A semi-supervised support vector machine approach for parameter setting in motor imagery-based brain computer interfaces

Jinyi Long · Yuanqing Li · Zhuliang Yu

Received: 25 January 2010 / Revised: 26 April 2010 / Accepted: 17 May 2010 / Published online: 8 June 2010
© Springer Science+Business Media B.V. 2010

Abstract Parameter setting plays an important role for improving the performance of a brain computer interface (BCI). Currently, parameters (e.g. channels and frequency band) are often manually selected. It is time-consuming and not easy to obtain an optimal combination of parameters for a BCI. In this paper, motor imagery-based BCIs are considered, in which channels and frequency band are key parameters. First, a semi-supervised support vector machine algorithm is proposed for automatically selecting a set of channels with given frequency band. Next, this algorithm is extended for joint channel-frequency selection. In this approach, both training data with labels and test data without labels are used for training a classifier. Hence it can be used in small training data case. Finally, our algorithms are applied to a BCI competition data set. Our data analysis results show that these algorithms are effective for selection of frequency band and channels when the training data set is small.

Keywords Electroencephalogram (EEG) · Motor imagery · Brain computer interface (BCI) · Channel · Frequency band · Semi-supervised learning

This work was supported by National Natural Science Foundation of China under Grants 60825306 and 60802068; Natural Science Foundation of Guangdong Province, China under grant 9251064101000012 and supported by the Fundamental Research Funds for the Central Universities, SCUT under grant 2009ZZ0055.

J. Long · Y. Li (✉) · Z. Yu
The College of Automation Science and Engineering, South China University of Technology, 510640 Guangzhou, China
e-mail: auyqli@scut.edu.cn

J. Long
e-mail: long.jinyi@mail.scut.edu.cn

Z. Yu
e-mail: zlyu@scut.edu.cn

Introduction

Brain computer interface (BCI) is a system that allows a brain to control a computer directly through a non-muscular communication channel, particularly for the paralyzed people who suffer severe neuromuscular disorders (Wolpaw et al. 2002; Dornhege 2007; Freeman 2007). BCIs can be invasive or non-invasive. Brain signals for non-invasive BCIs include electroencephalogram (EEG), magnetoencephalogram (MEG) and functional magnetic resonance imaging (fMRI), while brain signals for invasive BCIs include electrocorticogram (ECoG) and neuron spikes. A class of EEG-based BCIs which rely on the motor imagery of users are of particular interest to the BCI community since this type of BCIs have relative robust communication performance and can help to understand the underlying mechanism of motor imagery (Pfurtscheller et al. 2006).

Because of many facts such as low topographical resolution and high noise level, it has been a big challenge to extract effective features from EEG signals and perform classification. Until now, a lot of methods have been proposed for EEG signal analysis (see e.g. Muller et al. 2004; Blankertz et al. 2008) as well as its feature extraction and classification of EEG signals (Shenoy et al. 2006; Vidaurre et al. 2007). For a BCI system, there usually exist several parameters. For instance, channels and frequency band are two important parameters in motor-imagery based BCI systems. Parameter setting plays an important role for improving the performance of a BCI system.

This paper discusses channel and frequency band selection for motor-imagery based BCI systems. Generally, channels and frequency band are set separately. For instance, frequency band selection was discussed in Pre-genzer and Pfurtscheller (1999), Garrett et al. (2003) and

Blankertz et al. (2008), while channel selection was discussed in Schroder et al. (2003), Al-Ani and Al-Sukker (2006), Lal et al. (2004) and Wang et al. (2005). Blankertz et al. (2008) presented a heuristic procedure to automatically select the discriminant frequency band with a small number of channels. In Lal et al. (2004), channel selection was carried out by ranking all channels according to their relevance to experimental tasks and selecting a subset of channels. The state-of-the-art approaches for rating the relevance of channels include zero-norm optimization based-support vector machine (SVM), recursive feature elimination (RFE) and Fisher ratio (Lal et al. 2004). After ranking all channels, the number of selected channels was generally determined by the average classification accuracy obtained in cross-validation of training data set (Al-Ani and Al-Sukker 2006; Lal et al. 2004; Wang et al. 2005). In Schroder et al. (2003), genetic algorithm was used to directly select a subset of channels without channel ranking. However, heavy computational burden is a disadvantage of genetic algorithm.

In addition, the discriminant frequency band depends on individual subject and electrodes' position (Pregenzer and Pfurtscheller 1999). Hence, a BCI system benefits from joint channel-frequency selection (Pregenzer and Pfurtscheller 1999; Garrett et al. 2003; Lemm et al. 2005; Dornhege et al. 2006; Wu et al. 2008). In Pregenzer and Pfurtscheller (1999), several channel-frequency pairs were selected by distinction sensitive learning vector quantization (DSLQV) method. However, only two channels ('C3' and 'C4') were considered. Genetic algorithm was used to select several channel-frequency pairs, in which only six channels were considered (Garrett et al. 2003). In Lemm et al. (2005), Dornhege et al. (2006), Wu et al. (2008), common spatial pattern (CSP) was modified to exploit temporal structure in EEG for determining the subject-specific 'optimal' temporal filters. In Wu et al. (2008), an iterative spatio-spectral patterns learning (ISSPL) method was proposed for improving the classification performance by optimizing the spatial and spectral filter, in which channels and frequency band were jointly selected through the weights of the spatial and spectral filter. However, these modified CSP methods need sufficient training dataset that plays an important role to obtain a reliable transformation matrix. Thus these algorithms are supervised ones. Considering that it is either time consuming or even intractable to collect sufficient training data (Wolpaw et al. 2000), it is worthy to discuss parameter setting for a BCI system when the training data set is insufficient.

Generally, if the number of trials in a training set is insufficient for training a reliable classifier through a supervised learning method, then we regard it as a small training set. In (Li and Guan 2008), an iterative semi-supervised SVM algorithm was proposed for joint feature

extraction and classification for small training data set. In this paper, we extend this algorithm for joint channel-frequency band selection in the case when the training data set is small. In our proposed algorithms, both training data with labels and testing data without labels are used to jointly select the channels and frequency band. Fisher ratio is applied to rank channels, while Rayleigh coefficient distance is applied to determine the number of selected channels and frequency band. The computation of Fisher ratio and Rayleigh coefficient is based on both training data and test data.

To validate the proposed algorithm, we apply it to the data set IVa of BCI competition III in 2005 (Dornhege et al. 2004). Furthermore, we compare our algorithm with an existing method (Fisher ratio based cross-validation) for channel selection. Data analysis results show that our algorithm can effectively perform channel and frequency-band selection with small training data set and provide satisfactory classification accuracy.

The remaining part of this paper is organized as follows. In Section "A semi-supervised SVM algorithm for channel selection", an iterative semi-supervised SVM algorithm is proposed for channel selection with given frequency band. In Section "Frequency selection", the algorithm in Section "A semi-supervised SVM algorithm for channel selection" is first modified for frequency band selection, which is presented in Algorithm 2. In Section "Joint channel-frequency selection", an iterative algorithm is obtained through combining the previous two algorithms for joint channel-frequency selection. Data analysis results including the comparison with an existing method (Fisher ratio based cross-validation) are presented in Section "Data analysis". Finally, conclusions in Section "Conclusions" review the approach of this paper.

A semi-supervised SVM algorithm for channel selection

In this section, we first briefly describe Fisher ratio and Rayleigh coefficient, which are used for rating the relevance of each channel and determining the number of selected channels, respectively. Then we present the steps of a semi-supervised SVM algorithm for channel selection in the case of small training data set.

Fisher ratio

Fisher ratio is defined as the ratio of the interclass difference to the intraclass spread (Bishop 1995). Denote $P_{l,i}(f)$ ($l = 1, \dots, N$, $i = 1, \dots, d$) as the discrete power spectral density function of a segment of EEG signal which is extracted from the l th channel and i th trial. N denotes the

number of channels and d represents the number of trials. Then the power feature of each channel is calculated as,

$$\bar{P}_{l,i} = \sum_{f \in F} P_{l,i}(f), \quad (1)$$

where F is the selected set of frequency band indices.

The Fisher ratio score of the l th channel is defined as

$$FR_l = \frac{(\text{mean}(\bar{P}_{l,i}, i \in Cl_1) - \text{mean}(\bar{P}_{l,i}, i \in Cl_2))^2}{\sigma^2(\bar{P}_{l,i}, i \in Cl_1) + \sigma^2(\bar{P}_{l,i}, i \in Cl_2)}, \quad (2)$$

where Cl_1 and Cl_2 denote two classes of trials with labels being +1 and -1, respectively. Mean and σ represent mean and standard deviation, respectively.

The relevance of the l th channel is quantified by the Fisher ratio score FR_l , and $\{FR_l | l = 1, \dots, N\}$ are used for ranking all channels.

Rayleigh coefficient

After the channels are ranked through Fisher ratio, we then determine the number of selected channels. This is performed by Rayleigh coefficient.

Define $n-1$ subsets of channels Q_j ($j = 2, \dots, n$) which contain j channels with the highest Fisher ratio scores, where $n \leq N$. Here $j \geq 2$ means that we select at least two channels. We now define Rayleigh coefficient of Q_j .

Using a band-pass filter with frequency band F , we first filter a segment of EEG signal from the l th channel and i th trial, where $l \in Q_j$. The filtered EEG signal is denoted as a row vector $\mathbf{x}_{l,i}$. Furthermore, denote $X_i = [\mathbf{x}_{1,i}^T, \dots, \mathbf{x}_{j,i}^T]^T$, where $(\cdot)^T$ represents the transposition operators.

We can define the discriminant activity Γ^d and common activity Γ^c between two conditions, respectively, as follows

$$\begin{aligned} \Gamma^d &= \frac{1}{|Cl_1|} \sum_{i \in Cl_1} X_i(X_i)^T - \frac{1}{|Cl_2|} \sum_{i \in Cl_2} X_i(X_i)^T, \\ \Gamma^c &= \frac{1}{|Cl_1|} \sum_{i \in Cl_1} X_i(X_i)^T + \frac{1}{|Cl_2|} \sum_{i \in Cl_2} X_i(X_i)^T, \end{aligned} \quad (3)$$

where $|Cl_1|$ and $|Cl_2|$ denote the sizes of the sets Cl_1 and Cl_2 , respectively.

The definition of Rayleigh coefficient is

$$RC = \max_u \frac{u^T \Gamma^d u}{u^T \Gamma^c u}, \quad (4)$$

where u is designed to maximize the discriminative activity Γ^d simultaneously to minimize the common activity Γ^c (Blanchard and Blankertz 2004; Ramoser et al. 2000). u can be obtained by jointly diagonalizing the matrices Γ^d and Γ^c .

Suppose that U is a matrix which jointly diagonalizes the matrices Γ^d and Γ^c . Note that U can be obtained

Fig. 1 An illustrative example: The curve of Rayleigh coefficients with respect to the number of channels

through a standard procedure for joint diagonalization of two matrices (Blankertz et al. 2008). The first column $\mathbf{u}_1$ and the last column $\mathbf{u}_j$ are two generalized eigenvectors of Γ^d and Γ^c corresponding to the largest and the smallest generalized eigenvalue, respectively. Using $\mathbf{u}_1$ and $\mathbf{u}_j$, the Rayleigh coefficient of the subset of channels Q_j is calculated as follows (Li and Guan 2008):

$$RC_j = \frac{\mathbf{u}_1^T \Gamma^d \mathbf{u}_1}{\mathbf{u}_1^T \Gamma^c \mathbf{u}_1} + \left| \frac{\mathbf{u}_j^T \Gamma^d \mathbf{u}_j}{\mathbf{u}_j^T \Gamma^c \mathbf{u}_j} \right|. \quad (5)$$

From our experience of data analysis, the Rayleigh coefficient value would increase when the number j of channels increases. For instance, Fig. 1 shows the curve of Rayleigh coefficient value with respect to the number j of selected channels.

Hence, we cannot simply compare the Rayleigh coefficient values of the subsets of channels to determine the number j of selected channels (or the subset Q_j of channels). We first calculate difference,

$$IR(j) = RC_j - RC_{j-1}, \quad j = 3, \dots, n. \quad (6)$$

Then the number of channels j_0 corresponding to the maximal difference in (7) is determined as follows

$$IR(j_0) = \max(IR(j), j = 3, \dots, n). \quad (7)$$

Hence, the subset of selected channels is Q_{j_0} .

Channel selection for the case of small training data set

When the training data set is small, the subset of channels selected by Fisher ratio and Rayleigh coefficient are not reliable. This is because reliable Fisher ratio and Rayleigh coefficient rely on sufficient training data with labels. In this case, we propose to use semi-supervised SVM

technique for channel selection in which both labeled training data and unlabeled test data are used.

Our approach relies on a standard SVM classifier, which is commonly used in BCIs (Thulasidas et al. 2006; Meinicke et al. 2003; Kaper et al. 2004). More details about SVM can be found in (Vapnik 1998; Burges 1998). A standard SVM classifier for two-class problem can be described as

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \dots, d, \end{aligned} \quad (8)$$

where $\{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^n, i = 1, \dots, d\}$ are feature vectors extracted from a training data set, $\{y_i \mid y_i \in \{+1, -1\}, i = 1, \dots, d\}$ are the corresponding labels of $\{\mathbf{x}_i\}$, $\{\xi_i \mid \xi_i \geq 0, i = 1, \dots, d\}$ are slack variables, $C > 0$ is a regularization constant, $\mathbf{w}$ is a weight vector, b is a bias value.

Given two raw EEG data sets, including a small training data set D_c containing N_c trials of EEG signals with labels $y_i \in \{+1, -1\}$, ($i = 1, \dots, N_c$) and a test data set D_t containing N_t trials of EEG signals without labels, in the following, we present the outline of an iterative semi-supervised SVM algorithm for channel selection and motor imagery classification.

Algorithm 1 Summary of channel selection algorithm

Define: [1] Data vector construction function: $FV(\mathbf{X}(j), Q_i)$ consisting the operations of constructing new data set by removing the EEG signals of those channels not in Q_i .
[2] Iteration stopping criterion: the normalized difference between labels predicted in two successive iterations being less than a predefined threshold δ_0 .
Input: the training set $D_c = \{\mathbf{X}(1), \mathbf{X}(2), \dots, \mathbf{X}(N_c)\}$ and their corresponding labels $\{y_1, y_2, \dots, y_{N_c}\}$ the test set $D_t = \{\mathbf{X}(N_c + 1), \mathbf{X}(N_c + 2), \dots, \mathbf{X}(N_c + N_t)\}$ threshold δ_0 for stopping the iterations

$iter = 0$

$\mathbf{x}_k = \bar{\mathbf{W}}^T \mathbf{X}(k)$, where $k = 1, \dots, N_c + N_t$, and $\bar{\mathbf{W}}$ is a transformation matrix, which is composed by the first and last three columns of CSP spatial transformation matrix.

$\{\mathbf{w}, b\} = \text{SVMtrain}\{\{\mathbf{x}_j, y_j\} \mid j = 1, 2, \dots, N_c\}$ by solving Eq. 8

For $j = N_c + 1$ to $N_c + N_t$

$y_j = \text{sign}\{\mathbf{w}^T \mathbf{x}_j + b\}$

end

Repeat

$iter = iter + 1$

For $l = 1$ to N

with $\{\{\mathbf{X}_l(j), y_j\} \mid j = 1, \dots, N_c\}$ and $\{\{\mathbf{X}_l(j), y_j\} \mid j = N_c + 1, \dots, N_c + N_t\}$, obtain $FR(l)$ by Eq. 2

Algorithm 1 continued

End

rank $FR(1), \dots, FR(N)$ in a descending order, and the corresponding channel sequence is denoted as a vector Q

For $i = 2$ to N

$Q_i = \{Q(1), Q(2), \dots, Q(i)\}$, which defines an i -channel subset

Obtain the Rayleigh coefficient RC_{Q_i} according to Eq. 5.

end

$Q^{(s)} = \arg \max_{Q_i} \{RC_{Q_3} - RC_{Q_2}, \dots, RC_{Q_i} - RC_{Q_{i-1}}, \dots, RC_{Q_N} - RC_{Q_{N-1}}\}$

$\mathbf{X}_{Q^s}(k) = FV(\mathbf{X}(k), Q^s)$, for $k = 1$ to $N_c + N_t$

$\mathbf{x}_k = \bar{\mathbf{W}}^T \mathbf{X}_{Q^s}(k)$, where $k = 1, \dots, N_c + N_t$ and $\bar{\mathbf{W}}$ is a transformation matrix, which is composed by the first and last three columns of CSP spatial transformation matrix based on

$\{(\mathbf{X}_{Q^s}(1), y_1), \dots, (\mathbf{X}_{Q^s}(N_c + N_t), y_{N_c + N_t})\}$.

$\{\mathbf{w}, b\} = \text{SVMtrain}\{\{\mathbf{x}_k, y_k\} \mid k = 1, 2, \dots, N_c + N_t\}$ by solving Eq. 8, where y_j ,

$j = N_c + 1, \dots, N_c + N_t$ are the labels predicted in the previous iteration.

For $j = N_c + 1$ to $N_c + N_t$

$y_j = \text{sign}\{\mathbf{w}^T \mathbf{x}_j + b\}$

end

Until stopping criterion satisfied

Output: the subset of channels $Q^{(s)}$ and the corresponding labels $y_{Q^{(s)}j}$, $j = N_c + 1, \dots, N_c + N_t$

Remarks 1 (1) Another way for termination of the algorithm is to compare the set of channels $Q^{(k-1)}$ with $Q^{(k)}$. Suppose that the number of different channels between $Q^{(k-1)}$ and $Q^{(k)}$ is $n_{ch}(k)$, and calculate the ratio:

$$r_{ch}(k) = \frac{n_{ch}(k)}{|Q^{(k-1)}|}, \quad (9)$$

where $|Q^{(k-1)}|$ is the number of channels in $Q^{(k-1)}$. Obviously, $r_{ch}(k) = 0$ implies $Q^{(k-1)}$ and $Q^{(k)}$ are equal. From our data analysis, we found that this case often happens. If $r_{ch}(k)$ is less than a given constant saying δ_1 , the algorithm is terminated. Otherwise, perform the $(k + 1)$ th iteration. (2) In Algorithm 1, the frequency band F_0 is generally set by experience. In this paper, F_0 is 8–14 Hz for *mu* rhythm extraction.

Frequency selection

As mentioned above, there are two key parameters related to the performance of motor imagery-based BCIs: frequency band and channel. In the previous section, an

iterative semi-supervised SVM algorithm is presented to select channels. In the following, we present an algorithm for frequency selection with a given subset of channels which is similar to Algorithm 1.

Algorithm 2 Summary of frequency band selection algorithm

Define: [1] Data vector construction function: $FV(\mathbf{X}(j), Q_i)$ consisting the operations of constructing new data set by removing the EEG signals of those channels not in Q_i .
 [2] Iteration stopping criterion: the normalized difference between labels predicted in two successive iterations being less than a predefined threshold δ_2 .

Input: the training set $D_c = \{\mathbf{X}(1), \mathbf{X}(2), \dots, \mathbf{X}(N_c)\}$ and their corresponding labels $\{y_1, y_2, \dots, y_{N_c}\}$ the test set $D_t = \{\mathbf{X}(N_c + 1), \mathbf{X}(N_c + 2), \dots, \mathbf{X}(N_c + N_t)\}$ frequency sub-bands $\{F_1, \dots, F_{n_f}\}$ threshold δ_1 for stopping the iterations

$iter = 0$

For $i = 1$ to n_f

$\mathbf{X}_i(k) = BF(\mathbf{X}(k)), F_i$, where $k = 1, \dots, N_c + N_t$.

$\mathbf{x}_{i,k} = \bar{W}^T \mathbf{X}_i(k)$, where $k = 1, \dots, N_c + N_t$ and $\bar{W}$ is a transformation matrix, which is composed by the first and last three columns of CSP spatial transformation matrix.

$\{\mathbf{w}, b\} = \text{SVMtrain}\{(\mathbf{x}_{i,k}, y_k) | k = 1, 2, \dots, N_c\}$ by solving Eq. 8

For $k = N_c + 1$ to $N_c + N_t$

$y_{i,k} = \text{sign}\{\mathbf{w}^T \mathbf{x}_{i,k} + b\}$

end

end

Repeat

$iter = iter + 1$

For $i = 1$ to n_f

If $i = 1$

with $\{(\mathbf{X}_i(k), y_k) | k = 1, \dots, N_c\}$ and $\{(\mathbf{X}_i(k), y_{i,k}) | k = N_c + 1, \dots, N_c + N_t\}$,

obtain RC_{F_i} by Eq. 5

Else

with $\{(\mathbf{X}_i(k), y_k) | k = 1, \dots, N_c\}$ and $\{(\mathbf{X}_i(k), y_k) | k = N_c + 1, \dots, N_c + N_t\}$,

obtain RC_{F_i} by Eq. 5

End

End

$F^{(s)} = \arg \max_{F_i} \{RC_{F_1}, \dots, RC_{F_{n_f}}\}$

corresponding predicted labels $y_j = y_{F^{(s)}j}$,

$j = N_c + 1, \dots, N_c + N_t$

$\mathbf{X}_{F^{(s)}}(k) = BF(\mathbf{X}(k), F^{(s)})$, for $k = 1$ to $N_c + N_t$

$\mathbf{x}_k = \bar{W}^T \mathbf{X}_{F^{(s)}}(k)$, where $k = 1, \dots, N_c + N_t$ and $\bar{W}$ is a transformation matrix, which is composed by the first and last three columns of CSP spatial transformation matrix based on $\{(\mathbf{X}_{F^{(s)}}(1), y_1), \dots, (\mathbf{X}_{F^{(s)}}(N_c), y_{N_c}), (\mathbf{X}_{F^{(s)}}(N_c + 1), y_{N_c+1}^{(iter-1)}), \dots, (\mathbf{X}_{F^{(s)}}(N_c + N_t), y_{N_c+N_t}^{(iter-1)})\}$

$\{\mathbf{w}, b\} = \text{SVMtrain}\{(\mathbf{x}_k^{iter}, y_k) | k = 1, 2, \dots, N_c + N_t\}$ by solving Eq. 8, where $y_k, k = N_c + 1, \dots, N_c + N_t$ are the labels predicted in the previous iteration.

Algorithm 2 continued

For $k = N_c + 1$ to $N_c + N_t$

$y_k = \text{sign}\{\mathbf{w}^T \mathbf{x}_k + b\}$

end

Until stopping criterion satisfied

Output: the frequency sub-band $F^{(s)}$ and the corresponding labels $y_{F^{(s)}k}, k = N_c + 1, \dots, N_c + N_t$

Remark 2 The time interval is also an important parameter for motor imagery BCI system. Similar to frequency band selection, Algorithm 2 can be extended for determining the time interval in the current motor imagery BCI system. If the training data set is insufficient, semi-supervised learning can be used.

Joint channel-frequency selection

Until now, channels and frequency band selection have been presented separately. In this section, we combine Algorithms 1 and 2 to perform joint channel-frequency selection in an alternative manner. Initially given a frequency band F_0 , we apply Algorithm 1 to determine a subset of channels. This subset is denoted as $Q_s^{(1)}$ here. Based on Rayleigh coefficient, we use Algorithm 2 to select a frequency sub-band denoted as $F_s^{(1)}$. Repeating this process for several times, the final subset of channels and frequency band are obtained. The following is the outline of our algorithm.

Algorithm 3 Summary of joint channel-frequency selection algorithm

Define: [1] Iteration stopping criterion: the normalized difference between labels predicted in two successive iterations being less than a predefined threshold δ_3 .

[2] $Algorithm1\{F_0\}$ represents performing Algorithm 1 with a frequency band F_0 ; similarly, $Algorithm2\{Q_0\}$ means performing Algorithm 2 with a given subset of channels Q_0 .

Input: a frequency band F_0

$iter = 1$

Repeat

$Q_s = Algorithm1\{F_0\}$,

$F_s = Algorithm2\{Q_s\}$, corresponding predicted labels

$y_{N_c+1}, \dots, y_{N_c+N_t}$,

$iter = iter + 1$

Until stopping criterion satisfied

Output: the frequency band F_s , the subset of channels Q_s and the corresponding labels $y_{N_c+1}, \dots, y_{N_c+N_t}$.

Remarks 3 (1) In Algorithm 3, channel selection and frequency band selection are alternatively carried out. We also can define a Fisher ratio for a subset of channels and a frequency band, which is used as an objective function. We then perform joint channel-frequency selection using genetic method to maximize the objective function. However, the computational burden will increase significantly. For reducing the computational complexity, we apply the iterative strategy to alternatively select the channels and frequency band. (2) To avoid selecting the termination parameters δ_0 and δ_2 in Algorithms 1 and 2, respectively, we fix the number of iterations of these two algorithms to be 5 in this paper. Data analysis results show the effectiveness of this setting (see Example 2 in Section “Data analysis”). (3) Furthermore, we can extend Algorithm 3 for joint channel-frequency-time interval selection by selecting these three parameters alternatively.

Note that if the subset of selected channels do not change between the k th and $(k-1)$ th iteration, the selected frequency band will not change. According to our data analysis, Algorithm 3 often converges in several iterations (e.g. 6 iterations) even if we set $\delta_3 = 0$.

Data analysis

In this section, we apply Algorithm 1 and 3 to data set IVa provided by the BCI competition III (Dornhege et al. 2004) to demonstrate the effectiveness of our algorithms. Two examples are presented, in which the first example illustrates channel selection by Algorithm 1 and the second example illustrates joint selection of channels and frequency band by Algorithm 3 for small training data case. All SVM classifications in the iterations of Algorithm 1 and Algorithm 3 are performed by the LibSVM toolbox (Chang and Lin 2001).

Data description: The aim of data set IVa in BCI competition III was to discriminate different motor imagery tasks with small training data set. This data set contains EEG signals from five subjects (aa, al, av, aw, ay). During every trial of the experiment, the subjects were prompted by visual cues to imagine for 3.5 s one of the following three motor imageries: (L) left hand, (R) right hand and (F) foot. 280 trials of EEG data (118 channels, down-sampled to 100 Hz) of two tasks i.e. imageries of right hand and foot were provided for competition. The details of this data set can be referred to the website: <http://www.ida.fraunhofer.de/projects/bci/competition-ii>.

Data partition: In the following two examples, the data set for each subject is partitioned into two subsets: the first one containing the first 200 trials is used for eightfold cross-validation and the other one containing the last 80

trials is used as an independent test set which is not involved in the training process. In each fold of cross-validation, 25 trials of EEG signals are used as the training data set, while the other 175 trials of EEG signals are used as the test data set.

Example 1 In this example, we apply Algorithm 1 to the above data set to demonstrate its effectiveness in channel selection for small training data set. We also compare Algorithm 1 with an existing method (Fisher ratio based cross-validation) mentioned in Lal et al. (2004).

Data preprocessing steps include spatial filtering (Common Average Reference, CAR), band-pass filtering in the range of 8–14 Hz corresponding to mu rhythm. During motor imagery tasks, event-related desynchronization (ERD) occurs in mu and beta frequency band which can be used as the discriminative frequency bands (Wolpaw et al. 2002). Since ERD generally happens in sensorimotor cortex, we only use 60 EEG channels located in the central area of scalp.

Now we apply Algorithm 1 for eightfold cross-validation. In each fold of cross-validation, a subject-specific subset of channels is selected. At the same time, we obtain the predicted labels for the test dataset. Table 1 shows the average number of selected channels for all subjects (averaged across eightfolds).

In each iteration of Algorithm 1, we perform channel selection based on Fisher ratios and Rayleigh coefficients. We also perform prediction of labels for the test data set and the independent test set. Figure 2 displays the topographies of Fisher ratios for all channels obtained in the first six iterations of the first fold of cross-validation for Subject al. We can find out from this figure that, in the last three iterations, large Fisher ratio appear more concentrative in sensorimotor cortex than that in the previous iterations. Furthermore, the topographies of Fisher ratios do not vary in the last three iterations. This implies that the selected subset of channels in the last three iterations keeps unchanged and the convergence of Algorithm 1 is demonstrated. For the other folds of cross-validation of this subject and for the other subjects, similar phenomena can be observed.

In each fold of cross-validation of one subject, we obtain two iterative curves of prediction accuracies for the test data set and the independent test set, respectively. We average these prediction accuracies across eightfolds. Five subplots of Fig. 3 show the iteration curves of the averaged classification accuracies obtained by Algorithm 1 (solid lines) for the five subjects, respectively. In each subplot, the curve depicted by solid line with “*” corresponds to the independent test set, while the curve depicted by solid line with “o” corresponds to the test data sets which are involved in semi-supervised learning. From these subplots, we can see: (1) satisfactory prediction accuracies can be

Table 1 The average numbers of selected channels for all subjects over the eightfold cross-validation

Subject	aa	al	av	aw	ay
Number of channels	17	16	21	25	15

Fig. 2 Topographies of Fisher ratios for all channels obtained in the first fold of cross-validation for subject al

obtained by Algorithm 1; (2) most of these curves of prediction accuracies obtained by Algorithm 1 (solid lines) show an increasing tendency; (3) Algorithm 1 converges generally in about 10 iterations. Thus the effectiveness of Algorithm 1 is demonstrated here.

For comparison, firstly, we now modify Algorithm 1 as follows. (1) Channel selection: Fisher ratio is used to rank the channels. The number of selected channels is determined by a leave one (trial) out cross-validation approach on the small training data set, where the classifier is a standard SVM. (2) Once the subset of channels is determined, an iterative semi-supervised SVM algorithm is used for joint feature extraction and classification. This algorithm is obtained by removing all steps related to channel selection in Algorithm 1. Thus it can be seen as a simplified version of Algorithm 1. Now we apply this approach to the data set and perform similar data analysis as above. In each fold of the eightfold cross-validation of one subject, we also obtain two average iteration curves of prediction accuracies for the test data set and the independent test set, respectively. Five subplots of Fig. 3 show the iterative curves of the averaged classification accuracies (dashed lines) for five subjects, respectively. In each subplot, the curve depicted by dashed line with “*” corresponds to the independent test set, while the curve depicted by dashed line with “o” corresponds to test data sets which are involved in semi-supervised learning. Secondly, we further apply a standard semi-supervised SVM (SSSVM) algorithm without channel selection to the data set and perform

similar data analysis as above, where the initial frequency band and initial channels in Algorithm 1 are used. The obtained classification results are shown in Fig. 3.

From the above data analysis, we have the following comparison results: (1) In fact, the results of the first iteration of Algorithm 1 in Fig. 3 can be seen as ones obtained by a supervised SVM (without using the test dataset). The increasing tendency of the curves in Fig. 3 show that the performance of our algorithm is better than that of the standard SVM. (2) Through Fig. 3, we can see that the performance of Algorithm 1 for channel selection is better than that of the modified algorithm and the standard semi-supervised SVM algorithm. The main reason is that channel selection of the modified algorithm is based on the small training data set, while the channel selection of Algorithm 1 is performed by a semi-supervised approach in which both the small training data set and the test data set are involved.

Example 2 In this example, we evaluate Algorithm 3 for joint channel-frequency selection using the same data set as above. In the first iteration of Algorithm 3, we set the initial frequency band F_0 to be 8–14 Hz as in Example 1. Using the frequency band 8–21 Hz, we generate the sub-bands as follows: $F_{\text{sub}} = [f, f + \Delta f]$, where $f \in \{8, 9, \dots, 16\}$, $\Delta f \in \{2, 3, 4, 5\}$. Thus totally 36 sub-bands denoted as $F_1, \dots, F_{36}$ are obtained for frequency band selection.

Using the first 200 trials of EEG data of each subject, Algorithm 3 is applied to perform eightfold cross-validation. At the same time, the last 80 trials of EEG data are used as an independent test set. In each fold of cross-validation, we obtain two iterative curves of prediction accuracy for the test set and the independent test set, respectively. For each subject, we average the prediction accuracies for these test sets across eightfolds and obtain an average iteration curve of prediction accuracy. Similarly, we obtain an average iteration curve of prediction accuracy for independent test sets. The two average iteration curves of prediction accuracy are shown in each subplot of Fig. 4, in which five subplots correspond to five subjects, respectively.

According to Fig. 4, satisfactory accuracy rates have been achieved for all subjects. Meanwhile in most of the cases, we see an increasing tendency in the curves of prediction accuracy as the numbers of iterations increase. Furthermore, these curves show that Algorithm 3 converges fast (generally in about 6 iterations).

After Algorithm 3 converges, the selected subset of channels and frequency band does not vary in the following iterations. The effectiveness of Algorithm 3 for joint channel-frequency selection is demonstrated in this example.

From our data analysis, we found that the selected subset of channels and a frequency band depend on the

Fig. 3 Curves of averaged prediction accuracies for five subjects. The *solid lines* are the results of Algorithm 1, while the *dashed lines* are for the modified algorithm in which the channel selection is performed by a leave-one-out cross-validation method. And the *dot-*

dashed lines are the results of standard semi-supervised SVM (SSSVM) algorithm without channel selection. The curves with “o” correspond to test data sets used in eightfold cross-validation, while the curves with “*” correspond to independent test sets

training dataset and subject. However, most of selected channels and frequency bands are consistent for different folds. For example, 75% of selected channels on average are consistent for two different folds and the frequency band 12–13 Hz is always included in the selected frequency bands in all eightfolds for subject *al*.

For comparison, firstly, we plot the average iteration curves of prediction accuracy of Algorithm 1 with dashed lines and standard semi-supervised SVM algorithm with dot-dashed lines in Fig. 4. The accuracy of the first iteration of these curves is obtained with standard supervised algorithm without using test dataset. Through Fig. 4, we can see that most of the curves show an increasing tendency. Thus the performance of our algorithm is better than that of the standard supervised SVM. Moreover, we can see that the performance of Algorithm 3 is better than that

Algorithm 1 and standard semi-supervised SVM algorithm without channel selection. The main reason is that both channel selection and frequency band selection are performed in Algorithm 3, and a semi-supervised approach based on both training data and test data is used to select these two parameters.

Secondly, we also use a genetic algorithm to jointly select channels and frequency band under the same data partition as above. We perform channel and frequency band searching over all the combinations of the sixty channels and 36 frequency bands as described above. Hence, there are totally 2,160 (60×36) choices. The average accuracy of fivefold cross-validation was used as the individual's fitness measure for optimization. More details about this algorithm can be found in Garrett et al. (2003). Using the data of Subject *al* as an example,

Fig. 4 Iteration curves of average prediction accuracy rates (%) obtained by Algorithm 3 with solid lines, while the dashed lines are the results of Algorithm 1 and the dot-dashed lines are the results of

SSSVM. The five subplots correspond to the five subjects, respectively. In each subplot, the curve with “o” is for test data sets, while the curve with “*” is for the independent test set

we performed channel and frequency band selection. Then the labels were predicted for the test dataset and the independent test dataset through a semi-supervised learning procedure. The final accuracy rates are 60.2 and 65.3% for the test data set and the independent test data set, respectively, which are much lower than the results of our proposed Algorithm 3 (93.2 and 98.4%) presented in the second subplot in Fig. 4. Such results unleash that parameter selection based on small training dataset is unreliable even if genetic algorithm is adopted. Strategically involving unlabeled test dataset may improve the effectiveness of parameter setting, which ultimately results in significantly better accuracy. On the other hand, regarding computational efficiency, our semi-supervised approach for parameter setting is also much superior to the genetic algorithm.

Note that data set IVa in BCI competition III includes five data sets corresponding to five subjects. Each data set including a training data set and a test data set contains 280

trials totally. The numbers of trials in training data sets are 168, 224, 84, 56 and 28 for subjects *aa*, *al*, *av*, *aw* and *ay*, respectively. In this paper, we use the same data partition for all subjects. The number of trials of the training data is always set to 25 for all subjects, which is much smaller than that in BCI competition III except subject *ay*. Furthermore, an independent data set containing 80 trials of EEG data is used here for each subject. This independent data set is not involved in semi-supervised learning. Therefore, our data partition is much different from that in BCI competition III and the accuracy rates obtained in this paper are not comparable to those obtained by the winners of this data set (Wang et al.). Additionally, Y. Li (one of authors of this paper) and his colleagues attended BCI competition III. They analyzed this data set using an extended EM algorithm and obtained the second. At that time, they selected channels and frequency band manually (Li and Guan 2006).

Conclusions

In this paper, we first presented a semi-supervised SVM algorithm for channel selection. This algorithm is designed for motor-imagery based BCI. Next, a modified version of this algorithm is obtained for frequency band selection. Combining these two algorithms, an iterative algorithm is obtained for joint selection of channels and frequency band. Based on semi-supervised learning strategy, these algorithms are tailored for the case of small training data set. The effectiveness of our algorithms were demonstrated through analyzing the data set IVa in BCI competition III.

Acknowledgments The authors would like to thank the Berlin BCI group for providing the data set IVa in BCI competition III.

References

- Al-Ani A, Al-Sukker A (2006) Effect of feature and channel selection on EEG classification. In: Proceedings of international conference on IEEE EMBS, pp 2171–2174
- Bishop C (1995) Neural networks for pattern recognition. Plus, Oxford University Press, USA
- Blanchard G, Blankertz B (2004) BCI competition 2003-data set IIa: spatial patterns of self-controlled brain rhythm modulations. *IEEE Trans Biomed Eng* 51(6):1062–1066
- Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller K (2008) Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process Magazine* 25(1):41
- Burges C (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* 2(2):121–167
- Chang C, Lin C (2001) LIBSVM: a library for support vector machines [online]. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Dornhege G (2007) Toward brain-computer interfacing. MIT Press
- Dornhege G, Blankertz B, Curio G, Muller K (2004) Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Trans Biomed Eng* 51(6):993–1002
- Dornhege G, Blankertz B, Krauledat M, Losch F, Curio G, Muller K (2006) Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Trans Biomed Eng* 53(11):2274–2281
- Freeman W (2007) Definitions of state variables and state space for brain-computer interface. *Cogn Neurodyn* 1(1):3–14
- Garrett D, Peterson D, Anderson C, Thaut M (2003) Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Trans Neural Syst Rehabil Eng* 11(2):141–144
- Kaper M, Meinicke P, Grossekhoefer U, Lingner T, Ritter H (2004) BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm. *IEEE Trans Biomed Eng* 51(6):1073–1076
- Lal T, Schroder M, Hinterberger T, Weston J, Bogdan M, Birbaumer N, Scholkopf B (2004) Support vector channel selection in BCI. *IEEE Trans Biomed Eng* 51(6):1003–1010
- Lemm S, Blankertz B, Curio G, Muller K (2005) Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans Biomed Eng* 52(9):1541–1548
- Li Y, Guan C (2006) An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces. *Neural Comput* 18(11):2730–2761
- Li Y, Guan C (2008) Joint feature re-extraction and classification using an iterative semi-supervised support vector machine algorithm. *Mach Learn* 71(1):33–53
- Meinicke P, Kaper M, Hoppe F, Heumann M, Ritter H (2003) Improving transfer rates in brain computer interfacing: a case study. In: Proceedings of advanced neural information processing system, pp 1131–1138
- Muller K, Krauledat M, Dornhege G, Curio G, Blankertz B (2004) Machine learning techniques for brain-computer interfaces. *Biomed Tech* 49(1):11–22
- Pfurtscheller G, Brunner C, Schlögl A, Lopes da Silva F (2006) Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks. *Neuroimage* 31(1):153–159
- Pregenzer M, Pfurtscheller G (1999) Frequency component selection for an EEG-based brain to computer interface. *IEEE Trans Neural Syst Rehabil Eng* 7(4):413–419
- Ramoser H, Muller-Gerking J, Pfurtscheller G (2000) Optimal spatial filtering of single trial EEG during imagined handmovement. *IEEE Trans Rehabil Eng* 8(4):441–446
- Schroder M, Bogdan M, Hinterberger T, Birbaumer N (2003) Automated EEG feature selection for brain computer interfaces. In: Proceedings of 1st International IEEE EMBS Conference on Neural Engineering, pp 626–629
- Shenoy P, Krauledat M, Blankertz B, Rao R, Müller K (2006) Towards adaptive classification for BCI. *J Neural Eng* 3(R13–R23):1–1
- Thulasidas M, Guan C, Wu J (2006) Robust classification of EEG signal for brain-computer interface. *IEEE Trans Neural Syst Rehabil Eng* 14(1):24–29
- Vapnik V (1998) Statistical learning theory. New York, Wiley
- Vidaurre C, Schlögl A, Cabeza R, Scherer R, Pfurtscheller G (2007) Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Trans Biomed Eng* 54(3):550–556
- Wang Y, Gao S, Gao X (2005) Common spatial pattern method for channel selection in motor imagery based brain-computer interface. In: Proceedings of international conference on IEEE EMBS, pp 5392–5395
- Wang Y, Gao X, Zhang Z, Hong B, Gao S. BCI competition III data set IVa: classifying single-trial EEG during motor imagery with a small training set. *IEEE Trans Neural Syst Rehabil Eng*, to be published
- Wolpaw J, McFarland D, Vaughan T (2000) Brain-computer interface research at the Wadsworth Center. *IEEE Trans Rehabil Eng* 8(2):222–226
- Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G, Vaughan T (2002) Brain-computer interfaces for communication and control. *Clin Neurophysiol* 113(6):767–791
- Wu W, Gao X, Hong B, Gao S (2008) Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL). *IEEE Trans Biomed Eng* 55(6):1733–1743