Expert Systems With Applications 96 (2018) 302–310

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces

Yu Zhang[a,*], Yu Wang[b], Guoxu Zhou[c], Jing Jin[a], Bei Wang[a], Xingyu Wang[a], Andrzej Cichocki[d,e]

[a] *Key Laboratory for Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai, China*
[b] *Shanghai Ruanzhong Information Technology Co., Ltd., Shanghai, China*
[c] *School of Automation at Guangdong University of Technology, Guangzhou, China*
[d] *Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan*
[e] *Skolkovo Institute of Science and Technology (SKOLTECH), 143026 Moscow, Russia*

A R T I C L E I N F O

Article history:
Received 9 March 2017
Revised 6 December 2017
Accepted 7 December 2017
Available online 8 December 2017

Keywords:
Brain-computer interface
Electroencephalogram
Extreme learning machine
Multi-kernel learning
Motor imagery

A B S T R A C T

One of the most important issues for the development of a motor-imagery based brain-computer interface (BCI) is how to design a powerful classifier with strong generalization capability. Extreme learning machine (ELM) has recently proven to be comparable or more efficient than support vector machine for many pattern recognition problems. In this study, we propose a multi-kernel ELM (MKELM)-based method for motor imagery electroencephalogram (EEG) classification. The kernel extension of ELM provides an elegant way to circumvent calculation of the hidden layer outputs and inherently encode it in a kernel matrix. We investigate effects of two different kernel functions (i.e., Gaussian kernel and polynomial kernel) on the performance of kernel ELM. The MKELM method is subsequently developed by integrating these two types of kernels with a multi-kernel learning strategy, which can effectively explore the supplementary information from multiple nonlinear feature spaces for more robust classification of EEG. An extensive experimental comparison with two public EEG datasets indicates that the MKELM method gives higher classification accuracy than those of the other competing algorithms. The experimental results confirm that superiority of the proposed MKELM-based method for accurate classification of EEG associated with motor imagery in BCI applications. Our method also provides a promising and generalized solution to investigate the complex and nonlinear information for various applications in the fields of expert and intelligent systems.

1. Introduction

A brain-computer interface (BCI) is an advanced technique to establish a direct communication between a human brain and a computer (Jin et al., 2015; Li et al., 2016; Wolpaw, Birbaumer, McFarland, Pfutscheller, & Vaughan, 2002). By recognizing a task-related electroencephalogram (EEG) pattern, BCI translates the mental state of human into computer command and provides a promising approach to recover environmental control capabilities of disabled people (Wang et al., 2016). Currently, the mostly adopted EEG patterns for BCI development include sensorimotor rhythms (SMRs), event-related potentials and steady-state visual evoked potentials (Chen et al., 2015; Jiao et al., 2017; Jin, Daly, Zhang, Wang, & Cichocki, 2014; Ma, Zhang, Cichocki, & Mastuno, 2015; Pfurtscheller, Brunner, Schlögl, & Lopes, 2006; Shi, Wang, & Zhang, 2015; Yu, Li, Long, & Gu, 2012; Zhang, Zhou, Jin, Wang, & Cichocki, 2014). Event-related desynchronization (ERD) is a significant power decrease of SMRs occurring at the contralateral sensorimotor area during the imagination of unilateral hand movements. Accordingly, motor-imagery (MI) based BCI is designed to detect the desired commands by classifying MI tasks according to ERD features (Yu et al., 2015).

A direct approach for recognizing ERD features is to measure the variance difference between the left and right hemispheres (i.e., electrodes C3 and C4). However, the simple method is likely to give a poor classification accuracy if signals of the two electrodes are contaminated by noises. Until now, extensive research efforts have been dedicated to improving EEG feature extraction and classification for BCI applications (Cong et al., 2015; Das, Suresh, & Sundararajan, 2016; García-Laencina, Rodríguez-Bermudez, & Roca-

* Corresponding author.
E-mail address: yuzhang@ecust.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.eswa.2017.12.015

Dorda, 2014; Liu, Yu, Wu, Gu, & Li, 2015; Nguyen, Khosravi, Creighton, & Nahavandi, 2015; da Silveira, Kozakevicius, & Rodrigues, 2016; Wu, Chen, Gao, & Brown, 2011; Zhang, Zhou et al., 2016; Zhou et al., 2016). Common spatial pattern (CSP) seeks spatial filters for multichannel optimization of EEG to maximize the variance of projected signal from one class while to minimize it from another class. In recent years, CSP and its variants have been most popularly applied to robust feature extraction for improving MI-related EEG classification (Arvaneh, Guan, Ang, & Quek, 2013; Blankertz, Tomioka, Lemm, Kawanabe, & Müller, 2008; Lotte & Guan, 2011; Wu et al., 2015; 2014).

Another important issue for EEG classification of MI tasks is how to design an powerful classifier with a strong generalization capability (Lotte, Congedo, Lecuyer, Lamarche, & Arnaldi, 2007; Zhang et al., 2014; Zhang, Wang, Jin, & Wang, 2017). Linear discriminant analysis (LDA) is a relatively simple algorithm and generally works well for pattern classification if the sample covariance matrices are similar among different classes (Krusienski et al., 2006). However, this assumption might not be met for ERD features so that a good classification accuracy could hardly be achieved by LDA due to the possible overfitting problem. Some regularization-based classification algorithms have been recently developed and increasingly applied to EEG analysis. For instance, shrinkage LDA remedies the ill-conditioned covariance matrices with shrinkage covariance estimator and significantly enhances classification accuracy, especially in small sample size scenarios (Blankertz, Lemm, Treder, Haufe, & Müller, 2011). The well-known support vector machine (SVM) adopts a soft margin regularization to achieve good generalization capability (Lotte et al., 2007). The kernel extension of SVM usually provides good effects on the classification of nonlinear features in EEG signals. Currently, combination of CSP and SVM has become one of the most popular methods for MI classification (Qiu et al., 2017; Zhang, Zhou, Jin, Wang, & Cichocki, 2015). On the other hand, a sparse representation-based scheme was also proposed for MI classification by l_1-norm regularization and showed better performance than LDA (Li et al., 2014; Shin, Lee, Lee, & Lee, 2012).

In recent years, extreme learning machine (ELM) has attracted increasing attention from researchers in the pattern recognition field (Avci, 2013; Avci & Coteli, 2012; Huang, Huang, Song, & You, 2015). ELM was originally proposed by Huang, Zhu, and Siew (2006) for training single hidden layer feedforward neural networks, such as multilayer perceptron (MLP). The hidden layer weights in ELM are randomly initialized and fixed without iteratively tuning. The output weights are optimized by solving the Moore–Penrose generalized inverse of hidden matrix so that ELM achieves not only the smallest training error but also the smallest norm of output weights. ELM is basically formed by two processing steps: (1) random mapping of input space to ELM feature space and (2) learning of an appropriate linear projection for classification. Some empirical studies have suggested that ELM provides comparable or even better generalization capability than that of SVM and its variants (Chen & Ou, 2011; Huang, 2014). Furthermore, a probabilistic version of ELM has been developed to estimate the probability distribution of output values instead of fitting data, thereby alleviating the data overfitting problem. Our previous study (Zhang, Jin, Wang, & Wang, 2016) validated the effectiveness of Bayesian ELM for EEG classification. However, the randomly assigned node parameters generally result in a large variation in classification accuracy for different runnings with the same number of hidden nodes (Pal, Maxwell, & Warner, 2013). Also, the optimal dimensionality of ELM feature space varies for different applications and is usually determined by experience or a time-consuming procedure (Iosifidis, Tefas, & Pitas, 2015). To overcome the problems, kernel extension of ELM has been increasingly studied to circumvent calculation of the hidden layer outputs and inherently encode it in a kernel matrix (Huang, Zhou, Ding, & Zhang, 2012). Instead of using a single kernel, ELM with multi-kernel learning has recently arisen and is able to achieve improved classification performance by combining different kernels (Liu, Wang, Huang, Zhang, & Yin, 2015).

Inspired by these studies, we propose a multi-kernel ELM (MKELM)-based method for accurate classification of EEG associated to MI tasks in BCI applications. Two different types of kernels, i.e., Gaussian kernel and polynomial kernel are exploited to map the original CSP features to different nonlinear feature spaces. The two nonlinear feature spaces provide richer discriminant information that may be supplementary to each other. Accordingly, we integrated them using a multi-kernel learning strategy to achieve more robust classification of EEG in MI tasks. With two public EEG datasets, an extensive experimental comparison is carried out among the proposed MKELM-based method and several other competing approaches. The experimental results demonstrate that the MKELM method is a promising candidate for the development of an improved MI-based BCI.

The rest of the paper is structured as follows. In Section 2, feature extraction and classification procedures are described. Some basic concepts of ELM are briefly reviewed. Multi-kernel ELM is introduced. Extensive experimental results are given in Section 3. A discussion is given in Section 4. Finally, in Section 5, some conclusions are given.

2. Methods

2.1. Feature extraction

Common spatial pattern (CSP) has proven to be an effective method for feature extraction in classifying two classes of motor imagery EEG data. Let $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2} \in \mathbb{R}^{C\times P}$ denote EEG samples of two classes recorded from the i-th trial with C and P being the number of channels and samples, respectively. Assume both the EEG samples have been bandpass filtered at a specified frequency band and mean-removed. CSP aims at finding spatial filter $\mathbf{w} \in \mathbb{R}^C$ to transform the EEG data so that the ratio of data variance between the two classes is maximized

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \boldsymbol{\Sigma}_1 \mathbf{w}}{\mathbf{w}^T \boldsymbol{\Sigma}_2 \mathbf{w}} \quad s.t. \quad \|\mathbf{w}\|_2 = 1, \tag{1}$$

where $\boldsymbol{\Sigma}_l = \sum_{i=1}^{N_l} \mathbf{X}_{il}\mathbf{X}_{il}^T/N_l$ with N_l being the number of trials belonging to class l $(l = 1, 2)$. The optimum solution of (1) can be obtained by solving a generalized eigenvalue problem (Blankertz et al., 2008).

A matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{2M}] \in \mathbb{R}^{C\times 2M}$ including the spatial filters is formed by the eigenvectors corresponding to the M largest and smallest eigenvalues. For a given EEG sample $\mathbf{X}$, the feature vector is constructed as $\mathbf{x} = [x_1, x_2, \ldots, x_{2M}]$ with entries

$$x_m = \log\left(\mathrm{var}(\mathbf{w}_m^T \mathbf{X})\right), \quad m = 1, 2, \ldots, 2M, \tag{2}$$

where var(·) denotes the variance operation.

2.2. SVM for EEG classification

In recent years, support vector machine (SVM) has been most frequently adopted with CSP features for MI classification. SVM is to learn a hyperplane maximizing the separating margin between two classes (Chen, Yang, & Chen, 2014; Chen, Yang, Ye, & Liang, 2011). Consider a set of training samples $\mathbf{x}_i \in \mathbb{R}^D$ and the corresponding class labels $y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, N$. The primal SVM optimization problem is formulated as

$$\min_{\mathbf{u},b} \ \frac{1}{2}\|\mathbf{u}\|_2^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } y_i(\mathbf{u}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{3}$$

where $\mathbf{u} \in \mathbb{R}^D$ is the learned discriminant vector, C is a regularization parameter for soft margin, and ξ_i denotes the slack variable. The primal optimization (3) can be equivalent to the following problem by Lagrangian multipliers

$$\max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\}$$
$$\text{s.t. } \alpha_i \geq 0, \quad i = 1, \ldots, N$$
$$\sum_{i=1}^{N} \alpha_i y_i = 0, \tag{4}$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ denotes a kernel function. As a result, a kernel SVM can be derived to exploit the potential nonlinear features for EEG classification. In recent years, SVM classification preceded by CSP feature extraction has been a state-of-the-art method for MI-related EEG classification (Arvaneh, Guan, Ang, & Quek, 2011; Zhang et al., 2015).

2.3. ELM for EEG classification

ELM was proposed as an extension of the single layer feed-forward network (Huang, Zhu, & Siew, 2006). In mathematics, extreme learning machine (ELM) with L hidden nodes and a feature mapping $h(\mathbf{x})$ is modeled as

$$\sum_{j=1}^{L} \boldsymbol{\beta}_j h(\mathbf{x}_i) = y_i, \quad i = 1, \ldots, N. \tag{5}$$

where $\boldsymbol{\beta}_j$ is a weight vector connecting the jth hidden neuron and the output neurons. Different functions may be used for the ELM feature mapping in different applications. A generally effective feature mapping is the sigmoid function

$$h(\mathbf{x}_i) = \frac{1}{1 + \exp(-(\mathbf{a}_j^T \mathbf{x}_i + b_j))}, \tag{6}$$

where $\mathbf{a}_j = [a_{j1}, a_{j2}, \ldots, a_{jD}]^T \in \mathbb{R}^D$ is the weight vector connecting the jth hidden neuron and the input neurons, b_j is the bias of the jth hidden node.

The N equations in (5) can be written in a compact form

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{y}, \tag{7}$$

with $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_L]^T \in \mathbb{R}^L$, $\mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$ and

$$\mathbf{H}(\mathbf{a}_1, \ldots, \mathbf{a}_L, b_1, \ldots, b_L, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \begin{bmatrix} h_1(\mathbf{x}_1) & \cdots & h_L(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ h_1(\mathbf{x}_N) & \cdots & h_L(\mathbf{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times L}, \tag{8}$$

where $\mathbf{H}$ is the so called hidden layer output matrix (Huang, 2003). The solution of (7) is formally equivalent to the following optimization problem

$$\min_{\mathbf{a}_i, b_i, \boldsymbol{\beta}} \|\mathbf{H}(\mathbf{a}_1, \ldots, \mathbf{a}_K, b_1, \ldots, b_K)\boldsymbol{\beta} - \mathbf{y}\|_2^2. \tag{9}$$

Since no gain is possible for the solution of (9) by adjusting the input weights and hidden layer biases, ELM randomly assigns and fixes the input weights $\mathbf{a}_i$ and hidden layer biases b_i. The optimization problem (9) is then transformed into

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{H}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \frac{1}{2}\|\boldsymbol{\beta}\|_2^2. \tag{10}$$

It can be seen that ELM theory aims to reach not only the smallest training error but also the smallest norm of output weights. We can derive the least-squares solution with minimum norm by

$$\tilde{\boldsymbol{\beta}} = \mathbf{H}^\dagger \mathbf{y}, \tag{11}$$

where † denotes the Moore–Penrose generalized inverse. For a new test sample $\hat{\mathbf{x}}$, the decision function of ELM is then given by

$$y = \text{sign}(h(\hat{\mathbf{x}})\tilde{\boldsymbol{\beta}}). \tag{12}$$

Fig. 1 illustrates the ELM calibration stage for EEG classification. ELM algorithm for EEG classification is summarized in Algorithm 1.

Algorithm 1: ELM algorithm.

Input: A set of EEG training samples $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, N$, a test sample $\hat{\mathbf{x}}$, specified feature mapping $h(\mathbf{x})$ and the number of hidden nodes L.
Output: Predicted label y.
1. Randomly specify weight vector $\mathbf{a}_l$ and bias b_l for $\forall l \in \{1, \ldots, L\}$
2. Compute the hidden layer output matrix $\mathbf{H}$
3. Compute the output weight vector $\tilde{\boldsymbol{\beta}}$
4. Compute the predicted label by (12).

2.4. Multi-kernel ELM for EEG classification

In a more general case, the feature mapping $g(\mathbf{x})$ is unknown or not specified before. The hidden layer output matrix $\mathbf{H}$ cannot be calculated from (8). We derive a kernel version of ELM by defining a kernel matrix as follows

$$\mathbf{K} = \mathbf{H}\mathbf{H}^T \tag{13}$$

with

$$(\mathbf{K})_{ij} = h(\mathbf{x}_i)h(\mathbf{x}_j)^T = k(\mathbf{x}_i, \mathbf{x}_j), \tag{14}$$

where $\mathbf{K}$ is called ELM kernel matrix. The output function of kernel ELM for a new test sample $\hat{\mathbf{x}}$ is then given by

$$\begin{aligned} S &= h(\hat{\mathbf{x}})\tilde{\boldsymbol{\beta}} \\ &= h(\hat{\mathbf{x}})\mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{y} \\ &= \begin{bmatrix} k(\hat{\mathbf{x}}, \mathbf{x}_1) \\ \vdots \\ k(\hat{\mathbf{x}}, \mathbf{x}_N) \end{bmatrix}^T \mathbf{K}^\dagger \mathbf{y}. \end{aligned} \tag{15}$$

Note that the kernel matrix $\mathbf{K}$ is likely to be singular since the number of hidden neurons L is usually much less than the number of training samples N. Thus, we used the Moore–Penrose generalized inverse in (15). The decision function of kernel ELM is then given by $y = \text{sign}(S)$. A feature mapping function $h(\mathbf{x})$ is actually no longer needed in kernel ELM. Instead, by using the 'kernel trick', we only need to define an appropriate kernel function for classification. Fig. 2 depicts the feature mapping to reproducing kernel Hilbert space via a specified kernel function.

Instead of using a single kernel mapping, we define multiple kernels and combine them under a multi-kernel learning strategy to obtain more robust classification performance. Assume $k_1(\mathbf{x}_i, \mathbf{x}_j), k_2(\mathbf{x}_i, \mathbf{x}_j), \ldots, k_Q(\mathbf{x}_i, \mathbf{x}_j)$ denote Q pre-defined kernels. These kernels are able to map the original CSP features to different nonlinear feature spaces that may provide supplementary discriminant information to each other. MKELM is naturally derived based the single kernel ELM by using a mixed kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{q=1}^{Q} \lambda_q k_q(\mathbf{x}_i, \mathbf{x}_j), \tag{16}$$

where λ_q is a parameter for the balance among different kernels. With MKELM, multiple nonlinear feature spaces are integrated for more robust classification of EEG. Note that, the same multi-kernel

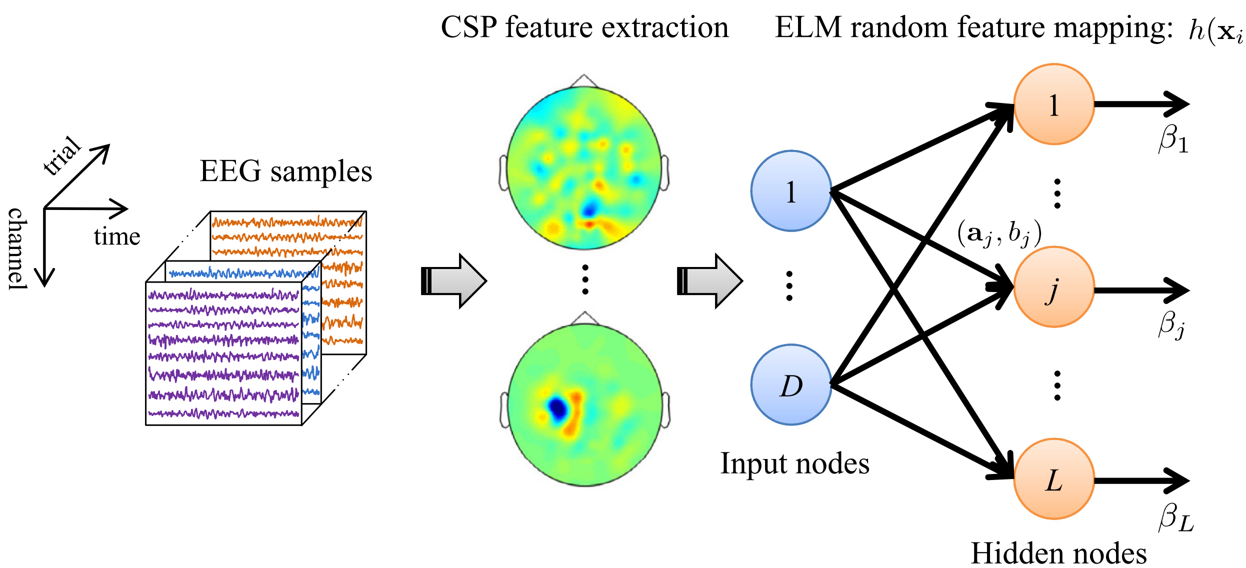

Fig. 1. Illustration of ELM calibration stage for EEG classification.

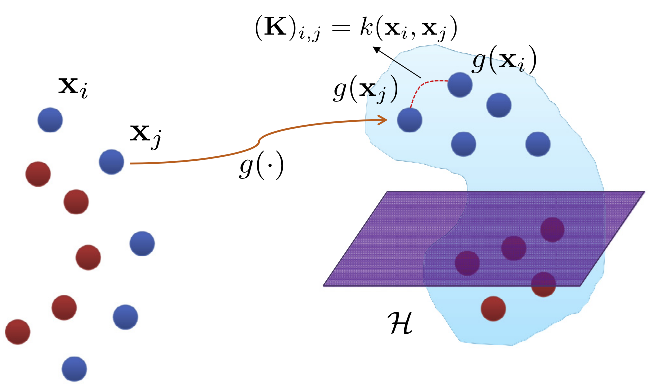

Fig. 2. Feature mapping to reproducing kernel Hilbert $\mathcal{H}$ space by a nonlinear function $h(\cdot)$. $k(\mathbf{x}_i, \mathbf{x}_j)$ denotes the kernel function. (Zhao et al., 2013).

learning strategy can be straightly applied to (4) to obtain a multi-kernel SVM method. Kernel ELM algorithm for EEG classification can be summarized in Algorithm 2.

Algorithm 2: Multi-kernel ELM algorithm.

Input: A set of EEG training samples $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, N$, a test sample $\hat{\mathbf{x}}$, specified Q kernels $k_q(\mathbf{x}_i, \mathbf{x}_j)$, parameters $\lambda_1, \ldots, \lambda_Q$.

Output: Predicted label y.

1. Compute the mixed kernel matrix $\mathbf{K}$ with $(\mathbf{K})_{ij} = \sum_{q=1}^{Q} \lambda_q k_q(\mathbf{x}_i, \mathbf{x}_j)$ for $\forall i, j \in \{1, \ldots, N\}$
2. Compute the mixed kernel $k(\hat{\mathbf{x}}, \mathbf{x}_i) = \sum_{q=1}^{Q} \lambda_q k_q(\hat{\mathbf{x}}, \mathbf{x}_i)$ for $\forall i \in \{1, \ldots, N\}$
3. Compute the output S by (15) and obtain the predicted label by $y = \mathrm{sign}(S)$.

In this study, we investigate effectiveness of MKELM using two popularly used kernels, i.e., Gaussian kernel and polynomial kernel

$$\text{Gaussian kernel: } k_1(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right), \quad (17)$$

$$\text{Polynomial kernel: } k_2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d, \quad (18)$$

where σ and d are kernel width and polynomial degree, respectively.

3. Experimental study

3.1. Data description

Two public EEG datasets were used for the experimental study. The first dataset was available from BCI Competition III dataset IVa. The EEG data were collected at the sampling rate of 100 Hz from 118 electrodes for five subjects (named “aa”, “al”, “av”, “aw”, and “ay”), during the imaginations of right hand or foot movements. Each subject completed 280 trials (half for each class of MI) and each trial lasted for 3.5 s. See http://www.bbci.de/competition/iii/ for more details about the dataset.

The second dataset was available from BCI Competition IV dataset IIb. The EEG data were collected from nine subjects at three electrodes C3, Cz and C4 with sampling rate of 250 Hz, during right hand and left hand MI tasks. This study only uses the third training sessions of the dataset, i.e., “B0103T”, “B0203T”, ..., “B0903T”. Each subject completed 160 trials (half for each class of MI). In each trial, the subject was indicated by a visual cue to perform MI task for 4.5 s. See http://www.bbci.de/competition/iv/ for more details about the dataset.

3.2. Experimental evaluation

For both of the two datasets, EEG data of each trial were band-pass filtered between 4–40 Hz that covered most ERD features related to MI. Features were then extracted by CSP with parameter $M = 1$. In this study, MI classification accuracy is compared among the following eight algorithms:

(1) MLP: multilayer perceptron with a single hidden layer;
(2) SVM: the conventional SVM;
(3) GKSVM: SVM with Gaussian kernel;
(4) PKSVM: SVM with polynomial kernel;
(5) MKSVM: Multi-kernel SVM using both Gaussian and polynomial kernels;
(6) ELM: the conventional ELM;

Table 1
Classification accuracies (%) obtained by MLP, SVM, GKSVM, PKSVM, ELM, GKELM, and PKELM, respectively, from 5 × 5-fold cross validation, for BCI Competition III dataset IVa. For each subject, the highest accuracy is marked in boldface.

Subject	MLP	SVM	GKSVM	PKSVM	MKSVM	ELM	GKELM	PKELM	MKELM
aa	81.2	80.8	82.3	81.2	82.9	82.5	82.6	82.9	**83.3**
al	97.5	97.8	97.5	97.8	98.3	97.9	98.0	98.0	**98.5**
av	69.4	69.3	69.8	69.1	70.6	69.6	69.8	70.1	**71.4**
aw	88.9	89.8	90.2	89.6	91.0	90.5	90.8	90.7	**91.3**
ay	91.6	91.6	92.0	91.8	92.9	92.3	92.5	92.7	**93.1**
Average	85.7 ± 10.8	85.9 ± 11.1	86.4 ± 10.7	85.9 ± 11.1	87.1 ± 10.8	86.6 ± 11.0	86.7 ± 11.0	86.9 ± 10.9	**87.5** ± 10.5

Table 2
Classification accuracy (%) obtained by MLP, SVM, GKSVM, PKSVM, ELM, GKELM, and PKELM, respectively, from 5 × 5-fold cross validation, for BCI Competition IV dataset IIb. For each subject, the highest accuracy is marked in boldface.

Subject	MLP	SVM	GKSVM	PKSVM	MKSVM	ELM	GKELM	PKELM	MKELM
B0103T	76.3	76.5	76.1	76.3	77.0	76.9	76.8	76.5	**77.5**
B0203T	56.6	56.8	57.1	57.8	62.6	59.6	60.6	62.1	**64.4**
B0303T	50.0	51.6	52.0	51.1	53.1	51.4	52.8	53.6	**54.3**
B0403T	98.8	98.8	99.3	98.6	**99.3**	98.8	98.8	99.0	**99.3**
B0503T	83.0	83.3	83.6	83.6	84.0	84.1	84.0	83.8	**84.6**
B0603T	66.4	67.8	67.8	68.1	68.8	68.3	69.0	68.4	**69.5**
B0703T	82.6	83.1	84.3	84.1	85.0	84.1	83.6	85.5	**86.8**
B0803T	87.8	86.9	87.4	87.5	88.3	87.6	89.1	89.5	**89.9**
B0903T	81.8	81.5	82.4	81.7	82.8	82.6	83.1	83.1	**83.7**
Average	75.9 ± 15.5	76.3 ± 15.0	76.7 ± 15.2	76.5 ± 15.0	77.9 ± 14.2	77.0 ± 14.8	77.5 ± 14.4	77.9 ± 14.3	**78.9** ± 14.0

(7) GKELM: ELM with Gaussian kernel;
(8) PKELM: ELM with polynomial kernel;
(9) MKELM: Multi-kernel ELM using both Gaussian and polynomial kernels.

For each of the algorithms, classification accuracy was evaluated by a 5×5 cross validation. Several parameters need to be pre-specified: the number of hidden nodes and the learning rate in MLP, the regularization parameter C for SVM; the number of hidden nodes L for ELM; the kernel width σ for GKSVM, GKELM, MKSVM and MKELM; the polynomial degree d for PKSVM, PKELM, MKSVM and MKELM; the kernel balance parameter λ for Gaussian kernel in MKELM ($1-\lambda$ for polynomial kernel). All of the required parameters were determined according to the cross validation on training set.

3.3. Results

Tables 1 and 2 summarizes the classification accuracies obtained by different algorithms for BCI Competition III dataset IVa and BCI Competition IV dataset IIb, respectively. These results indicate that ELM and its kernel extensions improved the accuracy in comparison with not only MLP but also SVM and its kernel extensions. By exploiting multi-kernel learning, both the MKSVM and MKELM outperformed those methods using a single kernel (either Gaussian kernel or polynomial kernel). Furthermore, MKELM provided the highest classification accuracy among all of the methods for all subjects.

Paired t-test was further implemented on the second dataset to investigate the significant differences among different methods. The result revealed that GKELM and PKELM worked comparably ($p = .18$) and yielded significantly higher classification accuracies than those of SVM and its kernel extensions (GKELM > SVM: $p < .05$, GKELM > GKSVM: $p = .07$, GKELM > PKELM: $p < .05$, PKELM > SVM: $p < .05$, PKELM > GKSVM: $p < .05$, PKELM > GKSVM: $p < .05$). For both SVM and ELM, multi-kernel learning significantly improved the accuracy compared with those using a single kernel (MKSVM > SVM: $p < .05$, MKSVM > GKSVM: $p = .058$, MKSVM > PKSVM: $p < .05$, MKELM > ELM: $p < .01$, MKELM > GKELM: $p < .05$, MKELM > PKELM: $p < .01$). More importantly, MKELM significantly outperformed MKSVM ($p < .01$).

Instead of tuning the hidden layer parameters, ELM adopts randomly generated ones and solve the output weights using regularized least square. Consistent with other existing studies (Huang et al., 2015; Huang, Zhu, & Siew, 2006), our results indicate that ELM achieves better generalization capability than MLP with hidden layer parameters tuned by back propagation. Through inherently encoding the estimation of hidden layer output with the 'kernel trick', the kernel version of ELM provides a promising approach with universal approximation capability to explore the potential nonlinear features, which outperforms SVM and its kernel extensions. Furthermore, by incorporating the multi-kernel learning strategy, the proposed MKELM is able to not only circumvent the determination of hidden layer size but also fuse the supplementary discriminant information from multiple nonlinear feature spaces.

4. Discussion

The pros and cons of different classification algorithms are summarized in Table 3. Similar to ELM, the well-known MLP presents universal approximation capability for continuous functions (Tang, Deng, & Huang, 2016). However, the hidden layer weights of MLP need to be tuned typically by the time-consuming back propagation (a gradient descent-based algorithm). SVM can deal with high dimensional data and generally provide good generalization performance, but its parameter selection is data dependent. ELM is tuning-free for the hidden layer weights and has shown better generalization capability than SVM for many pattern recognition problems (Huang et al., 2015). However, a potential issue for ELM is how to determine the number of hidden nodes for a specific application (Huang et al., 2015). As an example, Fig. 3 depicts effects of varying number of hidden nodes on MI classification accuracy derived by ELM for a typical subject. The classification accuracy does not necessarily increase with increasing the number of hidden nodes. The optimal size of hidden layer is usually subject-specific and determined by experience or a time-consuming procedure (Iosifidis et al., 2015). Also, randomly generated hidden layer weights that result in a large variation in classification accuracy for different runnings (Pal et al., 2013). Incremental ELM (Huang, Chen, & Siew, 2006) and adaptive ELM (Zhang, Lan, Huang, & Xu, 2012) were developed to dynami-

Table 3
Pros and cons of different classification algorithms.

Method	Pros	Cons
MLP	• Universal approximation capability	• Hidden layer parameters need to be tuned • Slow due to the gradient descent • Easy to overfit the limited training data
SVM	• Can deal with high dimensional data • Good generalization performance	• Parameter selection is data dependent
ELM	• Tuning-free for the hidden layer weights • High efficiency and good generalization performance	• Need to specify the size of hidden layer
KELM	• Not need to determine the size of hidden layer • Can explore nonlinear features	• Need to specify a good kernel function
MKELM	• Can combine multiple types of kernels • Explore supplementary nonlinear features • Better generalization performance	• Need to specify a good balance between kernels

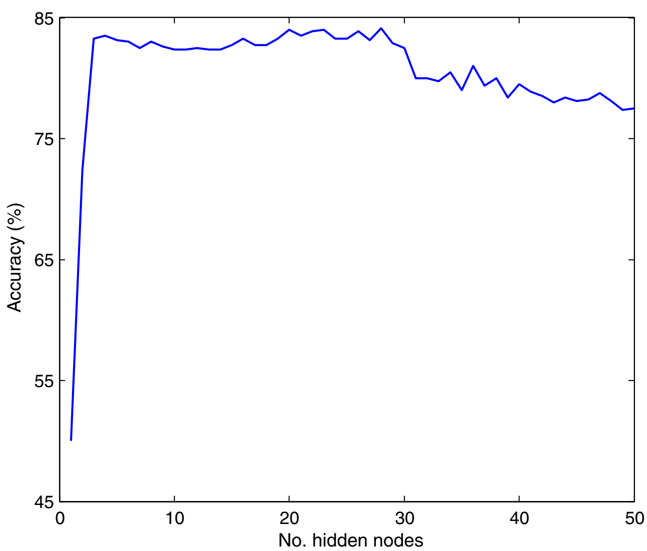

Fig. 3. Effects of varying numbers L of hidden nodes on classification accuracy obtained by ELM for a typical subject.

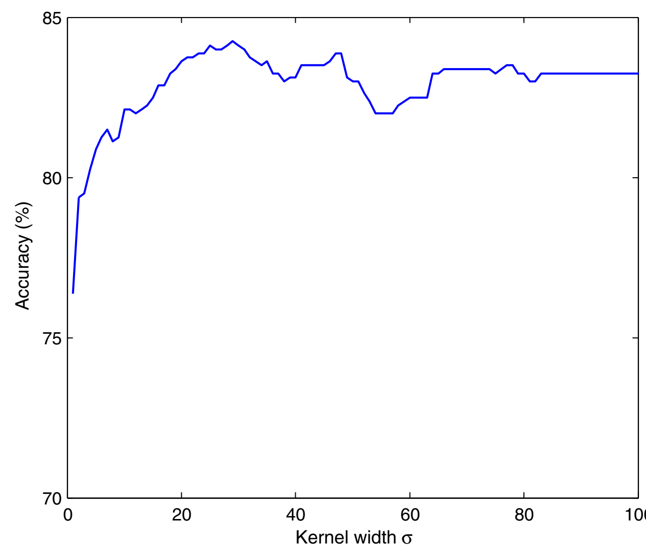

Fig. 4. Effects of varying kernel width σ on classification accuracy obtained by GKELM for a typical subject.

cally adjust the network size of ELM during the training process, while sparse Bayesian learning-based ELM (Luo, Vong, & Wong, 2014; Wong, Vong, Wong, & Luo, 2015) was designed to achieve a compact network by automatically tuning most of the output weights to zeros. Although these algorithms provide effective ways to determine the number of hidden nodes, they increase the model complexity of classical ELM to varying degrees.

In this study, we investigated effects of another alternative, i.e., kernel ELM, for performance improvement in motor imagery EEG classification. Kernel extension of ELM is an elegant approach to circumvent calculation of the hidden layer outputs and inherently encode it in a kernel matrix (Huang et al., 2012). ELM can be implemented by using the 'kernel trick' without the need of specifying the number of hidden nodes. As a result, we only need to select some appropriate kernel parameters (e.g., kernel width for GKELM, constant term and polynomial degree for PKELM) and avoid determining the size of hidden layer. Figs. 4 and 5 show effects of varying kernel parameters on MI classification accuracy obtained by GKELM and PKELM, respectively. On the other hand, kernel ELM also provides a promising way to mining the potential nonlinear features of EEG for improving classification performance. However, a single kernel may not be sufficient to completely reveal the complex nonlinear features of EEG. Instead, multi-kernel learning (Liu, Wang et al., 2015; Wang, Huang, Sun, & Gao, 2015; Zhu, Xie, Zhu, Liu, & Zhang, 2015) has been shown powerful strength for improved classification performance by combining different types of kernels. In our proposed MKELM method, two different types of kernels (e.g., Gaussian kernel and polynomial kernel) are exploited to map the original CSP features to different nonlinear feature spaces that bring richer discriminant information in comparison with using only a single kernel. We further investigate effects of varying the kernel balance parameter λ on classification accuracy obtained by MKELM (see Fig. 6). By choosing an appropriate kernel balance parameter λ, MKELM outperforms both GKELM and PKELM. This confirms that the two types of kernels indeed provide supplementary information to each other. With multi-kernel learning, improved classification accuracy of EEG (see Tables 1 and 2) is finally achieved by integrating the two nonlinear feature spaces.

In our method, a weighting strategy was adopted for the fusion of multiple kernels, which provides a relatively simple approach to integrate the supplementary information from different nonlinear feature spaces. However, a potential issue is how to find an optimal balance between different kernels for a specific application since the kernel balance parameter is typically data dependent. In this study, we selected this parameter using the cross-validation procedure that is typically time-consuming and may not be feasible in a small sample size scenario. An efficient method (Yeh, Su, & Lee, 2013) has recently been developed to derive the weights in multi-kernel learning by transforming the learning into a semidefinite programming problem with an induction setting. Such a more sophisticated strategy could be utilized to enhance the efficiency of the MKELM method, which is worth our further study.

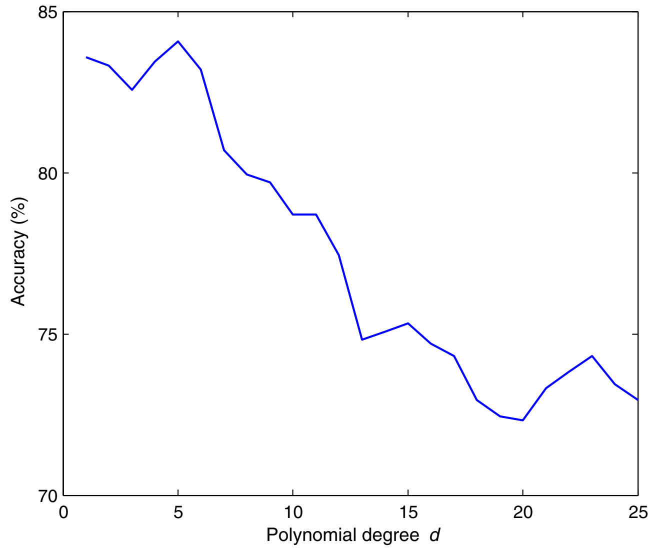

Fig. 5. Effects of varying the polynomial degree d on classification accuracy obtained by PKELM for a typical subject.

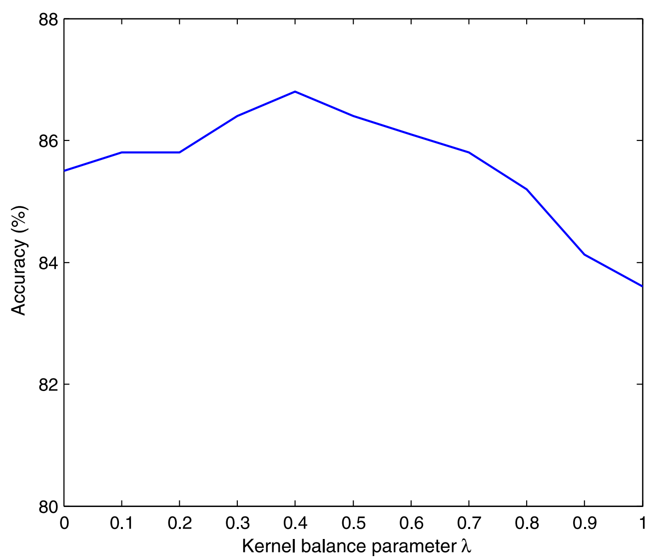

Fig. 6. Effects of varying the kernel balance parameter λ on classification accuracy obtained by MKELM for a typical subject. MKELM will become GKELM at $\lambda = 1$ while PKELM at $\lambda = 0$.

In recent years, deep learning, as a leading trend in the artificial intelligence field, has confirmed that pattern recognition can significantly benefit from the abstract information learned via hierarchical feature representation (LeCun, Bengio, & Hinton, 2015). Although ELM and its variants have proven their strengths on solving various pattern recognition problems, most of them are based on a single hidden-layer model. Due to the shallow architecture, the ELM-based methods can hardly explore the high-level feature representation that might be crucial for analysis of the complex neural data. A hierarchical learning framework has recently arisen for the multilayer extension of ELM (Tang et al., 2016). By exploiting a sparse autoencoder, the hierarchical ELM achieves high-level feature representation that is more meaningful than that of the original ELM. Accordingly, we consider that the extension of our MKELM method to a deeper architecture could be promising for further improving the classification performance.

As most of the matrix factorization algorithms, the proposed MKELM method ignores the multilinear structure of data, which can serve as a priori information for more accurate decomposition of the data with good generalization capability (Chen, Wang, & McKeown, 2016; Wang et al., 2017; Zhang et al., 2017; Zhang, Zhou, Zhao, Cichocki, & Wang, 2016; Zhou, Cichocki, Zhao, & Xie, 2014; Zhu, Suk, Huang, & Shen, 2017). In recent years, kernel method has been introduced into tensor analysis for learning multiway nonlinear features (Zhao, Zhou, Adali, Zhang, & Cichocki, 2013) while a combination of ELM and tensor analysis is beginning to emerge. (Wang, Chen, Yan, Chen, & Fu, 2014). Hence, it is reasonable to expect that a new ELM algorithm by simultaneously exploiting multi-kernel learning and tensor analysis could be feasibly developed to improve the classification performance further. These will be investigated in our future study.

5. Conclusions

In this study, we proposed a multi-kernel ELM-based method for EEG classification in motor-imagery BCIs. Two different types of kernels, i.e., Gaussian kernel and polynomial kernel were exploited to map the original CSP features to different nonlinear feature spaces that provide richer discriminant information. With multi-kernel learning, the two nonlinear feature spaces were integrated to achieve more robust classification of EEG. An extensive experimental comparison was carried out on two public EEG datasets to validate the effectiveness of the proposed MKELM-based method, in comparison with several other competing approaches. The superior performance based on the experimental results demonstrate that the MKELM method is a promising candidate for the development of an improved MI-based BCI.

It should be noted that our proposed method is not limited to BCI but could also be straightly applied to various other applications. Especially, the development of expert systems could significantly benefit from more extensions of MKELM. The following directions may be worthy of our further studies:

(1) Studying on sophisticated strategy for accurate estimation of kernel balance parameters. Developing improved kernel fusion algorithms such as semidefinite programming or ensemble learning is important for further improving the classification performance of MKELM.
(2) Extending the MKELM to a deeper architecture. Follow the deep learning theory, studying the multiple layer encoding of kernel matrix will be an interesting topic. It is also worth investigating the deep feature based multi-kernel learning ELM algorithm for improved classification performance.
(3) Studying on tensorial kernel function and feature fusion algorithm fr tensor kernels. Designing kernel-based multilinear learning algorithm for high-order extension of ELM could be potentially useful for exploring more discriminant features from the multi-dimensional structure of data.
(4) Improving classification accuracy in small sample size, a very common issue not only in BCI but also various pattern recognition applications. Incorporating advanced learning algorithms such as transfer learning or reinforcement learning into ELM to reduce the required number of samples could be considerably important and worthy of further study.

Acknowledgments

This work was supported in part by the grant National Natural Science Foundation of China, under Grant nos. 91420302, 61573142, 61673124. This work was also supported by the Fundamental Research Funds for the Central Universities WH1516018, Shanghai Chenguang Program under Grant 14CG3 and Shanghai Natural Science Foundation under Grant 16ZR1407500, the Programme of Introducing Talents of Discipline to Universities (the 111 Project) under Grant B17017, the MES RF grant 14.756.31.0001, and the PNSC grant 2016/20/W/NZ/00354.

References

Arvaneh, M., Guan, C., Ang, K., & Quek, C. (2011). Optimizing the channel selection and classification accuracy in EEG-based BCI. *IEEE Transactions on Biomedical Engineering, 58*(6), 1865–1873.

Arvaneh, M., Guan, C., Ang, K., & Quek, C. (2013). Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems, 24*(4), 610–619.

Avci, E. (2013). A new method for expert target recognition system: genetic wavelet extreme learning machine (GAWELM). *Expert Systems with Applications, 40*(10), 3984–3993.

Avci, E., & Coteli, R. (2012). A new automatic target recognition system based on wavelet extreme learning machine. *Expert Systems with Applications, 39*(16), 12340–12348.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K. (2011). Single-trial analysis and classification of ERP components – A tutorial. *NeuroImage, 56*(2), 814–825.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Müller, K. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine, 25*(1), 41–56.

Chen, F., & Ou, T. (2011). Sales forecasting system based on gray extreme learning machine with Taguchi method in retail industry. *Expert Systems with Applications, 38*(3), 1336–1345.

Chen, X., Wang, Y., Nakanishi, M., Gao, X., Jung, T.-P., & Gao, S. (2015). High-speed spelling with a noninvasive brain-computer interface. *Proceedings of the National Academy of Sciences of the United States of America, 112*(44), E6058–E6067.

Chen, X., Wang, Z., & McKeown, M. (2016). Joint blind source separation for neurophysiological data analysis: Multiset and multimodal methods. *IEEE Signal Processing Magazine, 33*(3), 86–107.

Chen, X., Yang, J., & Chen, L. (2014). An improved robust and sparse twin support vector regression via linear programming. *Soft Computing, 18*(12), 2335–2348.

Chen, X., Yang, J., Ye, Q., & Liang, J. (2011). Recursive projection twin support vector machine via within-class variance minimization. *Pattern Recognition, 44*(10), 2643–2655.

Cong, F., Lin, Q., Kuang, L., Gong, X., Astikainen, P., & Ristaniemi, T. (2015). Tensor decomposition of EEG signals: A brief review. *Journal of Neuroscience Methods, 248*, 59–69.

Das, A., Suresh, S., & Sundararajan, N. (2016). A discriminative subject-specific spatio-spectral filter selection approach for EEG based motor-imagery task classification. *Expert Systems with Applications, 64*, 375–384.

García-Laencina, P. J., Rodríguez-Bermudez, G., & Roca-Dorda, J. (2014). Exploring dimensionality reduction of EEG features in motor imagery task classification. *Expert Systems with Applications, 41*(11), 5285–5295.

Huang, G., Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks, 61*, 32–48.

Huang, G.-B. (2003). Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transations on Neural Networks, 14*(2), 274–281.

Huang, G.-B. (2014). An insight into extreme learning machines: Random neurons, random features and kernels. *Cognitive Computation, 6*(3), 376–390.

Huang, G.-B., Chen, L., & Siew, C. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks, 17*(4), 879–892.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, Cybernetics, PartB: Cybernetics, 42*(2), 513–529.

Huang, G.-B., Zhu, Q., & Siew, C. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1), 489–501.

Iosifidis, A., Tefas, A., & Pitas, I. (2015). On the kernel extreme learning machine classifier. *Pattern Recognition Letters, 54*, 11–17.

Jiao, Y., Zhang, Y., Wang, Y., Wang, B., Jin, J., & Wang, X. (2017). A novel multilayer correlation maximization model for improving CCA-based frequency recognition in SSVEP brain-computer interface. *International Journal of Neural Systems*. In press.

Jin, J., Daly, I., Zhang, Y., Wang, X., & Cichocki, A. (2014). An optimized ERP brain–computer interface based on facial expression changes. *Journal of Neural Engineering, 11*(3), 036004.

Jin, J., Sellers, E., Zhou, S., Zhang, Y., Wang, X., & Cichocki, A. (2015). A P300 brain-computer interface based on a modification of the mismatch negativity paradigm. *International Journal of Neural Systems, 25*(3), 1550011.

Krusienski, D., Sellers, E., Cabestaing, F., Bayoudh, S., McFarland, D., Vaughan, T., & Wolpaw, J. (2006). A comparison of classification techniques for the P300 speller. *Journal of Neural Engineering, 3*(4), 299–305.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Li, Y., Pan, J., Long, J., Yu, T., Wang, F., Yu, Z., & Wu, W. (2016). Multimodal bcis: Target detection, multidimensional control, and awareness evaluation in patients with disorder of consciousness. *Proceedings of IEEE, 104*(2), 332–352.

Li, Y., Yu, Z., Bi, N., Xu, Y., Gu, Z., & Amari, S. (2014). Sparse representation for brain signal processing: A tutorial on methods and applications. *IEEE Signal Processessing Magazine, 31*(3), 96–106.

Liu, K., Yu, Z. L., Wu, W., Gu, Z., & Li, Y. (2015). Straps: A fully data-driven spatio-temporally regularized algorithm for M/EEG patch source imaging. *International Journal of Neural Systems, 25*(4), 1550016.

Liu, X., Wang, L., Huang, G.-B., Zhang, J., & Yin, J. (2015). Multiple kernel extreme learning machine. *Neurocomputing, 149*, 253–264.

Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for eeg-based brain-computer interfaces. *Journal of Neural Engineering, 4*(2), R1–R13.

Lotte, F., & Guan, C. (2011). Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms. *IEEE Transactins on Biomedical Engineering, 58*(2), 355–362.

Luo, J., Vong, C., & Wong, P. (2014). Sparse bayesian extreme learning machine for multi-classification. *IEEE Transactions on Neural Networks and Learning Systems, 25*(4), 836–843.

Ma, J., Zhang, Y., Cichocki, A., & Mastuno, F. (2015). A novel EOG/EEG hybrid human-machine interface adopting eye movements and ERP: application to robot control. *IEEE Transactions on Biomedical Engineering, 62*(3), 876–889.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). EEG signal classification for BCI applications by wavelets and interval type-2 fuzzy logic systems. *Expert Systems with Applications, 42*(9), 4370–4380.

Pal, M., Maxwell, A., & Warner, T. (2013). Kernel-based extreme learning machine for remote-sensing image classification. *Remote Sensing Letters, 4*(9), 853–862.

Pfurtscheller, G., Brunner, C., Schlögl, A., & Lopes, S. (2006). Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage, 31*(1), 153–159.

Qiu, Z., Allison, B., Jin, J., Zhang, Y., Wang, X., Li, W., & Cichocki, A. (2017). Optimized motor imagery paradigm based on imagining chinese characters writing movement. *IEEE Transactions on Neural Systems and Rehabilitaion Engineering, 25*(7), 1009–1017.

Shi, T., Wang, H., & Zhang, C. (2015). Brain computer interface system based on indoor semi-autonomous navigation and motor imagery for unmanned aerial vehicle control. *Expert Systems with Applications, 42*(9), 4196–4206.

Shin, Y., Lee, S., Lee, J., & Lee, H. (2012). Sparse representation-based classification scheme for motor imagery-based brain-computer interface systems. *Journal of Neural Engineering, 9*(5), 056002.

da Silveira, T. L., Kozakevicius, A. J., & Rodrigues, C. R. (2016). Automated drowsiness detection through wavelet packet analysis of a single eeg channel. *Expert Systems with Applications, 55*, 559–565.

Tang, J., Deng, C., & Huang, G.-B. (2016). Extreme learning machine for multilayer perceptron systems. *IEEE Transactions on Neural Networks and Learning Systems, 27*(4), 809–821.

Wang, H., Zhang, Y., Waytowich, N., Krusienski, D., Zhou, G., Jin, J., et al. (2016). Discriminative feature extraction via multivariate linear regression for SSVEP-based BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24*(5), 532–541.

Wang, J. J.-Y., Huang, J. Z., Sun, Y., & Gao, X. (2015). Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization. *Expert Systems with Applications, 42*(3), 1278–1286.

Wang, J., Wang, Q., Peng, J., Nie, D., Zhao, F., Kim, M., et al. (2017). Multitask diagnosis for autism spectrum disorders using multimodality features: A multicenter study. *Human Brain Mapping, 38*(6), 3081–3097.

Wang, S., Chen, H., Yan, W., Chen, Y., & Fu, X. (2014). Face recognition and micro–expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Processing Letters, 39*(1), 25–43.

Wolpaw, J., Birbaumer, N., McFarland, D., Pfutscheller, G., & Vaughan, T. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*(6), 767–791.

Wong, K., Vong, C., Wong, P., & Luo, J. (2015). Sparse Bayesian extreme learning machine and its application to biofuel engine performance prediction. *Neurocomputing, 149*, 397–404.

Wu, W., Chen, Z., Gao, S., & Brown, E. (2011). A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG. *NeuroImage, 56*(4), 1929–1945.

Wu, W., Chen, Z., Gao, X., Li, Y., Brown, E., & Gao, S. (2015). Probabilistic common spatial patterns for multichannel EEG analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 639–653.

Wu, W., Wu, C., Gao, S., Liu, B., Li, Y., & Gao, X. (2014). Bayesian estimation of ERP components from multicondition and multichannel EEG. *NeuroImage, 88*, 319–339.

Yeh, C. Y., Su, W. P., & Lee, S. J. (2013). An efficient multiple-kernel learning for pattern classification. *Expert Systems with Applications, 40*(9), 3491–3499.

Yu, T., Li, Y., Long, J., & Gu, Z. (2012). Surfing the internet with a BCI mouse. *Journal of Neural Engineering, 9*(3), 036012.

Yu, T., Xiao, J., Wang, F., Zhang, R., Gu, Z., Cichocki, A., & Li, Y. (2015). Enhanced motor imagery training using a hybrid BCI with feedback. *Journl of Neuroscience Methods, 62*(7), 1706–1717.

Zhang, R., Lan, Y., Huang, G., & Xu, Z. (2012). Universal approximation of extreme learning machine with adaptive growth of hidden nodes. *IEEE Transactions on Neural Networks and Learning Systems, 23*(2), 365–371.

Zhang, Y., Jin, J., Wang, X., & Wang, Y. (2016). Motor imagery EEG classification via bayesian extreme learning machine. In *Information science and technology (ICIST), 2016 sixth international conference on* (pp. 27–30). IEEE.

Zhang, Y., Wang, Y., Jin, J., & Wang, X. (2017). Sparse bayesian learning for obtaining sparsity of EEG frequency bands based feature vectors in motor imagery classification. *International Journal of Neural Systems, 27*(2), 1650032.

Zhang, Y., Zhou, G., Jin, J., Wang, X., & Cichocki, A. (2014). Frequency recognition in SSVEP-based BCI using multiset canonical correlation analysis. *International Journal of Neural Systems, 24*(3), 1450013.

Zhang, Y., Zhou, G., Jin, J., Wang, X., & Cichocki, A. (2015). Optimizing spatial patterns with sparse filter bands for motor-imagery based brain-computer interface. *Journal of Neuroscience Methods, 255*, 85–91.

Zhang, Y., Zhou, G., Jin, J., Zhang, Y.-S., Wang, X., & Cichocki, A. (2017). Sparse bayesian multiway canonical correlation analysis for EEG pattern recognition analysis. *Neurocomputing, 225*, 103–110.

Zhang, Y., Zhou, G., Jin, J., Zhao, Q., Wang, X., & Cichocki, A. (2016). Sparse Bayesian classification of EEG for brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems, 27*(11), 2256–2267.

Zhang, Y., Zhou, G., Zhao, Q., Cichocki, A., & Wang, X. (2016). Fast nonnegative tensor factorization based on accelerated proximal gradient and low-rank approximation analysis. *Neurocomputing, 198*, 148–154.

Zhao, Q., Zhou, G., Adali, T., Zhang, L., & Cichocki, A. (2013). Kernelization of tensor-based models for multiway data analysis: Processing of multidimensional structured data. *IEEE Signal Processing Magazine, 30*(4), 137–148.

Zhang, Y., Zhou, G., Jin, J., Zhao, Q., Wang, X., & Cichocki, A. (2014). Aggregation of sparse linear discriminant analysis for event-related potential classification in brain-computer interface. *International Journal of Neural Systems, 24*(1), 1450003.

Zhou, G., Cichocki, A., Zhao, Q., & Xie, S. (2014). Nonnegative matrix and tensor factorizations: An algorithmic perspective. *IEEE Signal Processing Magazine, 31*(3), 54–65.

Zhou, G., Zhao, Q., Zhang, Y., Adalı, T., Xie, S., & Cichocki, A. (2016). Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proceedings of the IEEE, 104*(2), 310–331.

Zhu, X., Suk, H. I., Huang, H., & Shen, D. (2017). Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers. *IEEE Transactions on Big Data, 3*(4), 405–414.

Zhu, X., Xie, Q., Zhu, Y., Liu, X., & Zhang, S. (2015). Multi-view multi-sparsity kernel reconstruction for multi-class image classification. *Neurocomputing, 169*, 43–49.