performs better than all the other results except the ISSPL one. The reason might be, like the author said, for ISSPL the temporal filter, channels, time window and the number of spatial filter were tuned according to the winning entry of the dataset [11]. The selection of channel and time segment is unclear in the paper. These manually tuned parameters are sensitive to researchers' experience and are hard to be generalized to other applications. However, we report median performance of SWCSP features with wrapper method to select the features. Our method only use SWCSP features and 13 typical channel configuration and 4 rough split of time segments. By the way, the computation of our method might be much more efficient than the ISSPL one. The ISSPL algorithm solves a regularized quadratic programming problem to optimize the spectral filters. It is time consuming to optimize the regularization term for ISSPL [35].

To further investigate the performance of proposed algorithm with the competition dataset, we report the results following the exact competition procedure by the proposed algorithm. The preprocessing procedure is exactly the same as those described in Section 3.2, the difference is that we use the same training dataset provided in BCI competition to perform feature selection and the same test set to validate the proposed algorithm. The wrapper method is used to perform feature selection. The challenge of dataset IVa is to validate the efficiency of algorithms to deal with various sizes of training dataset. For subjects 'aa', 'al', 'av', 'aw' and 'ay' the corresponding training dataset contains 168, 224, 84, 56 and 28 trials, respectively. In Table 5, we list the first four results of 14 submissions of the dataset IVa of BCI competition III. Without the need of exhausting tuning parameters, the SWCSP with feature selection from 4 given time segments and 13 channel configurations takes the third place compared to all the submissions. Note that we do not take any adaptation or extension of training dataset by classified test samples as those reported by the first and second contributors [36], hence, our results are still to be improved by considering such operations. On the other hand, the combination of several features including readiness potential might be the reason for the winner to have superior results.

5. Discussion

5.1. Best time segment and channel configuration

Based on the results of the last section, we focus our discussion on SWCSP features. The best time segment and channel configuration are subject and feature selection method dependent. Take subjects 'aa' and 'av' as examples shown in Fig. 4. The left column shows the selection results of channel configuration and the right column corresponds to the selection results of time segment. The horizontal axis corresponds to the label of channel configuration or time segment. Since the channel configuration or time segment selected by Mutual information or Fisher ratio is dependent on the training dataset which is split by a 10×10 cross-validation procedure. The value of the vertical axis shows the accumulated number where the corresponding channel configuration or time segment is selected. While only the first training dataset is used to select the time segment and channel configuration by 2×10 nested cross-validation on training dataset for wrapper method. The reason is explained in the last section. Hence, the selection of best setting is determined by the first splitting of training dataset and its accumulated number is always 100.

We can see that the electrodes around central and posterior motor cortex is preferred by subject 'aa' from Fig. 4. Although the 118 full channels contain more information, the classification accuracy is 85.8 ± 6.5 (Table 2), which is much lower than that 94.1 ± 4.5 in Table 3. The similar phenomenon is found in subject 'av'. From the next section, we can see that the topography of spatial patterns is not neurophysiological plausible for 'aa' and 'av' with 118 full channel configuration. We make a simple discussion on this. The CSP is sensitive to outliers, which has been reported by several researchers. Unfortunately, the SWCSP shares the same disadvantage, since it uses the same strategy as CSP to optimize spatial filters. The discriminative criterion in

Table 5
Comparison of classification accuracies by standard competition procedure. The training trials for subject 'aa', 'al', 'av', 'aw' and 'ay' are 168, 224, 84, 56 and 28, respectively. The wrapper method is used to perform feature selection for SWCSP and CSP.

Method	Mean acc (%)	aa (%)	al (%)	av (%)	ay (%)	aw (%)
First	94.17	95.5	100.0	80.6	100.0	97.6
Second	85.12	89.3	98.2	76.5	92.4	80.6
SWCSP	83.92	83.0	100.0	73.5	91.5	82.1
Third	83.45	82.1	94.6	70.4	87.5	88.1
CSP	77.98	83.0	100.0	63.8	78.1	81.7
Fourth	72.62	83.9	100.0	63.3	50.9	88.1

Table 4
Comparison of average classification errors $\pm$ standard deviation (%) provided by other researchers.

Subject	SWCSP	CSP [33]	SBCSP (MC) [12]	SBCSP (RBE) [12]	SWCSP [11]	ISSPL [11]	FBCSP [34]	DFBCSP [34]
aa	5.8 ± 4.5	8.5 ± 5.4	10.7 ± 5.6	9.2 ± 4.5	13.2 ± 6.3	6.4 ± 3.3	6.9 ± 5.8	9.8 ± 5.6
al	0.8 ± 1.8	0.8 ± 1.8	1.4 ± 1.8	2.2 ± 3.4	2.5 ± 2.9	0.0 ± 0.0	1.0 ± 2.4	1.3 ± 2.9
av	22.0 ± 8.5	29.1 ± 8.2	29.6 ± 5.3	31.0 ± 7.3	23.2 ± 9.4	20.7 ± 7.1	31.0 ± 14.2	22.2 ± 9.9
aw	2.3 ± 2.9	3.1 ± 2.8	4.3 ± 4.0	4.2 ± 3.3	6.1 ± 3.8	0.4 ± 1.1	4.9 ± 8.9	2.1 ± 3.7
ay	4.4 ± 3.8	5.3 ± 3.8	4.3 ± 2.8	5.0 ± 3.4	7.5 ± 3.9	1.4 ± 3.0	6.2 ± 9.7	5.8 ± 5.3
Average	7.0 ± 4.3	9.4 ± 4.4	10.1 ± 3.9	10.3 ± 4.4	10.5 ± 5.3	5.8 ± 2.9	10.0 ± 8.2	8.2 ± 5.5

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

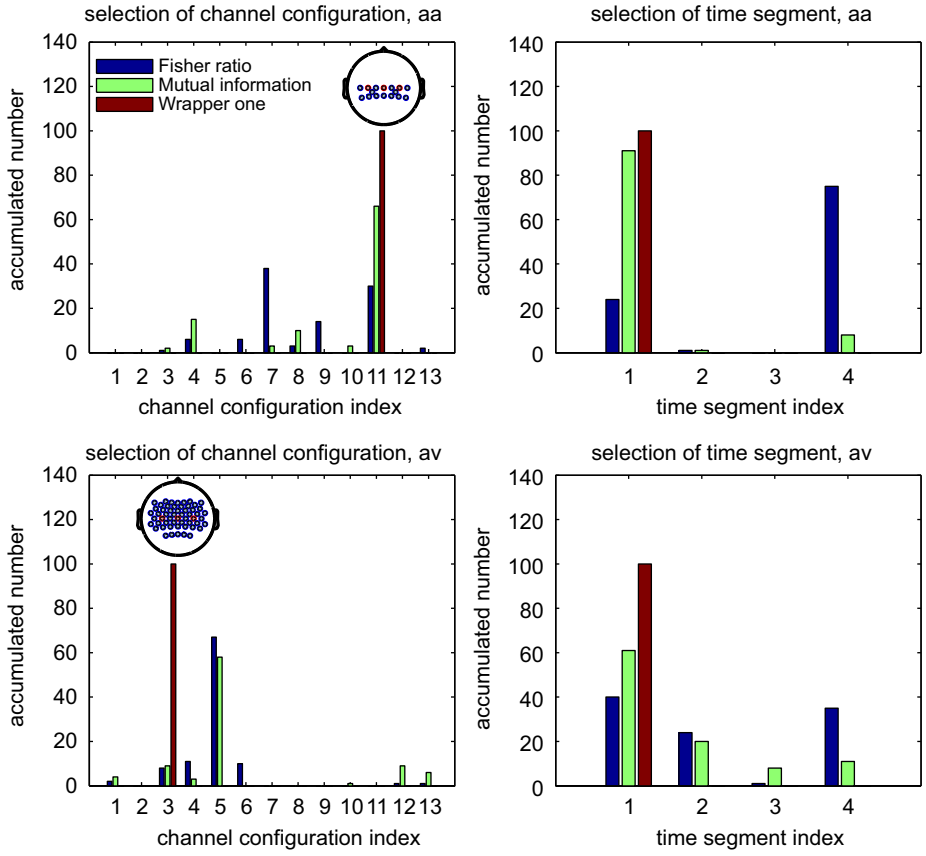


Fig. 4. Selection results of channel configuration and time segment by three different algorithms (mutual information, Fisher ratio and wrapper method) for subject 'aa' (first row) and 'av' (second row). In the left column, the horizontal axis shows the sequence of channel configuration (corresponding to the number of channel configuration in Fig. 2); the vertical axis shows 'accumulated number', where 'accumulated number' means how many times the corresponding channel configuration is selected by 10×10-fold cross-validation (the sum of accumulated number for each method is 100). The best channel configuration selected by wrapper one is shown above the corresponding number of label. In the right column, the horizontal axis shows the label of time segment; the vertical axis shows accumulated number.

Eq. (4) uses the separation of mean power of two classes, which might be sensitive to outliers [37]. Sometimes, artifacts like blinking or other muscle movements happen to be unevenly distributed in two class conditions. Then CSP algorithm, which uses channel configuration containing distorted EEG signals in some channels, will capture the artifact with very high eigenvalue. Taking subject 'aa' as an example, we plot the log power of signals in the surrogate channels, which is filtered by the most significant spatial and spectral filters, in Fig. 5. The maximizing power of right hand versus right foot in full channel configuration shows nearly no discriminability because of artifacts which is shown on the left plot of Fig. 5(a). The similar results have been reported in paper [37, Fig. 8]. The class-specific box-plots in Fig. 5(a) show no difference in median of the variances. While the discriminability for optimal channel configuration becomes better (see Fig. 5(c)). This might explain why fewer channels in configuration 11 improve the classification accuracy. The situation is a little different for subject 'av'. The electrodes around central, posterior and pre-motor cortex are all selected, see the left plot in lower row of Fig. 4. More electrodes may provide more information for SWCSP algorithm. However, the full channel configuration is still not the best choice in this example.

The results for best time segment are simpler than channel configuration. Time segments [0.5,2.5] s and [0.5,3.5] s are chosen by the algorithms most frequently, see the right column of Fig. 4. For real application, time segment [0.5,2.5] s is preferred since the time window is short and the response is more quickly. Nonetheless, more or less differences exist among different methods and different persons.

5.2. Optimal spatial spectral patterns

From the results of Section 4, the SWCSP features selected by the wrapper method provide the best average performance. Hence, this section discusses the spatial and spectral patterns selected by the wrapper method. Let us continue the examples of 'aa' and 'av', whose topography of spatial patterns and spectral filters are plotted in Figs. 6 and 7, respectively. Only the most significant spatial patterns, which are corresponding to the spatial filters with largest and smallest eigenvalues, are shown in the figures. The spatial patterns of subject 'aa' and 'av' with full channel configuration (the plots are shown in upper rows of Fig. 6 and 7, respectively), derived by SWCSP, show similar concentration areas as those that are reported in Fig. 5 of paper [13]. However, the counterparts with channel configuration selected by the wrapper method show more neurophysiological plausibility, see the lower rows of Figs. 6 and 7. At the same time, these SWCSP features get higher classification accuracy, where the average accuracy improves 8.4% and 2.4% for 'aa' and 'av', respectively. From Fig. 6, we may conclude that the ERD/ERS rhythm activity is contaminated by the abnormal EEG activity for subject 'aa', which leads to focus around electrodes 'FC6' and 'FT8'. After channel configuration selection by the wrapper method, the best channel configuration excludes the abnormal EEG electrodes. Hence, the ERD/ERS rhythm activity concentrates around the mid-central region (near foot representation area [4,38]) in optimal channel configuration rather than focus on the electrodes 'FC6' and 'FT8' (see plots on the left column of Fig. 6). The similar effect can be found in subject 'av' shown in Fig. 7, but this time the

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

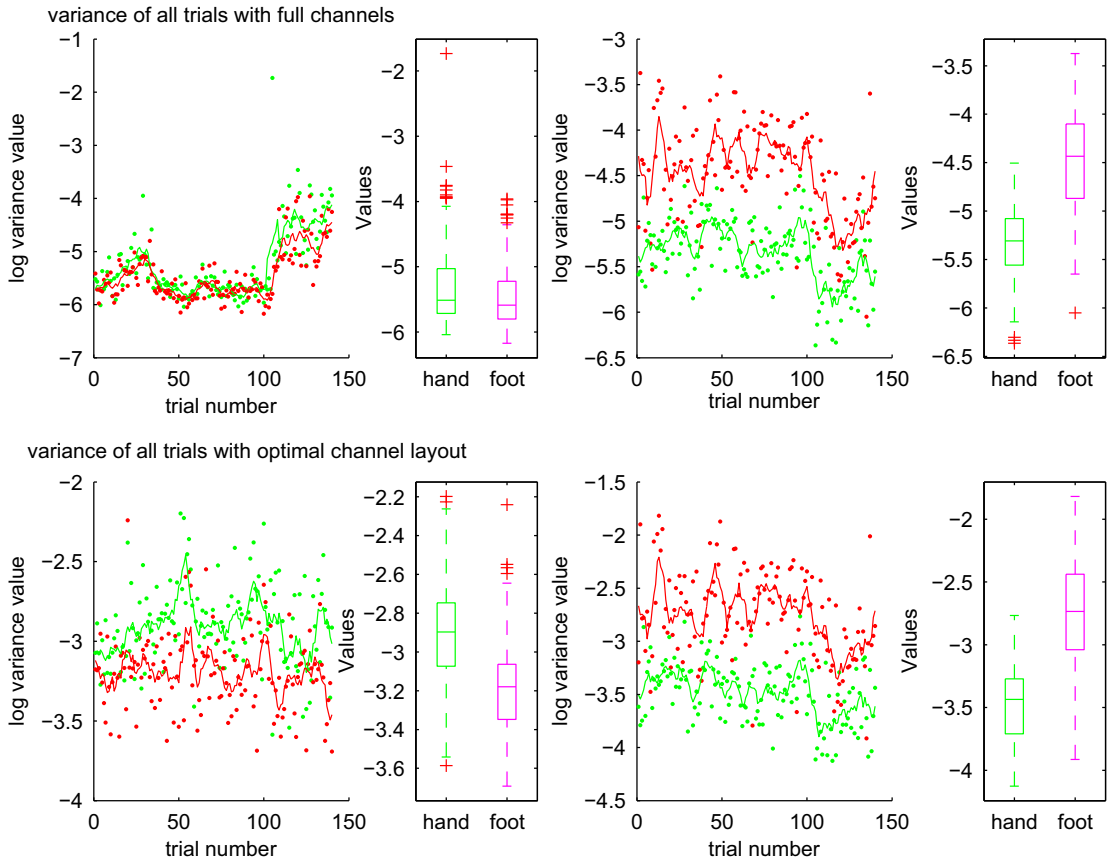


Fig. 5. This example is drawn from dataset 'aa'. In the first row, the common spatial filters are derived from the 118 full channel configuration. In the second row, the common spatial filters are derived from best channel configuration selected by wrapper method. The left column (a) and (c) shows log variance of the CSP surrogate channel which is corresponding to the largest eigenvalue of maximizing power of right hand versus right foot (green: right hand imagery, red: right foot imagery). The horizontal axis shows the number of trial in chronological order. The vertical axis shows log variance value of the trial. The right plot of (a) and (c) shows class-specific box-plots. The right column (b) and (d) shows log variance of the reverse CSP surrogate channel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

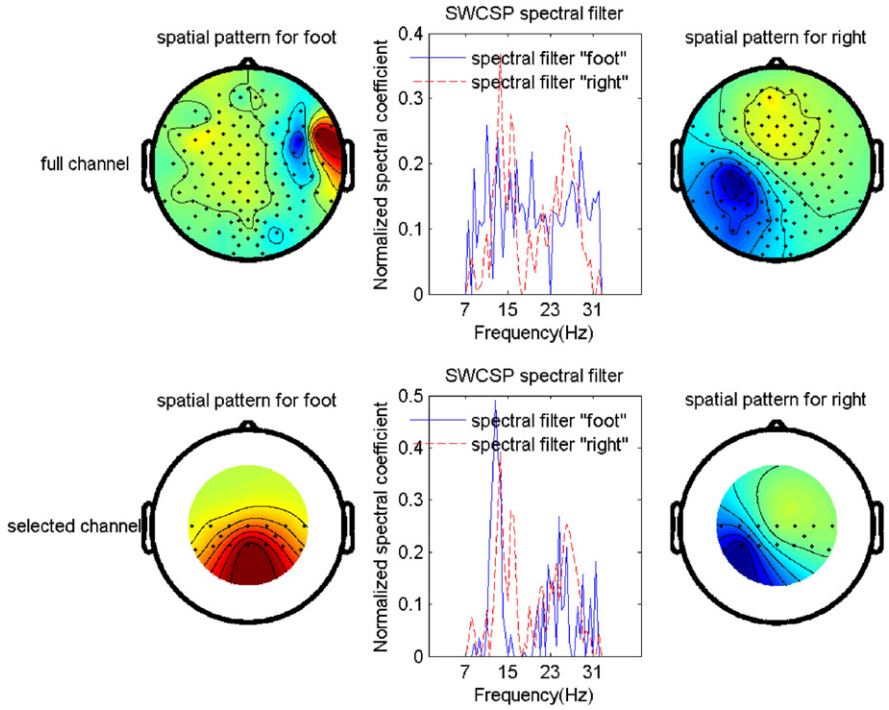


Fig. 6. Comparison of spatial patterns and spectral filters with 118 full channels (first row) and selected channels by wrapper method (second row). The topography is generated from dataset 'aa'. The most significant spatial patterns, which are corresponding to the spatial filters with largest and smallest eigenvalues, are shown as 'spatial pattern for foot' and 'spatial pattern for right', respectively. Their corresponding spectral filters are shown in the middle plots. The SWCSP spatial patterns with full channels for right foot are highly influenced and concentrate on electrodes 'FC6' and 'FT8'. While the counterparts with selected channels shift to the mid-central foot representation area.

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

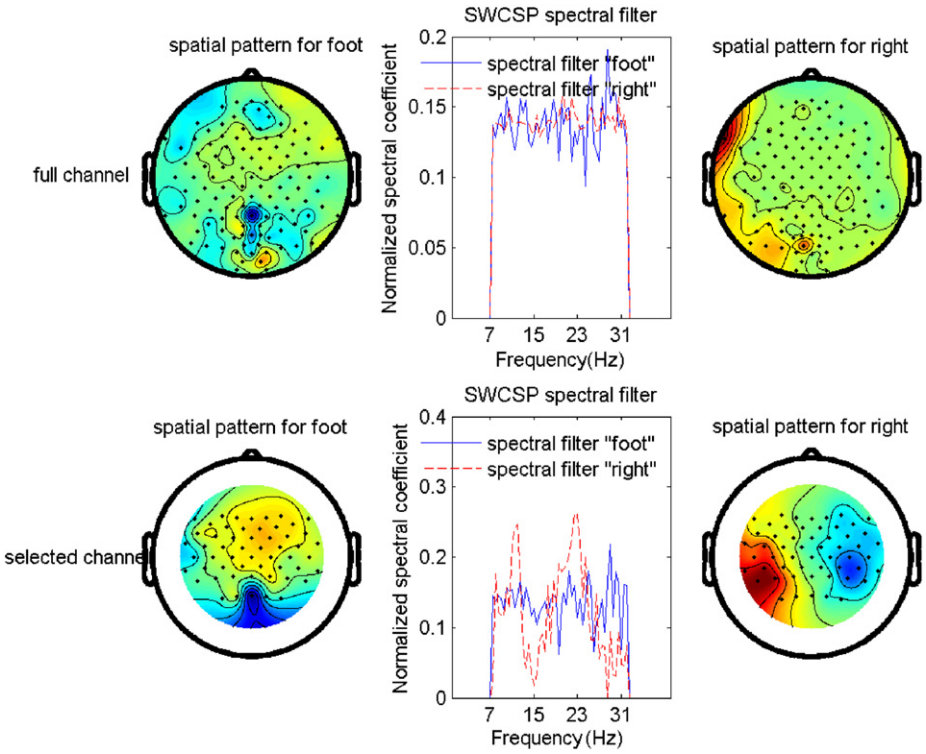


Fig. 7. Comparison of spatial patterns and spectral filters with 118 full channels (first row) and selected channels by wrapper method (second row). The topography is generated from dataset 'av'. The most significant spatial patterns, which are corresponding to the spatial filters with largest and smallest eigenvalues, are shown as 'spatial pattern for foot' and 'spatial pattern for right', respectively. Their corresponding spectral filters are shown in the middle plots. The SWCSP spatial patterns with full channels for right hand are highly influenced and concentrate on electrodes 'FT9'. While the counterparts with selected channels shift to the hand representation area. The spectral filters for subject 'av' are rather flat compared to those for subject 'aa'.

unreasonable spatial pattern is for right hand. With the help of feature selection by the wrapper method, the spatial pattern for right hand is more interpretable(shift focus on electrode 'FT9' to hand representation area). Note that, like CSP, SWCSP is also not a source localization algorithm [37], the result presented here cannot correspond to the physiological phenomenon exactly.

The spectral filter in ordinary CSP takes homogeneous weighting of spectrum, while SWCSP takes nonhomogeneous weighting of spectrum (see middle plots of Figs. 6 and 7). The weighting coefficient is adjusted according to subject-specific EEG signals. More important frequency bands like mu rhythm (8–12) Hz and beta rhythm (16–24) Hz usually get larger weight coefficients, while signals in other frequency bands (might be noise) usually get smaller or even zero weight coefficients. This might be the reason for improvement of the classification accuracy by SWCSP algorithm, especially for some subjects like 'aa' and 'aw' in this dataset. While for subject 'av', the spectral filters are apt to homogeneous weighting of spectrum (see middle plots of Fig. 7. The weightings of spectrum for subject 'av' are relatively flat compared to those of 'aa'.) that might be the reason why the difference between SWCSP and CSP is not statistically significant for some subjects.

6. Conclusion

Machine learning approach used in BCI aims to translate humans' intention as accurately as possible. In two-class motor imagery classification, there are several factors that play important roles in improving the subjects' performance such as time segment, channel configuration and frequency bands proposed in this paper. Several previous studies propose methods to solve these problems separately. Few of papers have considered to optimize these parameters jointly to further improve subjects' performance. In this paper, we attempt to propose a machine learning approach to solve these optimization jointly. Optimal frequency bands are automatically weighted by SWCSP algorithm. The best time segment and channel configuration are selected by employing various feature selection algorithms. In the motor imagery BCI setting, the feature selection by the wrapper method shows superiority than filter ones. After feature selection by the wrapper method, the spatial patterns show more neurophysiological interpretable results and also better cross-validation average accuracy. Although part of the algorithm relies on some statistical assumption, the improvement of classification suggests that the mild assumption might be reasonable for two-class motor imagery BCI.

Acknowledgment

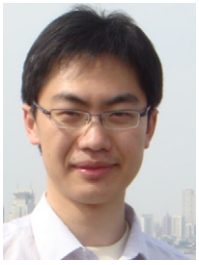
This work was supported by National Basic Research Program (973 Program) of China (No. 2011CB013305), National Natural Science Foundation of China (No. 51075265), the Science and Technology Commission of Shanghai Municipality (Grant No. 11JC1406000).

References

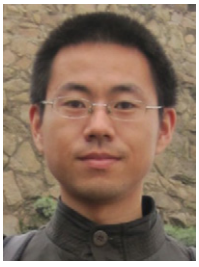
[1] G. Dornhege, B. Blankertz, G. Curio, K. Muller, Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms, IEEE Trans. Biomed. Eng. 51 (6) (2004) 993–1002.
[2] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, T. Vaughan, Brain–computer interfaces for communication and control, Clin. Neurophysiol. 113 (6) (2002) 767–791.
[3] Available from: 〈http://www.bbci.de/competition/〉.
[4] G. Pfurtscheller, F. Lopes da Silva, Event-related EEG/MEG synchronization and desynchronization: basic principles, Clin. Neurophysiol. 110 (11) (1999) 1842–1857.

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004

[5] M. Grosse-Wentrup, M. Buss, Multiclass common spatial patterns and information theoretic feature extraction, IEEE Trans. Biomed. Eng. 55 (8) (2008) 1991–2000.
[6] H. Ramoser, J. Müller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial EEG during imagined handmovement, IEEE Trans. Rehab. Eng. 8 (4) (2000) 441–446.
[7] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[8] S. Lemm, B. Blankertz, G. Curio, K. Muller, Spatio-spectral filters for improving the classification of single trial EEG, IEEE Trans. Biomed. Eng. 52 (9) (2005) 1541–1548.
[9] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, K. Muller, Combined optimization of spatial and temporal filters for improving brain–computer interfacing, IEEE Trans. Biomed. Eng. 53 (11) (2006) 2274–2281.
[10] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, K. Müller, Spectrally Weighted Common Spatial Pattern Algorithm for Single Trial EEG Classification, Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep. 40.
[11] W. Wu, X. Gao, B. Hong, S. Gao, Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL), IEEE Trans. Biomed. Eng. 55 (6) (2008) 1733–1743.
[12] Q. Novi, C. Guan, T. Dat, P. Xue, Sub-band common spatial pattern (SBCSP) for brain–computer interface, in: The Third International IEEE/EMBS Conference on Neural Engineering, 2007 (CNE'07, 2007), pp. 204–207.
[13] K. Ang, Z. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (FBSCP) in brain–computer interface, in: Proceedings of IEEE International Joint Conference on Neural Networks, IEEE, June 2008, pp. 2390–2397.
[14] M. Grosse-Wentrup, K. Gramann, M. Buss, Adaptive spatial filters with predefined region of interest for eeg based brain–computer-interfaces, Adv. Neural Inf. Process. Syst. 19 (2007) 537–544.
[15] X. Lei, P. Yang, P. Xu, T. Liu, D. Yao, Common spatial pattern ensemble classifier and its application in brain–computer interface, J. Electr. Sci. Technol. China 7 (1) (2009) 17–21.
[16] G. Huang, G. Liu, J. Meng, D. Zhang, X. Zhu, Model based generalization analysis of common spatial pattern in brain–computer interfaces, Cognit. Neurodyn. 4 (3) (2010) 217–223.
[17] J. Farquhar, N. Hill, T. Lal, B. Schölkopf, Regularised CSP for sensor selection in BCI, in: Proceedings of the Third International Brain–Computer Interface Workshop and Training Course, 2006, pp. 14–15.
[18] X. Yong, R. Ward, G. Birch, Sparse spatial filter optimization for EEG channel reduction in brain–computer interface, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008), pp. 417–420.
[19] M. Arvaneh, C. Guan, K. Ang, H. Quek, Optimizing the channel selection and classification accuracy in EEG-based BCI, IEEE Trans. Biomed. Eng. 58 (6) (2011) 1865–1873.
[20] C. Sannelli, T. Dickhaus, S. Halder, E. Hammer, K. Müller, B. Blankertz, On optimal channel configurations for SMR-based brain–computer interfaces, Brain Topogr. 23 (2) (2010) 186–193.
[21] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[22] Z. Xianda, Matrix Analysis and Applications, Tsinghua University Press, Beijing, 2004.
[23] H. Zhang, Z. Chin, K. Ang, C. Guan, C. Wang, Optimum spatio-spectral filtering network for brain–computer interface, IEEE Trans. Neural Networks 22 (1) (2011) 52–63.
[24] K. Ang, Z. Chin, H. Zhang, C. Guan, Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs, Pattern Recognition 45 (6) (2012) 2137–2144.
[25] K. Ang, C. Quek, Rough set-based neuro-fuzzy system, in: International Joint Conference on Neural Networks (IJCNN'06), IEEE, 2006, pp. 742–749.
[26] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, New York, 2001.
[27] A. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, vol. 18, Oxford University Press, USA, 1997.
[28] R. Kohavi, G. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[29] J. Li, L. Zhang, D. Tao, H. Sun, Q. Zhao, A Prior Neurophysiologic Knowledge Free Tensor-based Scheme for Single Trial EEG Classification, IEEE Trans. Neural Syst. Rehabil. Eng 17 (2) (2009) 107-115.
[30] C.-C. Chang, C.-J. Lin, LIBSVM—A Library for Support Vector Machines, Available from: 〈http://www.csie.ntu.edu.tw/cjlin/libsvm/〉.
[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2001.
[32] F.F.I.D. A. Group, C.B.F. of the Charité University Medicine Berlin Neurophysics Group, URL 〈http://www.bbci.de/competition/iii/desc_IVa.html〉.
[33] L. Song, J. Epps, Classifying EEG for brain–computer interface: learning optimal filters for dynamical system features, in: ICML '06 Proceedings of the 23rd International Conference on Machine Learning, Hindawi Publishing Corp., 2006, p. 8.
[34] K. Thomas, C. Guan, C. Lau, A. Vinod, K. Ang, A new discriminative common spatial pattern method for motor imagery brain–computer interfaces, IEEE Trans. Biomed. Eng. 56 (11) (2009) 2730–2733.
[35] D. Fang, G. Xiaorong, An iterative algorithm for learning spatio-temporal filters for motor imagery-based brain–computer interfaces, Chin. J. Biomed. Eng. 30 (1) (2011) 11–16.
[36] Available from: 〈 http://www.bbci.de/competition/iii/results/index.html〉.
[37] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K. Müller, Optimizing spatial filters for robust EEG single-trial analysis, IEEE Signal Process. Mag. 25 (1) (2008) 41–56.
[38] G. Pfurtscheller, C. Brunner, A. Schlögl, F. Lopes da Silva, Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks, Neuroimage 31 (1) (2006) 153–159.



Jianjun Meng got his Bachelor degree from School of Mechanical Engineering at Shanghai Jiao Tong University, China, in 2005. Now he is pursuing his PhD degree. His research interest is focused on BCI technology, EEG signal processing, feature extraction and classification.



Gan Huang got his Bachelor and Master degrees from Department of Mathematics at Southeast University, China, in 2005 and 2008, respectively. Now he is a PhD student in Shanghai Jiao Tong University, China. His research interests include BCI technology, EEG signal modeling, processing and classification.

Dingguo Zhang received the B.E. degree in Electrical Engineering from Jilin University and the M.E. degree in Control Engineering from Harbin Institute of Technology, China, in 2000 and 2002, respectively. He got the PhD degree from Nanyang Technological University, Singapore, in 2007. From 2006 to 2007, he worked as a research fellow in Biorobotics Lab of Nanyang Technological University, Singapore. In 2008, he was a postdoctoral fellow in LIRMM of CNRS, France. He is currently an Associate Professor in Institute of Robotics at Shanghai Jiao Tong University, China. His research interests include biorobotics, biomechatronics, biological cybernetics, and BCI/EMG technique.

Xiangyang Zhu obtained the B.S. degree from the Department of Automatic Control Engineering, Nanjing Institute of Technology, Nanjing, China, in 1985, the Master degree in instrumentation engineering from Southeast University, Nanjing, China, in 1989, and the PhD degree in Automatic Control Engineering from the same university in 1992. He joined the Department of Mechanical Engineering, Southeast University in 1995, after two years work as a postdoctoral fellow in Huazhong University of Science and Technology, Wuhan, China. Since June 2002, he has been a Professor of Shanghai Jiao Tong University, with a joint appointment in the Robotics Institute and the State Key Laboratory of Mechanical Systems and Vibrations. His current research interests include robotic manipulation planning, manufacturing automation, and biomechatronics. Zhu was awarded the National Science Fund for Distinguished Young Scholars in 2005, and appointed as a "Cheung Kong" Chair Professor in 2007.

Please cite this article as: J. Meng, et al., Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface, Neurocomputing (2012), http://dx.doi.org/10.1016/j.neucom.2012.11.004